The Impact of Predictor Variable(s) with Skewed Cell Probabilities on the Wald Test in Binary Logistic Regression

by

Arwa A. Alkhalaf

B.Sc., King Abdulaziz University, 2006
MA, The University of British Columbia, 2012

A DISSERTATION SUBMITTED IN PARTIAL FULLFILMENT OF THE REQUIRMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

The Faculty of Graduate and Postdoctoral Studies

(Measurement, Evaluation and Research Methodology)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

April 2017

Abstract

What happens to the parameter estimates and test operating characteristics when the predictor variables in a logistic regression are skewed? The statistics literature provides relatively few answers to this question. A series of simulation studies are reported that investigated the impact of a skewed predictor (s) on the Type I error rate and power of the Wald test in a logistic regression model. Five simulations were conducted for three different models: a simple logistic regression with a binary predictor, a simple logistic regression with a continuous predictor, and a multiple logistic regression with two dichotomous predictors. The results show that the Type I error rate and power were affected by severe predictor skewness, but that the effect was moderated by sample size. The Type I error rate was consistently deflated for all three models. Also, power improved with less skewness. A detailed description of the impact of skewed cell predictor probabilities and sample size provide guidelines for practitioners as to where to expect the greatest problems. These findings highlight the importance of the effects of predictor characteristics on statistical analysis of a logistic regression.

Preface

This dissertation is original work done by the author Arwa Alkhalaf. I was the lead investigator responsible for all major ideas of concept formation, simulation design and data analysis, as well as the manuscript composition. During my program and the creation of this dissertation I was mentored and supervised by Dr. Bruno Zumbo, who was involved throughout this dissertation in concept formation and dissertation edits. A version of Chapter 3 is currently under review in a journal.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to express my deepest gratitude and appreciation to my advisor Professor Bruno Zumbo, who is not only a brilliant teacher but a remarkable intellectual philosopher. He has continuously challenged me to reach my full potential to become a better scientist. He has, also, supported me in many ways. Without his guidance and persistent help this dissertation would not have been possible.

My thanks must also be extended to all my family and friends, who have been there for me through thick and thin. To my parents, whom their prayers have guided my way. To my dearest friends who have been an invaluable support system.

Dedication

To my companion in this journey, my dearest son, Faris.

Chapter 1: Introduction

Real-life data rarely satisfy the assumptions of most statistical methods. Modified methods that perform well when these assumptions are not met are therefore crucial to achieving sound results. Known as robust statistics, these theories and techniques allow statisticians to estimate the characteristics of a parametric model while dealing with deviations from idealized conditions. Examples of deviations include the contamination of data by outliers, rounding and grouping errors, and the departure from an assumed sample distribution. Therefore, important questions to be considered are when and under what conditions do conventional approaches breakdown.

In computational statistics and applied analysis, researchers focus on finding robust estimators and tests that can provide mathematically sound solutions in a logistic regression (e.g., Ahmad, Ramli, & Midi, 2010; Bianco & Martínez, 2009). Some extensively tested and documented issues with logistic regression are separation, a low number of events per predictor, non-normal residuals, and outlying observations.

Logistic regression models are widely used in the medical and behavioural sciences and describe the effect of predictor variables on a dichotomous outcome variable. The logistic regression model assumes independent outcome variables that follow a Bernoulli distribution, with the probability of a positive response modeled as

$$P(y_i = 1|X = x_i) = F(x_i^T \beta),$$

where F is the logistic link function, $x_i \in R^P$ are the predictor variable vectors, $\beta \in R^P$ is the unknown parameter vector, $i = 1, \dots, j < n$, and $P = i$. Such models are usually estimated by the maximum likelihood estimator (MLE), which is very sensitive to violations in sample size and

normality.

There are a few assumptions in logistic regression related to the independence of observations and unassociated predictors. Testing violations of logistic regression assumptions is not common practice, simply because the aforementioned assumptions are usually assumed to be met. However, post-hoc diagnostics are an essential part of the analysis. Diagnostics in generalized linear models are based on the assumptions that the residuals are normally distributed (Pergibon, 1981). The robust statistics literature, however, does not articulately describe how the violation of an assumption occurs, and how much of the violation can be tolerated. Unlike much of the research in the field of robust statistics, this dissertation does not focus on overcoming the violations of assumptions in a logistic regression, but rather, on articulating and documenting the impact of problematic situations in data configuration and complex data structures in logistic regression and estimation assumptions.

The problematic data characteristic that is the highlight of this dissertation is skewness in the predictor variable of a logistic regression model. While we examine skewness in both continuous and dichotomous predictors, we concentrate primarily on the latter. Traditionally, skewness is defined as asymmetry in the probability distribution of a random continuous variable about its mean. But since the term is rarely used in categorical data analysis, this definition does not carry through in the same manner for categorical variables. Instead, throughout this dissertation, we adopt the term "skewed probabilities" to describe the row (or column) marginal distribution of categorical predictor variables. We define skewness in a categorical variable as a severe inequality in the probability of the occurrence of its categories. This dissertation seeks to identify what degree of skewed probability affects the logistic regression model or maximum likelihood estimation. To what extent can a researcher trust the test results when the predictors

are skewed? This dissertation also documents the Type I and II errors and the power of the Wald test. Relatively little is known about the impact of skewed probabilities on the later statistical decisions of a logistic regression model.

The remainder of the dissertation is divided into three chapters. Chapter two provides an overview of logistic regression models and identifies relevant definitions that are used in this body of work. It then examines the recent application of logistic regression in higher education research[1] to understand current practices and reporting practices. A few recommendations conclude this part of the chapter. The final section describes the methodology that was used to analyze the data, paying particular attention to the Monte Carlo simulation experiment.

At the core of this dissertation, chapter three offers a rationale for the research question and problem statement[2]. It describes the concept of skewness in cell probability and how it is linked to the experimental design. It then examines skewness through three main models: a simple logistic regression model with a binary predictor, a simple logistic regression model with a continuous predictor, and a multiple logistic regression model with two dichotomous predictors (similar to randomized-control experimental models). Skewness, sample size, and the interaction between the two are the main factors that were studied. The research also analyzed the rate of Type I and II error and statistical power. The last chapter concludes this dissertation and describes the future direction of research in this area.

[1] The field of Higher Education Research was chosen as the focus for two reasons. First, in additional to statistical methods, this is my area of substantive research and in which I hold an academic appointment. Second, logistic regression is a widely used method in this field – perhaps among the largest adoption in sub-fields of Educational research.

[2] This dissertation is written in the University of British Columbia's 'manuscript' style dissertation format wherein, in my case, the central contribution is a free-standing manuscript that is prepared for publication at a journal. Please see the Preface of this dissertation for details.

Chapter 2: Background

The number of publications using logistic regression is increasing dramatically, especially in higher education research. Peng, So, Stage, and St John (2002) found that 52 such articles were published from 1988 to 1999 in three leading education journals. Moreover, a search of the Education Research Complete database using the keyword "logistic regression" showed that from 2000 to 2013, the journal of *Research in Higher Education* published 134 articles using this term, compared with only 37 in the previous decade.

The growing appeal of logistic regression is due to advances in its ability to measure and allocate discrete variables. Unlike OLS regression, it has no assumptions regarding the distribution of predictors, their relationship with the dependent variable, or homogeneity within groups, making it more suitable for higher education research. In addition, a logistic regression model may have a mix of continuous and categorical predictors (Tabachnick & Fidell, 2013).

All statistical software packages can now perform logistic regression. SPSS, STATA, SAS and R are the most popular in the behavioural sciences. Other programs include Microsoft Excel, MedCalc, LogXact and many free online calculators that are based on R or other platforms. Similarly, many books and articles on this technique are published for educational researchers. The abundance of data on logistic regression and the availability of the appropriate software have not only increased its use, but have created possibilities for very intricate modeling.

This chapter has multiple purposes. The first is to provide an overview of logistic regression and to ease us into the language in which it is discussed in the next chapter. It also sheds light on the practices and current use of this technique and summarizes the methods employed in this dissertation.

What Is Logistic Regression?

Regression methods are an integral part of most analyses concerned with understanding the relationship between an outcome variable and one or more predictor variables. Logistic regression models are the most common technique used when the outcome variable is discrete.

In a dichotomous case, the relationship between the outcome and predictor is not linear, but follows a logistic function. The outcome variable y is coded $y = 1$ if it is in a category and $y = 0$ otherwise. The probability distribution associated with a dichotomous y is a Bernoulli distribution with a mean of the proportion p of cases falling in the category and a variance of $p\ (p - 1)$. The linear regression model with one predictor, x, is:

$$y_i' = \beta_0 + \beta_1 x_i,$$

where y_i' is the estimated outcome variable, β_1 is the parameter estimate for the predictor x_i and is interpreted as the slope of the linear relationship, and β_0 is both the constant and the intercept.

Estimating parameters in this model using OLS will violate many assumptions. For instance, as a result of the distribution of the outcome variable, the conditional mean or predicted score can range from zero to 1. The additive nature of the linear OLS regression can generate a predicted score that falls outside this range, thus producing an inappropriate estimate for the population's probability of being in a category. Furthermore, the dichotomous nature of the outcome variable has two undesirable constraints on the residuals. Firstly, the residuals are not independent because their variance is based on the predicted scores ($y_i'(1 - y_i')$). Secondly, the discrete nature of the outcome variable results in residuals that do not satisfy the normality assumption, but follow, in the binary case, a binomial distribution (Cohen, Cohen, West, & Aiken, 2013).

To overcome the limitations associated with a dichotomous outcome in OLS regression, a logistic link function is used to predict the probability of being in a category p_i' from one or more predictors.

Probability $(y = 1|x = x_i) = p_i' = 1/[1 + e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_k x_k)}]$,

where x_i is a discrete or continuous predictor.

with simplification:

$$\frac{p_i'}{1 - p_i'} = e^{(\beta_0+\beta_1 x_1+\cdots+\beta_k x_k)}$$

The natural log of the previous equation is identical to that of the OLS linear regression model:

$Logit = \ln\left(\frac{p_i'}{1-p_i'}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$,

where the logit can take any value.

With a dichotomous outcome variable, one can perform the simplest type of logistic regression. In cases where the outcome variable is polytomous or ordinal, the model becomes more complicated. In the polytomous case, for an outcome variable y with three categories, for example, two logit functions are needed. The researcher must decide which category to use as the reference. When the outcome variable is ordinal, the researcher must also consider their rank ordering. The relationship between ordered categories becomes a movement along a latent continuum as a function of the predictors (Cohen et al., 2013).

Parameter estimation is not direct in logistic regression, but depends on an iterative numerical process. This estimation process is called maximum likelihood, where the likelihood of a sample given a chosen estimate is calculated based on a statistical criterion for stopping. The estimates are modified in a manner that increases the likelihood of a sample. This process is repeated until the coefficients slightly differ by a convergence criterion. A solution is said to

converge when the amount of change from one iteration to another is less than the convergence criterion (Cohen et al., 2013).

Evaluation of logistic regression models.

The beta coefficients that are estimated through maximum likelihood are analogous to the interpretation of the coefficients in an OLS regression. That is, the predicted logit increases by β_i for an increase of one unit in the predictor. Hence, together with the standard errors, the magnitude of a predictor can be assessed. The statistical significance of each predictor is evaluated using the Wald test, $= \frac{\beta_i^2}{SE_{\beta_i}^2}$. The test has a chi-square distribution with one degree of freedom under the null hypothesis that the coefficient is equal to zero. The square root of the Wald test follows the normal distribution.

Predicted scores can be transformed into odds ratios that range from 0 to ∞. An odds ratio (OR) measures the association between a predictor and an outcome. It represents the odds that an outcome will occur given a particular predictor, compared to the odds of the outcome occurring in its absence. The OR is calculated by the exponentiation of the beta coefficient. In the multivariate predictor case, it is interpreted, similar to OR, while holding other coefficients constant. An OR of one is equivalent to a beta of zero, indicating the absence of a relationship with the outcome variable.

In case-control studies, cohort studies, or clinical trials, odds ratios are a common measure of the effect size. Odds ratio are usually interpreted in relation to the relative risk, which is calculated as the ratio of the risks in the two groups for an outcome of interest. When the risks in the two groups being compared are both small (say less than 20%) then the odds ratio will approximate to the relative risk, and interpretation is straightforward. However, as the risk in either group rises above 20% the gap between the odds ratio and the relative risk will increase.

When the effect is size is large the odds ratio will exaggerate the effect in comparison to the relative risk (Oakley Davies, Crombie, & Tavakoli, 1998). That is, if the odds ratio estimate is less than one then it is always smaller than the relative risk. On the other hand, if the odds ratio is greater than one then it is overestimated and always larger than the relative risk. To better estimated the odds ratio and the accompanied relative risk, an adjusted odds ratio calculation is recommended by Mcnutt, Wu, Xue, & Hafner (2003) to better approximate the measure of relative risk.

An assessment of the adequacy of the model should precede the interpretation of parameters. There are generally two approaches to assessing the degree to which the model fits the data: measures of predictive power (like R^2) and goodness of fit tests (like the Pearson chi-square).

The coefficient of determination or R^2 is popular mainly due to its interpretability as the proportion of the variance of the dependent variable that can be explained by a given regression model and its predictor variables. It quantifies predictability and gives the strength of a regression relationship. Although there are many ways to calculate R^2 for logistic regression, no consensus exists on which one is best for two reasons (Menard, 2000; Mittlbock & Schemper, 1996). First, there is no clear description of how to calculate and determine the corresponding measures of the strength of association between the discrete dependent variable and the total set of predictors. The coefficient of determination for quantitative dependent variables is based on residual variation, which is calculated through sums of squares. However, several possible residual variation criteria can be proposed for binary dependent variables (Menard, 2000). As a result, there are numerous mathematical equivalents to R^2 in an OLS regression, which are not mathematically or conceptually equivalent to R^2 in a logistic regression. The second reason is

that with binary responses, the various measures of R^2 tend to be low even for an underlying perfect regression relationship. The three measures of R^2 that are most often reported by statistical software are: (a) one proposed by McFadden (1974); (b) one attributed to Cox and Snell (1989), along with its corrected version (Allison, 2014); and (c) Nagelkerke's (1991) generalization of R^2 to include discrete outcome variables.

The second class of measures that assess suitability of a logistic model are goodness-of-fit statistics. These can help researchers decide whether their model fits the data well. Classic goodness-of-fit tests are available for logistic regression when the data can be aggregated into covariate patterns, which are groups of cases that have exactly the same values on the predictors. In the goodness-of-fit tests, the fitted model is compared against a saturated model. There are two widely used tests of model adequacy: the deviance or likelihood ratio test and the Pearson chi-square test. The calculations for the two tests are based on the observed and expected frequencies of the covariate patterns in the k-predictor model. Both the deviance and the Pearson chi-square have good statistical properties when the expected number of events and the expected number of non-events for each covariate pattern is at least 5 (Hosmer, Lemeshow, & Sturdivant, 2013). However, most applications of logistic regression use data that do not allow for aggregation into covariate patterns because they include one or more continuous predictors. Lemeshow and Hosmer (1982) proposed a goodness-of-fit test that groups cases together according to their predictive values from the model. The predicted values are arranged from lowest to highest, and then separated into several groups of approximately equal size.

Assessing a model's adequacy by looking at its three main components—that is, parameter estimates, measures of predictive power, and goodness of fit—is crucial to understanding its suitability of the model to the given data.

Assumptions and diagnostic measures.

Logistic regression, unlike linear regression, does not require distribution assumptions of predictors, although multivariate normality and linearity among predictors may enhance its power (Tabachnick & Fidell, 2013). It assumes independence of observations and that a linear relationship exists between the predictors and the logit (Hosmer et al., 2013; Tabachnick & Fidell, 2013). It also assumes a between-subjects' analysis—that is, responses from different cases in the sample must be independent from each other to ensure independence of errors. Although assumption tests are not required in this context, a few diagnostic measures should be performed to avoid potential bias and to ensure accurate parameter estimation. These diagnostic tools are originally derived from categorical data analysis, but can be extended to apply to logistic regression. A discussion of some of these concepts follows.

Sparse tables. Sparse tables are a major concern in categorical data analysis, including logistic regression. They occur when the sample size is small, when a variable contains a large number of categories, or when a table has many variables (Agresti, 2002). Agresti (2002) and Cohen et al. (2013) describe them as contingency tables having small or zero cell counts and offer several ways of dealing with them. Sparse tables are problematic in logistic regression since they affect estimation and the accuracy of the estimates required to conduct the chi-square or goodness-of-fit test (Cohen et al., 2013). Since most statistical software packages fail to converge or produce an estimate for the odds ratio, Hosmer et al. (2013) recommend collapsing the categories in the independent variable in a way that eliminates the zero cell.

Separation. Separation describes a problematic data configuration between the categorical outcome and predictor variables, and has been found to have a negative effect on maximum likelihood estimation (MLE). Albert and Anderson (1984) identified three types of

data configuration that may affect estimation: complete separation, quasi-complete separation, and overlap. They proved that while overlap yields a finite and unique solution, MLEs do not exist for the other two data patterns. As a result, methodological and psychometric researchers were motivated to introduce new estimation techniques to overcome this problem.

Separation is caused by a linear combination of continuous or dichotomous predictors that perfectly separates events from non-events. Complete separation occurs when one or more of a model's predictors perfectly predict the outcome variable. Therefore, no variance is left to be explained in the outcome variable by the model's other predictors. More commonly, quasi-complete separation ensues when only one covariate pattern has a size equal to zero. Under such conditions, the parameter estimate for the separating variable will also be infinite in size, but the model's other predictors may remain unaffected (Zorn, 2005).

Events per predictors (EVP). The key element in determining whether the sample size adequately fits the model involves the number of events per predictor. An event is defined as the frequency of the least common outcome $m = m\min(n_0, n_1)$, where n_0 is the frequency of the outcome $y = 0$ and n_1 is the frequency of the outcome $y = 1$. Peduzzi, Concato, Kemper, Holford, and Feinstein (1996) concluded that the rule of thumb of 10 events per predictor can be too conservative, but that it is needed to avoid problems of over or under estimation of variance. They devised a formula that allows for the number of parameters that are estimated as a more suitable method to determine the adequacy of a sample size:

$$EVP = Number\ of\ paramters + 1 < \frac{m}{10}, where\ m = m\min(n_0, n_1).$$

Similarly, Vittinghoff and Mcculloch (2007) found that problems associated with confidence interval coverage, Type I error, and relative bias in 5-9 events per parameter were

comparable to those in 10-16 events per parameter. Although the study showed that the former number is acceptable, the authors stated that the results should be interpreted with caution.

Application of Logistic Regression in Education Research: A Sample from 2000 to 2013 in Higher Education

Although logistic regression is frequently employed, the variability in the application, presentation and interpretation of the results makes it difficult to compare findings across articles. Confusion continues to exist over practice, most importantly in reporting and the interpretation of parameters. This review, which was conducted in 2013, was a follow-up of a study by Peng et al. (2002) of articles published from 1988 to 1999 that employed logistic regression. Peng et al. (2002) explored the following questions: "Is logistic regression analysis conducted appropriately and are results interpreted accurately?" This question is still relevant because the application of logistic regression has recently changed to include complex modeling.

The present review draws on a sample of published articles on higher education research from 2000 to 2013 and aims to examine differences in the implementation of dichotomous, multinomial, and ordinal logistic regression and the effects of the availability of better knowledge and software. Like Peng et al. (2002), I hoped to promote standards for reporting practices in this field. Because many policy decisions in higher education are based on research that uses logistic regression, it is important to report all relevant information and accurately interpret the results, especially given the added complexity of models and methods.

The application of logistic regression was examined in the higher education research literature from 2000 to 2013. Ten leading journals, shown in Table 1, were individually searched for articles using the keyword "logistic regression." The search yielded 323 articles. This number has increased dramatically from the 52 papers Peng et al. (2002) discovered in the previous

decade. Table 1 shows the number of articles referring to the term "logistic regression" published in each journal during 2000-2013. These years span a period in which all of the papers are available electronically. The coverage after 2013 is incomplete because of the recency of publication and publishing embargoes. The total number of published articles in Table 1 includes scientific papers, but excludes book reviews and editorials. *Research in Higher Education* published the most articles on logistic regression, while *Higher Education* published the most articles overall, although only 5.8% of them included the term. For an article to meet the inclusion criteria, it must have conducted at least one analysis, either primary or secondary, to derive a dichotomous, multinomial, or ordinal logistic model and its parameters. Hence, tobit, probit, and other forms of regression models were excluded.

Table 1. Number of published articles that include the term "logistic regression" in text.

Journal	Search Results	Total No. of Published Articles	Percentage
Research in Higher Education	134	525	25.5%
Higher Education	60	1024	5.8%
The Journal of Higher Education	50	472	10.6%
The Review of Higher Education	29	279	10.4%
Studies in Higher Education	22	755	3%
Innovative Higher Education	12	409	3%
The Canadian Journal of Higher Education	7	258	2.7%
Assessment and Evaluation in Higher Education	6	711	0.84%
International Journal of Higher Education	2	81	2.4%
The Journal of General Education	1	273	0.366%

Because the total number of articles was large, for the purposes of this paper, a random sample was drawn. Citations for all the literature found in the keyword search were downloaded and each was given an ID number from 1 to 323. The statistical package R was used to draw a random sample of 65 articles from the reference list. The list of all articles included in this

review is in Appendix A. Table 2 shows the breakdown of articles by year and indicates an increased use of logistic regression in higher education literature after 2008. This may be explained by the widespread inclusion of the method in statistical software and the availability of books and articles discussing it.

Table 2. Frequency of studies published per year.

Year of Publication	Number of Articles in the Sample	Year of Publication	Number of Articles in the Sample
2000	2	2007	2
2001	1	2008	5
2002	3	2009	7
2003	2	2010	9
2004	2	2011	5
2005	1	2012	4
2006	3	2013	3

Of the 65 articles reviewed, 50 (77%) employed logistic regression methods, while the 15 that were excluded referred to them as a means of comparison with other models. Table 3 shows the methods used in the excluded studies. Most of these employed OLS regression, although two replaced it with dichotomous or ordinal logistic regression.

Table 3. Methods used in excluded articles.

