DIGITALCOMMONS
@WAYNESTATE

Journal of Modern Applied Statistical Methods

Volume 18 | Issue 2 Article 25

9-25-2020

Logistic Regression Under Sparse Data Conditions

David A. Walker
Northern Illinois University, dawalker@niu.edu

Thomas J. Smith
Northern Illinois University, tjsmith@niu.edu

Recommended Citation

Walker, D. A., & Smith, T. J. (2019). Logistic regression under sparse data conditions. Journal of Modern Applied Statistical Methods, 18(2), eP3372. doi: 10.22237/jmasm/1604190660

Journal of Modern Applied Statistical Methods
November 2019, Vol. 18, No. 2, eP3372.
doi: 10.22237/jmasm/1604190660

ISSN 1538 − 9472

Logistic Regression Under Sparse Data Conditions

David A. Walker
Northern Illinois University
DeKalb, IL

Thomas J. Smith
Northern Illinois University
DeKalb, IL

The impact of sparse data conditions was examined among one or more predictor variables in logistic regression and assessed the effectiveness of the Firth (1993) procedure in reducing potential parameter estimation bias. Results indicated sparseness in binary predictors introduces bias that is substantial with small sample sizes, and the Firth procedure can effectively correct this bias.

Keywords: Sparse, data, logistic regression, Firth

Introduction

Binary logistic regression, an analytic approach that uses one or more continuous or categorical variables to predict the log-odds of a binary event's occurrence, is a commonly employed technique in education and the social sciences. Logistic regression identifies an optimally-weighted linear combination of the predictors, where each regression weight (β_i) typically is estimated using maximum likelihood (ML) estimation, specifically maximizing the log-likelihood function, ln L($\beta \mid y$). The ML estimate of each slope parameter, $\hat{\beta}_i$ indicates the predicted change in the log-odds of the event's occurrence per unit of change in its associated predictor, adjusting for other predictors in the model.

Although logistic regression is a relatively robust technique in the sense that it does not require characteristics such as normality of continuous predictors, linearity, or homoscedasticity, estimation difficulties can occur if sparseness is evident in the data, typically viewed as a condition in which one of the two outcome categories has a very small number of observed values. For example, if an analyst is interested in predicting the likelihood of an individual becoming a professional athlete using a set of three personal characteristics as predictors and, among the

doi: 10.22237/jmasm/1604190660 | Accepted: March 5, 2020; Published: September 25, 2020.
Correspondence: David A. Walker, dawalker@niu.edu

1000 observed individuals, only 15 report themselves as professional athletes, a sparse data condition is evident. As another example, suppose the analyst wishes to predict, among a set of high school seniors, the probability of acceptance into a top tier (e.g., Ivy League) college/university, where only a small percentage of such seniors have achieved acceptance. Sparse occurrence of an outcome category often, although not always, is referenced in terms of the occurrence of this outcome relative to the number of predictor variables in the model. A general rule is that at least 10 events per variable (EPV) is necessary—sometimes referred to as the "Rule of Ten" (Hair et al., 2011). Considerable debate exists, however, concerning the reliability of this rule (e.g., van Smeden et al., 2016). Some authors (e.g., Vittinghoff & McColloch, 2007) suggested EPV may be relaxed and, in certain contexts, results from regression with EPV values of 5-9 should not summarily be discounted.

Several undesirable phenomena can occur under sparse data conditions. One of these is the risk of complete separation, a condition in which a predictor variable predicts the outcome variable perfectly. For example, in the data condition represented in Table 1, the predictor variable x_1 perfectly predicts the binary outcome, y. That is, all observed values of $y = 0$ have associated values of x_1 that are less than 5. Conversely, all observed values of $y = 1$ have associated values of x_1 that are greater than or equal to 5. In this situation, y is perfectly predicted by x_1, and $\hat{\beta}_1$ is thus not estimable and, in fact, is an infinite value. A similar condition, known as "quasi-complete separation," can occur when a predictor variable predicts an outcome variable to a considerable extent (see UCLA Statistical Consulting Group, 2017). The variable x_2 in Table 1, for example, predicts the outcome variable (y) very well, with less than perfect prediction evident only for values of $x_2 = 5$. In this situation, too, $\hat{\beta}_1$ is not uniquely estimable.

Even when the risk of complete separation is not high (i.e., a sufficient EPV value is evident), bias in the predicted probabilities can occur when the incidence of the event is small relative to the observed sample size. For example, suppose once again that three predictors were used to estimate the likelihood of an individual becoming a professional athlete. If data from 10,000 individuals were collected, and among those 10,000 athletes, 150 became professional athletes, the EPV is sufficiently high (EPV = 150/3 = 50), but the relative likelihood of the event of interest is still small (150/10,000 = .015). In this case, the risk is not of complete separation but, rather, of bias in the predicted probability of becoming a professional athlete.

