SPRINGER BRIEFS IN STATISTICS

Gauss M. Cordeiro
Francisco Cribari-Neto

An Introduction to Bartlett Correction and Bias Reduction

Springer

SpringerBriefs in Statistics

For further volumes:
http://www.springer.com/series/8921

Gauss M. Cordeiro · Francisco Cribari-Neto

An Introduction to Bartlett Correction and Bias Reduction

Springer

Gauss M. Cordeiro
Francisco Cribari-Neto
Departamento de Estatística
Universidade Federal de Pernambuco
Recife, Pernambuco
Brazil

ISSN 2191-544X ISSN 2191-5458 (electronic)
ISBN 978-3-642-55254-0 ISBN 978-3-642-55255-7 (eBook)
DOI 10.1007/978-3-642-55255-7
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014938215

Mathematics Subject Classification (2010): 62F03, 62F10

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

This book is dedicated to Maurice S. Bartlett

Preface

Statistical inference is oftentimes based on first-order asymptotic theory. In particular, it is a common practice to perform likelihood ratio, score and Wald tests using approximate critical values. Such critical values are obtained from the test statistic limiting distribution when the null hypothesis is true. The approximation holds when the number of observations in the sample tends to infinity, and it is thus expected to deliver reliable inferences in large samples. When the sample is not large, however, size distortions are likely to arise. That is, the effective type I error probability may not be close to the nominal size selected by the practitioner. It is thus important to have at hand alternatives that deliver more reliable inference in small samples. In this monograph, we cover analytical corrections known as Bartlett and Bartlett-type corrections. Bartlett corrections are applied to likelihood ratio test statistics whereas Bartlett-type corrections are applied to score test statistics and also to other asymptotically χ^2 criteria. The corrections deliver modified tests with error rates that decay faster toward zero. Thus, such tests can be expected to display superior finite sample behavior.

Practitioners are usually also interested in parameter estimation. Maximum likelihood estimators are typically consistent and asymptotically normal, but are usually biased. That is, the estimator expected value differs from the true parameter value, which implies the existence of a systematic error. We provide analytical and numerical approaches that can be used to reduce the bias of the maximum likelihood estimator. Preventive and corrective bias reduction schemes are presented and discussed. The former entails modifying the log likelihood or score function prior to obtaining the estimator whereas the latter entails obtaining the estimator in the usual fashion and then bias correcting it. These bias corrections can deliver modified estimators that have much smaller systematic errors than the corresponding unmodified estimators.

The material we present in this book is a compilation of analytical results and numerical evidence available in the literature. We do not include new results. Our goal is to present, in a coherent way, strategies that can be used to achieve more accurate inferences. Our main focus lies on obtaining analytical corrections to tests that are based on a first-order asymptotic approximation and also on achieving bias reduction in small samples. Numerical (Monte Carlo) results are presented in order to illustrate the gains involved in using such finite sample

corrections. We also point out that the details involved in many of the derivations were not included in the text since we intend to provide readers with a concise monograph. Further details can be found in the references listed at the end of each chapter.

The structure of our monograph was dictated by three main choices we made. First, we intend to provide readers with a concise overview of the topic. Second, we decided to focus on point estimation and testing inference. We do so by focusing on bias reduction of estimators and corrections that can be applied to test statistics. Additionally, even though our focus lies on analytical corrections we also include material on bootstrap-based inference since it is often cited as an appealing alternative to analytically corrected estimators and tests.

Finally, we would like to thank Klaus Vasconcellos for suggestions on parts of our monograph. We also gratefully acknowledge financial support from CNPq/Brazil.

Recife, February 2014

Gauss M. Cordeiro
Francisco Cribari-Neto

Contents

Acronyms

BS	Birnbaum–Saunders
GLM	Generalized linear model
LBC	Linear-bias-correction
LR	Likelihood ratio
ML	Maximum likelihood
MLE	Maximum likelihood estimator
OLS	Ordinary least squares
OLS	Ordinary least squares estimator
SAR	Synthetic aperture radar

Chapter 1
Likelihood-Based Inference and Finite-Sample Corrections: A Brief Overview

Abstract This chapter introduces the likelihood function and estimation by maximum likelihood. Some important properties of maximum likelihood (ML) estimators are outlined. We also briefly present several important concepts that will be used throughout the book. Three asymptotic testing criteria are also introduced. The chapter also motivates the use of Bartlett and Bartlett-type corrections. The underlying idea is to transform the test statistic in such a way that its null distribution is better approximated by the reference χ^2 distribution. We also investigate the use of bias corrections. They are used to reduce systematic errors in the point estimation process. Finally, we motivate the use of a data resampling method: the bootstrap.

Keywords Bartlett correction · Bartlett-type correction · Bias correction · Bootstrap · Likelihood ratio test · Maximum likelihood · Score test · Wald test

1.1 Introduction

Statistics deals with measurement under uncertainty. Its ultimate goal is to perform inference on a population or phenomenon from which data can be sampled. This is achieved by first considering a model that represents the phenomenon of interest. A model is a simplified representation of a more comprehensive reality. A good model must retain the most important features of the phenomenon it represents. A statistical model has a stochastic component—since it represents a phenomenon that occurs in uncertain fashion—and is typically indexed by fixed and (usually) unknown quantities known as parameters. Statistical inference is then performed on such parameters using data previously collected. By performing inference on the model parameters, we make inference on the model and hence on the phenomenon it is supposed to describe.

Inference can be carried out in three different ways, namely (1) point estimation, (2) interval estimation, and (3) hypothesis testing. Several approaches for

G. M. Cordeiro and F. Cribari-Neto, *An Introduction to Bartlett Correction and Bias Reduction*, SpringerBriefs in Statistics, DOI: 10.1007/978-3-642-55255-7_1,

parameter point estimation were proposed in the literature, the maximum likelihood (ML) method being the most commonly employed. The maximum likelihood estimator (MLE) enjoys desirable properties and can be used when constructing confidence intervals and regions and also in test statistics.

In what follows, let Y be a random variable whose density function with respect to the Lebesgue or the counting measure on the real line is $f_Y(\cdot;\theta)$ which we shall also write as $f(\cdot;\theta)$. Here, θ is the parameter vector that indexes the distribution. It is usually unknown and takes values in the parameter space Θ. We shall also consider random samples obtained from $f(\cdot;\theta)$, which shall be denoted as $Y_1,\ldots,Y_n$. We wish to perform inference on θ using the n-dimensional sample $(Y_1,\ldots,Y_n)$. An estimator is a quantity that depends on the data and optionally on known quantities that can be used to estimate θ (or a given function of θ).

1.2 Likelihood Inference

One of the most widely used estimation methods is the ML method. Its underlying motivation is simple and intuitive. Let $Y_1,\ldots,Y_n$ be an n-dimensional sample and assume that each variate has probability density function (pdf) $f(\cdot;\theta)$, θ being a p-dimensional vector, i.e., $\theta = (\theta_1,\ldots,\theta_p)^\top \in \Theta$. Assume that $Y_1,\ldots,Y_n$ are independent and identically distributed (i.i.d.). Their observed values are denoted as $y_1,\ldots,y_n$.

The likelihood function is the joint density function $f_{Y_1,\ldots,Y_n}(y_1,\ldots,y_n;\theta)$ considered as a function of the parameter $\theta = (\theta_1,\ldots,\theta_p)^\top$. Since the n variates in our sample are i.i.d., the likelihood function is the product of the n marginal pdfs. We shall denote the likelihood function as $L(\theta; Y_1,\ldots,Y_n)$. The MLE is the value of θ in the parameter space Θ which maximizes the likelihood function, if such a value exists. It is noteworthy that the MLE also maximizes the log-likelihood function, $\ell = \ell(\theta; Y_1,\ldots,Y_n) = \log L(\theta; Y_1,\ldots,Y_n)$. The log-likelihood derivative with respect to θ is known as the score function. It is possible to show that

$$\mathbb{E}\left(\frac{\partial \ell}{\partial \theta}\right) = 0,$$

that is, the score function has mean zero. The Fisher information matrix is

$$K(\theta) = \mathbb{E}\left(\frac{\partial \ell}{\partial \theta}\frac{\partial \ell}{\partial \theta^\top}\right).$$

Under the conditions given below, it can be shown that $K(\theta) = \mathbb{E}(-\partial^2\ell/\partial\theta\partial\theta^\top)$. Notice that the Fisher information equals the score function variance.

The relevant assumptions for ML inference can be stated as follows:

1. The parameter space Θ is open, and the log-likelihood function assumes a global maximum in Θ.
2. For almost all y, the fourth-order derivatives $\partial^4 f(y;\theta)/\partial\theta_h\partial\theta_j\partial\theta_k\partial_l$ $(h, j, k, l \in \{1, \ldots, p\})$ exist and are continuous in an open neighborhood $M \subset \Theta$ that contains the true parameter value.
3. Consider the integral with respect to y of a function that can be represented as a polynomial in one or more variables in $f, \log f$ and their derivatives of any order with respect to θ. The derivative of such an integral with respect to any component of θ can be obtained by differentiating inside the integral.
4. Consider the $p \times p$ matrix $K = K(\theta)$, whose (h, j) element is

$$K_{hj}(\theta) = \mathbb{E}\left(\frac{\partial \ell}{\partial \theta_h}\frac{\partial \ell}{\partial \theta_j}\right),$$

   $h, j = 1, \ldots, p$. Here, K is positive definite and finite for all $\theta \in M$.
5. There exist functions M_{hjk} such that

$$\left|\frac{\partial^3 \log f(y;\theta)}{\partial\theta_h\partial\theta_j\partial\theta_k}\right| \leq M_{hjk}(y),$$

   for all $\theta \in M$ and for all $h, j, k = 1, \ldots, p$, and $\mathbb{E}_0[M_{hjk}(Y)] < \infty$, for all $h, j, k = 1, \ldots, p$, where $\mathbb{E}_0$ denotes expectation under the true parameter value.

These regularity conditions hold in most, nearly all applications. They are thus not restrictive.

1.3 Some Properties of Maximum Likelihood Estimators

An important property of the MLE is its invariance. Let $\hat{\theta}$ be the MLE of θ, and let $g(\cdot)$ be a function from $\mathbb{R}^p$ to $\mathbb{R}^s$ (not necessarily one to one). Then, $g(\hat{\theta})$ is the MLE of $g(\theta)$. For instance, suppose that $\hat{\sigma}^2$ is the MLE of a given variance. Then, $\sqrt{\hat{\sigma}^2}$ is the MLE of σ (standard deviation).

The MLE enjoys other important properties, including large sample ones. It is noteworthy that it is consistent for θ, i.e., $\hat{\theta}_n \xrightarrow{p} \theta$, where $\xrightarrow{p}$ denotes convergence in probability and the subscript n indicates dependence on the sample size. This property means that in large samples, the estimator will be close to the true parameter with high probability.

Another important property is related to the asymptotic distribution of the MLE. It follows that

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{\mathscr{D}} \mathcal{N}\left(0, K^{-1}(\theta)\right),$$

where $\stackrel{\mathscr{D}}{\rightarrow}$ denotes convergence in distribution (or law). Notice that the inverse of Fisher's information equals the MLE asymptotic variance and that the estimator is asymptotically Gaussian. It should also be noted that the MLE is asymptotically efficient; that is, its asymptotic variance achieves the Cramér-Rao lower bound; see Bickel and Doksum (2001, pp. 181–182).

1.4 A Simple Example

Let $y_1, \ldots, y_n$ be a random sample from the beta distribution $\mathscr{B}(a, b)$. The density of y_i, for each $i = 1, \ldots, n$, is

$$f(y; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1},$$

where $0 < y < 1, a > 0, b > 0$ and $\Gamma(\cdot)$ is the gamma function. The interest lies in the estimation of $\theta = (a, b)^\top$. The log-likelihood function is

$$\ell(a, b) = n \left\{ (a-1)\log g_1 + (b-1)\log g_2 + \log\left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\right)\right\},$$

where the sufficient statistics g_1 and g_2 are the geometric means of the y_i's and $(1-y_i)$'s, respectively.

The MLEs $\hat{a}$ and $\hat{b}$ of a and b are the solution to the nonlinear system

$$\begin{aligned} \psi(a) - \psi(a+b) &= \log g_1 \\ \psi(b) - \psi(a+b) &= \log g_2, \end{aligned}$$

where $\psi(\cdot)$ denotes the digamma function, i.e., the first derivative of the log-gamma function. The MLEs of a and b cannot be expressed in closed form. They are obtained by numerically maximizing the log-likelihood function using a nonlinear optimization algorithm.

We shall now compute the estimates of the parameters that index the beta law using the R software (http://www.R-project.org). To that end, we shall use the `fitdistr` function of the `MASS` package.

```
> library(MASS)
> set.seed(16851750)  # random number generator seed
> randomsample <- rbeta(100, shape1=0.5, shape2=1.5)
> MLfit = fitdistr(random, dbeta, start=list(shape1=1,
+ shape2=1), lower=c(0,0))
> MLfit
     shape1        shape2
  0.50056877   1.50584321
 (0.05925167) (0.22864059)
```

In this example, the true parameter values are $a = 0.5$ and $b = 1.5$ and the sample size is $n = 100$. The MLEs are $\hat{a} = 0.501$ and $\hat{b} = 1.506$. The number in parentheses are the standard errors of the estimates, which are obtained as the square roots of the diagonal elements of Fisher's information matrix inverse after the unknown parameters are replaced by their MLEs.

1.5 Likelihood-Based Testing Inference

Let θ denote the p-dimensional parameter vector that indexes the model used to represent the population or phenomenon of interest and partition it as $\theta = (\theta_1^\top, \theta_2^\top)^\top$, $\dim(\theta_1) = q$ and $\dim(\theta_2) = p - q$. Suppose our interest lies in making testing inference on θ_1, i.e., we wish to test $H_0 : \theta_1 = \theta_1^0$ against $H_1 : \theta_1 \neq \theta_1^0$, where θ_1^0 is a given q-vector. We say θ_1 is the parameter of interest and θ_2 is the nuisance parameter. For instance, we have a random sample from the beta distribution $\mathscr{B}(a, b)$ and wish to test $H_0 : b = 1$ against a two-sided alternative. Here, b is the parameter of interest and a is the nuisance parameter; additionally, $\theta_1^0 = 1$.

Let $\hat{\theta} = (\hat{\theta}_1^\top, \hat{\theta}_2^\top)^\top$ and $\tilde{\theta} = (\theta_1^{0\top}, \tilde{\theta}_2^\top)^\top$ denote the unrestricted and restricted MLEs of θ, respectively. The restricted MLE is obtained by imposing $\theta_1 = \theta_1^0$ and maximizing $\ell(\theta)$ over θ_2. The likelihood ratio (LR) test statistic is

$$LR = 2\left[\ell(\hat{\theta}) - \ell(\tilde{\theta})\right]. \tag{1.1}$$

When the null hypothesis is true, little is lost by imposing it when estimating the parameter vector and, as a consequence, the likelihood function evaluated at the unrestricted and restricted MLEs should be approximately the same; that is, the difference between the likelihood functions evaluated at the two point estimates is small. It then follows that LR is small and the null hypothesis is not rejected. Notice that likelihood ratio testing inference requires the estimation of both null and non-null models.

Alternatively, the null hypothesis can be tested using Rao's score test (Rao 1948). The underlying idea is that if the null hypothesis is true, the score function (i.e., the log-likelihood derivative) should be close to zero when evaluated at the restricted MLE, $\tilde{\theta}$. The score test statistic is

$$S_R = s(\tilde{\theta})^\top K^{-1}(\tilde{\theta}) s(\tilde{\theta}), \tag{1.2}$$

where $s(\theta) = \partial\ell(\theta)/\partial\theta$ is the score function. Notice that the score function and Fisher's information matrix are evaluated at the restricted MLE, $\tilde{\theta}$. It then follows that only the null model is estimated.

It is also possible to base our testing inference on a different test statistic, namely the Wald statistic. It is given by

$$W = (\hat{\theta}_1 - \theta_1^0)^\top K^{11}(\hat{\theta})^{-1}(\hat{\theta}_1 - \theta_1^0), \tag{1.3}$$

where $K^{11}(\theta)^{-1}$ is the upper $q \times q$ block of Fisher's information matrix inverse; i.e., it is the upper q-dimensional block of $K^{-1}(\theta)$. Thus, the null hypothesis is not rejected when the distance between $\hat{\theta}_1$ and θ_1^0 is small. Notice that in order to compute the Wald test statistic, one only estimates the non-null (unrestricted) model.

The exact null distributions of the test statistics given in (1.1), (1.2) and (1.3) are usually unknown. Under certain regularity conditions (Bickel and Doksum 2001, Chap. 6; Serfling 1978, Chap. 4), however, it can be established that they converge to χ^2_q as $n \to \infty$. It is then possible to base our testing inference on critical values obtained from such a distribution. It follows that, for $T = LR, S_R, W$ and under the null hypothesis, $\Pr(T \leq \chi^2_{1-\alpha;q}) = \alpha + o(1)$, where α is the test nominal significance level and $\chi^2_{1-\alpha;q}$ is the $(1-\alpha)$th χ^2_q upper quantile. That is, the difference between the probability that $T \leq \chi^2_{1-\alpha;q}$ (null rejection rate) and α (the nominal significance level) vanishes as $n \to \infty$.

1.6 Some Remarks on Bartlett and Bartlett-Type Corrections

The likelihood function $L(\theta) = L(\theta; y)$ is the basis for most methods of statistical inference. A natural rule is to base inference on $L(\hat{\theta})/L(\theta) > c$ in order to decide what is the range of 'plausible' values of θ, where θ is assumed to have dimension p and $\hat{\theta}$ is the MLE of θ. Inference based on the likelihood function can also be calibrated with reference to the probability model $f(y; \theta)$, by examining the distribution of $L(\theta)$ as a random function, or more usually, by examining the distribution of various associated quantities. Let $\ell(\theta) = \log[L(\theta)]$ be the log-likelihood function. The asymptotic likelihood theory is based on a version of the central limit theorem for the score function $U(\theta) = \partial\ell(\theta)/\partial\theta$ and also on Fisher's information matrix $K(\theta)$. These quantities were introduced in Sect. 1.2. If $Y = (Y_1, \ldots, Y_n)^\top$ has independent components, then $U(\theta)$ is a sum of n independent components, which, under mild regularity conditions, is asymptotically normal. If $\hat{\theta}$ is consistent for θ and $L(\theta)$ has sufficient regularity, the quantities $(\hat{\theta} - \theta)^\top K(\theta)(\hat{\theta} - \theta)$, $U(\theta)^\top K(\theta)^{-1} U(\theta)$ and $2\{\ell(\hat{\theta}) - \ell(\theta)\}$ converge in distribution to χ^2_p. It is noteworthy, however, that the use of χ^2_p as an approximation to the true underlying distributions can lead to inaccurate inferences when the sample size is small. The book by Cox and Hinkley (1974) gives a detailed account of likelihood inference and principles of statistical inference. Other good book-length treatments of likelihood inference are Barndorff-Nielsen and Cox (1994), Pawitan (2000), Severini (2000), and Brazzale et al. (2000).

Large sample tests are commonly used in the applied statistics since exact tests are not always available. These tests rely on what is called 'first-order asymptotics'; that is, they employ critical values obtained from a known limiting null distribution. Generally speaking, the main difficulty of testing a null hypothesis using the

LR statistic lies not so much in deriving its closed-form expression—when it has one—but in finding its exact null distribution, or at least a good approximation to it. In a very influential paper, Bartlett (1937) pioneered the correction to the LR statistic in the context of comparing the variances of several populations. For regular problems, Lawley (1956), through a heroic series of calculations, obtained a general formula for the null expected value of LR and demonstrated that all cumulants of the Bartlett-corrected statistic for testing a composite hypothesis agree with those of the reference χ^2 distribution with error of order $n^{-3/2}$.[1] Alternative expressions for the Bartlett corrections were developed by DiCiccio and Stern (1993), McCullagh and Cox (1986), and Skovgaard (2001). In particular, Cordeiro (1983, 1987) was the first to provide matrix expressions for generalized linear models.

Cordeiro and Ferrari (1991) extended the idea of Bartlett corrections to other test statistics, such as the score (S_R) and Wald (W) statistics. In fact, they derived a general formula for Bartlett-type corrections to improve any test statistic that is, under the null hypothesis, asymptotically distributed as χ^2. The standard Bartlett correction is a special case of their general result. Bartlett and Bartlett-type corrections intend to bring the empirical sizes of asymptotic tests close to the corresponding nominal sizes. In most cases, they do so quite effectively. It is important to bear in mind that these corrections can lead to a loss in power. However, an important result is that the untransformed statistic and its Bartlett-corrected version have the same local power to order n^{-2}. More precisely, let S be a test statistic which is χ^2 distributed under the null hypothesis and let $S^\star$ denote the Bartlett-corrected statistic obtained as a transformation of S. Then, under local (Pitman) alternatives, $\Pr(S^\star \geq x) = \Pr(S \geq x) + \mathcal{O}(n^{-2})$.

In this book, we shall restrict ourselves to the LR, score test, and Wald test, since they are the most commonly used large sample testing inference. As is well known, these three statistics are asymptotically distributed as χ^2 when the null hypothesis H_0 is true, where q is the number of restrictions under test. However, it is also well known that this first-order approximation may not be accurate in finite samples, thus leading to size distortions. We address the issue of evaluating such approximation and designing more accurate tests. The question 'Can we do better?' can be approached from two distinct viewpoints. First, we can obtain a new test statistic whose null distribution is better approximated by the first-order limiting distribution. Second, we can obtain a new distribution which is 'closer' to the test statistic exact null distribution. In this monograph, we shall focus on the former approach. Readers interested in the latter approach are referred to Barndorff-Nielsen and Cox (1979, 1989), Reid (1988, 1991), and Hall (1992) and the references therein.

One of the main goals of our monograph is to provide a unified review of the literature on Bartlett and Bartlett-type corrections, i.e., corrections that can be applied to test statistics (not to critical values). An issue of interest is how to define Bartlett-type corrections since it is possible to write the correction in different ways which are equivalent up to a certain order of magnitude. We address this issue by Monte Carlo simulation. We also include discussions on how to obtain the corrections in

[1] Henceforth, 'to order n^{-k}' means that terms of order smaller than n^{-k} are neglected.

regression models, such as generalized linear models, Birnbaum-Saunders nonlinear regression models, and heteroskedastic linear regressions. We use the linear regression framework to address two important issues through simulation: the influence of the covariate values and of the number of nuisance parameters on the first-order asymptotic approximation used in some asymptotic tests.

Bartlett corrections constitute an important topic of research among statisticians. However, they have not yet found their appropriate space and usage in several applied areas of statistics, in which size corrections are almost always based on transformations of critical values obtained from Edgeworth expansions. We hope this book will help narrow this gap. The authors have established general results and explicit expressions for Bartlett and Bartlett corrections in a series of joint publications, as can be seen in their Web pages: http://www.de.ufpe.br/~gauss (Gauss M. Cordeiro) and http://www.de.ufpe.br/~cribari (Francisco Cribari-Neto). Some applications of Bartlett-type corrections in regression models include score tests for generalized linear models with known dispersion (Cordeiro et al. 1993) and unknown dispersion (Cribari-Neto and Ferrari 1995), exponential family nonlinear models (Ferrari and Cordeiro 1996), and heteroskedastic t regression models (Barroso et al. 2002), among several others. A detailed account of Bartlett and Bartlett-type corrections can be found in Cribari-Neto and Cordeiro (1996).

1.7 Some Remarks on Bias Corrections

We shall also be concerned with point estimation. To that end, we shall review the literature on bias correction. Bias is a systematic error, and there are strategies that can be used to reduce it. The MLEs can be quite biased in small samples. It is thus important to evaluate the n^{-1} biases of these estimators, where n is the sample size, and then define modified estimators that are bias free to this order of approximation. In particular, it is important to derive closed-form expressions for the second-order biases of estimators in some classes of models which can be used in practical applications in order to evaluate the accuracy of these estimators and also to define estimators with smaller biases.

One of our goals in this monograph is to review the literature on bias correction of MLEs. The obvious difficulty is that many MLEs cannot be expressed as explicit functions of the data. Over the last 25 years, there have been many advances with respect to bias calculation of nonlinear MLEs in special distributions and wider classes of regression models such as generalized linear models and heteroskedastic regressions. The computation of higher-order biases is perhaps one of the most important approximations in the theory of estimation by ML in regression models.

There has been considerable interest in finding simple closed-form expressions for second-order biases of MLEs in some classes of regression models. By 'closed-form' we mean expressions that do not involve cumulants of log-likelihood derivatives. In fact, the $\mathcal{O}(n^{-1})$ biases of the MLEs have been derived in homoscedastic normal nonlinear models (Cook et al. 1986), generalized log-gamma regression models (Young

and Bakir 1987), generalized linear models (Cordeiro and McCullagh 1991), multiplicative regression models (Cordeiro 1993), ARMA models (Cordeiro and Klein 1994), multivariate nonlinear regression models with normal errors (Cordeiro and Vasconcellos 1997), univariate nonlinear Student's t-regression models (Cordeiro et al. 1998), multivariate Student's t-regression models (Vasconcellos and Cordeiro 2000), heteroskedastic models (Vasconcellos et al. 2000), beta regression models (Ospina et al. 2006), and heteroscedastic normal linear models (Cordeiro 2008). These results were obtained using the general formula given by Cox and Snell (1968). Simulation results on bias corrections can be found in Cordeiro and Cribari-Neto (1993). An appealing alternative approach to computer-intensive bias correction is described by MacKinnon and Smith (1998). For alternative methods, see Cadigan (1994) and Taniguchi and Puri (1995). Second- and third-order bias corrections for one-parameter models were obtained by Ferrari et al. (1996). More recent general results on bias corrections in regression models can be found in Patriota and Lemonte (2009).

1.8 Some Remarks on the Bootstrap

Even though our focus is on analytical corrections to test statistics and estimators, we also cover alternatives that are based on data resampling, more specifically on bootstrap resampling (Efron 1979). The underlying idea is that additional artificial samples can be obtained by sampling from the original sample as if we were sampling from the population. The random drawing mechanism can be of parametric or nonparametric nature. Higher precision can be achieved by using nested bootstrap schemes; see Hall and Martin (1988). For further details on the bootstrap method, see Efron and Tibshirani (1986, 1993), Hall (1992), Shao and Tu (1995) and the references therein. We also refer readers to Young (1994), who also lists the shortcomings of using data resampling. The relationship between Edgeworth expansions and the bootstrap is discussed in generality by Hall (1992). For an econometric example of this relationship, see Rayner (1990). Rocke (1989) suggested the use of a bootstrap Bartlett adjustment for the log-likelihood ratio statistic in the context of seemingly unrelated regressions. As we shall see, his proposal is to use data resampling to estimate the Bartlett correction factor.

In what follows, we shall describe how the bootstrap can be used as an alternative to analytical finite sample corrections to estimators and tests. Empirical researchers can then choose which method is more appropriate to the application at hand. It is important to note that two researchers who use the same data and perform the same analytical correction will arrive at exactly the same result. The same does not hold true, however, for the bootstrap since the final exact result will depend on the number of bootstrap replication, on the random number generator used, and on other factors. The data resampling mechanism used can also be a discrepancy source. The two approaches are thus different in nature, but they aim at the same goal: delivering more accurate and reliable inferences.

References

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society A*, *160*, 268–282.

Barroso, L. P., Cordeiro, G. M., & Vasconcellos, K. L. P. (2002). Second-order asymptotic for score tests in heteroskedastic t regression models. *Communications in Statistics, Theory and Methods*, *31*, 1515–1529.

Barndorff-Nielsen, O. E., & Cox, D. R. (1979). Edgeworth and saddle-point approximations with statistical applications. *Journal of the Royal Statistical Society B*, *41*, 279–312.

Barndorff-Nielsen, O. E., & Cox, D. R. (1989). *Asymptotic techniques for use in statistics*. London: Chapman and Hall.

Barndorff-Nielsen, O. E., & Cox, D. R. (1994). *Inference and asymptotics*. London: Chapman and Hall.

Bickel, P. J., & Doksum, K. A. (2001). *Mathematical statistics: Basic ideas and selected topics* (2nd ed.). Upper Saddle River: Prentice Hall.

Brazzale, A. R., Davison, A. C., & Reid, N. (2000). *Applied asymptotics*. Cambridge: Cambridge University Press.

Cadigan, N. G. (1994). Bias approximation for maximum likelihood estimates. *Journal of Statistical Computation and Simulation*, *51*, 89–95.

Cook, D. R., Tsai, C. L., & Wei, B. C. (1986). Bias in nonlinear regression. *Biometrika*, *73*, 615–623.

Cordeiro, G. M. (1983). Improved likelihood ratio statistics for generalized linear models. *Journal of the Royal Statistical Society B*, *45*, 404–413.

Cordeiro, G. M. (1987). On the corrections to the likelihood ratio statistics. *Biometrika*, *74*, 265–274.

Cordeiro, G. M. (1993). Bartlett corrections and bias correction for two heteroscedastic regression models. *Communications in Statistics, Theory and Methods*, *22*, 169–188.

Cordeiro, G. M. (2008). Corrected maximum likelihood estimators in linear heteroscedastic regression models. *Brazilian Review of Econometrics*, *28*, 53–67.

Cordeiro, G. M., & Cribari-Neto, F. (1993). On Bartlett corrections, bias reduction and a new class of transformations. *Brazilian Journal of Probability and Statistics*, *7*, 179–200.

Cordeiro, G. M., & Ferrari, S. L. P. (1991). A modified score statistic having chi-squared distribution to order n^{-1}. *Biometrika*, *78*, 573–582.

Cordeiro, G. M., Ferrari, S. L. P., & Paula, G. A. (1993). Improved score tests for generalized linear models. *Journal of the Royal Statistical Society B*, *55*, 661–674.

Cordeiro, G. M., & Klein, R. (1994). Bias correction in ARMA models. *Statistics and Probability Letters*, *19*, 169–176.

Cordeiro, G. M., & McCullagh, P. (1991). Bias correction in generalized linear models. *Journal of the Royal Statistical Society B*, *53*, 629–643.

Cordeiro, G. M., & Vasconcellos, K. L. P. (1997). Bias correction for a class of multivariate nonlinear regression models. *Statistics and Probability Letters*, *35*, 155–164.

Cordeiro, G. M., Vasconcellos, K. L. P., & Santos, M. L. F. (1998). On the second-order bias of parameter estimates in nonlinear regression models with student t errors. *Journal of Statistical Computation and Simulation*, *60*, 363–378.

Cribari-Neto, F., & Cordeiro, G. M. (1996). On Bartlett and Bartlett-type corrections. *Econometric Reviews*, *15*, 339–367.

Cribari-Neto, F., & Ferrari, S. L. P. (1995). Second order asymptotics for score tests in generalized linear models. *Biometrika*, *82*, 426–432.

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical Statistics*. London: Chapman and Hall.

Cox, D. R., & Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society B*, *30*, 248–275.

DiCiccio, T. J., & Stern, S. E. (1993). On Bartlett adjustments for approximate bayesian inference. *Biometrika*, *80*, 731–740.

Efron, B. (1979). Bootstrap methods: another look at the jacknife. *Annals of Statistics*, *7*, 1–26.

Efron, B., & Tibshirani, R. J. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, *1*, 54–96.

Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Ferrari, S. L. P., Botter, D. A., Cordeiro, G. M., & Cribari-Neto, F. (1996). Second and third order bias reduction in one-parameter family models. *Statistics and Probability Letters*, *30*, 339–345.

Ferrari, S. L. P., & Cordeiro, G. M. (1996). Corrected score tests for exponential family nonlinear models. *Statistics and Probability Letters*, *26*, 7–12.

Hall, P. (1992). *The bootstrap and the Edgeworth expansion*. New York: Springer.

Hall, P., & Martin, M. A. (1988). On bootstrap resampling and iteration. *Biometrika*, *75*, 661–671.

Lawley, D. N. (1956). A general method for approximating to the distribution of the likelihood ratio criteria. *Biometrika*, *71*, 233–244.

MacKinnon, J. G., & Smith, J. A. A. (1998). Approximate bias correction in econometrics. *Journal of Econometrics*, *85*, 205–230.

McCullagh, P., & Cox, D. R. (1986). Invariants and the likelihood ratio statistic. *Annals of Statistics*, *14*, 1419–1430.

Ospina, R., Cribari-Neto, F., & Vasconcellos, K. L. P. (2006). Improved point and interval estimation for a beta regression model. *Computational Statistics and Data Analysis*, *51*, 960–981. [Errata: vol. 55, p. 2445, 2011].

Patriota, A. G., & Lemonte, A. J. (2009). Bias correction in a multivariate regression model with general parameterization. *Statistics and Probability Letters*, *79*, 1655–1662.

Pawitan, Y. (2000). *In all likelihood*. Oxford: Oxford University Press.

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, *44*, 50–57.

Rayner, R. K. (1990). Bootstrap tests for generalized least squares regression models. *Economics Letters*, *34*, 261–265.

Rocke, D. M. (1989). Bootstrap Bartlett adjustment in seemingly unrelated regression. *Journal of the American Statistical Association*, *84*, 598–601.

Reid, N. (1988). Saddlepoint methods and statistical inference. *Statistical Science*, *3*, 213–238.

Reid, N. (1991). Statistical theory and modeling: In honour of Sir David Cox, FRS. In D. V. Hinkley, N. Reid, & E. J. Snell (Eds.), *Approximations and asymptotics*. London: Chapman and Hall.

Serfling, R. J. (1978). *Approximation theorems of mathematical statistics*. New York: Wiley.

Shao, J., & Tu, D. (1995). *The jackknife and bootstrap*. New York: Springer.

Severini, T. A. (2000). *Likelihood methods in statistics*. Oxford: Oxford University Press.

Skovgaard, I. (2001). Likelihood asymptotic. *Scandinavian Journal of Statistics*, *28*, 3–32.

Taniguchi, M., & Puri, M. L. (1995). Higher order asymptotic theory for normalizing transformations of maximum likelihood estimators. *Annals of the Institute of Statistical Mathematics*, *47*, 581–600.

Vasconcellos, K. L. P., & Cordeiro, G. M. (2000). Bias corrected estimates in multivariate Student-*t* regression models. *Communications in Statistics, Simulation and Computation*, *29*, 797–822.

Vasconcellos, K. L. P., Cordeiro, G. M., & Barroso, L. P. (2000). Improved estimation for robust econometric regression models. *Brazilian Journal of Probability and Statistics*, *14*, 141–157.

Young, G. A. (1994). Bootstrap: More than a stap in the dark? *Statistical Science*, *9*, 382–415.

Young, D. H., & Bakir, S. T. (1987). Bias correction for a generalized log-gamma regression model. *Technometrics*, *29*, 183–191.

Chapter 2
Bartlett Corrections and Bootstrap Testing Inference

Abstract This chapter introduces the Bartlett correction to the likelihood ratio test statistic. The likelihood ratio test typically employs critical values that are only asymptotically correct and, as consequence, size distortions arise. The null distribution of Bartlett-corrected test statistic is typically better approximated by the limiting distribution than that of the corresponding unmodified test statistics. The correction reduces the test error rate, that is, size discrepancies of the corrected test vanish at a faster rate. We also show how to estimate the exact null distribution of a test statistic using data resampling (bootstrap). It is noteworthy that the bootstrap can be used to estimate the Bartlett correction factor.

Keywords Bartlett correction · Bootstrap · Likelihood ratio test · Size distortion · Power · Type I error

2.1 Introduction

Statistical large-sample theory is concerned with the behavior of statistical procedures as the sample size increases to infinity. It is important to statisticians because it usually delivers simple approximations that work well in finite samples. Statisticians often seek to approximate quantities, such as the density of a test statistic, that depend on the sample size in order to obtain better approximate distributions. The resulting approximation should be easy to handle either analytically or numerically. Asymptotic expansions are usually assessed by examining the error behavior as the sample size increases to infinity.

The LR statistic is one of the most commonly used statistics for performing testing inference in parametric models. Let w be the LR statistic for testing some composite or simple null hypothesis H_0 against an alternative hypothesis H. It is well known that under the null hypothesis, w is asymptotically distributed as χ^2_q, where q is the difference between the dimensions of the parameter spaces under the two

G. M. Cordeiro and F. Cribari-Neto, *An Introduction to Bartlett Correction and Bias Reduction*, SpringerBriefs in Statistics, DOI: 10.1007/978-3-642-55255-7_2,

hypotheses (alternative and null). Since the test uses an approximate critical value, which is obtained from the limiting null distribution of w, size distortions may take place in small samples.

Generally speaking, the main difficulty of testing a null hypothesis using the LR criterion lies not so much in deriving its closed form, when it has one, but in finding its exact distribution, or at least a good approximation, when the null hypothesis is true. In a paper that later became quite influential, Bartlett (1937) proposed an improved LR statistic. His argument goes as follows. Suppose that under the null hypothesis $\mathbb{E}(w) = q + b + \mathcal{O}(n^{-2})$, where b is a constant of order $\mathcal{O}(n^{-1})$ that can be consistently estimated under H, n is the number of observations or some related quantity and q is the difference between the dimensions of the parameter spaces under the alternative and null hypotheses. Then, the expected value of the transformed statistic $w^* = w/(1 + b/q)$ is closer to that of the limiting null χ^2 distribution than that of w. For the test of homogeneity of variances, he showed that the first three cumulants of w^* agree with those of the χ^2_q distribution with error of order $\mathcal{O}(n^{-2})$, thus providing strong grounds for one to believe that the density of w^* is better approximated by the limiting null χ^2 distribution than that of w.

It is well known that a way of improving the χ^2 approximation to the LR statistic is by dividing w by the correction factor $c = (1 + b/q)$; this is known as *Bartlett correction* (Lawley 1956; Hayakawa 1977; Cordeiro 1987). This idea was pioneered by Bartlett (1937) and later generalized by Lawley (1956). Bartlett obtained a number of these corrections in a series of papers on multivariate analysis that were published between 1938 and 1955. The correction factors obtained by Bartlett were widely used for improving the large-sample χ^2 approximation to the null distribution of w. The Bartlett correction $c = 1 + b/q$ now represents an important tool for improving the χ^2 approximation used when performing LR tests. The expected value of the Bartlett-corrected statistic $w^* = w/c$ is closer to that of χ^2_q than that of w. Moreover, for continuous data, the null distribution of w^* is, in general, closer to χ^2_q than the null distribution of w. Box (1949) used Bartlett's approach to investigate the moments of w in the following cases: the test of constancy of variances and covariances of k sets of p-variate samples and the Wilks test for the independence of k sets of residuals, where the ith set contains p_i variables. For these cases, he showed that the modified statistic w^* follows a χ^2_q distribution more closely than does the unmodified statistic w. Box's results are applicable whenever the Laplace transform of the test statistic can be explicitly written in terms of gamma and reciprocal gamma functions.

A general method to obtain Bartlett corrections for regular statistical models was developed in full generality by Lawley (1956), who obtained a general formula for the correction factor c as function of covariant tensors. He derived expressions for the moments of certain log-likelihood derivatives and, through an exceedingly complicated calculation, obtained a general formula for the null expected value of w. Further, he showed that all cumulants of the corrected statistic w^* for testing composite hypotheses agree with those of the reference χ^2_q distribution with error of order $\mathcal{O}(n^{-2})$; see Hayakawa (1977) and Cordeiro (1987). The analytical derivation of the Bartlett corrections using Lawley's approach is, however, notoriously cumbersome

since it requires the computation of some joint cumulants of log-likelihood derivatives. See, also, Eqs. (5.30)–(5.32) in Barndorff-Nielsen and Cox (1994).

It is noteworthy that the expected value needed for determining w^* may be very difficult or even impossible to compute. A general matrix formula for c was derived by Cordeiro (1993a). His matrix formula can be useful when it comes for implementing Bartlett corrections. Such corrections can substantially reduce size distortions when used with continuous data. However, for discrete data, the Bartlett correction may not yield a clear improvement in the asymptotic error rate of the χ^2 approximation. Several papers have focused on deriving Bartlett corrections for special regression models using matrix formulae for specific models, bypassing the traditional machinery of calculating the required cumulants. One can always obtain these matrix formulae when the joint cumulants of log-likelihood derivatives are invariant under permutation of parameters. These formulae can be easily handled by computer algebra systems (e.g., MATHEMATICA and MAPLE) and programming languages with support for matrix operations (e.g., GAUSS, OX, and R).

2.2 Bartlett Identities

Let $L = L(\theta)$ and $\ell = \ell(\theta) = \log[L(\theta)]$ be the total likelihood and total log-likelihood functions for a regular parametric model depending on a $p \times 1$ vector θ of unknown parameters having continuous partial derivatives up to the fourth order. We assume that the model is regular in the sense that we can interchange differentiation and integration. The derivatives at an arbitrary point θ are denoted by $U_r = \partial\ell/\partial\theta_r$, $U_{rs} = \partial^2\ell/\partial\theta_r\partial\theta_s$, $U_{rst} = \partial^3\ell/\partial\theta_r\partial\theta_s\partial\theta_t$, and so on. Hereafter, the moments of the log-likelihood derivatives are assumed finite and are denoted by $\mu_r = \mathbb{E}(U_r)$, $\mu_{rs} = \mathbb{E}(U_{rs})$, $\mu_{r,s} = \mathbb{E}(U_rU_s)$, $\mu_{rst} = \mathbb{E}(U_{rst})$, $\mu_{r,st} = E(U_rU_{st})$, and so on.