Method	Frequency
OLS regression	8
Event history modeling	3
Time hazard model	1
Literature review	1
ANOVA	1

This review is guided by books authored by Hosmer et al. (2013) and Cohen et al. (2013) and by a similar survey authored by Peng et al. (2002). It investigates the following research questions:

- What was the purpose of the sampled studies? Were the research questions appropriate for the methods used? What was the research design? How was data collected? Was logistic regression a primary or secondary form of analysis?
- Was the description of the model clear? What type of logistic regression was used? How was model fit assessed?
- What was the events per parameter ratio? Was it adequate?
- What parameters were reported? How were they interpreted? What types of inferences were made?
- Were interactions examined? How were they interpreted?
- What type of diagnostic analysis was used to examine the effects of outliers? How were missing data handled?
- What statistical software and references were available?

Results.

Study design. In this sample, logistic regression answered questions related to the importance of predictors (38 studies), the prediction of group membership (13 studies), and modeling (2 studies). The two studies that utilized logistic regression for modeling applied it to predict student behaviour, such as cheating and university enrolment, given a set of known variables. These models aim to help decision makers create preventive strategies in the case of cheating, or redefine policies in the case of attracting students.

The dependent variables in the sample were dichotomous (68), polytomous (18), and ordinal (2). Some authors chose to change the nature of the variables to customize them to fit logistic regression. For example, in an article analyzing a dependent variable that allowed a respondent to check all that applies or have more than one answer, the authors decided to

dichotomize each choice to perform a dichotomous logistic regression for each dependent variable. In a different article, two nominal dependent variables with four categories were dichotomized to adjust them to the technical requirements of logistic regression. However, a few authors did not follow variable assumptions in logistic regression. For example, a study treated a four-category ordinal dependent variable as nominal. A different group of researchers conducted a multinomial logistic regression on a three-category ordinal dependent variable because the parallel regression assumption was violated when performing ordinal logistic regression. In an exploratory study that analyzed strengths of association rather than prediction, the authors carried out OLS regression with a dichotomous dependent variable because R^2 from OLS was identical to Cox and Snell's measure R^2 in dichotomous logistic regression.

The authors drew their data mostly from secondary sources (36 studies, 72%), taking 36% of their information from the National Center for Education Statistics (NCES). The rest was gathered from grants, universities, and regulated governmental data collections. Surveys and records are the two major types of data, providing information on pre-college demographics, number of publications, academic progress, enrolment, and financial awards. A couple of studies conducted interviews. In addition, five studies employed validated Psychological questionnaires to measure attitudes and actions such as anxiety, motivation, fraudulent behaviours, depression, and student readiness.

Logistic regression was used as a primary method of analysis (28 studies), a secondary method of analysis as part of a larger set of techniques (10 studies), or in a complex design (12 studies). Studies that employed logistic regression as a primary method of analysis centered their methodology on logistic regression. There were also a group of studies that employed logistic regression as part of a larger methodology. One of these studies used it as a weighting technique

to correct for potential bias due to low survey response rates in a larger analysis that was centered on the structured equation modeling of a measure. Other studies used complex modelling, such as path models and cluster analysis, to answer their research questions. Logistic regression was utilized to accommodate the categorical nature of the dependent variable or to test the predictive validity of clusters. A few other studies used it as a secondary way to analyze data through t-tests, ANOVA, and OLS regression.

The complex studies included five multilevel designs, one multilevel path model, one path model, two matching model designs, one trend analysis, and one repeated measures logistic regression. The multilevel designs in this sample were mostly multinomial (4 studies) or two-level (5 studies), with student characteristics as the first level and school or college characteristics as the second. In path model studies, the authors performed a number of appropriate dichotomous or multinomial logistic regressions, then calculated the direct, indirect, and total effects. In matching model designs, the authors sought to pair students with university or program characteristics and to understand the impact of these pairings on their future. In the repeated measures dichotomous logistic regression, the dependent variable was measured at three points in time. Parameter estimation, in this study, accounted for the violation of the independent observations assumption.

Reporting practice. All studies carefully depicted the variables: what they were and how they were found and measured. The sample included dichotomous (37 studies), multinomial (15 studies), and proportional odds (one study) logistic regression models. Two studies did not disclose the nature of the dependent variable, and hence the model type, or did not give enough information. A clear description of the model itself was, in a few studies, absent.

This review established a rating scale to measure how clearly models were explained, where 1 represents a model that is not described or graphed, 2 a model that is either explicitly defined or described, or graphed or written in mathematical form, and 3 a model that is both described and illustrated. It was found that 42% of the studies (21) did not describe the model, but clearly identified the dependent variable(s). This was particularly problematic when the design included multiple dependent variables and a large set of predictors. The authors did not list the steps they took when conducting each regression, or the number of models they estimated. Mostly, however, they included results tables that showed the estimated parameters of each coefficient. The next 20 studies, 40% of the sample, clearly described the statistical analysis they conducted or offered a graph, and usually explained the types of results that were included and how to interpret them. The final 18% (9 studies) both summarized and graphed or wrote the model in mathematical form, making the transition to the results section smooth and easy to follow.

While a handful of studies explicitly indicated their modelling strategies, otherwise they were usually inferred from the text. These strategies generally involved purposeful selection (38 studies) and forward stepwise selection (14 studies) methods. One study did not specify the method because it used logistic regression as a weighting technique and provided no information on the estimated parameters. Most studies employed purposeful selection of the predictor variables, basing their decision on theory and/or prior research. Studies using forward selection were exploratory in nature and applied both theory and prior research information to their models. None of the studies eliminated coefficients after assessing model fit.

Less than 50% of the studies did not provide goodness-of-fit indices, while the rest (26) reported one or more. The most commonly reported indices are the percentage of correctly

classified cases (10 studies), the likelihood ratio of two models (9 studies), Nagalkerke's R^2 (8 studies), pseudo-R^2 (7 studies), the deviance of a full model from an intercept-only model (5 studies), the Hosmer and Lemeshow chi-square (3 studies), and the Cox and Snell R^2 (2 studies). Other indices that were reported include the adjusted R^2 (did not indicate if it was based on ML estimation or OLS), McFadden's R^2, the adjusted McFadden R^2, the score test, Bayesian information criteria, and Akaike information criteria. Seventeen studies clearly interpreted the goodness-of-fit indices, which proved to fit the data well. Others included them in results tables with no clear interpretation of the implications. One study did not interpret the results, but provided guidelines on how to understand them.

Events per predictor. Almost all the models were multivariable (two studies were bivariate), with a collection of continuous and categorical predictors reflecting the complex nature of the data in the higher education research field. All the predictor variables were presented and explained very clearly.

I found that the range of sample sizes and the number of predictors was large. Sample sizes varied from 63 to 630913, with the number of predictor variables ranging from 1 to 95. Three studies had access to complete populations of 85894, 165921, and 630913 individuals, respectively. In general, the number of predictors was reasonable, given the large sample size. Table 4 shows the events per predictor and observations per predictor ratios. Hosmer et al. (2013), Peduzzi et al. (1994), and Vittinghoff and McCulloch (2006) recommend the use of events per predictor instead of the total number of observations per predictor. Events were either given in the studies or calculated from descriptive statistics and unweighted means. Of the 62 analyses in this sample, 20 analyses in 20 studies did not include a sample of each category in the outcome variable or enough information to calculate events. Therefore, the total sample size

was used to calculate the observations per predictor ratio. We followed the 10:1 rule recommended by Hosmer et al. (2013). Two studies were not included in Table 4, one because it did not disclose the sample size due to restrictions from the data providers, and another because it lacked the results of the logistic regression analysis. Articles that relied on the NCES have approximate sample sizes since data providers ask them to round them up to the nearest 10.

Five studies had events per predictor ratios below 10:1. Those with the ID numbers 2, 20, 22, and 29 had ratios of 7.27, 7.87, 2.46, 5.21, and 6.28, respectively. These models contain a collection of continuous and categorical predictors and the results should be interpreted with caution. Study 9 had the largest events per predictor ratio at 4231.7. Observations per predictor ratios averaged 654.37, with 5 ratios below 50. These should be interpreted carefully, since the frequency of categories in the outcome variable is not known. A small count in one category in the outcome variable can affect the stability of estimates and introduce bias.

A small cell count in this context refers to the frequency of observations in a cross-tabulation of a dependent variable and one or more predictors. In this sample, two studies included cross-tabulation tables with cell frequencies as low as one; study 22 had a sample size of 482 and 23 predictors, and study 38 had a sample size of 85894 and 4 predictors. Neither paper mentioned its unusually low cell count. Another five articles noted their small cell counts, two of which dropped the category, two of which combined it with another category, and one of which interpreted the results with caution. Of these studies, none examined convergence issues and one used bootstrapping to overcome the possible lack of convergence.

Table 4. Events/observations per predictor ratio for each study.

Article ID	Sample Size	Lowest Number of Events in the DV	Number of Predictors	Event/Observation to Predictor Ratio
1. Allen, Robins, Casillas, & Oh (2008)	6872	952	13	73.2
2. Anderson, Sun, & Alfonso (2006)	517	126	16	7.88
	315	109	15	7.27
3. Bahr (2008)	85894 (had access to population)	1897	66	28.7
4. Bahr (2010a)	165921 (had access to population)	4691	52	90.2
5. Bahr (2010b)	68884	-	54	1275.6
6. Bahr (2012)	133482	12786	20	639.3
	101871	3297	20	164.85
7. Bailey, Calcagno, Jenkins, Leinbach, & Kienzl (2006)	915	-	68 (largest)	13.4
8. Belloc, Maruotti, & Petrella (2009)	9725	1518	20	75.9
9. Berggren (2006)	630913 (had access to population)	29622	7	4231.7
10. Bieri & Schuler (2011)	147	59	2	29.5
11. Bonilla, Buch, & Johnson (2013)	2063	381	1	381
12. Callender & Jackson (2008)	817	212	26 (largest)	8.15
13. Craney, McKay, Mazzeo, Morris, Prigodich, & Groot (2011)	465	274	2	137
14. Crisp & Nora (2010)	570	200	20	10
15. DesJardins (2001)	9604	-	36	266.77
16. DesJardins (2002)	3801	1237	26	47.5
17. Eggens, Van Der Werf, & Bosker (2008)	1451	-	36 (largest)	40.3
18. Engberg (2007)	4697	-	2	2348.5
19. Flashman (2013)	35770	-	16 (largest)	2235.6
20. Girard (2010)	1474	37	15	2.46
21. Hovdhaugen (2009)	1780	319	18	17.7
	1583	755	18	42
22. Hu & Hossler (2000)	482	73	14	5.21
23. Jansen (2004)	5151	1751	12	145.9

Article ID	Sample Size	Lowest Number of Events in the DV	Number of Predictors	Event/Observation to Predictor Ratio
24. Kim, Bankart, & Isdell (2011)	20295	6874	37	185.78
25. Klien & Weiss (2011) Will be problematic if the lowest number of cases is less than 30% or 8:1.	2594	-	95	27.3
26. Klugman (2012)	9880	889	18	49.38
27. Konecny, Basl, Myslivecek, & Simonova (2012)	37713	18097	9	2010.82
28. Marks (2009)	7415 (largest of four in size and predictors)	-	7	1059.28
	2877	604	8	75.5
	8810	475	8	59.4
	8450	566	12 (largest)	47
	5371	687	11	62.4
29. Martin & Spenner (2009)	1178	132	21	6.28
30. Masjuan & Troiano (2009)	1823	-	7	260.4
		-	5	364.6
31. Melguizo (2008)	3000	300	13	23
32. Newman & Petrosko (2011)	4332	685	26	26.35
33. Outcalt & Skewes-Cox (2002)	855	-	9	95
34. Perna (2005)	8982	1401	37	37.86
		844	37	22.8
35. Perna & Titus (2005)	9810	-	59	166.27
36. Riegle-Crumb (2010)	1635	491	18	27.27
	2006	632	18	35.11
37. Roksa (2010)	2789	-	26	107.26
38. Seelen (2002)	63 (smallest)	-	2 (all models had the same set of variables)	31.5

Article ID	Sample Size	Lowest Number of Events in the DV	Number of Predictors	Event/Observation to Predictor Ratio
39. Shankland, Genolini, Franca, Guelfi, & Ionescu (2010)	130	-	7	18.57
40. St. John, Musoba, Simmons, Chung, Schmit, & Peng (2004)	65588	-	23 (largest)	2851
41. Stassen (2003)	3948	1169	16	73
	3580	1138	16	71
42. Teixeira & Rocha (2010)	7213	-	34 (largest)	212.14
43. Tien (2000)	1017	474	1	474
44. Toutkoushian & Bellas (2003)	24441	10500	19	552.6
		4515	15	301
45. Wells, Lynch, & Seifert (2011)	10027	-	11	911.54
		7709	11	700.8
46. Wolnaik, Mayhew, & Engberg (2012)	2439	195	18	10.83
47. Xu (2013)	11192	-	48	233.16
48. Zimdars (2007)	568	-	7	81.14

Interpretation of results. The results were presented in seven different but related formats: odds ratios (34 studies), parameters in the logistic regression model (32 studies), predicted probability (10 studies), delta-p (4 studies), marginal effects (4 studies), odds (2 studies), and the inverse of the odds ratio (one study). Over half of the studies reported two formats (26 studies) and 6 reported even more. The most reported logistic regression finding was odds ratios. The predicted probability was interpreted 20%, delta-p 8%, and odds 2% of the time. These are the most useful ways to interpret logistic regression results. When these formats were reported, they were accurately interpreted. Not all studies, however, explored important logistic regression findings; some simply explained how to understand and calculate predicted probabilities and odds ratios.

The parameters of the logistic regression models were included in tables in the results or discussion sections of the studies. The interpretation of the beta coefficients in those studies implied confusion between their roles in OLS and logistic regression. Indeed, many papers that included the beta parameter and marginal effects only interpreted this parameter in an OLS manner, such that beta itself represented magnitude and direction. The logit link function between predictors and the dependent variable must be reflected in the interpretation of the beta coefficients, odds, odds ratios, and predicted probabilities. Marginal effects are also not favoured in logistic regression since it is based on a linear relationship between the predictor and the dependent variable (Peng et al., 2002). The interpretation of parameters in complex models was no more sophisticated. The discussion of the design and results reflected the complexity of the designs, but the parameters were interpreted similarly to those in the rest of the sample.

Most studies (29) included standard errors. The magnitudes of the reported standard errors were reasonable, with a maximum value of 0.75 in the study with the lowest events per

predictor ratio. In papers that did not mention standard errors, the stability of the parameters was unknown. Less than half the sample included the intercept (21 studies), making it difficult for readers to infer other findings than those important to the authors or to compare the study with the rest of the literature. Five studies reported the Wald test, which shows the significance of a predictor. However, when the statistic was not presented, the parameter's significance was usually indicated in the results tables.

Lastly, 11 articles examined the interaction effects after investigating the corresponding main effects. All of these noted the significance of the interaction effects, while two studies reported their odds ratios and predicted probabilities and two others interpreted them in OLS terms (beta, marginal effects).

Diagnostic measures. Three studies conducted a diagnostic analysis of outliers. The first dropped them from the sample; the second examined Cook's influence statistics, leverage values and normalized residuals; while the last investigated boxplots and Cook's influence statistics. Over half the studies (25) performed a total of 31 analyses of missing data: three conducted a sensitivity analysis, seven determined multiple imputations of missing values, 15 omitted missing cases, five included a missing as a dummy variable, and one mentioned that it had dealt with missing data.

Resources used. Most of the studies (37) did not say which software they employed for the analysis. Based on the 14 studies that reported this information, STATA was used the most (9 studies). Other types of software are reported in Table 5.

Table 5. Software used for estimation.

Software	Frequency	Percentage
WinBugs	1	1.9%
STATA	9	17.6%
SPSS	2	3.9%
SAS	1	1.9%
R	1	1.9%
Unspecified	37	72.5%

A potential indicator of what sources authors might have used as guidelines for logistics regression analysis is the citation of multivariate resources. Most of the studies, as shown in Table 6, did not cite any multivariate references, while 20 mentioned multivariate statistics textbooks and methodology papers. Table 6 provides a list of references and how often they were cited. The most frequently cited books are Long (1994), Hosmer and Lemeshow (1989), and Hosmer and Lemeshow (2000). The most frequently mentioned papers are Cabrera (1994) and Peng et al. (2002). These references provide reputable information on logistic regression.

Table 6. Citation of books and articles used for methodology.

Citation	Type	Frequency
Agresti, A. (1990). *Categorical Data Analysis* (2nd Edition). Hoboken, New Jersey: John Wiley & Sons, Inc.	Book	1
Allison, P. D. (1999). *Logistic regression using the SAS system: Theory and application.* Cary, NC: SAS Institute, Inc.	Book	1
Cabrera, A. F. (1994). Logistic regression analysis in higher education: An applied perspective. In J. C. Smart (Ed.), *Handbook of theory and research* (Vol. 10, pp. 225–256). New York: Agathon Press.	Book	3
Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioural sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.	Book	1
Frank, K. A. (2000). Impact of a confounding variable on a regression coefficient. *Sociological Methods and Research, 29*, 147–194.	Article	1
Hanushek, E. A., & Jackson, J. E. (1977). *Statistical methods for social scientists.* San Diego, CA: Academic Press.	Book	2
Harel, O. (2009). The estimation of R2 and adjusted R2 in incomplete data sets using multiple imputation. *Journal of Applied Statistics, 36*, 1109–1118.	Article	1
Hocking, R. R. (2003). *Methods and applications of linear models: Regression and the analysis of variance.* Hoboken, NJ: Wiley.	Book	1
Hosmer, D. W., & Lemeshow, S. (1989). *Applied regression analysis.* New York: John Wiley and Sons.	Book	3
Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: Wiley-Inter- science Publications.	Book	3
Long, J. S. (1997). *Regression models for categorical and limited dependent variables.* Thousand Oaks, CA: Sage Publications.	Book	5

Citation	Type	Frequency
Long, J. S., & Freese, J. (2006). *Regression models for categorical dependent variables using Stata.* College Station, TX: Stata Press.	Book	1
Lottes, I. L., DeMaris, A., & Adler, M. A. (1996). Using and interpreting logistic regression: A guide for teachers and students. *Teaching Sociology, 24*(3), 284–298.	Book	1
McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models.* New York: Chapman & Hall.	Book	1
Miles, J., & Shevlin, M. (2001). *Applying regression & correlation: A guide for students and researcher* (1st Edition). London: Sage Publication Ltd.	Book	1
Pampel, F. C. (2000). *Logistic regression: A primer.* Thousand Oaks, CA: Sage.	Book	2
Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction.* Fort Worth, TX; London: Harcourt Brace.	Book	1
Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach.* Hillsdale, NJ: Lawrence Erlbaum.	Book	1
Peng, C. J., So, T. S. H., Stage, F. K., & St. John, E. P. (2002). The use and interpretation of logistic regression in higher education journals. *Research in Higher Education, 43*, 259–294.	Article	3
Petersen, T. (1985). A comment on presenting results from logit and probit models. *American Sociological Review, 50*(1), 130–131.	Article	2
Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods.* Thousand Oaks, CA: Sage.	Book	1
Statistics book in a non-English language	Book	1
None		30

Discussion.

This descriptive analysis was conducted to compare the findings of Peng et al. (2002) on the state of the application of logistic regression in higher education research to more recent work in this area. Indeed, Peng et al.'s (2002) research questions are very similar to those posed in this review. An important limitation of this review, like all reviews of this nature, is that what is reported in the articles does not, necessarily, reflect all of the analyses and assumption checks conducted by the authors. There are likely many 'file drawer' analyses. However, this review does reflect what is allowed (and required) by the community of reviewers and editors of these journals.

This review of the 50 articles showed an increase not only in the use of logistic regression, but also in its application to complex designs and research questions. Logistic regression was not necessarily the focus of these articles' methodology, but was often embedded in larger and more sophisticated designs.

Over the nearly 25 years covered by Peng et al. (2002) and by this recent review, some good practices have remained. Most studies deliberately selected the predictors in the model, which, as previously stated, is the most scientifically sound way to implement regression analysis. Fewer studies are performing stepwise selection, which is known for its overfitting issues. Another practice that persists in this sample is evaluating models using goodness-of-fit indices, such as detecting deviances and validities of predicted probabilities through classification tables. In this data set, none of the studies employing classification tables provided enough information on the choice of cutoff. However, almost all of those that provided goodness-of-fit indices presented more than one measure. The final practice that is found in both literature reviews is the correction of selection or response bias. Peng et al. (2002) reported

frequent correction of these biases through computer programs. This finding can be extended, as 19 studies (38%) emended such biases. However, as Peng et al. (2002) observed, studies here rarely resolved self-selection biases, limiting the generalizability of their findings.

This review also notes issues in the recent literature that differ from those found by Peng and colleagues (2002). The first is the use of the events per predictor ratio to calculate the adequacy of the sample size. It also shows that generally, large sample sizes and a reasonable number of predictors eliminated this problem. Six studies were found to have fewer than the recommended number of events, meaning that they have extremely large or small proportions of the reference category in the outcome variable.

Secondly, most studies did not contain cross-tabulation or distributional information on the variables of interest, which would have aided readers in evaluating the findings. Moreover, even with large samples, a couple had small cell counts, which can result in convergence problems or biases in estimates; coupled with a small sample size, they can also inflate standard errors.

While none of Peng et al.'s papers reported conducting analyses of outliers, a few more recent studies have adopted this practice. Another improvement is the growing trend of applying the analysis of interaction effects, after main effects were examined. Peng et al. (2002) found that although interactions were addressed after the main effects were found to be significant, they were typically handled by subgroup analyses.

To help validate the findings of this review, a complete examination of all 323 articles from 2002 to 2013 is needed. Additionally, since it is based on one individual's review, inter-rater reliability does not apply. It also depends on what has been reported in the literature, which may not reflect what has actually been done to produce the published material.

Recommendations based on the results of this review are built upon the idea of requiring better reporting practices. This may or may not affect the practices themselves.

Recommendations.

In light of the previous review, this section puts forward recommendations on the use and interpretation of logistic regression for researchers in the educational and behavioural sciences. The objective of these recommendations is to enhance reporting practices, thereby helping readers to understand and evaluate statistical findings. This section does not make judgements on how authors apply logistic regression, since they may not report all the steps they took in their articles. Rather, it aims to shed light on how reporting can be improved. The following is a list of recommendations that emerged from this review.

