DE GRUYTER

*Sergii Masiuk, Alexander Kukush, Sergiy Shklyar, Mykola Chepurny, Illya Likhtarov*

# RADIATION RISK ESTIMATION

## BASED ON MEASUREMENT ERROR MODELS

SERIES IN MATHEMATICS AND LIFE SCIENCES

DE G

S. V. Masiuk, A. G. Kukush, S. V. Shklyar, M. I. Chepurny, I. A. Likhtarov†
**Radiation Risk Estimation**

# De Gruyter Series in Mathematics and Life Sciences

Edited by
Alexandra V. Antoniouk, Kyiv, Ukraine
Roderick V. Nicolas Melnik, Waterloo, Ontario, Canada

## Volume 5

S. V. Masiuk, A. G. Kukush, S. V. Shklyar,
M. I. Chepurny, I. A. Likhtarov†

# Radiation Risk Estimation

Based on Measurement Error Models

DE GRUYTER

**Mathematics Subject Classification 2010**
Primary: 62P10; Secondary: 62J12

**Authors**
Dr. Sergii Masiuk
Ukrainian Radiation Protection Institute
National Research Center for Radiation Medicine
National Academy of Medical Sciences of Ukraine
Melnykova Street, 53
Kiev 04050, Ukraine
masja1979@gmail.com

Prof. Dr. Alexander Kukush
Taras Shevchenko National University of Kyiv
Faculty of Mechanics and Mathematics
Volodymyrska Street, 64
Kiev 01033, Ukraine
alexander_kukush@univ.kiev.ua

Dr. Sergiy Shklyar
Taras Shevchenko National University
Faculty of Mechanics and Mathematics
Volodymyrska Street, 64
Kiev 01033, Ukraine
shklyar@univ.kiev.ua

Mykola Chepurny
National Research Center for Radiation Medicine
National Academy of Medical Sciences of Ukraine
Melnykova Street, 53
Kiev 04050, Ukraine
ch@rpi.kiev.ua

Prof. Dr. Illya Likhtarov†
Ukrainian Radiation Protection Institute
National Research Center for Radiation Medicine
National Academy of Medical Sciences of Ukraine
Deceased on January 14, 2017

ISBN 978-3-11-044180-2
e-ISBN (PDF) 978-3-11-043366-1
e-ISBN (EPUB) 978-3-11-043347-0
Set-ISBN 978-3-11-043367-8
ISSN 2195-5530

**Library of Congress Cataloging-in-Publication Data**
A CIP catalog record for this book has been applied for at the Library of Congress.

**Bibliographic information published by the Deutsche Nationalbibliothek**
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de.


Typesetting: le-tex publishing services GmbH, Leipzig
Printing and binding: CPI books GmbH, Leck
♾ Printed on acid-free paper
Printed in Germany

www.degruyter.com

# List of authors

**Masiuk Sergii**
PhD, Head, Laboratory of Statistical Methods, Ukrainian Radiation Protection Institute, Kyiv, Ukraine; Head, Laboratory of Radiological Protection, State Institution “National Research Center for Radiation Medicine of the National Academy of Medical Sciences of Ukraine”, Kyiv, Ukraine.

**Kukush Alexander**
Doctor of Sciences in Physics and Mathematics, Professor, Department of Mathematical Analysis, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine.

**Shklyar Sergiy**
PhD, Senior Scientist, Department of Probability Theory, Statistics and Actuarial Mathematics, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine.

**Chepurny Mykola**
Scientist, Laboratory of Radiological Protection, State Institution “National Research Center for Radiation Medicine of the National Academy of Medical Sciences of Ukraine”, Kyiv, Ukraine.

**Likhtarov Illya**[†]
Doctor of Sciences in Physics and Mathematics, Professor, Head, Department of Dosimetry and Radiation Hygiene, State Institute “National Research Center for Radiation Medicine of the National Academy of Medical Sciences of Ukraine”, Kyiv, Ukraine; Principal Director, Ukrainian Radiation Protection Institute, Kyiv, Ukraine.

Dedicated to the Heavenly Hundred and all those who gave their lives for Ukraine

# Editor's Foreword

Radiation risks and tools for their estimation relate to the most fundamental concepts and methods used in designing a safe system of interaction between humans and nuclear radiation technologies. This holds true for normal operation of civilian and military facilities as well as for emergency situations (Chornobyl, Fukushima, Kyshtym, etc.) and extraordinary events (atomic bombings of Hiroshima and Nagasaki). As a matter of fact, without knowing the quantitative value of radiation risk it is impossible to construct an acceptable system of safety standards for the personnel of industrial facilities employing nuclear radiation technologies and for the population involved to some extent in contact with sources of ionizing radiation.

Quantitative estimation of radiation risks has a long and productive history that undoubtedly deserves a separate monograph. Here, it is worth to dwell on a rather specific feature of the risk estimates that we have available and widely use in modern international and national documents regulating acceptable levels of radiation for an individual and for the human population as a whole. This peculiarity consists in that when analyzing the results of numerous radio-epidemiological studies – the main purpose of which is precisely to determine the radiation risk value – consideration has always been given to a stochastic link between the effects (i.e. the distribution of various radiation-induced pathologies), on the one hand, and the "exact" values of the exposure doses, on the other hand. That is, only the stochastic nature of the effects was taken into account, while ignoring the obvious fact that the "exact" dose values are unknown to us and that they are substituted for by a point statistical parameter (e.g. expectation) of the true dose distribution. It is clear that this results in disregard for errors inevitably arising in the instrumental measurements and in the computations of doses and their components. This particular approach to risk estimation is implemented in the best-known and popular interpretive software package *EPICURE*. Below, this approach as well as the estimates themselves will be referred to as "naive."

As regards risk analysis methods they usually involve mathematical tools that were quite comprehensively developed already in the fundamental works of David Cox, where the naive approach was also employed. Further developments, per se, merely refined the Cox models for various versions of epidemiological studies (ecological, cohort, "case-control" ones). Again, however, their analysis of results was always based on the naive approach; therefore, by definition, the resulting risk estimates were also naive.

It is perfectly obvious that under the naive approach, with its disregard for exposure dose errors, the obtained risk estimates can be distorted; the extent of the distortions, however, is a priori unclear. Naturally, the consequences of the naive approach automatically apply to the bounds of permissible doses and their derivatives. It should be noted that this problem did not go unnoticed; and so in the past twenty years publications began to appear in which attempts were made to take into account the dose

uncertainty in the risk analysis. Unfortunately, the problem has not been completely solved so far.

These are some of the difficulties that are bound to arise once we try to substitute dose distributions for their point estimates in the risk analysis procedure.

- When determining dose estimates, one inevitably has to use the results of different types of measurements, each type involving its own classical or Berkson errors; and so, special statistical procedures need to be developed for obtaining the final dose as a result of an overlap of individual distributions. In this case, the dose distributions are formed due to errors of two types: classical and Berkson. Therefore, proper risk analysis requires separate estimation of the contributions of the classical and Berkson errors to the total dose error.
- At present, more or less established methods for risk estimation in the presence of a mixture of classical and Berkson errors in the exposure doses are still unavailable.

It is clear that the above problems cannot be addressed by using the results of field epidemiological studies with different types of dose error, as there are no such field studies and none can be conducted in principle. The only way is to widely use the so-called stochastic experiment which involves simulation modeling.

Since this concerns the estimation of not just risks, but of risks associated with exposure, all the above-stated problems can be resolved only through the joint efforts of dosimetry physicists and mathematical statisticians. That is why the team of authors of this monograph is made up of experts in the aforementioned fields of science. At the same time, the material in this book is radically focused on mathematical problems of estimation of radiation risks in the presence of errors in exposure doses, while the error level estimation methods (which definitely deserve a separate monograph) are presented in a shorthand form.

Finally, it should be emphasized that, based on the results obtained in the book, a software product similar to the above-mentioned *EPICURE* is worth creating, provided that it includes risk estimation procedures having regard for dose errors. In that case, experts engaged in epidemiological data processing would have a convenient tool for obtaining not only naive risk estimates, but also estimates taking into account the classical and Berkson errors in covariate.

The material and results presented in this monograph will be useful to epidemiologists, dosimetrists, experts engaged in statistical processing of data or working in the field of modern methods of Mathematical Statistics, as well as to undergraduate and graduate university students.

Doctor of Sciences in Physics and Mathematics, Professor I. A. Likhtarov†

# Preface

As a result of the Chornobyl accident in 1986, most of the territories of Ukraine, Belarus, and Russia were radio-contaminated and the residents of these areas underwent radioactive exposure. The most affected by the radiation was the thyroid, due to the intake of iodine radioisotopes, primarily $^{131}$I (Likhtarov et al., 1993a, 2005, 2006b).

As early as 5–6 years after the accident, a sharp increase in thyroid cancer incidence among children and adolescents residing in areas with a rather high radiation exposure of this organ was revealed (Likhtarov et al., 1995a; Buglova et al., 1996; Jacob et al., 2006). In fact, the increase in thyroid cancer incidence among children and adolescents due to internal thyroid irradiation resulting from Chornobyl's radioactive emissions was the main statistically significant long-term effect of the Chornobyl accident. Not surprisingly, this phenomenon generated great interest among radio-epidemiologists around the world and led to a series of epidemiological studies in Ukraine, Belarus, and Russia (Likhterev et al., 2006a; Tronko et al., 2006; Kopecky et al., 2006; Zablotska et al., 2011). The exceptional interest in this problem is also accounted for by the presence of sufficiently complete and reliable information about the risk of radiation-induced thyroid cancer in case of exposure of this organ to external radiation (Ron et al., 1995). As to internal exposure, data on the related radiation risk value are extremely scarce (Likhterev et al., 2006a; Tronko et al., 2006; Kopecky et al., 2006; Zablotska et al., 2011).

When interpreting the results of the most radio-epidemiological studies described in the above-cited papers, a series of general assumptions were made, primarily concerning the estimation of the radiation factor, namely the exposure dose:

- It was noted that the exposure dose estimates contain errors which are usually significant.
- Even in the case where the variances of dose errors were determined, the analytical tools of risk analysis ignored this fact.
- In the dosimetric support for radio-epidemiological studies, instrumental measurements of any types were practically always absent.

Thus, the interpretation of the most radio-epidemiological studies was based on risk estimation methods failing to take into account the presence of errors in the exposure doses. One of the consequences of the assumption about the absence of errors in the doses is a bias of the risk coefficient estimates and a distortion of the form of the "dose–effect" curve. Note that such distortions result not only from systematic errors in dose estimates, which is obvious, but also from random errors as well.

It is known that a measured or estimated dose is inevitably accompanied by errors of the classical or Berkson type, or by a mixture of them (Lyon et al., 2006; Li et al., 2007; Kukush et al., 2011; Mallick et al., 2002; Masuk et al., 2016). And at the same time, there is still no final conclusion as to the impact of a classical, Berkson, or mixed error

in dose estimates on the end result of risk analysis, which is usually expressed in terms of relative (*ERR*) or absolute (*EAR*) risk (Health Risks from Exposure to Low Levels of Ionizing radiation, 2006).

A prominent example of the importance and urgency of this problem is the interpretation of the results of long-term radio-epidemiological studies of a cohort of children with thyroid exposure caused by the accident at the Chornobyl nuclear power plant (Jacob et al., 2006; Likhtarov et al., 2006a; Tronko et al., 2006; Zablotska et al., 2011). It is vital to note that the incidence of thyroid cancer in this cohort were determined quite accurately. Also, there were obtained not only determined (i.e., point) estimates, but stochastic (i.e., interval) dose estimates as well (Likhtarov et al., 2005, 2006b). However, no more or less acceptable mathematically reasonable computational procedure for combining two-dimensional error in dose and in effect within a unified procedure of risk analysis is available at present. *EPICURE*, the most popular software package in radio-epidemiology (Preston et al., 1993), operates upon determined dose values and is not adapted to account for any uncertainty of the input data.

This book is devoted to the problem of estimation of the radiation risk as a result of the thyroid exposure by radioactive iodine. The focus is primarily on the binary model of disease incidence in which the odds function is linear in exposure dose. The thyroid exposure dose is not measured directly by a device; the estimated dose is based on primary individual and environmental data, the model of atmospheric radioactivity transfer, the biokinetic model of radioiodine transport, individualized thyroid masses, and lastly, data from direct individual radioiodine measurements of the thyroid made in May and June, 1986. As a result, the final estimates of exposure doses contain both classical and Berkson errors. The mixture of measurement errors of different types makes risk estimation quite a hard task to accomplish. The main goal of the book is to develop modern methods of risk analysis that would allow taking into consideration such uncertainties in exposure doses.

This book describes known methods of risk estimation in the binary model of disease incidence in the presence of dose errors: maximum likelihood, regression calibration, and also develops original estimation methods for this model, namely, the corrected score method and *SIMEX* (simulation–extrapolation method). The efficiency of the methods was tested by a stochastic experiment based on the results of radio-epidemiological studies of thyroid cancer incidence rate after the Chornobyl accident. The essence of the experiment is as follows: the real thyroid doses were contaminated with generated measurement errors, and also thyroid cancer cases were generated based on the binary model of disease incidence with realistic risk coefficients. After that, radiation risk estimation was performed in the presence of errors in exposure doses. Such a risk analysis requires a deep study of observation models with errors in covariates which conceptually are not reduced to ordinary regression models and are characterized by most complicated parameter estimation. Such models are widely used in various fields of science, particularly in epidemiology, meteorology, econometrics, signal processing, and identification of dynamic systems. Recently, a series

of fundamental papers devoted to this subject matter was published. Thus, the books by Schneeweiss and Mittag (1986) and Fuller (1987) study linear models, both scalar and vector ones; the manual by Cheng and Van Ness (1999) investigates the linear and polynomial models; the book by Wansbeek and Meijer (2000) discloses the use of linear and quadratic models in econometrics; finally, both editions of the book by Carroll et al. (1995, 2006) describe various nonlinear models and their applications in epidemiology.

This book comprises two parts, a list of references and appednices. The first part of the book (Chapters 1–4) is based on a special course "Regression Measurement Error Models" that one of the authors has been teaching for a long time at the Mechanics and Mathematics Faculty of Taras Shevchenko National University of Kyiv. The second part (Chapters 5–7) contains the results of long-term studies performed at the Department of Dosimetry and Radiation Hygiene of the Institute of Radiation Hygiene and Epidemiology of the National Research Center for Radiation Medicine of the National Academy of Medical Sciences of Ukraine. Chapter 1 provides a general overview of regression errors-in-variables models and a comparison of the main methods for estimating regression parameters. Chapter 2 presents the mostly used linear model with the classical error, Chapter 3 analyzes the polynomial regression model, and Chapter 4 studies other popular nonlinear models, including the logistic one. Chapter 5 makes an overview of risk models implemented in the software package *EPICURE*. Chapter 6 deals directly with radiation risk estimation in the binary model with linear risk in the presence of measurement errors in thyroid doses. It analyzes in detail combined multiplicative errors in exposure doses as a mixture of the classical and Berkson errors. Finally, Chapter 7 undertakes a thorough analysis of procedures for thyroid dose estimation and considers a more realistic model for errors in doses, namely, a mixture of the classical additive and the multiplicative Berkson errors. The four appendices contain the mathematical foundations for the proposed estimation methods. In particular, Appendix A outlines with mathematical rigor the elements of the theory of unbiased estimating equations, including the conditions for the existence and uniqueness of a solution, which defines the parameter estimator, and asymptotic properties of the estimators.

Knowledge of the basics of calculus and probability theory as presented in the standard obligatory courses (Burkill, 1962; Kartashov, 2007) is sufficient to understand the material. It is desirable to know the Lebesgue integral theory (Halmos, 2013), although utilizing conventional mathematical formalism when calculating expectations is enough to comprehend most of the contents.

The book will be useful to experts in probability theory, mathematical and applied statistics, specialists in biomedical data processing, epidemiologists, dosimetrists, and university students enrolled in specialties, "statistics," "applied statistics," or "mathematics".

The authors express their sincere gratitude to Doctor of Technical Sciences Leonila Kovgan for comprehensive support and assistance in writing this book.

# *In memoriam*
# Illya Likhtarov (1935–2017)

Prof. Dr. Illya A. Likhtarov, an outstanding Ukrainian biophysicist, an expert in radiation dosimetry, radiological protection and risk analysis, and a scientist of world level, passed away suddenly and unexpectedly on January 14, 2017.

Illya Likhtarov was born on February 1, 1935 in the town of Pryluky in Chernihiv Oblast of Ukraine. He spent his childhood in Kyiv, which became his lifelong hometown. He started his carrier in 1960 with the radiology group of Kyiv Regional Sanitary–Epidemiological Station. In 1962 he graduated with honors from the All-Union Correspondence Polytechnic Institute in Moscow as an engineer-physicist.

In 1964, I. Likhtarov enrolled in graduate school at the Leningrad Institute of Radiation Hygiene (IRH). Illya's early charge was experimental and theoretical work on the safety of radioactive iodine. This included studies in animals and in human volunteers, the development of a model of iodine metabolism in the body, the application of protective agents of stable iodine, and the radiobiological effect of radioiodine in the thyroid gland. Upon successful completion of this work in 1968, I. Likhtarov received his Ph.D. degree.

From 1966 to 1986 Dr. Likhtarov led the Laboratory of Radiation Biophysics within the IRH responsible for studying radionuclide metabolism and dosimetry of internal human exposure. Under his leadership studies were conducted on the metabolism of tritium, iodine, strontium, calcium, plutonium and other radionuclides in humans and animals, and mathematical models and methods for calculation of internal doses were developed. Radiation safety standards were developed and implemented for workers and the public. In 1976, Dr. Likhtarov obtained the degree of Doctor of Sciences with a specialty in Biophysics.

Immediately after the Chernobyl accident on April 26, 1986 Dr. Likhtarov returned to Ukraine, where he was an expert advisor to the Ukrainian Minister of Health. He supervised the work under emergency conditions of numerous radiation measurements, assessing current and future doses of the populations of the affected areas, and in the development and implementation of protective measures.

In October 1986, Illya created and headed the Department of Dosimetry and Radiation Hygiene of the newly established All-Union Scientific Center for Radiation Medicine (now the National Research Center for Radiation Medicine of Ukraine). The Department became the base of the prolific Ukrainian scientific school of dosimetry and radiological protection, which has been functioning for over 30 years. As a part of his legacy, Dr. Likhtarov trained many young professionals in this field.

In the aftermath of the Chernobyl accident Illya Likhtarov and his team were faced with the task of large-scale assessment of the radiation situation in more than 2,200 towns that were home to more than 3.5 million people. This task was complicated due to a variety of environmental and social conditions. Under his guidance numer-

ous measurements of cesium and strontium radionuclides in the body of citizens of Ukraine were made, and a set of eco-dosimetric models was developed. Appropriate measures for radiation protection of the population and rehabilitation of Ukrainian territories were carried out. Dr. Likhtarov's team also developed and implemented a system of thyroid dose reconstruction of the entire population of Ukraine.

In 1995 Dr. Likhtarov and his co-workers founded the Ukrainian Radiation Protection Institute (RPI). The RPI has developed into the Ukrainian center of expertise for radiation protection and dosimetry: Core regulations have been drafted and support has been provided to the national authorities and industries. During the last 15 years, the RPI has been providing the occupational safety and internal dosimetry services for the international project for the erection of the Chernobyl's New Safe Confinement. Professor Likhtarov's team implemented an unprecedented large-scale program of individual monitoring of internal exposure, which is focuses on the intake of transuranic elements and covers more than 17,000 workers.

In the early 1990s, the incidence rate of thyroid cancer increased in children residing in the affected areas of Ukraine. Initial analysis showed significant correlation between the thyroid dose caused by ingestion of radioiodine and the cancer incidence rate. The results were published in 1995 in Nature and attracted the interest of researchers from many countries. In the mid-1990s, a cohort of about 13,000 children (as of 1986) was formed for a long-term Ukrainian–US epidemiological study of radiogenic thyroid cancer. The dosimetry team led by Dr. Likhtarov created an original model for dose assessment that considers individual behavior of subjects and the environmental characteristics in places of their residence. At the moment when point estimates of doses were obtained, it became clear that ignoring errors in exposure doses causes essential underestimation of radiation risks, and therefore, underestimation of harmful effect of ionizing exposure on human health. That is why Dr. Likhtarov organized a working group on elaboration of methods for radiation risk estimation, which take into account dose uncertainties. The activity of the group resulted in this monograph, which shows the way how to estimate correctly the risk of the radiation incidence rate of thyroid cancer in Ukraine after the Chornobyl accident.

In recognition of Dr. Likhtarov's scientific achievements, he was elected a member of the USSR National Radiological Protection Commission in 1978. Since 1992, he has headed the Commission on Radiation Standards of Ukraine that developed and implemented into practice basic national regulatory documents. In 1993 he was elected to Committee 2 (Dosimetry) of the International Commission of Radiation Protection (ICRP), where he worked successfully until 2005. Since 2002 he was also a member of the IAEA Radiation Safety Standards Committee (RASSC).

Dr. Likhtarov carried out extensive international cooperation with specialists from the USA, Europe and Japan since the Soviet times and afterwards. He participated in international congresses and conferences, where his papers and participation in discussions invariably aroused keen interest of the audience.

The scientific heritage of Dr. Likhtarov includes more than 600 scientific papers; among them are articles in prestigious journals, monographs, and documents of the ICRP, UNSCEAR, WHO and IAEA. Professor Likhtarov is included in the list of the 50 most cited scientists of Ukraine. Under his guidance, 25 students have earned Ph.D. and 10 have earned Doctor of Science degrees.

Illya Likhtarov is survived by his wife Lionella Kovgan, sons Mikhail and Dmitry, step daughter Tamila Kovgan, six grandchildren, and his sister Elena. He was preceded in death by his parents, twin sister Rosa, and his first wife Tamara Likhtarova.

# Contents

## Part II Radiation risk estimation under uncertainty in exposure doses

## Translated from the Ukrainian edition:

Моделі регресії з похибками вимірювання та їх застосування до оцінювання радіаційних ризиків / С. В. Масюк, О. Г. Кукуш, С. В. Шкляр, М. І. Чепурний, І. А. Ліхтарьов; за ред. д-ра фіз.-мат. наук, проф. І. А. Ліхтарьова – К.: ДІА, 2015. – 288 с.

## Summary

Ignoring errors in exposure doses leads to reducing of the radiation risk estimates, and therefore, is a reason for underestimation of unhealthy action of the exposure.

The first part of the book is devoted to nonlinear measurement error models and parameter estimation in the models. The second part deals with the problem of risk estimation in the models with errors in exposure doses. Well-known methods and original statistical methods of risk estimation are described in the presence of measurement errors in covariates. Efficiency of the methods is verified based on real radio-epidemiological studies.

The book will be useful to experts in mathematical and applied statistics, specialists in biomedical data processing, epidemiologists, dosimetrists, and university students enrolled in specialties, “statistics,” “applied statistics,” or “mathematics”.

# List of symbols, abbreviations, units, and terms

## Symbols

| | |
|---|---|
| $\mathbf{R}$ | Set of real numbers |
| $\mathbf{R}^m$ | Denotes $m$-dimensional Euclidean space of column-vectors |
| $\mathbf{R}^{n\times m}$ | Set of real matrices of size $n \times m$ |
| $(x, y) = \sum_{i=1}^{n} x_i y_i$ | Inner product of two vectors in Euclidean space |
| $\|\|z\|\| = \sqrt{\sum_{i=1}^{n} z_i^2}$ | Euclidean norm of a vector $z$ |
| $B(x, r)$ | Open ball with center $x$ and radius $r$ |
| $\overline{B}(x, r)$ | Closed ball with center $x$ and radius $r$ |
| $A^0$ | Set of all interior points of a set $A$ |
| $\mathbf{I}_A$ | Indicator (or characteristic) function of a set $A$ |
| $Z^{\mathrm{T}}$ | Transposed vector $Z$ or transposed matrix $Z$ |
| $A^{-\mathrm{T}}$ | Matrix $(A^{\mathrm{T}})^{-1} = (A^{-1})^{\mathrm{T}}$, defined for a nonsingular matrix $A$ |
| $\lambda_{\min}$ | Least eigenvalue of a matrix |
| trace | Trace of a square matrix |
| $A \geq B$, $A > B$ | For square matrices, it means that $A - B$ is positive semidefinite (positive definite); in particular, for a matrix $A$, $A > 0$ means that $A$ is positive definite |
| $\|\|Z\|\|_{\mathrm{F}} = \sqrt{\sum_{i,j} z_{ij}^2}$ | Frobenius norm of a matrix $Z$ |
| $\|\|C\|\|$ | Operator norm of a matrix $C$, $\|\|C\|\| = \sup\{\frac{\|\|Cx\|\|}{\|\|x\|\|} : x \neq 0\}$ |
| $\mathbf{P}(A)$ | Probability of a random event $A$ |
| $\mathbf{E}$ | Expectation of a random variable, a random vector or a random matrix |
| $\mathbf{D}$ or $\mathbf{V}$ | Variance of a random variable |
| $\mathrm{cov}(x, y)$ | Covariance of random variables $x$ and $y$ |
| $\mathrm{cov}(z)$ | Covariance matrix of a random vector $z$ |
| $\mathbf{E}_\beta$ | Expectation given that $\beta$ is the true value of a parameter; similarly the next notations are introduced: $\mathbf{P}_\beta(A)$, $\mathbf{D}_\beta$, $\mathbf{V}_\beta$, $\mathrm{cov}_\beta$ |
| $\overline{x}$ | Mean value of observations $x_1, \ldots, x_n$ |
| $\mathrm{med}\, x$ | Median of a random variable $x$ |
| $\mathrm{GSD}(x)$ | Geometric standard deviation of a positive random variable $x$, $\mathrm{GSD}(x) = \mathrm{e}^{\sigma}$, $\sigma = (\mathbf{D} \ln x)^{1/2}$ |
| $x =^{\mathrm{d}} y$ | Means that random variables $x$ and $y$ have the same distribution |
| $\mathrm{corr}(x, y)$ | Correlation coefficient of random variables $x$ and $y$ |
| $x \perp\!\!\!\perp y$ | Stochastic independence of random variables or random vectors $x$ and $y$ |

| | |
|---|---|
| $S_{xx}$ | Sample variance of a variable $x$ |
| $S_{xy}$ | Sample covariance of variables $x$ and $y$ |
| $N(\mu, \sigma^2)$ | Normal distribution with mean $\mu$ and variance $\sigma^2$ |
| $LN(\mu, \sigma^2)$ | Lognormal distribution with parameters $\mu$ and $\sigma^2$; if a random variable $X \sim LN(\mu, \sigma^2)$, then $\ln X \sim N(\mu, \sigma^2)$ |
| $Pois(\lambda)$ | Poisson distribution with parameter $\lambda$ |
| $\chi^2_n$ | Chi-squared distribution with $n$ degrees of freedom |
| $\xrightarrow{P1}$ | Convergence almost surely (a.s.), that is with probability 1 |
| $\xrightarrow{\mathbf{P}}$ | Convergence in probability |
| $\xrightarrow{d}$ | Convergence in distribution |
| $O_p(1)$ | Stochastically bounded sequence of random variables or random vectors |
| $o_p(1)$ | Sequence of random variables or random vectors, which tends to 0 in probability |
| $K$ | Reliability ratio in a model with the classical error, $K = \frac{\sigma^2_\xi}{\sigma^2_\xi + \sigma^2_\delta}$ |
| $\lambda$ | Total risk or total incidence rate |
| $\lambda_0$ | Baseline risk or background incidence rate |
| $D^{\text{mes}}$ | Instrumental (i.e., measured) exposure dose |
| $D^{\text{tr}}$ | Actual (or true) exposure dose |
| $Q^{\text{mes}}$ | Measured value of $^{131}$I activity in thyroid gland |
| $Q^{\text{tr}}$ | True value of $^{131}$I activity in thyroid gland |
| $\delta_Q$ | Relative classical additive error of the result of direct measurement of thyroid radioactivity in May and June, 1986 |
| $\sigma^2_F$ | Variance of the logarithm of Berkson multiplicative error in absorbed dose |
| $\sigma^2_Q$ | Variance of the logarithm of the classical multiplicative error in absorbed dose |
| $GSD_Q = \exp(\sigma_Q)$ | Geometric standard deviation of the classical multiplicative error in absorbed dose |
| $GSD_F = \exp(\sigma_F)$ | Geometric standard deviation of Berkson multiplicative error in absorbed dose |

## Abbreviations

| | |
|---|---|
| ACM | Asymptotic covariance matrix of an estimator |
| AM | Arithmetic mean |
| ALS | Adjusted least squares |
| a.s. | Almost surely, i.e., with probability 1 |
| cdf | Cumulative distribution function |

| | |
|---|---|
| CLT | Central limit theorem |
| CNPP | Chornobyl Nuclear Power Plant |
| CSE | Corrected score estimator, i.e., the estimator obtained by the method of correction an estimating function |
| CS | Corrected score method |
| *95% DI* | Deviance interval, based on 2.5% and 97.5% quantiles of the estimates |
| *EAR* | Excess absolute risk |
| *EPICURE* | Package of applied programs for parameter estimation in the models of absolute and relative risks without taking into consideration errors in covariates |
| *ERR* | Excess relative risk |
| GLM | Generalized linear model |
| GSD | Geometric standard deviation |
| ML | Maximum likelihood |
| MLE | Maximum likelihood estimator |
| NPFML | Non-parametric method of full maximum likelihood |
| NPRC | Non-parametric method of regression calibration |
| NRC | New regression calibration |
| ORE | Orthogonal regression estimator |
| pdf | Probability density function |
| PFML | Parametric method of full maximum likelihood |
| PRC | Parametric method of regression calibration |
| QLE | Quasi-likelihood estimator |
| q50% | 50% quantile of a probability distribution (i.e., median) |
| RC | Regression calibration |
| *SIMEX* | Simulation and extrapolation method |
| SLLN | Strong law of large numbers |

## Units

***Becquerel (Bq)*** is a unit of radioactivity, in SI. One Bq is defined as the activity of a quantity of radioactive material in which one nucleus decays per second.

***Gray (Gy)*** is a unit of absorbed dose of ionizing radiation, in SI; 1 Gy = 1 J $kg^{-1}$.

## Terms

***Absolute risk*** is the difference between the frequency of adverse effect among persons exposed to a factor that is studied (e.g., exposure dose) and the frequency of the effect in a group of persons who are not exposed to the factor.

***Absorbed dose (D)*** is the energy absorbed per unit mass at a given point. The unit is the J per kilogram (J $kg^{-1}$) and is given the special name gray (Gy).

***Background (or spontaneous) incidence rate*** is a component of incidence rate not associated with the effect of ionizing radiation on humans.

***Consistent estimator*** is a statistical estimator of regression parameters that converges in probability to the true values of the parameters, as the sample size tends to infinity.

***Efficient estimator*** is the unbiased statistical estimator with the least variance within the class of all unbiased estimators of a given parameter.

***Eventually*** means with probability 1 for any sample size exceeding certain random number (see Definition 2.12).

***Excess absolute risk (EAR)*** is the coefficient at exposure dose in the linear model of absolute risk, which coincides with radiation induced incidence rate if the exposure dose is 1 Gy. In SI, the unit is $Gy^{-1}$.

***Excess relative risk (ERR)*** is the coefficient at exposure dose in the linear model of relative risk, which shows for how many times the radiation induced incidence rate is higher than the background incidence rate if the exposure dose is 1 Gy. In SI, the unit is $Gy^{-1}$.

***Incidence rate*** is the number of cases of a disease, e.g., cancer, recorded during a year per a certain number of persons (10 000, 100 000 or 1 000 000).

***Internal exposure*** is the exposure of the human body (its separate organs and tissues) by ionizing radiation sources being in the body.

***Ionizing radiation*** is radiation (either electromagnetic or corpuscular), which when being interacted with substance causes ionization (directly or not) and excitation of its atoms and molecules.

***Naive estimator*** is a statistical estimator of regression parameters under errors in covariates, which is obtained by the method where such errors are ignored (e.g., by the ordinary maximum likelihood method).

***Nuisance parameter*** is a parameter of regression model but not a regression coefficient (e.g., a parameter of regressor's distribution or the variance of error in response). Nuisance parameter estimation is not an ultimate goal of statistical study.

***Organ dose ($D_T$)*** is a quantity defined in ICRP (1991) in relation to the probability of stochastic effects (mainly cancer induction) as the absorbed dose averaged over an organ, i.e., the quotient of the total energy imparted to the organ and the total mass of the organ:

$$D_T = E_T/m_T\,,$$

where $E_T$ is total ionization energy imparted to the organ or tissue $T$ and $m_T$ is mass of the organ or tissue.

***Radiation incidence rate*** is a component of incidence rate caused by exposure of the human body (its separate organs and tissues) by sources of ionizing radiation.

***Radiation risk*** is the probability for a person or his/her descendants to get any harmful effect caused by ionizing radiation.

***Regression*** is a form of link between random variables. It is a law of change for expectation of a random variable in dependency on the values of another one. There are linear, polynomial, nonlinear, and other kinds of regression.

***Regressor*** is a covariate in regression analysis.

***Relative risk*** is the ratio of the frequency of adverse effect among persons exposed to a factor that is studied (e.g., radiation dose) to the frequency of the effect in a group of persons who are not exposed to the factor.

***Response*** is a dependent variable in regression analysis.

***Stochastic effects*** are nonthreshold effects of radiation influence, the probability of which is positive at any dose of ionizing radiation and increases with dose. The stochastic effects include malignancies (i.e., somatic stochastic effects) and genetic changes transmitted to descendants (i.e., hereditary effects).

***Unbiased estimator*** is a statistical estimator, with expectation equal to the estimated parameter.

# Part I: **Estimation in regression models with errors in covariates**

# 1 Measurement error models

Consider an ordinary model of nonlinear regression

$$y_i = f(\xi_i, \beta) + \varepsilon_i\,, \quad i = \overline{1, n}\,. \tag{1.1}$$

Here, $\xi_i \in \mathbf{R}^d$ are known (observable) values of *regressors*;

$\beta$ is an unknown *vector of regression parameters* that belongs to the parameter set $\Theta \subset \mathbf{R}^p$;

$f\colon \mathbf{R}^d \times \Theta \to \mathbf{R}$ is a known (given) *regression function*;

$\varepsilon_i$, $i = \overline{1, n}$ are random observation errors usually assumed to be independent, centered (i.e., with zero mean), and having finite variance;

$y_i$, $i = \overline{1, n}$ are observable values of the dependent variable, or response.

The regression parameter $\beta$ should be estimated in frames of the model (1.1) by observations $\{y_i, \xi_i, i = \overline{1, n}\}$.

As another example of regression model, consider a binary logistic model. Let the response $y_i$ take two values, 0 and 1, depending on the true value of the regressor $\xi_i$, namely

$$\mathbf{P}\{y_i = 1|\xi_i\} = \frac{\lambda(\xi_i)}{1 + \lambda(\xi_i)}\,, \quad \mathbf{P}\{y_i = 0|\xi_i\} = \frac{1}{1 + \lambda(\xi_i)}\,, \quad i = \overline{1, n}\,, \tag{1.2}$$

where

$$\lambda(\xi_i) = \lambda(\xi_i, \beta) = \mathrm{e}^{\beta_0 + \beta_1 \xi_i} \tag{1.3}$$

is the odds function and $\beta = (\beta_0, \beta_1)^{\mathrm{T}}$ is the regression parameter. The observed couples $\{(y_i, \xi_i), i = \overline{1, n}\}$ are assumed stochastically independent, and the parameter $\beta$ should be estimated by the observations.

The model (1.2) is widely used in epidemiology and can be interpreted as follows. $y_i$ is an indicator of disease for the subject $i$ of a cohort; in the case $y_i = 1$, the subject $i$ has obtained a given type of disease during a fixed period of observations; in the case $y_i = 0$, the subject $i$ has not demonstrated any symptoms of the disease during the period; $\xi_i$ is the true value of the regressor that affects on disease incidence.

Another model of the odds function is widely used in radio-epidemiology, namely, linear one

$$\lambda(\xi_i, \beta) = \beta_0 + \beta_1 \xi_i\,, \quad i = \overline{1, n}\,. \tag{1.4}$$

The model (1.2) and (1.4) serves for the simulations of the incidence rate of oncological diseases caused by radiation exposure; the regressor $\xi_i$ is the radiation dose recieved by the subject $i$ of the cohort during the fixed period of observations. This period includes also all the registered cases of cancer.

The parameters $\beta_0$ and $\beta_1$ to be estimated are the ones of radiation risk. From mathematical point of view, the model (1.2) and (1.4) is a generalized linear model

(GLM) of binary observations; it resembles a logistic model (1.2) and (1.3) but somewhat differs from it. Herein the term "generalized linear model" means that the parameters of the conditional distribution $y_i|\xi_i$ depend on the regressor $\xi_i$ through a linear expression $\beta_0 + \beta_1\xi_i$, where $\beta_0$ and $\beta_1$ are unknown regression parameters. Of course, the logistic model (1.2) and (1.3) is also a generalized linear regression model.

The main object of this book is a model with errors in the regressor. It is called also *errors-in-variables model* or *measurement error model*.

A distinguishing feature of such models is that the true value of the regressor $\xi_i$ is unknown to us, but instead we observe a surrogate value $x_i$ including apart of $\xi_i$, a random measurement error, i.e.,

$$x_i = \xi_i + \delta_i\,, \quad i = \overline{1,n}\,. \tag{1.5}$$

We assume that the errors $\delta_i$ are centered, stochastically independent, and have finite variance; moreover, the values $\{\xi_i, \delta_i, i = \overline{1,n}\}$ are stochastically independent. Such errors are called *classical* (additive) measurement errors. In the model (1.5), the unobservable regressor $\xi_i$ is called also a latent variable.

In particular, equalities (1.1) and (1.5) describe a *nonlinear regression model* with *measurement errors*. Exact assumptions about base units are the following:

- Random vectors and random variables $\{\xi_i, \varepsilon_i, \delta_i, i = \overline{1,n}\}$ are independent.
- Vector $\beta$ should be estimated based on observations $\{y_i, x_i, i = \overline{1,n}\}$.

In both logistic model with errors in the covariates (1.2), (1.3), and (1.5) and binary GLM with errors in the covariates (1.2), (1.4), and (1.5), exact assumptions about basic values are the following:

- Random variables $\{(y_i, \xi_i), \delta_i, i = \overline{1,n}\}$ are independent.
- Vector $\beta$ has to be estimated based on observations $\{y_i, x_i, i = \overline{1,n}\}$.

From a general point of view, the model with measurement errors consists of three parts:

(a) the regression model that links the response variable with unobservable $\xi$ and observable $z$ regressors, respectively;
(b) the measurement error model that connects $\xi$ with the observable surrogate variable $x$;
(c) the distribution model of $\xi$.

In particular, the binary model (1.2), (1.4), and (1.5) gives: (1.2) and (1.4) to be the regression model in which there is no regressor $z$ observable without errors; and (1.5) to be the model of measurement errors. In regards to the distribution of regressors $\xi_i$, one can assume, e.g., that the $\xi_i$, $i = \overline{1,n}$ are independent and identically distributed with normal distribution $\mathrm{N}(\mu_x, \sigma_\xi^2)$, wherein the parameters $\mu_x$, $\sigma_\xi^2$ (also called nuisance parameters) can be either known or unknown.

A regressor $z$ can be included in the risk model as follows:

$$\lambda(\xi_i, \beta, \gamma) = e^{\gamma^T z_i}(\beta_0 + \beta_1 \xi_i)\,, \quad i = \overline{1, n}\,. \tag{1.6}$$

Here, $\{z_i, i = \overline{1, n}\}$ are independent random vectors in $\mathbf{R}^q$, and $\gamma \in \mathbf{R}^q$ is a vector of additional regression parameters. The estimators of the regression parameters $\beta$ and $\gamma$ are constructed based on observations $\{y_i, z_i, x_i, i = \overline{1, n}\}$ in frames of the model (1.2), (1.5), and (1.6). Within the radio-epidemiological risk models, the vector $z_i$ components may present the age of person $i$ as well as his/her gender along with individual features of the person. Usually the regressor $z$ is the categorical variable, namely the one that takes discrete values.

## 1.1 Structural and functional models, linear, and nonlinear models

In this section, we assume for simplicity that the observed part of regressors $z$ to be absent in underlying measurement error models.

Consider the regression model linking the response $y_i$ with the regressors $\xi_i$. This model is called *structural* if the true values of $\xi_i$, $i = \overline{1, n}$, are random, moreover, they are independent and identically distributed in $\mathbf{R}^d$. Usually, we know the form of the regressor $\xi$ distribution, i.e., the probability density function (pdf) of $\xi$ is known up to certain parameters. Those values (nuisance parameters) can be either known or unknown. If they are known, the distribution of $\xi$ is given exactly.

For example, the structural logistic model (1.2), (1.4), and (1.5) usually requires $\{\xi_i, i = \overline{1, n}\}$ to be independent and identically distributed random variables with common lognormal distribution $\mathrm{LN}(\mu_x, \sigma_\xi^2)$, where the nuisance parameters $\mu_\xi \in \mathbf{R}$ and $\sigma_\xi > 0$ are unknown.

Unlike the structural model, the functional regression model assumes the true values of $\{\xi_i, i = \overline{1, n}\}$ to be nonrandom. Within the functional errors-in-variables models, the true regressor values $\xi_i$ become the nuisance parameters; their number grows as the sample size $n$ increases making it difficult to perform research.

Actually, the structural models impose stringent restrictions on the behavior of regressor values. Thus, the assumption that $\{\xi_i, i \geq 1\}$ is a sequence of independent identically distributed random variables with finite variance ensures, using the law of large numbers, the existence of finite limits for expressions $\overline{\xi} = \frac{1}{n}\sum_{i=1}^n \xi_i$, $\overline{\xi^2} = \frac{1}{n}\sum_{i=1}^n \xi_i^2$, and even for $\frac{1}{n}\sum_{i=1}^n g(\xi_i)$, where $g$ is a Borel measurable function such that $|g(t)| \leq \mathrm{const}\,(1 + t^2)$, $t \in R$.

In the framework of the functional models, if, e.g., it is required only to have additional stabilization of the expressions $\overline{\xi}$ and $\overline{\xi^2}$, then it does not follow the existence of finite limit for the average $\frac{1}{n}\sum_{i=1}^n \sin \xi_i$.

The choice between the structural and functional models relies on an accurate analysis of the measurement procedure. In epidemiology, the structural models are

more popular, while modeling physical or chemical experiment where monotonously changing regressor $\xi$ takes place is more relevant to the functional models.

We give an example of the functional model. In the adiabatic expansion, the gas pressure and gas volume are related to each other according to Boyle's law:

$$pV^{\gamma} = c\,. \tag{1.7}$$

Here, $\gamma$ and $c$ are some positive constants that we want to estimate by experiment. Rewrite (1.7) as

$$\ln p = -\gamma \ln V + \ln c\,. \tag{1.8}$$

Denote

$$y = \ln p\,, \quad \xi = \ln V\,; \quad \beta_0 = \ln c\,, \quad \beta_1 = -\gamma\,. \tag{1.9}$$

Then the response $y$ is related linearly to the regressor $\xi$:

$$y = \beta_0 + \beta_1 \xi\,. \tag{1.10}$$

Suppose that we observe $y$ and $\xi$ with additive errors

$$\begin{aligned} y_i &= \beta_0 + \beta_1 \xi_i + \varepsilon_i\,, \\ x_i &= \xi_i + \delta_i\,, \quad i = \overline{1,n}\,. \end{aligned} \tag{1.11}$$

Naturally, regressors $\xi_i = \ln V_i$ can be supposed nonrandom because the adiabatic expansion affects on gas volume $V_i$ being increased in time. Thus, the observation model (1.8) is naturally assumed to be the functional one, since the values $\xi_i$ are nonrandom although unobservable. The errors $\{\varepsilon_i, \delta_i, i = \overline{1,n}\}$ are assumed independent. By observations $\{y_i, x_i, i = \overline{1,n}\}$ the regression parameters $\beta_0$ and $\beta_1$ are estimated and next the parameters of equation (1.7), namely $c = \mathrm{e}^{\beta_0}$ and $\gamma = -\beta_1$, are estimated as well.

Note that in terms of the original variables $p_i$ and $V_i$, we have a model with multiplicative errors

$$\begin{aligned} p_i^{\mathrm{mes}} &= c(V_i^{\mathrm{tr}})^{-\gamma} \cdot \tilde{\varepsilon}_i\,, \\ V_i^{\mathrm{mes}} &= V_i^{\mathrm{tr}} \cdot \tilde{\delta}_i\,, \quad i = \overline{1,n}\,. \end{aligned} \tag{1.12}$$

Here $\tilde{\varepsilon}_i = \mathrm{e}^{\varepsilon_i}$ and $\tilde{\delta}_i = \mathrm{e}^{\delta_i}$ are the multiplicative errors; $V_i^{\mathrm{tr}}$ is the true value of gas volume; $p_i^{\mathrm{mes}}$ and $V_i^{\mathrm{mes}}$ are measured values of the pressure and volume, respectively. As this, the unknown values $V_i^{\mathrm{tr}}$ are assumed to be nonrandom.

Now, we consider an example of the structural model. Let us investigate the crime rate $\eta$ in dependence on the average annual income $\xi$ in a certain district of a country. This dependence we model using a given regression function $\eta = f(\xi, \beta)$, where $\beta$ is the vector of regression paramers. We randomly select the area $i$ (e.g., the area of a big city), measure the crime rate (e.g., the number of registered crimes per capita), and take into account the average annual income $\xi_i$ (e.g., by interviewing residents). For obvious reasons, the measurement will be inaccurate. One can assume that measurement errors are additive:

$$\begin{aligned} y_i &= f(\xi_i, \beta) + \varepsilon_i\,, \\ x_i &= \xi_i + \delta_i\,, \quad i = \overline{1,n}\,. \end{aligned} \tag{1.13}$$

It is appropriate to assume that the latent variables $\xi_i$ are random because we randomly select a particular area for research and not moving, say, from poor to wealthier areas. The structural regression model naturally describes our experiment.

The regression model (1.1) in which the function $f$ depends linearly on the regressor $\xi_i$ is called *linear*. All other models of the regression of response $y$ on the regressor $\xi$ are called *nonlinear*. These are, in particular, the model (1.1) in which the regression function $f$ depends nonlinearly (e.g., polynomially) on $\xi_i$, and the binary model (1.2) with any dependence of the odds function $\lambda$ on $\xi_i$. The linear measurement error models will be studied in Chapter 2, and the nonlinear ones in all subsequent chapters.

## 1.2 Classical measurement error and Berkson error

As already noted, the error $\delta_i$ is called the *classical* one in the context of the model (1.5), if $\xi_i$ and $\delta_i$ are stochastically independent. The error describes instrumental measurements when some physical quantity to be measured using a device is characterized by a certain fluctuation error.

There is also the *Berkson* measurement model

$$\xi_i = x_i + \delta_i^{\mathrm{B}}\,, \quad i = \overline{1,n}\,. \tag{1.14}$$

Here, $\xi_i$ is the true value of regressor (random and unobservable), $x_i$ is the result of the observation (random), $\delta_i^{\mathrm{B}}$ is the Berkson error (centered), and it is known that $x_i$ and $\delta_i^{\mathrm{B}}$ are stochastically independent. The model was named after the American Joseph Berkson (1899–1982), who first examined it in 1950.

It seems that one can transform the model (1.14) to the classical model (1.5) as follows:

$$x_i = \xi_i + \tilde{\delta}_i\,, \quad \tilde{\delta}_i = -\delta_i^{\mathrm{B}}\,, \tag{1.15}$$

but then the new error $\tilde{\delta}_i$ becomes correlated with the regressor $\xi_i = x_i - \tilde{\delta}_i$ (remember that now $x_i$ and $\tilde{\delta}_i$ are stochastically independent). Thus, the model (1.14) and the classical model (1.5) are considerably different measurement models.

The Berkson model occurs particularly in situations where the observation $x_i$ is formed by averaging. Imagine that some part of the observations $x_i$, $i = \overline{1,m}$, $2 \le m < n$ is an average quantity $x_{\mathrm{c}}$:

$$x_i = x_{\mathrm{c}}\,, \quad i = \overline{1,m}\,, \quad x_{\mathrm{c}} = \frac{1}{m}\sum_{i=1}^{m} \xi_i\,. \tag{1.16}$$

Then

$$\xi_i = x_i + \delta_i\,, \quad x_i = x_{\mathrm{c}}\,, \quad \delta_i = \left(1 - \frac{1}{m}\right)\xi_i - \frac{1}{m}\sum_{\substack{j=\overline{1,m}\\ j\ne i}} \xi_j\,, \quad i = \overline{1,m}\,. \tag{1.17}$$

Let $\{\xi_i,\, i = \overline{1,m}\}$ be independent with common expectation $\mu_\xi$ and positive variance $\sigma_\xi^2$. Then

$$\mathbf{E}\delta_i = 0\,, \quad \mathbf{D}\delta_i = \frac{m-1}{m}\sigma_\xi^2\,, \quad \mathbf{D}x_{\mathrm{c}} = \frac{\sigma_\xi^2}{m}\,. \tag{1.18}$$

Since $\sum_{i=1}^n \delta_i = 0$, and due to symmetry

$$\operatorname{cov}(x_{\mathrm{c}}, \delta_i) = \frac{1}{m}\operatorname{cov}\left(x_{\mathrm{c}}, \sum_{i=1}^m \delta_i\right) = 0\,, \tag{1.19}$$

$$\operatorname{cov}(\delta_i, \delta_j) = -2\left(1 - \frac{1}{m}\right)\frac{1}{m}\sigma_\xi^2 + \frac{m-2}{m^2}\sigma_\xi^2 = -\frac{\sigma_\xi^2}{m}\,, \quad i \neq j\,. \tag{1.20}$$

Hence the correlation coefficients are

$$\operatorname{corr}(x_{\mathrm{c}}, \delta_i) = \frac{\operatorname{cov}(x_{\mathrm{c}}, \delta_i)}{\sqrt{\mathbf{D}x_{\mathrm{c}} \cdot \mathbf{D}\delta_i}} = 0\,, \tag{1.21}$$

$$\operatorname{corr}(\delta_i, \delta_j) = -\frac{\sigma_\xi^2}{m}\left(\frac{m-1}{m}\sigma_\xi^2\right)^{-1} = -\frac{1}{m-1}\,, \quad i \neq j\,. \tag{1.22}$$

The values $x_{\mathrm{c}}$ and $\delta_i$ are uncorrelated. And $\delta_i$ are almost uncorrelated when $m$ is large, that is, $\operatorname{corr}(\delta_i, \delta_j) = \frac{1}{m-1}$ tends to 0. Therefore, in the model (1.17), the values $\{x_{\mathrm{c}}, \delta_i, i = \overline{1,m}\}$ can be considered approximately uncorrelated. If the probability distributions of $\xi_i$ are normal, then these variables are close to be independent.

This reasoning shows that the model (1.17), in a certain approximation, can be considered as the Berkson model (1.14). The latter can be realized when the entire sample $x_1, \ldots, x_n$ is divided into several groups. Inside each group, the data have been taken as a result of appropriate averaging, namely when one assigns the observation $x_i$ to an average which is close to the arithmetic mean of the corresponding values of the true regressor. We will often deal with such situations in the measurement of exposure doses within radiation risk models.

The Berkson and classical measurement errors can be compared with regard to the efficient estimation of regression parameters. Consider the structural model in two modifications: the first will have the classical errors (1.5) and the second will just have the Berkson ones (1.14), and let

$$\mathbf{D}\delta_i = \mathbf{D}\delta_i^{\mathrm{B}} = \sigma_\delta^2\,, \quad i = \overline{1,n}\,. \tag{1.23}$$

For the first and second modifications, we construct the adequate estimates $\hat\beta_{\mathrm{cl}}$ and $\hat\beta_{\mathrm{B}}$ of the regression parameter $\beta$ (we will explain later how to construct such estimates). Then, the deviation of $\hat\beta_{\mathrm{cl}}$ from the true value will (very likely) be more than the corresponding deviation of $\hat\beta_{\mathrm{B}}$

$$\left\|\hat\beta_{\mathrm{cl}} - \beta\right\| > \left\|\hat\beta_{\mathrm{B}} - \beta\right\|\,. \tag{1.24}$$

Here we use the Euclidean norm. The inequality (1.24) is qualitative in its nature and indicates the following trend: for the same level of errors, the classical errors stringently complicate the efficient estimation compared with the Berkson ones. Our experience with the binary model (1.2) and (1.4) confirms this conclusion.

## 1.3 Explicit and implicit models

The models discussed above are explicit regression models in which the response (dependent variable) and the regressor (independent variable) are separated variables. The implicit models are more general where all variables under observation are treated equally. These models have the form

$$\begin{aligned} &G(\eta_i, \beta) = 0\,, \\ &z_i = \eta_i + \gamma_i\,, \quad i = \overline{1,n}\,. \end{aligned} \tag{1.25}$$

Here the latent variables $\eta_i$ belong to $\mathbf{R}^m$ $m \geq 2$, $\gamma_i$ are the random (classical) measurement errors, the regression parameter $\beta$ belongs to a parameter set $\Theta \subset \mathbf{R}^p$, the link function $G\colon \mathbf{R}^m \times \Theta \to \mathbf{R}$ is known. One has to estimate the vector $\beta$ by the observations $\{z_i, i = \overline{1,n}\}$.

Thus, the true values $\eta_i$ lie on a given hypersurface

$$S_\beta = \{\eta \in \mathbf{R}^m\colon G(\eta, \beta) = 0\}\,. \tag{1.26}$$

In fact, we want to retrieve this surface using the observed points. We impose the restriction $m \geq 2$ in order to have at least two scalar variables, components of the vector $\eta$, among which there are some regression relations.

The implicit model (1.25) might be either the functional one where the $\eta_i$ are nonrandom points on the surface $S_\beta$, or the structural one where $\{\eta_i, i = \overline{1,n}\}$ are independent and identically distributed on the surface $S_\beta$.

Here is an example of an implicit model. Suppose that we have to restore a circle using the observations of points on it. For the true points $\eta_\mathrm{i} = (x_i, y_i)^\mathrm{T}$, it holds that

$$(x_i - x_0)^2 + (y_i - y_0)^2 = r^2\,, \quad i = \overline{1,n}\,. \tag{1.27}$$

Here, $c = (x_0; y_0)^\mathrm{T}$ is the center and $r > 0$ is the radius. Instead of the vectors $\eta_i$ there are observed vectors $z_i = \eta_i + \gamma_i$, $i = \overline{1,n}$, where $\gamma_i$ are independent normally distributed random vectors with variance–covariance matrix $\sigma^2 I_2$ ($I_2$ is unit $2\times 2$ matrix), where $\sigma > 0$ is nuisance parameter. Based on the observations $\{z_i, i = \overline{1,n}\}$ one has to estimate the vector $\beta = (x_0; y_0; r)^\mathrm{T}$. This observation model fits exactly the scheme (1.25). The link function is

$$G(x_i, y_i; \beta) = (x_i - x_0)^2 + (y_i - y_0)^2 - r^2\,, \quad i = \overline{1,n}\,, \tag{1.28}$$

and the curve $S_\beta = \{(x; y)^\mathrm{T} \in \mathbf{R}^2\colon (x - x_0)^2 + (y - y_0)^2 = r^2\}$ is just the circle.

If the vectors $\eta_i$ are nonrandom, then we have a functional model, and if, say, $\{\eta_i, i = \overline{1,n}\}$ are independent and uniformly distributed over $S_\beta$ then the model is structural. Similar models occur in meteorology, computer vision as well as in pattern recognition problems. The book by Chernov (2010) deals with the described problem on restoring the circle.

It should be noted that the explicit model (1.1) and (1.5) can be transformed to the implicit one. For this purpose, denote $\alpha_i = f(\xi_i, \beta)$, $G(\alpha_i, \xi_i; \beta) = \alpha_i - f(\xi_i, \beta)$, $\eta_i = (\alpha_i; \xi_i^{\mathrm{T}})^{\mathrm{T}}$, $z_i = (y_i; x_i^{\mathrm{T}})^{\mathrm{T}}$, and $\gamma_i = (\varepsilon_i; \delta_i^{\mathrm{T}})^{\mathrm{T}}$. Relation (1.25) describing the implicit model holds true for the new variables. After such transformation, the response and the regressors are treated on an equal basis.

## 1.4 Estimation methods

In regression errors-in-variables models, there are several reasonable estimation methods. Some of them are *consistent*, i.e., they yield estimators $\hat{\beta}_n$ of the parameter $\beta$ that converge in probability to the true value $\beta$ as the sample size tends to infinity. Others yield estimators with significant deviation from $\beta$; there are also estimators with reduced deviation. For small and moderate samples, the inconsistent estimators may even be advantageous because the consistent ones sometimes converge to $\beta$ too slowly. The data of cohort radioepidemiologic studies include samples of rather moderate size, because if the number $n$ of surveyed persons can reach tens of thousands then the number of oncological cases, fortunately, will be sufficiently smaller (about a hundred). From this point of view, the most promising methods are those that significantly reduce the deviation $\hat{\beta}_n - \beta$ of the estimators compared with "rough" estimation methods.

### 1.4.1 Naive estimators

The naive estimation method constructs the estimators by algorithms that lead to consistent estimation in case of the absence of measurement errors, i.e., when the regressor is observed precisely.

Start with either the nonlinear regression model (1.1) and (1.5) or the model (1.1) and (1.14) with Berkson error. We can construct the naive estimator by ordinary least squares method using the objective function

$$Q_{\mathrm{OLS}}(y_1, x_1, \ldots, y_n, x_n; \beta) = \sum_{i=1}^{n} q_{\mathrm{OLS}}(y_i, x_i; \beta), \quad q_{\mathrm{OLS}}(y, x; \beta) = (y - f(x, \beta))^2 . \quad (1.29)$$

The corresponding naive estimator is as follows:

$$\hat{\beta}_{\mathrm{naive}} = \hat{\beta}_{\mathrm{OLS}} = \arg\min_{\beta \in \Theta} Q_{\mathrm{OLS}}(y_1, \ldots, x_n; \beta) . \quad (1.30)$$

If the set $\Theta$ is open in $\mathbf{R}^p$, then it is convenient to define the estimator by an equation (the so-called *normal equation*) instead of optimization (1.30). For this, we introduce the estimating function

$$s_{\mathrm{OLS}}(y, x; \beta) = -\frac{1}{2}\,\frac{\partial q_{\mathrm{OLS}}(y, x; \beta)}{\partial \beta} = (y - f(x, \beta))\frac{\partial f(x, \beta)}{\partial \beta} \tag{1.31}$$

and define the estimator $\tilde{\beta}_{\mathrm{OLS}}$ as one of the solutions to the equation

$$\sum_{i=1}^{n} s_{\mathrm{OLS}}(y_i, x_i; \beta) = 0\,, \quad \beta \in \Theta\,. \tag{1.32}$$

The estimators $\hat{\beta}_{\mathrm{OLS}}$ and $\tilde{\beta}_{\mathrm{OLS}}$ do not differ very much from each other. The naive researcher would follow this way if he/she knows nothing about the theory of measurement errors in covariates. The mentioned researcher just neglects the existence of such errors and constructs the estimator to be consistent when the errors are absent.

Now, consider the binary model (1.2) and (1.5), where the odds function $\lambda$ has the form either (1.3) or (1.4). Should $\{y_i, \xi_i, i = \overline{1, n}\}$ be observed the likelihood function will be equal to

$$L(\beta) = \prod_{i=1}^{n} \mathbf{P}^{y_i}\{y_i = 1|\xi_i\} \cdot \mathbf{P}^{1-y_i}\{y_i = 0|\xi_i\} = \prod_{i=1}^{n} \frac{\lambda_i^{y_i}}{1 + \lambda_i}\,, \tag{1.33}$$

with $\lambda_i = \lambda(\xi_i, \beta)$. The loglikelihood function is

$$l(\beta) = \sum_{i=1}^{n} (y_i \ln \lambda_i - \ln(1 + \lambda_i))\,. \tag{1.34}$$

The score function is

$$\begin{aligned} s_{\mathrm{OML}}(y, \xi; \beta) &= \frac{\partial}{\partial \beta}(y \ln \lambda(\xi, \beta) - \ln(1 + \lambda(\xi, \beta))) = \\ &= \left(\frac{y}{\lambda(\xi, \beta)} - \frac{1}{1 + \lambda(\xi, \beta)}\right)\frac{\partial \lambda(\xi, \beta)}{\partial \beta}\,. \end{aligned} \tag{1.35}$$

Here, the index "OML" hints at the ordinary maximum likelihood method. The usual maximum likelihood estimator is given by the equation

$$\sum_{i=1}^{n} s_{\mathrm{OML}}(y_i, \xi_i; \beta) = 0\,, \quad \beta \in \Theta. \tag{1.36}$$

Here $\Theta$ is an open parameter set in $\mathbf{R}^2$.

However, the measurement errors (1.5) (or (1.14)) being present, the regressors $\xi_i$ are unavailable. Using the observations $\{y_i, x_i, i = \overline{1, n}\}$, we can construct the naive estimator $\hat{\beta}_{\mathrm{naive}} = \hat{\beta}_{\mathrm{OML}}$ as one of the solutions to the system of equations

$$\sum_{i=1}^{n} s_{\mathrm{OML}}(y_i, x_i; \beta) = 0\,, \quad \beta \in \Theta. \tag{1.37}$$

For example, in the logistic model (1.2), (1.3), and (1.5) we have the following equations for the naive estimator:

$$\sum_{i=1}^{n} \left( y_i - \frac{1}{1 + \mathrm{e}^{-\beta_0 - \beta_1 x_i}} \right) \cdot \begin{bmatrix} 1 \\ x_i \end{bmatrix} = 0 \,, \quad \beta \in \Theta \,. \tag{1.38}$$

In the model (1.2), (1.4), and (1.5) with linear odds function, a counterpart of equations (1.38) is as follows:

$$\sum_{i=1}^{n} \left( \frac{y_i}{\beta_0 + \beta_1 x_i} - \frac{1}{1 + \beta_0 + \beta_1 x_i} \right) \cdot \begin{bmatrix} 1 \\ x_i \end{bmatrix} = 0 \,, \quad \beta \in \Theta \,. \tag{1.39}$$

In more general regression models, the conditional pdf $\rho_{y|\xi}(y_i, \xi_i; \beta)$ of $y_i$ given $\xi_i$ is defined. This is a density function of the conditional probability with respect to some measure $\nu$ on the real line. It contains the unknown regression parameter $\beta$. Then the score function is equal to

$$s_{\mathrm{OML}}(y_i, \xi_i; \beta) = \frac{\partial \ln \rho_{y|\xi}(y_i, \xi_i; \beta)}{\partial \beta} = \frac{1}{\rho_{y|\xi}} \, \frac{\partial \rho(y|\xi)}{\partial \beta} \,, \tag{1.40}$$

and the naive estimator is defined by (1.37) and (1.40).

Note that all the naive estimators being discussed in Section 1.4.1 fit into the overall scheme with the estimating function (1.40).

If, in the model (1.1), the errors $\varepsilon_i$ follow the normal distribution $\mathrm{N}(0, \sigma_\varepsilon^2)$, then $\rho_{y|\xi}(y_i, \xi_i, \beta)$ is the density of the normal law $\mathrm{N}(f(\xi_i, \beta), \sigma_\varepsilon^2)$ w.r.t. the Lebesgue measure on the real line, and equations (1.37) and (1.40) are reduced to the normal equations (1.32).

In the binary model (1.2), we have the conditional pdf

$$\rho_{y|\xi}(y_i, x_i; \beta) = \left( \frac{\lambda_i}{1 + \lambda_i} \right)^{y_i} \left( \frac{1}{1 + \lambda_i} \right)^{1 - y_i} \,, \quad \lambda_i = \lambda(\xi_i, \beta) \,. \tag{1.41}$$

This is the density of $y_i$ given $\xi_i$ w.r.t. the counting measure $\nu$ to be concentrated at the points 0 and 1, namely $\nu(A) = \mathbf{I}_A(0) + \mathbf{I}_A(1)$, $A \subset \mathbf{R}$. Herein $\mathbf{I}_A$ denotes an indicator function

$$\mathbf{I}_A(x) = 1 \,, \quad \text{if } x \in A \,, \quad \text{and } \mathbf{I}_A(x) = 0 \,, \quad \text{if } x \notin A \,. \tag{1.42}$$

Then the estimating equations (1.38) and (1.39) for the naive estimator correspond to the general score function (1.40).

The naive estimators $\hat{\beta}_{\mathrm{naive}}$ are inconsistent, as the sample size grows; they have a significant deviation $\hat{\beta}_{\mathrm{naive}} - \beta$ for moderate samples as well. However, this does not mean that they cannot be applied: usually they are evaluated by simple numerical procedures and under small measurement errors (i.e., when the variance of the errors is small), the naive estimators are quite accurate and can estimate better than the consistent ones.

### 1.4.2 Maximum likelihood estimator (MLE)

In this section, we consider only the *structural* model with errors in the regressor. Let the regression of the response $y$ on the scalar covariate $\xi$ be given by the conditional pdf $\rho(y|\xi) = \rho(y|\xi;\beta)$ w.r.t. some measure $\nu$ on real line and the covariate $\xi$ is random and has the pdf $\rho(\xi) = \rho(\xi;\gamma)$, with nuisance parameter $\gamma$.

**A model with Berkson error**

Suppose we have the Berkson measurement error, i.e., $x$ is measured instead of $\xi$ and

$$\xi = x + \delta^{\mathrm{B}}\,, \tag{1.43}$$

the random vector $(y, x)^{\mathrm{T}}$ and $\delta^{\mathrm{B}}$ are stochastically independent, and the probability law of the Berkson error is known, namely, we know the pdf $\rho(\delta^{\mathrm{B}})$. Consider the independent copies of model $(y_i, x_i, \delta_i^{\mathrm{B}}, \xi_i)$, $i = \overline{1,n}$ (i.e., all these sets are independent and have the same distribution as the set $(y, x, \delta^{\mathrm{B}}, \xi)$ from the model). By the observations $\{(y_i, x_i), i = \overline{1,n}\}$, we estimate the model parameters.

Construct the joint pdf of observed variables

$$\rho(y, x;\beta) = \int\limits_{\mathbf{R}} \rho(y|\xi)\big|_{\xi=x+\delta^{\mathrm{B}}} \cdot \rho(\delta^{\mathrm{B}})\, d\delta^{\mathrm{B}}. \tag{1.44}$$

Then the score function is

$$s_{\mathrm{ML}}(y, x;\beta) = \frac{\partial \ln \rho(y, x;\beta)}{\partial \beta} = \frac{1}{\rho}\,\frac{\partial \rho}{\partial \beta}\,. \tag{1.45}$$

In case $\Theta$ is open, the MLE satisfies the equation

$$\sum_{i=1}^{n} s_{\mathrm{ML}}(y_i, x_i;\beta) = 0\,, \quad \beta \in \Theta. \tag{1.46}$$

We note that often the integral (1.44) is not calculated in the closed form, e.g., this is the case in binary models. Therefore, evaluating the estimating function (1.45) in the observable points with varying $\beta$ can be a daunting problem of numerical methods. For approximate calculation of the integral (1.44), we can apply the Monte Carlo method.

**A model with classical error**

Let the relationship between $y$ and $\xi$ be such as described above (in particular, the conditional pdf $\rho_{y|\xi}(y_i, \xi_i;\beta)$ is given, see Section 1.4.1), but instead of the Berkson error (1.43), we have a classical measurement error

$$x = \xi + \delta\,, \tag{1.47}$$

such that random vector $(y, \xi)^{\mathrm{T}}$ and $\delta$ are stochastically independent and the pdf $\rho(\delta)$ of the classical error is known. Consider independent copies of the model $(y_i, \xi_i, \delta_i, x_i)$, $i = \overline{1, n}$. By the observations $\{(y_i, x_i), i = \overline{1, n}\}$, we estimate the model parameters.

The joint pdf $\rho(y, x; \beta)$ is written more complicate than in (1.44) and the nuisance parameter $\gamma$ is included in the pdf $\rho(\xi; \gamma)$:

$$\rho(y, x; \beta, \gamma) = \int\limits_{\mathbf{R}} \rho(y|\xi)\big|_{\xi=x-\delta} \cdot \rho(x|\delta) \cdot \rho(\delta)\, d\delta\,. \tag{1.48}$$

Because $\xi$ and $\delta$ are independent, we have

$$\rho(x|\delta) = \rho(\xi)\big|_{\xi=x-\delta}\,. \tag{1.49}$$

For example, if $\xi \sim \mathrm{N}(\mu_\xi, \sigma_\xi^2)$, then the conditional law $x|\delta \sim \mathrm{N}(\mu_\xi + \delta, \sigma_\xi^2)$, thus

$$\rho(x|\delta) = \frac{1}{\sqrt{2\pi}\sigma_\xi} \mathrm{e}^{-\frac{(x-\mu_X-\delta)^2}{2\sigma_\xi^2}} = \frac{1}{\sqrt{2\pi}\sigma_\xi} \mathrm{e}^{-\frac{(\xi-\mu_\xi)^2}{2\sigma_\xi^2}}\Bigg|_{\xi=x-\delta} = \rho(\xi)\big|_{\xi=x-\delta}\,. \tag{1.50}$$

Relations (1.48) and (1.49) imply

$$\rho(y, x; \beta, \gamma) = \int\limits_{\mathbf{R}} \rho(y|\xi; \beta)\big|_{\xi=x-\delta} \cdot \rho(\xi; \gamma)\big|_{\xi=x-\delta} \cdot \rho(\delta)\, d\delta\,. \tag{1.51}$$

The score function is

$$s_{\mathrm{ML}}(y, x; \beta, \gamma) = \frac{1}{\rho} \frac{\partial \rho}{\partial(\beta, \gamma)}\,, \quad \rho = \rho(y, x; \beta, \gamma)\,. \tag{1.52}$$

Here, the partial derivatives are taken in both $\beta$ and $\gamma$. The MLE $\hat{\beta}_{\mathrm{ML}}, \hat{\gamma}_{\mathrm{ML}}$ is a solution to the system of equations

$$\sum_{i=1}^{n} s_{\mathrm{ML}}(y_i, x_i; \beta, \gamma) = 0\,, \quad \beta \in \Theta\,, \quad \gamma \in \Theta_\gamma\,. \tag{1.53}$$

Here, $\Theta_\gamma$ is a parameter set for the parameter $\gamma$.

As we see, the expression (1.51) for the joint pdf is quite complicated, making it problematic to use this method in models other than linear. For example, in the model (1.1) and (1.5) with $\xi_i \sim \mathrm{N}(\mu_\xi, \sigma_\xi^2)$, $\varepsilon_i \sim \mathrm{N}(0, \sigma_\varepsilon^2)$, $\delta_i \sim \mathrm{N}(0, \sigma_\delta^2)$ under known $\sigma_\delta^2$ and $\sigma_\varepsilon^2$, it holds that

$$\rho(y, x; \beta, \gamma) = \int\limits_{\mathbf{R}} \frac{1}{\sqrt{2\pi}\sigma_\varepsilon} \mathrm{e}^{-\frac{(y-f(x-\delta))^2}{2\sigma_\varepsilon^2}} \times \frac{1}{\sqrt{2\pi}\sigma_\xi} \mathrm{e}^{-\frac{(x-\delta-\mu_\xi)^2}{2\sigma_\xi^2}} \times \frac{1}{\sqrt{2\pi}\sigma_\delta} \mathrm{e}^{-\frac{\delta^2}{2\sigma_\delta^2}}\, d\delta\,. \tag{1.54}$$

Here $\gamma = (\mu_\xi, \sigma_\xi^2)^{\mathrm{T}}$.

In the linear model, $f(\xi) = f(\xi, \beta) = \beta_0 + \beta_1 \xi$ and the integral is calculated in the closed form, but it is not the case in the incomplete quadratic model $f(\xi, \beta) = \beta \xi^2$.

So far, besides the computer challenges are still taking place, it is unknown whether the loglikelihood function $l = \ln \rho$ has regularity properties, which provide the following well-known good features of the maximum likelihood method:

(a) efficiency (realization of the Cramér–Rao inequality), and
(b) consistency of the estimator as the sample size grows.

Since the existence of moments for $\hat{\beta}_{\mathrm{ML}}$ is not guaranteed, the efficiency should be replaced by asymptotic efficiency. That is, we are looking for the smallest asymptotic covariance matrix (ACM) of the estimator $\hat{\beta}_{\mathrm{ML}}$ within a quite broad class of estimators. However, even this asymptotic efficiency for $\hat{\beta}_{\mathrm{ML}}$ is not at all guaranteed.

In the end, we want to draw a conclusion: in models with the Berkson error, the Maximum Likelihood method is worth realizing, but in nonlinear models with classical measurement error, using the method seems doubtful. We will see that in the latter models, much more reliable estimation methods are developed.

### 1.4.3 Quasi-likelihood estimator (QLE)

**Mean–variance model**

Consider the general structural regression model described at the beginning of Section 1.4.2 with the vector regressor $\xi$ and the classical error (1.47). One can add conditional mean and conditional variance of $y$ given $\xi$:

$$\mathbf{E}(y|\xi) = m^*(\xi, \beta) = \int_{\mathbf{R}} y \, \rho(y|\xi; \beta) \, dy \,, \tag{1.55}$$

$$\mathbf{V}(y|\xi) = v^*(\xi, \beta) = \int_{\mathbf{R}} (y - m^*(\xi, \beta))^2 \, \rho(y|\xi; \beta) \, dy \,. \tag{1.56}$$

Now, consider the conditional mean of $y$ given the observable variable $x$:

$$m(x, \beta) = \mathbf{E}(y|x) = \mathbf{E}[\mathbf{E}(y|x, \xi)|x] \,. \tag{1.57}$$

Classical error in the model (1.47) is *nondifferentiable*, so it means that under $x$ and $\xi$ known, only $\xi$ alone contains all the information about the response $y$ (this is a consequence of the independence of the couple $(y, \xi)$ from $\delta$). Then

$$\mathbf{E}[y|x, \xi] = \mathbf{E}(y|\xi) = m^*(\xi, \beta) \,, \tag{1.58}$$

$$m(x, \beta) = \mathbf{E}[m^*(\xi, \beta)|x] = \int_{\mathbf{R}^d} m^*(\xi, \beta) \, \rho(\xi|\beta) \, d\xi \,. \tag{1.59}$$

Here $\rho(\xi|x)$ is the conditional pdf of $\xi$ given $x$ (being evaluated at the point $\xi$). Later on, we will consider how to find it.

Pass to the conditional variance

$$v(x,\beta) = \mathbf{V}(y|x) = \mathbf{E}[(y - m(x,\beta))^2|x] \,. \tag{1.60}$$

Hereafter, we write briefly $m^*(\xi) = m^*(\xi,\beta)$, $m(x) = m(x,\beta)$ and apply that the classical error is indifferentiable:

$$v(x,\beta) = \mathbf{E}\{\mathbf{E}[(y - m(x))^2|x,\xi]|x\} \,, \tag{1.61}$$

$$\begin{aligned} v_1 &= \mathbf{E}[(y - m(x))^2|x,\xi] = \mathbf{E}[(y - m^*(\xi) + m^*(\xi) - m(x))^2|x,\xi] = \\ &= \mathbf{E}[(y - m^*(\xi))^2|\xi] + \mathbf{E}[(m^*(\xi) - m(x))^2|x,\xi] = \\ &= \mathbf{V}(y|\xi) + \mathbf{E}[(m^*(\xi) - m(x))^2|x,\xi] \,. \end{aligned} \tag{1.62}$$

We made use of the fact that $y - m^*(\xi)$ and $m^*(\xi) - m(x)$ are conditionally uncorrelated:

$$\begin{aligned} &\mathbf{E}[(y - m^*(\xi))(m^*(\xi) - m(x))|x,\xi] \\ &= (m^*(\xi) - m(x))\ \mathbf{E}[(y - m^*(\xi))|x,\xi] \\ &= (m^*(\xi) - m(x))\ \mathbf{E}[(y - m^*(\xi))|\xi] = 0 \,. \end{aligned} \tag{1.63}$$

Further, by the tower property of conditional expectations (Kartashov, 2007)

$$v(x,\beta) = \mathbf{E}(v_1|x) = \mathbf{E}[v^*(\xi)|x] + \mathbf{E}[(m^*(\xi) - m(x))^2|x] \,. \tag{1.64}$$

Finally, we have a remarkable equality

$$v(x,\beta) = \mathbf{V}(y|x) = \mathbf{E}[\mathbf{V}(y|\xi)|x] + \mathbf{V}[\mathbf{E}(y|\xi)|x] \,. \tag{1.65}$$

In the last step, we used the fact that

$$\mathbf{E}[m^*(\xi)|x] = \mathbf{E}[\mathbf{E}(y|\xi)|x] = \mathbf{E}[\mathbf{E}(y|x,\xi)|x] = \mathbf{E}(y|x) = m(x) \,. \tag{1.66}$$

Using the conditional distribution of $\xi$ given $x$, one can rewrite the conditional variance as follows:

$$v(x,\beta) = \int\limits_{\mathbf{R}^d} v^*(\xi,\beta)\ \rho(\xi|x)\ d\xi + \int\limits_{\mathbf{R}^d} (m^*(\xi) - m(x))^2 \rho(\xi|x)\ d\xi \,. \tag{1.67}$$

The functions (1.59) and (1.67) designate the so-called *mean–variance model*, see Cheng and Van Ness (1999) and Wansbeek and Meijer (2000).

In this model, the regression of $y$ on the observable covariate $x$ can be written as

$$y = m(x,\beta) + \varepsilon \,, \tag{1.68}$$

where $\varepsilon = \varepsilon(x,y)$ plays a role of error, such that

$$\mathbf{E}(\varepsilon|x) = 0 \,, \quad \mathbf{V}(\varepsilon|x) = v(x,\beta) \,, \quad \mathbf{E}[m(x,\beta)\,\varepsilon|x] = 0 \,, \tag{1.69}$$

i.e., the error is conditionally uncorrelated with the new regression function $m(x,\beta)$, moreover the error is conditionally centered. Such error is also called conditionally unbiased and its conditional variance depends on a new covariate $x$ and unknown parameter $\beta$. The idea to transform the structural errors-in-variables model to the form (1.68) and (1.69) is due to Gleser (1990).

**Construction of the estimator**

For the model (1.68) and (1.69) (which is a consequence of initial regression model with classical error), the quasi-likelihood estimator $\hat{\beta}_{\mathrm{QL}}$ is constructed (Carroll et al., 2006). For this purpose, an estimating function is introduced

$$s_{\mathrm{QL}}(y, x; \beta) = \frac{y - m(x, \beta)}{v(x, \beta)} \cdot \frac{\partial m(x, \beta)}{\partial \beta} , \tag{1.70}$$

and the estimator $\hat{\beta}_{\mathrm{QL}}$ is defined as one of the solutions to the following equation:

$$\sum_{i=1}^{n} s_{\mathrm{QL}}(y_i, x_i; \beta) = 0 , \quad \beta \in \Theta . \tag{1.71}$$

Let us explain why the estimating function $s_{\mathrm{QL}}$ is selected in such a way. In the regression model (1.68) and (1.69), it seems natural to use the weighted least squares method, i.e., to take the objective function

$$q_{\mathrm{WLS}}(y, x; \beta) = \frac{(y - m(x, \beta))^2}{v(x, \beta)} \tag{1.72}$$

and define the estimator $\hat{\beta}_{\mathrm{WLS}}$ as a minimum point of the function

$$Q_{\mathrm{WLS}}(y_1, \ldots, x_n; \beta) = \sum_{i=1}^{n} q_{\mathrm{WLS}}(y_i, x_i; \beta) , \quad \beta \in \Theta . \tag{1.73}$$

If the set $\Theta$ is open, then one can consider the estimating function

$$s_{\mathrm{WLS}}(y, x; \beta) = -\frac{1}{2} \frac{\partial q_{\mathrm{WLS}}}{\partial \beta} = \frac{y - m(x, \beta)}{v(x, \beta)} \frac{\partial m(x, \beta)}{\partial \beta} + (y - m(x, \beta))^2 \frac{\partial}{\partial \beta} \left( \frac{1}{v} \right) \tag{1.74}$$

and define the estimator $\tilde{\beta}_{\mathrm{WLS}}$ as a root of the equation

$$\sum_{i=1}^{n} s_{\mathrm{WLS}}(y_i, x_i; \beta) = 0 . \tag{1.75}$$

However, the estimating function will be biased (see discussion of unbiasedness in Appendix A1). Indeed, we have

$$\mathbf{E}_\beta s_{\mathrm{WLS}}(y, x; \beta) = \mathbf{E}_\beta \left( \frac{y - m(x, \beta)}{v} \frac{\partial m(x, \beta)}{\partial \beta} \right) - \mathbf{E}_\beta \left[ \frac{(y - m(x, \beta))^2}{v^2} \frac{\partial v}{\partial \beta} \right] . \tag{1.76}$$

Further, the first summand is

$$\mathbf{E}\mathbf{E}_\beta \left( \frac{y - m(x, \beta)}{v} \cdot \frac{\partial m(x, \beta)}{\partial \beta} \,\middle|\, x \right) = \mathbf{E} \left[ \mathbf{E}_\beta [(x - m(x, \beta) | x] \cdot \frac{\partial m}{\partial \beta} \right] = 0 , \tag{1.77}$$

and the second one is

$$\mathbf{E}\mathbf{E}_\beta \left( \frac{(y - m(x, \beta))^2}{v^2} \cdot \frac{\partial v}{\partial \beta} \,\middle|\, x \right) = \mathbf{E} \left[ \mathbf{E}_\beta [(y - m(x, \beta))^2 | x] \cdot \frac{1}{v^2} \cdot \frac{\partial v}{\partial \beta} \right] = \mathbf{E} \left( \frac{1}{v} \frac{\partial v}{\partial \beta} \right) . \tag{1.78}$$

Thus,

$$\mathbf{E}_\beta s_{\mathrm{WLS}}(y, x; \beta) = -\mathbf{E}\left( \frac{1}{v(x, \beta)} \frac{\partial v}{\partial \beta} \right) \neq 0 \,. \tag{1.79}$$

The latter is true actually for nonlinear models wherein the conditional variance $v = \mathbf{V}(y|x)$ does depend on $\hat{\beta}_{\mathrm{WLS}}$. Biasedness (1.79) implies the inconsistency of the estimator $\hat{\beta}_{\mathrm{WLS}}$ (see Appendix A).

At the same time, equality (1.77) states unbiasedness of the estimating function $s_{\mathrm{QL}}$ (see (1.70)) that under regularity conditions ensures the consistency of the estimator $\hat{\beta}_{\mathrm{QL}}$. We obtain the estimating function $s_{\mathrm{QL}}$ from $s_{\mathrm{WLS}}$ if we neglect the dependence of $v$ on the parameter $\beta$, and then the relations $\partial v/\partial \beta \approx 0$ and $s_{\mathrm{WLS}} \approx s_{\mathrm{QL}}$ hold. One can see that this neglect is productive.

**Explanation of term**

The term *quasi-likelihood* is a hint at the likelihood function. To clarify this, consider an idealized situation where the conditional distribution of $y$ given $x$ is normal,

$$y|x \sim \mathrm{N}(m, v) \,, \quad m = m(x, \beta) \,, \quad v = v(x, \beta) \,. \tag{1.80}$$

Then

$$\rho(y|x) = \frac{1}{\sqrt{2\pi v}} \mathrm{e}^{-\frac{(y-m)^2}{2v}} \,, \tag{1.81}$$

$$\ln \rho(y|x) = -\frac{(y-m)^2}{2v} - \frac{1}{2} \ln v + \mathrm{const}, \tag{1.82}$$

and the score function is

$$s_{\mathrm{ML}}(x, \beta) = \frac{\partial}{\partial \beta} \rho(y|x) = s_{\mathrm{QL}} + \frac{1}{2} \left( \frac{(y-m)^2}{v^2} - \frac{1}{v} \right) \frac{\partial v}{\partial \beta} \,. \tag{1.83}$$

If the $s_{\mathrm{QL}}$ is a linear function in the response $y$, then the extra summand in (1.83) is a quadratic function in $y$. This quadratic term is unbiased

$$\mathbf{E}_\beta \left[ \left( \frac{(y-m)^2}{v^2} - \frac{1}{v} \right) \frac{\partial v}{\partial \beta} \right] = 0 \,. \tag{1.84}$$

If we delete it from $s_{\mathrm{ML}}$, then we come to the estimating function $s_{\mathrm{QL}}$. We can state that in the normal case (1.80), the estimating function $s_{\mathrm{QL}}$ almost surely coincides with the score function $s_{\mathrm{ML}}$, namely, they differ by a quadratic term which is small if the conditional variance $v$ depends slightly on $\beta$. In this sense, $s_{\mathrm{QL}}$ is almost surely the estimating function of the MLE (for the normal case (1.80)).

**Finding distribution of $\xi$ given $x$**

To use the formulas (1.59) and (1.67), one should know $\rho(\xi|x)$. In some cases, this conditional pdf can be found explicitly.

Let $\xi$ and $\delta$ be independent scalar variables and

$$x = \xi + \delta\,, \quad \xi \sim \mathrm{N}(\mu_\xi, \sigma_\xi^2)\,, \quad \delta \sim \mathrm{N}\left(0, \sigma_\delta^2\right)\,. \tag{1.85}$$

Then the conditional distribution of $\xi$ given $x$ will be normal as well (Anderson, 2003):

$$\xi|x \sim \mathrm{N}(\mu_1(x), \tau^2)\,, \quad \mu_1(x) = Kx + (1-K)\mu_\xi\,, \quad \tau^2 = K\sigma_\delta^2\,. \tag{1.86}$$

Here $K$ is the so-called *reliability ratio* in the classical linear measurement error model

$$K = \frac{\sigma_\xi^2}{\mathbf{D}x} = \frac{\sigma_\xi^2}{\sigma_\xi^2 + \sigma_\delta^2}\,. \tag{1.87}$$

For the sake of completeness, we present a simple proof of the relations (1.86) and (1.87).

We have

$$\rho(\xi|x) = \frac{\rho(\xi, x)}{\int_{\mathbf{R}} \rho(\xi, x)\, d\xi}\,. \tag{1.88}$$

Write the joint pdf

$$\rho(\xi, x) = \rho(x|\xi)\,\rho(\xi) = \frac{1}{\sqrt{2\pi}\sigma_\delta} \mathrm{e}^{-\frac{(x-\xi)^2}{2\sigma_\delta^2}} \times \frac{1}{\sqrt{2\pi}\sigma_\xi} \mathrm{e}^{-\frac{(\xi-\mu_\xi)^2}{2\sigma_\xi^2}}\,, \tag{1.89}$$

$$\rho(\xi, x) = \exp\left\{-\left(\frac{\xi^2}{2\tau^2} - A(x)\xi\right)\right\} \times C(x)\,. \tag{1.90}$$

Here

$$\frac{1}{\tau^2} = \frac{1}{\sigma_\delta^2} + \frac{1}{\sigma_\xi^2}\,, \quad \tau^2 = \frac{\sigma_\xi^2\sigma_\delta^2}{\sigma_\xi^2 + \sigma_\delta^2} = K\sigma_\delta^2\,, \tag{1.91}$$

where $K$ is the reliability ratio (1.87). Further

$$A(x) = \frac{x}{\sigma_\delta^2} + \frac{\mu_\xi}{\sigma_\xi^2}\,, \quad \rho(\xi, x) = C_1(x) \exp\left\{-\frac{(\xi - \tau^2 A(x))^2}{2\tau^2}\right\}\,; \tag{1.92}$$

$$\tau^2 A(x) := \mu_1(x) = \frac{\sigma_\xi^2}{\sigma_\xi^2 + \sigma_\delta^2} x + \frac{\sigma_\delta^2}{\sigma_\xi^2 + \sigma_\delta^2}\mu_\xi\,, \tag{1.93}$$

$$\mu_1(x) = K\,x + (1-K)\,\mu_\xi\,. \tag{1.94}$$

Then

$$\rho(\xi, x) = C_2(x) \frac{1}{\sqrt{2\pi}\tau} \mathrm{e}^{-\frac{(\xi-\mu_1(x))^2}{2\tau^2}}\,. \tag{1.95}$$

It immediately follows that $\int_{\mathbf{R}} \rho(\xi, x)\, d\xi = C_2(x)$ and taking into account (1.88),

$$\rho(\xi|x) = \frac{1}{\sqrt{2\pi}\tau} \mathrm{e}^{-\frac{(\xi-\mu_1(x))^2}{2\tau^2}}\,. \tag{1.96}$$

This proves the desired relationships (1.86) and (1.87).

We can interpret the conditional distribution (1.86) in the following way. If one observes $x = \xi + \delta$, see (1.85), then how to estimate the latent variable $\xi$? The Bayes estimator of $\xi$ is a random variable $\hat{\xi} \sim \mathrm{N}(\mu_1(x), \tau^2)$, or

$$\hat{\xi} = \mu_1(x) + \tau\gamma\,, \quad \gamma \sim \mathrm{N}(0,1)\,, \quad x \perp\!\!\!\perp \gamma\,. \tag{1.97}$$

Hereafter, the symbol $\perp\!\!\!\perp$ denotes the stochastic independence.

As a point estimator of $\xi$ by a single observation $x$, we will take

$$\mathbf{E}(\xi|x) = \mu_1(x) = Kx + (1-K)\mu_\xi\,. \tag{1.98}$$

Interestingly, this estimator takes the form of a convex combination a prior estimator $\mu_\xi$ (it is a natural estimator because $\mathbf{E}\xi = \mu_\xi$) and the observed value $x$ (in a sense it is close to $\xi$, because $\mathbf{E}\xi = \mathbf{E}x$). If $K$ is close to 1, then $\sigma_\delta^2$ is small, we really trust in our observation and $\mu_1(x) \approx x$ takes place. If $K$ is close to 0, then $\sigma_\delta^2$ is large making the observation unreliable, and $\mu_1(x) \approx \mu_\xi$. In an intermediate situation, the point estimator of $\xi$ ranges between two limit values $x$ and $\mu_\xi$.

The value (1.97) allows us to rewrite the formulas for $m(x,\beta)$ and $v(x,\beta)$ in a more compact form

$$m(x,\beta) = \mathbf{E}[m^*(\mu_1(x) + \tau\gamma;\ \beta)|x]\,, \tag{1.99}$$

$$v(x,\beta) = \mathbf{E}[v^*(\mu_1(x) + \tau\gamma;\ \beta)|x] + \mathbf{E}[(m^*(\mu_1(x) + \tau\gamma;\ \beta) - m(x,\beta))^2|x]\,. \tag{1.100}$$

In fact, we take the expectations w.r.t. $\gamma \sim \mathrm{N}(0,1)$; if necessary, they can be approximately evaluated by the Monte Carlo method.

#### Conditional distribution of $\xi$ given $x$: generalizations

The relation (1.86) can be extended to the case of a vector regressor (the so-called *multiple regression*). Let the regressor $\xi$ be distributed in $\mathbf{R}^d$,

$$x = \xi + \delta\,, \quad \xi \sim \mathrm{N}(\mu_\xi, \Sigma_\xi)\,, \quad \delta \sim \mathrm{N}(0, \Sigma_\delta)\,, \quad \xi \perp\!\!\!\perp \delta\,. \tag{1.101}$$

Here, $\mu_\xi \in \mathbf{R}^d$, $\Sigma_\xi$, and $\Sigma_\delta$ are the covariance matrices. Then we have the conditional distribution

$$\xi|x \sim \mathrm{N}(\mu_1(x), T)\,, \tag{1.102}$$

$$\mu_1(x) = K\,x + (I-K)\,\mu_\xi\,, \quad T = K\Sigma_\delta\,, \quad K = \Sigma_\xi\,(\Sigma_\xi + \Sigma_\delta)^{-1}\,. \tag{1.103}$$

The latter matrix is an analog of the reliability ratio for the multiple model with the classical measurement error.

Another generalization of the formulas (1.86) deals with a model in which the scalar latent variable is distributed according to a mixture of normal laws. Let

$$x = \xi + \delta\,, \quad \xi \sim \sum_{i=1}^{N} p_i\,\mathrm{N}(\mu_i, \sigma_{\xi,i}^2)\,, \quad \delta \sim \mathrm{N}(0, \sigma_\delta^2)\,, \quad \xi \perp\!\!\!\perp \delta. \tag{1.104}$$

Here, $\{p_i, i = \overline{1, N}\}$ is a full set of positive probabilities and the distribution of $\xi$ is a mixture of $N$ normal laws. We can interpret this as follows. We have $N$ classes of objects $A_1, \ldots, A_N$ characterized by pdfs $\rho_i(\xi) \sim \mathrm{N}(\mu_i, \sigma^2_{\xi,i})$, $i = \overline{1, N}$. The number $p_i$ is a prior probability that the object $\xi$ belongs to the class $A_i$. Then the unconditional pdf of $\xi$ is $\rho(\xi) = \sum_{i=1}^N p_i \rho_i(\xi)$.

After obtaining the observation $x$, we have a posterior probability of membership $\xi$ to the classes $A_1, \ldots, A_N$

$$q_i(x) = \frac{p_i \rho_i(x)}{\sum_{j=1}^N p_j \rho_j(x)}, \quad i = \overline{1, N}. \tag{1.105}$$

Here, $\rho_i(x)$ is the pdf of $x$ provided $\xi$ belongs to the class $A_i$, $\rho_i(x) \sim \mathrm{N}(\mu^{(i)}_\xi, \sigma^2_{\xi,i} + \sigma^2_\delta)$. Then

$$\xi | x \sim \sum_{i=1}^N q_i(x) \times (\xi | x, A_i). \tag{1.106}$$

Here, $(\xi | x, A_i)$ is the distribution of $\xi$ provided that $\xi$ belongs to $A_i$ and one gets the observation $x$. According to (1.86) we have:

$$(\xi | x, A_i) \sim \mathrm{N}(\mu_i(x), \tau_i^2), \quad i = \overline{1, N}, \tag{1.107}$$

$$\mu_i(x) = K_i x + (1 - K_i)\mu^{(i)}_\xi, \quad \tau_i^2 = K_i \sigma^2_\delta, \quad K_i = \frac{\sigma^2_{\xi,i}}{\sigma^2_{\xi,i} + \sigma^2_\delta}. \tag{1.108}$$

Therefore,

$$\rho(\xi, x) \sim \sum_{i=1}^N q_i(x) \rho(\xi | x, A_i), \quad \rho(\xi | x, A_i) \sim \mathrm{N}(\mu_i(x), \tau_i^2). \tag{1.109}$$

The conditional distribution of $\xi$ given $x$ coincides with the distribution of random variable

$$\sum_{i=1}^N I_i(x)(\mu_i(x) + \tau_i \gamma), \quad \gamma \sim \mathrm{N}(0, 1), \tag{1.110}$$

where $I_i(x) = 1$ with probability $q_i(x)$ and $I_i(x) = 0$ with probability $1 - q_i(x)$, $\sum_{i=1}^N I_i(x) = 1$ and $\gamma$ is stochastically independent of both $x$ and a set of indicators $\{I_i(x), i = \overline{1, N}\}$.

The sum (1.110) can be generated as follows: among the numbers from 1 to N select a random number $I = I(x)$ with posterior probabilities $q_1(x), \ldots, q_N(x)$, and then evaluate $\mu_I(x) + \tau_I \gamma$, where $\gamma \sim \mathrm{N}(0, 1)$ and $\gamma$ is independent of the couple $(x;\ I(x))$. As a result $\xi | x \sim \mu_I(x) + \tau_I \gamma$.

If $\mathbf{E}(f(\xi)|x)$ is to be computed, we will have

$$\mathbf{E}(f(\xi)|x) = \int_{\mathbf{R}} f(\xi)\rho(\xi|x)\, d\xi = \sum_{i=1}^N q_i(x) \int_{\mathbf{R}} f(\xi)\rho(\xi|x, A_i)\, d\xi =$$

$$= \sum_{i=1}^N q_i(x)\mathbf{E}[f(\mu_i(x) + \tau_i \gamma)|x], \quad \gamma \sim \mathrm{N}(0, 1), \quad \gamma \perp\!\!\!\perp x. \tag{1.111}$$

This is helpful for calculation by formulas (1.59) and (1.67).

**Pre-estimation of distribution parameters of $\xi$**

Consider the structural model described at the beginning of Section 1.4.2, with the classical error (1.47). Let the distribution of measurement error $\delta$ be known and the distribution of $\xi$ be known up to the vector parameter $\gamma$. The estimating function $s_{\mathrm{QL}}$ presented in (1.70) contains the conditional mean and variance $m$ and $v$ that depend not only on the regression parameter $\beta$, but also on the parameter $\gamma$. To derive the QLE, it is possible to pre-estimate $\gamma$ by the MLM based on the observations $\{x_i, i = \overline{1, N}\}$ and then substitute the obtained estimator $\hat{\gamma}_{\mathrm{ML}}$ in formula (1.70). Thus, we get the estimating function

$$\hat{s}_{\mathrm{QL}}(x, y;\, \beta) = \frac{y - \hat{m}(x, \beta)}{\hat{v}(x, \beta)} \cdot \frac{\partial \hat{m}(x, \beta)}{\partial \beta}\,, \tag{1.112}$$

$$\hat{m}(x, \beta) = m(x; \beta, \hat{\gamma}_{\mathrm{ML}})\,, \quad \hat{v}(x, \beta) = v(x; \beta, \hat{\gamma}_{\mathrm{ML}})\,. \tag{1.113}$$

Then the estimator $\hat{\beta}_{\mathrm{QL}}$ constructed using the pre-estimation is one of the solutions to the equation

$$\sum_{i=1}^{n} \hat{s}_{\mathrm{QL}}(y_i, x_i; \beta) = 0\,, \quad \beta \in \Theta\,. \tag{1.114}$$

The estimating function $\hat{s}_{\mathrm{QL}}$ is asymptotically unbiased, i.e.,

$$\mathbf{E}_\beta \hat{s}_{\mathrm{QL}}(y, x; \beta) \xrightarrow{P} 0\,, \quad \text{as } n \to \infty\,. \tag{1.115}$$

Hereafter, $\to^P$ means convergence in probability. The convergence (1.115) holds, because the estimating function (1.70) (under known $\gamma$) is unbiased and the estimator $\hat{\gamma}_{\mathrm{ML}}$ is a consistent estimator of $\gamma$. The convergence (1.115) under mild regularity conditions implies the consistency of $\hat{\beta}_{\mathrm{QL}}$ being the root of equation (1.114).

In particular, the normal model (1.85) makes use of the nuisance parameter $\gamma = (\mu_\xi; \sigma_\xi^2)^{\mathrm{T}}$. The MLE of $\gamma$ takes the form

$$\hat{\mu}_{\xi,\mathrm{ML}} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,, \quad \hat{\sigma}^2_{\xi,\mathrm{ML}} = \overline{(x - \hat{\mu}_{\xi,\mathrm{ML}})^2} - \sigma_\delta^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 - \sigma_\delta^2\,. \tag{1.116}$$

Hereafter, the bar means the arithmetic mean being calculated by the observed sample. Instead of the estimator $\hat{\sigma}^2_{\xi,\mathrm{ML}}$, it is better to use an unbiased modification

$$\hat{\sigma}_\xi^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 - \sigma_\delta^2\,, \quad n \geq 2\,. \tag{1.117}$$

Here, the summand $\hat{\sigma}_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is unbiased estimator of the variance of $x$.

### 1.4.4 Corrected score (CS) method

Methods considered in Sections 1.4.2 and 1.4.3 are *structural*, i.e., they work in the structural models and utilize information about the shape of distribution of the latent

variable $\xi$. In this section, we describe the *functional* method, which can be used in the functional models. This method is suitable for the structural models as well, in situations where the shape of the distribution of $\xi$ is unknown.

Suppose we have a functional regression model with the classical error (1.47). Remember that in this model, the latent variable $\xi_i$ from $\mathbf{R}^d$ is nonrandom, $i = \overline{1, n}$. The idea of the corrected score (CS) method is as follows.

Start with some unbiased estimating function $s(y, \xi; \beta)$ providing consistent estimation of $\beta$ in the absence of the measurement error in the regressor. Next, the corrected estimating function $s_{\mathrm{C}}(y, x; \beta)$ is constructed such that for all $b \in \Theta$,

$$\mathbf{E}[s_{\mathrm{C}}(y, x; b)|y, \xi] = s(y, \xi; b) \,. \tag{1.118}$$

We define the CS estimator $\hat{\beta}_{\mathrm{C}}$ as one of the solutions to the folowing equation:

$$\sum_{i=1}^{n} s_{\mathrm{C}}(y_i, x_i, \beta) = 0 \,, \quad \beta \in \Theta \,. \tag{1.119}$$

The corrected estimating function is unbiased because

$$\mathbf{E}_\beta s_{\mathrm{C}}(y, x; \beta) = \mathbf{E}_\beta \mathbf{E}(s_{\mathrm{C}}(y, x; \beta)|y, \xi) = \mathbf{E}_\beta s(y, \xi; \beta) = 0 \,, \tag{1.120}$$

and therefore, under some regularity conditions the estimator $\hat{\beta}_{\mathrm{C}}$ is consistent. For the first time this method was applied by Stefanski (1989) and Nakamura (1990).

Equation (1.118) with unknown function $s_{\mathrm{C}}$ is called *deconvolution equation*. If $\rho(\delta)$ is the pdf of the error $\delta$, then (1.118) can be written as

$$\int\limits_{\mathbf{R}^d} s_{\mathrm{C}}(y, \xi + \delta;\ b)\, \rho(\delta)\, d\delta = s(y, \xi; b) \,, \quad y \in \mathbf{R} \,, \quad \xi \in \mathbf{R}^d \,, \quad b \in \Theta \,. \tag{1.121}$$

The left-hand side of (1.121) is a convolution of the desired function $s_{\mathrm{C}}(y, x; \beta)$ and $\rho(-x)$ with respect to $x$. Here, $\rho(-x)$ is the pdf of $-\delta$. That is why equation (1.121) is called deconvolution equation, i.e., $s_{\mathrm{C}}$ is defined by the inverse operation of the convolution operation.

If we make the linear replacement $x = \xi + \delta$ in the integral (1.121), then (1.121) takes the form:

$$\int\limits_{\mathbf{R}^d} s_{\mathrm{C}}(y, x; b)\, \rho(x - \xi)\, dx = s(y, \xi; b) \,, \quad y \in \mathbf{R} \,, \quad \xi \in \mathbf{R}^d \,, \quad b \in \Theta \,. \tag{1.122}$$

This is a Fredholm integral equation of the first kind with the kernel $K(\xi, x) = \rho(x - \xi)$. It is known that those integral equations, unlike equations of the second kind, are not always effectively resolved. That is why the CS method is not universal.

