Robust Methods in Biostatistics

Stephane Heritier

The George Institute for International Health, University of Sydney, Australia

Eva Cantoni

Department of Econometrics, University of Geneva, Switzerland

Samuel Copt

Merck Serono International, Geneva, Switzerland

Maria-Pia Victoria-Feser

HEC Section, University of Geneva, Switzerland

A John Wiley and Sons, Ltd, Publication

Robust Methods in Biostatistics

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

A complete list of the titles in this series appears at the end of this volume.

Robust Methods in Biostatistics

Stephane Heritier

The George Institute for International Health, University of Sydney, Australia

Eva Cantoni

Department of Econometrics, University of Geneva, Switzerland

Samuel Copt

Merck Serono International, Geneva, Switzerland

Maria-Pia Victoria-Feser

HEC Section, University of Geneva, Switzerland

A John Wiley and Sons, Ltd, Publication

This edition first published 2009

Registered office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

Library of Congress Cataloging-in-Publication Data

Robust methods in biostatistics / Stephane Heritier … [et al.].
p. cm.
Includes bibliographical references and index.
ISBN 978-0-470-02726-4 (cloth)
1. Biometry–Statistical methods. I. Heritier, Stephane.
[DNLM: 1. Biometry–methods. WA 950 R667 2009]
QH323.5.R615 2009
570.1'5195–dc22

2009008863

A catalogue record for this book is available from the British Library.

ISBN 9780470027264

Set in 10/12pt Times by Sunrise Setting Ltd, Torquay, UK.
Printed in Great Britain by CPI Antony Rowe, Chippenham, Wiltshire.

To Anna, Olivier, Cassandre, Oriane, Sonia,
Johannes, Véronique, Sébastien and Raphaël,
who contributed in their ways…

Contents

Preface

The use of statistical methods in medicine, genetics and more generally in health sciences has increased tremendously in the past two decades. More often than not, a parametric or semi-parametric model is used to describe the data and standard estimation and testing procedures are carried out. However, the validity and good performance of such procedures generally require strict adherence to the model assumptions, a condition that is in stark contrast with experience gained from field work. Indeed, the postulated models are often chosen because they help to understand a phenomenon, not because they fit *exactly* the data at hand. Robust statistics is an extension of classical statistics that specifically takes into account the fact that the underlying models used by analysts are only approximate. The basic philosophy of robust statistics is to produce statistical procedures that are stable with respect to small changes in the data or to small model departures. These include 'outliers', influential observations and other more sophisticated deviations from the model or model misspecifications.

There has been considerable work in robust statistics in the last forty years following the pioneering work of Tukey (1960), Huber (1964) and Hampel (1968) and the theory now covers all models and techniques commonly used in biostatistics. However, the lack of a simple introduction of the basic concepts, the absence of meaningful examples presented at the appropriate level and the difficulty in finding suitable implementation of robust procedures other than robust linear regression have impeded the development and dissemination of such methods. Meanwhile, biostatisticians continue to use 'ad-hoc' techniques to deal with outliers and underestimate the impact of model misspecifications. This book is intended to fill the existing gap and present robust techniques in a consistent and understandable manner to all researchers in the health sciences and related fields interested in robust methods. Real examples chosen from the authors' experience or for their relevance in biomedical research are used throughout the book to motivate robustness issues, explain the central ideas and concepts, and illustrate similarities and differences with the classical approach. This material has previously been tested in several short and regular courses in academia from which valuable feedback has been gained. In addition, the R-code and data used for all examples discussed in the book are available on the supporting website (http://www.wiley.com/go/heritier). The data-based approach presented here makes it possible to acquire both the conceptual framework and practical tools for not only a good introduction but also a practical training in robust methods for a large spectrum of statistical models.

The book is organized as follows. Chapter 1 pitches robustness in the history of statistics and clarifies what it is supposed to do and not to do. Concepts and results are introduced in a general framework in Chapter 2. This chapter is more formalized as it presents the ideas and the results in their full generality. It presents in a more mathematical manner the basic concepts and statistical tools used throughout the book, to which the interested reader can refer when studying a particular model presented in one of the following chapters. Fundamental tools such as the influence function, the breakdown point and M-estimators are defined here and illustrated through examples. Chapters 3 to 7 are structured by model and include specific elements of theory but the emphasis is on data analysis and interpretation of the results. These five chapters deal respectively with robust methods in linear regression, mixed linear models, generalized linear models, marginal longitudinal data models, and models for survival analysis. Techniques presented in this book focus in particular on estimation, uni- and multivariate testing, model selection, model validation through prediction and residual analysis, and diagnostics. Chapters can be read independently of each other but starting with linear regression (Chapter 3) is recommended. A short introduction to the corresponding classical procedures is given at the beginning of each chapter to facilitate the transition from the classical to the robust approach. It is however assumed that the reader is reasonably familiar with classical procedures. Finally, some of the computational aspects are discussed in the appendix.

The intended audience for this book includes: biostatisticians who wish to discover robust statistics and/or update their knowledge with the more recent developments; applied researchers in medical or health sciences interested in this topic; advanced undergraduate or graduate students acquainted with the classical theory of their model of interest; and also researchers outside the medical sciences, such as scientists in the social sciences, psychology or economics. The book can be read at different levels. Readers mainly interested in the potential of robust methods and their applications in their own field should grasp the basic statistical methods relevant to their problem and focus on the examples given in the book. Readers interested in understanding the key underpinnings of robust methods should have a background in statistics at the undergraduate level and, for the understanding of the finer theoretical aspects, a background at the graduate level is required. Finally, the datasets analyzed in this book can be used by the statistician familiar with robustness ideas as examples that illustrate the practice of robust methods in biostatistics. The book does not include all the available robust tools developed so far for each model, but rather a selected set that has been chosen for its practical use in biomedical research. The emphasis has been put on choosing only one or two methods for each situation, the methods being selected for their efficiency (at different levels) and their practicality (i.e. their implementation in the R package `robustbase`), hence making them directly available to the data analyst. This book would not exist without the hard work of *all* the statisticians who have contributed directly or indirectly to the development of robust statistics, not only the ones cited in this book but also those that are not.

Acknowledgements

We are indebted to Elvezio Ronchetti and Chris Field for stimulating discussions, comments on early versions of the manuscript and their encouragement during the writing process, to Tadeusz Bednarski for valuable exchanges about robust methods for the Cox model and for providing its research code, and to Steve Portnoy for his review of the section on the censored regression quantiles. We also thank Sally Galbraith, Serigne Lo and Werner Stahel for reading some parts of the manuscript and for giving useful comments, Dominique Couturier for his invaluable help in the development of R code for the regression model and the mixed linear model, and Martin Mächler and Andreas Ruckstuhl for implementing the robust GLM in the `robustbase` package. The GFR data have been provided by Judy Simpson and the cardiovascular data by Pascal Bovet. Finally, we would like to thank the staff at Wiley for their support, as well as our respective institutions, our understanding colleagues and students who had to endure our regular 'blackouts' from daily work during the writing process of this book.

1

Introduction

1.1 What is Robust Statistics?

> The scientific method is a set of principles and procedures for the systematic pursuit of knowledge involving the recognition and formulation of a problem, the collection of data through observation and experiment, and the formulation and testing of hypotheses (*Merriam-Webster online dictionary, http://merriam-webster.com*).

Although procedures may be different according to the field of study, scientific researchers agree that hypotheses need to be stated as explanations of phenomena, and experimental studies need to be designed to test these hypotheses. In a more philosophical perspective, the hypothetico-deductive model for scientific methods (Whewell, 1837, 1840) was formulated as the following four steps: (1) characterizations (observations, definitions and measurements of the subject of inquiry); (2) hypotheses (theoretical, hypothetical explanations of observations and measurements of the subject); (3) predictions (possibly through a model, logical deduction from the hypothesis or theory); (4) experiments (test (2) and (3), essentially to disprove them). It is obvious that statistical theory plays an important role in this process. Not only are measurements usually subject to uncertainty, but experiments are also set using the theory of experimental designs and predictions are often made through a statistical model that accounts for the uncertainty or the randomness of the measurements. As statisticians, however, we are aware that models can at best be approximated (at least for the random part), and this introduces another type of uncertainty into the process. G. E. P. Box's famous citation that 'all models are wrong, some models are useful' (Box, 1979) is often cited by the researcher when faced with the data to analyze. Hence, for truly honest scientific

Robust Methods in Biostatistics S. Heritier, E. Cantoni, S. Copt and M.-P. Victoria-Feser

research, statistics should offer methods that not only deal with uncertainty of the collected information (sampling error), but also with the fact that models are at best an approximation of reality. Consequently, statistics should be in 'some sense' *robust* to model misspecifications. This is important since the aim of scientific research is the pursuit of knowledge that is used *in fine* to improve the wellbeing of people as is obviously the case, for example, in medical research.

Robust methods date back to the prehistory of statistics and they naturally start with outlier detection techniques and the subsequent treatment of the data. Mathematicians of the 18th century such as Bernoulli (1777) were already questioning the appropriateness of rejection rules, a common practice among astronomers of the time. The first formal rejection rules are suggested in the second part of the 19th century; see Hampel *et al.* (1986, p. 34), for details. Student (1927) proposes repetition (additional observations) in the case of outliers, combined with rejection. Independently, the use of mixture models and simple estimators that can partly downweight observations appears from 1870 onwards; see Stone (1873); Edgeworth (1883); Newcomb (1886) and others. Newcomb even imagines a procedure that can be posthumously described as a sort of one-step Huber estimator (see Stigler, 1973). These attempts to reduce the influence of outliers, to make them harmless instead of discarding them, are in the same spirit as modern robustness theory; see Huber (1972); Harter (1974–1976); Barnett and Lewis (1978) and Stigler (1973). The idea of a 'supermodel' is proposed by Pearson (1916) who embedded the normal model that gained a central role at the turn of the 20th century into a system of Pearson curves derived from differential equations. The curves are actually distributions where two additional parameters are added to 'accommodate' most deviations from normality. The discovery of the drastic instability of the test for equality of variance by Pearson (1931) sparked the systematic study of the non-robustness of tests. Exact references on these developments can be found in Hampel *et al.* (1986, pp. 35–36).

The term *robust* (strong, sturdy, rough) itself appears to have been proposed in the statistical literature by Box (1953). The field of modern *robust statistics* finally emerged with the pioneering works of Tukey (1960), Huber (1964) and Hampel (1968), and has been intensively developed ever since. Indeed, a rough bibliographic search in the *Current Index to Statistics*[1] revealed that since 1960 the number of articles having the word 'robust' in their title and/or in their keywords list has increased dramatically (see Figure 1.1). Compared with other well-established keywords, 'robust' appears to be quite popular: roughly half as popular as 'Bayesian' and 'design', but more popular than 'survival', 'bootstrap', 'rank' and 'smoothing'. Is robust statistics really as popular as it appears to be, in that it is used fairly routinely in practical data analysis? We do not really believe so. It might be that the word 'robust' is associated with other keywords such as 'rank', 'smoothing' or 'design' because of the perceived nature of these methods or procedures. We also performed a rough bibliographic search under the same conditions as before, but with the combination of the words 'robust' and each of the other words. The result is presented in Figure 1.2. It appears that although 'robust' is relatively more associated

[1] http://www.statindex.org/

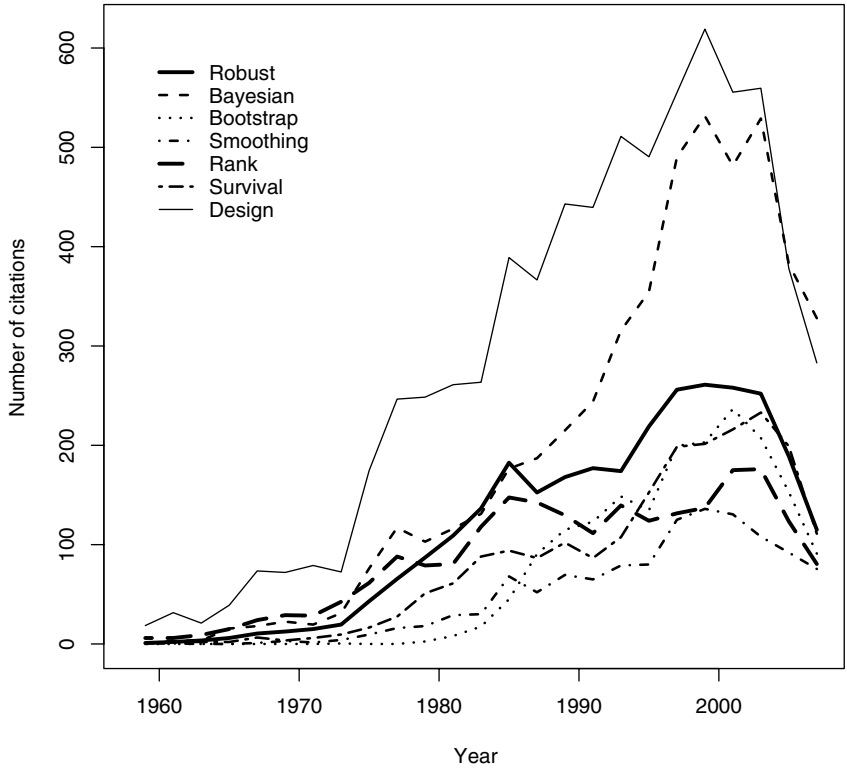

Figure 1.1 Number of articles (average per 2 years) citing the selected words in the title or in the keywords list according to the *Current Index to Statistics* (http://www.statindex.org/), December 2007.

with 'design' and 'Bayesian', when we remove all of the combined associations there are 4367 remaining articles citing the word 'robust' (group 'other'), a fairly large number.

We believe that this rather impressive number of articles have often used the term 'robust' in quite different manners. At this point, it could be worth searching more deeply, for example by taking a sample or articles and looking at the possible meanings or uses of the statistical term 'robust', but we do not attempt that here. Instead, we will clarify in what sense we use the term 'robust' or 'robustness' in the present book. We hope that this will help in clarifying the extent and limitations of the theory of robust statistics for the scientist as set by Tukey (1960), Huber (1964) and Hampel (1968).

1.2 Against What is Robust Statistics Robust?

Robust statistics aims at producing consistent and reasonably efficient estimators, test statistics with stable level and power, when the model is slightly misspecified.

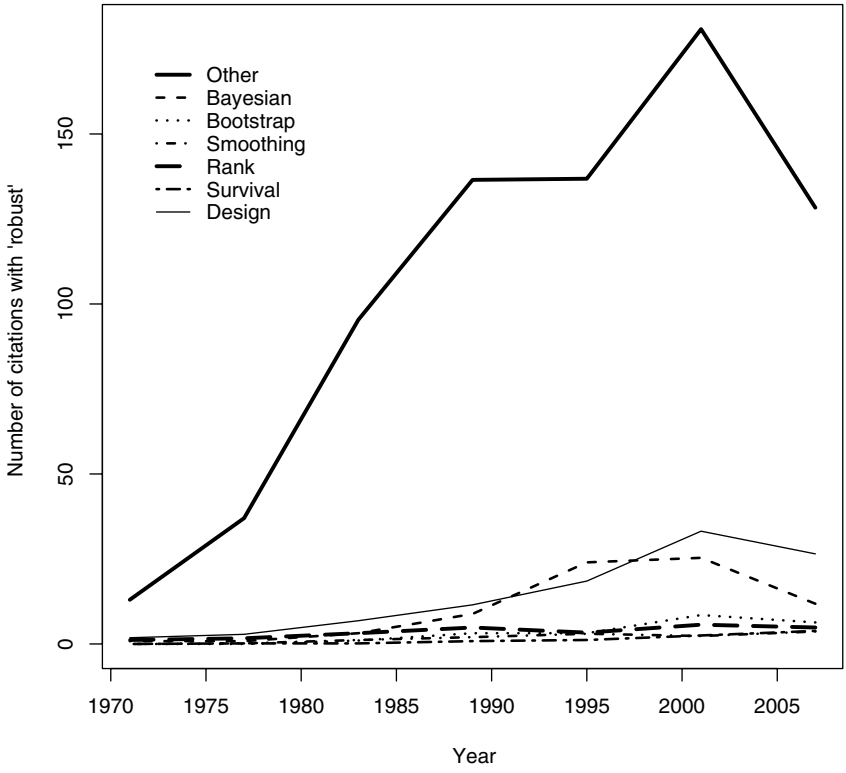

Figure 1.2 Number of articles (average per 6 years) citing the selected words together with 'robust' in the title or in the keywords list according to the *Current Index to Statistics* (http://www.statindex.org/), December 2007.

Model misspecifications encompass a relatively large set of possibilities, and robust statistics cannot deal with all types of model misspecifications. First we characterize the model using a cumulative probability distribution $F_{\boldsymbol{\theta}}$ that captures the structural part as well as the random part of the model. The parameters needed for the structural part and/or the random part are included in the parameter's vector $\boldsymbol{\theta}$. For example, in the regression model that is thoroughly studied in Chapter 3, $\boldsymbol{\theta}$ contains the (linear) regression coefficients (structural part) as well as the residual error variance (random part) and $F_{\boldsymbol{\theta}}$ is the (conditional) normal distribution of the response variable (given the set of explanatory variables). Here $F_{\boldsymbol{\theta}}$ does not need to be fully parametric, e.g. the Cox model presented in Chapter 7 can also be used. Then, by 'slight model misspecification', we assume that the data-generating process lies in a neighborhood of the true (postulated) model $F_{\boldsymbol{\theta}}$ that is considered as 'useful' for the problem under investigation. This notion of a neighborhood, due originally to Huber (1964),

is formalized as

$$F_\varepsilon = (1-\varepsilon)F_{\boldsymbol{\theta}} + \varepsilon G, \tag{1.1}$$

where $F_{\boldsymbol{\theta}}$ is the postulated model, $\boldsymbol{\theta}$ is a set of parameters of interest, G is an arbitrary distribution and $0 \leq \varepsilon \leq 1$.[2] The form of G is not really important, but there are some interesting special cases. For example, G can be a gross error-generating process (or point mass distribution), i.e.

$$G(x) = \Delta_z(x) = \begin{cases} 0 & x < z, \\ 1 & x \geq z. \end{cases} \tag{1.2}$$

In other words, data generated from F_ε are from $F_{\boldsymbol{\theta}}$ with probability $1-\varepsilon$ and from G with probability ε. Since G is an arbitrary distribution, and $\varepsilon \in (0, 1)$, the neighborhood F_ε is very general. However, the crucial quantity is actually ε, which in a sense measures the 'amount' of model misspecification. When $\varepsilon = 0$, then there is no model misspecification in that the data-generating process is exactly the postulated model. This is the fundamental hypothesis in classical estimation based, for example, on the maximum likelihood estimator (MLE) and classical testing based, for example, on the F-test in analysis of variance (ANOVA). For a data analysis practitioner, experience shows that the chosen 'useful' model ($F_{\boldsymbol{\theta}}$) is very rarely equal to the data-generating process (F_ε). Assuming that the data analyst does not ignore this fact, he/she is faced with the problem of 'what to do next'. There exist many 'practical strategies' that have been developed over the years to process the data in an *ad-hoc* fashion or tweak the model to ultimately resort to classical inference. Most of the strategies may fail in that what is sought is not necessarily what is found. Indeed, when $0 < \varepsilon < 1$, the situation becomes murkier. If one truly believes that the data-generating process F_ε is the true model, then inference should be carried out at F_ε. A mixture distribution should then be used, assuming that G can be chosen adequately. For instance, in injury prevention, researchers are interested in modeling the number of crashes involving young drivers or the number of serious injuries. Such outcomes often display a large number of zeros and are typically modeled by a zero-inflated Poisson or negative binomial distribution (see, e.g., Lambert, 1992) or alternatively using hurdle, two-step or conditional models (see, e.g., Welsh *et al.*, 1996). Inference is at the zero-inflated model (or one of the other models), represented as a mixture model with ε, as the probability of an observation being part of the (excess) spike at zero, commonly described on the logistic scale through a set of covariates. This is a reasonable model if ε is relatively large, but there is no guarantee that the resulting mixture (after choosing G) is the exact data-generating process. Inference is then sought simultaneously on $\boldsymbol{\theta}$, ε and the other parameters (or directly on the quantiles of G if it is a non-parametric model). The

[2]Note that (1.1) is not exactly a neighborhood in the mathematical sense. However, Huber's idea was to imagine a workable set of distributions that were 'close enough' to the assumed model, hence the use of the term neighborhood. He proved that any distribution in (1.1) is within a distance ε of $F_{\boldsymbol{\theta}}$ for a proper metric on the distribution space such as the Kolmogorov or Prohorov distance; see Huber (1981, Chapter 2).

procedure can become very cumbersome because the number of parameters can become very large.

If ε is relatively small and G is not obvious to define, then another strategy should be chosen. Indeed, very often $F_{\boldsymbol{\theta}}$ is chosen because it makes sense with respect to the problem under investigation, so that another model is less 'interpretable'. Moreover, when focusing on $F_{\boldsymbol{\theta}}$ and $\varepsilon > 0$, it is very difficult to define G in order to use F_ε as the 'true' model. In practice, discovering the form of G from the data is often impossible, unless the sample size is very large. Most of the time, the best 'guess' is $F_{\boldsymbol{\theta}}$ and the data can be assumed to have been generated approximately by $F_{\boldsymbol{\theta}}$. As stated previously, even if one has a fairly good idea about G, you still cannot be sure that the mixture $(1-\varepsilon)F_{\boldsymbol{\theta}} + \varepsilon G$ is the exact data-generating process. Finally, the mixture can be so complicated that one may wonder whether it is even worth using F_ε for small ε when one is actually interested in inference about $F_{\boldsymbol{\theta}}$.

Another situation (at least in theory) occurs when $\varepsilon = 1$. In this case it would make no sense to still seek inference about $F_{\boldsymbol{\theta}}$, so the postulated model should be changed to G. However, in generalized linear mixed models, for example, several authors have studied the 'robustness' of the Wald test to the assumption of normality of the random effects. Hence, in these cases, $F_{\boldsymbol{\theta}}$ is the mixed model with normal random effects, G is the mixed model with non-normal random effects and $\varepsilon = 1$. One then seeks inference about $\boldsymbol{\theta}$ when F_ε with large ε is the data-generating process. Some of the proposed procedures have been found to be 'robust' in some particular situations and for some specific distributions for the random effects;[3] see e.g. Litière *et al.* (2007b) for details. Although this type of robustness is also important, it is limited to some particular instances of G (i.e. for some distributions for the random effects). This actually concerns a type of model misspecification that can be called structural misspecification in that the form of G is known (and ε is large). The robust procedures we propose here are robust in the sense that inference remains correct at $F_{\boldsymbol{\theta}}$ even if F_ε is the data-generating process and ε is unknown but small and G can be of any form. The type of model misspecification in this case can be called distributional misspecification in that the best that can be said is that the data-generating process is approximately $F_{\boldsymbol{\theta}}$ (small ε).

Seeking inference about $F_{\boldsymbol{\theta}}$ when F_ε is the actual data-generating process is not the same as seeking inference about F_ε when $F_{\boldsymbol{\theta}}$ is fitted. Indeed, sometimes classical procedures (with possible added corrections) are said to be robust to model misspecification in the sense that the estimator of $\boldsymbol{\theta}$ (when $F_{\boldsymbol{\theta}}$ is fitted) still provides consistent estimators for (some of) the parameters of F_ε. For example, in the important case of the omission of covariates, we would have G (assuming that it exists) such that $F_\varepsilon = F_{(\boldsymbol{\theta},\boldsymbol{\theta}')}$ where $\boldsymbol{\theta}'$ is the added parameter corresponding to the missing predictors. This is another case of structural misspecification that is not covered by the robustness theory introduced in this book. In the 1980s there were some important studies of the conditions under which a consistent estimate of $\boldsymbol{\theta}$, assuming $F_{(\boldsymbol{\theta},\boldsymbol{\theta}')}$ (as the true model) but fitting $F_{\boldsymbol{\theta}}$ could still be obtained; see Gail *et al.* (1984) and Bretagnolle and Huber-Carol (1988) for instance. They essentially

[3]This type of 'robustness' is such that the level of the test is preserved under these assumptions.

Table 1.1 Models at which inference can at best be made.

Inference	G	$\varepsilon = 1$	$0 \ll \varepsilon < 1$	$0 < \varepsilon \ll 1$	$\varepsilon = 0$
Classical	Arbitrary	?	?	?	$F_{\boldsymbol{\theta}}$
	$G = \Delta_z$	?	F_ε	F_ε	$F_{\boldsymbol{\theta}}$
	G specified	G	F_ε	F_ε	$F_{\boldsymbol{\theta}}$
Robust	Arbitrary	?	?	$F_{\boldsymbol{\theta}}$	$F_{\boldsymbol{\theta}}$
	$G = \Delta_z$	?	F_ε	$F_{\boldsymbol{\theta}}$	$F_{\boldsymbol{\theta}}$
	G specified	G	F_ε	$F_{\boldsymbol{\theta}}$	$F_{\boldsymbol{\theta}}$

showed that a small residual bias remains although in some simple cases such as the linear model this situation does not occur.

In Table 1.1 we summarize some of the possible situations discussed so far regarding the postulated model, the data-generating process, the value of ε, the form of G and the estimation method. Except in the case when $\varepsilon = 0$, classical inference is not (at least *a priori*) suitable in most of the situations considered here. Moreover, even if theoretically one could postulate F_ε instead of $F_{\boldsymbol{\theta}}$, the former is often difficult to find and/or to estimate. The robust methods we propose in this book provide an alternative (and more effective) approach when ε is relatively small. We propose a set of statistical tools for correct estimation and inference about $F_{\boldsymbol{\theta}}$ when the data-generating process is F_ε, not only when $\varepsilon = 0$, as with classical methods, but also for relatively small ε and *any* G. As a by-product, data not fitting $F_{\boldsymbol{\theta}}$ exactly can be easily identified, and the model can possibly be changed and refitted.

Hence, one possibility is to manipulate the data so that they 'fit' the postulated model, but as argued below, this is not a good method. Another possibility is to change the model, but it is not always clear what a suitable alternative model may be. One could also use a more flexible model (e.g. using non-parametric statistics), but care should be taken as to what the underlying assumptions really are (see the discussion in Section 1.3). The alternative we propose here is to use robust statistics, which allows one to make inferences about $F_{\boldsymbol{\theta}}$, when the data-generating process is actually F_ε, with small ε and arbitrary G. We spend the remainder of this chapter explaining how robust statistics work in general to achieve these goals.

1.3 Are Diagnostic Methods an Alternative to Robust Statistics?

Since classical methods, when $F_{\boldsymbol{\theta}}$ is the postulated model, only work when $\varepsilon = 0$, one could be tempted to modify the data by removing 'dubious' observations from the sample. By 'dubious' observations we mean here that they are in some sense far from the bulk of the data generated by $F_{\boldsymbol{\theta}}$. The measure of how far an observation is from the bulk of the data is highly dependent on the problem (hence the model). For the problem of estimating the mean of a population, also

called the location problem, measures such as standard deviations are sometimes used to build thresholds (e.g. three standard deviations around the mean) outside which observations are considered as outliers. Hampel (1985) provides an account of several measures for the location problem and compares the properties of the mean estimator computed after the removal of outliers. In more complex situations, the measure can be based on graphical tools such as boxplots and/or scatterplots and constructed before the model is fitted. Alternatively, one could rely on some sort of 'residual analysis' (i.e. estimation of the random part of the model, once the estimated fixed part has been removed) for checking the distributional assumptions. More sophisticatedly, in regression models, the so-called 'diagnostic' techniques could be used (see, e.g. Atkinson, 1985; Belsley *et al.*, 1980; Cook and Weisberg, 1982). One such well-known tool is the Cook distance. The strategy of removing observations, although apparently simple, can be not only unpractical, but also very misleading. The main arguments are as follows.

- Graphical tools used before the model is fitted are only suitable for simple problems, such as when comparing groups (testing differences in mean responses), without control variables or in correlation settings (such as regression models) when there are only at most three variables (the response and two explanatory variable) if one uses three-dimensional graphs. When the dimension is higher, then combinations of (often pairs of) variables could be used, but multivariate effects could be masked.

- In practice, however, it is not always obvious how to quantify 'far from', and some observations might appear to be just at the (imaginary) border. Then the analyst is left with a rather arbitrary decision to make.

- Sometimes raw measures of 'outlyingness' are used that are based on standard deviations (e.g. remove observations that are three standard deviations away from the mean). This leaves the question of how the standard deviations (and the mean) are estimated open. The chosen scale estimator could be inflated and the mean itself biased by outlying observations generating a masking effect (see, e.g. Rousseeuw and Leroy, 1987).

- Moreover, when this is done in a univariate fashion, outlying observations are found only with respect to one variable at a time masking the effects of other covariates (see the example below).

- A 'residual analysis' can be used to detect 'outlying' observations once the model is fitted. This is commonly done in regression models. However, this procedure is not flawless as it does not take into account how the residuals are estimated. Indeed, if classical estimators for the model parameters are computed, they can be seriously biased by model deviations such as outliers. Hence, residuals obtained through these biased estimates will, in turn, be biased. This is another illustration of the masking effect. Removing observations on the basis of potentially biased residual estimates can become a very dangerous strategy.

- The same argument applies to other diagnostic tools based on classical estimators of the model parameters. Even if the diagnostic tools are based on the comparison of fitted models with and without one observation at a time such as the Cook distance, the simultaneous effect of multiple outliers could be masked.

- The 'data-cleaning' process can become very cumbersome in that one or some of the observations are removed on the basis of some criteria, then the model refitted, the criteria calculated again, new data are removed, etc. The process may never end at a satisfactory stage and a large proportion of data are removed before the process is stabilized. It is also unfeasible for large datasets.

- However, perhaps the most important argument is inference. A proper inferential procedure should take into account the data manipulation. In other terms, inference (e.g. significance tests) should be conditional on the criteria used for the removal of the observations. The use of classical inference (e.g. t-test) after case deletion and refit ignores this problem and is therefore dubious and, in some cases, completely wrong (see also Welsh and Ronchetti, 2002).

In our view the true purpose of diagnostic methods should be to identify genuine structural model misspecifications, e.g. adding a quadratic term, a missing covariate or an interaction in the model, identifying a systematic violation to the proportional hazard assumption in the Cox model or an incorrect formulation of the random component of a mixed linear model. They do not oppose robust methods, they are just complementary.

To illustrate the danger of a relatively naive data-cleaning process used before fitting, we consider the following dataset which will also be reanalyzed in Chapter 3 using robust techniques. The data (kindly provided by Dr Pascal Bovet, IUMSP, Lausanne, Switzerland) come from a study investigating the prevalence of hyperuricemia and the association of uric acid levels with various cardiovascular risk factors in the Seychelles Islands. A total of 998 participants from this population, mainly of African origin, were included in the study; see Conen *et al.* (2004). The primary outcome, serum uric (`uric`), is typically analyzed by linear regression with predictors such as the triglycerides level in body fat (`trig`) and the low-density lipoprotein (`ldl`) cholesterol; see Section 3.5 for a complete description of all covariates. Before the regression model is fitted, a descriptive data analysis should be performed. Boxplots of each of the variables can be drawn to detect extreme measurements, as well as scatterplots of pairs of variables to study their relationship and possibly detect outlying observations. In Figures 1.3 and 1.4 we present the scatterplots of `uric` versus `trig` and `ldl` versus `trig`, respectively. The vertical and horizontal lines represent the values of the sample means plus three standard deviations (i.e. the quantile 0.999 on the normal distribution) for each variable. As is routinely performed, a 'cleaning' mechanism based on this univariate criterion would remove from the sample all of the data represented by the points lying to the right and above these lines on both graphs. This rather rough mechanism does not take into account the possible correlation between the variables, especially between the

response (`uric`) and the explanatory variables. A more sophisticated but still rough mechanism is to consider as extreme values those with low probability under the bivariate normal model, or in other words, observations lying outside the quantiles with equal density (say with corresponding cumulative probability of 0.999) as illustrated by the ellipses on the scatter plots. To draw these ellipses, one needs to estimate the bivariate center and the covariance matrix between the pairs of variables. The classical estimators are the sample means, variances and Pearson correlation. If there are extreme observations (away from the bulk of data), these estimators can be artificially inflated, and this is the case with both examples in Figures 1.3 and 1.4. Alternatively, one can compute the ellipses based on a robust estimation of the center and covariance matrix between the pairs of variables (see Section 2.3.3), which, in the examples taken here, lead to better 'centered' ellipses (with respect to the bulk of data) of smaller volume. Using the ellipse-based criterion for 'cleaning' the data does not lead to the same decisions regarding the data to discard. In particular, observations that have not been removed with the three standard deviations would be removed with the classical ellipse, while one outside the three standard deviations on the `ldl` variable in Figure 1.4 would not be discarded with the robust ellipse. One can also notice that more observations would be discarded with the robust ellipses.

The difference in the appreciation of the 'outlyingness' of one observation between a univariate and a multivariate approach is due to the fact that the underlying model upon which the decisions are made is not the same. Indeed, with the ellipses, the correlation between the two variables is taken into account, which is not the case with a univariate approach. One could also consider multivariate criteria, i.e. criteria based on the relationship between all of the variables simultaneously. Such a criterion is given by the Mahalanobis distances (Mahalanobis, 1936) based on the multivariate normal assumption of the data; see (2.34). In the bivariate case as in Figures 1.3 and 1.4, the points on the same ellipses have equal Mahalanobis distances. Hence, a limit distance could be chosen and points with corresponding Mahalanobis distance exceeding this limit could be discarded from the sample. This would lead to the rejection of yet different observations. Even if this approach takes into account the relationships between all of the variables simultaneously, and hence is better than a univariate or bivariate approach, it is not satisfactory for the chosen example. Given that the data are actually used to explain the response variable `uric` through a regression model, extreme observations should be chosen with respect to the regression model, and this can only be done through a robust estimation of the latter. The complete robust analysis of the cardiovascular risk factors data by means of a regression model will be presented in Section 3.5. In this analysis, observations are down-weighted according to their degree of 'outlyingness'. In Figures 1.3 and 1.4 extreme observations (weight less than or equal to 0.3) with respect to the final regression model estimated in Table 3.10 have been drawn using the symbol **o**. The striking feature is that, although most of them correspond to observations that would have been discarded using one of the previous *ad-hoc* methods, they do not correspond to all of them (hence, more data are used for the estimators, and consequently inference is more powerful). More dramatically, some observations that would not have been removed with the *ad-hoc* method are strongly downweighted by

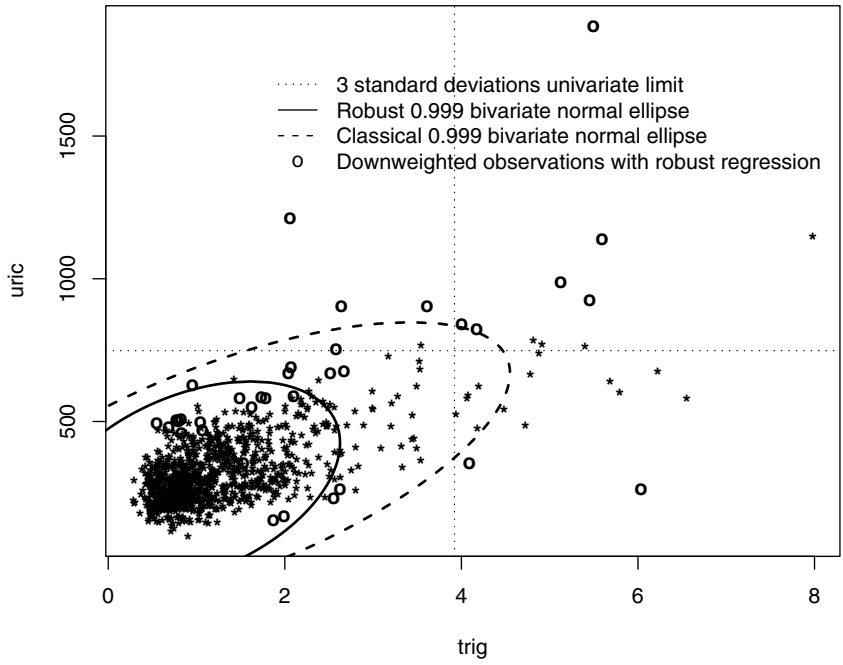

Figure 1.3 Scatterplot of the variables `uric` and `trig` for the cardiovascular risk factors data, together with the three standard deviations univariate limits, the robust and classical 0.975 bivariate normal contours. The symbol o denotes observations that have been downweighted by a robust regression analysis.

means of the robust regression estimator, questioning the validity of the procedure. In other words, the 'outlyingness' of an observation is relative to a model, and procedures that do not take this fact into account are not good procedures.

1.4 How do Robust Statistics Compare with Other Statistical Procedures in Practice?

Robust methods have not seen much use in clinical trials or epidemiological studies, except in a few cases such as Conen *et al.* (2004), Kurttio *et al.* (2005), Tashkin *et al.* (2006) and Wen *et al.* (2004). The only areas were the penetration of such methods is not anecdotal are medical imaging and genetics where robust regression (and smoothing techniques) are commonly used successfully; see, for instance, Ma and Elis (2003), Wager *et al.* (2003) and Jain *et al.* (2005). Apart from these particular

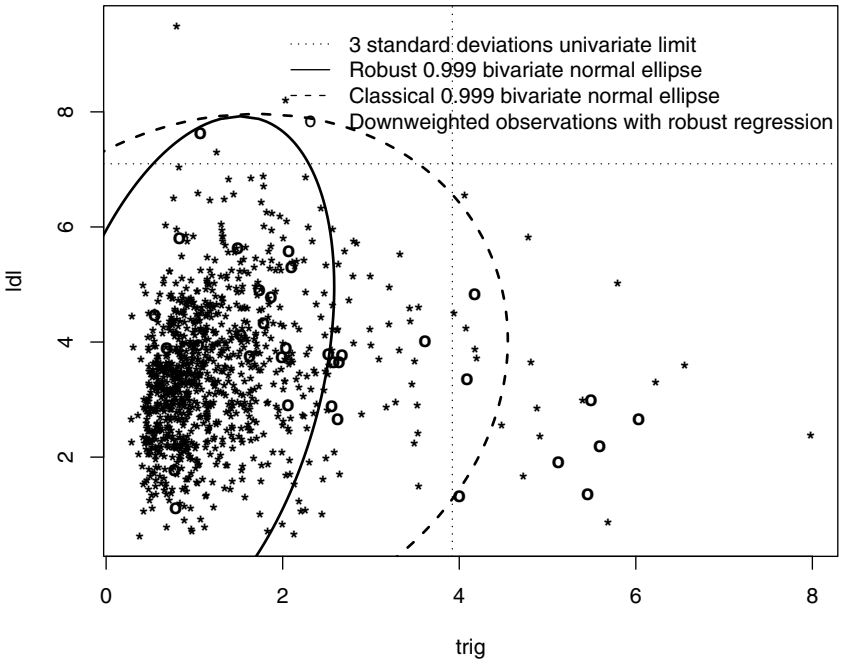

Figure 1.4 Scatterplot of the variables `trig` and `ldl` for the cardiovascular risk factors data, together with the three standard deviations univariate limits, the robust and classical 0.975 bivariate normal contours. The symbol o denotes observations that have been downweighted by a robust regression analysis.

cases, there seems to be a general feeling that outliers rarely occur in clinical trials, and if present they can properly be dealt with by means of traditional methods (e.g. rank-based techniques, described below). We believe that there is no reason to think that distributional model misspecifications are less present in clinical research than in any other area. If it is true that in regression settings covariates can be well controlled, extreme responses may nevertheless be present and hence ruin the interpretation of many standard procedures. A sense of false security arises as binary endpoints such as the occurrence of a specific event (e.g. death, disease progression, relapse) or the corresponding times to event are routinely studied. Even if procedures such as the chi-squared test typically used to analyze binary data or the log-rank test for survival times are less sensitive to extreme responses, one often forgets that a treatment estimate generally has to be given and some form of modeling assumed. If, as is done most of time, the Cox proportional hazard model is used, we show in Chapter 7

that the classical estimator for the hazard ratio based on the partial likelihood can be ruined by a small number of abnormal long-term survivors. In addition, if an adjusted analysis is performed as a second step, or if the outcome is simply continuous, the situation is about the same as any other area in biostatistics.

In clinical research, the case deletion and refit is not a satisfactory alternative to robust statistics, not only for the reasons discussed in Section 1.3, but also because such a procedure violates the intention-to-treat (ITT) principle. This principle (Hollis and Campbell, 1999) states that patients must be analyzed as randomized, irrespective of what actually happens and usually assumes that everybody randomized on the trial is included in the analyzed datasets. In contrast, robust techniques have solid theoretical underpinnings, protect against outliers and other model misspecifications and offer an elegant way to preserve the ITT principle by automatically downweighting extreme observations instead of deleting them in an *ad-hoc* fashion. In addition, there is no reason to think that results obtained from a robust fit (when appropriately used) will favor a particular outcome, say a positive effect of the drug under investigation, which is precisely what the guidelines request as a proper way to deal with outliers.

Rank-based methods (see e.g. Hettmansperger and McKean, 1998) are more sensible and this may explain why they are so popular for some types of analysis, such as survival analysis where the log-rank test is systematically used. However, they are not always available for more complex techniques, e.g. generalized estimating equations, complex models with covariates or mixed linear models. In addition, power issues have to be taken into consideration, an issue that is often overlooked. As described later in this book, it is often possible to calibrate the robust procedures we propose to achieve a pre-specified efficiency in the model (e.g. 90–95%). In other words the price to pay for the use of such techniques with respect to the MLE and related tests is a small loss in efficiency if the model holds.

It is also often believed that resampling methods such as the bootstrap can be used as an alternative to robust methods as 'one does not need to specify the distribution'. One should first recall that the bootstrap method of Efron (1982) (see also e.g. Davison and Hinkley, 1997) is a technique allowing the computation of standard errors, confidence intervals and p-values that are based on given estimators or test statistics. It thus does not provide new estimators or test statistics 'per se'. This method can be used for parametric, semi-parametric or even non-parametric analyses. What is understood behind this distribution-free assumption is that the sampling error distribution does not usually need to be a given parametric model (such as the normal distribution). One simply assumes instead that the observations are 'independent and identically distributed' (i.i.d.). The bootstrap and other non-parametric methods do not become naturally robust to model misspecification just because the model sampling error distribution is not specified.

We do not believe that the applied statistician has no 'model' in mind when stating the i.i.d. condition. Indeed, in order to summarize a group response by a 'mean parameter', e.g. the mean cholesterol per treatment arm in a statin[4] trial, it is

[4]Statins are drugs that improve blood cholesterol levels.

implicitly assumed that the response distribution is somewhat symmetric around this mean. Otherwise it would make little sense to summarize the outcome in this way but it would be more sensible to compare the (whole) response distribution across the treatment arm. In other words, even if the bootstrap provides good inference techniques without the need to specify a data-generating process (e.g. sampling error distribution), what is tested (choice of the parameters such as the 'mean' response) is not necessarily appropriate in all situations (e.g. bimodal distributions instead of symmetric distributions) and the conclusions can be very misleading.

Resampling techniques can be particularly sensitive to some types of model deviations such as outliers. Indeed, some of the bootstrap samples will invariably have a larger proportion of outliers and therefore heavily bias the estimators or test statistics computed on these samples. Confidence intervals or p-values derived from these bootstrapped statistics will then represent an 'average' between estimates (or tests statistics) computed from samples with different proportions of contaminated data (outliers). One might then wonder whether these confidence intervals (or p-values) are really informative, since they are representative of neither the 'clean' data nor the outliers. A standard bootstrap procedure may even fail when applied to a robust estimator as it may not necessarily withstand more than a certain proportion of outliers (i.e. the breakdown point of the procedure is reached). A solution to this problem would then be to use a robust bootstrap procedure applied to a robust estimator as originally suggested by Salibian-Barrera and Zamar (2002) in the linear regression model.

Finally, in non-parametric regression, it is often thought that robustness is automatically achieved, given that the approach relies on relaxed hypotheses (no normality assumption for the error term). This feeling is reinforced by the fact that the non-parametric regression estimators (smoothers) are local averages and it is therefore wrongly believed that an outlier occurring in a given subspace of the design only affects the estimation around this region. In fact, quite long ago Huber (1979) had already warned against the non-robustness of non-parametric regression and proposed a robust version of smoothing splines. There are also other alternatives, for example the M-kernels of Härdle (1990) or the Locally Weighted Regression and Smoothing Scatterplots (LOWESS) of Cleveland (1979). More recently, Cantoni and Ronchetti (2001a) have shown that the data-driven choice of the smoothing parameter pertaining to smoothers also needs to be made robust. They propose both a cross validation and a C_p-like criterion to cope with this issue.

To conclude, the robust methods we propose in this book are based on the specification of a core (parametric) model $F_{\boldsymbol{\theta}}$ such as, for example, the linear regression model, the mixed linear model, the generalized linear model (GLM), models for longitudinal data or a model which might contain some non-parametric parts such as the Cox model for survival data. We assume, however, that the data are generated by a distribution in a neighborhood (1.1). In order to avoid the potential bias on classical estimators, test statistics and other inferential procedures of this type of model misspecification, we propose instead the use of alternative robust statistics which provide correct inference at the core model $F_{\boldsymbol{\theta}}$.

2

Key Measures and Results

2.1 Introduction

Prior to the introduction of robust estimation and testing techniques for many common models used in biostatistics such as linear regression, mixed linear models, generalized linear models, models for longitudinal data and the Cox regression, it is important to lay the foundations of modern robust statistics. Hence, this chapter formalizes the concepts introduced in Chapter 1, reviews fundamental tools and presents key results that are used throughout the book. Historically, the development of a formal robustness theory only started in the 1960s although evidence that standard techniques lacked stability had been provided since the early days of statistics; see Section 1.1 and also Huber (1981, pp. 1–5) and Hampel *et al.* (1986, pp. 34–36), for details. Tukey (1960) and his group reignited the interest in such problems by proposing stable alternatives to the sample mean that is known to be badly affected by outliers. The pioneer work of Huber (1964) forms the first solid basis for a theory of robust estimation. In Huber's approach, the estimation problem is seen as a game between nature (which chooses a distribution in a neighborhood) and the statistician (who chooses an estimator in a given class). The statistician can achieve robustness by constructing an estimator which minimizes a loss criterion (such as the bias or the variance) in the worst possible case in the full neighborhood. Huber calls this approach the minimax problem and solves it in the class of M-estimators (see Section 2.3.1) for a location model, a simple model where only the central parameter (typically the mean for a normal model) has to be estimated. Despite the elegance of this theory, its extension to more complex models has proved challenging in general parametric models, in particular when no invariance structure is available. The key concept of M-estimators was soon extended to any parametric model by Huber (1967, 1981). The development of the influence function (IF) by Hampel (1968, 1974) (see Section 2.2.1) was another breakthrough in the development of the robustness theory that is available today. These tools paved the

Robust Methods in Biostatistics S. Heritier, E. Cantoni, S. Copt and M.-P. Victoria-Feser

way for a formal treatment of robust estimation in general, soon followed by the problem of robust inference. A summary of the pioneer work by Huber, Hampel and colleagues can be found in Huber (1981), Hampel *et al.* (1986) and Huber and Ronchetti (2009) or for a presentation at an intermediate level see Staudte and Sheather (1990). Comprehensive reviews on robust inference until the late 1990s are given by Markatou and Ronchetti (1997) or Ronchetti (1997a). A more recent reference on robust statistics is Maronna *et al.* (2006).

This chapter is organized as follows. We first review in Section 2.2 Hampel's *IF*, which generalizes the concept of a sensitivity curve. This function measures the asymptotic bias caused to an estimator by an infinitesimal contamination in a neighborhood of the assumed model. It is usually completed by the breakdown point that measures, loosely speaking, the percentage of contamination of the data is required to drive the estimator to any arbitrary value. We then introduce general approaches for robust estimation in Section 2.3. In particular, we show that M-estimators constitute a large class of consistent estimators that are convenient to work with. They have an asymptotically normal distribution, a simple *IF* and can subsequently lead to the construction of robust alternatives to the MLE or other classical estimators. Issues related to robust testing are considered next. The concepts of the *IF* and the breakdown point are extended to tests, first in the one-dimensional setting in Section 2.4, and then in the multidimensional case in Section 2.5. We then introduce extensions of the likelihood ratio, score and Wald tests based on M-estimators and show that the good robustness properties of these estimators can be carried over to the tests. Results are presented here in their generality, with the details of the specifics of robust estimation or testing techniques in a particular model being given in subsequent chapters. Finally, issues related to model selection are treated, when appropriate, in the corresponding chapters.

2.2 Statistical Tools for Measuring Robustness Properties

Robust statistics aims at producing consistent and possibly efficient estimators, test statistics with stable level and power and so forth, when the model is slightly misspecified. Before developing such procedures, one needs to introduce tools to formally assess the robustness properties of any statistical procedure with respect to model misspecification (or model deviation) that we have already defined through the concept of a neighborhood given in (1.1). Such tools could depend on the choice of the contaminating distribution G for given values of ε, be maximized over all possible G or be directly dependent on ε. In any case, any robustness measure will depend on the postulated model $F_{\boldsymbol{\theta}}$. The effect of model misspecifications on an estimator $\hat{\boldsymbol{\theta}}$ or an a test procedure is *a priori* a vague concept. While the case of testing is discussed in Section 2.4, one might ask on which estimator characteristic(s) should the effect of model misspecification be measured? Indeed, the effect could happen on the (asymptotic) distribution of $\hat{\boldsymbol{\theta}}$, through its first moment, second moment, etc. The theory developed so far focuses on the first asymptotic moment

of $\hat{\boldsymbol{\theta}}$, more precisely on the asymptotic bias caused by model misspecifications. Tools that target the asymptotic bias for small (infinitesimal values of) ε are said to measure infinitesimal or local robustness. Measures of the highest contamination ε that an estimator can tolerate without causing the asymptotic bias to be out of control assess global robustness. In this section, we present the most popular measures that have been used to assess the robustness properties of statistical procedures and build robust estimators or related tests that will be used in subsequent chapters.

2.2.1 The Influence Function

A simple way to assess the influence of an arbitrary point x on a particular statistic is to compute the difference between its value *with* observation x and its value *without* it. This is sometimes called the empirical influence function. As a better solution, one can standardize the difference through the proportion of contamination in the resulting sample, yielding the sensitivity curve $\mathrm{SC}_n(x)$ originally due to Tukey (1970). Assume that we initially have $n-1$ observations $x_1, x_2, \ldots, x_{n-1}$ from the normal model $\mathcal{N}(\mu, \sigma^2)$, so that by adding one arbitrary value x we have a sample of size n, then the sensitivity curve for the sample mean $\bar{x}_n := (x_1 + x_2 + \cdots + x_n)/n$ is

$$\mathrm{SC}_n(x) = \frac{(x_1 + x_2 + \cdots + x)/n - \bar{x}_{n-1}}{1/n} = x - \bar{x}_{n-1},$$

a linear function in x. When $n \to \infty$, $\mathrm{SC}_n(x)$ tends to $x - \mu$ almost everywhere and the limit is the 'asymptotic influence' of x on the sample mean. The extension of $\mathrm{SC}_n(x)$ to any statistics is straightforward but generally $\mathrm{SC}_n(x)$ is sample-dependent. To overcome this difficulty Hampel (1968, 1974) introduced the influence curve or *IF* by exploiting the fact that most estimators can actually be viewed as a functional (or function of a distribution). Let $F^{(n)}$ be the empirical distribution function of a sample of n observations $(x_1, \ldots, x_n)$ i.i.d., i.e

$$F^{(n)}(x) = \frac{1}{n} \sum_{i=1}^{n} \Delta_{x_i}(x)$$

and, as an illustration, consider $\hat{\boldsymbol{\theta}} = (x_1 + x_2 + \cdots + x_n)/n$, the sample mean. One can easily rewrite $\hat{\boldsymbol{\theta}}$ as $\hat{\boldsymbol{\theta}}(F^{(n)}) = \int x \, dF^{(n)}(x)$. Thus, the functional associated with the sample mean is $\hat{\boldsymbol{\theta}}(F) = \int x \, dF(x)$. A similar construction is generally possible for any estimator $\hat{\boldsymbol{\theta}}$ of the parameter $\boldsymbol{\theta}$, typically rewritten as $\hat{\boldsymbol{\theta}}(F)$.

Hampel's *IF* can be seen as a natural and elegant generalization of the sensitivity curve approach for functionals. The equivalent of the contaminated sample is the F_ε-neighborhood or 'gross error model'. It is obtained by choosing the contaminating distribution G in (1.1) as the point-mass distribution $\Delta_z(x)$ at a particular point z.[1] The contamination fraction in this gross error model is ε and replaces $1/n$.

[1] The point mass distribution was defined earlier in the univariate case. It can be extended to the multivariate case Δ_z which is the distribution for which $\Delta_z(A) = 1$ if the set A contains the q-vector z and zero otherwise.

In Hampel's definition, the *IF* measures the infinitesimal variation on the standardized difference $(\hat{\boldsymbol{\theta}}(F_\varepsilon) - \hat{\boldsymbol{\theta}}(F_{\boldsymbol{\theta}}))/\varepsilon$ when ε tends to zero ($\varepsilon \downarrow 0$), or more formally[2]

$$IF(z; \hat{\boldsymbol{\theta}}, F_{\boldsymbol{\theta}}) = \lim_{\varepsilon \downarrow 0} \frac{\hat{\boldsymbol{\theta}}(F_\varepsilon) - \hat{\boldsymbol{\theta}}(F_{\boldsymbol{\theta}})}{\varepsilon}, \tag{2.1}$$

with F_ε given in (1.1) and $G(x) = \Delta_z(x)$. The *IF* can simply be obtained by computing the right-sided derivative $(\partial/\partial\varepsilon)\hat{\boldsymbol{\theta}}(F_\varepsilon)$ at $\varepsilon = 0$ when the derivative exists. As an illustration, the *IF* of the sample mean at the normal model $F_{\boldsymbol{\theta}} = \mathcal{N}(\mu, \sigma^2)$ (i.e. with $\boldsymbol{\theta} = (\mu, \sigma^2)^T$) is simply $z - \mu$. Not surprisingly, *IF* coincides in that case with the limit of the sensitivity curve. As the normal model is symmetric around μ the sample median could also be considered as a possible (but inefficient) estimator of μ. Its *IF* is $\sigma \operatorname{sign}(z - \mu)/(2\varphi(0))$ where φ is the standard normal density function; see Huber (1981, p. 137). The fact that the *IF* is bounded illustrates the good local robustness property of the sample median. In Figure 2.1 we plot the *IF* of the sample mean and median together with the *IF* of the robust Huber estimator ($c = 1.345$) that will be presented in Section 2.3.1. While the *IF* of the sample mean is unbounded and the *IF* of the median has only two values, the *IF* of the Huber estimator can be seen as a compromise. It remains identical to the sample mean *IF* in the middle and then is truncated symmetrically beyond a certain threshold. As explained later, the form of the *IF* has implications on the efficiency of the corresponding estimator: the Huber estimator is more efficient than the (sample) median.

In general, if $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}$, then the *IF* measures the asymptotic bias of $\hat{\boldsymbol{\theta}}$ due to infinitesimal model deviations of the type (1.1). The choice for G as the gross error generating process is not restrictive. Indeed, Hampel *et al.* (1986, p. 175), showed that for small ε (and general F_ε),

$$\sup_G \|\hat{\boldsymbol{\theta}}(F_\varepsilon) - \hat{\boldsymbol{\theta}}(F_{\boldsymbol{\theta}})\| = \operatorname{bias}(\hat{\boldsymbol{\theta}}, F_{\boldsymbol{\theta}}, \varepsilon) \simeq \varepsilon \sup_z \|IF(z; \hat{\boldsymbol{\theta}}, F_{\boldsymbol{\theta}})\|, \tag{2.2}$$

where $\|\cdot\|$ denotes the Euclidean norm. This means that an estimator $\hat{\boldsymbol{\theta}}$ with a bounded *IF* has a bounded asymptotic bias for any type of contaminating distribution G. This property explains why the *IF* has been a central tool in the development of robustness theory in the infinitesimal sense.

The *IF* is not only a useful tool for assessing the robustness properties of an estimator $\hat{\boldsymbol{\theta}}$, but also for deriving its asymptotic variance. Indeed, Hampel *et al.* (1986, p. 85) showed that

$$V(\hat{\boldsymbol{\theta}}, F_{\boldsymbol{\theta}}) = \int IF(x; \hat{\boldsymbol{\theta}}, F_{\boldsymbol{\theta}}) IF^T(x; \hat{\boldsymbol{\theta}}, F_{\boldsymbol{\theta}})\, dF_{\boldsymbol{\theta}}(x), \tag{2.3}$$

[2]Note that one traditionally uses the notation in z for the arguments of the *IF* while x is used for the sensitivity curve. In what follows, we use z, x or y depending on the context. These notational choices do not change the interpretation of the *IF* which remains the infinitesimal variation of the estimator due to an infinitesimal amount of contamination at any point denoted by z, x or y or, in the multivariate setting, by $\boldsymbol{z}$, $\boldsymbol{x}$ or $\boldsymbol{y}$.

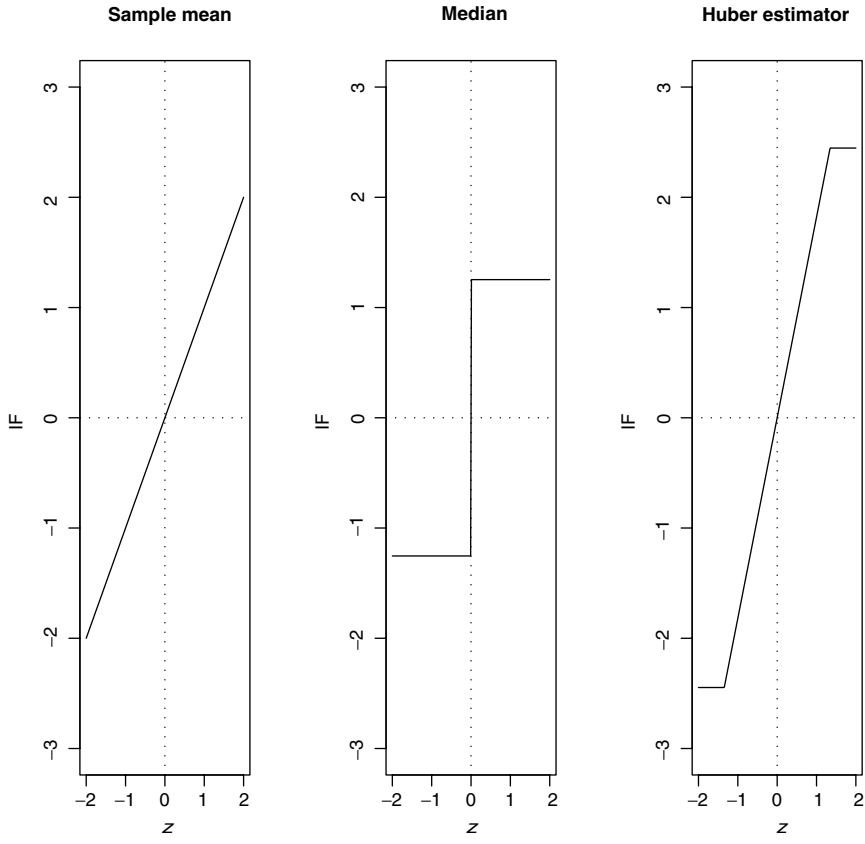

Figure 2.1 The *IF* of the sample mean, median and Huber estimators for the normal population mean with $\mu = 0$ (and $\sigma^2 = 1$).

which can be estimated empirically with

$$\widehat{V}(\hat{\boldsymbol{\theta}}, F_{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n} IF(x_i; \hat{\boldsymbol{\theta}}, F_{\boldsymbol{\theta}}) IF^T(x_i; \hat{\boldsymbol{\theta}}, F_{\boldsymbol{\theta}})$$

with $\boldsymbol{\theta}$ replaced by a consistent estimator[3] (see also Efron, 1982).

A by-product derived from the *IF* is the gross error sensitivity (GES) given by

$$\text{GES}(\hat{\boldsymbol{\theta}}, F_{\boldsymbol{\theta}}) = \sup_z \|IF(z; \hat{\boldsymbol{\theta}}, F_{\boldsymbol{\theta}})\| \tag{2.4}$$

which, according to (2.2), is proportional to the maximal bias on the estimator $\hat{\boldsymbol{\theta}}$ due to a model misspecification of the F_ε-neighborhood type (for small ε). The *IF* (and

[3]Note that to get var($\hat{\boldsymbol{\theta}}$), one has to divide $V(\hat{\boldsymbol{\theta}}, F_{\boldsymbol{\theta}})$ by n.

indeed the GES) can be used not only to assess the robustness properties of a given estimator, but also to build robust estimators by choosing them in general classes among those with bounded *IF* (see Section 2.3).

Finally, other robustness measures derived from the *IF* can be built, such as the local-shift sensitivity which measures the influence of numerous but small deviations such as rounding effects; see Hampel *et al.* (1986, pp. 87–88).

2.2.2 The Breakdown Point

The breakdown point ε^* measures the robustness properties of an estimator in the global sense. It is defined as the maximal amount of model misspecification an estimator can withstand before its bias becomes too large (infinite), i.e. it breaks down. Again we have an empirical and theoretical version. For example, only one observation chosen arbitrarily can carry the sample mean above any given value, therefore its empirical breakdown point is zero. On the other hand, the (sample) median is highly resistant as we need to substitute $[n/2 + 1]/n$ of the observations to make it break down. Its breakdown point therefore tends to 50% for large enough samples. For a more formal definition, we can relate the amount of model misspecification to the quantity ε in the F_ε-neighborhood (1.1). There exists different formal definitions of ε^*, the first proposals being those of Hodges (1967) and Hampel (1968). One can define the breakdown point as

$$\varepsilon^* := \varepsilon^*(\hat{\boldsymbol{\theta}}, F_{\boldsymbol{\theta}}) = \inf\{\varepsilon \mid \text{bias}(\hat{\boldsymbol{\theta}}, F_{\boldsymbol{\theta}}, \varepsilon) = \infty\}. \tag{2.5}$$

A consequence is that if the GES of $\hat{\boldsymbol{\theta}}$ is infinite, then its ε^* is nil. Thus, the theoretical breakdown point of the sample mean is zero, which further illustrates its lack of robustness. In contrast, the median is globally robust as $\varepsilon^* = \frac{1}{2}$. In most cases, the theoretical breakdown point is also equal to the limit of its empirical version.

In practice, the breakdown point is very rarely used as such to assess the robustness properties of a statistic as it corresponds to the worst-case scenario. Usually, the *IF* and consequently the GES are first measured and, if they are bounded, ε^* is computed in a second step (as a complementary measure). This is important, because a robust estimator with bounded *IF* can become useless in practice if its breakdown point is too small. In Section 2.3 we present general classes of estimators with high breakdown points, also called globally robust estimators.

2.2.3 Geometrical Interpretation

The *IF*, the GES and the breakdown point ε^* are robustness measures that are linked together. In Figure 2.2 we illustrate these three robustness measures on a (theoretical) plot of the maximal bias of an estimator $\hat{\boldsymbol{\theta}}$ as a function of the amount of model deviation ε. While the GES, through the *IF*, measures a first-order approximation of the maximal bias (see (2.2) and (2.4)), the breakdown point measures the maximal amount of model deviation the estimator can withstand before its maximal bias becomes too large.

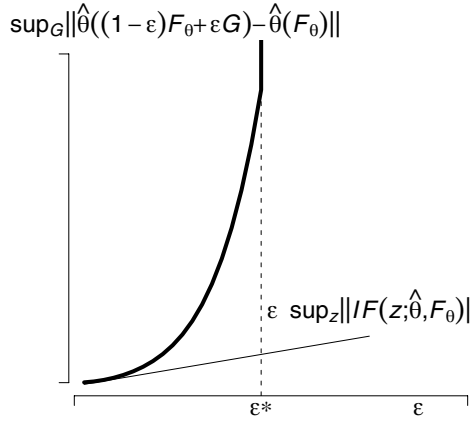

Figure 2.2 The relationship between the *IF*, GES, breakdown point ε^* and maximal bias of an estimator under model misspecification.

2.2.4 The Rejection Point

Another useful robustness measure that is useful for comparing different robust estimators is the rejection point. It is used in multivariate settings and, roughly speaking, it is the distance ρ^* from the center of the data such that points lying outside this distance have no influence on the asymptotic bias. More formally (see Hampel *et al.*, 1986, p. 88), assuming that $F_{\boldsymbol{\theta}}$ is symmetric (and centered at, say, $\boldsymbol{m}$), the rejection point is defined as

$$\rho^* := \rho^*(\hat{\boldsymbol{\theta}}, F_{\boldsymbol{\theta}}) = \inf\{r > 0 \mid IF(\boldsymbol{z}; \hat{\boldsymbol{\theta}}, F_{\boldsymbol{\theta}}) = 0 \text{ for some } \boldsymbol{z}, \text{ when } \delta(\boldsymbol{z}, \boldsymbol{m}) > r\}, \tag{2.6}$$

where δ is a suitable distance measure (e.g. the Euclidean norm) and $\boldsymbol{z}$ is a point in a multidimensional space. It is desirable that a robust estimator has a finite rejection point, meaning that points too far away from the center of the data receive a weight of zero. However, the size of ρ^* is somewhat arbitrary, unless it can be relative to the model $F_{\boldsymbol{\theta}}$ at which it is computed. Rocke (1996) proposes to relate ρ^* to the probability that a point lying 'outside' ρ^* has been generated by $F_{\boldsymbol{\theta}}$. If this probability is too small, then only very improbable points under the model have no influence on the estimator. It is therefore important to also control ρ^* such that points with a probability of, say, α^* of being 'outside' ρ^* have no influence on $\hat{\boldsymbol{\theta}}$. Rocke (1996) defines α^* as the asymptotic rejection probability.

2.3 General Approaches for Robust Estimation

Estimation is an important aspect of statistical inference. Given a sample of observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ from, presumably, a parametric model $F_{\boldsymbol{\theta}}$ with corresponding

density $f(\boldsymbol{x}_i; \boldsymbol{\theta})$, one is typically interested in finding an estimator for the q-dimensional vector of parameters $\boldsymbol{\theta}$, say $\hat{\boldsymbol{\theta}}$. The choice for $\hat{\boldsymbol{\theta}}$ is usually quite large, and should therefore be based on the properties that $\hat{\boldsymbol{\theta}}$ possesses. A classical property is consistency relative to the postulated model $F_{\boldsymbol{\theta}}$. Consistency generally means that $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}$ in some sense when n goes to infinity. One way to extend this concept to estimators seen as functionals is the strong Fisher consistency, defined naturally as

$$\hat{\boldsymbol{\theta}}(F_{\boldsymbol{\theta}}) = \boldsymbol{\theta}. \tag{2.7}$$

In other words, a Fisher consistent estimator computed at the model produces the value of the model parameters.

Another classical property is efficiency, i.e. the minimal variance or minimal mean squared error (MSE) when the estimator is not consistent. Moreover, in the robustness paradigm, we are also interested in estimators that are robust either in the infinitesimal sense (bounded *IF*) or in the global sense (high breakdown point). In Section 2.3.1 we introduce a very general class of estimators in which estimators that fulfill these properties (consistency, reasonable efficiency and robustness) can be found.

At the model $F_{\boldsymbol{\theta}}$, i.e. when one assumes that the data have been generated exactly from $F_{\boldsymbol{\theta}}$, a consistent and efficient estimator is the MLE or $\hat{\boldsymbol{\theta}}_{[MLE]}$, obtained by maximizing the log-likelihood of the data. Mathematically, $\hat{\boldsymbol{\theta}}_{[MLE]}$ is then the solution of

$$\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log f(\boldsymbol{x}_i; \boldsymbol{\theta}) \tag{2.8}$$

or, alternatively, the solution for $\boldsymbol{\theta}$ of the first-order equation

$$\sum_{i=1}^{n} s(\boldsymbol{x}_i; \boldsymbol{\theta}) = \mathbf{0}, \tag{2.9}$$

where $s(\boldsymbol{x}; \boldsymbol{\theta}) = (\partial/\partial\boldsymbol{\theta}) \log f(\boldsymbol{x}; \boldsymbol{\theta})$ is the q-dimensional score function. Its *IF* can be easily computed and is given by

$$IF(z; \hat{\boldsymbol{\theta}}_{[MLE]}, F_{\boldsymbol{\theta}}) = \left[\int s(\boldsymbol{x}; \boldsymbol{\theta}) s^T(\boldsymbol{x}; \boldsymbol{\theta}) \, dF_{\boldsymbol{\theta}}(\boldsymbol{x}) \right]^{-1} s(z; \boldsymbol{\theta}), \tag{2.10}$$

which is unbounded if the score function is unbounded. This will have disastrous consequences if $F_{\boldsymbol{\theta}}$ is not the exact model (i.e. the exact data-generating process is somehow different from what was assumed initially).

For example, suppose again that the (univariate) normal model $\mathcal{N}(\mu, \sigma^2)$ is a good working model for the data. The sample mean $\hat{\mu} = (1/n) \sum x_i$ and the sample variance $\hat{\sigma}^2 = (1/n) \sum (x_i - \hat{\mu})^2$ are the MLEs of $\boldsymbol{\theta} = (\mu, \sigma^2)^T$. Indeed, simple

calculations show that the score function for that model is

$$\begin{aligned} s(x;\mu,\sigma^2) &= \frac{\partial}{\partial(\mu,\sigma^2)^T}\log f(x;\mu,\sigma^2) \\ &= \frac{\partial}{\partial(\mu\sigma^2)^T}\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)\right) \\ &= \begin{bmatrix} \frac{1}{\sigma^2}(x-\mu) \\ -\frac{1}{2\sigma^2}+\frac{1}{2\sigma^4}(x-\mu)^2 \end{bmatrix}, \end{aligned}$$

and $\hat{\boldsymbol{\theta}} = (\hat{\mu}, \hat{\sigma}^2)^T$ satisfies (2.9) with the score function given above. As $s(x;\mu,\sigma^2)$ can become arbitrarily large, when x is far away from the mean μ, so is the *IF* of both $\hat{\mu}$ and $\hat{\sigma}^2$ and, as a result, the MLE of both parameters is not robust. This rather formal result will not surprise anyone, as a single observation far away from the bulk of the data (i.e. far away from the mean or center of the distribution) inflates the sample mean as illustrated with the sensitivity curve. The same can also be observed with the sample variance. Here $\hat{\mu}$ and $\hat{\sigma}^2$ are both biased estimators of, respectively, the population mean and variance when the data-generating process is not exactly the normal distribution.

The MLE is not the only estimator that does not have, in general, good local robustness property. Moment-based estimators are not robust either since they rely on sample means, variances, etc. For some (simple) models, *ad-hoc* or simple intuitive estimators can be proposed that are robust in addition to being consistent and reasonably efficient. For example, for the normal model, a α-trimmed mean, i.e. the sample mean of the data from which the α proportion of the smallest and largest observations have been discarded, is a consistent and robust estimator of the population mean μ, in that it has a bounded *IF*. Its breakdown point is equal to α.

2.3.1 The General Class of M-estimators

An estimator is often chosen as a member of a general class of estimators that is optimal in some sense or fulfills a set of good properties. We present below the class of M-estimators, give their *IF* so that it is then easier to choose a robust estimator within this class and consider a subclass of so-called weighted MLEs with different forms of weights. We also investigate their properties of consistency, efficiency and asymptotic normality. We do not present all of the robust M-estimators available in the literature, but concentrate on the most well-known estimators that will also be used throughout the book. Assume we have n (multivariate) i.i.d. observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ from a common model $F_{\boldsymbol{\theta}}$. In his pioneering work, Huber (1964, 1967) proposed a class of M-estimators that naturally generalize the MLE. An M-estimator of $\boldsymbol{\theta}$ is given by the solution $\hat{\boldsymbol{\theta}}_{[M]}$ of the minimization problem

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \rho(\boldsymbol{x}_i;\boldsymbol{\theta}), \tag{2.11}$$

or, alternatively, by the solution for $\boldsymbol{\theta}$ of

$$\sum_{i=1}^{n} \Psi(\boldsymbol{x}_i; \boldsymbol{\theta}) = \mathbf{0}, \tag{2.12}$$

for suitable ρ and Ψ functions where $\Psi(\boldsymbol{x}; \boldsymbol{\theta}) = \partial\rho(\boldsymbol{x}; \boldsymbol{\theta})/\partial\boldsymbol{\theta}$. Setting $\rho = -\log f$ in (2.11) or $\Psi = s$ in (2.12) gives back (2.8) and (2.9), respectively. Hence, the MLE is just a particular case in the class of M-estimators when ρ is the negative log-density and Ψ the score function. In general, $\Psi(\boldsymbol{x}; \boldsymbol{\theta})$ needs not be the derivative of some ρ-function with respect to the parameter of interest, therefore (2.12) is more general and is often referred as the proper definition of an M-estimator.[4]

The *IF* of an M-estimator is given by

$$IF(\boldsymbol{z}; \hat{\boldsymbol{\theta}}_{[M]}, F_{\boldsymbol{\theta}}) = M^{-1}(\Psi, F_{\boldsymbol{\theta}})\Psi(\boldsymbol{z}; \boldsymbol{\theta}) \tag{2.13}$$

with

$$M(\Psi, F_{\boldsymbol{\theta}}) = -\int \frac{\partial}{\partial\boldsymbol{\theta}}\Psi(\boldsymbol{x}; \boldsymbol{\theta})\, dF_{\boldsymbol{\theta}}(\boldsymbol{x}), \tag{2.14}$$

see e.g. Huber (1981). Formula (2.13) shows that, in a similar fashion to the MLE, the *IF* of an M-estimator is proportional to its defining Ψ-function. This is a powerful result since it suffices to choose a bounded Ψ-function to obtain a robust M-estimator. On the other hand, if the model score function $s(\boldsymbol{x}; \boldsymbol{\theta})$ or any $\Psi(\boldsymbol{x}; \boldsymbol{\theta})$ function turns out to be unbounded in its argument $\boldsymbol{x}$, then the corresponding estimator is not robust for the parameter of interest, as illustrated with the MLE of the normal model in (2.10).

Relatively simple M-estimators include the so-called weighted MLE (WMLE)[5] defined as the solution for $\boldsymbol{\theta}$ of

$$\sum_{i=1}^{n} w(\boldsymbol{x}_i; \boldsymbol{\theta})s(\boldsymbol{x}_i; \boldsymbol{\theta}) - a(\boldsymbol{\theta}) = \mathbf{0}, \tag{2.15}$$

where $a(\boldsymbol{\theta})$ is a consistency correction factor (see Section 2.3.2). The MLE corresponds to the choice $w(\boldsymbol{x}_i; \boldsymbol{\theta}) = 1$ for all i and, consequently, $a(\boldsymbol{\theta}) = \mathbf{0}$. To construct a robust WMLE, one simply chooses weights that make $w(\boldsymbol{x}; \boldsymbol{\theta})s(\boldsymbol{x}; \boldsymbol{\theta})$ bounded. The weights can depend on the observations only, on a quantity that depends itself on the observations and the parameters or, more generally, directly on the score function. As an illustration, consider again the univariate normal model (with unit scale). The score function for μ is $s(x; \mu) = x - \mu$; obviously the quantity to bound is the score function itself. In the linear regression model $y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \varepsilon_i$ (with again $\text{var}(\varepsilon_i) = 1$ for simplicity), the score function has a similar expression $s(y, \boldsymbol{x}; \boldsymbol{\beta}) = r \cdot \boldsymbol{x}$ but is

[4] It should also be stressed that an M-estimator can also be defined for weighted data such as data produced by some type of stratified sampling. Denoting the weights due to the sampling scheme by $\omega(\boldsymbol{x}_i), i, \ldots, n$, (2.12) becomes $\sum_{i=1}^{n} \omega(\boldsymbol{x}_i)\Psi(\boldsymbol{x}_i; \boldsymbol{\theta}) = \mathbf{0}$. The subsequent definitions and results are presented assuming $\omega(\boldsymbol{x}_i) = 1$, for all i, but can be easily extended to the more general case by simply adding $\omega(\boldsymbol{x}_i)$ after the summation symbol.

[5] Robust WMLEs were first formalized by Field and Smith (1994).

proportional to $r = y - \boldsymbol{x}^T\boldsymbol{\beta}$, the residual, and $\boldsymbol{x}$, the covariate; see also Chapter 3 for details. In this case, there are two quantities to bound, the residual r and covariate value $\boldsymbol{x}$. The situation described here in a simple model is actually representative of many regression models where some univariate residual appears in the score function. In that univariate case, a popular choice are Huber's weights

$$w_{[Hub]}(r; \boldsymbol{\beta}, c) = \psi_{[Hub]}(r; \boldsymbol{\beta}, c)/r = \min\{1; c/|r|\}, \tag{2.16}$$

i.e. the weight is equal to one for all (small) values of r satisfying $|r| < c$ and $c/|r|$ otherwise. The Huber ψ-function (Huber, 1964) is simply

$$\psi_{[Hub]}(r; \boldsymbol{\beta}, c) = \frac{\partial}{\partial r}\rho_{[Hub]}(r; \boldsymbol{\beta}, c) = \min\{c, \max\{r, -c\}\}$$

with the corresponding ρ-function

$$\rho_{[Hub]}(r; \boldsymbol{\beta}, c) = \begin{cases} \frac{1}{2}r^2 & \text{if } |r| \le c, \\ c|r| - \frac{1}{2}c^2 & \text{if } |r| > c. \end{cases} \tag{2.17}$$

Note that in a regression model with known scale, the Huber estimator is an M-estimator associated with $\Psi_{[Hub]}(r, \boldsymbol{x}; \boldsymbol{\beta}, c) = \psi_{[Hub]}(r, \boldsymbol{x}; \boldsymbol{\beta}, c)\boldsymbol{x}$, a bounded function of r (or the response y). The ρ- and ψ-functions of the MLE and Huber proposal are depicted in Figure 2.3, left and middle panels.

As explained in the subsequent chapters, Huber's weights can be used with many different models where the corresponding M-estimator is defined through

$$\Psi_{[Hub]}(r, \boldsymbol{x}; \boldsymbol{\theta}, c) = w_{[Hub]}(r; \boldsymbol{\theta}, c) \cdot r\frac{\partial r}{\partial \boldsymbol{\theta}} + a(\boldsymbol{\theta}),$$

where $a(\boldsymbol{\theta})$ is a consistency correction factor (see Section 2.3.2). As long as the argument r of the weight function, whether it be a residual, a score or some other quantity, has a reasonable value the weights are equal to one and no downweighting is performed. The observation is downweighted only if the argument exceeds some threshold value c. The latter is chosen on the basis of robustness arguments (the lower c is, the lower the weights, the more robust the estimator) and efficiency arguments (the lower c is, the more observations are downweighted, the less efficient the estimator); see also Section 2.3.2.

More generally, in the multivariate case, the weights of the WMLE in (2.15) can be chosen as (see e.g. Hampel *et al.*, 1986, p. 239)

$$w_{[Hub]}(s(\boldsymbol{x}; \boldsymbol{\theta}); c) = \min\{1; c/\|s(\boldsymbol{x}; \boldsymbol{\theta})\|\} \tag{2.18}$$

and the corresponding WMLE with Huber's weights is defined through

$$\Psi_{[Hub]}(\boldsymbol{x}; \boldsymbol{\theta}, c) = w_{[Hub]}(s(\boldsymbol{x}; \boldsymbol{\theta}); c)s(\boldsymbol{x}; \boldsymbol{\theta}) + a(\boldsymbol{\theta}). \tag{2.19}$$

It is theoretically possible to define an optimal M-estimator in the sense that it has maximal efficiency among all M-estimators with bounded *IF* measured in an

appropriate metric; see Hampel *et al.* (1986). This estimator is called the optimal B-robust estimator (OBRE) but is often difficult to compute. Its Ψ-function has a similar form to the WMLE with Huber's weights (2.19), except that the Ψ-function in (2.19) is multiplied by a matrix $A(\boldsymbol{\theta})$ carefully chosen on efficiency grounds (for more details, see Hampel *et al.* (1986, p. 240)).

Although the previous weighting schemes may appear the most natural, they have been criticized because the resulting M-estimators may possess too small a breakdown point in relatively high dimensions. Such estimators have actually been shown to have a breakdown of at best $1/\dim(\boldsymbol{\theta})$ (see Hampel *et al.*, 1986; Maronna, 1976). Fortunately, the breakdown point can be improved by considering the 'so-called' redescending Ψ-functions, i.e. functions that become nil for large values of their argument or, in other terms, their rejection point (2.6) is finite. This is generally equivalent to saying that the corresponding ρ-function is constant for large enough values. A good example of such redescending score function is the popular bisquare or biweight proposed by Beaton and Tukey (1974) associated with

$$\rho_{[bi]}(r;\boldsymbol{\theta},c)=\begin{cases}3\left(\dfrac{r}{c}\right)^2-3\left(\dfrac{r}{c}\right)^4+\left(\dfrac{r}{c}\right)^6 & \text{if } |r|\le c,\\ 1 & \text{if } |r|>c.\end{cases} \tag{2.20}$$

The function $\rho_{[bi]}$ is indeed bounded for values of $r:=r(\boldsymbol{\theta})$ larger than the tuning constant c, and this feature helps in constructing a high breakdown point estimator. Tukey's biweight $\rho_{[bi]}$ was first proposed in the context of the normal model and linear regression and, hence, the classical (non-robust) counterpart of (2.20) is $\rho(r;\boldsymbol{\theta})=r^2/2$ (i.e. the squared residuals). To define the weights, one uses

$$\psi_{[bi]}(r;\boldsymbol{\theta},c)=\frac{c^2}{6}\frac{\partial}{\partial r}\rho_{[bi]}(r;\boldsymbol{\theta},c),$$

i.e.

$$\psi_{[bi]}(r;\boldsymbol{\theta},c)=\begin{cases}\left(\left(\dfrac{r}{c}\right)^2-1\right)^2 r & \text{if } |r|\le c,\\ 0 & \text{if } |r|>c,\end{cases} \tag{2.21}$$

$$=w_{[bi]}(r;\boldsymbol{\theta},c)r, \tag{2.22}$$

so that

$$w_{[bi]}(r;\boldsymbol{\theta},c)=\begin{cases}\left(\left(\dfrac{r}{c}\right)^2-1\right)^2 & \text{if } |r|\le c,\\ 0 & \text{if } |r|>c.\end{cases} \tag{2.23}$$

The corresponding M-estimator is therefore defined through

$$\Psi_{[bi]}(r,\boldsymbol{x};\boldsymbol{\theta},c)=w_{[bi]}(r;\boldsymbol{\theta},c)\cdot r\frac{\partial r}{\partial\boldsymbol{\theta}}+a(\boldsymbol{\theta}), \tag{2.24}$$

where $a(\boldsymbol{\theta})$ is again a consistency correction factor[6] and $\boldsymbol{x}$ is a (possible) set of covariates that enter in the definition of the residuals r. When r tends to the tuning constant c, the weight $w_{[bi]}(r; c)$ tends to zero and so does $\Psi_{[bi]}$. Formula (2.23) applies to unidimensional arguments r, but this is not too restrictive, as most common uses of such objective functions have been in regression models or mixed models where the problem can be made unidimensional by exploiting the dependence of the score on the residual or Mahalanobis distance; see Section 2.3.3 for details. Instead of the biweight function, one can also choose other functions. For example, Hampel proposes a 'three-part' redescending function and Andrews proposes a sine function (see Andrews *et al.*, 1972), most of these being effectively used for the linear regression model. Figure 2.3 displays the ρ- and ψ-functions for the MLE, Huber and biweight estimators. For the MLE, the ψ-function (score) and corresponding $\rho = -\log(f)$ are unbounded. In contrast, the Huber estimator has a score function bounded by c, but it remains constant beyond that threshold. Its ρ-function is quadratic in the middle and linear in the tails. Finally, the biweight ψ-function score is redescending with the corresponding ρ-function being symmetric and constant for $|r| > c$. As a result, the MLE is not robust, the Huber estimator is robust but does not have a large breakdown point in high dimensions q, while the biweight estimator is a high breakdown point estimator.

2.3.2 Properties of *M*-estimators

For an M-estimator to be consistent, the Ψ-function must satisfy some regularity conditions given in Huber (1967). Fisher consistency for an M-estimator at the model F_θ is ensured if Ψ satisfies[7]

$$\int \Psi(\boldsymbol{x}; \boldsymbol{\theta})\, dF_{\boldsymbol{\theta}}(\boldsymbol{x}) = \mathbf{0}. \tag{2.25}$$

In particular, the WMLE given in (2.15) is consistent if $a(\boldsymbol{\theta})$ is equal to $n \int w(\boldsymbol{x}; \boldsymbol{\theta})\, s(\boldsymbol{x}; \boldsymbol{\theta})\, dF_{\boldsymbol{\theta}}(\boldsymbol{x})$. The (consistent) WMLE is then the solution for $\boldsymbol{\theta}$ of

$$\sum_{i=1}^{n} w(\boldsymbol{x}_i; \boldsymbol{\theta}) s(\boldsymbol{x}_i; \boldsymbol{\theta}) - n \int w(\boldsymbol{x}; \boldsymbol{\theta}) s(\boldsymbol{x}; \boldsymbol{\theta})\, dF_{\boldsymbol{\theta}}(\boldsymbol{x}) = \mathbf{0}. \tag{2.26}$$

Fortunately, in symmetric models such as the normal model or the regression model, we have $a(\boldsymbol{\theta}) = 0$, which makes the WMLE rather easy to compute, using for example an iteratively reweighted least squares (IRWLS) algorithm; see Section 3.2.4 for an example. In other cases, the computation of a consistent WMLE through (2.26) is harder as the integral term has to be evaluated. Moustaki and Victoria-Feser (2006) (see also Victoria-Feser, 2007) propose the use of the method of indirect inference (see Gallant and Tauchen, 1996; Genton and Ronchetti, 2003;

[6] With the models in this book for which we use $\Psi_{[bi]}$, we have that $a(\boldsymbol{\theta}) = 0$.

[7] The rationale for this is simply that, by dividing (2.12) by n we see that the functional defining the M-estimator is $\hat{\boldsymbol{\theta}}_{[M]}(F)$ satisfying $\int \Psi(\boldsymbol{x}_i; \hat{\boldsymbol{\theta}}_{[M]}(F))\, dF(\boldsymbol{x}) = 0$. The general Fisher condition (2.7) is then equivalent to (2.25).

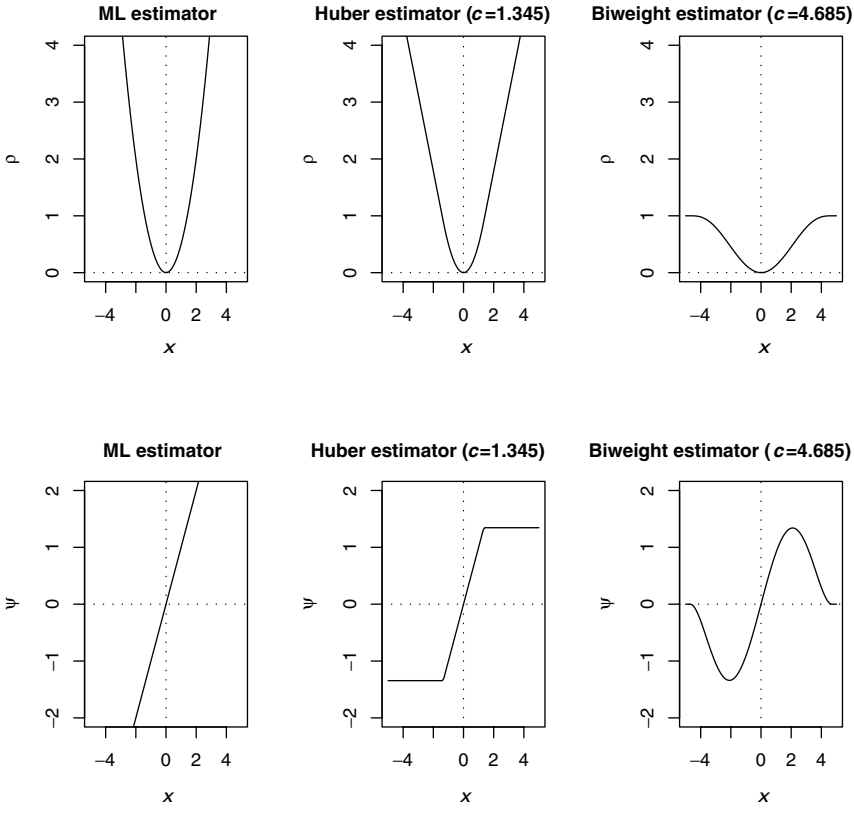

Figure 2.3 ρ-functions (top panels) and ψ-functions (bottom panels) for the MLE, Huber and biweight estimators.

Gouriéroux *et al.*, 1993), to remove the bias of a WMLE in (2.15) in complex models. The basic idea behind the method is as follows. An inconsistent estimator $\hat{\boldsymbol{\theta}}^0$ is first computed from the data using the WMLE without the bias correction $nE[w(\boldsymbol{x};\boldsymbol{\theta})s(\boldsymbol{x};\boldsymbol{\theta})]$. Then the bias is corrected via simulation even though its exact form may not be known;[8] for a short discussion on Fisher consistency correction, see also Section 5.3.2.

If consistency is certainly a desirable property, a reasonable level of efficiency is also required. As the MLE is the most efficient estimator at the model among all (asymptotically) consistent estimators, some loss of efficiency is expected when using a robust alternative. This is the price to pay for robustness if the model that generated the data is indeed $F_{\boldsymbol{\theta}}$. We certainly want to keep that loss small in general.

[8]More specifically, at the model $F_{\boldsymbol{\theta}}$, asymptotically we have that $\int w(\boldsymbol{x};\hat{\boldsymbol{\theta}}^0(F_{\boldsymbol{\theta}}))$ $s(\boldsymbol{x};\hat{\boldsymbol{\theta}}^0(F_{\boldsymbol{\theta}}))\,dF_{\boldsymbol{\theta}}(\boldsymbol{x}) = \mathbf{0}$; then, using simulated data, the value of the estimator is found solving the previous equation in $\boldsymbol{\theta}$ and taking for $\hat{\boldsymbol{\theta}}^0(F_{\boldsymbol{\theta}})$ its sample value $\hat{\boldsymbol{\theta}}^0$.

The concept of efficiency is closely related to the (asymptotic) variance of an M-estimator that can be obtained simply by combining (2.3) and (2.13), yielding

$$V(\hat{\boldsymbol{\theta}}_{[M]}, F_{\boldsymbol{\theta}}) = M^{-1}(\Psi, F_{\boldsymbol{\theta}})Q(\Psi, F_{\boldsymbol{\theta}})M^{-T}(\Psi, F_{\boldsymbol{\theta}}) \tag{2.27}$$

with

$$Q(\Psi, F_{\boldsymbol{\theta}}) = \int \Psi(\boldsymbol{x}; \boldsymbol{\theta})\Psi^T(\boldsymbol{x}; \boldsymbol{\theta})\, dF_{\boldsymbol{\theta}}(\boldsymbol{x}), \tag{2.28}$$

and $M(\Psi, F_{\boldsymbol{\theta}})$ given in (2.14).[9] Moreover, when $\hat{\boldsymbol{\theta}}_{[M]}$ is Fisher consistent, i.e. its Ψ-function satisfies (2.25), we have the following simpler expression for M

$$M(\Psi, F_{\boldsymbol{\theta}}) = \int \Psi(\boldsymbol{x}; \boldsymbol{\theta})s^T(\boldsymbol{x}; \boldsymbol{\theta})\, dF_{\boldsymbol{\theta}}(\boldsymbol{x}). \tag{2.29}$$

Note that, in the special case of the MLE, i.e. when $\Psi = s$, (2.27) reduces to the inverse of the Fisher information matrix, i.e.

$$V(\hat{\boldsymbol{\theta}}_{[MLE]}, F_{\boldsymbol{\theta}}) = \left[\int s(\boldsymbol{x}; \boldsymbol{\theta})s^T(\boldsymbol{x}; \boldsymbol{\theta})\, dF_{\boldsymbol{\theta}}(\boldsymbol{x})\right]^{-1} = I(\boldsymbol{\theta})^{-1}, \tag{2.30}$$

which is the minimal variance an unbiased estimator can achieve at the model $F_{\boldsymbol{\theta}}$. As the robust M-estimator has an asymptotic variance that is always greater than the asymptotic variance of the MLE at the core model $F_{\boldsymbol{\theta}}$, one way to quantify the loss in efficiency (and possibly fine tune the robust estimator) is to use the ratio of the traces (denoted by tr) of the asymptotic variances

$$\frac{\text{tr}([\int s(\boldsymbol{x}; \boldsymbol{\theta})s^T(\boldsymbol{x}; \boldsymbol{\theta})\, dF_{\boldsymbol{\theta}}(\boldsymbol{x})]^{-1})}{\text{tr}(M^{-1}(\Psi, F_{\boldsymbol{\theta}})Q(\Psi, F_{\boldsymbol{\theta}})M^{-T}(\Psi, F_{\boldsymbol{\theta}}))}, \tag{2.31}$$

which can be interpreted as the asymptotic MSE (see Hampel *et al.*, 1986, p. 242). Note that (2.31) is not necessarily constant and may depend on the parameter $\boldsymbol{\theta}$ (or $\hat{\boldsymbol{\theta}}$); in that case, the same value should be used both in the numerator and the denominator. The efficiency of an M-estimator is related to the form of the score function Ψ, which in turn depends on tuning constants that regulate the robustness properties of the estimator. This is the case for example for the WMLE (2.15) with Huber-type weights (2.18) or with the biweight M-estimator (2.22) through the lone tuning constant c. A strategy consists of choosing the tuning constant(s) that make(s) the efficiency (2.31) reach a given level, typically 90% or 95%. In Figure 2.3, the ρ- and ψ-functions of the robust estimators were tuned to achieve 95% efficiency at the normal model. Finally, under some (mild) regularity conditions on ψ, we have that $\sqrt{n}(\hat{\boldsymbol{\theta}}_{[M]} - \boldsymbol{\theta})$ is asymptotically normal with zero mean and variance equal to (2.27). The regularity conditions can be found in, for example, Huber (1981) or Welsh (1996).

[9] Note that to obtain $\text{var}(\hat{\boldsymbol{\theta}}_{[M]})$, one has to divide $V(\hat{\boldsymbol{\theta}}_{[M]}, F_{\boldsymbol{\theta}})$ by n.

2.3.3 The Class of *S*-estimators

S-estimators were first proposed by Rousseeuw and Yohai (1984) in the context of regression models. They are useful in problems where the estimation of some scale parameter is an issue and constitute an alternative to *M*-estimators. One of their key features is a high breakdown point (as defined in Section 2.2.2). In regression models, for example, *S*-estimators generalize the LS estimator which is based on the scale of the residuals (i.e. the mean of the squared residuals). In a multivariate normal context, *S*-estimators are used for the (robust) estimation of covariance matrices $\boldsymbol{\Sigma}$ and mean vectors $\boldsymbol{\mu}$. Formally, and very generally, an *S*-estimator is defined by the minimization of a dispersion function, depending on the unknown parameters collected in the vector $\boldsymbol{\theta}$, under the constraint that

$$\frac{1}{n}\sum_{i=1}^{n}\rho(d_i;\boldsymbol{\theta}) = E[\rho(d;\boldsymbol{\theta})], \tag{2.32}$$

where the quantities $d_i := d_i(\boldsymbol{\theta})$ also depend on the observations and where the expectation is taken at the distribution of d (for consistency of the *S*-estimator). The function ρ must satisfy some regularity conditions (see e.g. Rousseeuw and Leroy, 1987, p. 136), in particular, ρ must be symmetric and bounded (i.e. constant for large values of its argument). This actually guarantees that the *S*-estimator has a high breakdown point obtained through

$$\varepsilon^* = \frac{E[\rho(d;\boldsymbol{\theta})]}{\max_x \rho(x;\boldsymbol{\theta})} \tag{2.33}$$

(see Rousseeuw and Leroy, 1987; Lopuhaä and Rousseeuw, 1991). A typical example of an appropriate ρ-function is Tukey's biweight. Tuning constants help in calibrating the ρ-function in (2.33) to achieve a pre-specified value of ε^*, for example, 25% or even 50%. In the regression setting of Chapter 3, *S*-estimators will also be used. In that case, the d_i are the standardized residuals and the dispersion function is the residual scale.

S-estimators are especially useful in the multivariate normal model, i.e. $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma}$ a symmetric and positive-definite matrix. Estimates of $\boldsymbol{\Sigma}$ are used in many analyses such as principal component analysis, factor analysis, structural equation models or simply as a preliminary data analysis. As will be shown in Chapter 4, even mixed linear models can be formalized as a multivariate normal model. In all of these settings, an important aspect is the (robust) estimation of the (population) covariance matrix with the multivariate normal model as the postulated model. *S*-estimators are very often used for their robustness properties (high ε^*) and their good efficiency relative to other robust estimators (see below). The dispersion function that is chosen for the multivariate normal model is the determinant of the covariance matrix $\det(\boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|$ and the argument d_i of the ρ-function are Mahalanobis distances

$$d_i = \sqrt{(\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})}. \tag{2.34}$$

The Mahalanobis distance is a natural measure of 'outlyingness' of an observation. Recall that points (observations) at equal Mahalanobis distances from the center of the data ($\boldsymbol{\mu}$) form an ellipse of equal density (see e.g. Figure 1.3). Consequently, and provided that the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in (2.34) are estimated robustly, Mahalanobis distances can be used to detect 'outlying' observations in that the latter correspond to large Mahalanobis distances. More precisely, since under the multivariate normal model, we have that $d_i^2 \sim \chi_q^2$ (the chi-square distribution with q degrees of freedom), then one could consider that an observation is also 'outlying' if it exceeds a cut-off value of say $(\chi_q^2)^{-1}(0.975)$.

In the multivariate normal setting, S-estimators are a generalization of the MLE. Indeed, for the latter, the log-likelihood function is

$$\log L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{n}{2} q \log(2\pi) - \frac{n}{2} \log|\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{i=1}^{n} d_i^2.$$

Hence, the MLE minimizes the quantity $\log|\boldsymbol{\Sigma}| + (1/n) \sum_{i=1}^{n} \rho(d_i)$, with $\rho(d) = d^2$. An S-estimator for the multivariate normal model minimizes $|\boldsymbol{\Sigma}|$ (or, equivalently, $\log|\boldsymbol{\Sigma}|$) for a fixed value of $(1/n) \sum_{i=1}^{n} \rho(d_i)$ and a bounded ρ-function.

The S-estimator, say $(\hat{\boldsymbol{\mu}}_{[S]}, \widehat{\boldsymbol{\Sigma}}_{[S]})$, has several important properties. It possesses a high breakdown point and the property of affine equivariance, i.e. a linear transformation $\boldsymbol{A}\boldsymbol{x}_i + \boldsymbol{a}$ of the data yields the estimator $(\boldsymbol{A}\hat{\boldsymbol{\mu}}_{[S]} + \boldsymbol{a}, \boldsymbol{A}\widehat{\boldsymbol{\Sigma}}_{[S]}\boldsymbol{A}^T)$. Conditions for its existence, consistency, asymptotic normality and high breakdown point have been investigated by Davies (1987).

Most known S-estimators can actually be written as M-estimators (2.12) (see Rousseeuw and Yohai, 1984; Lopuhaä, 1989; Rocke, 1996; Copt and Victoria-Feser, 2006). They therefore have a similar asymptotic distribution. In particular, S-estimators for multivariate location and covariance matrix are the solution for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ of

$$\sum_{i=1}^{n} \Psi_S^{\boldsymbol{\mu}}(\boldsymbol{x}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i=1}^{n} w(d_i)(\boldsymbol{x}_i - \boldsymbol{\mu}) = \boldsymbol{0}, \tag{2.35}$$

$$\sum_{i=1}^{n} \Psi_S^{\boldsymbol{\Sigma}}(\boldsymbol{x}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{q \sum_{i=1}^{n} w(d_i)(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T}{\sum_{i=1}^{n} w(d_i) d_i^2} - \boldsymbol{\Sigma} = \boldsymbol{0}, \tag{2.36}$$

with $w(d; \boldsymbol{\theta}) = \rho'(d; \boldsymbol{\theta})/d$ and $\rho'(d; \boldsymbol{\theta}) = \psi(d; \boldsymbol{\theta}) = \partial\rho(d; \boldsymbol{\theta})/\partial d$. The Ψ-functions $\Psi_S^{\boldsymbol{\mu}}$ and $\Psi_S^{\boldsymbol{\Sigma}}$ are redescending because they depend on weights w based on the first derivative of the ρ-function which is, by definition, bounded. S-estimators therefore form a subclass of M-estimators built on a redescending ψ-function and, hence, have a positive breakdown point that can be set to a chosen value.

Again, a popular choice for the ρ-function is Tukey's biweight function (2.20) with argument $r = d(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = d$, the Mahalanobis distance (2.34). Rocke (1996) proposes to extend Tukey's biweight function to a translated biweight function to control for the asymptotic rejection probability α^*.

A less desirable feature of S-estimators inherent to redescending ψ-functions is that the system (2.35) and (2.36) admits more than one solution (corresponding

to different local minima). To overcome this difficulty a good initial estimator is first computed and then used as starting point in the computation of the S-estimator through (2.35) and (2.36); see Maronna *et al.* (2006) for a possible algorithm. For the multivariate normal model, a popular choice for the initial estimator is the minimum covariance determinant (MCD) of Rousseeuw (1984). It is defined as the mean and covariance of the $h < n$ points for which the determinant of their (sample) covariance matrix is minimal. It is also affine equivariant and has a high breakdown point. To compute it, one needs algorithms that resample subsets of data on which a search for the optimal subsample is performed. A fast algorithm has been proposed by Rousseeuw and Van Driessen (1999). S-estimators will be used in the linear regression (Chapter 3) and the mixed linear models (Chapter 4). As stated previously, they can however be used in many multivariate data analyses; for examples of robust multivariate data analysis, see e.g. Reimann *et al.* (2008, Chapter 14).