Differentiation of $\int L\,dy = 1$ with respect to θ_r and reversing the order of differentiation gives $\mu_r = \mathbb{E}(U_r) = 0$. This is the well-known result that the mean of the score function equals zero. From this basic relation, we can obtain a sequence of balance equations known as *Bartlett identities*. In particular, differentiation with respect to θ_s yields $\int(U_{rs} + U_rU_s)L\,dy = 0$, which can be expressed as $\mu_{rs} + \mu_{r,s} = 0$. This equation provides two alternative formulae for computing the expected information matrix for θ: $K = \{\mu_{r,s}\} = \{-\mu_{rs}\}$.

In similar fashion, we obtain the third Bartlett identity: $\mu_{r,s,t} + \mu_{rst} + \Sigma_{(3)}\mu_{r,st} = 0$, where the notation (m) indicates the sum of m permutations of indices. The fourth Bartlett identity is

$$\mu_{r,s,t,u} + \mu_{rstu} + \Sigma_{(4)}\mu_{rst,u} + \Sigma_{(3)}\mu_{rs,tu} + \Sigma_{(6)}\mu_{rs,t,u} = 0.$$

We now introduce the cumulants (denoted from now on by κ's) of log-likelihood derivatives which can be defined in terms of the moments by

$$\begin{aligned}
\kappa_r &= \mu_r = 0,\ \kappa_{rs} = \mu_{rs},\ \kappa_{r,s} = \mu_{r,s},\\
\kappa_{r,s,t} &= \mathrm{cum}(U_r, U_s, U_t) = \mu_{r,s,t},\ \kappa_{rs,t} = \mathrm{Cov}(U_{rs}, U_t) = \mu_{rs,t},\\
\kappa_{rs,tu} &= \mathrm{Cov}(U_{rs}, U_{tu}) = \mu_{rs,tu} - \mu_{rs}\mu_{tu},\\
\kappa_{r,s,tu} &= \mathrm{cum}(U_r, U_s, U_{tu}) = \mu_{r,s,tu} - \mu_{r,s}\mu_{tu},\\
\kappa_{r,s,t,u} &= \mathrm{cum}(U_r, U_s, U_t, U_u) = \mu_{r,s,t,u} - \Sigma_{(3)}\mu_{r,s}\mu_{t,u},
\end{aligned}$$

and so on.

These cumulants satisfy the *Bartlett identities*

$$\kappa_{rs} + \kappa_{r,s} = 0,\ \kappa_{r,s,t} + \kappa_{rst} + \Sigma_{(3)}\kappa_{r,st} = 0$$

and

$$\kappa_{rstu} + \Sigma_{(4)}\kappa_{r,stu} + \Sigma_{(3)}\kappa_{rs,tu} + \Sigma_{(6)}\kappa_{r,s,tu} + \kappa_{r,s,t,u} = 0.$$

The order of the identity is defined by the number of indices in its terms. So, $\kappa_{r,s,t} + \kappa_{rst} + \Sigma_{(3)}\kappa_{r,st} = 0$ is a Bartlett identity of third order.

In addition, the derivatives of the cumulants are denoted by $\kappa_{rs}^{(t)} = \partial\kappa_{rs}/\partial\theta_t$, $\kappa_{rs}^{(tu)} = \partial\kappa_{rs}/\partial\theta_t\theta_u$, etc. From these definitions, we can obtain new Bartlett identities involving the κ's:

$$\kappa_{r,st} + \kappa_{rst} - \kappa_{st}^{(r)} = 0,\ \kappa_{r,s,t} - 2\kappa_{rst} + \Sigma_{(3)}\kappa_{rs}^{(t)} = 0,\ \kappa_{rs}^{(t)} + \kappa_{r,s,t} + \kappa_{r,st} + \kappa_{s,rt} = 0,$$

$$\kappa_{rst}^{(u)} = \kappa_{rstu} + \kappa_{rst,u},\ \kappa_{r,stu} + \kappa_{rstu} - \kappa_{stu}^{(r)} = 0,$$

$$\kappa_{r,s,t,u} = -3\kappa_{rstu} + 2\Sigma_{(4)}\kappa_{rst}^{(u)} - \Sigma_{(6)}\kappa_{rs}^{(tu)} + \Sigma_{(3)}\kappa_{rs,tu},$$

$$\kappa_{r,s,tu} = \kappa_{rstu} - \kappa_{rtu}^{(s)} - \kappa_{stu}^{(r)} + \kappa_{tu}^{(rs)} - \kappa_{rs,tu},$$

etc. These identities usually simplify the derivation of several asymptotic quantities in regular likelihood theory.

For the one-parameter model (i.e., when θ is a scalar), we obtain $\kappa_{\theta,\theta\theta} + \kappa_{\theta\theta\theta} - \kappa_{\theta\theta}^{(\theta)} = 0$, $\kappa_{\theta,\theta,\theta} - 2\kappa_{\theta\theta\theta} + 3\kappa_{\theta\theta}^{(\theta)} = 0$, $\kappa_{\theta\theta\theta}^{(\theta)} = \kappa_{\theta\theta\theta\theta} + \kappa_{\theta\theta\theta,\theta}$, and so on.

As an example, consider a simple derivation of the joint cumulants. For the $N(\mu, \sigma^2)$ distribution, the log likelihood $\ell = \ell(\theta)$ for $\theta = (\mu, \sigma^2)^\top$ from a sample of n i.i.d. random variables is

$$\ell = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2.$$

The joint cumulants are $\kappa_{\mu\mu} = -n/\sigma^2$, $\kappa_{\sigma^2\sigma^2} = -n/2\sigma^4$, $\kappa_{\mu\sigma^2} = 0, \kappa_{\mu,\mu,\mu} = \kappa_{\mu,\mu\mu} = \kappa_{\mu\mu\mu} = 0, \kappa_{\sigma^2,\sigma^2,\sigma^2} = -\kappa_{\sigma^2,\sigma^2\sigma^2} = n/\sigma^6, \kappa_{\sigma^2\sigma^2\sigma^2} = 2n/\sigma^6, \kappa_{\mu,\mu\sigma^2} =$

$-\kappa_{\mu\mu\sigma^2} = -n/\sigma^4, \kappa_{\mu,\mu,\sigma^2} = 3n/\sigma^4, \kappa_{\mu\mu\sigma^2\sigma^2} = -2n/\sigma^6$, etc. Some of these cumulants can be easily computed using Bartlett identities.

2.3 Lawley's Expansion

Bartlett (1938, 1947, 1954) obtained a number of adjustment factors in the area of multivariate analysis, and these factors became widely used for improving the large-sample χ^2 approximation to the null distribution of the LR statistic. Box (1949) used Bartlett's (1937) results to investigate in detail the general expression for the moments of LR statistics in the following cases: the test of constancy of variance and covariance of m sets of p-variate samples and the Wilks test for the independence of k sets of residuals, the ith set having p_i variates. He has shown that in these cases and under the null hypothesis, the modified statistic w^* follows a χ^2 distribution more closely than the unmodified statistic w. Box's results are applicable to all tests for which the Laplace transform of the test statistic can be explicitly written in terms of gamma and reciprocal gamma functions. In particular, it is possible to use these results to obtain $\mathbb{E}(w)$ and $\mathrm{Var}(w)$. The results in Lawley (1956), McCullagh and Cox (1986) and Cordeiro (1993a) are, however, more useful for deriving Bartlett corrections in regression and time series models.

For regular problems, Lawley (1956) obtained expressions for the moments of certain derivatives of the log-likelihood function and, using an exceedingly complicated derivation, gave a general formula for the null expected value of the log-likelihood criterion and showed that all cumulants of the Bartlett-corrected statistic w^* for testing a composite hypothesis agree with those of the reference χ^2 distribution with error of order $\mathcal{O}(n^{-2})$. A related reference is Beale (1960), who obtained an approximation to the asymptotic distribution of the residual sum of squares in the normal non-linear regression model and gave an interpretation for the correction factor in terms of the curvature of a surface. Beale's paper has three noteworthy contributions: It defined a measure of the intrinsic non-linearity of a regression model as a function of the covariates and of the parameter values, it showed how improved confidence regions for the parameter values of the model can be obtained, and it showed how to select a suitable transformation of the parameters that delivers near linearity in the neighborhood of the MLEs. His results, however, are limited to normal models. In terms of Bartlett correction, its main contribution was to give a geometric interpretation of the correction for normal models. This interpretation was later generalized to non-normal models by McCullagh and Cox (1986).

Suppose we have n independent but not necessarily identically distributed variates $Y = (Y_1, \ldots, Y_n)^\top$ and that the total log-likelihood function $\ell(\theta)$ is a function of the $p \times 1$ parameter vector θ. Further, we assume that $\ell = \ell(\theta)$ is regular (Cox and Hinkley 1974) with respect to all θ derivatives up to and including those of fourth order. We assume that the MLE $\hat{\theta}$ of θ is a consistent solution of the non-linear equations $\hat{U}_r = 0$ for $r = 1, \ldots, p$. In what follows, we shall use the Einstein summation convention, which is useful for dealing with coordinate formulae. By

expanding $\hat{U}_r = 0$ around θ, we obtain

$$U_r + U_{rs}(\hat{\theta}_s - \theta_s) + \frac{1}{2}U_{rst}(\hat{\theta}_s - \theta_s)(\hat{\theta}_t - \theta_t) + \frac{1}{6}U_{rstu}(\hat{\theta}_s - \theta_s)(\hat{\theta}_t - \theta_t)(\hat{\theta}_u - \theta_u) + \cdots .$$

Let $-U^{rs}$ denotes the (r, s) element of the inverse observed information matrix $-U_{rs}$. By inverting the previous expansion, we obtain

$$\begin{aligned}\hat{\theta}_r - \theta_r = &-U^{rs}U_s - \frac{1}{2}U^{rs}U^{tu}U^{vw}U_{stv}U_uU_w + \frac{1}{6}U^{rs}U^{tu}U^{vw}U^{xy}(U_{suyw} \\ &- 3U^{pq}U_{swp}U_{quy})U_tU_vU_x + \cdots . \end{aligned} \tag{2.1}$$

The quantity $-U^{rs}$ generally exists and admits the following expansion in terms of the (r, s)th element $-\kappa^{rs}$ of the inverse information matrix:

$$U^{rs} = -\kappa^{rs} + \kappa^{rt}\kappa^{su}(U_{tu} - \kappa tu) - \kappa^{rt}\kappa^{su}\kappa^{vw}(U_{tv} - \kappa_{tv})(U_{uw} - \kappa_{uw}) + \cdots . \tag{2.2}$$

We now consider the expansion of $\ell(\hat{\theta}) - \ell(\theta)$ given by

$$\begin{aligned}\ell(\hat{\theta}) - \ell(\theta) = &U_r(\hat{\theta}_r - \theta_r) + \frac{1}{2}U_{rs}(\hat{\theta}_r - \theta_r)(\hat{\theta}_s - \theta_s) \\ &+ \frac{1}{6}U_{rst}(\hat{\theta}_r - \theta_r)(\hat{\theta}_s - \theta_s)(\hat{\theta}_t - \theta_t) \\ &+ \frac{1}{24}U_{rstu}(\hat{\theta}_r - \theta_r)(\hat{\theta}_s - \theta_s)(\hat{\theta}_t - \theta_t)(\hat{\theta}_u - \theta_u) + \cdots . \end{aligned} \tag{2.3}$$

By inserting Eqs. (2.1) and (2.2) into (2.3) and rearranging terms according to their asymptotic orders, we obtain, after lengthy algebra, the expected value of $2[\ell(\hat{\theta}) - \ell(\theta)]$ to order $\mathcal{O}(n^{-1})$, where $\ell(\theta)$ is the log likelihood at the true parameter value. This result was first derived by Lawley (1956). His expansion can be expressed as $2\mathbb{E}[\ell(\hat{\theta}) - \ell(\theta)] = p + \varepsilon_p + \mathcal{O}(n^{-2})$, where ε_p, which is of order $\mathcal{O}(n^{-1})$, is given by

$$\varepsilon_p = \sum{}' (\lambda_{rstu} - \lambda_{rstuvw}), \tag{2.4}$$

where

$$\lambda_{rstu} = \kappa^{rs}\kappa^{tu}\left\{\frac{\kappa_{rstu}}{4} - \kappa^{(u)}_{rst} + \kappa^{(su)}_{rt}\right\} \tag{2.5}$$

and

$$\begin{aligned}\lambda_{rstuvw} = &\kappa^{rs}\kappa^{tu}\kappa^{vw}\left\{\kappa_{rtv}\left(\frac{\kappa_{suw}}{6} - \kappa^{(u)}_{sw}\right) + \kappa_{rtu}\left(\frac{\kappa_{svw}}{4} - \kappa^{(v)}_{sw}\right)\right. \\ &\left. + \kappa^{(v)}_{rt}\kappa^{(u)}_{sw} + \kappa^{(u)}_{rt}\kappa^{(v)}_{sw}\right\}. \end{aligned} \tag{2.6}$$

Here, $\sum'$ denotes summation over all components of θ, i.e., the indices r, s, t, u, v, and w vary over all p parameters. All individual terms in the sums in (2.5) and (2.6) are of order $\mathcal{O}(n^{-1})$. The main difficulty with these sums is that the individual terms are not invariants, and therefore, they have no geometrical interpretation independent of the coordinate system chosen.

Lawley (1956) showed that the rth cumulant of $2[\ell(\hat{\theta}) - \ell(\theta)]$, say τ_r, can be expressed as

$$\tau_r = 2^{r-1}\,(r-1)!\,p\left(1+\frac{\varepsilon_p}{p}\right)^r + \mathcal{O}(n^{-2}). \tag{2.7}$$

The leading term in (2.7) is the rth cumulant of the χ^2_p distribution. So, all cumulants of $2\,[\ell(\hat{\theta}) - \ell(\theta)]$ may be matched with those of the appropriate χ^2_p random variable as far as terms of order $\mathcal{O}(n^{-1})$. The matching of cumulants in (2.7) is more obviously appropriate for continuous than discrete random variables.

Several papers have focused on deriving matrix formulae for Bartlett corrections in general classes of regression models based on Eqs. (2.4)–(2.6). Sharp (1975) used these equations to obtain corrections for testing the following hypotheses in Markov chains: that the transition probabilities are stable over time, that the chain is of a given order, and that several samples come from the same chain. Sharp's results cover most of the tests on Markov parameters used in practice. Williams (1976) derived Bartlett correction factors for log-linear models in complete multidimensional tables with closed-form estimators by expanding the LR criterion in a Taylor series instead of using these equations. Cordeiro (1983, 1987) obtained Bartlett corrections for generalized linear models (GLMs) when the dispersion parameter is known and unknown, respectively.

Barndorff-Nielsen and Cox (1984) gave an indirect method for computing Bartlett corrections under rather general parametric models by establishing a simple connection between the correction term b and the normalizing constants of the general expression for the conditional distribution of the MLE, namely $b = (A_0/A)^q\,(n/2\pi)$, where A and A_0 are the normalizing constants of the general formula for the density of the MLE conditional on an exact or approximate ancillary statistic when this formula is applied to the unrestricted and null (restricted) models, respectively. It is usually easier to obtain the Bartlett correction for special cases using Lawley's formula than using Barndorff-Nielsen and Cox's expression, since the former involves only moments of log-likelihood derivatives, whereas the latter requires exact or approximate computation of the conditional distribution of the MLE. When there are many nuisance parameters, it may not be easy to obtain ancillary statistics for these parameters, and hence, the evaluation of Barndorff-Nielsen and Cox's formula can be quite cumbersome. The constants A_0 and A are usually functions of the maximal ancillary statistic, although to the relevant order of magnitude, w^* is independent of the ancillary statistic selected. The authors have also obtained various expressions for these quantities and, in particular, an approximation that does not require integration over the sample space for the one-parameter case.

Since the statistic w is invariant under reparameterization, it is possible to obtain large-sample expansions for it and for its expectation in terms of invariants. McCullagh and Cox (1986) used this fact to represent the Bartlett correction as a function of invariant combinations of cumulants of the first two log-likelihood derivatives and gave it a geometric interpretation in full generality in terms of the model curvature. It is also noteworthy that McCullagh and Cox's (1986) formula is in agreement with Lawley's (1956) formula. The advantage of McCullagh and Cox's formula lies in its geometric interpretation, whereas the main advantage of Lawley's result is that it can be more easily implemented to obtain Bartlett corrections for special models.

Considerable attention in the literature has been given to the computation of Bartlett corrections, both using alternative methods to Lawley's formula and by means of simpler formulas for specific models; see, for example, the references in Cribari-Neto and Cordeiro (1996).

2.4 Bartlett-Corrected Likelihood Ratio Tests

Consider a parametric model $f(y;\theta)$, whose probability or density function is indexed by the parameter vector $\theta = (\psi^\top, \lambda^\top)^\top$, with $\dim(\psi) = q$ and $\dim(\lambda) = p - q$ for $q < p$. The interest lies in testing the composite null hypothesis $H_0 : \psi = \psi^{(0)}$ against the two-sided alternative hypothesis $H : \psi \neq \psi^{(0)}$, where λ is a vector of nuisance parameters. The LR statistic w is defined as

$$w = 2[\ell(\hat{\psi}, \hat{\lambda}) - \ell(\psi^{(0)}, \tilde{\lambda})],$$

where $\hat{\psi}$ and $\hat{\lambda}$ are the MLEs of ψ and λ under the alternative hypothesis and $\tilde{\lambda}$ is the restricted MLE of λ subject to $\psi = \psi^{(0)}$.

The expected value of w can be expressed as

$$\mathbb{E}(w) = 2\mathbb{E}[\ell(\hat{\psi}, \hat{\lambda}) - \ell(\psi, \lambda)] - 2\mathbb{E}[\ell(\psi^{(0)}, \tilde{\lambda}) - \ell(\psi, \lambda)],$$

and then using (2.4) it follows that

$$\mathbb{E}(w) = q + \varepsilon_p - \varepsilon_{p-q}, \tag{2.8}$$

where ε_p is obtained from Eqs. (2.5) and (2.6) by summing over the parameters in ψ and λ. The term ε_{p-q} is calculated analogously, the only difference being that the summations run over the parameters in λ since ψ is fixed at $\psi^{(0)}$. For further details, see Lawley (1956) and Cordeiro (1993a).

A further step on the improvement of the statistic w was taken by Hayakawa (1977), who derived an asymptotic expansion for the null distribution of w to order $\mathcal{O}(n^{-1})$, under the null hypothesis $H_0 : \psi = \psi^{(0)}$, given by

$$\Pr(w \leq z) = F_q(z) + \frac{1}{24}\left[A_2 F_{q+4}(z) - (2A_2 - A_1)F_{q+2}(z) + (A_2 - A_1)F_q(z)\right], \tag{2.9}$$

where $F_q(\cdot)$ is the cumulative distribution function (cdf) of a χ^2 random variable with q degrees of freedom and the quantities A_1 and A_2 are of order $\mathcal{O}(n^{-1})$. Here, A_1 is a function of expected values of the first four log-likelihood derivatives and of the first two derivatives of these expected values with respect to the model parameters. When there are nuisance parameters, A_1 can be determined as the difference between two functions identical to (2.4), evaluated under the null and alternative hypotheses. The error in Eq. (2.9) is $\mathcal{O}(n^{-2})$ and not $\mathcal{O}(n^{-3/2})$ as it is sometimes reported. Recall, however, that the Bartlett correction factor is given by $c = 1 + (12q)^{-1}A_1$, which differs from the one that follows from (2.9) unless $A_2 = 0$. This points to a conflict between Hayakawa's and Lawley's results. This puzzle was solved by Harris (1986) and Cordeiro (1987). Harris showed that A_2 should not be present in (2.9), whereas Cordeiro showed that A_2 always equals zero; see also Chesher and Smith (1995). The main contribution of Eq. (2.9) with $A_2 = 0$ is that it provides a relatively simple proof that $w^* = w/c$ has a χ^2_q null distribution with error $\mathcal{O}(n^{-2})$. In fact, Cordeiro (1987) demonstrated that the simple correction of the first moment of w to order $\mathcal{O}(n^{-1})$ causes the removal of the term of the same order in the asymptotic expansion of the corrected statistic w^*. This result was a starting point for numerous subsequent research efforts in the direction of establishing several explicit expressions for Bartlett corrections in various classes of statistical models. Let $f_q(\cdot)$ be the density function of a χ^2_q distribution. Differentiation of (2.9) with $A_2 = 0$ yields the density function of w to order $\mathcal{O}(n^{-1})$, which is given by

$$f_w(x) = f_q(x)\left\{1 + \frac{b}{2}\left(\frac{x}{q} - 1\right)\right\}, \tag{2.10}$$

where $b = b(\psi^{(0)}, \lambda) = A_1/12 = \varepsilon_p - \varepsilon_{p-q}$ is evidently a quantity of order $\mathcal{O}(n^{-1})$ to be estimated under the null hypothesis H_0. A general formula for the constant b can be obtained using Eqs. (2.4), (2.5), and (2.6).

Clearly, $f_w(\cdot)$ depends only on the dimension of ψ, on the reference density function $f_q(\cdot)$ and on the term of order $\mathcal{O}(n^{-1})$ in the expected value of w. Using Eq. (2.10), it is possible to show that the null density function of the modified statistic $w^* = w/(1 + b/q)$ or $w(1 - b/q)$, up to terms of order $\mathcal{O}(n^{-1})$, is $f_{w^*}(x) = f_q(x)$. Hence, $\Pr(w \leq x) = F_q(x) + \mathcal{O}(n^{-2})$, whereas $\Pr(w^* \leq x) = F_q(x) + \mathcal{O}(n^{-1})$. In other words, the error of the χ^2_q approximation to the null distribution of w is of order $\mathcal{O}(n^{-1})$, which is reduced to order $\mathcal{O}(n^{-2})$ when the limiting χ^2 distribution is used to approximate the null distribution of w^*. Thus, the modified statistic w^*

has a χ^2_q null distribution, except for terms of order $\mathcal{O}(n^{-2})$ as first suggested by Lawley's Eq. (2.7). In general, convergence of the cumulants implies convergence in distribution, because the asymptotic cumulants uniquely determine a distribution.

In practice, we can obtain the Bartlett correction from the joint cumulants of the log-likelihood derivatives. Such cumulants can, however, be cumbersome in some statistical models. In certain regressions models, they are invariant under permutation of parameters and that fact considerably simplifies the computations.

When testing a simple null hypothesis $H_0 : \theta = \theta^{(0)}$ against a simple composite hypothesis $H: \theta \neq \theta^{(0)}$, all indices are equal to θ, and the Bartlett correction simplifies to

$$\varepsilon_1 = \kappa^{\theta\theta^2}\{\kappa_{\theta\theta\theta}/4 - \kappa^{(\theta)}_{\theta\theta\theta} + \kappa^{(\theta\theta)}_{\theta\theta}\} - \kappa^{\theta\theta^3}\{\kappa_{\theta\theta\theta}(5\kappa_{\theta\theta\theta}/12 - 2\kappa^{(\theta)}_{\theta\theta}) + 2\kappa^{(\theta)^2}_{\theta\theta}\}. \tag{2.11}$$

The corrected LR statistic $w^* = w/(1 + \varepsilon_1)$ is χ^2_1 distributed under the null hypothesis to order $\mathcal{O}(n^{-1})$. An important non-regression case is that of the one-parameter exponential family model. A simple closed-form Bartlett correction for testing the null hypothesis that its parameter equals a given scalar was obtained by Cordeiro et al. (1995). They then applied their result to a number of distributions in the exponential family, some of which are widely used in empirical applications in a variety of fields.

We now provide three simple examples. First, we consider n i.i.d. observations from the exponential distribution with mean μ. The log-likelihood function is $\ell(\mu) = -n\,\log(\mu) - n\overline{y}/\mu$, where $\overline{y}$ is the sample mean. The LR statistic for testing $H_0 : \mu = \mu^{(0)}$ against $H : \mu \neq \mu^{(0)}$ is $w = 2n\{\overline{Y}\log(\overline{Y}/\mu^{(0)}) - (\overline{Y} - \mu^{(0)})\}$. The cumulants are $\kappa_{\mu,\mu} = n/\mu^2$, $\kappa_{\mu,\mu,\mu} = -\kappa_{\mu,\mu\mu} = 2n/\mu^3$, $\kappa_{\mu\mu\mu} = 4n/\mu^3$, $\kappa_{\mu\mu\mu\mu} = -30n/\mu^4$, $\kappa_{\mu,\mu\mu\mu} = 18n/\mu^4$, etc. By plugging these cumulants into (2.11), we obtain the Bartlett correction $c = 1 + 1/(6n\mu^{(0)})$.

Next, we take n i.i.d. observations from the normal distribution $N(\mu, \sigma^2)$. The log-likelihood function $\ell = \ell(\theta)$ for $\theta = (\mu, \sigma^2)^T$ reduces to

$$\ell = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2.$$

For testing $H_1 : \mu = \mu^{(0)}$ against $A_1 : \mu \neq \mu^{(0)}$ (σ^2 unknown) and $H_2 : \sigma^2 = \sigma^{(0)^2}$ against $A_2 : \sigma^2 \neq \sigma^{(0)^2}$ (μ unknown), the LR statistics reduce to

$$w_1 = 2\{\ell(\hat{\mu}, \hat{\sigma}^2) - \ell(\mu^{(0)}, \tilde{\sigma}^2)\} = n\log\left\{\frac{\Sigma(Y_i - \mu^{(0)})^2}{\Sigma(Y_i - \overline{Y})^2}\right\}$$

and

$$w_2 = 2\{\ell(\hat{\mu}, \hat{\sigma}^2) - \ell(\tilde{\mu}, \sigma^{(0)^2})\} = n\left[\log\left(\frac{\sigma^{(0)^2}}{\hat{\sigma}^2}\right) + \frac{\hat{\sigma}^2 - \sigma^{(0)^2}}{\sigma^{(0)^2}}\right],$$

respectively, where $\hat{\mu} = \tilde{\mu} = \overline{Y}/n$, $\hat{\sigma}^2 = \Sigma(Y_i - \overline{Y})^2/n$ and $\tilde{\sigma}^2 = \Sigma(Y_i - \mu^{(0)})^2/n$.

The cumulants κ's required for computing the Bartlett corrections are $\kappa_{\mu\mu} = -n/\sigma^2$, $\kappa_{\sigma^2\sigma^2} = -n/2\sigma^4$, $\kappa_{\mu\sigma^2} = 0$, $\kappa_{\mu,\mu,\mu} = \kappa_{\mu,\mu\mu} = \kappa_{\mu\mu\mu} = 0$, $\kappa_{\sigma^2,\sigma^2,\sigma^2} = -\kappa_{\sigma^2,\sigma^2\sigma^2} = n/\sigma^6$, $\kappa_{\sigma^2\sigma^2\sigma^2} = 2n/\sigma^6$, $\kappa_{\mu,\mu\sigma^2} = -\kappa_{\mu\mu\sigma^2} = -n/\sigma^4$, $\kappa_{\mu,\mu,\sigma^2} = 3n/\sigma^4$, $\kappa_{\mu\mu\sigma^2\sigma^2} = -2n/\sigma^6$, etc. Several of them are obtained using Bartlett identities. From Eqs. (2.4)–(2.6), we have

$$\mathbb{E}(w_1) = 1 + \sum_{\mu,\sigma^2}(\ell_{rstu} - \ell_{rstuvw}) - (\ell_{\sigma^2\sigma^2\sigma^2\sigma^2} - \ell_{\sigma^2\sigma^2\sigma^2\sigma^2\sigma^2\sigma^2})$$

and

$$\mathbb{E}(w_2) = 1 + \sum_{\mu,\sigma^2}(\ell_{rstu} - \ell_{rstuvw}) - (\ell_{\mu\mu\mu\mu} - \ell_{\mu\mu\mu\mu\mu\mu}).$$

After some algebra, we obtain

$$\mathbb{E}(w_1) = 1 + \frac{3}{2n} \quad \text{and} \quad \mathbb{E}(w_2) = 1 + \frac{11}{6n}.$$

Thus, the modified LR statistics are $w_1^* = w_1/(1 + 3/2n)$ and $w_2^* = w_2/(1 + 11/6n)$ (for testing H_1 and H_2, respectively). The corrections can also be obtained from first principles by noting that $n\hat{\sigma}^2/\sigma^2 \sim \chi^2_n$ and $n\tilde{\sigma}^2/\sigma^2 \sim \chi^2_{n-1}$ and then approximating $\mathbb{E}[\log(\chi^2_n)]$ by $\log(n) - n^{-1}$.

Bartlett corrections represent an important area of research in asymptotic theory because of their widely applicability. Corrected LR statistics for exponential family non-linear models were obtained by Cordeiro and Paula (1989). They gave general matrix expressions for Bartlett corrections in these models involving an unpleasant looking quantity which may be regarded as a measure of non-linearity of the model systematic component. Attfield (1991) and Cordeiro (1993b) have shown how to correct LR statistics used in heteroscedasticity tests. Computer code for calculating Bartlett corrections was developed by Andrews and Stafford (1993). Cordeiro et al. (1994) derived matrix formulas for Bartlett corrections in dispersion models, thus extending previous results by Cordeiro (1983) and Cordeiro and Paula (1989). Cordeiro (1995) presented extensive simulation results on the performance of the corrected statistic w^* in GLMs with focus on gamma and log-linear models. Zucker et al. (2000) obtained Bartlett correction formulas for LR tests for the regression parameters in the general mixed model and investigated the performance of the Bartlett-corrected tests. More recently, using Eqs. (2.4)–(2.6), Giersbergen (2009) derived Bartlett corrections for testing hypotheses on the autoregressive parameter in the stable AR(1) model, in the AR(1) model with intercept and in the AR(1) model with intercept and linear trend. Melo et al. (2009) addressed the issue of improving

LR tests in mixed linear models. For a detailed account of the applicability of Bartlett corrections, see Cribari-Neto and Cordeiro (1996).

2.5 Generalized Linear Models

The class of GLMs is particularly useful for fitting non-normal models, typically by the ML method. A wide variety of models can be studied within the framework of GLMs when the classical assumptions of normal theory are violated. The unified theory of these models, including a general algorithm for computing the MLEs, is extremely important for data analysis. This class of models is based on the exponential family. The use of GLMs has become very common in recent years, and it is thus useful to develop second-order asymptotic theory for inference and diagnostics. The statistical analysis of such models is generally based on the asymptotic properties of the MLEs. Standard references on GLMs are McCullagh and Nelder (1989) and Dobson and Barnett (1998).

In these models, the random variables $Y_1, \ldots, Y_n$ are assumed to be independent, each Y_i having distribution in the linear exponential family given by

$$\pi(y; \theta_i, \phi) = \exp\{\phi\,[y\,\theta_i - b(\theta_i) + a(y)] + c(y, \phi)\}, \tag{2.12}$$

where $b(\cdot)$ and $c(\cdot, \cdot)$ are known appropriate functions. The parameter ϕ is said to be the precision parameter and is assumed constant throughout the observations. Let $\sigma^2 = \phi^{-1}$ be the dispersion parameter. If Y is continuous π is assumed to be a density with respect to the Lebesgue measure, whereas if Y is discrete π is assumed to be a density with respect to the counting measure. Several important distributions are special cases of exponential family models.

The mean and variance of Y_i are $\mathbb{E}(Y_i) = \mu_i = db(\theta_i)/d\theta_i$ and $\mathrm{Var}(Y_i) = \phi^{-1}V_i$, where $V = d\mu/d\theta$ is the variance function. The parameter $\theta = \int V^{-1}d\mu = q(\mu)$ is a known one-to-one function of μ. The exponential family model (2.12) is uniquely characterized by its variance function V, which plays a key role in the study of its mathematical properties and in estimation. For gamma models, the dispersion parameter σ^2 is the reciprocal of the index, whereas for normal and inverse Gaussian models, σ^2 is the variance and $\mathrm{Var}(Y)/\mathbb{E}(Y)^3$, respectively. If the distribution of Y involves only one unknown parameter, as in the binomial and Poisson models, then ϕ can be taken to be equal to one.

For two-parameter full exponential family distributions with canonical parameters ϕ and $\phi\,\theta$, the decomposition $c(y, \phi) = d_1(\phi) + d_2(y)$ holds. Here, $d_1(\phi) = \log(\phi)/2$ and $d_2(y) = -\log(2\pi)/2$ for the normal distribution with variance ϕ^{-1}, $d_1(\phi) = \phi\log(\phi) - \log[\Gamma(\phi)]$ and $d_2(y) = -y$ for the gamma distribution with index ϕ and $d_1(\phi) = \log(\phi)/2$ and $d_2(y) = -\log(2\pi y^3)/2$ for the inverse Gaussian distribution with $\phi = \mathbb{E}(Y)^3/\mathrm{Var}(Y)$, where $\Gamma(\cdot)$ is the gamma function.

A GLM is defined by the family of distributions in (2.12) and by the systematic component $g(\mu) = \eta = X\beta$, where $g(\cdot)$ is a known one-to-one continuously

twice-differentiable function, X is a specified $n \times p$ model matrix of full rank p ($p < n$) and $\beta = (\beta_1, \ldots, \beta_p)^\top$ is a set of unknown linear parameters. The link function defined by $\theta = \eta$ is known as the canonical link function. The canonical link functions for the most commonly used distributions are: normal $g(\mu) = \mu$, Poisson $g(\mu) = \log(\mu)$, gamma $g(\mu) = -\mu^{-1}$, binomial $g(\mu) = \log[\mu/(1-\mu)]$, and inverse Gaussian $g(\mu) = -\mu^{-2}$.

Denote the n observations by $y_1, \ldots, y_n$, the total log-likelihood for β by $\ell = \ell(\beta)$ and the MLE of β by $\hat{\beta}$. The information matrix for β is given by $K = \{-\kappa_{rs}\} = \phi\,(X^\top W X)$, where $W = \text{diag}\{w_i\}$ and $w_i = V_i^{-1}\,(d\mu_i/d\eta_i)^2$. Since X has full rank and w and ϕ are positive, the information matrix is positive definite and so is its inverse, $K^{-1} = \{-\kappa^{rs}\} = \phi^{-1}\,(X^\top W X)^{-1}$. Let $\hat{\eta} = X\,\hat{\beta}$ and $\hat{\mu} = g^{-1}(\hat{\eta})$ be the MLEs of η and μ, respectively. The precision parameter ϕ does not enter into the estimating equations $(X^\top \hat{W}\, X)\,\hat{\beta} = X^\top \hat{W}\,\hat{z}$, which have the form of a linear-weighted least-squares regression with weight matrix given by W. The dependent variable in the weighted regression is $z = (z_1, \ldots, z_n)^\top$, where $z_i = \eta_i + (y_i - \mu_i) d\eta_i/d\mu_i$, $i = 1, \ldots, n$.

For two-parameter exponential models when ϕ is unknown, we have $c(y, \phi) = d_1(\phi) + d_2(y)$. Hence, the MLE $\hat{\phi}$ of ϕ can be obtained as the solution of

$$2n\, d_1'(\hat{\phi}) + 2\sum_{i=1}^{n}[v(y_i) + a(y_i)] = D_p(y, \hat{\mu}),$$

where

$$D_p(y, \hat{\mu}) = 2\sum_{i=1}^{n}\{v(y_i) - v(\hat{\mu}_i) + (\hat{\mu}_i - y_i)\, q(\hat{\mu}_i)\}$$

is the deviance of the model, $v(\mu) = \mu\, q(\mu) - b(q(\mu))$ and primes here denote derivatives with respect to ϕ. The deviance can be computed from the observations and from the MLEs $\hat{\mu}_1, \ldots, \hat{\mu}_n$. Thus, $\hat{\phi}$ is a function of the model deviance.

2.5.1 Bartlett Correction

In what follows, dashes denote derivatives of the mean with respect to the linear predictor. So, $\mu' = d\mu/d\eta$, $\mu'' = d^2\mu/d\eta^2$, etc. Further, let $V^{(r)} = d^r V/d\mu^r$ for $r = 1, 2$. We introduce the scalars $f = V^{-1}\mu'\mu''$, $g = V^{-1}\mu'\mu'' - V^{-2}V^{(1)}\mu'^3$ and

$$h = V^{-1}\mu''\left(\mu'' - 4wV^{(1)}\right) + w^2\left(2V^{-1}V^{(1)2} - V^{(2)}\right),$$

and the corresponding diagonal matrices $F = \text{diag}\{f_1, \ldots, f_n\}$, $G = \text{diag}\{g_1, \ldots, g_n\}$, and $H = \text{diag}\{h_1, \ldots, h_n\}$. We define the $n \times n$ positive semi-definite matrix $Z = \{z_{ij}\} = X\,(X^\top W X)^{-1}\,X^\top$ of rank p which is, apart from the multiplier ϕ^{-1}, the asymptotic covariance matrix of the estimators $\hat{\eta}_1, \ldots, \hat{\eta}_n$ of the linear

predictors. Additionally, $Z_d = \text{diag}\{z_{11}, \ldots, z_{nn}\}$ is a diagonal matrix with the diagonal elements of Z, $Z^3 = \{z_{ij}^3\}$ and $\mathbf{1}$ is an $n \times 1$ vector of ones. The joint cumulants corresponding to the β components in Eqs. (2.5) and (2.6) can be easily derived. Some of them are $\kappa_{rs} = E(\partial^2\ell/\partial\beta_r\partial\beta_s) = -\phi\sum_{i=1}^n w_i\, x_{ir}x_{is}$, $\kappa_{rst} = E(\partial^3\ell/\partial\beta_r\partial\beta_s\partial\beta_t) = -\phi\sum_{i=1}^n(f_i + 2g_i)\, x_{ir}x_{is}x_{it}$, $\frac{1}{4}\kappa_{rstu} - \kappa_{rst}^{(u)} + \kappa_{rt}^{(su)} = \frac{\phi}{4}\sum_{i=1}^n h_i\, x_{ir}x_{is}x_{it}x_{iu}$, and so on. All κ's refer to a total over the sample and are, in general, of order n. For GLMs, these cumulants are invariant under permutation of parameters, for example, $\kappa_{rs,t} = \kappa_{rt,s} = \kappa_{st,r}$ holds but not in general models.

For a GLM, we can easily obtain the cumulants κ's. The key to obtain a simple expression for the Bartlett correction in GLMs is that the log-likelihood derivatives are linear functions of y and the invariance of the cumulants κ's under permutation of the β parameters. Let ε_p be the $\mathcal{O}(n^{-1})$ term in the expected value of $2[\ell(\hat{\beta}) - \ell(\beta)]$. Plugging the expressions for κ^{rs}, κ_{rst}, κ_{rstu}, $\kappa_{rs}^{(t)}$, $\kappa_{rs}^{(tu)}$, and $\kappa_{rst}^{(u)}$ in Eqs. (2.4)–(2.6), carrying out the sums over the sample after evaluating the sums over the parameters, Cordeiro (1983) obtained a simple matrix formula for ε_p given by

$$\varepsilon_p = \varepsilon_p(\phi, X, \mu) = \frac{1}{4\phi}\,\text{tr}(\text{H}\,\text{Z}_\text{d}^2) - \frac{1}{3\phi}\,\mathbf{1}^\top\,\text{G}\,\text{Z}^{(3)}\,(\text{F} + \text{G})\,\mathbf{1} + \frac{1}{12\phi}\,\mathbf{1}^\top\,F(2Z^{(3)} + 3Z_d Z Z_d)\,F\,\mathbf{1}, \tag{2.13}$$

where tr is the trace operator. Equation (2.13) depends only on the model matrix X, the precision parameter ϕ and the variance and link functions with their first and second derivatives. Equation (2.13) only involves simple operations on matrices and vectors and can be easily applied in practice using a computer algebra system such as MATHEMATICA or MAPLE, or using a programming language with support for matrix operations, such as OX or R. Numerical computation involving higher-order joint cumulants as in Eqs. (2.5) and (2.6) is thereby avoided. For GLMs with closed-form expressions for Z, it is possible to obtain simpler expressions for ε_p. This formula for ε_p is very important to derive corrected LR tests for these models. It can be applied to several special cases of GLMs as discussed by Cordeiro (1983, 1987).

Consider now a partition of the $p \times 1$ vector $\beta = (\beta_1^\top, \beta_2^\top)^\top$ of the linear parameters of the GLM, where $\beta_1 = (\beta_1, \ldots, \beta_q)^\top$ and $\beta_2 = (\beta_{q+1}, \ldots, \beta_p)^\top$ for $q \leq p$, and an induced partition of the model matrix as $X = (X_1, X_2)$. Our interest is in testing the composite null hypothesis $H_0 : \beta_1 = \beta_1^{(0)}$ against a two-sided alternative hypothesis, where $\beta_1^{(0)}$ is a vector of known constants. Usually, $\beta_1^{(0)} = 0$. Assume for the moment that ϕ is known. The LR statistic for testing H_0 reduces to $w = 2\{\ell(\hat{\beta}_1, \hat{\beta}_2) - \ell(\beta_1^{(0)}, \tilde{\beta}_2)\}$, where $\tilde{\beta}_2$ is the MLE of β_2 restricted to $\beta_1 = \beta_1^{(0)}$. Using (2.13), we can write $\mathbb{E}(w) = q + \varepsilon_p - \varepsilon_{p-q} + \mathcal{O}(n^{-2})$, where $\varepsilon_{p-q} = \varepsilon_{p-q}(\phi, X_2, \mu)$ can be determined from this equation with X_2 in place of X. Therefore, the corrected LR statistic is defined by $w^* = w/c$, where

$$c = 1 + \frac{\varepsilon_p - \varepsilon_{p-q}}{q}. \tag{2.14}$$

After forming the matrices $Z = X\,(X^\top W X)^{-1}\,X^\top$ and $Z_2 = X_2\,(X_2^\top W X_2)^{-1}\,X_2^\top$, it is then straightforward to evaluate ε_p and ε_{p-q}. A possible motivation for developing simple formulae for Bartlett corrections to LR statistics is that these formulae can reveal which aspects of the model contribute to the quality of the first-order χ^2 approximation.