As a special case, consider the functional model (1.1) and (1.5). We will correct the least squares estimating function $s_{\mathrm{OLS}} = (y - f(\xi, \beta))\, \frac{\partial f(\xi, \beta)}{\partial \beta}$. Split the deconvolution equation

$$\mathbf{E}[s_{\mathrm{C}}(y, x; b)|y, \xi] = s_{\mathrm{OLS}}(y, \xi; b) \tag{1.123}$$

in two equations with unknown vector functions $g$ and $h$

$$\mathbf{E}[g(x, b)|\xi] = \frac{\partial f(\xi, b)}{\partial \beta}\,, \tag{1.124}$$

$$\mathbf{E}[h(x, b)|\xi] = f(\xi, b)\ \frac{\partial f(\xi, b)}{\partial \beta}\,. \tag{1.125}$$

If one solves these equations, then the function

$$s_{\mathrm{C}}(y, x; b) = yg(x, b) - h(x, b) \tag{1.126}$$

satisfies (1.123). Indeed, we will have

$$\begin{aligned}\mathbf{E}[yg(x, b) - h(x, b)|y, \xi] &= y\mathbf{E}[g(x, b)|\xi] - \mathbf{E}[h(x, b)|\xi] = \\ &= y\,\frac{\partial f(\xi, b)}{\partial \beta} - f(\xi, b)\,\frac{\partial f(\xi, b)}{\partial \beta} = s_{\mathrm{OLS}}(y, \xi; b)\,.\end{aligned} \tag{1.127}$$

In Section 3, it will be shown that the deconvolution equations (1.124) and (1.125) for polynomial regression function

$$f(\xi, \beta) = \sum_{i=0}^{k} \beta_i \xi^i \tag{1.128}$$

have a unique solution within the class of polynomial functions in $x$. In Stefanski (1989), a quite broad class of cases is studied where it is possible to solve the deconvolution equations of type either (1.124) or (1.125) for the normal error $\delta$.

Now, consider a general regression model of $y$ on $\xi$ described at the beginning of Section 1.4.2, which is a functional case with the classical error (1.47). Usually the score function

$$s_{\mathrm{OML}}(y, \xi; \beta) = \frac{1}{\rho(y, \xi; \beta)}\ \frac{\partial \rho(y, \xi; \beta)}{\partial \beta} \tag{1.129}$$

is corrected. The deconvolution equation will look like

$$\mathbf{E}[s_{\mathrm{C}}(y, x; b)|y, \xi] = s_{\mathrm{OML}}(y, \xi; b)\,. \tag{1.130}$$

### 1.4.5 Regression calibration (RC)

This is a purely structural method in the presence of the classical error (1.5). Let $s(y, \xi; \beta)$ be an unbiased estimating function, which was mentioned at the beginning of Section 1.4.4. Another way to adjust the naive estimator generated by the estimating function $s(y, x; \beta)$ consists in evaluating the conditional expectation

$$\xi^* = \xi^*(x) = \mathbf{E}(\xi|x) \tag{1.131}$$

(see Section 1.4.3) after which the estimator $\hat{\beta}_{\mathrm{RC}}$ is defined as one of the solutions to the following equation:

$$\sum_{i=1}^{n} s(y_i, \xi_i^*; \beta) = 0\,, \quad \beta \in \Theta\,, \tag{1.132}$$

where $\xi_i^* = \xi^*(x_i) = \mathbf{E}(\xi | x = x_i)$.

The method is convenient because it does not need an adjustment of the initial estimating function, instead the observed values $x_i$ are corrected. If the nuisance parameter $\gamma$ of distribution of $\xi$ is unknown, then it can be estimated as in Section 1.4.3 and substituted in the expression for the conditional expectation (1.132)

$$\hat{\xi}_i^* = \hat{\mathbf{E}}(\xi | x = x_i)\,, \tag{1.133}$$

where cap above the expectation means that we substituted the estimator $\hat{\gamma}$ instead of $\gamma$ when evaluated the conditional expectation; then the estimator $\hat{\beta}_{\mathrm{RC}}$ is defined as one of the solution to the equation

$$\sum_{i=1}^{n} s(y_i, \hat{\xi}_i^*; \beta) = 0\,, \quad \beta \in \Theta\,. \tag{1.134}$$

Generally speaking, the RC method does not yield a consistent estimator of the parameter $\beta$. However, deviation of the estimator from the true value is much less than the deviation $\hat{\beta}_{\mathrm{naive}} - \beta$. Remember that the naive estimator $\hat{\beta}_{\mathrm{naive}}$ satisfies the uncorrected equation

$$\sum_{i=1}^{n} s(y_i, x_i; \beta) = 0\,, \quad \beta \in \Theta\,. \tag{1.135}$$

The RC method is very popular, and it gives good results when the variance $\sigma_\delta^2$ is relatively small. However, at large $\sigma_\delta^2$, this method can produce rather large deviation $\hat{\beta}_{\mathrm{RC}} - \beta$.

### 1.4.6 SIMEX estimator

#### Modeling and extrapolation steps

The idea of the method is due to the American mathematicians Cook and Stefanski (1994). As in Section 1.4.4, consider a general structural model with the classical error (1.47). Suppose for simplicity that the regressors $\xi_i$ are scalar and we know the measurement error variance $\sigma_\delta^2$.

The unbiased evaluating function $s(y, \xi; \beta)$ providing consistent estimation of the parameter $\beta$ at $\delta_i \equiv 0$ allows, based on the sample $\{y_i, x_i, i = \overline{1, n}\}$, to construct the naive estimator $\hat{\beta}_{\mathrm{naive}} = \hat{\beta}_{\mathrm{naive}}(\sigma_\delta)$ as one of the solutions to the equation

$$\frac{1}{n} \sum_{i=1}^{n} s(y_i, x_i, b) = 0\,, \quad b \in \Theta\,. \tag{1.136}$$

By the Strong Law of Large Numbers, the left-hand side of (1.136) a.s., as $n$ tends to infinity, converges to the limit function

$$S_\infty(b) = \mathbf{E}_\beta s(y, x; \beta) , \quad b \in \Theta . \tag{1.137}$$

If the variance $\sigma_\delta^2$ is small compared to $\sigma_\xi^2$ (and we further demand it in this section), then the naive estimator converges almost surely to a nonrandom vector $\beta^* = \beta^*(\sigma_\delta)$ satisfying the limit equation

$$S_\infty(\beta^*) = \mathbf{E}_\beta(y, x; \beta^*) = 0 , \quad \beta^* \in \Theta . \tag{1.138}$$

The main idea of the SIMEX method (pronounced ['sɪm eks] from SIMulation–EXtrapolation) is to experimentally investigate the effect of measurement error on the $\hat\beta_{\text{naive}}(\sigma_\delta)$ by imposing additional noise.

SIMEX is a randomized estimation method, i.e., a method that requires additionally generated random numbers. The estimating procedure consists of two steps. In the simulation step, we generate additional data with greater measurement error variance $(1 + \lambda)\sigma_\delta^2$. At each $\lambda \geq 0$, we define

$$x_{b,i}(\lambda) = x_i + \sqrt{\lambda}\delta_{b,i} , \quad i = \overline{1, n} , \quad b = \overline{1, B} , \tag{1.139}$$

where $\{\delta_{b,i}, i = \overline{1, n}, b = \overline{1, B}\}$ are computer-generated pseudoerrors to be independent of all observable data and independent normally distributed with distribution $\mathrm{N}(0, \sigma_\delta^2)$. Let $\hat\beta_b(\lambda)$ be the naive estimate evaluated based on $\{y_i, x_{b,i}(\lambda), i = \overline{1, n}\}$, i.e. the root of equation (1.136) where instead of $x_i$ we substitute $x_{b,i}(\lambda)$, $i = \overline{1, n}$.

Evaluating the mean

$$\hat\beta(\lambda) = \frac{1}{B}\sum_{b=1}^{B} \hat\beta_b(\lambda) \tag{1.140}$$

completes the simulation step in the SIMEX algorithm.

In the extrapolation step, we approximate the evaluated values $\{\hat\beta(\lambda_m), \lambda_m;\ m = \overline{1, M}\}$ by some function $G(\lambda, \Gamma)$, $\lambda = \lambda_m$, $m = \overline{1, M}$, where $\lambda_1 = 0 < \lambda_2 < \cdots < \lambda_M$ and $\Gamma \in \Theta_\Gamma$ is a vector parameter of a chosen family of analytic functions in $\lambda$. The parameter $\Gamma$ is estimated by the least squares method:

$$\hat\Gamma = \arg\min_{\Gamma\in\Theta_\Gamma} \sum_{m=1}^{M} \left\|\hat\beta(\lambda_m) - G(\lambda_m, \Gamma)\right\|^2 . \tag{1.141}$$

Now, we have an approximate equality

$$\hat\beta(\lambda) \approx G(\lambda, \hat\Gamma) , \quad \lambda \geq 0 . \tag{1.142}$$

Remember that $\hat\beta(\lambda)$ corresponds to the averaged naive estimator at the aggregate measurement error variance $(1 + \lambda)\sigma_\delta^2$. We define the SIMEX estimator as the extrapolated value

$$\hat\beta_{\text{SIMEX}} = G(-1, \hat\Gamma) . \tag{1.143}$$

This corresponds to the error variance $(1+\lambda)\sigma_\delta^2|_{\lambda=-1} = 0$, and at zero error, the naive estimator begins to be consistent.

Thus, the SIMEX estimator provides significant compensation of deviations from the true value. The estimator is designed mainly for small and moderate samples.

**Choice of extrapolating function and values $\lambda_m$**

If one is able to find a family of extrapolating functions $\{G(\bullet, \Gamma)\ ,\ \ \Gamma \in \Theta_\Gamma\}$ so that at some $\Gamma \in \Theta_\Gamma$ the approximate equality (1.142) is accurate, then the estimator $\hat{\beta}_{\text{SIMEX}}$ is consistent. For the case of *linear regression*

$$y = \beta_0 + \beta_1 \xi + \varepsilon\ , \quad x = \xi + \delta \tag{1.144}$$

the equality in (1.75) is provided by a rational linear function (Cheng and Van Ness, 1999; Wansbeek and Meijer, 2000)

$$G_{\text{RL}}(\lambda, \Gamma) = \gamma_1 + \frac{\gamma_2}{\gamma_3 + \lambda}\ , \tag{1.145}$$

and all the components of the function $G$ have the form (1.145).

In nonlinear models, in addition to the functions (1.145), we can apply either a quadratic function

$$G_{\text{Q}}(\lambda, \Gamma) = \gamma_0 + \gamma_1 \lambda + \gamma_2 \lambda^2\ , \tag{1.146}$$

or a polynomial function of higher order

$$G_{\text{P},k}(\lambda, \Gamma) = \gamma_0 + \gamma_1 \lambda + \cdots + \gamma_k \lambda^k\ . \tag{1.147}$$

To select the exact class of extrapolating functions is difficult because, in practice, the SIMEX estimator is inconsistent. However, for small and moderate samples the inconsistent estimator usually behaves better than common consistent estimators.

As for the choice of values $\lambda = \lambda_m$ in the simulation step, the monographs Cheng and Van Ness (1999) and Wansbeek and Meijer (2000) suggest using uniform partition of the interval $[0, \lambda_{\max}]$, with $1 \le \lambda_{\max} \le 2$.

Note that the SIMEX estimator can perform quite well in functional models, because in its construction it is not necessary to know the distribution form of the latent variable.

**Confidence ellipsoid**

The books by Cheng and Van Ness (1999) and Wansbeek and Meijer (2000) proposed a method for constructing the ACM of the estimator $\hat{\beta}_{\text{SIMEX}}$ based on the asymptotic normality of $\hat{\beta}_{\text{SIMEX}}$ at small $\sigma_\delta^2$ (Carroll et al., 1996). In this section, we describe the method.

If there were no measurement errors, then under regularity conditions, the naive estimator would be asymptotically normal:

$$\sqrt{n}(\hat{\beta}_{\text{naive}}(0) - \beta) \xrightarrow{d} \text{N}(0, \Sigma(0))\ , \tag{1.148}$$

where the ACM $\Sigma(0)$ is given by the sandwich formula (see Appendix A2):

$$\Sigma(0) = A^{-1}BA^{-\mathrm{T}}\,, \tag{1.149}$$

$$A^{-\mathrm{T}} = (A^{-1})^{\mathrm{T}}\,, \quad A = \mathbf{E}_\beta \frac{\partial s(y,x;\beta)}{\partial \beta^{\mathrm{T}}}\,, \tag{1.150}$$

$$\text{where } B = \operatorname{cov}(s(y,x;\beta)) = \mathbf{E}_\beta s(y,x;\beta)\cdot s^{\mathrm{T}}(y,x;\beta)\,. \tag{1.151}$$

As a consistent estimator of the matrix $\Sigma(0)$, the following matrix is used:

$$\tau^2(0) = A_n^{-1}B_nA_n^{-\mathrm{T}}\,, \tag{1.152}$$

$$\begin{aligned} A_n &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial s(y_i,x_i;\hat\beta_{\mathrm{naive}}(0))}{\partial\beta^{\mathrm{T}}}\,,\\ B_n &= \frac{1}{n}\sum_{i=1}^{n} s(y_i,x_i;\beta)\cdot s^{\mathrm{T}}(y_i,x_i;\beta)\Big|_{\beta=\hat\beta_{\mathrm{naive}}(0)}\,. \end{aligned} \tag{1.153}$$

Similar approximate ACMs can be evaluated for $\hat\beta_b(\lambda_m)$, $b = 1,\dots,B$, $m = 1,\dots,M$:

$$\tau_b^2(\lambda) = A_n(\lambda)^{-1}B_n(\lambda)A_n(\lambda)^{-\mathrm{T}}\,, \tag{1.154}$$

$$A_n(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial S(y_i,x_{i,b};\hat\beta_b(\lambda))}{\partial\beta^{\mathrm{T}}}\,, \tag{1.155}$$

$$B_n(\lambda) = \frac{1}{n}\sum_{i=1}^{n} s(y_i,x_{i,b};\hat\beta_b(\lambda))\; s(y_i,x_{i,b};\hat\beta_b(\lambda))^{\mathrm{T}}. \tag{1.156}$$

Let $\tau^2(\lambda)$ be the mean of the matrices $\tau_b^2(\lambda)$ over $b = \overline{1,B}$, and $S_\Delta^2(\lambda)$ be a sample covariance matrix of the vectors $\hat\beta_b(\lambda)$, $b = \overline{1,B}$ (remember that $p = \dim\beta$),

$$S_\Delta^2(\lambda) = \frac{1}{n-p}\sum_{b=1}^{B}(\hat\beta_b(\lambda)-\hat\beta(\lambda))(\hat\beta_b(\lambda)-\hat\beta(\lambda))^{\mathrm{T}}\,. \tag{1.157}$$

Then, we can estimate the ACM of the SIMEX estimator by extrapolation of differences $\tau^2(\lambda) - S_\Delta^2(\lambda)$, $\lambda = \lambda_1,\dots,\lambda_m$ to the value $\lambda = -1$. Of course for finite sample, the obtained matrix need not be positive definite. If it is, then the asymptotic confidence ellipsoid can be constructed by a standard procedure based on the approximate relation

$$\sqrt{n}(\hat\beta_{\mathrm{SIMEX}} - \beta) \approx \mathrm{N}(0, \tau^2_{\mathrm{SIMEX}})\,. \tag{1.158}$$

If the matrix $\tau^2_{\mathrm{SIMEX}}$ does not come out positive definite but its diagonal entries are $(\tau^2_{\mathrm{SIMEX}})_{ii} > 0$, then the asymptotic confidence interval for the $i$-th component $\beta_i$ of the vector $\beta$ is constructed based on the approximate relation

$$\sqrt{n}(\hat\beta_{\mathrm{SIMEX},i} - \beta_i) \approx \mathrm{N}(0, (\tau^2_{\mathrm{SIMEX}})_{ii})\,. \tag{1.159}$$

**Asymptotic expansion of the estimator for small $\sigma_\delta^2$**

For small $\sigma_\delta^2$, we can explain the SIMEX estimator efficiency by the following asymptotic expansions of the estimators $\hat\beta_{\text{naive}}(\sigma_\delta)$ and $\hat\beta_{\text{SIMEX}} = \hat\beta_{\text{SIMEX}}(\sigma_\delta)$, proved in Gontar and Kuechenhoff (2008).

Let the naive estimator within the structural model be defined by equation (1.136), where $\Theta$ is a convex compact set in $\mathbf{R}^p$; $\beta$ is an interior point of $\Theta$; the regressors are scalar, and for all $\lambda \in \mathbf{R}$, $\mathbf{E}e^{\lambda\xi_1} < \infty$ holds. Regarding the classical error $\delta$, we require that $\delta \sim \mathrm{N}(0, \sigma_\delta^2)$ where $\sigma_\delta^2 > 0$ is given.

**Theorem 1.1** (Expansion of the naive estimator). *With fixed $l \geq 0$, assume the following.*

(1) *A measurable, over the set of variables, estimating function is given, $s(y, \xi; \beta) \in C^{2l+2}(\mathbf{R}\times\mathbf{R}\times U \to \mathbf{R}^p)$ (here the smoothness is required in $\xi$ and $\beta$), with some open set $U \supset \Theta$.*

(2) *For any partial derivative $D_q s(y, \xi; \beta)$ of order $0 \leq q \leq 2l+2$ in $\xi$ and components $\beta$, it holds that $\|D_q s(y, \xi; \beta)\| \leq c_1 e^{C_2|\xi|}$, where $c_1$ and $c_2$ are positive constants; $D_0 s = s$.*

(3) *There is a unique solution to the equation $\mathbf{E}_\beta s(y, \xi; \beta) = 0$, $b \in \Theta$, and the solution is $b = \beta$.*

(4) *The matrix $\mathbf{E}_\beta \frac{\partial s(y,\xi;\beta)}{\partial \beta^{\mathrm{T}}}$ is nonsingular.*

*Then, there exists a $\sigma > 0$ such that for all $\sigma_\delta \in (0, \sigma)$,*

$$\hat\beta_{\text{naive}}(\sigma_\delta) \xrightarrow{P1} \beta^*(\sigma_\delta)\,, \quad as \;\; n \to \infty\,, \tag{1.160}$$

*where $\beta^*(\sigma_\delta)$ is a unique root of the limit equation* (1.138)*, moreover*

$$\beta^*(\sigma_\delta) = \beta + \sum_{j=1}^{l} \frac{1}{(2j)!} \times \frac{\partial^{2j}\beta^*(0)}{\partial(\sigma_\delta)^{2j}} + O\left(\sigma_\delta^{2l+2}\right)\,, \quad as \;\; \sigma_\delta^2 \to 0\,. \tag{1.161}$$

We comment Theorem 1.1. The conditions (1) and (2) provide the desired smoothness and unboundedness of the estimating function; the condition (3) is a basis for the strong consistency of the estimator $\hat\beta_{\text{naive}}(0)$ in the absence of measurement error; the condition (4) allows applying the implicit function theorem (Burkill, 1962) and get the desired smoothness of the root $\beta^* = \beta^*(\sigma_\delta)$ as a function in $\sigma_\delta \in [0, \sigma)$. The expansions (1.160) and (1.161) demonstrate that the asymptotic deviation of the naive estimator $\beta^*(\sigma_\delta) - \beta$ begins with members of order $\sigma_\delta^2$ (if $\frac{\partial^2\beta^*(0)}{\partial(\sigma_\delta)^2} \neq 0$, which is realistic).

**Theorem 1.2** (Expansion of the SIMEX estimator). *Let the extrapolating function $G(\lambda, \Gamma)$ be used in the SIMEX procedure, being composed of polynomials of degree not higher than $m$, and $l \leq m \leq M$. Suppose that the conditions of Theorem* 1.1 *hold. Then*

$$\hat\beta_{\text{SIMEX}}(\sigma_\delta) \xrightarrow{P1} \beta^*_{\text{SIMEX}}(\sigma_\delta)\,, \quad as \quad n \to \infty\,, \tag{1.162}$$

*with nonrandom limit*

$$\beta^*_{\text{SIMEX}}(\sigma_\delta) = \beta + O(\sigma_\delta^{2l+2})\,, \quad as \quad \sigma_\delta^2 \to 0\,. \tag{1.163}$$

Theorem 1.2 shows that the asymptotic deviation $\beta^*_{\text{SIMEX}}(\sigma_\delta) - \beta$ does not contain terms of order $(\sigma_\delta)^k$, $k = \overline{1, 2l+1}$. Thus, at small $\sigma^2_\delta$ the asymptotic deviation for the SIMEX estimator is less than for the naive one.

### 1.4.7 Comparison of various estimators in models with the classical error

Relative to other estimators, the naive ones are suitable only for small variances $\sigma^2_\delta$ but they are less accurate for larger $\sigma^2_\delta$.

The structural estimators are: the MLE, the QLE, the estimator obtained by RC, and the functional ones are the CS and SIMEX estimators. The functional estimators are more stable (*robust*) against violation of assumption about the distribution of the latent variable. Therefore, if there is no certainty in the latter assumption, the functional estimators will be the best choice. It is worth mentioning the paper by Schneeweiss and Cheng (2006), which shows that violation of the assumption about the normal distribution of $\xi$ disturbs the consistency of the QLE based on the requirement of such normality.

Now, let us know reliably that $\xi$ has a normal distribution. Under such circumstances, it is worth using the QLE estimator, which is optimal in a broad class of consistent estimators (Kukush et al., 2007, 2009). In nonlinear models, it is quite difficult to compute the MLE and we cannot guarantee its asymptotic properties. The RC is easy to use but it does not yield a consistent estimator.

Compare the functional methods. The CS method yields the consistent estimator, but we cannot always implement it, because rather often the deconvolution equation (1.118) is difficult or even impossible to solve. Even if it was done, the CS is unstable and requires a small sample modification (Cheng et al., 2000). At the same time, the SIMEX estimator is numerically stable and reduces the deviation of estimator, although in general it is inconsistent.

As we can see, the choice of an appropriate estimator is an art. It depends on the sample size and on how we trust the assumptions of the observation model.

# 2 Linear models with classical error

In this chapter, we illustrate the problems arising when estimating regression parameters with classical measurement errors using an example of the simplest regression model, the linear one. It is described by the following two equations:

$$y_i = \beta_0 + \beta_1 \xi_i + \varepsilon_i \,, \tag{2.1}$$

$$x_i = \xi_i + \delta_i \,, \quad i = \overline{1, n} \,. \tag{2.2}$$

All the variables are scalar. This model has been used in an example of Section 1.1. Here $\xi_i$ is unobservable value of the latent variable (the $\xi_i$ acts as a regressor), $x_i$ is the observed surrogate data, $y_i$ is the observable response, $\varepsilon_i$ is the response error, and $\delta_i$ is the measurement error in the covariate. We have to estimate the intercept $\beta_0$ and the slope $\beta_1$. Typical are the following assumptions about the observation model.

(i) Random variables $\xi_i$, $\varepsilon_i$, $\delta_i$, $i \geq 1$, are independent.

(ii) The errors $\varepsilon_i$ are identically distributed and centered with finite and positive variance $\sigma_\varepsilon^2$.

(iii) The errors $\delta_i$ are identically distributed and centered with finite and positive variance $\sigma_\delta^2$.

In the structural case, the latent variable is random. Or speaking more precisely, we require the following:

(iv) Random variables $\xi_i$ are identically distributed, with $\mathbf{E}\xi_i = \mu_\xi$ and $\mathbf{D}\xi_i = \sigma_\xi^2 > 0$.

Note that under conditions (i) and (iv), the random variables $\xi_i$ are independent and identically distributed.

In the functional case, the next condition holds instead of (iv).

(v) The values $\xi_i$ are nonrandom.

We try to transform the model (2.1) and (2.2) to an ordinary linear regression. Excluding the $\xi_i$ from equation (2.1) we have

$$y_i = \beta_0 + \beta_1 x_i + \tau_i \,, \quad \tau_i = \varepsilon_i - \beta_1 \delta_i \,, \quad i = \overline{1, n} \,. \tag{2.3}$$

The model (2.3) resembles an ordinary linear regression, with errors $\tau_i$. Given (i)–(iii), the random variables $\tau_i$ are centered, independent, and identically distributed. We get

$$\mathbf{D}\tau_i = \mathbf{D}\varepsilon_i + \mathbf{D}(-\beta_1 \delta_i) = \sigma_\varepsilon^2 + \beta_1^2 \sigma_\delta^2 > 0 \,. \tag{2.4}$$

As we can see, the new error variance depends on the slope $\beta_1$. This is the first difference of the model (2.3) from the ordinary regression, where usually the error variance does not depend on regression parameters.

We now find the covariance between the new regressor $x_i$ and the error $\tau_i$ using the assumptions (i)–(iii):

$$\operatorname{cov}(x_i, \tau_i) = \operatorname{cov}(\xi_i + \delta_i, \varepsilon_i - \beta_1 \delta_i) = \operatorname{cov}(\delta_i, -\beta_1 \delta_i) = \mathbf{E}(-\beta_1 \delta_i^2) = -\beta_1 \sigma_\delta^2 \,. \tag{2.5}$$

Thus, in case of $\beta_1 \neq 0$ (i.e., when in equation (2.1) there is a real dependence between the response and regressor $\xi_i$), we have

$$\operatorname{cov}(x_i, \tau_i) \neq 0 \,. \tag{2.6}$$

Inequality (2.6) is the second main difference of the model (2.3) from the ordinary regression, where it is strictly required that the regressor and the error are uncorrelated. One can see that in essence the model (2.1) and (2.2) does not come to the ordinary regression. Therefore, the theory of linear measurement error models is substantial.

In the following the structural errors-in-variables model with assumptions (i)–(iv) will be mainly under consideration.

## 2.1 Inconsistency of the naive estimator: the attenuation effect

Choose $\Theta = \mathbf{R}^2$ in the model (2.1) and (2.2) as a parameter set for the augmented regression parameter $\beta = (\beta_0, \beta_1)^{\mathrm{T}}$. Such a choice of $\Theta$ corresponds to the absence of a prior information about possible values of regression parameters. According to Section 1.4.1, the naive estimator $\hat{\beta}_{\mathrm{naive}} = \hat{\beta}_{\mathrm{OLS}}$ is defined by minimizing in $\mathbf{R}^2$ the objective function

$$Q_{\mathrm{OLS}}(\beta) = \sum_{i=1}^{n} q_{\mathrm{OLS}}(y_i, x_i; \beta) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \,. \tag{2.7}$$

Therefore,

$$\hat{\beta}_{\mathrm{naive}} = \operatorname*{arg\,min}_{\beta \in \mathbf{R}^2} Q_{\mathrm{OLS}}(\beta) \,. \tag{2.8}$$

Find an explicit expression for the naive estimator. Denote the sample mean as

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \,. \tag{2.9}$$

Hereafter bar means averaging, e.g.,

$$\overline{xy} = \frac{1}{n} \sum_{i=1}^{n} x_i y_i \,, \quad \overline{y^2} = \frac{1}{n} \sum_{i=1}^{n} y_i^2 \,. \tag{2.10}$$

The sample variance of the values $x_i$ is defined by the expression

$$\mathrm{S}_{xx} = \overline{(x - \bar{x})^2} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \,, \tag{2.11}$$

and the sample covariance between $x_i$ and $y_i$ by the expression

$$S_{xy} = \overline{(x - \bar{x})(y - \bar{y})} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) . \tag{2.12}$$

There are other useful representations of the statistics (2.11) and (2.12):

$$S_{xx} = \overline{x^2} - (\bar{x})^2 , \quad S_{xy} = \overline{xy} - \bar{x}\cdot\bar{y} . \tag{2.13}$$

**Theorem 2.1.** *In the model* (2.1) *and* (2.2)*, let the sample size* $n \geq 2$*, and moreover, not all the observed values* $x_i$ *coincide. Then the objective function* (2.7) *has a unique minimum point in* $\mathbf{R}^2$*, with components*

$$\hat{\beta}_{1,\text{naive}} = \frac{S_{xy}}{S_{xx}} , \quad \hat{\beta}_{0,\text{naive}} = \bar{y} - \bar{x}\cdot\hat{\beta}_{1,\text{naive}} . \tag{2.14}$$

*Proof.* A necessary condition for $\hat{\beta}$ to be the minimum point of the function $Q_{\text{OLS}}$ is the so-called *system of normal equations*:

$$\frac{\partial Q_{\text{OLS}}(\hat{\beta})}{\partial \beta_0} = 0 , \quad \frac{\partial Q_{\text{OLS}}(\hat{\beta})}{\partial \beta_1} = 0 . \tag{2.15}$$

It has the form

$$\begin{cases} \hat{\beta}_0 + \hat{\beta}_1\bar{x} = \bar{y} , \\ \hat{\beta}_0\bar{x} + \hat{\beta}_1\overline{x^2} = \overline{xy} . \end{cases} \tag{2.16}$$

Eliminate $\hat{\beta}_0$ from the second equation:

$$\hat{\beta}_1\left(\overline{x^2} - (\bar{x})^2\right) = \overline{xy} - \bar{x}\cdot\bar{y} , \tag{2.17}$$

$$S_{xx}\cdot\hat{\beta}_1 = S_{xy} . \tag{2.18}$$

Since not all the values $x_i$ coincide, $S_{xx} \neq 0$. Hence,

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} , \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} . \tag{2.19}$$

Thus, if a global minimum of the function $Q_{\text{OLS}}$ is attained, then a minimum point is $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)^{\text{T}}$ from (2.19). Now, it is enough to show that

$$Q_{\text{OLS}}(\beta) > Q_{\text{OLS}}(\hat{\beta}) , \quad \beta \neq \hat{\beta} . \tag{2.20}$$

The function $Q_{\text{OLS}}$ is a polynomial of the second order in $\beta_0$ and $\beta_1$. This function can be exactly expanded in the neighborhood of $\hat{\beta}$ using Taylor's formula

$$Q_{\text{OLS}}(\beta) = Q_{\text{OLS}}(\hat{\beta}) + Q'_{\text{OLS}}(\hat{\beta})(\beta - \hat{\beta}) + \frac{1}{2}(\beta - \hat{\beta})^{\text{T}}Q''_{\text{OLS}}\cdot(\beta - \hat{\beta}) . \tag{2.21}$$

The derivative $Q'_{\mathrm{OLS}}(\hat\beta) = 0$ because $\hat\beta$ satisfies (2.15), and the matrix of the second derivatives $Q''_{\mathrm{OLS}}$ does not depend on the point where it is evaluated:

$$\frac{1}{2}\left(\frac{1}{n}Q_{\mathrm{OLS}}\right)'' = \begin{pmatrix} 1 & \bar x \\ \bar x & \overline{x^2} \end{pmatrix} = S\,. \tag{2.22}$$

The matrix $S$ is positive definite, because $S_{11} = 1 > 0$ and $\det S = \overline{x^2} - (\bar x)^2 = \mathrm{S}_{xx} > 0$. Therefore, at $\beta \neq \hat\beta$ from (2.21), we have

$$\frac{1}{n}(Q_{\mathrm{OLS}}(\beta) - Q_{\mathrm{OLS}}(\hat\beta)) = (\beta - \hat\beta)^{\mathrm{T}} S\,(\beta - \hat\beta) > 0\,. \tag{2.23}$$

This proves (2.20) and the statement of the theorem. □

**Remark 2.2.** Based on the naive estimator, one can draw a straight line

$$y = \hat\beta_{0,\mathrm{naive}} + \xi \cdot \hat\beta_{1,\mathrm{naive}}\,. \tag{2.24}$$

The line will pass through the center of mass $M(\bar x, \bar y)$ of the observed points $M_i(x_i; y_i)$, $i = \overline{1, n}$. In fact, it follows from the first equation of system (2.16). The line (2.24) is an estimator of the true line $y = \beta_0 + \beta_1\xi$ in the coordinate system $(\xi;\ y)$.

Study the limit of the naive estimator, as the sample size grows. Here we consider the structural model.

Recall that the reliability ratio $K = (\sigma_\xi^2)/(\sigma_\xi^2 + \sigma_\delta^2)$ has been introduced in Section 1.4.3. The almost sure (a.s.) convergence is denoted by $\xrightarrow{P1}$.

**Theorem 2.3.** *Given* (i)–(iv), *it holds that*

$$\hat\beta_{1,\mathrm{naive}} \xrightarrow{P1} K\beta_1\,, \quad \text{as } n \to \infty\,. \tag{2.25}$$

*Proof.* Using the strong law of large numbers (SLLN), we find limit of the denominator of expression (2.14):

$$\mathrm{S}_{xx} = \overline{x^2} - (\bar x)^2 \xrightarrow{P1} \mathbf{E}x^2 - (\mathbf{E}x)^2 = \mathbf{D}x = \sigma_\xi^2 + \sigma_\delta^2 > 0\,. \tag{2.26}$$

Therefore, the statistic $\mathrm{S}_{xx}$ becomes positive with probability 1 at $n \geq n_0(\omega)$ (i.e., starting from some random number); then one can apply Theorem 2.1. Again by the SLLN,

$$\mathrm{S}_{xy} = \overline{xy} - \bar x \cdot \bar y \xrightarrow{P1} \mathbf{E}xy - (\mathbf{E}x)(\mathbf{E}y) = \mathrm{cov}(x, y)\,, \quad \text{as } n \to \infty\,. \tag{2.27}$$

Here, random variables $(x, \xi, \delta, \varepsilon, y)$ have the same joint distribution as $(x_i, \xi_i, \delta_i, \varepsilon_i, y_i)$. As $\xi, \delta, \varepsilon$ are jointly independent, we have further

$$\mathrm{cov}(x, y) = \mathrm{cov}(\xi + \delta, \beta_0 + \beta_1\xi + \varepsilon) = \mathrm{cov}(\xi, \beta_1\xi) = \beta_1\sigma_\xi^2\,. \tag{2.28}$$

Thus, with probability 1 at $n \geq n_0(\omega)$ it holds that

$$\hat\beta_{1,\mathrm{naive}} = \frac{\mathrm{S}_{xy}}{\mathrm{S}_{xx}} \xrightarrow{P1} \beta_1\frac{\sigma_\xi^2}{\sigma_\xi^2 + \sigma_\delta^2} = K\beta_1\,, \quad \text{as } n \to \infty\,. \tag{2.29}$$

The theorem is proved. □

Relation (2.25) shows that the estimator $\hat{\beta}_{\text{naive}}$ is not consistent, because for $\hat{\beta}_1 \neq 0$, the estimator $\beta_{1,\text{naive}}$ does not converge in probability to $\beta_1$ (in fact, it converges a.s., and therefore, in probability, to $K\beta_1 \neq \beta_1$). The phenomenon that the limit of the slope estimator is the value $K\beta_1$ located between 0 and $\beta_1$, is called *attenuation effect*. It characterizes the behavior of the naive estimator in any regression model with the classical error. The effect is plausible by the following argument. Consider the predictive line (2.24). Thanks to the convergence (2.25), the slope of the line is close to $K\beta_1$ at large $n$. If $\beta_1 \neq 0$ and $n$ is large, the line (2.24) passes more flat than the true line $y = \beta_0 + \beta_1\xi$. Thus, we conclude that the estimated dependence $y$ of $\xi$ is weaker than the true one.

In particular, using the naive estimator in the binary incidence model (1.2), (1.4), and (1.5) leads to the attenuation of dependence of the odds function $\lambda$ on the exposure dose $\xi_i$: thus, we will have

$$\hat{\lambda}_{\text{naive}}(\xi_0) = \hat{\beta}_{0,\text{naive}} + \hat{\beta}_{1,\text{naive}} \cdot \xi_i\,, \quad i = \overline{1, n}\,, \tag{2.30}$$

with the estimated excess absolute risk (EAR) $\hat{\beta}_{1,\text{naive}} < \beta_1$. It should be noted that the latter inequality will hold at sufficiently large sample size.

## 2.2 Prediction problem

Now, it is required that the latent variable and the measurement error be normally distributed. Thus, we specify the conditions (iii) and (iv).

(vi) The errors $\delta_i$ are identically distributed, with distribution $\mathrm{N}(0, \sigma_\delta^2)$, $\sigma_\delta^2 > 0$.

(vii) Random variables $\xi_i$ are identically distributed, with distribution $\mathrm{N}(\mu_\xi, \sigma_\xi^2)$, $\sigma_\xi^2 > 0$.

The conditions (i), (ii), (vi), and (vii) are imposed. *The prediction problem* in the model (2.1) and (2.2) is the following: if the next observation $x_{n+1}$ of the surrogate variable comes, the corresponding observation $y_{n+1}$ of the response should be predicted. The optimal predictor $\hat{y}_{n+1}$ is sought as a Borel measurable function of the sample and the value $x_{n+1}$ , for which the mean squared error $\mathbf{E}(\hat{y}_{n+1} - y_{n+1})^2$ is minimal.

From probability theory, it is known that such an optimal predictor is given by the conditional expectation:

$$\hat{y}_{n+1} = \mathbf{E}(y_{n+1}|y_1, x_1, y_2, x_2, \ldots, y_n, x_n;\ x_{n+1})\,. \tag{2.31}$$

The vector $(y_{n+1}, x_{n+1})^{\mathrm{T}}$ is stochastically independent of the sample $y_1, x_1, \ldots, y_n, x_n$, therefore,

$$\hat{y}_{n+1} = \mathbf{E}\,(y_{n+1}|x_{n+1}) = \beta_0 + \beta_1\,\mathbf{E}\,(\xi_{n+1}|x_{n+1}) + \mathbf{E}(\varepsilon_{n+1}|x_{n+1})\,. \tag{2.32}$$

Since $\varepsilon_{n+1}$ and $x_{n+1}$ are independent, we have

$$\mathbf{E}\,(\varepsilon_{n+1}|x_{n+1}) = \mathbf{E}\varepsilon_{n+1} = 0\,. \tag{2.33}$$

Further, using the normality assumptions (vi) and (vii) we can utilize the results of Section 1.4.3 and get

$$\mathbf{E}\,(\xi_{n+1}|x_{n+1}) = Kx_{n+1} + (1-K)\mu_\xi\ . \tag{2.34}$$

Equalities (2.32)–(2.34) yield the optimal predictor

$$\hat{y}_{n+1} = (\beta_0 + \beta_1(1-K)\mu_\xi) + K\beta_1 x_{n+1}\ . \tag{2.35}$$

At the same time, this predictor is unfeasible because the model parameters $\beta_0, \beta_1, K, \mu_\xi$ are unknown. Relying on the naive estimator one can offer the predictor

$$\tilde{y}_{n+1} = \hat{\beta}_{0,\text{naive}} + x_{n+1}\hat{\beta}_{1,\text{naive}}\ . \tag{2.36}$$

Unexpectedly, this predictor is little different from $\hat{y}_{n+1}$, for large $n$.

**Theorem 2.4.** *Assume* (i), (ii), (vi), *and* (vii). *Then*

$$\hat{\beta}_{0,\text{naive}} \xrightarrow{P1} \beta_0 + \beta_1\,(1-K)\,\mu_\xi\ , \tag{2.37}$$

$$\hat{\beta}_{1,\text{naive}} \xrightarrow{P1} K\beta_1\ . \tag{2.38}$$

*Therefore, the predictor $\tilde{y}_{n+1}$ is close to the optimal one*

$$\tilde{y}_{n+1} = \hat{y}_{n+1} + o(1) + x_{n+1}\cdot o(1)\ . \tag{2.39}$$

*(Hereafter $o(1)$ denotes a sequence of random variables which tends to 0, a.s.)*

*Proof.* The convergence (2.38) follows from Theorem 2.3. Consider

$$\hat{\beta}_{0,\text{naive}} = \bar{y} - \hat{\beta}_{1,\text{naive}}\cdot\bar{x} \xrightarrow{P1} \mathbf{E}y - K\beta_1\cdot\mathbf{E}x = \beta_0 + \beta_1\mu_\xi - K\beta_1\mu_\xi\ , \tag{2.40}$$

whence (2.37) follows. Thus, we finally have

$$\begin{aligned}\tilde{y}_{n+1} - \hat{y}_{n+1} &= (\hat{\beta}_{0,\text{naive}} - \beta_0 - \beta_1(1-K)\mu_\xi) + x_{n+1}(\hat{\beta}_{1,\text{naive}} - K\beta_1) = \\ &= o(1) + x_{n+1}\cdot o(1).\end{aligned} \tag{2.41}$$

The theorem is proved. □

Interestingly, a consistent estimator $\hat{\beta}$ is worse applicable to the prediction problem. Indeed, we would then construct a predictor

$$y'_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1}\ , \tag{2.42}$$

which at large $n$, approaches to $\beta_0 + \beta_1 x_{n+1}$ that is significantly different from the optimal predictor (2.35). The reason for this phenomenon is as follows: The precise estimation of model coefficients and accurate prediction of the next value of the response feedback are totally different statistical problems. The more so as we need a prediction based on the value $x_{n+1}$ of the surrogate variable rather than the value $\xi_{n+1}$ of the latent variable. In the latter case the next conditional expectation would be the unfeasible optimal predictor:

$$\mathbf{E}\,(y_{n+1}|\xi_{n+1}) = \beta_0 + \beta_1\xi_{n+1}\ ,$$

and then the predictor $\hat{\beta}_0 + \hat{\beta}_1\xi_{n+1}$ based on the consistent estimator would be more accurate than the predictor $\hat{\beta}_{0,\text{naive}} + \hat{\beta}_{1,\text{naive}}\cdot\xi_{n+1}$ (see Remark 2.2).

## 2.3 The linear model is not identifiable

In Section 2.2, random variables $\xi_i$ and $\delta_i$ were normal but $\varepsilon_i$ was not necessarily. Now, consider the normal model (2.1) and (2.2), i.e., the $\varepsilon_i$ will be normal as well.

(viii)Random variables $\varepsilon_i$ are identically distributed, with distribution $\mathrm{N}(0, \sigma_\varepsilon^2)$ and $\sigma_\varepsilon^2 > 0$.

In this section, we solve the following problem: *In the model* (2.1) *and* (2.2)*, assume the conditions* (i) *and* (vi)–(viii) *are satisfied; is it possible to estimate the parameters* $\beta_0$ *and* $\beta_1$ *consistently if the nuisance parameters* $\mu_\xi$, $\sigma_\xi^2$, $\sigma_\delta^2$, $\sigma_\varepsilon^2$ *are unknown?*

In total, we have six unknown model parameters that fully describe the distribution of the observed normal vector $(y; x)^{\mathrm{T}}$.

Let us give a general definition of a not identifiable observation model. Let the observed vectors be

$$Z_1, Z_2, \dots, Z_n\,, \tag{2.43}$$

with cumulative distribution function (cdf)

$$F(z_1, z_2, \dots, z_n;\ \theta)\,, \tag{2.44}$$

which depends on unknown parameter $\theta \in \Theta$, $\Theta \subset \mathbf{R}^m$.

**Definition 2.5.** The model (2.43) and (2.44) is *not identifiable* if there exist $\theta^1, \theta^2 \in \Theta$, $\theta^1 \neq \theta^2$, such that $F(z_1, \dots, z_n;\ \theta^1) \equiv F(z_1, \dots, z_n;\ \theta^2)$. Conversely, the model (2.43) and (2.44) is *identifiable* if such a couple of the parameter values does not exist.

Suppose further that the observed vectors (2.43) are independent and identically distributed. Then

$$F(z_1, \dots z_n;\ \theta) = \prod_{i=1}^{n} F(z_i; \theta)\,, \quad \theta \in \Theta\,, \tag{2.45}$$

where $F(z_i; \theta)$ is cdf of a single observation $Z_i$. In this case, the model is identifiable if, and only if, the model for a single observation $Z_i$ is identifiable.

Remember that by definition a statistical estimator of a model parameter is a Borel measurable function of the observed sample.

**Theorem 2.6.** *Consider the model* (2.43) *and* (2.44)*. If this model is not identifiable, then there is no consistent estimator of the parameter* $\theta$*.*

*Proof.* Suppose that there exists a consistent estimator $\hat\theta_n$. Let $\theta^1 \neq \theta^2$ be the values of the parameter from Definition 2.5. Then, it is true that $\hat\theta_n \to \theta^1$ in probability $\mathbf{P}_{\theta^1}$ (i.e., provided that $\theta^1$ is the true value) as well as $\hat\theta_n \to \theta^2$ in probability $\mathbf{P}_{\theta^2}$. The latter means that for each $\varepsilon > 0$,

$$\mathbf{P}_{\theta^2}\left\{\left\|\hat\theta_n(Z_1, \dots, Z_n) - \theta^2\right\| > \varepsilon\right\} \to 0\,, \quad \text{as } n \to \infty\,. \tag{2.46}$$

Identical equality of the cdf of random vectors implies the equality for the corresponding probability distributions. Therefore, the next follows from the nonidentifiability of model:

$$\mathbf{P}_{\theta^2}\{(Z_1,\dots,Z_n)\in A\}=\mathbf{P}_{\theta^1}\{(Z_1,\dots,Z_n)\in A\}\,,$$

for any Borel measurable set $A\subset\mathbf{R}^{(\dim Z)\times n}$. In particular,

$$\mathbf{P}_{\theta^2}\left\{\left\|\hat\theta_n(Z_1,\dots,Z_n)-\theta^2\right\|>\varepsilon\right\}=\mathbf{P}_{\theta^1}\left\{\left\|\hat\theta_n(Z_1,\dots,Z_n)-\theta^2\right\|>\varepsilon\right\}\,. \tag{2.47}$$

That is why $\hat\theta_n\to\theta^2$ also in probability $\mathbf{P}_{\theta^1}$. Thus, the estimator $\hat\theta_n$ simultaneously tends to $\theta^1$ and $\theta^2$ in probability $\mathbf{P}_{\theta^1}$. A limit in probability is unique, and therefore, $\theta^1=\theta^2$. The resulting contradiction proves the statement of the theorem. □

This theorem shows that a necessary condition for the existence of consistent estimator is the identifiability of the model. Thus, before one tries to construct a consistent estimator, it is worth to ensure that the model is identifiable.

It should be noted that the inverse statement to Theorem 2.6 is false: There exists an identifiable model in which it is impossible to estimate consistently all the model parameters.

It can happen that some components of the vector parameter $\theta$ are identifiable and some are not.

**Definition 2.7.** Consider the model (2.43) and (2.44), with the parameter $\theta=(\theta_1,\dots,\theta_m)^{\mathrm T}\in\Theta$. The component $\theta_i$ is called *identifiable* one if there is no $\theta^1=(\theta_1^1,\dots,\theta_m^1)^{\mathrm T}\in\Theta$ and $\theta^2=(\theta_1^2,\dots,\theta_m^2)^{\mathrm T}\in\Theta$ such that $\theta_i^1\neq\theta_i^2$ and the identity from Definition 2.5 is correct.

Clearly, the model is not identifiable if, and only if, there exists a nonidentifiable component of the parameter $\theta$.

Reasoning as in the proof of Theorem 2.6, it is possible to prove the next statement.

**Theorem 2.8.** *Assume the conditions of Theorem* 2.6. *If some component $\theta_i$ is not identifiable, then there is no consistent estimator of the component $\theta_i$.*

Thus, the identifiability of the component $\theta_i$ is necessary condition for the existence of consistent estimator for this component.

**Example 2.9.** Consider the normal linear model (2.1) and (2.2) under the conditions (i) and (vi)–(viii). The parameter $\mu_\xi$ is identifiable because $\mathbf{E}x_i=\mu_\xi$, i.e., the cdf of the observation $x_i$ will be affected by change of $\mu_\xi$. Moreover the $\mu_\xi$ can be consistently estimated, because $\bar x\xrightarrow{\mathbf{P}1}\mu_\xi$, as $n\to\infty$.

The question is whether the normal model is identifiable or not, with the model parameter

$$\theta=(\beta_0,\beta_1,\mu_\xi,\sigma_\xi^2,\sigma_\delta^2,\sigma_\varepsilon^2)^{\mathrm T}\,,\quad \Theta=\mathbf{R}^3\times(0,+\infty)^3\,. \tag{2.48}$$

**Theorem 2.10.** *The normal linear model* (2.1) *and* (2.2) *under the conditions* (i) *and* (vi)–(viii), *which has six unknown parameters, is not identifiable.*

*Proof.* Independent identically distributed normal vectors $Z_i = (x_i; y_i)^{\mathrm{T}}$, $i = \overline{1, n}$, are observed, so a single observation $Z_i$ is enough to consider when the identifiability is touched upon. The latter is a normal vector and its distribution is uniquely defined by the mean and the covariance matrix $C$. We have

$$\mathbf{E}x = \mu_\xi\,, \quad \mathbf{E}y = \beta_0 + \beta_1\mu_\xi\,, \tag{2.49}$$

$$C = \begin{pmatrix} \mathbf{D}x & \operatorname{cov}(x,y) \\ \operatorname{cov}(x,y) & \mathbf{D}y \end{pmatrix} = \begin{pmatrix} \sigma_\xi^2 + \sigma_\delta^2 & \beta_1\sigma_\xi^2 \\ \beta_1\sigma_\xi^2 & \beta_1^2\sigma_\xi^2 + \sigma_\varepsilon^2 \end{pmatrix}. \tag{2.50}$$

To prove the nonidentifiability, two different sets of model parameters should be considered:

(a)

$$\beta_0 = \mu_\xi = 0\,, \quad \beta_1 = \sigma_\xi^2 = \sigma_\varepsilon^2 = \sigma_\delta^2 = 1\,. \tag{2.51}$$

Then,

$$\mathbf{E}x = \mathbf{E}y = 0\,, \quad C = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}. \tag{2.52}$$

(b)

$$\beta_0 = \mu_\xi = 0\,, \quad \beta_1 = \frac{2}{3}\,, \quad \sigma_\xi^2 = \frac{3}{2}\,, \quad \sigma_\varepsilon^2 = \frac{4}{3}\,, \quad \sigma_\delta^2 = \frac{1}{2}\,. \tag{2.53}$$

For the letter set of parameters, equalities (2.52) hold as well. In particular,

$$\mathbf{D}y = \left(\frac{2}{3}\right)^2 \cdot \frac{3}{2} + \frac{4}{3} = \frac{2}{3} + \frac{4}{3} = 2\,. \tag{2.54}$$

Thus, for both sets of parameters (2.51) and (2.53), the distribution of the observed vector $Z_1$ is the same, and therefore, the model is not identifiable. □

**Remark 2.11.** Example 2.9 contains a nonidentifiable model having an identifiable parameter.

Explanation of the model nonidentifiability inferred from Theorem 2.10 is quite simple: the joint distribution of Gaussian observations is uniquely defined by five characteristics $\mathbf{E}x$, $\mathbf{E}y$, $\mathbf{D}x$, $\mathbf{D}y$, and $\operatorname{cov}(x, y)$, whereas the model has more free parameters (namely six).

Theorem 2.10 implies that without additional information, it is impossible to estimate consistently all the six unknown parameters in the context of the normal linear measurement error model (this can be done only for the parameter $\mu_\xi$). Imposing additional (prior) relationships among the parameters should improve the situation.

The most common are the two alternative constraints:

(a)

$$\text{the ratio } \lambda = \frac{\sigma_\varepsilon^2}{\sigma_\delta^2} \text{ is known, or} \tag{2.55}$$

(b)

$$\text{the measurement error variance } \sigma_\delta^2 \text{ is known.} \tag{2.56}$$

As we will see later, in both cases the model becomes identifiable (in fact, we narrow the parameter set $\Theta$). Moreover for these cases, we will construct consistent estimators for all the parameters of the model.

The assumption (2.55) is more common than (2.56). For example, if the response and regressor have the same physical dimension, at that measuring the response and regressor is being carried out by the same (or similar) physical device, the observations may be considered as equally precise, i.e., with $\sigma_\varepsilon^2 = \sigma_\delta^2$ or $\lambda = 1$.

## 2.4 The model with known error variance

This section studies the model (2.1) and (2.2) under the conditions (i)–(iv). The value $\sigma_\delta^2$ is assumed known. We will derive the same consistent estimator of the slope $\beta_1$ using three ways. Also, the consistent estimators for the other model parameters will be written down.

### 2.4.1 The adjusted naive estimator of the slope

According to (2.38), $\hat{\beta}_{1,\text{naive}}$ converges to $K\beta_1$ , a.s. We want to adjust this estimator so that a new estimator is consistent. For this, estimate consistently the reliability ratio

$$K = \frac{\sigma_\xi^2}{\sigma_\xi^2 + \sigma_\delta^2} \,. \tag{2.57}$$

The value $\sigma_\delta^2$ is known to us, and as a consistent estimator one may take

$$\hat{\sigma}_\xi^2 = \mathrm{S}_{xx} - \sigma_\delta^2 \,; \quad \hat{\sigma}_\xi^2 \xrightarrow{P1} \mathbf{D}x - \sigma_\delta^2 = \sigma_\xi^2 \,. \tag{2.58}$$

In case of the normal model, this estimator is the maximum likelihood estimator (MLE), see Section 1.4.3.

Then the consistent estimator of the $K$ is

$$\hat{K} = \frac{\hat{\sigma}_\xi^2}{\hat{\sigma}_\xi^2 + \sigma_\delta^2} = \frac{\mathrm{S}_{xx} - \sigma_\delta^2}{\mathrm{S}_{xx}} \,. \tag{2.59}$$

The next definition is useful while studying the asymptotic behavior of estimators.

**Definition 2.12.** Consider the sequence of statements $A_n = A_n(\omega)$ that depend on an elementary event $\omega$ from a probability space $\Omega$. The statement $A_n$ is said to hold *eventually* if there exists such a random event $\Omega_0$, with $\mathbf{P}(\Omega_0) = 1$, that for each $\omega \in \Omega_0$, there exists a number $n_0 = n_0(\omega)$ such that for all $n \geq n_0$, the statement $A_n(\omega)$ holds true.

In particular, $S_{xx} > 0$ *eventually*, because $S_{xx} \xrightarrow{P1} \mathbf{D}x > 0$. Thus, *eventually* the estimator $\hat{K}$ is correctly defined by (2.59). Moreover, $\hat{K} > 0$ *eventually*, because $\hat{K} \xrightarrow{P1} K > 0$.

Taking into account (2.38), the adjusted estimator is given by the following expression:

$$\hat{\beta}_1 = \frac{1}{\hat{K}}\hat{\beta}_{1,\text{naive}} = \frac{S_{xx}}{S_{xx} - \sigma_\delta^2} \cdot \frac{S_{xy}}{S_{xx}}\,, \tag{2.60}$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx} - \sigma_\delta^2}\,. \tag{2.61}$$

It is clear that $S_{xx} - \sigma_\delta^2 > 0$,*eventually*. If it so happens that $S_{xx}(\omega) = \sigma_\delta^2$ for some $n$ and $\omega$, then one may set $\hat{\beta}_1 = 0$. Our estimator $\hat{\beta}_1$ can be considered as defined by (2.61), *eventually*.

**Definition 2.13.** The estimator (2.61) is called the adjusted least squares (ALS) estimator.

The next theorem follow from Theorem 2.4 and from our procedure of constructing the estimator.

**Theorem 2.14.** *In the linear model, assume the conditions* (i)–(iv). *Then the ALS estimator* (2.61) *is a strongly consistent estimator of the slope, i.e.,*

$$\hat{\beta}_1 = \hat{\beta}_{1,\text{ALS}} \xrightarrow{P1} \beta_1\,, \quad \textit{as} \;\; n \to \infty\,. \tag{2.62}$$

### 2.4.2 The corrected score estimator of regression parameters

We apply the CS method described in Section 1.4.4. Our linear model is a particular case of the model (1.1) and (1.5), with regression function $f(\xi_i, \beta) = \beta_0 + \beta_1\xi_i$.

The deconvolution equations (1.124) and (1.125) take the form

$$\mathbf{E}[g(x, b)|\xi] = \frac{\partial f}{\partial \beta} = \begin{pmatrix} 1 \\ \xi \end{pmatrix}\,, \tag{2.63}$$

$$\mathbf{E}[h(x, b)|\xi] = f\frac{\partial f}{\partial \beta} = \begin{pmatrix} b_0 + b_1\xi \\ b_0\xi + b_1\xi^2 \end{pmatrix}\,. \tag{2.64}$$

We search for the solutions in the class of polynomial functions in $x$. It is clear that $g(x, b) = (1;\, x)^{\mathrm{T}}$. To find the second function $h$ is to construct a polynomial $t_2(x)$ such that

$$\mathbf{E}[t_2(x)|\xi] = \xi^2. \tag{2.65}$$

It is easy to verify that $t_2(x) = x^2 - \sigma_\delta^2$ satisfies (2.65). Indeed

$$\begin{aligned}\mathbf{E}[(\xi + \delta)^2 - \sigma_\delta^2|\xi] &= \xi^2 + 2\xi \cdot \mathbf{E}[\delta|\xi] + \mathbf{E}[\delta^2|\xi] - \sigma_\delta^2 = \\ &= \xi^2 + 2\xi \cdot \mathbf{E}\delta + \mathbf{E}\delta^2 - \sigma_\delta^2 = \xi^2\,.\end{aligned} \tag{2.66}$$

Therefore,

$$h(x, b) = \begin{pmatrix} b_0 + b_1 x \\ b_0 x + b_1 x^2 - b_1 \sigma_\delta^2 \end{pmatrix} . \tag{2.67}$$

The estimating function (1.126) takes the form

$$s_{\mathrm{C}}(y, x; b) = y \cdot \begin{pmatrix} 1 \\ x \end{pmatrix} - \begin{pmatrix} b_0 + b_1 x \\ b_0 x + b_1 x^2 - b_1 \sigma_\delta^2 \end{pmatrix} . \tag{2.68}$$

Then the CS estimator of $\beta$ is given by the equation

$$\frac{1}{n} \sum_{i=1}^{n} s_{\mathrm{C}}(y_i, x_i; b) = 0 , \quad b \in \mathbf{R}^2 , \tag{2.69}$$

or a system of equations

$$\begin{cases} \bar{y} - b_0 - b_1 \bar{x} = 0 , \\ \overline{xy} - b_0 \bar{x} - b_1 (\overline{x^2} - \sigma_\delta^2) = 0 . \end{cases} \tag{2.70}$$

Eliminating $b_0$ from the second equation, we get

$$b_1 (\overline{x^2} - (\bar{x})^2 - \sigma_\delta^2) = \overline{xy} - \bar{x} \cdot \bar{y} , \tag{2.71}$$

$$\hat{\beta}_{1,\mathrm{CS}} = \frac{\mathrm{S}_{xy}}{\mathrm{S}_{xx} - \sigma_\delta^2} . \tag{2.72}$$

As we can see, this estimator coincides with the ALS estimator (2.61), *eventually*.

The estimator $\hat{\beta}_{0,\mathrm{CS}}$ can be now found from the first equation of the system (2.70)

$$\hat{\beta}_{0,\mathrm{CS}} = \bar{y} - \hat{\beta}_{1,\mathrm{CS}} \cdot \bar{x} . \tag{2.73}$$

As with the naive estimator, the estimated straight line $y = \hat{\beta}_{0,\mathrm{CS}} + \hat{\beta}_{1,\mathrm{CS}} \xi$ goes through the center of mass $M(\bar{x}; \bar{y})$ for systems of points $M_i(x_i; y_i)$, $i = \overline{1, n}$ (see Remark 2.2). It is clear that as a consequence of Theorem 2.14, the estimator (2.73) is a strongly consistent estimator of the parameter $\beta_0$.

### 2.4.3 Maximum likelihood estimator of all the parameters

Consider the normal linear model, i.e., assume the conditions (i) and (vi)–(viii). To write down an equation for the MLE of five unknown parameters, there is no need to write out the pdf of the observed normal vector $(y; x)^{\mathrm{T}}$. Instead, one can use the so-called *functional invariance* of the MLE (Kendall and Stuart, 1979, Chapter 18). For this, the equalities (2.49) and (2.50) should be rewritten in the following manner: the expectations are replaced by the sample means, and the variance and covariance are

replaced by the sample counterparts as well. Thus, we have a system for the MLE:

$$\bar{x} = \hat{\mu}_\xi\,, \quad \bar{y} = \hat{\beta}_0 + \hat{\beta}_1\hat{\mu}_\xi\,, \tag{2.74}$$

$$\mathrm{S}_{xx} = \hat{\sigma}_\xi^2 + \sigma_\delta^2\,, \quad \mathrm{S}_{xy} = \hat{\beta}_1\hat{\sigma}_\xi^2\,, \quad \mathrm{S}_{yy} = \hat{\beta}_1^2\hat{\sigma}_\xi^2 + \hat{\sigma}_\varepsilon^2\,, \tag{2.75}$$

$$\sigma_\xi^2 > 0\,, \quad \sigma_\varepsilon^2 > 0\,. \tag{2.76}$$

So, the solution to equations (2.69) and (2.70) is as follows:

$$\hat{\mu}_\xi = \bar{x}\,, \quad \hat{\sigma}_\xi^2 = \mathrm{S}_{xx} - \sigma_\delta^2\,, \tag{2.77}$$

$$\hat{\beta}_1 = \frac{\mathrm{S}_{xy}}{\mathrm{S}_{xx} - \sigma_\delta^2}\,, \quad \hat{\beta}_0 = \bar{y} - \bar{x}\cdot\hat{\beta}_1\,, \tag{2.78}$$

$$\hat{\sigma}_\varepsilon^2 = \mathrm{S}_{yy} - \frac{(\mathrm{S}_{xy})^2}{\mathrm{S}_{xx} - \sigma_\delta^2}\,. \tag{2.79}$$

The inequality $\mathrm{S}_{xx} > \sigma_\delta^2$ holds *eventually* (see Section 2.4.1), so

$$\hat{\sigma}_\xi^2 > 0\,, \quad \textit{eventually}\,. \tag{2.80}$$

Further, by the SLLN,

$$\hat{\sigma}_\varepsilon^2 \xrightarrow{P1} \mathbf{D}y - \frac{(\mathrm{cov}(x,y))^2}{\mathbf{D}x - \sigma_\delta^2} = \beta_1^2\sigma_\xi^2 + \sigma_\varepsilon^2 - \frac{\beta_1^2\sigma_\xi^4}{\sigma_\xi^2} = \sigma_\varepsilon^2 > 0\,, \tag{2.81}$$

$$\hat{\sigma}_\varepsilon^2 > 0\,, \quad \textit{eventually}\,. \tag{2.82}$$

Thus, *eventually* both inequalities (2.80) and (2.82) hold and the expressions (2.77)–(2.79) are those for the MLE, *eventually*. The expressions (2.77) and (2.78) were met by us before, but the estimator (2.79) is new for us. The latter estimator is consistent without the requirement of normality of errors and $\xi_i$.

**Theorem 2.15.** *Assume the conditions* (i)–(iv). *Then the expression* (2.81) *is a strongly consistent estimator of the error variance* $\sigma_\varepsilon^2$.

*The proof* follows from the convergence (2.81) which is valid under the assumptions (i)–(iv).

**Remark 2.16.** Under the assumptions (vi) and (vii), when $n \geq 2$, the expressions (2.78) and (2.79) are well-defined, almost surely, i.e.,

$$\mathrm{S}_{xx} \neq \sigma_\delta^2\,, \ \text{a.s.} \tag{2.83}$$

*Proof.* Provided (vi) and (vii), the surrogate variables $x_i$ are independent with the distribution $\mathrm{N}(\mu_\xi, \sigma_x^2)$, $\sigma_x^2 = \sigma_\xi^2 + \sigma_\delta^2 > 0$. Therefore,

$$\mathrm{S}_{xx} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \sim \frac{\sigma_x^2}{n}\chi_{n-1}^2\,, \tag{2.84}$$

where $\chi^2_\nu$ is a random variable having *chi-square distribution with $\nu$ degrees of freedom*, i.e., $\chi^2_\nu$ has the same distribution as $\sum_{i=1}^{\nu} \gamma_i^2$, where $\gamma_i$, $i = \overline{1,\nu}$, are independent standard normal random variables (Seber and Lee, 2003). It is known that $\chi^2_\nu$ has a pdf. Therefore,

$$\mathbf{P}\{\mathrm{S}_{xx} = \sigma_\delta^2\} = \mathbf{P}\left\{\frac{\sigma_x^2}{n}\chi^2_{n-1} = \sigma_\delta^2\right\} = \mathbf{P}\left\{\chi^2_{n-1} = \frac{n\sigma_\delta^2}{\sigma_x^2}\right\} = 0\,, \tag{2.85}$$

which proves the statement (2.83). □

### 2.4.4 Asymptotic normality of the estimator for the slope

By the *sandwich formula* (see Appendix A2), it can be shown that the augmented estimator $\hat\theta = (\hat\beta_0, \hat\beta_1, \hat\sigma_\varepsilon^2, \hat\mu_\xi, \hat\sigma_\xi^2)^{\mathrm{T}}$ of the parameter $\theta = (\beta_0, \beta_1, \sigma_\varepsilon^2, \mu_\xi, \sigma_\xi^2)^{\mathrm{T}}$ is asymptotically normal in the normal linear model, i.e.,

$$\sqrt{n}(\hat\theta - \theta) \xrightarrow{d} \mathrm{N}(\bar{0}, \Sigma_\theta)\,. \tag{2.86}$$

Here, $\Sigma_\theta$ being dependent on $\theta$ is the asymptotic covariance matrix (ACM). It is positive definite. The condition that the model is normal can be relaxed significantly. The expression (2.86) permits to construct an asymptotic confidence ellipsoid for the vector $\theta$.

However, the use of sandwich formula, in this instance, necessitates cumbersome calculations. We derive the asymptotic normality only for the ALS estimator of $\hat\beta_{1,\mathrm{ALS}}$ by utilizing its explicit formula.

Here are some known facts from stochastic analysis (Schervish, 1995).

**Definition 2.17.** A sequence $\{x_n\}$ of random variables is called *stochastically bounded* if

$$\sup_{n\geq 1} \mathbf{P}\{|x_n| > c\} \to 0\,, \quad \text{as } c \to +\infty\,. \tag{2.87}$$

For such a sequence, we write

$$x_n = O_p(1)\,. \tag{2.88}$$

Also denote $y_n = o_p(1)$ if $y_n \xrightarrow{P} 0$.

**Lemma 2.18** (Slutsky's lemma). *Let $\xi_n \xrightarrow{d} \xi$ and $\eta_n \xrightarrow{P} 0$ hold true. Then $\xi_n + \eta_n \xrightarrow{d} \xi$.*

**Lemma 2.19.** *If $\xi_n \xrightarrow{d} \xi$, then $\xi_n = O_p(1)$.*

**Lemma 2.20.** *If $\xi_n = o_p(1)$, $\eta_n = O_p(1)$, then $\xi_n\eta_n = o_p(1)$.*

**Corollary 2.21.** *If $x_n \xrightarrow{\mathbf{P}} a$ and $y_n \xrightarrow{\mathrm{d}} y$, then $x_n y_n \xrightarrow{\mathrm{d}} ay$. Here $a$ is a real number.*

*Proof.* We have

$$x_n y_n = (x_n - a)\, y_n + a y_n \,. \tag{2.89}$$

By Lemmas 2.19 and 2.20, the first summand is $o_p(1) \cdot O_P(1) = o_p(1)$, and the second summand converges in distribution to $ay$. Then by Slutsky's lemma, $x_n y_n \xrightarrow{\mathrm{d}} ay$. □

**Theorem 2.22.** *Assume the conditions* (i), (ii), (iv), *and* (vi). *Then the estimator* $\hat\beta_1 = \hat\beta_{1,\mathrm{ALS}}$ *given by formula* (2.61) *is asymptotically normal, in more detail*

$$\sqrt{n}(\hat\beta_1 - \beta_1) \xrightarrow{d} \mathrm{N}(0, \sigma_1^2)\,, \tag{2.90}$$

$$\sigma_1^2 = \frac{1}{\sigma_\xi^4}\left(\sigma_\varepsilon^2 \sigma_x^2 + \beta_1^2(\sigma_x^2 \sigma_\delta^2 + \sigma_\delta^4)\right)\,, \tag{2.91}$$

$$\sigma_x^2 = \mathbf{D}x = \sigma_\xi^2 + \sigma_\delta^2\,. \tag{2.92}$$

*Proof.* The sample covariance

$$\mathrm{S}_{uv} = \overline{(u - \bar u)\,(v - \bar v)} \tag{2.93}$$

between two samples $u_1, \dots, u_n$ and $v_1, \dots v_n$ is linear in both $u$ and $v$. In particular, this means that for the sum of two samples $w = u + z = (u_1 + z_1, u_2 + z_2, \dots, u_n + z_n)$, it holds that

$$\mathrm{S}_{wv} = \mathrm{S}_{u+z,v} = \mathrm{S}_{uv} + \mathrm{S}_{zv}\,. \tag{2.94}$$

Moreover, if $u_i = \mathrm{const}$, $i = \overline{1, n}$, then $\mathrm{S}_{uv} = 0$.

Since in the linear model,

$$x = \xi + \delta\,, \quad y = \beta_0 + \beta_1 \rho + \beta_1 \mu_\xi + \varepsilon\,, \quad \rho = \xi - \mu_\xi\,, \tag{2.95}$$

then, by the linearity of the operator (2.88), we obtain

$$\mathrm{S}_{xy} = \beta_1\, \mathrm{S}_{\rho\rho} + \mathrm{S}_{\rho\varepsilon} + \beta_1\, \mathrm{S}_{\delta\rho} + \mathrm{S}_{\delta\varepsilon}\,, \tag{2.96}$$

$$\mathrm{S}_{xx} = \mathrm{S}_{\rho\rho} + 2\, \mathrm{S}_{\rho\delta} + \mathrm{S}_{\delta\delta}\,. \tag{2.97}$$

Consider the denominator of the fraction (2.61):

$$\mathrm{S}_{xx} - \sigma_\delta^2 \xrightarrow{\mathbf{P1}} \sigma_x^2 - \sigma_\delta^2 = \sigma_\xi^2\,, \tag{2.98}$$

$$\mathrm{S}_{xx} - \sigma_\delta^2 = \sigma_\xi^2 + o(1)\,. \tag{2.99}$$

Remember that $o(1)$ denotes a sequence of random variables $z_n$ that tends to 0, a.s. It is clear that $z_n = o_p(1)$ as well. We have

$$\sqrt{n}(\hat\beta_1 - \beta_1) = \frac{-\beta_1 \sqrt{n}(\mathrm{S}_{xx} - \sigma_\delta^2) + \sqrt{n}\, \mathrm{S}_{xy}}{\mathrm{S}_{xx} - \sigma_\delta^2}\,, \tag{2.100}$$

$$\sqrt{n}(\hat\beta_1 - \beta_1) = \frac{-\beta_1 \sqrt{n}(\mathrm{S}_{\delta\rho} + \mathrm{S}_{\delta\delta} - \sigma_\delta^2) + \sqrt{n}(\mathrm{S}_{\rho\varepsilon} + \mathrm{S}_{\delta\varepsilon})}{\sigma_\xi^2 + o(1)}\,. \tag{2.101}$$

Next,

$$\sqrt{n}\,\mathrm{S}_{\delta\rho} = \sqrt{n}\cdot\overline{\delta\rho} - \sqrt{n}\cdot\bar{\delta}\cdot\bar{\rho} = \sqrt{n}\cdot\overline{\delta\rho} - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\delta_i\cdot\bar{\rho}\,. \tag{2.102}$$

Using the central limit theorem (CLT) and SLLN, we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\delta_i \xrightarrow{\mathrm{d}} \mathrm{N}(0,\sigma_\delta^2)\,, \qquad \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\delta_i = O_p(1)\,, \tag{2.103}$$

$$\begin{aligned}\bar{\rho} \xrightarrow{\mathbf{P1}} \mathbf{E}\rho = 0\,, \quad \bar{\rho} = o(1)\,, \quad \bar{\rho} = o_p(1)\,,\\ (\sqrt{n}\cdot\bar{\delta})\cdot\bar{\rho} = O_p(1)\cdot o_p(1) = o_p(1)\,.\end{aligned} \tag{2.104}$$

Therefore,

$$\sqrt{n}\,\mathrm{S}_{\delta\rho} = \sqrt{n}\cdot\overline{\delta\rho} + o_p(1)\,. \tag{2.105}$$

Similarly:

$$\begin{aligned}\sqrt{n}\,\mathrm{S}_{\rho\varepsilon} &= \sqrt{n}\cdot\overline{\rho\varepsilon} + o_p(1)\,,\\ \sqrt{n}\,\mathrm{S}_{\delta\varepsilon} &= \sqrt{n}\cdot\overline{\delta\varepsilon} + o_p(1)\,,\\ \sqrt{n}\,\mathrm{S}_{\delta\delta} &= \sqrt{n}\cdot\overline{\delta^2} + o_p(1)\,.\end{aligned} \tag{2.106}$$

Substituting (2.105) and (2.106) into (2.101), we get

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) = \frac{-\beta_1\sqrt{n}\left(\overline{\delta\rho} + \left(\overline{\delta^2} - \sigma_\delta^2\right)\right) + \sqrt{n}(\overline{\rho\varepsilon} + \overline{\delta\varepsilon})}{\sigma_\xi^2 + o(1)} + o_p(1)\,. \tag{2.107}$$

Prove the convergence in distribution for the numerator in (2.107). According to the CLT (Kartashov, 2007),

$$\sqrt{n}\left(\overline{\delta\rho}, \overline{\delta^2 - \sigma_\delta^2}, \overline{\rho\varepsilon}, \overline{\delta\varepsilon}\right)^{\mathrm{T}} \xrightarrow{\mathrm{d}} \gamma = (\gamma_1,\ldots,\gamma_4)^{\mathrm{T}} \sim \mathrm{N}(0, S)\,, \tag{2.108}$$

$$S = \mathrm{diag}(\sigma_\delta^2\sigma_\xi^2, \mathbf{D}\delta^2, \sigma_\xi^2\sigma_\varepsilon^2, \sigma_\delta^2\sigma_\varepsilon^2)\,. \tag{2.109}$$

Here the diagonal entries of the matrix $S = (S_{ij})_{i,j=1}^4$ are composed of the variances of random variables that were averaged. In particular,

$$S_{11} = \mathbf{D}(\delta\rho) = \mathbf{E}(\delta\rho)^2 = \mathbf{E}\delta^2\cdot\mathbf{E}\rho^2 = \sigma_\delta^2\sigma_\xi^2\,. \tag{2.110}$$

The off-diagonal elements of the matrix $S$ are equal to 0, because $\delta$, $\rho$, and $\varepsilon$ are independent, e.g.,

$$S_{12} = \mathbf{E}\delta\rho(\delta^2 - \sigma_\delta^2) = \mathbf{E}\rho\cdot\mathbf{E}\delta(\delta^2 - \sigma_\delta^2) = 0\,. \tag{2.111}$$

Moreover, since $\delta \sim \mathrm{N}(0,\sigma_\delta^2)$,

$$\mathbf{D}\delta^2 = \mathbf{E}\delta^4 - (\mathbf{E}\delta^2)^2 = 3\sigma_\delta^4 - \sigma_\delta^4 = 2\sigma_\delta^4\,. \tag{2.112}$$

The convergence (2.108) implies that the numerator in (2.107) converges in distribution to

$$-\beta_1(\gamma_1+\gamma_2)+\gamma_3+\gamma_4 \sim \mathrm{N}(0,\beta_1^2(S_{11}+S_{22})+S_{33}+S_{44})\,. \tag{2.113}$$

Then, by Corollary 2.21 and Lemma 2.18, we have the following from relationships (2.107), (2.109), and (2.113):

$$\sqrt{n}(\hat\beta_1-\beta_1) \xrightarrow{\mathrm{d}} \mathrm{N}(0,\sigma_1^2)\,, \tag{2.114}$$

$$\sigma_1^2=\frac{1}{\sigma_\xi^4}(\beta_1^2(\sigma_\xi^2\sigma_\delta^2+\sigma_\delta^4)+\sigma_\varepsilon^2\sigma_\xi^2+\sigma_\varepsilon^2\sigma_\delta^2)\,, \tag{2.115}$$

$$\sigma_1^2=\frac{1}{\sigma_\xi^4}(\beta_1^2(\sigma_x^2\sigma_\delta^2+\sigma_\delta^4)+\sigma_\varepsilon^2\sigma_x^2)\,. \tag{2.116}$$

This proves the theorem. □

The asymptotic variance $\sigma_1^2$ contains unknown parameters (only $\sigma_\delta^2$ is assumed known). The strongly consistent estimator of the $\sigma_1^2$ can be constructed as follows:

$$\hat\sigma_1^2=\frac{1}{\hat\sigma_\xi^4}(\hat\beta_1^2(\hat\sigma_x^2\sigma_\delta^2+\sigma_\delta^4)+\hat\sigma_\varepsilon^2\hat\sigma_x^2)\,,\quad \hat\sigma_x^2=\hat\sigma_\xi^2+\sigma_\delta^2\,, \tag{2.117}$$

with estimators of model parameters being set in (2.77)–(2.79). Under the conditions of Theorem 2.22, these estimators converge almost surely to the corresponding true values, and therefore,

$$\hat\sigma_1^2 \xrightarrow{\mathbf{P}1} \sigma_1^2\,,\quad \hat\sigma_1=\sqrt{\hat\sigma_1^2} \xrightarrow{\mathbf{P}1} \sigma_1=\sqrt{\sigma_1^2}\,. \tag{2.118}$$

To construct a confidence interval of the parameter $\beta_1$, consider the statistic

$$\frac{\sqrt{n}(\hat\beta_1-\beta_1)}{\hat\sigma_1}=\frac{\sqrt{n}(\hat\beta_1-\beta_1)}{\sigma_1}\cdot\frac{\sigma_1}{\hat\sigma_1} \xrightarrow{\mathrm{d}} \mathrm{N}(0,1)\,. \tag{2.119}$$

We took advantage of the convergence in (2.114) and (2.118) and of Corollary 2.21. If the confidence probability is specified to be 0.99, then the asymptotic confidence interval can be taken as

$$I_n=\left[\hat\beta_1-\frac{3\hat\sigma_1}{\sqrt{n}},\hat\beta_1+\frac{3\hat\sigma_1}{\sqrt{n}}\right]\,. \tag{2.120}$$

Indeed, (2.119) implies that for $\gamma \sim \mathrm{N}(0,1)$,

$$\lim_{n\to\infty}\mathbf{P}\{\beta_1\in I_n\}=\lim_{n\to\infty}\mathbf{P}\left\{\left|\frac{\sqrt{n}(\hat\beta_1-\beta_1)}{\hat\sigma_1}\right|\le 3\right\}=\mathbf{P}\{|\gamma|\le 3\}\ge 0.99\,. \tag{2.121}$$

Thus, the asymptotic confidence interval has been constructed.

A disadvantage of measurement error models is that even under the conditions (i), (vi)–(viii) of the model normality with known $\sigma_\delta^2>0$, it is in principle impossible to

construct a finite nonasymptotic confidence interval $J_n$ for $\beta_1$. Such an interval should satisfy

$$\inf_{\theta\in\Theta} \mathbf{P}_\theta\{\beta_1 \in J_n\} \geq 1-\gamma\,. \tag{2.122}$$

Here $1-\gamma > 0$ is the confidence probability, $\theta$ is the augmented vector of unknown model parameters. In fact, the left-hand side of (2.122) is equal to 0. This phenomenon is called *Gleser–Hwang effect* (Cheng and Van Ness, 1999). It is explained by the fact that the structural model (2.1) and (2.2), at small $\sigma_\xi^2$, can be arbitrarily close to the degenerate model

$$y_i = \beta_0 + \beta_1\mu_\xi + \varepsilon_i\,, \quad x_i = \mu_\xi + \delta_i\,, \quad i = \overline{1,n}\,. \tag{2.123}$$

But the latter is not identifiable because for fixed $\mu_\xi$, there are many couples $(\beta_0, \beta_1)$ providing the fixed value of $\mathbf{E}y_i = \beta_0 + \beta_1\mu_\xi$.

The Gleser–Hwang effect disappears if the parameter set $\Theta$ is specified so that $\sigma_\xi^2$ is separated away from 0. Namely, it should hold

$$\sigma_\xi^2 \geq \text{const} > 0\,, \quad \text{for any } \theta \in \Theta\,. \tag{2.124}$$

### 2.4.5 Bias of the naive estimator and nonexistence of expectation of ALS estimator

Consider a question on the existence of expectation for the naive estimator and the ALS estimator.

**Bias of the naive estimator**

**Theorem 2.23.** *Let the conditions* (i), (ii), (vi), *and* (vii) *hold true. Then with* $n \geq 3$,

$$\mathbf{E}\hat\beta_{1,\text{naive}} = \beta_1 K\,, \quad K = \frac{\sigma_\xi^2}{\sigma_\xi^2+\sigma_\delta^2}\,. \tag{2.125}$$

*Proof.* Using the linearity (2.94) of the sample covariance operator, we have a.s., that

$$\hat\beta_{1,\text{naive}} = \frac{\mathrm{S}_{xy}}{\mathrm{S}_{xx}} = \beta_1\frac{\mathrm{S}_{x\xi}}{\mathrm{S}_{xx}} + \frac{\mathrm{S}_{x\varepsilon}}{\mathrm{S}_{xx}}\,. \tag{2.126}$$

Verify that the expectations of both summands are finite. It holds that

$$|\,\mathrm{S}_{x\varepsilon}\,| = \frac{1}{n}\left|\sum_{i=1}^n (x_i-\bar x)\,(\varepsilon_i-\bar\varepsilon)\right| \leq \sqrt{\frac{1}{n}\sum_{i=1}^n (x_i-\bar x)^2}\,\sqrt{\frac{1}{n}\sum_{i=1}^n (\varepsilon_i-\bar\varepsilon)^2} = \mathrm{S}_{xx}^{1/2}\,\mathrm{S}_{\varepsilon\varepsilon}^{1/2}\,. \tag{2.127}$$

Therefore, as a result of mutual independence of $\{x_i\}$ and $\{\varepsilon_i\}$

$$\mathbf{E}\frac{|\,\mathrm{S}_{x\varepsilon}\,|}{\mathrm{S}_{xx}} \leq \mathbf{E}\frac{\mathrm{S}_{\varepsilon\varepsilon}^{1/2}}{\mathrm{S}_{xx}^{1/2}} = \mathbf{E}\frac{1}{\mathrm{S}_{xx}^{1/2}}\cdot\mathbf{E}\,\mathrm{S}_{\varepsilon\varepsilon}^{1/2}\,. \tag{2.128}$$

Here the second factor is finite but it is not so obvious for the first one. Explain why this is so. According to (2.84), $S_{xx} \sim \sigma_x^2 \cdot n^{-1} \cdot \chi^2_{n-1}$, thus,

$$\mathbf{E}\frac{1}{S_{xx}^{1/2}} = \frac{\sqrt{n}}{\sigma_x}\mathbf{E}\frac{1}{\sqrt{\chi^2_{n-1}}} . \tag{2.129}$$

Remember that $\chi^2_{n-1} = \sum_{i=1}^{n-1} \gamma_i^2$, where $\gamma_i$ are independent standard normal random variables. The augmented vector $(\gamma_1, \dots, \gamma_{n-1})^{\mathrm{T}}$ has a pdf

$$\frac{1}{(\sqrt{2\pi})^{n-1}} \mathrm{e}^{-\frac{\|x\|^2}{2}} , \quad x \in \mathbf{R}^{n-1} . \tag{2.130}$$

Therefore,

$$\mathbf{E}\frac{1}{\sqrt{\chi^2_{n-1}}} = \mathbf{E}\left(\sum_{i=1}^{n-1} \gamma_i^2\right)^{-1/2} = \mathrm{const} \cdot \int\limits_{\mathbf{R}^{n-1}} \frac{1}{\|x\|} \mathrm{e}^{-\frac{\|x\|^2}{2}} dx . \tag{2.131}$$

In the integral, move the generalized spherical coordinates. The transition Jacobian contains the factor $r^{n-2}$ in which $r = \|x\|$ is the Euclidean vector norm. Then for $n \geq 3$,

$$\mathbf{E}\frac{1}{\sqrt{\chi^2_{n-1}}} = \mathrm{const} \cdot \int\limits_0^\infty \frac{1}{r} \mathrm{e}^{-\frac{r^2}{2}} \cdot r^{n-2} dr = \mathrm{const} \cdot \int\limits_0^\infty r^{n-3} \mathrm{e}^{-\frac{r^2}{2}} dr < \infty. \tag{2.132}$$

This justifies the equality in (2.128), and that

$$\mathbf{E}\frac{|S_{x\varepsilon}|}{S_{xx}} < \infty , \quad \text{and similarly } \mathbf{E}\frac{|S_{x\xi}|}{S_{xx}} < \infty . \tag{2.133}$$

Further, for $\vec{x} = (x_1, \dots, x_n)^{\mathrm{T}}$, we have

$$\mathbf{E}\frac{S_{x\varepsilon}}{S_{xx}} = \mathbf{E}\mathbf{E}\left[ \frac{S_{x\varepsilon}}{S_{xx}} \,\middle|\, \vec{x} \right] = 0 , \tag{2.134}$$

because

$$\begin{aligned} \mathbf{E}\left[ \frac{S_{x\varepsilon}}{S_{xx}} \,\middle|\, \vec{x} \right] &= \frac{1}{S_{xx}} \cdot \frac{1}{n} \sum_{i=1}^{n} \mathbf{E}[(x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})|\vec{x}] = \\ &= \frac{1}{n S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) \mathbf{E}(\varepsilon_i - \bar{\varepsilon}) = 0 . \end{aligned} \tag{2.135}$$

In addition, under the normality conditions (vi) and (vii), the relations (1.86) from Section 1.4.3 hold. Let $\{\gamma_i, i = \overline{1, n}\}$ be a set of $n$ independent standard normal random variables being independent of $\{x_i\}$. Then,

$$\begin{aligned} \mathbf{E}\frac{S_{x\xi}}{S_{xx}} &= \mathbf{E}\mathbf{E}\left[ \frac{S_{x\xi}}{S_{xx}} \,\middle|\, \vec{x} \right] = \mathbf{E}\frac{1}{S_{xx}}\mathbf{E}[S_{x\xi}|\vec{x}] = \\ &= \mathbf{E}\frac{1}{S_{xx}}\mathbf{E}[S_{x,Kx+(1-K)\mu_\xi+\tau\gamma}|\vec{x}] = \mathbf{E}\frac{S_{x,Kx}}{S_{xx}} + \mathbf{E}[\frac{S_{x,\tau\gamma}}{S_{xx}}|\vec{x}] = K\mathbf{E}\frac{S_{xx}}{S_{xx}} = K . \end{aligned} \tag{2.136}$$

From the equalities (2.126), (2.134), and (2.136), we finally get (2.125). □

**Corollary 2.24.** *If $\beta_1 \neq 0$ and under the conditions of Theorem* 2.23*, the naive estimator is biased, i.e.,*

$$\mathbf{E}\hat{\beta}_{1,\text{naive}} \neq \beta_1 \,. \tag{2.137}$$

*Theorem* 2.23 *strengthens the attenuation effect (see Section* 2.1*): not only for large $n$, the naive estimator is closer to zero than $\beta_1$, but having fixed $n$ it is also shifted to zero, i.e., $\mathbf{E}\hat{\beta}_1$ is closer to zero (actually it is located between 0 and $\beta_1$).*

**Remark 2.25.** Consider the ordinary linear regression (2.1) under the conditions (i), (ii), and (vi). Then with $n \geq 3$, the least squares estimator

$$\hat{\beta}_{1,\text{ML}} = \frac{\mathrm{S}_{\xi y}}{\mathrm{S}_{\xi\xi}} \tag{2.138}$$

is unbiased, i.e.,

$$\mathbf{E}\hat{\beta}_{1,\text{ML}} = \beta_1 \,. \tag{2.139}$$

*The proof* follows from fragments of the proof of Theorem 2.23.

**Nonexistence of expectation of the adjusted least squares estimator**

**Theorem 2.26.** *Assume the conditions* (i), (ii), (vi), *and* (vii). *If, in addition, at $\beta_1 \neq 0$ and $n \geq 2$, then*

$$\mathbf{E}|\hat{\beta}_{1,\text{ALS}}| \;= \infty \,. \tag{2.140}$$

*Proof.* Suppose that

$$\mathbf{E}\left|\hat{\beta}_{1,\text{ALS}}\right| < \infty \,. \tag{2.141}$$

Then the next expectation is finite

$$\mathbf{E}\hat{\beta}_{1,\text{ALS}} = \mathbf{E}[\hat{\beta}_{1,\text{ALS}}|\vec{x},\vec{\xi}] \,, \quad \vec{\xi} = (\xi_1,\dots,\xi_n)^{\mathrm{T}} \,. \tag{2.142}$$

From formula (2.61), we have almost surely, that

$$\hat{\beta}_{1,\text{ALS}} = \beta_1 \frac{\mathrm{S}_{x\xi}}{\mathrm{S}_{xx} - \sigma_\delta^2} + \frac{\mathrm{S}_{x\varepsilon}}{\mathrm{S}_{xx} - \sigma_\delta^2} \,. \tag{2.143}$$

This implies that

$$\mathbf{E}[\hat{\beta}_{1,\text{ALS}}|\vec{x},\vec{\xi}] = \beta_1 \frac{\mathrm{S}_{x\xi}}{\mathrm{S}_{xx} - \sigma_\delta^2} + \frac{1}{\mathrm{S}_{xx} - \sigma_\delta^2}\mathbf{E}[\mathrm{S}_{x\varepsilon}\,|\vec{x}] \,. \tag{2.144}$$

As we saw in (2.135), the latter conditional expectation is zero. Then, as in (2.136), we have

$$\mathbf{E}[\hat{\beta}_{1,\text{ALS}}|\vec{x},\vec{\xi}] = \beta_1 \frac{\mathrm{S}_{x\xi}}{\mathrm{S}_{xx} - \sigma_\delta^2} \,, \tag{2.145}$$

$$\mathbf{E}\hat{\beta}_{1,\text{ALS}} = \beta_1 \,\mathbf{E}\frac{\mathrm{S}_{x\xi}}{\mathrm{S}_{xx} - \sigma_\delta^2} = \beta_1 \cdot K \cdot \mathbf{E}\frac{\mathrm{S}_{xx}}{\mathrm{S}_{xx} - \sigma_\delta^2} = \beta_1 \, K \left(1 + \sigma_\delta^2 \,\mathbf{E}\frac{1}{\mathrm{S}_{xx} - \sigma_\delta^2}\right) \,, \tag{2.146}$$

at that, all the expectations are finite. However, similar to (2.129) and (2.132), we have

$$\mathbf{E}\frac{1}{\left|S_{xx}-\sigma_\delta^2\right|} = \text{const}\cdot\int\limits_{\mathbf{R}^{n-1}}^{\infty}\frac{1}{\left|\frac{\sigma_x^2}{n}\,\|x\|^2-\sigma_\delta^2\right|}\mathrm{e}^{-\frac{\|x\|^2}{2}}\,dx = \\ = \text{const}\cdot\int\limits_{0}^{\infty}\frac{1}{\left|r^2-\frac{n\sigma_\delta^2}{\sigma_x^2}\right|}r^{n-2}\mathrm{e}^{-\frac{r^2}{2}}\,dr\,. \tag{2.147}$$

The latter improper integral has got a singularity at point $r_0 = \sqrt{\frac{n\sigma_\delta^2}{\sigma_x^2}}$. As $r \to r_0$, the integrand $f(r)$ has such a behavior:

$$f(r) \sim \text{const}\cdot\frac{1}{|r-r_0|}\,. \tag{2.148}$$

Here, the equivalence means that the ratio of the left-hand and right-hand sides tends to 1, as $r \to r_0$. Therefore, the integral (2.147) diverges simultaneously with the improper integral $\int_{r_0}^{r_0+1}\frac{dr}{r-r_0}$. Thus, expectation in (2.147) is infinite and the last expectation in (2.146) is not finite. The resulting contradiction shows that our assumption (2.141) is wrong. The theorem is proved. □

**Remark 2.27.** It can be shown that under the conditions of Theorem 2.26, expectation $\mathbf{E}\hat{\beta}_{1,\mathrm{ALS}}$ is not well-defined as Lebesgue integral, i.e.,

$$\mathbf{E}\hat{\beta}_{1,\mathrm{ALS}}^{+} = \mathbf{E}\hat{\beta}_{1,\mathrm{ALS}}^{-} = +\infty\,, \tag{2.149}$$

$$\hat{\beta}_{1,\mathrm{ALS}}^{+} = \max(\hat{\beta}_{1,\mathrm{ALS}},0)\,,\quad \hat{\beta}_{1,\mathrm{ALS}}^{-} = -\min(\hat{\beta}_{1,\mathrm{ALS}},0)\,. \tag{2.150}$$

Here, $\hat{\beta}_{1,\mathrm{ALS}}^{+}$ and $\hat{\beta}_{1,\mathrm{ALS}}^{-}$ are the positive and negative parts of the function $\hat{\beta}_{1,\mathrm{ALS}}(\omega)$, $\omega \in \Omega$ ($\omega$ is an elementary random event, and $\Omega$ is a total space of elementary events).

As we see, the naive estimator has the advantage over the consistent Adjusted Least Squares estimator that it has finite expectation. One can talk about the bias of the naive estimator, but it makes no sense to talk about the bias of the estimator $\hat{\beta}_{1,\mathrm{ALS}}$ has no sense taking into account Theorem 2.26 and Remark 2.27. Therein lies an important difference between measurement error models and ordinary regression models. In the first models, reasonable estimators do not have expectations and in the second ones they do have (see Remark 2.25).

Theorem 2.26 should be considered in the analysis of numerical simulation results. Suppose we have $N$ independent realizations of the model (2.1) and (2.2), i.e., have $N$ samples of $n$ observations each. For $k$th realization, we compute the estimate $\hat{\beta}_{1,\mathrm{ALS}}^{(k)}$, $k = \overline{1,N}$; these estimators are independent and identically distributed. If the averaged estimate $\frac{1}{N}\sum_{k=1}^{N}\hat{\beta}_{1,\mathrm{ALS}}^{(k)}$ is considered with regard to Remark 2.27, then that average behaves chaotically, as $N \to \infty$. At the same time, by Theorem 2.23 we have for the average of the naive estimator:

$$\frac{1}{N}\sum_{k=1}^{N}\hat{\beta}_{1,\mathrm{naive}}^{(k)} \xrightarrow{\mathbf{P}1} \mathbf{E}\hat{\beta}_{1,\mathrm{naive}} = K\beta_1\,,\quad \text{as } N\to\infty\,. \tag{2.151}$$

Thus, a table of dependency from $n$ of the values $\frac{1}{N}\sum_{k=1}^{N}\hat{\beta}_{1,\mathrm{naive}}^{(k)}$ may illustrate a bias of the naive estimator and the attenuation effect.

Instead for the ALS estimator, it is better to use a median for the set of the estimates $\{\hat{\beta}_{1,\mathrm{ALS}}^{(k)}, k=\overline{1,N}\}$. For this purpose, the estimates are arranged in increasing order:

$$\hat{\beta}_{1,\mathrm{ALS}}^{(k(1))} \le \hat{\beta}_{1,\mathrm{ALS}}^{(k(2))} \le \cdots \le \hat{\beta}_{1,\mathrm{ALS}}^{(k(N))}\,, \tag{2.152}$$

and we set

$$\operatorname{med}\hat{\beta}_{1,\mathrm{ALS}}^{(\bullet)} = \begin{cases} \hat{\beta}_{1,\mathrm{ALS}}^{(k(m))} & \text{if } N=2m-1\,, \\ \frac{\hat{\beta}_{1,\mathrm{ALS}}^{(k(m))}+\hat{\beta}_{1,\mathrm{ALS}}^{(k(m+1))}}{2} & \text{if } N=2m\,. \end{cases} \tag{2.153}$$

Under mild conditions (Beirlant et al., 2004), this value tends to the median of the estimator $\hat{\beta}_{1,\mathrm{ALS}}$ of a single realization:

$$\operatorname{med}\hat{\beta}_{1,\mathrm{ALS}}^{(\bullet)} \xrightarrow{\mathbf{P}1} \operatorname{med}\hat{\beta}_{1,\mathrm{ALS}}\,, \quad \text{as } N\to\infty\,. \tag{2.154}$$

The estimator $\hat{\beta}_{1,\mathrm{ALS}}$ is strongly consistent, so

$$\operatorname{med}\hat{\beta}_{1,\mathrm{ALS}} \xrightarrow{\mathbf{P}1} \beta_1\,, \quad \text{as } n\to\infty\,. \tag{2.155}$$

Relations (2.154) and (2.155) show that $\operatorname{med}\hat{\beta}_{1,\mathrm{ALS}}^{(\bullet)}$ approaches to the true value $\beta_1$, when $N$ and $n$ are becoming large enough. Thus, when $N$ is taken large enough, a table of dependence of $\operatorname{med}\hat{\beta}_{1,\mathrm{ALS}}^{(\bullet)}$ on $n$ can illustrate the strong consistency of the ALS estimator.

In addition to the bias, another common characteristic of the accuracy of an estimator $\hat{\theta}$ is the mean squared error $\mathbf{E}_\theta(\hat{\theta}-\theta)^2$. However, for the ALS estimator one cannot use the empirical mean squared error $\frac{1}{N}\sum_{k=1}^{N}(\hat{\beta}_{1,\mathrm{ALS}}^{(\mathrm{k})}-\beta_1)^2$, because by Theorem 2.26,

$$\frac{1}{N}\sum_{k=1}^{N}(\hat{\beta}_{1,\mathrm{ALS}}^{(k)}-\beta_1)^2 \xrightarrow{\mathbf{P}1} \mathbf{E}(\hat{\beta}_{1,\mathrm{ALS}}-\beta_1)^2 = +\infty\,, \quad \text{as } N\to\infty\,. \tag{2.156}$$

Instead, it is quite possible to use the empirical median $\operatorname{med}(\hat{\beta}_{1,\mathrm{ALS}}^{(\bullet)}-\beta_1)^2$, which tends to $\operatorname{med}(\hat{\beta}_{1,\mathrm{ALS}}-\beta_1)^2$, almost surely, as $N\to\infty$. The latter tends to 0, as $n\to\infty$, because of the strong consistency of the estimator. For this reason, a table of dependence of $\operatorname{med}(\hat{\beta}_{1,\mathrm{ALS}}^{(\bullet)}-\beta_1)^2$ on $n$ illustrates the strong consistency of the ALS estimator as well. It is better even to apply the empirical median of absolute deviation,

$$\operatorname{med}|\hat{\beta}_{1,\mathrm{ALS}}^{(\bullet)}-\beta_1| = \sqrt{\operatorname{med}(\hat{\beta}_{1,\mathrm{ALS}}^{(\bullet)}-\beta_1)^2}\,, \tag{2.157}$$

because it has the same physical dimension as $\beta_1$.

### 2.4.6 The adjusted least squares estimator in the vector model

**The vector linear model**

Generalize the model (2.1) and (2.2) for the vector case. Suppose, we have the observations

$$y_i = B^{\mathrm{T}}\xi_i + \varepsilon_i\,, \tag{2.158}$$

$$x_i = \xi_i + \delta_i\,, \quad i = \overline{1,n}\,. \tag{2.159}$$

Here, $\xi_i$ is unobservable random vector in $\mathbf{R}^d$, $y_i$ is observed response in $\mathbf{R}^m$, $\varepsilon_i$ is a vector of observation errors of the response, and $\delta_i$ is a vector of measurement errors in the covariate. The regression coefficients matrix $B$ of size $d\times m$ should be estimated.

The model (2.158) and (2.159) is called the *vector* (structural) measurement error model. In particular case when the response is scalar ($m = 1$), we call this model as *multiple* linear model, meaning that a collection of several scalar regressors is considered, on which the response depends linearly. In multiple regression models, the matrix $B$ becomes a column vector and then $B^{\mathrm{T}}\xi_i$ is just an inner product of two column vectors.

Another particular case of the model (2.158) and (2.159) is the vector model with intercept:

$$y_i = y_0 + C^{\mathrm{T}}\psi_i + \varepsilon_i\,, \tag{2.160}$$

$$w_i = \psi_i + \upsilon_i\,, \quad i = \overline{1,n}\,. \tag{2.161}$$

Here, $\psi_i$ is a random vector in $\mathbf{R}^{d-1}$ ($d \geq 2$), $y_i$ is an observed response in $\mathbf{R}^m$, $\varepsilon_i$ is a vector of errors in the response, $\upsilon_i$ is a vector of measurement errors in the covariates. The intercept $y_0 \in \mathbf{R}^m$ and the regression coefficients matrix $C$ of size $(d-1)\times m$ have to be estimated. This model is reduced to the model (2.158) and (2.159) if we put

$$\xi_i = \begin{pmatrix} 1 \\ \psi_i \end{pmatrix}\,, \quad B = \begin{bmatrix} y_0^{\mathrm{T}} \\ C \end{bmatrix}\,, \quad \delta_i = \begin{pmatrix} 0 \\ \upsilon_i \end{pmatrix}\,, \quad x_i = \begin{pmatrix} 1 \\ w_i \end{pmatrix}\,. \tag{2.162}$$

The model (2.158) and (2.159) can be written in a matrix form. Introduce the matrices

$$Y = \begin{bmatrix} y_1^{\mathrm{T}} \\ \vdots \\ y_n^{\mathrm{T}} \end{bmatrix}\,, \quad X = \begin{bmatrix} x_1^{\mathrm{T}} \\ \vdots \\ x_n^{\mathrm{T}} \end{bmatrix}\,, \quad X^{\mathrm{tr}} = \begin{bmatrix} \xi_1^{\mathrm{T}} \\ \vdots \\ \xi_n^{\mathrm{T}} \end{bmatrix}\,, \tag{2.163}$$

$$Y^{\mathrm{e}} = \begin{bmatrix} \varepsilon_1^{\mathrm{T}} \\ \vdots \\ \varepsilon_n^{\mathrm{T}} \end{bmatrix}\,, \quad X^{\mathrm{e}} = \begin{bmatrix} \delta_1^{\mathrm{T}} \\ \vdots \\ \delta_n^{\mathrm{T}} \end{bmatrix}\,, \quad Y^{\mathrm{tr}} = X^{\mathrm{tr}}B\,. \tag{2.164}$$

Since the vector model can be rewritten as

$$y_i^{\mathrm{T}} = \xi_i^{\mathrm{T}}B + \varepsilon_i^{\mathrm{T}}\,, \quad x_i^{\mathrm{T}} = \xi_i^{\mathrm{T}} + \delta_i^{\mathrm{T}}\,, \quad i = \overline{1,n}\,, \tag{2.165}$$

it is equivalent to the matrix equations:

$$Y = Y^{\mathrm{tr}} + Y^{\mathrm{e}}\,, \quad X = X^{\mathrm{tr}} + X^{\mathrm{e}}\,, \quad Y^{\mathrm{tr}} = X^{\mathrm{tr}}B\,. \tag{2.166}$$

Here, $X$ is the observed input matrix, $Y$ is the observed output matrix, $X^{\mathrm{tr}}$ and $Y^{\mathrm{tr}}$ are the corresponding true matrices, and $X^{\mathrm{e}}$ and $Y^{\mathrm{e}}$ are the corresponding error matrices. The matrix model (2.166) is also recorded by means of the approximate equality

$$\underset{n\times d}{X}\ \underset{d\times m}{B} \approx \underset{n\times m}{Y}\,. \tag{2.167}$$

The given matrices $X$ and $Y$ contain additive measurement errors and the unknown matrix $B$ is estimated. The latter has a fixed size and does not change with increasing number $n$ of observed rows of the matrices $X$ and $Y$. The model (2.167) can be interpreted as overdetermined system of linear equations. Namely, we have $dm$ unknown entries of $B$ and $n$ vector equations $x_i^{\mathrm{T}}B \approx y_i^{\mathrm{T}}$, $i = \overline{1,n}$; if $n > dm$ then there are more equations than unknowns. Problem of finding $B$ from this system touches upon the computational linear algebra while the estimation of $B$ under the observation model (2.158) and (2.159) relates to the theory of linear regression. As evident, from a mathematical point of view these two problems are equivalent, if the uncertainties in the observable matrices $X$ and $Y$ are modeled by means of additive random errors.

**The adjusted least squares estimator**

The following assumptions about the model (2.158) and (2.159) are common.

(a) Random vectors $\{\xi_i, \varepsilon_i, \delta_i, i \geq 1\}$ are independent.

(b) The errors $\varepsilon_i$ are identically distributed in $\mathbf{R}^m$ and centered, with finite second moments.

