ICSA Book Series in Statistics
Series Editor: Ding-Geng Chen

Jeffrey R. Wilson
Kent A. Lorenz
Lori P. Selby

Modeling Binary Correlated Responses

Using SAS, SPSS, R and STATA

Second Edition

Springer

ICSA Book Series in Statistics

Series Editor

Ding-Geng Chen, College of Health Solutions
Arizona State University
Chapel Hill, NC, USA

The ICSA Book Series in Statistics showcases research from the International Chinese Statistical Association that has an international reach. It publishes books in statistical theory, applications, and statistical education. All books are associated with the ICSA or are authored by invited contributors. Books may be monographs, edited volumes, textbooks and proceedings.

Jeffrey R. Wilson • Kent A. Lorenz
Lori P. Selby

Modeling Binary Correlated Responses

Using SAS, SPSS, R and STATA

Second Edition

Springer

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

Kent A. Lorenz
Department of Kinesiology
San Francisco State University
San Francisco, CA, USA

Lori P. Selby
Arizona State University, School of
Mathematics and Statistics
Tempe, AZ, USA

ISSN 2199-0980 ISSN 2199-0999 (electronic)
ICSA Book Series in Statistics
ISBN 978-3-031-62426-1 ISBN 978-3-031-62427-8 (eBook)
https://doi.org/10.1007/978-3-031-62427-8

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2015, 2024
This work is subject to copyright. All rights are solely and exclusively licensed 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

If disposing of this product, please recycle the paper.

This book is dedicated to my grandson Willem Wilson-Ellis, my three daughters, Rochelle, Roxanne, and Rhonda, and all my statistics students past and present at Arizona State University.

Jeffrey R. Wilson

My parents, Rose and Arnold, and all my family and friends.

Kent A. Lorenz

My parents Patricia Selby, Urick Selby, mentors Joseph L. and Teresa Abraham, my family and friends.

Lori P. Selby

Preface

The main focus of this book is on the modeling of binary response data. Binary outcomes can be observed directly or through the dichotomization of a continuous variable; however, binary data analysis has some unique challenges when compared to continuous data analysis. Some potential issues a researcher needs to consider when analyzing binary data are as follows:

- Are the trials based on a mechanism that produces independent or correlated observations?
- Are the data based on repeated measures or are they cross-sectional?
- Are the covariates time dependent or time independent?
- Are the covariates entered into the model as a fixed or random effect?
- Are there marginal models being fitted or subject-specific models? In other words, is the interest to model the mean or to be subject specific?

This book is based on real examples and data we have encountered over several years of research and teaching statistics at the Master's and PhD levels at Arizona State University. In fact, several of the chapters are based on the applied projects and theses of Master's and PhD students in the university's statistics programs. The examples in this book are analyzed whenever possible using SAS, SPSS, R, and STATA. While the SAS, SPSS, R, and STATA outputs are contained in the text with partial data tables, the completed datasets can be found at the web address https://www.public.asu.edu/~jeffreyw.

The aim of this book is to concentrate on making complicated ideas and propositions comprehendible, specifically those ideas related to modeling different types of binary response data (Fig. 1). The chapters in this book are designed to help guide researchers, practitioners, and students (at the senior or Master's degree levels who have some basic experience with regression as well as some knowledge of statistical packages like SAS, SPSS, R, and STATA) in understanding binary regression models using a variety of application areas.

This book presents existing studies and recent developments in statistical methods, focusing on their applications to correlated binary data and other related research. The data and computer programs used throughout the text and analyzed

