

Emerging Topics in Statistics and Biostatistics

Jeffrey R. Wilson
Elsa Vazquez-Arreola
(Din) Ding-Geng Chen

Marginal Models in Analysis of Correlated Binary Data with Time Dependent Covariates

 Springer

Emerging Topics in Statistics and Biostatistics

Series Editor

(Din) Ding-Geng Chen, University of North Carolina, Chapel Hill, NC, USA

Editorial Board Members

Andriëtte Bekker, University of Pretoria, Pretoria, South Africa

Carlos A. Coelho, Universidade Nova de Lisboa, Caparica, Portugal

Maxim Finkelstein, University of the Free State, Bloemfontein, South Africa

Jeffrey R. Wilson, Arizona State University, Tempe, AZ, USA

More information about this series at <http://www.springer.com/series/16213>

Jeffrey R. Wilson • Elsa Vazquez-Arreola
(Din) Ding-Geng Chen

Marginal Models in Analysis of Correlated Binary Data with Time Dependent Covariates

 Springer

Jeffrey R. Wilson
Department of Economics
W. P. Carey School of Business
Arizona State University
Chandler, AZ, USA

Elsa Vazquez-Arreola
School of Mathematical and Statistical
Sciences
Arizona State University
Tempe, AZ, USA

(Din) Ding-Geng Chen
School of Social Work &
Department of Biostatistics
University of North Carolina
Chapel Hill, NC, USA

Department of Statistics
University of Pretoria
Pretoria, South Africa

ISSN 2524-7735

ISSN 2524-7743 (electronic)

Emerging Topics in Statistics and Biostatistics

ISBN 978-3-030-48903-8

ISBN 978-3-030-48904-5 (eBook)

<https://doi.org/10.1007/978-3-030-48904-5>

© Springer Nature Switzerland AG 2020

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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

I dedicate this to my students, present and past. Their insight had a great deal to do with the materials covered in this book.

Jeffrey R. Wilson

I dedicate this to my family for their unconditional support.

Elsa Vazquez-Arreola

I dedicate this to my family for their support.

(Din) Ding-Geng Chen

Preface

In the analysis of correlated data, it is often desirable to evaluate the effect of the time-dependent covariates. However, the changing nature of time-dependent covariates may have delayed effects or feedback. If the relation goes unchecked, one can have a differential effect on the response, and the conventional models may not be appropriate.

The focus of this book is the modeling of correlated response data with time-dependent covariates. We have been accustomed to models for correlated data with time-independent covariates, but modeling correlated data with time-dependent covariates brings some added challenges. These include delayed effects, feedback between responses and covariates, and relation among the responses. This book is then the first book designed to address these challenges with a compilation of research and publications we developed in the past years to address the analysis of correlated data with time-dependent covariates.

Chandler, AZ, USA
Tempe, AZ, USA
Chapel Hill, NC, USA
Pretoria, South Africa

Jeffrey R. Wilson
Elsa Vazquez-Arreola
(Din) Ding-Geng Chen

Acknowledgments

The authors of this book owe a great deal of gratitude to many who helped in the completion of the book. We have been fortunate enough to work with a number of graduate students at Arizona State University: Many thanks to the staff in the Department of Economics and the computing support group in the W. P. Carey School of Business. We also gratefully acknowledge the professional support of Ms. Laura Aileen Briskman from Springer, who made the publication of this book a reality. A special thanks to Dr. Kyle Irimata, Dr. Katherine Irimata, and Dr. Trent Lalonde. To everyone involved in the making of this book, we say thank you!. This work is based on the research supported partially by the National Research Foundation of South Africa (Grant Number 127727) and the South African National Research Foundation (NRF) and South African Medical Research Council (SAMRC) (South African DST-NRF-SAMRC SARChI Research Chair in Biostatistics, Grant Number 114613).

JRW
EVA
DC

About the Book

In this book, we focus on time-dependent covariates in the fit of marginal models. We use five data sets to demonstrate these models fitted throughout the book. This book consists of eight chapters, and they represent the model development from time-independent to time-dependent developed over the last few years of our research and teaching of statistics at the master's and PhD level at Arizona State University. The aim of this book is to concentrate on using marginal models with their developed theory and the associated practical implementation. The examples in this book are analyzed whenever possible using SAS, but when possible, the R code is provided. The SAS outputs are given in the text with partial tables. The completed data sets and the associated SAS/R programs can be found at the web address www.public.asu.edu/~jeffreyw.