1. Researchers should define their terminology. They should describe their parameter estimates clearly by providing brief mathematical definitions and examples of how they are commonly interpreted in logistic regression or are used in the article. Similarly, indices such as R^2, deviance, and classification of cases should be accompanied by brief descriptions of what they are and how they assess model fit.
2. For readers to understand the purpose of the study, it must thoroughly explain the logistic regression model. For a model to be clear, all its parts must be described very well. Firstly, the outcome and predictor variables should be identified explicitly. With categorical variables, the reference category should be explicated. Secondly, the type of model that is used must be mentioned. Thirdly, the article should specify the number of models that will be estimated and the set of predictors in each model. Fourthly, it should identify the modeling strategy,

whether purposeful, stepwise, or block selection. Lastly, the paper should clearly state how the model will be assessed, what goodness-of-fit indices will be used, and whether any predictors will be eliminated as a result.

3. Since the goodness-of-fit indices are essential for understanding the model, they should be presented and explained well. If deviances are used, the deviance value for both the saturated and unsaturated models should be given and interpreted. If R^2 is used, the type should be clear and caution should be taken with the interpretation. If the study employs classification of cases, it should include a precise description of the probability cutoff value. This reasoning should be applied to all other indices.
4. When reporting logistic regression results, authors should provide all parts of the equation, including the intercept, standard errors or confidence intervals, and Wald test. In other words, all evaluative statistics should be presented.
5. Care must be taken when interpreting beta coefficients. Researchers should keep the logistic link function in mind. Also, interpreting odds ratios instead of beta provides more meaningful information about predictors.
6. The presentation of delta-p should be accompanied by the initial probability and the specified values of other predictors. It must be clarified that delta-p is not a parameter estimate, but is an index describing the change in the probability.
7. Most studies did not report performing diagnostic analyses. This is an important step to mention. Also, the exact types of analyses and results should be noted.
8. Small cell count is an issue that can affect the convergence of the estimation procedure. It may also lead to unstable estimates or inflated standard errors. A

cross-tabulation of the outcome variable and predictors in the model can help readers judge if this problem can influence the estimation.

9. Authors should offer descriptive statistics such as distributions, means, and the number of events of categorical variables to reveal more information about the performance of the predictors in the models.
10. Given the increasing number of complex models using this technique, extra care should be taken when performing and documenting logistic regression. A complete description of where logistic regression falls in the model design is required. Also, the interpretation of parameters must suit the larger complex model.
11. Authors should specify the resources they use. Statistical software can help readers replicate certain findings, as well as understand the defaults that led to them. Also, references to multivariate articles or books can inform readers about the school of thought the authors are using to make sense of the logistic models.

The Monte Carlo Simulation Analysis

A Monte Carlo simulation is a process that examines "what if" scenarios for factors or phenomena being considered. In this simulation, a probability distribution is created by building models of possible results and substituting the range of values. It then calculates the results, each time using a different set of fixed or random values depending on the research question. A Monte Carlo simulation could involve tens of thousands of recalculations before it is complete. During the process, values are sampled at random from the input probability distributions. Each set of samples is called a replication, and the resulting outcome from that sample is recorded. In this way, a Monte Carlo simulation provides a comprehensive view of what may occur. It shows not

only what could happen, but how likely it is to happen.

The sampling distributions in a logistic regression context for an outcome variable are usually binomial, multinomial, or Poisson, depending on the outcome variable type. A dichotomous outcome variable follows a binomial distribution with one trial, or what is called a Bernoulli distribution. Binomial distributions typically represent the number of successes in a series of n trials. The observations in this distribution must meet a few assumptions. The total number of observations n must be fixed in advance. Each is a dichotomy that falls into either success or failure (in a multinomial distribution, more categories are present). The outcomes of all n observations are statistically independent, and all n observations have the same probability of "success" p. The mean and standard deviation for a variable that follows a binomial distribution x are $\mu = np$, and $\sigma = np(1-p)$. For small samples, binomial distributions are skewed when p is different from 0.5. When the sample is highly skewed or p is very high or low, an MLE is not sufficient to estimate the parameters. Historically, median-unbiased estimators were preferred (Hirji, Tsiatis, & Mehta, 1989).

In this study, a Monte Carlo simulation was used to examine the impact of skewness in the probability of a dichotomous predictor at the population level in a few different logistic regression models—that is, when skewness is not a sampling artifact, but rather the result of a population imbalance. The predictor(s) was drawn from a Bernoulli distribution with a skewed p, which is different from 0.5. These variables are applied to a logistic regression, and the results are examined in the next chapter. The flowchart of the simulation experiment is in Appendix B. Throughout the next chapter, the focus is on the Type I error rates and the statistical power of the Wald test for the predictor(s).

Type I and II error and statistical power.

A Type I error occurs when the null hypothesis is rejected when it is in fact true. The null hypothesis is related to the statement being tested, either because it is believed to be true or because it is used as a basis for argument. In common practice, the significance level, often denoted as alpha (α), for which the null hypothesis is rejected is decided before performing a test. In general, the least amount of Type I error accepted, in other words, α, is 0.001, 0.01, or 0.05. In our research, a few situations were simulated where estimates agree with the null hypothesis.

Conversely, accepting the null hypothesis when it is untrue is a Type II error, which is often denoted as beta (β). The power of a hypothesis test is the probability of not committing a Type II error or 1-β. In other words, the power measures the test's ability to reject the null hypothesis when it is actually false, that is, to make a correct decision. Similarly, I simulated a few situations where estimates deviate from the null hypothesis.

Two techniques can help compare the nominal and empirical Type I error rates and statistical power. The first is to examine the empirical distributions and hypothesis tests. The second is to use Bradley's (1978) well-established criteria, which was done throughout this dissertation. Bradley (1978) specified two criteria of robustness, one stringent and one liberal. His stringent criterion states that for a robust test, the empirical Type I error should lie within a range of $\alpha \pm 0.l\alpha$, whereas his liberal criterion requires a range of $\alpha \pm 0.5\alpha$. Since a nominal Type I error rate of 0.05 is specified, the interval for an acceptable empirical Type I error rate lies between 0.025 and 0.075 for a liberal study and between 0.045 and 0.055 for a stringent one.

The following chapter answers the research question on the effect of the skewed probability of a predictor on the eventual statistical conclusions of a logistic regression model.

Five interrelated studies were conducted wherein I simulated outcome and predictor variables with varying degrees of skewness, sample size, and predictor variable type (i.e., dichotomous and continuous).

Chapter 3: The Impact of Predictor Variable(s) with Skewed Cell Probabilities on Wald Tests in Binary Logistic Regression

Logistic regression modeling is growing in popularity in psychological and educational research (Cohen et al., 2013; Tabachnick & Fidell, 2013). In these disciplines, data analysts commonly encounter skewed predictor variables: either categorical predictor variables that reflect skewed cell probabilities or skewed continuous predictors. The purposes of this chapter are to describe the issues surrounding skewed predictors and to document their consequences on parameter estimation, as well as on the Types I and II error (and statistical power) of their Wald tests.

The skewness of predictors is rarely discussed in statistical treatments of logistic regression for educational and psychological researchers. Moreover, while the mathematical statistics literature does mention skewed variables, as will be seen below, they are typically used as a motivation for employing alternative estimators, test statistics, and analysis strategies—which is quite reasonable given the purpose of those studies. What is not found in either the methodological or the mathematical/statistical literature is a detailed documentation of the impact of predictor skewness on the convergence of estimators, and on the Types I and II error and statistical power of the hypothesis tests. There is no detailed information to guide researchers on the impact of skewed predictors in logistic regression. It should be noted that throughout this dissertation, the term "operating characteristics" is used to refer to the Type I and II error rates and statistical power of a hypothesis test (Ferris, Grubbs, & Weaver, 1946).

With an eye toward filling this gap in the literature, this chapter reports the results of five simulation studies that aim to provide a comprehensive investigation of the convergence in

maximum likelihood estimation (MLE) of the regression parameters (b-weights) and the operating characteristics of the Wald statistic for predictors in logistic regression with skewed cell probabilities. With this purpose in mind the remainder of this chapter is organized into four sections. The first offers an overview of problematic data configurations in categorical data analyses and a summary of what is known to date about their impact on the Wald statistic. The second section describes the general simulation methodology that is used throughout. The third includes the methods, results, and conclusions for five simulation studies for three classes of widely used binary logistic regression models. The simulation studies focused on examining the convergence rates of the MLE and the operating characteristics of the Wald tests. The last section is the overall discussion of the findings.

Problematic Data Structures: Sparse Tables, Skewness, and Separation in Logistic Regression, and Statement of the Problem

There are very few discussions of the issue of skewed or unequal cell probabilities in the logistic regression literature (Jennings, 1986; Larntz, 1978). A review of the literature on broad categorical data analysis reveals three types of data patterns that provide a context for issues potentially related to the impact of skewed cell probabilities and hence may offer insights on the problem.

Three types of data patterns.

To understand problematic data patterns, we must first be able to visualize the data. In addition to the conventional data matrix (in which rows are participants and columns are variables), we can display categorical data as a multi-way table in which the cells are counts of occurrences of the corresponding row and column elements. The former display allows one to gain insight on the variety of covariate patterns, whereas the latter allows one to learn about

potential small sample sizes in the cells of the table that result in sparse data. The statistical literature on categorical data analysis uses both of these data visualization tools, though it focuses more on cross-tabulation and the language of cell counts, and provides a few descriptions of problematic data structures and an extensive number of remedies (i.e. smoothing techniques and robust estimation procedures).

Sparse tables are a common concern in categorical data analysis. From the perspective of the cross-tabulation of the data, one is fitting a logistic regression model with categorical predictors in this table as is common in experimental designs (Agresti, 2002). In his discussion of empty cells and sparse tables, Agresti (2002) describes them as contingency tables having small or zero cell counts. Sparse tables may occur when the sample size is small, when a variable contains a large number of categories, or when a model has many predictor variables and hence a high dimensional multi-way table. A sparse table, although only a concern in cross-classifiaction analyses in categorical data analysis, may manifest in a logistic regression with categorical predictors (Agresti, 2002). For example, a logistic regression with a single dichotomous predictor can be thought of as a two-way (row-by-column) table that has a similar format to Table 7. In Table 7, we can see that even though the outcome variable (Y) is symmetrically distributed and the predictor variable has a small skew in the marginal cell counts, there is a cell with zero occurrences—an empty cell. As such, it is clear that the marginal distributions are not necessarily indicative of the covariate pattern in the data.

Table 7. Two-way table with a zero count cell, an example of a sparse table or quasi-complete separation.

		X		
		0	1	Total
Y	0	40	10	50
	1	0	50	50
Total		40	60	100

The issue of separation was first introduced by Day and Kerridge (1967) to describe a problematic data configuration between the categorical outcome and predictor variables that negatively affects the MLE. Refining these earlier findings, Albert and Anderson (1984) identified three types of data configurations that may affect estimation: complete separation, quasi-complete separation, and overlap. They proved that, while overlap yields a finite and unique solution, MLEs do not exist for the other two data patterns, although future researchers introduced new techniques to overcome this obstacle (e.g., Barreto, Russo, Brasil, & Simon, 2014; Gordóvil-merino, Guàrdia-olmos, & Peró-cebollero, 2012; Heinze & Puhr, 2010; Mîndrilã, 2010; Rousseeuw & Christmann, 2003).

Although skewness is a term rarely used in categorical data analysis, following Larntz's (1978) classic study, we will adopt the phrase 'skewed probabilities' to describe the row (or column) marginal distribution of the categorical predictor variables. It should be noted that this phrase has two uses in the statistical literature of interest. Larntz considers a case where the binary or multinomial predictor variables have an implicit order (what he describes in his motivating example as 'help grade', or otherwise an ordered categorical variable of "help"), and where the marginal probabilities of the predictor variable are therefore "distributed" in a skewed manner. On the other hand, Jennings (1986) does not use the phrase "skewed probabilities," but

instead describes the marginal probabilities of the outcome variable as equal or unequal. The characterization Larntz describes is more in line with the one adopted in this dissertation, in good part because we have observed that it corresponds more closely to how data analysts in education and Psychology conceptualize such distributions.

Although sparseness is a term more formally adopted in the categorical data analysis literature and has a format similar to the example shown in Table 7, this term was consciously avoided. The reason was that the covariate structure is quite irrelevant when a multiple logistic regression model including both continuous and categorical predictors is fit. From a modelling perspective, the logistic regression is fit over the data arranged in the table, therefore sparseness as defined by the low count for some covariate pattern(s) does not apply. The framework adopted in this dissertation is based on the generalized linear modelling and regression perspective. Although sparseness is a term very closely related to the issue at hand it is firmly situated in cross-classification framework as well as independence-type of tests such as Cochran-Mantel-Heanszel test. Separation, a condition that is specific to generalized linear models, was also avoided here since it is a sample property rather than a population characteristic. Because of that skewness here is used in lieu of sparse tables or separation. In addition, sparse tables and separation are descriptive of a relationship between two variables (i.e., outcome and predictor), whereas skewness in the probability of occurrence for the categories in a predictor is a unique descriptor of one variable. Also, the term skewness can be generalized to the continuous cases.

Since I am specifically interested in the skewness of the predictor variable, the methods in conducting this study reflect this concept. It should be noted, however, that in severe cases of skewed probabilities, sparse tables or separation may occur. On the other hand, similar to the

example shown in Table 7, a sparse table or separation does not indicate skewed marginal probabilities in a variable.

Relatively little is known about the impact of skewed probabilities on later statistical decisions of a logistic regression model. Therefore, if the skewness in the probabilities of a predictor is not severe enough to disrupt the MLEs in terms of convergence, to what extent could a researcher trust the test results and make valid decisions? Table 8 represents an example of this problem, wherein the predictor X's probability of obtaining category 0 is nine times more likely to occur than category 1, all while the probability of obtaining both categories in the outcome variable Y is approximately 0.5. The cell counts for the adjacent cells of (X=0, Y=1) and (X=1, Y=1) are very different. The question is, although estimation will yield a finite solution, to what extent are these estimates to be trusted? Is there bias? How large or small are the standard errors? And ultimately, how much can we trust the results of test statistics such as the Wald test?

Table 8. An example of the data structure examined in this study.

		X		
		0	1	Total
Y	0	89	18	107
	1	91	2	93
Total		180	20	200

What is Known to Date

Skewed probabilities of a categorical variable.

Although we know of no studies that have investigated the skewness of predictor variables in logistic regression, we have found three others on related issues. Jennings (1986) examined the impact of skewed probabilities (in the outcome variable) in a dichotomous logistic

regression, where one category in the outcome variable was more likely to occur than the other. The author found that the MLEs of the parameter coefficients are upwardly biased as the cell with the lowest count in the row-by-column table becomes smaller. As a result, Jennings introduced a measure that detects inadequacies in estimation. The second study by Larntz (1978) focused on the case of goodness of fit of binary and multinomial variables with two- and three-way tables and compared the performance of three multinomial goodness-of-fit statistics with varying sample sizes and degrees of skewness of cell probabilities. Working particularly with small samples because they often generate sparse tables, Larntz used a Monte Carlo simulation to induce skewness in the probabilities in the binary and multinomial variable. The author found that the fit statistics generally performed well. In the third study, which aimed to find a solution to the separation problem, Anderson and Richardson (1979) conducted a simulation study to investigate the effectiveness of a bias reduction method within MLEs. What is interesting in their study is the recognition of potential skewness in the data set; as they state, "the distribution of the maximum likelihood estimators would be skew, particularly when the number of sample points from at least one population was disproportionately small" (p. 72). Because simulating complete separation or a cell with zero frequency would result in estimates that are extremely large (characterized as $\pm\infty$), these were eliminated, while only those data sets that were "acceptable" were included (p. 74). Bias in the MLE of parameters as a result of skewness and kurtosis is not unique to logistic regression, but also extends to structural equation modeling (Yuan, Bentler, & Zhang, 2005).

Separation and MLE.

Viewing the impact of skewed predictor cell probabilities from the different but potentially related lenses of separation and sparse tables resulting from particular data

configurations (Anderson & Richardson, 1979; Jennings, 1986; Larntz, 1978), it is predicted that when these probabilities are skewed, the Type I error rate will be deflated and effect sizes will, in some cases, be inflated and may be infinite, however, the extent and under what conditions are unknown. There will be cases in the simulation when separation is inevitable—that is, when the sample size is small and the predictor variable is highly skewed. More generally, separation is caused by a linear combination of continuous or dichotomous predictors that perfectly separates events from non-events (the 1 and 0 of the outcome variable). Complete separation occurs when one or more of a model's predictors perfectly predict the outcome variable, therefore, no variance is left to be explained in the outcome variable by the model's other predictors. More commonly, quasi-complete separation occurs when only one covariate pattern has a zero count—expressed differently, when, for example, only one cell of the implied 2x2 table of X and Y is empty (Zorn, 2005, p. 161). Under such conditions, the parameter estimate for the separating variable will also be infinite, but the model's other predictors may remain unaffected (Zorn, 2005). Both complete and quasi-complete separation may be present in this simulation experiment as a by-product of the data configuration.

It is well documented that the problem with small samples and separated data lies in the estimation process—that is, a finite and unique MLE in logistic regression may not exist. The resulting estimates of the log odds ratios are biased, and the bias increases as the ratio of the number of observations to the number of parameters decreases (Cordeiro & McCullagh, 1991). The astronomically large estimates that are produced indicate that a variable perfectly predicts the outcome, which is in essence very desirable, but is an artifact of the data configuration. However, in small data sets, we must assume that separation is not due to truly infinite estimates, but is instead caused by random variation or the nature of the data configuration.

What is even more interesting is the effect of separation on test statistics, specifically the Wald test. Hauck and Donner (1977) demonstrate that for any sample size, the Wald test statistics decrease to zero as the distance between the parameter estimates and null values increases. Consequently, in all tests for model validation, validation variables are biased and the confidence intervals of the parameter estimates and the odds ratio are not efficient. In cases of separation, the distance between the parameter estimates and their null value is very large, resulting in an insignificant Wald statistic. In a simulation study conducted by Peduzzi, Concato, Kemper, Holford, and Feinstein (1996), which examined the effects of the number of events per predictor variable in a logistic regression model, they found that with two and five events per predictor, the MLE did not converge. Moreover, when the MLE did converge, the Type I error was deflated (i.e., became more conservative), the power decreased, and the empirical distribution of the Wald statistic was not normally distributed. These problems did not exist with 10 or more events per predictor. On the other hand, Barreto et al. (2014) found that the Wald test can detect which variables are significant individually, but fails to determine the significance of the variable that presents separation. The maximum likelihood estimates become inefficient, providing inflated variances.

The Wald test has been criticized for its limitations under both ideal (Pawitan, 2000) and problematic circumstances (Fears, Benichou, & Gail, 1996; Gregory & Veall, 1986; Lütkepohl & Burda, 1997; Vaeth, 1985). However, it is still widely reported and used to this day. In a recent review (Alkhalaf, 2014) of 323 articles in higher education research that use logistic regression, it was found that all of them reported the significance of parameters via the Wald test or z-statistic. Moreover, widely used software packages such as R, SAS, Stata, and SPSS provide the Wald statistic as output. For these reasons, the focus is on the Wald test in this study.

Simulation Studies

The results of five simulation studies are reported and organized around three logistic regression models.

- The first model examined simple logistic regression with skewed probabilities of a dichotomous predictor. The results of two studies are reported. The first focused on the quality of the parameter coefficient estimates, including the convergence rates of the MLEs, as well as Type I error. The second simulation study investigated statistical power.
- The second model considered skewness in simple logistic regression with a continuous predictor. Because this model was included to check the generalizability to a continuous predictor case (rather than a categorical predictor), only the MLE convergence and Type I error rate were investigated.
- The final model included two simulations that explored multiple logistic regression with skewed cell probabilities of two dichotomous predictors. Like the first model, the first simulation study focused on the convergence rates of the MLEs and Type I error, and the second on statistical power.

General methods.

Herein I describe the simulation method that is common to all the five simulation studies. In this series of studies, Monte Carlo simulations were used to examine the skewness of a predictor at the population level. That is, I look at what happens when skewness is not a sampling artifact, but is rather the result of a population imbalance of the marginal probabilities of the predictor(s), which is sometimes called "naturally occurring skewness." Examples of variables that are naturally skewed in the population include (a) the number of visually impaired undergraduate students in a certain discipline; (b) in clinical, psychological, health, or medical

research, the presence of a rare diagnostic ailment; and (c) in the social sciences, a large gender imbalance of the participants in a study due to culturally sensitive issues.

To directly answer the research question of the effect of a skewed predictor on the eventual statistical conclusions of a logistic regression model, I simulated outcome and predictor(s) variables with varying degrees of skewness, sample size, and predictor type (i.e., dichotomous and continuous). In all cases, the same statistical model that generated the data was fitted to the simulated sample using conventional MLE and Wald tests—that is, all of the models are correctly specified. Throughout this dissertation, the Type I error rates and statistical power of the Wald test for the predictor(s) are the focus. As is common practice, the nominal Type I error rate (α) was set at 0.05.

Accordingly, the overall research question can be stated more formally as: What is the empirical Type I error rate and statistical power for the Wald test when the predictor variable(s) has a skewed cell probability from a generalized linear model perspective, specifically the logistic regression model? The five simulation studies, including their methods, analysis, results, and conclusions, are discussed in the next section. To organize the findings clearly, the simulations were grouped into three parts representing each model. In each part the specified simulation studies and results and conclusions are included. The chapter concludes with a general discussion.

Model 1: Single Binary Predictor

The first model of interest involves simple logistic regression with one dichotomous predictor:

$$g(y) = \beta_0 + \beta_1 x,$$

where x is a predictor variable with skewed cell probability, β_0 and β_1 are fixed, $g(y)$ is a logit function, and y is a balanced outcome variable. This model acts as a baseline for comparing the results of the forthcoming studies.

Study A: Type I error rates and parameter estimates.

Purpose of the study. The purpose of this first simulation experiment is to document the impact of skewed cell probability in a dichotomous predictor variable on the MLE, parameter estimates, and Type I error rate of the Wald test of the β_1 parameter. It must be noted that the outcome variable of the regression model throughout this dissertation is balanced or nearly balanced (i.e., not skewed). A secondary aim of this study is to provide researchers with diagnostic information by documenting the situations where skewness may affect decisions and inferences.