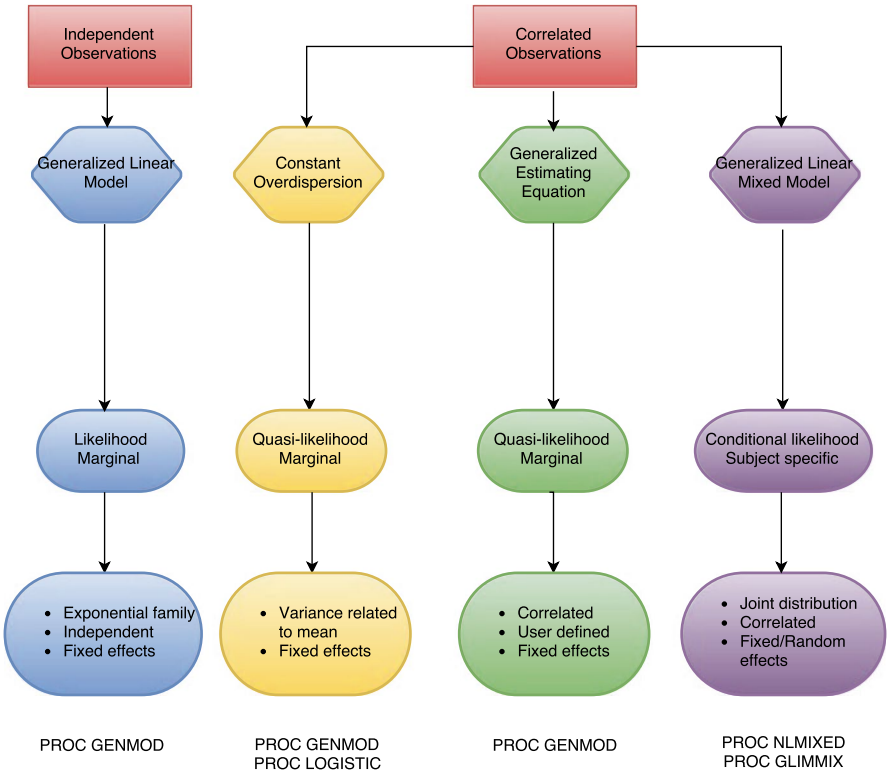

Fig. 1 Types of binary models

using SAS, SPSS, R, and STATA are publicly available so that readers can replicate the models and the results presented in each chapter. This allows the reader to easily apply the data and methods to his or her own research. The book strives to bring together in one place the key methods used in the analysis of dependent observations with binary outcomes, and presents and discusses recent issues in statistical methodological development, as well as their applications. The book is timely and has the potential to impact model development and correlated binary data analyses of health and health-related research, education, banking, and social studies, among others. In an academic setting, the book could serve as a reference guide for a course on binary data with overdispersion, particularly for students at the graduate level (master's or doctoral students) seeking degrees in related quantitative fields of study, though not necessarily in statistics. In addition, this book could serve as a reference for researchers and data analysts in education, the social sciences, public health, and biomedical research.

Each chapter consists of seven sections and is organized as follows:

Section 1: Motivating Example

1.1 Description of the Case Study
1.2 Study Hypotheses

Section 2: Definitions and Notations
Section 3: Exploratory Analyses
Section 4: Statistical Model
Section 5: Analysis of Data
Section 6: Conclusions
Section 7: Examples

The book comprises five major parts, and all of the chapters are arranged within them. Below, we provide a short summary for each of the chapters found within the four major parts of the book.

Part I: Introduction and Review of Modeling Uncorrelated Observations

1. Introduction to Binary Logistic Regression

   Statistical inference with binary data presents many challenges, whether or not the observations are dependent or independent. Studies involving dependent observations tend to be longitudinal or clustered in nature and, therefore, provide inefficient estimates if the correlation in the data is ignored. This chapter, then, reviews binary data under the assumption that the observations are independent. It provides an overview of the issues to be addressed in the book, as well as the different types of binary correlated data. It introduces SAS, SPSS, R, and STATA as the statistical programs used to analyze the data throughout the book and concludes with general recommendations.
2. Short History of the Logistic Regression Model

   The logistic regression model, as compared to the probit, Tobit, log–log, and complementary log–log models, is worth revisiting based upon the work of Cramer (2002, 2003). The ability to model the odds has made the logistic regression model a popular method of statistical analysis, in addition to the fact that the model can be used for prospective, retrospective, or cross-sectional data while the probit, Tobit, log–log, and the complementary log–log models can only be used with prospective data to model probability. This chapter provides a summary of Cramer's work (2002, 2003) and relies heavily on Cramer's own excellent but terse history of the evolution of the logistic regression model.
3. Standard Binary Logistic Regression Model

   The logistic regression model is a type of predictive model that can be used when the response variable is binary, as in the cases of: live/die, disease/no disease, purchase/no purchase, win/lose, etc. In short, we want to model the probability of getting a certain outcome by modeling the mean of the variable (which

is the same as the probability in the case of binary variables). A logistic regression model can be applied to response variables with more than two categories; however, those cases, though mentioned in this text, are less common. This chapter also addresses the fact that the logistic regression model is more effective and accurate when analyzing binary data as opposed to the simple linear regression. We will therefore present three significant problems that a researcher may encounter if the linear regression model was fitted to binary data:

1. There are no limits on the values predicted by a linear regression, so the predicted response (mean) might be less than 0 or greater than 1, which is clearly outside the realm of possible values for a response probability.
2. The variance for each subpopulation is different and therefore not constant. Since the variance of a binary response is a function of the mean, if the mean changes from subpopulation to subpopulation, the variance will also change.
3. Usually, the response is binary and so the assumption of normal distribution is not appropriate in these cases.

The chapter provides an example using cross-sectional data and a binary (two-level) response, and then fits the model in SAS, SPSS, R, and STATA. The models are based on data collected for one observation per sampling unit, and the chapter also summarizes the application to independent binary outcomes. There are several excellent texts on this topic, including Agresti (2002), which is referenced in the chapter.

Part II: Analyzing Correlated Data Through Random Component

4. Overdispersed Logistic Regression Model

When binary data are obtained through simple random sampling, the covariance of the responses follows the binomial model (two possible outcomes from independent observations with constant probability). However, when the data are obtained under other circumstances, the covariances of the responses differ substantially from the binomial case. For example, clustering effects or subject effects in repeated measure experiments can cause the variance of the observed proportions to be much larger than the variances observed under the binomial assumption. The phenomenon is generally referred to as overdispersion or extra variation. The presence of overdispersion can affect the standard errors and, therefore, also affect the conclusions made about the significance of the predictors. This chapter presents a method of analysis based on work presented in:

Wilson, J. R., & Koehler, K. J. (1991). Hierarchical models for cross-classified overdispersed multinomial data. *Journal of Business and Economic Statistics*, *9*(1), 103–110.