2.4 Statistical Tools for Measuring Tests Robustness

While (robust) estimation is an important aspect of a statistical analysis, (robust) inference is often pivotal in the decision-making process. Many questions arising in practice, such as whether an experimental treatment is effective or not, are often stated in terms of hypotheses to be tested. A common strategy is to define a null hypothesis H_0, in this case that there is 'no difference between treatments', with the hope that it can be refuted by the experiment. An hypothesis test is a procedure based on the data to decide whether there is enough evidence to reject H_0. This is usually done in favor of an alternative hypothesis H_1 that specifies the existence of an effect. In many cases, those null hypotheses can be recast by imposing restrictions to a parameter of interest $\boldsymbol{\theta}$ of an underlying model $F_{\boldsymbol{\theta}}$ having allegedly generated the data $\boldsymbol{x}_i, i = 1, \ldots, n$. Formally, this amounts to testing that $k < q$ components of $\boldsymbol{\theta}$ are zero. If we denote $\boldsymbol{a}^T = (\boldsymbol{a}_{(1)}^T, \boldsymbol{a}_{(2)}^T)$ the partition of a q-dimensional vector $\boldsymbol{a}$ into $q - k$ and k components, and by $A_{(ij)}, i, j = 1, 2$ the corresponding partition of a $q \times q$ matrix A, the null hypothesis is then $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, where $\boldsymbol{\theta}_{0(2)} = 0$ and $\boldsymbol{\theta}_{0(1)}$ unspecified, against the alternative $H_1 : \boldsymbol{\theta}_{0(2)} \neq 0$ and $\boldsymbol{\theta}_{(1)}$ unspecified. Testing H_0 is usually achieved through a test statistic $T = T(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ that rejects H_0 for extreme values[10] of T. This decision can lead to two types of error, i.e. an incorrect conclusion in a situation where the null hypothesis is true, and an incorrect conclusion in a situation where the alternative hypothesis H_1 is true.

To control for the first type or type I error, we define α, the test nominal level, typically 5%, and the corresponding critical region $R = [|T| > t_\alpha]$ as the set of values for which H_0 is rejected. The cut-off point, t_α, is the critical value of the test specified by

$$P(|T| > t_\alpha \mid \boldsymbol{x} \sim F_{\boldsymbol{\theta}_0}) = \alpha. \tag{2.37}$$

Thus, α is the probability of rejecting the null hypothesis H_0 when in fact it is true. To derive the exact value of t_α, one needs the null distribution of T, that is, the

[10] We consider here only a two-sided test, but definition (2.37) can be easily adapted to a one-sided test.

distribution of T under $F_{\boldsymbol{\theta}_0}$. As the latter is not always known, a common practice is to replace it by its asymptotic limit, i.e. use an asymptotic value for t_α independent of the sample size. Note that this asymptotic distribution in general depends on the assumed model $F_{\boldsymbol{\theta}_0}$ (under the null hypothesis) for the data. Alternatively, one can compute, instead of t_α, the test p-value $= P(|T| > |t_{obs}| \mid \boldsymbol{x} \sim F_{\boldsymbol{\theta}_0})$, where t_{obs} is the observed value of T over the sample, and compare it with α. The null hypothesis H_0 is then rejected if the p-value is smaller than α.

To control for the second type of error (or type II error) we usually define a given power the test needs to achieve at the alternative H_1 of interest, say $\boldsymbol{x} \sim F_{\boldsymbol{\theta}_1}$ and $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_0$. The power is defined as the probability of the critical region when H_1 is true, i.e. $P(|T| > t_\alpha \mid \boldsymbol{x} \sim F_{\boldsymbol{\theta}_1})$. Since power is just one minus the probability of a type II error, controlling the type II error to be acceptably low is equivalent to controlling the power to be acceptably high. The power of the test can also be used to choose among several test statistics T, for a given (nominal) level α and a given hypothesis H_0. The statistic with the best power at the alternative of interest is often the recommended option.

In the same way as classical estimators, standard testing procedures are however based on the (strong) assumption that the data-generating process is exactly $F_{\boldsymbol{\theta}_0}$ under H_0 (and $F_{\boldsymbol{\theta}_1}$ under H_1). If instead the data have been generated from a distribution in some neighborhood of the null model, say F_ε, then there is no guarantee that the critical point or the p-value are actually correct. In fact, the actual level of the test $P(|T| > t_\alpha \mid \boldsymbol{x} \sim F_\varepsilon)$ can be substantially greater than the nominal level α, which casts doubts over the validity of the procedure. For example, consider the F-test for equality of two variances and its generalization to m samples (Bartlett's test). Box (1953) investigated its actual level when the data-generating process was the t-distribution (instead of the normal distribution) with different degrees of freedom including the normal as a limit case.[11] For $m = 2$, the actual test level is several times as large as the nominal level 5%, e.g. 11% when the data come from the Student distribution with 10 degrees of freedom t_{10}, and 16.6% when t_7 is the data-generating process. More extreme actual test levels up to 48% are even obtained for $m = 10$ groups, see Hampel *et al.* (1986, p. 388) for details. The Bartlett's test is very unstable under small departures from the model (through departures from the data-generating process).

The lack of level stability is called non-*robustness of validity*. Other classical testing procedures show a less dramatic behavior but are still problematic from a robustness standpoint. For instance, the one-sample t-test cannot maintain its (nominal) level in the presence of asymmetric contamination (see Beran (1981), Subrahmanian *et al.* (1975) and others), the effect being worse in small samples. This also happens with confidence intervals (their length become unreliable) since they are based on the same sample information as t-tests. Even if the t-test is approximately valid (actual level approximately equal to the nominal level) when the data-generating process is symmetric but non-normal, it cannot be considered

[11] The t-distribution can be considered as a particular distribution belonging to the neighborhood F_ε of the normal distribution.

as a robust testing procedure. Moreover, the t-test can suffer from a drastic drop in power when the data-generating process under H_1 is contaminated; see Hampel *et al.* (1986, p. 405). This type of instability (i.e. affecting power) is called non-*robustness of efficiency*. The two-sample t-test also exhibits similar deficiencies and this is better illustrated through the following example.

2.4.1 Sensitivity of the Two-sample t-test

The lack of stability of many classical inferential procedures also affects the corresponding p-value. Everitt (1994, pp. 25–27), provides the IQ scores of 94 children of age 5; fifteen of these children have mothers suffering from post-natal depression (group A) whereas the 79 others have healthy mothers (group B). Researchers are interested in testing whether group A children have a different IQ to those of group B. The null hypothesis H_0 of 'no difference between the mean IQs across the two groups' is typically tested by means of a two-sample t-test. Unfortunately, in that particular sample, most IQ values are between 80 and 144 with the exception of two small values of 22 and 48 for one child in each group A and B, respectively. Applying the Student t-test to the data returns a p-value of 0.016 supporting a difference in IQ between the two groups, whereas an inconclusive result ($p = 0.07$) is observed when the two outliers are deleted. To further illustrate the erratic behavior of the p-value, we moved around the IQ score of the kid with the lowest score in group B (case 2 in that group). A plot of the two-sample t-test p-value versus hypothetical IQ values between 20 and 100 for that child's IQ score is given in Figure 2.4. This includes the original value of 22. We also added the p-value of a Wilcoxon test for comparison. The t-test p-value varies from 0.015 to 0.12, which further demonstrates its instability. Even if the observed IQ score of 22 is a genuine observation, it has too big an impact on the test decision. Hence, relying on the corresponding p-value of 0.0156 (and, thus, rejecting the null hypothesis) is extremely hazardous. In contrast, the Wilcoxon rank-sum test is a safer procedure that returns a relatively stable p-value around 6–8%. Note that it starts to move up a bit when the hypothetical score for child 2 in group B is set to 85 or above. The reason is that the ranks that the Wilcoxon test are based on are more disrupted when the IQ score is pushed back to the bulk of the data. This example also allows us to illustrate the behavior of the Bartlett test (or F-test) mentioned above on real data. When applied to the original sample it strongly rejects the null hypothesis of equal variance (p-value $= 0.0003$) but it is clearly inconclusive (p-value $= 0.13$) when the two lowest IQ values are removed. This is a blatant illustration on how unreliable this procedure can be and why, in our view, it should be avoided.

In the following section, we present statistical tools for measuring the stability of a testing procedure, both in terms of robustness of validity and efficiency.

2.4.2 Local Stability of a Test: the Univariate Case

In the robustness paradigm, the objective is twofold: a test must have (i) a stable type I error (or level) under small, arbitrary departures from the null hypothesis (robustness of validity) and (ii) a good power under small arbitrary departures from

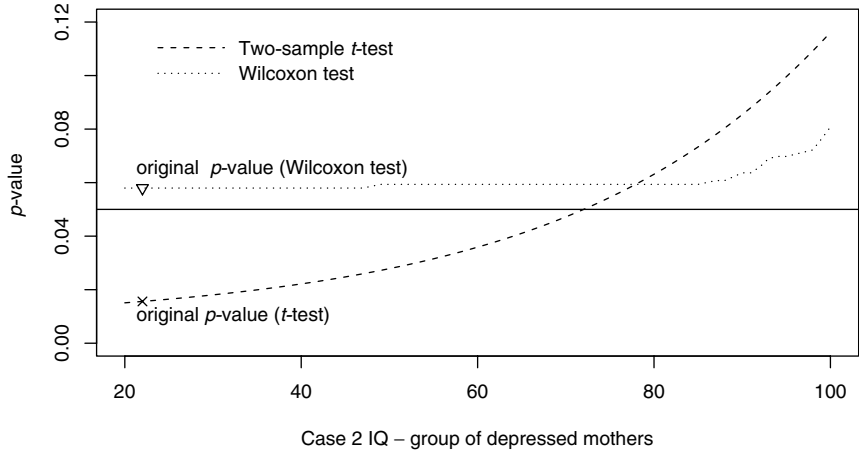

Figure 2.4 p-value for the two-sample t-test and Wilcoxon test when the IQ score of case 2 in group B is changed.

the specified alternative (robustness of efficiency) to be declared locally robust. The wording 'small departures' generally refers to a neighborhood of the data-generating process, under the null or the alternative. In this section, we formalize these concepts in univariate parametric models following the pioneering work of Ronchetti (1982b,a), and Rousseeuw and Ronchetti (1979, 1981). The (natural) extension to the multivariate case is discussed in Section 2.5.5.

Consider T a test statistic computed on a sample $x_1, x_2, \ldots, x_n$, for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$. Also, let the (nominal) level of the test be α. A simple example is the test of the mean $\theta = \mu$ of the (univariate) normal distribution with known variance σ. A standard test could then be the z-statistic $T := z = \sqrt{n}(\bar{x} - \theta_0)/\sigma$ with critical region $\{z \geq z_\alpha\}$. To measure the stability of the testing procedure under (small) model misspecifications, the idea is first to compute the asymptotic level of the test under model misspecification and compare it with its nominal level. For that, we need the so-called 'level contaminated model'

$$F_{\varepsilon,n}^{\mathrm{L}} = (1 - \varepsilon/\sqrt{n})F_{\theta_0} + \varepsilon/\sqrt{n}G, \tag{2.38}$$

where G is an arbitrary distribution. This is similar to (1.1) and constitutes a neighborhood of the null hypothesis. The only difference is that (2.38) 'shrinks' at a rate $1/\sqrt{n}$, i.e. $\varepsilon_n = \varepsilon/\sqrt{n}$ replaces ε in (1.1). This type of neighborhood is often chosen (Hampel *et al.*, 1986, Huber-Carol 1970; Rieder 1978) because it depends on the sample size n in the same manner as the sequence of contiguous alternatives are formulated, i.e. $H_1 : \theta = \theta_1 = \theta_0 + \delta/\sqrt{n}$ with $\delta > 0$, mainly for power determination. Indeed, the power of the test is computed for alternatives close to the null hypothesis, so that the test always become more powerful as the sample size increases. It is then necessary to choose an amount of contamination that converges to zero at the same rate that θ_1 converges to θ_0 to avoid overlapping

between the neighborhood of the null $F^{\mathrm{L}}_{\varepsilon,n}$, and that of the alternative, say $F^{\mathrm{P}}_{\varepsilon,n}$. The latter is called the 'power contaminated model' and is defined as (2.38) with the alternative model F_{θ_1} replacing F_{θ_0} in the formula.

Let $\alpha(F^{\mathrm{L}}_{\varepsilon,n})$ be the *actual* level of the test when the data are generated by (2.38) and $\alpha(F_{\theta_0}) = \alpha_0$ the nominal level, i.e. the (asymptotic) level of the test when the data are generated exactly from F_{θ_0}. The question of interest is how $\alpha(F^{\mathrm{L}}_{\varepsilon,n})$ compares to α_0. An answer is provided by Ronchetti (1982a,b), and Rousseeuw and Ronchetti (1979, 1981) who derived the following approximation[12] valid for small values of ε and large sample sizes n:

$$\alpha(F^{\mathrm{L}}_{\varepsilon,n}) \simeq \alpha_0 + \varepsilon \int IF(z; \alpha, F_{\theta_0})\, dG(z), \tag{2.39}$$

where $IF(z; \alpha, F_{\theta_0})$ is the level influence function given by

$$IF(z; \alpha, F_{\theta_0}) = \varphi(\Phi^{-1}(1-\alpha_0))IF(z; T, F_{\theta_0})/[V(T, F_{\theta_0})]^{1/2}, \tag{2.40}$$

with $\Phi^{-1}(1-\alpha_0)$ the $1-\alpha_0$ quantile of the standard normal distribution Φ with density φ, $IF(z; T, F_{\theta_0})$ the IF of T and $V(T, F_{\theta_0}) = \int IF(z; T, F_{\theta_0})^2\, dF_{\theta_0}(z)$ its asymptotic variance (see also (2.3)). A direct consequence of this result is that the stability of the test level under small departures from the assumed model (i.e. under (2.38)) is proportional to the IF of the test statistic T.

A similar approach can be followed for the test (asymptotic) power. Let $\beta(F^{\mathrm{P}}_{\varepsilon,n})$ be the actual power and β_0 the nominal power. The following approximation can also be obtained:

$$\beta(F^{\mathrm{P}}_{\varepsilon,n}) \simeq \beta_0 + \varepsilon \int IF(z; \beta, F_{\theta_0})\, dG(z), \tag{2.41}$$

where $IF(z; \beta, F_{\theta_0})$, the power influence function, is again proportional to the IF of the test statistic T, i.e.

$$IF(z; \beta, F_{\theta_0}) = \varphi(\Phi^{-1}(1-\alpha_0) - \delta\sqrt{E})IF(z; T, F_{\theta_0})/[V(T, F_{\theta_0})]^{1/2}, \tag{2.42}$$

where $E = [\xi'(\theta_0)]^2 / V(T, F_{\theta_0})$ is the Pitman's efficacy of the test with $\xi(\theta) = T(F_\theta)$; see Hampel *et al.* (1986, (3.2.14)). The nominal (asymptotic) power (under the true model) is $\beta_0 = 1 - \Phi(\Phi^{-1}(1-\alpha_0) - \delta\sqrt{E})$ (defined as the limit of the power at the alternative θ_1 when $n \to \infty$). Note that, as the sequence of contiguous alternatives tends to the null hypothesis, the power influence function depends on F_{θ_0}.

An important consequence of these results is that bounding the IF of the test statistic guarantees the local stability of the level and power with respect to small deviations from the assumed model. More specifically, if we denote $\gamma^*_{test} = \sup_z |IF(z; T, F_{\theta_0})/[V(T, F_{\theta_0})]^{1/2}|$ the (standardized) gross error sensitivity of the test, then by means of (2.39)–(2.42), we obtain a maximum asymptotic level α_{max}

[12] This approximation is based on functional expansions of von Mises (1947) and Fernholz (1983).

and minimum asymptotic power β_{min} over the neighborhood, i.e.

$$\alpha_{max} \simeq \alpha_0 + \varepsilon\varphi(\Phi^{-1}(1-\alpha_0))\gamma^*_{test},$$
$$\beta_{min} \simeq \beta_0 - \varepsilon\varphi(\Phi^{-1}(1-\alpha_0) - \delta\sqrt{E})\gamma^*_{test}.$$

Unlike the z-test, the sign and one-sample Wilcoxon tests statistics have a bounded *IF* and, hence, are robust testing procedures with stable level and power under (slight) model misspecification. Details and numerical values of α_{max} and β_{min} for various values of ε and δ can be found in Hampel *et al.* (1986, pp. 200–201). They also show that the level of the sign test is slightly more stable than that of the Wilcoxon test. The latter has more power at the normal model (uncontaminated case) but it quickly loses its advantage as the percentage of contamination increases. For $\varepsilon \geq 15\%$ the sign test will generally outperform the Wilcoxon test in terms of robustness of the level and power.

2.4.3 Global Reliability of a Test: the Breakdown Functions

Good local robustness properties of a test do not guarantee that the level and power of the test will remain stable in the presence of large deviations (also called global robustness). This concept is vague but includes situations where a large proportion of outliers arise in the data (e.g. large values of ε in (2.38)), or when the working model is further away from the assumed model. A second step in the robustness analysis would then be the computation of the breakdown point of the tests, in the same spirit as is done for estimators (see Section 2.2.2). We feel, however, that 'local considerations are relevant for inference, which is more meaningful in smaller neighborhoods of the assumed model'; see He *et al.* (1990, p. 447). In other words, trying to draw inference using a model that is grossly misspecified is hazardous. In our view, the notion of a high breakdown point is not as critical for tests as it is for estimators.

The first important contribution in this respect is due to Ylvisaker (1977) who defined the concept of resistance of a test. We describe it here for a one-sided test with critical region $\{T \geq t_\alpha\}$. The resistance to rejection ρ_R (respectively resistance to acceptance ρ_A) of the test is defined as the smallest proportion m_0/n of sample observations $x_1, \ldots, x_{m_0}$ for which the observed test statistic T (computed on the whole sample) is such that $T \geq t_\alpha$ (respectively $T < t_\alpha$), no matter what $x_{m_0+1}, \ldots, x_n$ are. In other words, for ρ_A, there is at least one sample of size $n - (n\rho_A - 1)$ which suggests rejection so strongly that this decision cannot be overruled by the remaining $n\rho_A - 1$ observations. This definition attempts to capture the strength of the reject–not reject decision of the test that should not be reversed by extreme observations if the test is robust. For instance, the z-test defined by the critical region $\bar{x}_n \geq t_\alpha$ has a resistance to acceptance and a resistance to rejection of $1/n$. Extensions of this idea can be found in Hettmansperger (1984) and Zhang (1996).

A more general approach can be found in He *et al.* (1990) and He (1991) in a more mathematical framework. They introduce the concept of level breakdown

function $\varepsilon^{\mathrm{L}}(\theta)$ and power breakdown function $\varepsilon^{\mathrm{P}}(\theta)$.[13] Here $\varepsilon^{\mathrm{L}}(\theta)$ is the smallest amount of contamination to the null distribution that drives the test statistic to any particular value it could take under the alternative. In a similar fashion, $\varepsilon^{\mathrm{P}}(\theta)$ is the smallest amount of contamination to a specific alternative distribution F_θ necessary to drive the test to a particular value of the test statistic under the null distribution. An interesting consequence of these definitions is that the directional derivatives of $\varepsilon^{\mathrm{L}}(\theta)$ at a boundary point between the null and the alternative (here θ_0) indicates the local robustness stability of the test level. In particular, if this derivative is zero, then the test does not have a stable level in a neighborhood of the null model (2.38). By analogy with the breakdown point of an estimator, one can also compute the level breakdown point ε^{L} as the supremum of $\varepsilon^{\mathrm{L}}(\theta)$ over all values of the parameter of interest θ. The power breakdown point ε^{P} is then $\sup_\theta \varepsilon^{\mathrm{P}}(\theta)$. These two quantities are less precise than their respective breakdown functions but can constitute useful summary measures.

As an illustration, consider the one-sample t-test $H_0 : \theta = 0$ in a normal model $\mathcal{N}(\theta, \sigma^2)$. He *et al.* (1990) showed that $\varepsilon^{\mathrm{L}}(\theta) = (\theta/\sigma)^2/(1 + \theta/\sigma)^2)$ and $\varepsilon^{\mathrm{P}}(\theta) = 0$. The slope of $\varepsilon^{\mathrm{L}}(\theta)$ at $\theta = 0$ is zero and the test does not have robustness of validity. This confirms theoretically the empirical findings of Subrahmanian *et al.* (1975) and Beran (1981) reported at the beginning of Section 2.4. It is worth noting that the level breakdown point ε^{L} tends to one ($\varepsilon^{\mathrm{L}}(\theta) \to 1$) when θ goes to infinity. This can partially explain why many researchers still believe that the level of the t-test is robust. However, $\varepsilon^{\mathrm{P}}(\theta) = 0$ which reflects its lack of robustness of efficiency. On the other hand, $\varepsilon^{\mathrm{L}}(\theta)$ for the Wilcoxon (or sign) test has a positive slope at zero (see Figure 1 in He *et al.* (1990)), which further justifies the level stability reported at the beginning of Section 2.4.

The concepts presented in this section can be extended to multivariate hypothesis in a general parametric model but it is often difficult to derive the exact values of $\varepsilon^{\mathrm{P}}(\boldsymbol{\theta})$ and $\varepsilon^{\mathrm{L}}(\boldsymbol{\theta})$ in that setting. In the next section, we therefore return only to the local robustness of tests, extending the ideas of Section 2.4.2 to the more general case.

2.5 General Approaches for Robust Testing

As stated before, many hypotheses of interest can be formulated as restrictions to a parameter of interest $\boldsymbol{\theta}$ from the model $F_{\boldsymbol{\theta}}$, i.e. $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}_{0(2)} = 0$, $\dim \boldsymbol{\theta}_{0(2)} = k$ and $\boldsymbol{\theta}_{0(1)}$ unspecified, against the alternative $H_1 : \boldsymbol{\theta}_{0(2)} \neq 0$ and $\boldsymbol{\theta}_{0(1)}$ unspecified. For such a general hypothesis, three tests, namely the Wald, score and likelihood ratio tests (LRTs) are available from the classical inference theory; see, for instance, Rao (1973), Cox and Hinkley (1992) or Welsh (1996). We start with a brief review and a geometrical interpretation in Sections 2.5.1 and 2.5.2. It is intuitively clear that the three tests lack stability as they are based on the MLE, which is itself non-robust. In Section 2.5.3, an extension of these three classes

[13] We omit here the exact definitions based on functionals and the reader is referred to He *et al.* (1990) and He (1991) for details.

based on M-estimators is presented, among which more stable alternative test statistics can be found. A more formal treatment of robustness issues is given in Section 2.5.5. Therefore, this section gives a general framework for robust testing; specific inferential procedures for each particular model (i.e. linear regression, GLM, mixed models, etc.) are discussed in later chapters.

2.5.1 Wald Test, Score Test and LRT

Consider again a sample $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ of n i.i.d. observations from a parametric model $F_{\boldsymbol{\theta}}$, with corresponding density $f(\boldsymbol{x}; \boldsymbol{\theta})$. To compute the different tests statistics, one needs to estimate $\boldsymbol{\theta}$. There are actually two estimates, $\hat{\boldsymbol{\theta}}_{[MLE]}$ and $\dot{\boldsymbol{\theta}}_{[MLE]}$, which are the MLE at the full model $F_{\boldsymbol{\theta}}$ and the MLE at the reduced model $F_{\boldsymbol{\theta}_0}$, respectively. Weak regularity conditions are usually needed for these two estimators to exist, be consistent and asymptotically normally distributed with asymptotic variance $V = V(\hat{\boldsymbol{\theta}}_{[MLE]}, F_{\boldsymbol{\theta}})$ given in (2.30).

The Wald test statistic originally proposed by Wald (1943) is a quadratic form of the MLE second component $\hat{\boldsymbol{\theta}}_{[MLE](2)}$ at the full model, i.e.

$$W^2 = n\hat{\boldsymbol{\theta}}_{[MLE](2)}^T V_{(22)}^{-1} \hat{\boldsymbol{\theta}}_{[MLE](2)}. \tag{2.43}$$

In practice, the inverse of the block (22) of $V = V(\hat{\boldsymbol{\theta}}_{[MLE]}, F_{\boldsymbol{\theta}}) = I(\boldsymbol{\theta})^{-1}$, where $I(\boldsymbol{\theta})$ is the Fisher information matrix, needs to be estimated to compute (2.43). This can be done by taking the sample analog for I computed at the MLE in the full model, i.e.

$$\hat{I} = \hat{I}(\hat{\boldsymbol{\theta}}_{[MLE]}) = \frac{1}{n} \sum s(\boldsymbol{x}_i; \hat{\boldsymbol{\theta}}_{[MLE]}) s(\boldsymbol{x}_i; \hat{\boldsymbol{\theta}}_{[MLE]})^T.$$

A score (or Rao) test is based on the test statistic

$$R^2 = n Z_n^T C^{-1} Z_n, \tag{2.44}$$

where

$$Z_n = \frac{1}{n} \sum_{i=1}^{n} s(\boldsymbol{x}_i; \dot{\boldsymbol{\theta}}_{[MLE]})_{(2)}$$

and $\dot{\boldsymbol{\theta}}_{[MLE]}$ is the MLE at the reduced model (i.e. under H_0), and is defined as the solution in $\boldsymbol{\theta}_{(1)}$ of

$$\sum_{i=1}^{n} s(\boldsymbol{x}_i; \boldsymbol{\theta})_{(1)} = 0, \tag{2.45}$$

with $\boldsymbol{\theta}_{(2)} = 0$. The matrix C is $C = I_{22.1} V_{(22)} I_{22.1}^T$ with $I_{22.1} = I_{(22)} - I_{(21)} I_{(11)}^{-1} I_{(12)}$, and is estimated by its sample analog computed at the MLE under the null hypothesis, i.e. $\dot{\boldsymbol{\theta}}_{[MLE]}$. It is worth noting that, from (2.44), R^2 is simply a quadratic form in Z_n, standardized by its asymptotic variance. The score test is also known as the Lagrange multiplier test (see e.g. Welsh (1996, p. 223) for details).

Let $l(\boldsymbol{\theta}; \boldsymbol{x}) = \log f(\boldsymbol{x}; \boldsymbol{\theta})$ be the log-likelihood function, the LRT statistic is given by

$$\text{LRT} = 2 \sum_{i=1}^{n} [l(\hat{\boldsymbol{\theta}}_{[MLE]}; \boldsymbol{x}_i) - l(\dot{\boldsymbol{\theta}}_{[MLE]}; \boldsymbol{x}_i)], \tag{2.46}$$

i.e. twice the logarithm of the likelihood ratio, hence the name LRT. As a simple illustration, we consider the example of linear regression[14] $y = \boldsymbol{x}^T \boldsymbol{\theta} + \varepsilon$, where the error term ε follows a normal distribution $\mathcal{N}(0, \sigma^2)$ (with for simplicity $\sigma = 1$). Elementary algebra shows that, up to a constant, the log-likelihood is $l(\hat{\boldsymbol{\theta}}_{[MLE]}; \boldsymbol{x}, y) = -r^2/2$ where $r = y - x^T \hat{\boldsymbol{\theta}}_{[MLE]}$. The LRT statistic is then the difference in the sums of squared residuals computed at the full and reduced models, respectively.

Under the null hypothesis, the three tests are asymptotically distributed as a χ^2 distribution with $k = \dim(\boldsymbol{\theta}_{(2)})$ degrees of freedom. Large values of the test statistics on that scale indicate evidence for the rejection of the null hypothesis. This result, however, is based on the (strong) assumption that the data have been generated exactly from $F_{\boldsymbol{\theta}_0}$, which might not be exactly true. As all three tests are asymptotically equivalent, any of them could, in principle, be used for testing a multivariate hypothesis of the type $H_0 : \boldsymbol{\theta}_{(2)} = 0$ for any parametric model.

2.5.2 Geometrical Interpretation

In order to better understand the difference between the LRT, score and Wald tests, we present here their geometrical interpretation. Without loss of generality, we assume that $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$ is a two-dimensional vector and $H_0 : \theta_2 = 0$ is the null hypothesis to be tested. A plot of the log-likelihood $l(\boldsymbol{\theta}; \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \sum \log f(\boldsymbol{x}_i; \boldsymbol{\theta})$ versus $\boldsymbol{\theta}$ is displayed in Figure 2.5.[15] The LRT statistic is equal to twice the vertical distance AB which measures the difference between the overall maximum and the maximum under the null. The Wald test is based on the horizontal distance OC properly standardized. In the one-dimensional case, the Wald test statistic is the square of the z-statistic where $z = \hat{\theta}_{[MLE]2}/SE(\hat{\theta}_{[MLE]2})$, i.e. the MLE of θ_2 (estimated together with θ_1) divided by its standard error. This means that the estimate has to be 'measured' in the metric given by its asymptotic variance. The score test is based on the slope of the log-likelihood at E or, more precisely, on some norm of the total score computed at $\dot{\boldsymbol{\theta}}_{[MLE]}$, the MLE at the reduced model. Once again the total score is standardized. Each of these tests rejects H_0 when the corresponding distance or norm is sufficiently large.

2.5.3 General Ψ-type Classes of Tests

The classical tests often have poor robustness properties as they are based on the MLE, which is usually non-robust. A natural way to overcome this deficiency is to

[14] This example is purely illustrative as the classical LRT test is never used in that setting: the exact F-test is available and is routinely used instead.

[15] Reproduced up to a notational difference, with permission, from Heritier and Victoria-Feser (1997).

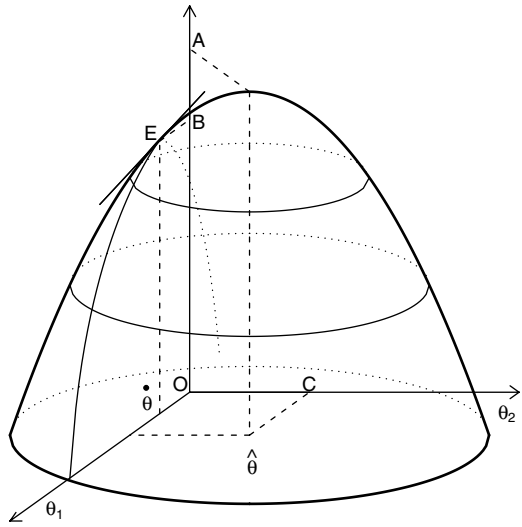

Figure 2.5 Geometry of the three classical tests.

rely upon M-estimators presented in Section 2.3.1 for the construction of the tests statistics. Intuitively, one can expect their good robustness properties to be carried over to the tests. Specifically, the log-likelihood function $\log f(\boldsymbol{x};\boldsymbol{\theta})$ is replaced by a suitable $\rho(\boldsymbol{x};\boldsymbol{\theta})$ function, the score function by $\Psi(\boldsymbol{x};\boldsymbol{\theta})$ which may be equal to the derivative of ρ with respect to $\boldsymbol{\theta}$ and use M-estimators $\hat{\boldsymbol{\theta}}_{[M]}$ based on ρ (or Ψ) instead of the MLE. This generates three classes of test statistics, extending the three classical tests.

The Wald-type test statistic is a quadratic form of the second component $\hat{\boldsymbol{\theta}}_{[M](2)}$ of an M-estimator of $\boldsymbol{\theta}$ (based on a function Ψ)

$$W_{\Psi}^2 = n\hat{\boldsymbol{\theta}}_{[M](2)}^T V(\hat{\boldsymbol{\theta}}_{[M]}, F_{\boldsymbol{\theta}})_{(22)}^{-1}\hat{\boldsymbol{\theta}}_{[M](2)} \tag{2.47}$$

with $V(\hat{\boldsymbol{\theta}}_{[M]}, F_{\boldsymbol{\theta}})_{(22)}^{-1}$ the inverse of the block (22) of the asymptotic variance of the M-estimator (2.27). As before, $V(\hat{\boldsymbol{\theta}}_{[M]}, F_{\boldsymbol{\theta}})_{(22)}$ needs to be estimated to obtain a numerical value for W_{Ψ}^2; this is typically achieved by computing its sample analog with $\boldsymbol{\theta}$ replaced by the robust estimate $\hat{\boldsymbol{\theta}}_{[M]}$. If $\Psi = s$, the score function, (2.47) is the classical Wald test (2.43). In the regression model, (2.47) is also the robust Wald test discussed by Markatou *et al.* (1991, p. 205) and Silvapulle (1992).

A score- (or Rao-) type test is based on the test statistic

$$R_{\Psi}^2 = nZ_n^T C^{-1} Z_n, \tag{2.48}$$

where $Z_n = (1/n)\sum_{i=1}^{n} \Psi(\boldsymbol{x}_i; \dot{\boldsymbol{\theta}}_{[M]})_{(2)}$, $\dot{\boldsymbol{\theta}}_{[M]}$ is the M-estimator in the reduced model, i.e. the solution in $\boldsymbol{\theta}_{(1)}$ of

$$\sum_{i=1}^{n} \Psi(\boldsymbol{x}_i; \boldsymbol{\theta})_{(1)} = \mathbf{0}, \tag{2.49}$$

with $\boldsymbol{\theta}_{(2)} = 0$. The matrix C is $C = M_{22.1} V_{(22)} M_{22.1}^T$ and $M_{22.1} = M_{(22)} - M_{(21)} M_{(11)}^{-1} M_{(12)}$ is derived from $M = M(\Psi, F_{\boldsymbol{\theta}_0})$ given by (2.29) and $V_{(22)} = V(\hat{\boldsymbol{\theta}}_{[M]}, F_{\boldsymbol{\theta}})_{(22)}$. As before for the Wald-type test, C has to be estimated and its sample analog computed at $\dot{\boldsymbol{\theta}}_{[M]}$ can be used. The test statistic is a quadratic form in Z_n standardized by its asymptotic variance. When $\Psi = s$, (2.48) is the classical score test (2.44). In the regression model, (2.48) is the robust score test proposed by Markatou and Hettmansperger (1990).

The LRT-type test (or ρ-test) statistic is given by

$$\mathrm{LRT}_\rho = 2\sum_{i=1}^{n} [\rho(\boldsymbol{x}_i; \hat{\boldsymbol{\theta}}_{[M]}) - \rho(\boldsymbol{x}_i; \dot{\boldsymbol{\theta}}_{[M]})], \tag{2.50}$$

where $\hat{\boldsymbol{\theta}}_{[M]}$ and $\dot{\boldsymbol{\theta}}_{[M]}$ are the M-estimators in the full and reduced models, respectively, and ρ is such that $\Psi(\boldsymbol{x}; \boldsymbol{\theta}) = \partial\rho(\boldsymbol{x}; \boldsymbol{\theta})/\partial\boldsymbol{\theta}$. When ρ is the log-likelihood function, (2.50) is the classical LRT statistic (2.46).

The choice of the ρ-function is model-specific and should be guided by robustness considerations. For a robust test, ρ should be such that $\Psi(\boldsymbol{x}; \boldsymbol{\theta})$, the derivative of $\rho(\boldsymbol{x}; \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, is bounded (see Section 2.3.1). Since it is often easier to define a robust estimator through its Ψ function rather than through its ρ function, then the ρ-test does not necessarily exist as Ψ may not be the derivative of a ρ-function. On the other hand, the computation of the Wald- (respectively score-)type test statistic requires only the estimation of $\boldsymbol{\theta}$ at the full (respectively reduced) model, which may be advantageous in complex models. In other words, no specification of a ρ-function is needed to derive W_Ψ^2 and R_Ψ^2. This makes these two classes fairly general in their definition and applicability. However, LRT_ρ is potentially more powerful in small samples as it captures more information from the data through the ρ-function (corresponding to the log-likelihood function in the classical LRT). Fortunately, for the models presented in this book, solutions exist for defining robust LRT_ρ test statistics.

2.5.4 Asymptotic Distributions

Heritier (1993) and Heritier and Ronchetti (1994) show that, under weak regularity conditions, W_Ψ^2 and R_Ψ^2 have the same asymptotic distribution as their classical counterparts, namely a (central) χ_k^2 under the null hypothesis. If, in addition, $\rho(\boldsymbol{x}; \boldsymbol{\theta})$ is defined in such a way that $M = M(\Psi; F_{\boldsymbol{\theta}})$ in (2.29) is symmetric positive definite, the LRT-type statistic LRT_ρ has asymptotically the same null distribution as a weighted sum of k independent χ_1^2 random variables. The weights are the positive

eigenvalues of the matrix

$$Q[M^{-1} - (M^*)^+],$$

where Q is given by (2.28), $(M^*)^+$ is a $q \times q$ matrix derived from M where blocks (12), (21), (22) are zero, and block (11) is $M_{(11)}^{-1}$. The p-values are computed using standard algorithms (see e.g. Wood, 1989; Farebrother, 1990). These results generalize earlier findings obtained for the linear regression model by Ronchetti (1982b), Markatou and Hettmansperger (1990) and Silvapulle (1992). It is worth stressing that, unlike in the classical case, the three tests are not asymptotically equivalent.

2.5.5 Robustness Properties

Local robustness properties of the classes of tests introduced in Section 2.5.3 can be studied by extending the approach based on the *IF* presented in Section 2.4.2 in the one-dimensional case. We follow here the developments of Heritier and Ronchetti (1994) and Cantoni and Ronchetti (2001b). As with the one-dimensional case, we start with robustness of validity. Again we consider a 'shrinking' neighborhood of the model under the null hypothesis, defined in the exact same manner as (2.38). The only difference is that now $\boldsymbol{\theta}_0$ is a multidimensional parameter. Data generated from $F_{\varepsilon,n}^{\mathrm{L}} = (1 - \varepsilon/\sqrt{n})F_{\boldsymbol{\theta}_0} + \varepsilon/\sqrt{n}G$ will occasionally contain observations coming from the contaminating distribution G. For instance, if the true model under the null is a Poisson distribution with parameter $\boldsymbol{\theta}_0$ possibly depending on some covariates, G could be another Poisson distribution unrelated to the covariates, a 'nastier' contamination such as a point-mass distribution as in (1.2) or any model for count data. As explained earlier, the comparison of the actual asymptotic level of the test $\alpha(F_{\varepsilon,n}^{\mathrm{L}})$ and the nominal level α_0 is of primary interest.

Heritier and Ronchetti (1994) studied the Wald- and score-type tests and derive the following level expansion, valid for a large sample size n and small amount of contamination ε:

$$\alpha(F_{\varepsilon,n}^{\mathrm{L}}) \simeq \alpha_0 + \varepsilon^2 \cdot \mu \left\| \int IF(x;\, U,\, F_{\boldsymbol{\theta}_0})\, dG(x) \right\|^2, \tag{2.51}$$

where $\mu = -(\partial/\partial\delta) H_k(\eta_{1-\alpha_0};\, \delta)|_{\delta=0}$, $H_k(.;\, \delta)$ is the cumulative distribution function of a $\chi_k^2(\delta)$ distribution, $\eta_{1-\alpha_0}$ the $1 - \alpha_0$ quantile of the central χ_k^2 distribution, and U is the functional defining the Wald- or score-type test statistics.[16] It is worth stressing an important difference with the level approximation (2.39) for the one-dimensional case, i.e. the difference $\alpha(F_{\varepsilon,n}^{\mathrm{L}}) - \alpha_0$ derived from (2.51) now has a leading term proportional to a quadratic term in ε. The proper quantity to bound to have a stable level in a neighborhood around the null is the *IF* of the functional $U(F)$.

[16] Both test statistics W_Ψ^2 and R_Ψ^2 can be written as a quadratic form $nU(F^{(n)})U^T(F^{(n)})$ where $F^{(n)}$ is the empirical distribution. The functional $U(F)$ is then $U_W(F) = V_{(22)}^{-1/2}\hat{\boldsymbol{\theta}}_{[M](2)}(F)$ for the Wald-type test, and $U_R(F) = C^{-1/2}Z(F)$ for the score-type test. In this definition, $\hat{\boldsymbol{\theta}}_{[M]}(F)$ and $Z(F)$ are, respectively, the functionals associated to the M-estimator and Z_n, and $A^{-1/2}$ is the Choleski root of A^{-1}, the inverse of a symmetric positive-definite matrix A.

If, in particular, we choose $G = \Delta_z$ a point mass contamination, (2.51) reduces to

$$\alpha(F^{\mathrm{L}}_{\varepsilon,n}) \simeq \alpha_0 + \varepsilon^2 \cdot \mu \| IF(z; \hat{\boldsymbol{\theta}}_{[M](2)}, F_{\boldsymbol{\theta}_0}) \|^2_s,$$

where

$$\| IF(z; \hat{\boldsymbol{\theta}}_{[M](2)}, F_{\boldsymbol{\theta}_0}) \|_s = [IF(z; \hat{\boldsymbol{\theta}}_{[M](2)}, F_{\boldsymbol{\theta}_0})^T V^{-1}_{(22)} IF(z; \hat{\boldsymbol{\theta}}_{[M](2)}, F_{\boldsymbol{\theta}_0})]^{1/2}$$

is the self-standardized *IF* of the second component $\hat{\boldsymbol{\theta}}_{[M](2)}$ (see Hampel *et al.*, 1986, p. 228) and $V_{(22)} = V(\hat{\boldsymbol{\theta}}_{[M]}, F_{\boldsymbol{\theta}})_{(22)}$. Therefore, to obtain robust Wald- and score-type tests we have to bound the *IF* of the underlying M-estimator.[17] As mentioned earlier, (2.47) and (2.48) cannot be computed without estimating the asymptotic variance $V_{(22)}$ and the matrix C. This also has to be done in a robust fashion.

The case of the LRT-type test is more complicated but a similar level approximation to (2.51) can be derived (see Cantoni and Ronchetti, 2001b)[18] yielding a similar conclusion.

Next we can consider the problem of robustness of efficiency. Again, the idea is to define a neighborhood of the contiguous alternatives $H_1 : \boldsymbol{\theta}_{(2)} = \boldsymbol{\theta}_{0(2)} + \Delta/\sqrt{n}$, where Δ is any k-dimensional vector, and compare the actual power $\beta(F^{\mathrm{P}}_{\varepsilon,n})$ to the nominal power β_0. Similar derivations to that performed for the level can be carried out; see Section 4 of Ronchetti and Trojani (2001). They show that the same condition, i.e. a bounded Ψ-function, is required to ensure that the power of the three tests remains stable in the neighborhood. Robustness of efficiency is then achieved without additional requirements. It is worth noting that, although theoretical results have only been obtained in shrinking neighborhoods, simulations (under different models and settings) show that stability of the robust tests is often satisfied over full neighborhoods of the form (1.1).

Finally, as for the one-dimensional case, global robustness of the tests can be investigated as a complement. In general, if both the M-estimator and the covariance matrix estimate of V (and also C) have a good global robustness property (i.e. a high breakdown point), the Wald- (respectively score-)type test also have global robustness properties; see Markatou and He (1994), Copt and Victoria-Feser (2006) and Copt and Heritier (2007). Theoretical results are not available in general with the exception of Markatou and He (1994) for the linear regression. They clearly show that the robust tests cannot do better than the estimators in terms of breakdown point; see also He *et al.* (1990).

[17] More exactly, robustness of validity is guaranteed if the second part of the *IF* of the underlying M-estimator, i.e. the component related to the parameter to be tested $\boldsymbol{\theta}_{(2)}$, is bounded. However, as the first part of the parameter $\boldsymbol{\theta}_{(1)}$ is generally unknown, and hence needs to be estimated simultaneously with $\boldsymbol{\theta}_{(2)}$, it is highly recommended to bound the whole $\Psi(x; \boldsymbol{\theta})$ function.

[18] They use the fact that LRT_ρ can be approximated asymptotically by a quadratic form under the additional condition that $M = M(\Psi; \boldsymbol{\theta})$ of formula (2.29) is symmetric positive definite.

3

Linear Regression

3.1 Introduction

The linear regression model is probably the most widely used model in many sciences such as the biological, medical, economics, behavioral and social sciences. It is the simplest model used to describe possible relationships between variables, more precisely, between a response variable and a set of so-called explanatory variables that supposedly explain it. The relationship between the variables is assumed to be linear, and this is why the linear regression model has existed for such a long time. Indeed the term regression in association with a linear relationship between two variables was used by Sir Francis Galton in his famous study of the average heights of sons of very tall and very short parents, called *Regression Toward Mediocrity in Hereditary Stature* and published in 1885. Although the linearity of the relationship between the response and explanatory variables can be restrictive and hence models such as GLMs (see Chapter 5) have since been developed, many studies still use the linear regression model as a core model.

For the linear regression model, classical estimation includes the LS and MLE methods. The MLE and the LS are optimal in the sense that they are the most efficient (and consistent) estimators (see also Section 2.3.2) but only under the relatively strong assumption that the distribution of the error term (see Section 3.2.1) is exactly normal. For the LS estimator, this optimality is also achieved under the hypothesis of i.i.d. residual error (with common mean of zero and common variance) but only within the class of linear estimators (in the observations).[1] The i.i.d. case encompasses a wide variety of distributions and the LS can be very inefficient outside the normality case, where better estimators (more efficient) exist but are not linear.

[1] Although the LS is also known as the BLUE, i.e. best linear unbiased estimator, one often forgets that 'best' is only for linear estimators. For a remainder of this feature, see e.g. Ronchetti (2006). Huber (1973) also discusses the limitations of the LS.

Robust Methods in Biostatistics S. Heritier, E. Cantoni, S. Copt and M.-P. Victoria-Feser

For example, suppose that the data are i.i.d. according to a Student distribution with three degrees of freedom, then a better estimator is given by the MLE under the Student distribution (with three degrees of freedom) which is a nonlinear estimator. The efficiency loss of the LS with respect to the MLE can be as large as 50%. This is of course just a particular example in which the non-normal data-generating process is known, but in general it is unknown and can best be said to be approximatively normal. As shown in Section 3.2.2, the effect on the classical LS or MLE estimators of even small model deviations from normality can be disastrous, leading to biased estimators and hence wrong interpretations of the postulated regression models. The biases induced by model deviations also have consequences for the calculation of the residuals (i.e. predicted response minus observed response), and hence on the analysis of the fit of the model.

Testing hypotheses is also an important aspect of the analysis of regression models. In particular, testing the significance of the regression parameters is routinely applied and is used for the interpretation of the model. Classically, a t-statistic is used and compared with a Student distribution for the computation of the p-values. When an explanatory variable is categorical with more than two categories, say K, it can be introduced into the model by means of $K-1$ dummy variables such as ANOVA models (see also Section 3.2.1). In such situations, an hypothesis of interest is the one of significance of the variable as a whole, i.e. the simultaneous significance of the regression parameters for all dummy variables corresponding to the categorical variable. The hypotheses of interest to be tested in this case is a multivariate hypothesis for which different testing procedures exist, such as the LRT-, Wald- or F-tests (see also Section 2.5.3). As is the case with estimation, testing procedures can be seriously affected by small model deviations, in that the actual level of the classical tests can be very far from the usual 5% level for which the tests are built (see Section 2.4). As a result, and as illustrated through the examples of this chapter, the conclusions drawn from classical tests of significance can be different when the model assumptions are not met and when a robust method is used in place of a classical method.

Finally, when the postulated regression model includes several explanatory variables, it is often observed in practice that at least some of them are more or less strongly correlated. They then explain the same part of the response variable, which makes their inclusion in the model altogether useless, or worse lead to the conclusion that these variables are not significant. Variable selection procedures are then necessary to objectively choose a suitable subset of regression variables. Classical procedures include those that are based on the likelihood function penalized for the number of parameters, e.g. the Akaike information criterion (AIC) (Akaike, 1973), and others based on prediction error criteria such as Mallows's C_p (Mallows, 1973). Both criteria are constructed on the normality of the errors hypothesis, which means that when they are computed from a sample in which not all of the observations have been generated by the postulated model, these criteria can lead to the choice of an inappropriate model. Model selection in regression is treated in Section 3.4.

This chapter is organized as follows. In Section 3.2 we present formally the regression model, study the robustness properties of the classical MLE and LS estimators and propose alternative robust estimators. In particular, we consider a Huber-type estimator as well as a high breakdown point estimator. Robust testing is developed in Section 3.3, while robust residual data analysis and model selection are treated in Section 3.4. All of the robust (and classical) methods are illustrated through the analysis of three different datasets.

3.2 Estimating the Regression Parameters

3.2.1 The Regression Model

One way to define the regression model is to postulate that the response variable y follows, conditionally on a set of regressors $\boldsymbol{x}$, a normal distribution of mean $\mu = \boldsymbol{x}^T\boldsymbol{\beta}$ and variance σ^2. The $(q+1)$-dimensional vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_q)$ contains the regression parameters or slopes with β_0 the intercept, and consequently $\boldsymbol{x} = (1, x_1, \ldots, x_q)^T$. For a sample of n observations, this amounts to postulating that $y_i|\boldsymbol{x}_i \sim \mathcal{N}(\mu_i, \sigma^2)$ with $\mu_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$ and $\boldsymbol{x}_i^T = (1, x_{i1}, \ldots, x_{iq})$. We will further assume that the variance is constant across observations and do not consider here the possible cases of heteroscedastic models (for possible robust estimators in these cases, see e.g. Bianco *et al.* (2000), Carroll and Ruppert (1982) and Giltinan *et al.* (1986)). The regression model can also alternatively be defined by means of

$$y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, \sigma^2) \quad \text{for all } i = 1, \ldots, n. \tag{3.1}$$

The multivariate form is written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n),$$

where $\boldsymbol{X} = [\boldsymbol{x}_i^T]_{i=1,\ldots,n}$, also called the design matrix, and $\boldsymbol{I}_n$ is the identity matrix of dimension n.

A particular regression model is one for which $\boldsymbol{x}$ is a set of dummy variables, i.e. taking the values of zero or one. This model can be used when a response variable is compared across categories of subjects. For example, one could study a physiological measure taken across K categories of patients. The response is the physiological measure and the explanatory variables are $x_1 = 1$ if the response corresponds to a patient in category one and zero otherwise, $x_2 = 1$ if the response corresponds to a patient in category two and zero otherwise, etc., and $x_{K-1} = 1$ if the response corresponds to a patient in category $K-1$ and zero otherwise. One could also cross two sets of categories, i.e. factors, such as the category of patients and their gender. This particular type of regression model actually corresponds to what is better known as the ANOVA models of Fisher. The values given to the dummy variables can in principle be changed, and this defines so-called contrasts. Hence, with a bit of reformulation, the robust techniques presented in this chapter can also be used for

ANOVA models[2] (for independent groups). ANOVA for repeated measures is treated in the mixed linear models of Chapter 4.

3.2.2 Robustness Properties of the LS and MLE Estimators

For a sample of n observations $(y_i, \boldsymbol{x}_i^T)$, $i = 1, \ldots, n$, and for the linear regression model $y_i|\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{x}_i^T\boldsymbol{\beta}, \sigma^2)$, the score function defining the MLE is given by

$$s(y_i, \boldsymbol{x}_i; \boldsymbol{\beta}, \sigma^2) = \frac{\partial}{\partial(\boldsymbol{\beta}^T, \sigma^2)^T} \log f(y_i, \boldsymbol{x}_i; \boldsymbol{\beta}, \sigma^2)$$
$$= \begin{bmatrix} \frac{1}{\sigma^2}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})\boldsymbol{x}_i \\ -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 \end{bmatrix}, \tag{3.2}$$

where f in (3.2) is the density of the normal distribution with mean $\boldsymbol{x}_i^T\boldsymbol{\beta}$ and variance σ^2. The LS estimator of $\boldsymbol{\beta}$ is equal to the MLE of $\boldsymbol{\beta}$ and the LS estimator of σ^2 is equal to the MLE of σ^2 up to a multiplicative constant of $n/(n-1)$. Setting $r_i = (y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})/\sigma$ as the standardized residuals, and simplifying the score function (i.e. multiplying it by σ), the LS estimator of $\boldsymbol{\beta}$ can be seen as an M-estimator (2.12) with

$$\Psi_{[LS]}(r_i, \boldsymbol{x}_i) = r_i\boldsymbol{x}_i. \tag{3.3}$$

Since the *IF* of an M-estimator is proportional to its Ψ-function, we can easily see that the LS and indeed the MLE are not robust as their *IF* is unbounded; in essence, arbitrary values in either the responses y_i (through r_i) or in the design matrix $\boldsymbol{X}$ can bias the estimators. By inspection of (3.2), one can see that this is also true for the MLE (or the LS) of the residual variance σ^2.

It could be tempting to remove observations to avoid possible bias with classical estimators. The criteria could be based on graphical analysis, residual analysis or more sophisticated diagnostic tools such as the Cook and Weisberg (1982) distance. However, as thoroughly argued in Section 1.3, this strategy, although apparently simple, is not only impractical, but also can be very misleading. It should, however, be stressed that more recently, procedures based on forward searches have been proposed for outlier detection in regression with no or limited masking effect (see Atkinson and Riani, 2000). Although this type of method is quite appealing, there remains an important problem that has not yet been solved: what about inference? Indeed, classical inference (e.g. t-test) is not valid when the data have been manipulated in some way. If observations have been removed from the sample on the basis of objective criteria, then inference should be conditional on these criteria, which is not the case with classical inference. The latter could be corrected (in a rather complicated fashion) using the results of Welsh and Ronchetti (2002). On the other hand, a robust approach based on robust estimators, also provides robust

[2]For practical examples of specific robust testing methods in the ANOVA setting, see e.g. Wilcox (1997).

testing procedures that take into account the fact that some observations have either been downweighted or, more drastically, removed from the sample (i.e. weighted with weight equal to zero).

Hence, we propose here a 'global' approach for estimation and inference based on robust bounded influence or bounded Ψ-function estimators and the corresponding inferential tools. Before presenting such estimators and testing procedures, we give an example to illustrate both the effect of model deviation on classical estimators and the properties of robust estimators.

3.2.3 Glomerular Filtration Rate (GFR) Data Example

The dataset considered here contains measurements or estimation of the glomerular filtration rate (`gfr`) and serum creatinine (`cr`). The `gfr` is the volume of fluid filtered from the renal glomerular capillaries into the Bowman's capsule per unit of time (typically in milliliters per minute) and clinically it is often used to determine renal function. Its estimation, when not measured, is of clinical importance and several techniques are used for that purpose. One of them is based on `cr`, an endogenous molecule, synthesized in the body, which is freely filtered by the glomerulus (but also secreted by the renal tubules in very small amounts). `cr` is therefore supposed to be a good predictor of `gfr` and, based on empirical evidence, their relationship is nonlinear. Several models have been proposed in the literature to explain the logarithm of `gfr` as a function of `cr` and possibly other explanatory variables such as age and sex; some of them are (see, e.g. Rule *et al.*, 2004)

$$\log(\texttt{gfr}) = \beta_0 + \beta_1 \log(\texttt{cr}) + \beta_2 \log(\texttt{age}) + \beta_3 \texttt{sex} + \epsilon \tag{3.4}$$

and

$$\log(\texttt{gfr}) = \beta_0 + \beta_{11}\texttt{cr}^{-1} + \beta_{12}\texttt{cr}^{-2} + \beta_2 \texttt{age} + \beta_3 \texttt{sex} + \epsilon. \tag{3.5}$$

The data we have at hand is the `gfr`, `cr` and `age` measured on a random sample of 30 men out of the 180 patients included in the Brochner-Mortensen *et al.* (1977) study of renal function, and analyzed by Ingelfinger *et al.* (1987, Table 9b-2, p. 229). As all subjects are males, we consider models (3.4) and (3.5) without the variable `sex`. In this sample, the median age of the participants is 50, their median serum creatinine is 1.395 and their median glomerular filtration rate 68 milliliters per minute.

Suppose for simplicity that as a first step we are interested in the linear relationship between $\log(\texttt{gfr})$ and $\texttt{crinv} = \texttt{cr}^{-1}$ (i.e. model (3.5) without the quadratic term and the variables `age` and `sex`). The data are plotted in Figure 3.1 together with three regression lines, one estimated by means of the LS estimator, another with the LS estimator without two observations and the other by means of a robust estimator (see Section 3.2.4). Note that the line of the LS estimator without 2 observations is nearly undistinguishable from that of the robust one. One can spot two observations at the far right of the graph which are actually extremes with respect to the linear regression model. A quadratic term could be added to the model (as in (3.5)) to fit these two observations, but we leave this option for later and use this

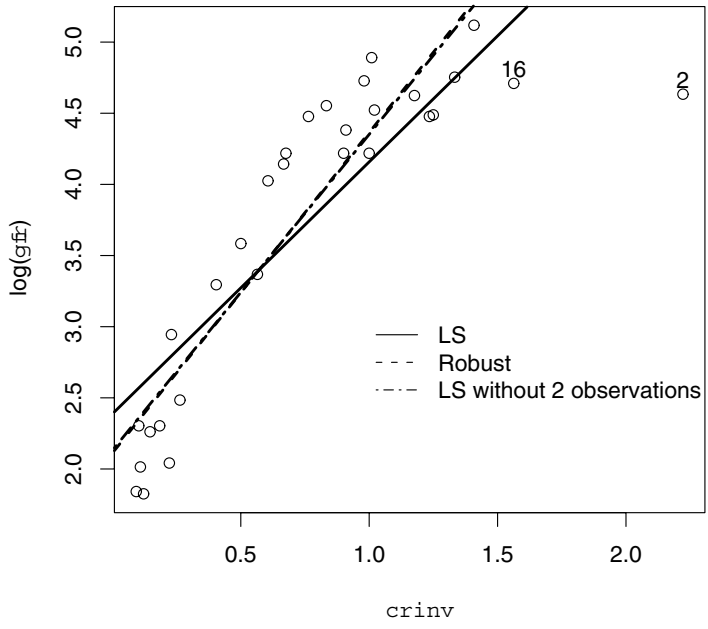

Figure 3.1 Estimated regression lines for the model $\log(\texttt{gfr}) = \beta_0 + \beta_1 \texttt{cr}^{-1} + \epsilon$.

example to illustrate the difference between a classical and a robust analysis. The two observations 2 and 16 have an important effect on the LS estimator: the fit is not satisfactory. The robust estimator on the other hand is able to capture the linear relationship between the two variables as illustrated by the majority of the data. We also estimated the regression parameter by means of the LS on the sample without the two extreme observations (2 and 16) and plotted the corresponding line. Without these two observations, the LS estimator provides an estimate similar to the robust estimator. The estimated regression parameter for $\texttt{cr}^{-1}$ is 1.76 with the LS estimator, 2.17 with the robust estimator and 2.36 with the LS on the reduced sample. The scale estimates are 0.57 with the LS estimator, 0.53 with the robust estimator and 0.39 with the LS on the reduced sample. The estimated impact of $\texttt{cr}^{-1}$ on $\log(\texttt{gfr})$ is hence not the same, although it is very similar between the robust estimator and the LS estimator on the reduced sample. Note, however, that the estimated standard errors of the robust estimator and of the LS on the reduced sample differ noticeably.

3.2.4 Robust Estimators

3.2.4.1 Huber's Estimator

Robust estimators for the regression coefficients have been proposed regularly in the statistical literature since the proposal of Huber (1973). He considered a WMLE (see (2.15)) estimator with Huber's weights (2.16) applied to (3.3). The resulting

estimator is hence an M-estimator with Ψ-function

$$\Psi_{[Hub]}(r_i, \boldsymbol{x}_i; \boldsymbol{\beta}, \sigma^2, c) = \psi_{[Hub]}(r_i; \boldsymbol{\beta}, \sigma^2, c)\boldsymbol{x}_i = w_{[Hub]}(r_i; \boldsymbol{\beta}, \sigma^2, c) r_i \boldsymbol{x}_i, \quad (3.6)$$

and corresponding ρ-function

$$\rho_{[Hub]}(r_i; \boldsymbol{\beta}, \sigma^2, c) = \begin{cases} \frac{1}{2} r_i^2 & \text{for } |r_i| \leq c, \\ c|r_i| - \frac{1}{2}c^2 & \text{for } |r_i| > c, \end{cases}$$

for the regression parameter $\boldsymbol{\beta}$. Since the corresponding ρ function is convex the solution in $\boldsymbol{\beta}$ of $\sum_{i=1}^n \Psi_{[Hub]}(r_i, \boldsymbol{x}_i; \boldsymbol{\beta}, \sigma^2, c) = 0$ is unique and is obtained using iteratively reweighted least squares (IRWLS). At iteration t, one obtains $\hat{\boldsymbol{\beta}}^t$ which is used to compute the residuals r_i^t and weights $w_{[Hub]}(r_i^t; \boldsymbol{\beta}, \sigma^2, c)$ and then $\sum_{i=1}^n \Psi_{[Hub]}(r_i, \boldsymbol{x}_i; \boldsymbol{\beta}, \sigma^2, c) = 0$ is solved for $\boldsymbol{\beta}$ (in r_i) to obtain an updated estimate $\hat{\boldsymbol{\beta}}^{t+1}$.

The scale parameter σ (in the r_i) also needs to be estimated and one can take, for example, Huber's Proposal 2 (Huber, 1981, p. 137), also a weighted estimator, defined through

$$\frac{1}{n}\sum_{i=1}^n w_{[Hub]}^2(r_i; \boldsymbol{\beta}, \sigma^2, c) r_i^2 - \int w_{[Hub]}^2(r; \boldsymbol{\beta}, \sigma^2, c) r^2 \varphi(r)\, dr = 0, \quad (3.7)$$

with φ the density of the standard normal distribution and

$$\begin{aligned} \int r^2 w_{[Hub]}^2(r; \boldsymbol{\beta}, \sigma^2, c)\varphi(r)\, dr &= E_\Phi[r^2 w_{[Hub]}^2(r; \boldsymbol{\beta}, \sigma^2, c)] \\ &= 2\Phi(c) - 1 - 2c\varphi(c) + 2c^2(1 - \Phi(c)), \quad (3.8) \end{aligned}$$

that ensures the consistency of the resulting estimator at the standardized normal model Φ (the hypothetical model for the standardized residuals). If the bounding constant c of $w_{[Hub]}$ in (3.6) and (3.7) is large enough, then all weights are equal to or tend to one and the integral (3.8) is equal to one and the estimators of $\boldsymbol{\beta}$ and σ^2 are the MLE. This constant c hence controls the degree of robustness of the procedure. It can be chosen on efficiency arguments, i.e. so that the ratio between the traces of the covariance matrices of the LS (or MLE) and the robust estimators achieves a given value, typically 90–95% (see (2.31) and also (3.20)). The simultaneous estimation of $\boldsymbol{\beta}$ and σ^2 can be obtained by an IRWLS in which both estimators are updated at each iteration (given the weights $w_{[Hub]}(r_i; \boldsymbol{\beta}, \sigma^2, c)$).

3.2.4.2 Robust Weighted Estimators in the Design Space

Huber's regression coefficient estimator cannot protect against bad leverage points, or in other terms, it is not robust in the design space. Indeed, the weights in (3.6) control for extreme residuals, but not for extreme values in the design matrix. This deficiency has led to several other proposals for robust regression coefficients estimators in both the response and the design matrix. These estimators can be

written as M-estimators (also called generalized M-estimators or GM-estimators) with

$$\Psi(r_i, \boldsymbol{x}_i; \boldsymbol{\beta}, \sigma^2, c) = v_1(\boldsymbol{x}_i)\Psi_{[Hub]}(r_i \cdot v_2(\boldsymbol{x}_i), \boldsymbol{x}_i; c). \tag{3.9}$$

There exist, in fact, many different robust estimators for the regression coefficients. In the class given in (3.9) many variations can be imagined for v_1 and v_2. One can cite e.g. Mallows's class (Mallows, 1975) and Hampel–Krasker–Welsch class (Krasker and Welsch, 1982). Others can be found in Hampel *et al.* (1986, pp. 315–316).

The Mallows's class has become more popular, especially because of its simplicity. It is given by

$$\Psi(r_i, \boldsymbol{x}_i; \boldsymbol{\beta}, \sigma^2, c) = w(\boldsymbol{x}_i)\Psi_{[Hub]}(r_i, \boldsymbol{x}_i; \boldsymbol{\beta}, \sigma^2, c). \tag{3.10}$$

One has to choose the weight function $0 \leq w(\boldsymbol{x}_i) \leq 1$ on the design space. This is not an obvious task since the choice will depend on the type of explanatory variables. Indeed, $w(\boldsymbol{x}_i)$ should downweight points in the design space that are in some sense 'large'. This 'largeness' can be measured by means of distances of each $\boldsymbol{x}_i$ with respect to a center and possibly a scatter matrix. If one has only continuous covariates, one can use a robust Mahalanobis distance $d_i = \sqrt{(\boldsymbol{x}_{i(2)} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_{i(2)} - \boldsymbol{\mu})}$ based on a robust estimator for the center $\boldsymbol{\mu}$ and scatter $\boldsymbol{\Sigma}$ (see Section 2.3.3). Note that the distances are taken on $\boldsymbol{x}_{i(2)} = (x_{i1}, \ldots, x_{iq})$, i.e. without the intercept part. The weights can be defined e.g. as

$$w(\boldsymbol{x}_i) = \begin{cases} 1 & \text{if } d_i^2 \leq (\chi_q^2)^{-1}(0.975), \\ 0 & \text{otherwise.} \end{cases}$$

In other words, observations in the design space that have a squared Mahalanobis distance that is larger than the 0.975 quantile of the χ_q^2 are given a weight of zero. Note that q is the number of explanatory variables, and $d_i^2 \sim \chi_q^2$ if the $\boldsymbol{x}_{i(2)}$ are multivariate normal.

When the (approximate) normality of the $\boldsymbol{x}_{i(2)}$ cannot be assumed, for example when some or all explanatory variables are categorical, one can rely on the so-called 'hat matrix'

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \tag{3.11}$$

and its diagonal elements or leverages $h_{ii} = \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i \in (0; 1)$. The latter have actually been extensively used in regression diagnostics; for general references, see e.g. Belsley *et al.* (1980), Cook and Weisberg (1982), Atkinson (1985), Chatterjee and Hadi (1988). A simple weighting scheme based on h_{ii} is given by

$$w(\boldsymbol{x}_i) = \sqrt{1 - h_{ii}}. \tag{3.12}$$

Indeed, since for the LS estimator $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$ and $\boldsymbol{r} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\epsilon}$, then in particular $r_i = (1 - h_{ii})\epsilon_i - \sum_{j \neq i} h_{ij}\epsilon_j$. This means that when h_{ii} is large (near one), an unexpected error in the response (i.e. large ϵ_i) might not be reflected in the residuals r_i. Therefore, the weights (3.12) compensate for the extreme responses

not captured by the residuals, i.e. the extreme observations in the design matrix (see also Staudte and Sheather, 1990, p. 209). These extreme observations are also called leverage points.

For the scale parameter σ^2, one can choose Huber's Proposal 2 in (3.7) or a modified version of it that includes weights $w(\boldsymbol{x}_i)$ as in (3.10). As for Huber's estimator, a *GM*-estimator can be found using an IRWLS in which at each iteration the weights $w_{[Hub]}(r_i; \boldsymbol{\beta}, \sigma^2, c)$ and $w(\boldsymbol{x}_i)$ are updated.

Although *GM*-estimators are relatively simple to compute and also rather intuitive, they are limited because of the arbitrary choice for the weights $w(\boldsymbol{x}_i)$. They are, however, adapted for GLM and marginal longitudinal data models in Chapters 5 and 6, for which computer intensive estimators (see Section 3.2.4.3) are not yet available.

3.2.4.3 High Breakdown Estimators

A high breakdown point is not achieved when Huber's weights are used (see Maronna *et al.*, 1979) which is especially a concern when a large number of predictors are included in the model. One way to overcome this is to use instead of $\psi_{[Hub]}$ a redescending ψ-function (see Section 2.3), i.e. a ψ-function that can become nil when the residuals are too large. One can use, for example, Tukey's biweight function (2.22) which, for the regression model, corresponds to the following Ψ-function

$$\Psi_{[bi]}(r_i, \boldsymbol{x}_i; \boldsymbol{\beta}, \sigma^2, c) = \psi_{[bi]}(r_i; c)\boldsymbol{x}_i = w_{[bi]}(r_i; c) r_i \boldsymbol{x}_i \tag{3.13}$$

with weights

$$w_{[bi]}(r_i; c) = \begin{cases} \left(\left(\dfrac{r_i}{c}\right)^2 - 1\right)^2 & \text{if } |r_i| \leq c, \\ 0 & \text{if } |r_i| > c, \end{cases} \tag{3.14}$$

hence, defining a WMLE. When r_i tends to the value of c, the tuning constant, the $\Psi_{[bi]}$-function tends to the value of zero. Note that the ρ-function corresponding to the weights (3.14), i.e. such that $w(r; \boldsymbol{\beta}, \sigma^2, c) = (1/r)\partial\rho(r; \boldsymbol{\beta}, \sigma^2, c)/\partial r$ is

$$\rho_{[bi]}(r_i; c) = \begin{cases} \left(\dfrac{c^2}{6}\right)\left(3\left(\dfrac{r_i}{c}\right)^2 - 3\left(\dfrac{r_i}{c}\right)^4 + \left(\dfrac{r_i}{c}\right)^6\right) & \text{if } |r_i| \leq c, \\ 1 & \text{if } |r_i| > c. \end{cases} \tag{3.15}$$

A problem arises which concerns the computation of such estimators. Indeed, even if one assumes that the scale parameter σ^2 is known, $\sum_{i=1}^{n} \Psi_{[bi]}(r_i, \boldsymbol{x}_i; \boldsymbol{\beta}, \sigma^2, c) = 0$ admits more than one solution for $\boldsymbol{\beta}$. In an iterative procedure, the starting point is hence crucial. By choosing a high breakdown point estimator (even with low efficiency) as the starting point one can define a high breakdown point estimator that achieves a chosen level of efficiency. To be more specific, one first computes a starting consistent estimator with high breakdown point $\hat{\boldsymbol{\beta}}^0$ (even with low efficiency) together with a robust residual scale estimator $\hat{\sigma}^2$ and uses $\hat{\boldsymbol{\beta}}^0$ as the

starting point in an iterative procedure (IRWLS) to find the solution $\hat{\boldsymbol{\beta}}_{[bi]}$ in $\boldsymbol{\beta}$ of $\sum_{i=1}^{n} \Psi_{[bi]}(r_i, \boldsymbol{x}_i; \boldsymbol{\beta}, \sigma^2, c) = 0$ with σ^2 replaced by $\hat{\sigma}^2$. The resulting estimator for $\boldsymbol{\beta}$ is called the *MM*-estimator by Yohai (1987); it solves an *M*-type equation with an *M*-type estimator as starting point. The estimator $\hat{\boldsymbol{\beta}}_{[bi]}$ has the same breakdown point than $\hat{\boldsymbol{\beta}}^0$ but an efficiency that can be chosen with a suitable value for c in (3.13) (see also Yohai *et al.*, 1991). It should also be noted that because of the redescending nature of the ψ-function, there is no need for a weighting scheme on the design space as is necessary with *GM*-estimators. For a discussion about the starting point and computational aspects, see Appendix A.

Instead of the biweight function defining $\hat{\boldsymbol{\beta}}_{[bi]}$, one can alternatively choose other ψ-functions (or corresponding ρ-functions; see e.g. Hampel *et al.* (1986, Section 2.6)). Yohai and Zamar (1998) proposed what they called the 'optimal' ρ-function which has an advantage (over the other functions) of minimizing the GES (see (2.4)) for the same breakdown point and efficiency. In practice, however, the differences are marginal and hence we propose to use the biweight function. We also fix the efficiency level at 90% (with corresponding $c = 3.8827$). This is the robust estimator used to estimate the regression line in Figure 3.1 with the GFR data example.

3.2.5 GFR Data Example (continued)

Let us come back to the GFR dataset and estimate the regression parameters using the classical LS and the robust estimator, with all explanatory variables (i.e. also with the quadratic term and the variable `age`) and considering both models (3.4) and (3.5). The different estimated values (together with corresponding p-values for significance testing, see Section 3.3.1) are presented in Table 3.1. The variable `age` is clearly not significant in either model or either method, whereas the variable `cr` plays a significant role in explaining the level of (log) `gfr` in both models and with both methods. The only difference between the LS and the robust estimation lies in the different estimated values for the parameters. For example, with model (3.5), the regression coefficient for $\texttt{cr}^{-1}$ is estimated to 4.27 with the LS and to 5.06 with the robust estimators. In terms of prediction (we exclude the variable `age`), given that

$$\begin{aligned} \log(\texttt{gfr}) = \beta_0 + \beta_{11}\texttt{cr}^{-1} + \beta_{12}\texttt{cr}^{-2} + \epsilon \Leftrightarrow \\ \texttt{gfr} = \exp\{\beta_0 + \beta_{11}\texttt{cr}^{-1} + \beta_{12}\texttt{cr}^{-2}\}\exp\{\epsilon\} \Leftrightarrow \\ E[\texttt{gfr}] = \exp\{\beta_0 + \beta_{11}\texttt{cr}^{-1} + \beta_{12}\texttt{cr}^{-2}\}E[\exp\{\epsilon\}] \end{aligned}$$

and using the moment-generating function of a $\mathcal{N}(\mu, \sigma^2)$ variable x, i.e. $E[\exp(tx)] = \exp(\mu t + \sigma^2 t^2/2)$ at $t = 1$ (with $\mu = 0$), we obtain

$$\widehat{\texttt{gfr}} = \exp\left\{\widehat{\beta}_0 + \frac{\widehat{\beta}_{11}}{\texttt{cr}} + \frac{\widehat{\beta}_{12}}{\texttt{cr}^2}\right\}\exp\left(\frac{\hat{\sigma}^2}{2}\right).$$

Differences in estimated values for the regression coefficients can lead to important practical differences in predictions, as illustrated in Figure 3.2. In this example, if

Table 3.1 Estimated regression parameters and significance tests for two models for the GFR data.

	Model (3.4)			
	LS		Robust	
	Estimate (*SE*)	*p*-value	Estimate (*SE*)	*p*-value
`intercept`	4.57 (0.72)	$<10^{-4}$	4.92 (0.75)	$<10^{-4}$
log(`cr`)	−1.12 (0.07)	$<10^{-4}$	−1.12 (0.06)	$<10^{-4}$
log(`age`)	−0.04 (0.19)	0.823	−0.13 (0.19)	0.503
$\widehat{\sigma}$	0.31		0.32	
R^2	0.924		0.940	
	Model (3.5)			
	LS		Robust	
	Estimate (*SE*)	*p*-value	Estimate (*SE*)	*p*-value
`intercept`	1.80 (0.25)	$<10^{-4}$	1.56 (0.33)	$<10^{-4}$
$\texttt{cr}^{-1}$	4.27 (0.28)	$<10^{-4}$	5.06 (0.44)	$<10^{-4}$
$\texttt{cr}^{-2}$	−1.38 (0.14)	$<10^{-4}$	−1.96 (0.27)	$<10^{-4}$
`age`	−0.003 (0.004)	0.45	−0.001 (0.004)	0.755
$\widehat{\sigma}$	0.27		0.30	
R^2	0.943		0.956	

The estimates are the LS and the biweight *MM*-estimator with $c = 3.8827$ (90% efficiency).

the model holds, the predicted `gfr` is quite different for values of `cr` up to one (and even over one). Given that the sample median value for `cr` is 1.395, this amounts to saying that the predicted `gfr` is quite different for half of the participants, and hence heavily dependent on the chosen estimation method. As we will see in Section 3.4, the difference is due to only one observation, namely observation 2.

3.3 Testing the Regression Parameters

3.3.1 Significance Testing

One aspect of inference in regression models is testing the significance of the regression parameters, i.e. $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$, $j = 1, \ldots, q$. When the LS or MLE estimator is chosen to estimate the β_j, then one uses the t-statistic

$$t\text{-statistic} = \frac{\widehat{\beta}_{[LS]j}}{SE(\widehat{\beta}_{[LS]j})}, \tag{3.16}$$

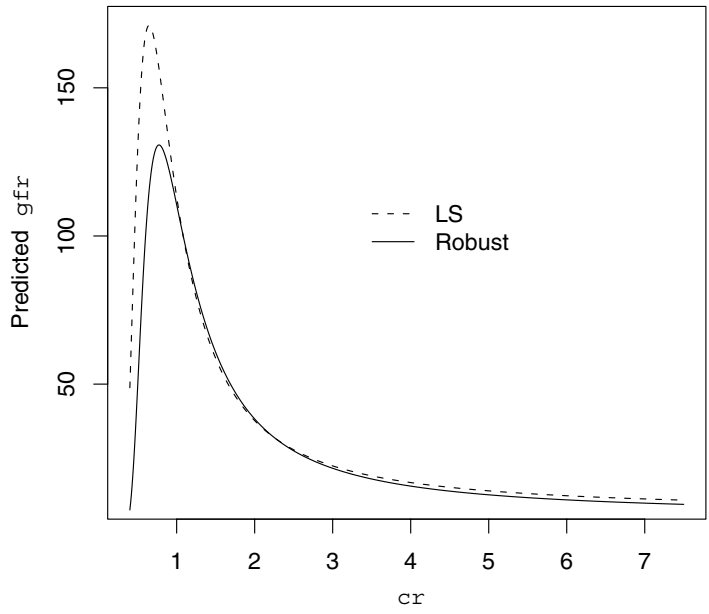

Figure 3.2 Predicted `gfr` value using the estimated regression model (3.5) without the variable `age`.

where $\widehat{\beta}_{[LS]j}$ is the LS estimator (or MLE) of β_j,

$$SE(\widehat{\beta}_{[LS]j}) = \sqrt{\widehat{\sigma}^2[(\boldsymbol{X}^T\boldsymbol{X})^{-1}]_{(j+1)(j+1)}},$$

$$\widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}_{[LS]})^2$$

is the classical residual scale estimate and $[\boldsymbol{A}]_{(j+1)(j+1)}$ denotes the element at the $(j+1)$th line and the $(j+1)$th column of the matrix $\boldsymbol{A}$. In fact we have that

$$\begin{aligned}
\mathrm{var}(\widehat{\boldsymbol{\beta}}_{[LS]})^{-1} &= n\int \frac{1}{\sigma}\Psi_{[LS]}(r,\boldsymbol{x};\boldsymbol{\beta},\sigma^2)\frac{1}{\sigma}\Psi_{[LS]}(r,\boldsymbol{x}_i;\boldsymbol{\beta},\sigma^2)^T\, d\Phi(r)\, dF(\boldsymbol{x}) \\
&= n\frac{1}{n}\frac{1}{\sigma^2}\sum_{i=1}^{n}\int \Psi_{[LS]}(r,\boldsymbol{x}_i;\boldsymbol{\beta},\sigma^2)\Psi_{[LS]}(r,\boldsymbol{x}_i;\boldsymbol{\beta},\sigma^2)^T\, d\Phi(r) \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T E_\Phi[r^2] = \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X},
\end{aligned}$$

with $\Psi_{[LS]}(r, \boldsymbol{x}; \boldsymbol{\beta}, \sigma^2)$ given in (3.3).[3] For the robust estimator, for example $\widehat{\boldsymbol{\beta}}_{[bi]}$, we proceed in the same way, i.e. using (2.27) in Chapter 2,

$$\operatorname{var}(\widehat{\boldsymbol{\beta}}_{[bi]}) = \frac{1}{n} M(\Psi_{[bi]}, \Phi)^{-1} Q(\Psi_{[bi]}, \Phi) M(\Psi_{[bi]}, \Phi)^{-T} \tag{3.17}$$

with

$$\begin{aligned} Q(\Psi_{[bi]}, \Phi) &= \frac{1}{n} \sum_{i=1}^{n} \int \Psi_{[bi]}(r, \boldsymbol{x}_i; \boldsymbol{\beta}, \sigma^2, c) \Psi_{[bi]}(r, \boldsymbol{x}_i; \boldsymbol{\beta}, \sigma^2, c)^T \, d\Phi(r) \\ &= \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T \int_{-c}^{c} r^2 \left(\left(\frac{r}{c} \right)^2 - 1 \right)^4 d\Phi(r) \end{aligned} \tag{3.18}$$

and

$$\begin{aligned} M(\Psi_{[bi]}, \Phi) &= -\frac{1}{n} \sum_{i=1}^{n} \int \frac{\partial}{\partial \boldsymbol{\beta}^T} \Psi_{[bi]}(r, \boldsymbol{x}_i; \boldsymbol{\beta}, \sigma^2, c) \, d\Phi(r) \\ &= \frac{1}{\sigma} \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T \int_{-c}^{c} \left(5 \left(\frac{r}{c} \right)^4 - 6 \left(\frac{r}{c} \right)^2 + 1 \right) d\Phi(r) \end{aligned} \tag{3.19}$$

so that

$$\begin{aligned} \operatorname{var}(\widehat{\boldsymbol{\beta}}_{[bi]}) &= \sigma^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1} \int_{-c}^{c} r^2 \left(\left(\frac{r}{c} \right)^2 - 1 \right)^4 d\Phi(r) \Bigg/ \\ &\qquad \left[\int_{-c}^{c} \left(5 \left(\frac{r}{c} \right)^4 - 6 \left(\frac{r}{c} \right)^2 + 1 \right) d\Phi(r) \right]^2 \\ &= \sigma^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1} e_c^{-1}. \end{aligned}$$

The variance of $\widehat{\boldsymbol{\beta}}_{[bi]}$ is larger than the variance of $\widehat{\boldsymbol{\beta}}_{[LS]}$ by a factor of e_c^{-1}. In other words, the efficiency of $\widehat{\boldsymbol{\beta}}_{[bi]}$ is[4]

$$e_c = \left[\int_{-c}^{c} \left(5 \left(\frac{r}{c} \right)^4 - 6 \left(\frac{r}{c} \right)^2 + 1 \right) d\Phi(r) \right]^2 \Bigg/ \int_{-c}^{c} r^2 \left(\left(\frac{r}{c} \right)^2 - 1 \right)^4 d\Phi(r). \tag{3.20}$$

The efficiency given in (3.20) is actually used to choose a value for the tuning constant c to achieve a given level of efficiency.

To estimate the variance (3.17) it is not wise to just replace σ by a robust $\hat{\sigma}$. Indeed, extreme values in the design matrix $\boldsymbol{X}$ are not automatically downweighted. Alternatively, one can use the asymptotic variance (3.17) and replace the matrices M and Q by their empirical counterparts, i.e. by removing in (3.18) and (3.19) the integrals and putting r_i instead of r;[5] see also Simpson *et al.* (1992). Another

[3] Note that to obtain the derivative of the log-likelihood function, one has to multiply $\Psi_{[LS]}$ by $1/\sigma$.

[4] An analytical expression is given in Appendix B.

[5] This estimator might however produce negative values for the variance because of (3.19).

estimator for (3.17) is given by

$$\widehat{\text{var}}(\widehat{\boldsymbol{\beta}}_{[bi]}) = \widehat{\sigma}^2 \left(\frac{1}{\sum_{i=1}^{n} w_{[bi]}(r_i; \boldsymbol{\beta}, \sigma^2, c)} \sum_{i=1}^{n} w_{[bi]}(r_i; \boldsymbol{\beta}, \sigma^2, c) \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1} e_c^{-1} \tag{3.21}$$

with $w_{[bi]}(r_i; \boldsymbol{\beta}, \sigma^2, c)$ the biweight weights given in (3.14) and $\widehat{\sigma}^2$ the corresponding robust residual variance estimator (see e.g. Maronna *et al.*, 2006, p. 140). The advantage of $\widehat{\text{var}}(\widehat{\boldsymbol{\beta}}_{[bi]})$ in (3.21) is that the resulting estimated standard errors for $\widehat{\boldsymbol{\beta}}_{[bi]}$ are also robust to extreme observations in the regressors. Finally, Croux *et al.* (2003) also proposed alternative estimators of $\text{var}(\widehat{\boldsymbol{\beta}}_{[bi]})$, one of which ($A\ \text{var}_1$) is currently implemented in the R package `robustbase`. Let

$$SE(\widehat{\beta}_{[bi]j}) = \sqrt{[\widehat{\text{var}}(\widehat{\boldsymbol{\beta}}_{[bi]})]_{(j+1)(j+1)}}$$

with $\widehat{\text{var}}(\widehat{\boldsymbol{\beta}}_{[bi]})$ the chosen estimate for (3.17), the test statistic for testing significance of regression parameters is then the ratio

$$z\text{-statistic} = \frac{\widehat{\beta}_{[bi]j}}{SE(\widehat{\beta}_{[bi]j})}. \tag{3.22}$$

While (3.16) can be compared with a Student distribution with $n - q - 1$ degrees of freedom under $H_0 : \beta_j = 0$, for (3.22) one can only rely on the asymptotic normality of M-estimators (see Section 2.3.1). Hence, the p-value for the robust significance test is obtained by comparing (3.22) with the standard normal distribution.

The (small) loss of efficiency of the robust estimator at the exact regression model induces a (small) reduction of the power of the test. In other words, when all of the data have been *exactly* generated by the postulated regression model, then one needs more data to detect significant parameters when a robust estimator is used. This is the price to pay in order to have a testing procedure that works when the model is not exact. Indeed, small model deviations, such as gross errors, not only bias the classical regression estimates but also the classical scale estimate, both used to construct the test statistic, and consequently the p-value that is computed and upon which a decision about the significance of the corresponding parameter is taken (see Section 2.4.2). However, in practice, the robustness issue largely overcomes the power issue, in that a small model deviation has a more dramatic effect on classical testing than the loss of efficiency of the robust estimator at the exact (normal) regression model.

For the GFR data example presented in Table 3.1, with both models, although the p-values are different between the classical and the robust approach, they all lead to the same conclusion: all regression coefficients are significant except that of the variable `age`.

3.3.2 Diabetes Data Example

The dataset illustrates the relationship between diabetes and obesity measured through the body mass index and the waist/hip ratio and controlled for the body frame

of the participants. The data come from the R package `Hmisc` (dataset diabetes) and consist of 19 variables on 403 subjects (see Harrell (2001, p. 379) for explanations on the origin of the dataset). The response variable is glycosolated hemoglobin (`gh`) which is usually taken as a positive diagnosis of diabetes when it exceeds the value of 7. We use as potential explanatory variables the `age`, the gender (`sex`), a dummy variable with one for male and zero for female, the body mass index weight/height^2 (`bmi`), the waist/hip ratio (`whip`), the body frame with three levels (small, medium and large), hence modeled using two dummy variables `bfmed` with one for medium frame and zero otherwise and `bflar` with one for large frame and zero otherwise, the stabilized glucose (`stabg`), as well as the location of the subject (`loc`), a dummy variable with zero for Buckingham County and one for Louisa County (two rural Virginia counties). We consider the following model

$$\begin{aligned}\texttt{gh}_i = \beta_0 &+ \beta_1 \texttt{age}_i + \beta_2 \texttt{sex}_i + \beta_3 \texttt{bmi}_i + \beta_4 \texttt{whip}_i \\ &+ \beta_5 \texttt{bfmed}_i + \beta_6 \texttt{bflar}_i + \beta_7 \texttt{stabg}_i + \beta_8 \texttt{loc}_i + \epsilon_i \end{aligned} \tag{3.23}$$

for which a variable selection method will be used later to select a suitable subset of explanatory variables. We use a missing-values-free subsample of size 372 (out of the original 403 observations).

In this sample, the median age of the participants is 45, there are 156 male (hence 216 female) participants, with median `bmi` of 27.6, median waist/hip ratio of 0.88 and, median stabilized glucose of 90. Moreover, 100 of the respondents have a small frame while 176 a medium frame (hence, 96 a large frame), and 180 live in Buckingham County (hence 192 in Luisa County).

The classical LS and robust estimates and corresponding p-values are provided in Table 3.2. While with the classical estimation, only the variables `age` and `stabg` are significant, with the robust estimation the variable location is also significant at the 5% level. The two approaches give similar coefficients (-0.21 for the LS estimate and -0.22 for the robust one), but the residual scale estimate is quite different (1.47 versus 0.763). Hence, a small model deviation such as one or a few outliers for this variable have an effect on the significance test but not on the estimate here.

It should, however, be stressed that the complete model is not necessarily the best model for the dataset at hand, and before concluding on the relationship between the response variable and the explanatory variables, one should first proceed with model checking and model selection (see Section 3.4).

3.3.3 Multiple Hypothesis Testing

With regression models, one could also in principle be interested in multiple hypothesis testing. The classical F-test for $H_0 : \beta_j = \cdots = \beta_{j'} = 0$, for some $j, j' > 0$ is one example. More generally, a LRT can be used to compare two nested models. Formally, let $\boldsymbol{\beta} = (\boldsymbol{\beta}_{(1)}, \boldsymbol{\beta}_{(2)})$ with $\dim(\boldsymbol{\beta}_{(2)}) = k$; suppose that we want to test $H_0 : \boldsymbol{\beta}_{(2)} = \boldsymbol{\beta}^0_{(2)}$ (reduced model) against $H_1 : \boldsymbol{\beta}_{(2)} \neq \boldsymbol{\beta}^0_{(2)}$ (full model) with $\boldsymbol{\beta}_{(1)}$ unspecified, and let also $\widehat{\boldsymbol{\beta}}_{[LS]}$, be the LS estimators of $\boldsymbol{\beta}$ (in the full model) and $\dot{\boldsymbol{\beta}}_{[LS]} = (\dot{\boldsymbol{\beta}}_{[LS](1)}, \boldsymbol{\beta}^0_{(2)})$ with $\dot{\boldsymbol{\beta}}_{[LS](1)}$ the LS estimator of $\boldsymbol{\beta}_{(1)}$ in the reduced model.

Table 3.2 Estimates and significance tests in model (3.23) for the diabetes data.

	LS		Robust	
	Estimate (*SE*)	p-value	Estimate (*SE*)	p-value
`intercept`	0.31 (1.0)	0.76	1.72 (0.52)	9.6×10^{-4}
`age`	0.018 (0.005)	4.8×10^{-4}	0.011 (0.003)	8.8×10^{-4}
`sex`	−0.11 (0.18)	0.54	−0.037 (0.12)	0.76
`bmi`	0.01 (0.01)	0.44	0.01 (0.01)	0.25
`whip`	1.2 (1.2)	0.30	0.59 (0.66)	0.38
`bfmed`	0.2 (0.2)	0.31	0.06 (0.11)	0.60
`bflar`	−0.08 (0.26)	0.75	0.19 (0.16)	0.24
`stabg`	0.029 (0.0015)	$<10^{-4}$	0.021 (0.005)	1.2×10^{-4}
`loc`	−0.21 (016)	0.18	−0.22 (0.10)	0.035
$\widehat{\sigma}$	1.47		0.763	
R^2	0.575		0.651	

The estimates are the LS and the biweight *MM*-estimator with $c = 3.8827$ (90% efficiency).

Using (2.46) with the regression model, the LRT statistic is given by

$$\begin{aligned}
\text{LRT} &= 2\sum_{i=1}^{n}(\log f(y_i, \boldsymbol{x}_i; \widehat{\boldsymbol{\beta}}_{[LS]}, \widehat{\sigma}^2_{[LS]}) - \log f(y_i, \boldsymbol{x}_i; \dot{\boldsymbol{\beta}}_{[LS]}, \widehat{\sigma}^2_{[LS]})) \\
&= \sum_{i=1}^{n}\left(\left(\frac{y_i - \boldsymbol{x}_i^T \dot{\boldsymbol{\beta}}_{[LS]}}{\widehat{\sigma}_{[LS]}}\right)^2 - \left(\frac{y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}_{[LS]}}{\widehat{\sigma}_{[LS]}}\right)^2\right) \\
&= \frac{\sum_{i=1}^{n}((y_i - \boldsymbol{x}_i^T \dot{\boldsymbol{\beta}}_{[LS]})^2 - (y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}_{[LS]})^2)}{\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}_{[LS]})^2}(n - q - 1) \\
&= \frac{\text{RSS}^0 - \text{RSS}}{\text{RSS}}(n - q - 1), \qquad (3.24)
\end{aligned}$$

where RSS, respectively RSS^0, is the residual sum of squares for the complete model (under H_1), respectively the reduced model (under H_0). Under H_0, asymptotically $\text{LRT} \sim \chi^2_k$. Up to a multiplicative constant, the LRT is equal to the F-test statistic, i.e. $F = \text{LRT}/k$ which under H_0 and under the normality assumption of the errors has an exact Fisher $F_{(k,n-q-1)}$ distribution.

The classical LRT (or indeed the F-test) is obviously not robust in the sense that small model deviations can lead to under or over rejection at a given level, or in other terms, that the actual test level does not correspond to the nominal one (see Section 2.4.2). The reasons are twofold: first the LS estimator is used for both the regression parameters and residual scale estimates (under the reduced and the full models) which as seen previously is not robust to small model deviations and, second, even if a robust estimator replaces the LS, the RSS in (3.24) would be inflated by the presence of model deviations in the form of e.g. outlying observations

(in both the response and the explanatory variables). Hence, a class of robust tests that controls both types of potential deviations is needed. For multiple hypothesis testing, a natural choice is the class of LRT-type tests (2.50) (see also Hampel *et al.* (1986, p. 354) for the regression model)

$$\mathrm{LRT}_\rho = 2\sum_{i=1}^{n}(\rho(r_i;\hat{\boldsymbol{\beta}}_{[M]},\sigma^2,c)-\rho(r_i;\dot{\boldsymbol{\beta}}_{[M]},\sigma^2,c)), \tag{3.25}$$

with $\widehat{\boldsymbol{\beta}}_{[M]}$ the M-estimator of $\boldsymbol{\beta}$ with Ψ-function $\Psi(r,\boldsymbol{x};\boldsymbol{\beta},\sigma^2,c)=\partial\rho(r;\boldsymbol{\beta},\sigma^2,c)/\partial\boldsymbol{\beta}$, $\dot{\boldsymbol{\beta}}_{[M]}$ the M-estimator of $\boldsymbol{\beta}$ under the reduced model in H_0 (i.e. solution of $\sum_{i=1}^{n}\Psi(r_i,\boldsymbol{x}_i;\boldsymbol{\beta},\sigma^2,c)_{(1)}=\mathbf{0}$ with $\boldsymbol{\beta}_{(2)}=\boldsymbol{\beta}^0_{(2)}$) and σ^2 replaced in practice by a consistent robust scale estimator (at the full model). As already stated in Section 2.5.4, under H_0, LRT_ρ[6] is in general asymptotically distributed as a weighted sum of χ^2_1 with weights that depend on the matrices Q and M in (3.18) and (3.19). For the regression model, inference is simplified in that (3.25) can be multiplied by[7]

$$\int \frac{\partial}{\partial r\partial r}\rho(r;\boldsymbol{\beta},\sigma^2,c)\,d\Phi(r)\Big/\int\left(\frac{\partial}{\partial r}\rho(r;\boldsymbol{\beta},\sigma^2,c)\right)^2 d\Phi(r) \tag{3.26}$$

and compared with a χ^2_k. This test is also known as the τ-test as in Hampel *et al.* (1986, Chapter 7).

An alternative testing procedure for the same null hypothesis is the Wald-type test statistic (2.47)

$$W^2_\Psi = n(\widehat{\boldsymbol{\beta}}_{[M](2)}-\boldsymbol{\beta}^0_{(2)})^T V(\Psi,\Phi)^{-1}_{(22)}(\widehat{\boldsymbol{\beta}}_{[M](2)}-\boldsymbol{\beta}^0_{(2)}),$$

with $V(\Psi,\Phi)=M(\Psi,\Phi)^{-1}Q(\Psi,\Phi)M(\Psi,\Phi)^{-T}$ evaluated at $\hat{\boldsymbol{\beta}}_{[M]}$ and at a consistent and robust estimator $\widehat{\sigma}^2$ of σ^2. Under H_0, W^2_Ψ follows asymptotically a χ^2_k distribution. As a choice for the Ψ-functions, we propose to take the same as for the estimators, i.e. $\Psi_{[bi]}$ defined through (3.13). It should be noted that W^2_Ψ is simpler than LRT_ρ to calculate because there is no need to compute $\dot{\boldsymbol{\beta}}_{[M]}$.

3.3.4 Diabetes Data Example (continued)

With the diabetes data example, an interesting hypothesis to test is the overall effect of the categorical variable frame, i.e. $H_0:\beta_5=\beta_6=0$ in model (3.23). Table 3.3 gives the result of the classical F-test and the robust LRT_ρ and Wald tests. We note that with this example, the classical and robust procedures lead to the same conclusion, i.e. the non-significance of H_0. This is not really surprising, at least with respect to the significance of each of the two dummy variables for the variable frame as given in Table 3.2.

[6]The LRT_ρ corresponds to a difference of deviances test in certain GLM families, see Section 5.2.2.

[7]For an analytical expression in the case of the biweight estimator, see Appendix B.

Table 3.3 Classical F-test, robust LRT_ρ and Wald tests for testing the significance of the body frame in the diabetes data example.

	df residuals $(n-q-1)$	df (k)	Test statistic	p-value
F-test	363	2	1.2772	0.2801
W_Ψ test $(\Psi_{[bi]})$		2	1.4640	0.4809
LRT_ρ test $(\rho_{[bi]})$		2	1.7466	0.4176

The ρ- and Ψ-functions are the biweight with $c = 3.8827$.

3.4 Checking and Selecting the Model

3.4.1 Residual Analysis

An important aspect of the analysis of the model is checking the model assumptions. This includes the (at least approximate) normality of the errors, the linearity of the explanatory variables in explaining the response variable and the possible presence of poorly fitted observations. This check can be successfully done through the analysis of the (estimated) standardized residuals. These are defined as $r_i = (y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})/\hat{\sigma}$ where $\hat{\boldsymbol{\beta}}$ is a chosen estimator, i.e. either the LS or a robust estimator, and $\hat{\sigma}$ is the classical or a robust estimator according to the choice of $\hat{\boldsymbol{\beta}}$. Under the regression model, the errors are assumed to be normally distributed, and hence the standardized residuals are expected to lie approximately within the bounds $(-1.96, 1.96)$ which correspond to 95% of the values of a standard normal variable. In other words, residuals exceeding (in absolute value) the bound of 1.96 can be considered as suspicious and, hence, as potential outliers.

Clearly, if the LS estimator is chosen, a residual analysis based on it can lead to wrong conclusions. In particular, one or a few outliers can attract the LS estimated regression line to them and making the resulting residuals relatively small, so that the outliers are 'masked'. This underlines the potential danger of a classical residual analysis.

3.4.2 GFR Data Example (continued)

As an illustration, we examine the GFR data and consider the simple model $\log(\texttt{gfr}) = \beta_0 + \beta_1 \texttt{cr}^{-1} + \epsilon$. In Figure 3.1 one can see that the LS regression line estimate is 'attracted' by observations 16 and 2. A standardized residual analysis is provided in Figure 3.3 for the LS and for the robust estimator. One can notice that both the LS and the robust estimator detect observation 2 as extreme in the sense that the value of its corresponding standardized residual (in absolute value) exceeds the value of 1.96. For observation 16, the story is different: it has a larger residual in the robust analysis, but still within the expected bounds, while it has a pretty average residual with a classical analysis. One also notices that the 'moon'-shaped residuals

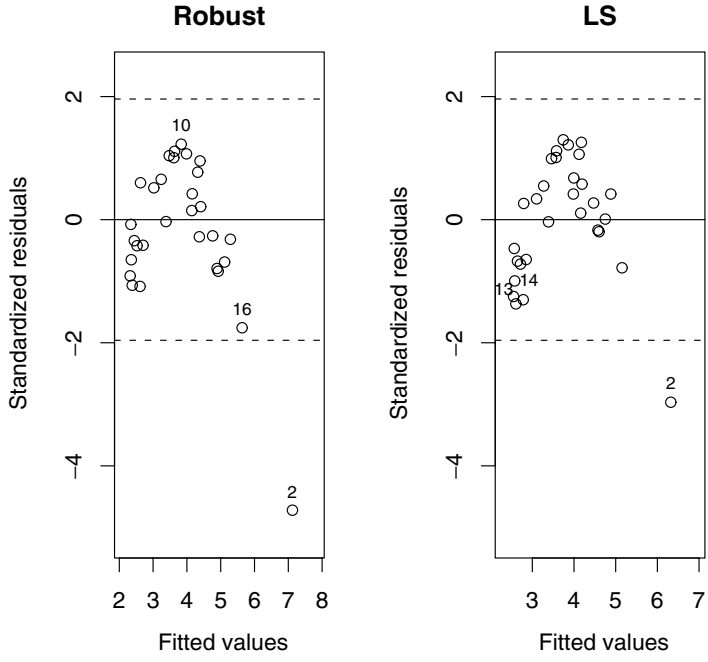

Figure 3.3 Residual analysis for the model $\log(\texttt{gfr}) = \beta_0 + \beta_1 \texttt{cr}^{-1}$ (robust residuals computed using the biweight *MM*-estimator with 90% efficiency).

cloud indicates (with both analyzes) that a more suitable model would be one that includes a quadratic term (i.e. $\texttt{cr}^{-2}$). We analyze this model later.

Before that, and just as an exercise, we now change the value of `cr` for observation 16 to match the value of `cr` for observation 2. The estimated regression lines are now given in Figure 3.4 (compare with Figure 3.1). The estimated regression parameters for $\texttt{cr}^{-1}$ are 1.56 with the LS estimator and 2.38 with the robust estimator. The scale estimates are 0.63 with the LS estimator and 0.53 with the robust estimator. We note that the robust scale estimate has not changed with the transformation of observation 16. A residual analysis is provided in Figure 3.5. With this modified sample, the LS estimator shows signs of breakdown, since the two extreme observations cannot really be considered as extreme on the basis of their standardized residuals. The robust estimator, on the other hand, clearly flags observations 2 and 16 as outliers (large standardized residuals). Moreover, the 'moon'-shaped cloud is now not so clear with the LS analysis, showing that a classical residual analysis can be very misleading. This is an important aspect of the non-robustness of LS-based residuals, i.e. the overestimation of the residual scale. Indeed, if there are outliers in the sample, then the residual scale is overestimated, which makes the detection of extreme observations more difficult.