Methods

Simulation factors. For this simulation, two experimental factors varied: sample size and skewness of the predictor variable. Sample sizes ranged across 13 levels from 10 to 5000. The expected probability p of the predictor variable, described in more detail below, ranged from 0.01 to 0.45 across 17 levels. Two extra conditions were investigated for comparison purposes, the case where the predictor variable is balanced, and when the probability of occurrence for both categories is 0.5. The resulting experiment is an 18×13 fully-crossed factorial design involving 234 cells, as depicted in Table 9. This large range of sample sizes and skewness levels is necessary to more fully document the impact of skewed cell probabilities.

Table 9. Simulation experiment.

β_0, β_1 fixed		Sample Size												
		10	50	100	200	300	400	500	600	700	800	900	1000	5000
Probability	0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5	1000 replications in each cell												

Simulation procedure. The simulation and analyses were conducted using the R software. There were 1000 replications in each cell of the experimental design (depicted in Table 9), resulting in an empirical probability (either a Type I error rate or statistical power) per cell, as well as an empirical representation of the sampling distribution of the parameter estimate. For each replication in each cell, the simulation algorithm consists of multiple loops that achieve different purposes. (For detailed descriptions and flowcharts of each loop, see Appendix B.) There are a few important steps in this process.

Step 1. The experiment is built upon a random sampling model, mimicking what happens in research practice. The data was generated from a Bernoulli distribution.

$$f(p) = \begin{cases} p, & k = 1 \\ 1 - p = q, & k = 0 \end{cases}$$

with the expected probability $E(x) = p$ and the variance $V(x) = p(1-p)$. The predictor is randomly drawn from a Bernoulli distribution with a specified sample size and expected probability. Similarly, the outcome variable was randomly chosen from a Bernoulli distribution with the same sample size and an expected probability that is calculated from the model as follows:

1. The mean of the Bernoulli distribution is a function of β_0 and β_1, which are fixed to zero. The intercept term is fixed to zero because the balanced outcome variable results in a natural log of one, which is zero.
2. The logit was calculated where $Logit = \beta_0 + \beta_1 X$ for the simple case.
3. The predicted probability was then calculated as $Predicted\ Probability = P/(1 - P) = e^{Logit}/(1 + e^{Logit})$. The predicted probability serves as the expected value for the Bernoulli distribution from which the outcome variable is drawn.
4. This process is repeated until the number of replications is complete.

Step 2. All the variables are aggregated in a data frame in preparation for analysis. The generalized linear models (glm) function in R is used to run the logistic regression. The parameter estimates and hypothesis test statistics are stored for each replication. In replications where the estimation does not converge (which is likely in this case due to separation[3]), an N/A is recorded and the simulation outcome (e.g., rejecting the null hypothesis using the Wald test) for that instance in the experimental design is computed from the remaining converging replications in that cell of the simulation experimental design.

[3] In some replications the simulation experiment would stop and break (meaning that the rest of the commands in the simulation would not be read by the interface). To understand why this was occurring I examined the simulated dataset that results in breaking the simulation. From the characteristics of this dataset I generated a population. Through cross tabulation, since our model is a simple logistic regression with a dichotomous predictor, the covariate structure of the population proved that complete separation was the reason why the simulation broke. As a result, the MLE would reach an unidentifiable solution and break. By overriding this problem I was able to count the number of non-convergences due to complete separation.

Step 3. The final step is to vary the sample size and skewed probability. Each combination of conditions is stored and analyzed separately. The Type I error rates are computed as the number of rejections of the null hypothesis out of the converged 1000 replications. The nominal significance level was 0.05 throughout this study. Therefore, the empirical Type I error is defined as the proportion of times that a true null hypothesis was falsely rejected at a critical value of 0.05.

Analysis of Type I error. Type I error rate was calculated for each condition. I used Bradley's (1978) approach to compare the nominal and empirical Type I error rates for each condition. Bradley specifies two criteria of robustness, one stringent and one liberal. His stringent criterion states that for a robust test, the empirical Type I error should fall within the range of $\alpha \pm 0.1\alpha$, whereas his liberal criterion stated that for a robust test, the empirical Type I error should lie in a range of $\alpha \pm 0.5\alpha$. Given that a nominal Type I error rate of 0.05 was specified, the interval for an accepted empirical Type I error rate lies between 0.025 and 0.075 for a liberal study and between 0.045 and 0.055 for a stringent one.

Results and conclusions.

Number of MLEs that do not converge. An important issue that was encountered in this study was the non-convergence of some replications, as indicated in Table 10. Table 10 is organized in the same manner as Table 9 and depicts the simulation experimental design, wherein each element is the number of non-convergences out of 1000 replications. For example, for a sample size of 200 and $p = 0.02$, 21 of the 1000 replications in that cell of the experimental design did not converge using conventional MLE. As expected, in the case of small sample sizes and a high degree of skewness in cell probability (i.e., small values of p), most of the replications did not converge. When the sample size was 10, non-convergence was present even when the

predictor was balanced (i.e., $p = 0.5$). With a sample size of 50, the issue of non-convergence diminished as the predictor became less skewed. As the sample size expanded, all replications converged, even with high levels of skewness. From the table we can see that a sample size of 500 is sufficient to ensure that the skewness of the predictor variable does not affect estimation for the single predictor model.

Table 10. Number of non-convergences from 1000 replications for Model 1.

	Sample Size												
	10	50	100	200	300	400	500	600	700	800	900	1000	5000
Probability													
0.01	893	610	388	148	50	21	2	0	0	0	0	0	0
0.02	790	380	154	21	6	1	0	0	0	0	0	0	0
0.03	710	230	58	7	1	0	0	0	0	0	0	0	0
0.04	647	143	19	1	1	0	0	0	0	0	0	0	0
0.05	575	83	7	1	1	0	0	0	0	0	0	0	0
0.06	508	44	5	0	0	0	0	0	0	0	0	0	0
0.07	481	29	2	0	0	0	0	0	0	0	0	0	0
0.08	422	12	1	0	0	0	0	0	0	0	0	0	0
0.09	392	7	0	0	0	0	0	0	0	0	0	0	0
0.1	346	3	0	0	0	0	0	0	0	0	0	0	0
0.15	213	0	0	0	0	0	0	0	0	0	0	0	0
0.2	114	0	0	0	0	0	0	0	0	0	0	0	0
0.25	62	0	0	0	0	0	0	0	0	0	0	0	0
0.3	40	0	0	0	0	0	0	0	0	0	0	0	0
0.35	24	0	0	0	0	0	0	0	0	0	0	0	0
0.4	10	0	0	0	0	0	0	0	0	0	0	0	0
0.45	6	0	0	0	0	0	0	0	0	0	0	0	0
0.5	7	0	0	0	0	0	0	0	0	0	0	0	0

It should be noted that the summary statistics reflecting the outcomes of the simulation (i.e., the empirical Type I error rates, odds ratios (ORs), parameter estimates, and standard errors) are computed based solely on the replications that converged. Non-convergent replicates are excluded, mimicking what would go on in daily research practice.

Type I error rate. Table 11 is structured in the same way as Tables 9 and 10 and provides Type I error rates for each experimental condition. These Type I error rates are compared against Bradley's criteria, which are shown in Table 12. Table 11 is greyscale coded to highlight two

important areas. The darkly shaded area falls below the liberal criterion, while the unshaded area falls within it. Given the interaction of the sample size and the skewness of the cell probability of the predictor, researchers and practitioners should be careful when interpreting results with variable characteristics that are included in the darkly shaded part of the table. As will be shown in the next study, statistical power is greatly affected for these values. By the same token, to consider the dark area a safe zone, we must also consider statistical power. The Type I error rate rarely met the stringent criterion. Most of the time, it ranged from 0 to 0.044, falling below the lower limit of the stringent threshold of 0.045.

Two baseline conditions were included to serve as a check on the simulation methodology. In the first case, the Type I error rate for different sample sizes was computed for a balanced predictor to establish baselines for comparison with the conditions wherein various levels of probability (i.e., skewness in probability) were manipulated. In the second case, the Type I error rates for various levels of probability were computed for a large sample of 5000. As expected, in both cases, the empirical Type I error rate did not exceed the liberal criterion, for the nominal level of .05 and hence verifying that the algorithm works as expected. In the balanced case, as shown in the last row of Table 11, all Type I error rates ranged from 0.052 to 0.062, meeting the liberal criterion. Also, the Type I error rates for the sample of 5000 varied from 0.031 to 0.059.

Table 11. Type I error rate for Model 1.

		Sample Size												
		10	50	100	200	300	400	500	600	700	800	900	1000	5000
Probability	0.01	0	0	0	0	0	0	0	0	0	.02	.02	.02	0.05
	0.02	0	0	0	0	0	.01	.02	.03	.03	.04	.04	.04	0.05
	0.03	0	0	0	0	.01	.03	.05	.04	.03	.04	.04	.05	0.05
	0.04	0	0	0	.01	.03	.03	.05	.03	.03	.04	.05	.04	0.04
	0.05	0	0	0	.02	.04	.04	.05	.04	.04	.03	.05	.05	0.03
	0.06	0	0	0	.03	.03	.03	.05	.04	.04	.03	.05	.05	0.04
	0.07	0	0	.01	.04	.04	.04	.05	.05	.04	.04	.05	.05	0.05
	0.08	0	.01*	.01	.04	.04	.04	.05	.04	.05	.05	.06	.05	0.06
	0.09	0	.01	.02	.04	.04	.04	.05	.04	.05	.05	.06	.04	0.05
	0.1	0	.01	.02	.04	.04	.04	.04	.04	.04	.05	.05	.06	0.05
	0.15	0	.01	.03	.06	.04	.03	.04	.04	.04	.06	.05	.05	0.05
	0.2	0	.03	.04	.05	.06	.04	.04	.05	.04	.05	.05	.05	0.05
	0.25	0	.05	.05	.06	.05	.05	.05	.05	.05	.06	.05	.06	0.06
	0.3	0	.05	.04	.06	.05	.05	.06	.04	.04	.06	.05	.05	0.05
	0.35	0	.05	.03	.06	.05	.06	.05	.04	.04	.05	.05	.05	0.05
	0.4	0	.05	.04	.05	.06	.05	.05	.04	.04	.06	.04	.05	0.06
	0.45	0	.05	.04	.05	.05	.05	.06	.05	.05	.05	.05	.06	0.05
	0.5	0	.06	.05	.06	.06	.06	.05	.05	.05	.05	.06	.05	0.06

* Rounded to decimal points.
Note: Cells depicted in grey have deflated Type I error rates, whereas those with no shading meet the adequacy condition using Bradley's criteria (see Table 12).

Table 12. Bradley's criteria.

Bradley's (1978) Criterion	Type I Error Rate
Violates liberal criterion, therefore deflated	$\alpha < 0.025$
Meets the liberal criterion	$0.025 < \alpha < 0.075$
Meets the stringent criterion	$0.045 < \alpha < 0.055$

In general, the Type I error rates ranged from 0 to 0.062, meaning that all conditions met Bradley's liberal criterion. Regardless of the sample size, the rates were consistently deflated with lower probabilities and closer to nominal values as they became more balanced. As documented in the literature, sample size plays an important role in MLE and therefore arriving at more precise parameter estimates. For example, a sample of 600 and a probability level of 0.02, results in a Type I error rate of 0.026. On the other hand, as the sample size decreased, the

level of skewed probability did not inflate the empirical Type I error rate greatly. For instance, sample sizes of 50 and 200 can tolerate skewed cell probabilities of 0.2 and 0.06, respectively. Of particular note is the tolerance of the skewed probability of the predictor in this model. Even in the most extreme case of skewness (i.e., a probability of 0.01), with the largest sample size (5000), the empirical Type I error rate is at the nominal value.

Effect size. Table 13 is structured similarly to the previous tables, each element being the average odds ratio (OR) over the replications that converged. The average OR values for small samples and a highly skewed cell probability of the predictor are astronomical with values in the millions—whereas their true value is 1. Clearly, the degree of bias caused by the skewed predictor is very high. In cases where there was bias in the OR estimate, the sampling distribution of the OR was skewed and occasionally contained large gaps in the distribution. Because of the statistical nature of the sampling distribution, it is also useful to examine its median OR in each cell, as shown in Table 14 (Birnbaum, 1964). This is referred to as median-unbiasedness.

Given the skewed nature of the sampling distribution of the OR, the OR medians are closer to the expected value of one. The median is biased upwards when the sample size is 10. The ORs displayed in Table 13 follow the trend in Table 11, wherein as the sample size and probability (i.e., skewness in probability) increase, the estimated ORs are closer to the simulated population value of one. For example, sample sizes of at least 400 perform very well and provide OR estimates closer to the simulated value when the skewed probability is at least 0.04.

Table 13. Average odds ratio, reflecting the widely used "mean unbiasedness."

	Sample Size												
	10	50	100	200	300	400	500	600	700	800	900	1000	5000
Probability													
0.01	≈∞	≈∞	≈∞	≈∞	≈∞	≈∞	≈∞	≈∞	≈∞	≈∞	≈∞	≈∞	1.05
0.02	≈∞	≈∞	≈∞	≈∞	≈∞	≈∞	≈∞	≈∞	1.2	1.2	1.2	1	1
0.03	≈∞	≈∞	≈∞	≈∞	≈∞	≈∞	≈∞	1.1	1.1	1.1	1.1	1.1	1
0.04	≈∞	≈∞	≈∞	≈∞	≈∞	1.2	1.2	1.1	1.1	1.1	1.1	1.1	1
0.05	≈∞	≈∞	≈∞	≈∞	1.2	1.1	1.1	1.1	1.1	1	1.1	1	1
0.06	≈∞	≈∞	≈∞	≈∞	1.2	1.1	1.1	1.1	1.1	1	1.1	1	1
0.07	≈∞	≈∞	≈∞	≈∞	1.2	1.1	1.1	1	1	1	1	1	1
0.08	≈∞	≈∞	≈∞	1.2	1.1	1.1	1.1	1	1	1	1	1	1
0.09	≈∞	≈∞	≈∞	1.2	1.1	1.1	1.1	1	1	1	1	1	0.99
0.1	≈∞	≈∞	≈∞	1.2	1.1	1.1	1.1	1.03	1.02	1.02	1.03	1.01	1
0.15	≈∞	≈∞	1.2	1.1	1.1	1	1	1	1	1	1	1	0.99
0.2	≈∞	≈∞	1.2	1.08	1.05	1.03	1.03	1.01	1	1	1	1	1
0.25	≈∞	≈∞	1.1	1.1	1.1	1	1	1	1	1	1	1	1
0.3	≈∞	1.2	1.1	1.1	1	1	1	1	1	1	1	1	1
0.35	≈∞	1.2	1.1	1.1	1	1	1	1	1	1	1	1	1
0.4	≈∞	1.2	1.1	1.1	1	1	1	1	1	1	1	1	1
0.45	≈∞	1.2	1.1	1.1	1.03	1	1	1	1	1	1	1	1
0.5	≈∞	1.2	1.1	1.1	1	1.01	1.01	1.01	1.01	1	1	1	1

Note that ≈∞ denotes ORs in the millions.

Table 14. Median odds ratios, reflecting "median unbiasedness" for skewed sampling distributions.

	Sample Size												
	10	50	100	200	300	400	500	600	700	800	900	1000	5000
Probability													
0.01	0	1.08	0.95	0.98	1	0.98	1	1	1	0.99	1.03	0.99	1.01
0.02	≈∞	1	0.96	0.95	0.98	0.97	1	0.98	1	0.97	1	0.99	0.99
0.03	≈∞	1	1	0.97	0.97	0.99	1	1	0.99	0.98	1	1	0.99
0.04	1.8	1	1	1	0.97	1.01	1	1	1	0.97	1	0.99	0.99
0.05	1.7	1	1	1	1	0.98	1	0.98	1	0.98	1	1	0.99
0.06	1.7	1	1	0.97	1.01	0.99	1	0.99	1	0.98	0.99	1	0.99
0.07	1.7	1	1	0.97	1.01	1	0.99	0.98	1	0.98	0.99	0.99	0.99
0.08	1.7	0.92	1	1	1	0.99	1	0.98	1	0.97	1	0.99	0.99
0.09	1.5	1	1	1	1	0.99	1	0.98	1	0.97	1	0.99	0.99
0.1	1.4	1	1	1	1.02	1	1	0.99	1	0.99	0.99	0.99	0.99
0.15	1	1	1	1	1	0.99	1	0.99	0.99	0.98	1	0.99	0.99
0.2	1	1	1.03	1	1	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
0.25	1	1	1	1	1	0.98	0.99	0.99	1	0.99	1	1	1
0.3	1	0.97	1	1	1	1	1	0.98	0.99	1	0.99	0.99	0.99
0.35	1	1	1	1	1	0.99	0.99	0.99	0.99	0.99	1	1	0.99
0.4	1	1	1	1	1	0.99	0.99	0.99	1	1	0.99	1	1
0.45	1	1	1	1	0.99	0.99	0.99	1	1	0.99	0.99	0.99	0.99
0.5	1	1	1	1	0.99	0.99	0.99	1	0.99	0.99	0.99	0.99	0.99

Why is the Type I error rate consistently conservative?

The reader should be reminded that the results in Tables 11 through 14 are based on the converged replications. Overall, the simulation agrees with the previous findings on parameter estimates (Peduzzi et al., 1996), that is, with a small sample size and few events per predictor, the standard error and slope estimates are highly biased. Figure 1 is a line graph that shows the slope (on the left) and standard error (on the right), where the y-axis is the slope or standard error and the x-axis is the skewed probability of the predictor. The line colours represent different sample sizes. Tables 15 and 16 contain the values from which these graphs were derived.

Figure 1. Slope and standard error averages.

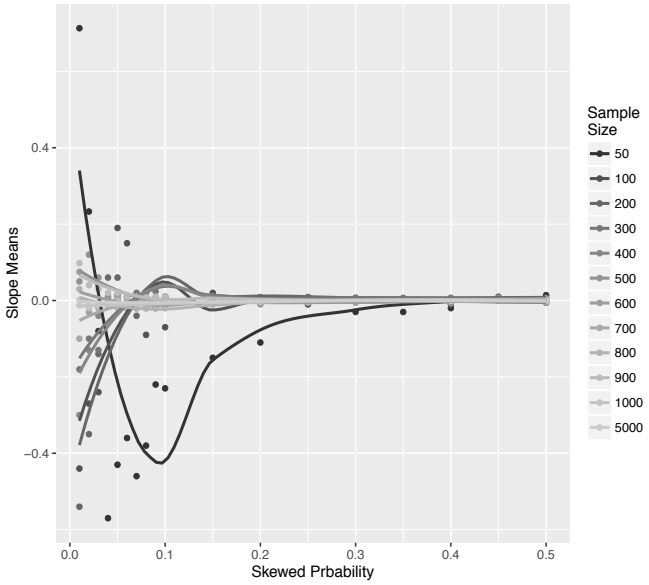


Figure 1.a. Estimated slope means.

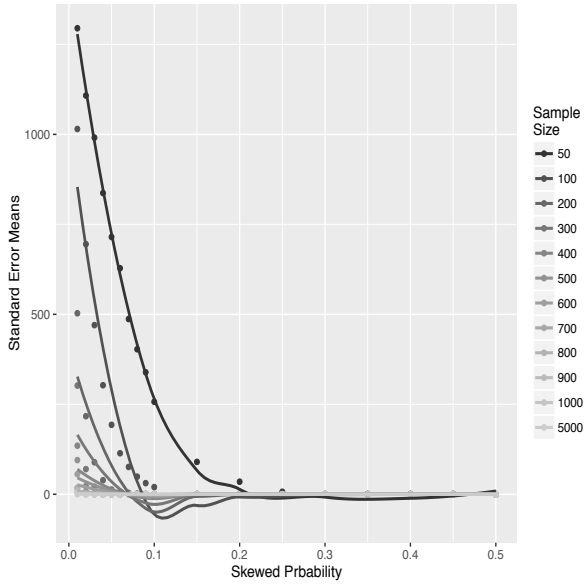


Figure 1.b. Estimated standard error means.

As seen in Figure 1 and Table 15, the bias for the slope is both positive and negative when the sample size and highly skewed cell probability. It should be noted, however, that the bias in slope is not as great as the bias in the ORs (as seen in Table 13). Let us look at a few examples to understand the distribution of the slope parameter estimate, and why it averages out

to a small bias. Let us contrast a small sample size of 50 and a large one of 500 at three levels of skewed cell probability, 0.01, 0.25, and 0.50. The first of these levels represents a highly skewed predictor, the second is moderate, and the last is a balanced probability of both categories in the predictor. Figure 2 shows a stacked density plot for a sample size of 50 and the three levels of skewed probability. For the first level of probability of 0.01, the slope estimate's range is [-17.58, 18.04] with a mean of 0.71, as shown in Table 15. The 25^{th}, 50^{th} and 75^{th} quantiles are -15.52, 0.083, and 15.52, respectively. As indicated in Figure 2, the distribution of the slope estimates from this simulation is fragmented into three parts, such that there are no slope estimates that lie between them. Most of the slope estimates were in the range of [-17.58, -14.75]; the least number were in the range of [-1.17, 0.78]; and the rest, which comprised the last part, ranged from [14.93, 18.04]. For the same sample size and a skewed cell probability of 0.25, the shape of the distribution of the slope estimates mostly varies around zero, with a few outliers in the tails. The range is [-18.42, 18.62] and the mean is -0.012. The 25^{th}, 50^{th}, and 75^{th} quantiles are -0.51, -1.0×e-16, and 0.43, respectively.

Figure 2. Distribution functions for the experimental condition: Sample size = 50, skewed probability = 0.01, 0.25, and 0.5.

