Journal of Modern Applied Statistical Methods

Volume 16 | Issue 2 Article 4

December 2017

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

Arwa Alkhalaf
University of British Columbia, arwaalkhalaf@hotmail.com

Bruno D. Zumbo
University of British Columbia, bruno.zumbo@ubc.ca

Recommended Citation
Alkhalaf, A., & Zumbo, B. D. (2017). The Impact of Predictor Variable(s) with Skewed Cell Probabilities on Wald Tests in Binary Logistic Regression. Journal of Modern Applied Statistical Methods, 16(2), 40-80. doi: 10.22237/jmasm/1509494640

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

Cover Page Footnote

The first author would like to acknowledge the supporting and funding from King Abdulaziz University. This project is funded by the UBC-Paragon Research Agreement.

Journal of Modern Applied Statistical Methods
November 2017, Vol. 16, No. 2, 40-80.
doi: 10.22237/jmasm/1509494640

ISSN 1538 − 9472

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

Arwa Alkhalaf
University of British Columbia
Vancouver, BC, Canada

Bruno D. Zumbo
University of British Columbia
Vancouver, BC, Canada

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 regression models. A detailed description of the impact of skewed cell predictor probabilities and sample size provide guidelines for practitioners wherein to expect the greatest problems.

Keywords: logistic regression, skewed cell probability, simulation study, categorical predictor, skewed predictor

Introduction

Logistic regression modeling is growing in popularity in psychological and educational research (Cohen, Cohen, West, & Aiken, 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 paper 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

Dr. Alkhalaf is a graduate of the University of British Columbia. Email her at arwaalkhalaf@hotmail.com. Dr. Zumbo is the Paragon UBC Professor of Psychometrics and Measurement and an editor of this journal. Email him at bruno.zumbo@ubc.ca.

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. With the aim of filling this gap in the literature, consider 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. Note operating characteristics is used to here refer to the Type I and II error rates and statistical power of a hypothesis test (Ferris, Grubbs, & Weaver, 1946).

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, it is necessary to be able to visualize the data. In addition to the conventional data matrix (in which rows are participants and columns are variables), categorical data may be displayed as a multi-way table in which the cells are counts of occurrences of the corresponding row and column elements. The former display allows insight on the variety of covariate patterns, whereas the latter allows learning 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 data in this table. In his discussion of empty cells and sparse tables, Agresti (2002) described 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 in a logistic regression with a dichotomous predictor can be thought of as a two-way (row-by-column) table that has a similar format to Table 1. In Table 1, 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 1. 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 mathematically proved that although overlap yields a finite and unique solution, MLEs do not exist for the other two data patterns, although it was left to future researchers to develop 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, skewed probabilities is adopted to describe the row (or column) marginal distribution of the categorical predictor variables. This phrase has two uses in the statistical literature of interest. Larntz considered a case where the binary or multinomial predictor variables have an implicit order (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. Jennings (1986) did not use the phrase skewed probabilities, but instead described the marginal probabilities of the outcome variable as equal or unequal. The characterization Larntz described is more in line with the one adopted here, in good part because it corresponds more closely to how data analysts in education and psychology conceptualize such distributions.

Skewness here is used in lieu of sparse tables or separation for a few reasons. 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. Because the interest is in the skewness of the predictor variable, the methods in conducting this study reflect this concept. In severe cases of skewed probabilities, however, sparse tables or separation may occur. Nevertheless, similar to the example shown in Table 1, 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? The information in Table 2 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 2. 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, there are three others on related issues.

1) 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 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.

2) 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. It was found that the fit statistics generally performed well.

3) In a study 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. The recognition of potential skewness in the data set was interesting. They stated, “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 ±∞), these were eliminated, while only those data sets that were “acceptable” were included (p. 74).

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), we predict 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. Therefore, given the lack of an analytic solution, computer simulation experiments are needed to more fully explore the impact of predictor skewness.

There will be cases in this 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 2×2 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 our simulation experiment as a by-product of the data configuration.

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 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 a nonsignificant Wald statistic.

In a simulation study conducted by Peduzzi, Concato, Kemper, Holford, and Feinstein (1996), in which they examined the effects of the number of events per predictor variable in a logistic regression model, it was found the MLE did not converge with two and five events per predictor. 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. However, Barreto et al. (2014) found that the Wald test can detect which variables are individually significant, but fails to determine the significance of the variable that presents separation. The maximum likelihood estimates become inefficient, providing inflated variances.

The Wald test was 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, we focus on the Wald test in this study.

Simulation Studies

The results of five simulation studies are reported, 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

In this series of studies Monte Carlo simulations were used to examine the skewness of a predictor at the population level, meaning 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). This 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; (c) in the social sciences, a large gender imbalance of the participants in a study due to culturally sensitive issues; and (d) as is well known, binary predictors in models 1 and 3 can be interpreted as being a design matrix in an experiment or clinical trial – note that the imbalance in the experiment or clinical trial reflects population imbalance or what is sometimes called unequal cell sample sizes reflecting population characteristics and therefore not due to selection bias or attrition (e.g., Christensen, 2016).

To directly answer the research question of the effect of a skewed predictor on the eventual statistical conclusions of a logistic regression model, outcome and predictor(s) variables were simulated 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. The focus is on the Type I error rates and statistical power of the Wald test for the predictor(s). 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 for a binary logistic regression when the predictor variable(s) has a skewed cell probability?

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. The outcome variable of the regression model, throughout, is balanced or nearly balanced (i.e., not skewed). A secondary aim is to provide diagnostic information by documenting the situations where skewness may affect decisions and inferences.

Method

Simulation factors. For this simulation, two experimental factors were varied: sample size and skewness of the predictor variable. Sample sizes ranged across 13 levels from 10 to 5000. This large range represents a wide space of sample sizes starting from very small at 10 and 50. Then the range includes sample sizes that are seen more frequently in educational research of 100 to 1,000 in increments of 100. A sample size of 5,000 was included to verify the simulation experiment. The expected probability p of the predictor variable, described in more detail below, ranged from 0.01 to 0.45 across 17 levels. Expected probabilities are directly linked to the degree of skewness as can be seen later in Model 2. The degrees of expected probability range from extremely skewed to non-skewed distributions. These levels of skewed probability were chosen to reflect an array of distributional characteristics. Similar to sample size, the slight increments in skewness levels provide a wide range of distributional characteristics for variables. In addition, for comparison purposes, the case where the predictor variable is balanced was considered, 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. This large range of sample sizes and skewness levels is necessary to more fully document the impact of skewed cell probabilities.

Simulation procedure. The simulation and analyses were conducted using the R software package. There were 1,000 replications in each cell of the experimental design, 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. The results based on one thousand replications were compared to 10,000 replications and we found that both yielded the same results in terms of Type I errors, percentage of non-convergences, standard errors, and statistical power. Therefore, there was no marginal gain from the additional replications and we report the results based on one thousand replications herein. For each replication in each cell, the simulation algorithm consists of multiple loops that achieve different purposes. There are a few important steps in this process.

Step 1. The experiment is built upon data 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), 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.