We provide several examples to allow the reader to mimic some of the models used. The chapters in this book are designed to help guide researchers, practitioners, and graduate students to analyze longitudinal data with time-dependent covariates.

The book is timely and has the potential to impact model fitting when faced with correlated data analyses. In an academic setting, the book could serve as a reference guide for a course on time-dependent covariates, particularly for students at the graduate-level statistics or for those seeking degrees in related quantitative fields of study. In addition, this book could serve as a reference for researchers and data analysts in education, social sciences, public health, and biomedical research or wherever clustered and longitudinal data are needed for analysis.

The book is composed of different opportunities for readers. Those interested in quick read can go from Chaps. 1, 2, 5 to 7. While others who wish to know all the details of time-dependent covariates may read Chaps. 1, 2, 5 to 8. However, once the reader is familiar with Chaps. 1 and 2, they can move in different directions as illustrated below (Fig. 1).

When analyzing longitudinal binary data, it is essential to account for both the correlation inherent from the repeated measures of the responses and the correlation realized because of the feedback created between the responses at a particular time and the covariates at other times (Fig. 2). Ignoring any of these correlations can lead to invalid conclusions. Such is the case when the covariates are time-dependent and

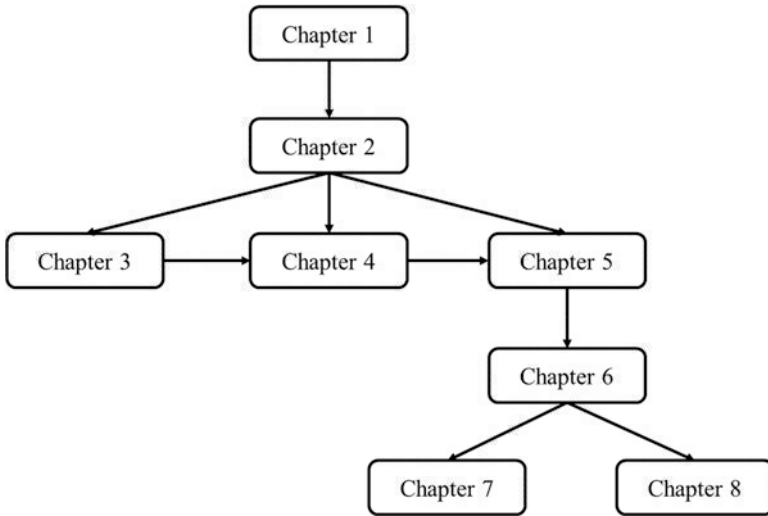


Fig. 1 Suggested system of chapter reading

Covariate over time

Responses over time

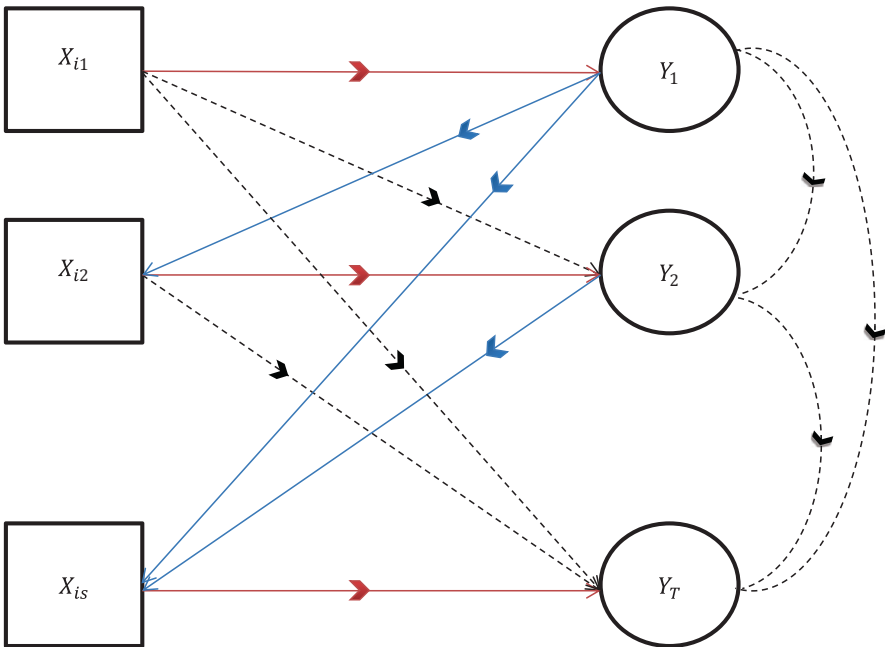


Fig. 2 Two types of correlation structures

the standard logistic regression model is used. Figure 2 describes two types of correlations: responses with responses and responses with covariates. We need a model that addresses both types of relationships. In Fig. 2, the different types of correlation presented are:

1. The correlation among the responses which are denoted by y_1, \dots, y_T as time t goes from 1 to T and
2. The correlation between response Y_t and covariate X_s :
 - (a) When responses at time t impact the covariates in time $t+s$.
 - (b) When the covariates in time t impact the responses in time $t+s$.

These correlations regarding feedback from Y_t to the future X_{t+s} and vice versa are important in obtaining the estimates of the regression coefficients.

This book provides a means of modeling repeated responses with time-dependent and time-independent covariates. The coefficients are obtained using the generalized method of moments (GMM). We fit these data using SAs and SAS Macro and at times used R.