Table 1. Example of data set demonstrating complete separation on x_1 and quasi-complete separation on x_2

y	x_1	x_2
0	1	3
0	3	2
0	4	4
0	2	5
1	5	5
1	6	8
1	5	7
1	7	8

Manski and Lerman (1977) and Prentice and Pyke (1979) independently proposed a correction to the estimated intercept term in the logistic regression equation to correct for this bias in predicted probabilities,

$$\beta_0 - \ln\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right),$$

where β_0 is the estimated intercept parameter, τ is an estimate of the proportion of successes in the population based on prior information, and $\bar{y}$ is the proportion of successes observed in the sample.

Rather than maximizing the log-likelihood function to obtain regression parameter estimates, another approach involves maximizing a weighted log-likelihood function,

$$\ln \mathrm{L}_{\mathrm{w}}\left(\beta \mid y\right) = -\sum_{i=1}^{n} w_i \ln\left(1+\mathrm{e}^{(1-2y_i)x_i\beta}\right),$$

where

$$w_i = w_1 Y_i + w_0\left(1-Y_i\right),\; w_1 = \frac{\tau}{\bar{y}},\; \text{and } w_0 = \frac{(1-\tau)}{(1-\bar{y})}.$$

This approach, like the intercept-correction approach discussed above, corrects the estimates for bias due to sparseness, but does so by adjusting the loss function.

Generally, in the presence of rare events, the estimated probability of the rare event tends to be underestimated, while the probability of the alternative event typically is overestimated. However, even in the presence of rare events, applied researchers seldom correct for the biases that can occur in these situations (King & Zeng, 2001).

Firth (1993) proposed correcting the bias introduced by the presence of sparse outcomes through the use of a penalized log-likelihood function,

$$\ln \mathrm{L}_{\mathrm{F}}\left(\beta \mid y\right) = \ln \mathrm{L}\left(\beta \mid y\right) + \frac{1}{2}\ln\left|\mathbf{I}\left(\beta\right)\right|,$$

where $\mathbf{I}(\beta)$ is the Fisher information matrix (equivalently, minus the second derivative of the log-likelihood). The Firth procedure, which is an available option in SAS, Stata, and the R package logistf, can be used to address situations with sparse data conditions, either when the EPV value is small, or the observed proportion of an outcome is small. Thus, it can address issues of complete separation and/or bias in predicted probabilities.

Although emphasis on sparse data conditions typically has focused on the distribution of the binary outcome variable in logistic regression, little research has investigated how sparse data conditions in the predictor variables may result in complete/quasi-complete separation or other estimation bias. The present study employs data simulation methods to explore this issue.

The purpose of this study was to examine the impact of sparse data conditions among predictor variables on the estimated parameters obtained from logistic regression analyses. Sparse data conditions are defined in this study as situations in which the distribution of one or more binary (0/1) predictors reflects very low frequency for one of the two possible values, either $\mathrm{p}(x_i = 1) = 0.05$ or $\mathrm{p}(x_i = 1) = 0.10$.

Methods

To explore the role of sparse data conditions among predictor variables in binary logistic regression, we simulated a series of data sets, where each data set consisted of a single, binary (0/1) outcome variable, and one or more predictor variables. Depending upon the specific simulation condition, the predictor variables consisted of either: (1) one or more binary (0/1) variables, where one of the two data values occurred with low frequency (i.e., were sparse); or (2) a combination of one or more sparse binary variables in combination with a normally-distributed continuous predictor (see Table 2 for the complete set of data conditions). For each data condition, the distribution of the binary dependent variable was non-sparse and uniform (proportion of “successes” ≈ proportion of “non-successes” ≈ .50). Data were generated using a data generation process with underlying intercept and slope population parameters of $\beta_0 = .20$ and $\beta_1 = \beta_2 = 0.50$, respectively, and using sample sizes of $N = 100$, 200, or 500. Regression weights were estimated using both (1) maximum likelihood estimation (*ML*); and (2) the Firth (1993) penalized maximum likelihood estimation procedure (ML_F). To eliminate variation due to sampling error, the same simulated data set was used within each simulation condition (i.e., in each condition, *ML* and ML_F were fitted to the same data), while simulated data were allowed to vary randomly across conditions. The distributions of estimated slope estimates then were examined, confidence intervals for each computed, and coverage probabilities (i.e., the proportion of intervals that contained the true regression parameters, $\beta_0 = 0.20$ and $\beta_1 = 0.50$) determined. Additionally, for each estimated regression coefficient, two indices were computed to assess bias: (1) absolute bias, computed as $AB = \left|\hat{\beta}_i - \beta_i\right|$; and (2) mean squared error, computed as $MSE = \left(\hat{\beta}_i - \beta_i\right)^2$. Although both statistics tend to produce similar patterns of results, *MSE* offers a better balance between bias and efficiency (Carsey & Harden, 2013). All analyses were carried out using R (version 3.5.1).