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 1,000 replications. (To highlight the matter of non-convergence for day-to-day researchers, non-convergence rates based on the 1,000 replications were reported. The resulting Type I error rates are therefore based on the number of convergences and the simulation results are unbalanced (i.e., every empirical Type I error rate is not based on the same denominator). The reported results were compared against the findings wherein the number of replications within a cell continued until 1,000 convergences. The findings did not change, therefore the reported Type I error rates and statistical power results would not change. In the worst cases it took up to one million replications to achieve the 1,000 convergences, so the marginal computational gain was minimal.)

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 (Mood, Graybill, & Boes, 1974).

Analysis of Type I error. Type I error rate was calculated for each condition. Bradley's (1978) approach was used to compare the nominal and empirical Type I error rates for each condition. Bradley specifies two criteria of robustness, one stringent and one liberal. The stringent criterion is for a robust test, the empirical Type I error should fall within the range of $\alpha \pm 0.1\alpha$, whereas for the liberal criterion 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 was the non-convergence of some replications, as indicated in Table 3. Table 3 depicts the simulation experimental design, wherein each element is the number of non-convergences out of 1,000 replications. For example, for a sample size of 200 and $p = 0.02$, 21 of the 1,000 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 increased, 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.

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.

Table 3. Number of non-convergences from 1,000 replications for Model 1.

Probability	Sample Size												
	10	*50*	*100*	*200*	*300*	*400*	*500*	*600*	*700*	*800*	*900*	*1,000*	*5,000*
.01	893	610	388	148	50	21	2	0	0	0	0	0	0
.02	790	380	154	21	6	1	0	0	0	0	0	0	0
.03	710	230	58	7	1	0	0	0	0	0	0	0	0
.04	647	143	19	1	1	0	0	0	0	0	0	0	0
.05	575	83	7	1	1	0	0	0	0	0	0	0	0
.06	508	44	5	0	0	0	0	0	0	0	0	0	0
.07	481	29	2	0	0	0	0	0	0	0	0	0	0
.08	422	12	1	0	0	0	0	0	0	0	0	0	0
.09	392	7	0	0	0	0	0	0	0	0	0	0	0
.10	346	3	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
.20	114	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
.30	40	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
.40	10	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
.50	7	0	0	0	0	0	0	0	0	0	0	0	0

Type I error rate. Table 4 is structured in the same way as Table 3 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 5. Table 4 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. Similarly, to consider the shaded area a safe zone, 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 our 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 4, 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 4. Type I error rate for Model 1.

Probability	**Sample Size**												
	10	*50*	*100*	*200*	*300*	*400*	*500*	*600*	*700*	*800*	*900*	*1,000*	*5,000*
.01	0	0	0	0	0	0	0	0	0	.02	.02	.02	.05
.02	0	0	0	0	0	.01	.02	.03	.03	.04	.04	.04	.05
.03	0	0	0	0	.01	.03	.05	.04	.03	.04	.04	.05	.05
.04	0	0	0	.01	.03	.03	.05	.03	.03	.04	.05	.04	.04
.05	0	0	0	.02	.04	.04	.05	.04	.04	.03	.05	.05	.03
.06	0	0	0	.03	.03	.03	.05	.04	.04	.03	.05	.05	.04
.07	0	0	.01	.04	.04	.04	.05	.05	.04	.04	.05	.05	.05
.08	0	.01*	.01	.04	.04	.04	.05	.04	.05	.05	.06	.05	.06
.09	0	.01	.02	.04	.04	.04	.05	.04	.05	.05	.06	.04	.05
.10	0	.01	.02	.04	.04	.04	.04	.04	.04	.05	.05	.06	.05
.15	0	.01	.03	.06	.04	.03	.04	.04	.04	.06	.05	.05	.05
.20	0	.03	.04	.05	.06	.04	.04	.05	.04	.05	.05	.05	.05
.25	0	.05	.05	.06	.05	.05	.05	.05	.05	.06	.05	.06	.06
.30	0	.05	.04	.06	.05	.05	.06	.04	.04	.06	.05	.05	.05
.35	0	.05	.03	.06	.05	.06	.05	.04	.04	.05	.05	.05	.05
.40	0	.05	.04	.05	.06	.05	.05	.04	.04	.06	.04	.05	.06
.45	0	.05	.04	.05	.05	.05	.06	.05	.05	.05	.05	.06	.05
.50	0	.06	.05	.06	.06	.06	.05	.05	.05	.05	.06	.05	.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 5).

Table 5. 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. 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 6 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 7 (Birnbaum, 1964). This is referred to as median-unbiasedness.

Table 6. Average odds ratio, reflecting the widely used “mean unbiasedness.”

Probability	Sample Size												
	10	*50*	*100*	*200*	*300*	*400*	*500*	*600*	*700*	*800*	*900*	*1,000*	*5,000*
.01	≈∞	≈∞	≈∞	≈∞	≈∞	≈∞	≈∞	≈∞	≈∞	≈∞	≈∞	≈∞	1.05
.02	≈∞	≈∞	≈∞	≈∞	≈∞	≈∞	≈∞	≈∞	1.20	1.20	1.20	1.00	1.00
.03	≈∞	≈∞	≈∞	≈∞	≈∞	≈∞	≈∞	1.10	1.10	1.10	1.10	1.10	1.00
.04	≈∞	≈∞	≈∞	≈∞	≈∞	1.20	1.20	1.10	1.10	1.10	1.10	1.10	1.00
.05	≈∞	≈∞	≈∞	≈∞	1.20	1.10	1.10	1.10	1.10	1.00	1.10	1.00	1.00
.06	≈∞	≈∞	≈∞	≈∞	1.20	1.10	1.10	1.10	1.10	1.00	1.10	1.00	1.00
.07	≈∞	≈∞	≈∞	≈∞	1.20	1.10	1.10	1.00	1.00	1.00	1.00	1.00	1.00
.08	≈∞	≈∞	≈∞	1.20	1.10	1.10	1.10	1.00	1.00	1.00	1.00	1.00	1.00
.09	≈∞	≈∞	≈∞	1.20	1.10	1.10	1.10	1.00	1.00	1.00	1.00	1.00	0.99
.10	≈∞	≈∞	≈∞	1.20	1.10	1.10	1.10	1.03	1.02	1.02	1.03	1.01	1.00
.15	≈∞	≈∞	1.20	1.10	1.10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
.20	≈∞	≈∞	1.20	1.08	1.05	1.03	1.03	1.01	1.00	1.00	1.00	1.00	1.00
.25	≈∞	≈∞	1.10	1.10	1.10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.30	≈∞	1.20	1.10	1.10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.35	≈∞	1.20	1.10	1.10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.40	≈∞	1.20	1.10	1.10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.45	≈∞	1.20	1.10	1.10	1.03	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.50	≈∞	1.20	1.10	1.10	1.00	1.01	1.01	1.01	1.01	1.00	1.00	1.00	1.00

Note that by ≈∞, we are indicating ORs in the millions.

Table 7. Median odds ratios, reflecting “median unbiasedness” for skewed sampling distributions.