(c) The errors $\delta_i$ are identically distributed in $\mathbf{R}^d$ and centered, with finite and known variance–covariance matrix

$$V_\delta = \mathbf{E}\delta_1\delta_1^{\mathrm{T}}\,. \tag{2.168}$$

(d) Random vectors $\xi_i$ are identically distributed in $\mathbf{R}^d$, with finite positive definite (unknown) correlation matrix

$$M_\xi = \mathbf{E}\xi_1\xi_1^{\mathrm{T}}\,. \tag{2.169}$$

Hereafter, the inequality $A > B$ for symmetric matrices means that the matrix $A - B$ is positive definite (this partial order is called *Loewner order*). In particular, notation $A > 0$ indicates that the matrix $A$ is positive definite.

To construct a consistent estimator of the matrix $B$ under conditions (a)–(d), we apply the Corrected Score method (see Section 2.4.2).

In the case $\delta_i \equiv 0$, the elementary objective function of the least squares method is

$$q_{\mathrm{LS}}(y, \xi; B) = \left\|y - B^{\mathrm{T}}\xi\right\|^2\,, \quad B \in \mathbf{R}^{d\times m}\,. \tag{2.170}$$

The estimating function of the least squares method is equal to

$$s_{\mathrm{LS}}(y, \xi; B) = \frac{1}{2} \frac{\partial q_{\mathrm{LS}}}{\partial B} . \tag{2.171}$$

Here, $\partial q_{\mathrm{LS}}/\partial B$ is a linear functional in the space of matrices $\mathbf{R}^{d\times m}$ (see the technique of matrix derivatives in the book by Cartan, 1970). In this space, linear functionals have a representation

$$f_V(U) = \operatorname{trace}(U^{\mathrm{T}} V) , \quad U \in \mathbf{R}^{d\times m} , \quad V \in \mathbf{R}^{d\times m} , \tag{2.172}$$

where $\operatorname{trace} A = \sum_i a_{ii}$ denotes the trace of a square matrix. The functional $f_V$ can be identified with the matrix $V$ that represents it. Then for arbitrary matrix $H \in \mathbf{R}^{d\times m}$,

$$\begin{aligned} \frac{1}{2} \frac{\partial q_{\mathrm{LS}}}{\partial B}(H) &= (B^{\mathrm{T}}\xi - y, H^{\mathrm{T}}\xi) = \\ &= \operatorname{trace}(H^{\mathrm{T}}\xi(B^{\mathrm{T}}\xi - y)^{\mathrm{T}}) = \operatorname{trace}(H^{\mathrm{T}}(\xi\xi^{\mathrm{T}}B - \xi y^{\mathrm{T}})) . \end{aligned} \tag{2.173}$$

Therefore, the linear functional (2.171) can be identified with the matrix $\xi\xi^{\mathrm{T}}B - \xi y^{\mathrm{T}}$. Next, we construct the corrected estimating function $s_{\mathrm{c}}$ as a solution to the deconvolution equation

$$\mathbf{E}[s_{\mathrm{c}}(y, x; B)|y, \xi] = \xi\xi^{\mathrm{T}}B - \xi y^{\mathrm{T}} , \quad B \in \mathbf{R}^{d\times m} . \tag{2.174}$$

The solution in the class of matrix-valued functions to be polynomial in $\xi$ looks like

$$s_{\mathrm{c}}(y, x; B) = (xx^{\mathrm{T}} - V_\delta)B - xy^{\mathrm{T}} . \tag{2.175}$$

Indeed, for the function (2.175), with $x = \xi + \delta$, we have

$$\begin{aligned} \mathbf{E}[s_{\mathrm{c}}|y, \xi] &= (\mathbf{E}[xx^{\mathrm{T}}|\xi] - V_\delta)B - \mathbf{E}[x|\xi] \cdot y^{\mathrm{T}} = \\ &= (\xi\xi^{\mathrm{T}} + \mathbf{E}\delta\delta^{\mathrm{T}} - V_\delta)B - \xi y^{\mathrm{T}} = \xi\xi^{\mathrm{T}}B - \xi y^{\mathrm{T}} , \quad B \in \mathbf{R}^{d\times m} . \end{aligned} \tag{2.176}$$

Then, according to the CS method, the ALS estimator is a measurable solution to the equation

$$\left[ \frac{1}{n} \sum_{i=1}^{n} (x_i x_i^{\mathrm{T}} - V_d) \right] B - \frac{1}{n} \sum_{i=1}^{n} x_i y_i^{\mathrm{T}} = 0 , \quad B \in \mathbf{R}^{d\times m} . \tag{2.177}$$

By the SLLN and in accordance with (2.169), we have

$$\frac{1}{n} \sum_{i=1}^{n} (x_i x_i^{\mathrm{T}} - V_\delta) \xrightarrow{\mathbf{P}1} \mathbf{E}xx^{\mathrm{T}} - V_\delta = \mathbf{E}\xi\xi^{\mathrm{T}} > 0 , \quad \text{as } n \to \infty . \tag{2.178}$$

So *eventually* the matrix $\frac{1}{n}\sum_{i=1}^{n} (x_i x_i^{\mathrm{T}} - V_\delta)$ is nonsingular, and the formula for the ALS estimator is valid *eventually*

$$\hat{B}_{\mathrm{ALS}} = \left( \sum_{i=1}^{n} x_i x_i^{\mathrm{T}} - nV_\delta \right)^{-1} \sum_{i=1}^{n} x_i y_i^{\mathrm{T}} . \tag{2.179}$$

In terms of matrices (2.163), we have the compact formula

$$\hat{B}_{\mathrm{ALS}} = (XX^{\mathrm{T}} - nV_\delta)^{-1} X^{\mathrm{T}} Y \,. \tag{2.180}$$

This formula is correct *eventually*, and for any sample size, the estimator can be defined as follows:

$$\hat{B}_{\mathrm{ALS}} = (XX^{\mathrm{T}} - nV_\delta)^{+} X^{\mathrm{T}} Y \,, \tag{2.181}$$

where $A^+$ is the Moore–Penrose pseudoinverse (Seber and Lee, 2003). If a square matrix $A$ is nonsingular then $A^+ = A^{-1}$ and the formula (2.181) is converted to (2.180).

**Theorem 2.28.** *Assume the conditions* (a)–(d). *Then* $\hat{B}_{\mathrm{ALS}}$ *is a strongly consistent estimator, i.e.*

$$\hat{B}_{\mathrm{ALS}} \xrightarrow{\mathbf{P}1} B \,, \quad as \;\; n \to \infty \,. \tag{2.182}$$

*Proof.* By the SLLN, as $n \to \infty$, we have

$$\frac{1}{n} \sum_{i=1}^{n} x_i y_i^{\mathrm{T}} \xrightarrow{\mathbf{P}1} \mathbf{E}(xy^{\mathrm{T}}) = \mathbf{E}(\xi\xi^{\mathrm{T}} B) = \mathbf{E}(\xi\xi^{\mathrm{T}}) B. \tag{2.183}$$

Then the formulas (2.179), (2.178), and (2.183) give *eventually*

$$\hat{B}_{\mathrm{ALS}} = \left( \frac{1}{n} \sum_{i=1}^{n} x_i x_i^{\mathrm{T}} - V_\delta \right)^{-1} \cdot \frac{1}{n} \sum_{i=1}^{n} x_i y_i^{\mathrm{T}} \xrightarrow{\mathbf{P}1} (\mathbf{E}\xi\xi^{\mathrm{T}})^{-1} (\mathbf{E}\xi\xi^{\mathrm{T}}) B = B \,. \tag{2.184}$$

The theorem is proved. □

**Parameter estimation in the vector model with intercept**

In the model (2.160) and (2.161), the formula for the estimator (2.179) can be specified because the matrix $B$ contains $y_0$ and $C$, see (2.162).

Introduce the appropriate conditions.

(a1) The random vectors $\psi_i, \varepsilon_i, v_i$, $i \geq 1$, are independent.

(c1) The errors $v_i$ are identically distributed in $\mathbf{R}^{d-1} (d \geq 2)$ and centered, with finite and known correlation matrix

$$V_v = \mathbf{E} v_1 v_1^{\mathrm{T}} \,. \tag{2.185}$$

(d1) The random vectors $\psi_i$ are identically distributed in $\mathbf{R}^{d-1}$, with finite (unknown) covariance matrix

$$S_\psi = \mathrm{cov}(\psi_1) = \mathbf{E}(\psi_1 - \mu_\psi)(\psi_1^{\mathrm{T}} - \mu_\psi^{\mathrm{T}}) \,, \quad \mu_\psi = \mathbf{E}\psi_1 \,; \tag{2.186}$$

$$S_\psi > 0 \,. \tag{2.187}$$

Remember that inequality (2.187) means positive definiteness of the matrix $S_\psi$.

**Theorem 2.29.** *Assume the conditions* (a1), (b), (c1), *and* (d1). *Then in the model* (2.160) *and* (2.161), *we have* eventually

$$\hat{C}_{\mathrm{ALS}} = (\mathrm{S}_{ww} - V_v)^{-1}\,\mathrm{S}_{wy}\,, \tag{2.188}$$

$$\hat{y}_{0,\mathrm{ALS}} = \bar{y} - \hat{C}_{\mathrm{ALS}}\bar{w}\,; \tag{2.189}$$

$$\begin{aligned} \textit{where}\quad & \mathrm{S}_{ww} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n}(w_i - \bar{w})(w_i - \bar{w})^{\mathrm{T}}\,, \\ & \mathrm{S}_{wy} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n}(w_i - \bar{w})(y_i - \bar{y})^{\mathrm{T}} \end{aligned} \tag{2.190}$$

*are the sample covariances;*

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i\,, \quad \bar{w} = \frac{1}{n}\sum_{i=1}^{n} w_i \tag{2.191}$$

*are the sample means. In addition, the estimators* (2.188) *and* (2.189) *are strongly consistent, i.e., as* $n \to \infty$, *it holds that*

$$\hat{C}_{\mathrm{ALS}} \xrightarrow{\mathbf{P}1} C\,, \quad \hat{y}_{0,\mathrm{ALS}} \xrightarrow{\mathbf{P}1} y_0\,. \tag{2.192}$$

*Proof.* According to formulas (2.162), the model (2.160) and (2.161) is reduced to the vector model (2.158) and (2.159). Moreover, the conditions (a)–(c) hold true obviously. It remains to verify the inequality (2.169) in condition (d).

We have $\xi = \binom{1}{\psi}$, and therefore, the second moments of $\xi$ exist. The required inequality

$$M_\xi = \mathbf{E}\xi\xi^{\mathrm{T}} > 0 \tag{2.193}$$

is equivalent to the linear independence of random variables 1, $\psi^{(1)}, \ldots, \psi^{(d-1)}$, where $\psi^{(k)}$ are components of the random vector $\psi$. Let $a_0, \ldots, a_{d-1}$ be real numbers such that

$$a_0 + \sum_{k=1}^{d-1} a_k \psi^{(k)} = 0\,, \quad \text{a.s.} \tag{2.194}$$

Then $0 = \mathbf{D}\sum_{k=1}^{d-1} a_k\psi^{(k)} = a^{\mathrm{T}}S_\psi a$, $a = (a_1, \ldots, a_{d-1})^{\mathrm{T}}$. Since by the condition (d1) it holds $S_\psi > 0$, then from here $a = 0$ and further from (2.194) we get $a_0 = 0$. Thus, linear independence of the random variables 1, $\psi^{(1)}, \ldots, \psi^{(d-1)}$ is proved, and therefore, the inequality (2.193) is justified. Thus, the condition (d) follows from the condition (d1).

As we can see, the conditions of Theorem 2.29 imply the conditions of the previous theorem. Hence, the convergence (2.192) is obtained.

To prove the formula for the estimators, transform the estimating equation (2.177). In view of (2.162), we obtain a couple of estimating equations (cf. with (2.70)):

$$\bar{y} = y_0 + C^{\mathrm{T}}\bar{w}\,, \tag{2.195}$$

$$\bar{w}y_0^{\mathrm{T}} + \left(\overline{ww^{\mathrm{T}}} - V_v\right) C = \overline{wy^{\mathrm{T}}}\,. \tag{2.196}$$

Eliminating $y_0$ from here, we get the estimating equation for $\hat{C}_{\mathrm{ALS}}$:

$$(\mathrm{S}_{ww} - V_v)\,C = \mathrm{S}_{wy}\,. \tag{2.197}$$

As $n \to \infty$,

$$S_{ww} - V_v \xrightarrow{\mathbf{P1}} \operatorname{cov}(w) - V_v = S_\psi > 0\,, \tag{2.198}$$

hence, it holds *eventually* that $S_{ww} - V_v > 0$. Then (2.197) implies equality (2.188), and from equation (2.195) equality (2.189) follows. The theorem is proved. □

Note that the formula for the estimators is a direct generalization of the scalar formulas (2.72) and (2.73).

It is necessary to mention the papers where the results of Section 2.4.6 are expanded. In Sen'ko (2013), more general conditions for the consistency of $\hat{B}_{\mathrm{ALS}}$ are given, and Sen'ko (2014) proves the asymptotic normality of the estimator and constructs its small sample modification. The modification is computationally more stable than $\hat{B}_{\mathrm{ALS}}$ for small and moderate sample, and it has the same ACM as $\hat{B}_{\mathrm{ALS}}$. Finally, in Cheng and Kukush (2006) it is shown that in the vector model with intercept, it holds that $\mathbf{E}\|\hat{C}_{\mathrm{ALS}}\| = \infty$, i.e., a generalization of Theorem 2.26 holds true.

## 2.5 The model with known ratio of error variances

Consider the normal linear model (2.1) and (2.2) under the conditions (i) and (vi)–(viii) (the conditions were given at the beginning of Chapter 2 and in Sections 2.2 and 2.3). Currently, we assume that $\sigma_\delta^2$ is unknown but the ratio is given

$$\lambda = \frac{\sigma_\varepsilon^2}{\sigma_\delta^2}\,. \tag{2.199}$$

This permits to overcome the nonidentifiability of the model (see Theorem 2.10). Since $\sigma_\varepsilon^2 = \lambda\sigma_\delta^2$, then five parameters, namely $\beta_0$, $\beta_1$, $\sigma_\delta^2$, $\mu_\xi$, and $\sigma_\xi^2$ should be estimated.

### 2.5.1 The MLE and its consistency

According to the method described in Section 2.4.3, the system of equations for the MLE is as follows:

$$\bar{x} = \hat{\mu}_\xi\,, \quad \bar{y} = \hat{\beta}_0 + \hat{\beta}_1\hat{\mu}_\xi\,, \tag{2.200}$$

$$\begin{aligned} S_{xx} &= \hat{\sigma}_\xi^2 + \hat{\sigma}_\delta^2\,, \\ S_{xy} &= \hat{\beta}_1\hat{\sigma}_\xi^2\,, \\ S_{yy} &= \hat{\beta}_1^2\hat{\sigma}_\xi^2 + \lambda\hat{\sigma}_\delta^2\,, \end{aligned} \tag{2.201}$$

$$\hat{\sigma}_\delta^2 > 0\,, \quad \hat{\sigma}_\xi^2 > 0\,. \tag{2.202}$$

From here

$$\hat{\mu}_\xi = \bar{x}\,. \tag{2.203}$$

From equations (2.201), we exclude $\hat{\sigma}_\xi^2$:

$$\hat{\beta}_1 \, \mathrm{S}_{xx} = \mathrm{S}_{xy} + \hat{\beta}_1 \hat{\sigma}_\delta^2 \,, \quad \mathrm{S}_{yy} = \hat{\beta}_1 \, \mathrm{S}_{xy} + \lambda \hat{\sigma}_\delta^2 \,. \tag{2.204}$$

Eliminating $\hat{\sigma}_\delta^2$ we get a quadratic equation in $\hat{\beta}_1$

$$\hat{\beta}_1^2 \, \mathrm{S}_{xy} + \hat{\beta}_1 (\lambda \, \mathrm{S}_{xx} - \mathrm{S}_{yy}) - \lambda \, \mathrm{S}_{xy} = 0 \,. \tag{2.205}$$

In the normal linear model, $\mathrm{S}_{xy} \neq 0$ when $n \geq 2$, a.s. Hence the quadratic equation does not degenerate to a linear one. From here, it holds almost surely, that

$$\hat{\beta}_1 = \frac{\mathrm{S}_{yy} - \lambda \, \mathrm{S}_{xx} \pm \sqrt{(\mathrm{S}_{yy} - \lambda \, \mathrm{S}_{xx})^2 + 4\lambda \, \mathrm{S}_{xy}^2}}{2 \, \mathrm{S}_{xy}} = \frac{U}{2 \, \mathrm{S}_{xy}} \,. \tag{2.206}$$

Further, from the second equation in (2.201),

$$\hat{\sigma}_\xi^2 = \frac{\mathrm{S}_{xy}}{\hat{\beta}_1} = \frac{2 \, \mathrm{S}_{xy}^2}{U} > 0 \,, \tag{2.207}$$

hence $U > 0$, a.s., and in (2.206) for $\hat{\beta}_1$ one needs to take "plus" before the root. Thus,

$$\hat{\beta}_1 = \frac{\mathrm{S}_{yy} - \lambda \, \mathrm{S}_{xx} + \sqrt{(\mathrm{S}_{yy} - \lambda \, \mathrm{S}_{xx})^2 + 4\lambda \, \mathrm{S}_{xy}^2}}{2 \, \mathrm{S}_{xy}} \,. \tag{2.208}$$

From the second equation (2.204), we have

$$\hat{\sigma}_\delta^2 = \frac{1}{\lambda} (\mathrm{S}_{yy} - \hat{\beta}_1 \, \mathrm{S}_{xy}) \,, \tag{2.209}$$

and from (2.200):

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \,. \tag{2.210}$$

To ensure that the estimators (2.203) and (2.207)–(2.210) determine a.s. the solution to the system (2.200)–(2.202), one must also check the following:

$$\mathrm{S}_{yy} - \hat{\beta}_1 \, \mathrm{S}_{xy} > 0 \,, \quad \text{a.s.} \tag{2.211}$$

We have almost surely:

$$\begin{aligned} 2\hat{\beta}_1 \, \mathrm{S}_{xy} = U = \mathrm{S}_{yy} - \lambda \, \mathrm{S}_{xx} + \sqrt{(\mathrm{S}_{yy} - \lambda \, \mathrm{S}_{xx})^2 + 4\lambda \, \mathrm{S}_{xy}^2} < \\ < \mathrm{S}_{yy} - \lambda \, \mathrm{S}_{xx} + \sqrt{(\mathrm{S}_{yy} - \lambda \, \mathrm{S}_{xx})^2 + 4\lambda \, \mathrm{S}_{xx} \, \mathrm{S}_{yy}} = \\ = \mathrm{S}_{yy} - \lambda \, \mathrm{S}_{xx} + \mathrm{S}_{yy} + \lambda \, \mathrm{S}_{xx} = 2 \, \mathrm{S}_{yy} \,. \end{aligned} \tag{2.212}$$

Hence, (2.211) holds true, and so do inequalities (2.202) for our solutions, a.s. We have proved the following statement.

**Theorem 2.30.** *Assume the conditions* (i), (vi)–(viii) *and let the ratio* (2.199) *be known in the linear model* (2.1) *and* (2.2). *Then the MLEs are given by the equalities* (2.203) *and* (2.207)–(2.210), *almost surely.*

As in Section 2.4, these estimators are strongly consistent without the normality assumption.

**Theorem 2.31.** *Assume the conditions* (i)–(iv) *and let the ratio* (2.199) *be known. Then the estimators* (2.203) *and* (2.207)–(2.210) *are strongly consistent.*

*Proof.* We verify only the consistency of the estimator (2.208); then the consistency for the estimators of the other parameters can be verified without any problem.

(a) *Case* $\beta_1 \neq 0$. We have

$$\begin{aligned} &\mathrm{S}_{xx} \xrightarrow{\mathbf{P1}} \sigma_\xi^2 + \sigma_\delta^2 \,, \\ &\mathrm{S}_{xy} \xrightarrow{\mathbf{P1}} \operatorname{cov}(x, y) = \beta_1 \sigma_\xi^2 \,, \\ &\mathrm{S}_{yy} \xrightarrow{\mathbf{P1}} \beta_1^2 \sigma_\xi^2 + \lambda \sigma_\delta^2 \,. \end{aligned} \tag{2.213}$$

Then *eventually* $\mathrm{S}_{xy} \neq 0$;

$$\begin{aligned} \hat\beta_1 \xrightarrow{\mathbf{P1}} \frac{(\beta_1^2 - \lambda)\,\sigma_\xi^2 + \sqrt{(\beta_1^2 - \lambda)^2 \sigma_\xi^4 + 4\beta_1^2 \sigma_\xi^4}}{2\beta_1 \sigma_\xi^2} = \\ = \frac{(\beta_1^2 - \lambda)\,\sigma_\xi^2 + (\beta_1^2 + \lambda)\,\sigma_\xi^2}{2\beta_1 \sigma_\xi^2} = \beta_1 \,. \end{aligned} \tag{2.214}$$

(b) *Case* $\beta_1 = 0$. It is convenient to convert the formula (2.208) by removing irrationality in the numerator (we assume now $\mathrm{S}_{xy} \neq 0$; if not, then $\hat\beta_1 = 0$):

$$\hat\beta_1 = -\frac{2\lambda\,\mathrm{S}_{xy}}{\mathrm{S}_{yy} - \lambda\,\mathrm{S}_{xx} - \sqrt{(\mathrm{S}_{yy} - \lambda\,\mathrm{S}_{xx})^2 + 4\lambda\,\mathrm{S}_{xy}^2}} \,. \tag{2.215}$$

In view of (2.196), while $\beta_1 = 0$, it holds that

$$\mathrm{S}_{xy} \xrightarrow{\mathbf{P1}} 0 \,, \quad \mathrm{S}_{yy} - \lambda\,\mathrm{S}_{xx} \xrightarrow{\mathbf{P1}} -\lambda\sigma_\xi^2 \,; \tag{2.216}$$

$$\hat\beta_1 \xrightarrow{\mathbf{P1}} -\frac{0}{-2\lambda\sigma_\xi^2} = 0 \,. \tag{2.217}$$

The strong consistency of the estimator $\hat\beta_1$ is proved in all the cases. □

Mention the following. Under the conditions of Theorem 2.30, the estimators (2.203) and (2.207)–(2.210) are specified by these expressions *eventually*. In particular, this means that the expression (2.209) is positive, *eventually*. For real data, it can happen that $\mathrm{S}_{yy} - \hat\beta_1\,\mathrm{S}_{xy} = 0$, and then we set $\hat\sigma_\delta^2 = 0$.

### 2.5.2 Asymptotic normality of the slope estimator

Reasoning as in Section 2.4.4, one can show that the augmented estimator $(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_\delta^2, \hat{\mu}_\xi, \hat{\sigma}_\xi^2)^{\mathrm{T}}$ taken from Section 2.5.1 is asymptotically normal. Here, we derive the asymptotic normality only for $\hat{\beta}_1$ using its explicit formula. We require the normality of $\varepsilon$ and $\delta$, but now the normality of $\xi$ is not mandatory. We exploit the stochastic calculus introduced in Section 2.4.4.

**Theorem 2.32.** *Assume the conditions* (i), (iv), (vi), *and* (vii). *Then for the estimator* (2.208), *it holds true that*

$$\sqrt{n}\,(\hat{\beta}_1 - \beta_1) \xrightarrow{d} \mathrm{N}(0, \sigma_{1,\lambda}^2)\,, \tag{2.218}$$

$$\sigma_{1,\lambda}^2 = \frac{1}{\sigma_\xi^4}\,(\sigma_\varepsilon^2\sigma_x^2 + \beta_1^2\sigma_\xi^2\sigma_\delta^2)\,. \tag{2.219}$$

*Proof.* We have

$$\hat{\beta}_1 - \beta_1 = \frac{\mathrm{S}_{yy} - \lambda\,\mathrm{S}_{xx} - 2\beta_1\,\mathrm{S}_{xy} + \sqrt{(\mathrm{S}_{yy} - \lambda\,\mathrm{S}_{xx})^2 + 4\lambda\,\mathrm{S}_{xy}^2}}{2\,\mathrm{S}_{xy}}\,. \tag{2.220}$$

We confine ourselves to the case $\beta_1 \neq 0$ (if $\beta_1 = 0$, then irrationality in (2.220) has to be moved in the denominator).

Further, in the proof we write down $u_n \approx v_n$ for sequences $\{u_n, n \geq 1\}$ and $\{v_n, n \geq 1\}$ of random variables, if

$$u_n - v_n = \frac{O_p(1)}{n}\,. \tag{2.221}$$

Then from the condition (2.221) it follows that

$$\sqrt{n}\,(u_n - v_n) = \frac{O_p(1)}{\sqrt{n}} \xrightarrow{\mathbf{P}} 0\,, \quad n \to \infty\,. \tag{2.222}$$

If exploiting the approximate equality, one converts the right-hand side of (2.220), then the terms, being neglected by us, will not affect the convergence in distribution after normalization by factor $\sqrt{n}$ (see Slutsky's Lemma 2.18).

Then in view of (2.96), (2.97), (2.105), and (2.106) we have, with $\rho = \xi - \mu_\xi$:

$$\mathrm{S}_{xy} \approx \beta_1\sigma_\xi^2 + \left[\beta_1\,(\overline{\rho^2} - \sigma_\xi^2) + \beta_1\overline{\delta\rho} + \overline{\rho\varepsilon} + \overline{\delta\varepsilon}\right] = \beta_1\sigma_\xi^2 + r_{xy}\,, \tag{2.223}$$

$$\mathrm{S}_{yy} = \beta_1^2\,\mathrm{S}_{\rho\rho} + 2\beta_1\,\mathrm{S}_{\rho\varepsilon} + \mathrm{S}_{\varepsilon\varepsilon} \approx \beta_1^2\sigma_\xi^2 + \sigma_\varepsilon^2 + r_{yy}\,, \tag{2.224}$$

$$r_{yy} = \beta_1^2(\overline{\rho^2} - \sigma_\xi^2) + 2\beta_1\overline{\rho\varepsilon} + (\overline{\varepsilon^2} - \sigma_\varepsilon^2)\,, \tag{2.225}$$

$$\begin{aligned} \mathrm{S}_{xx} &\approx \sigma_x^2 + r_{xx} = \sigma_\xi^2 + \sigma_\delta^2 + r_{xx}\,, \\ r_{xx} &= (\overline{\rho^2} - \sigma_\xi^2) + 2\overline{\rho\delta} + (\overline{\delta^2} - \sigma_\delta^2)\,. \end{aligned} \tag{2.226}$$

All the residual terms above have the order $O_p(1)/\sqrt{n}$. From here, we get

$$\mathrm{S}_{yy} - \lambda\,\mathrm{S}_{xx} \approx (\beta_1^2 - \lambda)\,\sigma_\xi^2 + r_{yy} - \lambda r_{xx}\,, \tag{2.227}$$

$$\mathrm{S}_{yy} - \lambda\,\mathrm{S}_{xx} - 2\beta_1\,\mathrm{S}_{xy} \approx -(\beta_1^2 + \lambda)\,\sigma_\xi^2 + r_{yy} - \lambda r_{xx} - 2\beta_1 r_{xy}\,, \tag{2.228}$$

$$(\mathrm{S}_{yy} - \lambda\,\mathrm{S}_{xx})^2 + 4\lambda\,\mathrm{S}_{xy}^2 \approx (\beta_1^2 + \lambda)^2\sigma_\xi^4 + 2\sigma_\xi^2 + 2\sigma_\xi^2\{(\beta_1^2 - \lambda)(r_{yy} - \lambda r_{xx}) + 4\beta_1\lambda r_{xy}\}\,. \tag{2.229}$$

To convert the root, the following expansion is applied, with $A > 0$:

$$\sqrt{A^2 + t} = A + \frac{t}{2A} + O(t^2)\,, \quad \text{as } t \to 0\,. \tag{2.230}$$

Therefore, putting $A = (\beta_1^2 + \lambda)\,\sigma_\xi^2$ we get

$$\sqrt{(\mathrm{S}_{yy} - \lambda\,\mathrm{S}_{xx})^2 + 4\lambda\,\mathrm{S}_{xy}^2} \approx \\ \approx (\beta_1^2 + \lambda)\,\sigma_\xi^2 + \frac{\sigma_\xi^2\{(\beta_1^2 - \lambda)\,(r_{yy} - \lambda r_{xx}) + 4\beta_1\lambda r_{xy}\}}{(\beta_1^2 + \lambda)\,\sigma_\xi^2}\,. \tag{2.231}$$

Substitute (2.223), (2.228), and (2.231) in equation (2.220):

$$\hat\beta_1 - \beta_1 \approx \frac{\beta_1(r_{yy} - \lambda r_{xx}) + (\lambda - \beta_1^2)\,r_{xy}}{\sigma_\xi^2(\beta_1^2 + \lambda)} = \frac{\Lambda_n}{\sigma_\xi^2\,(\beta_1^2 + \lambda)}\,. \tag{2.232}$$

From the expansions of the residual terms from (2.223), (2.225), and (2.226), we have

$$\Lambda_n = (\beta_1^2 + \lambda)\,\overline{\rho\varepsilon} + \beta_1(\overline{\varepsilon^2} - \sigma_\varepsilon^2) - \beta_1(\beta_1^2 + \lambda)\overline{\rho\delta} - \beta_1\lambda\,(\overline{\delta^2} - \sigma^2) + (\lambda - \beta_1^2)\,\overline{\delta\varepsilon}\,. \tag{2.233}$$

According to the CLT (Kartashov, 2007),

$$\sqrt{n}\,(\overline{\rho\varepsilon},\, \overline{\varepsilon^2} - \sigma_\varepsilon^2,\, \overline{\rho\delta},\, \overline{\delta^2} - \sigma_\delta^2,\, \overline{\delta\varepsilon})^{\mathrm{T}} \xrightarrow{\mathrm{d}} \gamma = (\gamma_1, \ldots, \gamma_5)^{\mathrm{T}} \sim \mathrm{N}(0, S)\,, \tag{2.234}$$

$$S = \mathrm{diag}(\sigma_\xi^2\sigma_\varepsilon^2,\, 2\sigma_\varepsilon^4,\, \sigma_\xi^2\sigma_\delta^2,\, 2\sigma_\delta^4,\, \sigma_\delta^2\sigma_\varepsilon^2) = (s_{ij})_{i,j=1}^5\,. \tag{2.235}$$

The normality of $\varepsilon$ and $\delta$, equality (2.112), and similar equality for $\mathbf{D}\varepsilon^2$ were used here; this situation is similar to the proof of Theorem 2.22. □

Further, the expansion (2.232) implies

$$\sqrt{n}(\hat\beta_1 - \beta_1) = \frac{\sqrt{n}\Lambda_n}{\sigma_\xi^2(\beta_1^2 + \lambda)} + o_P(1)\,. \tag{2.236}$$

Using the convergence in (2.234) and (2.235), we conclude that the numerator in (2.236) converges in distribution to

$$(\beta_1^2 + \lambda)\gamma_1 + \beta_1\gamma_2 - \beta_1(\beta_1^2 + \lambda)\gamma_3 - \beta_1\lambda\gamma_4 + (\lambda - \beta_1^2)\gamma_5 \sim \mathrm{N}(0, v^2)\,, \tag{2.237}$$

$$v^2 = (\beta_1^2 + \lambda)^2 s_{11} + \beta_1^2 s_{22} + \beta_1^2(\beta_1^2 + \lambda)^2 s_{33} + \beta_1^2\lambda^2 s_{44} + (\lambda - \beta_1^2)^2 s_{55} = \\ = (\beta_1^2 + \lambda)^3\sigma_\xi^2\sigma_\delta^2 + (\beta_1^2 + \lambda)^2\sigma_\delta^2\sigma_\varepsilon^2\,. \tag{2.238}$$

Then, by Lemma 2.18 and exploiting the relations (2.236)–(2.238), it follows that

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \sigma^2_{1,\lambda}) , \tag{2.239}$$

$$\sigma^2_{1,\lambda} = \frac{v^2}{\sigma^4_\xi (\beta_1^2 + \lambda)^2} = \frac{1}{\sigma^4_\xi} (\sigma^2_\varepsilon \sigma^2_x + \beta_1^2 \sigma^2_\xi \sigma^2_\delta) . \tag{2.240}$$

The theorem is proved (for the case $\beta_1 \neq 0$).

**Corollary 2.33.** *The estimator $\hat{\beta}_{1,\mathrm{ALS}}$ (under known $\sigma^2_\delta$) has larger asymptotic variance than the estimator $\hat{\beta}_{1,\mathrm{ML}}$ (under known $\lambda = \sigma^2_\varepsilon / \sigma^2_\delta$), if $\beta_1 \neq 0$:*

$$\sigma^2_{1,\mathrm{ALS}} - \sigma^2_{1,\lambda} = \frac{2\beta_1^2 \sigma^4_\delta}{\sigma^4_\xi} > 0 . \tag{2.241}$$

*Here $\sigma^2_{1,\mathrm{ALS}} = \sigma^2_1$ is given by formula* (2.91).

This result illustrates the following. If a statistician deals with a linear errors-in-variables model, where $\varepsilon$ and $\delta$ are normally distributed, it is better to design an experiment with known ratio $\sigma^2_\varepsilon / \sigma^2_\delta$ rather than with known $\sigma^2_\delta$, because in the first case, the parameter $\beta_1$ can be estimated more accurately.

Based on Theorems 2.32 and 2.31, it is possible to construct the asymptotic confidence interval for $\beta_1$ as we constructed the confidence interval (2.121). For that we demand the normality of errors $\varepsilon$ and $\delta$.

### 2.5.3 Orthogonal regression estimator (ORE)

To understand the geometric meaning of the estimators (2.208) and (2.210), we reduce the *explicit* linear model (2.1) and (2.2) to the *implicit* one (we follow the way outlined in the end of Section 1.3).

Denote

$$\eta_y = \beta_0 + \beta_1 \xi . \tag{2.242}$$

Then the explicit model can be rewritten as

$$y = \eta_y + \varepsilon , \quad x = \xi + \delta , \quad \eta_y - \beta_1 \xi - \beta_0 = 0 . \tag{2.243}$$

This is an implicit model of the form (1.25) where

$$z = (y;\ x)^{\mathrm{T}} , \quad \eta = (\eta_y;\ \xi)^{\mathrm{T}} , \quad \beta = (\beta_0;\ \beta_1)^{\mathrm{T}} , \tag{2.244}$$

In Cartesian coordinate system $(\eta_y;\ \xi)$, the straight line equality

$$\eta_y - \beta_1 \xi - \beta_0 = 0 \tag{2.245}$$

can be rewritten in a canonical form

$$\eta_y \cdot \tau_y + \xi \cdot \tau_x = d\,,$$
$$\tau_y = \frac{1}{\sqrt{1+\beta_1^2}}\,, \quad \tau_x = -\frac{\beta_1}{\sqrt{1+\beta_1^2}}\,, \quad d = \frac{\beta_0}{\sqrt{1+\beta_1^2}}\,. \tag{2.246}$$

Here $\tau_y$ and $\tau_x$ are the straight line directional cosines.

For arbitrary $\beta_0$ and $\beta_1$, the line (2.245) takes any position except vertical. At the same time, if $\tau_y$ and $\tau_x$ are allowed to get arbitrary values such that $\tau_y^2 + \tau_x^2 = 1$, then the line can take any position, including the vertical one.

The transition from coefficients $\beta_0$, $\beta_1$ to $\tau_y$, $\tau_x$, $d$ makes variables $y$, $x$ equal in rights. This is an advantage of the implicit model compared with the explicit.

Consider the implicit linear model in Euclidean space $\mathbf{R}^m$, with $m \geq 2$:

$$z_i = \eta_i + \gamma_i\,, \tag{2.247}$$
$$(\eta_i, \tau) = d\,, \quad i = \overline{1,n}\,. \tag{2.248}$$

Here $z_i$ are observed random vectors in $\mathbf{R}^m$; $\eta_i$ are latent variables that lie on the hyperplane:

$$\Gamma_{\tau d} = \{u \in \mathbf{R}^m \colon (u, \tau) = d\}\,, \tag{2.249}$$

where $\tau$ is unit normal vector to the hyperplane and $d \in \mathbf{R}$; $\gamma_i$ are random errors. By the observations $z_i$, $i = \overline{1,n}$, we want to estimate the hyperplane (2.249).

Note that the sets $\tau = \tau_0$, $d = d_0$ and $\tau = -\tau_0$, $d = -d_0$, where $\|\tau_0\| = 1$, specify the same hyperplane.

**Definition 2.34.** A random vector $\hat{\tau}$ and random variable $\hat{d}$ define the ORE of the parameters $\tau$ and $d$, if they provide a minimum of the objective function

$$Q(\tau, d) = \sum_{i=1}^{n} \rho^2(z_i, \Gamma_{\tau d})\,, \quad \|\tau\| = 1\,, \quad d \in \mathbf{R}\,. \tag{2.250}$$

The ORE is constructed as follows: we search a hyperplane $\Gamma_{\tau d}$ for which the sum of squared distances to the observed points $z_i$ is minimal.

**Theorem 2.35.** *In the model* (2.247) *and* (2.248)*, the ORE $\hat{\tau}$ is a normalized eigenvector of the sample covariance matrix*

$$\mathrm{S}_{zz} = \frac{1}{n}\sum_{i=1}^{n} (z_i - \bar{z})\,(z_i - \bar{z})^{\mathrm{T}}\,, \quad \bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i\,, \tag{2.251}$$

*$\hat{\tau}$ corresponds to the smallest eigenvalue $\lambda_{\min}(\mathrm{S}_{zz})$, and the estimator of $d$ is*

$$\hat{d} = (\bar{z}, \hat{\tau})\,. \tag{2.252}$$

*Proof.* We have

$$Q(\tau, d) = \sum_{i=1}^{n} ((z_i, \tau) - d)^2\,, \tag{2.253}$$

which for fixed $\tau$, attains its minimum at

$$d = d_{\min} = \frac{1}{n}\sum_{i=1}^{n}(z_i, \tau) = (\bar{z}, \tau)\,; \tag{2.254}$$

$$Q(\tau, d_{\min}) = \sum_{i=1}^{n}(z_i - \bar{z}, \tau)^2 = \tau^{\mathrm{T}}\,\mathrm{S}_{zz}\,\tau\,. \tag{2.255}$$

Therefore,

$$\min_{(\|\tau\|=1,\, d\in\mathbf{R})} Q(\tau, d) = \min_{\|\tau\|=1} \tau^{\mathrm{T}}\,\mathrm{S}_{zz}\,\tau = \lambda_{\min}(\mathrm{S}_{zz})\,, \tag{2.256}$$

and minimum is attained on the normalized eigenvector $\hat{\tau}$ that corresponds to $\lambda_{\min}(\mathrm{S}_{zz})$ (the eigenvector $\hat{\tau} = \hat{\tau}(\omega)$ can be chosen so that it was a random vector). Finally, equality (2.252) follows from equality (2.254). The theorem is proved. □

**Remark 2.36.** In the model (2.247) and (2.248), the ORE specifies a hyperplane $(z - \bar{z}, \hat{\tau}) = 0$ containing the center of mass $\bar{z}$ of the observed points $z_i$, $i = \overline{1, n}$.

Consider a particular case of the implicit linear model in the plane:

$$m = 2\,, \quad z_i = (y_i;\ x_i)^{\mathrm{T}}\,, \quad i = \overline{1, n}\,, \quad \tau = (\tau_x;\ \tau_y)^{\mathrm{T}}\,. \tag{2.257}$$

It is to this model, we can bring the linear scalar errors-in-variables model.

**Theorem 2.37.** *If in the two-dimensional model* (2.247), (2.248), *and* (2.257), *for some elementary event* $\omega$

$$\mathrm{S}_{xy} \neq 0\,, \tag{2.258}$$

*then for this* $\omega$, *there is a single straight line, which corresponds to the ORE, and this is the straight line*

$$y = \hat{\beta}_0 + \hat{\beta}_1 x\,, \tag{2.259}$$

*having the coefficients given by equalities* (2.208) *and* (2.210) *at* $\lambda = 1$.

*Proof.* Assume (2.258). The symmetric matrix

$$\mathrm{S}_{zz} = \begin{pmatrix} \mathrm{S}_{yy} & \mathrm{S}_{xy} \\ \mathrm{S}_{xy} & \mathrm{S}_{xx} \end{pmatrix} \tag{2.260}$$

is not diagonal, hence its two eigenvalues are distinct, and in view of Theorem 2.35, the ORE specifies a unique straight line.

The number $\lambda_{\min} = \lambda_{\min}(\mathrm{S}_{zz})$ is the smallest root of the characteristic equation

$$\det\begin{pmatrix} \mathrm{S}_{yy} - \lambda & \mathrm{S}_{xy} \\ \mathrm{S}_{xy} & \mathrm{S}_{xx} - \lambda \end{pmatrix} = 0\,; \tag{2.261}$$

$$\lambda_{\min} = \frac{\mathrm{S}_{xx} + \mathrm{S}_{yy} - \sqrt{(\mathrm{S}_{xx} - \mathrm{S}_{yy})^2 + 4\,\mathrm{S}_{xy}^2}}{2}\,. \tag{2.262}$$

The sought-for unit vector $(\hat\tau_y, \hat\tau_x)^{\mathrm{T}}$ satisfies the equation

$$\begin{pmatrix} S_{yy} - \lambda_{\min} & S_{xy} \\ S_{xy} & S_{xx} - \lambda_{\min} \end{pmatrix} \begin{pmatrix} \hat\tau_y \\ \hat\tau_x \end{pmatrix} = 0 \,. \tag{2.263}$$

From here

$$(S_{yy} - \lambda_{\min})\hat\tau + S_{xy}\,\hat\tau_x = 0 \,. \tag{2.264}$$

Since now $S_{xy} \neq 0$, then $\hat\tau_y \neq 0$, i.e., the sought-for straight line is not vertical; its equation is either $y\hat\tau_y + x\hat\tau_x = \hat d$ or

$$y = -\frac{\hat\tau_x}{\hat\tau_y}x + \frac{\hat d}{\hat\tau_y} = \tilde\beta_1 x + \tilde\beta_0 \,. \tag{2.265}$$

Find the slope $\tilde\beta_1$ from equality (2.264):

$$\tilde\beta_1 = -\frac{\hat\tau_x}{\hat\tau_y} = \frac{S_{yy} - \lambda_{\min}}{S_{xy}} = \frac{-S_{xx} + S_{yy} + \sqrt{(S_{xx} - S_{yy})^2 + 4\,S_{xy}^2}}{2\,S_{xy}} \,. \tag{2.266}$$

This coincides with $\hat\beta_1$ from the formula (2.208), with $\lambda = 1$. Since the desired straight line goes through the center of mass (see Remark 2.36), it will be exactly the straight line (2.259). The theorem is proved. □

**Corollary 2.38.** *With $n \geq 2$, consider the explicit linear model* (2.1) *and* (2.2)*, assuming that $\varepsilon_i$ and $\delta_i$ are normally distributed and $\lambda = \sigma_\varepsilon^2/\sigma_\delta^2 = 1$ (the model can be either structural or functional). Then almost surely, the estimators $\hat\beta_1$ and $\hat\beta_0$ specified by equalities* (2.208) *and* (2.210) *coincide with the OREs, i.e., the straight line* (2.259) *minimizes the objective function* (2.250).

*Proof.* Because of the normality of errors, $S_{xy} \neq 0$ almost surely. Now, the desired statement follows from Theorem 2.37. □

Further the estimators (2.208) and (2.210) are called the OREs. The corresponding objective function can be set as follows:

$$Q_{\mathrm{OR}}(\beta_0, \beta_1) = \sum_{i=1}^{n} \rho^2(M_i, \Gamma_\beta) \,, \quad \beta = (\beta_0, \beta_1)^{\mathrm{T}} \in \mathbf{R}^2 \,. \tag{2.267}$$

Here, $M_i = (x_i, y_i)$ and $\Gamma_\beta$ is the straight line $y = \beta_0 + \beta_1 x$. Thus, the estimator $\hat\beta_{\mathrm{OR}}$ is a minimum point of the function (2.267). If $S_{xy} \neq 0$, then the minimum point exists and is unique.

If the normality of errors is dropped, then the minimum of function (2.267) does not necessarily exist. This happens when the minimum of the corresponding objective function (2.250) is attained only by the vertical straight line. From Theorem 2.35 and in view of matrix (2.260), we conclude that it happens when

$$S_{xy} = 0 \,, \quad S_{yy} > S_{xx} \,. \tag{2.268}$$

In the case

$$\mathrm{S}_{xy} = 0\,, \quad \mathrm{S}_{yy} < \mathrm{S}_{xx}\,, \tag{2.269}$$

the ORE is the horizontal straight line, and if

$$\mathrm{S}_{xy} = 0\,, \quad \mathrm{S}_{yy} = \mathrm{S}_{xx}\,, \tag{2.270}$$

then the matrix $\mathrm{S}_{zz}$ is proportional to the unit one, and then any straight line passing through the center of mass produces the minimum of the function (2.250). This occurs, e.g., when the points $M_i(x_i;\ y_i)$ are the vertices of a regular polygon.

The comparison of the objective function (2.7) with $Q_{\mathrm{OR}}$ shows that the OLS estimator minimizes the sum of squared vertical distances from the points $M_i$ to the straight line. At the same time, the ORE minimizes the sum of squared "orthogonal" distances to the straight line.

The objective function $Q_{\mathrm{OR}}$ can be written more explicitly using the formula of distance from a point to a straight line:

$$Q_{\mathrm{OR}}(\beta) = \sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{1 + \beta_1^2}\,. \tag{2.271}$$

The presence of denominator distinguishes this objective function from the function (2.7).

Alternatively, the ORE is called the total least squares (TLS) Estimator. This is due to the fact that the optimization problem from Definition 2.34 can be restated as follows:

$$\min \sum_{i=1}^{n} \|\Delta z_i\|^2\,, \tag{2.272}$$

provided there exist such $\tau$, with $\|\tau\| = 1$, and $d \in \mathbf{R}$ that for all $i = \overline{1, n}$,

$$(z_i - \Delta z_i, \tau) = d\,. \tag{2.273}$$

As a result of the minimization there is formed an estimator of the desired hyperplane $(z, \hat{\tau}) = \hat{d}$ and estimators of the true points $\hat{\eta}_i = z_i - \Delta\hat{z}_i$ lying on the hyperplane.

The ORE estimator is very common in the vector model $XB \approx Y$, see (2.167). For independent errors $\varepsilon_i$ and $\delta_i$ stemming from equations (2.165), assume that the variance–covariance matrices have the form

$$V_\varepsilon = \mathbf{E}\varepsilon_i\varepsilon_i^{\mathrm{T}} = \sigma^2 I_m\,, \quad V_\delta = \mathbf{E}\delta_i\delta_i^{\mathrm{T}} = \sigma^2 I_d\,, \tag{2.274}$$

where $\sigma > 0$ is unknown and $I_m$, $I_d$ are unit matrices of corresponding size. Then a natural estimator of the matrix $B$ is the ORE estimator $\hat{B}_{\mathrm{TLS}}$, which is a solution to the optimization problem:

$$\min \left(\|\Delta X\|_{\mathrm{F}}^2 + \|\Delta Y\|_{\mathrm{F}}^2\right), \tag{2.275}$$

provided that there exists a matrix $B \in \mathbf{R}^{d\times m}$ such that

$$(X - \Delta X)\, B = Y - \Delta Y\,. \tag{2.276}$$

Here $\|Z\|_{\mathrm{F}}$ is the Frobenius norm of a matrix $Z = (z_{ij})$: $\|Z\|_{\mathrm{F}} = \sqrt{\sum_{i,j} z_{ij}^2}$.

In functional case, where the matrix $X^{\mathrm{tr}}$ is nonrandom, the estimator $\hat{\beta}_{\mathrm{TLS}}$ is the MLE, if the errors $\varepsilon_i$ and $\delta_i$ are normal and (2.274) holds.

General conditions for the consistency of $\hat{\beta}_{\mathrm{TLS}}$ are given in Kukush and Van Huffel (2004). If the matrices $X$ and $Y$ are structured (i.e., they obey an additional structure like Toeplitz or Hankel matrix), then the estimation method for the matrix $B$ taking into account the availability of such a structure is called the structured total least squares (STLS). The paper by Markovsky et al. (2004) is devoted to computation of the estimates and the article by Kukush et al. (2005a) deals with their consistency.

### 2.5.4 The ORE in implicit linear model: equivariance and consistency

We will study some properties of the ORE in the model (2.247)–(2.249).

**Theorem 2.39.** *Let $\{\eta_i, i = \overline{1,n}\}$ be nonrandom, and $\{y_i, i = \overline{1,n}\}$ be independent and identically distributed in $\mathbf{R}^m$, with distribution $\mathrm{N}(0, \sigma_y^2 I_m)$, where $\sigma_y^2 > 0$ is unknown. Then the ORE coincides with the MLE of the parameters $\tau$ and $d$.*

*Proof.* The random vector $z_i$ has the pdf

$$\rho(z_i) = \frac{1}{(\sqrt{2\pi})^m \sigma_y^m} \mathrm{e}^{-\frac{\|z_i - \eta_i\|^2}{2\sigma_y^2}}, \quad z_i \in \mathbf{R}^m . \tag{2.277}$$

The log-likelihood function is as follows:

$$L(z_1, \dots, z_n;\ \eta_1, \dots, \eta_n) = -\frac{1}{2\sigma_y^2} \sum_{i=1}^{n} \|z_i - \eta_i\|^2 + f_n(\sigma_y) . \tag{2.278}$$

Its maximization in $\eta_1, \dots, \eta_n$ leads to the optimization problem (2.272) and (2.273), with $\Delta z_i = z_i - \eta_i$, $i = \overline{1,n}$. Therefore, the MLEs $\hat{\tau}_{\mathrm{ML}}$ and $\hat{d}_{\mathrm{ML}}$ are the OREs. The theorem is proved. □

The following statement stems from the geometric meaning of the ORE.

**Theorem 2.40.** *Let $\Gamma_{\hat{\tau}\hat{d}}$ be the hyperplane* (2.249)*, which is the ORE in the model* (2.247) *and* (2.248)*; the estimator is based on the sample $z_i$, $i = \overline{1,n}$. Let $U$ be either an orthogonal operator, or a translation operator $Uz = z + c$, $z \in \mathbf{R}^m$, with $c \in \mathbf{R}^m$, or a homothetic transformation $Uz = kz$, $z \in \mathbf{R}^m$, with $k \in \mathbf{R}$, $k \neq 0$. Then the ORE based on a sample $Uz_i$, $i = \overline{1,n}$, is the transformed hyperplane $U\Gamma_{\hat{\tau}\hat{d}}$.*

The theorem demonstrates the following: a position of the hyperplane $\Gamma_{\hat{\tau}\hat{d}}$ relative to the observable points $z_i$ does not depend of the choice of neither a Cartesian coordinate system nor a scale (at the same time, the scale has to be the same in all directions). Due to these properties of the ORE, this estimator looks natural in pattern recognition problems. Such concordant variability of an estimator regarding certain transformation group is called *equivariance* (Schervish, 1995, Chapter 6).

Now, we prove the ORE consistency in the structural model.

**Theorem 2.41.** *Consider the model* (2.247)–(2.249). *Assume the following.*
(a) *Random vectors* $\eta_i$, $\gamma_i$, $i \geq 1$ *are independent.*
(b) *Random vectors* $\gamma_i$, $i \geq 1$ *are identically distributed in* $\mathbf{R}^m$, *with zero mean and the variance–covariance matrix* $S_\gamma = \sigma_\gamma^2 I_m$, *where* $\sigma_\gamma^2 > 0$ *is unknown.*
(c) *Random vectors* $\eta_i$, $i \geq 1$ *are identically distributed in* $\Gamma_{\tau d} \subset \mathbf{R}^m$, *with covariance matrix* $S_\eta$ *of rank* $m - 1$.
*Then the ORE* $(\hat\tau\,;\ \hat d)$ *is strongly consistent, i.e., as* $n \to \infty$,

$$\min\left\{\left\|\hat\tau - \tau\right\| + \left|\hat d - d\right|, \left\|\hat\tau + \tau\right\| + \left|\hat d + d\right|\right\} \xrightarrow{\mathbf{P1}} 0\,. \tag{2.279}$$

**Remark 2.42.** The convergence (2.279) is due to the fact that the couple $\tau = \tau_0$, $d = d_0$ determines the same hyperplane as the couple $\tau = -\tau_0$, $d = -d_0$.

*Proof of the theorem.* We have by Theorem 2.35 that

$$\mathrm{S}_{zz}\,\hat\tau = \lambda_{\min}(\mathrm{S}_{zz}) \cdot \hat\tau\,, \quad \left\|\hat\tau\right\| = 1\,, \quad \hat d = (\bar z, \hat\tau)\,. \tag{2.280}$$

By the SLLN,

$$\mathrm{S}_{zz} \xrightarrow{\mathbf{P1}} \mathrm{cov}(z_1) = S_\eta + S_\gamma = S_\eta + \sigma_\gamma^2 I_m = S_\infty\,. \tag{2.281}$$

Since $S_\eta$ is positive semidefinite matrix, then

$$\lambda_{\min}(S_\infty) \geq \lambda_{\min}(\sigma_\gamma^2 I_m) = \sigma_\gamma^2\,. \tag{2.282}$$

Next, $(\eta_1, \tau) = d$, because

$$0 = \mathbf{D}(\eta_1, \tau) = \tau^{\mathrm{T}} S_\eta \tau\,, \quad S_\eta \tau = 0\,, \quad S_\infty \tau = \sigma_\gamma^2 \tau\,. \tag{2.283}$$

So $\tau$ is a normalized eigenvector of the matrix $S_\infty$, which corresponds to the smallest eigenvalue

$$\lambda_{\min}(S_\infty) = \sigma_\gamma^2\,. \tag{2.284}$$

Show that the eigenvalue has multiplicity 1. Indeed, let $S_\infty \upsilon = \sigma_\gamma^2 \upsilon$, $\|\upsilon\| = 1$. Then $S_\eta \upsilon = 0$. But the kernel $Ker S_\eta = \{z \in \mathbf{R}^m\colon S_\eta z = 0\}$ has dimension $m - rk(S_\eta) = 1$, hence, $\upsilon = \pm\tau$.

By Wedin's theorem (Stewart and Sun, 1990) on stability of eigenvectors for a matrix, the convergence (2.281) implies that almost surely a sequence $\{\hat\tau = \hat\tau_n(\omega), n \geq 1\}$ of normalized eigenvectors corresponding to $\lambda_{\min}(\mathrm{S}_{zz})$ may have only two limit points $\pm\tau$, which are the normalized eigenvectors of the matrix $S_\infty$ corresponding to the simple eigenvalue $\lambda_{\min}(S_\infty)$.

Suppose that for a fixed $\omega$, the sequence $\hat\tau = \hat\tau_n(\omega)$ is divided into two subsequences $\hat\tau_{n'}(\omega)$ and $\hat\tau_{n''}(\omega)$, moreover $\hat\tau_{n'}(\omega) \to \tau$ and $\hat\tau_{n''}(\omega) \to -\tau$. Then $\hat d_{n'}(\omega) = (\bar z, \hat\tau_{n'}) \to (\mathbf{E}z_1, \tau) = d$. The latter is true because $\mathbf{E}z_1 \in \Gamma_{\tau d}$. Similarly, $\hat d_{n''}(\omega) \to -(\mathbf{E}z_1, \tau) = -d$. The resulting convergences justify (2.279). □

# 3 Polynomial regression with known variance of classical error

In Chapter 1, it was mentioned that the binary observation model (1.2), (1.4), and (1.5) is widespread in radio-epidemiology. This binary model is the so-called *generalized linear model* (GLM). This means that the conditional distribution of the response $y$ given the true value of regressor $\xi$ is expressed through a linear function in $\xi$, with some unknown parameters. The linear function defines the odds function (1.4). For more information on the generalized linear models, see Chapter 4.

However, a quadratic odds function is used in radio-epidemiology as well:

$$\lambda(\xi_i, \beta) = \beta_0 + \beta_1 \xi_i - \beta_2 \xi_i^2 , \quad i = \overline{1, n} , \quad \beta = (\beta_0, \beta_1, \beta_2)^{\mathrm{T}} . \tag{3.1}$$

The model (1.2), (3.1), and (1.5) serves for modeling the thyroid cancer incidence as a result of exposure by radioactive iodine: $\xi_i$ is a radiation dose received by a subject $i$ from a cohort during a fixed observation period. Positive radiation risk parameters $\beta_0, \beta_1$, and $\beta_2$ are to be estimated. The presence of a negative term member in the quadratic odds function (3.1) describes the effect of burning out cancer cells at high exposure doses, which may even lead to some reduction of disease incidence.

The binary regression model (1.2) and (3.1) is no longer the GLM model, because now the conditional distribution of $y_i$ given $\xi_i$ is expressed through the quadratic (3.1), but not through a linear function in $\xi_i$.

To get a feeling for the effect of polynomial influence of regressor on response, consider the *polynomial regression model* with classical measurement error. It is described by the two equations:

$$y_i = \beta_0 + \beta_1 \xi_i + \cdots + \beta_k \xi_i^k + \varepsilon_i , \tag{3.2}$$

$$x_i = \xi_i + \delta_i , \quad i = \overline{1, n} . \tag{3.3}$$

Here, $k \geq 1$ is a given degree of the polynomial (at $k = 1$ we get the linear model from Chapter 2), $\xi_i$ is a latent variable, $x_i$ is an observed surrogate data, $y_i$ is an observed response, $\varepsilon_i$ and $\delta_i$ are observation errors. The regression parameter

$$\beta = (\beta_0, \beta_1, \ldots, \beta_k)^{\mathrm{T}} \tag{3.4}$$

has to be estimated.

Introduce the vector function

$$\rho(\xi) = (1, \xi, \ldots, \xi^k)^{\mathrm{T}} , \quad \xi \in \mathbf{R} . \tag{3.5}$$

Now, equality (3.2) can be rewritten in the compact form

$$y_i = \rho^{\mathrm{T}}(\xi_i)\beta + \varepsilon_i . \tag{3.6}$$

We consider the structural model in which the $\xi_i$ are random variables. Make the following assumptions about the observation model.

(i) Random variables $\xi_i$, $\varepsilon_i$, and $\delta_i$ are independent.
(ii) The errors $\varepsilon_i$ are identically distributed with distribution $\mathrm{N}(0, \sigma_\varepsilon^2)$, $\sigma_\varepsilon^2 > 0$.
(iii) The errors $\delta_i$ are identically distributed with distribution $\mathrm{N}(0, \sigma_\delta^2)$, $\sigma_\delta^2 > 0$.
(iv) Random variables $\xi_i$ are identically distributed with distribution $\mathrm{N}(\mu_\xi, \sigma_\xi^2)$, $\sigma_\xi^2 > 0$.

The model (3.2) and (3.3), with assumptions (i)–(iv), is called *normal* structural polynomial model with classical measurement errors. Such models can be used, e.g., in econometrics (Carroll et al., 2006).

Section 2.3 states that at $k = 1$ without additional assumptions about the parameters, this model (in this case being linear) is not identifiable. From here, it follows as well that at $k \geq 2$, the normal polynomial model is not identifiable, because in the proof of Theorem 2.10 for the polynomial model, an additional condition for the rest of the true parameters can be imposed:

$$\beta_2 = \cdots = \beta_k = 0 \,. \tag{3.7}$$

In the absence of further restrictions, the normal polynomial model is determined by a vector parameter $\theta$ and the corresponding parameter set $\Theta$:

$$\theta = (\beta^{\mathrm{T}}, \mu_\xi, \sigma_\xi^2, \sigma_\delta^2, \sigma_\varepsilon^2)^{\mathrm{T}} \in \mathbf{R}^{k+5} \,, \quad \Theta = \mathbf{R}^{k+1} \times \mathbf{R} \times (0, +\infty)^3 \,. \tag{3.8}$$

Let it be assumed additionally that $k \geq 2$ and condition (3.7) is violated, i.e., the true regression function

$$f(\xi_i, \beta) = \beta_0 + \beta_1 \xi_i + \cdots + \beta_k \xi_i^k \tag{3.9}$$

is not linear in $\xi_i$. Then the model becomes identifiable due to the fact that the distribution of the response $y_i$ is never normal. In this situation, one can construct consistent estimators of all the parameters using the method of moments (see discussion for the quadratic model ($k = 2$) in Carroll et al., 2006, Section 5.5.3).

If we allow the degeneracy (3.7), then we need some restriction of the parameter set (3.8) to get the identifiability of the model. *In this section, it is required that the measurement error variance $\sigma_\delta^2$ is known.*

Rarely, the ratio of error variances

$$\lambda = \frac{\sigma_\varepsilon^2}{\sigma_\delta^2} \,. \tag{3.10}$$

is assumed known. In Shklyar (2008), consistent estimators of model parameters are constructed for this case.

Note that for $k \geq 2$, the ML method in the normal model with assumptions (i)–(v) is not feasible (see discussion in Section 1.4.2). In particular, even under known $\sigma_\varepsilon^2$ (then $\sigma_\delta^2$ is known as well as a result of (v)) the joint pdf of the observed variables $y$ and $x$ is given by the integral (1.54), with the polynomial regression function $f(\xi) = f(\xi, \beta)$ as defined in (3.9). The integral is not calculated analytically, which complicates the usage of the ML method and makes it problematic to study properties of the estimator.

## 3.1 The adjusted least squares estimator

### 3.1.1 The formula for the estimator

The corrected score method was described in Section 1.4.4. Apply it to the model (3.6) and (3.3). The regression function is

$$f(\xi, \beta) = \rho^{\mathrm{T}}(\xi)\,\beta\,, \quad \xi \in \mathbf{R}\,, \quad \beta \in \mathbf{R}^{k+1}\,. \tag{3.11}$$

Since $\partial f/\partial\beta = \rho(\xi)$, the basic deconvolution equations (1.124) and (1.125) take the form

$$\mathbf{E}[g(x, b)|\xi] = \rho(\xi)\,, \tag{3.12}$$

$$\mathbf{E}[h(x, b)|\xi] = \rho(\xi)\rho^{\mathrm{T}}(\xi)\beta\,. \tag{3.13}$$

Within the class of polynomials in $\xi$, the solutions are unique:

$$g(x, b) = t(x)\,, \quad h(x, b) = H(x)\beta\,, \tag{3.14}$$

where the vector function $t(x)$ and matrix-valued function $H(x)$ satisfy the deconvolution equations

$$\mathbf{E}[t(x)|\xi] = \rho(\xi)\,, \quad \mathbf{E}[H(x)|\xi] = \rho(\xi)\rho^{\mathrm{T}}(\xi)\,. \tag{3.15}$$

Hereafter, all the equalities for conditional expectations hold almost surely (a.s.). For the $j$th component of the function $t(x)$, it holds that

$$\mathbf{E}[t_j(x)|\xi] = \xi^j\,, \quad j \geq 0\,. \tag{3.16}$$

For an entry $H_{ij}(x)$ of the matrix $H(x)$, we have

$$\mathbf{E}[H_{ij}(x)|\xi] = \xi^i \cdot \xi^j = \xi^{i+j}\,, \quad 0 \leq i, j \leq k\,. \tag{3.17}$$

Below we show that deconvolution equation (3.16) has a unique polynomial solution $t_j(x)$, and then

$$H_{ij}(x) = t_{i+j}(x)\,, \quad 0 \leq i, j \leq k\,, \tag{3.18}$$

is the only polynomial solution to equation (3.17).

A solution to equation (3.16) is given by the *Hermite polynomial* $H_j(x)$ closely related to normal distribution. The polynomial can be specified by an explicit formula through higher derivatives:

$$H_j(x) = (-1)^n e^{\frac{x^2}{2}} \left(\mathrm{e}^{-\frac{x^2}{2}}\right)^{(j)}\,, \quad x \in \mathbf{R}\,, \quad j \geq 0\,. \tag{3.19}$$

In particular,

$$H_0(x) = 1\,, \quad H_1(x) = x\,. \tag{3.20}$$

The recurrence relation is

$$H_j(x) = xH_{j-1}(x) - (j-1)H_{j-2}(x)\,, \quad x \in \mathbf{R}\,, \quad j \geq 2\,. \tag{3.21}$$

Applying the formula for $j = 2$ and $j = 3$, we get

$$H_2(x) = x^2 - 1\,, \quad H_3(x) = x^3 - 3x\,, \quad x \in \mathbf{R}\,. \tag{3.22}$$

The recurrence relation (3.21) allows us to compute consequently the next Hermite polynomials; all of them have unit leading coefficient. Let

$$\gamma \sim \mathrm{N}(0,1)\,. \tag{3.23}$$

Then for all $n, m \geq 0$, the equality holds:

$$\mathbf{E}H_n(\gamma)H_m(\gamma) = n!\,\delta_{nm}\,. \tag{3.24}$$

Here, $\delta_{nm}$ is the Kronecker symbol:

$$\delta_{nm} = \begin{cases} 1 & \text{if } n = m\,, \\ 0 & \text{if } n \neq m\,. \end{cases} \tag{3.25}$$

The next Hermite polynomial property is due to Stulajter (1978). We give a simple proof.

**Lemma 3.1.** *For a standard normal random variable $\gamma$, it holds that*

$$\mathbf{E}H_n(\mu + \gamma) = \mu^n\,, \quad n \geq 0\,, \quad \mu \in \mathbf{R}\,. \tag{3.26}$$

*Proof.* We use induction. Denote

$$I_n = I_n(\mu) = \mathbf{E}H_n(\mu + \gamma)\,. \tag{3.27}$$

(a) For $n = 0$, we have taken into account (3.20):

$$I_0(\mu) = \mathbf{E}H_0(\mu + \gamma) = 1 = \mu^0\,, \tag{3.28}$$

and (3.26) holds true. For $n = 1$, we have, see (3.20):

$$I_1(\mu) = \mathbf{E}H_1(\mu + \gamma) = \mathbf{E}(\mu + \gamma) = \mu = \mu^1\,, \tag{3.29}$$

and (3.26) is fulfilled as well.

(b) Derive a recurrence relation for expressions (3.27). If $n \geq 1$, we have in view of the fact that $\mu + \gamma \sim \mathrm{N}(\mu, 1)$:

$$I_n = \int\limits_{\mathbf{R}} H_n(t)\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{(t-\mu)^2}{2}}\,dt\,, \tag{3.30}$$

$$I_n'(\mu) = \frac{1}{\sqrt{2\pi}}\int\limits_{\mathbf{R}} H_n(t)\frac{\partial}{\partial\mu}\mathrm{e}^{-\frac{(t-\mu)^2}{2}}\,dt = \frac{1}{\sqrt{2\pi}}\int\limits_{\mathbf{R}} H_n(t)\mathrm{e}^{-\frac{(t-\mu)^2}{2}}(t-\mu)dt\,. \tag{3.31}$$

Using equation (3.30) and identity (3.21), with $j = n + 1$, further we get

$$I_n'(\mu) = -\mu I_n + \frac{1}{\sqrt{2\pi}} \int_{\mathbf{R}} (H_{n+1}(t) + nH_{n-1}(t)) \mathrm{e}^{-\frac{(t-\mu)^2}{2}} dt = -\mu I_n + I_{n+1} + nI_{n-1} \,, \tag{3.32}$$

$$I_{n+1} = I_n' + \mu I_n - nI_{n-1} \,, \quad n \geq 1 \,. \tag{3.33}$$

(c) Assume that (3.26) holds true, for all $n \leq k$, where $k \geq 1$ is fixed. Then from equation (3.33), we will have

$$I_{k+1}(\mu) = (\mu^k)' + \mu \cdot \mu^k - k\mu^{k-1} = \mu^{k+1} \,. \tag{3.34}$$

Thus, we obtain (3.26), with $n = k + 1$.

According to the method of mathematical induction, (3.26) has been proved for all $n \geq 0$, with $\mu \in \mathbf{R}$. □

**Corollary 3.2.** *In case $\sigma_\delta^2 = 1$, equality* (3.16) *is valid, with $t_j(x) = H_j(x)$.*

*Proof.* By equality (3.26), it follows (now, both $\xi$ and $\gamma$ are independent and $\gamma \sim \mathrm{N}(0, 1)$) that:

$$\mathbf{E}[H_j(x)|\xi] = \mathbf{E}[H_j(\xi + \gamma)|\xi] = \xi^j \,, \quad j \geq 0 \,, \tag{3.35}$$

which proves the desired statement. □

**Lemma 3.3.** *In case of arbitrary $\sigma_\delta^2 > 0$, equality* (3.16) *holds true, with*

$$t_j(x) = (\sigma_\delta)^j H_j\left(\frac{x}{\sigma_\delta}\right) \,, \quad j \geq 0 \,. \tag{3.36}$$

*Proof.* Put $\delta = \sigma_\delta \gamma$, $\gamma \sim \mathrm{N}(0, 1)$, then by Corollary 3.2,

$$\mathbf{E}\left[ (\sigma_\delta)^j H_j\left(\frac{\xi + \sigma_\delta \gamma}{\sigma_\delta}\right) \,\middle|\, \xi \right] = (\sigma_\delta)^j \mathbf{E}\left[ H_j\left(\frac{\xi}{\sigma_\delta} + \gamma\right) \,\middle|\, \frac{\xi}{\sigma_\delta} \right] = (\sigma_\delta)^j \left(\frac{\xi}{\sigma_\delta}\right)^j = \xi^j \,, \tag{3.37}$$

which proves the statement of the lemma. □

As we can see, the function (3.36) is the only solution to the deconvolution equation (3.16) in the class of polynomials in $\xi$. Now, construct the estimating function (1.126) by means of the ALS method:

$$s_{\mathrm{C}}(y, x;\ b) = g(x, b)\, y - h(x, b) = t(x)\, y - H(x)\, b \,. \tag{3.38}$$

The ALS estimator $\hat{\beta}_{\mathrm{C}}$ is found from the equation

$$\frac{1}{n} \sum_{i=1}^{n} t(x_i) y_i - \left( \frac{1}{n} \sum_{i=1}^{n} H(x_i) \right) \hat{\beta}_{\mathrm{C}} = 0 \,, \tag{3.39}$$

$$\hat{\beta}_{\mathrm{C}} = \left(\overline{H(x)}\right)^{-1} \overline{t(x) y} \,. \tag{3.40}$$

Here like in previous chapters, bar means averaging over a given sample; the formula (3.40) is valid when the matrix $\overline{H(x)}$ is nonsingular. One can weaken the condition (iv) about the normality of $\xi$ and provide nonsingularity of the matrix *eventually*, i.e., almost surely for all $n \geq n_0(\omega)$. Consider the following milder condition.

(v) Random variables $\xi_i$ are identically distributed, with $\mathbf{E}(\xi_1)^{2k} < \infty$; moreover, the distribution of $\xi_1$ is not concentrated at $k$ or even fewer points.

The latter requirement about the distribution means the following: for each set $\{a_1, \dots, a_k\} \subset \mathbf{R}$,

$$\mathbf{P}\{\xi_1 \in \{a_1, \dots, a_k\}\} < 1 \,. \tag{3.41}$$

**Lemma 3.4.** *Assume the conditions* (iii) *and* (v). *Then,* eventually *the matrix* $\overline{H(x)}$ *is nonsingular.*

*Proof.* From (3.15), using the SLLN, we obtain (here $x =^{\mathrm{d}} x_1$):

$$\overline{H(x)} = \frac{1}{n}\sum_{i=1}^{n} H(x_i) \xrightarrow{\mathbf{P}1} \mathbf{E}H(x) = \mathbf{E}\mathbf{E}[H(x)|\xi] = \mathbf{E}\rho(\xi)\rho^{\mathrm{T}}(\xi) \,, \quad \text{as } n \to \infty \,. \tag{3.42}$$

The limit matrix is the Gram matrix for random variables $1, \xi, \dots, \xi^k$ in the space $L_2(\Omega, \mathbf{P})$ of random variables on $\Omega$ having finite second moment. In this space, an inner product is

$$(\xi, \eta) = \mathbf{E}\xi\eta \,. \tag{3.43}$$

This Gram matrix is nonsingular if, and only if, the random variables $1, \xi, \dots, \xi^k$ are linearly independent in $L_2(\Omega, \mathbf{P})$. Prove that the latter holds.

Suppose that for some real numbers $a_0, \dots, a_k$, we have

$$a_0 + a_1\xi + \cdots + a_k\xi^k = 0 \,, \quad \text{a.s.} \tag{3.44}$$

Then, $\xi$ coincides almost surely with some root of the polynomial $p(z) = a_0 + a_1 z + \dots a_k z^k$. If not all coefficients of the polynomial are zeros, then the polynomial has no more than $k$ real roots, and therefore, $\xi$ almost surely belongs to the set of roots. Thus, we got a contradiction to condition (v) about the distribution of $\xi$. So $a_0 = a_1 = \cdots = a_k = 0$ proving the linear independence of $1, \xi, \dots, \xi^k$.

Then the matrix $\mathbf{E}\rho(\xi)\rho^{\mathrm{T}}(\xi)$ is nonsingular, and (3.42) implies that

$$\det \overline{H(x)} \xrightarrow{\mathbf{P}1} \det(\mathbf{E}\rho(\xi)\rho^{\mathrm{T}}(\xi)) \neq 0 \,. \tag{3.45}$$

Thus, *eventually* the matrix $\overline{H(x)}$ is nonsingular.

Lemma 3.4 shows that under conditions (iii) and (v), the ALS estimator is *eventually* given by formula (3.40). In particular, in the case (iv), $\xi$ has a continuous distribution and then condition (v) holds true. □

### 3.1.2 Consistency of the estimator

The ALS estimator remains strongly consistent without the assumption on normality of errors $\varepsilon_i$ and regressors $\xi_i$. Introduce a weaker assumption.
(vi) The errors $\varepsilon_i$ are centered and identically distributed.

**Theorem 3.5.** *Assume the conditions* (i), (iii), (v), *and* (vi). *Then*

$$\hat{\beta}_{\mathrm{C}} \xrightarrow{\mathbf{P}1} \beta\,, \quad as \;\; n \to \infty\,. \tag{3.46}$$

*Proof.* Use equality (3.40) that holds *eventually*.

By the SLLN and the first equality in (3.15), we have

$$\begin{aligned}\overline{t(x)y} \xrightarrow{\mathbf{P}1} \mathbf{E}_\beta t(x)y &= \mathbf{E}t(x)(\rho^{\mathrm{T}}(\xi)\beta + \varepsilon) = (\mathbf{E}t(x)\rho^{\mathrm{T}}(\xi))\beta + \mathbf{E}t(x)\cdot\mathbf{E}\varepsilon \\ &= [\mathbf{E}\mathbf{E}(t(x)\rho^{\mathrm{T}}(\xi)|\xi)]\beta = [\mathbf{E}\mathbf{E}(t(x)|\xi)\rho^{\mathrm{T}}(\xi)]\beta = [\mathbf{E}\rho(\xi)\rho^{\mathrm{T}}(\xi)]\beta\,.\end{aligned} \tag{3.47}$$

The condition (v) holds. Thus, according to (3.45), the matrix $\mathbf{E}\rho\rho^{\mathrm{T}}$ is nonsingular. In equation (3.40), let us tend $n$ to infinity:

$$\hat{\beta}_{\mathrm{C}} \xrightarrow{\mathbf{P}1} (\mathbf{E}\rho\rho^{\mathrm{T}})^{-1}(\mathbf{E}\rho\rho^{\mathrm{T}})\beta = \beta\,. \tag{3.48}$$

The proof is accomplished. □

### 3.1.3 Conditional expectation and conditional variance of response

To ensure asymptotic normality of the estimator, we need the existence of the second moment of errors $\varepsilon_i$. Assume the following.
(vii)The errors $\varepsilon_i$ are centered with variance $\sigma_\varepsilon^2 > 0$.

Given (i), (iii), (iv), and (vii), write down the conditional expectation $m(x, \beta) = \mathbf{E}(y|x)$ and conditional variance $v(x;\ \beta, \sigma_\varepsilon^2) = \mathbf{V}(y|x)$. Hereafter $(y, x, \xi, \varepsilon, \delta)$ are the copies of random variables $(y_1, x_1, \xi_1, \varepsilon_1, \delta_1)$ from the model (3.2), (3.3), in particular,

$$y = \rho^{\mathrm{T}}(\xi)\beta + \varepsilon\,, \quad x = \xi + \delta\,. \tag{3.49}$$

We use relations (1.86) and (1.97). Denote

$$\mu(x) = \mathbf{E}[\rho(\xi)|x] = \mathbf{E}[\rho(\mu_1(x) + \tau\gamma)|x]\,. \tag{3.50}$$

Here, $x \perp\!\!\!\perp \gamma$, $\gamma \sim \mathrm{N}(0, 1)$, and $\mu_1(x)$ and $\tau$ are given in (1.86). We have

$$\mu(x) = (\mu_i(x))_{i=0}^k\,, \quad \mu_0(x) = 1\,, \quad \mu_1(x) = Kx + (1-K)\mu_\xi\,, \tag{3.51}$$

$$\mu_2(x) = \mathbf{E}[(\mu_1(x) + \tau\gamma)^2|x] = \mu_1^2(x) + \tau^2\,, \tag{3.52}$$

$$\mu_3(x) = \mathbf{E}[(\mu_1(x) + \tau\gamma)^3|x] = \mu_1^3(x) + 3\mu_1(x)\tau^2\,, \tag{3.53}$$

$$\mu_4(x) = \mathbf{E}[(\mu_1(x) + \tau\gamma)^4|x] = \mu_1^4(x) + 6\mu_1^2(x)\tau^2 + 3\tau^4\,. \tag{3.54}$$

If one needs, it is easy to find further values of $\mu_i(x)$ using the moments of $\gamma$. Next,

$$m(x,\beta) = \mathbf{E}[\rho^{\mathrm{T}}(\xi)\beta|x] = \mathbf{E}[\rho^{\mathrm{T}}(\xi)|x]\cdot\beta = \mu^{\mathrm{T}}(x)\beta\,. \tag{3.55}$$

To find the conditional variance, we apply the formula (1.65):

$$v(x;\ \beta,\sigma_\varepsilon^2) = \mathbf{E}[\mathbf{V}(y|\xi)|x] + \mathbf{V}[\mathbf{E}(y|\xi)|x] = \sigma_\varepsilon^2 + \mathbf{V}(\rho^{\mathrm{T}}(\xi)\beta|x)\,, \tag{3.56}$$

$$\begin{aligned}\mathbf{V}(\rho^{\mathrm{T}}(\xi)\beta|x) &= \mathbf{E}[\beta^{\mathrm{T}}(\rho(\xi)-\mu(x))(\rho(\xi)-\mu(x))^{\mathrm{T}}\beta|x] = \\ &= \beta^{\mathrm{T}}(M(x)-\mu(x)\mu^{\mathrm{T}}(x))\beta\,.\end{aligned} \tag{3.57}$$

Here,

$$M(x) = \mathbf{E}[\rho(\xi)\rho(\xi)^{\mathrm{T}}|x] = (M_{ij}(x))_{i,j=0}^{k}\,, \tag{3.58}$$

$$M_{ij}(x) = \mathbf{E}(\xi^{i+j}|x) = \mu_{i+j}(x)\,,\quad 0\le i,j\le k\,. \tag{3.59}$$

Thus,

$$v(x;\ \beta,\sigma_\varepsilon^2) = \sigma_\varepsilon^2 + \beta^{\mathrm{T}}(M(x)-\mu(x)\mu^{\mathrm{T}}(x))\beta\,. \tag{3.60}$$

In the polynomial model (3.49), the variance of $y$ given $x$ does not depend of the intercept $\beta_0$, because adding a constant to $y$ does not change the conditional variance. Therefore, the formula for $v(x;\ \beta,\sigma_\varepsilon^2)$ can be rewritten. Denote

$$\beta_0 = (\beta_1,\dots,\beta_k)^{\mathrm{T}}\,,\quad M_{-0} = (M_{ij}(x))_{i,j=1}^{k} = (\mu_{i+j}(x))_{i,j=1}^{k}\,, \tag{3.61}$$

$$\mu_{-0}(x) = (\mu_i(x))_{i=1}^{k}\,. \tag{3.62}$$

Thus, we deleted the null coordinate in vectors $\beta$ and $\mu(x)$ and deleted both zero row and zero column in the matrix $M(x)$. Then

$$v(x;\ \beta,\sigma_\varepsilon^2) = \sigma_\varepsilon^2 + \beta_{-0}^{\mathrm{T}}(M_{-0}(x)-\mu_{-0}(x)\mu_{-0}^{\mathrm{T}}(x))\beta_{-0}\,. \tag{3.63}$$

Consider the particular cases of linear ($k=1$) and quadratic ($k=2$) model.

**Lemma 3.6.** *Assume the conditions* (i), (iii), (iv), *and* (vii). *Then in the linear model,*

$$m(x,\beta) = \beta_0 + \mu_1(x)\beta_1 = \beta_0 + (Kx + (1-K)\mu_\xi)\beta_1\,, \tag{3.64}$$

$$v(x,\beta) = v(\beta) = \sigma_\varepsilon^2 + \tau^2\beta_1^2 = \sigma_\varepsilon^2 + K\sigma_\delta^2\beta_1^2\,; \tag{3.65}$$

*and in the square model,*

$$m(x,\beta) = \beta_0 + \mu_1(x)\beta_1 + (\mu_1^2(x)+\tau^2)\beta_2\,, \tag{3.66}$$

$$v(x;\ \beta,\sigma_\varepsilon^2) = \sigma_\varepsilon^2 + \tau^2\beta_1^2 + 4\beta_1\beta_2\mu_1(x)\tau^2 + 2\beta_2^2(2\mu_1^2(x)\tau^2+\tau^4)\,. \tag{3.67}$$

*Proof.* The formulas for conditional means stem from equalities (3.55), (3.51), and (3.52). Next, we use equality (3.63) to find the conditional variances.
(a) $k=1$. By formula (3.52) we obtain

$$v(x;\ \beta,\sigma_\varepsilon^2) = \sigma_\varepsilon^2 + \beta_1^2(\mu_2(x)-\mu_1^2(x)) = \sigma_\varepsilon^2 + \tau^2\beta_1^2\,. \tag{3.68}$$

Therefore, the conditional variance does not depend on $x$ in the linear model.

(b) $k = 2$. Find the entries of the symmetric matrix $M_{-0}(x) - \mu_{-0}(x)\mu_{-0}^{\mathrm{T}}(x) =: N(x)$ of size $2 \times 2$ :

$$N_{12}(x) = \mu_3(x) - \mu_1(x)\mu_2(x) = 2\mu_1(x)\tau^2 , \tag{3.69}$$

$$N_{22}(x) = \mu_4(x) - \mu_2^2(x) = 4\mu_1^2(x)\tau^2 + 2\tau^4 . \tag{3.70}$$

Here, we used the calculations (3.52)–(3.54). From here,

$$v(x;\ \beta, \sigma_\varepsilon^2) = \sigma_\varepsilon^2 + \tau^2\beta_1^2 + 2\beta_1\beta_2 N_{12}(x) + \beta_2^2 N_{22}(x) , \tag{3.71}$$

and formula (3.67) is proved. As we can see in the square model, the conditional variance does not depend of $x$. The lemma is proved. □

### 3.1.4 Asymptotic normality of the estimator

**Theorem 3.7.** *Assume the conditions* (i), (iii), (iv), *and* (vii). *Then*

$$\sqrt{n}(\hat{\beta}_{\mathrm{C}} - \beta) \ \xrightarrow{\mathrm{d}}\ \mathrm{N}(0, \Sigma_{\mathrm{C}}) , \tag{3.72}$$

*where the asymptotic covariance matrix (ACM)* $\Sigma_{\mathrm{C}}$ *is nonsingular and depends on unknown parameters* $\beta_{-0}$ *and* $\sigma_\varepsilon^2$,

$$\Sigma_{\mathrm{C}} = A_{\mathrm{C}}^{-1} B_{\mathrm{C}} A_{\mathrm{C}}^{-\mathrm{T}} , \tag{3.73}$$

$$A_{\mathrm{C}} = \mathbf{E}\rho\rho^{\mathrm{T}} , \quad B_{\mathrm{C}} = \mathbf{E}vtt^{\mathrm{T}} + \mathbf{E}(t\mu_{-0}^{\mathrm{T}} - H_{-0})\beta_{-0}\beta_{-0}^{\mathrm{T}}(t\mu_{-0}^{\mathrm{T}} - H_{-0})^{\mathrm{T}} . \tag{3.74}$$

*Here* $v$ *is given in* (3.63) *and* $t = t(x) = (t_j(x))_{j=0}^k$ *is determined in* (3.36),

$$H_{-0} = H_{-0}(x) = (H_{ij}(x))_{\substack{0\le i\le k,\\ 1\le j\le k}} = (t_{i+j}(x))_{\substack{0\le i\le k,\\ 1\le j\le k}} . \tag{3.75}$$

*Proof.* From formula (3.40), we have *eventually*:

$$\sqrt{n}(\hat{\beta}_{\mathrm{C}} - \beta) = (\overline{H(x)})^{-1} \cdot \sqrt{n}(\overline{t(x)y} - \overline{H(x)}\beta) . \tag{3.76}$$

According to (3.42),

$$\overline{H(x)} \ \xrightarrow{\mathbf{P}1}\ \mathbf{E}\rho\rho^{-1} = A_{\mathrm{C}} > 0 . \tag{3.77}$$

Remember that the latter notation means that a matrix is positive definite. Further, as a result of calculation (3.47),

$$\mathbf{E}_\beta(t(x)y - H(x)\beta) = 0 \tag{3.78}$$

(actually, this is unbiasedness of the estimating function (3.38)), and the CLT can be applied to the factor in (3.76):

$$\sqrt{n}(\overline{t(x)y} - \overline{H(x)\beta}) = \frac{1}{\sqrt{n}}\sum_{i=1}^n (t(x_i)y_i) - H(x_i)\beta) \ \xrightarrow{\mathrm{d}}\ \eta \sim \mathrm{N}(0, B_{\mathrm{C}}) , \tag{3.79}$$

$$B_{\mathrm{C}} = \mathrm{cov}_\beta(t(x)y - H(x)\beta) . \tag{3.80}$$

Below, for brevity, we omit the argument $x$; thus, we write $m = m(x, \beta)$, $v = v(x;\ \beta, \sigma_\varepsilon^2)$. Next,

$$B_{\mathrm{C}} = \mathbf{E}_\beta (t\ (y - m) + tm - H\beta)\,(t\ (y - m) + tm - H\beta)^{\mathrm{T}} = \\ = \mathbf{E}_\beta (y - m)^2 tt^{\mathrm{T}} + \mathbf{E}(t\mu^{\mathrm{T}} - H)\beta\beta^{\mathrm{T}}(t\mu^{\mathrm{T}} - H)^{\mathrm{T}}\,. \tag{3.81}$$

We used equality (3.55) and the following relation:

$$\mathbf{E}_\beta t\ (y - m)\,(tm - H\beta)^{\mathrm{T}} = \mathbf{E}\mathbf{E}_\beta [t\ (y - m)\ (tm - H\beta)^{\mathrm{T}} | x] = \\ = \mathbf{E}\left\{ t \cdot \mathbf{E}_\beta [(y - m)|x] \cdot (tm - H\beta)^{\mathrm{T}} \right\} = 0\,. \tag{3.82}$$

Finally,

$$\mathbf{E}_\beta (y - m)^2 tt^{\mathrm{T}} = \mathbf{E}\mathbf{E}_\beta [(y - m)^2 tt^{\mathrm{T}} | x] = \mathbf{E}\left\{ \mathbf{E}_\beta [(y - m)^2 | x] \cdot tt^{\mathrm{T}} \right\} = \mathbf{E} v tt^{\mathrm{T}}\,. \tag{3.83}$$

Therefore,

$$B_{\mathrm{C}} = \mathbf{E} v tt^{\mathrm{T}} + \mathbf{E}(t\mu^{\mathrm{T}} - H)\beta\beta^{\mathrm{T}}(t\mu^{\mathrm{T}} - H)^{\mathrm{T}}\,. \tag{3.84}$$

For an entry in zero row, we get

$$(t\mu^{\mathrm{T}} - H)_{i0} = t_i \mu_0 - t_{i+0} = t_i - t_i = 0\,, \tag{3.85}$$

because the second term in (3.84) does not depend on $\beta_0$. This fact allows us to represent the matrix (3.84) in the form (3.74).

By the vector analogue of Slutsky's lemma (see Corollary 2.21), we can move to the limit in distribution using the convergences (3.77) and (3.79):

$$\sqrt{n}(\hat{\beta}_{\mathrm{C}} - \beta) \xrightarrow{\mathrm{d}} A_{\mathrm{C}}^{-1}\eta \sim \mathrm{N}(0, A_{\mathrm{C}}^{-1} B_{\mathrm{C}} A_{\mathrm{C}}^{-\mathrm{T}})\,. \tag{3.86}$$

The convergence (3.72)–(3.75) is proved.

To prove nonsingularity of the matrix (3.73), it is enough to show that

$$B_{\mathrm{C}} > 0\,. \tag{3.87}$$

The second term in (3.84) is the covariance matrix of a random vector $tm - H\beta$, so this term is a positive semidefinite matrix. From here and from (3.56), we obtain

$$B_{\mathrm{C}} \geq \mathbf{E} v tt^{\mathrm{T}} \geq \sigma_\varepsilon^2 \mathbf{E} tt^{\mathrm{T}}\,. \tag{3.88}$$

The latter matrix is the Gram matrix for random variables $1, t_1(x), \ldots, t_k(x)$ in the space $L_2(\Omega, \mathbf{P})$ (see proof of Lemma 3.4). Show that they are linearly independent in this space. Suppose that we have $a_0 + a_1 t_1(x) + \cdots + a_k t_k(x) = 0$, a.s., for some real numbers $a_0, \ldots, a_k$. Then

$$0 = \mathbf{E}\left( \sum_{i=0}^{k} a_i t_i(x) \;\middle|\; \xi \right) = \sum_{i=0}^{k} a_i \xi^i\,. \tag{3.89}$$

But $\xi$ is a normal random variable and has a continuous distribution. Therefore, like in the proof of Lemma 3.4, it follows that $a_0 = a_1 = \cdots = a_k = 0$. This proves the linear independence of the random variables $1, t_1(x), \ldots, t_k(x)$. Then $\mathbf{E} tt^{\mathrm{T}} > 0$ and $B_{\mathrm{C}} > 0$ stems from (3.88). The theorem is proved. □

**Remark 3.8.** Equality (3.73) is the so-called *sandwich formula* for the estimating function (3.38) (we could prove Theorem 3.7 based on Theorem A.26 from Appendix A2, but instead we demonstrated a straightforward derivation of the formula). Here:

$$A_{\mathrm{C}} = A_{\mathrm{C}}^{\mathrm{T}} = -\mathbf{E}_{\beta}\frac{\partial s_{\mathrm{C}}(y, x; \beta)}{\partial b^{\mathrm{T}}}\,, \tag{3.90}$$

$$B_{\mathrm{C}} = \operatorname{cov}_{\beta} s_{\mathrm{C}}(x, y; \beta)\,. \tag{3.91}$$

We state Theorem 3.7 in greater detail for the linear model.

**Corollary 3.9.** *Let $k = 1$ and the conditions of Theorem 3.7 hold. Then* (3.72) *and* (3.73) *are satisfied, with*

$$A_{\mathrm{C}} = \begin{pmatrix} 1 & \mu_{\xi} \\ \mu_{\xi} & \mu_{\xi}^2 + \sigma_{\xi}^2 \end{pmatrix}, \tag{3.92}$$

$$B_{\mathrm{C}} = (\sigma_{\varepsilon}^2 + \tau^2\beta_1^2)\cdot\begin{pmatrix} 1 & \mu_{\xi} \\ \mu_{\xi} & \mu_{\xi}^2 + \sigma_{x}^2 \end{pmatrix} + \beta_1^2\mathbf{E}\begin{pmatrix} \mu_1 - t_1 \\ t_1\mu_1 - t_2 \end{pmatrix}\begin{pmatrix} \mu_1 - t_1 \\ t_1\mu_1 - t_2 \end{pmatrix}^{\mathrm{T}}. \tag{3.93}$$

*Proof.* Explain only equality (3.93). We have as a result of (3.68):

$$\mathbf{E}\upsilon tt^{\mathrm{T}} = \upsilon\mathbf{E}\begin{pmatrix} 1 & t_1 \\ t_1 & t_1^2 \end{pmatrix} = \upsilon\mathbf{E}\begin{pmatrix} 1 & x \\ x & x^2 \end{pmatrix} = (\sigma_{\varepsilon}^2 + \tau^2\beta_1^2)\cdot\begin{pmatrix} 1 & \mu_{\xi} \\ \mu_{\xi} & \mu_{\xi}^2 + \sigma_{x}^2 \end{pmatrix}. \tag{3.94}$$

Next, examine the second term in (3.74): $\beta_{-0} = \beta_1$,

$$t\mu_{-0}^{\mathrm{T}} - H_{-0} = \begin{pmatrix} 1 \\ t_1 \end{pmatrix}\mu_1 - \begin{pmatrix} H_{01} \\ H_{11} \end{pmatrix} = \begin{pmatrix} \mu_1 - t_1 \\ t_1\mu_1 - t_2 \end{pmatrix}, \tag{3.95}$$

$$(t\mu_{-0}^{\mathrm{T}} - H_{-0})\beta_{-0}\beta_{-0}^{\mathrm{T}}(t\mu_{-0}^{\mathrm{T}} - H_{-0})^{\mathrm{T}} = \beta_1^2\begin{pmatrix} \mu_1 - t_1 \\ t_1\mu_1 - t_2 \end{pmatrix}\begin{pmatrix} \mu_1 - t_1 \\ t_1\mu_1 - t_2 \end{pmatrix}^{\mathrm{T}}. \tag{3.96}$$

Now, equality (3.93) follows from formulas (3.74), (3.94), and (3.96). The corollary is proved. □

Note that this result generalizes Theorem 2.22.

### 3.1.5 Confidence ellipsoid for regression parameters

Under the conditions of Theorem 3.7, we will construct consistent estimators for matrices $A_{\mathrm{C}}$ and $B_{\mathrm{C}}$. Denote

$$\hat{A}_{\mathrm{C}} = \overline{H(x)}\,, \tag{3.97}$$

$$\hat{B}_{\mathrm{C}} = \frac{1}{n}\sum_{i=1}^{n}\left(t(x_i)y_i - H(x_i)\hat{\beta}_{\mathrm{C}}\right)\left(t(x_i)y_i - H(x_i)\hat{\beta}_{\mathrm{C}}\right)^{\mathrm{T}}. \tag{3.98}$$

Then $\hat{A}_{\mathrm{C}} \xrightarrow{\mathbf{P}1} A_{\mathrm{C}}$, since $\hat{A}_{\mathrm{C}}$ is a strongly consistent estimator of the matrix $A_{\mathrm{C}}$. In particular, $\hat{A}_{\mathrm{C}} > 0$, *eventually*. Further, by Theorem 3.5, $\hat{\beta}_{\mathrm{C}} \xrightarrow{\mathbf{P}1} \beta$, and therefore,

$$\begin{aligned} &\hat{B}_{\mathrm{C}} = \frac{1}{n}\sum_{i=1}^{n} (t(x_i)y_i - H(x_i)\beta)\,(t(x_i)y_i - H(x_i)\beta)^{\mathrm{T}} + r_n = B(n) + r_n\,, \\ &\|r_n\| \xrightarrow{\mathbf{P}1} 0\,. \end{aligned} \tag{3.99}$$

By the SLLN, as $n \to \infty$,

$$B(n) \xrightarrow{\mathbf{P}1} \mathbf{E}(t(x)y - H(x)\beta)\,(t(x)y - H(x)\beta)^{\mathrm{T}} = B_{\mathrm{C}}\,, \tag{3.100}$$

$$\hat{B}_{\mathrm{C}} \xrightarrow{\mathbf{P}1} B_{\mathrm{C}}\,. \tag{3.101}$$

Hence, $\hat{B}_{\mathrm{C}}$ is strongly consistent estimator of the matrix $B_{\mathrm{C}}$. From here

$$\hat{\Sigma}_{\mathrm{C}} = \hat{A}_{\mathrm{C}}^{-1}\hat{B}_{\mathrm{C}}\hat{A}_{\mathrm{C}}^{-\mathrm{T}} \xrightarrow{\mathbf{P}1} A_{\mathrm{C}}^{-1}B_{\mathrm{C}}A_{\mathrm{C}}^{-\mathrm{T}} = \Sigma_{\mathrm{C}}\,, \tag{3.102}$$

and the matrix $\hat{\Sigma}_{\mathrm{C}}$ is *eventually* positive definite.

Convergences (3.72) and (3.102) imply that

$$\sqrt{n}(\hat{\Sigma}_{\mathrm{C}})^{-1/2}\cdot(\hat{\beta}_{\mathrm{C}} - \beta) \xrightarrow{\mathrm{d}} \mathrm{N}(0, I_{k+1})\,. \tag{3.103}$$

If $\hat{\Sigma}_{\mathrm{C}} > 0$, then $(\hat{\Sigma}_{\mathrm{C}})^{-1/2} = (\sqrt{\hat{\Sigma}_{\mathrm{C}}})^{-1}$, where $\sqrt{\hat{\Sigma}_{\mathrm{C}}}$ is the only positive definite matrix, with $(\sqrt{\hat{\Sigma}_{\mathrm{C}}})^2 = \hat{\Sigma}_{\mathrm{C}}$. Further, by the convergence (3.103) we have

$$\left\|\sqrt{n}(\hat{\Sigma}_{\mathrm{C}})^{-1/2}(\hat{\beta}_{\mathrm{C}} - \beta)\right\|^2 = n(\hat{\beta}_{\mathrm{C}} - \beta)^{\mathrm{T}}(\hat{\Sigma}_{\mathrm{C}})^{-1}(\hat{\beta}_{\mathrm{C}} - \beta) \xrightarrow{\mathrm{d}} \chi^2_{k+1}\,. \tag{3.104}$$

Here $\chi^2_{k+1}$ is $\chi^2$ distribution with $k+1$ degrees of freedom.

Fix the confidence probability $1-\alpha$ (e.g., 0.95). For the asymptotic confidence ellipsoid for $\beta$, we take the random set

$$E_n = \left\{z \in \mathbf{R}^{k+1} \colon (z - \hat{\beta}_{\mathrm{C}})^{\mathrm{T}}(\hat{\Sigma}_{\mathrm{C}})^{-1}(z - \hat{\beta}_{\mathrm{C}}) \le \frac{1}{n}(\chi^2_{k+1})_\alpha\right\}\,. \tag{3.105}$$

Here $(\chi^2_{k+1})_\alpha$ is the quantile of the $\chi^2_{k+1}$ distribution, with

$$\mathbf{P}\{\chi^2_{k+1} > (\chi^2_{k+1})_\alpha\} = \alpha\,. \tag{3.106}$$

We construct the set $E_n$ only in the case $\hat{\Sigma}_{\mathrm{C}} > 0$ (this *eventually* takes place). Then, as $n \to \infty$, we get

$$\begin{aligned} &\mathbf{P}\{\beta \in E_n\} = \\ &\quad = \mathbf{P}\{n(\beta - \hat{\beta}_{\mathrm{C}})^{\mathrm{T}}(\hat{\Sigma}_{\mathrm{C}})^{-1}(\beta - \hat{\beta}_{\mathrm{C}}) \le (\chi^2_{k+1})_\alpha\} \to \mathbf{P}\{\chi^2_{k+1} \le (\chi^2_{k+1})_\alpha\} = 1-\alpha\,. \end{aligned} \tag{3.107}$$

The convergence (3.107) means that $E_n$ is the asymptotic confidence ellipsoid for $\beta$ with confidence probability $1-\alpha$.

### 3.1.6 Estimator for variance of error in response

We will treat the model (3.6) and (3.3) as a multiple model of regression of the response $y$ on the vector regressor $\rho$, where $\rho = \rho(\xi)$ is given in (3.6):

$$y = \rho^{\mathrm{T}}\beta + \varepsilon\,, \quad t = \rho + e\,. \tag{3.108}$$

Here the vector $t = t(x)$ has the components (3.36) satisfying the deconvolution equation (3.16). Assume the conditions (i), (iii), (v), and (vii). In the model (3.108), the observed values are $y_i$ and $t_i = t(x_i)$ (denote also $\rho_i = \rho(\xi_i)$ and $e_i = t_i - \rho_i$), such that

$$y_i = \rho_i^{\mathrm{T}}\beta + \varepsilon_i\,, \quad t_i = \rho_i + e_i\,, \quad i = 1, \dots, n\,. \tag{3.109}$$

The observations (3.109) are independent copies of the model (3.108).

The "error" $e$ is not stochastically independent of $\rho = \rho(\xi)$, but

$$\mathbf{E}(e|\rho) = \mathbf{E}(t - \rho|\rho) = \mathbf{E}(t|\rho) - \rho = \mathbf{E}(t|\xi) - \rho = 0\,. \tag{3.110}$$

Thus, the "error" $e$ is conditionally centered given $\rho$. This fact shows that the vector $e$ can be viewed analogous to an additive error.

Further, in the model (3.109), we construct the estimator of $\sigma_\varepsilon^2$ by the *method of moments*. In so doing, it is important that by Theorem 3.5, we have the consistent estimator $\hat{\beta}_{\mathrm{C}}$ of the parameter $\beta$. Find the second moments in the model (3.108):

$$\mathbf{E}y^2 = \beta^{\mathrm{T}}(\mathbf{E}\rho\rho^{\mathrm{T}})\beta + \sigma_\varepsilon^2\,, \tag{3.111}$$

$$\mathbf{E}ty = (\mathbf{E}t\rho^{\mathrm{T}})\beta = \mathbf{E}\mathbf{E}(t\rho^{\mathrm{T}}|\xi)\cdot\beta = \mathbf{E}\left[\mathbf{E}(t|\xi)\,\rho^{\mathrm{T}}\right]\beta = (\mathbf{E}\rho\rho^{\mathrm{T}})\beta\,. \tag{3.112}$$

From here the unknown matrix $\mathbf{E}\rho\rho^{\mathrm{T}}$ is excluded, and finally

$$\sigma_\varepsilon^2 = \mathbf{E}y^2 - \beta^{\mathrm{T}}\mathbf{E}ty\,. \tag{3.113}$$

As an estimator, we take

$$\hat{\sigma}_\varepsilon^2 = \overline{y^2} - \hat{\beta}_{\mathrm{C}}^{\mathrm{T}}\overline{ty}\,. \tag{3.114}$$

**Theorem 3.10.** *Assume the conditions* (i), (iii), (v), *and* (vii). *The estimator* (3.114) *is strongly consistent for the parameter* $\sigma_\varepsilon^2$.

