Robust Methods in Biostatistics

Stephane Heritier
*The George Institute for International Health, University of Sydney, Australia*

Eva Cantoni
*Department of Econometrics, University of Geneva, Switzerland*

Samuel Copt
*Merck Serono International, Geneva, Switzerland*

Maria-Pia Victoria-Feser
*HEC Section, University of Geneva, Switzerland*
Robust Methods in Biostatistics
WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

Editors
David J. Balding, Noel A. C. Cressie, Garrett M. Fitzmaurice, Iain M. Johnstone, Geert Molenberghs, David W. Scott, Adrian F. M. Smith, Ruey S. Tsay, Sanford Weisberg, Harvey Goldstein.

Editors Emeriti
Vic Barnett, J. Stuart Hunter, Jozef L. Teugels

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

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

Eva Cantoni
Department of Econometrics, University of Geneva, Switzerland

Samuel Copt
Merck Serono International, Geneva, Switzerland

Maria-Pia Victoria-Feser
HEC Section, University of Geneva, Switzerland
To Anna, Olivier, Cassandre, Oriane, Sonia, Johannes, Véronique, Sébastien and Raphaël, who contributed in their ways...
Contents

Preface xiii
Acknowledgments xv

1 Introduction 1
1.1 What is Robust Statistics? .......................... 1
1.2 Against What is Robust Statistics Robust? .......... 3
1.3 Are Diagnostic Methods an Alternative to Robust Statistics? ... 7
1.4 How do Robust Statistics Compare with Other Statistical Procedures in Practice? ....................... 11

2 Key Measures and Results 15
2.1 Introduction ........................................ 15
2.2 Statistical Tools for Measuring Robustness Properties ...... 16
  2.2.1 The Influence Function .......................... 17
  2.2.2 The Breakdown Point ............................ 20
  2.2.3 Geometrical Interpretation ....................... 20
  2.2.4 The Rejection Point ............................. 21
2.3 General Approaches for Robust Estimation ............... 21
  2.3.1 The General Class of $M$-estimators ................ 23
  2.3.2 Properties of $M$-estimators ..................... 27
  2.3.3 The Class of $S$-estimators ..................... 30
2.4 Statistical Tools for Measuring Tests Robustness .......... 32
  2.4.1 Sensitivity of the Two-sample $t$-test ............. 34
  2.4.2 Local Stability of a Test: the Univariate Case .......... 34
  2.4.3 Global Reliability of a Test: the Breakdown Functions ... 37
2.5 General Approaches for Robust Testing .................. 38
  2.5.1 Wald Test, Score Test and LRT .................. 39
  2.5.2 Geometrical Interpretation ...................... 40
  2.5.3 General $\Psi$-type Classes of Tests ............. 40
  2.5.4 Asymptotic Distributions ....................... 42
  2.5.5 Robustness Properties ......................... 43
3 Linear Regression

3.1 Introduction .................................. 45
3.2 Estimating the Regression Parameters ................. 47
  3.2.1 The Regression Model ...................... 47
  3.2.2 Robustness Properties of the LS and MLE Estimators ..... 48
  3.2.3 Glomerular Filtration Rate (GFR) Data Example ..... 49
  3.2.4 Robust Estimators .......................... 50
  3.2.5 GFR Data Example (continued) ............... 54
3.3 Testing the Regression Parameters .................... 55
  3.3.1 Significance Testing ........................ 55
  3.3.2 Diabetes Data Example ....................... 58
  3.3.3 Multiple Hypothesis Testing .................. 59
  3.3.4 Diabetes Data Example (continued) ............. 61
3.4 Checking and Selecting the Model .................... 62
  3.4.1 Residual Analysis .......................... 62
  3.4.2 GFR Data Example (continued) ............... 62
  3.4.3 Diabetes Data Example (continued) ............. 65
  3.4.4 Coefficient of Determination ................ 66
  3.4.5 Global Criteria for Model Comparison .......... 69
  3.4.6 Diabetes Data Example (continued) ............. 75
3.5 Cardiovascular Risk Factors Data Example ............ 78