5. Weighted Logistic Regression Model

Binary responses, which are common in surveys, can be modeled through binary models that can provide a relationship between the probability of a response and a set of covariates. However, as explained in Chap. 4, when the data are not obtained by simple random sampling, the standard logistic regression is not valid. Rao and Scott (1984) show that when the data come from a complex survey designed with stratification, clustering, and/or unequal weighting, the usual estimates are not appropriate. In these cases, specialized techniques must be applied in order to produce the appropriate estimates and standard errors. Clustered data are frequently encountered in fields such as health services, public health, epidemiology, and education research. Data may consist of patients clustered within primary care practices or hospitals, or households clustered within neighborhoods, or students clustered within schools. Subjects nested within the same cluster often exhibit a greater degree of similarity, or homogeneity of outcomes, compared to randomly selected subjects from different clusters (Austin et al., 2001; Goldstein, 1995; Kreft & De Leeuw, 1998; Raudenbush & Bryk, 2002; Snijders & Boskers, 1999). Due to the possible lack of independence of subjects within the same cluster, traditional statistical methods may not be appropriate for the analysis of clustered data. While Chap. 4 uses the overdispersed logistic regression and the exchangeability logistic regression model to fit correlated data, this chapter incorporates a series of weights or design effects to account for the correlation. The logistic regression model on the analysis of survey data takes into account the properties of the survey sample design, including stratification, clustering, and unequal weighting. The chapter fits this model in SAS, SPSS, and R, using methods based on:

Koehler, K. J., & Wilson, J. R. (1986). Chi-square tests for comparing vectors of proportions for several cluster samples. *Communications in Statistics*, *15* (10), 2977–2990.
Wilson, J. R. (1986). Approximate distribution and test of fit for the clustering effect in Dirichlet multinomial model. *Communications in Statistics*, *15*(4), 1235–1249.
Wilson, J. R. (1989). Chi-square tests for overdispersion with multiparameter estimates. *Journal of Royal Statistics Society Series C, Applied Statistics*, *38*(3), 441–454.

6. Generalized Estimating Equation Logistic Regression Model

Many fields of study use longitudinal datasets, which usually consist of repeated measurements of a response variable, often accompanied by a set of covariates for each of the subjects/units. However, longitudinal datasets are problematic because they inherently show correlation due to a subject's repeated set of measurements. For example, one might expect a correlation to exist when looking at a patient's health status over time or a student's performance over time. But in those cases, when the responses are correlated, we cannot readily obtain the underlying joint distribution; hence, there is no closed-form joint likelihood function to present, as with the standard logistic regression model. One remedy is to fit a generalized estimating equations (GEE) logistic regression model for the data, which is explored in this chapter. This chapter addresses repeated measures of the sampling unit, showing how the GEE method allows missing values within a subject without losing all the data from the subject, and

time-varying predictors that can appear in the model. The method requires a large number of subjects and provides estimates of the marginal model parameters. We fit this model in SAS, SPSS, R, and STATA, basing our work on the method best presented by Ziang and Leger (1986), and Liang and Zeger (1986).

7. Generalized Method of Moments Logistic Regression Model

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 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 as follows:

1. There is the correlation among the responses which are denoted by $y_1, \ldots, y_T$ as time t goes from 1 to T.
2. There is 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 $t + s$

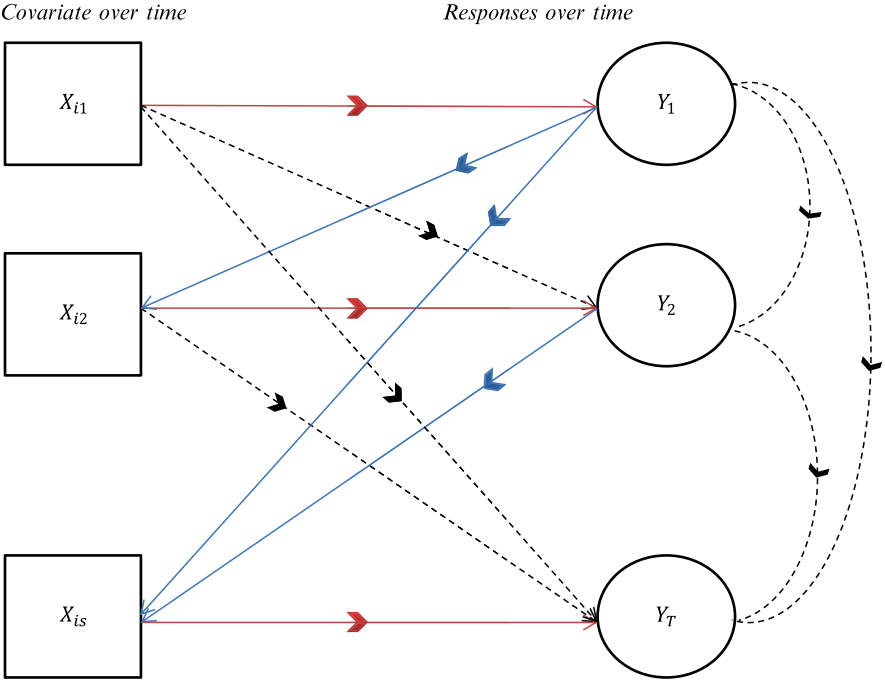

Fig. 2 Two types of correlation structures

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 chapter 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 with SAS Macro (Cai & Wilson, 2015) using methods based on:

> LaLonde, T., Wilson, J. R., & Yin, J. (2014). GMM logistic regression models for longitudinal data with time-dependent covariates and extended classifications. *Statistics in Medicine*, *33*(27).

8. Exact Logistic Regression Model

As computers' abilities to do tedious calculations have increased, using exact logistic regression models has become more popular in healthcare, banking, and other industries. Traditional methods (which are based on asymptotic theory) when used for analyzing small, skewed, or sparse datasets are not usually reliable. When sample sizes are small or the data are sparse or skewed, exact conditional inference is necessary and applicable (Derr, 2008). Exact methods of inferences are based on enumerating the exact distributions of certain statistics to estimate the parameters of interest in a logistic regression model, conditional on the remaining parameters. This is a method of testing and estimation that uses conditional methods to obtain exact tests of parameters in binary and nominal logistic models. Exact methods are appropriate for small-sample or sparse data situations that often result in the failure (nonconvergence or separation) of the usual unconditional maximum likelihood estimation method. However, exact methods can take a great deal of time and memory as sample or model sizes increase. For sample sizes too large for the default exact method, a Monte Carlo method is provided. The chapter uses EXACT statement in PROC LOGISTIC or PROC GENMOD, and we also fit models in SAS, C+, and R. Our methods are based on:

> Troxler, S., Lalonde, T. L., & Wilson, J. R. (2011). Exact logistic models for nested binary data. *Statistics in Medicine*, *30*(8).

