Data contamination versus model deviation

by

Viviane Grunert da Fonseca

A thesis submitted to the University of Sheffield
in candidature for the degree
of
Doctor of Philosophy

Department: Probability and Statistics
submitted: February 1999

Summary

Statistical inference is based on two sources of information: empirical data and assumptions which are presented in the form of a statistical model. Inference procedures commonly process the data and (together with the model assumptions) produce the desired description of the unknown aspect of interest. Ideally, statistical inference is carried out with *correct* information and inference procedures which offer corresponding 'optimal' performance. However, this is rather seldom the case and concern in the statistical literature is more and more devoted to inferential performance under *distortion*. The most relevant subject areas are "robustness" and "diagnostics".

The present work develops a basic classification scheme for distortion in the framework of classical statistical inference. In particular, it emphasizes the still outstanding and consequent distinction between *data contamination* and *model deviation*. The claim that statistical inference is ultimately aimed at real-world description, and not data description, is in this respect of key importance. Concern in this work is further given to the *performance* of statistical inference procedures under distortion. It is explored when *differences* in performance under data contamination and model deviation are possible and how these can be detected. Methodology is developed for the study of inferential performance under *increasing* distortion. A critical review of some important references in the robustness and diagnostics literature moreover indicates which approach towards inferential performance assessment is aimed at data contamination and which at model deviation. The concepts and conclusions of the thesis are finally illustrated by two detailed simulation examples. The first studies the performance of the Abdushukurov-Chen-Lin (ACL) estimator under increasing distortion from the Koziol-Green proportional hazards model, and the second in like manner considers estimation problems related to a parametric linear regression model with correlated errors for longitudinal data.

Acknowledgements

I would like to thank my supervisor Dr. Nick Fieller for his constant help and patient support throughout this research work. Many thanks also to the Department of Probability and Statistics, University of Sheffield, for creating a friendly and personal atmosphere, both in professional and in social terms. I am grateful to my colleagues for their friendship, particularly Sofia Lopes, Sivarajalingam Janarthanan, Coomeran Vencatasawmy, Mojgan Naaeni, Vedide Rezzan Uslu, Andrew Kyprianou, and Ben Gordor, and to Valceres Silva and Illy Khatib also for their hospitality.

An acknowledgement is due to Prof. Ursula Gather, University of Dortmund, Germany, for first introducing me to the Koziol-Green model considered in chapter 5.

I thank my parents for their love and support, my brothers Veit Peter and Simon Peter for encouragement, and my husband and colleague Carlos Fonseca for his dedication and love, and for several enlightening discussions. A special word of appreciation goes to my parents-in-law, Maria Liseta and António Manuel, for their great and dedicated help on several occasions of need.

Finally, I wish to acknowledge the Graduiertenkolleg of the Department of Statistics, University of Dortmund, Germany, and subsequently the bursary scheme of the University of Sheffield in collaboration with the Department of Probability Statistics, University of Sheffield, for financial support.

Statement of Originality

Unless otherwise stated in the text, the work described in this thesis was carried out solely by the candidate. None of this work has already been accepted for any other degree, nor is it being concurrently submitted in candidature for any degree.

Candidate: _____

Viviane Grunert da Fonseca

Supervisor: _____

Nick Fieller

Für meine Eltern, und für meine Brüder Veit Peter und Simon Peter.

Ao Carlos e ao Filipe.

Contents

Chapter 1

Introduction

1.1 Motivation

The concern of *robust statistics* seems to be defined with some ambiguity. Following a key-reference by Hampel *et al.* (1986) we learn the following:

> *"Robust statistics, in a loose, nontechnical sense, is concerned with the fact that many assumptions commonly made in statistics (such as normality, linearity, independence) are at most approximations to reality"* (p. 1).

Hence, we conclude that robust statistics deals with incorrect *assumptions*. However, the next sentence in (Hampel *et al.*, 1986) already causes doubts:

> *"One reason is the occurrence of gross errors, such as copying or keypunch errors"*.

This rather refers to defects in the *data*, which are the second source of information for statistical inference. Also the key-reference by Huber (1981) does not clear our incomprehension. He defines robustness as "insensitivity to small deviations from the assumptions" (p. 1) and remarks later that "a

primary goal of robust procedures is to safeguard against gross errors" (p. 5). We still cannot understand why wrong (prior) assumptions and gross errors in the data should relate to the same thing.

Hampel *et al.* (1986) claim that "robustness theories as such cannot be applied in a nonparametric situation" (p. 9) because "all of the present theories of robustness consider deviations from the various assumptions of parametric models" (p. 7). However, they also admit that the concepts can "still provide valuable insight into the behaviour of nonparametric methods" (p. 9) when used in parametric situations. The arithmetic mean, for example, is shown to be non-robust against outliers relative to the normal model (p. 88f). But would we relax using the arithmetic mean in a non-parametric situation just because we have *not* made any distributional assumptions? If not, the arithmetic mean must (also) be sensitive to something other than the wrong normal *assumption*. In fact, an outlier may reflect "an inadequate basic model" assumption or may be due to a "measurement or execution error" in the data (Barnett and Lewis, 1995, p. 34), and it is indeed the *second* fault which concerns us when using the arithmetic mean in a non-parametric situation.

Hence robust statistics seems to refer to both, wrong assumptions *and* wrong data, though not 'officially'. This quite naturally leads us to the following questions: Is there *really* a difference between the two situations of *distortion*? And maybe even more important: Can it *make* a difference in terms of the performance of statistical inference procedures? In other words: Does 'robustness against wrong model assumptions' and 'robustness against wrong data' always mean the same thing? If not, *when* not?

The work in this thesis tries to give answers to the questions above from the point of view of classical inference. It will define the distinction between the two situations of distortion, from now on called *model deviation* and *data contamination*, and explore when and how such a distinction becomes relevant for inferential performance.

1.2 Summary of the thesis

With the claim that the real-world (and not the data) is the ultimate aim and reference point for statistical inference, three types of distortion are classified. Two of them can be seen as 'antonyms' and are emphasized, representing either model deviation or data contamination for the *same* potential data-generating process. We present methodology for the study of inferential performance under increasing distortion, and explore when differences in performance are possible (under the two types of distortion) and how they can be detected. Important approaches for robustness and diagnostics are critically reviewed, and two simulation studies illustrate the overall concepts and conclusions.

Chapter 2 presents the notions of *distortion* and *performance* in statistical inference as currently seen in the literature. It is meant to motivate and prepare for the identification of ideas developed in later chapters. The two most relevant subject areas dealing with distortion are "robustness" and "diagnostics". After their brief introduction, several citations illustrate the existing views of distortion in the literature. It is concluded that a distinction between data contamination and model deviation seems to have mostly been ignored. The chapter continues with the presentation of general concepts formalizing distortion and possibilities for distortion with some common statistical models. A few typical or especially recent references are provided. The second part of the chapter is devoted to the performance of statistical inference procedures. The various ideas are discriminated by the formal representation of such procedures and the preferred attribute of 'good' performance. Performance descriptions based on (so called) performance statistics are given particular attention. Some aspects of performance under distortion (as dealt with in the literature) are finally outlined, while the main discussion is postponed to the end of chapter 4.

Chapter 3 describes our view of distortion. It begins with a prelude claiming that statistical inference is aimed at real-world description (and not data

description). In further preparation the *model* and *inference framework* are introduced. The former recalls the way from a real-world situation to the data-generating process via the *aspect of interest* and the (set of) *data units*. Several examples accompany the description. It further employs a threefold interpretation of the data-generating process referring to real-world, data, and model assumptions, which correspondingly leads to three individual representations through statistical models (the *model triplet*). The inference framework relates the logical elements of a statistical inference process, emphasizing the roles of inference procedure(s) and model assumptions. The remainder of the chapter is devoted to the new classification of distortion. With attribution to the inference framework, disagreement within the model triplet is identified as three different types of distortion. Two affect aspects of the model 'within data units' and represent comparable situations of data contamination and model deviation. The third type relates to aspects of the model 'between data units' as another separate form of model deviation. Examples from an earlier stage are re-considered to illustrate and discuss the findings.

Chapter 4 develops methodology for the assessment of inferential performance under increasing distortion. In a preparatory section potential implications of data contamination and model deviation are considered. When are differences possible and how can they be detected? The concept of model expansion is suggested to quantify distortion of the three types. Moreover, a new notation for performance statistics is introduced to underline the reference to real-world, data, and model assumptions. A subsequent section presents the idea of *influence* and *preference graphs*. They study the change of performance under increasing distortion and likewise compare performance among combinations of inference procedure(s) and/or assumptions. The final part of the chapter is devoted to a critical review of important approaches for robustness and diagnostics in the literature. Independently from what has been claimed, what do they actually consider – data contamination or model deviation?

Chapter 5 presents the first simulation example. It studies performance of the Abdushukurov-Chen-Lin (ACL) estimator (Abdushukurov, 1984; Cheng and Lin, 1984) under distortion from the semi-parametric Koziol-Green (KG) proportional hazards model (Koziol and Green, 1976). The ACL-estimator is the maximum-likelihood estimator for the true survival function (*aspect of interest*) under the KG-model. The latter is used in the censored survivals and competing risks framework. Distortion affects the independence requirement between the observed survival times and the censoring 0-1-indicator (and thus relates to aspects of the model *within* data units). The corresponding situations of data contamination and model deviation are distinguished. Distortion as such is modelled in a novel way using the concept of model expansion. The relationship to other formalization approaches for the same kind of distortion is considered and the simulation set-up described. The study itself is based on influence and preference graphs where the latter compare the ACL-performance with that of the competing Kaplan-Meier estimator (Kaplan and Meier, 1958). Overall, a trade-off in performance between corresponding situations of data contamination and model deviation is exhibited.

Chapter 6 is devoted to the second simulation example. It considers estimation problems related to a parametric linear regression model with correlated errors for longitudinal data (Diggle *et al.*, 1995). Distortion affects the (sub)model describing the covariance/correlation structure between observations of the same data unit (and thus relates to aspects of the model *within* data units). Measurement errors are comprised in the *distorted* model and are ignored in the corresponding ideal reference model. Corresponding situations of data contamination and model deviation are again discriminated. The variogram and mean response profile (Diggle *et al.*, 1995) are considered to be the two (main) *aspects of interest*. Performance of their estimators is the prime objective of the simulation study. In addition, the individual parameter estimators for the variogram are examined. After the description of the simulation set-up the results are discussed. Influence

graphs show in which case performance under data contamination and model deviation can be distinguished. Preference graphs compare the respective performances under the two model assumptions which do or do not include the measurement error component.

Chapter 7 presents the overall conclusions and gives some directions for future work.

1.3 Contributions

- The identification of inconsistencies in the meaning of *distortion* in the statistical literature and the subsequent formal distinction between *data contamination* and *model deviation* through a novel classification scheme of distortion (chapters 2 and 3): We claim that statistical inference is aimed at real-world description. The real-world, data, and model assumptions are made comparable by associating them with respective data-generating processes which in turn are represented by statistical models forming the *model triplet*. Different conditions of disagreement within the model triplet define distortion types ① to ③. The first two affect aspects of the model 'within data units' and are identified as *comparable* situations of data contamination and model deviation. Type ③ represents a separate form of model deviation affecting aspects of the model 'between data units'.

- The discussion of seven concise examples to illustrate (finally) the new classification of distortion (chapter 3).

- The identification of conditions for which the performance of statistical inference procedures is different under data contamination and type ② model deviation (chapter 4): A new notation for performance statistics underlines the reference to real-world, data, and model assumptions. A difference in inferential performance is generally *possible* if distortion refers to the aspect of interest as represented in the statistical model.

It can be *detected* if the performance description refers to the aspect of interest as part of the real-world (as e.g. the bias of a point estimator).

- The formulation of influence and preference graphs which study the performance of statistical inference procedures under increasing distortion of types ① to ③ (chapter 4): The *discrepancy magnitude* of a model defines the distance to some ideal reference model. The model triplet associated with three such magnitudes serves as the underlying 'moving' reference frame for a performance statistic. This determines the type and current amount of distortion for each graph.

- The critical review and subsequent relation of current robustness and diagnostics approaches to either data contamination or model deviation, independently from what has been claimed in the respective references (chapter 4).

- A new formulation of the *distorted* Koziol-Green proportional hazards model (chapter 5): Invalidation of the characteristic independence requirement between observed survivals and censoring indicator is expressed in a novel way via model expansion. The new idea is brought into relation with other existing approaches.

- The completion of two detailed simulation studies addressing the effects of data contamination and model deviation on two specific estimation problems of representative character (chapters 5 and 6): The new ideas and methodology are broadly illustrated and implemented using the program environment S-PLUS.

- The specification of a 'duality-problem' for the model assumptions (chapter 7): Assumptions need to be directed to the real-world for nominal inference and in addition to the data for stochastic inference.

Chapter 2

Distortion and performance in the statistical literature

2.1 Introduction

Under pressure of time there is frequently the wish to analyse data as quickly as possible by using well-implemented standard methods such as classical regression. Many studies, however, do not meet the basic requirements associated with these procedures. For example, there are outliers in the data and the assumption of a normal distribution does not hold. In other words, one is often faced with a situation of *distortion*.

This defect in the statistical setup can affect many nice properties of the inference procedures involved. In the case of a linear regression model, for instance, failure of the normal assumption can ruin the well-known UMVU-property (uniformly minimum-variance unbiased) of the maximum-likelihood estimator (Graybill, 1961, p. 113f). Overall, it is therefore the *performance* of statistical inference procedures which potentially suffers under distortion.

This chapter will consider the aspects *distortion* and *performance* as they appear in the statistical literature. Having briefly introduced the two most relevant subject areas called *diagnostics* and *robustness* in section 2.2, the

concept of distortion will be discussed in section 2.3. Several citations therein will point out the different existing views. In addition, general concepts of formalizing distortion will be presented as well as possibilities to relate distortion to some common statistical models. The chapter will continue with a discussion of the performance of statistical inference procedures (section 2.4). After some general background information in order to discriminate between the various ideas, performance descriptions based on (so called) performance statistics will be presented. The chapter will close with a short discussion about related aspects under distortion.

2.2 Diagnostics and robustness

Statistical publications referring to the idea of *distortion* can most often be associated with at least one of the two subject areas *diagnostics* and *robustness*. While both areas are aimed at sensible inferences in the presence of distortion, their methods are being developed from a different point of view (Stahel and Weisberg, 1991, p. xi). See also Huber (1991).

In diagnostic studies, distortion is (first) tried to be identified. This should enable the decision of subsequent actions such as changing the assumptions or the data, or choosing an alternative inference procedure. Typical examples are discordancy tests for outliers which allow the detection, and afterwards if desired, the rejection of outliers (Barnett and Lewis, 1995). Further there are the so called perturbation or influence diagnostics which have been mainly applied in regression analysis. A motivating reference in this particular area is Cook (1986).

Contrastingly, robust statistics *directly* seeks new inferential methods which are insensitive to, or robust against, potential distortion. In line with this idea are inference procedures which accommodate outliers (Barnett and Lewis, 1995). Overall, the main approaches to robustness are the following:

- The notion of (asymptotic) *qualitative robustness* going back to Ham-

pel (1971) is motivated by the idea that small changes in the underlying sample only produce small changes in the corresponding inference results. This requirement is also known as *resistance*. The formal definition for qualitative robustness itself can be related to the continuity notion of functionals (Huber, 1981, p. 9f).

- Huber's (1981) asymptotic *minimax-approach* considers quantitative aspects (quantitative robustness). Here, a robust estimator needs to show maximum performance (in comparison to alternative estimators) for the "least-favourable distribution minimizing the Fisher-information in a chosen distortion neighbourhood" (Huber, 1981, p. 73).

- The *approach based on influence functions* concentrates on infinitesimal distortion (therefore also *infinitesimal approach*). Robustness criteria are derived from the influence function which, roughly speaking, corresponds to the first derivative of a functional (inference procedure) at some ideal model distribution. The latter is embedded in the space of all probability distributions contemplating an unrestricted full, though infinitesimal, distortion neighbourhood. In addition to the influence function itself, the concepts of qualitative robustness (see above) and that of the breakdown point are important (Hampel *et al.*, 1986).

- *Configural polysampling* is a small sample theory of robustness which is aimed at invariant models such as the ones of location/scale or regression type. In a first step a few "over-diverse" alternatives are selected which reflect the type of deviation considered such as the Gaussian or some heavy-tailed distribution. Then, inference procedures with good performance at these situations are constructed in the hope that their behaviour is also similar at intermediate cases. See Morgenthaler and Tukey (1991) and Morgenthaler (1991).

- Huber's *capacities approach* to robust testing and confidence intervals produces "exact finite sample results". It robustifies the Neyman-Pearson lemma and yields interval estimates of location. While being

mathematically deep and elegant, the approach is less successful in terms of general applicability (Huber, 1981; Hampel *et al.*, 1986).

- Rieder's (1994) approach to robust asymptotic statistics deals with "optimally robust functionals and their unbiased estimators and tests" by "linking up nonparametric statistical optimality with infinitesimal robustness criteria". It addresses the following two questions: "Which functional to chose?" and "Which statistical procedure to use for the selected functional?" (p. vii).

Apart from statistical inference procedures the term "robust" is also employed for objects such as parameters (Godambe and Thompson, 1984) and samples (Cook, 1986; Barnard, 1980). In addition, expressions such as "model robustification" and "model robustness" are known in the area of Bayesian statistics (Box, 1980; Draper and Parmigiani, 1995).

Instead of now getting involved with the actual methods of diagnostics and robustness, attention will be directed to the following, more fundamental aspects:

- What is the kind and nature of potential distortion, and
- what are the means of assessing the performance of statistical inference procedures (under distortion)?

Thus, the remainder of the chapter is devoted to corresponding views in the statistical literature. Different approaches towards explaining distortion will be presented (§ 2.3), and the way in which the performance of statistical inference procedures has been described up to now will be discussed (§ 2.4).

2.3 Distortion

2.3.1 Meaning and origin

Several types of *error* can occur in statistical inference. First of all, there is the inherent, random error (statistical error) associated with every sample based inference procedure. Approximately standard normal distributed, it is the repeated sum of elementary random measurement errors (central limit theorem), or it can be associated with the fact that the chosen random sample is not absolutely representative. Other random, not necessarily normal distributed errors are also possible, for example due to (unbalanced) rounding or grouping. Errors might also systematically contaminate the data such as regular recording errors. Finally there could be specification errors (mistakes) in the model assumptions, since the latter are nothing else than 'human made' simplifications of nature.

In the present work, any situation which inherits an error in either the data or the assumptions will be denoted by the term *distortion*. This includes all the errors mentioned above, except the statistical error due to unrepresentative, but still correct samples. The corresponding errors may be of any size. Alternatively referring to Huber (1981, p. 1):

> *Distortion shall mean any kind of deviations from the assumptions,*

where the latter can be parametric, semi- or non-parametric. In this way distortion could imply *data contamination*, i.e. wrong data which deviate from the correct assumptions (due to recording errors, for example). Or, it could be *model deviation* (model misspecification), i.e. wrong model assumptions where the truth deviates from the assumptions (or better: the assumptions deviate from the truth). Both concepts of distortion will be introduced in detail in chapter 3.

Distortion as a concept has a long pedigree in a variety of descriptions and interpretations in the statistical literature. This is in particular reflected

in the 'target' of distortion which at times even seems to be unclear: Does distortion operate on the data, the assumptions, or in turn possibly on both? Some examples from the literature illustrate the diversity:

1. The following references claim to consider errors in the data:

   - Donoho and Liu (1988, p. 557f) refer to a measure of distortion denoted as $\delta = \delta(P, P_0)$, where P_0 is the 'ideal' distribution which "holds for physical or other reasons" and where "the real data ... have a distribution P distorted through gross-errors, nonlinearities of measurement, rounding errors and other factors outside our control".
   - Millar (1981) explains that "due to contaminations ..., the data actually collected ... follow a distribution that is ... possibly distinct from the P_θ's" belonging to "a fixed parametric family" (p. 73). And, this family "is to be regarded as a theoretically correct probabilistic model of the phenomenon at hand; ... perhaps forced on us by accepted principles of theoretical physics" (p. 74).

2. The following references claim to consider errors in the model assumptions:

   - Stahel and Weisberg (1991, p. xi) introduce the theme of robust statistics and diagnostics by referring to the situation when "... the parametric model is not completely correct" or " ... is not correctly specified".
   - Morgenthaler (1991, p. 49) is "concerned with the stability of inference procedures with regard to distributional and other structural assumptions".
   - According to Cook (1986, p. 133) "models, ... are nearly always wrong", and as a consequence one is often faced with "model perturbation".

3. The following references refer to errors in *both*, data and assumptions. Even though the latter are mentioned separately, their distinction does not lead to different methodological considerations.

   - In the introduction of Copas (1988) topics such as robustness, outliers, leverage, diagnostics, and resistant fitting are mentioned in relation to "bad data and misspecified models". The remaining discussions therein, however, deal with outliers in the sense of *contaminated data* – as the title of the paper already specifies.
   - Barnett and Lewis (1995, p. 21) explain that outliers "...may reflect deterministic factors (errors of measurement, misrecording, etc.) or be probabilistic in nature (causing us to question distributional assumptions)". In other words, they can be due to "measurement errors", "execution errors" or unrecognized "inherent variability" (p. 33f). While the two former error types obviously affect the data, the latter as a "natural feature of the population" (p. 33) may reflect an "inadequate basic model" (p. 34), i.e. wrong model assumptions. The authors acknowledge the fact that the omission of data contaminating outliers might be sensible, but that it "is hardly a robust policy" when "outliers arise because our initial model does not reflect the appropriate degree of inherent variation" (p. 36). Still, they do not intend to distinguish "in terminology" between those "sources of variation" (p. 34). "Of course, we have no way of knowing whether or not any observation is a contaminant. All we can do is concentrate attention on outliers as *the possible manifestation of contamination* ..." (p. 9).

4. In the references below the target of the errors seems to be unclear. In the first two publications both types of errors are mentioned and this independently from the immediate context and just as if the errors in data and assumptions were the same. The last reference uses a neutral formulation of distortion.

- Hampel (1971, p. 1887) deals with situations where "the parametric model is not quite true", i.e. when the assumptions are in error. Nevertheless, the "reasons for deviations from the parametric model" are according to him: "(i) rounding of the observations; (ii) occurrence of gross errors" – referring to errors in the data – and "(iii) the model itself may only be an approximation of the underlying chance mechanism" – which seems to mean wrong (model) assumptions.
- Huber (1981, p. 1) talks about the situation when "assumptions are not supposed to be exactly true" and when there are "deviations from the assumptions". However, he later explains that "the occurrence of gross errors in a ... fraction of the observations is to be regarded as a ... deviation" (p. 5).
- He and Simpson (1993, p. 314) use the vague expression "contamination of F_θ".

The references above support an important conclusion: The potential difference between data contamination and model deviation is often not consequently emphasized in the statistical literature. Rather the opposite seems to be true, i.e. a distinction as such has mostly been ignored. Only a few publications could be traced which indeed seem to stress the latter:

- In Cabrera *et al.* (1997) two aspects of robustness are mentioned: "(I) model robustness" ..."where the reality does not exactly agree with the assumed model" and "(II) data robustness" where there are outliers in the data. The overall paper, however, deals only with the second kind of 'robustness'.
- "Closer study of the idea of perturbations suggests that it is important to distinguish between those of the *data* and those of the *model*" (Billor and Loynes, 1993, p. 1595). In line with this thought the authors develop a new method for regression diagnostics.

- In the context of diagnostics Lawrance (1991) discusses "perturbations", where "three sorts" are "distinguished: Perturbations to assumptions, perturbations to data values and perturbations to case weights". With such a *threefold* distinction the author's intention still seems to be different.

- Hettmansperger and Sheather (1992) refer to robustness when "... the true underlying model is in a neighbourhood of the assumed model" (p. 145) and to resistance when there is "a small amount of data contamination" (p. 146). Then, however, they quantify resistance with the influence function and the breakdown point, both concepts from the infinitesimal *robustness* approach. Hence, the distinction becomes less clear again (though the authors point into the right direction, see the discussion in § 4.4).

The distinction between data contamination and model deviation will also find only little notice in the remaining discussions of this chapter which presents ideas mainly as they appear in the literature. Yet, as one of the main points of the present work, this problem will be treated at length in chapter 3.

2.3.2 Formalization concepts

Most formalization concepts of distortion are based on the idea of a probability model. This is obvious for incorrect model assumptions and randomly contaminated data, but also sensible in deterministic situations. According to Barnett and Lewis (1995, p. 32f) for example, many deterministic outliers cannot be traced back to their origin, and are therefore regarded as random.

The ideal model distribution representing a situation of no distortion (which in cases could be a univariate distribution or a more complex model) can be 'extrapolated' in various ways. While some approaches are mainly of mathematical character, others appear to be more heuristic and are

therefore easier to apply. Whether such a concept is realistic or not, nevertheless, depends on the flexibility and size of the corresponding model or neighbourhood. In this respect a range of possibilities can be found in the literature.

The known approaches towards formalizing distortion will be briefly discussed now. Note, that a review similar to the following is also presented in Hampel *et al.* (1986, p. 8ff) and Ronchetti (1997, p. 60f).

2.3.2.1 Finite number of alternatives

The simplest way of describing distortion is by taking a finite number of alternative models which are more or less different from the ideal one representing a situation of no distortion. Many studies which compare the performance (robustness) of certain estimators, tests, etc. are based on this approach. The most famous application is certainly the Princeton study by Andrews *et al.* (1972), where properties of a large number of location estimators were compared under "about 40 different sampling situations" which "were at least as long-tailed as the normal distribution" (p. 67). In addition the methodology known as *configural polysampling* (see § 2.2) uses selected alternatives in order to develop robust inference procedures. Finally, a single, so called, *inherent* alternative can be used to explain outliers. The latter may be a different fully specified distribution or it may represent a distinct parametric family (Barnett and Lewis, 1995, p. 46).

2.3.2.2 Model expansion

Also intuitive seems to be the method of model expansion where, in order to explain distortion, further (distortion-) parameters are added to the ideal model. Hampel *et al.* (1986, p. 9) call the resulting enlarged model a "supermodel". A normal regression model, for instance, could be extended by considering the t-distribution with k degrees of freedom for the errors. The $N(0,1)$-error distribution is then 'included' as the limiting distribution

for $k \to \infty$ (see e.g. in Lange *et al.* (1989) for a multivariate version and the references therein). Other examples are self-similar processes which model long-term correlations by adding a so called self-similarity parameter to the original i.i.d. model (independent identically distributed). See for example Hampel *et al.* (1986), p. 8 and p. 389ff. Finally note, that also our approach will be based on this idea (chapter 4). Corresponding detailed examples will be discussed in chapters 5 and 6 where the ideal model is even 'properly' embedded in the supermodel.

2.3.2.3 Model mixing

The idea of model mixing leads to distortion models known as

- the *gross-error model* (also ϵ-contamination neighbourhood) which is used in the minimax approach to robustness initiated by Huber (see § 2.2). It contains all distributions 'around' some ideal distribution F_0 which are composed of a mixture of F_0 and some other arbitrary distribution H according to a ratio of $(1-\epsilon)$ to ϵ (Huber, 1981, p. 11).
- *mixture models* where "outliers reflect the (small) chance λ that observations arise from distribution G, quite different from the initial model F" (Barnett and Lewis, 1995, p. 46ff).

Note, that unlike the mixing distribution H in the gross-error model, the distribution G in an (outlier) mixture-model is considered to be fixed.

2.3.2.4 Neighbourhoods based on a distance

Distortion neighbourhoods around some ideal model distribution can be defined based on an underlying distance such as the Lévy, Prohorov, total variation, or the Kolmogorov distance. Also goodness-of-fit measures can be used as for example one of Cramér-von Mises type (Huber, 1981; Donoho and Liu, 1988).

2.3.2.5 Infinitesimal neighbourhoods

In the infinitesimal approach towards robustness, an inference procedure is replaced at the ideal model by a linear approximation based on the corresponding influence function. While the approximation itself 'reaches' into the full neighbourhood of all probability distributions, robustness behaviour is only studied at the ideal model. This (indirectly) defines a "full but infinitesimal neighbourhood" (Hampel *et al.*, 1986, p. 41f, 273).

2.3.2.6 Distortion as changes in the sample

The concept of qualitative robustness is motivated by distortion which is simply seen as small changes in the underlying sample, i.e. "...small changes in all of the observations x_i (rounding, grouping) or large changes in a few of them ..." (Huber, 1981, p. 9). A model distribution as such is not involved for this kind of distortion.

2.3.2.7 Slippage models

Slippage models aim at describing the occurrence of outliers. They suggest that "all observations apart from some prescribed small number k ...arise independently from the initial model F ..., whilst the remaining k are independent observations from a modified version of F". The so called *exchangeable models* are also related to this idea (Barnett and Lewis, 1995, p. 49ff).

2.3.3 Distortion with some particular statistical models

Distortion is usually considered in relation to a statistical probability model (see above). The more complex this model, the larger is the number of (model) aspects which distortion could address. The following paragraphs will present theoretical possibilities for some common statistical models and

point to a few typical or especially recent references. Several of the former and many more can be found in a survey paper by Stahel (1991). Note that it will remain to be seen whether distortion characterized in the following ways is also likely in practical situations.

2.3.3.1 Simple stochastic models

A simple stochastic model is characterized by random variables $X_1, \ldots, X_n$ which are i.i.d. like X with $X \sim F_\theta$ ('distributed like') or $X \sim F$. This includes all parametric, semi- or non-parametric models with a single (univariate or multivariate) distribution and no covariate information. Potential distortion of the model could affect

- the independence between the $X_1, \ldots, X_n$ and their distribution equality (homogeneity),
- the distribution of X itself with respect to distribution type or shape, continuity, symmetry, parameter value, and dependence structure of its components (if X is multivariate).

Deviations from independence have been described in terms of moving-average schemes in Portnoy (1977; 1979). Hampel (1986) discusses the use of self-similar processes (see § 2.3.2.2) which are especially suitable for long-range dependencies, and Künsch (1991) mentions various stationary processes in his discussion on "robustness against dependence". Recent work with "dependent Gaussian random variables" has been published in Genton (1998).

Deviations from distribution equality can be addressed by slippage and exchangeable models as defined in Barnett and Lewis (1995, p. 49ff).

Most of the early robustness studies were concerned with *deviations from the assumed distribution of* X (distributional robustness). In particular, location and/or scale models have been widely considered. The book by

Huber (1981) is certainly one of the most important references in this area. It uses the gross-error model (§ 2.3.2.3) and distortion neighbourhoods based on a distance (§ 2.3.2.4). See also Morgenthaler and Tukey (1991) and Morgenthaler (1991) who take into account selected model alternatives for the distribution of X. Infinitesimal (full) neighbourhoods are further the essence of the well-known book by Hampel *et al.* (1986). Moreover, it is worth mentioning the outlier models of inherent and mixture type (Barnett and Lewis, 1995, p. 46ff), and the theory of robust minimum distance estimation using distortion neighbourhoods based on distance measures (Donoho and Liu, 1988). On the whole the literature in the area is immense, since also many *empirical* comparisons between individual model alternatives have to be included. For three recent contributions see Collins and Wu (1998), Wiens *et al.* (1998), and Wu and Zhou (1998). While the first paper compares M-estimators in asymmetric ϵ-contamination neighbourhoods of a symmetric unimodal distribution, the last two study L- and M-estimators in Kolmogorov neighbourhoods of the normal distribution, respectively. Finally, note Copas and Stride (1997) who study a local maximum likelihood estimator which adapts "to local departures from the assumed model" (in their paper the normal distribution) and the related paper by Eguchi and Copas (1998).

Violations of *continuity* are addressed by the local-shift sensitivity (Hampel *et al.*, 1986).

2.3.3.2 Linear regression models

In a linear regression model the random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ of response variables is modelled according to

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.1}$$

where X is a $(n \times p)$–matrix of non-random explanatory variables (carriers), $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a vector of parameters, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$ represents the random errors. While the term $X\boldsymbol{\beta}$ describes the structural (deterministic)

part of the model, $\boldsymbol{\epsilon}$ constitutes the distributional part – usually with $\epsilon_1, \ldots, \epsilon_n$ i.i.d. like $\epsilon \sim N(0, \sigma^2)$.

Related models are for example random-carrier models, analysis-of-variance (ANOVA) models, where the design matrix X contains only the values 0 and 1, errors-in-variables models, where X is observed with superimposed errors, random effect models, where $\boldsymbol{\beta}$ is (partially) random, and the Cox-regression model for survival times. Potential distortion of the model in (2.1) could affect the

- *structural part* of the model, i.e. deviations from linearity and wrong entries in $\boldsymbol{\beta}$ or X.
- *distributional part* of the model, i.e.
  - the independence between the $\epsilon_1, \ldots, \epsilon_n$ and their distribution equality, e.g. hetero- instead of homo-skedasticity,
  - the distribution of ϵ itself with respect to distribution type and shape, continuity, and parameter values.

The related models above could suffer further (other) possibilities of distortion which will not be discussed here.

For distortion with respect to the *structural part* of the regression model the following references can be noted: Diagnostics for perturbations of the explanatory variables are discussed in Cook (1986) and later e.g. in Lawrance (1991). Departures from linearity have been considered in the context of robust designs by Huber (1981, minimax approach) and recently by Wiens and Zhou (1997, infinitesimal approach). Interestingly, Stahel (1991, p. 245) remarks in his review paper, that he is not aware of any recent research on "*deviations from linearity* in regression".

Deviations from independence of the error terms is studied in Künsch (1991) using an AR(1)-autoregressive process, while Lawrance (1991) addresses this aspect within the framework of perturbation diagnostics. A comprehensive

paper is also presented by Künsch *et al.* (1993) giving several examples for long-range correlation structures. The robust designs in Wiens and Zhou (1997) are developed under "small departures from the assumption of uncorrelated errors".

Invalidation of the *distribution equality* is often connected with the idea of non-constant variances (hetero-skedasticity). Several references are given in Stahel (1991, p. 245). The majority of them formulate the error-variance in dependence on the explanatory variable x_i. Nanayakkara and Cressie (1991) give an overview on research related to departures from homo-skedasticity. Further, the diagnostics approach introduced by Cook (1986) can be used to study unequal variances in the standard linear model. See also e.g. Lawrance (1991). Distribution inequality in more general terms can be described by outlier models (of slippage type). Note Barnett and Lewis (1995) and the references therein.

As in the simple model case, most of the literature deals with distortion affecting the *error-distribution* itself. Several publications are based on outlier models (Barnett and Lewis, 1995). Further important references are Huber (1981, minimax-approach), Hampel *et al.* (1986, infinitesimal approach), Rousseeuw and Leroy on "Robust regression and outlier detection" (1987), Morgenthaler and Tukey (1991, configural polysampling), and Rieder (1994). Nanayakkara and Cressie (1991) refer to "departures from normality" in their overview paper. Moreover, it is worth mentioning Lawrance (1991) who is concerned with "perturbations to response values". Note papers using the idea of model expansion/alternatives for direct inference purposes such as Lange *et al.* (1989), Taylor (1992), and Morgenthaler (1994). Finally, many empirical studies can be found which compare inference procedures under a finite number of model alternatives. A recent one is for example Meintanis and Donatos (1997).

Continuity aspects are again addressed by the infinitesimal approach, see e.g. Hampel *et al.* (1986).

2.3.3.3 Time series models

A time series model is represented by a stochastic process $(X_t)_{t\in T}$ which is a series of (dependent) random variables, where usually $T = \mathbb{N}$ or $T = \mathbb{Z}$. A popular assumption is (weak) stationarity, i.e. $\mathrm{E}(X_t) = \mu$ and $\mathrm{Cov}(X_{t+h}, X_t) = \gamma(h)$. Potential distortion could address the

- marginal distribution of single X_t,
- the underlying correlation structure of the series, or
- the observation times t.

Distortion of time series models is often addressed by outlier models (Stahel, 1991, p. 256f). In general, the presence of an outlier at time t is associated with changes in the corresponding *distribution of* X_t. *Additive outliers* occur superimposed and unrelated to the underlying time series model. *Innovation outliers* are reflected in subsequent observations following the correlation structure of the series. See Barnett and Lewis (1995), and Rousseeuw and Leroy (1987) for a general reference. The potential effects of additive outliers are discussed for example in Ledolter (1991, in terms of forecast errors) and Lucas (1997, in terms of influence functions). Outliers in connection with a state space representation of a time series are studied in Taplin (1993). The author formalizes the problem by changing the distribution of the corresponding error term in the observation equation. Finally, Hampel *et al.* (1986) discuss the influence function developed for time series. See also the references therein.

Taplin (1993) models level shifts of non-stationary processes by modifying the *dependence between* X_t *and* X_{t-1}. In the case of stationarity, however, the phenomenon can again be explained by use of the marginal distribution (remark by Taplin). Hampel (1986, p. 397, 422) and Stahel (1991, p. 257) mention the problem of deviations from an assumed correlation structure but do not give any further references (neither are we aware of any recent work on this topic).

Many more complex models exist in the literature which exhibit additional possibilities for distortion. It is beyond the scope of this thesis to discuss all of them but chapters 5 and 6 will give details of two specific models related to § 2.3.3.1 and § 2.3.3.2. Having reviewed the concept of distortion, we will now continue to consider the notion of *performance* as it appears in the literature.

2.4 Performance of statistical inference procedures

In the present work a statistical inference procedure will be seen as a *descriptive* process, and thus in distinction to any decision-making procedure (compare with Barnett (1982), p. 13). Further, the overall considerations will be limited to classical inference and there to (point) estimation. Hence, other approaches to statistical inference such as Bayesian inference as well as the problem area of hypothesis testing will be neglected in the following discussions. Being aware of these restrictions, the reader may already keep in mind the potential perspectives this might offer for future research (see § 7.5).

There are several ways of describing the performance of statistical inference procedures. All of them somehow try to assess the reliability of the resulting information. Nevertheless, performance description of inference procedures (particularly estimators), can be classified by e.g. taking into account the following two criteria: the formal representation of the inference procedure under study (§ 2.4.1) – leading to the distinction of finite sample and asymptotic performance descriptions, and the preferred meaning of its 'good' performance – here in terms of performance attributes (§ 2.4.2).

Quantitative performance measures will further be considered in § 2.4.3 and some related aspects under distortion will be pointed out in § 2.4.4.

2.4.1 Representation of statistical inference procedures

In the theory of classical inference, estimating as well as testing procedures can be considered as statistics, as sequences of statistics, or in many cases as functionals. The different representations which motivate either finite sample or asymptotic performance descriptions, are explained now in more detail by putting most emphasis on functionals (general definition and their use in robust statistics). For the discussion below let $X_1, \ldots, X_n$ be a random sample of size n from a population with distribution function F.

2.4.1.1 Statistics

In the simplest case, a statistical inference procedure T_n is represented as a (real-valued) statistic, which is both a random variable $T_n(X_1, \ldots, X_n)$ and a function $T_n(\cdot, \ldots, \cdot)$ with domain $\mathbb{R}^n$ and dependent on some fixed sample size n. Finite sample considerations are based on this representation.

2.4.1.2 Sequences of statistics

A statistical inference procedure can also be seen as a sequence of statistics $\{T_n; n \geq 1\}$ where n is a running-index and as such *not* fixed. Even though T_n is still dependent on n, the inference procedure itself can be considered – from an asymptotic point of view – *independently* of any sample size n.

2.4.1.3 Functionals

Finally, a statistical inference procedure can be viewed as a statistical functional T (also: statistical function) if each T_n of the corresponding sequence $\{T_n; n \geq 1\}$ can be written as a composition of T, which is independent of n, and the empirical distribution function F_n, i.e.

$$T_n(X_1, \ldots, X_n) = T[F_n(X_1, \ldots, X_n; z)] = T(F_n)$$

for all n. The domain of T is the set of distribution functions which contains F_n for all $n \geq 1$ and F (Fernholz, 1983, p. 5). Note, that the empirical distribution function, in detail denoted as $F_n(X_1, \ldots, X_n; z)$, is interpreted as a random variable with realizations in the set of distributions functions, where z is arbitrary and *not* fixed. In other words, $F_n(X_1, \ldots, X_n; z)$ is a random distribution function. This has to be distinguished from F_n as a real-valued random variable where z is fixed, and also from F_n as a deterministic distribution function where $x_1, \ldots, x_n$ is a fixed sample. Compare with the known stochastic-deterministic features of F_n as e.g. explained in Serfling (1980, p. 56).

