Lecture Notes in Statistics

Proceedings

Volume 211

For further volumes:
http://www.springer.com/series/8440

Brajendra C. Sutradhar
Editor

ISS-2012 Proceedings Volume On Longitudinal Data Analysis Subject to Measurement Errors, Missing Values, and/or Outliers

Springer

Editor
Brajendra C. Sutradhar
Department of Mathematics & Statistics
Memorial University of Newfoundland
St. John's, NL, Canada

ISSN 0930-0325
ISBN 978-1-4614-6870-7 ISBN 978-1-4614-6871-4 (eBook)
DOI 10.1007/978-1-4614-6871-4
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2013941150

Printed on acid-free paper

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

To
Bhagawan Sri Sathya Sai Baba

Preface

This special proceedings volume contains nine selected papers that were presented in the *International Symposium in Statistics (ISS) on Longitudinal Data Analysis Subject to Outliers, Measurement Errors, and/or Missing Values*, held at Memorial University, Canada from July 16–18, 2012. Three years ago the ISS-2009 was organized focussing *on inferences in generalized linear longitudinal mixed models (GLLMMs)*, and a special issue of the Canadian Journal of Statistics (2010, Vol. 38, June issue, Wiley) was published with seven selected papers from this symposium. These seven papers from ISS-2009 dealt with progress and challenges in the area of discrete longitudinal data analysis. As a reflection of the theme of the ISS-2012, the papers in the present volume deal with inferences for longitudinal data with additional practical issues such as measurement errors, missing values, and/or outliers. The inferences for this type of complex longitudinal data become more challenging than the inferences for standard longitudinal data following generalized linear longitudinal models (GLLMs). The present volume with nine papers makes a significant contribution toward such challenging inferences. To make it as precise and clear as possible, the papers are grouped into three parts along the theme of the symposium. Part I contains four papers in longitudinal data analysis subject to measurement errors, similarly Part II also contains four papers but they are in longitudinal data analysis subject to missing values, and Part III has one paper dealing with inferences for longitudinal data subject to outliers.

In a longitudinal setup, repeated responses along with a set of multidimensional time-dependent covariates are collected over a small period of time. There are situations where it is realized that the observed multidimensional covariates at a given time point differ from the corresponding true covariates by some measurement errors, but it is of interest to find the regression effects of the true covariates on the repeated responses. The first paper in Part I, by B. C. Sutradhar, begins with a discussion on this measurement error problem for scalar responses. This setup with scalar responses is referred to as the independent setup. In the first part of the paper, the author considers the independent setup and provides an overview of the existing vast literature on inferences for various bias correction approaches. In the longitudinal setup, repeated responses are, however, likely to follow a true

time-dependent covariates-based correlation structure. Because the true covariates are unobserved, this involvement of the true covariates in the correlation structure makes the bias correction to the observed covariates-based regression estimator very difficult, specially for longitudinal discrete such as count and binary data. In the second part of the paper, the author gives a brief discussion on the existing bias corrected generalized method of moments (BCGMM) and generalized quasi-likelihood (BCGQL) approaches in the linear longitudinal models (LLMs) setup. The author then discusses the progress and challenges in obtaining bias corrected inferences in the generalized LLMs (GLLMs) setup, mainly for repeated count and binary data. An overview is given on how to develop a BCGQL inference approach for longitudinal count data when corresponding covariates are subject to measurement errors. The bias correction inference for longitudinal binary data appears to be extremely difficult. The author has discussed the progress and emphasized on further investigations to resolve this challenging problem. In the second paper of Part I, Laine Thomas, Leonard A. Stefanski, and Marie Davidian consider a binary logistic regression model in independent setup where, on top of baseline covariates, the binary responses are also influenced by the mean and variance parameters of a scalar covariate which is repeatedly measured over a period of time. The authors have used a moment approach for the prediction of the variance components involved in the linear regression measurement error models for the repeated values of the covariate, and these predicted variances are used in turn in a conditional scores-based bias correction approach for the estimation of the main regression parameters of the binary outcome model. As opposed to the longitudinal setup, the third paper in Part I, by John P. Buonaccorsi, deals with measurement errors in time series. The author assumes that the true time series follows a dynamic model of interest but the series itself is unobserved. Instead, a series with measurement error is observed. Thus, the observed response at a given time is not necessarily the true response, rather, it follows a suitable distribution with its conditional mean as a function of the true response. The author has discussed various bias correction approaches including moments and likelihood methods for the estimation of the parameters of the dynamic model for the true but unobserved time series. In the fourth paper of Part I, Erik Meijer, Laura Spierdijk, and Tom Wansbeek consider a linear dynamic mixed model in panel data setup, but the true responses satisfying the underlying model are not observed. The observed responses, which are subject to measurement errors, are used to obtain consistent and efficient estimators for the parameters of the model for the true responses. Thus, this paper considers the measurement error in responses, whereas the first and the second paper in this part considered the measurement errors in covariates.