Part III: Analyzing Correlated Data Through Systematic Components

9. Two-Level Nested Logistic Regression Model

Studies including repeated measures are expected to give rise to correlated data. Such data are common in many disciplines including healthcare, banking, poll tracking, and education. Subjects or units are followed over time and are repeatedly observed under different experimental conditions, or are observed in clusters. Often times, such data are available in hierarchical structures consisting of a subset of a population of units at several levels. We review methods that include the clustering directly in the model (systematic component) as opposed

to methods that include the clustering within the random component. These methods belong to the class of generalized linear mixed models. The basic idea behind generalized linear mixed models is conceptually straightforward (McCulloch, 2003) and incorporates random effects into the systematic component of a generalized linear model to account for the correlation. Such approaches are most useful when researchers wish to account for both fixed and random effects in the model. The desire to address the random effects in a logistic model makes it a subject-specific model. This is a conditional model that can also be used to model longitudinal or repeated measures data. We fit this model in SAS, SPSS, and R. Our method of modeling is based on:

> Lalonde, T., Nguyen, A. Q., Yin, J., Irimata, K., & Wilson, J. R. (2013). Modeling correlated binary outcomes with time-dependent covariates. *Journal of Data Science*, *11*(4), 715–738

10. Hierarchical Logistic Regression Model

    This chapter expands upon the results of Chap. 9. It is common to come into contact with data that have a hierarchical or clustered structure. Examples include patients within a hospital, students within a class, factories within an industry, or families within a neighborhood. In such cases, there is variability between the clusters, as well as variability between the units, which are nested within the clusters. Hierarchical models take into account the variability at each level of the hierarchy and thus allow for the cluster effects at different levels to be analyzed within the models (Shahian et al., 2001). This chapter tells how one can use the information from different levels to produce a subject-specific model. We concentrate on fitting logistic regression models to these kinds of nested data at three levels and higher (Fig. 3). In Fig. 3, as an example, patients are nested within doctors and doctors are nested within hospitals. This is a three-level nested design but can be expanded to higher levels, though readily available computing may be challenge.
11. Fixed Effects Logistic Regression Model

    If a researcher wants to know whether having a job reduces recidivism among chronic offenders, that researcher could compare an individual's arrest rate when he/she is employed with his/her arrest rate when unemployed.

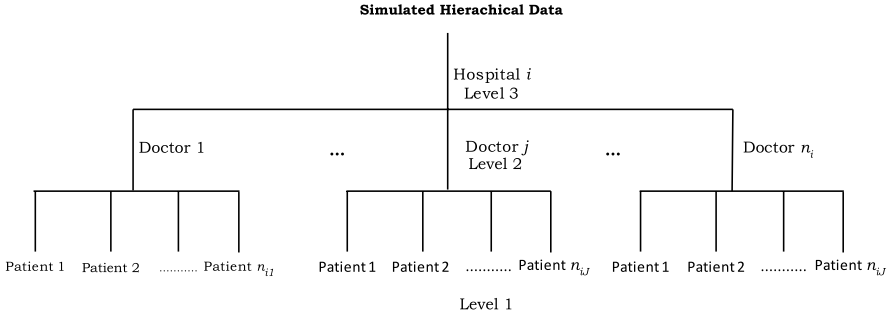

Fig. 3 Hierarchical structure of three levels

The difference in arrest rates between the two periods is an estimate of the employment effect for that individual. Similarly, a researcher might want to know how a child's performance in school differs depending on how much time he/she spends watching television. The researcher could compare how the child does when spending significant time watching television versus when he/she does not watch television. Fixed effects logistic regression models can be used for both of these scenarios. Such models are used to analyze longitudinal data with repeated measures on both the response and the covariates. These models treat each measurement on each subject as a separate observation, and the set of subject coefficients that would appear in an unconditional model are eliminated by conditional methods. This is a conditional, subject- specific model (as opposed to a population-averaged model like the GEE model). We fit this model in SAS, SPSS, and R. An excellent discussion with examples can be found in P. D. Allison (2005), *Fixed Effects Regression Methods for Longitudinal Data Using SAS*. For binary response data, we use the STRATA statement in PROC LOGISTIC.

Part IV: Analyzing Correlated Data Through the Joint Modeling of Mean and Variance

12. Heteroscedastic Logistic Regression Model

    Correlated binomial data can be modeled using a mean model if the interest is only on the mean and the dispersion is considered a nuisance parameter. However, if the intraclass correlation is of interest, then there is a need to apply a joint modeling of the mean and the dispersion. Efron (1986) was one of the first to model both the mean and the variance. The dispersion sub-model allows extra parameters to model the variance independent of the mean, thus allowing covariates to be included in both the mean and variance sub-models. In this chapter, we present a sub-model that analyzes the mean and a sub-model that analyzes the variance. This model allows both the dispersion and the mean to be modeled. We use the MODEL statement in the SAS/ETS procedure QLIM to specify the model for the mean and use the HETERO statement to specify the dispersion model. We fit this model in SAS and SPSS. Our results and presentation are based on work done in some recent masters' research papers at Arizona State University.