Statistics which, explicitly or implicitly, can be expressed in terms of F_n (and are therefore functionals) are easy to find. Keeping in mind that

$$F_n(X_1, \ldots, X_n; z) = \frac{1}{n} \sum_{i=1}^{n} \imath_{\{X_i \leq z\}},$$

where $\imath$ is the indicator function, a simple example is the linear functional

$$T_n(X_1, \ldots, X_n) = \int h(z)\ dF_n(X_1, \ldots, X_n; z) = \frac{1}{n} \sum_{i=1}^{n} h(X_i),$$

where $h(\cdot)$ is some real-valued function. Other examples are given in Serfling (1980, p. 211f) – among them M-estimators, i.e. especially maximum-likelihood estimators, and the generalized Cramér-von Mises statistic.

Functional representations are generally useful for the theoretical investigation of many statistics (Serfling, 1980, p. 58), and they play an important role in the theory of robust statistics (Fernholz, 1983, p. 6).

2.4.1.4 Functionals in robust statistics

For representations in robust statistics, inference procedures need to satisfy *additional* requirements: The functionals are generally considered to be (weakly) consistent, i.e.

$$T(F_n) \xrightarrow{n \to \infty} T(F)$$

in probability, and asymptotically normal, i.e.

$$\sqrt{n}\ [T(F_n) - T(F)] \xrightarrow{n\to\infty} N[0, V(T,F)]$$

in distribution, where $V(T,F)$ is the asymptotic variance of the corresponding inference procedure at F. See Hampel *et al.* (1986, p. 82f) and Huber (1981, p. 11). Note that the former also consider inference procedures which "can asymptotically be replaced by functionals". The asymptotic theory of von Mises (1947) ensures that plenty of inference procedures satisfy these conditions, e.g. linear functionals and M-estimators (see Serfling (1980), p. 212 and Huber (1981), p. 49f).

With these assumptions statistical inference procedures are then referred to as $T(F)$ – meaning, that they operate on data which are produced by F. The asymptotic representation is again independent from any sample size n.

The theoretical representation of a statistical inference procedure is overall only one aspect which decides on how to describe corresponding inferential performance in a potential study. Another more obvious motivation for the performance description is the (preferred) meaning of 'good' performance. In the following, possible attributes of 'good' performance will be discussed.

2.4.2 Attributes of 'good' performance

Just a few attributes can summarize the different interpretations of 'good' performance of estimators, which form the basis for the various existing finite sample and asymptotic performance descriptions in the literature. Some of the latter will be considered in § 2.4.3. The attributes, in short referred to as *closeness*, *nice type of distribution*, and *stability* are explained now in more detail.

Closeness requires, that estimation results are in some way concentrated near the unknown instance to be estimated (estimand), see for example Mood *et al.* (1974, p. 289). This can address the location and/or the spread of

- the distribution associated with the estimator as a statistic, or
- the limiting distribution associated with the estimator as a sequence of statistics.

Several performance measures are known which are aimed at quantifying the goodness of location and spread. If these are used to compare different estimators, one also refers to *efficiency* properties (measures). Finally, *closeness* could imply that estimators – as sequences of statistics – converge to the correct estimand. This interpretation leads to the qualitative property known as consistency.

A nice type of distribution refers to the distribution/limiting distribution of inference procedures viewed as statistics or sequences of statistics, respectively. In general, it is desirable that the type of distribution is well known and easy to deal with. For instance, the qualitative property of (asymptotic) normality results from this attribute of 'good' performance.

Stability requires that the quality of inference behaviour is somehow persistent under (growing) distortion. However, unless a stable procedure is also 'good' in some other way, it is usually only of little use. Therefore, the concept of *stability* generally motivates the transferring and relating of each of the two attributes *closeness* and *nice type of distribution* from ideal situations of no distortion to situations of distortion.

The distinction of the three attributes of 'good' performance and the theoretical representation of statistical inference procedures (§ 2.4.1) should help to classify the ways in which inferential performance is described in the literature. The present thesis, with its special interest laid upon the situation of distortion, will concentrate on performance in relation to the attributes *closeness* and *stability*. Besides and independently from this choice, it should be remarked that there seems to exist only little work on *stability* pertaining to a *nice type of distribution*.

A short account of some common performance descriptions aimed at the attribute *closeness* will now follow.

2.4.3 Common measures of performance

Apart from the classifications made to this point, the performance of estimators as statistical inference procedures can be described in two principal ways:

1. by using measures which quantify performance and from which qualitative properties can be derived, or
2. by directly defining (qualitative) properties such as normality or consistency.

The following discussion will concentrate on quantitative performance descriptions referring to the attribute *closeness*. Especially designed statistics, which will be called *performance statistics*, serve as the main methodological tool. While in the finite-sample case their expected values are considered as a performance measure (in dependence on the underlying sample size n), asymptotic studies concentrate on the corresponding limiting values.

Point estimation Denote the estimator and corresponding estimand as $T = \hat{\theta}$ and θ, respectively. While the former may be represented as either a statistic or a sequence of statistics, the latter should be seen as a parameter in the widest sense. Measures of performance based on the idea of *closeness* are then aimed at assessing the (limiting) distribution of $\hat{\theta}$ in terms of its *location* relative to θ, and/or in terms of *spread*.

For univariate point estimators some of the performance statistics (associated with their expected values) are

- the expected difference between the estimator and the unknown estimand, known as the *bias* of T (addresses location)

$$\mathrm{E}\left[\hat{\theta} - \theta\right],$$

- the expected squared difference between the estimator and its expected value, known as the *variance* of T (addresses spread)

$$\mathrm{E}\left\{\left[\hat{\theta} - \mathrm{E}(\hat{\theta})\right]^2\right\},$$

- the expected squared difference between the estimator and the unknown estimand, known as the *mean squared error* of T (addresses location *and* spread)

$$\mathrm{E}\left\{\left[\hat{\theta} - \theta\right]^2\right\}, \text{and}$$

- the expected absolute difference between the estimator and the unknown estimand, known as the *mean absolute deviation* of T (addresses location)

$$\mathrm{E}\left\{\left|\hat{\theta} - \theta\right|\right\}.$$

The list could be continued at length – taking into account that performance statistics and their expected values are sometimes also referred to as loss and risk functions, regardless of their 'classical' intentions. See for example Mood *et al.* (1974, p. 297). Asymptotic versions of the above performance descriptions are also common.

For multivariate point estimators generalizations of the above measures are available which are either also single-valued, or are vector-valued and accordingly accompanied with a suitable order principle. See e.g. Mood (1974, p. 351ff) for generalizations of the variance.

Curve estimation Finally, consider performance descriptions of curve estimators $T = \hat{g}(\cdot)$, where g is some unknown function such as a density,

distribution, regression, or survival function. Here, the notion of *closeness* is either viewed point-wise, or can be interpreted by addressing the estimated curve as a whole. While the former approach leads to performance descriptions as the ones listed above, the latter might be associated with measures such as

- the expected average sum of squares

$$\mathrm{E}\left\{\frac{1}{k}\sum_{i=1}^{k}\left[\,\hat{g}(X_i)-g(X_i)\right]^2\right\}.$$

  The performance statistic with $g = F$ is known as the Cramér-von Mises statistic.

- the expected supremum-norm

$$\mathrm{E}\Big\{\sup_{i=1,..,k}\big|\,\hat{g}(X_i)-g(X_i)\big|\Big\}.$$

  The performance statistic with $g = F$ is known as the Kolmogorov-Smirnov statistic.

Further performance measures might be derived from goodness-of-fit tests and other general distance measures.

So far, performance descriptions have been discussed which verify the degree of *closeness*. On their own, these are usually used in ideal situations of *no* distortion. The next subsection will now briefly outline some particularities under distortion.

2.4.4 Aspects under distortion

When distortion has to be allowed for, the notion of 'good' performance becomes more a matter of general *stability*, rather than just *closeness* under the ideal situation. The two requirements together usually turn out to be a trade-off problem since they cannot be optimized simultaneously. As a

consequence, various compromise solutions have been developed in the area of robust statistics which combine both objectives and arrive at a single criteria for 'good' performance in the face of potential distortion. Here, this line of thought is of minor importance and the interested reader may e.g. refer to the books of Huber (1981) and Hampel *et al.* (1986).

Instead, attention will be devoted to the simple fact that performance descriptions in terms of *closeness* need to be transferred to situations of distortion. This is because *stability* can generally be addressed by considering and comparing performance under ideal and non-ideal situations (§ 2.4.2). Accordingly, alternative performance descriptions (in terms of *closeness*) have been formulated for the case of distortion.

Rather than presenting specific measures of performance at this point, note the following: Most of the main approaches towards robustness rely on the use of *functionals* in order to formulate measures of performance. The reason behind this development is found in the work of von Mises (1947) who developed notions of differentiability in the space $\mathcal{F}$ of all distribution functions.

As described in § 2.3.2, distortion is generally seen in relation to an ideal model distribution F, so that it is sensible to also express the inference procedure itself in terms of F by using the representation $T(F)$. This facilitates to link the ideas of *stability* with the notion of continuity on $\mathcal{F}$, because the inference procedure is interpreted as a function on $\mathcal{F}$ rather than on some n-dimensional space of sample-outcomes. The latter principle is related to the idea of *qualitative robustness* (see Huber (1981) and § 2.2).

The theory of differentiability on $\mathcal{F}$, moreover, offers mathematical tools for the influence function and the breakdown point, both well-known concepts within the *infinitesimal approach* (see Hampel *et al.* (1986) and § 2.2). While the former is defined as a kind of first derivative, the latter is based on the idea of "the distance to its nearest pole" of a function (Hampel *et al.*, 1986, p. 40).

Finally, the fact that F can be considered as a 'point' in $\mathcal{F}$ allows to construct neighbourhoods (around that point) which are used in Huber's *minimax approach* to quantitative robustness (see Huber (1981) and § 2.2). Performance of $T(F)$ under distortion can then be studied by observing the functional behaviour of $T(\cdot)$ in these neighbourhoods.

Final remark Up to this point the concepts of distortion and performance have been presented as they appear in the literature. In the end of chapter 4 they will be re-considered and discussed (in relation to each other) from our point of view. Prior to this, the notion of distortion will be revisited by emphasizing the fundamental distinction between data contamination and model deviation in chapter 3, and a suitable way of describing (and comparing) performance will be introduced in chapter 4.

Chapter 3

Distortion from a revised point of view

3.1 Introduction

The previous chapter pointed out different interpretations of distortion in the literature. In most cases distortion has been seen either in relation to just the data or to just the model assumptions. In some references, moreover, both views are loosely combined. Overall however, a *distinction* between distortion due to contaminated data and distortion due to a misspecified model has mostly been ignored. This is surprising, especially when taking into account the long-established position of robust statistics and diagnostics in statistical research.

This chapter will introduce distortion from a revised point of view. A distinction will be made in definitions for data contamination and model deviation by emphasizing that statistical inference is aimed at explaining some unknown aspect of the real-world representing the truth rather than just describing the underlying data (section 3.2). Following the presentation of model and inference framework in section 3.3, the two types of distortion will be discussed in detail in section 3.4.

3.2 Prelude: Statistical inference about *what*?

Statistical methodology can be distinguished into descriptive and inferential statistics (Hartung, 1995). The former is devoted to the description of data "entirely or largely independent of a probability model" (Cox, 1978). We will concentrate on the latter aspect, also called *statistical inference*. According to Barnett (1982, p. 13) this is the study of procedures which utilize "information to obtain a description of the practical situation". Unlike descriptive statistics, it uses the notion of a probability model. However, what is the *aim* of statistical inference? Is it also the data, or might the interest be laid upon some other higher order 'instance'? In other words: Statistical inference is made about *what*? A look into the literature does not give a clear answer:

- Barnett (1982, p. 6f) refers to "real data which has arisen from the practical situation", but also to "data arising from the real situation". Hence, do "real situation" and "practical situation" mean the same thing? In addition the author mentions the "real-world problem" from which the model is motivated (Figure 1.1). It is not apparent (to us) whether the "practical situation" coincides with the underlying situation of the real-world or whether it corresponds to the process actually generating the data.

- "Statistical inference is the theory and methods concerned with the way that background information and current data make implications concerning unknowns in a system under investigation" (Fraser, 1983).

- Cox and Hinkley (1974) explain in their book about the logical aspects of statistical inference: "Statistical methods of analysis are intended to aid the interpretation of data" (p. 1).

- Any book on *data* analysis most certainly involves methodology from statistical inference.

Most references seem to address the data or the process generating the data as the ultimate aim of statistical inference. Still, we will adopt the view that statistical inference is aimed at *real-world* description – and *not* data description. First of all this means that data and the real-world are *not* the same, although ideas could get mixed up when thinking of 'real data' etc. (see above in the citation of Barnett). To be rather philosophical, one could say that the real-world is always true (in the sense of correct), whereas the data do not necessarily need to be.

Secondly, this implies that the basic ingredients of statistical inference, the *data* and *model assumptions*, should be oriented towards the real-world in some way. Concerning the data this seems to be generally accepted, i.e. data should be correct and representative. However, there appear to be different opinions with respect to the model assumptions. Many statisticians claim that model assumptions should address the underlying *data*, i.e. ultimately the model should fit the data. This becomes apparent in a paper by Chatfield (1995) discussing problems arising with the common practice of *data*-dependent (model) specification searches (not only in the context of Bayesian statistics). Others, such as Tukey (1997, p. 21), understand that the assumptions should try to explain the *real-world* (ignoring the possibility that for them real-world and data might be identical).

This work will adopt an intermediate position: Model assumptions should address the *real-world* (independently from any information in the data) and they should take into account the (dependence) mechanism which actually generates the *data* (e.g. assuming independence for the sampling procedure).

Explaining the real-world with the help of statistical model assumptions can be justified if one believes in the existence of a so called *true model*. The latter shall exist in some 'objective reality' reflecting a natural (physical) phenomenon, *independently* from any data-generating mechanism. In most cases such a true model is infinitely complex and only arbitrarily large data sets (produced under constant circumstances) could even theoretically contain all its information. Thus as a rule, (finite) data cannot be absolutely

representative, and model assumptions can only mean simplifications of the true model. Hampel *et al.* (1986, p. 409ff) give a critical discussion about several alternative views of such a true model (including the one introduced here).

We will concentrate on the fact that statistical inference aims at describing the *real-world*. It provides the starting point for a distinction of distortion into data contamination and model deviation. In preparation for the details the next section will introduce the model framework which includes the triplet *data, model assumptions* and *real-world*.

3.3 The model and inference framework

A basis for the discussion of data contamination and model deviation will be developed by introducing the so called *model framework* and *inference framework*. The two concepts together form the logical structure of a scientific experiment involving statistical inference. Already existing ideas will be reinterpreted and related in order to accommodate later the revised point of view of distortion.

3.3.1 The model framework

The model framework describes the route from a real-world situation to the corresponding data-generating process (Figure 3.1) and implements different instances associated with the statistical model (see later). The basic ideas will be described first by a simple example (Barnett, 1982, p. 5):

Example 1 *Interest is given to the unknown emission rate of α-particles of some radio-active substance. In an experiment the number of these particles is counted over several fixed and non-overlapping time intervals of the same length t. Experience suggests that the counts are Poisson(λt)-distributed.*

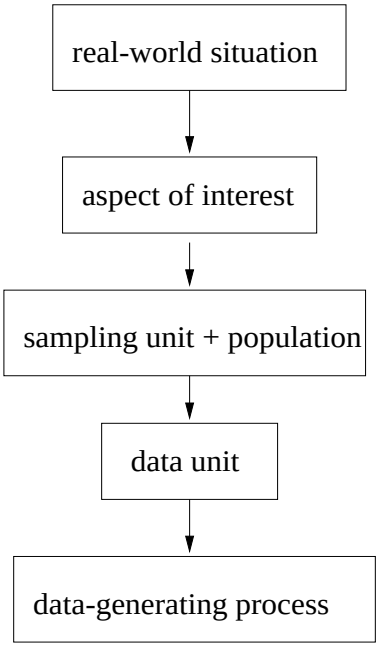

Figure 3.1: *From the real-world to the data-generating process.*

3.3.1.1 From the real-world to the data-generating process

The first part of the model framework consists of the directed route through

real-world situation describing the scientific context in which the statistical experiment is embedded. In the above example this would be the fact that the radio-active substance emits α-particles.

aspect of interest which is specified by the scientific problem, here the emission rate of α-particles.

sampling unit and population The real-world situation and the aspect of interest naturally indicate the kind of *sampling unit* which is derived from some underlying *population*. The latter can be either finite or infinite, and concrete or fictitious. In the example a sampling unit is represented by the collection of α-particles in a fixed time interval.

The potential collections of all possible time intervals of the same length constitute the infinite, fictitious population.

data unit The observable outcome of a sampling unit determines the corresponding *data unit* (datum). In the above example a data unit consists of a single value and corresponds to the number of α-particles recorded in the particular time interval.

data-generating process Knowing the characteristics of data unit and underlying population, a *data-generating process* can be described in terms of a statistical model, i.e. now using a statistical language. In the above example the data units are explained by the random variables $X_1, \ldots, X_n$, where X_i represents the number of α-particles in the i-th fixed time interval. They can be considered as independently and identically like Poisson(λt)-distributed.

The characteristics of *data units* can generally be distinguished by aspects such as the number of observations (univariate or multivariate) per unit, and associated covariates. For instance, there is just one univariate observation per data unit for the emission rate of α-particles, and there are no associated covariates. Data units may also entail the chance of representing missing values. Finally, a censoring indicator may be attached to their observations, and the data units themselves may refer to unobservable aspects such as censoring times. Missing values and censoring mechanisms are, however, not relevant in the α-particle example.

Additional examples shall now explain other particularities of a model framework.

Example 2 *In order to analyse to what extent certain fertilizers influence the quantity of wheat crop, average crop results from four different fertilizers are compared.*

The real-world situation in example 2 may be seen as the general dependence of fertilizers and crop results, while the specific comparison of average

crop results from those four fertilizers formulates the particular aspect of interest. A sampling unit consists of a single field treated with a certain fertilizer producing one univariate observation for the corresponding data unit (quantity of wheat crop per square meter). The factor 'fertilizer' and other variables such as temperature serve as covariates. The theoretical population of all fertilized fields is infinite and fictitious. An ANOVA-model (or ANCOVA-model) could explain the underlying data-generating process.

Example 3 *In a social survey students in Sheffield are asked about their daily consumption of cigarettes. Some of them, however, do not fill in the questionnaire.*

In this example the real-world situation is reflected by the smoking habits of students. The corresponding average number of cigarettes consumed daily constitutes the aspect of interest. As is typical for social surveys, the sample is taken from a *finite* and concrete population, here, composed of the students in Sheffield. To be more precise, the idea of a population could be distinguished into *target* population, about which information is desired, and *sampled* population, from which the sample is taken (see e.g. Mood *et al.* (1974), p. 222f). That is, the target population could be the fictitious "super-population" of students in Europe at an arbitrary point of time, but it could also be identical with the finite sampled population of all students in Sheffield. Each student represents one sampling unit. The single observations of each corresponding data unit are multivariate due to additional information such as age, sex, etc. or represent missing values in some cases. A model for the data-generating process could take into account the potential missing value status, and also the dependencies between data units (sampling units) if the sample is drawn without replacement. In case, the target population is considered to be a "super-population" a classical normal regression model might be suitable, eventually accounting for dependencies ("super-population model" as termed e.g. by Barnard (1971)).

Example 4 *The effects of a new treatment for cancer patients is tested in a clinical trial. Patients may enter the study at different times, they may not be followed-up until death occurs, and they may discontinue the treatment prematurely (random censoring, see e.g. Miller (1981), p. 5f).*

Here, the real-world situation is described by the existence of a new cancer treatment and the fact that patients may or may not show a positive treatment effect. The survival time distribution of an 'average' patient under the new treatment formulates the aspect of interest. For each cancer patient (sampling unit) two items are recorded: the observed survival time and the information of whether this time is censored or not (censoring indicator). The latter indicates whether the time of the patient leaving the study while being alive (censoring time) or the actual unknown survival time is smaller. The 2-dimensional vector of the observed survival time and the censoring indicator forms the corresponding data unit. The population of all cancer patients is considered to be infinite and fictitious (also taking into account future patients).

A statistical model for the data-generating process could be the Koziol-Green model (see the detailed example of chapter 5): The positive random variables $X_1, \ldots, X_n$, $Y_1, \ldots, Y_n$, and $Z_1, \ldots, Z_n$ representing the actual, censored, and observed survival times are i.i.d. like X, Y, and Z with continuous distribution functions. The random variables $\Delta_1, \ldots, \Delta_n$ representing the censoring indicator are i.i.d. like Δ, where the latter is $\mathrm{Bin}(1, p)$-distributed. X and Y, as well as Z and Δ are independent.

Example 5 *In a longitudinal study several protein measurements are taken from cows milk during a fixed period of time. Each donating cow is treated with one of three different diets and is presumed to be healthy (Diggle et al., 1995).*

It is known that in the real-world the diet of cows may in some way influence the protein content of milk. The aspect of interest is therefore a comparison

of the resulting average protein contents from the three different diets. Measurements from the same cow (one sampling unit) are summarized to a single data unit. This means that the data units are (usually) composed of *more than one* observation – a typical characteristic of longitudinal data. The observations themselves are univariate and associated with the factor 'diet' and a covariate indicating the time of measurement. The population of healthy milk cows is considered to be infinite.

A statistical model for the data-generating process might be the following (see the detailed example in chapter 6 for a specific case): The N observations are summarized into m independent random vectors $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in})'$ corresponding to the data units with $m \cdot n = N$ and

$$\boldsymbol{Y}_i = X_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i, \tag{3.1}$$

where X_i is a $(n \times p)$–matrix of covariates – of the same structure within each diet group, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a vector of coefficients, and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{in})'$ is a vector of random errors distributed like $\mathrm{mN}(\mathbf{0}, \sigma^2 V)$ with some matrix V taking into account the natural correlation structure between measurements from the same cow.

Example 6 *Interest is given to the development of deaths due to bronchitis, emphysema, and asthma in the U.K. Basis for a study are the corresponding monthly registered deaths over a fixed number of years (Diggle, 1990).*

Since the above diseases are well-known causes of death in the real-world, the aspect of interest is devoted to the numeric development of those cases. The number of monthly registered deaths over the years considered is a time series and as such corresponds to only a *single* data unit with several (univariate) observations. The population could be defined in various ways, e.g. be identical with the single sampling unit (collection of registered deaths) or describe the corresponding records in several countries. One of the known time series models might be appropriate to represent the data-generating process.

Example 7 *In order to get knowledge about some physical constant ρ, several measurements are taken in an experiment.*

Here, the real-world situation corresponds to the physical phenomenon, and the aspect of interest is the physical constant itself. The sampling units (= data units) with single univariate observations are 'drawn' from the infinite, fictitious population of all possible measurement values. As in example 3 a distinction could be made between target and sampled population. That is, the target population contains the unknown constant as its single element, while the set of all possible measurement values constitutes the sampled population. The corresponding data-generating process is intuitively associated with a $N(\rho, \sigma^2)$-distribution model. Nevertheless, the unknown constant of interest is deterministic and would correspond to a one-point distribution at ρ. This fact will be given particular attention in section 3.4 discussing model deviation and data contamination.

The above examples are used to explain the concepts and ideas developed throughout the present chapter. In addition, the examples 4 and 5 will be considered in much more detail in later chapters.

3.3.1.2 Aspects of the model between and within data units

A data-generating process can be described in terms of a statistical model (see § 3.3.1.1). In order to specify distortion it is important to distinguish aspects of the model *between* data units (describing dependencies) from those *within* data units.

The independence assumption of the often cited i.i.d.–statement of a statistical model refers to *aspects between data units* and is usually a natural consequence of random sampling from an infinite population. For finite populations the statistician might choose sampling schemes which imply dependence between the data (sampling) units, see example 3. Aspects between data units *do not* exist in cases with only a single unit as in example 6 (time series model).

Aspects within data units can be modelled by specifying a distribution (usually the same) for each data unit as in examples 1 and 7. When covariates are present (examples 2 and 5) this is replaced by a more complex regression model consisting of a deterministic and a stochastic (distributional) part. Then, a single identical distribution for all data units is obviously no longer valid. In addition, dependence structures are to be defined within data units when the number of observations per unit is greater than one (examples 5 and 6), and maybe within observations when the units themselves are of multivariate or censored nature (examples 3 and 4).

Note, that an i.i.d.–statement is also used in the context of regression models. Here, the identical distribution becomes valid after 'subtracting' the deterministic part of the model. It then means that the i.i.d.–statement addresses aspects between *residuals* (and not data units).

3.3.1.3 REALITY, DATA, and ASSUMPTIONS

So far, a single model has been used to represent the data-generating process. For reasons of simplicity the three different views of a data-generating process represented by the statistical models called REALITY, DATA, and ASSUMPTIONS (see Figure 3.2) have still been neglected. This shall be accomplished now.

In § 3.2 it was pointed out that an experiment involving statistical inference is strongly related to the triplet consisting of real-world, data, and model assumptions. While the ultimate aim of such an experiment is the description of the *real-world*, the relevant tools used for it (in addition to the inference procedure itself) are the *data* and *model assumptions*. It is important that somehow the last two elements are oriented towards the real-world. In other words, real-world, data, and model assumptions are supposed to conform with each other in some way, but with the real-world regarded as central.

A difficulty arises when trying to specify this correspondence. The three components are of completely different natures, so that a direct comparison

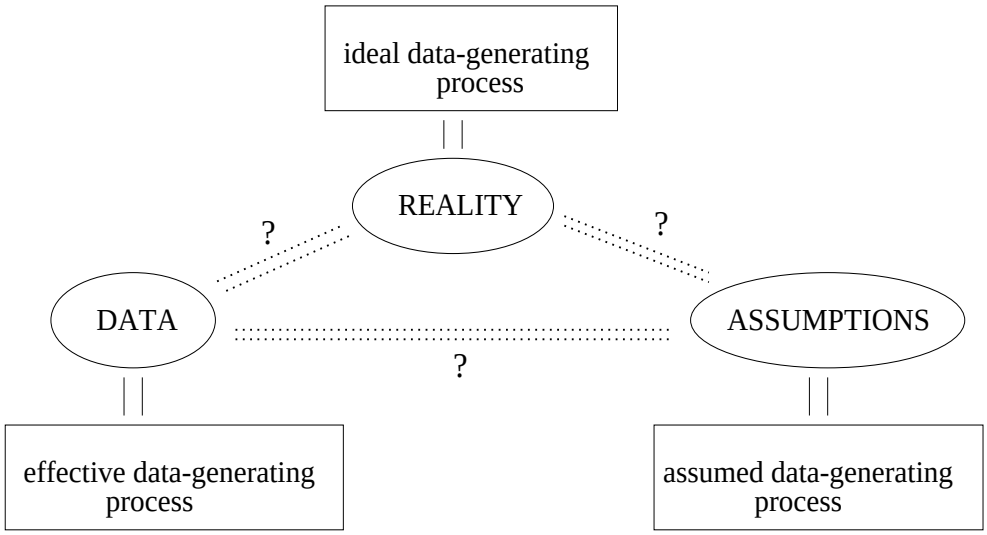

Figure 3.2: *Instances of the data-generating process and the model triplet.*

between them is not possible. However, the problem can be avoided when viewing each component as a *data-generating process*, or vice versa, viewing the data-generating process of an experiment in three different forms (see Figure 3.2). In detail, the

- *ideal data-generating process* complies with the *real-world* form. It involves the theoretically true, but *not* necessarily actually observed data units and the sampling scheme which is either planned or ideally results from the underlying population.

  In example 1 the ideal data-generating process refers to the true emission of α-particles from the radio-active substance. Since the number of particles is naturally independent and equally distributed in arbitrary fixed and non-overlapping time intervals, the ideal 'sampling scheme' is i.i.d.

- *effective data-generating process* complies with the form actually producing the *data*. It involves the data units really observed and the

sampling scheme which finally is brought to effect, also when not necessarily being planned in the first place.

In example 1 the effective data-generating process corresponds to the actual recording by the counter of particles activating the sensor. This is also associated with the possibilities that particles arrive undetected, or spurious particles are recorded.

- *assumed data-generating process* is modelled by the statistician giving the *model assumptions*. He attempts to identify the theoretically true data units resulting from the real-world (reflecting the *true model*, see § 3.2) and the finally effective sampling scheme actually producing the data.

  In example 1 this corresponds to assuming that the particles arrive in a Poisson process where the Poisson distribution itself serves to explain the *true model*.

A basis for the comparison of real-world, data, and model assumptions can now be created by representing each corresponding form of the data-generating process by a *statistical model*. These will be called REALITY, DATA, and ASSUMPTIONS, respectively. The models describe aspects *within* data units, e.g. $X_i \sim F_i$ as well as aspects *between* data units, e.g. the independence of the X_i. Overall, the above approach can be justified by

- interpreting REALITY as the unknown *true model* which is extended through the superimposed ideal sampling scheme e.g. associated with the characteristic of independence. In other words, the true model refers to a *single* theoretically true data unit, and the model REALITY describes the *overall* random sample of corresponding true data units. While the true model corresponds to aspects *within* data units, the independence of the ideal sampling scheme forms the aspects *between* data units.

- considering the data as a *repetitive event* which is produced always according to the same probability mechanism as reflected in the model called DATA.

Like the true model and due to it, the model called REALITY is most often infinitely complex. For instance, the true crop results of wheat are usually *not* only dependent on the underlying type of fertilizer and the temperature. In fact, there may be many more additional (unknown) influence factors. Thus, REALITY in example 2 could be some kind of ANOVA-model with infinitely many covariates. The same usually applies to the model denoted as DATA. Taking again example 2, the really observed crop results may be influenced by infinitely many side-effects. In addition to the actually recorded covariate information the corresponding model DATA may therefore also include infinitely many other (unobserved) covariates.

In general, infinite complexity of a model may mean that the model itself finally turns out to be essentially deterministic. One could imagine that nature can always provide some additional influence factor for a phenomenon which in principle seems to be random. The outcome of a dice throw, for example, does obviously also depend on the angle of throw, direction and strength of wind, etc. Thus, after all, the dice throw might be absolutely deterministic, if one could encounter for all potentially possible influence factors (compare with Hartung (1995)). In some cases, however, a true model may already be deterministic in *finite* terms as e.g. the unknown physical constant in example 7.

Despite the often infinite complexity of REALITY and DATA, it is common practice to use simple and thus finite ASSUMPTIONS, i.e. *statistical* model assumptions. The above interpretation of REALITY and DATA is hence only of philosophical importance in the present work. Here, each of the models REALITY, DATA, and ASSUMPTIONS will be considered to be some finite 'conventional' statistical model. In the case of an ANOVA-model this means, for instance, that all additional influence factors are summarized

in the normal-error distribution.

After the description of the inference framework in the next subsection, the 'correspondence issue' of REALITY, DATA, and ASSUMPTIONS will be discussed in the context of data contamination and model deviation. Note, that *model assumptions* and ASSUMPTIONS are synonymous expressions and will be used interchangeably from now on according to context.

3.3.2 The inference framework

The inference framework relates the logical elements involved in a statistical inference process (see Figure 3.3). Throwing light on the *aspect of interest* of a real-world situation is the aim of statistical inference. To pursue this aim the statistician is provided with the following 'tools':

- *data* as a collection of data units,
- *model assumptions* as prior belief about the *true model* and the finally effective sampling scheme, and
- *inference procedures* which process the data and give information about the relevant unknown parameters of the model assumptions, parameters taken in the widest sense. Note, that in the present work inference procedures are generally considered to be estimators.

One of the two following ways may then lead to the desired description of the aspect of interest (see Figure 3.3):

1. Use of just the result of a single inference procedure, e.g. an estimate for the parameter as the only aspect of interest in a parametric model or a non-parametric density estimate.
2. Use of both, the results of the inference procedure(s) and the model assumptions. For example, the dependence between a variable of interest and some covariate is described by the combination of parameter estimates and the assumed structural part of a linear model.

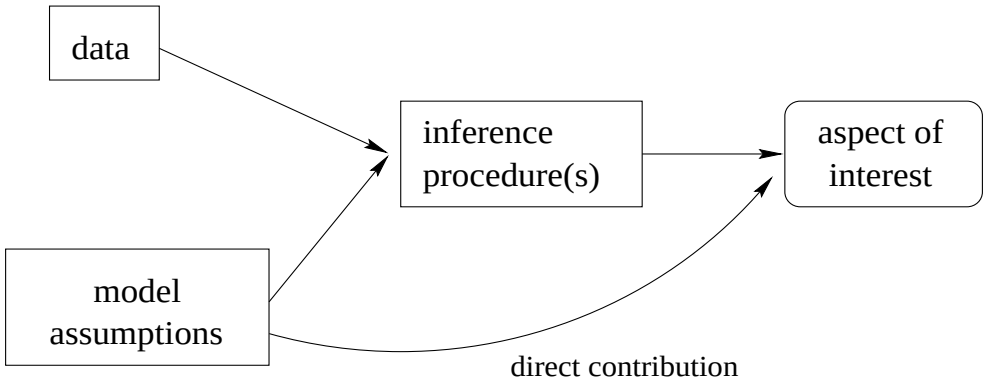

Figure 3.3: *The inference framework.*

Besides any direct contribution to the real-world description, the model assumptions obviously give a formal representation of the aspect of interest and they suggest the kind(s) of inference procedure to be used e.g. a location or density estimator. In addition, they are responsible for proposing statistical properties of the inference procedures involved. Aspects of the model assumptions which are *only* used to propose statistical properties belong to the so called *dummy-part* of the model and parameters associated with them are usually known as *nuisance parameters*. For further reference note that 'real-world description' and 'proposal of performance properties' translates to *nominal* and *stochastic* inference, respectively, in Dawid (1983).

The idea of an inference framework will now be illustrated with the previous examples 7, 4, and 5:

- In example 7 the measurements are assumed to be $N(\rho, \sigma^2)$-distributed. The unknown parameter of interest ρ is therefore estimated by the maximum-likelihood (ML-) estimator $\hat{\rho} = 1/n \cdot \sum X_i$. Under the above normal model this estimator shows properties such as unbiasedness and BAN (best asymptotically normal), see for instance in Mood *et al.* (1974). Since $\hat{\rho}$ yields a value for the unknown constant and thus completely describes the (only) aspect of interest, no *direct* contribution

from the model assumptions is used for this particular real-world description.

- In example 4 the unknown survival function of X, denoted as $S_X(\cdot)$ (aspect of interest), is estimated by the so called ACL-estimator

$$S_X^{ACL}(z) = \left[1 - \hat{F}_Z(z)\right]^{p_n},$$

which is the corresponding ML-estimator under the Koziol-Green model (Csörgő, 1988). At first sight, the model assumptions do not *directly* contribute to the interpretation of the aspect of interest. However, if one considers the empirical distribution function $\hat{F}_Z(\cdot)$ and $p_n = 1/n \cdot \sum_{i=1}^{n} \Delta_i$ as two individual estimators, the model assumptions *do* make a direct contribution since the equation $S_X(z) = [1 - F_Z(z)]^p$ with $p = P(\Delta = 1)$ is equivalent to the Koziol-Green model assumption.

- In example 5 the relevant unknown parameters $\boldsymbol{\beta}$ and the nuisance parameters σ^2 and V are estimated via ML-estimation (Diggle *et al.*, 1995, p. 63f). While the deterministic part of the regression model in equation (3.1) directly contributes to the interpretation of the average protein contents (aspect of interest), the stochastic part implementing a Gaussian model only supports the use of the ML-method and thus proposes ML-associated properties such as BAN.

In experiments with model and inference framework, the real-world (comprising the aspect of interest), data, and model assumptions are supposed to correspond to each other. In other words, the statistical models REALITY, DATA, and ASSUMPTIONS should be in harmony as a triplet by agreeing with each other in some way. Any situation *not* satisfying this condition shall be called *distortion*. The latter will be discussed in detail in the next section.

3.4 Data contamination and model deviation

Distortion as the *inter*ference factor of an experiment with statistical inference has found several definitions in the literature (see chapter 2). A fundamental inconsistency can be notified therein in the 'target' of distortion: Is it the data, the model assumptions, or perhaps both, which could suffer distortion? In chapter 2, the question has only been roughly answered by considering distortion as *any kind of deviations from the assumptions* (p. 12). Thus, it has been agreed that distortion might refer to

- 'bad' data – the data deviate from the (correct) assumptions (*data contamination*), or to
- 'bad' model assumptions – the assumptions deviate from the truth (*model deviation*).

With the present chapter this distinction could moreover be supported by the claim that statistical inference is aimed at real-world description (§ 3.2). It is the real-world which in most instances can provide a reference point for the corresponding data and model assumptions (see the following discussions about exceptions). Hence, in these cases the latter two do not necessarily need to be compared directly with each other. This allows a distinction between distortion 'targeting' the data and distortion 'targeting' the model assumptions.

Distortion in general can be described as a kind of non-correspondence between real-world, data, and model assumptions. More formally this implies that there is some *disagreement* among the statistical models called REALITY, DATA, and ASSUMPTIONS (§ 3.3.1.3). As a first step, the present section will specify this disagreement in more detail by taking into account model structures and origins, and by devoting attention to aspects of the model *within* and *between* data units. Afterwards the two situations known as data contamination and model deviation will be distinguished and discussed with several examples.

3.4.1 Model disagreement

Relative to the model structure A disagreement between REALITY, DATA, and ASSUMPTIONS can be explained intuitively by considering the structure of the statistical model itself. Thus, principally different model structures, such as a linear instead of a non-linear term, or a Weibull in the place of an exponential distribution could result in model disagreement. As a consequence, the number of parameters may differ, or they even inherit different meanings as in the case of a uniform and a normal distribution. However, the model triplet could also be incompatible due to differences in just local model aspects. Possibilities with some common statistical models are presented in § 2.3.3. Moreover, known formalization concepts of distortion (see § 2.3.2) which consider some kind of *distorted* model beside the ideal model refer to the model structure (finite number of alternatives, model expansion, and model mixing, see § 2.3.2.1 – § 2.3.2.3).

Relative to the model origin Further aspects become clear when the different origins of the model triplet are taken into account. The ASSUMPTIONS already agree with REALITY or DATA when the former model *equals* or *embraces* the other two models. This means that a more general model for the ASSUMPTIONS, e.g. the model of right random censorship, does *not* disagree with a more specialized model such as the Koziol-Green model for the REALITY or DATA (see example 4). The same applies to a non-parametric model for the ASSUMPTIONS which embraces a parametric model for the REALITY or DATA. Model assumptions which are too general cannot be wrong.

Due to their origin the three models give a different status to the parameters involved. As a rule, parameters are fixed to a specific value in REALITY and DATA, and not fixed in ASSUMPTIONS. This has some special implications for model disagreement relative to a *parameter value* (important for the formulation of influence and preference graphs in the next chapter). While

DATA and REALITY may simply differ in this respect, a relation with the ASSUMPTIONS has to consider the following:

- If the ASSUMPTIONS include the corresponding parameter, they naturally comprise and therefore agree with the models DATA and REALITY.
- If the ASSUMPTIONS do *not* include the corresponding parameter and as a result are more specific, this agreement does not necessarily follow any more.

An artificial example may illustrate this problem: Consider a Weibull(a, b)-distribution describing the aspects within data units. With the parameter value b not further specified, the ASSUMPTIONS embrace (and agree with) all model formulations with fixed b – in particular the corresponding fixed Weibull-distributions in DATA and REALITY. This is *not* necessarily the case, however, if the ASSUMPTIONS refer to an exponential(a)-distribution, i.e. with b 'frozen' to the value 1.

Relative to the data unit It is also important to study the meaning of model disagreement within the inference framework (§ 3.3.2). Starting with the opposite problem: to what extent does statistical inference, and in particular nominal inference, require the *agreement* of REALITY, DATA, and ASSUMPTIONS?