In many biomedical, clinical, and socioeconomic studies, a response and its corresponding multidimensional covariates are collected repeatedly over time from a large number of independent individuals. In this setup, it is assumed that the repeated responses from an individual marginally follow a linear or nonlinear regression model and jointly they follow a longitudinal correlation structure. It is of interest to estimate the regression effects of the time-dependent covariates on the repeated responses. For varieties of reasons it may, however, happen that a portion

of responses are missing from some individuals. In practice, in general, there are three types of missing mechanism such as missing completely at random (MCAR), missing at random (MAR), or missing non-ignorably. Further the nonresponse may occur in a monotonic fashion or they may be intermittent. The analysis of this type of incomplete longitudinal data, specially the inferences for the regression effects by using incomplete data, is complicated. This is because, to develop proper inferences, one requires to accommodate both missing mechanism and the correlation structure for the available longitudinal responses. The first paper in Part II, by B. C. Sutradhar, provides an overview on incomplete data analysis both in independent and aforementioned longitudinal setup. In the independent setup, attempts are made to collect multidimensional responses from a large number of independent individuals, but it may happen that a small portion of individuals do not provide complete multidimensional responses leading to incomplete data. The author first discusses some of the widely used existing estimation methods including the imputation technique for such incomplete data analysis in the independent setup. The author then discusses the progress made and the difficulties encountered, by the existing inference techniques, which do not appear to accommodate the missing mechanism and longitudinal correlations properly. Details are given for some remedies to overcome this anomaly in order to develop proper estimating equations for the regression effects. An unconditional as well as a conditional approach is discussed to develop estimating equations for consistent regression estimates. In the second paper of Part II, Taslim Mallick, Patrick Farrell, and B. C. Sutradhar proposed a GQL approach along the lines of the first paper by B. C. Sutradhar that provides consistent regression estimates. When the responses are MAR, the authors have further demonstrated that the existing generalized estimating equations (GEE) approach encounters serious convergence problems specially when missing proportion is large. This breakdown shows the inconsistency of the GEE-based approaches, whereas the proposed GQL approach does not encounter such convergence problems unless the missing proportion is unreasonably high, and it produces almost unbiased regression estimates with smaller standard errors. In the third paper of Part II, Paul S. Albert, Rajeshwari Sundaram, and Alexander C. McLain discuss a random effects approach to analyze longitudinal data subject to missing. The authors introduce suitable random effects and assume that they cause the correlations among repeated data and also determine the missing mechanism. More specifically, they assume that conditional on the same random effects, the responses do not dependent on the missing data status, also the repeated responses are independent. This yields a simple probability model for the observed random variables, that is, responses and missing indicators, which in turn leads to fairly simple likelihood and/or conditional likelihood inferences for the regression parameters. In the last paper of Part II, Michael A. McIsaac and Richard J. Cook consider a two-phase sampling-based dropout model, where in the first phase a vector of clustered or repeated responses along with a vector of multidimensional auxiliary covariates are collected from a large number of independent individuals. However, in the second phase, a vector of multidimensional expensive covariates are collected only from a portion of individuals. It is of interest to examine the effects

of both auxiliary and expensive covariates on the clustered responses. The authors discuss the incomplete data-based likelihood, mean score, and weighted pseudo-likelihood methods for the estimation of such regression effects.

Part III of the volume contains one paper by B. C. Sutradhar, on the inferences for longitudinal data subject to outliers. It is known in the independent setup that a few outlying responses mainly caused by the associated contaminated covariates may adversely influence the valid inferences for the regression effects. The author first gives an overview of the existing robust approaches in the independent setup for the estimation of the regression effects in linear, count, and binary data models. These approaches include a recently developed fully standardized Mallow's type quasi-likelihood (FSMQL) method that provides almost unbiased regression estimates. The author then extends the overview to the longitudinal setup. The robust inferences for longitudinal binary and count data are, however, not adequately discussed in the literature. The author discusses a robust GQL approach for unbiased regression estimation for count and binary longitudinal data models.