We welcome readers' comments, including notes on typos or other errors, and look forward to receiving suggestions for improvements to future editions. Please send comments and suggestions to Professor Jeffrey Wilson (email: jeffrey.wilson@asu.edu).

Contents

1	Review of Estimators for Regression Models	1
1.1	Notation	1
1.2	Introduction to Statistical Models	1
1.2.1	The General Linear Model.	1
1.2.2	Generalized Linear Models (GLMs)	3
1.2.3	Transformation Versus GLM	5
1.2.4	Exponential Family	7
1.2.5	Estimation of the Model Parameters	8
1.3	Review of Generalized Method of Moments Estimates	14
1.3.1	Generalized Method of Moments (GMM).	14
1.3.2	Method of Moments (MM) Estimator	14
1.3.3	Some Comparisons Between ML Estimators and GMM Estimators	20
1.4	Review of Bayesian Intervals.	22
1.4.1	Bayes Theorem	23
	References.	30
2	Generalized Estimating Equation and Generalized Linear Mixed Models	31
2.1	Notation	31
2.2	Introduction to Correlated Data	31
2.2.1	Longitudinal Data	31
2.2.2	Repeated Measures	32
2.2.3	Advantages and Disadvantages of Longitudinal Data.	32
2.2.4	Data Structure for Clustered Data	33
2.3	Models for Correlated Data	34
2.3.1	The Population-Averaged or Marginal Model.	34
2.3.2	Parameter Estimation of GEE Model	36
2.3.3	GEE Model Fit.	37

- 2.3.4 The Subject-Specific Approach 40
- 2.4 Remarks 47
- References. 48
- 3 GMM Marginal Regression Models for Correlated Data with Grouped Moments 49**
 - 3.1 Notation 49
 - 3.2 Background 49
 - 3.3 Generalized Estimating Equation Models 50
 - 3.3.1 Problems Posed by Time-Dependent Covariates. 52
 - 3.4 Marginal Models with Time-Dependent Covariates 54
 - 3.4.1 Types of Covariates 55
 - 3.4.2 Model 57
 - 3.4.3 GMM Versus GEE 61
 - 3.4.4 Identifying Covariate Type. 61
 - 3.5 GMM Implementation in R 63
 - 3.6 Numerical Example 64
 - 3.6.1 Philippines: Modeling Mean Morbidity 64
 - 3.7 Further Comments 65
 - References. 66
- 4 GMM Regression Models for Correlated Data with Unit Moments 67**
 - 4.1 Notation 67
 - 4.2 Introduction 67
 - 4.3 Generalized Method Moment Models 69
 - 4.3.1 Valid Moments. 69
 - 4.3.2 Multiple Comparison Test 72
 - 4.3.3 Obtaining GMM Estimates 73
 - 4.4 SAS Marco to Fit Data. 75
 - 4.5 Numerical Examples 77
 - 4.6 Some Remarks 80
 - References. 81
- 5 Partitioned GMM Logistic Regression Models for Longitudinal Data 83**
 - 5.1 Notation 83
 - 5.2 Introduction 83
 - 5.3 Model 84
 - 5.3.1 Partitioned GMM Estimation. 86
 - 5.3.2 Types of Partitioned GMM Models. 88
 - 5.4 SAS Macro to Fit Data. 88
 - 5.5 Numerical Examples 89
 - 5.6 Some Remarks 96
 - References. 98