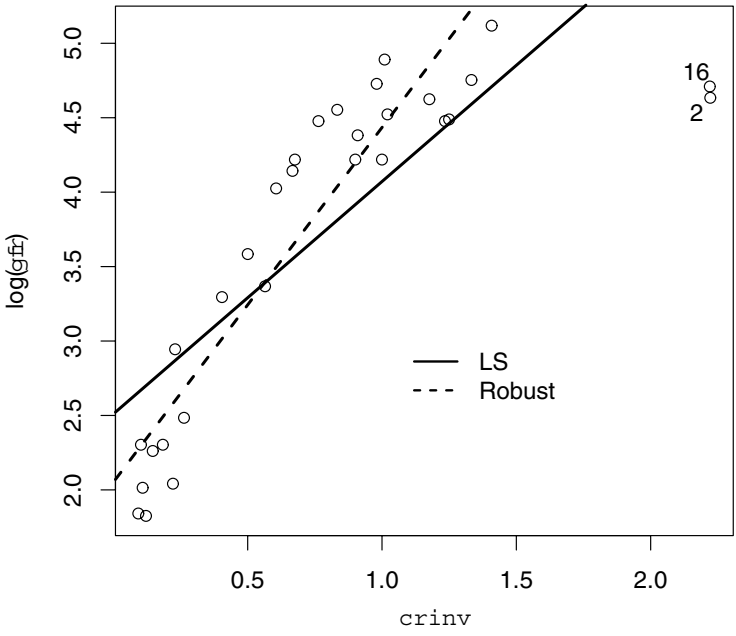

Figure 3.4 Estimated regression lines for the model $\log(\texttt{gfr}) = \beta_0 + \beta_1 \texttt{cr}^{-1}$ with $\texttt{cr}_{16} = \texttt{cr}_2$.

As we have already noticed before, a more suitable model for these data is the model $\log(\texttt{gfr}) = \beta_0 + \beta_{11}\texttt{cr}^{-1} + \beta_{12}\texttt{cr}^{-2} + \epsilon$. In Table 3.4 we present the estimated regression parameters (and significance tests) using the robust estimator $\widehat{\beta}_{[bi]}$, and the classical $\widehat{\beta}_{[LS]}$ both on the complete sample and on the sample without observation 2. In Figures 3.6 and 3.7 we present the corresponding residual analysis. The aim is to study the stability of both estimators when one observation is deleted. The first noticeable difference lies in the estimates for β_{11} and β_{12}. Indeed, the robust estimator on both the complete and reduced sample (without observation 2) and the LS estimator on the reduced sample provide approximately the same estimates, whereas the LS on the full sample provides values that are different. This difference is quite important as illustrated in the predicted values for gfr in Figure 3.8. In this figure, the predictions curves for the robust and the LS without observation 2 are confounded. A close inspection of the residuals on the complete sample (Figure 3.6) shows that observation 2 is not an extreme observation for the LS estimator. This indicates that the quadratic model estimated by means of the LS is able to 'fit' the extreme observation 2. However, at the same time it does not capture so well the quadratic nature of the relationship between $\log(\texttt{gfr})$ and $\texttt{cr}^{-1}$ and $\texttt{cr}^{-2}$ as illustrated in Figure 3.8. On the other hand, the robust estimator still considers observation 2 as extreme (at least with respect to the postulated model) and is able to capture the quadratic relationship. Therefore, in

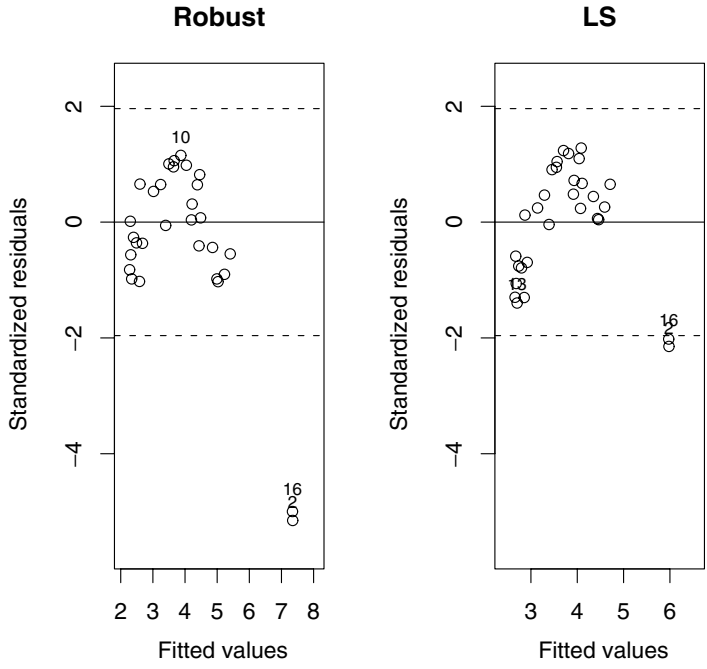

Figure 3.5 Residual analysis for the model $\log(\texttt{gfr}) = \beta_0 + \beta_1 \texttt{cr}^{-1}$ with $\texttt{cr}_{16} = \texttt{cr}_2$.

trying to accommodate observation 2, the LS fit tends to predict higher gfr values for large values of $\texttt{cr}^{-1}$. The equation found by Rule *et al.* (2004), who fitted a similar model is $\log(\texttt{gfr}) = 1.911 + 5.249\texttt{cr}^{-1} - 2.114\texttt{cr}^{-2} - 0.00686\texttt{age}$ for males. This finding is consistent with the robust fit, especially for large values of $\texttt{cr}^{-1}$. Finally, note that the residual analyses in Figure 3.7 coincide since the LS fit is performed on the sample without observation 2.

3.4.3 Diabetes Data Example (continued)

As another example, consider the diabetes data and the estimated model given in Table 3.2. The residual scale estimates are quite different between the LS and the robust estimators and this has an implication for the residual analysis as illustrated in Figure 3.9: some robust standardized residuals are larger, although for the bulk of the data they are similar between the two analyses. In particular, observations 180, 309 and 336 are found to be extreme with both analyses, but far more extreme (larger standardized residual) with the robust analysis. These three observations correspond respectively to a 68-year-old man with an average response but an extreme value for the stabilized glucose, a 26-year-old man with a larger than average response and a

Table 3.4 Estimated regression parameters and significance tests for model $\log(\texttt{gfr}) = \beta_0 + \beta_{11}\texttt{cr}^{-1} + \beta_{12}\texttt{cr}^{-2}$ for the GFR data.

	Complete sample			
	LS		Robust	
	Estimate (*SE*)	*p*-value	Estimate (*SE*)	*p*-value
`intercept`	1.63 (0.12)	$<10^{-4}$	1.48 (0.12)	$<10^{-4}$
$\texttt{cr}^{-1}$	4.29 (0.27)	$<10^{-4}$	5.11 (0.38)	$<10^{-4}$
$\texttt{cr}^{-2}$	−1.36 (0.14)	$<10^{-4}$	−1.98 (0.25)	$<10^{-4}$
$\widehat{\sigma}$	0.27		0.28	
R^2	0.942		0.957	
	Sample without observation 2			
	LS		Robust	
	Estimate (*SE*)	*p*-value	Estimate (*SE*)	*p*-value
`intercept`	1.48 (0.12)	$<10^{-4}$	1.47 (0.12)	$<10^{-4}$
$\texttt{cr}^{-1}$	5.06 (0.40)	$<10^{-4}$	5.12 (0.39)	$<10^{-4}$
$\texttt{cr}^{-2}$	−1.95 (0.26)	$<10^{-4}$	−1.99 (0.25)	$<10^{-4}$
$\widehat{\sigma}$	0.25		0.26	
R^2	0.952		0.957	

The estimates are the LS and the biweight *MM*-estimator with $c = 3.8827$ (90% efficiency).

very small value for the stabilized glucose and a 60-year-old woman with an extreme response.

A residual analysis can also be used to check the linearity of the relationship between each (non-categorical) explanatory variable. In Figure 3.10 we present the standardized residuals versus each of the non-categorical explanatory variables of model (3.23) for the LS and robust regression. Except for the extreme outliers that have been spotted in the residual analysis in Figure 3.9, there is no apparent non-linearity of the relationship between each explanatory variable and the response variable.

3.4.4 Coefficient of Determination

A summary measure for the goodness of fit of the model is given by the coefficient of determination R^2 which estimates the percentage of variance of the response variable explained by its (linear) relationship with the explanatory variables. Classically, it is

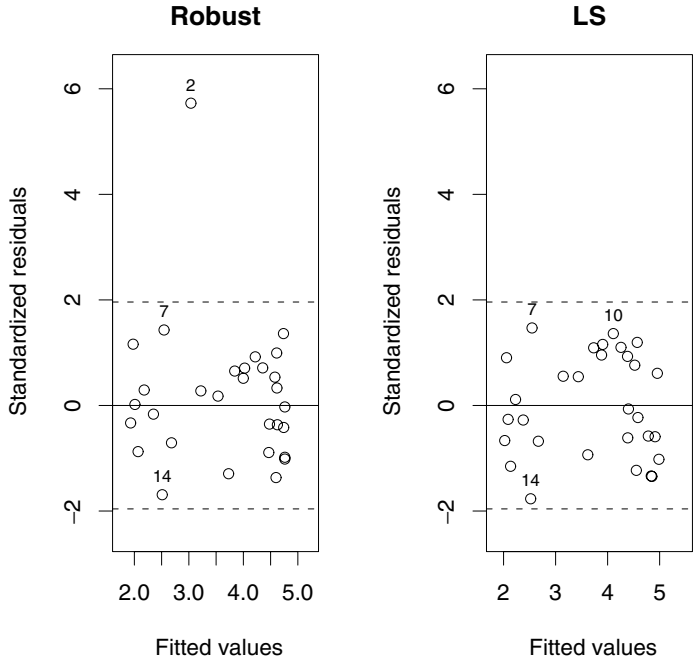

Figure 3.6 Residual analysis for the model $\log(\texttt{gfr}) = \beta_0 + \beta_{11}\texttt{cr}^{-1} + \beta_{12}\texttt{cr}^{-2}$, full GFR data sample.

computed by means of the ratio

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{\text{TSS} - \text{RSS}}{\text{TSS}}$$
$$= \frac{\sum_{i=1}^{n}(y_i - \overline{y})^2 - \sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}_{[LS]})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}, \tag{3.27}$$

where ESS, TSS and RSS are the explained, total and residual sum of squares, respectively. The coefficient of determination is actually equal to the square of the correlation coefficient between y_i, and the predicted response $\hat{y}_i$ (if there is an intercept term), i.e. (see e.g. Greene, 1997, p. 253)

$$R^2 = \left(\frac{\sum_{i=1}^{n}(y_i - \overline{y})(\hat{y}_i - \overline{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \overline{\hat{y}})^2}} \right)^2, \tag{3.28}$$

with $\overline{\hat{y}}$ the mean predicted responses. Definition (3.28) has a nice interpretation in that R^2 measures the goodness of fit of the regression model by its ability to predict the response variable. The R^2 is often adjusted for the sample size and the number of explanatory variables, i.e. $R^2_{adj} = 1 - (1 - R^2)((n-1)/(n-q))$.

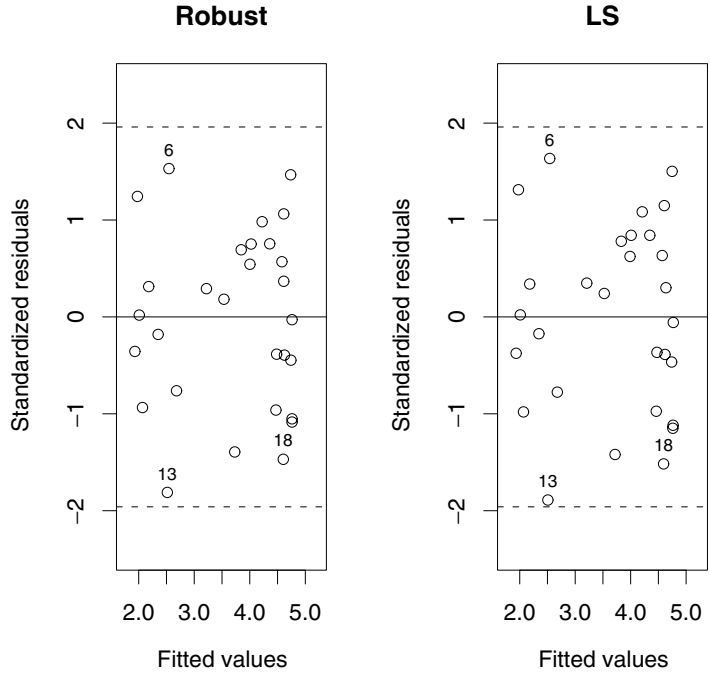

Figure 3.7 Residual analysis for the model $\log(\texttt{gfr}) = \beta_0 + \beta_{11}\texttt{cr}^{-1} + \beta_{12}\texttt{cr}^{-2}$, GFR data sample without observation 2.

Again, it is obvious that the R^2 can be driven by extreme observations, not only through the LS estimator $\widehat{\boldsymbol{\beta}}_{[LS]}$, but also through the average response $\overline{y}$ and the possible large residuals or deviations $y_i - \overline{y}$. We propose here to measure the goodness of fit of the model by means of the robust R^2 proposed by Renaud and Victoria-Feser (2009)

$$R^2_w = \left(\frac{\sum_{i=1}^n w_i (y_i - \overline{y}_w)(\hat{y}_i - \overline{\hat{y}}_w)}{\sqrt{\sum_{i=1}^n w_i (y_i - \overline{y}_w)^2 \sum_{i=1}^n w_i (\hat{y}_i - \overline{\hat{y}}_w)^2}} \right)^2, \tag{3.29}$$

where $\overline{y}_w = (1/\sum w_i) \sum w_i y_i$, $\overline{\hat{y}}_w = (1/\sum w_i) \sum w_i \hat{y}_i$ and the weights w_i are produced by the robust regression estimator, for example Tukey's biweight. Renaud and Victoria-Feser (2009) showed that under some conditions, (3.29) can be written as a robust extension of (3.27), i.e.[8]

$$R^2_w = \frac{\sum_{i=1}^n w_i (y_i - \overline{y}_w)^2 - \sum_{i=1}^n w_i (y_i - \widehat{y}_i)^2}{\sum_{i=1}^n w_i (y_i - \overline{y}_w)^2}. \tag{3.30}$$

[8] R^2_w can possibly be corrected for consistency, see Renaud and Victoria-Feser (2009).

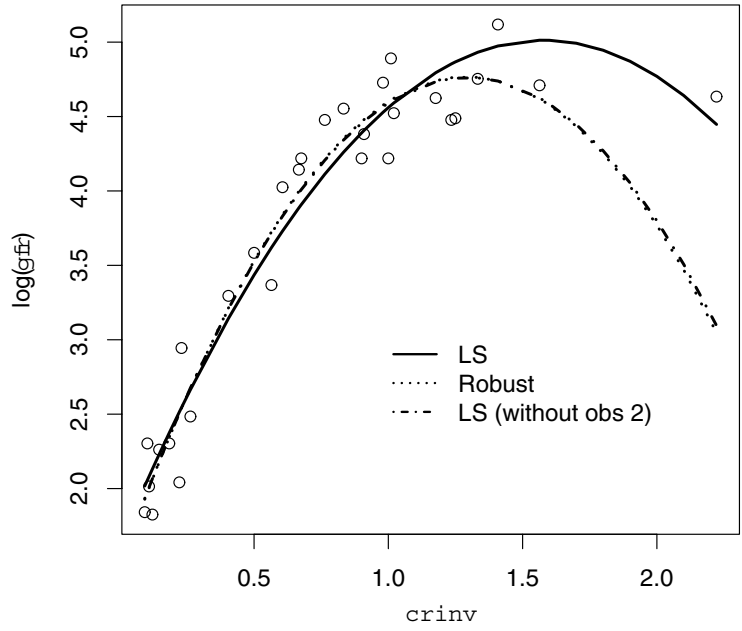

Figure 3.8 Predicted gfr values using the estimated regression models $\log(\texttt{gfr}) = \beta_0 + \beta_{11}\texttt{cr}^{-1} + \beta_{12}\texttt{cr}^{-2}$.

In the diabetes data example (see Table 3.2), the R^2 are different between a classical and a robust analysis. It is larger with a robust fit, hence reflecting the fact that there are extreme observations that lead to a poorer fit for all data with the LS estimator and a better fit for the majority of the data with the robust estimator. This can be better understood by means of a scatterplot between the responses and their predictions as in Figure 3.11. The scatter of the data reflects the degree of correlation between the response and its prediction, and hence the coefficient of determination. In the robust version, the deviations in (3.29) are weighted, and observations receiving a small weight (below 0.3) have been spotted in the (robust) scatter by means of the symbol '○'. It is clear that the correlation between the response and its prediction is stronger in the robust analysis than in the classical analysis, not only because extreme observations (with respect to the regression model) have been downweighted, but also because the estimated line is not biased by these extreme observations.

3.4.5 Global Criteria for Model Comparison

When the postulated regression model includes several explanatory variables, more often than not can one observe that some of them are correlated. Such variables then explain the same part of the response variable, which makes their inclusion altogether in the model useless, or worse can lead to the conclusion that these variables are

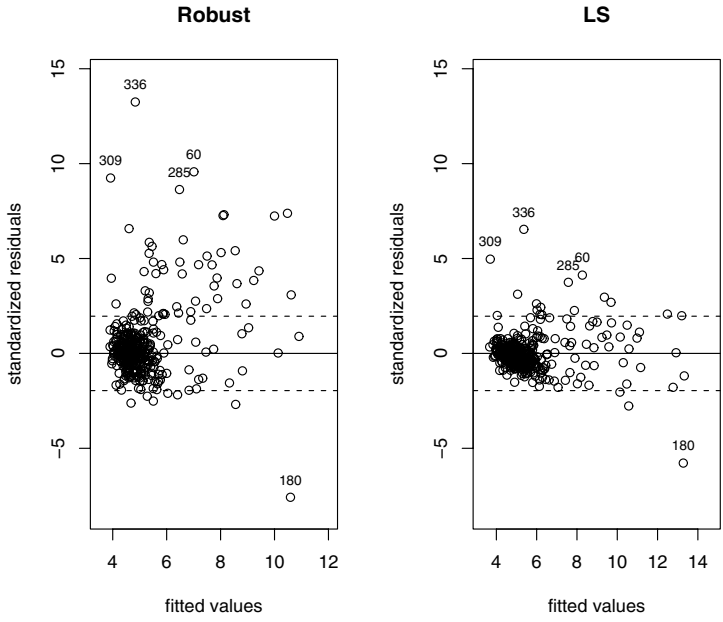

Figure 3.9 LS and robust standardized residual analysis of model (3.23) for the diabetes data.

not significant. This phenomenon is known as the problem of multicollinearity. Variable selection procedures are then necessary to objectively choose a suitable subset of regression variables. These procedures are different in spirit than a test for comparing different models such as the LRT (3.24). They are based on the optimization of a criterion that in some sense gives an indication of the fit of the data to the postulated model. In practice, all possible models are built with all possible combinations of the available explanatory variables, then the criterion is computed for each of these models and the 'best' model is that with the best value (minimum or maximum) for the criterion. One has however to be cautious with the 'best' choice, since the criterion is actually computed from the data and hence is also a random variable, which means that two models can in principle have two criteria that are not significantly different. Hence, it is safer to consider a few (two or three) 'best' models.

We propose in this section robust alternatives to classical model selection criteria that are not (or less) influenced by data that cannot be considered as having been generated by the postulated regression model. One could argue, however, that when assessing a model, the chosen criterion should also consider these extreme values and hence in some sense represent the bad fit of the model to some data. However, this goes against the fundamental ideas supporting a robust approach. Indeed, one

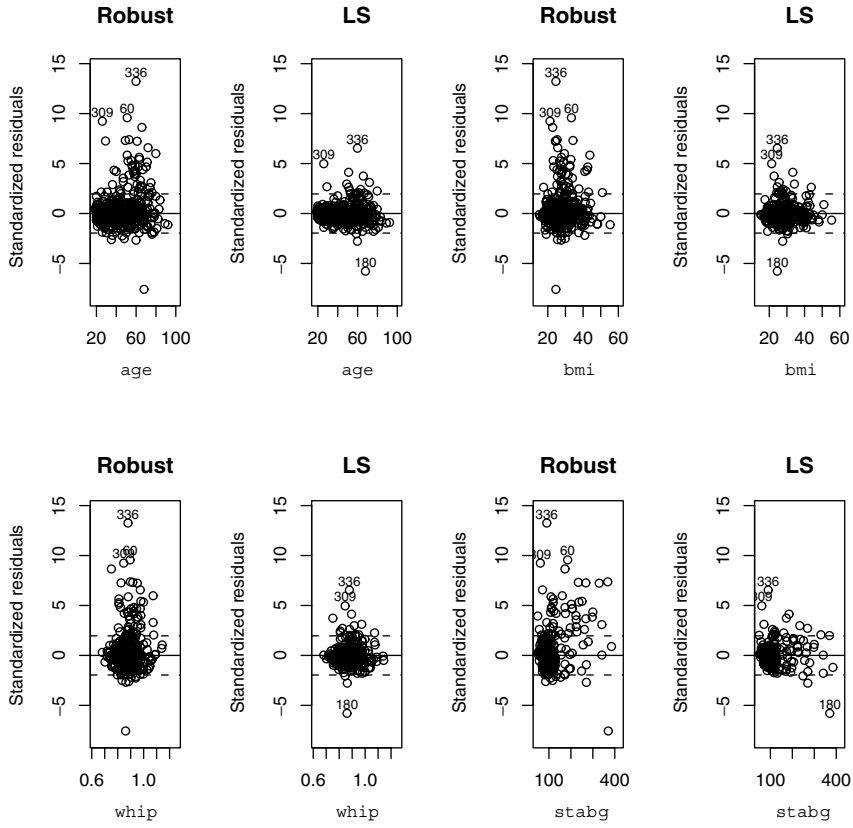

Figure 3.10 LS and robust standardized residuals of model (3.23) versus covariates for the diabetes data.

is interested in the model that fits the data in general, not all of the data because we suppose, in robust statistics, that not all of the data have been generated by the postulated model. In other words, it is better to have a good model for the majority of the data than an 'average' model for all of the data. Therefore, a robust procedure should also be used for selecting the 'best' models.

3.4.5.1 AIC

The AIC (Akaike, 1973) is one of these criteria. Let $\boldsymbol{x}_{(p)}$ be a subset of $p \leq q+1$ explanatory variables taken from the complete set $\boldsymbol{x}$ (including the intercept) and $\boldsymbol{\beta}^p$ the corresponding regression parameters in the regression model with $\boldsymbol{x}_{(p)}$. The AIC

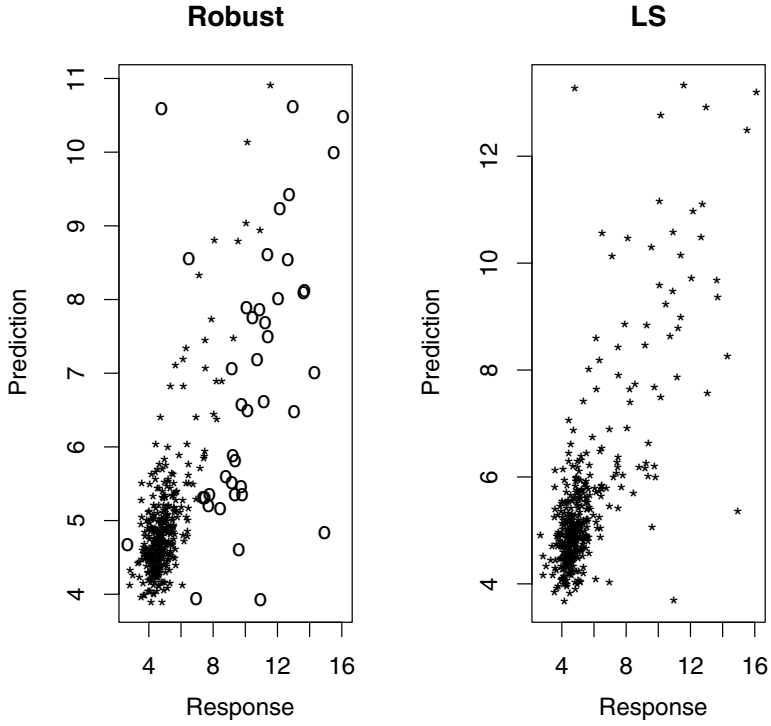

Figure 3.11 Response versus robust and classical (LS) prediction of model (3.23) for the diabetes data. The symbol '○' corresponds to observations that have received a robust weight below 0.3.

is given by

$$
\begin{aligned}
n\text{AIC} &= -2\sum_{i=1}^{n} \log f(y_i, \boldsymbol{x}_{i(p)}; \widehat{\boldsymbol{\beta}}^p_{[LS]}, (\widehat{\sigma}^p_{[LS]})^2) + 2p \\
&= -2\sum_{i=1}^{n} \left[-\log(\sqrt{2\pi}\widehat{\sigma}^p_{[LS]}) - \frac{1}{2}(r^p_{[LS]i})^2 \right] + 2p,
\end{aligned}
$$

with $r^p_{[LS]i} = (y_i - \boldsymbol{x}^T_{i(p)}\hat{\boldsymbol{\beta}}^p_{[LS]})/\hat{\sigma}^p_{[LS]}$. AIC is actually the value of (minus twice) the log-likelihood function at the estimated model with a subset of p explanatory variables (including the intercept) penalized by the number of regression coefficients p. The 'best' model is the model with the smallest corresponding value for the AIC. The penalty is necessary, because the greater the number of explanatory variables in the model (even if they do not explain the response), the greater the value of the log-likelihood function and hence, without penalization, the minimization would be achieved at the model with all of the available explanatory variables.

Again, as for the LRT statistic, it is obvious that the AIC is not robust against extreme values in the data. This is not only due to the fact the LS estimator for the regression parameters and for the residual scale is not robust, but also because one extreme residual can make a too negative contribution to the log-likelihood function. A robust version for the AIC was first proposed by Ronchetti (1982b) (see also Ronchetti, 1997b). For the regression model it is given generally by

$$\text{RAIC} = 2\sum_{i=1}^{n} \rho(r_i^p; \boldsymbol{\beta}_{[M]}^p, \sigma^2, c) + 2p\frac{a}{b} \tag{3.31}$$

with

$$a = \int \left(\frac{\partial}{\partial r}\rho(r; \boldsymbol{\beta}, \sigma^2, c)\right)^2 d\Phi(r)$$

$$b = \int \frac{\partial^2}{\partial r \partial r}\rho(r; \boldsymbol{\beta}, \sigma^2, c)\, d\Phi(r)$$

and r_i^p computed using a robust M-estimator $\hat{\boldsymbol{\beta}}_{[M]}^p$ and σ^2 replaced by a robust estimator of residual scale at the model with all explanatory variables. One can, in principle, choose any ρ function defining a robust estimator. We propose here to take the biweight $\rho_{[bi]}$ in (3.15) and corresponding $\Psi_{[bi]}$.[9]

3.4.5.2 Mallows's C_p

Another variable selection procedure is given by Mallows's C_p (Mallows, 1973) which is based on a prediction error criterion. It is given by

$$\begin{aligned} C_p &= \frac{1}{\widehat{\sigma}_{[LS]}^2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_{i(p)}^T\widehat{\boldsymbol{\beta}}_{[LS]}^p)^2 - n + 2p \\ &= \frac{1}{\widehat{\sigma}_{[LS]}^2}\text{RSS}_p - n + 2p, \end{aligned}$$

where $\widehat{\boldsymbol{\beta}}_{[LS]}^p$ is the LS estimator of $\boldsymbol{\beta}^p$ and $\widehat{\sigma}_{[LS]}^2 = \text{SSR}_{q+1}/(n-q-1)$ is the LS residual scale estimate at the complete model. One can notice that when $p = q+1$, then $C_p = p$, i.e. $C_{q+1} = q+1$. Here C_p actually estimates the prediction error of the model measured by

$$\frac{1}{\sigma^2}\text{E}\left[\sum_{i=1}^{n}(\widehat{y}_i^p - \text{E}[y_i \mid \boldsymbol{x}_{i(p)}])^2\right]$$

where $\widehat{y}_i^p = \boldsymbol{x}_{i(p)}^T\widehat{\boldsymbol{\beta}}^p$ is the predicted value at the model with $\boldsymbol{x}_{(p)}$ for a chosen estimator $\widehat{\boldsymbol{\beta}}^p$.

[9] In this case the analytical expressions for a and b are given in Appendix B.

Like for the AIC, the C_p is sensitive to small model deviations from (each of) the assumed models. As for the AIC, it is important that a selection procedure is not affected by a few extreme observations, otherwise the 'best' models would represent an average model for all of the data, rather than a good model for the majority of the data. Ronchetti and Staudte (1994) propose a robust alternative to the C_p as an estimator of

$$\Gamma_p = \frac{1}{\sigma^2}\mathrm{E}\left[\sum_{i=1}^{n}(\widehat{w}^p_{[M]i}(\widehat{y}^p_{[M]i} - \mathrm{E}[y_i \mid \boldsymbol{x}_{i(p)}]))^2\right]$$

where $\widehat{y}^p_{[M]i} = \boldsymbol{x}^T_{i(p)}\widehat{\boldsymbol{\beta}}^p_{[M]}$ and $\widehat{\boldsymbol{\beta}}^p_{[M]}$ is the M-estimator (with corresponding Ψ-function) for the model with $\boldsymbol{x}_{(p)}$, and

$$\widehat{w}^p_{[M]i} = \left[\frac{\partial}{\partial r}\rho(r; \hat{\boldsymbol{\beta}}^p_{[M]}, \sigma^2, c)\Big/ r\right]_{r=r_i}$$

are the weights given to each residual by the M-estimator, and ρ is such that $\Psi(r; \boldsymbol{x}; \boldsymbol{\beta}, \sigma^2, c) = \partial\rho(r; \boldsymbol{\beta}, \sigma^2, c)/\partial\boldsymbol{\beta}$. The weights are different for each model since an observation can be outlying with respect to one model and have full weight in another. The weighting scheme therefore not only has the effect of downweighting the outlying observations with respect to model with $\boldsymbol{x}_{(p)}$ in estimating $\boldsymbol{\beta}^p$, but also limiting their influence on Γ_p and therefore on the model selection procedure itself. Ronchetti and Staudte (1994) show that a suitable estimator for Γ_p is

$$RC_p = \frac{1}{\widehat{\sigma}^2}W_p - (U_p - V_p) \tag{3.32}$$

with $W_p = \sum_{i=1}^{n}(\widehat{w}^p_{[M]i}(y_i - \widehat{y}^p_{[M]i}))^2$ a weighted RSS, $\widehat{\sigma}^2 = W_{q+1}/U_{q+1}$ a robust residual and consistent scale estimator at the full model, and U_p and V_p are quantities given in Ronchetti and Staudte (1994) or Ronchetti (1997b).[10] At $p = q + 1$, by definition $RC_p = V_p$, i.e. $RC_{q+1} = V_{q+1}$. Models for which $RC_p \approx V_p$ are among the 'best' models. When the weights $\widehat{w}^p_{[M]i}$ are all equal to one (e.g. if the M-estimator is the LS estimator), then $RC_p = C_p$, for all p.

Finally, other criteria computed in a robust fashion exist for variable selection. As an alternative to the RC_p, Machado and Machado (1993) propose a robust version of the Bayesian information criterion (BIC) (Schwarz, 1978) on objective functions defining M-estimators, and Sommer and Huggins (1996) propose a criterion based on the Wald test statistic. Ronchetti *et al.* (1997) develop a robust criterion based on cross-validation. However, it requires the splitting of the dataset and hence the estimation of all of the models for all of the splits, which, with robust estimators, can be computationally intensive. Müller and Welsh (2005) propose a RAIC-like criterion but instead of using the same ρ-function in (3.31) as that for the estimator for the regression coefficients $\boldsymbol{\beta}^p$, they use a simple bounded $\rho(r; c) = \min\{r^2; c^2\}$ with $c = 2$, and the expected value is estimated by means of a stratified bootstrap in which the strata are built according to the value of the residuals for a given fit. It is not however clear whether this procedure is better than the standard RAIC in (3.31) in terms of probability of choosing the correct model.

[10] For the biweight ρ-function, $U_p - V_p$ is given in Appendix B.

Table 3.5 AIC and RAIC of the 'best' models for the diabetes data. The values in brackets are the ranks (up to 10) of the models with each criteria, from best to worse. The last column is the mean of the robust weights (3.14) corresponding to each model.

Models	AIC	(rank)	RAIC	(rank)	Mean weights
AG	1345.62	(3)	354.13		0.797
ACG	1346.29	(9)	353.21		0.798
AEG	1345.24	(2)	355.65		0.797
AGH	1345.84	(5)	350.99	(4)	0.799
ACEG	1346.11	(7)	354.49		0.798
ACGH	1346.37	(10)	349.70	(1)	0.799
ADEG	1345.96	(6)	355.99		0.797
AEGH	1345.06	(1)	352.67		0.798
AFGH	1347.24		350.60	(3)	0.801
ABCGH	1348.10		351.28	(6)	0.800
ABFGH	1348.79		351.34	(8)	0.802
ACDGH	1347.88		351.03	(5)	0.800
ACEGH	1345.80	(4)	351.30	(7)	0.798
ACFGH	1346.81		350.48	(2)	0.800
ADEGH	1346.22	(8)	353.70		0.798
AEFGH	1347.02		351.77	(9)	0.798
ABCFGH	1348.79		351.85	(10)	0.801

A is for `age`, B for `sex`, C for `bmi`, D for `whip`, E for `bfmed`, F for `bflar`, G for `stabg`, and H for `loc`. The ρ-function is the biweight with $c = 3.8827$.

3.4.6 Diabetes Data Example (continued)

Some of the explanatory variables in the diabetes dataset are probably correlated (e.g. `bmi` and `whip`), hence a (robust) selection procedure is necessary to choose among the 'best' models. For all possible models, the AIC and RAIC are computed and the best ones, i.e. those with smallest AIC and RAIC are presented in Figures 3.12 and 3.13. The corresponding values are given in Table 3.5.

The classical AIC proposes for the three 'best' models, the models with the variables `age`, `bfmed`, `stabg` and `loc` (1), `age`, `bfmed` and `stabg` (2) and `age` and `stabg` (3). The RAIC on the other hand, proposes for the three 'best' models the models with the variables `age`, `bmi`, `stabg` and `loc` (1), `age`, `bmi`, `bflar`, `stabg` and `loc` (2) and `age`, `bflar`, `stabg` and `loc` (3). The variables `age` and `stabg` are in all selected models, whatever the method. The variables `loc` and either `bmi` or `bflar` are always selected with the robust method. Finally `sex` and `whip` are never selected. The models selected by means of the classical AIC are different from those selected by the RAIC and this difference is certainly due to extreme data in the sample that were spotted in the residual analysis. In Tables 3.6 and 3.7 we present the estimated parameters for both the 'best' model chosen by the

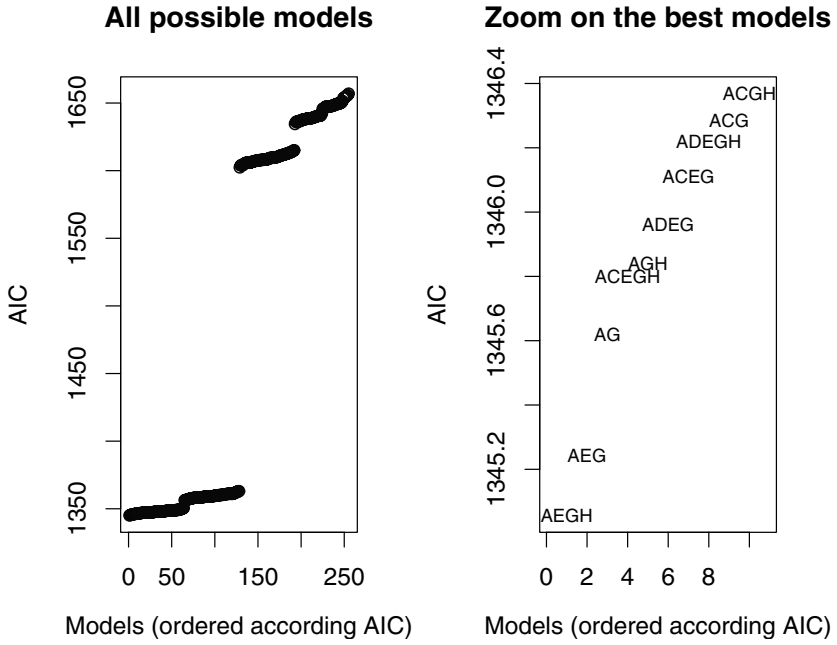

Figure 3.12 AIC for the 'best' models. A is for `age`, B for `sex`, C for `bmi`, D for `whip`, E for `bfmed`, F for `bflar`, G for `stabg`, and H for `loc`.

Table 3.6 LS estimated parameters of the 'best' model selected by the AIC.

Coefficient	Estimate (*SE*)	*p*-value
`intercept`	1.6 (0.27)	$<10^{-4}$
`age`	0.019 (0.005)	1.2×10^{-4}
`bfmed`	0.25 (0.15)	0.10
`stabg`	0.029 (0.001)	$<10^{-4}$
`loc`	−0.22 (0.15)	0.14
$\widehat{\sigma}$	1.46	
R^2	0.572	

classical AIC and estimated by the LS estimator, and the 'best' model selected by the RAIC and estimated by the robust biweight estimator.

It is not necessarily expected that the selected models lead to explanatory variables that are all significant, since the criteria used for selection do not contain the same information as the test statistics used for significance testing. However, one can notice that the model selected by means of the classical AIC and estimated by the LS estimator contains variables that are not significant (`bfmed` and `loc`), while all

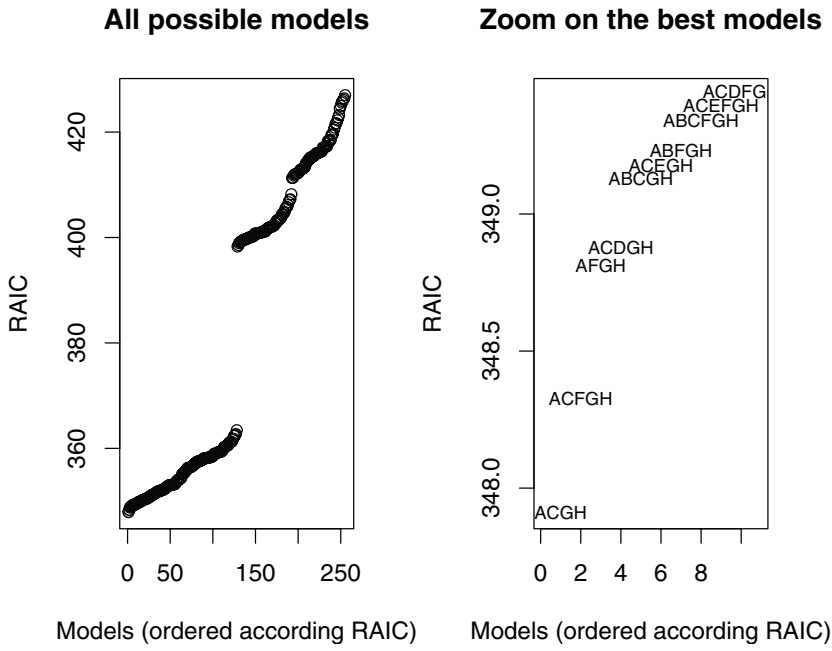

Figure 3.13 RAIC for the 'best' models. A is for `age`, B for `sex`, C for `bmi`, D for `whip`, E for `bfmed`, F for `bflar`, G for `stabg`, and H for `loc`.

Table 3.7 Robust estimated parameters of the 'best' model selected by the RAIC.

Coefficient	Estimate (*SE*)	p-value
`intercept`	2.02 (0.40)	$<10^{-4}$
`age`	0.012 (0.003)	$<10^{-4}$
`bmi`	0.017 (0.007)	0.011
`stabg`	0.021 (0.004)	$<10^{-4}$
`loc`	−0.25 (0.10)	0.015
$\widehat{\sigma}$	0.760	
R^2	0.656	

The estimates are computed using the biweight *MM*-estimator with $c = 3.8827$ (90% efficiency).

coefficients are significant with the model selected by means of the robust RAIC and estimated by the robust biweight estimator. This might indicate that a full robust procedure (estimation, testing and model choice) is more stable than a classical procedure, or, in other words, that small model deviations such as outliers do affect the different steps of the full classical procedure in different ways.

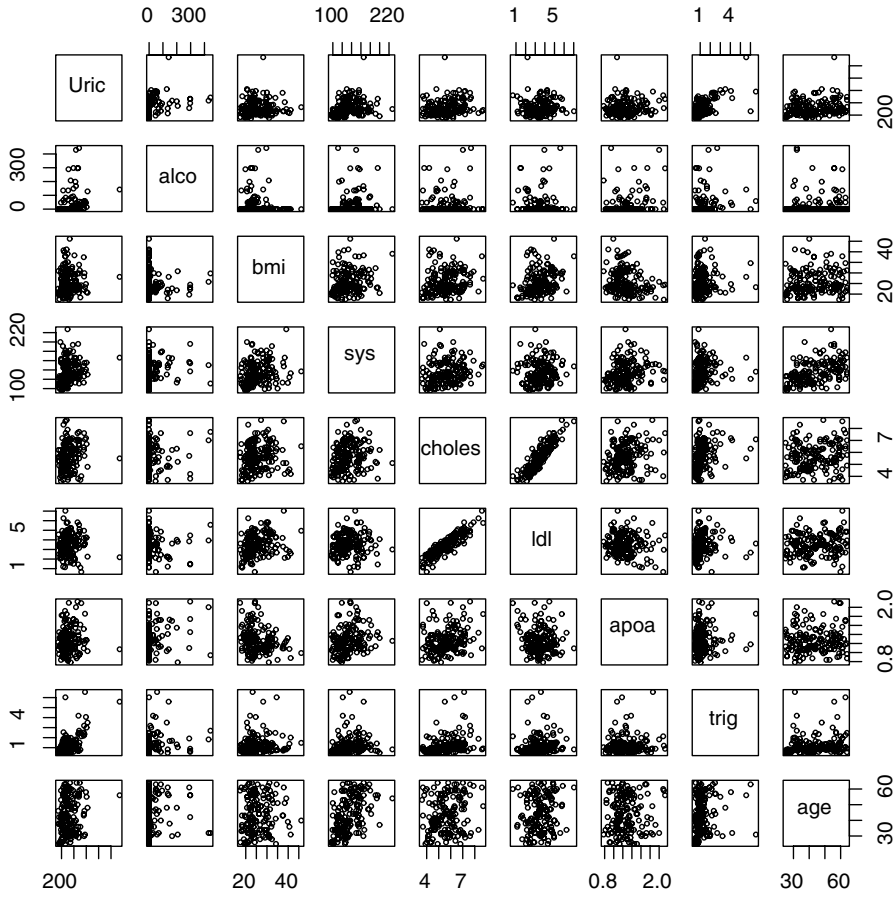

Figure 3.14 Scatter diagram for the cardiovascular risk factors data.

3.5 Cardiovascular Risk Factors Data Example

In this section, we fully analyze the dataset on cardiovascular risk factors briefly presented in Section 1.3. This dataset comes from a study aimed at investigating the prevalence of hyperuricemia and the association between uric acid levels and various cardiovascular risk factors in a developing country with high average blood pressures; see Conen *et al.* (2004). The 998 participants aged 25 to 64 years (mean age of 45) live in the Seychelles, a group of 115 islands lying in the Indian Ocean, and belong to a population mainly of African origin. There are 474 men (hence 524 women) and different measures were taken that concern physiological as well as behavioral characteristics of the participants. We consider here as potential risk factors (explanatory variables) the body mass index (`bmi`), systolic blood

Table 3.8 Robust estimates of the regression parameters and significance tests for the cardiovascular risk factors data (full model).

	Estimate (*SE*)	*p*-value
`intercept`	5.17 (18)	0.775
`smok`	−9.3 (7)	0.1848
`alco`	0.108 (0.046)	0.019
`bmi`	2.947 (0.454)	$<10^{-4}$
`sys`	0.432 (0.101)	$<10^{-4}$
`choles`	22.70 (10.27)	0.027
`ldl`	−22.93 (10.12)	0.024
`apoa`	−12.04 (16.8)	0.475
`trig`	60.51 (6.02)	$<10^{-4}$
`age`	0.367 (0.208)	0.078
`sex`	110.6 (5.28)	$<10^{-4}$
$\widehat{\sigma}$	59.6	
R^2	0.725	

The estimates are computed using the biweight *MM*-estimator with $c = 3.8827$ (90% efficiency).

pressure (`sys`), low-density lipoprotein cholesterol (`ldl`), triglycerides level in body fat (`trig`), total cholesterol (`choles`), apoprotein A (`apoa`) which is highly correlated with high-density lipoprotein cholesterol, the smoking habit (`smok`) a dummy variable with one for regular smoker, the alcohol intake (`alco`) in milliliters per day, together with `age` and `sex`, a dummy variable with one for men. Conen *et al.* (2004) consider slightly different explanatory variables and estimate a regression model for men and a separate model for women. They use Stata's *rreg* command for robust regression estimation. *rreg* uses an IRWLS algorithm with a Huber-type estimator (3.6), but with biweight weights (2.23) with $c = 4.685$ (i.e. an efficiency of 95%). More precisely, for the first two steps of the IRWLS, Huber weights (2.16) with $c = 1.345$ are used, and then the procedure switches to biweight weights. The scale parameter (for scaling the residuals) is chosen as the median absolute deviation (MAD) of the residuals, i.e.

$$\hat{\sigma}_{\text{MAD}} = 1.483\,\text{med}|r_i - \text{med}(r_i)|$$

where the factor 1.483 ensures consistency at the normal model (see e.g Hampel *et al.*, 1986, p. 107). This kind of hybrid robust estimator is different from the one which we use here, in that it does not share the same properties (it does not have a high breakdown point). However, the results we find below are in accordance with the analysis of Conen *et al.* (2004), even if they cannot really be compared. Indeed, the models are (slightly) different, as is the robust estimator, and they do not use the same variable selection procedure (they use a stepwise procedure).

A scatter diagram of the continuous variables is given in Figure 3.14. One may notice that there appears to be some extreme observations and that some of the

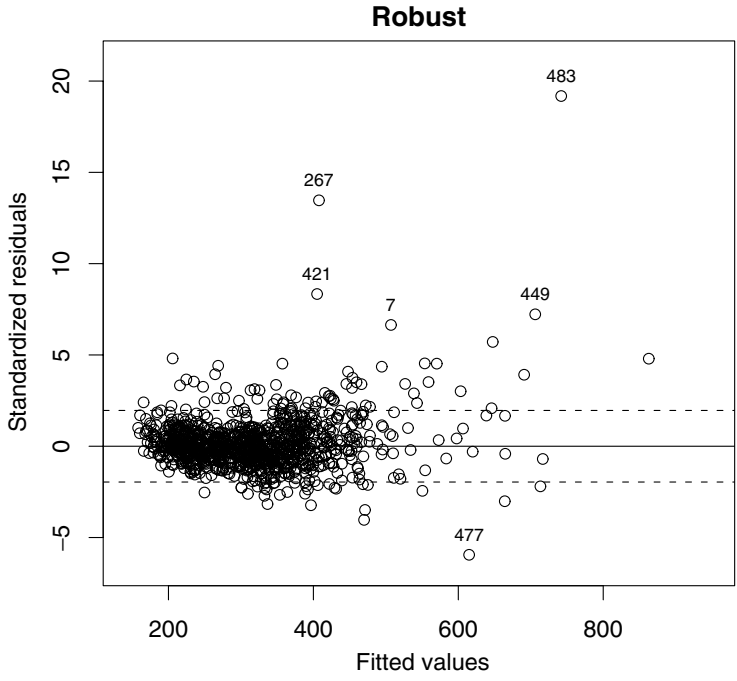

Figure 3.15 Residual analysis of the full regression model for the cardiovascular risk factors data.

explanatory variables are highly correlated (in particular `choles` and `ldl`). Hence, to fit and test a regression model we not only need a robust procedure but we also need to perform a variable selection procedure.

The robust estimates are given in Table 3.8. The variables `alco`, `bmi`, `sys`, `choles`, `ldl`, `trig` and `sex` are found to be significant at the 5% level. However, before any conclusion can be drawn, a residual analysis and a model selection procedure should be performed.

In Figure 3.15 we present the residual analysis corresponding to the full model. The observations with the six most extreme residuals are identified. The analysis shows several extreme observations with two being quite extreme.

In order to avoid the effect of correlated explanatory variables on the regression estimation and testing, we perform a variable selection using the RAIC (see Section 3.4.5). The results are summarized in Table 3.9. The RAIC selects models containing, for the first three models, mainly the variables `sys`, `ldl` and `sex`, and for some also the variables `bmi`, `smok`, `alco`, `apoa` and `age`. The estimation of the 'best' selected model is presented in Table 3.10. According to a robust analysis, the level of uric acid depends on the body mass index, the systolic blood pressure total cholesterol, the low-density lipoprotein cholesterol and gender. Except for the

Table 3.9 RAIC of the 'best' model for the cardiovascular risk factors data. The values in brackets are the ranks (up to 10) of the models, from best to worse. The last column is the mean of the weights (3.14) corresponding to each model.

Models	RAIC	(rank)	Mean weights
DEIJ	4875.56	(10)	0.839
ADFIJ	4867.88	(2)	0.840
BDEIJ	4870.33	(5)	0.835
CDEFJ	4865.57	(1)	0.837
ABDFIJ	4872.76	(6)	0.837
ABCFGIJ	4875.28	(8)	0.834
ABDFGIJ	4869.04	(3)	0.837
ABCEFHIJ	4875.52	(9)	0.845
ABCFGHIJ	4870.12	(4)	0.846
ACEFGHIJ	4874.83	(7)	0.848

A is for `smok`, B for `alco`, C for `bmi`, D for `sys`, E for `choles`, F for `ldl`, G for `apoa`, H for `trig`, I for `age` and J for `sex`. The ρ-function is the biweight with $c = 3.8827$ (90% efficiency).

Table 3.10 Robust estimates of the regression parameters and significance tests for the cardiovascular risk factors data (model selected by the RAIC).

	Estimate (*SE*)	*p*-value
`intercept`	−70.12 (21.8)	0.0014
`bmi`	4.74 (0.51)	$<10^{-4}$
`sys`	0.66 (0.11)	$<10^{-4}$
`choles`	41.11 (7.6)	$<10^{-4}$
`ldl`	−33.97 (7.8)	$<10^{-4}$
`sex`	122.2 (5.6)	$<10^{-4}$
$\widehat{\sigma}$	66.24	
R^2	0.585	

The estimates are computed using the biweight *MM*-estimator with $c = 3.8827$ (90% efficiency).

low-density lipoprotein cholesterol, the larger the values of these factors, the higher the level of uric acid. The latter is also higher for men in general.

At this stage, it is interesting to check which observations have been considered as too extreme by the robust analysis. For the robust selected model (see Table 3.10), the weights (i.e. (2.23)) are given in Figure 3.16. Some of the observations appear to

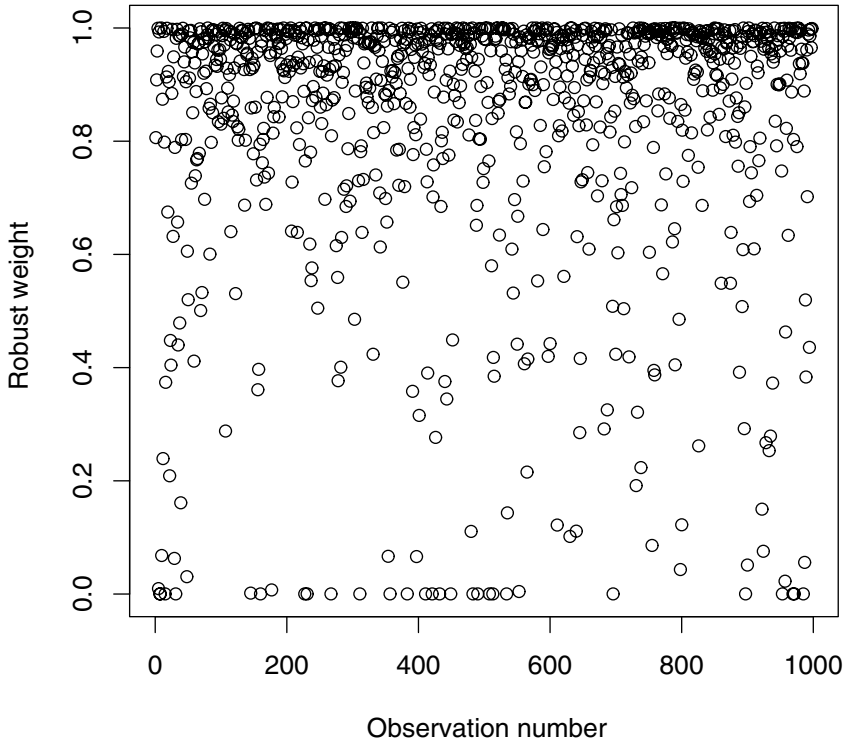

Figure 3.16 Robust weights (2.23) of the robust selected model of Table 3.10 for the cardiovascular risk factors data.

be receiving a weight nil or near zero. These are actually observations[11] numbers 7, 8, 15, 31, 60, 227, 231, 267, 311, 357, 383, 421, 432, 449, 483, 490, 508, 513, 534, 696, 897, 953, 969, 971 and 985. Recall that a residual analysis of the full model revealed that observations 477, 7, 449, 421, 267 and 483 were the most extreme observations (see Figure 3.15). In the reduced model, observation 477 no longer appears as extreme. It is indeed possible that observations appear extreme (large residual, low weight) in one model while being considered as not so extreme in another model. This is because the robust estimator produces weights that are relative to the chosen model, and hence for different models but the same dataset, the weights (and hence the residuals) are in general different, and sometimes even very different.

[11] Not shown on the graph.

4

Mixed Linear Models

4.1 Introduction

Mixed linear models (MLMs) nicely extend the regression model by including in the linear predictor a set of unobserved random terms making the response a sum of fixed and random effects; hence, the use of the term 'mixed'. The aim is generally to capture the response variability, possibly due to the random effect of groups of observations (e.g. experimental units) that share similar characteristics or correlation over time between repeated measures. MLM therefore apply to settings where several measurements are taken on the same experimental unit, such as the same subject, individuals of the same cluster, or multilevel data, etc. The statistical literature contains numerous references on MLMs, among which one can cite the recent books by Pinheiro and Bates (2000), Verbeke and Molenberghs (2000) and Diggle *et al.* (2002).

Historically, the seminal work of Fisher (1925) on the ANOVA can be considered as the beginning of the development of the MLMs. Until the mid-1950s, the focus is on variance components where description and quantification of variability of the data is of primary interest with applications in animal breeding, experimental designs and industrial quality control; see Searle *et al.* (1992) for a review. The years from 1950 to 1969 saw major developments in methods estimating variance components with, in particular, the pioneering work of Henderson (1953) for the unbalanced case. Since then, research in the field has slightly evolved to a broader use of MLM, pushed by research in medicine, health sciences, psychology and so forth. Such developments were made possible by the enormous progress in computing hardware and software. Classical estimation procedures to fit MLMs are the MLE or restricted (or residual) MLE of Patterson and Thompson (1971) and related tests, e.g. Fisher's F-test or standard Wald or LRT tests; see Searle *et al.* (1992) and Verbeke and Molenberghs (2000) for details. Unfortunately, these statistical procedures rely heavily on the normality assumption as small departures can have disastrous effects on classical estimators (bias) and tests (increased type I

Robust Methods in Biostatistics S. Heritier, E. Cantoni, S. Copt and M.-P. Victoria-Feser

or II errors). Historically, robustness in MLMs follows the general development of robustness theory, pushed by the ever-increasing number of applications of such models. Some early attempts are based on maximizing a robustified likelihood (Huggins, 1993; Huggins and Staudte, 1994) or the Student t-likelihood (Pinheiro *et al.*, 2001; Stahel and Welsh, 1992, 1997). Alternatives solving a weighted score equation are also proposed by Bednarski and Zontek (1996), Richardson (1997), Richardson and Welsh (1995) and Stahel and Welsh (1997). They generalize the influence function approach of Hampel *et al.* (1986) to the MLM setting but only focus on the estimation problem. A review of these methods can be found in Stahel and Welsh (1997). In recent years, Copt and Victoria-Feser (2006) reconsider the problem by using the normal multivariate formulation of MLMs and propose high breakdown point estimators (i.e. S-estimators; see Section 2.3.3). This work also allows the construction of robust Wald and score tests. Building on this idea, Copt and Heritier (2007) propose MM-estimators for MLMs and systematically focus on inferential issues. In particular, a robust LRT-type test is proposed as an alternative to the F-test. These latter alternatives have the advantages of being easier to compute (even for highly structured models) and allow robust inference on the fixed effects to be performed.

The chapter is organized as follows. In Section 4.2 we present a multivariate normal formulation of the MLMs and introduce several datasets that will be used throughout the chapter. The MLE, restricted maximum likelihood (REML) and related tests are reviewed in Section 4.3. We show both theoretically and empirically through a sensitivity analysis that these procedures are not robust. Different robust estimators are presented as an alternative in Section 4.4, including bounded-influence estimators (Richardson and Welsh, 1995; Richardson, 1997) that robustify the MLE/REML. The S- and MM-estimators of Copt and Victoria-Feser (2006) and Copt and Heritier (2007) are largely illustrated as they constitute simple alternatives with a high breakdown point. They pave the way for the definition of robust tests for the fixed effects that we present in Section 4.5. In particular, single hypotheses and contrasts can be tested through a robust Wald-type test while more general multivariate hypotheses are investigated with a LRT-type test. Inferential issues linked to testing hypotheses on the variance components are briefly discussed. Robust residuals and predictions are presented in Section 4.6 and illustrated through real examples. Three datasets are used to illustrate the theory throughout the different sections and a fourth dataset is thoroughly analyzed in Section 4.7, using the robust procedures introduced earlier. Section 4.8 finally discusses current limitations and extensions.

4.2 The MLM

4.2.1 The MLM Formulation

MLMs were originally introduced by Laird and Ware (1982) to better analyze longitudinal data. Fixed effects are used to explain the population average relationship

between the response and a set of predictors while heterogeneity across subjects under study or more generally clusters is accounted for by the inclusion of random effects. The model can be seen as an extension of the linear model specified by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{4.1}$$

where $\boldsymbol{y}$ is the N-vector of all measurements (observations) for all subjects, $\boldsymbol{\beta}$, a $(q+1)$-vector of unknown fixed (regression) parameters (including the intercept), $\boldsymbol{X}$ is a known $N \times (q+1)$ design matrix for the fixed effects and $\boldsymbol{\epsilon}$ is the N-vector of independent errors with $E[\boldsymbol{\epsilon}] = 0$ and $\text{var}(\boldsymbol{\epsilon}) = \sigma_\epsilon^2 \boldsymbol{I}_N$. In the MLM setting, the error terms $\boldsymbol{\epsilon}$ can no longer remain independent as typically one want to capture through the modeling some correlation present in the data, possibly due to repeated measurements on the same sampling unit (or cluster). A possible way to achieve this is to specify $\text{var}(\boldsymbol{\epsilon})$ differently, so that it reflects this heterogeneity in the data, or simply add some random terms to the model.

To fix ideas, consider a very simple example given in Berry (1987). The data come from an experiment in which five types of electrodes are applied to the arms of 16 subjects and their skin resistance measured (this example will be fully presented in Section 4.2.2). The skin resistance is the response that is assumed to depend on the electrode type which acts here as a fixed effect. Figure 4.1 displays a profile plot per subject of the skin resistance for the five electrode types. One can see that the 'mean' skin resistance per subject varies across subjects. In other words, there is a subject effect on the response that should be taken into account when building the model. This effect is introduced in the model as a random effect for the subject (which here is the cluster).

In other words, let y_{ij} be the response (i.e. resistance) of subject i on electrode j, λ_j the (fixed) effect of the jth electrode and s_i the (random) effect of the ith subject on the response variable, the MLM can be written as

$$y_{ij} = \mu + \lambda_j + s_i + \epsilon_{ij}, \quad i = 1, \ldots, n,\ j = 1, \ldots, p \tag{4.2}$$

where μ is the grand mean, a parameter that is often introduced in a MLM, which implies then that the λ_j must be constrained by means of $\sum_{j=1}^{p} \lambda_j = 0$, and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$ is the residual error term. In this example, $p = 5$. Since the subjects were randomly selected from some population, they are considered as random, and hence their effect on the response (the resistance) cannot be considered as systematic. Therefore, it is assumed that $s_i \sim \mathcal{N}(0, \sigma_s^2)$ and that they are independent of the residual error. Note that $\boldsymbol{s} = (s_1, \ldots, s_n)^T$ is one random effect with n (unknown) levels. Using the notation in (4.1), we have that $\boldsymbol{\beta} = (\mu, \lambda_1, \lambda_2, \lambda_3, \lambda_4)^T$ and $\boldsymbol{X}$ is made of the n-times stacking (columnwise) of the 5×5 matrix

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{e}_4 & \boldsymbol{I}_4 \\ 1 & -\boldsymbol{e}_4^T \end{bmatrix} \tag{4.3}$$

with $\boldsymbol{e}_4$ a four-dimensional vector of ones (more generally, $\boldsymbol{e}_p$ will stand for a p-dimensional vector of ones). The matrix $\boldsymbol{x}$ actually defines a set of contrasts for

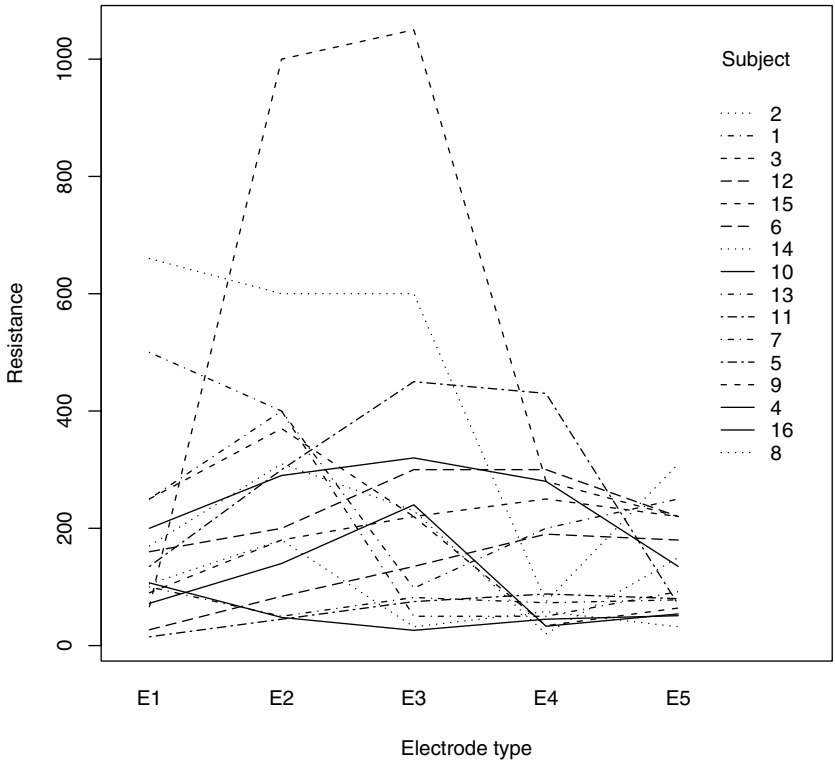

Figure 4.1 Profile plot for the skin resistance data.

the fixed effects parameters $\boldsymbol{\beta}$. The choice of these contrasts is arbitrary and mostly depends on the problem. They will be explained in more detail in Section 4.2.2.

Depending on the problem, there might be more than one different random effect (see examples in Sections 4.2.2 and 4.7.1), so that the latter are introduced in the MLM in a more general fashion. Let $\boldsymbol{\gamma}_j$, $j = 1, \ldots, r$ be a q_j-dimensional vector of random effects levels for the jth random effect. The latter can possibly be pre-multiplied by a design matrix $\boldsymbol{Z}_j$. Incorporating these random effects into (4.1) gives the general model formulation for a MLM

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sum_{j=1}^{r} \boldsymbol{Z}_j \boldsymbol{\gamma}_j + \boldsymbol{\epsilon}. \tag{4.4}$$

We usually assume that the q_j levels of each random effect $\boldsymbol{\gamma}_j$ are independent normal with zero mean and variance σ_j^2; each of the N random error terms in $\boldsymbol{\epsilon}$ is normal independent with zero mean and variance σ_ϵ^2; and $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_r$ and $\boldsymbol{\epsilon}$ are independent. Alternatively, one could bind all $\boldsymbol{Z}_j$ matrices together in a large matrix

$\boldsymbol{Z}$ and stack all vectors $\boldsymbol{\gamma}_j$ into a vector $\boldsymbol{\gamma}$ yielding the compact form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}. \tag{4.5}$$

Model (4.5) is similar in structure to the linear model (4.1) where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the design matrices for the fixed and random effects, respectively, and $\boldsymbol{\epsilon}$ is the vector of independent error term.

Under the MLM assumptions stated above, we have that

$$E[\boldsymbol{y}] = \boldsymbol{X}\boldsymbol{\beta} \tag{4.6}$$

and, if we put $\boldsymbol{Z}_0 = \boldsymbol{I}_N$, $\boldsymbol{\gamma}_0 = \boldsymbol{\epsilon}$ and $\sigma_\epsilon^2 = \sigma_0^2$ for convenience,

$$\text{var}(\boldsymbol{y}) = \sum_{j=0}^{r} \sigma_j^2 \boldsymbol{Z}_j \boldsymbol{Z}_j^T = \boldsymbol{V}. \tag{4.7}$$

To avoid problems of identifiability, we also assume that we have chosen a parametrization where the overall parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma_0^2, \ldots, \sigma_r^2)^T$ is identifiable. In many situations, one can split the response vector $\boldsymbol{y}$ into n independent clusters of observations (e.g. the different subjects taking part in an experiment) $\boldsymbol{y_i}$ for $i = 1, \ldots, n$. Then (4.6) and (4.7) become $E[\boldsymbol{y}_i] = \boldsymbol{x}_i\boldsymbol{\beta}$ and $\text{var}(\boldsymbol{y}_i) = \boldsymbol{\Sigma}_i$, $i = 1, \ldots, n$ with $\boldsymbol{y}_i$ the p_i-vector of observations, $\boldsymbol{x}_i$ the corresponding $(p_i \times (q+1))$ sub-matrix of $\boldsymbol{X}$, and $\boldsymbol{\Sigma}_i$ is a $p_i \times p_i$ matrix of the form

$$\boldsymbol{\Sigma}_i = \sum_{j=0}^{r} \sigma_j^2 [\boldsymbol{Z}_j \boldsymbol{Z}_j^T]_{(ii)}. \tag{4.8}$$

The quantity $[\boldsymbol{Z}_j \boldsymbol{Z}_j^T]_{(ii)}$ stands for the ith block-diagonal element of $\boldsymbol{Z}_j \boldsymbol{Z}_j^T$. Note that because of the independence assumption of the clusters or subjects $\boldsymbol{y}_i$ the covariance matrix $\boldsymbol{V}$ in (4.7) is block-diagonal, with diagonal elements given by (4.8). With the normality assumption for the random effect $\boldsymbol{\gamma}$ and the error $\boldsymbol{\epsilon}$ in (4.5), we then have that, at the cluster level,

$$\boldsymbol{y_i} \sim \mathcal{N}(\boldsymbol{x}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i), \tag{4.9}$$

where the variance matrix $\boldsymbol{\Sigma}_i$ is given by (4.8). The model is presented here under the 'conditional-independence' assumption, that is, all of the random terms are independent of each other. A more general formulation allowing for correlated errors and random effects is possible; see McLean *et al.* (1991) and Searle *et al.* (1992). It is usually done by assuming that the random component $\boldsymbol{\gamma}$ and random error $\boldsymbol{\epsilon}$ have respectively a variance $\boldsymbol{D}$ and $\boldsymbol{R}$ parametrized by a small number of parameters. This leads to a general variance structure $\text{var}(\boldsymbol{y}) = \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}^T + \boldsymbol{R}$. When a partition in n independent clusters is available, the same structure appears at the cluster level, yielding the MLM of Laird and Ware (1982). We use the simpler model (4.9) in this book as robustness theory for MLM has mainly been developed in that setting.

Before presenting classical and robust estimation and inference, and in order to give some examples of different models belonging to the general formulation given in (4.9), we study three different datasets in more detail that will also be used to illustrate the theory presented in this chapter.

4.2.2 Skin Resistance Data

Berry (1987) describes a dataset resulting from an experiment in which five types of electrode are applied to the arms of 16 subjects and their skin resistance measured. The experiment is conducted to assess whether the type of electrode can affect the response (see Berry, 1987). In this example, the electrode type is assumed to be a fixed effect and the subjects, which were randomly selected from some population, are considered as a random effect, i.e. their responses (the skin resistance) may vary across subjects although a common pattern per subject is expected. This experiment can be easily analyzed using a one factor within-subject ANOVA represented by (4.2) and the assumptions given below this equation. In this example researchers are primarily interested in testing the null hypothesis of 'no effect of the electrode type' on skin resistance, and a standard F-test is typically the default approach. To do this, reliable estimates of the fixed effects and the different variance components parameters (σ_s^2, σ_ϵ^2) are also needed.

As can be seen in Figure 4.1, one subject (case 15) clearly behaves differently than the other subjects, in that two measurements (type 2 and 3 electrodes) are much larger than the others. This raises the question of the reliability of the mean or contrast estimates and the corresponding F-test, a question that is discussed later.

The multivariate formulation is obtained by constructing n p-vectors of observations $\boldsymbol{y}_i$ for each subject (with here $p = 5$), so that

$$\boldsymbol{y}_i = \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{e}_5 s_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, 16. \tag{4.10}$$

Note that in this example we do not need to specify a different contrast matrix $\boldsymbol{x}_i$ for each subject as it is the same for all subjects. Also note that $\boldsymbol{x}$ given in (4.3) corresponds to the contrast matrix 'sum to zero contrast'. The name comes from the constraint $\sum_{j=1}^{5} \lambda_j = 0$ we are using. One could specify other types of contrast such as the 'treatment contrast'. With these contrasts, one of the groups j is chosen as the reference level and set to zero. For example, if the reference level is the first level of the factor electrode, then first one would set $\boldsymbol{\beta} = (\mu, \lambda_2, \lambda_3, \lambda_4, \lambda_5)^T$, and then the corresponding design matrix $\boldsymbol{x}$ using the 'treatment contrast' would be

$$\boldsymbol{x} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{0}_4^T \\ \boldsymbol{e}_4 & \boldsymbol{I}_4 \end{bmatrix}, \tag{4.11}$$

with $\mathbf{0}_p$ a p-dimensional vector of zeros. For the skin resistance data, the covariance matrix $\boldsymbol{\Sigma}$ is constant for all subjects and is defined as

$$\boldsymbol{\Sigma} = \text{var}(\boldsymbol{y}_i) = \text{var}(\boldsymbol{e}_5 s_i + \boldsymbol{\epsilon}_i) = \sigma_s^2 \boldsymbol{e}_5 \boldsymbol{e}_5^T + \sigma_\epsilon^2 \boldsymbol{I}_5 = \sigma_s^2 \boldsymbol{J}_5 + \sigma_\epsilon^2 \boldsymbol{I}_5, \tag{4.12}$$

with $\boldsymbol{J}_5$ being a 5×5 matrix of ones. In other words, $\boldsymbol{\Sigma}$ is the compound symmetric (or exchangeable) 5×5 matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_\varepsilon^2 + \sigma_s^2 & \sigma_s^2 & \sigma_s^2 & \sigma_s^2 & \sigma_s^2 \\ \sigma_s^2 & \sigma_\varepsilon^2 + \sigma_s^2 & \sigma_s^2 & \sigma_s^2 & \sigma_s^2 \\ \sigma_s^2 & \sigma_s^2 & \sigma_\varepsilon^2 + \sigma_s^2 & \sigma_s^2 & \sigma_s^2 \\ \sigma_s^2 & \sigma_s^2 & \sigma_s^2 & \sigma_\varepsilon^2 + \sigma_s^2 & \\ \sigma_s^2 & \sigma_s^2 & \sigma_s^2 & \sigma_s^2 & \sigma_\varepsilon^2 + \sigma_s^2 \end{bmatrix}.$$

Finally we can retrieve the general MLM formulation (4.5) by writing

$$\boldsymbol{y} = (\boldsymbol{e}_{16} \otimes \boldsymbol{x})\boldsymbol{\beta} + (\boldsymbol{I}_{16} \otimes \boldsymbol{e}_5)\boldsymbol{\gamma} + \boldsymbol{\epsilon} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon},$$

where $\boldsymbol{y}$ is a 16×5-vector of responses, $\boldsymbol{\epsilon}$ is a 16×5-vector of errors, and $\otimes$ is the Kronecker product.[1] We also have $\boldsymbol{\gamma} = \boldsymbol{\gamma}_1 = (s_1, \ldots, s_{16})^T$, $\boldsymbol{Z}_1 = \boldsymbol{I}_{16} \otimes \boldsymbol{e}_5$, so that $\boldsymbol{Z}_1\boldsymbol{Z}_1^T = (\boldsymbol{I}_{16} \otimes \boldsymbol{e}_5)(\boldsymbol{I}_{16} \otimes \boldsymbol{e}_5)^T$. It follows that $E[\boldsymbol{y}] = \boldsymbol{X}\boldsymbol{\beta}$, and that

$$\boldsymbol{V} = \text{var}(\boldsymbol{y}) = (\boldsymbol{I}_{16} \otimes \boldsymbol{e}_5)\sigma_s^2\boldsymbol{I}_5(\boldsymbol{I}_{16} \otimes \boldsymbol{e}_5)^T + \sigma_\epsilon^2\boldsymbol{I}_{80} \tag{4.13}$$

$$= \sigma_s^2(\boldsymbol{I}_{16} \otimes \boldsymbol{J}_5) + \sigma_\epsilon^2(\boldsymbol{I}_{16} \otimes \boldsymbol{I}_5) = \boldsymbol{I}_{16} \otimes (\sigma_s^2\boldsymbol{J}_5 + \sigma_\epsilon^2\boldsymbol{I}_5). \tag{4.14}$$

The expression of $\boldsymbol{V}$ is hence simply a block-diagonal matrix with block element $\boldsymbol{\Sigma}$ repeated 16 times.

4.2.3 Semantic Priming Data

The study of semantic and associative priming in picture naming is well known in psychology (see e.g. Alario and Ferrand, 2000; Holcomb and McPherson, 1994). The data we have come from an experiment in which 21 subjects had to decide as quickly as possible whether a target (object's drawing), which appeared after a prime (action of a pantomime), was a real object or not. The delay between the pantomime and the showing of the object was either short or long and the pantomime was either related, neutral or unrelated. Several different real objects were used. In psychology, this type of experiment is performed in order to study the effect of a prime in naming objects known as 'semantic and associative priming in picture naming'. For each combination of real object and type of prime, five measures (time to decide whether the object was real or not) were taken on each subject, of which the first one (trial) and the errors (wrong object decision) were discarded and the means of the remaining were taken as the response variable. The primary hypothesis is that the reaction time changes when a link between the priming and the object (i.e. in the related item) exists. Delay between the prime and the object is also assumed to affect the response. The data were collected at the University of Geneva (see Moy and Mounoud, 2003). We consider here a subsample involving the object 'broom', with 21 elderly subjects

[1] The Kronecker product between an $m \times p$ matrix $\boldsymbol{A}$ (with elements a_{ij}) and a $q \times r$ matrix $\boldsymbol{B}$ yields the $m \cdot q \times p \cdot r$ matrix $[a_{ij}\boldsymbol{B}]_{i=1,\ldots,m,\, j=1,\ldots,p}$.

(aged 70 and over). A two-way ANOVA model with repeated measures can be fitted to these data, model given by

$$y_{ijk} = \mu + \lambda_j + \gamma_k + (\lambda\gamma)_{jk} + s_i + (\lambda s)_{ij} + (\gamma s)_{ik} + \epsilon_{ijk}, \tag{4.15}$$

with λ_j, $j = 1, 2$ the fixed effect for the factor delay, $\gamma_k, k = 1, 2, 3$ the fixed effect for the factor condition (i.e. if the pantomime is either related, neutral or unrelated) with $\sum_j \lambda_j = 0$, $\sum_k \gamma_k = 0$, $(\lambda\gamma)_{jk}$, $j = 1, 2$, $k = 1, 2, 3$ the fixed effect for the interaction between the two factors, with $\sum_j \sum_k (\lambda\gamma)_{jk} = 0$. The independent normal random effects are given by the subject effect, s_i, the interaction between the subject effect and the pantomime type $(\lambda s)_{ij}$, the interaction between the subject effect and the factor delay $(\gamma s)_{ik}$ and ϵ_{ijk}. Here an interaction term is added to the model because it is assumed that the difference in reaction time between a short and a long delay depends on the type of object that is shown as a prime (related or not). Assuming the independence between all random effects and that the responses are ordered as $y_{i11}, y_{i12}, y_{i13}, y_{i21}, \ldots, y_{i23}, i = 1, \ldots, 21$, model (4.15) can be written as $\boldsymbol{y}_i \sim \mathcal{N}(\boldsymbol{x}\boldsymbol{\beta}, \boldsymbol{\Sigma})$, a multivariate normal model of dimension $2 \times 3 = 6$ where

$$\boldsymbol{\beta} = (\mu, \lambda_1, \gamma_1, \gamma_2, (\lambda\gamma)_{11}, (\lambda\gamma)_{12})^T,$$

$$\boldsymbol{x} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \end{bmatrix}.$$

If we express (4.15) through a multivariate formulation (as in (4.10)), we can show that, in this example, the subject covariance matrix $\boldsymbol{\Sigma}$ is

$$\boldsymbol{\Sigma} = \sigma^2_\epsilon \boldsymbol{I}_6 + \sigma^2_s \boldsymbol{J}_6 + \sigma^2_{\lambda s}(\boldsymbol{I}_2 \otimes \boldsymbol{J}_3) + \sigma^2_{\gamma s}(\boldsymbol{J}_2 \otimes \boldsymbol{I}_3). \tag{4.16}$$

As before the overall matrix $\boldsymbol{V}$ in (4.7) is block-diagonal with block element $\boldsymbol{\Sigma}$ repeated 21 times.

4.2.4 Orthodontic Growth Data

These data come from an orthodontic growth study where a set of different measurements were collected from X-rays of 27 children's skulls (16 males and 11 females). The response variable is the distance in millimeters between the pituitary and the pterygomaxillary fissure, two points that can be easily located on the X-rays. The distance was measured at 8, 10, 12 and 14 years of age for each child. These data were originally reported by Potthoff and Roy (1964) and subsequently analyzed by several authors. A preliminary profile plot given in Figure 4.2 shows that the distance grows linearly with age, each participant having their own intercept and slope. The within-subject variability seems also slightly larger for boys than for girls. This features conducted Pinheiro *et al.* (2001) to use the following model for

this data

$$y_{ijt} = \beta_0 + \beta_1 t + (\beta_{0g} + \beta_{1g} t) J_i(j) + \gamma_{0i} + \gamma_{1i} t + \epsilon_{ijt} \tag{4.17}$$

with y_{ijt} the response for the ith subject ($i = 1, \ldots, 27$) of gender j ($j = 1$ for boys and $j = 2$ for girls) at age $t = 8, 10, 12, 14$, and

$$J_i(j) = \begin{cases} 0 & j = 1, \\ 1 & j = 2, \end{cases}$$

a dummy variable for gender. The vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_{0g}, \beta_{1g})^T$ is the fixed effects parameter, and $\gamma_{0i}, \gamma_{1i}, \epsilon_{ijt}$ are the random effects levels for the ith observation. All random effects are independent normal with zero mean and respective variance $\sigma^2_{\gamma_0}, \sigma^2_{\gamma_1}$ and σ^2_ϵ. This model is technically a random slope and intercept model. Each subject's line varies around the group average line, $y = \beta_0 + \beta_1 t$ for the boys and $y = (\beta_0 + \beta_{0g}) + (\beta_1 + \beta_{1g})t$ for the girls. The multivariate formulation of model (4.17) is achieved by writing the mean vector $\boldsymbol{\mu}_{j(i)} = \boldsymbol{x}_{j(i)}\boldsymbol{\beta}$, where the subject design matrix is $\boldsymbol{x}_{j(i)} = (\boldsymbol{e}_4, \boldsymbol{e}_4 J_i(j), \boldsymbol{t}, \boldsymbol{t} J_i(j))$ and $\boldsymbol{t}$ is the common age vector $\boldsymbol{t} = (8, 10, 12, 14)^T$. Note that because the subject is nested in gender (a subject can be either a boy or a girl), the design matrix is then different between boys and girls. The covariance matrix at the cluster level (child) is $\boldsymbol{\Sigma} = \sigma^2_{\gamma_0} \boldsymbol{J}_4 + \sigma^2_{\gamma_1} \boldsymbol{t}\boldsymbol{t}^T + \sigma^2_\epsilon \boldsymbol{I}_4$.

While the traditional analysis relies on normality assumptions for all random components, Pinheiro *et al.* (2001) identify a few outliers in the data and propose to use a different estimator based on the multivariate t-distribution (see also Lange *et al.*, 1989; Welsh and Richardson, 1997).

4.3 Classical Estimation and Inference

4.3.1 Marginal and REML Estimation

Two classical techniques are used to estimate MLMs, namely the MLE and the REML. Denote by $\boldsymbol{\alpha} = (\sigma^2_0, \ldots, \sigma^2_r)^T$ the vector of all variance parameters and by $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)^T$ the overall parameter. The likelihood for n observations $\boldsymbol{y}_i$ with model (4.9) is given by

$$L(\boldsymbol{\theta} \mid \boldsymbol{y}) = (2\pi)^{-1/(2n)} |\boldsymbol{\Sigma}_i|^{-1/2} \prod_{i=1}^{n} \exp\{(\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta})^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta})\}, \tag{4.18}$$

where $|\boldsymbol{\Sigma}_i|$ denotes the determinant of $\boldsymbol{\Sigma}_i$. The MLE, $\hat{\boldsymbol{\theta}}_{[MLE]}$, maximizes (4.18) or, equivalently, solves for the fixed effects parameter $\boldsymbol{\beta}$ and variance components σ^2_j respectively

$$\sum_{i=1}^{n} \boldsymbol{x}_i^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta}) = 0 \tag{4.19}$$

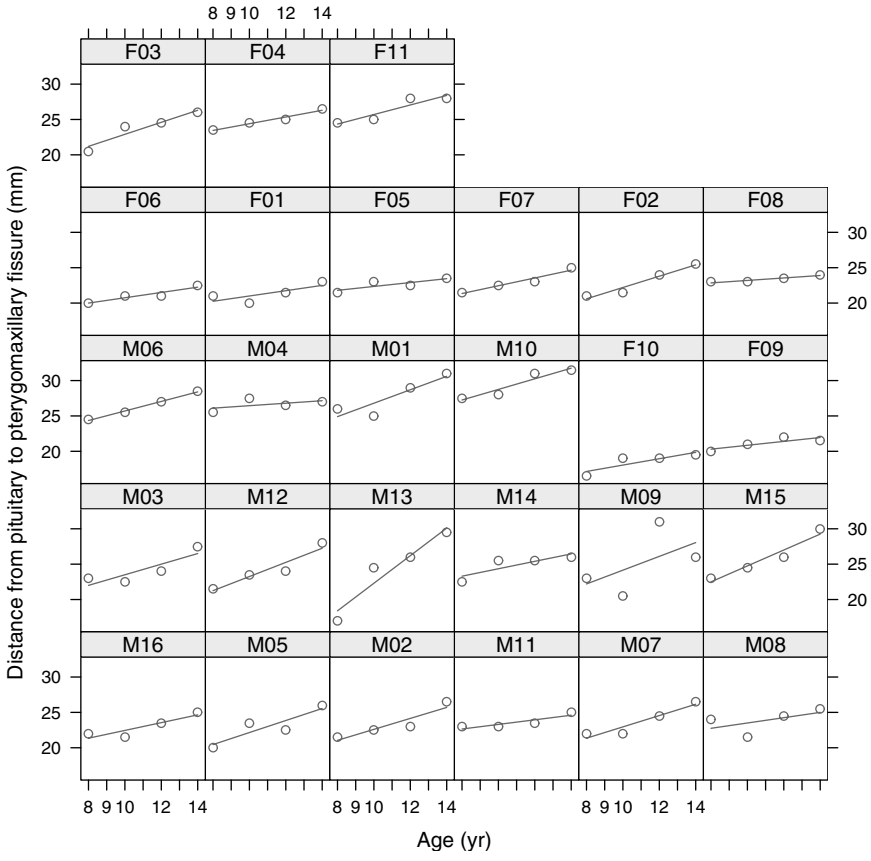

Figure 4.2 Orthodontic growth patterns in 16 boys (M) and 11 girls (F).

and

$$\sum_{i=1}^{n}\{(\boldsymbol{y}_i-\boldsymbol{x}_i\boldsymbol{\beta})^T\boldsymbol{\Sigma}_i^{-1}[\boldsymbol{Z}_j\boldsymbol{Z}_j^T]_{(ii)}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i-\boldsymbol{x}_i\boldsymbol{\beta})-\mathrm{tr}(\boldsymbol{\Sigma}_i^{-1}[\boldsymbol{Z}_j\boldsymbol{Z}_j^T]_{(ii)})\}=0 \quad (4.20)$$

for $j=0,\ldots,r$, where $[\boldsymbol{Z}_j\boldsymbol{Z}_j^T]_{(ii)}$ again stands for the ith block-diagonal element of $\boldsymbol{Z}_j\boldsymbol{Z}_j^T$.

To solve this system of equations, let us first assume that $\boldsymbol{\alpha}$ is known. Then, only (4.19) is necessary, yielding

$$\boldsymbol{\beta}(\boldsymbol{\alpha})=\left(\sum_{i=1}^{n}\boldsymbol{x}_i^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{x}_i\right)^{-1}\sum\boldsymbol{x}_i^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{y}_i=(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y}. \quad (4.21)$$

When $\boldsymbol{\alpha}$ is unknown, but an estimate $\hat{\boldsymbol{\alpha}}$ is available, one can simply estimate $\boldsymbol{\beta}$ by replacing $\boldsymbol{\Sigma}_i$ by $\widehat{\boldsymbol{\Sigma}}_i(\hat{\boldsymbol{\alpha}})$ in (4.21). A common choice for $\hat{\boldsymbol{\alpha}}$ is the MLE, denoted by

$\hat{\boldsymbol{\alpha}}_{[MLE]}$, typically obtained by solving (4.20) for $j = 0, 1, \ldots, r$, after $\boldsymbol{\beta}$ is replaced by its expression (4.21). The resulting estimate is the MLE for the fixed effects parameter, $\hat{\boldsymbol{\beta}}_{[MLE]}$ which follows asymptotically a multivariate normal distribution with mean $\boldsymbol{\beta} := \boldsymbol{\beta}(\boldsymbol{\alpha})$ and covariance matrix

$$\text{var}(\hat{\boldsymbol{\beta}}_{[MLE]}) = \left(\sum_{i=1}^{n} \boldsymbol{x}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{x}_i \right)^{-1} = (\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1}. \tag{4.22}$$

In practice, the covariance matrix of $\hat{\boldsymbol{\beta}}_{[MLE]}$ has to be estimated by replacing $\boldsymbol{\Sigma}_i$ by $\widehat{\boldsymbol{\Sigma}}_i$, itself obtained by substituting $\boldsymbol{\alpha}$ by $\hat{\boldsymbol{\alpha}}_{[MLE]}$, in (4.22). Under regularity conditions, one can show that $\hat{\boldsymbol{\alpha}}_{[MLE]}$ and $\hat{\boldsymbol{\beta}}_{[MLE]}$ are asymptotically uncorrelated and their asymptotic covariances being given by the inverse of their respective Fisher information matrix; see, for instance, Searle *et al.* (1992, pp. 238–240). Several methods for the actual computation of the MLE have been proposed in the literature. Nowadays, Newton–Raphson procedures or clever implementations of the EM algorithm are used to estimate all parameters in the model; see Lindstrom and Bates (1988) and Searle *et al.* (1992, Chapter 8) for details.

Although the MLE is asymptotically efficient when the normality assumptions are met, the variance components MLE is only asymptotically unbiased. In small samples, the (finite) bias of $\hat{\boldsymbol{\alpha}}_{[MLE]}$ can be large and become even larger when q increases. To overcome this problem, Patterson and Thompson (1971) and Harville (1977) introduce the REML of $\boldsymbol{\alpha}$ from the MLE of independent contrasts of the data, i.e. variables $\boldsymbol{L}\boldsymbol{y}$ where $\boldsymbol{L}$ is any $(N - q - 1) \times N$ matrix of full rank satisfying $\boldsymbol{L}\boldsymbol{X} = 0$. The choice of $\boldsymbol{L}$ is unimportant as the log-likelihood never differs from

$$L_R(\boldsymbol{\alpha}|\boldsymbol{y}) = -\tfrac{1}{2}\{(\boldsymbol{L}\boldsymbol{y}^T \boldsymbol{V}^{-1} \boldsymbol{L}\boldsymbol{y} + \log|\boldsymbol{V}| + \log|\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X}|\}, \tag{4.23}$$

by more than a constant. Then $\hat{\boldsymbol{\alpha}}_{[REML]}$ is obtained by maximizing (4.23) or, equivalently, by solving for (the elements of) $\boldsymbol{\alpha}$ the first-order equations $\partial/\partial\sigma_j^2 L_R(\boldsymbol{\alpha}|\boldsymbol{y}) = 0$ for $j = 0, \ldots, r$. This can be rewritten after algebraic manipulations as

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{V}^{-1} \boldsymbol{Z}_j \boldsymbol{Z}_j^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) - \text{tr}(\boldsymbol{P}\boldsymbol{Z}_j \boldsymbol{Z}_j^T) = 0, \quad j = 0, \ldots, r, \tag{4.24}$$

where $\boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1}$; see Harville (1977) for details. The matrix $\boldsymbol{P}$ is not block-diagonal, so (4.24) cannot be rewritten as a sum over independent subvectors. The other difference with (4.20) lies in the trace term that includes $\boldsymbol{P}$ instead of a block-diagonal element of $\boldsymbol{V}$. The REML estimator of the fixed effects, $\hat{\boldsymbol{\beta}}_{[REML]}$, is given by the formula (4.21) with $\boldsymbol{\Sigma}_i$ being replaced by $\widehat{\boldsymbol{\Sigma}}_i(\hat{\boldsymbol{\alpha}}_{[REML]})$, its REML estimate using the relationship (4.8). Under mild conditions given by Cressie and Lahiri (1993) and Richardson and Welsh (1994), $(\hat{\boldsymbol{\beta}}_{[REML]}^T, \hat{\boldsymbol{\alpha}}_{[REML]}^T)^T$ is asymptotically normally distributed with asymptotic variance given by the inverse of the Fisher information matrix. The two off-diagonal blocks in this matrix are equal to zero, which proves that the REML fixed effects and variance estimates are asymptotically independent. Like for the MLE, $\text{var}(\hat{\boldsymbol{\beta}}_{[REML]})$ is estimated by (4.22) where, conventionally, $\boldsymbol{\Sigma}_i$ is replaced by its REML estimate.

4.3.2 Classical Inference

In this section we primarily consider methods for making inference about the mixed effect parameters as the variance component $\boldsymbol{\alpha}$ is often considered as a nuisance parameter. Given that point estimates and their standard errors are available for both the MLE and REML, one can easily test a single hypothesis, say $H_0 : \beta_j = 0$, by computing $z = \hat{\beta}_j / SE(\hat{\beta}_j)$, where $SE(\hat{\beta}_j)$ is the (estimated) standard error of $\hat{\beta}_j$, and compare it with a normal distribution. To simplify notation, we omit the subscript [MLE] or [REML] and we only specify dependence on the method if necessary. The z-statistic corresponds to a special case of the Wald test statistics for a null hypothesis[2] of the type $H_0 : \boldsymbol{L\beta} = 0$, given by

$$W^2 = (\boldsymbol{L}\hat{\boldsymbol{\beta}})^T [\boldsymbol{L}\widehat{\boldsymbol{V}}\boldsymbol{L}^T]^{-1} \boldsymbol{L}\hat{\boldsymbol{\beta}}, \tag{4.25}$$

where $\widehat{\boldsymbol{V}} = (\sum_{i=1}^{n} \boldsymbol{x}_i^T \widehat{\boldsymbol{\Sigma}}_i^{-1} \boldsymbol{x}_i)^{-1}$ is the estimated variance–covariance matrix of the fixed effects estimate, and $\boldsymbol{L}$ is a contrast matrix.[3] Under the null hypothesis, W^2 is asymptotically distributed as a χ^2 distribution with rank($\boldsymbol{L}$) degrees of freedom. As noted by Dempster *et al.* (1981), the Wald test statistic does not take into account the uncertainty in the estimation of the variance components in $\widehat{\boldsymbol{V}}$, and as a result the estimated standard errors are too small in small samples. One possible way to correct for this is to use approximate F-statistic,[4] e.g.

$$F = W^2 / \text{rank}(\boldsymbol{L}), \tag{4.26}$$

the numerator degrees of freedom being rank($\boldsymbol{L}$). In general, (4.26) is not directly related to any particular ANOVA F-statistic but it seems reasonable to use the F distribution as an approximation. The denominator degrees of freedom have to be estimated from the data and several procedures have been proposed, among which, the Satterthwaite approximation (Satterthwaite, 1941) is the most commonly used; see Verbeke and Molenberghs (1997) and Verbeke and Molenberghs (2000, p. 57). Another is to use the scaled Wald test statistics of Kenward and Roger (1997) based on an adjusted variance matrix. Its small sample distribution was found to be well approximated by a F distribution using Satterthwaite's method.

The LRT test based on the MLE, as defined in Section 2.5.1, is also available for testing canonical hypotheses in mixed models. It must be stressed though that such a test is not possible with REML for hypotheses on fixed effects. Indeed, the mean structure fitted under the null $H_0 : \boldsymbol{\beta}_{(2)} = 0$ is not the same as that fitted with the full model, leading to different contrasts. Therefore, the two restricted likelihoods are based on different observations making them non-comparable. It is also possible to use the score test to test $H_0 : \boldsymbol{\beta}_{(2)} = 0$ but this test is apparently not commonly used for that model.

[2] We use a slightly more general formulation than the canonical presentation of Section 2.5.1 as with MLMs, it is more convenient to test contrasts.

[3] Note that although we use the same notation, $\boldsymbol{L}$ here is not necessarily the same as the contrast matrix used to define the REML.

[4] An exact F-test exists in some cases, e.g. balanced ANOVA models; see Littell (2002, p. 482).

The main focus of inference in the mixed model is probably hypotheses about the fixed effects; however, in some applications, such as in genetics, the variance components are of direct interest. In general, over-parametrization of the variance structure also leads to inefficient and potentially poor assessment of standard errors of the fixed effects parameters. It is therefore important to propose valid model-based inference for the variance parameters. In principle, we know from the classical theory, that $\hat{\boldsymbol{\alpha}}_{[MLE]}$ is asymptotically normally with (asymptotic) covariance matrix given by the inverse of the Fisher information matrix. So Wald tests or LRT tests could be used to test restrictions to the variance parameters. However, the problem is made complicated by the fact that null hypotheses of interest typically involve constraints of the type $\sigma_j^2 = 0$, i.e. the parameter of interest lies on the boundary of the parameter space. In this situation, regularity conditions required for the asymptotic distribution to be valid are not met. As a result, under H_0, the normal approximation for $\hat{\boldsymbol{\alpha}}_{[MLE]}$ fails and the χ^2 distribution for LRT or Wald tests is no longer valid. The distribution of the small variance components when the true parameter are not on the boundary but are close to it can also be affected; see Stern and Welsh (1998). This situation is known in the literature as non-standard asymptotics. However, Stram and Lee (1994) were able to prove that, when the number of fixed effects remain constant, the LRT test statistic is often distributed as a mixture of χ^2 distributions. This work is based on the theory by Self and Liang (1987) and assumes conditional independence of the error term in the model, i.e. $\text{var}(\boldsymbol{\epsilon}) = \boldsymbol{I}_N$ in (4.5).

This result is better explained through an example. Consider the orthodontic growth data where the working mixed model only includes a random intercept (model (4.17) without $\gamma_{1i}t$). We are interested in testing whether adding a random slope (for time) effect is necessary to capture a possible increase of variance over time. To illustrate the exact same situation described in Stram and Lee (1994), we add a non-null correlation between the two random effects by assuming that a covariance $\sigma_{\gamma_{01}}$ between γ_{0i} and γ_{1i} exists in model (4.17). The null hypothesis is then $H_0\colon \sigma_{\gamma_1}^2 = 0, \sigma_{\gamma_{01}} = 0$. In terms of covariance structure, we are testing that the covariance matrix for $(\gamma_{0i}, \gamma_{1i})^T$ changes from

$$\boldsymbol{D} = \begin{bmatrix} \sigma_{\gamma_0}^2 & 0 \\ 0 & 0 \end{bmatrix}$$

to the alternative H_1:

$$\boldsymbol{D} = \begin{bmatrix} \sigma_{\gamma_0}^2 & \sigma_{\gamma_{01}} \\ \sigma_{\gamma_{01}} & \sigma_{\gamma_1}^2 \end{bmatrix},$$

with $\sigma_{\gamma_1}^2 > 0$ to guarantee that D is positive-definite. As two additional parameters $\sigma_{\gamma_{01}}$ and $\sigma_{\gamma_1}^2$ have been added to the model, a naive application of the classical theory would compare the corresponding LRT test with a χ_2^2 distribution. The exact theory states that a mixture with equal weights 0.5 for χ_1^2 and χ_2^2 must be used. Therefore, a naive analysis could lead to larger p-values and, hence, acceptance of oversimplified variance structures. This result also holds for the REML-based LRT (Morrell, 1998).

Table 4.1 Estimates and standard errors for the REML for the skin resistance data using model (4.2)–(4.3) with and without observation 15.

	REML		REML without observation 15	
Parameter	Estimate (*SE*)	*p*-value	Estimate (*SE*)	*p*-value
μ	2.030 (0.341)	$<10^{-4}$	1.817 (0.284)	$<10^{-4}$
λ_1	−0.213 (0.334)	0.525	0.076 (0.246)	0.756
λ_2	0.842 (0.334)	0.014	0.580 (0.246)	0.221
λ_3	0.549 (0.334)	0.105	0.234 (0.246)	0.345
λ_4	−0.526 (0.334)	0.120	−0.399 (0.246)	0.110
σ_s	1.190		0.994	
σ_ϵ	1.495		1.068	

As it is more accurate with a small sample we use this variant on the orthodontic growth data. The LRT statistic for testing $H_0 : \sigma^2_{\gamma_1} = 0, \sigma_{\gamma_{01}} = 0$ returns a value of $2(-216.3 + 216.9) = 1.2$. A correct p-value is therefore $p = 0.5 \times P(\chi^2_2 > 1.2) + 0.5 \times P(\chi^2_1 > 1.2) = 0.41$, whereas the naive calculation yields $p = 0.55$. In this case, both procedures conclude that a second random effect is probably not necessary (assuming that no robustness issue arises here).

Stram and Lee (1994) also consider the case of testing k versus $k + 1$ random effects. In that case, a mixture with equal weights 0.5 for χ^2_k and χ^2_{k+1} is obtained for the asymptotic distribution. A more complex mixture is also available when $l > 1$ random effects are added to the model but requires complex calculations. Again, extensions of these results to LRT tests based on the REML is possible (see Morrell, 1998). Recent work by Scheipl *et al.* (2008) show that they are generally more powerful and should therefore be preferably used. The Wald test and classical confidence intervals for the variance parameter must also be corrected bearing in mind that they are generally outperformed by their LRT counterparts. Finally, a good account of these problems with applications can be found in Verbeke and Molenberghs (2000, pp. 64–74).

4.3.3 Lack of Robustness of Classical Procedures

To illustrate the sensitivity of the classical estimators introduced in Section 4.3.1, let us go back to the skin resistance data. In Figure 4.1 we saw that, out of the 80 readings, two measurements (resistance of electrodes of type 2 and 3) taken on subject 15 were much larger than the others. The experimenter discovered later that the reason for these two rather large readings was the excessive amount of hair on the subject's arm (see Berry, 1987). Table 4.1 presents the classical (REML) estimates and standard errors with and without case 15.[5]

One may notice that there is considerable variation in the estimates of the different electrode types (significant fixed effects) when observation 15 is present in the data.

[5]The raw data have been divided by 100.

These differences are less obvious when case 15 is removed from the data. Also a large difference is observed in the residual error variance estimate $\hat{\sigma}_\epsilon^2$ when it is computed with and without case 15. This clearly illustrates the lack of robustness of the REML.

To quantify in a more formal way the sensitivity of the MLE and REML, we use the *IF* (see Section 2.2.1) which offers an elegant way to justify theoretically these empirical findings. Indeed, both the MLE and REML are *M*-estimators defined through estimating (4.19)–(4.20) and (4.19)–(4.24). Their *IF* is therefore proportional to their defining Ψ-functions. Specifically, the influence of the ith independent cluster (i.e. the four measurements of the ith subject in the skin resistance experiment) for both classical estimators of $\boldsymbol{\beta}$ is proportional to the score function for that parameter

$$s(\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\theta}) = \boldsymbol{x}_i^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{x}_i \boldsymbol{\beta}). \tag{4.27}$$

This quantity is unbounded in $\boldsymbol{y}_i$ and in $\boldsymbol{x}_i$, which proves theoretically that both the MLE and REML estimates for the fixed effects are not robust. The situation is even worse for the variance component. The *IF* of $\hat{\boldsymbol{\alpha}}_{[MLE]}$ is proportional to the summand in (4.20), a quadratic form of $\boldsymbol{y_i}$ and, as a result, a single abnormal response (such as case 15's readings for type 2 and 3 electrodes) can ruin $\hat{\boldsymbol{\alpha}}_{[MLE]}$. It is not possible to assess directly the effect of a single cluster on the REML variance estimates as the estimating equation cannot be defined at that level. However, a quadratic form appears in the left-hand side of (4.24), proving that $\hat{\boldsymbol{\alpha}}_{[REML]}$ is just as sensitive as $\hat{\boldsymbol{\alpha}}_{[MLE]}$ to contamination.

