Statistics for Social and Behavioral Sciences

Maria Kateri
Irini Moustaki *Editors*

Trends and Challenges in Categorical Data Analysis

Statistical Modelling and Interpretation

Springer

Statistics for Social and Behavioral Sciences

Statistics for Social and Behavioral Sciences (SSBS) includes monographs and advanced textbooks relating to education, psychology, sociology, political science, public policy, and law.

Maria Kateri • Irini Moustaki
Editors

Trends and Challenges in Categorical Data Analysis

Statistical Modelling and Interpretation

Springer

Editors
Maria Kateri
Department of Mathematics
RWTH Aachen University
Aachen, Germany

Irini Moustaki
Department of Statistics
London School of Econ. & Polit. Science
London, UK

ISSN 2199-7357 ISSN 2199-7365 (electronic)
Statistics for Social and Behavioral Sciences
ISBN 978-3-031-31185-7 ISBN 978-3-031-31186-4 (eBook)
https://doi.org/10.1007/978-3-031-31186-4

Mathematics Subject Classification: 62H, 62J

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The analysis of categorical data has led to the development of a whole new set of methods, tools, and theory. The methodological developments have been followed by commercial and open source software, which has facilitated the spread and use of the methods in many substantive areas of application. The book aims to bring together and provide a comprehensive review of a selected list of topics connected to recent advances in statistical modelling and interpretation of categorical data. The focus is on cross-sectional as well as time-dependent data.

We consider research questions of both symmetrical and regression-type nature, such as studying and modelling the association of a number of categorical variables, as well as regression-type analysis of a categorical response variable explained by a number of observed covariates.

Categorical data predominate in social surveys and their analysis, from descriptive and exploratory to statistical modelling, require special treatments that take into account the nature and information included in these data. Traditionally, categorical data analysis (CDA) methodology has focused on two- and three-way contingency tables, while for higher dimensional tables, it is usually commented that they are analysed analogously. Many of today's real applications involve high-dimensional and complex data with many more than three variables. Categorical data methods have been extended to handle multivariate data of higher dimensions, addressing issues of sparseness, model estimation, fit, and model selection. More specifically, binary and ordinal response models have become the focus of attention in areas of supervised machine learning. Graphical models and networks involving categorical data have applications, in social sciences, biology, and natural language processing, among others. Developments and problems in data science necessitate special treatment for different types of categorical data and impose new challenges on CDA.

To tackle problems in contemporary applications of categorical data, a thoughtful revisiting of traditional methods of CDA is required.

Serving this goal, the current volume covers nine distinct topics, underlining, when necessary, their inter-relationships and helping the reader to place methods and tools for categorical data into a general framework. It reviews association models for multi-way contingency tables and their connection to item response

theory models and graphical models, marginal models, regression type models with categorical responses, and/or categorical covariates including simple measures of interpretation, time series models for count and binary data, models for binary panel data, as well as methodology for bias correction and Bayesian inference.

The volume is intended for statisticians, data scientists, graduate students of statistics, but also computer scientists or researchers with a strong interest in methods and tools used for the analysis of categorical data. The chapters include applications from economics, education, psychiatry, medicine, and finance, but the applicability of the methods discussed go beyond those areas.

The volume is organised into three parts. Part I (Chaps. 1–4) focuses on modelling multivariate (multiple response variables) categorical data through their joint and marginal distributions. Chapter 1 reviews classical association models and establishes the connection with item response theory models and graphical models that provide multiple insights into the data problem. A computationally feasible composite likelihood estimation method and testing framework are proposed. Real data examples from massively open online courses (MOOC) and from the Depression, Anxiety and Stress Scale (DASS) are included, as well as information on the R packages logmulti and pleLMA. Graphical models are discussed in more detail in Chap. 2, which covers undirected graphical log-linear models, directed graphical models, and graphical chain models for modelling complex multivariate associations. The infant survival data, presented in other seminal books on categorical data, are used to illustrate the various graphical models. Graphical models already covered in Chaps. 1 and 2 are shown to be connected to the class of marginal models presented in Chap. 3. In this chapter, a thorough overview of marginal models is provided. Marginal models are helpful for testing hypotheses about relations among correlated categorical marginal distributions. The content of this chapter is motivated with examples from repeated measurements/panel data, missing data, and graphical data in which marginal distributions of higher-dimensional joint distributions play an important role. Potential estimation methods are thoroughly discussed. Information on three available R packages (cmm, mph.fit, and hmmm) for marginal modelling is provided. The chapter concludes with a list of further theoretical and methodological developments in the area of marginal modelling and extensions for the future. Chapter 4 offers a Bayesian treatment of multivariate categorical data with emphasis on estimation, choices of priors, and model selection. The explored tools are applied to two-way contingency tables from three medical areas of research, namely risk for coronary heart disease, lymphoma and chemotherapy, and toxaemia in pregnancy.