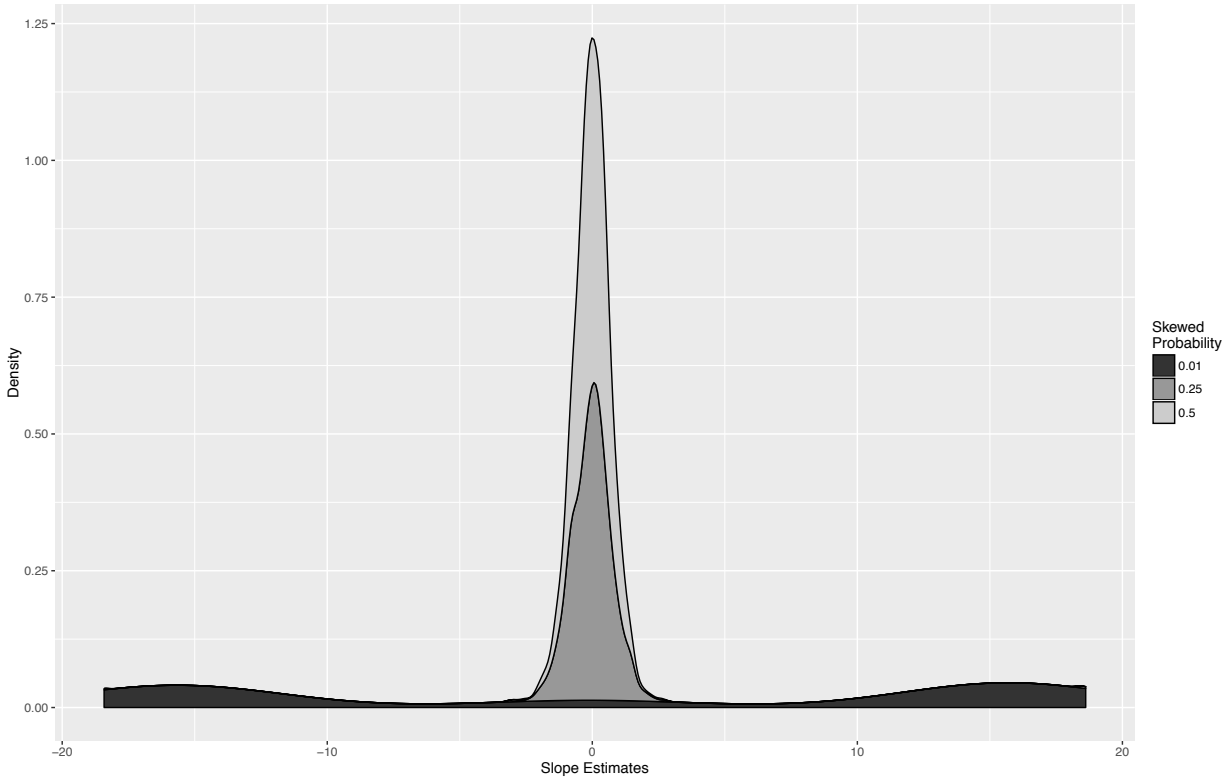


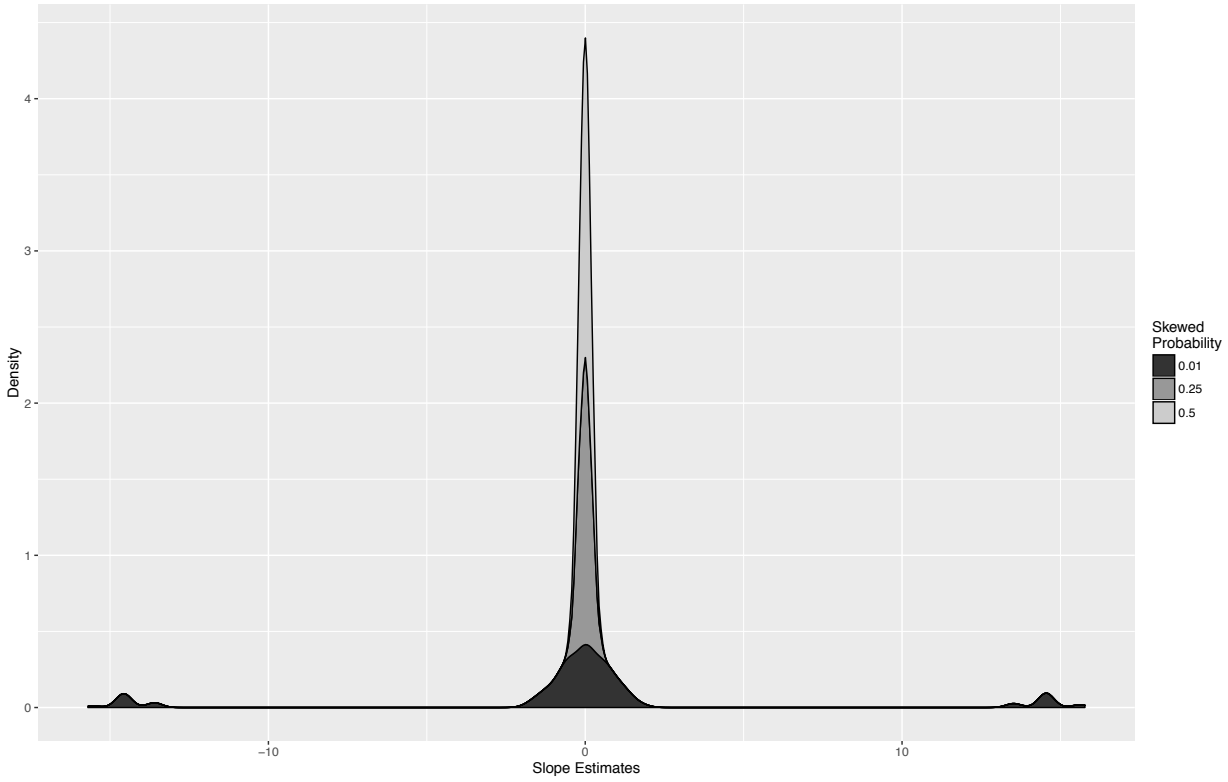
Table 15. Average slope in each cell of the simulation design for Model 1.

	Sample Size												
	10	50	100	200	300	400	500	600	700	800	900	1000	5000
Probability													
0.01	-.36	.71	-.44	-.54	-.18	-.3	.05	.08	-.1	.03	.1	-.01	0
0.02	1.2	.23	-.27	-.35	-.1	-.13	.12	-.03	-.01	-.01	.04	-.02	-.01
0.03	1.7	-.08	-.24	-.14	-.13	-.04	.06	-.01	-.02	-.02	.03	-.01	0
0.04	1.3	-.57	-.01	.06	-.01	.01	.01	0	-.01	-.01	.02	-.01	0
0.05	1.4	-.43	.19	.06	.01	0	.01	-.01	0	-.02	.02	-.01	0
0.06	1.1	-.36	.15	-.02	.01	0	.01	-.01	0	-.02	.01	-.01	-.01
0.07	1	-.46	-.02	-.04	.02	.01	.01	-.01	0	-.02	.01	-.01	0
0.08	1	-.38	-.09	-.02	.02	.01	0	-.01	0	-.02	.01	-.01	0
0.09	.9	-.22	-.02	0	.02	.01	0	-.01	-.01	-.02	.01	-.01	-.01
0.1	.8	-.23	-.07	.01	.01	.01	0	-.01	-.01	-.02	.01	-.01	0
0.15	.3	-.15	.02	.01	0	.01	0	0	-.01	-.01	.01	0	0
0.2	.02	-.11	0	.01	.01	0	-.01	0	0	0	0	0	0
0.25	-.04	-.01	0	.01	.01	0	0	-.01	0	0	0	0	0
0.3	.12	-.03	0	.01	.01	0	0	-.01	0	0	0	0	0
0.35	.24	-.03	0	.01	.01	0	0	0	0	0	0	0	0
0.4	.27	-.02	-.01	.01	.01	-.01	-.01	0	0	0	0	0	0
0.45	-.01	0	0	.01	0	0	.01	0	0	0	0	0	0
0.5	-.13	.01	0	.01	0	-.01	0	0	.01	0	0	0	0

For comparison purposes, I also examined the distribution of the slope estimates when the probability of the predictor is balanced (i.e., p = 0.5). As Table 15 shows, the estimated slopes are close to the simulated values of zero. As shown in Figure 2, the distribution in this experimental condition is nearly symmetrical, with a range of [-2.59, 2.58] and a mean of 0.014. All of this suggests that the distribution of the simulated slope estimates is disrupted by the skewness in the probability of the predictor.

Now let us take an example where the sample size is large, in this case 500, and examine the extent to which the distribution of the slope parameter changes with the aforementioned three levels of probability. Figure 3 demonstrates the stacked density plots for the three experimental conditions. For a probability of 0.01, the distribution is fragmented into three parts that cluster around zero. The distribution range is [-15.7, 15.77] and the mean is 0.052. The 25^{th}, $50^{th,}$ and 75^{th} quantiles are -0.66, 0.008, and 0.696, respectively. For the same sample size and a moderate probability of 0.25, I find that the distribution is symmetrical and nearly resembles a normal distribution. The slope estimates vary close to zero (the actual value), with a mean of 0.004 and a range of [-0.63, 0.62]; the 25^{th}, $50^{th,}$ and 75^{th} quantiles are -0.15, -0.005, and 0.14, respectively. Finally, when the sample is 500 and the predictor is balanced, the distribution is tighter and varies closer to zero. It has a range of [-0.57, 0.55] with three outliers equal to 1.57, 4.81, and 7.87. The mean, as seen in Table 15, is -0.004 and the 25^{th}, $50^{th,}$ and 75^{th} quantiles are -0.123, 0.0003, and 0.12, respectively.

Figure 3. Distribution functions for the experimental condition: Sample size = 500, skewed probability = 0.01, 0.25, and 0.5.



This wide range of the slope estimate when the skewed probability of a predictor is small clarifies a few things about the aforementioned small bias, and the largely upward bias of the ORs. Since the ORs are the exponentiation of the slope estimate, slopes with large positive values can create ORs that are in the order of magnitude of tens of millions, whereas negative slopes can result in ORs that tend toward zero. Therefore, the upwardly tending slopes will result in very large bias in the ORs.

Table 16 depicts the average standard errors over replicates of the simulation. These standard errors range from highly biased to unbiased, with the concentration of high bias for small sample sizes and highly skewed cell probabilities (i.e., the top left corner of the table). For example, for sample sizes of 50 to 300 and probability levels of 0.01-0.2, the average standard

errors are in the thousands and range from [0.24, 4500], as shown in Table 16. As we learned from examining the distributions of the slope estimates above, sample sizes of at least 400 perform very well and provide estimates closer to the simulated values when the skewed probability is at least 0.04. This supports the claim that the maximum likelihood estimation is affected by the skewed probabilities of the predictor. That is, even in replications where the MLE produced finite estimates, there was bias in the parameter estimates and standard error. However, as the sample size increases, the estimation becomes less influenced by the skewness.

Table 16. Average standard error of the slope in each cell of the simulation design for Model 1.

	Sample Size												
	10	50	100	200	300	400	500	600	700	800	900	1000	5000
Probability													
0.01	3956	1295	1015	503	302	135	95	56	18	16	8	4	.28
0.02	4522	1108	695	217	70	23	8	3	1.1	.54	.5	.48	.2
0.03	4221	991	470	89	21	4	2	.5	.46	.43	.4	.38	.16
0.04	4360	837	303	39	6	.54	.48	.43	.39	.37	.35	.33	.15
0.05	414	715	193	13	.57	.48	.43	.38	.36	.33	.31	.29	.13
0.06	4053	628	114	7	.51	.44	.38	.35	.32	.3	.28	.27	.12
0.07	3914	487	76	4	.47	.4	.36	.33	.3	.28	.27	.25	.11
0.08	3836	403	49	2	.44	.38	.34	.31	.28	.26	.25	.23	.1
0.09	3830	339	31	.52	.42	.36	.32	.29	.27	.25	.24	.22	.1
0.1	3525	257	20	.49	.4	.34	.3	.28	.26	.24	.22	.21	.09
0.15	3617	90	.59	.41	.33	.28	.25	.23	.21	.2	.18	.18	.08
0.2	3076	35	.52	.36	.29	.25	.22	.21	.19	.18	.17	.16	.07
0.25	2434	7	.47	.33	.27	.23	.21	.19	.17	.16	.15	.15	.07
0.3	2067	.65	.44	.32	.25	.22	.19	.18	.17	.15	.14	.14	.06
0.35	1744	.62	.43	.3	.24	.2	.18	.17	.16	.15	.14	.13	.06
0.4	1608	.6	.42	.29	.24	.21	.18	.17	.15	.14	.14	.13	.06
0.45	1535	.59	.41	.28	.23	.2	.18	.16	.15	.14	.13	.13	.06
0.5	1483	.59	.41	.28	.23	.2	.18	.16	.15	.14	.13	.12	.06

Regardless of the low bias in the slope estimates, when the Wald statistic is calculated, the bias of the denominator is very high and outweighs the negligible bias of the numerator. This results in a Wald statistic that will likely not reject the null hypothesis, resulting in a conservative test. For instance, for a sample size of 100 and a skewed probability of .04 (which is quite a skewed predictor), the numerator of the Wald statistic is not highly biased, but the denominator is, resulting in a Type I error rate of zero.

Bringing all of this evidence together, it should be noted that even though some modest bias exists in the parameter estimates, the conservative Type I error rates are clearly driven by the large standard errors. The apparent contradiction between a conservative Type I error rate and a highly inflated OR is best understood by examining the shape of the sampling distribution of the slope, wherein the large values of some of the replications with a cell of the experimental design influence the average value of the ORs. This is best seen by contrasting Tables 13 and 14 with the mean and median ORs, respectively.

In the extreme case of a sample size of 10, the average slope deviates far from the simulated value, even when the predictor is balanced. Likewise, the standard errors are always upwardly biased in the order of magnitude of the thousands. Because of these obvious biases and the impracticality of such a small sample size, it was removed from further analyses.

Study B: Power.

Purpose of the study. Usually, a low probability of Type I errors is accompanied by low statistical power. Therefore, the next step is to examine the statistical power of the Wald test of the slope parameter for this model. Although there is no agreement on what magnitude of effect (i.e., effect size) is necessary to establish practical significance, Ferguson (2009) suggests three values related to risk estimates, i.e., measures comparing relative risk for a particular outcome

between two or more groups. According to Ferguson (2009), ORs of 2, 3, and 4 represent small, moderate, and large effect sizes, respectively.

Methods.

Simulation factors and methodology. In addition to skewness and sample size, a third factor was manipulated in this study. As in the previous simulation, the sample size varied across 13 levels ranging from 50 to 5000, and the probability of the occurrence of a category in the predictor from 0.01 to 0.45. For comparison purposes, I investigated what happens when the predictor variable is balanced. The third factor that was added is effect size, which varied from small, moderate, and large. The resulting experiment is an 18×12×3 completely crossed factorial design.

In this simulation, the estimation is built on the assumption that this model has an effect. In other words, we are assuming beforehand that β_0 and β_1 are fixed to a number different from zero. The intercept parameter was fixed to -2. The slope parameter was fixed at three levels of effect size: small effect of 0.683 (equivalent to OR=2), moderate effect of 1.1 (equivalent to OR=3), and large effect of 1.38 (equivalent to OR = 4). As in Study A, each cell includes 1000 replications of the same model.

Analysis of simulation results. To assess power estimates, I adopted a framework similar to Bradley's for Type I error rates. That is, we investigated at what level of skewness we lost 10% and 50% of the expected statistical power. The expected statistical power was identified as the power in the case of the balanced cell probability of the predictor. Hence, the estimated statistical power for each cell in the experiment is compared to the power for the same sample and effect size but with no skewness in the predictor's probability.

Results and conclusions.

Tables 17, 18, and 19 follow the structure of previous tables and show the statistical power for each effect size level. The last row in each table is the power estimate for the balanced predictor variable. The tables are greyscale coded: no shading reflects losing 10% of power or less, light shading reflects losing 10%-50% of power, and dark shading reflects losing over 50% of power.

Because sample size and effect size both significantly influence power, it is not surprising that as these two factors increase, power also increases. From the tables, we can see that for a balanced predictor, an effect size of OR = 2, and a sample size of 200 or less, the power of the Wald test is less than 50%. It exceeds 50% after a sample size of 300, exceeds 75% after a sample size of 500, and reaches one after a sample size of 1000. Moreover, the tables indicate that as the effect size grows, there is less of a need for larger sample sizes to detect the effects. For example, with a balanced predictor and a sample of 100, the statistical power is nearly 75% to detect an OR of 4, while it is 52% and 21% for ORs of 3 and 2, respectively.

For a low effect size, samples from 100 to 1000, and a skewed cell probability less than or equal to 0.2, over 10% of power is lost compared to the balanced cases. As the effect size increases, the level of skewed probability that is tolerated slightly increases. For example, to retain 10% of power for sample sizes of 100-300, the level of probability tolerated for a low effect size ranges between 0.2 and 0.3. However, the level of skewed cell probability needed to retain 10% of power for a high effect is 0.15-0.25 for the same sample sizes. It can be concluded that power is highly influenced by skewed probabilities in small sample sizes, even when the effect size is moderate to large and hence highly detectable.

On the other hand, sample sizes of 400 and over can retain 10% of power with low levels of skewness. For example, a sample of 500 with an OR of 2 retains 10% of power at a probability of 0.3. As the effect size increases to an OR of 3, 10% of power is retained at level 0.15; at an OR of 4, 10% is retained at level 0.07. A sample size of 1000 with an OR of 2 retains 10% of power at level 0.2. The probability level that retains the same percentage of power quickly jumps to 0.06 and 0.04 for ORs of 3 and 4, respectively.

Power is largely affected by the skewed probability of the predictor variable. That is, high levels of skewness [p = 0.01, 0.1] result in the loss of over 50% of power in most sample sizes, with the exception of 1000 and 5000. Although the previous study showed that Type I error is acceptable for the majority of factors examined, it has implications on the empirical power of the Wald test. We have provided evidence that suggests that statistical power is biased downwards with the combination of smaller samples and higher degrees of skewed probability. Clearly, the skewed probability of a predictor diminishes the statistical power of the Wald test.

Table 17. Power with low effect size (OR = 2).

	Sample Size											
	50	100	200	300	400	500	600	700	800	900	1000	5000
Probability												
0.01	0	.01	.04	.06	.08	.1	.13	.13	.15	.17	.17	.5
0.02	.01	.04	.09	.12	.13	.17	.18	.19	.22	.25	.27	.75
0.03	.03	.05	.11	.16	.16	.21	.25	.27	.29	.32	.32	.87
0.04	.03	.07	.12	.18	.21	.24	.3	.32	.34	.38	.41	.94
0.05	.05	.09	.14	.21	.24	.28	.35	.39	.4	.45	.47	.98
0.06	.06	.11	.16	.24	.26	.33	.38	.43	.45	.49	.52	.99
0.07	.08	.13	.17	.26	.28	.36	.42	.45	.49	.55	.59	.99
0.08	.08	.14	.2	.28	.32	.4	.46	.51	.54	.59	.63	.99
0.09	.09	.15	.21	.3	.31	.41	.49	.54	.58	.64	.67	1
0.1	.09	.17	.22	.32	.36	.45	.52	.57	.62	.67	.72	1
0.15	.12	.18	.3	.39	.49	.59	.65	.7	.74	.81	.85	1
0.2	.13	.22	.34	.46	.56	.67	.73	.8	.85	.87	.91	1
0.25	.13	.21	.37	.52	.64	.72	.78	.86	.9	.91	.94	1
0.3	.13	.22	.39	.55	.68	.76	.83	.89	.92	.94	.96	1
0.35	.11	.23	.42	.58	.69	.78	.84	.9	.93	.94	.98	1
0.4	.1	.22	.43	.6	.72	.8	.85	.91	.95	.96	.97	1
0.45	.09	.22	.44	.6	.73	.81	.87	.93	.96	.96	.98	1
0.5	.06	.21	.42	.61	.7	.8	.87	.92	.96	.96	.98	1

Table 18. Power with moderate effect size (OR = 3).

	Sample Size											
	50	100	200	300	400	500	600	700	800	900	1000	5000
Probability												
0.01	.01	.02	.07	.11	.15	.19	.26	.26	.29	.35	.36	.88
0.02	.01	.07	.15	.23	.29	.35	.41	.43	.45	.54	.54	.99
0.03	.04	.11	.22	.31	.38	.44	.51	.55	.6	.67	.69	1
0.04	.05	.14	.27	.37	.45	.53	.63	.66	.71	.79	.79	1
0.05	.07	.18	.34	.44	.53	.61	.68	.76	.78	.85	.85	1
0.06	.09	.21	.36	.5	.58	.69	.74	.82	.84	.89	.9	1
0.07	.12	.24	.41	.56	.64	.73	.78	.86	.88	.93	.93	1
0.08	.14	.27	.44	.6	.69	.77	.83	.9	.92	.95	.95	1
0.09	.15	.29	.48	.62	.74	.82	.85	.92	.93	.97	.97	1
0.1	.16	.32	.52	.66	.78	.85	.89	.94	.96	.98	.98	1
0.15	.22	.4	.65	.78	.88	.94	.97	.98	.99	1	1	1
0.2	.25	.46	.7	.86	.95	.97	.99	.99	1	1	1	1
0.25	.26	.47	.77	.9	.97	.99	.99	1	1	1	1	1
0.3	.28	.51	.8	.94	.98	.99	1	1	1	1	1	1
0.35	.28	.53	.83	.94	.99	1	1	1	1	1	1	1
0.4	.27	.52	.85	.95	.99	1	1	1	1	1	1	1
0.45	.27	.52	.87	.95	.99	1	1	1	1	1	1	1
0.5	.24	.52	.87	.96	.99	1	1	1	1	1	1	1

Table 19. Power with large effect size (OR = 4).