- 6 Partitioned GMM for Correlated Data with Bayesian Intervals** 99
 - 6.1 Notation 99
 - 6.2 Background 99
 - 6.2.1 Composite Likelihoods 101
 - 6.3 Partition GMM Marginal Model 102
 - 6.3.1 Partitioned GMM Estimation. 102
 - 6.4 Partitioned GMM Model with Bayesian Intervals. 104
 - 6.5 Properties of Model 105
 - 6.6 Code for Fit Model. 106
 - 6.7 Numerical Example 106
 - 6.8 Some Remarks 113
 - References. 113

- 7 Simultaneous Modeling with Time-Dependent Covariates and Bayesian Intervals** 117
 - 7.1 Notation 117
 - 7.2 Introduction 117
 - 7.3 Background 118
 - 7.4 Marginal Regression Modeling with Time-Dependent Covariates. 120
 - 7.4.1 Partitioned Coefficients with Time-Dependent Covariates. 121
 - 7.4.2 Partitioned Data Matrix 121
 - 7.5 MVM Marginal Model with Bayesian Intervals 122
 - 7.5.1 Simultaneous Responses with Nested Working Correlation Matrix 123
 - 7.5.2 Special Case: Single Response MVM Models with Bayesian Intervals 126
 - 7.6 Simulation Study 126
 - 7.7 Computing Code 126
 - 7.8 Numerical Examples 127
 - 7.9 Some Remarks 132
 - References. 134

- 8 A Two-Part GMM Model for Impact and Feedback for Time-Dependent Covariates.** 137
 - 8.1 Notation 137
 - 8.2 Introduction 137
 - 8.2.1 General Framework 138
 - 8.3 Two-Part Model for Feedback 140
 - 8.3.1 Stage 1: Model 140
 - 8.3.2 Feedback of Responses on Time-Dependent Predictors Model 142

8.4	Coefficients and Interpretation of the Model	146
8.5	Implementation in SAS: Code and Program	146
8.6	Numerical Examples	147
8.7	Remarks	155
	References	155
Appendix A: Introduction of Major Data Sets Analyzed in this Book . . .		157
	Index	163

List of Figures

Fig. 1.1	The variance of Athlete's salary increases with career hits.	6
Fig. 1.2	Residual plots for readmissions.	21
Fig. 1.3	Trace plots indicating convergence.	28
Fig. 3.1	A diagram of the four types of covariates.	57
Fig. 4.1	A diagram of the covariate types.	70
Fig. 5.1	Regression coefficients and Confidence intervals for time-dependent covariates when modeling using Lalonde, Wilson, and Yin moment conditions.	93
Fig. 5.2	Regression coefficients and 95% Confidence Intervals for time-dependent covariates when modeling social alcohol use with Partitioned GMM model with Lai and Small moment conditions.	94
Fig. 6.1	Markov Chains for coefficients' posterior distributions for obesity status model. *Note: beta[1] = intercept, beta[2] = white, beta[3] = depression, beta[4] = TV hours, beta[5] = physical activity, beta[6] = alcohol, beta[7] = lag-1 depression, beta[8] = lag-1 TV hours, beta[9] = lag-1 physical activity, beta[10] = lag-1 alcohol, beta[11] = lag-2 physical activity, beta[12] = lag-2 alcohol, beta[13] = lag-3 physical activity	108
Fig. 6.2	Prior and posterior distributions for coefficients in obesity status model. *Note: Red dotted line represents prior distribution, black solid density represents the posterior distribution	109