*Proof.* By the SLLN and Theorem 3.5, we get

$$\hat{\sigma}_\varepsilon^2 \xrightarrow{\mathbf{P1}} \mathbf{E}y^2 - \beta^{\mathrm{T}}\mathbf{E}ty = \sigma_\varepsilon^2\,. \tag{3.115}$$

This proves the theorem. □

Find the ACM of the estimator (3.114). Denote $\theta = \binom{\beta}{\sigma_\varepsilon^2}$ to be the augmented vector of parameters to be estimated. *Eventually*, the estimator $\hat{\theta}_{\mathrm{C}} = \binom{\hat{\beta}_{\mathrm{C}}}{\hat{\sigma}_\varepsilon^2}$ is a solution to the system of equations $\frac{1}{n}\sum_{i=1}^n s_{\mathrm{C}}^{(\theta)}(y_i, x_i;\ \theta) = 0$, with

$$s_{\mathrm{C}}^{(\theta)} = \begin{pmatrix} s_{\mathrm{C}}^{(\beta)} \\ s_{\mathrm{C}}^{(\sigma_\varepsilon^2)} \end{pmatrix}\,, \quad s_{\mathrm{C}}^{(\beta)} = t(x)y - H(x)\beta\,, \tag{3.116}$$

$$s_{\mathrm{C}}^{(\sigma_\varepsilon^2)} = y^2 - \beta^{\mathrm{T}}t(x)y - \sigma_\varepsilon^2\,. \tag{3.117}$$

**Theorem 3.11.** *Assume the conditions* (i), (iii), (iv), *and also the following condition:*
(viii) *The errors $\varepsilon_i$ are identically distributed and centered with $\mathbf{E}\varepsilon_1^4 < \infty$, and moreover, the distribution of $\varepsilon_1$ is not concentrated at two or fewer points.*
*Then*

$$\sqrt{n}(\hat{\theta}_{\mathrm{C}} - \theta) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \Sigma_{\mathrm{C}}^{(\theta)}) , \tag{3.118}$$

*where the matrix $\Sigma_{\mathrm{C}}^{(\theta)}$ is nonsingular, and*

$$\Sigma_{\mathrm{C}}^{(\theta)} = \begin{pmatrix} \Sigma_{\mathrm{C}}^{(\beta)} & \Sigma_{\mathrm{C}}^{(\beta,\sigma_\varepsilon^2)} \\ \Sigma_{\mathrm{C}}^{(\sigma_\varepsilon^2,\beta)} & \Sigma_{\mathrm{C}}^{(\sigma_\varepsilon^2)} \end{pmatrix} , \tag{3.119}$$

*where $\Sigma_{\mathrm{C}}^{(\beta)}$ is ACM of the estimator $\hat{\beta}_{\mathrm{C}}$ given in* (3.73) *and* (3.74), *and $\Sigma_{\mathrm{C}}^{(\beta)}$ is asymptotic variance of the estimator of $\sigma_\varepsilon^2$,*

$$\Sigma_{\mathrm{C}}^{(\sigma_\varepsilon^2)} = \mathrm{cov}_\theta\, s_{\mathrm{C}}^{(\sigma_\varepsilon^2)}(y, x;\ \theta) . \tag{3.120}$$

*Proof.* (1) By Theorems 3.5 and 3.10, the estimator $\hat{\theta}_{\mathrm{C}}$ is strongly consistent, i.e., $\hat{\theta}_{\mathrm{C}} \xrightarrow{\mathbf{P}1} \theta$. The convergence (3.118) follows from the sandwich formula (see Appendix A2), and

$$\Sigma_{\mathrm{C}}^{(\theta)} = A_\theta^{-1} B_\theta A_\theta^{-1} , \tag{3.121}$$

$$A_\theta = -\mathbf{E}\frac{\partial s_{\mathrm{C}}^{(\theta)}(y, x;\ \theta)}{\partial \theta^{\mathrm{T}}} = \begin{pmatrix} \mathbf{E}H & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{E}\rho\rho^{\mathrm{T}} & 0 \\ 0 & 1 \end{pmatrix} > 0 ; \tag{3.122}$$

$$B_\theta = \mathrm{cov}_\theta\, s_{\mathrm{C}}^{(\theta)}(y, x;\ \theta) . \tag{3.123}$$

Here the condition $\mathbf{E}\varepsilon^4 < \infty$ and the normality of $\xi$ and $\delta$ provide finiteness of second moments for the estimating function $s_{\mathrm{C}}^{(\theta)}(y, x;\ \theta)$.

(2) In order to prove that $B_\theta$ is nonsingular, it is enough to prove the linear independence of the components of $s_{\mathrm{C}}^{(\theta)}(y, x;\ \theta)$ in the space $L_2(\Omega, \mathbf{P})$ of random variables. For this purpose, we make a linear combination of components of this estimating function $s_{\mathrm{C}}^{(\theta)}(y, x;\ \theta)$ at the true point $\theta$, and let the combination be equal to 0, a.s.:

$$a^{\mathrm{T}} s_{\mathrm{C}}^{(\beta)} + b s_{\mathrm{C}}^{(\sigma_\varepsilon^2)} = a^{\mathrm{T}}(t(x)y - H(x)\beta) + b(y^2 - \beta^{\mathrm{T}} t(x) y - \sigma_\varepsilon^2) = 0 . \tag{3.124}$$

Here $a \in \mathbf{R}^{k+1}$ and $b \in \mathbf{R}$. Thus, we obtain, for nonrandom $c_1$ and $c_2$:

$$0 = \mathbf{E}[a^{\mathrm{T}} s_{\mathrm{C}}^{(\beta)} + b s_{\mathrm{C}}^{(\sigma_\varepsilon^2)} | \varepsilon] = b\varepsilon^2 + c_1\varepsilon + c_2 , \tag{3.125}$$

and by the condition (viii) about the distribution of $\varepsilon$, it follows that $b = 0$. Further, proving Theorem 3.7, we established that the matrix (3.80) is nonsingular, and therefore, the components of $s_{\mathrm{C}}^{(\beta)}(y, x;\ \beta)$ are linearly independent in $L_2(\Omega, \mathbf{P})$. Then by equation (3.124) at $b = 0$, it is deduced that $a = 0$. Thus, the components of $s_{\mathrm{C}}^{(\theta)}(y, x;\ \theta)$ are linearly independent, and $B_\theta > 0$. This proves nonsingularity of the matrix $\Sigma_{\mathrm{C}}^{(\theta)}$ from formula (3.121).

(3) It remains to prove (3.120). Let

$$B_\theta = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}, \quad B_{22} = \operatorname{cov}_\theta s_{\mathrm{C}}^{(\sigma_\varepsilon^2)}(y, x;\ \theta)\,. \tag{3.126}$$

Then

$$\Sigma_{\mathrm{C}}^{(\theta)} = A_\theta^{-1} B_\theta A_\theta^{-\mathrm{T}} = \begin{pmatrix} (\mathbf{E}H)^{-1} & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} \begin{pmatrix} (\mathbf{E}H)^{-1} & 0 \\ 0 & 1 \end{pmatrix} =$$
$$= \begin{pmatrix} (\mathbf{E}H)^{-1} B_{11} (\mathbf{E}H)^{-1} & (\mathbf{E}H)^{-1} B_{12} \\ B_{21} (\mathbf{E}H)^{-1} & B_{22} \end{pmatrix}. \tag{3.127}$$

From here and the convergence (3.118), it follows that

$$\sqrt{n}(\hat\sigma_\varepsilon^2 - \sigma_\varepsilon^2) \xrightarrow{\mathrm{d}} \mathrm{N}(0, B_{22}) = \mathrm{N}(0, \Sigma_{\mathrm{C}}^{(\sigma_\varepsilon^2)})\,. \tag{3.128}$$

Taking into account (3.126), we obtain the desired relation (3.120). The theorem is proved. □

Based on the theorem one can, like in Section 3.1.5, construct the confidence ellipsoid for $\theta$. Now, we indicate only how to construct the asymptotic confidence interval for $\sigma_\varepsilon^2$.

A strongly consistent estimator for the positive number $B_{22}$ is the statistic

$$\hat B_{22} = \frac{1}{n} \sum_{i=1}^{n} (y_i^2 - \hat\beta_{\mathrm{C}}^{\mathrm{T}} t(x_i) y_i - \hat\sigma_\varepsilon^2)^2\,, \tag{3.129}$$

which is positive, *eventually*. Then from (3.128) it follows that

$$\sqrt{\frac{n}{\hat B_{22}}}\,(\hat\sigma_\varepsilon^2 - \sigma_\varepsilon^2) \xrightarrow{\mathrm{d}} \mathrm{N}(0, 1)\,. \tag{3.130}$$

When the confidence probability is equal to $1 - \alpha$, the asymptotic confidence interval for parameter $\sigma_\varepsilon^2$ is constructed in the form (in the case $\hat B_{22} > 0$):

$$I_n = \left\{ z > 0 \colon\ |z - \hat\sigma_\varepsilon^2| \le \sqrt{\frac{\hat B_{22}}{n}} \cdot n_{\alpha/2} \right\}. \tag{3.131}$$

Here $n_{\alpha/2}$ is a quantile of normal distribution, with

$$\mathbf{P}\{N(0, 1) > n_{\alpha/2}\} = \frac{\alpha}{2}\,. \tag{3.132}$$

### 3.1.7 Modifications of the ALS estimator

(1) Remember that the ALS estimator $\hat\beta_{\mathrm{C}}$ is unstable for a small sample, see Section 1.4.7. In Cheng et al. (2000) for polynomial models, a modified estimator $\hat\beta_{\mathrm{C}}^{\mathrm{M}}$ is

proposed, which is also consistent and $\sqrt{n}\,(\hat{\beta}_{\mathrm{C}}^{\mathrm{M}} - \hat{\beta}_{\mathrm{C}}) \xrightarrow{\mathbf{P}} 0$. Such two estimators, $\hat{\beta}_{\mathrm{C}}$ and $\hat{\beta}_{\mathrm{C}}^{\mathrm{M}}$, are called *asymptotically equivalent*. In view of Theorem 3.7, we will have

$$\sqrt{n}\,(\hat{\beta}_{\mathrm{C}}^{\mathrm{M}} - \beta) = \sqrt{n}\,(\hat{\beta}_{\mathrm{C}} - \beta) + \sqrt{n}\,(\hat{\beta}_{\mathrm{C}}^{\mathrm{M}} - \hat{\beta}_{\mathrm{C}}) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \Sigma_{\mathrm{C}})\,. \tag{3.133}$$

Slutsky's lemma 2.18 is exploited here. Obviously, the estimator $\hat{\beta}_{\mathrm{C}}^{\mathrm{M}}$ has the same ACM as the estimator $\hat{\beta}_{\mathrm{C}}$. Using $\hat{\beta}_{\mathrm{C}}^{\mathrm{M}}$ instead of the $\hat{\beta}_{\mathrm{C}}$, we do not lose accuracy of estimation (especially for large samples) and for small and moderate samples, get more stable numerical procedures.

(2) The estimator $\hat{\beta}_{\mathrm{C}}$ (and also the $\hat{\beta}_{\mathrm{C}}^{\mathrm{M}}$) does not use the information on the form of distribution of $\xi$. Therefore, this estimator can be well exploited in the functional polynomial model, where the latent variables $\xi_i$ are nonrandom. The estimator remains consistent and asymptotically normal under mild conditions.

(3) One can abandon the normality of the measurement error $\delta$. Instead, for the construction of the ALS estimator, it is necessary that $\mathbf{E}\delta = 0$, $\mathbf{E}\delta^{2k} < \infty$, and the following moments are known:

$$m_{\delta}^{(i)} = \mathbf{E}\delta^{i}\,, \quad 2 \le i \le 2k\,. \tag{3.134}$$

Then one can construct reduced polynomials $t_j(x)$ of degree $j$ which are solutions to the deconvolution problem (3.16), with $j \le 2k$. This allows us to construct the ALS estimator of $\beta$, see Cheng and Schneeweiss (1998).

## 3.2 Quasi-likelihood estimator

In the polynomial model (3.6) and (3.3) under the conditions (i), (iii), (iv), and (vii), the conditional expectation $m(x, \beta)$ and conditional variance $v(x; \beta, \sigma_\varepsilon^2)$ of the response are written down in Section 3.1.3. This allows to construct the estimating function (1.70) for the QLE $\hat{\beta}_{\mathrm{QL}}$ (the estimating function depends on nuisance parameters $\sigma_\varepsilon^2$, $\mu_\xi$, and $\sigma_\xi^2$):

$$s_{\mathrm{QL}}^{(\beta)}(y, x; \beta, \sigma_\varepsilon^2) = \frac{\mu(x)(y - \mu^{\mathrm{T}}(x)\beta)}{v(x; \beta, \sigma_\varepsilon^2)}\,. \tag{3.135}$$

If the nuisance parameters are known, the estimator $\hat{\beta}_{\mathrm{QL}}$ is defined as a solution to the equation

$$\frac{1}{n}\sum_{i=1}^{n} \frac{1}{v(x_i; \beta, \sigma_\varepsilon^2)}(\mu(x_i)y_i - \mu(x_i)\mu^{\mathrm{T}}(x_i)\beta) = 0\,, \quad \beta \in \mathbf{R}^{k+1}\,. \tag{3.136}$$

This nonlinear equation not always has a solution. Define the estimator more accurately.

**Definition 3.12.** The estimator $\hat{\beta}_{\mathrm{QL}}$ is a Borel measurable function of observations $y_1$, $x_1, \ldots, y_n, x_n$, for which:

(a) if equation (3.136) has no solution, then $\hat{\beta}_{\mathrm{QL}} = 0$;
(b) if there exists a solution to equation (3.136), then $\hat{\beta}_{\mathrm{QL}}$ is a solution with minimal norm (if there are several such solutions, then we take any of them).

Note that the definition is correct. Indeed, because of continuity in $\beta$ of the left-hand side of (3.136), the set $A_\omega$ of solutions to the equation is closed for each elementary event $\omega$, and in the case $A_\omega \neq \emptyset$, $\min_{\beta \in A_\omega} \|\beta\|$ is attained.

### 3.2.1 The case of known nuisance parameters

#### Asymptotic properties of the estimator

In the following we consider only the normal polynomial model.

**Theorem 3.13.** *Let the nuisance parameters $\sigma_\varepsilon^2$, $\mu_\xi$, and $\sigma_\xi^2$ be known and the conditions* (i)–(iv) *hold true in the model* (3.6) *and* (3.3). *Denote by $b = (b_i)_{i=0}^k$ the true values of the polynomial coefficients. Then the following statements hold.*

(a) *For any $R > \|b\|$, equation* (3.136) *has a unique solution in the ball*

$$\bar{B}_R = \{\beta \in \mathbf{R}^{k+1} \colon \ \|\beta\| \ \leq R\}\,, \tag{3.137}$$

(b)

$$\hat{\beta}_{\mathrm{QL}} \ \xrightarrow{\mathbf{P}1} \ b\,, \quad \text{as } \ n \to \infty\,, \tag{3.138}$$

(c)

$$\sqrt{n}(\hat{\beta}_{\mathrm{QL}} - b) \ \xrightarrow{\mathrm{d}} \ \mathrm{N}(0, B_b)\,, \tag{3.139}$$

$$B_b = \left(\mathbf{E}\frac{\mu(x)\mu^{\mathrm{T}}(x)}{v(x,b)}\right)^{-1}\,. \tag{3.140}$$

*Proof.* The statements (a) and (b) follow from the theory of estimating equations (see Appendix A1). We verify only the basic condition about the uniqueness of solution to the limit equation.

On the set (3.137), the left-hand side of equation (3.136) converges a.s. uniformly to the function

$$\begin{aligned}
s_\infty(\beta; b) &= \mathbf{E}_b \frac{\mu(x)y - \mu(x)\mu^{\mathrm{T}}(x)\beta}{v(x,\beta)}\,, \quad \beta \in \bar{B}_R\,; \\
s_\infty(\beta, b) &= \mathbf{E}\mathbf{E}_b\left(\frac{\mu(x)y}{v(x,\beta)} \,\middle|\, x\right) - \mathbf{E}\frac{\mu(x)\mu^{\mathrm{T}}(x)}{v(x,\beta)} \cdot \beta = \\
&= \mathbf{E}\frac{\mu(x)\mu^{\mathrm{T}}(x)\beta}{v(x,\beta)} - \mathbf{E}\frac{\mu(x)\mu^{\mathrm{T}}(x)}{v(x,\beta)}\beta\,, \qquad (3.141) \\
s_\infty(\beta, b) &= \mathbf{E}\frac{\mu(x)\mu^{\mathrm{T}}(x)}{v(x,\beta)} \cdot (b - \beta)\,. \qquad (3.142)
\end{aligned}$$

The limit equation is

$$\mathbf{E}\frac{\mu(x)\mu^{\mathrm{T}}(x)}{v(x,\beta)}\cdot(b-\beta)=0\,,\qquad \beta\in\overline{B}_R\,. \tag{3.143}$$

Make sure that for each $\beta\in\mathbf{R}^{k+1}$, the matrix

$$\Phi_\beta=\mathbf{E}\frac{\mu(x)\mu^{\mathrm{T}}(x)}{v(x,\beta)} \tag{3.144}$$

is nonsingular. In fact, this is the Gram matrix of random variables

$$\frac{\mu_0(x)}{\sqrt{v(x,\beta)}}\,,\quad \frac{\mu_1(x)}{\sqrt{v(x,\beta)}}\,,\quad \dots\,,\quad \frac{\mu_k(x)}{\sqrt{v(x,\beta)}} \tag{3.145}$$

in the space $L_2(\Omega,\mathbf{P})$. Let $\{a_i,\,i=\overline{0,k}\}$ be real numbers such that almost surely

$$\sum_{i=0}^{k}\frac{a_i\mu_i(x)}{\sqrt{v(x,\beta)}}=0\,. \tag{3.146}$$

Then

$$\sum_{i=0}^{k}a_i\mu_i(x)=0\,,\quad \text{a.s.} \tag{3.147}$$

Since $x$ has normal distribution and $\mu_i(x)$ is a polynomial of $i$th degree in $x$, it follows that $a_0=a_1=\dots=a_k=0$ (see proof of Lemma 3.4). Thus, the random variables (3.145) are linearly independent, and the matrix (3.144) is positive definite. Then $\Phi_\beta$ has zero kernel, and equation (3.143) has a unique solution $\beta=b\in\bar{B}_R$. We have just proved the validity of statements (a) and (b).

The statement (c) follows from the sandwich formula (see Appendix A2). For the ACM $\Sigma_b$ of the estimator $\hat{\beta}_{\mathrm{QL}}$, we have

$$\Sigma_b=A_{\mathrm{Q}}^{-1}B_{\mathrm{Q}}A_{\mathrm{Q}}^{-\mathrm{T}}\,, \tag{3.148}$$

$$A_{\mathrm{Q}}=-\mathbf{E}_b\frac{\partial s_{\mathrm{QL}}^{(\beta)}}{\partial\beta^{\mathrm{T}}}(y,x;b)=\mathbf{E}\frac{\mu(x)\mu^{\mathrm{T}}(x)}{v(x,b)}-\mathbf{E}_b\mu(x)(y-\mu^{\mathrm{T}}(x)b)\frac{\partial v^{-1}}{\partial\beta^{\mathrm{T}}}(x,b)\,. \tag{3.149}$$

The latter expectation is equal to

$$\begin{aligned}\mathbf{E}\mathbf{E}_b\left(\mu(x)(y-\mu^{\mathrm{T}}(x)b)\frac{\partial v^{-1}}{\partial\beta^{\mathrm{T}}}(x,b)\;\middle|\;x\right)=\\ =\mathbf{E}\left\{\mu(x)\cdot\mathbf{E}_b(y-\mu^{\mathrm{T}}(x)b|x)\cdot\frac{\partial v^{-1}}{\partial\beta^{\mathrm{T}}}(x,b)\right\}=0\,,\end{aligned} \tag{3.150}$$

because here the conditional expectation is zero: $\mathbf{E}_b(y-\mu^{\mathrm{T}}(x)b|x)=0$. Then

$$A_{\mathrm{Q}}=B_b^{-1}=\mathbf{E}\frac{\mu(x)\mu^{\mathrm{T}}(x)}{v(x,b)}\,. \tag{3.151}$$

For a middle part of the "sandwich" (3.148) we get

$$B_{\mathrm{Q}} = \operatorname{cov}_b s_{\mathrm{QL}}^{(\beta)}(y, x; b) = \mathbf{E}_b \frac{\mu(x)(y - \mu^{\mathrm{T}}(x)b)^2 \mu^{\mathrm{T}}(x)}{v^2(x, b)}\,, \tag{3.152}$$

$$B_{\mathrm{Q}} = \mathbf{E}\left\{\frac{\mu(x)\cdot \mathbf{E}_b[(y - \mu^{\mathrm{T}}(x)b)^2 | x]\cdot \mu^{\mathrm{T}}(x)}{v^2(x, b)}\right\} = \mathbf{E}\frac{(\mu(x)\mu^{\mathrm{T}}(x))v(x, b)}{v^2(x, b)} = B_b^{-1}\,. \tag{3.153}$$

Then $\Sigma_b = B_b B_b^{-1} B_b = B_b$, and the theorem is proved. □

**Remark 3.14.** According to the recommendations in Cheng and Van Ness (1999) and Wansbeek and Meijer (2000), as a strongly consistent estimator for the matrix $\Sigma_b$ one can take

$$\hat{\Sigma}_b = \hat{A}_{\mathrm{Q}}^{-1} \hat{B}_{\mathrm{Q}} \hat{A}_{\mathrm{Q}}^{-\mathrm{T}}\,, \tag{3.154}$$

$$\hat{A}_{\mathrm{Q}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\mu(x_i)\mu^{\mathrm{T}}(x_i)}{v(x_i, \hat{\beta}_{\mathrm{QL}})}\,, \tag{3.155}$$

$$\hat{B}_{\mathrm{Q}} = \frac{1}{n}\sum_{i=1}^{n} \frac{(y_i - \mu^{\mathrm{T}}(x_i)\hat{\beta}_{\mathrm{QL}})^2}{v^2(x_i, \hat{\beta}_{\mathrm{QL}})}\mu(x_i)\mu^{\mathrm{T}}(x_i)\,. \tag{3.156}$$

The convergence $\hat{\Sigma}_b \xrightarrow{\mathbf{P}1} \Sigma_b$ stems from the strong consistency of the estimator $\hat{\beta}_{\mathrm{QL}}$ and from the SLLN. Now, like in Section 3.1.5, one can construct the confidence ellipsoid for the vector $b$ based on the estimator $\hat{\beta}_{\mathrm{QL}}$.

### Calculation technique for the estimate

For numerical solution of the estimating equation (3.136), the iteratively reweighted least squares method is used. To describe the essence of the method, let us rewrite (3.136) in another form:

$$\beta = \phi_n(\beta)\,, \quad \phi_n(\beta) := \left(\frac{1}{n}\sum_{i=1}^{n} \frac{\mu(x_i)\mu^{\mathrm{T}}(x_i)}{v(x_i, \beta)}\right)^{-1} \cdot \frac{1}{n}\sum_{i=1}^{n} \frac{\mu(x_i)y_i}{v(x_i, \beta)}\,. \tag{3.157}$$

Actually in equation (3.136), the denominators $v(x_i, \beta)$ are fixed and the equation is solved as linear in $\beta$.

Now, we describe the above-mentioned iterative algorithm in more detail.

(1) The initial value $\beta_n^{(0)} = \beta^{(0)}$ is taken arbitrary (e.g., $\beta^{(0)} = 0$).

(2) Given $\beta^{(j)}$ from the $j$th iteration of the algorithm, we find

$$\beta^{(j+1)} = \phi_n(\beta^{(j)})\,. \tag{3.158}$$

The algorithm leads to the desired solution $\hat{\beta}_{\mathrm{QL}}$ to equation (3.136). It turns out that for large $n$, all the values $\phi_n(\beta^{(j)})$, $j \geq 0$, are well-defined, despite the need to invert some matrix in (3.158).

**Theorem 3.15.** *Assume the conditions* (i)–(iv). *Suppose that the initial values of the algorithm* $\beta_n^{(0)}$ *are random, and moreover for a real number* $R > 0$,

$$\left\|\beta_n^{(0)}\right\| < R\,, \quad \textit{eventually}\,. \tag{3.159}$$

*Then*

$$\lim_{j\to\infty} \beta^{(j)} = \hat{\beta}_{\mathrm{QL}}\,, \quad \textit{eventually}\,. \tag{3.160}$$

**Remark 3.16.** As the initial value $\beta^{(0)}$, one can take the estimator $\hat{\beta}_{\mathrm{C}}$ defined in (3.40). Since it is strongly consistent, then the condition (3.159) is satisfied.

*Sketch of the proof of Theorem* 3.15. (1) Let $b$ be the true value of the parameter $\beta$. We will consider a ball $\bar{B}_R$ given in (3.137), where $R > \|b\|$ and $R$ satisfies (3.159).

(2) For a sequence of matrices

$$A_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} \frac{\mu(x_i)\mu^{\mathrm{T}}(x_i)}{v(x_i,\beta)}\,, \quad \beta \in \mathbf{R}^{k+1}\,, \tag{3.161}$$

the following is carried out: with probability 1, the $A_n(\beta)$ converges uniformly in $\beta \in \bar{B}_R$ to a positive definite matrix

$$B_Q(\beta) = \mathbf{E}\frac{\mu(x)\mu^{\mathrm{T}}(x)}{v(x,\beta)}\,. \tag{3.162}$$

So *eventually* the matrix $A_n(\beta)$ is positive definite simultaneously for all $\beta \in \bar{B}_R$, and *eventually* the function $\phi_n(\beta)$ is well-defined on $\bar{B}_R$.

(3) Using an expansion $y_i = \rho^{\mathrm{T}}(\xi_i)b + \varepsilon_i$ and formula (3.157), we can verify directly that

$$\lim_{n\to\infty} \sup_{\beta\in\bar{B}_R} \left\|\frac{\partial\phi_n(\beta)}{\partial\beta^{\mathrm{T}}}\right\| = 0\,, \quad \text{a.s.} \tag{3.163}$$

Denote

$$\lambda_n = \sup_{\beta\in\bar{B}_R} \left\|\frac{\partial\phi_n(\beta)}{\partial\beta^{\mathrm{T}}}\right\|\,; \tag{3.164}$$

$$\lambda_n \xrightarrow{\mathbf{P}1} 0\,. \tag{3.165}$$

Then *eventually* for all $\beta_1, \beta_2 \in \bar{B}_R$,

$$\left\|\phi_n(\beta_1) - \phi_n(\beta_2)\right\| \le \lambda_n \cdot \left\|\beta_1 - \beta_2\right\|\,. \tag{3.166}$$

(4) By Theorem 3.13, equation (3.157) has *eventually* a unique solution $\hat{\beta}_{\mathrm{QL}}$ on the ball $\bar{B}_R$. By the same theorem, $\hat{\beta}_{\mathrm{QL}} \xrightarrow{\mathbf{P}1} b$, that is why for certain ball

$$\bar{B}(b,\varepsilon) = \{\beta \in \mathbf{R}^{k+1}\colon\ \|\beta - b\| \le \varepsilon\} \subset \bar{B}_R\,, \tag{3.167}$$

it holds true that $\hat{\beta}_{\mathrm{QL}} \in \bar{B}(b,\varepsilon)$, *eventually*. We have *eventually* that for all $\beta \in \bar{B}_R$,

$$\left\|\phi_n(\beta) - \hat{\beta}_{\mathrm{QL}}\right\| = \left\|\phi_n(\beta) - \phi_n(\hat{\beta}_{\mathrm{QL}})\right\| \le \lambda_n \left\|\beta - \hat{\beta}_{\mathrm{QL}}\right\| \le \varepsilon \left\|\beta - \hat{\beta}_{\mathrm{QL}}\right\|\,. \tag{3.168}$$

Thus *eventually*,

$$\phi_n \colon \bar{B}_R \to \bar{B}_R\,, \quad \phi_n(\bar{B}_R) \subset \bar{B}_R\,. \tag{3.169}$$

(5) From relations (3.165) and (3.166), it follows that $\lambda_n < \frac{1}{2}$ *eventually*, and then *eventually* for all $\beta_1, \beta_2 \in \bar{B}_R$,

$$\|\phi_n(\beta_1) - \phi_n(\beta_2)\| \;\le\; \frac{1}{2}\,\|\beta_1 - \beta_2\|\,. \tag{3.170}$$

The relations (3.169) and (3.170) mean that *eventually* $\phi_n$ is a contraction mapping on $\bar{B}_R$. Then by the Banach contraction principle, the convergence of iterations (3.160) is *eventually* realized. Here, $\hat{\beta}_{\mathrm{QL}}$ is a fixed point of the mapping (3.169). □

**Remark 3.17.** Instead of condition (3.159), one can impose a somewhat weaker condition

$$\sup_{n\ge k+1} \left\|\beta_n^{(0)}\right\| < \infty\,, \quad \text{a.s.} \tag{3.171}$$

Then convergence (3.160) remains true.

**Asymptotic optimality of the estimator**

We study two competitive estimators $\hat{\beta}_{\mathrm{C}}$ and $\hat{\beta}_{\mathrm{QL}}$. Which one is more efficient? In classical regression theory, estimators are compared by variance. However, this is not appropriate in errors-in-variables models, because in such models, reasonable estimators do not possess finite second moment (see Theorem 2.26). Instead, estimators can be compared by their ACM.

Remember that for symmetric matrices $A$ and $B$ of the same size, notation $A \ge B$ means that $A - B$ is a positive semidefinite matrix, and notation $A > B$ means that $A - B$ is a positive definite matrix. The partial order $A \ge B$ is the so-called Loewner order in the space of symmetric matrices of fixed size.

Suppose we have two estimators $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(2)}$ of the parameter $\beta$ constructed by the same sample, moreover both are asymptotically normal:

$$\sqrt{n}\,(\hat{\beta}^{(i)} - \beta) \;\xrightarrow{\mathrm{d}}\; \mathrm{N}(0, \Sigma_i)\,, \quad i = 1, 2\,. \tag{3.172}$$

The matrix $\Sigma_i = \Sigma_i(\beta)$ is ACM of the estimator $\hat{\beta}^{(i)}$. From (3.172) the consistency of estimators follows: $\hat{\beta}^{(i)} \xrightarrow{\mathbf{P}} \beta$, $i = 1, 2$.

**Definition 3.18.** The estimator $\hat{\beta}^{(1)}$ is *asymptotically more efficient* than $\hat{\beta}^{(2)}$, if for each true value of $\beta$ taken from a parameter set, it holds that $\Sigma_1(\beta) \le \Sigma_2(\beta)$. If the inequality is always strict (in terms of Loewner order), then the estimator $\hat{\beta}^{(1)}$ is called *strictly asymptotically more efficient* than $\hat{\beta}^{(2)}$.

The strict asymptotic efficiency is related to making the asymptotic confidence ellipsoids. Usually, the ACM is nonsingular and continuous in $\beta$. Then the convergence (3.172) implies

$$\sqrt{n}\hat{\Sigma}_i^{-1/2}(\hat{\beta}^{(i)} - \beta) \;\xrightarrow{\mathrm{d}}\; \mathrm{N}(0, I_m)\,, \quad i = 1, 2\,, \quad m = \dim\beta\,. \tag{3.173}$$

Here $\hat{\Sigma}_i = \Sigma_i(\hat{\beta}^{(i)}) \xrightarrow{\mathbf{P}} \Sigma_i$.

Similar to Section 3.1.5, we construct the asymptotic confidence ellipsoids for $\beta$:

$$E_n^{(i)} = \left\{ z \in \mathbf{R}^m \colon (z - \hat{\beta}^{(i)})^{\mathrm{T}} \hat{\Sigma}_i^{-1} (z - \hat{\beta}^{(i)}) \le \frac{1}{n} (\chi^2_m)_\alpha \right\} , \quad i = 1, 2 . \tag{3.174}$$

Here, $1 - \alpha$ is the confidence probability, and $(\chi^2_m)_\alpha$ is a quantile of the $\chi^2_m$ distribution, see (3.106). Consider the centered ellipsoids

$$E_{n,c}^{(i)} = \left\{ z \in \mathbf{R}^m \colon z^{\mathrm{T}} \hat{\Sigma}_i^{-1} z \le \frac{1}{n} (\chi^2_m)_\alpha \right\} , \quad i = 1, 2 . \tag{3.175}$$

The boundary of ellipsoid (3.175), i.e., the elliptic surface is given by the equation

$$\partial E_{n,c}^{(i)} = \left\{ z \in \mathbf{R}^m \colon z^{\mathrm{T}} \hat{\Sigma}_i^{-1} z = \frac{1}{n} (\chi^2_m)_\alpha \right\} . \tag{3.176}$$

**Lemma 3.19.** *Suppose that* (3.172) *holds true and the ACMs $\Sigma_i(\beta)$ are continuous in $\beta$. Assume also that the estimator $\hat{\beta}^{(1)}$ is strictly asymptotically more effecient than $\hat{\beta}^{(2)}$. Then*

(a)

$$\mathbf{P}\{E_{n,c}^{(1)} \subset E_{n,c}^{(2)}, \partial E_{n,c}^{(1)} \cap \partial E_{n,c}^{(2)} = \emptyset\} \to 1 , \quad \text{as } n \to \infty , \tag{3.177}$$

(b) *if in addition both estimators $\hat{\beta}^{(i)}$ are strongly consistent, then it holds* eventually*:*

$$E_{n,c}^{(1)} \subset E_{n,c}^{(2)} , \quad \partial E_{n,c}^{(1)} \cap \partial E_{n,c}^{(2)} = \emptyset . \tag{3.178}$$

**Remark 3.20.** Relation (3.178) means that one ellipsoid is "strictly" located inside the other. Lemma 3.19 can be interpreted as follows: the estimator, which is strictly asymptotically more efficient, generates "strictly less" asymptotic confidence ellipsoid.

*Proof of Lemma 3.19.* Because

$$\Sigma_1(\beta) < \Sigma_2(\beta) , \tag{3.179}$$

the matrices are continuous in $\beta$, and $\hat{\beta}^{(i)} \xrightarrow{\mathbf{P}} \beta$, then

$$\mathbf{P}\{\hat{\Sigma}_1 = \Sigma_1(\hat{\beta}_{(1)}) < \hat{\Sigma}_2 = \Sigma_2(\hat{\beta}_{(2)})\} \to 1 , \quad \text{as } n \to \infty . \tag{3.180}$$

By Theorem 16.E.3.b from the book by Marshall et al. (2011) and inequality $0 < \hat{\Sigma}_1 < \hat{\Sigma}_2$, it follows

$$\hat{\Sigma}_1^{-1} > \hat{\Sigma}_2^{-1} . \tag{3.181}$$

Let $z \in E_{n,c}^{(1)}$, then

$$\frac{1}{n} (\chi^2_m)_\alpha \ge z^{\mathrm{T}} \hat{\Sigma}_1^{-1} z \ge z^{\mathrm{T}} \hat{\Sigma}_2^{-1} z \quad \Rightarrow \quad \frac{1}{n} (\chi^2_m)_\alpha \ge z^{\mathrm{T}} \hat{\Sigma}_2^{-1} z , \tag{3.182}$$

and we have $z \in E_{n,c}^{(2)}$. From here, in this case, we get $E_{n,c}^{(1)} \subset E_{n,c}^{(2)}$. Next, let $u \in \partial E_{n,c}^{(1)}$, and we have

$$\frac{1}{n} (\chi^2_m)_\alpha = u^{\mathrm{T}} \hat{\Sigma}_1^{-1} u > u^{\mathrm{T}} \hat{\Sigma}_2^{-1} u \quad \Rightarrow \quad u \notin \partial E_{n,c}^{(2)} . \tag{3.183}$$

Thus,

$$\mathbf{P}\{\hat{\Sigma}_1 < \hat{\Sigma}_2\} \le \mathbf{P}\{E^{(1)}_{n,c} \subset E^{(2)}_{n,c},\, \partial E^{(1)}_{n,c} \cap \partial E^{(2)}_{n,c} = \emptyset\}\,, \tag{3.184}$$

and now the convergence (3.180) implies the convergence (3.177).

The statement (b) is proved similarly, one should just note that under the strong consistency of estimators $\hat{\Sigma}_i \xrightarrow{\mathbf{P}1} \Sigma_i(\beta)$, $i = 1, 2$, it holds *eventually* $\hat{\Sigma}_1 < \hat{\Sigma}_2$. The lemma is proved. □

Next, we are going to show that the estimator $\hat{\beta}_{\mathrm{QL}}$ is asymptotically effecient in a broad class of estimators. We deal with the normal polynomial model (3.6) and (3.3) under the conditions (i)–(iv).

Consider a linear in $y$ and unbiased estimating function

$$s_{\mathrm{L}}(y, x;\ \beta) = g(x, \beta)\cdot y - h(x, \beta), \quad y, x \in \mathbf{R}, \quad \beta \in \bar{B}_R = \{b \in \mathbf{R}^{k+1}:\ \|b\| \le R\}. \tag{3.185}$$

Here $g$ and $h$ are Borel measurable functions valued in $\mathbf{R}^{k+1}$.

Remember that the unbiasedness of $s_{\mathrm{L}}$ means the following: for every $\beta \in \bar{B}_R$,

$$\mathbf{E}_\beta s_{\mathrm{L}}(y, x; \beta) = 0\,. \tag{3.186}$$

The estimating functions $s_{\mathrm{C}}$ from (3.38) and $s_{\mathrm{QL}}$ from (3.135) satisfy the relations (3.185) and (3.186), respectively. The unbiasedness of $s_{\mathrm{C}}$ is provided by the equalities

$$\mathbf{E}[s_{\mathrm{C}}(y, x; \beta)|y, \xi] = s_{\mathrm{ML}}(y, \xi; \beta)\,, \quad \mathbf{E}_\beta s_{\mathrm{ML}}(y, \xi; \beta) = 0\,. \tag{3.187}$$

where $s_{\mathrm{ML}}$ is the estimating function of the ML method in the absence of measurement error $\delta$, and the unbiasedness of $s_{\mathrm{QL}}$ is fulfilled due to the equality

$$\mathbf{E}_\beta[s_{\mathrm{QL}}(y, x; \beta)|x] = 0\,. \tag{3.188}$$

Based on $s_{\mathrm{L}}$, the estimator $\hat{\beta}_{\mathrm{L}}$ is constructed as a measurable solution to the equation

$$\sum_{i=1}^{n} s_{\mathrm{L}}(y_i, x_i; \beta) = 0\,, \quad \beta \in \bar{B}_R\,. \tag{3.189}$$

We assume that the true value $b$ of the regression parameter belongs to $B_R = \{z \in \mathbf{R}^{k+1}: \|z\| < R\}$. Under mild general conditions (see Appendix A1), equation (3.189) has a solution with probability that tends to 1, as $n \to \infty$; a solution allows to define well the estimator $\hat{\beta}_{\mathrm{L}}$ to be consistent (i.e., $\hat{\beta}_{\mathrm{L}} \xrightarrow{\mathbf{P}} b$) and asymptotically normal (see Appendix A2). The ACM of the estimator $\hat{\beta}_{\mathrm{L}}$ is given by the sandwich formula

$$\Sigma_{\mathrm{L}} = A_{\mathrm{L}}^{-1} B_{\mathrm{L}} A_{\mathrm{L}}^{-\mathrm{T}}\,, \quad A_{\mathrm{L}} = -\mathbf{E}\frac{\partial s_{\mathrm{L}}}{\partial \beta}(y, x; b)\,, \quad B_{\mathrm{L}} = \mathrm{cov}_b\, s_{\mathrm{L}}(y, x; b)\,. \tag{3.190}$$

It is required that the matrix $A_{\mathrm{L}}$ is nonsingular. Denote by $L$ the class of all estimating functions of the form (3.185) and (3.186), for which the corresponding estimator $\hat{\beta}_{\mathrm{L}}$ has all abovementioned asymptotic properties. It is clear that $s_{\mathrm{C}}$ and $s_{\mathrm{QL}}$ belong to $L$.

**Theorem 3.21** (about asymptotic efficiency of the QL estimator). *Let $s_{\mathrm{L}} \in L$, $\Sigma_{\mathrm{L}}$ and $\Sigma_{\mathrm{QL}}$ be the ACM of the estimators $\hat{\beta}_{\mathrm{L}}$ and $\hat{\beta}_{\mathrm{QL}}$, respectively. Then*

$$\Sigma_{\mathrm{QL}} \leq \Sigma_{\mathrm{L}} \,. \tag{3.191}$$

*If additionally $\Sigma_{\mathrm{L}} = \Sigma_{\mathrm{QL}}$, for all the true values $b \in \bar{B}_R$, then $\hat{\beta}_{\mathrm{L}} = \hat{\beta}_{\mathrm{QL}}$* eventually.

*Proof* is given in Kukush et al. (2009).

From this theorem, it follows that the QL estimator is asymptotically more efficient than the ALS estimator. But for these two estimators one can state even more. Remember the notation $\beta_{-0} = (\beta_1, \ldots, \beta_k)^{\mathrm{T}}$.

**Theorem 3.22.** *Let $\Sigma_{\mathrm{C}}$ and $\Sigma_{\mathrm{QL}}$ be ACMs of the estimators $\hat{\beta}_{\mathrm{C}}$ and $\hat{\beta}_{\mathrm{QL}}$ in the model* (3.6) *and* (3.3) *under the assumptions* (i)–(iv).
(a) *If $\beta_{-0} = 0$, then $\Sigma_{\mathrm{QL}} = \Sigma_{\mathrm{C}}$.*
(b) *If $\beta_{-0} \neq 0$, then $\Sigma_{\mathrm{QL}} < \Sigma_{\mathrm{C}}$.*

*Proof* can be found in Kukush et al. (2006). The statement (a) is verified directly, but the statement (b) is nontrivial.

Next, consider the behavior of matrices $\Sigma_{\mathrm{QL}}$ and $\Sigma_{\mathrm{C}}$ for small $\sigma_\delta^2$. Each of the matrices is expanded into series w.r.t. the measurement error variance $\sigma_\delta^2$, as $\sigma_\delta^2 \to 0$. We will see that the difference of the matrices starts only with the terms of order $\sigma_\delta^4$.

**Theorem 3.23.** *Under the conditions of Theorem* 3.22, *for ACMs $\Sigma_{\mathrm{QL}} = \Sigma_{\mathrm{QL}}(\beta)$ and $\Sigma_{\mathrm{QL}} = \Sigma_{\mathrm{QL}}(\beta)$, it holds true that:*
(a)

$$\Sigma_{\mathrm{QL}} = \sigma_\varepsilon^2 (\mathbf{E}\rho\rho^{\mathrm{T}})^{-1} + O(\sigma_\delta^2) \,, \quad \text{as} \quad \sigma_\delta^2 \to 0 \,; \tag{3.192}$$

(b)

$$\Sigma_{\mathrm{C}} = \Sigma_{\mathrm{QL}} + O(\sigma_\delta^4) \,, \quad \text{as} \quad \sigma_\delta^2 \to 0 \,; \tag{3.193}$$

(c) *if $\beta_{-0} \neq 0$, then the matrix $\lim_{\sigma_\delta^2 \to 0} \sigma_\delta^{-4}(\Sigma_{\mathrm{C}} - \Sigma_{\mathrm{QL}})$ is positive definite.*

*Proof* of the statements (a) and (b) is given in Kukush et al. (2005b) and of the statement (c) in Malenko (2007).

Note that the statement (c) stems from Theorem 3.22 (b). Theorem 3.23 shows that for small measurement errors, we almost do not lose the efficiency of estimation, if instead of the QL estimator, we use the ALS estimator. The reason is as follows: the distinction of the two ACMs manifests only in the terms of order $\sigma_\delta^4$, and the terms of order $(\sigma_\delta^2)^0$ and $\sigma_\delta^2$ do not possess such a distinction. For small $\sigma_\delta^2$, it is advisable to use the ALS estimator instead of the QL estimator, because the ALS estimator does not require the assumption of normality of the latent variable $\xi$ (see discussion in Section 1.4.7).

### 3.2.2 The case of unknown error variance in response and known parameters of regressor's distribution

Now, we assume that in the polynomial model (3.6) and (3.3) under the assumptions (i)–(iv), the variance $\sigma_\delta^2$ and the parameters $\mu_\xi$ and $\sigma_\xi^2$ of the distribution of $\xi$ are known, and at the same time the variance $\sigma_\varepsilon^2$ is unknown.

**Estimation in linear model**

Consider the case $k = 1$. The estimating function (3.135) for $\beta = (\beta_0, \beta_1)^{\mathrm{T}}$ takes the form (see the formula (3.65)):

$$s_{\mathrm{QL}}^{(\beta)} = \frac{\mu(x)y - \mu(x)\mu^{\mathrm{T}}(x)\beta}{v(\sigma_\varepsilon^2, \beta_1)}\,, \quad \mu(x) = (\mu_0(x), \mu_1(x))^{\mathrm{T}}\,, \tag{3.194}$$

$$v(\sigma_\varepsilon^2, \beta_1) = \sigma_\varepsilon^2 + \tau^2\beta_1^2\,, \quad \tau^2 = K\sigma_\delta^2\,. \tag{3.195}$$

Since in the linear model, the conditional variance $v$ does not depend of $x$, it is not necessary to know $\sigma_\varepsilon^2$ for estimation of $\beta$ by the QL method. The reason is as follows: the estimator $\hat\beta_{\mathrm{QL}}$ is defined as a solution to the equation

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\mu(x_i)y_i - \mu(x_i)\mu^{\mathrm{T}}(x_i)\beta}{v(\sigma_\varepsilon^2, \beta_1)} = 0\,, \quad \beta \in \mathbf{R}^2\,, \tag{3.196}$$

and it is equivalent to the equation

$$\frac{1}{n}\sum_{i=1}^{n}(\mu(x_i)y_i - \mu(x_i)\mu^{\mathrm{T}}(x_i)\beta) = 0\,, \quad \beta \in \mathbf{R}^2\,. \tag{3.197}$$

Therefore, the equivalent estimating function for $\beta$ is equal to

$$\tilde s_{\mathrm{QL}}^{(\beta)} = \mu(x)y - \mu(x)\mu^{\mathrm{T}}(x)\beta\,. \tag{3.198}$$

The estimator $\hat\beta_{\mathrm{QL}}$ has the form

$$\hat\beta_{\mathrm{QL}} = \left(\overline{\mu\mu^{\mathrm{T}}}\right)^{-1}\overline{\mu y}\,. \tag{3.199}$$

(In our normal model, the matrix $\overline{\mu\mu^{\mathrm{T}}}$ is nonsingular, with probability 1.)

Based on equality (3.65), the estimator of $\sigma_\varepsilon^2$ can be written explicitly. We have:

$$\mathbf{E}[(y - \mu^{\mathrm{T}}(x)\beta)^2|x] = \sigma_\varepsilon^2 + \tau^2\beta_1^2\,, \tag{3.200}$$

$$\mathbf{E}(y - \mu^{\mathrm{T}}(x)\beta)^2 = \mathbf{E}\mathbf{E}[(y - \mu^{\mathrm{T}}(x)\beta)^2|x] = \sigma_\varepsilon^2 + \tau^2\beta_1^2\,. \tag{3.201}$$

Therefore, by the method of moments the estimating function for $\sigma_\varepsilon^2$ is the following:

$$\tilde s_{\mathrm{QL}}^{(\sigma_\varepsilon^2)} = (y - \mu^{\mathrm{T}}(x)\beta)^2 - \sigma_\varepsilon^2 - \tau^2\beta_1^2\,. \tag{3.202}$$

The estimators $\hat{\beta}_{\mathrm{QL}}$ and $\hat{\sigma}^2_{\varepsilon,\mathrm{QL}}$ are defined as a solution to the system of equations

$$\frac{1}{n}\sum_{i=1}^{n} \tilde{s}_{\mathrm{QL}}(x_i, y_i; \beta) = 0\,, \tag{3.203}$$

$$\frac{1}{n}\sum_{i=1}^{n} \tilde{s}^{(\sigma^2_\varepsilon)}_{\mathrm{QL}}(y_i, x_i; \beta, \sigma^2_\varepsilon) = 0\,, \quad \beta \in \mathbf{R}^2\,, \quad \sigma^2_\varepsilon > 0\,. \tag{3.204}$$

From here

$$\hat{\sigma}^2_{\varepsilon,\mathrm{QL}} = \frac{1}{n}\sum_{i=1}^{n} \left(y_i - \mu^{\mathrm{T}}(x_i)\hat{\beta}_{\mathrm{QL}}\right)^2 - \tau^2 \cdot \hat{\beta}^2_{1,\mathrm{QL}}\,, \tag{3.205}$$

if the expression is positive. Note that in Cheng and Schneeweiss (1998), it is proposed to use another estimator:

$$\tilde{\sigma}^2_{\varepsilon,\mathrm{QL}} = \frac{1}{n-2}\sum_{i=1}^{n} \left(y_i - \mu^{\mathrm{T}}(x_i)\hat{\beta}_{\mathrm{QL}}\right)^2 - \tau^2 \cdot \hat{\beta}^2_{1,\mathrm{QL}}\,, \tag{3.206}$$

which is written similarly to the variance estimator in ordinary regression (here, in the denominator, $2 = \dim\beta$). However, the estimators (3.205) and (3.206) are *asymptotically equivalent*, i.e.,

$$\sqrt{n}\,(\hat{\sigma}^2_{\varepsilon,\mathrm{QL}} - \tilde{\sigma}^2_{\varepsilon,\mathrm{QL}}) \xrightarrow{\mathbf{P}} 0\,. \tag{3.207}$$

This leads to the fact that the ACMs of the estimators $\hat{\sigma}^2_{\varepsilon,\mathrm{QL}}$ and $\tilde{\sigma}^2_{\varepsilon,\mathrm{QL}}$ coincide.

Remember that the sample covariance of two samples $U_1, \ldots, U_n$ and $V_1, \ldots, V_n$ is denoted by $\mathrm{S}_{UV}$, see the proof of Theorem 2.22. In particular for $\mu_1 = \mu_1(x)$, we get

$$\mathrm{S}_{\mu_1 y} = \frac{1}{n}\sum_{i=1}^{n} (\mu_1(x_i) - \overline{\mu_1(x_i)})(y_i - \bar{y})\,, \tag{3.208}$$

$$\mathrm{S}_{\mu_1\mu_1} = \frac{1}{n}\sum_{i=1}^{n} (\mu_1(x_i) - \overline{\mu_1(x)})^2\,. \tag{3.209}$$

**Theorem 3.24.** *Consider the linear model* (3.6) *and* (3.3)*, with* $k = 1$*, for which the conditions* (i)–(iv) *are fulfilled and* $n \geq 2$.

(a) *Components of the estimator* (3.199) *can be found from the equalities, being performed almost surely:*

$$\hat{\beta}_{1,\mathrm{QL}} = \frac{\mathrm{S}_{\mu_1 y}}{\mathrm{S}_{\mu_1\mu_1}}\,, \quad \bar{y} = \hat{\beta}_{0,\mathrm{QL}} + \hat{\beta}_{1,\mathrm{QL}} \times \overline{\mu_1(x)}\,. \tag{3.210}$$

(b) *The estimators* (3.199) *and* (3.205) *are strongly consistent, i.e., it holds for the true values* $\beta$ *and* $\sigma^2_\varepsilon$*:*

$$\hat{\beta}_{\mathrm{QL}} \xrightarrow{\mathbf{P}1} \beta\,, \quad \hat{\sigma}^2_{\varepsilon,\mathrm{QL}} \xrightarrow{\mathbf{P}1} \sigma^2_\varepsilon\,. \tag{3.211}$$

(c) *The estimators* (3.199) *and* (3.205) *of the parameter* $\theta = (\beta^{\mathrm{T}}, \sigma^2_\varepsilon)^{\mathrm{T}}$ *are asymptotically normal, namely:*

$$\Sigma_\theta = A_\theta^{-1} B_\theta A_\theta^{-\mathrm{T}}\,, \quad A_\theta = \begin{pmatrix} \mathbf{E}\mu\mu^{\mathrm{T}} & 0 \\ 0; & 2\tau^2\beta_1 \quad 1 \end{pmatrix}\,, \quad B_\theta = \begin{pmatrix} v \cdot \mathbf{E}\mu\mu^{\mathrm{T}} & 0 \\ 0 & 2v^2 \end{pmatrix}\,, \tag{3.212}$$

*where* $v$ *is given by equalities* (3.195).

*Proof.* (a) Since $\mu_0(x) = 1$, the system of equations (3.197) for $\hat\beta_{0,\mathrm{QL}}$ and $\hat\beta_{1,\mathrm{QL}}$ has the form

$$\begin{cases} \bar y - \beta_0 - \overline{\mu_1(x)}\cdot\beta_1 = 0\,, \\ \overline{\mu_1(x)\cdot y} - \overline{\mu_1(x)}\cdot\beta_0 - \overline{\mu_1^2(x)}\cdot\beta_1 = 0\,. \end{cases} \tag{3.213}$$

Therefore, the solutions to the system are presented in (3.210), provided $\mathrm{S}_{\mu_1\mu_1} \neq 0$ (here the reasoning is similar to the proof of Theorem 2.1). But the latter condition holds a.s., with $n \geq 2$.

(b) The strong consistency of the estimator $\hat\beta_{\mathrm{QL}}$ follows from Theorem 3.13(b). However, we will carry out a straightforward simple proof based on the formula (3.199).

By the SLLN,

$$\overline{\mu\mu^{\mathrm{T}}} \xrightarrow{\mathbf{P}1} \mathbf{E}\mu\mu^{\mathrm{T}} > 0\,. \tag{3.214}$$

The positive definiteness of the latter matrix was explained in the proof of Theorem 3.13. Next, $\overline{\mu y} = \overline{\mu\rho^{\mathrm{T}}}\cdot\beta + \overline{\mu\varepsilon}$,

$$\overline{\mu y} \xrightarrow{\mathbf{P}1} \left(\mathbf{E}\mu(x)\rho^{\mathrm{T}}(\xi)\right)^{\mathrm{T}} + \mathbf{E}\mu(x)\varepsilon = (\mathbf{E}\mu\rho^{\mathrm{T}})\cdot\beta\,. \tag{3.215}$$

But

$$\mathbf{E}\mu\rho^{\mathrm{T}} = \mathbf{E}\mathbf{E}(\mu\rho^{\mathrm{T}}|x) = \mathbf{E}\left[\mu(x)\mathbf{E}(\rho^{\mathrm{T}}(\xi)|x)\right] = (\mathbf{E}\mu\mu^{\mathrm{T}})\beta\,, \tag{3.216}$$

$$\hat\beta_{\mathrm{QL}} \xrightarrow{\mathbf{P}1} (\mathbf{E}\mu\mu^{\mathrm{T}})^{-1}\cdot(\mathbf{E}\mu\mu^{\mathrm{T}})\beta = \beta\,. \tag{3.217}$$

Then by the formula (3.205), we have

$$\hat\sigma^2_{\varepsilon,\mathrm{QL}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu^{\mathrm{T}}(x_i)\beta)^2 + o(1) - \tau^2\hat\beta^2_{1,\mathrm{QL}}\,, \tag{3.218}$$

$$\begin{aligned} \hat\sigma^2_{\varepsilon,\mathrm{QL}} &\xrightarrow{\mathbf{P}1} \mathbf{E}_\beta(y - \mu^{\mathrm{T}}(x)\beta)^2 - \tau^2\beta_1^2 = \\ &= \mathbf{E}\mathbf{E}_\beta\left[(y - \mu^{\mathrm{T}}(x)\beta)^2|x\right] - \tau^2\beta_1^2 = \upsilon - \tau^2\beta_1^2 = \sigma^2_\varepsilon\,. \end{aligned} \tag{3.219}$$

(c) Use the sandwich formula from Appendix A2. The estimating function $\tilde{\mathrm{s}}^{(\theta)}_{\mathrm{QL}}$ has components (3.198) and (3.202). We have

$$A_\theta = -\mathbf{E}\frac{\partial\tilde{\mathrm{s}}^{(\theta)}_{\mathrm{QL}}}{\partial\theta} = \left(\begin{array}{cc|c} & \mathbf{E}\mu\mu^{\mathrm{T}} & 0 \\ \hline 0; & 2\tau^2\beta_1 & 1 \end{array}\right)\,, \tag{3.220}$$

$$B_\theta = \mathrm{cov}_\theta\,\tilde{\mathrm{s}}^{(\theta)}_{\mathrm{QL}}(y, x; \beta) = \begin{pmatrix} B_\beta & B_{\beta,\sigma^2_\varepsilon} \\ B_{\sigma^2_\varepsilon,\beta} & B_{\sigma^2_\varepsilon} \end{pmatrix}\,. \tag{3.221}$$

Here the diagonal blocks correspond to the autocovariance of components of the estimating functions $\tilde{\mathrm{s}}^{(\beta)}_{\mathrm{QL}}$ and $\tilde{\mathrm{s}}^{(\sigma^2_\varepsilon)}_{\mathrm{QL}}$, and the off-diagonal blocks to the mutual covariance of the components. Note that in our linear model, normal random variables $\xi$, $\delta$, and $\varepsilon$ are independent, and therefore, the conditional distribution of $y$ given $x$ is normal,

$$y|x \sim \mathrm{N}(\mu^{\mathrm{T}}(x)\beta, \upsilon)\,, \quad \upsilon = \sigma^2_\varepsilon + \tau^2\beta_1^2\,. \tag{3.222}$$

We obtain

$$\begin{aligned} B_\beta &= \operatorname{cov}_\theta \mu(x)(y-\mu^{\mathrm{T}}(x)\beta) = \mathbf{E}_\theta \mu(x)(y-\mu^{\mathrm{T}}(x)\beta)^2\mu^{\mathrm{T}}(x) = \\ &= \mathbf{E}\mathbf{E}_\theta\left[\mu(x)(y-\mu^{\mathrm{T}}(x)\beta)^2\mu^{\mathrm{T}}(x)|x\right] = v\cdot\mathbf{E}\mu\mu^{\mathrm{T}}\,, \end{aligned} \tag{3.223}$$

$$\begin{aligned} B_{\sigma_\varepsilon^2} &= \operatorname{cov}_\theta\left[(y-\mu^{\mathrm{T}}(x)\beta)^2 - v\right] \\ &= \mathbf{E}_\theta(y-\mu^{\mathrm{T}}(x)\beta)^4 - 2v\mathbf{E}_\theta(y-\mu^{\mathrm{T}}(x)\beta)^2 + v^2 \\ &= \mathbf{E}\mathbf{E}_\theta\left[(y-\mu^{\mathrm{T}}(x)\beta)^4|x\right] - v^2 = 3v^2 - v^2 = 2v^2\,. \end{aligned} \tag{3.224}$$

Here we used the relation (3.222) and the fact that the the fourth central moment of normal random variable $y \sim \mathrm{N}(m, v)$ is the following:

$$\mathbf{E}(y-m)^4 = 3v^2\,. \tag{3.225}$$

Finally,

$$\begin{aligned} B_{\sigma_\varepsilon^2,\beta} &= \mathbf{E}_\theta\mu(x)\,(y-\mu^{\mathrm{T}}(x)\beta)\,((y-\mu^{\mathrm{T}}(x)\beta)^2 - v) = \\ &= \mathbf{E}_\theta\mu(x)\,(y-\mu^{\mathrm{T}}(x)\beta)^3 - v\,\mathbf{E}_\theta\mu(x)\,(y-\mu^{\mathrm{T}}(x)\beta) = 0\,. \end{aligned} \tag{3.226}$$

We made use of the fact that, as a result of (3.222), it holds

$$\mathbf{E}\left[(y-\mu^{\mathrm{T}}(x)\beta)^3|x\right] = 0\,. \tag{3.227}$$

Then $B_{\beta,\sigma_\varepsilon^2} = B_{\sigma_\varepsilon^2,\beta}^{\mathrm{T}} = 0$, and the matrix (3.221) indeed has the form (3.212). Now, the statement (b) follows from the sandwich formula and equality (3.220). The theorem is proved. □

**Remark 3.25.** Consider the estimated straight line $y = \hat\beta_{0,\mathrm{QL}} + \hat\beta_{1,\mathrm{QL}}\xi$. In general the center of mass of the sample $M(\bar x;\bar y)$ does not lie on this straight line (cf. the estimated straight lines from Chapter 2). Instead, the point $\tilde M(\overline{\mu_1(x)};\bar y)$ is located on the straight line. Here

$$\overline{\mu_1(x)} = K\bar x + (1-K)\mu_\xi \tag{3.228}$$

is a posterior estimator of the expectation $\xi$ (a prior estimator is $\mu_\xi$).

**Estimation in polynomial model**

Consider the model (3.6) and (3.3) under the conditions (i)–(iv), with arbitrary $k \geq 1$. Now, only the parameters $\sigma_\delta^2$, $\mu_\xi$, and $\sigma_\xi^2$ are assumed known. When $k \geq 2$ the conditional variance $v = \mathbf{V}(y|x)$ depends not only on $\sigma_\varepsilon^2$ and $\beta$ but also on $x$, see the formula (3.67) for $k = 2$. Therefore under unknown $\sigma_\varepsilon^2$, the estimating equation (3.136) can not be exploited for estimation of $\beta$. In addition to the estimating function (3.135) one should form a separate estimating function corresponding to the parameter $\sigma_\varepsilon^2$.

To do this, we use equality (3.60). We have

$$\begin{aligned} \mathbf{E}_\beta(y-\mu^{\mathrm{T}}(x)\beta)^2 &= \mathbf{E}\mathbf{E}_\beta\left[(y-\mu^{\mathrm{T}}(x)\beta)^2|x\right] = \mathbf{E}v(x;\beta,\sigma_\varepsilon^2) \\ &= \sigma_\varepsilon^2 + \beta^{\mathrm{T}}\mathbf{E}[M(x)-\mu(x)\mu^{\mathrm{T}}(x)]\cdot\beta\,. \end{aligned} \tag{3.229}$$

The additional estimating function is specified as follows:

$$s_{\mathrm{QL}}^{(\sigma_\varepsilon^2)}(y, x; \beta, \sigma_\varepsilon^2) = (y - \mu^{\mathrm{T}}(x)\beta)^2 - \sigma_\varepsilon^2 - \beta^{\mathrm{T}}[M(x) - \mu(x)\mu^{\mathrm{T}}(x)]\,\beta\,. \tag{3.230}$$

Respectively, we compose an additional equation to define the ALS estimator:

$$\sigma_\varepsilon^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu^{\mathrm{T}}(x_i)\beta)^2 - \beta^{\mathrm{T}} \cdot \frac{1}{n}\sum_{i=1}^{n}[M(x_i) - \mu(x_i)\mu^{\mathrm{T}}(x_i)] \cdot \beta\,,$$
$$\beta \in \mathbf{R}^{k+1}\,, \quad \sigma_1^2 \le \sigma_\varepsilon^2 \le \sigma_2^2\,. \tag{3.231}$$

Here, $0 < \sigma_1^2 < \sigma_2^2$ are given thresholds for the variance of errors in response.

**Definition 3.26.** The estimators $\hat\beta_{\mathrm{QL}}$ and $\hat\sigma^2_{\varepsilon,\mathrm{QL}}$ are Borel measurable functions of observations $y_1, x_1, \ldots, y_n, x_n$ such that:

(a) If the system of equations (3.136) and (3.231) has no solution for $\beta \in \mathbf{R}^{k+1}$ and $\sigma_\varepsilon^2 \in [\sigma_1^2, \sigma_2^2]$, then $\hat\beta_{\mathrm{QL}} = 0$, $\hat\sigma^2_{\varepsilon,\mathrm{QL}} = \sigma_1^2$;

(b) If the system has a solution for the mentioned set, then $\hat\beta_{\mathrm{QL}}$ is a solution for $\beta$ with the smallest possible norm (if there are several such solutions, any of them can be taken).

We mention that the definition is correct. Indeed, in consequence to the continuity of the estimating functions $s_{\mathrm{QL}}^{(\beta)}$ and $s_{\mathrm{QL}}^{(\sigma_\varepsilon^2)}$ in both $\beta$ and $\sigma_\varepsilon^2$, the set of solutions $A_\omega$ to the system (3.136) and (3.231) is closed, for each elementary event $\omega$; hence if $A_\omega \neq \emptyset$, then $\min_{(\beta,\sigma_\varepsilon^2)\in A_\omega}\|\beta\|$ is attained.

Denote

$$\theta = \begin{pmatrix}\beta\\ \sigma_\varepsilon^2\end{pmatrix}\,, \quad s_{\mathrm{QL}}^{(\theta)} = \begin{pmatrix} s_{\mathrm{QL}}^{(\beta)} \\ s_{\mathrm{QL}}^{(\sigma_\varepsilon^2)}\end{pmatrix}\,. \tag{3.232}$$

The latter is the augmented estimating function of the parameter $\theta$.

**Theorem 3.27.** *Let the parameters $\sigma_\delta^2$, $\mu_\xi$, and $\sigma_\xi^2$ be known in the model* (3.6) *and* (3.3) *under the conditions* (i)–(iv), *and moreover we know that $\sigma_\varepsilon^2 \in (\sigma_1^2, \sigma_2^2)$, with $\sigma_1^2 > 0$. Denote by $b = (b_i)_{i=0}^k$ the true values of polynomial coefficients, and by $s_\varepsilon^2$ the true value of the parameter $\sigma_\varepsilon^2$. Then the following statements are valid:*

(a) *for any $R > \|b\|$, the system of equations* (3.136) *and* (3.231) *has* eventually *a unique solution in $\bar B_R \times [\sigma_1^2, \sigma_2^2]$, where the ball $\bar B_R$ is given in* (3.137)*;*

(b)
$$\hat\beta_{\mathrm{QL}} \xrightarrow{\mathbf{P1}} b\,, \quad \hat\sigma^2_{\varepsilon,\mathrm{QL}} \xrightarrow{\mathbf{P1}} s_\varepsilon^2\,, \quad \text{as } n \to \infty\,; \tag{3.233}$$

(c)
$$\sqrt{n}\begin{pmatrix}\hat\beta_{\mathrm{QL}} - b\\ \hat\sigma^2_{\varepsilon,\mathrm{QL}} - s_\varepsilon^2\end{pmatrix} \xrightarrow{\mathrm{d}} \mathrm{N}(0, \Sigma_\theta)\,, \tag{3.234}$$

$$\Sigma_\theta = \Sigma_\theta(b, s_\varepsilon^2) > 0\,, \quad \Sigma_\theta = A_\theta^{-1} B_\theta A_\theta^{-\mathrm{T}}\,, \tag{3.235}$$

$$A_\theta = -\mathbf{E}_{b,s_\varepsilon^2} \frac{\partial s_{\mathrm{QL}}^{(\theta)}(y, x; b, s_\varepsilon^2)}{\partial \theta^{\mathrm{T}}}\,, \tag{3.236}$$

$$B_\theta = \mathrm{cov}_{b,s_\varepsilon^2}\, s_{\mathrm{QL}}^{(\theta)}(y, x; b, s_\varepsilon^2)\,; \tag{3.237}$$

(d)

$$\sqrt{n}(\hat\beta_{\mathrm{QL}} - b) \xrightarrow{\mathrm{d}} \mathrm{N}\left(0, \left(\mathbf{E}\frac{\mu\mu^{\mathrm{T}}}{v(x; b, s_\varepsilon^2)}\right)^{-1}\right)\,. \tag{3.238}$$

**Remark 3.28.** Thus, the ACM of the QL estimator for the parameter $\beta$ is just the same if the variance $\sigma_\varepsilon^2$ is known or not, see formulas (3.139), (3.140), and (3.238).

*Proof of the theorem.* The statements (a) and (b) follow from the theory of estimating equations (see Appendix A1). Here it is important that $s_\varepsilon^2$ lies inside the interval $[\sigma_1^2, \sigma_2^2]$. We will explain only why the limit equation has a unique solution.

On the set $\bar B_R \times [\sigma_1^2, \sigma_2^2]$, almost surely, the left-hand side of equation (3.136) converges uniformly to the function (see (3.142)):

$$s_\infty^{(\beta)}(\beta, \sigma_\varepsilon^2) = \mathbf{E}\frac{\mu\mu^{\mathrm{T}}}{v(x; \beta, \sigma_\varepsilon^2)}(b - \beta)\,. \tag{3.239}$$

On the same set almost surely, the function (3.230) converges uniformly to

$$s_\infty^{(\sigma_\varepsilon^2)}(\beta, \sigma_\varepsilon^2) = \mathbf{E}_{b,s_\varepsilon^2}(y - \mu^{\mathrm{T}}\beta)^2 - \sigma_\varepsilon^2 - \beta^{\mathrm{T}} \cdot \mathbf{E}[M(x) - \mu\mu^{\mathrm{T}}] \cdot \beta\,. \tag{3.240}$$

Consider the limit system of equations

$$s_\infty^{(\beta)}(\beta, \sigma_\varepsilon^2) = 0\,, \quad s_\infty^{(\sigma_\varepsilon^2)}(\beta, \sigma_\varepsilon^2) = 0\,, \quad (\beta, \sigma_\varepsilon^2) \in \bar B_R \times [\sigma_1^2, \sigma_2^2]\,. \tag{3.241}$$

From the first equation, we have $\beta = b$ (see proof of Theorem 3.13). Then the second equation takes the form

$$\mathbf{E}_{b,s_\varepsilon^2}(y - \mu^{\mathrm{T}}b)^2 - \sigma_\varepsilon^2 - b^{\mathrm{T}}\mathbf{E}[M(x) - \mu\mu^{\mathrm{T}}]b = 0\,, \tag{3.242}$$

or $s_\varepsilon^2 - \sigma_\varepsilon^2 = 0$, $s_\varepsilon^2 = \sigma_\varepsilon^2$. Thus, the system (3.241) has a unique solution $\theta = \binom{b}{s_\varepsilon^2}$. This fact is the basis for the statements (a) and (b) being satisfied.

The statement (c) follows from the sandwich formula (see Appendix A2). We explain only the nonsingularity of the matrices $A_\theta$ and $B_\theta$. We have

$$A_\theta = -\mathbf{E}_{b,s_\varepsilon^2} \frac{\partial s_{\mathrm{QL}}^{(\theta)}}{\partial \theta^{\mathrm{T}}}(y, x; b, s_\varepsilon^2) = \begin{pmatrix} \mathbf{E}\frac{\mu\mu^{\mathrm{T}}}{v(x;b,s_\varepsilon^2)} & 0 \\ A_{21} & 1 \end{pmatrix}\,. \tag{3.243}$$

We used the fact that

$$\mathbf{E}_{b,s_\varepsilon^2}\frac{\partial s_{\mathrm{QL}}^{(\beta)}(b,s_\varepsilon^2)}{\partial\sigma_\varepsilon^2}=\mathbf{E}_{b,s_\varepsilon^2}\mu\,(y-\mu^{\mathrm{T}}b)\,\frac{\partial}{\partial\sigma_\varepsilon^2}\frac{1}{v}(b,s_\varepsilon^2)=$$

$$=\mathbf{E}\left[\mu\,\mathbf{E}_b(y-\mu^{\mathrm{T}}b|x)\cdot\frac{\partial}{\partial\sigma_\varepsilon^2}\frac{1}{v}(b,s_\varepsilon^2)\right]=0\,; \tag{3.244}$$

$$A_{21}=-\mathbf{E}_{b,s_\varepsilon^2}\frac{\partial s_{\mathrm{QL}}^{(\sigma_\varepsilon^2)}}{\partial\beta^{\mathrm{T}}}(b,s_\varepsilon^2)=2b^{\mathrm{T}}\mathbf{E}(M(x)-\mu\mu^{\mathrm{T}})\,. \tag{3.245}$$

(As we saw in the proof of Theorem 3.24, this vector is not identically zero even for $k=1$). The matrix (3.243) is nonsingular because

$$\det A_\theta=\det\mathbf{E}\frac{\mu\mu^{\mathrm{T}}}{v(x;b,s_\varepsilon^2)}\neq 0\,. \tag{3.246}$$

Further,

$$B_\theta=\mathrm{cov}_{b,s_\varepsilon^2}\,s_{\mathrm{QL}}^{(\theta)}(b,s_\varepsilon^2)=\begin{pmatrix}B_{11}&B_{12}\\B_{21}&B_{22}\end{pmatrix}\,, \tag{3.247}$$

$$B_{11}=\mathrm{cov}_{b,s_\varepsilon^2}\,s_{\mathrm{QL}}^{(\beta)}(b,s_\varepsilon^2)=\mathbf{E}\frac{\mu\mu^{\mathrm{T}}}{v(x;b,s_\varepsilon^2)}>0\,, \tag{3.248}$$

$$B_{22}=\mathrm{cov}_{b,s_\varepsilon^2}\,s_{QL}^{(\sigma_\varepsilon^2)}\,,\qquad B_{12}=B_{21}^{\mathrm{T}}=\mathbf{E}_{b,s_\varepsilon^2}s_{\mathrm{QL}}^{(\beta)}\cdot s_{\mathrm{QL}}^{(\sigma_\varepsilon^2)}\,. \tag{3.249}$$

We explain why $B_\theta>0$. The determinant of $B_\theta$ is the Gram determinant for components of $s_{\mathrm{QL}}^{(\theta)}$ in the space $L_2(\Omega,\mathbf{P})$ of random variables. Thus, it is enough to prove the linear independence of the components in $L_2(\Omega,\mathbf{P})$.

Suppose that for some $c\in\mathbf{R}^{k+1}$ and $d\in\mathbf{R}$, it holds

$$c^{\mathrm{T}}s_{\mathrm{QL}}^{(\beta)}(y,x;b,s_\varepsilon^2)+d\cdot s_{\mathrm{QL}}^{(\sigma_\varepsilon^2)}(y,x;b,s_\varepsilon^2)=0\,,\quad\text{a.s.} \tag{3.250}$$

Then $\mathbf{E}(c^{\mathrm{T}}s_{\mathrm{QL}}^{(\beta)}+d\cdot s_{\mathrm{QL}}^{(\sigma_\varepsilon^2)}|x,\xi)=0$. The components of $s_{\mathrm{QL}}^{(\beta)}$ are linear in $y$, and $s_{\mathrm{QL}}^{(\sigma_\varepsilon^2)}$ is a square function in $y$. So we get from here (since $\varepsilon$ is stochastically independent of $x$ and $\xi$):

$$d\cdot\varepsilon^2+\cdots=0\,,\quad\text{a.s.} \tag{3.251}$$

Here, the unwritten terms are either linear in $\varepsilon$ or do not contain $\varepsilon$. Since $\varepsilon$ has a continuous (normal) distribution, it follows from (3.251) that $d=0$. Then in (3.250) we have

$$c^{\mathrm{T}}s_{\mathrm{QL}}^{(\beta)}=0\,,\quad\text{a.s.} \tag{3.252}$$

But the components of $s_{\mathrm{QL}}^{(\beta)}$ are linearly independent due to (3.248), whence $c=0$. For this reason, the components of $s_{\mathrm{QL}}^{(\theta)}$ are linearly independent, and $B_\theta>0$.

Now, formulas (3.234)–(3.237) follow from the sandwich formula. The nonsingularity of the matrix $\Sigma_\theta$ follows from the nonsingularity of components of $A_\theta$ and $B_\theta$.

(d) It is necessary to compute the "top left" block of ACM $\Sigma_\theta$, which corresponds to the ACM of the estimator $\hat{\beta}_{\mathrm{QL}}$. According to (3.243),

$$A_\theta = \begin{pmatrix} A_{11} & 0 \\ A_{21} & 1 \end{pmatrix}. \tag{3.253}$$

It is easy to verify that

$$A_\theta^{-1} = \begin{pmatrix} A_{11}^{-1} & 0 \\ -A_{21}A_{11}^{-1} & 1 \end{pmatrix}. \tag{3.254}$$

The matrix $B_\theta$ is written in a block form (3.247). Then

$$A_\theta^{-1}B_\theta = \begin{pmatrix} A_{11}^{-1}B_{11} & * \\ * & * \end{pmatrix}. \tag{3.255}$$

Hereafter, "$*$" denote the unwritten blocks. Finally,

$$\Sigma_\theta = A_\theta^{-1}B_\theta A_\theta^{-\mathrm{T}} = \begin{pmatrix} A_{11}^{-1}B_{11} & * \\ * & * \end{pmatrix} \begin{pmatrix} A_{11}^{-1} & -A_{11}^{-1}A_{21}^{\mathrm{T}} \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} A_{11}^{-1}B_{11}A_{11}^{-1} & * \\ * & * \end{pmatrix}. \tag{3.256}$$

Then from the convergence (3.234) and formula (3.256), we obtain

$$\sqrt{n}(\hat{\beta}_{\mathrm{QL}} - b) \xrightarrow{\mathrm{d}} \mathrm{N}(0, A_{11}^{-1}B_{11}A_{11}^{-1}) = \mathrm{N}\left(0, \left(\mathbf{E}\frac{\mu\mu^{\mathrm{T}}}{v}\right)^{-1}\right). \tag{3.257}$$

This completes the proof of the theorem. □

The estimator $\hat{\theta}_{\mathrm{QL}}$ can be computed by the iteratively reweighted least squares method (see Section 3.2.1; we introduce all necessary corrections to the description of the method when the variance $\sigma_\varepsilon^2$ is unknown).

Rewrite (3.136) as

$$\beta = \phi_n(\beta, \sigma_\varepsilon^2), \tag{3.258}$$

$$\phi_n(\beta, \sigma_\varepsilon^2) := \left(\frac{1}{n}\sum_{i=1}^n \frac{\mu(x_i)\mu^{\mathrm{T}}(x_i)}{v(x_i;\beta,\sigma_\varepsilon^2)}\right)^{-1} \cdot \frac{1}{n}\sum_{i=1}^n \frac{\mu(x_i)y_i}{v(x_i;\beta,\sigma_\varepsilon^2)}. \tag{3.259}$$

Denote by $\psi_n(\beta)$ the right-hand side of (3.231).

The iterative algorithm is as follows.

(1) We determine arbitrary initial values $\beta_n^{(0)} = \beta^{(0)}$ (e.g., $\beta^{(0)} = 0$) and $(\sigma_{\varepsilon,n}^2)^{(0)} = (\sigma_\varepsilon^2)^{(0)} \in [\sigma_1^2, \sigma_2^2]$.

(2) Given $\beta^{(j)}$ and $(\sigma_\varepsilon^2)^{(j)}$ from the $j$th iteration of the algorithm, we find

$$\beta^{(j+1)} = \phi_n(\beta^{(j)}, (\sigma_\varepsilon^2)^{(j)}), \tag{3.260}$$

$$(\sigma_\varepsilon^2)^{(j+1)} = P(\psi_n(\beta^{(j+1)})). \tag{3.261}$$

Here $P$ is a projector on $[\sigma_1^2, \sigma_2^2]$,

$$P(x) = \begin{cases} x & \text{if } x \in [\sigma_1^2, \sigma_2^2] \,, \\ \sigma_1^2 & \text{if } x < \sigma_1^2 \,, \\ \sigma_2^2 & \text{if } x > \sigma_2^2 \,. \end{cases} \tag{3.262}$$

As in Theorem 3.15, one can prove the statement about convergence of the algorithm.

**Theorem 3.29.** *Assume the conditions of Theorem* 3.27. *Suppose that initial values* $\beta_n^{(0)}$ *and* $(\sigma_\varepsilon^2)^{(0)}$ *of the algorithm are random, moreover for some* $R > 0$, *it holds that* $\|\beta_n^{(0)}\| < R$, eventually. *Then* eventually,

$$\lim_{j\to\infty} \beta^{(j)} = \hat{\beta}_{\mathrm{QL}} \,, \quad \lim_{j\to\infty} (\sigma_\varepsilon^2)^{(j)} = \hat{\sigma}^2_{\varepsilon,\mathrm{QL}} \,. \tag{3.263}$$

**Remark 3.30.** In Cheng and Schneeweiss (1998) instead of the additional equation (3.231), it is proposed to use a slightly modified equation:

$$\sigma_\varepsilon^2 = \frac{1}{n-(k+1)} \sum_{i=1}^{n} (y_i - \mu^{\mathrm{T}}(x_i)\beta)^2 - \beta^{\mathrm{T}} \cdot \frac{1}{n} \sum_{i=1}^{n} [M(x_i) - \mu(x_i)\mu^{\mathrm{T}}(x_i)] \cdot \beta \,, \tag{3.264}$$

being written by analogy with the variance estimator in ordinary regression (here in the denominator, $k + 1 = \dim \beta$). However, such a change of the estimating function yields the asymptotically equivalent estimator $\tilde{\theta}_{\mathrm{QL}}$, i.e.,

$$\sqrt{n} \left\| \hat{\theta}_{\mathrm{QL}} - \tilde{\theta}_{\mathrm{QL}} \right\| \xrightarrow{\mathbf{P}} 0 \,. \tag{3.265}$$

Therefore, the ACMs of the estimators $\hat{\theta}_{\mathrm{QL}}$ and $\tilde{\theta}_{\mathrm{QL}}$ coincide. As a result, in the iterative algorithm, one can use the right-hand side of (3.264) as $\psi_n(\beta)$, whereupon the statement of Theorem 3.29 remains valid.

### 3.2.3 The case where all the nuisance parameters are unknown

Now, the only parameter $\sigma_\delta^2$ is known in the model (3.6) and (3.3) under the conditions (i)–(iv). The augmented parameter is

$$\theta = \begin{pmatrix} \beta \\ \sigma_\varepsilon^2 \\ \gamma \end{pmatrix} \,, \quad \gamma = \begin{pmatrix} \mu_\xi \\ \sigma_\xi^2 \end{pmatrix} \,. \tag{3.266}$$

$\theta$ has to be estimated. Here, $\gamma$ is a vector of nuisance parameters in the distribution of $\xi$.

**Pre-estimation of the parameters of the distribution of $\xi$**

The vector $\mu(x) = \mathbf{E}(\rho(\xi)|x)$ depends on the parameter $\gamma$, which enters the conditional distribution

$$\xi|x \sim \mathrm{N}(\mu_1(x), \tau^2)\,, \tag{3.267}$$

$$\mu_1(x) = Kx + (1-K)\mu_\xi\,, \quad K = \frac{\sigma_\xi^2}{\sigma_\xi^2 + \sigma_\delta^2}\,, \quad \tau^2 = K\sigma_\delta^2\,. \tag{3.268}$$

Therefore, to estimate $\beta$ and $\sigma_\varepsilon^2$ we have to construct an aditional estimating function for $\gamma$. In this situation, it is natural to use a preliminary estimation (*pre-estimation*) of the parameter $\gamma$. Hence, the estimator $\hat\gamma = \hat\gamma_{\mathrm{QL}}$ can be constructed based on the sample $x_1, \dots, x_n$. We use the MLE $\hat\gamma_{\mathrm{QL}}$, with components

$$\hat\mu_{\xi,\mathrm{QL}} = \bar x\,, \tag{3.269}$$

$$\hat\sigma^2_{\xi,\mathrm{QL}} = \frac{1}{n}\sum_{i=1}^n (x_i - \bar x)^2 - \sigma_\delta^2\,. \tag{3.270}$$

The estimators are strongly consistent and asymptotically normal. (Note that instead of the estimator (3.270), one can take an unbiased estimator $S_\xi^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar x)^2 - \sigma_\delta^2$, but at the same time it is asymptotically equivalent to the above estimator $\hat\sigma^2_{\xi,\mathrm{QL}}$.) Denote by $\hat{\mathrm{s}}^{(\beta)}_{\mathrm{QL}}$ and $\hat{\mathrm{s}}^{(\sigma_\varepsilon^2)}_{\mathrm{QL}}$ the estimating functions (3.135) and (3.230), where instead of $\mu(x)$, $M(x)$, and $v(x;\beta,\sigma_\varepsilon^2)$, we substitute the following:

$$\hat\mu(x) = \mu(x)\big|_{\gamma=\hat\gamma_{\mathrm{QL}}}\,, \quad M(x) = M(x)|_{\gamma=\hat\gamma_{\mathrm{QL}}}\,, \quad \hat v(x;\beta,\sigma_\varepsilon^2) = v(x;\beta,\sigma_\varepsilon^2)\big|_{\gamma=\hat\gamma_{\mathrm{QL}}}\,. \tag{3.271}$$

The estimators $\hat\beta_{\mathrm{QL}}$ and $\hat\sigma^2_{\varepsilon,\mathrm{QL}}$ are now defined according to Definition 3.26, where the new functions $\hat{\mathrm{s}}^{(\beta)}_{\mathrm{QL}}$ and $\hat{\mathrm{s}}^{(\sigma_\varepsilon^2)}_{\mathrm{QL}}$ are utilized instead of the former estimating functions $s^{(\beta)}_{\mathrm{QL}}$ and $s^{(\sigma_\varepsilon^2)}_{\mathrm{QL}}$.

**Asymptotic normality of augmented estimator**

**Theorem 3.31.** *Let the only parameter $\sigma_\delta^2$ be known in the model* (3.6) *and* (3.3) *under the conditions* (i)–(iv), *and moreover it is also known that $\sigma_\varepsilon^2 \in (\sigma_1^2, \sigma_2^2)$, with $\sigma_1^2 > 0$. Denote by $b = (b_i)_{i=0}^k$ the true values of polynomial coefficients, and by $s_\varepsilon^2$ the true value of the parameter $\sigma_\varepsilon^2$. Then the following statements are valid:*

(a) *for any $R > \|b\|$, the system of equations $\hat{\mathrm{S}}^{(\beta)}_{\mathrm{QL}} = 0$, $\hat{\mathrm{S}}^{(\sigma_\varepsilon^2)}_{\mathrm{QL}} = 0$,* eventually *has a unique solution on $\bar B_R \times [\sigma_1^2, \sigma_2^2]$, where the ball $\bar B_R$ is given in* (3.137),

(b)

$$\hat\beta_{\mathrm{QL}} \xrightarrow{\mathbf{P}1} b\,, \quad \hat\sigma^2_{\varepsilon,\mathrm{QL}} \xrightarrow{\mathbf{P}1} s_\varepsilon^2\,, \quad \text{as } n \to \infty\,, \tag{3.272}$$

(c)

$$\sqrt{n}(\hat{\beta}_{\mathrm{QL}} - b) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \tilde{\Sigma}_\beta)\,, \tag{3.273}$$

$$\tilde{\Sigma}_\beta = A_\beta^{-1} + A_\beta^{-1} A_{\beta\gamma} B_\gamma A_{\beta\gamma}^{\mathrm{T}} A_\beta^{-1}\,, \tag{3.274}$$

$$A_\beta = \mathbf{E}\frac{\mu\mu^{\mathrm{T}}}{v(x; b, s_\varepsilon^2, \gamma)}\,, \tag{3.275}$$

$$A_{\beta\gamma} = -\mathbf{E}\frac{\partial s_{\mathrm{QL}}^{(\beta)}}{\partial \gamma^{\mathrm{T}}}(y, x; b, \sigma_\varepsilon^2, \gamma)\,, \tag{3.276}$$

*where* $B_\gamma = cov s_{\mathrm{QL}}^{(\gamma)}(x, \gamma)$ *is the ACM of the estimator* $\hat{\gamma}_{\mathrm{QL}}$ *defined in* (3.269) *and* (3.270).

**Remark 3.32.** Formula (3.274) was obtained in Kukush et al. (2005b) and Shklyar et al. (2007). It shows that the ACM for the estimator $\hat{\beta}_{\mathrm{QL}}$ increases (in Loewner sense) if the nuisance parameter $\gamma$ is unknown. (If $\gamma$ is known, the ACM of the estimator is equal to $A_\beta^{-1} \le \tilde{\Sigma}_\beta$.)

*Proof of the theorem.* The statements (a) and (b) are proved in the same way as Theorem 3.27, and in so doing the strong consistency of the estimator $\hat{\gamma}_{\mathrm{QL}}$ is exploited.
(c) Estimators (3.269) and (3.270) correspond to the estimating function

$$s_{\mathrm{QL}}^{(\gamma)} = \begin{pmatrix} x - \mu_\xi \\ (x - \mu_\xi)^2 - \sigma_\delta^2 - \sigma_\xi^2 \end{pmatrix}, \tag{3.277}$$

i.e., the estimator $\hat{\gamma}_{\mathrm{QL}}$ is defined as a solution to a system of equations

$$\frac{1}{n}\sum_{i=1}^{n} s_{QL}^{(\gamma)}(x_i; \mu_\xi, \sigma_\xi^2) = 0\,, \quad \mu_\xi \in \mathbf{R}\,, \quad \sigma_\xi^2 > 0\,. \tag{3.278}$$

Then $\hat{\theta}_{\mathrm{QL}}$ is defined by the estimating function

$$s_{\mathrm{QL}}^{(\theta)} = \begin{pmatrix} s_{\mathrm{QL}}^{(\beta)} \\ s_{\mathrm{QL}}^{(\sigma_\varepsilon^2)} \\ s_{\mathrm{QL}}^{(\gamma)} \end{pmatrix}. \tag{3.279}$$

By the sandwich formula (see Appendix A2; the formula can be applied due to the strong consistency of $\hat{\theta}_{\mathrm{QL}}$), we have for the true value $\theta = (b^{\mathrm{T}}, s_\varepsilon^2, \gamma^{\mathrm{T}})^{\mathrm{T}}$:

$$\sqrt{n}(\hat{\theta}_{\mathrm{QL}} - \theta) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \Sigma_\theta)\,, \tag{3.280}$$

$$\Sigma_\theta = A_\theta^{-1} B_\theta A_\theta^{-\mathrm{T}}\,. \tag{3.281}$$

Here

$$A_\theta = -\mathbf{E}_\theta \frac{\partial s_{\mathrm{QL}}^{(\theta)}}{\partial \theta^{\mathrm{T}}}(y, x; \theta) = \begin{pmatrix} A_\beta & 0 & A_{\beta\gamma} \\ A_{\sigma_\varepsilon^2\beta} & 1 & A_{\sigma_\varepsilon^2\gamma} \\ 0 & 0 & I_2 \end{pmatrix}, \tag{3.282}$$

$I_2$ is the identity matrix of size $2 \times 2$,

$$A_\beta = -\mathbf{E}_\theta \frac{\partial s_{QL}^{(\beta)}}{\partial \beta^{\mathrm{T}}}(y, x; \theta) = \mathbf{E} \frac{\mu \mu^{\mathrm{T}}}{v(x, \theta)} , \tag{3.283}$$

$$A_{\beta\gamma} = -\mathbf{E}_\theta \frac{\partial s_{QL}^{(\beta)}}{\partial \gamma^{\mathrm{T}}}(y, x; \theta) . \tag{3.284}$$

The matrix $A_\theta^{-1}$ exists and has the form

$$A_\theta^{-1} = \begin{pmatrix} A_\beta^{-1} & 0 & Y \\ X & 1 & Z \\ 0 & 0 & I_2 \end{pmatrix} , \tag{3.285}$$

$$X = -A_{\sigma_\varepsilon^2 \beta} A_\beta^{-1} , \quad Y = -A_\beta^{-1} A_{\beta\gamma} , \quad Z = -A_{\sigma_\varepsilon^2 \beta} Y - A_{\sigma_\varepsilon^2 \gamma} . \tag{3.286}$$

The middle part of the "sandwich" (3.281) is equal to

$$B_\theta = \mathbf{E}_\theta \operatorname{cov} s_{\mathrm{QL}}^{(\theta)}(y, x; \theta) = \begin{pmatrix} B_\beta & B_{\beta\sigma_\varepsilon^2} & 0 \\ B_{\sigma_\varepsilon^2 \beta} & B_{\sigma_\varepsilon^2} & 0 \\ 0 & 0 & B_\gamma \end{pmatrix} , \tag{3.287}$$

$$B_\beta = \mathbf{E}_\theta \operatorname{cov} s_{\mathrm{QL}}^{(\beta)}(y, x; \theta) = A_\beta = \mathbf{E} \frac{\mu \mu^{\mathrm{T}}}{v(x; \theta)} , \tag{3.288}$$

$$B_{\beta\sigma_\varepsilon^2} = \mathbf{E}_\theta s_{\mathrm{QL}}^{(\beta)} s_{\mathrm{QL}}^{(\sigma_\varepsilon^2)} = B_{\sigma_\varepsilon^2 \beta}^{\mathrm{T}} , \quad B_\gamma = \operatorname{cov}_\theta s_{\mathrm{QL}}^{(\gamma)} . \tag{3.289}$$

By the way according to the sandwich formula,

$$\sqrt{n}(\hat{\gamma}_{\mathrm{QL}} - \gamma) \xrightarrow{\mathrm{d}} \mathrm{N}(0, I_2^{-1} B_\gamma I_2^{-1}) = \mathrm{N}(0, B_\gamma) , \tag{3.290}$$

so that $B_\gamma$ is indeed the ACM of the estimator $\hat{\gamma}_{\mathrm{QL}}$ as stated in the theorem.

Then by the formulas (3.281), (3.285), and (3.287) we have

$$\begin{aligned} \Sigma_\theta &= \begin{pmatrix} A_\beta^{-1} & 0 & Y \\ X & 1 & Z \\ 0 & 0 & I_2 \end{pmatrix} \begin{pmatrix} B_\beta & * & 0 \\ * & * & 0 \\ 0 & 0 & B_\gamma \end{pmatrix} \begin{pmatrix} A_\beta^{-1} & X^{\mathrm{T}} & 0 \\ 0 & 1 & 0 \\ Y^{\mathrm{T}} & Z^{\mathrm{T}} & I_2 \end{pmatrix} = \\ &= \begin{pmatrix} A_\beta^{-1} B_\beta & * & Y B_\gamma \\ * & * & Z B_\gamma \\ 0 & 0 & B_\gamma \end{pmatrix} \begin{pmatrix} A_\beta^{-1} & X^{\mathrm{T}} & 0 \\ 0 & 1 & 0 \\ Y^{\mathrm{T}} & Z^{\mathrm{T}} & I_2 \end{pmatrix} , \end{aligned} \tag{3.291}$$

$$\Sigma_\theta = \begin{pmatrix} A_\beta^{-1} B_\beta A_\beta^{-1} + Y B_\gamma Y & * & * \\ * & * & * \\ * & * & * \end{pmatrix} . \tag{3.292}$$

The written out block of matrix $\Sigma_\theta$ is the ACM of the estimator $\hat{\beta}_{\mathrm{QL}}$. And taking into account formulas (3.288) and (3.286), the ACM is equal to

$$\tilde{\Sigma}_\beta = A_\beta^{-1} + A_\beta^{-1} A_{\beta\gamma} B_\gamma A_{\beta\gamma}^{\mathrm{T}} A_\beta^{-1} . \tag{3.293}$$

Finally, we mention that the matrix $B_\theta$ is nonsingular, because it is block-diagonal with nonsingular blocks $B_\gamma$ (this is strightforward) and

$$\begin{pmatrix} B_\beta & B_{\beta\sigma_\varepsilon^2} \\ B_{\sigma_\varepsilon^2\beta} & B_{\sigma_\varepsilon^2} \end{pmatrix} > 0 \tag{3.294}$$

(see Theorem 3.27). The nonsingularity of $B_\theta$ justifies application of the sandwich formula to the estimator $\hat{\theta}_{\mathrm{QL}}$. The theorem is proved. □

**Iterative procedure for computation of the estimate**

To compute the estimates $\hat{\beta}_{\mathrm{QL}}$ and $\hat{\sigma}^2_{\varepsilon,\mathrm{QL}}$, when $\gamma$ is unknown, an iterative procedure is used similar to the one described in Section 3.2.2. It is only necessary, instead of the functions $\phi_n$ and $\psi_n$, to apply the functions:

$$\hat{\phi}_n = \phi_n\big|_{\gamma=\hat{\gamma}_{\mathrm{QL}}}\,, \quad \hat{\psi}_n = \psi_n\big|_{\gamma=\hat{\gamma}_{\mathrm{QL}}}\,. \tag{3.295}$$

In particular (see formula (3.259)),

$$\hat{\phi}_n(\beta,\sigma_\varepsilon^2) := \left(\frac{1}{n}\sum_{i=1}^n \frac{\hat{\mu}(x_i)\hat{\mu}^{\mathrm{T}}(x_i)}{\hat{v}(x_i,\beta,\sigma_\varepsilon^2)}\right)^{-1} \times \frac{1}{n}\sum_{i=1}^n \frac{\hat{\mu}(x_i)y_i}{\hat{v}(x_i,\beta,\sigma_\varepsilon^2)}\,. \tag{3.296}$$

Here, $\hat{\mu}$ and $\hat{v}$ are the function from formula (3.271).

For such a modified iterative procedure, the statement of Theorem 3.29 on the convergence of iterations to the QL estimator remains true.

**Asymptotic efficiency of the estimator**

In the framework of Remark 3.32, it is unclear whether the estimator $\hat{\beta}_{\mathrm{QL}}$ is more asymptotically efficient than $\hat{\beta}_{\mathrm{C}}$, under unknown nuisance parameter $\gamma$ (see Theorem 3.22 for the case of known $\gamma$). The answer is positive.

**Theorem 3.33.** *Assume the conditions of Theorem* 3.31. *The model in* (3.5) *can handle polynomials of degree up to k, but suppose that the true polynomial is of degree s (that is* $0 \le s \le k$*; for b the true value of the parameter* $\beta$, $b_s \neq 0$*;* $b_i = 0$ *for all i such that* $s < i \le k$*). Let* $\Sigma_{\mathrm{C}}$ *be the ACM of estimator* $\hat{\beta}_{\mathrm{C}}$*, and* $\Sigma_{\mathrm{QL}}$ *be the ACM of estimator* $\hat{\beta}_{\mathrm{QL}}$ *under unknown* $\sigma_\varepsilon^2$, $\sigma_\xi^2$, *and* $\mu_\xi$ *(both matrices are of size* $(k+1)\times(k+1)$*). Then:*

(a) *in all the cases* $\Sigma_{\mathrm{QL}} \le \Sigma_{\mathrm{C}}$,
(b) *if* $b_1 = b_2 = \cdots = b_k = 0$, *then* $\Sigma_{\mathrm{QL}} = \Sigma_{\mathrm{C}}$,
(c) *if* $s = 1$, *then* $\operatorname{rank}(\Sigma_{\mathrm{C}} - \Sigma_{\mathrm{QL}}) = k - 1$,
(d) *if* $s = 2$, *then* $\operatorname{rank}(\Sigma_{\mathrm{C}} - \Sigma_{\mathrm{QL}}) = k$,
(e) *if* $s \ge 3$, *then* $\Sigma_{\mathrm{QL}} < \Sigma_{\mathrm{C}}$.

*Proof* can be found in Kukush et al. (2009).

**Remark 3.34.** The statement (c) implies that in the case $k = s = 1$, it holds that $\Sigma_{\mathrm{QL}} = \Sigma_{\mathrm{C}}$. This is natural because one can show that in the linear model under unknown nuisance parameters, $\hat{\beta}_{\mathrm{QL}}$ and $\hat{\beta}_{\mathrm{C}}$ coincide.

**Theorem 3.35.** *Under the conditions of Theorem* 3.33*, the following holds true:*
(a)

$$\Sigma_{\mathrm{QL}} = \sigma_\varepsilon^2(\mathbf{E}\rho\rho^{\mathrm{T}})^{-1} + O(\sigma_\delta^2)\,, \quad \text{as } \sigma_\delta^2 \to 0\,, \tag{3.297}$$

(b)

$$\Sigma_{\mathrm{C}} = \Sigma_{\mathrm{QL}} + O(\sigma_\delta^4)\,, \quad \text{as } \sigma_\delta^2 \to 0\,. \tag{3.298}$$

*Proof* is given in Kukush et al. (2005b).

As one can see (cf. Theorem 3.23), these two statements are valid for both known and unknown nuisance parameters.

Consider the situation where $\sigma_\varepsilon^2$ is known, and the nuisance parameter $\gamma$ remains unknown. In this situation, we estimate the parameter

$$\upsilon = \begin{pmatrix} \beta \\ \gamma \end{pmatrix} = \begin{pmatrix} \beta \\ \mu_\xi \\ \sigma_\xi^2 \end{pmatrix} \tag{3.299}$$

using the estimating function

$$s_{\mathrm{QL}}^{(\upsilon)} = \begin{pmatrix} s_{\mathrm{QL}}^{(\beta)} \\ s_{\mathrm{QL}}^{(\gamma)} \end{pmatrix}\,, \tag{3.300}$$

i.e., the component $s_{\mathrm{QL}}^{(\sigma_\varepsilon^2)}$ is deleted from the estimating function (3.279). Then the estimator $\hat{\upsilon}_{\mathrm{QL}}$ remains asymptotically normal, and moreover for a component $\hat{\beta}_{\mathrm{QL}}$, the ACM is specified with the formula (3.274). As we can see, knowledge or lack of knowledge of $\sigma_\varepsilon^2$ does not affect the asymptotic efficiency of estimator $\hat{\beta}_{\mathrm{QL}}$.

The estimating function (3.300) is linear in $y$ and unbiased. Like in Section 3.2.1, consider arbitrary unbiased estimating function being linear in $y$:

$$s_{\mathrm{L}}(y, x; \upsilon) = g(x, \upsilon)y - h(x, \upsilon)\,, \quad y, x \in \mathbf{R}\,, \quad \upsilon \in \mathbf{R}^{k+1} \times \mathbf{R} \times (0, +\infty)\,. \tag{3.301}$$

Here, $g$ and $h$ are the Borel measurable functions valued in $\mathbf{R}^{k+3}$, which can involve $\sigma_\varepsilon^2$. Under mild conditions, as it is demonstrated in Appendix A1, the estimating function (3.301) constructed from the observations $y_1, x_1, \ldots, y_n, x_n$, yields the estimator $\hat{\upsilon}_{\mathrm{L}}$ which is consistent and asymptotically normal, and moreover its ACM is given by a sandwich formula like (3.190). Denote by $L$ the class of all estimating functions (3.301) for which everything just described above is valid. The class $L$ contains the estimating function (3.300) and also the estimation function

$$s_{\mathrm{C}}^{(\upsilon)} = \begin{pmatrix} s_{\mathrm{C}}^{(\beta)} \\ s_{\mathrm{QL}}^{(\gamma)} \end{pmatrix}\,. \tag{3.302}$$

**Theorem 3.36** (On asymptotic efficiency of the QL estimator). *Let the parameters $\sigma_\delta^2$ and $\sigma_\varepsilon^2$ be known but $\beta$ and $\gamma$ unknown in the model* (3.6) *and* (3.3) *under the conditions* (i)–(iv). *Consider an arbitrary estimating function $s_{\mathrm{L}} \in L$, and let $\Sigma_{\mathrm{L}}^{(\upsilon)}$ and $\Sigma_{\mathrm{QL}}^{(\upsilon)}$ be the ACMs of the estimators $\hat{\upsilon}_{\mathrm{L}}$ and $\hat{\upsilon}_{\mathrm{QL}}$, respectively. Then*

$$\Sigma_{\mathrm{QL}}^{(\upsilon)} \le \Sigma_{\mathrm{L}}^{(\upsilon)}\,. \tag{3.303}$$

*If in addition* $\Sigma_{\mathrm{QL}}^{(v)} = \Sigma_{\mathrm{L}}^{(v)}$, *for all the true values of the parameter* $v$, *then almost surely,* $\hat{v}_{\mathrm{L}} = \hat{v}_{\mathrm{QL}}$.

*Proof* can be found in Kukush et al. (2009).

In particular, statement (a) of Theorem 3.33 follows from inequality (3.303). Thus, the QL estimator is asymptotically efficient in a broad class of estimators.

**Remark 3.37.** A number of results from Chapter 3 do not require the condition (ii) on normal distribution of errors $\varepsilon_i$. Instead, it is often enough to require a weaker assumption (vii) about the centrality of errors and finiteness of the second moment $\mathbf{E}\varepsilon_i^2$.

**Remark 3.38.** The QL estimator of the parameter $\beta$ can be constructed under the assumption that $\xi_i$ are identically distributed and the distribution of $\xi_i$ is a mixture of several normal distributions (see Section 1.4.3).

# 4 Nonlinear and generalized linear models

In previous chapters, we studied some types of nonlinear regression (1.1) and considered the binary logistic regression models (1.2) and (1.3). It turns out that in these models, conditional distribution of $y$ given $\xi$ belongs to the so-called exponential family.

## 4.1 Exponential family of densities

Let $Y$ be a random variable, with distribution depending on the parameter $\eta \in I$. Here $I = \mathbf{R}$ or $I$ is a given open interval on the real line being either finite or infinite. Let $\mu$ be a *$\sigma$-finite measure* on the Borel $\sigma$-algebra $B(\mathbf{R})$; $\sigma$-finiteness of a measure means that the real line can be decomposed as $\mathbf{R} = \bigcup_{i=1}^{\infty} A_n$, where $A_n$ are Borel measurable sets, moreover $\mu(A_n) < \infty$, for all $n \geq 1$.