Statistical inference is aimed at real-world description (§ 3.2). This implies that the data, as a collection of data units and *independently* from any sampling scheme, should be oriented towards the real-world. It only matters that the data units themselves are without mistakes – *how* they are related does not influence the 'correctness' of the data.

Similarly, the model assumptions should address the real-world in order to give the correct formal representation of the aspect of interest, suggest the right kind(s) of inference procedure, and give a valid direct contribution to

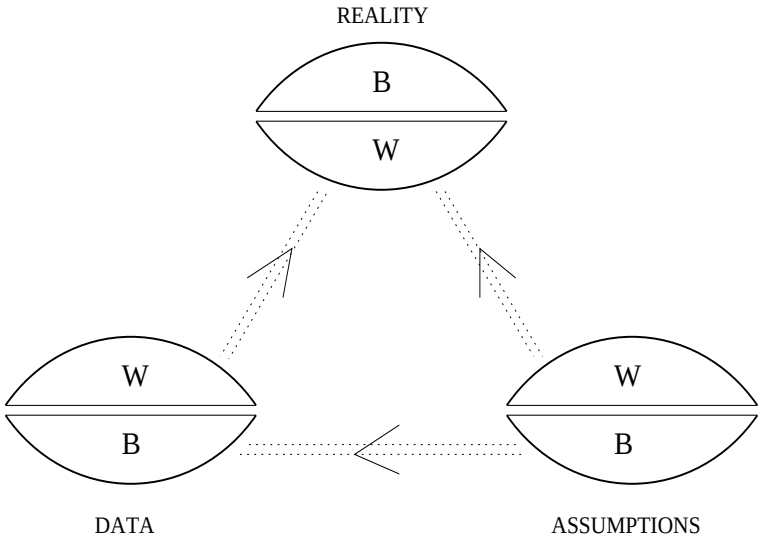

Figure 3.4: *Directed model agreement required for statistical inference, W - within and B - between data units.*

the real-world description. However, model assumptions should *also* take into account the dependence structure between actually generated data units. This is because, they are used for proposing statistical properties of the inference procedures involved. It does not matter which sampling scheme has been planned initially. Only the dependence structure of the *effective* sampling scheme counts and needs to be assumed correctly. For inference purposes model agreement is therefore required in the following directions (see Figure 3.4):

1. from DATA to REALITY relative to aspects *within* data units,
2. from ASSUMPTIONS to REALITY relative to aspects *within* data units, and
3. from ASSUMPTIONS to DATA relative to aspects *between* data units.

Model aspects *between* data units of the model REALITY are not involved. They have only been introduced as an extension to the *true model* in order to simplify the foregoing theoretical discussions.

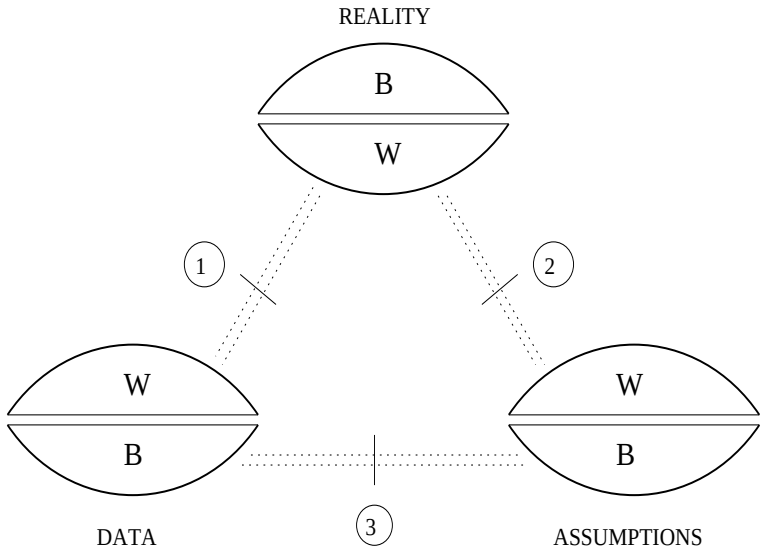

Figure 3.5: *Model disagreement relevant for distortion, W - within and B - between data units.*

Overall, we come to the following conclusion: Model *dis*agreement which is disadvantageous for statistical inference and implies *distortion* is restricted to the corresponding three cases as indicated in Figure 3.5.

3.4.2 Which type of distortion – when?

Up to this point distortion has been discussed, without further classifications, as some disagreement within the model triplet of REALITY, DATA, and ASSUMPTIONS. A next step will now clarify

- when model disagreement means data contamination and when model deviation,
- what kind of errors can be associated with each situation of distortion, and
- when it is of particular interest to *compare* potential consequences of distortion.

Model disagreement due to wrong DATA defines *data contamination* and due to wrong ASSUMPTIONS defines *model deviation*. The following important conclusions can then be formulated, referring to the arrow directions in Figure 3.4 and the numbers in Figure 3.5:

- Model disagreement of type ① can be identified as *data contamination*. It is due to a conflict between DATA and REALITY relative to aspects *within* data units.
- Model disagreement of type ② can be identified as *model deviation*. It is due to a conflict between ASSUMPTIONS and REALITY relative to aspects *within* data units.
- Model disagreement of type ③ can be identified as *model deviation*. It is due to a conflict between ASSUMPTIONS and DATA relative to aspects *between* data units.

A distinction between data contamination and model deviation can *only* be made when distortion relates to aspects *within* data units. Data contamination relative to aspects *between* data units is *not* possible. Again, this is because the dependence structure of data units actually observed does not influence the 'correctness' of the data. As long as the data units themselves are correct there are *no* grounds for the presence of data contamination.

Each type of distortion further implies the following basic errors (Figure 3.6):

- Data contamination (distortion of type ①) can be explained by incorrectly imitating the distribution of (some of) the theoretically true data units with the effective data-generating process, i.e. in a potential sample (some of) the actually observed data units are wrong.
- Type ② model deviation can be explained by making incorrect assumptions about the distribution of (some of) the theoretically true data units.

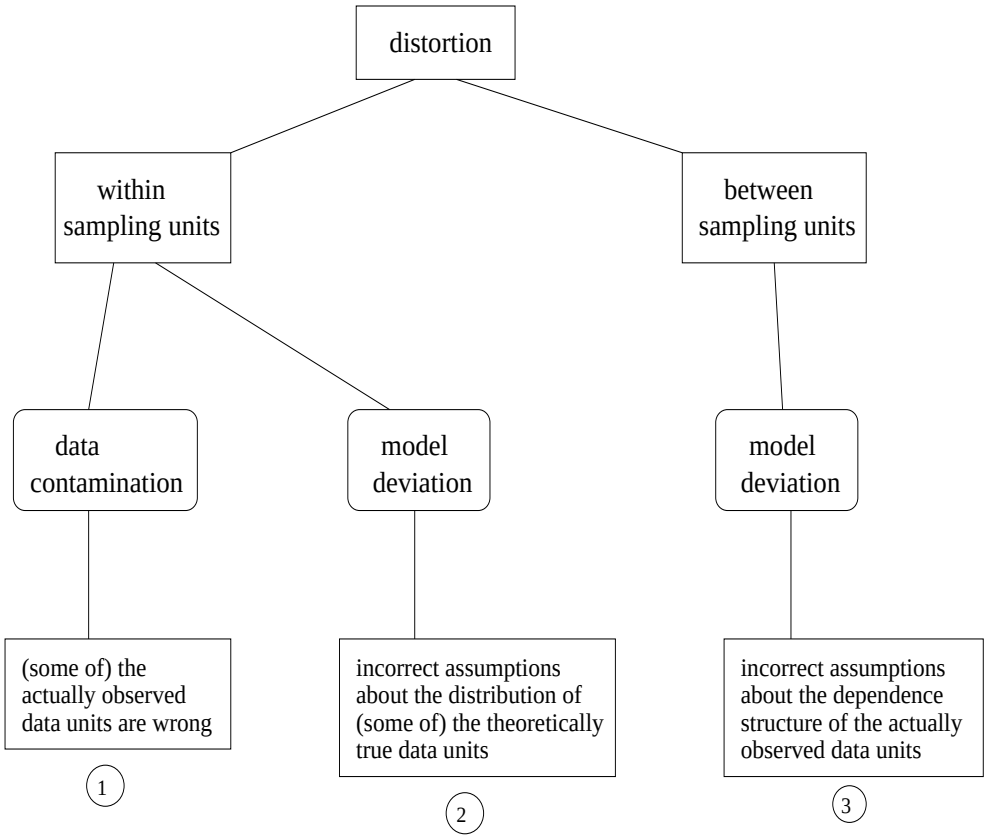

Figure 3.6: *Distortion – associated errors and their classification.*

- Type ③ model deviation can be explained by incorrect assumptions about the dependence structure of the actually observed data units, i.e. a wrong specification of the actual effective sampling mechanism.

The subsequent chapters will focus on distortion of the types ① and ② which both affect aspects of the model *within* data units and allow a distinction between model deviation and data contamination. For these cases a great variety of potential examples can be found (see the next subsection). This is because the complexity of a statistical model and therefore its 'vulnerability' are determined *within* the data units. Also, most of the statistical literature deals with this kind of distortion, however, usually without distinguishing between model deviation and data contamination (see chapter 2).

Nevertheless, it is *this* possible distinction, which motivates the *comparison* of distortion of types ① and ②. Both situations have the potential to influence the performance of inference procedures in a *different* way (see the subsequent chapters). Still, a statistician confronted with one or the other type of distortion, will not necessarily be able to *identify* correctly the presence of data contamination or model deviation in order to take appropriate precautions. He can often *only* recognize that overall the model denoted as DATA does not conform with the ASSUMPTIONS. Which of the two models is actually wrong, however, is not always easy to decide. The third model instance called REALITY is (usually) out of perception.

It is therefore of interest to assess the danger of this ignorance: To what extent do the results of statistical experiments differ under corresponding data contamination and model deviation? What consequences should this imply for the experimental set-up? These and other problems will be addressed in the context of two particular examples in chapters 5 and 6.

This chapter concludes with examples and a general discussion covering distortion of types ① to ③. For reasons of simplicity consideration will only be given to 'plain' distortion of one or the other type. It is also realistic, however, to expect distortion to be of mixed type, i.e. simultaneous data contamination and model deviation or a conflict where aspects of between as well as within data units are involved.

3.4.3 Examples and discussion

The theoretical considerations above will now be illustrated and discussed with the examples already used throughout this chapter. At some points additional examples may serve for further clarification.

Data contamination as distortion of type ① implies that some of the actually observed data units are wrong (refer to Figure 3.6). It can occur as measurement, recording, rounding, or grouping error in the process of

observing the outcome of a sampling unit. Alternatively, data contamination may be due to some sampling units themselves which are inappropriate. If the errors only appear occasionally but are "quite powerful", they are also known as *gross errors* (see e.g. Hampel *et al.* (1986), p. 21ff).

In example 1 data contamination could suggest that some α-particles arrive undetected or that other spurious observations (background radiation) are recorded. As a result some of the counts (data units) are incorrect. In probabilistic terms this could mean that the (identical) Poisson-distribution of the X_i in the model called DATA inherits a wrong parameter value (relative to the true value in REALITY). It is also possible that the distributions of only single X_i are affected if the occurrence of errors is restricted to only certain time intervals.

The example can further explain why data contamination relative to aspects of the model *between* data units is not prevalent. One could imagine that radio-active particles already counted are somehow still around and might be recorded again by mistake in one of the following time intervals. This overlap would first of all imply a dependence *between* subsequent counts, i.e. data units. However, the dependence effect as such does not principally mean incorrect data. These are rather the data units (counts) *themselves* which are wrong and imply data contamination. Thus, the effect which initially appears to operate between data units turns out to be data contamination *within* data units.

Wrong data units leading to incorrect distributions in the DATA-model can (do) also occur

- in example 2 when the quantity of wheat crop is weighted improperly or mistakes have been made in the transfer of measurement values to a data base. The incorrect data entries could become obvious as extreme outliers, but they can also be hidden among the 'good' values. Similarly, errors could arise in recording the covariate information. As a consequence the structural part of DATA (i.e. the regression model)

might reflect a different relationship between the dependent variable (crop results) and the covariates of interest. Further, the corresponding error-distribution could be affected in terms of distribution equality, distribution type and parameter values (see also § 2.3.3.2). Rounding errors of the weight results could finally invalidate the continuity of the error-distribution in DATA.

Note, that *without* the presence of data contamination the error-distribution of the DATA-model is considered to account for the additional, non-contaminating influences which are generally out of interest (see the discussion on page 48). This excludes in particular the (elementary) random measurement errors.

- in example 3 when students enter wrong answers into the questionnaires.

- in example 7 where the actual measurements (generated according to DATA) are approximately normal-distributed around the true constant ρ. Since REALITY inherits a one-point distribution at ρ, data contamination is an obvious and but also *unavoidable* consequence.

- in example 4 where mistakes may occur in recording the observed survival time Z (due to errors in noting the censored or the actual survival time) or the censoring indicator Δ. As a result the corresponding distributions in DATA may be different from their true 'originals'. Data contamination can also affect parameter values (see the discussion on page 54). Here, errors associated with the censoring indicator Δ may corrupt the value of the (only) parameter p in DATA away from the theoretically true value in REALITY.

- in example 5 due to measurement or recording errors. Moreover, external circumstances could badly influence the dependence structure of the observations *within* data units. This might e.g. be a cow in poor health (independently from the diet) during the course of the

experiment, i.e. an inappropriate sampling unit. Again, the possibility that some data units are dependent does not contaminate the data. The latter could be due to related cows, which perfectly well belong to the target population associated with REALITY, or might be caused by the effective sampling scheme itself. If, on the other hand, related cows were to be excluded from the target population, they *themselves* would have to be considered as inappropriate sampling units (i.e. again data contamination *within* data units).

- in example 6 due to errors in recording and transmitting the data.

Model deviation as distortion of type ② occurs when the theoretically true distribution of at least one data unit is wrongly specified. This may apply to any aspect of the model *within* data units (see Figure 3.6). Every statistical model assumption is first of all to be seen as an approximation to the real-world phenomenon at hand. If nothing else seems to be incorrect, this fact is very often an initial cause of model deviation. Example 1, for instance, does not serve very well as a candidate for 'substantial' distortion of type ②. The Poisson assumption for the α-particle counts is overall reasonably realistic (individual emissions occur with very low probability). However, also here the approximative character of the model can be regarded as model deviation.

Further situations of type ② model deviation can be met with

- example 2: The structural part of the assumed ANOVA-model certainly excludes several known and unknown covariates. The latter have been summarized and *approximated* by the corresponding error-distribution. This kind of model simplification again shows the first source of model deviation. The structural part of the model assumptions could further be wrong with respect to linearity (e.g.). Also, errors might occur in the distributional part with respect to distribution equality and type. Deviations from continuity, however, seem to be less visible.

Finally, model deviation relative to actual parameter values is *not* possible, since the latter remain pending in the ASSUMPTIONS (see the discussion on page 54).

Note, that the error distribution of the model assumptions, according to common practice, takes into account the (elementary) random measurement errors in the data. However, as long as the situation of *no* measurement error is included as a special case, this does not imply model deviation (see also the comments below for example 7).

- example 3 when a super-population model is assumed: A classical normal regression model could be affected as described above. In addition, the continuity requirement would not be satisfied since the experiment deals with discrete count data (number of cigarettes).

- example 4: The Koziol-Green model is a rather complex *semi*-parametric model and as such does allow model deviation to just a certain extent. The distributions of X, Y, and Z are not further specified apart from continuity and the limitation to positive realisations. The Bernoulli-distribution for Δ is certainly realistic. Model deviation, therefore, seems to be only justifiable with respect to the independence assumption between X and Y, and Z and Δ. Violations of the latter will be studied in detail in chapter 5.

- example 5 where the possible mistakes are similar to the ones in example 2. In addition, the dependence/correlation structure *within* data units (measurements from the same cow) could be assumed wrongly. Consequences of distortion with a particular correlation model will be considered in chapter 6.

- example 6: The ARMA-model does not provide a correct description for the underlying, theoretically true stochastic process.

Example 7 should be considered separately. While REALITY is a one-point distribution at ρ (within data units), the ASSUMPTIONS are formulated

with a $N(\rho, \sigma^2)$-distribution. However, even though orientation is given to the *data* rather than the real-world there is *no* situation of model deviation. The assumed normal model with pending parameter values includes the true constant as a special case, namely when the variance equals zero. Hence, the model assumptions approximate the real-world phenomenon by *generalization*, a principle which does not lead to model deviation.

Model deviation as distortion of type ③ relates to aspects of the model *between* data units. It prevails when incorrect assumptions are being made about the dependence structure of the actually observed data units. In example 3, where sampling is carried out from a finite population, actual (short-term) *dependencies* among the data (sampling) units might be improperly specified. In most cases however, it is the *independence* of the famous i.i.d.-assumption which causes doubts.

The problem can be explained again with example 1: Under very strict scientific considerations it turns out that the particle counts are not independent over the subsequent time intervals. There are only a limited number of α-particles to be released from the radio-active substance. Thus, the emission rate *must* decrease with the remaining particles in the substance (reduction of radioactivity). In other words, what seemed to be a Poisson (λ)-process with independent and stationary increments, is in fact some state-dependent generalisation with dependent increments (see e.g. Fahrmeier (1981), p. 93ff).

Further violations of the independence assumption between data units are discussed in Hampel *et al.* (1986, p. 387ff). One is the so called "Hurst phenomenon" where the yearly flood heights of the river Nile are correlated according to "seven fat years and seven lean years" (p. 390, referring to the Bible). The aspect of interest may be the average flood height of the river. Assuming independence between each of the n height values (data units) again results in type ③ model deviation. Obviously the dependence structure is influenced by the way the data is collected. Observing flood heights only every 14 years might as well eliminate the dependence effect.

3.4.4 Further remarks

- Even though the theory of robustness has been mainly developed for parametric models (Hampel *et al.*, 1986; Huber, 1981), a situation of distortion may also exist in the context of *non-parametric* inference. In fact, data contamination is generally possible independently from any kind of (non-parametric) model assumptions. The weight of 10-year old boys, for example, may be normal-distributed in the real world. A faulty scale, though, could reduce readings of very heavy weights by 10 %. As a consequence the data are contaminated, i.e. the model DATA does no longer inherit the 'pure' normal distribution of REALITY.

  Model deviation, on the other hand, is indeed less of a problem in the non-parametric context. *Only* what is actually assumed can be subject to model deviation. If the ASSUMPTIONS contain nothing other than continuity it does not matter how skewed the true distribution in REALITY turns out to be (e.g. a log-normal distribution for article prices in a stock, see Hartung (1995)). In this case, just an additional assumption of symmetry could imply type ② model deviation.

- Data contamination generally reduces the quality of the underlying data. However, not all lower-quality data are necessarily contaminated. Missing values, such as non-responses in a questionnaire (example 3) or missing covariate entries (example 2) are not inherently wrong. They cannot be related to a conflict between DATA and REALITY. The same applies to the fact that an observation is censored.

  Also, a 'non-representative' sample with many extreme and less typical sampling units is low in quality, and this *without* necessarily being contaminated. In this case the corresponding empirical distribution turns out to be too skewed. Still, each resulting data unit might be correct, and just the current sample could be very unlikely relative to REALITY. The situation may be described as *pseudo data contam-*

ination. Along this line, extremely heavy smokers could tend to be non-respondents in the social survey of example 3. Even though the actually observed data set becomes less representative due to missing values, it is not contaminated when the answers still available are all correct.

- The philosophy of aspects of the model *between* data units can be extended: As part of DATA the dependence structure between data units is determined by the effective sampling scheme. The latter is (directly or indirectly) controlled by the statistician, for which reason the corresponding 'truth' is based on the statistician's choice and is *not* originated in the real-world. Model aspects *between* data units should therefore not be seen in relation to REALITY (page 55). Apart from the sampling scheme, the statistician could also control other features of the effective data-generating process (if they are relevant). He can e.g. choose the censoring mechanism. Again, the affected aspects of DATA – even though they might be localized *within* data units – should not be compared to REALITY.

  Working with the model triplet might become complicated when a change of the DATA-model (within data units) could be due to either of the two possibilities: data contamination or controlled interference by the statistician. Then it is not clear anymore whether the relevant model aspects of DATA and ASSUMPTIONS should or should *not* be compared with REALITY. The example in chapter 5 is chosen to illustrate and at the same time avoid this difficulty.

In the present chapter the notion of *distortion* has been formalized and classified by putting particular emphasis on the distinction between data contamination and model deviation. Means for assessing the *performance* of statistical inference procedures under such distortion will now be developed in the following chapter.

Chapter 4

Performance under distortion

4.1 Introduction

The previous chapter introduced the concepts of model and inference framework which allowed us to reconsider and formalize the notion of *distortion*. With the real-world as the aim and reference point of statistical inference, distortion could be classified and, in particular, identified as data contamination or model deviation. The present chapter considers the *implications* of distortion. First of all, it is the performance of the inference procedure(s) which distortion might affect. However, the overall inference process could also be influenced through the *direct* contributions of possibly wrong model assumptions (if they are relevant, see § 3.3.2).

Section 4.2 will commence with some preparatory discussions and notation. What is the impact of distortion on an experiment with statistical inference? Special attention will be devoted to the possibly different implications of data contamination and model deviation. One might expect that the implications become more serious with *increasing* distortion. Hence, distortion will be explained in a quantitative way by implementing the idea of *amount of distortion*. The section concludes with a basic notation for *performance statistics*.

In section 4.3, the notion of *influence* and *preference graphs* is introduced. Both serve the performance assessment of inference procedures under increasing distortion from the frequency point of view, i.e. contemplating distortion as a repetitive event. While influence graphs consider the change of inferential performance under varying levels of distortion, preference graphs likewise compare (combined) choices of inferences procedure(s) and/or assumptions. Especially the *simultaneous* study of either inference or preference graphs under data contamination and model deviation can promise interesting results (see the chapters 5 and 6 for applications of the methodology).

The chapter will close with a brief and critical reflection on the various existing approaches towards inferential performance assessment under distortion (section 4.4). Also the approach of the present work will be summarized at this point in order to prepare for the 'applied' chapters 5 and 6.

4.2 Preliminaries

4.2.1 The impact of distortion

The success of an experiment based on statistical inference is dependent on the 'correspondence' of real-world, data, and model assumptions (agreement of REALITY, DATA, and ASSUMPTIONS, see chapter 3). Only in this case can one hope to reach (absolute) optimal inferential performance. This relates to the attributes *closeness* and *nice type of distribution* discussed for inference procedures in § 2.4.2. Optimal performance of the overall inference process, when real-world description is based on both resulting estimates and model assumptions (refer to Figure 3.3), may be seen in the same way. In this case, the latter is considered as a single *compound* inference procedure.

In any case, distortion means that the optimal inferential performance is potentially missed. With the distinction drawn in chapter 3, the question is now: When is such an influence *different* under data contamination and (type ②) model deviation?

Even though proposed by the model assumptions, the performance of an inference procedure is finally determined by the data (with reference to the real-world if the location is of interest). This is because inference procedures operate on the sample which represents the underlying data. Distortion might therefore imply the following:

- Under *data contamination* the actual results of the inference procedures, and hence their performance, could be modified. This also has consequences for the entire inference process.
- Under *model deviation* the inference procedures could aim at something which in the end is *not* the aspect of interest (due to a wrong formal representation of the latter through the model assumptions). The outcome of the overall inference process might moreover be affected through the *direct* contribution of the wrong model assumptions.

This seems to mean the same when considering the data (and hence any situation of data contamination) as a repetitive event. Still, a difference becomes apparent with performance descriptions based on closeness in terms of *location* (location-closeness). Given the DATA, the two types of distortion differ with respect to the REALITY – in the case of (plain) type ② model deviation the REALITY agrees with the DATA, whereas under data contamination it does *not*.

Potential differences in the influence of distortion can therefore be detected, if performance is assessed *relative to* REALITY and, within reason, in particular relative to the unknown aspect of interest. This also means that *especially* the parts of DATA or ASSUMPTIONS, which represent the aspect of interest, must be subject to distortion. In other words: The influence under data contamination and model deviation *cannot* be distinguished when

- the performance assessment does not refer to REALITY.
- the distortion only affects the dummy-part of the model.

If in example 5 the stochastic part of the regression model is subject to distortion, the estimator $\hat{\boldsymbol{\beta}}$, presumably no longer the ML-estimator, might loose the asymptotic normal property (see chapter 3). However, independently from how the performance of $\hat{\boldsymbol{\beta}}$ is assessed, the influence will be the *same* for both types of distortion, since the deterministic part of the model which incorporates the aspect of interest (average protein contents) remains unaffected.

Data contamination and model deviation targeting the independence of Z and Δ in the Koziol-Green model (described in example 4) is associated with different realities relative to DATA. Here, the description of the aspect of interest (survival time distribution) is affected by distortion since it is directly related to the above independence requirement. In fact, a corresponding difference in performance of the ACL-estimator becomes apparent when studying the former in terms of *location* (see chapter 5 for a detailed discussion of this example). It also shows, that this difference becomes more obvious, the further away Z and Δ are from independence. The implied notion of *increasing distortion* will be introduced in the following.

4.2.2 Quantification of distortion

A study of inferential performance under *increasing* distortion requires a quantification of the notion of model disagreement (§ 3.4.1). For this, the respective models need to be 'numerically comparable'. That is, they can be different from each other as long as their parameters still inherit the same meanings. A simple example would be the Weibull and the exponential model, where the latter is a sub-model of the former.

Model discrepancy

The term *discrepancy* shall describe the 'distance' from some ideal reference model. Then, the *discrepancy magnitude* (a positive value) or the *discrepancy*

structure from which a *discrepancy magnitude* can be derived, is embedded into the model in the form of a parameter (taken in the widest sense).

A discrepancy magnitude of value zero (or one) shall represent the ideal reference model. As the 'neutral element' of the corresponding parametrisation, it does not explicitly appear in the (ideal) model. The model is therefore of simpler structure. This conforms to the idea of model expansion as discussed in § 2.3.2.2. Note that here the expression 'ideal' is not to be seen in the sense of 'true' and that the ideal reference model may or may *not* represent REALITY.

In example 4, the *discrepancy* from the ideal Koziol-Green model corresponds to how 'far away' the random variables Z and Δ are from independence. An appropriate *discrepancy structure* could therefore be loosely defined as the function $s(z) =$ 'probability($\Delta = 1$ given that $Z = z$)'. The latter is constant under independence of Z and Δ. A measure related to the area between $s(z)$ and this constant line might further describe a *discrepancy magnitude*. See chapter 5 for a detailed discussion.

Amount of distortion

Model disagreement between REALITY, DATA, and ASSUMPTIONS can now be quantified by assigning a discrepancy magnitude (`dm`) to each of the three models in comparison to a chosen reference model. Since parameters are generally *not* fixed in ASSUMPTIONS (§ 3.4.1), the discrepancy magnitude in this model needs to be 'frozen' to a specific value, say $\alpha = 0$ as the 'neutral element', or it just remains pending ($\alpha = +$). Then, the *amount of distortion* can be defined as the absolute difference between

- `dm`(DATA) and `dm`(REALITY) under data contamination,
- `dm`(ASSUMPTIONS) and `dm`(REALITY) under type ② model deviation, and

- `dm`(ASSUMPTIONS) and `dm`(DATA) under type ③ model deviation. In this case, a reference to REALITY is *not* made.

The difference between $\alpha \in [0, \infty)$ and $\alpha = +$ shall be defined as 0, which is reasonable since a model with $\alpha = +$ embraces all other models with positive discrepancy magnitudes. Suitable quotients (on a logarithmic scale) could be taken, if $\alpha = 1$ reflects the 'neutral element' of the parametrisation. This case, however, will be neglected for reasons of simplicity.

The *amount of distortion* is one of the requirements for inferential performance description under *increasing* distortion. Also needed are so called performance statistics which will be re-introduced in the following.

4.2.3 Performance statistics

Performance statistics, as they appear in the literature for the non-distorted (ideal) case, have already been discussed in § 2.4.3. They describe the performance (in terms of *closeness*) of statistical inference procedures in a quantifiable way, either as expected values (finite sample view), or as limiting values (asymptotic view).

In the present section the idea of a performance statistic will be re-introduced and in particular cover inferential performance *under distortion*. The discrepancy magnitudes of REALITY, DATA, and/or ASSUMPTIONS relative to some ideal model will serve as the underlying reference frame. Performance will address the *closeness* of inference procedures and attention will be restricted to the finite sample view of performance, i.e. to the expected values of performance statistics.

Altogether, performance statistics shall be denoted as

$$\pi(R = \alpha_1,\ D = \alpha_2\ |A = \alpha_3,\ \Delta), \tag{4.1}$$

where α_1, α_2 and α_3 are discrepancy magnitudes, and where R stands for REALITY, D for DATA, A for ASSUMPTIONS, and Δ for the inference

procedure used. Here and elsewhere the vertical line should *not* be read as introducing a conditional event. Instead it separates the components A and Δ from R and D, where the two former are chosen by the statistician. Depending on the problem and the interests of the study, the R and/or A component may not be relevant and can be ignored. Note that the expected value of π must refer to the model DATA, since inference procedures are functions of the underlying data.

Performance statistics for *location*, associated with the expected value, of some selected examples from chapter 3 are the following (the meaning of the discrepancy magnitudes shall not be discussed here; instead refer to the detailed examples in chapters 5 and 6).

Example 7

$$\mathrm{E}\left[\pi(R, D \mid A, \hat{\rho})\right] = \mathrm{E}_{\mathrm{DATA}}(\hat{\rho}) - \rho,$$

which is simply the bias of estimator $\hat{\rho}$. The reference to REALITY is given by the unknown constant ρ, and the ASSUMPTIONS are not directly involved. The expected value is based on DATA.

Example 4

$$\mathrm{E}\left[\pi(R, D \mid A, S_X^{ACL})\right] = \mathrm{E}_{\mathrm{DATA}}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[S_X(Z_i) - S_X^{ACL}(Z_i)\right]^2\right\},$$

where the unknown survival function S_X reflects the REALITY and the ACL-estimator S_X^{ACL} operates on a DATA-generated sample. The ASSUMPTIONS contribute as described in § 3.3.2, example 4. See the following chapter for a study based on this performance statistic.

Example 5

$$\mathrm{E}\left[\pi(R, D \mid A, \widehat{\boldsymbol{Y}})\right] = \mathrm{E}_{\mathrm{DATA}}\left\{\frac{1}{n}\sum_{j=1}^{n}\left[\mathrm{E}(Y(j)) - \widehat{Y}(j)\right]^2\right\}, \qquad (4.2)$$

which is the expected average squared distance between the true and estimated average protein content of one diet group. The term $\mathrm{E}(Y(j))$

is the expected (under REALITY) protein content at measurement time j in that group, and hence part of the unknown aspect of interest. $\widehat{Y}(j)$ receives contributions from $\hat{\boldsymbol{\beta}}$ and the deterministic part of the regression model as defined in ASSUMPTIONS and can be seen as a compound estimator (as well as $\widehat{\boldsymbol{Y}}$). See also the study in chapter 6.

Note that the expected value of a performance statistic may also *not* refer to REALITY at all, such as the *variance* of an estimator.

The idea of performance statistics in this section does not seem to be very different from what has been already presented in chapter 2. Indeed, it is rather the new notation, which should call attention to the fairly hidden involvement of the models REALITY, DATA, and ASSUMPTIONS. The influence and preference graphs introduced in the next section will make extensive use of the latter.

4.3 Performance assessment under distortion

Influence and *preference graphs* provide tools for the study of inferential (finite sample) performance under increasing distortion. While influence graphs consider changes in performance under varying amounts of distortion, preference graphs likewise compare performance with alternative inference procedures and/or model assumptions. Overall, the performance is assessed in the *long-term*, meaning that a situation of distortion for a given sample size is to be considered as a repetitive event. For this purpose, both influence and preference graphs are based on expected values.

Note beforehand, that influence graphs as discussed by Cook (1986) and others do follow a different, though related, intention (see later in § 4.4.7).

4.3.1 Influence graphs

An *influence graph* indicates the expected change in performance for a given choice of inference procedure and model assumptions under increasing distortion. Performance statistics are denoted as in (4.1) and their expected values refer to DATA. Then, an influence graph can be defined as a function of the *amount of distortion* according to

$$g_i^D(\alpha) = \mathrm{E}_\alpha\Big[\pi(R=0, D=\alpha \mid A=0, \Delta)\Big] - \mathrm{E}_0\Big[\pi(R=0, D=0 \mid A=0, \Delta)\Big] \tag{4.3}$$

for studies under *data contamination*,

$$g_i^M(\alpha) = \mathrm{E}_\alpha\Big[\pi(R=\alpha, D=\alpha \mid A=0, \Delta)\Big] - \mathrm{E}_0\Big[\pi(R=0, D=0 \mid A=0, \Delta)\Big] \tag{4.4}$$

for studies under type ② *model deviation*, and

$$g_i^*(\alpha) = \mathrm{E}_\alpha\Big[\pi(R, D=\alpha \mid A=0, \Delta)\Big] - \mathrm{E}_0\Big[\pi(R, D=0 \mid A=0, \Delta)\Big]$$

for studies under type ③ *model deviation*.

The component R in $g_i^*(\cdot)$ remains unspecified, since type ③ distortion occurs independently from the real-world. Still, performance assessment could refer to REALITY, e.g. in terms of the bias. In return, the R (or A) component may be ignored for the performance description in the above influence graphs, though being relevant for the specification of distortion.

The discrepancy magnitude of $\alpha = 0$ shall correspond to an ideal reference model such as the Koziol-Green model in example 4. Any other (strictly positive) value for α belongs to some *distorted model* which is 'α discrepancy-units' away from the ideal reference model (see § 5.3 for a distorted Koziol-Green model). Note that here the expression 'distorted' is seen in the sense of 'modified' and that a distorted model, like the ideal reference model, may or may not represent REALITY.

The first part of each equation refers to performance under distortion, and the discrepancy magnitudes of R, D, and A therein determine the 'current' *amount of distortion* for the influence graph (see § 4.2.2). The second part of each equation reflects the undistorted situation. Hence, based on differences from the ideal situation, an influence graph considers *absolute* changes in performance. The differences might also be replaced by quotients in order to consider *relative* changes in performance.

In $g_i^D(\cdot)$ it turned out to be reasonable to see REALITY as the simpler ideal model ($\alpha = 0$) and DATA as the more complex (distorted) model. This is because data contamination usually contributes to the complication of the given data structure. The ASSUMPTIONS in $g_i^M(\cdot)$ and $g_i^*(\cdot)$, at the same time, were associated with $\alpha = 0$ because of the common wish to use and verify simple model assumptions. Finally, the ASSUMPTIONS in $g_i^D(\cdot)$ were represented by $\alpha = 0$ in order to facilitate a direct comparison with the influence graph $g_i^M(\cdot)$.

4.3.2 Preference graphs

While influence graphs study performance as such, *preference graphs* serve to *compare* performance under increasing distortion. They aim at indicating which choice of inference procedure and/or model assumptions is expected to perform better (or worse) under a given amount and type of distortion. As before, this can be done in either absolute or relative terms. Depending on the final objective of comparison three different types of preference graphs can be defined.

4.3.2.1 Comparison of two inference procedures

For the comparison of two inference procedures/methods Δ_1 and Δ_2 both associated with the *same* ASSUMPTIONS, preference graphs can be expressed

by the following:

$$g_p^D(\alpha) = \mathrm{E}_\alpha\Big[\pi(R=0, D=\alpha \mid A=0, \Delta_1) - \pi(R=0, D=\alpha \mid A=0, \Delta_2)\Big]$$

for studies under *data contamination*,

$$g_p^M(\alpha) = \mathrm{E}_\alpha\Big[\pi(R=\alpha, D=\alpha \mid A=0, \Delta_1) - \pi(R=\alpha, D=\alpha \mid A=0, \Delta_2)\Big]$$

for studies under type ② *model deviation*, and

$$g_p^*(\alpha) = \mathrm{E}_\alpha\Big[\pi(R, D=\alpha \mid A=0, \Delta_1) - \pi(R, D=\alpha \mid A=0, \Delta_2)\Big]$$

for studies under type ③ *model deviation*. As before with the corresponding influence graph, the component R in $g_p^*(\cdot)$ remains unspecified.

In example 5 preference graphs of the kind $g_p^D(\cdot)$ and $g_p^M(\cdot)$ based on the performance measure in equation (4.2) could compare the ML- and REML-estimation method (Diggle *et al.*, 1995, p. 63ff) under increasing distortion. However, only if the distortion affects the deterministic part of the model (addressing the aspect of interest), the comparison will actually lead to different results under corresponding data contamination and type ② model deviation (see § 4.2.1). An alternative comparison *not* referring to REALITY could be based on the (local) estimator variances.

4.3.2.2 Comparison of two choices of ASSUMPTIONS

Preference graphs comparing two choices of ASSUMPTIONS $A = 0$ and $A = +$ (the first is a sub-model of the second) and both associated with the *same* inference procedure/method are derived as follows:

$$g_p^D(\alpha) = \mathrm{E}_\alpha\Big[\pi(R=0, D=\alpha \mid A=0, \Delta) - \pi(R=0, D=\alpha \mid A=+, \Delta)\Big] \tag{4.5}$$

for studies under *data contamination*,

$$g_p^M(\alpha) = \mathrm{E}_\alpha\Big[\pi(R=\alpha, D=\alpha \mid A=0, \Delta) - \pi(R=\alpha, D=\alpha \mid A=+, \Delta)\Big] \tag{4.6}$$

for studies under type ② *model deviation*, and

$$g_p^*(\alpha) = \mathrm{E}_\alpha\Big[\pi(R, D = \alpha \mid A = 0, \Delta) - \pi(R, D = \alpha \mid A = +, \Delta)\Big]$$

for studies under type ③ *model deviation*. Note that only the first part of $g_p^M(\cdot)$ and $g_p^*(\cdot)$ actually refers to performance under distortion.

Consider, for instance, the model of example 5 with extra parametrized covariance structure between repeated measurements. In this case, random variation *within* data units could be explained in terms of random effects, serial correlation, and measurement error (Diggle *et al.*, 1995, p. 79ff). Preference graphs of the types $g_p^D(\cdot)$ and $g_p^M(\cdot)$ can compare resulting performances of the REML-method under ASSUMPTIONS which either include or *do not* include $N(0, \tau^2)$-distributed measurement errors. The discrepancy magnitude could be represented by the parameter τ, i.e. $A = 0$ corresponds to the model *without* a measurement error component (see chapter 6 for further details).

4.3.2.3 Comparison of two inference procedures based on alternative ASSUMPTIONS

Certain inference procedures are known on their own and, at the same time, *imply* the ASSUMPTIONS made. Then, performance comparisons distinguished so far simplify to a comparison of two single inference procedures Δ_1 and Δ_2 which are based on alternative assumptions. Preference graphs for this case are determined according to

$$g_p^D(\alpha) = \mathrm{E}_\alpha\Big[\pi(R = 0, D = \alpha \mid A = 0, \Delta_1) - \pi(R = 0, D = \alpha \mid A = +, \Delta_2)\Big] \tag{4.7}$$

for studies under *data contamination*, and

$$g_p^M(\alpha) = \mathrm{E}_\alpha\Big[\pi(R = \alpha, D = \alpha \mid A = 0, \Delta_1) - \pi(R = \alpha, D = \alpha | A = +, \Delta_2)\Big] \tag{4.8}$$

for studies under type ② *model deviation*, and

$$g_p^*(\alpha) = \mathrm{E}_\alpha\Big[\pi(R, D = \alpha \mid A = 0, \Delta_1) - \pi(R, D = \alpha \mid A = +, \Delta_2)\Big]$$

for studies under type ③ *model deviation*.

The ASSUMPTIONS represented by the Koziol-Green model in example 4 are directly related to the ACL-estimator, i.e. the latter implies the model assumptions. A comparison of the ACL-estimator (Δ_1) with the Kaplan-Meier estimator (Δ_2) is therefore naturally associated with a comparison of the Koziol-Green model assumption ($A = 0$) and the more general right random censorship model assumption ($A = +$). See chapter 5 for a detailed discussion.