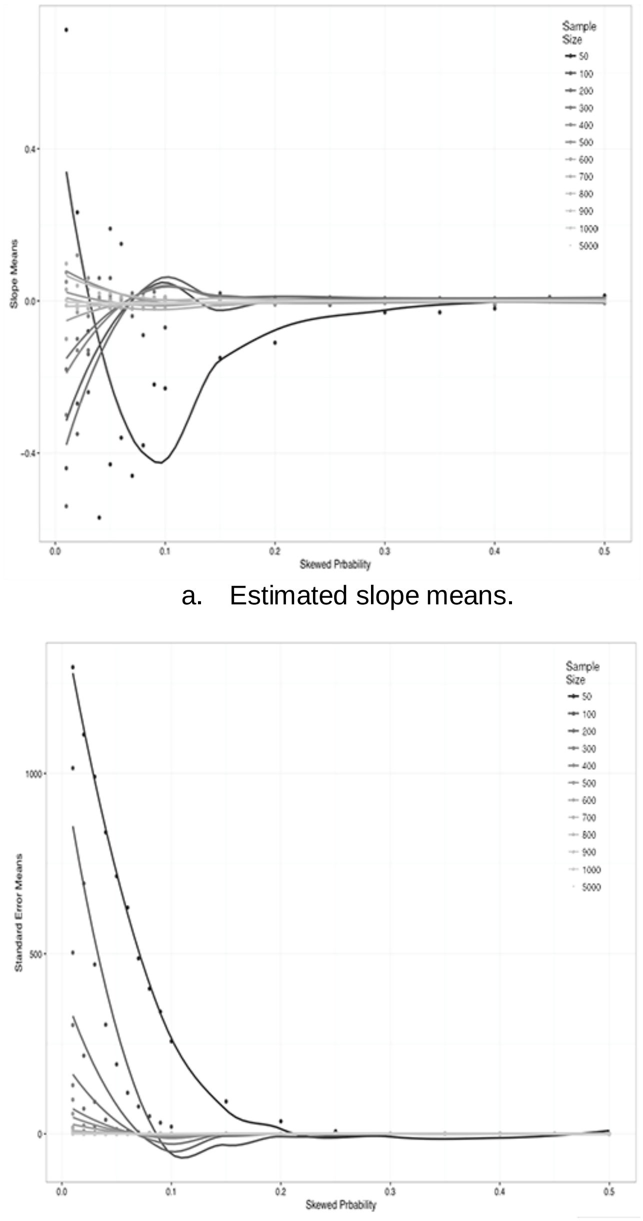
Probability	**Sample Size**												
	10	*50*	*100*	*200*	*300*	*400*	*500*	*600*	*700*	*800*	*900*	*1,000*	*5,000*
.01	0.00	1.08	0.95	0.98	1.00	0.98	1.00	1.00	1.00	0.99	1.03	0.99	1.01
.02	≈∞	1.00	0.96	0.95	0.98	0.97	1.00	0.98	1.00	0.97	1.00	0.99	0.99
.03	≈∞	1.00	1.00	0.97	0.97	0.99	1.00	1.00	0.99	0.98	1.00	1.00	0.99
.04	1.80	1.00	1.00	1.00	0.97	1.01	1.00	1.00	1.00	0.97	1.00	0.99	0.99
.05	1.70	1.00	1.00	1.00	1.00	0.98	1.00	0.98	1.00	0.98	1.00	1.00	0.99
.06	1.70	1.00	1.00	0.97	1.01	0.99	1.00	0.99	1.00	0.98	0.99	1.00	0.99
.07	1.70	1.00	1.00	0.97	1.01	1.00	0.99	0.98	1.00	0.98	0.99	0.99	0.99
.08	1.70	0.92	1.00	1.00	1.00	0.99	1.00	0.98	1.00	0.97	1.00	0.99	0.99
.09	1.50	1.00	1.00	1.00	1.00	0.99	1.00	0.98	1.00	0.97	1.00	0.99	0.99
.10	1.40	1.00	1.00	1.00	1.02	1.00	1.00	0.99	1.00	0.99	0.99	0.99	0.99
.15	1.00	1.00	1.00	1.00	1.00	0.99	1.00	0.99	0.99	0.98	1.00	0.99	0.99
.20	1.00	1.00	1.03	1.00	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
.25	1.00	1.00	1.00	1.00	1.00	0.98	0.99	0.99	1.00	0.99	1.00	1.00	1.00
.30	1.00	0.97	1.00	1.00	1.00	1.00	1.00	0.98	0.99	1.00	0.99	0.99	0.99
.35	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99	0.99	1.00	1.00	0.99
.40	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99	1.00	1.00	0.99	1.00	1.00
.45	1.00	1.00	1.00	1.00	0.99	0.99	0.99	1.00	1.00	0.99	0.99	0.99	0.99
.50	1.00	1.00	1.00	1.00	0.99	0.99	0.99	1.00	0.99	0.99	0.99	0.99	0.99

Note that by ≈∞, we are indicating ORs in the millions.

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 6 follow the trend in Table 4, 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.

Why is the Type I error rate consistently conservative?

The results in Tables 4 through 7 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 average standard error and average slope estimates are highly biased. Figure 1 is a line graph that shows the slope and standard error, where the y-axis is the slope or standard error and the x-axis is the skewed probability of the predictor. The line shades represent different sample sizes. Tables 8 and 9 contain the values from which these graphs were derived.



a. Estimated slope means.

b. Estimated standard error means.

Figure 1. Slope and standard error averages.

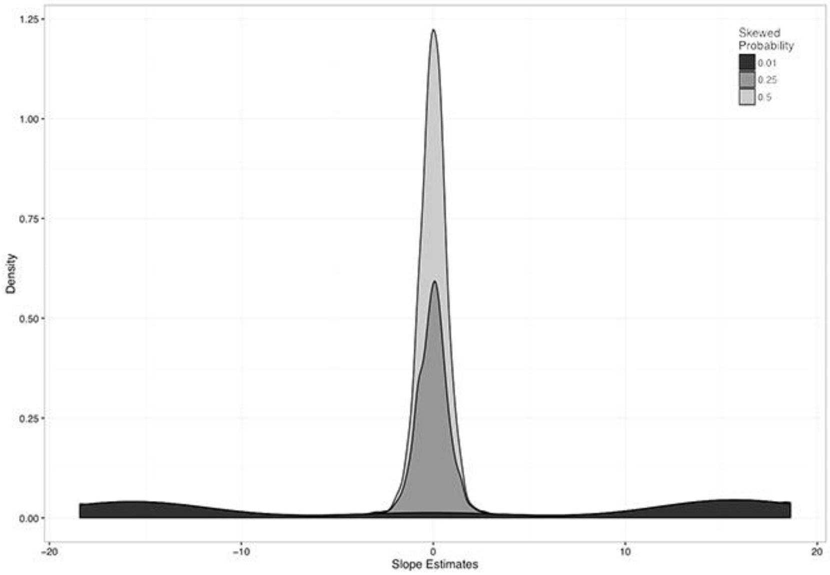


Figure 2. Distribution functions for the experimental condition: Sample size = 50, skewed probability = 0.01, 0.25, and 0.5.

As seen in Figure 1 and Table 8, the bias for the slope is both positive and negative when the sample size and highly skewed cell probability. The bias in slope is not as great as the bias in the ORs (as seen in Table 6). Consider a few examples to understand the distribution of the slope parameter estimate, and why it averages out to a small bias. 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. Shown in Figure 2 is 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 8. The 25th, 50th and 75th 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 25th, 50th, and 75th quantiles are −0.51, $-1.0 \times e{-16}$, and 0.43, respectively.

Table 8. Average slope in each cell of the simulation design for Model 1.