For calculating the Bartlett corrections in GLMs when ϕ is unknown, we have to take into account the joint cumulants between the components of β and ϕ. In this case, for testing $H_0 : \beta_1 = \beta_1^{(0)}$, Cordeiro (1987) demonstrated, after intensive algebraic developments, that the Bartlett correction has an extra quantity and it reduces to

$$c = 1 + \frac{\varepsilon_p - \varepsilon_{p-q}}{q} + \frac{2[\phi d_1'''(\phi) + d_1''(\phi)] - (p+q)d_1''(\phi)}{4n\phi^2 d_1''(\phi)^2}. \tag{2.15}$$

Finally, we consider the composite null hypothesis $H_0 : \phi = \phi^{(0)}$ against the alternative $H : \phi \neq \phi^{(0)}$, where now β denotes a vector of nuisance parameters. For testing $H_0 : \phi = \phi^{(0)}$, the LR statistic is $w = 2[\ell(\hat{\phi}) - \ell(\phi^{(0)})]$ and the Bartlett correction, which is equal to the expected value of w to order $\mathcal{O}(n^{-1})$, is given by Cordeiro (1987)

$$c = 1 + \frac{1}{2nd_1''(\phi)^2}\left\{\frac{d_1''''(\phi)}{2} - \frac{5d_1'''(\phi)^2}{6d_1''(\phi)} + \left[\frac{\phi d_1'''(\phi) + d_1''(\phi)}{\phi^2}\right] p - \frac{d_1''(\phi)}{2\phi^2}\, p^2\right\}.$$

For the normal model with variance σ^2 and for the inverse Gaussian model with precision parameter ϕ, $d_1(\phi) = \log(\phi)/2$, which yields $c = \mathbb{E}(w) = 1 + (6n)^{-1}(3p^2 + 6p + 2)$.

2.5.2 Special Models

Equation (2.13) can be simplified for several important special models as shown in Cordeiro (1983, 1987). Normal models apply to data with constant variance over the entire range of parameter values. For the normal model ($\theta = \mu$, $\phi = \sigma^{-2}$), we obtain

$$\varepsilon_p = \frac{\sigma^2}{4}\left[\mathrm{tr}(\mathrm{H}\,\mathrm{Z}_{\mathrm{d}}^2) + \mathbf{1}^\top \mathrm{F}(\mathrm{Z}_{\mathrm{d}}\,\mathrm{Z}\,\mathrm{Z}_{\mathrm{d}} - 2\mathrm{Z}^{(3)})\,\mathrm{F}\,\mathbf{1}\right],$$

where $W = \mathrm{diag}\{\mu'^2\}$, $F = \mathrm{diag}\{\mu'\mu''\}$, and $H = \mathrm{diag}\{\mu''^2\}$. Log-linear models are appropriate for analyzing count data. For the log-linear model ($\theta = \log(\mu)$, $\phi = 1$), it follows that

$$\varepsilon_p = -\frac{1}{4}\,\mathrm{tr}(\mathrm{H}\,\mathrm{Z}_{\mathrm{d}}^2) + \frac{1}{6}\,\mathbf{1}^\top \mathrm{W}\,\mathrm{Z}^{(3)}\,\mathrm{W}\,\mathbf{1} + \frac{1}{4}\,\mathbf{1}^\top \mathrm{W}\,\mathrm{Z}_{\mathrm{d}}\,\mathrm{Z}\,\mathrm{Z}_{\mathrm{d}}\,\mathrm{W}\,\mathbf{1},$$

where $W = \mathrm{diag}\{\mu\}$. Gamma models are widely used for data, including continuous measurements as well as discrete data, with constant coefficient of variation. For the gamma model ($\theta = -\mu^{-1}$) with power link function $\eta = \mu^\alpha$ ($\alpha = 0$ interpreted as the logarithm link), we obtain

$$\varepsilon_p = \frac{(\alpha^2 - 6\alpha - 1)}{4\phi}\,\mathrm{tr}(\mathrm{H}\,\mathrm{Z}_{\mathrm{d}}^2) - \frac{(3\alpha^2 + 6\alpha - 1)}{6\phi}\,\mathbf{1}^\top \mathrm{W}^{3/2}\,\mathrm{Z}^{(3)}\,\mathrm{W}^{3/2}\,\mathbf{1} + \frac{(\alpha - 1)^2}{4\phi}\,\mathbf{1}^\top W^{3/2}\,Z_d\,Z\,Z_d\,W^{3/2}\,\mathbf{1},$$

where $W = \mathrm{diag}\{\mu^{-2\alpha}\}$.

Suppose now that p populations follow the density given in (2.12) and that independent random samples of sizes $n_1, \ldots, n_p$ ($n_i \geq 1, i = 1, \ldots, n$) are taken from such populations. In each population, the observations have the same dispersion parameter, which may be unknown. The vector of responses is written as $Y = (Y_{11}, \ldots, Y_{1n_1}, \ldots, Y_{p1}, \ldots, Y_{pn_p})^\top$, and the linear structure is given by $\eta_i = \beta + \beta_i$ for $i = 1, \ldots, n$, where β is the overall mean and β_i is the effect on the response of the ith population. Here, $\sum \beta_i = 0$ and η is functionally related to $\mu = \mathbb{E}(Y)$. It can be shown that the form of $X^\top W X$ in the general one-way classification model is $X^\top W X = \mathrm{diag}\{n_i w_i\}$ and that $\mathrm{rank}(X^\top W X) = p$. The matrix Z has order $\sum n_i$ with typical element $\delta_{ij}(n_i w_i)^{-1}$, where $\delta_{ij} = 1$ if i and j index observations in the same population and zero otherwise. Let f_i, g_i, and h_i be the functions defined in the matrices F, G, and H, respectively, for the ith population. We can obtain $Z^{(3)} = Z_d\,Z\,Z_d = \{\delta_{ij}(n_i w_i)^{-3}\}$ and then, for example, $\mathrm{tr}(H\,Z_d^2) = \sum n_i^{-1} h_i w_i^{-2}$, $\mathbf{1}^\top G\,Z^{(3)}\,(F + G)\,\mathbf{1} = \sum n_i^{-1} g_i (f_i + g_i) w_i^{-3}$ and $\mathbf{1}^\top F\,Z^{(3)}\,F\,\mathbf{1} = \sum n_i^{-1} f_i^2 w_i^{-3}$, where all the summations range from 1 to p. Substitution into (2.13) gives

$$\varepsilon_p = \frac{1}{12\phi} \sum_{i=1}^{p} n_i^{-1} \left[\frac{2}{V} \left(\frac{dV}{d\mu} \right)^2 - 3\,\frac{d^2 V}{d\mu^2} \right]_i .$$

The subscript i in the right-hand side of the above equation indicates that the quantity inside brackets is evaluated at the ith population. Clearly, because $\hat{\mu}_i$ is the sample mean in the ith population, the quantity ε_p does not depend on the link function. For normal and inverse Gaussian models, $\varepsilon_p = 0$, which is in agreement with the exact χ^2 distribution of the LR statistic in both cases.

2.5.3 Computer Codes for Calculating Bartlett Corrections

Silva and Cordeiro (2009) provided computer codes for calculating Bartlett corrections in GLMs using the R software. They gave empirical examples where their computer codes are used to calculate Bartlett-corrected statistics. Their first example employs data given by Feigl and Zelen (1965) on survival time to death, in weeks, from diagnosis (y) and log(10) of initial blood cell count (x) for leukemia patients. Such data were analyzed by McCullagh and Nelder (1989). Silva and Cordeiro (2009) fitted an exponential regression model with the systematic component $\log(\mu_i) = \beta_0 + \beta_1 x_i$. Their interest lies in testing $H_0 : \beta_1 = 0$. At the outset, the data are entered and the null and non-null models are fitted:

```
x <- c(3.36, 2.88, 3.63, 3.41, 3.78, 4.02, 4.00, 4.23, 3.73,
       3.85, 3.97, 4.51, 4.54, 5.00, 5.00, 4.72, 5.00)
y <- c(65, 156, 100, 134, 16, 108, 121, 4, 39, 143, 56, 26,
       22, 1, 1, 5, 65)
d <- data.frame(x, y)
fit1 <- glm(y ~ 1, family=Gamma(link="log"), x=TRUE, data=d)
fit2 <- glm(y ~ x, family=Gamma(link="log"), x=TRUE, data=d)
anova(fit2, test="Chisq", dispersion=1)
```

The (uncorrected) LR test is then performed:

```
Analysis of Deviance Table
Model: Gamma, link: log
Response: y

Terms added sequentially (first to last)
     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                    16    26.2821
x     1   6.8256        15    19.4565    0.0090
```

To obtain the improved LR test for testing $H_0 : \beta_1 = 0$, they considered the variance function $V(\mu) = \mu^2$ and $\phi = 1$. Since the logarithm link function was considered, the `modlrt` function is called as

```
modlrt(fit1, fit2, V="mu^2", linkfun="log(mu)", phi=1)
```

The output is

```
    Likelihood Ratio Tests

 Error distribution:  gamma
 Link function      :  log
 Model 1            :  y ~ 1
 Model 2            :  y ~ x

 Model Residual Df Deviance Dispersion
     1          16   26.282          1
     2          15   19.457          1

            LR criterion Df P(>|Chi|)
Uncorrected        6.826  1    0.0090
```

```
Corrected          6.743  1    0.0094

 Bartlett Correction:  1.0122
 ----------------------------------------
```

The Bartlett correction is estimated as 1.0122 and thus reduces the LR statistic from 6.826 to 6.743. In this case, the correction is small, and the p value changes from 0.90 to 0.94 %.

Their second empirical illustration considers the 2^4 (unreplicated) factorial experiment presented by Myers et al. (2002, p. 176), where the response variable is the resistivity of test wafers in a semiconductor manufacturing process. The goal is to test the hypothesis of no interaction between the third and fourth factors. This can be achieved by comparing the model with main effects and all second-order interactions to the model that does not include the interactions specified in the null hypothesis. The data are entered and the model are fitted as

```
x1 <- gl(2, 1, labels=c(-1,1), 16)
x2 <- gl(2, 2, labels=c(-1,1), 16)
x3 <- gl(2, 4, labels=c(-1,1), 16)
x4 <- gl(2, 8, labels=c(-1,1), 16)
y  <- c(193.4, 247.6, 168.2, 205, 303.4, 339.9, 226.3, 208.3,
        220, 256.4, 165.7, 203.5, 285, 268, 169.1, 208.5)
d <- data.frame(x1, x2, x3, x4, y)
fit1 <- glm(y ~ x1 + x2 + x3 + x4 + x1*x2 + x1*x3 + x1*x4 +
            x2*x4, family=Gamma(log), x=TRUE, data=d)
summary(fit1)
fit2 <- glm(y ~ x1 + x2 + x3 + x4 + x1*x2 + x1*x3+ x1*x4 +
            x2*x4 + x3*x4, family=Gamma(log), x=TRUE, data=d)
summary(fit2)
```

The instruction

```
modlrt(fit1, fit2, V="mu^2", linkfun="log(mu)")
```

yields the output corresponding to the LR test:

```
    Likelihood Ratio Tests

 Error distribution:  gamma
 Link function      :  log
 Model 1            :  y ~ x1 + x2 + x3 + x4 + x1 * x2 + x1 *
                       x3 + x1 * x4 + x2 * x4
 Model 2            :  y ~ x1 + x2 + x3 + x4 + x1 * x2 + x1 *
                       x3 + x1 * x4 + x2 * x4 + x3 * x4

 Model Residual Df Deviance Dispersion
     1           7    0.098     0.0061
     2           6    0.065     0.0041

            LR criterion Df P(>|Chi|)
Uncorrected        5.413  1    0.0200
Corrected          3.269  1    0.0706
```

```
Bartlett Correction:  1.656
----------------------------------------
```

First, notice that the precision parameter ϕ was not specified in the call to `modlrt`, which caused it to be estimated by the ML method and the Bartlett correction was computed accordingly. It is clear from the above R output that the uncorrected LR test rejects the null hypothesis of no interaction between the third and fourth factors at the 5 % nominal level, since the test p value equals 0.02. The corrected LR test, however, does not reject H_0 at the 5 % nominal level (its p value equals 0.07).

A final example is based on a 3×4 factorial experiment with four replicates that was carried out to evaluate the effects of toxic agents on survival times of rats. The experiment is described in Box and Cox (1964), and the data set can be made available into `R` from the object `rats` in the package `faraway`. The factors in the experiment are `poison` and `treat` having three and four levels, respectively. Silva and Cordeiro (2009) considered the inverse Gaussian model for the survival times with a canonical link function and tested the significance of an interaction between `poison` and `treat`. The null hypothesis under test is that there is no such interaction. The two models are fitted as follows:

```
require(faraway)
fit1 <- glm(time ~ poison + treat,
            family=inverse.gaussian(link = "1/mu^2"),
            x=TRUE, data=rats)
summary(fit1)
fit2 <- glm(time ~ treat * poison,
            family=inverse.gaussian(link = "1/mu^2"),
            x=TRUE, data=rats)
summary(fit2)
```

The instruction

```
modlrt(fit1, fit2, V="mu^3", linkfun="1/mu^2")
```

is used to perform LR inference:

```
    Likelihood Ratio Tests

 Error distribution:  inverse.gaussian
 Link function     :  1/mu^2
 Model 1           :  time ~ poison + treat
 Model 2           :  time ~ treat * poison

 Model Residual Df Deviance Dispersion
     1          42    5.455     0.1299
     2          36    3.642     0.1012

            LR criterion Df P(>|Chi|)
Uncorrected        13.96  6    0.0301
Corrected          11.52  6    0.0736

 Bartlett Correction:  1.2118
 ----------------------------------------
```

The estimated Bartlett correction equals 1.2118. It reduces the value of the LR statistic from 13.96 to 11.52, thus increasing the test p value from 0.0301 to 0.0736.

2.6 Birnbaum–Saunders Non-linear Regression Models

The random variable T is said to be Birnbaum–Saunders (BS) distributed with parameters $\alpha, \eta > 0$, say $\mathcal{BS}(\alpha, \eta)$, if its cdf is given by

$$F_T(t) = \Phi\left[\frac{1}{\alpha}\left(\sqrt{\frac{t}{\eta}} - \sqrt{\frac{\eta}{t}}\right)\right], \quad t > 0,$$

where $\Phi(\cdot)$ is the standard normal distribution function and α and η are shape and scale parameters, respectively. The BS distribution can be used to model lifetime data and typically yields satisfactory tail fitting. It was originally obtained from a model in which failure follows from the development and growth of a dominant crack (Birnbaum and Saunders 1969). It is easy to show that η is the median, i.e., $F_T(\eta) = \Phi(0) = 1/2$. For all $k > 0$, it follows that $kT \sim \mathcal{BS}(\alpha, k\eta)$.

Rieck and Nedelman (1991) introduced a log-linear regression model based on the $\mathcal{BS}(\alpha, \eta)$ distribution. They showed that if $T \sim \mathcal{BS}(\alpha, \eta)$, then $Y = \log(T)$ is sinh-normal distributed with shape, location, and scale parameters given by α, $\mu = \log(\eta)$ and $\sigma = 2$, respectively, say $Y \sim \mathcal{SN}(\alpha, \mu, 2)$. The density function of Y is given by

$$\pi(y) = \frac{1}{\alpha\sqrt{2\pi}} \cosh\left(\frac{y-\mu}{2}\right) \exp\left\{-\frac{2}{\alpha^2}\sinh^2\left(\frac{y-\mu}{2}\right)\right\}, \quad y \in \mathbb{R}, \tag{2.16}$$

which has a number of interesting properties. For example, it is symmetric around the location parameter μ, the mean of Y is $\mathbb{E}(Y) = \mu$ and if $Y_\alpha \sim \mathcal{SN}(\alpha, \mu, \sigma)$, then $Z_\alpha = 2(Y_\alpha - \mu)/(\alpha\sigma)$ converges in distribution to the standard normal distribution when $\alpha \to 0$. Likelihood-based inference in BS linear regression models can be found in several articles.

Lemonte and Cordeiro (2009) proposed the class of BS non-linear regression models given by

$$Y_i = \mu_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.17}$$

where Y_i is the logarithm of the ith lifetime, x_i is an $m \times 1$ vector of explanatory variables values associated with the ith observable response y_i, $\beta = (\beta_1, \ldots, \beta_p)^\top$ is a vector of unknown non-linear parameters ($m \leq p < n$) and $\varepsilon_i \sim \mathcal{SN}(\alpha, 0, 2)$ for $i = 1, \ldots, n$. They considered a non-linear structure for the mean parameter $\mu_i = f_i(x_i; \beta)$, where $f_i(\cdot)$ is assumed to be a known and twice continuously differentiable function such that the derivative matrix $X = X(\beta) = \partial\mu/\partial\beta^\top$ has rank p for all β, where $\mu = (\mu_1, \ldots, \mu_n)^\top$. The non-linear predictors $x_1, \ldots, x_n$ are embedded in an infinite sequence of $m \times 1$ vectors that must satisfy these

regularity conditions for the asymptotics to be valid. Under these assumptions, the MLEs have the usual desirable properties, such as consistency, sufficiency, and asymptotic normality. The $n \times p$ local matrix X has elements that are, in general, functions of the unknown parameter vector β. As the name suggests, the class of the BS non-linear regression models extends Rieck and Nedelman's model to allow for non-linear parameters β's.

The log-likelihood function for the parameter vector $\theta = (\beta^\top, \alpha)^\top$ from a random sample $Y = (Y_1, \ldots, Y_n)^\top$ obtained from (2.16), with observed values $y = (y_1, \ldots, y_n)^\top$, except for constants, can be expressed as

$$\ell(\theta) = \sum_{i=1}^{n} \log(\xi_{i1}) - \frac{1}{2} \sum_{i=1}^{n} \xi_{i2}^2, \tag{2.18}$$

where

$$\xi_{i1} = \xi_{i1}(\theta) = \frac{2}{\alpha} \cosh\left(\frac{y_i - \mu_i}{2}\right), \quad \xi_{i2} = \xi_{i2}(\theta) = \frac{2}{\alpha} \sinh\left(\frac{y_i - \mu_i}{2}\right),$$

for $i = 1, \ldots, n$. The derivatives of $\ell(\theta)$ with respect to the components of β and α are given by $U_r = \partial \ell(\theta)/\partial \beta_r$, $U_\alpha = \partial \ell(\theta)/\partial \alpha$, $U_{rs} = \partial^2 \ell(\theta)/\partial \beta_r \partial \beta_s$, $U_{r\alpha} = \partial^2 \ell(\theta)/\partial \beta_r \partial \alpha$, $U_{rs\alpha} = \partial^3 \ell(\theta)/\partial \beta_r \partial \beta_s \partial \alpha$, etc. The joint cumulants of log-likelihood derivatives are $\kappa_{rs} = \mathbb{E}(U_{rs})$, $\kappa_{r,\alpha} = \mathbb{E}(U_r U_\alpha)$, $\kappa_{rst} = \mathbb{E}(U_{rst})$, etc. Let $\kappa_{rs}^{(t)} = \partial \kappa_{rs}/\partial \beta_t$, etc. All κ's and their derivatives are of order $\mathcal{O}(n)$. In what follows, we use the notation: $d_{ir} = \partial \mu_i/\partial \beta_r$ and $g_{irs} = \partial^2 \mu_i/\partial \beta_r \partial \beta_s$ for the first and second partial derivatives of μ_i with respect to the elements of β. We assume that some standard regularity conditions on $\ell(\theta)$ and its first four derivatives hold as n tends to infinity. We use the standard notation where joint cumulants are denoted by indices $r, s, t, \ldots$ if they correspond to β parameters, whereas the index α corresponds to the α parameter. It follows from the differentiation of (2.18) that

$$\kappa_{rs} = -\frac{\psi_1(\alpha)}{4} \sum_{i=1}^{n} d_{ir}\, d_{is}, \quad \kappa_{r\alpha} = 0, \quad \kappa_{\alpha\alpha} = -\frac{2n}{\alpha^2},$$

where

$$\psi_1(\alpha) = 2 + \frac{4}{\alpha^2} - \frac{\sqrt{2\pi}}{\alpha} \psi_0(\alpha), \quad \psi_0(\alpha) = \left\{1 - \mathtt{erf}\left(\frac{\sqrt{2}}{\alpha}\right)\right\} \exp\left(\frac{2}{\alpha^2}\right).$$

Here, $\mathtt{erf}(\cdot)$ is the error function defined by $\mathtt{erf}(x) = (2/\sqrt{\pi}) \int_0^x \mathrm{e}^{-t^2}\, \mathrm{d}t$.

Notice that the parameters β and α are globally orthogonal (Cox and Reid 1987) since $\kappa_{r\alpha} = 0$ for all $r = 1, \ldots, p$. Thus, the joint information matrix K_θ for $\theta = (\beta^\top, \alpha)^\top$ is block diagonal, say $K_\theta = \mathrm{diag}\{K_\beta, 2n/\alpha^2\}$, where $K_\beta = \psi_1(\alpha)(X^\top X)/4$. In view of the block diagonality of K_θ, the Fisher scoring

method can be used to obtain the MLEs $\hat{\beta}$ and $\hat{\alpha}$ simultaneously by iteratively solving the following equations:

$$(X^{(m)\top}X^{(m)})\beta^{(m+1)} = X^{(m)\top}\zeta^{(m)}, \quad \alpha^{(m+1)} = \frac{1}{2}\alpha^{(m)}(1+\bar{\xi}_2^{(m)}), \quad m = 0, 1, \ldots,$$

where $\zeta^{(m)} = X^{(m)}\beta^{(m)} + [2/\psi_1(\alpha^{(m)})]\, s^{(m)}$, $s = (s_1, \ldots, s_n)^\top$, $\bar{\xi}_2^{(m)} = n^{-1}$ $\sum_{i=1}^n \xi_{i2}^{2(m)}$ and $s_i = \xi_{i1}\xi_{i2} - \xi_{i2}/\xi_{i1}$. Any software with a weighted linear regression routine can be used to calculate the MLEs of β and α iteratively. Starting values $\beta^{(0)}$ and $\alpha^{(0)}$ for the iterative algorithm are required. These new values can update ζ and $\bar{\xi}_2$ and so on. The iterating mechanism goes on until convergence is achieved.

In what follows, we present the Bartlett correction for testing nested hypotheses in BS non-linear regression models. The joint cumulants for β and α required for the Bartlett correction are easily calculated. They are not given here but can be obtained from the authors upon request. Our aim is to present the Bartlett corrections in a more readily computable form by exploiting special properties of these κ's. The calculations follow Eqs. (2.4)–(2.6). We can write $2\mathbb{E}\{\ell(\hat{\beta}, \hat{\alpha}) - \ell(\beta, \alpha)\} = p + 1 + \varepsilon_{p+1} + \mathcal{O}(n^{-2})$, where ε_{p+1} can be calculated using Eq. (2.4) with all indices varying in β and α. Some additional matrices are introduced:

$$Z = X(X^\top X)^{-1}X^\top = \{z_{ij}\}, \quad Z_d = \text{diag}\{z_{11}, \ldots, z_{nn}\},$$

$$D_d = \text{diag}\{d_1, \ldots, d_n\}, \quad B = \{b_{ij}\}, \quad B_d = \text{diag}\{b_{11}, \ldots, b_{nn}\},$$

where $d_i = \text{tr}\{X_i(X^\top X)^{-1}\}$, $b_{ij} = \text{tr}\{X_i(X^\top X)^{-1}X_j(X^\top X)^{-1}\}$, and X_i denotes a $p \times p$ matrix whose elements are $\partial^2\mu_i/\partial\beta_r\partial\beta_s$ for $r, s = 1, \ldots, p$. Also, $Z^{(2)} = Z \odot Z$, $Z_d^{(2)} = Z_d \odot Z_d$, and so on, where '$\odot$' denotes the Hadamard product of matrices.

It is possible to write, after lengthy algebra, $\varepsilon_{p+1} = \varepsilon(\alpha, p, Z, B, D_d)$ as (Lemonte et al. 2012)

$$\varepsilon(\alpha, p, Z, B, D_d) = \varepsilon_L(\alpha, p, Z) + \varepsilon_{NL}(\alpha, Z, B, D_d), \tag{2.19}$$

where

$$\varepsilon_L(\alpha, p, Z) = \frac{1}{n}\left\{\frac{1}{3} + \delta_1(\alpha)p + \delta_2(\alpha)p^2\right\} + \delta_3(\alpha)\text{tr}(Z_d^{(2)})$$

and

$$\varepsilon_{NL}(\alpha, Z, B, D_d) = -\frac{1}{\psi_1(\alpha)}\text{tr}(D_d^{(2)} - 2B_d + ZB).$$

Additionally,

$$\delta_0(\alpha) = \frac{2+\alpha^2}{\psi_1(\alpha)\alpha^2}, \quad \delta_1(\alpha) = 4\delta_0(\alpha)\left\{\frac{2}{2+\alpha^2} + \delta_0(\alpha) - \frac{2\alpha\psi_3(\alpha)}{\psi_1(\alpha)}\right\},$$
$$\delta_2(\alpha) = 2\delta_0(\alpha)^2, \quad \delta_3(\alpha) = \frac{4\psi_2(\alpha)}{\psi_1(\alpha)^2},$$
$$\psi_2(\alpha) = -\frac{1}{4}\left\{2 + \frac{7}{\alpha^2} - \sqrt{\frac{\pi}{2}}\left(\frac{1}{2\alpha} + \frac{6}{\alpha^3}\right)\psi_0(\alpha)\right\},$$
$$\psi_3(\alpha) = \frac{3}{\alpha^3} - \frac{\sqrt{2\pi}}{4\alpha^2}\left(1 + \frac{4}{\alpha^2}\right)\psi_0(\alpha).$$

The details of the calculations can be found in Lemonte et al. (2012).

Equation (2.19) provides a simple decomposition for the Bartlett correction. A brief commentary on this equation seems in order. The quantity $\varepsilon_L(\alpha, p, Z)$ is identical to the expression for the BS linear regression models derived by Lemonte et al. (2010). On the other hand, the quantity $\varepsilon_{NL}(\alpha, Z, B, D_d)$ may be regarded as the amount of non-linearity in the null expected LR induced by the non-linear parameters in $f_i(x_i; \beta)$. In particular, if $f_i(x_i; \beta)$ is linear for $i = 1, \ldots, n$, we obtain $\varepsilon_{NL}(\alpha, Z, B, D_d) = 0$, since d_i and b_{ij} vanish and thus ε_{p+1} reduces to the result by Lemonte et al. (2010). Finally, it should be noted that Eq. (2.19) is quite simple and can be easily implemented in any mathematical or statistical/econometric programming environment, such as MAPLE, OX, and R.

The interest typically lies in testing restrictions on a subset of the regression parameters. Consider the partition $\beta = (\beta_1^\top, \beta_2^\top)^\top$, where $\beta_1 = (\beta_1, \ldots, \beta_q)^\top$ and $\beta_2 = (\beta_{q+1}, \ldots, \beta_p)^\top$ are vectors of dimensions $q \times 1$ and $(p-q) \times 1$, respectively, and the test of $H_0 : \beta_1 = \beta_1^{(0)}$ against $H : \beta_1 \neq \beta_1^{(0)}$, where $\beta_1^{(0)}$ is a q-vector of known constants, β_2 and α being nuisance parameters. The local model matrix X is partitioned following the partition of β, say $X = (X_1 \; X_2)$, the dimensions of X_1 and X_2 being $n \times q$ and $n \times (p-q)$, respectively. The Bartlett correction factor is $c = 1 + b/q$, where $b = \varepsilon_p(\alpha, Z, B, D_d) - \varepsilon_{p-q}(\alpha, Z_1, B_2, D_{2d})$. It can be shown that

$$b = \frac{1}{n}\left\{q\,\delta_1(\alpha) + q(2p-q)\,\delta_2(\alpha)\right\} + \delta_3(\alpha)\mathrm{tr}(\mathrm{Z}_{\mathrm{d}}^{(2)} - \mathrm{Z}_{2\mathrm{d}}^{(2)})$$
$$- \frac{1}{\psi_1(\alpha)}\mathrm{tr}\{(\mathrm{D}_{\mathrm{d}}^{(2)} - \mathrm{D}_{2\mathrm{d}}^{(2)}) - 2(\mathrm{B_d} - \mathrm{B_{2d}}) + (\mathrm{ZB} - \mathrm{Z_2B_2})\}.$$

Here,

$$Z_2 = X_2(X_2^\top X_2)^{-1}X_2^\top = \{z_{2ij}\}, \quad Z_{2d} = \mathrm{diag}\{z_{211}, \ldots, z_{2nn}\},$$

$$D_{2d} = \mathrm{diag}\{d_{21}, \ldots, d_{2n}\}, \quad B_2 = \{b_{2ij}\}, \quad B_{2d} = \mathrm{diag}\{b_{211}, \ldots, b_{2nn}\},$$

where $d_{2i} = \mathrm{tr}\{X_{22i}(X_2^\top X_2)^{-1}\}$, $b_{2ij} = \mathrm{tr}\{X_{22i}(X_2^\top X_2)^{-1}X_{22j}(X_2^\top X_2)^{-2}\}$, and X_{22i} is a $(p-q) \times (p-q)$ matrix obtained from the $p \times p$ partitioned matrix following the partition of β,

$$X_i = \left\{ \frac{\partial^2 \mu_i}{\partial \beta_r \partial \beta_s} \right\} = \begin{bmatrix} X_{11i} & X_{12i} \\ X_{21i} & X_{22i} \end{bmatrix},$$

for $i = 1, \ldots, n$.

Consider now the test of $H_0 : \alpha = \alpha^{(0)}$ against $H : \alpha \neq \alpha^{(0)}$, where $\alpha^{(0)}$ is a given positive scalar and β is a vector of nuisance parameters. The Bartlett correction factor reduces to

$$c = 1 + \frac{1}{n} \left\{ \frac{1}{3} + \delta_1(\alpha^{(0)})\, p + \delta_2(\alpha^{(0)})\, p^2 \right\}.$$

The correction c depends only on the non-linear structure through the rank of X (i.e., p) and it is exactly the same given by Lemonte et al. (2010) for the BS linear regression. Thus, the Bartlett correction for testing $H_0 : \alpha = \alpha^{(0)}$ is the same for any non-linear regression structure with the same p.

2.7 Bootstrap-Based Hypothesis Testing

An alternative strategy for improving on LR testing inference (and also on testing inference based on other criteria) is to use data resampling to estimate the test statistic null distribution, thus avoiding the use of an asymptotic approximation. This can be done using Efron's (1979) bootstrap in its parametric version. The main idea is to sample from the data as if we were sampling from the population, and then use the information contained in the pseudo-samples to improve the statistical inference.

Consider a random sample denoted by $Y = (Y_1, \ldots, Y_n)^\top$, where each Y_i is a random draw from the random variable Y. We denote the distribution function of Y by $F = F_\theta = F_\theta(y)$, where θ is a scalar- or vector-valued parameter. The parameter θ can be viewed as a functional of F: $\theta = t(F)$. Suppose θ is vector-valued and we wish to test $H_0 : \psi = \psi^0$ against $H : \psi \neq \psi^0$, where $\theta = (\psi^\top, \lambda^\top)^\top$. Here, ψ is a q-vector of parameters of interest and λ is a $(p-q)$-vector of nuisance parameters. Hence, θ contains p unknown parameters. The (unrestricted) MLE of θ is $\hat{\theta} = (\hat{\psi}^\top, \hat{\lambda}^\top)^\top$, with $\hat{\psi}$ and $\hat{\lambda}$ being the MLEs of ψ and λ, respectively. The restricted MLE of θ (i.e., obtained by imposing the null hypothesis) is $\tilde{\theta} = (\psi^0, \tilde{\lambda}^\top)^\top$, where $\tilde{\lambda}$ is the MLE of λ given $\psi = \psi^0$. The LR test statistic becomes

$$w = 2\{\ell(\hat{\theta}) - \ell(\tilde{\theta})\},$$

which, under the null hypothesis, has a limiting χ^2_q distribution. As before, ℓ denotes the log-likelihood function. The null hypothesis is rejected if $w > \chi^2_{1-\alpha,q}$, where $\chi^2_{1-\alpha,q}$ is the $(1-\alpha)$th quantile from the χ^2_q distribution and α is the test significance level (e.g., $\alpha = 0.05$). Since the test is based on an asymptotic (approximate) critical value size, distortions are likely to occur when the sample size is not large.

Testing inference based on asymptotic χ^2 criteria (such as the LR test statistic) can be made more reliable by a critical value obtained from the test statistic null distribution estimated by bootstrap resampling. This can be accomplished as follows. We obtain, from the original sample y, a large number (say, R) of pseudo-samples $y^* = (y_1^*, \ldots, y_n^*)^\top$. The null hypothesis is imposed when generating the artificial samples, which are obtained by taking random draws from $F_{\tilde{\theta}}$. Notice that we sample from the distribution function F after replacing the unknown parameter vector by its *restricted* MLE. The test statistic is computed for each of the pseudo-samples $w_1^*, \ldots, w_R^*$. The bootstrap statistics are then used to estimate the null distribution of w. It is important to note that they were computed from samples generated by imposing the null hypothesis, and hence, they can be used to estimate the *null* distribution of w. A bootstrap critical value corresponding to the $(1-\alpha)$th nominal level $(0 < \alpha < 1)$ can be obtained as the $(1-\alpha)$th quantile of the $R+1$ test statistics (R bootstrap statistics and the test statistic computed using the original sample). Denote such a critical value by $\text{cvb}_{1-\alpha}$. The null hypothesis is thus rejected if $w > \text{cvb}_{1-\alpha}$. Notice that the bootstrap test does not use a critical value obtained from the test statistic *limiting* (asymptotic) null distribution, which may yield a poor approximation to the exact critical value; it uses a critical value from the *estimated* null distribution of w.

Alternatively, we can state the rejection rule using the bootstrap p value, which is given by $(k+1)/(R+1)$, where k is the number of bootstrap replications in which w^* (the LR statistic computed using the pseudo-sample) is greater than w (the LR statistic computed using the original sample). The null hypothesis is rejected if such a p-value is smaller than or equal to α, the significance level.

As $R \to \infty$, the bootstrap p value tends to the ideal p value, say p^*, which leads to the rejection of the null hypothesis whenever $p^* \leq \alpha$. However, a 'feasible bootstrap' must be based on a finite number of bootstrap replications ($R < \infty$). This causes a loss of power. Such a loss, however, tends to be small when R (the number of bootstrap resamples) is large. It is possible to obtain a bound for the loss in power that follows from using a finite number of resamples. Let π_R and π_∞ denote the powers of the feasible and ideal bootstrap schemes, respectively (that is, the probability that we reject H_0 when H_0 is false based on $R < \infty$ and on $R = \infty$). It can be shown that (Jöeckel 1986)

$$\frac{\pi_R}{\pi_\infty} \geq 1 - \sqrt{\frac{1-\alpha}{2\pi\,(R+1)\,\alpha}}.$$

This bound can give us a rough idea of the magnitude of the power loss that moderate values of R may introduce. For details, see Davison and Hinkley (1997, pp. 155–156).

Consider the linear regression model, which is commonly used in a wide variety of fields:

$$Y = X\beta + \varepsilon,$$

where Y is an n-vector of responses, ε is an n-vector of (random, unobservable) errors, X is a fixed $n \times p$ model matrix of covariate values ($\text{rank}(X) = p < n$), and

$\beta = (\beta_1, \ldots, \beta_p)^\top$ is a p-vector of unknown regression parameters. The model can be written as

$$Y_i = x_i^\top \beta + \varepsilon_i, \quad i = 1, \ldots, n,$$

where x_i is the ith row of X. Each error ε_i has mean zero, constant variance σ^2 ($0 < \sigma^2 < \infty$), and is pairwise uncorrelated with all other errors. The errors covariance matrix is $\sigma^2 I_n$, where I_n denotes the n-dimensional identity matrix. The ordinary least-squares (OLS) estimators of β and σ^2 are $\hat{\beta} = (X^\top X)^{-1} X^\top Y$ and $\hat{\sigma}^2 = (n-p)^{-1} \hat{\varepsilon}^\top \hat{\varepsilon} = (n-p)^{-1} \sum_{i=1}^n \hat{\varepsilon}_i^2$, respectively. Here, $\hat{\varepsilon} = Y - X\hat{\beta}$ is the vector of OLS residuals, $\hat{\varepsilon}_i$ denoting its ith component. It is easy to show that the covariance matrix of $\hat{\beta}$ is $\sigma^2 (X^\top X)^{-1}$, which can be easily estimated by $\hat{\sigma}^2 (X^\top X)^{-1}$.

Suppose we wish to test $H_0 : \beta_j = \beta_j^0$ against $H : \beta_j \neq \beta_j^0$, for some $j = 1, \ldots, p$. This is usually done by assuming that the errors are normally distributed, computing the test statistic given by

$$t = \frac{\hat{\beta}_j - \beta_j^0}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_j)}}, \tag{2.20}$$

where $\widehat{\mathrm{Var}}(\hat{\beta}_j)$ is the jth diagonal element of $\hat{\sigma}^2 (X^\top X)^{-1}$, and then rejecting the null hypothesis if $|t| > t_{1-\alpha/2, n-p}$. Here, $t_{1-\alpha/2, n-p}$ denotes the $(1-\alpha/2)$th quantile of the Student t_{n-p} distribution and α is the test significance level. This is an exact test, but it is heavily dependent on the assumptions that the errors are normally distributed and homoscedastic (i.e., that the errors have constant variance).

Let us use bootstrap resampling to perform the test of $H_0 : \beta_j = \beta_j^0$ without having to resort to the normality assumption. We proceed as follows. First, compute the t test statistic in (2.20) and estimate the restricted model. Then:

1. For each i, $i = 1, \ldots, n$, sample $\tilde{\varepsilon}_i^*$ from $\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_n$ with replacement, where $\tilde{\varepsilon}_i$ is the ith *restricted* OLS residual.
2. Construct a bootstrap sample (Y^*, X), where $Y_i^* = x_i^\top \tilde{\beta} + \tilde{\varepsilon}_i^*$, $i = 1, \ldots, n$. Here, $\tilde{\beta}$ is the restricted parameter estimate.
3. Compute the OLS estimate of β, $\hat{\beta}^* = (X^\top X)^{-1} X^\top Y^*$, and compute the associated test statistic, t^*.
4. Execute steps 1–3 a large number (say, R) of times.
5. Compute the quantile of interest of the empirical distribution of the $R+1$ realizations of the test statistic.
6. Perform the test using the t test statistic computed from the original sample together with the bootstrap critical value obtained in Step 5.

Note that in the bootstrap test we do not rely on critical values from the Student t distribution (nor from the standard normal limiting null distribution). Instead, we use critical values obtained from the bootstrapping scheme. The bootstrap method described above is usually referred to as the *unweighted* or *naïve* bootstrap.

Let us now consider testing inference on the regression parameters under heteroscedasticity. Our goal is to perform inference that is valid under both

homoscedasticity and heteroscedasticity of unknown form. In this setting, $\varepsilon_1, \ldots, \varepsilon_n$ have variances $\sigma_1^2, \ldots, \sigma_n^2$ ($0 < \sigma_i^2 < \infty \forall\, i$), respectively, the error covariance matrix being $\Omega = \text{diag}\{\sigma_1^2, \ldots, \sigma_n^2\}$. Note that the n variances show up in the main diagonal and that all off-diagonal elements equal zero (since each error is uncorrelated with all other errors). The OLSE $\hat{\beta}$ remains unbiased and consistent when the errors are heteroscedastic and it is easy to show that its covariance matrix becomes $(X^\top X)^{-1} X^\top \Omega X (X^\top X)^{-1}$. As explained in Sect. 4.10, this matrix can be consistently estimated using a heteroscedasticity-consistent covariance matrix estimator, such as, for instance, the HC0 estimator (White 1980)

$$\text{HC0} = \hat{\Psi} = (X^\top X)^{-1} X^\top \hat{\Omega} X (X^\top X)^{-1}, \tag{2.21}$$

where $\hat{\Omega} = \text{diag}\{\hat{\varepsilon}_1^2, \ldots, \hat{\varepsilon}_n^2\}$, with $\hat{\varepsilon}_i$ being the ith least-squares residual, i.e., $\hat{\varepsilon}_i = Y_i - x_i^\top \hat{\beta}, i = 1, \ldots, n$. The quasi-$t$ test statistic is as in (2.20), but now the estimated variance of $\hat{\beta}_j$ is the jth diagonal element of (2.21). Under the null hypothesis, it has a limiting standard normal distribution, and hence, a test can be performed by comparing the test statistic to the appropriate standard normal quantile. It has been shown, however, that the White estimator is typically quite biased in small samples and that the associated quasi-t test is usually liberal (oversized).

It is noteworthy that in the naïve bootstrap, the errors are taken to be identically distributed, which is not the case when they display heteroscedasticity. A *weighted* bootstrap for handling heteroscedastic data was introduced by Wu (1986). Let τ denote the quasi-t test statistic computed from the original sample, i.e., from (Y, X). Then:

1. For each i, $i = 1, \ldots, n$, draw a random number t_i^* from a population that has *mean zero* and *variance one*.
2. Construct a bootstrap sample (Y^*, X), where $Y_i^* = x_i^\top \tilde{\beta} + t_i^* \tilde{\varepsilon}_i / (1 - h_i)$. Here, $\tilde{\beta}$ and $\tilde{\varepsilon}$ are the restricted parameter estimates and the associated restricted least-squares residuals from the regression of Y on X, respectively. Also, h_i is the ith diagonal element of $H = X(X^\top X)^{-1} X^\top$ (the 'hat matrix').
3. Compute the OLSE of β, $\hat{\beta}^* = (X'X)^{-1} X'Y^*$, and the associated quasi-t test statistic, τ^*.
4. Execute steps 1–3 a large number (say, R) of times.
5. Compute the quantile of interest of the empirical distribution of the $R + 1$ realizations of the test statistic.
6. Perform the test using the quasi-t statistic computed initially (τ) together with the bootstrap critical value obtained in step 5 above.