4.4 Robust Estimation

4.4.1 Bounded Influence Estimators

It is possible to extend the bounded-influence approach of Section 2.3.2 to MLMs. Most of these methods are based on a weighted version of the likelihood, either directly (Huggins, 1993; Huggins and Staudte, 1994) where a robustified likelihood is maximized, or through a weighted score equation (Richardson and Welsh, 1995; Richardson, 1997; Stahel and Welsh, 1997). Summarizing the previous work, Welsh and Richardson (1997) introduce a very general class that encompasses most of the previous proposals through

$$\sum_{i=1}^{n} \boldsymbol{x}_i^T \boldsymbol{W}_{0i} \boldsymbol{\Sigma}_i^{-1/2} \psi_{0i}(\boldsymbol{\Sigma}_i^{-1/2} \boldsymbol{U}_{0i}(\boldsymbol{y}_i - \boldsymbol{x}_i \boldsymbol{\beta})) = 0 \tag{4.28}$$

for the fixed effects, and

$$\frac{1}{2} \sum_{i=1}^{n} \{\psi_{1i}(\boldsymbol{\Sigma}_i^{-1/2} \boldsymbol{U}_{1i}(\boldsymbol{y}_i - \boldsymbol{x}_i \boldsymbol{\beta}))^T \boldsymbol{W}_{1i} \boldsymbol{\Sigma}_i^{-1/2} [\boldsymbol{Z}_j \boldsymbol{Z}_j^T]_{(ii)} \boldsymbol{\Sigma}_i^{-1/2} \boldsymbol{W}_{1i}$$
$$\cdot \psi_{2i}(\boldsymbol{\Sigma}_i^{-1/2} \boldsymbol{U}_{1i}(\boldsymbol{y}_i - \boldsymbol{x}_i \boldsymbol{\beta})) - \mathrm{tr}(\boldsymbol{K}_{2i} \boldsymbol{\Sigma}_i^{-1} [\boldsymbol{Z}_j \boldsymbol{Z}_j^T]_{(ii)})\} = 0 \tag{4.29}$$

for each variance component σ_j^2. The matrices $\boldsymbol{K}_{2i}$ are needed to ensure consistency at the normal model; see Welsh and Richardson (1997) for details. Equations (4.28) and (4.29) generalize the score equations (4.19) and (4.20) for the MLE. The choice of the weight matrices $\boldsymbol{W}_{0i}$, $\boldsymbol{W}_{1i}$, $\boldsymbol{U}_{0i}$, $\boldsymbol{U}_{1i}$ and functions ψ_{0i}, ψ_{1i}, ψ_{2i} defines each particular estimator including Huggins' earlier proposals. The ψ functions are typically chosen as Huber functions applied to all components but other choices are also possible. The robust estimator with all weights equal to one and $\psi_0 = \psi_1 = \psi_2$ is called robust MLE II in Richardson and Welsh (1995), as (4.29) is analogous to Huber's Proposal 2 in linear regression. Likewise, the choice $\psi_0 = \psi_2$ and $\psi_1(z) = z$ gives the robust MLE I of Richardson and Welsh (1995). It is also possible to define robust versions of the REML by using similar weighted equations to (4.29), the difference being a more complex trace term.[6] As before two variants exist and are called robust REML I and II in Richardson and Welsh (1995) and Welsh and Richardson (1997). As all the proposals discussed here are defined through estimating equations of the type $\sum \Psi(\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)^T$, the general asymptotic theory for M-estimators applies. Although these developments generalize the bounded-influence approach of Section 3.2.4 in a considerable level of generality, several limitations can be mentioned. First, computation is generally complicated by the presence of complex matrices $\boldsymbol{K}_{2i}$ required for consistency. The problem may even become intractable for redescending Ψ or complex variance structures. Second, in the presence of contaminated data, some small residual bias to the robust variance estimates remains even for the robust REML proposals; see the simulation results in Richardson and Welsh (1995, pp. 1437–1438). Finally, the breakdown point of such bounded influence estimators can be low and this may be an issue in complex models.

4.4.2 *S*-estimators

The reformulation of the MLM as a multivariate normal model offers an elegant way to tackle the robustification problem. Specifically, S-estimators introduced earlier in Section 2.3.3 for their good breakdown properties can easily be generalized to balanced MLMs , i.e. models of type (4.8)–(4.9) where the cluster size $p_i = p$ and $\boldsymbol{\Sigma_i} = \boldsymbol{\Sigma}$ for all clusters. This assumption is certainly not desirable from a practical perspective as the number of applications involved unbalanced data or variable repeated measures over time. As this theory is new (Copt and Victoria-Feser, 2006), there is however hope that this limitation will be relaxed in the near future.

In the multivariate normal setting, one can define an S-estimator for the mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ as the solution for these parameters that minimizes $\det(\boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|$ subject to

$$n^{-1} \sum_{i=1}^{n} \rho(d_i) = b_0, \tag{4.30}$$

[6]The equation is similar to (4.29) with the trace term $\mathrm{tr}(\boldsymbol{K}_2 \boldsymbol{P} \boldsymbol{Z}_j \boldsymbol{Z}_j^T)$ where $\boldsymbol{K}_2 = \mathrm{diag}(\boldsymbol{K}_{21}, \ldots, \boldsymbol{K}_{2n})$ sitting outside the summation for all i.

where

$$d_i^2 = (\boldsymbol{y}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}) \tag{4.31}$$

are the Mahalanobis distances, ρ is a bounded function and $b_0 = E_\Phi[\rho(d)]$ ensures consistency at the normal model. Using the relationship (2.33), the tuning parameter of the ρ-function can be chosen to achieve a pre-specified breakdown point ε^* (see Section 2.3.3). A typical choice for ρ is Tukey's biweight given in (2.20). For the balanced case, the marginal MLM (4.9) simply becomes $\boldsymbol{y}_i \sim \mathcal{N}(\boldsymbol{x}_i\boldsymbol{\beta}; \boldsymbol{\Sigma})$ where the common covariance matrix is

$$\boldsymbol{\Sigma} = \sum_{j=0}^{r} \sigma_j^2 \boldsymbol{z}_j \boldsymbol{z}_j^T \tag{4.32}$$

and $\boldsymbol{z}_j$ is the (common) element of the design matrix $\boldsymbol{Z}_j$ for a particular cluster. In the skin resistance data example, $\boldsymbol{\Sigma}$ is given by (4.32) (see also (4.12)), with $\boldsymbol{z}_0\boldsymbol{z}_0^T = \boldsymbol{I}_5$ (for the residual variance) and $\boldsymbol{z}_1\boldsymbol{z}_1^T = \boldsymbol{e}_5\boldsymbol{e}_5^T$ (for the subject random effect variance). Likewise, for the semantic priming data example, according to (4.16), we have that $\boldsymbol{z}_0\boldsymbol{z}_0^T = \boldsymbol{I}_6$ (for the residual variance), $\boldsymbol{z}_1\boldsymbol{z}_1^T = \boldsymbol{J}_6$, $\boldsymbol{z}_2\boldsymbol{z}_2^T = \boldsymbol{I}_2 \otimes \boldsymbol{J}_3$ and $\boldsymbol{z}_3\boldsymbol{z}_3^T = \boldsymbol{J}_2 \otimes \boldsymbol{I}_3$ for the subject and its factors' interactions random effects variances.

The additional structure on the mean and covariance matrix implied by the MLM formulation does not create additional difficulties to extend the definition of an S-estimator to that setting. Indeed, it can be defined as the solution for $\boldsymbol{\beta}$, σ_j^2, $j = 0, \ldots, r$ of the same minimization problem under the constraint (4.30), with

$$d_i = d_i(\boldsymbol{\beta}) = \sqrt{(\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta})} \tag{4.33}$$

and $\boldsymbol{\Sigma}$ having the particular structure (4.32). The problem can be restated as solving the estimating equations

$$\sum w(d_i) \boldsymbol{x}_i^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta}) = \sum \Psi_{\boldsymbol{\beta}}(\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\theta}) = 0, \quad j = 0, \ldots, r \tag{4.34}$$

for $\boldsymbol{\beta}$, and

$$\begin{aligned} &\sum \{p w(d_i) (\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta})^T \boldsymbol{\Sigma}^{-1} \boldsymbol{z}_j \boldsymbol{z}_j^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta}) - w(d_i) d_i^2 \operatorname{tr}[\boldsymbol{\Sigma}^{-1} \boldsymbol{z}_j \boldsymbol{z}_j^T]\} \\ &\quad = \sum \Psi_{\sigma_j^2}(\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\theta}) = 0, \end{aligned} \tag{4.35}$$

for the variance component $\boldsymbol{\alpha} = (\sigma_0^2, \ldots, \sigma_r^2)^T$ (see Copt and Victoria-Feser, 2006). Here $w(d) = (\partial/\partial d)\rho(d)/d$ is the robust weight given to each observation. Equations (4.34) and (4.35) can be rewritten in a more compact form as

$$\sum \Psi(\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\theta}) = 0$$

where $\Psi = (\Psi_{\boldsymbol{\beta}}^T, \Psi_{\sigma_1^2}, \ldots, \Psi_{\sigma_r^2})^T$. We propose to use Tukey's biweight ρ-function (2.20) and call the resulting robust estimator CBS for constrained biweight S-estimator.

Like for *S*-estimators in the linear regression model in Section 3.2.4, (4.34) and (4.35) may have multiple roots, and hence a good high breakdown point estimator is needed as a starting point to find the solution to (4.34) and (4.35) with a high breakdown point. A simple algorithm has been suggested by Copt and Victoria-Feser (2006) and is given in Appendix C; the way the high breakdown point starting estimator is obtained is also detailed.

Following Davies (1987) and Lopuhaä (1992) for the normal multivariate case, Copt and Victoria-Feser (2006) prove that, under mild regularity conditions, a (constrained) *S*-estimator defined through (4.34) and (4.35) of $\boldsymbol{\theta}$ is consistent and asymptotically normal distributed. In particular, if the inverse of $\sum_{i=1}^{n} \boldsymbol{x}_i^T \boldsymbol{x}_i$ exists, $\hat{\boldsymbol{\beta}}_{[S]}$ has an asymptotic variance given by

$$\frac{e_1}{e_2^2}\left(\sum_{i=1}^{n} \boldsymbol{x}_i^T \boldsymbol{x}_i\right)^{-1} \sum_{i=1}^{n} \boldsymbol{x}_i^T \boldsymbol{\Sigma} \boldsymbol{x}_i \left(\sum_{i=1}^{n} \boldsymbol{x}_i^T \boldsymbol{x}_i\right)^{-1}, \tag{4.36}$$

where

$$e_1 = \frac{1}{p} E_\Phi[d^2 w(d)^2] \tag{4.37}$$

and

$$e_2 = E_\Phi\left[w(d) + \frac{1}{p} d \frac{\partial}{\partial d} w(d)\right] \tag{4.38}$$

and w is the weight function associated with the ρ-function. The constrained *S*-estimator of $\boldsymbol{\alpha}$ is also asymptotically normally distributed with variance given by a complex sandwich formula (omitted here for simplicity); see Copt and Victoria-Feser (2006, p. 294).

4.4.3 *MM*-estimators

In the same spirit as in the regression setting (see Section 3.2.4), Copt and Heritier (2007) propose *MM*-estimators for the main effects parameter. They possess many good properties, i.e. a high breakdown point even in the presence of leverage points, a good efficiency and, unlike *S*-estimators, they can be used to build a robust LRT-type test. This last property was the key incentive for their introduction; see also Section 4.5. The class of *MM*-estimators was first introduced by Yohai (1987) in the linear regression setting and was then generalized by Lopuhaä (1992) and Tatsuoka and Tyler (2000) to the multivariate linear model. The idea is to dissociate the estimation of the regression parameter (fixed effects) and variance component (random effects), and proceed in two steps. In the MLM setting, one can first obtain a high breakdown point estimator for the covariance matrix via the CBS estimator ($\widehat{\boldsymbol{\Sigma}}_{CBS}$) then use a better tuned ρ function to obtain a more efficient *M*-estimator for the fixed effects parameter (i.e. $\boldsymbol{\beta}$). In practice the initial variance estimator is based on a ρ-function $\rho_0(d; c_0)$, the final estimator on $\rho_1(d; c_1)$. The tuning constants are usually chosen to achieve a specific breakdown point (through c_0) and efficiency

(through c_1) at the model. Technically, the second step amounts to solving for $\boldsymbol{\beta}$

$$\sum_{i=1}^{n} \Psi(\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\beta}) = \sum_{i=1}^{n} w_1(d_i) \boldsymbol{x}_i^T \widehat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{y}_i - \boldsymbol{x}_i \boldsymbol{\beta}) = 0, \tag{4.39}$$

where e.g. $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Sigma}}_{[CBS]}$, $w_1(d) = (\partial/\partial d)\rho_1(d; c_1)/d$ is the weight function associated with ρ_1, i.e. the ρ-function in the M-step. The solution of (4.39) is the *MM*-estimator $\widehat{\boldsymbol{\beta}}_{[MM]}$ of $\boldsymbol{\beta}$.

Two natural choices for $w_1(\cdot)$ (and, hence, ρ_1) naturally arise from the regression setting, either the Huber's ρ-function (see Equation (2.17)) or the bounded Tukey's biweight ρ-function (see Equation (2.20)) leading to $\widehat{\boldsymbol{\beta}}_{[Hub]}$ and $\widehat{\boldsymbol{\beta}}_{[bi]}$, respectively. The corresponding weights are

$$w_1(d) = \min(1, c_1/|d|), \tag{4.40}$$

for Huber's weights and

$$w_1(d) = \begin{cases} \left(\left(\dfrac{d}{c_1} \right)^2 - 1 \right)^2 & \text{if } |d| \leq c_1 \\ 0 & \text{if } |d| > c_1 \end{cases} \tag{4.41}$$

for Tukey's biweight weights (see also (3.14)). These two proposals serve different purposes. Huber's estimator is well adapted to the cases when model deviations occur in the response variable only such as in ANOVA or models with well-controlled covariates. It can, however, be severely biased in the presence of (bad) leverage points. This is not the case with Tukey's biweight which is robust to both response and covariate extreme observations. Note that for Huber's weights (4.40), the associated ρ-function is (2.17), and for biweight weights (4.41) it is (3.15) with c replaced by c_1 in both cases.

Copt and Heritier (2007) show that, under mild conditions on ρ_1, $\sqrt{n}(\hat{\boldsymbol{\beta}}_{[MM]} - \boldsymbol{\beta})$ has a limiting normal distribution with zero mean and $\text{var}(\hat{\boldsymbol{\beta}}_{[MM]}) = H = (1/n) M^{-1} Q M^{-T}$ where M and Q are proportional to $\Gamma = E_K[\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x}]$ and K is the covariates' distribution.[7] A simpler representation for H can thus be given by

$$H = \frac{1}{n} \frac{e_1}{e_2^2} E_K[\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x}]^{-1}, \tag{4.42}$$

where e_1 and e_2 are given in (4.37) and (4.38), respectively, with $w(d) = (\partial/\partial d)$ $\rho_1(d)/d$. In the case of fixed covariates, K can be replaced by the covariates' empirical distribution in (4.42) yielding an asymptotic variance matrix H proportional to the asymptotic variance of the MLE (4.22). The multiplicative constant e_1/e_2^2 will

[7] In this section, we work under slightly more general conditions than in Section 4.4.2 by assuming that the covariates are not necessarily fixed but have a common distribution K. The rationale for this is to be able to account for leverage points or other problems in the covariates space. If one does not want to specify a particular model for K and therefore get back to the previous setting, one only needs to replace K by the empirical distribution of $\boldsymbol{x}$.

Table 4.2 Values for c_0 and c_1 for Tukey's biweight ρ-function (2.20) for the multivariate normal model.

Constant c_0 for a breakdown point of 50%										
p	1	2	3	4	5	6	7	8	9	10
c_0	1.56	2.66	3.45	4.09	4.65	5.14	5.59	6.01	6.40	6.77
Constant c_1 for 95% efficiency										
c_1	4.68	5.12	5.51	5.82	6.10	6.37	6.60	6.83	7.04	7.25

be used to calibrate the efficiency of the *MM*-estimator (see below). However, we prefer to ignore the reduced form (4.42) to derive an estimate of H and use instead the sample analog of the sandwich formula

$$\widehat{H} = \frac{1}{n}\widehat{M}^{-1}\widehat{Q}\widehat{M}^{-1},$$

where $\widehat{M}$ and $\widehat{Q}$ are the empirical versions of (2.28) and (2.29) for the MLM. For instance, $\widehat{M} = (1/n)\sum_{i=1}^{n}\Psi(\boldsymbol{y_i}, \boldsymbol{x}_i; \boldsymbol{\beta})s(\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\beta})^T$ with Ψ as in (4.39), and s is the score function (4.27) where again $\hat{\boldsymbol{\Sigma}}_{[CBS]}$ has been plugged in for $\boldsymbol{\Sigma}$. Such an estimator is usually more robust when extreme covariate values are observed. Numerical values are obtained by replacing $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}_{[MM]}$.

4.4.4 Choosing the Tuning Constants

As mentioned earlier, the constant c_0 of ρ_0 is chosen to ensure a high (asymptotic) breakdown point ε^* (50% in our case) for the initial estimate $\widehat{\boldsymbol{\Sigma}}_{[CBS]}$. For that purpose, the relationship

$$E[\rho_0(d; c_0)] = \varepsilon^* \max_x \rho_0(x; c_0)$$

is solved for c_0 to achieve a pre-specified breakdown point ε^* with (in our examples) Tukey's biweight ρ_0. To determine the constant c_1, an efficiency level (typically 95%) needs to be specified *a priori*. As discussed earlier, formula (4.42) shows that the relative efficiency of the *MM*-estimator relative to the MLE is given by the ratio

$$\frac{e_2^2}{e_1} = p\frac{E_\Phi[w_1(d) + (1/p)d(\partial/\partial d)w_1(d)]^2}{E_\Phi[d^2 w_1(d)^2]} \tag{4.43}$$

with $w_1(d)$ given in (4.40) or (4.41) depending on the choice for ρ_1 (Huber or biweight) and $c = c_1$. The constant c_1 is then found by equating (4.43) to the desired efficiency level (e.g. 95%). Note that in the univariate case ($p = 1$), (4.43) reduces to (3.20).

Both constants depend on the dimension p of the response vector and can be obtained by Monte Carlo simulations. They are summarized in Table 4.2 for Tukey's

Table 4.3 Estimates and standard errors for the REML and the CBS–*MM* for the skin resistance data using model (4.2).

	REML		CBS–*MM*	
Parameter	Estimate (*SE*)	p-value	Estimate (*SE*)	p-value
μ	2.030 (0.341)	$<10^{-4}$	1.440 (0.233)	$<10^{-4}$
λ_1	−0.213 (0.334)	0.525	−0.161 (0.175)	0.356
λ_2	0.842 (0.334)	0.014	0.403 (0.175)	0.021
λ_3	0.549 (0.334)	0.105	0.243 (0.175)	0.163
λ_4	−0.526 (0.334)	0.120	−0.169 (0.175)	0.332
σ_s	1.190		0.842	
σ_ϵ	1.459		0.761	

CBS computed with $c_0 = 4.65$ and *MM* (biweight) computed with $c_1 = 6.10$.

biweight ρ-functions (for ρ_0 and ρ_1). When p becomes large enough, an asymptotic approximation given by Rocke (1996, p. 1330) can be used for Tukey's biweight which yields $c_1 = \sqrt{p}/m$ where m is defined through $\rho_{[bi]}(m) = 0.5\rho_{[bi]}(1)$, with $\rho_{[bi]}$ given in (2.20) with $c = 1$. This approximation gives reasonable results from $p > 10$. Finally note that the values of c_0 and c_1 given here obviously depend on the choice of the ρ-function and would need to be recomputed had other ρ-functions been used. Another option is available for the Huber estimator, i.e. when Huber weights (4.40) are chosen. It stems from the fact that ρ in (2.17) is a function of d, the Mahalanobis distance. As d^2 has a chi-squared distribution with p degrees of freedom χ^2_p, c_1 can be chosen as the square-root of a specific quantile of this distribution.

4.4.5 Skin Resistance Data (continued)

As an illustration, we go back to the skin resistance data. Table 4.3 presents the robust *MM* estimates $\widehat{\boldsymbol{\beta}}_{[bi]}$ and robust CBS estimates $\hat{\boldsymbol{\alpha}}_{[CBS]}$[8] and standard errors for the electrode resistance data[9] along with the REMLs obtained earlier. The *MM* contrast estimates are not affected by case 15's extreme readings for electrodes of type 2 and 3. They are actually close to what was observed with case 15 removed from the analysis. The CBS variances estimates, especially the residual estimate, are much smaller confirming the previous findings that the REML estimates are unduly inflated by the two abnormal readings.

To limit the influence of potential outlying observations, Berry (1987) actually proposes to use a $\log(y + c)$ ($c = 32$) transformation of the data. A profile plot of the transformed data is presented in Figure 4.3. Graphically, the log-transformation limits the effect of the potential outliers (in particular observation number 15). The estimated model parameters using the transformed data and the classical (REML)

[8]For simplicity, we call this set of robust estimators the CBS–*MM*.

[9]The raw data have been divided by 100.

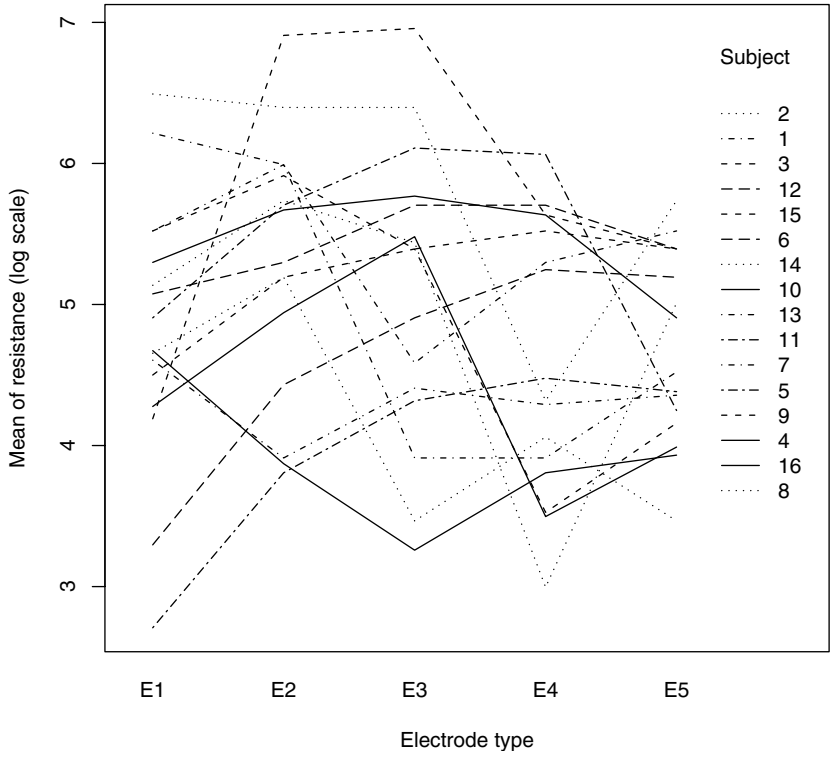

Figure 4.3 Profile plot for the skin resistance data (log-transformed).

and robust estimators are presented in Table 4.4. The overall mean, the contrasts and variance components and p-values this time are similar in the two methods. This illustrates the fact that outliers are model specific, i.e. the two abnormal readings on the original scale do not appear as so extreme on the log-transformed one. This was not the case with the non-transformed data. We defer the discussion on the effect of the electrode type to Section 4.5.3.

4.5 Robust Inference

The *MM*-estimators were introduced earlier to offer more options for testing hypotheses on the main effects. Typical tests usually involve contrasts or multidimensional hypotheses that a component of the main effects parameter is null.

4.5.1 Testing Contrasts

A contrast test occurs when a linear combination of the elements of $\boldsymbol{\beta}$, typically represented by a $(q+1)$-vector $\boldsymbol{L}$, is tested. For example, suppose that we have a

Table 4.4 Estimates and standard errors for the REML and the CBS–*MM* for the skin resistance data using model (4.2) with a log-transformed response.

	REML		CBS–*MM*	
Parameter	Estimate (*SE*)	p-value	Estimate (*SE*)	p-value
μ	4.913 (0.166)	$<10^{-4}$	4.918 (0.176)	$<10^{-4}$
λ_1	−0.097 (0.158)	0.542	−0.058 (0.161)	0.718
λ_2	0.396 (0.158)	0.015	0.376 (0.161)	0.019
λ_3	0.179 (0.158)	0.262	0.167 (0.161)	0.299
λ_4	−0.289 (0.158)	0.072	−0.282 (0.161)	0.079
σ_s	0.585		0.610	
σ_ϵ	0.701		0.689	

CBS computed with $c_0 = 4.65$ and *MM* (biweight) computed with $c_1 = 6.10$.

one-factor within-subject ANOVA model with three levels, i.e. $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2) = (\mu, \lambda_1, \lambda_2)$ and suppose that the design matrix $\boldsymbol{x}$ is parametrized as 'treatment' contrasts (see e.g. (4.11)) with the third level as the reference level. Suppose also that our goal is to test for differences among elements of the mean vector (μ_1, μ_2, μ_3). The corresponding null hypotheses are

$$
\begin{aligned}
H_0 &: \mu_1 - \mu_3 = \beta_2 = \lambda_1 = 0 \\
H_1 &: \mu_1 - \mu_3 = \beta_2 = \lambda_1 \neq 0, \\
H_0 &: \mu_2 - \mu_3 = \beta_3 = \lambda_2 = 0 \\
H_1 &: \mu_2 - \mu_3 = \beta_3 = \lambda_2 \neq 0, \\
H_0 &: \mu_2 - \mu_1 = \beta_3 - \beta_2 = \lambda_2 - \lambda_1 = 0 \\
H_1 &: \mu_2 - \mu_1 = \beta_3 - \beta_2 = \lambda_2 - \lambda_1 \neq 0.
\end{aligned}
$$

The corresponding contrasts $\boldsymbol{L}$ are $\boldsymbol{L}^T = (0, 1, 0)$, $\boldsymbol{L}^T = (0, 0, 1)$ and $\boldsymbol{L}^T = (0, 1, -1)$.

Simple robust inference for contrasts can be performed using an estimate of the asymptotic covariance of $\widehat{\boldsymbol{\beta}}_{[MM]}$ given in (4.42). For $H_0 : \boldsymbol{L}^T\boldsymbol{\beta} = 0$, a robust z-test statistic is given by

$$
z\text{-statistic} = \frac{\boldsymbol{L}^T\widehat{\boldsymbol{\beta}}_{[MM]}}{SE(\boldsymbol{L}^T\widehat{\boldsymbol{\beta}}_{[MM]})} \tag{4.44}
$$

with

$$
SE(\boldsymbol{L}^T\hat{\boldsymbol{\beta}}_{[MM]}) = \sqrt{\boldsymbol{L}^T\hat{H}\boldsymbol{L}}.
$$

The corresponding p-value is obtained by comparing (4.44) with the standard normal distribution. Note that, although we compute the z-statistic with the *MM*-estimator, the same sort of calculation can be done with the *S*-estimator using the appropriate asymptotic variance.

4.5.2 Multiple Hypothesis Testing of the Main Effects

Tests involving multiple hypothesis can, for instance, be used to compare models with the same variance structure or to assess the statistical significance of a factor with several levels such as the type of electrode in model (4.2). Denote again by $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_{(1)}^T, \boldsymbol{\beta}_{(2)}^T)$ the partition of the vector $\boldsymbol{\beta}$ into $q+1-k$ and k components and by $A_{(ij)}$, $i, j = 1, 2$ the corresponding partition of $(q+1) \times (q+1)$ matrices. The hypothesis to be tested can usually be formulated as

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 \text{ where } \boldsymbol{\beta}_{0(2)} = 0, \boldsymbol{\beta}_{0(1)} \text{ unspecified},$$

$$H_1 : \boldsymbol{\beta}_{0(2)} \neq 0, \boldsymbol{\beta}_{(1)} \text{ unspecified}.$$

The need for robust testing in this setting is obvious as the classical F-test has reportedly been found to be unreliable under sometimes mild model deviations (see e.g. (Copt and Heritier, 2007)). Robust alternatives to the classical Wald or score tests are readily available through (2.47) for the robust Wald test, (2.48) for the robust score test for any model. Robustifying the LRT is probably the most natural route to build a robust alternative to the F-test but, as alluded to in Section 4.4.3, such a test does not always exist for S-estimators. The reason is that the corresponding test statistic is by construction zero. To see this, just note that the robust LRT in (2.50) is based on the difference in $\sum \rho(d_i)$ for both the full and reduced models. As the definition of S-estimators (4.30) sets both sums to b_0 (up to a $1/n$ factor) the difference is simply zero. As shown in Copt and Heritier (2007), MM-estimators circumvent the problem by using *another* loss function ρ_1, different from that used to build the S-estimator, therefore the LRT statistic exits.

A direct application of the general theory of robust testing introduced in Section 2.5 can then be used. Formally, the LRT statistic is computed in the same way as in the general case. Again let $d_i(\boldsymbol{\beta}) = \sqrt{(\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta})^T \widehat{\boldsymbol{\Sigma}}_{[S]}^{-1} (\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta})}$ be the Mahalanobis distance for observation i with $\hat{\boldsymbol{\Sigma}}_{[S]}$ a chosen S-estimator of $\boldsymbol{\Sigma}$ (e.g. $\hat{\boldsymbol{\Sigma}}_{[CBS]}$). The robust LRT -type test statistic is given by

$$\text{LRT}_\rho = 2 \sum_{i=1}^{n} [\rho(d_i(\dot{\boldsymbol{\beta}}_{[MM]})) - \rho(d_i(\widehat{\boldsymbol{\beta}}_{[MM]}))], \tag{4.45}$$

where $\widehat{\boldsymbol{\beta}}_{[MM]}$ and $\dot{\boldsymbol{\beta}}_{[MM]}$ are the robust estimators in the full and reduced models, respectively, with corresponding loss function ρ_1. More specifically LRT_ρ associated with the Huber estimator, respectively the biweight estimator, is defined through (4.39) with weight function (4.40), respectively (4.41), with corresponding ρ_1-function given in (2.17), respectively in (3.15). In both cases the covariance matrix estimate is the CBS $\widehat{\boldsymbol{\Sigma}}_{[CBS]}$.

An estimate of a robust Wald-type test statistic is naturally defined by

$$W_\Psi^2 = \widehat{\boldsymbol{\beta}}_{[MM](2)}^T \widehat{H}_{(22)}^{-1} \widehat{\boldsymbol{\beta}}_{[MM](2)},$$

where $\widehat{\boldsymbol{\beta}}_{[MM](2)}$ is the robust MM-estimator of $\boldsymbol{\beta}_{(2)}$ in the full model and $\hat{H}_{(22)}$ the corresponding covariance estimate. Finally, a score-(or Rao-)type test statistic is

given by

$$R^2_\Psi = Z_n^T \widehat{C}^{-1} Z_n,$$

where $Z_n = (1/n) \sum_{i=1}^n \Psi(\boldsymbol{y}_i, \boldsymbol{x}_i; \dot{\boldsymbol{\beta}}_{[MM]})_{(2)}$, $\dot{\boldsymbol{\beta}}_{[MM]}$ is the *MM*-estimator in the reduced model with corresponding Ψ-function given in (4.39) and weights in (4.40) for Huber's estimator and in (4.41) for Tukey's biweight estimator. The $k \times k$ positive-definite matrix $\widehat{C}$ is $\widehat{C} = \widehat{M}_{22.1} \widehat{H}_{(22)} \widehat{M}_{22.1}^T$ with $\widehat{M}_{22.1} = \widehat{M}_{(22)} - \widehat{M}_{(21)} \widehat{M}_{(11)}^{-1} \widehat{M}_{(12)}$ from the partitioning of the matrix $\widehat{M}$. Again we have defined the three test statistics for the *MM*-estimators but it is also possible to define the robust Wald and score test in a similar fashion for the CBS estimators. Under the null hypothesis, their asymptotic distribution is the same as in the general parametric settings (see Section 2.5.4).

4.5.3 Skin Resistance Data Example (continued)

We now return to the problem of testing the multivariate hypothesis of equality of mean resistances given by

$$H_1 : H_0 \text{ is not true} \quad \mu \text{ unspecified}$$

irrespective of the chosen contrast matrix.

The classical F-test statistic is 3.1455. When compared with an $F_{4,60}$ distribution, we find a p-value of 0.020 so that the test is significant at the 5% level. We could conclude that there is a difference between the five electrode types. Using the Tukey's biweight ρ-function, the robust LRT test statistics yields a p-value of 0.086 at the same 5% level. The test is, hence, not significant. Observations 15 and possibly 2 seem to have an influence on the MLE (or REML) estimates and consequently on the F-test.

If the responses are log-transformed, the F-test statistic is 2.87 corresponding to a p-value of 0.03 and the robust LRT test gives a p-value of 0.061. Although the log-transformation gives similar results for the parameters' estimates (see Table 4.4), it does not completely reduce the influence of the outlying observations (number 15 and possibly number 2) on the classical F-test: we still reject the null hypothesis of equal resistances. Note that Berry (1987) analyzes this dataset with subject 15 deleted, and finds a significant F-test on the original data (p-value of 0.044) and a non-significant F-test on the log-transformed data (p-value of 0.10).

4.5.4 Semantic Priming Data Example (continued)

The model used to analyze this dataset is given in (4.15) with λ_j, $j = 1, 2$, the fixed effect for the delay and γ_k, $k = 1, 2, 3$, the fixed effect for the condition. Table 4.5 gives the estimates for the REML and the CBS–*MM* and the standard errors for the fixed effects computed using Tukey's biweight weights. The contrasts for each factor are the 'sum'-type contrasts. We can see that both methods detect a significant effect for the delay but with a borderline p-value of 0.046 for the REML whereas the message is clearer with the robust method yielding a p-value of 0.003. Another

Table 4.5 Estimates and standard errors for the REML and the CBS–*MM* for the semantic priming data using model (4.15).

	REML		CBS–*MM*	
Parameter	Estimate (*SE*)	p-value	Estimate (*SE*)	p-value
μ	633.436 (28.465)	$<10^{-4}$	586.420 (18.817)	$<10^{-4}$
λ_1	−18.071 (8.974)	0.046	−17.876 (6.082)	0.003
γ_1	18.563 (13.732)	0.179	14.317 (11.691)	0.221
γ_2	−51.222 (13.732)	$<10^{-4}$	−56.994 (11.691)	$<10^{-4}$
$\lambda\gamma_{11}$	−3.690 (12.691)	0.771	12.706 (10.582)	0.230
$\lambda\gamma_{12}$	16.809 (12.691)	0.188	(8.844 (10.582)	0.403
σ_s	122.622		77.991	
$\sigma_{\lambda s}$	0.006		N/A	
$\sigma_{\gamma s}$	29.433		27.199	
σ_ϵ	100.73		81.885	

CBS computed with $c_0 = 5.14$ and *MM* (biweight) computed with $c_1 = 6.37$.

important feature of this model is the estimation of the random effects. The robust estimate of the variance for the interaction between subject and delay is not reported. This is because the robust estimator gives a negative value. This can sometimes happen as some of the variance components correspond to covariances between responses on the same subject and, hence, can in principle be negative. Standard algorithms included in common statistical packages work around this solution by imputing very small values close to zero each time a variance is found to be negative. In this example, using the R package `lme`, one obtains a small value (0.006) for the corresponding classical estimator (REML).[10]

We also tested the significance of each factor and of the interactions, using the F-test and the robust LRT-type test. Results are presented in Table 4.6. The classical F-test and robust LRT test give similar results for the three hypotheses with, again, a stronger effect for the delay variable. In this example, the presence of possible outlier does not seem to influence the results of the tests.

With this type of data, one can also consider a log-transformation, although in this domain one usually prefers the original scale, mainly for interpretation reasons. In Table 4.7 we give the REML and the CBS–*MM* estimates with corresponding standard errors. The estimates and p-values for significance testing are quite similar and lead to the same conclusions. Also note that again the variance of the random effect for the interaction between the subject and the delay is set to zero with the REML and found to be negative (and hence reported as N/A) with the CBS estimator.

We can also test the significance of each factor and of the interactions, using the F-test and the robust LRT-type test. Results are presented in Table 4.8. Both approaches lead to similar conclusions.

[10]The problem of negative variances is not specific to robust approaches but is a common problem in the general ANOVA/MLM setting; see, for example, Searle *et al.* (1992).

Table 4.6 Classical F-test and robust LRT for the fixed effects of the semantic priming data using model (4.15).

	p-value	
Variable	Classical F-test	Robust LRT test
Delay	0.046	0.005
Condition	0.001	0.001
Delay:Condition	0.383	0.131

Robust LRT test computed using the CBS with $c_0 = 5.14$ and the *MM* (biweight) with $c_1 = 6.37$.

Table 4.7 Estimates and standard errors for the REML and the CBS–*MM* for the semantic priming data using model (4.15) with log-transformed data.

	REML		CBS–*MM*	
Parameter	Estimate (*SE*)	p-value	Estimate (*SE*)	p-value
μ	6.421 (0.040)	$<10^{-4}$	6.386 (0.036)	$<10^{-4}$
λ_1	-0.027 (0.012)	0.025	-0.028 (0.010)	0.007
γ_1	0.032 (0.021)	0.127	0.029 (0.022)	0.177
γ_2	-0.088 (0.021)	$<10^{-4}$	-0.092 (0.022)	$<10^{-4}$
$\lambda\gamma_{11}$	0.002 (0.017)	0.873	0.011 (0.018)	0.526
$\lambda\gamma_{12}$	0.018 (0.017)	0.271	0.017 (0.018)	0.336
σ_s	0.173		0.148	
$\sigma_{\lambda s}$	0.000		N/A	
$\sigma_{\gamma s}$	0.069		0.069	
σ_ϵ	0.136		0.137	

CBS computed with $c_0 = 5.14$ and *MM* (biweight) computed with $c_1 = 6.37$.

Table 4.8 Classical F-test and robust LRT for the fixed effects of the semantic priming data using model (4.15), with log-transformed data.

	p-value	
Variable	Classical F-test	Robust LRT test
Delay	0.025	0.008
Condition	0.0003	0.001
Delay:Condition	0.390	0.272

Robust LRT test computed using the CBS with $c_0 = 5.14$ and the *MM* (biweight) with $c_1 = 6.37$.

4.5.5 Testing the Variance Components

Most of the effort in robust testing in MLMs has focused on the main effects because the variance parameters are often considered as nuisance parameters. If one is truly interested in testing whether some random effects could be removed, the same problem mentioned above arises. As the null hypothesis typically involves restrictions of the type $\sigma_j^2 = 0$, the overall null parameter vector $\boldsymbol{\theta}_0$ is on the boundary of the parameter space and, as a result, the general theory of Section 2.5.3 breaks down. One could conjecture that the same kind of mixture of χ^2 distributions could be used for the robust Wald test. However such tests are known to perform poorly in the classical case and a similar behavior is expected in the robust case. The LRT test could constitute a better alternative but no such robust LRT test exists to the best of our knowledge as the only proposal to date, the robust LRT test (4.45), only targets hypotheses on the fixed effects. At this stage, the only viable option seems to use bootstrapping techniques with the warning mentioned in Chapter 2 that the simple bootstrap can fail when applied to robust estimators (as the breakdown point may be reached in some bootstrap samples). Our practical recommendation in that case is to use a robust estimator with a 50% breakdown point to have a good chance of avoiding the problem.

4.6 Checking the Model

4.6.1 Detecting Outlying and Influential Observations

Since the MLM can be seen as a multivariate normal model, multivariate tools can be used to measure in some sense at which point the observations are far from the bulk of data. Such a tool is given by the Mahalanobis distances in (4.31) in which $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are replaced by suitable estimates. In order for the estimated Mahalanobis distances not to be influenced (hence biased) by extreme observations, it is necessary that $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are replaced by their robust estimators, namely $\widehat{\boldsymbol{\beta}}_{[MM]}$ and $\widehat{\boldsymbol{\Sigma}}_{[CBS]}$. One then can rely on the asymptotic result that d_i in (4.31) has an asymptotic χ_p^2 distribution and, hence, compare the estimated Mahalanobis distances to, say, the corresponding 0.975 quantile. One can also, for comparison, estimate the Mahalanobis distances using the MLE or the REML for $\boldsymbol{\beta}$ and the variance components of $\boldsymbol{\Sigma}$. A scatterplot of the robust versus classical Mahalanobis distances would reveal the outlying observations, i.e. the observations with corresponding robust and classical Mahalanobis distance above the 0.975 quantile of the χ_p^2, as well as the influential observations, i.e. the observations with corresponding robust Mahalanobis distances above and the corresponding classical Mahalanobis distances below the 0.975 quantile of the χ_p^2. These influential observations are such that the classical estimator is not able to detect them but is influenced by them. In multivariate setting such as with MLM, Mahalanobis distances are usually preferred to the weights per se to detect outlying observations.

As an example, consider the skin resistance dataset estimated in Section 4.4.5. In Figure 4.1 we saw that out of the 80 readings, two measurements (electrodes of

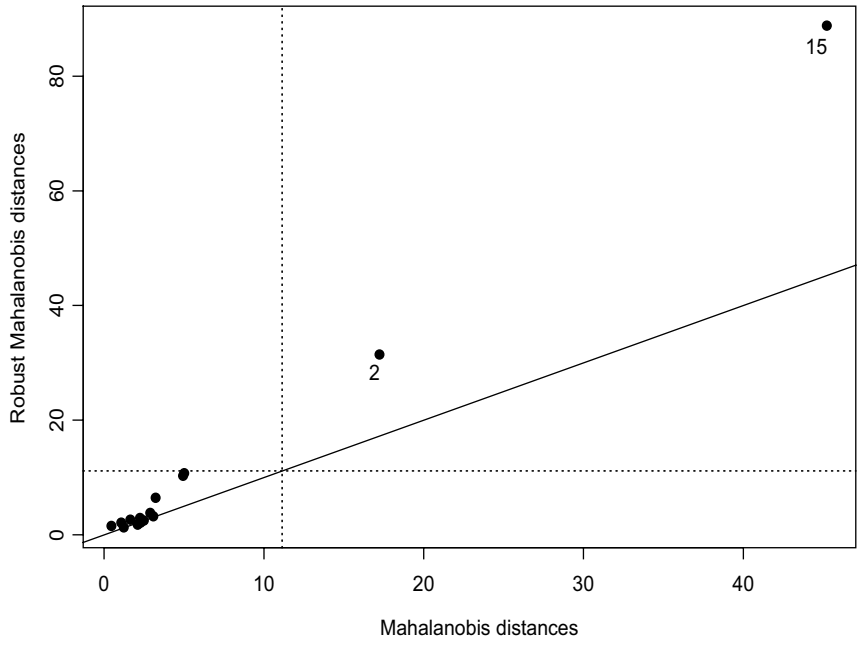

Figure 4.4 Scatterplot of the Mahalanobis distances for the skin resistance data. CBS computed with $c_0 = 4.65$ and *MM* (biweight) with $c_1 = 6.10$.

type 2 and 3) taken on subject 15 were much larger than the others. Observation number 2 corresponds to the second largest response. In Figure 4.4 we give the scatterplot of the Mahalanobis distances computed with the REML and the CBS–*MM*. The horizontal and vertical dotted lines correspond to the 0.975 quantile on the χ^2_5 distribution, to detect outlying observations. The REML and CBS–*MM* estimators detect observations 15 and 2 as outlying observations. No influential observations is present in the sample. With the log-transformed data, the scatterplot of the Mahalanobis distances given in Figure 4.5, shows that the CBS–*MM* detects observation 15 as an influential observation, and observation 2 is no longer considered as extreme.

As another example, consider the semantic priming dataset estimated in Section 4.5.4. In Figure 4.6 we give the scatterplot of the Mahalanobis distances computed with the REML and the CBS–*MM*. One can see that the REML and CBS–*MM* detect one outlier (observation 3) and the CBS–*MM* detects two influential observations (observations 8 and 16). These observations are certainly the cause of the differences found between the classical and robust estimates. With the log-transformed data, the scatterplot of the Mahalanobis distances for the corresponding REML and CBS–*MM* estimates is given in Figure 4.7. One can see that there are two outliers detected

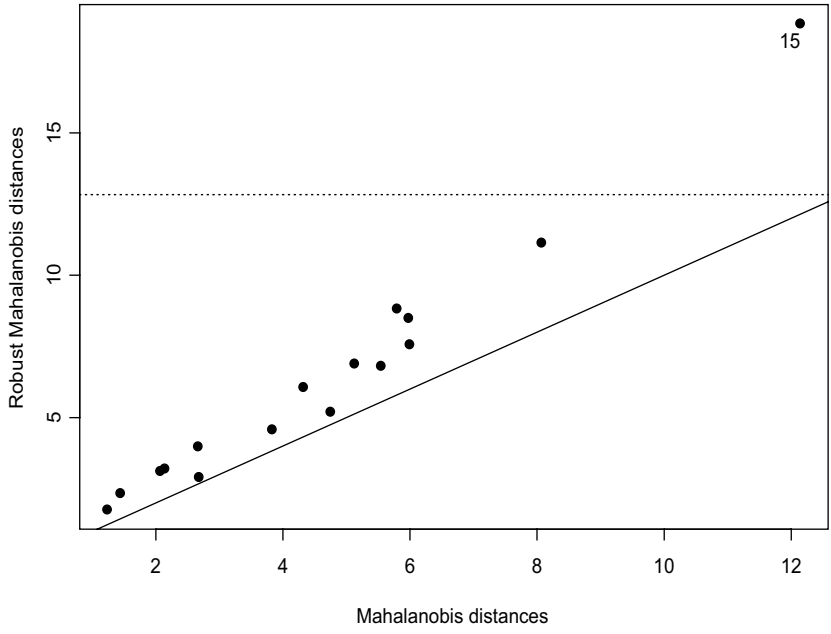

Figure 4.5 Scatterplot of the Mahalanobis distances for the skin resistance data (log-transformed). CBS with $c_0 = 4.65$ and *MM* (biweight) with $c_1 = 6.10$.

by the REML and CBS–*MM*, and they do not seem to have much influence on the estimates.

4.6.2 Prediction and Residual Analysis

As for the regression model of Chapter 3, residual analysis with MLMs is used to check the model fit and also the model assumptions. In order to compute residuals, one needs to be able to compute predicted values for the response vector $\boldsymbol{y}$. For that, and with MLM, one also needs to compute estimates for the random effects levels. Actually, one can define predicted (or fitted) response values at different levels of nesting or directly at the population level. Given estimated values for $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma_0^2, \ldots, \sigma_r^2)^T$, the predictions at the so-called population level are

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} \tag{4.46}$$

and the predictions at the so-called cluster (lowest) level are

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}\widehat{\boldsymbol{\gamma}}. \tag{4.47}$$

We note that depending on the problem and for hierarchical models, there might be different cluster levels, so that $\boldsymbol{Z}\widehat{\boldsymbol{\gamma}}$ in (4.47) can be modified accordingly. In all cases,

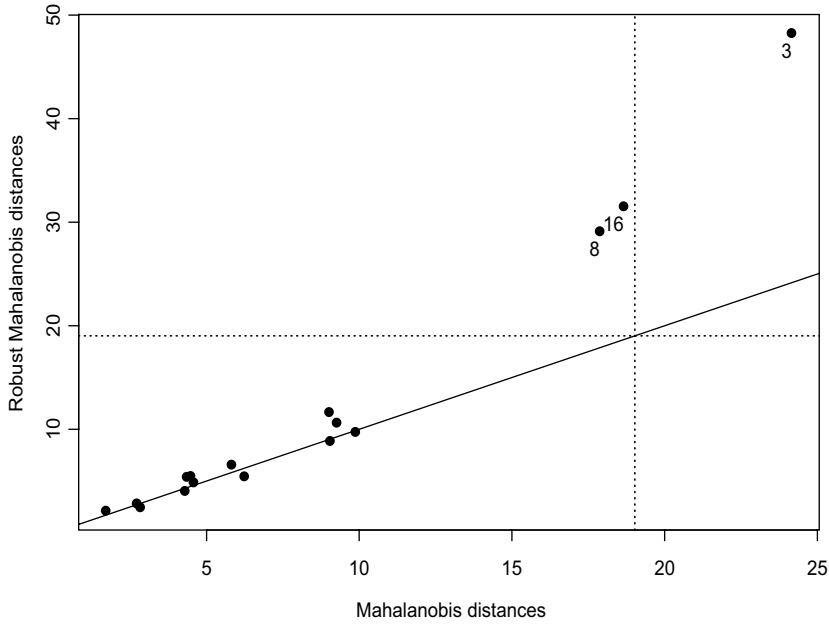

Figure 4.6 Scatterplot of the Mahalanobis distances for the semantic priming data. CBS with $c_0 = 5.14$ and *MM* (biweight) with $c_1 = 6.37$.

when predicting at the cluster levels, an estimate for $\widehat{\boldsymbol{\gamma}}$ is needed so that the first step is to define estimators for the random effects levels.

Recall that random effects are unobservable variables. However, given the information contained in a sample and given a model, it is possible to predict (an expected value of) the vector of random effects for each response. Classically, one uses the Best Linear Unbiased Predictor (BLUP) given by[11]

$$\widehat{\boldsymbol{\gamma}} = \boldsymbol{D}\boldsymbol{Z}^T\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{4.48}$$

where $\boldsymbol{D} = \text{cov}(\boldsymbol{\gamma})$. Given values for the variance components $\boldsymbol{\alpha}$, (4.48) is computed using (4.21) for $\boldsymbol{\beta}$. An interesting interpretation of $\widehat{\boldsymbol{\gamma}}$ is that it is the MLE based on the likelihood of the joint distribution of $f(\boldsymbol{y}, \boldsymbol{\gamma}) = f(\boldsymbol{y}|\boldsymbol{\gamma})f(\boldsymbol{\gamma})$ (for fixed values of $\boldsymbol{\alpha}$). Henderson *et al.* (1959) propose a set of equations for the simultaneous estimation of $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\beta}}$ indeed based on the joint distribution of $\boldsymbol{y}$ and $\boldsymbol{\gamma}$.

Prediction and residual analysis with robust estimators is not as straightforward as replacing all parameters in (4.48) by their robust estimates. If we choose this simple approach, we face the risk that a random effect corresponding to a particular observation $y_{ijk\ldots}$ could be overestimated or underestimated if this observation is considered as an outlier in terms of the Mahalanobis distance. Indeed Copt and

[11] See e.g. McCulloch and Searle (2001, Chapter 9).

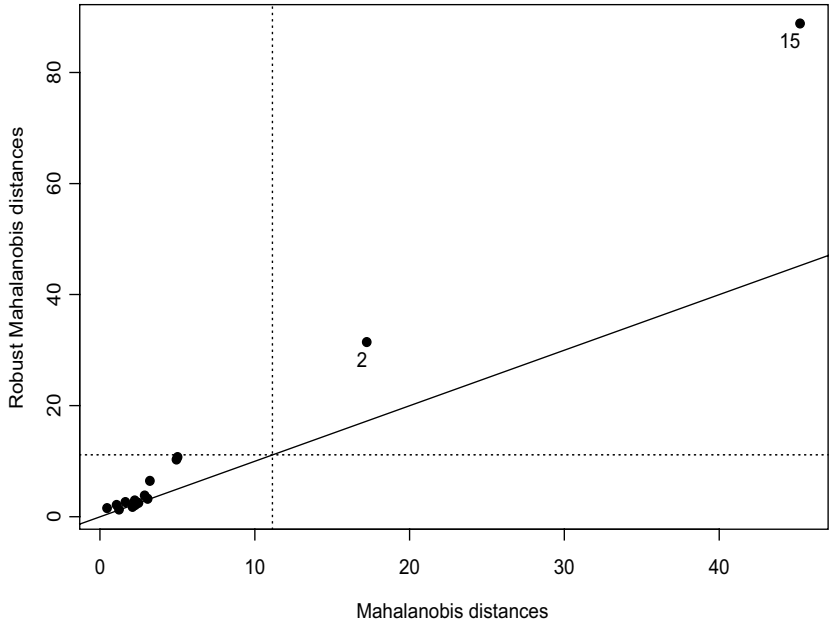

Figure 4.7 Scatterplot of the Mahalanobis distances for the semantic priming data (log-transformed). CBS computed with $c_0 = 5.14$ and *MM* (biweight) with $c_1 = 6.37$.

Victoria-Feser (2009) show that the *IF* of $\widehat{\boldsymbol{\gamma}}$ in (4.48) depends on the robustness properties of $\widehat{\boldsymbol{\beta}}(\boldsymbol{\alpha})$ and also on the deviations $(\boldsymbol{y} - \boldsymbol{X\beta})$. This means that in order to make the predictions robust to model deviations, one needs not only a robust estimator such as the CBS–*MM*, but also to bound (4.48). Copt and Victoria-Feser (2009) propose the use of ψ-based prediction defined as[12]

$$\hat{\boldsymbol{\gamma}}_{\psi} = e_{\psi,c}\boldsymbol{DZ}^T\boldsymbol{V}^{-1/2}\psi(\boldsymbol{V}^{-1/2}(\boldsymbol{y} - \boldsymbol{X\beta})),$$

where $\psi(r) = (\partial/\partial r)\rho(r)$ is a bounded function such as the Huber's of Tukey's biweight functions, and $e_{\psi,c}$ is a correction factor (see below). A bounded ψ-function is necessary to guarantee the robustness of the corresponding prediction estimator. Moreover, in order for $\widehat{\boldsymbol{\gamma}}_{\psi}$ to behave similarly to $\widehat{\boldsymbol{\gamma}}$ at the normal model, we also need to impose that $E[\widehat{\boldsymbol{\gamma}}_{\psi}] = 0$ and $\text{var}(\widehat{\boldsymbol{\gamma}}_{\psi}) = \text{var}(\widehat{\boldsymbol{\gamma}})$. These constraints define (implicitly) the correction factor $e_{\psi,c}$. For Tukey's biweight ψ-function, Copt and Victoria-Feser (2009) show that

$$e_{\psi_{[bi]},c} = \left(I_2(c) - \frac{4}{c^2}I_4(c) + \frac{6}{c^4}I_8(c) - \frac{4}{c^6}I_8(c) + \frac{1}{c^8}I_{10}(c)\right)^{-1/2}$$

[12]To compute $\boldsymbol{V}^{-1/2}$, we follow Richardson and Welsh (1995) and chose $\boldsymbol{V}^{-1/2}$ to be symmetric with the same additive structure as $\boldsymbol{V}$ and $\boldsymbol{V}^{-1}$ and with the property that $\boldsymbol{V}^{-1/2}\boldsymbol{V}^{-1/2} = \boldsymbol{V}^{-1}$.

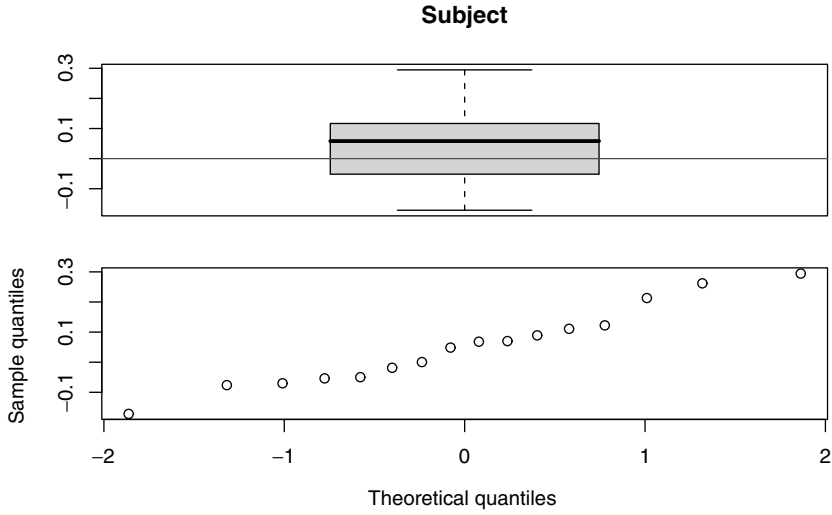

Figure 4.8 Boxplot and Q-Q plot of the (estimated) subject random effect for the skin resistance data. CBS computed with $c_0 = 4.65$ and *MM* (biweight) with $c_1 = 6.10$.

where

$$I_k(c) = \int_{-c}^{c} r^k \, d\Phi(r);$$

see Appendix B for the computation of these truncated normal moments. For Huber's ψ-function, $e_{\psi_{[Hub],c}} = (1 - 2c^2(1 - \Phi(c)))^{-1/2}$. Finally, to compute $\hat{\gamma}_\psi$ in practice, one replaces $\boldsymbol{\alpha}$ in ($\boldsymbol{V}$ and $\boldsymbol{D}$) and $\boldsymbol{\beta}$ by their robust estimates.

Estimated random effects can be used to check the model assumptions. Recall that the random effects are assumed to be normally distributed and independent of each other. A normal probability plot (normal quantiles against ordered estimated random effects) or a boxplot can be used to assess the normality assumption. Again, consider as an example the skin resistance dataset estimated in Section 4.2.2. This model has only one random effect, the subject. Figure 4.8 suggests that the normality of the subject random effect is fairly respected.

As in the linear regression setting, residuals are defined as the difference between the response and the predicted value, i.e. $\boldsymbol{y} - \widehat{\boldsymbol{y}}$ where $\hat{\boldsymbol{y}}$ is given in (4.47) and possibly also (4.46). They thus depend on the choice of predicted response. However, since random effects have been introduced into the model, it is more sensible to use the subject predicted values to define residuals as population fitted values may produce a structure in the residuals which is simply due to the random effects. The residuals can also be standardized by means of the (estimated) covariance matrix of $\boldsymbol{y}$, yielding $\boldsymbol{V}^{-1/2}(\boldsymbol{y} - \widehat{\boldsymbol{y}})$. Figure 4.9 displays the standardized residuals versus fitted values at the subject level. We can see that there is no particular structure in the residuals.

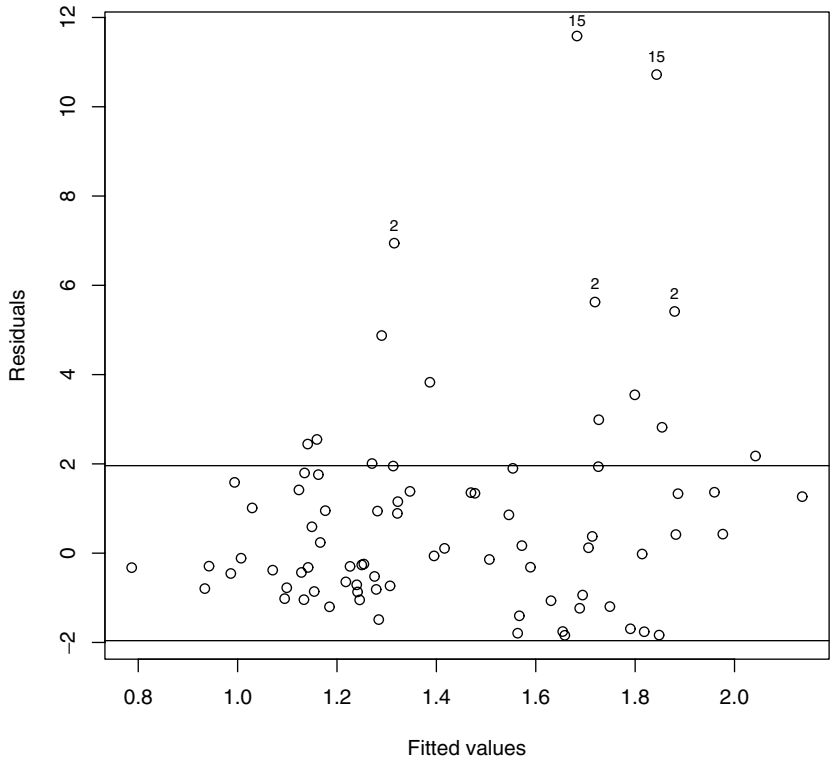

Figure 4.9 Standardized residuals (subject level) versus fitted values for the skin resistance data. CBS computed with $c_0 = 4.65$ and *MM* (biweight) with $c_1 = 6.10$.

4.7 Further Examples

4.7.1 Metallic Oxide Data

Until now, we have only presented models in which each level of a factor is combined with every level of another factor. Hierarchical models are models where only some levels of a factor are combined with the levels of another factor. More formally, suppose that we have two treatments $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$ with l and g levels, respectively. In the language of experimental design, if each level of treatment $\boldsymbol{\gamma}$ appears only in one level of treatment $\boldsymbol{\lambda}$, $\boldsymbol{\gamma}$ is said to be nested in $\boldsymbol{\lambda}$.

One can also extend the models so as to include so-called between subjects factors. For example, we have the typical experiment in which a measurement is taken from n_1 samples of type $j = 1$ and n_2 samples of type $j = 2$, and in each sample the measure is taken on g 'objects'. For example, the 'objects' can be rats, the samples cages, n_1 of which are given treatment $j = 1$ and others n_2 given treatment $j = 2$. This type of design is called a nested design. The rats are nested within the

cages. A rat belongs either to cage 1 or cage 2. We use different notation to represent nested factors. For example, suppose that $\boldsymbol{\gamma}$ is the parameter for the cage, then $\gamma_{j(i)}$ would represent rat i nested within cage j. The between subjects factor here is the treatment.

In this section, we analyze a dataset originating from a sampling study designed to explore the effects of process and measurement variation on the properties of lots of metallic oxides (Bennet, 1954). Two samples were drawn from each lot. Duplicate analyses were then performed by each of two chemists, with a pair of chemists randomly selected for each sample. Hence, the response y_{ijklm} corresponds to the metal content (percent by weight) measured on the ith metallic oxide type, on the jth lots, on the kth sample, by the lth chemist for the mth analysis. The model can be written as

$$y_{ijklm} = \mu + \lambda J_i(j) + \gamma_{j(i)} + \delta_{j(i(k))} + \xi_{j(i(k(l)))} + \epsilon_{j(i(k(l(m))))}, \tag{4.49}$$

where

$$J_i(j) = \begin{cases} 0 & j = 1, \\ 1 & j = 2, \end{cases}$$

and with $\mu + \lambda J_i(j)$ the fixed effect and $\gamma_{j(i)}, i = 1, \ldots, n = n_1 + n_2$ the random effect due to the lot, $\delta_{j(i(k))}, k = 1, \ldots, 2n$, the random effect due to the sample and $\xi_{j(i(k(l)))}, l = 1, \ldots, 4n$, the random effect due to the chemist. We then have

$$\boldsymbol{\mu}_i = \boldsymbol{e}_8(\mu + \lambda J_i(j)) = \boldsymbol{e}_8 \otimes (1, J_i(j))(\mu, \lambda)^T = \boldsymbol{x}_i \boldsymbol{\beta}$$

and $\boldsymbol{Z}_1 = \boldsymbol{I}_n \otimes \boldsymbol{e}_8$ for σ^2_γ, $\boldsymbol{Z}_2 = \boldsymbol{I}_n \otimes \boldsymbol{I}_2 \otimes \boldsymbol{e}_4$ for σ^2_λ, $\boldsymbol{Z}_3 = \boldsymbol{I}_n \otimes \boldsymbol{I}_4 \otimes \boldsymbol{e}_2$ for σ^2_δ, so that

$$\boldsymbol{\Sigma} = \sigma^2_\gamma \boldsymbol{J}_8 + \sigma^2_\lambda \boldsymbol{I}_2 \otimes \boldsymbol{J}_4 + \sigma^2_\delta \boldsymbol{I}_4 \otimes \boldsymbol{J}_2 + \sigma^2_\epsilon \boldsymbol{I}_8.$$

Thus, the parameters to be estimated are the means for each type of metallic oxide and the variances associated with lots, samples and chemists. This dataset contains 248 observations. We can then make $n = 31$ independent sub-vectors y_i of size 8. A plot of the responses by sample and chemist is given in Figure 4.10. One may notice that whatever the sample or the chemist, the responses are rather low for lots (observations) numbers 24 and 25 relative to the other lots.

Table 4.9 presents the estimates and standard errors for the CBS–*MM*. The mean effect of the metallic oxide type is significant (p-value of 0.005), and the variances are larger for the lot and the chemist, and smaller for the sample. As a comparison, the REML gives larger estimates for the variance components of the lot and sample, while a smaller estimate for the chemist (results not presented here). An analysis of the Mahalanobis distances reveals that there are a few potential outlying observations (see Figure 4.11). One can see that the REML and CBS–*MM* detect two outliers (observations 24 and 30) and possibly observation 17 as well, while the CBS–*MM* detects two influential observations (observations 12 and 25). The analysis based on the classical Malahanobis distance alone is certainly misleading.

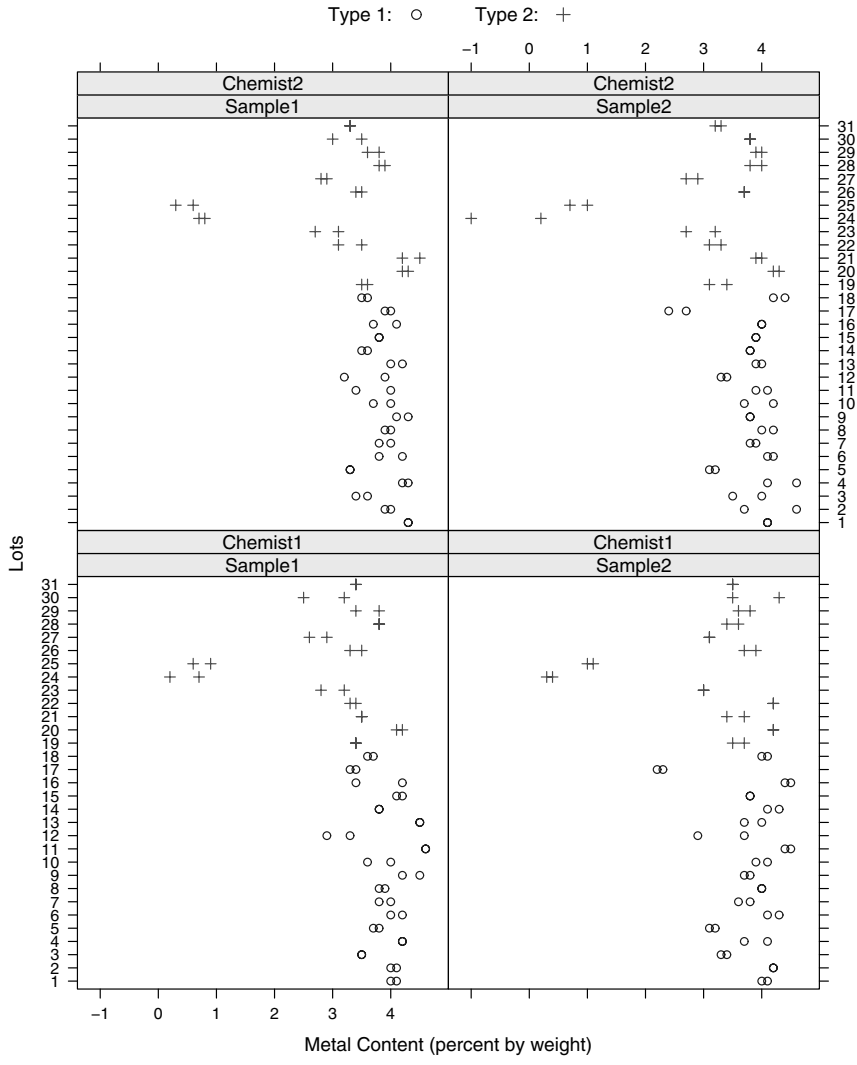

Figure 4.10 Metal content response for each lot by sample and chemist.

4.7.2 Orthodontic Growth Data (continued)

The orthodontic growth data introduced in Section 4.2 are summarized in Figure 4.2 where individual scatterplots of the distance (between the pituitary and the pterygomaxillary fissure) versus age are displayed. Individual LS fits based on simple linear regression are added to each scatterplot. They reveal that the estimated slope for subject M13 is far larger than the other estimated slopes. Overall, it seems that the

Table 4.9 Estimates and standard errors for the CBS–*MM* for the metallic oxide data using model (4.49).

	CBS–*MM*	
Parameter	Estimate (*SE*)	p-value
μ	3.726 (0.066)	$<10^{-4}$
λ	0.184 (0.066)	0.005
σ_{lot}	0.317	
σ_{sample}	0.144	
σ_{chemist}	0.188	
σ_{ϵ}	0.186	

CBS computed with $c_0 = 6.01$ and *MM* (biweight) computed with $c_1 = 6.83$.

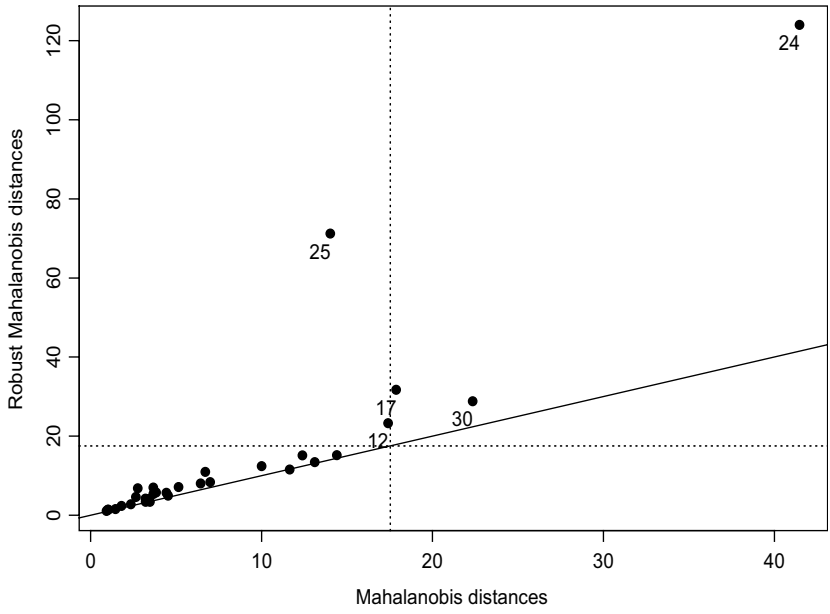

Figure 4.11 Scatterplot of the Mahalanobis distances for the metallic oxide data. CBS computed with $c_0 = 6.01$ and *MM* (biweight) with $c_1 = 6.83$.

responses for the boys vary more than those for the girls. Moreover, the plot suggests that two observations on subject M09 are outliers. These potential outliers are also detected in Figure 4.12 that presents the LS residuals plots by gender.

As discussed in Section 4.2, a plausible working model is thought to be

$$y_{ijt} = \beta_0 + \beta_1 t + (\beta_{0g} + \beta_{1g} t) J_i(j) + \gamma_{0i} + \gamma_{1i} t + \epsilon_{ijt} \tag{4.50}$$

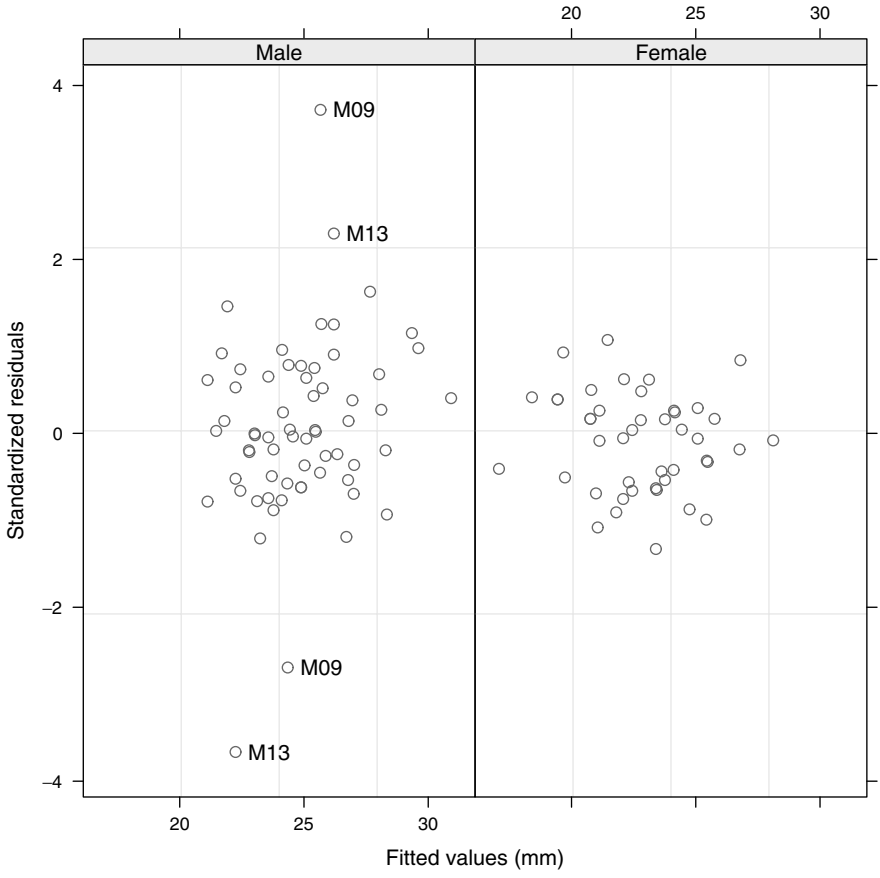

Figure 4.12 Residuals versus fitted values by gender, corresponding to individual LS fits.

with y_{ijt} the response for the ith subject ($i = 1, \ldots, 27$) of gender j ($j = 1$ for boys and $j = 2$ for girls) at age $t = 8, 10, 12, 14$, and $J_i(j) = 0$ for boys ($j = 1$) and 1 for girls ($j = 2$).

Table 4.10 presents the CBS–*MM* estimates and standard errors for the model parameters. The estimates show that the there is no significant mean intercept difference between boys and girls (p-value of 0.896), while there is a significant mean slope difference (p-value of 0.036). The random slope variance is found to be relatively small compared with the random intercept variance. As a comparison, the REML gives similar results, with a larger random slope variance estimate and residual variance. The robust Mahalanobis distances detect observations corresponding to the 9th and 13th boys as extreme, as was already found in the graphical data

Table 4.10 Estimates and standard errors for the CBS–*MM* for the orthodontic data using model (4.50).

	CBS–*MM*	
Parameter	Estimate (*SE*)	p-value
β_0	17.395 (0.613)	$<10^{-4}$
β_{0g}	0.080 (0.613)	0.896
β_1	0.581 (0.052)	0.000
β_{1g}	−0.110 (0.052)	0.036
σ_{γ_0}	1.584	
σ_{γ_1}	0.115	
σ_ϵ	1.04	

CBS computed with $c_0 = 4.09$ and *MM* (biweight) computed with $c_1 = 5.82$.

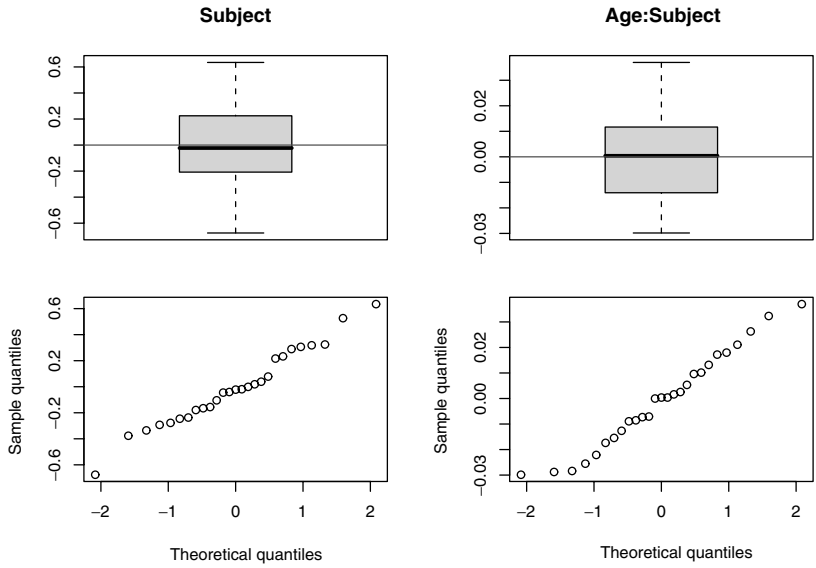

Figure 4.13 Boxplot and Q-Q plot of the random effects for the orthodontic data. CBS computed with $c_0 = 4.09$ and *MM* (biweight) with $c_1 = 5.82$.

analysis in Figure 4.2. It should be noted that Pinheiro *et al.* (2001) also find the same outlying observations.

A plot of the estimated random effects (see Figure 4.13) shows that both the random slope and the random intercept estimated with the robust estimator are normally distributed.

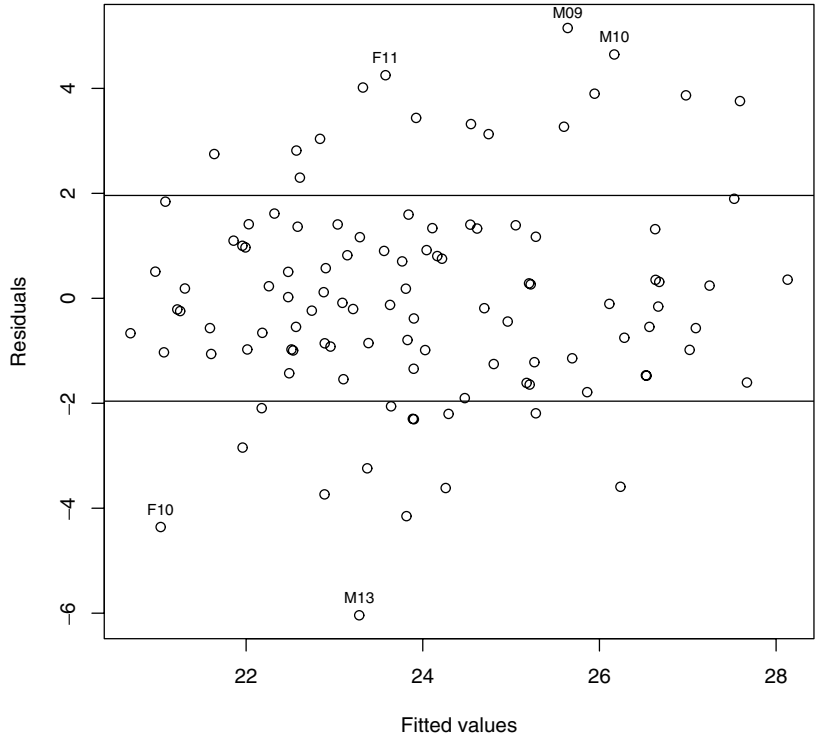

Figure 4.14 Standardized residuals (subject level) versus fitted values for the orthodontic data. CBS computed with $c_0 = 4.09$ and *MM* (biweight) with $c_1 = 5.82$.

Figure 4.14 displays the standardized (Pearson) residuals versus fitted values at the subject level. We can see that there is no particular structure in the residuals and that subjects M13 and M09 are the largest outliers.

4.8 Discussion and Extensions

Despite its good robustness properties and the fact that it does not suffer from computational problems when applied to complex data structures (as is often the case when modeling longitudinal data with fixed covariates), the CBS-MM estimator has a few limitations. The first limitation is, as stated earlier in this chapter, that the CBS-MM estimator cannot handle unbalanced data at the moment unlike the very general bounded influence approach of Richardson and Welsh (1995). This is particularly annoying as balanced data are usually not the rule and one is more likely to encounter unbalanced data especially in medical research.

The second limitation is the lack of inference theory for the variance components. We have seen (see Sections 4.3.2 and 4.5.5) that no proper solution to this problem

exists in the current robustness theory. Robust inferential procedures presented in this book fail as they all assume the null hypothesis to be an interior point of the parameter space. In addition, the robust LRT test defined in Section 4.5 targets only hypotheses on the fixed effects. It even cannot be defined for a simple testing problem on the variance parameters σ_j^2, e.g. testing the equality $\sigma_j^2 = \sigma_{j'}^2$. Its extension to more general hypotheses on $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)^T$ may be proved challenging. In general, further research work is needed in this area.

One possible robust extension of the MLM is to assume that the data follows a t distribution instead of the normal distribution assumed throughout this chapter. For example, Pinheiro *et al.* (2001) incorporate multivariate t distributed random components for the t MLMs. More recently Lin and Lee (2006) propose a model based on multivariate t distribution for autocorrelated longitudinal data by incorporating first an autoregressive dependence structure in the variance components and extend the work of Pinheiro *et al.* (2001) to allow for inference about the random effects and predictions.

The next natural extension of robustness in the MLM environment is to extend it to the class of generalized linear mixed models (GLMMs). Yau and Kuk (2002) introduce robust maximum quasi-likelihood and residuals maximum quasi-likelihood estimation to limit the influence of outlying observations. The way they introduce robustness in the GLMM follows the same line of thoughts as used by Richardson and Welsh (1995) in the MLM. Other attempts at robustifying the GLMM can be found in Mills *et al.* (2002) or Sinha (2004). More recently Litière *et al.* (2007a) study the impact of an incorrectly specified probability model on the maximum likelihood estimation in GLMM. They study the impact of misspecifying the random-effects distribution on the estimation and inference and show that the MLE are inconsistent in the presence of such misspecifications.

5

Generalized Linear Models

5.1 Introduction

The framework of GLMs allows us to extend the class of models considered in Chapter 3 and to address situations with non-normal (non-Gaussian) responses. In particular, it allows us to consider continuous and discrete distributions for the response, both symmetric and asymmetric. From the practical point of view, this unified framework opens many perspectives formalized under the same setting and sharing a number of properties. The fields of application are quite wide: certainly biostatistics, but also medicine, economics, ecology, demography, psychology and many more. The family of possible distributions for the response is quite large, but the most common settings with no doubts include binary or binomial responses (e.g. presence or absence of a characteristic, see the example in Section 5.5, or the number of 'successes' in a sequence), count data (for example, the number of visits to the doctor, see the example in Section 5.6) and positive responses (e.g. hospital costs, see the example in Section 5.3.5).

All of the classical theory of GLMs is likelihood based, and the gain in popularity of GLMs has helped in reinforcing the central role of the likelihood in statistical inference. We will see that the robust versions of GLM presented in this chapter move away from the likelihood setting, but retain almost all of its advantages in terms of statistical properties and interpretation.

The route to the definition of the unified class of GLMs has been long and the steps to it went through multiple linear regression (Legendre, Gauss, early 19th century), the ANOVA of designed experiments (Fisher: 1920–1935), the likelihood function (Fisher, 1922), dilution assay (Fisher, 1922), the exponential family of distributions (Fisher, 1934), the probit analysis (Bliss, 1935), the logit models for proportions (Berkson, 1944; Dyke and Patterson, 1952), the item analysis (Rasch, 1960), log-linear models for counts (Birch, 1963) and inverse polynomials (Nelder 1966; see McCullagh and Nelder (1989, Chapter 1), for additional information). Nelder and

Robust Methods in Biostatistics S. Heritier, E. Cantoni, S. Copt and M.-P. Victoria-Feser

Wedderburn (1972) show that the above problems can all be treated in the same way. They also show that the MLE for all of these models can be obtained using the same algorithm (IRWLS; see Appendix E.3).

Binary logistic regression has received quite a lot of attention in the robust literature. In fact, one can find several robust contributions that follow different approaches: the early contributions of Pregibon (1982), Copas (1988) and Carroll and Pederson (1993), the L_1-norm quasi-likelihood approach of Morgenthaler (1992), the weighted likelihood approaches of Markatou *et al.* (1997) and Victoria-Feser (2002), and the high breakdown approaches of Bianco and Yohai (1997) and Christmann (1997). This wide contribution is certainly due to the fact that addressing the binary framework is simpler than addressing the general GLM class. This more general class has nevertheless been addressed with the work of Stefanski *et al.* (1986) and Künsch *et al.* (1989), who derive optimal (OBRE, see Section 2.3.1) and conditionally unbiased estimators for the entire GLM class. This theory is quite complex (even in its simpler conditional approach) and only the case of logistic regression can be implemented easily. More recently, Cantoni and Ronchetti (2001b) define Huber and Mallows-type estimators and quasi-deviance functions for application within the GLM framework, see also Cantoni (2003, 2004a) and Cantoni and Ronchetti (2006). Here we present this last piece of work which seems to us the most promising for use in the entire GLM class. In fact, it has the advantage over other proposals of having computationally tractable expressions (that allow us to consider the entire class of GLM and not only the logistic application) and of jointly providing a solution to the variable selection question through the definition of quasi-deviance functions.

The present chapter is organized as follows. In Section 5.2 we set up the notation and define the model. We continue in Section 5.3 where we define the class of (robust) estimators and give their properties. The technique is illustrated on a real example in Section 5.3.5. The variable selection issue is addressed in Section 5.4.2 where a family of quasi-deviance functions are defined and its distribution studied. Section 5.4.3 considers the application to the previous studied example. Two additional complete data analyses with robust model selection are presented in Sections 5.5 and 5.6. Finally, Section 5.7 discusses the possible extensions of this work.

5.2 The GLM

5.2.1 Model Building

We introduce here the GLM modeling approach without necessarily giving a complete and exhaustive treatment of the subject. Instead, we refer the interested reader to the general references treating GLM modeling, which include Dobson (2001) (a good starting point for beginners), Lindsey (1997) (an applied approach), McCullagh and Nelder (1989) (with additional technical details) and Fahrmeir and Tutz (2001) (more focused on discrete data).

Table 5.1 Properties of some distributions belonging to the exponential family.

Distribution		$\theta_i(\mu_i)$	ϕ	$E[y_i]$	$\text{var}(y_i)$
Normal	$\mathcal{N}(\mu_i, \sigma^2)$	μ_i	σ^2	$\mu_i = \theta_i$	σ^2
Bernoulli	$\mathcal{B}(1, p_i)$	$\log(p_i/(1-p_i))$	1	$p_i = \dfrac{\exp(\theta_i)}{1+\exp(\theta_i)}$	$p_i(1-p_i)$
Scaled binomial	$\mathcal{B}(m, p_i)/m$	$\log(p_i/(1-p_i))$	$1/m$	$p_i = \dfrac{\exp(\theta_i)}{1+\exp(\theta_i)}$	$p_i(1-p_i)$
Poisson	$\mathcal{P}(\lambda_i)$	$\log(\lambda_i)$	1	$\lambda_i = \exp(\theta_i)$	λ_i
Gamma	$\Gamma(\mu_i, \nu)$	$-1/\mu_i$	$1/\nu$	$\mu_i = -1/\theta_i$	μ_i^2/ν

See Appendix D for the distributions definitions.

Consider a sample of n individuals, for which we define the three following ingredients.

- **The random component.** n independent random variables $y_1, \ldots, y_n$ which are assumed to share the same distribution from the exponential family, that is with density that can be written as

$$f(y_i; \theta_i, \phi) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right] \tag{5.1}$$

  for some specific functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$. We denote $\mu_i = E[y_i]$ and $\text{var}(y_i) = \phi v_{\mu_i}$, where the specific form of v_{μ_i} depends on the distributional assumption on y_i, see the last column of Table 5.1.

  The most common families of distributions such as the normal, the binomial, the Poisson, the exponential, and the Gamma belong to the exponential family of distributions. Some of these distributions will be considered more closely here.

  The parameter θ_i, which is a function of μ_i, is called the natural parameter and ϕ is an additional scale or dispersion parameter, usually considered as a nuisance parameter. We note that ϕ is a constant in certain models (for example, $\phi = 1/m$ for the scaled binomial and $\phi = 1$ for the Poisson distribution), and coincides with σ^2 in the normal model, see the fourth column of Table 5.1.

- **The systematic component.** A set of parameter $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \ldots, \beta_q)$ and q explanatory variables or covariates that can either be quantitative (numerical) or qualitative (levels of a factor, then coded with dummy variables as in linear regression). For each individual $i = 1, \ldots, n$, the covariates are stored in the vector $\boldsymbol{x}_i^T = (1, x_{i1}, \ldots, x_{iq})$, from which the linear predictor $\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$ is constructed. The parameter β_0 therefore identifies the intercept. The pooled

covariate information is collected in a design matrix $\mathbf{X}$ as follows:

$$\mathbf{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_n^T \end{pmatrix}. \tag{5.2}$$

As in the linear model, linearity in GLM is intended with respect to the parameters. We note that one could introduce transformed covariates, $\log(x_{ij})$ or x_{ij}^2, for example, as well as interactions. Moreover, there are situations where a parameter β_j is known *a priori*: the corresponding term in the linear structure is called an offset in the GLM terminology.

- **The link.** A monotone link function g which links the random and the systematic components of the model

$$g(\mu_i) = \eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}. \tag{5.3}$$

The link function defines the form of the relationship between the mean μ_i of the response and the assumed linear predictor η_i. It needs to be monotonic and differentiable. Moreover, it can be chosen to ensure that the estimated parameter lies in the admissible space of values (for example, the interval $(0, 1)$ for the binomial distribution and $(0, \infty)$ for the Poisson distribution). The natural or canonical link function is that relating the natural parameter directly to the linear predictor ($\theta_i = \theta_i(\mu_i) = \eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$). Models making use of the canonical link enjoy convenient mathematical and statistical properties, but the canonical link can be easily replaced with a more appropriate link function from the practical or interpretation point of view (see Example 5.3.5).

The definition of model (5.3) may be surprising at first to people used essentially to the linear model setting, but the connection with the linear model appears more evident when this latter (as defined in (3.1)) is rewritten in the equivalent form

$$E[y_i] = \mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta},$$

with $y_i \sim \mathcal{N}(\mu_i, \sigma^2)$. In this case, the link function is the identity function. In the GLM setting, the distributional assumptions are defined with respect to the response itself (conditionally on the set of explanatory variables) and not with respect to an additive error term. Table 5.1 provides an overview of the components of a GLM model for the most common situations.

5.2.2 Classical Estimation and Inference for GLM

The parameters of model (5.3) are usually estimated by maximizing the corresponding log-likelihood (with respect to $\boldsymbol{\beta}$)

$$l(\boldsymbol{\beta}; \mathbf{y}) = l(\boldsymbol{\mu}; \mathbf{y}) = \log\left(\prod_{i=1}^{n} f(y_i; \theta_i, \phi)\right)$$
$$= \sum_{i=1}^{n}\left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right] = \sum_{i=1}^{n} l_i(\mu_i; y_i), \tag{5.4}$$

where $\mu_i = g^{-1}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$ and $\theta_i = \theta_i(\mu_i) = \theta_i(g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta}))$ are functions of $\boldsymbol{\beta}$.

The maximization of the log-likelihood (5.4) is performed numerically, either directly or via an IRWLS, see McCullagh and Nelder (1989, Section 2.5) and Appendix E.3. The resulting estimator $\hat{\boldsymbol{\beta}}_{[MLE]}$ enjoys the general properties of maximum likelihood estimation, in particular the normal asymptotic distribution with variance given by the inverse of the Fisher information matrix $I(\boldsymbol{\beta})$ (see (2.30)), that is $\sqrt{n}(\hat{\boldsymbol{\beta}}_{[MLE]} - \boldsymbol{\beta}) \sim \mathcal{N}(0, I^{-1}(\boldsymbol{\beta}))$.

Based on this asymptotic result, one can construct univariate test statistics for the coefficients β_j, $j = 0, \ldots, q$ as

$$\frac{\hat{\beta}_{[MLE]j}}{SE(\hat{\beta}_{[MLE]j})} \tag{5.5}$$

with

$$SE(\hat{\beta}_{[MLE]j}) = \sqrt{\frac{1}{n}[\hat{I}^{-1}(\hat{\boldsymbol{\beta}}_{[MLE]})]_{(j+1)(j+1)}},$$

and using an estimator $\hat{I}$ for the Fisher information matrix

$$\hat{I}(\hat{\boldsymbol{\beta}}_{[MLE]}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\beta}} l_i(\mu_i; y_i)\frac{\partial}{\partial\boldsymbol{\beta}^T} l_i(\mu_i; y_i)\bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{[MLE]}}.$$

The statistic (5.5) is labeled the t-statistic if the dispersion parameter (ϕ) is estimated (for example, for the Gaussian and Gamma distributions), and is labelled the z-statistic if the dispersion parameter is known (for example, for the binomial and Poisson distributions). The test statistic (5.5) has a $t_{n-(q+1)}$ distribution under the null hypothesis $H_0 : \beta_j = 0$ in the first case and the standard normal in the second. The p-value for a two-sided alternative hypothesis $H_1 : \beta_j \neq 0$ is therefore computed as $P(|z\text{-statistic}| > |z_{\text{obs}}|) = 2(1 - \Phi(|z_{\text{obs}}|))$ or $P(|t\text{-statistic}| > |t_{\text{obs}}|) = 2(1 - t_{n-(q+1)}(|t_{\text{obs}}|))$, where z_{obs} and t_{obs} are the values taken by the statistic (5.5) on the sample.

Note that the z/t-statistic is a Wald approximation of the log-likelihood (second-order Taylor expansion of the log-likelihood at the MLE) to test $H_0 : \beta_j = 0$ and is sometimes misleading with binomial GLMs. In fact, a small value for the z/t-statistic can either correspond to a small LRT statistic or to a situation where $|\beta_j|$ is large, the

Wald approximation is poor and the likelihood ratio statistic is large. These problems can occur in cases when the fitted probabilities are extremely close to zero or one. This is called the Hauck–Donner phenomenon, see Hauck and Donner (1977).

The asymptotic result is also useful in constructing approximate $(1-\alpha)$ confidence intervals (CIs), according to the formula

$$(\hat{\beta}_{[MLE]j} - q_{(1-\alpha/2)}SE(\hat{\beta}_{[MLE]j}); \hat{\beta}_{[MLE]j} + q_{(1-\alpha/2)}SE(\hat{\beta}_{[MLE]j})),$$

where $q_{(1-\alpha/2)}$ is either the $(1-\alpha/2)$ quantile of the standard normal distribution or of the $t_{n-(q+1)}$ distribution, depending on whether ϕ is known or not.

For binomial and Poisson models, it sometimes happens that data do not satisfy the variance assumption of the model, but rather that $\text{var}(y_i) = \tau v_{\mu_i}$ (recall that $\phi = 1$ for binomial and Poisson models). This phenomenon is called over- or under-dispersion depending on whether τ is larger or smaller than one. One of the main reasons for over-dispersion is clustering in the population (the parameter θ_i varies from cluster to cluster, as a function of cluster size for example). This means that the parameter θ_i is regarded as random rather than fixed. Beyond normality, specifying the expectation and the variance structure separately (first and second moment) does not correspond to a distribution function, therefore preventing the definition of a likelihood function. In this case, the model is fitted via the estimating equations

$$\sum_{i=1}^{n}\left(\frac{y_i - \mu_i}{\tau v_{\mu_i}}\right)\mu_i' = \mathbf{0}, \tag{5.6}$$

where $\mu_i' = \partial\mu_i/\partial\boldsymbol{\beta}$.

Equation (5.6) corresponds to the maximization of the so-called quasi-likelihood function

$$Q(\boldsymbol{\mu}; \mathbf{y}) = \sum_{i=1}^{n} Q(\mu_i; y_i) = \sum_{i=1}^{n}\int_{y_i}^{\mu_i}\frac{y_i - t}{\tau v_t}\,dt, \tag{5.7}$$

where $\boldsymbol{\mu}^T = (\mu_1, \ldots, \mu_n)$ and $\mathbf{y}^T = (y_1, \ldots, y_n)$. Under some general conditions (see Wedderburn, 1974) the quasi-likelihood estimator is asymptotically normally distributed. Moreover, the MLE and the maximum quasi-likelihood estimator (MQLE) are the same for all of the models of the one-parameter exponential family (binomial and Poisson, for example).

Note that τ has no impact on (5.6) because it cancels out, but does have an impact on the computation of the standard errors of the coefficients. The estimation of τ is based on the RSS as follows

$$\hat{\tau} = \frac{1}{n-(q+1)}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{v_{\hat{\mu}_i}},$$

where $\hat{\mu}_i$ are the fitted values $g^{-1}(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{[MQLE]})$ on the response scale. The estimator $\hat{\tau}$ is an unbiased estimator of τ if the fitted model is correct.

A particular function based on the log-likelihood plays an important role in GLM modeling. It is called the deviance, which, assuming that $a_i(\phi)$ in (5.1) can be

decomposed as ϕ/w_i, is defined by

$$D(\hat{\boldsymbol{\mu}}; \mathbf{y}) = 2\phi[l(\mathbf{y}; \mathbf{y}) - l(\hat{\boldsymbol{\mu}}; \mathbf{y})] = \sum_{i=1}^{n} 2\phi[l_i(y_i; y_i) - l_i(\hat{\mu}_i; y_i)] = \sum_{i=1}^{n} \phi\, d_i, \tag{5.8}$$

where $\hat{\boldsymbol{\mu}}$ is the vector of fitted values $g^{-1}(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{[MLE]})$ and where $l(\hat{\boldsymbol{\mu}}; \mathbf{y})$ is the log-likelihood of the postulated model and $l(\mathbf{y}; \mathbf{y})$ is the saturated log-likelihood for a full model with n parameters. The deviance measures the discrepancy between the performance of the current model via its log-likelihood and the maximum log-likelihood achievable. It can therefore be used for goodness-of-fit purposes. Large values of $D(\hat{\boldsymbol{\mu}}; \mathbf{y})$ indicate that the model is not good. On the other hand, small values of $D(\hat{\boldsymbol{\mu}}; \mathbf{y})$ arise when the log-likelihood $l(\hat{\boldsymbol{\mu}}; \mathbf{y})$ is close to the saturated log-likelihood $l(\mathbf{y}; \mathbf{y})$.

The distribution of the deviance is exactly $\chi^2_{n-(q+1)}$ for normally distributed responses, and this distribution can be taken as an approximation for other distributions, for example binomial and Poisson. However, $D(\hat{\boldsymbol{\mu}}; \mathbf{y})$ is not usable for goodness-of-fit for Bernoulli responses, because it only depends on the observations $\mathbf{y}$ through the fitted probabilities $\hat{\boldsymbol{\mu}}$ and as such does not carry information about the agreement between the observations and the fitted probabilities (see Collett 2003a, Section 3.8.2). The deviance can be regarded as a LRT statistic for testing a specific model within the saturated model, assuming $\phi = 1$. This is the case for binomial and Poisson models, but for other distributions, e.g. normal or Gamma, the deviance is not directly related to a LRT statistic.

The deviance is also used to construct a difference of deviance statistics to compare nested models. Suppose that a model $\mathcal{M}_{q-k+1}$ with $(q-k)$ explanatory variables (plus intercept) is nested into a larger model $\mathcal{M}_{q+1}$ with q explanatory variables (plus intercept). To test the null hypothesis, which states that the smallest model suffices to describe the data, one can test whether the parameters associated with the variables not included in the smallest model are equal to zero with the test statistic

$$\Delta D(\hat{\boldsymbol{\mu}}, \dot{\boldsymbol{\mu}}) = D(\dot{\boldsymbol{\mu}}; \mathbf{y}) - D(\hat{\boldsymbol{\mu}}; \mathbf{y}) = 2\phi[l(\hat{\boldsymbol{\mu}}; \mathbf{y}) - l(\dot{\boldsymbol{\mu}}; \mathbf{y})], \tag{5.9}$$

where $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}_{[MLE]})$ and $\dot{\boldsymbol{\mu}} = \boldsymbol{\mu}(\dot{\boldsymbol{\beta}}_{[MLE]})$ are the MLE estimates in the full model $\mathcal{M}_{q+1}$ and the reduced model $\mathcal{M}_{q-k+1}$, respectively.

If ϕ is known and, under the null hypothesis that the smaller model is good enough to represent the data, the distribution of $\Delta D(\hat{\boldsymbol{\mu}}, \dot{\boldsymbol{\mu}})$ can be approximated by a $\phi\ \chi^2_k$ (it is the LRT statistic up to a factor ϕ). This approximation is more accurate than the approximation of the deviance itself by a $\chi^2_{n-(q+1)}$ distribution. When ϕ is not known (e.g. normal, Gamma) the usual approximation under H_0 uses an F-type statistic:

$$\frac{(D(\dot{\boldsymbol{\mu}}; \mathbf{y}) - D(\hat{\boldsymbol{\mu}}; \mathbf{y}))/k}{\hat{\phi}} \sim F_{k,n-(q+1)},$$

where $\hat{\phi} = D(\hat{\boldsymbol{\mu}}; \mathbf{y})/(n-(q+1))$. Note that for the normal case with identity link this is an exact result, but for the Gamma model the accuracy of this approximation is not well known.

A natural definition of a quasi-deviance function follows from the definition (5.7) of a quasi-likelihood function:

$$\mathrm{QD}(\hat{\boldsymbol{\mu}}; \mathbf{y}) = Q(\mathbf{y}; \mathbf{y}) - Q(\hat{\boldsymbol{\mu}}; \mathbf{y}). \tag{5.10}$$

By analogy with the deviance function, one can use the quasi-deviance function for inference purposes to test whether a smaller model $\mathcal{M}_{q-k+1}$ nested into a larger model $\mathcal{M}_{q+1}$ is a good enough representation of the data with the difference of quasi-deviances statistics:

$$\Delta\mathrm{QD}(\hat{\boldsymbol{\mu}}, \dot{\boldsymbol{\mu}}) = Q(\dot{\boldsymbol{\mu}}; \mathbf{y}) - Q(\hat{\boldsymbol{\mu}}; \mathbf{y}), \tag{5.11}$$

where $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}_{[MQLE]})$ and $\dot{\boldsymbol{\mu}} = \boldsymbol{\mu}(\dot{\boldsymbol{\beta}}_{[MQLE]})$ are the MQLE estimates in the full model $\mathcal{M}_{q+1}$ and the reduced model $\mathcal{M}_{q-k+1}$, respectively.