Probability	**Sample Size**												
	10	*50*	*100*	*200*	*300*	*400*	*500*	*600*	*700*	*800*	*900*	*1,000*	*5,000*
.01	-0.36	0.71	-0.44	-0.54	-0.18	-0.30	0.05	0.08	-0.10	0.03	0.10	-0.01	0.00
.02	1.20	0.23	-0.27	-0.35	-0.10	-0.13	0.12	-0.03	-0.01	-0.01	0.04	-0.02	-0.01
.03	1.70	-0.08	-0.24	-0.14	-0.13	-0.04	0.06	-0.01	-0.02	-0.02	0.03	-0.01	0.00
.04	1.30	-0.57	-0.01	0.06	-0.01	0.01	0.01	0.00	-0.01	-0.01	0.02	-0.01	0.00
.05	1.40	-0.43	0.19	0.06	0.01	0.00	0.01	-0.01	0.00	-0.02	0.02	-0.01	0.00
.06	1.10	-0.36	0.15	-0.02	0.01	0.00	0.01	-0.01	0.00	-0.02	0.01	-0.01	-0.01
.07	1.00	-0.46	-0.02	-0.04	0.02	0.01	0.01	-0.01	0.00	-0.02	0.01	-0.01	0.00
.08	1.00	-0.38	-0.09	-0.02	0.02	0.01	0.00	-0.01	0.00	-0.02	0.01	-0.01	0.00
.09	0.90	-0.22	-0.02	0.00	0.02	0.01	0.00	-0.01	-0.01	-0.02	0.01	-0.01	-0.01
.10	0.80	-0.23	-0.07	0.01	0.01	0.01	0.00	-0.01	-0.01	-0.02	0.01	-0.01	0.00
.15	0.30	-0.15	0.02	0.01	0.00	0.01	0.00	0.00	-0.01	-0.01	0.01	0.00	0.00
.20	0.02	-0.11	0.00	0.01	0.01	0.00	-0.01	0.00	0.00	0.00	0.00	0.00	0.00
.25	-0.04	-0.01	0.00	0.01	0.01	0.00	0.00	-0.01	0.00	0.00	0.00	0.00	0.00
.30	0.12	-0.03	0.00	0.01	0.01	0.00	0.00	-0.01	0.00	0.00	0.00	0.00	0.00
.35	0.24	-0.03	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
.40	0.27	-0.02	-0.01	0.01	0.01	-0.01	-0.01	0.00	0.00	0.00	0.00	0.00	0.00
.45	-0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
.50	-0.13	0.01	0.00	0.01	0.00	-0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00

For comparison purposes, the distribution of the slope estimates was examined when the probability of the predictor is balanced (i.e., $p = 0.5$). As shown in Table 8, the estimated slopes are close to the simulated values of zero. 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. This suggests that the distribution of the simulated slope estimates is disrupted by the skewness in the probability of the predictor.

Consider 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 25th, 50th, and 75th quantiles are −0.66, 0.008, and 0.696, respectively. For the same sample size and a moderate probability of 0.25, we 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 25th, 50th, and 75th 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 8, is −0.004 and the 25th, 50th, and 75th quantiles are −0.123, 0.0003, and 0.12, respectively.

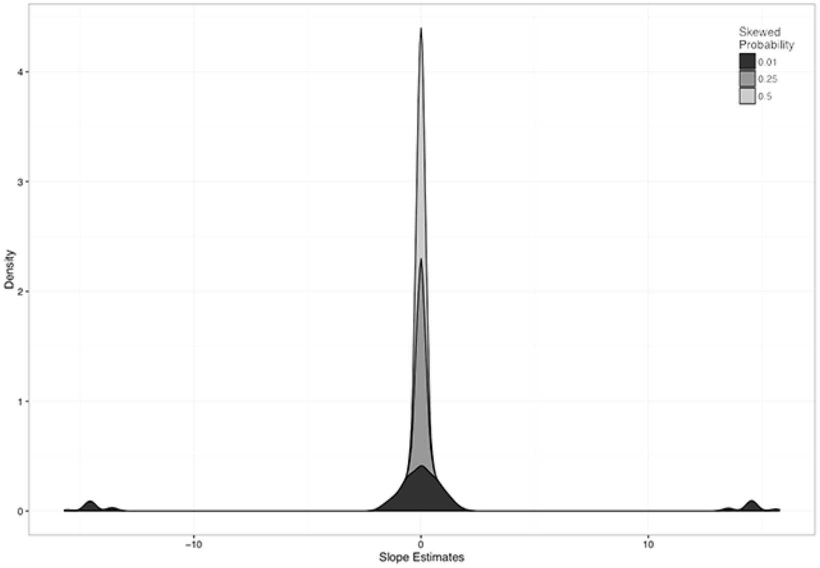


Figure 3. Distribution functions for the experimental condition: Sample size = 500, skewed probability = 0.01, 0.25, and 0.5.

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.

Depicted in Table 9 are 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 through 0.2, the average standard errors are in the thousands and range from [0.24, 4500], as shown in Table 9. As 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 9. Average standard error of the slope in each cell of the simulation design for Model 1.

Probability	Sample Size												
	10	*50*	*100*	*200*	*300*	*400*	*500*	*600*	*700*	*800*	*900*	*1,000*	*5,000*
.01	3956	1295	1015	503	302	135	95	56	18	16	8	4	0.28
.02	4522	1108	695	217	70	23	8	3	1.10	0.54	0.50	0.48	0.20
.03	4221	991	470	89	21	4	2	0.50	0.46	0.43	0.40	0.38	0.16
.04	4360	837	303	39	6	0.54	0.48	0.43	0.39	0.37	0.35	0.33	0.15
.05	414	715	193	13	0.57	0.48	0.43	0.38	0.36	0.33	0.31	0.29	0.13
.06	4053	628	114	7	0.51	0.44	0.38	0.35	0.32	0.30	0.28	0.27	0.12
.07	3914	487	76	4	0.47	0.40	0.36	0.33	0.30	0.28	0.27	0.25	0.11
.08	3836	403	49	2	0.44	0.38	0.34	0.31	0.28	0.26	0.25	0.23	0.10
.09	3830	339	31	0.52	0.42	0.36	0.32	0.29	0.27	0.25	0.24	0.22	0.10
.10	3525	257	20	0.49	0.40	0.34	0.30	0.28	0.26	0.24	0.22	0.21	0.09
.15	3617	90	0.59	0.41	0.33	0.28	0.25	0.23	0.21	0.20	0.18	0.18	0.08
.20	3076	35	0.52	0.36	0.29	0.25	0.22	0.21	0.19	0.18	0.17	0.16	0.07
.25	2434	7	0.47	0.33	0.27	0.23	0.21	0.19	0.17	0.16	0.15	0.15	0.07
.30	2067	0.65	0.44	0.32	0.25	0.22	0.19	0.18	0.17	0.15	0.14	0.14	0.06
.35	1744	0.62	0.43	0.30	0.24	0.20	0.18	0.17	0.16	0.15	0.14	0.13	0.06
.40	1608	0.60	0.42	0.29	0.24	0.21	0.18	0.17	0.15	0.14	0.14	0.13	0.06
.45	1535	0.59	0.41	0.28	0.23	0.20	0.18	0.16	0.15	0.14	0.13	0.13	0.06
.50	1483	0.59	0.41	0.28	0.23	0.20	0.18	0.16	0.15	0.14	0.13	0.12	0.06

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.