The null hypothesis is rejected if $(k + 1)/(R + 1) < \alpha$, where α is the nominal level of the test and k is the number of times (out of R) such that $|\tau^*| > |\tau|$. Note that $(k + 1)/(R + 1)$ can be viewed as a bootstrap p value.

In the above bootstrap scheme, t_i^* must be sampled from a population that has mean zero and unit variance (step 1). This population is sometimes referred to as 'the pick distribution' (Godfrey 2009). Obvious choices for the pick distribution are:

(1) the regression residuals (standardized to have mean zero and variance one) and (2) the standard normal distribution. Liu (1988) showed that when the pick distribution third non-central moment equals one, weighted bootstrap enjoys second-order optimality in the sense that the test statistic first three moments are estimated correctly up to order $\mathcal{O}(n^{-1})$. A pick distribution that has mean zero, variance one and third non-central moment equal to one can be defined as follows. Let Z_1 and Z_2 be independent normal random variables with variance $1/2$ and means given by

$$\frac{1}{2}\left(\sqrt{\frac{17}{6}}+\sqrt{\frac{1}{6}}\right) \quad \text{and} \quad \frac{1}{2}\left(\sqrt{\frac{17}{6}}-\sqrt{\frac{1}{6}}\right),$$

respectively. We then sample from the pick distribution $Z_1 \times Z_2 - \mathbb{E}(Z_1) \times \mathbb{E}(Z_2)$. The Monte Carlo evidence, however, suggests that the best pick distribution is a very simple one: the Rademacker distribution. It is defined as: -1 with probability $1/2$ and $+1$ with probability $1/2$; see, e.g., Flachaire (2005).

It is possible to obtain a more accurate bootstrap p value using the *double bootstrap*, which is, however, more computer intensive. Here, we nest a bootstrap sampling within each bootstrap replication, that is, we perform a second level of bootstrap resampling for each original bootstrap replication; see Davison and Hinkley (1997, §4.5). Let $\tau_1^*, \ldots, \tau_R^*$ denote the R bootstrap realizations of the test statistic. We proceed as follows, where C denotes the number of bootstrap replications in the second level of bootstrapping, and $b = 1, \ldots, R$ indexes the first level of bootstrapping:

1. For each i, $i = 1, \ldots, n$, draw a random number t_i^{**} from a population that has mean zero and variance one.
2. Construct a bootstrap sample (Y^{**}, X), where $Y_i^{**} = x_i^\top \tilde{\beta}^\dagger + t_i^{**}\tilde{\varepsilon}_i^\dagger/(1-h_i)$. Here, $\tilde{\beta}^\dagger$ and $\tilde{\varepsilon}^\dagger$ are the restricted parameter estimates and the associated restricted least-squares residuals from the regression of Y^* on X.
3. Compute the OLSE of β, $\hat{\beta}^{**} = (X'X)^{-1}X'Y^{**}$, and the associated quasi-t statistic, τ^{**}.
4. Compute p_b^* using (2.22); see below.
5. Use the realizations from the two levels of bootstrapping to obtain an adjusted p value for the test (see below).

Steps 1–4 described above must be performed for each outer bootstrap replication ($b = 1, \ldots, R$). The adjusted bootstrap p value is given by

$$p_{\text{adj}} = \frac{1 + \#\{p_b^* \leq p\}}{R+1},$$

where, for each b,

$$p_b^* = \frac{1 + \#\{|\tau_{bc}^{**}| \geq |\tau_b^*|\}}{C+1}, \tag{2.22}$$

$c = 1, \ldots, C$. We reject the null hypothesis, tested against a two-sided alternative hypothesis, if $p_{\text{adj}} \leq \alpha$, where α is the nominal level of the test. Note that the total

number of bootstrap replications is now $R \times C$, thus implying a heavier computational burden. Typically, $C < R$, i.e., we use fewer replications in the inner bootstrap.

Monte Carlo results on the finite sample performance of weighted bootstrap and weighted double bootstrap tests in heteroscedastic linear regressions can be found in Cribari-Neto (2004). The number of bootstrap replications used in his simulations were $R = 999$ and $C = 249$. The double bootstrap test typically displays size distortions that are slightly smaller than those of the bootstrap test. For instance, when the sample contains 50 observations, the model is given by $Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, the interest lies in the test of $H_0 : \beta_2 = 0$ against $H : \beta_2 \neq 0$, and the largest error variance is approximately 95 times larger than the smallest one, and the null rejection rates of the single and double bootstrap tests at the 5 % nominal level are 8.5 and 7.0 %, respectively. When $n = 150$, the corresponding null rejection rates are 7.0 and 6.5 %.

Rocke (1989) introduced the bootstrap Bartlett adjustment. His proposal is to use bootstrap resampling to estimate the Bartlett correction factor used to improve the LR test and not to obtain a critical value or a p value. Recall that the Bartlett-corrected test statistic can be written as w/c, where $c = \mathbb{E}(w)/q$, q being the number of restrictions imposed by the null hypothesis. Rocke (1989) recommended the use of parametric bootstrap resampling to estimate c. R bootstrap samples are produced using the parametric bootstrap and imposing the null hypothesis, the LR test statistic is computed for each artificial sample (w^*), and the bootstrap Bartlett-corrected test statistic is computed as

$$w_{\text{boot}} = \frac{wq}{\overline{w}^*},$$

where $\overline{w}^*$ is the average of all bootstrap statistics, i.e.,

$$\overline{w}^* = \frac{1}{R} \sum_{b=1}^{R} w_b^*,$$

w_b^* being the LR statistic computed using the bth pseudo-sample ($b = 1, \ldots, R$). The main advantage of Rocke's approach over the standard bootstrap testing strategy is the smaller number of bootstrap resamples that are needed to achieve a good approximation: 200 bootstrap replications are usually enough to accurately estimate the Bartlett adjustment factor, whereas 1,000 bootstrap replications are typically recommended when estimating the test critical value (or, equivalently, when obtaining a bootstrap p value). It is noteworthy that in the former, we use the bootstrap to estimate the test statistic null distribution *mean*, whereas in the latter, we use it to estimate a *tail quantity* (an upper quantile), which is considerably harder to estimate and thus demands more replications.

References

Andrews, D., & Stafford, J. E. (1993). Tools for the symbolic computation of asymptotic expansions. *Journal of the Royal Statistical Society B*, *55*, 613–627.

Attfield, C. L. F. (1991). A Bartlett-adjustment to the likelihood ratio test for homoskedasticity in the linear model. *Economics Letters*, *37*, 119–123.

Barndorff-Nielsen, O. E., & Cox, D. R. (1984). Bartlett adjustments to the likelihood ratio statistic and the distribution of the maximum likelihood estimator. *Journal of the Royal Statistical Society B*, *46*, 484–495.

Barndorff-Nielsen, O. E., & Cox, D. R. (1994). *Inference and asymptotics*. London: Chapman and Hall.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society A*, *160*, 268–282.

Bartlett, M. S. (1938). Further aspects of the theory of multiple regression. *Proceedings of the Cambridge Society*, *34*, 33–40.

Bartlett, M. S. (1947). Multivariate analysis. *Journal of the Royal Statistical Society*, *9*, 176–197. (Supplement).

Bartlett, M. S. (1954). A note on the multiplying factors for various χ^2 approximations. *Journal of the Royal Statistical Society B*, *16*, 296–298.

Beale, E. M. L. (1960). Confidence regions in non-linear estimation. *Journal of the Royal Statistical Society B*, *26*, 41–88.

Birnbaum, Z. W., & Saunders, S. C. (1969). A new family of life distributions. *Journal of Applied Probability*, *6*, 319–327.

Box, G. E. P. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika*, *36*, 317–346.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society B*, *26*, 211–252.

Chesher, A., & Smith, R. (1995). Bartlett corrections to likelihood ratio tests. *Biometrika*, *82*, 433–436.

Cordeiro, G. M. (1983). Improved likelihood ratio statistics for generalized linear models. *Journal of the Royal Statistical Society B*, *45*, 404–413.

Cordeiro, G. M. (1987). On the corrections to the likelihood ratio statistics. *Biometrika*, *74*, 265–274.

Cordeiro, G. M. (1993a). General matrix formula for computing Bartlett corrections. *Statistics and Probability Letters*, *16*, 11–18.

Cordeiro, G. M. (1993b). Bartlett corrections and bias correction for two heteroscedastic regression models. *Communications in Statistics, Theory and Methods*, *22*, 169–188.

Cordeiro, G. M. (1995). Performance of a Bartlett-type modification for the deviance. *Journal of Statistical Computation and Simulation*, *51*, 385–403.

Cordeiro, G. M., Cribari-Neto, F., Aubin, E. C. Q., & Ferrari, S. L. P. (1995). Bartlett corrections for one-parameter exponential family models. *Journal of Statistical Computation and Simulation*, *53*, 211–231.

Cordeiro, G. M., & Paula, G. A. (1989). Improved likelihood ratio statistics for exponential family nonlinear models. *Biometrika*, *76*, 93–100.

Cordeiro, G. M., Paula, G. A., & Botter, D. A. (1994). Improved likelihood ratio tests for dispersion models. *International Statistical Review*, *62*, 257–276.

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. London: Chapman and Hall.

Cox, D. R., & Reid, N. (1987). Approximations to noncentral distributions. *Canadian Journal of Statistics*, *15*, 105–114.

Cribari-Neto, F. (2004). Asymptotic inference under heteroskedasticity of unknown form. *Computational Statistics and Data Analysis*, *45*, 215–233.

Cribari-Neto, F., & Cordeiro, G. M. (1996). On Bartlett and Bartlett-type corrections. *Econometric Reviews*, *15*, 339–367.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge: Cambridge University Press.

Dobson, A. J., & Barnett, A. (1998). *An introduction to generalized linear models* (3rd ed.). London: Chapman and Hall/CRC.

Efron, B. (1979). Bootstrapping methods: Another look at the jackknife. *Annals of Statistics*, *7*, 1–26.

Feigl, P., & Zelen, M. (1965). Estimation of exponential survival probabilities with con-comitant information. *Biometrics*, *21*, 826–838.

Flachaire, E. (2005). Bootstrapping heteroskedastic regression models: Wild bootstrap versus pairs bootstrap. *Computational Statistics and Data Analysis*, *49*, 361–376.

Godfrey, L. (2009). *Bootstrap tests for regression models*. New York: Palgrave MacMillan.

Harris, P. (1986). A note on Bartlett adjustments to likelihood ratio tests. *Biometrika*, *73*, 735–737.

Hayakawa, T. (1977). The likelihood ratio criterion and the asymptotic expansion of its distribution. *Annals of the Institute of Statistical Mathematics*, *29*, 359–378.

Jöeckel, K.-H. (1986). Finite sample properties and asymptotic efficiency of Monte Carlo tests. *Annals of Statistics*, *14*, 336–347.

Lawley, D. N. (1956). A general method for approximating to the distribution of the likelihood ratio criteria. *Biometrika*, *71*, 233–244.

Lemonte, A. J., & Cordeiro, G. M. (2009). Birnbaum–Saunders nonlinear regression models. *Computational Statistics and Data Analysis*, *53*, 4441–4452.

Lemonte, A. J., Ferrari, S. L. P., & Cribari-Neto, F. (2010). Improved likelihood inference in Birnbaum–Saunders regressions. *Computational Statistics and Data Analysis*, *54*, 1307–1316.

Lemonte, A. J., Cordeiro, G. M., & Moreno, G. (2012). Bartlett corrections in Birnbaum–Saunders nonlinear regression models. *Journal of Statistical Computation and Simulation*, *82*, 927–935.

Liu, R. (1988). Bootstrap procedures under non i.i.d. models. *Annals of Statistics*, *16*, 1696–1708.

McCullagh, P., & Cox, D. R. (1986). Invariants and likelihood ratio statistics. *Annals of Statistics*, *14*, 1419–1430.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London: Chapman and Hall.

Melo, T. F. N., Ferrari, S. L. P., & Cribari-Neto, F. (2009). Improved testing inference in mixed linear models. *Computational Statistics and Data Analysis*, *53*, 2573–2582.

Myers, R. H., Montgomery, D. C., & Vining, G. G. (2002). *Generalized linear models: With applications in engineering and the science*. New York: Wiley.

Rieck, J. R., & Nedelman, J. R. (1991). A log-linear model for the Birnbaum–Saunders distribution. *Technometrics*, *33*, 51–60.

Rocke, D. M. (1989). Bootstrap Bartlett adjustment in seemingly unrelated regression. *Journal of the American Statistical Association*, *84*, 598–601.

Sharp, S. A. (1975). Correction to likelihood ratio tests of hypotheses concerning the parameters of Markov chains. *Biometrika*, *62*, 595–598.

Silva, D., & Cordeiro, G. M. (2009). A computer program to improve LR tests for generalized linear models. *Communications in Statistics, Simulation and Computation*, *38*, 2184–2197.

van Giersbergen, N. P. A. (2009). Bartlett corrections in the stable AR(1) model with intercept and trend. *Econometric Theory*, *3*, 857–872.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*, 817–838.

Williams, D. A. (1976). Improved likelihood ratio tests for complete contingency tables. *Biometrika*, *63*, 33–37.

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Annals of Statistics*, *14*, 1261–1295.

Zucker, D. M., Lieberman, O., & Manor, O. (2000). Improved small sample inference in the mixed linear model: Bartlett correction and adjusted likelihood. *Journal of the Royal Statistical Society B*, *62*, 827–838.

Chapter 3
Bartlett-Type Corrections

Abstract This chapter introduces Bartlett-type corrections. They extend the Bartlett correction to chi-squared asymptotic criteria other than the likelihood ratio statistic. Bartlett-type corrections are typically applied to score and Wald test statistics. The corrected tests are usually more accurate than the uncorrected ones. As with the correction described in the previous chapter, the test error vanishes faster after the correction to the test statistic has been applied. A key difference between Bartlett and Bartlett-type corrections is that the latter may involve a polynomial of second order on the test statistic.

Keywords Bartlett-type correction · Score test · Size distortion · Power · Type I error · Wald test

3.1 Introduction

The problem of developing a correction similar to the Bartlett correction to other test statistics was posed by Cox (1988) and addressed three years later in full generality by Cordeiro and Ferrari (1991), and by Chandra and Mukerjee (1991) and Taniguchi (1991) for certain special cases; see also Mukerjee (1992). We shall focus on Cordeiro and Ferrari's results since they are more general in the sense that they allow for nuisance parameters. For a comparison of these corrections, see Rao and Mukerjee (1995). Bartlett-type corrections constitute an extension of Bartlett corrections to statistics other than LR statistics. We describe some of the main results involving Bartlett-type corrections in a unified framework and provide simulation studies that show how the independent variables and the number of nuisance parameters can affect the first-order asymptotic approximation to some test statistics in regression models.

G. M. Cordeiro and F. Cribari-Neto, *An Introduction to Bartlett Correction and Bias Reduction*, SpringerBriefs in Statistics, DOI: 10.1007/978-3-642-55255-7_3,

3.2 Bartlett-Type Correction to the Score Statistic

Suppose we have n independent random variables $Y = (Y_i, \ldots, Y_n)^\top$ whose probability or density function is indexed by a parameter vector $\theta = (\theta_1^\top, \theta_2^\top)^\top$, where $\theta_1 = (\theta_1, \ldots, \theta_q)^\top$ and $\theta_2 = (\theta_{q+1}, \ldots, \theta_p)^\top$, and hence $\dim(\theta) = p$, $\dim(\theta_1) = q$, and $\dim(\theta_2) = p - q$, for $q \leq p$. We want to test $H_0 : \theta_1 = \theta_1^{(0)}$ against a two-sided alternative hypothesis $H_1 : \theta_1 \neq \theta_1^{(0)}$, where $\theta_1^{(0)}$ is a q-vector of constants and θ_2 is a vector of nuisance parameters. Let $\ell = \ell(\theta)$ be the total log-likelihood function, and define (as in Sect. 2.2) the log-likelihood derivatives $U_i = \partial\ell/\partial\theta_i$, $U_{ij} = \partial^2\ell/\partial\theta_i\partial\theta_j$, $U_{ijk} = \partial^3\ell/\partial\theta_i\partial\theta_j\partial\theta_k$, and $U_{ijkr} = \partial^4\ell/\partial\theta_i\partial\theta_j\partial\theta_k\partial\theta_r$. The respective cumulants are $\kappa_{ij} = \mathbb{E}(U_{ij})$, $\kappa_{i,j} = \mathbb{E}(U_iU_j)$, $\kappa_{ijk} = \mathbb{E}(U_{ijk})$, $\kappa_{i,jk} = \mathbb{E}(U_iU_{jk})$, $\kappa_{i,j,k} = \mathbb{E}(U_iU_jU_k)$, $\kappa_{ijkr} = \mathbb{E}(U_{ijkr})$, $\kappa_{i,jkr} = \mathbb{E}(U_iU_{jkr})$, $\kappa_{ij,kr} = \mathbb{E}(U_{ij}U_{kr}) - \kappa_{ij}\kappa_{kr}$, and $\kappa_{i,j,k,r} = \mathbb{E}(U_iU_jU_kU_r) - \kappa_{i,j}\kappa_{k,r} - \kappa_{i,k}\kappa_{j,r} - \kappa_{i,r}\kappa_{j,k}$. Let $U = (U_1^\top, U_2^\top)^\top$ be the score function assumed partitioned in the same way as θ. Further, the expected information matrix $K = \{-\kappa_{ij}\}$ and its inverse $K^{-1} = \{-\kappa^{ij}\}$ partitioned as θ are given by

$$K = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix}, \quad K^{-1} = \begin{pmatrix} K^{11} & K^{12} \\ K^{21} & K^{22} \end{pmatrix} \quad \text{and} \quad A = \begin{pmatrix} 0 & 0 \\ 0 & K_{22}^{-1} \end{pmatrix}.$$

The unrestricted MLE of θ is $(\hat\theta_1^\top, \hat\theta_2^\top)^\top$, and the restricted estimate of θ_1 is denoted by $\tilde\theta_1$. Functions evaluated at the point $(\theta_1^{(0)\top}, \tilde\theta_2^\top)^\top$ will be distinguished by the addition of a tilde. The score statistic S (also known as the Lagrange multiplier statistic), for testing $H_0 : \theta_1 = \theta_1^{(0)}$ versus $H_0 : \theta_1 \neq \theta_1^{(0)}$, has the simple form $S = \tilde U_2^\top \tilde K^{22} \tilde U_2$. The score statistic S is one of the most used test statistics in Statistics and Econometrics due to its computational simplicity. It is especially useful when estimation under the alternative hypothesis is computationally costly since it only requires estimation under the null hypothesis.

An asymptotic expansion to the null distribution of S up to order $\mathcal{O}(n^{-1})$ was derived by Harris (1985) as

$$\begin{aligned} \Pr(S \leq x) = F_q(x) + \frac{1}{24}\,[&A_3\,F_{q+6}(x) + (A_2 - 3A_3)\,F_{q+4}(x) \\ &+ (3A_3 - 2A_2 + A_1)\,F_{q+2}(x) + (A_2 - A_1 + A_3)\,F_q(x)] + \mathcal{O}(n^{-2}), \end{aligned} \tag{3.1}$$

where $F_q(\cdot)$ denotes the cumulative distribution of the χ^2_q random variable and A_1, A_2, and A_3 are complicated functions of order $\mathcal{O}(n^{-1})$ of some joint cumulants of log-likelihood derivatives. In Appendix A.1, we present general formulae for A_1, A_2, and A_3 in expansion (3.1) and demonstrate that A_1 can be expressed in terms of the quantity ε_p given by Eqs. (2.4–2.6). Harris (1985) obtained the first three cumulants of S to order $\mathcal{O}(n^{-1})$ given by $\kappa_1(S) = q + A_1/12$, $\kappa_2(S) = 2q + (A_1 + A_2)/3$, and $\kappa_3(S) = 8q + (A_1 + 2A_2 + A_3)$. It is well known that the first three cumulants

of the χ^2_q distribution are $\kappa_1(\chi^2_q) = q$, $\kappa_2(\chi^2_q) = 2q$, and $\kappa_3(\chi^2_q) = 8q$, and then, if we know A_1, A_2, and A_3, we can obtain the first three cumulants of S to order $\mathcal{O}(n^{-1})$ and compare them with those cumulants of the reference χ^2_q random variable. Equation (3.1) holds for both simple and composite null hypotheses. More importantly, this result implies that there exists no scalar transformation based on the test statistic, which corrects all cumulants to a certain order of precision, as it is the case with the Bartlett correction to the LR statistic. Harris's results enable us to apply Hill and Davis (1968) inverse formula to (3.1) in order to obtain transformed critical values to be used in the score test (Harris 1985, p. 657). The A's can be used to obtain corrections for models based on independent, but not necessarily identically distributed observations, thus covering a number of linear and nonlinear regression models.

A correction to be directly applied to the test statistic itself was obtained by Cordeiro and Ferrari (1991). If $f_q(z)$ denotes the density function of the χ^2_q random variable, we have the recurrence formula $f_{m+2}(x) = m^{-1}\, x\, f_m(x)$. By differentiating (3.1), it is possible to demonstrate that the density function expansion of S is

$$f_S(x) = f_q(x)\,(1 + B_0 + B_1\, x + B_2\, x^2 + B_3\, x^3) + \mathcal{O}(n^{-2}), \tag{3.2}$$

where

$$B_0 = (A_2 - A_1 - A_3)/24,\ \ B_1 = (3A_3 - 2A_2 + A_1)/(24q),$$
$$B_2 = (A_2 - 3A_3)/\{24q(q+2)\},\ \ B_3 = A_3/\{24q(q+2)(q+4)\}.$$

The density function of S given in (3.2) involves a multiplicative polynomial of third degree with coefficients which depend on three constants and suggests the corrected score statistic

$$S^* = S\left(1 - \sum_{j=1}^{3} \alpha_j\, S^{j-1}\right), \tag{3.3}$$

where the multiplying factor in braces is a kind of Bartlett-type adjustment as a function of the score statistic S itself. The coefficients α_1, α_2, and α_3 can be determined as functions of the A's such that the density function of the modified statistic S^* is identical to a χ^2_q density function, when terms of order smaller than n^{-1} are neglected. Cordeiro and Ferrari (1991) determined these coefficients using two different methods. A simple one is based on the generating function of S, say $M_S(t)$, and follows by expanding $\exp\left\{-t\, \sum_{j=1}^{3} \alpha_j\, S^{j-1}\right\}$ in Taylor series up to order $\mathcal{O}(n^{-1})$ and setting the new variable $y = x(1-2t)$. We have

$$M_S(t) = M_q(t) + \frac{(1-2t)^{-r/2}}{\Gamma(r/2)\, 2^{r/2}}\, I,$$

where $M_q(t)$ is the moment generating function of the χ^2_q random variable and

$$I = \int_0^\infty y^{(r-2)/2}\, e^{-y/2} \left[\frac{(B_3 - \alpha_3 t)}{(1-2t)^3}\, y^3 + \frac{(B_2 - \alpha_2 t)}{(1-2t)^2}\, y^2 + \frac{(B_1 - \alpha_1 t)}{(1-2t)}\, y + B_0 \right].$$

The relation $M_S(t) = M_q(t)$ holds to order $\mathcal{O}(n^{-1})$ if and only if $I = 0$. The unique solution is given by $\alpha_1 = (A_1 - A_2 + A_3)/(12q)$, $\alpha_2 = (A_2 - 2A_3)/\{12q(q+2)\}$, and $\alpha_3 = A_3/\{12q(q+2)(q+4)\}$.

When the A's involve unknown parameters, such parameters should be replaced by their MLEs under H_0, which does not affect the order of approximation of the correction. The Bartlett-type correction in (3.3) is a function of the unmodified statistic S, and then, it is not a *Bartlett correction* in the classical sense. Given its similarity with the Bartlett correction, however, it is called the *Bartlett-type correction*.

Based on formulae (1) and (2) of Cox and Reid (1987), Cordeiro and Ferrari (1991) demonstrated that, under certain regularity conditions and to order $\mathcal{O}(n^{-1})$, $\Pr(S^* \leq x) = \Pr(S \leq z)$, where z is a modified critical value given by $z = x\left(1 + \sum_{j=1}^3 \alpha_j\, x^{j-1}\right)$. Under the null hypothesis, the modified test based on S^*, with the χ^2 distribution as a reference, is equivalent to the test based on the original statistic S with the modified critical value z defined as above. Nonetheless, the test based on the corrected statistic is more intuitive and easier to implement.

3.3 An Extended Result

Cordeiro and Ferrari (1991) derived a more general result which can be described as follows. Let T be a general test statistic which is asymptotically distributed as χ^2_q. Under mild regularity conditions, Chandra (1985) demonstrated that it is possible to expand $\Pr(T \leq z)$ as

$$\Pr(T \leq z) = F_q(z) + \sum_{i=0}^{k} a_i\, F_{q+2i}(z), \tag{3.4}$$

when terms of order $\mathcal{O}(n^{-2})$ or smaller are neglected. Equation (3.4) implies that the distribution function to $\mathcal{O}(n^{-1})$ of a test statistic asymptotically χ^2 distributed is, under certain conditions, a linear combination of χ^2's with $q, q+2, \ldots, q+2k$ degrees of freedom. The a_i's are linear functions of some joint cumulants of log-likelihood derivatives of the model for which T is defined. For the LR ($k = 1$) and score S ($k = 3$) statistics, the a_i's are linear functions of the A's in (3.1).

Let $\mu'_i = 2^i \Gamma(i + q/2)/\Gamma(q/2)$ be the ith moment about zero of the χ^2_q distribution, where $\Gamma(p) = \int_0^p x^{p-1}\, \mathrm{e}^{-x}\mathrm{dx}$ is the gamma function. Cordeiro and Ferrari (1991) demonstrated that the modified test statistic

$$T^* = T\left\{1 - 2\sum_{i=1}^{k}\left(\sum_{j=i}^{k} a_j\right)(\mu_i')^{-1}\, T^{i-1}\right\} \tag{3.5}$$

is distributed as χ^2_q to order $\mathcal{O}(n^{-1})$. Equation (3.5) is a very general result which can be used to improve many important tests in Econometrics and Statistics. Cordeiro and Ferrari (1991) proof is based on a theorem of Cox and Reid (1987); see their formula (1). The Bartlett-corrected test statistic T^* given in (3.5) converges, under the null hypothesis, to χ^2_q faster than the unmodified statistic, and hence, it should deliver empirical sizes closer to the nominal ones in finite samples. An extension of this result to Bartlett-type adjustments of order higher than a second order of approximation was proposed by Kakizawa (1996).

Building upon Eq. (3.5), Cordeiro et al. (1993) and Cribari-Neto and Ferrari (1995b) obtained Bartlett-type corrections to score tests in GLMs for the cases of known and unknown dispersion, respectively. Bartlett-corrected score tests for heteroskedastic linear models were considered by Cribari-Neto and Ferrari (1995a). Similar corrections for score tests in multivariate regression models were derived by Cribari-Neto and Zarkos (1995). Bartlett-type corrections to score tests for heteroskedasticity were obtained by Cribari-Neto and Ferrari (1995c). Ferrari and Arellano-Valle (1993) proposed improved score statistics for regression models with Student-t errors. Corrections to score tests that can be used in proper dispersion models were derived by Cordeiro and Ferrari (1996). Bartlett-type corrections to the class of information matrix tests, which are score tests, were considered by Cribari-Neto (1997) building upon the Edgeworth expansion in Chesher and Spady (1991).

Bartlett-type corrections are usually defined as $T^* = T\,(1-B)$, where $B = B(T)$ is a polynomial on the unmodified statistic T of the order $\mathcal{O}(n^{-1})$, such as that one in Eq. (3.5). Although most of the literature has focused on a particular form of the Bartlett-type correction given by this equation, we also consider two other forms which are equivalent to order n^{-1} to T^* and compare them through Monte Carlo simulation. There are two alternative definitions of Bartlett-type corrections, namely $T_1^* = T\,(1+B)^{-1}$ and $T_2^* = T\,\exp(-B)$, which are equivalent to T^* when terms of order smaller than $\mathcal{O}(n^{-1})$ are ignored. The latter form has the advantage of always delivering non-negative corrected statistics. Quite generally, these three forms are clearly preferable to the unmodified statistic T. They can be applied to regression models and compared through Monte Carlo simulation (see Sect. 3.8).

The modified statistic (3.5) corresponding to $k = 3$ reduces to $T^* = T\left\{1-(\alpha_1 + \alpha_2 T + \alpha_3 T^2)\right\}$. The statistic T^* is a general result to improve many important tests in Econometrics and Statistics. We can demonstrate that $\alpha_3 \geq 0$ for score statistics. It follows by local orthogonal re-parametrization of the model to make the information matrix for all parameters at the true parameter point equal to the identity matrix. In this case, Harris'(1985) expression for a_3, with $k = 3$, implies $a_3 \geq 0$ and, therefore, $\alpha_3 \geq 0$.

The statistic T^* is not always a monotone transformation of the original statistic T. To overcome this problem, the monotone transformation $K(T) = T^* + P(T)$ was

suggested by Kakizawa (1996) involving the unmodified statistic T itself and the coefficients α_1, α_2, α_3, where $P(T)$ is a polynomial of fifth degree in the original statistic T and is of order $\mathcal{O}_p(n^{-2})$. He proved that $P(T)$ reduces to

$$P(T) = \frac{1}{4}\left\{\alpha_3^2\, T + 2\alpha_2\alpha_3\, T^2 + \left(2\alpha_1\alpha_3 + \frac{4}{3}\alpha_2^2\right) T^3 + 3\alpha_1\alpha_2\, T^4 + \frac{9}{5}\alpha_1^2\, T^5\right\}. \tag{3.6}$$

A further alternative monotone transformation $\tilde{T}$ was developed by Cordeiro et al. (1998) in terms of the standard normal cumulative distribution $\Phi(\cdot)$. We shall now derive the statistic $\tilde{T}$, which is asymptotically equivalent to T^*, whose monotonicity in T is immediate. By differentiating T^* with respect to T and then integrating, we obtain a modified statistic of the form

$$\tilde{T} = \int_0^T \exp\{-(\alpha_1 + 2\alpha_2\, T + 3\alpha_3\, T^2)\}\mathrm{d}T,$$

where the integral (assuming $\alpha_3 > 0$) can be expressed in terms of the normal cumulative distribution $\Phi(\cdot)$. We obtain

$$\tilde{T} = \sqrt{\frac{\pi}{3\alpha_3}}\exp\left(\frac{\alpha_2^2}{3\alpha_3} - \alpha_1\right)\left\{\Phi\left(\sqrt{6\alpha_1}\, T + \sqrt{\frac{2}{3\alpha_3}}\alpha_2\right) - \Phi\left(\sqrt{\frac{2}{3\alpha_3}}\alpha_2\right)\right\}, \tag{3.7}$$

if $\alpha_3 > 0$ (α_3 is always non-negative), and

$$\tilde{T} = \frac{1}{2\alpha_2}\exp(-\alpha_1)\{1 - \exp(-2\alpha_2 T)\},$$

if $\alpha_3 = 0$ and $\alpha_2 \neq 0$. Note that, if $\alpha_2 = \alpha_3 = 0$, T^* is a monotone transformation of T and there is no need to define an alternative corrected statistic. The three statistics T^*, $K(T)$, and $\widetilde{T}$ are equivalent to second order, i.e., they typically differ by terms of order $\mathcal{O}_p(n^{-2})$.

The first two terms in (3.7) are $\mathcal{O}(n^{1/2})$ and $1 + \mathcal{O}(n^{-1})$, respectively, but the last term (in braces) as a function of T itself is $O_p(n^{-1/2})$. We may obtain a partial check of (3.7) by using it to derive the expression for T^*. The bracketed stochastic quantity in (3.7) can be expanded in the neighborhood of zero up to order $\mathcal{O}_p(n^{-3/2})$. As mentioned earlier, corrected chi-squared tests with better finite size properties can be performed by using the corrected statistics defined by T^*, $K(T)$, and $\tilde{T}$ and the reference χ_q^2 distribution. It can be shown that the unmodified statistics and their three corrected versions have the same powers only to order $\mathcal{O}(n^{-1/2})$.

Bartlett and Bartlett-type corrections are designed to bring the actual size of asymptotic tests close to their corresponding nominal sizes. In most cases, they are effective in doing so. However, they are not intended, however, to be corrections to increase the power of the test. It is important to bear in mind that these corrections can lead to a loss in power, much in the same way as the power of Durbin's h statistic (Durbin 1970), a transformation of the traditional Durbin–Watson statistic, can be lower than the power of the Durbin–Watson test in regression models with lagged dependent variables; see Inder (1984, 1986). However, an important result is that the untransformed statistic and its Bartlett-corrected version have the same local power to order $n^{-1/2}$. This result follows from Theorem 1 in Cox and Reid (1987). More precisely, let T be a test statistic with null distribution χ^2_q, and T^* a Bartlett-corrected statistic obtained as a transformation of T. Then, under local (Pitman) alternatives, $\Pr(T^* \geq x) = \Pr(T \geq x) + o(n^{-1/2})$.

3.4 Barttlett-Type Correction to the Wald Statistic

The Wald test is convenient to test nonlinear restrictions in linear models since it does not require estimation of the null model and therefore avoids nonlinear estimation. However, it has been shown by Gregory and Veall (1985), Lafontaine and White (1986), and others that a major drawback of this test is that it is not invariant to alternatively equivalent forms of the null hypothesis. Since many hypotheses of interest in economics are nonlinear (e.g., restrictions implied by rational expectations models), it is important to develop corrections that can be reliably applied in finite samples. Let the data generating mechanism of a random variable Y depend on a p-vector β of parameters. The following hypothesis $H_0 : h(\beta) = 0$ is to be tested against a two-sided alternative, where $h(\cdot)$ is a continuously differentiable vector function (at least to third order) in $\mathbb{R}^q$, where $q \leq p$. Let $\hat{\beta}$ be the MLE of β obtained from a sample of size n and define $q = \sqrt{n}(\hat{\beta} - \beta)$. We assume that the covariance matrix of the limiting distribution of q is the identity matrix I_p of order p.

Given this setup, the Wald statistic for testing H_0 becomes $W = n\hat{h}^\top(\hat{H}\hat{H}^\top)^{-1}\hat{h}$, where H is the $q \times p$ matrix of the first derivatives of $h(\cdot)$ with respect to the components of β. The statistic W is asymptotically distributed as χ^2_q. Phillips and Park (1988) obtained an expansion to the null distribution of W in agreement with (3.4) with $k = 3$, where the quantities a_i's are given in their paper. A Bartlett-type correction to the Wald test of nonlinear restrictions was further obtained by Ferrari and Cribari-Neto (1993). They demonstrated that the corrected Wald statistic is given by (3.5) with $k = 3$, i.e., $W^* = W\,(1 - \sum_{i=0}^{3} \alpha_i\, W^{i-1})$ is distributed as chi-squared to order n^{-1}. In other words, $\Pr(W^* \leq x) = \Pr(\chi^2_q \leq x) + \mathcal{O}(n^{-2})$.

As an example, consider the model in Lafontaine and White (1986) $Y_i = \beta_0 + \beta_1\, x_i + \epsilon_i$, where $\epsilon_i \sim \text{NID}(0, \sigma^2)$ for $i = 1, \ldots, n$. The null hypothesis of interest is $H_0 : \beta^p = 1$ against a two-tailed alternative, where p is a non-zero integer. The Wald statistic for this test reduces to $W = (\hat{\beta}^p - 1)/\left\{p\hat{\beta}^{p-1}\,\widehat{\text{var}}(\hat{\beta})\right\}$, where

$\widehat{\text{var}}(\hat{\beta}) = \hat{\sigma}^2 \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2\right)^{-1}$ is the estimated variance of $\hat{\beta}$, $\bar{x} = \sum_{i=1}^n x_i/n$ and $\hat{\sigma}^2$ is a consistent estimator of the error variance. As shown by Lafontaine and White (1986), the size of the Wald test is highly sensitive to the value of p. For this test, we can obtain $\alpha_0 = \alpha_1 = 0$, $\alpha_2 = -(2/3)(p-1)(p-2)$, and $\alpha_3 = (1/4)(p-1)^2$; see Phillips and Park (1988) and Ferrari and Cribari-Neto (1993). The Bartlett-type correction to W should be effective for moderately small values of p. The χ_q^2 approximation for the Wald statistic becomes very poor when the nonlinearity increases. It is also possible to design Bartlett-type corrections for other test statistics. For example, Cribari-Neto and Ferrari (1995a) obtained improved Wald tests for heteroskedastic linear models and Cribari-Neto and Zarkos (1995) derived similar corrections to be used in multivariate regressions.

3.5 One-Parameter Model

We consider a set of n i.i.d. random variables $Y_1, \ldots, Y_n$ following an arbitrary regular continuous or discrete one-parameter distribution indexed by an unknown scalar parameter θ. Let $\ell(\theta) = \log[\pi(y;\theta)]$ be the log-likelihood for the unknown parameter θ given one observation y. We assume that $l(\theta)$ satisfies the usual regularity conditions stated in Serfling (1980, p. 144). Let $U_\theta = d\ell(\theta)/d\theta$, $U_{\theta\theta} = d^2\ell(\theta)/d\theta^2$, etc. In what follows, we use the standard notation for the cumulants of log-likelihood derivatives (see Sect. 2.2): $\kappa_{\theta\theta} = \mathbb{E}(U_{\theta\theta})$, $\kappa_{\theta\theta\theta} = \mathbb{E}(U_{\theta\theta\theta})$, $\kappa_{\theta,\theta} = \mathbb{E}(U_\theta^2) = -\kappa_{\theta\theta}$, $\kappa_{\theta,\theta\theta} = \mathbb{E}(U_\theta U_{\theta\theta})$, $\kappa_{\theta\theta,\theta\theta} = \mathbb{E}(U_{\theta\theta}^2) - \kappa_{\theta\theta}^2$, $\kappa_{\theta\theta\theta\theta} = E(U_{\theta\theta\theta\theta})$, $\kappa_{\theta,\theta,\theta\theta} = \mathbb{E}(U_\theta^2 U_{\theta\theta}) - \kappa_{\theta,\theta}\,\kappa_{\theta\theta}$, $\kappa_{\theta,\theta,\theta,\theta} = \mathbb{E}(U_\theta^4) - 3\kappa_{\theta,\theta}^2$, and $\kappa_{\theta,\theta\theta\theta} = \mathbb{E}(U_\theta U_{\theta\theta\theta})$. We also denote the derivatives of the cumulants with superscripts as $\kappa_{\theta\theta}^{(\theta)} = d\kappa_{\theta\theta}/d\theta$, $\kappa_{\theta\theta}^{(\theta\theta)} = d^2\kappa_{\theta\theta}/d\theta^2$, etc. All κ's refer here to a single observation and then are of order $\mathcal{O}(1)$. Under these regularity conditions, the asymptotic distribution of the MLE $\hat{\theta}$ is normal $\mathcal{N}(\theta, n^{-1}\,\kappa_{\theta,\theta}^{-1})$, with an error of order $\mathcal{O}(n^{-1/2})$. The cumulants κ's satisfy certain Bartlett identities which facilitate their computation as presented in Sect. 2.2; see, also, Lawley (1956) and Cordeiro (1987).