Part V: Case Studies: USA Election 2020 and Netherlands COVID

13. Case Studies: Analysis of Election Data and Covid Data

    We take a practitioner view of 2020 Election Data in the United States of America and address certain questions with views to the 2024 election. We also did a similar approach to COVID data as they data were collected over time.

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: Anh Nguyen, who provided the graphics and had a lot to do with extracting and analyzing the Medicare dataset in the initial stages; Hong Xiang, who, through her master's applied paper, contributed to findings regarding PROC NLMIXED and hierarchical analyses; Jianqiong Yin, who provided insight and unwavering contributions to our statistical programming through her thesis and associated work; Katherine Cai, who helped with SAS Macro; and Chad Mehalechko, who provided overwhelming support in doing the SAS, SPSS, and R programming for all chapters. Many thanks to the staff in the Department of Economics and the computing support group in the W. P. Carey School of Business. To everyone involved in the making of this book, we say thank you!

Finally, a special thanks to our families, who have provided us with the support needed to achieve this great endeavor.

Tempe, AZ, USA — Jeffrey R. Wilson
San Francisco, CA, USA — Kent A. Lorenz
Tempe, AZ, USA — Lori P. Selby

Contents

Part I Introduction and Review of Modeling Uncorrelated Observations

1 Introduction to Binary Logistic Regression 3
1.1 Motivating Example. 3
1.2 Definition and Notation 4
1.2.1 Notation 4
1.2.2 Definitions 4
1.3 Exploratory Analyses. 6
1.4 Statistical Models. 8
1.4.1 Chapter 3: Standard Binary Logistic Regression Model . . . 8
1.4.2 Chapter 4: Overdispersed Logistic Regression Model 8
1.4.3 Chapter 5: Survey Data Logistic Regression Model 8
1.4.4 Chapter 6: Generalized Estimating Equations (GEE) Logistic Regression Model 9
1.4.5 Chapter 7: Generalized Method of Moments (GMM) Logistic Regression Model 9
1.4.6 Chapter 8: Exact Logistic Regression Model 9
1.4.7 Chapter 9: Two-Level Nested Logistic Regression Model. 9
1.4.8 Chapter 10: Hierarchical Logistic Regression Model 10
1.4.9 Chapter 11: Fixed Effects Logistic Regression Model 10
1.4.10 Chapter 12: Heteroscedastic Logistic Regression Model . . 10
1.4.11 Chapter 13: Case Studies—Election Data and COVID Data 10
1.5 Analysis of Data. 12
1.5.1 SAS Programming. 13
1.5.2 SPSS Programming 13
1.5.3 R Programming 13
1.5.4 STATA Programming. 14
1.5.5 Fitting Models 14
1.6 Conclusions 15

1.7 Related Examples . . . 15
1.7.1 Medicare Data . . . 16
1.7.2 Philippines Data . . . 16
1.7.3 Household Satisfaction Survey . . . 16
1.7.4 NHANES: Treatment for Osteoporosis . . . 17
1.7.5 COVID Data . . . 17
1.7.6 Election Data . . . 17
References . . . 18

2 Short History of the Logistic Regression Model . . . 19
2.1 Motivating Example . . . 19
2.2 Definition and Notation . . . 20
2.2.1 Notation . . . 20
2.2.2 Definition . . . 20
2.3 Exploratory Analyses . . . 21
2.4 Statistical Model . . . 21
2.5 Analysis of Data . . . 24
2.6 Conclusions . . . 24
References . . . 24

3 Standard Binary Logistic Regression Model . . . 27
3.1 Motivating Example . . . 27
3.1.1 Study Hypotheses . . . 27
3.2 Definition and Notation . . . 28
3.3 Exploratory Analyses . . . 30
3.4 Statistical Models . . . 33
3.4.1 Probability . . . 33
3.4.2 Odds . . . 34
3.4.3 Logits . . . 34
3.4.4 Logistic Regression Versus Ordinary Least Squares . . . 34
3.4.5 Generalized Linear Models (GLMs) . . . 35
3.4.6 Response Probability Distributions . . . 36
3.4.7 Log-Likelihood Functions . . . 36
3.4.8 Maximum Likelihood Fitting . . . 37
3.4.9 Goodness-of-Fit . . . 37
3.4.10 Other Fit Statistics . . . 38
3.4.11 Assumptions for Logistic Regression Model . . . 38
3.4.12 Interpretation of Coefficients . . . 38
3.4.13 Interpretation of Odds Ratio (OR) . . . 38
3.4.14 Model Fit . . . 39
3.4.15 Null Hypothesis . . . 40
3.4.16 Predicted Probabilities . . . 40
3.4.17 Computational Issues Encountered with Logistic Regression . . . 41

3.5 Analysis of Data . . . 42
3.5.1 Medicare Data . . . 42
3.5.2 Analysis of Medicare Data with SAS Computing . . . 43
3.5.3 Analysis of Medicare Data with SPSS Computing . . . 51
3.5.4 Analysis of Medicare Data with R Computing . . . 54
3.5.5 Analysis of Medicare Data with STATA Computing . . . 55
3.6 Conclusion . . . 56
3.7 Related Examples . . . 57
Appendix . . . 58
References . . . 59