Table 2. Data simulation conditions

Data condition	Sample size (*N*)	Number of binary predictors	Number of continuous predictors	Distribution of binary (0/1) predictor(s) i.e., p(*x*)=1
1	100	1	0	5%
2	100	1	0	10%
3	100	1	1	5%
4	100	1	1	10%
5	100	2	0	5%
6	100	2	0	10%
7	200	1	0	5%
8	200	1	0	10%
9	200	1	1	5%
10	200	1	1	10%
11	200	2	0	5%
12	200	2	0	10%
13	500	1	0	5%
14	500	1	0	10%
15	500	1	1	5%
16	500	1	1	10%
17	500	2	0	5%
18	500	2	0	10%

Results

Shown in Table 3 are descriptive statistics for the estimated regression parameters from the logistic regression model fitted to data of size *N* = 100, using one binary predictor with sparseness = 5% (i.e., the first simulation condition). When results for the model fitted using maximum likelihood (*ML*) are compared to the results using the Firth penalized maximum likelihood (*ML*F), *ML* estimation resulted in a number of instances in which the slope was severely overestimated ($\hat{\beta}_1 > 15$, see Figure 1). In contrast, *ML*F estimation resulted in much more consistent estimation of b_1 than *ML* estimation, with less bias as indicated by both bias indices [E(*AB*) = 0.72 and E(*MSE*) = 0.86 for *ML*F estimation vs. E(*AB*) = 4.09 and E(*MSE*) = 89.87 for *ML* estimation]. The observed coverage probability of the computed 95% confidence intervals estimating β_1 (based on 10,000 replicated samples) were .991 when using *ML* estimation and .996 when using *ML*F estimation. Because 95% confidence intervals were constructed, these probabilities would be expected to equal .95 in unbiased estimation. Thus, the standard error of β_1 appears to have been underestimated with both *ML* and *ML*F, although to a slightly lesser extent with the *ML* than with *ML*F.

Table 3. Descriptive Statistics for estimated regression parameters ($\hat{\beta}_0$ and $\hat{\beta}_1$) from binary logistic regression model fitted to simulated data (*N* = 100) with one binary predictor with sparseness = 5%

	$\hat{\beta}_0$					
Estimation method	***M***	***Med***	***SD***	**95% CI**	**E(*AB*)**	**E(*MSE*)**
ML	0.200	0.190	0.209	(0.196, 0.204)	0.167	0.044
*ML*F	0.198	0.188	0.207	(0.195, 0.203)	0.165	0.043

	$\hat{\beta}_1$					
Estimation method	***M***	***Med***	***SD***	**95% CI**	***AB***	***MSE***
ML	2.403	0.469	5.579	(1.310, 3.497)	2.745	34.749
*ML*F	0.482	0.399	0.928	(0.481, 0.484)	0.734	0.861

Note: Simulations based on 10,000 replicated samples; *ML* = maximum likelihood estimation, *ML*F = Firth penalized maximum likelihood estimation; true population parameters from generative model are β_0 = 0.20 and β_1 = 0.50; E(*AB*) = mean absolute bias = $\mathrm{E}\left[\left|\hat{\beta}_i - \beta_i\right|\right]$; E(*MSE*) = mean of the mean squared error = $\mathrm{E}\left[\left(\hat{\beta}_i - \beta_i\right)^2\right]$

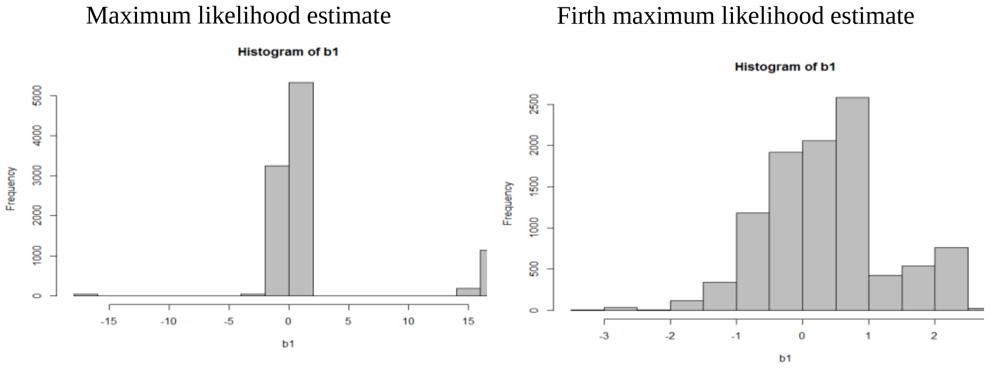

Figure 1. Distribution of estimated regression slope parameter $\left(\hat{\beta}_1\right)$ from binary logistic regression model fitted to simulated data (*N* = 100) with one binary predictor with sparseness = 5%; simulation based on 10,000 replicated samples