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 6 and 7 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, our 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. As Cohen (1988) clearly stated, all effect size guidelines are research context dependent and should only be used in the research settings from which they were derived.

Method

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 was 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, the investigation pertained to what happens when the predictor variable is balanced. The third factor added is effect size, which was 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. Hence, it was assumed β_0 and β_1 are fixed to a number different from zero. The intercept parameter was fixed to −2. For the slope parameter, three levels of effect size were considered: 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 1,000 replications of the same model.

Analysis of simulation results. To assess power estimates, a framework was adopted similar to Bradley's for Type I error rates, to determine at what level of skewness 10% and 50% of expected statistical power was lost. 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 10, 11, and 12 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% to 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, it can be seen 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 1,000. 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 1,000, 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 to 0.25 for the same sample sizes. Power is highly influenced by skewed probabilities in small sample sizes, even when the effect size is moderate to large and hence highly detectable.

Table 10. Power with low effect size (OR = 2).

Probability	**Sample Size**											
	50	*100*	*200*	*300*	*400*	*500*	*600*	*700*	*800*	*900*	*1,000*	*5,000*
.01	0.00	0.01	0.04	0.06	0.08	0.10	0.13	0.13	0.15	0.17	0.17	0.50
.02	0.01	0.04	0.09	0.12	0.13	0.17	0.18	0.19	0.22	0.25	0.27	0.75
.03	0.03	0.05	0.11	0.16	0.16	0.21	0.25	0.27	0.29	0.32	0.32	0.87
.04	0.03	0.07	0.12	0.18	0.21	0.24	0.30	0.32	0.34	0.38	0.41	0.94
.05	0.05	0.09	0.14	0.21	0.24	0.28	0.35	0.39	0.40	0.45	0.47	0.98
.06	0.06	0.11	0.16	0.24	0.26	0.33	0.38	0.43	0.45	0.49	0.52	0.99
.07	0.08	0.13	0.17	0.26	0.28	0.36	0.42	0.45	0.49	0.55	0.59	0.99
.08	0.08	0.14	0.20	0.28	0.32	0.40	0.46	0.51	0.54	0.59	0.63	0.99
.09	0.09	0.15	0.21	0.30	0.31	0.41	0.49	0.54	0.58	0.64	0.67	1.00
.10	0.09	0.17	0.22	0.32	0.36	0.45	0.52	0.57	0.62	0.67	0.72	1.00
.15	0.12	0.18	0.30	0.39	0.49	0.59	0.65	0.70	0.74	0.81	0.85	1.00
.20	0.13	0.22	0.34	0.46	0.56	0.67	0.73	0.80	0.85	0.87	0.91	1.00
.25	0.13	0.21	0.37	0.52	0.64	0.72	0.78	0.86	0.90	0.91	0.94	1.00
.30	0.13	0.22	0.39	0.55	0.68	0.76	0.83	0.89	0.92	0.94	0.96	1.00
.35	0.11	0.23	0.42	0.58	0.69	0.78	0.84	0.90	0.93	0.94	0.98	1.00
.40	0.10	0.22	0.43	0.60	0.72	0.80	0.85	0.91	0.95	0.96	0.97	1.00
.45	0.09	0.22	0.44	0.60	0.73	0.81	0.87	0.93	0.96	0.96	0.98	1.00
.50	**0.06**	**0.21**	**0.42**	**0.61**	**0.70**	**0.80**	**0.87**	**0.92**	**0.96**	**0.96**	**0.98**	**1.00**

Note: 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.

However, 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 1,000 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.

Table 11. Power with moderate effect size (OR = 3).

Probability	**Sample Size**											
	50	*100*	*200*	*300*	*400*	*500*	*600*	*700*	*800*	*900*	*1,000*	*5,000*
.01	0.01	0.02	0.07	0.11	0.15	0.19	0.26	0.26	0.29	0.35	0.36	0.88
.02	0.01	0.07	0.15	0.23	0.29	0.35	0.41	0.43	0.45	0.54	0.54	0.99
.03	0.04	0.11	0.22	0.31	0.38	0.44	0.51	0.55	0.60	0.67	0.69	1.00
.04	0.05	0.14	0.27	0.37	0.45	0.53	0.63	0.66	0.71	0.79	0.79	1.00
.05	0.07	0.18	0.34	0.44	0.53	0.61	0.68	0.76	0.78	0.85	0.85	1.00
.06	0.09	0.21	0.36	0.50	0.58	0.69	0.74	0.82	0.84	0.89	0.90	1.00
.07	0.12	0.24	0.41	0.56	0.64	0.73	0.78	0.86	0.88	0.93	0.93	1.00
.08	0.14	0.27	0.44	0.60	0.69	0.77	0.83	0.90	0.92	0.95	0.95	1.00
.09	0.15	0.29	0.48	0.62	0.74	0.82	0.85	0.92	0.93	0.97	0.97	1.00
.10	0.16	0.32	0.52	0.66	0.78	0.85	0.89	0.94	0.96	0.98	0.98	1.00
.15	0.22	0.40	0.65	0.78	0.88	0.94	0.97	0.98	0.99	1.00	1.00	1.00
.20	0.25	0.46	0.70	0.86	0.95	0.97	0.99	0.99	1.00	1.00	1.00	1.00
.25	0.26	0.47	0.77	0.90	0.97	0.99	0.99	1.00	1.00	1.00	1.00	1.00
.30	0.28	0.51	0.80	0.94	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00
.35	0.28	0.53	0.83	0.94	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.40	0.27	0.52	0.85	0.95	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.45	0.27	0.52	0.87	0.95	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.50	**0.24**	**0.52**	**0.87**	**0.96**	**0.99**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**

Note: 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.

Table 12. Power with large effect size (OR = 4).

Probability	**Sample Size**											
	50	*100*	*200*	*300*	*400*	*500*	*600*	*700*	*800*	*900*	*1,000*	*5,000*
.01	0.00	0.02	0.07	0.15	0.23	0.29	0.36	0.40	0.42	0.49	0.51	0.98
.02	0.01	0.08	0.21	0.34	0.42	0.52	0.58	0.61	0.66	0.73	0.76	1.00
.03	0.05	0.15	0.31	0.44	0.55	0.65	0.71	0.76	0.81	0.87	0.87	1.00
.04	0.07	0.20	0.40	0.53	0.65	0.75	0.81	0.85	0.89	0.94	0.94	1.00
.05	0.10	0.25	0.49	0.62	0.74	0.82	0.86	0.90	0.95	0.97	0.98	1.00
.06	0.13	0.30	0.54	0.69	0.80	0.87	0.90	0.94	0.97	0.99	0.98	1.00
.07	0.16	0.34	0.59	0.74	0.84	0.91	0.93	0.96	0.98	1.00	0.99	1.00
.08	0.19	0.40	0.65	0.70	0.88	0.93	0.95	0.98	0.99	1.00	1.00	1.00
.09	0.21	0.41	0.69	0.83	0.91	0.95	0.97	0.98	1.00	1.00	1.00	1.00
.10	0.24	0.46	0.72	0.85	0.93	0.97	0.98	0.99	1.00	1.00	1.00	1.00
.15	0.32	0.58	0.84	0.93	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.20	0.38	0.66	0.90	0.97	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.25	0.39	0.69	0.93	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.30	0.44	0.73	0.95	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.35	0.45	0.74	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.40	0.44	0.75	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.45	0.46	0.75	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.50	**0.43**	**0.75**	**0.98**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**

Note: 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.