Suppose that a non-negative statistic T for testing $H_0 : \theta = \theta^{(0)}$ in any regular one-parameter distribution, where $\theta^{(0)}$ is a given scalar, is asymptotically distributed as χ_1^2 under the null hypothesis H_0, with the error of the approximation being $\mathcal{O}(n^{-1})$. Denote the total log-likelihood by $\ell_T(\theta)$ and the total score function by $U_T(\theta) = d\ell_T(\theta)/d\theta$. Consider that T can take the form of any of the statistics LR (w), Rao score (S), Wald (W), and modified Wald (MW) given by $w = 2[\ell_T(\hat{\theta}) - \ell_T(\theta^{(0)})]$, $S = U_T(\theta^{(0)})^2/(n\,\tilde{\kappa}_{\theta,\theta})$, $W = n(\hat{\theta} - \theta^{(0)})^2\,\hat{\kappa}_{\theta,\theta}$, and $MW = n(\hat{\theta} - \theta^{(0)})^2\,\tilde{\kappa}_{\theta,\theta}$, respectively, where $\hat{\kappa}_{\theta,\theta}$ and $\tilde{\kappa}_{\theta,\theta}$ represent the expected information for one observation evaluated at $\hat{\theta}$ and $\theta^{(0)}$, respectively. The statistic T can be substantially improved through the Bartlett-type correction given by (3.5) with $k = 3$, which yields the corrected statistic in terms of the coefficients α_1, α_2, α_3, and of the original statistic T as

$$T^* = T\left[1 - \frac{1}{n}\left(\alpha_1 + \alpha_2\, T + \alpha_3\, T^2\right)\right]. \tag{3.8}$$

Expressions for the coefficients in (3.8) as functions of cumulants of log-likelihood derivatives, when T equals w, S, W, or MW, in one-parameter models, can be found in Cordeiro et al. (1995), Ferrari et al. (1996), and Santos and Cordeiro (1999). The expressions for α_1 for the Wald and modified Wald statistics are identical, and therefore, in what follows, we provide results only for the Wald statistic. The expressions are given explicitly by:

LR statistic (w)

$$\alpha_1 = \frac{5\kappa_{\theta\theta\theta}^2 + 24\kappa_{\theta\theta}^{(\theta)}\left(\kappa_{\theta\theta}^{(\theta)} - \kappa_{\theta\theta\theta}\right)}{12\,\kappa_{\theta\theta}^3} - \frac{\kappa_{\theta\theta\theta\theta} + 4\left(\kappa_{\theta\theta}^{(\theta\theta)} - \kappa_{\theta\theta\theta}^{(\theta)}\right)}{4\kappa_{\theta\theta}^2}, \tag{3.9}$$

$$\alpha_2 = \alpha_3 = 0. \tag{3.10}$$

Score statistic (S)

$$\alpha_1 = \frac{-\kappa_{\theta,\theta,\theta}^2}{36\,\kappa_{\theta\theta}^3}, \tag{3.11}$$

$$\alpha_2 = \frac{10\kappa_{\theta,\theta,\theta}^2 + 3\,\kappa_{\theta\theta}\kappa_{\theta,\theta,\theta,\theta} - 9\kappa_{\theta\theta}^3}{36\,\kappa_{\theta\theta}^3}, \tag{3.12}$$

$$\alpha_3 = \frac{-5\kappa_{\theta,\theta,\theta}^2 - 3\kappa_{\theta\theta}\kappa_{\theta,\theta,\theta,\theta} + 9\kappa_{\theta\theta}^3}{12\,\kappa_{\theta\theta}^3}. \tag{3.13}$$

Wald statistic (W)

$$\alpha_1 = \frac{-44\kappa_{\theta\theta\theta}^2 + 120\kappa_{\theta\theta\theta}\kappa_{\theta\theta}^{(\theta)} - 81(\kappa_{\theta\theta}^{(\theta)})^2 + 12\kappa_{\theta\theta}\kappa_{\theta,\theta,\theta\theta} - 3\kappa_{\theta\theta}\kappa_{\theta,\theta,\theta,\theta}}{12\,\kappa_{\theta\theta}^3}, \tag{3.14}$$

$$\alpha_2 = \frac{-10{\kappa_{\theta\theta\theta}}^2 + 48\left(2\kappa_{\theta\theta\theta} - 3\kappa_{\theta\theta}^{(\theta)}\right)^2 + 6\left(\kappa_{\theta\theta}^{(\theta)} - \kappa_{\theta\theta\theta}\right)\left(17\kappa_{\theta\theta\theta} - 45\kappa_{\theta\theta}^{(\theta)}\right)}{72\kappa_{\theta\theta}^3} + \frac{3\kappa_{\theta\theta,\theta\theta} + 20\kappa_{\theta\theta\theta}^{(\theta)} - 11\kappa_{\theta\theta\theta\theta} - 12\kappa_{\theta\theta}^{(\theta\theta)}}{12\kappa_{\theta\theta}^2}, \tag{3.15}$$

$$\alpha_3 = -\frac{\kappa_{\theta\theta\theta}^2}{36\,\kappa_{\theta\theta}^3}. \tag{3.16}$$

Modified Wald statistic (MW)

$$\alpha_2 = \frac{63\kappa_{\theta\theta\theta}\kappa_{\theta\theta}^{(\theta)} - 22\kappa_{\theta\theta\theta}^2 - 45(\kappa_{\theta\theta}^{(\theta)})^2}{18\,\kappa_{\theta\theta}^3} + \frac{4\kappa_{\theta\theta\theta\theta} - 4\kappa_{\theta\theta\theta}^{(\theta)} - 4\kappa_{\theta,\theta,\theta\theta} - 3\kappa_{\theta,\theta,\theta,\theta}}{12\,\kappa_{\theta\theta}^2}, \tag{3.17}$$

$$\alpha_3 = -\frac{\left(3\kappa_{\theta\theta}^{(\theta)} - \kappa_{\theta\theta\theta}\right)^2}{36\,\kappa_{\theta\theta}^3}. \tag{3.18}$$

Some of the third- and fourth-order cumulants that appear in (3.9)–(3.18) can be more easily computed using the Bartlett identities

$$\begin{aligned}
&\kappa_{\theta,\theta} = -\kappa_{\theta\theta},\ \kappa_{\theta,\theta,\theta} = 2\kappa_{\theta\theta\theta} - 3\kappa_{\theta\theta}^{(\theta)},\ \kappa_{\theta,\theta\theta} = \kappa_{\theta\theta}^{(\theta)} - \kappa_{\theta\theta\theta},\\
&\kappa_{\theta,\theta,\theta,\theta} = -3\kappa_{\theta\theta\theta\theta} + 8\kappa_{\theta\theta\theta}^{(\theta)} - 6\kappa_{\theta\theta}^{(\theta\theta)} + 3\kappa_{\theta\theta,\theta\theta},\\
&\kappa_{\theta,\theta,\theta\theta} = \kappa_{\theta\theta\theta\theta} - 2\kappa_{\theta\theta\theta}^{(\theta)} + \kappa_{\theta\theta}^{(\theta\theta)} - \kappa_{\theta\theta,\theta\theta},\ \kappa_{\theta,\theta\theta\theta} = \kappa_{\theta\theta\theta}^{(\theta)} - \kappa_{\theta\theta\theta\theta}.
\end{aligned} \tag{3.19}$$

Equation (3.19) usually facilitate the computation of the cumulants $\kappa's$. We can use an algebraic manipulation software such as Mathematica to evaluate the $\kappa's$ for several one-parameter continuous and discrete distributions and then obtain the coefficients α_1, α_2, and α_3 for all four test statistics. So, we can derive corrected statistics from Eqs. (3.9–3.18) by evaluating these coefficients at $\theta^{(0)}$. All four corrected statistics defined by T^* have a χ_1^2 distribution to order $\mathcal{O}(n^{-1})$ under the null hypothesis.

Cordeiro and Stosic (2008) developed a simple program (script) that may be used with algebraic manipulation software MATHEMATICA (the script was written and tested on MATHEMATICA version 5.2.0.0) to obtain, from Eqs. (3.9–3.19), the coefficients α_1, α_2, and α_3 in closed form for the four statistics for testing $H_0 : \theta = \theta^{(0)}$. While the MATHEMATICA symbolic computation software has currently the ability to deal with analytic expressions of formidable size and complexity, limitations still exist, and it turns out that the complexity of the formulae involved in calculating the cumulants of log-likelihood derivatives for some distributions exceed its capacity. Even for these cases, the current script may be expected to produce results on future versions of the software (under the assumption that backward compatibility of the scripting language is maintained).

It should be pointed out that the four statistics under study depend on the functional form of the density function. In particular, while the LR statistic and its Bartlett correction are both invariant under re-parametrization, the score, Wald, and modified Wald statistics together with their corresponding Bartlett-type corrections are not invariant. Therefore, different choices of the parametrization of the model in general yield different analytical expressions for these statistics and their corresponding

Bartlett-type corrections. Moreover, it may turn out that a special choice is more amenable to algebraic manipulations then another. Hence, it may be expected that in some cases, MATHEMATICA produces closed-form expressions for a given functional form of the parameter, but not for other choices. Nonetheless, the presented procedure is of quite general nature and should yield correct corresponding results for arbitrary selected parameter. In summary, if for a given choice, MATHEMATICA does not yield a closed-form expression, others re-parameterizations should be tested.

After specifying the form and the domain of the density function $f = f(y; \theta)$ as well as the assumptions to be made on y and θ (e.g., $y \in R$ or $\theta > 0$), the program first defines and evaluates (analytically) all the cumulants (κ's), which are then inserted into Eqs. (3.9–3.18) to produce in regular one-parameter distributions closed-form expressions for the coefficients α_1, α_2, and α_3 for all corrected test statistics defined from Eqs. (3.5–3.7).

The MATHEMATICA script that can be used to obtain the corrected LR, score, Wald, and modified Wald statistics is described in Appendix A.2. In this script, the function *KK* embedded inside the module *corrections* performs the actual calculation of most of the cumulants, the only exception being the cumulant $\kappa_{\theta\theta,\theta\theta}$ (denoted by symbol *kt22* in the computer code) which is calculated separately. The module *corrections* receives as arguments the form of the density function $f(\cdot)$, the lower bound p and upper bound q of integration (summation), the required condition on the parameter θ (as will be shown in Appendix A2 for some special cases), and a flag indicating whether the distribution is continuous ($cont = 1$) or discrete ($cont = 0$). The expressions for correction terms, evaluated by straightforward implementation of Eqs. (3.9–3.18), are printed on the screen and stored in global variables $LR1$, $S1$, $S2$, $S3$, $W1$, $W2$, $W3$, $MW1$, $MW2$, and $MW3$, for possible posterior manipulation. It should be noted that each invocation of module *corrections* overwrites the results stored in the global variables representing the correction terms, and therefore, the results should be stored under different names if further processing is necessary (in particular if one wants to simultaneously manipulate results for different density functions).

From a programming viewpoint, the above implementation may be considered quite elementary: This choice was made in order to keep the transparency of the code and facilitate possible modifications. In particular, we use a single module, with a single embedded function. We also use only one local variable for intermediate calculations, while all the cumulants are kept as global, in order to facilitate their individual inspection and interactive manipulation. The *print* statements have the sole purpose of immediate visualization of the results, and they may be commented out, if the module is used within a concrete application.

After the above module has been copied into a MATHEMATICA notebook, it should be activated (e.g., on the Windows platform by clicking the mouse on any line, and pressing 'Shift+Enter' key combination). In order to obtain the expressions of the corrections for a given distribution, the user should specify the form of the density function and then invoke module *corrections* with the correct range of integration, the conditions on the parameter θ, and the type (discrete or continuous). Each item starts with the name of the distribution, followed by several lines of MATHEMATICA

code, and then the resulting expressions for the corrections. As the output format and notation implemented for special functions by the software platform is somewhat unusual, we display the results in a notionally simplified form in comparison with the actual output. In particular, we use the notation $\psi'(x)$ and $\psi''(x)$ for the first and second derivatives of the digamma function $\psi(x) = \mathrm{d}\log\Gamma(x)/\mathrm{d}x$, respectively, whereas MATHEMATICA output notation for the nth polygamma function is `PolyGamma[n,x]`. Also, we use notation $\zeta(x)$ for the Riemann zeta function and γ for the Euler's constant, which are denoted in MATHEMATICA output by `Zeta[x]` and `EulerGamma`, respectively. In some cases, analytical expressions obtained turn out too large to be given here explicitly, so we present results in tabular form for some values of the parameter θ. After running the module *corrections* for a given distribution, the user may create a table of numerical values by invoking the script

```
TableForm[
  Transpose[
    Table[{x, LR1 /. \[Theta] -> x,
    S1 /. \[Theta] -> x, S2 /. \[Theta] -> x,
    S3 /. \[Theta] -> x, W1 /. \[Theta] -> x,
    W2 /. \[Theta] -> x, W3 /. \[Theta] -> x,
    MW2 /. \[Theta] -> x, MW3 /. \[Theta] -> x},
    {x, 0.5, 3.0, 0.5}]],
  TableHeadings -> {{"\[Theta]", "LR1", "S1", "S2", "S3",
  "W1", "W2", "W3", "MW2", "MW3"}, {}}],
```

where θ in the above example varies between 0.5 and 3.0, with step of 0.5.

Finally, the coefficients for the corrected statistics obtained for five continuous distributions (Cauchy, chi-squared, Maxwell, Rayleigh, and Student t) and two discrete distributions (binomial and Poisson) using the script developed by Cordeiro and Stosic (2008) agree with previous results reported for the LR statistic (Cordeiro et al. 1995), for the score statistic (Ferrari et al. 1996), and for the Wald and modified Wald statistics (Santos and Cordeiro 1999).

3.6 The p^* Approximation

For a scalar parameter θ of interest, there are two familiar first-order statistics to measure the departure of $\hat{\theta}$ from θ: (i) the Wald departure $q = (\hat{\theta} - \theta)\,|\hat{J}_T|^{1/2}$, where $\hat{J}_T$ is the total observed information $J_T = -\mathrm{d}^2\ell_T(\theta)/\mathrm{d}\theta^2$ for θ evaluated at $\hat{\theta}$, and (ii) the directed LR statistic $r = \mathrm{sgn}(\hat{\theta} - \theta)\sqrt{w}$, where $w(\theta) = 2\left\{\ell_T(\hat{\theta}) - \ell_T(\theta)\right\}$. The corresponding first-order p-values are $\Phi(q)$ and $\Phi(r)$. However, for small sample sizes, q and r can be very misleading, since the errors of the approximations for both statistics to the standard normal distribution are of order $\mathcal{O}(n^{-1/2})$.

A major development in likelihood-based inference is that the likelihood function can be used directly to provide a more accurate approximation for inference about θ than the two above-mentioned first-order normal approximations (i)

and (ii). The main result for this development is the Barndorff-Nielsen (1983) p^* approximation for the density function of the MLE $\hat{\theta}$ given by

$$p^*(\hat{\theta}; \theta) = c(\theta)\, |\hat{J}_T|^{1/2} \exp\{-w(\theta)/2\},$$

where the normalizing constant $c(\theta)$ is determined numerically. The accuracy of the p^* approximation is usually of order $\mathcal{O}(n^{-3/2})$. Most of the higher-order asymptotic theory for likelihood inference can be justified by means of the p^* formula, and therefore, a variety of alternative corrections have been proposed to improve the asymptotic standard normal approximation to the distribution of the statistic r. By integrating the p^* approximation, the cdf of r with an error usually of order $\mathcal{O}(n^{-3/2})$ can be expressed as

$$F(r) = \Phi(r^*) = \Phi(r) + (r^{-1} - u^{-1})\, \phi(r), \tag{3.20}$$

where $u = \hat{J}_T^{-1/2}\, w(\theta)/2$, $\phi(\cdot)$ is the standard normal density function and $r^* = r + r^{-1} \log(u/r)$ is the modified directed likelihood due to Barndorff -Nielsen (1990) (the term $r^{-1} \log(u/r)$ in r^* is of order $\mathcal{O}_p(n^{-1/2})$).

The expression for r^* has been prominent in likelihood theory. Equation (3.20) offers two alternatives for the approximate calculation of tail probabilities (the first and second terms on the right-hand side of (3.20)), both being extremely accurate over the range of r, representing quite simple means to compute the p-value for inference on θ. Equation (3.20), known as the Lugannani and Rice (1980) formula, applied to tail areas of one-parameter distributions, also provides an approximate cdf for $\hat{\theta}$ with relative error $\mathcal{O}(n^{-3/2})$. Usually, computing probabilities from the two formulae on the right-hand side of Eq. (3.20) yield slightly different results.

Under moderate regularity conditions, and assuming that the log-likelihood has the usual asymptotic properties as $n \to \infty$, the p-value calculated from (3.20) is accurate to third order only when the distribution of y is continuous. However, we can apply this result for the analysis of discrete data. The statistics r and r^* are in principle easy to be implemented in software packages with algebraic capabilities and are generally quite accurate. However, they require ML estimation and are data dependent. Bartlett and Bartlett-type corrections are based on the geometry of the model (independent of the data). They thus shed some light on for which regions of the parameter space the χ^2 approximation can be expected to work well, without previous knowledge of $\hat{\theta}$ (in fact, in some cases, they are actually independent of $\hat{\theta}$). Finally, it should be mentioned that the approach of Sect. 3.5, applied to improve chi-squared statistics for one-parameter distributions, can be extended to multiparameter cases in a more straightforward way than the normal distribution-based statistics (see Sects. 3.7 and 3.9).

3.7 Generalized Linear Models

A fairly general framework of regression models is the GLMs described in Sect. 2.5. Consider a GLM, where $Y = (Y_1, \ldots, Y_n)^\top$ is a vector of independent variables and each y_i has a probability or density function in the exponential family (2.12). The mean and variance of y_i are $\mathbb{E}(Y_i) = \mu_i = b'(\theta_i)$ and $\text{var}(Y_i) = \phi^{-1}V_i$, where ϕ^{-1} is the dispersion parameter, $V = V(\mu) = d\mu/d\theta$ is the variance function, and $\theta = \int V^{-1}d\mu = q(\mu)$ is a strictly monotonic function of the mean. The linear predictor is given by $\eta = \sum_{j=1}^{p} \beta_j x_j = X\beta$, where X is an $n \times p$ matrix that contains the values of explanatory variables (of rank p) and β is a p-vector of unknown parameters to be estimated. The mean of the dependent variable is then related to the linear predictor through a strictly monotonic twice differentiable link function $d(\mu) = \eta$, which is usually assumed known. GLMs include as special cases the normal linear regression, gamma, inverse Gaussian, Poission, logit, and probit models. For example, $V = 1$ and $\mu = \eta$ for the normal linear model with variance ϕ^{-1}. In this section, we develop Bartlett-type corrections for score tests in GLMs. Similar Bartlett corrections for LR statistics are discussed in Sect. 2.5.1 (see, also, Cordeiro 1983, 1987).

Suppose the vector β is partitioned as $\beta = (\beta_1^\top, \beta_2^\top)$, where $\beta_1 = (\beta_1, \ldots, \beta_q)^\top$ $(q \leq p)$ and $\beta_2 = (\beta_{q+1}, \ldots, \beta_p)^\top$, thus inducing a corresponding partition of the model matrix $X = (X_1\ X_2)$. We want to test the null hypothesis $H_0 : \beta_1 = \beta_1^{(0)}$, where $\beta_1^{(0)}$ is a q-vector of known constants, against a two-sided alternative hypothesis. The score statistic for this test is given by

$$S = \tilde{s}^\top \tilde{W}^{1/2} X_1 (\tilde{R}^\top \tilde{W} \tilde{R})^{-1} X_1^\top \tilde{W}^{1/2} \tilde{s},$$

where $W = \text{diag}\{w_1, \ldots, w_n\}$, $s = (s_1, \ldots, s_n)^\top$, $R = X_1 - X_2(X_2^\top W X_2)^{-1} X_2^\top W X_1$, and tildes denote evaluation at the restricted MLEs. For $i = 1, \ldots, n$, we have $w_i = V_i^{-1}(d\mu_i/d\eta_i)^2$ and $s_i = \phi^{1/2} V_i^{-1/2}(y_i - \mu_i)$. When the dispersion parameter is unknown, we obtain a two-parameter full exponential family with canonical parameters ϕ and $\phi\theta$, and the quantity $a(y, \phi)$ in (2.12) can be decomposed as $a(y, \phi) = d_1(\phi) + d_2(\phi)$. Each distribution in (2.12) has specific functions for $d_1(\phi)$ and $d_2(y)$. For example, $d_1(\phi) = \log(\phi/2)/2$ and $d_2(y) = 0$ for the normal distribution with variance ϕ^{-1}.

Following the general expressions in Appendix A.1, Cordeiro et al. (1993) derived the A's that define the Bartlett-type correction to the score statistic S when ϕ is known. The case of ϕ unknown was discussed further by Cribari-Neto and Ferrari (1995b). They demonstrated that $A_1 = A_{1,\beta} + A_{1,\beta\phi}$, $A_2 = A_{2,\beta} + A_{2,\beta\phi}$ and $A_3 = A_{3,\beta} + A_{3,\beta\phi}$, where $A_{1,\beta}$, $A_{2,\beta}$, and $A_{3,\beta}$ are the A's for the known dispersion case (Cordeiro et al. 1993), and $A_{1,\beta\phi}$, $A_{2,\beta\phi}$, and $A_{3,\beta\phi}$ are some extra terms that account for the uncertainty involved in the estimation of ϕ^{-1}. Cordeiro et al. (1993) demonstrated that

$$A_{1,\beta} = \phi^{-1}\{3\,\mathbf{1}^\top F Z_{2d}(Z - Z_2)Z_{2d}F\,\mathbf{1} + 6\,\mathbf{1}^\top F Z_{2d}Z_2(Z - Z_2)d(F - G)\mathbf{1} - 6\,\mathbf{1}^\top F\{Z^{(2)}2 \odot (Z - Z_2)\}(2G - F)\mathbf{1} - 6\,\mathbf{1}^\top H(Z - Z_2)_d Z_{2d}\mathbf{1}\},$$

$$A_{2,\beta} = \phi^{-1}\{-3\,\mathbf{1}^\top (F - G)(Z - Z_2)_d Z_2(Z - Z_2)_d(F - G)\mathbf{1} - 6\,\mathbf{1}^\top F Z_{2d}(Z - Z_2)(Z - Z_2)_d(F - G)\mathbf{1} - 6\,\mathbf{1}^\top (F - G)\{(Z - Z_2)^{(2)} \odot Z_2\}(F - G)\mathbf{1} + 3\,\mathbf{1}^\top B(Z - Z_2)_d^{(2)}\mathbf{1}\}$$

and

$$A_{3,\beta} = \phi^{-1}\{3\,\mathbf{1}^\top (F - G)(Z - Z_2)_d(Z - Z_2)(Z - Z_2)_d(F - G)\mathbf{1} + 2\,\mathbf{1}^\top (F - G)(Z - Z_2)^{(3)}(F - G)\mathbf{1}\},$$

where

$$Z = X(X^\top W X)^{-1}X^\top,\ Z_2 = X_2(X_2^\top W X_2)^{-1}X_2^\top,\ Z_d = \mathrm{diag}\{z_{ll}, \ldots, z_{nn}\},$$
$$Z_{2d} = \mathrm{diag}\{z_{2ll}, \ldots, z_{2nn}\},\ F = \mathrm{diag}\{f_1, \ldots, f_n\},\ G = \mathrm{diag}\{g_l, \ldots, g_n\},$$
$$B = \mathrm{diag}\{b_1, \ldots, b_n\} \text{ and } H = \mathrm{diag}\{h_1, \ldots, h_n\},$$

with $\mathbf{1}$ being an $n \times 1$ vector of ones, '$\odot$' denoting the Hadamard product of matrices and

$$f = \frac{1}{V}\frac{\mathrm{d}\mu}{\mathrm{d}\eta}\frac{\mathrm{d}^2\mu}{\mathrm{d}\eta^2},\quad g = \frac{1}{V}\frac{\mathrm{d}\mu}{\mathrm{d}\eta}\frac{\mathrm{d}^2\mu}{\mathrm{d}\eta^2} - \frac{1}{V^2}\frac{\mathrm{d}V}{\mathrm{d}\mu}\left(\frac{\mathrm{d}\mu}{\mathrm{d}\eta}\right)^3,$$
$$b = \frac{1}{V^3}\left(\frac{\mathrm{d}\mu}{\mathrm{d}\eta}\right)^4\left\{\left(\frac{\mathrm{d}V}{\mathrm{d}\mu}\right)^2 + V\frac{\mathrm{d}^2V}{\mathrm{d}\mu^2}\right\},$$
$$h = \frac{1}{V^2}\frac{\mathrm{d}V}{\mathrm{d}\mu}\left(\frac{\mathrm{d}\mu}{\mathrm{d}\eta}\right)^2\frac{\mathrm{d}^2\mu}{\mathrm{d}\eta^2} + \frac{1}{V^2}\frac{\mathrm{d}^2V}{\mathrm{d}\mu^2}\left(\frac{\mathrm{d}\mu}{\mathrm{d}\eta}\right)^4.$$

Further, we have Cribari-Neto and Ferrari (1995b)

$$A_{1,\beta\phi} = \frac{6q\left\{d_{(3)} - (p - q - 2)d_{(2)}\right\}}{nd_{(2)}^2},\quad A_{2,\beta\phi} = \frac{3q(q + 2)}{nd_{(2)}},$$

and $A_{3,\beta\phi} = 0$, where $d_{(2)} = d_{(2)}(\phi) = \phi^2 d''_1(\phi)$ and $d_{(3)}(\phi) = \phi^3 d'''_1(\phi)$.

For the normal linear model, the A's are obtained as special cases by taking $V = 1$ and $\eta = \mu$. It should also be noted that similar results for Poisson regression and logit and probit models that are commonly used in the econometrics literature can also be obtained as special cases of the formulae above. For Poisson models, $V = \mu$, and for logit and probit models, $V = \mu(1 - \mu)$. A generalization of the result presented above to nonlinear models can be found in Ferrari et al. (1997).

We can also consider the test of the null hypothesis $H_0 : \phi = \phi^{(0)}$ against the alternative $H_1 : \phi \neq \phi^{(0)}$, where $\phi^{(0)}$ is a given scalar. For example, in Poisson regression models, one might want to test the hypothesis that $\phi = 1$ against the alternative of overdispersion or underdispersion. The A's for the Bartlett-type correction of the score statistic are (Cordeiro et al. 1993)

$$A_1 = -\frac{3p(p-2)}{d_{(2)}}, \quad A_2 = -\frac{3\{2pd_{(3)} + d_{(4)}\}}{d_{(2)}^2}, \quad A_3 = -\frac{5d_{(3)}^2}{d_{(2)}^3},$$

where $d_{(4)} = d_{(4)}(\phi) = \phi^4 d\prime\prime_1(\phi)$.

3.8 Simulation Results

In this section, we report some simulation results comparing the sizes and powers of the LR and score tests and of the tests based on their modified statistics. We adopt three versions for the LR statistics (see Sect. 2.4), namely w, $w^* = w/(1+b/q)$, and $w_1^* = w(1-b/q)$, and six versions for the score statistics as follows. From Eq. (3.3), we define $B = (c_1 + c_2 S + c_3 S^2)$ and the score tests based on S, $S^* = S(1-B)$, $S_1^* = S/(1+B)$, $S_2^* = S\exp(-B)$, $\tilde{S}$ given by (3.7) and the score test based on the modified critical value z defined at the end of Sect. 3.2. For a given critical value x_α, the labels (1), (2), (3), (4), (5), (6), (7), (8), and (9) refer to $\Pr(w \geq x_\alpha)$, $\Pr(w^* \geq x_\alpha)$, $\Pr(w_1^* \geq x_\alpha)$, $\Pr(S \geq x_\alpha)$, $\Pr(S^* \geq x_\alpha)$, $\Pr(S_1^* \geq x_\alpha)$, $\Pr(S_2^* \geq x_\alpha)$, $\Pr(\tilde{S} \geq x_\alpha)$, and $\Pr(S \geq z_\alpha)$, respectively.

For the simulations, we consider eight normal models with mean μ and variance $\sigma^2 = 1$ and eight gamma models with mean μ and variance μ^2/ϕ, where $\phi = 2$ and μ in both cases is related to unknown regression parameters as

$$\begin{aligned}
&(i) \quad \eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2, \\
&(ii) \quad \eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3, \\
&\qquad\qquad \vdots \\
&(viii) \quad \eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9.
\end{aligned}$$

For the normal and gamma models, we adopt the identity and reciprocal link functions, respectively. The null hypothesis is $H_0 : \beta_1 = \beta_2 = 0$. Denoting the number of regression parameters by p and the number of restrictions under H_0 by q, we vary the number of nuisance parameters as $p-q = 1, 2, \ldots, 8$ for the two models described before. Ten thousand samples of $n = 30$ observations were generated for each model with $\beta_0 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = 0.05$. The covariates were taken as 3 replicates of 10 random draws from the Cauchy, χ_3^2, $F(2,5)$, $F(3,3)$, $LN(0,1)$, $U(0,1)$, $\mathcal{N}(0,1)$, $\mathcal{N}(0,2)$, and t_3 distribution. The values for x's were kept constant throughout the experiment.

Table 3.1 Sizes of tests for the normal model with $\sigma^2 = 1$, $q = 2$, $p - q = 1, \ldots, 8$

$p-q$	Nominal levels (%)	LR			S					
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	1.0	1.5	1.0	1.0	0.7	0.9	0.9	0.9	0.9	0.9
2	1.0	1.8	1.1	1.0	0.8	0.9	0.9	0.9	0.9	0.9
3	1.0	2.1	1.1	1.0	2.6	1.1	1.3	1.2	1.3	1.1
4	1.0	2.3	1.0	1.0	1.1	1.0	0.9	0.9	0.9	0.9
5	1.0	2.6	1.2	1.0	1.3	1.0	1.0	1.0	1.0	0.9
6	1.0	3.4	1.4	1.0	1.8	1.0	1.0	1.0	1.0	1.0
7	1.0	3.7	1.6	1.0	2.3	1.1	1.1	1.1	1.1	1.0
8	1.0	5.0	1.6	1.0	2.6	1.1	1.3	1.2	1.3	1.1
1	5.0	6.8	4.9	4.8	4.8	4.8	4.8	4.8	4.8	4.8
2	5.0	7.4	5.0	4.7	5.3	4.7	4.7	4.7	4.8	4.7
3	5.0	7.7	5.0	4.7	5.8	4.7	4.7	4.7	4.7	4.7
4	5.0	8.7	5.4	4.7	6.5	4.8	5.0	4.9	4.9	4.8
5	5.0	10.0	5.6	4.7	7.5	4.9	5.1	5.0	5.1	5.0
6	5.0	11.0	6.1	5.0	8.5	5.3	5.6	5.4	5.5	5.4
7	5.0	12.0	6.2	4.6	9.3	5.4	5.8	5.7	5.7	5.7
8	5.0	14.0	7.6	5.4	11.1	6.1	7.0	6.6	6.6	6.6
1	10.0	12.3	10.0	10.0	10.6	10.0	10.0	10.0	10.0	10.0
2	10.0	13.5	10.0	10.1	11.5	10.1	10.2	10.2	10.2	10.1
3	10.0	14.1	10.0	9.4	11.9	10.0	10.0	10.0	10.0	10.0
4	10.0	15.3	10.1	9.6	12.8	10.0	10.0	10.0	10.0	10.0
5	10.0	16.4	10.1	9.9	14.4	10.2	10.7	10.5	10.5	10.5
6	10.0	18.3	10.1	9.7	15.9	10.2	10.9	10.7	10.7	10.7
7	10.0	19.5	12.0	10.0	17.3	10.4	11.4	10.9	10.9	11.0
8	10.0	22.0	13.0	10.4	19.1	11.0	12.4	11.7	11.8	12.0

Tables 3.1 and 3.2 display the estimated sizes of the LR and score tests and their modified versions above for $p - q = 1, \ldots, 8$ corresponding to the nominal sizes $\alpha = 1, 5$ and 10 %. It is clear from these figures that the size performance of the usual LR and score tests deteriorates as the number of nuisance regression parameters increases. In fact, for $p - q = 8$, both tests are quite oversized. The corrected statistics w^*, w_1^*, S^*, S_1^*, S_2^*, and $\tilde{S}$ are quite effective in bringing the sizes of the modified tests closer to the nominal sizes especially if $p - q$ is not small. It is clear that the four Bartlett-corrected score statistics have a similar size behavior and that all corrected tests outperform the original score test, especially when $p - q$ is large. In particular, S_2^* has a slightly superior behavior for small samples followed by $\tilde{S}$ and then S_1^*. The corrected score test based on the modified critical value z_α provides good χ_2^2 approximation for the modified score test. In Tables 3.3 and 3.4, we present the estimated sizes of the LR and score tests and their modified versions when $p = 10$ and $q = 2$ for the normal and gamma models, respectively, by varying the number of observations $n = 20$, 30, 40, and 50. As the sample size increases, all

Table 3.2 Sizes of tests for the gamma model with $\phi = 2$, $q = 2$, $p - q = 1, \ldots, 8$

$p-q$	Nominal levels (%)	LR				S				
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	1.0	1.6	1.1	1.0	1.0	1.1	1.1	1.1	1.1	1.1
2	1.0	1.8	1.0	0.9	0.9	1.0	1.0	1.0	1.0	1.0
3	1.0	2.1	1.1	1.1	1.1	0.9	0.9	0.9	1.0	1.0
4	1.0	2.7	1.2	1.1	1.2	1.0	1.0	1.0	1.0	1.1
5	1.0	2.9	1.3	1.0	1.3	1.0	1.0	1.0	1.1	1.1
6	1.0	3.4	1.3	1.0	1.2	0.9	0.9	0.9	0.9	1.0
7	1.0	4.3	1.7	1.1	1.6	1.1	1.1	1.1	1.1	1.1
8	1.0	5.6	1.7	0.8	2.1	1.1	1.1	1.1	1.1	1.0
1	5.0	6.4	5.0	4.9	4.5	4.8	4.8	4.8	4.8	4.8
2	5.0	7.0	5.2	5.1	4.9	4.9	4.9	4.9	4.9	4.9
3	5.0	7.7	5.1	4.8	5.8	4.8	4.8	4.8	4.8	4.8
4	5.0	8.9	5.3	4.8	6.5	4.9	5.0	5.0	5.0	5.0
5	5.0	10.2	5.8	5.1	7.3	5.1	5.2	5.1	5.2	5.1
6	5.0	10.9	6.0	4.9	7.2	4.9	5.1	5.0	5.1	5.0
7	5.0	12.7	6.7	5.0	8.3	5.3	5.6	5.5	5.6	5.3
8	5.0	14.7	7.5	5.1	9.8	5.6	6.1	5.9	6.0	5.7
1	10.0	12.0	10.0	9.7	9.7	9.9	9.9	9.9	9.9	9.9
2	10.0	13.3	10.2	9.9	10.5	10.0	10.0	10.0	10.0	9.9
3	10.0	13.6	10.2	9.7	11.8	10.0	10.1	10.0	10.0	10.0
4	10.0	16.1	10.9	9.8	13.2	10.0	10.2	10.1	10.1	10.1
5	10.0	17.0	11.3	10.1	13.6	10.4	10.7	10.6	10.6	10.6
6	10.0	18.1	11.5	9.8	14.4	10.2	10.5	10.4	10.5	10.3
7	10.0	20.1	12.5	10.0	16.0	10.7	11.3	11.0	11.1	10.9
8	10.0	23.1	13.2	9.7	18.4	10.5	11.6	11.1	11.2	11.0

statistics converge to the χ^2_2 distribution but the rate of convergence of the modified statistics is much higher.

Some power simulations not reported here for the above experiment were conducted using tabulated and not estimated critical values. This was done mainly because none of the tests is oversized. We are then comparing the powers of *level* α (as opposed to *size* α) tests. The results showed that S_2^* has the best power performance. For the normal model, all four corrected score tests had slightly higher power than the original test. The power behavior of the corrected tests was similar. For the gamma model, S_2^* was followed by $\tilde{S}$, S, and S_1^*. Although S^* is the most used version of the Bartlett-type corrected score statistic, the other alternative forms considered here were slightly more powerful under the alternative hypothesis. It should be remarked that when the power comparisons are based on estimated critical values so that all tests are forced to have the same size, some corrected tests become considerably less powerful than the original test. This illustrates the fact that in some cases, the size adjustment comes at the expense of some loss in power.

A second experiment study conducted to verify the superiority of the tests based on the monotone corrected statistics $\tilde{S}$ and $K(S)$ over the usual corrected score

Table 3.3 Sizes of tests for the normal model ($p = 10, q = 2$)

n	Nominal levels (%)	LR			S				
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	1.0	9.5	3.0	1.0	4.3	1.1	1.5	1.2	1.3	1.1
20	5.0	21.6	10.1	4.7	15.9	5.9	7.8	6.9	7.1	6.7
	10.0	30.7	17.2	9.5	25.8	11.4	14.5	13.7	13.7	13.6
	1.0	5.0	1.6	1.0	2.6	1.1	1.3	1.2	1.3	1.1
30	5.0	13.8	7.6	5.4	11.1	6.1	7.0	6.6	6.6	6.6
	10.0	22.0	13.0	10.4	19.1	11.0	12.4	11.8	11.8	12.0
	1.0	3.5	1.5	1.1	2.2	1.1	1.3	1.2	1.3	1.1
40	5.0	10.8	6.2	5.3	8.9	5.5	6.0	5.8	5.8	5.8
	10.0	18.0	11.7	10.1	16.2	10.4	11.3	11.0	11.0	11.1
	1.0	2.7	1.2	1.1	1.7	1.1	1.2	1.1	1.2	1.1
50	5.0	9.4	5.8	5.2	8.1	5.3	5.6	5.5	5.5	5.5
	10.0	16.2	11.4	10.4	14.7	10.6	11.1	10.9	10.9	10.9

Table 3.4 Sizes of tests for the gamma model ($p = 10, q = 2$)

n	Nominal levels (%)	LR			S				
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	1.0	11.9	3.2	1.0	4.1	1.2	1.5	1.4	1.4	1.6
20	5.0	24.6	11.6	4.8	16.2	6.3	8.0	7.3	7.4	7.4
	10.0	34.0	18.8	9.8	26.4	12.3	15.5	14.2	14.3	14.6
	1.0	5.6	1.7	1.0	2.1	1.1	1.1	1.1	1.1	1.0
30	5.0	14.7	7.5	5.1	9.8	5.6	6.1	5.9	6.0	5.7
	10.0	23.1	13.2	9.7	18.4	10.5	11.6	11.1	11.2	11.0
	1.0	3.6	1.4	1.0	1.5	1.0	1.0	1.0	1.0	1.0
40	5.0	11.7	6.2	4.8	8.1	5.4	5.6	5.5	5.5	5.4
	10.0	18.5	12.1	9.7	15.5	10.3	11.0	10.7	10.7	10.6
	1.0	2.8	1.2	1.0	1.3	1.0	1.0	1.0	1.0	1.0
50	5.0	9.9	5.6	5.0	7.3	6.5	5.3	5.2	5.2	5.0
	10.0	17.0	11.5	10.1	14.5	10.5	10.9	10.7	10.8	10.7

test becomes clear when one compares their powers. Ten thousand samples of sizes $n = 20, 40, \ldots, 100$ were generated from the multiplicative heteroskedastic normal linear model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n,$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ and $\sigma_i^2 = \exp\{\alpha_0 + \alpha_1 x_i\}$. The values of the covariate x were taken as $n\, t_3$ random draws. The null hypothesis under test is $H_0 : \alpha_1 = 0$. Table 3.5 reports the estimated sizes of the tests. The results indicate that the corrected score

Table 3.5 Sizes of score tests for the heteroskedastic linear model

n	Nominal levels (%)	S	S^*	$\tilde{S}$	$K(S)$
20	10	7.2	9.0	9.1	9.0
	5	3.2	4.2	4.3	4.2
40	10	8.3	10.1	10.2	10.1
	5	4.0	4.8	4.8	4.8
60	10	7.5	10.6	11.1	10.9
	5	3.7	5.6	5.4	5.3
80	10	7.5	10.1	10.4	10.2
	5	3.7	5.1	5.1	5.0
100	10	8.3	10.2	10.4	10.3
	5	4.2	5.0	5.1	5.0

Table 3.6 Powers of score tests for the heteroskedastic linear model

n	Nominal levels (%)	S	S^*	$\tilde{S}$	$K(S)$
20	10	34.4	38.5	38.8	38.7
	5	23.2	26.6	26.8	26.7
40	10	73.3	76.2	76.6	76.5
	5	62.2	65.1	65.5	65.4
60	10	97.2	45.6	98.3	98.3
	5	94.2	40.0	96.0	95.9
80	10	99.4	39.5	99.6	99.6
	5	98.7	35.9	99.1	99.1
100	10	99.9	34.6	99.9	99.9
	5	99.7	32.0	99.8	99.8

tests perform much better than the uncorrected score test in terms of size especially if n is not very large. The figures in Table 3.6 reveal that the usual corrected score test performs poorly in terms of power in large samples. In fact, the behavior of the statistic $S^\star$ in terms of power is very different of the monotonic statistics $\tilde{S}$ and $K(S)$. The power of the corrected statistic $S^\star$ does not increase when n increases. For instance, when $n = 100$ and at a 10 % nominal size, the power of the statistic $S^\star$ is 34.6 % for $\alpha_1 = 0.6$, while the powers of all the other tests are nearly 100 %.

3.9 Heteroskedastic Regression

Most econometric applications involve regression models where the mean of a dependent variable is related to a linear or nonlinear predictor which is defined by unknown parameters and independent variables. There are a number of Bartlett and Bartlett-type corrections that can applied to heteroskedastic regression models, and this section looks at some of them. It also sheds some light on the effect of covariate

values and nuisance parameters on the convergence to the limiting null distribution of some test statistics using Monte Carlo simulation. We consider the linear regression model $Y = X\beta + \epsilon$, where Y, the dependent variable, and ϵ, the random disturbance, are n-vectors, X is an $n \times p$ matrix of values of covariates and β is a p-vector of unknown parameters. For each i, $i = 1, 2, \ldots, n$, we assume $\epsilon_i \sim \text{NID}(0, \sigma_i^2)$, where $\sigma_i^2 = h(w_i^\top \alpha)$, $w_i^\top = (1\, v_i^\top)$ is a $1 \times (q+1)$ vector of exogenous variables, α is a $(q+1)$-vector of parameters and $h(\cdot)$, the skedastic function, is any positive-valued function independent of i. It is common practice to use Breusch and Pagan (1979) score statistic to test the null hypothesis of homoskedasticity $H_0 : \alpha_1 = \cdots = \alpha_q = 0$ against the alternative of heteroskedasticity of unknown form. A well-known problem associated with this test is its tendency to under-reject the null hypothesis when heteroskedasticity is not present. The score statistic for this test becomes $S = \tilde{u}^\top W(W^\top W)^{-1} W^\top \tilde{u}$, where $W = (w_1, \ldots, w_n)$ is a $(q+1) \times n$ matrix, $\tilde{u}$ is an n-vector with typical element $\tilde{\epsilon_i}^2 - \tilde{\sigma}^2$, $\tilde{\epsilon_i}$ are the OLS residuals and $\tilde{\sigma}^2 = n^{-1} \sum_{i=1}^n \tilde{\epsilon_i}^2$. Closed-form expressions for the $A's$ for this test can be found in Honda (1988) and Cribari-Neto and Ferrari (1995c). In particular, in the latter paper, it is shown that

$$A_1 = \frac{24q(p-1)}{n} - 24\, tr(H_d J_d) + 6\mathbf{1}^\top J_d H\, J_d \mathbf{1} + 12\mathbf{1}^\top (H \odot J \odot J)\mathbf{1},$$
$$A_2 = -\frac{24q(q+2)}{n} + 36\, tr(H_d \odot H_d) - 24\mathbf{1}^\top H_d H J_d \mathbf{1},$$
$$A_3 = 24\mathbf{1}^\top H_d H\, H_d \mathbf{1} + 16n\mathbf{1}^\top H \odot H \odot H\mathbf{1},$$

where

$$J = X(X^\top X)^{-1} X^\top,\ H = V(V^\top V)^{-1} V^\top,\ V = (v_1 - \overline{v}, \ldots, v_n - \overline{v})^\top,$$

$J_d = \text{diag}\{j_{11}, \ldots j_{nn}\}$, $H_d = \text{diag}\{h_{11}, \ldots, h_{nn}\}$, $\mathbf{1}$ is an n-vector of ones, and '$\odot$' as before denotes the Hadamard product. These formulae can be used to obtain numerical values for A_1, A_2, and A_3 in empirical applications or closed-form expressions for special models. These expressions for the A's provide indication of which features of the model affect the finite-sample behavior of the score test (to order n^{-1}). Cribari-Neto and Ferrari (1995c) provided a program written in the S- PLUS language to compute the A's above. Bartlett corrections for LR tests for heteroskedasticity can be found in Attfield (1991) and Cordeiro (1993).

