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:

- 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 Ph.D. levels at Arizona State University. In fact, several of the chapters are based on the applied projects and theses of Master's and Ph.D. students in the university's statistics programs. The examples in this book were analyzed whenever possible using SAS, SPSS, and R. While the SAS, SPSS, and R outputs are contained in the text with partial data tables, the completed datasets can be found at the web address 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, and R) 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

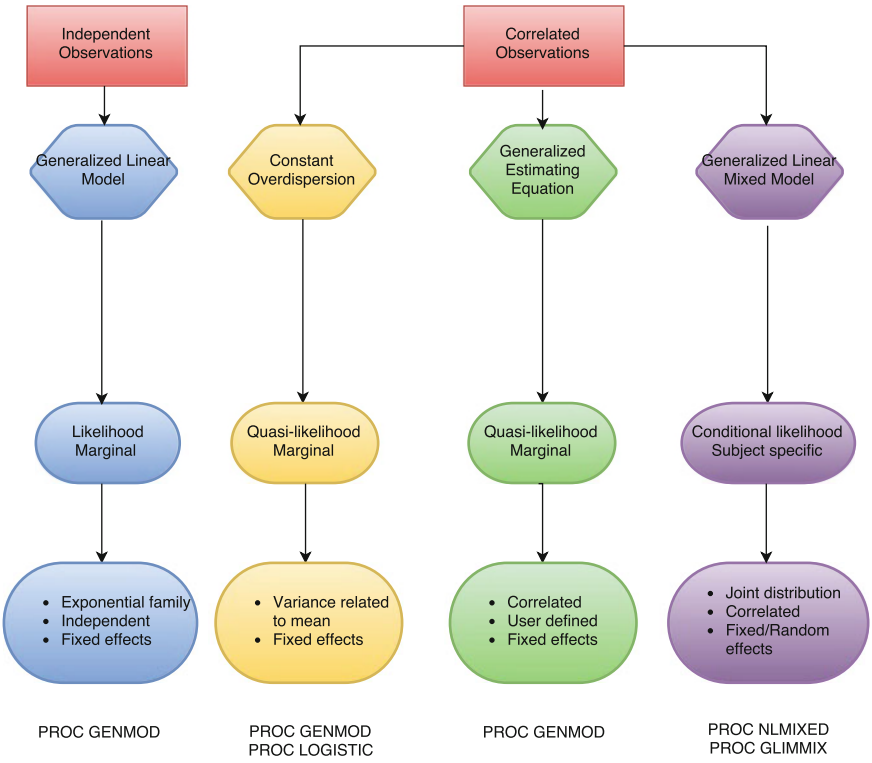

Fig. 1 Types of binary models

research. The data and computer programs used throughout the text and analyzed using SAS, SPSS, and R 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 present and discuss 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 four 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, and R 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, and R. 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, A15* (10), 2977–2990.

Wilson, J. R. (1986). Approximate distribution and test of fit for the clustering effect in Dirichlet multinomial model. *Communications in Statistics, A15*(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 Equations Logistic Regression
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, and R, 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

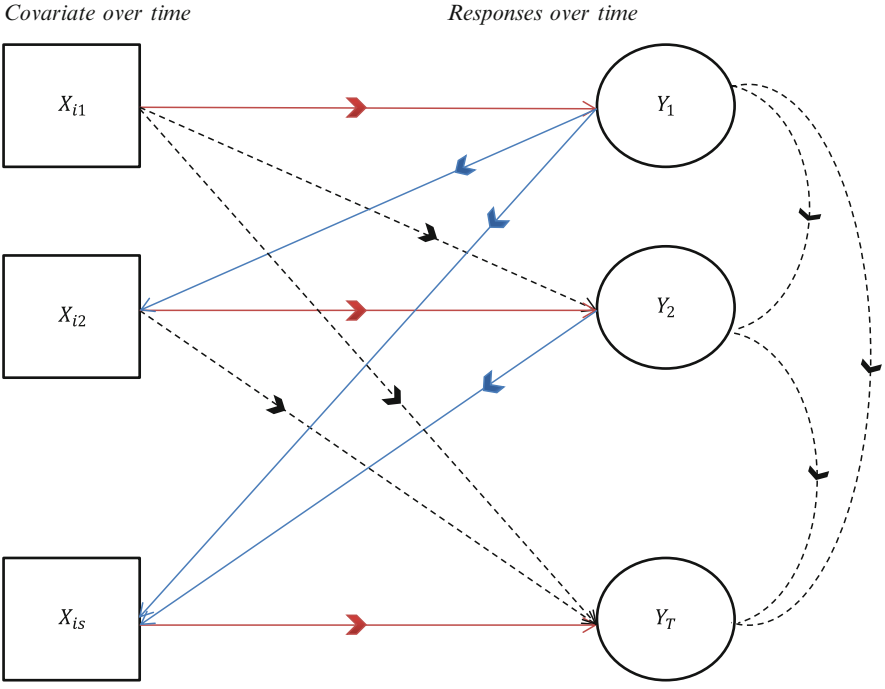

Fig. 2 Two types of correlation structures

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:

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 and
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 $t+s$.

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.

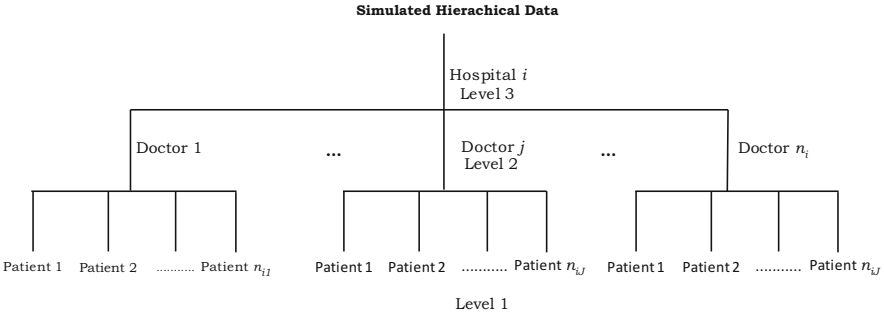

Fig. 3 Hierarchical structure of three levels

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.

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 both of us with the support needed to achieve this great endeavor.

Tempe, AZ — Jeffrey R. Wilson
Phoenix, AZ — Kent A. Lorenz

http://www.springer.com/978-3-319-23804-3

Modeling Binary Correlated Responses using SAS, SPSS and R

Wilson, J.R.; Lorenz, K.A.

2015, XXIII, 264 p. 26 illus., Hardcover
ISBN: 978-3-319-23804-3