The test statistic $\Delta\mathrm{QD}(\hat{\boldsymbol{\mu}}, \dot{\boldsymbol{\mu}})$ is then compared with a $\chi^2_{n-(q+1)}$ distribution, at least when ϕ is known. As with the likelihood, an F-type test is more appropriate if ϕ is unknown, see above.

5.2.3 Hospital Costs Data Example

We introduce here a dataset on health care expenditures previously analyzed by Marazzi and Yohai (2004) and Cantoni and Ronchetti (2006). The aim is to explain the cost of stay (`cost` in Swiss francs) of 100 patients hospitalized at the Centre Hospitalier Universitaire Vaudois in Lausanne (Switzerland) during 1999 for 'medical back problems' (APDRG 243). The following explanatory variables have been measured: length of stay (`los`, in days), admission type (`adm`: 0 = planned, 1 = emergency), insurance type (`ins`: 0 = regular, 1 = private), age in years (`age`), sex (`sex`: 0 = female, 1 = male) and discharge destination (`dest`: 1 = home, 0 = another health institution). The median age over the 100 patients is 56.5 years (the youngest patient is 16 years old and the oldest is 93 years old). Moreover, 60 individuals out of the 100 in the sample were admitted as emergencies and only 9 patients had private insurance. Also, both sexes are well represented in the sample with 53 men and 47 women. After being treated, 82 patients went home directly. Modeling medical expenses is an important step in cost management and health care policy. Establishing the relationship between the cost and the above explanatory variables could, for example, help in reducing costs in health care expenditures which are increasing extremely fast everywhere and are therefore a matter of concern.

In addition to be positive, cost measurements are known to be highly skewed. Moreover, it is also known that the thickness of the tail of their distribution is often determined by a small number of heavy users. Several authors (e.g. Blough *et al.*, 1999; Gilleskie and Mroz, 2004) report that the variance of health care expenditures data can be considered as proportional to the squared mean. We therefore consider fitting a Gamma GLM model with a logarithmic link. Note that this model can be seen as issued from a multiplicative model $y_i = \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}) \cdot u_i$, where the error term u_i has constant variance. This is the reason why we use the logarithmic link instead of the canonical link $g(\mu_i) = 1/\mu_i$ (the inverse function), which, by the way, does

Table 5.2 Classical estimates for model (5.12).

Variable	Estimate (*SE*)	95% CI	p-value
`intercept`	7.234 (0.147)	(6.940; 7.528)	$<10^{-4}$
log(`los`)	0.822 (0.028)	(0.766; 0.878)	$<10^{-4}$
`adm`	0.214 (0.050)	(0.114; 0.314)	$<10^{-4}$
`ins`	0.093 (0.079)	(−0.065; 0.252)	0.2414
`age`	−0.0005 (0.001)	(−0.003; 0.002)	0.6790
`sex`	0.095 (0.050)	(−0.005; 0.195)	0.0602
`dest`	−0.104 (0.069)	(−0.243; 0.034)	0.1353
$1/\nu$ (scale)	0.0496		

The estimates are obtained by maximum likelihood, see (5.4) (CI, confidence interval).

not guarantee that $\mu_i > 0$. More specifically, we consider a parameterization of the Gamma density function such that one parameter identifies μ_i and the variance structure is defined by $v(\mu_i) = \mu_i^2/\nu$, see the top of page 201 in Cantoni and Ronchetti (2006).

We start by fitting the full model, that is the model with all of the available explanatory variables, as follows

$$\log(E[\texttt{cost}]) = \beta_0 + \beta_1 \log(\texttt{los}) + \beta_2 \texttt{adm} + \beta_3 \texttt{ins} + \beta_4 \texttt{age} + \beta_5 \texttt{sex} + \beta_6 \texttt{dest}. \tag{5.12}$$

The MLE parameter estimates, their standard errors and the p-values of the significance tests (5.5) are given in Table 5.2. Before proceeding with any interpretation, it is recommended to validate the model. In this example, the deviance statistic (5.8) takes the value 5.07, which yields a p-value $P(D > 5.07) \simeq 1$ when compared with a $\chi^2_{n-(q+1)} = \chi^2_{93}$ distribution. This large p-value provides no evidence against the null hypothesis that the postulated model is better than the saturated model.

5.2.4 Residual Analysis

Residual diagnostic plots are an alternative to formal tests. In the GLM setting several types of residuals can be defined, between which the most common are:

- the Pearson residuals $r_{i\mathrm{P}} = (y_i - \hat{\mu}_i)/\sqrt{\hat{\phi} v_{\hat{\mu}_i}}$;
- the standardized Pearson residuals $r_{i\mathrm{PS}} = (y_i - \hat{\mu}_i)/\sqrt{\hat{\phi} v_{\hat{\mu}_i}(1 - h_{ii})}$, where the leverages h_{ii} are the diagonal entries of the hat matrix, see (3.11);
- the deviance residuals $r_{i\mathrm{D}} = \mathrm{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}$;
- the standardized deviance residuals $r_{i\mathrm{DS}} = r_{i\mathrm{D}}/\sqrt{\hat{\phi}(1 - h_{ii})}$.

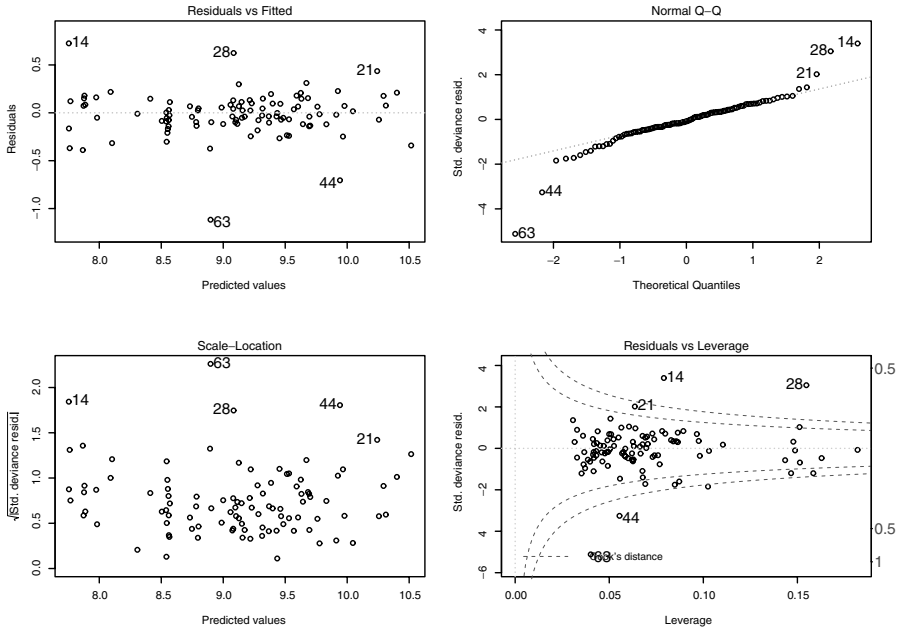

Figure 5.1 Diagnostic plots for the Gamma model (5.12), estimated with a MLE.

Residual plots can help in identifying departures from the linearity assumption (when plotted against continuous covariates), serial correlation (when plotted against the order in which the observations are collected, if known) and particular structures (when plotted against predicted values). In addition, it is usual to look at a Q-Q plot of the residuals against the normal quantiles. Note, that for binary logistic models very often structures appear on the residuals plots which are due to the discrete nature of the response variable but do not indicate fitting problems.

Since the diagnostic approach is based on a classical fit, it has therefore to be used with caution. In fact, masking can occur, where a single large outlier may mask others. It is worth noting that in the GLM setting, an outlier or extreme observation would be an observation $(y_i, \boldsymbol{x}_i^T)$ such that, under the GLM model that fits the majority of the data, y_i is in some sense far from its fitted value $g^{-1}(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})$. The quantity $y_i - g^{-1}(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})$ can be large because y_i is an extreme response and/or the covariates $\boldsymbol{x}_i$ are (at least for one of them) extreme themselves. A classical residual analysis can suffer from the masking effect in that the distorted data appear to be the norm rather than the exception. For instance, consider a regression setting where an outlier may have such a large effect on a slope estimated by a MLE that its residual (or any other measure used for diagnostic) will tend to be small, whereas other observations will have corresponding relatively large residuals. This behavior is due to the fact that classical estimates are affected by outlying points and are pulled in the direction of them. We advocate later for the use of a robust analysis in

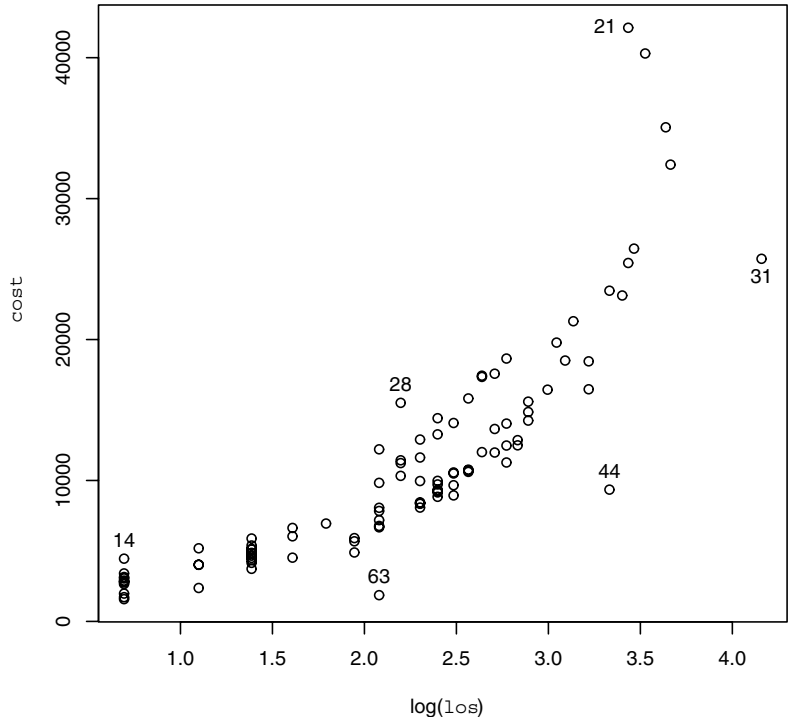

Figure 5.2 `cost` versus log(`los`) for the Gamma example of Section 5.2.3.

the first place (see also the discussion in Section 1.3). We nevertheless propose as a starting point to look at a few plots. In Figure 5.1 we present the diagnostic plots for the fitted Gamma model as per (5.12). In this figure we represent the Pearson residuals as a function of the fitted values (top left panel), a normal Q-Q plot of the standardized deviance residuals (top right panel), a scale–location plot of the standardized deviance residuals as a function of the fitted values (bottom left panel) and a residuals versus leverage plot, that is, a plot of the standardized deviance residuals as a function of the leverage h_{ii} (bottom right panel). This last plot comes with added contour lines of equal Cooks distances (see Cook and Weisberg, 1982). Note that the `plot` function in R can also produce two extra plots, namely the Cook's distances and the Cook's distances as a function of the leverage.

From Figure 5.1, we can see that there seems to be few outlying/influential data points with large residuals, in particular those identified with their observation number. To see why these observations are extreme, one can for example look at the plot of the variable `cost` as a function of the variable log(`los`), as in Figure 5.2. We see from this figure that the points with large residuals are in fact points which are extreme with respect to observations with the same or similar values of log(`los`). Even though the Gamma model admits variance increasing with the covariates (of the

order of $\mu_i^2 = \exp(2\boldsymbol{x}_i^T\boldsymbol{\beta})$), observations 14, 28, 63, 44 and 21 are considered too extreme with respect to the bulk of the data. On the other hand, observation 31 could be a leverage point, but is not otherwise worrying given that its y-value lies in a region covered by the model assumptions.

The more extreme observations identified with this diagnostic analysis can potentially have a very bad impact on the parameter estimates and this issue needs to be investigated further. We reanalyze this dataset in Section 5.6 with a robust technique.

5.3 A Class of M-estimators for GLMs

Deviations from the model can also occur for GLMs. The nature of possible deviations in the GLM class of models are close to what one can see in the regression setting: outliers in the response (producing large residuals) and leverage points in the design space. A notable exception is the binary response setting where deviations in the response space take the form of misclassification (a zero instead than a one, or vice versa), and where the difference between an outlier and a leverage point is less clearcut.

To address the potential problem of deviating points in real data, or more generally the problem of slight model misspecification, we propose here a general class of M-estimators (see Section 2.3.1) for the GLM model as defined in Section 5.2. Given the Pearson residuals $r_i = (y_i - \mu_i)/\sqrt{\phi v_{\mu_i}}$, the M-estimating equations for $\boldsymbol{\beta}$ of model (5.3) are given by the solution of the following estimating equations

$$\sum_{i=1}^{n}\left[\psi(r_i;\boldsymbol{\beta},\phi,c)w(\boldsymbol{x}_i)\frac{1}{\sqrt{\phi v_{\mu_i}}}\mu_i' - a(\boldsymbol{\beta})\right] = \sum_{i=1}^{n}\Psi(y_i,\boldsymbol{x}_i;\boldsymbol{\beta},\phi,c) = \mathbf{0}, \quad (5.13)$$

where $\mu_i' = \partial\mu_i/\partial\boldsymbol{\beta} = \partial\mu_i/\partial\eta_i\boldsymbol{x}_i$ and $a(\boldsymbol{\beta}) = (1/n)\sum_{i=1}^{n}E[\psi(r_i;\boldsymbol{\beta},\phi,c)]w(\boldsymbol{x}_i)/\sqrt{\phi v_{\mu_i}}\ \mu_i'$, with the expectation taken over the distribution of $y_i|\boldsymbol{x}_i$. The constant $a(\boldsymbol{\beta})$ is a correction term to ensure Fisher consistency; see Sections 2.3.2 and 5.3.2.

The function $\psi(r_i;\boldsymbol{\beta},\phi,c)$ and the weights $w(\boldsymbol{x}_i)$ are the new ingredients with respect to the classical GLM estimators obtained by maximum quasi-likelihood: compare with the estimating equations (5.6), which are obtained with $\psi(r_i;\boldsymbol{\beta},\phi,c) = r_i$ and $w(\boldsymbol{x}_i) = 1$ for all i. The function ψ is introduced to control deviations in the y-space and leverage points are downweighted by the weights $w(\boldsymbol{x})$. Conforming to the usage in robust linear regression, we call the estimator issued from (5.13) a Mallows-type estimator. It simplifies to a Huber-type estimator when $w(\boldsymbol{x}_i) = 1$ for all i.

It is worth noting that the estimating equations (5.13) can be conveniently rewritten as

$$\sum_{i=1}^{n}\left[\tilde{w}(r_i;\boldsymbol{\beta},\phi,c)w(\boldsymbol{x}_i)r_i\frac{1}{\sqrt{\phi v_{\mu_i}}}\mu_i' - a(\boldsymbol{\beta})\right] = \mathbf{0}, \quad (5.14)$$

where $\tilde{w}(r;\boldsymbol{\beta},\phi,c) = \psi(r;\boldsymbol{\beta},\phi,c)/r$. In this form, the estimating equations (5.13) can be interpreted as the classical estimating equations weighted (both with respect

to r_i and $\boldsymbol{x}_i$) and re-centered via $a(\boldsymbol{\beta})$ to ensure consistency. The particular weighting scheme considered in (5.14) is multiplicative in its design and residuals components ($w_i = \tilde{w}(r_i; \boldsymbol{\beta}, \phi, c)w(\boldsymbol{x}_i)$). Alternatively, one could consider a global weighting scheme of the form $w_i(r_i, \boldsymbol{x}_i)$, as for example in Künsch *et al.* (1989). It should nevertheless be stressed that such a scheme increases the difficulty in calculating the Fisher consistency correction $a(\boldsymbol{\beta})$.

The estimation procedure issued from (5.13) can be written as an IRWLS, in the same manner as it is usually presented for the classical GLM estimating equations. We give the algorithm in Appendix E.3. The IRWLS algorithm has been a particularly convincing 'selling argument' when GLMs have been proposed. Thanks to this representation, the estimation procedure only requires software that allows the computation of weighted LS (or even only matrix computation). Nowadays computer power is a less crucial issue and other numerical procedures can be considered. For example, one can use a Newton–Raphson or a quasi-Newton algorithm.

Finally, one can see that if we write $\mathbf{y}^T = (y_1, \ldots, y_n)$ and $\boldsymbol{\mu}^T = (\mu_1, \ldots, \mu_n)$, the estimating equations (5.13) correspond to the minimization of the quantity

$$Q_M(\boldsymbol{\mu}; \mathbf{y}) = \sum_{i=1}^{n} Q_M(\mu_i; y_i), \tag{5.15}$$

with respect to $\boldsymbol{\beta}$, where the functions $Q_M(y_i; \mu_i)$ can be written as

$$\begin{aligned} &Q_M(\mu_i; y_i) \\ &\quad = \int_{\tilde{s}}^{\mu_i} \psi\left(\frac{y_i - t}{\phi v_t}; c\right) w(\boldsymbol{x}_i) \frac{1}{\sqrt{\phi v_t}}\, dt \\ &\qquad - \frac{1}{n} \sum_{j=1}^{n} \int_{\tilde{t}}^{\mu_j} E\left[\psi\left(\frac{y_i - t}{\phi v_t}; c\right) w(\boldsymbol{x}_i) \frac{1}{\sqrt{\phi v_t}}\right] dt, \end{aligned} \tag{5.16}$$

with $\tilde{s}$ such that $\psi((y_i - \tilde{s})/(\phi v_{\tilde{s}}); c) = 0$, and $\tilde{t}$ such that $E[\psi((y_i - \tilde{t})/(\phi v_{\tilde{t}}); c)] = 0$. The function $Q_M(\mu_i; y_i)$ in (5.16) plays the same role as the function $Q(\mu_i; y_i)$ in (5.7), and is used later to define a difference of quasi-deviance type statistic, see Section 5.4.2.

5.3.1 Choice of ψ and $w(x)$

The role of the function ψ is to control the effect of large residuals, therefore it has to be bounded. Common choices for ψ are functions that level off such as the Huber function or functions that are redescending, see Section 2.3.1 for a discussion of the possible options. The function ψ is usually tuned with a constant c, which is typically chosen to guarantee a given level of asymptotic efficiency (which is computed as the ratio of traces of the asymptotic variances of the classical and the robust estimators, see, for example, (2.31)). The exact computation of the value of c that guarantees a certain level of efficiency in GLM models is more complicated than in linear regression because the asymptotic efficiency also depends here on the design and no

general result can be derived. It is always possible to inspect the estimated efficiency *a posteriori* and refit the model with a different value of c if it is not satisfactory. In practice, if the Huber ψ-function is used (and this is the case in the `glmrob` function of the `robustbase` R package and therefore in our examples), a value of c between 1.2 and 1.8 is often adequate. The default value is set to 1.345, the value that guarantees 95% efficiency for the normal-identity link GLM model. This value is also often a reasonable choice for the other models, such as the binomial and Poisson models. Note that when $c \to \infty$, the classical GLM estimators are reproduced. In practice, very large values of c (e.g. ≥ 10) have the same effect.

The choice of $w(\boldsymbol{x}_i)$ is also suggested by robust estimators in linear models: the simplest approach is to use $w(\boldsymbol{x}_i) = \sqrt{1 - h_{ii}}$, where h_{ii} is the leverage. More sophisticated choices for $w(\boldsymbol{x}_i)$ are available, in particular some that in addition do have high breakdown properties (see Section 3.2.4 for linear regression). The current implementation of the `robustbase` package in addition to equal weights ($w(\boldsymbol{x}_i) = 1$, for all i, the default) and $w(\boldsymbol{x}_i) = \sqrt{1 - h_{ii}}$, allows one to choose weights based on the Mahalanobis distances d_i (see (2.34)) of the form

$$w(\boldsymbol{x}_i) = \frac{1}{\sqrt{1 + 8\ \max(0, (d_i^2 - q)/\sqrt{2q})}}.$$

A few options are available to estimate the center and the scatter in d_i robustly, either by the MCD estimator of Rousseeuw (1984) or a more efficient S-estimator, see Section 2.3.3. Note, however, that these high breakdown estimators are not well suited for categorical or binary covariates, and their use only makes sense if all of the explanatory variables are continuous. A variation of this kind of weights is given in Victoria-Feser (2002).

The weighting scheme issued from a robust fitting procedure can be used for diagnostic purposes. In fact, inspecting the observations that received a low weight allows the user to identify the outlying observations. For an illustration, see Section 5.5 (Figure 5.3) and Section 5.6 (Figure 5.7).

5.3.2 Fisher Consistency Correction

The term $a(\boldsymbol{\beta})$ in the estimating equations (5.13) guarantees that the estimator is Fisher consistent, that is, asymptotically unbiased under the postulated model (normal, binomial, etc.). This term can sometimes be difficult to compute. Note, however, that it can be computed explicitly for GLM models where the responses are binomial and Poisson (cf. Cantoni and Ronchetti (2001b, p. 1028) with the change in notation $V(\mu_i) = \phi v_{\mu_i}$), and Gamma (see Cantoni and Ronchetti (2006, pp. 210–211) with the change in notation $v(\mu_i) = \phi v_{\mu_i}$). The expression of $a(\boldsymbol{\beta})$ for these models in the unified notation of this book are given in Appendix E.1.

When $a(\boldsymbol{\beta})$ cannot be computed analytically, its estimation by simulation can be considered: the expectation involved in its computation is replaced by the empirical mean of a simulated sample.[1]

A different strategy is to compute a simpler biased estimator of $\boldsymbol{\beta}$ by solving the uncorrected estimating equations

$$\sum_{i=1}^{n} \psi(r_i; \boldsymbol{\beta}, \phi, c) w(\boldsymbol{x}_i) \frac{1}{\sqrt{\phi v_{\mu_i}}} \mu_i' = \sum_{i=1}^{n} \widetilde{\Psi}(y_i, \boldsymbol{x}_i; \boldsymbol{\beta}, \phi, c) = \mathbf{0} \tag{5.17}$$

and correct the bias *a posteriori*. In fact, the asymptotic bias of the estimator solving (5.17) can be approximated by a Taylor expansion and takes the form

$$-E\left[\frac{\partial \sum_{i=1}^{n} \widetilde{\Psi}(y_i, \boldsymbol{x}_i; \boldsymbol{\beta}, \phi, c)}{\partial \boldsymbol{\beta}}\right]^{-1} E\left[\sum_{i=1}^{n} \widetilde{\Psi}(y_i, \boldsymbol{x}_i; \boldsymbol{\beta}, \phi, c)\right]. \tag{5.18}$$

This bias has to be estimated. One can either compute the expectations by numerical integration and evaluate them at $\tilde{\boldsymbol{\beta}}$ (solution of (5.17)), or replace expectations with averages with respect to the data. Given that $\sum_{i=1}^{n} \widetilde{\Psi}(y_i, \boldsymbol{x}_i; \boldsymbol{\beta}, \phi, c)$ evaluated at the solution $\tilde{\boldsymbol{\beta}}$ of (5.17) equals zero, a robust pilot estimator, that is a robust estimator obtained by other means, is needed. For further details on the comparison of the estimator obtained from (5.17)–(5.18) and the estimator obtained from (5.13), see Dupuis and Morgenthaler (2002), in particular their Section 2.2.

Using indirect inference (Gallant and Tauchen, 1996; Gouriéroux *et al.*, 1993) is another possible approach that can be implemented to correct the bias *a posteriori* as is done in e.g. Moustaki and Victoria-Feser (2006). For illustrations of the use of indirect inference with robust estimators, see also Genton and Ronchetti (2003).

5.3.3 Nuisance Parameters Estimation

As stated previously, ϕ is known to be constant for Bernoulli, (scaled) binomial and Poisson models. In other models, this parameter has to be estimated, and this should be done by paying attention to maintaining the robustness properties gained in the estimation of $\boldsymbol{\beta}$. In other words, it is necessary to also use a robust estimator for ϕ.

We address here the normal and the Gamma distribution settings. In both cases the nuisance parameter is a scale parameter (for the Gamma, one may notice that $\text{var}((y_i - \mu_i)/\mu_i) = \nu$), and we suggest borrowing one of the robust scale estimators available in the literature. Namely, we propose to use the Huber's Proposal 2 estimator (Huber, 1981, p. 137) defined by (see also (3.7) for the regression model)

$$\sum_{i=1}^{n} \chi\left(\frac{y_i - \mu_i}{\sqrt{\phi v_{\mu_i}}}; \boldsymbol{\beta}, \phi, c\right) = 0, \tag{5.19}$$

where $\chi(u; \boldsymbol{\beta}, \phi, c) = \psi^2(u; \boldsymbol{\beta}, \phi, c) - \delta$, and $\delta = E[\psi^2(u; \boldsymbol{\beta}, \phi, c)]$ is a constant that ensures Fisher consistency for the estimation of ϕ, see Hampel *et al.* (1986,

[1] Care should be taken that in the iterative estimation process, the value of $\boldsymbol{\beta}$ used to simulate the data is not equal to the current value of $\hat{\boldsymbol{\beta}}$.

p. 234). The function ψ can be chosen to be the same as that in (5.13).[2] The expectation in δ is computed under normality for u, see (3.8) for its computation for $\psi^2(u; \boldsymbol{\beta}, \phi, c) = \psi^2_{[Hub]}(u; \boldsymbol{\beta}, \phi, c) = u^2 w^2_{[Hub]}(u; \boldsymbol{\beta}, \phi, c)$.

Ideally, (5.19) has to be solved simultaneously with (5.13), but in practice a two-step procedure is often used. Starting from a first guess for ϕ, an estimate of $\boldsymbol{\beta}$ is obtained, which in turn is used in (5.19), and so on until convergence.

5.3.4 *IF* and Asymptotic Properties

The estimator defined by (5.13) is an M-estimator $\hat{\boldsymbol{\beta}}_{[M]}$ characterized by the Ψ-function $\Psi(y_i, \boldsymbol{x}_i; \boldsymbol{\beta}, \phi, c) = \psi(r_i; \boldsymbol{\beta}, \phi, c) w(\boldsymbol{x}_i) / \sqrt{\phi v_{\mu_i}}\ \mu_i' - a(\boldsymbol{\beta})$. Its *IF* is then

$$IF(y, \boldsymbol{x}; \hat{\boldsymbol{\beta}}, F_{\boldsymbol{\beta}}) = M(\Psi, F_{\boldsymbol{\beta}})^{-1} \Psi(y, \boldsymbol{x}; \boldsymbol{\beta}, \phi, c), \tag{5.20}$$

where $M(\Psi, F_{\boldsymbol{\beta}}) = -E[(\partial/\partial \boldsymbol{\beta}) \Psi(y, \boldsymbol{x}; \boldsymbol{\beta}, \phi, c)]$. Moreover, $\sqrt{n}(\hat{\boldsymbol{\beta}}_{[M]} - \boldsymbol{\beta})$ has an asymptotic normal distribution with asymptotic variance $M(\Psi, F_{\boldsymbol{\beta}})^{-1} Q(\Psi, F_{\boldsymbol{\beta}})$ $M(\Psi, F_{\boldsymbol{\beta}})^{-1}$, where $Q(\Psi, F_{\boldsymbol{\beta}}) = E[\Psi(y, \boldsymbol{x}; \boldsymbol{\beta}, \phi, c) \Psi(y, \boldsymbol{x}; \boldsymbol{\beta}, \phi, c)^T]$ (see also (2.27)). The matrices $M(\Psi, F_{\boldsymbol{\beta}})$ and $Q(\Psi, F_{\boldsymbol{\beta}})$ for the Mallows quasi-likelihood estimator (5.13) can be easily computed as

$$Q(\Psi, F_{\boldsymbol{\beta}}) = \frac{1}{n} \mathbf{X}^T A \mathbf{X} - a(\boldsymbol{\beta}) a(\boldsymbol{\beta})^T, \tag{5.21}$$

where A is a diagonal matrix with elements $a_i = E[\psi(r_i; \boldsymbol{\beta}, \phi, c)^2] w^2(\boldsymbol{x}_i) / (\phi v_{\mu_i})$ $(\partial \mu_i / \partial \eta_i)^2$, and

$$M(\Psi, F_{\boldsymbol{\beta}}) = \frac{1}{n} \mathbf{X}^T B \mathbf{X}, \tag{5.22}$$

where B is a diagonal matrix with elements b_i as defined in Appendix E.1, and where the expectations are taken at the conditional distribution of $y_i | \boldsymbol{x}_i$. Cantoni and Ronchetti (2001b) have computed these matrices for binomial and Poisson models and Cantoni and Ronchetti (2006) for Gamma models. These results are presented in Appendix E.2 in a unified notation.

Estimated versions of the matrices $M(\Psi, \boldsymbol{\beta})$ and $Q(\Psi, \boldsymbol{\beta})$ are obtained by replacing the parameters by their M-estimates.

5.3.5 Hospital Costs Example (continued)

Consider again the hospital costs example introduced in Section 5.2.3. Model (5.12) is now refitted via the robust estimating equations (5.13) with $c = 1.5$ and $w(\mathbf{x}_i) = 1$, that is, with a Huber estimator. The scale estimator (5.19) is used for the nuisance parameter with the same value of c. The estimated parameters, standard errors, CIs and p-values of the significance test statistics (5.23) are given in Table 5.3, to be compared with Table 5.2 (classical estimates). Only small differences appear on the values of the estimated coefficients between the classical and the robust analysis

[2] The Huber ψ-function is the one used in the implementation in the `robustbase` package.

Table 5.3 Robust estimates for model (5.12).

Variable	Estimate (*SE*)	95% CI	p-value
`intercept`	7.252 (0.105)	(7.042; 7.462)	$<10^{-4}$
log(`los`)	0.839 (0.020)	(0.799; 0.879)	$<10^{-4}$
`adm`	0.222 (0.036)	(0.151; 0.294)	$<10^{-4}$
`ins`	0.009 (0.057)	(−0.104; 0.122)	0.869
`age`	−0.001 (0.001)	(−0.003; 0.001)	0.257
`sex`	0.073 (0.036)	(0.001; 0.144)	0.042
`dest`	−0.123 (0.050)	(−0.222; −0.024)	0.013
$1/\nu$ (scale)	0.0243		

The estimates are obtained solving (5.13) with $c = 1.5$ and $w(\mathbf{x}_i) = 1$ for all i (Huber's estimator), and (5.19) with $c = 1.5$.

except for the variable `ins`, where there is a difference by a factor of 10 (which is not a typo). This large difference is certainly due to the small number of patients (only nine) with private insurance, one of which is heavily downweighted in the robust analysis (patient 28, $\tilde{w}(r_i; \boldsymbol{\beta}, \phi, c) = 0.24$). On the other hand, there are major discrepancies between the estimated standard errors by the two estimators, those based on the robust approach being much smaller. These differences are mainly due to the fact that the scale estimate from the classical analysis is twice as large as that from the robust analysis (see also the simulation results of Cantoni and Ronchetti (2006, Section. 4)). This will also have an impact on the CIs and significance tests, as we will see in Section 5.4.3.

Meanwhile, we look at what the robust fit tells us. The observations that are heavily downweighted, that is, with weights $\tilde{w}(r_i; \boldsymbol{\beta}, \phi, c)$ smaller than 0.5 are $\tilde{w}(r_{14}; \boldsymbol{\beta}, \phi, c) = 0.23$, $\tilde{w}(r_{21}; \boldsymbol{\beta}, \phi, c) = 0.50$, $\tilde{w}(r_{28}; \boldsymbol{\beta}, \phi, c) = 0.24$, $\tilde{w}(r_{44}; \boldsymbol{\beta}, \phi, c) = 0.42$ and $\tilde{w}(r_{63}; \boldsymbol{\beta}, \phi, c) = 0.32$, which in this case are the same observations as identified in Section 5.2.3.

Very similar results in terms of coefficient and standard error estimates are obtained if weights $w(\boldsymbol{x}_i) = \sqrt{1 - h_{ii}}$ are used (not shown). This indicates that we can be confident that there are no bad leverage points (see Section 3.2.4.2) in the sample and, therefore, we can use a Huber-type estimator to avoid any additional loss in efficiency. Indeed, if one computes the weights $w(\boldsymbol{x}_i) = \sqrt{1 - h_{ii}}$, they would range from 0.9 to 1, with the first quartile equal to 0.96, the median equal to 0.97 and the third quartile equal to 0.98. It is particularly interesting to look at the weight of observation 31 (a potential influential point, as can be seen in Figure 5.2) which is $w(\boldsymbol{x}_{31}) = 0.96$, indicating that there is no leverage effect.

5.4 Robust Inference

5.4.1 Significance Testing and CIs

With the asymptotic result of Section 5.3.4, it is possible to draw approximate inference for $\boldsymbol{\beta}$, either by constructing approximate $(1 - \alpha)$ CIs or by computing

univariate z-statistics, namely

$$z\text{-statistic} = \frac{\hat{\beta}_{[M]j}}{SE(\hat{\beta}_{[M]j})}, \tag{5.23}$$

where $SE(\hat{\beta}_{[M]j}) = \sqrt{\widehat{\text{var}}(\hat{\beta}_{[M]j})}$ and

$$\widehat{\text{var}}(\hat{\beta}_{[M]j}) = \frac{1}{n}[\hat{M}(\Psi, F_{\boldsymbol{\beta}})^{-1}\hat{Q}(\Psi, F_{\boldsymbol{\beta}})\hat{M}(\Psi, F_{\boldsymbol{\beta}})^{-1}]_{(j+1)(j+1)}$$

in which the matrices $\hat{Q}$ and $\hat{M}$ are estimated using $\hat{\boldsymbol{\beta}}_{[M]}$ in (5.21) and (5.22), respectively. The z-statistic can then be compared with a standard normal distribution to test the null hypothesis $H_0 : \beta_j = 0$ and compute the corresponding p-value.

As in the classical setting, the asymptotic distribution can be used to define approximate $(1 - \alpha)$ CIs for each parameter β_j. Here, they write

$$(\hat{\beta}_{[M]j} - z_{(1-\alpha/2)}SE(\hat{\beta}_{[M]j}); \hat{\beta}_{[M]j} + z_{(1-\alpha/2)}SE(\hat{\beta}_{[M]j})),$$

where $z_{(1-\alpha/2)}$ is the $(1 - \alpha/2)$ quantile of the standard normal distribution.

5.4.2 General Parametric Hypothesis Testing and Variable Selection

The general parametric theory on robust testing (e.g. Heritier and Ronchetti, 1994), i.e. robust LRT, Wald and Rao or score tests, can also be used in the GLMs setting using the results presented in Section 2.5.3. However, since historically with GLMs the deviance has been used for inference purposes, we prefer to concentrate on the possibilities offered by a robust version of the deviance. Note, however, that in the classical setting the difference of deviances statistic to compare two nested models coincides with the LRT statistic when ϕ (the scale parameter) is known.

When confronted with data, it is common practice to fit a first model that includes all available explanatory variables (the full model). The p-values associated with the univariate test statistics (z-statistics) on each coefficient separately give a first broad impression on the important variables impacting the response. However, this information has to be interpreted with caution, given the possible correlation between explanatory variables and non-orthogonality of the tests. It is therefore preferable to conduct a proper variable selection analysis by means of adequate tools. Tools for variable selection, e.g. test statistics, are as much affected by extreme observations as estimators. This effect manifests itself in terms of level (for example, an actual level which does not correspond to the nominal level) and in terms of loss of power; see discussions in Sections 2.4.2, 2.4.3 and 2.5.5.

Consider a larger model $\mathcal{M}_{q+1}$ with q explanatory variables (plus intercept) and a sub-model $\mathcal{M}_{q-k+1}$ with only $(q - k)$ explanatory variables (plus intercept). The question that arises is whether the sub-model is a good enough representation of the data. Testing that some explanatory variables are not significantly contributing to

the model amounts to testing that a subset of $\boldsymbol{\beta}$ is equal to zero. Therefore, without loss of generality, we split $\boldsymbol{\beta} = (\boldsymbol{\beta}_{(1)}^T, \boldsymbol{\beta}_{(2)}^T)^T$ with $\boldsymbol{\beta}_{(1)}$ of dimension $(q - k + 1)$ and $\boldsymbol{\beta}_{(2)}$ of dimension k, and we test the null hypothesis $H_0 : \boldsymbol{\beta}_{(2)} = \mathbf{0}$. We propose (see Cantoni and Ronchetti, 2001b) a robust counterpart to the difference of deviances statistic

$$\Lambda_{QM} = 2\left[\sum_{i=1}^{n} Q_M(\hat{\mu}_i; y_i) - \sum_{i=1}^{n} Q_M(\dot{\mu}_i; y_i)\right], \tag{5.24}$$

where the quasi-likelihood functions $Q_M(\mu_i; y_i)$ are defined by (5.16), $\hat{\mu}_i = \mu_i(\hat{\boldsymbol{\beta}}_{[M]})$ is the M-estimate under model $\mathcal{M}_{q+1}$ and $\dot{\mu}_i = \mu_i(\dot{\boldsymbol{\beta}}_{[M]})$ is the M-estimate under model $\mathcal{M}_{q-k+1}$. Note that this difference of deviances is independent of $\tilde{s}$ and $\tilde{t}$, see (5.16), because their contributions cancel out.

Computing Λ_{QM} implies the computation of the functions $Q_M(\mu_i; y_i)$ which are integral forms and for which there is no general analytical expression. They can easily be approximated numerically and they have been implemented in this way. In situations where the evaluation of these integrals is problematic, an asymptotic approximation can be used, see Section 5.4.2.1.

The same forms for the functions ψ and $w(\boldsymbol{x}_i)$ as for the M-estimator $\hat{\boldsymbol{\beta}}_{[M]}$ can be used in (5.24), see the discussion in Section 5.3.1. The test statistic Λ_{QM} can be used to compare two nested models predefined by the analyst, but can also be used for a more automatic analysis, either sequential (see the example in Section 5.6.2) or marginal (stepwise, see the example in Section 5.5.2).

The test statistic (5.24) is in fact a generalization of the quasi-deviance test for GLMs ((5.11), which is recovered by taking $Q_M(\mu_i; y_i) = \int_{y_i}^{\mu_i} ((y_i - t)/\tau v_t)\, dt)$. Moreover, when the link function is the identity (linear regression), the statistic (5.24) becomes the τ-test statistic given by Hampel *et al.* (1986, Chapter 7), see also Section 3.3.3.

5.4.2.1 Asymptotic Distribution and Robustness Properties

Let $A_{(ij)}$, $i, j = 1, 2$ be the partitions of a $(q + 1) \times (q + 1)$ matrix A according to the partition of $\boldsymbol{\beta}$ into $\boldsymbol{\beta}_{(1)}$ and $\boldsymbol{\beta}_{(2)}$. Under technical conditions discussed in Cantoni and Ronchetti (2001b) and under $H_0 : \boldsymbol{\beta}_{(2)} = \mathbf{0}$, the test statistic Λ_{QM} defined by (5.24) is asymptotically equivalent to

$$n\mathbf{L}_n^T C(\Psi, F_{\boldsymbol{\beta}})\mathbf{L}_n = n\mathbf{R}_{n(2)}^T M(\Psi, F_{\boldsymbol{\beta}})_{22.1}\mathbf{R}_{n(2)}, \tag{5.25}$$

where $C(\Psi, F_{\boldsymbol{\beta}}) = M^{-1}(\Psi, F_{\boldsymbol{\beta}}) - \tilde{M}^{+}(\Psi, F_{\boldsymbol{\beta}})$ (with $\tilde{M}^{+}(\Psi, F_{\boldsymbol{\beta}})$ given below), $\sqrt{n}\,\mathbf{L}_n$ (of dimension $(q + 1)$) is normally distributed $\mathcal{N}(\mathbf{0}, Q(\Psi, F_{\boldsymbol{\beta}}))$,

$$M(\Psi, F_{\boldsymbol{\beta}})_{22.1} = M(\Psi, F_{\boldsymbol{\beta}})_{(22)} - M(\Psi, F_{\boldsymbol{\beta}})_{(12)}^T M(\Psi, F_{\boldsymbol{\beta}})_{(11)}^{-1} M(\Psi, F_{\boldsymbol{\beta}})_{(12)},$$

and $\sqrt{n}\,\mathbf{R}_n$ (of dimension $(q + 1)$) is normally distributed

$$\mathcal{N}(\mathbf{0}, M^{-1}(\Psi, F_{\boldsymbol{\beta}})Q(\Psi, F_{\boldsymbol{\beta}})M^{-1}(\Psi, F_{\boldsymbol{\beta}}))$$

(see Cantoni and Ronchetti, 2001b). Note that $\mathbf{R}_{n(2)}$ is of dimension k.

This means that Λ_{QM} is asymptotically equivalent to a quadratic form in normal variables and that Λ_{QM} is asymptotically distributed as $\sum_{i=1}^{k} d_i N_i^2$, where $N_1, \ldots, N_k$ are independent standard normal variables, $d_1, \ldots, d_k$ are the k positive eigenvalues of the matrix $Q(\Psi, F_{\boldsymbol{\beta}})(M^{-1}(\Psi, F_{\boldsymbol{\beta}}) - \tilde{M}^{+}(\Psi, F_{\boldsymbol{\beta}}))$, and $\tilde{M}^{+}(\Psi, F_{\boldsymbol{\beta}})$ is equal to

$$\tilde{M}^{+}(\Psi, F_{\boldsymbol{\beta}}) = \begin{pmatrix} M(\Psi, F_{\boldsymbol{\beta}})^{-1}_{(11)} & 0_{(q-k+1)\times k} \\ 0_{k\times(q-k+1)} & 0_{k\times k} \end{pmatrix},$$

where $0_{a\times b}$ is a matrix of dimension $a \times b$ with only zero entries.

The above results imply that the asymptotic distribution of Λ_{QM} is a linear combination of χ_1^2, for which theoretical results (e.g. Imhof, 1961) and algorithms (see Davies, 1980; Farebrother, 1990) exist. Moreover, if necessary, the distribution of the variable $\sum_{i=1}^{k} d_i N_i^2$ can be approximated with a $\bar{d}\chi_k^2$, distribution quite well, where $\bar{d} = 1/k \sum_{i=1}^{k} d_i$. No formal proof exists for the asymptotic distribution of this test statistic, but we expect that the results for linear models by Markatou and Hettmansperger (1992) carry over, at least approximately. Our experience shows that it is often the case in practice. Other approximations exist, see Wood (1989), Wood *et al.* (1993) and Kuonen (1999).

In addition to providing the asymptotic distribution of Λ_{QM}, result (5.25) states that Λ_{QM} is asymptotically equivalent to the quadratic form

$$\hat{\boldsymbol{\beta}}_{[M](2)}^{T} M(\Psi, \boldsymbol{\beta})_{22.1} \hat{\boldsymbol{\beta}}_{[M](2)}.$$

This suggests that Λ_{QM} can be approximated with this easier to compute quadratic form to avoid the numerical integrations in $Q_M(\mu_i; y_i)$, in particular when n is large.[3]

The robustness properties of a test statistic are measured on the level and on the power scale, see Section 2.2. Cantoni and Ronchetti (2001b) work out the expressions of the level and of the power of Λ_{QM} under contamination. These results show in particular that the asymptotic level of Λ_{QM} under contamination is stable as long as a bounded influence M-estimator $\hat{\boldsymbol{\beta}}_{[M](2)}$ is used in its definition.

5.4.3 Hospital Costs Data Example (continued)

If we look back at Tables 5.2 and 5.3, we can see that the conclusions from both the classical and the robust analyses on the basis of the univariate test statistics (p-values in Table 5.2 and 5.3) are quite different: if no doubt arises as to the significance of the `intercept`, and the variables log(`los`) and `adm` on both analyses, the robust analysis would suggest a significant effect also for `dest`, and less clearly for `sex`, making the role of these two variables less clear (see also the corresponding CIs). A more complete variable selection procedure is therefore recommended before proceeding with any interpretation and conclusion. We now investigate this variable selection issue a little bit further.

[3]The `anova.glmrob` function in the package `robustbase` in R (called by the generic function `anova`), implements both the test statistic Λ_{QM} and its asymptotic quadratic approximation, in addition to a Wald test.

We first start by comparing the full model to the reduced model without the variables `ins` and `age`. This amounts to testing $H_0 : \beta_3 = \beta_4 = 0$ in (5.12). We keep the same robustness tuning parameters for the robust test as in Section 5.3.5, that is, $c = 1.5$ and $w(\boldsymbol{x}_i) = 1$. The difference of quasi-deviances Λ_{QM} is equal to 1.23 (p-value $= 0.5$), which confirms the fact that these two variables do not have a significant impact on the cost of stay significantly. We go on by comparing the model including log(`los`), `adm`, `sex` and `dest` to the nested sub-model that excludes `sex`. The hypothesis that the coefficient corresponding to variable `sex` is equal to zero is rejected at the 5% level ($\Lambda_{QM} = 5.26$ and p-value $= 0.015$). Similarly, we compare the model including log(`los`), `adm`, `sex` and `dest` to the nested sub-model that discards `dest`. The difference of quasi-deviances statistic Λ_{QM} is equal to 4.82 and the p-value is 0.02, which implies the rejection of the null hypothesis that the coefficient of `dest` is equal to zero at the 5% level. This means that the models without either `sex` and `dest` are not enough to describe the data.

As a comparison, a classical analysis would also fail to reject the sub-model without `ins` and `age` compared with the full model (5.12) (p-value $= 0.44$). Starting from this sub-model, the classical analysis would reject the sub-model without `sex`, but not the sub-model without `dest`. This confirms the preliminary differences between the classical and robust analysis observed with the full fit in Section 5.3.5.

The final model obtained from the robust analysis has the following estimated linear predictor (with standard errors of the coefficients within parentheses)

$$\underset{(0.067)}{7.168} + \underset{(0.020)}{0.839}\,\log(\texttt{los}) + \underset{(0.035)}{0.231}\,\texttt{adm} + \underset{(0.034)}{0.082}\,\texttt{sex} - \underset{(0.047)}{0.104}\,\texttt{dest}\,.$$

The estimate of the scale parameter is 0.024.

The analysis suggests that hospital costs of stay for back problems are heavily dependent on length of stay, but also on the type of admission, the sex of the patient and their destination when leaving the hospital. The age of the patient and the type of insurance do not impact the costs significantly for this pathology.

The impact of the significant covariates on the average costs $E[y_i \mid \boldsymbol{x}_i] = \mu_i$ is described by $\mu_i = g^{-1}(\boldsymbol{x}_i^T \boldsymbol{\beta})$. Having used a logarithmic link in this example, we have that $\mu_i = \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})$. The interpretation of each coefficient uses this relationship and can be done separately under the circumstance that all of the other variables are kept fixed. In this respect, the above constructed model tells us that an emergency admission has a multiplicative effect of $\exp(0.231) = 1.26$ on the average cost, which means a 26% increase. Patients that go home directly after the hospital stay (with respect to those that go to another institution) have lower costs (about $90\% = \exp(-0.104)$). One could expect the converse to be true, but the patient destination after hospital is probably an indicator of how severe the back problems under treatment are: a patient that can be independent and go home directly is probably treated for a lighter problem in the beginning. Of course, the longer the stay, the higher the costs, as expected: if log(`los`) increases by 1, that is if `los` increases by 2.7, the average cost is multiplied by $\exp(0.839) = 2.31$. Finally, costs

for male patients seem to be slightly higher than those for female patients by a factor of $\exp(0.082) = 1.09$.

In this example, the estimated parameters of the variables appearing in this final model are quite close to the corresponding estimates in the full model, see Tables 5.2 and 5.3. This is due to the low correlation between the covariates.

5.5 Breastfeeding Data Example

5.5.1 Robust Estimation of the Full Model

We now look at a binary response example. The data come from a study conducted in a UK hospital on the decision of pregnant women to breastfeed their babies or not, see Moustaki *et al.* (1998). For the study, 135 expectant mothers were asked what kind of feeding method they would use for their coming baby. The responses were classified into two categories (variable `breast`), the first including breastfeeding, try to breastfeed and mixed breast- and bottle-feeding (coded 1), and the second for exclusive bottle-feeding (coded 0). The available covariates are the advancement of the pregnancy (`pregnancy`, end or beginning), how the mothers were fed as babies (`howfed`, some breastfeeding or only bottle-feeding), how the mother's friend fed their babies (`howfedfriends`, some breastfeeding or only bottle-feeding), if they had a partner (`partner`, no or yes), their age (`age`), the age at which they left full-time education (`educat`), their ethnic group (`ethnic`, white or non-white) and if they have ever smoked (`smokebf`, no or yes) or if they had stopped smoking (`smokenow`, nor or yes). All of the factors are two-level factors. The first listed level of each factor is used as the reference (coded 0).

The sample characteristics are as follows: out of the 135 observations, 99 were from mothers that have decided at least to try to breastfeed, 54 mothers were at the beginning of their pregnancy, 77 were themselves breastfed as a baby, 85 of the mother's friend had breastfed their babies, 114 mothers had a partner, median age was 28.17 (with minimum equal 17 and maximum equal 40), median age at the end of education was 17 (minimum = 14, maximum = 38), 77 mothers were white and 32 mothers were smoking during the pregnancy, whereas 51 had smoked before.

The aim of the study was to determine the factors impacting the decision to at least try to breastfeed in order to target breastfeeding promotion toward women with a lower probability of choosing it. We fitted the following model:

$$\begin{aligned}\text{logit}(E[\texttt{breast}]) = \text{logit}(P(\texttt{breast})) &= \beta_0 + \beta_1\texttt{pregnancy} + \beta_2\texttt{howfed} \\ &+ \beta_3\texttt{howfedfr} + \beta_4\texttt{partner} + \beta_5\texttt{age} + \beta_6\texttt{educat} \\ &+ \beta_7\texttt{ethnic} + \beta_8\texttt{smokenow} + \beta_9\texttt{smokebf}, \qquad (5.26)\end{aligned}$$

where $\text{logit}(p) = \log(p/(1-p))$, with $p/(1-p)$ being the odds of a success, and $P(\texttt{breast})$ is the probability of at least try to breastfeed.

Table 5.4 gives the robust estimates, standard errors and p-values for the z-test (5.23) of model (5.26) for a Huber-type estimator ($w(\boldsymbol{x}_i) = 1$) and for a Mallows-type estimator with $w(\boldsymbol{x}_i) = \sqrt{1 - h_{ii}}$. The value $c = 1.5$ has been used in both

Table 5.4 Robust estimates for model (5.26).

	Huber		Mallows	
Variable	Estimate (*SE*)	p-value	Estimate (*SE*)	p-value
`intercept`	−7.782 (3.365)	0.021	−7.778 (3.363)	0.021
`pregnancy` beginning	−0.816 (0.695)	0.241	−0.815 (0.694)	0.241
`howfed` breast	0.545 (0.710)	0.443	0.540 (0.708)	0.445
`howfedfr` breast	1.479 (0.690)	0.032	1.482 (0.689)	0.032
`partner` yes	0.772 (0.816)	0.344	0.775 (0.816)	0.342
`age`	0.030 (0.060)	0.611	0.031 (0.060)	0.608
`educat`	0.377 (0.186)	0.042	0.376 (0.185)	0.042
`ethnic` non-white	2.712 (1.125)	0.016	2.705 (1.122)	0.016
`smokenow` yes	−3.476 (1.129)	0.002	−3.468 (1.127)	0.002
`smokebf` yes	1.507 (1.103)	0.172	1.507 (1.102)	0.171

The estimates are obtained by solving (5.13) with $c = 1.5$ (Huber's estimator) and with $c = 1.5$ and $w(\mathbf{x}_i) = \sqrt{1 - h_{ii}}$ (Mallows's estimator).

cases. The coefficient estimates from both analyses are quite close, even though individual 18 (see the top panel of Figure 5.3) is considered as a potential leverage point. This mother is 38 years old and is still in education (`educat`=38). This is possible, but is certainly not common to the majority of the population. This remark raises the question of the rationale behind the definition of the variable `educat` (age at the end of full-time education). What information are we trying to measure with this variable? If it is educational level, maybe it is not what the variable `educat` really measures. In other studies, the number of years of education is recorded, which can also be seen as a proxy for social status.

From Figure 5.3 (bottom panel) we can also see that a small set of observations are downweighted on the grounds of their residuals, in particular observations 11, 14, 63, 75, 90 and 115 receive a weight of less than 0.6. Note that 6 observations out of 135 constitute about 4.5% of the total information. For these mothers the fitted model (5.26) would predict a probability of at least try to breastfeed which is not consistent with the behavior of the majority of the mothers in the sample on the basis of the covariates (see Figure 5.4): for instance, for observations 75, 11, 115 and 14 the predicted probability of trying to breastfeed is larger than 0.90, whereas these mothers have decided to bottlefeed. On the other hand, mothers 90 and 63 are given a low probability of only 0.02 and 0.11 respectively of trying to breastfeed by the model, whereas they have chosen to do so.

According to the p-values of Table 5.4, the variables that have the greatest impact on the decision to at least try to breastfeed are whether the ethnic group is non-white, whether currently smoking and less strongly the age at which one left education and whether friends have chosen to breastfeed. A more formal variable selection procedure follows in Section 5.5.2.

Note that a classical analysis would have yield different estimates and conclusions, see also Section 5.5.2. A slightly different estimation method for this dataset

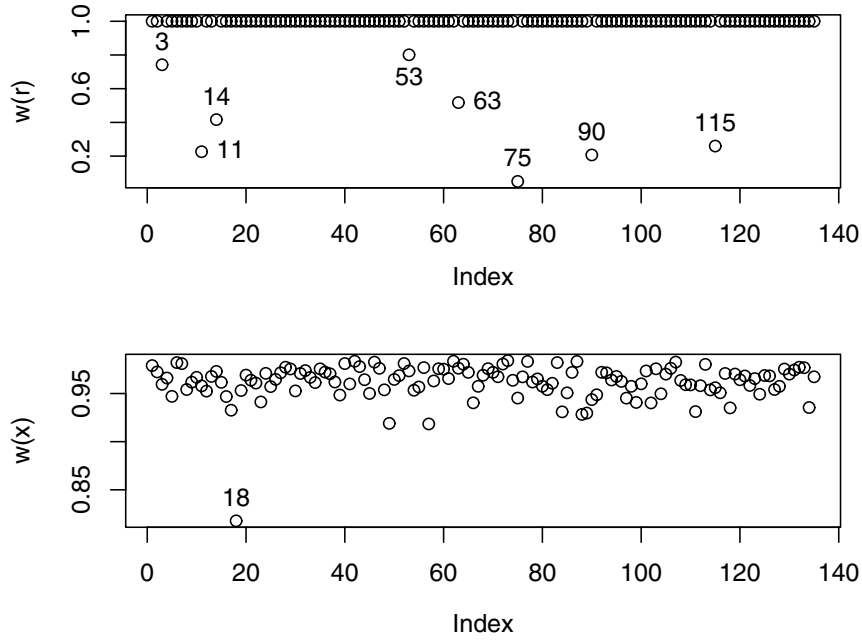

Figure 5.3 Robustness weights on the design and on the residuals for model (5.26), when estimated by (5.13) with $c = 1.5$ and $w(\boldsymbol{x}_i) = \sqrt{1 - h_{ii}}$.

has been used in Victoria-Feser (2002), in particular a model-based weighting scheme. The conclusions are similar between our proposal and her Mallows-type estimator.

5.5.2 Variable Selection

When analysing the full model, on the basis of p-values corresponding to the z-statistics, the variables `howfedfr`, `smokenow`, `ethnic` and `educat` have an important impact on the decision to at least try to breastfeed. Here we investigate further the variable selection issue. With this dataset, we illustrate a backward stepwise procedure. We start with the full model and we use the test statistic Λ_{QM} to test each sub-model with one variable removed. All of the sub-models for which the p-value of such a test is larger than 5% are candidates for removal, and we choose between them the sub-model which has generated the larger p-value. We then repeat the procedure by taking this sub-model as the new reference model and testing all of its sub-models. The procedure is stopped when all of the p-values are larger than 0.05.

Table 5.5 gives the p-values at the first steps of the procedure. For comparison, we also put the results for a classical analysis. A comparison of the p-values from the classical and the robust approaches confirms that the robustness issues related to

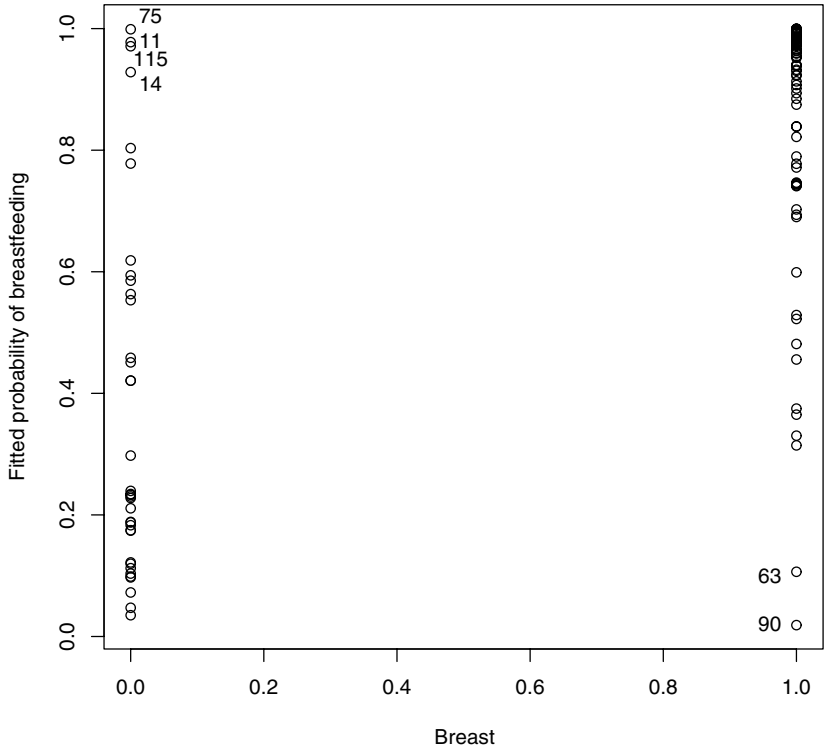

Figure 5.4 Fitted values versus actual values for model (5.26), when estimated by (5.13) with $c = 1.5$ and $w(\boldsymbol{x}_i) = \sqrt{1 - h_{ii}}$. Observations with $w(r_i; \boldsymbol{\beta}, \phi, c) < 0.6$ are spotted.

the presence of deviating data points are also a concern for inference. In fact, large discrepancies (as large as 0.2) appear between the two approaches in terms of p-values. Some of these differences do not really have an impact on the significance decision at a usual level of 5% or 10% (e.g. `howfed` or `partner`), but some others do (e.g. `educat`).

The complete robust stepwise procedure yields the following final model (with standard errors of the coefficients within parentheses):

$$\underset{(2.973)}{-6.417} + \underset{(0.622)}{1.478}\,\texttt{howfedfr} + \underset{(1.199)}{3.260}\,\texttt{ethnic} + \underset{(0.177)}{0.403}\,\texttt{educat} - \underset{(0.664)}{2.421}\,\texttt{smokenow}.$$

From the robust analysis, the non-significant variables have been removed in the following order: `age` (p-value $= 0.58$, the largest p-value at the first step, see Table 5.5), `howfed` (p-value $= 0.40$), `pregnancy` (p-value $= 0.26$), `partner` (p-value $= 0.41$) and `smokebf` (p-value $= 0.25$).

As a comparison, a classical backward stepwise procedure would have discarded (in the order) `age` (p-value $= 0.60$, the largest p-value at the first step, see

Table 5.5 p-values of the first step of a backward stepwise procedure for variable selection for the breastfeeding data example of Section 5.5.

Variable	Classical	Robust
`pregnancy` beginning	0.08134	0.20600
`howfed` breast	0.60261	0.39778
`howfedfr` breast	0.00951	0.02820
`partner` yes	0.12219	0.32888
`age`	0.60271	0.58512
`educat`	0.14075	0.02283
`ethnic` non-white	0.00012	0.00187
`smokenow` yes	$<10^{-4}$	$<10^{-4}$
`smokebf` yes	0.05157	0.08605

Classical p-values obtained with $c = \infty$ and $w(\boldsymbol{x}_i) = 1$ and robust p-values with $c = 1.5$ and $w(\boldsymbol{x}_i) = \sqrt{1 - h_{ii}}$ in (5.24).

Table 5.5), `howfed` (p-value $= 0.58$), `educat` (p-value $= 0.10$), `pregnancy` (p-value $= 0.20$), `smokebf` (p-value $= 0.10$) and `partner` (p-value $= 0.0577$). The classical final model would therefore include only `howfedfr`, `ethnic` and `smokenow`, which is a smaller and different set of covariates than obtained by the robust analysis.

From the model identified and fitted by the robust technique we learn that the way a mother has been fed as a child does not play a role in her decision of whether to breastfeed, whereas the choice of friends is more important and has an effect on the expectant mother's decision. A mother's choice to try to breastfeed does not evolve during the pregnancy. This choice is also not affected by the mother being single. Having smoked before being pregnant has no effect on the decision to breastfeed, but being a smoker during the pregnancy significantly reduces the probability to at least try to breastfeed. Ethnicity and age at which a mother leaves education are also factors that have an impact on a mother's decision.

The coefficient values allow us to quantify the identified effects on the decision to at least try to breastfeed. As opposed to the Gamma model of Section 5.3.5 or to a Poisson model (see Section 5.6), the interpretation of the impact of covariates on the probability $P(\texttt{breast})$ is more difficult due to the nature of the logit transformation. In fact,

$$P(\texttt{breast}_i) = \mu_i = \frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}. \tag{5.27}$$

With these models it is therefore more common to interpret the coefficients on the odds or odd-ratios scale. The robust estimation procedure has no impact on the way the model is interpreted. The only difference is that the coefficients are estimated differently.

For a continuous variable, the effect of a unit change on the odds is equal to the exponential of the corresponding coefficient. For example, leaving education a year later increases the odds of at least try to breastfeed by a factor of $\exp(0.403) = 1.50$, if all the other covariates are kept fixed. On the other hand, for two-level factors the logit model leads to the interpretation of the odds-ratio (the ratio of the odds). For instance, the odds-ratio of at least try to breastfeed for a non-white expecting mother relative to a white mother is equal to $\exp(3.260) = 26.05$. Similarly, the odds-ratio of at least try to breastfeed for a smoking mother relative to a non-smoking one is $\exp(-2.421) = 0.09$. Being a smoker during pregnancy has the strongest (negative) effect on the model. Finally, the odds-ratio of at least try to breastfeed for an expectant mother whose friends have chosen to breastfeed relative to friends bottlefeeding is $\exp(1.478) = 4.38$.

The interpretation of odds and odds-ratios pertains to the logistic model (that is, the binomial model with logit link), but does not apply to models with the probit or complementary log–log link. This fact is one of the reasons that makes logistic models more popular than the two other alternatives, in addition to their more convenient computational aspects.

To summarize, let us recall that the aim of the study was to better target the expectant mothers when promoting breastfeeding. The analysis of this dataset suggests that if one wants to increase the average probability of choosing to at least try to breastfeed, directed effort should be towards white mothers and towards mothers that leave education earlier. Pregnant women that smoke tend to avoid breastfeeding: investigating this phenomenon further could help increase the average probability of expectant mothers choosing to breastfeed.

5.6 Doctor Visits Data Example

5.6.1 Robust Estimation of the Full Model

Count data are an important subclass of data that fits into the GLM framework. For this application we use data from the Health and Retirement Study (HRS),[4] which surveys more than 22 000 Americans over the age of 50 every 2 years. The study paints an emerging portrait of an aging America's physical and mental health, insurance coverage, financial status, family support systems, labor market status and retirement planning.

The original full dataset from RAND HRS Data (Version D) distribution (six waves: 1992, 1994, 1996, 1998, 2000 and 2002) contains 26 728 observations and 4140 variables per individual. Individuals were separated in four cohorts:

- HRS cohort (born between 1931 and 1941);
- AHEAD cohort (born before 1924);

[4] Sponsored by the National Institute of Aging (grant number NIA U01AG09740) and conducted by the University of Michigan, see http://hrsonline.isr.umich.edu/.

- CODA cohort (born between 1924 and 1930);
- WB cohort (born between 1942 and 1947).

In addition to respondents from eligible birth years, the survey interviewed the spouses of married respondents or the partner of a respondent, regardless of age.

We focus on a subsample of 3066 individuals of the AHEAD cohort for wave 6 (year 2002). Note that only individuals with full information have been retained, to avoid issues with missing values.

The aim is to identify variables impacting on equity in health care utilization. When the information about costs themselves is not available (in contrast to the example in Section 5.2.3), a proxy variable is used to measure health care consumption, for example the number of visits to the doctor in the previous 2 years. A set of potentially interesting explanatory variables has been retained on the basis of previous studies from the literature, e.g. Dunlop *et al.* (2002) and Gerdtham (1997), see Table 5.6. These variables are classified into three categories: predisposing variables, health needs and economic access. The first category includes age, gender, race and marital status. Health needs are represented by chronic conditions and functional limitations. In the economic access category, years of education and parents' education measure human capital, whilst income and health insurance from a current or previous employer measure financial ability to pay.

A potential concern with count data in the setting of health consumption is the excess of zeros, that is, a large presence of zero values among the responses, which cannot be modeled with standard distributions (see Ridout *et al.* (1998) and Section 5.7.1). Given that we target here a population of regular users (elderly) this issue can be excluded. In fact, only about 4% of the counts are equal to zero, see the histogram in Figure 5.5. We therefore confidently proceeded with a GLM Poisson model with log-link including all of the available covariates:

$$\begin{aligned}
&\log(E[\texttt{visits}]) \\
&\quad = \beta_0 + \beta_1\texttt{age} + \beta_2\texttt{gender} + \beta_3\texttt{race} + \beta_4\texttt{hispan} \\
&\qquad + \beta_5\texttt{marital} + \beta_6\texttt{arthri} + \beta_7\texttt{cancer} + \beta_8\texttt{hipress} \\
&\qquad + \beta_9\texttt{diabet} + \beta_{10}\texttt{lung} + \beta_{11}\texttt{hearth} + \beta_{12}\texttt{stroke} \\
&\qquad + \beta_{13}\texttt{psych} + \beta_{14}\texttt{iadla1} + \beta_{15}\texttt{iadla2} + \beta_{16}\texttt{iadla3} \\
&\qquad + \beta_{17}\texttt{adlwa1} + \beta_{18}\texttt{adlwa2} + \beta_{19}\texttt{adlwa3} + \beta_{20}\texttt{edyears} \\
&\qquad + \beta_{21}\texttt{feduc} + \beta_{22}\texttt{meduc} + \beta_{23}\log(\texttt{income} + 1) + \beta_{24}\texttt{insur}. \qquad (5.28)
\end{aligned}$$

We fitted both a classical MLE and a Mallows' robust estimator according to (5.13) with $c = 1.6$ and $w(\mathbf{x}_i) = \sqrt{1 - h_{ii}}$.

Given the large number of covariates, the results are presented graphically. Figure 5.6 shows approximate 95% CIs for each variable resulting from a classical fit (on the left, gray line) and from a robust fit (on the right, black line). The intervals are symmetric and the coefficient itself is represented in the middle with a dot.

Table 5.6 HRS data variables description. Note that `iadla` sums the answer to 'can use the phone', 'can manage money', 'can take medication', where the answer to each question is coded 1 = difficulty or 0 = no difficulty. Similarly, `adlwa` sums the response to being able to 'bath', 'eat' and 'dress'. Finally, 'med' stands for median. Sample size is 3066.

Name	Description	Sample values
Response		
`visits`	Number of visits to the doctor	0–750 (med = 8)
Predisposing		
`age`	Age in years	42–109 (med = 82)
`gender`	Gender (0 = male, 1 = female)	2079 females
`race`	Race (1 = white/Caucasian, 0 = other)	2714 whites
`hispan`	Hispanic (1 = Hispanic, 0 = other)	183 Hispanic
`marital`	Marital status (1 = married, 0 = other)	1203 married
Health needs		
`arthri`	Ever had arthritis (1 = yes, 0 = no)	'yes': 2200
`cancer`	Ever had cancer (1 = yes, 0 = no)	'yes': 594
`hipress`	Ever had high blood pressure (1 = yes, 0 = no)	'yes': 1856
`diabet`	Ever had diabetes (1 = yes, 0 = no)	'yes': 524
`lung`	Ever had lung disease (1 = yes, 0 = no)	'yes': 312
`hearth`	Ever had hearth problems (1 = yes, 0 = no)	'yes': 1206
`stroke`	Ever had a stroke (1 = yes, 0 = no)	'yes': 492
`psych`	Ever had psychiatric problems (1 = yes, 0 = no)	'yes': 479
`iadla`	Instr. activities of daily leaving (0,1,2,3)	'0': 2433, '1': 258 '2': 178, '3': 197
`adlwa`	Activities of daily leaving (0,1,2,3)	'0': 2284, '1': 361 '2': 234, '3': 187
Econ. access		
`edyears`	Education years	0–17 (med = 12)
`feduc`	Father education (years)	0–17 (med = 8.5)
`meduc`	Mother education (years)	0–16 (med = 8.5)
`income`	Total household income	0–725 600 (med = 21 540)
`insur`	Ins. from current/prev. empl. (1 = yes, 0 = no)	'yes': 649

Note that the magnitude of the coefficients is not comparable between all of the variables. In fact, some of them are measured in years, e.g. `age`, `meduc`, `feduc` and `edyears`, one is measured in log-dollars (log(`income` + 1)) and all of the other variables are dummies.

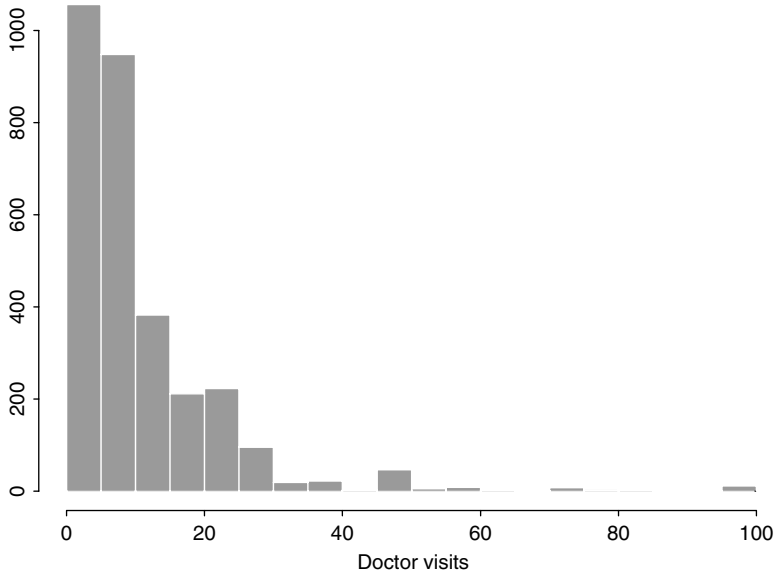

Figure 5.5 Histogram of `visits`. Note that the abscissa has been limited to (0, 100) (there are 21 observations out of 3066 outside this range, the largest value being 750).

As one can see, the coefficients of the classical and the robust analyses are sometimes quite different. Also, the standard errors estimates tend to be a bit larger in the robust analysis.

The CIs from the classical analysis indicate that all of the variables are highly significant (no crossing of the horizontal line at zero), except for `marital`. From the robust analysis it seems, however, that the variables `race`, `meduc`, $\log(\texttt{income}+1)$ and `insur` are not significant. For additional variable significance tests, see Section 5.6.2.

The dataset here is much larger than the previous dataset both in sample size and in the number of covariates. For this reason, the plot of the weights (see Figure 5.7) shows what seems to be a large number of downweighted observations. Note, however, that the average of the weights with respect to the total number of observations is $\sum_{i=1}^{3066} \tilde{w}(r_i; \boldsymbol{\beta}, \phi, c) w(\boldsymbol{x}_i)/3066 = 79.4\%$, which reflects, loosely speaking, an average degree of 'outlyingness' of about 20%. This may seem a lot, possibly indicating that extra covariates should be added or that the distributional assumptions should be modified. Also, the weights on the design are all close to one.

5.6.2 Variable Selection

As can be seen in Figure 5.6, almost all of the (preselected) variables for this study seem significant. We would like to confirm whether the variables `race`, `meduc`, $\log(\texttt{income}+1)$ and `insur` can be excluded from the model. For this purpose

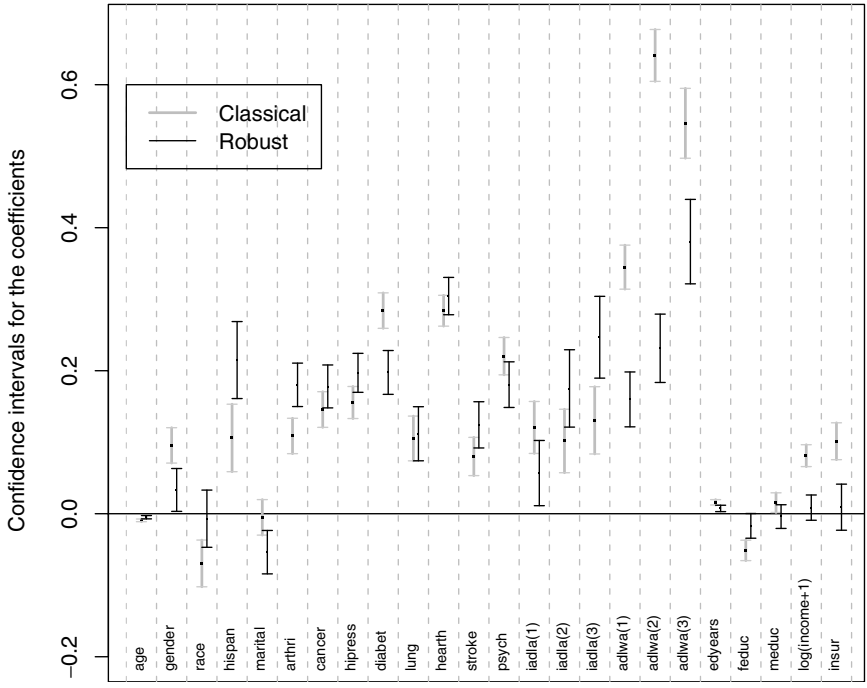

Figure 5.6 Coefficient estimates and approximate 95% CIs for the log-link Poisson model (5.28), estimated by maximum likelihood (classical) and by (5.13) with $c = 1.6$ and $w(\boldsymbol{x}_i) = \sqrt{1 - h_{ii}}$ (robust). For each variable, the results on the left are from the classical analysis and on the right from the robust analysis.

we use the difference of quasi-deviance statistic Λ_{QM} with $c = 1.6$ and $w(\boldsymbol{x}_i) = \sqrt{1 - h_{ii}}$. We first test the null hypothesis $H_0 : \beta_3 = 0$ in the full model, which is not rejected (p-value $= 0.73$). We therefore remove the variable `race`. We test next whether `meduc` is significant in the sub-model that has already `race` removed. This variable is not significant (p-value $= 0.62$) and we remove it. We go on with testing whether we can in addition remove $\log(\texttt{income} + 1)$, which is not significant (p-value $= 0.35$). We last test the removal of `insur`. The p-value is 0.50, and we decide to remove also `insur`.

The above approach is called a sequential approach and differs from a marginal/stepwise approach in that it does not test all of the sub-models at each step. The drawback is that the final model is heavily dependent on the order in which the variables are considered for removal, in particular when the covariates are far from being independent.

Table 5.7 gives the estimates on the final model retained above. The factors explaining the number of visits to the doctor are numerous, as confirmed by the long list of variables in Table 5.7. We have already learned that being Caucasian, the

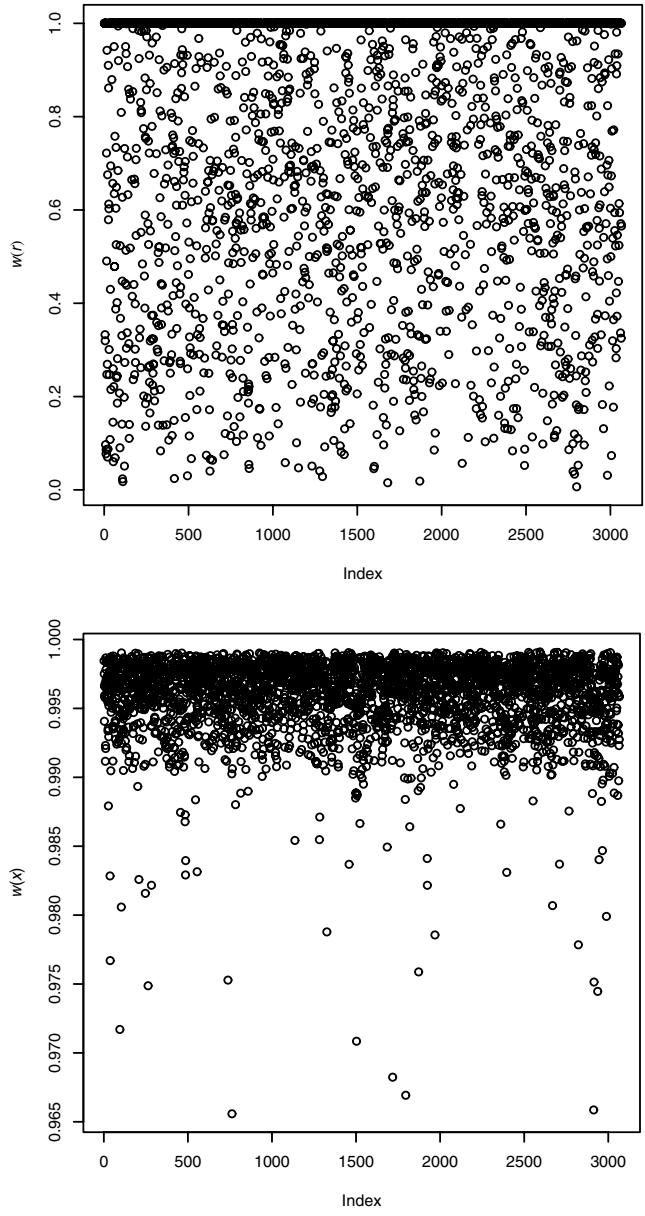

Figure 5.7 Robustness weights from the fit of model (5.28) estimated by (5.13) with $c = 1.6$ and $w(\mathbf{x}_i) = \sqrt{1 - h_{ii}}$.

Table 5.7 Final model estimates for the doctor visits data.

Variable	Estimate (*SE*)	p-value
`intercept`	1.989 (0.114)	$<10^{-4}$
`age`	−0.005 (0.001)	$<10^{-4}$
`gender`	0.030 (0.015)	0.0409
`hispan`	0.213 (0.027)	$<10^{-4}$
`marital`	−0.050 (0.014)	0.0006
`arthri`	0.180 (0.015)	$<10^{-4}$
`cancer`	0.178 (0.015)	$<10^{-4}$
`hipress`	0.197 (0.014)	$<10^{-4}$
`diabet`	0.198 (0.015)	$<10^{-4}$
`lung`	0.110 (0.019)	$<10^{-4}$
`hearth`	0.304 (0.013)	$<10^{-4}$
`stroke`	0.125 (0.016)	$<10^{-4}$
`psych`	0.180 (0.016)	$<10^{-4}$
`iadla1`	0.056 (0.023)	0.0143
`iadla2`	0.176 (0.027)	$<10^{-4}$
`iadla3`	0.244 (0.029)	$<10^{-4}$
`adlwa1`	0.160 (0.019)	$<10^{-4}$
`adlwa2`	0.231 (0.024)	$<10^{-4}$
`adlwa3`	0.382 (0.029)	$<10^{-4}$
`edyears`	0.008 (0.002)	$<10^{-4}$
`feduc`	−0.020 (0.006)	0.0025

The estimates are obtained by (5.13) with $c = 1.6$ and $w(\boldsymbol{x}_i) = \sqrt{1 - h_{ii}}$ (Mallows' estimator).

level of mother's education, total household income and having a health insurance plan from a previous employer do not have a statistically significant impact on health consumption (doctor visits).

The Poisson GLM model used for this example has a logarithmic link. Interpretation of the coefficient is therefore done through the relationship $\mu_i = \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})$, as in the Gamma model with logarithmic link in Section 5.4.3. For example, a patient who is five years older would have a number of visits to the doctor multiplied by $\exp(-0.005 \cdot 5) = 0.975$ on average, that is, reduced by 2.5%. It is surprising to see that the coefficient of `age` is negative, meaning that older patients consume less. However, the effect is really small (no practical significance), even though statistically significant. Interpretation of education level via years of education (`edyears`) and father's education (`feduc`) is puzzling. On the one hand, an extra year of father's education decreases the number of visits by 2% ($\exp(-0.02) = 0.98$). On the other hand, years of education of the patient himself tend to increase the doctor needs by 1% ($\exp(0.008) = 1.01$). Married individuals do visit the doctor less on average: $\exp(-0.05) = 95\%$. All of the effects of 'health needs' category are positive, indicating, as expected, that if some conditions are

present (arthritis, diabetes, high blood pressure, etc.), the number of doctor visits is larger on average with respect to an individual with absence of these conditions.

5.7 Discussion and Extensions

The GLM class encompasses a large variety of data distributions, but of course it has its own limitations. Therefore, GLM have been extended in various ways. The linear component structure has been relaxed and non-parametric functions have been considered in generalized additive models (GAMs; see Hastie and Tibshirani (1990)). The exponential family restriction can be overcome by using quasi-likelihood functions instead of proper likelihoods. The asymptotic results for the estimators derived in this way have to be adapted, essentially by changing the asymptotic variance estimator (sandwich formula, see Fahrmeir and Tutz (2001, pp. 55–58)). Finally, in GLM the responses are assumed to be independent and therefore do not include, for instance, longitudinal or clustered data, where there are typically several observations per subject for which it is not reasonable to assume independence (even though the subjects themselves can be considered independent), see in particular Chapter 6.

In the following sections we discuss some ideas for extensions of the approach presented in this chapter and some open areas of research.

5.7.1 Robust Hurdle Models for Counts

A particular feature of count data is the fact that they sometimes show an excess of zeros. Typical examples include the number of visits to the doctor on a given period (see Cameron and Trivedi, 1998) or abundance of species (see Barry and Welsh, 2002).

Data with an excess of zeros have been modeled in various ways: with mixture models, with more flexible distributions than the more common Poisson (e.g. negative binomial, Neyman type-α, see for instance Dobbie and Welsh (2001b)), with zero-inflated distributions (zero-inflated Poisson or zero-inflated negative binomial, see Lambert (1992)), or with hurdle models (also called two-step or conditional models, see Mullahy (1986)). Ridout *et al.* (1998) and Min and Agresti (2002) give extensive reviews.

From our perspective, hurdle models are quite attractive because they possess nice orthogonality properties, they fit nicely in the GLM framework and in its robust approach presented in this chapter. A hurdle model is characterized by a two-stage procedure. First, the presence ($y_i > 0$) or absence ($y_i = 0$) is modeled through a set of covariates $\boldsymbol{x}_i$ with a logistic-type of model. Then, conditional on the presence, the positive values are modeled through a set of covariates $\tilde{\boldsymbol{x}}_i$ (possibly equal to $\boldsymbol{x}_i$) with a truncated distribution (e.g. a truncated Poisson) and corresponding model (a log-linear type of model). This implies that $y_i = 0$ with probability $1 - p(\boldsymbol{x}_i)$ and

$y_i \sim$ truncated Poisson with probability $p(\boldsymbol{x}_i)$. In summary,

$$P(Y_i = y_i \mid \boldsymbol{x}_i, \tilde{\boldsymbol{x}}_i) = \begin{cases} (1 - p(\boldsymbol{x}_i)) & y_i = 0, \\ p(\boldsymbol{x}_i) \dfrac{\exp(-\lambda(\tilde{\boldsymbol{x}}_i))\lambda(\tilde{\boldsymbol{x}}_i)^{y_i}}{y_i!(1 - \exp(-\lambda(\tilde{\boldsymbol{x}}_i)))} & y_i = 1, 2, \ldots, \end{cases}$$

with $\text{logit}(p(\boldsymbol{x}_i)) = \boldsymbol{x}_i^T \boldsymbol{\beta}$ and $\log(\lambda(\tilde{\boldsymbol{x}}_i)) = \tilde{\boldsymbol{x}}_i^T \boldsymbol{\alpha}$.

The log-likelihood $l(\boldsymbol{\alpha}, \boldsymbol{\beta})$ of the above model factorizes as $l(\boldsymbol{\alpha}) + l(\boldsymbol{\beta})$, which has the double advantage of splitting the fitting into two subproblems of smaller size and rendering the interpretation easier (each set of parameter impact only one part of the model).

A robust procedure for the hurdle model can be derived by robustifying each submodel separately. The logistic presence/absence model can be fitted robustly by the approach presented in the previous sections and the truncated Poisson modeling part has been addressed in Zedini (2007). Routines in R are currently under preparation and will be made available either within the `robustbase` package or as a standalone package.

5.7.2 Robust Akaike Criterion

The principle of the AIC (see Section 3.4.5) is to use the likelihood information at a given model penalized by its number of parameters to identify the best model(s), that is, the best compromise(s) between parsimony and goodness of fit. The smaller the value of AIC, the better. In fact, AIC is an estimate of the expected entropy, that one would like to maximize. A robust version of AIC is available for linear models, see Section 3.31, but not (yet) for GLMs, where a generalized version of AIC can be constructed based on the quasi-likelihood functions defined in this chapter. We briefly sketch the idea here.

The log-likelihood in the original definition of AIC can be replaced by the quasi-likelihood function (5.7) with the penalization term adapted, see Ronchetti (1997b) and Stone (1977). This yields the final generalized criterion:

$$\text{GAIC} = -2 \sum_{i=1}^{n} Q_M(\hat{\mu}_i; y_i) + 2 \operatorname{tr}(M^{-1}(\Psi, F_{\boldsymbol{\beta}}) Q(\Psi, F_{\boldsymbol{\beta}})),$$

with $M(\Psi, F_{\boldsymbol{\beta}})$ and $Q(\Psi, F_{\boldsymbol{\beta}})$ given in (5.21) and (5.22).

5.7.3 General C_p Criterion for GLMs

The Mallows' C_p criterion (Mallows, 1973) has been mainly used in linear regression. A robust version of it for linear models exist thanks to Ronchetti and Staudte (1994) (see (3.32)). It is constructed upon the idea that the C_p criterion is an unbiased estimator of some sort of measure of prediction error. Following the same reasoning, Cantoni *et al.* (2005) develop a similar criterion, called GC_p, to be used for GEE models to address various issues (missingness, heteroscedasticity) including

robustness. The GLM setting being the limiting case of a longitudinal setting where there is only one observation per subject, GC_p for GLM can be deduced from the original proposal of Cantoni *et al.* (2005). If we define the rescaled weighted predictive squared error by

$$\Gamma_p = \sum_{i=1}^{n} E\left[w^2(r_i^p) \cdot \left(\frac{\hat{y}_i^p - E[y_i \mid \boldsymbol{x}_{i(p)}]}{\sqrt{\phi \hat{v}_{\mu_i}}} \right)^2 \right], \tag{5.29}$$

where $r_i^p = (y_i - \hat{y}_i^p)/\sqrt{\phi \hat{v}_{\mu_i}}$ are the Pearson residuals, $\hat{y}_i^p$ are the fitted values at the model with $p \leq (q+1)$ explanatory variables $\boldsymbol{x}_{i(p)}$ (including the intercept), $\hat{v}_{\mu_i}$ are 'external' variance estimates (held fixed) and where $w(\cdot)$ is a weighting function to downweight atypical observations, then a general form of an unbiased estimator for Γ_p is

$$GC_p = \sum_{i=1}^{n} (w(r_i^p) r_i^p)^2 - \sum_{i=1}^{n} E[(w(r_i^p)\epsilon_i)^2] + 2 \sum_{i=1}^{n} E[w^2(r_i^p)\epsilon_i \delta_i], \tag{5.30}$$

with $\epsilon_i = (y_i - E[y_i \mid \boldsymbol{x}_{i(p)}])/\sqrt{\phi v_{\mu_i}}$ and $\delta_i = (\hat{y}_i^p - E[y_i \mid \boldsymbol{x}_{i(p)}])/\sqrt{\phi v_{\mu_i}}$ and where the two latter terms are corrections to achieve unbiasedness. Computing these two terms for GLM and for our particular (robust) M-estimator (5.13) would yield the final form of GC_p.

5.7.4 Prediction with Robust Models

The goals of model fitting are numerous, but they certainly include prediction. For example, in the hospital costs example of Section 5.2.3, health insurances could be interested in forecasting costs for the following year in order to establish their budget. If in this example the robust fitted model is used naively to obtain predictions, the reproducibility of the outliers, that is the fact that individuals with high abnormal costs will likely appear again in the future, would imply potential severe bias in prediction (e.g. underestimation). This particular feature is shared by all of the models where the outliers are characterized by particularly large values with respect to the bulk of the data (this is not the case in examples with binary responses, for example).

In this kind of situation, one should therefore correct the predictions for possible reproducible outliers, by considering shrinkage robust estimators, see for example Welsh and Ronchetti (1998) and Genton and Ronchetti (2008).

6

Marginal Longitudinal Data Analysis

6.1 Introduction

Longitudinal data models are a step further away from linear models. Beyond GLMs, longitudinal studies are those where individuals are measured repeatedly over time. So, with respect to the GLM modeling of Chapter 5, a second dimension is added, where each subject can be measured several times. With respect to the (normal) MLMs of Chapter 4, the extension broadens the nature of responses considered. Here we allow the response to come from any distribution of the exponential family (discrete or continuous), as in Chapter 5. Note that The terminology 'longitudinal data' is used mostly in medicine, biology and health sciences, whereas sociologists and economists would mostly use the term 'panel data'.

It has to be stressed that even though the most common applications are for situations where the main units are individuals (e.g. the example in Section 6.5), the methodology can also be applied to otherwise clustered data where there are units in which measurements cannot be considered independent (e.g. the example in Section 6.2.3).

When there is only one observation per subject, inference solely about the population average is possible. In contrast, longitudinal studies can distinguish between changes over time within individuals (called aging effects) and differences among people in their baseline levels (called cohort effects). Otherwise said, longitudinal studies are able to distinguish between the degree of variation of the response across time for one person and the variation in the response among people. Statistically speaking, one has to take into account the correlation within

Robust Methods in Biostatistics S. Heritier, E. Cantoni, S. Copt and M.-P. Victoria-Feser

measurements of the same subject (even if the subjects themselves can be considered independent). The same pattern/behavior is assumed across subjects and strength is borrowed from this.

The literature about marginal longitudinal models is wide, also because models are developed in at least three main directions, see Section 6.2. The bases of the generalized estimating equations (GEE) approach that we follow here have been introduced with the seminal work of Liang and Zeger (1986) and Zeger and Liang (1986). Since then, many extensions and variations have been considered, in particular including an extension to the mixed linear type of models (Zeger *et al.*, 1988), polytomous responses (Heagerty and Zeger, 1996; Liang *et al.*, 1992; Stram *et al.*, 1988), survival responses (Heagerty and Zeger, 2000, and references therein), weighted GEE (Preisser *et al.*, 2000) and zero-inflated count data (Dobbie and Welsh, 2001a). A nice book on longitudinal data is Diggle *et al.* (2002), which is an extension of an earlier edition. The book by Molenberghs and Verbeke (2005) is another interesting reference. A more focused book on the GEE approach is Hardin and Hilbe (2003) and a recent book addressing correlated data is Song (2007).

The theory around the GEE approach is sometimes sparse, in particular when it comes to the nuisance parameters, where the inferential aspects have not been well treated. The variable selection issues with these models have been addressed only recently, when Pan (2001) define an Akaike-type criterion for GEE, called QIC. Moreover, Cantoni *et al.* (2005) introduce a general C_p-like criterion for variable selection for marginal longitudinal models that can also address robustness issues.

Robust alternatives to GEE-type of fits have been first proposed by Preisser and Qaqish (1999), who define a set of resistant estimating equations. Wang *et al.* (2005) propose a robust GEE-type bias corrected estimator, where the bias is estimated using a classical GEE estimator. Qu and Song (2004) show that their estimating equations proposal based on quadratic inference functions (Qu *et al.*, 2000) has some nice robustness properties for the estimation of the regression parameters in some cases. Cantoni (2004b) propose a more general and improved version of the estimating equations of Preisser and Qaqish (1999) that also allows quasi-likelihood functions to be defined for inference, which puts the user in a position to carry a full analysis. We have chosen to present this approach given our familiarity with it, because of its extensions that make variable selection possible along the same lines as the approach for GLM and because of its forthcoming availability in R.

In this chapter, after discussing the possible approaches to longitudinal data (Section 6.2), we go on to introduce marginal longitudinal models in more detail and present the classical estimation procedure (GEE) to fit them and the associated inference in Section 6.2.1. The robust counterpart, as per Cantoni (2004b), is introduced and illustrated in Section 6.3. It is based on a weighted set of estimating equations. In addition, quasi-deviance functions are defined for inference purposes and robust model selection. Three different examples serve as motivation and illustration of the theoretical elements introduced in this chapter, especially in Sections 6.3.4, 6.5 and 6.6.

6.2 The Marginal Longitudinal Data Model (MLDA) and Alternatives

We assume that we have measurements y_{it} for individual (or unit or cluster) $i = 1, \ldots, n$ at time (or occasion or occurrence) $t = 1, \ldots, n_i$. We additionally define $\boldsymbol{y}_i^T = (y_{i1}, \ldots, y_{in_i})$ as the collection of measurements for subject i and we assume independence between subjects. We assume that $E[\boldsymbol{y}_i] = \boldsymbol{\mu}_i$ and that $\text{var}(\boldsymbol{y}_i)$ is non-diagonal. At each time point, a set of covariates $\boldsymbol{x}_{it}^T = (1, x_{it1}, \ldots, x_{itq})$ is also measured for each individual. The covariates information on subject i is collected in a $n_i \times (q+1)$ matrix

$$\mathbf{X}_i = \begin{pmatrix} \boldsymbol{x}_{i1}^T \\ \vdots \\ \boldsymbol{x}_{in_i}^T \end{pmatrix} = \begin{pmatrix} 1 & x_{111} & \cdots & x_{11q} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n_i1} & \cdots & x_{1n_iq} \end{pmatrix}.$$

The complete set of data comprises $N = \sum_{i=1}^{n} n_i$ observations.

As with GLMs, the response y_{it} will be allowed to come from any distribution of the exponential family, see Table 5.1 in Chapter 5. However, using the GLM methodology would not be appropriate here because it ignores the correlation between the measurements of the same subject. Ignoring this correlation has consequences at different levels: inference about the regression parameters is incorrect, estimation of the regression parameters is inefficient and there is suboptimal protection against biases caused by missing data.

The difficulty with the analysis of non-Gaussian longitudinal data was the lack of a rich class for the joint distribution of $(y_{i1}, \ldots, y_{in_i})$. There are essentially three strategies to address the issue. All three approaches model both the dependence of the response on the explanatory variables and the correlation among the responses. In the following we give a brief overview.

1. **Marginal models.** Via this approach one models parametrically not only the marginal mean of y_{it} (as in GLMs and in cross-sectional studies in general) but also the correlation matrix $\text{corr}(\boldsymbol{y}_i)$, by imposing a relationship $g(E[y_{it}]) = \boldsymbol{x}_{it}^T \boldsymbol{\beta}$ for a link function g, and by modeling the covariance matrix with extra parameters τ and $\boldsymbol{\alpha}$: $V_{\boldsymbol{\mu}_i, \tau, \boldsymbol{\alpha}} = \tau A_{\boldsymbol{\mu}_i}^{1/2} R_{\boldsymbol{\alpha},i} A_{\boldsymbol{\mu}_i}^{1/2}$, with $A_{\boldsymbol{\mu}_i} = \text{diag}(v_{\mu_{i1}}, \ldots, v_{\mu_{in_i}})$, where $v_{\mu_{it}} = \text{var}(y_{it})$, $R_{\boldsymbol{\alpha},i}$ is the working correlation matrix and τ is a scale parameter. Only inference about the population mean is possible (population average inference). The parameters are estimated via a set of estimating equations, because there is no likelihood available in this setting.

2. **Random effects models.** With these models it is assumed that the correlation arising among repeated responses is due to the variation of the regression coefficients across individuals. One therefore models the conditional expectation of y_{it} given $\boldsymbol{\gamma}_i$ (the individuals unexplained variations) by assuming $g(E[y_{it} \mid \boldsymbol{\gamma}_i]) = \boldsymbol{x}_{it}^T \boldsymbol{\beta} + \boldsymbol{z}_{it}^T \boldsymbol{\gamma}_i$, with $\boldsymbol{\gamma}_i$ issued from a distribution F

(usually Gaussian) such that $E[\boldsymbol{\gamma}_i] = \mathbf{0}$ and $\text{var}(\boldsymbol{\gamma}_i) = \sigma_\gamma^2 \boldsymbol{I}$. This modeling approach allows for inference about individuals (subject-specific inference). Parameters estimation is performed via likelihood maximization.

3. **Transition models.** In this case, the conditional expectation given the past $E[y_{it} \mid y_{i(t-1)}, \ldots, y_{i1}]$ is modeled. The assumptions about the dependence of y_{it} on the past responses and on $\boldsymbol{x}_{it}$ are combined into a single equation, that is, the conditional expectation of y_{it} is written as an explicit function of $y_{i(t-1)}, \ldots, y_{i1}$ and $\boldsymbol{x}_{it}$. The likelihood is also the estimation method here.

6.2.1 Classical Estimation and Inference in MLDA

In this chapter, we focus on marginal models, where the final goal is to describe the population average and for which a robust procedure similar to that in Chapter 5 is available. We note at this point that some robust options exist for random effects models as well, see e.g. Mills *et al.* (2002), Sinha (2004) and Noh and Lee (2007).

The model assumptions under which we work are partially common with the main ingredients defined for GLM.

- The marginal expectation of the response $E[y_{it}] = \mu_{it}$ depends on a set of explanatory variables $\boldsymbol{x}_{it}$ via $g(\mu_{it}) = \boldsymbol{x}_{it}^T \boldsymbol{\beta}$, where g is the link function.

- The marginal variance depends on the marginal mean through the relationship $\text{var}(y_{it}) = \tau v_{\mu_{it}}$. The scale parameter τ allows for over- or under-dispersion, in the same manner as for GLMs, see Section 5.2.2.

- The correlation between y_{it} and $y_{it'}$ $(t \neq t')$ is a function of the corresponding marginal means and possibly of additional parameters $\boldsymbol{\alpha}$. This goal is achieved by parameterizing the correlation matrix with a parameter $\boldsymbol{\alpha}$ yielding a modeled covariance matrix $V_{\boldsymbol{\mu}_i, \tau, \boldsymbol{\alpha}} = \tau A_{\boldsymbol{\mu}_i}^{1/2} R_{\boldsymbol{\alpha},i} A_{\boldsymbol{\mu}_i}^{1/2}$, with $A_{\boldsymbol{\mu}_i} = \text{diag}(v_{\mu_{i1}}, \ldots, v_{\mu_{in_i}})$, where $v_{\mu_{it}} = \text{var}(y_{it})$. The modeled correlation matrix $R_{\boldsymbol{\alpha},i}$ is called the 'working' correlation matrix, as opposed to the true underlying and unknown correlation matrix $\text{corr}(\boldsymbol{y}_i)$.

The regression parameters $\boldsymbol{\beta}$ have the same interpretation as in GLM. They are regarded as the parameters of interest, whereas τ and $\boldsymbol{\alpha}$ are considered nuisance parameters. This may not be appropriate when the time course for each subject is the focus, in which case one would need to consider either the extension proposed by Zeger *et al.* (1988) or a random effects model.

Marginal models are natural extensions of GLM for dependent data. Therefore, the same or similar choices for the marginal distributions (within the exponential family) and the same link functions as in GLMs are used, see Chapter 5. However, even if a marginal distribution for y_{it} is postulated (e.g. Bernoulli, binomial, Poisson), it does not define a (unique) joint multivariate distribution for $\boldsymbol{y}_i$, making it impossible to define a likelihood function to work with. The regression parameters $\boldsymbol{\beta}$ are therefore estimated by the GEE approach of Liang and Zeger (1986). Note, however, that the GEE reduce to maximum likelihood when the $\boldsymbol{y}_i$ are multivariate

Gaussian distributed. In addition, GEE can be viewed as an extension of the quasi-likelihood approach where the variance cannot be specified only through the expectation $\boldsymbol{\mu}_i$ but rather with additional correlation parameters $\boldsymbol{\alpha}$. This similarity with the quasi-likelihood approach explains why the parameter τ is directly included in the definition of $V_{\boldsymbol{\mu}_i,\tau,\boldsymbol{\alpha}}$.

The quasi-likelihood approach used in (5.6) for GLM can be extended by solving for $\boldsymbol{\beta}$ the GEE (assuming τ and $\boldsymbol{\alpha}$ are given):

$$\sum_{i=1}^{n}(D_{\boldsymbol{\mu}_i,\boldsymbol{\beta}})^T(V_{\boldsymbol{\mu}_i,\tau,\boldsymbol{\alpha}})^{-1}(\mathbf{y}_i-\boldsymbol{\mu}_i)=\mathbf{0}, \tag{6.1}$$

where $D_{\boldsymbol{\mu}_i,\boldsymbol{\beta}}=\partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}$ and $V_{\boldsymbol{\mu}_i,\tau,\boldsymbol{\alpha}}=\tau A_{\boldsymbol{\mu}_i}^{1/2}R_{\boldsymbol{\alpha},i}A_{\boldsymbol{\mu}_i}^{1/2}$. The resulting GEE estimator for $\hat{\boldsymbol{\beta}}_{[GEE]}$ can be obtained through an IRWLS by implementing a Fisher scoring algorithm. This algorithm is given in Appendix F.1 in its more general robust form.

As said before, $R_{\boldsymbol{\alpha},i}$ is called the 'working' correlation, as opposed to the true (unknown) correlation matrix $\text{corr}(\mathbf{y}_i)$. The working correlation is imposed by the user and possible choices are as follows.

- **Independence.** Here $R_{\boldsymbol{\alpha},i}=\mathbf{I}_{n_i}$, where $\mathbf{I}_{n_i}$ is the identity matrix of size n_i. In this case, all of the set of $N=\sum_{i=1}^{n}n_i$ measurements are considered independent even within the same subject, and therefore we can treat this situation with a simple GLM model as if each observation y_{it} corresponds to independent subjects.