Power is largely affected by the skewed probability of the predictor variable. 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 1,000 and 5,000. Although it was shown in the previous study Type I error is acceptable for the majority of factors examined, it has implications on the empirical power of the Wald test. This is evidence 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.

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. The aim is 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, focus 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, the aim was to document the impact of a skewed continuous predictor variable on the estimation, parameter estimates, and Type I error rate of the Wald test.

Method

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 were matched to the expected probabilities in Model 1. Shown in Table 13 are the shape parameter values used and the equivalent skewness levels. To create a baseline, the case investigated was where the predictor variable is drawn from a standard normal distribution with a mean of zero and a standard deviation of one. The resulting simulation experiment is a 12 (sample size) by 18 (skewness) completely crossed factorial design.

Table 13. Shape parameter and equivalent skewness level.

Shape Parameter	Skewness	Probability
0.047	9.250	0.010
0.086	6.870	0.020
0.130	5.550	0.030
0.200	4.750	0.040
0.250	4.040	0.050
0.300	3.700	0.060
0.370	3.250	0.070
0.400	3.190	0.080
0.500	2.830	0.090
0.600	2.720	0.100
1.000	1.960	0.150
1.750	1.500	0.200
3.000	1.150	0.250
5.500	0.873	0.300
10.000	0.630	0.350
25.000	0.410	0.400
50.000	0.200	0.450
Standard Normal	0.000	0.500

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 14, 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 14, 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 (9.23).

Table 14. Liberal Type I error rate model.

Skewness	**Sample Size**											
	50	*100*	*200*	*300*	*400*	*500*	*600*	*700*	*800*	*900*	*1,000*	*5,000*
9.25	0.00	0.00	0.01	0.01	0.02	0.03	0.03	0.05	0.02	0.03	0.04	0.04
6.87	0.00	0.01	0.02	0.03	0.03	0.05	0.03	0.04	0.03	0.04	0.04	0.06
5.55	0.01	0.02	0.03	0.04	0.04	0.04	0.05	0.05	0.04	0.05	0.04	0.05
4.75	0.01	0.03	0.04	0.04	0.03	0.05	0.05	0.04	0.05	0.06	0.05	0.04
4.04	0.01	0.03	0.04	0.04	0.04	0.05	0.04	0.06	0.04	0.04	0.05	0.05
3.70	0.02	0.03	0.04	0.04	0.04	0.04	0.05	0.04	0.04	0.06	0.05	0.05
3.25	0.02	0.03	0.04	0.04	0.06	0.04	0.05	0.06	0.05	0.03	0.06	0.06
3.19	0.02	0.04	0.04	0.05	0.04	0.04	0.04	0.04	0.05	0.04	0.06	0.05
2.83	0.03	0.03	0.05	0.04	0.05	0.04	0.05	0.05	0.04	0.06	0.05	0.05
2.72	0.02	0.03	0.04	0.05	0.05	0.05	0.05	0.05	0.06	0.05	0.04	0.05
1.96	0.04	0.03	0.04	0.04	0.04	0.06	0.06	0.05	0.06	0.04	0.05	0.05
1.50	0.05	0.03	0.05	0.04	0.04	0.04	0.05	0.05	0.05	0.05	0.06	0.05
1.15	0.04	0.05	0.05	0.04	0.05	0.06	0.06	0.05	0.06	0.07	0.06	0.04
.873	0.04	0.06	0.04	0.05	0.05	0.05	0.05	0.04	0.04	0.04	0.04	0.06
.63	0.04	0.04	0.04	0.05	0.04	0.05	0.05	0.05	0.06	0.05	0.05	0.06
.41	0.04	0.04	0.04	0.05	0.06	0.04	0.04	0.07	0.06	0.06	0.06	0.05
.20	0.03	0.04	0.04	0.06	0.05	0.04	0.06	0.05	0.05	0.05	0.04	0.05
.00	**0.05**	**0.06**	**0.06**	**0.04**	**0.05**	**0.06**	**0.05**	**0.05**	**0.06**	**0.06**	**0.05**	**0.06**

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 5).

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 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 the 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, a 2×2 (two-factor) randomized experiment or randomized clinical trial.

Study A: Type I error rates and non-convergences

Method

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

There were 4,212 experimental conditions. Many results were identical or only a couple of decimals apart. Because of the sheer volume and the lack of variation, only a few sample sizes are presented.

Number of non-convergences. Shown in Tables 15 and 16 are 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 5,000.

Table 15. Number of non-convergences from 1,000 replications for Model 3 when sample size is 100.

x_1 Probability	x_2 Probability								
	0.01	*0.02*	*0.03*	*0.04*	*0.05*	*0.06*	*0.07*	…	*0.5*
.01	610	463	412	397	390	389	388	…	388
.02	474	260	195	165	157	155	154	…	154
.03	407	176	102	70	62	60	58	…	58
.04	384	143	68	33	23	21	19	…	19
.05	374	131	56	21	11	9	7	…	7
.06	373	129	54	19	9	7	5	…	5
.07	371	127	52	16	6	4	2	…	2
.08	371	127	52	16	6	3	1	…	1
.09	370	126	51	15	5	2	0	…	0
…	…	…	…	…	…	…	…	⋱	0
.50	370	126	51	15	5	2	0	0	0

Table 16. Number of non-convergences from 1,000 replications for Model 3 when sample size is 400.

x_1 Probability	x_2 Probability				
	.01	*.02*	*.03*	…	*.50*
.01	41	21	21	…	21
.02	21	1	1	…	1
.03	20	0	0	…	0
…	…	…	…	⋱	0
.50	20	0	0	0	0

Type I error rate. The effect of the skewness of x_2 on the Type I error rate was compared for the Wald test of x_1. The results in Table 17 are formatted somewhat differently from those in other tables in this paper. Following Conover, Johnson, and Johnson (1981), average Type I error rates were used. Presented in the table are 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 17 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.

Table 17. 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.