Consider a simple linear regression model given by $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, for $i = 1, 2, \ldots, n$, with $\epsilon \sim \text{NID}(0, \sigma_i^2)$, where $\sigma_i^2 = h(\alpha_0 + \alpha_1 x_1)$, which is a special case of the heteroskedastic model introduced above. For this simple regression model, Cribari-Neto and Ferrari (1995c) derived the following expressions for the A's: $A_1 = 6(8 + 4\gamma_{2x} - 3\gamma_{1x}^2)/n$, $A_2 = 12(3 + 3\gamma_{2x} - 2\gamma_{1x}^2)/n$, and $A_3 = 40\gamma_{1x}^2/n$, where γ_{1x} and γ_{2x} are the sample measures of skewness and excess kurtosis of the independent variable. Then, the improved score statistic S^* can be easily obtained from (3.3) using these A's and $q = 1$. In fact, the sample skewness and the sample

excess kurtosis of the independent variable affect the first-order approximation of the test. So, the covariate values can play an important role in the quality of the asymptotic χ^2 approximation that is used to perform the Breusch–Pagan test. Cribari-Neto and Ferrari (1995c) demonstrated that, in some cases, the covariate values can affect the size performance of asymptotic tests considerably.

Another important factor that can affect the first-order approximation of asymptotic econometric criteria is the number of nuisance parameters. To illustrate this point, we consider a normal linear regression model and test restrictions on the components of β, the vector of regression parameters. The A's obtained for the test are $A_1 = 12q(p-q)/n$, $A_2 = -6q(q+2)/n$, and $A_3 = 0$ (Cribari-Neto and Ferrari 1995b). It is then clear that the number of nuisance parameters $p-q$, where q is the number of restrictions imposed by H_0, has an impact on A_1 and thus on the finite-sample performance of the score test. The simulation results reported by Cribari-Neto and Ferrari (1995b) indicate that the corrected statistics are not sensitive to the values of the nuisance parameters (as expected) and that the score test is slightly undersized when $p-q=0$ and becomes oversized as $p-q$ increases, being extremely oversized when $p-q$ becomes large. The Bartlett-corrected test holds its size close to the nominal levels remarkably well.

References

Attfield, C. L. F. (1991). A Bartlett adjustment to the likelihood ratio test for homoskedasticity in the linear model. *Economics Letters*, *37*, 119–123.

Barndorff-Nielsen, O. E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, *70*, 343–365.

Barndorff -Nielsen, O. E. (1990). Approximate interval probabilities. *Journal of the Royal Statistical Society B*, *52*, 485–496.

Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroskedasticity and random coefficient variation. *Econometrica*, *47*, 1287–1294.

Chandra, T. K. (1985). Asymptotic expansions of perturbed chi-square variables. *Sankhya A*, *47*, 100–110.

Chandra, T. K., & Mukerjee, R. (1991). Bartlett-type modification for Rao's efficient score statistic. *Journal of Multivariate Analysis*, *36*, 103–112.

Chesher, A., & Spady, R. (1991). Asymptotic expansions of the information matrix test statistic. *Econometrica*, *59*, 787–815.

Cordeiro, G. M. (1983). Improved likelihood ratio statistics for generalized linear models. *Journal of the Royal Statistical Society B*, *45*, 404–413.

Cordeiro, G. M. (1987). On the corrections to the likelihood ratio statistics. *Biometrika*, *74*, 265–274.

Cordeiro, G. M. (1993). Bartlett corrections and bias correction for two heteroscedastic regression models. *Communications in Statistics, Theory and Methods*, *22*, 169–188.

Cordeiro, G. M., Cribari-Neto, F., Aubin, E. C. Q., & Ferrari, S. L. P. (1995). Bartlett corrections for one-parameter exponential family models. *Journal of Statistical Computation and Simulation*, *53*, 211–231.

Cordeiro, G. M., & Ferrari, S. L. P. (1991). A modified score test statistic having chi-squared distribution to order n^{-1}. *Biometrika*, *78*, 573–582.

Cordeiro, G. M., & Ferrari, S. L. P. (1996). Bartlett-type corrections for some score tests in proper dispersion models. *Communications in Statistics, Theory and Methods*, *25*, 29–48.

Cordeiro, G. M., Ferrari, S. L. P., & Cysneiros, A. H. M. A. (1998). A formula to improve score test statistics. *Journal of Statistical Computation and Simulation*, *62*, 123–136.

Cordeiro, G. M., Ferrari, S. L. P., & Paula, G. A. (1993). Improved score tests for generalized linear models. *Journal of the Royal Statistical Society B*, *55*, 661–674.

Cordeiro, G. M., & Stosic, B. (2008). Correcting four test statistics for one-parameter distributions using mathematica. *Communications in Statistics, Simulation and Computation*, *37*, 1663–1681.

Cox, D. R. (1988). Some aspects of conditional and asymptotic inference: A review. *Sankhya A*, *50*, 314–337.

Cox, D. R., & Reid, N. (1987). Approximations to noncentral distributions. *Canadian Journal of Statistics*, *15*, 105–114.

Cribari-Neto, F. (1997). On the corrections to information matrix tests. *Econometric Reviews*, *16*, 39–53.

Cribari-Neto, F., & Ferrari, S. L. P. (1995a). Bartlett-corrected tests for heteroskedastic linear models. *Economics Letters*, *48*, 113–118.

Cribari-Neto, F., & Ferrari, S. L. P. (1995b). Second order asymptotics for score tests in generalized linear models. *Biometrika*, *82*, 426–432.

Cribari-Neto, F., & Ferrari, S. L. P. (1995c). An improved Lagrange multiplier test for heteroskedasticity. *Communications in Statistics, Simulation and Computation*, *24*, 31–44.

Cribari-Neto, F., & Zarkos, S. (1995). Improved test statistics for multivariate regression. *Economics Letters*, *49*, 113–120.

Durbin, J. (1970). Testing for serial correlation in least squares regression when some of the regressors are lagged dependent variables. *Econometrica*, *38*, 410–421.

Ferrari, S. L. P., & Arellano-Valle, R. B. (1993). Bartlett-corrected tests for regression models with Student-*t* independent errors. Working Paper 9310, Department of Statistics, University of Sao Paulo.

Ferrari, S. L. P., Botter, D. A., Cordeiro, G. M., & Cribari-Neto, F. (1996). Second and third order bias reduction in one-parameter family models. *Statistics and Probability Letters*, *30*, 339–345.

Ferrari, S. L. P., & Cribari-Neto, F. (1993). On the corrections to the Wald test of non-linear restrictions. *Economics Letters*, *42*, 321–326.

Ferrari, S. L. P., Uribe-Opazo, M. A., & Cribari-Neto, F. (1997). Second order asymptotics for score tests in exponential family nonlinear models. *Journal of Statistical Computation and Simulation*, *59*, 179–194.

Gregory, A. W., & Veall, M. R. (1985). Formulating Wald tests of nonlinear restrictions. *Econometrica*, *53*, 1465–1468.

Harris, P. (1985). An asymptotic expansion for the null distribution of the efficient score statistic. *Biometrika*, *72*, 653–659 (Erratum in vol. 74, p. 667).

Hill, G. W., & Davis, A. W. (1968). Generalized asymptotic expansions of Cornish-Fisher type. *Annals of Mathematical Statistics*, *39*, 1264–1273.

Honda, Y. (1988). A size correction to the Lagrange multiplier test for heteroskedasticity. *Journal of Econometrics*, *38*, 375–386.

Inder, B. A. (1984). Finite-sample power of tests for autocorrelation in models containing lagged dependent variables. *Economics Letters*, *14*, 179–185 (Erratum in vol. 16, pp. 401–402).

Inder, B. A. (1986). An approximation to the null distribution of the Durbin-Watson statistic in models containing lagged dependent variables. *Econometric Theory*, *2*, 413–428.

Kakizawa, Y. (1996). Higher order Monotone Bartlett-Type adjustment for some multivariate test statistics. *Biometrika*, *83*, 923–927.

Lafontaine, F., & White, K. J. (1986). Obtaining any Wald statistic you want. *Economics Letters*, *21*, 35–40.

Lawley, D. N. (1956). A general method for approximating to the distribution of likelihood ratio criteria. *Biometrika*, *71*, 233–244.

Lugannani, R., & Rice, S. (1980). Saddlepoint approximation for the distribution of the sum of independent random variables. *Advances Applied Probabability*, *12*, 475–490.

Mukerjee, R. (1992). Parametric orthogonality and a Bartlett-type modification for Rao's statistic in the presence of a nuisance parameter. *Statistics and Probability Letters*, *13*, 397–400.
Phillips, P. C. B., & Park, J. Y. (1988). On the formulation of Wald tests of nonlinear restrictions. *Econometrica*, *56*, 1065–1083.
Rao, C. R., & Mukerjee, R. (1995). Comparison of Bartlett-type adjustments for the efficient score statistic. *Journal of Statistical Planning and Inference*, *46*, 137–146.
Santos, S. J. P., & Cordeiro, G. M. (1999). Corrected Wald test statistics for one-parameter exponential family models. *Communications in Statistics, Theory and Methods*, *28*, 1391–1414.
Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley.
Taniguchi, M. (1991). Third-order asymptotic properties of a class of test statistics under a local alternative. *Journal of Multivariate Analysis*, *37*, 223–238.

Chapter 4
Analytical and Bootstrap Bias Corrections

Abstract Maximum likelihood estimators are usually biased: In finite samples, their expected value differs from the true parameter value. This is a systematic error. It typically vanishes as the sample size increases, but it can be large in small samples. Different strategies can be employed to reduce such systematic error. In this chapter, we present two analytical bias corrections. We also show how the bootstrap can be used to bias-correct estimators. Bias corrections in different statistical models are presented and discussed. In particular, we address the issue of bias-correcting covariance matrix estimators in heteroskedastic linear regressions.

Keywords Bias · Bias correction · Bootstrap · Heteroskedasticity · Likelihood · Maximum likelihood · Regression

4.1 Introduction

A main object in asymptotic likelihood theory is to calculate the second-order biases of the MLEs. These estimators typically have biases of order $\mathcal{O}(n^{-1})$, where n is the sample size, which are commonly ignored in practice, the justification being that they are small when compared to the standard errors of the parameter estimators that are of order $\mathcal{O}(n^{-1/2})$. For small samples sizes, however, these biases can be appreciable and of the same magnitude as the corresponding standard errors. In such cases, the biases should not be neglected. Bias reduction approaches based on the $\mathcal{O}(n^{-1})$ bias function can be quite effective. The usual normal approximation can be oftentimes improved by making a simple bias adjustment to the MLE.

Approximations to the bias of the MLE in simple models may be obtained analytically. Bias correction typically does a very good job in reducing the bias. However, it may either increase the mean-squared error. Whether bias correction is useful in practice depends basically on the shape of the bias function and on the variance of the MLE.

G. M. Cordeiro and F. Cribari-Neto, *An Introduction to Bartlett Correction and Bias Reduction*, SpringerBriefs in Statistics, DOI: 10.1007/978-3-642-55255-7_4,

In order to improve the accuracy of MLEs using analytical bias reduction, one needs to obtain several cumulants of log-likelihood derivatives, which are notoriously cumbersome. Bias correction has been extensively studied in the statistical literature, and there has been considerable interest in finding simple matrix expressions for second-order biases of MLEs in a number of classes of regression models, which do not involve cumulants of log-likelihood derivatives. This approach has been applied to several regression models. We cite the following models: normal nonlinear models (Cook et al. 1986), generalized log-gamma regression model (Young and Bakir 1987), generalized linear models (Cordeiro and McCullagh 1991), ARMA models (Cordeiro and Klein 1994), multivariate nonlinear regression models (Cordeiro and Vasconcellos 1997), generalized linear models with dispersion covariates (Botter and Cordeiro 1998), Poisson regression (Giles and Feng 2011), symmetric nonlinear regression models (Cordeiro et al. 2000), Student t regression model with unknown degrees of freedom (Vasconcellos and Silva 2005), beta regression models (Ospina et al. 2006), and a class of multivariate normal model where the mean vector and the covariance matrix have parameters in common (Patriota and Lemonte 2009). It is noteworthy that the leading term in the asymptotic bias can be computed using a weighted linear regression. Stosic and Cordeiro (2009) showed how to symbolically compute the biases of the MLEs in general two-parameter continuous distributions, thus bypassing the traditional computation of joint cumulants of log-likelihood derivatives. Bias corrections for the MLEs of the parameters that index several distributions have also been derived; for instance, Cordeiro et al. (1997) and Cribari-Neto and Vasconcellos (2002) addressed bias correction for the MLEs in the beta law, Giles (2012) obtained such a correction for the half-logistic distribution, and Giles et al. (2013) derived the correction for the Lomax distribution.

One can easily obtain a *bias-reduced* estimator by subtracting the $\mathcal{O}(n^{-1})$ bias from the MLE. Alternatively, an examination of the form of the bias may suggest a reparametrization of the model that results in less biased estimators.

4.2 A General Formula

Consider that the total log-likelihood function $\ell(\theta)$, based on n observations not necessarily i.i.d., is a function of a $p \times 1$ vector θ of unknown parameters. We assume that $\ell = \ell(\theta)$ is regular (Cox and Hinkley 1974) with respect to all θ derivatives up to and including those of third order. We consider the notation for the log-likelihood derivatives in which we reserve lower-case subscripts $r, s, t, \ldots$ to denote components of the vector θ: $U_r = \partial l/\partial\theta_r$, $U_{rs} = \partial^2 l/\partial\theta_r\partial\theta_s$, and so on. The standard notation will be adopted for the cumulants of log-likelihood derivatives: $\kappa_{rs} = \mathbb{E}(U_{rs})$, $\kappa_{r,s} = \mathbb{E}(U_r U_s)$, $\kappa_{rs,t} = \mathbb{E}(U_{rs}U_t)$, etc., where all κ's refer to a total over the sample and are, in general, of order n. The elements of Fisher's information matrix K are $\kappa_{r,s} = -\kappa_{rs}$ and let $\kappa^{r,s} = -\kappa^{rs}$ denote the corresponding elements of the inverse matrix K^{-1}, which is of order $\mathcal{O}(n^{-1})$.

The MLE $\hat{\theta}$ of θ can be obtained as a solution of a system of nonlinear equations $\hat{U}_r = 0$ for $r = 1, \ldots, p$. A general formula for the $\mathcal{O}(n^{-1})$ bias of $\hat{\theta}$ for a regular statistical model with p unknown parameters was given by Cox and Snell (1968) and Cordeiro and McCullagh (1991). Hereafter, we shall use Einstein's summation convention with indices varying over the corresponding parameters. Assuming standard regularity conditions (Cox and Hinkley 1974), we can expand $\hat{U}_r = 0$ to obtain $U_r + \sum_s U_{rs} (\theta_s - \theta_s) + \mathcal{O}_p(1) = 0$ and then use matrix notation to write $U = J (\hat{\theta} - \theta) + \mathcal{O}_p(1)$, where U is the score vector and J is the observed information matrix. Since $J = K + \mathcal{O}_p(n^{1/2})$, it follows that $U = K (\hat{\theta} - \theta) + \mathcal{O}_p(1)$ and

$$\hat{\theta} - \theta = K^{-1} U + \mathcal{O}_p(n^{-1}). \tag{4.1}$$

Equation (4.1) is important because it can be used when computing higher-order moments and cumulants of the estimator $\hat{\theta}$. By expanding $\hat{U}_r$ up to terms of second order, we have

$$U_r + \sum_s U_{rs} (\hat{\theta}_s - \theta_s) + \frac{1}{2} \sum_{s,t} U_{rst} (\hat{\theta}_s - \theta_s) (\hat{\theta}_t - \theta_t) + o_p(1) = 0.$$

Taking expected values, we can write

$$\sum_s \kappa_{rs} \, \mathbb{E}(\hat{\theta}_s - \theta_s) + \sum_s \mathrm{Cov}(U_{rs}, \hat{\theta}_s - \theta_s) + \frac{1}{2} \sum_{s,t} \kappa_{rst} (-\kappa^{st}) + o(1) = 0. \tag{4.2}$$

Using (4.1) we obtain, up to terms of order $\mathcal{O}(n^{-1})$,

$$\mathrm{Cov}(U_{rs}, \hat{\theta}_s - \theta_s) = \mathrm{Cov}\left(U_{rs}, -\sum_t \kappa^{st} U_t \right) = -\sum_t \kappa_{rs,t} \, \kappa^{st}. \tag{4.3}$$

Let $B(\hat{\theta}_a)$ be the $\mathcal{O}(n^{-1})$ bias of the estimator $\hat{\theta}_a$ for $a = 1, \ldots, p$. Plugging (4.3) into (4.2), we obtain

$$\sum_s \kappa_{rs} \, B(\hat{\theta}_s) - \sum_{s,t} \kappa^{st} \left(\kappa_{rs,t} + \frac{1}{2} \kappa_{rst} \right) + o(1) = 0$$

which leads to

$$B(\hat{\theta}_a) = \sum_{r,s,t} \kappa^{ar} \, \kappa^{st} \left(\kappa_{rs,t} + \frac{1}{2} \kappa_{rst} \right) = \sum_{r,s,t} \kappa^{ar} \, \kappa^{st} \left(\kappa_{rs}^{(t)} - \frac{1}{2} \kappa_{rst} \right). \tag{4.4}$$

We can verify that the two alternative formulae for $B(\hat{\theta}_a)$ are equivalents using a Bartlett identity. For general regression models, we can derive matrix expressions

for the bias of the MLE $\hat{\theta}$, say $B(\hat{\theta})$, from Eq. (4.4) as long as the cumulants κ's are invariant under permutations of parameters (see Cordeiro and McCullagh 1991).

Let $\hat{B}(\hat{\theta})$ denote the estimated second-order bias, i.e., $\hat{B}(\hat{\theta})$ is obtained by replacing θ by $\hat{\theta}$ in $B(\hat{\theta})$. We can now define the bias-corrected estimator $\tilde{\theta} = \hat{\theta} - \hat{B}(\hat{\theta})$. The corrected estimator $\tilde{\theta}$ is expected to have better sampling properties than the original estimator $\hat{\theta}$. In fact, several simulation studies presented in the literature (Botter and Cordeiro 1998; Cordeiro et al. 2000; Vasconcellos and Silva 2005; Ospina et al. 2006; Patriota and Lemonte 2009) have shown that the corrected estimators $\tilde{\theta}$ have smaller biases than their corresponding uncorrected estimators, thus suggesting that these bias corrections have the effect of bringing the corrected estimates closer on average to the true parameter values. It is noteworthy, nonetheless, that bias correction can lead to variance inflation.

A simple illustration of (4.4) is provided by taking n i.i.d. observations from the normal distribution with mean μ and variance σ^2, namely $\mathcal{N}(\mu, \sigma^2)$. Suppose the interest lies in computing the n^{-1} biases of the estimators of μ and σ. The information matrix elements are as follows: $\kappa_{\mu,\mu} = n/\sigma^2$, $\kappa_{\mu,\sigma} = 0$, and $\kappa_{\sigma,\sigma} = 2n/\sigma^2$. The third-order cumulants are easily obtained as $\kappa_{\mu\mu\mu} = \kappa_{\mu,\mu\mu} = \kappa_{\sigma,\mu\mu} = \kappa_{\sigma,\mu\sigma} = \kappa_{\mu,\sigma\sigma} = \kappa_{\mu\sigma\sigma} = 0$, $\kappa_{\mu\mu\sigma} = -\kappa_{\mu,\mu\sigma} = 2n/\sigma^3$, $\kappa_{\sigma,\sigma\sigma} = -6n/\sigma^3$, and $\kappa_{\sigma\sigma\sigma} = 10n/\sigma^3$. Thus, $B(\hat{\mu}) = 0$ since $\hat{\mu} = \Sigma y_i/n$ has no bias. Further, after some algebra, $B(\hat{\sigma}) = -3\sigma/4n$. This approximate result is in agreement with the exact expected value of $\hat{\sigma} = \{\Sigma(y_i - \overline{y})^2/n\}^{1/2}$ given by $\mathbb{E}(\hat{\sigma}) = b(n)\,\sigma$, where $b(n) = (\sqrt{2/n})\,\Gamma(n/2)/\Gamma((n-1)/2)$. This exact value can be easily obtained by noting that $(n-1)\hat{\sigma}^2/\sigma^2$ is χ^2_{n-1} distributed. In fact, using Stirling expansion in $\mathbb{E}(\hat{\sigma})$ yields $\mathbb{E}(\hat{\sigma}) = \sigma\,[1 - 3/(4n) + \mathcal{O}(n^{-2})]$. The bias-corrected estimator of σ is then $\tilde{\sigma} = [1 + 3/(4n)]\hat{\sigma}$. Clearly, an unbiased estimator of σ can be obtained by dividing $\hat{\sigma}$ by $b(n)$. As n grows, $b(n)$ approaches one, but for small values of n, the correction can be important. For example, for $n = 4$, 10 and 50, we obtain $b(4) = 0.797884$, $b(10) = 0.922745$, and $b(50) = 0.984912$, respectively. If the calculation of $b(n)$ is cumbersome, one can then use $\tilde{\sigma}$.

4.3 One-Parameter Distributions

For a one-parameter model, the n^{-1} bias of $\hat{\theta}$ follows from Eq. (4.4) by setting all parameters equal to θ. We then obtain the formula first derived by Bartlett (1953):

$$B(\hat{\theta}) = \kappa^{\theta\theta^2}\left(\kappa_{\theta\theta,\theta} + \frac{1}{2}\kappa_{\theta\theta\theta}\right) = \kappa^{\theta\theta^2}\left(\kappa^{(\theta)}_{\theta\theta} - \frac{1}{2}\kappa_{\theta\theta\theta}\right). \tag{4.5}$$

Let $Y_1, \ldots, Y_n$ be a set of n i.i.d. random variables having distribution in the one-parameter exponential family defined by

$$\pi(y;\theta) = \frac{1}{\zeta(\theta)}\,\exp\{-\alpha(\theta)\,d(y) + \nu(y)\}, \tag{4.6}$$

where θ is a scalar parameter, $\zeta = \zeta(\theta)$, $\alpha = \alpha(\theta)$, $d(y)$, and $\nu(y)$ are known functions. It is assumed that the support of $\pi(y;\theta)$ does not depend upon θ. It is also assumed that α and ζ have continuous first three derivatives with respect to θ and that ζ is positive valued. Let $\beta = \beta(\theta) = \zeta'(\zeta\,\alpha')^{-1}$ and assume that α' and β' are different from zero for all values of θ in the parameter space, primes denoting derivatives with respect to θ. Many commonly used distributions are special cases of the family of distributions (4.6). Examples are the binomial, exponential, extreme value, gamma, inverse Gaussian, Laplace, log-normal with only one unknown parameter, normal, Pareto and Poisson distributions.

The score function for a single observation is given by $U_\theta = -\alpha'[\beta + d(y)]$. Since $\mathbb{E}(U_\theta) = 0$, it follows that $\mu = \mathbb{E}[d(Y)] = -\beta$. The MLE $\hat{\theta}$ comes from $n^{-1}\sum_i d(y_i) = -\beta(\hat{\theta})$. Its computation may require the use of numerical methods. The second and third log-likelihood derivatives for a single observation are $U_{\theta\theta} = -\alpha''[\beta + d(y)] - \alpha'\beta'$ and $U_{\theta\theta\theta} = -\alpha'''[\beta + d(y)] - 2\alpha''\beta' - \alpha'\beta''$, respectively. It is now easy to obtain the cumulants as $\kappa_{\theta\theta} = -n\alpha'\beta'$, $\kappa_{\theta\theta\theta} = -2n\alpha''\beta' - n\alpha'\beta''$, etc. The asymptotic variance of $\hat{\theta}$ is $\mathrm{Var}(\hat{\theta}) = (\alpha'\beta' n)^{-1}$. Using the cumulants in (4.5), we obtain

$$B(\hat{\theta}) = -\frac{\beta''}{2\alpha'\,\beta'^2\,n}. \tag{4.7}$$

It is noteworthy that (4.7) only requires knowledge of α and ζ and their first three derivatives with respect to θ. It can be easily implemented in a computer algebra system, such as MAPLE and MATHEMATICA, to obtain bias-corrected estimates with minimal effort.

It is possible to check Eq. (4.7) from first principles for special distributions. The simplest special case is the normal distribution with known mean μ and variance θ for which $B(\hat{\theta})$ vanishes. Here, $\hat{\theta} = \sum(y_i - \mu)^2/n \sim \theta\chi^2_n/n$ is clearly an unbiased estimator. It is easy to verify that the same happens for the following distributions: binomial, exponential, inverse Gaussian with unknown scale parameter θ, Laplace, Poisson and truncated extreme value. We consider two additional examples. For the inverse Gaussian distribution with known mean $\mu > 0$ and scale parameter $\theta > 0$, we have $\alpha = \theta$, $\zeta = \theta^{-1/2}$, $d(y) = (y-\mu)^2/(2\mu^2 y)$ and $\nu(y) = -[\log(2\pi\, y^3)]/2$ and then $\hat{\theta} = n\mu^2\left[\sum(Y_i - \mu)^2/y_i\right]^{-1} \sim n\theta/\chi^2_n$. A Taylor series expansion to order n^{-1} gives $\mathbb{E}(\hat{\theta}) = 2\theta/n$, which is in agreement with the result obtained from (4.7). Consider now the gamma distribution with known index $k > 0$ and scale parameter $\theta > 0$. Here, $\alpha = \theta$, $\zeta = \theta^{-k}$, $d(y) = y$, and $\nu(y) = (k-1)\log(y) - \log[\Gamma(k)]$. It is easy to show that $\hat{\theta}$ is $2\,n\,k\theta/\chi^2_{2kn}$ distributed. Thus, by direct expansion to order n^{-1}, we establish that $\mathbb{E}(\hat{\theta}) = \theta/(kn)$, which agrees with the result obtained using (4.7).

4.4 Two-Parameter Distributions

Stosic and Cordeiro (2009) presented computer codes that may be used with MAPLE and MATHEMATICA to obtain closed-form expressions for the bias corrections B_μ and B_ϕ of the MLEs of the parameters μ and ϕ, for arbitrary two-parameter continuous distributions, through a straightforward application of Eq. (4.4).

The symbolic computation software MAPLE and MATHEMATICA are quite useful for dealing with analytic expressions of formidable size and complexity, but we note that limitations still exist. It turns out that the complexity of the formulae involved in calculating the cumulants of log-likelihood derivatives for some distributions exceeds the software capacity. In some cases, neither MAPLE nor MATHEMATICA were able to produce closed-form expressions for the bias corrections. It should be noted, nonetheless, that the programs provided by the authors are still useful in such cases since future versions of the software may be able to handle them. It should be also pointed out that such a limitation does not diminish the usefulness of the scripts provided by Stosic and Cordeiro (2009), since both software have produced closed-form expressions for most two-parameter continuous density functions considered. Moreover, whenever both software yielded a closed-form expression, the results were found to be identical.

For both MAPLE and MATHEMATICA, the user must specify the form and the domain of the density function $f = f(y; \mu, \phi)$, as well as the relevant constraints on μ and ϕ (e.g., $\mu \in \mathbb{R}$ or $\mu > 0$), and the program first defines and analytically computes the cumulants (κ's). Then, the second-order cumulants are subsequently inserted into the expression for the information matrix, the inverse information matrix is computed, and the results are used together with the third-order cumulants to produce the final result using Eq. (4.4). In what follows, we denote the first and second derivatives of the digamma function $\psi(p) = \mathrm{d}\log\{\Gamma(p)\}/\mathrm{d}p$ by $\psi'(p)$ and $\psi''(p)$, respectively. Also, $\gamma = 1 - \psi(2)$ is Euler's constant and $\zeta(p) = \sum_{n=1}^{\infty} n^{-p}$ is the Riemann Zeta function. The formulae for the examples listed below were obtained using the MAPLE and MATHEMATICA scripts of Stosic and Cordeiro (2009):

1. Normal distribution with mean μ and variance ϕ^2:

$$B_\mu = 0, \quad B_\phi = -\frac{3\,\phi}{4n}.$$

2. Reciprocal normal distribution with mean μ and variance ϕ^2:

$$B_\mu = 0, \quad B_\phi = -\frac{3\,\phi}{4n}.$$

3. Gamma distribution with mean μ and shape parameter ϕ:

$$B_\mu = 0, \quad B_\phi = -\frac{2 - \phi\psi'(\phi) + \phi^2\psi''(\phi)}{2\,[\phi\psi'(\phi) - 1]^2\,n}.$$

We note that the expression for B_ϕ given above is a special case of Eq. (5.1) in Cordeiro and McCullagh (1991) and corrects their Eq. (5.2).

4. Inverse gamma distribution with scale μ and shape ϕ:

$$B_\mu = \frac{\mu\,[2\,\phi\,\psi'(\phi)^2 - 3\,\psi'(\phi) - \phi\,\psi''(\phi)]}{2[\phi\,\psi'(\phi) - 1]^2 n},$$
$$B_\phi = \frac{\phi\,\psi'(\phi) - \phi^2\psi''(\phi) - 2}{2[\phi\,\psi'(\phi) - 1]^2 n}.$$

5. Weibull distribution with scale μ and shape ϕ [here, $\mathbb{E}(Y) = \mu\Gamma(1+\phi^{-1})$]:

$$B_\mu = \frac{\mu}{2\,\pi^4\,\phi^2\,n}\left\{\pi^4\,(1-2\,\phi) + 6\,\pi^2\left[1+\gamma^2+5\,\phi-2\,\gamma\,(1+2\,\phi)\right]\right.$$
$$\left. +\ 72\,(\gamma-1)\,\phi\,\zeta(3)\right\},$$
$$B_\phi = \frac{18\,\phi\,\left(\pi^2 - 2\,\zeta(3)\right)}{\pi^4\,n}.$$

6. Logistic distribution with mean μ and variance $\pi^2\,\phi^2/6$:

$$B_\mu = 0, \quad B_\phi = -\frac{9\phi\,(4\,\pi^2+3)}{(\pi^2+3)^2\,4n}.$$

7. Extreme value distribution with mean $\mu + \gamma\,\phi$ and variance $\pi^2\,\phi^2/6$:

$$B_\mu = \frac{\phi\,[3\,(-5+4\,\gamma)\,\pi^2 + \pi^4 - 36\,(-1+\gamma)\,\zeta(3)]}{4\pi^4\,n},$$
$$B_\phi = \frac{-12\,\phi\,(\pi^2 - 3\,\zeta(3))}{4\pi^4\,n}.$$

8. Random walk distribution:

$$B_\mu = 0, \quad B_\phi = \frac{3\,\phi}{n}.$$

9. Student's t-distribution with location parameter μ and dispersion parameter ϕ:

$$B_\mu = 0, \quad B_\phi = \frac{-3\,(-3+2\,\nu+\nu^2)\,\phi}{4\,\nu\,(5+\nu)\,n}.$$

Here, ν denotes the number of degrees of freedom. For $\nu = 1$ (Cauchy distribution), we obtain $B_\phi = 0$. When $\nu \to \infty$, we obtain $B_\phi = -3\phi/4$, which is the bias function for the normal distribution.

10. Fisher–Tippett distribution with mode μ and variance $\pi^2\,\phi^2/6$:

$$B_\mu = \frac{\phi\,[3\,(-5+4\,\gamma)\,\pi^2+\pi^4-36\,(-1+\gamma)\,\zeta(3)]}{\pi^4\,n},$$
$$B_\phi = \frac{-12\,\phi\,[\pi^2-3\,\zeta(3)]}{\pi^4\,n}.$$

Bias correction is easily carried out and tends to work quite well whenever the bias function is approximately flat. If the bias function is approximately linear, the $\mathcal{O}(n^{-1})$ bias can still be easily computed, but the resulting bias correction may not be as effective. In particular, if the bias function slopes downward, the bias-corrected estimators will display larger variances than the uncorrected estimators, and they may thus display larger mean-squared errors. If the bias function slopes upward, the bias-corrected estimators will have smaller variances than the uncorrected ones.

4.5 Generalized Linear Models

For the two-parameter linear exponential family distributions defined in (2.12), with canonical parameters ϕ and $\phi\,\theta$, the decomposition $c(y,\phi) = d_1(\phi) + d_2(y)$ holds. As discussed in Sect. 2.5, a GLM is defined by the family of distributions in (2.12) and by the systematic component $g(\mu) = \eta = X\,\beta$, where $g(\cdot)$ is a known one-to-one continuously twice-differentiable function, X is a specified $n \times p$ model matrix of full rank p ($p < n$) and $\beta = (\beta_1, \ldots, \beta_p)^\top$ is a set of unknown linear parameters to be estimated.

Denote the n observations by $y_1, \ldots, y_n$ and the total log-likelihood for β and ϕ by $\ell = \ell(\beta, \phi)$. The parameters β and ϕ are orthogonal since $\mathbb{E}(\partial^2\ell/\partial\beta\,\partial\phi) = 0$. Let $\hat\beta$ and $\hat\phi$ be the MLEs of β and ϕ, respectively. The joint cumulants are

$$\kappa_{rs} = \mathbb{E}(\partial^2\ell/\partial\beta_r\partial\beta_s),\ \kappa_{r\phi} = \mathbb{E}(\partial^2\ell/\partial\beta_r\,\partial\phi),\ \kappa_{rst} = \mathbb{E}(\partial^3\ell/\partial\beta_r\partial\beta_s\partial\beta_t),$$
$$\kappa_{r,st} = \mathbb{E}(\partial\ell/\partial\beta_r\,\partial^2\ell/\partial\beta_s\partial\beta_t),\ \kappa_{rs,\phi} = \mathbb{E}(\partial^2\ell/\partial\beta_r\,\partial\beta_s\,\partial\ell/\partial\phi),$$

$\kappa_{rs}^{(t)} = \partial\kappa_{rs}/\partial\beta_t$, etc., with the indices being replaced by ϕ when derivatives are taken with respect to this parameter. All κ's refer to a total over the sample and are, in general, of order n. The joint information matrix for $(\beta^\top, \phi)^\top$ is $K = \mathrm{diag}\{\phi\,(X^\top W X), -n d_1''(\phi)\}$, where $K_\beta = \{-\kappa_{rs}\} = \phi\,(X^\top W X)$ is the information for β and $\{-n d_1''(\phi)\}$ is the information for ϕ. Here, $K^{-1} = \mathrm{diag}\{\{-\kappa^{rs}\}, -\kappa^{\phi\phi}\} = \mathrm{diag}\{\phi^{-1}\,(X^\top W X)^{-1}, -[n d_1''(\phi)]^{-1}\}$ is the inverse of the joint information matrix.

Let $\hat\phi$, $\hat\beta$, $\hat\eta = X\,\hat\beta$ and $\hat\mu = g^{-1}(\hat\eta)$ be the MLEs of ϕ, β, η, and μ, respectively. Estimation of β and ϕ was discussed in Sect. 2.5.

The MLEs of β and ϕ are asymptotically independent due to their asymptotic normality and the block diagonal structure of the joint information matrix K. For a GLM, we can easily obtain the cumulants κ's. The key to obtain a simple expression for the bias $B(\hat\beta)$ in GLMs is the invariance of the κ's under permutation of the β

parameters and the orthogonality between ϕ and β. Let $S = \{1, \ldots, p\}$ be the set that indexes the β parameters. Since $\kappa_{r\phi} = 0$ and $\kappa_{r\phi\phi} = \kappa_{r\phi}^{(\phi)} = \kappa_{r\phi,\phi} = 0$ for $r \in S$, we only need to take into account one summation term involving the various combinations of β parameters in Eq. (4.4). It can be shown that the crucial quantity for the n^{-1} bias of $\hat{\beta}$ is equal to

$$\kappa_{rs}^{(t)} - \frac{1}{2}\kappa_{rst} = -\frac{\phi}{2}\sum_{i=1}^{n} f_i \, x_{ir} \, x_{is} x_{it},$$

where $f = V^{-1} \, d\mu/d\eta \, d^2\mu/d\eta^2$ is a typical element of the diagonal matrix $F = \text{diag}\{f_1, \ldots, f_n\}$. Rearranging the summation terms in (4.4), we obtain

$$B(\hat{\beta}_a) = -\frac{\phi}{2}\sum_{i} f_i \left(\sum_{r} \kappa^{ar} \, x_{ir}\right)\left(\sum_{s,t} \kappa^{st} \, x_{is} x_{it}\right), \tag{4.8}$$

where r, s, and t vary in S, and i runs over the observations. We define the matrix $Z = \{z_{ij}\} = X \, (X^\top W X)^{-1} \, X^\top$ which is, apart from the multiplier ϕ^{-1}, the asymptotic covariance matrix of the estimators $\hat{\eta}_1, \ldots, \hat{\eta}_n$ of the model linear predictors. Additionally, $Z_d = \text{diag}\{z_{11}, \ldots, z_{nn}\}$ is a diagonal matrix with the diagonal elements of Z, and $\mathbf{1_n}$ is an $n \times 1$ vector of ones. It is now possible to write (4.8) in simple matrix form as (Cordeiro and McCullagh 1991)

$$B(\hat{\beta}) = -\frac{1}{2\phi}(X^\top \, W \, X)^{-1} \, X^\top \, Z_d \, F \, \mathbf{1_n}. \tag{4.9}$$

We define the $n \times 1$ vector $\xi = -(2\phi)^{-1} \, W^{-1} \, Z_d \, F \, \mathbf{1_n}$, whose components are $\xi_i = -(2\phi)^{-1} \, \mu_i'' \, \mu_i'^{-1} \, z_{ii}$, where $\mu_i' = d\mu_i/d\eta_i$ and $\mu_i'' = d^2\mu_i/d\eta_i^2$ are the derivatives of the inverse link function and z_{ii} is the asymptotic variance of $\hat{\eta}_i$ except for the multiplier ϕ^{-1}. The components of ξ are 0 for the identity link, $-z_{ii}/2$ for the logarithm link, $z_{ii}(\mu_i - 1/2)$ for the logit link, and $(z_{ii}\eta_i)/2$ for the probit link. Then, Eq. (4.9) reduces to

$$B(\hat{\beta}) = (X^\top \, W \, X)^{-1} \, X^\top \, W \, \xi. \tag{4.10}$$

Equation (4.10) is easily obtained as the vector of regression coefficients in the formal linear regression of $\hat{\xi}$ on X using $\hat{W}$ as a weight matrix. We retain the weights and the model formula from the GLM, but the link function becomes the identity and the response vector becomes $\hat{\xi}$. In order to evaluate $B(\hat{\beta})$, one only needs the variance and link functions and the first two derivatives. We can now replace unknown parameters by their MLEs on the right-hand side of Eq. (4.10) to obtain the bias-corrected estimator $\tilde{\beta} = \hat{\beta} - \hat{B}(\hat{\beta})$, where $\hat{B}(\hat{\beta})$ is the value of $B(\hat{\beta})$ at the vector $(\hat{\beta}^\top, \hat{\phi})^\top$.

We now provide the $\mathcal{O}(n^{-1})$ bias of the MLE of the parameter ϕ. Using Eq. (4.4) and the orthogonality between ϕ and β, we can write

$$B(\hat{\theta}_a) = \kappa^{\phi\phi} \sum_{r,s} \kappa^{rs} \left(\kappa_{\phi r,s} + \frac{1}{2}\kappa_{\phi rs} \right) + \kappa^{\phi\phi^2} \left(\kappa_{\phi\phi,\phi} + \frac{1}{2}\kappa_{\phi\phi\phi} \right).$$

The cumulants required are $\kappa_{\phi rs} = -\kappa_{\phi r,s} = \sum_{i=1}^{n} w_i\, x_{ir} x_{is}$, $\kappa_{\phi\phi,\phi} = 0$ and $\kappa_{\phi\phi\phi} = d_1'''(\phi)$. We have $\sum_{i=1}^{n} w_i\, z_{ii} = \text{tr}(WZ) = \text{rank}(X) = p$, and then, the $\mathcal{O}(n^{-1})$ bias of $\hat{\phi}$ can be expressed as

$$B(\hat{\phi}) = \frac{\phi\, d_1'''(\phi) - p\, d_1''(\phi)}{2\phi\, d_1''(\phi)^2 n}. \tag{4.11}$$

Equation (4.11) depends on the model matrix only through its rank. The corrected estimator of the precision parameter is then $\tilde{\phi} = \hat{\phi} - \hat{B}(\hat{\phi})$. For the normal model with variance given by the reciprocal of ϕ, (4.11) reduces to $B(\hat{\phi}) = (p+2)\phi/n$. For the gamma model with index ϕ,

$$B(\hat{\phi}) = \frac{p\,[\phi\,\psi'(\phi) - 1] - [1 + \phi\,\psi''(\phi)]^2}{2\,[\phi\,\psi'(\phi) - 1]\,n},$$

where $\psi'(\phi)$ and $\psi''(\phi)$ are the digamma and trigamma functions, respectively.