	Sample Size											
	50	100	200	300	400	500	600	700	800	900	1000	5000
Probability												
0.01	0	.02	.07	.15	.23	.29	.36	.4	.42	.49	.51	.98
0.02	.01	.08	.21	.34	.42	.52	.58	.61	.66	.73	.76	1
0.03	.05	.15	.31	.44	.55	.65	.71	.76	.81	.87	.87	1
0.04	.07	.2	.4	.53	.65	.75	.81	.85	.89	.94	.94	1
0.05	.1	.25	.49	.62	.74	.82	.86	.9	.95	.97	.98	1
0.06	.13	.3	.54	.69	.8	.87	.9	.94	.97	.99	.98	1
0.07	.16	.34	.59	.74	.84	.91	.93	.96	.98	1	.99	1
0.08	.19	.4	.65	.7	.88	.93	.95	.98	.99	1	1	1
0.09	.21	.41	.69	.83	.91	.95	.97	.98	1	1	1	1
0.1	.24	.46	.72	.85	.93	.97	.98	.99	1	1	1	1
0.15	.32	.58	.84	.93	.99	1	1	1	1	1	1	1
0.2	.38	.66	.9	.97	.99	1	1	1	1	1	1	1
0.25	.39	.69	.93	.99	1	1	1	1	1	1	1	1
0.3	.44	.73	.95	.99	1	1	1	1	1	1	1	1
0.35	.45	.74	.98	1	1	1	1	1	1	1	1	1
0.4	.44	.75	.97	1	1	1	1	1	1	1	1	1
0.45	.46	.75	.98	1	1	1	1	1	1	1	1	1
0.5	.43	.75	.98	1	1	1	1	1	1	1	1	1

Model 2: Single Continuous Predictor

The second model of interest involves simple logistic regression with one continuous predictor.

$$g(y) = \beta_0 + \beta_1 x,$$

where x is a skewed continuous variable, β_0 and β_1 are fixed, $g(y)$ is a logit function, and y is a balanced outcome variable. The purpose of this model was to investigate whether findings from Model 1 (which used one binary predictor) would generalize to a skewed continuous predictor. That is, we wish to determine whether issues with the skewness of the predictor are related to the categorical versus the numeric aspect of the variable—whether it is the skew or the binary nature that is causing the effect on the Type I error. To confirm this, I focused only on the Type I error rate because its reduction is accompanied by a corresponding reduction in statistical power. Therefore, a decreased Type I error rate is diagnostic of a problem with decreased power.

Study A: Type I error rates.

Purpose of the study. Similarly to the previous study, I wish to document the impact of a skewed continuous predictor variable on the estimation, parameter estimates, and Type I error rate of the Wald test.

***Methods*.**

The simulation factors, methodology, and analysis of the Type I error rate are exactly the same as in the previous study. The only difference is in the nature of the predictor. This variable was generated from a Gamma distribution with the rate and scale parameters fixed to 1 and varying the shape parameter across 17 levels. Skewness in the gamma distribution is a function of the shape parameter. To enable comparison, the skewness levels for this model to the expected probabilities in Model 1 were matched. Table 20 shows the shape parameter values used and the

equivalent skewness levels. To create a baseline, the case where the predictor variable is drawn from a standard normal distribution with a mean of zero and a standard deviation of one was also examined. The resulting simulation experiment is a 12 (sample size) by 18 (skewness) completely crossed factorial design.

Table 20. Shape parameter and equivalent skewness level.

Shape Parameter	Skewness	Probability
0.047	9.25	0.01
0.086	6.87	0.02
0.13	5.55	0.03
0.2	4.75	0.04
0.25	4.04	0.05
0.3	3.7	0.06
0.37	3.25	0.07
0.4	3.19	0.08
0.5	2.83	0.09
0.6	2.72	0.1
1	1.96	0.15
1.75	1.5	0.2
3	1.15	0.25
5.5	0.873	0.3
10	0.63	0.35
25	0.41	0.4
50	0.2	0.45
Standard N	0	0.5

Results and conclusions.

It is not surprising that with a continuous predictor, all of the replications converged for the 216 conditions of the simulation experiment. Table 21, which is formatted similarly to the previous tables, shows that the Type I error rates ranged from 0.001 to 0.066, with an average of 0.043. The majority of the conditions met the liberal criterion, but not the stringent one. As can be seen from Table 21, in only a few cases did the Type I error rate fall below 0.025, as dictated by the liberal criterion. These instances are with sample size 50 with skewness ≥ 2.6, sample size 100 with skewness ≥ 5.54, and sample sizes 200 and 300 with the highest skewness level (9.25).

Table 21. Liberal Type I error rate model.

	Sample Size											
	50	100	200	300	400	500	600	700	800	900	1000	5000
Skewness												
9.25	0	0	.01	.01	.02	.03	.03	.05	.02	.03	.04	.04
6.87	0	.01	.02	.03	.03	.05	.03	.04	.03	.04	.04	.06
5.55	.01	.02	.03	.04	.04	.04	.05	.05	.04	.05	.04	.05
4.75	.01	.03	.04	.04	.03	.05	.05	.04	.05	.06	.05	.04
4.04	.01	.03	.04	.04	.04	.05	.04	.06	.04	.04	.05	.05
3.7	.02	.03	.04	.04	.04	.04	.05	.04	.04	.06	.05	.05
3.25	.02	.03	.04	.04	.06	.04	.05	.06	.05	.03	.06	.06
3.19	.02	.04	.04	.05	.04	.04	.04	.04	.05	.04	.06	.05
2.83	.03	.03	.05	.04	.05	.04	.05	.05	.04	.06	.05	.05
2.72	.02	.03	.04	.05	.05	.05	.05	.05	.06	.05	.04	.05
1.96	.04	.03	.04	.04	.04	.06	.06	.05	.06	.04	.05	.05
1.5	.05	.03	.05	.04	.04	.04	.05	.05	.05	.05	.06	.05
1.15	.04	.05	.05	.04	.05	.06	.06	.05	.06	.07	.06	.04
0.873	.04	.06	.04	.05	.05	.05	.05	.04	.04	.04	.04	.06
0.63	.04	.04	.04	.05	.04	.05	.05	.05	.06	.05	.05	.06
0.41	.04	.04	.04	.05	.06	.04	.04	.07	.06	.06	.06	.05
0.2	.03	.04	.04	.06	.05	.04	.06	.05	.05	.05	.04	.05
0	.05	.06	.06	.04	.05	.06	.05	.05	.06	.06	.05	0.06

It can be observed that the estimation tolerated a skewed continuous predictor a great deal better than a dichotomous one. The same conclusions from the previous study can be drawn here in that as the sample size increases and the skewness becomes smaller, the Type I error rate gets closer to the nominal value. Hence, as with a dichotomous predictor, a highly skewed continuous predictor affects the estimation and inferences in the extreme case of a small sample size.

Model 3: Multiple Logistic Regression with Two Independent Binary Predictors

The last model investigated in this dissertation is a multiple logistic regression with two dichotomous predictors,

$$g(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

where x_1 and x_2 are independent dichotomous variables with skewed probabilities; β_0, β_1 , and β_2 are fixed, $g(y)$ is a logit function and y is a balanced outcome variable. The goal in including

this model was to discover whether the skewed probability of one predictor could alter the parameter estimates of other variables in the model when the two predictors are independent. This model reflects, for example, what one would have with a 2×2 (two-factor) randomized experiment or randomized clinical trial.

Study A: Type I error rates and non-convergences.

Methods.

The simulation methodology and analysis of the type I error rate were the same as in Models 1 and 2. Three factors were manipulated in this experiment. The first two are sample size and the probability of x_1, while the additional factor is the probability of x_2. The probability for each predictor varied from 0.01 to 0.045. As in previous studies, the case where the variables were balanced was also examined for comparison purposes. The resulting experiment is a 12×18×18 completely crossed factorial design. Again, as in earlier models, the empirical and nominal Type I error rates were compared using Bradley's criteria.

Results and conclusions.

We investigated 4212 experimental conditions. Many of these results are identical or a couple of decimals apart. Because of the sheer volume and the lack of variation, a few sample sizes were presented as examples, while noting the others in the text.

Number of non-convergences. Tables 22 and 23 show the number of non-converging replications with varying degrees of skewed probability on both predictors for samples of 100 and 400, respectively. As in previous studies, for a sample of 100 and a probability of 0.01, most of the replications did not converge. All replications converged when the probability of x_1 is 0.09 or higher and the probability of x_2 is equal or higher than 0.07. For sample sizes 100 and

400, the number of non-converging replications decreases as the probability of both variables becomes more balanced. This issue ceases to be important for samples of 900 and 5000.

Table 22. Number of non-convergences from 1000 replications for Model 3 when sample size is 100.

		x_2 *Probability*								
		0.01	0.02	0.03	0.04	0.05	0.06	0.07	⋯	0.5
x_1 *Probability*	0.01	610	463	412	397	390	389	388	⋯	388
	0.02	474	260	195	165	157	155	154	⋯	154
	0.03	407	176	102	70	62	60	58	⋯	58
	0.04	384	143	68	33	23	21	19	⋯	19
	0.05	374	131	56	21	11	9	7	⋯	7
	0.06	373	129	54	19	9	7	5	⋯	5
	0.07	371	127	52	16	6	4	2	⋯	2
	0.08	371	127	52	16	6	3	1	⋯	1
	0.09	370	126	51	15	5	2	0	⋯	0
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱	0
	0.5	370	126	51	15	5	2	0	0	0

Table 23. Number of non-convergences from 1000 replications for Model 3 when sample size is 400.

		x_2 *Probability*				
		0.01	0.02	0.03	⋯	0.5
x_1 *Probability*	0.01	41	21	21	⋯	21
	0.02	21	1	1	⋯	1
	0.03	20	0	0	⋯	0
	⋮	⋮	⋮	⋮	⋱	0
	0.5	20	0	0	0	0

Type I error rate. I compared the effect of the skewness of x_2 on the Type I error rate for the Wald test of x_1. The results in Table 24 are formatted somewhat differently from those in other tables in this chapter. Following Conover, Johnson, and Johnson (1981), average Type I error rates were used. The table presents the Type I error rate of the Wald test for x_1 averaged

across all levels of skewness of x_2 for sample size 100, 400, 900, and 5000 to represent the small, medium, and large sample sizes found in the literature. Like the tables in Models 1 and 2, Table 24 is greyscale coded, the darkly shaded areas falling below Bradley's liberal criterion and the unshaded ones falling within it. The Type I error was consistently deflated, with lower levels of probability for x_1 and x_2, growing closer to the nominal value as the skewed probability for both predictors became more balanced. It satisfies the liberal criterion with ranges of [0, 0.06] for a sample size of 100, [0.002, 0.068] for a sample size of 400, [0.017, 0.065] for a sample size of 900, and [0.033, 0.057] for a sample size of 5000.

The range of average Type I error rates for each cell in Table 24 does not vary greatly. Therefore, it is clear that the degree of skewness of the cell probability of x_2, has little to no impact on the Type I error rate of x_1. In other words, the Type I error rate of x_1 with low probability on x_2 does not differ from the Type I error rate of x_1 when x_2 is balanced. For example, for a sample of 400, the Type I error rate for x_1 is 0.035 when the probability of $x_1 = 0.04$ and the probability of x_2 = 0.03 or 0.45. Comparing the Type I error rate of x_1 in Model 1 with Model 3, we can see that factoring in the skewness of an additional variable that is completely independent from other variables in this model has minimal impact.

Table 24. Type I error rate for x_1 averaged across all levels of x_2, and the range of Type I errors across all levels of the skewed probability of x_2.

		Sample Size			
		100	400	900	5000
x_1 Probability	0.01	0 (0,0)	.002 (.001,.003)	.019 (.017,.021)	.053 (.051,.057)
	0.02	0 (0,0)	.015 (.012,.018)	.039 (.035,.043)	.041 (.035,.042)
	0.03	0 (0,0)	.023 (.021,.026)	.043 (.04,.044)	.043 (.041,.045)
	0.04	0 (0,0)	.036 (.033,.039)	.045 (.042,.046)	.036 (.033,.052)
	0.05	.003 (.001,.005)	.041 (.038,.043)	.058 (.056,.06)	.04 (.039,.041)
	0.06	.01 (.005,.011)	.039 (.036, .046)	.047 (.045,.049)	.047 (.046,.049)
	0.07	.01 (.01,.02)	.046 (.043,.049)	.056 (.053,.059)	.045 (.043,.047)
	0.08	.019 (.016,.022)	.05 (.048,.054)	.063 (.061,.065)	.049 (.049,.051)
	0.09	.019 (.017,.023)	.05 (.048,.055)	.062 (.06,.064)	.038 (.038,.04)
	0.1	.023 (0,.03)	.049 (.047,.052)	.052 (.05,.055)	.047 (.046,.049)
	0.15	.037 (.034,.04)	.062 (.06,.065)	.055 (.052,.061)	.049 (.049,.051)
	0.2	.045 (.034,.048)	.049 (.047, .057)	.045 (.042,.047)	.042 (.041,.043)
	0.25	.041 (.035,.044)	.048 (.046,.049)	.047 (.045,.05)	.052 (.051,.053)
	0.3	.047 (.039,.053)	.054 (.052,.056)	.038 (.036,.041)	.055 (.054,.056)
	0.35	.056 (.039,.06)	.053 (.052,.055)	.048 (.046,.05)	.047 (.046,.048)
	0.4	.049 (.047,.052)	.066 (.062,.069)	.053 (.05,.054)	.05 (.035,.055)
	0.45	.049 (.037,.052)	.058 (.055,.061)	.04 (.038,.045)	.049 (.047,.051)
	0.5	.056 (.052,.059)	.057 (.052,.06)	.047 (.046,.049)	.042 (.042,.044)

Study B: Power.

Purpose of the study. Pursuant to Study A, this simulation was designed to investigate the impact of two independent dichotomous predictors with skewed probabilities on the power of the Wald test. Ferguson's (2009) suggestion for a small, moderate, and large effect size were applied to ORs.

Methods.

In addition to the three factors mentioned in Study A, a fourth factor was added. The manipulated variables were sample size, the probability of both predictors, and effect size, which was either small, moderate, or large. The resulting experiment is a 12×18×18×3 completely crossed factorial design.

The simulation procedure is the same, with the added assumption of model effect. We already assume that β_0, β_1, and β_2 are fixed to a number different from zero. The intercept parameter was fixed to -2. Three levels of effect size were examined for β_1: small effect: 0.683 (equivalent to OR=2), moderate effect: 1.1 (equivalent to OR=3), and large effect: 1.38 (equivalent to OR=4), However, the effect size for β_2 to a moderate value of 1.1 was fixed. The simulation methodology is similar to that in Study A, with the addition of an extra loop to account for effect size.

Results and conclusions.

To analyze power estimates, I compared the best achievable power in the case of a balanced design to other combinations of probabilities for each sample size. As well as examined 10% and 50% loss of power. The resulting number of conditions was 12636. Similar to Study A, four typical sample sizes were presented: 100, 400, 900, and 5000. Tables 25 through 27 show

power estimates for x_1 averaged over all levels of probability of x_2. As in Table 24, Tables 25 through 27 include the range of statistical power for each condition.

The first direct finding is that the statistical power of x_1 is not affected by changes in the probability of x_2, but rather, is affected by its own skewed probability regardless of the sample and effect sizes. Looking at the range of statistical power for each condition, we can see that changes in the power of x_1 when the probability of x_2 is at its extreme are within [0.01, 0.03] of the power when the probability of x_2 is balanced. On the other hand, it is evident that the power of x_1 is highly influenced by its own skewed probability. For example, the highest achievable power for this model under the circumstances identified in this simulation for x_1 with a sample of 100 and a small effect size is 0.3, as seen in Table 25. Power is dramatically affected by the deflation in the Type I error rate. This study shows that with a skewed probability of 0.01, the power of the Wald test for the same predictor plummets to a range of [0, 0.5] for all sample and effect sizes. Similar to the findings in Model 1, as the sample and effect sizes increase, the skewed probability tolerance accelerates significantly.

Table 25. Statistical power at an OR = 2 for x_1 averaged across all levels of x_2, and the range of statistical power errors across all levels of the skewed probability of x_2.

		Sample Size			
		100	400	900	5000
x_1 Probability	0.01	.01 (0,.013)	.1 (.074,.11)	.17 (.16,.18)	.53 (.5,.6)
	0.02	.04 (.02,.21)	.14 (.14,.15)	.26 (.24,.28)	.8 (.77,.85)
	0.03	.05 (.04,.055)	.166 (.15,.18)	.33 (.31,.36)	.92 (.89,.95)
	0.04	.08 (.065,.085)	.213 (.2,.23)	.41 (.34,.45)	.97 (.95,1)
	0.05	.09 (.011,.11)	.26 (.23,.28)	.47 (.44,.53)	.99 (.98,1)
	0.06	.1 (.096,.011)	.3 (.27,.397)	.53 (.49.59)	.99 (.99,1)
	0.07	.12 (.11,.13)	.32 (0.3,0.35)	.57 (.53,.66)	1 (.99,1)
	0.08	.13 (.12,.14)	.36 (.34,.39)	.62 (.58,.7)	1 (.99,1)
	0.09	.15 (.13,.16)	.38 (.36,.41)	.66 (.62,.73)	1 (1,1)
	0.1	.15 (.14,.17)	.4 (.38,.44)	.7 (.66,.78)	1 (1,1)
	0.15	.18 (.16,.2)	.52 (.48,.58)	.84 (.81,.89)	1 (1,1)
	0.2	.197 (.17,.23)	.6 (.56,.67)	.91 (.89,.95)	1 (1,1)
	0.25	.218 (.19,.26)	.67 (.62,.75)	.95 (.93,.98)	1 (1,1)
	0.3	.23 (.19,.29)	.72 (.68,.79)	.96 (.95,.98)	1 (1,1)
	0.35	.23 (.2,.31)	.74 (.69,.82)	.97 (.96,.98)	1 (1,1)
	0.4	.25 (.22,.31)	.76 (.72,.83)	.98 (.97,.99)	1 (1,1)
	0.45	.25 (.2,.3)	.76 (.71,.84)	.98 (.96,.99)	1 (1,1)
	0.5	.246 (0.21,.3)	.76 (.71,.85)	.98 (.97,.99)	1 (1,1)

Table 26. Statistical power at an OR = 3 for x_1 averaged across all levels of x_2, and the range of statistical power errors across all levels of the skewed probability of x_2.

		Sample Size			
		100	400	900	5000
x_1 Probability	0.01	.02 (.01,.3)	.15 (.12,.17)	.33 (.32,.35)	.9 (.88,.93)
	0.02	.05 (.04,.06)	.28 (.27,.29)	.54 (.52,.56)	.99 (.98,1)
	0.03	.09 (.08,.1)	.38 (.37,.4)	.68 (.66,.73)	.99 (.99,1)
	0.04	.14 (.12,.15)	.46 (.44,.48)	.79 (.77,.83)	.99 (.99,1)
	0.05	.18 (.18,.2)	.53 (.51,.57)	.86 (.84,.9)	1 (1,1)
	0.06	.22 (.21,.23)	.6 (.57,.63)	.9 (.88,.94)	1 (1,1)
	0.07	.25 (.24,.27)	.66 (.64,.71)	.94 (.93,.95)	1 (1,1)
	0.08	.27 (.26,.29)	.71 (.69,.75)	.96 (.95,.97)	1 (1,1)
	0.09	.3 (.29,.33)	.76 (.74,.78)	.97 (.96,.98)	1 (1,1)
	0.1	.32 (.3,.36)	.8 (.75,.92)	.98 (.97,.99)	1 (1,1)
	0.15	.4 (.32,.44)	.9 (.8,.93)	.99 (0.98,1)	1 (1,1)
	0.2	.46 (.43,.52)	.96 (.93,.98)	0.99 (0.99,1)	1 (1,1)
	0.25	.49 (.45,.56)	.98 (.94,.99)	1 (1,1)	1 (1,1)
	0.3	.55 (.5,.62)	.99 (.98,.99)	1 (1,1)	1 (1,1)
	0.35	.56 (.52,.63)	.99 (.98,.99)	1 (1,1)	1 (1,1)
	0.4	.58 (.53,.67)	.99 (.99,.99)	1 (1,1)	1 (1,1)
	0.45	.59 (.54,.67)	.99 (.99,.99)	1 (1,1)	1 (1,1)
	0.5	.59 (.44,.68)	.99 (.99,.99)	1 (1,1)	1 (1,1)

Table 27. Statistical power at an OR = 4 for x_1 averaged across all levels of x_2, and the range of statistical power errors across all levels of the skewed probability of x_2.

		Sample Size			
		100	400	900	5000
x_1 Probability	0.01	.02 (.01,.04)	.2 (.17,.23)	.5 (.45,0.99)	.98 (.97,1)
	0.02	.07 (.05,.09)	.39 (.38,.41)	.74 (.72,.77)	.99 (0.99,1)
	0.03	.13 (.11,.16)	.55 (.53,.57)	.87 (.85,.9)	1 (1,1)
	0.04	.19 (.17,.22)	.65 (.63,.68)	.95 (.94,.97)	1 (1,1)
	0.05	.26 (.24,.27)	.74 (.7,.78)	.97 (.97,.9)	1 (1,1)
	0.06	.31 (.29,.33)	.8 (.77,.83)	.98 (.97,.99)	1 (1,1)
	0.07	.36 (.34,.37)	.86 (.83,.89)	.99 (.99,1)	1 (1,1)
	0.08	.39 (.38,.41)	.89 (.86,.91)	.99 (.99,1)	1 (1,1)
	0.09	.44 (.42,.46)	.92 (.91,.94)	.99 (.99,1)	1 (1,1)
	0.1	.47 (.45,.5)	.94 (.93,.96)	.97 (.98,1)	1 (1,1)
	0.15	.59 (.57,.63)	.99 (.98,.99)	1 (1,1)	1 (1,1)
	0.2	.67 (.63,.72)	.99 (.99,1)	1 (1,1)	1 (1,1)
	0.25	.72 (.6,.78)	.99 (.99,1)	1 (1,1)	1 (1,1)
	0.3	.77 (.72,.83)	.99 (0.99,1)	1 (1,1)	1 (1,1)
	0.35	.79 (.73,.85)	1 (1,1)	1 (1,1)	1 (1,1)
	0.4	.81 (.76,.87)	1 (1,1)	1 (1,1)	1 (1,1)
	0.45	.82 (.77,.88)	1 (1,1)	1 (1,1)	1 (1,1)
	0.5	.82 (.77,.88)	1 (1,1)	1 (1,1)	1 (1,1)

Discussion and Conclusion

It is not uncommon for researchers to encounter data from skewed populations. In these cases, the skewness of the sample and predictor variables reflects the true character of the

population rather than a sampling bias. Hence, the skewness in the predictor(s) may influence estimation if separation occurs or decrease the reliability of parameter estimates. Detecting separation through data configurations, infinite parameter estimates, and the non-convergence of the MLE is straightforward. However, with a skewed predictor, these clear indicators are not present. This leaves us with the question of the impact of skewed predictors on the eventual statistical results of a logistic regression. To answer this general question, I conducted five inter-related simulation studies, which, to my knowledge, are the first of their kind to be done for skewed dichotomous predictors. This study offers advice to practicing researchers on how dramatically skewness will affect their conclusions given their sample size. Although they were deliberately simulated in idealized situations, these five related studies offer guidelines that may apply in a practical context.