For example, the Lebesgue measure $\lambda_1$ on $B(\mathbf{R})$ is $\sigma$-finite because $\mathbf{R} = \bigcup_{n=1}^{\infty} [-n, n]$ and $\lambda_1([-n, n]) = 2n < \infty, \quad n \geq 1$.

Suppose that $Y$ has a density function $\rho(y|\eta)$ w.r.t. the measure $\mu$. This means that for every set $A \in B(\mathbf{R})$,

$$\mathbf{P}\{Y \in A|\eta\} = \int_A \rho(y|\eta)\, d\mu(y)\,. \tag{4.1}$$

The latter integral is the Lebesgue integral (Halmos, 2013).

**Definition 4.1.** A density function $\rho(y|\eta)$, $\eta \in I$, belongs to the *exponential family* if

$$\rho(y|\eta) = \exp\left\{\frac{y\eta - C(\eta)}{\phi} + c(y, \phi)\right\}\,, \quad \eta \in I\,. \tag{4.2}$$

Here $C(\cdot) \in C^2(I)\,, \quad C''(\eta) > 0$ for all $\eta$; $\phi > 0$ is the so-called *dispersion parameter*; $c(y, \phi)$ is a Borel measurable function of two variables.

It appears that the conditional mean $\mathbf{E}(Y|\eta)$ and conditional variance $\mathbf{D}(Y|\eta)$ are expressed through the function $C(\cdot)$ and $\phi$; moreover, $\mathbf{E}(Y|\eta)$ does not depend on $\phi$.

**Lemma 4.2.** *Assume that $\rho(y|\eta)$, $\eta \in I$, belongs to the exponential family* (4.2). *Then for each $\eta \in I$,*

$$\mathbf{E}(Y|\eta) = C'(\eta)\,. \tag{4.3}$$

$$\mathbf{D}(Y|\eta) = \phi C''(\eta)\,. \tag{4.4}$$

*Proof.* Putting $A = \mathbf{R}$ in (4.1), we have the identity

$$\int_{\mathbf{R}} \exp\left\{\frac{y\eta - C(\eta)}{\varphi} + c(y, \varphi)\right\} d\mu(y) = 1\,, \quad \eta \in I\,. \tag{4.5}$$

Differentiating both sides of (4.5) with respect to $\eta$ (we assume that the standard conditions allowing to apply the Leibniz rule for differentiation of an integral w.r.t. a parameter are fulfilled):

$$\int_{\mathbf{R}} \frac{\partial \rho(y|\eta)}{\partial \eta} d\mu(y) = 0 \,, \tag{4.6}$$

$$\int_{\mathbf{R}} \rho(y|\eta) \cdot \left( \frac{y - C'(\eta)}{\phi} \right) d\mu(y) = 0 \,, \tag{4.7}$$

$$\int_{\mathbf{R}} y\rho(y|\eta) d\mu(y) = C'(\eta) \int_{\mathbf{R}} \rho(y|\eta) d\mu(y) \,, \tag{4.8}$$

$$\mathbf{E}(Y|\eta) = C'(\eta) \,. \tag{4.9}$$

To prove formula (4.4), let us differentiate the identity having been found above:

$$\int_{\mathbf{R}} y\rho(y|\eta) d\mu(y) = C'(\eta) \,, \quad \eta \in I \,. \tag{4.10}$$

Again, using the Leibniz rule we get

$$\int_{\mathbf{R}} y\rho(y|\eta) \cdot \left( \frac{y - C'(\eta)}{\phi} \right) d\mu(y) = C''(\eta) \,, \tag{4.11}$$

$$\int_{\mathbf{R}} y^2 \rho(y|\eta) d\mu(y) - C'(\eta) \cdot \int_{\mathbf{R}} y\rho(y|\phi) d\mu(y) = \phi C''(\phi) \,, \tag{4.12}$$

$$\mathbf{E}(Y^2|\eta) - [\mathbf{E}(y|\eta)]^2 = \phi C''(\eta) \,, \tag{4.13}$$

$$\mathbf{D}(Y|\eta) = \phi C''(\eta) \,, \quad \eta \in I \,. \tag{4.14}$$

The lemma is proved. □

We give several examples of exponential families.

**Example 4.3** (Normal distribution). Let $Y \sim \mathrm{N}(m, \sigma^2)$, with $m \in \mathbf{R}$ and $\sigma > 0$. As the measure $\mu$, we take the Lebesgue measure $\lambda_1$ on the Borel $\sigma$-algebra $B(\mathbf{R})$. Then, for $Y$, the density function $\rho(y|m)$ with respect to $\lambda_1$ is just the ordinary probability density function (pdf),

$$\rho(y|m) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{(y-m)^2}{2\sigma^2}} \,, \quad y \in \mathbf{R} \,. \tag{4.15}$$

Transform it to the form (4.2)

$$\rho(y|m) = \exp\left\{ \frac{ym - m^2/2}{\sigma^2} - \left( \frac{y^2}{2\sigma^2} + \ln(\sqrt{2\pi}\sigma) \right) \right\} \,. \tag{4.16}$$

Thus, the pdf satisfies Definition 4.1. Here

$$\eta = m \,, \quad \phi = \sigma^2 \,, \quad C(\eta) = \frac{1}{2}\eta^2 \,, \quad c(y, \phi) = -\left( \frac{y^2}{2\phi} + \ln(\sqrt{2\pi\phi}) \right) \,. \tag{4.17}$$

For the normal density function, it holds that $\phi = \sigma^2 = \mathbf{D}Y$, and this fact justifies the name of $\phi$ as dispersion parameter. We apply Lemma 4.2:

$$\begin{aligned} \mathbf{E}Y &= C'(\eta) = \eta = m \,, \\ \mathbf{D}Y &= \phi C''(\eta) = \phi = \sigma^2 \,. \end{aligned} \tag{4.18}$$

So, we have come to the correct results.

**Example 4.4** (Poisson distribution). Suppose, $Y$ has Poisson distribution Pois($\lambda$), with parameter $\lambda = \mathrm{e}^{\eta}$, $\eta \in \mathbf{R}$. This means that $Y$ takes nonnegative values and

$$\mathbf{P}\{Y = y\} = \frac{\mathrm{e}^{-\lambda}}{y!}\lambda^y \,, \quad y = 0, 1, 2, \dots \,. \tag{4.19}$$

As the measure $\mu$, we use the counting measure concentrated at integer points,

$$\mu(A) = \ |A \cap (N \cup \{0\})| \ , \quad A \in B(\mathbf{R}) \,, \tag{4.20}$$

i.e., $\mu(A)$ is the number of points in the latter intersection. Then expression (4.19) specifies the density function $\rho(y|\eta)$ w.r.t. the measure $\mu$

$$\rho(y|\eta) = \exp\{y \ln \lambda - \lambda - \ln(y!)\} = \exp\left\{y\eta - \mathrm{e}^{\eta} - \ln(y!)\right\} \,, \quad y \in N \cup \{0\} \,. \tag{4.21}$$

Thus, the density function satisfies Definition 4.1. Here $\eta \in \mathbf{R}$, $\phi = 1$, $C(\eta) = \mathrm{e}^{\eta}$, and $c(y, \phi) = c(y) = -\ln(y!)$.

We apply Lemma 4.2:

$$\mathbf{E}Y = C'(\eta) = \mathrm{e}^{\eta} = \lambda \,, \quad \mathbf{D}Y = \phi C''(\eta) = \mathrm{e}^{\eta} = \lambda \,. \tag{4.22}$$

Thus, we have obtained the correct result.

**Example 4.5** (Gamma distribution). Suppose, $Y$ has the gamma distribution $\Gamma(\alpha, \lambda)$, with the shape parameter $\alpha$ and scale parameter $\lambda > 0$. This means that $Y$ takes positive values, and the pdf of $Y$ is equal to

$$\rho(y|\lambda) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1}\mathrm{e}^{-\lambda y} \,, \quad y > 0 \,. \tag{4.23}$$

The parameter $\alpha$ is assumed fixed.

As the measure $\mu$, we take the Lebesgue measure concentrated on the positive semiaxis:

$$\mu(A) = \lambda_1(A \cap (0, +\infty)) \,, \quad A \in B(\mathbf{R}) \,. \tag{4.24}$$

Then expression (4.23) will give us the density function of $Y$ w.r.t. the measure $\mu$. Transform $\rho(y|\lambda)$ to the exponential form

$$\rho(y|\lambda) = \exp\{-\lambda y + \alpha \ln \lambda + ((\alpha - 1) \ln y - \ln \Gamma(\alpha))\} \,. \tag{4.25}$$

Set

$$\eta = -\lambda \,, \quad \eta < 0 \,. \tag{4.26}$$

Then

$$\rho(y|\eta) = \exp\{y\eta - (-\alpha \ln(-\eta)) + c(y)\}\,, \quad y > 0\,, \tag{4.27}$$

$$c(y) = (\alpha - 1)\ln y - \ln \Gamma(\alpha)\,, \quad y > 0\,. \tag{4.28}$$

The density function $\rho(y|\lambda)$, $\eta < 0$, belongs to the exponential family, with

$$\phi = 1\,, \quad C(\eta) = -\alpha \ln(-\eta)\,. \tag{4.29}$$

Let us utilize Lemma 4.2:

$$\mathbf{E}Y = C'(\eta) = -\frac{\alpha}{\eta} = \frac{\alpha}{\lambda}\,, \quad \mathbf{D}Y = C''(\eta) = \frac{\alpha}{\eta^2} = \frac{\alpha}{\lambda^2}\,. \tag{4.30}$$

**Example 4.6** (Exponential distribution). Consider the distribution $\Gamma(\alpha, \lambda)$, with $\alpha = 1$. This is exponential distribution, with the parameter $\lambda > 0$:

$$\rho(y|\lambda) = \lambda \mathrm{e}^{-\lambda y}\,, \quad y > 0\,. \tag{4.31}$$

The density function satisfies (4.2), where

$$\eta = -\lambda\,, \quad \phi = 1\,, \quad C(\eta) = -\ln(-\eta)\,, \quad c(y, \phi) = 0\,. \tag{4.32}$$

According to formulas (4.30), we have

$$\mathbf{E}Y = \frac{1}{\lambda}\,, \quad \mathbf{D}Y = \frac{1}{\lambda^2}\,. \tag{4.33}$$

**Example 4.7** (Binary distribution, with $\lambda = \mathrm{e}^{\eta}$). Suppose that $Y$ has the binary distribution (cf. (1.2)):

$$\mathbf{P}\{Y = 1\} = \frac{\lambda}{1+\lambda}\,, \quad \mathbf{P}\{Y = 0\} = \frac{1}{1+\lambda}\,. \tag{4.34}$$

Here $\lambda > 0$. The distribution is very common in epidemiology (see discussion at the beginning of Chapter 1).

As the measure $\mu$, we take a point measure concentrated at the points 0 and 1:

$$\mu(A) = \mathbf{I}_A(0) + \mathbf{I}_A(1)\,, \quad A \in B(\mathbf{R})\,. \tag{4.35}$$

Here $\mathbf{I}_A$ is the indicator function,

$$\mathbf{I}_A(x) = \begin{cases} 1\,, & x \in A\,, \\ 0\,, & x \notin A\,. \end{cases} \tag{4.36}$$

With respect to $\mu$, the density function of $Y$ is given as follows:

$$\rho(y|\lambda) = \left(\frac{\lambda}{1+\lambda}\right)^{y} \left(\frac{1}{1+\lambda}\right)^{1-y}\,, \quad y = 0;\ 1\,; \tag{4.37}$$

$$\rho(y|\lambda) = \exp\{y \ln \lambda - \ln(1+\lambda)\}\,, \quad y = 0;\ 1\,. \tag{4.38}$$

Put $\eta = \ln\lambda \in \mathbf{R}$:

$$\rho(y|\eta) = \exp\{y\eta - \ln(1+\mathrm{e}^{\eta})\}\,, \quad y = 0;\ 1\,. \tag{4.39}$$

For this density function, (4.2) holds with

$$\phi = 1\,, \quad C(\eta) = \ln(1+\mathrm{e}^{\eta})\,, \quad c(y,\phi) = 0\,. \tag{4.40}$$

According to Lemma 4.2, we have

$$\mathbf{E}Y = C'(\eta) = \frac{\mathrm{e}^{\eta}}{1+\mathrm{e}^{\eta}} = \frac{\lambda}{1+\lambda}\,, \quad \mathbf{D}Y = C''(\eta) = \frac{\mathrm{e}^{\eta}}{(1+\mathrm{e}^{\eta})^2} = \frac{\lambda}{(1+\lambda)^2}\,. \tag{4.41}$$

The latter equality can be transformed to the classical form

$$\mathbf{D}Y = \mathbf{P}\{Y=1\}\cdot\mathbf{P}\{Y=0\}\,. \tag{4.42}$$

Examples 4.3–4.7 demonstrate that various distributions belong to the exponential family.

## 4.2 Regression model with exponential family of densities and measurement errors

Given the exponential family (4.2), suppose that the parameter $\eta$ depends on the regressor $\xi$ and unknown parameter $\beta$,

$$\eta = \eta(\xi,\beta)\,, \quad \xi \in \mathbf{R}\,, \quad \beta \in \Theta_\beta \subset \mathbf{R}^p\,. \tag{4.43}$$

Here $\Theta_\beta$ is a given parameter set, and the function (4.43) is assumed to be smooth enough. Formulas (4.2) and (4.43) define the regression of $Y$ on $\xi$, and $\beta$ is a regression parameter, while $\phi$ is a nuisance parameter.

**Definition 4.8.** The abovementioned regression model of $Y$ on $\xi$ is called *generalized linear model* (GLM), if the relation (4.43) has the form

$$\eta = h(\beta_0+\beta_1\xi)\,, \quad \xi \in \mathbf{R}\,, \quad \beta = (\beta_0,\beta_1)^{\mathrm{T}} \in \Theta_\beta \subset \mathbf{R}^2\,. \tag{4.44}$$

Here $h$ is a given smooth function.

Usually in the GLM, the parameter set $\Theta_\beta = \mathbf{R}^2$, i.e., there is no prior information about the parameters $\beta_0$ and $\beta_1$.

### 4.2.1 Maximum likelihood estimator in the absence of measurement errors

Let $\{(y_i,\xi_i),\ i=\overline{1,n}\}$ be independent observations in the model (4.2) and (4.43). This implies that the observed couples are stochastically independent; moreover, the density function of $y_i$ w.r.t. the measure $\mu$ is given by equality (4.2), with $\eta = \eta_i$ and $\eta_i = \eta(\xi_i,\beta)$.

The values $\xi_i$ are nonrandom in the functional case, and they are independent identically distributed random variables in the structural case.

In both cases it holds that

$$\rho(y|\xi,\beta) = \rho(y|\eta(\xi,\beta)) \,. \tag{4.45}$$

Then

$$\frac{\partial \rho(y|\xi,\beta)}{\partial \beta} = \frac{1}{\phi}(y\cdot\eta'_\beta - C'(\eta)\cdot\eta'_\beta)\cdot\rho(y|\xi,\beta)\,, \quad \eta=\eta(\xi,\beta)\,. \tag{4.46}$$

The score function $s_{\mathrm{ML}}$ is equal to

$$s_{\mathrm{ML}}(y,\xi;\beta) = \eta'_\beta\cdot(y-C'(\eta))\,, \quad \eta=\eta(\xi;\beta)\,. \tag{4.47}$$

If the parameter set $\Theta_\beta$ is open, then the maximum likelihood estimator (MLE) $\hat\beta_{\mathrm{ML}}$ is a measurable solution to the equation

$$\sum_{i=1}^{n} s_{\mathrm{ML}}(y_i,\xi_i;\beta) = 0\,, \quad \beta\in\Theta_\beta\,. \tag{4.48}$$

Under mild conditions, the estimator is strongly consistent:

$$\hat\beta_{\mathrm{ML}} \xrightarrow{\mathrm{P1}} \beta\,, \quad \text{as } n\to\infty\,. \tag{4.49}$$

It is also asymptotically normal:

$$\sqrt{n}\,(\hat\beta_{\mathrm{ML}}-\beta) \xrightarrow{\mathrm{d}} \mathrm{N}(0,\Phi^{-1})\,. \tag{4.50}$$

Here $\Phi$ is a positive definite matrix of size $p\times p$. In the structural case, $\Phi$ is the so-called *Fisher information matrix*,

$$\Phi = -\mathbf{E}_\beta \frac{\partial s_{\mathrm{ML}}(y,\xi;\beta)}{\partial \beta^{\mathrm{T}}}\,. \tag{4.51}$$

Using (4.47), we evaluate the matrix

$$\Phi = -\mathbf{E}_\beta(y-C'(\eta))\eta''_{\beta\beta} + \mathbf{E}_\beta C''(\eta)\eta'_\beta(\eta'_\beta)^{\mathrm{T}}\,. \tag{4.52}$$

The first term is zero, because

$$\mathbf{E}\mathbf{E}_\beta\left[(y-C'(\eta))\,\eta''_{\beta\beta}|\xi\right] = \mathbf{E}\left\{\mathbf{E}_\beta[y-C'(\eta)|\xi]\cdot\eta''_{\beta\beta}\right\} = 0\,. \tag{4.53}$$

Therefore,

$$\Phi = \mathbf{E}_\beta C''(\eta)\eta'_\beta(\eta'_\beta)^{\mathrm{T}}\,. \tag{4.54}$$

Since $C''>0$, the matrix $\Phi$ is positive definite, if the components of vector $\eta'_\beta$ are linearly independent in the space $L_2(\Omega,\mathbf{P})$ of random variables. Usually, the latter holds true.

Indeed, in the GLM (4.44),

$$\eta'_\beta = h'(\beta_0+\beta_1\xi)\cdot(1;\ \xi)^{\mathrm{T}}\,. \tag{4.55}$$

If $h'(t)\neq 0$, $t\in\mathbf{R}$, and the random variable $\xi$ is not constant, then the components of vector $\eta'_\beta(\xi,\beta)$ are linearly independent in $L_2(\Omega,\mathbf{P})$, and thus, the information matrix (4.54) is positive definite.

**Remark 4.9.** In the functional case (when $\xi_i$ are nonrandom), the matrix $\Phi$ from relation (4.50) is found as follows:

$$\Phi = -\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}_\beta\frac{\partial s_{\mathrm{ML}}(y,\xi_i;\beta)}{\partial\beta^{\mathrm{T}}}\,, \tag{4.56}$$

$$\Phi = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}C''(\eta_i)\eta'_\beta(\eta'_\beta)^{\mathrm{T}}\,,\quad \eta_i=\eta(\xi_i,\beta)\,,\quad \eta'_\beta=\eta'_\beta(\xi_i,\beta)\,. \tag{4.57}$$

Come back to the structural case. Having estimated $\beta$ one can, if necessary, estimate the dispersion parameter $\phi$. From formulas (4.3) and (4.4), we obtain

$$\mathbf{E}(Y-C'(\eta))^2=\phi\cdot\mathbf{E}C''(\eta)\,,\quad \eta=\eta(\xi,\beta)\,. \tag{4.58}$$

Therefore, the unbiased estimating equation for $\phi$ takes the form

$$\sum_{i=1}^{n}(y_i-C'(\eta_i))^2=\phi\sum_{i=1}^{n}C''(\eta_i)\,,\quad \eta_i=\eta(\xi_i,\beta)\,. \tag{4.59}$$

This equation should be considered in tandem with equation (4.48). Then the estimator $\hat\phi$ of the parameter $\phi$ is equal to

$$\hat\phi=\frac{1}{n}\sum_{i=1}^{n}(y_i-C'(\hat\eta_i))^2:\left(\frac{1}{n}\sum_{i=1}^{n}C''(\hat\eta_i)\right)\,,\quad \hat\eta_i=\eta(\xi_i,\hat\beta_{\mathrm{ML}})\,. \tag{4.60}$$

Due to the convergence (4.49) and the SLLM, one can show that a.s.,

$$\lim_{n\to\infty}\hat\phi=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(y_i-C'(\eta_i))^2:\left(\frac{1}{n}\sum_{i=1}^{n}C''(\eta_i)\right)=\frac{\mathbf{E}(Y-C'(\eta))^2}{\mathbf{E}C''(\eta)}=\phi\,, \tag{4.61}$$

i.e., $\hat\phi$ is strongly consistent. This fact holds in the functional case as well.

### 4.2.2 Quasi-likelihood estimator in the presence of measurement errors

Consider the regression model (4.2) and (4.43). Let a surrogate variable $x$ be observed instead of $\xi$

$$x=\xi+\delta\,, \tag{4.62}$$

the random vector $(y,\xi)^{\mathrm{T}}$ and $\delta$ be stochastically independent, and also the pdf $\rho(\delta)$ of the classical error be known. It is assumed that $\mathbf{E}\delta=0$. By the observations $\{(y_i,x_i),\ i=\overline{1,n}\}$, we estimate the model parameters.

Note that in Section 1.4, we presented estimation methods in the model (1.1) and (1.5), which is a particular case of the model (4.2), (4.43), and (4.62). This is due to the fact that the normal pdf belongs to the exponential family (see Example 4.3). Now, we apply these estimation methods to the general model.

The naive estimator $\hat{\beta}_{\text{naive}}$ is defined by the estimating function (4.47), in which the observed values of the surrogate variable $x$ are substituted instead of the unobserved values of the latent variable $\xi$. Thus, $\hat{\beta}_{\text{naive}}$ is a measurable solution to the equation

$$\sum_{i=1}^{n} s_{\text{ML}}(y_i, x_i; \beta) = 0\,, \quad \beta \in \Theta_\beta\,. \tag{4.63}$$

The estimator is not consistent even in the linear measurement error model (Theorem 2.3). The reason for the inconsistency is bias of the estimating function $s_{\text{ML}}(y, x; \beta)$ (yet the unbiasedness of an estimating function is a prerequisite for the consistency of an estimator, see Appendix A1). Indeed,

$$\begin{aligned}\mathbf{E}_\beta s_{\text{ML}}(y, x; \beta) &= \mathbf{E}_\beta \eta'_\beta (y - C'(\eta)) = \mathbf{E}\mathbf{E}_\beta \left[\eta'_\beta (y - C'(\eta))|x, \xi\right] = \\ &= \mathbf{E}\left\{\eta'_\beta \mathbf{E}_\beta \left(y - C'(\eta)|x, \xi\right)\right\} = \mathbf{E}\eta'_\beta (C'(\eta(\xi, \beta)) - C'(\eta(x, \beta))\,.\end{aligned} \tag{4.64}$$

Here $\eta = \eta(x, \beta)$ and $\eta'_\beta = \eta'_\beta(x, \beta)$; we have exploited the indifferentiability of error $\delta$ (Section 1.4.3) in the calculation

$$\mathbf{E}_\beta(y|x, \xi) = \mathbf{E}_\beta(y|\xi) = C'(\eta(\xi, \beta))\,. \tag{4.65}$$

From equality (4.64), it is obvious that in general case, $\mathbf{E}_\beta s_{\text{ML}}(y, x; \beta) \neq 0$, so the estimating function $s_{\text{ML}}(y, x; \beta)$ is biased.

The naive estimator can be used under relatively small measurement error variances $\sigma^2_\delta$, then the asymptotic deviation of the estimate $\hat{\beta}_{\text{naive}}$ from the true value will be small.

To construct the quasi-likelihood (QL) estimator, we write down the conditional mean and conditional variance of $y$ given $x$. To do this, assume the following:

(i) The errors $\delta_i$ are identically distributed, with distribution $\text{N}(0, \sigma^2_\delta)$, where $\sigma^2_\delta$ is positive and known.

(ii) Random variables $\xi_i$ are identically distributed, with distribution $\text{N}(\mu_\xi, \sigma^2_\xi)$, $\sigma^2_\xi > 0$.

Given (i) and (ii), in the model of regressor's observations (4.62), we get

$$\xi \sim \text{N}(\mu_\xi, \sigma^2_\xi)\,, \quad \delta \sim \text{N}(0, \sigma^2_\delta)\,, \quad \xi \perp\!\!\!\perp \delta\,. \tag{4.66}$$

Then the conditional distribution of $\xi$ given $x$ is determined by (1.86). This allows writing down the conditional expectation and conditional variance of $y$ given $x$.

Let $\theta$ be the total vector of unknown parameters , $\theta^{\text{T}} = (\beta^{\text{T}}, \phi, \mu_\xi, \sigma^2_\xi)$. According to formulas (1.59) and (4.9) we obtain

$$\begin{aligned}m(x, \theta) &= \mathbf{E}(y|x) = \mathbf{E}\left[m^*(\xi, \theta)|x\right] = \\ &= \mathbf{E}\left[C'(\eta(\xi, \beta))|x\right] = \mathbf{E}\left[C'(\eta(\mu_1(x) + \tau\gamma, \beta))|x\right]\,.\end{aligned} \tag{4.67}$$

Here $\gamma \sim \text{N}(0, 1)$, $\gamma \perp\!\!\!\perp x$, and the values $\mu_1(x)$ and $\tau^2$ are given in (1.86); $\tau > 0$.

Notice that the conditional mean $m(x, \theta)$ does not depend on $\phi$. In fact, the expectation in (4.67) is taken w.r.t. the variable $y$ under a fixed $x$.

Further, according to equalities (1.65), (4.9), and (4.14), we get

$$\begin{aligned} v(x, \theta) = \mathbf{V}(y|x) &= \phi \mathbf{E}[C''(\eta)|x] + \mathbf{V}[C'(\eta)|x] = \\ &= \phi \mathbf{E}[C''(\eta)|x] + \mathbf{E}[(C'(\eta))^2|x] - m^2(x, \theta) \, . \end{aligned} \tag{4.68}$$

Here $\eta = \eta(\xi, \beta)$; the latter conditional expectations can be rewritten like in equality (4.67). As is seen, the conditional variance $v(x, \theta)$ depends linearly on $\phi$. Because $\phi > 0$ and $C''(\eta) > 0$, then $v(x, \theta) > 0$.

**Estimation of regression parameter only**

Assume now that the parameters $\phi$, $\mu_\xi$, and $\sigma_\xi^2$ are known (remember that $\sigma_\delta^2$ is also assumed known by assumption (i)). The QL estimator $\hat{\beta}_{\mathrm{QL}}$ is defined by formulas (1.71) and (1.70), with $m(x, \beta)$ and $v(x, \beta)$ given in (4.67) and (4.68), respectively. The parameter set $\Theta_\beta$ is assumed to be compact in $\mathbf{R}^p$. Thus, now the estimating function is as follows:

$$s_{\mathrm{QL}}^{(\beta)}(y, x; \beta) = \frac{y - m(x, \theta)}{v(x, \theta)} \cdot \frac{\partial m(x, \theta)}{\partial \beta} \, . \tag{4.69}$$

We write separately the latter column vector of partial derivatives:

$$\frac{\partial m(x, \theta)}{\partial \beta} = \mathbf{E}\left[C''(\eta) \cdot \eta'_\beta | x\right] \, . \tag{4.70}$$

This formula has been obtained from (4.67) by differentiating with respect to the parameter being under the expectation sign (we assume that the regularity conditions being imposed allow to do so); in formula (4.70), the functions $\eta$ and $\eta'_\beta$ are evaluated at point $(\xi, \beta)$.

Under mild conditions, the estimator $\hat{\beta}_{\mathrm{QL}}$ is strongly consistent (see Appendix A1). Indeed,

$$\mathbf{E}_\beta s_{\mathrm{QL}}^{(\beta)}(y, x; \beta) = 0 \tag{4.71}$$

(see calculation (1.77)), and the estimating function $s_{\mathrm{QL}}^{(\beta)}$ is unbiased. In addition, when the strong law of large numbers (SLLN) holds true, we have

$$\frac{1}{n} \sum_{i=1}^{n} s_{\mathrm{QL}}^{(\beta)}(y_i, x_i; b) \xrightarrow{\mathrm{P1}} \mathbf{E}_\beta \frac{y - m(x, b)}{v(x, b)} \cdot \frac{\partial m(x, b)}{\partial \beta} \, , \tag{4.72}$$

$$\frac{1}{n} \sum_{i=1}^{n} s_{\mathrm{QL}}^{(\beta)}(y_i, x_i; b) \xrightarrow{\mathrm{P1}} S_\infty^{(\beta)}(\beta, b) = \mathbf{E} m'_\beta(x, b) \cdot \frac{m(x, \beta) - m(x, b)}{v(x, b)} \, . \tag{4.73}$$

Moreover, the convergence is uniform in $\beta \in \Theta_\beta$, almost surely.

The asymptotic equation,

$$S_\infty^{(\beta)}(\beta, b) = 0 \, , \quad b \in \Theta_\beta \, , \tag{4.74}$$

should have a unique solution $b = \beta$. This is our demand on the regression model, which holds, in particular, for polynomial regression (see the proof of Theorem 3.13). Under the demand, it holds indeed that $\hat{\beta}_{\mathrm{QL}} \xrightarrow{\mathrm{P1}} \beta$.

Then just as in the proof of Theorem 3.13, using the sandwich formula it can be shown that if $\beta$ is an interior point of $\Theta_\beta$, then

$$\sqrt{n}\,(\hat{\beta}_{\mathrm{QL}} - \beta) \xrightarrow{\mathrm{d}} \mathrm{N}(0, B_\beta)\,, \tag{4.75}$$

$$B_\beta = \Phi_\beta^{-1} = \left(\mathbf{E}\frac{1}{v}\frac{\partial m}{\partial \beta}\left(\frac{\partial m}{\partial \beta}\right)^{\mathrm{T}}\right)^{-1}. \tag{4.76}$$

Here, $v$ and $\frac{\partial m}{\partial v}$ are evaluated at the point $(x, \beta)$. The matrix $\Phi_\beta$ is nonsingular, if components of the random vector $\frac{\partial m}{\partial \beta}(x, \beta)$ are linearly independent in the space $L_2(\Omega, \mathbf{P})$. We urgently require the latter; this holds true for the polynomial regression (see the proof of Theorem 3.13).

**Estimation of regression parameter and dispersion parameter**

Now, suppose that $\mu_\xi$ and $\sigma_\xi^2$ are known and the parameters $z = (\beta^{\mathrm{T}}, \phi)^{\mathrm{T}}$ are to be estimated. The estimating function $s_{\mathrm{QL}}^{(\beta)}(y, x; \beta, \phi)$ is given by equality (4.69). However, an additional estimating function $s_{\mathrm{QL}}^{(\phi)}$ has to be constructed for taking $\phi$ into account. In view of (4.68), we put

$$s_{\mathrm{QL}}^{(\phi)}(y, x; \beta, \phi) = (y - m(x, \theta))^2 - \phi\mathbf{E}[C''(\eta)|x] - \mathbf{E}[(C'(\eta))^2|x] + m^2(x, \theta)\,. \tag{4.77}$$

Construct the total estimating function

$$s_{\mathrm{QL}}^{(z)} = \begin{pmatrix} s_{\mathrm{QL}}^{(\beta)} \\ s_{\mathrm{QL}}^{(\phi)} \end{pmatrix}. \tag{4.78}$$

The estimator $\hat{z}_{\mathrm{QL}} = (\hat{\beta}_{QL}^{\mathrm{T}}, \hat{\phi}_{\mathrm{QL}})^{\mathrm{T}}$ is defined as a Borel measurable function of observations $y_1, x_1, \ldots, y_n, x_n$, which *eventually* satisfies the estimating equation

$$\sum_{i=1}^{n} s_{\mathrm{QL}}^{(z)}(y_i, x_i; z) = 0\,, \quad z \in \Theta_z\,. \tag{4.79}$$

Here $\Theta_z$ is a given parameter set in $\mathbf{R}^p \times (0, +\infty)$. Usually we set $\Theta_z = \Theta_\beta \times [\phi_1, \phi_2]$, where $\Theta_\beta$ is a compact set in $\mathbf{R}^p$ and $0 < \phi_1 < \phi_2$. Thus, we assume that the dispersion parameter $\phi$ lies between certain bounds $\phi_1$ and $\phi_2$.

Explain why the estimating function (4.78) is unbiased. We have

$$\mathbf{E}_z s_{\mathrm{QL}}^{(\beta)}(y, x; z) = \mathbf{E}\frac{m(x, \beta) - m(x, \beta)}{v(x; \beta, \phi)} \cdot \frac{\partial m(x, \beta)}{\partial \beta} = 0\,, \tag{4.80}$$

$$\mathbf{E}_z s_{\mathrm{QL}}^{(\phi)}(y, x; z) = \mathbf{E}[v(x; z) - v(x; z)] = 0\,, \tag{4.81}$$

and the estimating function $s_{\mathrm{QL}}^{(z)}$ is indeed unbiased. Next, let $z_0 = (b^{\mathrm{T}}, \phi_0)^{\mathrm{T}}$ be the true value of $z$. By the SLLN,

$$\frac{1}{n}\sum_{i=1}^{n} s_{\mathrm{QL}}^{(z)}(y_i, x_i; z) \xrightarrow{\mathrm{P1}} s_{\infty}^{(z)}(z_0, z) = \mathbf{E}_{z_0} s_{\mathrm{QL}}^{(z)}(y, x; z)\,. \tag{4.82}$$

Moreover, here the convergence is uniform in $z \in \Theta_z$.

Consider the limit equation

$$s_{\infty}^{(z)}(z_0, z) = \begin{pmatrix} s_{\infty}^{(\beta)}(z_0, z) \\ s_{\infty}^{(\phi)}(z_0, z) \end{pmatrix} = 0\,, \quad z \in \Theta_z\,. \tag{4.83}$$

The first equation of this system takes the form

$$\mathbf{E}\frac{m(x, b) - m(x, \beta)}{v(x; \beta, \phi)} \cdot \frac{\partial m(x, \beta)}{\partial \beta} = 0\,, \quad b \in \Theta_\beta\,, \quad \phi \in [\phi_1, \phi_2]\,. \tag{4.84}$$

We demand that for each $\phi \in [\phi_1, \phi_2]$, the equation being an equation in $b \in \Theta_\beta$ should have a unique solution $b = \beta$ (cf. the proof of Theorem 3.13). Given $b = \beta$, the second equation of the system (4.83) is simplified as

$$(\phi_0 - \phi)\,\mathbf{E}C''(\eta(x, \beta)) = 0\,, \quad \phi \in [\phi_1, \phi_2]\,. \tag{4.85}$$

This expectation is positive, because $C'' > 0$ due to Definition 4.1. So, $\phi = \phi_0$ is the only solution to equation (4.85).

Hence, the limit equation (4.83) has a unique solution $z = z_0$.

This fact and the unbiasedness of estimating function ensure the strong consistency of the estimator $\hat{z}_{\mathrm{QL}}$ (Appendix A1). Moreover, by the sandwich formula (Appendix A2) we have the following. If $\beta_0$ is an interior point of $\Theta_\beta$ and $\phi_0 \in (\phi_1, \phi_2)$, then

$$\sqrt{n}\,(\hat{z}_{\mathrm{QL}} - z_0) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \Sigma_z)\,, \tag{4.86}$$

$$\Sigma_z = A_z^{-1} B_z A_z^{-\mathrm{T}}\,, \quad A_z = -\mathbf{E}_{z_0}\frac{\partial s_{QL}^{(z)}}{\partial z^{\mathrm{T}}}(y, x; z_0)\,, \tag{4.87}$$

$$B_z = \mathbf{E}_{z_0} s_{\mathrm{QL}}^{(z)} (s_{\mathrm{QL}}^{(z)})^{\mathrm{T}}\,. \tag{4.88}$$

Here $s_{\mathrm{QL}}^{(z)} = s_{\mathrm{QL}}^{(z)}(y, x; z_0)$.

The matrix $A_z$ has the form

$$A_z = \left(\begin{array}{c|c} -\mathbf{E}\frac{\partial s_{QL}^{(\beta)}}{\partial \beta^{\mathrm{T}}} & 0 \\ \hline * & \mathbf{E}C''(\eta) \end{array}\right)\,. \tag{4.89}$$

The matrix has got a zero upper block because

$$\mathbf{E}_{z_0}\frac{\partial s_{QL}^{(\beta)}}{\partial \phi}(y, x; z_0) = \mathbf{E}_{z_0}(y - m(x, b))\frac{\partial m(x, b)}{\partial \beta} \cdot \frac{\partial}{\partial \beta}\left(\frac{1}{v(x; b, \phi_0)}\right) = 0\,. \tag{4.90}$$

In addition,

$$-\mathbf{E}\frac{\partial s_{\mathrm{QL}}^{(\beta)}}{\partial \beta^{\mathrm{T}}} = \Phi_\beta > 0\,, \tag{4.91}$$

where the matrix $\Phi_\beta$ was introduced in (4.76). Hence,

$$\det A_z = (\det \Phi_\beta)\mathbf{E}C''(\eta) > 0\,, \tag{4.92}$$

and the matrix $A_z$ is nonsingular. Under mild conditions, the matrix $B_z$ is nonsingular as well. The nonsingularity of the two matrices justifies validity of the sandwich formula (4.86)–(4.88).

Further, the matrix $B_z$ has a block structure

$$B_z = \begin{pmatrix} \Phi_\beta & * \\ * & * \end{pmatrix}. \tag{4.93}$$

From relations (4.89), (4.91), (4.93), and the sandwich formula (4.87), we obtain, as in the proof of Theorem 3.27, that

$$\Sigma_z = \begin{pmatrix} \Phi_\beta^{-1}\Phi_\beta\Phi_\beta^{-1} & * \\ * & * \end{pmatrix} = \begin{pmatrix} \Phi_\beta^{-1} & * \\ * & * \end{pmatrix}. \tag{4.94}$$

Given the convergence (4.86), this means for the component $\hat\beta_{\mathrm{QL}}$ of the estimator $\hat z_{\mathrm{QL}}$, that

$$\sqrt{n}\,(\hat\beta_{\mathrm{QL}} - b) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \Phi_\beta^{-1})\,, \tag{4.95}$$

i.e., when $\mu_\xi$ and $\sigma_\xi^2$ are known, the asymptotic covariance matrix (ACM) of the QL estimator for the parameter $\beta$ is equal to $\Phi_\beta^{-1}$ and does not depend on whether the parameter $\phi$ is known or not (see (4.75) and (4.76)).

The system (4.79) of equations w.r.t. $z = (\beta^{\mathrm{T}}, \phi)^{\mathrm{T}}$ can be solved numerically by an iterative method, similar to the one from Theorem 3.29. We just describe the method.

Denote by $P$ a *projector* on $[\phi_1, \phi_2]$,

$$P(\phi) = \begin{cases} \phi, & \phi \in [\phi_1, \phi_2], \\ \phi_1, & \phi < \phi_1, \\ \phi_2, & \phi > \phi_2. \end{cases} \tag{4.96}$$

(1) Take arbitrary initial values $\beta^{(0)} \in \Theta_\beta$ and $\phi^{(0)} \in [\phi_1, \phi_2]$.
(2) Given $\beta^{(j)}$ and $\phi^{(j)}$ from the $j$th iteration of the algorithm, we evaluate $\beta^{(j+1)}$ as a solution to the equation

$$\sum_{i=1}^{n} \frac{y_i - m(x_i, \beta)}{v(x_i; \beta^{(j)}, \phi^{(j)})} \cdot \frac{\partial m}{\partial \beta}(x_i, \beta) = 0\,, \quad \beta \in \Theta_\beta\,. \tag{4.97}$$

(If the equation has no solution, then a point from the compact set $\Theta_\beta$, with the smallest Euclidean norm of the left-hand side of (4.97), should be taken as $\beta^{(j+1)}$.)

Then we find $\tilde{\phi}^{(j+1)}$ as a solution to the linear equation

$$\sum_{i=1}^{n} s_{\mathrm{QL}}^{(\phi)}(y_i, x_i; \beta^{(j+1)}, \phi) = 0\,, \quad \phi \in \mathbf{R}\,, \tag{4.98}$$

and put $\phi^{(j+1)} = P(\tilde{\phi}^{(j+1)})$.

Under mild conditions, we have *eventually*

$$\lim_{j\to\infty} \beta^{(j)} = \hat{\beta}_{\mathrm{QL}}\,, \quad \lim_{j\to\infty} \phi^{(j)} = \hat{\phi}_{\mathrm{QL}}\,. \tag{4.99}$$

Note that unlike the case of polynomial regression, (4.97) can be a nonlinear equation, which has to be solved by corresponding numerical methods.

**The case where all the nuisance parameters are unknown**

Now, we have to estimate the vector parameter

$$\theta = (z^{\mathrm{T}}, \gamma^{\mathrm{T}})^{\mathrm{T}} = (\beta^{\mathrm{T}}, \phi, \mu_\xi, \sigma_\xi^2)^{\mathrm{T}}\,. \tag{4.100}$$

Here $\gamma = (\mu_\xi, \sigma_\xi^2)^{\mathrm{T}}$ is the vector of parameters of the distribution of $\xi$.

As a preliminary, construct the estimator $\hat{\gamma}_{\mathrm{QL}}$ by formulas (3.269) and (3.270). Next, denote

$$\hat{s}_{\mathrm{QL}}^{(z)}(y, x; z) = s_{\mathrm{QL}}^{(z)}(y, x; z, \hat{\gamma}_{\mathrm{QL}})\,. \tag{4.101}$$

The estimator $\hat{z}_{\mathrm{QL}} = (\hat{\beta}_{\mathrm{QL}}^{\mathrm{T}}, \hat{\phi}_{\mathrm{QL}})^{\mathrm{T}}$ is defined as a measurable solution to the equation

$$\sum_{i=1}^{n} \hat{s}_{\mathrm{QL}}^{(z)}(y_i, x_i; z) = 0\,, \quad z \in \Theta_\beta \times [\phi_1, \phi_2]\,. \tag{4.102}$$

The iterative numerical algorithm described in Section 4.2.2 can be obviously adapted to compute the estimate $\hat{z}_{\mathrm{QL}}$.

The estimator is strongly consistent. If the true value of $\beta$ is an interior point of $\Theta_\beta$ and the true value $\phi \in (\phi_1, \phi_2)$, then $\hat{z}_{\mathrm{QL}}$ is asymptotically normal estimator. To write down the ACM of the estimator, we reflect in a manner similar to the proof of Theorem 3.31.

The estimator $\hat{\gamma}_{\mathrm{QL}}$ is a solution to the system, (3.278) and (3.277), hence $\hat{\theta}_{\mathrm{QL}} = (\hat{z}_{\mathrm{QL}}^{\mathrm{T}}, \hat{\gamma}_{\mathrm{QL}}^{\mathrm{T}})^{\mathrm{T}}$ is determined by the estimating function

$$s_{\mathrm{QL}}^{(\theta)} = \begin{pmatrix} s_{\mathrm{QL}}^{(z)} \\ s_{\mathrm{QL}}^{(\gamma)} \end{pmatrix}\,. \tag{4.103}$$

Then utilizing the sandwich formula, we get

$$\sqrt{n}\,(\hat{\theta}_{\mathrm{QL}} - \theta) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \Sigma_\theta)\,, \quad \Sigma_\theta = A_\theta^{-1} B_\theta A_\theta^{-\mathrm{T}}\,. \tag{4.104}$$

Here

$$A_\theta = -\mathbf{E}_\theta \frac{\partial s_{\mathrm{QL}}^{(\theta)}}{\partial \theta^{\mathrm{T}}}(y, x; \theta) = \begin{pmatrix} A_\beta & 0 & A_{\beta\gamma} \\ A_{\phi\beta} & A_{\phi\phi} & A_{\phi\gamma} \\ 0 & 0 & I_2 \end{pmatrix}, \tag{4.105}$$

$$A_\beta = \Phi_\beta\,, \quad A_{\phi\phi} = \mathbf{E} C''(\eta) > 0\,. \tag{4.106}$$

By Sylvester's criterion, matrix $A_\theta$ is positive definite. The inverse matrix has the form

$$A_\theta^{-1} = \begin{pmatrix} A_\beta^{-1} & 0 & Y \\ * & A_{\phi\phi}^{-1} & * \\ 0 & 0 & I_2 \end{pmatrix}, \quad Y = -A_\beta^{-1} A_{\beta\gamma}\,. \tag{4.107}$$

The middle part of the "sandwich" (4.104) is equal to

$$B_\theta = \mathbf{E}_\theta \operatorname{cov} s_{\mathrm{QL}}^{(\theta)}(y, x; \theta) = \begin{pmatrix} B_\beta & B_{\beta\phi} & 0 \\ B_{\phi\beta} & B_\phi & 0 \\ 0 & 0 & B_\gamma \end{pmatrix}, \tag{4.108}$$

$$B_\beta = \Phi_\beta\,, \quad B_\gamma = \operatorname{cov}_\gamma s_{\mathrm{QL}}^{(\gamma)}\,. \tag{4.109}$$

Then similarly to the calculations (3.291) and (3.292), we obtain

$$\Sigma_\theta = \begin{pmatrix} A_\beta^{-1} B_\beta A_\beta^{-1} + Y B_\gamma Y^{\mathrm{T}} & * & * \\ * & * & * \\ * & * & * \end{pmatrix}. \tag{4.110}$$

The written out block of the matrix $\Sigma_\theta$ is the ACM of the estimator $\hat\beta_{\mathrm{QL}}$, and the ACM is equal to (see formulas (4.106), (4.107), and (4.109)):

$$\tilde\Sigma_\beta = \Phi_\beta^{-1} + \Phi_\beta^{-1} A_{\beta\gamma} B_\gamma A_{\beta\gamma}^{\mathrm{T}} \Phi_\beta^{-1}\,. \tag{4.111}$$

We obtained a formula similar to (3.293). It shows that lack of knowledge about the nuisance parameters $\gamma$ worsens the estimation quality for the regression parameter $\beta$.

### 4.2.3 Corrected score estimator

We consider the structural model (4.2), (4.43), and (4.62), under the assumption (i). The distribution of $\xi_i$ is not necessarily normal.

Construct the CS estimator $\hat\beta_{\mathrm{C}}$ (Section 1.4.4). We start with the unbiased score $s_{\mathrm{ML}}(y, \xi; \beta)$ presented in (4.47), which yields the consistent estimator of $\beta$ under the absence of error in the regressor. Let $g_{\mathrm{C}}(x, \beta)$ and $h_{\mathrm{C}}(x, \beta)$ be solutions to the vector deconvolution equations:

$$\mathbf{E}[g_{\mathrm{C}}(x, \beta)|\xi] = \eta'_\beta(\xi, \beta)\,, \tag{4.112}$$

$$\mathbf{E}[h_{\mathrm{C}}(x, \beta)|\xi] = C'(\eta) \cdot \eta'_\beta(\xi, \beta)\,, \quad \eta = \eta(\xi, \beta)\,; \quad \beta \in \Theta_\beta\,. \tag{4.113}$$

Now, we assume that there exist smooth enough solutions to these equations. (Later on, we will reveal that it is not true for the logistic errors-in-variables model, based on Example 4.7. However, in many concrete models, these deconvolution equations can be solved.)

Consider the estimating function

$$S_{\mathrm{C}}(y, x; \beta) = y \cdot g_{\mathrm{C}}(x, \beta) - h_{\mathrm{C}}(x, \beta) , \quad \beta \in \Theta_\beta . \tag{4.114}$$

We have

$$\begin{aligned} \mathbf{E}[s_{\mathrm{C}}(y, x; b)|y, \xi] &= y\ \mathbf{E}[g_{\mathrm{C}}(x, b)|\xi] - \mathbf{E}[h_{\mathrm{C}}(x, b)|\xi] \\ &= y \cdot \eta'_\beta - C'(\eta) \cdot \eta'_\beta = s_{\mathrm{ML}}(y, \xi; b) , \quad b \in \Theta_\beta . \end{aligned} \tag{4.115}$$

The corrected score estimator $\hat{\beta}_{\mathrm{C}}$ *eventually* satisfies the equation

$$\sum_{i=1}^{n} s_{\mathrm{C}}(y_i, x_i; \beta) = 0 , \quad \beta \in \Theta_\beta . \tag{4.116}$$

**Asymptotic properties of the estimator**

Under mild conditions, the estimator $\hat{\beta}_{\mathrm{C}}$ is strongly consistent and asymptotically normal, with the ACM

$$\Sigma_{\beta,C} = A^{-1}_{\beta,\mathrm{C}} B_{\beta,\mathrm{C}} A^{-1}_{\beta,\mathrm{C}} , \tag{4.117}$$

$$A_{\beta,\mathrm{C}} = -\mathbf{E}_\beta \frac{\partial s_{\mathrm{C}}(y, x; \beta)}{\partial \beta^{\mathrm{T}}} , \quad B_{\beta,\mathrm{C}} = \mathrm{cov}_\beta\, s_{\mathrm{C}}(y, x; \beta) . \tag{4.118}$$

The latter two matrices can be written through the conditional mean (4.67) and conditional variance (4.68). Here we assume that either the condition (ii) on the normality of $\xi$ holds or the distribution of $\xi$ is a mixture of several normal distributions (see Section 1.4.3). Thus, we get

$$A_{\beta,\mathrm{C}} = \mathbf{E}\frac{\partial h_{\mathrm{C}}}{\partial \beta^{\mathrm{T}}} - \mathbf{E}_\beta y \frac{\partial g_{\mathrm{C}}}{\partial \beta^{\mathrm{T}}} = \mathbf{E}\left( \frac{\partial h_{\mathrm{C}}}{\partial \beta^{\mathrm{T}}} - m(x, \beta) \cdot \frac{\partial g_{\mathrm{C}}}{\partial \beta^{\mathrm{T}}} \right) , \tag{4.119}$$

$$\begin{aligned} B_{\beta,\mathrm{C}} &= \mathrm{cov}_\beta[(y - m(x, \beta))g_{\mathrm{C}} + (m(x, \beta)g_{\mathrm{C}} - h_{\mathrm{C}})] = \\ &= \mathbf{E}(y - m(x, \beta))^2 g_{\mathrm{C}} g_{\mathrm{C}}^{\mathrm{T}} + \mathrm{cov}_\beta(m(x, \beta)g_{\mathrm{C}} - h_{\mathrm{C}}) = \\ &= \mathbf{E}v(x, \beta) g_{\mathrm{C}} g_{\mathrm{C}}^{\mathrm{T}} + \mathrm{cov}_\beta(m(x, \beta)g_{\mathrm{C}} - h_{\mathrm{C}}) . \end{aligned} \tag{4.120}$$

Here we used the fact that for $m = m(x, \beta)$, it holds that

$$\mathbf{E}(y - m)\, g_{\mathrm{C}}(m g_{\mathrm{C}} - h)^{\mathrm{T}} = \mathbf{E}[g_{\mathrm{C}}(m g_{\mathrm{C}} - h)^{\mathrm{T}} \cdot \mathbf{E}(y - m|x)] = 0 . \tag{4.121}$$

Of course, the nonsingularity of matrix $A_{\beta,\mathrm{C}}$ is required. By formula (4.120), we obtain

$$B_{\beta,\mathrm{C}} \geq \mathbf{E}v(x, \beta) g_{\mathrm{C}} g_{\mathrm{C}}^{\mathrm{T}} > 0 , \tag{4.122}$$

if the components of the gradient $\eta'_\beta(\xi, \beta)$ are linearly independent random variables in the space $L_2(\Omega, \mathbf{P})$. The latter we certainly demand. Then the matrix $B_{\beta,\mathrm{C}}$ is nonsingular as well, and the sandwich formula (4.117) can be applied.

Notice that in the generalized linear model (4.44), if $h'(t) \neq 0$, $t \in \mathbf{R}$, then the components of $\eta'_\beta(\xi, \beta)$ are linearly independent (Section 4.2.1).

**Quasi-likelihood estimator is more efficient than corrected score estimator**

Now, assume both conditions (i) and (ii) in the model (4.2), (4.43), and (4.62). The nuisance parameters $\phi$, $\mu_\xi$, and $\sigma^2_\xi$ are assumed known. We are going to compare the ACMs of estimators $\hat\beta_{\mathrm{QL}}$ and $\hat\beta_{\mathrm{C}}$.

Theorem 3.21 on the asymptotic efficiency of the QL estimator is generalized. Consider a linear in $y$ unbiased estimating function

$$s_{\mathrm{L}}(y, x; \beta) = g(x, \beta) \cdot y - h(x, \beta),\ y \in \mathbf{R},\ x \in \mathbf{R},\ \beta \in \Theta_\beta \subset \mathbf{R}^p\ . \tag{4.123}$$

Here $g$ and $h$ are Borel measurable functions with values in $\mathbf{R}^p$. Based on $s_{\mathrm{L}}$, we define the estimator $\hat\beta_{\mathrm{L}}$ as a measurable solution to equation

$$\sum_{i=1}^{n} s_L(y_i, x_i; \beta) = 0\ , \quad \beta \in \Theta_\beta\ . \tag{4.124}$$

Under mild conditions, $\hat\beta_{\mathrm{L}}$ is consistent and asymptotically normal estimator, with the ACM given in (3.190) (see Corollary A.31 in Appendix A2).

Denote by $L$ the class of all such estimating functions that the corresponding estimator $\hat\beta_{\mathrm{L}}$ has all abovementioned asymptotic properties. It is clear that $s_{\mathrm{C}}$ and $s_{\mathrm{QL}}$ belong to $L$.

**Theorem 4.10.** *Let $s_{\mathrm{L}} \in L$, and $\Sigma_{\mathrm{L}}$, $\Sigma_{\mathrm{QL}}$ be the ACMs of estimators $\hat\beta_{\mathrm{L}}$ and $\hat\beta_{\mathrm{QL}}$, respectively. Then $\Sigma_{\mathrm{QL}} \leq \Sigma_{\mathrm{L}}$. If in addition, $\Sigma_L = \Sigma_{QL}$ holds true, for all the true values $b \in \Theta_\beta$, then $\hat\beta_{\mathrm{L}} = \hat\beta_{\mathrm{QL}}$, almost surely.*

*Proof* is given in Kukush et al. (2009).

The theorem shows that the QL estimator is asymptotically more efficient than the CS estimator. Later on for concrete models, we will give conditions that ensure a strict inequality $\Sigma_{\mathrm{QL}} < \Sigma_{\mathrm{C}}$.

Further, consider the behavior of the matrices $\Sigma_{\mathrm{QL}}$ and $\Sigma_{\mathrm{C}}$ for small $\sigma^2_\delta$. Each of those is expanded in powers of the variance $\sigma^2_\delta$, as $\sigma^2_\delta \to 0$. We reveal that the expansion for difference of the matrices starts only with terms of order $\sigma^4_\delta$ or even higher order.

**Theorem 4.11.** *For the ACMs $\Sigma_{\mathrm{QL}}$ and $\Sigma_{\mathrm{C}}$, it holds that*
(a)

$$\Sigma_{\mathrm{QL}} = \Phi^{-1} + O(\sigma^2_\delta)\ , \quad \text{as}\ \ \sigma^2_\delta \to 0\ , \tag{4.125}$$

*with the information matrix $\Phi$ given in* (4.54)*, and*

(b)

$$\Sigma_{\mathrm{C}} = \Sigma_{\mathrm{QL}} + O(\sigma_\delta^4)\,, \quad as \;\; \sigma_\delta^2 \to 0\,. \tag{4.126}$$

*Proof* can be found in Kukush and Schneeweiss (2005a, 2005b).

Please pay attention to the discussion just after Theorem 3.23. The discussion remains in force regarding the models with exponential family of densities.

**Remark 4.12.** Suppose that $\phi$ is known, while $\beta$, $\mu_\xi$, and $\sigma_\xi^2$ are unknown. Then, in general, one cannot state that $\tilde{\Sigma}_\beta \le \Sigma_{\mathrm{C}}$, where $\tilde{\Sigma}_\beta$ is the ACM of the estimator $\hat{\beta}_{\mathrm{QL}}$ (see formula (4.111)). But we will see later on that in some models, the inequality holds true. We have already seen that this is true in the polynomial model.

**Estimation of dispersion parameter**

Let the form of distribution of $\xi_i$ be unknown. To construct $\hat{\beta}_{\mathrm{C}}$, it is unnecessary to know the parameter $\phi$. However, one might need to estimate $\phi$, for example, in order to estimate consistently the ACM of the estimator $\hat{\beta}_{\mathrm{C}}$.

To construct the estimator $\hat{\beta}_{\mathrm{C}}$, we write down the conditional second moment of response on the basis of formulas (4.9) and (4.14):

$$\mathbf{E}(y^2|\xi) = (C')^2 + \phi C''\,. \tag{4.127}$$

Hereafter $C'$ and $C''$ are evaluated at the point $\eta(\xi, \beta)$.

Let $p_{\mathrm{C}}(x, \beta)$ and $q_{\mathrm{C}}(x, \beta)$ be solutions to deconvolution equations

$$\mathbf{E}[p_{\mathrm{C}}(x, \beta)|\xi] = (C')^2\,, \tag{4.128}$$

$$\mathbf{E}[q_{\mathrm{C}}(x, \beta)|\xi] = C''\,, \quad \beta \in \Theta_\beta\,. \tag{4.129}$$

Introduce the estimating function

$$s_{\mathrm{C}}^{(\phi)}(y, x; \beta, \phi) = y^2 - p_{\mathrm{C}}(x, \beta) - \phi q_{\mathrm{C}}(x, \beta)\,, \quad \beta \in \Theta_\beta\,. \tag{4.130}$$

Denote by $\tilde{\phi}_{\mathrm{C}}$ a solution to the linear equation

$$\sum_{i=1}^{n} s_{\mathrm{C}}^{(\phi)}(y_i, x_i; \hat{\beta}_{\mathrm{C}}, \phi) = 0\,, \quad \phi \in \mathbf{R}\,. \tag{4.131}$$

Set

$$\hat{\phi}_{\mathrm{C}} = \begin{cases} \tilde{\phi}_{\mathrm{C}} & \text{if } \tilde{\phi}_{\mathrm{C}} \ge 0\,, \\ 0 & \text{if } \tilde{\phi}_{\mathrm{C}} < 0\,. \end{cases} \tag{4.132}$$

The estimators $\hat{\beta}_{\mathrm{C}}$ and $\hat{\phi}_{\mathrm{C}}$ *eventually* satisfy the system of equations

$$\sum_{i=1}^{n} s_{\mathrm{C}}^{(z)}(y_i, x_i; \beta, \phi) = 0\,, \quad \beta \in \Theta_\beta\,, \quad \phi > 0\,, \tag{4.133}$$

$$s_{\mathrm{C}}^{(z)} = \begin{pmatrix} s_{\mathrm{C}}^{(\beta)} \\ s_{\mathrm{C}}^{(\phi)} \end{pmatrix}\,. \tag{4.134}$$

The estimating function $s_{\mathrm{C}}^{(z)}$ is unbiased:

$$\begin{aligned}\mathbf{E}_{\beta,\phi} s_{\mathrm{C}}^{(\phi)}(y, x; \beta, \phi) &= \mathbf{E}\mathbf{E}_{\beta,\phi}[s_{\mathrm{C}}^{(\phi)}(y, x; \beta, \phi)|\xi] = \\ &= \mathbf{E}[\mathbf{E}_{\beta,\phi}(y^2|\xi) - (C')^2 - \phi C''] = 0 \,. \end{aligned} \tag{4.135}$$

Further, as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} s_{\mathrm{C}}^{(z)}(y_i, x_i; b, f) \xrightarrow{\mathrm{P1}} S_{\mathrm{C},\infty}^{(z)}(\beta, \phi; b, f) = \mathbf{E}_{\beta,\phi} s_{\mathrm{C}}^{(z)}(y, x; b, f) \,. \tag{4.136}$$

Consider the asymptotic equation

$$S_{\mathrm{C},\infty}^{(z)}(\beta, \phi;\ b, f) = 0 \,, \quad b \in \Theta_\beta \,, \quad f > 0 \,. \tag{4.137}$$

From the first equation $S_{\mathrm{C},\infty}^{(\beta)}(\beta; b) = 0$, $b \in \Theta_\beta$, we get $b = \beta$. Then the second equation of (4.137) is simplified as

$$(\phi - f) \cdot \mathbf{E}C''(\eta(\xi, \beta)) = 0 \,, \quad f > 0 \,, \tag{4.138}$$

whence $f = \phi$. Therefore, the asymptotic equation (4.137) has a unique solution $b = \beta$ and $f = \phi$. This fact and the unbiasedness of the estimating function $s_{\mathrm{C}}^{(z)}$ provide the strong consistency of the estimator $\hat{z}_{\mathrm{C}} = (\hat{\beta}_{\mathrm{C}}^{\mathrm{T}}, \hat{\phi}_{\mathrm{C}})^{\mathrm{T}}$, in particular, $\hat{\phi}_{\mathrm{C}}$ is the strongly consistent estimator of the dispersion parameter.

The ACM of $\hat{z}_{\mathrm{C}}$ can be found by the sandwich formula.

### 4.2.4 Other methods for estimation of regression parameter

Regression calibration method described in Section 1.4.5 can be utilized in the model (4.2), (4.43), and (4.62) under the conditions (i) and (ii). As unbiased estimating function $s(y, \xi; \beta)$, one can take the function $s_{\mathrm{ML}}$ from formula (4.47), and the vector $(\mu_\xi, \sigma_\xi^2)^{\mathrm{T}}$ plays the role of a nuisance parameter $\gamma$. The same method can be used under unknown $\phi$, $\mu_\xi$, and $\sigma_\xi^2$ (see the discussion in Section 1.4.5).

In the latter situation, the SIMEX estimator can be applied as well (see Section 1.4.6), again with $s_{\mathrm{ML}}(y, \xi; \beta)$ taken as an unbiased estimating function.

Remember that both methods do not yield consistent estimator, but they can significantly reduce the deviations of the estimate from the true value, compared to the naive estimate $\hat{\beta}_{\mathrm{naive}}$ (see Section 4.2.2).

## 4.3 Two consistent estimators in Gaussian model

Now, we turn to the concrete structural regression models with exponential family of densities and errors in the covariate. Consider Example 4.3, in which we set

$$\eta = \beta^{\mathrm{T}} \phi(\xi) \,, \quad \beta \in \mathbf{R}^p \,, \quad \phi \colon \mathbf{R} \to \mathbf{R}^p \,. \tag{4.139}$$

The regression model (4.16), (4.17), and (4.139) is a linear in $\beta$ Gaussian model:

$$y = \beta^{\mathrm{T}}\phi(\xi) + \varepsilon\,, \quad \varepsilon \sim \mathrm{N}(0, \sigma^2)\,, \quad \xi \perp\!\!\!\perp \varepsilon\,. \tag{4.140}$$

Additionally, we assume that instead of $\xi$, we observe

$$x = \xi + \delta\,, \quad \delta \sim \mathrm{N}(0, \sigma_\delta^2)\,. \tag{4.141}$$

Moreover, it is supposed that $\xi$, $\varepsilon$, and $\delta$ are independent and $\sigma_\delta^2$ is known.

The polynomial model studied in Chapter 3 is a special case of the observation model (4.140) and (4.141). But we will consider another choice of the function $\phi$:

$$\phi(\xi) = \left(\mathrm{e}^{\lambda_1\xi}, \ldots, \mathrm{e}^{\lambda_p\xi}\right)^{\mathrm{T}}\,, \quad \lambda_i \neq \lambda_j\,, \quad i, j = \overline{1, p}\,. \tag{4.142}$$

For such a function $\phi$, we concretize the consistent estimators from Section 4.2.

### 4.3.1 Corrected score estimator

In the model (4.140), the score $s_{\mathrm{ML}}(y, \xi; \beta)$, corresponding to the observations $y$ and $\xi$, is as follows:

$$s_{\mathrm{ML}} = \phi y - (\phi\phi^{\mathrm{T}})\beta\,. \tag{4.143}$$

The corrected score $s_{\mathrm{C}}(y, \xi; \beta)$ has the form

$$s_{\mathrm{C}} = ty - H\beta\,, \tag{4.144}$$

with

$$\mathbf{E}(t(x)|\xi) = \phi(\xi)\,, \quad \mathbf{E}(H(x)|\xi) = \phi(\xi)\,\phi^{\mathrm{T}}(\xi)\,. \tag{4.145}$$

The basic equation is the following:

$$\mathbf{E}(t_\lambda(x)|\xi) = \phi_\lambda(\xi) = \mathrm{e}^{\lambda\xi}\,, \quad \lambda \in \mathbf{R}\,. \tag{4.146}$$

The function

$$t_\lambda(x) = \frac{\mathrm{e}^{\lambda x}}{\mathbf{E}\mathrm{e}^{\lambda\delta}} = \mathrm{e}^{-\frac{\lambda^2\sigma_\delta^2}{2}} \cdot \mathrm{e}^{\lambda x}\,, \quad x \in \mathbf{R}\,, \tag{4.147}$$

satisfies (4.146), because

$$\mathbf{E}(t_\lambda(x)|\xi) = \frac{1}{\mathbf{E}\mathrm{e}^{\lambda\delta}}\mathbf{E}(\mathrm{e}^{\lambda\xi} \cdot \mathrm{e}^{\lambda\delta}|\xi) = \frac{\mathrm{e}^{\lambda\xi} \cdot \mathbf{E}\mathrm{e}^{\lambda\delta}}{\mathbf{E}\mathrm{e}^{\lambda\delta}} = \mathrm{e}^{\lambda\xi}\,. \tag{4.148}$$

Then solutions to equations (4.145) are

$$t(x) = (t_{\lambda_i}(x))_{i=1}^p\,, \quad H(x) = (H_{ij}(x))_{i,j=1}^p = (t_{\lambda_i+\lambda_j}(x))_{i,j=1}^p\,. \tag{4.149}$$

The estimating equation for the estimator $\hat{\beta}_{\mathrm{C}}$ has the form

$$\overline{t(x)y} - \overline{H(x)}\,\beta = 0\,, \quad \beta \in \mathbf{R}^p\,. \tag{4.150}$$

**Theorem 4.13.** *In the model* (4.140)–(4.142)*, assume the following:*
(a)

$$\mathbf{E}\mathrm{e}^{2\lambda_* \xi} < \infty\,, \quad \mathbf{E}\mathrm{e}^{2\lambda^* \xi} < \infty\,, \tag{4.151}$$

*where* $\lambda_* = \min\{0;\ \lambda_i,\ i = \overline{1,p}\}$, $\quad \lambda^* = \min\{0;\ \lambda_i,\ i = \overline{1,p}\}$.
(b) *The cdf* $F_\xi$ *of the regressor* $\xi$ *is strictly increasing on certain interval* $[a, b]$.
*Then* eventually *the matrix* $\overline{H(x)}$ *is nonsingular, the estimator* $\hat{\beta}_{\mathrm{C}}$ eventually *satisfies the equality*

$$\hat{\beta}_{\mathrm{C}} = \left(\overline{H(x)}\right)^{-1} \overline{t(x)y}\,, \tag{4.152}$$

*and* $\hat{\beta}_{\mathrm{C}}$ *is the strongly consistent estimator of the parameter* $\beta$.

*Proof.* Due to condition (a) and equality (4.148), we have

$$\mathbf{E}t_{\lambda_i+\lambda_j}(x) = \mathbf{E}\mathrm{e}^{(\lambda_i+\lambda_j)\xi} < \infty\,, \quad i, j = \overline{1,p}\,. \tag{4.153}$$

By the SLLN,

$$\mathbf{E}\overline{H(x)} \xrightarrow{\mathrm{P1}} H_\infty = \mathbf{E}H(x) = \mathbf{E}\phi(x)\phi^{\mathrm{T}}(x)\,. \tag{4.154}$$

The matrix $H_\infty$ is the Gram matrix of random variables $\phi_{\lambda_i}(\xi) = \mathrm{e}^{\lambda_i \xi}$, $i = \overline{1,p}$, in the space $L_2(\Omega, \mathbf{P})$. We will show that they are linearly independent.

Let $a_1\mathrm{e}^{\lambda_1 \xi} + \cdots + a_p\mathrm{e}^{\lambda_p \xi} = 0$, almost surely, where $a_1, \ldots, a_p$ are some real numbers. Because of condition (b), for all $t \in [a, b]$ it holds that $a_1\mathrm{e}^{\lambda_1 t} + \cdots + a_p\mathrm{e}^{\lambda_p t} = 0$. But it is known that the functions $\phi_{\lambda_i}(t)$, $i = \overline{1,p}$, are linearly independent on each interval, when $\lambda_i'$s are distinct. Hence, we get $a_1 = \cdots = a_p = 0$, which proves the linear independence of the random variables $\phi_{\lambda_i}(\xi)$, $i = \overline{1,p}$. This implies that the Gram matrix $H_\infty$ is nonsingular.

From the convergence (4.154), we infer that the matrix $\overline{H(x)}$ is nonsingular *eventually*, and the solution to (4.150) is *eventually* given by equality (4.152).

The proof of the strong consistency of the estimator $\hat{\beta}_{\mathrm{C}}$ is similar to the proof of Theorem 3.5 and omitted here. The proof of Theorem 4.13 is accomplished. □

We write down the ACM of the estimator under the following condition about the normality of latent variable
(c)

$$\xi \sim \mathrm{N}(\mu_\xi, \sigma_\xi^2)\,, \quad \sigma_\xi^2 > 0\,. \tag{4.155}$$

To do this, at first the conditional mean $m(x, \beta)$ and conditional variance $v(x; \beta, \sigma_\varepsilon^2)$ of the response $y$ given $x$ should be found. So, we have from equality (4.140):

$$m(x, \beta) = \mathbf{E}(y|x) = \beta^{\mathrm{T}}\mathbf{E}[\phi(\xi)|x]\,. \tag{4.156}$$

Remember that under the additional condition (c), the conditional distribution of $\xi$ given $x$ is presented in (1.86) and (1.87). Then for $\lambda \in \mathbf{R}$, it holds (here $\gamma \sim \mathrm{N}(0, 1)$ and

$y \perp\!\!\!\perp x$)

$$f_\lambda(x) := \mathbf{E}(\mathrm{e}^{\lambda\xi}|x) = \mathbf{E}(\mathrm{e}^{\lambda(\mu_1(x)+\tau\gamma)}|x) = \mathrm{e}^{\lambda\mu_1(x)}\mathbf{E}\mathrm{e}^{\lambda\cdot\tau\gamma}\,, \tag{4.157}$$

$$f_\lambda(x) = \exp\left\{\lambda\mu_1(x) + \frac{\lambda^2\tau^2}{2}\right\}\,. \tag{4.158}$$

Denote

$$f(x) = (f_{\lambda_1}(x), \ldots, f_{\lambda_p}(x))^{\mathrm{T}}\,, \quad x \in \mathbf{R}\,. \tag{4.159}$$

Then from equations (4.156) and (4.142) it follows

$$m(x,\beta) = \beta^{\mathrm{T}} f(x)\,. \tag{4.160}$$

Next, we introduce a matrix

$$F(x) = \mathbf{E}[\phi(\xi)\,\phi^{\mathrm{T}}(\xi)|x] = (F_{ij}(x))_{i,j=1}^{p}\,, \tag{4.161}$$

$$F_{ij}(x) = \mathbf{E}(\mathrm{e}^{(\lambda_i+\lambda_j)\xi}|x) = f_{\lambda_i+\lambda_j}(x)\,. \tag{4.162}$$

Similarly to the formula (3.60) being found for the polynomial model, we deduce the following in the model (4.140)–(4.142):

$$v(x;\beta,\sigma_\varepsilon^2) = \mathbf{V}(y|x) = \sigma_\varepsilon^2 + \beta^{\mathrm{T}}(F(x) - f(x)f(x)^{\mathrm{T}})\beta\,. \tag{4.163}$$

Like in the polynomial model, it holds that

$$v(x;\beta,\sigma_\varepsilon^2) \geq \sigma_\varepsilon^2 > 0\,. \tag{4.164}$$

**Theorem 4.14.** *In the model* (4.140)–(4.142), *assume the conditions* (b) *and* (c). *Then*

$$\sqrt{n}(\hat\beta_{\mathrm{C}} - \beta) \xrightarrow{d} \mathrm{N}(0, \Sigma_{\mathrm{C}})\,, \tag{4.165}$$

*where $\Sigma_{\mathrm{C}}$ is the nonsingular matrix which depends on $\beta$, $\sigma_\varepsilon^2$, $\mu_\xi$, and $\sigma_\xi^2$:*

$$\Sigma_{\mathrm{C}} = A_{\mathrm{C}}^{-1} B_{\mathrm{C}} A_{\mathrm{C}}^{-1}\,, \quad A_{\mathrm{C}} = \mathbf{E}\phi\phi^{\mathrm{T}}\,, \quad B_{\mathrm{C}} = \quad \mathbf{E}vtt^{\mathrm{T}} + \mathbf{E}(tf^{\mathrm{T}} - H)\beta\beta^{\mathrm{T}}(tf^{\mathrm{T}} - H)^{\mathrm{T}}\,. \tag{4.166}$$

*Here $v$ is given in* (4.163), *$f$ in relations* (4.158), (4.159), *and $t$ and $H$ in equalities* (4.147), (4.148).

*Proof* is quite similar to the proof of Theorem 3.7 and omitted here.

We mention that formula (4.166) for $B_{\mathrm{C}}$ is similar to formula (3.84). The matrix $A_{\mathrm{C}}$ is positive definite as the Gram matrix of linearly independent random variables $\mathrm{e}^{\lambda_i\xi}$, $i = \overline{1,p}$, in the space $L_2(\Omega, \mathbf{P})$; in addition,

$$B_{\mathrm{C}} \geq \mathbf{E}vtt^{\mathrm{T}} > 0\,, \tag{4.167}$$

because the latter matrix is the Gram matrix of linearly independent random variables $\sqrt{v}\,t_{\lambda_i}(x)$, $i = \overline{1,p}$, in $L_2(\Omega, \mathbf{P})$.

**Remark 4.15.** Assume the conditions of Theorem 4.14. Similarly to Theorem 3.10,

$$\hat\sigma_{\varepsilon,C}^2 = \overline{y^2} - \hat\beta_{\mathrm{C}}^{\mathrm{T}}\overline{ty} \tag{4.168}$$

is a strongly consistent estimator of the parameter $\sigma_\varepsilon^2$.

### 4.3.2 Quasi-likelihood estimator

Consider the model (4.140)–(4.142) under the conditions (b) and (c). In this model, we have written down the conditional mean and conditional variance of the response variable, see (4.160) and (4.163). As in Section 3.2, this allows us to write down the estimating function for the QL estimator $\hat{\beta}_{\mathrm{QL}}$ (the estimating function depends on the nuisance parameters $\sigma_\varepsilon^2$, $\mu_\xi$, and $\sigma_\xi^2$):

$$s_{\mathrm{QL}}^{(\beta)}(y, x; \beta, \sigma_\varepsilon^2) = \frac{f(x)\,(y - f^{\mathrm{T}}(x)\beta)}{v(x; \beta, \sigma_\varepsilon^2)}\,. \tag{4.169}$$

If the nuisance parameters are known, the estimator $\hat{\beta}_{\mathrm{QL}}$ is defined as a solution to equation

$$\frac{1}{n}\sum_{i=1}^{n} \frac{1}{v(x_i; \beta, \sigma_\varepsilon^2)}(f(x_i)y_i - f(x_i)f^{\mathrm{T}}(x_i)\beta) = 0\,, \quad \beta \in \mathbf{R}^p\,. \tag{4.170}$$

Nonlinear equation (4.170) does not always possess a solution. Define the estimator more accurately.

**Definition 4.16.** The estimator $\hat{\beta}_{\mathrm{QL}}$ is a Borel measurable function of the observations $y_1, x_1, \ldots, y_n, x_n$, such that

(a) if equation (4.170) has no solution, then $\hat{\beta}_{\mathrm{QL}} = 0$;

(b) if the equation has a solution, then $\hat{\beta}_{\mathrm{QL}}$ is a solution with minimal norm (if there are several such solutions then we take any of them).

(Concerning correctness of the definition, see the discussion just after Definition 3.12.)

Without proof, we state an analog of Theorem 3.13.

**Theorem 4.17.** *In the model* (4.140)–(4.142), *assume that the nuisance parameters* $\sigma_\varepsilon^2$, $\mu_\xi$ *and* $\sigma_\xi^2$ *are known and the conditions* (b) *and* (c) *hold true. Denote by* $b = (b_i)_{i=1}^p$ *the true value of regression coefficients. Then the following statements hold true.*

(a) *For any* $R > \|b\|$, eventually *there exists a unique solution to equation* (4.170) *on the ball*

$$\overline{B}_R = \{\beta \in \mathbf{R}^p\colon\ \|\beta\| \le R\}\,. \tag{4.171}$$

(b)

$$\hat{\beta}_{\mathrm{QL}} \xrightarrow{\mathrm{P1}} b\,. \tag{4.172}$$

(c)

$$\sqrt{n}\,(\hat{\beta}_{\mathrm{QL}} - b) \xrightarrow{d} \mathrm{N}(0, B_b)\,, \tag{4.173}$$

$$B_b = \left(\mathbf{E}\frac{f(x)f^{\mathrm{T}}(x)}{v(x, b)}\right)^{-1}\,. \tag{4.174}$$

*The estimate can be evaluated similarly to Section* 3.2.1. *Using methods of Kukush et al. (2009), one can prove the next analog of Theorem* 3.22.

**Theorem 4.18.** *Assume the conditions of Theorem* 4.17. *Let* $\Sigma_{\mathrm{C}}$ *and* $\Sigma_{\mathrm{QL}} = B_b$ *be the ACMs of the estimators* $\hat{\beta}_{\mathrm{C}}$ *and* $\hat{\beta}_{\mathrm{QL}}$ *in the model* (4.140)–(4.142).

(a) *If the true regression function* $b^{\mathrm{T}}\phi(\xi)$ *is constant (i.e., the dependence of* $y$ *on* $\xi$ *vanishes), then* $\Sigma_{\mathrm{QL}} = \Sigma_{\mathrm{C}}$.

(b) *If the true regression function is not constant, then* $\Sigma_{\mathrm{QL}} < \Sigma_{\mathrm{C}}$.

Now, suppose that the variance $\sigma_\varepsilon^2$ is unknown and nuisance parameters $\mu_\xi$ and $\sigma_\xi^2$ are still known. Then we estimate simultaneously the parameters $\beta$ and $\sigma_\varepsilon^2$, and the following analog of (3.231) is added to equation (4.170):

$$\sigma_\varepsilon^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - f^{\mathrm{T}}(x_i)\beta)^2 - \beta^{\mathrm{T}}\cdot\frac{1}{n}\sum_{i=1}^{n}[F(x_i) - f(x_i)f^{\mathrm{T}}(x_i)]\,\beta\,,$$
$$\beta \in \mathbf{R}^{k+1}\,, \quad \sigma_1^2 \le \sigma_\varepsilon^2 \le \sigma_2^2\,. \quad (4.175)$$

Here $0 < \sigma_1^2 < \sigma_2^2$ are given thresholds for the variance of error in response.

The estimators $\hat{\beta}_{\mathrm{QL}}$ and $\hat{\sigma}^2_{\varepsilon,\mathrm{QL}}$ are defined by equations (4.170) and (4.175) similarly to Definition 3.26. The analog of Theorem 3.27 is correct, where the ACM of the QL estimator for the parameter $\beta$ does not depend on whether we know the variance $\sigma_\varepsilon^2$ or not.

Now, let all the nuisance parameters $\sigma_\varepsilon^2$, $\mu_\xi$, and $\sigma_\xi^2$ be unknown. Preliminary estimators $\hat{\mu}_{\xi,\mathrm{QL}}$ and $\hat{\sigma}^2_{\xi,\mathrm{QL}}$ are constructed by formulas (3.269) and (3.270). Further, these estimators are substituted in the expressions for $f(x)$, $F(x)$, and $v(x;\beta,\sigma_\varepsilon^2)$ in the system (4.170) and (4.175); then the system defines new estimators $\hat{\beta}_{\mathrm{QL}}$ and $\hat{\sigma}^2_{\varepsilon,\mathrm{QL}}$. The analog of Theorem 3.31 is correct, which shows that the ACM of the estimator for the parameter $\beta$ increases compared to expression (4.174) (cf. formula (3.274) for the polynomial regression). At that, the technique of the paper by Kukush et al. (2009) does not allow to state that the ACM of the QL estimator for the parameter $\beta$ does not exceed the ACM of the CS estimator. The situation here differs from the polynomial regression, where the inequality $\Sigma_{\mathrm{QL}}^{(\beta)} \le \Sigma_{\mathrm{C}}^{(\beta)}$ for the ACMs is still valid after pre-estimation of $\mu_\xi$ and $\sigma_\xi^2$ (see Theorem 3.36 and discussion above it).

**Remark 4.19.** If the function $\phi$ from the regression equation (4.140) is known to be periodic, say, with a period $2\pi$, then the function $\beta^{\mathrm{T}}\phi(\xi)$ can be modeled by the trigonometric polynomial

$$\frac{a_0}{2} + \sum_{k=1}^{p}(a_k\cos k\xi + b_k\sin k\xi)\,.$$

For the corresponding model, one can create a theory similar to developed in this Section 4.3. The matter is that the sines and cosines are expressed through complex exponents by Euler's formula; this makes it easy to find analogs to the functions $t_\lambda(x)$, see (4.147), and $f_\lambda(x)$, see (4.157).

## 4.4 Three consistent estimators in Poisson log-linear model

We construct a regression model based on Example 4.4. Suppose $y$ has Poisson distribution (4.21):

$$y \sim \mathrm{Pois}(\lambda)\,, \quad \lambda = \mathrm{e}^{\eta}\,, \quad \eta = \beta_0 + \beta_1 \xi\,. \tag{4.176}$$

Additionally, we demand that instead of $\xi$, we observe $x$ in accordance with relations (4.141); moreover $\xi$ and $\delta$ are independent, and $\sigma_\delta^2$ is assumed known. We observe independent copies of the model $(y_i, x_i)$, $i = \overline{1,n}$. The parameter $\beta = (\beta_0, \beta_1)^{\mathrm{T}}$ is under estimation.

### 4.4.1 Corrected score Estimator

By formula (4.21), we have

$$\begin{aligned} &\ln \rho(y|\xi) = y\eta - \mathrm{e}^{\eta} - \ln y!\,, \\ &y \in N \cup \{0\}\,, \quad \eta = \beta_0 + \beta_1 \xi\,. \end{aligned} \tag{4.177}$$

In the model (4.176), the score function $s_{\mathrm{ML}}(y, \xi; \beta)$ corresponding to the observations $y$ and $\xi$ is equal to

$$s_{\mathrm{ML}} = \frac{\partial \ln \rho(y|\xi)}{\partial \beta} = (y - \mathrm{e}^{\eta}) \frac{\partial \eta}{\partial \beta}\,, \tag{4.178}$$

$$s_{\mathrm{ML}} = y\,(1, \xi)^{\mathrm{T}} - (\mathrm{e}^{\eta}, \xi \mathrm{e}^{\eta})^{\mathrm{T}}\,. \tag{4.179}$$

The corrected score of type (4.114) has the form

$$s_{\mathrm{C}}(y, x; \beta) = y g_{\mathrm{C}}(x) - h_{\mathrm{C}}(x, \beta)\,, \tag{4.180}$$

$$g_{\mathrm{C}}(x) = (1, x)^{\mathrm{T}}\,, \quad h_{\mathrm{C}} = \exp\left\{\beta_0 + \beta_1 x - \frac{1}{2}\beta_1^2 \sigma_\delta^2\right\} \begin{pmatrix} 1 \\ x - \sigma_\delta^2 \beta_1 \end{pmatrix}. \tag{4.181}$$

Indeed,

$$\mathbf{E}[g_{\mathrm{C}}(\xi + \delta)|\xi] = \mathbf{E}[(1, x)^{\mathrm{T}}|\xi] = (1, \xi)^{\mathrm{T}}\,, \tag{4.182}$$

$$\mathbf{E}[h_{\mathrm{C},0}(\xi + \delta, \beta)|\xi] = \mathbf{E}\left[\exp\left\{\beta_0 + \beta_1(\xi + \delta) - \tfrac{1}{2}\beta_1^2\sigma_\delta^2\right\} \,\middle|\, \xi\right] = \mathrm{e}^{\eta} = \mathrm{e}^{\beta_0 + \beta_1 \xi}\,. \tag{4.183}$$

We used an expression for the exponential moment of the normal random variable:

$$\mathbf{E}\mathrm{e}^{\beta_1 \delta} = \exp\left\{\tfrac{1}{2}\beta_1^2 \sigma_\delta^2\right\}\,. \tag{4.184}$$

Next, we differentiate the identity (4.183) with respect to $\beta_1$:

$$\mathbf{E}\left[\frac{\partial}{\partial \beta_1} h_{\mathrm{C},0}(x, \beta) \,\middle|\, \xi\right] = \mathbf{E}[h_{\mathrm{C},0}(x, \beta)\,(x - \sigma_\delta^2 \beta_1)|\xi] = \xi \mathrm{e}^{\eta}\,. \tag{4.185}$$

Thus, for the other component of the function $h_{\mathrm{C}}$, it holds that

$$\mathbf{E}[h_{\mathrm{C},1}(x,\beta)|\xi] = \mathbf{E}[h_{\mathrm{C},0}(x,\beta)\,(x-\sigma_\delta^2\beta_1)|\xi] = \xi \mathrm{e}^{\eta}\,. \tag{4.186}$$

From equalities (4.182), (4.183), and (4.186), we get the following relation for the function (4.180):

$$\mathbf{E}[s_{\mathrm{C}}(y,x;b|y,\xi] = s_{\mathrm{ML}}(y,\xi;b)\,, \quad b\in\mathbf{R}^2\,, \tag{4.187}$$

and therefore, the function (4.180) is estimating function of the corrected score method (see formula (1.118) in a general description of the method).

The estimator $\hat\beta_{\mathrm{C}}$ is defined by the nonlinear equation

$$\frac{1}{n}\sum_{i=1}^{n}(y_i g_{\mathrm{C}}(x_i) - h_{\mathrm{C}}(x_i,\beta)) = 0\,, \quad \beta\in\mathbf{R}^2\,. \tag{4.188}$$

**Definition 4.20.** The estimator $\hat\beta_{\mathrm{C}}$ is a Borel measurable function of observations $x_1,\ldots,y_n$, for which:

(a) if equation (4.188) has no solution, then $\hat\beta_{\mathrm{C}} = 0$,

(b) if the equation has a solution, then $\hat\beta_{\mathrm{C}}$ is a solution with the smallest norm (if there are several such solutions, then we take any of them).

**Theorem 4.21.** *Let the random latent variable $\xi$ be not constant and for each $c\in\mathbf{R}$,*

$$\mathbf{E}\mathrm{e}^{c\xi} < \infty\,. \tag{4.189}$$

*Then $\hat\beta_{\mathrm{C}}$ is strongly consistent estimator of the parameter $\beta$.*

*Proof* is based on Theorem A.15 from Appendix A1. Examine only the condition (e) of the latter theorem about the uniqueness of solution to the limit equation.

Let $b = (b_0,b_1)^{\mathrm{T}}$ be true value of the parameter $\beta$. The left-hand side of (4.188) converges almost surely to the function

$$S_\infty(b,\beta) = \mathbf{E}_b(yg_{\mathrm{C}}(x) - h_{\mathrm{C}}(x,\beta)) = \mathbf{E}(\mathrm{e}^{\eta_0} - \mathrm{e}^{\eta})(1,\xi)^{\mathrm{T}}\,. \tag{4.190}$$

Here $\eta_0 = b_0 + b_1\xi$ and $\eta = \beta_0 + \beta_1\xi$. The limit estimating equation is as follows:

$$S_\infty(b,\beta) = 0\,, \quad \beta\in\mathbf{R}^2\,. \tag{4.191}$$

We shall demonstrate that it has the unique solution $\beta = b$.

Denote

$$\Phi(\beta) = \mathbf{E}\mathrm{e}^{\beta_0+\beta_1\xi}(1,\xi)^{\mathrm{T}}\,, \quad \beta\in\mathbf{R}^2\,. \tag{4.192}$$

By condition (4.189), this vector function is well-defined. Calculate Jacobi's matrix

$$\Phi'(\beta) = \begin{pmatrix} \mathbf{E}\mathrm{e}^{\beta_0+\beta_1\xi} & \mathbf{E}\xi\mathrm{e}^{\beta_0+\beta_1\xi} \\ \mathbf{E}\xi\mathrm{e}^{\beta_0+\beta_1\xi} & \mathbf{E}\xi^2\mathrm{e}^{\beta_0+\beta_1\xi} \end{pmatrix}. \tag{4.193}$$

The matrix is positive definite, because its top left entry is positive and

$$\det \Phi'(\beta) = (\mathbf{E}e^{\beta_0+\beta_1\xi})(\mathbf{E}\xi^2 e^{\beta_0+\beta_1\xi}) - (\mathbf{E}\xi e^{\beta_0+\beta_1\xi})^2 > 0\,. \tag{4.194}$$

The latter relation follows from the Cauchy–Schwartz inequality. Here, the strict inequality holds, because the random variables $\xi \exp(\frac{\beta_0+\beta_1\xi}{2})$ and $\exp(\frac{\beta_0+\beta_1\xi}{2})$ are linearly independent.

Consider the inner product

$$(\Phi(b) - \Phi(\beta), b - \beta) = (b - \beta)^{\mathrm{T}} \Phi'(u)(b - \beta)\,. \tag{4.195}$$

Here $u$ is an intermediate point between $b$ and $\beta$; the Lagrange theorem has been applied to the function

$$q(t) = (\Phi(\beta + t \cdot \Delta\beta) - \Phi(\beta), \Delta\beta)\,, \quad \Delta\beta = b - \beta\,, \quad t \in [0, 1]\,. \tag{4.196}$$

Let $\beta$ be a solution to (4.191). Then from equality (4.195) we get $(\Delta\beta)^{\mathrm{T}} \Phi'(u) \cdot \Delta\beta = 0$. Now, positive definiteness of the matrix $\Phi'(u)$ leads to $\Delta\beta = 0$ and $\beta = b$.

We have checked that the limit equation (4.191) has a unique solution. The proof is accomplished.

Under the conditions of Theorem 4.21, the estimator $\hat{\beta}_{\mathrm{C}}$ is asymptotically normal. If $\xi$ has normal distribution, then the ACM of the estimator can be found by general formulas from Section 4.2.3.

**Remark 4.22.** From the first equation of system (4.188), $e^{\beta_0}$ can be expressed through $\beta_1$ and substituted to the second one. In so doing, an equation with respect to $\beta_1$ is obtained. It can be solved by numerical methods. And it can have many solutions. Among them one can take a solution with the smallest absolute value. If the estimator is defined in this way, then it is strongly consistent (although it may differ from the estimator described in Definition 4.20).

### 4.4.2 Simplified quasi-likelihood estimator

In Section 4.4.1, we did not specify the shape of distribution of $\xi$. Now, assume additionally the following:

$$\xi \sim \mathrm{N}(\mu_\xi, \sigma_\xi^2)\,, \quad \sigma_\xi^2 > 0\,. \tag{4.197}$$

The condition allows constructing an estimator of $\beta$ which has a smaller ACM than the CS estimator.

Using formulas (4.67) and (4.68), we write down the functions $m(x, \theta)$ and $v(x, \theta)$, with $\theta = (\beta_0, \beta_1, \mu_\xi, \sigma_\xi^2)^{\mathrm{T}}$. In the Poisson model, $C(\eta) = e^\eta$, $\eta = \beta_0 + \beta_1\xi$, and $\phi = 1$.

Then

$$m(x,\theta) = \mathbf{E}(y|x) = \mathbf{E}(\mathrm{e}^{\eta}|x) = \mathrm{e}^{\beta_0}\mathbf{E}(\mathrm{e}^{\beta_1\xi}|x)\,, \tag{4.198}$$

$$m(x,\theta) = \exp\left\{\beta_0 + \beta_1\mu_1(x) + \frac{\beta_1^2\tau^2}{2}\right\}\,, \quad \mu_1(x) = Kx + (1-k)\mu_\xi\,, \quad \tau^2 = K\sigma_\delta^2\,. \tag{4.199}$$

We used (4.157) and (4.158). Furthermore,

$$\begin{aligned} v(x,\theta) = \mathbf{V}(y|x) &= \mathbf{E}[C''(\eta)|x] + \mathbf{E}[(C'(\eta))^2|x] - m^2(x,\theta) = \\ &= m(x,\theta) + \exp\{2\beta_0 + 2\beta_1\mu_1(x) + 2\beta_1^2\tau^2\} - m^2(x,\theta)\,, \end{aligned} \tag{4.200}$$

$$v(x,\theta) = m^2(x,\theta)\,(\mathrm{e}^{\beta_1^2\tau^2} - 1) + m(x,\theta)\,. \tag{4.201}$$

Following Shklyar and Schneeweiss (2005), introduce a simplified QL estimating function

$$s_{\mathrm{S}}(y,x;\theta) = \frac{y - m(x,\theta)}{m(x,\theta)} \cdot \frac{\partial m(x,\theta)}{\partial\beta}\,. \tag{4.202}$$

The logic is as follows. The QL method requires the presence of $v(x,\theta)$ instead of $m(x,\theta)$ in the denominator. However, in the absence of measurement errors (i.e., when $\delta \equiv 0$) it holds that

$$v(x,\theta) = \mathbf{V}(y|\xi) = \mathrm{e}^{\eta} = \mathbf{E}(y|\xi) = m(x,\theta)\,. \tag{4.203}$$

We made use of a well-known property of the Poisson distribution: its variance and expectation coincide. Replacing $v(x,\theta)$ by $m(x,\theta)$, we lose a bit in efficiency of the estimator. However, the estimate produced by the estimating function (4.202) is easier to evaluate.

The simplified QL estimator $\hat\beta_{\mathrm{S}}$, under $\mu_\xi$ and $\sigma_\xi^2$ known, is given by the following equation:

$$\sum_{i=1}^{n} s_{\mathrm{S}}(y_i, x_i; \theta) = 0\,, \quad \beta \in \mathbf{R}^2\,. \tag{4.204}$$

Formally, the estimator $\hat\beta_{\mathrm{S}}$ can be defined similarly to Definition 4.20.

If the nuisance parameters $\mu_\xi$ and $\sigma_\xi^2$ are unknown, they are pre-estimated by formulas (3.269) and (3.270). The corresponding estimators are substituted in the estimating function $s_{\mathrm{S}}$, and then the estimator $\hat\beta_{\mathrm{S}}$ is defined by the estimating equation

$$\sum_{i=1}^{n} s_{\mathrm{S}}(y_i, x_i; \beta, \hat\mu_\xi, \hat\sigma_\xi^2) = 0\,, \quad \beta \in \mathbf{R}^2\,. \tag{4.205}$$

A formal definition of the estimator is similar to Definition 4.20.

The estimating function $s_{\mathrm{S}}$ is unbiased, because

$$\begin{aligned} \mathbf{E}_\theta s_{\mathrm{S}}(y,x;\theta) &= \mathbf{E}\mathbf{E}_\theta[s_{\mathrm{S}}(y,x;\theta)|x] = \\ &= \mathbf{E}\left\{\frac{1}{m(x,\theta)}\frac{\partial m(x,\theta)}{\partial\beta}\mathbf{E}_\theta[y - m(x,\theta)|x]\right\} = 0\,. \end{aligned} \tag{4.206}$$

It is, therefore, natural that the estimator $\hat{\beta}_S$ is strongly consistent. The estimator is asymptotically normal as well. In case of known $\mu_\xi$ and $\sigma_\xi^2$, its ACM is determined by the corresponding sandwich-formula. When the nuisance parameters are unknown, its ACM becomes a bit larger, namely, an additional "sandwich" appears, like in formula (4.111).

**Theorem 4.23.** *In the Poisson model, assume the condition* (4.197). *Let $\Sigma_C$ and $\Sigma_S$ be the ACMs of estimators $\hat{\beta}_C$ and $\hat{\beta}_S$, under $\mu_\xi$ and $\sigma_\xi^2$ known. If the true value $\beta_1 = 0$ then $\Sigma_C = \Sigma_S$, and if $\beta_1 \neq 0$ then $\Sigma_C < \Sigma_S$.*

*Proof* is given in Shklyar and Schneeweiss (2005).

Thus, if the nuisance parameters are assumed known, the estimator $\hat{\beta}_S$ is more efficient compared with $\hat{\beta}_C$. But under unknown $\mu_\xi$ and $\sigma_\xi^2$, the latter statement may be wrong.

To evaluate the estimator $\hat{\beta}_S$, we find $\beta_0$ from the first equation of the system (4.204):

$$e^{\beta_0} = e^{-\beta_1^2\tau^2} \cdot \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n e^{\beta_1\mu_1(x_i)}} . \tag{4.207}$$

Substituting this into the second equation of the system, we obtain the equation in $\beta_1$:

$$\frac{\sum_{i=1}^n e^{\beta_1\mu_1(x_i)} \cdot \mu_1(x_i)}{\sum_{i=1}^n e^{\beta_1\mu_1(x_i)}} \cdot \sum_{i=1}^n y_i = \sum_{i=1}^n y_i\mu_1(x_i) . \tag{4.208}$$

**Lemma 4.24.** *Denote $I = \{i = \overline{1,n}\colon y_i \geq 1\}$. If not all $y_i$, $i = \overline{1,n}$, are equal to 0 and not all $x_i$, $i \in I$, coincide, then equation* (4.208) *has a unique solution.*

*Proof.* Introduce the function

$$f(\beta_1) = \frac{\sum_{i=1}^n e^{\beta_1\mu_1(x_i)}\mu_1(x_i)}{\sum_{i=1}^n e^{\beta_1\mu_1(x_i)}} , \quad \beta_1 \in \mathbf{R} . \tag{4.209}$$

We have

$$f'(\beta_1) = \frac{\left(\sum_{i=1}^n \mu_1^2(x_i)e^{\beta_1\mu_1(x_i)}\right)\left(\sum_{i=1}^n e^{\beta_1\mu_1(x_i)}\right) - \left(\sum_{i=1}^n e^{\beta_1\mu_1(x_i)}\mu_1(x_i)\right)^2}{\left(\sum_{i=1}^n e^{\beta_1\mu_1(x_i)}\right)^2} , \quad \beta_1 \in \mathbf{R} . \tag{4.210}$$

From the Cauchy–Schwartz inequality, it follows $f'(\beta_1) \geq 0$; the equality is achieved only when the sets of numbers

$$\exp\left\{\tfrac{1}{2}\beta_1\mu_1(x_i)\right\}, i = \overline{1,n} , \quad \text{and} \quad \mu_1(x_i)\exp\left\{\tfrac{1}{2}\beta_1\mu_1(x_i)\right\} , \quad i = \overline{1,n} ,$$

are proportional. By the lemma's condition, this does not happen, and therefore, $f'(\beta_1) > 0$, $\beta_1 \in \mathbf{R}$, and the continuous function $f$ strictly increases. Hence, equation (4.208) has no more than one solution.

In addition,

$$\lim_{\beta_1\to+\infty} f(\beta_1) = \max_{i=\overline{1,n}} \mu_1(x_i) , \quad \lim_{\beta_1\to-\infty} f(\beta_1) = \min_{i=\overline{1,n}} \mu_1(x_i) . \tag{4.211}$$

With increasing $\beta_1$ from $-\infty$ to $+\infty$, the left-hand side of (4.208) takes all values from an open interval $(A, B)$, with

$$A = S \cdot \min_i \mu_1(x_i)\,, \quad B = S \cdot \max_i \mu_i(x_i)\,, \quad S = \sum_{i=1}^{n} y_i > 0\,. \tag{4.212}$$

At the same time, the right-hand side of equation (4.208) falls in the interval, because by the condition, not all $\mu_1(x_i)$, $i \in I$, coincide. Thus, there exists a solution to equation (4.208) and it is unique. The lemma is proved. □

Note that the event "not all $x_i$, $i \in I$, coincide" happens *eventually*. So, *eventually* the estimator $\hat{\beta}_S$ is uniquely defined.

Equations (4.208) can be solved numerically using standard dichotomy. It is as follows. One finds a segment $[\beta_{1*}, \beta_1^*]$ such that at the point $\beta_{1*}$, the left-hand side of (4.208) is smaller than the right-hand side, and the situation is opposite at the point $\beta_1^*$; then one takes the midpoint of the segment and finds the sign of inequality at it; then one takes the midpoint of that one of the two segments, where the desired root is located, etc.

### 4.4.3 Quasi-likelihood estimator

Consider the Poisson model under the condition (4.197). The QL estimating function $s_{\mathrm{QL}}^{(\beta)}(y, x; \beta)$ is given by equality (4.69), where the functions $m(x, \beta)$ and $v(x, \beta)$ are given in (4.199) and (4.201). Let the true value $\beta$ belong to the parameter set $\Theta \subset \mathbf{R}^2$. The nuisance parameters $\gamma = (\mu_\xi, \sigma_\xi^2)^{\mathrm{T}}$ are now assumed known.

**Definition 4.25.** The QL estimator $\hat{\beta}_{\mathrm{QL}}$ is a Borel measurable function of the observations $y_1, x_1, \ldots, y_n, x_n$, such that

(a) if the equation

$$\frac{1}{n} \sum_{i=1}^{n} s_{\mathrm{QL}}^{(\beta)}(y_i, x_i; \beta) = 0\,, \quad \beta \in \Theta\,, \tag{4.213}$$

has no solution then $\hat{\beta}_{\mathrm{QL}} = 0$;

(b) if the equation has solutions, then $\hat{\beta}_{\mathrm{QL}}$ is one of them.

**Definition 4.26.** A set $A \subset \mathbf{R}^p$ is called *convex* if for any $a, b \in A$, the segment

$$[a, b] = \{\lambda a + (1 - \lambda) b \colon 0 \le \lambda \le 1\} \tag{4.214}$$

is included in $A$ as well.

For instance, for a convex function $f \colon \mathbf{R} \to \mathbf{R}$, its "overgraph" $\{(x, y) \colon x \in \mathbf{R}, y \ge f(x)\}$ is a convex set.

**Theorem 4.27.** *In the Poisson model, assume the condition* (4.197). *Let the parameter set $\Theta$ be compact and convex in $\mathbf{R}^2$, and moreover, for all $b, \beta \in \Theta^0$, the next matrix $\Phi_{b\beta}$*

*be nonsingular:*

$$2\Phi_{b\beta} = \mathbf{E}\frac{\frac{\partial m(x,b)}{\partial \beta}\cdot\left(\frac{\partial m(x,\beta)}{\partial \beta}\right)^{\mathrm{T}}}{v(x,\beta)} + \mathbf{E}\frac{\frac{\partial m(x,\beta)}{\partial \beta}\cdot\left(\frac{\partial m(x,b)}{\partial \beta}\right)^{\mathrm{T}}}{v(x,\beta)} . \tag{4.215}$$

*Let the true value $\beta$ be an interior point of $\Theta$.*

*Then* eventually *equation* (4.213) *has a solution, and $\hat{\beta}_{\mathrm{QL}}$ is the strongly consistent estimator of $\beta$. In addition,*

$$\sqrt{n}\,(\hat{\beta}_{\mathrm{QL}} - \beta) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \Phi_{\beta}^{-1}) , \tag{4.216}$$

*with $\Phi_{\beta} = \Phi_{\beta\beta}$, i.e., $\Phi_{\beta}$ is obtained from* (4.215)*, with $b = \beta$.*

*Proof* is based on the appropriate theorems from Appendices A1 and A2. Verify only the basic condition on solutions to the limit equation. Now, let $\beta \in \Theta^0$ be true value of regression parameter.

For all $\beta \in \Theta$, as $n \to \infty$, almost surely the left-hand side of (4.213) tends to the function

$$S_{\infty}(b,\beta) = \mathbf{E}_b s_{\mathrm{QL}}^{(\beta)}(y,x;\beta) = \mathbf{E}\frac{m(x,b) - m(x,\beta)}{v(x,\beta)}\cdot\frac{\partial m(x,\beta)}{\partial \beta} . \tag{4.217}$$

The limit equation is

$$S_{\infty}(b,\beta) = 0 , \quad \beta \in \Theta . \tag{4.218}$$

We demonstrate that it has the unique solution $\beta = b$.