4.4 Approaches in the literature

The concepts of distortion and performance, as they appear in the literature, have been described – with only little connection – in chapter 2. Our view of distortion and our approach of describing performance (*under* distortion) have been introduced in chapter 3 and in the present chapter. Now, we are in a good position to reflect (from our point of view) on the still outstanding connection of distortion and performance in the statistical literature. Independently from what has been claimed, which type of distortion is actually considered in the various approaches? In particular, is it *data contamination* or *model deviation*? If not already obvious by the formalization of distortion itself, the question(s) can usually be answered by looking at the specific performance descriptions which refer to the aspect of interest in the real-world.

An interesting observation will be made: While Cook's (1986) approach to "perturbation diagnostics" and the concept of configural polysampling seem to address model deviation, the other more classical approaches towards robustness deal with data contamination. Nevertheless, only the main

approaches and references will be discussed. It remains to be seen whether other publications might be oriented in a different way. Before the actual discussion, we will briefly summarize our approach.

4.4.1 Our approach (summary)

Distortion is modelled and quantified according to the idea of model expansion by taking into account a so called discrepancy magnitude. The amount and type of distortion can be specified by the comparison of corresponding discrepancy magnitudes associated with the model triplet of REALITY, DATA, and ASSUMPTIONS. Overall, *any* type of distortion, i.e. in particular model deviation as well as data contamination, can be considered.

Inferential performance in terms of *closeness* is described by the expected value of suitable performance statistics. The latter can be defined for the ideal situation and also under distortion, again just by referring to the corresponding model triplet. Performance under *increasing* distortion can be studied with so called influence and/or preference graphs.

4.4.2 Qualitative robustness

With the condition that the inference procedure under study is a consistent functional T, the formal definition of *qualitative robustness* is equivalent to continuity of T at some ideal model distribution F_0. This heuristically corresponds to stable behaviour of $T(F_n)$ under small changes in the underlying sample (Huber, 1981, p. 9f, 41). Qualitative robustness addresses therefore distortion in form of *data contamination*, while inferential performance itself is globally expressed by the behaviour of $T(F_n)$, i.e. without involving values from a performance statistic.

4.4.3 Quantitative robustness

Huber's minimax approach is based on distortion neighbourhoods around some ideal model distribution F_0. The neighbourhoods themselves are either defined by a distance (§ 2.3.2.4) or they are described by the so called gross-error model (§ 2.3.2.3). Originally, they address the *shape* of a model distribution (Huber, 1981) and hence could principally refer to distortion of type ① or ② (data contamination or model deviation, see § 3.4). Extensions for type ③ distortion have also been proposed, but they will not be of interest in the subsequent discussion.

The inference procedure under study is represented as a functional T which is asymptotically normal and consistent at some 'underlying' distribution F (Huber, 1981, p. 10f). Note, that it is not very clear whether T should be consistent under F_0 as well (also remarked in Hampel *et al.* (1986), p. 48).

Inferential performance (under the ideal *and* distorted situation) is then described in terms of the asymptotic bias and variance. The former, in particular, is formulated as $T(F) - T(F_0)$. While $T(F_0)$ is to be estimated and as such reflects the real-world, the term $T(F)$ as the limit in probability of $T(F_n)$ relates to the *data*. A reference to any model assumptions is not provided, unless the inference procedure itself would imply the latter. Thus, it seems that the present approach considers *data contamination* only, and this even though Huber (1981) also refers to assumptions which "are not supposed to be exactly true" (compare with the citations in chapter 2).

In detail, the possibility of model deviation seems to be excluded because of the following argumentation: Under (plain) model deviation the limit of $T(F_n)$ would coincide with the quantity $T(F_0)$ to be estimated, so that the asymptotic bias as formulated by Huber would be zero. This is because the data is generated correctly, i.e. both DATA and REALITY are represented by F_0, *and* in case T is consistent at F_0. The difference $T(F) - T(F_0)$ should *not* be interpreted as "asymptotic bias" at all, if after all consistency at F_0 *cannot* be guaranteed. In this case the term $T(F)$ almost certainly does *not*

relate to F_0-distributed data.

Under model deviation, it would therefore be sensible to use a bias description which does not rely on the consistency of T such as the expression

$$\mathrm{E}_{F_0}\Big[T(F_n) - T(F_0)\Big]$$

in either finite or asymptotic terms. The mathematical implications of the latter shall not be discussed in this work.

4.4.4 Robustness approach based on influence functions

The approach based on influence functions studies inferential performance in full, but infinitesimal distortion neighbourhoods (§ 2.3.2.5). This means that independently from the reference model, the neighbourhoods can principally address *any* type of distortion. Larger amounts of distortion are also considered, especially by the use of the breakdown point. At this point, however, we will restrict ourselves to the discussion of the influence function and distortion affecting the shape of a model distribution. As a first derivative, the influence function contributes to extrapolate the functional T, representing the inference procedure under study, into the distortion neighbourhood (Hampel *et al.*, 1986, p. 42, Figure 2). Moreover, it provides the basic performance description for T, which in particular, is considered to be (generally) consistent.

According to Hampel *et al.* (1986, p. 84) the influence function of T at some reference model F is defined as

$$IF(x; T, F) = \lim_{t \to 0} \frac{T\{(1-t)F + t\Delta_x\} - T(F)}{t},$$

where Δ_x is the "probability measure which puts mass 1 at the point x". The point x plays roughly "the role of the coordinate in the infinite-dimensional space of probability distributions" (Hampel *et al.*, 1986, p. 41) and as such describes a particular location in the infinitesimal distortion neighbourhood.

In other words, distortion is quantified in terms of location, and *not* in terms of distance from some ideal reference model (our approach).

Overall, the influence function is supposed to measure "the asymptotic bias caused by contamination in the observations" (Hampel *et al.*, 1986, p. 84). While $T\{(1-t)F + t\Delta_x\}$ is the asymptotic value of the inference procedure operating on $(1-t)F + t\Delta_x$-distributed *data*, the term $T(F)$ represents the unknown quantity from the real-world. Again, no explicit reference is ascribed to the underlying model assumptions, unless the inference procedure itself would imply the latter. Thus, also Hampel's influence function seems to be devoted to *data contamination* only (as already indicated in the previous citation). The same line of argument as for Huber's approach can be used: An asymptotic bias formulation in the above way is only sensible as long as DATA and REALITY correspond to *different* statistical models. This is however only the case under data contamination.

4.4.5 Outliers

An active area related to distortion is that of *outliers* in statistical data. It is treated in robustness as well as diagnostics and is therefore not associated with just a single (robustness) approach. Still, one of the main references is the book by Barnett and Lewis (1995), which will be considered in the following brief discussion.

Outliers are seen as observations which appear to be inconsistent with the remainder of the data and are a surprise to the data analyst (Barnett and Lewis, 1995, p. 7, 460). In principle they could imply data contamination *or* model deviation (as type ② distortion) which has already been indicated by respective citations of Barnett and Lewis in chapter 2. The two possibilities are also reflected in the way outliers are being modelled. While some outlier models explain rather a situation of data contamination such as the ones of slippage type (§ 2.3.2.7), others seem to be more realistic for model deviation (e.g. the inherent alternative, § 2.3.2.1). A mixture model (§ 2.3.2.3) might

finally address *any* of the two distortion types.

In terms of methodology, outliers are handled in two different ways. Firstly, they may or may not be rejected after detection through a (so called) discordancy test. The act of rejection is obviously only sensible when outliers actually *contaminate* the *data*. Alternatively, there might be the wish to accommodate outliers with the use of robust procedures. Hence, the study of inferential performance becomes important, which is this time under distortion through outliers.

Even though outliers could either contaminate the data *or* reflect incorrect model assumptions, Barnett and Lewis (1995) seem to restrict performance descriptions exclusively to the situation of *data contamination*. This becomes apparent when they discuss the bias of estimators. According to the authors, for example, the estimator $\bar{X} = 1/n \cdot \sum X_i$ is biased under "asymmetric contamination" (p. 62f). This is however only possible in the face of (asymmetric) data contamination. In fact, $\bar{X}$ would remain unbiased under corresponding model deviation, since DATA and REALITY would agree with each other. Moreover, Barnett and Lewis refer to the approach based on influence functions, which already before has been associated with data contamination only.

4.4.6 Configural polysampling

Configural polysampling is a finite sample approach towards robustness (Morgenthaler and Tukey, 1991; Morgenthaler, 1991). It is aimed at invariant models such as the ones of location/scale or regression type. Distortion is expressed by a finite number of model alternatives called "confrontations" which usually address the shape of a distribution (§ 2.3.2.1). For the estimation a location-parameter this could be the pair of distributions {Gaussian, Slash} where the Slash is a heavy-tailed alternative to the normal distribution (Morgenthaler, 1991). Note that *no* quantification of distortion has been undertaken. At first sight the approach might again handle *either*

data contamination *or* type ② model deviation, since aspects of the model *within* data units are considered to be distorted.

Performance of an estimator is described in terms of the mean-squared error (MSE). A 'robust' estimator can be found by choosing a set of alternative distributions such as $\{F, G\}$ and by minimizing the vector criteria

$$\big(\mathrm{MSE}_F(T_n), \mathrm{MSE}_G(T_n)\big)$$

over all potential estimators T_n. The approach is based on numerical sampling techniques and presented in Morgenthaler and Tukey (1991). Here, we will stress upon the following fact:

$$\mathrm{MSE}_F(T_n) = \mathrm{E}_F\left\{\big[T_n(X_1, \ldots, X_n) - \mu(F)\big]^2\right\} \tag{4.9}$$

is "the mean-squared error under sampling from distribution F ... where $\mu(F)$ denotes the center of symmetry of F or some other prespecified target" (Morgenthaler and Tukey, 1991, p. 38f, in the original $F(y)$ instead of F). Thus, the aspect of interest $\mu(F)$ belongs to the *same* model distribution F from which also the data is generated (the expected value refers to F). In other words, the models DATA and REALITY correspond to each other and *no* data contamination can be considered.

Even though the model assumptions are not explicitly spelled out in (4.9) we conclude that configural polysampling addresses the issue of *model deviation*. The approach is used "to select, to fine-tune, and to design procedures" (Morgenthaler, 1991, p. 50) and it tries to *avoid* model deviation. This is achieved by not making specific model assumptions, but to choose a *set* of alternative assumptions which retain the required interpretation for the aspect of interest (e.g. mean, median, or mode for the location problem). The set of assumptions is then used to define the above optimization problem.

4.4.7 Perturbation diagnostics

The methodology of perturbation diagnostics as presented by Cook (1986) studies the influence of "model perturbation" (his words). As similar to

our approach, a distorted model is described by some instance ω which is additionally associated with the model. However, the instance does not seem to be a parameter in the usual sense (see below), and it might not necessarily reflect a *distance* from the ideal reference model. Overall, "ω can reflect any well-defined perturbation scheme" (p. 136), i.e. it might at first sight apply to any of the distortion types ① to ③.

Cook (1986) introduced the so called likelihood displacement

$$LD(\omega) = 2 \cdot \Big[L(\hat{\theta}) - L(\hat{\theta}_\omega) \Big]$$

and its alternative

$$LD'(\omega) = 2 \cdot \Big[L(\hat{\theta}_\omega |\ \omega) - L(\hat{\theta} |\ \omega) \Big],$$

where $L(\cdot |\ \omega)$ and $L(\cdot)$ are the log-likelihood functions of the distorted and the ideal model, respectively, and $\hat{\theta}_\omega$ and $\hat{\theta}$ are the corresponding model-associated ML-estimators for θ.

The value of a log-likelihood of $\hat{\theta}$ quantifies how well this estimator 'reached' the true value θ given the underlying data. It therefore describes the *performance* of $\hat{\theta}$ in terms of location-closeness, while the *true* value θ refers to REALITY (considering it as part of aspects of the model within data units, see chapter 3). Hence, using the notation of performance statistics (§ 4.2.3) and indicating the ideal reference model by $\omega_0 = 0$ (for simplicity), the right side of the equations above could be expressed as (multiplied the factor -2)

$$\pi(R = 0, D\ |A = \omega,\ \hat{\theta}_\omega\ ) - \pi(R = 0, D\ |A = 0,\ \hat{\theta}\ ) \tag{4.10}$$

and

$$\pi(R = \omega, D\ |A = 0,\ \hat{\theta}\ ) - \pi(R = \omega, D\ |A = \omega,\ \hat{\theta}_\omega\ ), \tag{4.11}$$

respectively.

Obvious parallels to preference graphs of our approach, which study *model deviation*, become clear: The first term of each expression describes performance under some kind of *type* ② *model deviation* (disagreement between ASSUMPTIONS and REALITY). Moreover, ML-estimators are generally defined through the underlying model assumptions, and one might also say that, vice versa, they *imply* the ASSUMPTIONS made. On the other hand, the estimators are *not* known on their own, as such, and it is rather the comparison of different model assumptions which seems to be of interest. Thus, Cook's likelihood displacements are related to the preference graphs introduced in § 4.3.2.2 and § 4.3.2.3. Still there are differences:

- Rather than comparing the *same* two estimators/model assumptions for *all* ω, the estimators themselves change with varying values for ω and as well do the ASSUMPTIONS. This is because the latter are associated with *fixed* values for ω. Thus, ω cannot be seen as a parameter in the usual sense, it rather represents some mathematical quantity introduced into the model (as also suggested by Lawrance (1991), p. 142).

- The component D for DATA remains unspecified. In contrast to the interpretation of a likelihood, the expressions in (4.10) and (4.11) could therefore be seen as functions of an immediate data sample, and further also as some statistic based on the (unspecified) DATA-model. In line with the idea of a preference graph one would then have to consider its expected value with respect to DATA.

As remarked before, the likelihood displacement as such is aimed at type ② model deviation. Distortion of type ③ is generally excluded, since a likelihood principally addresses the distributional part of a model and thus model aspects *within* data units. In order to apply Cook's approach to data contamination, a model for DATA needs to be specified and this in corresponding disagreement with REALITY and ASSUMPTIONS.

A final remark shall be made about Billor and Loynes's paper (1993). They suggest an alternative to Cook's likelihood displacement according to

$$LD^*(\omega) = -2 \cdot \Big[L(\hat{\theta}) - L(\hat{\theta}_\omega |\ \omega) \Big]$$

which in our notation leads to the expression (ignoring the factor -2)

$$\pi(R = 0, D\ |A = 0,\ \hat{\theta}\,) - \pi(R = \omega, D\ |A = \omega,\ \hat{\theta}_\omega\,). \tag{4.12}$$

The authors claim to study perturbations in form of data contamination or model deviation. Still, neither of the two terms in (4.12) do actually reflect a situation of distortion. From our point of view, it rather compares performance under two ideal (non-distorted) situations.

Final remark Up to this point, theoretical and methodological aspects of distortion and performance have been discussed. The following two chapters will now, in contrast to the preceding, present the *practical* study of two specific examples. Both are based on simulation.

Chapter 5

First detailed example: The Koziol-Green model

5.1 Introduction

In the previous chapters, distortion and inferential performance assessment (under distortion) have been discussed on solely theoretical grounds. The present chapter will now illustrate the ideas with a first detailed and *practical* example. It will consider the Koziol-Green (KG) proportional hazards model (Koziol and Green, 1976) which can be used in the censored survivals and in the competing risks framework.

Being semi-parametric, the KG-model is a special case of the right random censorship model. The latter is characterized by the independence of the actual unknown survival and censoring times. The *additional* requirement of independence between the observed survival times and them being or not being censored is the essence of the KG-model. While the Abdushukurov-Chen-Lin (ACL) estimator (Abdushukurov, 1984; Cheng and Lin, 1984) is the maximum-likelihood estimator for the true survival function in the KG-model, its counterpart in the right random censorship model is the well-known Kaplan-Meier (KM) product limit estimator (Kaplan and Meier,

1958). Both estimators have been compared (point-wise asymptotically) *under* the KG-model in Cheng and Lin (1987) and Csörgő (1988).

Not surprisingly, it is the additional (KG-) requirement which shall be subject to distortion. In this way both types of distortion, data contamination as well as model deviation, are theoretically possible, since the KG-requirement relates to aspects of the model *within* data units. It becomes sensible to study performance of the ACL-estimator under the above types of distortion, in particular in comparison with the competing KM-estimator. Note, that in practice the latter is usually preferred. A simulation study based on influence and preference graphs will in fact indicate a trade-off in finite-sample performance of the two estimators between data contamination and model deviation.

After defining the KG-model and the ACL-estimator, and presenting the corresponding model and inference framework in section 5.2, distortion of the KG-model is discussed and formalized in section 5.3. The simulation study addressing the ACL-performance supported by influence and preference graphs (section 5.4) follows in section 5.5. The chapter concludes with some final remarks (section 5.6).

5.2 The Koziol-Green model and the ACL-estimator

5.2.1 The model and the model framework

The Koziol-Green (KG) model can be used in experiments with censored survival or competing risks data. Respective *real-world situations* in the first case are typically circumstances in which survival times of a "biological unit (patient, animal, cell, etc.)" or failure times of a "physical component (mechanical or electrical)" are of interest (citations from Miller (1981), p. 1). Parallel to the unknown survival distribution F_X (often the *aspect of interest*)

one has to take into account a second positive distribution, the censoring distribution F_Y, usually reflecting times at which the follow-up process of an individual *sampling unit* (e.g. patient) is interrupted due to reasons other than death or failure. This might for example be the patient's decision to "drop out", but could also be the statistician's decision to terminate the study. See also example 4 in chapter 3 for a typical censored survival problem.

If the KG-model is used in the *competing risks framework* both distribution functions, F_X and F_Y, refer to 'proper' survival or failure times. As such they could for instance address two exclusive causes of death for a patient, or they might respectively be assigned to the failure time of two individual sub-components in a physical system. According to the scientific problem at hand, the *aspect of interest* may then e.g. be the overall survival, or the survival of the individual risks.

Despite their conceptual differences, problems as above with censored survivals as well as with competing risks deal with the same type of data structure: In both cases there is one univariate observation per *data unit* (a positive survival or failure time) which is associated with a censoring indicator Δ. While Δ tells whether the respective observation is censored or not in survival data, it indicates which component of the two under study failed at the particular point of time in competing risks data.

Model definition The KG-model for a corresponding *data-generating process* is defined as follows (using the more popular notation from the censored survival context):

Let $X_1, \ldots, X_n$ represent the true survival times, $Y_1, \ldots, Y_n$ the censored, and $Z_1, \ldots, Z_n$ the observed values, where each sequence of these continuous, positive random variables is i.i.d. like X, Y, and Z, respectively. The corresponding distribution and survival functions are given by $F_X(\cdot) = 1 - S_X(\cdot)$, $F_Y(\cdot) = 1 - S_Y(\cdot)$, and $F_Z(\cdot) = 1 - S_Z(\cdot)$. Further, let the sequence of censoring indicators $\Delta_1, \ldots, \Delta_n$ be i.i.d. like the $\mathrm{Bin}(1, p)$-distributed random variable Δ, where

$\Delta = \imath(Z = X)$ with $\imath(\cdot)$ as the indicator function,

$Z = \min\{X, Y\}$, and

$p = P(\Delta = 1)$.

The right random censorship model is (already) characterized by the independence between X and Y, which implies that

$$S_Z(z) = S_X(z) \cdot S_Y(z) \quad \text{for } z \in [0, \infty). \tag{5.1}$$

The *additional* KG-requirement of independence between Z and Δ leads to the *Koziol-Green model*, which is equivalent to the existence of a positive constant c, so that $S_Y(z) = S_X(z)^c$ for $z \in [0, \infty)$. Thus, under the KG-model $p = (1 + c)^{-1}$ and

$$S_X(z) = S_Z(z)^p \quad \text{for } z \in [0, \infty). \tag{5.2}$$

Since the (cumulative) hazard functions $\Lambda_X(z) = -\log S_X(z)$ and $\Lambda_Y(z) = -\log S_Y(z)$ are proportional under the KG-model, the latter is also known as the "proportional hazards model of random censorship" (Csörgő, 1988).

In contrast to the non-parametric right random censorship model, the KG-model can be seen as semi-parametric with one parameter (the censoring parameter p). Even though it does not involve covariates of any kind and addresses data units with only single observations, the KG-model is rather complex due to censoring. While the independence within the four sequences $X_1, \ldots, X_n$, $Y_1, \ldots, Y_n$, $Z_1, \ldots, Z_n$, and $\Delta_1, \ldots, \Delta_n$, respectively, corresponds to aspects of the model *between* data units, all remaining conditions refer to aspects *within* data units.

Model interpretation In clinical trials, where patients may enter the study at different times and losses to follow-up occur randomly, the random censorship model seems to be justified (Miller, 1981, p. 5f). The above independence requirement between the observed survivals Z and the censoring

indicator Δ for the KG-model would *additionally* mean, that the chance for a loss to appear remains *constant* beyond the start of every follow-up period. This seems to be a quite unrealistic assumption in medical practice, especially when taking into account that the statistician will most likely terminate the study at some point with certain patients still being alive.

However, the KG-model might be more plausible in a competing risks framework as the following examples indicate.

- Suppose a two-component series system which fails when at least one of the components fail. Apart from a power-transformation, their failure time distributions F_X and F_Y are identical.
- Consider the previous situation where each component i, itself, is a series system of k_i independent sub-components, $i = 1, 2$. All sub-components are identically distributed (Chen *et al.*, 1982, p. 142).

The interpretation of the KG-model in terms of the models REALITY, DATA, and ASSUMPTIONS (see § 3.3.1.3) motivates moreover the following remarks:

- The parameter p and the respective survival functions remain *unspecified* in ASSUMPTIONS, but are considered to be *fixed* in DATA and REALITY (compare with § 3.4.1).
- The traditional naming of X, Y, and Z as true, censored, and observed survivals could be misleading. As such they might only be suitable for the DATA-model with *no* data contamination being present (otherwise the 'true' survival time in DATA would not necessarily be true). In the case of REALITY and ASSUMPTIONS one might better refer to true, censored, and *observable* survivals, since Z is not actually observed in REALITY.
- In the context of censored survival data the interpretation of the model triplet may imply a problem. Censoring times produced by the effective

data-generating process (represented by DATA) could be partly under control of the statistician, for example when the statistician decides to terminate the study. Thus, it is not obvious anymore whether the corresponding 'truth' of F_Y should be found in the real-world or in the statistician's choice (see also the remarks in § 3.4.4).

The KG-model has received theoretical attention in the statistical literature in order to develop inferential procedures, such as the ACL-estimator (see below), but also to study properties of procedures formulated for the more general model of right random censorship. Examples for the latter are Koziol and Green (1976) themselves, who derive the asymptotic distribution of a Cramér-von Mises type (goodness-of-fit) statistic, Chen *et al.* (1982) obtaining small-sample results for the Kaplan-Meier (KM) estimator such as the exact bias and variance, Ghorai and Pattanaik (1993) who study the least-squares cross-validation method of bandwidth selection for a kernel density estimator, and more recently Chang (1996) deriving the overall exact distribution of the KM-estimator and Gather and Pawlitschko (1998) who study versions of the Kaplan-Meier integral (estimator of $\int \phi dF$).

However, despite the theoretical interest, the KG-model seems to have been fitted to *actual* data rather seldom. The so called Channing House data (Hyde, 1977) appears to be one of the only cited references in this respect (two recent ones are moreover given in Csörgő (1998)). Nevertheless, the KG-model became known especially because of its maximum-likelihood estimator for the unknown survival function, the ACL-estimator. The latter will be introduced in the following.

5.2.2 The estimator and the inference framework

If the *aspect of interest* is the unknown survival function $S_X = 1 - F_X$ then a suitable estimator is the so called ACL-estimator. According to Csörgő (1988) the original references are Abdushukurov (1984) and Cheng and Lin

(1984). As the maximum-likelihood estimator for S_X under KG-model, the ACL-estimator is defined as (referring to equation (5.2))

$$S_X^{ACL}(z) = \left[1 - \hat{F}_Z(z)\right]^{p_n} \quad \text{for } z \in [0, \infty), \tag{5.3}$$

where $\hat{F}_Z$ is the empirical distribution function based on the random sample of Z and $p_n = 1/n \cdot \sum_{i=1}^{n} \Delta_i > 0$. In the unlikely event of $p_n = 0$ the estimator is defined as $S_X^{ACL}(z) = 1 - \imath(z \geq Z_{n:n})$, where $Z_{n:n}$ is the corresponding n-th order statistic and $\imath(\cdot)$ the indicator function (Csörgő, 1988, p. 440). The estimator, as a single inference procedure, provides a full description of the *aspect of interest*. Still, it is composed of the two individual estimators p_n and $\hat{F}_Z$ through a direct contribution of the KG-assumptions (see § 3.3.2 and the discussion of example 4 therein).

The ACL-estimator is generally smoother than the KM-estimator, since it takes into account uncensored *as well as* censored observations. Further, under the Koziol-Green model, it is (asymptotically) more efficient than the KM-estimator (Csörgő, 1988). Up to a certain quantile of F_X, it outperforms (again asymptotically) the empirical distribution function, which is based on the respective *un*censored sample (Csörgő and Faraway, 1998). Other large-sample properties such as asymptotic unbiasedness and asymptotic normality (weak convergence to a Gaussian process) have been discussed in Csörgő (1988), Cheng and Lin (1987) and Mi (1990).

Like the empirical distribution function and the KM-estimator, the ACL-estimator has served to derive new inference procedures according to the 'plug-in' principle. An example is Csörgő (1988) for the estimation of quantiles, densities, hazards rates and some specific reliability functions. The author also constructed confidence bands based on the ACL-estimator. Further examples are Herbst (1992a) for the estimation of moments, Gijbels and Veraverbeke (1989), Ghorai (1991b), Gronen (1993), and Grunert (1993) for quantile estimation, Stute (1992) and Dikta (1995) for the estimation of $\int \phi dF$, and finally Ghorai (1991a) and Ebrahimi and Habibullah (1992) for goodness-of-fit test statistics based on the ACL-estimator.

Despite its theoretical importance, the ACL-estimator has not, as yet, found substantial practical interest. This is to be seen in clear contrast to the widely used KM-estimator. Obviously not much 'trust' has been placed in the validity of the additional KG-requirement (independence of Z and Δ) and thus to the performance of the ACL-estimator under this particular kind of model deviation. The simulation study in the present chapter will investigate whether or not this concern is justified. Equally of interest therein will be the corresponding, but fundamentally distinct, situation of data contamination. The next section will provide the necessary formalization of distortion.

5.3 The distorted Koziol-Green model

Attention will be given to distortion of the KG-model affecting the independence between the observed survival times Z and the censoring indicator Δ (previously denoted as KG-requirement). The latter is part of the model aspects *within* data units, so that principally distortion of type ① (data contamination) as well as distortion of type ② (model deviation) are possible.

An example for respective *data contamination*

- in the competing risks framework could occur with the two-component series system with, apart from a power-transformation, identical F_X and F_Y as mentioned on page 93. The KG-requirement is not satisfied in DATA (even though true in REALITY), if one of the components is exposed to an *external* influence such as a higher temperature which converts its failure time distribution by a non-power manner during the course of the experiment.

- in the censored survival context is less obvious. First of all, it is generally difficult to find a situation (in clinical trials) where the KG-assumptions seem to be justified. Moreover there is the statistician himself who, to a certain extent, is in control of the independence

between Z and Δ in the DATA-model. Thus, it is not always clear whether a certain change in DATA is actually due to data contamination or is just an effect of interference by the statistician (see also the remarks in § 3.4.4 and § 5.2.1).

Corresponding *model deviation*, on the other hand, is present *whenever* the additional KG-requirement cannot properly reflect the truth. This may be the case in a competing risks framework and is most likely in the context of censored survivals. It particularly occurs in situations where the probability of a failure, e.g. 'eventual conception' in fertility studies, is less than one. For further examples see also Peña and Rohatgi (1989), p. 372.

While data contamination reflects model disagreement between DATA and REALITY, type ② model deviation is due to a conflict between ASSUMPTIONS and REALITY (§ 3.4.2). Still in common, both cases imply a mutual disagreement between the KG-model and some *distorted* KG-model. The latter will be specified in the following by implementing the idea of *model discrepancy* (§ 4.2.2).

5.3.1 Discrepancy structure and magnitude

Discrepancy structure Departure from the KG-model will first of all be explained in form of a *discrepancy structure*. This shall be the function

$$\mathbf{s} : [0, \infty) \longrightarrow [0, 1]$$

with

$$\mathbf{s}(z) = \frac{\frac{\partial}{\partial z}\left[P(Z \leq z, \Delta = 1)\right]}{f_Z(z)} = \text{`}P(\Delta = 1 | Z = z)\text{'} \tag{5.4}$$

for all $z \in [0, \infty)$ with $f_Z(z) > 0$, assuming the derivative exists at all but a finite number of points. From a probabilistic point of view, $\mathbf{s}(\cdot)$ is the conditional mass function of Δ given $Z = z$ at the margin $\Delta = 1$. Compare e.g. with Woodroofe (1975, p. 269).

If $\mathbf{s}(\cdot)$ is constant for all z the KG-model is retrieved. A '*non*-constant behaviour' of $\mathbf{s}(\cdot)$ for some $z \in [0, \infty)$ indicates the dependence of Z and Δ, and thus a violation of the KG-requirement. Under an $\mathbf{s}(\cdot)$-distorted KG-model the survival function of the true survivals S_X can be represented, like in (5.2), by an expression which involves Z as the only random variable. For $z \in [0, \infty)$ with $f_Z(z) > 0$ and $S_Z(z) > 0$

$$S_X(z) = \exp\Big[- \int_0^z \mathbf{s}(t) \cdot \lambda_Z(t)\ dt \Big], \tag{5.5}$$

where $\lambda_Z(t) = f_Z(t)/S_Z(t)$ is the hazard rate of the observed survivals Z. With the assumption that all relevant densities exist, (5.5) is correct since $\frac{\partial}{\partial z}\left[P(Z \leq z, \Delta = 1)\right] = f_X(z) \cdot S_Y(z)$ under independence of X and Y which in particular leads to

$$\mathbf{s}(z) = \frac{f_X(z) \cdot S_Y(z)}{f_Z(z)} = \frac{f_X(z)}{S_X(z)} \cdot \frac{S_Z(z)}{f_Z(z)} = \frac{\lambda_X(z)}{\lambda_Z(z)}, \tag{5.6}$$

and since in general $S(z) = \exp\left[-\int_0^z \lambda(t)\ dt\right]$. The equation (5.5) reduces to (5.2) in the case of $\mathbf{s}(z) \equiv p$.

Discrepancy magnitude The function $\mathbf{s}(\cdot)$ in (5.4) characterizes a distorted KG-model which does not fulfill the KG-requirement. In order to describe the resulting *distance* from the ideal KG-model, a *discrepancy magnitude* $\gamma = \sup_{z \in [0,\infty)} |\gamma(z)|$ is defined by

$$\gamma(z) = \int_0^z \left[\mathbf{s}(t) - p\right] dF_Z(t), \quad z \in [0, \infty). \tag{5.7}$$

The measure is related to the area between '$P(\Delta = 1 | Z = z)$' and $P(\Delta = 1)$. The particular form of integration has been chosen to eliminate the effect of high 'fluctuations' of $\mathbf{s}(\cdot)$ around p (similar to white noise). This is sensible because such a behaviour of $\mathbf{s}(\cdot)$ would again come close to independence of Z and Δ.

With the existence of f_Z, the expression $\mathbf{s}(z) \cdot f_Z(z)$ can be defined for *all* $z \in [0, \infty)$, which ensures that the function $\gamma(\cdot)$ converges to zero for $z \to \infty$.

This is because $\frac{\partial}{\partial z}\left[P(Z \leq z, \Delta = 1)\right] = f_{Z|\Delta}(z,1) \cdot P(\Delta = 1)$ where $f_{Z|\Delta}(\cdot,1)$ is a density (Woodroofe, 1975, p. 269), and hence

$$\int_0^\infty \mathbf{s}(t) \cdot f_Z(t)\ dt = p. \tag{5.8}$$

Note that a discrepancy magnitude as above *quantifies* the distance from the KG-model but does *not* usually characterize a distorted KG-model, since different $\mathbf{s}(\cdot)$ can be summarized by the same distance γ. This also applies to the maximum value of γ: Depending on p and taking into account all possible functions $\mathbf{s}(\cdot)$, the discrepancy magnitude γ ranges from 0 to $p(1-p)$. It can be shown that the maximum discrepancy $\gamma^* = p(1-p)$ is reached by structures

$$\mathbf{s}(z) = \imath_{[z^*,\infty)}(z) \quad \text{with} \quad F_Z(z^*) = 1 - p \tag{5.9}$$

and

$$\mathbf{s}(z) = \imath_{[0,z^*)}(z) \quad \text{with} \quad F_Z(z^*) = p, \tag{5.10}$$

where z^* is the location of the supremum of $|\gamma(\cdot)|$ and $\imath(\cdot)$ is the indicator function.

PROOF: It is for $z \in [0,\infty)$

$$\begin{aligned}
\gamma^* &= \max_{\mathbf{s}(\cdot)} \left\{ \sup_{z\in[0,\infty)} \left| \int_0^z [\mathbf{s}(t) - p]\ dF_Z(t) \right| \right\} \\
&= \max \left\{ \sup_{z\in[0,\infty)} \left[\ p \cdot F_Z(z) \ \middle|\ \mathbf{s}(t) = 0,\ t \in [0,z)\ \right], \right. \\
&\qquad \left. \sup_{z\in[0,\infty)} \left[\ (1-p) \cdot F_Z(z) \ \middle|\ \mathbf{s}(t) = 1,\ t \in [0,z)\ \right] \right\} \\
&\overset{(5.8)}{=} \max \left\{ p \cdot \sup \left[\ F_Z(z) \ \middle|\ z:\ \int_z^\infty \mathbf{s}(z)\ dF_Z(z) = p\ \right], \right. \\
&\qquad \left. (1-p) \cdot \sup \left[\ F_Z(z) \ \middle|\ z:\ F_Z(z) + \int_z^\infty \mathbf{s}(z)\ dF_Z(z) = p\ \right] \right\} \\
&= \max \left\{ p \cdot (1-p)\ \wedge\ \mathbf{s}(\cdot) \text{ as in (5.9)}, \right. \\
&\qquad \left. (1-p) \cdot p\ \wedge\ \mathbf{s}(\cdot) \text{ as in (5.10)} \right\}.
\end{aligned}$$

□

Finally, note that the measure γ has *not* been normalized, since eliminating the dependence on p would prejudge the potential discrepancy from the KG-model possible in a practical situation and would make a comparison between different values of p impossible.

5.3.2 Specific parametrisations

Since the class of possible discrepancy structures $\mathbf{s}(\cdot)$ is large, for practical reasons only two subclasses will be considered in the simulation study below.

EXP is the subclass taking functions of the form

$$\mathbf{s}(z) = a \cdot \exp(-c \cdot z) + d, \quad z \in [0, \infty)$$

where $c > 0$, $0 \leq d \leq 1$ and $0 \leq a + d \leq 1$. This parametrisation is reasonable because the conditional chance of observing a censored subject is likely to follow a trend over time in cases where the latter is not constant. The exponential form ensures a convergence for $z \to \infty$.

STEP is the subclass taking functions of the form

$$\mathbf{s}(z) = a \cdot \imath_{[0,z^*)}(z) + b \cdot \imath_{[z^*,\infty)}(z), \quad z \in [0, \infty)$$

where $\imath(\cdot)$ is the indicator function and z^* is such that

$$\begin{aligned} F_Z(z^*) &= p \qquad \text{and} \quad 0 \leq b \leq p \leq a \leq 1, \quad \text{or} \\ F_Z(z^*) &= 1 - p \quad \text{and} \quad 0 \leq a \leq p \leq b \leq 1. \end{aligned}$$

The choice is sensible because the maximum discrepancy magnitude γ^* is represented by $a = 0$ and $b = 1$ for $F_Z(z^*) = 1 - p$, and by $a = 1$ and $b = 0$ for $F_Z(z^*) = p$, which respectively corresponds to (5.9) and (5.10).

In order to finally enforce an *injective* mapping from $\mathbf{s}(\cdot)$ to the discrepancy magnitude γ, the subclasses EXP and STEP have to be further subdivided

according to decreasing and increasing slopes. For structures of kind EXP also one of the parameters needs to be kept fixed. These restrictions entail some limitations to the simulation study below, but are necessary to produce results.

5.3.3 Other formalization approaches

1. Peña and Rohatgi (1989) represent the survival function of the censored survivals S_Y conditioned on the realization b of some random variable β satisfying a condition for model identification. It is

$$S_Y(z|b) = P(Y > z|\beta = b) = [S_X(z)]^b \quad \text{for } z \in [0, \infty).$$

For $z \in [0, \infty)$ with $f_Z(z) > 0$ the relationship to the $\mathbf{s}(\cdot)$-approach is given by

$$\mathbf{s}(z) = \frac{\mathrm{E}_\beta\{[S_X(z)]^\beta\} \cdot f_X(z)}{f_Z(z)}.$$

The above follows immediately from the first part of (5.6). For a simulation study the authors use a special case with

$$S_X(z) = \exp(-z) \qquad \text{and} \qquad S_Z(z) = \frac{\theta \cdot S_X(z)}{1 - S_X(z) \cdot (1 - \theta)},$$

where $\theta \in (0, 1]$. The latter corresponds to the discrepancy structure $\mathbf{s}(z) = (\theta - 1) \cdot \exp(-z) + 1$ which is increasing in z and belongs to EXP taking $a = (\theta - 1)$, $c = 1$, and $d = 1$.

2. Beirlant *et al.* (1992) generalize the KG-model by defining

$$S_Y(z) = [S_X(z)]^\theta \cdot \mathbf{L}(S_X(z)) \quad \text{for } z \in [0, \infty),$$

where $\mathbf{L}$ is some slowly varying function at the origin with $\mathbf{L}(1) = 1$ and θ some constant in $[0, \infty)$. If $\lim_{z\to\infty} \mathbf{s}(z) = s_\infty > 0$ exists, the relationship to the $\mathbf{s}(\cdot)$-approach is given by $\theta = (1 - s_\infty)/s_\infty$ and the slowly varying function $\mathbf{L}^*$ with

$$\mathbf{L}^*(u) = \exp\left[\frac{s_\infty - 1}{s_\infty} \cdot \ln u - \int_0^{S_X^{-1}(u)} \frac{1 - \mathbf{s}(z)}{\mathbf{s}(z)} \cdot \frac{f_X(z)}{S_X(z)}\, dz\right], \quad (5.11)$$

where S_X^{-1} is the inverse of S_X (the proof is given in Appendix A.1).

3. Ebrahimi and Kirmani (1996) have found an alternative characterization of the condition (5.2) which demands the constancy of

$$I(t) = \int_t^\infty \frac{f_X(x)}{S_X(t)} \cdot \log \left[\frac{f_X(x) \cdot S_Y(t)}{f_Y(x) \cdot S_X(t)} \right] \, dx.$$

Hence, a non-constant behaviour of $I(\cdot)$ could be used to formalize a distorted KG-model.

4. Instead of conditioning on the random variable Z, Herbst (1992b) conditions on Δ and compares the two resulting conditional functions $P(Z < z|\ \Delta = 0)$ and $P(Z < z|\ \Delta = 1)$. He then formulates a goodness-of-fit test of Kolmogorov-Smirnov type for the Koziol-Green model. Moreover, graphical checks of the KG-requirement have been proposed by Ebrahimi (1985) and Ebrahimi and Habibullah (1992) comparing the two sub-survival distributions $P(Z > z, \Delta = 0)$ and $P(Z > z, \Delta = 1)$, and Csörgő (1988) and Beirlant *et al.* (1992) studying $P(\Delta = 1|\ Z \leq z)$ and $P(\Delta = 1|\ Z > z)$, respectively.

Since the interpretation and conditions of the $\mathbf{s}(\cdot)$-approach are fairly simple and the formulation of the discrepancy magnitude γ makes it easy to define the *amount of distortion* from the KG-model (see later), this formulation is preferred for the simulation study below.