4 Mixed Linear Models

4.1 Introduction .................................. 83
4.2 The MLM ..................................... 84
  4.2.1 The MLM Formulation ...................... 84
  4.2.2 Skin Resistance Data ........................ 88
  4.2.3 Semantic Priming Data ....................... 89
  4.2.4 Orthodontic Growth Data .................... 90
4.3 Classical Estimation and Inference .................. 91
  4.3.1 Marginal and REML Estimation ............... 91
  4.3.2 Classical Inference ........................ 94
  4.3.3 Lack of Robustness of Classical Procedures ..... 96
4.4 Robust Estimation ................................ 97
  4.4.1 Bounded Influence Estimators ................ 97
  4.4.2 S-estimators ................................ 98
  4.4.3 MM-estimators .............................. 100
  4.4.4 Choosing the Tuning Constants ............... 102
  4.4.5 Skin Resistance Data (continued) ............. 103
4.5 Robust Inference ................................ 104
  4.5.1 Testing Contrasts .......................... 104
  4.5.2 Multiple Hypothesis Testing of the Main Effects .... 106
  4.5.3 Skin Resistance Data Example (continued) ...... 107
  4.5.4 Semantic Priming Data Example (continued) .... 107
  4.5.5 Testing the Variance Components ............... 110
CONTENTS

4.6 Checking the Model .......................... 110
   4.6.1 Detecting Outlying and Influential Observations  110
   4.6.2 Prediction and Residual Analysis ............... 112
4.7 Further Examples ................................ 116
   4.7.1 Metallic Oxide Data .......................... 116
   4.7.2 Orthodontic Growth Data (continued) ......... 118
4.8 Discussion and Extensions .......................... 122

5 Generalized Linear Models 125
   5.1 Introduction .................................. 125
   5.2 The GLM ...................................... 126
      5.2.1 Model Building ............................ 126
      5.2.2 Classical Estimation and Inference for GLM  129
      5.2.3 Hospital Costs Data Example ............... 132
      5.2.4 Residual Analysis .......................... 133
   5.3 A Class of M-estimators for GLMs .................. 136
      5.3.1 Choice of ψ and w(x) ...................... 137
      5.3.2 Fisher Consistency Correction ............... 138
      5.3.3 Nuisance Parameters Estimation ............. 139
      5.3.4 IF and Asymptotic Properties ............... 140
      5.3.5 Hospital Costs Example (continued) ......... 140
   5.4 Robust Inference ............................... 141
      5.4.1 Significance Testing and CIs ............... 141
      5.4.2 General Parametric Hypothesis Testing and
            Variable Selection .......................... 142
      5.4.3 Hospital Costs Example (continued) ......... 144
   5.5 Breastfeeding Data Example ....................... 146
      5.5.1 Robust Estimation of the Full Model ......... 146
      5.5.2 Variable Selection .......................... 148
   5.6 Doctor Visits Data Example ...................... 151
      5.6.1 Robust Estimation of the Full Model ......... 151
      5.6.2 Variable Selection .......................... 154
   5.7 Discussion and Extensions ......................... 158
      5.7.1 Robust Hurdle Models for Counts .......... 158
      5.7.2 Robust Akaike Criterion .................... 159
      5.7.3 General \( C_p \) Criterion for GLMs ........ 159
      5.7.4 Prediction with Robust Models ............... 160