The second-order bias of the MLE of the mean vector μ can also be obtained. Since μ_i is a one-to-one function of η_i, we can expand $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$ in Taylor series to order n^{-1} as

$$B(\hat{\mu}_i) = B(\hat{\eta}_i)\, \frac{\text{d}\mu_i}{\text{d}\eta_i} + \text{Var}(\hat{\eta}_i)\, \frac{\text{d}^2\mu_i}{\text{d}\eta_i^2},$$

where $\text{Var}(\hat{\eta}_i)$ is the $\mathcal{O}(n^{-1})$ term in the variance of $\hat{\eta}_i$. Let $G_1 = \text{diag}\{\text{d}\mu_i/\text{d}\eta_i\}$ and $G_2 = \text{diag}\{d^2\mu_i/d\eta_i^2\}$. It then follows that

$$B(\hat{\mu}) = (2\phi)^{-1}(G_2 - G_1\, Z\, F)\, Z_d \mathbf{1_n},$$

and the corrected mean estimators are defined by $\tilde{\mu} = \hat{\mu} - \hat{B}(\hat{\mu})$.

4.6 The Birnbaum–Saunders Model

The material contained in this section and in the next three sections is based on Lemonte and Cordeiro (2010). The two-parameter BS distribution, in short $\mathcal{BS}(\alpha, \eta)$, was defined in Sect. 2.6.

Lemonte et al. (2007) derived the second-order biases of the MLEs of α and η, and obtained a corrected LR statistic for testing the parameter α. Lemonte et al. (2008) proposed several bootstrap bias-corrected estimates of α and η.

Rieck and Nedelman (1991) proposed a log-linear regression model based on the BS distribution. They showed that if $T \sim \mathcal{BS}(\alpha, \eta)$, then $Y = \log(T)$ is sinh-normal distributed, say $Y \sim \mathcal{SN}(\alpha, \mu, \sigma)$, with shape, location, and scale parameters given by α, $\mu = \log(\eta)$ and $\sigma = 2$, respectively. Their model has been widely used as an alternative model to the gamma, log-normal, and Weibull regression models, see Rieck and Nedelman(1991, § 7). The density function of Y can be expressed as

$$\pi(y; \alpha, \mu, \sigma) = \frac{2}{\alpha\sigma\sqrt{2\pi}} \cosh\left(\frac{y-\mu}{\sigma}\right) \exp\left\{-\frac{2}{\sigma^2}\sinh^2\left(\frac{y-\mu}{\sigma}\right)\right\}, \quad y \in \mathbb{R}. \tag{4.12}$$

The distribution with density given in (4.12) has a number of interesting properties (Rieck 1989): (1) it is symmetric around the location parameter μ; (2) it is unimodal for $\alpha \leq 2$ and bimodal for $\alpha > 2$; (3) the mean and variance of Y are $\mathbb{E}(Y) = \mu$ and $\mathrm{Var}(Y) = \sigma^2 w(\alpha)$, respectively. There is no closed-form expression for $w(\alpha)$, but Rieck (1989) obtained asymptotic approximations for both small and large values of α; and (4) if $Y_\alpha \sim \mathcal{SN}(\alpha, \mu, \sigma)$, then $S_\alpha = 2(Y_\alpha - \mu)/(\alpha\sigma)$ converges in distribution to the standard normal distribution when $\alpha \to 0$.

As explained in Sect. 2.6 [see Eq. (2.17)], Lemonte and Cordeiro (2009) proposed the nonlinear regression model

$$Y_i = f_i(x_i; \beta) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{4.13}$$

where Y_i is the logarithm of the ith observed lifetime, x_i is an $m \times 1$ vector of values of explanatory variables associated with the ith response Y_i, $\beta = (\beta_1, \ldots, \beta_p)^\top$ is a vector of unknown nonlinear parameters to be estimated, and $\varepsilon_i \sim \mathcal{SN}(\alpha, 0, 2)$. We assume a nonlinear structure for the location parameter $\mu_i = f_i(x_i; \beta)$ in model (4.13), where f_i is a known and twice continuously differentiable function with respect to β. For the linear regression $\mu_i = x_i^\top \beta$, the model (4.13) reduces to Rieck and Nedelman's (1991) model.

The log-likelihood function $\ell(\theta)$ given in (2.18) is assumed to be regular (Cox and Hinkley 1974, Chap. 9) with respect to all β and α derivatives up to third order. It is well known that, under general regularity conditions (Cox and Hinkley 1974, Chap. 9), the MLEs are consistent, asymptotically efficient, and asymptotically normal. Let $\hat{\theta} = (\hat{\beta}^\top, \hat{\alpha})^\top$ be the MLE of $\theta = (\beta^\top, \alpha)^\top$. We can write $\hat{\theta} \overset{a}{\sim} \mathcal{N}_{p+1}(\theta, K_\theta^{-1})$ for large n, where $\overset{a}{\sim}$ denotes approximately distributed, K_θ is the block diagonal Fisher information matrix given by $K_\theta = \mathrm{diag}\{K_\beta, \kappa_{\alpha,\alpha}\}$, K_θ^{-1} is its inverse, $K_\beta = \psi_1(\alpha)(D^\top D)/4$ is the information for β, and $\kappa_{\alpha,\alpha} = 2n/\alpha^2$ is the information for α. Here, the $n \times p$ local matrix $D = D(\beta) = \partial\mu/\partial\beta$ of partial derivatives of μ with respect to β is assumed to be of full rank p for all β and

$$\psi_1(\alpha) = 2 + \frac{4}{\alpha^2} - \frac{\sqrt{2\pi}}{\alpha}\left[1 - \text{erf}\left(\frac{\sqrt{2}}{\alpha}\right)\right]\exp\left(\frac{2}{\alpha^2}\right),$$

where $\text{erf}(z) = (2/\pi)\int_0^z e^{-t^2}dt$ is the error function. Since K_θ is block diagonal, the vector β and the scalar α are globally orthogonal (Cox and Reid 1987), and $\hat{\beta}$ and $\hat{\alpha}$ are asymptotically independent. It can be shown (Rieck 1989) that $\psi_1(\alpha) \approx 1+4/\alpha^2$ for α small and $\psi_1(\alpha) \approx 2$ for α large.

In what follows we shall use the notation $d_{ir} = \partial\mu_i/\partial\beta_r$ and $g_{irs} = \partial^2\mu_i/\partial\beta_r\partial\beta_s$ for the first and second partial derivatives of μ_i with respect to the elements of β. The joint cumulants and their derivatives are

$$\kappa_{rs} = -\frac{\psi_1(\alpha)}{4}\sum_{i=1}^{n} d_{ir}d_{is}, \quad \kappa_{r\alpha} = \kappa_{r\alpha\alpha} = 0, \quad \kappa_{\alpha\alpha} = -\frac{2n}{\alpha^2}, \quad \kappa_{\alpha\alpha\alpha} = \frac{10n}{\alpha^3},$$

$$\kappa_{rst} = -\frac{\psi_1(\alpha)}{4}\sum_{i=1}^{n}(g_{irs}d_{it} + g_{irt}d_{is} + d_{ir}g_{ist}), \quad \kappa_{rs\alpha} = \frac{(2+\alpha^2)}{\alpha^3}\sum_{i=1}^{n} d_{ir}d_{is},$$

$$\kappa_{rs}^{(t)} = -\frac{\psi_1(\alpha)}{4}\sum_{i=1}^{n}(g_{irt}d_{is} + d_{ir}g_{ist}), \quad \kappa_{r\alpha}^{(\alpha)} = \kappa_{r\alpha}^{(s)} = 0 \quad \text{and} \quad \kappa_{\alpha\alpha}^{(\alpha)} = \frac{4n}{\alpha^3}.$$

Let $B(\hat{\beta}_a)$ and $B(\hat{\alpha})$ be the n^{-1} biases of $\hat{\beta}_a$ $(a = 1, \ldots, p)$ and $\hat{\alpha}$, respectively. The use of Eq. (4.4) to obtain these biases is greatly simplified, since β and α are globally orthogonal and the cumulants corresponding to the parameters in β are invariant under their permutation. We have

$$B(\hat{\beta}_a) = \sum_{s,t,u}{}'\kappa^{a,s}\kappa^{t,u}\left(\kappa_{st}^{(u)} - \frac{1}{2}\kappa_{stu}\right) + \kappa^{\alpha,\alpha}\sum_{s}{}'\kappa^{a,s}\left(\kappa_{s\alpha}^{(\alpha)} - \frac{1}{2}\kappa_{s\alpha\alpha}\right) \quad (4.14)$$

and

$$B(\hat{\alpha}) = (\kappa^{\alpha,\alpha})^2\left(\kappa_{\alpha\alpha}^{(\alpha)} - \frac{1}{2}\kappa_{\alpha\alpha\alpha}\right) + \kappa^{\alpha,\alpha}\sum_{t,u}{}'\kappa^{t,u}\left(\kappa_{\alpha t}^{(u)} - \frac{1}{2}\kappa_{\alpha tu}\right), \quad (4.15)$$

where $\kappa^{r,s}$ is the (r, s)th element of K_β^{-1} (the inverse of the information matrix for β), $\kappa^{\alpha,\alpha} = \kappa_{\alpha,\alpha}^{-1}$ and $\sum'$ denotes, here and from now on, the summation over all combinations of parameters $\beta_1, \ldots, \beta_p$.

First, we consider Eq. (4.14) from which we readily have that the second sum is zero since $\kappa_{s\alpha\alpha} = \kappa_{s\alpha}^{(\alpha)} = 0$. By rearranging the summation terms, we can write

$$B(\hat{\beta}_a) = -\frac{\psi_1(\alpha)}{8}\sum_{i=1}^{n}\sum_{s}{}'\kappa^{a,s}d_{is}\sum_{t,u}{}'\kappa^{t,u}g_{itu}.$$

Let $d_i^\top$ $(1 \times p)$ and $g_i^\top$ $(1 \times p^2)$ be vectors containing the first and second partial derivatives of the mean μ_i with respect to the β's. In matrix notation,

$$B(\hat{\beta}_a) = -\frac{\psi_1(\alpha)}{8}\rho_a^\top K_\beta^{-1} D^\top G \text{vec}(K_\beta^{-1}),$$

where $\rho_a^\top$ is the ath row of the $p \times p$ identity matrix, vec$(\cdot)$ is the operator which transforms a matrix into a vector by stacking the columns of the matrix one underneath the other, and $G = \partial^2\mu/\partial\beta^\top\partial\beta = (g_1, \ldots, g_n)^\top$ is a $n \times p^2$ matrix of second partial derivatives of the mean vector μ with respect to β. The n^{-1} bias vector $B(\hat{\beta})$ of $\hat{\beta}$ can then be expressed as

$$B(\hat{\beta}) = (D^\top D)^{-1} D^\top d, \tag{4.16}$$

where d is an $n \times 1$ vector defined as $d = -[2/\psi_1(\alpha)]\, G\, \text{vec}\{(D^\top D)^{-1}\}$.

We can now obtain the n^{-1} bias of $\hat{\alpha}$. Using (4.15), we can write

$$\begin{aligned} B(\hat{\alpha}) &= -\frac{\alpha}{4n} - \frac{(2+\alpha^2)}{4\alpha n}\sum_{i=1}^{n}\sum_{t,u}{}'\kappa^{t,u} d_{it} d_{iu} = -\frac{\alpha}{4n} - \frac{(2+\alpha^2)}{4\alpha n}\sum_{i=1}^{n} d_i^\top K_\beta^{-1} d_i \\ &= -\frac{\alpha}{4n} - \frac{(2+\alpha^2)}{4\alpha n}\text{tr}(D K_\beta^{-1} D^\top). \end{aligned}$$

Since $\text{tr}(D K_\beta^{-1} D^\top) = 4p/\psi_1(\alpha)$, we can rewrite $B(\hat{\alpha})$ as

$$B(\hat{\alpha}) = -\frac{1}{n}\left\{ p\left[\frac{2+\alpha^2}{\alpha\psi_1(\alpha)}\right] + \frac{\alpha}{4}\right\}. \tag{4.17}$$

The bias vector $B(\hat{\beta})$ can be determined from a simple OLS regression of d on the columns of D. It depends on the nonlinearity of the regression function f and on the parameter α. The bias $B(\hat{\beta})$ is small when d is orthogonal to the columns of D. It may be large when $\psi_1(\alpha)$ and n are both small. Equation (4.16) is easily handled algebraically for any type of nonlinear regression, since it involves simple operations on matrices and vectors. For special models with closed-form information matrix for β, it is possible to obtain closed-form expressions for $B(\hat{\beta})$. For linear models, the matrix G and the vector d vanish and hence $B(\hat{\beta}) = 0$, which is in agreement with the result due to Rieck and Nedelman (1991, p. 54). Equation (4.17) depends on the nonlinear structure of the regression model only through the rank p of D. It reveals that the bias is always a linear function of the dimension p of β.

By replacing the unknown parameters on the right-hand sides of (4.16) and (4.17), which are both of order n^{-1}, by the corresponding MLEs, we obtain the bias-corrected estimators $\tilde{\beta} = \hat{\beta} - \hat{B}(\hat{\beta})$ and $\tilde{\alpha} = \hat{\alpha} - \hat{B}(\hat{\alpha})$, where $\hat{B}(\hat{\beta})$ and $\hat{B}(\hat{\alpha})$ are the values of $B(\hat{\beta})$ and $B(\hat{\alpha})$, respectively, at $\hat{\theta} = (\hat{\beta}^\top, \hat{\alpha})^\top$. The bias-corrected estimates $\tilde{\beta}$ and $\tilde{\alpha}$ are expected to have better sampling properties than the classical MLEs $\hat{\beta}$ and $\hat{\alpha}$. In fact, simulation results presented in Sect. 4.8 show that $\tilde{\beta}$ and $\tilde{\alpha}$ have smaller biases

than their unmodified counterparts, thus indicating that the bias corrections have the effect of shrinking the modified estimates toward the true parameter values.

We now calculate the second-order bias $B(\hat{\mu}_i)$ of the MLE $\hat{\mu}_i$ of the ith mean $\mu_i = f_i(x_i; \beta)$. We can easily verify by Taylor series expansion that

$$B(\hat{\mu}_i) = d_i^\top B(\hat{\beta}) + \frac{1}{2}\,\mathrm{tr}[M_i \mathrm{Cov}(\hat{\beta})],$$

where M_i is a $p \times p$ matrix of second partial derivatives $\partial^2 \mu_i / \partial\beta_r \partial\beta_s$ (for $r, s = 1, \ldots, p$), $\mathrm{Cov}(\hat{\beta}) = K_\beta^{-1}$ is the asymptotic covariance matrix of $\hat{\beta}$, and the vectors d_i and $B(\hat{\beta})$ are as defined before. All quantities in the above equation should be evaluated at $\hat{\beta}$.

The asymptotic variance of $\hat{\mu}_i$ can also be expressed explicitly in terms of the covariance of $\hat{\beta}$:

$$\mathrm{Var}(\hat{\mu}_i) = \mathrm{tr}[(d_i d_i^\top)\mathrm{Cov}(\hat{\beta})].$$

4.7 Special Models

Equation (4.16) is easily handled algebraically for any type of nonlinear model, since it involves simple operations on matrices and vectors. This equation, in conjunction with a computer algebra system such as MATHEMATICA or MAPLE, can be used to compute $B(\hat{\beta})$ algebraically with minimal effort. In particular, (4.16) can be considerably simplified when the number of nonlinear parameters is small. Moreover, for any special nonlinear model, we can calculate the bias $B(\hat{\beta})$ numerically using a software with numerical linear algebra facilities such as OX (Doornik 2009) and R (R Development Core Team 2006).

First, we consider a nonlinear regression model which depends on a single nonlinear parameter β. Equation (4.16) gives

$$B(\hat{\beta}) = -\frac{2}{\psi_1(\alpha)} \frac{\kappa_2}{\kappa_1^2},$$

where $\kappa_1 = \sum_{i=1}^n (\mathrm{d}f_i/\mathrm{d}\beta)^2$ and $\kappa_2 = \sum_{i=1}^n (\mathrm{d}f_i/\mathrm{d}\beta)(\mathrm{d}^2 f_i/\mathrm{d}\beta^2)$. The constants κ_1 and κ_2 are evaluated at $\hat{\beta}$ and $\hat{\alpha}$ to yield $\hat{B}(\hat{\beta})$ and the corrected estimator $\tilde{\beta} = \hat{\beta} - \hat{B}(\hat{\beta})$. For example, the simple exponential model $f_i = \exp(\beta x_i)$ yields $\kappa_1 = \sum_{i=1}^n x_i^2 \exp(2\beta x_i)$ and $\kappa_2 = \sum_{i=1}^n x_i^3 \exp(2\beta x_i)$.

As a second application, we consider a partially nonlinear regression model defined by

$$\mu = Z\lambda + \eta\, g(\gamma), \tag{4.18}$$

where Z is a known $n \times (p-2)$ matrix of full rank, $g(\gamma)$ is an $n \times 1$ vector, $\beta = (\lambda^\top, \eta, \gamma)^\top$, $\lambda = (\lambda_1, \ldots, \lambda_{p-2})^\top$ and η and γ are scalar parameters. This class of

models occurs very often in statistical modeling, see Cook et al. (1986) and Cordeiro et al. (2000). Here, we consider three examples: $\mu = \lambda_1 z_1 + \lambda_2 z_2 + \eta \exp(\gamma x)$, $\mu = \lambda - \eta \log(x_1 + \gamma x_2)$, and $\mu = \lambda + \eta \log(x_1/(\gamma + x_2))$. Ratkowsky (1983, Chap. 5) discussed several models of the form (4.18) which include the asymptotic regression and Weibull-type models given by $\mu = \lambda - \eta\gamma^x$ and $\mu = \lambda - \eta \exp(-\gamma x)$, respectively.

The $n \times p$ local model matrix D takes the form $D = [Z, g(\gamma), \eta(\mathrm{d}g(\gamma)/\mathrm{d}\gamma)]$. After some algebra, we obtain from (4.16) $B(\hat{\beta}) = (D^\top D)^{-1} D^\top (\mathrm{d}^2 g(\gamma)/\mathrm{d}\gamma^2)$, which is simply the set of coefficients from the ordinary regression of the vector $\mathrm{d}^2 g(\gamma)/\mathrm{d}\gamma^2$ on the matrix D. Clearly, the vector $B(\hat{\beta})$ does not depend explicitly on the linear parameters in λ. Further, the covariance term $\mathrm{Cov}(\hat{\eta}, \hat{\gamma})$ only contributes to the bias of $\hat{\gamma}$.

4.8 Monte Carlo Simulation Evidence

We shall now present some Monte Carlo simulation results on the finite-sample performance of the unmodified and bias-reduced MLEs. Parameter estimates are calculated by maximizing the log-likelihood function using the BFGS quasi-Newton method with analytical derivatives. The covariate values are selected as random draws from the standard uniform $\mathscr{U}(0, 1)$ distribution, and, for each sample size considered, those values are kept constant throughout the experiment. The number of Monte Carlo replications is 10,000. All simulations are performed using the Ox matrix programming language (Doornik 2009).

In order to analyze the performance of the estimators, we compute, for each sample size and for each estimate, the relative bias (the relative bias of an estimate $\hat{\theta}$, defined as $\{\mathbb{E}(\hat{\theta}) - \theta\}/\theta$, is obtained by empirically calculating $\mathbb{E}(\hat{\theta})$ by Monte Carlo) and the root-mean-square error ($\sqrt{\mathrm{MSE}}$), where MSE is the estimated mean-square error from the 10,000 Monte Carlo replications.

First, consider the nonlinear regression model

$$\mu_i = \lambda_1 z_{i1} + \lambda_2 z_{i2} + \eta \exp(\gamma x_i),$$

where $\varepsilon_i \sim \mathscr{SN}(\alpha, 0, 2)$ for $i = 1, \ldots, n$. The sample sizes are $n = 15$, 30 and 45. Without loss of generality, the true values of the regression parameters are taken as $\lambda_1 = 4$, $\lambda_2 = 5$, $\eta = 3$, $\gamma = 1.5$, and $\alpha = 0.5$ and 1.5.

Table 4.1 displays the relative biases of both uncorrected and corrected estimators. (In the table, BCE stands for 'bias-corrected estimator'.) Notice that the bias-corrected estimates are much closer to the true parameters than the unadjusted estimates. For instance, when $n = 15$ and $\alpha = 1.5$, the average of the estimated relative biases for the model parameters estimators is -0.03224, whereas the average of the estimated relative biases for the corrected estimates is -0.0086. Hence, the average bias (in absolute value) of the MLEs is almost four times greater than the

Table 4.1 Relative biases of the uncorrected and corrected estimators

α	n		λ_1	λ_2	η	γ	α
0.5	15	MLE	0.0006	−0.0013	0.0011	0.0020	−0.1691
		BCE	0.0007	−0.0011	0.0001	0.0008	−0.0395
	30	MLE	0.0001	−0.0013	0.0013	0.0009	−0.0811
		BCE	0.0002	−0.0012	0.0007	−0.0001	−0.0092
	45	MLE	0.0003	−0.0012	0.0007	0.0008	−0.0537
		BCE	0.0003	−0.0011	0.0003	0.0001	−0.0042
1.5	15	MLE	−0.0068	−0.0083	0.0248	0.0197	−0.1916
		BCE	−0.0055	−0.0046	0.0113	0.0056	−0.0481
	30	MLE	−0.0016	−0.0034	0.0079	0.0078	−0.0933
		BCE	−0.0011	−0.0018	0.0027	0.0012	−0.0116
	45	MLE	−0.0028	−0.0027	0.0052	0.0026	−0.0614
		BCE	−0.0023	−0.0018	0.0023	−0.0005	−0.0048

Table 4.2 Root-mean-square errors of the uncorrected and corrected estimators

α	n		λ_1	λ_2	η	γ	α
0.5	15	MLE	0.4093	0.4920	0.2707	0.0924	0.1234
		BCE	0.4093	0.4921	0.2709	0.0922	0.1067
	30	MLE	0.3006	0.3806	0.2113	0.0688	0.0763
		BCE	0.3006	0.3806	0.2114	0.0686	0.0702
	45	MLE	0.2434	0.2874	0.1768	0.0567	0.0590
		BCE	0.2434	0.2874	0.1769	0.0566	0.0555
1.5	15	MLE	1.6302	1.1230	0.9756	0.3235	0.3938
		BCE	1.6333	1.1274	0.9819	0.3152	0.3315
	30	MLE	0.9684	0.7003	0.5785	0.1931	0.2399
		BCE	0.9693	0.7011	0.5807	0.1908	0.2155
	45	MLE	0.6505	0.5575	0.3895	0.1318	0.1837
		BCE	0.6507	0.5577	0.3901	0.1311	0.1700

average bias of the corrected estimates. This indicates that the second-order biases of the MLEs should not be ignored in samples of small to moderate size, since they can be non-negligible.

When the value of α increases, the finite-sample performance of the MLEs deteriorates (see Tables 4.1 and 4.2). For instance, when $n = 15$, the relative biases of $\hat{\gamma}$ (MLE) and $\tilde{\gamma}$ (BCE) are 0.0020 and 0.0008 (for $\alpha = 0.5$) and 0.0197 and 0.0056 (for $\alpha = 1.5$), which indicate an increase in the relative biases of nearly 10 and 7 times, respectively. Also, the root-mean-square errors in the same order are 0.0924 and 0.0922 (for $\alpha = 0.5$) and 0.3235 and 0.3152 (for $\alpha = 1.5$). In addition, all estimators have similar root-mean-square errors (see Table 4.2).

Next, we consider the Michaelis–Menton model, which is very useful for estimating growth curves, where it is common for the response to approach an asymptote as the stimulus increases. The Michaelis–Menton model (McCullagh and Nelder 1989, p. 16) provides an hyperbolic form for μ_i against x_i given by

Table 4.3 Relative biases and root-mean-squared erros of uncorrected and corrected estimators; $\alpha = 0.5$ and different sample sizes

n		Relative bias			$\sqrt{\text{MSE}}$		
		η	γ	α	η	γ	α
20	MLE	0.0476	0.1718	−0.0669	0.6984	0.3947	0.0859
	BCE	−0.0016	−0.0081	−0.0061	0.5264	0.2783	0.0847
30	MLE	0.0313	0.1077	−0.0439	0.5245	0.2750	0.0684
	BCE	0.0004	0.0012	−0.0024	0.4478	0.2252	0.0678
40	MLE	0.0215	0.0754	−0.0330	0.4222	0.2207	0.0582
	BCE	−0.0001	−0.0003	−0.0015	0.3835	0.1954	0.0578
50	MLE	0.0160	0.0558	−0.0259	0.3609	0.1862	0.0516
	BCE	0.0000	−0.0001	−0.0005	0.3380	0.1710	0.0514

$$\mu_i = \frac{\eta x_i}{\gamma + x_i}, \quad i = 1, 2, \ldots, n,$$

where the curve has an asymptote at $\mu = \eta$. Here, the sample sizes are $n = 20, 30, 40$ and 50. Also, the true values of the regression parameters are $\eta = 3$ and $\gamma = 0.5$, with $\alpha = 0.5$.

Table 4.3 lists the relative biases and root-mean-squared erros of both uncorrected and corrected estimators. The figures in this table indicate that the MLEs of the model parameters can be substantially biased, even when $n = 50$, and that the bias correction presented in the previous section is quite effective. This shows the importance of using a bias correction. In addition, all estimators have similar root-mean-square errors.

4.9 An Application

Here, we consider an application to a biaxial fatigue data set reported by Rieck and Nedelman (1991) on the life of a metal piece in cycles to failure. The response N is the number of cycles prior to failure, and the explanatory variable w is the work per cycle (mJ/m^3). The data contain forty-six observations and were taken from Table 4.1 of Galea et al. (2004). We consider the nonlinear regression model

$$Y_i = \beta_1 + \beta_2 \exp(\beta_3 / w_i) + \varepsilon_i, \quad i = 1, \ldots, 46, \tag{4.19}$$

where $Y_i = \log(N_i)$ and $\varepsilon_i \sim \mathcal{SN}(\alpha, 0, 2)$. The uncorrected estimates (estimated standard errors in parentheses) are $\hat{\beta}_1 = 8.988$ (0.744), $\hat{\beta}_2 = -5.180$ (0.508), $\hat{\beta}_3 = -22.520$ (7.378), and $\hat{\alpha} = 0.40$ (0.042). The bias-corrected estimates are as follows: $\tilde{\beta}_1 = 8.781$ (0.773), $\tilde{\beta}_2 = -4.936$ (0.527), $\tilde{\beta}_3 = -22.171$ (7.655), and $\tilde{\alpha} = 0.42$ (0.043). Hence, the uncorrected estimates are slightly different from the bias-corrected estimates even for large samples ($n = 46$ observations).

4.10 Linear Heteroskedastic Regression

The linear regression model is commonly used in many different fields, such as chemistry, economics, engineering, finance, medicine, and psychology. The model is

$$Y = X\beta + \varepsilon,$$

where Y and ε are $n \times 1$ vectors of responses and random errors, respectively, X is a full rank $n \times p$ matrix of fixed explanatory variables ($\text{rank}(X) = p < n$), and $\beta = (\beta_1, \ldots, \beta_p)^\top$ is a $p \times 1$ vector of unknown regression parameters to be estimated, with n being the sample size. It can also be expressed as

$$Y_i = x_i^\top \beta + \varepsilon_i, \quad i = 1, \ldots, n,$$

where x_i is the ith row of X. The error ε_i has mean zero, variance $0 < \sigma_i^2 < \infty$, $i = 1, \ldots, n$, and is uncorrelated with ε_j for all $j \neq i$. The error covariance matrix is $\Omega = \text{Cov}(\varepsilon) = \text{diag}\{\sigma_1^2, \ldots, \sigma_n^2\}$. Note that the n variances show up in the main diagonal and that all off-diagonal elements equal zero (since each error is uncorrelated with all other errors).

The parameter vector β can be estimated using the method of ordinary least squares, i.e., finding the value of β that minimizes the sum of squared errors $\sum_{i=1}^n \varepsilon_i^2 = (Y - X\beta)^\top (Y - X\beta)$. It is easy to prove that the OLSE of β can be expressed in closed form as $\hat{\beta} = (X^\top X)^{-1} X^\top Y$. Its covariance matrix is $\Psi = \text{Cov}(\hat{\beta}) = P \Omega P^\top$, where $P = (X^\top X)^{-1} X^\top$. The main diagonal of Ψ contains the variances of $\hat{\beta}_1, \ldots, \hat{\beta}_p$, and the off-diagonal elements are the covariances. Under homoskedasticity (i.e., when all errors share the same variance), $\sigma_i^2 = \sigma^2$, for $i = 1, \ldots, n$, where $\sigma^2 > 0$, and hence $\Psi = \sigma^2 (X^\top X)^{-1}$. The covariance matrix Ψ can then be easily estimated by $\hat{\Psi} = \hat{\sigma}^2 (X^\top X)^{-1}$, where $\hat{\sigma}^2 = (Y - X\hat{\beta})^\top (Y - X\hat{\beta})/(n - p)$.

Under unequal error variances, the OLSE of β is unbiased, consistent, and asymptotically normal, although it is no longer the best (least variance) linear unbiased estimator. A common practice is to estimate β by least squares and to base interval estimation and hypothesis testing inference on an estimator of $\text{Cov}(\hat{\beta})$ that is consistent under both homoskedasticity and heteroskedasticity of unknown form. The most commonly used estimator was proposed by Halbert White in a highly influential paper (White 1980). His estimator is commonly referred to as 'HC0' and is given by

$$\text{HC0} = \hat{\Psi} = P \hat{\Omega} P^\top,$$

where $\hat{\Omega} = \text{diag}\{\hat{\varepsilon}_1^2, \ldots, \hat{\varepsilon}_n^2\}$. Here, $\hat{\varepsilon}_i$ is the ith least-squares residual, i.e., $\hat{\varepsilon}_i = Y_i - x_i^\top \hat{\beta}$, $i = 1, \ldots, n$. The vector of OLS residuals is $\hat{\varepsilon} = (\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n)^\top = (I - H)Y$, where $H = X(X^\top X)^{-1} X^\top = XP$ and I is the $n \times n$ identity matrix. White's estimator is consistent under both equal and unequal error variances: $\text{plim}(\Psi^{-1} \hat{\Psi})$

equals the $p \times p$ identity matrix in both cases, where plim denotes limit in probability. That is, $\Psi^{-1}\hat{\Psi}$ converges in probability to I_p. It has, nonetheless, an important shortcoming: It tends to be considerably biased in small to moderately large samples. More specifically, it tends to underestimate the true variances, more so when the data contain leverage points; see, e.g., Chesher and Jewitt (1987).

Cribari-Neto et al. (2000) proposed an interative bias-correcting scheme for HC0. Their sequence of estimators was obtained by correcting HC0, then correcting the resulting adjusted estimator, and so on. Let $(A)_d$ denotes the diagonal matrix obtained by setting the non-diagonal elements of the square matrix A equal to zero. The authors demonstrated that the biases of $\hat{\Omega}$ and $\hat{\Psi}$ as estimators of Ω and Ψ are $B_{\hat{\Omega}}(\Omega) = \mathbb{E}(\hat{\Omega}) - \Omega = \{H\Omega(H-2I)\}_d$ and $B_{\hat{\Psi}}(\Omega) = \mathbb{E}(\hat{\Psi}) - \Psi = PB_{\hat{\Omega}}(\Omega)P^\top$, respectively. They then defined the bias-corrected estimator $\hat{\Omega}^{(1)} = \hat{\Omega} - B_{\hat{\Omega}}(\hat{\Omega})$. This new estimator can also be adjusted for bias as $\hat{\Omega}^{(2)} = \hat{\Omega}^{(1)} - B_{\hat{\Omega}^{(1)}}(\hat{\Omega})$. It is also possible to adjust $\hat{\Omega}^{(2)}$ for bias. After k iterations of the bias-correcting scheme, one obtains

$$\hat{\Omega}^{(k)} = \hat{\Omega}^{(k-1)} - B_{\hat{\Omega}^{(k-1)}}(\hat{\Omega}).$$

The kth order bias-corrected estimator and its respective bias are given by $\hat{\Omega}^{(k)} = \sum_{j=0}^{k}(-1)^j M^{(j)}(\hat{\Omega})$ and $B_{\hat{\Omega}^{(k)}}(\Omega) = (-1)^k M^{(k+1)}(\Omega)$, for $k = 1, 2, \ldots$. The above notation uses the recursive function of an $n \times n$ diagonal matrix A given by $M^{(k+1)}(A) = M^{(1)}(M^{(k)}(A))$, for $k = 0, 1, \ldots$, where $M^{(0)}(A) = A$ and $M^{(1)}(A) = \{HA(H-2I)\}_d$.

A sequence of bias-corrected covariance matrix estimators can be defined as $\{\hat{\Psi}^{(k)}, k = 1, 2, \ldots\}$, where $\hat{\Psi}^{(k)} = P\hat{\Omega}^{(k)}P^\top$. The bias of $\hat{\Psi}^{(k)}$ is

$$B_{\hat{\Psi}^{(k)}}(\Omega) = (-1)^k PM^{(k+1)}(\Omega)P^\top, \quad k = 1, 2, \ldots.$$

Assume that the matrix of explanatory variables X is such that P and H are $\mathcal{O}(n^{-1})$ and that Ω is $\mathcal{O}(1)$. It can then be shown that $B_{\hat{\Psi}}(\Omega) = \mathcal{O}(n^{-2})$, i.e., the bias of HC0 is of order $\mathcal{O}(n^{-2})$. Cribari-Neto et al. (2000) have also shown that $B_{\hat{\Psi}^{(k)}}(\Omega) = \mathcal{O}(n^{-(k+2)})$. That is, the bias of the kth corrected estimator is of order $\mathcal{O}(n^{-(k+2)})$, whereas the bias of White's estimator is $\mathcal{O}(n^{-2})$. Notice that the biases of the corrected estimators decay faster than that of HC0, as the sample size increases, more so for large values of k.

An alternative sequence of bias-corrected estimators of Ψ was obtained by Cribari-Neto and Lima (2011). They obtained faster convergence rates by estimating the bias in each step of the sequence in different fashion. Notice that when constructing the sequence of estimators adjusted for systematic error, one subtracts *the estimated bias* from the estimator and then proceeds to bias-correct it. The estimated bias used by Cribari-Neto et al. (2000) is obtained by evaluating the bias function $B(\Omega)$ at $\hat{\Omega}$, where $\hat{\Omega}$ is a diagonal matrix containing the vector of squared OLS residuals. Cribari-Neto and Lima (2011) estimated the biases of the corrected estimators by evaluating the bias functions at a more accurate estimator of Ω, which in turn yields a more accurate estimator of $X^\top \Omega X$. They defined the sequence of modified

estimators of Ω

$$\hat{\Omega}_M^{(k)} = \hat{\Omega}_M^{(k-1)} - B_{\hat{\Omega}_M^{(k-1)}}(\hat{\Omega}_M^{(k-1)}), \quad k = 1, 2, \ldots.$$

Note that the true biases are no longer evaluated at $\hat{\Omega}$. They are instead evaluated at the estimate of Ω obtained in the previous step of the iterative scheme. As expected, $B_{\hat{\Omega}_M^{(k-1)}}(\hat{\Omega}_M^{(k-1)})$ estimates the true bias of $\hat{\Omega}_M^{(k-1)}$ much more accurately than $B_{\hat{\Omega}_M^{(k-1)}}(\hat{\Omega})$.

The first corrected estimator is $\hat{\Omega}_M^{(1)} = \hat{\Omega} - B_{\hat{\Omega}}(\hat{\Omega})$. It equals the first bias-corrected estimator in the sequence of estimators proposed by Cribari-Neto et al. (2000). It was shown by Cribari-Neto and Lima (2011) that the remaining elements of the sequence are

$$\hat{\Omega}_M^{(k)} = \sum_{j=0}^{2^k-1} (-1)^j M^{(j)}(\hat{\Omega}), \quad k = 2, 3, \ldots,$$

and

$$B_{\hat{\Omega}_M^{(k)}}(\Omega) = -M^{(2^k)}(\Omega), \quad k = 1, 2, \ldots$$

The authors then obtained a new sequence of corrected estimators for $\hat{\Psi}$: $\{\hat{\Psi}_M^{(k)}, k = 1, 2, \ldots\}$, where

$$\hat{\Psi}_M^{(k)} = P\hat{\Omega}_M^{(k)}P^\top \quad \text{and} \quad B_{\hat{\Psi}_M^{(k)}}(\Omega) = -PM^{(2^k)}(\Omega)P^\top.$$

Assume again that the matrix of explanatory variables X is such that $P = (X^\top X)^{-1}X^\top$ and $H = X(X^\top X)^{-1}X^\top$ are $\mathcal{O}(n^{-1})$. Cribari-Neto and Lima (2011) demonstrated that $M^{(k)}(\Omega) = \mathcal{O}(n^{-k})$ and $PM^{(k)}(\Omega)P^\top = \mathcal{O}(n^{-(k+1)})$. They have also shown that $B_{\hat{\Omega}_M^{(k)}}(\Omega) = \mathcal{O}(n^{-2^k})$ and $B_{\hat{\Psi}_M^{(k)}}(\Omega) = \mathcal{O}(n^{-(2^k+1)})$. It is noteworthy that the biases of the estimators in the above sequence vanish at a much faster rate than those of the estimators proposed by Cribari-Neto et al. (2000). The bias order of the kth estimator is $\mathcal{O}(n^{-(2^k+1)})$, whereas the corresponding bias order of their estimator is $\mathcal{O}(n^{-(k+2)})$. For example, when $k = 4$, the bias orders are $\mathcal{O}(n^{-6})$ (for Cribari-Neto et al. 2000) and $\mathcal{O}(n^{-17})$ (for Cribari-Neto and Lima 2011).

The numerical evidence reported by Cribari-Neto and Lima (2011) favors the sequence of estimators relative to that of Cribari-Neto et al. (2000). They considered a simple linear regression model and computed the total relative biases. For each estimator, they calculated and reported

$$\frac{|\mathbb{E}\{\hat{\text{var}}(\hat{\beta}_1)\} - \text{var}(\hat{\beta}_1)|}{\text{var}(\hat{\beta}_1)} + \frac{|\mathbb{E}\{\hat{\text{var}}(\hat{\beta}_2)\} - \text{var}(\hat{\beta}_2)|}{\text{var}(\hat{\beta}_2)},$$

where 'vâr' denotes the relevant variance estimator. When the data are heteroskedastic and the sample contains only 20 observations, the total relative biases of HC0 (White's estimator), the fourth estimator in the Cribari-Neto et al. (2000) sequence, and the fourth estimator in Cribari-Neto and Lima (2011) sequence are, respectively, 0.825, 0.338, and 0.036. It is noteworthy that the total relative bias of Cribari-Neto and Lima (2011) estimator (four iterations, $\hat{\Psi}_M^{(4)}$) is approximately ten times smaller than that of the corresponding estimator of Cribari-Neto et al. (2000), $\hat{\Psi}^{(4)}$, and nearly 23 times smaller than the total relative bias of HC0.

4.11 Beta Regressions

Oftentimes, one wishes to model random variables that assume values in the standard unit interval $(0, 1)$, such as rates and proportions. Ferrari and Cribari-Neto (2004) proposed a beta regression model which allows such a modeling to be conditioned on a set of explanatory variables. They have used an alternative parameterization in which the beta density is indexed by mean and precision parameters. In their model, the mean of the response is related to a linear predictor that involves explanatory variables and unknown regression parameters through a link function.

The random variable Y is said to be beta distributed, denoted by $Y \sim \mathscr{B}(p, q)$, if its density function is given by

$$f(y; p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}, \quad 0 < y < 1, \quad p, q > 0, \tag{4.20}$$

where $\Gamma(\cdot)$ is the gamma function. Ferrari and Cribari-Neto (2004) introduced an alternative beta parameterization for (4.20). Specifically, let

$$\mu = p/(p+q) \quad \text{and} \quad \phi = p+q,$$

i.e.,

$$p = \mu\phi \quad \text{and} \quad q = (1-\mu)\phi.$$

It is easy to verify that

$$\mathbb{E}(Y) = \mu \quad \text{and} \quad \text{Var}(Y) = \frac{V(\mu)}{1+\phi},$$

where $V(\mu) = \mu(1-\mu)$. Here, μ is the mean and ϕ can be regarded as a precision parameter in the sense that, for fixed μ (i.e., for a given mean value), the larger the value of ϕ, the smaller the variance of Y. Then, the density of Y can then be expressed as

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad 0 < y < 1,$$
$$0 < \mu < 1 \text{ and } \phi > 0.$$

It is noteworthy that the more general case in which the random variable assumes values in (a, b), where a and b are known constants such that $a < b$, can be easily handled by modeling $(Y - a)/(b - a)$, which assumes values in the standard unit interval. In what follows, we focus, without loss of generality, on the responses that assume values in $(0, 1)$.