St. John's, NL, Canada Brajendra C. Sutradhar

ISS-2012 Delegates

ISS-2012 Welcome Address by Brajendra C. Sutradhar (Organizer)

With the name of Lord, we welcome all of you, to Memorial University, the host for the International Symposium in Statistics, 2012 (ISS-2012) on Longitudinal Data Analysis Subject to Outliers, Measurement Errors, and/or Missing Values. It gives us a pleasure to note that we have been able to keep up the spirit of the first symposium (ISS-2009) that took place here in Memorial University, in organizing the present symposium covering extended and more challenging research areas in the longitudinal setup, mainly for discrete data such as count and binary data.We thank all of you for your interest and response to this symposium that has attempted to attract the researchers deeply involved in the inferences for longitudinal data those encounter practical difficulties due to measurement errors, nonresponse, and/or outliers. We hope that you will find the symposium stimulating and will derive spirits for doing more and more quality research in these challenging areas as a service to the society and mankind at large. We also hope that the symposium generates and enhances the spirit of collaborative research among the participants which also might reconfirm our sense of achievement in a greater horizon of life, as the proverb goes: "Life is a march from I to We to He (Sri Sathya Sai Baba, India)". It is indeed a pleasure to note that we have delegates in this specialized symposium from many countries such as Australia, Bangladesh, Brazil, Canada, Mauritius, the Netherlands, Saudi Arabia, Spain, and the USA covering a large part of the globe. We extend our hearty welcome to all of you.

We also welcome you to St. John's, the oldest city of North America, known as the City of Legends, where you can view icebergs, watch whales, and experience Newfoundland and Labrador's unique culture. It is a progressive city and is the site of many world class facilities. A mosaic of fishing villages, cultural festivals, and wildlife tours bring variety to the city. Also, the Cape Spear, the most easterly point of North America, is not far from the city, where one can experience the unique beauty of sunrise. We hope that you have planned for an extended stay in St. John's following the symposium to enjoy these and other endless options!

St. John's, NL, Canada Brajendra C. Sutradhar

Acknowledgments

This proceedings volume (lecture note) is a collection of selected papers that were presented in ISS-12 (International Symposium in Statistics, 2012) held at Memorial University from July 16–18, 2012. Organizing this symposium would not, however, have been possible without the generous contributions from Memorial University and the Atlantic Association for Research in Mathematical Sciences. I wish to express my special thanks to these two institutes for their support.

All papers in this volume were refereed. Prior to the symposium, the papers were sent to the referees who were also supposed to be present during the presentation. The authors were also benefitted from the warm discussion by the audience of the symposium and prepared the revision of the paper by addressing all suggestions and comments from the referees and the audience. My sincere thanks go to the delegates and referees to make the symposium and this volume a grand success. Some of the contributed papers were also considered for their publication in this volume. A special thanks go to Dr. Alwell Oyet for his warm service in processing all contributed papers for the symposium. It has been a pleasure to work with Marc Strauss, Hannah Bracken, Mary Helena, and Lesley Poliner of Springer-Verlag in preparing this volume.

Contents

Part III Longitudinal Data Analysis Subject to Outliers

Contributors

Paul S. Albert *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, Bethesda, MD, USA

John P. Buonaccorsi University of Massachusetts, Amherst, MA, USA

Richard J. Cook Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada

Marie Davidian Department of Statistics, North Carolina State University, Raleigh, NC, USA

Patrick J. Farrell School of Mathematics and Statistics, Carleton University, Ottawa, ON, Canada

Taslim S. Mallick School of Mathematics and Statistics, Carleton University, Ottawa, ON, Canada

Michael A. McIsaac Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada

Alexander C. McLain University of South Carolina, Columbia, SC, USA

Erik Meijer RAND Corporation, Santa Monica, CA, USA

Laura Spierdijk University of Groningen, Groningen, The Netherlands

Leonard A. Stefanski Department of Statistics, North Carolina State University, Raleigh, NC, USA

Rajeshwari Sundaram *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, Bethesda, MD, USA

Brajendra C. Sutradhar Memorial University, St. John's, NL, Canada

Laine Thomas Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA

Tom Wansbeek University of Groningen, Groningen, The Netherlands