- **Fixed.** The correlation matrix $R_{\boldsymbol{\alpha},i}$ (or R) has a predefined form (either through a known parameter $\boldsymbol{\alpha}$ or in general). This case is rare in practice, but could be implied by a formal theory or a result of previous studies.

- **Exchangeable (or compound symmetry).** All of the correlations $(R_{\alpha,i})_{tt'}$ between two occurrences t and t' ($t\neq t'$) are assumed to be equal to a scalar value α to be estimated. Formally, $R_{\alpha,i}=\alpha\mathbf{e}_{n_i}\mathbf{e}_{n_i}^T+(1-\alpha)\mathbf{I}_{n_i}$, where $\mathbf{e}_{n_i}^T$ is a vector of ones of dimension n_i and $\mathbf{I}_{n_i}$ is the $n_i\times n_i$ identity matrix. This hypothesis may not be fulfilled when the repeated measurements are issued from subjects measured on several occasions over time, but is more appropriate in data where units are 'natural' clusters, such as children in the same class, members of a family or patients of the same practice, see e.g. the example in Section 6.2.3. Note that assuming exchangeable correlation in the normal-identity link setting corresponds to a random intercept MLM.

- **Autoregressive (AR).** The correlation decreases with time difference, e.g. $(R_{\alpha,i})_{tt'}=\alpha^{|t-t'|}$, for an unknown scalar value α. This hypothesis is quite commonly used for measurements on the same subject over time because it can accommodate an arbitrary number and spacing of observations.

- ***m*-dependence.** Observations are correlated up to time distance m, and therefore correlation is set to zero for observations that are more than m units

apart. Formally, for $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$

$$(R_{\boldsymbol{\alpha},i})_{tt'} = \begin{cases} 1 & t = t', \\ \alpha_d & d = |t - t'| \leq m, \\ 0 & \text{otherwise.} \end{cases}$$

- **Unstructured/unspecified.** The correlation matrix $R_{\boldsymbol{\alpha},i}$ is completely free (apart from a diagonal of ones and the symmetry constraint), which gives many parameters to estimate. Obviously, this option requires clusters to be of the same size, that is, $n_i = n^*$ for all i.

We refer the reader to Table 1 in Horton and Lipsitz (1999) for a description of the possible correlation structures and recommendations. Moreover, Hardin and Hilbe (2003, pp. 141–142) give additional guidelines when choosing the correlation structure, as a function of the nature of the data at hand (e.g. size of the clusters, balanced data, characteristics defining the clusters).

6.2.2 Estimators for τ and $\boldsymbol{\alpha}$

The GEE (6.1) are defined for given values of τ and $\boldsymbol{\alpha}$. A procedure that iterates between the estimation of the regression parameters $\boldsymbol{\beta}$ and the (moment) estimation of the nuisance parameters τ and $\boldsymbol{\alpha}$ is implemented in all good software and therefore used in practice. Given that τ and $\boldsymbol{\alpha}$ are nuisance parameters, less attention has been paid to their estimation and almost no theoretical results for inference exist for these parameters.

The estimation of τ is based on the fact that τ is equal to $\text{var}(\sqrt{\tau} r_{it})$, where $r_{it} = (y_{it} - \mu_{it})/\sqrt{\tau v_{\mu_{it}}}$ are the Pearson residuals for unit i at occurrence t. Therefore, a simple estimator of τ is derived from the variance estimator based on all of the N residuals, i.e.

$$\hat{\tau} = \sum_{i=1}^{n} \sum_{t=1}^{n_i} \frac{(y_{it} - \hat{\mu}_{it})^2 / v_{\hat{\mu}_{it}}}{N - (q+1)}. \tag{6.2}$$

On the other hand, the estimator of the correlation parameter $\boldsymbol{\alpha}$ depends on the choice of the correlation structure $R_{\boldsymbol{\alpha},i}$. The general approach is to estimate $\boldsymbol{\alpha}$ by a simple function of all of the pairs of residuals $\hat{r}_{it}, \hat{r}_{it'}$ that share the same correlation (t and t' defined accordingly). Below, we give some of the solutions implemented in software for the most common correlation structures.[1]

- If $(R_{\boldsymbol{\alpha},i})_{tt'} = \alpha$ (exchangeable correlation) for all $t \neq t'$, then we have

$$\hat{\alpha} = \sum_{i=1}^{n} \sum_{t>t'} \hat{r}_{it} \hat{r}_{it'} / (K - (q+1)), \tag{6.3}$$

where $K = 1/2 \sum_{i=1}^{n} n_i (n_i - 1)$ and $\hat{r}_{it} = (y_{it} - \hat{\mu}_{it})/\sqrt{\hat{\tau} v_{\hat{\mu}_{it}}}$.

[1] Note that this list is not exhaustive, and different software implement different solutions.

- If $(R_{\alpha,i})_{tt'} = \alpha_{t,t'} = \alpha^{|t-t'|}$ (AR correlation), then given that $E[r_{it}r_{it'}] \simeq \alpha^{|t-t'|}$ (because $E[r_{it}r_{it'}] \simeq \text{cov}(r_{it}, r_{it'})$), one estimates α by the slope of the regression of $\log(\hat{r}_{it}\hat{r}_{it'})$ on $\log(|t-t'|)$. Another option (see Hardin and Hilbe, 2003, p. 66) is to use

$$\hat{\alpha}_{t,t'} = \sum_{i=1}^{n} \frac{\sum_{t=1}^{n_i-(t-t')} \hat{r}_{it}\hat{r}_{it'}}{n_i}.$$

- If $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{n^*-1})$, where $\alpha_t = (R_{\boldsymbol{\alpha},i})_{t(t+1)}$ and n^* is such that $n_1 = \cdots = n_n = n^*$, then

$$\hat{\alpha}_t = \sum_{i=1}^{n} \hat{r}_{it}\hat{r}_{i(t+1)}/(n-(q+1)).$$

  In particular, if $R_{\boldsymbol{\alpha},i}$ is tridiagonal with $(R_{\boldsymbol{\alpha},i})_{t(t+1)} = \alpha_t$ (one-dependent model), then if we let $\alpha_t = \alpha$, we can estimate it by

$$\hat{\alpha} = \sum_{t=1}^{n^*-1} \hat{\alpha}_t/(n^*-1).$$

  The extension to m-dependence is possible.

- If $R_{\boldsymbol{\alpha},i}$ is totally unspecified, that is $(R_{\boldsymbol{\alpha},i})_{tt'} = \alpha_{tt'}$ for $t \neq t'$, one uses

$$\hat{R} = \frac{1}{\hat{\tau}n} \sum_{i=1}^{n} (A_{\hat{\boldsymbol{\mu}}_i})^{-1/2}(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)^T (A_{\hat{\boldsymbol{\mu}}_i})^{-1/2}.$$

For the independence, exchangeable and m-dependence correlation structure, τ does not need to be computed to solve the estimating equations (it cancels out). In contrast, it is needed when $R_{\boldsymbol{\alpha},i}$ is AR. Liang and Zeger (1986, Section 4) give further details.

The above-described estimators for τ and $\boldsymbol{\alpha}$ are moment estimators that have a closed-form, but can be expressed in an estimating equation form to be solved simultaneously with the estimating equations for $\boldsymbol{\beta}$, see Liang *et al.* (1992, pp. 9–10). The GEE approach operates as if $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ were orthogonal to each other, even when they are not, yielding less efficient estimates of $\boldsymbol{\beta}$ when the correlation structure is misspecified. Zhao and Prentice (1990) introduce a modified version of GEE, called GEE2, that relaxes the orthogonality hypothesis. The price to pay is an increased computational burden and a larger sensitivity to the misspecification of the correlation structure, see Song (2007, p. 96). The GEE2 approach is usually not what is implemented in most software and for this reason, we do not pursue this theory further.

If $\sqrt{n}$-consistent estimators are used to estimate τ and $\boldsymbol{\alpha}$, it can be proved that $\sqrt{n}(\hat{\boldsymbol{\beta}}_{[GEE]} - \boldsymbol{\beta})$ is asymptotically normally distributed with zero mean and variance

$$\Omega = \lim_{n\to\infty} M^{-1} Q M^{-1},$$

where

$$M = \frac{1}{n}\sum_{i=1}^{n}(D_{\boldsymbol{\mu}_i,\boldsymbol{\beta}})^T(V_{\boldsymbol{\mu}_i,\tau,\boldsymbol{\alpha}})^{-1}D_{\boldsymbol{\mu}_i,\boldsymbol{\beta}},$$

and

$$Q = \frac{1}{n}\sum_{i=1}^{n}(D_{\boldsymbol{\mu}_i,\boldsymbol{\beta}})^T(V_{\boldsymbol{\mu}_i,\tau,\boldsymbol{\alpha}})^{-1}\,\mathrm{var}(\boldsymbol{y}_i)(V_{\boldsymbol{\mu}_i,\tau,\boldsymbol{\alpha}})^{-1}D_{\boldsymbol{\mu}_i,\boldsymbol{\beta}},$$

see Liang and Zeger (1986, Theorem 2). Note that the asymptotic theory here is intended with respect to the number of subjects (n) and for fixed numbers of occurrences (n_i).

The estimator used for Ω is $\hat{\Omega} = \hat{M}^{-1}\hat{Q}\hat{M}^{-1}$, where

$$\hat{M} = \frac{1}{n}\sum_{i=1}^{n}(D_{\hat{\boldsymbol{\mu}}_i,\hat{\boldsymbol{\beta}}})^T(V_{\hat{\boldsymbol{\mu}}_i,\hat{\tau},\hat{\boldsymbol{\alpha}}})^{-1}D_{\hat{\boldsymbol{\mu}}_i,\hat{\boldsymbol{\beta}}}, \tag{6.4}$$

and

$$\hat{Q} = \frac{1}{n}\sum_{i=1}^{n}(D_{\hat{\boldsymbol{\mu}}_i,\hat{\boldsymbol{\beta}}})^T(V_{\hat{\boldsymbol{\mu}}_i,\hat{\tau},\hat{\boldsymbol{\alpha}}})^{-1}(\boldsymbol{y}_i-\hat{\boldsymbol{\mu}}_i)(\boldsymbol{y}_i-\hat{\boldsymbol{\mu}}_i)^T(V_{\hat{\boldsymbol{\mu}}_i,\hat{\tau},\hat{\boldsymbol{\alpha}}})^{-1}D_{\hat{\boldsymbol{\mu}}_i,\hat{\boldsymbol{\beta}}}, \tag{6.5}$$

where $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{[GEE]}$, $\hat{\boldsymbol{\mu}}_i = \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}_{[GEE]})$, $\hat{\tau}$ is defined by (6.2) and $\hat{\boldsymbol{\alpha}}$ is one of the estimators defined in the list above, depending on the assumed correlation structure.

Note that an estimator for $\mathrm{var}(\hat{\boldsymbol{\beta}}_{[GEE]})$ is $n^{-1}\widehat{\Omega}$. This is what is called in the literature a 'robust' variance estimator, in contrast to a 'naive' variance estimator that would be obtained by assuming that the working correlation is true, and hence $\mathrm{var}(\boldsymbol{y}_i) = V_{\boldsymbol{\mu}_i,\tau,\boldsymbol{\alpha}}$. This would yield $\widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}}_{[GEE]}) = n^{-1}\hat{M}^{-1}$. So, here 'robust' is intended with respect to the misspecification of the correlation structure. For a similar use of 'robust', see also the discussion in Section 7.2.4.

Approximate z-statistics and $(1-\alpha)$ CIs can be defined in the usual manner, i.e.

$$z\text{-statistic} = \frac{\hat{\beta}_{[GEE]j}}{SE(\hat{\beta}_{[GEE]j})}, \tag{6.6}$$

with $SE(\hat{\beta}_{[GEE]j}) = \sqrt{\widehat{\mathrm{var}}(\hat{\beta}_{[GEE]j})}$ and $\widehat{\mathrm{var}}(\hat{\beta}_{[GEE]j}) = n^{-1}\widehat{\Omega}_{(j+1)(j+1)}$. In the same manner, we obtain

$$(\hat{\beta}_{[GEE]j} - z_{(1-\alpha/2)}SE(\hat{\beta}_{[GEE]j});\ \hat{\beta}_{[GEE]j} + z_{(1-\alpha/2)}SE(\hat{\beta}_{[GEE]j})),$$

where $z_{(1-\alpha/2)}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution.

The GEE estimator $\hat{\boldsymbol{\beta}}_{[GEE]}$ of $\boldsymbol{\beta}$ is attractive because it presents some nice theoretical properties. For instance, the asymptotic variance of $\hat{\boldsymbol{\beta}}_{[GEE]}$ does not depend on the choice of the estimators for τ and $\boldsymbol{\alpha}$ among the $\sqrt{n}$-consistent estimators. In addition, the consistency of $\hat{\boldsymbol{\beta}}_{[GEE]}$ and $\hat{\Omega}$ depends only on the correct specification of the means $\boldsymbol{\mu}_i$ and not on the correct specification of the correlation structure. In fact, inference about $\boldsymbol{\beta}$ is valid even when the correlation matrix is not

specified correctly (see Liang and Zeger (1986) for a more detailed discussion and for the proofs of these theoretical aspects). However, a careful choice of $R_{\boldsymbol{\alpha},i}$, close to the true correlation matrix corr($\boldsymbol{y}_i$), increases efficiency, even though simulations results in Liang and Zeger (1986, Tables 1 and 2, p. 19) and Liang *et al.* (1992, Table 1, p. 15) tend to suggest otherwise. In these references the loss of efficiency is important only for highly correlated responses, but is limited for situations with moderate correlation.

The drawbacks of the GEE approach are mostly related to the lack of a likelihood function for these models, which makes diagnostic and inference limited, and to the poor theory for the nuisance parameters.

6.2.3 GUIDE Data Example

We consider the dataset of the GUIDE study (Guidelines for Urinary Incontinence Discussion and Evaluation[2] as used by Preisser and Qaqish (1999). The response variable is the coded answer (`bothered`: 1 for 'yes', 0 for 'no') of a patient to the question: 'Do you consider this accidental loss of urine a problem that interferes with your day to day activities or bothers you in other ways?'. There are five explanatory variables: gender, coded as an indicator for women (`female`), age (scaled by subtracting 76 and dividing by 10: `age`), the average number of leaking accidents per day (`dayacc`), the degree of the leak (`severe`: coded '1' for 'just create some moisture', '2' for 'wet their underwear (or pad)', '3' for 'trickle down their thigh', '4' for 'wet the floor') and the daily number of visits to the toilet to urinate (`toilet`). A total of 137 patients divided into 38 practices participated in the study.

Figure 6.1 shows the responses for each cluster. Note that here the cluster sizes are different, ranging from one to eight. On the other hand, Figure 6.2 presents a summary of all of the covariates for all of the individuals. The series of plots in the left column is for observations such that $y_{it} = 1$, that is, for patients that are bothered by their incontinence. The right column of plots is for patients with $y_{it} = 0$. We observe a strong presence of female patients in the sample and a slightly larger proportion of female (90% versus 80%) within the subsample for which $y_{it} = 0$. The age distribution is quite comparable between the two groups. On the other hand, as one can expect, the three indicators of the severity of the incontinence (`dayacc`, `severe` and `toilet`) show larger values for patients that declare themselves `bothered` by their problem (left column). For example, the median number of visits to the toilet is 6.5 for patients for which $y_{it} = 1$ versus 5 for the other group. Similarly, the median number of leaking accidents per day for the first group is 4.6 against 1 for the second group.

The model considered for this dataset is a binary logit-link model ($\tau = 1$) defined by

$$\begin{aligned}\text{logit}(E[\texttt{bothered}]) &= \text{logit}(P(\texttt{bothered})) \\ &= \beta_0 + \beta_1 \texttt{female} + \beta_2 \texttt{age} + \beta_3 \texttt{dayacc} \\ &\quad + \beta_4 \texttt{severe} + \beta_5 \texttt{toilet}, \end{aligned} \tag{6.7}$$

[2] Available at http://www.bios.unc.edu/~jpreisse/personal/uidata/preqaq99.dat

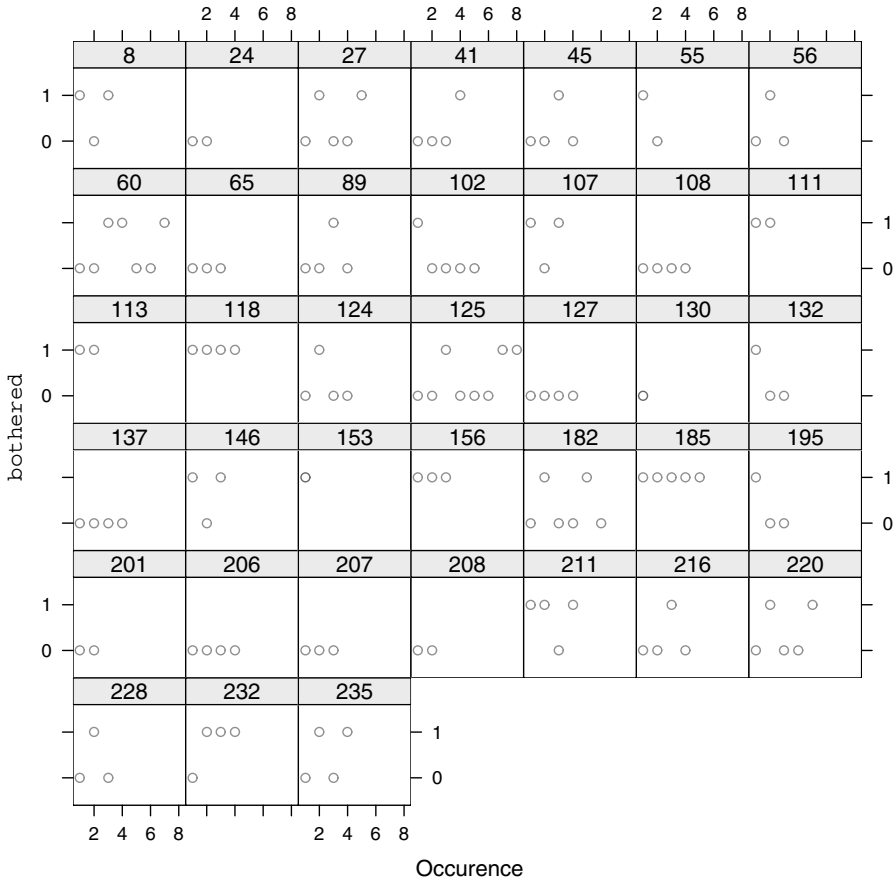

Figure 6.1 The response for the GUIDE dataset for each practice (labeled by an increasing number, appearing in the shaded box).

where $\text{logit}(p) = \log(p/(1-p))$, with $p/(1-p)$ being the odds and $p = P(\texttt{bothered})$ the probability of being `bothered`. The clusters are defined by the practice, which means that patients from different practices are assumed independent. We assume common exchangeable correlation α between any two patients of a same practice. This hypothesis makes sense *a priori* in the context of this example: in fact, even though the patients of the same practice behave independently, correlation could be induced by the fact that a physician tends to prescribe similar treatments for their patients under treatment for the same problem.

Note that the scaling of the variable `age` is not necessary but it is kept for consistency with the original analysis in Preisser and Qaqish (1999). Also, `severe` is used as a count (again for consistency with the original analysis) but should probably be put in the model as a four-level factor.

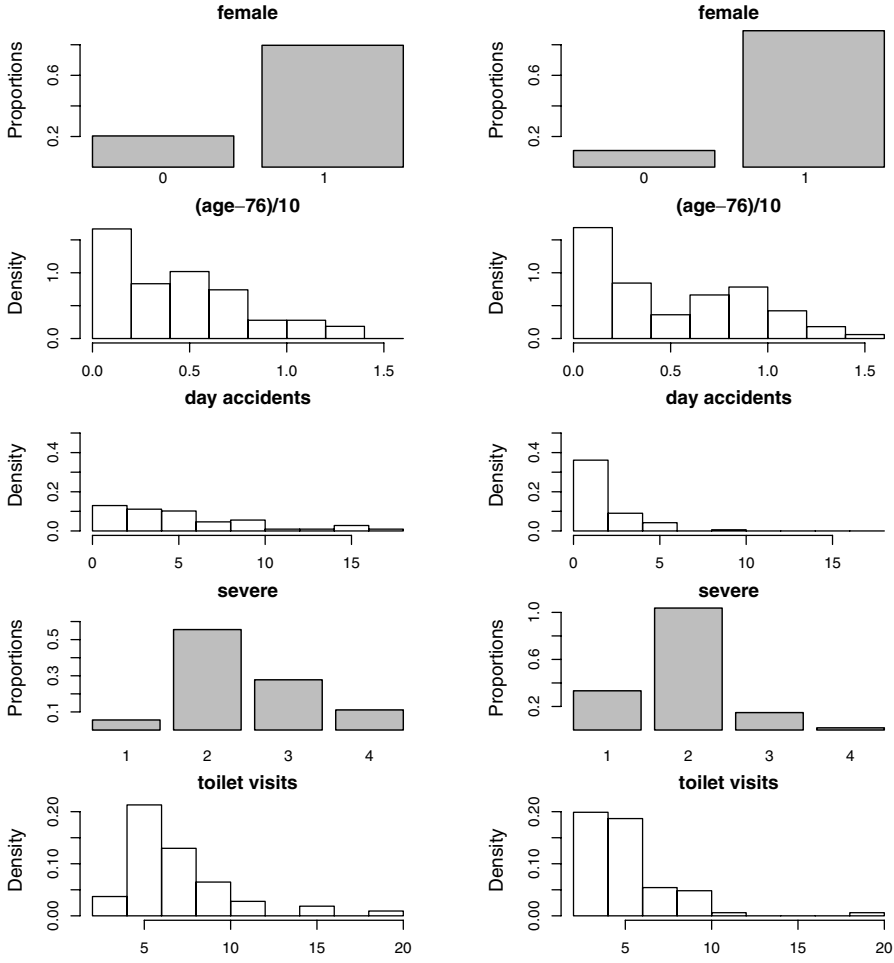

Figure 6.2 Covariates pattern for the GUIDE dataset. The left column is for observations such that $y_{it} = 1$ (54 observations out of 137) and the right column is for observations such that $y_{it} = 0$ (83 observations).

The fitted parameters of model (6.7) via classical GEE with exchangeable correlation are given in the first column of Table 6.1. We interpret the results in Section 6.3.4.

6.2.4 Residual Analysis

Residuals with longitudinal data can be considered at the observation level or at the cluster level. In both cases, the residuals proposed for GEE are similar to those used for GLMs with the additional requirement that the cluster structure has to be

Table 6.1 Estimates of α and $\boldsymbol{\beta}$ by classical and robust GEE for model (6.7).

Variable	Classical coefficient (*SE*)	Huber coefficient (*SE*)	Mallows coefficient (*SE*)
$\hat{\alpha}$	0.09	0.11	0.10
`intercept`	−3.05 (0.96)	−3.62 (1.30)	−3.63 (1.28)
`female`	−0.75 (0.60)	−1.45 (0.80)	−1.41 (0.78)
`age`	−0.68 (0.56)	−1.48 (0.71)	−1.39 (0.69)
`dayacc`	0.39 (0.09)	0.51 (0.13)	0.52 (0.13)
`severe`	0.81 (0.36)	0.71 (0.42)	0.69 (0.41)
`toilet`	0.11 (0.10)	0.36 (0.13)	0.35 (0.13)

The classical estimates are the solution of (6.1)–(6.3). The robust estimates are obtained by solving (6.8), (6.10) and (6.11) with $c = 1.5$ and $k = 2.4$ (Huber's estimator), and with $c = 1.5$, $w(\boldsymbol{x}_{it}) = \sqrt{1 - h_{i,tt}}$ and $k = 2.4$ (Mallows' estimator).

considered, see Hammill and Preisser (2006), Hardin and Hilbe (2003, Section 4.2) and Chapter 4.

As in GLMs, we define the Pearson residuals

$$\hat{r}_{it} = \frac{y_{it} - \hat{\mu}_{it}}{\sqrt{\hat{\tau} v_{\hat{\mu}_{it}}}}.$$

They can be plotted to identify outliers and other violation of the assumptions like in other regression settings (e.g. heteroscedasticity, functional form of the regression, etc.).

Figure 6.3 is a plot of the Pearson residuals for the GEE fit of the GUIDE dataset. It shows some large residuals, in particular for observations 8, 19, 42, 87 and 88. Given the fact that residuals estimated through non-robust estimators have to be analyzed with caution, in particular in regard of the possible masking effects, we defer the detailed interpretation of this residual analysis and introduce first the robust estimators.

6.3 A Robust GEE-type Estimator

6.3.1 Linear Predictor Parameters

The robust counterpart to the GEE approach is built upon the theory of optimally weighted estimating equations (see Hanfelt and Liang 1995; McCullagh and Nelder 1989, p. 334). In the class of all estimating equations based on $(\boldsymbol{y}_i - \boldsymbol{\mu}_i)$ the optimal (that is, with smallest asymptotic dispersion) estimating equations are given by

$$\sum_{i=1}^{n} (D_{\boldsymbol{\mu}_i,\boldsymbol{\beta}})^T \boldsymbol{\Gamma}_i^T (V_{\boldsymbol{\mu}_i,\tau,\boldsymbol{\alpha}})^{-1} (\boldsymbol{\psi}_i - \boldsymbol{c}_i) = \sum_{i=1}^{n} \Psi_1(\boldsymbol{y}_i, \boldsymbol{X}_i; \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau, c) = \boldsymbol{0}, \qquad (6.8)$$

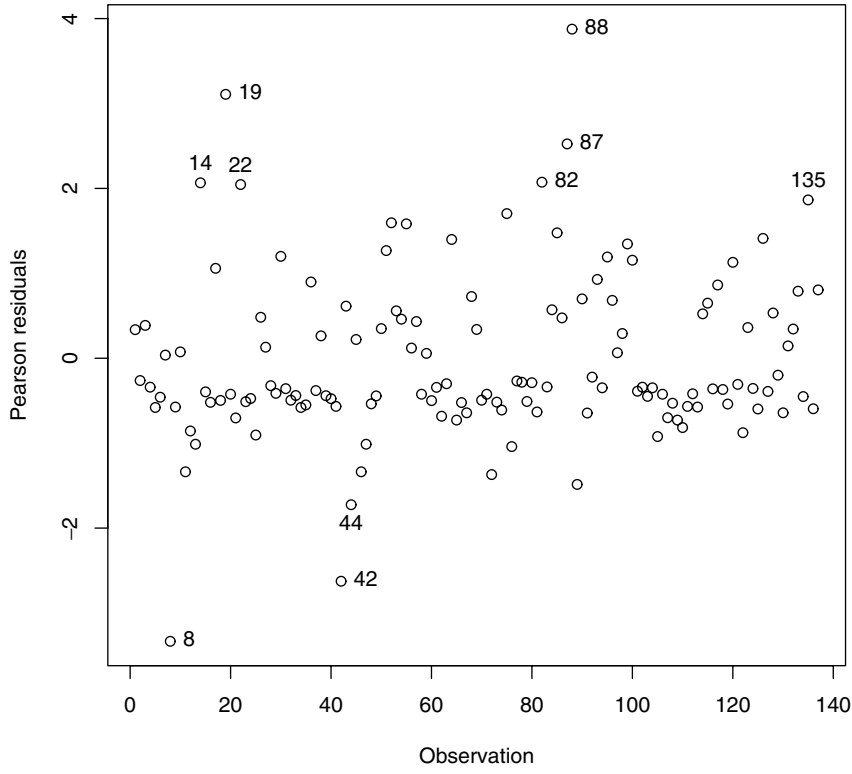

Figure 6.3 Pearson residuals corresponding to the classical GEE fit of the GUIDE dataset (first column of Table 6.1).

where $D_{\boldsymbol{\mu}_i,\boldsymbol{\beta}} = D_i(\mathbf{X}_i, \boldsymbol{\beta}) = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}$ is a $n_i \times (q+1)$ matrix,

$$V_{\boldsymbol{\mu}_i,\tau,\boldsymbol{\alpha}} = \tau A_{\boldsymbol{\mu}_i}^{1/2} R_{\boldsymbol{\alpha},i} A_{\boldsymbol{\mu}_i}^{1/2}$$

is a $n_i \times n_i$ matrix. Moreover, $\boldsymbol{\psi}_i = \mathbf{W}_i \cdot (\boldsymbol{y}_i - \boldsymbol{\mu}_i)$, where the matrix $\mathbf{W}_i = \mathbf{W}(\boldsymbol{y}_i, \mathbf{X}_i; \boldsymbol{\mu}_i) = \text{diag}(w_{i1}, \ldots, w_{in_i})$ is a $n_i \times n_i$ diagonal weight matrix containing robustness weights w_{it} for $t = 1, \ldots, n_i$, and $\boldsymbol{c}_i = E[\boldsymbol{\psi}_i]$. Finally, $\boldsymbol{\Gamma}_i = E[\tilde{\boldsymbol{\psi}}_i - \tilde{\boldsymbol{c}}_i]$ with $\tilde{\boldsymbol{\psi}}_i = \partial \boldsymbol{\psi}_i / \partial \boldsymbol{\mu}_i$ and $\tilde{\boldsymbol{c}}_i = \partial \boldsymbol{c}_i / \partial \boldsymbol{\mu}_i$. Note that the set of estimating equations in (6.8) are a slightly modified version of the estimating equations in Preisser and Qaqish (1999) in that it includes the matrix $\boldsymbol{\Gamma}_i$, which, for a given choice of weights $\mathbf{W}_i$ and 'working' correlation $R_{\boldsymbol{\alpha},i}$, makes it optimal (in the sense of smallest asymptotic dispersion) in the class of all estimating equations based on $(\boldsymbol{y}_i - \boldsymbol{\mu}_i)$, see Hanfelt and Liang (1995). The computational details of $\boldsymbol{c}_i$ and $\boldsymbol{\Gamma}_i$ for binary responses are given in Appendix F.2.

We assume that the weights $\mathbf{W}_i$ downweight each observation separately, even though it is possible to consider a cluster downweighting scheme, see the discussion

about observation versus cluster outliers in Section 6.2.4. Possible choices for the weights are $w(r_{it}; \boldsymbol{\beta}, \tau, c)$ as a function of the Pearson residuals $r_{it} = (y_{it} - \mu_{it})/\sqrt{\tau v_{\mu_{it}}}$, for example Huber's weight (see also (2.16))

$$w(r_{it}; \boldsymbol{\beta}, \tau, c) = \begin{cases} c/|r_{it}/\sqrt{\tau}| & \text{if } |r_{it}/\sqrt{\tau}| > c, \\ 1 & \text{otherwise,} \end{cases} \tag{6.9}$$

to ensure robustness with respect to outlying points in the response space (Huber's estimator), or $w(\boldsymbol{x}_{it})$ as a function of the diagonal elements $h_{i,tt}$ of the hat matrix $\mathbf{H}_i$ (see (3.11)) for subject i (for example, $w(\boldsymbol{x}_{it}) = \sqrt{1 - h_{i,tt}}$) to handle leverage points. In practice, it often makes sense to combine both types of weights multiplicatively: $w_{it} = w(r_{it}; \boldsymbol{\beta}, \tau, c)w(\boldsymbol{x}_{it})$ (Mallows' esitmator). The classical GEE are obtained with $\mathbf{W}_i$ equal to the identity matrix. We refer to Cantoni and Ronchetti (2001b) for a detailed discussion on the choice of the weights.

For simplicity, our weighting scheme (as in Preisser and Qaqish, 1999) does not take into account the within-subject correlation and is therefore not suitable for the situation where this correlation is high, in which case it has to be redefined properly, see for example Huggins (1993) and Richardson and Welsh (1995). Doing so will change the definition in (6.8) and affect the distributional properties. Note, however, that protection against outliers affecting all of the observations of a cluster can be handled by our approach by specifying a cluster downweighting scheme, that is, with $w_{it} = w_i^*$ for all $t = 1, \ldots, n_i$, where the w_i^* have to be defined to take into account the information of the entire cluster.

The estimating equations (6.8) do not simplify exactly to the estimating equations (5.13) for GLMs owing to the presence of the matrix $\boldsymbol{\Gamma}_i$ in the former. The presence of this matrix in the GEE setting is necessary to allow the construction of the quasi-deviance functions for inference (see Section 6.4.2).

6.3.2 Nuisance Parameters

The estimators of the dispersion parameter τ and of the correlation parameter $\boldsymbol{\alpha}$ also have to be made robust to avoid harmful consequences on the estimation of the regression parameters. We build again on the fact that the parameter τ is the variance of $(y_{it} - \mu_{it})/\sqrt{\tau v_{\mu_{it}}} = \sqrt{\tau} r_{it}$, see Section 6.2.2. We therefore proceed similarly as for GLM and choose Huber's Proposal 2 estimator of variance (see Section 5.3.3), which is written here as

$$\sum_{i=1}^{n} \sum_{t=1}^{n_i} \chi(r_{it}; \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau, c) = \sum_{i=1}^{n} \Psi_2(\boldsymbol{r}_i; \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau, c) = 0, \tag{6.10}$$

where $\chi(u; \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau, c) = \psi^2(u; \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau, c) - \delta$. In addition, $\delta = E[\psi^2(u; \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau, c)]$ (under normality for u) is a constant that ensures Fisher consistency of the estimation of τ. For its computation for $\psi^2(u; \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau, c) = \psi_{[Hub]}(u; \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau, c)$ (our preferred choice), see (3.8), while noticing that $\psi_{[Hub]}(u; \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau, c) = u w_{[Hub]}(u; \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau, c)$.

As in the classical GEE theory, the estimator of the correlation parameter $\boldsymbol{\alpha}$ depends on the assumed correlation structure. To build a robust estimator of $\boldsymbol{\alpha}$, the

idea is to base this estimator on appropriate pairs of residuals, along the same line as for the classical estimators (see Section 6.2.2), but to consider additional weighting schemes to downweight outlying observations.

In the following we discuss in detail the case of exchangeable correlation and explain how one can deal with two other common situations, namely the m-dependence and the AR correlation structures. Let us recall that the exchangeable correlation structure defines $R_{\boldsymbol{\alpha},i} = \alpha \boldsymbol{e}_{n_i} \boldsymbol{e}_{n_i}^T + (1-\alpha)\mathbf{I}_{n_i}$, with $\boldsymbol{e}_{n_i}$ a vector of ones of length n_i, and $\mathbf{I}_{n_i}$ the identity matrix of size $n_i \times n_i$, which means that $\text{corr}(y_{it}, y_{it'}) = \alpha$ for $t \neq t'$, and one otherwise. A simple M-estimator of covariance can be defined through Huber's type of weights (based on $\psi_{[Hub]}(\cdot; \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau, c)$) which we define as functions of the Mahalanobis distance $d^i_{tt'}$ (see (2.34)) between the pair of residuals $\hat{r}_{it}$ and $\hat{r}_{it'}$. The Mahalanobis distance is given in this case by $(d^i_{tt'})^2 = (\hat{r}_{it}\ \hat{r}_{it'})\ \hat{\boldsymbol{\Sigma}}^{-1}\ (\hat{r}_{it}\ \hat{r}_{it'})^T$ with

$$\hat{\Sigma} = \hat{\tau}_{[M]} \begin{pmatrix} 1 & \hat{\alpha}_{[M]} \\ \hat{\alpha}_{[M]} & 1 \end{pmatrix}.$$

We define Huber's weights on the Mahalanobis distances by

$$u_{1,k}(d^i_{tt'}) = \begin{cases} 1 & \text{if } d^i_{tt'} \leq k, \\ k/d^i_{tt'} & \text{otherwise.} \end{cases}$$

We then put $u_{2,k}(d^i_{tt'}) = u_{1,k^2}(d^i_{tt'})/\gamma$ with $\gamma = E[r u_{1,k^2}(|r|)]/2$ where the expectation is computed under normality for r. This yields $\gamma = F_{\chi^2_4}(k^2) + k^2/2(1 - F_{\chi^2_2}(k^2))$, where $F_{\chi^2_4}$ and $F_{\chi^2_2}$ are the cumulative distribution function of a χ^2 distribution with four and two degrees of freedom, respectively. Let $\boldsymbol{B}_i = (\hat{r}_{i1} \cdot \hat{r}_{i2}, \hat{r}_{i1} \cdot \hat{r}_{i3}, \ldots, \hat{r}_{i(n_i-1)} \cdot \hat{r}_{in_i})^T$ be the vector of the product of all of the pairs of residuals for cluster i and let $\boldsymbol{G}_i = (u_{2,k}(d^i_{12}), u_{2,k}(d^i_{13}), \ldots, u_{2,k}(d^i_{(n_i-1)n_i}))^T$ be the vector of weights, then our robust estimator of α is defined as the solution $\hat{\alpha}_{[M]}$ of

$$\sum_{i=1}^{n} \left(\boldsymbol{G}_i^T \boldsymbol{B}_i - \frac{K}{n} \alpha \tau \right) = \sum_{i=1}^{n} \Psi_3(\boldsymbol{r}_i; \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau, c) = 0, \tag{6.11}$$

with $K = \sum_{i=1}^{n} n_i(n_i - 1)/2$. For more details on all of the above computations we refer to Maronna (1976), Devlin *et al.* (1981) and Marazzi (1993, p. 225).

M-estimators are known to have a low breakdown point, namely one over the dimension of the problem, which is equal to two here (see the discussion of this point in Section 2.3.1). Nevertheless, high breakdown point estimators could be considered to estimate $\boldsymbol{\Sigma}$. An *ad hoc* estimator of α in the case of binary responses with exchangeable correlation inspired by the classical moment estimator is considered by Preisser and Qaqish (1999). This proposal relies on the hypothesis that $\text{var}(\boldsymbol{\psi}_i)$ can be decomposed as $C_i\ \text{var}(\boldsymbol{y}_i) C_i$ and therefore the proposal cannot be extended to other settings, e.g. Poisson. Our proposal is more general and has the advantage of inheriting the whole set of distributional properties of M-estimators. It is also worth mentioning that all $u_{1,k}(d^i_{tt'}) = 1$, and therefore all $u_{2,k}(d^i_{tt'}) = 1$ gives the usual (classical) moment estimators for these situations.

Two other common correlation structures are the m-dependence correlation structure, which assumes that $\text{corr}(y_{it}, y_{i,t+j}) = \alpha_j$, for $j = 1, \ldots, m$, and the AR correlation structure which assumes that $\text{corr}(y_{it}, y_{i,t+j}) = \alpha^j$ for $j = 0, 1, \ldots, n_i - t$. The procedure described above can be adapted to these cases by constructing $\boldsymbol{B}_i$ appropriately, that is, with all of the products $\hat{r}_{it} \cdot \hat{r}_{i,t+j}$ in the first case, and $\hat{r}_{it} \cdot \hat{r}_{i,t+1}$ in the latter. The correction terms have to be defined accordingly.

6.3.3 *IF* and Asymptotic Properties

Under standard regularity conditions we have that $(\sqrt{n}(\hat{\boldsymbol{\beta}}_{[M]} - \boldsymbol{\beta})^T, \sqrt{n}(\hat{\tau}_{[M]} - \tau)^T, \sqrt{n}(\hat{\boldsymbol{\alpha}}_{[M]} - \boldsymbol{\alpha})^T)^T$, with $\hat{\boldsymbol{\beta}}_{[M]}$, $\hat{\tau}_{[M]}$ and $\hat{\boldsymbol{\alpha}}_{[M]}$ defined through (6.8), (6.10) and (6.11), respectively, follows an asymptotic normal distribution with mean zero and covariance matrix

$$\lim_{n\to\infty} \begin{pmatrix} F & 0 & 0 \\ G & H & 0 \\ J & L & N \end{pmatrix}^{-1} \begin{pmatrix} \Theta_{(11)} & \Theta_{(12)} & \Theta_{(13)} \\ \Theta_{(12)}^T & \Theta_{(22)} & \Theta_{(23)} \\ \Theta_{(13)}^T & \Theta_{(23)}^T & \Theta_{(33)} \end{pmatrix} \begin{pmatrix} F & 0 & 0 \\ G & H & 0 \\ J & L & N \end{pmatrix}^{-\mathrm{T}}, \qquad (6.12)$$

where all of the sub-matrices in (6.12) are given in Cantoni (2004b) (up to a factor $1/n$, with $\Theta = \Phi$), where the proof of the distributional result is also given. The particular form of the matrices in (6.12) implies that the marginal asymptotic distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}}_{[M]} - \boldsymbol{\beta})$ is normal with mean zero and variance equal to

$$\Upsilon = \lim_{n\to\infty} F^{-1}\Theta_{(11)}F^{-T}, \qquad (6.13)$$

where

$$F = \frac{1}{n}\sum_{i=1}^{n}(D_{\boldsymbol{\mu}_i,\boldsymbol{\beta}})^T \boldsymbol{\Gamma}_i^T (V_{\boldsymbol{\mu}_i,\tau,\boldsymbol{\alpha}})^{-1}\boldsymbol{\Gamma}_i D_{\boldsymbol{\mu}_i,\boldsymbol{\beta}}, \qquad (6.14)$$

and

$$\Theta_{(11)} = \frac{1}{n}\sum_{i=1}^{n}(D_{\boldsymbol{\mu}_i,\boldsymbol{\beta}})^T \boldsymbol{\Gamma}_i^T (V_{\boldsymbol{\mu}_i,\tau,\boldsymbol{\alpha}})^{-1} \text{var}(\boldsymbol{\psi}_i)(V_{\boldsymbol{\mu}_i,\tau,\boldsymbol{\alpha}})^{-1}\boldsymbol{\Gamma}_i D_{\boldsymbol{\mu}_i,\boldsymbol{\beta}}. \qquad (6.15)$$

The distributional result in (6.12) generalizes the results of Prentice (1988): it applies to other types of responses than Bernoulli trials, it allows for an over-dispersion parameter (τ) and is developed in the more general setting of robust estimating equations defined by (6.8), (6.10) and (6.11).

In addition, the estimating equations (6.8), (6.10) and (6.11) define a set of M-estimators (Huber, 1981), with the corresponding score functions $\Psi_1(\boldsymbol{y}_i, \boldsymbol{X}_i; \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau, c)$, $\Psi_2(\boldsymbol{r}_i; \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau, c)$, $\Psi_3(\boldsymbol{r}_i; \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau, c)$ in Appendix F.1. From general theory on M-estimation, we know that the *IF* of these estimators is proportional to the score functions defining them. Therefore, the estimators obtained by our procedure are robust as long as the functions Ψ_i are bounded in the design and in the response. This is in particular achieved if $\boldsymbol{\psi}_i$ in (6.8) and χ in (6.10) are bounded, and $u_{2,k}$ through $\boldsymbol{G}_i$ in (6.11) are allowed to be less than one.

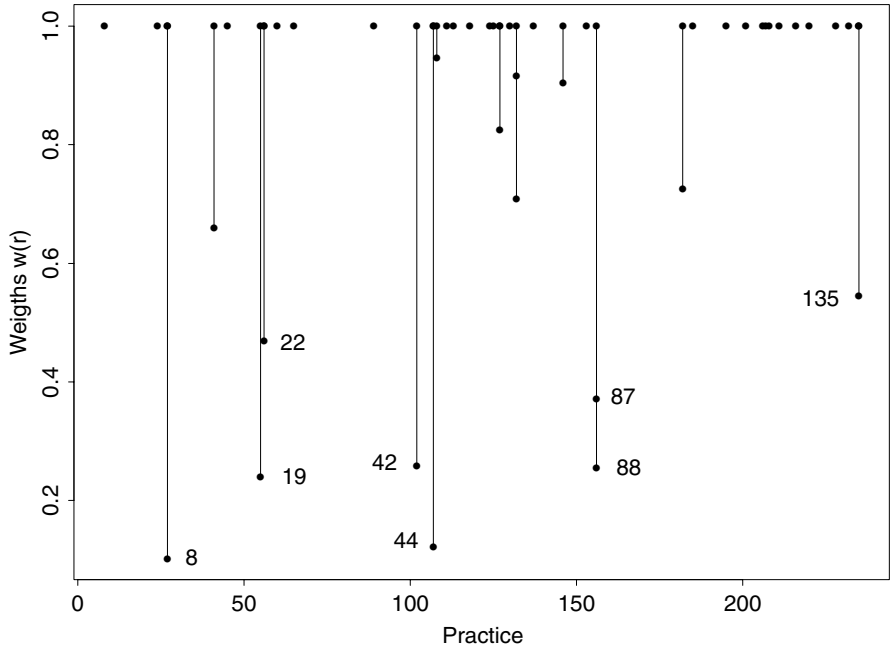

Figure 6.4 Robustness weights on the response, grouped by practice, for the fit corresponding to the middle column of Table 6.1 (Huber's estimator).

6.3.4 GUIDE Data Example (continued)

We estimate the regression parameters with the set of equations in (6.8), where we consider both a Mallows' estimator with $w(\boldsymbol{x}_{it}) = \sqrt{1 - h_{i,tt}}$ and $c = 1.5$ and a Huber estimator with $w(\boldsymbol{x}_{it}) = 1$ and $c = 1.5$. In both cases, the exchangeable correlation parameter α is estimated through (6.11) with $k = 2.4$, which is approximately the 95%-quantile of a χ^2_2 distribution. In addition to the classical results already presented in Section 6.2.3, Table 6.1 gives the estimated coefficients for the two robust alternatives. First note that the results of the second and third column (robust analyses) are quite close, whereas they differ noticeably from the classical analysis. This means that the additional weights on the design are probably not crucial in the sense that the dataset does not seem to contain leverage points. By looking at approximate CIs (see their definition in Section 6.4.1), the variables `female` and `age` are not significant in the classical analysis, but are borderline in the robust analysis. The significance of the variable `dayacc` seems to be equally well assessed in both types of analysis. The variable `severe` is no longer significant in the robust analysis, whereas the variable `toilet` seems to play an important role that was hidden in the classical approach.

The robust procedure also gives information on how many and which observations are downweighted. For example, in the analysis with weights on the response only (middle column of Table 6.1), there are 15 observations out of 137 that do not receive full weight, 8 of which have weight less than 0.6, see Figure 6.4. This group of observations is partially the same as that identified in Preisser and Qaqish (1999) with their robust procedure. The diagnostic approach in Hammill and Preisser (2006) identify as potential outliers a smaller group of observations, in particular patients 8 and 44. These two patients, together with patient 42, report not being bothered despite their high frequency of visits to the toilet (10 for patients 8 and 42, and 20 for patient 44) and their large average number of leaking accidents per day (9.3 for patient 8, 6 for patient 42 and 3 for patient 44). On the other hand, patients 19 and 88 declared themselves bothered, even though the severity of their symptoms (variables `severe` and `toilet`) are pretty low with respect to the other sample values.

Only two of these heavily downweighted observations belong to the same practice (cluster), namely observations 87 and 88 from practice 156, confirming that the individual downweighting scheme is justified with this dataset.

6.4 Robust Inference

6.4.1 Significance Testing and CIs

The z-test for significance testing and $(1-\alpha)$ CIs for the regression parameters $\boldsymbol{\beta}$ can be constructed based on the asymptotic distribution of the estimator, see Section 6.3.3.

The z-statistics and $(1-\alpha)$ CIs are given by

$$z\text{-statistic} = \frac{\hat{\beta}_{[M]j}}{SE(\hat{\beta}_{[M]j})},$$

and

$$(\hat{\beta}_{[M]j} - z_{(1-\alpha/2)}SE(\hat{\beta}_{[M]j}); \hat{\beta}_{[M]j} + z_{(1-\alpha/2)}SE(\hat{\beta}_{[M]j})),$$

with

$$SE(\hat{\beta}_{[M]j}) = \sqrt{\frac{1}{n}[\hat{\Upsilon}]_{(j+1)(j+1)}},$$

where $z_{(1-\alpha/2)}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution, and where $\hat{\Upsilon} = \hat{F}^{-1}\hat{\Theta}_{(11)}\hat{F}^{-T}$, with

$$\hat{F} = \frac{1}{n}\sum_{i=1}^{n}(D_{\hat{\boldsymbol{\mu}}_i,\hat{\boldsymbol{\beta}}})^T\boldsymbol{\Gamma}_i^T(V_{\hat{\boldsymbol{\mu}}_i,\hat{\tau},\hat{\boldsymbol{\alpha}}})^{-1}\boldsymbol{\Gamma}_i D_{\hat{\boldsymbol{\mu}}_i,\hat{\boldsymbol{\beta}}}, \tag{6.16}$$

and

$$\hat{\Theta}_{(11)} = \frac{1}{n}\sum_{i=1}^{n}(D_{\hat{\boldsymbol{\mu}}_i,\hat{\boldsymbol{\beta}}})^T\boldsymbol{\Gamma}_i^T(V_{\hat{\boldsymbol{\mu}}_i,\hat{\tau},\hat{\boldsymbol{\alpha}}})^{-1}(\boldsymbol{\psi}_i - \boldsymbol{c}_i)(\boldsymbol{\psi}_i - \boldsymbol{c}_i)^T(V_{\hat{\boldsymbol{\mu}}_i,\hat{\tau},\hat{\boldsymbol{\alpha}}})^{-1}\boldsymbol{\Gamma}_i D_{\hat{\boldsymbol{\mu}}_i,\hat{\boldsymbol{\beta}}}, \tag{6.17}$$

where $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{[M]}$, $\hat{\boldsymbol{\mu}}_i = \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}_{[M]})$, $\hat{\tau} = \hat{\tau}_{[M]}$ and $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}_{[M]}$.

6.4.2 Variable Selection

Variable selection is performed here by comparing the adequacy of a submodel $\mathcal{M}_{q-k+1}$ with $(q-k+1)$ regression parameters with respect to a larger model $\mathcal{M}_{q+1}$ with $(q+1)$ regression parameters. This is done either in a stepwise procedure, or by comparing two predefined nested models. For that we define a class of test statistics based on differences of quasi-likelihoods, in the same spirit as the difference of quasi-deviances in (5.24) for GLMs in Chapter 5:

$$\Lambda_{\boldsymbol{t}(s)} = 2\left\{\sum_{i=1}^{n} Q_{\boldsymbol{t}_i(s)}(\boldsymbol{y}_i; \hat{\boldsymbol{\mu}}_i) - \sum_{i=1}^{n} Q_{\boldsymbol{t}_i(s)}(\boldsymbol{y}_i; \dot{\boldsymbol{\mu}}_i)\right\}, \tag{6.18}$$

where $\hat{\boldsymbol{\mu}}_i = \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}_{[M]}, \hat{\boldsymbol{\alpha}}_{[M]}, \hat{\tau}_{[M]})$ is the estimation under model $\mathcal{M}_{q+1}$, and where

$$\dot{\boldsymbol{\mu}}_i = \boldsymbol{\mu}_i(\dot{\boldsymbol{\beta}}_{[M]}, \dot{\boldsymbol{\alpha}}_{[M]}, \dot{\tau}_{[M]})$$

is the estimation under model $\mathcal{M}_{q-k+1}$ and where the quasi-likelihood functions take the multidimensional form

$$\begin{aligned} &Q_{\boldsymbol{t}_i(s)}(\boldsymbol{y}_i; \boldsymbol{\mu}_i) \\ &\quad = \frac{1}{\tau}\int_{\boldsymbol{y}_i}^{\boldsymbol{\mu}_i} (\boldsymbol{y}_i - \boldsymbol{t}_i)^T \boldsymbol{W}(\boldsymbol{y}_i, \boldsymbol{X}_i; \boldsymbol{t}_i(s))(V_{\boldsymbol{t}_i(s),\tau,\boldsymbol{\alpha}})^{-1}\boldsymbol{\Gamma}_i(\boldsymbol{t}_i(s))\, d\boldsymbol{t}_i(s) \\ &\qquad - \frac{1}{\tau}\int_{\boldsymbol{y}_i}^{\boldsymbol{\mu}_i} E[(\boldsymbol{y}_i - \boldsymbol{t}_i(s))^T \boldsymbol{W}(\boldsymbol{y}_i, \boldsymbol{X}_i; \boldsymbol{t}_i(s))](V_{\boldsymbol{t}_i(s),\tau,\boldsymbol{\alpha}})^{-1}\boldsymbol{\Gamma}_i(\boldsymbol{t}_i(s))\, d\boldsymbol{t}_i(s), \end{aligned} \tag{6.19}$$

with the integrals possibly path-dependent. This means that there are several paths to go from a point $\boldsymbol{y}_i$ to a point $\boldsymbol{\mu}_i$ and implies, therefore, that the integrals in (6.19) are not uniquely defined. It is common practice to parameterize this path and a typical set of integration paths is given for example by $t_{it}(s) = y_{it} + (\mu_{it} - y_{it})s^{c_{it}}$, for $s \in [0,1]$, $c_{it} \geq 1$ and $t = 1, \ldots, n_i$. For instance, when $c_{it} \equiv 1$ for all t (see for example McCullagh and Nelder, 1989, p. 335), we have that

$$\begin{aligned} &Q_{\boldsymbol{t}_i(s)}(\boldsymbol{y}_i; \boldsymbol{\mu}_i) \\ &\quad = -\frac{1}{\tau}(\boldsymbol{y}_i - \boldsymbol{\mu}_i)^T\left[\int_0^1 s\boldsymbol{W}(\boldsymbol{y}_i, \boldsymbol{X}_i; \boldsymbol{t}_i(s))(V_{\boldsymbol{t}_i(s),\tau,\boldsymbol{\alpha}})^{-1}\boldsymbol{\Gamma}_i(\boldsymbol{t}_i(s))\, ds\right](\boldsymbol{y}_i - \boldsymbol{\mu}_i) \\ &\qquad + \frac{1}{\tau}\int_0^1 E[\{\boldsymbol{y}_i - \boldsymbol{t}_i(s)\}^T \boldsymbol{W}(\boldsymbol{y}_i, \boldsymbol{X}_i; \boldsymbol{t}_i(s))](V_{\boldsymbol{t}_i(s),\tau,\boldsymbol{\alpha}})^{-1}\boldsymbol{\Gamma}_i(\boldsymbol{t}_i(s))(\boldsymbol{y}_i - \boldsymbol{\mu}_i)\, ds, \end{aligned}$$

which involves only univariate integrations, uniquely defined. The asymptotic result from Section 6.4.2.1 shows that the path-dependence of the integrals in (6.19) vanishes asymptotically. In addition, Hanfelt and Liang (1995) showed that the path of integration does not play an important role in finite-sample situations. These results support the use of the difference of robust quasi-likelihoods for inference.

6.4.2.1 Multivariate Testing

Multivariate testing of the type $H_0 : \boldsymbol{\beta}_{(2)} = 0$ with $\boldsymbol{\beta} = (\boldsymbol{\beta}_{(1)}, \boldsymbol{\beta}_{(2)})$ and with $\boldsymbol{\beta}_{(1)}$ of dimension $(q + 1 - k)$ and $\boldsymbol{\beta}_{(2)}$ of dimension k, can be performed using $\Lambda_{t(s)}$ defined by (6.18) as a test statistic. Cantoni (2004b) proves that under quite general conditions (and under H_0), $\Lambda_{t(s)}$ is asymptotically equivalent to the following quadratic forms in normal variables

$$n\mathbf{L}_n^T(M^{-1} - \tilde{M}^+)\mathbf{L}_n = n\mathbf{R}_{n(2)}^T M_{22.1}\mathbf{R}_{n(2)}, \tag{6.20}$$

where

$$M = \lim_{n\to\infty} F = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} (D_{\boldsymbol{\mu}_i,\boldsymbol{\beta}})^T \boldsymbol{\Gamma}_i^T (V_{\boldsymbol{\mu}_i,\tau,\boldsymbol{\alpha}})^{-1} \boldsymbol{\Gamma}_i D_{\boldsymbol{\mu}_i,\boldsymbol{\beta}}$$

is partitioned into four blocks according to the partition of $\boldsymbol{\beta}$:

$$\begin{pmatrix} M_{(11)} & M_{(12)} \\ M_{(12)}^T & M_{(22)} \end{pmatrix}$$

and

$$\tilde{M}^+ = \begin{pmatrix} M_{(11)}^{-1} & 0_{(q-k+1)\times k} \\ 0_{k\times(q-k+1)} & 0_{k\times k} \end{pmatrix},$$

where $0_{a\times b}$ is a matrix of dimension $a \times b$ with only zero entries.

The variables $\sqrt{n}\mathbf{L}_n$ and $\sqrt{n}\mathbf{R}_n$ are asymptotically normally distributed $\mathcal{N}(0, Q)$ and $\mathcal{N}(0, M^{-1}QM^{-1})$, respectively, where

$$Q = \lim_{n\to\infty} \Theta_{(11)} = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} D_{\boldsymbol{\mu}_i,\boldsymbol{\beta}}^T \boldsymbol{\Gamma}_i^T (V_{\boldsymbol{\mu}_i,\tau,\boldsymbol{\alpha}})^{-1} \operatorname{var}(\boldsymbol{\psi}_i)(V_{\boldsymbol{\mu}_i,\tau,\boldsymbol{\alpha}})^{-1} \boldsymbol{\Gamma}_i D_{\boldsymbol{\mu}_i,\boldsymbol{\beta}}.$$

This implies that $\Lambda_{t(s)}$ is asymptotically distributed as linear combination of χ_1^2 variables, similarly as for GLMs (see Section 5.4.2). More precisely, $\Lambda_{t(s)}$ is asymptotically distributed as $\sum_{i=1}^{k} d_i N_i^2$, where $N_1, \ldots, N_k$ are independent standard normal variables, $d_1, \ldots, d_k$ are the k positive eigenvalues of the matrix $Q(M^{-1} - \tilde{M}^+)$. In practice, the empirical version of M and Q are used, that is, $\hat{M} = \hat{F}$ (see (6.16)) and $\hat{Q} = \hat{\Theta}_{(11)}$ (see (6.17)).

In addition to giving the asymptotic distribution, the above result provides an asymptotically equivalent quadratic form to $\Lambda_{t(s)}$, which can be used as an asymptotic approximation when the integrals involved in the definition of $\Lambda_{t(s)}$ are problematic to compute. More precisely, one computes $n\hat{\boldsymbol{\beta}}_{M(2)}^T \hat{M}_{22.1} \hat{\boldsymbol{\beta}}_{M(2)}$.

Finally, Cantoni (2004b) proves that the level and the power of $\Lambda_{t(s)}$ under contamination are bounded provided that $\hat{\boldsymbol{\beta}}_{M(2)}$ has a bounded *IF*.

6.4.3 GUIDE Data Example (continued)

Let us consider a backward stepwise procedure based on the difference of quasi-likelihoods functions defined by (6.18) to check more carefully the issues related

Table 6.2 p-values of the backward stepwise procedure on the GUIDE dataset.

	Variable	Step 1	Step 2	Step 3	Step 4
Classical	`female`	0.224	0.270	–	–
	`age`	0.249	–	–	–
	`dayacc`	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
	`severe`	0.089	0.081	0.061	0.011
	`toilet`	0.224	0.164	0.165	–
Robust	`female`	0.070	0.095	–	–
	`age`	0.045	0.041	0.068	–
	`dayacc`	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$
	`severe`	0.092	–	–	–
	`toilet`	0.006	0.004	0.004	0.002

The robust test statistics (6.18) are computed by applying Huber's-type weights ($c = 1.5$), and by using $k = 2.4$ for the estimation of α in (6.11) (exchangeable correlation). The classical test statistics are computed with $c = \infty$ and $k = \infty$.

to model selection. We use the same weights and the same set of parameters as for the Huber's estimator of Section 6.3.4, and compute the quadratic form (6.20) asymptotically equivalent to $\Lambda_{t(s)}$. At each step of the procedure, we remove the variable that is the least significant by looking at the p-value or, equivalently, at the value of the test statistic. The procedure is stopped when all of test statistics are significant at the 5% level. The classical counterpart is computed with the same quadratic form, by using $c = \infty$ and $k = \infty$ to compute the estimators.

Table 6.2 gives the p-values of this backward stepwise procedure. It is impressive to see how the classical p-values differ from the robust p-values. This highlights the heavy influence of outlying observations on the test procedure and not only on the estimation procedure. Finally, the robust procedure ends up by retaining the variables `dayacc` and `toilet`, whereas the classical analysis would retain the variables `dayacc` and `severe`. On the basis of the theoretical properties of the robust estimator, and also on the simulations results in Cantoni (2004b), the conclusions from the robust analysis are more reliable. We therefore robustly refit the model with only `dayacc` and `toilet` and proceed with interpretations from this model. The estimated coefficients and standard errors for the linear predictors are as follows:

$$\underset{(0.76)}{-3.67} \;+\; \underset{(0.12)}{0.49\,\texttt{dayacc}} \;+\; \underset{(0.10)}{0.29\,\texttt{toilet}}.$$

The estimated model in this clustered setting can be interpreted in the same way as for GLMs. In this example, the response is binary, and therefore the discussion of Section 5.5.2 about interpreting the coefficients on the odds scale still holds. The effect of an additional leaking accident per day is to increase by 63% ($\exp(0.49) = 1.63$) the odds of a patient being bothered by their incontinence problem. Similarly, the effect of an extra visit to the toilet results in a 34% increase ($\exp(0.29) = 1.34$)

on the same odds. This second effect is smaller in magnitude, which seems compliant with common sense.

6.5 LEI Data Example

We consider here a dataset on direct laryngoscopic endotracheal intubation (LEI), a potentially life-saving procedure in which many health care professionals are trained. We examine data from a prospective longitudinal study on LEI at Dalhousie University, previously analyzed by Mills *et al.* (2002). Variable selection is an important step as the model(s) chosen will include only those covariates significant in predicting successful completion of LEI.

A total of 436 LEIs were analyzed. We let $y_{it} = 1$ if trainee i performs a complete LEI in less than 30 seconds on trial t, and zero otherwise. The correlation between observations on the same trainee is taken to be exchangeable. An AR correlation structure would be another option with these data. We judge trainees based on the following nine binary covariates taking the value one if the condition is satisfied: whether the head and neck were in optimal position (`neckflex` and `extoa`); whether they inserted the scope properly (`proplgsp`); whether they performed the lift successfully (`proplift`); whether there was appropriate request for help (`askas`); whether there was unsolicited intervention by the attending anesthesiologist (`help`); whether there were no complications (`comps`); and the trainee's handedness (`trhand`) and gender (`trgend`). Nineteen trainees performed anywhere from 18 to 33 trials.

Figure 6.5 gives the pattern profiles for the 19 trainees. These patterns tend to show that training results in better performance over time, see for example profiles of trainees K, L, VV and Z. It seems less evident for other individuals, namely AA and S. Table 6.3 presents a summary of all of the (binary) covariates for all of the individuals. As naturally expected, the proportion of ones (indicating that successful action has been taken or that no complications were observed) is larger for individual that have succeeded in performing a complete LEI.

We fitted robustly a GEE model with exchangeable correlation to the above data. The estimates and test results are given in Table 6.4. The robust GEE model uses a Huber's estimator with $c = 1.5$ for the Huber function and $k = 2.4$ for the Huber's Proposal 2 (6.10). No weights on the design has been used here given the binary nature of all the covariates.

A priori we would expect all of the coefficients to be positive, which would indicate that if proper action is taken, the probability of success in performing LEI increases. It is indeed the case expect for a few non-significant coefficients (`askas` and `extoa`). A classical approach (not shown here) would give substantially different estimated coefficients. The standard errors of the MLE would also be quite larger, which is a serious drawback when performing significance testing.

Figure 6.6 gives the weights $w(r_{it}; \boldsymbol{\beta}, \tau, c)$ from the robust fit. The two main outliers are observations 11 (11th trial of trainee AA) and 273 (14th trial of trainee T). The first observation corresponds to the only successful LEI for trainee AA in

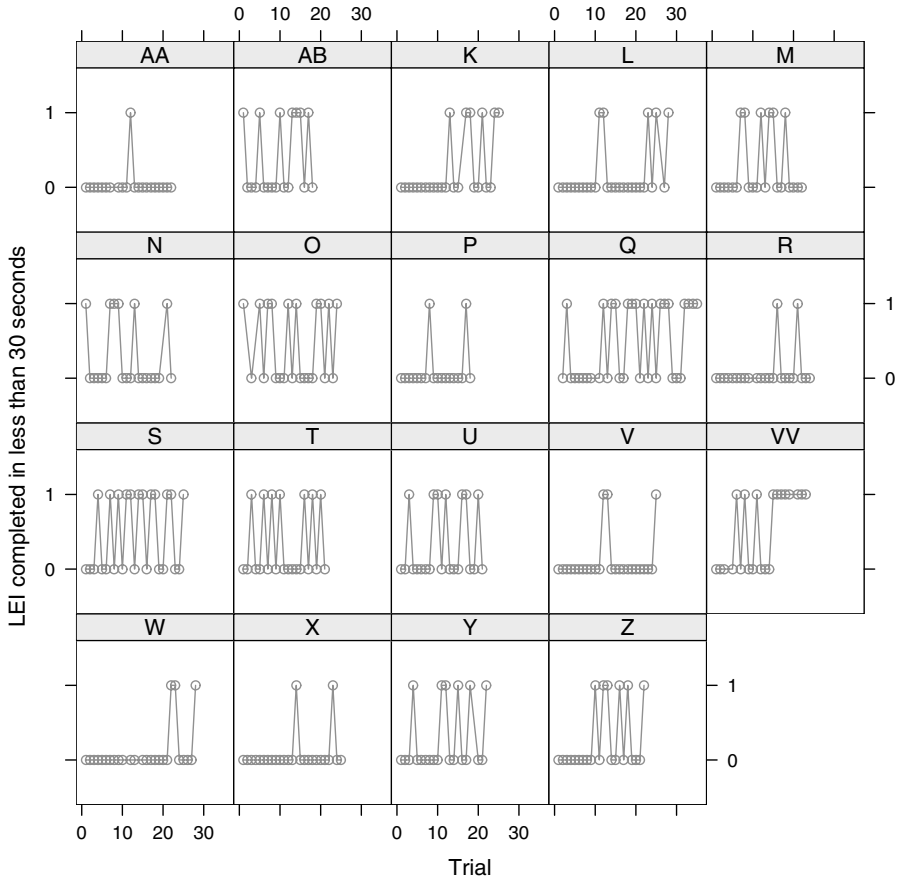

Figure 6.5 LEI responses (one for completed in less than 30 seconds) for each trainee, labeled by capital letters.

21 trials (see Figure 6.5) for a covariate pattern for this trainee which is quite stable through the trials (not shown) and can therefore not explain the different response. The second observation is an unsuccessful LEI, even though the covariates pattern would have called for a success.

The significant variables stemming from the robust approach on the basis of their z-statistic are `neckflex`, `proplgsp`, `proplift`, `help` and perhaps `comps`. In the classical analysis, `neckflex` would have been considered non-significant, as would `comps`.

The significance of all of the variables except `comps` is clearcut. We therefore only test three particular nested models with the difference of quasi-deviances (6.18) with Huber's weights with $c = 1.5$: the full model including all of the available covariates, against the submodel without `extoa`, `askas`, `trhand`, `trgend`

Table 6.3 Covariates characteristics for the LEI dataset.

	Successful LEI (118 observations)		Unsuccessful LEI (318 observations)	
Variable	Proportion of ones	Proportion of zeros	Proportion of ones	Proportion of zeros
neckflex	0.99	0.01	0.95	0.05
extoa	0.99	0.01	0.97	0.03
proplgsp	0.86	0.14	0.52	0.48
proplift	0.88	0.12	0.39	0.61
askas	0.15	0.85	0.20	0.80
help	0.70	0.30	0.37	0.63
comps	0.95	0.05	0.78	0.22
trhand	0.82	0.18	0.84	0.16
trgend	0.77	0.23	0.69	0.31

Table 6.4 Robust GEE fits for the LEI dataset.

Variable	Coefficient (*SE*)	*p*-value
intercept	−4.18 (0.51)	$<10^{-4}$
neckflex	1.52 (0.39)	$<10^{-4}$
extoa	−0.24 (0.41)	0.56
proplgsp	0.69 (0.20)	0.0007
proplift	0.98 (0.25)	$<10^{-4}$
askas	−0.42 (0.26)	0.11
help	0.34 (0.12)	0.004
comps	0.99 (0.49)	0.04
trhand	0.04 (0.26)	0.89
trgend	0.05 (0.24)	0.84
α	0.061	

The estimates are obtained by solving (6.8), (6.10) and (6.11) with $c = 1.5$ and $k = 2.4$ (Huber's estimator).

(the clearly non-significant covariates) and the submodel that in addition remove comps. Table 6.5 gives the *p*-values associated with these tests. It confirms that the submodel without extoa, askas, trhand and trgend is enough to represents the relationship that describes a successful LEI. The robust analysis also shows the importance of the variable comps, given the rejection of the null hypothesis that its coefficient is equal to zero.

The estimated final model therefore yields the following coefficients and standard errors for the linear predictor:

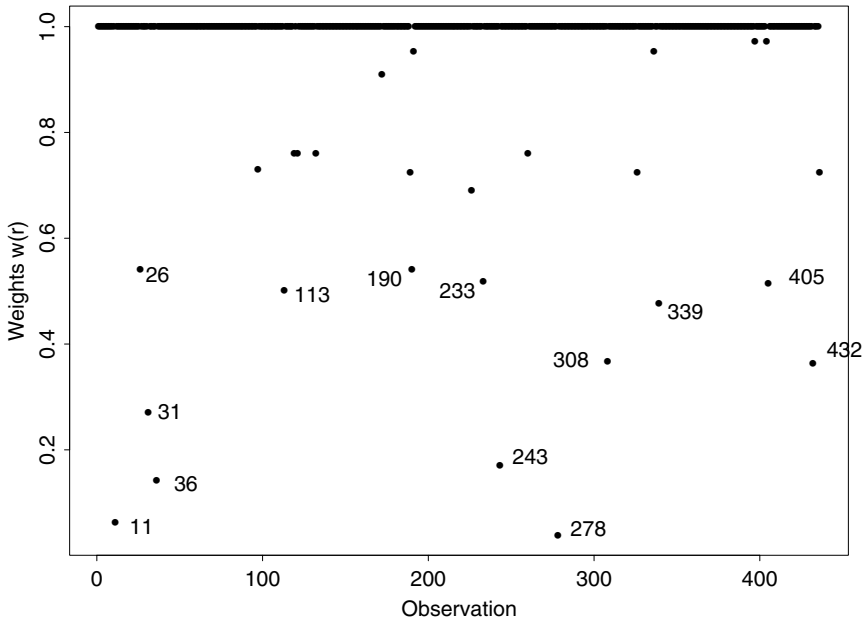

Figure 6.6 Robust weights for the LEI data example.

Table 6.5 Robust p-values for comparison of models based on the difference of quasi-deviances (6.18).

Model	$\Lambda_{\boldsymbol{t}(s)}$	p-value
Full		
- `extoa` - `askas` - `trhand` - `trgend`	4.49	0.36
- `extoa` - `askas` - `trhand` - `trgend` - `comps`	2.76	0.01

The robust test statistics (6.18) are computed by applying Huber's-type weights ($c = 1.5$) and by using $k = 2.4$ for the estimation of α in (6.11) (exchangeable correlation).

$$\underset{(1.23)}{-9.17} + \underset{(0.75)}{3.78}\,\texttt{neckflex} + \underset{(0.35)}{1.33}\,\texttt{proplgsp} + \underset{(0.50)}{1.93}\,\texttt{proplift} + \underset{(0.26)}{0.60}\,\texttt{help} + \underset{(0.77)}{1.93}\,\texttt{comps}.$$

The multiplicative effects of a positive action taken by the trainee or the fact that there was no complications (in which cases the covariate is equal to one) on the odds of succeeding in performing a LEI are as follows (exponential of the coefficient):

`neckflex`	`proplgsp`	`proplift`	`help`	`comps.`
43.69	3.79	6.89	1.82	6.90

In addition to the statistical significance, we can see that the strongest effect on the odds of a successful LEI is definitely the proper positioning of the neck, followed

by the correct lift and the absence of complications. Inserting the scope properly and asking for help were also positively associated with a successful LEI, but the associations were somehow weaker.

6.6 Stillbirth in Piglets Data Example

Genetic selection is an important research domain in animal science. It allows species to be selected with 'stronger' characteristics. For example, for most mammalian species, farrowing is a critical period. In pigs, for example, up to 8% of newborns are stillborn. Limiting or reducing the number of stillbirths requires its major determinants to be investigated.

This section is devoted to the study of stillbirth in four genetic types of sow: Duroc $\times$ Large White (DU $\times$ LW), Large White (LW), Meishan (MS) and Laconie (LA). Data are from the INRA GEPA experimental unit (France) and have been kindly provided by L. Canario and Y. Billon. Related publications are Canario (2006) and Canario *et al.* (2006) where the reader can find a more extensive discussion of the modeling issues for this dataset. Previous studies have shown that parity number, piglet birth weight, sex and birth assistance were associated with perinatal mortality. The aim of the study is to establish whether there is a genotype effect, in view of possible genetic selection (e.g. development of crossed-synthetic lines).

Our dataset comprises 80 litters for the genetic type (coded `gentype`) DU $\times$ LW, 633 litters for LW, 59 litters for MS and 168 litters for LA, for a total of 940 litters and 11 638 observations. There were 565 deaths (coded = 1) out of the 11 638. The genetic type LW is taken as the reference. Parity number, the number of times a mother has given birth (variable `parity`, taken as a factor), ranges from one to six with the following corresponding frequencies (35%, 26%, 15%, 12%, 8%, 4%), with one taken as the reference. Birth assistance (variable `birthassist`) is coded zero for no assistance and one for one or several assistances. The cluster is defined as the litter, which size varies from 5 to 23.

We fit a binary logit-link model with exchangeable correlation. For the robust fit we use weights $w(r_{it}; \boldsymbol{\beta}, \tau, c)$ on the residuals with a tuning constant $c = 1.5$ for the Huber function. We use $k = 2.4$ for the Huber's Proposal 2 (6.10). The estimated coefficients, standard errors and p-values for z-test for significance on each coefficient ($H_0 : \beta_j = 0$) are given in Table 6.6. The robust analysis shows that piglets born from the MS genetic type have a lower risk of stillbirth with respect to LW. The odds ratio of a stillbirth for the MS genotype with respect to the LW genotype is equal to $\exp(-1.71) = 0.18$. Also, the mortality increases with parity, at least for the 5th and 6th parity, which could result from the fatness of old sows or aging of the uterus (or both). The estimated exchangeable correlation is $\hat{\alpha} = 0.035$, which is low.

The conclusions, however, have to be taken with caution. A careful inspection of the weights associated by the robust technique to the observations show a particular pattern. Indeed, in Figure 6.7, one can see that the downweighted observations identify a subpopulation of the data, in fact all of the 565 observations corresponding

Table 6.6 Robust estimates for the piglets dataset.

Variable	Coefficient (*SE*)	*p*-value
`intercept`	−3.00 (0.11)	$<10^{-4}$
factor(`gentype`)DU × LW	−0.20 (0.19)	0.31
factor(`gentype`)MS	−1.71 (0.43)	$<10^{-4}$
factor(`gentype`)LA	0.11 (0.14)	0.45
factor(`parity`)2	−0.23 (0.15)	0.12
factor(`parity`)3	0.10 (0.17)	0.57
factor(`parity`)4	0.15 (0.17)	0.38
factor(`parity`)5	0.38 (0.21)	0.08
factor(`parity`)6	0.55 (0.20)	0.005
`birthassist`	0.13 (0.12)	0.30

The estimates are obtained by solving (6.8), (6.10) and (6.11) with $c = 1.5$ and $k = 2.4$ (Huber's estimator).

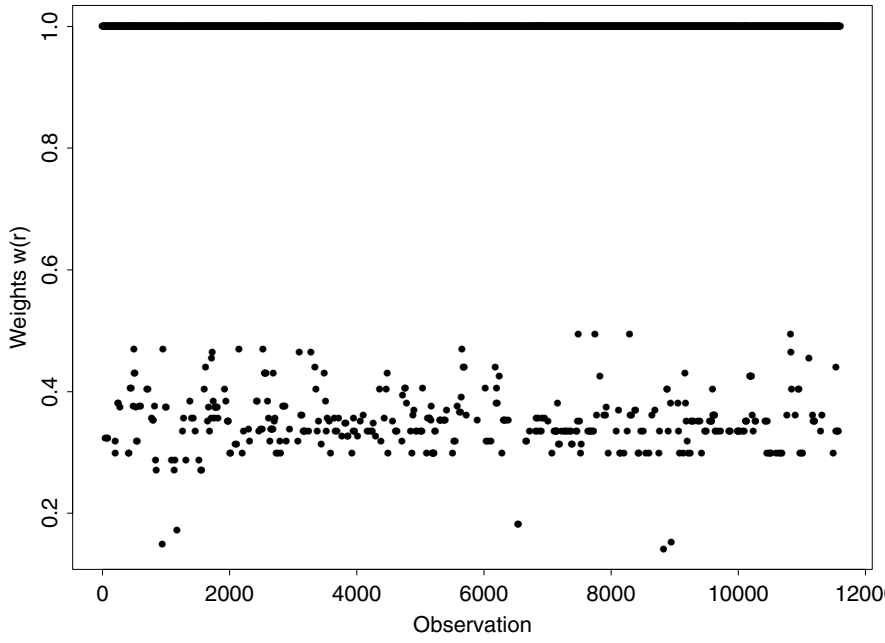

Figure 6.7 Robustness weights on the response for the piglets dataset.

to a death (response = 1). Further investigations allowed us to identify suspected separation or near-separation in the data. This peculiarity of binary regression is a situation where the design space of the observations for which $y = 1$ and the observations for which $y = 0$ can be completely separated by a hyperplan.

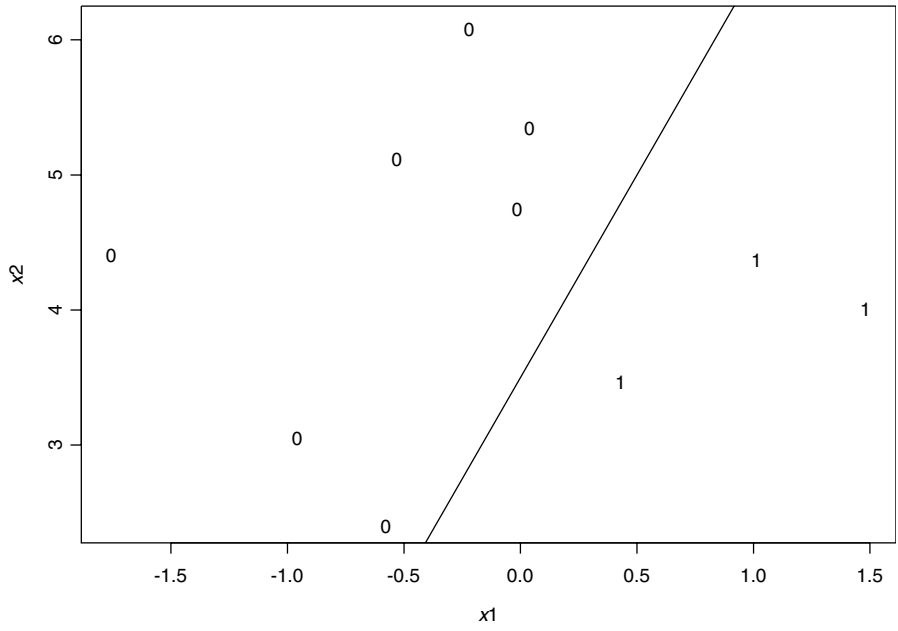

Figure 6.8 Illustration of a situation with no overlap in binary regression: the observations for which $y = 0$ and the observations for which $y = 1$ can be completely separated by a hyperplan.

For example, if there are two covariates x_1 and x_2, this would correspond to the situation depicted in Figure 6.8. We say that there is no overlap for this dataset (see also the illustration in Christmann and Rousseeuw (2001, pp. 67–69)).

Christmann and Rousseeuw (2001) give an algorithm to compute the overlap, that is, the smallest number of observations whose removal yields complete or quasi-complete separation. In these cases, most estimators do not exist. In cases where the overlap is very small, the estimators exist but can potentially be very unstable. In addition, robust estimators work by downweighting (or sometimes removing) outlying points. It can therefore happen that the whole dataset has overlap, but that the robust estimators do not exist. The methodology by Christmann and Rousseeuw (2001) was used on the piglets dataset to compute the overlap, which is equal to eight. This is particularly related to the binary/categorical nature of the data. A (limited) sensitivity analysis has nevertheless shown that some stability is present and that therefore the study provides useful conclusions.

In this analysis the robust methodology has helped in highlighting a peculiar feature of the data that could lead to disastrous conclusions if it remains undetected.

6.7 Discussion and Extensions

At the time of writing, only the Bernoulli family has been implemented for the robust estimation and inference for GEE. Note, however, that the theory presented in this chapter is general and includes all GLM distributions. The difficulty arising in practice is the computation of the correction term $\boldsymbol{c}_i$ in (6.8). This difficulty can be circumvented by computing the correction term by simulation. This is currently work in progress.

As mentioned in Section 5.7.3, Cantoni *et al.* (2005) develop a criterion, called GC_p, inspired by Mallows's C_p for general model comparisons. It is given in (5.29) and the general form for an unbiased estimator GC_p is given in (5.30). The particular form of GC_p for a Mallows's quasi-likelihood estimator as defined by (6.8) is given by Cantoni *et al.* (2005), where their extensive simulation study shows that the GC_p is very effective in handling contaminated data.

7

Survival Analysis

7.1 Introduction

Survival analysis is central to biostatistics and modeling such data is an important part of the work carried out daily by statisticians working with clinicians and medical researchers. Basically, survival data analysis is necessary each time a survival time or a time to a specific event (failure) such as organ dysfunction, disease progression or relapse is the outcome of interest. Such data are often censored as not all subjects enrolled in the study experience the event. When investigators are interested in testing the effect of a particular treatment on failure time the default method of analysis is the log-rank test, usually supplemented by Kaplan–Meier survival estimates. The log-rank test is, by definition, based on ranks and therefore offers some degree of protection against outliers. Criticisms have been raised (Kim and Bae, 2005) but the test is not as sensitive as most of the standard testing procedures in other models. When the outcome has to be explained by a set of predictors, the standard approach is the Cox (1972) proportional hazard model. Cox regression is appealing owing to its flexibility in modeling the instantaneous risk of failure (e.g. death) or hazard, even in the presence of censored observations. This interest toward the Cox model goes well beyond the world of medicine and biostatistics. Applications in biology, engineering, psychology, reliability theory, insurance and so forth can easily be found in the literature. Its uniqueness also stems from the fact that it is not, strictly speaking, based on maximum likelihood theory but on the concept of partial likelihood. This notion was introduced by Cox in his original paper to estimate the parameters of interest in a semi-parametric formulation of the instantaneous risk of failure at a specific time point, given that such an event has not occurred so far.

Over the years many papers dealing with various misspecifications in the Cox model have been published. Diagnostic techniques have also flourished boosted by the ever-growing number of applications related to that model; see, for instance,

Robust Methods in Biostatistics S. Heritier, E. Cantoni, S. Copt and M.-P. Victoria-Feser

Chen and Wang (1991), Nardi and Schemper (1999), Therneau and Grambsch (2000), Collett (2003b), Wang *et al.* (2006) for a review. In the 1980s, researchers were typically interested in whether a consistent estimate of a treatment effect could be obtained when omitting a covariate. Work by Gail *et al.* (1984), Bretagnolle and Huber-Carol (1988) and others show both theoretically and through simulations that if important predictors are omitted from the Cox model, a small resulting bias occurs in the estimate. Later, Lin and Wei (1989) propose a sandwich formula for the treatment effect estimator's variance, which they call 'robust' in that significance testing of the treatment effect has approximately the desired level, even if important predictors are omitted from the model. They also claim that their variance estimator can cope with possible misspecifications of the hazard function. As argued in Section 1.2, this type of robustness is different from that discussed in this book. Robustness methods in the modern sense of the word have been relatively slow to emerge in survival analysis, hindered by the presence of censoring that is unaccounted for by the general robustness theory. Regarding the Cox model, another complication stems from its semi-parametric nature. This is in essence different from the fully parametric setting discussed at length in the previous chapters. In the early 1990s, researchers such as Hjort (1992) and Schemper (1992) started to tackle the problem but the first real attempts to robustify Cox's partial likelihood appeared in Bednarski (1993), Sasieni (1993b,a) and Minder and Bednarski (1996). A complex methodology is generally required to cope with censoring. Bednarski's work is based on a doubly weighted partial likelihood and extends the *IF* approach presented in Chapter 2. Later Grzegorek (1993) and Bednarski (1999, 2007) refined their weighting estimation technique to make it adaptive and invariant to time-transformation. A good account on how outliers affect the estimation process in practical terms with an illustration on clinical data is given in Valsecchi *et al.* (1996). A comparison of Bednarski's approach and the work by Sasieni and colleagues is carried out in Bednarski and Nowak (2003). It essentially shows that none of these estimators clearly outperforms the others as far as problems in the response is the primary target. This technical literature focuses only on the estimation problem prompting questions about the robustness of tests as defined in Chapter 2. Recent work by Heritier and Galbraith (2008) illustrates the current limitations of robust testing for this model and clarifies the link with the theory by Lin and Wei (1989).

Independently of all of these developments related to the Cox model, an innovative technique called regression quantiles appears in the late 1970s that seems totally unrelated to survival analysis. That pioneer work due to Koenker and Bassett (1978) is introduced in the econometric literature as a robust alternative to linear regression. In this work, any percentile of a particular outcome (e.g a survival time), or a transformation of it, can be regressed on a set of explanatory variables. This work in itself and many subsequent papers would not be sufficient to be mentioned in this chapter, if an extension to the censored case had not been proposed. Fortunately, such a method, called censored regression quantiles, now exists. The extension, due to Portnoy (2003), has a great potential in practice. It is easily computable, inherits the robustness of the sample quantiles, and constitutes a viable alternative to the Cox model, especially when the proportional hazard assumption is not met.

This chapter is organized as follows. Cox's partial likelihood and the classical theory is reviewed in Section 7.2. The so-called robust sandwich formula of Lin and Wei (1989) and its link to the *IF* is presented. The lack of robustness properties of standard estimation and inferential procedures are motivated by the analysis of the myeloma data. A robust (adaptive) estimator based on the work of Bednarski and colleagues is presented and illustrated in Section 7.3. Issues related to robust testing in the Cox model and its current limitations are also discussed. A complete worked-out example using the well-known veterans' administration lung cancer data (see e.g. Kalbfleisch and Prentice, 1980) is described in Section 7.4. Other issues including model misspecifications are outlined in Section 7.5. Finally, Section 7.6 is devoted to censored regression quantiles. We first introduce quantile regression, discuss its extension to censored data and apply the method to the lung cancer dataset.

7.2 The Cox Model

7.2.1 The Partial Likelihood Approach

As mentioned earlier, the proportional hazard model introduced by Cox (1972) is probably the most commonly used model to describe the relationship between a set of covariates and survival time or another time-to-event, possibly censored. Let (t_i, δ_i) be independent random variables recording the survival time and absence of censoring ($\delta_i = 1$ if t_i is observed, 0 otherwise) for a sample of n individuals. It is convenient to write $t_i = \min(t_i^0, c_i)$ where t_i^0 is the possibly unknown survival time and c_i the censoring time. The t_i^0 are independent random variables from a cumulative distribution $F(\cdot \mid \boldsymbol{x}_i)$ with density $f(\cdot \mid \boldsymbol{x}_i)$, where $\boldsymbol{x}_i$ is a q-dimensional vector of fixed covariates. For simplicity we consider the standard case where all time points are different and ordered, i.e. $t_1 < t_2 < \cdots < t_n$. We also assume that the censoring mechanism is non-informative. The Cox model relates the survival time t to the covariates $\boldsymbol{x}$ through the hazard function of F[1]

$$\lambda(t \mid \boldsymbol{x}) = \lambda_0(t) \exp(\boldsymbol{x}^T \boldsymbol{\beta}), \tag{7.1}$$

where $\lambda_0(t)$ is the so-called baseline hazard, usually unspecified, and $\boldsymbol{\beta}$ the regression parameter vector.[2] In essence, $\lambda(t \mid \boldsymbol{x})$ measures the instantaneous risk of death (or hazard rate) at time t for an individual with specific characteristics described by $\boldsymbol{x}$, given that they have survived so far. The interesting feature of formulation (7.1) is that $\lambda(t \mid \boldsymbol{x})$ is the product of a baseline hazard $\lambda_0(t)$ and an exponential term depending on the covariates. This has two major advantages. First, as we will see in Section 7.2.1, it is not necessary to know the baseline hazard $\lambda_0(t)$ to estimate the coefficients $\boldsymbol{\beta}$. Second, we can derive immediately the effect of an increase of one unit in a particular covariate x_j (e.g. the effect of an experimental treatment represented by a binary indicator: one for treatment, zero for placebo) on survival.

[1] The hazard function of a distribution function F with density f is $f(t)/(1 - F(t))$.

[2] Note that, by writing (7.1) on the log scale, $\log(\lambda_0(t))$ can be seen as the intercept term added to the linear predictor $\boldsymbol{x}^T \boldsymbol{\beta}$, so that $\dim(\boldsymbol{\beta}) = q$.

Indeed, such an increase translates into a constant relative change $\exp(\beta_j)$ of the hazard $\lambda(t)$. This quantity is the hazard ratio (HR) and is usually interpreted as the relative risk of death related to an increment of 1 of that particular predictor. This property justifies the terminology proportional hazard model, commonly used for the Cox model. Model (7.1) encompasses two important parametric models, namely the exponential regression model for which $\lambda_0(t) = \lambda$ and the Weibull regression model for which $\lambda_0(t) = \lambda\gamma t^{\gamma-1}$. However, in a full parametric setting, the additional parameters λ and/or γ need to be estimated along with the slopes $\boldsymbol{\beta}$ for these models to be fully specified. In the proposal of Cox (1972), this is not necessary.

Equation (7.1) can be expressed equivalently through the survival function $S(t \mid \boldsymbol{x}) = 1 - F(t \mid \boldsymbol{x})$ (see, for instance, Collett (2003b) and Therneau and Grambsch (2000)) given by

$$S(t \mid \boldsymbol{x}) = \{S_0(t)\}^{\exp(\boldsymbol{x}^T\boldsymbol{\beta})}, \tag{7.2}$$

where $S_0(t)$ is defined through

$$-\log(S_0(t)) = \int_0^t \lambda_0(u)\,du = \Lambda_0(t), \tag{7.3}$$

and $\Lambda_0(t)$ is the baseline cumulative hazard obtained by integrating $\lambda_0(u)$ between zero and t.

The usual estimate of $\boldsymbol{\beta}$ is the parameter value that maximizes the partial likelihood

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n}\left[\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\sum_{j\geq i}\exp(\boldsymbol{x}_j^T\boldsymbol{\beta})}\right]^{\delta_i} \tag{7.4}$$

or equivalently the solution of the first order equation

$$\sum_{i=1}^{n}\delta_i\left[\boldsymbol{x}_i - \frac{S^{(1)}(t_i;\boldsymbol{\beta})}{S^{(0)}(t_i;\boldsymbol{\beta})}\right] = 0 \tag{7.5}$$

where

$$S^{(0)}(t_i;\boldsymbol{\beta}) = \sum_{j\geq i}\exp(\boldsymbol{x}_j^T\boldsymbol{\beta}) \quad \text{and} \quad S^{(1)}(t_i;\boldsymbol{\beta}) = \sum_{j\geq i}\exp(\boldsymbol{x}_j^T\boldsymbol{\beta})\boldsymbol{x}_j$$

as in Minder and Bednarski (1996) and Lin and Wei (1989).[3] The solution of (7.5) is the partial likelihood estimator (PLE) also denoted by $\hat{\boldsymbol{\beta}}_{[PLE]}$. Equation (7.5) is a simple rewriting of a more conventional presentation as in, for instance, Collett (2003b, Chapter 3). There, the risk set $R(t_i)$ at time t_i is used, i.e. the set of all patients who have not yet achieved the event by time t_i and are then still 'at risk' of dying. It is formed of all observations with indices greater than or equal to i.

[3]The idea is to base the likelihood function on the probability for a subject to achieve the event by time t_i. This is given by the ratio of the hazard at time t_i of the subject i over the hazard of all subjects who have not yet experienced the event by time t_i, i.e. the set $j \geq i$ (also called the risk set). In this ratio, the baseline hazard cancels out and one obtains the expression given in (7.4).

The purpose of writing (7.5) in that way is to stress the similarity with the definition of an M-estimator. Indeed, let

$$U_i = U(t_i, \delta_i, \boldsymbol{x}_i; \boldsymbol{\beta}) = \delta_i \left[\boldsymbol{x}_i - \frac{S^{(1)}(t_i; \boldsymbol{\beta})}{S^{(0)}(t_i; \boldsymbol{\beta})} \right]$$

be the individual contribution (or score) then, at least literally, (7.5) looks like an M-estimator with Ψ-function $U(t, \delta, \boldsymbol{x}; \boldsymbol{\beta})$. The main difference is that the scores U_i are no longer independent since the two sums $S^{(0)}(t_i; \boldsymbol{\beta})$ and $S^{(1)}(t_i; \boldsymbol{\beta})$ depend on subsequent time points t_j for $j \geq i$.

We have assumed, for simplicity, that all observed time points are different in the sample. The partial likelihood approach is generally modified to handle ties. We refer the reader to Therneau and Grambsch (2000) and Collett (2003b) for a more general introduction and Kalbfleisch and Prentice (1980) for technical details.

Under some regularity conditions, the PLE is asymptotically normally distributed with asymptotic variance $V = I(\boldsymbol{\beta})^{-1}$, where $I(\boldsymbol{\beta})$ is the information matrix for that model (see Kalbfleisch and Prentice (1980, Chapter 4) for details). Here $I(\boldsymbol{\beta})$ is usually estimated by minus the second derivative of the average log partial likelihood, i.e.

$$\widehat{I}(\boldsymbol{\beta}) = -\frac{1}{n} \sum_{i}^{n} \frac{\partial U_i}{\partial \boldsymbol{\beta}}. \tag{7.6}$$

Numerical values are obtained by replacing $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}_{[PLE]}$ in (7.6). An alternative formula for the variance of $\hat{\boldsymbol{\beta}}_{[PLE]}$ will be given in Section 7.2.4. The asymptotic distribution can then be used for testing single hypothesis $H_0 : \beta_j = 0$ in a standard way. One just defines a z-statistic as

$$z\text{-statistic} = \frac{\hat{\beta}_{[PLE]j}}{SE(\hat{\beta}_{[PLE]j})}, \tag{7.7}$$

where

$$SE(\hat{\beta}_{[PLE]j}) = \sqrt{n^{-1}[\widehat{I}(\hat{\boldsymbol{\beta}}_{[PLE]})^{-1}]_{jj}} \tag{7.8}$$

is the (estimated) standard error of $\hat{\boldsymbol{\beta}}_{[PLE]j}$, i.e. the square root of the jth diagonal element of (7.6). Here z is compared with a standard normal distribution.

More generally, standard asymptotic tests such as the LRT, score and Wald tests are available to test a composite null hypothesis of the type $H_0 : \boldsymbol{\beta}_{(2)} = \boldsymbol{\beta}^0_{(2)}$, with $\boldsymbol{\beta}_{(1)}$ unspecified and with $\boldsymbol{\beta} = (\boldsymbol{\beta}^T_{(1)}, \boldsymbol{\beta}^T_{(2)})^T$. Specifically, the LRT is equal to twice the difference in the maximum log-partial-likelihood obtained at the full and reduced model, i.e.

$$\text{LRT} = 2(\log(L(\hat{\boldsymbol{\beta}}_{[PLE]})) - \log(L(\dot{\boldsymbol{\beta}}_{[PLE]})) \tag{7.9}$$

where $\hat{\boldsymbol{\beta}}_{[PLE]}$ denotes the PLE at the full, model and $\dot{\boldsymbol{\beta}}_{[PLE]}$ its value under the null hypothesis (at the reduced model).

The Wald test is based on $\hat{\boldsymbol{\beta}}_{[PLE](2)}$, the second component of the PLE in the full model, i.e.

$$W = n(\hat{\boldsymbol{\beta}}_{[PLE](2)} - \boldsymbol{\beta}^0_{(2)})^T \widehat{V}^{-1}_{(22)} (\hat{\boldsymbol{\beta}}_{[PLE](2)} - \boldsymbol{\beta}^0_{(2)}), \tag{7.10}$$

where $\widehat{V}_{(22)}$ is the block (22) of the estimated asymptotic variance $\widehat{V} = \widehat{I}^{-1}(\hat{\boldsymbol{\beta}}_{[PLE]})$. For the score test, the general approach outlined in Section 2.5.1 can be extended to the Cox model. Under H_0, all three tests are asymptotically distributed as a χ^2_k distribution, where $k = \dim(\boldsymbol{\beta}_{(2)})$. They are generally provided by all common statistical packages.

7.2.2 Empirical Influence Function for the PLE

The *IF* for the PLE is based on complex functionals taking into account the semi-parametric nature of the Cox model and the presence of censoring. We give here its empirical version $\widehat{IF}_i$ evaluated at the observation $(t_i, \delta_i, \boldsymbol{x}_i)$ as originally derived by Reid and Crépeau (1985). Here $\widehat{IF}_i$ can be used as a diagnostic tool to assess the effect of a particular observation on the PLE. It is a q-dimensional vector proportional to a shifted score, i.e.

$$\widehat{IF}_i = \widehat{I}^{-1}(\boldsymbol{\beta})(U_i - C_i(\boldsymbol{\beta})), \tag{7.11}$$

where $C_i(\boldsymbol{\beta})$ (or to be more specific $C(t_i, \delta_i, \boldsymbol{x}_i; \boldsymbol{\beta})$) is a term depending on the observations in a complicate way; see Section 7.2.4. This 'shift' is not needed for consistency but to account for the dependence across the individual scores U_i. As noted by Reid and Crépeau (1985), the first component in $\widehat{IF}_i$ is similar to the usual *IF* for M-estimators in the uncensored case, and the second component $-\widehat{I}^{-1}(\boldsymbol{\beta})C_i(\boldsymbol{\beta})$ represents the influence of the ith observation on the risk set of other subjects. A similar expression with a two-part *IF* is generally found for estimators for censored data. The first term is unbounded in $\boldsymbol{x}_i$, which means that spurious observations in the covariates can ruin the PLE. The second term shows the same deficiency and, as a function of t_i's only, can be large enough to compromise the estimation process. It captures the influence of a particular observation (e.g. an abnormal long-term survivor) on the risk set of the others subjects. Valsecchi *et al.* (1996) give a good explanation on the acting mechanism. Abnormal long-term survivors *'exert influence in two ways. First, that individual forms part of the very many risks sets (for all preceding failures). Secondly, whereas early failures will be matched to a large risk set, individuals failing toward the end of the study may, depending on the censoring, be matched to a very small risk set. Two groups may be initially of similar size but as time progresses the relative size of the two groups may steadily change as individuals in the high risk group die at a faster rate than those in the other group. Eventually the risk set may be highly imbalanced with just one or two individuals from the high risk group, so that removal of one such individual will greatly affect the hazard ratio.'* Atypical long-term survivors are not the only type of abnormal response that can be encountered but they are by far the most dangerous. Another possibility occurs when a low-risk individual fails early. As pointed out by Sasieni (1993a) this type of outlier is less harmful as their early disappearance from the risk set reduces their contribution to the score equation. Despite its relative complexity the *IF* for the PLE has similar properties to that given for the M-estimators of Chapter 2. It still measures the worst asymptotic bias caused

to the estimator by some infinitesimal contamination in the neighborhood of the Cox model. It is therefore desirable to find estimators which bound that influence in some way. The two-part structure of (7.11) rules out a similar weighting to that used earlier in a fully parametric model. Innovative ways have to be imagined to control both components, in particular the influence on the risk set. This will be developed further in Section 7.2.4 in relation to the asymptotic variance.

7.2.3 Myeloma Data Example

Krall *et al.* (1975) discuss the survival of 65 multiple myeloma patients and its association with 16 potential predictors all listed in their Table 2. They originally selected the logarithms of blood urea nitrogen (`bun`), serum calcium at diagnosis (`calc`) and hemoglobin (`hgb`) as significant covariates. Chen and Wang (1991) used their diagnostic plot and found that case 40 is an influential observation. They also concluded that no log-transformation of the three predictors was necessary. We also use the data without transformation to illustrate the *IF* approach.

Table 7.1 presents the most influential data points as detected by the change $\Delta_i \hat{\boldsymbol{\beta}}$ in the regression coefficient $\hat{\boldsymbol{\beta}}_{[PLE]}$ when the ith observation is deleted. Figures are given as percentages to make changes comparable across coefficients: percentages were simply obtained by dividing the raw change by the absolute value of the corresponding estimate obtained on all data points. The deletion of any of the remaining observations did not change the coefficients by more than $\pm 11\%$ for these two variables, and even less for `bun`. These values can be seen as a handy approximation of the *IF* itself as $\widehat{IF}_i \approx (n-1)\Delta_i \hat{\boldsymbol{\beta}}$ as pointed out by Reid and Crépeau (1985). This result is generally true for all models but is particularly useful here when *IF* has a complicated expression. Clearly case 40 is influential confirming the analysis by Chen and Wang (1991). Other observations might also be suspicious, e.g cases 3 or 48. A careful look at all exact values of the *IF* (not shown here) shows that the approximation works reasonably well justifying the use of $\Delta_i \hat{\boldsymbol{\beta}}$ as proxy for $\widehat{IF}_i$. A word of caution must be added here. The empirical *IF* in (7.11) is typically computed at the PLE, itself potentially biased. This can cloud its ability to detect outliers as pointed out by Wang *et al.* (2006). However, extreme observations are generally correctly identified by this simple diagnostic technique. To illustrate how they can distort the estimation and testing procedures, we deleted case 40 and refitted the data. PLE estimates, standard errors and p-values for significance testing (z-statistic in (7.7)) are displayed in Table 7.2. Case 40 is actually a patient with high levels of serum calcium who survived much longer than similar patients. For that reason this subject tends on his own to determine the fit, an undesirable feature as the aim of the analysis is to identify associations that hold for the majority of the subjects. When all observations are included in the analysis, calcium is not significant ($p = 0.089$). After removal of case 40 a highly significant increase in risk of death of $\exp(0.31) - 1.0 = 0.36$, 95% CI (0.1;0.7), per additional unit of serum calcium appears. This clearly illustrates the dramatic effect of case 40 on the test. The differences are even more pronounced if both cases 40 and 48 are removed making the need of a robust analysis even greater. However, as the dataset is relatively

Table 7.1 Diagnostics $\Delta_i \hat{\boldsymbol{\beta}}$ for myeloma data.