The following section will now specify the performance assessment needed for the ACL-estimator, both under the ideal situation of the KG-model and in particular under distortion.

5.4 ACL-performance assessment

With the above formulation of a distorted KG-model and additional limitations on the discrepancy structure $\mathbf{s}(\cdot)$, one can easily characterize and

quantify situations of data contamination and type ② model deviation. The respective *amount of distortion* results from corresponding differences between discrepancy magnitudes of REALITY, DATA and ASSUMPTIONS (see § 4.2.2). Distortion affects parts of the model which relate to S_X (the aspect of interest in REALITY), since the latter is different for the distorted KG-model according to (5.5). Hence, ACL-performance is potentially distinct under corresponding data contamination and model deviation.

Influence and preference graphs (§ 4.3) are now defined which try to reveal this difference in ACL-performance under *increasing* distortion. They are based on suitable performance statistics (§ 4.2.3) which refer to the true S_X of REALITY. The overall methodology will be used for the simulation study in the subsequent section.

5.4.1 Performance statistics

Performance of the ACL-estimator S_X^{ACL}, seen as a finite sample statistic, shall be described in terms of location-closeness (§ 2.4.2). As a curve estimator, S_X^{ACL} will be assessed globally as well as more locally. Here, the performance statistics are chosen to be of Cramér-von Mises type, i.e. good performance will correspond to small realizations of the statistic.

For an entirely *global* assessment define

$$\begin{aligned} \pi_o &= \pi_o\,(R = \alpha_1,\ D = \alpha_2\ |A = \alpha_3,\ \Delta) \\ &= \int_0^\infty [S_X(t) - S_X^{ACL}(t)]^2\ d\hat{F}_Z(t) = \frac{1}{n}\sum_{i=1}^{n}\ [S_X(Z_i) - S_X^{ACL}(Z_i)]^2, \end{aligned}$$

where S_X is the true survival function to be estimated in REALITY, $S_X^{ACL}(t)$ is the ACL-estimate at time t determined from a DATA-generated sample, and $\hat{F}_Z$ the empirical distribution function based on the random sample of Z in DATA. The values α_1, α_2, and α_3 will be specified with the definition of influence and preference graphs (see below). Note that the statistic π_o refers to the aspect of interest in REALITY, which makes it possible to distinguish

and compare influence on ACL-performance under data contamination and type ② model deviation (§ 4.2.1).

A more *localized* assessment can be carried out by assigning twice the weight to observations within certain quantile-ranges of Z in DATA and no weights to the remaining ones. It is

$$\begin{aligned}\pi_l &= 2 \cdot \int_{\mathcal{A}_l} [S_X(t) - S_X^{ACL}(t)]^2 \, d\hat{F}_Z(t) \\ &= \frac{2}{\sharp\{Z_i : Z_i \in \mathcal{A}_l\}} \sum_{Z_i \in \mathcal{A}_l} [S_X(Z_i) - S_X^{ACL}(Z_i)]^2,\end{aligned}$$

where $\mathcal{A}_l = [0, z_{[0.5]}]$ with $z_{[\alpha]}$ as the α-quantile of the distribution of Z in DATA, and where the sign $\sharp$ means 'number of elements'. In the same way the performance statistics π_u and π_m are defined with $\mathcal{A}_u = [z_{[0.5]}, \infty)$ and $\mathcal{A}_m = [z_{[0.25]}, z_{[0.75]}]$.

Although a further reduction of the interval lengths would effect stricter localisations, this will not be done here. For a completely localized investigation under model deviation see instead the simulation study in Peña and Rohatgi (1989).

5.4.2 Influence and preference graphs

Influence graphs which study the *absolute* performance of the ACL-estimator in terms potential change under increasing distortion are of the form

$$\begin{aligned}g_i^D(\gamma) = \mathrm{E}_\gamma\Big[\pi(R = 0, D = \gamma \mid A = 0, S_X^{ACL})\Big] \\ - \mathrm{E}_0\Big[\pi(R = 0, D = 0 \mid A = 0, S_X^{ACL})\Big]\end{aligned}$$

for *data contamination*, and

$$\begin{aligned}g_i^M(\gamma) = \mathrm{E}_\gamma\Big[\pi(R = \gamma, D = \gamma \mid A = 0, S_X^{ACL})\Big] \\ - \mathrm{E}_0\Big[\pi(R = 0, D = 0 \mid A = 0, S_X^{ACL})\Big]\end{aligned}$$

for type ② *model deviation* (compare with (4.3) and (4.4)). While π stands for one of the previously defined performance statistics π_o, π_l, π_u, or π_m, the value γ represents corresponding *discrepancy magnitudes* of REALITY and/or DATA on the one hand, and the resulting *amount of distortion*, on the other. Note that all expected values refer to a DATA-model.

Increasing influence graphs of the above kind indicate that the performance of the ACL-estimator gets worse under increasing distortion. By definition they originate with the value zero at $\gamma = 0$.

Preference graphs which study the *absolute* performance of the ACL-estimator in comparison with the Kaplan-Meier (KM) estimator S_X^{KM} under increasing distortion are of the form

$$g_p^D(\gamma) = \mathrm{E}_\gamma\Big[\pi(R=0, D=\gamma \mid A=0, S_X^{ACL}) - \pi(R=0, D=\gamma \mid A=+, S_X^{KM})\Big]$$

for *data contamination*, and

$$g_p^M(\gamma) = \mathrm{E}_\gamma\Big[\pi(R=\gamma, D=\gamma \mid A=0, S_X^{ACL}) - \pi(R=\gamma, D=\gamma | A=+, S_X^{KM})\Big]$$

for type ② *model deviation*. This corresponds to preference graphs for two inference procedures based on alternative assumptions (see (4.7) and (4.8)). It is suitable since both estimators are known on their own and, as maximum-likelihood estimators, imply the model assumptions made. The latter are the Koziol-Green model ($A = 0$) for the ACL-estimator and the right random censorship model ($A = +$) for the KM-estimator. Again, π may be one of the statistics π_o, π_l, π_u, or π_m, and the expected values refer to the 'current' DATA-model.

Overall, a preference graph as above at distortion γ indicates advantages for the ACL-estimator when the corresponding (preference) value is below zero.

The foregoing influence and preference graphs provide the methodological tools for the following simulation study which exclusively will study *absolute* performance of the ACL-estimator. In other words, relative (changes

of) performance, assessable via influence and preference graphs based on *quotients*, will be neglected.

5.5 Simulation study

Using the influence and preference graphs introduced in the previous section, the finite-sample behaviour of the ACL-estimator is *compared* under particular situations of model deviation and data contamination (distortion of types ① and ②). The latter are characterized as corresponding model disagreement (of REALITY, DATA, and ASSUMPTIONS) between the KG-model and some distorted KG-model (§ 5.3). It is of interest whether the influence on the performance of the ACL-estimator is indeed different, as has been claimed in § 3.4.2. In *practical* terms such a comparison makes sense when both types of distortion are at the same time principally justifiable. This is the case in a competing risks framework (see the beginning of § 5.3), so that in the first place, the latter should be kept in mind.

The method of *simulation* is chosen, since direct interest is in the *finite* sample case where theoretical evaluations of influence and preference graphs appear to be difficult: in contrast with the goodness-of-fit idea which compares two distributions *both* from the KG-model (Ghorai, 1991a), a performance statistic here compares 'distributions' where at least one of them does *not* belong to the KG-model. A discrepancy magnitude γ, moreover, summarizes alternative distributions corresponding to (Δ, Z), which cannot be represented by traditional model parametrisations.

The simulation study is carried out on a Sun-compatible work station using S-PLUS (version 3.3, release 1).

5.5.1 Set-up

Problems addressed

- Is there a difference between the consequences of model deviation and data contamination – in particular under *increasing* distortion?
- Does the kind of discrepancy structure $\mathbf{s}(\cdot)$ matter?
- What is the influence of the sample size n and the probability p of an observation being uncensored?
- When does the ACL-estimator outperform the KM-estimator and vice versa?
- What is the behaviour of the ACL-estimator in certain regions such as the tails?

Program organization

1. Specification of a discrepancy structure $\mathbf{s}(\cdot)$ representing a γ-distorted KG-model.

   After having chosen p and F_Z as the Weibull (shape=2, scale=1)-distribution, decide on the subclass STEP or EXP (§ 5.3.2) and derive the corresponding parameters for $\mathbf{s}(\cdot)$ using the information in (5.7) and (5.8).

   STEP $\mathbf{s}(\cdot)$ is determined by the parameter settings

   $$a = \frac{\gamma}{p} + p \quad \text{and} \quad b = \frac{p(1-a)}{1-p} \quad \text{for} \quad F_Z(z^*) = p$$

   and

   $$a = p - \frac{\gamma}{1-p} \quad \text{and} \quad b = 1 - \frac{a(1-p)}{p} \quad \text{for} \quad F_Z(z^*) = 1-p.$$

EXP Given c and the sign of a, the parameters a and d are specified by approximating a common (single) solution for (5.7) and (5.8) via numerical integration and optimization. For convenience, the value of c is taken which corresponds to the maximum discrepancy explained by EXP given p and the sign of a (see Table 5.1). A few more details are given in Appendix B.

2. Simulation of B samples $(z_1, \delta_1)', \ldots, (z_n, \delta_n)'$ according to DATA which is a distorted KG-model as specified above. For each sample and for all $i = 1, \ldots, n$ this implies: simulation of z_i according to F_Z, evaluation of $\mathbf{s}(z_i)$, and finally generation of δ_i according to a $\text{Bin}[1, \mathbf{s}(z_i)]$-distribution.

3. Evaluation of individual influence and preference values $g_i(\gamma)$ and $g_p(\gamma)$.

   - Choose one of the performance statistics π_o, π_l, π_u, or π_m and the type of distortion (data contamination or type ② model deviation).
   - For each sample evaluate the *true* and *estimated* survival probabilities at $z_{1:n}, \ldots, z_{n-1:n}$. While the former are based on the KG-model for data contamination, and on the γ-distorted KG-model for model deviation (further details in Appendix B), the respective ACL- and KM-estimates are derived from the DATA-generated samples which, in any case, reflect the γ-distorted KG-model.

     In addition, note that the maximum survival times $z_{n:n}$ are omitted to secure conformity between the ACL- and KM-estimator, since $S_X^{ACL}(z_{n:n}) = 0$, but $S_X^{KM}(z_{n:n}) = 0$ for uncensored $z_{n:n}$ and $S_X^{KM}(z_{n:n}) = S_X^{KM}(z_{n-1:n})$, otherwise.
   - With $R = 0$ for data contamination and $R = \gamma$ for model deviation, determine for each simulated sample

     $$\pi(R, D = \gamma \mid A = 0, S_X^{ACL}) \tag{5.12}$$

     and

     $$\pi(R, D = \gamma \mid A = 0, S_X^{ACL}) - \pi(R, D = \gamma \mid A = +, S_X^{KM}). \tag{5.13}$$

Table 5.1: *Maximum discrepancy γ in STEP and EXP for different values of p (d=decreasing and i=increasing, i.e., $a \geq 0$ and $a \leq 0$, respectively).*

p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
STEP	0.09	0.16	0.21	0.24	**0.25**	0.24	0.21	0.16	0.09
EXP (d)	0.05	0.07	0.08	0.08	0.07	0.07	0.05	0.04	0.02
EXP (i)	0.02	0.04	0.05	0.07	0.07	0.08	0.08	0.07	0.05

- Take the average of (5.12) over the B simulated samples and subtract from it a corresponding average for $\gamma = 0$. The result serves as contribution for the influence graph.
- Take the average of (5.13) over the B simulated samples as contribution for the preference graph.

4. Evaluation of corresponding one standard-deviation limits for each of the above contributions by adding and subtracting the empirical standard-deviation of (5.12) and (5.13), respectively, over the B simulated samples.

5. For the *overall* influence and preference graphs the previous steps are repeated with different values of γ.

Design

While restricting F_Z to a Weibull (shape=2, scale=1)-distribution, the parameter values for the study are chosen as $p \in \{0.1, \ldots, 0.9\}$, $n \in \{20, 60, 150\}$ and $B = 200$. The discrepancy magnitudes range from 0 to their possible limits (see Table 5.1) with a step size of 0.01. Decreasing and increasing 'trends', respectively, are considered for $\mathbf{s}(\cdot)$ in each of the subclasses STEP and EXP.

5.5.2 Results

Consider the Figures 5.1 and 5.2 showing an arbitrary ACL- and KM-estimate (dashed-thick and dotted line, respectively) under high in fact the maximum distortion and compare them with the corresponding true survival function (solid line). The *different* performance of the ACL-estimator under data contamination and model deviation (type ②), explained by decreasing $\mathbf{s}(\cdot)$ and the function $\gamma(\cdot)$ in Figures 5.3 and 5.4, is very obvious. While the ACL-estimator performs well under data contamination (Figure 5.1), it seems to be clearly overestimating for small times z and underestimating for large z under model deviation (Figure 5.2).

The behaviour under corresponding distortion with *increasing* $\mathbf{s}(\cdot)$ again shows good performance under data contamination (Figure 5.5) and bad performance under model deviation, but this time by underestimating for small z and overestimating for large z (Figure 5.6). Pena and Rohatgi (1989) also came to this result by carrying out a localized simulation study for a distortion model based on an increasing discrepancy structure from the class EXP under model deviation (see 5.3.3).

The following study of *influence and preference graphs* will confirm this contradictory performance of the ACL-estimator.

Figures 5.7 and 5.8 exemplify corresponding influence graphs (solid line), here with *decreasing* discrepancy structure $\mathbf{s}(\cdot)$ of kind STEP, $p = 0.5$ and $n = 150$, taking the performance statistic π_o. While the 'average closeness' of the ACL-estimator under data contamination remains more or less the same when compared to the undistorted situation, the performance under model deviation clearly declines with increasing amount of distortion. The two dashed lines represent the one standard-deviation bounds.

Comparing the ACL-estimator with the KM-estimator using a preference graph, the effects of data contamination and model deviation reveal a *trade-off* (see Figures 5.9 and 5.10). The more the ACL-estimator outperforms the KM-estimator under data contamination, the worse it behaves under

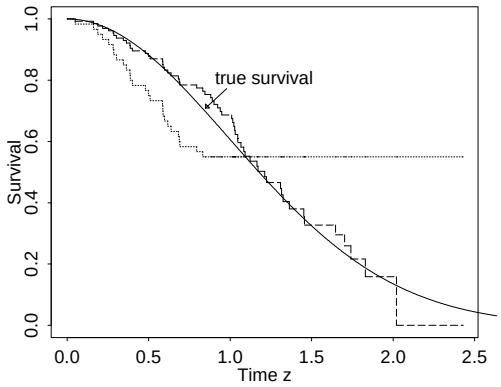

Figure 5.1: *Estimates under data contamination:* $\mathrm{s}(\cdot)$ *decreas. and* $\gamma = 0.25$.

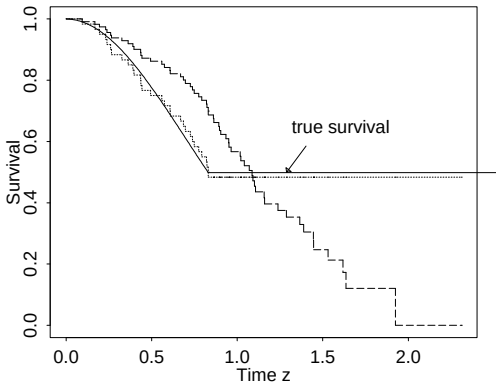

Figure 5.2: *Estimates under model deviation:* $\mathrm{s}(\cdot)$ *decreasing and* $\gamma = 0.25$.

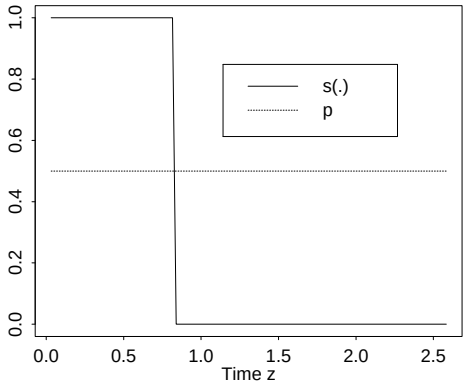

Figure 5.3: *Decreasing discrepancy structure* $\mathrm{s}(\cdot)$ *corresponding to* $\gamma = 0.25$.

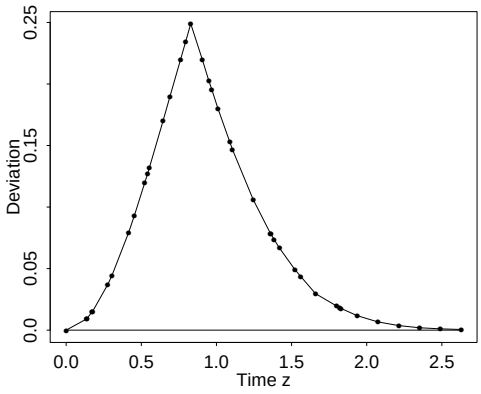

Figure 5.4: *Function* $\gamma(\cdot)$ *resulting from the previous Figure.*

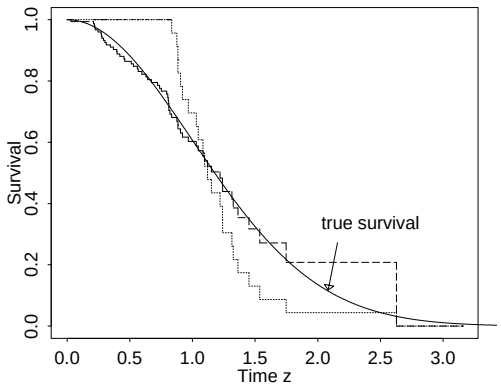

Figure 5.5: *Estimates under data contamination:* $\mathrm{s}(\cdot)$ *increas. and* $\gamma = 0.25$.

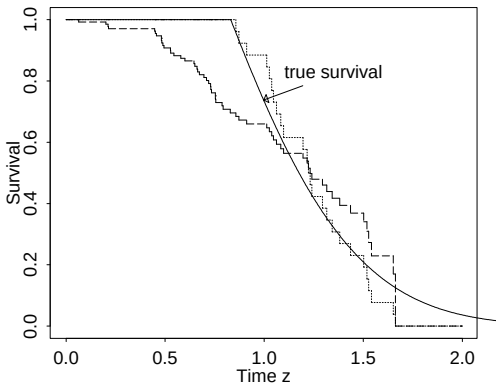

Figure 5.6: *Estimates under model deviation:* $\mathrm{s}(\cdot)$ *increasing and* $\gamma = 0.25$.

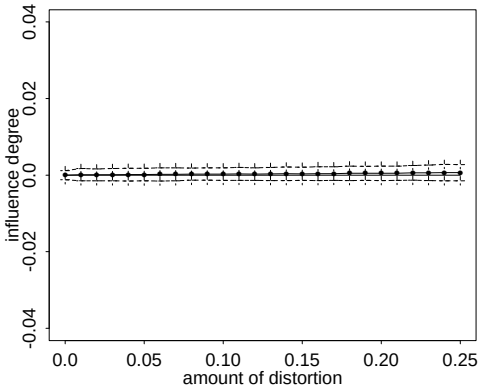

Figure 5.7: $g_i^D(\gamma)$ *for data contamination: STEP (d),* $p = 0.5$*,* $n = 150$*.*

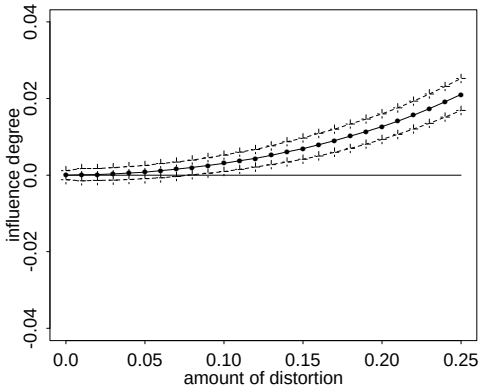

Figure 5.8: $g_i^M(\gamma)$ *for model deviation: STEP (d),* $p = 0.5$*,* $n = 150$*.*

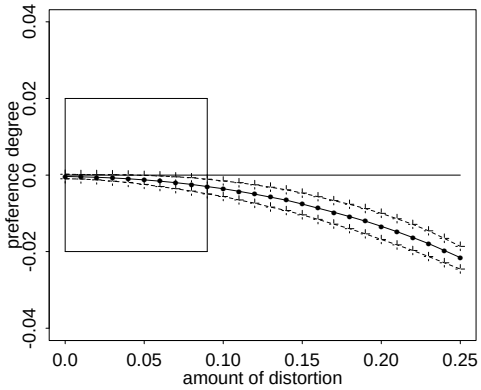

Figure 5.9: $g_p^D(\gamma)$ *for data contamination: STEP (d),* $p = 0.5$*,* $n = 150$*.*

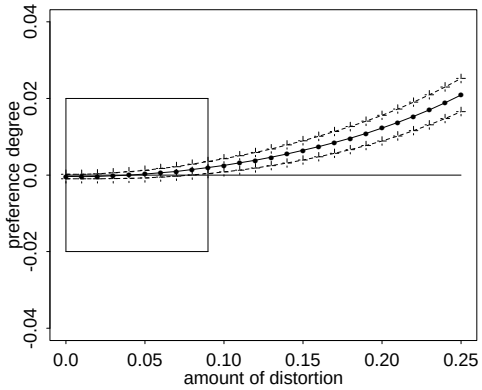

Figure 5.10: $g_p^M(\gamma)$ *for model deviation: STEP (d),* $p = 0.5$*,* $n = 150$*.*

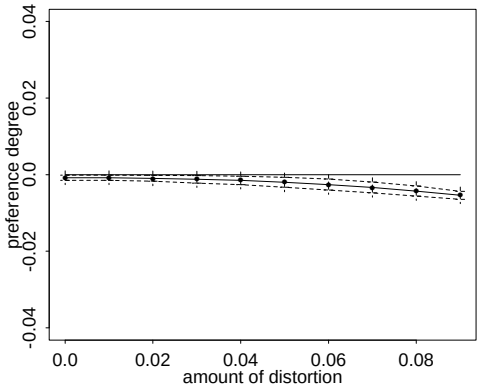

Figure 5.11: $g_p^D(\gamma)$ *for data contamination: STEP (d),* $p = 0.1$*,* $n = 150$*.*

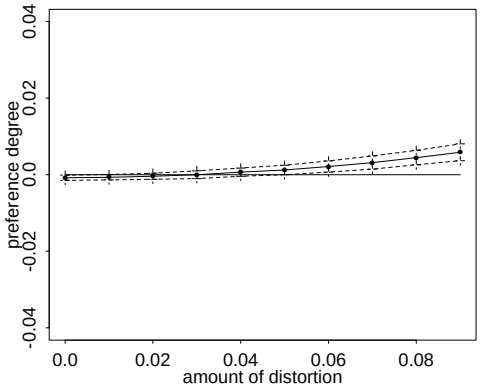

Figure 5.12: $g_p^M(\gamma)$ *for model deviation: STEP (d),* $p = 0.1$*,* $n = 150$*.*

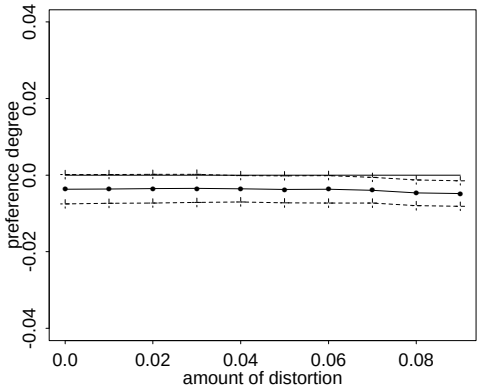

Figure 5.13: $g_p^D(\gamma)$ *for data contamination: STEP (d),* $p = 0.1$*,* $n = 20$*.*

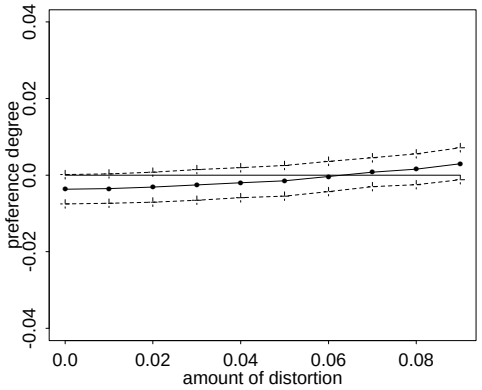

Figure 5.14: $g_p^M(\gamma)$ *for model deviation: STEP (d),* $p = 0.1$*,* $n = 20$*.*

corresponding model deviation.

Over the range of different probabilities p, the consequences of data contamination and model deviation appear to be less dramatic for larger p. No obvious change in any of the influence and preference graphs of the type considered so far is noticeable any more for a value of $p = 0.9$. Still, the consequences are slightly stronger for smaller p. However, because of smaller maximum amounts of distortion (corresponding to the maximum discrepancy magnitudes in Table 5.1), high absolute differences as above for $p = 0.5$ do not exist for smaller p (compare Figures 5.11 and 5.12 with the corresponding partitions in Figures 5.9 and 5.10).

An increased effect of distortion accompanied by smaller values of p is also apparent for decreasing distortion structures of kind EXP. Even though the corresponding changes are slightly larger, the high absolute influence and preference values for p around 0.5 are again not possible because of smaller maximum amounts of distortion within the class EXP (Table 5.1).

Until now, only *decreasing* $\mathbf{s}(\cdot)$ have been considered. The consequences indicated by influence and preference graphs under corresponding *increasing* distortion structures (STEP and EXP) appear to be nearly equivalent for $n = 150$. Again, the ACL-estimator (in comparison with the KM-estimator)

Table 5.2: *The effect of distortion indicated by the performance statistics π_l, π_u and π_m in comparison to π_o (taking into account influence as well as preference graphs).*

$\mathbf{s}(\cdot)$		small p	intermediate p	large p
decreasing	π_l	less	less	less
	π_u	more	more	more
	π_m	less	more	more
increasing	π_l	less	more	more
	π_u	more	less	less
	π_m	less	more	less

shows advantages under increasing data contamination but performs poorly under increasing model deviation. As before, this effect becomes more important as p decreases, but now to an even greater extent.

Performance statistics restricted to lower, upper or intermediate time-intervals can be used to examine a more *local* behaviour of the ACL-estimator. Table 5.2 summarizes a comparison with influence and preference graphs based on the non-restricted performance statistic π_o.

Finally, a reduction of the sample size ($n = 60, 20$) leads to an increase of the variability associated with the individual influence and preference values over the B samples. As a consequence, any effect of distortion, if present, becomes more indistinct. Neglecting the high variability, performance in general also changes to a smaller extent (influence graphs). Moreover, the disadvantages under model deviation seem to be less serious for small sample sizes (preference graphs, see Figures 5.13 and 5.14). This is because the ACL-estimator performs better than the KM-estimator under the *un*distorted KG-model. The overall phenomenon almost disappears when considering preference graphs based on π_l and is very clear when instead using π_u.

Despite the numerous details pointed out in the previous paragraphs, the

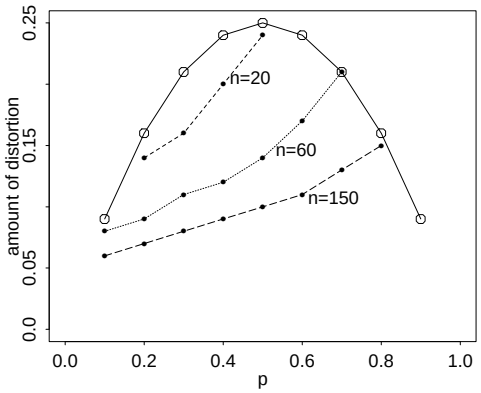

Figure 5.15: *Minimum distortion γ with* $\mathbf{s}(\cdot)$ *decreasing, above which* S_X^{ACL} *shows disadvantages under model deviation.*

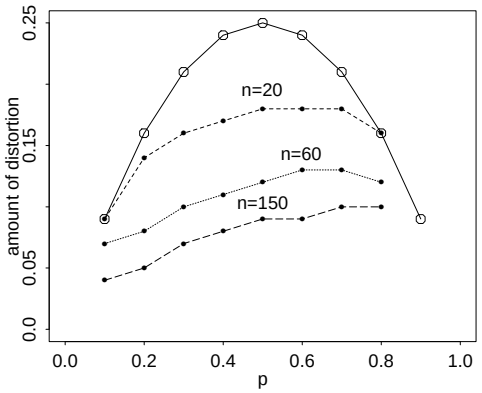

Figure 5.16: *Minimum distortion γ with* $\mathbf{s}(\cdot)$ *increasing, above which* S_X^{ACL} *shows disadvantages under model deviation.*

simulation results allow the following conclusions in comparison with the KM-estimator (considering preference graphs for STEP based on π_o and taking into account a rounding error of 0.001).

Data contamination: The ACL-estimator shows *advantages*, i.e. the upper one standard-deviation bound is below zero for all $\gamma > 0.01$:

- – for $0.1 \leq p \leq 0.8$ with $n = 150$ and $\mathbf{s}(\cdot)$ decreasing.
  – for $0.1 \leq p \leq 0.9$ with $n = 150$ and $\mathbf{s}(\cdot)$ increasing.
- – for $0.1 \leq p \leq 0.5$ with $n = 60$ and $\mathbf{s}(\cdot)$ decreasing.
  – for $0.1 \leq p \leq 0.4$ with $n = 60$ and $\mathbf{s}(\cdot)$ increasing.
- – for $0.1 \leq p \leq 0.4$ with $n = 20$ and $\mathbf{s}(\cdot)$ decreasing.
  – for $0.1 \leq p \leq 0.3$ with $n = 20$ and $\mathbf{s}(\cdot)$ increasing.

In all other cases, 'just' the preference value itself is below zero.

Model deviation: The ACL-estimator shows *disadvantages* from certain amounts of distortion γ onwards, as indicated in the Figures 5.15 and 5.16. In these cases the corresponding lower one standard-deviation

bound is above zero. In both Figures, the solid line represents the maximum possible amount of distortion (according to Table 5.1). The differences between decreasing and increasing $\mathbf{s}(\cdot)$ are mainly due to broader one standard-deviation bounds for the former, given large p.

5.5.3 Discussion

The fact that the performances of the ACL-estimator under data contamination and model deviation are *different* is of great interest. Broadly speaking, S_X^{ACL} seems to be resistant against data contamination of the $\mathbf{s}(\cdot)$-kind, but shows problems under corresponding model deviation. As a consequence, it tends to outperform the KM-estimator in the first case and be inferior in the second. This is because the ACL-estimator relying on the KG-model assumptions *uses less information from the data*, that is, ignoring the information given in the distribution of the censored observations over the time and replacing this missing bit by the extra KG-requirement.

Fortunately, the above phenomenon is only half of the truth. The study of influence and preference graphs could reveal *more* details about the effect of distortion and its potential dependence on aspects such as p and n.

The consequences of distortion become less important with increasing p, i.e. with fewer censored observations expected in the data. Given a value of p greater than about 0.8, an effect on the overall ACL-performance is in most cases no longer convincing – also because distortion is only possible with smaller amounts. Moreover, the influence becomes less distinct with smaller sample sizes. Thus, a concern about the global non-robustness of the ACL-estimator under model deviation should not be too overemphasized, since most practical studies (hopefully) deal with larger percentages of uncensored data and (unfortunately) are based on smaller sample sizes.

The consideration based on the restricted performance statistics π_l, π_u and π_m revealed a further interesting aspect. For *decreasing* discrepancy structures, influence is mainly due to changes in the upper (right) tail. Hence,

estimators for the lower quantiles based on the ACL-estimator are expected to be more robust against misspecification of the KG-assumptions than corresponding estimators for the upper quantiles. However, the situation turns out to be different for *increasing* $\mathbf{s}(\cdot)$. Here, the main location of influence is strongly dependent on the value p (see Table 5.2).

Apart from the latter aspect, any kind of discrepancy structure $\mathbf{s}(\cdot)$, i.e. whether from STEP or EXP, or whether decreasing or increasing, seem to cause roughly equivalent influence and preference graphs when taking into account the *whole* time scale.

Finally, the simulation study has been carried out with observed survivals Z from a Weibull(2, 1)-distribution. Even though this is a limitation of the study, there are reasons to believe that other distributions would produce similar results, since the discrepancy magnitude γ, and hence the amount of distortion, is grounded on information from F_Z, see (5.7). A 'pilot-run' with a Weibull (1, 2)-distribution confirms this position. Still, an extension of the simulation study which includes other (types of) distributions for Z might be of interest.

5.6 Final remarks

Apart from the actual results of the simulation study, the present chapter received importance through its exemplifying character for the theoretical ideas and concepts in the preceding chapters.

The particular kind of distortion from the KG-model affects aspects *within* data units and hence should involve REALITY as a reference point (note, however, that exceptions have been pointed out in the censored survival case). As a consequence it may refer to either *data contamination* (type ① distortion) or *model deviation* as type ② distortion. In other words, model disagreement can be considered between DATA and REALITY, *or* between ASSUMPTIONS and REALITY which in both cases is described as

a conflict between the KG-model and some *distorted* KG-model. Note that this distinction has as yet not found attention in the statistical literature.

The distorted KG-model as such has been modelled in a novel way by the *discrepancy structure* $\mathbf{s}(\cdot)$ from which then the *discrepancy magnitude* γ could be derived. The latter allowed to formulate the *amount of distortion* by corresponding comparison of REALITY, DATA, and ASSUMPTIONS. A characterization of a distorted model *solely* based on a discrepancy magnitude has *not* been chosen – due to the complexity of the overall distortion neighbourhood (expressing dependence between Z and Δ) and since it was of interest to what *extremes* distortion affecting the KG-requirement can go.

The maximum limit of possible discrepancy magnitudes and hence amounts of distortion depends on p and can be associated with the specific structures $\mathbf{s}(\cdot)$ in (5.9) and (5.10). The *overall* maximum distortion of amount $\gamma^* = 0.25$ is only possible for $p = 0.5$ (see Table 5.1). Here, $\mathbf{s}(\cdot)$ is given the maximum 'freedom' to vary around p without violating (5.8). It is further interesting to note that under high model deviation with *decreasing* $\mathbf{s}(\cdot)$, the true survival function S_X in REALITY can become fairly improper (see Figure 5.2). In this case the probability of finally failing, e.g. becoming pregnant, is less than one.

Performance assessment of the ACL-estimator has been carried out with reference to REALITY, which allowed to reveal the difference in performance under data contamination and model deviation. The Cramér-von Mises type performance statistics have been chosen for convenience, but also other statistics addressing the closeness of the estimator would be appropriate such as the ones of Kolmogorov-Smirnov type. The one standard-deviation bounds associated with each influence and preference graph reflect the variability of their evaluated values, but due to the underlying integral and difference structures, do not generally represent the variability of the estimators involved. Nevertheless, very obvious changes in width could be due to the latter.

In terms of *preference graphs* the simulation study could compare performance of the ACL-estimator with that of the KM-estimator, and hence two estimators based on alternative assumptions (the KG-model and the right random censorship model, respectively). The following chapter will now present a simulation example where preference graphs are being used to compare the consequences under two different *assumptions* both associated with the *same* inference method.

Chapter 6

Second detailed example: A model for longitudinal data

6.1 Introduction

The previous chapter dealt with distortion from the restricted Koziol-Green model, where the distorted model was embedded in the more general model of right random censorship. In terms of inference procedures the problem led to a comparison of the performances of two model-specific maximum-likelihood estimators. The example discussed in this chapter will be of a somewhat different nature. Here, the idea of distortion addresses issues of a more complex model-building process. Are measurement errors to be included into the model assumptions or not? What different consequences might this have on the real-world description of interest given a particular inference method for the unknown model parameters?

In more detail, we will consider a general linear model with correlated errors for longitudinal data, where the (sub)model for the covariance structure *within* data units is subject to distortion. That is, normally distributed measurement errors are comprised in some *distorted* covariance model and are ignored in the corresponding ideal reference model. *Data contamination*

is then present when the above measurement errors are indeed real, and *model deviation* occurs when the assumptions do not account for the extra random variation in the data, even though it is a reflection of some other (relevant) phenomenon in the real-world.

Section 6.2 introduces the model and the model framework, and devotes attention to the theoretical function known as the *variogram* (Diggle, 1990). The latter is often used to describe the underlying covariance structure in longitudinal data. In addition, the inference method/procedures are explained, this time with mention of the (empirical) *sample variogram*. Section 6.3 interprets and formalizes distortion from the longitudinal data model. This is followed by the methodology for suitable inferential performance assessment in section 6.4 (influence and preference graphs). As in chapter 5, a simulation study illustrates some finite-sample consequences under corresponding data contamination and model deviation for the real-world descriptions of interest (§ 6.5). Some final remarks close the chapter (§ 6.6).

6.2 The model and inference procedures

6.2.1 The model and model framework

A *real-world situation* in the context of longitudinal data analysis is most commonly a phenomenon which changes over time. This could for instance be the varying protein content of cows milk under a certain diet (see example 5 in chapter 3), or it might be the progressing health status of a patient under a particular medical treatment. The *aspect of interest* is often primarily the average development of the phenomenon over time, usually allowing for dependence on some covariate information. According to Diggle *et al.* (1995) this is called the *mean response profile*.

A second aspect of interest becomes obvious when considering the typical structure of longitudinal data: The outcome of every *sampling unit* (e.g. a

cow) consists of a *sequence* of univariate observations with additional covariate information (e.g. measurements of the protein content under a certain diet). In other words, each of the m *data units* is associated with a particular time series of length $n_i \geq 1$, $i = 1, \ldots, m$. This is to be seen in contrast to the 'one observation per data unit' situation with non-longitudinal, i.e. cross-sectional data (Diggle *et al.*, 1995). Chapter 5 gave an example of the latter. The longitudinal character of a study implies a further *aspect of interest* which is the correlation structure of measurements *within* data units.

Depending on the scientific problem and the dimension of the data, either *aspect of interest* may be more important than the other (though the correlation structure is generally pertinent for inference purposes, see later). For the study and comparison of treatment effects the number of data units m is usually much greater than the number of observations per unit n, and the *mean response profile* is of prime interest. Other examples exist where more emphasis is laid upon the estimation and prediction of *single* response developments such as individual AIDS-disease progressions, and where n is comparatively large. In these cases the correlation structure *within* data units acquires priority (Diggle *et al.*, 1995, p. 20).

Model definition Taking into account available covariate information a *data-generating process* for the above problem is represented by a general linear model with correlated random errors as follows (Diggle *et al.*, 1995):

The N potential observations are summarized into m independent random vectors $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$ corresponding to the m data units and continuously measured at time points $\boldsymbol{t}_i = (t_{i1}, \ldots, t_{in_i})'$ with $\sum_{i=1}^{m} n_i = N$. Then

$$\boldsymbol{Y}_i = X_i \boldsymbol{\beta} + \boldsymbol{\epsilon}_i, \tag{6.1}$$

where X_i is a $(n_i \times p)$–matrix of explanatory variables, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a vector of coefficients, and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{in_i})'$ is a vector of random errors.

For our purposes the model is further specified by all $\boldsymbol{Y}_i$'s measured at the same set of equally spaced time points $\boldsymbol{t}_i = (1, \ldots, n)'$ with $n_i = n$, and by taking X_i either as

$$\begin{pmatrix} 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & n \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & n \end{pmatrix}, \quad \text{or} \quad \begin{pmatrix} 0 & 0 & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & n \end{pmatrix}. \tag{6.2}$$

Thus, the model incorporates three 'treatment' effects and a linear time trend, and is based on a balanced data structure. A simpler formulation for this *mean model* is the following:

$$\mathrm{E}(\text{response in g-th treatment group at time } t) = b_{0g} + b_1 \cdot t, \tag{6.3}$$

where $g = 1, 2, 3$ and $t = 1, \ldots, n$. Referring to Diggle (1988) and Diggle *et al.* (1995, p. 87) the random errors will moreover separately be expressed as

$$\epsilon_{ij} = U_i + W_i(j) + Z_{ij}, \tag{6.4}$$

with $i = 1, \ldots, m$ and $j = 1, \ldots, n$ and where

- U_i are i.i.d. like $N(0, \nu^2)$ representing random effects.
- $\{W_i(t) : t \in \mathbb{R}\}$ are independent stationary Gaussian processes representing the serial correlation within each data unit. That is, $W_i(t)$ is multivariate normal distributed with $\mathrm{E}\left[W_i(t)\right] = 0$ and

  $$\mathrm{Cov}\big[W_i(t), W_i(s)\big] = \sigma^2 \cdot \rho_W(|t - s|).$$

  The autocorrelation function $\rho_W(\cdot)$ shall be of the form

  $$\rho_W(u) = \exp(-\phi u), \quad u \geq 0. \tag{6.5}$$

- Z_{ij} are i.i.d. like $N(0, \tau^2)$ representing measurement errors.