Let $\beta$ satisfy this equation at a given $b \in \Theta^0$. Consider the function

$$q(t) = (S_{\infty}(\beta + t(b-\beta), \beta), b - \beta) , \quad t \in [0,1] . \tag{4.219}$$

This scalar function is well-defined due to the covexity of $\Theta$. We have $q(0) = q(1) = 0$. Then by Rolle's theorem, for some intermediate point $\bar{\beta} = \beta + \bar{t}(b-\beta)$, $0 < \bar{t} < 1$, it holds that

$$0 = q'(\bar{t}) = (b-\beta)^{\mathrm{T}}\frac{\partial S_{\infty}(\bar{\beta},\beta)}{\partial b^{\mathrm{T}}}(b-\beta) . \tag{4.220}$$

Since $\beta \in \Theta^0$, then $\bar{\beta}$ belongs to $\Theta^0$ as well (here we use the convexity of $\Theta$). But by the theorem's condition, the matrix

$$\frac{\partial S_{\infty}(\bar{\beta},\beta)}{\partial b^{\mathrm{T}}} + \left(\frac{\partial S_{\infty}(\bar{\beta},\beta)}{\partial b^{\mathrm{T}}}\right)^{\mathrm{T}} = 2\Phi_{\bar{\beta}\beta} \tag{4.221}$$

is positive definite. Then equation (4.220) implies $\beta = b$. Thus, the limit equation (4.218) indeed has a unique solution.

**Remark 4.28.** The condition of Theorem 4.27 concerning the matrix $\Phi_{b\beta}$ is quite restrictive. At $b = \beta$, the matrix (4.215) is already positive definite. Therefore, Theorem 4.27 can be applied when the parameter set $\Theta$ does not allow the parameter $b$ to be much removed from the true value $\beta$.

**Theorem 4.29.** *Assume the conditions of Theorem* 4.27. *Let* $\Sigma_{\mathrm{S}}$ *and* $\Sigma_{\mathrm{QL}} = \Phi_\beta^{-1}$ *be the ACMs of estimators* $\hat\beta_{\mathrm{S}}$ *and* $\hat\beta_{\mathrm{QL}}$, *under known* $\mu_\xi$ *and* $\sigma_\xi^2$. *If the true value* $\beta_1 = 0$, *then* $\Sigma_{\mathrm{QL}} = \Sigma_{\mathrm{S}}$, *and if* $\beta_1 \neq 0$, *then* $\Sigma_{\mathrm{QL}} < \Sigma_{\mathrm{S}}$.

*Proof* is given in Shklyar and Schneeweiss (2005).

Thus, if $\mu_\xi$ and $\sigma_\xi^2$ are known, the QL estimator is more efficient than the estimator $\hat\beta_{\mathrm{S}}$. But the estimator $\hat\beta_{\mathrm{QL}}$ is much more difficult to compute. Based on the estimator $\hat\beta_{\mathrm{S}}$, one can construct a new estimator $\tilde\beta_{\mathrm{QL}}$, which is easy to evaluate and has the same ACM as $\hat\beta_{\mathrm{QL}}$ has.

The idea is as follows. By Newton's method, we try to solve equation (4.213):

$$S_{\mathrm{QL}}(\beta) = 0\,, \quad S_{\mathrm{QL}}(\beta) = \frac{1}{n}\sum_{i=1}^{n} s_{\mathrm{QL}}^{(\beta)}(y_i, x_i; \beta)\,. \tag{4.222}$$

As initial approximation, we take the asymptotically normal estimator $\hat\beta_{\mathrm{S}}$. Let us make one step according to Newton's method. For this, we write down an approximation to $S_{\mathrm{QL}}(\beta)$ in the neighborhood of $\hat\beta_{\mathrm{S}}$ following Taylor's formula:

$$S_{\mathrm{QL}}(\beta) \approx S_{\mathrm{QL}}(\hat\beta_{\mathrm{S}}) + S'_{\mathrm{QL}}(\hat\beta_{\mathrm{S}})(\beta - \hat\beta_{\mathrm{S}})\,. \tag{4.223}$$

Equating the latter expression to zero, we find a modified estimator $\tilde\beta_{\mathrm{QL}}$:

$$\tilde\beta_{\mathrm{QL}} = \hat\beta_{\mathrm{S}} - [S'_{\mathrm{QL}}(\hat\beta_{\mathrm{S}})]^{-1} S_{\mathrm{QL}}(\hat\beta_{\mathrm{S}})\,. \tag{4.224}$$

Because of the strong consistency of $\hat\beta_{\mathrm{S}}$, we obtain

$$S'_{\mathrm{QL}}(\hat\beta_{\mathrm{S}}) \xrightarrow{\mathrm{P1}} \mathbf{E}_\beta \frac{\partial s_{\mathrm{QL}}^{(\beta)}(y, x; \beta)}{\partial \beta^{\mathrm{T}}} = -\Phi_\beta < 0\,. \tag{4.225}$$

Hence *eventually* the Jacobian matrix $S'_{\mathrm{QL}}(\hat\beta_{\mathrm{S}})$ is nonsingular, and the estimator $\tilde\beta_{\mathrm{QL}}$ is *eventually* well-defined by equality (4.224).

**Theorem 4.30.** *In the Poisson model, assume the condition* (4.197). *Then* $\tilde\beta_{\mathrm{QL}}$ *is a strongly consistent estimator of* $\beta$, *and moreover* $\sqrt{n}\,(\tilde\beta_{\mathrm{QL}} - \beta) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \Phi_\beta^{-1})$.

The statement of this theorem stems from Theorem 7.75 of the textbook by Schervish (1995).

We give a practical recommendation: one need not solve the quite complicated nonlinear equation (4.213), but instead it is much better to compute the estimator $\tilde\beta_{\mathrm{QL}}$ having previously calculated $\hat\beta_{\mathrm{S}}$. For that, no prior information on the parameters $\beta_0$ and $\beta_1$ is required.

Now let the nuisance parameters $\mu_\xi$ and $\sigma_\xi^2$ be unknown. Then the estimator $\hat\beta_{\mathrm{QL}}$ has to be modified in accordance with recommendations of Section 4.2.2. Pre-estimate $\gamma = (\mu_\xi, \sigma_\xi^2)^{\mathrm{T}}$ by means of (3.269) and (3.270) and denote

$$\hat s_{\mathrm{QL}}^{(\beta)}(y, x; \beta) = s^{(\beta)}(y, x; \beta, \hat\gamma_{\mathrm{QL}})\,. \tag{4.226}$$

The estimator $\hat{\beta}_{\mathrm{QL}}$ corresponds to the equation

$$\frac{1}{n}\sum_{i=1}^{n}\hat{s}_{\mathrm{QL}}^{(\beta)}(y_i, x_i; \beta) = 0\,, \quad \beta \in \Theta\,. \tag{4.227}$$

A formal definition of the estimator is quite similar to Definition 4.25. An analog of Theorem 4.27 holds true, but the ACM of the estimator is larger than the matrix $\Phi_{\beta}^{-1}$ ("additional sandwiches" appear like in formula (4.111)).

**Theorem 4.31.** *Assume the conditions of Theorem* 4.27 *and suppose that the parameters* $\mu_\xi$ *and* $\sigma_\xi^2$ *are unknown. Let* $\Sigma_{\mathrm{C}}$, $\Sigma_{\mathrm{S}}$, *and* $\Sigma_{\mathrm{QL}}$ *be the ACMs of estimators* $\hat{\beta}_{\mathrm{C}}$, $\hat{\beta}_{\mathrm{S}}$, *and* $\hat{\beta}_{\mathrm{QL}}$, *respectively. If the true value* $\beta_1 = 0$, *then* $\Sigma_{\mathrm{C}} = \Sigma_{\mathrm{S}} = \Sigma_{\mathrm{QL}}$, *and if* $\beta_1 \neq 0$, *then the following inequalities hold:*

$$\Sigma_{\mathrm{QL}} < \Sigma_{\mathrm{C}}\,, \quad \Sigma_{\mathrm{QL}} < \Sigma_{\mathrm{S}}\,. \tag{4.228}$$

*Proof* is carried out by technique of the paper by Kukush et al. (2009), see also Kukush et al. (2007).

In case $\beta_1 \neq 0$, Shklyar (2006) proved a stronger inequality $\Sigma_{\mathrm{QL}} < \Sigma_{\mathrm{S}} < \Sigma_{\mathrm{C}}$.

Thus, if the nuisance parameters are unknown, the estimator $\hat{\beta}_{\mathrm{QL}}$ is more efficient than the other two estimators. However, it is difficult to compute. Instead, it is possible to propose an analog of the estimator (4.224):

$$\tilde{\beta}_{\mathrm{QL}} = \hat{\beta}_{\mathrm{S}} - [\hat{S}'_{\mathrm{QL}}(\hat{\beta}_{\mathrm{S}})]^{-1}\hat{S}_{\mathrm{QL}}(\hat{\beta}_{\mathrm{S}})\,, \tag{4.229}$$

$$\hat{S}_{\mathrm{QL}}(\beta) = \frac{1}{n}\sum_{i=1}^{n}\hat{s}_{QL}^{(\beta)}(y_i, x_i; \beta)\,, \quad \beta \in \mathbf{R}^2\,. \tag{4.230}$$

By Theorem 7.75 from the monograph by Schervish (1995), the estimator $\tilde{\beta}_{\mathrm{QL}}$ is strongly consistent and asymptotically normal, and its ACM coincides with the ACM of the estimator $\hat{\beta}_{\mathrm{QL}}$. So, we suggest using the estimator (4.229), if it is known for certain that the latent variable $\xi$ has normal distribution (though with unknown parameters of the distribution).

## 4.5 Two consistent estimators in logistic model

We construct a logistic regression model based on Example 4.7. Suppose that $y$ has Bernoulli distribution (4.34), with $\lambda = \mathrm{e}^{\eta}$:

$$\mathbf{P}\{y = 1\} = H(\eta)\,, \quad \mathbf{P}\{y = 0\} = 1 - H(\eta)\,, \quad H(\eta) = \frac{\mathrm{e}^{\eta}}{1 + \mathrm{e}^{\eta}}\,; \tag{4.231}$$

$$\eta = \beta_0 + \beta_1\xi\,. \tag{4.232}$$

The regressor $\xi$ is random variable.

### 4.5.1 Evaluation of MLE in model without measurement error

Let $(y_i, \xi_i)$, $i = \overline{1, n}$, be observed independent copies of the model (4.231) and (4.232). We construct the MLE of the regression coefficients $\beta = (\beta_0, \beta_1)^{\mathrm{T}}$.

By formula (4.38) we get

$$\ln \rho(y|\xi) = y\eta - \ln(1 + \mathrm{e}^{\eta}) , \quad y = 0;\ 1 . \tag{4.233}$$

The score function is

$$s_{\mathrm{ML}}(y, \xi; \beta) = \frac{\partial \ln \rho(y|\xi)}{\partial \beta} = (y - H(\eta)) \frac{\partial \eta}{\partial \beta} , \tag{4.234}$$

$$s_{\mathrm{ML}}(y, \xi; \beta) = (y - H(\eta)) \cdot \begin{pmatrix} 1 \\ \xi \end{pmatrix} . \tag{4.235}$$

The MLE $\hat{\beta}_{\mathrm{ML}}$ satisfies the estimating equation

$$\sum_{i=1}^{n} (y_i - H(\eta_i)) \cdot \begin{pmatrix} 1 \\ \xi_i \end{pmatrix} = 0 , \quad \eta_i = \beta_0 + \beta_1 \xi_i , \quad \beta \in \mathbf{R}^2 . \tag{4.236}$$

A natural assumption is that the regressor $\xi$ has finite variance $\mathbf{D}\xi = \sigma_{\xi}^2 > 0$. Under this condition, equation (4.236) has *eventually* a unique solution. Formally, the estimator $\hat{\beta}_{\mathrm{ML}}$ can be defined by equation (4.236) like in Definition 4.25. Then it is the strongly consistent estimator, and moreover,

$$\sqrt{n}\,(\hat{\beta}_{\mathrm{ML}} - \beta) \xrightarrow{d} \mathrm{N}(0, \Phi^{-1}) , \tag{4.237}$$

$$\Phi = -\mathbf{E} \frac{\partial s_{\mathrm{ML}}(y, \xi; \beta)}{\partial \beta^{\mathrm{T}}} = \mathbf{E} H'(\eta) \cdot \begin{pmatrix} 1 \\ \xi \end{pmatrix} \cdot (1, \xi) , \tag{4.238}$$

$$\Phi = \mathbf{E} H(\eta)\,(1 - H(\eta))\,(1, \xi)^{\mathrm{T}} (1, \xi) . \tag{4.239}$$

Equation (4.236) can be solved by the iteratively reweighted least squares method. For the logistic model, the method is described in Myers et al. (2002).

The idea of the method is as follows. Notice that

$$\mathbf{E}_{\beta}(y|\xi) = H(\eta) , \quad \eta = \beta_0 + \beta_1 \xi ; \tag{4.240}$$

$$v(\xi, \beta) = \mathbf{V}_{\beta}(y|\xi) = H(\eta)\,(1 - H(\eta)) . \tag{4.241}$$

We applied formula (4.42) to evaluate the conditional variance.

Further, given the identity $H'(\eta) = H(\eta)\,(1 - H(\eta))$, rewrite the estimating function (4.235) in the form

$$s_{\mathrm{ML}}(y, \xi; \beta) = \frac{y - H(\eta)}{v(\xi, \beta)} \frac{\partial H(\eta)}{\partial \beta} , \tag{4.242}$$

$$s_{\mathrm{ML}} = -\frac{1}{2} \frac{\partial}{\partial \beta} \left( \frac{(y - H(\eta))^2}{v(\xi, b)} \right) \Bigg|_{b=\beta} , \quad \eta = \beta_0 + \beta_1 \xi . \tag{4.243}$$

Thus, making differentiation we "freeze" the denominator regarding it independent of the regression parameter $\beta$.

Equality (4.243) hints the next iteration procedure. Start with some value $\beta^{(0)} \in \mathbf{R}^2$. Put $w_1(\xi) = v(\xi, \beta^{(0)})$. The value $\beta^{(1)}$ is found as a global minimum point of the objective function

$$Q(\beta; w_1) = \sum_{i=1}^{n} \frac{(y_i - H(\eta_i))^2}{w_1(\xi_i)} \ . \tag{4.244}$$

If $\beta^{(k)}$ have been already constructed, then put $w_{k+1}(\xi) = v(\xi, \beta^{(k)})$ and find $\beta^{(k+1)}$ as a global minimum point of the renewed objective function $Q(\beta; w_k)$. Construction of the iterations should last until the convergence criterion is fulfilled:

$$\frac{\left\|\beta^{(k+1)} - \beta^{(k)}\right\|}{\left\|\beta^{(k)}\right\|} < \delta \ . \tag{4.245}$$

Here $\delta$ is some fixed small number, say $10^{-6}$.

Constructing $\beta^{(k+1)}$ as a minimum point of the function $Q(\beta, w_k)$ can be carried out by its own iterative procedure using standard linearization of the new "regression function" $f(\beta) = H(\eta)$, see Section 3.3.4 in Myers et al. (2002). Thus, we obtain two nested iterative procedure for calculation of the estimate $\hat{\beta}_{\mathrm{ML}}$.

Notice that $\beta^{(0)}$ is worth determining by the ordinary least squares method, i.e., by minimization of the function

$$Q_0(\beta) = \sum_{i=1}^{n} (y_i - H(\eta_i))^2 \ , \quad \beta \in \mathbf{R}^2 . \tag{4.246}$$

### 4.5.2 Conditional score estimator

Now, assume that regressor is observed with the classical error:

$$x = \xi + \delta \ , \quad \delta \sim \mathrm{N}(0, \sigma_\delta^2) \ , \quad \sigma_\delta^2 > 0 \ . \tag{4.247}$$

Here $\xi$ and $\delta$ are independent, and $\sigma_\delta^2$ is assumed known.

An attempt to construct the corrected score estimator in the logistics model does not succeed. Indeed, according to this method, it is necessary to construct the corrected score $s_{\mathrm{C}}(y, x; \beta)$ such that for all $\beta \in \mathbf{R}^2$:

$$\mathbf{E}[s_{\mathrm{C}}(y, x; \beta)|y, \xi] = s_{\mathrm{ML}}(y, \xi; \beta) = (y - H(\eta)) \begin{pmatrix} 1 \\ \xi \end{pmatrix} . \tag{4.248}$$

To do this, one should find solutions $g_{\mathrm{C}}$ and $h_{\mathrm{C}}$ to the following deconvolution equations:

$$\mathbf{E}[g_{\mathrm{C}}(x, \beta)|\xi] = H(\eta) = \frac{1}{1 + \mathrm{e}^{-\beta_0 - \beta_1 \xi}} \ , \tag{4.249}$$

$$\mathbf{E}[h_{\mathrm{C}}(x, \beta)|\xi] = \xi H(\eta) \ , \quad \beta \in \mathbf{R}^2 . \tag{4.250}$$

Solutions to such equations are searched as entire functions in $x$, i.e., the functions which can be extended to the complex plane $\mathbf{C}$ preserving its analyticity. But the right-hand sides of equations (4.249) and (4.250) at $\beta_1 \neq 0$ have complex roots in the denominator, in particular, $-\beta_0 - \beta_1\xi = i\pi$, $\xi = -\beta_1^{-1}(\beta_0 + i\pi)$ where $i$ is imaginary unit. This fact does not allow solving the deconvolution equations and makes the CS method not applicable in this case.

Instead, Stefanski and Carroll (1987) proposed the conditional score method. The method is as follows. Put

$$z = x + y\sigma_\delta^2\beta_1\,, \quad \eta_* = \beta_0 + \beta_1 z\,. \tag{4.251}$$

It turns out that in the model (4.231), (4.232), and (4.247),

$$m_* = \mathbf{E}_\beta(y|z) = H(\eta_* - \tfrac{1}{2}\beta_1^2\sigma_\delta^2)\,, \tag{4.252}$$

$$v_* = \mathbf{V}_\beta(y|z) = H\,(1-H)\,, \quad H = H(\eta_* - \tfrac{1}{2}\beta_1^2\sigma_\delta^2)\,. \tag{4.253}$$

The proof of (4.252) is given in Carroll et al. (2006, p. 158); formula (4.253) is just a simple consequence of (4.252). A new estimating function $s_{\mathrm{D}}(y, x; \beta)$ is formed out of the estimating function (4.234) by substitution $\eta_*$ instead of $\eta$:

$$s_{\mathrm{D}}(y, x; \beta) = (y - m_*)\begin{pmatrix}1\\ z\end{pmatrix}\,. \tag{4.254}$$

The conditional score estimator $\hat\beta_{\mathrm{D}}$ is defined by the equation

$$\sum_{i=1}^{n} s_{\mathrm{D}}(y_i, x_i; \beta) = 0\,, \quad \beta \in \Theta \subset \mathbf{R}^2\,. \tag{4.255}$$

Estimating function (4.254) is unbiased because

$$\mathbf{E}_\beta s_{\mathrm{D}}(y, x; \beta) = \mathbf{E}\mathbf{E}_\beta[s_{\mathrm{D}}(y, x; \beta)|z] = 0\,. \tag{4.256}$$

Here we utilized the equality (4.252). The unbiasedness of the estimating function causes (under appropriate restrictions on $\Theta$) the consistency and asymptotic normality of the estimator $\hat\beta_{\mathrm{D}}$. At that, it is necessary that the true value $\beta$ be interior point of $\Theta$ and the regressor $\xi$ have a positive and finite variance $\sigma_\xi^2$. The ACM $\Sigma_{\mathrm{D}}$ of the estimator $\hat\beta_{\mathrm{D}}$ can be found by the sandwich formula.

Nonlinear equation (4.255) can be solved numerically using an iterative procedure like Newton–Raphson, see details in Carroll et al. (2006, p. 175). As initial approximation one can take the naive estimator $\hat\beta_{\mathrm{naive}}$, which is a solution to the equation

$$\sum_{i=1}^{n} s_{\mathrm{ML}}(y_i, x_i; \beta) = 0\,, \quad \beta \in \mathbf{R}^2\,. \tag{4.257}$$

The method of solving the latter equation was discussed in Section 4.5.1.

### 4.5.3 Quasi-likelihood estimator

Additionally, the normality condition is imposed:

$$\xi \sim \mathrm{N}(\mu_\xi, \sigma_\xi^2)\,, \quad \sigma_\xi^2 > 0\,. \tag{4.258}$$

Using formula (4.67), we write down the conditional mean $m(x, \theta)$, with $\theta^{\mathrm{T}} = (\beta, \alpha)$, $\alpha = (\mu_\xi, \sigma_\xi^2)^{\mathrm{T}}$:

$$m(x, \theta) = \mathbf{E}(y|x) = \mathbf{E}[H(\beta_0 + \beta_1\mu_1(x) + \beta_1\tau\gamma)|x]\,. \tag{4.259}$$

Here $\gamma \sim \mathrm{N}(0, 1)$, $x \perp\!\!\!\perp \gamma$, $\mu_1(x) = Kx + (1-K)\mu_\xi$, and $\tau = \sigma_\delta\sqrt{K}$. Let $g(t)$ be the standard normal pdf, $g(t) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}t^2)$. The conditional mean is expressed by an integral being not calculated in a closed form but only approximately:

$$m(x, \theta) = \int_{\mathbf{R}} H(\beta_0 + \beta_1\mu_1(x) + \beta_1\tau t)\, g(t)\, dt\,. \tag{4.260}$$

The conditional variance is

$$v(x, \theta) = \mathbf{V}(y|x) = \mathbf{E}(y^2|x) - m^2(x, \theta) = m(x, \theta)\,(1 - m(x, \theta))\,. \tag{4.261}$$

The QL estimating function is

$$s_{\mathrm{QL}} = \frac{y - m(x, \theta)}{v(x, \theta)}\frac{\partial m(x, \theta)}{\partial \beta} = \frac{y - m}{m\,(1-m)}\frac{\partial m}{\partial \beta}\,. \tag{4.262}$$

**The case where the nuisance parameters $\mu_\xi$ and $\sigma_\xi^2$ are known**

In this case, the estimating function (4.262) is just the estimating function of the ML method in the logistic model with measurement error (4.247).

Indeed, similarly to formula (4.37), we have

$$\ln \rho(y|x) = y(\ln m - \ln(1-m)) + \ln(1-m)\,, \quad m = m(x, \theta)\,. \tag{4.263}$$

Further, the estimating function of the ML method is

$$s = \frac{\partial \ln \rho(y|x)}{\partial \beta} = \frac{\partial \ln \rho(y|x)}{\partial m}\frac{\partial m}{\partial \beta} = \frac{y - m}{m\,(1-m)}\frac{\partial m}{\partial \beta} = s_{\mathrm{QL}}\,. \tag{4.264}$$

The QL estimator coincides with the MLE and can be formally defined as in Definition 4.25, with an open parameter set $\Theta \subset \mathbf{R}^2$. A complete analog of Theorem 4.27 holds true. The ACM of the estimator $\hat{\beta}_{\mathrm{QL}}$ is given by the expression

$$\Sigma_{\mathrm{QL}} = \Phi_\beta^{-1}\,, \quad \Phi_\beta = \mathbf{E}\frac{\frac{\partial m}{\partial \beta}\left(\frac{\partial m}{\partial \beta}\right)^{\mathrm{T}}}{m\,(1-m)}\,, \quad m = m(x; \beta, \alpha)\,.$$

**Theorem 4.32.** *Let the parameters $\mu_\xi$ and $\sigma_\xi^2$ be known, and $\Sigma_{\mathrm{D}}$ and $\Sigma_{\mathrm{QL}}$ be the ACMs of the conditional score estimator and QL estimator, respectively. If $\beta_1 = 0$, then $\Sigma_{\mathrm{D}} = \Sigma_{\mathrm{QL}}$, and if $\beta_1 \neq 0$, then $\Sigma_{\mathrm{QL}} < \Sigma_{\mathrm{D}}$.*

This theorem follows from the results of Kukush et al. (2009).

**The case of unknown nuisance parameters**

In this case, $\mu_\xi$ and $\sigma_\xi^2$ are pre-estimated by formulas (3.269) and (3.270). Denote

$$\hat{s}_{\mathrm{QL}}(y, x; \beta) = s_{\mathrm{QL}}(y, x; \beta, \hat{\alpha}) , \quad \hat{\alpha} = (\hat{\mu}_\xi, \hat{\sigma}_\xi^2)^{\mathrm{T}} . \tag{4.265}$$

A new estimator $\hat{\beta}_{\mathrm{QL}}$ corresponds to the equation

$$\sum_{i=1}^{n} \hat{s}_{\mathrm{QL}}(y_i, x_i; \beta) = 0 , \quad \beta \in \Theta . \tag{4.266}$$

The ACM of this estimator is larger than the matrix $\Phi_\beta^{-1}$, see formula (4.111). The estimator $\hat{\beta}_{\mathrm{QL}}$ constructed in this way will no longer be the MLE. We cannot guarantee that the ACM of estimator $\hat{\beta}_{\mathrm{QL}}$ does not exceed the ACM of the estimator $\hat{\beta}_{\mathrm{D}}$. But in case $\beta_1 = 0$, the latter ACMs coincide.

## 4.6 Two consistent estimators in log-linear gamma model

In Example 4.5, put $\lambda = \alpha/\omega$, $\omega > 0$, and $\eta = -\omega^{-1}$. Then by formula (4.25),

$$\ln \rho(y|\eta, \alpha) = \frac{y\eta + \ln(-\eta)}{\alpha^{-1}} + c(y, \alpha) , \quad y > 0 . \tag{4.267}$$

This pdf belongs to the exponential family (4.2), with $\eta < 0$, $C(\eta) = -\ln(-\eta)$, and dispersion parameter $\phi = \alpha^{-1}$. We get

$$\mathbf{E}y = C'(\eta) = -\frac{1}{\eta} , \quad \mathbf{D}y = \alpha^{-1} C''(\eta) = \frac{1}{\alpha\eta^2} . \tag{4.268}$$

Log-linear gamma model is given by the relation

$$\omega = \mathrm{e}^{\beta_0 + \beta_1 \xi} . \tag{4.269}$$

Instead of $\xi$, we observe $x$ according to (4.247).

### 4.6.1 Corrected score estimator

We have $\eta = -\mathrm{e}^{-\beta_0 - \beta_1 \xi}$. By (4.47),

$$s_{\mathrm{ML}}(y, \xi; \beta) = (y - \mathrm{e}^{\beta_0 + \beta_1 \xi}) \eta'_\beta = (y e^{-\beta_0 - \beta_1 \xi} - 1) \begin{pmatrix} 1 \\ \xi \end{pmatrix} . \tag{4.270}$$

Using formulas (4.183) and (4.185) and reasoning as in Section 4.4.1, we find that the corrected score is as follows:

$$s_{\mathrm{C}}(y, x; \beta) = y g_{\mathrm{C}}(x, \beta) - h_{\mathrm{C}}(x) , \tag{4.271}$$

$$g_{\mathrm{C}}(x, \beta) = \exp\{-\beta_0 - \beta_1 x - \tfrac{1}{2}\beta_1^2 \sigma_\delta^2\} (1, x + \sigma_\delta^2 \beta_1)^{\mathrm{T}} , \tag{4.272}$$

$$h_{\mathrm{C}}(x) = (1, x)^{\mathrm{T}} . \tag{4.273}$$

Writing formula (4.272), we utilized the second function of (4.181), where $-\beta_0$ was substituted for $\beta_0$ and $-\beta_1$ was substituted for $\beta_1$. The estimating function (4.271) satisfies equation (4.187), where the right-hand side is taken from (4.270).

The estimator $\hat{\beta}_{\mathrm{C}}$ is defined by the nonlinear equation

$$\frac{1}{n}\sum_{i=1}^{n}(y_i g_{\mathrm{C}}(x_i,\beta)-h_{\mathrm{C}}(x_i))=0\,,\quad \beta\in\mathbf{R}^2\,. \tag{4.274}$$

A formal definition of $\hat{\beta}_{\mathrm{C}}$ is now given by Definition 4.20 word for word.

**Theorem 4.33.** *Let the random latent variable $\xi$ be not constant and for each $c\in\mathbf{R}$, the condition* (4.189) *holds true. Then the estimator $\hat{\beta}_{\mathrm{C}}$ is strongly consistent.*

*Proof* is conducted in a manner similar to the proof of Theorem 4.21. Let $b=(b_0,b_1)^{\mathrm{T}}$ be the true value of the parameter $\beta$. The left-hand side of (4.274) converges almost surely to the function

$$S_\infty(b,\beta)=\mathbf{E}_b(yg_{\mathrm{C}}(x,\beta)-h_{\mathrm{C}}(x))=\mathbf{E}_b s_{\mathrm{ML}}(y,\xi;\beta)\,, \tag{4.275}$$

$$S_\infty(b,\beta)=\mathbf{E}(\mathrm{e}^{(b-\beta_0)+(b-\beta_1)\xi}-1)\begin{pmatrix}1\\ \xi\end{pmatrix}. \tag{4.276}$$

Denote

$$\Phi(\beta)=\mathbf{E}(\mathrm{e}^{\beta_0+\beta_1\xi}-1)\begin{pmatrix}1\\ \xi\end{pmatrix},\quad \beta\in\mathbf{R}^2\,. \tag{4.277}$$

The limit equation (4.191) takes the form

$$\Phi(b-\beta)=0\,,\quad \beta\in\mathbf{R}^2\,. \tag{4.278}$$

The derivative $\Phi'(\beta)$ is given by (4.193), whence $\Phi'(\beta)>0$, $\beta\in\mathbf{R}^2$. As a result of this, as in the proof of Theorem 4.21, we obtain that the limit equation (4.278) has a unique solution $b=\beta$. This completes the proof.

Everything we wrote above concerning the asymptotic normality and computation of the estimator $\hat{\beta}_{\mathrm{C}}$ in the Poisson model (see Section 4.4.1) is transferred to the estimator $\hat{\beta}_{\mathrm{C}}$ in the gamma model.

### 4.6.2 Quasi-likelihood estimator

Assume the normality condition

$$\xi\sim\mathrm{N}(\mu_\xi,\sigma_\xi^2)\,,\quad \sigma_\xi^2>0\,. \tag{4.279}$$

Now, in the gamma model

$$m^*(\xi;\beta)=\mathbf{E}(y|\xi)=\mathrm{e}^{\beta_0+\beta_1\xi}\,, \tag{4.280}$$

$$v^*(\xi;\beta,\alpha)=\frac{1}{\alpha}\mathrm{e}^{2\beta_0+2\beta_1\xi}\,. \tag{4.281}$$

The conditional mean $m(x, \theta)$, $\theta^{\mathrm{T}} = (\beta^{\mathrm{T}}, \alpha, \mu_\xi, \sigma_\xi^2)$ is the same as in the Poisson log-linear model, see (4.199):

$$m(x, \theta) = \exp\left\{\beta_0 + \beta_1\mu_1(x) + \frac{\beta_1^2\tau^2}{2}\right\} . \tag{4.282}$$

The conditional variance $v(x, \theta)$ can be found by equality (1.65):

$$v(x, \theta) = \mathbf{V}(y|x) = \alpha^{-1}\mathbf{E}[\mathrm{e}^{2\beta_0+2\beta_1\xi}|x] + \mathbf{E}[(m^*)^2|x] - m^2(x, \theta) , \tag{4.283}$$

$$\begin{aligned} v(x, \theta) &= (1 + \alpha^{-1})\exp\{2\beta_0 + 2\beta_1\mu_1(x) + 2\beta_1^2\tau^2\} - \exp\{2\beta_0 + 2\beta_1\mu_1(x) + \beta_1^2\tau^2\} = \\ &= \mathrm{e}^{2\beta_0+2\beta_1\mu_1(x)+\beta_1^2\tau^2}((1 + \alpha^{-1})\mathrm{e}^{\beta_1^2\tau^2} - 1) . \end{aligned} \tag{4.284}$$

As seen in Section 4.2.2, the ACM of the QL estimator for the parameter $\beta$ does not change if the dispersion parameter $\alpha^{-1}$ is assumed to be either known or unknown. Hence, we will only deal with the case of $\alpha$ known.

The estimating function for parameter $\beta$ is equal to

$$s_{\mathrm{QL}}^{(\beta)}(y, x; \beta, \mu_\xi, \sigma_\xi^2) = \frac{y - m(x, \theta)}{v(x, \theta)} \frac{\partial m(x, \theta)}{\partial \beta} . \tag{4.285}$$

Under known $\mu_\xi$ and $\sigma_\xi^2$, a formal definition of the estimator $\hat{\beta}_{\mathrm{QL}}$ is the same as in Definition 4.25, with some parameter set $\Theta \subset \mathbf{R}^2$. Theorem 4.27 on the consistency and asymptotic normality of the estimator $\hat{\beta}_{\mathrm{QL}}$ is word for word transferred to the case of gamma model. It is possible to compute the estimator $\hat{\beta}_{\mathrm{QL}}$ by Newton–Raphson method.

**Theorem 4.34.** *In the log-linear gamma model, assume the condition* (4.279). *Let the parameters $\alpha$, $\mu_\xi$, and $\sigma_\xi^2$ be known, the true value of $\beta$ is an interior point of the convex compact set $\Theta \subset \mathbf{R}^2$, moreover for all $b, \beta \in \Theta^0$, the matrix* (4.215) *be positive definite, where the functions $m(x, \beta) = m(x, \theta)$ and $v(x, \beta) = v(x, \theta)$ are given in* (4.282) *and* (4.284), *respectively. Let $\Sigma_{\mathrm{C}}$ and $\Sigma_{\mathrm{QL}}$ be the ACMs of estimators $\hat{\beta}_{\mathrm{C}}$ and $\hat{\beta}_{\mathrm{QL}}$, respectively. If the true value $\beta_1 = 0$ then $\Sigma_{\mathrm{QL}} = \Sigma_{\mathrm{C}}$, but if $\beta_1 \neq 0$ then $\Sigma_{\mathrm{QL}} < \Sigma_{\mathrm{C}}$.*

The statement follows from results of Kukush et al. (2007).

Now, let the nuisance parameters $\mu_\xi$ and $\sigma_\xi^2$ be unknown. Then the estimator $\hat{\beta}_{\mathrm{QL}}$ can be modified by formulas (4.226) and (4.227). In doing so, the ACM of the new estimator will increase.

**Theorem 4.35.** *Assume the conditions of Theorem* 4.34, *but the parameters $\mu_\xi$, $\sigma_\xi^2$ are unknown, while the parameter $\alpha$ is known. Let $\Sigma_{\mathrm{C}}$ and $\Sigma_{\mathrm{QL}}$ be the ACMs of the estimators $\hat{\beta}_{\mathrm{C}}$ and $\hat{\beta}_{\mathrm{QL}}$, respectively. If the true value $\beta_1 = 0$ then $\Sigma_{\mathrm{QL}} = \Sigma_{\mathrm{C}}$, but if $\beta_1 \neq 0$ then $\Sigma_{\mathrm{QL}} < \Sigma_{\mathrm{C}}$.*

*Proof* is given in Kukush et al. (2007).

**Remark 4.36.** For most concrete models of this section containing the classical measurement error, the pre-estimation of parameters $\mu_\xi$ and $\sigma_\xi^2$ yields the asymptotically

efficient estimator $\hat{\beta}_{\mathrm{QL}}$. Here the asymptotic efficiency is understood in the sense of Theorem 4.10. In particular, the estimator $\hat{\beta}_{\mathrm{QL}}$ is asymptotically more efficient than $\hat{\beta}_{\mathrm{C}}$ for such specific models. The exceptions are the Gaussian model with exponential regression function and the logistic model. In the latter models, it is more efficient to estimate the parameters $\beta$, $\mu_\xi$, and $\sigma_\xi^2$ simultaneously, see Kukush et al. (2007).

# Part II: **Radiation risk estimation under uncertainty in exposure doses**

# 5 Overview of risk models realized in program package EPICURE

*EPICURE* is a package of applied interactive computer programs designed on the base of original programs *AMFIT* and *PYTAB*, which were created by D. Preston and D. Pierce for the analysis of radiation effects in victims of the atomic bombing of Japanese cities Hiroshima and Nagasaki (Preston et al., 1993). The software package allows estimating parameters in generalized risk models and analyzing the data of epidemiological and experimental studies. *EPICURE* consists of four modules, each of which is designed for a particular type of data processing:

- binomial data (module *GMBO*),
- matched data for the case-control study (module *PECAN*),
- survival data (module *PEANUTS*),
- grouped data that have Poisson distribution (module *AMFIT*).

Each module of the software package includes statistical models to estimate the parameters of the generalized risk $\lambda_i(x_i, \theta)$, which is a function of the vector of covariates $x_i = \{x_{i,0}, x_{i,1}, \ldots, x_{i,4}\}$ and the parameters $\theta = \{\beta_0, \beta_1, \ldots, \beta_4\}$ of a regression model for observations with numbers $i = 1, 2, \ldots, n$. Each of $x_{i,0}, x_{i,1}, \ldots, x_{i,4}$ and $\beta_0, \beta_1, \ldots, \beta_4$ is a vector. Mathematical content of $\lambda_i(x_i, \theta)$ depends on a type of statistical regression model. The content of each module in *EPICURE* is described as follows:

- *GMBO* – the binomial odds or a function of the odds,
- *PECAN* – the odds ratio for cases and controls,
- *PEANUTS* – the relative risk or hazard ratio modifying a nonparametric underlying hazard function for censored survival data,
- *AMFIT* – the Poisson mean or a piecewise constant hazard function for grouped survival data.

Formally, being programmed in *EPICURE* a regression model can be specified as the relative risk

$$\lambda_i(x_i, \theta) = T_0(x_{i,0}, \beta_0) \cdot \left(1 + \sum_{j=1}^{4} T_j(x_{i,j}, \beta_j)\right) \tag{5.1}$$

or the absolute risk

$$\lambda_i(x_i, \theta) = T_0(x_{i,0}, \beta_0) + \sum_{j=1}^{4} T_j(x_{i,j}, \beta_j)\,. \tag{5.2}$$

Here $T_0(x_{i,0}, \beta_0)$ and $T_j(x_{i,j}, \beta_j)$ are products of linear and log-linear functions of regression parameters.

## 5.1 Risk analysis of individual data (GMBO module)

Mathematical basis of the software *GMBO* module is regression model with binary response $Y_i$ that takes two values, usually 0 and 1. The model is used typically if an epidemiologist has data of individual observations. This situation is common for *cohort studies*, the essence of which is that there is some group (cohort) of individuals exposed to radiation or other factor. Later on, some subjects of the cohort may get cancer disease, i.e., each subject of the cohort can be corresponded to a binary variable that takes the value 0 ("*i*th person is not diseased") or 1 ("*i*th person is diseased"). It is assumed that the researcher has complete information on each individual from the cohort (i.e., the researcher knows his/her age, sex, individual dose, the implementation time of disease, time elapsed since exposure, age at exposure, etc.). Having data on each subject of the cohort and using one of the models of absolute or relative risk, it is possible to write down the risk of disease as[1].

$$\lambda_i(x_i, \theta) = \frac{\mathbf{P}(Y_i = 1)}{\mathbf{P}(Y_i = 0)} = \frac{P_i(x_i, \theta)}{1 - P_i(x_i, \theta)} \,. \tag{5.3}$$

Here $P_i(x_i, \theta)$ is probability of 1, or expectation of $Y_i$:

$$P_i(x_i, \theta) = \mathbf{P}(Y_i = 1) = \mathbf{E}(Y_i) \,. \tag{5.4}$$

It is evident that

$$P_i(x_i, \theta) = \frac{\lambda_i(x_i, \theta)}{1 + \lambda_i(x_i, \theta)} \,. \tag{5.5}$$

Further, based on these probabilities, one can construct the likelihood function, whose maximum point defines the estimate of the vector $\theta$ of unknown coefficients. The likelihood function for logistic model is given as a product (see Example 4.7):

$$\prod_i P_i(x_i, \theta)^{Y_i} (1 - P_i(x_i, \theta))^{1 - Y_i} \,. \tag{5.6}$$

Respectively, the log-likelihood function is equal to

$$\begin{aligned} l(\theta) &= \sum\nolimits_i [Y_i \ln P_i(x_i, \theta) + (1 - Y_i) \ln(1 - P_i(x_i, \theta))] = \\ &= \sum\nolimits_i [Y_i \ln \lambda_i(x_i, \theta) - (1 - Y_i) \ln(1 + \lambda_i(x_i, \theta))] \,. \end{aligned} \tag{5.7}$$

The gradient of the log-likelihood function is given by the equality

$$\begin{aligned} g^{\mathrm{T}}(\theta) = \frac{\partial l}{\partial \theta} &= \sum_i \left[ \frac{Y_i}{P_i(x_i, \theta)} - \frac{1 - Y_i}{1 - P_i(x_i, \theta)} \right] \frac{\partial P_i(x_i, \theta)}{\partial \theta} = \\ &= \sum \frac{Y_i - P_i(x_i, \theta)}{P_i(x_i, \theta)(1 - P_i(x_i, \theta))} \frac{\partial P_i(x_i, \theta)}{\partial \theta} \,. \end{aligned} \tag{5.8}$$

**1** In the software *EPICURE*, it is possible also to specify the risk as $\lambda_i(x_i, \theta) = P_i(x_i, \theta)$ or $\lambda(x_i, \theta) = -\ln(1 - P_i(x_i, \theta))$.

Using (5.5), transforming (5.8):

$$g^{\mathrm{T}}(\theta) = \sum_i \left[ \frac{Y_i}{\lambda_i(x_i, \theta)} - \frac{1}{1 + \lambda_i(x_i, \theta)} \right] \frac{\partial \lambda_i(x_i, \theta)}{\partial \theta} . \tag{5.9}$$

The Hessian matrix is given as

$$\begin{aligned} H(\theta) = &\sum_i \left[ \frac{Y_i}{P_i(x_i, \theta)} - \frac{1 - Y_i}{1 - P_i(x_i, \theta)} \right] \frac{\partial^2 P_i(x_i, \theta)}{\partial \theta \partial \theta^{\mathrm{T}}} + \\ &+ \sum_i \left[ \frac{1 - Y_i}{(1 - P_i(x_i, \theta))^2} - \frac{Y_i}{P_i(x_i, \theta)^2} \right] \frac{\partial P_i(x_i, \theta)}{\partial \theta} \frac{\partial P_i(x_i, \theta)}{\partial \theta^{\mathrm{T}}} . \end{aligned} \tag{5.10}$$

Expressing $P_i(x_i, \theta)$ via the function $\lambda_i(x_i, \theta)$, we get

$$\begin{aligned} H(\theta) = &\sum_i \left[ \frac{Y_i}{\lambda_i(x_i, \theta)} - \frac{1}{1 - \lambda_i(x_i, \theta)} \right] \frac{\partial^2 \lambda_i(x_i, \theta)}{\partial \theta \partial \theta^{\mathrm{T}}} + \\ &+ \sum_i \left[ \frac{1}{(1 - \lambda_i(x_i, \theta))^2} - \frac{Y_i}{\lambda_i(x_i, \theta)^2} \right] \frac{\partial \lambda_i(x_i, \theta)}{\partial \theta} \frac{\partial \lambda_i(x_i, \theta)}{\partial \theta^{\mathrm{T}}} . \end{aligned} \tag{5.11}$$

Having expressions for the gradient and Hessian, one can maximize the function $l(\theta)$ using the Newton–Raphson numerical method or another optimization method.

## 5.2 Case-control study (PECAN module)

The *PECAN* module is designed for data processing of epidemiological case-control studies. In contrast to cohort studies, the binary outcome variable in a case-control study is fixed by stratification. The dependent variables in this setting are one or more primary covariates, exposure variables in $x$. In this type of study design, samples of fixed size are chosen from the two strata defined by the outcome variable. The values of the primary exposure variables and the relevant covariates are then measured for each subject selected. At this, the main covariate (dose) and other significant covariates (such as gender, age, etc.) are assumed known. The total likelihood function is the product of stratum-specific likelihood functions, which are dependent on the probability of getting a subject to the sample with given distribution of covariates. After some simple transformations, one can obtain the logistic (or similar to logistic) regression model, where the response variable will be in reality interesting for a researcher (Hosmer et al., 2013). A key point of the mentioned transformations is Bayes' theorem.

Let a variable $s$ mean selection ($s = 1$) or not selection ($s = 0$) of a subject. The likelihood function for a sample of $n^1$ cases (subjects with realization of the effect $y = 1$) and $n^0$ controls (subjects without realization of the effect $y = 0$) can be written as follows:

$$L = \prod_{i=1}^{n^1} P(x_i | y_i = 1, s_i = 1) \prod_{i=1}^{n^0} P(x_i | y_i = 0, s_i = 1) . \tag{5.12}$$

After applying Bayes' theorem to the individual probabilities from (5.12), we get

$$P(x|y, s = 1) = \frac{P(y|x, s = 1)P(x|s = 1)}{P(y|s = 1)} . \tag{5.13}$$

Applying Bayes' theorem to the first factor in the numerator of (5.13) for $y = 1$, we have

$$\mathbf{P}(y = 1|x, s = 1) = \frac{\mathbf{P}(y = 1|x)\mathbf{P}(s = 1|x, y = 1)}{\mathbf{P}(y = 0|x)\mathbf{P}(s = 1|x, y = 0) + \mathbf{P}(y = 1|x)\mathbf{P}(s = 1|x, y = 1)} . \tag{5.14}$$

Similarly for $y = 0$,

$$\mathbf{P}(y = 0|x, s = 1) = \frac{\mathbf{P}(y = 0|x)\mathbf{P}(s = 1|x, y = 0)}{\mathbf{P}(y = 0|x)\mathbf{P}(s = 1|x, y = 0) + \mathbf{P}(y = 1|x)\mathbf{P}(s = 1|x, y = 1)} . \tag{5.15}$$

Suppose that the selection of cases and controls is independent from covariates that influence the disease incidence, i.e., from the vector $x$. Denote the probability of selection of the case and control by $\tau_1$ and $\tau_0$, respectively, i.e.,

$$\begin{aligned} \tau_1 &= \mathbf{P}(s = 1|\, y = 1, x) = \mathbf{P}(s = 1|\, y = 1) , \\ \tau_0 &= \mathbf{P}(s = 1|\, y = 0, x) = \mathbf{P}(s = 1|\, y = 0) . \end{aligned} \tag{5.16}$$

Denote by $\eta(x)$ the conditional probability of the case:

$$\eta(x) = \mathbf{P}(y = 1|x) = \frac{\lambda(x)}{1 + \lambda(x)} , \tag{5.17}$$

where $\lambda(x)$ is the total incidence rate.

Substituting (5.16) and (5.17) to (5.14) and (5.15), we obtain

$$\begin{aligned} \mathbf{P}(y = 1|x, s = 1) &= \frac{\tau_1\eta(x)}{\tau_0(1 - \eta(x)) + \tau_1\eta(x)} , \\ \mathbf{P}(y = 0|x, s = 1) &= \frac{\tau_0(1 - \eta(x))}{\tau_0(1 - \eta(x)) + \tau_1\eta(x)} . \end{aligned} \tag{5.18}$$

Introduce a notation

$$\eta^*(x) = \frac{\tau_1\eta(x)}{\tau_0(1 - \eta(x)) + \tau_1\eta(x)} = \frac{\tau_1\lambda(x)}{\tau_0 + \tau_1\lambda(x)} = \frac{\frac{\tau_1}{\tau_0}\lambda(x)}{1 + \frac{\tau_1}{\tau_0}\lambda(x)} = \frac{\lambda^*(x)}{1 + \lambda^*(x)} , \tag{5.19}$$

with $\lambda^*(x) = \frac{\tau_1}{\tau_0}\lambda(x)$ .

Substituting (5.18) with the notation (5.19) to (5.13) and bearing in mind that the selection of cases and controls is independent of $x$, we get

$$\begin{aligned} P(x|y = 1, s = 1) &= \frac{\eta^*(x)P(x)}{\mathbf{P}(y = 1|s = 1)} , \\ P(x|y = 0, s = 1) &= \frac{(1 - \eta^*(x))P(x)}{\mathbf{P}(y = 0|s = 1)} . \end{aligned} \tag{5.20}$$

If we denote

$$L^* = \prod_{i=1}^{n^1+n^0} \eta^*(x)^{y_i}(1 - \eta^*(x))^{1-y_i} , \tag{5.21}$$

then the likelihood function (5.12) is written as

$$L = L^* \prod_{i=1}^{n^1+n^0} \frac{P(x_i)}{P(y_i|s_i = 1)} . \tag{5.22}$$

The first factor $L^*$ on the right-hand side of (5.22) is constructed in the same manner as the likelihood function for cohort studies, but by the data obtained within the case-control study. If the distribution of covariate $P(x_i)$ does not depend on the model parameters, and the selection of cases and controls is carried out randomly from the same subset, i.e., the conditions $\mathbf{P}(y = 1|s = 1) = \frac{n^1}{n^1+n^0}$ and $\mathbf{P}(y = 0|s = 1) = \frac{n^0}{n^1+n^0}$ hold true, then the likelihood function $L^*$ can be used for the risk coefficients estimation.

Let the total incidence rate take the form:

$$\lambda_i = \mathrm{e}^{\alpha_0+\alpha v_i}(1 + D_i \mathrm{e}^{\beta_0+\beta z_i}) , \tag{5.23}$$

where $\alpha_0$ is baseline risk coefficient, $D_i$ is exposure dose, $\alpha$, $\beta_0$, $\beta$ are risk coefficients, $z_i$ is risk modifier, $v_i$ is confounder, i.e., the covariate affecting the baseline risk. If $v_i$ and $z_i$ are column vectors, then $\alpha$ and $\beta$ are row vectors.

Then

$$\lambda_i^* = \mathrm{e}^{\ln \frac{\tau_1}{\tau_0}} \lambda_i = \mathrm{e}^{\ln \frac{\tau_1}{\tau_0}+\alpha_0+\alpha v_i}(1 + D_i \mathrm{e}^{\beta_0+\beta z_i}) = \mathrm{e}^{\alpha_0^*+\alpha v_i}(1 + D_i \mathrm{e}^{\beta_0+\beta z_i}) . \tag{5.24}$$

Thus, optimizing the function $L^*$, it is possible to estimate the incidence rate of spontaneous $\alpha_0^* = \ln \frac{\tau_1}{\tau_0} + \alpha_0$ instead of the true baseline risk $\alpha_0$.

### 5.2.1 Matched case-control study

An important special case of stratified case-control study is *matched study*. This kind of study is justified in Breslow and Day (1980), Schlesselman (1982), Kelsey et al. (1996), and Rothman et al. (2008). In this type of study, subjects are stratified by main covariates influencing the response. Typically, such covariates are sex and age. Each stratum is a sample consisting of cases and controls. The number of cases and controls in different strata can be different. However, in most studies of this type, from one to five controls are included in each stratum. Such research is called *1-M matched case-control study*.

Theoretically the stratum-specific covariates can be included to the regression model and their influence can be estimated. This approach works well when the number of subjects in each stratum is large. But usually in matched case-control study, a stratum contains a few subjects. For example, in the 1-1 matched design with $n$ case-control pairs we have only two subjects per stratum. Then for analysis of the model with $p$ covariates, $n + p$ parameters should be estimated (including constant term, $n - 1$ stratum-specific parameters, and $p$ risk coefficients), based on the sample of

size $2n$. It is known that under increasing sample size, the properties of the likelihood function are improved if the number of estimated parameters is fixed. In Breslow and Day (1980), it is shown that ignorance of this recommendation can lead to 100% bias of the estimates. If we consider the stratum-specific parameters as nuisance, then the conditional likelihood function can be used for estimation of the risk coefficients. In so doing, the obtained estimators will be consistent and asymptotically normal (Cox and Hinkley, 1974).

Suppose that all the data consist of $K$ strata, and in $k$th stratum, $k = 1, 2, \ldots, K$, there are $n_k^1$ cases and $n_k^0$ controls, with $n_k = n_k^1 + n_k^0$.

Assume also that the total incidence rate in $k$th stratum takes the form

$$\lambda_i = \mathrm{e}^{\alpha_k + \alpha v_i}(1 + D_i \mathrm{e}^{\beta_0 + \beta z_i}) \,. \tag{5.25}$$

Here $\alpha_k$ is stratum-specific parameter, $D_i$ is exposure dose, $\alpha$, $\beta_0$, and $\beta$ are risk coefficients, $z_i$ is risk modifier, and $v_i$ is confounder.

The conditional likelihood function for $k$th stratum is probability of the observed data, provided $n_k^1$ cases and $n_k^0$ controls got to the stratum. Equivalently, it is the ratio of probability of the observed data and probability of the observed data under all possible combinations of $n_k^1$ cases and $n_k^0$ controls. The number of combinations in $k$th stratum is given by the following formula:

$$c_k = \frac{n_k!}{n_k^1!(n_k - n_k^1)!} \,. \tag{5.26}$$

Let in any combination, subjects from 1 to $n_k^1$ correspond to cases and ones from $n_k^1 + 1$ to $n_k$ correspond to controls. Then the conditional likelihood function for $k$th stratum can be written as

$$L_k = \frac{\prod_{i=1}^{n_k^1} P(x_i|y_i = 1) \prod_{i=n_k^1+1}^{n_k} P(x_i|y_i = 0)}{\sum_{j=1}^{c_k} \left( \prod_{i=1}^{n_k^1} P(x_i|y_i = 1) \prod_{i=n_k^1+1}^{n_k} P(x_i|y_i = 0) \right)} \,. \tag{5.27}$$

The complete likelihood function is the product of all $L_k$:

$$L = \prod_{k=1}^{K} L_k \,. \tag{5.28}$$

By Bayes' theorem, we find multipliers in the right-hand side of (5.27):

$$\begin{aligned} P(x_i|y_i = 1) &= \frac{\mathbf{P}(y_i = 1|x_i)P(x_i)}{\mathbf{P}(y_i = 1)} = \frac{\lambda_i}{1 + \lambda_i} \frac{P(x_i)}{\mathbf{P}(y_i = 1)} \,, \\ P(x_i|y_i = 0) &= \frac{\mathbf{P}(y_i = 0|x_i)P(x_i)}{\mathbf{P}(y_i = 0)} = \frac{1}{1 + \lambda_i} \frac{P(x_i)}{\mathbf{P}(y_i = 0)} \,. \end{aligned} \tag{5.29}$$

Substituting (5.29) to (5.27), we get

$$L_k = \frac{\prod_{i=1}^{n_k^1} \frac{\lambda_i}{1+\lambda_i} \frac{P(x_i)}{\mathbf{P}(y_i=1)} \prod_{i=n_k^1+1}^{n_k} \frac{1}{1+\lambda_i} \frac{P(x_i)}{\mathbf{P}(y_i=0)}}{\sum_{j=1}^{c_k} \left( \prod_{i=1}^{n_k^1} \frac{\lambda_i}{1+\lambda_i} \frac{P(x_i)}{\mathbf{P}(y_i=1)} \prod_{i=n_k^1+1}^{n_k} \frac{1}{1+\lambda_i} \frac{P(x_i)}{\mathbf{P}(y_i=0)} \right)} = \frac{\prod_{i=1}^{n_k^1} \lambda_i}{\sum_{j=1}^{c_k} \left( \prod_{i=1}^{n_k^1} \lambda_i \right)} \,. \tag{5.30}$$

Substituting expression (5.25) in (5.30) instead of $\lambda_i$, we find

$$L_k = \frac{\prod_{i=1}^{n_k^1} RR_i}{\sum_{j=1}^{c_k}\left(\prod_{i=1}^{n_k^1} RR_i\right)}, \tag{5.31}$$

where $RR_i = e^{\alpha v_i}(1 + D_i e^{\beta_0+\beta z_i})$ is the relative risk.

After substitution (5.31) in (5.28), finding the logarithm of (5.31), and writing in addition index $k$ to indicate the stratum, we obtain the final expression for log-likelihood function:

$$l = \sum_{k=1}^{K}\left(\sum_{i=1}^{n_k^1} \ln RR_{i,k} - \ln \sum_{j=1}^{c_k}\left(\prod_{i=1}^{n_k^1} RR_{i,k}\right)\right). \tag{5.32}$$

For example, the log-likelihood function $l(\alpha, \beta_0, \beta)$ has the following form for 1-$M$ matched case-control study (i.e., for 1 case and $M$ controls):

$$l = \sum_{k=1}^{K}\left(\ln RR_{1,k} - \ln \sum_{i=1}^{M+1} RR_{i,k}\right), \tag{5.33}$$

where $RR_{i,k}$ is the corresponding relative risk for a subject of $k$th stratum.

Thus, a case-control study can be performed in the following two versions:

- Unmatched, or ordinary, case-control study,
- Matched case-control study.

In the first version, the controls are selected randomly from the same subset as the cases, in order to reflect with sufficient accuracy the distribution of covariates influencing the incidence. At this, the same numerical algorithms and computer procedures can be used as for ordinary cohort study. Such version of case-control study allows estimating all risk coefficients (in particular, the confounders and risk modifiers), except of the baseline risk $\alpha_0$. If obtaining estimates for nuisance parameters of the incidence rate of spontaneous (i.e., confounders) is not a principle question for a researcher, then it is possible to use the second version, namely matched case-control study. In the latter version, for each case a selection of controls is performed, with the same values of nuisance parameters, as for the case; controls are selected from the same set as the case. After that, the conditional likelihood function is constructed, which however, does not allow obtaining estimates of nuisance parameters.

## 5.3 Survival models (PEANUTS module)

Let a nonnegative random variable $T$ be the waiting time for an event to occur. For simplicity, we use the terminology from survival analysis. In so doing, the underlying event will be called *death*, and the waiting time will be named *survival time*, although

the technique being discussed below has much broader application. It can be used, for example, to analyze the morbidity, migration, life expectancy, etc. (Rodríguez, 2008).

Suppose $T$ is continuous random variable, with probability density function (pdf) $f(t)$ and cumulative distribution function (cdf) $F(t) = \mathbf{P}\{T \le t\}$, the latter is probability that the event occurred before the moment $t$. It is often convenient to use a complement to the cdf, which is called the *survival function*:

$$S(t) = \mathbf{P}\{T > t\} = 1 - F(t) = \int_t^\infty f(x)dx \,. \tag{5.34}$$

It determines probability to survive till the moment $t$, or in a broad sense, probability that the event did not occur until the moment $t$.

An alternative way to characterize the distribution of $T$ is to define the *hazard function*, or instantaneous intensity of event's realization:

$$\lambda(t) = \lim_{dt \to 0} \frac{\mathbf{P}\{t < T \le t + dt \mid T > t\}}{dt} \,. \tag{5.35}$$

The numerator of this expression is conditional probability that the event will take place in the interval $(t, t + dt)$ being not happened earlier, and the denominator is the width of the interval. Dividing one to another, we obtain the intensity of event's realization per time unit. Tending the width to zero, we get the instant intensity of the event's realization. Conditional probability in the numerator can be written as ratio of joint probability that $T$ belongs to the interval $(t, t + dt)$ and $T > t$ (of course, this coincides with probability of belonging $T$ to abovementioned interval) and probability that $T > t$. The first of these is equal to $f(t)dt$ for small $dt$, and the latter is $S(t)$ by definition. Thus,

$$\lambda(t) = \frac{f(t)}{S(t)} \,, \tag{5.36}$$

i.e., the intensity of event's realization at moment $t$ is equal to ratio of the pdf at that moment to probability of survival till that moment.

Equality (5.34) demonstrates that $f(t)$ is a derivative of the function equal to $-S(t)$. Then equality (5.36) can be written as

$$\lambda(t) = -\frac{d}{dt} \ln S(t) \,. \tag{5.37}$$

By integrating both sides of (5.37) from 0 to $t$ and entering a boundary condition $S(0) = 1$ (this holds true because the event cannot occur before moment 0), we can write down the probability to survive till the moment $t$ via the hazard function:

$$S(t) = \exp\left( - \int_0^t \lambda(x)dx \right) \,. \tag{5.38}$$

The integral in parentheses is called *cumulative hazard* and denoted as

$$\Lambda(t) = \int_0^t \lambda(x)dx \,. \tag{5.39}$$

### 5.3.1 Censoring

A peculiarity of survival analysis is censoring, i.e., the phenomenon that the investigated event has occurred for some subjects, and therefore, an exact waiting time is known, while for others, this event has not occurred and it is only known that the waiting time exceeds the observation time.

There are several types of censing. In the first type, the sample of $n$ subjects is observed for a fixed time. That is, each subject has a maximal possible pre-fixed observation period, which can vary from one subject to another, and the total number of deaths is random.

In censoring the second type, the sample of $n$ subjects is observed until the event is not realized for $d$ subjects. In this scheme, the number of deaths $d$ is fixed in advance and it can be used as a parameter, but the total duration of study is random and cannot be known in advance.

Within a more general scheme called *random censoring*, every subject has a potential censoring moment $C_i$ and potential duration of life $T_i$, which are assumed to be independent random variables. The value $Y_i = \min\{C_i, T_i\}$ is observed, as well as the censoring indicator, often denoted as $d_i$ or $\delta_i$, which points out how the observation was finalized: as a result of death or censoring. All these schemes are united by the fact that the censoring mechanism is noninformative, and all of them, in fact, yield the same likelihood function. The weakest assumption, which is required for getting the likelihood function, is that censoring does not provide any information on the prospects of survival of a subject beyond the censoring date. That is, all what is known about the observation being censored at time $t$, is that the duration of life for the subject exceeds $t$.

### 5.3.2 Likelihood function for censored data

Suppose there are $n$ subjects under observation, with duration of life characterized by survival function $S(t)$, probability density function $f(t)$, and hazard function $\lambda(t)$. Assume also that the subject $i$ is being observed until the moment $t_i$. If the subject has died at the moment $t_i$, then his/her contribution to the likelihood function is the value of pdf at this moment, which can be written as the product of survival function and hazard function: $L_i = f(t_i) = S(t_i)\lambda(t_i)$. If the subject is still alive at the moment $t_i$, all what is known under noninformative censoring is that duration of his/her life exceeds $t_i$. Probability of the latter event is equal to $L_i = S(t_i)$ and shows the contribution of censored observation to the likelihood function. That is, both contributions contain the survival function $S(t_i)$, because in both cases the subject has survived until the moment $t_i$. Death multiplies this contribution by the hazard function $\lambda(t_i)$ and censoring does not.

Let $d_i$ be the censoring indicator equal to 1 if the object has died at the moment $t_i$, and 0 otherwise. Then the likelihood function takes the form:

$$L = \prod_i L_i = \prod_i \lambda(t_i)^{d_i} S(t_i) \,. \tag{5.40}$$

Finding the logarithm of (5.40) and using (5.38) and (5.39), we obtain the log-likelihood function for censored data:

$$l = \sum_i^n (d_i \ln \lambda(t_i) - \Lambda(t_i)) \,. \tag{5.41}$$

### 5.3.3 Cox proportional hazards model

The Cox proportional hazards model was first proposed by Cox (1972). In the model, the hazard function for an individual having characteristics $x_i$ at the moment $t$ is given as

$$\lambda_i(t\,|x_i) = \lambda_0(t)\mathrm{e}^{\beta x_i} \,. \tag{5.42}$$

Here $\lambda_0(t)$ is the baseline hazard function (the incidence rate of spontaneous), and $\mathrm{e}^{\beta x_i}$ is the relative risk, i.e., proportional increase or decrease of the hazard function associated with a set of covariates $x_i$. If $x_i$ is a column vector, then $\beta$ is a row vector. Note that increase or decrease of the hazard function is the same for all moments $t$.[2]

Note that the proportional hazards model separates the effect of time from the effect of covariates explicitly. Finding the logarithm of (5.42), it is easy to see that the proportional hazards model is just a simple additive model for the logarithm of hazard function:

$$\ln \lambda_i(t\,|\,x_i) = \alpha_0(t) + \beta x_i \,, \tag{5.43}$$

where $\alpha_0(t) = \ln \lambda_0(t)$ is the logarithm of baseline hazard function. Here like in every additive model, the same effect of covariates is provided for all moments $t$.

Integrating both sides of (5.42) from 0 to $t$, one can obtain the proportional cumulative hazards:

$$\Lambda_i(t\,|x_i) = \Lambda_0(t)\mathrm{e}^{\beta x_i} \,. \tag{5.44}$$

---

**2** For instance, if the dummy covariate $x_i$ means membership of an individual to the first of null group, the hazards model takes the form

$$\lambda_i(t|x) = \begin{cases} \lambda_0(t), & \text{if } x = 0 \,, \\ \lambda_0(t)\mathrm{e}^{\beta}, & \text{if } x = 1 \,. \end{cases}$$

Here $\lambda_0(t)$ is the hazard function in the null group, and $\mathrm{e}^{\beta}$ is ratio of hazard functions in the first group and the null group. If $\beta = 0$, then the hazard functions coincide. If $\beta = \ln 2$, then the hazard function in the first group is twice larger compared with the hazard function in the null group.

Taking exponent of the equality with the opposite sign, we get the survival function:

$$S_i(t\,|x_i) = S_0(t)^{\exp(\beta x_i)}\,, \tag{5.45}$$

where $S_0(t)$ is the baseline survival function.

Thus, in the proportional hazards model the effect of covariates $x_i$ on the baseline survival function consists in raising it to power equal to the relative risk.

### 5.3.4 Partial likelihood function

To estimate the relative risk parameters $\beta$ in the proportional hazards model (5.42), Cox (1972) proposed a method of partial likelihood, which lies in maximizing the partial likelihood function.

Let $n$ be the number of observations of which $k$ cases fall to the event being as a result of the subject's death, and $n - k$ cases do as a result of censoring. Also suppose that $t_i, i = 1, \ldots, n$ is the ordered time array of observation points, i.e., $t_1 < t_2 < \ldots < t_n$. Then the partial likelihood function for observation $i$ is the probability of happening of the event at moment $t_i$, provided the number of subjects being under risk until the moment $t_i$ is known. In other words, if the event happened, then what is the probability that $i$th individual died of those who were under risk? The partial likelihood function answers this question.

Let $R(t_i)$ be a set of subjects being under risk until the moment $t_i$. Then the probability of death (i.e., realization of the event) of $i$th object is given by the following formula (Preston et al., 1993):

$$P(t_i|\,R(t_i)) = \frac{\mathrm{e}^{\beta x_i}}{\sum_{j\in R(t_i)} \mathrm{e}^{\beta x_j}}\,. \tag{5.46}$$

Taking into account the contribution of all observations and censoring, we obtain

$$L = \prod_i \left( \frac{\mathrm{e}^{\beta x_i}}{\sum_{j\in R(t_i)} \mathrm{e}^{\beta x_j}} \right)^{d_i}. \tag{5.47}$$

Finding the logarithm, we get

$$l = \sum_i^n \beta x_i + \sum_{i=1}^n \ln \left( \sum_{j\in R(t_i)} \mathrm{e}^{\beta x_j} \right). \tag{5.48}$$

## 5.4 Risk analysis of grouped data (AMFIT module)

Mathematical basis for the analysis of grouped data is the Poisson regression model. Often in the environmental and epidemiological studies, individual characteristics of

the population are unavailable. Therefore, the population under consideration is divided into groups according to some features (such as sex, age, residence place, etc.), and for each group, there are known observation time, estimates of mean group exposure doses (or other factors), number of cases, and perhaps some other characteristics (e.g., level of examination, territorial specificity of location). The incidence in each group is given in the form of absolute or relative risk. The response variable reflects the number of realizations of the diseases in $i$th group and can take values $0, 1, 2, \ldots, n$. It is assumed that it has Poisson distribution:

$$\mathbf{P}(Y_i = k) = \frac{\mu_i^k}{k!} \mathrm{e}^{-\mu_i} , \tag{5.49}$$

where $\mu_i$ is the distribution parameter (see Example 4.4).

Such a distribution has a random variable $Y$ being equal to the number of events that occurred during certain period of time, if the events are independent and happen at a constant speed (i.e., uniformly in time). This could be, for example, a number of radioactive decays having occurred during a second or a number of persons who developed cancer in a year.

In Holford (1980) and Oliver and Laird (1981), it is shown that the Poisson regression model is equivalent to the proportional hazards model, with piecewise constant baseline hazard function $\lambda_0(t)$.

Both expectation and variance of the Poisson distribution coincide with the parameter $\mu_i$:

$$\mathbf{E}(Y_i) = \sigma^2(Y_i) = \mu_i . \tag{5.50}$$

The latter depends on a set of covariates and on unknown parameters:

$$\mu_i = n_i \lambda_i(x_i, \theta) . \tag{5.51}$$

Here, $\lambda_i(x_i, \theta)$ is the hazard function given in (5.1) or (5.2), and $n_i$ is the number of person-years of observations in $i$th group.

The log-likelihood function for the Poisson regression is

$$l(\theta) = \sum\nolimits_i [Y_i \ln n_i \lambda_i(x_i, \theta) - n_i \lambda_i(x_i, \theta) - \ln Y_i!] . \tag{5.52}$$

The gradient and Hessian are as follows:

$$g^{\mathrm{T}}(\theta) = \frac{\partial l}{\partial \theta} = \sum_i \left[ \frac{Y_i}{\lambda_i(x_i, \theta)} - n_i \right] \frac{\partial \lambda_i(x_i, \theta)}{\partial \theta} , \tag{5.53}$$

$$\begin{aligned} H(\theta) = {} & \sum\nolimits_i \left[ \frac{Y_i}{\lambda_i(x_i, \theta)} - n_i \right] \frac{\partial^2 \lambda_i(x_i, \theta)}{\partial \theta \partial \theta^{\mathrm{T}}} - \\ & - \sum\nolimits_i \frac{1}{\lambda_i(x_i, \theta)^2} \frac{\partial \lambda_i(x_i, \theta)}{\partial \theta} \frac{\partial \lambda_i(x_i, \theta)}{\partial \theta^{\mathrm{T}}} . \end{aligned} \tag{5.54}$$

Optimizing the log-likelihood function, one can estimate the unknown regression coefficients in (5.1) or (5.2).

# 6 Estimation of radiation risk under classical or Berkson multiplicative error in exposure doses

As well known, today the most common methods for estimation of radiation risks that associated with human exposure (Preston et al., 1993; Ron et al., 1995; Jacob et al., 2006; Likhtarov et al., 2006a; Tronko et al., 2006; Zablotska et al., 2011) use a number of principle approximations. In particular, the assumption of no uncertainty in individual dose, i.e., it is assumed that we have the determined value for the exposure dose of a subject. It is clear that such a statement is fundamentally wrong, since there are practically no situations in which being estimated by any method dose would not have some statistical distribution (Likhtarov et al., 2012, 2013a, 2014). One of the consequences from the assumption of the absence of errors in exposure doses is the bias of risk estimates and distortion of the shape of curve "dose–effect" (Carroll et al., 1995, 2006; Kukush et al., 2011; Masiuk et al., 2013, 2016). Note that such distortions of the risk estimates can be caused not only by systematic errors in dose estimates which is obvious, but by random errors as well. And although recently repeated attempts to include dose errors in the risk analysis have been made (Mallick et al., 2002; Kopecky et al., 2006; Lyon et al., 2006; Carroll et al., 2006; Li et al., 2007; Masiuk et al., 2008, 2011, and 2013; Little et al., 2014), the problem is not fully resolved until the present time.

It is known that exposure doses estimation is inevitably accompanied by either the classical or Berkson type errors, or a combination of them (Mallick et al., 2002; Kopecky et al., 2006; Lyon et al., 2006; Li et al., 2007; Likhtarov et al., 2014, 2015; Masiuk et al., 2016). However, at the moment there is no final conclusion on the impact of the classical, Berkson, or mixed error in dose estimates to the final result of risk analysis, usually being expressed in values of either excess relative risk (*ERR*) or excess absolute risk (*EAR*); see *Health risks from exposure to low levels of ionizing radiation* (2006).

One of the bright examples of importance of this problem is risk analysis of the results of long-term radio-epidemiology cohort studies of children with exposed thyroid due to the accident at Chornobyl nuclear power plant (Tronko et al., 2006; Bogdanova et al., 2015). It is vital to note that in the studies, absolute and relative frequencies of thyroid cancer in this cohort have been identified fairly. Not only point but also interval (in a statistical sense) doses estimates have been obtained (Likhtarov et al., 2005, 2006b, 2012, 2013a, 2014). However, due to the lack of a more or less acceptable mathematically grounded computational procedure for combining two-dimensional error in dose and effect within a single procedure of risk analysis, the risks estimation of radiation-induced effects was performed by the popular in radio-epidemiology computer package *EPICURE* (Preston et al., 1993). The latter operates with deterministic dose values and is not adapted to take into account any uncertainty of input data.

Chapters 6 and 7 considers method taking into account both classical and Berkson errors in radiation doses for risk estimation in regression with binary response. The quality of the estimates is verified by means of stochastic simulation experiment for linear two-parameter risk model.

## 6.1 General principles for construction of radiation risk models

As known (*Health risks from exposure to low levels of ionizing radiation*, 2006), at relatively low doses the risk of radiation-induced effect either depends on the dose linearly or contains both linear and quadratic term in dose. Radiobiological theory indicates that at low doses, the risk of a biological lesion being formed should depend linearly on dose if a single event is required or on the square of dose if two events are required. It is commonly held that high linear energy transfer radiation can cause lesions by the transversal of a single particle, and low linear energy transfer radiation does it by either one or two photons (or energetic beta particles). At higher doses of radiation, cell sterilization and cell death compete with the process of malignant transformation, thereby attenuating the risk of cancer at higher doses. The probability of cell death is subject to ordinary survival laws, i.e., it has a negative exponential dependence on the dose (or on squared dose). Combining these principles, one can get a general model for dependence of the radiation risk on the dose $D$ that is widely used in radio-epidemiology for low linear energy transfer radiation:

$$f(D) = (\alpha_0 + \alpha_1 D + \alpha_2 D^2)\mathrm{e}^{-\beta_1 D - \beta_2 D^2} \ . \tag{6.1}$$

Here, $\alpha_0$, $\alpha_1$, $\alpha_2$, $\beta_1$, and $\beta_2$ are model parameters to be estimated from the data.

The models for dependence on dose are generally incorporated into risk models by assuming that the excess risk functions are proportional to $f(D)$, where the multiplicative constant (in dose) depends on such risk modifiers of radio-induction as sex and age at the moment of exposure. Moreover, for most malignant tumors (other than leukemia and bone cancer) the risk of disease increases over time of surveillance. Therefore, as a rule, most of risk estimates are based on the assumption that the risk increases during the life span of the population.

The most radio-epidemiologic studies (Likhtarov et al., 2006a; Tronko et al., 2006; Zablotska et al., 2011; Little et al., 2014) use the following linear models for risk estimation:

– relative risk model:

$$\lambda(D, s_1, \ldots, s_p, z_1, \ldots, z_q) = \mathrm{e}^{\sum_i \alpha_i s_i}\left(1 + ERR\, D\, e^{\sum_j \gamma_j z_j}\right) , \tag{6.2}$$

– absolute risk model:

$$\lambda(D, s_1, \ldots, s_p, z_1, \ldots, z_q) = \mathrm{e}^{\sum_i \alpha_i s_i} + EAR\, D\, e^{\sum_j \gamma_j z_j} \ . \tag{6.3}$$

Here, $\alpha_i$, $ERR$, $EAR$ and $\gamma_j$ are regression coefficients to be estimated, $D$ is radiation covariate (dose), $s_i$ are covariates (confounders) that affect the level of background incidence rate (as a confounder there may be the age, sex, level of examination, etc.), $z_j$ are modifying covariates making an effect on the risk of radio-induction (e.g., the age at the moment of exposure, sex, or the time elapsed since the moment of exposure). $ERR$ and $EAR$ are treated as excess relative and absolute risk per Gray, respectively.

## 6.2 Linear two-parameter model

Consider the two-parameter linear in dose regression model with binary response:

$$\begin{aligned} \mathbf{P}(Y_i = 1|D_i) &= \frac{\lambda_i}{1+\lambda_i}\,, \\ \mathbf{P}(Y_i = 0|D_i) &= \frac{1}{1+\lambda_i}\,, \end{aligned} \tag{6.4}$$

where $\lambda_i$ is the total risk or total incidence rate,

$$\lambda_i = \lambda_0 + EAR \cdot D_i\,, \tag{6.5}$$

or as a version with relative risk:

$$\lambda_i = \lambda_0(1+\beta D_i) = \lambda_0(1+ERR \cdot D_i)\,. \tag{6.6}$$

Here $D_i$ is the individual exposure dose, $\lambda_0$ is the background incidence rate (i.e., in the absence of the dose factor), $\beta = ERR$ is excess relative risk, $EAR = \lambda_0\beta = \lambda_0 \cdot ERR$ is excess absolute risk.

In this instance, $\lambda_0$ and $EAR$ (or $ERR$) are positive model parameters to be estimated. The observed sample consists of couples $(Y_i, D_i)$, $i = 1, \ldots, N$, where $D_i$ are the doses (nonnegative numbers); $Y_i = 1$ in the case of disease within some time interval, and $Y_i = 0$ in the absence of disease within the interval.

Model (6.4) resembles the logistic model (4.231) and (4.232), but the latter has total incidence rate being exponentially (not linearly) dependent on the dose: $\lambda_i = \exp(\mu_0 + \mu_1 D_i)$.

### 6.2.1 Efficient estimators of parameters in linear model

From relations (6.4)–(6.6) it follows that

$$\begin{aligned} \mathbf{E}((1-Y)(1+ERR \cdot D)) &= \mathbf{E}\left(\frac{Y}{\lambda_0}\right)\,, \\ \mathbf{E}(1-Y) &= \mathbf{E}\left(\frac{Y}{\lambda_0(1+ERR \cdot D_)}\right)\,. \end{aligned} \tag{6.7}$$

Replacing the expectations in (6.7) by the empirical means, we get the unbiased equation (see Appendix A1) for estimation of regression parameters $\lambda_0$ and $ERR$:

$$\begin{aligned} \sum_{i=1}^{N} \left( (1 - Y_i)(1 + ERR \cdot D_i) - \frac{Y_i}{\lambda_0} \right) = 0 \,, \\ \sum_{i=1}^{N} \left( 1 - Y_i - \frac{Y_i}{\lambda_0 (1 + ERR \cdot D_i)} \right) = 0 \,. \end{aligned} \tag{6.8}$$

From the first equation (6.8) we have

$$\hat{\lambda}_0 = \hat{\lambda}_0(ERR) = \frac{\sum_{i=1}^{N} Y_i}{\sum_{i=1}^{N} (1 - Y_i)(1 + ERR \cdot D_i)} \,. \tag{6.9}$$

Excluding $\lambda_0$ from system (6.8), we obtain a relation for the estimator of the parameter $ERR$:

$$\sum_{i=1}^{N} \frac{Y_i (D_i - D_{\text{av}})}{1 + ERR \cdot D_i} = 0 \,, \quad \text{ERR} > 0 \,. \tag{6.10}$$

Here $D_{\text{av}}$ is mean exposure dose for healthy subjects from the sample:

$$D_{\text{av}} := \frac{\sum_{i=1}^{N} D_i (1 - Y_i)}{\sum_{i=1}^{N} (1 - Y_i)} \,. \tag{6.11}$$

In more detail, the properties of estimators (6.9) and (6.10) are studied in Appendix B.

Compared to the maximum likelihood estimates (MLEs), the estimates (6.9) and (6.10) are efficiently computed because their evaluation is reduced to a nonlinear equation with one variable. As shown in the simulation study, the estimates (6.9) and (6.10) possess good asymptotic properties as the sample size increases. (This follows as well from a general theory of estimating equations (see Appendices A1 and A2).) Therefore, they can be used as initial approximation in computation of the MLE.

## 6.3 Two types of dose errors

Analyzing the impact of dose errors on the estimation of radiation risks, it is important to establish the essence mechanism of such errors. There are two basic models of the errors: the classical and Berkson ones (see Section 1.2). In practice, these two types of errors are usually realized jointly (Mallick et al., 2002; Masiuk et al., 2016). However, their impact on the radiation risk estimates cardinally differs. Therefore, it is advisable to consider the classical and Berkson errors separately from each other.

As a rule, the errors in doses are multiplicative in their nature. Hence, we will consider both classical and Berkson errors regarding the logarithms of exposure doses.

Let $D^{\text{tr}}$ be the true value of exposure dose (usually $D^{\text{tr}}$ is unknown), and $D^{\text{mes}}$ be its measured value.

### 6.3.1 Berkson multiplicative error

If we have a multiplicative log-normal Berkson dose error, then

$$\begin{aligned} x_i &= w_i + u_i \,, \\ u_i &\sim \mathrm{N}(0, \sigma_i^2) \,. \end{aligned} \tag{6.12}$$

Here, $w_i = \ln(D_i^{\mathrm{mes}})$ is the logarithm of measured doses (known quantity), $x_i = \ln(D_i^{\mathrm{tr}})$ is the logarithm of true dose (unknown quantity), and $u_i$ is normal error with zero expectation and known variance $\sigma_i^2$.

The variable $w_i$ can be either random or deterministic. If $w_i$ is random, then both $w_i$ and $u_i$ are assumed stochastically independent.

In case of Berkson error, the conditional distribution of the logarithm $x_i$ of true dose is given as

$$x_i | w_i \sim \mathrm{N}(w_i, \sigma_i^2) \,, \tag{6.13}$$

i.e., for each observation there is known the conditional distribution of the true random variable (i.e., of the logarithm of dose), but its exact realization is unknown (see Figure 6.1 (a)). The Berkson error occurs every time when the dose mean value is applied instead of the dose true value. In particular, if the individual dose values are unknown, but approximated values of their expectations are known (e.g., mean group estimates of individual doses (Likhtarov et al., 2005, 2014) obtained by numerical Monte Carlo procedure), then replacement of the true doses by their approximate expectations leads to measurement errors of Berkson type.

The Berkson error (unlike the classical error) possesses a convenient property that the use of doses with moderate Berkson errors in the linear (w.r.t. the dose) risk model practically does not bias the risk estimates (Carroll et al., 2006; Kukush et al., 2011; Masiuk et al. 2013, 2016).

### 6.3.2 The classical multiplicative error

Let the dose be observed with the classical log-normally distributed multiplicative error. Then

$$\begin{aligned} w_i &= x_i + u_i \,, \\ u_i &\sim \mathrm{N}(0, \sigma_i^2) \,. \end{aligned} \tag{6.14}$$

Here, $x_i = \ln(D_i^{\mathrm{tr}})$ is unknown value of the logarithm of true dose, $w_i = \ln(D_i^{\mathrm{mes}})$ is known logarithm of the measured dose; $u_i$ is independent of $x_i$ normal random variable, with zero expectation and known variance (in other words $u_i$ is a random error).

In case of the classical error,

$$w_i | x_i \sim \mathrm{N}(x_i, \sigma_i^2) \,. \tag{6.15}$$

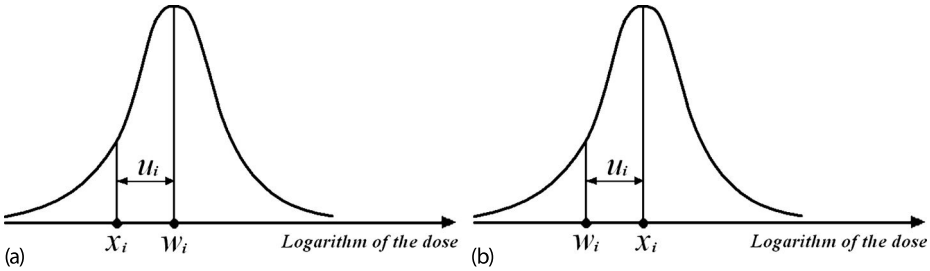


**Fig. 6.1:** Illustration to the concepts of Berkson (a) and classical (b) multiplicative errors.

From (6.15), it follows that the conditional distribution of the logarithm of measured dose $w_i$ has known variance and unknown expectation, which is regarded as the logarithm of true dose (see Figure 6.1 (b)).

The classical error occurs when the measured dose value (i.e., some computational and instrumental dose realization including the error) is used instead of the true dose value. For example, measurements of radioactivity and the results of questionnaires fluctuate just because of the presence of classical error.

### 6.3.3 Comparison of the classical and Berkson errors

At first glance, it seems that the classical error stems out from the Berkson one by simply transferring the value $u_i$ to the right-hand side of (6.12):

$$\begin{aligned} w_i &= x_i - u_i \,, \\ u_i &\sim \mathrm{N}(0, \sigma_i^2) \,. \end{aligned} \tag{6.16}$$

In view of the fact that $u_i$ and $-u_i$ are equally distributed, it seems that (6.16) is equivalent to (6.14). However, there is a significant difference: in (6.14), $x_i$ and $u_i$ are independent random variables, while in (6.16), $w_i$ and $u_i$ are independent. Therefore, the two error models (6.14) and (6.16) are quite different.

Thus, in the presence of Berkson error the logarithm $w_i$ of measured dose is independent covariate, which characterizes the distribution of the logarithm of true dose, or more precisely, $w_i$ is its conditional expectation given the measured dose. On the other hand, in the presence of the classical error the logarithm $w_i$ of measured dose is a random variable correlated with the logarithm $x_i$ of true dose and the error $u_i$.

Berkson error compared with the classical error gives more information about the true dose. With Berkson error, at least we know the conditional distribution of the true dose (or of its logarithm $x_i$). And with the classical error, we know only the logarithm of true dose up to the random error $u_i$.

## 6.4 Methods of risk estimation in the models with Berkson multiplicative error in exposure dose

### 6.4.1 Full maximum likelihood method

Consider the full maximum likelihood (FML) method for the risk estimation in the presence of multiplicative error of Berkson type in exposure doses. Assume that in measured doses, only Berkson multiplicative error is present. Then

$$\begin{aligned} &\ln D_i^{\mathrm{tr}} = \ln D_i^{\mathrm{mes}} + u_i\,, \quad i = 1, \ldots, N\,, \\ &u_i \sim \mathrm{N}(0, \sigma_{F,i}^2)\,. \end{aligned} \tag{6.17}$$

Here, $D_i^{\mathrm{tr}}$ is the true dose (unknown), $D_i^{\mathrm{mes}}$ is the estimated dose (known), and $u_i$ is the normal error.

Find the likelihood function $L(Y_i, D_i^{\mathrm{mes}}, \theta)$ using the following steps. Suppose $P(Y_i, D_i^{\mathrm{mes}}, D_i^{\mathrm{tr}})$ is joint probability distribution of random variables $Y_i$, $D_i^{\mathrm{mes}}$, and $D_i$. Then by the formula of conditional probabilities we obtain

$$P(Y_i, D_i^{\mathrm{mes}}, D_i^{\mathrm{tr}}) = P(Y_i \mid D_i^{\mathrm{mes}}, D_i^{\mathrm{tr}}) \cdot P(D_i^{\mathrm{mes}}, D_i^{\mathrm{tr}}) = P(Y_i \mid D_i^{\mathrm{tr}}) \cdot P(D_i^{\mathrm{tr}} \mid D_i^{\mathrm{mes}}) \cdot P(D_i^{\mathrm{mes}})\,. \tag{6.18}$$

Since the multiplier $P(D_i^{\mathrm{mes}})$ does not bear any additional information on the vector parameter $\theta$, it can be omitted. In (6.18), passing from the probability distributions to the likelihood function and using the fact that the conditional distribution $D_i^{\mathrm{tr}} | D_i^{\mathrm{mes}}$ is log-normal: $D_i^{\mathrm{tr}} | D_i^{\mathrm{mes}} \sim \mathrm{LN}(\ln D_i^{\mathrm{mes}}, \sigma_{F,i}^2)$, we obtain

$$L(Y_i, D_i^{\mathrm{mes}}, D_i^{\mathrm{tr}}, \theta) = \left(\frac{\lambda_i(D_i^{\mathrm{tr}}, \theta)}{1 + \lambda_i(D_i^{\mathrm{tr}}, \theta)}\right)^{Y_i} \left(\frac{1}{1 + \lambda_i(D_i^{\mathrm{tr}}, \theta)}\right)^{1-Y_i} \frac{\exp\left(-\frac{(\ln D_i^{\mathrm{tr}} - \ln D_i^{\mathrm{mes}})^2}{2\sigma_{F,i}^2}\right)}{D_i^{\mathrm{tr}}\sqrt{2\pi}\sigma_{F,i}}\,. \tag{6.19}$$

To get rid of unknown variable $D_i^{\mathrm{tr}}$, it is necessary to do convolution with respect to this variable:

$$L(Y_i, D_i^{\mathrm{mes}}, \theta) = \int_0^\infty \left(\frac{\lambda_i(t, \theta)}{1 + \lambda_i(t, \theta)}\right)^{Y_i} \left(\frac{1}{1 + \lambda_i(t, \theta)}\right)^{1-Y_i} \frac{\exp\left(-\frac{(\ln t - \ln D_i^{\mathrm{mes}})^2}{2\sigma_{F,i}^2}\right)}{t\sqrt{2\pi}\sigma_{F,i}} dt\,. \tag{6.20}$$

Thus, the likelihood function, which takes into account the contributions of all observations, is equal to

$$L(\theta) = \prod_{i=1}^N \int_0^\infty \left(\frac{\lambda_i(t, \theta)}{1 + \lambda_i(t, \theta)}\right)^{Y_i} \left(\frac{1}{1 + \lambda_i(t, \theta)}\right)^{1-Y_i} \frac{\exp\left(-\frac{(\ln t - \ln D_i^{\mathrm{mes}})^2}{2\sigma_{F,i}^2}\right)}{t\sqrt{2\pi}\sigma_{F,i}} dt. \tag{6.21}$$

The corresponding log-likelihood function is

$$l(\theta) = \sum_{i=1}^N \ln \int_0^\infty \left(\frac{\lambda_i(t, \theta)}{1 + \lambda_i(t, \theta)}\right)^{Y_i} \left(\frac{1}{1 + \lambda_i(t, \theta)}\right)^{1-Y_i} \frac{\exp\left(-\frac{(\ln t - \ln D_i^{\mathrm{mes}})^2}{2\sigma_{F,i}^2}\right)}{t\sqrt{2\pi}\sigma_{F,i}} dt\,. \tag{6.22}$$

In view of the fact that the variable $Y_i$ takes only two values 0 and 1, the latter equality can be written as:

$$l(\theta) = \sum_{i=1}^{N} \left[ Y_i \ln \int_0^{\infty} \left( \frac{\lambda_i(t,\theta)}{1+\lambda_i(t,\theta)} \right) \frac{\exp\left( -\frac{(\ln t - \ln D_i^{\text{mes}})^2}{2\sigma_{F,i}^2} \right)}{t\sqrt{2\pi}\sigma_{F,i}} dt + \right.$$
$$\left. + (1 - Y_i) \ln \int_0^{\infty} \left( \frac{1}{1+\lambda_i(t,\theta)} \right) \frac{\exp\left( -\frac{(\ln t - \ln D_i^{\text{mes}})^2}{2\sigma_{F,i}^2} \right)}{t\sqrt{2\pi}\sigma_{F,i}} dt \right] . \tag{6.23}$$

For each $i = 1, \ldots, N$, change the variable in integrals (6.23):

$$dz = \frac{1}{t\sqrt{2\pi}\sigma_{F,i}} \exp\left( -\frac{(\ln t - \ln D_i^{\text{mes}})^2}{2\sigma_{F,i}^2} \right) dt,$$
$$z = \int_0^t \frac{1}{t\sqrt{2\pi}\sigma_{F,i}} \exp\left( -\frac{(\ln t - \ln D_i^{\text{mes}})^2}{2\sigma_{F,i}^2} \right) dt = G_i(t) . \tag{6.24}$$

Here $t = G_i^{-1}(z)$, and $z = G_i(t)$ is the cumulative distribution function (cdf) of lognormal law. Hence,

$$z(0) = 0 , \quad z(+\infty) = 1 . \tag{6.25}$$

Substituting (6.24) and (6.25) to (6.23), we obtain

$$l(\theta) = \sum_{i=1}^{N} \left[ Y_i \ln \int_0^1 \left( \frac{\lambda_i(G_i^{-1}(z),\theta)}{1+\lambda_i(G_i^{-1}(z),\theta)} \right) dz + (1 - Y_i) \ln \int_0^1 \left( \frac{1}{1+\lambda_i(G_i^{-1}(z),\theta)} \right) dz \right] . \tag{6.26}$$

The integrals in (6.26) have a singularity at the point $z = 1$, since $G_i^{-1}(z) \to +\infty$, as $z \to 1$. Nevertheless, the integrals exist as the absolutely convergent improper Riemann integrals. One can use the Monte Carlo method for their evaluation. In so doing, it is possible to use the following relation:

$$G_i^{-1}(z) = \exp(D_i^{\text{mes}} + \sigma_{F,i}\Phi^{-1}(z)) , \tag{6.27}$$

where $\Phi(z)$ is the cdf of standard normal law.

The *FML* method lies in finding a vector parameter $\theta$ at which the likelihood function (6.26) attains its maximum.

### 6.4.2 Simulated stochastic experiment

To check the efficiency of the proposed method for estimation the regression parameters, the simulated stochastic experiment was done. The simulation was performed based on epidemiological studies of thyroid cancer incidence in Ukraine (Likhtarov

et al., 2006a; Tronko et al., 2006; Bogdanova et al., 2015). The absorbed doses of internal thyroid exposure correspond to those published in Likhtarov et al. (2006b) and Likhtarov et al. (2014) doses for a real subpopulation of children and adolescents aged from 0 to 18 years (13,204 persons in total) resided in settlements of Zhytomyr, Kyiv, and Chernihiv Oblasts of Ukraine, where direct measurements of thyroid radioactivity were conducted in May–June, 1986.

In simulation of the thyroid cancer incidence rate at fixed time interval, the two-parameter logistic linear model of absolute risk (6.4)–(6.5) was used.[1] The true model parameters were chosen being close to the estimates obtained during epidemiological studies of thyroid cancer in Ukraine (Likhtarov et al., 2006a; Tronko et al., 2006), namely:

$$\begin{aligned} \lambda_0 &= 2 \times 10^{-4} \, \frac{\text{cases}}{\text{person years}}, \\ EAR &= 5 \times 10^{-4} \, \frac{\text{cases}}{\text{Gy} \cdot (\text{person years})} \, . \end{aligned} \tag{6.28}$$

In addition, it was assumed that Berkson multiplicative error in dose is distributed by lognormal law (6.17). The Berkson error value was set so that the geometric standard deviation GSD $= \exp(\sigma_{F,i})$ took values from 2 to 10, for all $i = 1, \ldots, N$. In the simulation, 1000 data sets were generated for each error value.

To estimate the regression parameters $\lambda_0$ and *EAR*, the naive estimation method (i.e., the one that ignores the presence of errors in doses) and the FML method were used.

Simulation results are presented in Figure 6.2. From it, one can see that at high errors in doses, the naive estimates of background incidence rate $\lambda_0$ and excess absolute risk *EAR* deviate significantly from the corresponding model values (i.e., true values).

The deviations depend on the error variance. Thus, if GSD < 3, the naive estimate of *EAR* is close to the model value, but further increase of the GSD yields practically linear decrease of the naive estimate. At GSD = 10, the naive estimate is less than the true value of *EAR* almost twice (see Figure 6.2 (b)). As mentioned above, this phenomenon is called *attenuation effect*, i.e., the effect of underestimation of the excess absolute risk in the presence of measurement errors in doses. At the same time, the opposite effect is observed for the background incidence rate. With significant errors in absorbed doses of internal thyroid exposure, the naive estimates of $\lambda_0$ are somewhat larger than the model (i.e., true) values of this parameter (see Figure 6.2 (a)).

**1** In the simulations, the authors used the risk model (6.5) in terms of background incidence rate and *EAR*. In the context of measurement errors, this model is more natural compared to the model of relative risk (6.6). In addition, the *EAR* characterizes the slope of the dose–effect curve, and for the naive estimate, there is well-known *attenuation effect* (Carroll et al., 2006; Kukush et al., 2011; Masiuk et al., 2016), i.e., effect of attraction of the estimate for *EAR* to zero; see also Section 2.1. However, it is quite possible to compute the estimates of *ERR* and construct the appropriate confidence intervals. A rough estimate of *ERR* could be the ratio of the estimates of *EAR* and $\lambda_0$.

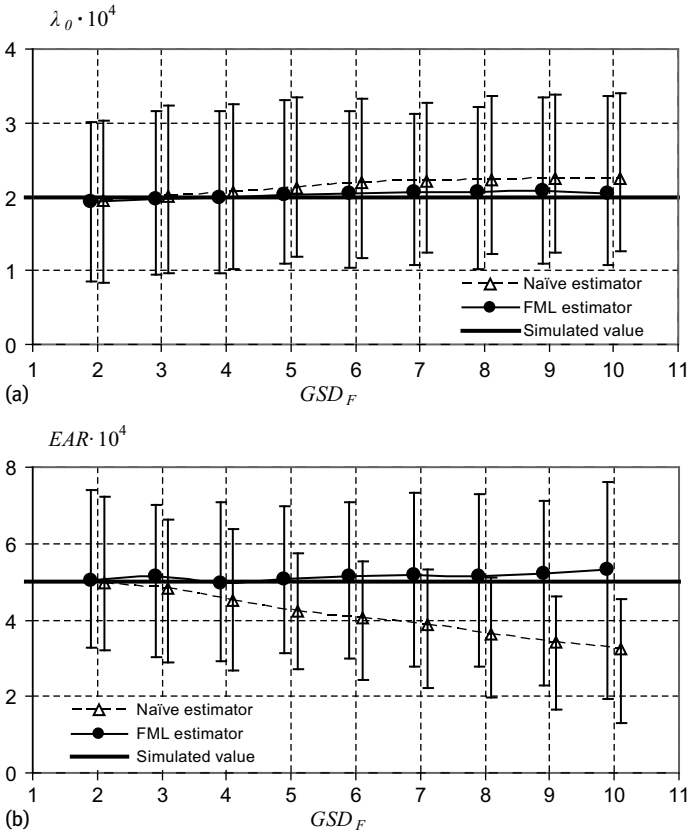


**Fig. 6.2:** Dependence of the estimates of background incidence rate (a) and excess absolute risk (b) on the level of Berkson multiplicative error in the thyroid absorbed doses.

Thus, in the presence of Berkson errors in doses, the naive estimates redistribute as follows: for *EAR* naive estimate is lower than the true value, whereas for $\lambda_0$ it is higher. This can be related to the nonlinearity of the log-likelihood function. It should be noted that in the case of Berkson error, the *attenuation effect* occurs only when the error is large, and it is not as significant as in the case of the classical error (Carroll et al., 2006; Kukush et al., 2011; Masiuk et al. 2013, 2016).

Even with significant errors in doses, the estimates of both regression parameters $\lambda_0$ and *EAR* are significantly improved when we use the FML method, which takes into account the Berkson error with help of the convolution integral (6.20).

Thus, stochastic simulation demonstrates that ignoring significant Berkson errors in doses causes the bias in the estimates of background incidence rate $\lambda_0$ and excess absolute risk *EAR*. At the same time, the biases are much smaller than in the case of the classical error. Using the FML method improves significantly both estimates.

## 6.5 Methods of risk estimation in the models with classical multiplicative error in exposure doses

In practice, to account for errors in measured dose[2], the Regression Calibration (Carroll et al., 2006; Kukush et al., 2011; Masiuk et al., 2016) is often used. Within this estimation method, before statistical processing of epidemiological data the measured dose is replaced by $D_i^* = \mathbf{E}(D_i^{\text{tr}} | D_i^{\text{mes}})$, the conditional expectation of the true dose given the measured dose; then classical regression analysis of epidemiological data is applied (e.g., using the software package *EPICURE*).

### 6.5.1 Parametric regression calibration

Let the doses be observed only with the classical multiplicative error, i.e.,

$$\begin{aligned} &\ln D_i^{\text{mes}} = \ln D_i^{\text{tr}} + u_i\,, \quad i = 1, \dots, N, \\ &u_i \sim \text{N}(0, \sigma_{Q,i}^2)\,. \end{aligned} \tag{6.29}$$

Within the functional approach, the true value of the covariate $D_i^{\text{tr}}$ is nonrandom, and then we get the classical functional errors-in-variables model. If $D_i^{\text{tr}}$ were assumed to be identically distributed random variables being independent of the errors $u_i$, we would obtain the so-called classical structural model. Within the lognormal structural model,[3]

$$\ln D_i^{\text{tr}} \sim \text{N}(\mu_1, \sigma_1^2)\,. \tag{6.30}$$

Since in the case of classical multiplicative error, the conditional distribution $\ln D_i^{\text{tr}} | D_i^{\text{mes}}$ is normal, then

$$\ln D_i^{\text{tr}} \Big| D_i^{\text{mes}} \sim \text{N}\left( \frac{\sigma_1^2 \cdot \ln D_i^{\text{mes}} + \mu_1 \cdot \sigma_{Q,i}^2}{\sigma_{Q,i}^2 + \sigma_1^2}, \frac{\sigma_{Q,i}^2 \cdot \sigma_1^2}{\sigma_{Q,i}^2 + \sigma_1^2} \right). \tag{6.31}$$

The moments of distribution (6.31) are found by the formulas for the moments of lognormal distribution (Koroliuk et al., 1985). If $\ln \eta \sim \text{N}(\mu, \sigma^2)$, then $\mathbf{E}\eta = \exp(\mu + 0,5\sigma^2)$ and $\mathbf{E}\eta^2 = \exp(2\mu + 2\sigma^2)$. Thus,

$$\mathbf{E}(D_i^{\text{tr}} \Big| D_i^{\text{mes}}) = \exp\left( \frac{\sigma_1^2 \cdot \ln D_i^{\text{mes}} + \mu_1 \cdot \sigma_{Q,i}^2 + 0.5\sigma_{Q,i}^2 \cdot \sigma_1^2}{\sigma_{Q,i}^2 + \sigma_1^2} \right), \tag{6.32}$$

$$\mathbf{E}[(D_i^{\text{tr}})^2 \Big| D_i^{\text{mes}}] = \exp\left( 2 \cdot \frac{\sigma_1^2 \cdot \ln D_i^{\text{mes}} + \mu_1 \cdot \sigma_{Q,i}^2 + \sigma_{Q,i}^2 \cdot \sigma_1^2}{\sigma_{Q,i}^2 + \sigma_1^2} \right). \tag{6.33}$$

---

**2** The term *measured dose* means the dose obtained by direct measurement of the thyroid radioactivity and using the ecological and dosimetric model of radioactivity transfer through the food chains.

**3** The parameters $\mu_1, \sigma_1^2$ can be estimated by observations $D_i^{\text{mes}}$: $\hat{\mu}_1 = \frac{1}{N}\sum_{i=1}^N \ln(D_i^{\text{mes}})$, $\hat{\sigma}_1^2 = \frac{1}{N-1}\sum_{i=1}^N (\ln(D_i^{\text{mes}}) - \hat{\mu}_1)^2 - \frac{1}{N}\sum_{i=1}^N \sigma_{Q,i}^2$, where $N$ is number of subjects in the cohort.

### 6.5.2 Nonparametric regression calibration

Let a discrete approximation to distribution of the set of variables $D_i^{\mathrm{tr}}$, $i = 1, 2, \ldots, N$, be searched. Assume that according to the estimation results[4]

$$\mathbf{P}(D_i^{\mathrm{tr}} = \exp(x_k)) = p_k\,, \quad \sum_{k=1}^{K} p_k = 1\,, \tag{6.34}$$

where $K$ is number of points at which the distribution of $D_i^{\mathrm{tr}}$ is concentrated.