Case	hgb	calc
40	+17%	−48%
48	0%	−16%
44	+13%	+12%
3	−1%	+24%
2	+2%	+35%

The regression coefficients are estimated by means of the PLE $\hat{\boldsymbol{\beta}}_{[PLE]}$.

Table 7.2 PLE estimates and standard errors for the myeloma data.

	All data		Case 40 removed	
Variable	Estimate (*SE*)	p-value	Estimate (*SE*)	p-value
bun	0.02 (0.005)	0.000	0.02 (0.005)	0.000
hgb	−0.14 (0.059)	0.019	−0.19 (0.063)	0.003
calc	0.17 (0.099)	0.089	0.31 (0.112)	0.006

Ties treated by Efron's method. Model-based SEs computed using (7.8).

small ($n = 65$), influential observations are more harmful and case-deletion and refit becomes a difficult exercise. We do not pretend to give a definitive analysis of these data here. The purpose was simply to illustrate the sensitivity of the PLE with respect to unexpected perturbations especially for small to moderate sample sizes.

7.2.4 A Sandwich Formula for the Asymptotic Variance

A different estimate for the asymptotic variance of the PLE has been proposed by Lin and Wei (1989). It is often called 'robust' variance in common statistical packages, but as we argue below, it is not robust in the sense used in this book. We therefore name it the LW formula or classical sandwich formula. Perhaps, the best way to introduce the LW formula is through its link to the *IF*, something that is generally overlooked. A careful reading of Reid and Crépeau (1985, p. 3) shows that $n^{-1} \sum \widehat{IF}_i \widehat{IF}_i^T$, where $\widehat{IF}_i$ is as in (7.11), provides another asymptotic variance estimate for the PLE. Elementary algebra shows that this can be rewritten as

$$\widehat{V}_{\mathrm{LW}}(\boldsymbol{\beta}) = \widehat{I}^{-1}(\boldsymbol{\beta}) \widehat{J}(\boldsymbol{\beta}) \widehat{I}^{-1}(\boldsymbol{\beta}) \tag{7.12}$$

where $\widehat{I}(\boldsymbol{\beta})$ is the information matrix estimator given in (7.6) and

$$\widehat{J}(\boldsymbol{\beta}) = \sum U_i^* U_i^{*T}, \tag{7.13}$$

Table 7.3 Estimates and standard errors for the PLE for the myeloma data.

	All data		Case 40 removed	
Variable	Estimate (SE_{LW})	p-value	Estimate (SE_{LW})	p-value
`bun`	0.02 (0.004)	0.000	0.02 (0.004)	0.000
`hgb`	−0.14 (0.059)	0.019	−0.19 (0.060)	0.002
`calc`	0.17 (0.127)	0.186	0.31 (0.103)	0.003

Ties treated by Efron's method. *SE* computed using (7.12).

where $U_i^* = U_i - C_i(\boldsymbol{\beta})$ is a shifted score. If we write down the correcting factor (shift)

$$C_i(\boldsymbol{\beta}) = \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})\boldsymbol{x}_i \sum_{j\le i} \frac{\delta_j}{S^{(0)}(t_j;\boldsymbol{\beta})} - \exp(\boldsymbol{x}_i^T\boldsymbol{\beta}) \sum_{j\le i} \frac{\delta_j S^{(1)}(t_j;\boldsymbol{\beta})}{[S^{(0)}(t_j;\boldsymbol{\beta})]^2}$$

and replace $\boldsymbol{\beta}$ by the PLE in (7.12) we obtain the variance estimate proposed by Lin and Wei (1989, p. 1074). Lin and Wei's derivation is actually more general as it also covers the case of time-dependent covariates. Although the formula presented here assumes n different time points, its extension to data with ties is straightforward (see Lin and Wei (1989) and Reid and Crépeau (1985) for technical details).

As an illustration, we refitted the myeloma data using the exact same model as before but use (7.12) as a variance estimate. PLE estimates, standard errors and p-values are displayed in Table 7.3. Note that the coefficients reported in Table 7.3 are the same as those reported in Table 7.2 since the estimation procedure is still the PLE. On the other hand, the standard errors differ as they are now based on the LW formula. The p-values reported here refer to the individual significance z-tests, i.e. for $H_0 : \beta_j = 0$,

$$z_{\mathrm{LW}}\text{-statistic} = \frac{\hat{\beta}_{[PLE]j}}{SE_{\mathrm{LW}}(\hat{\beta}_{[PLE]j})}, \tag{7.14}$$

where $SE_{\mathrm{LW}}(\hat{\beta}_{[PLE]j})$ is the standard error of $\hat{\beta}_{[PLE]j}$ based on the LW formula (7.12), i.e. the square root of $n^{-1}[\widehat{V}_{\mathrm{LW}}(\hat{\boldsymbol{\beta}}_{[PLE]})]_{jj}$.

Results are very similar to those obtained in Table 7.2. It is clear that case 40 is influential even if the LW formula is used. In other words, the LW formula offers no kind of protection against extreme (influential) observations. For example, no effect of calcium appears when all data are fitted (p-value = 0.186), and after removal of case 40 the deleterious effect of this observation on the significance of serum calcium seems obvious as a p-value of 0.003 is reported.

So we may legitimately ask the question 'What is the LW formula robust against?'. Lin and Wei (1989) motivate their approach by mentioning some structural misspecifications, in particular covariate omission. As an example they consider a randomized clinical trial in which the effectiveness of a particular treatment on survival time is assessed. The true model is thought to be the Cox model with

parameter $\boldsymbol{\beta}$. We can split $\boldsymbol{\beta}$ into two parts $\boldsymbol{\nu}$ and $\boldsymbol{\eta}$ where these components represent, respectively, the treatment parameters and the covariate effects. A valid test of no treatment effect is sought even if some of the predictors may be missing in the working model. Alternatively, investigators may simply prefer an unadjusted analysis for generalizabilty purposes, in which case only $\boldsymbol{\nu}$ will be included in the analysis. Lin and Wei (1989) showed that approximate valid inference can still be achieved using their formula. This, of course, assumes that no treatment by covariate interaction exists. To test the null hypothesis of no treatment effect (i.e. $H_0 : \boldsymbol{\nu} = 0$), one then uses (7.14). Lin and Wei (1989) also considered more serious departures from the Cox model, e.g. misspecification of the hazard form. This includes models with hazard defined on the log-scale or even a multiplicative model. Their simulation study shows that their approach allows approximate valid inference in the sense that the type I error (empirical level) of the Wald test using the LW formula (7.12) is close to the nominal level. The term 'robust' formula is hence used in that sense. This type of robustness however does not protect against biases induced by extreme (influential) observations. The reader is referred to Section 7.5 for further discussion on this topic in a more general setting.

7.3 Robust Estimation and Inference in the Cox Model

7.3.1 A Robust Alternative to the PLE

The robust alternative to the PLE we present here has emerged over the years from Bednarski's research. It is based on a doubly weighted PLE that astutely modifies the estimating equation (7.5) without fundamentally changing its structure. It also has the advantage of being easily computable with some code available and included in the R `Coxrobust` package. Following Bednarski (1993) and Minder and Bednarski (1996), we assume that a smooth weight function $w(t, \boldsymbol{x})$ is available. Denote by $w_{ij} = w(t_i, \boldsymbol{x}_j)$ and $w_i = w_{ii} = w(t_i, \boldsymbol{x}_i)$ the weights for all $1 \leq i \leq j \leq n$ and set all other weights to zero by construction. Define the two sums

$$S_w^{(0)}(t_i; \boldsymbol{\beta}) = \sum_{j \geq i} w_{ij} \exp(\boldsymbol{x}_j^T \boldsymbol{\beta}) \tag{7.15}$$

$$S_w^{(1)}(t_i; \boldsymbol{\beta}) = \sum_{j \geq i} w_{ij} \exp(\boldsymbol{x}_j^T \boldsymbol{\beta}) \boldsymbol{x}_j \tag{7.16}$$

in a similar way to their unweighted counterparts of Section 7.2. A natural extension of the PLE is the solution for $\boldsymbol{\beta}$ of

$$\sum_{i=1}^{n} w_i \delta_i \left[\boldsymbol{x}_i - \frac{S_w^{(1)}(t_i; \boldsymbol{\beta})}{S_w^{(0)}(t_i; \boldsymbol{\beta})} \right] = 0. \tag{7.17}$$

The weight function $w(t, \boldsymbol{x})$ enters at two points: (i) in the main sum with w_i downweighting the uncensored observations; (ii) in the inner sums $S_w^{(0)}$ and $S_w^{(1)}$ with

all the w_{ij} for $i \leq j \leq n$. Equation (7.17) clearly has a similar structure to (7.5). Moreover, when all of the weights are chosen equal to one, the solution of (7.17) is the PLE, so that (7.17) can be literally seen as an extension of equation (7.5). By analogy with the notation of Section 7.2 we also denote by the individual score $U_{w,i}$, i.e. the contribution of the ith observation to the sum in (7.17)

$$U_{w,i} = w_i \delta_i \left[\boldsymbol{x}_i - \frac{S_w^{(1)}(t_i; \boldsymbol{\beta})}{S_w^{(0)}(t_i; \boldsymbol{\beta})} \right], \tag{7.18}$$

and by U_w the total score or left-hand side of (7.17). A proper choice of $w(t, \boldsymbol{x})$ is pivotal to make the solution of (7.17) both consistent and robust. The weights we consider here truncate large values of $g(t) \exp(\boldsymbol{x}^T \boldsymbol{\beta})$ where $g(t)$ is an increasing function of time.[4] Indeed, Bednarski (1993) and Minder and Bednarski (1996), considering the exponential model, argued that the PLE often fails when $t_i \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})$ is too large. They hence proposed weight functions based on truncations of such quantities (i.e. with $g(t) = t$). Bednarski (1999), however, pointed out that a better choice for $g(t)$ is the baseline cumulative hazard $\Lambda_0(t)$ in (7.3). The rationale for this is that $\Lambda_0(t_i) \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})$, given the covariate vector $\boldsymbol{x}_i$, has an exponential one distribution if the Cox model holds and t_i is not censored. This gives rise to the following weights

$$w(t, \boldsymbol{x}) = \begin{cases} K - \min(K, \Lambda_0(t) \exp(\boldsymbol{x}^T \boldsymbol{\beta})) & \text{(linear)}, \\ \exp(-\Lambda_0(t) \exp(\boldsymbol{x}^T \boldsymbol{\beta})/K) & \text{(exponential)}, \\ \max(0, K - \Lambda_0(t) \exp(\boldsymbol{x}^T \boldsymbol{\beta}))^2/K^2 & \text{(quadratic)}, \end{cases}$$

where K is a known cut-off value that can be chosen on robustness and efficiency grounds. Such weights have been used successfully ever since and are now implemented in the R `Coxrobust` package. In practice, two additional difficulties occur: first, the truncation value K is generally difficult to specify *a priori*, especially for censored data;[5] second, the unknown cumulative baseline hazard $\Lambda_0(t)$ is needed to compute the weights. This hazard is not often estimated in the Cox model as it is not actually needed to obtain the PLE and related tests. To overcome the first problem, Bednarski and colleagues proposed the use of an adaptive procedure that adjusts K at each step. They deal with the second problem by jointly (and robustly) estimating $\Lambda_0(t)$ and $\boldsymbol{\beta}$; see Grzegorek (1993) and Bednarski (1999, 2007).

To compute a robust estimator defined through (7.17) with one of the proposed weighting schemes updated adaptively, one can use the following algorithm. Given a specific quantile value τ, e.g. $\tau = 90\%$, used to derive the truncation value adaptively, one proceeds through the following steps.

- Initialization: obtain an initial estimator $\hat{\boldsymbol{\beta}}^0$, e.g. the PLE, compute the cut-off K as the pre-specified quantile τ of the empirical distribution $t_i \exp(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}^0)$,

[4] Note that the notation above did not mention any dependence of $w(t, \boldsymbol{x})$ on the regression parameter $\boldsymbol{\beta}$ and the weights should be more seen as 'fixed'. However, Bednarski (1999) showed that, under stringent conditions, the dependence on $\boldsymbol{\beta}$ does not modify the asymptotic distribution of the resulting estimator.

[5] One could argue that a quantile of the exponential one distribution could be used, at least in the absence of censoring.

$i = 1, \ldots, n$, set-up the current estimate $\boldsymbol{b}$ at $\hat{\boldsymbol{\beta}}^0$ and initialize the set of weights.

- Take the current estimate $\boldsymbol{b}$, evaluate K as the same quantile τ of the empirical distribution $\widehat{\Lambda}_w(t_i)\exp(\boldsymbol{x}_i^T\boldsymbol{b})$ with

$$\widehat{\Lambda}_w(t) = \sum_{i \le t} \frac{w_i \delta_i}{\sum_{j \ge i} w_{ij} \exp(\boldsymbol{x}_j^T \boldsymbol{b})}. \tag{7.19}$$

- Update $\boldsymbol{b}$ by solving (7.17) and then recompute the set of weights.
- Repeat the previous two steps until convergence.

Technical details about the adaptive process and formula (7.19) are omitted for simplicity but can be found in Bednarski (2007). Note though that $\widehat{\Lambda}_w(t)$ is a robust adaptation of the Breslow estimator.[6] The final value obtained through this algorithm is the adaptive robust estimator (ARE) or $\hat{\boldsymbol{\beta}}_{[ARE]}$. It can generally be obtained within a few iterations even for relatively large datasets. An advantage of this adaptive weighting scheme based on the cumulative hazard estimate (7.19) is that the ARE is invariant with respect to time transformations. It can also better cope with censored data by the way the cut-off value is updated. The price to pay for this flexibility is purely computational. Bednarski (1999) shows that the ARE has the same asymptotic distribution as its 'fixed-weight' counterpart defined in (7.17) and performs similarly in terms of robustness. The issue of the choice of weight function or quantile τ is more a matter of efficiency and/or personal preference. This question is discussed in the next section. Finally, it should be stressed that other possible weights have been proposed by Sasieni (1993a,b). Although the spirit of his approach is essentially the same, the proposed weights cannot handle abnormal responses for patients with extreme values in the covariates such as elevated blood cell counts or laboratory readings. Such extreme but still plausible data points are harmful to classical procedures. In contrast the ARE is built to offer some kind of protection in that case. A more formal treatment of these alternative weighting schemes can be found in Bednarski and Nowak (2003) along with a comparison with the ARE.

7.3.2 Asymptotic Normality

Under regularity conditions on $w(t, \boldsymbol{x})$ given in Bednarski (1993, 1999), the ARE is consistent and has the following asymptotic distribution

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{[ARE]} - \boldsymbol{\beta}) \to \mathcal{N}(0, V_w(\boldsymbol{\beta})), \tag{7.20}$$

where the asymptotic variance is given by a sandwich formula[7]

$$V_w(\boldsymbol{\beta}) = M_w^{-1} Q_w M_w^{-T}. \tag{7.21}$$

[6] Formula (7.19) gives back the Breslow estimator of baseline cumulative hazard when all weights are equal to one and thus $\boldsymbol{b}$ is the PLE; see, for instance, Collett (2003b, p. 101).

[7] Again, to obtain the variance of $\hat{\boldsymbol{\beta}}_{[ARE]}$, one needs to divide (7.21) by n.

The matrices $M_w = M_w(\boldsymbol{\beta})$, $Q_w = Q_w(\boldsymbol{\beta})$ are complicated expectations that we omit for simplicity; see Bednarski (1993) for details. Tedious but straightforward calculations show that their empirical versions have a much simpler form, i.e.

$$\widehat{M}_w(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial U_{w,i}}{\partial \boldsymbol{\beta}} \tag{7.22}$$

and

$$\widehat{Q}_w(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} U^*_{w,i}U^{*T}_{w,i} \tag{7.23}$$

with $U_{w,i}$ given in (7.18) and $U^*_{w,i} = U_{w,i} - C_{w,i}(\boldsymbol{\beta})$ a shifted weighted score with shift given in (7.24). A final estimate for the asymptotic variance follows easily by replacing $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}_{[ARE]}$ in (7.21)–(7.23). The asymptotic distribution (7.21) is valid not only under the assumption that the weights are fixed, i.e. K and $g(t)$ are pre-specified independently of the regression parameters,[8] but also for the adaptive weighting scheme described above; see Bednarski (1993, 1999, 2007) for details. Technical developments essentially show that the asymptotic result is not altered when smooth adaptive weight functions with bounded support and a robust hazard estimate are used.

There is a clear link between the LW sandwich formula (7.12) and the asymptotic variance (7.21). The first thing to note is that for both the classical and robust estimators the sandwich formula stems from the same property, i.e. from $n^{-1}\sum \widehat{IF}_i\widehat{IF}_i^T$ that is another consistent variance estimator. We have seen this for the PLE in Section 7.2.2 and the same has been shown by Bednarski (1993, 1999) for the ARE. Second, tedious rewriting show that the empirical *IF* for the ARE is again proportional to a shifted score as follows:

$$\widehat{IF}_{w,i} = \widehat{M}_w^{-1}(\boldsymbol{\beta})U^*_{w,i}(\boldsymbol{\beta}) = \widehat{M}_w^{-1}(\boldsymbol{\beta})(U_{w,i}(\boldsymbol{\beta}) - C_{w,i}(\boldsymbol{\beta}))$$

with $U_{w,i}(\boldsymbol{\beta})$ given in (7.18) and where the shift has an 'ugly' but computable expression given by

$$C_{w,i}(\boldsymbol{\beta}) = \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})\boldsymbol{x}_i\sum_{j\le i}\frac{w_j\delta_j w_{ji}}{S_w^{(0)}(t_j;\boldsymbol{\beta})} - \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})\sum_{j\le i}\frac{w_j w_{ji}\delta_j S_w^{(1)}(t_j;\boldsymbol{\beta})}{[S_w^{(0)}(t_j;\boldsymbol{\beta})]^2}. \tag{7.24}$$

A careful look at all of the quantities involved in (7.18) and (7.24) and in the equations of Section 7.2.4 shows that, if all of the weights w_{ij} and w_i are equal to one, then not only is the ARE identical to the PLE but their *IF* are the same and consequently its asymptotic variance reduces to the LW formula (7.12). The robust approach proposed in this chapter can literally be seen as an extension of the PLE combined with its LW sandwich variance. In practice, this never happens as the weights cannot be set to one if one wants the ARE to be robust. This analogy is nevertheless useful as it helps in understanding both formulas and properties.

[8]The function $g(t)$ is discussed on page 201.

The use of AREs is desirable from a robustness standpoint. However, an expected loss of efficiency with respect to the PLE is observed when the Cox model assumptions hold. Unlike in simpler models, e.g. linear regression or mixed models, it is impossible to calibrate the tuning constant (i.e. the quantile τ) to achieve a specific efficiency at the model for all designs. However, some hints can be given. First, it is clear that by construction linear adaptive weights of Section 7.3 automatically set a certain percentage of weights w_i to zero. If we choose $\tau = 90\%$, then roughly 10% of the weights will be zero even though the data arise from a proportional hazard model. This automatically generates a loss of efficiency at the model as genuine observations will be ignored. A similar argument holds for quadratic weights. In contrast, the exponential weighting scheme of Section 7.3 is smoother and the ARE with exponential weights performs generally better in terms of efficiency. Second, simulations can provide valuable information. Our (limited) experience with the exponential distribution indicates that adaptive exponential weights do reasonably well in terms of robustness and efficiency when τ is chosen in the range 80–90%. An asymptotic efficiency relative to the PLE of at least 90% can easily be obtained even in the presence of a small amount of censoring, while linear weights achieve an efficiency of at most 80–90%. However, both weighting schemes perform equally well from a robustness point of view. For these reasons we tend to prefer exponential weights, with τ in the range 80–90%. Previous references by Bednarski and colleagues also used similar values of τ successfully. We would not recommend the use of much smaller quantiles. Finally a choice for τ for a given weighting scheme could in principle be computed by simulations to achieve a predetermined loss of efficiency at a parametric model (i.e. a parametric form for the hazard). For that purpose, one would need an idea of the censoring level, a rough idea of the true parameter values, and a distribution for the covariates or conditioning. In situations where data-driven experimentations are suspicious, experience with previous similar data can also help in selecting τ without having to use the current dataset.

7.3.3 Handling of Ties

We have assumed so far that all observed time points were different. Unfortunately this is rarely the case as even intrinsically continuous data are rounded to their nearest time unit (days, months) in practice. This situation occurs for the datasets analyzed in this chapter. As mentioned in Section 7.2 the partial likelihood approach is generally modified to handle this situation. Four methods are available, namely Efron's or Breslow's, discrete or exact approach. We again refer to the literature, e.g. Collett (2003b, pp. 67–69) or Kalbfleisch and Prentice (1980, Chapter 4) for details. For the robust ARE, we simply suggest either: (i) the use of the 'jittering' method of Tai *et al.* (2002) which randomly adds or subtracts a small value to each tied event time to randomly break the tie; or (ii) do 'as if there are no ties'. This basically means that the way the data are sorted by increasing observed survival time is the way ties are broken, which is not necessarily at random. Approach (i) is, for instance, used by Sasieni (1993a) as a practical solution for continuous data where

Table 7.4 Estimates and standard errors for the PLE and the ARE for the myeloma data.

	PLE		ARE	
Variable	Estimate (*SE*)	p-value	Estimate (*SE*)	p-value
`bun`	0.02 (0.005)	0.000	0.02 (0.005)	0.000
`hgb`	−0.14 (0.059)	0.019	−0.17 (0.073)	0.020
`calc`	0.17 (0.099)	0.089	0.28 (0.122)	0.023

PLE: ties treated with Efron's method; ARE: exponential weights, $\tau = 0.80$.

time rounding created equal values. Its main disadvantage is that the analyses are not fully reproducible. Note, however, that the case of fundamentally discrete data cannot be dealt with, using the methods proposed here and further work for a proper treatment of ties is certainly desirable.

7.3.4 Myeloma Data Example (continued)

To illustrate how the ARE works in practice, we go back to the myeloma data example where diagnostic techniques such as the empirical *IF* of Section 7.2.2 identified at least case 40 as influential. As the sample size is small ($n = 65$) we choose exponential weights for the ARE to avoid too large a loss of efficiency. Such weights are smooth and do not trim the data excessively. The quantile value τ was set at 80% although $\tau = 90\%$ was also tried and gave comparable results. Table 7.4 presents the PLE and ARE estimates, standard errors and the corresponding p-values for individual z-tests of significance; see Section 7.3.5 for details. The ARE fit confirms the results obtained in the case-deletion and refit approach identifying serum calcium (`calc`) as an important predictor in the model. The deleterious effect of case 40 has been automatically reduced by the method without having to delete it or equivalently put a weight of zero. The advantage of this analysis is that the weighting has been accounted for in the asymptotic distribution of the (robust) estimator.

We can gain some insight into how the ARE works by inspecting the weights w_i.[9] Unlike weights provided by M-estimators they decrease with the survival time by construction so that a log-transformation is more suitable for exponential weights. Indeed, as noted earlier $-K \log(w_i) = \Lambda_0(t_i) \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})$ has an exponential distribution Exp(1) if the model holds and t_i is not censored. Large values on this scale are unlikely and therefore represent downweighted observations. Figure 7.1 depicts the negative logarithm of the weight (up to a multiplicative constant K) versus the observation number in the initial dataset. One may notice that roughly 20% of the observations are above the horizontal line with intercept K. This is expected as we chose $\tau = 0.80$, therefore $1 - \tau = 20\%$ of the observations must

[9] For an analysis of the weights w_{ij}, see Section 7.4.2.

be downweighted by construction. Only two observations emerge on this plot as abnormal cases: numbers 40 and 48 with the latter being more extreme with a value for $-K\log(w_i)$ close to eight. It is not really surprising to detect patient 40 as outlier as it was clearly identified by diagnostic techniques. The presence of the other observation (number 48) deserves further explanation. A look at case 48 shows that his risk factors are similar to those of case 40 with a relatively high level of serum calcium (11) and he has a high survival time (92 months). This is actually the longest survivor among all participants with a level of serum calcium that was at least comparable. This is considered as unlikely by the ARE, given that most subjects with similar profile die much earlier, even patient 40. This observation therefore receives a nearly zero weight (0.0008) reducing its influence drastically. It is interesting to see that diagnostic techniques of Section 7.2.2 were less adamant at classifying this subject as so influential. This could be due to some masking effect due to case 40 on the PLE of $\boldsymbol{\beta}$ used to compute the diagnostic tools and/or the fact that the exponential weight penalizes such abnormal responses more directly. Conversely, the same argument explains probably why cases 2 and 3 do not really appear on the plot. Those two patients have very short survival times, 1.25 and 2.0 months, respectively. They are not considered as really harmful with this type of weight although they might be considered as (too) early failures. Globally, this example illustrates the importance of a robust fit as both a safer technique and a new diagnostic tool. To conclude with the myeloma data example, we do not pretend that this analysis is final. One might obtain different results and, hence, conclusions if predictors, in particular serum calcium, are log-transformed. Other factors could also have been brought in to explain the discrepancy caused by these two long-term survivors. We do not, however, pursue this issue further.

7.3.5 Robust Inference and its Current Limitations

A robust Wald test for testing hypotheses of the type $H_0 : \boldsymbol{\beta}_{(2)} = \boldsymbol{\beta}^0_{(2)}$ as in Section 7.2 is directly available from the asymptotic distribution of $\hat{\boldsymbol{\beta}}_{[ARE]}$ of Section 7.3.2. The test statistic is given by

$$W_w = n(\hat{\boldsymbol{\beta}}_{[ARE(2)]} - \boldsymbol{\beta}^0_{(2)})^T \widehat{V}^{-1}_{w(22)}(\hat{\boldsymbol{\beta}}_{[ARE(2)]} - \boldsymbol{\beta}^0_{(2)}), \tag{7.25}$$

where $\widehat{V}_{w(22)}$ is the block (22) of the asymptotic variance $\widehat{V}_w(\boldsymbol{\beta})$. It is the natural counterpart of the classical Wald test (7.10). As always $\hat{\boldsymbol{\beta}}_{[ARE]}$ is used to replace $\boldsymbol{\beta}$ in the asymptotic variance $\widehat{V}_w(\boldsymbol{\beta}) = \widehat{M}_w^{-1}(\boldsymbol{\beta})\widehat{Q}_w(\boldsymbol{\beta})\widehat{M}_w^{-T}(\boldsymbol{\beta})$ obtained through (7.22)–(7.23) to obtain numerical values. In the special case of a single hypothesis, the Wald test reduces to a z-test as used above on the myeloma data. Under H_0, W_w is asymptotically distributed according to a χ^2_k distribution with $k = \dim(\boldsymbol{\beta}^0_{(2)})$. It is not clear whether a score-type or LRT-type test (see Section 2.5.3) can be used because of the dependence of the scores functions $U_{w,i}$ and the adaptive weighting process. This issue has not been investigated so far.

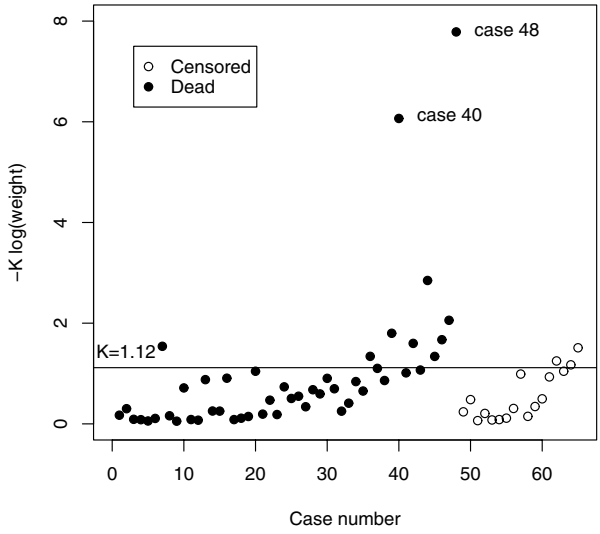

Figure 7.1 Plot of robust ARE exponential weight (log-transformed) versus case number for the myeloma data.

The robustness properties of the z- (or Wald) test based on the PLE (with or without the LW formula) and the ARE have been examined by Heritier and Galbraith (2008), yielding contrasting results. As we work here with Wald-type tests there is equivalence between CIs and tests. The notions of robustness of validity and efficiency for tests introduced in Section 2.5.3 can immediately be translated into similar concepts for CIs. In essence, a CI must maintain its nominal coverage (e.g. 95%) and its length in a neighborhood of the assumed model for the procedure to be declared robust. Heritier and Galbraith (2008) therefore studied the coverage probability of CIs based on the (estimated) variances of $\hat{\boldsymbol{\beta}}_{[ARE]}$ and $\hat{\boldsymbol{\beta}}_{[PLE]}$. Two types of model deviations are investigated, namely a shrinking neighborhood of the type $(1-\varepsilon_n)F_{\boldsymbol{\beta}} + \varepsilon_n G$ where $F_{\boldsymbol{\beta}}$ is the assumed (Cox) model, G a contaminating distribution and $\varepsilon_n = \varepsilon/\sqrt{n}$ and a full neighborhood (1.1), i.e. when ε does not depend on n. The first type of neighborhood is the one assumed in robust testing theory to avoid overlapping between the neighborhood of the null and that of a sequence of alternative contiguous hypotheses (see Section 2.4.2). In practice the proportion of (potential) extreme observations does not necessarily decrease with the sample size and it is therefore important to check whether the Wald-type test given by (7.25) is robust in a full neighborhood. Heritier and Galbraith (2008) generate data from an exponential model with hazard $\lambda(t \mid \boldsymbol{x}) = 3\exp(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)$ with independent standard normal covariates and $(\beta_1, \beta_2, \beta_3) = (0, 0.5, 1)$. A percentage ε_n (or ε) of observations is then replaced by data coming from an exponential model with hazard $\lambda(t) = 3$ (i.e. not depending on the covariates' values). Such contamination is assumed to mimic the effect of abnormal long-term survivors.

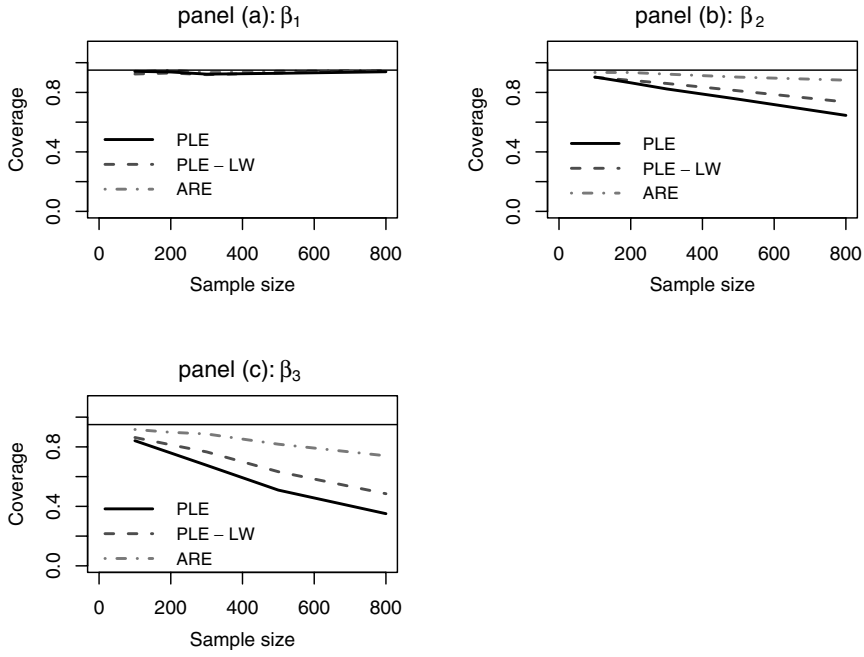

Figure 7.2 Coverage of 95% CIs under a full neighborhood contamination. The CIs are based on the (estimated) variances of $\hat{\boldsymbol{\beta}}_{[ARE]}$ (ARE), $\hat{\boldsymbol{\beta}}_{[PLE]}$ (classical, PLE) and $\hat{\boldsymbol{\beta}}_{[PLE]}$ with LW formula (PLE–LW)

In the shrinking neighborhood setting, the CIs based on the ARE with exponential weights and $\tau = 90\%$ perform well (coverage probability close to the target of 95%). So does the level of corresponding Wald-type test because of the equivalence between CIs and tests. In this simulation the percentage of contamination ε_n is set to 5% for $n = 100$ observations. As n increased progressively to 300, 500 and 1000, ε_n drops to 2.9%, 2.2% and 1.6% respectively. Alternatively, this corresponds to 16, 11 and 9 spurious cases generated from the contaminating distribution Exp(3). In contrast, the CIs based on the PLE (with or without the LW formula) do not show the same stability and display a coverage that slightly deteriorates as n increases. CIs using model-based standard errors are uniformly the worst performers with a coverage that drops below 80% for $n = 1000$ while, when the LW formula is used, a slightly better coverage of 87% is observed.

If instead, the simulations are repeated in a full neighborhood of the model, i.e. $\varepsilon = 5\%$ irrespective of the sample size, less favorable results are observed. Figure 7.2 clearly shows that none of the methods considered here can generally maintain appropriate coverage unless the true parameter is indeed zero as shown in panel (a). The confidence intervals based on the ARE always outperform their classical counterparts but this is little consolation as coverage values as

low as 80% can be observed when $n \geq 800$ and $\beta = 1$. For moderate sample sizes or small effect size, a current situation in clinical research, the CIs based on the ARE might still be reliable and one may see this as a relief. From a more theoretical perspective, it is quite clear that the procedure breaks down and does not meet the requirements of the robustness paradigm. This unusual feature[10] can be explained relatively simply. Since the scores given in (7.18) are not independent, the robustification process is much harder than in the i.i.d. setting. Looking back at Bednarski's theory, we clearly see that the ARE has been proved valid from a robustness point of view in only a shrinking neighborhood (see Bednarski, 1993; Minder and Bednarski, 1996). Simulations show that, in a full neighborhood, a small residual bias usually toward the null remains (even asymptotically). This is no big deal as far as estimation is concerned but it is when inference is the primary target. Indeed, this attenuation effect causes the CI to be shifted slightly. As the sample size increases, so does its accuracy, which in turns leads to a CI coverage that appears comparatively further off the mark. This dramatic loss of coverage corresponds to an equally important increase in type I error for the Wald tests investigated here. This emphasizes the need for further work in this area as the robust theory developed so far is clearly unsatisfactory.

7.4 The Veteran's Administration Lung Cancer Data

7.4.1 Robust Estimation

We describe in this section a complete analysis based on a benchmark in survival analysis: the veteran's administration lung cancer data. This dataset originated from a controlled clinical trial where investigators were interested in assessing the effect of treatment (standard/test) and several predictors on survival in males with advanced inoperable lung cancer. Covariates include both continuous predictors, i.e. Karnofsky (`karnofsky`) performance status (a scale index allowing patients to be classified as to their functional impairment), disease duration (`dduration`), age (`age`) and indicators such as prior therapy (`ptherapy`) or cell type (`cell`) with four levels, *Adeno*, *Squamous*, *Small*, *Large*, which is taken as the reference. Details can be found in Kalbfleisch and Prentice (1980, p. 60) who were the first to discuss that example: 137 male patients were randomized to either the placebo or the experimental arm and only 9 of them survived to the end of the study. This dataset is interesting as it contains abnormal long-term survivors and the Cox model is generally considered as a good working model. Outliers identified by many authors including Cain and Lange (1984), Sasieni (1993a,b), Minder and Bednarski (1996) and Bednarski (1999) have been dealt with in various ways. Minder and Bednarski (1996) analyzed the veteran's administration lung cancer data using an early version of their estimator (i.e. not using the adaptive weights of the ARE) while Bednarski

[10]Note that this undesirable feature is not completely new: it has been observed for specific location and regression estimators by Adrover *et al.* (2004).

Table 7.5 Estimates and standard errors for the Veterans' Administration lung cancer data.

	PLE		ARE	
Variable	Estimate (*SE*)	*p*-value	Estimate (*SE*)	*p*-value
`karnofsky`	−0.033 (0.006)	0.000	−0.039 (0.005)	0.000
`dduration`	−0.000 (0.008)	0.995	−0.004 (0.009)	0.708
`age`	−0.008 (0.009)	0.388	−0.009 (0.011)	0.392
`ptherapy`	0.065 (0.232)	0.779	0.109 (0.236)	0.643
`cell`				
Squamous	−0.397 (0.283)	0.160	−0.247 (0.296)	0.402
Small	0.483 (0.265)	0.068	0.806 (0.278)	0.004
Adeno	0.800 (0.303)	0.008	0.943 (0.253)	0.000
`treatment`	0.286 (0.217)	0.166	0.214 (0.206)	0.298

PLE: ties treated with Efron's method; ARE: exponential weights, $\tau = 0.90$.

(1999) used the ARE with linear weights. Irrespective of the robust method used the results are very consistent. Table 7.5 presents the ARE estimates with exponential weights, their standard errors and the corresponding p-values of individual tests for significance. The ties are ignored for simplicity and reproducibility. For the sake of completion we also tried to break the ties at random and obtained pretty similar results. Both estimation procedures identify `karnofsky` as a significant predictor while neither `dduration`, `treatment`, `age` nor `ptherapy` appear to be. Differences appear, however, on the importance of the type of cell. The PLE does not clearly identify small cell patients as having a worse prognosis than those with large cell carcinomas. The PLE for small cell is 0.48 (0.27) leading to a hazard ratio of $\exp(0.48) = 1.62$ with a 95% CI of $(-0.95;2.74)$, a somehow inconclusive result ($p = 0.068$). In contrast, the ARE sorts out the issue with a nearly twofold robust estimate 0.81 (0.28) or equivalently a hazard ratio of 2.2 (0.3;2.9), a clearly significant result ($p = 0.004$). The effect of adeno carcinomas is also larger leading to a neater effect of `cell` overall. Had a smaller quantile been used e.g. $\tau = 0.80$ an even stronger effect of small cells would have been observed. Similar conclusions were reached in the above references where linear weights were used.

7.4.2 Interpretation of the Weights

To identify potential outliers we again use a plot of the weights on (minus) a logarithm scale, with the plot presented in Figure 7.3. Clearly, two cases (numbers 17 and 44) are identified as influential observations as they receive a raw weight of 0.04 and 0.002, respectively. These are by far the smallest values even on an untransformed scale. A careful look at the profiles of these two patients shows that both are among the 7% highest survivors. This is considered as abnormal by the ARE given their risk factors and other subjects with similar characteristics. They are

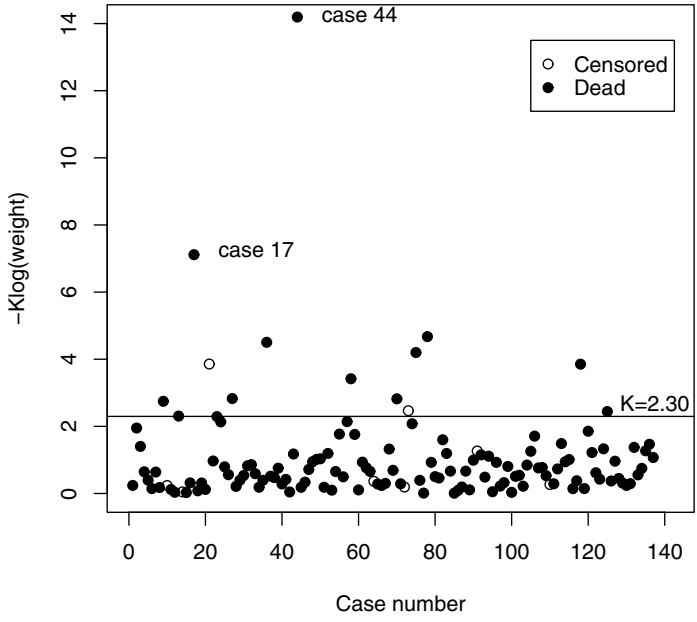

Figure 7.3 Plot of robust ARE exponential weight (log-transformed) versus case number for the lung cancer data.

therefore automatically downweighted (especially case 44 who receives the smallest weight). To see this on the plot, note that if the Cox model holds, the exponential one model is a reasonable model for the distribution of the transformed weight $-K \log(w_i) = \Lambda_0(t_i) \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})$ (as the level of censoring is low in these data). Then the probability of observing a value of 14.2 (i.e. case 44's transformed weight) or higher is $\exp(-14.2) = 7 \cdot 10^{-7}$. The method clearly identifies case 44's response as totally inconsistent with the rest of the data, given his risk factors. A similar calculation for case 17 returns a probability of 0.0008, which is still very small. This analysis agrees well with the work of Sasieni (1993a) and Cain and Lange (1984) who also identified influential cases, especially case 44.

It is more difficult to make sense of the triangular matrix of weights arising in the two internal sums (7.15) and (7.16). Roughly speaking these weights try to 'repair' the risk set and related sums contaminated by outliers. As an attempt to illustrate this point, we examine the matrix weights attributed to case 44 for all patients with a shorter survival time (≤ 392 days). Case 44 is actually recorded as the 131st patient when patients are ordered by increasing survival time. It therefore contributes to the risk set of more than 90% of the sample. Figure 7.4 displays the matrix weights on (minus) the logarithm scale versus survival time for $i = 131$ (case 44, solid line) and the six patients who outlived him ($i = 132$–137, dashed lines). As before, points in the upper part of the graph correspond to small weights. As a rough indicator we have

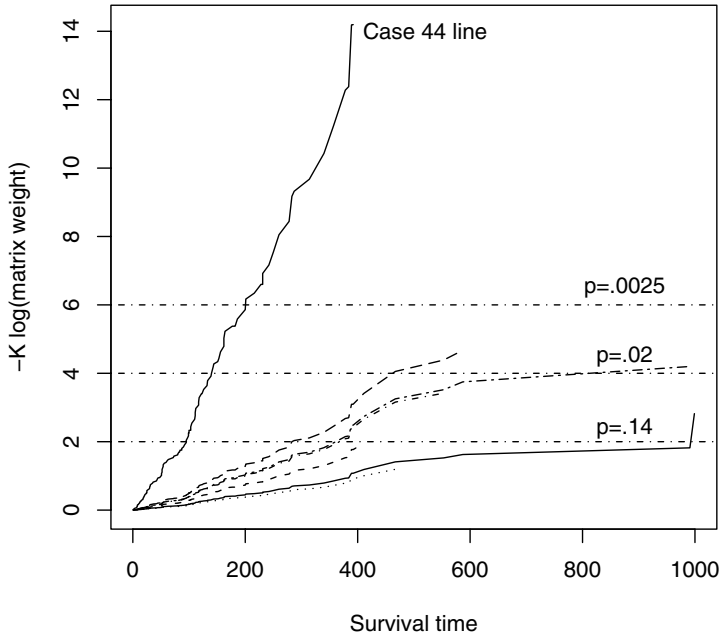

Figure 7.4 Plot of the ARE (transformed) matrix weights versus survival time for the last seven survivors.

added horizontal lines for several quantiles of the exponential one distribution with corresponding probabilities for this distribution to be above that line. Most of the contributions of case 44 to the risk set of other patients are considered as abnormal and should therefore be small compared with the others. The way ARE achieves this is by giving them a small exponential weight, the weight becoming smaller as time increases. This is certainly a desirable property as ideally subject 44 should not be part of the risk set of many other observations beyond a specific time point. This can be seen on the plot as the line for case 44 has a steep slope and goes further up leading to tinier weights. In contrast the dashed lines for the last 6 survivors stay reasonably low as these subjects are considered as 'normal' by the robust approach. The fact that all of the lines are roughly straight is due to the form of the cumulative hazard that looks linear for these data. They also have different lengths, e.g the line for case 44 does not go beyond time 392 days, to reflect that beyond their own survival time the patient is no longer in the risk set of other subjects.

7.4.3 Validation

The literature on how to validate a robust fit in the Cox model is sketchy with the exception of Valsecchi *et al.* (1996) and Minder and Bednarski (1996). This is also due to the relatively new development of the theory. It is still not clear how to

extend the many residuals and related plots developed for a classical fit (PLE); see Therneau and Grambsch (2000, Chapter 4), for a review. In an attempt to validate their approach, Minder and Bednarski (1996) proposed to compare the Kaplan–Meier survival curve with their counterparts obtained when the Cox model is fitted with the PLE and the ARE. For increasing survival times t_i, the Kaplan–Meier estimate[11] is simply

$$\widehat{S}_{[\mathrm{KM}]}(t) = \prod_{t_i<t}\left(\frac{n_i - d_i}{n_i}\right), \tag{7.26}$$

where n_i is the number of subjects still 'at risk' just prior to t_i and d_i is the number of deaths at time t_i; see Kalbfleisch and Prentice (1980, p. 12) or Collett (2003b, p. 20). The model-based survival curves are obtained as follows. First, note that by combining (7.2) and (7.3) we obtain the usual (but rarely used) expression of $S(t \mid \boldsymbol{x})$ as a function of $\boldsymbol{\beta}$ and the cumulative baseline hazard $\Lambda_0(t)$, i.e.

$$S(t \mid \boldsymbol{x}) = \exp(-\Lambda_0(t)\exp(\boldsymbol{x}^T\boldsymbol{\beta})). \tag{7.27}$$

Second, an overall survival curve estimate can be simply computed by averaging over the sample the predictions of individual survival time $S(t \mid \boldsymbol{x}_i)$ for $t = t_j$, $j = 1, \ldots, n$. For the ARE, the ith patient's survival prediction is obtained by replacing in formula (7.27) the true cumulative baseline hazard by its estimate (7.19), and the linear predictor by $\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{[ARE]}$. The same can be done for the PLE by using the corresponding classical estimates. The comparison between $\hat{S}_{[\mathrm{KM}]}(t)$ and its Cox-based counterparts proceeds by plotting their 'standardized' differences versus the logarithm of the survival time, possibly by categories (e.g. quartiles) of the linear predictor $\boldsymbol{x}^T\hat{\boldsymbol{\beta}}_{[PLE]}$. For the standardization factor, we follow Minder and Bednarski (1996) and use the square-root of $\hat{S}_{[\mathrm{KM}]}(t)(1 - \hat{S}_{[\mathrm{KM}]}(t))$. Figure 7.5 displays the standardized difference per tertile of linear predictor $\boldsymbol{x}^T\hat{\boldsymbol{\beta}}_{[PLE]}$. The horizontal lines represent plus or minus twice the standard error of the Kaplan–Meier estimate obtained through the Greenwood formula (see Collett, 2003b, pp. 24–25) to take into account the sample variability, at least approximately. A good agreement between the Kaplan–Meier and ARE survival curves can be observed for all panels. In contrast some discrepancy appears when the PLE is used to fit the Cox model, in particular in panels (a) and (c). This lack of fit disappears after deletion of the extreme observations identified earlier and repeat of the procedure (Figures not shown). This is a compelling argument in favor of the robust fit assuming that the model is structurally correct. Other plots can also be found in Minder and Bednarski (1996) and Bednarski (1999). Note as well that separate plots for each treatment arm could also be drawn, but this is not done here as the experimental treatment was found to be ineffective.

[11] In the presence of ties, formula (7.26) still applies by replacing the t_i, $i = 1, \ldots, n$, by the $k < n$ distinct ordered survival times $t_1 < t_2 < \ldots < t_k$.

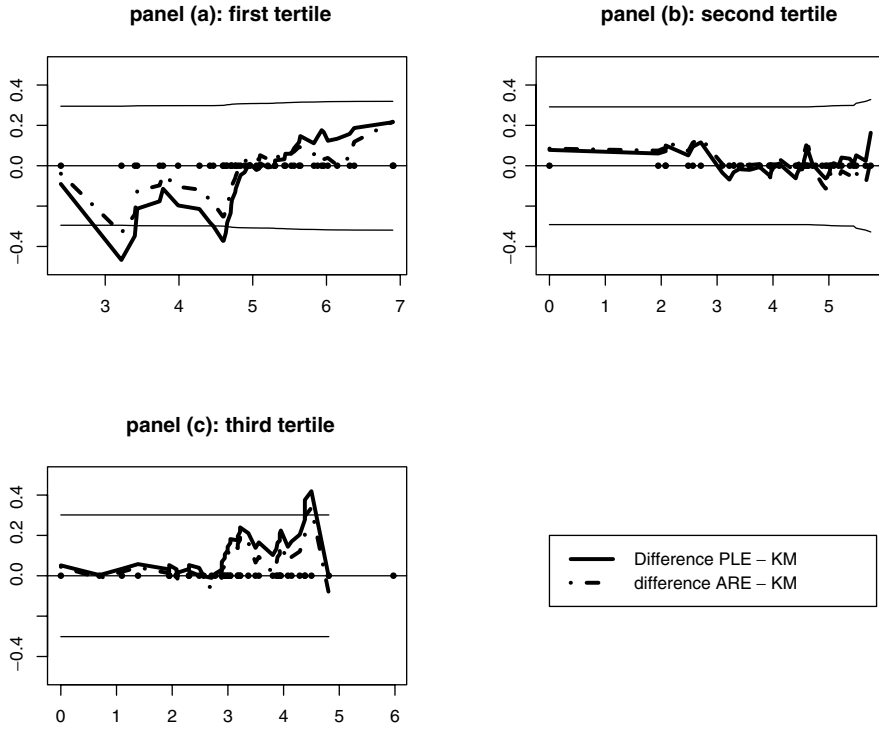

Figure 7.5 Standardized differences between the Kaplan–Meier estimate (KM) and the model-based survival curves (PLE or ARE).

7.5 Structural Misspecifications

7.5.1 Performance of the ARE

The main objective of this book is to present robust techniques dealing with distributional robustness. In essence, we assume a specific model (e.g. the Cox model) and propose estimation or testing procedures that are meant to be stable and more efficient than the classical procedures in a 'neighborhood' of the working model. We normally speak of model misspecification in that sense. This sometimes creates confusion, in particular for the proportional hazard model where the effect of many departures have been studied over the years, e.g. covariate omission, deviations from the proportional hazard assumption or measurement error in the variables. These can be seen as structural model misspecifications and are not the scope of the robustness theory. It is however important to discuss the performance of the robust procedures (estimation, tests) presented so far in that setting.

Historically, researchers first studied the impact of covariate omission on the estimation process, in particular in randomized experiments where the main endpoint

is a specific time to event. Typically, the question of whether an unadjusted analysis of a two-arm randomized clinical trial (RCT) still provides a consistent estimate of the treatment effect was of primary interest. Work by Gail *et al.* (1984), Bretagnolle and Huber-Carol (1988) and others showed both theoretically and by simulations that if important predictors were omitted from the Cox model the classical estimate (PLE) was slightly biased toward the null. They even showed that this situation could happen when the data were perfectly balanced as in RCTs and was worsened by the presence of censoring. The key reason for this is that the PLE does not generally converge toward the true regression parameter unless the treatment effect is itself zero. This type of problem being structural, a similar situation arises with the ARE. No formulas have ever been established but a hint was given by Minder and Bednarski (1996) who explicitly investigated the problem.[12] They show in their simulation (type B) that the robust proposal is indeed biased toward the null but tend to be less biased than the PLE. A similar situation is encountered in Section 7.3.1 with measurement error problems. If a predictor x_1 cannot be measured exactly but instead $v_1 = x_1 + u$ is used where u is some random term added independently, it is well known that β_1, the slope of x_1, is not estimated consistently; see Carroll *et al.* (1995) for instance. An attenuation effect or dilution is observed if the naive approach is used (i.e. regressing the outcome on v_1 and the other covariates) for most types of regression including Cox's. In addition estimates of other slopes can also be affected. Again the ARE is not specifically built to remove such bias resulting more from a key feature of the data, i.e. a structural model misspecification, ignored in a naive analysis (classical or robust). In another simulation (type C), Minder and Bednarski (1996) showed that the ARE tended to be less biased than its classical counterpart. In such a case it is highly recommended to directly correct for measurement error using one of the many techniques described, for instance, by Carroll *et al.* (1995) and in the abundant literature dealing with this issue. Robust methods could then also be specifically developed in that setting, i.e. with a model that includes possible measurement error.

The problem of 'what to do when the hazard is not proportional' often arises and is even more important in practice. Here two elements of the answer can be brought in. First, if non-proportionality is caused by a subgroup of patients responding differently then ARE will certainly provide safer results. Second, if the problem is more structural, e.g. a multiplicative model captures more the inherent nature of the data, the ARE will not perform any better than the classical technique. The reason is that (i) both methods still assume proportional hazards; (ii) this type of departure is not in a neighborhood of the working model. In other words it will not be 'within the range' of what the robust method can handle. By definition, the problem is more structural than distributional and beyond the scope of the current method.

Finally, one may wonder how large amounts of censoring affects the ARE or if something similar is possible for the time-dependent Cox model. The robust approach presented here is only valid under the assumption of fixed predictors.

[12] The estimator used in this reference is a simpler version of the ARE: the weighting scheme is based on $g(t) = t$, not the cumulative hazard function; see Section 7.3.1 for the definitions of the weights. However, the results are illustrative of what could be obtained with the ARE.

Its extension to time-varying covariates has not been attempted, even under simple circumstances, and seem a considerable challenge. Regarding the impact of censoring, no work has been carried out to illustrate the performance of the ARE in the presence of heavy censoring.

7.5.2 Performance of the robust Wald test

It is probably legitimate to wonder whether the robust Wald test defined in Section 7.3.5 provides some kind of protection against structural misspecifications. This question arises naturally as we know that the asymptotic variance (7.21) is literally a generalization of (7.12), the LW formula is supposed to be better at dealing with that type of problem; see the discussion in Section 7.2.4 and the link between the two formulas in Section 7.3.2. An insight is given in Heritier and Galbraith (2008) who carried out simulations similar to those undertaken by Lin and Wei (1989) with the addition of the ARE as genuine contender. We report here the results for covariate omission, a particularly relevant situation in RCTs as discussed earlier. The data were not, however, generated to mimic that situation as in Lin and Wei (1989). Survival times come from an exponential model with hazard $\lambda(t \mid \boldsymbol{x}) = \exp(x_1^2)$ where x_1 follows a standard normal distribution. This is supposed to be an even worse scenario than simply ignoring the predictors in a RCT. The working model is a Cox model with two predictors x_1 and x_2, generated independently of each other with the same distribution. This model is misspecified as x_1^2 has been omitted from the fitted model and x_2 is unnecessary. The primary objective is the performance of tests of $H_0 : \beta_1 = 0$ at the true model. The standard z-test (with model-based *SE*) cannot maintain its nominal level of 5% and instead exhibits an inflated type I error around 13%. In contrast the LW z-test has a type I error around 6–6.5% while ARE's is around 3.5–4.5%. These results stand for a sample size of 50–100 and are consistent with those initially reported by Lin and Wei (1989). The ARE-based Wald test thus seems to perform well in that particular setting; if anything the test seems to be conservative. A similar performance to the LW approach is also observed by Heritier and Galbraith (2008) for the other designs studied by Lin and Wei (1989), including misspecified hazards, e.g. fitting the Cox model to data generated with a logarithmic type of hazard. These conclusions are seriously limited by the fact that we are only focusing on the test level. Nothing is said about the loss of power of such procedures compared with those of inferential (robust) procedures developed in a structurally correct model. We therefore strongly recommend sorting out structural problems before carrying out robust inference. Distributional robustness deals with small deviations from the assumed (core) model, and this statement is even more critical for inferential matters. This is clearly not the case if, for instance, the right scale for the data is multiplicative as opposed to additive (i.e. one of the scenarios considered here). Using testing procedures in a Cox model fitted with the ARE should not be done if linearity on the log-hazard scale is clearly violated. The same kind of conclusion holds for violations from the proportional hazard assumption. This recommendation could only be waived if such departures are caused by a few abnormal cases, in which case the use of a robust Wald test can be beneficial. Finally, the LW approach is also

used when correlation (possibly due to clustering) is present in the data. It generally outperforms its model-based counterpart and maintains its level close to the nominal level. The properties of (7.21) in that setting have not been investigated.

7.5.3 Other Issues

Robust methods in survival data have just started their development. As mentioned earlier the presence of censoring creates a considerable challenge. In the uncensored case robust methods in fully parametric models are readily available. One could, for instance, use robust Gamma regression as described in Chapter 5. Specific methods have also been proposed for the (log)-Weibull or (log)-Gamma distributions by Marazzi (2002), Marazzi and Barbati (2003), Marazzi and Yohai (2004) and Bianco *et al.* (2005). Interesting applications to the modeling of length of stay in hospital or its cost are given as an illustration. The inclusion of covariates is considered in the last two references. Marazzi and Yohai (2004) can also deal with right truncation but, unfortunately, these methods are not yet general enough to accommodate random censoring. In addition, the theory developed in this chapter for the Cox model assumes non-informative censoring. Misspecifications of the censoring mechanism have recently received attention, at least in the classical case; see Kong and Slud (1997) and DiRienzo and Lagakos (2001, 2003). Whether modern robustness ideas can valuably contribute to that type of problem is still an open question. Robust model choice selection for censored data is still a research question with an attempt in that direction by Bednarski and Mocarska (2006) for the Cox model.

7.6 Censored Regression Quantiles

7.6.1 Regression Quantiles

In this section we introduce an approach that is a pure product of robust statistics in the sense that it does not have a classical counterpart. The seminal work dates back to Koenker and Bassett (1978) who were to first to propose to model any pre-specified quantile of a response variable instead of modeling the conditional mean. By doing so they offered statisticians a unique way to explain the entire conditional distribution. As the quantiles themselves can be modeled as a linear function of covariates they are called regression quantiles and the approach is termed quantile regression (QR). This technique was historically introduced as a robust alternative approach to linear regression in the econometric literature. Before presenting the extension to censored data, we present here the basic ideas underlying the QR approach.

The basic idea is to estimate the conditional quantile of an outcome y given a vector of covariates $\boldsymbol{x}$ defined as

$$Q(y, \boldsymbol{x}; \tau) = \inf\{u : P(y \leq u \mid \boldsymbol{x}) = \tau\} \tag{7.28}$$

for any pre-specified level $0 \leq \tau \leq 1$. We further assume that $Q(y, \boldsymbol{x}; \tau)$ is a linear combination of the covariates, i.e.

$$Q(y, \boldsymbol{x}; \tau) = \boldsymbol{x}^T \boldsymbol{\beta}(\tau) \tag{7.29}$$

with $\boldsymbol{\beta}(\tau)$ the τth regression parameter. The rationale for (7.29) is that in many problems the way small or large quantiles depend on the covariates might be quite different from the median response. This will be particularly true in the heteroscedastic data common in the econometric literature where this approach gained rapid popularity. On the other hand, the ability to detect structures for different quantiles is appealing irrespective of the context. The linear specification is the simplest functional form we can imagine and corresponds to the problem of finding regression quantiles in a linear, possibly heterogeneous, regression model. Of course the response function need not be linear and $f(\boldsymbol{x}, \boldsymbol{\beta}(\tau))$ is the obvious extension of the linear predictor in that case. For $0 \leq \tau \leq 1$ define the piecewise-linear function $\rho(u; \tau) = u(\tau - \iota(u < 0))$ where $\iota(u < 0)$ is one when $u < 0$ and zero otherwise. Koenker and Bassett (1978) then showed that a consistent estimator of $\boldsymbol{\beta}(\tau)$ is the value $\hat{\boldsymbol{\beta}}(\tau)$ that minimizes the objective function

$$r(\boldsymbol{\beta}(\tau)) = \sum_{i=1}^{n} \rho(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}(\tau); \tau), \tag{7.30}$$

for an i.i.d. sample $(y_i, \boldsymbol{x}_i)$. When $\tau = 1/2$, $\rho(u; \tau)$ reduces to the absolute value up to a multiplicative factor 1/2. Thus, for the special case of the median this estimator is the so-called L_1-estimator in reference to the absolute (or L_1) norm. For that reason, this approach is also referred to as the L_1 regression quantiles. An introduction to this approach at a low level of technicality with a telling example for a biostatistical audience can be found in Koenker and Hallock (2001).

In their pioneering work Koenker and Bassett (1978) provided an algorithm based on standard linear programming to compute $\hat{\boldsymbol{\beta}}(\tau)$ that was later refined by Koenker and D'Orey (1987). They also proved that this estimator is consistent and asymptotically normal under mild conditions. For instance, in the classical i.i.d. setting we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}(\tau)) \to \mathcal{N}(0, \omega(\tau)\Omega^{-1}) \tag{7.31}$$

where $\omega = \tau(1 - \tau)/f^2(F^{-1}(\tau))$, $\Omega = E[\boldsymbol{x}\boldsymbol{x}^T]$ and f and F are the density and cumulative distribution functions for the error term, respectively. Conditions on f include $f(F^{-1}(\tau)) > 0$ in a neighborhood of τ. It should be stressed that the fact that the asymptotic distribution of $\hat{\boldsymbol{\beta}}(\tau)$ depends on the (unspecified) error distribution can create some difficulties in computing it. Indeed, the density needs to be estimated non-parametrically and the resulting estimates may suffer from a lack of stability. Inferential methods based on the bootstrap might then be preferred. We refer the reader interested in the technical aspects of this work to Koenker and Bassett (1982) for details and for a more comprehensive account discussing inferential aspects to Koenker (2005).

The QR technique took two decades to make its way into survival data analysis, probably because of the lack of flexibility of QR to deal with censoring. A step in

the right direction was suggested by Koenker and Geling (2001). It is based on a simple idea: a transformation of the survival time y_i, e.g. the log-transformation, is used providing a regression quantile approach to accelerated failure time model. This is straightforward when all survival times are indeed observed, see Koenker and Geling (2001) for an instructive example. However, this approach is insufficient for most applications in medical research where censoring occurs.

7.6.2 Extension to the Censored Case

Early attempts to deal with censoring required too strict assumptions making their use relatively limited; see Powell (1986), Buchinsky and Hahn (1998), Honore *et al.* (2002) and Chernozhukov and Hong (2002), among others. The important breakthrough came with Portnoy (2003) who was able to accommodate general forms of censoring. He also made available a user-friendly R package called `CRQ` for censored regression quantiles (directly accessible on his website). We can then expect a rapid development of this innovative approach in biostatistics and medical research where it could be used as a valuable complement to the Cox model.

CRQ involve more technical aspects since it combines both the elements of regression quantiles and the modeling of censored survival times. The reader may decide to skip this section in the first instance and just accept the existence of the extension to the censored case. Let c_i, $i = 1, \ldots, n$ be the censoring times and y_i^0 the possibly unobserved response (e.g. survival time t_i^0) for the ith subject. We have $y_i = \min(y_i^0, c_i)$ (e.g. $y_i = t_i$ the survival time), and $\delta_i = \iota(y_i^0 \leq c_i)$ the indicator of censoring. We can even allow c_i to depend on $\boldsymbol{x}_i$ but require y_i^0 and c_i to be independent conditionally on $\boldsymbol{x}_i$. The model now stipulates that the conditional quantiles of y_i^0 are a linear combination of the covariates but will not impose any particular functional form on those of y_i. Portnoy (2003) astutely noticed that QR is actually a generalization of the one-sample Kaplan–Meier approach. Two key ingredients combine here: (1) the Kaplan–Meier estimator (7.26) can be viewed as a 'recursively reweighted' empirical survival estimate; (2) a more technical argument linked to the regression quantiles computation, i.e. the weighted gradient used in the programming remains piecewise linear in τ. This simple remark permits the use of simplex pivoting techniques. Point (1) follows from Efron (1967) who shows that the Kaplan–Meier estimator can be computed by redistributing the mass of each censored observation to subsequent non-censored observations. In other words, the mass $P(y_i^0 > c_i)$ can be redistributed to observations above c_i. This is done by exploiting a key point of QR, i.e. the estimator $\hat{\boldsymbol{\beta}}(\tau)$ depends on the sign of the residuals at any given point and not on the actual value of the response. The procedure for estimating $\hat{\boldsymbol{\beta}}(\tau)$ when there is censoring works then in the following way. First, it is easy to start with a low quantile τ. We might not know the exact value of y_i^0 but we do know that it is beyond the censoring time c_i. Then, when c_i lies above the τth regression line, so does y_i^0. The true residual $y_i^0 - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}(\tau)$ will be positive irrespective of y_i^0 and we can just use the ordinary QR for such a small quantile value. Of course as τ becomes larger sooner or later a censored observation

will have a negative residual $c_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}(\tau)$. We do not know for sure whether the true residual is positive or negative but as the sign has changed we call such an observation crossed from now on. The level at which the observation is crossed is denoted $\hat{\tau}_i$, thus

$$c_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}(\hat{\tau}_i) \geq 0 \quad \text{and} \quad c_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}(\tau) \leq 0 \quad \text{for all } \tau > \hat{\tau}_i.$$

As explained by Portnoy (2003) and Debruyne *et al.* (2008), the critical idea is 'to estimate the probability of crossed censored observations having a positive, respectively negative residual and then use these estimates as weights further on'. This can be achieved by splitting such an observation into two weighted pseudo-observations, one at $(c_i, \boldsymbol{x}_i)$ with weight $w_i(\tau) \approx P(y_i^0 - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}(\tau) \geq 0)$ and one at $(+\infty, \boldsymbol{x}_i)$ with weight $1 - w_i(\tau)$. The weight itself comes from quantile regression as $1 - \hat{\tau}_i$ is a rough estimate of the censoring probability $P(y_i^0 > c_i)$, i.e.

$$w_i(\tau) = \frac{\tau - \hat{\tau}_i}{1 - \hat{\tau}_i} \quad \text{for } \tau > \hat{\tau}_i.$$

Then we can proceed recursively to obtain the CRQ estimate; the exact algorithm as detailed in Debruyne *et al.* (2008) is given in Appendix G. This process is technically equivalent to one minus the Kaplan–Meier estimate with the Efron recursive reweighting scheme; see the example given in Portnoy (2003, p. 1004), for details. Improvements to the computation of CRQ can also be found in Fitzenberger and Winker (2007) and may prove useful for large datasets.

7.6.3 Asymptotic Properties and Robustness

Establishing asymptotic results for CRQ is a considerable task as the weighting scheme sketched above must be taken into account. The most accurate result so far is that $\hat{\boldsymbol{\beta}}(\tau)$ converges to $\boldsymbol{\beta}(\tau)$ at the rate $n^{-1/2}$ as shown by Neocleous *et al.* (2006). The asymptotic normality with a closed form for the asymptotic variance is still a work in progress. The current way to compute standard errors or CIs is the bootstrap. This technique is computer intensive but stable and provides an effective way to perform inference in the i.i.d. setting; it is also the default method in the R package `CRQ` provided by Portnoy. More generally, even if an asymptotic result were available, it would not necessarily lead to an accurate estimate. Indeed as indicated earlier in (7.31), the asymptotic variance for the regression quantiles estimates in the non-censored case depends on the underlying (unspecified) error distribution and hence bootstrap methods can provide more reliable standard errors estimates. This is also certainly true in the presence of censoring. Elaboration must be made on the exact implementation of the bootstrap for CRQ as a few complications arise. First, when the survival distribution presents many censored observations in its right tail, it is virtually impossible to estimate the conditional quantile above the last τ value corresponding to the last uncensored observation. When bootstrapping the problem is even more serious as this cut-off is random. In one bootstrap sample the observed cut-off can be 0.9 whereas in another one it is about 0.7 due to the

presence of more censored observations from the right-hand tail. Thus, the simple percentile CI possibly fails. Portnoy (2003) introduced a hybrid approach called the 2.906 IQR bootstrap to cope with this problem: simply take the bootstrap estimate of the interquartile range (IQR) and use normality to obtain the relevant percentiles. Technically, this amounts to computing the bootstrap sample interquartile values $\hat{\boldsymbol{\beta}}^*_{0.75} - \hat{\boldsymbol{\beta}}^*_{0.5}$ and $\hat{\boldsymbol{\beta}}^*_{0.5} - \hat{\boldsymbol{\beta}}^*_{0.25}$, multiplying them by 2.906 for consistency and adding the values to the median estimate $\hat{\boldsymbol{\beta}}^*_{0.5}$ to get upper and lower bounds of the 95% CI for all $\boldsymbol{\beta}(\tau)$. This approach seems to work reasonably well both in simulations and examples. Second, as the computational time can be prohibitive for large samples discouraging users, a possible solution has been implemented in the R package `CRQ`. It is called the 'n-choose m' bootstrap whereby replicates of size $m < n$ are chosen to compute the estimates and then adjust the CIs for the smaller sample size. Improvements on the CRQ implementation are work in progress and limitations will certainly be relaxed in the near future.

Regression quantiles inherit the robustness of ordinary sample quantiles, and thus present some form of robustness to distributional assumptions. As pointed out by Koenker and Hallock (2001) the estimates have 'an inherent distribution-free character because quantile estimation is influenced only by the local behavior of the conditional distribution near the specified quantile'. This is equally true for CRQ as long as perturbations in the response only are considered. However, both regression quantiles and CRQ break down in the presence of bad leverage points or problems in the covariates. Robust inference has not been specifically studied but it is safe to say that the bootstrap-based approach probably works well for low levels of contamination and central values of τ (which is probably where most applied problems sit). In contrast extreme values of τ or a higher percentage of spurious data in the sample cause more trouble. Indeed, in that case the standard bootstrap approach breaks down as more outliers can be generated in the bootstrap sample. This is even more critical when extreme τ are the target as the breakdown point of $\hat{\boldsymbol{\beta}}(\tau)$ is automatically lower.

7.6.4 Comparison with the Cox Proportional Hazard Model

Straightforward computations based on the survival function and cumulative hazard given in Section 7.2 show that the conditional quantile for the survival time t given a particular covariate vector $\boldsymbol{x}$ is

$$Q(t, \boldsymbol{x}; \tau) = \Lambda_0^{-1}[-\log(1-\tau)\exp(-\boldsymbol{x}^T\boldsymbol{\beta})]. \tag{7.32}$$

Thus, the exponential form of the Cox model imposes a specific form on the conditional quantiles. More specifically (7.32) shows that they are all monotone in $\log(1-\tau)$ and depend on Λ_0 in a complicated way. As the conditional quantiles are not linear in the covariates the Cox model does not provide a direct analog of $\hat{\boldsymbol{\beta}}(\tau)$. However, Koenker and Geling (2001) and Portnoy (2003) suggested that a good proxy for $\hat{\boldsymbol{\beta}}(\tau)$ is the derivative of (7.32) evaluated at $\bar{\boldsymbol{x}}$, the average covariate

vector, i.e.

$$\boldsymbol{b}(\tau) = \frac{\partial}{\partial \boldsymbol{x}} Q(t, \boldsymbol{x}; \tau)\Big|_{\boldsymbol{x}=\bar{\boldsymbol{x}}}. \tag{7.33}$$

If we now plug in the PLE for $\boldsymbol{\beta}$ into formula (7.33) we obtain $\hat{\boldsymbol{b}}(\tau)$ that we can now compare with the censored regression quantile estimate $\hat{\boldsymbol{\beta}}(\tau)$. It is worth noting that (7.33) implies that

$$b_j(\tau) = -\frac{(1-\tau)^{\gamma(\boldsymbol{x})} \log(1-\tau)\gamma(\boldsymbol{x})}{S_0'[Q(t, \boldsymbol{x}; \tau)]}\beta_j,$$

where $\gamma(\boldsymbol{x}) = \exp(-\boldsymbol{x}^T\boldsymbol{\beta})$. So the effect of the various covariates as a function of τ are all identical up to a scaling factor depending on $\boldsymbol{x}$. In particular, the quantile treatment effect for the Cox model must have the same sign as β_j precluding any form of effect that would allow crossings of the survival functions for different settings of covariates. This can be seen as a lack of flexibility of the Cox model imposed by the proportional hazard assumption.

7.6.5 Lung Cancer Data Example (continued)

Figure 7.6 displays a concise summary of the results for a censored quantile regression analysis of log(time), i.e. an accelerated failure rate model, on the lung cancer data. The model includes eight estimated coefficients but `ptherapy` and `age` were omitted as the same flat non-significant pattern appears for all values of τ and methods. The shaded area represents the 95% pointwise band for each CRQ coefficient obtained by bootstrapping. The dashed line represents the analog of $\hat{\boldsymbol{\beta}}(\tau)$ for the Cox model given by (7.33). The Karnofsky performance status (`karnofsky`) is a standard score of 0–100 assessing the functional ability of a patient to perform tasks; 0 represents death and 100 a normal ability with no complaints. Its effect depicted in the first panel is highly significant at all levels and for both the Cox and CRQ models. Around median values, e.g. $\tau = 0.50$, the CRQ estimate is roughly 0.04 which translates into a multiplicative effect of $\exp(0.04 * 10) = 1.49$ on median survival for a 10 point increase on that scale (holding all other factors constant). The effect looks somehow higher for smaller quantiles and weaker for larger values of τ, a decreasing trend that is not detected by the Cox model. `dduration` and `treatment` have little impact on the outcome for all values of τ strengthening the previous findings that these predictors are not important in these data.

`cell` is a more interesting predictor. No clear effect of squamous versus large cells appears although it seems that in the tails things could be different with possibly a crossover. With the 95% CI also being larger towards the ends, we do not pursue this interpretation. The situation is much neater for small cells where a significant constant effect appears at all levels except perhaps for larger values, $\tau \geq 0.80$ say. An estimate of -0.70 is obtained for $\tau = 0.50$; this means that the presence of small cells reduces the median survival by $1 - \exp(-0.70) = 50\%$ in comparison with large cells. In contrast, the QR estimate (7.33) for the Cox model represented by the

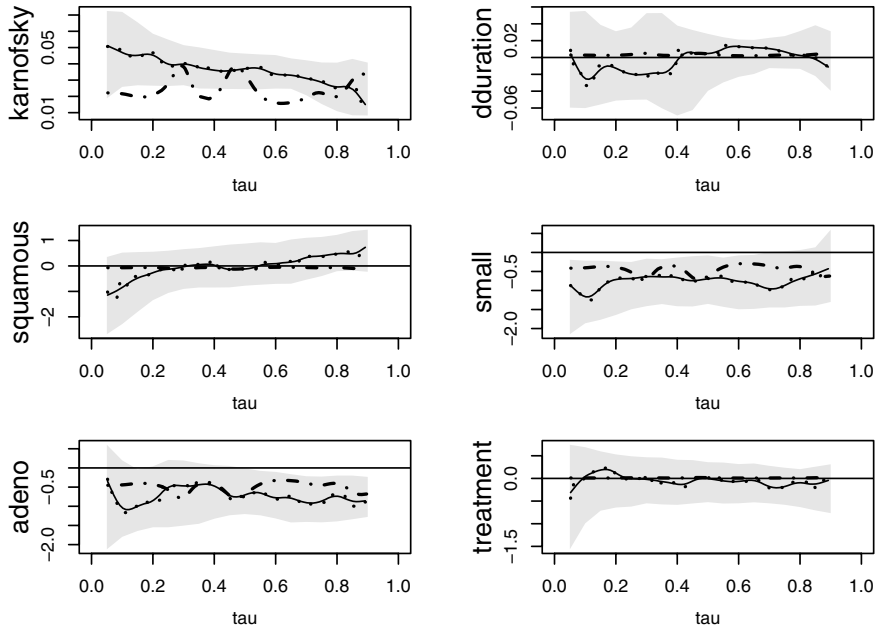

Figure 7.6 The CRQ coefficient, $\hat{\boldsymbol{\beta}}(\tau)$, with shaded 95% band for the lung cancer data. The Cox coefficient effect (7.33) is represented by the dashed line.

dashed line in the same panel is higher, more variable and its significance uncertain, probably for the same reasons mentioned earlier. Adeno cells seem to act similarly to small cells on survival although their effect looks clearer towards the upper end of the distribution. Finally we would like to mention some robustness concerns. As the CRQ approach is based on quantiles, it is robust to outliers in the response or vertical outliers as indicated earlier. It is therefore not influenced by the two long-term survivors (cases 17 and 44). This explains why the robust analysis of Section 7.4 is more in line with the current findings, especially on the role of the cell type.

For the sake of completeness we also give the CRQ fit at $\tau = 0.50$ in Table 7.6. It can be seen as a snapshot of Figure 7.6 at a particular level, here the median. The 95% CIs provided in this table are based on the bootstrap with $B = 1000$ replicates. The p-values correspond to the z-statistic obtained by studentizing by the bootstrap IQR as directly implemented in the R package developed by Portnoy.

It is worth noting that the coefficients are similar to those given in the robust analysis of Section 7.4 up to the minus sign. The systematic reversing of the signs for significant predictors is generally observed. This is due to the fact that CRQ explains a specific quantile of the logarithm of time whereas in a Cox model the classical interpretation with hazard ratios relates more to survival. It is actually possible to obtain similar tables for other values of τ but the graphical summary is usually more informative unless an investigator is interested in one particular quantile of the

Table 7.6 Estimates, 95% CIs and p-values for significance testing for the Veteran's Administration lung cancer data.

Variable	Estimate	95% CI	p-value
Intercept	2.297	(0.45; 4.12)	0.01
`karnofsky`	0.036	(0.02; 0.05)	0.00
`dduration`	0.005	(−0.02; 0.06)	0.80
`age`	0.003	(−0.02; 0.03)	0.83
`ptherapy`	−0.010	(−0.07; 0.04)	0.71
`cell`			
Squamous	−0.117	(−0.81; 0.78)	0.77
Small	−0.685	(−1.28; −0.05)	0.03
Adeno	−0.751	(−1.33; −0.06)	0.02
`treatment`	0.018	(−0.54; 0.44)	0.94

The regression coefficients are estimated by means of the CRQ at $\tau = 0.50$.

distribution. To conclude, it is useful to note that although the differences between the quantile method and the Cox model may not be considered as important in this example, it is not always the case. As pointed out by Portnoy (2003), CRQ generally provides new insight on the data with the discovery of substantial differences when a greater signal-to-noise ratio exists in the data.

7.6.6 Limitations and Extensions

Despite its uniqueness and originality combined with both good local robustness properties and direct interpretation, CRQ have a few limitations. Unlike the proportional hazard model it cannot be extended to time-varying predictors as the whole algorithm is based on fixed $\boldsymbol{x}$. This must be played down as many of the time-dependent covariates used in the extended Cox model are actually introduced when the proportional hazard assumption itself is violated. As the proportional hazard assumption is no longer needed in QR, the problem is no object. From a robustness perspective CRQ is resistant to vertical outliers, i.e. abnormal responses in time, but not to leverage points. Recent work by Debruyne *et al.* (2008) shows that this difficulty can be overcome by introducing censored depth quantiles. More research is needed to study their asymptotic properties and compare them with CRQ. More importantly some more work is needed to sort out inferential issues even though the bootstrap approach described above offers a workable solution. Recently Peng and Huang (2008) introduced a new approach for censored QR based on the Nelson–Aalen estimator of the cumulative hazard function. Implementation of this technique has been provided in the R package `quantreg`; see Koenker (2008). This work is promising as Peng and Huang's estimator admits a Martingale representation providing a natural route for an asymptotic theory. A key assumption of all of these techniques, however, is that $Q(t, \boldsymbol{x}; \tau)$ depends linearly on the regression

parameter $\boldsymbol{\beta}$. This condition can be relaxed in partially linear models as investigated by Neocleous and Portnoy (2006). This could constitute a valuable alternative for intrinsically non-linear data. Irrespective of the method, we would like to stress the potential of QR in biostatistics as it constitutes an original complement to the Cox model. It has the advantage of being naturally interpretable and does not assume any form of proportionality of the hazard function. Results obtained by CRQ can sometimes contradict those derived from the Cox model. This should not be seen as a deficiency but more as a major strength. It can often capture new structures that were hidden behind the proportional hazard assumption. In general, its greater flexibility suggests that the corresponding results are more reliable, but we encourage users to carry out additional work to better understand how such differences can be explained.

Appendices

A

Starting Estimators for *MM*-estimators of Regression Parameters

For the starting point $\widehat{\boldsymbol{\beta}}^0$, one can choose an estimator among the class of *S*-estimators as proposed by Rousseeuw and Yohai (1984) (see also Section 2.3.3). A popular choice for the corresponding ρ-function is the biweight function (2.20), hence leading to the solution $\widehat{\boldsymbol{\beta}}^0_{[bi]}$ for $\boldsymbol{\beta}$ and $\widehat{\sigma}^2_{[bi]}$ for σ^2 which minimize σ^2 subject to

$$\frac{1}{n}\sum_{i=1}^{n}\rho_{[bi]}(r_i;\boldsymbol{\beta},\sigma^2,c) - E_{\Phi}[\rho_{[bi]}(r;\boldsymbol{\beta},\sigma^2,c)] = 0 \tag{A.1}$$

where the expected value ensures Fisher consistency of the resulting estimator. The breakdown point of this *S*-estimator can be chosen through the value of c that satisfies for $\rho_{[bi]}$ the condition $E_{\Phi}[\rho_{[bi]}(r;\boldsymbol{\beta},\sigma^2,c)] = \varepsilon^*\rho_{[bi]}(c;\boldsymbol{\beta},\sigma^2,c)$, where ε^* is the desired breakdown point (see Rousseeuw and Yohai, 1984). When $\varepsilon^* = 0.5$ (the maximal value), then $c = 1.547$ (see Rousseeuw and Leroy, 1987, p. 136). However, its efficiency, i.e. the ratio between the traces of the asymptotic variances of respectively the LS and the *S*-estimator under the exact regression model, is equal to 0.287 (see Yohai *et al.*, 1991), hence it is roughly four times more variable than the LS.

The solution can be found by a random resampling algorithm, followed by a local search (see Yohai *et al.*, 1991), by a genetic algorithm in place of the resampling algorithm, by an exhaustive form of sampling algorithm for small problems (see Marazzi (1993) for details on the numerical algorithms) and by faster algorithm for large problems (see Pena and Yohai, 1999).

The computational speed is still an issue for computing $\hat{\boldsymbol{\beta}}^0$ in general. When some of the explanatory variables are actually categorical (i.e. factors) as is the case

Robust Methods in Biostatistics S. Heritier, E. Cantoni, S. Copt and M.-P. Victoria-Feser

with the diabetes data (variables `bfmed`, `bflar` and `loc`), Maronna and Yohai (2000) propose splitting the estimation procedure into an M-estimation part for the categorical variables and an S-estimation part for the other variables, resulting into what they call an MS-estimator. Basically, consider the following regression model

$$y_i = \boldsymbol{x}_{i(1)}^T \boldsymbol{\beta}_{(1)} + \boldsymbol{x}_{i(2)}^T \boldsymbol{\beta}_{(2)} + \epsilon_i, \quad i = 1, \ldots, n,$$

where $\boldsymbol{x}_{i(1)}$ are 0–1 vectors (i.e. dummy variables) of dimension q_1 and $\boldsymbol{x}_{i(2)}$ are real-valued vectors of dimension q_2. An estimator for $\boldsymbol{\beta}_{(1)}$ is defined conditionally on a value for $\boldsymbol{\beta}_{(2)}$, i.e. the solution $\dot{\boldsymbol{\beta}}_{(1)}(\boldsymbol{\beta}_{(2)})$ in $\boldsymbol{\beta}_{(1)}$ of

$$\sum_{i=1}^{n} \Psi(\tilde{r}_i, \boldsymbol{x}_{i(1)}) = \mathbf{0}, \tag{A.2}$$

with $\tilde{r} = (\tilde{y} - \boldsymbol{x}_{(1)}^T \boldsymbol{\beta}_{(1)})/\sigma$ and $\tilde{y} = y - \boldsymbol{x}_{(2)}^T \boldsymbol{\beta}_{(2)}$. As an estimator for $\boldsymbol{\beta}_{(2)}$ one uses e.g. the S-estimator (A.1) in which $r_i = y_i - \boldsymbol{x}_{i(1)}^T \dot{\boldsymbol{\beta}}_{(1)}(\boldsymbol{\beta}_{(2)}) + \boldsymbol{x}_{i(2)}^T \boldsymbol{\beta}_{(2)}$. For a discussion on the choice for the Ψ-function in (A.2) and simplified numerical procedures, see Maronna and Yohai (2000).

One can also choose different ρ-functions and/or other objective functions to define high breakdown point estimators for the starting point. Indeed, one can cite the least median of squares estimator (LMS) and the least trimmed squares estimator (LTS), both from Rousseeuw (1984), and the least absolute deviations estimator (LAD) of Edgeworth (1887) (see also Bloomfield and Steiger, 1983) also known under L_1-regression. They can be seen in their definition as natural adaptation of the LS estimator or as a particular case of S-estimators. Indeed, the LS (for a given σ^2) is defined as the solution of

$$\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} r_i^2, \tag{A.3}$$

i.e. the minimization of a scale estimate of the residuals, in a similar manner as for S-estimators (the square of the residuals is generalized to a function ρ). Replacing the mean by the median leads to the LMS, using a trimmed mean leads to the LTS and taking the absolute value instead of the square in (A.3) leads to the LAD. All of these estimators require a robust estimator for the scale σ and special algorithms to compute them. They have progressively been abandoned in favor of $\widehat{\boldsymbol{\beta}}_{[bi]}^0$ (and $\widehat{\sigma}_{[bi]}^2$).

B

Efficiency, LRT_ρ, RAIC and RC_p with Biweight ρ-function for the Regression Model

To develop the efficiency (3.20) and other quantities for the LRT_ρ, RAIC and the RC_p with the biweight estimator with ρ-function (3.15), we make use of

$$E[r^k] = \frac{(k)!}{2^{k/2}(k/2)!}$$

to compute the moments of a $\mathcal{N}(0, 1)$, and of

$$\int_{-\infty}^{c} r^k \, d\Phi(r) = L_k = -c^{k-1}\phi(c) + (k-1)\Phi(c)L_{k-2},$$

with $L_0 = \Phi(c)$ and $L_1 = -\phi(c)$. We need (even) moments up to the order 14, i.e.

$$\begin{aligned}
L_2 &= -c\phi(c) + \Phi(c)^2 \\
L_4 &= -(c^3 + 3c\Phi(c))\phi(c) + 3\Phi(c)^3 \\
L_6 &= -(c^5 + 5c^3\Phi(c) + 15c\Phi(c)^2)\phi(c) + 15\Phi(c)^4 \\
L_8 &= -(c^7 + 7c^5\Phi(c) + 35c^3\Phi(c)^2 + 105c\Phi(c)^3)\phi(c) + 105\Phi(c)^5 \\
L_{10} &= -(c^9 + 9c^7\Phi(c) + 63c^5\Phi(c)^2 + 315c^3\Phi(c)^3 + 945c\Phi(c)^4)\phi(c) \\
&\quad + 945\Phi(c)^6 \\
L_{12} &= -(c^{11} + 11c^9\Phi(c) + 99c^7\Phi(c)^2 + 693c^5\Phi(c)^3 + 3465c^3\Phi(c)^4 \\
&\quad + 10\,395c\Phi(c)^5)\phi(c) + 10\,395\Phi(c)^7 \\
L_{14} &= -(c^{13} + 13c^{11}\Phi(c) + 143c^9\Phi(c)^2 + 1287c^7\Phi(c)^3 + 9009c^5\Phi(c)^4 \\
&\quad + 45\,045c^3\Phi(c)^5 + 135\,135c\Phi(c)^6)\phi(c) + 135\,135\Phi(c)^8
\end{aligned}$$

and, therefore,

$$\begin{aligned}
\int_{-c}^{c} d\Phi(r) &= 1 - 2\Phi(-c) \\
\int_{-c}^{c} r^2\, d\Phi(r) &= \int r^2\, d\Phi(r) - 2\int_{-\infty}^{-c} r^2\, d\Phi(r) = 1 - 2\phi(c)c - 2\Phi(-c)^2 \\
\int_{-c}^{c} r^4\, d\Phi(r) &= 3 - 2\phi(c)(c^3 + 3c\Phi(-c)) - 6\Phi(-c)^3 \\
\int_{-c}^{c} r^6\, d\Phi(r) &= 15 - 2\phi(c)(c^5 + 5c^3\Phi(-c) + 15c\Phi(-c)^2) - 30\Phi(-c)^4 \\
\int_{-c}^{c} r^8\, d\Phi(r) &= 105 - 2\phi(c)(c^7 + 7c^5\Phi(-c) + 35c^3\Phi(-c)^2 \\
&\quad + 105c\Phi(-c)^3) - 210\Phi(-c)^5 \\
\int_{-c}^{c} r^{10}\, d\Phi(r) &= 945 - 2\phi(c)(c^9 + 9c^7\Phi(c) + 63c^5\Phi(c)^2 \\
&\quad + 315c^3\Phi(c)^3 + 945c\Phi(c)^4) - 1890\Phi(c)^6 \\
\int_{-c}^{c} r^{12}\, d\Phi(r) &= 10\,395 - 2\phi(c)(c^{11} + 11c^9\Phi(-c) + 99c^7\Phi(-c)^2 \\
&\quad + 693c^5\Phi(-c)^3 + 3465c^3\Phi(-c)^4 + 10\,395c\Phi(-c)^5) \\
&\quad - 20\,790\Phi(c)^7 \\
\int_{-c}^{c} r^{14}\, d\Phi(r) &= 135\,135 - 2\phi(c)(c^{13} + 13c^{11}\Phi(-c) + 143c^9\Phi(-c)^2 \\
&\quad + 1287c^7\Phi(-c)^3 + 9009c^5\Phi(-c)^4 + 45\,045c^3\Phi(-c)^5 \\
&\quad + 135\,135c\Phi(-c)^6) - 270\,270\Phi(-c)^8.
\end{aligned}$$

For the efficiency (3.20), we have

$$\begin{aligned}
e_c = &\left[\frac{5}{c^4}\int_{-c}^{c} r^4\, d\Phi(r) - \frac{6}{c^2}\int_{-c}^{c} r^2\, d\Phi(r) + \int_{-c}^{c} d\Phi(r)\right]^2 \\
&\times \left(\frac{1}{c^8}\int_{-c}^{c} r^{10}\, d\Phi(r) - \frac{4}{c^6}\int_{-c}^{c} r^8\, d\Phi(r) + \frac{6}{c^4}\int_{-c}^{c} r^6\, d\Phi(r)\right. \\
&\left. - \frac{4}{c^2}\int_{-c}^{c} r^4\, d\Phi(r) + \int_{-c}^{c} r^2\, d\Phi(r)\right)^{-1}.
\end{aligned}$$

For the LRT$_\rho$, and using the ρ-function given in (3.15), we have that (3.26) reduces to

$$\left(\frac{5}{c^4}\int_{-c}^{c} r^4\,d\Phi(r) - \frac{6}{c^2}\int_{-c}^{c} r^2\,d\Phi(r) + \int_{-c}^{c} d\Phi(r)\right) \times \left(\frac{1}{c^8}\int_{-c}^{c} r^{10}\,d\Phi(r) - \frac{4}{c^6}\int_{-c}^{c} r^8\,d\Phi(r) + \frac{6}{c^4}\int_{-c}^{c} r^6\,d\Phi(r) - \frac{4}{c^2}\int_{-c}^{c} r^4\,d\Phi(r) + \int_{-c}^{c} r^2\,d\Phi(r)\right)^{-1}.$$

For the RAIC given in (3.31), and using the ρ-function given in (3.15), we have

$$a = \left(\frac{1}{c^8}\int_{-c}^{c} r^{10}\,d\Phi(r) - \frac{4}{c^6}\int_{-c}^{c} r^8\,d\Phi(r) + \frac{6}{c^4}\int_{-c}^{c} r^6\,d\Phi(r) - \frac{4}{c^2}\int_{-c}^{c} r^4\,d\Phi(r) + \int_{-c}^{c} r^2\,d\Phi(r)\right)$$

$$b = \left(\frac{5}{c^4}\int_{-c}^{c} r^4\,d\Phi(r) - \frac{6}{c^2}\int_{-c}^{c} r^2\,d\Phi(r) + \int_{-c}^{c} d\Phi(r)\right).$$

For the RC_p, Ronchetti and Staudte (1994) have shown that

$$\begin{aligned} U_p - V_p = {} & n\int\left(\frac{\partial}{\partial r}\rho(r)\right)^2 d\Phi(r) \\ & - 2p\int\left(\frac{\partial}{\partial r}\rho(r)\right)^2 \frac{\partial^2}{\partial r\,\partial r}\rho(r)\,d\Phi(r)\left[\int \frac{\partial^2}{\partial r\,\partial r}\rho(r)\,d\Phi(r)\right]^{-1} \\ & + p\left(\int\left(\frac{\partial^2}{\partial r\,\partial r}\rho(r)\right)^2 d\Phi(r) + 2\int \frac{1}{r}\frac{\partial}{\partial r}\rho(r)\frac{\partial^2}{\partial r\,\partial r}\rho(r)\,d\Phi(r)\right. \\ & \left. - 3\int \frac{1}{r^2}\left(\frac{\partial}{\partial r}\rho(r)\right)^2 d\Phi(r)\right)\int\left(\frac{\partial}{\partial r}\rho(r)\right)^2 d\Phi(r) \\ & \times\left[\int \frac{\partial^2}{\partial r\,\partial r}\rho(r)\,d\Phi(r)\right]^{-2} \end{aligned}$$

and

$$V_P = p\int \frac{1}{r^2}\left(\frac{\partial}{\partial r}\rho(r)\right)^2 d\Phi(r)\int\left(\frac{\partial}{\partial r}\rho(r)\right)^2 d\Phi(r)\left[\int \frac{\partial^2}{\partial r\,\partial r}\rho(r)\,d\Phi(r)\right]^{-2}.$$

For the biweight ρ-function (3.15), we have

$$
\begin{aligned}
(U_p - V_p) = n&\left(\frac{1}{c^8}\int_{-c}^{c} r^{10}\,d\Phi(r) - \frac{4}{c^6}\int_{-c}^{c} r^{8}\,d\Phi(r) + \frac{6}{c^4}\int_{-c}^{c} r^{6}\,d\Phi(r)\right)\\
&- n\left(\frac{4}{c^2}\int_{-c}^{c} r^{4}\,d\Phi(r) - \int_{-c}^{c} r^{2}\,d\Phi(r)\right)\\
&- 2p\left(\frac{5}{c^{12}}\int_{-c}^{c} r^{14}\,d\Phi(r) - \frac{26}{c^{10}}\int_{-c}^{c} r^{12}\,d\Phi(r)\right.\\
&+ \frac{55}{c^8}\int_{-c}^{c} r^{10}\,d\Phi(r) - \frac{60}{c^6}\int_{-c}^{c} r^{8}\,d\Phi(r) + \frac{35}{c^4}\int_{-c}^{c} r^{6}\,d\Phi(r)\\
&\left.- \frac{10}{c^2}\int_{-c}^{c} r^{4}\,d\Phi(r) + \int_{-c}^{c} r^{2}\,d\Phi(r)\right)\\
&\times\left[\frac{5}{c^4}\int_{-c}^{c} r^{4}\,d\Phi(r) - \frac{6}{c^2}\int_{-c}^{c} r^{2}\,d\Phi(r) + \int_{-c}^{c} d\Phi(r)\right]^{-1}\\
&+ p\left(\frac{32}{c^8}\int_{-c}^{c} r^{8}\,d\Phi(r) - \frac{80}{c^6}\int_{-c}^{c} r^{6}\,d\Phi(r)\right.\\
&\left.+ \frac{64}{c^4}\int_{-c}^{c} r^{4}\,d\Phi(r) - \frac{16}{c^2}\int_{-c}^{c} r^{2}\,d\Phi(r)\right)\\
&\times\left(\frac{1}{c^8}\int_{-c}^{c} r^{10}\,d\Phi(r) - \frac{4}{c^6}\int_{-c}^{c} r^{8}\,d\Phi(r) + \frac{6}{c^4}\int_{-c}^{c} r^{6}\,d\Phi(r)\right.\\
&\left.- \frac{4}{c^2}\int_{-c}^{c} r^{4}\,d\Phi(r) + \int_{-c}^{c} r^{2}\,d\Phi(r)\right)\\
&\times\left[\frac{5}{c^4}\int_{-c}^{c} r^{4}\,d\Phi(r) - \frac{6}{c^2}\int_{-c}^{c} r^{2}\,d\Phi(r) + \int_{-c}^{c} d\Phi(r)\right]^{-2}
\end{aligned}
$$

and

$$
\begin{aligned}
V_P = p&\left(\frac{1}{c^8}\int_{-c}^{c} r^{8}\,d\Phi(r) - \frac{4}{c^6}\int_{-c}^{c} r^{6}\,d\Phi(r) + \frac{6}{c^4}\int_{-c}^{c} r^{4}\,d\Phi(r)\right.\\
&\left.- \frac{4}{c^2}\int_{-c}^{c} r^{2}\,d\Phi(r) + \int_{-c}^{c} d\Phi(r)\right)\left(\frac{1}{c^8}\int_{-c}^{c} r^{10}\,d\Phi(r) - \frac{4}{c^6}\int_{-c}^{c} r^{8}\,d\Phi(r)\right.\\
&\left.+ \frac{6}{c^4}\int_{-c}^{c} r^{6}\,d\Phi(r) - \frac{4}{c^2}\int_{-c}^{c} r^{4}\,d\Phi(r) + \int_{-c}^{c} r^{2}\,d\Phi(r)\right)\\
&\times\left[\frac{5}{c^4}\int_{-c}^{c} r^{4}\,d\Phi(r) - \frac{6}{c^2}\int_{-c}^{c} r^{2}\,d\Phi(r) + \int_{-c}^{c} d\Phi(r)\right]^{-2}.
\end{aligned}
$$

C

An Algorithm Procedure for the Constrained *S*-estimator

The following is a pseudo code of the algorithm for computing the constrained *S*-estimator.

- Given a model, define the design matrices $z_j z_j^T$ to obtain the structure of the covariance matrix and the matrices $\boldsymbol{x}_i$ that define the mean vectors $\boldsymbol{x}_i\boldsymbol{\beta}$, so that

$$\boldsymbol{\Sigma} = \sum_{j=0}^{r} \sigma_j^2 z_j z_j^T.$$

- Compute the starting point of the constrained estimator, that is

$$\boldsymbol{x}_i\boldsymbol{\beta}_{\text{start}} \quad \text{and} \quad \boldsymbol{\Sigma}_{\text{start}}.$$

  In principle one can choose any high breakdown point estimator as starting point. It can be made 'constrained' to match the MLM model by averaging out the elements of the estimated covariance matrix that are equal under the MLM. We use the MCD estimator (see Section 2.3.3).

- Compute the constrained estimator through the following iterative procedure:

  1. Compute the Mahalanobis distances

$$d_i^{(1)} = \sqrt{(\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta}_{\text{start}})^T \boldsymbol{\Sigma}_{\text{start}}^{-1} (\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta}_{\text{start}})}.$$

  2. Compute the weights $w(d_i^{(1)})$.
  3. Compute the fixed effects parameters $\boldsymbol{\beta}^{(1)}$ by solving

$$\sum w(d_i^{(1)}) \boldsymbol{x}_i^T \boldsymbol{\Sigma}_{\text{start}}^{-1} (\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta}_{\text{start}}).$$

Robust Methods in Biostatistics S. Heritier, E. Cantoni, S. Copt and M.-P. Victoria-Feser

4. Let $\boldsymbol{\alpha} = (\sigma_0^2, \ldots, \sigma_r^2)^T$, an iterative expression for the variance components $\boldsymbol{\alpha}^{(1)}$ is given by

$$\boldsymbol{\alpha}^{(1)} = \left(\frac{1}{n} \sum_{i=1}^{n} w(d_i^{(1)})(d_i^{(1)})^2 \right)^{-1} \boldsymbol{Q}^{-1} \boldsymbol{U}$$

with $\boldsymbol{U}$ defined as

$$\boldsymbol{U} = \left(\frac{1}{n} \sum pw(d_i^{(1)})(\boldsymbol{y}_i - \boldsymbol{x}_i \boldsymbol{\beta}_{start})^T \times \boldsymbol{\Sigma}_{start}^{-1} \boldsymbol{z}_j \boldsymbol{z}_j^T \boldsymbol{\Sigma}_{start}^{-1} (\boldsymbol{y}_i - \boldsymbol{x}_i \boldsymbol{\beta}_{start}) \right)_{j=0,\ldots,r}$$

and

$$\boldsymbol{Q} = tr(\boldsymbol{M}_j \boldsymbol{M}_k)_{j,k=0,\ldots,r}$$

with

$$\boldsymbol{M}_j = \boldsymbol{\Sigma}_{start}^{-1} \boldsymbol{z}_j \boldsymbol{z}_j^T.$$

5. Using the design matrices $\boldsymbol{z}_j \boldsymbol{z}_j^T$, update the constrained matrix by

$$\boldsymbol{\Sigma}^{(1)} = \sum_{j=0}^{r} \sigma_j^{2(1)} \boldsymbol{z}_j \boldsymbol{z}_j^T.$$

6. Update the fixed effects by

$$\boldsymbol{x}_i \boldsymbol{\beta}^{(1)}.$$

7. Compute some convergence criterion. If the conditions of the criterion are met, stop; otherwise put $\boldsymbol{\beta}_{\text{start}} = \boldsymbol{\beta}^{(1)}$, $\boldsymbol{\Sigma}_{\text{start}} = \boldsymbol{\Sigma}^{(1)}$ and start again at point 1 by computing $d_i^{(2)}$, the weights $w(d_i^{(2)})$ then $\boldsymbol{\beta}^{(2)}$ and $\boldsymbol{\Sigma}^{(2)}$. Repeat the procedure until convergence.

D

Some Distributions of the Exponential Family

We give here the definitions of some of the distributions belonging to the exponential family, as listed in Table 5.1.

- **Normal.** The density function of a variable distributed as $y_i \sim \mathcal{N}(\mu_i, \sigma^2)$ is

$$f(y; \mu_i, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(y - \mu_i)^2\right),$$

  for y in $\mathcal{R}$.