The model, as a whole, is of *parametric* nature and includes the mean parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_4)' = (b_{01}, b_{02}, b_{03}, b_1)'$ in the structural part and

the variance parameters ν, σ, ϕ, and τ in the distributional part. The second group of parameters will be transformed in the next section in order to formalize distortion. Note that, except for the independence between responses $\boldsymbol{Y}_i$ (or errors $\boldsymbol{\epsilon}_i$) of each data unit, *all* model aspects describe features *within* data units.

Model interpretation The simple mean model in (6.3) suggests parallel profiles of the mean responses in the three treatment groups and corresponds to the case $t \leq 3$ of a more complex model used in Diggle *et al.* (1995, p. 96). The error model in (6.4) summarizes the random variation which cannot be explained by the deterministic mean model. It requires that the errors are composed (in an additive way) of the components denoted as random effects, serial correlation, and measurement errors (Diggle *et al.*, 1995, p. 79f).

- The *random effects* represented by U_i describe variation among levels of individual response profiles.
- The *serial correlation* represented by $\{W_i(t)\}$ refers to the correlation among pairs of observations of the *same* sampling unit which again depends on the interval length h between corresponding measurement times. With $\rho_W(\cdot)$ as in (6.5) and equally spaced discrete time points for all sampling units, the serial correlation of the present model follows a first-order autoregressive process (Diggle, 1988, p. 961) and becomes weaker with increasing h.
- The *measurement errors* represented by Z_{ij} describe the remaining random variation for each single Y_{ij}. According to Diggle and co-authors this kind of variation is due to (real) measurement errors, i.e. "repeated measurements arbitrarily close in time" which "are sometimes imperfectly correlated" (Diggle, 1988, p. 960). However, it could also address 'extra' variation in form of interactions between the individual unit and the time of measurement, not accounted for in the mean

model. This inconsistency motivates to relate the model component Z_{ij} to distortion (see later).

The above combination of mean and error model shall stand for REALITY, DATA, and ASSUMPTIONS accordingly. Note once again that parameter values are considered to be fixed for the two former models and remain pending with the latter. In terms of REALITY and DATA the mean model is most certainly a simplification of the respective 'real models', since many more factors may influence the *theoretically true* as well as *really observed* responses. The error model consequently covers all the unexplained influence factors (see also the discussion in § 3.3.1.3). The ASSUMPTIONS, finally, are meant to be a simplification of the truth 'by default'.

In the context of longitudinal data analysis the error model is usually referred to as model for the *covariance/correlation* structure (Diggle *et al.*, 1995). This is because it is the non-trivial covariance structure of the responses (or errors) which distinguishes the overall model from a classical linear model with independent errors. While equation (6.4) formally defines such a covariance model, a more illustrative description of it can further be provided by the corresponding *variogram* (see below).

6.2.1.1 The variogram

The *variogram* of a general stochastic process $\{S(t) : t \in \mathbb{R}\}$ is a function which describes the expected association among subsequent realizations of the process. It is defined as

$$\mathrm{V}_S(u) = \frac{1}{2} \cdot \mathrm{E}\left\{[S(t) - S(t-u)]^2\right\}, \qquad u \geq 0.$$

If $\{S(t)\}$ is stationary it holds further

$$\begin{aligned} \mathrm{V}_S(u) &= \gamma(0) \cdot [1 - \rho_S(u)] \\ &= \gamma(0) - \gamma(u), \qquad u \geq 0, \end{aligned} \tag{6.6}$$

where $\gamma(u) = \gamma_S(u) = \mathrm{Cov}\,[S(t), S(t-u)]$ is the autocovariance function and $\rho_S(u) = \gamma(u)/\gamma(0)$ the autocorrelation function of the process (Diggle *et al.*, 1995, p. 51). Note that the variogram originates in zero and afterwards increases with decreasing correlation (not necessarily continuously).

For the specific error model in (6.4) it is

$$\mathrm{Var}(\epsilon_{ij}) = \nu^2 + \sigma^2 + \tau^2 \qquad \text{and} \tag{6.7}$$

$$\begin{aligned} \mathrm{Cov}(\epsilon_{ij}, \epsilon_{ik}) &= \mathrm{Var}(U_i) + \mathrm{Cov}\,[W_i(j), W_i(k)] \\ &= \nu^2 + \sigma^2 \cdot \rho_W(|j-k|) \end{aligned} \tag{6.8}$$

and hence for the autocovariance function of the errors ϵ_{ij} seen as a stationary stochastic process $\{\epsilon_{ij}\} = \{\epsilon(t_{ij})\} = \{\epsilon(j)\}$

$$\gamma_\epsilon(u) = \begin{cases} \nu^2 + \sigma^2 + \tau^2 & \text{for} \quad u = 0 \\ \nu^2 + \sigma^2 \cdot \rho_W(u) & \text{for} \quad u > 0. \end{cases}$$

With (6.6) this leads to the following variogram for the error-process characterizing the specific covariance structure underlying in (6.4)

$$\mathrm{V}_\epsilon(u) = \begin{cases} 0 & \text{for} \quad u = 0 \\ \tau^2 + \sigma^2 \cdot [1 - \rho_W(u)] & \text{for} \quad u > 0. \end{cases} \tag{6.9}$$

Note that the above formulation of $\mathrm{V}_\epsilon(\cdot)$ differs slightly from the one in Diggle *et al.* (1995, p. 87). The authors ignored the case $u = 0$.

Overall, the variogram $\mathrm{V}_\epsilon(\cdot)$ is important for inference purposes: Since

$$\lim_{u \to 0} \mathrm{V}_\epsilon(u) = \tau^2 \tag{6.10}$$

and with $\lim_{u\to\infty} \rho_W(u) = 0$ also

$$\lim_{u \to \infty} \mathrm{V}_\epsilon(u) = \tau^2 + \sigma^2 \tag{6.11}$$

initial estimates for the (squared) variance parameters can be derived by estimating $\mathrm{V}_\epsilon(\cdot)$ and the corresponding process variance $\nu^2 + \tau^2 + \sigma^2$. The next subsection will describe the respective inference problem as a whole.

6.2.2 Estimators and inference framework

The present chapter considers an inference framework which involves longitudinal model assumptions of the type introduced above (with or without Z_{ij}), and which simultaneously is aimed at the following two *aspects of interest*:

- the treatment specific *mean response profiles* parametrized in (6.3), and
- the *variogram* representing the (same) correlation structure *within* each data unit, parametrized in (6.9).

Both aspects of interest are described by a combination of relevant parameter estimates with assumed mean model or variogram formulation, respectively. Note that the latter has been preferred to the actual error model expression for illustrative purposes. In the simulation study to follow, the so called restricted maximum-likelihood (REML) method will be used for the estimation of the unknown model parameters. Diggle *et al.* (1995, p. 64ff and 92f) give a detailed description of the overall procedure, so here just a brief outline will be presented.

6.2.2.1 The REML-method

The REML-method is motivated by the fact that the ordinary maximum-likelihood (ML) estimators for the variance parameters may be biased, especially when the length p of the parameter vector $\boldsymbol{\beta}$ is long. Moreover, the ML-estimators may fail to be consistent, when a wrong form for X is assumed. Thus, following the general recommendation of Diggle *et al.* (1995) the simulation study in this chapter will be based on the REML-method, even though the mean model for the corresponding artificial reality/data-situations is always considered to be correct.

Overall, "the REML estimator is defined as a maximum likelihood estimator based on a linearly transformed set of data $\boldsymbol{Y}^* = A\boldsymbol{Y}$ such that the distribution of $\boldsymbol{Y}^*$ does not depend on $\boldsymbol{\beta}$" (Diggle *et al.*, 1995, p. 65).

While the matrix A is not explicitly involved in the actual REML-estimation procedure, the following steps need to be implemented (Diggle, 1988; Diggle *et al.*, 1995, p. 64ff and 92).

1. transform the variance parameters into σ^2 and $\boldsymbol{\alpha} = (\tau^2/\sigma^2,\ \nu^2/\sigma^2,\ \phi)'$,
2. for *given* $\boldsymbol{\alpha}$, evaluate the ML-estimator for $\boldsymbol{\beta}$ and the REML-estimator for σ^2 based on $\mathrm{RSS}(\boldsymbol{\alpha})/(nm-p)$, where RSS are the resulting residual sum of squares,
3. determine the REML-estimator of $\boldsymbol{\alpha}$ through maximization of the corresponding reduced log-likelihood based on the previous two estimates,
4. repeat step 2. given the REML-estimate of $\boldsymbol{\alpha}$.

A suitable algorithm for the above REML-estimation procedure is implemented in the OSWALD software (Smith *et al.*, 1996). It requires the provision of initial estimates for ν^2, τ^2, and ϕ which can be derived from the *sample variogram*, see below. Note that the parametric variogram-estimate $\widehat{\mathrm{V}}_\epsilon(\cdot)$, which will finally describe the corresponding *aspect of interest*, is based on the variance estimates obtained by the REML-method and the assumed variogram formulation.

6.2.2.2 The sample variogram and initial variance estimators

The *sample variogram* is " the empirical counterpart of the variogram" (Diggle *et al.*, 1995, p. 51). When measurements are taken at the same set of time points $t = 1, \ldots, n$ for each sampling unit as in the present example, the sample variogram is essentially unbiased and can be calculated as follows:

$$\widetilde{\mathrm{V}}_\epsilon(u) = \frac{1}{2 \cdot m \cdot |I_u|} \sum_{i=1}^{m} \sum_{(j,k) \in I_u} (r_{ij} - r_{ik})^2, \qquad u = 1, \ldots, n$$

where $I_u = \big\{(j,k) \in \{1, \ldots, n\}^2 : |\, j-k| = u\big\}$, and where r_{ij} are the residuals "obtained by subtracting from each measurement the ordinary least-squares

estimate of the corresponding mean response" assuming a saturated groups-by-times model (Diggle *et al.*, 1995, p. 51 and 91). Note that in contrast to the REML-method, the ordinary least-squares method ignores the underlying correlation structure.

Initial estimates of the squared variance parameters ν^2 and τ^2 can now be derived from the (smoothed) sample variogram by taking into account (6.10) and (6.11) as well as the estimated variance of the residual process as an estimate for $\nu^2 + \tau^2 + \sigma^2$. Together with a rough guess for ϕ (choose a larger value, if $\widetilde{\mathrm{V}}_\epsilon(\cdot)$ levels-off quickly), the initial estimates are employed to start the corresponding REML-estimation algorithm for the unknown model parameters (Smith *et al.*, 1996, p. 24f).

6.3 A situation of distortion

To focus the problem, distortion will affect the error model in (6.4), and in particular the random variables Z_{ij} which have been denoted as *measurement errors* by Diggle *et al.* (1995). Since the error model describes the correlation structure between measurements from the *same* sampling unit and hence model aspects *within* data units, again both types of distortion, data contamination (type ①) and model deviation (type ②), are in principle possible.

By including Z_{ij} into the model assumptions, Diggle and co-authors seem to follow the traditional approach of formulating the ASSUMPTIONS according to the DATA (compare with § 3.2). This is because measurement errors in their literal meaning are associated with the *effective data-generating process* (§ 3.3.1.3). An obvious question is now, whether one really should do so in case the Z_{ij} are *indeed* measurement errors and reflect a situation of *data contamination*. In example 5 (see chapter 3) the measurement errors are real for instance, because the protein content of cows milk is measured by "an assay technique which itself introduces a component of random variation"

(Diggle *et al.*, 1995, p. 80). Still, the Z_{ij} could also express extra random variation which in fact exists in the real-world situation under study. For example, there might be a patient-specific reaction to the treatment at a particular time t. Thus, dropping the Z_{ij} from the model assumptions would result in a situation of *model deviation*.

In order to address the issue of REML-performance under the above kind of (increasing) distortion, corresponding model disagreement between REALITY, DATA, and ASSUMPTIONS needs to be formalized. Again, this can be done by introducing a suitable notion of *model discrepancy*.

6.3.1 Model discrepancy

The longitudinal model *without* measurement error component Z_{ij} will be taken as the (non-distorted) reference model. Note that this correlation structure has already been used for data-analysis purposes in Pantula and Pollock (1985). *Discrepancy* from the (overall) model could intuitively be expressed by the variance of Z_{ij} denoted as τ^2. However, this choice shows the disadvantage of τ^2 ranging from zero to infinity. Instead, it seems to be reasonable to use the transformation $\alpha = \tau^2/(\nu^2 + \sigma^2 + \tau^2)$ as *discrepancy magnitude*. The latter is element of $[0, 1]$ and can easily be interpreted as the proportion of process variability which is *not* explained by serial correlation and random effects. The variance parameters ν, σ, τ, and ϕ will therefore be re-parametrized as follows:

$$\alpha = \frac{\tau^2}{(\nu^2 + \sigma^2 + \tau^2)} \tag{6.12}$$

and in addition

$$\xi_1 = \nu^2 + \sigma^2 + \tau^2, \qquad \xi_2 = \frac{\nu^2}{\sigma^2}, \qquad \text{and} \qquad \phi = \phi,$$

where ξ_1 represents the process variance and ξ_2 the random effect variance as a proportion of the serial component variance. Note that for $\alpha = 1$, the

original parametrisation with $\nu = \sigma = 0$ must be used since ξ_2 cannot be defined.

A discrepancy magnitude of $\alpha = 0$ represents the reference model *without* measurement errors, for which the corresponding variogram reduces to

$$\mathrm{V}_\epsilon(u) = \begin{cases} 0 & \text{for} \quad u = 0 \\ \sigma^2 \cdot [1 - \rho_W(u)] & \text{for} \quad u > 0. \end{cases} \tag{6.13}$$

As such, it is a *sub-model* of the original model in (6.3) and (6.4). In terms of the ASSUMPTIONS, where parameter values are *not* fixed, the latter again is associated with $\alpha = +$ (compare with § 4.2.2). The maximum discrepancy of $\alpha = 1$, at last, corresponds to the situation where the Z_{ij} fully explain the random errors ϵ_{ij} and thus imply a classical linear model with independent errors.

The following section will specify the inferential performance assessment for the present model and inference framework, in particular, by allowing for the above kind of distortion.

6.4 Inferential performance assessment

In contrast to chapter 5 where a single estimator (ACL-estimator) was used to entirely describe the only aspect of interest (survival function S_X), the inferential circumstances in the current example are more complex. Here, one can study the two compound curve estimators for the mean response profile and the variogram (the actual *aspects of interest*), but also the individual estimators for the model parameters, each on its own. Hence, a variety of *performance statistics* addressing the attribute *closeness* in terms of location (see § 2.4.2) will be defined in the following. As before, they are generally denoted as $\pi\,(R = \alpha_1,\ D = \alpha_2\ |A = \alpha_3,\ \Delta)$, where the discrepancy magnitudes α_1, α_2, and α_3 are to be specified in connection with suitable influence and preference graphs (see later).

6.4.1 Performance statistics

Variogram Global as well as local performance assessment will be carried out for the variogram estimator $\widehat{\mathrm{V}}_\epsilon(\cdot)$ seen as a finite-sample statistic: As the *average squared distance* between the true and the estimated variogram

$$\pi_{\bar{d}} = \frac{1}{n}\sum_{j=1}^{n}\left[\mathrm{V}_\epsilon(j) - \widehat{\mathrm{V}}_\epsilon(j)\right]^2, \tag{6.14}$$

and as the *squared distance* between the true and estimated variogram at the *local* measurement time $t = j$

$$\pi_{d(j)} = \left[\mathrm{V}_\epsilon(j) - \widehat{\mathrm{V}}_\epsilon(j)\right]^2.$$

While the true variogram refers to REALITY, the estimated variogram is based on the *assumed* variogram formulation and receives contributions from the corresponding variance estimates (determined from a DATA-generated sample).

Mean response Similar, a *global* performance statistic for the treatment-specific mean response profiles $\mathrm{E}(Y_g(\cdot))$ will be defined as

$$\pi_{Y(g)} = \frac{1}{n}\sum_{j=1}^{n}\left[\mathrm{E}(Y_g(j)) - \widehat{Y}_g(j)\right]^2,$$

which is the *average squared distance* between the true and estimated mean response, both for treatment group g. Reference to REALITY, DATA, and ASSUMPTIONS is provided as above.

Variance parameters Performance statistics for the individual variance parameters are finally given by

$$\pi_\sigma = (\sigma - \hat{\sigma})^2, \qquad \pi_\nu = (\nu - \hat{\nu})^2 \qquad \text{and} \qquad \pi_\phi = (\phi - \hat{\phi})^2.$$

Note that unlike before, the ASSUMPTIONS are *not* directly involved.

6.4.2 Influence and preference graphs

Once again, the methodology of *influence* and *preference* graphs will be used to study respective inferential performance under increasing distortion. This time, however, the latter may be based either on *differences* or on *quotients*, in order to assess performance in *absolute* and *relative* terms (see § 4.3).

Definitions will generally be formulated for both data contamination and model deviation. Still, it will not be possible to distinguish performance of the mean response estimator $\widehat{Y}_g(\cdot)$ for the two cases, since distortion does *not* relate to the mean model (see § 4.2.1). The "interpretation" of the mean parameters "is essentially independent of the correlation structure" (Diggle *et al.*, 1995, p. 131). The same also applies to the variance parameter ϕ which can be expressed independently from the discrepancy magnitude α (see § 6.3.1).

Influence graphs study performance of the various REML-estimators in terms of potential change under increasing distortion. For *data contamination* (type ① distortion) they address the question

→ What difference does it make, when the data contain more and more measurement errors, while the analysis is still based on the assumptions of the sub-model $(A = 0)$?

They are denoted as

$$g_i^D(\alpha) = \left\{ \begin{array}{c} \mathrm{E}_{\,\alpha}\Big[\pi(R=0, D=\alpha \;|A=0, \mathrm{REML})\Big] \\ \boxed{\;-\quad \text{or} \quad :\;} \\ \mathrm{E}_{\,0}\Big[\pi(R=0, D=0 \;|A=0, \mathrm{REML})\Big], \end{array} \right.$$

where π stands for one of the previously defined performance statistics, α represents a discrepancy magnitude or the resulting amount of distortion, respectively, and the symbol $\boxed{\;-\quad \text{or} \quad :\;}$ stands for one of the two

arithmetic operations indicated. Again, the expected values are based on a DATA-model.

Influence graphs for type ② *model deviation* address the question

→ What difference does it make, when the additional (true) random variation increases while still being ignored in the model assumptions?

They are denoted as

$$g_i^M(\alpha) = \left\{ \begin{array}{c} \mathrm{E}_\alpha\Big[\pi(R=\alpha, D=\alpha \mid A=0, \mathrm{REML})\Big] \\ \boxed{-\quad \text{or} \quad :} \\ \mathrm{E}_0\Big[\pi(R=0, D=0 \mid A=0, \mathrm{REML})\Big]. \end{array} \right.$$

On the whole, the influence graphs *increase* when the performance of the corresponding REML-estimator deteriorates under growing distortion. By definition, they originate with the value zero or one.

Preference graphs in the present example will compare REML-performance under the two assumptions $A = 0$ and $A = +$, which either *ignore* or *include* the measurement errors Z_{ij}. They correspond to the preference graphs introduced in (4.5) and (4.6). For *data contamination* they address the question

→ Should one ignore measurement errors (of a certain amount) in the data or should one include them into the model assumptions?

They are denoted as

$$g_p^D(\alpha) = \left\{ \begin{array}{c} \mathrm{E}_\alpha\Big[\pi(R=0, D=\alpha \mid A=0, \mathrm{REML}) \\ \boxed{-\quad \text{or} \quad :} \\ \pi(R=0, D=\alpha \mid A=+, \mathrm{REML})\Big]. \end{array} \right.$$

Preference graphs for type ② *model deviation* address the question

→ Should one take into account the true extra random variation (of a certain amount) when formulating the assumptions or not?

They are denoted as

$$g_p^M(\alpha) = \left\{ \begin{array}{l} \mathrm{E}_\alpha\Big[\pi(R=\alpha, D=\alpha \mid A=0, \mathrm{REML}) \\ \qquad\qquad \boxed{\quad - \quad \text{or} \quad : \quad} \\ \qquad\qquad \pi(R=\alpha, D=\alpha \mid A=+, \mathrm{REML})\Big]. \end{array} \right.$$

Overall, absolute preference values below 0 (or 1) indicate advantages for $A = 0$, and they signify a preference for $A = +$ when resulting above 0 (or 1). In the simulation study below, this rule will obviously be weakened due to the inaccuracy associated with averaging finite sample values.

6.5 Simulation study

The problem of 'measurement errors in the data' will now be considered from a finite sample point of view via simulation. Should model assumptions take into account measurement errors if the latter indeed imply wrong data? What happens in the case when they are ignored while being a correct reflection of the real-world? These and other questions are addressed by means of influence and preference graphs which study inferential performance under increasing distortion. For the present example they have been specified in the preceding section. Once again, the distinction of data contamination and type ② model deviation is of central importance. As before the two situations are expressed by corresponding model disagreement between REALITY, DATA, and ASSUMPTIONS using a particular notion of *model discrepancy* (§ 6.3).

The simulation study is carried out in S-PLUS (version 3.4, release 1) on a Sun-compatible work station. In particular it uses the OSWALD software (version 2.6) written for the S environment by Smith *et al.* (1996).

6.5.1 Set-up

6.5.1.1 Problems addressed

Primary questions Longitudinal data are explained by the sub-model which involves *no* measurement error component, i.e. $A = 0$, while (increasing) random variation of the kind Z_{ij} is present in the sample: How is the statistical analysis for the *correlation structure*, based on REML-estimation, influenced when this kind of distortion reflects

- growing data contamination, i.e. the random variation is due to measurement errors?
- growing model deviation, i.e. the random variation is a true reflection of the unknown real-world situation?

Is there a difference between the influence under data contamination and model deviation? When is it sensible to include Z_{ij} into the model assumptions and when not?

Further details of interest

- What effect do other elements of the model acquire, in particular
  - the length n of each measurement series (data unit),
  - the total number of data units m,
  - the values of the (transformed) variance parameters ξ_2 and ϕ, which are fixed in REALITY and DATA?

  When do they work against and when do they worsen the (bad) effects of distortion?
- Do certain regions of the variogram-estimator suffer more under distortion than others?

- How much is the estimation of the mean-response and the variance parameter ϕ influenced under increasing discrepancy between DATA and ASSUMPTIONS? Note, a distinction between performance under data contamination and model deviation is not possible. When is it important that data and model assumptions correspond to each other with respect to Z_{ij}?

6.5.1.2 Program organization

1. Simulation of B data sets according to an α-distorted DATA-model: Each data set is a $(m \times n)$–matrix of observations, where the n measurements of a single sampling unit correspond to one row of the matrix.

   - Choose ξ_1 and ξ_2 and evaluate the original variance parameters ν, σ, and τ. For $\alpha = 1$ the values of ν and σ always result to zero, while τ is determined by ξ_1. In addition, choose ϕ.
   - Simulate B single $(m \times n)$–data matrices Y according to

   $$Y = \begin{pmatrix} (X_1 \cdot \boldsymbol{\beta})' \\ \vdots \\ (X_m \cdot \boldsymbol{\beta})' \end{pmatrix} + \epsilon,$$

   where X_i is one of the $(n \times 4)$–matrices of explanatory variables in (6.2), $\boldsymbol{\beta}$ is a fixed parameter vector of length 4, and ϵ is the $(m \times n)$–matrix of random errors with components ϵ_{ij} as in (6.4). The latter is derived by

   $$\epsilon = \begin{pmatrix} \boldsymbol{Z}'_1 \\ \vdots \\ \boldsymbol{Z}'_m \end{pmatrix} \cdot W, \tag{6.15}$$

   where $\boldsymbol{Z}_i$ are independent column-vectors of n independent $N(0,1)$-distributed random variables, each, and where W is an upper-triangular $(n \times n)$–matrix of coefficients (see Appendix A.2).

2. Evaluation of the *true variogram* according to (6.9) under model deviation ($R = \alpha$, $D = \alpha$, and $A = 0$) and according to (6.13) under data contamination ($R = 0$, $D = \alpha$, and $A = 0$).

3. Evaluation of the treatment specific *true mean response profiles* according to (6.3) using the pre-assigned mean parameters b_{01}, b_{02}, b_{03} and b_1. Note that *no* distinction is made between the two types of distortion, since they do not affect the mean model.

4. Determination of *initial estimates* for ν^2 and τ^2 by

   - fitting a saturated groups-by-times model to each data set using the ordinary least-squares method (OSWALD-function "olsres"),
   - evaluating a sample variogram and estimating the process variance $\nu^2 + \tau^2 + \sigma^2$ from each resulting set of residuals (OSWALD-function "variogram"), and finally
   - deriving the estimates according to (6.10) and (6.11).

   The 'initial estimate' for ϕ shall always be the true, in a practical situation unknown, value which has been chosen in the beginning.

5. Estimation of the mean and variance parameters via the REML-method (OSWALD-function "pcmid"): For each (α-distorted) data set the procedure is applied twice, first *without* the assumption of measurement errors ($A = 0$), i.e. the starting value for τ^2 equals zero, and second *with* the assumption of measurement errors ($A = +$), i.e. the starting value for τ^2 is the initial estimate determined in the previous step. The starting values for ν^2 and ϕ are the same in both cases.

6. Estimation of the *variogram* and the treatment specific *mean response profiles* via substitution of the above REML-estimates into the assumed variogram and mean response formula, respectively. Note that for each data set one obtains estimates based on $A = 0$ as well as $A = +$.

7. Evaluation of individual influence and preference values $g_i(\alpha)$ and $g_p(\alpha)$.

   - Choose one of the performance statistics in § 6.4.1 and the type of distortion (data contamination or type ② model deviation).
   - With $R = 0$ for data contamination and $R = \alpha$ for model deviation, determine for each simulated sample

     $$\pi(\alpha) \;=\; \pi(R, D = \alpha \mid A = 0, \mathrm{REML}),$$

     and

     $$\pi_d^*(\alpha) \;=\; \pi(R, D = \alpha \mid A = 0, \mathrm{REML}) - \pi(R, D = \alpha \mid A = +, \mathrm{REML})$$

     or

     $$\pi_q^*(\alpha) \;=\; \pi(R, D = \alpha \mid A = 0, \mathrm{REML}) \,/\, \pi(R, D = \alpha \mid A = +, \mathrm{REML}).$$

   - Take the average of $\pi(\alpha)$ over the B simulated samples and subtract from it (or divide by) the corresponding average $\pi(0)$. The result serves as contribution for the influence graph based on differences (or based on quotients).
   - Take the average of $\pi_d^*(\alpha)$ over the B simulated samples as contribution for the preference graph based on differences, and take the median of $\pi_q^*(\alpha)$ when the preference graph is based on quotients.

8. Assessment of the sample-based variability for each of the above contributions:

   - Evaluate the corresponding 'one standard-deviation limit' over the B simulated samples by adding and subtracting the estimated value of
     - $\sqrt{\mathrm{Var}_\alpha\,[\pi(\alpha)]}$ for influence graphs based on differences,

  - $\sqrt{\mathrm{Var}_\alpha\left[\pi(\alpha)\right]}/\,\mathrm{E}_0\left[\pi(0)\right]$ for influence graphs based on quotients.
  - $\sqrt{\mathrm{Var}_\alpha\left[\pi_d^*(\alpha)\right]}$ for preference graphs based on differences,
- Evaluate the corresponding (empirical) interquartile range of $\pi_q^*(\alpha)$ over the B simulated samples for preference graphs based on quotients.

9. For the *overall* influence and preference graphs the previous steps are repeated with different values of α.

6.5.1.3 Design

- The parameters of the *mean model* in DATA and REALITY are set to

$$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_4)' = (b_{01}, b_{02}, b_{03}, b_1)' = (1, 3, 5, 0.3)'.$$

- The transformed parameters of the *error model* (correlation structure) in DATA and REALITY are chosen as
  - $\xi_1 = 1$ standardizing of the variance of the error process.
  - $\xi_2 \in \{0.1, 1, 10\}$ providing some possibilities for the proportion of the random effect and serial component variance.
  - $\phi \in \{0.1, 0.5, 1\}$ covering stronger as well as weaker autocorrelation between observations of the same data unit.
- The discrepancy magnitude of DATA, and under model deviation also of REALITY, varies according to $\alpha = 0, 0.1, \ldots, 0.9, 1$. This likewise applies to the *amount of distortion* when considering the model triplet as a whole.
- The data is generated according to
  - $B = 100$, number of simulated data sets, and

– $(m,n) \in \{(\mathbf{48},6),(\mathbf{24},6),(\mathbf{12},6),(\mathbf{6},\mathbf{6}),(6,\mathbf{12}),(6,\mathbf{24}),(6,\mathbf{48})\}$ for the consideration of growing numbers of subjects with fixed numbers of measurements and vice versa.

or $(m,n) \in \{(6,\mathbf{3}),(12,\mathbf{3}),(24,\mathbf{3})\}$ for cross-over like situations with a very small number of measurements,

or $(m,n) \in \{(6,6),(12,12),(24,24)\}$ for balanced samples of increasing size.

Each data set has $m/3$ data units per treatment group.

6.5.2 Results

Considerations will begin with a balanced 'starting situation' which is followed by respective studies under increasing data dimension (m,n) and under different values for ξ_2 and ϕ.

6.5.2.1 Starting situation

A first analysis will be based on data with $(m,n) = (12,12)$, $\xi_1 = \xi_2 = 1$, and $\phi = 0.5$. See the Figures C.1 – C.4 in Appendix C for simulated mean responses and sample variograms.

Parametric variogram estimates derived from α-distorted samples and based on $A = 0$ and $A = +$ are shown in Figures 6.1, 6.3, and 6.5. They can be compared with the corresponding true variogram curves under $R = 0$ and $R = \alpha$. Note the following representations:

–○–: estimated variogram with $A = 0$,

–●–: estimated variogram with $A = +$,

·····: true variogram under $R = 0$, and

——: true variogram under $R = \alpha$.

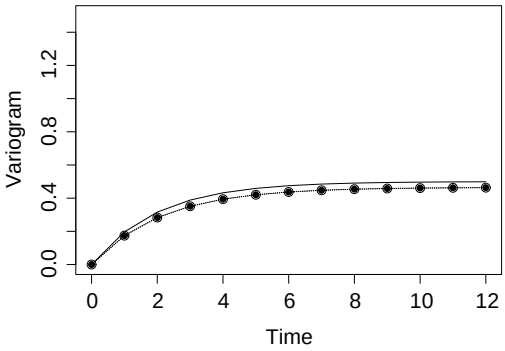

Figure 6.1: *Variogram estimates under no distortion* $\alpha = 0$.

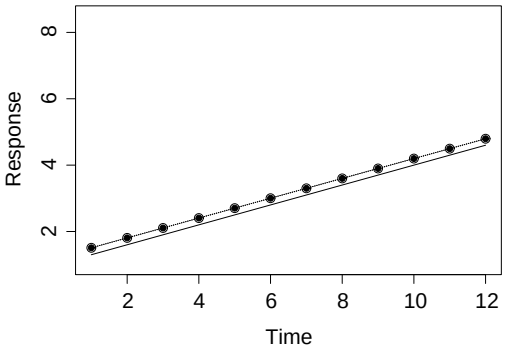

Figure 6.2: *Mean response estimates (group 1) under no distortion* $\alpha = 0$.

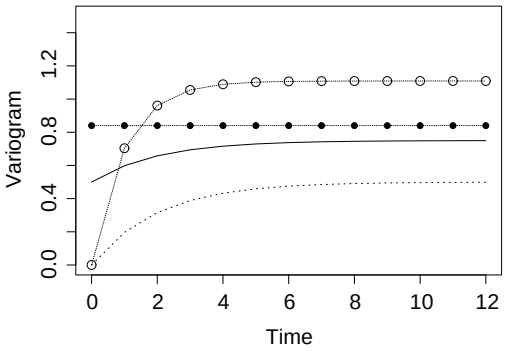

Figure 6.3: *Variogram estimates under distortion* $\alpha = 0.5$.

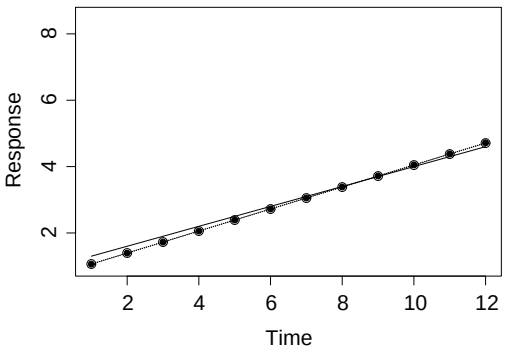

Figure 6.4: *Mean response estimates (group 1) under distortion* $\alpha = 0.5$.

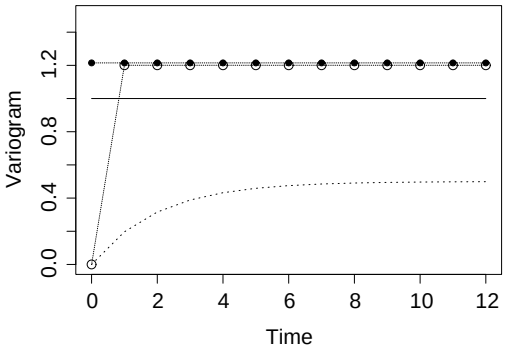

Figure 6.5: *Variogram estimates under distortion* $\alpha = 1$.

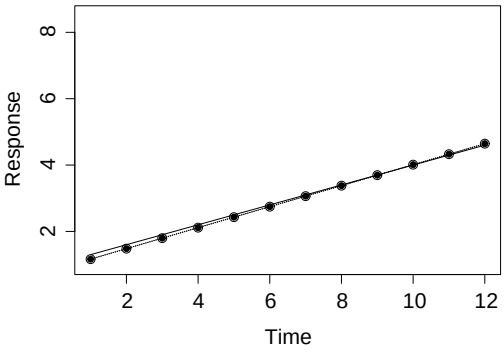

Figure 6.6: *Mean response estimates (group 1) under distortion* $\alpha = 1$.

Consideration of estimates based on $A = 0$ indicates the following: Under growing data contamination ($R = 0$) the variogram tends to be more and more overestimated, while the origin remains without bias. Under growing model deviation ($R = \alpha$) it is overestimated, too, but a trend does not become obvious. The origin, in this case, is clearly *under*estimated. Consideration of estimates based on $A = +$ suggest that under growing data contamination ($R = 0$) the variogram is increasingly overestimated, and this on the whole time scale. Similarly, when $R = \alpha$ (no model deviation!) it seems to be (slightly) overestimated, again on the whole time scale.

Respective estimates of the mean response in treatment group 1 are shown in Figures 6.2, 6.4, and 6.6. Performance is identical under both *types* of distortion (by default), and seems to be very similar also under the different *amounts* of distortion (see the Figures C.5 – C.10 in Appendix C for group 2 and 3).

The above conclusions, each based on a *single* simulated sample, can only give first insights to the problem. The study of *influence* and *preference graphs* will now present more reliable answers using *repeated* simulations.

Influence Figure 6.7 exemplifies an influence graph based on differences (D-influence graph) for the parametric variogram estimator $\widehat{\mathrm{V}}_\epsilon(\cdot)$ under data contamination. The two —|—lines represent the one standard-deviation bounds. The bad influence on the estimator performance becomes quite obvious. Indeed, a change for the worse of factor 49 at $\alpha = 1$ is indicated by the respective influence graph based on quotients (no Figure). *No* influence, however, becomes apparent under model deviation (Figure 6.8). Note that the statistic $\pi_{\bar{d}}$ studies (global) performance from time $t = 1$ onwards. Behaviour down to the origin ($t = 0$) is being neglected.

Pointwise performance considerations for $\widehat{\mathrm{V}}_\epsilon(\cdot)$ based on $\pi_{d(j)}$ will provide further details: Under data contamination, maximum differences in performance commence with 0.6 for time $t = 1$ and decrease with 0.3 for $t = 7$ down to 0.2 for $t = 12$ (Figures C.11 – C.13 in Appendix C). This corresponds

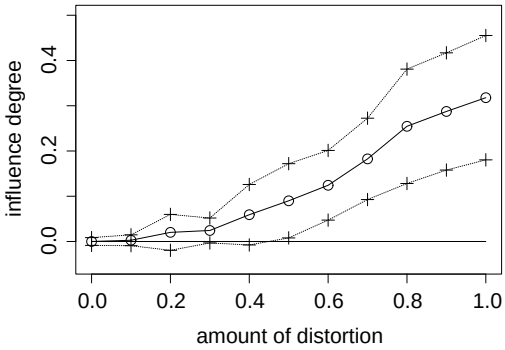

Figure 6.7: *D-influence graph for data contamination based on* $\pi_{\bar{d}}$.

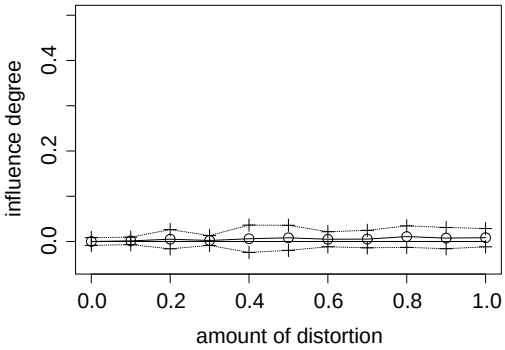

Figure 6.8: *D-influence graph for model deviation based on* $\pi_{\bar{d}}$.

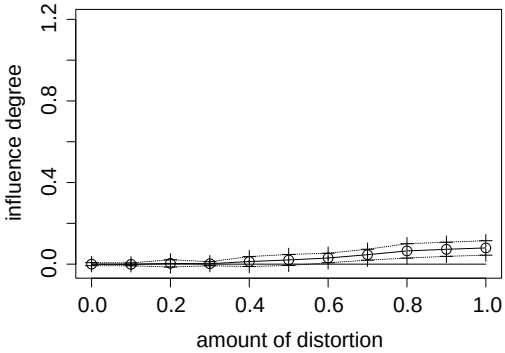

Figure 6.9: *D-influence graph for data contamination based on* π_{σ}.

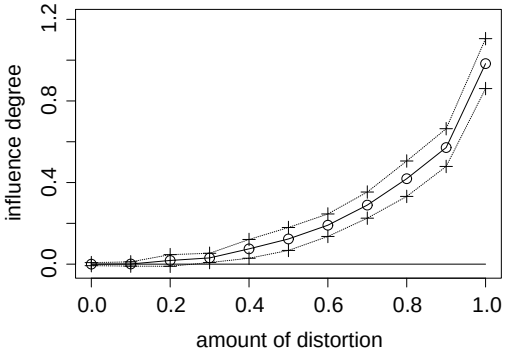

Figure 6.10: *D-influence graph for model deviation based on* π_{σ}.

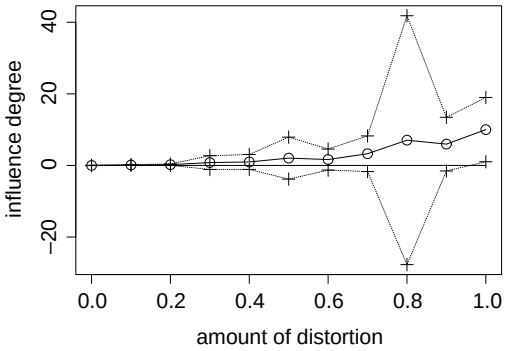

Figure 6.11: *D-influence graph based on* π_{ϕ}.