Supposing also that (6.34) is the true distribution of $D_i^{\mathrm{tr}}$, we obtain the conditional discrete distribution $D_i^{\mathrm{tr}}|D_i^{\mathrm{mes}}$:

$$\mathbf{P}(\,D_i^{\mathrm{tr}} = \exp(x_k)\Big|\, D_i^{\mathrm{mes}}) = \frac{p_k l_{i,k}}{\sum_{j=1}^{K} p_j l_{i,j}}\,, \tag{6.35}$$

where $l_{i,k} = \frac{1}{\sqrt{2\pi}\sigma_{Q,i}} \exp(-\frac{(\ln D_i^{\mathrm{mes}} - x_k)^2}{2\sigma_{Q,i}^2})$ is the probability density function (pdf) of conditional distribution $\ln D_i^{\mathrm{mes}}|\,[D_i^{\mathrm{tr}} = \exp(x_k)]$, which is evaluated at the point $D_i^{\mathrm{mes}}$.

Moments of the conditional distribution $D_i^{\mathrm{tr}}|D_i^{\mathrm{mes}}$ are equal to

$$\mathbf{E}(\,D_i^{\mathrm{tr}}\Big|\, D_i^{\mathrm{mes}}) = \frac{\sum_{k=1}^{K} \exp(x_k) p_k l_{i,k}}{\sum_{k=1}^{K} p_k l_{i,k}}\,, \tag{6.36}$$

$$\mathbf{E}[\,(D_i^{\mathrm{tr}})^2 \;\Big|\; D_i^{\mathrm{mes}}] = \frac{\sum_{k=1}^{K} \exp(2x_k) p_k l_{i,k}}{\sum_{k=1}^{K} p_k l_{i,k}}\,. \tag{6.37}$$

### 6.5.3 Full maximum likelihood method and its modification

Evaluate the likelihood function $L(Y_i, D_i^{\mathrm{mes}}, \theta)$ by the following steps. Suppose that $P(Y_i, D_i^{\mathrm{mes}}, D_i^{\mathrm{tr}})$ is joint probability distribution of the variables $Y_i$, $D_i^{\mathrm{mes}}$, $D_i^{\mathrm{tr}}$. Then by the formula for conditional probabilities we get

$$P(Y_i, D_i^{\mathrm{mes}}, D_i^{\mathrm{tr}}) = P(Y_i \,|\, D_i^{\mathrm{mes}}, D_i^{\mathrm{tr}}) \cdot P(D_i^{\mathrm{mes}}, D_i^{\mathrm{tr}}) = P(Y_i \mid D_i^{\mathrm{tr}}) \cdot P(D_i^{\mathrm{mes}} \mid D_i^{\mathrm{tr}}) \cdot P(D_i^{\mathrm{tr}})\,. \tag{6.38}$$

In (6.38), passing from probabilities to the likelihood function, given the fact that the conditional distribution $D_i^{\mathrm{mes}}|D_i^{\mathrm{tr}}$ is lognormal, and being within the structural model

4 Here the discrete probabilities $p_k$ are estimated by the ML method which lies in maximization of the functional

$$\prod_{i=1}^{N} \mathbf{P}(D^{\mathrm{tr}} = D_i^{\mathrm{tr}}) = \prod_{i=1}^{N} \sum_{k=1}^{K} p_k l_{i,k}\,,$$

provided $p_k \geq 0$, $k = 1, 2, \ldots, K$, $p_1 + p_2 + \ldots + p_K = 1$. The weights $l_{i,k}$ are given further.

(6.30), we obtain

$$L(Y_i, D_i^{\text{mes}}, D_i^{\text{tr}}, \theta) = \left(\frac{\lambda(D_i^{\text{tr}}, \theta)}{1+\lambda(D_i^{\text{tr}}, \theta)}\right)^{Y_i} \left(\frac{1}{1+\lambda(D_i^{\text{tr}}, \theta)}\right)^{1-Y_i} \times \\ \times \frac{\exp\left(-\frac{(\ln D_i^{\text{tr}} - \mu_1)^2}{2\sigma_1^2}\right)}{D_i^{\text{tr}}\sqrt{2\pi}\sigma_1} \frac{\exp\left(-\frac{(\ln D_i^{\text{mes}} - \ln D_i^{\text{tr}})^2}{2\sigma_{Q,i}^2}\right)}{D_i^{\text{mes}}\sqrt{2\pi}\sigma_{Q,i}}. \tag{6.39}$$

To get rid of unknown covariate $D_i^{\text{tr}}$ it is necessary to make convolution with respect to this variable:

$$L(Y_i, D_i^{\text{mes}}, \theta) = \int_{-\infty}^{\infty} \left(\frac{\lambda(t, \theta)}{1+\lambda(t, \theta)}\right)^{Y_i} \left(\frac{1}{1+\lambda(t, \theta)}\right)^{1-Y_i} \times \\ \times \frac{\exp\left(-\frac{(\ln t - \mu_1)^2}{2\sigma_1^2}\right)}{t\sqrt{2\pi}\sigma_1} \frac{\exp\left(-\frac{(\ln D_i^{\text{mes}} - \ln t)^2}{2\sigma_{Q,i}^2}\right)}{D_i^{\text{mes}}\sqrt{2\pi}\sigma_{Q,i}} dt. \tag{6.40}$$

Thus, the likelihood function taking into account the contributions of all observations is equal to

$$L(\theta) = \prod_{i=1}^{N} \int_0^{\infty} \left(\frac{\lambda(t, \theta)}{1+\lambda(t, \theta)}\right)^{Y_i} \left(\frac{1}{1+\lambda(D^{\text{tr}}, \theta)}\right)^{1-Y_i} \times \\ \times \frac{\exp\left(-\frac{(\ln t - \mu_1)^2}{2\sigma_1^2}\right)}{t\sqrt{2\pi}\sigma_1} \frac{\exp\left(-\frac{(\ln D_i^{\text{mes}} - \ln t)^2}{2\sigma_{Q,i}^2}\right)}{D_i^{\text{mes}}\sqrt{2\pi}\sigma_{Q,i}} dt. \tag{6.41}$$

The corresponding log-likelihood function is written as

$$l(\theta) = \sum_{i=1}^{N} \ln \int_0^{\infty} \left(\frac{\lambda(t, \theta)}{1+\lambda(t, \theta)}\right)^{Y_i} \left(\frac{1}{1+\lambda(t, \theta)}\right)^{1-Y_i} \times \\ \times \frac{\exp\left(-\frac{(\ln t - \mu_1)^2}{2\sigma_1^2}\right)}{t\sqrt{2\pi}\sigma_1} \frac{\exp\left(-\frac{(\ln D_i^{\text{mes}} - \ln t)^2}{2\sigma_{Q,i}^2}\right)}{D_i^{\text{mes}}\sqrt{2\pi}\sigma_{Q,i}} dt. \tag{6.42}$$

Given the fact that the response $Y_i$ takes only two values 0 and 1, the latter equality can be written in the form:

$$l(\theta) = \sum_{i=1}^{N} \left[ Y_i \ln \int_0^{\infty} \left(\frac{\lambda(t, \theta)}{1+\lambda(t, \theta)}\right) \frac{\exp\left(-\frac{(\ln t - \mu_1)^2}{2\sigma_1^2}\right)}{t\sqrt{2\pi}\sigma_1} \frac{\exp\left(-\frac{(\ln D_i^{\text{mes}} - \ln t)^2}{2\sigma_{Q,i}^2}\right)}{D_i^{\text{mes}}\sqrt{2\pi}\sigma_{Q,i}} dt + \right. \\ \left. + (1-Y_i) \ln \int_0^{\infty} \left(\frac{1}{1+\lambda(t, \theta)}\right) \frac{\exp\left(-\frac{(\ln t - \mu_1)^2}{2\sigma_1^2}\right)}{t\sqrt{2\pi}\sigma_1} \frac{\exp\left(-\frac{(\ln D_i^{\text{mes}} - \ln t)^2}{2\sigma_{Q,i}^2}\right)}{D_i^{\text{mes}}\sqrt{2\pi}\sigma_{Q,i}} dt \right]. \tag{6.43}$$

In the integrals (6.43), change variables:

$$\begin{aligned} dz &= \frac{1}{t\sqrt{2\pi}\sigma_1^2}\exp\left(-\frac{(\ln t-\mu_1)^2}{2\sigma_1^2}\right)dt, \\ z &= \int\limits_0^t \frac{1}{t\sqrt{2\pi}\sigma_1^2}\exp\left(-\frac{(\ln t-\mu_1)^2}{2\sigma_1^2}\right)dD^{\mathrm{tr}} = G(t)\ , \\ t &= G^{-1}(z)\ . \end{aligned} \tag{6.44}$$

Since $z = G(t)$ is the cdf of lognormal law, it holds that

$$z(0) = 0, \quad z(\infty) = 1\ . \tag{6.45}$$

Substituting (6.44) and (6.45) into (6.43), we obtain

$$\begin{aligned} l(\theta) = \sum_{i=1}^{N}\Bigg[ & Y_i \ln \int\limits_0^1 \left(\frac{\lambda(G^{-1}(z),\theta)}{1+\lambda(G^{-1}(z),\theta)}\right)\frac{\exp\left(-\frac{(\ln D_i^{\mathrm{mes}}-\ln G^{-1}(z))^2}{2\sigma_{Q,i}^2}\right)}{D_i^{\mathrm{mes}}\sqrt{2\pi}\sigma_{Q,i}}\,dz + \\ & + (1-Y_i)\ln\int\limits_0^1\left(\frac{1}{1+\lambda(G^{-1}(z),\theta)}\right)\frac{\exp\left(-\frac{(\ln D_i^{\mathrm{mes}}-\ln G^{-1}(z))^2}{2\sigma_{Q,i}^2}\right)}{D_i^{\mathrm{mes}}\sqrt{2\pi}\sigma_{Q,i}}dz\Bigg]\ . \end{aligned} \tag{6.46}$$

Integrals (6.46) are improper and have singularities at the points $z = 0$ and $z = 1$, since $G^{-1}(0) = 0$ and $G^{-1}(z) \to +\infty$, as $z \to 1$. However, these integrals exist as the absolutely convergent improper Riemann integrals. For their computation, one can apply the Monte Carlo method. For the computation of (6.46), it is convenient to use the equality

$$G^{-1}(z) = \exp(\mu_1 + \sigma_1\Phi^{-1}(z))\ , \tag{6.47}$$

where $\Phi(z)$ is the cdf of standard normal law.

The FML method lies in finding such a vector parameter $\theta$ at which the likelihood function attains its maximum:

$$\theta:\ l(\theta) \to \max\ . \tag{6.48}$$

Within the parametric version of the method, the distribution of true doses is parameterized as $D^{\mathrm{tr}} \sim \mathrm{LN}(\mu_1, \sigma_1^2)$, and therefore, the problem is reduced to the optimization of expression (6.46) using (6.47).

Peculiarity of the nonparametric modification of the FML method lies in refusal to parameterize the distribution of $D^{\mathrm{tr}}$, and the empirical cdf of $D^{\mathrm{tr}}$ is found by means of relation (6.34). Then in expression (6.46), the function $G^{-1}(z)$ has to be replaced by the inverse empirical cdf of $D^{\mathrm{tr}}$.

### 6.5.4 SIMEX method and its modification

The estimate obtained by the *SIMEX* method is randomized, i.e., it is a random function of observations. Such method of risk parameters estimation was used in Kopecky et al. (2006), but the approach proposed below allows taking into account more accurately the structure of the measured doses. Note that the *SIMEX* method does not require knowledge of the true dose distribution (see overview of the method in Section 1.4.6).

In order to take into account the classical error, the following algorithm is used.

Select a natural number $B \geq 2$. It is desirable that $B$ is large enough, e.g., $B = 100$. Then we generate random perturbations of the logarithms of doses

$$U^*_{b,i} \sim \mathrm{N}(0, \sigma^2_{Q,i}) , \quad b = 1, \ldots, B , \quad i = 1, \ldots, N . \tag{6.49}$$

The perturbations $U^*_{b.i} \sim \mathrm{N}(0, \sigma^2_{Q,i})$, $b = 1, \ldots, B$, $B = 100$ are generated so that, for fixed $i$, the random variables $U^*_{b,i}$ have normal joint distribution, with

$$\mathbf{E}[U^*_{b_1,i} U^*_{b_2,i}] = -\frac{\sigma^2_{Q,i}}{B-1} , \quad b_1 \neq b_2 . \tag{6.50}$$

Then it holds that $\sum_{b=1}^{B} U^*_{b,i} = 0$, $i = 1, \ldots, N$. The latter requirement yields less span and less deviation of the estimates from the true value (see Appendix B).

Select a set $\Lambda = \{0, 0.2, 0.4, 0.6\}$. The perturbed doses are computed for each $\kappa \in \Lambda$:

$$D^*_{b,i}(\kappa) = D^{\mathrm{mes}}_i \exp(\sqrt{\kappa} U^*_{b,i}), \kappa \in \Lambda . \tag{6.51}$$

Compute the ordinary (naive) estimates for $\kappa = 0, 0.2, 0.4, 0.6$ and average the result over $b$. For example, for the linear model of absolute risk (6.5), the estimates $\hat{\lambda}^*_0(\kappa)$ and $\widehat{EAR}^*(\kappa)$ are computed as follows:

$$\begin{aligned} \hat{\lambda}^*_0(\kappa) &= \frac{1}{B} \sum_{b=1}^{B} \hat{\lambda}^*_{0b}(\kappa) , \\ \widehat{EAR}^*(\kappa) &= \frac{1}{B} \sum_{b=1}^{B} \widehat{EAR}^*_b(\kappa), \kappa \in \Lambda . \end{aligned} \tag{6.52}$$

The functions $\hat{\lambda}^*_0(\kappa)$ and $\widehat{EAR}^*(\kappa)$ are extrapolated numerically to the point $\kappa = -1$, and finally we get the *SIMEX* estimates for the parameters $\lambda_0$ and $EAR$.

The numerical extrapolation of the functions $\hat{\lambda}^*_0(\kappa)$ and $\widehat{EAR}^*(\kappa)$ can be performed using the least squares method for approximation by the second degree polynomial. Exact formula for the value of extrapolation polynomial at the point $\kappa = -1$ is as follows:

$$\hat{\theta}^*(-1) = 12.45\, \hat{\theta}^*(0) - 9.35\, \hat{\theta}^*(0.2) - 10.65\, \hat{\theta}^*(0.4) + 8.55\, \hat{\theta}^*(0.6) . \tag{6.53}$$

Here $\hat{\theta}^*(\kappa)$ denotes either $\hat{\lambda}^*_0(\kappa)$ or $\widehat{EAR}^*(\kappa)$.

Confidence intervals for the *SIMEX* estimate are constructed using the estimate of the covariance matrix. The latter estimate is computed by recommendations outlined in the monograph by Carroll et al. (2006) (see also Section 1.4.6). We have estimated the variance of estimates $\hat{\theta}_b^*(\kappa)$ using $Y_i$ and $D_{b,i}^*(\kappa)$ as the data and applying the sandwich formula (see Appendix A2). Denote by $\hat{\tau}_b^2(\kappa)$ the estimate for the variance of $\hat{\theta}_b^*(\kappa)$. Then the estimate of the variance for the *SIMEX* estimate is the value of function $\frac{1}{B}\sum_{b=1}^{B}\hat{\tau}_b^2(\kappa) - \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b^*(\kappa) - \hat{\theta}^*(\kappa))^2$ being extrapolated to the point $\kappa = -1$.

The efficient *SIMEX* is a combination of the *SIMEX* method and the corrected score method. The efficient *SIMEX* is characterized by the circumstance that instead of (6.51) we use the doses

$$D_{b,i}^*(\kappa) = \begin{cases} D_i^{\text{mes}} \exp\left(-\frac{\sigma_{Q,i}^2}{2}\right), & \text{if } Y_i = 0\,, \\ D_i^{\text{mes}} \exp(\sqrt{\kappa} U_{b,i}^*), & \text{if } Y_i = 1\,. \end{cases} \tag{6.54}$$

As a result, the system of equations for estimates of the parameters of linear model (6.8) for each $b = 1, \ldots, B\,, \quad B = 100$, takes the form

$$\sum_{i=1}^{N}(1 - Y_i)\left(1 + \widehat{ERR}_b^*(\kappa) D_i^{\text{mes}} \exp\left(-\frac{\sigma_{Q,i}^2}{2}\right)\right) = \sum_{i=1}^{N}\frac{Y_i}{\hat{\lambda}_{0,b}^*(\kappa)}\,, \tag{6.55}$$

$$\sum_{i=1}^{N}(1 - Y_i) = \sum_{i=1}^{N}\frac{Y_i}{\hat{\lambda}_{0,b}^*(\kappa)(1 + \widehat{ERR}_b^*(\kappa) D_{b,i}^*(\kappa))}\,. \tag{6.56}$$

Here, the left-hand side of (6.55) is obtained by virtue of taking into account errors in doses and correction of the expression $\sum_{i=1}^{N}(1 - Y_i)(1 + ERR \cdot D_i)$ in the first equation of (6.8), and the right-hand side of equation (6.56) is done by the disturbance of $D_i^{\text{mes}}$ in the expression

$$\sum_{i=1}^{N}\frac{Y_i}{\hat{\lambda}_0(1 + ERR \cdot D_i^{\text{mes}})}\,.$$

The above-proposed modification of the *SIMEX* method is efficient in the computational sense for the following reasons:

- every naive estimate being derived by solving equations (6.55) and (6.56) is efficient from the computational point of view,
- some computations are common to all naive estimates and are performed only once.

A more detailed justification of the proposed procedure is presented in Appendix B.

### 6.5.5 Stochastic simulation of the classical multiplicative error

To check efficiency of the developed (not naive) methods for regression parameters estimation, the stochastic simulation study was conducted. The simulation was based

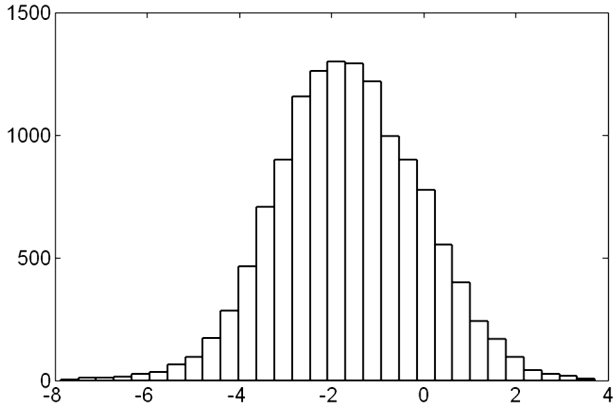


**Fig. 6.3:** Histogram of $\ln(D^{\mathrm{tr}})$.

on the epidemiological studies described in Section 6.4.2. When simulating the thyroid cancer incidence rate at a fixed time interval, the two-parameter linear model (6.5) of absolute risk was used. The true model parameters were chosen to be close to the estimates obtained during the epidemiological studies of thyroid cancer in Ukraine (Likhtarov et al., 2006a; Tronko et al., 2006), namely:

$$\begin{aligned} \lambda_0 &= 10^{-4}\,\frac{\text{cases}}{\text{person years}}\,, \\ EAR &= 5\times 10^{-4}\,\frac{\text{events}}{\text{Gy}\cdot(\text{person years})}\,. \end{aligned} \tag{6.57}$$

In the simulation process, it was assumed that thyroid exposure doses were observed only with the classical multiplicative error. In so doing, various values of measurement error variances were used. Geometric standard deviation $\mathrm{GSD}_Q = \exp(\sigma_{Q,i})$ of the classical multiplicative error in the absorbed thyroid dose varied in the range from 1.5 to 5.0, for all $i = 1, \ldots, N$.

During the simulation, 1000 data sets were generated. However, the unperturbed doses $D_i^{\mathrm{tr}}$ (Figure 6.3) were based on real data and coincided in all realizations.

Estimation of the absolute risk parameters was performed by several methods:

- *Naive method* is the ordinary Maximum Likelihood one, in which the thyroid doses were assumed free of errors,
- *Parametric full maximum likelihood (PFML)*, the method which takes into account the errors in exposure doses using the integral convolution (6.46) under the assumption that the value $D^{\mathrm{tr}}$ has lognormal distribution,
- *Nonparametric full maximum likelihood (NPFML)*, the method in which, unlike the *PFML*, the distribution of groups $D^{\mathrm{tr}}$ is not parameterized and the empirical distribution of $D^{\mathrm{tr}}$ is found using relation (6.34),
- *Parametric regression calibration (PRC)* described in Section 6.5.1,
- *Nonparametric regression calibration (NPRC)* described in Section 6.5.2,
- *Ordinary SIMEX* and *efficient SIMEX* presented in Section 6.5.4.

Each estimate was computed for 1000 realizations of doses and cases. After this, the corresponding risk values were averaged (in the form of arithmetical mean).

It is known that statistical inconsistency of estimators (i.e., nonconvergence of obtained estimators to the true values, as the sample size tends to infinity) takes place for the naive method and all methods of Regression Calibration. Therefore, instead of the confidence interval, the deviance interval (95% DI) was computed based on the 2.5 and 97.5 percent quantiles of the estimates for 1000 realizations.

The simulation results are presented in Tables 6.1 and 6.2 and Figures 6.4 and 6.5.

### 6.5.6 Simulation results

**The naive method**

Analysis of the simulation results showed that in case of the classical measurement error, the naive method underestimates the excess absolute risk *EAR*. At the same time, the naive estimate of the background incidence rate is overestimated. Similar effect is known in statistical literature (Carroll et al., 2006) as "*attenuation effect*" (see Section 2.1 for the case of linear model).

For the naive estimate, the attenuation effect increases, as the variance of the classical error grows. Notice that for sufficiently large variances of the classical errors (for instance $\mathrm{GSD}_Q = 5$), the naive errors of excess absolute risk and baseline risk may differ from the model (true) values in several times, moreover the first ones will be underestimated, and the second ones will be overestimated. This effect is clearly seen in Figures 6.4 and 6.5.

**Regression calibration and full maximum likelihood**

The estimates obtained by the parametric and nonparametric Regression Calibration and by the FML method have relatively small bias (see Tables 6.1 and 6.2). When using parametric methods, the distribution of group doses was approximated by lognormal law. For nonparametric methods, there is no need to approximate group doses because these methods use the empirical distribution of the doses. Since the parametric and nonparametric methods gave similar results, we infer that the parameterization of true doses distribution in the methods of regression calibration and FML is adequate. The estimates of background incidence rate and excess absolute risk for different variances of the classical error are shown in Figures 6.4 and 6.5.

**SIMEX method**

Although the *SIMEX* method is robust, i.e., stable at violation of the assumption about the distribution of dose group, its behavior deteriorates in case of large errors. Typically the *SIMEX* estimates have significant bias for large measurement errors. The reason is as follows: we use square extrapolation function, but the naive MLE as a

**Table 6.1:** Estimates of $\lambda_0 \times 10^4$ for different variances of classical multiplicative error in absorbed thyroid doses.

| $GSD_Q$ | Estimation method (model value of $\lambda_0 \times 10^4$ is 1.0) | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Naive (EPICURE) | | Full maximum likelihood | | | | Regression calibration | | | | *SIMEX* | | | |
| | | | Parametric | | Nonparametric | | Parametric | | Nonparametric | | Ordinary | | Efficient | |
| | AM | 95% DI | AM | 95% DI | AM | 95% DI | AM | 95% DI | AM | 95% DI | AM | 95% DI | AM | 95% DI |
| 1.5 | 1.16 | 0.83–1.49 | 1.05 | 0.73–1.36 | 1.06 | 0.77–1.33 | 1.04 | 0.72–1.35 | 1.04 | 0.73–1.35 | 1.02 | 0.66–1.38 | 1.02 | 0.67–1.37 |
| 2 | 1.39 | 1.06–1.73 | 1.07 | 0.75–1.38 | 1.04 | 0.71–1.34 | 1.05 | 0.73–1.36 | 1.04 | 0.75–1.34 | 1.01 | 0.64–1.39 | 1.02 | 0.66–1.37 |
| 3 | 1.87 | 1.49–2.21 | 1.11 | 0.79–1.43 | 1.05 | 0.71–1.34 | 1.06 | 0.76–1.39 | 1.07 | 0.76–1.41 | 0.99 | 0.58–1.41 | 1.02 | 0.64–1.40 |
| 5 | 2.56 | 2.10–2.96 | 1.16 | 0.80–1.50 | 1.03 | 0.70–1.39 | 1.13 | 0.75–1.49 | 1.06 | 0.74–1.46 | 1.15 | 0.65–1.64 | 1.05 | 0.65–1.46 |

**Table 6.2:** Estimates of $EAR \times 10^4$ for different variances of classical multiplicative error in absorbed thyroid doses.

| $GSD_Q$ | Estimation method (model value of $EAR \times 10^4$ is 5.0) | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Naive (EPICURE) | | Full maximum likelihood | | | | Regression calibration | | | | *SIMEX* | | | |
| | | | Parametric | | Nonparametric | | Parametric | | Nonparametric | | Ordinary | | Efficient | |
| | AM | 95% DI | AM | 95% DI | AM | 95% DI | AM | 95% DI | AM | 95% DI | AM | 95% DI | AM | 95% DI |
| 1.5 | 4.33 | 3.57–5.31 | 4.82 | 3.97–5.87 | 5.08 | 4.33–6.35 | 4.88 | 4.03–5.95 | 5.00 | 4.12–6.09 | 4.95 | 4.09–5.80 | 4.95 | 4.09–5.80 |
| 2 | 3.39 | 2.81–4.21 | 4.69 | 3.88–5.72 | 5.17 | 4.40–6.44 | 4.84 | 4.04–5.96 | 4.97 | 4.09–6.10 | 4.91 | 4.04–5.77 | 4.95 | 4.08–5.82 |
| 3 | 1.92 | 1.49–2.40 | 4.49 | 3.72–5.55 | 5.14 | 4.30–6.32 | 4.78 | 3.94–5.87 | 4.88 | 4.00–6.05 | 4.44 | 3.67–5.22 | 4.95 | 4.04–5.87 |
| 5 | 0.62 | 0.44–0.82 | 4.30 | 3.56–5.34 | 5.12 | 4.16–6.63 | 4.57 | 3.73–5.75 | 4.81 | 3.91–5.93 | 2.65 | 2.20–3.11 | 4.90 | 3.88–5.92 |

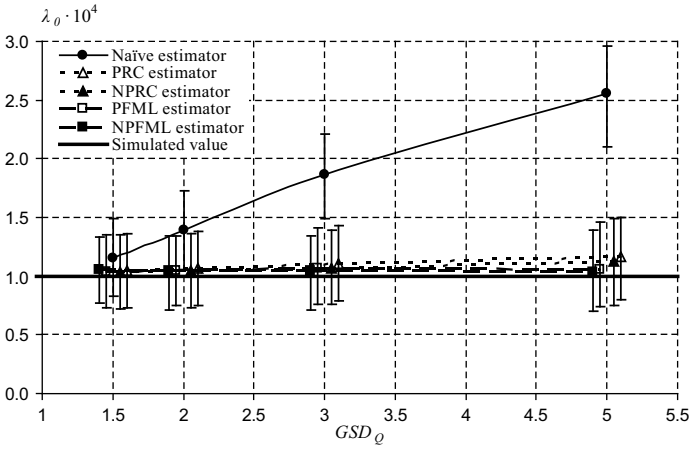


**Fig. 6.4:** Estimates of background incidence rate for different variances of classical multiplicative error in absorbed thyroid doses.

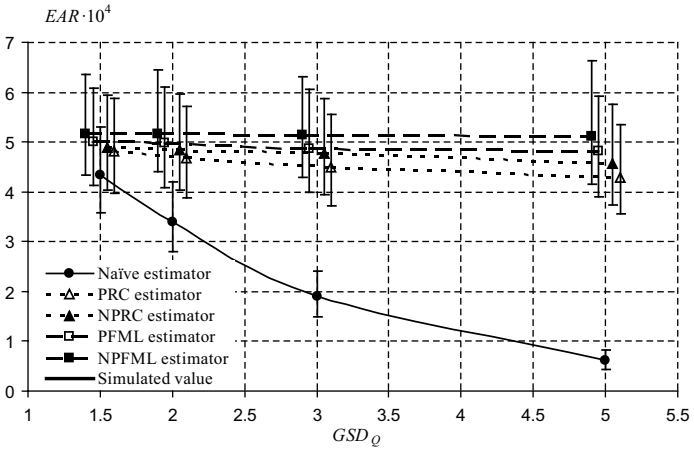


**Fig. 6.5:** Estimates of excess absolute risk for different variances of classical multiplicative error in absorbed thyroid doses.

function of the variance of additional error deviates from quadratic law. In the efficient computational *SIMEX* procedure, the modified naive estimates are used being defined by equations (6.55) and (6.56). Such naive estimates depend on the variance of additional error almost like a quadratic function. As a result, the efficient *SIMEX* method yields a relatively small bias even for quite large errors. In simulations, the maximal level of classical error was $\mathrm{GSD}_Q = 5$, and for larger errors of classical type, the efficient *SIMEX* estimates can become worse.

# 7 Radiation risk estimation for persons exposed by radioiodine as a result of the Chornobyl accident

As a result of the Chornobyl accident in 1986, much of the territories of Ukraine, Belarus, and Russia were subjected to radioactive contamination, and inhabitants of these territories to radioactive exposure. The most significant was the thyroid exposure due to intake of iodine radioisotopes, primarily of $^{131}$I (Likhtarev et al., 1993a, 1993b, 1995b).

Already in 5–6 years after the accident a dramatic increase began to be exhibited in the thyroid cancer incidence of children and adolescents, which resided in the areas where the estimates of thyroid exposure doses occurred to be quite high (Likhtarev et al., 1995a; Buglova et al., 1996).

In fact, the growth of thyroid cancer incidence of children and adolescents caused by internal thyroid exposure from Chornobyl fallout was the main statistically significant remote effect of the Chornobyl accident. It comes as no surprise that the phenomenon caused enormous interest of radio-epidemiologists all over the world so that a series of studies was conducted in Ukraine, Belarus, and Russia (Jacob et al., 2006; Likhtarov et al., 2006a; Tronko et al., 2006; Kopecky et al., 2006; Zablotska et al., 2011).

The interpretation of results of most radio-epidemiologic studies was based on a number of assumptions, primarily on the estimates of exposure doses. The assumptions include the following:

- It was recognized that the dose estimates include uncertainty which is typically significant.
- Even in the cases where the level of dose errors turned out to be determined, the analytical procedures of risk analysis ignored that fact.

As a result of the above-mentioned general properties of the dosimetric support for epidemiological studies, it was the merely stochastic nature of thyroid cancer cases that was taken into account in analytical procedures of risk analysis, whereas the exposure doses of subjects were assumed precise.

Studies performed by Kukush et al. (2011) and Masiuk et al. (2016) demonstrated that the dose uncertainties can be quite correctly taken into account in the process of risk analysis. Some difficulty lies in the fact that the main sources of dose uncertainties are related to errors in estimation of the weight of exposed organ, instrumental measurements of radioactivity of the organ at certain moment, and the ecological component of dose. The papers by Kukush et al. (2011) and Masiuk et al. (2016) show that the estimates of thyroid weight and ecological component of thyroid dose include Berkson error, and the instrumental measurements contain the classical error. The size of Berkson error is easily estimated by the Monte Carlo method (Likhtarov et al., 2014), while a specific analysis is required to estimate the size of the classical error.

## 7.1 Estimation of error level in direct thyroid radioactivity measurements for children and adolescents exposed by $^{131}$I in May–June 1986

### 7.1.1 Peculiarities of organization and performance of the mass thyroid dosimetric monitoring

Conducting any campaign of mass measurement of radioiodine isotopes uptake in thyroid gland (further called *thyroid dosimetric monitoring*) is always limited in time for solely physical reasons. This is due to the fact that the most long-lived radioisotope of iodine $^{131}$I has a half-life of 8 days, and within 5–8 weeks after the accidental emission of radioisotope mixture in the environment the gamma radiation from the radioiodine gets invisible against the background of cesium radioisotopes. Since the deployment of thyroid dosimetric monitoring is always associated with necessity to solve a lot of organizational and technical issues, consequently the monitoring itself should be carried out in less than a month.

The need for coverage by measurements as many people in the areas undergone by the accident as possible and being combined with short term of the monitoring leads to the fact that such a monitoring has significant differences from laboratory studies. First of all they are: attraction of staff without any experience in dosimetry, use of nonspecialized equipment and many types of devices, reducing endurance in the measurements, and simplified calibration of devices. All this leads to the values of measurement errors that are significantly larger than in laboratory studies, and also to the appearance of new error components, which would be avoided during laboratory tests.

### 7.1.2 Implementation of thyroid dosimetric monitoring at the territory of Ukraine in 1986

The thyroid dosimetric monitoring in the areas affected by the Chornobyl accident was conducted by special emergency teams under supervision of the Ministry of Public Health of the UkSSR. Advisory assistance was provided to them by a team from the Research Institute of Hygiene of Maritime Transport (Leningrad city), which developed a general method of measurement and provided emergency crews by the referent sources of radioisotope $^{131}$I necessary for calibration of devices.

The first measurements were taken in mobile radiometric laboratories (MRL), in which the spectrometers brought from the city of Leningrad were used. The purpose of these measurements was to obtain initial estimates of doses and estimates of risk level due to the radioiodine pollution. After obtaining the first results, it was decided to expand the scope of measurements by connecting to them local medical institutions and available equipment.

Immediately after the accident, the Ministry of Public Health of the UkSSR got several specialized gamma-thyroid-radiometers GTRM-01ts; in addition, medical institutions, sanitary, and epidemiological stations provided a lot of nonprofessional devices of the two main classes: the so-called window radiometers, which were able to record gamma exposure in a narrow energy window (DSU, UR, NC), and integrated radiometers, mainly the scintillation field radiometers SRP-68-01 intended for geological prospecting. The gauge sources were prepared especially at the Research Institute of Endocrinology and Metabolism (city of Kyiv) using solutions with $^{131}$I obtained from the production association "Isotope."

Thus, during the thyroid dosimetric monitoring the measurements were carried out by devices of two main types: energy-selective spectrometers (usually single-channel ones) and integrated radiometers (Likhtarev et al., 1995b; Likhtarov et al., 2015). Table 7.1 presents data of all the models of spectrometric and nonspectrometric devices that were used in various oblasts of Ukraine, as well as the appropriate number of measurements performed by the devices.

**Table 7.1:** The peculiarities of the thyroid dosimetric monitoring in 1986: types and models of the devices used, the number of measurements, the duration of the monitoring.

| Type of devices | Model | Number of measurements | Percentage of all measurements | Duration of measurements in 1986 |
|---|---|---|---|---|
| One-channel and multichannel spectrometers | NC-150, NC-350 | 19 321 | 13.2 | 14.05–11.06 |
| | GTRM-01ts | 17 834 | 12.2 | 17.05–30.06 |
| | UR | 9745 | 6.7 | 17.05– 6.06 |
| | DSU-68 | 4452 | 3.0 | 25.05–31.05 |
| | DSU-2 | 1345 | 0.9 | 18.05–29.05 |
| | MRL | 7 | < 0.1 | 8.05–12.05 |
| Integrated radiometers | SRP 68-01 | 93 717 | 64.0 | 30.04–25.06 |
| | DP-5B | 5 | < 0.1 | 8.05–19.05 |
| All devices | | 146 426 | 100 | 30.04–30.06 |

As seen from Table 7.1, more than half of all the measurements were made by integrated radiometers SRP-68-01. Using devices of this type, the greatest amount of the measurements was performed in Zhytomyr, Odessa, and Chernihiv Oblasts, and also in the Crimea. As to measurements made by the spectrometric instruments, the most of them were done by specialized thyroid radiometer GTRM-01ts (mainly in Chernihiv and Zhytomyr Oblasts) and also by universal window radiometers NC-150 and NC-350 produced by the plant "Gamma" (Hungary). In whole, the most number of the measurements were performed in Zhytomyr, Chernihiv, Odessa Oblasts, and in the Crimea.

As a rule, the scheme of measurements was as follows. The measurements were carried out in well-ventilated premises with hourly wet cleaning. To reduce background of radiation, the detectors of devices were protected with lead collimators. For field radiometers SRP 68-01 with factory kitting without collimator, the homemade collimators were used made of scrap materials[1]. During the measurement, the detector device was being brought to the neck of a person measured, before having cleaned with a cotton wool wet in alcohol, then single counting the number of pulses or their intensity was written down in sheet. Every hour or every day the background in the same point of space was measured and its value was written down to the measuring list. Also, in order to make a calibration, the count of pulses coming from a bottle phantom (cylindrical bottle of 10 ml containing a reference solution of isotope $^{131}$I) was being fulfilled either hourly or daily.

The results of measurements performed in the same settlement in the same day and by the same device and by the same team were recorded in so-called measuring list. (A typical list contained the results of 100–200 individual measurements, although in some cases the number of measurements performed by the team during the day could be about thousand.) In conformity with the established requirements to the measurement data, there were also to be recorded: the personal data of person (name, date or even year of birth), information on dosimetric team, on type of measuring device, the results of the device's calibration (calibration factor), and the value of radiation background in the room. Unfortunately, usually not all the data mentioned above were written down in the lists. Some of them had to be recovered during several cycles of the data processing.

Throughout June 1986, the bodies of people residing in contaminated areas were continuing to accumulate cesium radioisotopes, while $^{131}$I was continuing to disintegrate quickly. In this regard, in early June the thyroid dosimetric monitoring was decided to be finalized. Some measurements of thyroid activity against the background of growing cesium exposure lasted until the end of June, but 98% of all measurements were made up to June 6, i.e., the bulk of the monitoring was held during a period less than a month.

Thus, a huge organizational work done in short terms gave medical workers a unique array of data with more than 150 000 measurements of $^{131}$I content in the thyroid of residents from the most contaminated areas of northern Ukraine: Zhytomyr, Kyiv, and Chernihiv Oblasts. Of these, about 112 thousand of measurements were conducted among children and adolescents aged from 0 to 18 years (Likhtarev et al., 1993a, 1993b, 1995b; Likhtarov et al., 2015). At the beginning of mass measurements of $^{131}$I content in the thyroid, a considerable part of children and adolescents

**1** Usually a thin sheet of lead wrapped around the detector unit served as collimator. Collimator could shift to a few centimeters relatively to the end face of detector. The value of this shift is called "collimator depth" or "collimator shift".

from the suffered areas was removed to summer vacation spots in southern, the least suffered Oblasts of Ukraine, and the population of the 30 km zone adjacent to the Chornobyl nuclear power plant was evacuated completely. Therefore, about 47,000 measurements were performed at the territory of 10 Oblasts being rather far from Chornobyl, whereas about 103 000 measurements were made within the three northern Oblasts of Ukraine, in the areas distinguished by significant radionuclide contamination.

### 7.1.3 Calibration of measuring devices

In general, the estimate of the content $Q$ of iodine radioisotopes in the thyroid being the result of a direct measurement is defined as

$$Q = K_{\mathrm{b}} \cdot G(I_{\mathrm{th}} - f_{\mathrm{sh}} \cdot I_{\mathrm{bg}}) \,, \tag{7.1}$$

where $K_{\mathrm{b}}$ is a calibration factor (CF) of measuring device from the bottle phantom; $G$ is correcting coefficient to the CF which takes into account the difference of the measurements geometry between reference source (the bottle phantom) and a subject of the measurements; $I_{\mathrm{th}}$ and $I_{\mathrm{bg}}$ show the device indication[2] during the measurement of thyroid gland and gamma background, respectively; $f_{\mathrm{sh}}$ is the coefficient of gamma background screening by body of the subject and is a function dependent on both the subjects' anthropometric parameters and spectral characteristics of the background. According to the literature, $f_{\mathrm{sh}}$ is in the range 0.9–1 (Pitkevich et al., 1996; Zvonova et al., 1997).

The CF is determined by measurement of the reference radiation source with its activity $Q_{\mathrm{ref}}$ being known in advance. When calibrating by the bottle phantom $G = 1$ (due to the definition of correction coefficient) and $f_{\mathrm{sh}} = 1$, then from (7.1) we have

$$K_{\mathrm{b}} = \frac{Q_{\mathrm{ref}}}{I_{\mathrm{ref}} - I_{\mathrm{bg}}} \,, \tag{7.2}$$

i.e., CF is numerically equal to radioactivity, which corresponds to the device's indication unit and is the value reversed to the sensitivity of the device.

In general, the device sensitivity can vary in time for many reasons (e.g., because of the temperature dependence of the parameters of electronics). Therefore, to minimize errors in the measurement results it is advisable to calibrate the device immediately prior to the measurements of the subjects. This in turn implies that the dosimetric team should have a reference radiation source.

**2** Devices with three types of indication were used during the monitoring: devices with needle indicators registered intensity of pulses, devices with indicators showed percent from intensity of reference source, and the ones that indicated the number of accumulated pulses.

Wide geography of the measurements (they were held simultaneously in 14 Oblasts of Ukraine) made it almost impossible to supply all dosimetric teams with the reference radiation sources for calibration of the instruments. Thus, first of all such sources were delivered to the teams that worked with spectrometers. Therefore, for a large part of the radiometers SRP-68-01, the calibration factor was unknown and required recovery.

During mass monitoring in 1986, the devices' calibration was performed using a bottled phantom. The calibration factors $K_b$ obtained through the bottle phantoms should be corrected by a factor $G$, which accounts for the influence of geometry of the measurements, physical characteristics of the thyroid, and coating tissue in the neck area to a signal of the device's detector (equation (7.1)).

Since the size of thyroid is a function of person's mass, and therefore, a function of age and gender, the coefficient $G$ is also dependent on age and gender. For calculation of the correction coefficients for devices with collimators of any depth, the human thyroid was modeled as two double-axis ellipsoids of revolution. In this case, the $G$ is the ratio of intensities registered by a detector of radiation coming from a model of thyroid and a model of bottled phantom, provided they have the same content of radioiodine in them.

Figure 7.1 demonstrates the CF value received by phantom experiments and corrected by the value of correcting coefficient $G$ (Likhtarov et al., 2015). For comparison, empirical values of the CF are also presented for three age groups obtained in Kaidanovsky and Dolgirev (1997) on volunteers using SRP-68-01 with absent collimator. Thus, in the absence of calibration procedure for the radiometers SRP-68-01 as a value of the CF, we used the value $K_b = 90\,\text{Bq h}/\mu\text{R}$ adjusted for the age-dependent correction factor $G$.

### 7.1.4 Estimation of errors for direct measurements of the content of radioiodine in the thyroid

It is known (Gol'danskii et al., 1959) that at the fixed intensity of emission for a radioactive source $n$, the probability to register $k$ counts using a measuring device (such as the Geiger–Muller counter) for the time $t$ is defined by the Poisson distribution with parameter $nt$:

$$p_n(k) = \frac{(nt)^k}{k!} e^{-nt},\ k = 0, 1, 2, \ldots . \tag{7.3}$$

Based on (7.3) and the described above measurement methods of radioactivity $^{131}$I in the thyroid, we obtain

$$Q = K\left(\frac{k_{th}}{t_{th}} - f_{sh}\frac{k_{bg}}{t_{bg}}\right), \tag{7.4}$$

where $Q$ is the radioactivity of $^{131}$I in the thyroid, $k_{th}$ is the number of pulses registered by the device when measuring the radioactivity of $^{131}$I in the thyroid during the

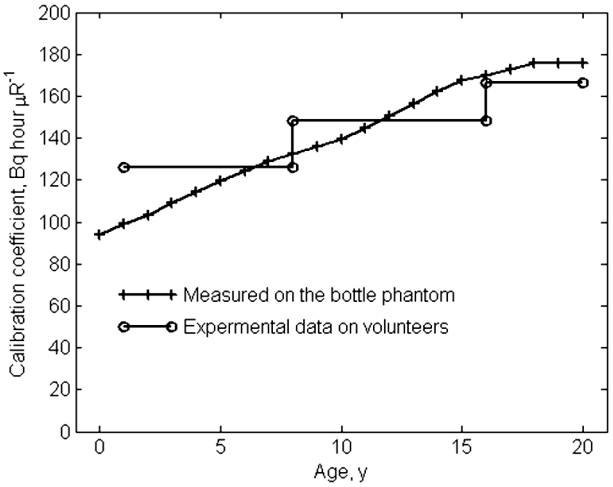


**Fig. 7.1:** Comparison between the experimental age-dependent calibration factors $K$ (see formula (7.5)) for the radiometers SRP-68-01 obtained on volunteers and gained using bottled phantom and taking into account the age-correcting coefficient $G$ (Likhtarov et al., 2015).

time $t_{\text{th}}$ of measurement, $k_{\text{bg}}$ is the number of pulses registered by the device when measuring radioactivity background during the time $t_{\text{bg}}$ of measurement, $f_{\text{sh}}$ is screening coefficient for the background radiation, and $K$ is the age-dependent calibration factor. The latter is the CF of device from the bottle phantom $K_{\text{b}}$ adjusted with the age-dependent geometric correcting factor $G$,

$$K = K_{\text{b}} \cdot G\,. \tag{7.5}$$

Because for large enough $n$, the Poisson distribution (7.3) is close to normal law (Molina, 1973), we can write

$$n_{\text{th}}^{\text{mes}} \sim \text{N}\left(n_{\text{th}}^{\text{tr}}, \sigma_{\text{th}}^2\right), n_{\text{bg}}^{\text{mes}} \sim \text{N}\left(n_{\text{bg}}^{\text{tr}}, \sigma_{\text{bg}}^2\right)\,, \tag{7.6}$$

where $\text{N}(m, \sigma^2)$ is normal law with expectation $m$ and variance $\sigma^2$; $n_{\text{th}}^{\text{mes}} = (k_{\text{th}})/(t_{\text{th}})$, $n_{\text{bg}}^{\text{mes}} = (k_{\text{bg}})/(t_{\text{bg}})$ are intensities of a radioactive source registered during the measurement of thyroid and background, respectively, and $\sigma_{\text{th}}^2 = (n_{\text{th}}^{\text{tr}})/(t_{\text{th}})$, $\sigma_{\text{bg}}^2 = (n_{\text{bg}}^{\text{tr}})/(t_{\text{bg}})$ are the variances of measurement errors. Index *tr* means the true value, while *mes* means the measured one.

In addition to the statistical error of registration, the values $n_{\text{th}}^{\text{mes}}$ and $n_{\text{bg}}^{\text{mes}}$ contain one more instrumental error, with variance $\sigma_{\text{dev}}^2$. One can estimate the full variances of measurement errors for both thyroid and background:

$$\hat{\sigma}_{\text{th}}^2 = \frac{n_{\text{th}}^{\text{mes}}}{t_{\text{th}}} + \sigma_{\text{dev}}^2\,, \quad \hat{\sigma}_{\text{bg}}^2 = \frac{n_{\text{bg}}^{\text{mes}}}{t_{\text{bg}}} + \sigma_{\text{dev}}^2 \tag{7.7}$$

Based on the method of calibrating the measuring device, one can write the approximate relation as follows:

$$K^{\text{mes}} \approx K^{\text{tr}}(1 + \delta_K \gamma_1) , \quad \gamma_1 \sim \text{N}(0, 1) , \tag{7.8}$$

where $\delta_K$ is based on the error of reference source $^{131}$I and the device's error.

Using (7.6)–(7.8), expression (7.4) can be modified as

$$Q^{\text{mes}} \approx K^{\text{tr}}(1 + \delta_K \gamma_1)(n_{\text{th}}^{\text{tr}} - f_{\text{sh}} n_{\text{bg}}^{\text{tr}} + \sigma_n \gamma_2) , \tag{7.9}$$

where $\sigma_n = \sqrt{\hat{\sigma}_{\text{th}}^2 + f_{\text{sh}}^2 \hat{\sigma}_{\text{bg}}^2}$ and $\gamma_2 \sim \text{N}(0, 1)$.

From (7.9), we get

$$Q^{\text{mes}} \approx K^{\text{tr}}(n_{\text{th}}^{\text{tr}} - f_{\text{sh}} n_{\text{bg}}^{\text{tr}} + (n_{\text{th}}^{\text{tr}} - f_{\text{sh}} n_{\text{bg}}^{\text{tr}})\delta_K \gamma_1 + \sigma_n \gamma_2 + \delta_K \sigma_n \gamma_1 \gamma_2) . \tag{7.10}$$

Since

$$Q^{\text{tr}} = K^{\text{tr}}(n_{\text{th}}^{\text{tr}} - f_{\text{sh}} n_{\text{bg}}^{\text{tr}}) , \tag{7.11}$$

then substituting (7.11) in (7.10), we get

$$Q^{\text{mes}} \approx Q^{\text{tr}} + K^{\text{tr}}(\sigma_n \gamma_2 + (n_{\text{th}}^{\text{tr}} - f_{\text{sh}} n_{\text{bg}}^{\text{tr}})\delta_K \gamma_1 + \delta_K \sigma_n \gamma_1 \gamma_2) \approx Q^{\text{tr}} + \sigma_Q^{\text{tr}} \gamma , \tag{7.12}$$

where $\sigma_Q^{\text{tr}} = K^{\text{tr}} \sqrt{\sigma_n^2 + \sigma_n^2 \delta_K^2 + (n_{\text{th}}^{\text{tr}} - f_{\text{sh}} n_{\text{bg}}^{\text{tr}})^2 \delta_K^2}$, $\gamma \sim \text{N}(0, 1)$.

Inasmuch as $n_{\text{th}}^{\text{tr}}$ and $n_{\text{bg}}^{\text{tr}}$ are unknown, the estimate of $\sigma_Q^{\text{tr}}$ will be the following:

$$\sigma_Q^{\text{mes}} = K^{\text{mes}} \sqrt{\sigma_n^2 + \sigma_n^2 \delta_K^2 + (n_{\text{th}}^{\text{mes}} - f_{\text{sh}} n_{\text{bg}}^{\text{mes}})^2 \delta_K^2} . \tag{7.13}$$

Finally, we get the observation model of thyroid radioactivity with the classical additive error:

$$Q^{\text{mes}} = Q^{\text{tr}} + \sigma_Q^{\text{mes}} \gamma . \tag{7.14}$$

### 7.1.5 Errors of the device calibration

The error $\delta_K$ of the age-dependent calibration factor $K$ specified in (7.5) can be found as

$$\delta_K = \sqrt{\delta_{\text{b}}^2 + \delta_G^2} , \tag{7.15}$$

where $\delta_{\text{b}}$ is relative error of the device's calibration using bottle phantom, and $\delta_G$ is relative error of age-correcting factor $G$.

For the radiometers SRP-68-01, the value of the relative error $\delta_G$ for geometric correction is based on empirical data obtained in Kaidanovsky and Dolgirev (1997) and is estimated as 15%. Since scintillation crystals of the detector spectrometers are located significantly farther from the thyroid, therefore the influence of measurement geometry is less for the detectors. So, for spectrometers, $\delta_G$ was expertly estimated as 5%.

The goal of device's calibration by a bottle phantom is to determine its sensitivity, i.e., to find out the values $n_{\text{ref}}^{\text{mes}} - n_{\text{bg}}^{\text{mes}}$ caused by radioactivity $Q_{\text{ref}}$ of a reference radiation source. Therefore, $\delta_{\text{b}}$ is specified as

$$\delta_{\text{b}} = \sqrt{\delta_{\text{ref}}^2 + \left(\frac{\sigma_S}{n_{\text{ref}} - n_{\text{bg}}}\right)^2}, \tag{7.16}$$

where $\delta_{\text{ref}}$ is relative error of activity for the reference radioactive source, which is known from the technical documentation of the provider (here: Production Association "Isotope"), and $\sigma_S$ is error in measuring the intensity of the reference source.

Since the process of calibration by a bottle phantom is similar to measurement of radioactivity in the thyroid, then the error $\sigma_S$ is calculated similarly:

$$\sigma_S = \sqrt{\hat{\sigma}_{\text{ref}}^2 + \hat{\sigma}_{\text{bg}}^2}, \tag{7.17}$$

where $\hat{\sigma}_{\text{ref}}^2 = (n_{\text{ref}}^{\text{mes}})/(t_{\text{ref}}) + \sigma_{\text{dev}}^2$ is the estimate of error variance for measuring the intensity of the reference source during the time period $t_{\text{ref}}$.

For devices with missing information about the calibration, $\delta_{\text{b}}$ was expertly estimated as 0.2 (i.e., 20%).

### 7.1.6 Analysis of relative errors in direct measurements of thyroid radioactivity content

The distribution of activities of radioiodine in the thyroid calculated according to the aforementioned methodology has clearly expressed lognormal nature (Figure 7.2) with a geometric mean (GM) equal to 4.8 kBq and geometric standard deviation (GSD) equal to 3.8. The spread of the activities occurred to be significant, namely, 90% of all values of the thyroid radioactivity, are in the range 0.58–47 kBq.

It should be emphasized that there is a deviation of the distribution from the lognormal one in the region of small values of radioiodine content. This deviation (the left side of Figure 7.2) is due to the fact that the results of unreliable measurements were censored. The measurements were considered reliable, if the probability to detect a net signal (that is difference between thyroid signal and background signal) on the assumption that its true value equals zero was not more than 25%. This is equivalent to the condition $(n_{\text{th}} - f_{\text{sh}} \cdot n_{\text{bg}}) \geq 0.68\sigma_n$. In other words, the critical limit of radioiodine in the thyroid was accepted at level $0.68\sigma_n$. In that case, if the result of the measurement was less than the critical limit, it is replaced by a half of the critical limit. With the proviso that $(n_{\text{th}} - f_{\text{sh}} \cdot n_{\text{bg}}) < 0.68\sigma_n$, it was accepted that $n_{\text{th}} - f_{\text{sh}} \cdot n_{\text{bg}} = 0.34\sigma_n$.

The distribution of relative errors of measurements of the thyroid radioactivity is depicted in Figure 7.3, and its characteristics such as the mean, median, and 5% and 95% percentiles are presented in Table 7.2.

For all the data set, the mean relative error is 0.33, which is essentially higher than the values of the formal instrument errors given in the technical documentation

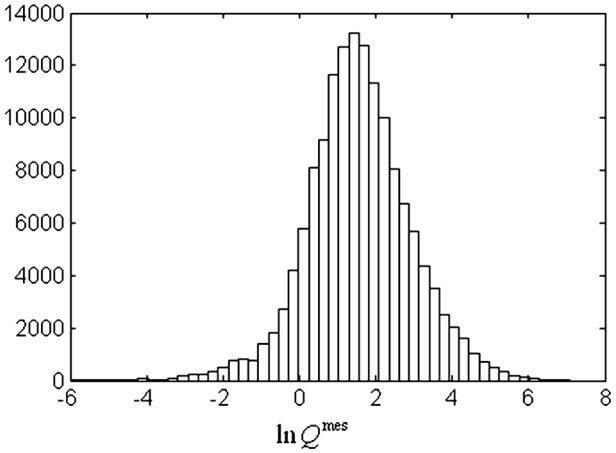


**Fig. 7.2:** Histogram of the estimates on content of radioiodine in the thyroid in the logarithmic scale (Likhtarov et al., 2015).

**Table 7.2:** Characteristics of the distribution of relative errors for the thyroid radioactivity in dependence on the type of device.

| Measurements by device's types | Percent of all measurements | Relative error AM | Median | 5th–95th percentiles |
|---|---|---|---|---|
| All measurements | 100% | 0.33 | 0.26 | 0.10–0.61 |
| Measurements performed by spectrometric devices | 36% | 0.27 | 0.22 | 0.09–0.54 |
| Measurements performed by radiometers | 64% | 0.37 | 0.29 | 0.20–0.66 |

of the measuring devices. The majority of values (90%) of the relative errors of the direct measurements of thyroid radioactivity being calculated according to (7.13) and (7.15)–(7.17) are in the range 0.1–0.6 (see Figure 7.3).

The complex nature of distribution shown in Figure 7.3 is explained by the combination inside a single array of measurements made by different types of instruments (spectrometers and integrated radiometers). Radiometers are the less accurate devices with relative error for them beginning with magnitudes of 0.2, while spectrometers show a significant number of measurements with less error. For measurements made with both types of the devices, there is an important characteristic expressed by significant right asymmetry of the distributions of relative errors. The analysis demonstrates that the main component of the relative errors with values above 0.5 is the $\sigma_n$ being the measurement error of the net signal (i.e., the measurement error of difference between thyroid signal and background signal), which reaches significant values in relation to the useful signal, is close to the background signal. It should be noted that a small net signal testifies about a negligible content of radioactivity $^{131}$I in the thyroid

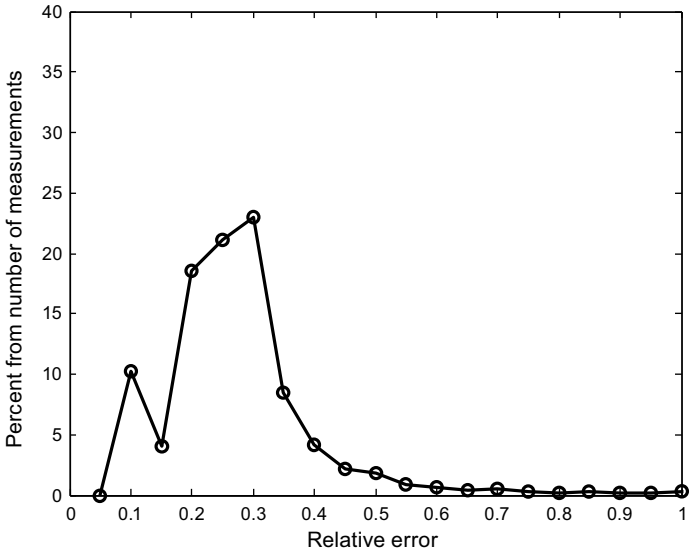


**Fig. 7.3:** Distribution of the relative measurement errors on the content of radioiodine in the thyroid (Likhtarov et al., 2015).

and respectively about a small dose of the thyroid irradiation. Thus, the large relative errors correspond to low absolute values of radioactivity $^{131}$I in the thyroid and consequently low exposure doses to the thyroid. The errors of the thyroid measurements for subjects who received significant doses are at the left side of the distribution shown in Figure 7.3.

## 7.2 A model of absorbed thyroid dose with classical additive and Berkson multiplicative errors

Estimation of absorbed doses of internal thyroid exposure for residents of Ukrainian regions suffered from radioactive $^{131}$I exposure after the accident on the Chornobyl nuclear power plant is often a complicated process involving mathematical modeling. First it is necessary to estimate the dynamics of $^{131}$I fallouts over the area, then using the model of radioiodine transfer along ecological chain to estimate its content in milk and other food, then based on the interview to determine the amount and types of foods consumed by the population, and only then to estimate the ecological dose (Likhtarov et al., 2015).

Reconstruction of individual exposure doses for people who resided (or still living) in the contaminated areas requires detailed information concerning their routine of behavior: location and environmental conditions, diet, preventive, and prophylactic measures. A framework for computing individual thyroid instrumental doses is pre-

sented in Figure 7.4. In this scheme, there are two types of the examined doses: the instrumental absorbed dose taking into account the data of direct thyroid radioactivity measurements, and ecological absorbed dose based only on the ecological model of radioiodine transportation.

As seen from the figure, the individual instrumental absorbed dose of internal thyroid exposure is computed using primary dosimetry, individual and ecological data (Kaidanovsky and Dolgirev, 1997), individualized thyroid masses (Likhtarov et al., 2013b), biokinetic radioiodine transportation models, and models of atmospheric transportation of radioactivity (Talerko, 2005a, 2005b).

According to Likhtarov et al. (2014), the measured individual instrumental absorbed thyroid dose for the $i$th person can be represented as

$$D_i^{\text{mes}} = \frac{f_i^{\text{mes}} Q_i^{\text{mes}}}{M_i^{\text{mes}}} , \tag{7.18}$$

where $M_i^{\text{mes}}$ is the measured thyroid mass, $Q_i^{\text{mes}}$ is the measured $^{131}$I radioactivity in the thyroid, $f_i^{\text{mes}}$ is a multiplier derived from the ecological model of radioactivity transition along the links of a food chain.

Ecological coefficient $f_i^{\text{mes}}$ includes the error of Berkson type (Likhtarov et al., 2014). Denote the factor with Berkson error as $\frac{f_i^{\text{mes}}}{M_i^{\text{mes}}} = F_i^{\text{mes}}$. Then relation (7.18) takes the form

$$D_i^{\text{mes}} = F_i^{\text{mes}} Q_i^{\text{mes}} . \tag{7.19}$$

The unknown true dose $D_i^{\text{tr}}$ is decomposed as

$$D_i^{\text{tr}} = F_i^{\text{tr}} Q_i^{\text{tr}} . \tag{7.20}$$

The connection between $F_i^{\text{tr}}$ and $F_i^{\text{mes}}$ is determined by Berkson multiplicative error:

$$F_i^{\text{tr}} = F_i^{\text{mes}} \cdot \delta_{F,i} , \quad \mathbf{E}\delta_{F,i} = 1 , \quad \ln \delta_{F,i} \sim \mathrm{N}\left( -\frac{\sigma_{F,i}^2}{2}, \sigma_{F,i}^2 \right) . \tag{7.21}$$

Here $F_i^{\text{mes}}$ and $\delta_{F,i}$ are stochastically independent, and $\sigma_{F,i}^2$ is the variance of $\ln \delta_{F,i}$. Further the values $\sigma_{F,i}^2$ are assumed to be known. Values $F_i^{\text{mes}}$ and $\sigma_{F,i}^2$ can be obtained by the Monte Carlo procedure described in Likhtarov et al. (2014).

According to (7.14), the individual measured radioactivity in thyroid $Q_i^{\text{mes}}$ can be written as

$$Q_i^{\text{mes}} = Q_i^{\text{tr}} + \sigma_{Q,i}^{\text{mes}} \gamma_i , \quad i = 1, \ldots, N . \tag{7.22}$$

Here $\gamma_1, \ldots, \gamma_N$ are independent standard normal variables and $\sigma_{Q,i}^{\text{mes}}$ are individual standard deviations of errors in direct measurements of thyroid radioactivity, which are determined according to (7.13). The quantities $\sigma_{Q,i}^{\text{mes}}$ and $Q_i^{\text{tr}}$ are independent random variables.

Substituting (7.22) to (7.20) and denoting $\overline{D}_i^{\text{tr}} = F_i^{\text{mes}} Q_i^{\text{tr}}$, we get

$$D_i^{\text{mes}} = F_i^{\text{mes}} Q_i^{\text{mes}} = F_i^{\text{mes}} (Q_i^{\text{tr}} + \sigma_{Q,i}^{\text{mes}} \gamma_i) = F_i^{\text{mes}} \cdot Q_i^{\text{tr}} + F_i^{\text{mes}} \sigma_{Q,i}^{\text{mes}} \gamma_i . \tag{7.23}$$

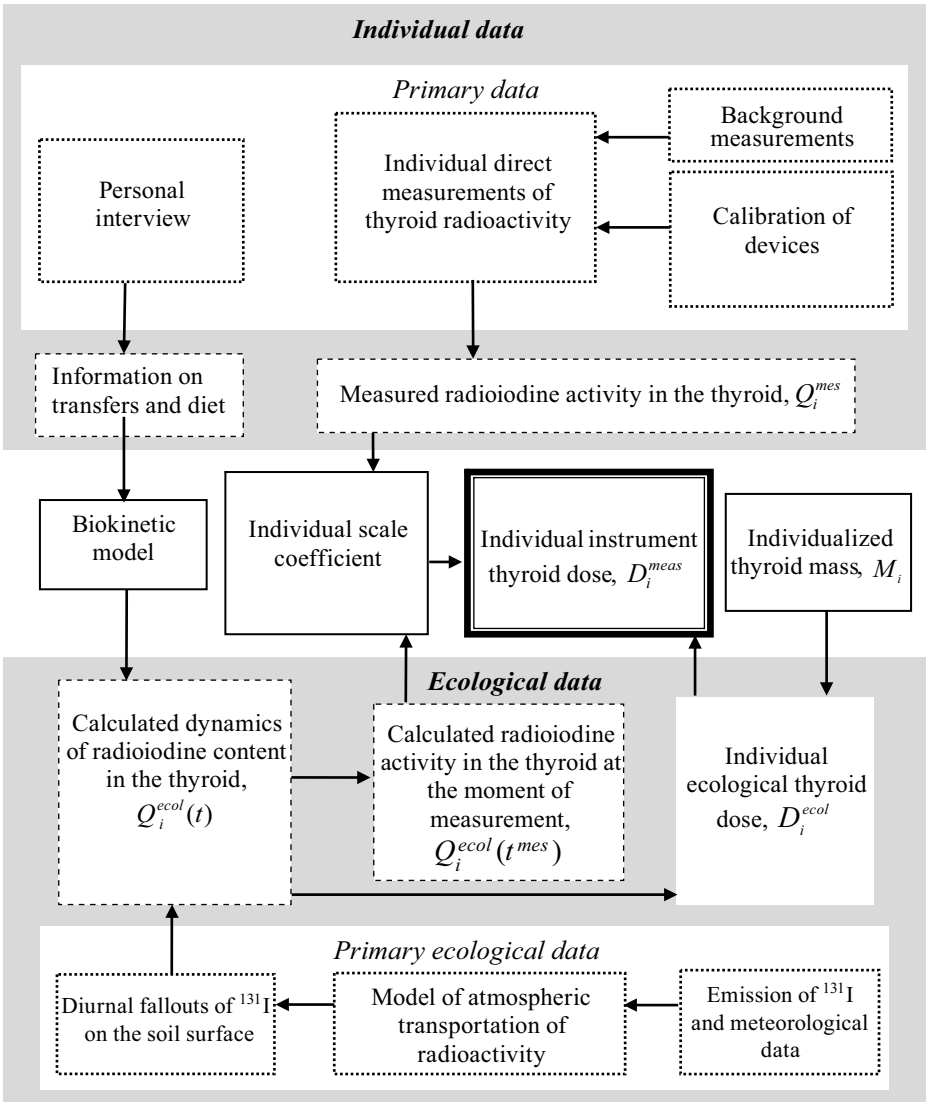


**Fig. 7.4:** Framework for computing individual instrumental doses of internal thyroid exposure for subjects under direct measurements of thyroid radioactivity in May–June 1986

Random variables $\{\delta_i, i \geq 1\}$, $\{\gamma_i, i \geq 1\}$ and random vectors $\{(F_i^{\text{mes}}, Q_i^{\text{tr}}), i \geq 1\}$ are jointly independent, but $F_i^{\text{mes}}$ and $Q_i^{\text{tr}}$ can be correlated. Introduce notations $\sigma_i$ =

$F_i^{\text{mes}}\sigma_{Q,i}^{\text{mes}}$ and $\overline{D}_i^{\text{tr}} = F_i^{\text{mes}} \cdot Q_i^{\text{tr}}$. Then (7.20)–(7.23) takes the form

$$D_i^{\text{mes}} = \overline{D}_i^{\text{tr}} + \sigma_i \gamma_i \,. \tag{7.24}$$

$$D_i^{\text{tr}} = \overline{D}_i^{\text{tr}} \delta_{F,i} \tag{7.25}$$

In fact, (7.24) and (7.25) constitute a model of dose observations with the classical additive error and Berkson multiplicative error.

It is easy to show that $\mathbf{E}(\overline{D}_i^{\text{tr}} | D_i^{\text{mes}}) = \mathbf{E}(D_i^{\text{tr}} | D_i^{\text{mes}})$.

## 7.3 Methods of risk estimation under classical additive and Berkson multiplicative errors in dose

In spite of the fact that Regression Calibration and the full maximum likelihood (FML) method for the classical *multiplicative* errors showed good results (see Section 6.5), for the classical *additive* errors in doses they proved to be inefficient (see Appendix D). Authors are inclined to relate this phenomenon with "unnatural" combination of the normal law of the classical dose errors in (7.23) and the lognormal law of population (see Figure 6.3). Therefore, the authors elaborated more appropriate methods of risk estimation which are presented in this section.

### 7.3.1 Corrected score method

Let the total incidence rate be given by (6.5). In the absence of dose errors, the likelihood function takes the form (see Example 7.4):

$$L = \left(\frac{\lambda}{1+\lambda}\right)^Y \left(\frac{1}{1+\lambda}\right)^{1-Y} = \frac{\lambda^Y}{1+\lambda}\,, \tag{7.26}$$

or

$$\ln L = Y \ln \lambda - \ln(1+\lambda)\,. \tag{7.27}$$

Then for unknown vector parameter $\theta = (\lambda_0, \beta)^{\text{T}}$, the score function is

$$S_{\text{ML}} = \frac{Y}{\lambda} \cdot \frac{\partial \lambda}{\partial \theta} - \frac{1}{1+\lambda} \cdot \frac{\partial \lambda}{\partial \theta}\,. \tag{7.28}$$

Notice that $\frac{\partial \lambda}{\partial \theta}$ is a linear function in $D^{\text{tr}}$. In order to apply the corrected score method and construct the estimating function depending on observations $(Y, D^{\text{mes}})$, we get rid of the denominator in (7.28) and consider the following function:

$$\tilde{S}_{\text{ML}} = \lambda(1+\lambda) S_{\text{ML}} = Y(1+\lambda)\frac{\partial \lambda}{\partial \theta} - \lambda \frac{\partial \lambda}{\partial \theta}\,. \tag{7.29}$$

The estimating function $\tilde{S}_{\mathrm{ML}}$ is unbiased, because it has zero expectation at the true point:

$$\begin{aligned}\mathbf{E}_\theta \tilde{S}_{\mathrm{ML}}(\theta) &= \mathbf{E}_\theta \left[ \left( Y - \frac{\lambda}{1+\lambda} \right) (1+\lambda) \frac{\partial \lambda}{\partial \theta} \right] = \\ &= \mathbf{E}\mathbf{E}_\theta \left[ \left( Y - \frac{\lambda}{1+\lambda} \right) (1+\lambda) \frac{\partial \lambda}{\partial \theta} \,\middle|\, D^{\mathrm{tr}} \right] = \\ &= \mathbf{E} \left[ (1+\lambda) \frac{\partial \lambda}{\partial \theta} \mathbf{E}_\theta \left[ \left( Y - \frac{\lambda}{1+\lambda} \right) \,\middle|\, D^{\mathrm{tr}} \right] \right] = 0 \,.\end{aligned}$$

A new estimating function $\tilde{S}_C(Y, D^{\mathrm{mes}})$ should satisfy the relation

$$\mathbf{E}[\tilde{S}_C(Y, D^{\mathrm{mes}}) | Y, D^{\mathrm{tr}}] = \tilde{S}_{\mathrm{ML}}(Y, D^{\mathrm{tr}}) = Y(1+\lambda) \frac{\partial \lambda}{\partial \theta} - \lambda \frac{\partial \lambda}{\partial \theta} \,, \tag{7.30}$$

for all $\theta$. For this purpose, the two deconvolution problems have to be solved:

$$\begin{aligned}\mathbf{E}[h_1(D^{\mathrm{mes}}) | D^{\mathrm{tr}}] &= \lambda \frac{\partial \lambda}{\partial \theta} \,, \\ \mathbf{E}[h_2(D^{\mathrm{mes}}) | D^{\mathrm{tr}}] &= \frac{\partial \lambda}{\partial \theta} \,,\end{aligned} \tag{7.31}$$

i.e., we have to find corresponding functions $h_1$ and $h_2$ and set

$$\tilde{S}_C = Y(h_1 + h_2) - h_1 = (Y-1)h_1 + Y h_2 \,. \tag{7.32}$$

Since $\frac{\partial \lambda}{\partial \theta}$ is linear in $D^{\mathrm{tr}}$, then $h_2 = \frac{\partial \lambda}{\partial \theta}|_{D^{\mathrm{tr}} = D^{\mathrm{mes}}} = \frac{\partial \lambda}{\partial \theta}(D^{\mathrm{mes}})$.

Introduce notation for coefficients in the gradient:

$$\frac{\partial \lambda}{\partial \theta} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} \beta \\ \lambda_0 \end{pmatrix} D^{\mathrm{tr}} = A + B D^{\mathrm{tr}} \,, \tag{7.33}$$

then

$$\begin{aligned} h_1 &= \lambda_0 A + D^{\mathrm{mes}}(\lambda_0 \beta A + \lambda_0 B) + ((D^{\mathrm{mes}})^2 - \sigma^2)\lambda_0 \beta B \,, \\ h_2 &= A + B D^{\mathrm{mes}} \,. \end{aligned} \tag{7.34}$$

The corrected score estimator $\hat{\theta}_N$ is found from equation

$$\sum_{i=1}^{N} \tilde{S}_C(Y_i, D_i^{\mathrm{mes}};\, \theta_N) = 0 \,. \tag{7.35}$$

The estimating function $\tilde{S}_C$ is unbiased as well, because

$$\mathbf{E}_\theta \tilde{S}_C = \mathbf{E}_\theta \mathbf{E}[\tilde{S}_C | Y, D^{\mathrm{tr}}] = \mathbf{E}_\theta \tilde{S}_{\mathrm{ML}}(Y, D^{\mathrm{mes}}) = 0 \,. \tag{7.36}$$

Therefore, as $N \to \infty$, the estimator $\hat{\theta}_N$ is strong consistent, i.e., almost surely $\hat{\theta}_N \to \theta$, as $N \to \infty$, where $\theta$ is the true vector parameter. The estimator is asymptotically

normal, namely $\sqrt{N}(\hat{\theta}_N - \theta) \xrightarrow{d} \mathrm{N}(0, \Sigma)$, where the asymptotic covariance matrix $\Sigma$ can be found by the sandwich formula (see Appendix A2):

$$\Sigma = U^{-1} V U^{-\mathrm{T}}, \quad U = \mathbf{E}\left(-\frac{\partial \tilde{S}_C}{\partial \theta^{\mathrm{T}}}\right), \quad V = \mathbf{E}\tilde{S}_C \tilde{S}_C^{\mathrm{T}}. \tag{7.37}$$

The matrices U and V can be estimated consistently by the formulas

$$\begin{aligned} \hat{U}_N &= \frac{1}{N}\sum_{i=1}^{N}\left(-\frac{\partial \tilde{S}_C}{\partial \theta^{\mathrm{T}}}(Y_i, D_i^{\mathrm{mes}};\, \hat{\theta})\right) = \\ &= \frac{1}{N}\sum_{i=1}^{N}\left((1-Y_i)\frac{\partial h_1(D_i^{\mathrm{mes}};\, \hat{\theta})}{\partial \theta^{\mathrm{T}}} - Y_i \frac{\partial h_2(D_i^{\mathrm{mes}};\, \hat{\theta})}{\partial \theta^{\mathrm{T}}}\right), \\ \hat{V}_N &= \frac{1}{N}\sum_{i=1}^{N}\left((Y_i - 1)h_1 + Y_i h_2\right)\left((Y_i - 1)h_1^{\mathrm{T}} + Y_i h_2^{\mathrm{T}}\right). \end{aligned} \tag{7.38}$$

In the latter equality $h_k = h_k(D_i^{\mathrm{mes}};\, \hat{\theta}_N)$, $k = 1, 2$. Then the estimator for the matrix $\Sigma$ is $\hat{\Sigma}_N = \hat{U}_N^{-1}\hat{V}_N\hat{U}_N^{-\mathrm{T}}$, and approximate relation holds: $\hat{\theta}_N - \theta \approx \frac{1}{\sqrt{N}}\mathrm{N}(0, \hat{\Sigma}_N)$. This makes it possible to construct the asymptotic confidence region for $\theta$.

### 7.3.2 Ordinary SIMEX and efficient SIMEX estimates

The estimate obtained by the *SIMEX* method is randomized, i.e., it is a random function of observations. Such a method of risk estimation was used by Kopecky et al. (2006), but in Masiuk et al. (2016) an approach which can more accurately take into account the structure of measured doses was proposed. A similar method was used in the presence of the classical *multiplicative* error (see Section 6.5.4).

In order to take into consideration the classical additive error in exposure doses, the following algorithm for Ordinary *SIMEX* is proposed:

(1) Choose a natural number $B \geq 2$. It is necessary that $B$ be large enough, for example, $B = 100$. Generate random perturbations for the logarithm of activities

$$U_{b,i}^{*} \sim \mathrm{N}(0, \sigma_i^2), \quad b = 1, \ldots, B, \quad i = 1, \ldots, N. \tag{7.39}$$

The perturbations are generated so that the condition $\sum_{b=1}^{B} U_{b,i}^{*} = 0$, $i = 1, \ldots, N$, holds true, which provides less spread and less deviation of the estimates (see Appendix B).

(2) Choose a set, for instance $\Lambda = \{0;\ 0.2;\ 0.4;\ 0.6\}$.

(3) Using the perturbed activities, compute the perturbed doses for each $\kappa \in \Lambda$:

$$D_{b,i}^{*} = D_i^{\mathrm{mes}} + \sqrt{\kappa} U_{b,i}^{*}, \kappa \in \Lambda. \tag{7.40}$$

(4) Compute the ordinary (naive) estimates $\hat{\lambda}^*_{0b}(\kappa)$ and $\hat{\beta}^*_b(\kappa)$ for $\kappa = 0; 0.2; 0.4; 0.6$ and average in $b$:

$$\begin{aligned} \hat{\lambda}^*_0(\kappa) &= \frac{1}{B}\sum_{b=1}^{B} \hat{\lambda}^*_{0b}(\kappa)\,, \\ \widehat{EAR}^*(\kappa) &= \frac{1}{B}\sum_{b=1}^{B} \hat{\lambda}^*_{0b}(\kappa)\hat{\beta}^*_b(\kappa),\ \kappa \in \Lambda\,. \end{aligned} \tag{7.41}$$

Extrapolate numerically the functions $\hat{\lambda}^*_0(\kappa)$ and $\widehat{EAR}^*(\kappa)$ to the point $\kappa = -1$ and finally get the *SIMEX* estimates for the parameters $\lambda_0$ and $EAR = \lambda_0\beta$. In extrapolation, we approximate $\hat{\lambda}^*_0(\kappa)$ and $\widehat{EAR}^*(\kappa)$ with quadratic polynomial. Such a choice of extrapolate function is the simplest one, and it allows to express the estimates explicitly through $\hat{\lambda}^*_0(\kappa)$ and $\widehat{EAR}^*(\kappa)$, see (6.53).

In Kukush et al. (2011), the "Efficient SIMEX estimator" of the risk parameters of the model with multiplicative error was derived as an alternative to the Ordinary *SIMEX*. It differed in the way that $D_i^{\text{mes}}$ is perturbed only in case $Y_i = 1$. In Masiuk et al. (2016), this idea in the model with additive errors was developed. In case of the Efficient *SIMEX* method, the system of equations for estimation the model parameters takes the form:

$$\begin{cases} \sum_{i=1}^N (1-Y_i)\left(1+\hat{\beta}^*_b(\kappa)D_i^{\text{mes}}\right) = \sum_{i=1}^N \frac{Y_i}{\hat{\lambda}^*_{0,b}(\kappa)}, \\ \sum_{i=1}^N (1-Y_i) = \frac{1}{\hat{\lambda}^*_{0,b}(\kappa)}\sum_{i=1}^N \frac{Y_i}{\left(1+\hat{\beta}^*_b(\Lambda)\right)\max(0,D^*_{b,i}(\kappa))}\,. \end{cases} \tag{7.42}$$

For significant perturbations, the modified dose $D^*_{b,i}(\kappa) = D_i^{\text{mes}} + \sqrt{\kappa}U^*_{b,i}$, $\kappa \in \Lambda$ may be negative, which may break down the estimation procedure. Therefore, the negative doses are changed to zeros, i.e., $\max(0, D^*_{b,i}(\kappa))$ is used instead of $D^*_{b,i}(\kappa)$.

### 7.3.3 New regression calibration

Because of the "unnatural" combination of the normal law of dose errors and the lognormal law of population $\overline{D}^{\text{tr}}$, the traditional regression calibration, including the linear and parametric ones, did not give an acceptable result (see Appendix D). Then new regression calibration (NRC) was developed. The idea of the method is as follows: the additive normal error in doses is replaced with the multiplicative lognormal one, but with nearly the same conditional variance (Masiuk et al., 2016).

Denote the lognormal error as $\delta_{L,i}$, $\log(\delta_{L,i}) \sim \text{N}(0, \sigma^2_{L,i})$. Equating the variance of the multiplicative error $\delta_{L,i}$ to the relative variance of the dose error $\sigma_i^2/(\overline{D}_i^{\text{tr}})^2$ and replacing the unknown dose $\overline{D}_i^{\text{tr}}$ with the measured one $D_i^{\text{mes}}$, we obtain an expression

for the parameter $\sigma^2_{L,i}$:

$$\mathbf{D}(\delta_{L,i}) = \exp(2\sigma^2_{L,i}) - \exp(\sigma^2_{L,i}) = \left(\frac{\sigma_i}{D_i^{\text{mes}}}\right)^2 . \tag{7.43}$$

From here,

$$\sigma^2_{L,i} = \ln\left(\frac{1}{2} + \sqrt{\frac{1}{4} + \left(\frac{\sigma_i}{D_i^{\text{mes}}}\right)^2}\right) . \tag{7.44}$$

Unlike Masiuk et al. (2016) in this version of NRC we are correcting $D_i^{\text{mes}}$ to improve the estimates for large classical error:

$$D_{L,i}^{\text{mes}} = D_i^{\text{mes}} \exp\left(-\frac{\sigma^2_{L,i}}{2}\right) . \tag{7.45}$$

After this, the calibration is carried out by the way described in Kukush et al. (2011):

$$\mathbf{E}(\overline{D}_i^{\text{tr}} \Big| D_i^{\text{mes}}) \approx \exp\left(\frac{\sigma^2_{\overline{D}^{\text{tr}}} \log D_{L,i}^{\text{mes}} + \sigma^2_{L,i}\mu_{\overline{D}^{\text{tr}}} + \frac{\sigma^2_{\overline{D}^{\text{tr}}}\sigma^2_{L,i}}{2}}{\sigma^2_{\overline{D}^{\text{tr}}} + \sigma^2_{L,i}}\right) . \tag{7.46}$$

The parameters $\mu_{\overline{D}^{\text{tr}}}$ and $\sigma^2_{\overline{D}^{\text{tr}}}$ are estimated by formulas from Koroliuk et al. (1985):

$$\hat{\mu}_{\overline{D}^{\text{tr}}} = \log\left(\frac{\left(\hat{m}_{\overline{D}^{\text{tr}}}\right)^2}{\sqrt{\hat{v}_{\overline{D}^{\text{tr}}} + \left(\hat{m}_{\overline{D}^{\text{tr}}}\right)^2}}\right), \hat{\sigma}^2_{\overline{D}^{\text{tr}}} = \log\left(\frac{\hat{v}_{\overline{D}^{\text{tr}}}}{\left(\hat{m}_{\overline{D}^{\text{tr}}}\right)^2} + 1\right) , \tag{7.47}$$

where

$$\begin{aligned} \hat{m}_{\overline{D}^{\text{tr}}} &= \frac{1}{N}\sum_{i=1}^{N} D_i^{\text{mes}} , \\ \hat{v}_{\overline{D}^{\text{tr}}} &= \frac{1}{N-1}\sum_{i=1}^{N}\left(D_i^{\text{mes}} - \hat{m}_{\overline{D}^{\text{tr}}}\right)^2 - \frac{1}{N}\sum_{i=1}^{N}\sigma_i^2 . \end{aligned} \tag{7.48}$$

### 7.3.4 Taking into account Berkson error

The corrected score method takes into account only the classical error but not the Berkson one. At the same time, the ordinary and efficient *SIMEX* methods and the NRC allow to take into consideration the presence of both types of errors.

In order to take into account Berkson multiplicative error when the ordinary *SIMEX* method is used, the estimates $\hat{\lambda}^*_{0b}(\kappa)$ and $\hat{\beta}^*_b(\kappa)$ are also computed by the FML method described in Section 6.4.

In Masiuk et al. (2016), it is shown that in order to take into account Berkson multiplicative error in the efficient SIMEX method we must replace equations (7.42) with (7.49)

$$\begin{cases} \sum_{i=1}^{N} \frac{Y_i}{m(D^*_{b,i}(\kappa);\lambda_0,\beta,\sigma^2_{F,i})} = N\,, \\ \sum_{i=1}^{N} \frac{Y_i \cdot D^*_{b,i}(\kappa)}{m(D^*_{b,i}(\kappa);\lambda_0,\beta,\sigma^2_{F,i})} = \frac{1}{\hat{\lambda}^*_{0,b}(\kappa)} \sum_{i=1}^{N} D_i^{\mathrm{mes}}\,, \end{cases} \tag{7.49}$$

where $m(D^*_{b,i}(\kappa);\lambda_0,\beta,\sigma^2_{F,i}) = P[Y = 1|\, \overline{D}_i^{\mathrm{tr}} = D^*_{b,i}(\kappa)]$.

The perturbed dose $D^*_{b,i}(\kappa) = D_i^{\mathrm{mes}} + \sqrt{\kappa}U^*_{b,i}$, $\kappa \in \Lambda$ can be negative. To prevent the effect of negative doses on the naive estimates, the perturbed doses are censored by zero from the left.

When the NRC was applied to account for the classical additive errors in thyroid doses, the latter were precalibrated using (7.44)–(7.48). Further, in order to take into account Berkson multiplicative dose error, the FML method described in Section 6.4 was applied.

### 7.3.5 Stochastic simulation of classical additive and Berkson multiplicative errors

A simulation was carried out based on the epidemiological studies of thyroid cancer incidence in Ukraine (Likhtarov et al., 2006a, 2006b, and 2014; Tronko et al., 2006). The absorbed doses of internal thyroid exposure correspond to doses for real subpopulation of children and adolescents aged from 0 to 18 (totally 13,204 subjects) from settlements of Zhytomyr, Kyiv, and Chernihiv Oblasts of Ukraine, where the direct measurements of thyroid radioactivity were being performed in May and June, 1986. In simulation of thyroid cancer total incidence rate at a fixed time interval, the absolute risk model (6.5) was used, with parameters close to the estimates obtained during epidemiological studies of thyroid cancer in Ukraine (Likhtarov et al., 2006a; Tronko et al., 2006), namely:

$$\begin{aligned} \lambda_0 &= 2 \times 10^{-4}\, \frac{\text{cases}}{\text{person years}}\,, \\ EAR &= 5 \times 10^{-4}\, \frac{\text{cases}}{\text{Gy} \cdot (\text{person years})}\,. \end{aligned} \tag{7.50}$$

In the framework of the study, there were modeled measured doses (7.24) and (7.25) observed with the classical additive normal error and Berkson multiplicative error. The size of classical error was determined by a constant $\delta_Q = \frac{\sigma^{\mathrm{mes}}_{Q,i}}{Q_i^{\mathrm{tr}}}$, for all $1 \le i \le 13{,}204$, and varied from 0.2 to 1. The size of Berkson error was set so that the geometric standard deviation $\mathrm{GSD}_F = \exp(\sigma_F)$ of the parameter $F^{\mathrm{tr}}$, with observed $F^{\mathrm{mes}}$, took the values: 1 (no error), 1.5, 2, 3, 5, and 8, for each $1 \le i \le 13{,}204$. All the listed values are realistic (Likhtarov et al., 2013a).

*Simulation study is performed in four steps:*

(1) Initial doses $\overline{D}_i^{\mathrm{tr}}$ are taken from the real thyroid doses of children and adolescents internally exposed to $^{131}$I in 1986 (see Figure 6.3).

(2) True dose values $D_i^{\mathrm{tr}}$ are generated for the cohort by using $\overline{D}_i^{\mathrm{tr}}$ and taking into account the uncertainty levels $\mathrm{GSD}_F$ given in the first column of Tables 7.3 and 7.4.
(3) Using the data from step 2, as well as the model in equations (6.4)–(6.6) with the parameter values $\lambda_0$ and *EAR* in (7.49), a disease vector is generated.
(4) Initial doses $\overline{D}_i^{\mathrm{tr}}$ were perturbed, and thus, the measured doses $D_i^{\mathrm{mes}}$ were generated according to equation (7.24), with the error standard deviation $\sigma_i = \delta_Q \cdot \overline{D}_i^{\mathrm{tr}}$, where $\delta_Q$ enters the second column of Tables 7.3 and 7.4. As a result, we obtain an observation model with classical additive and Berkson multiplicative errors in doses.

It should be noted that under sizable additive errors, some of the generated measured doses $D_i^{\mathrm{mes}}$ could be negative. In the latter case, the doses were censored, i.e., negative dose values were substituted with certain small positive number. In the simulation, 1000 data sets were generated.

Based on the measured doses $D_i^{\mathrm{mes}}$, the information of measurement errors $\mathrm{GSD}_F$ and $\delta_Q$, and the disease vector generated in step 3, the parameter values $\lambda_0$ and *EAR* are estimated by the following methods:
(1) The naive method (using the package *EPICURE*).
(2) The NRC.
(3) Ordinary *SIMEX*.
(4) Efficient *SIMEX*.
(5) The corrected score method.

Each estimate was computed for 1000 realizations of doses and cases. Then the median and the deviance interval (95% DI) were calculated, based on the 2.5% and 97.5% quantiles of estimates over 1000 realizations. In the cases where the 2.5 percent quantile occurred to be negative, its value was replaced with zero by mere physical reasons (because the risk coefficients cannot be negative).

### 7.3.6 Discussion of results

#### Naive method

The simulation results are given in Tables 7.3 and 7.4. At the same time, Figures 7.5 and 7.6 show the risk estimates behavior for the case $\mathrm{GSD}_F = 1$, i.e., when Berkson error is absent. Analysis of the simulation results shows that the naive estimates are biased both in the case of the classical additive measurement error and the case of Berkson multiplicative error. As this, the estimates of excess absolute risk *EAR* are underestimated, while the estimates of background incidence rate $\lambda_0$ are overestimated. The bias of the naive estimate increases, as the variance of the classical or Berkson error grows. The impact of the classical error on the risk estimates is quite significant.

**Table 7.3:** Medians of estimates (q50%) and 95% deviance intervals (95% DI) of background incidence rate $\lambda_0 \times 10^4$ for various levels of classical additive and Berkson multiplicative errors.

| Error level | | Estimation method (model value of $\lambda_0 \times 10^4$ is 2.0) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Naive (EPICURE) | | NRC | | Ordinary *SIMEX* | | Efficient *SIMEX* | |
| $GSD_F$ | $\delta_Q$ | q50% | 95% DI | q50% | 95% DI | q50% | 95% DI | q50% | 95% DI |
| 1 | 0.0 | 1.95 | 0.99–3.08 | 1.97 | 1.00–2.99 | 1.91 | 0.65–3.06 | 1.95 | 0.99–3.08 |
| | 0.2 | 2.04 | 1.02–3.20 | 1.97 | 0.90–3.03 | 1.98 | 0.62–3.06 | 1.92 | 0.99–3.08 |
| | 0.4 | 2.23 | 1.19–3.42 | 2.11 | 1.00–3.26 | 2.28 | 1.22–3.56 | 1.87 | 0.46–3.24 |
| | 0.6 | 2.58 | 1.47–3.79 | 2.43 | 1.24–3.60 | 3.59 | 2.26–5.15 | 2.46 | 0.20–4.65 |
| | 0.8 | 2.91 | 1.80–4.13 | 2.68 | 1.56–3.90 | 4.75 | 3.51–5.88 | 3.52 | 0.77–6.28 |
| | 1 | 3.14 | 2.01–4.35 | 2.90 | 1.75–4.14 | 4.97 | 3.75–6.01 | 4.46 | 1.28–7.46 |
| 2 | 0.0 | 1.95 | 0.84–3.08 | 1.95 | 0.92–3.00 | 1.88 | 0.82–2.95 | 1.92 | 0.93–2.98 |
| | 0.2 | 2.06 | 1.21–3.49 | 1.96 | 0.91–3.02 | 1.93 | 0.78–3.01 | 1.92 | 0.95–3.00 |
| | 0.4 | 2.27 | 1.39–3.70 | 2.10 | 1.00–3.18 | 2.21 | 1.12–3.47 | 1.90 | 0.61–3.27 |
| | 0.6 | 2.61 | 1.70–4.02 | 2.42 | 1.14–3.49 | 2.90 | 1.80–4.48 | 2.40 | 0.07–4.47 |
| | 0.8 | 2.94 | 1.98–4.33 | 2.71 | 1.40–3.87 | 3.61 | 2.31–5.24 | 3.56 | 0.67–6.38 |
| | 1 | 3.18 | 2.21–4.55 | 2.90 | 1.56–4.09 | 4.03 | 2.69–5.58 | 4.47 | 1.57–7.70 |
| 3 | 0.0 | 2.01 | 0.96–3.23 | 1.95 | 0.94–3.05 | 1.91 | 0.75–2.88 | 1.94 | 0.86–3.09 |
| | 0.2 | 2.08 | 1.06–3.27 | 1.95 | 0.93–3.07 | 1.94 | 0.71–2.97 | 1.92 | 0.91–3.06 |
| | 0.4 | 2.26 | 1.24–3.46 | 2.07 | 0.94–3.22 | 2.18 | 0.96–3.22 | 1.94 | 0.51–3.48 |
| | 0.6 | 2.61 | 1.55–3.79 | 2.41 | 1.22–3.50 | 2.73 | 1.71–3.94 | 2.40 | 0.06–4.74 |
| | 0.8 | 2.93 | 1.87–4.11 | 2.70 | 1.44–3.80 | 3.31 | 2.14–4.58 | 3.55 | 0.59–6.11 |
| | 1 | 3.17 | 2.09–4.33 | 2.88 | 1.59–4.03 | 3.75 | 2.46–5.11 | 4.41 | 1.03–7.46 |
| 5 | 0.0 | 2.12 | 1.19–3.35 | 1.99 | 0.90–3.20 | 1.87 | 0.66–2.90 | 1.93 | 0.88–3.06 |
| | 0.2 | 2.17 | 1.07–3.33 | 1.98 | 0.86–3.12 | 1.90 | 0.61–2.98 | 1.94 | 0.87–3.08 |
| | 0.4 | 2.34 | 1.28–3.52 | 2.09 | 0.91–3.23 | 2.13 | 0.88–3.29 | 1.95 | 0.56–3.38 |
| | 0.6 | 2.65 | 1.50–3.81 | 2.38 | 1.17–3.55 | 2.58 | 1.56–3.77 | 2.38 | 0.19–4.59 |
| | 0.8 | 2.94 | 1.81–4.06 | 2.67 | 1.38–3.80 | 3.03 | 1.83–4.18 | 3.53 | 0.70–5.74 |
| | 1 | 3.14 | 2.01–4.24 | 2.85 | 1.51–3.99 | 3.38 | 2.09–4.62 | 4.36 | 1.21–7.19 |
| 8 | 0.0 | 2.23 | 1.23–3.37 | 1.98 | 0.95–3.11 | 1.85 | 0.76–2.97 | 1.94 | 0.86–3.13 |
| | 0.2 | 2.27 | 1.22–3.42 | 2.00 | 0.90–3.07 | 1.87 | 0.77–2.99 | 1.94 | 0.87–3.13 |
| | 0.4 | 2.43 | 1.36–3.58 | 2.09 | 1.02–3.17 | 2.07 | 1.00–3.13 | 1.91 | 0.44–3.43 |
| | 0.6 | 2.71 | 1.62–3.84 | 2.34 | 1.20–3.46 | 2.43 | 1.40–3.60 | 2.36 | 0.18–4.31 |
| | 0.8 | 2.96 | 1.87–4.04 | 2.59 | 1.38–3.74 | 2.79 | 1.66–4.01 | 3.27 | 0.38–5.63 |
| | 1 | 3.13 | 2.07–4.24 | 2.79 | 1.47–3.86 | 3.12 | 1.99–4.32 | 4.08 | 0.91–6.70 |

Namely, for $\delta_Q = 0.4$, the bias of *EAR* (to smaller side) and $\lambda_0$ (to larger side) is approximately 10%. Note that for sufficiently high variance of the classical error, the naive estimates may differ from the model values (i.e., from the true ones) up to several times. This effect is clearly seen in Figures 7.5 and 7.6. At the same time, the impact of Berkson error on the risk analysis results is significantly smaller. Namely, for $GSD_F \leq 2$, the impact of the latter error is negligible.

**Table 7.4:** Medians of estimates (q50%) and 95% deviance intervals (95% DI) of excess absolute risk $EAR \times 10^4$ for various levels of classical additive and multiplicative Berkson error.

| Error level | | Estimation method (model value of $EAR \times 10^4$ is 5.0) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Naive (EPICURE) | | NRC | | Ordinary *SIMEX* | | Efficient *SIMEX* | |
| $GSD_F$ | $\delta_Q$ | q50% | 95% DI | q50% | 95% DI | q50% | 95% DI | q50% | 95% DI |
| 1 | 0.0 | 5.01 | 3.00–7.04 | 5.00 | 3.07–6.74 | 5.09 | 3.09–7.21 | 5.01 | 3.00–7.04 |
| | 0.2 | 4.90 | 3.09–7.00 | 5.09 | 2.98–7.41 | 5.03 | 3.03–7.17 | 5.03 | 2.99–7.04 |
| | 0.4 | 4.59 | 2.83–6.63 | 5.21 | 3.02–7.66 | 4.51 | 2.62–6.51 | 5.10 | 2.60–7.61 |
| | 0.6 | 4.01 | 2.33–5.83 | 4.92 | 2.70–7.32 | 2.18 | 0.45–7.27 | 4.13 | 1.20–7.44 |
| | 0.8 | 3.41 | 1.85–5.11 | 4.40 | 2.31–6.85 | 0.86 | 0.00–1.28 | 2.33 | 0.00–6.19 |
| | 1 | 2.90 | 1.57–4.49 | 4.02 | 1.93–6.36 | 0.53 | 0.04–1.11 | 0.83 | 0.00–5.47 |
| 2 | 0.0 | 4.97 | 3.20–7.23 | 4.99 | 3.07–7.04 | 5.01 | 2.92–7.43 | 5.03 | 3.27–7.21 |
| | 0.2 | 4.81 | 3.22–7.21 | 5.22 | 3.19–7.37 | 4.92 | 2.79–7.25 | 5.05 | 3.21–7.24 |
| | 0.4 | 4.52 | 2.96–6.85 | 5.35 | 3.20–7.63 | 4.48 | 2.54–6.88 | 5.08 | 2.59–7.79 |
| | 0.6 | 3.89 | 2.48–6.10 | 4.97 | 2.89–7.23 | 3.30 | 1.52–5.36 | 4.29 | 1.11–8.35 |
| | 0.8 | 3.28 | 2.02–5.34 | 4.50 | 2.46–6.75 | 2.21 | 0.60–4.46 | 2.47 | 0.00–6.96 |
| | 1 | 2.83 | 1.69–4.72 | 4.12 | 2.02–6.30 | 1.46 | 0.21–3.46 | 0.92 | 0.00–5.01 |
| 3 | 0.0 | 4.85 | 2.88–6.64 | 5.02 | 2.87–7.14 | 5.10 | 2.87–7.47 | 5.07 | 3.20–7.37 |
| | 0.2 | 4.69 | 2.82–6.68 | 5.19 | 2.98–7.43 | 4.99 | 2.83–7.36 | 5.06 | 3.22–7.34 |
| | 0.4 | 4.41 | 2.58–6.31 | 5.38 | 2.96–7.92 | 4.58 | 2.63–7.07 | 5.06 | 2.63–7.98 |
| | 0.6 | 3.82 | 2.19–5.47 | 4.98 | 2.70–7.54 | 3.55 | 1.90–6.09 | 4.30 | 0.42–8.21 |
| | 0.8 | 3.24 | 1.77–4.77 | 4.47 | 2.33–7.03 | 2.62 | 1.18–4.96 | 2.36 | 0.00–7.13 |
| | 1 | 2.79 | 1.43–4.24 | 4.10 | 1.86–6.57 | 1.94 | 0.66–3.64 | 0.87 | 0.00–5.72 |
| 5 | 0.0 | 4.24 | 2.71–5.73 | 5.11 | 2.77–7.18 | 5.21 | 2.72–8.46 | 5.07 | 2.88–7.48 |
| | 0.2 | 4.26 | 2.47–6.17 | 5.32 | 2.89–7.56 | 5.10 | 2.69–8.50 | 5.04 | 2.83–7.52 |
| | 0.4 | 3.94 | 2.24–5.84 | 5.47 | 2.90–8.07 | 4.69 | 2.54–8.03 | 4.99 | 2.34–8.30 |
| | 0.6 | 3.42 | 1.84–5.17 | 5.06 | 2.39–7.78 | 3.77 | 2.00–6.59 | 4.34 | 0.13–8.53 |
| | 0.8 | 2.91 | 1.50–4.49 | 4.57 | 1.97–7.28 | 2.91 | 1.37–5.70 | 2.22 | 0.00–7.33 |
| | 1 | 2.51 | 1.22–3.97 | 4.17 | 1.76–6.93 | 2.28 | 0.92–4.62 | 0.72 | 0.00–6.00 |
| 8 | 0.0 | 3.62 | 1.99–5.10 | 5.22 | 2.51–7.67 | 5.24 | 2.58–9.08 | 5.01 | 2.48–7.78 |
| | 0.2 | 3.57 | 1.79–5.42 | 5.46 | 2.53–7.95 | 5.19 | 2.60–9.07 | 5.01 | 2.42–7.82 |
| | 0.4 | 3.30 | 1.67–5.09 | 5.68 | 2.48–8.48 | 4.78 | 2.30–8.27 | 5.01 | 2.34–8.36 |
| | 0.6 | 2.79 | 1.34–4.50 | 5.25 | 2.15–8.33 | 3.78 | 1.84–7.61 | 4.09 | 0.03–8.59 |
| | 0.8 | 2.38 | 1.04–3.90 | 4.67 | 1.88–7.85 | 3.03 | 1.35–6.06 | 2.28 | 0.00–7.80 |
| | 1 | 2.04 | 0.83–3.40 | 4.23 | 1.50–7.35 | 2.39 | 0.94–5.36 | 0.71 | 0.00–6.48 |

### New regression calibration and SIMEX

Although the parametric regression calibration introduced in Likhtarov et al. (2013) takes into account the shape of the distribution of $\overline{D}^{\mathrm{tr}}$, the estimates computed by this method are considerably biased, with underestimated background incidence rate and overestimated excess absolute risk (the results are shown in Appendix D). This is unexpected effect compared with simulation results from Kukush et al. (2011), where in case of multiplicative measurement errors in doses, the parametric estimates were

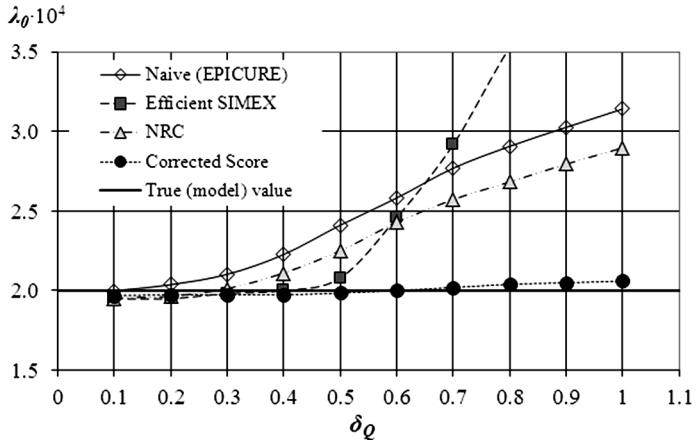


**Fig. 7.5:** Background incidence rate estimates for various relative additive errors of classical type in thyroid absorbed doses.