To begin with, the relationship between skewed cell probability and separation is noteworthy. I found that when the simulated data were highly skewed, this likely is characterized by complete separation, and hence the MLE did not converge. Webb, Wilson, and Chong (2004) showed that when the predictor variable is binary, some statistical software may still produce results in the presence of quasi-complete separation. Like other programs examined by Webb et al. (2004), R, which was used in this study, provides results with quasi-complete separation for binary predictors. They also demonstrated that the MLE does not converge in the presence of quasi-complete separation when the predictor is continuous. The implications for the findings are that the investigated parameter space in Study 2 does not likely include any setting with quasi-complete separation. In this dissertation, the non-convergence of the estimation is most likely due to complete separation, and perhaps in some cases to quasi-complete separation, although the latter was not detected.

The first model that was simulated examined simple logistic regression. The MLE did not converge for all conditions with high levels of skewed probability and small sample sizes. By examining the simulated populations of the highly skewed cell probabilities of a predictor, I saw that separation was a product of this attribute. Albert and Anderson (1984) proved that with a separated data set, a finite MLE does not exist. This was also the case in this study, in replications where the data were separated and the MLE did not converge. However, it was showen that in replications where separation did not occur and the MLE converged for the same set of conditions, parameter estimates were biased upwards and the Type I error rate was reduced, often approaching zero. As sample sizes increased and skewed probabilities decreased, the chances of separation occurring in the data quickly receded (as evidenced by the convergence of all replications), and empirical parameter estimates became close or equal to the population values. Type I error rates, surprisingly, were consistently deflated to close or equal to zero, but reached a nominal value as sample sizes increased and skewed probabilities decreased. As a result, the Type I error rate met only the liberal, not the stringent, criterion. It was also consistently deflated as a direct result of the enormous upward distortion of the standard errors. Standard errors were upwardly biased as the sample size decreased and the skewed probability increased. This inflation in standard errors was likely due to the undetected quasi-complete separation.

By examining the distributions of the estimated slope, it was observed that with a small sample and a large degree of skewed probability on the predictor, the slope distribution is far from what was expected, which is leptokurtic around zero. The distributions are closer to what was simulated as the skewed probability decreases. As a result, the ORs can become tremendously distorted.

Sample size plays a crucial role in the estimation. Many researchers may opt to increase their sample when their predictors are skewed. However, while this can solve the problem, the size required depends on the degree of skewness in the predictor. A sample size of 400 may not be large enough when the predictor has a skewed probability of 0.01, but suffices when the skewness is 0.07.

Although the Type I error rate was small (lower than nominal), indicating an acceptable Wald test, the accompanying power analysis completes the picture. With a very large sample and a potentially large effect size, researchers should be confident in the Wald test. On the other hand, statistical power is reduced, as anticipated from the Type I error results. The power for sample sizes from 50 to 400 usually suffers from a skewed probability in a predictor.

The second model used simple logistic regression, but with a skewed continuous predictor. We learned that the findings from the first model can be extended to this case. The only difference was that separation was not an issue, which means that the MLE converged for all replications. The Type I error rate reached a nominal value much more quickly and the violation of the lower end of the liberal criterion was 9% less than in Model 1.

The third model, similar to a randomized (clinical) trial experiment with unequal cell sample sizes, examined the impact of two skewed predictors on the analysis. The simultaneous skewness of both predictors had no effect on the Type I error rates. The skewed probability of one of the predictors had a negative impact on its Wald test, but not on the Wald test of the other variable. In terms of non-convergence, Type I error, and power from Model 1 are generalized in this case.

These findings are similar to research that examines the number of events per variable on a logistic regression and the Wald statistic. Peduzzi et al. (1996) and Vittinghoff and McCulloch

(2006) showed that as the events per variable decreased, Type I error and statistical power were deflated. Skewed probabilities may potentially decrease the number of events per variable, yielding similar findings. In addition, this dissertation supports the claim that the skewed probabilities of a predictor affect only the Wald test for that same predictor, with the assumption of independence among predictors being demonstrated here and in Barreto et al. (2014) and Zorn (2005).

The present study contributes to the literature by providing a broad picture of the effects of skewed cell probabilities in dichotomous predictors on the logistic regression model. More precisely, it was thoroughly describe how a categorical predictor's statistical characteristics affect estimation, parameter estimates, and the Wald test. It is important to note that in many cases, the estimator came to a convergence and results were produced, but there is no warning that a potential problem may exist. Data analysts can carry on without being aware that the standard errors are greatly inflated, resulting in low to no statistical power and (at times) greatly enlarged ORs. This study adds to the body of research on the characteristics of the Wald test in logistic regression by taking a step back and shedding light on some instances when it is biased by skewed predictor probabilities.

This study was conducted with methodologists and researchers—who read others' work and analyze their own data—in mind. It highlights a few important points, the first being that skewed probabilities can induce separation, which automatically affects estimation and results in non-convergence (Albert & Anderson, 1984). Secondly, when separation does not occur—even in severe cases of skewed probability—ML converges and estimates are produced. Thirdly, MLEs are biased upwards in severe conditions of small samples and highly skewed probability. Lastly, when skewness is less severe, with a range of [0.25, 0.5], or the sample size is

sufficiently large, Type I error rates reach a nominal value and power is high. Overall, these findings demonstrate why it is important to consider the descriptive characteristics of the predictor(s) before conducting a logistic regression analysis. Researchers may encounter situations wherein the Type I error rate of their hypothesis test is highly deflated, ostensibly declaring a strong test when this may not be the case. Also, the power of the hypothesis test performs in a complementary manner to the Type I error rates. That is, the power is deflated when the Type I error is, and reaches full power when the rate achieves a nominal value.

Several studies have investigated the effects of separation and found ways to overcome them through various estimation strategies. Similarly, a few have explored how separation can alter the Wald statistic, and their findings are echoed in this dissertation. However, more research is needed on the effects of predictor characteristics on real-life data and on more complicated logistic regression models. The association among predictors, path models, and generalized linear mixed models also requires further study.

Chapter 4: Conclusion and Future Research Directions

Scientific and technological advances have made calculations of intricate models with complicated data fairly easy to accomplish. Because statistical modelling is imperative in many research contexts, methodology courses are taught at all major universities. With the constant use of these techniques, statistical researchers are increasingly identifying and solving problems—finding ways of alleviating compromising data patterns, constructing or enhancing estimation procedures, and recommending alternative modelling strategies. This dissertation follows the same line of thought by conducting a study that examines the impact of a problematic data structure on the logistic regression model. This concluding chapter begins with a brief re-statement of the problem and methodology, followed by a review of the results, recommendations, and future research.

Problem Re-statement and Methodology

The **s**kewness of a categorical predictor in a logistic regression model is a notion that has not been discussed in recent literature. The description of skewness as a statistical characteristic of a categorical variable is also rarely used. But what happens if a researcher finds an imbalance in the sample size in the categories of a categorical predictor? That is, he or she may discover that the probability of obtaining each category is skewed. Imagine an ordinal variable of three categories. For this variable to be balanced or "not skewed," the probability of obtaining each category should be equal to one-third. This same variable is said to have *skewed cell probability* when the probability of each category ranges from a low to a high chance of occurrence or vice versa, which would change the direction of skewness. For example, an ordered skewed cell probability of category 1 might have a probability of occurring equal to 0.02, category 2 might have one equal to 0.25, and category 3 might have one equal to 0.73. The idea of skewness in

categorical variables is extended to dichotomous variables. The skewness of variables is due to either biased sampling or to a population artifact, the latter of which case is the focus of this research. The question then becomes, what happens when such variables are included in models? In particular, I wonder what happens when a dichotomous variable with skewed cell probabilities is a predictor in a logistic regression model. In this context, the skewed cell probabilities may, for example, reflect a natural (population based) imbalance in the sample sizes of two groups in an experiment – in this case, the dichotomous predictor variable represents a design vector of group differences.

In the literature reviews, I found no detailed information on the impact of skewed predictors (either continuous or categorical) in logistic regression. Mathematical and statistical literature do not document their effect on the convergence of estimators, or on the Type I and II error rates and statistical power of hypothesis tests. However, from personal experience, I have learned that estimating a logistic regression model with two or more dichotomous predictors with skewed cell probabilities yields a very high p-value, a large standard error estimate, and an astronomical effect size. In some cases, the model did not converge. The research question at the heart of this dissertation is: If the skewness in the probabilities of a predictor is not severe enough to disrupt the maximum likelihood estimation (MLE) in terms of convergence, to what extent could a researcher trust the test results?

To directly answer this question, five inter-related simulation studies were conducted. The simulations were based on constructing outcome and predictor(s) variables with varying degrees of skewness, sample size, and predictor variable type (i.e., dichotomous or continuous). In all cases, the models were correctly specified. Throughout, the Type I error rates and statistical power of the Wald tests for the predictor(s) were recorded. As is common practice, we

set the nominal Type I error rate (α) at 0.05.

The five simulation studies were organized around three logistic regression models. The first model examined simple logistic regression with a dichotomous predictor having skewed probabilities. The first study focused on the quality of the parameter coefficient estimates, including the convergence rates of the MLEs, as well as on Type I error. The second analyzed statistical power. To check the generalizability of skewness, the third study examined the MLE convergence and Type I error rate of a simple logistic regression with a continuous predictor. The final model included two simulations exploring a multiple logistic regression with the skewed cell probabilities of two dichotomous predictors. One looked at the convergence rates of the MLEs and Type I error rates for this model, and the other focused on its statistical power.

Review of Results and Discussion

The skewness of categorical predictors is seldom examined in the literature. What is interesting is the connection it has with a different but closely related concept, separation. Separation is inherently linked with logistic regression and pertains to a problematic data configuration that results in the non-convergence of the MLE and/or biased parameter estimates. It is caused by a linear combination of continuous or dichotomous predictors that perfectly separates events from non-events. There are two types of separation: 1) complete separation, where one or more of a model's predictors perfectly explains the outcome variable; and 2) quasi-complete separation, where only one covariate pattern has a zero count. It has been proven that with these two data configurations, the MLE does not converge or a finite estimate does not exist. A data configuration that is defined as one where estimation converges and estimates are acceptable is called overlap. In this dissertation, I found that when the simulated data were highly skewed, there is often complete separation and the MLE did not converge. In other cases,

where the MLE did converge, the simulation may have been characterized by overlap or quasi-complete separation; however there is not enough information to be certain.

The MLE for the simple logistic regression model did not converge for all conditions with high levels of skewed probability and small sample sizes. By examining the simulated populations of the highly skewed cell probabilities of a predictor, I saw that this factor in the predictor generated separation. In simulation replications where the data were separated, an MLE did not converge. However, in replications where separation did not occur and the MLE converged for the same set of conditions, parameter estimates were biased upwards and the Type I error rate was reduced, often approaching zero. As sample sizes increased and skewed probabilities decreased, the possibility of separation quickly receded (as evidenced by the convergence of all replications), and empirical parameter estimates became close or equal to the population values. Type I error rates, surprisingly, were consistently close or equal to zero, but reached the nominal value as the sample size increased and the skewed probability decreased. The Type I error rate was consistently deflated as a direct result of the enormous upward distortion of the standard errors. With increases in the sample size and declines in the skewed probability, standard errors became upwardly biased. This inflation was likely due to the undetected quasi-complete separation. On the other hand, statistical power is also reduced, as anticipated from the Type I error results. The power for sample sizes of 50 to 400 mostly suffers from a skewed cell probability in a predictor. The findings from the first model can be extended to the second and third, though with a few differences. Separation was not present when the variable was continuous. Also, the simultaneous skewness of both predictors in the third model had no impact on the Type I error rates.

It must be noted that the experiment conducted for this dissertation skewed cell

probabilities (or skewness) of the predictor was studied in isolation to better understand this phenomenon. In practice, however, both predictor and outcome variable may be slightly skewed. This may compound the issue of skewed probabilities to include both outcome and predictor. In an extended but related study on the impact of group imbalance on the application of logistic regression in differential item functioning with assessment data , Alkhalaf and Zumbo (2016) found that the results are a complex interaction of sample size, group imbalance (predictor skewness) , and difficulty of the item being modeled (outcome skewness). Preliminary results showed that in most cases, the Type I error rate is well below the nominal 0.05.

This study adds to the body of work that examines how variable characteristics interact with logistic regression. This dissertation shows that skewness, a unique characteristic of a single variable, induces bias in the Wald statistic and hence the hypothesis of the same variable. Not only that, but I have also demonstrated that separation may be induced by skewness. This mirrors the findings of a few other studies that have investigated the effects of separation on the Wald test, and the number of events per parameter in a logistic regression, which their findings are echoed in this dissertation. However, more research is needed in this area, investigating the impact of predictor characteristics on real life data and on more complicated models.

Recommendations

The results of this dissertation confirm the importance of conducting a thorough diagnostic analysis of the variables in a logistic regression. Indeed, reporting the descriptive characteristics of these variables should be common practice. In the future, articulated guidelines should be developed to ease the application of diagnostic methods and broaden the availability of this knowledge.

The implications of this study for practitioners and readers can be condensed into two points. Firstly, the impact of the skewness of a predictor(s) is independent of variable type. Continuous and dichotomous variables may result in deflated Type I error rates with high levels of skewness (3.2 or higher when the variable is continuous and 0.15 or lower when the variable is categorical), specifically when the sample size is less than or equal to 400. An interesting methodological twist occurs in that although a deflated Type I error rate results in a valid hypothesis test, the hypothesis test for the skewed variable and the interpretation of the resulting analyses should not be trusted because of the related increase in Type II errors and subsequent reduction in statistical power. In short, the statistical test is valid but not particularly useful to the researcher. Secondly, if separation occurs due to severe skewness, it is well established that an alternative estimator should be used. The best alternative estimator for skewed data is outside the scope of this dissertation.

Direction of Future Research

By now, it is obvious that this topic is vast and almost untouched. As a result, a future line of research is very large and may span a few years. The research plan is composed of two parts: 1) continuing along the lines of simulation studies; and 2) examining the implications of this problem using real-life data. This program is discussed in the next few paragraphs.

Simulation Studies. Simulations create an environment where factors are confined to a set of prespecified conditions. Researchers can identify and potentially solve statistical dilemmas through systematic simulations of problematic data structures. That being said, isolating all the factors that could potentially influence such analyses takes time and effort. In this dissertation, I examined a few factors—related to the degree of skewed cell probability (or skewness), sample size, and effect size—that impinge on three models. Our results point to the need for a further

series of simulation studies that deal with problems associated with innumerable other factors that include, but are not limited to, model type, variable type, model misspecification, and estimation method.

1) *Simple vs. mixed effects models and the quadratic model.* The first obvious extension of the work done in this dissertation is moving away from the simple fixed-effects model to the quadratic and mixed-effects models. How does the variation in the skewness of a predictor affect these complicated logistic regressions for different sample sizes? Can the findings from this dissertation be generalized to the mixed-effects and quadratic cases?
2) *Ordered, discrete, and polytomous variables.* In this dissertation, the focus is on the skewness of a dichotomous predictor, ignoring other types of categorical variables that are frequently seen in daily research. Firstly, ordered, nominal, and discrete categorical variables have distinctive features in terms of skewness in cell probabilities. Discrete variables do not assume an order of categories, indicating that the definition of skewness that is adopted in this dissertation may not apply to them. Is the ordering of a categorical variable imperative to understanding skewness in cell probabilities? Secondly, categorical predictors with more than two categories are modelled by first creating a design matrix, where each category is assigned a variable that contains two values, one indicating affiliation to this category and another indicating no affiliation. Then each category is entered into the logistic model, while leaving out a reference category. Will design matrix variables with extremely high or low counts be skewed and hence influence estimation, Type I and II errors, and statistical power similarly to dichotomous variables?

3) *Applications of the logistic regression model.* Our efforts, thus far, were dedicated to examining the impact of skewed predictors on a simulated version of the widely used randomized clinical trial. Logistic regression is also applied in psychometric and item response theory when assessing differential item functioning (DIF) in groups (e.g., Zumbo, 2007) In many cases, groups are unequal in size and are perhaps sampled from highly skewed populations. How will a skewed group membership variable alter the DIF analysis? Moreover, what happens when an item is too easy or hard and group membership is skewed? In this case, both the predictor and outcome variables are skewed. How will the skewness of cell probabilities in both variables affect the DIF analysis?
4) *Misspecification.* In simulations for this dissertation, the logistic regression models fitted to the data were correctly specified. Misspecification refers to fitting the simulated data to an incomplete or different model. In everyday research, an ideal model where the model being fit is the same as the generating model in the population. Therefore, fitting a misspecified model is a realistic way of examining the impact of the skewness of a predictor on logistic regression.
5) *Maximum likelihood vs. alternative methods of estimation.* Maximum likelihood was used as the estimation procedure in this study to gain insight on convergence and the quality of estimates. It is widely criticized for its unadaptability to violations of its assumptions. The literature offers a vast array of alternative estimators that are accurate and can tolerate problematic data structures such as separation. Comparing the performance of alternative estimators to the maximum likelihood in cases of skewed predictors or skewed cell probabilities in a logistic regression is integral to

providing sound recommendations for best practices, specifically since the skewness I am interested in is inherent in the variable and not a sampling artifact.

6) *Impact of skewness on classification tables.* Classification tables of model-predicted versus observed data in logistic regression are frequently used in educational and psychological research for both model adequacy and in some cases for model fit. How will skewness and quasi-complete separation impact the results of classification accuracy? Will the choice of cutoff be biased?
7) *Impact of skewness on the Likelihood Ratio Test.* In this dissertation we focus on the Wald test. It has been criticized frequently under ideal and problematic circumstances. A few authors (e.g. Agresti, 2002; Hauck, Jr, Walter & Donner, 1977) have recommended the use of the likelihood ratio test. Although tedious, it sometimes proves powerful and more efficient than the Wald test. However, what has not been examined is whether skewness of the predictor variable may have an impact on the likelihood ratio test. Is the likelihood ratio test immune to the skewed distributional characteristic of the predictor variables, or will it be effected similar to the Wald test?
8) *Examining marginal probabilities from a cross-classification perspective of categorical data.* Although the results of this research are clearly situated within a generalized linear modeling and regression framework., there is some value in looking at this problem from a cross-classification perspective. That is, the essential difference would be between the sparseness of covariate data and the skewness of that data. A skewed marginal probability –as defined in this dissertation- may lead to sparse tables. At the same time, sparse tables are not necessarily indicative of skewed marginal probabilities. The cross-classification and regression approaches reflect two

schools of thought wherein discussion would be around 'sparseness' in the multi-way table resulting from the cross-classification and not 'separation'. The focus on regression and separation, here, brought a new lens in the social science and educational research literature. Although the issue of sparse tables is historical and rooted in categorical data analysis and theory, new insights may be gained by adopting the cross-classification framework in examining the marginal probabilities of the categorical variables.

9)

Real-Life Data Application. Simulations help us understand the phenomena under investigation directly and in a sterile context. However, actual situations are usually more complicated and sometimes involve multiple dilemmas. Therefore, examining the skewness of a predictor in a logistic model (fixed effects, mixed effects, path model, etc.) alongside other contextual real-life factors will provide a realistic understanding of this problem and how it presents itself in everyday research situations. This parallels debates in social and behavioural science research methodology about internal and external validity, and want which should take precedent.

References

Agresti, A. (2002). *Categorical Data Analysis* (3rd Editio). Hoboken, New Jersey: John Wiley & Sons, Inc. http://doi.org/10.1007/s007690000247

Ahmad, S., Ramli, N. M., & Midi, H. (2010). Robust Estimators in Logistic Regression : A Comparative Simulation Study. *Journal of Modern Applied Statistical Metods*, *9*(2).

Albert, A., & Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, *71*(1), 1–10.

Alkhalaf, A. A. (2014). *Application of Logistic Regression in Higher Education Research: A sample from 2000-2013*.

Alkhalaf, A. A., & Zumbo, B. D. (2016). The Impact of Group Imbalance on Logistic Regression Analyses with Assessment Data. In *10th International Test Comission Conference*.

Allison, P. D. (2014). Measures of Fit for Logistic Regression. *SAS Global Forum 2014*, (1970), 1–12.

Anderson, J. A., & Richardson, S. C. (1979). Logistic Discrimination and Bias Correction in Maximum Likelihood Estimation. *American Society of Quality*, *21*(1), 71–78.

Barreto, I. D. de C., Russo, S. L., Brasil, G. H., & Simon, V. H. (2014). Seperation Phenomena Logistic Regression. *Revista GEINTEC*, *4*(1), 716–728. Retrieved from http://srmo.sagepub.com/view/logistic-regression/SAGE.xml

Bianco, A. M., & Martínez, E. (2009). Robust testing in the logistic regression model. *Computational Statistics and Data Analysis*, *53*(12), 4095–4105. http://doi.org/10.1016/j.csda.2009.04.015

Birnbaum, A. (1964). Median-Unbiased Estimators. *Bulletin of Mathematical Statistics (Tokyo)*,

11, 25–34.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144–152.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (3rd ed.). Routledge.

Conover, A. W. J., Johnson, M. E., & Johnson, M. M. (1981). A Comparative Study of Tests for Homogeneity of Variances , with Applications to the Outer Continental Shelf Bidding Data. *Technometrics*, *23*(4), 351–361.

Cordeiro, G. M., & McCullagh, P. (1991). Bias Correction in Generalized Linear Models. *Journal of the Royal Statistical Society. Series B (Methodological)*, *53*(3), 629–643. http://doi.org/10.1038/203024b0

Cox, D. R., & Snell, D. J. (1989). *Analysis of binary data*. London; New Yprk: Chapman & Hall.

Day, N. E., & Kerridge, D. F. (1967). A General Likelihood Discriminant. *Biometrics*, *23*(2), 313–323.

Fears, T. R., Benichou, J., & Gail, M. H. (1996). A Reminder of the Fallibility of the Wald Statistic. *The American Statistician*, *50*(3), 226–227. http://doi.org/10.1080/00031305.1996.10474384

Ferguson, C. J. (2009). An Effect Size Primer : A Guide for Clinicians and Researchers. *Professional Psychology: Research and Practice*, *40*(5), 532–538. http://doi.org/10.1037/a0015808

Ferris, C. D., Grubbs, F. E., & Weaver, C. L. (1946). Operating Characteristics for the Common Statistical Tests of Significance. *The Annals of Mathematical Statistics*, *17*(2), 178–197.