6 Marginal Longitudinal Data Analysis 161
   6.1 Introduction .................................. 161
   6.2 The Marginal Longitudinal Data Model (MLDA) and Alternatives 163
      6.2.1 Classical Estimation and Inference in MLDA  164
      6.2.2 Estimators for \( \tau \) and \( \alpha \) ............ 166
      6.2.3 GUIDE Data Example .......................... 169
      6.2.4 Residual Analysis .......................... 171
6.3 A Robust GEE-type Estimator ........................................ 172
  6.3.1 Linear Predictor Parameters ............................... 172
  6.3.2 Nuisance Parameters ....................................... 174
  6.3.3 IF and Asymptotic Properties ........................... 176
  6.3.4 GUIDE Data Example (continued) .......................... 177
6.4 Robust Inference ..................................................... 178
  6.4.1 Significance Testing and CIs ........................... 178
  6.4.2 Variable Selection .................................... 179
  6.4.3 GUIDE Data Example (continued) ...................... 180
6.5 LEI Data Example ..................................................... 182
6.6 Stillbirth in Piglets Data Example ............................. 186
6.7 Discussion and Extensions ........................................ 189

7 Survival Analysis ..................................................... 191
  7.1 Introduction ...................................................... 191
  7.2 The Cox Model .................................................... 193
    7.2.1 The Partial Likelihood Approach ....................... 193
    7.2.2 Empirical Influence Function for the PLE ........... 196
    7.2.3 Myeloma Data Example .................................. 197
    7.2.4 A Sandwich Formula for the Asymptotic Variance ...... 198
  7.3 Robust Estimation and Inference in the Cox Model ............ 200
    7.3.1 A Robust Alternative to the PLE ..................... 200
    7.3.2 Asymptotic Normality .................................. 202
    7.3.3 Handling of Ties ....................................... 204
    7.3.4 Myeloma Data Example (continued) .................... 205
    7.3.5 Robust Inference and its Current Limitations .......... 206
  7.4 The Veteran’s Administration Lung Cancer Data ............... 209
    7.4.1 Robust Estimation ...................................... 209
    7.4.2 Interpretation of the Weights ....................... 210
    7.4.3 Validation ............................................ 212
  7.5 Structural Misspecifications .................................. 214
    7.5.1 Performance of the ARE .............................. 214
    7.5.2 Performance of the robust Wald test ................... 216
    7.5.3 Other Issues ........................................ 217
  7.6 Censored Regression Quantiles ................................ 217
    7.6.1 Regression Quantiles .................................... 217
    7.6.2 Extension to the Censored Case ...................... 219
    7.6.3 Asymptotic Properties and Robustness ................ 220
    7.6.4 Comparison with the Cox Proportional Hazard Model ... 221
    7.6.5 Lung Cancer Data Example (continued) .................. 222
    7.6.6 Limitations and Extensions ............................. 224
<table>
<thead>
<tr>
<th>CONTENTS</th>
<th>xi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appendices</td>
<td>227</td>
</tr>
<tr>
<td>A Starting Estimators for ( MM )-estimators of Regression Parameters</td>
<td>229</td>
</tr>
<tr>
<td>B Efficiency, LRT( \rho ), RAIC and ( RC \rho ) with Biweight ( \rho )-function for the Regression Model</td>
<td>231</td>
</tr>
<tr>
<td>C An Algorithm Procedure for the Constrained S-estimator</td>
<td>235</td>
</tr>
<tr>
<td>D Some Distributions of the Exponential Family</td>
<td>237</td>
</tr>
<tr>
<td>E Computations for the Robust GLM Estimator</td>
<td>239</td>
</tr>
<tr>
<td>E.1 Fisher Consistency Corrections</td>
<td>239</td>
</tr>
<tr>
<td>E.2 Asymptotic Variance</td>
<td>240</td>
</tr>
<tr>
<td>E.3 IRWLS Algorithm for Robust GLM</td>
<td>242</td>
</tr>
<tr>
<td>F Computations for the Robust GEE Estimator</td>
<td>245</td>
</tr>
<tr>
<td>F.1 IRWLS Algorithm for Robust GEE</td>
<td>245</td>
</tr>
<tr>
<td>F.2 Fisher Consistency Corrections</td>
<td>246</td>
</tr>
<tr>
<td>G Computation of the ( CRQ )</td>
<td>247</td>
</tr>
<tr>
<td>References</td>
<td>249</td>
</tr>
<tr>
<td>Index</td>
<td>265</td>
</tr>
</tbody>
</table>
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.