

Preface

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

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

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

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