- **Bernoulli.** A y_i Bernoulli distributed variable can take values $y = 0$ or $y = 1$ according to

$$P(y_i = y; p_i) = p_i^y (1 - p_i)^{1-y}.$$

- **Scaled binomial.** The scaled binomial distributed variables y_i/m take values $0, 1/m, 2/m, \ldots, 1$ and are derived from the binomial variables y_i with probabilities

$$P(y_i = y; p_i) = \binom{m}{y} p_i^y (1 - p_i)^{1-y},$$

  for $y = 0, 1, \ldots, m$.

- **Poisson.** For a Poisson variable $y_i \sim \mathcal{P}(\lambda)$, probabilities are computed according to

$$P(y_i = y; \lambda_i) = \exp(-\lambda_i)\frac{\lambda_i^y}{y!},$$

  for $y = 0, 1, 2, \ldots$.

Robust Methods in Biostatistics S. Heritier, E. Cantoni, S. Copt and M.-P. Victoria-Feser

- **Gamma.** Here y_i is said to be $\Gamma(\mu_i, \nu)$ distributed if its density is

$$f(y; \mu_i, \nu) = \frac{\nu/\mu_i \cdot \exp(-\nu y/\mu_i) \cdot (\nu y/\mu_i)^{\nu-1}}{\Gamma(\nu)},$$

for $y > 0$, with $\Gamma(\nu) = \int_0^\infty \exp(-u)u^{\nu-1}\, du$.

E

Computations for the Robust GLM Estimator

E.1 Fisher Consistency Corrections

We give here the Fisher consistency corrections

$$a(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} E[\psi(r_i; \boldsymbol{\beta}, \phi, c)]w(\boldsymbol{x}_i)\frac{1}{\sqrt{\phi v_{\mu_i}}}\mu_i',$$

for the binomial, Poisson and Gamma models. Note that for binomial and Poisson models, $\phi = 1$ and for the Gamma model $\phi = 1/\nu$, see Table 5.1. The only term to be computed for each model is $E[\psi(r_i; \boldsymbol{\beta}, \phi, c)]$, which is done below for $\psi(r_i; \boldsymbol{\beta}, \phi, c) = \psi_{[Hub]}(r_i; \boldsymbol{\beta}, \phi, c)$, see Section 3.6.

Let us first define $j_1 = \lfloor \mu_i - c\sqrt{\phi v_{\mu_i}} \rfloor$ and $j_2 = \lfloor \mu_i + c\sqrt{\phi v_{\mu_i}} \rfloor$, where $\lfloor u \rfloor$ denotes the largest integer not greater than u.

The binomial model states that $y_i \sim \mathcal{B}(m_i, p_i)$, so that $E[y_i] = \mu_i = m_i p_i$ and $\text{var}(y_i) = \mu_i((m_i - \mu_i)/m_i)$. Then we have

$$\begin{aligned} E[\psi_{[Hub]}(r_i; \boldsymbol{\beta}, \phi, c)] &= \sum_{j=-\infty}^{\infty} \psi_{[Hub]}\left(\frac{j - \mu_i}{\sqrt{v_{\mu_i}}}; \boldsymbol{\beta}, \phi, c\right) P(y_i = j)\, \iota(j \in [0, m_i]) \\ &= c[P(y_i \geq j_2 + 1) - P(y_i \leq j_1)] \\ &\quad + \frac{\mu_i}{\sqrt{v_{\mu_i}}}[P(j_1 \leq \tilde{y}_i \leq j_2 - 1) - P(j_1 + 1 \leq y_i \leq j_2)], \end{aligned}$$

with $\tilde{y}_i \sim \mathcal{B}(m_i - 1, p_i)$, and where $\iota(C)$ is the indicator function that takes the value one if C is true and zero otherwise.

Robust Methods in Biostatistics S. Heritier, E. Cantoni, S. Copt and M.-P. Victoria-Feser

The Poisson model states that $y_i \sim \mathcal{P}(\mu_i)$ and, hence, $E[y_i] = V(\mu_i) = \mu_i$. Then,

$$\begin{aligned} E[\psi_{[Hub]}(r_i; \boldsymbol{\beta}, \phi, c)] &= \sum_{j=-\infty}^{\infty} \psi_{[Hub]}\left(\frac{j-\mu_i}{\sqrt{v_{\mu_i}}}; \boldsymbol{\beta}, \phi, c\right) P(y_i = j)\iota(j \geq 0) \\ &= c(P(y_i \geq j_2 + 1) - P(y_i \leq j_1)) \\ &\quad + \frac{\mu_i}{\sqrt{v_{\mu_i}}}[P(y_i = j_1) - P(y_i = j_2)]. \end{aligned}$$

Finally, for the Gamma model, one remarks in the first place that $r_i = (y_i - \mu_i)/\sqrt{\phi v_{\mu_i}}$ has a Gamma distribution (independent of μ_i) with expectation equal to $\sqrt{\nu}$ and shifted origin to $-\sqrt{\nu}$. It holds that

$$\begin{aligned} E[\psi_{[Hub]}(r_i; \boldsymbol{\beta}, \phi, c)] &= \int_{-\sqrt{\nu}}^{\infty} \psi_{[Hub]}(r; \boldsymbol{\beta}, \phi, c) f(r; \sqrt{\nu}, \nu)\, \iota(r > -\sqrt{\nu})\, dr \\ &= c[P(r_i > c) - P(r_i < -c)] \\ &\quad + \frac{\nu^{(\nu-1)/2}}{\Gamma(\nu)}[G(-c, \nu) - G(c, \nu)], \end{aligned}$$

where $f(r; \sqrt{\nu}, \nu)$ is the Gamma density (see Appendix D) and

$$G(t, \kappa) = \exp(-\sqrt{\nu}(\sqrt{\nu} + t))(\sqrt{\nu} + t)^{\kappa}\, \iota(t > -\sqrt{\nu}).$$

E.2 Asymptotic Variance

Computing the asymptotic variance amounts to computing the matrices A and B of Section 5.3.4, and therefore of $E[\psi^2(r_i; \boldsymbol{\beta}, \phi, c)]$ and $E[\psi(r_i; \boldsymbol{\beta}, \phi, c)(\partial/\partial\mu_i) \log h(y_i \mid \boldsymbol{x}_i, \mu_i)]$ again for $\psi(r_i; \boldsymbol{\beta}, \phi, c) = \psi_{[Hub]}(r_i; \boldsymbol{\beta}, \phi, c)$, where $h(y_i \mid \boldsymbol{x}_i, \boldsymbol{\mu}_i)$ is the conditional density or probability of $y_i \mid \boldsymbol{x}_i$.

For the binomial model

$$\begin{aligned} &E[\psi^2_{[Hub]}(r_i; \boldsymbol{\beta}, \phi, c)] \\ &\quad = c^2(P(y_i \leq j_1) + P(y_i \geq j_2 + 1)) \\ &\qquad + \frac{1}{v_{\mu_i}}[\pi_i^2 m_i(m_i - 1)P(j_1 - 1 \leq \tilde{\tilde{y}}_i \leq j_2 - 2) \\ &\qquad + (\mu_i - 2\mu_i^2)P(j_1 \leq \tilde{y}_i \leq j_2 - 1) + \mu_i^2 P(j_1 + 1 \leq y_i \leq j_2)], \end{aligned}$$

with $y_i \sim \mathcal{B}(m_i, \pi_i)$, $\tilde{y}_i \sim \mathcal{B}(m_i - 1, \pi_i)$ and $\tilde{\tilde{y}}_i \sim \mathcal{B}(m_i - 2, \pi_i)$ $(m_i \geq 3)$.

Given that $(\partial/\partial\mu_i)\log h(y_i \mid \boldsymbol{x}_i, \mu_i)$ is equal to $(y_i - \mu_i)/v_{\mu_i}$, we have

$$
\begin{aligned}
&E\left[\psi_{[Hub]}(r_i; \boldsymbol{\beta}, \phi, c)\frac{\partial}{\partial\mu_i}\log h(y_i \mid \boldsymbol{x}_i, \mu_i)\right] \\
&\quad = E\left[\psi_{[Hub]}(r_i; \boldsymbol{\beta}, \phi, c)\frac{y_i - \mu_i}{v_{\mu_i}}\right] \\
&\quad = \frac{c\mu_i}{v_{\mu_i}}[P(y_i \le j_1) - P(\tilde{y}_i \le j_1 - 1) + P(\tilde{y}_i \ge j_2) - P(y_i \ge j_2 + 1)] \\
&\qquad + \frac{1}{v_{\mu_i}^{3/2}}[\pi_i^2 m_i(m_i - 1)P(j_1 - 1 \le \tilde{\tilde{y}}_i \le j_2 - 2) \\
&\qquad + (\mu_i - 2\mu_i^2)P(j_1 \le \tilde{y}_i \le j_2 - 1) + \mu_i^2 P(j_1 + 1 \le y_i \le j_2)],
\end{aligned}
$$

with $y_i \sim \mathcal{B}(m_i, \pi_i)$, $\tilde{y}_i \sim \mathcal{B}(m_i - 1, \pi_i)$ and $\tilde{\tilde{y}}_i \sim \mathcal{B}(m_i - 2, \pi_i)$ $(m_i \ge 3)$.

For the Poisson model,

$$
\begin{aligned}
E[\psi_{[Hub]}^2(r_i; \boldsymbol{\beta}, \phi, c)] = c^2[&P(y_i \le j_1) + P(y_i \ge j_2 + 1)] \\
&+ \frac{1}{v_{\mu_i}}[\mu_i^2 P(j_1 - 1 \le y_i \le j_2 - 2) \\
&+ (\mu_i - 2\mu_i^2)P(j_1 \le y_i \le j_2 - 1) \\
&+ \mu_i^2 P(j_1 + 1 \le y_i \le j_2)].
\end{aligned}
$$

We have

$$
\frac{\partial}{\partial\mu_i}\log h(y_i \mid \boldsymbol{x}_i, \mu_i) = \frac{y_i - \mu_i}{\mu_i} = \frac{y_i - \mu_i}{v_{\mu_i}},
$$

so that

$$
\begin{aligned}
&E\left[\psi_{[Hub]}(r_i; \boldsymbol{\beta}, \phi, c)\frac{\partial}{\partial\mu_i}\log h(y_i \mid \boldsymbol{x}_i, \mu_i)\right] \\
&\quad = E\left[\psi_{[Hub]}(r_i; \boldsymbol{\beta}, \phi, c)\frac{y_i - \mu_i}{v_{\mu_i}}\right] \\
&\quad = c[P(y_i = j_1) + P(y_i = j_2)] + \mu_i P(j_1 \le y_i \le j_2 - 1) \\
&\qquad + \frac{1}{v_{\mu_i}^{3/2}}\mu_i^2[P(y_i = j_1 - 1) - P(y_i = j_1) - P(y_i = j_2 - 1) + P(y_i = j_2)].
\end{aligned}
$$

For the Gamma model, we first note that

$$
\frac{\partial}{\partial\mu_i}\log h(y_i \mid \boldsymbol{x}_i, \mu_i) = (y_i - \mu_i)/(\mu_i^2/\nu) = \sqrt{\nu}r_i/\mu_i.
$$

This yields

$$
\begin{aligned}
&E(\psi_{[Hub]}(r_i;\boldsymbol{\beta},\phi,c)\frac{\partial}{\partial\mu_i}\log h(y_i \mid \boldsymbol{x}_i,\mu_i)\\
&\quad=\frac{\sqrt{v}}{\mu_i}E(\psi_{[Hub]}(r_i;\boldsymbol{\beta},\phi,c)r_i)\\
&\quad=\frac{v^{v/2}c}{\mu_i\Gamma(v)}[G(-c,v)+G(c,v)]+\frac{\sqrt{v}}{\mu_i}P(-c<r_i<c)\\
&\qquad+\frac{v^{v/2}}{\mu_i\Gamma(v)}[G(-c,v+1)-G(c,v+1)]\\
&\qquad+\frac{v^{(v+1)/2}}{\mu_i\Gamma(v)}\left(\frac{v+1}{v}-2\right)[G(-c,v)-G(c,v)].
\end{aligned}
$$

E.3 IRWLS Algorithm for Robust GLM

We show here how the estimation procedure issued from (5.13) can be written as an IRWLS algorithm. Given $\boldsymbol{\beta}^{t-1}$, the estimated value of $\boldsymbol{\beta}$ at iteration $t-1$, one can obtain $\boldsymbol{\beta}^t$, the value of $\boldsymbol{\beta}$ at iteration t, by regressing $\mathbf{Z}=\mathbf{X}^T\boldsymbol{\beta}^{t-1}+\boldsymbol{e}^{t-1}$ on $\mathbf{X}$ (see Definition (5.2)) with weights $B=\mathrm{diag}(b_1,\ldots,b_n)$ with

$$
b_i=E\left[\psi(r_i;\boldsymbol{\beta},\phi,c)\frac{\partial}{\partial\mu_i}\log h(y_i \mid \boldsymbol{x}_i,\mu_i)\right]\Big/\sqrt{\phi v_{\mu_i}}w(\boldsymbol{x}_i)\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2, \qquad \text{(E.1)}
$$

for $i=1,\ldots,n$, where $h(\cdot)$ is the conditional density or probability of $y_i \mid \boldsymbol{x}_i$ and $\boldsymbol{e}^{t-1}=(e_1^{t-1},\ldots,e_n^{t-1})$ with

$$
e_i^{t-1}=\frac{\psi(r_i^{t-1};\boldsymbol{\beta},\phi,c)-E[\psi(r_i^{t-1};\boldsymbol{\beta},\phi,c)]}{E[\psi(r_i^{t-1};\boldsymbol{\beta},\phi,c)(\partial/\partial\mu_i)\log h(y_i \mid \boldsymbol{x}_i,\mu_i^{t-1})]}. \qquad \text{(E.2)}
$$

To see the above, define $U(\boldsymbol{\beta})=\sum_{i=1}^n\Psi(y_i,\boldsymbol{x}_i;\boldsymbol{\beta},\phi,c)$, where $\Psi(y_i,\boldsymbol{x}_i;\boldsymbol{\beta},\phi,c)$ is given in (5.13).

The Fisher-scoring algorithm at step t writes

$$
\boldsymbol{\beta}^t=\boldsymbol{\beta}^{t-1}+H^{-1}(\boldsymbol{\beta}^{t-1})U(\boldsymbol{\beta}^{t-1})
$$

or, alternatively,

$$
H(\boldsymbol{\beta}^{t-1})\boldsymbol{\beta}^t=H(\boldsymbol{\beta}^{t-1})\boldsymbol{\beta}^{t-1}+U(\boldsymbol{\beta}^{t-1}),
$$

where

$$
H(\boldsymbol{\beta}^{t-1})=E\left[-\frac{\partial}{\partial\boldsymbol{\beta}}U(\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{t-1}}\right]=nM(\Psi,F_{\boldsymbol{\beta}})=\boldsymbol{X}^TB|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{t-1}}\boldsymbol{X}.
$$

Moreover, for $\mathbf{Z} = \mathbf{X}^T \boldsymbol{\beta}^{t-1} + \boldsymbol{e}^{t-1}$ with $\boldsymbol{e}^{t-1}$ as defined in (E.2), we have that $H(\boldsymbol{\beta}^{t-1})\boldsymbol{\beta}^{t-1} + U(\boldsymbol{\beta}^{t-1}) = \mathbf{X}^T B\mathbf{Z}$. In fact, for each $j = 1, \ldots, n$, it holds that

$$
\begin{aligned}
&[H(\boldsymbol{\beta}^{t-1})\boldsymbol{\beta}^{t-1} + U(\boldsymbol{\beta}^{t-1})]_j \\
&\quad = \sum_{k=1}^{p} \sum_{i=1}^{n} b_i x_{ij} x_{ik} \beta_k^{t-1} + \sum_{i=1}^{n} \psi(r_i; \boldsymbol{\beta}, \phi, c) w(\boldsymbol{x}_i) \frac{1}{\sqrt{\phi v_{\mu_i}}} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} \\
&\qquad - \sum_{i=1}^{n} E[\psi(r_i; \boldsymbol{\beta}, \phi, c)] w(\boldsymbol{x}_i) \frac{1}{\sqrt{\phi v_{\mu_i}}} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} \\
&\quad = \sum_{i=1}^{n} \left\{ \sum_{k=1}^{p} x_{ik} \beta_k^{t-1} + \frac{\psi(r_i; \boldsymbol{\beta}, \phi, c) w(\boldsymbol{x}_i) (1/\sqrt{\phi v_{\mu_i}}) (\partial \mu_i / \partial \eta_i)}{b_i} \right. \\
&\qquad \left. - \frac{E[\psi(r_i; \boldsymbol{\beta}, \phi, c)] w(\boldsymbol{x}_i) (1/\sqrt{\phi v_{\mu_i}}) (\partial \mu_i / \partial \eta_i)}{b_i} \right\} b_i x_{ij} \\
&\quad = \sum_{i=1}^{n} \left\{ \boldsymbol{x}_i \boldsymbol{\beta}^{t-1} + \frac{\psi(r_i; \boldsymbol{\beta}, \phi, c) - E[\psi(r_i; \boldsymbol{\beta}, \phi, c)]}{E[\psi(r_i; \boldsymbol{\beta}, \phi, c)(\partial / \partial \mu_i) \log h(y_i \mid \boldsymbol{x}_i, \mu_i)]} \frac{\partial \eta_i}{\partial \mu_i} \right\} b_i x_{ij} \\
&\quad = \sum_{i=1}^{n} Z_i b_i x_{ij} = [\mathbf{X}^T B \mathbf{Z}]_j,
\end{aligned}
$$

where the involved quantities are evaluated at $\boldsymbol{\beta}^{t-1}$.

F

Computations for the Robust GEE Estimator

F.1 IRWLS Algorithm for Robust GEE

The whole robust procedure consists of solving the three following sets of equations:

$$\sum_{i=1}^{n}(D_{\boldsymbol{\mu}_i,\boldsymbol{\beta}})^T \boldsymbol{\Gamma}_i^T (V_{\boldsymbol{\mu}_i,\tau,\boldsymbol{\alpha}})^{-1}(\boldsymbol{\psi}_i - \boldsymbol{c}_i) = \sum_{i=1}^{n}\Psi_1(\boldsymbol{y}_i, \boldsymbol{X}_i; \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau, c) = 0 \qquad \text{(F.1)}$$

$$\sum_{i=1}^{n}\sum_{t=1}^{n_i}\chi(r_{it}; \boldsymbol{\beta}, \boldsymbol{\alpha}, \phi, c) = \sum_{i=1}^{n}\Psi_2(\boldsymbol{r}_i; \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau, c) = 0 \qquad \text{(F.2)}$$

$$\sum_{i=1}^{n}\left(\boldsymbol{G}_i^{\mathrm{T}}\boldsymbol{B}_i - \frac{K}{n}\boldsymbol{\alpha}\tau\right) = \sum_{i=1}^{n}\Psi_3(\boldsymbol{r}_i; \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau, c) = 0. \qquad \text{(F.3)}$$

Ideally these equations should be solved simultaneously (as, for example, in Huggins (1993)). We implement a two-stage approach iterating between the estimation of the regression parameters via (F.1) and the estimation of the dispersion and correlation parameters via (F.2) and (F.3). In fact, for fixed values of the nuisance parameters τ and $\boldsymbol{\alpha}$, the estimation of the regression parameter $\boldsymbol{\beta}$ can be performed via an IRWLS algorithm by regressing the adjusted dependent variable

$$\mathbf{Z} = \mathbf{X}_{\text{tot}}\hat{\boldsymbol{\beta}} + D^*\boldsymbol{\Gamma}^{-1}(\boldsymbol{\psi}_{\text{tot}} - \boldsymbol{c}_{\text{tot}})$$

on $\mathbf{X}_{\text{tot}}$ with a block-diagonal weight matrix $\mathbf{W}^*$, where $\mathbf{X}_{\text{tot}} = (\mathbf{X}_1^T, \ldots, \mathbf{X}_n^T)^T$, $\boldsymbol{\psi}_{\text{tot}} = (\boldsymbol{\psi}_1^T, \ldots, \boldsymbol{\psi}_n^T)^T$, $\boldsymbol{c}_{\text{tot}} = (\boldsymbol{c}_1^T, \ldots, \boldsymbol{c}_n^T)^T$ are the combined informations for the entire sample. The ith block of $\mathbf{W}^*$ is the $n_i \times n_i$ matrix

$$\mathbf{W}_i^* = {D^*_{\boldsymbol{\mu}_i,\boldsymbol{\beta}}}^{-1}\boldsymbol{\Gamma}_i^{\mathrm{T}}(A_{\boldsymbol{\mu}_i})^{-1/2}(R_{\boldsymbol{\alpha},i})^{-1}(A_{\boldsymbol{\mu}_i})^{-1/2}\boldsymbol{\Gamma}_i {D^*_{\boldsymbol{\mu}_i,\boldsymbol{\beta}}}^{-1},$$

and D^* is a block-diagonal matrix with blocks $D^*_{\boldsymbol{\mu}_i,\boldsymbol{\beta}} = \text{diag}(\partial\eta_{i1}/\partial\mu_{i1}, \ldots, \partial\eta_{in_i}/\partial\mu_{in_i})$. We remark that $D_{\boldsymbol{\mu}_i,\boldsymbol{\beta}} = D^{*-1}_{\boldsymbol{\mu}_i,\boldsymbol{\beta}}\mathbf{X}_i$. The matrix $\boldsymbol{H}_i = \mathbf{X}_i(\mathbf{X}^T\mathbf{W}^*\mathbf{X})^{-1}\mathbf{X}_i^T\mathbf{W}_i^*$ defines the hat matrix for subject i. One then obtains an estimate of τ and next an estimate of $\boldsymbol{\alpha}$ from (F.2) and (F.3), respectively. Note that (F.3) can be solved explicitly when exchangeable correlation is assumed, yielding $\hat{\alpha} = 1/(\hat{\tau}K)\sum_{i=1}^n \boldsymbol{G}_i^T\boldsymbol{B}_i$.

F.2 Fisher Consistency Corrections

Let Y_{it} and $Y_{it'}$ be Bernoulli distributed with probability of success equal to μ_{it} and $\mu_{it'}$, respectively, and with correlation $\rho_{tt'}$. We assume that the robustness weight w_{it} associated with subject i at time t can be decomposed as $w(\boldsymbol{x}_{it})w(r_{it};\boldsymbol{\beta},\tau,c)$. The joint distribution of $(y_{it}, y_{it'})$ is multinomial with set of probabilities $(\pi_{11}, \pi_{10}, \pi_{01}, \pi_{00})$, where $\pi_{11} = \rho_{tt'}v_{it}^{1/2}v_{it'}^{1/2} + \mu_{it}\mu_{it'}$, $\pi_{10} = \mu_{it} - \pi_{11}$, $\pi_{01} = \mu_{is} - \pi_{11}$ and $\pi_{00} = 1 - \mu_{it} - \mu_{is} + \pi_{11}$.

The consistency correction vector $\boldsymbol{c}_i$ has elements $c_{it} = E[\psi_{it}]$ that takes the form:

$$c_{it} = w(r_{it}^{(1)};\boldsymbol{\beta},\tau,c)(w(r_{it}^{(0)};\boldsymbol{\beta},\tau,c) - w(r_{it}^{(0)};\boldsymbol{\beta},\tau,c))v(\mu_{it}),$$

where $w(r_{it}^{(j)};\boldsymbol{\beta},\tau,c) = w((j-\mu_{it})/v(\mu_{it})/\sqrt{\tau})$ is the weight for the tth measure of cluster i evaluated at $y_{it} = j$.

Moreover, the diagonal matrix $\boldsymbol{\Gamma}_i = E[\tilde{\boldsymbol{\psi}}_i - \tilde{\boldsymbol{c}}_i]$, with $\tilde{\boldsymbol{\psi}}_i = \partial\boldsymbol{\psi}_i/\partial\boldsymbol{\mu}_i$ and $\tilde{\boldsymbol{c}}_i = \partial\boldsymbol{c}_i/\partial\boldsymbol{\mu}_i$, has diagonal elements

$$\Gamma_{it} = -w(\boldsymbol{x}_{it})((1-\mu_{it})w(r_{it}^{(1)};\boldsymbol{\beta},\tau,c) + \mu_{it}w(r_{it}^{(0)};\boldsymbol{\beta},\tau,c)).$$

G

Computation of the *CRQ*

The global algorithm uses the notation and definitions introduced in Section 7.6.2. It is taken from Portnoy (2003) or Debruyne *et al.* (2008) and works as follows.

- As long as no censored observations are crossed use ordinary QR as in Koenker and Bassett (1978).
- When the ith censored observation is crossed at the τth quantile store this value as $\hat{\tau}_i = \tau$.
- When censored observations have been crossed for a specific τ, find the value in $\boldsymbol{\beta}$ that minimizes a weighted version of (7.30):

$$\sum_{i \in K_\tau^c} \rho(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}(\tau); \tau) + \sum_{i \in K_\tau} [w_i(\tau)\rho(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}(\tau); \tau) + (1 - w_i(\tau)\rho(y^* - \boldsymbol{x}_i^T \boldsymbol{\beta}(\tau); \tau)], \tag{G.1}$$

  where K_τ represents the set of crossed and censored observations at τ and K_τ^c its complementary. The weights $w_i(\tau)$ are defined in Section 7.6.2 and y^* is any value sufficiently large to exceed $\boldsymbol{x}_i^T \boldsymbol{\beta}$ for all i.

To compute the regression quantile objective function (G.1) in practice, a sequence of breakpoints $\tau_1^*, \tau_2^*, \ldots, \tau_L^*$ is defined so that $\hat{\boldsymbol{\beta}}(\tau)$ is piecewise constant between these breakpoints. Then, simplex pivoting techniques allow to move from one breakpoint to another using the gradients of (G.1). Portnoy (2003) points out that the resulting gradients are linear in τ making the whole thing tractable. The above reference contains a detailed algorithm and additional explanations. Recently a variant of this called the grid algorithm has been proposed by Neocleous and Portnoy (2006). It is more stable, faster and has already been implemented in the R package

provided by Portnoy. It should be preferably used for large datasets. The simplex pivoting algorithm is still available and works well for smaller samples, that is, n up to several thousands.

References

Adrover, J., Salibian-Barrera, M. and Zamar, R. (2004) Globally robust inference for the location and simple regression model. *Journal of Statistical Planning and Inference*, **119**, 353–375.

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. *Proceedings of the Second International Symposium on Information Theory* (eds Petrov, B.N. and Csaki, F.), Akademiai Kiado, Budapest, pp. 267–281.

Alario, F.J.S. and Ferrand, L. (2000) Semantic and associative priming in picture naming. *The Quarterly Journal of Experimental Psychology*, **53**, 741–764.

Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H. and Tukey, J.W. (1972) *Robust Estimates of Location: Survey and Advances*, Princeton University Press, Princeton, NJ.

Atkinson, A.C. (1985) *Plots, Transformations and Regression*, Oxford University Press, Oxford.

Atkinson, A.C. and Riani, M. (2000) *Robust Diagnostic Regression Analysis*, Springer, Berlin.

Barnett, V. and Lewis, T. (1978) *Outliers in Statistical Data*, John Wiley & Sons, New York.

Barry, S. and Welsh, A. (2002) Generalized additive modelling and zero inflated count data. *Ecological Modelling*, **157**, 179–188.

Beaton, A.E. and Tukey, J.W. (1974) The fitting of power series, meaning polynomials, illustrated on band-spectroscopicdata. *Technometrics*, **16**, 147–185.

Bednarski, T. (1993) Robust estimation in Cox's regression model. *Scandinavian Journal of Statistics*, **20**, 213–225.

Bednarski, T. (1999) Adaptive robust estimation in the Cox regression model. *Biocybernetics and Biomedical Engineering*, **19**, 5–15.

Bednarski, T. (2007) On a robust modification of Breslow's cumulated hazard estimator. *Computational Statistics and Data Analysis*, **52**, 234–238.

Bednarski, T. and Mocarska, E. (2006) On robust model selection within the Cox model. *Econometrics Journal*, **9**, 279–290.

Bednarski, T. and Nowak, M. (2003) Robustness and efficiency of Sasieni-type estimators in the Cox model. *Journal of Statistical Planning and Inference*, **115**, 261–272.

Bednarski, T. and Zontek, S. (1996) Robust estimation of parameters in a mixed unbalanced model. *Annals of Statistics*, **24**, 1493–1510.

Belsley, D.A., Kuh, E. and Welsch, R.E. (1980) *Regression Diagnostics*, John Wiley & Sons, New York.

Bennet, C.A. (1954) Effect on measurement error on chemical process control. *Industrial Quality Control*, **11**, 17–20.

Beran, R. (1981) Efficient robust tests in parametric models. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **57**, 73–86.

Berkson, J. (1944) Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, **39**, 357–365.

Bernoulli, D. (1777) Dijudicatio maxime probabilis plurium observationum discrepantium atque verisimillima inductio inde formanda. *Acta Acad, Sci. Petropolit.*, **1**, 3–33 (English translation by Allen, C.C. (1961), *Biometrika*, **48**, 3–13.)

Berry, D.A. (1987) Logarithmic transformations in ANOVA. *Biometrics*, **43**, 439–456.

Bianco, A., Boente, G. and di Rienzo, J. (2000) Some results for robust GM-based estimators in heteroscedastic regression models. *Journal of Statistical Planning and Inference*, **89**, 215–242.

Bianco, A.M. and Yohai, V.J. (1997) Robust estimation in the logistic regression model. *Robust Statistics, Data Analysis and Computer Intensive Methods* (ed. Rieder H), Springer, New York, pp. 17–34.

Bianco, A.M., Ben, M.G. and Yohai, V.J. (2005) Robust estimation for linear regression with asymmetric errors. *The Canadian Journal of Statistics*, **33**, 511–528.

Birch, M.W. (1963) Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society, Series B, Methodological*, **25**, 220–233.

Bliss, C.I. (1935) The calculation of the dosage-mortality curve. *Annals of Applied Biology*, **22**, 134–167.

Bloomfield, P. and Steiger, W.L. (1983) *Least Absolute Deviations: Theory, Applications, and Algorithms*, Birkhäuser, Boston, MA.

Blough, D.K., Madden, C.W. and Hornbrook, M.C. (1999) Modeling risk using generalized linear models. *Journal of Health Economics*, **18**, 153–171.

Box, G. (1979) Robustness in the strategy of scientific model building. *Robustness in Statistics* (ed. Launer, R. and Wilkinson, G.), Academic Press, New York.

Box, G.E.P. (1953) Non-normality and tests of variances. *Biometrika*, **40**, 318–335.

Bretagnolle, J. and Huber-Carol, C. (1988) Effects of omitting covariates in Cox's model for survival data. *Scandinavian Journal of Statistics*, **15**, 125–128.

Brochner-Mortensen, J., Jensen, S. and Rodbro, P. (1977) Assessment of renal function from plasma creatinine in adult patients. *Scandinavian Journal of Urology and Nephrology*, **11**, 263–270.

Buchinsky, M. and Hahn, J. (1998) An alternative estimator for the censored quantile regression model. *Econometrica*, **66**, 653–671.

Cain, K. and Lange, T. (1984) Approximate case influence for the proportional hazards regression model with censored data. *Biometrics*, **40**, 439–499.

Cameron, A.C. and Trivedi, P.K. (1998) *Regression Analysis of Count Data*, Cambridge University Press, Cambridge.

Canario, L. (2006) Genetic aspects of piglet mortality at birth and in early suckling period: relationships with sow maternal abilities and piglet vitality, PhD thesis, Institut National Agronomique Paris-Grignon, France.

Canario, L., Cantoni, E., Le Bihan, E., Caritez, J., Billon, Y., Bidanel, J. and Foulley, J. (2006) Between breed variability of stillbirth and relationships with sow and piglet characteristics. *Journal of Animal Science*, **84**, 3185–3196.

Cantoni, E. (2003) Robust inference based on quasi-likelihoods for generalized linear models and longitudinal data. *Developments in Robust Statistics. Proceedings of ICORS 2001* (eds. Dutter, R., Filzmoser, P., Gather, U. and Rousseeuw, P.J.), Springer, Heidelberg, pp. 114–124.

Cantoni, E. (2004a) Analysis of robust quasi-deviances for generalized linear models. *Journal of Statistical Software*., Vol. 10, Issue 4.

Cantoni, E. (2004b) A robust approach to longitudinal data analysis. *Canadian Journal of Statistics*, **32**, 169–180.

Cantoni, E. and Ronchetti, E. (2001a) Resistant selection of the smoothing parameter for smoothing splines. *Statistics and Computing*, **11**, 141–146.

Cantoni, E. and Ronchetti, E. (2001b) Robust inference for generalized linear models. *Journal of the American Statistical Association*, **96**, 1022–1030.

Cantoni, E. and Ronchetti, E. (2006) A robust approach for skewed and heavy-tailed outcomes in the analysis of health care expenditures. *Journal of Health Economics*, **25**, 198–213.

Cantoni, E., Mills Flemming, J. and Ronchetti, E. (2005) Variable selection for marginal longitudinal generalized linear models. *Biometrics*, **61**, 507–514.

Carroll, R., Ruppert, D. and Stefanski, L. (1995) *Measurement Error in Nonlinear Models*, Chapman & Hall, London.

Carroll, R.J. and Pederson, S. (1993) On robustness in the logistic regression model. *Journal of the Royal Statistical Society, Series B, Methodological*, **55**, 693–706.

Carroll, R.J. and Ruppert, D. (1982) Robust estimation in heteroscedastic linear models. *Annals of Statistics*, **10**, 1224–1233.

Chatterjee, S. and Hadi, A.S. (1988) *Sensitivity Analysis in Linear Regression*, John Wiley & Sons, New York.

Chen, C. and Wang, P. (1991) Diagnostic plots in Cox's regression model. *Biometrics*. **47**, 841–850.

Chernozhukov, V. and Hong, H. (2002) Three-step censored quantile regression and extramarital affairs. *Journal of the American Statistical Association*, **97**, 872–882.

Christmann, A. (1997) High breakdown point estimators in logistic regression. *Robust Statistics, Data Analysis and Computer Intensive Methods* (ed. Rieder H), Springer, New York, pp. 79–90.

Christmann, A. and Rousseeuw, P.J. (2001) Measuring overlap in binary regression. *Computational Statistics and Data Analysis*, **37**, 65–75.

Cleveland, W.S. (1979) Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, **74**, 829–836.

Collett, D. (2003a) *Modelling Binary Data*, Chapman & Hall, London.

Collett, D. (2003b) *Modelling Survival Data in Medical Research*, 2nd edn, Chapman & Hall, London.

Conen, D., Wietlisbach, V., Bovet, P., Shamlaye, C., Riesen, W., Paccaud, F. and Burnier, M. (2004) Prevalence of hyperuricemia and relation of serum uric acid with cardiovascular risk factors in a developing country. *BMC Public Health*, http://www.biomedcentral.com/1471-2458/4/9.

Cook, R.D. and Weisberg, S. (1982) *Residuals and Influence in Regression*, Chapman & Hall, New York.

Copas, J.B. (1988) Binary regression models for contaminated data. *Journal of the Royal Statistical Society, Series B, Methodological*, **50**, 225–265.

Copt, S. and Heritier, S. (2007) Robust alternative to the F-test in mixed linear models based on *MM*-estimates. *Biometrics*, **63**, 1045–1052.

Copt, S. and Victoria-Feser, M.P. (2006) High breakdown inference for mixed linear models. *Journal of the American Statistical Association*, **101**(473), 292–300.

Copt, S. and Victoria-Feser, M.P. (2009) Robust predictions in mixed linear models, Technical report, University of Geneva.

Cox, D. (1972) Regression models and life tables. *Journal of the Royal Statistical Society, Series B, Methodological*, **34**, 187–220.

Cox, D.R. and Hinkley, D.V. (1992) *Theoretical Statistics*, Chapman & Hall, London.

Cressie, N. and Lahiri, S. (1993) The asymptotic distribution of REML estimators. *Journal of Multivariate Analysis*, **45**, 217–233.

Croux, C., Dhaene, G. and Hoorelbeke, D. (2003) Robust standard errors for robust estimators, *Discussion Paper Series 03.16*, Center for Economic Studies, Catholic University of Leuven.

Davies, P.L. (1987) Asymptotic behaviour of S-estimators of multivariate location parameters and dispertion matrices. *Annals of Statistics*, **15**, 1269–1292.

Davies, R.B. (1980) [Algorithm AS 155] The distribution of a linear combination of χ^2 random variables (AS R53: 84V33 p366- 369). *Applied Statistics*, **29**, 323–333.

Davison, A.C. and Hinkley, D.V. (1997) *Bootstrap Methods and their Applications*, Cambridge University Press, Cambridge.

Debruyne, M., Hubert, M., Portnoy, S. and Vanden Branden, K. (2008) Censored depth quantiles. *Computational Statistics and Data Analysis*, **52**, 1604–1614.

Dempster, A.P., Rubin, D.B. and Tsutakawa, R.K. (1981) Estimation in covariance components models. *Journal of the American Statistical Association*, **76**, 341–353.

Devlin, S.J., Gnanadesikan, R. and Kettenring, J.R. (1981) Robust estimation of dispersion matrices and principal components. *Journal of the American Statistical Association*, **76**, 354–362.

Diggle, P.J., Heagerty, P., Liang, K.Y. and Zeger, S.L. (2002) *Analysis of Longitudinal Data*, Oxford University Press, New York.

DiRienzo, A.G. and Lagakos, S.W. (2001) Effects of model misspecification on tests of no randomized treatment effect arising from Coxs proportional hazards model. *Journal of the Royal Statistical Society Series B, Methodological*, **63**, 745–757.

DiRienzo, A.G. and Lagakos, S.W. (2003) The effects of misspecifying Coxs regression model on randomized treatment group comparisons. *Handbook of Statistics*, **23**, 1–15.

Dobbie, M.J. and Welsh, A.H. (2001a) Modelling correlated zero-inflated count data. *Australian and New Zealand Journal of Statistics*, **43**(4), 431–444.

Dobbie, M.J. and Welsh, A.H. (2001b) Models for zero-inflated count data using the Neyman type A distribution. *Statistical Modelling*, **1**(1), 65–80.

Dobson, A.J. (2001) *An Introduction to Generalized Linear Models*, Chapman & Hall/CRC, Boca Raton, FL.

Dunlop, D.D., Manheim, L.M., Song, J. and Chang, R.W. (2002) Gender and ethnic/racial disparities health care utilization among older adults. *Journal of Gerontology*, **57B**, S221–S233.

Dupuis, D.J. and Morgenthaler, S. (2002) Robust weighted likelihood estimators with an application to bivariate extreme value problems. *Canadian Journal of Statistics*, **30**, 17–36.

Dyke, G.V. and Patterson, H.D. (1952) Analysis of factorial arrangements when the data are proportions. *Biometrics*, **8**, 1–12.

Edgeworth, F.Y. (1883) The method of least squares. *Philosophical Magazine*, **23**, 364–375.

Edgeworth, F.Y. (1887) On observations relating to several quantities. *Hermathena*, **6**, 279–285.

Efron, B. (1967) The power of the likelihood ratio test. *The Annals of Mathematical Statistics*, **38**, 802–806.

Efron, B. (1982) *The Jackknife, the Bootstrap an Other Resampling Plans*, vol. 38, Society for Industrial and Applied Mathematics, Philadelphia, PA.

Everitt, B.S. (1994) *Statistical Analysis using S-Plus*, Chapman & Hall, London.

Fahrmeir, L. and Tutz, G. (2001) *Multivariate Statistical Modelling Based on Generalized Linear Models*, Springer, Berlin.

Farebrother, R.W. (1990) [Algorithm AS 256] The distribution of a quadratic form in normal variables. *Applied Statistics*, **39**, 294–309.

Fernholz, L.T. (1983) *Von Mises Calculus for Statistical Functionals* (*Lecture Notes in Statistics*, vol. 19), Springer, New York.

Field, C. and Smith, B. (1994) Robust estimation—a weighted maximum likelihood approach. *International Statistical Review*, **62**, 405–424.

Fisher, R. (1925) *Statistical Methods for Research Workers*, 1st edn, Oliver and Boyd, Edinburgh.

Fisher, R.A. (1922) On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society*, **222**, 309–368.

Fisher, R.A. (1934) Two new properties of mathematical likelihood. *Philosophical Transactions of the Royal Society A*, **144**, 285–307.

Fitzenberger, B. and Winker, P. (2007) Improving the computation of censored quantile regressions. *Computational Statistics and Data Analysis*, **52**, 88–108.

Gail, M., Wieand, S. and Piantodosi, S. (1984) Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, **71**, 431–444.

Gallant, A.R. and Tauchen, G. (1996) Which moments to match? *Econometric Theory*, **12**, 657–681.

Genton, M.G. and Ronchetti, E. (2003) Robust indirect inference. *Journal of the American Statistical Association*, **98**(461), 67–76.

Genton, M.G. and Ronchetti, E. (2008) Robust prediction of beta, in *Computational Methods in Financial Engineering - Essays in Honour of Manfred Gilli* (eds Kontoghiorghes, E.J., Rustem, B. and Winker, P.), Springer, Berlin pp. 147–161.

Gerdtham, U. (1997) Equity in health care utilization: further tests based on hurdle models and Swedish micro data. *Health Economics*, **6**, 303–319.

Gilleskie, D.B. and Mroz, T.A. (2004) A flexible approach for estimating the effect of covariates on health expenditures. *Journal of Health Economics*, **23**, 391–418.

Giltinan, D.M., Carroll, R.J. and Ruppert, D. (1986) Some new estimation methods for weighted regression when there are possible outliers. *Technometrics*, **28**, 219–230.

Gouriéroux, C., Monfort, A. and Renault, E. (1993) Indirect inference. *Journal of Applied Econometrics*, **8S**, 85–118.

Greene, W. (1997) *Econometric Analysis*, 3rd edn, Prentice Hall, Englewood Cliffs, NJ.

Grzegorek, K. (1993) On robust estimation of baseline hazard under the Cox model via Fréchet differentiability, PhD thesis, Preprint of the Institute of Mathematics of the Polish Academy of Sciences, 518.

Hammill, B.G. and Preisser, J.S. (2006) A SAS/IML software program for GEE and regression diagnostic. *Computational Statistics and Data Analysis*, **51**, 1197–1212.

Hampel, F.R. (1968) Contribution to the theory of robust estimation, PhD thesis, University of California, Berkeley, CA.

Hampel, F.R. (1974) The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, **69**, 383–393.

Hampel, F.R. (1985) The breakdown points of the mean combined with some rejection rules. *Technometrics*, **27**, 95–107.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986) *Robust Statistics: The Approach Based on Influence Functions*, John Wiley & Sons, New York.

Hanfelt, J.J. and Liang, K.Y. (1995) Approximate likelihood ratios for general estimating functions. *Biometrika*, **82**, 461–477.

Hardin, J.W. and Hilbe, J.M. (2003) *Generalized Estimating Equations*, Chapman & Hall, London.

Härdle, W. (1990) *Applied Nonparametric Regression*, Cambridge University Press, Cambridge.

Harrell, F.E.J. (2001) *Regression Modeling Strategies: With Application to Linear Models, Logistic Regression and Survival Analysis* (*Springer Series in Statistics*), Springer, Berlin.

Harter, H.L. (1974–1976) The method of least squares and some alternatives. *Reviews of International Institute of Statistics*, **42**, 147–174 (Part I); **42**, 235–264 (Part II); **43**, 1–44 (Part III); **43**, 125–190 (Part IV); **43**, 269–278 (Part V); **44**, 113–159 (Part VI).

Harville, D.A. (1977) Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, **72**, 320–340.

Hastie, T.J. and Tibshirani, R.J. (1990) *Generalized Additive Models*, Chapman & Hall, London.

Hauck, W.W. and Donner, A. (1977) Wald's test as applied to hypotheses in logit analysis (Corr **75**, p. 482). *Journal of the American Statistical Association*, **72**, 851–853.

He, X. (1991) A local breakdown property of robust tests in linear regression. *Journal of Multivariate Analysis*, **38**, 294–305.

He, X., Simpson, D. and Portnoy, S. (1990) Breakdown robustness of tests. *Journal of the American Statistical Association*, **85**, 446–452.

Heagerty, P.J. and Zeger, S.L. (1996) Marginal regression models for clustered ordinal measurements. *Journal of the American Statistical Association*, **91**, 1024–1036.

Heagerty, P.J. and Zeger, S.L. (2000) Multivariate continuation ratio models: connections and caveats. *Biometrics*, **56**(3), 719–732.

Henderson, C.R. (1953) Estimation of variance and covariance components. *Biometrics*, **9**, 226–252.

Henderson, C.R., Kempthorne, O., Searle, S.R. and von Krosigk, C.N. (1959) Estimation of environmental and genetic trends from records subject to culling. *Biometrics*, **15**, 192–218.

Heritier, S. (1993) Contribution to robustness in nonlinear models: application to economic data, PhD thesis, Faculty of Economic and Social Sciences, University of Geneva Switzerland.

Heritier, S. and Galbraith, S. (2008) A revisit of robust inference in the Cox model, Technical report, University of New South Wales, Australia.

Heritier, S. and Ronchetti, E. (1994) Robust bounded-influence tests in general parametric models. *Journal of the American Statistical Association*, **89**(427), 897–904.

Heritier, S. and Victoria-Feser, M.P. (1997) Practical applications of bounded-influence tests, in *Handbook of Statistics*, vol. 15 (eds Maddala, G. and Rao, C.), Elsevier Science, Amsterdam, pp. 77–100.

Hettmansperger, T.P. (1984) *Statistical Inference Based on Ranks*, John Wiley & Sons, New York.

Hettmansperger, T.P. and McKean, J.W. (1998) *Robust Nonparametric Statistical Methods*, Arnold, London.

Hjort, N. (1992) On inference in parametric survival models. *International Statistical Review*, **60**, 55–387.

Hodges, J.L.J. (1967) Efficiency in normal samples and tolerance of extreme values for some estimates of location, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, University of California Press, Berkeley, CA, pp. 163–186.

Holcomb, P. and McPherson, W. (1994) Event-related brain potentials reflect semantic priming in an object decision task. *Brain and Cognition*, **24**, 259–276.

Hollis, S. and Campbell, F. (1999) What is meant by intention to treat? survey of published randomised clinical trials. *British Medical Journal*, **319**, 670–674.

Honore, B., Khan, S. and Powell, J.L. (2002) Quantile regression under random censoring. *Journal of Econometrics*, **109**, 67–105.

Horton, N.J. and Lipsitz, S.R. (1999) Review of software to fit generalized estimating equation regression models. *The American Statistician*, **53**, 160–169.

Huber-Carol, C. (1970) Etude Asymptotique de Tests Robustes, PhD thesis, ETH Zürich, Switzerland.

Huber, P.J. (1964) Robust estimation of a location parameter. *Annals of Mathematical Statistics*, **35**, 73–101.

Huber, P.J. (1967) The behavior of the maximum likelihood estimates under non standard conditions, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, University of California Press, Berkeley, CA, pp. 221–233.

Huber, P.J. (1972) Robust statistics: a review. *Annals of Mathematical Statistics*, **43**, 1041–1067.

Huber, P.J. (1973) Robust regression: Asymptotics, conjectures and Monte Carlo. *Annals of Statistics*, **1**, 799–821.

Huber, P.J. (1979) Robust smoothing, in *Robustness in Statistics* (eds Launer RL and Wilkinson GN), Academic Press, New York, pp. 33–48.

Huber, P.J. (1981) *Robust Statistics*, John Wiley & Sons, New York.

Huber, P.J. and Ronchetti, E.M. (2009) *Robust Statistics*, 2nd edn, John Wiley & Sons, New York.

Huggins, R.M. (1993) A robust approach to the analysis of repeated measures. *Biometrics*, **49**, 715–720.

Huggins, R.M. and Staudte, R.G. (1994) Variance components models for dependent cell populations. *Journal of the American Statistical Association*, **89**, 19–29.

Imhof, J.P. (1961) Computing the distribution of quadratic forms in normal variables. *Biometrika*, **48**, 352–363.

Ingelfinger, J.A., Mosteller, F., Thibodeau, L.A. and Ware, J.H. (1987) *Biostatistics in Clinical Medicine*, 2nd edn, McMillan, New York.

Jain, A., Tindell, C.A., Laux, I., Hunter, J.B., Curran, J., Galkin, A., Afar, D.E., Aronson, N., Shak, S., Natale, R.B. and Agus, D.B. (2005) Epithelial membrane protein-1 is a biomarker of gefitinib resistance. *Proceedings of the National Academy of Science USA*, **102**, 11858–11863.

Kalbfleisch, J. and Prentice, R. (1980) *The Statistical Analysis of Failure Time Data*, John Wiley & Sons, Ltd., Chichester.

Kenward, M.G. and Roger, J.H. (1997) Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, **53**, 983–997.

Kim, C. and Bae, W. (2005) Case influence diagnostics in the Kaplan–Meier estimator and the log-rank test. *Computational Statistics*, **20**, 521–534.

Koenker, R. (2005) *Quantile Regression (Econometric Society Monographs)*, Cambridge University Press, New York.

Koenker, R. (2008) Censored quantile regression redux. *Journal of Statistical Software*, **27**(6), 2–25.

Koenker, R. and Bassett, G. (1978) Regression quantiles. *Econometrica*, **46**, 33–50.

Koenker, R. and Bassett, G. (1982) Robust test for heteroscedasticity based on regression quantiles. *Econometrica*, **50**, 43–62.

Koenker, R. and D'Orey, V. (1987) Computing regression quantiles. *Applied Statistics*, **36**, 383–393.

Koenker, R. and Geling, O. (2001) Reappraising medfly longevity: a quantile regression survival analysis. *Journal of the American Statistical Association*, **96**, 458–468.

Koenker, R. and Hallock, K. (2001) Quantile regression: an introduction. *Journal of Econometric Perspectives*, **15**, 143–156.

Kong, F.H. and Slud, E. (1997) Robust covariate-adjusted logrank tests. *Biometrika*, **84**, 847–862.

Krall, J.M., Uthoff, V.A. and Hareley, J.B. (1975) A step-up procedure for selecting variables associated with survival. *Biometrics*, **31**, 49–57.

Krasker, W.S. and Welsch, R.E. (1982) Efficient bounded-influence regression estimation. *Journal of the American Statistical Association*, **77**, 595–604.

Künsch, H.R., Stefanski, L.A. and Carroll, R.J. (1989) Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. *Journal of the American Statistical Association*, **84**, 460–466.

Kuonen, D. (1999) Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika*, **86**, 929–935.

Kurttio, P., Komulainen, H., Leino, A., Salonen, L., Auvinen, A. and Saha, H. (2005) Bone as a possible target of chemical toxicity of natural uranium in drinking water. *Environmental Health Perspectives*, **113**, 68–72.

Laird, N. and Ware, J. (1982) Random-effect models for longitudinal data. *Biometrics*, **38**, 963–974.

Lambert, D. (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1–14.

Lange, K.L., Little, R.J.A. and Taylor, J.M.G. (1989) Robust statistical modeling using the t-distribution. *Journal of the American Statistical Association*, **84**, 881–896.

Liang, K.Y. and Zeger, S.L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Liang, K.Y., Zeger, S.L. and Qaqish, B. (1992) Multivariate regression analyses for categorical data (Discussion: pp. 24–40). *Journal of the Royal Statistical Society, Series B, Methodological*, **54**, 3–24.

Lin, D.Y. and Wei, L.J. (1989) The robust inference for the Cox proportional hazard model. *Journal of the American Statistical Association*, **84**, 1074–1078.

Lin, T.I. and Lee, J.C. (2006) A robust approach to t linear mixed models applied to multiple sclerosis data. *Statistics in Medicine*, **25**, 1397–1412.

Lindsey, J.K. (1997) *Applying Generalized Linear Models*, Springer, Berlin.

Lindstrom, M.J. and Bates, A. (1988) Newton–Raphson and EM algorithms for linear mixed-effects models for repeated data (correction: **94**(89), 1572). *Journal of the American Statistical Association*, **83**, 1014–1022.

Litière, S., Alonso, A. and Molenberghs, G. (2007a) The impact of misspecified random-effect distribution on the estimation and performance of inferential procedures in generalized linear mixed models. *Statistics in Medicine*, **27**, 3125–3144.

Litière, S., Alonso, A. and Molenberghs, G. (2007b) Type I and type II error under random-effects misspecification in generalized linear mixed models. *Biometrics*, **63**, 1038–1044.

Littell, R.C. (2002) Analysis of unbalanced mixed model data: a case study comparison of ANOVA versus REML/GLS. *Journal of Agricultural, Biological and Environmental Statistics*, **7**, 472–490.

Lopuhaä, H.P. (1989) On the relation between S-estimators and M-estimators of multivariate location and covariance. *Annals of Statistics*, **17**, 1662–1683.

Lopuhaä, H.P. (1992) High efficient estimators of multivariate location with high breakdown point. *Annals of Statistics*, **20**, 398–413.

Lopuhaä, H.P. and Rousseeuw, P.J. (1991) Breakdown points of affine equivariant estimators of multivariate locution and covariance matrices. *Annals of Statistics*. **19**, 229–248.

Ma, B. and Elis, R.E. (2003) Robust registration for computer-integrated orthopedic surgery: laboratory validation and clinical experience. *Medical Image Analysis*, **7**(3), 237–250.

Machado, J.A.F. and Machado, J.A.F. (1993) Robust model selection and m-estimation. *Econometric Theory*, **9**, 478–493.

Mahalanobis, P.C. (1936) On the generalized distance in statistics. *Proceedings of the National Institute of Science of India*, **12**, 49–55.

Mallows, C.L. (1973) Some comments on c_p. *Technometrics*, **15**, 661–675.

Mallows, C.L. (1975) On some topics in robustness, Technical report, Bell Telephone Laboratories, Murray Hill, NJ.

Marazzi, A. (1993) *Algorithms, Routines and S-Functions for Robust Statistics*, Wadsworth and Brooks/Cole, Belmont, CA.

Marazzi, A. (2002) Bootstrap tests for robust means of asymmetric distributions with unequal shapes. *Computational Statistics and Data Analysis*, **39**, 503–528.

Marazzi, A. and Barbati, G. (2003) Robust parametric means of asymmetric distributions: estimation and testing. *Estadistica*, **54**, 47–72.

Marazzi, A. and Yohai, V. (2004) Adaptively truncated maximum likelihood regression with asymmetric errors. *Journal of Statistical Planning and Inference*, **122**, 271–291.

Markatou, M. and He, X. (1994) Bounded influence and high breakdown point testing procedures in linear models. *Journal of the American Statistical Association*, **89**, 543–549.

Markatou, M. and Hettmansperger, T.P. (1990) Robust bounded influence tests in linear models. *Journal of the American Statistical Association*, **85**, 187–190.

Markatou, M. and Hettmansperger, T.P. (1992) Applications of the asymmetric eigen value problem techniques to robust testing. *Journal of Statistical Planning and Inference*, **31**, 51–65.

Markatou, M. and Ronchetti, E. (1997) Robust inference: The approach based on influence functions, in *Handbook of Statistics, Vol. 15: Robust Inference* (eds Maddala, G. S. andRao C), Elsevier Science, New York, pp. 49–75.

Markatou, M., Basu, A. and Lindsay, B. (1997) Weighted likelihood estimating equations: the discrete case with application to logistic regression. *Journal of Statistical Planning and Inference*, **57**, 215–232.

Markatou, M., Stahel, W.A. and Ronchetti, E. (1991) Robust M-type testing procedures for linear models, in *Directions in Robust Statistics and Diagnostics, Part I* (eds Stahel WA and Weisberg S), Springer, New York, pp. 201–220.

Maronna, R.A. (1976) Robust *M*-estimators of multivariate location and scatter. *Annals of Statistics* **4**, 51–67.

Maronna, R.A. and Yohai, V.J. (2000) Robust regression with both continuous and categorical predictors. *Journal of Statistical Planning and Inference*, **89**, 197–214.

Maronna, R.A., Bustos, O.H. and Yohai, V.J. (1979) Bias- and efficiency-robustness of general *M*-estimators for regression with random carriers, in *Smoothing Techniques for Curve Estimation* (eds Gasser, T. and Rosenblatt, M.), Springer, New York, pp. 91–116.

Maronna, R.A., Martin, R.D. and Yohai, V.J. (2006) *Robust Statistics: Theory and Methods*, John Wiley & Sons, Ltd, Chichester.

McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models*, 2nd edn, Chapman & Hall, London.

McCulloch, C.E. and Searle, S.R. (2001) *Generalized, Linear, and Mixed Models*, John Wiley & Sons, Ltd, Chichester.

McLean, R.A., Sanders, W.L. and Stroup, W.W. (1991) A unified approach to mixed models. *The American Statistician*, **45**, 54–64.

Mills, J.E., Field, C.A. and Dupuis, D.J. (2002) Marginally specified generalized linear mixed models: a robust approach. *Biometrics*, **58**, 727–734.

Min, Y. and Agresti, A. (2002) Modeling nonnegative data with clumping at zero: a survey. *Journal of the Iranian Statistical Society*, **1**, 7–33.

Minder, C.E. and Bednarski, T. (1996) A robust method for proportional hazards regression. *Statistics in Medicine*, **15**, 1033–1047.

Molenberghs, G. and Verbeke, G. (2005) *Models for Discrete Longitudinal Data*, Springer, Berlin.

Morgenthaler, S. (1992) Least-absolute-deviations fits for generalized linear models. *Biometrika*, **79**, 747–754.

Morrell, C.H. (1998) Likelihood ratio testing of variance components in the linear mixed-effects model using restricted maximum likelihood. *Biometrics*, **54**, 1560–1568.

Moustaki, I. and Victoria-Feser, M.P. (2006) Bounded-influence robust estimation in generalized linear latent variable models. *Journal of the American Statistical Association*, **101**(474), 644–653.

Moustaki, I., Victoria-Feser, M.P. and Hyams, H. (1998) A UK study on the effect of socioeconomic background of pregnat women and hospital practice on the decision to breastfeed and the initiation and duration of breastfeeding, *Technical Report Statistics Research Report LSERR44*, London School of Economics, London.

Moy, G. and Mounoud, P. (2003) Object recognition in young adult: is priming with pantomimes possible?, in *Catalogue des abstracts : 8ème congrès de la société suisse de psychologie (SSP)*, Bern, Switzerland.

Mullahy, J. (1986) Specification and testing of some modified count data models. *Journal of Econometrics*, **33**, 341–365.

Müller, S. and Welsh, A.H. (2005) Outlier robust model selection in linear regression. *Journal of the American Statistical Association*, **100**, 1297–1310.

Nardi, A. and Schemper, M. (1999) New residuals for Cox regression and their application to outlier screening. *Biometrics*, **55**, 523–529.

Nelder, J.A. (1966) Inverse polynomials, a useful group of multi-factor response functions. *Biometrics*, **22**, 128–141.

Nelder, J.A. and Wedderburn, R.W.M. (1972) Generalized linear models. *Journal of the Royal Statistical Society—Series A*, **135**, 370–384.

Neocleous, T. and Portnoy, S. (2006) A partly linear model for censored regression quantiles, Technical report, Statistics Department, University of Illinois, IL, USA.

Neocleous, T., Vanden Branden, K. and Portnoy, S. (2006) Correction to censored regression quantiles by S. Portnoy, **98** (2003), 1001–1012, *Journal of the American Statistical Association*, **101**, 860–861.

Newcomb, S. (1886) A generalized theory of the combinations of observations so as to obtain the best result. *American Journal of Mathematics*, **8**, 343–366.

Noh, M. and Lee, Y. (2007) Robust modeling for inference from generalized linear model classes. *Journal of the American Statistical Association*, **102**(479), 1059–1072.

Pan, W. (2001) Akaike's information criterion in generalized estimating equations. *Biometrics*, **57**(1), 120–125.

Patterson, H.D. and Thompson, R. (1971) Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545–554.

Pearson, E.S. (1931) The analysis of variance in cases of non-normal variation. *Biometrika*, **23**, 114–133.

Pearson, K. (1916) Second supplement to a memoir on skew variation. *Philosophical Transactions A*, **216**, 429–457.

Pena, D. and Yohai, V. (1999) A fast procedure for outlier diagnostics in large regression problems. *Journal of the American Statistical Association*, **94**, 434–445.

Peng, R. and Huang, Y. (2008) Survival analysis with quantile regression models. *Journal of American Statistical Association*, **103**, 637–649.

Pinheiro, J.C. and Bates, D.M. (2000) *Mixed-Effects Models in S and S-PLUS*, Springer, New York.

Pinheiro, J.C., Liu, C. and Wu, Y.N. (2001) Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics*, **10**(2), 249–276.

Portnoy, S. (2003) Censored regression quantiles. *Journal of the American Statistical Association*, **98**, 1001–1012.

Potthoff, R.F. and Roy, S.N. (1964) A generalized multivariate analysis of variance model useful especially for growth curve problem. *Biometrika*, **51**, 313–326.

Powell, J.L. (1986) Censored regression quantiles. *Journal of Econometrics*, **32**, 143–155.

Pregibon, D. (1982) Resistant fits for some commonly used logistic models with medical applications. *Biometrics*, **38**, 485–498.

Preisser, J.S. and Qaqish, B.F. (1999) Robust regression for clustered data with applications to binary regression. *Biometrics*, **55**, 574–579.

Preisser, J.S., Galecki, A.T., Lohman, K.K. and Wagenknecht, L.E. (2000) Analysis of smoking trends with incomplete longitudinal binary responses. *Journal of the American Statistical Association*, **95**, 1021–1031.

Prentice, R.L. (1988) Correlated binary regression with covariates specific to each binary observation. *Biometrics*, **44**, 1033–1048.

Qu, A. and Song, P.X.K. (2004) Assessing robustness of generalised estimating equations and quadratic inference functions. *Biometrika*, **91**, 447–459.

Qu, A., Lindsay, B.G. and Li, B. (2000) Improving generalised estimating equations using quadratic inference functions. *Biometrika*, **87**(4), 823–836.

Rao, C.R. (1973) *Linear Statistical Inference and its Application*, John Wiley & Sons, New York.

Rasch, G. (1960) *Probabilistic Models for some Intelligence and Attainment Tests*, Danmarks Paedagogiske Institut, Copenhagen.

Reid, N. and Crépeau, H. (1985) Influence functions for proportional hazards regression. *Biometrika*, **72**, 1–9.

Reimann, C., Filzmoser, P., Garrett, R. and Dutter, R. (2008) *Statistical Data Analysis Explained*, John Wiley & Sons, Ltd, Chichester.

Renaud, O. and Victoria-Feser, M.P. (2009) Robust coefficient of determination, Technical report, University of Geneva.

Richardson, A.M. (1997) Bounded influence estimation in the mixed linear model. *Journal of the American Statistical Association*, **92**, 154–161.

Richardson, A.M. and Welsh, A.H. (1994) Asymptotic properties of the restricted maximum likelihood for hierarchical mixed models. *Australian Journal of Statistics*, **36**, 31–43.

Richardson, A.M. and Welsh, A.H. (1995) Robust restricted maximum likelihood in mixed linear models. *Biometrics*, **51**, 1429–1439.

Ridout, M., Demétrio, C.G.B. and Hinde, J. (1998) Models for count data with many zeros, in *Proceedings of the 19th International Biometrics Conference*, Cape Town pp. 179–190.

Rieder, H. (1978) A robust asymptotic testing model. *Annals of Statistics*, **6**, 1080–1094.

Rocke, D.M. (1996) Robustness properties of *S*-estimators of multivariate location and shape in high dimension. *Annals of Statistics*, **24**, 1327–1345.

Ronchetti, E. (1982a) Robust alternatives to the F-test for the linear model (STMA V24 1026), in *Probability and Statistical Inference* (eds Grossmann, W., Pflug, G.C. and Wertz, W.), Reidel, Dordrecht, pp. 329–342.

Ronchetti, E. (1982b) Robust Testing in Linear Models: The Infinitesimal Approach, PhD thesis, ETH, Zürich, Switzerland.

Ronchetti, E. (1997a) Robust influence by influence functions. *Journal of Statistical Planning and Inference*, **57**, 59–72.

Ronchetti, E. (1997b) Robustness aspects of model choice. *Statistica Sinica*, **7**, 327–338.

Ronchetti, E. (2006) Fréchet and robust statistics. *Journal de la Société Francaise de Statistique* **147**, 73–75. (Comment on 'Sur une limitation très générale de la dispersion de la médiane' by Maurice Fréchet.)

Ronchetti, E. and Staudte, R.G. (1994) A robust version of Mallows's C_p. *Journal of the American Statistical Association*, **89**, 550–559.

Ronchetti, E. and Trojani, F. (2001) Robust inference with GMM estimators. *Journal of Econometrics*, **101**(1), 37–69.

Ronchetti, E., Field, C. and Blanchard, W. (1997) Robust linear model selection by cross-validation. *Journal of the American Statistical Association*, **92**, 1017–1023.

Rousseeuw, P.J. (1984) Least median of squares regression. *Journal of the American Statistical Association*, **79**, 871–880.

Rousseeuw, P.J. and Leroy, A.M. (1987) *Robust Regression and Outlier Detection*, John Wiley & Sons, New York.

Rousseeuw, P.J. and Ronchetti, E. (1979) The influence curve for tests, *Research Report 21*, ETH Zürich, Switzerland.

Rousseeuw, P.J. and Ronchetti, E. (1981) Influence curves for general statistics. *Journal of Computational and Applied Mathematics*, **7**, 161–166.

Rousseeuw, P.J. and Van Driessen, K. (1999) A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, **41**, 212–223.

Rousseeuw, P.J. and Yohai, V.J. (1984) Robust regression by means of *S*-estimators, in *Robust and Nonlinear Time Series Analysis* (eds Franke JW, Hardle and Martin RD), Springer, New York, pp. 256–272.

Rule, A.D., Larson, T.S., Bergstralh, E.J., Slezak, J.M., Jacobsen, S.J. and Cosio, F.G. (2004) Using serum creatinine to estimate glomerular filtration rate: Accuracy in good health and in chronic kidney disease. *Annals of Internal Medicine*, **141**(12), 929–938.

Salibian-Barrera, M. and Zamar, R.H. (2002) Bootstrapping robust estimates of regression. *Annals of Statistics*, **30**, 556–582.

Sasieni, P.D. (1993a) Maximum weighted partial likelihood estimates in the Cox model. *Journal of the American Statistical Association*, **88**, 144–152.

Sasieni, P.D. (1993b) Some new estimators for Cox regression. *Annals of Statistics*, **21**, 1721–1759.

Satterthwaite, F.E. (1941) Synthesis of variance. *Psychometrika*, **6**, 309–316.

Scheipl, F., Greven, S. and Küchenhoff, H. (2008) Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Computational Statistics and Data Analysis*, **52**, 3283–3299.

Schemper, M. (1992) Cox analysis of survival data with non-proportional hazard functions. *The Statistician*, **41**, 455–465.

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Searle, S.R., Casella, G. and McCulloch, C.E. (1992) *Variance Components*, John Wiley and Sons, Ltd, Chichester.

Self, S.G. and Liang, K.Y. (1987) Asymptotic properties of the maximum likelihood estimators and likelihood tests under nonstandard conditions. *Journal of the American Statistical Association*, **82**, 605–610.

Silvapulle, M.J. (1992) Robust Wald-type tests of one sided hypothesis in the linear model. *Journal of the American Statistical Association*, **87**, 156–161.

Simpson, D.G., Ruppert, D. and Caroll, R.J. (1992) One-step *GM* estimates and stability of inferences in linear regression. *Journal of the American Statistical Association*, **87**, 439–450.

Sinha, S.K. (2004) Robust analysis of generalized linear mixed models. *Journal of the American Statistical Association*, **99**(466), 451–460.

Sommer, S. and Huggins, R.M. (1996) Variables selection using the Wald test and a robust C_p. *Applied Statistics*, **45**, 15–29.

Song, P.X.K. (2007) *Correlated Data Analysis: Modeling, Analytics, and Applications*, Springer, New York.

Stahel, W.A. and Welsh, A. (1997) Approaches to robust estimation in the simplest variance components model. *Journal of Statistical Planning and Inference*, **57**, 295–319.

Stahel, W.A. and Welsh, A.H. (1992) Robust estimation of variance components, *Research Report* 69, ETH, Zürich.

Staudte, R.G. and Sheather, S.J. (1990) *Robust Estimation and Testing*, John Wiley & Sons, New York.

Stefanski, L.A., Carroll, R.J. and Ruppert, D. (1986) Optimally bounded score functions for generalized linear models with applications to logistic regression. *Biometrika*, **73**, 413–424.

Stern, S.E. and Welsh, A.H. (1998) Likelihood inference for small variance components. *The Canadian Journal of Statistics*, **28**, 517–532.

Stigler, S.M. (1973) Simon Newcomb, Percy Daniell, and the history of robust estimation 1885–1920. *Journal of the American Statistical Association*, **68**, 872–879.

Stone, E.J. (1873) On the rejection of discordant observations. *Monthly Notices of the Royal Astronomical Society*, **34**, 9–15.

Stone, M. (1977) An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society, Series B, Methodological*, **39**, 44–47.

Stram, D.O. and Lee, J.W. (1994) Variance components testing in the longitudinal mixed effects model. *Biometrics*, **50**, 1171–1177.

Stram, D.O., Wei, L.J. and Ware, J.H. (1988) Analysis of repeated ordered categorical outcomes with possibly missing observations and time-dependent covariates. *Journal of the American Statistical Association*, **83**, 631–637.

Student, (1927) Errors of routine analysis. *Biometrika* **19**, 151–164.

Subrahmanian, K., Subrahmaniam, K. and Messeri, J.Y. (1975) On the robustness of some tests of significance in sampling from a normal population. *Journal of the American Statistical Association*, **70**, 435–438.

Tai, B.C., White, I.R., Gebski, V. and Machin, D. (2002) On the issue of 'multiple' first failures in competing risks analysis. *Statistics in Medicine*, **21**, 2243–2255.

Tashkin, D.P. *et al.* (2006) Cyclophosphamide versus placebo in scleroderma lung disease. *New England Journal of Medicine*, **354**(25), 2655–66.

Tatsuoka, K.S. and Tyler, D.E. (2000) On the uniqueness of S-functionals and M-functionals under nonelliptical distributions. *Annals of Statistics*, **28**(4), 1219–1243.

Therneau, T.M. and Grambsch, P.M. (2000) *Modeling Survival Data: Extending the Cox Model*, Springer, New York.

Tukey, J.W. (1960) A survey of sampling from contaminated distributions, in *Contributions to Probability and Statistics* (ed. Olkin, I.), Stanford University Press, Stanford, CA, pp. 448–485.

Tukey, J.W. (1970) *Exploratory Data Analysis*, Addison-Wesley, Reading, MA. (Mimeographed preliminary edition. Published in 1977.)

Valsecchi, M.G., Silvestri, D. and Sasieni, P. (1996) Evaluation of long-term survival: use of diagnostics and robust estimators with Cox's proportional hazards models. *Statistics in Medicine*, **15**, 2763–2780.

Verbeke, G. and Molenberghs, G. (1997) *Linear Mixed Model in Practice: a SAS-Oriented Approach* (*Lecture Notes in Statistics*, vol. 126), Springer, New York.

Verbeke, G. and Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data*, Springer, New York.

Victoria-Feser, M.P. (2002) Robust inference with binary data. *Psychometrica*, **67**, 21–32.

Victoria-Feser, M.P. (2007) De-biasing weighted MLE via indirect inference: The case of generalized linear latent variable models. *Revstat Statistical Journal*, **5**, 85–96.

von Mises, R. (1947) On the asymptotic distribution of differentiable statistical functions. *The Annals of Mathematical Statistics*, **18**, 309–348.

Wager, T.D., Keller, M.C., Lacey, S.C. and Jonides, J. (2003) Increased sensitivity in neuroimaging analyses using robust regression. *NeuroImage*, **26**, 99–113.

Wald, A. (1943) Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, **3**, 426–482.

Wang, H.M., Jones, M.P. and Storer, B.E. (2006) Comparison of case-deletion diagnostic methods for Cox regression. *Statistics in Medicine*, **25**, 669–683.

Wang, Y.G., Lin, X. and Zhu, M. (2005) Robust estimating functions and bias correction for longitudinal data analysis. *Biometrics*, **61**, 684–691.

Wedderburn, R.W.M. (1974) Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika*, **61**, 439–447.

Welsh, A. and Richardson, A. (1997) Approaches to the robust estimation of mixed models, in *Handbook of Statistics*, vol. 15 (eds Maddala G and Rao C), Elsevier Science, pp. 343–385.

Welsh, A.H. (1996) *Aspects of Statistical Inference* (*Wiley Series in Probability and Statistics*), John Wiley & Sons, New York.

Welsh, A.H. and Ronchetti, E. (1998) Bias-calibrated estimation from sample surveys containing outliers. *Journal of the Royal Statistical Society, Series B, Methodological*, **60**, 413–428.

Welsh, A.H. and Ronchetti, E. (2002) A journey in single steps: robust one-step m-estimation in linear regression. *Journal of Statistical Planning and Inference*, **103**(2), 287–310.

Welsh, A.H., Cunningham, R.B., Donnelly, C.F. and Lindenmayer, D.B. (1996) Modelling the abundance of rare species: statistical models for counts with extra zeros. *Ecological Modelling*, **88**, 297–308.

Wen, L., Parchman, M.L., Linn, W.D. and Lee, S. (2004) Association between self-monitoring of blood glucose and glycemic control in patients with type 2 diabetes mellitus. *American Journal of Health-System Pharmacy*, **61**, 2401–2405.

Whewell, W. (1837) *History of the Inductive Sciences, from the Earliest to the Present Time*, Parker, London.

Whewell, W. (1840) *Philosophy of the Inductive Sciences, Founded upon their History*, Parker, London.

Wilcox, R.R. (1997) *Introduction to Robust Estimation and Hypothesis Testing*, Academic Press, New York, London.

Wood, A.T.A. (1989) An F approximation to the distribution of a linear combination of chi-squared variables. *Communications in Statistics: Simulation and Computation*, **18**, 1439–1456.

Wood, A.T.A., Booth, J.G. and Butler, R.W. (1993) Saddlepoint approximations to the CDF of some statistics with nonnormal limit distributions. *Journal of the American Statistical Association*, **88**, 680–686.

Yau, K.K.W. and Kuk, A.Y.C. (2002) Robust estimation in generalized linear mixed models. *Journal of the Royal Statistical Society, Series B, Methodological*, **64**, 101–117.

Ylvisaker, D. (1977) Test resistance. *Journal of the American Statistical Association*, **72**, 551–556.

Yohai, V.J. (1987) High breakdown point and high efficiency robust estimates for regression. *Annals of Statistics*, **15**, 642–656.

Yohai, V.J. and Zamar, R.H. (1998) Optimal locally robust m-estimates of regression. *Journal of Statistical Planning and Inference*, **64**, 309–323.

Yohai, V.J., Stahel, W.A. and Zamar, R.H. (1991) A procedure for robust estimation and inference in linear regression, in *Directions in Robust Statistics and Diagnostics, part II* (eds Stahel WA and Weisberg S) (*The IMA Volumes in Mathematics and its Applications*, vol. 34), Springer, Berlin, pp. 365–374.

Zedini, A. (2007) Poisson hurdle model: Towards a robustified approach, Master's thesis, University of Geneva.

Zeger, S.L. and Liang, K.Y. (1986) Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42**, 121–130.

Zeger, S.L., Liang, K.Y. and Albert, P.S. (1988) Models for longitudinal data: a generalized estimating equation approach (Correction: **45**, 347). *Biometrics*, **44**, 1049–1060.

Zhang, J. (1996) The sample breakdown of tests. *Journal of Statistical Planning and Inference*, **52**, 161–181.

Zhao, L.P. and Prentice, R.L. (1990) Correlated binary regression using a quadratic exponential model. *Biometrika*, **77**, 642–648.

Index

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

The Wiley Series in Probability and Statistics is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches.

This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

ABRAHAM and LEDOLTER · Statistical Methods for Forecasting
AGRESTI · Analysis of Ordinal Categorical Data
AGRESTI · An Introduction to Categorical Data Analysis
AGRESTI · Categorical Data Analysis, Second Edition
ALTMAN, GILL and McDONALD · Numerical Issues in Statistical Computing for the Social Scientist
AMARATUNGA and CABRERA · Exploration and Analysis of DNA Microarray and Protein Array Data
ANDĚL · Mathematics of Chance
ANDERSON · An Introduction to Multivariate Statistical Analysis, Third Edition
*ANDERSON · The Statistical Analysis of Time Series
ANDERSON, AUQUIER, HAUCK, OAKES, VANDAELE and WEISBERG · Statistical Methods for Comparative Studies
ANDERSON and LOYNES · The Teaching of Practical Statistics
ARMITAGE and DAVID (editors) · Advances in Biometry
ARNOLD, BALAKRISHNAN and NAGARAJA · Records
*ARTHANARI and DODGE · Mathematical Programming in Statistics
*BAILEY · The Elements of Stochastic Processes with Applications to the Natural Sciences
BALAKRISHNAN and KOUTRAS · Runs and Scans with Applications
BALAKRISHNAN and NG · Precedence-Type Tests and Applications
BARNETT · Comparative Statistical Inference, Third Edition
BARNETT · Environmental Statistics: Methods & Applications
BARNETT and LEWIS · Outliers in Statistical Data, Third Edition
BARTOSZYNSKI and NIEWIADOMSKA-BUGAJ · Probability and Statistical Inference
BASILEVSKY · Statistical Factor Analysis and Related Methods: Theory and Applications
BASU and RIGDON · Statistical Methods for the Reliability of Repairable Systems
BATES and WATTS · Nonlinear Regression Analysis and Its Applications
BECHHOFER, SANTNER and GOLDSMAN · Design and Analysis of Experiments for Statistical Selection, Screening and Multiple Comparisons
BELSLEY · Conditioning Diagnostics: Collinearity and Weak Data in Regression
BELSLEY, KUH and WELSCH · Regression Diagnostics: Identifying Influential Data and Sources of Collinearity
BENDAT and PIERSOL · Random Data: Analysis and Measurement Procedures, Third Edition
BERNARDO and SMITH · Bayesian Theory
BERRY, CHALONER and GEWEKE · Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner
BHAT and MILLER · Elements of Applied Stochastic Processes, Third Edition

*Now available in a lower priced paperback edition in the Wiley Classics Library.

BHATTACHARYA and JOHNSON · Statistical Concepts and Methods
BHATTACHARYA and WAYMIRE · Stochastic Processes with Applications
BIEMER, GROVES, LYBERG, MATHIOWETZ and SUDMAN · Measurement Errors in Surveys
BILLINGSLEY · Convergence of Probability Measures, Second Edition
BILLINGSLEY · Probability and Measure, Third Edition
BIRKES and DODGE · Alternative Methods of Regression
BLISCHKE and MURTHY (editors) · Case Studies in Reliability and Maintenance
BLISCHKE and MURTHY · Reliability: Modeling, Prediction and Optimization
BLOOMFIELD · Fourier Analysis of Time Series: An Introduction, Second Edition
BOLLEN · Structural Equations with Latent Variables
BOLLEN and CURRAN · Latent Curve Models: A Structural Equation Perspective
BOROVKOV · Ergodicity and Stability of Stochastic Processes
BOSQ and BLANKE · Inference and Prediction in Large Dimensions
BOULEAU · Numerical Methods for Stochastic Processes
BOX · Bayesian Inference in Statistical Analysis
BOX · R. A. Fisher, the Life of a Scientist
BOX and DRAPER · Empirical Model-Building and Response Surfaces
*BOX and DRAPER · Evolutionary Operation: A Statistical Method for Process Improvement
BOX, HUNTER and HUNTER · Statistics for Experimenters: An Introduction to Design, Data Analysis and Model Building
BOX, HUNTER and HUNTER · Statistics for Experimenters: Design, Innovation and Discovery, Second Edition
BOX and LUCEÑO · Statistical Control by Monitoring and Feedback Adjustment
BRANDIMARTE · Numerical Methods in Finance: A MATLAB-Based Introduction
BROWN and HOLLANDER · Statistics: A Biomedical Introduction
BRUNNER, DOMHOF and LANGER · Nonparametric Analysis of Longitudinal Data in Factorial Experiments
BUCKLEW · Large Deviation Techniques in Decision, Simulation and Estimation
CAIROLI and DALANG · Sequential Stochastic Optimization
CASTILLO, HADI, BALAKRISHNAN and SARABIA · Extreme Value and Related Models with Applications in Engineering and Science
CHAN · Time Series: Applications to Finance
CHATTERJEE and HADI · Regression Analysis by Example, Fourth Edition
CHATTERJEE and HADI · Sensitivity Analysis in Linear Regression
CHATTERJEE and PRICE · Regression Analysis by Example, Third Edition
CHERNICK · Bootstrap Methods: A Practitioner's Guide
CHERNICK and FRIIS · Introductory Biostatistics for the Health Sciences
CHILÉS and DELFINER · Geostatistics: Modeling Spatial Uncertainty
CHOW and LIU · Design and Analysis of Clinical Trials: Concepts and Methodologies, Second Edition
CLARKE and DISNEY · Probability and Random Processes: A First Course with Applications, Second Edition
*COCHRAN and COX · Experimental Designs, Second Edition
CONGDON · Applied Bayesian Modelling
CONGDON · Bayesian Models for Categorical Data
CONGDON · Bayesian Statistical Modelling
CONGDON · Bayesian Statistical Modelling, Second Edition
CONOVER · Practical Nonparametric Statistics, Second Edition
COOK · Regression Graphics
COOK and WEISBERG · An Introduction to Regression Graphics
COOK and WEISBERG · Applied Regression Including Computing and Graphics
CORNELL · Experiments with Mixtures, Designs, Models and the Analysis of Mixture Data, Third Edition
COVER and THOMAS · Elements of Information Theory
COX · A Handbook of Introductory Statistical Methods
*COX · Planning of Experiments

*Now available in a lower priced paperback edition in the Wiley Classics Library.

CRESSIE · Statistics for Spatial Data, Revised Edition
CSÖRGÖ and HORVÁTH · Limit Theorems in Change Point Analysis
DANIEL · Applications of Statistics to Industrial Experimentation
DANIEL · Biostatistics: A Foundation for Analysis in the Health Sciences, Sixth Edition
*DANIEL · Fitting Equations to Data: Computer Analysis of Multifactor Data, Second Edition
DASU and JOHNSON · Exploratory Data Mining and Data Cleaning
DAVID and NAGARAJA · Order Statistics, Third Edition
*DEGROOT, FIENBERG and KADANE · Statistics and the Law
DEL CASTILLO · Statistical Process Adjustment for Quality Control
DEMARIS · Regression with Social Data: Modeling Continuous and Limited Response Variables
DEMIDENKO · Mixed Models: Theory and Applications
DENISON, HOLMES, MALLICK and SMITH · Bayesian Methods for Nonlinear Classification and Regression
DETTE and STUDDEN · The Theory of Canonical Moments with Applications in Statistics, Probability and Analysis
DEY and MUKERJEE · Fractional Factorial Plans
DILLON and GOLDSTEIN · Multivariate Analysis: Methods and Applications
DODGE · Alternative Methods of Regression
*DODGE and ROMIG · Sampling Inspection Tables, Second Edition
*DOOB · Stochastic Processes
DOWDY, WEARDEN and CHILKO · Statistics for Research, Third Edition
DRAPER and SMITH · Applied Regression Analysis, Third Edition
DRYDEN and MARDIA · Statistical Shape Analysis
DUDEWICZ and MISHRA · Modern Mathematical Statistics
DUNN and CLARK · Applied Statistics: Analysis of Variance and Regression, Second Edition
DUNN and CLARK · Basic Statistics: A Primer for the Biomedical Sciences, Third Edition
DUPUIS and ELLIS · A Weak Convergence Approach to the Theory of Large Deviations
EDLER and KITSOS (editors) · Recent Advances in Quantitative Methods in Cancer and Human Health Risk Assessment
*ELANDT-JOHNSON and JOHNSON · Survival Models and Data Analysis
ENDERS · Applied Econometric Time Series
ETHIER and KURTZ · Markov Processes: Characterization and Convergence
EVANS, HASTINGS and PEACOCK · Statistical Distribution, Third Edition
FELLER · An Introduction to Probability Theory and Its Applications, Volume I, Third Edition, Revised; Volume II, Second Edition
FISHER and VAN BELLE · Biostatistics: A Methodology for the Health Sciences
FITZMAURICE, LAIRD and WARE · Applied Longitudinal Analysis
*FLEISS · The Design and Analysis of Clinical Experiments
FLEISS · Statistical Methods for Rates and Proportions, Second Edition
FLEMING and HARRINGTON · Counting Processes and Survival Analysis
FULLER · Introduction to Statistical Time Series, Second Edition
FULLER · Measurement Error Models
GALLANT · Nonlinear Statistical Models.
GEISSER · Modes of Parametric Statistical Inference
GELMAN and MENG (editors) · Applied Bayesian Modeling and Casual Inference from Incomplete-data Perspectives
GEWEKE · Contemporary Bayesian Econometrics and Statistics
GHOSH, MUKHOPADHYAY and SEN · Sequential Estimation
GIESBRECHT and GUMPERTZ · Planning, Construction and Statistical Analysis of Comparative Experiments
GIFI · Nonlinear Multivariate Analysis
GIVENS and HOETING · Computational Statistics
GLASSERMAN and YAO · Monotone Structure in Discrete-Event Systems
GNANADESIKAN · Methods for Statistical Data Analysis of Multivariate Observations, Second Edition
GOLDSTEIN and LEWIS · Assessment: Problems, Development and Statistical Issues

*Now available in a lower priced paperback edition in the Wiley Classics Library.

GREENWOOD and NIKULIN · A Guide to Chi-Squared Testing
GROSS and HARRIS · Fundamentals of Queueing Theory, Third Edition
*HAHN and SHAPIRO · Statistical Models in Engineering
HAHN and MEEKER · Statistical Intervals: A Guide for Practitioners
HALD · A History of Probability and Statistics and their Applications Before 1750
HALD · A History of Mathematical Statistics from 1750 to 1930
HAMPEL · Robust Statistics: The Approach Based on Influence Functions
HANNAN and DEISTLER · The Statistical Theory of Linear Systems
HEIBERGER · Computation for the Analysis of Designed Experiments
HEDAYAT and SINHA · Design and Inference in Finite Population Sampling
HEDEKER and GIBBONS · Longitudinal Data Analysis
HELLER · MACSYMA for Statisticians
HINKELMANN and KEMPTHORNE · Design and Analysis of Experiments, Volume 1: Introduction to Experimental Design
HINKELMANN and KEMPTHORNE · Design and analysis of experiments, Volume 2: Advanced Experimental Design
HOAGLIN, MOSTELLER and TUKEY · Exploratory Approach to Analysis of Variance
HOAGLIN, MOSTELLER and TUKEY · Exploring Data Tables, Trends and Shapes
*HOAGLIN, MOSTELLER and TUKEY · Understanding Robust and Exploratory Data Analysis
HOCHBERG and TAMHANE · Multiple Comparison Procedures
HOCKING · Methods and Applications of Linear Models: Regression and the Analysis of Variance, Second Edition
HOEL · Introduction to Mathematical Statistics, Fifth Edition
HOGG and KLUGMAN · Loss Distributions
HOLLANDER and WOLFE · Nonparametric Statistical Methods, Second Edition
HOSMER and LEMESHOW · Applied Logistic Regression, Second Edition
HOSMER and LEMESHOW · Applied Survival Analysis: Regression Modeling of Time to Event Data
HUBER · Robust Statistics
HUBERTY · Applied Discriminant Analysis
HUNT and KENNEDY · Financial Derivatives in Theory and Practice, Revised Edition
HUSKOVA, BERAN and DUPAC · Collected Works of Jaroslav Hajek—with Commentary
HUZURBAZAR · Flowgraph Models for Multistate Time-to-Event Data
IMAN and CONOVER · A Modern Approach to Statistics
JACKSON · A User's Guide to Principle Components
JOHN · Statistical Methods in Engineering and Quality Assurance
JOHNSON · Multivariate Statistical Simulation
JOHNSON and BALAKRISHNAN · Advances in the Theory and Practice of Statistics: A Volume in Honor of Samuel Kotz
JOHNSON and BHATTACHARYYA · Statistics: Principles and Methods, Fifth Edition
JUDGE, GRIFFITHS, HILL, LU TKEPOHL and LEE · The Theory and Practice of Econometrics, Second Edition
JOHNSON and KOTZ · Distributions in Statistics
JOHNSON and KOTZ (editors) · Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present
JOHNSON, KOTZ and BALAKRISHNAN · Continuous Univariate Distributions, Volume 1, Second Edition
JOHNSON, KOTZ and BALAKRISHNAN · Continuous Univariate Distributions, Volume 2, Second Edition
JOHNSON, KOTZ and BALAKRISHNAN · Discrete Multivariate Distributions
JOHNSON, KOTZ and KEMP · Univariate Discrete Distributions, Second Edition
JUREČKOVÁ and SEN · Robust Statistical Procedures: Asymptotics and Interrelations
JUREK and MASON · Operator-Limit Distributions in Probability Theory
KADANE · Bayesian Methods and Ethics in a Clinical Trial Design
KADANE and SCHUM · A Probabilistic Analysis of the Sacco and Vanzetti Evidence
KALBFLEISCH and PRENTICE · The Statistical Analysis of Failure Time Data, Second Edition

*Now available in a lower priced paperback edition in the Wiley Classics Library.

KARIYA and KURATA · Generalized Least Squares
KASS and VOS · Geometrical Foundations of Asymptotic Inference
KAUFMAN and ROUSSEEUW · Finding Groups in Data: An Introduction to Cluster Analysis
KEDEM and FOKIANOS · Regression Models for Time Series Analysis
KENDALL, BARDEN, CARNE and LE · Shape and Shape Theory
KHURI · Advanced Calculus with Applications in Statistics, Second Edition
KHURI, MATHEW and SINHA · Statistical Tests for Mixed Linear Models
*KISH · Statistical Design for Research
KLEIBER and KOTZ · Statistical Size Distributions in Economics and Actuarial Sciences
KLUGMAN, PANJER and WILLMOT · Loss Models: From Data to Decisions
KLUGMAN, PANJER and WILLMOT · Solutions Manual to Accompany Loss Models: From Data to Decisions
KOTZ, BALAKRISHNAN and JOHNSON · Continuous Multivariate Distributions, Volume 1, Second Edition
KOTZ and JOHNSON (editors) · Encyclopedia of Statistical Sciences: Volumes 1 to 9 with Index
KOTZ and JOHNSON (editors) · Encyclopedia of Statistical Sciences: Supplement Volume
KOTZ, READ and BANKS (editors) · Encyclopedia of Statistical Sciences: Update Volume 1
KOTZ, READ and BANKS (editors) · Encyclopedia of Statistical Sciences: Update Volume 2
KOVALENKO, KUZNETZOV and PEGG · Mathematical Theory of Reliability of Time-Dependent Systems with Practical Applications
KUROWICKA and COOKE · Uncertainty Analysis with High Dimensional Dependence Modelling
LACHIN · Biostatistical Methods: The Assessment of Relative Risks
LAD · Operational Subjective Statistical Methods: A Mathematical, Philosophical and Historical Introduction
LAMPERTI · Probability: A Survey of the Mathematical Theory, Second Edition
LANGE, RYAN, BILLARD, BRILLINGER, CONQUEST and GREENHOUSE · Case Studies in Biometry
LARSON · Introduction to Probability Theory and Statistical Inference, Third Edition
LAWLESS · Statistical Models and Methods for Lifetime Data, Second Edition
LAWSON · Statistical Methods in Spatial Epidemiology, Second Edition
LE · Applied Categorical Data Analysis
LE · Applied Survival Analysis
LEE and WANG · Statistical Methods for Survival Data Analysis, Third Edition
LEPAGE and BILLARD · Exploring the Limits of Bootstrap
LEYLAND and GOLDSTEIN (editors) · Multilevel Modelling of Health Statistics
LIAO · Statistical Group Comparison
LINDVALL · Lectures on the Coupling Method
LINHART and ZUCCHINI · Model Selection
LITTLE and RUBIN · Statistical Analysis with Missing Data, Second Edition
LLOYD · The Statistical Analysis of Categorical Data
LOWEN and TEICH · Fractal-Based Point Processes
MAGNUS and NEUDECKER · Matrix Differential Calculus with Applications in Statistics and Econometrics, Revised Edition
MALLER and ZHOU · Survival Analysis with Long Term Survivors
MALLOWS · Design, Data and Analysis by Some Friends of Cuthbert Daniel
MANN, SCHAFER and SINGPURWALLA · Methods for Statistical Analysis of Reliability and Life Data
MANTON, WOODBURY and TOLLEY · Statistical Applications Using Fuzzy Sets
MARCHETTE · Random Graphs for Statistical Pattern Recognition
MARKOVICH · Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and practice
MARDIA and JUPP · Directional Statistics
MARKOVICH · Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and Practice
MARONNA, MARTIN and YOHAI · Robust Statistics: Theory and Methods
MASON, GUNST and HESS · Statistical Design and Analysis of Experiments with Applications to Engineering and Science, Second Edition

*Now available in a lower priced paperback edition in the Wiley Classics Library.

MCCULLOCH and SERLE · Generalized, Linear and Mixed Models
MCFADDEN · Management of Data in Clinical Trials
MCLACHLAN · Discriminant Analysis and Statistical Pattern Recognition
MCLACHLAN, DO and AMBROISE · Analyzing Microarray Gene Expression Data
MCLACHLAN and KRISHNAN · The EM Algorithm and Extensions
MCLACHLAN and PEEL · Finite Mixture Models
MCNEIL · Epidemiological Research Methods
MEEKER and ESCOBAR · Statistical Methods for Reliability Data
MEERSCHAERT and SCHEFFLER · Limit Distributions for Sums of Independent Random Vectors: Heavy Tails in Theory and Practice
MICKEY, DUNN and CLARK · Applied Statistics: Analysis of Variance and Regression, Third Edition
*MILLER · Survival Analysis, Second Edition
MONTGOMERY, PECK and VINING · Introduction to Linear Regression Analysis, Fourth Edition
MORGENTHALER and TUKEY · Configural Polysampling: A Route to Practical Robustness
MUIRHEAD · Aspects of Multivariate Statistical Theory
MULLER and STEWART · Linear Model Theory: Univariate, Multivariate and Mixed Models
MURRAY · X-STAT 2.0 Statistical Experimentation, Design Data Analysis and Nonlinear Optimization
MURTHY, XIE and JIANG · Weibull Models
MYERS and MONTGOMERY · Response Surface Methodology: Process and Product Optimization Using Designed Experiments, Second Edition
MYERS, MONTGOMERY and VINING · Generalized Linear Models. With Applications in Engineering and the Sciences
†NELSON · Accelerated Testing, Statistical Models, Test Plans and Data Analysis
†NELSON · Applied Life Data Analysis
NEWMAN · Biostatistical Methods in Epidemiology
OCHI · Applied Probability and Stochastic Processes in Engineering and Physical Sciences
OKABE, BOOTS, SUGIHARA and CHIU · Spatial Tesselations: Concepts and Applications of Voronoi Diagrams, Second Edition
OLIVER and SMITH · Influence Diagrams, Belief Nets and Decision Analysis
PALTA · Quantitative Methods in Population Health: Extentions of Ordinary Regression
PANJER · Operational Risks: Modeling Analytics
PANKRATZ · Forecasting with Dynamic Regression Models
PANKRATZ · Forecasting with Univariate Box-Jenkins Models: Concepts and Cases
*PARZEN · Modern Probability Theory and Its Applications
PEÑA, TIAO and TSAY · A Course in Time Series Analysis
PIANTADOSI · Clinical Trials: A Methodologic Perspective
PORT · Theoretical Probability for Applications
POURAHMADI · Foundations of Time Series Analysis and Prediction Theory
PRESS · Bayesian Statistics: Principles, Models and Applications
PRESS · Subjective and Objective Bayesian Statistics, Second Edition
PRESS and TANUR · The Subjectivity of Scientists and the Bayesian Approach
PUKELSHEIM · Optimal Experimental Design
PURI, VILAPLANA and WERTZ · New Perspectives in Theoretical and Applied Statistics
PUTERMAN · Markov Decision Processes: Discrete Stochastic Dynamic Programming
QIU · Image Processing and Jump Regression Analysis
RAO · Linear Statistical Inference and its Applications, Second Edition
RAUSAND and HØYLAND · System Reliability Theory: Models, Statistical Methods and Applications, Second Edition
RENCHER · Linear Models in Statistics
RENCHER · Methods of Multivariate Analysis, Second Edition
RENCHER · Multivariate Statistical Inference with Applications
RIPLEY · Spatial Statistics
RIPLEY · Stochastic Simulation
ROBINSON · Practical Strategies for Experimenting

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in a lower priced paperback edition in the Wiley - Interscience Paperback Series.

ROHATGI and SALEH · An Introduction to Probability and Statistics, Second Edition
ROLSKI, SCHMIDLI, SCHMIDT and TEUGELS · Stochastic Processes for Insurance and Finance
ROSENBERGER and LACHIN · Randomization in Clinical Trials: Theory and Practice
ROSS · Introduction to Probability and Statistics for Engineers and Scientists
ROSSI, ALLENBY and MCCULLOCH · Bayesian Statistics and Marketing
ROUSSEEUW and LEROY · Robust Regression and Outline Detection
RUBIN · Multiple Imputation for Nonresponse in Surveys
RUBINSTEIN · Simulation and the Monte Carlo Method
RUBINSTEIN and MELAMED · Modern Simulation and Modeling
RYAN · Modern Regression Methods
RYAN · Statistical Methods for Quality Improvement, Second Edition
SALEH · Theory of Preliminary Test and Stein-Type Estimation with Applications
SALTELLI, CHAN and SCOTT (editors) · Sensitivity Analysis
*SCHEFFE · The Analysis of Variance
SCHIMEK · Smoothing and Regression: Approaches, Computation and Application
SCHOTT · Matrix Analysis for Statistics
SCHOUTENS · Levy Processes in Finance: Pricing Financial Derivatives
SCHUSS · Theory and Applications of Stochastic Differential Equations
SCOTT · Multivariate Density Estimation: Theory, Practice and Visualization
*SEARLE · Linear Models
SEARLE · Linear Models for Unbalanced Data
SEARLE · Matrix Algebra Useful for Statistics
SEARLE and WILLETT · Matrix Algebra for Applied Economics
SEBER · Multivariate Observations
SEBER and LEE · Linear Regression Analysis, Second Edition
SEBER and WILD · Nonlinear Regression
SENNOTT · Stochastic Dynamic Programming and the Control of Queueing Systems
*SERFLING · Approximation Theorems of Mathematical Statistics
SHAFER and VOVK · Probability and Finance: Its Only a Game!
SILVAPULLE and SEN · Constrained Statistical Inference: Inequality, Order and Shape Restrictions
SINGPURWALLA · Reliability and Risk: A Bayesian Perspective
SMALL and MCLEISH · Hilbert Space Methods in Probability and Statistical Inference
SRIVASTAVA · Methods of Multivariate Statistics
STAPLETON · Linear Statistical Models
STAUDTE and SHEATHER · Robust Estimation and Testing
STOYAN, KENDALL and MECKE · Stochastic Geometry and Its Applications, Second Edition
STOYAN and STOYAN · Fractals, Random and Point Fields: Methods of Geometrical Statistics
STYAN · The Collected Papers of T. W. Anderson: 1943–1985
SUTTON, ABRAMS, JONES, SHELDON and SONG · Methods for Meta-Analysis in Medical Research
TANAKA · Time Series Analysis: Nonstationary and Noninvertible Distribution Theory
THOMPSON · Empirical Model Building
THOMPSON · Sampling, Second Edition
THOMPSON · Simulation: A Modeler's Approach
THOMPSON and SEBER · Adaptive Sampling
THOMPSON, WILLIAMS and FINDLAY · Models for Investors in Real World Markets
TIAO, BISGAARD, HILL, PEÑA and STIGLER (editors) · Box on Quality and Discovery: with Design, Control and Robustness
TIERNEY · LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics
TSAY · Analysis of Financial Time Series
UPTON and FINGLETON · Spatial Data Analysis by Example, Volume II: Categorical and Directional Data
VAN BELLE · Statistical Rules of Thumb

*Now available in a lower priced paperback edition in the Wiley Classics Library.

VAN BELLE, FISHER, HEAGERTY and LUMLEY · Biostatistics: A Methodology for the Health Sciences, Second Edition
VESTRUP · The Theory of Measures and Integration
VIDAKOVIC · Statistical Modeling by Wavelets
VINOD and REAGLE · Preparing for the Worst: Incorporating Downside Risk in Stock Market Investments
WALLER and GOTWAY · Applied Spatial Statistics for Public Health Data
WEERAHANDI · Generalized Inference in Repeated Measures: Exact Methods in MANOVA and Mixed Models
WEISBERG · Applied Linear Regression, Second Edition
WELISH · Aspects of Statistical Inference
WESTFALL and YOUNG · Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment
WHITTAKER · Graphical Models in Applied Multivariate Statistics
WINKER · Optimization Heuristics in Economics: Applications of Threshold Accepting
WONNACOTT and WONNACOTT · Econometrics, Second Edition
WOODING · Planning Pharmaceutical Clinical Trials: Basic Statistical Principles
WOOLSON and CLARKE · Statistical Methods for the Analysis of Biomedical Data, Second Edition
WU and HAMADA · Experiments: Planning, Analysis and Parameter Design Optimization
WU and ZHANG · Nonparametric Regression Methods for Longitudinal Data Analysis: Mixed-Effects Modeling Approaches
YANG · The Construction Theory of Denumerable Markov Processes
YOUNG, VALERO-MORA and FRIENDLY · Visual Statistics: Seeing Data with Dynamic Interactive Graphics
*ZELLNER · An Introduction to Bayesian Inference in Econometrics
ZELTERMAN · Discrete Distributions: Applications in the Health Sciences
ZHOU, OBUCHOWSKI and McCLISH · Statistical Methods in Diagnostic Medicine

*Now available in a lower priced paperback edition in the Wiley Classics Library.