Part II Analyzing Correlated Data Through Random Component

4 Overdispersed Logistic Regression Model . . . 63
4.1 Motivating Example . . . 63
4.2 Definition and Notation . . . 63
4.3 Exploratory Data Analyses . . . 64
4.4 Statistical Model . . . 65
4.4.1 Williams' Method of Analysis . . . 67
4.4.2 Overdispersion Factor . . . 68
4.4.3 Datasets . . . 69
4.4.4 Housing Satisfaction Survey . . . 69
4.5 Analysis of Data . . . 69
4.5.1 Standard Logistic Regression Model . . . 70
4.5.2 Overdispersed Logistic Regression Model . . . 73
4.5.3 Overdispersed Logistic Regression Model Using SAS Program . . . 74
4.5.4 Overdispersed Logistic Regression Model Using R Program . . . 77
4.5.5 Exchangeability Logistic Regression Model . . . 78
4.5.6 Exchangeability Logistic Regression Model Using SAS Program . . . 78
4.5.7 Exchangeability Logistic Regression Model Using R Program . . . 81
4.6 Conclusion . . . 82
4.7 Related Example . . . 83
4.7.1 Use of Word Einai . . . 83
References . . . 83

5 Weighted Logistic Regression Model . . . 85
5.1 Motivating Example . . . 85
5.2 Definition and Notation . . . 85
5.3 Exploratory Analyses . . . 87
5.3.1 Treatment for Osteoporosis . . . 87
5.4 Statistical Model . . . 89

5.5 Analysis of Data . . . 90
5.5.1 Weighted Logistic Regression Model with Survey Weights . . . 91
5.5.2 Weighted Logistic Regression Model with Survey Weights Using SAS Program . . . 91
5.5.3 Weighted Logistic Regression Model with Survey Weights Using SPSS Program . . . 94
5.5.4 Weighted Logistic Regression Model with Survey Weights Using R Program . . . 99
5.5.5 Weighted Logistic Regression Model with Strata and Clusters Identified . . . 101
5.5.6 Comparison of Weighted Logistic Regression Models . . . 104
5.6 Conclusion . . . 104
5.7 Related Examples . . . 105
References . . . 105

6 Generalized Estimating Equations Logistic Regression . . . 107
6.1 Motivating Example . . . 107
6.1.1 Description of the Rehospitalization Issues . . . 107
6.2 Definition and Notation . . . 108
6.3 Exploratory Analyses . . . 110
6.4 Statistical Models: GEE Logistic Regression . . . 113
6.4.1 Medicare Data . . . 113
6.4.2 Generalized Linear Model . . . 113
6.4.3 Generalized Estimating Equations . . . 114
6.4.4 Marginal Model . . . 115
6.4.5 Working Correlation Matrices . . . 115
6.4.6 Model Fit . . . 116
6.4.7 Properties of GEE Estimates . . . 116
6.5 Data Analysis . . . 117
6.5.1 GEE Logistic Regression Model . . . 117
6.5.2 GEE Logistic Regression Model with SAS Programming . . . 117
6.5.3 GEE Logistic Regression Model with SPSS Programming . . . 121
6.5.4 GEE Logistic Regression Model with R Programming . . . 123
6.5.5 GEE Logistic Regression Model with STATA Programming . . . 124
6.6 Conclusion . . . 125
6.7 Related Examples . . . 126
References . . . 127

7 Generalized Method of Moments Logistic Regression Model . . . 129
7.1 Motivating Example . . . 129
7.1.1 Description of the Case Study . . . 129
7.1.2 Study Hypotheses . . . 130

7.2 Definition and Notation 130
7.3 Exploratory Analyses 131
7.4 Statistical Model 133
7.4.1 GEE Models for Time-Dependent Covariates 135
7.4.2 Lai and Small GMM Method 135
7.4.3 Types of Classification of Time-Dependent Covariates 136
7.4.4 Lalonde Wilson and Yin Method 138
7.5 Analysis of Data 139
7.5.1 Modeling Probability of Rehospitalization 139
7.5.2 Modeling Probability of Rehospitalization SAS: Results 139
7.6 Conclusions 142
7.7 Related Examples 143
References 143

8 Exact Logistic Regression Model 145
8.1 Motivating Example 145
8.2 Definition and Notation 146
8.3 Exploratory Analysis 147
8.3.1 Artificial Data for Clustering 147
8.3.2 Standard Logistic Regression 147
8.3.3 Two-Stage Clustered Data 148
8.4 Statistical Models 149
8.4.1 Independent Observations 149
8.4.2 One-Stage Cluster Model 149
8.4.3 Two-Stage Cluster Exact Logistic Regression Model 151
8.5 Analysis of Data 152
8.5.1 Exact Logistic Regression for Independent Observations . . 153
8.5.2 Exact Logistic Regression for One-Stage Clustered Data . . 153
8.5.3 Exact Logistic Regression for Independent Observations with R Programming 159
8.5.4 Exact Logistic Regression for One-Stage Clustered Data with R Program 159
8.5.5 Exact Logistic Regression for One-Stage Clustered Data with C++ Program 160
8.5.6 Exact Logistic Regression for Two-Stage Clustered Data with C++ Program 160
8.6 Conclusions 161
8.7 Related Examples 161
8.7.1 Description of the Data 162
8.7.2 Clustering 162
References 162