Part II (Chaps. 5–7) focuses on regression type models for binary and ordinal responses. Chapter 5 proposes probability-based effect measures that provide a simpler interpretation of regression coefficients of logistic and probit models with linear and non-linear predictors, which are missing from the traditional literature on binary and ordinal regression. The proposed measures are used to compute effective measures for a class of generalised linear models with logit, log, and identity link functions, fitted to data from an Italian survey on employment status and a generalised additive model fitted to the horseshoe crab data. R code is provided

for replicating the analysis. Chapter 6 proposes mean and median bias reduction in adjacent-categories logit models with proportional odds and mean bias reduction in models with non-proportional odds. The methodology is illustrated using real examples, and the R code is provided to replicate all the numerical and graphical results. Chapter 7 gives an overview of regularised estimation methods for generalised additive models with ordinal covariates, considering predictor selection and merging of predictor categories with the effect of reducing the number of parameters and easing interpretability. The proposed method is compared to existing classical methods and is applied to a real data set from the International Classification of Functioning, Disability and Health study on chronic widespread pain. Information on R packages that perform the different types of analysis discussed in the chapter is provided.

Part III (Chaps. 8 and 9) discusses models for discrete time-dependent data. Chapter 8 presents an overview of a unified framework of ARMA-type models widely used for continuous time series for binary and count data, with emphasis on associated stochastic properties and likelihood-based inferential tools. The methodology is applied to two real data sets: the daily number of deaths from COVID-19 in Italy, for which a Poisson and a negative binomial distribution is assumed for the data; and a binary series of log-returns for the weekly closing prices of Johnson & Johnson. The code for replicating the analysis in the chapter is provided. Finally, Chap. 9 reviews the formulation and estimation of fixed-effects type models for binary panel data. In particular, the chapter reviews and illustrates, through an extensive simulation study, estimation methods for dealing with the inconsistency of the maximum likelihood estimator due to incidental parameters, embedding in a unified framework the target-corrected and conditional maximum likelihood estimators, including a pseudo conditional maximum likelihood estimator. The methodology is applied to data on female labour force participation from the US Panel Study of Income Dynamics. The chapter also includes a review of packages available to estimate the models discussed.

Each chapter makes its own methodological and distinct contribution to the modelling of categorical data and can be read independently. In some cases, connections are made among the topics covered in the edited volume, but these connections or overlaps do not imply that the reader needs to read the chapters in any particular order. The division of the book in three parts is also indicative and does not provide a strict separation of the contributions.

The seed for this volume was sown during the workshop “Challenges for Categorical Data Analysis” (CCDA2018) held in Aachen in 2018. We would like to thank all the participants of this workshop for the inspiring discussions and for motivating our book project. We specially thank Eva Hiripi, Senior Editor at Springer, for her continuous support and guidance in the process of preparing the volume.

Finally, we are grateful to the friends and colleagues who contributed chapters to this volume. Without their engagement and impressive work, this project would not have been possible.

Aachen, Germany — Maria Kateri
London, UK — Irini Moustaki
February, 2023

Contents

1 Log-Linear and Log-Multiplicative Association Models for Categorical Data 1
Carolyn J. Anderson, Maria Kateri, and Irini Moustaki

2 Graphical Models for Categorical Data 43
Peter W. F. Smith

3 Marginal Models: An Overview 67
Tamás Rudas and Wicher Bergsma

4 Bayesian Inference for Multivariate Categorical Data 117
Jonathan J. Forster and Mark E. Grigsby

5 Simple Ways to Interpret Effects in Modeling Binary Data 155
Alan Agresti, Claudia Tarantola, and Roberta Varriale

6 Mean and Median Bias Reduction: A Concise Review and Application to Adjacent-Categories Logit Models 177
Ioannis Kosmidis

7 Regularization and Predictor Selection for Ordinal and Categorical Data 199
Jan Gertheiss and Gerhard Tutz

8 An Overview of ARMA-Like Models for Count and Binary Data 233
Mirko Armillotta, Alessandra Luati, and Monia Lupparelli

9 Advances in Maximum Likelihood Estimation of Fixed-Effects Binary Panel Data Models 275
Francesco Valentini, Claudia Pigini, and Francesco Bartolucci

Contributors

Alan Agresti Department of Statistics, University of Florida, Gainesville, FL, USA

Carolyn J. Anderson Department of Educational Psycholgy, College of Education, University of Illinois at Urbana-Champaign, Champaign, IL, USA

Mirko Armillotta Department of Statistical Sciences, University of Bologna, Bologna, Italy

Francesco Bartolucci Department of Economics, University of Perugia, Perugia, Italy

Wicher Bergsma Department of Statistics, London School of Economics and Political Science, London, UK

Jonathan J. Forster Department of Statistics, University of Warwick, Coventry, UK

Jan Gertheiss Helmut Schmidt University, Hamburg, Germany

Mark E. Grigsby Proctor and Gamble, The Heights Weybridge, Surrey, UK

Maria Kateri Institute of Statistics, RWTH Aachen University, Aachen, Germany

Ioannis Kosmidis Department of Statistics, University of Warwick, Coventry, UK

Alessandra Luati Department of Statistical Sciences, University of Bologna, Bologna, Italy

Monia Lupparelli Department of Statistics, Computer Science, Applications, University of Florence, Florence, Italy

Irini Moustaki Department of Statistics, London School of Economics and Political Science, London, UK

Claudia Pigini Department of Economics and Social Sciences, Marche Polytechnic University, Ancona, Italy

Tamás Rudas Department of Statistics, Faculty of Social Sciences, Eötvös Loránd University, Budapest, Hungary

Peter W. F. Smith Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, UK

Claudia Tarantola Department of Economics and Management, University of Pavia, Pavia, Italy

Gerhard Tutz Ludwig Maximilians University, Munich, Germany

Francesco Valentini Department of Economics and Social Sciences, Marche Polytechnic University, Ancona, Italy

Roberta Varriale Istat, Rome, Italy