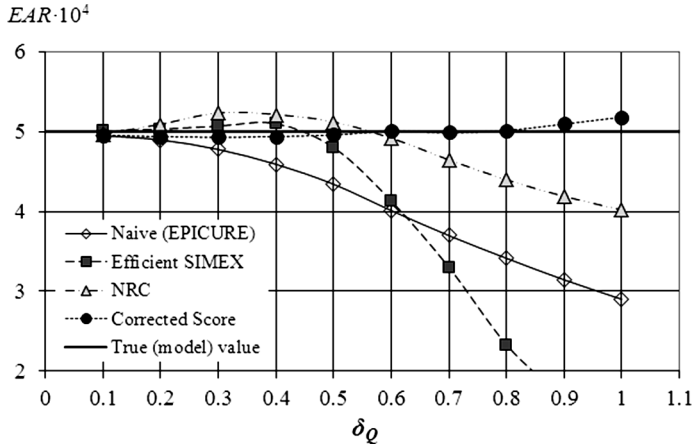


**Fig. 7.6:** Estimates of excess absolute risk for various relative additive errors of classical type in thyroid absorbed doses.

quite acceptable. It seems the reason for that lies in the combined structure of the normal measurement errors $\sigma_i \gamma_i$ and the lognormal distribution of $\overline{D}^{\mathrm{tr}}$, but we have no more definite explanation. Estimates obtained by the NRC are much more stable and less biased compared with the ones obtained by other methods of regression calibration (see Appendix D), and are quite satisfactory when the classical error in dose is not too large. Estimates of absolute risk model parameters obtained by the ordinary SIMEX method do not differ much from naive ones if the impact of the Berkson error is negligible. The efficient SIMEX method fits the model values only for relatively small classical errors. The estimates are satisfactory if $\delta_Q \leq 0.4$. If $\delta_Q \geq 0.6$, bias for the NRC

**Table 7.5:** Corrected score estimates of medians (q50%) and 95% deviance intervals (95% DI) of background incidence rate and excess absolute risk for various levels of classical additive errors. Model values of $\lambda_0 \times 10^4$ and $EAR \times 10^4$ is 2.0 and 5.0, respectively.

| Error level | Background incidence rate $\lambda_0 \times 10^4$ | | Excess absolute risk $EAR \times 10^4$ | |
|---|---|---|---|---|
| $\delta_Q$ | q50% | 95% DI | q50% | 95% DI |
| 0.0 | 1.93 | 0.00–3.71 | 5.03 | 2.34–8.53 |
| 0.2 | 1.95 | 0.00–3.77 | 4.97 | 2.38–8.80 |
| 0.4 | 1.98 | 0.00–3.89 | 4.93 | 2.24–9.16 |
| 0.6 | 1.93 | 0.00–4.12 | 4.98 | 2.00–9.95 |
| 0.8 | 1.90 | 0.00–4.34 | 5.02 | 1.66–11.1 |
| 1 | 1.98 | 0.00–4.64 | 5.02 | 1.28–12.7 |

method is less than one for the efficient SIMEX method. At the same time, those methods allow taking into account the presence of sizable Berkson errors in the radiation doses. Another advantage of the methods is as follows: they can be readily adapted to more complicated risk models.

### Influence of Berkson error

For moderate levels $\mathrm{GSD}_F \leq 2$, the effect of Berkson error on ultimate estimates is insignificant. But if $\mathrm{GSD}_F$ increases to 3 and more, then the influence of Berkson error becomes significant and should be taken into account. Simulation results showed that for the naive estimates, Berkson error as well as the classical error (but to a smaller extent) lead to underestimation of $EAR$ and overestimation of $\lambda_0$.

### Corrected score method

The least unbiased estimate (of all ones presented in this chapter) for regression model with the classical additive error in doses is the corrected score estimate, see Figures 7.5 and 7.6 and Table 7.5. In particular for $\mathrm{GSD}_F \leq 2$, the maximal bias of those estimates does not exceed 5% within the whole range of relative errors $0.1 \leq \delta_Q \leq 1$. The corrected score estimates have rather wide deviance intervals, because the bias correction leads to increasing the variability of estimates.

A disadvantage of this method is as follows: it ignores the presence of Berkson errors. Using this estimator, only classical error was taken into account. This yields biased estimates for large Berkson errors (Masiuk et al., 2016).

# A Elements of estimating equations theory

## A.1 Unbiased estimating equations: conditions for existence of solutions and for consistency of estimators

Before proceeding to stochastic estimating equations, we first consider deterministic equations.

### A.1.1 Lemma about solutions to nonrandom equations

We state two classical fixed-point theorems.

**Definition A.1.** For a mapping $f$, a point $x_0$ is called the *fixed point* if $f(x_0) = x_0$.

**Theorem A.2** (Banach fixed-point theorem). *Let $A$ be a closed set in $\mathbf{R}^n$, and $f$ be a contracting mapping from $A$ to $A$, i.e., such a mapping that for a fixed $\lambda < 1$, it holds that*

$$\|f(x_1) - f(x_2)\| \le \lambda \|x_1 - x_2\| . \tag{A.1}$$

*Then $f$ has a fixed point, and the point is unique.*

**Theorem A.3** (Brauer's theorem). *Let $\bar{B}(x_0, r)$ be a closed ball in $\mathbf{R}^n$, and $f\colon \bar{B}(x_0, r) \to \bar{B}(x_0, r)$ be a continuous mapping. Then $f$ has a fixed point.*

Remember some definitions and facts from the course on Calculus, see Burkill (1962).

**Definition A.4.** Let $A \subset \mathbf{R}^n$. A point $x_0 \in A$ is called the *interior point* of $A$ if a certain ball $B(x_0, r)$ is a part of $A$.

A set of all interior points of $A$ is denoted by $A^0$.

A set $K \subset \mathbf{R}^n$ is compact if, and only if, $K$ is closed and bounded.

Remember that in this book, all vectors are column ones.

**Definition A.5.** Let $A \subset \mathbf{R}^n$ and $x_0 \in A^0$. A vector function $f\colon A \to \mathbf{R}^m$ is called *differentiable* at a point $x_0$ if for some matrix $L$ of size $m \times n$, the following expansion holds:

$$f(x) = f(x_0) + L \cdot (x - x_0) + o(\|x - x_0\|) , \quad \text{as } x \to x_0 . \tag{A.2}$$

Then $L$ is a *derivative* (or *Jacobian matrix*) of the function $f$ at the point $x_0$.

Denote the derivative as $f'(x_0)$. Hereafter the notation $h(y) = o(\|y\|)$ means that $\|h(y)\|/\|y\| \to 0$, as $y \to 0$.

**Lemma A.6.** *Let $\Theta$ be a compact set in $\mathbf{R}^d$, $\beta_0 \in \Theta^0$, and $\{S_n, n \ge 1\}$ be a sequence of continuous functions from $\Theta$ to $\mathbf{R}^d$ being uniformly convergent to a function $S_\infty$. Assume also that the following conditions hold true.*

(1) $S_\infty(b) = 0$, $b \in \Theta$, *if, and only if,* $b = \beta_0$.
(2) *The function* $S_\infty$ *is differentiable at the point* $\beta_0$, *and its derivative*

$$\Phi = S'_\infty(\beta_0) \tag{A.3}$$

*is nonsingular matrix.*
*Then for all sufficiently large* $n$, *the equation*

$$S_n(b) = 0\,, \quad b \in \Theta \tag{A.4}$$

*has a solution. Denote by* $\hat\beta_n$ *an arbitrary solution to the equation (here* $\hat\beta_n$ *is well-defined for large* $n$*). Then*

$$\hat\beta_n \to \beta_0\,, \quad as \;\; n \to \infty\,. \tag{A.5}$$

*Proof. Existence of solution.* Introduce a function

$$f_n(b) = b - \Phi^{-1} \cdot S_n(b)\,, \quad b \in \Theta\,. \tag{A.6}$$

The function $S_\infty$ is differentiable at the point $\beta_0 \in \Theta^0$, i.e.,

$$S_\infty(b) = S_\infty(\beta_0) - S'_\infty(\beta_0)(b-\beta_0) + o(\|b-\beta_0\|) = \Phi \cdot (b-\beta_0) + o(\|b-\beta_0\|)\,, \tag{A.7}$$

as $b \to \beta_0$. Hence there exists such an $\varepsilon_1$ that $\bar B(\beta_0, \varepsilon_1) \subset \Theta$ and for all $b \in \bar B(\beta_0, \varepsilon_1)$, it holds that

$$\|S_\infty(b) - \Phi \cdot (b-\beta_0)\| \;\le \frac{\|b-\beta_0\|}{2\,\|\Phi^{-1}\|}\,. \tag{A.8}$$

Given the uniform convergence of the functions, we find a number $n_1$ such that for all $n \ge n_1$ and all $b \in \Theta$, it holds that

$$\|S_n(b) - S_\infty(b)\| \;\le \frac{\varepsilon_1}{2\,\|\Phi^{-1}\|}\,. \tag{A.9}$$

Then for $\|b-\beta_0\| \;\le \varepsilon_1$ and $n \ge n_1$ , the inequalities hold true:

$$\begin{aligned} \|f_n(b) - \beta_0\| &= \|\,\Phi^{-1}(\Phi(b-\beta_0) - S_n(b))\,\| \;\le \\ &\le \;\|\Phi^{-1}\| \cdot (\,\|S_\infty(b) - \Phi(b-\beta_0)\| + \|S_n(b) - S_\infty(b)\|\,)\,, \end{aligned} \tag{A.10}$$

$$\|f_n(b) - \beta_0\| \;\le\; \|\Phi^{-1}\| \left( \frac{\|b-\beta_0\|}{2\|\Phi^{-1}\|} + \frac{\varepsilon_1}{2\|\Phi^{-1}\|} \right) \le \varepsilon_1\,. \tag{A.11}$$

Thus, for $n \ge n_1$, we have $f_n(\bar B(\beta_0, \varepsilon_1) \subset \bar B(\beta_0, \varepsilon_1)$. By Theorem A.3, the function $f_n$ has a fixed point from the ball $\bar B(\beta_0, \varepsilon_1)$, and the point is a solution to equation (A.4).

*Convergence.* The function $S_\infty$ is continuous as a uniform limit of a sequence of continuous functions. Therefore, the minimum of function $\|S_\infty(b)\|$ is attained on any nonempty compact set.

Let us fix $\varepsilon > 0$ and show that for large enough $n$, it holds that

$$\|\hat\beta_n - \beta_0\| \;< \varepsilon\,. \tag{A.12}$$

If $\Theta \subset B(\beta_0, \varepsilon)$, then inequality (A.12) holds because $\hat{\beta}_n$ belongs to $\Theta$. Otherwise, if $\Theta$ is not a subset of $B(\beta_0, \varepsilon)$, then $\Theta \backslash B(\beta_0, \varepsilon)$ is a nonempty compact set where $S_\infty(b) \neq 0$. Hence,

$$\min_{b \in \Theta, \|b-\beta_0\| \geq \varepsilon} \|S_\infty(b)\| > 0 . \tag{A.13}$$

From uniform convergence we get that for large enough $n$,

$$\max_{b \in \Theta} \|S_n(b) - S_\infty(b)\| < \min_{b \in \Theta, \|b-\beta_0\| \geq \varepsilon} \|S_\infty(b)\| . \tag{A.14}$$

For such $n$, the equation $S_n(b) = 0$ has no solution on the set $\Theta \backslash B(\beta_0, \varepsilon)$. For large enough $n$, such that $S_n(\hat{\beta}_n) = 0$ and inequality (A.14) holds true, we get the desired (A.12). Convergence (A.5) is proved. □

**Remark A.7.** If among the conditions of Lemma A.6, the condition (1) is replaced by a weaker condition $S_\infty(\beta_0) = 0$, then for large $n$, it would be possible to guarantee the existence of a solution to equation (A.4). Indeed, in the first part of the proof of Lemma A.6 (about the existence of solution), we did not utilize the condition (1), but used only the equality $S_\infty(\beta_0) = 0$.

**Corollary A.8.** *Let $\Theta$ be an arbitrary closed set in $\mathbf{R}^d$, $\beta_0 \in \Theta^0$, and $\{S_n, n \geq 1\}$ be a sequence of continuous functions from $\Theta$ to $\mathbf{R}^d$ being uniformly convergent to a function $S_\infty$ on each compact set $K \subset \Theta$. Also assume that condition* (1) *(for a new $\Theta$) and condition* (2) *of Lemma* A.6 *hold true.*

*Then for all n large enough, the equation*

$$S_n(b) = 0 , \quad b \in \Theta , \tag{A.15}$$

*has a solution. Denote by $\tilde{\beta}_n$ a solution to the equation having the less norm (if there are several such numbers then we take any of them; $\tilde{\beta}_n$ is well-defined for n large enough). Then*

$$\tilde{\beta}_n \to \beta_0 , \quad as \ \ n \to \infty . \tag{A.16}$$

*Proof. Existence of solution.* The point $\beta_0 \in \Theta^0$, therefore, for some $\varepsilon_0$, it holds that $\bar{B}(\beta_0, \varepsilon_0) \subset \Theta$. On this closed ball, there is a solution to the equation $S_n(b) = 0$ for $n$ being large enough, say $n \geq n_0$, as a result of Lemma A.6.

*Convergence.* Let $n \geq n_0$ and $R = \|\beta_0\| + \varepsilon_0$. Then

$$\tilde{\beta} \in \Theta_1 = \bar{B}(0, R) \cap \Theta . \tag{A.17}$$

The set $\Theta_1$ is closed as an intersection of two closed sets, and it is bounded. Therefore, $\Theta_1$ is a compact subset of $\Theta$, and $\beta_0 \in \Theta_1^0$. Apply Lemma A.6 to the set $\Theta_1$ and get the desired convergence (A.16). □

The next statement is related to Lemma A.6. Assuming additionally the uniform convergence of derivatives $S_n'$ we ensure the uniqueness of solution to equation (A.4).

**Lemma A.9.** *Let $\Theta$ be a compact set in $\mathbf{R}^d$, $\{S_n, n \geq 1\}$ be a sequence of continuous functions from $\Theta$ to $\mathbf{R}^d$ being a uniformly convergent to a function $S_\infty$, and the equation $S_\infty(b) = 0$, $b \in \Theta$, have a unique solution $\beta_0$. Suppose that for some ball $B(\beta_0, \varepsilon) \subset \Theta$, the following holds true:*

(1) *For all $n \geq 1$, the functions $S_n$ are continuously differentiable on the set $B(\beta_0, \varepsilon)$, and the function $S_\infty$ is differentiable on this set as well.*
(2) *$S_n'(b)$ converges uniformly to $S_\infty'(b)$ on $B(\beta_0, \varepsilon)$.*
(3) *The derivative $\Phi = S_\infty'(\beta_0)$ is nonsingular matrix.*

*Then for large enough $n$, equation* (A.4) *has a unique solution. The sequence of solutions converges to $\beta_0$.*

*Proof.* Consider again the function (A.6) and also the function

$$R_n(b) = \Phi f_n(b) = \Phi b - S_n(b)\,. \tag{A.18}$$

We have

$$R_n'(b) = \Phi - S_n'(b)\,, \quad b \in B(\beta_0, \varepsilon)\,. \tag{A.19}$$

From condition (2), it follows that there exists a number $n_2$ such that for all $n \geq n_2$,

$$\sup_{b \in B(\beta_0, \varepsilon)} \|S_n'(b) - S_\infty'(b)\| < \frac{1}{4\|\Phi^{-1}\|}\,. \tag{A.20}$$

The derivative $S_\infty'(b)$, $b \in B(\beta_0, \varepsilon)$, is continuous as a uniform limit of continuous functions. In particular, $S_\infty'(b)$ is continuous at the point $\beta_0$. Thus, there exists such an $\varepsilon_2 < \varepsilon$ that

$$\sup_{b \in \bar{B}(\beta_0, \varepsilon_2)} \|S_\infty'(b) - S_\infty'(\beta_0)\| < \frac{1}{4\|\Phi^{-1}\|}\,. \tag{A.21}$$

Then for all $n \geq n_2$,

$$\sup_{b \in \bar{B}(\beta_0, \varepsilon_2)} \|R_n'(b)\| = \sup_{b \in \bar{B}(\beta_0, \varepsilon_2)} \|S_\infty'(\beta_0) - S_n'(b)\| < \frac{1}{2\|\Phi^{-1}\|}\,. \tag{A.22}$$

According to Lagrange's theorem for vector functions (Burkill, 1962, §12.2.2, Theorem 4), for all $n \geq n_2$ and $\|b_i - \beta_0\| \leq \varepsilon_2$, $i = 1, 2$, we get

$$\|R_n(b_1) - R_n(b_2)\| \leq \sup_{\tau \in (0,1)} \|R_n'(b_2 + \tau(b_1 - b_2))\| \cdot \|b_1 - b_2\| \leq \frac{\|b_1 - b_2\|}{2\|\Phi^{-1}\|}\,, \tag{A.23}$$

$$\|f_n(b_1) - f_n(b_2)\| \leq \|\Phi^{-1}\| \cdot \|R_n(b_1) - R_n(b_2)\| \leq \tfrac{1}{2}\|b_1 - b_2\|\,. \tag{A.24}$$

The conditions of Lemma A.6 are fulfilled on the set $\bar{B}(\beta_0, \varepsilon_2)$. From the proof of Lemma A.6, it follows that there exist $\varepsilon_1 \leq \varepsilon_2$ and $n_1 \geq 1$ such that for all $n \geq n_1$, it holds that $f_n(\bar{B}(\beta_0, \varepsilon_1)) \subset \bar{B}(\beta_0, \varepsilon_1)$. Then according to (A.24), the function $f_n$ is a contracting mapping of the closed ball $\bar{B}(\beta_0, \varepsilon_1)$, for $n \geq \max(n_1, n_2)$. By Theorem A.2, the function $f_n$ possesses a unique fixed point on the set $\bar{B}(\beta_0, \varepsilon_1)$ for such $n$. For the same $n$, the equation $S_n(b) = 0$ has a unique solution on the set $\bar{B}(\beta_0, \varepsilon_1)$.

In a similar manner as it was done in the proof of Lemma A.6, it is shown that the equation $S_n(b) = 0$ has no solution outside the ball $\bar{B}(\beta_0, \varepsilon_1)$ for $n$ being large enough.

Thus, for large enough $n$, equation (A.4) has a unique solution, and it lies on the ball $\bar{B}(\beta_0, \varepsilon_1) \subset \Theta$.

The convergence of the solution to $\beta_0$ follows from Lemma A.6. □

### A.1.2 Existence, uniqueness, and consistency of estimators defined by estimating equations

Remember that a concept of convex set in the Euclidean space was introduced in Definition 2.3.4.

Let $\Theta$ be a convex closed set in $\mathbf{R}^p$, $\{z_k, k \geq 1\}$ be a sequence of independent identically distributed random vectors, with distribution that depends on a parameter $\theta \in \Theta$, and the vectors be distributed in a Borel measurable set $Z \subset \mathbf{R}^m$. Further we assume that the true value $\theta \in \Theta^0$. The first $n$ vectors $z_1, \ldots, z_n$ are observed. In order to construct the estimator of the parameter $\theta$, we use the so-called *estimating function* $s(z, t)$, $z \in Z$, $t \in \Theta$, with its values in $\mathbf{R}^p$. The function has to be Borel measurable in the first argument. Form the *estimating equation*

$$S_n(t) = 0\,, \quad t \in \Theta\,; \quad S_n(t) := \frac{1}{n}\sum_{i=1}^{n} s(z_i, t)\,. \tag{A.25}$$

A random vector $\hat{\theta} = \hat{\theta}_n$ being a solution to equation (A.25) (if such solution exists) will be called the *estimator* of the parameter $\theta$ corresponding to the estimating function $s(z, t)$. A more precise definition of the estimator will be presented later.

Remember that for a sequence of random statements, the notion "it holds *eventually*" was introduced in Definition 12.2. "*Eventually*" means the following: something holds true with probability 1 for all $n$, beginning with certain random number $n_0(\omega)$, $\omega \in \Omega$. Hereafter $(\Omega, F, \mathbf{P})$ is a fixed probability space. We write $\mathbf{P} = \mathbf{P}_\theta$, if $\theta$ is the true value of desired parameter of underlying observation model.

Further denote by $z$ a stochastic copy of $z_1$, i.e., a random vector on $(\Omega, F, \mathbf{P})$ having the same distribution as the vector $z_1$.

**Theorem A.10** (Existence of a solution to equation). *Let the convex parameter set $\Theta$ be compact and the following conditions hold:*

(a) *almost surely $s(z, \cdot) \in C^1(\Theta)$, and for all $t \in \Theta$,*

$$\mathbf{E}_\theta \|s(z, t)\| \ < \infty\,, \tag{A.26}$$

(b)

$$S_\infty(t, \theta) := \mathbf{E}_\theta s(z, t)\,, \quad S_\infty(\cdot, \theta) \in C^1(\Theta)\,, \quad S_\infty(\theta, \theta) = 0\,, \tag{A.27}$$

(c)

$$\mathbf{E}_\theta \sup_{t \in \Theta} \left\| \frac{\partial s(z, t)}{\partial t^{\mathrm{T}}} \right\| < \infty\,, \tag{A.28}$$

(d) $V := .\frac{\partial S_\infty(t,\theta)}{\partial t^{\mathrm{T}}}|_{t=\theta}$ *is nonsingular matrix.*

*Then* eventually *with respect to* $\mathbf{P}_\theta$*, there exists* $\hat{\theta}_n$ *being a solution to equation* (A.25)*, i.e., the equality*

$$S_n(\hat{\theta}_n) = 0 \tag{A.29}$$

*holds almost surely (a.s.), for all* $n \geq n_0(\omega)$.

**Remark A.11.** Hereafter the notation $f \in C^k(\Theta)$ means that the function $f$ is actually defined on a wider open set $U \supset \Theta$, and moreover $f$ is $k$ times continuously differentiable on $U$.

*Proof of the theorem.* First we show that almost surely, $S_n(t) \to S_\infty(t,\theta)$ uniformly in $t \in \Theta$.

Indeed, according to the strong law of large numbers (SLLN) for each $t \in \Theta$, a.s.,

$$S_n(t) \to S_\infty(t,\theta)\,, \quad \text{as } n \to \infty\,. \tag{A.30}$$

Moreover, as a consequence of the condition (c), $\{S_n(\cdot), n \geq 1\}$ are equicontinuous, almost surely. The latter follows from the convexity of $\Theta$ and the relations:

$$\sup_{t\in\bar{\Theta}} \left\| \frac{\partial S_n}{\partial t^{\mathrm{T}}}(t) \right\| \leq \frac{1}{n}\sum_{i=1}^{n} \sup_{t\in\bar{\Theta}} \left\| \frac{\partial s}{\partial t^{\mathrm{T}}}(z_i,t) \right\| \overset{\text{P1}}{\longrightarrow} \mathbf{E}_\theta \sup_{t\in\bar{\Theta}} \left\| \frac{\partial s}{\partial t^{\mathrm{T}}}(z,t) \right\| < \infty\,, \tag{A.31}$$

and then a.s.,

$$\sup_{n\geq 1} \sup_{t\in\bar{\Theta}} \left\| \frac{\partial S_n}{\partial t^{\mathrm{T}}}(t) \right\| < \infty\,. \tag{A.32}$$

The equicontinuity on the convex compact set and point-wise convergence (A.30) ensure that a.s., $S_n(t) \to S_\infty(t,\theta)$ uniformly in $t \in \Theta$.

Now, we explain the latter convergence in more detail. Choose a countable set $T \subset \Theta$ being dense in $\Theta$. Then there exists a random event $\Omega_0$, $\mathbf{P}(\Omega_0) = 1$, such that for all $\omega_0 \in \Omega_0$, the following holds true: $\{S_n(t,\omega_0), n \geq 1, t \in \Theta\}$ are equicontinuous and $S_n(t_i,\omega_0) \to S_\infty(t_i,\theta)$, as $n \to \infty$, for all $t_i \in T$. Then by the Arzela–Ascoli theorem (Burkill, 1962), the sequence of functions $\{S_n(\cdot,\omega_0), n \geq 1\}$ is relatively compact in the space $C(\Theta)$ of continuous functions on the convex compact set $\Theta$ with uniform norm. Let a subsequence $S_{n(k)}(\cdot,\omega_0)$, $k \geq 1$, be uniformly convergent to $F(\cdot)$. This limit function is continuous together with pre-limit functions. Due to the point-wise convergence on $T$, we have $S_\infty(t_i,\theta) = F(t_i)$, $t_i \in T$. Now, because $T$ is dense and both functions $S_\infty(\cdot,\omega_0)$ and $F(\cdot)$ are continuous, we get the identity $S_\infty(t,\theta) = F(t)$, $t \in \Theta$. This fact and relative compactness of $\{S_n(\cdot,\omega_0), n \geq 1\}$ in the space $C(\Theta)$ ensure the uniform convergence of $S_n(\cdot)$ to $S_\infty(\cdot,\theta)$, with probability 1.

Next, fix $\omega_0 \in \Omega_0$. The sequence of functions $\{S_n(t,\omega_0), n \geq 1, t \in \Theta\}$ satisfies the conditions of Remark A.7, and then the equation $S_n(t,\omega_0) = 0$, $t \in \Theta$, has a solution, for all $n \geq n_0(\omega_0)$. This proves the statement. □

**Definition A.12.** In the case of compact set $\Theta$, the *estimator* $\hat{\theta}_n$ *defined by equation* (A.25) is a Borel measurable function of $z_1, \ldots, z_n$, such that for those $\omega \in \Omega$ at which equation (A.25) has a solution, the equality $S_n(\hat{\theta}_n(\omega)) = 0$ holds true.

**Remark A.13.** The existence of such a Borel measurable function follows from the theorem in Pfanzagl (1969, p. 252).

**Remark A.14.** For such $\omega \in \Omega$ that equation (A.25) has no solution, one can set $\hat{\theta}_n(\omega) = t_f$, where $t_f \in \Theta$ is a fixed value. If for some $\omega$, there exist several solutions, then $\hat{\theta}_n(\omega)$ coincides with one of them, but in such a manner that in whole $\hat{\theta}_n = f_n(z_1, \ldots, z_n)$ is a Borel measurable function $f_n$ of the observations.

Theorem A.10 demonstrates that under the conditions of the theorem, it holds *eventually* that

$$S_n(\hat{\theta}_n(\omega)) = 0\,. \tag{A.33}$$

**Theorem A.15** (Strong consistency of the estimator). *Let the conditions of Theorem* A.10 *be fulfilled, as well as the following condition.*
(e) *If* $S_\infty(t, \theta) = 0$ *for some* $t \in \Theta$, *then* $t = \theta$.
*Then the estimator* $\hat{\theta}_n$ *is strongly consistent, i.e.,*

$$\hat{\theta}_n \xrightarrow{\text{P1}} \theta\,, \quad \text{as } n \to \infty\,. \tag{A.34}$$

*Proof.* In the proof of Theorem A.10 we showed that for all $\omega_0 \in \Omega_0$, $\mathbf{P}_\theta(\Omega_0) = 1$, the sequence of functions $\{S_n(t, \omega_0), n \geq 1, t \in \Theta\}$ satisfies the conditions of Remark A.7. But taking into account the condition (e), this sequence satisfies the conditions of Lemma A.6 as well. Therefore, the sequence $\{\hat{\theta}_n(\omega_0), n \geq n_0(\omega_0)\}$ of solutions to equation (A.25) satisfies the relation $\hat{\theta}_n(\omega_0) \to \theta$, as $n \to \infty$. This proves the convergence (A.34). □

**Definition A.16.** The estimator $\hat{\theta}_n$ defined by equation (A.25) in the case of closed unbounded $\Theta$, is a Borel measurable function of $z_1, \ldots, z_n$, such that for those $\omega \in \Omega$ at which equation (A.25) has a solution, the equality $S_n(\hat{\theta}_n(\omega)) = 0$ holds true, and moreover $\hat{\theta}_n(\omega)$ has the lowest norm among the solutions.

**Remark A.17.** In addition let the estimating function $s(z, t)$ be continuous in $t$. Then the function $S_n(t)$ is continuous, and for fixed $\omega \in \Omega$, the set of solutions $\{t \in \Theta\colon S_n(t, \omega) = 0\}$ is closed. If the set is not empty, then there exists a solution (possibly not unique) and it has the lowest norm. Now, existence of the Borel measurable function from Definition A.16 follows from the theorem published in Pfanzagl (1969, p. 252).

**Theorem A.18** (The case of unbounded $\Theta$). *Let the parameter set* $\Theta$ *be convex, closed, and unbounded. Assume the conditions* (a), (b), (d), (e), *and the following condition:*
(c') *For each nonempty compact set* $K \subset \Theta$, *it holds that*

$$\mathbf{E}_\theta \sup_{t \in K} \left\| \frac{\partial s(z, t)}{\partial t^{\mathrm{T}}} \right\| < \infty\,. \tag{A.35}$$

*Then there exists an estimator $\hat{\theta}_n$ in terms of Definition* A.16*, moreover the equality* (A.29) *holds* eventually*, and*

$$\hat{\theta}_n \xrightarrow{\text{P1}} \theta\,, \quad \text{as } n \to \infty\,. \tag{A.36}$$

*Proof.* Existence of the estimator was explained in Remark A.17. Similarly to the proof of Theorem A.10, we get that for all $\omega_0 \in \Omega_0$, $\mathbf{P}(\Omega_0) = 1$, the sequence of functions $\{S_n(t, \omega_0), n \geq 1, t \in \Theta\}$ satisfies the conditions of Corollary A.8. (Here as compacts we take the sequence $K_m = \bar{B}(0, m) \cap \Theta$, $m \geq 1$, and for each of them construct $\Omega_m$, $\mathbf{P}(\Omega_m) = 1$, such that for all $\omega \in \Omega_m$, the sequence $\{S_n(t, \omega), n \geq 1, t \in K_m\}$ converges uniformly to $S_\infty(t)$; further we set $\Omega_0 = \bigcap_{m=1}^{\infty} \Omega_m$.) Then by the corollary we obtain $Q_n(\hat{\theta}_n(\omega_0)) = 0$ *eventually*, and $\hat{\theta}_n(\omega_0) \xrightarrow{\text{P1}} \theta$, as $n \to \infty$. This proves the theorem. □

**Definition A.19.** Let $s\colon Z \times \Theta \to \mathbf{R}^p$ be an estimating function, which is Borel measurable in the first argument. The function is called *unbiased* if for any $\theta \in \Theta$, it holds that

$$\mathbf{E}_\theta s(z, \theta) = 0\,. \tag{A.37}$$

In fact, the condition of unbiasedness appears in condition (b), which (together with other conditions) ensures the consistency of $\hat{\theta}_n$. We show that the unbiasedness of an estimating function is necessary for the consistency of an estimator.

**Theorem A.20.** *Let $\Theta$ be a convex set in $\mathbf{R}^p$, which has at least one interior point; $s\colon Z \times \Theta \to \mathbf{R}^p$ be a Borel measurable estimating function in argument $z$, moreover $s(z, \cdot) \in C(\Theta)$ almost surely; $\hat{\theta}_n$ is an estimator, i.e., a Borel measurable function of observations. For each $\theta \in \Theta^0$, assume the following:*

(1)

$$\hat{\theta}_n \xrightarrow{\mathbf{P}_\theta} \theta\,, \quad \text{as } n \to \infty\,. \tag{A.38}$$

(2)

$$\frac{1}{n} \sum_{i=1}^{n} s(z_i, \hat{\theta}_n) \xrightarrow{\mathbf{P}_\theta} 0\,, \quad \text{as } n \to \infty\,. \tag{A.39}$$

(3) *For each nonempty compact $K \subset \Theta$, it holds that*

$$\mathbf{E}_\theta \sup_{t \in K} \left\| \frac{\partial s(z, t)}{\partial t^{\mathrm{T}}} \right\| < \infty\,. \tag{A.40}$$

(4) *The function $\theta \mapsto \mathbf{E}_\theta s(z, \theta)$ is continuous on $\Theta$.*

*Then the estimating function $s(z, \theta)$ is unbiased.*

*Proof.* Let $\theta_0 \in \Theta^0$. Consider the difference, as $n \to \infty$ (due to the convergence $\hat{\theta}_n \xrightarrow{\mathbf{P}_{\theta_0}} \theta_0$ one can assume that $\hat{\theta}_n \in K = \bar{B}(\theta_0, \varepsilon) \subset \Theta$):

$$\left\| \frac{1}{n}\sum_{i=1}^{n} s(z_i, \hat{\theta}_n) - \frac{1}{n}\sum_{i=1}^{n} s(z_i, \theta_0) \right\| \le \frac{1}{n}\sum_{i=1}^{n} \left\| s(z_i, \hat{\theta}_n) - s(z_i, \theta_0) \right\| \le$$

$$\le \frac{1}{n}\sum_{i=1}^{n} \sup_{t\in K} \left\| \frac{\partial s(z_i, t)}{\partial t^{\mathrm{T}}} \right\| \cdot \|\hat{\theta}_n - \theta_0\| \xrightarrow{\mathbf{P}_{\theta_0}} \mathbf{E}_{\theta_0} \sup_{t\in K} \left\| \frac{\partial s(z, t)}{\partial t^{\mathrm{T}}} \right\| \cdot 0 = 0 \,. \quad \text{(A.41)}$$

Here we used the convexity of $\Theta$ and also the conditions (1), (3), and the SLLN.

Further, using the SLLN we get $\mathbf{P}_{\theta_0}$-almost surely (and therefore, in probability $\mathbf{P}_{\theta_0}$):

$$\frac{1}{n}\sum_{i=1}^{n} s(z_i, \theta_0) \to \mathbf{E}_{\theta_0} s(z, \theta_0) \,, \quad \text{as } n \to \infty \,. \quad \text{(A.42)}$$

From relations (A.41) and (A.42) we obtain, for all $\theta_0 \in \Theta^0$:

$$\mathbf{E}_{\theta_0} s(z, \theta_0) = 0 \,. \quad \text{(A.43)}$$

Now, let $\theta$ be a boundary point of $\Theta$. Since this is a convex set with nonempty interior, there exists a sequence $\{\theta_k\} \subset \Theta^0$ converging to $\theta$. Passing to the limit in equality

$$\mathbf{E}_{\theta_k} s(z, \theta_k) = 0 \,, \quad \text{(A.44)}$$

we have by condition (4):

$$\mathbf{E}_{\theta} s(z, \theta) = 0 \,. \quad \text{(A.45)}$$

Since $\theta$ is arbitrary boundary point of $\Theta_0$, then (A.43) together with (A.45) gives the desired. □

**Remark A.21.** From Theorem A.20, it follows the inconsistency of the naive estimators in structural models with the classic error in covariates, because the corresponding estimating function is biased (i.e., equality (A.37) holds not for all $\theta \in \Theta$).

### A.1.3 The case of pre-estimation of nuisance parameters

Let a part of components of the parameter $\theta \in \Theta \subset \mathbf{R}^d$ be consistently estimated by certain method, and the estimating equation be utilized for estimation of the rest of components of $\theta$.

Thus, let $\theta^{\mathrm{T}} = (\beta;\ \alpha)$, $\beta \in \Theta_\beta \subset \mathbf{R}^p$, $\alpha \in \Theta_\alpha \subset \mathbf{R}^k$, $p+k = d$, and $\Theta = \Theta_\beta \times \Theta_\alpha$. The set $\Theta_\beta$ is assumed closed and convex, and $\Theta_\alpha$ is assumed open. It is given an estimating function $s\colon\ Z \times \Theta \to \mathbf{R}^p$, where $Z$ is a Borel measurable set in $\mathbf{R}^m$; the function $s(z, t)$ is a Borel measurable in a couple of arguments. As before, we observe the first $n$ vectors of a sequence of independent identically distributed random vectors $\{z_k, k \ge 1\}$; the vectors are distributed in $Z$ and their distribution depends on $\theta \in \Theta$. Let $z =^d z_1$.

Suppose that there exists a strongly consistent estimator $\hat{\alpha}_n$ of the parameter $\alpha$, $\hat{\alpha}_n = \hat{\alpha}_n(z_1, \ldots, z_n)$, i.e., $\hat{\alpha}_n \to \alpha$, as $n \to \infty$, $\mathbf{P}_\theta$-almost surely. The estimator of the parameter $\beta$ will be defined by the equation

$$\hat{S}_n(b) = 0\,, \quad b \in \Theta_\beta\,; \quad \hat{S}_n(b) = S_n(b, \hat{\alpha}_n) = \frac{1}{n}\sum_{i=1}^{n} s(z_i; b, \hat{\alpha}_n)\,. \tag{A.46}$$

**Definition A.22.** Let $\Theta_\beta$ be a compact convex set. *The estimator $\hat{\beta}_n$ defined by equation* (A.46) is such a Borel measurable function of $z_1, \ldots, z_n$ that for those $\omega \in \Omega$ at which equation (A.46) has a solution, the following equality holds true:

$$\hat{S}_n(\hat{\beta}_n(\omega)) = 0\,. \tag{A.47}$$

**Theorem A.23.** *Let $\Theta_\beta$ be a compact convex set and $\beta \in \Theta_\beta^0$. Assume the following.*

(1) *$\mathbf{P}_\theta$-almost surely, $s(z, \cdot) \in C^1(\Theta)$, and for all $t \in \Theta$,* $\mathbf{E}_\theta \|s(z, t)\| < \infty$.

(2) $S_\infty(t, \theta) := \mathbf{E}_\theta s(z, t)$, $S_\infty(\cdot, \theta) \in C^1(\Theta)$, $S_\infty(b, \alpha; \theta) = 0$*;*

$$b \in \Theta_\beta \quad \textit{if, and only if,} \quad b = \beta\,. \tag{A.48}$$

(3) *There exists a ball $K_\alpha = \bar{B}(\alpha, r_\alpha) \subset \Theta_\alpha$, with*

$$\mathbf{E}_\theta \sup_{t \in \Theta_\beta \times K_\alpha} \left\| \frac{\partial s}{\partial \alpha^{\mathrm{T}}}(z, t) \right\| < \infty\,, \tag{A.49}$$

$$\mathbf{E}_\theta \sup_{b \in \Theta_\beta} \left\| \frac{\partial s}{\partial \beta^{\mathrm{T}}}(z; b, a) \right\| < \infty\,. \tag{A.50}$$

(4) $V_\beta := \frac{\partial S_\infty(\beta, \alpha; \theta)}{\partial \beta^{\mathrm{T}}}$ *is nonsingular matrix.*

*Then the estimator $\hat{\beta}_n$ (in the sense of Definition* A.22*) satisfies equality* (A.47) eventually*, and*

$$\hat{\beta}_n \xrightarrow{\mathrm{P1}} \beta\,, \quad \textit{as} \quad n \to \infty\,. \tag{A.51}$$

*Proof.* The estimator $\hat{\alpha}_n$ is a Borel measurable function of observations $z_1, \ldots, z_n$, hence as a result of the estimating function $s(z; \beta, \alpha)$ to be Borel measurable, the function $\hat{S}_n(z_1, \ldots, z_n; \beta)$ is a Borel measurable one of the observations and the parameter $\beta$. Then existence of the estimator $\hat{\beta}_n$ in terms of Definition A.22 follows from the theorem in Pfanzagl (1969, p. 252).

Further, we are interested in asymptotic properties of the estimator $\hat{\beta}_n$. So, $\hat{\alpha}_n$ can be replaced with a random variable $\tilde{\alpha}_n$, which is distributed in the ball $K_\alpha$ from condition (3):

$$\tilde{\alpha}_n = \hat{\alpha}_n\,, \quad \text{if } \hat{\alpha}_n \in K_\alpha\,; \quad \text{otherwise } \tilde{\alpha}_n = \alpha\,. \tag{A.52}$$

In consequence of the strong consistency of $\hat{\alpha}_n$, we have that $\hat{\alpha}_n = \tilde{\alpha}_n$ *eventually*. Therefore, we may and do assume that the estimator $\hat{\beta}_n$ fits Definition A.22 for the equation

$$\tilde{S}_n(b) = 0\,, \quad b \in \Theta_\beta\,; \quad \tilde{S}_n(b) = S_n(b, \tilde{\alpha}_n)\,. \tag{A.53}$$

Further arguments resemble the proof of Theorem A.10, with the corresponding changes. Consider

$$\left\| \frac{1}{n} \sum_{i=1}^{n} (s(z_i; b, \tilde{\alpha}_n) - s(z_i; b, \alpha)) \right\| \le \|\tilde{\alpha}_n - \alpha\| \cdot \frac{1}{n} \sum_{i=1}^{n} \sup_{t \in \Theta_\beta \times K} \left\| \frac{\partial s(z_i, t)}{\partial \alpha^{\mathrm{T}}} \right\| . \tag{A.54}$$

We have $\tilde{\alpha}_n \xrightarrow{\text{P1}} \alpha$, as $n \to \infty$, and we get from condition (A.49) by the SLLN:

$$\frac{1}{n} \sum_{i=1}^{n} \sup_{t \in \Theta_\beta \times K} \left\| \frac{\partial s(z_i, t)}{\partial \alpha^{\mathrm{T}}} \right\| \xrightarrow{\text{P1}} \mathbf{E}_\theta \sup_{t \in \Theta_\beta \times K} \left\| \frac{\partial s(z, t)}{\partial \alpha^{\mathrm{T}}} \right\| < \infty . \tag{A.55}$$

Therefore, the right-hand side of (A.54) tends to 0, a.s., and then a.s. uniformly in $b \in \Theta_\beta$,

$$\frac{1}{n} \sum_{i=1}^{n} s(z_i, b, \tilde{\alpha}_n) - \frac{1}{n} \sum_{i=1}^{n} s(z_i, b, \alpha) \to 0 , \quad \text{as } n \to \infty . \tag{A.56}$$

Further, by condition (1) and condition (A.50), it follows that a.s. uniformly in $b \in \Theta_\beta$,

$$\frac{1}{n} \sum_{i=1}^{n} s(z_i; b, \alpha) \to S_\infty(b, \alpha; \theta) = \mathbf{E}_\theta s(z; b, \alpha) . \tag{A.57}$$

From relations (A.56) and (A.57) it follows that a.s. uniformly in $b \in \Theta_\beta$,

$$\tilde{S}_n(b) \to S_\infty(b, \alpha; \theta) , \quad \text{as } n \to \infty . \tag{A.58}$$

Then this uniform convergence holds for all $\omega \in \Omega_0$, $\mathbf{P}(\Omega_0) = 1$.

Fix $\omega_0 \in \Omega_0$. The sequence of functions

$$\{\tilde{S}_n(b, \omega_0), n \ge 1, b \in \Theta_\beta\} \tag{A.59}$$

satisfies the conditions of Lemma A.6 (see conditions (2) and (4) of the theorem), and hence the equation $\tilde{S}_n(b, \omega_0) = 0$, $b \in \Theta_\beta$, has a solution for all $n \ge n_0(\omega_0)$. This proves that $\tilde{S}_n(\hat{\beta}_n(\omega)) = 0$ is performed *eventually*, and this implies the equality (A.47) *eventually*. Finally, by Lemma A.6 any sequence of solutions $\tilde{\beta}_n(\omega_0)$, $n \ge n_0(\omega_0)$, to the equation $\tilde{S}_n(b, \omega_0) = 0$ converges to $\beta$. This proves that $\tilde{\beta}_n(\omega) \xrightarrow{\text{P1}} \beta$, and then the desired convergence (A.51) is also valid. □

**Definition A.24.** Let $\Theta_\beta$ be an unbounded closed convex set. The estimator $\hat{\beta}_n$ defined by equation (A.46) is such a Borel measurable function of $z_1, \ldots, z_n$ that for those $\omega \in \Omega$ for which the equation (A.46) has a solution, the equality (A.47) holds true, and moreover $\hat{\beta}_n(\omega)$ has the lowest norm among the solutions.

**Theorem A.25** (the case of unbounded $\Theta_\beta$). *Let $\Theta_\beta$ be a convex, closed, and unbounded set. Assume the conditions* (1), (2), (4) *of Theorem* A.23 *and the following condition.*

(3′) *There exists such a ball* $K_\alpha = \bar{B}(\alpha, r_\alpha) \subset \Theta_\alpha$ *that for each nonempty compact* $K_\beta \subset \Theta_\beta$, *it holds that*

$$\mathbf{E}_\theta \sup_{t \in K_\beta \times K_\alpha} \left\| \frac{\partial s}{\partial \alpha^{\mathrm{T}}}(z, t) \right\| < \infty , \tag{A.60}$$

$$\mathbf{E}_\theta \sup_{b \in K_\beta} \left\| \frac{\partial s}{\partial \beta^{\mathrm{T}}}(z; b, a) \right\| < \infty . \tag{A.61}$$

*Then the estimator* $\hat{\beta}_n$ *(in the sense of Definition* A.24*) satisfies the equality* (A.47)*, and moreover the convergence* (A.51) *is performed.*

*Proof.* Similarly to the proof of Theorem A.18, we use the functions $\tilde{S}_n(b, \omega)$ defined in (A.53), with $b \in K_{\beta,m} = \bar{B}(0, m) \cap \Theta_\beta$ , $\quad m \geq 1$. □

## A.2 Asymptotic normality of estimators

### A.2.1 The sandwich formula

**Theorem A.26** (uniqueness and asymptotic normality). *Let the conditions of Theorem* A.15 *hold true. Furthermore, suppose the following.*
(f)

$$\mathbf{E}_\theta \| s(z, \theta) \|^2 < \infty . \tag{A.62}$$

(g) $s(z, \cdot) \in C^2(\Theta)$ *almost surely, and for all* $i, j = \overline{1, p}$,

$$\mathbf{E}_\theta \sup_{t \in \Theta} \left\| \frac{\partial^2 s(z, t)}{\partial t_i \partial t_j} \right\| < \infty . \tag{A.63}$$

*Then* eventually *the equation* (A.25) *has a unique solution that defines the estimator* $\hat{\theta}_n$ *in terms of Definition* A.12*. The estimator is asymptotically normal, i.e.,*

$$\sqrt{n}\,(\hat{\theta}_n - \theta) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \Sigma) , \quad \Sigma = V^{-1} B V^{-\mathrm{T}} . \tag{A.64}$$

*Here* $V$ *is specified in condition* (d) *of Theorem* A.10, $V^{-\mathrm{T}} := (V^{-1})^{\mathrm{T}}$, *and*

$$B := \mathbf{E}_\theta s(z, \theta) s^{\mathrm{T}}(z, \theta) . \tag{A.65}$$

**Remark A.27.** The formula (A.64) for the asymptotic covariance matrix $\Sigma$ is called the sandwich formula. For its validity, it is not necessary to have the strong consistency of the estimator. It is enough to have the *consistency* $\hat{\theta}_n \xrightarrow{\mathbf{P}_\theta} \theta$, as $n \to \infty$, and assume the condition (d) and other regularity conditions.

*Proof of the theorem. Uniqueness of the estimator.* Let $\Omega_0$ be a random event constructed in the proof of Theorem A.10, $\mathbf{P}(\Omega_0) = 1$. Fix $\omega_0 \in \Omega_0$. The sequence of

continuous functions $\{S_n(t, \omega_0), n \geq 1, t \in \Theta\}$ converges uniformly to $S_\infty(t; \theta)$, $t \in \Theta$, and conditions (b) and (e) for the limit function are fulfilled. In order to apply Lemma A.9, we verify its second condition. Due to condition (c), we have

$$\frac{\partial S_n(t)}{\partial t^{\mathrm{T}}} = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial s(z_i, t)}{\partial t^{\mathrm{T}}} \xrightarrow{\mathrm{P1}} \mathbf{E}_\theta \frac{\partial s(z, t)}{\partial t^{\mathrm{T}}} = \frac{\partial S_\infty(t; \theta)}{\partial t^{\mathrm{T}}}, \quad t \in \Theta. \tag{A.66}$$

Further, with probability 1 the sequence of matrices $\{\frac{\partial S_n(t,\omega)}{\partial t^{\mathrm{T}}}, n \geq 1, t \in \Theta\}$ is equicontinuous due to condition (g). This fact and the pointwise convergence (A.66) ensure that almost surely

$$\frac{\partial S_n(t)}{\partial t^{\mathrm{T}}} \to \frac{\partial S_\infty(t; \theta)}{\partial t^{\mathrm{T}}}, \quad \text{as } n \to \infty, \tag{A.67}$$

uniformly in $t \in \Theta$. Therefore, we may and do assume that such a convergence is performed for all $\omega_0 \in \Omega_0$.

By Lemma A.9 we have that for all $n \geq n_0(\omega_0)$, the equation $S_n(t, \omega_0) = 0$ has a unique solution. Thus, *eventually* equation (A.25) has a unique solution.

*Asymptotic normality.* The equality $S_n(\hat{\theta}_n) = 0$ holds true *eventually*, and by Theorem A.15, $\hat{\theta}_n \xrightarrow{\mathrm{P1}} \theta$, as $n \to \infty$. By Taylor's formula, we have

$$S_n(\theta) + \frac{\partial S_n}{\partial t^{\mathrm{T}}}(\theta) \cdot (\hat{\theta}_n - \theta) + r(n) \cdot (\hat{\theta}_n - \theta) = 0, \tag{A.68}$$

where the entries of the matrix $r(n)$ are defined as follows:

$$r_{ij}(n) = \frac{\partial S_{ni}}{\partial t_j}(\bar{\theta}_i) - \frac{\partial S_{ni}}{\partial t_j}(\theta), \quad i, j = \overline{1, p}. \tag{A.69}$$

Intermediate points $\bar{\theta}_i \in \Theta$ lie within the segment which connects $\theta$ and $\hat{\theta}_n$.

We show that

$$r_{ij}(n) \xrightarrow{\mathbf{P}_\theta} 0, \quad \text{as } n \to \infty. \tag{A.70}$$

We have

$$|r_{ij}(n)| \leq \sup_{t \in \Theta} \left\| \frac{\partial}{\partial t} \left( \frac{\partial S_{ni}}{\partial t_j} \right)(t) \right\| \cdot \|\hat{\theta}_n - \theta\| = a_{ij}(n) \cdot \|\hat{\theta}_n - \theta\|. \tag{A.71}$$

Due to condition (A.63), the sequence of random variables $\{a_{ij}(n), n \geq 1\}$ is stochastically bounded, and therefore, (A.70) stems from the convergence $\hat{\theta}_n \xrightarrow{\mathrm{P1}} \theta$. Thus,

$$\|r(n)\| \xrightarrow{\mathbf{P}_\theta} 0, \quad \text{as } n \to \infty. \tag{A.72}$$

From equality (A.68), we get

$$\left( \frac{\partial S_n}{\partial t^{\mathrm{T}}}(\theta) + r(n) \right) \sqrt{n} \, (\hat{\theta}_n - \theta) = -\sqrt{n} \, S_n(\theta). \tag{A.73}$$

From relations (A.67) and (A.72), we have

$$\sqrt{n} \, (\hat{\theta}_n - \theta) = -V^{-1} \sqrt{n} \, S_n(\theta) + o_{\mathrm{P}}(1). \tag{A.74}$$

Due to condition (f) and the CLT, we get

$$\sqrt{n}\, S_n(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} s(z_i, \theta) \xrightarrow{\mathrm{d}} \mathrm{N}(0, B)\,, \tag{A.75}$$

where the matrix $B$ is written in (A.65). Finally, from the expansion (A.74) and by Slutsky's Lemma 2.18, it holds that

$$\sqrt{n}\,(\hat{\theta}_n - \theta) \xrightarrow{\mathrm{d}} \mathrm{N}(0, V^{-1} B V^{-\mathrm{T}})\,. \tag{A.76}$$

The theorem is proved. □

**Corollary A.28.** *Assume the conditions of Theorem* A.18 *and condition* (f). *Furthermore, assume the following:*
(g') *Almost surely,* $s(z, \cdot) \in C^2(\Theta)$ *and for each nonempty compact* $K \subset \Theta$, *it holds that*

$$\mathbf{E}_\theta \sup_{t \in K} \left\| \frac{\partial^2 s(z, t)}{\partial t_i \partial t_j} \right\| < \infty\,, \quad i, j = \overline{1, p}\,. \tag{A.77}$$

*Then for the estimator* $\hat{\theta}_n$ *from Definition* A.16, *relations* (A.64) *and* (A.65) *hold true.*

*Proof.* By Theorem A.18, we have $\hat{\theta}_n \xrightarrow{\mathrm{P1}} \theta$. As a compact $K$, one can take the ball $\bar{B}(\theta, \varepsilon) \subset \Theta$, and we may and do assume that for all $n$, it holds that $\hat{\theta}_n \in K$. Next, we use Theorem A.26, with the convex compact $K$ taken instead of $\Theta$. □

**Remark A.29.** Under the conditions of Theorem A.26 or Corollary A.28, the following estimator of the ACM $\Sigma$ can be constructed:

$$\hat{\Sigma}_n = \hat{V}_n^{-1} \hat{B}_n \hat{V}_n^{-\mathrm{T}}\,, \quad \hat{V}_n = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial s(z_i, \hat{\theta}_n)}{\partial \theta^{\mathrm{T}}}\,, \tag{A.78}$$

$$\hat{B}_n = \frac{1}{n} \sum_{i=1}^{n} s(z_i, \hat{\theta}_n) s^{\mathrm{T}}(z_i, \hat{\theta}_n)\,, \quad n \geq 1\,. \tag{A.79}$$

The estimator $\hat{\Sigma}_n$ is strongly consistent, i.e., $\|\hat{\Sigma}_n - \Sigma\| \xrightarrow{\mathrm{P1}} 0$, as $n \to \infty$.

**Theorem A.30** (Asymptotic normality in the case of pre-estimation). *Assume the conditions of Theorem* A.23. *Furthermore, assume the following:*
(g") *Almost surely,* $s(z, \cdot) \in C^2(\Theta)$ *and for all* $i, j = \overline{1, p}$,

$$\mathbf{E}_\theta \sup_{t \in \Theta_\beta \times K_\alpha} \left\| \frac{\partial^2 s(z, t)}{\partial t_i \partial t_j} \right\| < \infty\,. \tag{A.80}$$

(e) *A convergence holds*

$$\sqrt{n} \begin{pmatrix} S_n(\beta, \alpha) \\ \hat{\alpha}_n - \alpha \end{pmatrix} \xrightarrow{\mathrm{d}} \mathrm{N}(0, C)\,, \tag{A.81}$$

*where* $C$ *is a positive semidefinite matrix.*

*Then*

$$\sqrt{n}\,(\hat{\beta}_n - \beta) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \Sigma_\beta)\,, \quad \Sigma_\beta = V_\beta^{-1}\,[I_p;\ V_\alpha]\ C\,[I_p;\ V_\alpha]^{\mathrm{T}}\,V_\beta^{-1}\,. \tag{A.82}$$

*Here $V_\beta$ is defined in condition* (4) *of Theorem* A.23 *and*

$$V_\alpha := \frac{\partial S_\infty}{\partial \alpha^{\mathrm{T}}}(\beta, \alpha; \theta)\,. \tag{A.83}$$

*Proof.* By Theorem A.23, it holds *eventually* that

$$S_n(\hat{\beta}_n, \hat{\alpha}_n) = 0\,, \tag{A.84}$$

and $\hat{\theta}_n = (\hat{\beta}_n^{\mathrm{T}}, \hat{\alpha}_n^{\mathrm{T}})^{\mathrm{T}} \xrightarrow{\mathrm{P1}} \theta = (\beta^{\mathrm{T}}, \alpha^{\mathrm{T}})^{\mathrm{T}}$. We may and do assume that for all $n \geq 1$, it holds that $\hat{\alpha}_n \in K_\alpha = \bar{B}(\alpha, r_\alpha)$, the latter set was introduced in condition (A.49). From equality (A.84), we obtain *eventually* using Taylor's formula:

$$S_n(\beta, \alpha) + \frac{\partial S_n}{\partial b^{\mathrm{T}}}(\beta, \alpha)\,(\hat{\beta}_n - \beta) + \frac{\partial S_n}{\partial \alpha^{\mathrm{T}}}(\beta, \alpha)\,(\hat{\alpha}_n - \alpha) + r(n)\,(\hat{\theta}_n - \theta) = 0\,. \tag{A.85}$$

Here $r(n)$ is a random matrix containing differences of partial derivatives of $S_n$ in the true point $\theta$ and intermediate points. Similarly, as in the proof of Theorem A.26, condition (g") implies that $\|r(n)\| \xrightarrow{\mathbf{P}_\theta} 0$, as $n \to \infty$. Then from equality (A.85), we find

$$V_\beta\,\sqrt{n}\,(\hat{\beta}_n - \beta) = -\sqrt{n}\,S_n(\beta, \alpha) - V_\alpha\,\sqrt{n}\,(\hat{\alpha}_n - \alpha) + o_{\mathrm{P}}(1)\,, \tag{A.86}$$

$$\sqrt{n}\,(\hat{\beta}_n - \beta) = -V_\beta^{-1}\,[I_p;\ V_\alpha]\,\sqrt{n}\begin{pmatrix} S_n(\beta, \alpha) \\ \hat{\alpha}_n - \alpha \end{pmatrix} + o_{\mathrm{P}}(1)\,. \tag{A.87}$$

By the condition (e) with Slutsky's lemma, the desired relation (A.82) is obtained. □

**Corollary A.31.** (Asymptotic normality in the case of unbounded set $\Theta_\beta$, with pre-estimation.) *Assume the conditions of Theorem* A.25, *condition* (e), *and the following condition:*

(g"')*Almost surely, $s(z, \cdot) \in C^2(\Theta)$ and for the ball $K_\alpha$ from condition* (A.60) *and each nonempty compact $K_\beta \subset \Theta_\beta$, it holds that*

$$\mathbf{E}_\theta \sup_{t \in K_\beta \times K_\alpha} \left\| \frac{\partial^2 s(z, t)}{\partial t_i \partial t_j} \right\| < \infty\,, \quad i, j = \overline{1, p}\,. \tag{A.88}$$

*Then the convergence* (A.82) *holds for the estimator $\hat{\beta}_n$, which satisfies Definition* A.24.

**Remark A.32.** For models with measurement errors, the formula for the ACM of the estimator $\hat{\beta}_n$ with pre-estimation of nuisance parameters was obtained in Section 4.2.2, see formula (4.111). It was written out based on the sandwich formula (A.64), with consideration of the total estimating function for a vector of all unknown parameters of the model. If this is impossible under the pre-estimation of nuisance parameters, then Theorem A.30 can be used instead.

### A.2.2 A class of asymptotically normal estimators in mean-variance model

Let the relationship between the response $y$ and covariate $x$ be given by the conditional mean and conditional variance:

$$\mathbf{E}(y|x) = m(x, y) , \quad V(y|x) = v(x, \theta) . \tag{A.89}$$

Here $\theta$ is a vector parameter to be estimated using observable independent realizations of the model $(x_i, y_i)$, $i = \overline{1, n}$.

The parameter $\theta$ belongs to a compact set $\Theta \subset \mathbf{R}^d$. The random variable $x$ has a density $\rho(x, \theta)$ with respect to a $\sigma$-finite measure on the Borel $\sigma$-algebra of the real line. We assume that $v(x, \theta) > 0$, for all $x$ and $\theta$, and that the functions (A.89) are smooth enough. This model is called the mean-variance model, see Carroll et al. (2006).

With the functions $g, h\colon \mathbf{R} \times \Theta \to \mathbf{R}^d$, we introduce the estimating function

$$s_{\mathrm{L}}(x, y; \theta) = yg(x, \theta) - h(x, \theta) . \tag{A.90}$$

Consider

$$\mathbf{E}_\theta s_{\mathrm{L}}(x, y; \theta) = \mathbf{E}\mathbf{E}_\theta[s_{\mathrm{L}}(x, y; \theta)|x] = \mathbf{E}(m(x, \theta)g(x, \theta) - h(x, \theta)) . \tag{A.91}$$

The function $s_{\mathrm{L}}$ is unbiased (i.e., for each $\theta$, it holds that $\mathbf{E}_\theta s_{\mathrm{L}}(x, y; \theta) = 0$ ) if, and only if, $\mathbf{E}(m(x, \theta)g(x, \theta) - h(x, \theta)) = 0$ for all $\theta$.

The estimating function (A.90) generates the estimator $\hat{\theta}_L$ specified by the equation

$$\sum_{i=1}^{n} s_{\mathrm{L}}(x_i, y_i; \theta) = 0 , \quad \theta \in \Theta . \tag{A.92}$$

**Theorem A.33.** *Consider the model* (A.89) *and assume the following:*

(1) *Parameter set $\Theta$ is a convex compact set in $\mathbf{R}^d$, and the true value $\theta \in \Theta^0$.*

(2) *Functions $g, h\colon \mathbf{R} \times U \to \mathbf{R}^d$ are Borel measurable, where $U$ is a neighborhood of $\Theta$, and moreover $g(x, \cdot)$ and $h(x, \cdot)$ belong to $C^2(U)$, almost surely.*

(3) $\mathbf{E}|m(x, \theta)| \cdot \|g(x, t)\| < \infty$, *for all $\theta \in \Theta^0$, $t \in \Theta$;*
$\mathbf{E}m^2(x, \theta) \cdot \|g(x, \theta)\|^2 < \infty$, *for all $\theta \in \Theta^0$.*

(4) $\mathbf{E}|m(x, \theta)| \cdot \sup_{t\in\Theta} |\mathrm{D}_t^{(j)} g_k(x, t)| < \infty$, *for all $\theta \in \Theta^0$, $k = \overline{1, d}$, $j = 1, 2$;*
$\mathbf{E}\sup_{t\in\Theta} |\mathrm{D}_t^{(j)} h_k(x, t)| < \infty$, *for all $k = \overline{1, d}$, $j = 1, 2$, where $g_k$ and $h_k$ are the corresponding components of $g$ and $h$; $\mathrm{D}_t^{(j)} g_k$ and $\mathrm{D}_t^{(j)} h_k$ denote partial derivatives of order $j$ with respect to the argument $t$ of the functions $g_k$ and $h_k$, respectively.*

(5) *For each $\theta \in \Theta^0$, the equality*

$$\mathbf{E}(m(x, \theta)g(x, t) - h(x, t)) = 0 , \quad t = \Theta , \tag{A.93}$$

*holds true if, and only if, $t = \theta$.*

(6) *The matrix $A_{\mathrm{L}} = -\mathbf{E}_\theta \frac{\partial s_{\mathrm{L}}(x,y,\theta)}{\partial \theta^{\mathrm{T}}}$ is nonsingular.*

*Then:*

(a) *there exists such a Borel measurable function* $\hat{\theta}_L$ *of observations* $x_1, y_1, \ldots, x_n, y_n$ *that* $\sum_{i=1}^{n} s_{\mathrm{L}}(x_i, y_i; \hat{\theta}_L) = 0$ eventually, *and*

(b) *for each such function* $\hat{\theta}_L$, *the following holds:*

$$\hat{\theta}_L \xrightarrow{\mathrm{P1}} \theta\,, \quad \text{as } n \to \infty, \tag{A.94}$$

$$\sqrt{n}\,(\hat{\theta}_L - \theta) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \Sigma_{\mathrm{L}}), \tag{A.95}$$

$$\Sigma_{\mathrm{L}} = A_{\mathrm{L}}^{-1} B_{\mathrm{L}} A_{\mathrm{L}}^{-\mathrm{T}}\,, \quad B_{\mathrm{L}} = \mathbf{E}_\theta s_{\mathrm{L}}(x, y; \theta) s_{\mathrm{L}}^{\mathrm{T}}(x, y; \theta)\,. \tag{A.96}$$

*Proof* is based on Theorems A.10, A.15, and A.26.

**Theorem A.34** (The case of unbounded $\Theta$). *Consider the model* (A.89) *and assume the following:*

(1) *Parameter set* $\Theta$ *is a convex closed set in* $\mathbf{R}^d$ *and the true value* $\theta \in \Theta^0$.

(2) *For each nonempty compact set* $K \subset \Theta$, *it holds that*
$\mathbf{E}|m(x, \theta)| \cdot \sup_{t \in K} |\mathrm{D}_t^{(j)} g_k(x, t)| < \infty$, *for all* $\theta \in \Theta^0$, $k = \overline{1, d}$, $j = \overline{1, 2}$;
$\mathbf{E} \sup_{t \in K} |\mathrm{D}_t^{(j)} h_k(x, t)| < \infty$, *for all* $k = \overline{1, d}$, $j = \overline{1, 2}$.

(3) *Conditions* (2), (3), (5), *and* (6) *of Theorem* A.33 *are repeated word-for-word.*

*Then:*

(a) *there exists such a Borel measurable function* $\hat{\theta}_L$ *of observations* $x_1, y_1, \ldots, x_n, y_n$ *that almost surely,* $\sum_{i=1}^{n} s_{\mathrm{L}}(x_i, y_i; \hat{\theta}_L) = 0$, *for all* $n \geq n_0(\omega)$, *and* $\hat{\theta}_L(\omega)$ *has the lowest norm of all solutions to equation* (A.92);

(b) *for each such function* $\hat{\theta}_L$, *relations* (A.94)–(A.96) *hold true.*

*Proof* is based on Theorem A.18 and Corollary A.28.

# B Consistency of efficient methods

Introduce a function

$$F_N(\beta) = \left( \sum_{i=1}^{N} \frac{Y_i}{1+\beta D_i} \right)^{-1} \sum_{i=1}^{N} \frac{Y_i\,(D_i - D_{\mathrm{av}})}{1+\beta D_i}\,, \quad \beta \geq 0\,, \tag{B.1}$$

where $D_{\mathrm{av}}$ is defined in the formula (6.11). Relation (6.10) is equivalent to the equation $F_N(\beta) = 0$, $\beta > 0$. The following conditions will be required:
(I) Among the observations with $Y_i = 1$, not all the doses $D_i$ coincide;
(II) $\sum_{i=1}^{N} Y_i\,(D_i - D_{\mathrm{av}}) > 0$;
(III) $\sum_{i=1}^{N} \frac{Y_i\,(D_i - D_{\mathrm{av}})}{D_i} < 0$.

Condition (I) means that not all sick subjects received the same dose. Condition (II) is natural as well and means that the mean dose for healthy subjects is less than the mean dose for sick subjects. And it can be shown that for the structural model in which the dose distribution is not degenerate to a single point (i.e., for every $D_0 \geq 0$, the inequality $\mathbf{P}(D = D_0) < 1$ holds true), condition (III) holds *eventually*.

**Lemma B.1.** *If the condition* (I) *holds true, then the function $F_N$ is strictly decreasing, and as a consequence, the equation* (6.10) *has no more than one solution. If conditions* (I)–(III) *are fulfilled then equation* (6.10) *has a unique solution $\beta > 0$.*

To *prove* the first statement of the lemma, we have to verify the inequality $F_N'(\beta) < 0$. Further, the inequality $F_N(0) > 0$ follows from the condition (II), and relation $\lim_{\beta\to+\infty} F_N(\beta) < 0$ stems from the condition (III). Therefore, the continuous function $F_N(\beta)$ equals 0 at some point $\beta > 0$.

**Theorem B.2.** *Assume that the dose distribution is not concentrated at a single point. Then* eventually *the efficient estimators $\hat{\lambda}_0^*$ and $\hat{\beta}^*$ of the model* (6.6) *exist and unique, and the estimators are strongly consistent, i.e., with probability 1 they tend to the true values:* $\lim_{N\to\infty} \hat{\lambda}_0^* = \lambda_0$ *and* $\lim_{N\to\infty} \hat{\beta}^* = \beta$, *almost surely.*

*Proof.* Lemma B.1 implies the existence of the estimators, for all $N \geq N_0(\omega)$, almost surely. Estimating equations for the efficient estimators can be written in the following form:

$$0 = \sum_{i=1}^{N} [1 - Y_i - Y_i/(\lambda_0(1+\beta D_i))]\,, \tag{B.2}$$

$$0 = \sum_{i=1}^{N} D_i[1 - Y_i - Y_i/(\lambda_0(1+\beta D_i))]\,. \tag{B.3}$$

(On the set $\{\lambda_0 > 0\,, \quad \beta > 0\}$ this system of equations is equivalent to (6.8); in particular, in order to get (B.3) it is necessary to divide by $\beta$ the difference between the first

and second equations (6.8).) Equations (B.2) and (B.3) are unbiased, i.e.,

$$\begin{aligned} 0 &= \mathbf{E}_{\lambda_0,\beta}[1 - Y_i - Y_i/(\lambda_0(1+\beta D_i))] \,, \\ 0 &= \mathbf{E}_{\lambda_0,\beta} D_i[1 - Y_i - Y_i/(\lambda_0(1+\beta D_i))] \,. \end{aligned} \tag{B.4}$$

Hereafter, $\mathbf{E}_{\lambda_0,\beta}$ denotes expectation under the condition that $\lambda_0$ and $\beta$ are the true values of parameters in the model (6.6). In addition, the limit system of equations

$$\begin{aligned} 0 &= \mathbf{E}_{\lambda_0,\beta}[1 - Y_i - Y_i/(\ell_0(1+bD_i))] \,, \\ 0 &= \mathbf{E}_{\lambda_0,\beta}[D_i(1 - Y_i - Y_i/(\ell_0(1+bD_i)))] \,, \quad \ell_0 > 0, b > 0 \,, \end{aligned} \tag{B.5}$$

has a unique solution $\ell_0 = \lambda_0, b = \beta$. This can be shown by using a convex function $q(\ell_0, a) = (1-Y)(\ell_0 + aD) - Y\ln(\ell_0 + aD)$. The functions under the sign of expectation are equal to the partial derivatives of $q(\ell_0, a)$ at point $(\ell_0, a) = (\ell_0, \ell_0 b)$. According to theory of estimating equations (see Appendix A1), these facts imply the strong consistency of the efficient estimators. □

# C Efficient SIMEX method as a combination of the SIMEX method and the corrected score method

The expression $\sum_{i=1}^{N}(1-Y_i)D_i^{\mathrm{mes}}\exp(-\frac{\sigma_{Q,i}^2}{2})$ is an unbiased estimator of the sum $\sum_{i=1}^{N}(1-Y_i)D_i^{\mathrm{tr}}$. Indeed,

$$\mathbf{E}\left[(1-Y_i)D_i^{\mathrm{mes}}\exp\left(-\frac{\sigma_{Q,i}^2}{2}\right)\,\middle|\,Y_i, D_i^{\mathrm{tr}}\right] = \\ = (1-Y_i)\mathbf{E}\left[D_i^{\mathrm{mes}}\exp\left(-\frac{\sigma_{Q,i}^2}{2}\right)\,\middle|\,D_i^{\mathrm{tr}}\right] = (1-Y_i)D_i^{\mathrm{tr}}\,. \quad \text{(C.1)}$$

In derivation of this relation, it was assumed that the multiplicative errors have log-normal distribution. Thus, equation (6.55) is obtained from the equation

$$\sum_{i=1}^{N}(1-Y_i)(1+\widehat{ERR}_b^*(\kappa)D_i^{\mathrm{tr}}) = \frac{\sum_{i=1}^{N}Y_i}{\hat{\lambda}_{0,b}^*(\kappa)}\,, \quad \text{(C.2)}$$

by the corrected score method, see Section 1.4.4.

We introduce the notations:

- $\theta = (\lambda_0, EAR)^{\mathrm{T}}$,
- $s(\theta, Y, D)$ is an elementary estimating function for the naive estimate,
- $\hat{\theta}_b^*(\kappa)$ and $\hat{\theta}^*(\kappa)$ are estimates of the parameter $\theta$ being used in the *SIMEX* method.

For ordinary *SIMEX*, with using fast estimates, the elementary estimating function can be written as:

$$s(\theta, Y, D) = \begin{pmatrix} 1-Y-Y/(\lambda_0+EAR\cdot D) \\ (1-Y-Y/(\lambda_0+EAR\cdot D))D \end{pmatrix}, \quad \text{(C.3)}$$

and for the efficient modification of *SIMEX*,

$$s(\theta, Y, D) = \begin{pmatrix} 1-Y-Y/(\lambda_0+EAR\cdot D) \\ (1-Y)\hat{D}_{\mathrm{av}}-YD/(\lambda_0+EAR\cdot D) \end{pmatrix}. \quad \text{(C.4)}$$

A random function $\hat{\theta}_b^*$ is searched as a solution to the equation

$$\sum_{i=1}^{N} s(\hat{\theta}_b^*(\kappa), Y_i, D_i^{\mathrm{mes}}\exp(\sqrt{\kappa}U_{b,i}^*)) = 0\,. \quad \text{(C.5)}$$

We find the derivative $\frac{d\hat{\theta}_b^*}{dr}|_{r=0}$, where $r=\sqrt{\kappa}$. For this purpose compute the partial derivatives $s'_D$ and $s'_\theta$ of the estimating function $s(\theta, Y, D)$:

$$\left.\frac{\partial\sum_{i=1}^{N}s(\theta, Y_i, D_i^{\mathrm{mes}}\mathrm{e}^{rU_{b,i}^*})}{\partial r}\right|_{r=0} = \sum_{i=1}^{N}s'_D(\theta, Y_i, D_i^{\mathrm{mes}})D_i^{\mathrm{mes}}U_{b,i}^*\,, \quad \text{(C.6)}$$

$$\left.\frac{\partial\sum_{i=1}^{N}s(\theta, Y_i, D_i^{\mathrm{mes}}\mathrm{e}^{rU_{b,i}^*})}{\partial\theta}\right|_{r=0} = \sum_{i=1}^{N}\frac{Y_i}{(\theta_1+\theta_2D_i^{\mathrm{mes}})^2}\begin{pmatrix} 1 & D_i^{\mathrm{mes}} \\ D_i^{\mathrm{mes}} & (D_i^{\mathrm{mes}})^2 \end{pmatrix}. \quad \text{(C.7)}$$

Under the conditions for existence of the efficient estimators, the derivative $\sum_{i=1}^{N} s'_{\theta}(\theta, Y_i, D_i^{\text{mes}})$ is nonsingular matrix.

By the implicit function theorem, it holds that

$$\left.\frac{d\hat{\theta}_b^*(r^2)}{dr}\right|_{r=0} = -\left(\sum_{i=1}^{N} s'_{\theta}(\hat{\theta}_b^*(0), Y_i, D_i^{\text{mes}})\right)^{-1} \sum_{i=1}^{N} s'_{D}(\hat{\theta}_b^*(0), Y_i, D_i^{\text{mes}}) D_i^{\text{mes}} U_{i,b}^* . \quad \text{(C.8)}$$

We mention that $\hat{\theta}_b^*(0) = \theta_1^*(0)$ does not depend of $b$ and find the derivative of the function $\hat{\theta}^*(r^2) = B^{-1} \sum_{i=1}^{N} \hat{\theta}_b^*(r^2)$ at zero point:

$$\left.\frac{d\hat{\theta}^*(r^2)}{dr}\right|_{r=0} = -\frac{\sum_{i=1}^{N} s'_{D}(\hat{\theta}_1^*(0), Y_i, D_i^{\text{mes}}) D_i^{\text{mes}} \sum_{b=1}^{B} U_{b,i}^*}{B \sum_{i=1}^{N} s'_{\theta}(\hat{\theta}_1^*(0), Y_i, D_i^{\text{mes}})} . \quad \text{(C.9)}$$

If the condition $\sum_{b=1}^{B} U_{b,i}^* = 0$ holds true, then the derivative $\frac{d\hat{\theta}^*(r^2)}{dr}$ is equal to 0 at zero point, and in the expansion

$$\hat{\theta}^*(\kappa) = \hat{\theta}^*(0) + \text{coef}_1 \sqrt{\kappa} + \text{coef}_2 \kappa + \text{coef}_3 \sqrt{\kappa^3} + \dots \quad \text{(C.10)}$$

the term $\text{coef}_1 \sqrt{\kappa}$ is vanishing. Then the expansion (C.10) resembles the Taylor series expansion of the function $\hat{\theta}^*(\kappa)$ with respect to the variable $\kappa$. Therefore, the extrapolated value $\hat{\theta}^*(-1)$, which is actually the *SIMEX* estimator, is calculated more stably.

# D Application of regression calibration in the model with additive error in exposure doses

Let the doses be observed with the classical additive error (7.23). The basic idea of regression calibration (RC) (see Carroll et al., 2006) lies in using the conditional expectations $\mathbf{E}(\overline{D}_i^{\text{tr}}|D_i^{\text{mes}})$ instead of the true doses within the framework of the radiation risk model.

## D.1 Parametric regression calibration

The method of parametric regression calibration assumes that probability distribution of the population $\overline{D}_i^{\text{tr}}$, $i = 1, \ldots, N$, is known (or it can be reliably estimated). As a result of the fact that the thyroid exposure doses are positive and possess left-skewed distribution (Likhtarov et al., 2014), the lognormal distribution serves as a satisfactory approximation for the distribution of doses. Therefore,

$$\ln \overline{D}^{\text{tr}} \sim \text{N}(\mu_{\overline{D}^{\text{tr}}}, \sigma^2_{\overline{D}^{\text{tr}}}) \,. \tag{D.1}$$

The parameters $\mu_{\overline{D}^{\text{tr}}}$ and $\sigma^2_{\overline{D}^{\text{tr}}}$ are reliably estimated by (7.44) and (7.45). Here we assume those parameters to be known.

Denote the pdf of $\overline{D}_i^{\text{tr}}$ by $\rho_{\text{tr}}$, the pdf of $D_i^{\text{mes}}$ by $\rho_{\text{mes}}$, the joint pdf of $\overline{D}_i^{\text{tr}}$ and $D_i^{\text{mes}}$ by $\rho_{\text{tr,mes}}$, and the pdf of the measurement error $\sigma_i\gamma_i$ by $\rho_\gamma$. Then

$$\rho_{\text{tr,mes}}(\overline{D}_i^{\text{tr}}, D_i^{\text{mes}}) = \rho_{\text{tr}}(\overline{D}_i^{\text{tr}}) \cdot \rho_\gamma(D_i^{\text{mes}} - \overline{D}_i^{\text{tr}}) \,, \tag{D.2}$$

and the conditional pdf is equal to:

$$\rho_{tr|\text{mes}}(\overline{D}_i^{\text{tr}}) = \frac{\rho_{\text{tr,mes}}(\overline{D}_i^{\text{tr}}, D_i^{\text{mes}})}{\int_0^\infty \rho_{\text{tr,mes}}(t, D_i^{\text{mes}})dt} = \frac{\rho_{\text{tr}}(\overline{D}_i^{\text{tr}}) \cdot \rho_\gamma(D_i^{\text{mes}} - \overline{D}_i^{\text{tr}})}{\rho_{\text{mes}}(D_i^{\text{mes}})} \,. \tag{D.3}$$

This implies that

$$\mathbf{E}(\overline{D}_i^{\text{tr}} \Big| D_i^{\text{mes}}) = \int_0^\infty t\rho_{\text{tr}|\text{mes}}(t)dt = \frac{1}{\rho_{\text{mes}}(D_i^{\text{mes}})} \int_0^\infty t\rho_{\text{tr}}(t) \cdot \rho_\gamma(D_i^{\text{mes}} - t)dt \,, \tag{D.4}$$

and thus, it holds (Likhtarov et. al, 2012):

$$\begin{aligned} \rho_{\text{mes}}(D_i^{\text{mes}}) &= \int_0^1 \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(D_i^{\text{mes}} - G^{-1}(z))^2}{2\sigma_i^2}\right)dz \,, \\ \int_0^\infty t\rho_{\text{tr}}(t) \cdot \rho_\gamma(D_i^{\text{mes}} - t)dt &= \int_0^1 \frac{G^{-1}(z)}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(D_i^{\text{mes}} - G^{-1}(z))^2}{2\sigma_i^2}\right)dz \,, \end{aligned} \tag{D.5}$$

where

$$z = G(t) = \int_0^t \frac{1}{t\sqrt{2\pi}\sigma_{\overline{D}^{\mathrm{tr}}}} \exp\left(-\frac{(\log t - \mu_{\overline{D}^{\mathrm{tr}}})^2}{2\sigma^2_{\overline{D}^{\mathrm{tr}}}}\right) dt \tag{D.6}$$

is the cdf of lognormal law, with parameters $\mu_{\overline{D}^{\mathrm{tr}}}$ and $\sigma_{\overline{D}^{\mathrm{tr}}}$.

## D.2 Linear regression calibration

In the case of linear regression calibration, the following linear approximations to conditional expectations of doses are used instead of the true doses:

$$\mathbf{E}(\overline{D}_i^{\mathrm{tr}} \Big| D_i^{\mathrm{mes}}) = a_i + b_i D_i^{\mathrm{mes}}\,, \tag{D.7}$$

where the coefficients $a_i$, $b_i$ are found by relations from Likhtarov et al. (2012):

$$a_i = \frac{\mu_{\overline{D}^{\mathrm{tr}}}\sigma_i^2}{\sigma^2_{\overline{D}^{\mathrm{tr}}} + \sigma_i^2},\; b_i = \frac{\sigma^2_{\overline{D}^{\mathrm{tr}}}}{\sigma^2_{\overline{D}^{\mathrm{tr}}} + \sigma_i^2}\,. \tag{D.8}$$

## D.3 Results of stochastic experiment

Figures D.1 and D.2 present the results of simulation performed in accordance with Section 7.3.4 for the case $\mathrm{GSD}_F = 1$, i.e., in the absence of Berkson errors in the exposure doses. The figures demonstrate that the shift of the estimates of *EAR* being obtained by the parametric and linear regression calibration is even larger than the shift of the naive estimates of this parameter. This is explained by an "unnatural" combination

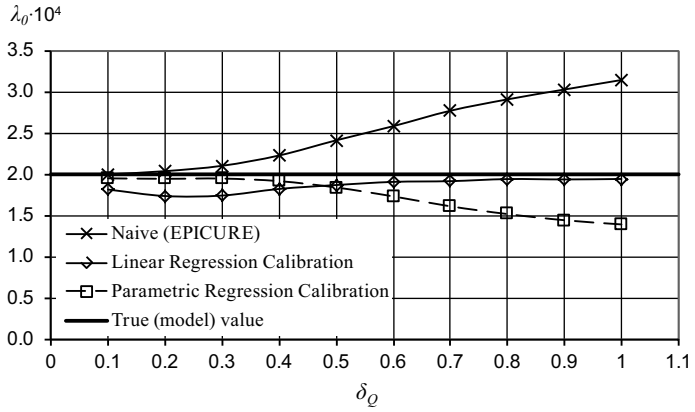


**Fig. D.1:** Estimates of background incidence rate obtained by the parametric and linear regression calibration in the model with classical additive error.

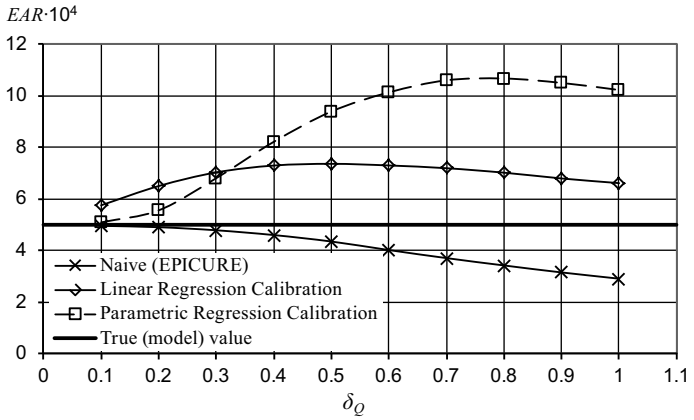


**Fig. D.2:** Excess absolute risk estimates obtained by the parametric and linear regression calibration in the model with classical additive error.

of the normal law of dose errors (7.23) and the lognormal law of the population $\overline{D}^{\mathrm{tr}}$ (D.1). Hence, the methods of regression parameter estimation described above are not applicable for the models, where the normal law of dose errors is combined with the lognormal law of the population of doses.

# Bibliography


[1] Anderson T. W., An Introduction to Multivariate Statistical Analysis. Wiley, New York (2003).

[2] Beirlant J., Goegebeur Yu., Segers J., Teugels J., De Waal D. and Ferro C., Statistics of Extremes: Theory and Applications. Wiley, Hoboken NJ (2004).

[3] Berkson J., Are there two regressions? *J. Amer. Statist. Assoc.* **45** (1950), 164–180.

[4] Bogdanova T. I., Zurnadzhy L. Yu., Nikiforov Y. E. et al, Histopathological features of papillary thyroid carcinomas detected during four screening examinations of a Ukrainian–American cohort. *British Journal of Cancer* **113** (2015), 1556–1564.

[5] Breslow N. E. and Day N. E., Statistical Methods in Cancer Research. Volume 1: The analysis of case-control studies. International Agency for Research on Cancer, Lyon (1980).

[6] Buglova E. E., Kenigsberg J. E. and Sergeeva N. V., Cancer risk estimation in Belarussian children due to thyroid irradiation as a consequence of the Chernobyl nuclear accident. *Health Phys.* **71** (1996), 45–49.

[7] Burkill J. C., A First Course in Mathematical Analysis, 6th edn. Cambridge University Press, Cambridge (1962).

[8] Carroll R. J., Ruppert D. and Stefanski L. A., Measurement Error in Nonlinear Models. Chapman and Hall / CRC, Boca Raton FL (1995).

[9] Carroll R. J., Küchenhoff H., Lombard F. and Stefanski L. A., Asymptotics for the SIMEX estimator in nonlinear measurement error models. *J. Amer. Statist. Assoc.* **91** (1996), 242–250.

[10] Carroll R. J., Ruppert D., Stefanski L. A. and Crainiceanu C., Measurement Error in Nonlinear Models: A Modern Perspective, 2nd edn. Chapman and Hall / CRC, New York (2006).

[11] Cartan H., Differential Forms. Dover, Mineola NY (1970).

[12] Cheng C.-L. and Schneeweiss H., Polynomial regression with errors in the variables. *J. Royal Statist. Society* Series B (Statist. Methodol.) **60** (1998), 189–199.

[13] Cheng C.-L. and Van Ness J. W., Statistical Regression with Measurement Error. Arnold, London (1999).

[14] Cheng C.-L., Schneeweiss H. and Thamerus M., A small sample estimator for a polynomial regression with errors in the variables. *J. Royal Statist. Society* Series B (Statist. Methodol.) **62** (2000), 699–709.

[15] Cheng C.-L. and Kukush A., Non-existence of the first moment of the adjusted least squares estimator in multivariate errors-in-variables model. *Metrika* **64** (2006), 41–46.

[16] Chernov N., Circular and Linear Regression: Fitting Circles and Lines by Least Squares. Chapman and Hall / CRC, Boca Raton FL (2010).

[17] Cook J. R. and Stefanski L. A., Simulation-extrapolation estimation in parametric measurement error models. *J. Amer. Statist. Assoc.* **89** (1994), 1314–1328.

[18] Cox D. R., Regression models and life-tables. *J. Royal Statist. Society* Series B. (Statist. Methodol.) **34**, 187–220. Reprinted in: S. Kotz and N. L. Johnson (Eds.), Breakthroughs in Statistics (p. 527–541), 1992, Springer, New York (1972).

[19] Cox D. R. and Hinkley D. V., Theoretical Statistics. Chapman and Hall, London (1974).

[20] Fuller W. A., Measurement Error Models. Wiley, New York (1987).

[21] Gleser L. J. and Hwang J. T., The nonexistence of 100(1-$\alpha$)% confidence sets of finite expected diameter in errors-in-variables and related models. *Ann. Statist.* **15** (1987), 1351–1362.

[22] Gleser L. J., Improvements of the naive approach to estimation in nonlinear errors-in-variables models. In: P. Brown and W. Fuller (Eds.), Statistical Analysis of Measurement Error Models and Applications (p. 99–114). American Mathematical Society, Providence (1990).

[23] Gol'danskii V. I., Kutsenko A. V. and Podgoretskii M. I., Statistics of Counts in Registration of Nuclear Particles. Moscow: Fizmatgiz. (in Russian) (1959).

[24] Gontar O. and Küchenhoff H., The expansion of a SIMEX estimator in the nonlinear errors-in-variables model with small measurement errors. *Theory of Stochastic Processes* **14** (30), no. 1 (2008), 39–48.
[25] Halmos P. R., Measure Theory. Springer, New York (2013).
[26] Health risks from exposure to low levels of ionizing radiation, BEIR VII Phase 2. National Academy Press, Washington DC (2006).
[27] Holford T. R., The analysis of rates and of survivorship using log-linear models. *Biometrics* **36** (1980), 299–305.
[28] Hosmer D. W., Lemeshow S. and Sturdivant R. X., Applied Logistic Regression, 3rd edn. Wiley, Hoboken NJ (2013).
[29] ICRP, 1990 Recommendations of the International Commission on Radiological Protection. ICRP Publication 60. Ann. ICRP 21 (1–3) (1991).
[30] Jacob P., Bogdanova T. I., Buglova E. et al, Thyroid cancer risk in areas of Ukraine and Belarus affected by the Chernobyl accident. *Radiation Research* **165** (2006), 1–8.
[31] Kaidanovsky G. N., Dolgirev E. I., Calibration of radiometers for mass control of incorporated $^{131}$I, $^{134}$Cs and $^{137}$Cs nuclides with the help of volunteers. *Radiat Prot Dosimetry*. **71** (1997), 187–194.
[32] Kartashov M. V., Probability, Processes, and Statistics. Kyiv University (in Ukrainian), Kyiv (2007).
[33] Kelsey J. L., Whittemore A. S., Evans A. S. and Thompson W. D., Methods in Observational Epidemiology, 2nd edn. Oxford University Press, New York (1996).
[34] Kendall M. and Stuart A., The Advanced Theory of Statistics, Volume 2: Inference and Relationship, 4th edn. Griffin, London (1979).
[35] Kopecky K. J., Stepanenko V., Rivkind N. et al, Childhood thyroid cancer, radiation dose from Chernobyl, and dose uncertainties in Bryansk oblast, Russia: A population-based case-control study. *Radiation Research* **166** (2006), 367–374.
[36] Koroliuk V. S., Portenko N. I., Skorohod A. V. and Turbin A. F., Handbook on Probability Theory and Mathematical Statistics. Nauka (in Russian), Moscow (1985).
[37] Kukush A. and Van Huffel S., Consistency of elementwise-weighted total least squares estimator in a multivariate errors-in-variables model *AX=B*. *Metrika* **59** (2004), 75–97.
[38] Kukush A., Markovsky I. and Van Huffel S., Consistency of the structured total least squares estimator in a multivariate errors-in-variables model. *J. Statist. Plann. Inference* **133** (2005a), 315–358.
[39] Kukush A., Schneeweiss H. and Wolf R., Relative efficiency of three estimators in a polynomial regression with measurement errors. *J. Statist. Plann. Inference* **127** (2005b), 179–203.
[40] Kukush A. and Schneeweiss H., Comparing different estimators in a nonlinear measurement error model. Part I. *Mathematical Methods of Statistics* **14** (2005c), 53–79.
[41] Kukush A. and Schneeweiss H., Comparing different estimators in a nonlinear measurement error model. Part II. *Mathematical Methods of Statistics* **14** (2005d), 203–223.
[42] Kukush A., Malenko A. and Schneeweiss H., Optimality of the quasi-score estimator in a mean-variance model with application to measurement error models. Discussion Paper 494, SFB 386, University of Munich (2006).
[43] Kukush A., Malenko A. and Schneeweiss H., Comparing the efficiency of estimates in concrete errors-in-variables models under unknown nuisance parameters. *Theory of Stochastic Processes* **13** (29), no. 4 (2007), 69–81.
[44] Kukush A., Malenko A. and Schneeweiss H., Optimality of the quasi-score estimator in a mean–variance model with applications to measurement error models. *J. Statist. Plann. Inference* **139** (2009), 3461–3472.

[45] Kukush A., Shklyar S., Masiuk S. et al, Methods for estimation of radiation risk in epidemiological studies accounting for classical and Berkson errors in doses. *The Internat. J. Biostatistics* **7**, no. 1 (2011), article 15.

[46] Li Y., Guolo A., Hoffman F. O. and Carroll R. J., Shared uncertainty in measurement error problems, with application to Nevada test site fallout data. *Biometrics* **63** (2007), 1226–1236.

[47] Likhtarev I. A., Shandala N. K., Gulko G. M. et al, Ukrainian thyroid doses after the Chernobyl accident. *Health Phys.* **64** (1993a), 594–599.

[48] Likhtarev I. A., Prohl G. and Henrichs K., Reliability and accuracy of the $^{131}$I thyroid activity measurements performed in the Ukraine after the Chornobyl accident in 1986. GSF-Bericht 19/93. Institut für Strahlenschutz, Munich (1993b).

[49] Likhtarev I. A., Sobolev B. G., Kairo I. A. et al, Thyroid cancer in the Ukraine. *Nature* **375** (1995a), p. 365.

[50] Likhtarev I. A., Grulko G. M., Sobolev B. G. et al, Evaluation of the $^{131}$I thyroid-monitoring measurements performed in Ukraine during May and June of 1986. *Health Phys.* **69** (1995b), 6–15.

[51] Likhtarov I., Kovgan L., Vavilov S. et al, Post-Chornobyl thyroid cancers in Ukraine. Report 1: Estimation of thyroid doses. *Radiation Research* **163** (2005), 125–136.

[52] Likhtarov I., Kovgan L., Vavilov S. et al, Post-Chornobyl thyroid cancers in Ukraine. Report 2: Risk analysis. *Radiation Research* **166** (2006a), 375–386.

[53] Likhtarov I., Bouville A., Kovgan L. et al, Questionnaire- and measurement-based individual thyroid doses in Ukraine resulting from the Chornobyl nuclear reactor accident. *Radiation Research* **166** (2006b), 271–286.

[54] Likhtarov I., Masiuk S., Chepurny M. et al., Error estimation for direct measurements in May–June 1986 of $^{131}$I radioactivity in thyroid gland of children and adolescents and their registration in risk analysis. In: A. V. Antoniouk and R. V. N. Melnik (Eds.), Mathematics and Life Sciences (p. 231–244). de Gruyter, Berlin (2012).

[55] Likhtarov I., Thomas G., Kovgan L. et al, Reconstruction of individual thyroid doses to the Ukrainian subjects enrolled in the Chernobyl Tissue Bank. *Radiat. Prot. Dosimetry* **156** (2013a), 407–423.

[56] Likhtarov I., Kovgan L., Masiuk S., Estimating thyroid masses for children, infants, and fetuses in Ukraine exposed to $^{131}$I from the Chernobyl accident. *Health Phys.* **104** (2013b), 78–86.

[57] Likhtarov I., Kovgan L., Masiuk S. et al, Thyroid cancer study among Ukrainian children exposed to radiation after the Chornobyl accident: improved estimates of the thyroid doses to the cohort members. *Health Phys.* **106** (2014), 370–396.

[58] Likhtarov I. A., Kovgan L. M., Chepurny M. I. et al, Interpretation of results of radioiodine measurements in thyroid for residents of Ukraine (1986). *Problems of Radiation Medicine and Radiobiology* **20** (2015), 185–203.

[59] Little M. P., Kukush A. G., Masiuk S. V. et al, Impact of Uncertainties in Exposure Assessment on Estimates of Thyroid Cancer Risk among Ukrainian Children and Adolescents Exposed from the Chernobyl Accident. *PLoS ONE* **9** (2014), e85723. doi:10.1371/journal.pone.0085723.

[60] Lyon J. L., Alder S. C., Stone M. B. et al, Thyroid disease associated with exposure to the Nevada nuclear weapons test site radiation: A reevaluation based on corrected dosimetry and examination data. *Epidemiology* **17** (2006), 604–614.

[61] Malenko A., Efficiency comparison of two consistent estimators in nonlinear regression model with small measurement errors. *Theory of Stochastic Processes* **13 (29)**, no. 1–2 (2007), 122–131.

[62] Mallick B., Hoffman F. O. and Carroll R. J., Semiparametric regression modeling with mixtures of Berkson and classical error, with application to fallout from the Nevada test site. *Biometrics* **58** (2002), 13–20.
[63] Markovsky I., Van Huffel S. and Kukush A., On the computation of the multivariate structured total least squares estimator. *Numer. Linear Algebra Appl.* **11** (2004), 591–608.
[64] Marshall A. W., Olkin I. and Arnold B. C., Inequalities: Theory of Majorization and Its Applications. Springer, Cham (2011).
[65] Masiuk S. V., Shklyar S. V., Kukush A. G. et al, Effect of uncertainty in doses on the radiation risk estimate. *Radiation and Risk* **17** (2008), 64–75 (in Russian).
[66] Masiuk S. V., Shklyar S. V. and Kukush A. G., Impact of measurement errors in thyroid doses on dose-response analysis. *Problems of Radiation Medicine and Radiobiology* **16** (2011), 25–29 (in Ukrainian).
[67] Masiuk S. V., Shklyar S. V. and Kukush A. G., Berkson errors in radiation dose assessments and their impact on radiation risk estimates. *Problems of Radiation Medicine and Radiobiology* **18** (2013), 119–126.
[68] Masiuk S. V., Shklyar S. V., Kukush A. G. et al, Estimation of radiation risk in presence of classical additive and Berkson multiplicative errors in exposure doses. *Biostatistics* **17** (2016), 422–436.
[69] Molina E. C., Poisson's Exponential Binomial Limit. Krieger, New York (1973).
[70] Myers R. H., Montgomery D. C., Vining G. G. et al., Generalized Linear Models with Applications in Engineering and the Science, 2nd edn. Wiley, Hoboken NJ (2010).
[71] Nakamura T., Corrected score function for errors-in-variables models: Methodology and application to generalized linear models. *Biometrika* **77** (1990), 127–137.
[72] Oliver D. and Laird N., Covariance analysis of censored survival data using log-linear analysis techniques. J. Amer. Statist. Assoc. **76** (1981), 231–240.
[73] Pfanzagl J., On the measurability and consistency of minimum contrast estimates. *Mertika* **14** (1969), 249–272.
[74] Pitkevich V. A., Khvostunov I. K., Shishkanov N. K., Influence of dynamics of $^{131}$I fallout due to the ChNPP accident on value of absorbed doses in thyroid for population of Bryansk and Kaluga regions of Russia. *Radiation and Risk* **7** (2008), 192–215 (in Russian).
[75] Preston D. L., Lubin J. H., Pierce D. A. et al., EPICURE User's Guide. Hirosoft Corporation, Seattle WA (1993).
[76] Rodríguez G., Survival models, *Quantile* **5** (2008), 1–27 (in Russian).
[77] Ron E., Lubin J. H., Shore R. E. et al, Thyroid cancer after exposure to external radiation: A pooled analysis of seven studies. *Radiation Research* **141** (1995), 259–277.
[78] Rothman K. J., Greenland S. and Lash T. L., Modern Epidemiology, 3rd edn. Kluwer, Lippincott Williams & Wilkins, Philadelphia (2008).
[79] Ruark A. and Devol L., The general theory of fluctuations in radioactive disintegration. *Phys. Rev.* **5** (1936), 355–367.
[80] Schervish M. J., Theory of Statistics. Springer, Cham (1995).
[81] Schlesselman J. J., Case-Control Studies: Design, Conduct, Analysis. Oxford University Press, New York (1982).
[82] Schneeweiss H. and Mittag H.-J., Lineare Modelle mit fehlerbehafteten Daten. Physica-Verlag, Heidelberg (1986).
[83] Schneeweiss H. and Cheng C. L., Bias of the structural quasi-score estimator of a measurement error model under misspecification of the regressor distribution. *J. Multivariate Anal.* **97** (2006), 455–473.
[84] Seber G. A. F. and Lee A. J., Linear Regression Analysis. Wiley, Hoboken NJ (2003).

[85] Sen'ko I. O., Consistency of an adjusted least-squares estimator in a vector linear model with measurement errors. *Ukrainian Mathematical Journal* **64** (2013), 1739–1751.

[86] Sen'ko I. O., The asymptotic normality of an adjusted least squares estimator in a multivariate vector errors-in-variables regression model. *Theor. Probability Math. Statist.* **88** (2014), 175–190.

[87] Shklyar S. and Schneeweiss H., A comparison of asymptotic covariance matrices of three consistent estimators in the Poisson regression model with measurement errors. *J. Multivariate Anal.* **94** (2005), 250–270.

[88] Shklyar S., Comparison of estimates in the Poisson regression model with measurement errors. *Bulletin of the University of Kyiv, Series: Physics and Mathematics* **2006**, no. 3 (2006), 60–67 (in Ukrainian).

[89] Shklyar S., Schneeweiss H. and Kukush A., Quasi score is more efficient than correted score in a polynomial measurement error model. *Metrika* **65** (2007), 275–295.

[90] Shklyar S. V., Consistence of an estimator of a polynomial regression with a known variance relation for errors in the measurement of the regressor and the echo. *Theor. Probability Math. Statist.***76** (2008), 181–197.

[91] Stefanski L. A. and Carroll R. J., Conditional scores and optimal scores for generalized linear measurement error models. *Biometrika* **74** (1987), 703–716.

[92] Stefanski L. A., Unbiased estimation of a nonlinear function of a normal mean with application to measurement error models. *Comm. Statist. Theory Methods* **18** (1989), 4335–4358.

[93] Stewart G. W. and Sun J., Matrix Perturbation Theory. Academic Press, Boston (1990).

[94] Stulajter F., Nonlinear estimators of polynomials in mean values of a Gaussian stochastic process. *Kybernetika* **14** (1978), 206–220.

[95] Talerko N., Mesoscale modelling of radioactive contamination formation in Ukraine caused by the Chernobyl accident. *J. Environmental Radioactivity* **78** (2005a), 311–329.

[96] Talerko N., Reconstruction of $^{131}$I radioactive contamination in Ukraine caused by the Chernobyl accident using atmospheric transport modelling. *J. Environmental Radioactivity* **84** (2005b), 343–362.

[97] Tronko M. D., Howe G. R., Bogdanova T. I. et al, A cohort study of thyroid cancer and other thyroid diseases after the Chornobyl accident: Thyroid cancer in Ukraine detected during first screening. *J. Natl. Cancer. Inst.* **98** (2006), 897–903.

[98] Tsyb A. F., Parshkov E. M., Shakhtarin V. V. et al., Thyroid cancer in children and adolescents of Bryansk and Kaluga regions. In: Proceedings of the first International Conference "The Radiological Consequences of the Chernobyl Accident", Minsk, Belarus, 691–698 (1996).

[99] Wansbeek T. and Meijer E., Measurement Error and Latent Variables in Econometrics. North Holland, Amsterdam (2000).

[100] Zablotska L. B., Ron E., Rozhko A. V. et al, Thyroid cancer risks in Belarus among children and adolescents exposed to radioiodine after the Chornobyl accident. *British Journal of Cancer* **104** (2011), 181–187.

[101] Zvonova I. A., Balonov M. I., Bratilova A. A. et al., Thyroid absorbed dose estimations for population of the Bryansk, Tula, Orel regions according to results of radiometry in 1986. *Radiation and Risk* **10** (1997), 95–116 (in Russian).

# Index

## Also of Interest

**Volume 4**
Sergey Vakulenko
*Complexity and Evolution of Dissipative Systems: An Analytical Approach, 2013*
ISBN 978-3-11-026648-1, e-ISBN 978-3-11-026828-7, Set-ISBN 978-3-11-026829-4

**Volume 3**
Zoran Nikoloski, Sergio Grimbs
*Network-based Molecular Biology: Data-driven Modeling and Analysis*
ISBN 978-3-11-026256-8, e-ISBN 978-3-11-026266-7, Set-ISBN 978-3-11-916541-9

**Volume 2**
Shair Ahmad, Ivanka M. Stamova (Eds.)
*Lotka-Volterra and Related Systems: Recent Developments in Population Dynamics, 2013*
ISBN 978-3-11-026951-2, e-ISBN 978-3-11-, Set-ISBN 978-02698-5-7

**Volume 1**
Alexandra V. Antoniouk, Roderick V. N. Melnik (Eds.)
*Mathematics and Life Sciences, 2012*
ISBN 978-3-11-027372-4, e-ISBN 978-3-11-028853-7, Set-ISBN 978-3-11-028854-4