x_1 Probability	***Sample Size***			
	100	*400*	*900*	*5000*
.01	0 (0,0)	0.002 (.001,.003)	0.019 (.017,.021)	0.053 (.051,.057)
.02	0 (0,0)	0.015 (.013,.018)	0.039 (.035,.043)	0.041 (.035,.042)
.03	0 (0,0)	0.023 (.021,.026)	0.043 (.040,.044)	0.043 (.041,.045)
.04	0 (0,0)	0.036 (.033,.039)	0.045 (.042,.046)	0.036 (.033,.052)
.05	0.003 (.001,.005)	0.041 (.038,.043)	0.058 (.056,.060)	0.040 (.039,.041)
.06	0.01 (.005,.011)	0.039 (.036,.046)	0.047 (.045,.049)	0.047 (.046,.049)
.07	0.01 (.010,.020)	0.046 (.043,.049)	0.056 (.053,.059)	0.045 (.043,.047)
.08	0.019 (.016,.022)	0.05 (.048,.054)	0.063 (.061,.065)	0.049 (.049,.051)
.09	0.019 (.017,.023)	0.05 (.048,.055)	0.062 (.060,.064)	0.038 (.038,.040)
.10	0.023 (0,.030)	0.049 (.047,.052)	0.052 (.050,.055)	0.047 (.046,.049)
.15	0.037 (.034,.040)	0.062 (.060,.065)	0.055 (.052,.061)	0.049 (.049,.051)
.20	0.045 (.034,.048)	0.049 (.047,.057)	0.045 (.042,.047)	0.042 (.041,.043)
.25	0.041 (.035,.044)	0.048 (.046,.049)	0.047 (.045,.050)	0.052 (.051,.053)
.30	0.047 (.039,.053)	0.054 (.052,.056)	0.038 (.036,.041)	0.055 (.054,.056)
.35	0.056 (.039,.060)	0.053 (.052,.055)	0.048 (.046,.050)	0.047 (.046,.048)
.40	0.049 (.047,.052)	0.066 (.062,.069)	0.053 (.050,.054)	0.050 (.035,.055)
.45	0.049 (.037,.052)	0.058 (.055,.061)	0.040 (.038,.045)	0.049 (.047,.051)
.50	**0.056** **(.052,.059)**	**0.057** **(.052,.060)**	**0.047** **(.046,.049)**	**0.042** **(.042,.044)**

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 5).

The range of average Type I error rates for each cell in Table 17 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.

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 was chosen, applying it to ORs.

Method

In addition to the three factors mentioned in Study A, a fourth factor was added: 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. Assume β_0, β_1, and β_2 are fixed to a number different from zero. The intercept parameter was fixed to −2. Examined three levels of effect size 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, we fixed the effect size for β_2 to a moderate value of 1.1. 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, the best achievable power in the case of a balanced design was compared with other combinations of probabilities for each sample size. Both 10% and 50% loss of power were considered. The resulting number of conditions was 12636. Similarly to Study A, four typical sample sizes: 100, 400, 900, and 5,000 are presented. Shown in Tables 18 through 20 are power estimates for x_1 averaged over all levels of probability of x_2. As in Table 17, Tables 18 through 20 include the range of statistical power for each condition.

Table 18. 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.

x_1 **Probability**	***Sample Size***			
	100	*400*	*900*	*5,000*
.01	0.01 (0,.013)	0.10 (.074,.110)	0.17 (.160,.180)	0.53 (.500,.600)
.02	0.04 (.020,.210)	0.14 (.140,.150)	0.26 (.240,.280)	0.80 (.770,.850)
.03	0.05 (.040,.055)	0.166 (.150,.180)	0.33 (.310,.360)	0.92 (.890,.950)
.04	0.08 (.065,.085)	0.213 (.200,.230)	0.41 (.340,.450)	0.97 (.950,1)
.05	0.09 (.011,.110)	0.26 (.230,.280)	0.47 (.440,.530)	0.99 (.980,1)
.06	0.10 (.096,.011)	0.30 (.270,.397)	0.53 (.490,.590)	0.99 (.990,1)
.07	0.12 (.110,.130)	0.32 (.300,.350)	0.57 (.530,.660)	1 (.990,1)
.08	0.13 (.120,.140)	0.36 (.340,.390)	0.62 (.580,.700)	1 (.990,1)
.09	0.15 (.130,.160)	0.38 (.360,.410)	0.66 (.620,.730)	1 (1,1)
.10	0.15 (.140,.170)	0.40 (.380,.440)	0.70 (.660,.780)	1 (1,1)
.15	0.18 (.160,.200)	0.52 (.480,.580)	0.84 (.810,.890)	1 (1,1)
.20	0.197 (.170,.230)	0.60 (.560,.670)	0.91 (.890,.950)	1 (1,1)
.25	0.218 (.190,.260)	0.67 (.620,.750)	0.95 (.930,.980)	1 (1,1)
.30	0.23 (.190,.290)	0.72 (.680,.790)	0.96 (.950,.980)	1 (1,1)
.35	0.23 (.200,.310)	0.74 (.690,.820)	0.97 (.960,.980)	1 (1,1)
.40	0.25 (.220,.310)	0.76 (.720,.830)	0.98 (.970,.990)	1 (1,1)
.45	0.25 (.200,.300)	0.76 (.710,.840)	0.98 (.960,.990)	1 (1,1)
.50	**0.246** **(0.210,.300)**	**0.76** **(.710,.850)**	**0.98** **(.970,.990)**	**1** **(1,1)**

Note: 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.

Table 19. 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.

x_1 Probability	***Sample Size***			
	100	*400*	*900*	*5,000*
.01	0.02 (.010,.300)	0.15 (.120,.170)	0.33 (.320,.305)	0.90 (.880,.930)
.02	0.05 (.040,.060)	0.28 (.270,.290)	0.54 (.520,.560)	0.99 (.980,1)
.03	0.09 (.080,.100)	0.38 (.370,.400)	0.68 (.660,.730)	0.99 (.990,1)
.04	0.14 (.120,.150)	0.46 (.440,.480)	0.79 (.770,.830)	0.99 (.990,1)
.05	0.18 (.180,.200)	0.53 (.510,.570)	0.86 (.840,.900)	0.99 (.980,1)
.06	0.22 (.210,.230)	0.6 (.570,.630)	0.90 (.880,.940)	0.99 (.990,1)
.07	0.25 (.240,.270)	0.66 (.640,.710)	0.94 (.930,.950)	1 (.990,1)
.08	0.27 (.260,.290)	0.71 (.690,.750)	0.96 (.950,.970)	1 (.990,1)
.09	0.3 (.290,.330)	0.76 (.740,.780)	0.97 (.960,.980)	1 (1,1)
.10	0.32 (.300,.360)	0.80 (.750,.920)	0.98 (.970,.990)	1 (1,1)
.15	0.40 (.320,.440)	0.90 (.800,.930)	0.99 (.980,1)	1 (1,1)
.20	0.46 (.430,.520)	0.96 (.930,.980)	0.99 (.990,1)	1 (1,1)
.25	0.49 (.450,.560)	0.98 (.940,.990)	1 (1,1)	1 (1,1)
.30	0.55 (.500,.620)	0.99 (.980,.990)	1 (1,1)	1 (1,1)
.35	0.56 (.520,.630)	0.99 (.980,.990)	1 (1,1)	1 (1,1)
.40	0.58 (.530,.670)	0.99 (.990,.990)	1 (1,1)	1 (1,1)
.45	0.59 (.540,.670)	0.99 (.990,.990)	1 (1,1)	1 (1,1)
.50	**0.59** **(.440,.680)**	**0.99** **(.990,.990)**	**1** **(1,1)**	**1** **(1,1)**

Note: 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.