Let $Y_1, \ldots, Y_n$ be a random sample such that $Y_i \sim \mathscr{B}(\mu_i, \phi)$, $i = 1, \ldots, n$. The beta regression model was defined by Ferrari and Cribari-Neto (2004) as

$$g(\mu_i) = \sum_{j=1}^{p} x_{ij}\beta_j = \eta_i,$$

where $\beta = (\beta_1, \ldots, \beta_k)^\top$ is a p-vector of unknown regression parameters ($\beta \in \mathbb{R}^p$) to be estimated, η_i is a linear predictor and $x_{i1}, \ldots, x_{ip}$ are (fixed) explanatory variables values ($p < n$). The link function $g : (0; 1) \to \mathbb{R}$ must be strictly monotone and twice differentiable. Some standard link functions are as follows:

1. Cauchy:
$$g(\mu) = \tan\{\pi(\mu - 0.5)\}.$$
2. complementary log–log:
$$g(\mu) = \log\{-\log(1-\mu)\};$$
3. log–log:
$$g(\mu) = -\log\{-\log(\mu)\};$$
4. logit:
$$g(\mu) = \log(\mu/(1-\mu));$$
5. probit:
$$g(\mu) = \Phi^{-1}(\mu).$$

The beta regression log-likelihood function is

$$\ell(\beta, \phi) = \sum_{t=1}^{n} \ell_i(\mu_i, \phi),$$

where

$$\ell_i(\mu_i, \phi) = \log \Gamma(\phi) - \log \Gamma(\mu_i \phi) - \log \Gamma((1-\mu_i)\phi) + (\mu_i \phi - 1) \log y_i + \{(1-\mu_i)\phi - 1\} \log(1-y_i).$$

Readers should notice that $\mu_i = g^{-1}(\eta_i)$ is a function of β. Parameter estimation is carried out by numerically maximizing the log-likelihood function, which can be done with the aid of a Newton (e.g., Newton–Raphson) or quasi-Newton (e.g., BFGS) algorithm.

Fisher's information matrix for the parameter vector (β, ϕ) can be shown to be

$$K = K(\beta, \phi) = \begin{pmatrix} K_{\beta\beta} & K_{\beta\phi} \\ K_{\phi\beta} & K_{\phi\phi} \end{pmatrix},$$

where $K_{\beta\beta} = \phi X^\top W X$, $K_{\beta\phi} = K_{\phi\beta}^\top = X^\top T c$ and $K_{\phi\phi} = \mathrm{tr}(D)$. Here, X is the $n \times p$ matrix of explanatory variables values, $T = \mathrm{diag}\{1/g'(\mu_1), \ldots, 1/g'(\mu_n)\}$, and $W = \mathrm{diag}\{w_1, \ldots, w_n\}$, with

$$w_i = \phi\{\psi'(\mu_i \phi) + \psi'((1-\mu_i)\phi)\} \frac{1}{g'(\mu_i)^2},$$

$\psi'(\cdot)$ being the trigamma function, i.e., the first derivative of the digamma function. Also, $D = \mathrm{diag}\{d_1, \ldots, d_n\}$, where $d_i = \psi'(\mu_i\phi)\mu_i^2 + \psi'((1-\mu_i)\phi)(1-\mu_i^2) - \psi'(\phi)$, and $c = (c_1, \ldots, c_n)^\top$, where $c_i = \phi\{\psi'(\mu_i\phi)\mu_i - \psi'((1-\mu_i)\phi)(1-\mu_i)\}$. In what follows, we denote the inverse information matrix by

$$K^{-1} = K(\beta, \phi)^{-1} = \begin{pmatrix} K^{\beta\beta} & K^{\beta\phi} \\ K^{\phi\beta} & K^{\phi\phi} \end{pmatrix}.$$

It is noteworthy that, unlike the class of GLMs, the parameters β and ϕ are not orthogonal.

Define $\tilde{X}$ as the $(n+1) \times (p+1)$ matrix given by

$$\tilde{X} = \begin{pmatrix} X & 0 \\ 0 & 1 \end{pmatrix}.$$

Also, let $\tilde{W}$ be the $(n+1) \times (n+1)$ matrix

$$\tilde{W} = \begin{pmatrix} W_{\beta\beta} & W_{\beta\phi} \\ W_{\phi\beta} & W_{\phi\phi} \end{pmatrix},$$

where

$$W_{\beta\beta} = \text{diag}\left\{\left(\phi\frac{d\mu_i}{d\eta_i}\right)^2 w_i\right\}, \quad W_{\beta\phi} = Tc,$$
$$W_{\phi\beta} = W_{\beta\phi}^\top, \quad W_{\phi\phi} = \text{tr}(\text{diag}(d_i)).$$

Here, $w_i = \psi'(\mu_i\phi) + \psi'((1-\mu_i)\phi)$. It can be proved that Fisher's information matrix for the parameter vector $\theta = (\beta^\top, \phi)^\top$ is given by

$$K(\theta) = \tilde{X}^\top \tilde{W} \tilde{X}.$$

For further details on the class of beta regression models, see Cribari-Neto and Zeileis (2010).

Ospina et al. (2006) obtained closed-form expressions for the second-order biases of the MLEs of β and ϕ. They demonstrated that the second-order bias of $\hat{\beta}$ can be expressed as

$$B(\hat{\beta}) = K^{\beta\beta}X^\top[W_1\delta_{\beta\beta} + (W_2 + W_3)XK^{\beta\phi} + \{\text{diagonal}(W_4)^\top\}K^{\phi\phi}]$$
$$+ K^{\beta\phi}[\text{tr}(W_3XK^{\beta\beta}X^\top) + K^{\phi\phi}\text{tr}(S)\{\text{diagonal}(W_4 + W_5)\}XK^{\beta\phi}],$$

where W_1–W_5 and S are defined in the appendix of their paper, diagonal$(\cdot)$ is the row vector formed with the entries in the main diagonal of a square matrix, and $\delta_{\beta\beta}$ is the $n \times 1$ dimensional vector defined by the main diagonal of $XK^{\beta\beta}X^\top$. Furthermore,

$$K^{\beta\beta} = (X^\top W_{\beta\beta}X)^{-1}\left\{I_p + \frac{X^\top Tcc^\top T^\top X(X^\top W_{\beta\beta}X)^{-1}}{\gamma}\right\},$$
$$\gamma = \text{tr}(\text{diag}(d_i)) - c^\top T^\top X(X^\top W_{\beta\beta}X)^{-1}X^\top Tc,$$
$$K^{\beta\phi} = (K^{\phi\beta})^\top = -\frac{1}{\gamma}(X^\top W_{\beta\beta}X)^{-1}X^\top Tc, \; K^{\phi\phi} = \frac{1}{\gamma},$$

where I_p denotes the p-dimensional identity matrix.

Define the $(n+1)$-vector $\tilde{\delta}$ as

$$\tilde{\delta} = \begin{pmatrix} W_1\delta_{\beta\beta} + (W_2 + W_3)XK^{\beta\phi} + \text{diagonal}(W_4)^\top \\ \text{tr}(W_3XK^{\beta\beta}X^\top) + K^{\phi\phi}\text{tr}(S) + \{\text{diagonal}(W_4 + W_5)\}XK^{\beta\phi} \end{pmatrix}.$$

The second-order bias of $\hat{\beta}$ can then be expressed as

$$B(\hat{\beta}) = K^{\beta *}\tilde{X}^\top\tilde{\delta},$$

where $K^{\beta *}$ is the $k \times (k+1)$ upper block of K^{-1}, i.e.,

$$K^{\beta *} = \begin{pmatrix} K^{\beta\beta} & K^{\beta\phi} \end{pmatrix}.$$

The authors have also shown that

$$B(\hat{\phi}) = K^{\phi\beta}X^\top\left[W_1\delta_{\beta\beta} + (W_2 + W_3)XK^{\beta\phi} + \{\text{diagonal}(W_4)^\top\}K^{\phi\phi}\right] + K^{\phi\phi}\left[\text{tr}(W_3XK^{\beta\beta}X^\top) + K^{\phi\phi}\text{tr}(S) + \{\text{diagonal}(W_4 + W_5)\}XK^{\beta\phi}\right].$$

Then, considering the $1 \times (p+1)$ lower block of the matrix $K(\theta)^{-1}$ given by

$$K^{\phi *} = \left(K^{\phi\beta}\ K^{\phi\phi}\right),$$

they wrote the second-order bias of $\hat{\phi}$ as

$$B(\hat{\phi}) = K^{\phi *}\tilde{X}^\top\tilde{\delta}.$$

Thus, the second-order bias of the MLE of the joint vector $\theta = (\beta^\top, \phi)^\top$ is

$$B(\hat{\theta}) = K(\theta)^{-1}\tilde{X}^\top\tilde{\delta} = (\tilde{X}^\top\tilde{W}\tilde{X})^{-1}\tilde{X}^\top\tilde{\delta}. \tag{4.21}$$

Defining $\tilde{\xi} = \tilde{W}^{-1}\tilde{\delta}$, the expression in (4.21) becomes

$$B(\hat{\theta}) = (\tilde{X}^\top\tilde{W}\tilde{X})^{-1}\tilde{X}^\top\tilde{W}\tilde{\xi}.$$

Therefore, the components of $B(\hat{\theta})$ can be estimated through a weighted linear regression.

The Monte Carlo evidence presented by Ospina et al. (2006) showed that the MLEs of $\beta_1, \ldots, \beta_p$ (the parameters in the linear predictor) are nearly unbiased in small samples, unlike the MLE of ϕ, which is considerably biased when the sample size is small. The precision parameter bias-corrected MLE, which was obtained by subtracting the estimated second-order bias from the MLE, displayed almost no bias even when the sample contained as few as 20 observations.

Finally, we note that the beta regression model can be extended to a more general setting considering non-constant precision, where the precision parameter is allowed to vary across observations. The model then consists of two (regression) submodels, namely a mean submodel and a precision submodel. The results in Ospina et al. (2006) were extended to cover non-constant precision by Simas et al. (2010).

4.12 An Alternative Analytical Bias Correction

The bias of the MLE of θ is $B(\theta) = \mathbb{E}(\hat{\theta}) - \theta$, depends on the true value of the parameter, θ, and can be expanded asymptotically as

$$B(\theta) = B_1(\theta)/n + B_2(\theta)/n^2 + \cdots . \tag{4.22}$$

As noted in the previous sections, one can obtain a second-order bias-corrected estimator by calculating the term $B_1(\theta)$ and then plugging the estimated parameter into this bias term, yielding the bias-corrected estimator

$$\hat{\theta}_{BC} = \hat{\theta} - B_1(\hat{\theta})/n. \tag{4.23}$$

This is a 'corrective' approach in the sense that one first obtains the MLE and then bias-correct it. An alternative approach consists of transforming the score function so that the resulting estimator will be unbiased to second order. This is a 'preventive' approach in the sense that one corrects the score function and not the MLE. It was introduced by Firth (1993).

The idea behind the preventive bias correction is the following. The expected value of the score function, $U_\theta = \partial \ell / \partial \theta$, evaluated at the true parameter value, θ,is zero, that is,

$$\mathbb{E}[(U_\theta)_r] = \int_{-\infty}^{+\infty} \frac{\partial f(x)}{\partial \theta_r} \frac{1}{f(x)} f(x)dx = \frac{\partial}{\partial \theta_r} \int_{-\infty}^{+\infty} f(x)dx = 0, \ \ \forall r \in \{1, \ldots, p\},$$

when θ is a p-component vector. However, the score is generally not linear in θ, and hence, when one calculates the MLE by equating the value of the score to zero, a bias typically arises. Through a simple geometrical argument, Firth suggests the score be shifted by $-K(\theta)B(\theta)$, where $K(\theta)$ denotes Fisher's information. A second-order bias-corrected estimator can then be obtained as the solution to the equation

$$U^*(\theta) = U(\theta) - K(\theta)B(\theta).$$

Notice that here one does not compute the MLE and then applies a bias correction to it. Instead, one modifies the score function, sets it equal to zero, and then solves for θ.

Firth (1993) showed that in exponential families with canonical parameterizations, his correction scheme consists in penalizing the likelihood by the Jeffreys invariant prior. His corrected estimator can then be obtained by numerically maximizing the modified log-likelihood function.

4.13 Bootstrap Bias Corrections

A different strategy for bias-correcting parameter estimators uses the bootstrap method pionered by Efron (1979). The main idea is to use data resampling in order to estimate the bias function. Let $Y = (Y_1, \ldots, Y_n)^\top$ be a set of independent and identically distributed random variables, each Y_i having the distribution function $F = F_\theta(y)$, where θ is the parameter that indexes the distribution and is viewed as a

functional of F, i.e., $\theta = t(F)$. Y is our original sample. Let $\hat{\theta}$ be an estimator of θ based on Y which we write as $\hat{\theta} = s(Y)$. We proceed as follows. We obtain, from the original sample Y, a large number (say, R) of pseudo-samples $Y^* = (Y_1^*, \ldots, Y_n^*)^\top$ and then use such artificial samples to improve the statistical inference on θ. We can obtain the bootstrap samples parametrically or nonparametrically. In the *parametric bootstrap*, we sample from $F = F_{\hat{\theta}}$, i.e., we sample from the model distribution function after replacing the unknown parameters by the corresponding MLEs. In the *nonparametric bootstrap*, we sample from the empirical distribution function $\hat{F}$. We can do so by sampling from the data $Y = (Y_1, \ldots, Y_n)^\top$ *with replacement*. Note that the nonparametric bootstrap does not entail parametric assumptions. The bootstrap samples and all statistics computed from them shall be denoted using '$*$'. Using each artificial sample y^*, we estimate θ, thus obtaining $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$. Next, we use these bootstrap parameter estimates to construct an estimate of the bias function. The bias of $\hat{\theta}$ is $B(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$. Notice that we can denote the bias of the estimator $\hat{\theta} = s(Y)$ by $B_F(\hat{\theta}, \theta)$, i.e.,

$$B_F(\hat{\theta}, \theta) = \mathbb{E}_F[\hat{\theta} - \theta] = \mathbb{E}_F[s(Y)] - t(F),$$

where the subscript F indicates that expectation is taken with respect to F. The parametric and nonparametric estimates of the bias are given, respectively, by

$$B_{F_{\hat{\theta}}}(\hat{\theta}, \theta) = \mathbb{E}_{F_{\hat{\theta}}}[s(Y)] - t(F_{\hat{\theta}}) \quad \text{and} \quad B_{\hat{F}}(\hat{\theta}, \theta) = \mathbb{E}_{\hat{F}}[s(Y)] - t(\hat{F}).$$

An alternative bootstrap bias estimator was introduced by Efron (1990). It is carried out nonparametrically and uses an auxiliary $(n \times 1)$ *resampling vector*, whose elements are the proportions of observations in the original sample $Y = (Y_1, \ldots, Y_n)^\top$ that were included in the bootstrap sample. Let $P^* = (P_1^*, P_2^*, \ldots, P_n^*)$ be the resampling vector. Its jth element ($j = 1, 2, \ldots, n$), P_j^*, is defined with respect to a given bootstrap sample $Y^* = (Y_1^*, \ldots, Y_n^*)^\top$ as $P_j^* = n^{-1}\big(\#\{Y_k^* = Y_j\}\big)$. It is important to note that the vector $P^0 = (1/n, 1/n, \ldots, 1/n)$ corresponds to the original sample. It should also be noted that any bootstrap replicate $\hat{\theta}^*$ can be defined as a function of the resampling vector. For example, if $\hat{\theta} = s(Y) = \overline{Y} = n^{-1}\sum_{i=1}^n Y_i$, then

$$\begin{aligned}\hat{\theta}^* &= \frac{Y_1^* + Y_2^* + \cdots + Y_n^*}{n} = \frac{\#\{Y_k^* = Y_1\}y_1 + \cdots + \#\{Y_k^* = Y_n\}Y_n}{n} \\ &= \frac{(nP_1^*)Y_1 + \cdots + (nP_n^*)Y_n}{n} = P^*Y.\end{aligned}$$

Suppose we can write the estimate of interest, obtained from the original sample Y, as $G(P^0)$. It is now possible to obtain bootstrap estimates $\hat{\theta}^{*b}$ using the resampling vectors P^{*b}, $b = 1, 2, \ldots, R$, as $G(P^{*b})$. Efron's (1990) bootstrap bias estimator, $\bar{B}_{\hat{F}}(\hat{\theta}, \theta)$, is defined as

$$\bar{B}_{\hat{F}}(\hat{\theta}, \theta) = \hat{\theta}^{*(\cdot)} - G(P^{*(\cdot)}), \quad \text{where} \quad P^{*(\cdot)} = \frac{1}{R}\sum_{b=1}^{R} P^{*b},$$

which differs from $\hat{B}_{\hat{F}}(\hat{\theta}, \theta)$, since $\hat{B}_{\hat{F}}(\hat{\theta}, \theta) = \hat{\theta}^{*(\cdot)} - G(P^0)$. Notice that this bias estimator uses an additional information, namely the proportions of the n observations that were selected in each nonparametric resampling.

After obtaining an estimator for the bias, it is easy to obtain a bias-adjusted estimator. Using the three bootstrap bias estimators presented above, we can define the estimators adjusted for bias as

$$\begin{aligned}
\tilde{\theta}_1 &= s(y) - \hat{B}_{\hat{F}}(\hat{\theta}, \theta) = 2\hat{\theta} - \hat{\theta}^{*(\cdot)}, \\
\tilde{\theta}_2 &= s(y) - \bar{B}_{\hat{F}}(\hat{\theta}, \theta) = \hat{\theta} - \hat{\theta}^{*(\cdot)} + G(P^{*(\cdot)}), \\
\tilde{\theta}_3 &= s(y) - \hat{B}_{F_{\hat{\theta}}}(\hat{\theta}, \theta) = 2\hat{\theta} - \hat{\theta}^{*(\cdot)}.
\end{aligned}$$

The modified estimates $\tilde{\theta}_1$ and $\tilde{\theta}_3$ are said to be constant-bias-correcting (CBC) estimates, see MacKinnon and Smith (1998).

It is important to note that the bias estimation procedure proposed by Efron (1990) requires the estimator $\hat{\theta}$ to have closed form. However, oftentimes the MLE of θ, the parameter that indexes the model used to represent the population, does not have a closed form. Rather, it needs to be obtained by numerically maximizing the log-likelihood function using a nonlinear optimization algorithm, such as a Newton or quasi-Newton algorithm. Cribari-Neto et al. (2002) proposed an adaptation of Efron's method that may be used with estimators that cannot be written in closed form. The authors used the resampling vector to modify the log-likelihood function and then maximize the modified log-likelihood. The main idea is to write the log-likelihood function in terms of P^0, replace this vector by $P^{*(\cdot)}$, and then maximize the resulting (modified) log-likelihood function. The maximizer of such a function is a bias-corrected MLE. It is noteworthy that this bootstrapping scheme only entails one nonlinear optimization, i.e., only one log-likelihood maximization is carried out. This occurs because the bootstrapping scheme is performed in order to obtain a vector that is used to modify the log-likelihood function which is then maximized. As a consequence, this resampling scheme is not as computationally intensive as alternative schemes in which a nonlinear optimization must be performed in each bootstrap replication. Cribari-Neto et al. (2002) used this bootstrapping scheme to bias-correcting the MLEs of the parameters that index a model used for SAR (synthetic aperture radar) image processing. Their simulation results showed that the bias-reduced estimator obtained from the maximization of the modified log-likelihood function using $P^{*(\cdot)}$ outperfomed other bias-corrected estimators. This approach was also considered by Lemonte et al. (2008) to reduce the biases of the MLEs of the two parameters that index the BS model.

According to MacKinnon and Smith (1998), estimators $\tilde{\theta}_1$ and $\tilde{\theta}_3$, which the authors call CBC, can be expected to work well whenever the bias function $B(\theta)$

is flat, i.e., when it is not a function of θ. They note that a different scheme can be developed for situations in which $B(\theta)$ is a linear function of θ, i.e.,

$$B(\theta) = a + c\theta.$$

It is clear that two constants must be estimated—a and b—using information available in the bootstrap samples. By doing so, one can estimate the bias of $\hat{\theta}$ and then obtain a bias-adjusted estimator.

At the outset, as in the previous bootstrapping schemes, we compute the estimate $\hat{\theta} = s(Y)$ (using the original sample Y). We then proceed to obtain point estimates for a and b. This is accomplished by estimating the bias function at two different points. First, we use a parametric bootstrapping scheme to obtain a bootstrap estimate for the bias of $\hat{\theta}$, which we denote by $\hat{B}$. This bias estimate is computed as $\hat{\theta}^{*(\cdot)} - \hat{\theta}$. Next, we use a second parametric bootstrapping scheme based on $\tilde{\theta}$, where $\tilde{\theta} = 2\hat{\theta} - \hat{\theta}^{*(\cdot)}$. Here, for each bootstrap sample, we compute $\hat{\theta}^{*b}_{F_{\tilde{\theta}}}$, for $b = 1, \ldots, R$. Therefore, we estimate the bias of $\tilde{\theta}$ as $\tilde{B} = \hat{\theta}^{*(\cdot)}_{F_{\tilde{\theta}}} - \tilde{\theta}$, where $\hat{\theta}^{*(\cdot)}_{F_{\tilde{\theta}}}$ is the average over all bootstrap replications of $\tilde{\theta}$. Notice that here one needs to perform $2R$ bootstrap replications, the double the number of bootstrap replications in the previous schemes. Finally, using the point estimates, $\hat{\theta}$ and $\tilde{\theta}$, and their respective estimated biases, $\hat{B}$ and $\tilde{B}$, we arrive at the system of two simultaneous equations

$$\hat{B} = \breve{a} + \breve{c}\hat{\theta} \quad \text{and} \quad \tilde{B} = \breve{a} + \breve{c}\tilde{\theta}.$$

The solution of this two equation system is

$$\breve{a} = \hat{B} - \frac{\hat{B} - \tilde{B}}{\hat{\theta} - \tilde{\theta}} \quad \text{and} \quad \breve{c} = \frac{\hat{B} - \tilde{B}}{\hat{\theta} - \tilde{\theta}}.$$

It is now straightforward to obtain the linear bias-correcting (LBC) estimator, say $\tilde{\theta}_4$ (MacKinnon and Smith 1998):

$$\tilde{\theta}_4 = \frac{1}{1 + \breve{c}}(\hat{\theta} - \breve{a}).$$

It is noteworthy that the variance of $\hat{\theta}_4$ is

$$\mathrm{Var}(\tilde{\theta}_4) = \frac{1}{(1 + \breve{c})^2}\mathrm{Var}(\hat{\theta}).$$

We thus conclude that the variance of $\tilde{\theta}_4$ will exceed that of $\hat{\theta}$ whenever $\breve{c}$ belongs to $\mathscr{A} = \big\{(-2, 0) \setminus \{-1\}\big\}$.

References

Bartlett, M. S. (1953). Confidence intervals II. *Biometrika*, *40*, 306–317.

Botter, D. A., & Cordeiro, G. M. (1998). Improved estimators for generalized linear models with dispersion covariates. *Journal of Statistical Computation and Simulation*, *62*, 91–104.

Chesher, A., & Jewitt, I. (1987). The bias of a heteroskedasticity consistent covariance matrix estimator. *Econometrica*, *55*, 1217–1222.

Cook, D. R., Tsai, C. L., & Wei, B. C. (1986). Bias in nonlinear regression. *Biometrika*, *73*, 615–623.

Cordeiro, G. M., & McCullagh, P. (1991). Bias correction in generalized linear models. *Journal of the Royal Statistical Society B*, *53*, 629–643.

Cordeiro, G. M., & Klein, R. (1994). Bias correction in ARMA models. *Statistics and Probability Letters*, *19*, 169–176.

Cordeiro, G. M., Rocha, E. C., Rocha, J. G. C., & Cribari-Neto, F. (1997). Bias corrected maximum likelihood estimation for the beta distribution. *Journal of Statistical Computation and Simulation*, *58*, 21–35.

Cordeiro, G. M., & Vasconcellos, K. L. P. (1997). Bias correction for a class of multivariate nonlinear regression models. *Statistics and Probability Letters*, *35*, 155–164.

Cordeiro, G. M., Ferrari, S. L. P., Uribe-Opazo, M. A., & Vasconcellos, K. L. P. (2000). Corrected maximum-likelihood estimation in a class of symmetric nonlinear regression models. *Statistics and Probability Letters*, *46*, 317–328.

Cox, D. R., & Snell, E. J. (1968). A general definition of residuals (with discussion). *Journal of the Royal Statistical Society B*, *30*, 248–275.

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical Statistics*. London: Chapman and Hall.

Cox, D. R., & Reid, N. (1987). Approximations to noncentral distributions. *Canadian Journal of Statistics*, *15*, 105–114.

Cribari-Neto, F., Ferrari, S. L. P., & Cordeiro, G. M. (2000). Improved heteroscedasticity-consistent covariance matrix estimators. *Biometrika*, *87*, 907–918.

Cribari-Neto, F., Frery, A. C., & Silva, M. F. (2002). Improved estimation of clutter properties in speckled imagery. *Computational Statistics and Data Analysis*, *40*, 801–824.

Cribari-Neto, F., & Vasconcellos, K. L. P. (2002). Nearly unbiased maximum likelihood estimation for the beta distribution. *Journal of Statistical Computation and Simulation*, *72*, 107–118.

Cribari-Neto, F., & Zeileis, A. (2010). Beta regressions in R. *Journal of Statistical Software*, *34*(2), 1–24.

Cribari-Neto, F., & Lima, M. G. A. (2011). A sequence of improved standard errors under heteroskedasticity of unknown form. *Journal of Statistical Planning and Inference*, *141*, 3617–3627.

Development Core Team, R. (2006). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Doornik, J. A. (2009). *An Object-Oriented Matrix Language Ox 6*. London: Timberlake Consultants Press.

Efron, B. (1979). Bootstrapping methods: Another look at the jackknife. *Annals of Statistics*, *7*, 1–26.

Efron, B. (1990). More efficient bootstrap computations. *Journal of the American Statistical Association*, *85*, 79–89.

Ferrari, S. L. P., & Cribari-Neto, F. (2004). Beta regression for modeling rates and proportions. *Journal of Applied Statistics*, *31*, 799–815.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, *80*, 27–38.

Giles, D. E. A. (2012). Bias reduction for the maximum likelihood estimator of the parameters in the half-logistic distribution. *Communications in Statistics, Theory and Methods*, *41*, 212–222.

Galea, M., Leiva, V., & Paula, G. A. (2004). Influence diagnostics in log-Birnbaum-Saunders regression models. *Journal of Applied Statistics*, *31*, 1049–1064.

Giles, D. E. A., & Feng, H. (2011). Reducing the bias of the maximum likelihood estimator for the Poisson regression model. *Economics Bulletin*, *31*, 2933–2943.

Giles, D. E. A., Feng, H., & Godwin, R. T. (2013). On the bias of the maximum likelihood estimator for the two-parameter Lomax distribution. *Communications in Statistics, Theory and Methods*, *42*, 1934–1950.

Lemonte, A. J., & Cordeiro, G. M. (2009). Birnbaum-Saunders nonlinear regression models. *Computational Statistics and Data Analysis*, *53*, 4441–4452.

Lemonte, A. J., & Cordeiro, G. M. (2010). Asymptotic skewness in Birnbaum-Saunders nonlinear regression models. *Statistics and Probability Letters*, *80*, 892–898.

Lemonte, A. J., Cribari-Neto, F., & Vasconcellos, K. L. P. (2007). Improved statistical inference for the two-parameter Birnbaum-Saunders distribution. *Computational Statistics and Data Analysis*, *51*, 4656–4681.

Lemonte, A. J., Simas, A. B., & Cribari-Neto, F. (2008). Bootstrap-based improved estimators for the two-parameter Birnbaum-Saunders distribution. *Journal of Statistical Computation and Simulation*, *78*, 37–49.

MacKinnon, J. G., & Smith, J. A. A. (1998). Approximate bias correction in econometrics. *Journal of Econometrics*, *85*, 205–230.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models* (2nd ed.). London: Chapman and Hall.

Ospina, R., Cribari-Neto, F., & Vasconcellos, K. L. P. (2006). Improved point and interval estimation for a beta regression model. *Computational Statistics and Data Analysis*, *51*, 960–981 (Errata: vol. 55, p. 2445, 2011).

Patriota, A. G., & Lemonte, A. J. (2009). Bias correction in a multivariate regression model with general parameterization. *Statistics and Probability Letters*, *79*, 1655–1662.

Ratkowsky, D. A. (1983). *Nonlinear Regression Modeling: A Unified Practical Approach*. New York: Marcel Dekker.

Rieck, J.R. (1989). Statistical analysis for the Birnbaum-Saunders fatigue life distribution. Ph.D. dissertation, Clemson University.

Rieck, J. R., & Nedelman, J. R. (1991). A log-linear model for the Birnbaum-Saunders distribution. *Technometrics*, *33*, 51–60.

Simas, A. B., Barreto-Souza, W., & Rocha, A. V. (2010). Improved estimators for a general class of beta regression models. *Computational Statistics and Data Analysis*, *54*, 348–366.

Stosic, B., & Cordeiro, G. M. (2009). Using Maple and Mathematica to derive bias corrections for two parameter distributions. *Journal of Statistical Computation and Simulation*, *75*, 409–423.

Vasconcellos, K. L. P., & Silva, S. G. (2005). Corrected estimates for Student *t* regression models with unknown degrees of freedom. *Journal of Statistical Computation and Simulation*, *79*, 751–767.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*, 817–838.

Young, D. H., & Bakir, S. T. (1987). Bias correction for a generalized log-gamma regression model. *Technometrics*, *29*, 183–191.

Appendix A
Supplementary Material

A.1 Bartett-Type Correction

Using the notation in Sect. 3.2, the general expressions for A_1, A_2, and A_3 are

$$\begin{aligned}
A_1 &= 3\sum{}'(\kappa_{ijk}+2\kappa_{i,jk})(\kappa_{rst}+2\kappa_{rs,t})\,a_{ij}\,a_{st}\,m_{kr}\\
&\quad-6\sum{}'(\kappa_{ijk}+2\kappa_{i,jk})\,\kappa_{r,s,t}\,a_{ij}\,a_{kt}\,m_{st}\\
&\quad+6\sum{}'(\kappa_{i,jk}-2\kappa_{i,j,k})\,(\kappa_{rst}+2\kappa_{rs,t})\,a_{js}\,a_{kt}\,m_{ir}\\
&\quad-6\sum{}'(\kappa_{i,j,k,r}+\kappa_{i,j,kr})\,a_{kr}\,m_{ij},\\
A_2 &= 3\sum{}'\kappa_{i,j,k}\kappa_{r,s,t}a_{kr}m_{ij}m_{st}\\
&\quad+6\sum{}'(\kappa_{ijk}+2\kappa_{i,jk})\,\kappa_{r,s,t}\,a_{ij}\,m_{kr}\,m_{st}\\
&\quad-6\sum{}'\kappa_{i,j,k}\,\kappa_{r,s,t}\,a_{kt}\,m_{ir}\,m_{js}\\
&\quad+3\sum{}'\kappa_{i,j,k,r}\,m_{ij}\,m_{kr},\\
A_3 &= 3\sum{}'\kappa_{i,j,k}\,\kappa_{r,s,t}\,m_{ij}\,m_{kr}\,m_{st}\\
&\quad+2\sum{}'\kappa_{i,j,k}\,\kappa_{r,s,t}\,m_{ir}\,m_{js}\,m_{kt}.
\end{aligned}$$

Here, a_{ij} and m_{ij} are the (i,j)th elements of the matrices A and M, respectively, and $\sum'$ denotes the sum over the specified components. The reader is referred to Harris (1985) for further details.

Now, we consider the expansion (3.1). In order to provide an alternative general formula for A_1, we adopt the same notation of Sect. 2.2. Let

G. M. Cordeiro and F. Cribari-Neto, *An Introduction to Bartlett Correction and Bias Reduction*, SpringerBriefs in Statistics, DOI: 10.1007/978-3-642-55255-7,

$$\kappa_{ij}^{(k)} = \frac{\partial \kappa_{ij}}{\partial \theta_k}, \quad \kappa_{ij}^{(kr)} = \frac{\partial^2 \kappa_{ij}}{\partial \theta_k \partial \theta_r}, \quad \kappa_{ijk}^{(r)} = \frac{\partial \kappa_{ijk}}{\partial \theta_r}.$$

Then, we can write $A_1 = 12(\varepsilon_p - \varepsilon_{p-q})$, where $\varepsilon_p = \sum{}'(\lambda_{ijkr} - \lambda_{ijkrst})$ and the quantities λ_{ijkr} and λ_{ijkrst} are defined in Eqs. (2.5) and (2.6). For calculating ε_p, the summations in $\sum{}'$ are over all components of θ. The quantity ε_{p-q} is defined analogously, but the summations run only from $q+1$ to p. For further details, see Cordeiro (1993) and Lawley (1956).

A.2 Bartlett-Type Corrections for Seven Distributions

(1) Maxwell distribution

```
f = Sqrt[2/Pi]*y^2*Exp[-y^2/(2*\[Theta]^2)]/\[Theta]^3
p = 0; q = \[Infinity]; cond := \[Theta] > 0;
corrections[f, p, q, cond, 1];
```

$$LR1 = -\frac{1}{9},$$
$$S1 = \frac{2}{27}, \; S2 = -\frac{11}{27}, \; S3 = \frac{1}{9},$$
$$W1 = \frac{1}{9}, \; W2 = -\frac{2}{27}, \; W3 = \frac{25}{216},$$
$$MW2 = -\frac{2}{27}, \; MW3 = \frac{1}{216}.$$

(2) Rayleigh distribution

```
f = y*Exp[-y^2/(2*\[Theta]^2)]/\[Theta]^2
p = 0; q = \[Infinity]; cond := \[Theta] > 0;
corrections[f, p, q, cond, 1];
```

$$LR1 = -\frac{1}{6},$$
$$S1 = \frac{1}{9}, \; S2 = -\frac{11}{18}, \; S3 = \frac{1}{6},$$
$$W1 = \frac{1}{6}, \; W2 = -\frac{1}{9}, \; W3 = \frac{25}{144},$$
$$MW2 = -\frac{1}{9}, \; MW3 = \frac{1}{144}.$$

(3) Cauchy distribution

```
f = \[Theta]/(Pi*(y^2 + \[Theta]^2))
p = -\[Infinity]; q = \[Infinity]; cond := \[Theta] > 0;
corrections[f, p, q, cond, 1];
```

$$
\begin{aligned}
&LR1 = -\frac{1}{4},\\
&S1 = 0,\ S2 = -\frac{1}{8},\ S3 = \frac{3}{8},\\
&W1 = \frac{7}{8},\ W2 = -\frac{17}{8},\ W3 = \frac{1}{2},\\
&MW2 = -\frac{19}{24},\ MW3 = \frac{1}{2}.
\end{aligned}
$$

(4) Chi-squared distribution

```
f = y^(\[Theta] - 1)*Exp[-y/2]/Gamma[\[Theta]]/2^(\[Theta])
p = 0; q = \[Infinity]; cond := \[Theta] > 0;
corrections[f, p, q, cond, 1];
```

$$
\begin{aligned}
LR1 &= \frac{3\psi^{(1)}(\theta)\,\psi^{(3)}(\theta) - 5\psi^{(2)}(\theta)^2}{12\psi^{(1)}(\theta)^3},\\
S1 &= \frac{\psi^{(2)}(\theta)^2}{36\psi^{(1)}(\theta)^3},\\
S2 &= \frac{3\psi^{(1)}(\theta)\,\psi^{(3)}(\theta) - 10\psi^{(2)}(\theta)^2}{36\psi^{(1)}(\theta)^3},\\
S3 &= \frac{5\psi^{(2)}(\theta)^2 - 3\psi^{(1)}(\theta)\,\psi^{(3)}(\theta)}{12\psi^{(1)}(\theta)^3},\\
W1 &= \frac{5\psi^{(2)}(\theta)^2 - 3\psi^{(1)}(\theta)\,\psi^{(3)}(\theta)}{12\psi^{(1)}(\theta)^3},\\
W2 &= \frac{9\psi^{(1)}(\theta)\,\psi^{(3)}(\theta) - 19\psi^{(2)}(\theta)^2}{36\psi^{(1)}(\theta)^3},\\
W3 &= \frac{\psi^{(2)}(\theta)^2}{36\psi^{(1)}(\theta)^3},\\
MW2 &= \frac{8\psi^{(2)}(\theta)^2 - 9\psi^{(1)}(\theta)\,\psi^{(3)}(\theta)}{36\psi^{(1)}(\theta)^3},\\
MW3 &= \frac{\psi^{(2)}(\theta)^2}{9\psi^{(1)}(\theta)^3}.
\end{aligned}
$$

Table A.1 Coefficients of the corrected statistics for the Student's t distribution

θ	0.5	1.0	1.5	2.0	2.5	3.0
LR1	−2.41879	−7.16703	−15.6808	−29.4776	−50.3429	−80.3236
S1	0.246807	0.407771	0.578902	0.751495	0.920521	1.08322
S2	−2.62419	−6.39687	−12.7119	−22.5462	−37.0711	−57.6417
S3	4.17048	13.074	29.452	56.3661	97.4056	156.677
W1	7.6835	24.9021	57.0206	110.191	191.611	309.504
W2	−4.62959	−13.2499	−28.6408	−53.3815	−90.4808	−143.37
W3	0.246033	0.477164	0.80764	1.24388	1.79162	2.45566
MW2	1.89066	8.09544	20.2045	40.794	72.8688	119.858
MW3	0.98568	1.76715	2.75409	3.92904	5.28059	6.80078

(5) Student's t distribution

```
f = Gamma[(\[Theta] + 1)/2]*(1 + y^2/\[Theta])^
    (-(\[Theta] + 1)/2)/Sqrt[Pi*\[Theta]]/Gamma[\[Theta]/2]
p = -\[Infinity]; q = \[Infinity]; cond := \[Theta] > 0;
corrections[f, p, q, cond, 1];
```

Since the resulting formulas are too cumbersome to be reported here, the numerical values of the coefficients of the corrected statistics for the Student's t distribution for some values of θ are given in Table A.1. The full expressions can be obtained from the authors upon request.

(6) Binomial distribution

```
f = Binomial[m, y]*\[Theta]^y*(1 - \[Theta])^(m - y)
p = 0; q = m; cond := \[Theta] > 0 && \[Theta] < 1 && m > 0
    && m \[Element] Integers && y \[Element] Integers;
corrections[f, p, q, cond, 0];
```

$$LR1 = -\frac{\theta^2 - \theta + 1}{6m\theta(1-\theta)},$$

$$S1 = \frac{4\theta^2 - 4\theta + 1}{36m\theta(1-\theta)},\ S2 = -\frac{22\theta^2 - 22\theta + 7}{36m\theta(1-\theta)},\ S3 = \frac{\theta^2 - \theta + 1}{6m\theta(1-\theta)},$$

$$W1 = \frac{\theta^2 - \theta + 1}{6m\theta(1-\theta)},\ W2 = \frac{-11\theta^2 + 11\theta + 1}{18m\theta(1-\theta)},\ W3 = \frac{4\theta^2 - 4\theta + 1}{9m\theta(1-\theta)},$$

$$MW2 = -\frac{22\theta^2 - 22\theta + 7}{36m\theta(1-\theta)},\ MW3 = \frac{4\theta^2 - 4\theta + 1}{36m\theta(1-\theta)}.$$

(7) Poisson's distribution

```
f = Exp[-\[Theta]]*\[Theta]^y/y!
p = 0; q = \[Infinity]; cond := \[Theta] > 0 &&
    y \[Element] Integers && y >= 0;
corrections[f, p, q, cond, 0];
```

$$
\begin{aligned}
LR1 &= -\frac{1}{6\theta}, \\
S1 &= \frac{1}{36\theta},\ S2 = -\frac{7}{36\theta},\ S3 = \frac{1}{6\theta}, \\
W1 &= \frac{1}{6\theta},\ W2 = \frac{1}{18\theta},\ W3 = \frac{1}{9\theta}, \\
MW2 &= -\frac{7}{36\theta},\ MW3 = \frac{1}{36\theta}.
\end{aligned}
$$

References

Cordeiro, G. M. (1993). General matrix formula for computing Bartlett corrections. *Statistics and Probability Letters*, *16*, 11–18.

Harris, P. (1985). An asymptotic expansion for the null distribution of the efficient score statistic. *Biometrika*, *72*, 653–659 (Erratum in vol. 74, p. 667).

Lawley, D. N. (1956). A general method for approximating to the distribution of the likelihood ratio criteria. *Biometrika*, *71*, 233–244.

Glossary

Bias The difference between the expected value of an estimator and the true parameter value.
Bias correction Method for removing the bias of an estimator, usually up to some order of accuracy.
Bootstrap Resampling method proposed by Bradley Efron that can be used, e.g., for bias correction, interval estimation, and hypothesis testing inference.
Heteroskedasticity Non-constant response variances in regression models.
Homoskedasticity Constant response variances in regression models.
Least-squares estimator Estimator obtained from the minimization of the sum of squared errors.
Maximum likelihood estimator Estimator obtained from the maximization of a likelihood function.
Quasi-t test Test similar to the usual t test in the linear regression model, but whose test statistic uses a heteroskedasticity-consistent standard error; the test is performed using standard normal critical values.

G. M. Cordeiro and F. Cribari-Neto, *An Introduction to Bartlett Correction and Bias Reduction*, SpringerBriefs in Statistics, DOI: 10.1007/978-3-642-55255-7,