Part III Analyzing Correlated Data Through Systematic Components

9 Two-Level Nested Logistic Regression Model 165
9.1 Motivating Example. 165
9.1.1 Description of the Case Study 165
9.1.2 Study Hypotheses. 165
9.2 Definition and Notation 166
9.3 Exploratory Analyses 168
9.3.1 Medicare 169
9.4 Statistical Model 170
9.4.1 Marginal and Conditional Models 171
9.4.2 Two-Level Nested Logistic Regression with Random Intercept Model 172
9.4.3 Interpretation of Parameter Estimates 173
9.4.4 Two-Level Nested Logistic Regression Model with Random Intercept and Slope 174
9.5 Analysis of Data. 175
9.5.1 Two-Level Nested Logistic Regression Model with Random Intercepts Using SAS (PROC NLMIXED Versus PROC GLIMMIX) 175
9.5.2 Two-Level Nested Logistic Regression Model with Random Intercepts Using SPSS. 186
9.5.3 Two-Level Nested Logistic Regression Model with Random Intercepts Using R. 189
9.5.4 Two-Level Nested Logistic Regression Model with Random Intercepts Using STATA 190
9.5.5 Two-Level Nested Logistic Regression Model Random Intercept and Slope with SAS 191
9.5.6 Two-Level Nested Logistic Regression Model Random Intercept and Slope with SPSS 195
9.5.7 Two-Level Nested Logistic Regression Model Random Intercept and Slope with STATA 198
9.6 Conclusions 198
9.7 Related Examples. 199
9.7.1 Multicenter Randomized Controlled Data (Beitler & Landis, 1985) 199
References. 200

10 Hierarchical Logistic Regression Models 201
10.1 Motivation 201
10.1.1 Description of Case Study 201
10.1.2 Study Hypotheses. 202
10.2 Definitions and Notations 202
10.3 Exploratory Analyses 203

10.4 Statistical Model 204
10.4.1 Multilevel Modeling Approaches with Binary Outcomes 205
10.4.2 Potential Problems 205
10.4.3 Three-Level Logistic Regression Models with Multiple Random Intercepts 206
10.4.4 Three-Level Logistic Regression Models with Random Intercepts and Random Slopes 207
10.4.5 Nested Higher Level Logistic Regression Models. 209
10.4.6 Cluster Sizes and Number of Clusters 209
10.4.7 Parameter Estimations 209
10.5 Analysis of Data. 210
10.5.1 Modeling Random Intercepts for Levels 2 and 3. 210
10.5.2 Modeling Random Intercepts for Levels 2 and 3. 210
10.5.3 Modeling Random Intercepts for Levels 2 and 3 Using STATA 214
10.5.4 Three-Level Logistic Regression Model with Random Slopes Using SAS 215
10.5.5 Modeling Random Intercepts for Levels 2 and 3 Using R 217
10.5.6 Three-Level Logistic Regression Model with Random Slopes Using R. 219
10.5.7 Three-Level Logistic Regression Model with Random Slopes at Doctor Level Using STATA 221
10.5.8 Interpretation 222
10.6 Conclusions 223
10.7 Related Examples. 224
References. 224

11 Fixed Effects Logistic Regression Model 227
11.1 Motivating Example. 227
11.2 Definition and Notation 228
11.3 Exploratory Analysis 228
11.3.1 Philippine's Data 228
11.4 Statistical Models. 229
11.4.1 Fixed Effects Regression Models with Two Observations Per Unit 230
11.4.2 Modeling More Than Two Observations Per Unit: Conditional Logistic Regression Model 232
11.5 Analysis of Data. 233
11.5.1 Fixed Effects Logistic Regression Model with Two Observations Per Unit 233
11.5.2 Fixed Effects Logistic Regression Model with Two Observations Per Unit Using SAS 234
11.5.3 Fixed Effects Logistic Regression Model with Two Observations Per Unit Using SPSS 237

11.5.4 Fixed Effects Logistic Regression Model with Two Observations Per Unit Using R 239
11.5.5 Fixed Effects Logistic Regression Model with Two Observations Per Unit Using STATA 240
11.6 Fixed Effects Logistic Regression Model with More Than Two Observations 241
11.6.1 Fixed Effects Logistic Regression Model with More Than Two Observations Using SAS 242
11.6.2 Fixed Effects Logistic Regression Model with More Than Two Observations Using SPSS 245
11.6.3 Fixed Effects Logistic Regression Model with More Than Two Observations Using R 247
11.6.4 Fixed Effects Logistic Regression Model with More Than Two Observations Using STATA 249
11.7 Conclusions 249
11.8 Related Examples 250
References 250

Part IV Analyzing Correlated Data Through the Joint Modeling of Mean and Variance

12 Heteroscedastic Logistic Regression Model 253
12.1 Motivating Example 253
12.2 Definitions and Notations 254
12.3 Exploratory Analyses 255
12.3.1 Dispersion Sub-model 257
12.4 Statistical Model 258
12.4.1 Joint Modeling 259
12.5 Analysis of Data 261
12.5.1 Heteroscedastic Logistic Regression Model 261
12.5.2 Standard Logistic Regression Model 264
12.5.3 Model Comparisons Mean Sub-model Versus Joint Modeling 265
12.6 Conclusions 266
12.7 Related Examples 266
12.7.1 Logistic Predicted 267
References 268

Part V Case Studies- USA Election 2020 and Netherlands COVID

13 Case Studies: Election Data and COVID Data 273
13.1 Two Case Studies- Election and COVID 273
13.2 USA 2020 Election Data 273
13.2.1 Election Data Questions 273
13.2.2 Election Variables and Data Analysis 274
13.2.3 Election Data Interpretation 276

13.3 The Netherlands 2020–2021 COVID Data 277
13.3.1 COVID Data Questions 278
13.3.2 COVID Data Variables and Analysis 278
13.3.3 COVID Data Interpretation 279
References 280

Index 281

Part I
Introduction and Review of Modeling Uncorrelated Observations

Chapter 1
Introduction to Binary Logistic Regression

1.1 Motivating Example

The need to profile or describe a unit/subject based on a binary outcome is often of utmost importance. While there are other models (e.g., probit, log-log, complementary log-log) that can be used to model binary responses, in this book, we concentrate on logistic regression models. The advantage of the logistic regression model lies in the model's ability to explain and predict the outcomes in terms of the odds and how applicable it is to cases where the data are obtained from prospective, retrospective, or cross-sectional settings. For example, we wish to detect the factors which determine whether patient's cancer is in remission. In that simulated study data, there is a three-level hierarchical structure with patients nested within doctors within hospitals. The study is part of a large study of lung cancer outcomes across multiple doctors and sites. The response variable is whether or not the cancer is in remission. The covariates at the patient level are age, length of stay, family history, and cancer stage. At the doctor's level, we have their identification and their years of experience. At the hospital level, we have the hospital identification.