Table 20. 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.

x_1 Probability	***Sample Size***			
	100	*400*	*900*	*5,000*
.01	0.02 (.010,.040)	0.20 (.170,.230)	0.50 (.450,.990)	0.98 (.970,1)
.02	0.07 (.050,.090)	0.39 (.380,.410)	0.74 (.720,.770)	0.99 (.990,1)
.03	0.13 (.110,.160)	0.55 (.530,.570)	0.87 (.850,.900)	1 (1,1)
.04	0.19 (.170,.220)	0.65 (.630,.680)	0.95 (.940,.970)	1 (1,1)
.05	0.26 (.240,.270)	0.74 (.700,.780)	0.97 (.970,.900)	1 (1,1)
.06	0.31 (.290,.330)	0.80 (.770,.830)	0.98 (.970,.990)	1 (1,1)
.07	0.36 (.340,.370)	0.86 (.830,.890)	0.99 (.990,1)	1 (1,1)
.08	0.39 (.380,.410)	0.89 (.860,.910)	0.99 (.990,1)	1 (1,1)
.09	0.44 (.420,.460)	0.92 (.910,.940)	0.99 (.990,1)	1 (1,1)
.10	0.47 (.450,.500)	0.94 (.930,.960)	0.97 (.980,1)	1 (1,1)
.15	0.59 (.570,.630)	0.99 (.980,.990)	1 (1,1)	1 (1,1)
.20	0.67 (.630,.720)	0.99 (.990,1)	1 (1,1)	1 (1,1)
.25	0.72 (.600,.780)	0.99 (.990,1)	1 (1,1)	1 (1,1)
.30	0.77 (.720,.830)	0.99 (.990,1)	1 (1,1)	1 (1,1)
.35	0.79 (.730,.850)	1 (1,1)	1 (1,1)	1 (1,1)
.40	0.81 (.760,.870)	1 (1,1)	1 (1,1)	1 (1,1)
.45	0.82 (.770,.880)	1 (1,1)	1 (1,1)	1 (1,1)
.50	**0.82** **(.770,.880)**	**1** **(1,1)**	**1** **(1,1)**	**1** **(1,1)**

Note: 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.

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. Considering the range of statistical power for each condition, 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. However, 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 18. Power is dramatically affected by the deflation in the Type I error rate. It is shown 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.

Conclusion

It is not uncommon 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 the question of the impact of skewed predictors on the eventual statistical results of a logistic regression. To answer this general question, five inter-related simulation studies were conducted, which to our knowledge are the first of their kind to be done for skewed dichotomous predictors.

A broad picture of the effects of skewed cell probabilities in dichotomous predictors on the logistic regression model is provided, specifically regarding how a categorical predictor's skewness in probabilities affect estimation, parameter estimates, and the Wald test. 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.

Skewed probabilities can induce separation, which automatically affects estimation and results in non-convergence (Albert & Anderson, 1984). When separation does not occur—even in severe cases of skewed probability—ML converges and estimates are produced.

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.

The skewed binary predictors can depending on the research design as (i) observed groups (e.g., gender differences or rare diagnostic or disease states) that have skewed probabilities of occurrence, or (ii) in other research settings these binary predictor(s) can be considered elements of a design matrix for experiments or clinical trials wherein the imbalance is group sizes is not due to selection bias or attrition.

Therefore, with data analysts and consumers of research materials in mind, guidelines are suggested on how to think about and handle skewed predictor(s) in a logistic regression analysis based on whether the skewness is severe or not.

1) If skewness is severe (i.e. the shaded areas in the tables included in this paper), there are two cases to consider. The first case under severe skewness is when separation (by inspecting the data structure) or zero cell counts (by looking at multi-way table) occur in the data. In this case, the conventional maximum likelihood estimator will not converge and estimates are not produced. Data analysts are encouraged to use alternative estimation procedures for example the ones found in Bull, Mak, and Greenwood (2002) and Heinze (2006). The second case under severe skewness occurs when estimates are produced by the conventional maximum likelihood estimator. This is due to the fact that one does not have zero cell counts nor separation. In this case of severe skewness, however, the Wald test is not reliable and an alternative test, such as the likelihood ratio test, is recommended.

2) If skewness is not severe (i.e. the non-shaded areas in the tables mentioned in this paper), the Wald statistic is reliable and the interpretation of the test statistic should follow general statistical analysis recommendations (i.e. do not rely on the test statistic

exclusively but also examine effect size, parameter estimates, standard errors, fit, etc.). In addition, because skew of the predictor may impact on the operating characteristics of the statistical test, when planning a study a researcher/reader should take this in to account when computing the level of power and assumed Type I error rate.

Acknowledgments

The first author would like to acknowledge the supporting and funding from King Abdulaziz University. This project is funded by the UBC-Paragon Research Agreement.

References

Agresti, A. (2002). *Categorical data analysis* (3rd Ed). Hoboken, NJ: John Wiley & Sons, Inc. doi: 10.1002/0471249688

Albert, A., & Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, *71*(1), 1–10. doi: 10.1093/biomet/71.1.1

Alkhalaf, A. A. (2014). *Application of Logistic Regression in Higher Education Research: A sample from 2000-2013*. Unpublished Manuscript, Measurement, Evaluation, & Research Methodology Program, University of British Columbia, Vancouver, Canada.

Anderson, J. A., & Richardson, S. C. (1979). Logistic discrimination and bias correction in maximum likelihood estimation. *Technometrics*, *21*(1), 71–78. doi: 10.2307/1268582

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. doi: 10.7198/s2237-0722201400010024

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*(2), 144–152. doi: 10.1111/j.2044-8317.1978.tb00581.x

Bull, S. B., Mak, C., & Greenwood, C. M. T. (2002). A modified score function estimator for multinomial logistic regression in small samples.

Computational Statistics and Data Analysis, *39*, 57–74. doi: 10.1016/s0167-9473(01)00048-2

Christensen, R. (2016). *Analysis of variance, design, and regression: linear modeling for unbalanced data* (2nd Ed.). Boca Raton, Florida: CRC Press, Taylor & Francis Group.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd Ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

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. doi: 10.2307/1268225

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.

Day, N. E., & Kerridge, D. F. (1967). A general likelihood discriminant. *Biometrics*, *23*(2), 313–323. doi: 10.2307/2528164

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. doi: 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. doi: 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. doi: 10.1214/aoms/1177730979

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. doi: 10.1016/0165-1765(86)90232-6

Hauck, Jr., W. W., & Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association*, *72*(360), 851–853. doi: 10.2307/2286473

Heinze, G. (2006). A comparative investigation of methods for logistic regression with separated or nearly separated data. *Statistics in Medicine*, *25*, 4216–4226. doi: 10.1002/sim.2687

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. doi: 10.1002/sim.3794

Jennings, D. E. (1986). Judging inference adequacy in logistic regression. *Journal of the American Statistical Association*, *81*(394), 471–476. doi: 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. doi: 10.2307/2286650

Lütkepohl, H., & Burda, M. M. (1997). Modified Wald tests under nonregular conditions. *Journal of Econometrics*, *78*, 315–332. doi: 10.1016/S0304-4076(97)80015-2

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. doi: 10.20533/ijds.2040.2570.2010.0010

Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the Theory of Statistics* (3rd Ed.). McGraw Hill Inc.

Pawitan, Y. (2000). A reminder of the fallibility of the Wald statistic : likelihood explanation. *The American Statistician*, *54*(1), 54–56. doi: 10.2307/2685612

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. doi: 10.1016/s0895-4356(96)00236-3

Rousseeuw, P. J., & Christmann, A. (2003). Robustness against separation and outliers in logistic regression. *Computational Statistics and Data Analysis*, *43*, 315–332. doi: 10.1016/s0167-9473(02)00304-3

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. doi: 10.2307/1402935

Zorn, C. (2005). A solution to separation in binary response models. *Political Analysis*, *13*(2), 157–170. doi: 10.1093/pan/mpi009