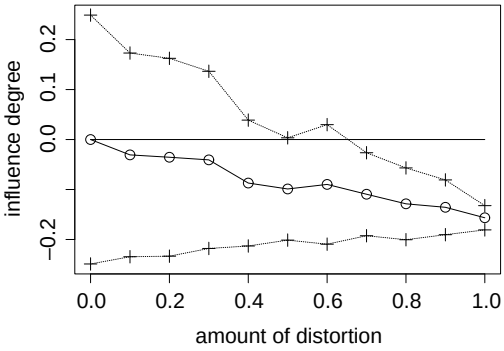

Figure 6.12: *D-influence graph based on* $\pi_{Y(1)}$.

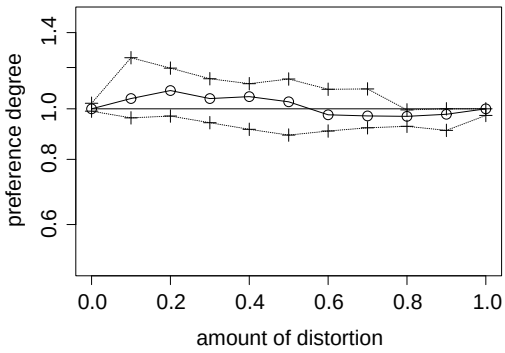

Figure 6.13: *Q-preference graph for data contamination based on* $\pi_{\bar{d}}$.

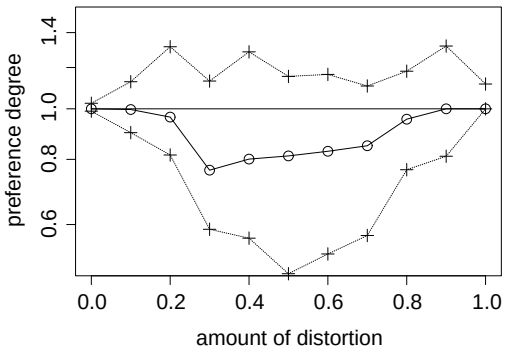

Figure 6.14: *Q-preference graph for model deviation based on* $\pi_{\bar{d}}$.

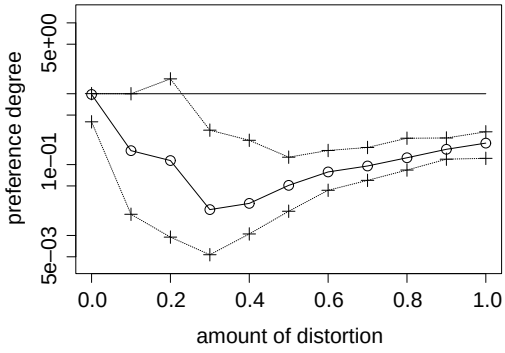

Figure 6.15: *Q-preference graph for data contamination based on* π_{σ}.

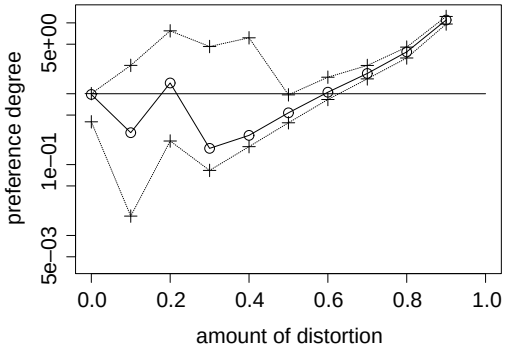

Figure 6.16: *Q-preference graph for model deviation based on* π_{σ}.

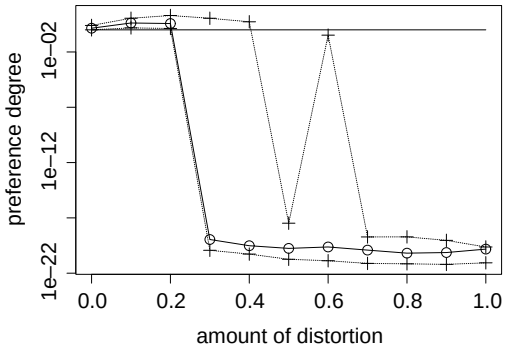

Figure 6.17: *Q-preference graph based on* π_{ϕ}.

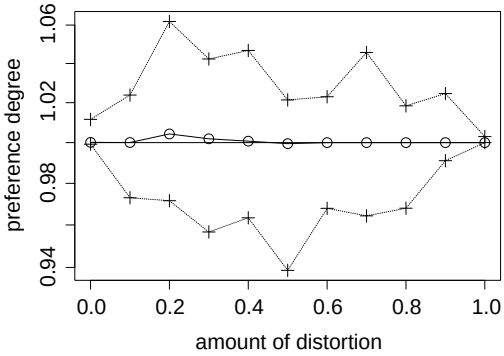

Figure 6.18: *Q-preference graph based on* $\pi_{Y(1)}$.

to maximum factors of degradation of about 980, 30, and 20, as indicated by respective influence graphs based on quotients (no Figure). Thus, local (negative) influence on the variogram estimator decreases with larger times of measurement. Model deviation shows the same phenomenon. However, any influence very soon becomes negligible (Figure C.14 in Appendix C). Note again that performance at the origin has *not* been considered.

Under $A = 0$ the variogram estimator $\widehat{\mathrm{V}}_\epsilon(\cdot)$ is composed of the variance estimators $\hat{\sigma}$ and $\hat{\phi}$ according to (6.13). Hence: Is the performance of the two individual estimators affected in a similar way? Figures 6.9 and 6.10 give the influence graphs for $\hat{\sigma}$ (based on differences). The estimator performance is badly influenced under both types of distortion, though much more seriously under model deviation. The degradation reaches factors up to 15 under increasing data contamination, and even up to 173(!) under model deviation (shown by respective influence graphs based on quotients, no Figure). The latter becomes visible through the abrupt 'jump' of the variogram estimate based on $A = 0$ (see the $-\circ-$line in Figure 6.5). Similarly, the estimator $\hat{\phi}$ seems to perform worse (Figure 6.11). The high variability of the influence values (as indicated by the one-standard deviation bounds), however, makes the influence less clear.

The parameter ν does not contribute to the variogram. Still, it is involved in the REML-estimation algorithm and it is of interest to notify negative influence under data contamination (performance becomes up to 6 times as bad) and slight positive(!) influence under model deviation (no Figure).

Final attention will be given to the three mean response estimators $\widehat{Y}_g(\cdot)$, $g = 1, 2, 3$. As with $\hat{\phi}$, performance *cannot* be distinguished under data contamination and model deviation. While an improvement can be seen for the first treatment group up to a factor of 10 (Figure 6.12), *no* influence is apparent for the two other groups (Figures C.15 and C.16 in Appendix C).

Preference Influence graphs studied the change in inferential performance when using $A = 0$. Are the more general model assumptions $A = +$ expected

to improve the real-world descriptions, especially in cases where $A = 0$ is negatively influenced? Which assumptions should be *preferred* under a given situation of distortion?

Figures 6.13 and 6.14 show preferences graphs for the variogram estimator $\widehat{\mathrm{V}}_\epsilon(\cdot)$. They are based on quotients and use a logarithmic scale for the y-axis (Q-preference graph). Note that variability among the simulated preference values is described by interquartile ranges. *None* of the two graphs can indicate a clear preference for $A = 0$ or $A = +$. Preference graphs based on differences confirm this result (Figures C.17 and C.18 in Appendix C). Pointwise considerations of the variogram estimator come to the same conclusions, with only one exception: Prefer $A = 0$ at measurement time $t = 1$ under moderate data contamination (no Figure).

The variance estimator $\hat{\sigma}$ shows more distinctive results. See Figures 6.15 and 6.16 for preference graphs based on quotients. While $A = 0$ appears to be preferable under data contamination, $A = +$ is expected to produce better results under more serious model deviation, i.e. $\alpha > 0.6$ (an extremely large preference value for $\alpha = 1$ has been excluded from the second figure). Note that arithmetic differences between performance levels do lead to smoother curves. The estimation of ϕ seems to give better results under increasing distortion given the assumptions $A = 0$ (Figure 6.17). This is because the use of $A = +$ tends to produce extraordinarily large estimates for ϕ (see the discussion in the next section).

Performance comparisons of $\hat{\nu}$ and the three mean response estimators do *not* indicate any particular preference (see Figure 6.18 for the first treatment group). The results so far are summarized in Tables 6.1 and 6.2. Note that a preference for $A = +$ with $\hat{\sigma}$ under model deviation is only given for $\alpha > 0.6$.

6.5.2.2 Increasing data dimension

Do the above results change with different number of data units m and/or observations n (per data unit)? It remains to be $\xi_1 = \xi_2 = 1$ and $\phi = 0.5$.

Table 6.1: *Influence and preference under data contamination and model deviation (starting situation).*

	data contamination		model deviation	
	influence	preference	influence	preference
$\widehat{\mathrm{V}}_\epsilon(\cdot)$	bad	no	no	no
$\hat{\sigma}$	bad	$A = 0$	very bad	$(A = +)$
$\hat{\nu}$	bad	no	slightly good	no

Table 6.2: *Influence and preference under distortion not to be distinguished (starting situation).*

	influence	preference
$\hat{\phi}$	bad	$A = 0$
$\widehat{Y}_1(\cdot)$	good	no
$\widehat{Y}_2(\cdot)$	no	no
$\widehat{Y}_3(\cdot)$	no	no

Increasing m and n: The consideration of balanced data situations with $(m, n) = (6, 6), (12, 12), (24, 24)$ indicates the following: While the *absolute* bad influence (differences) remains almost the same, the *relative* bad influence (quotients) increases clearly, both under data contamination and model deviation (see Table 6.3). This also applies to the estimator $\hat{\phi}$, though the influence continues to be less distinct due to the high variability. The positive influence of $\hat{\nu}$ under model deviation dies out, while it improves for the estimator $\widehat{Y}_1(\cdot)$. The overall preference pattern does not change very much. Still, there is a preference for $A = 0$ with the variogram estimator under data contamination when $(m, n) = (6, 6)$ and considerations are based on quotients. Other local irregularities might be eliminated if simulating a larger number of samples B (see the discussion in the next section).

Table 6.3: *Maximum influence factors for increasing m and n.*

	data contamination			model deviation		
(m,n)	$(6,6)$	$(12,12)$	$(24,24)$	$(6,6)$	$(12,12)$	$(24,24)$
$\widehat{\mathrm{V}}_\epsilon(\cdot)$	26	49	169	no influence		
$\hat{\sigma}$	4	15	78	38	173	890
$\hat{\nu}$	2	6	30	influence dies out		

Increasing m: The study of $n = 6$ and $m = 6, 12, 24, 48$ comes to very similar conclusions: The influence on inferential performance remains the same in *absolute* terms and increases in *relative* terms, though to a smaller extent. The preference pattern does not change and is the same as in the starting situation, with a difference for the variogram estimator and $\hat{\nu}$: For smaller m, the Q-preference graphs indicate a tendency to prefer $A = 0$ under data contamination.

Also cross-over like situations as represented by $n = 3$ and $m = 6, 12, 24$ confirm the overall trends in *relative* influence. However, influence factors are generally smaller and the high variability leads to less distinct results. A *clear* negative influence can only be verified for $\widehat{\mathrm{V}}_\epsilon(\cdot)$ under data contamination and for $\hat{\sigma}$ under model deviation. The preference behaviour can be described as in the previous paragraph.

Increasing n: Simulation results based on data with $m = 6$ and $n = 6, 12, 24, 48$ come again to very similar conclusions: While the *absolute* influence does not change, the *relative* influence increases with growing n. This time the latter applies also to positive influence. Compared to the balanced data situations (Table 6.3) the trends are weaker for negative and stronger for positive influence. The preference pattern is identical to the one observed in the starting situation.

6.5.2.3 Different values for the variance parameters

Taking the starting situation the variance parameters are varied, one at the time, according to $\xi_2 = 0.1, 1, 10$ or $\phi = 0.1, 0.5, 1$.

Increasing ξ_2: Absolute as well as relative influence increases with a growing proportion of ν^2/σ^2 in the DATA-model. This applies to positive as well as negative influence. The relative influence of the variogram estimator (under data contamination) and of $\hat{\sigma}$ even appears to explode. Maximum influence factors for $\widehat{\mathrm{V}}_\epsilon(\cdot)$ under data contamination are for example: 4, 49, and 4215(!) with $\xi_2 = 0.1, 1$, and 10, respectively. The preference pattern remains unchanged, except for $\hat{\sigma}$. For the latter, preference changes to $A = +$ under data contamination and still becomes more established for $A = +$ under model deviation ($\xi_2 = 10$: prefer $A = +$ already for $\alpha > 0.2$).

Increasing ϕ: Note that the serial correlation *de*creases with increasing parameter value ϕ. This seems to lead to increasing (relative) influence for $\hat{\sigma}$, $\hat{\nu}$ (data contamination), and $\hat{\phi}$, and to decreasing influence (absolute and relative) for the variogram estimator, the mean response estimator $\widehat{Y}_1(\cdot)$, and $\hat{\nu}$ (model deviation). Especially the results for $\widehat{\mathrm{V}}_\epsilon(\cdot)$ indicate that the underlying trends should not be considered to be linear. The preference pattern remains unchanged.

6.5.3 Discussion

The main conclusion of chapter 5 can also be drawn for the present study: When distortion is related to the estimand, inferential performance (in terms of location-closeness) is *different* under data contamination and model deviation. The variogram estimator $\widehat{\mathrm{V}}_\epsilon(\cdot)$ suffers *bad* influence under data contamination and *no* influence under model deviation. Still, *no* overall preference for either of the assumptions $A = 0$ or $A = +$ becomes obvious

(though there could be advantages for $A = 0$ under data contamination when n is small). This suggests that both approaches perform rather data-dependently and use less information from the model assumptions (the opposite has been concluded for the ACL-estimator in chapter 5).

Based on $A = 0$ the variogram *estimate* starts in the origin. With growing discrepancy magnitude α in DATA, it then tends to die out at a level which is increasingly high (see Figures 6.1, 6.3, and 6.5). This is related to the following facts (keeping in mind (6.10) and (6.11)):

- $\hat{\nu}$ seems to be mainly data-driven, i.e. the estimates are small when α in DATA is large. This leads to bad performance of $\hat{\nu}$ under data contamination and good performance under model deviation.
- Since τ^2 is set to zero in the REML-algorithm, the latter tries to express the process variance $\nu^2+\sigma^2+\tau^2$ with the remaining parameters ν^2 and σ^2. As a consequence, the estimate for σ tends to (strongly) increase with larger α in DATA. Especially under model deviation ($R = \alpha$) this leads to bad performance of $\hat{\sigma}$, since the true σ is small.

Consider also the *true* variogram curves. While the level-off value $\tau^2 + \sigma^2$ of the true variogram is moderate with respect to data contamination ($R = 0$, i.e. $\tau = 0$), it converges to the overall process variance with increasing α when referring to model deviation ($R = \alpha$). The conflict between a potential variogram estimate based on $A = 0$ and the true variogram under growing data contamination (and usually *not* under growing model deviation) becomes understandable.

The above conflict under data contamination becomes more serious when the true variogram dies out at a level σ^2 which gets closer to zero. This is the case for increasing $\xi_2 = \nu^2/\sigma^2$ (independently from α). A similar effect, but to a much smaller extent, seems to be produced with decreasing parameter ϕ. The true variogram levels-off more slowly and remains *below* the level σ^2 for a longer time. However, one can imagine that this trend in

influence becomes soon negligible in the opposite direction. This is due to the exponential structure of the autocorrelation function $\rho_W(\cdot)$ in (6.5) and the fact that performance assessment for the variogram estimator starts at $t = 1$ only.

Performance of the variogram estimator, for the same reasons, is only little affected by the increasing *over*estimation of ϕ. Large estimates of ϕ (up to a value of about 30 when $\alpha = 1$ and estimation is based on $A = 0$) very soon loose their weight over time t. The phenomenon becomes even more important with respect to the assumptions $A = +$. Here, the REML-algorithm more and more often produces *extremely* large estimates for ϕ (of around 10e20!) under increasing distortion (presumably because of the fixed starting value). Still, the two estimated variogram curves for $A = 0$ and $A = +$ tend to be similar beyond $t = 1$. Huge estimates of ϕ under $A = +$, nevertheless, seem to be (partly) responsible for the preference pattern of $\hat{\sigma}$: Given the former, estimates of σ are close to zero. This is of advantage for $\hat{\sigma}$ when α in REALITY gets large and the true σ tends to be small (preference for $A = +$ under more serious model deviation) and also shows when the true σ decreases with larger ξ_2.

Considerations under growing data dimension (m, n) indicate that bad influence (if present) increases in *relative* terms, but remains the same in *absolute* terms. In these cases distortion seems to have an *additive* effect independently from the data dimension. Hence, inferential performance can be expressed as $v_0 + \Delta(\alpha)$ where v_0 refers to the ideal situation of no distortion and converges to zero for $n, m \to \infty$ (assuming asymptotic unbiasedness at $\alpha = 0$), and where $\Delta(\alpha) > 0$ is independent of (m, n). Relative *positive* influence may however increase or decrease, since the corresponding additive effect $\Delta^*(\alpha) < 0$, if it exists, is dependent on (m, n) due to $|\Delta^*(\alpha)| \leq v_0$.

The global performance statistic $\pi_{\bar{d}}$ neglects curve-segments before time $t = 1$. This, obviously has a great impact on performance descriptions of the variogram estimator (local considerations indicated highest influence for small t). Alternative approaches might be of interest, which e.g. focus on the

location of the estimated level-off point (if it exists).

Up to this point, the discussion pertained to the performance of the variogram estimator. A further interesting conclusion refers to the mean response estimator(s): A discrepancy between ASSUMPTIONS and DATA in terms of α seems to influence the first treatment group only. There, performance of $\widehat{Y}_1(\cdot)$ appears to improve(!). Still, all treatment groups show the same performance quality with respect to $A = 0$ and $A = +$ and *no* preference is indicated. Note that performance cannot be distinguished between data contamination and model deviation here, since distortion is not related to the estimand. Estimators for the individual mean (response) parameters b_{01}, b_{02}, b_{03}, and b_1 have not been considered separately to limit the overall length of the study. Finally note the following:

- Some influence and preference graphs appeared rather irregular, especially the latter based on quotients. It might therefore be sensible to increase the number B of simulated samples for any future studies.

- It is generally difficult to simulate human interactions within the inference process (Chatfield, 1995, p. 434). The starting value ('initial estimate') for ϕ in the REML-algorithm has been fixed to the true value 0.5. In practice, however, the applied statistician would give a (more or less subjective) guess by looking at the resulting sample variogram. Presumably, the extremely large REML-estimates of ϕ could have been avoided by such a human decision. It is moreover common to reformulate the model assumptions when the sample variogram indicates non-stationarity in the data or when (final) diagnostic checks turn out to be unsatisfactory (Diggle *et al.*, 1995, p. 91ff). We simulated situations where the model is assumed to be known *a priori*.

6.6 Final remarks

As in chapter 5, the simulation study is meant to illustrate the ideas and concepts of the previous theoretical chapters.

Longitudinal data is characterized by the fact that data units are (usually) composed of *more than one* observation, measured through time and accompanied by covariates. The large amount of information *within* data units is reflected in our more complex regression model which allows for *correlated* random errors (§ 6.2.1). Various targets are possible for distortion of type ① (data contamination) and type ② (model deviation), both defined with reference to REALITY. For convenience, we chose one which involves the model (6.4) for a correlation structure *within* data units already implemented in the OSWALD software (Smith *et al.*, 1996). More precisely, distortion is considered to affect the measurement error component Z_{ij} and be quantified by the *discrepancy magnitude* $\alpha \in [0, 1]$ in (6.12).

The kind of distortion considered is not a 'classical' example in the robustness literature. In many applications the approach is to *include* a measurement error component into the model assumptions and a conflict between DATA and ASSUMPTIONS does not arise (at least not in this respect). Our choice was motivated by a different thought: The assumption of a measurement error component is literally directed to the *effective* data-generating process represented by DATA. At first sight, this seems to contradict our claim that model assumptions (within data units) should address the *real-world* (§ 3.2). However, the additional assumption described by Z_{ij} implies a *generalization* of $A = 0$ and as such does *not* cause a conflict between REALITY and ASSUMPTIONS. Thus, which of the two model assumptions $A = 0$ and $A = +$ give advantages in terms of *inferential performance*? – The simulation study based on location-closeness indicates in most cases that there is no preference, no matter whether the measurement errors are real or not. It remains an open question whether closeness in the meaning of spread would give a different answer.

Consideration has been given to the REML-estimation *method* from which several individual estimators are derived. The two main ones are the compound curve estimators for the variogram and the mean response profile (aspects of interest). Inferential performance assessment has again been carried out with reference to REALITY. However, a difference in performance under data contamination and model deviation could only be studied when the corresponding estimand was related to distortion. This applied to the variogram estimator, and to $\hat{\sigma}$ and $\hat{\nu}$.

Influence graphs studied changes in performance for the *simpler* assumptions $A = 0$. An investigation with $A = +$ might also be of interest, but has been omitted for conciseness. Preference graphs compared the performance of the REML-method in support of the various estimators under $A = 0$ and $A = +$. Contrastingly to chapter 5, emphasis was given to the comparison of *model assumptions* rather than individual estimators (which are defined differently but called the same under $A = 0$ and $A = +$). A further line of research could compare the performance of the REML- and the ML-method, both based on the same model assumptions (compare with § 4.3.2.1 and § 4.3.2.2).

Performance has been assessed in *absolute* as well as *relative* terms. Still, the objective joint interpretation of preference graphs, based on differences or quotients, at times gave difficulties. Q-preference graphs of the variogram estimator indicated advantages for $A = 0$ under data contamination when n is small, but due to high variability D-preference graphs did not (for a further example compare the Figures 6.13, 6.14, C.17, and C.18). We tried to avoid the problem by (only) reporting global impressions.

Chapter 7

Conclusions

7.1 Statistical inference and the model triplet

The present work has touched on one of the most fundamental issues in statistics. What is the aim of statistical inference? We claim it to be *real-world* description, instead of just data description. The real-world is always correct and reflects the ultimate truth, whereas the data could be wrong (§ 3.2). Chapter 3 gave several examples of *real-world situations* in which some *aspect of interest* could be described by statistical inference. An important consequence of our claim soon became clear and was explored in this thesis: The two sources of information available for statistical inference, the data and model assumptions, should somehow be validated with respect to the real-world. Hence we are confronted with a *triple* relation of data, model assumptions, and real-world where the latter is regarded as central (§ 3.2).

The way to the real-world description has been illustrated by the inference framework (§ 3.3.2). Model assumptions reflect our prior knowledge in the form of a statistical model and *inference procedures*, which process the data, give information about the unknown model parameters (seen in the widest sense). The aspect of interest from the real-world can then be described through one inference procedure alone, or through one or more inference

procedures in combination with the model assumptions. A non-parametric density estimator would be an example for the former, while the parametric *compound* estimator of the variogram in chapter 6 is typical for the latter. The semi-parametric ACL-estimator of chapter 5 could also be seen in the second way, even though here the individual 'sub-estimators' were not of interest. Overall, the model assumptions have been assigned to the following (optional) tasks:

- give a formal representation of the aspect of interest, suggest the kind(s) of inference procedure to be used, and directly contribute to the real-world description for *nominal* inference, and
- propose statistical (performance) properties of the inference procedures involved for *stochastic* inference.

Note that we restricted the illustration to the problem of classical inference (sampling theory) and in particular to estimation.

The *formal comparison* of real-world, data, and model assumptions was enabled through the idea of the model triplet REALITY, DATA, and ASSUMPTIONS. The statistical models respectively represent the ideal, effective, and assumed data-generating process (see the model framework in § 3.3.1). REALITY was considered to comprise the unknown *true model* which again was seen to exist in some 'objective reality' independently from any data-generating mechanism. Hence, statements such as "the operational 'true model' depends on the sample size" (Hampel *et al.*, 1986, p. 411) loose their accepted meaning in our approach. Even though REALITY and DATA are principally of infinite complexity, we restricted considerations to 'conventional' finite models. A statistical model for the data situation could be justified by regarding the generation of data as a repetitive event. The latter has to be seen in hypothetical terms when repetitions are not possible in practice, e.g. in observational studies. Note that strictly speaking, DATA and REALITY are only 'improper' statistical models. This is because their

parameters are considered to be fixed. We are neglecting this theoretical detail for convenience.

A data unit was defined as the observable outcome of a sampling unit (§ 3.3.1.1). The distinction between aspects of the model *within* and *between* data units (§ 3.3.1.2) was of vital importance for the following conclusion: Statistical inference requires correspondence of the model triplet between

1. DATA and REALITY relative to aspects *within* data units,
2. ASSUMPTIONS and REALITY relative to aspects *within* data units,
3. ASSUMPTIONS and DATA relative to aspects *between* data units.

Requirement 1 and 2 are necessary for the purpose of *nominal* inference, while item 3, together with 1 and 2, can be used for *stochastic* inference. Note already that the second purpose relates to an interesting controversial issue which will be discussed further below. At this point we first of all wish to focus on the implications of the above requirements.

7.2 Data contamination and model deviation

A well-known problem in statistical inference is the potential presence of *distortion*. It has already found extensive treatment in the literature, but with a variety of interpretations which may or may not mean the same thing (§ 2.3.1). An elementary classification of distortion, to our knowledge, did not exist so far and for the first time has been presented in this work (§ 3.4). Its main contribution is the *distinction between data contamination and model deviation*. In detail we defined (referring to the requirements on correspondence in the model triplet, see above)

- distortion of type ① as failure of requirement 1 (data contamination),
- distortion of type ② as failure of requirement 2 (model deviation), and

- distortion of type ③ as failure of requirement 3 (model deviation).

The distinction between data contamination and model deviation becomes most important for distortion types ① and ②. Both affect aspects of the model *within* data units. This could be a distributional specification as such (mostly studied in the literature) but may also apply to any dependence structure within vector observations (chapter 5) or between observations of the same data unit (chapter 6). Only what is actually *assumed* can be subject to model deviation, while data contamination is in principle always possible. Thus, highly informative parametric model assumptions together with data naturally give *more* possibilities to distinguish between data contamination and (type ②) model deviation than just simple non-parametric assumptions.

Going to the extreme, data contamination can be seen *independently* from the model assumptions while type ② model deviation still makes sense *without* the presence of data. This is because the reference point for DATA and ASSUMPTIONS (relative to aspects within data units) is represented by REALITY. Thus, in *theoretical* studies the model DATA could be formulated in any suitable way, independently from the model assumptions and just with reference to REALITY (if data contamination as such is the only interest). This is especially important in *non*-parametric statistics where the problem of data contamination is prevalent. Robustness studies, where the inference procedure is seen as a random variable with distribution derived from the model DATA, are therefore also possible in the latter area (and not just resistance studies where the inference procedure is seen as a function of the data independently from a statistical model, compare with Hettmansperger and Sheather (1992)).

A short review of (unclassified) distortion for some common statistical models has been given in § 2.3.3. Further in § 3.4.3, several examples discussed distortion of types ① to ③. They illustrated data contamination in the form of measurement, recording, and rounding errors, and due to inappropriate sampling units. They explained that model deviation can be caused by

approximation and misspecification, but not through generalization. They also indicated why aspects of the model *between* data units cannot be subject to data contamination. One particular example referred to the measurement problem of a (physical) constant. Elementary measurement errors, as part of the statistical error (§ 2.3.1), imply data contamination as an *unavoidable* consequence. The problem has been addressed in a more complex regression context in chapter 6. Finally, the term *pseudo data contamination* has been introduced for non-representative samples, which also reflect the statistical error but do not imply data contamination (§ 3.4.4).

Overall, the model triplet idea also has limitations. When a change in the DATA-model (within data units) can be due to *either* data contamination or controlled interference by the statistician, a reference to REALITY is no longer a clear case. This also applies to the corresponding parts of the ASSUMPTIONS. The particular choice of a censored survival/competing risks model for the example of chapter 5 could illustrate and at the same time avoid this problem. A censoring mechanism may be controlled by the statistician, but the model in the context of competing risks can perfectly be discussed using the model triplet approach.

7.3 Performance assessment under distortion

Distortion can affect the quality of the real-world description. This shows in the (bad) performance of individual inference procedures and could as well be due to direct contributions of wrong model assumptions (§ 4.2.1). To study the problem as a whole we generally considered *inference procedures*, possibly of compound structure.

Distortion was formalized and *quantified* by the idea of model expansion to study performance under *growing* distortion. It involved the implementation of an extra parameter (the discrepancy magnitude) or a parameter function (the discrepancy structure) into some ideal reference model. The former

was used in the longitudinal example of chapter 6, while the latter gave the novel description of a distorted Koziol-Green model in chapter 5. The generalization to discrepancy structures can first of all increase the dimension of potential distortion neighbourhoods and hence broaden the applicability of the approach. However, it is the discrepancy magnitude which ultimately specifies the distance from the reference model. Thus, distortion structures at the same distance need to be summarized again or consideration is given to just a few of them (as in § 5.3.2). The *comparison* of corresponding discrepancy magnitudes of REALITY, DATA, and ASSUMPTIONS finally determines the type and amount of distortion. See § 4.2.2, and for a general review of current formalization approaches of distortion also § 2.3.2.

Common interpretations of the performance quality of estimators have been classified in § 2.4. We concentrated on the attribute *closeness* (of the estimator distribution to the unknown estimand) and based performance descriptions on suitable *performance statistics*. A specific notation underlined their possible 'input' of REALITY, DATA, and ASSUMPTIONS (§ 4.2.3). *Influence graphs* were introduced which study the absolute or relative change of inferential (finite sample) performance under increasing distortion. They are determined by the comparison of expected performances in distorted and ideal (undistorted) situations for a given sample size. *Preference graphs* follow a related intention. They compare performance of alternative inference procedures and/or model assumptions under increasing distortion, this time through expected performance differences or ratios. Three types have been distinguished, as inference procedures and assumptions may or may not be directly linked to each other (§ 4.3).

The methodology of influence and preference graphs is applicable whenever distortion can be *quantified*. Each amount of distortion should however represent an 'easily accessible' (small) class of models. Hence, a class specified by a particular Lévy or Prohorov distance, or ϵ-contamination (gross-error model) seems to be rather general and less suitable for this purpose. From the robustness theory point of view, we considered quite

restricted distortion neighbourhoods (Hampel *et al.*, 1986, p. 9). Indeed, our aim was the interpretation of, and the study of inferential performance under, *particular forms of distortion*. It was not the development of universal robust procedures.

The model triplet REALITY, DATA, and ASSUMPTIONS as the reference frame for a performance statistic led to important conclusions (§ 4.2.1 and § 4.2.3): If performance is described as closeness in terms of

- *location*, the performance statistic refers to the aspect of interest in REALITY (the estimand). For theoretical studies the latter is considered to be *known*. The expected value of such a performance statistic is e.g. the bias of a point estimator.
- *spread*, the performance statistic does *not* refer to REALITY. The expected value of such a performance statistic is e.g. the variance of a point estimator.

Moreover, distortion refers to the aspect of interest (estimand) if the latter, seen as part of the model, is dependent on the discrepancy structure or magnitude. It especially means that the aspect of interest, as part of REALITY and relative to DATA, varies under data contamination and model deviation. This is the case for the unknown survival function of chapter 5, and becomes apparent for the parameters σ and ν and consequently for the variogram in chapter 6. Overall, a *difference* in performance under data contamination and type ② model deviation (if both are possible) can therefore be detected, if distortion relates to the aspect of interest, *and* assessment is (at least) carried out in terms of *location*-closeness. An estimator could be bias-robust under data contamination or under model deviation. The corresponding type of distortion, in any case, needs to be specified.

Having classified distortion and insisted on a reference to REALITY for performance in the sense of location-closeness, we critically reviewed some current approaches dealing with distortion (§ 4.4). We concluded that the more

classical approaches towards robustness refer to data contamination. These are qualitative and quantitative robustness, and the approach based on influence functions. Also, the accommodation approach for outliers as described in Barnett and Lewis (1995) seems to apply to data contamination. The recent robustness approach "configural polysampling" and the perturbation diagnostics by Cook (1986), on the other hand, seem to address type ② model deviation.

Chapters 5 and 6 presented simulation studies addressing performance in terms of location-closeness for two particular estimation examples. In both cases, distortion was considered to affect aspects of the model *within* data units. Thus, data contamination and type ② model deviation were principally possible (with an exception for the first example, however avoided in the simulation study – see above). Data for the two examples were generated according to DATA-models of growing discrepancy magnitude. The models REALITY and ASSUMPTIONS were chosen accordingly so that finally increasing distortion of the two types could be expressed. The first example represented a case of unwanted dependence *within* observations. The discrepancy structure used to explain this kind of distortion (§ 5.3.1) is applicable to any kind of multivariate data which involves a 0-1-variable. A generalization to scales of higher order might be possible, but would require further summaries within the distortion neighbourhood. The longitudinal example of chapter 6 addressed distortion affecting the dependence structure *between* observations of the same data unit. Distortion of this kind is especially a concern within the time series context. The particular (distorted) correlation structure of the example (§ 6.2.1 and § 6.3.1) applies to discrete as well as continuous time processes. As a linear regression model with dependent errors the example could further represent a rather complex parametric model which involves compound as well as individual estimators. It was important that distortion only affected *parts* of the model, here the distributional part but *not* the structural one.

The two studies revealed a *difference* in performance for the ACL-estimator

(chapter 5), the variogram estimator (chapter 6), and two individual parameter estimators (chapter 6). The results were not surprising, since *typical situations* have been addressed in which data contamination as well as type ② model deviation are possible and relate to the estimand, *and* simultaneously performance is assessed in terms of location-closeness. The mean-response estimator(s) and a third parameter estimator in chapter 6 moreover exemplified a case in which distortion does *not* relate to the estimand. As a natural consequence, performance could not be distinguished between data contamination and type ② model deviation.

The fact that data contamination *as well as* type ② model deviation are possible and relate to the aspect of interest occurs reasonably often. The examples chosen are representative for a wide range of estimation problems. First of all, data can always be generated with mistakes so that the true aspect of interest in the real-world is improperly reflected. Model deviation can 'easily' relate to the aspect of interest, when model assumptions *directly* contribute to the real-world description. This applies to the two compound estimators of the examples, but also to all other types of parametric or semi-parametric curve estimators. Further, this is relevant when a (parametric) model as a whole is of interest, e.g. an ARMA-model for prediction purposes. In addition, assumptions may give a wrong representation of the aspect of interest in terms of *model parameters*. Model assumptions not accounting for measurement errors ($A = 0$) in chapter 6 do *only* give a correct representation of the serial component variance σ^2, for example, if REALITY corresponds to the ideal reference model. Then σ^2 equals $\xi_1/(\xi_2 + 1)$. Otherwise, the correct representation of σ^2 involves the discrepancy magnitude α as a further parameter, which $A = 0$ obviously does not suggest. Another example would be the estimation of the mode under the assumption of a symmetric distribution with mean parameter (addressed by the approach of configural polysampling, § 4.4.6). Under the assumptions, the aspect of interest (the mode) is considered to be the mean parameter. This is however only correct as long as the true distribution in REALITY is symmetric.

Beside influence graphs, the two examples demonstrated the application of preference graphs. They analysed inferential performance in comparison with a competing estimator in chapter 5 and under alternative model assumptions in chapter 6. Some problems became apparent with the attempt to interpret those comparisons simultaneously in absolute and relative terms (chapter 6).

7.4 The duality of the model assumptions

The completion of this work allows an interesting conclusion about the objective of statistical model assumptions. Inferential performance descriptions in chapters 5 and 6 have been acquired through simulation. That is, the expected values of (differences or quotients of) performance statistics have been determined by averaging over DATA-generated samples. Similarly, one could try to derive the results on theoretical grounds (even though this might be difficult). In any way DATA is considered to be known, with fixed parameter values in the first case and maybe pending ones in the second. The situation is however different in *practical* circumstances. To anticipate potential performance behaviour of inference procedures (stochastic inference), we would additionally need to *assume* the DATA-model. The requirement 3 for model correspondence (page 158) does only serve for part of the purpose. Thus, we are confronted with some kind of duality-problem: The statistician has to assume the REALITY for *nominal* inference, and in addition the DATA for *stochastic* inference. Nevertheless, only one set of assumptions is ultimately formulated in statistical practice.

The problem becomes important when data contamination is suspected. The measurement error assumption in chapter 6 seems to be one good compromise to this problem (as the simulation results indicate). Unavoidable elementary measurement errors can be included in the ASSUMPTIONS through *generalization* which does not imply model deviation. Another obvious solution is to 'clean' the data in order to avoid data contamination. Of course, this is only possible within our limited (but hopefully correct)

knowledge about the truth in the real-world. Finally, one could indeed try to formulate two sets of model assumptions, one for the underlying real-world situation (referred to as ASSUMPTIONS in this thesis) and the other for the data. In other words, the first set is specified *before* data collection and is wholly based on prior knowledge, while the second is determined from the present data eventually accounting for knowledge gained from earlier (equally contaminated) data. We are aware that this sounds rather demanding, especially when thinking of how little prior information we often have available and that data-driven model formulation is in itself already associated with a more or less substantial amount of uncertainty (Chatfield, 1995).

Dawid (1983, p. 92) remarks that "conditioning on an ancillary does not change the estimate (nominal inference)". Thus, ancillary information as well as the dummy-part of the model assumptions (§ 3.3.2) do not need to be validated by REALITY and the duality-problem does not arise in this respect. Neither does the problem occur if after all statistical inference is used for data analysis in its *literal* meaning, i.e. information is exclusively desired about the data and *not* about any real-world situation. This case has *not* been considered in the present thesis, since it rather belongs to the rubric of descriptive statistics (see § 3.2).

7.5 Future perspectives

Some future perspectives are briefly listed below, either as statements or as questions:

- Any study dealing with the problem of distortion in statistical inference should indicate which type of distortion it refers to. This is especially important when data contamination and model deviation are simultaneously possible.

- The choice of robustness approach in a practical situation should be made in the judgement of which to fear more, data contamination or model deviation. That is, the statistician needs to decide whether it is more important for an inference procedure to be robust (in terms of location-closeness) against errors in the data or against errors in the model assumptions.

- Current approaches dealing with distortion seem to address exclusively *either* data contamination *or* model deviation (§ 4.4): If at all possible, how can these approaches be generalized to consider the two (three) types of distortion, each on its own or even both simultaneously?

- We restricted considerations to *classical* inference. What is distortion from the point of view of other schools of inference? Does the distinction between data contamination and model deviation have the same importance?

- Can we make a difference between data contamination and model deviation for missing data models?

- We have restricted considerations to (point) estimation problems. It is also of interest to study the potential difference of performance under data contamination and model deviation in the context of interval estimation and hypothesis testing, or even more generally in the decision-theoretic framework.

- The methodology of influence and preference graphs could be generalized to account for more than one performance criteria summarized by an 'appropriately' weighted average, or more than one kind of distortion in multiple dimensions.

- Influence and preference graphs compare inferential performance as expected in the long-term. Once evaluated, one could try to combine their information with some prior belief about the type (and amount) of

distortion for a *particular* data situation. The result could be expressed in form of a likelihood.

- The simulation studies in chapters 5 and 6 have been carried out under various restrictions. To what extent can their results be generalized? Are the results of ACL- and KM-performance in chapter 5 also valid for other discrepancy structures $\mathbf{s}(\cdot)$ and for other observed survival distributions F_Z? Would an alternative mean or error model in chapter 6 lead to different conclusions? What happens in situations of non-balanced data structures and irregularly spaced measurement times?

- The choice of discrepancy magnitudes γ and α in chapters 5 and 6 fail to measure the amount of distortion in a way which is *invariant* with respect to changes in scale (of γ and α). Hence, is it possible to formulate invariant measures of the amount of distortion in these examples? (Our choice of discrepancy magnitudes was mainly due to convenience. While $\alpha \in [0, 1]$ we preferred $\gamma \in [0, p(1-p)]$ to allow a comparison between different values of p.)

- We could extend the notion of distortion by relating to § 7.4. Then, a question which certainly has been addressed in the literature could find a new interpretation: How is the performance of a *performance statistic* affected under wrong assumptions of the DATA-model?