This book deals with the topic of logistic regression models concentrating on binary data. The statement leads naturally to the following questions:

- What is a binary logistic regression model?
- What do we mean by binary data?
- Why do we need logistic regression models?
- Why is it not adequate to use the standard regression analysis to analyze binary data?

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024

J. R. Wilson et al., *Modeling Binary Correlated Responses*, ICSA Book Series in Statistics, https://doi.org/10.1007/978-3-031-62427-8_1

1.2 Definition and Notation

1.2.1 Notation

Let N denote the number of units in the population, while n denotes the sample size, and T times $t = 1, \ldots T$; with J covariates $j = 1, \ldots J$.

1.2.2 Definitions

A *variable* is defined as any measure that varies from unit to unit (i.e., individual to individual or subject to subject). For example, if we measured the blood pressure for 10 individuals, we would expect the values to be different, making blood pressure the "variable" of interest. Another example of a variable would be measuring the ages of 8000 Japanese-American men to analyze the variable of age.

The *output* variable Y is also called the *response* variable, or *dependent* variable or *outcome* variable or *target* variable. It is the variable of interest.

A *quantitative* variable is an individual measure taken on a numerical scale. It is often referred to as a continuous variable, meaning that these observations may lie on a continuum. Thus, a *continuous* variable takes on an infinite number of possible values, and its reported value is limited only by physical measurement accuracy. We will make references to quantitative variables measured on two scales as follows:

An *interval scale* pertains to variables that can be ordered and give precise measures of the distances between categories. Measuring temperature in degrees Celsius is an example of this type of quantitative variable.

A *ratio scale* is for ratio variables, which are interval variables with a true zero value. Using temperature as our example, a reading of zero degrees on the Celsius scale does not actually denote an absence of heat or energy. Therefore, we classify it as an interval variable. However, zero degrees on the Kelvin scale refer to an absolute zero, or an absence of temperature or energy. Therefore, we classify it as a ratio variable. A key difference between interval and ratio scales is that, with ratio scales, the distance between measurements may be equal (i.e., the difference between 10 °C and 20 °C is the same as between 25 °C and 35 °C. However, 20 °C is not twice as much "temperature" as 10 °C. Only when using a ratio scale could you make those statements (i.e., 20 K is twice as much as 10 K)).

A *qualitative* variable is one whose values vary in type rather than in magnitude. It expresses a qualitative attribute, meaning that its values are not numerical. The observations cannot be said to vary in magnitude, and a natural hierarchical ordering is not possible.

A binary random variable is usually symbolized or coded as a 1 or 0. It is appropriate when the assessment of a characteristic is in terms of any two categories such as yes/no, present/absent, and favorable/unfavorable. These are called categories. The corresponding variable is called *binary* or *dichotomous* or *indicator*. The coded

values do not have any meaning or rank, so they cannot be interpreted numerically. However, several statistical programs require one or zero coding for analysis. With a zero or one coded value, a researcher can find the mean. Such a summary value is referred to as the proportion. Examples include: "Do you smoke?" Proportion: those who smoke. Or "Do you drink?" Proportion: those who drink. Thus, the mean of a binary variable is called the proportion.

A *categorical* variable is a variable measured on a nominal scale whose categories determine class or group membership. A *categorical* variable can take "values" which denote nonnumerical categories or classes. Some definitions are: A *categorical* variable represents a set of discrete events, such as groups, decisions, or anything else that can be classified into categories. Categorical variables are also referred to as discrete variables, meaning a limited set of numerical values that are typically counting numbers or whole numbers. Two main types of categorical variables are nominal and ordinal variables.

Ordinal variable consists of a rank, or a rating. One example would be: "How was the performance?" Some possible responses on a measurement scale could be: 1 for "Excellent," 2 for "Above Average," 3 for "Average," 4 for "Below Average," and 5 for "Poor." There is an obvious limitation to this type of measurement: The examiner cannot be assured of the preciseness of a measurement like satisfaction. In this example, "Excellent" is higher than "Above Average," but it is not very precise because it does not show how much higher an "Excellent" rating is compared to an "Above Average" rating. This limitation also means that there can be no guarantee that a difference between a score of 1 and 2 is the same as a difference between a score of 4 and 5.

Nominal variable has characteristics that are mutually exclusive and exhaustive. This means that nominal variables can only be measured in terms of whether an individual belongs to a certain category. Some examples are: "What region of the country are you from (Northeast, Northwest, Midwest, etc.)?" or "Are you single, married, widowed, or divorced?"

Categorical Variable in the Form of a Series of Binary Variables

Categorical variables can be reformulated and treated like a set of binary variables.

Consider the categorical variable "Region," with levels (East, West, North, and South).

When converted to four binary variables representing "Region," it looks like the following:

1. East with levels (yes, no)
2. West with levels (yes, no)
3. North with levels (yes, no)
4. South with levels (yes, no)

Example: Consider the following data given about "Region" in Table 1.1.