Gordóvil-merino, A., Guàrdia-olmos, J., & Peró-cebollero, M. (2012). Estimation of logistic

regression models in small samples . A simulation study using a weakly informative default prior distribution. *Psicologica*, *33*, 345–361.

Gregory, W., & Veall, R. (1986). Wald Tests of Common Factor Restrictions. *Economica Letters*, *22*, 203–208.

Hauck, Jr, Walter, W., & Donner, A. (1977). Wald's Test as Applied to Hypotheses in Logit Analysis. *Journal of the American Statistical Association*, *72*(360), 851–853. Retrieved from file:///C:/Documents and Settings/L?ny/Local Settings/Application Data/Mendeley Ltd./Mendeley Desktop/Downloaded/Donner - 2010 - Applied in to Logit Hypotheses.pdf

Heinze, G., & Puhr, R. (2010). Bias-reduced and separation-proof conditional logistic regression with small or sparse data sets. *Statistics in Medicine*, (29), 770–777. http://doi.org/10.1002/sim.3794

Hirji, K. F., Tsiatis, A. A., & Mehta, C. R. (1989). Median Unbiased Estimation for Binary Data. *The American Statistician*, *43*(1), 7–11.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). John Wiley & Sons, Inc.

Jennings, D. E. (1986). Judging inference adequacy in logistic regression. *Journal of the American Statistical Association*, *81*(394), 471–476. http://doi.org/10.2307/2289237

Larntz, K. (1978). Small-Sample Comparisons of Exact Levels for Chi-Squared Goodness-of-fit Statistics. *Journal of the American Statistical Association*, *73*(362), 253–263.

Lemeshow, S., & Hosmer, D. W. (1982). A Review of Goodness of Fit Statistics for Use in the Development of Logistic Regression Models. *American Journal of Epidemiology*, *115*(1), 92–106.

Lütkepohl, H., & Burda, M. M. (1997). Modified wald tests under nonregular conditions.

Journal of Econometrics, *78*, 315–332. http://doi.org/10.1016/S0304-4076(97)80015-2

McFadden, D. (1974). Conditional logit Analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in econometrics* (pp. 105–142). Academic Press.

Mcnutt, L., Wu, C., Xue, X., & Hafner, J. P. (2003). Estimating the Relative Risk in Cohort Studies and Clinical Trials of Common Outcomes. *American Journal of Epidemiology*, *157*(10), 940–943. http://doi.org/10.1093/aje/kwg074

Menard, S. (2000). Coefficients of Determination for Multiple Logistic Regression Analysis. *The American Statistician*, *54*(1), 17–24. http://doi.org/10.1080/00031305.2000.10474502

Mîndrilã, D. (2010). Maximum likelihood (ML) and diagonally wighted least squares (DWLS) estimation procedures: A comparison of estimation bias with ordinal and multivariate non-normal data. *International Journal of Digital Society*, *1*(1), 60–66.

Mittlbock, M., & Schemper, M. (1996). Explained Variation for Logistic Regression. *Statistics in Medicine*, *15*(19), 1987–1997. http://doi.org/10.1002/(SICI)1097-0258(19961015)15:19<1987::AID-SIM318>3.0.CO;2-9

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*(3), 691–692. http://doi.org/10.1093/biomet/78.3.691

Oakley Davies, H. T., Crombie, lain K., & Tavakoli, M. (1998). When can odds ratio mislead? *British Medical Journal*, *316*(7136).

Pawitan, Y. (2000). A Reminder of the Fallibility of the Wald Statistic : Likelihood Explanation. *The American Statistician*, *54*(1), 54–56.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A Simulation Study of the Number of Events per Variable in Logistic Regression Analysis. *Journal of Clinical Epidemiology*, *49*(12), 1373–1379.

Peng, C. J., So, T. H., Stage, F. K., & John, E. P. S. (2002). The Use and Interpretation of Logistic Regression in Higher Education Journals: 1988-1999. *Research in Higher Education*, *43*(3), 259–293.

Pergibon, D. (1981). Logistic Regression Diagnostics. *The Annals of Statistics*, *9*(4), 705–724.

Rousseeuw, P. J., & Christmann, A. (2003). Robustness against separation and outliers in logistic regression. *Computational Statistics and Data Analysis*, *43*, 315–332.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using Multivariate Statistics* (6th ed.). Pearson Education.

Vaeth, M. (1985). On the Use of Wald's Test in Exponential Families. *International Statistical Review*, *53*(2), 199–214.

Vittinghoff, E., & McCulloch, C. E. (2006). Relaxing the rule of ten events per variable in logistic and cox regression. *American Journal of Epidemiology*, *165*(6), 710–718. http://doi.org/10.1093/aje/kwk052

Webb, M. C., Wilson, J. R., & Chong, J. (2004). An Analysis of Quasi-complete Binary Data with Logistic Models : Applications to Alcohol Abuse Data. *Journal of Data Science*, *2*, 273–285.

Yuan, K.-H., Bentler, P. M., & Zhang, W. (2005). The Effect of Skewness and Kurtosis on Mean and Covariance Structure Analysis: The Univariate Case and Its Multivariate Implication. *Sociological Methods & Research*, *34*(2), 240–258. http://doi.org/10.1177/0049124105280200

Zorn, C. (2005). A Solution to Separation in Binary Response Models. *Political Analysis*, *13*(2), 157–170. http://doi.org/10.1093/pan/mpi009

Zumbo, B. D. (2007). Three Generations of DIF Analyses : Considering Where It Has Been ,

Where It Is Now, and Where It Is Going. *Language Assessment Quarterly*, *4*(2), 223–233.

Appendix A

List of Reviewed Articles

1. Allen, J., Robbins, S., Casillas, A., & Oh, I. (2008). Third-year college retention and transfer: Effects of academic performance, motivation, and social connectedness. *Research in Higher Education, 49*, 647-664. doi:10.1007/s11162-008-9098-3
2. Anderson, G., Sun, J. C., & Alfonso, M. (2006). Effectiveness of statewide articulation agreements on the probability of transfer: A preliminary policy analysis. *Review of Higher Education, 29*(3), 261-291. doi:10.1353/rhe.2006.0001
3. Bahr, P. (2008). Does mathematics remediation work? A comparative analysis of academic attainment among community college students. *Research in Higher Education, 49*, 420-450. doi:10.1007/s11162-008-9089-4
4. Bahr, P. (2010a). The Bird's eye view of community colleges: A behavioral typology of first-time students based on cluster analytic classification. *Research in Higher Education, 51*, 724-749. doi:10.1007/s11162-010-9180-5
5. Bahr, P. (2010b). Revisiting the efficacy of postsecondary remediation: The moderating effects of Depth/Breadth of deficiency. *Review of Higher Education, 33*(2), 177-205. doi:10.1353/rhe.0.0128
6. Bahr, P. (2012). Deconstructing remediation in community colleges: Exploring associations between course-taking patterns, course outcomes, and attrition from the remedial math and remedial writing sequences. *Research in Higher Education, 53*, 661–693. doi:10.1007/s11162-011-9243-2

7. Bailey, T., Calcagno, J. C., Jenkins, D., Leinbach, T., & Kienzl, G. (2006). Is student-right-to-know all you should know? An analysis of community college graduation rates. *Research in Higher Education, 47*(5), 491-519. doi:10.1007/s11162-005-9005-0
8. Belloc, F., Maruotti, A., & Petrella, L. (2010). University drop-out: An Italian experience. *Higher Education, 60*, 127–138. doi:10.1007/s10734-009-9290-1
9. Berggren, C. (2006). Labour market influence on recruitment to higher education: Gender and class perspectives. *Higher Education, 52*, 121-148. doi:10.1007/s10734-004-5793-y
10. Bieri, C., & Schuler, P. (2011). Cross-curricular competencies of student teachers: A selection model based on assessment centre admission tests and study success after the first year of teacher training. *Assessment & Evaluation in Higher Education, 36*(4), 399-415.
11. Bonilla, D., Buch, K. K., & Johnson, C. W. (2013). Effect of learning communities on student attitudes and corresponding behaviors: A mediated test of involvement theory. *International Journal of Higher Education, 2*(3), 107-114.
12. Callender, C., & Jackson, J. (2008). Does the fear of debt constrain choice of university and subject of study? *Studies in Higher Education, 33*(4), 405-429.
13. Cragg, K. M. (2009). Influencing the probability for graduation at four-year institutions: A multi-model analysis. *Research in Higher Education, 50*, 394–413. doi:10.1007/s11162-009-9122-2
14. Craney, C., McKay, T., Mazzeo, A., Morris, J., Prigodich, C., & de Groot, R. (2011). Cross-discipline perceptions of the undergraduate research experience. *Journal of Higher Education, 82*(1), 92-113.

15. Crisp, G., & Nora, A. (2010). Hispanic student success: Factors influencing the persistence and transfer decisions of Latino community college students enrolled in developmental education. *Research in Higher Education, 51*, 175–194. doi:10.1007/s11162-009-9151-x
16. DesJardins, S. (2001). Assessing the effects of changing institutional aid policy. *Research in Higher Education, 42*(6), 653-678.
17. DesJardins, S. (2002). An analytic strategy to assist institutional recruitment and marketing efforts. *Research in Higher Education, 43*(5), 531-553.
18. Eggens, L., Van der Werf, M. P., & Bosker, R. J. (2008). The influence of personal networks and social support on study attainment of students in university education. *Higher Education, 55*, 553–573. doi:10.1007/s10734-007-9074-4
19. Engberg, M. E. (2007). Educating the workforce for the 21st century: A cross-disciplinary Analysis of the impact of the undergraduate Experience on students' development of a pluralistic orientation. *Research in Higher Education, 48*(3), 283-317. doi:10.1007/s11162-006-9027-2
20. Flashmen, J. (2013). A cohort perspective on gender gaps in college attendance and completion. *Research in Higher Education, 54*, 545–570. doi:10.1007/s11162-013-9285-8
21. Girard, M. (2010). Match between pre- and postmigration education among new immigrants: Determinants and payoffs. *The Canadian Journal of Higher Education, 40*(3), 81-99.
22. Hovdhaugen, E. (2009). Transfer and dropout: Different forms of student departure in Norway. *Studies in Higher Education, 34*(1), 1-17.
23. Hu, S., & Hossler, D. (2000). Willingness to pay and preference for private institutions. *Research in Higher Education, 41*, 685-701.

24. Jansen, E. P. (2004). The influence of the curriculum organization on study progress in higher education. *Higher Education, 47*, 411-435.

25. Kim, D., Bankart, C., & Isdell, L. (2011). International doctorates: Trends analysis on their decision to stay in US. *Higher Education, 62*, 141-161. doi:10.1007/s10734-010-9371-1

26. Klein, M., & Weiss, F. (2011). Is forcing them worth the effort? Benefits of mandatory internships for graduates from diverse family backgrounds at labour market entry. *Studies in Higher Education, 36*(8), 969-987.

27. Klugman, J. (2012). How resource inequalities among high schools reproduce class advantages in college destinations. *Research in Higher Education, 53*, 803–830. doi:10.1007/s11162-012-9261-8

28. Konecˇny´, T., Basl, J., Myslivecˇek, J., & Simonova´, N. (2012). Alternative models of entrance exams and access to higher education: The case of the Czech Republic. *Higher Education, 63*, 219-235. doi:10.1007/s10734-011-9433-z

29. Marks, G. N. (2009). The social effects of the Australian higher education contribution scheme (HECS). *Higher Education, 57*, 71–84. doi:10.1007/s10734-008-9133-5

30. Martin, N. D., & Spenner, K. I. (2009). Capital conversion and accumulation: A social portrait of legacies at an elite university. *Research in Higher Education, 50*, 623–648. doi:10.1007/s11162-009-9136-9

31. Melguizo, T. (2008). Quality matters: Assessing the impact of attending more selective institutions on college completion rates of minorities. *Research in Higher Education, 49,* 214-236. doi:10.1007/s11162-007-9076-1

32. Newman, M. D., & Petrosko, J. M. (2011). Predictors of alumni association membership. *Research in Higher Education, 52*, 738–759. doi:10.1007/s11162-011-9213-8

33. Outcalt, C. L., & Skewes-Cox, T. E. (2002). Involvement, interaction, and satisfaction: The human environment at HBCUs. *Review of Higher Education, 25*(3), 331-347. doi:10.1353/rhe.2002.0015

34. Perna, L. (2005). Sex differences in faculty tenure and promotion: The contribution of family ties. *Research in Higher Education, 46*(3), 277-307. doi:10.1007/s11162-004-1641-2

35. Perna, L., & Titus, M. A. (2005). The relationship between parental involvement as social capital and college enrollment: An examination of Racial/Ethnic group differences. *Journal of Higher Education, 76*(5), 485-518.

36. Rettinger, D. A., & Kramer, Y. (2009). Situational and personal causes of student cheating. *Research in Higher Education, 50*, 293-313. doi:10.1007/s11162-008-9116-5

37. Riegle-Crumb, C. (2010). More girls go to college: Exploring the social and academic factors behind the female postsecondary advantage among Hispanic and white students. *Research in Higher Education, 51*, 573–593. doi:10.1007/s11162-010-9169-0

38. Roksa, J. (2010). Bachelor's degree completion across state contexts: Does the distribution of enrollments make a difference? *Research in Higher Education, 51*, 1-20. doi:10.1007/s11162-009-9146-7

39. Seelen, L. P. (2002). Is performance in English as a second language a relevant criterion for admission to an English medium university? *Higher Education, 44*, 213-232.

40. Shankland, R., Genolini, C., Riou França, L., Guelfi, J., & Ionescu, S. (2010). Student adjustment to higher education: The role of alternative educational pathways in coping with the demands of student life. *Higher Education, 59*, 353-366. doi:10.1007/s10734-009-9252-7

41. St. John, E. P., Musoba, G. D., Simmons, A., Chung, C., Schmit, J., & Peng, C. J. (2004). Meeting the access challenge: An examination of Indiana's twenty-first century scholars program. *Research in Higher Education, 45*(8), 829-871.

42. Stassen, M. L. (2003). Student outcomes: The impact of varying living-learning community models. *Research in Higher Education, 44*(5), 581-613.

43. Tein, F. (2000). To what degree does the desire for promotion motivate faculty to perform research? Testing the expectancy theory. *Research in Higher Education, 41*(6), 723-752.

44. Teixeira, A., & Rocha, M. F. (2010). Cheating by economics and business undergraduate students: An exploratory international assessment. *Higher Education, 59*, 663–701. doi:10.1007/s10734-009-9274-1

45. Toutkoushian, R. K., & Bellas, M. L. (2003). The effects of part-time employment and gender on faculty earnings and satisfaction. *Journal of Higher Education, 74*(2), 172-195.

46. Wells, R. S., Lynch, C. M., & Seifert, T. A. (2011). Methodological options and their implications: An example using secondary data to analyze Latino educational expectations. *Research in Higher Education, 52*, 693–716. doi:10.1007/s11162-011-9216-5

47. Wolniak, G. C., Mayhew, M. J., & Engberg, M. E. (2012). Learning's weak link to persistence. *Journal of Higher Education, 83*(6), 795-823.

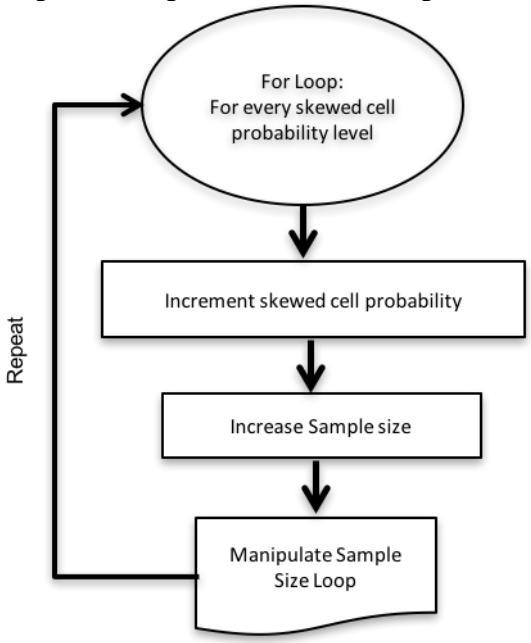
48. Xu, Y. J. (2013). Career outcomes of STEM and non-STEM college graduates: Persistence in majored-field and influential factors in career choices. *Research in Higher Education, 54*, 349-382. doi:10.1007/s11162-012-9275-2

49. Zimdars, A. K. (2007). Testing the spill-over hypothesis: Meritocracy in enrolment in postgraduate education. *Higher Education, 54*, 1-19. doi:10.1007/s10734-006-9043-3

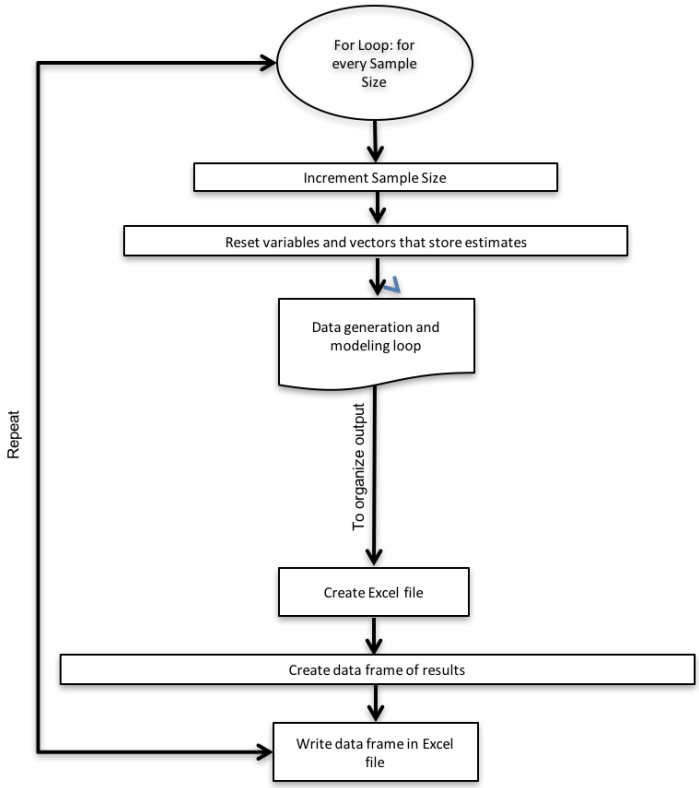
Appendix B

Simulation Flowcharts

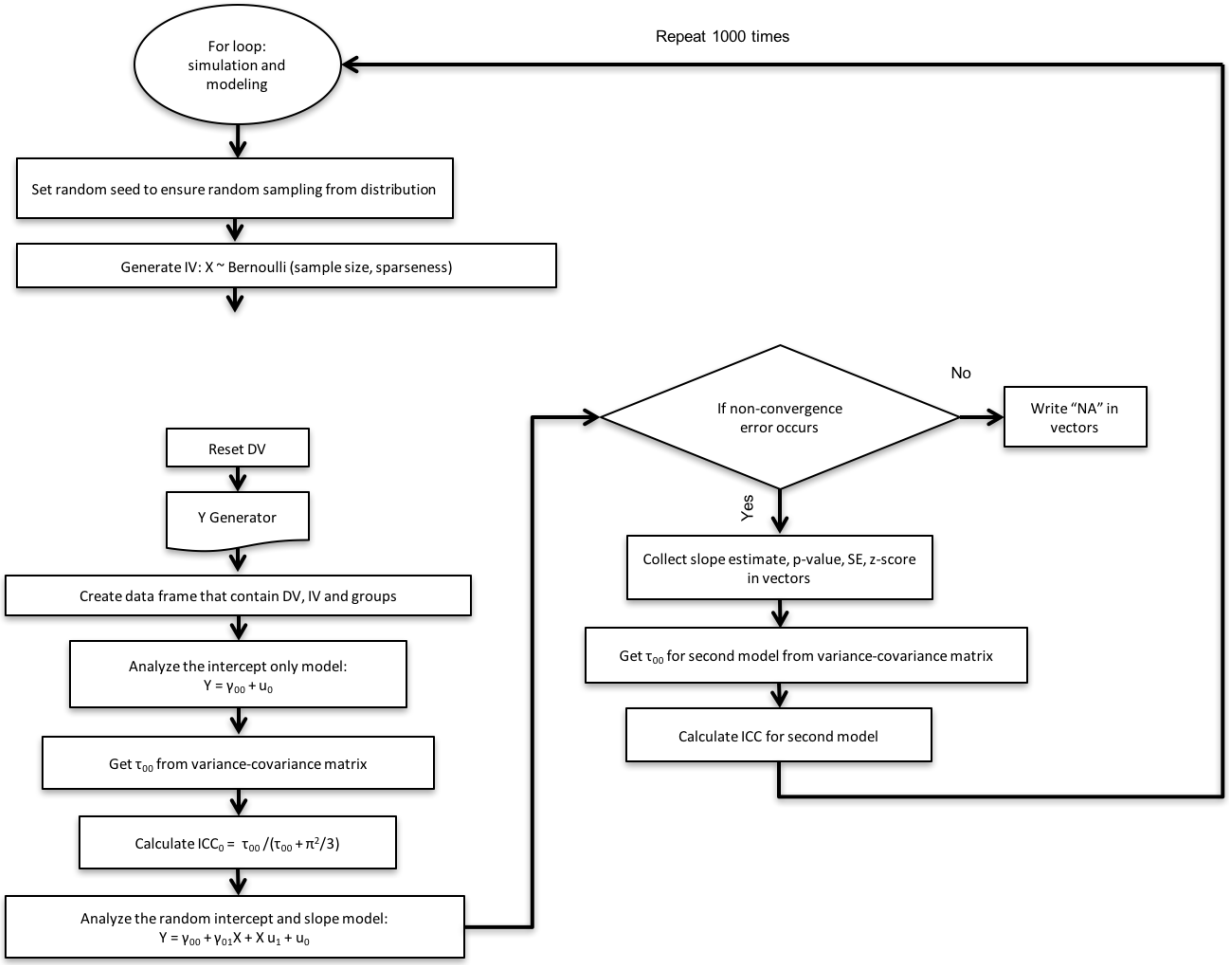
Step 1. Manipulate skewed cell probability.



Step 2. Manipulate sample size.



Step 3. Data Generation and Modelling Loop.



Step 4. Dependent Variable Generator.