Appendix A

Proofs

A.1 Relationship to an alternative distorted Koziol-Green model

The relationship of modelling a distorted KG-model according to our $\mathbf{s}(\cdot)$-approach and the approach by Beirlant *et al.* (1992) has been formulated in § 5.3.3. The corresponding proof will now follow.

Show that the function $\mathbf{L}^*(\cdot)$ as formulated in (5.11) is slowly varying at the origin with $\mathbf{L}^*(1) = 1$: Assuming the existence of the densities f_X, f_Y and f_Z and taking into account (5.1) it is

$$\begin{aligned} f_Z(z) = f_{\min\{X,Y\}}(z) &= \frac{\partial}{\partial z}\left[1 - (1 - F_X(z)) \cdot (1 - F_Y(z))\right] \\ &= (1 - F_X(z)) \cdot f_Y(z) + (1 - F_Y(z)) \cdot f_X(z). \end{aligned}$$

With the first part of (5.6) it further follows that the equation

$$\lambda_Y(z) = \frac{1 - \mathbf{s}(z)}{\mathbf{s}(z)} \cdot \lambda_X(z)$$

holds for $z \in [0, \infty)$ and $\mathbf{s}(z) > 0$. Therefore

$$S_Y(z) = \exp\left[-\int_0^z \frac{1 - \mathbf{s}(t)}{\mathbf{s}(t)} \cdot \lambda_X(t)\ dt\right] = \left[S_X(z)\right]^\theta \cdot \mathbf{L}^*\left[S_X(z)\right],$$

where $\theta = (1 - \mathbf{s}_\infty)/\mathbf{s}_\infty$ with $s_\infty = \lim_{z\to\infty} \mathbf{s}(z) > 0$ and the function $\mathbf{L}^*(\cdot)$ as defined in (5.11).

Since it is obvious that $\mathbf{L}^*(1) = 1$, it only needs to be shown that for any $\vartheta > 0$

$$\lim_{u\to 0} \frac{\mathbf{L}^*(\vartheta \cdot u)}{\mathbf{L}^*(u)} = 1$$

(Beirlant *et al.*, 1992, p. 27).

Consider the case $0 < \vartheta \leq 1$ (the proof for $\vartheta \geq 1$ is analogous):

Because $\mathbf{s}(z)$ is assumed to converge for large z, some $z^* \in (0, \infty)$ exists, so that $[1 - \mathbf{s}(z)]/\mathbf{s}(z)$ is bounded for $z \geq z^*$. Knowing that $f_X(z)/S_X(z) \geq 0$ for $z \in [0, \infty)$ and assuming that $\mathbf{s}(\cdot)$ is continuous within $[z^*, \infty)$, a general version of the mean value theorem for integrals can then be applied, which ensures the existence of some ξ_u with $z^* \leq S^{-1}(u) \leq \xi_u \leq S^{-1}(\vartheta \cdot u)$, so that

$$\begin{aligned} \int_{S^{-1}(u)}^{S^{-1}(\vartheta\cdot u)} \frac{1 - \mathbf{s}(z)}{\mathbf{s}(z)} \cdot \frac{f_X(z)}{S_X(z)}\, dz &= \frac{1 - \mathbf{s}(\xi_u)}{\mathbf{s}(\xi_u)} \cdot \int_{S^{-1}(u)}^{S^{-1}(\vartheta\cdot u)} \frac{f_X(z)}{S_X(z)}\, dz \\ &= \frac{1 - \mathbf{s}(\xi_u)}{\mathbf{s}(\xi_u)} \cdot (-1) \cdot \ln \vartheta. \end{aligned} \tag{A.1}$$

Now, it is

$$\begin{aligned} \lim_{u\to 0} \frac{\mathbf{L}^*(\vartheta \cdot u)}{\mathbf{L}^*(u)} &= \lim_{u\to 0} \left\{ \exp\left[\frac{\mathbf{s}_\infty - 1}{\mathbf{s}_\infty} \cdot \ln \vartheta - \int_{S^{-1}(u)}^{S^{-1}(\vartheta\cdot u)} \frac{1 - \mathbf{s}(z)}{\mathbf{s}(z)} \cdot \frac{f_X(z)}{S_X(z)}\, dz \right] \right\} \\ &\overset{(A.1)}{=} \exp\left[\frac{\mathbf{s}_\infty - 1}{\mathbf{s}_\infty} \cdot \ln \vartheta - \lim_{u\to 0} \frac{\mathbf{s}(\xi_u) - 1}{\mathbf{s}(\xi_u)} \cdot \ln \vartheta \right] \\ &= \exp(0) = 1, \end{aligned}$$

since $\lim_{u\to 0} \xi_u = \infty$. □

A.2 Expression for the longitudinal error-matrix

Show that the $(m \times n)$–error matrix of a single longitudinal data set can be simulated according to (6.15): The relations in (6.7) and (6.8) allow to describe the variance of an arbitrary row $\boldsymbol{\epsilon}_i' = (\epsilon_{i1}, \ldots, \epsilon_{in})$ of the matrix according to

$$\mathrm{Var}(\boldsymbol{\epsilon}_i') = \mathrm{Var}(\boldsymbol{\epsilon}_i) = \mathrm{Var}(\boldsymbol{\epsilon}) = \nu^2 J + \sigma^2 H + \tau^2 I$$

where J is the $(n \times n)$–unity matrix, I the $(n \times n)$–identity matrix, and

$$H = \Big(\rho_W(|j-k|)\Big)_{j,k=1,\ldots,n}$$

with $\rho_W(u) = \exp(-\phi u)$. The subscript i of $\boldsymbol{\epsilon}_i$ can be dropped since errors from different data units are considered to be i.i.d. (Diggle *et al.*, 1995, p. 81, 87).

The problem of simulating the components of the matrix ϵ can therefore be reformulated by setting $\boldsymbol{\epsilon} = W' \cdot \boldsymbol{Z}$, where $\boldsymbol{Z}$ is a column-vector of n independent $N(0,1)$-distributed random variables and W an upper-triangular $(n \times n)$–matrix of coefficients. The latter is evaluated by a Choleski decomposition as $W = \mathrm{chol}\,[\mathrm{Var}(\boldsymbol{\epsilon})]$, since

$$\begin{aligned} \mathrm{Var}(\boldsymbol{\epsilon}) &= \mathrm{Var}(W' \cdot \boldsymbol{Z}) \\ &= W' \cdot \mathrm{Var}(\boldsymbol{Z}) \cdot W = W' \cdot W \end{aligned}$$

and

$$\mathrm{chol}(W' \cdot W) = W.$$

Stacking independent and like $\boldsymbol{Z}$ distributed random vectors $\boldsymbol{Z}_i$ one obtains a $(m \times n)$–matrix, so that finally

$$\epsilon = \begin{pmatrix} \boldsymbol{\epsilon}_1' \\ \vdots \\ \boldsymbol{\epsilon}_m' \end{pmatrix} = \begin{pmatrix} \boldsymbol{Z}_1' \\ \vdots \\ \boldsymbol{Z}_m' \end{pmatrix} \cdot W.$$

□

Appendix B

Some programming details for chapter 5

Specification of a discrepancy structure

A suitable discrepancy structure $\mathbf{s}(\cdot)$ from the class EXP given F_Z, p, γ, and the sign of a is determined as follows:

1. Evaluation of c which produces a maximum discrepancy γ^* explained by EXP given F_Z, p and the sign of a:

   - A sufficient condition for maximizing $|\gamma(z)|$ according to (5.7) for all $z \in [0, \infty)$ with $f_Z(z) > 0$ is given by $(\mathbf{s}(z) - p) \cdot f_Z(z) = 0$. The latter is satisfied for

     $$z^* = \frac{\log\left[(p-d)/a\right]}{-c}.$$

     Since $\mathbf{s}(\cdot)$ is strictly monotonic, not more than one local maximum exists.

   - Maximize the discrepancy magnitude

     $$\gamma = u(a, c, d) = \left| \int_0^{z^*} \left[a \cdot \exp(-c \cdot t) + d - p\right] \, dF_Z(t) \right|$$

with z^* as above over all possible a, c and d (given F_Z, p and the sign of a). Each $u(a, c, d)$ is evaluated by numerical integration, and maximization is carried out by an evolutionary strategy. The resulting maximum discrepancies are presented in Table 5.1.

2. Note the corresponding value of c.

3. Evaluation of a and d (given a value for c as above and given F_Z, p, γ, and the sign of a):

   Solve the following optimization problem over a and d by an evolutionary strategy:

   - Evaluate by numerical integration

     $$\text{result } 1 = |p(a, d) - p|\,,$$

     where $p(a, d) = P(\Delta = 1)$ is calculated using (5.8) under some *preliminarily* specified a and d.

   - Evaluate by numerical integration

     $$\text{result } 2 = |\gamma(a, d) - \gamma|\,,$$

     where $\gamma(a, d)$ is the discrepancy magnitude under some *preliminarily* specified a and d.

   - Minimize: max(result 1, result 2) over a and d.

4. Note the corresponding values for a and d.

Evaluation of the true survival function

Under model deviation, the true survival function S_X of interest is derived according to the relationship (5.5), since in this case REALITY corresponds to a distorted KG-model. Depending on whether the respective discrepancy structures are taken from STEP or EXP, S_X is then evaluated as follows:

STEP The particular form of $\mathbf{s}(\cdot)$ in this class simplifies (5.5) to

$$S_X(z) = \begin{cases} \left[S_Z(z)\right]^a & \text{for} \quad z \leq z^* \\ \\ \left[S_Z(z^*)\right]^{(a-b)} \cdot \left[S_Z(z)\right]^b & \text{for} \quad z > z^*. \end{cases}$$

The above can be directly implemented.

EXP The product-integral notation for (5.5)

$$S_X(z) = \prod_0^z \left[1 - \mathbf{s}(t) \cdot \lambda_Z(t)\, dt\right]$$

(compare with Andersen *et al.* (1993), p. 49 and 89f) motivates the use of the following finite sample counterpart given as the finite product

$$\begin{aligned} \hat{S}_X(z) &= \prod_{j:z_j<z} \left[1 - \mathbf{s}(z_j) \cdot \lambda_Z(z_j)\right] \\ &= \prod_{z:z_j<z} \left[1 - \mathbf{s}(z_j) \cdot \left(1 - \frac{1 - F_Z(z_{j+1})}{1 - F_Z(z_j)}\right)\right]. \end{aligned} \tag{B.1}$$

The term (B.1) based on formula (1.2.7) in Lawless (1982, p. 10) is finally implemented in order to avoid expressions with the density f_Z.

Appendix C

Additional graphics for chapter 6

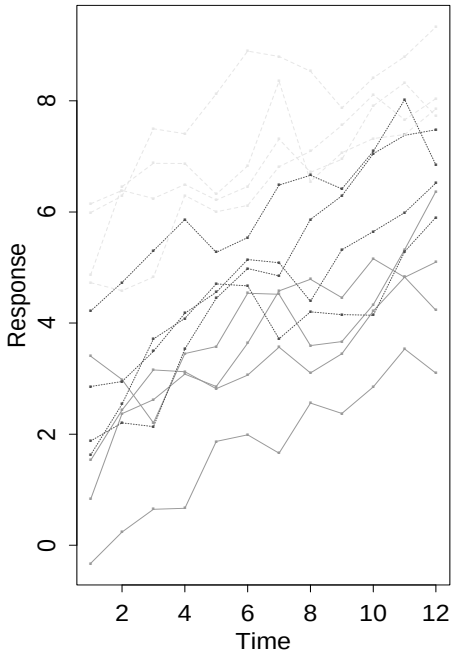

Figure C.1: *Simulated responses for the three treatment groups ($\alpha = 0$).*

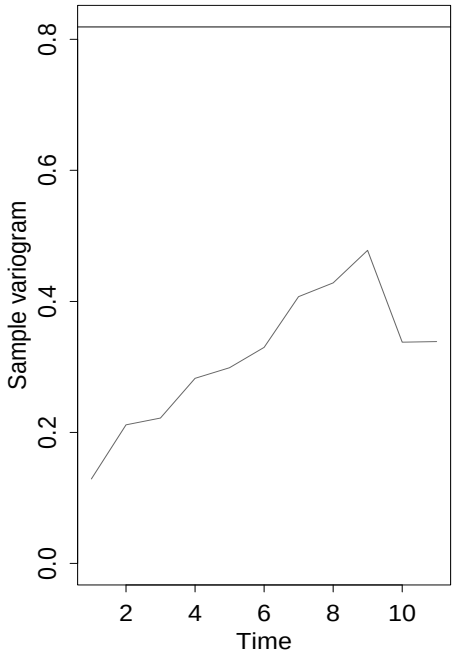

Figure C.2: *Sample variogram with estimated process variance ($\alpha = 0$).*

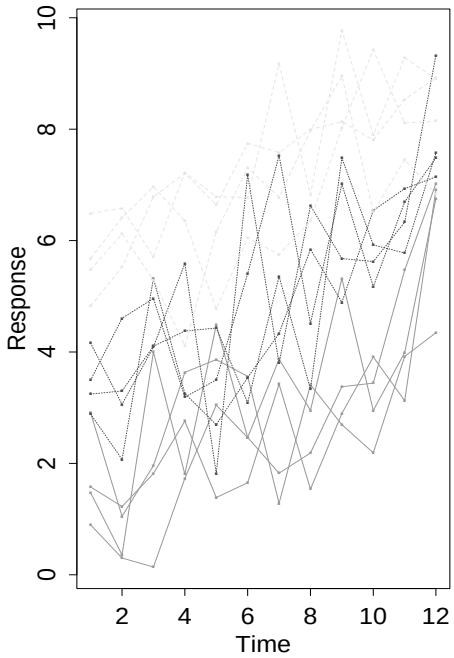

Figure C.3: *Simulated responses for the three treatment groups ($\alpha = 1$).*

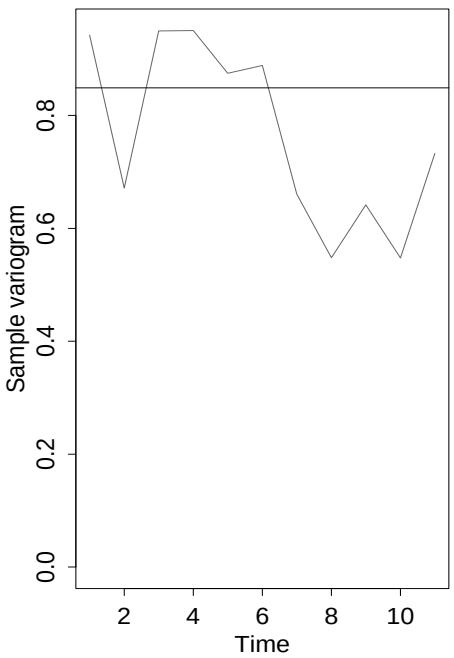

Figure C.4: *Sample variogram with estimated process variance ($\alpha = 1$).*

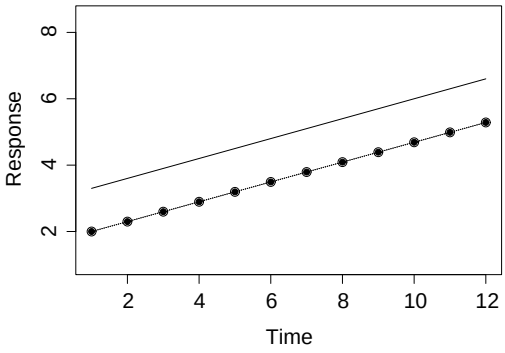

Figure C.5: *Mean response estimates (group 2) under no distortion* $\alpha = 0$.

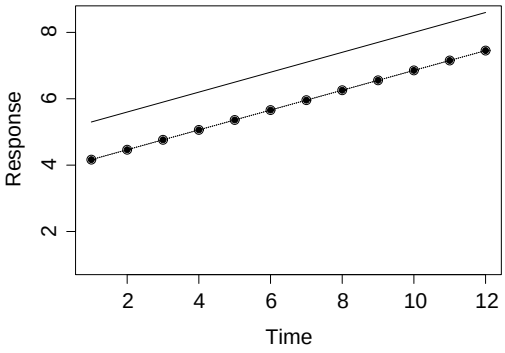

Figure C.6: *Mean response estimates (group 3) under no distortion* $\alpha = 0$.

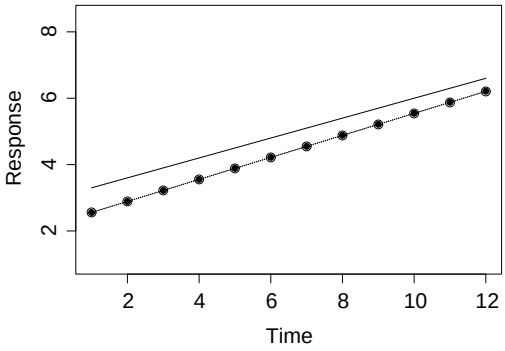

Figure C.7: *Mean response estimates (group 2) under distortion* $\alpha = 0.5$.

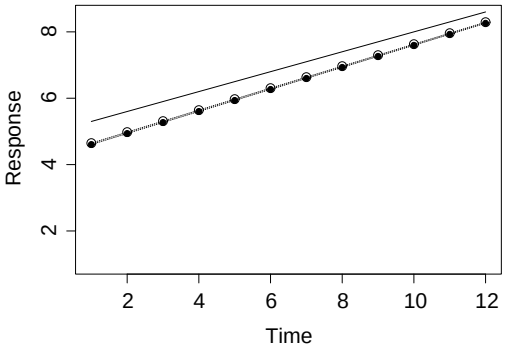

Figure C.8: *Mean response estimates (group 3) under distortion* $\alpha = 0.5$.

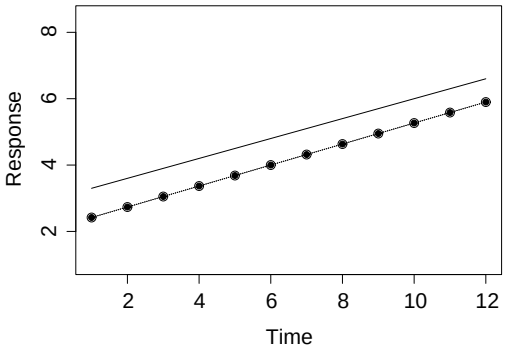

Figure C.9: *Mean response estimates (group 2) under distortion* $\alpha = 1$.

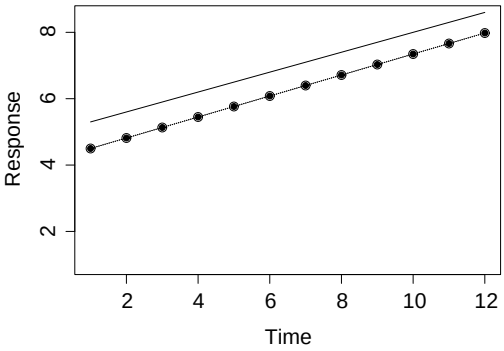

Figure C.10: *Mean response estim. (group 3) under distortion* $\alpha = 1$.

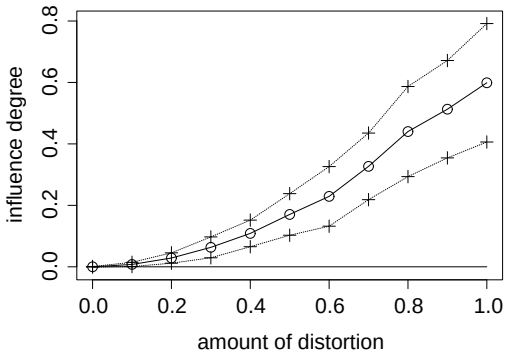

Figure C.11: *D-influence graph for data contamination based on* $\pi_{d(1)}$.

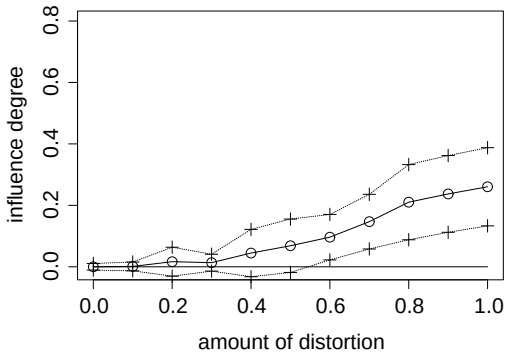

Figure C.12: *D-influence graph for data contamination based on* $\pi_{d(7)}$.

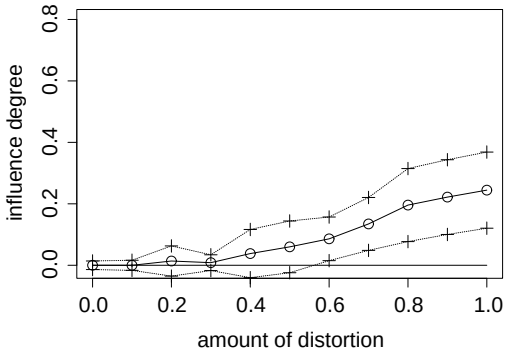

Figure C.13: *D-influence graph for data contamination based on* $\pi_{d(12)}$.

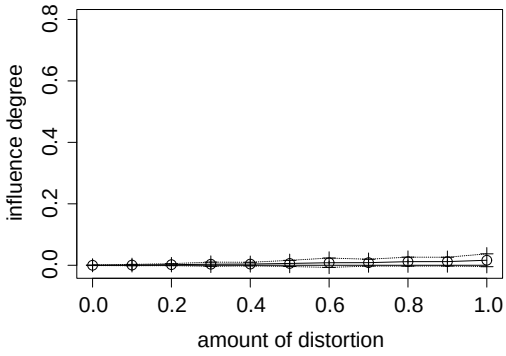

Figure C.14: *D-influence graph for model deviation based on* $\pi_{d(1)}$.

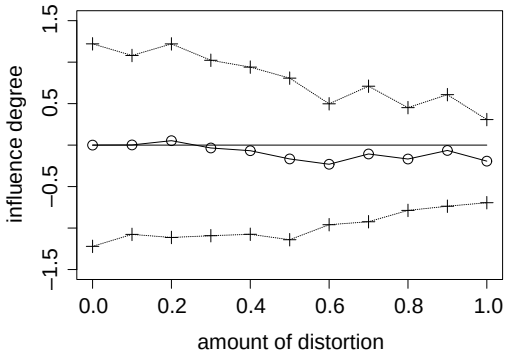

Figure C.15: *D-influence graph based on* $\pi_{Y(2)}$.

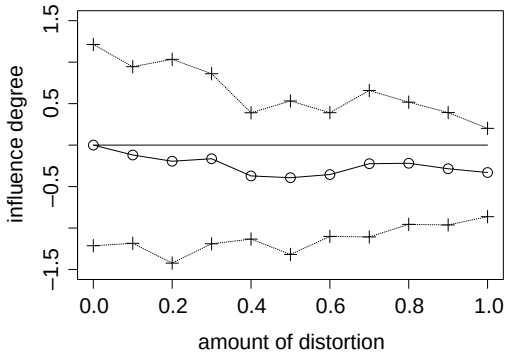

Figure C.16: *D-influence graph based on* $\pi_{Y(3)}$.

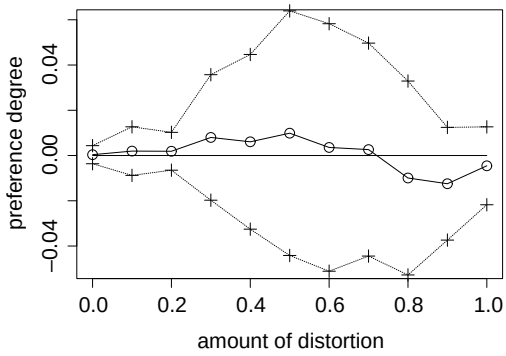

Figure C.17: *D-preference graph for data contamination based on* $\pi_{\bar{d}}$.

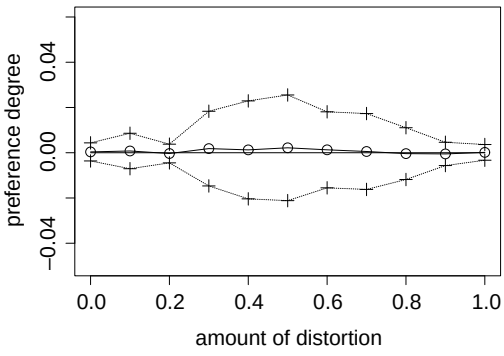

Figure C.18: *D-preference graph for model deviation based on* $\pi_{\bar{d}}$.

References

Abdushukurov, A. A. (1984). On some estimates of the distribution function under random censorship. In *Conference of Young Scientists*, Tashkent, Uzbek SSR. Math. Inst. Acad. Sci. VINITI No. 8756-V, in Russian.

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* Springer Series in Statistics. Springer Verlag, New York.

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey, J. W. (1972). *Robust Estimates of Location: Survey and Advances.* Princeton University Press, Princeton.

Barnard, G. A. (1971). Discussion of the paper by Professor Godambe and Dr. Thompson. *Journal of the Royal Statistical Society, Series B*, 33:376–378.

Barnard, G. A. (1980). Discussion of Professor Box's paper. *Journal of the Royal Statistical Society, Series A*, 143(4):404–406.

Barnett, V. (1982). *Comparative Statistical Inference.* John Wiley & Sons, New York, 2nd edition.

Barnett, V. and Lewis, T. (1995). *Outliers in Statistical Data.* John Wiley & Sons, New York, 3rd edition.

Beirlant, J., Carbonez, A., and van der Meulen, E. (1992). Long run proportional hazards models of random censorship. *Journal of Statistical Planning and Inference*, 32:25–44.

Billor, N. and Loynes, R. M. (1993). Local influence: A new approach. *Communications in Statistics – Theory and Methods*, 22(6):1595–1611.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society, Series A*, 143(4):383–430. with discussion.

Cabrera, J., Maguluri, G., and Singh, K. (1997). Indices of empirical robustness. *Statistics & Probability Letters*, 33:49–62.

Chang, M. N. (1996). Exact distribution of the Kaplan-Meier estimator under the proportional hazards model. *Statistics & Probability Letters*, 28:153–157.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society, Series A*, 158(3):419–466. with discussion.

Chen, Y. Y., Hollander, M., and Langberg, N. A. (1982). Small-sample results for the Kaplan-Meier estimator. *Journal of the American Statistical Association*, 77(377):141–144.

Cheng, P. E. and Lin, G. D. (1984). Maximum likelihood estimation of a survival function under the Koziol-Green proportional hazards model. Technical Report B-84-5, Institute of Statistics, Academia Sinica, Taipei, Taiwan.

Cheng, P. E. and Lin, G. D. (1987). Maximum likelihood estimation of a survival function under the Koziol-Green proportional hazards model. *Statistics & Probability Letters*, 5:75–80.

Collins, J. R. and Wu, B. (1998). Comparisons of asymptotic biases and variances of M-estimators of scale under asymmetric contamination. *Communications in Statistics – Theory and Methods*, 27(7):1791–1810.

Cook, R. D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society, Series B*, 48(2):133–169. with discussion.

Copas, J. B. (1988). Binary regression models for contaminated data. *Journal of the Royal Statistical Society, Series B*, 50(2):225–265. with discussion.

Copas, J. B. and Stride, C. B. (1997). Fitting a normal distribution when the model is wrong. *Annals of the Institute of Statistical Mathematics*, 49(4):601–614.

Cox, D. R. (1978). Foundations of statistical inference: The case for electicism. *Australian Journal of Statistics*, 20(1):43–59.

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

Csörgő, S. (1988). Estimation in the proportional hazards model of random censorship. In *Mathematische Operationsforschung und Statistik*, volume 19 of *Statistics*, pages 437–463. Akademie-Verlag.

Csörgő, S. (1998). Testing for the partial proportional hazards model of random censorship. In Hušková, M., Lachout, P., and Víšek, J. A., editors, *Prague Stochastics '98 Proceedings*, pages 87–92, Prague. Union of Czech Mathematicians and Physicists.

Csörgő, S. and Faraway, J. J. (1998). The paradoxical nature of the proportional hazards model of random censorship. *Statistics*, 31:67–78.

Dawid, A. P. (1983). Inference, statistical: I. In Kotz, S. and Johnson, N. L., editors, *Encyclopedia of Statistical Sciences*, volume 4, pages 89–105. John Wiley & Sons, New York.

Diggle, P. J. (1988). An approach to the analysis of repeated measurements. *Biometrics*, 44:959–971.

Diggle, P. J. (1990). *Time Series: A Biostatistical Introduction.* Oxford University Press, Oxford.

Diggle, P. J., Liang, K.-Y., and Zeger, S. L. (1995). *Analysis of Longitudinal Data.* Number 13 in Oxford Statistical Science Series. Oxford University Press, Oxford. reprint with corrections.

Dikta, G. (1995). Asymptotic normality under the Koziol-Green model. *Communications in Statistics – Theory and Methods*, 24(6):1537–1549.

Donoho, D. L. and Liu, R. C. (1988). The "automatic" robustness of minimum distance functionals. *The Annals of Statistics*, 16(2):552–586.

Draper, D. and Parmigiani, G., editors (1995). *Workshop on Model Uncertainty and Model Robustness*, Bath, England. electronic proceedings with abstracts of presentations and links to relevant papers on the Web at http://www.isds.duke.edu/conferences/bath/abstracts.html.

Ebrahimi, N. (1985). Nonparametric estimation of survival functions for incomplete observations when the life time distribution is proportionally related to the censoring time distribution. *Communications in Statistics – Theory and Methods*, 14(12):2887–2898.

Ebrahimi, N. and Habibullah, M. (1992). Testing to determine the underlying distribution using incomplete observations when the life time distribution is proportionally related to the censoring time distribution. *Journal of Statistical Computation and Simulation*, 40:109–118.

Ebrahimi, N. and Kirmani, S. N. U. A. (1996). A characterization of the proportional hazards model through a measure of discrimination between two residual life distributions. *Biometrika*, 83(1):233–235.

Eguchi, S. and Copas, J. (1998). A class of local likelihood methods and near-parametric asymptotics. *Journal of the Royal Statistical Society, Series B*, 60(4):709–724.

Fahrmeier, L., Kaufmann, H. L., and Ost, F. (1981). *Stochastische Prozesse. Eine Einführung in Theorie und Anwendungen.* Carl Hanser Verlag, Munich, Vienna.

Fernholz, L. T. (1983). *von Mises Calculus for Statistical Functions*, volume 19 of *Lecture Notes in Statistics.* John Wiley & Sons, New York.

Fraser, D. A. S. (1983). Inference, statistical: II. In Kotz, S. and Johnson, N. L., editors, *Encyclopedia of Statistical Sciences*, volume 4, pages 105–114. John Wiley & Sons, New York.

Gather, U. and Pawlitschko, J. (1998). On Efron's and Gill's version of the Kaplan-Meier integral. *Communications in Statistics – Theory and Methods*, 27(1):181–192.

Genton, M. G. (1998). Asymptotic variance of M-estimators for dependent Gaussian random variables. *Statistics & Probability Letters*, 38:255–261.

Ghorai, J. K. (1991a). Cramér-von Mises statistic for testing goodness of fit under the proportional hazards model. *Communications in Statistics – Theory and Methods*, 20(3):1107–1126.

Ghorai, J. K. (1991b). Estimation of a smooth quantile function under the proportional hazards model. *Annals of the Institute of Statistical Mathematics*, 43(4):747–760.

Ghorai, J. K. and Pattanaik, L. M. (1993). Asymptotically optimal bandwidth selection of the kernel density estimator under the proportional hazards model. *Communications in Statistics – Theory and Methods*, 22(5):1383–1401.

Gijbels, I. and Veraverbeke, N. (1989). Quantile estimation in the proportional hazards model of random censorship. *Communications in Statistics – Theory and Methods*, 18(5):1645–1663.

Godambe, V. P. and Thompson, M. E. (1984). Robust estimation through estimating equations. *Biometrika*, 71(1):115–125.

Graybill, F. A. (1961). *An Introduction to Linear Statistical Models, Volume 1.* McGraw-Hill Series in Probability and Statistics. McGraw-Hill Book Company, Singapore.

Gronen, S. (1993). Der Harrell–Davis–Quantilschätzer. Diplomarbeit, Department of Statistics, University of Dortmund, Dortmund, Germany.

Grunert, V. (1993). Der Kaigh–Lachenbruch–Quantilschätzer im unzensierten und im zensierten Fall. Diplomarbeit, Department of Statistics, University of Dortmund, Dortmund, Germany.

Hampel, F. R. (1971). A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, 42(6):1887–1896.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics. The Approach Based on Influence Functions.* Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York.

Hartung, J. (1995). *Statistik. Lehr- und Handbuch der angewandten Statistik.* R. Oldenbourg Verlag, Munich, Vienna, 10th edition.

He, X. and Simpson, D. G. (1993). Lower bounds for contamination bias: globally minimax versus locally linear estimation. *The Annals of Statistics*, 21(1):314–337.

Herbst, T. (1992a). Estimation of moments under Koziol-Green model for random censorship. *Communications in Statistics – Theory and Methods*, 21(3):613–624.

Herbst, T. (1992b). Test of fit with the Koziol-Green model for random censorship. *Statistics & Decisions*, 10:163–171.

Hettmansperger, T. P. and Sheather, S. J. (1992). Resistant and robust procedures. In Hoaglin, D. C. and Moore, D. S., editors, *Perspectives on Contemporary Statistics*, number 21 in MAA Notes, pages 145–170. Mathematical Association of America.

Huber, P. J. (1981). *Robust Statistics*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York.

Huber, P. J. (1991). Between robustness and diagnostics. In Stahel, W. and Weisberg, S., editors, *Directions in Robust Statistics and Diagnostics, Part I*, volume 33 of *The IMA Volumes in Mathematics and its Applications*, pages 121–130. Springer Verlag, New York.

Hyde, J. (1977). Testing survival under right censoring and left truncation. *Biometrika*, 64(2):225–230.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Koziol, J. A. and Green, S. B. (1976). A Cramér-von Mises statistic for randomly censored data. *Biometrika*, 63(3):465–474.

Künsch, H., Beran, J., and Hampel, F. (1993). Contrasts under long-range correlations. *The Annals of Statistics*, 21(2):943–964.

Künsch, H. R. (1991). Dependence among observations: Consequences and methods to deal with it. In Stahel, W. and Weisberg, S., editors, *Directions in Robust Statistics and Diagnostics, Part I*, volume 33 of *The IMA Volumes in Mathematics and its Applications*, pages 131–140. Springer Verlag, New York.

Lange, K. L., Little, R. J. A., and Taylor, J. M. G. (1989). Robust statistical modeling using the t-distribution. *Journal of the American Statistical Association*, 84(408):881–896.

Lawless, J. F. (1982). *Statistical Models and Methods for Lifetime Data*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York.

Lawrance, A. J. (1991). Local and deletion influence. In Stahel, W. and Weisberg, S., editors, *Directions in Robust Statistics and Diagnostics, Part I*, volume 33 of *The IMA Volumes in Mathematics and its Applications*, pages 141–157. Springer Verlag, New York.

Ledolter, J. (1991). Outliers in times series analysis: Some comments on their impact and their detection. In Stahel, W. and Weisberg, S., editors, *Directions in Robust Statistics and Diagnostics, Part I*, volume 33 of *The IMA Volumes in Mathematics and its Applications*, pages 159–165. Springer Verlag, New York.

Lucas, A. (1997). Asymptotic robustness of least median of squares for autoregressions with additive outliers. *Communications in Statistics – Theory and Methods*, 26(10):2363–2380.

Meintanis, S. G. and Donatos, G. S. (1997). A comparative study of some robust methods for coefficient-estimation in linear regression. *Computational Statistics & Data Analysis*, 23:525–540.

Mi, J. (1990). Asymptotic results for the Koziol-Green model. *Communications in Statistics – Theory and Methods*, 19(8):2767–2779.

Millar, P. W. (1981). Robust estimation via minimum distance methods. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 55:73–89.

Miller, R. G. (1981). *Survival Analysis*. Wiley Series in Applied Probability and Statistics. John Wiley & Sons, New York.

Mood, A. M., Graybill, F. A., and Boes, D. C. (1974). *Introduction to the Theory of Statistics.* McGraw-Hill Series in Probability and Statistics. McGraw-Hill Book Company, Singapore, 3rd edition.

Morgenthaler, S. (1991). Configural polysampling. In Stahel, W. and Weisberg, S., editors, *Directions in Robust Statistics and Diagnostics, Part II*, volume 34 of *The IMA Volumes in Mathematics and its Applications*, pages 49–63. Springer Verlag, New York.

Morgenthaler, S. (1994). Small sample efficiency and exact fit for Cauchy regression models. *Statistics & Probability Letters*, 19:381–385.

Morgenthaler, S. and Tukey, J. W., editors (1991). *Configural Polysampling. A Route to Practical Robustness.* Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York.

Nanayakkara, N. and Cressie, N. (1991). Robustness to unequal scale and other departures from the classical linear model. In Stahel, W. and Weisberg, S., editors, *Directions in Robust Statistics and Diagnostics, Part II*, volume 34 of *The IMA Volumes in Mathematics and its Applications*, pages 65–113. Springer Verlag, New York.

Pantula, S. G. and Pollock, K. H. (1985). Nested analysis of variance with autocorrelated errors. *Biometrics*, 41:909–920.

Peña, E. A. and Rohatgi, V. K. (1989). Survival function estimation for a generalized proportional hazards model of random censorship. *Journal of Statistical Planning and Inference*, 22:371–389.

Portnoy, S. L. (1977). Robust estimation in dependent situations. *The Annals of Statistics*, 5:22–43.

Portnoy, S. L. (1979). Further remarks on robust estimation in dependent situations. *The Annals of Statistics*, 7:224–231.

Rieder, H. (1994). *Robust asymptotic statistics*. Springer Series in Statistics. Springer Verlag, New York.

Ronchetti, E. (1997). Robust inference by influence functions. *Journal of Statistical Planning and Inference*, 57:59–72.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust regression and outlier detection*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York.

Smith, D. M., Robertson, W. H., and Diggle, P. J. (1996). Oswald: Object-oriented software for the analysis of longitudinal data in S. Technical Report MA96/192, Department of Mathematics and Statistics, Lancaster University, Lancaster, LA1 4YF, U.K.

Stahel, W. and Weisberg, S., editors (1991). *Directions in Robust Statistics and Diagnostics*, volume 33 and 34 of *The IMA Volumes in Mathematics and its Applications*. Springer Verlag, New York. Part I and II.

Stahel, W. A. (1991). Research directions in robust statistics. In Stahel, W. and Weisberg, S., editors, *Directions in Robust Statistics and Diagnostics, Part II*, volume 34 of *The IMA Volumes in Mathematics and its Applications*, pages 243–278. Springer Verlag, New York.

Stute, W. (1992). Strong consistency under the Koziol-Green model. *Statistics & Probability Letters*, 14:313–320.

Taplin, R. H. (1993). Robust likelihood calculation for time series. *Journal of the Royal Statistical Society, Series B*, 55(4):829–836.

Taylor, J. M. G. (1992). Properties of modelling the error distribution with an extra shape parameter. *Computational Statistics & Data Analysis*, 13:33–46.

Tukey, J. (1997). More honest foundations for data analysis. *Journal of Statistical Planning and Inference*, 57:21–28.

von Mises, R. (1947). On the asymptotic distribution of differentiable statistical functions. *The Annals of Mathematical Statistics*, 18:309–348.

Wiens, D. P., Wu, E. K. H., and Zhou, J. (1998). On the trimmed mean and minimax-variance L-estimation in Kolmogorov neighbourhoods. *Canadian Journal of Statistics–Revue*, 26(2):231–238.

Wiens, D. P. and Zhou, J. (1997). Robust designs based on the infinitesimal approach. *Journal of the American Statistical Association*, 92(440):1503–1511.

Woodroofe, M. (1975). *Probability with applications*. McGraw-Hill Book Company, Singapore.

Wu, E. K. H. and Zhou, J. (1998). Efficient bias–robust M–estimators of location in Kolmogorov normal neighbourhoods. *Canadian Journal of Statistics–Revue*, 26(2):257–266.