Fig. 6.3	Posterior OR and 95% credible intervals for coefficient across time for obesity status. *Note: Dots represent OR estimated based on posterior distributions' means, bands represent 95% credible intervals for OR based on 2.5% and 97.5% posterior distributions' quantiles. Lag 0 represents cross-sectional effects. Plots for lag-2 and lag-3 for depression and TV hours and for lag-3 for alcohol use are not provided because these covariates did not have valid moments at those lags.	110
Fig. 6.4	Posterior distributions for prior sensitivity for regression coefficients. Note: 1 = informative priors have variance 1; 5 = informative priors have variance 5; 10 = informative priors have variance 10; 10,000 = noninformative priors with mean 0 and variance 10,000; 0.25 = informative priors have variance 0.25.	112
Fig. 7.1	Regression coefficients and 95% credible intervals for time-dependent covariates effects on smoking	129
Fig. 7.2	Regression coefficients and 95% credible intervals for time-dependent covariates effects on social alcohol use	131
Fig. 7.3	Regression coefficients and 95% credible intervals for time-dependent covariates effects on obesity	133
Fig. 8.1	Stage 1: impact of X_{*j*} on Y_{i*}	142
Fig. 8.2	Stage 2: impact of Y_{*j} on X_{*j*}	143
Fig. 8.3	95% Confidence intervals for time-dependent covariates on obesity	149
Fig. 8.4	Regression coefficient estimates and 95% CI for effects of time-dependent covariates on interviewer-rated health using CLHLS data.	153
Fig. A.1	Suggested system of chapter reading	158
Fig. A.2	Two types of correlation structures	159

List of Tables

Table 2.1	General layout for repeated measurements	33
Table 2.2	Parameter estimates and standard errors logistic regression	45
Table 3.1	Parameter estimates and P-value based on GEE YWL-CUGMM YWL-2SGMM LS-CUGMM and LS-2SGMM	65
Table 4.1	Moment conditions for the Add Health Study	78
Table 4.2	Cross-sectional parameter estimates in the Add Health Data	79
Table 4.3	Moment conditions for the Medicare Data	80
Table 4.4	Parameter estimates for Medicare Data.	80
Table 5.1	Moment conditions for the Add Health Study	90
Table 5.2	Cross-sectional partitioned and lagged parameter estimates and p-values for the Add Health Study	92
Table 5.3	Moment conditions for the Medicare study	95
Table 5.4	Cross-sectional, partitioned and lagged parameter estimates and p-values for the Medicare study	96
Table 6.1	Results of the Partitioned GMM model with Bayesian intervals for Add Health Data	110
Table 6.2	Comparison of Partitioned GMM model with Bayesian intervals and Partitioned GMM (Add Health).	111
Table 6.3	Comparison of Partitioned MVM marginal models with Bayesian intervals and Partitioned GMM (Add Health).	111
Table 6.4	Hellinger distances between posterior distributions for obesity	113
Table 7.1	Working correlation matrix	128
Table 7.2	Parameter estimates and percentiles	128
Table 7.3	HPD intervals for parameters	129

Table 7.4 Parameter estimates and percentiles 130

Table 7.5 Parameter estimates and HPD interval 131

Table 7.6 Parameter estimates and percentiles 132

Table 7.7 Parameter estimates and HPD intervals 132

Table 7.8 Parameter estimates and percentiles 133

Table 7.9 Parameter estimates and HPD intervals 134

Table 8.1 Analysis of partial GMM estimates. 148

Table 8.2 Analysis feedback from outcome to activity
scale covariate 150

Table 8.3 Analysis feedback from outcome
to FEELINGSCALE covariate 150

Table 8.4 Analysis of partial GMM estimates. 152

Table 8.5 Analysis feedback from outcome
to OWN_DECISIONB covariate. 154

Table 8.6 Analysis feedback from outcome
to DRESSINGB covariate. 154

Table A.1 Cross classification of readmit by time 157

Table A.2 Cross classification of smoking by wave. 159

Table A.3 Cross classification of social alcohol use by wave 160

Table A.4 Cross classification of obesity by wave. 160

Table A.5 Cross classification of interviewer-rated health by wave 161

Table A.6 Cross classification of complete physical check by wave. 161

Table A.7 Cross classification of self-rated quality of life by wave 162

Table A.8 Cross classification of self-rated health by wave 162