Shown in Figures 2-4 are the mean estimates of the regression slope parameters (β_1 and β_2) for each of the experimental conditions described previously, based on simulated samples and using 10,000 replications. As is seen in these figures, in each condition ML_F estimation resulted in estimates of the parameters that were closer to the actual parameter values ($\beta_1 = \beta_2 = 0.5$) than were the *ML* estimates. That is, mean levels of bias as reflected by absolute bias [E(*AB*), Figures 5-7] and *MSE* (Figures 8-10) were lower when using ML_F estimation than when using *ML* estimation. For both estimation methods, as the sample size used in the regression increased, the observed level of bias decreased. Also, as the sample size increased, the difference in bias between the two estimation methods decreased. In the largest sample size condition ($n = 500$), both estimation methods showed little bias and also very little difference in bias. This suggests that the critical issue as it pertains to biased parameter estimates in the presence of sparse predictors is not the level of sparseness, but rather the absolute frequency of the sparse event. That is, 5 occurrences of a particular value of a binary predictor in a sample of $n = 100$ leads to more severe bias in the regression slopes than does 25 occurrences of a particular value of a binary predictor in a sample of $n = 500$.

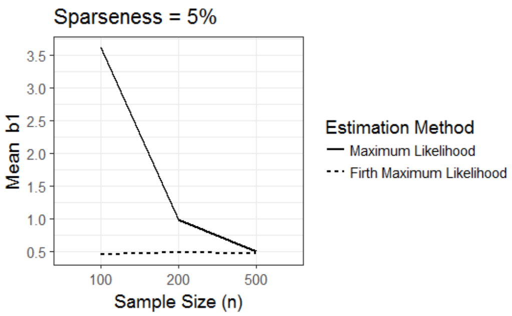

Figure 2. Estimated values of β_1 for maximum likelihood and Firth maximum likelihood logistic regression models fitted using one binary predictor with either 5% sparseness or 10% sparseness; actual population value of β_1 is 0.50

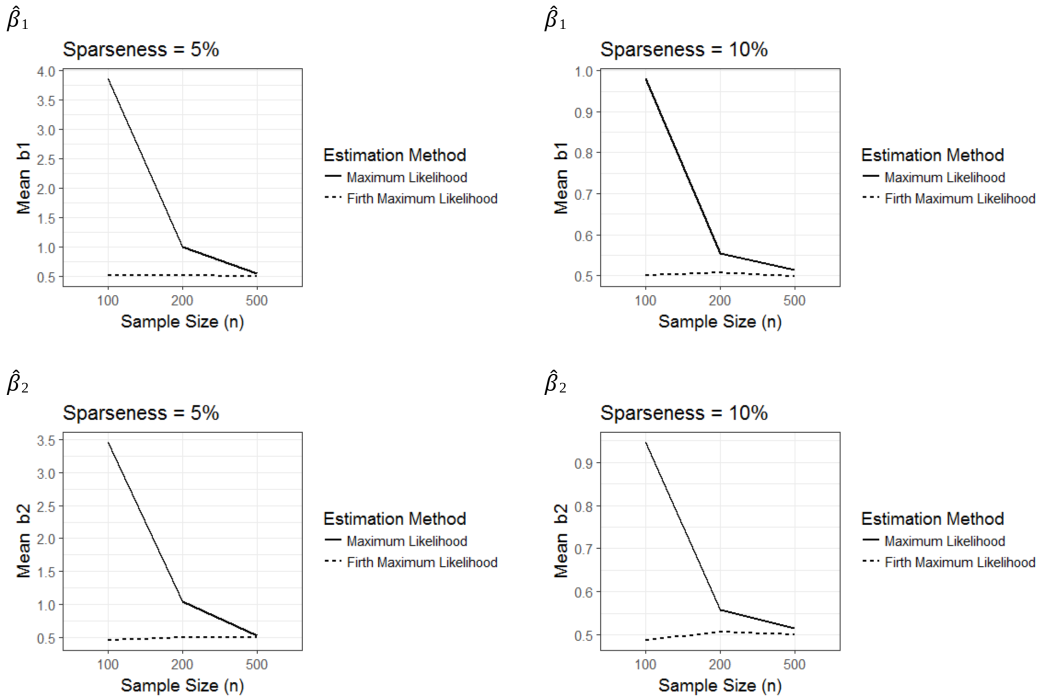

Figure 3. Estimated values of β_1 and β_2 for maximum likelihood and Firth maximum likelihood logistic regression models fitted using two binary predictors with either 5% sparseness or 10% sparseness; actual population values of β_1 and β_2 are 0.50

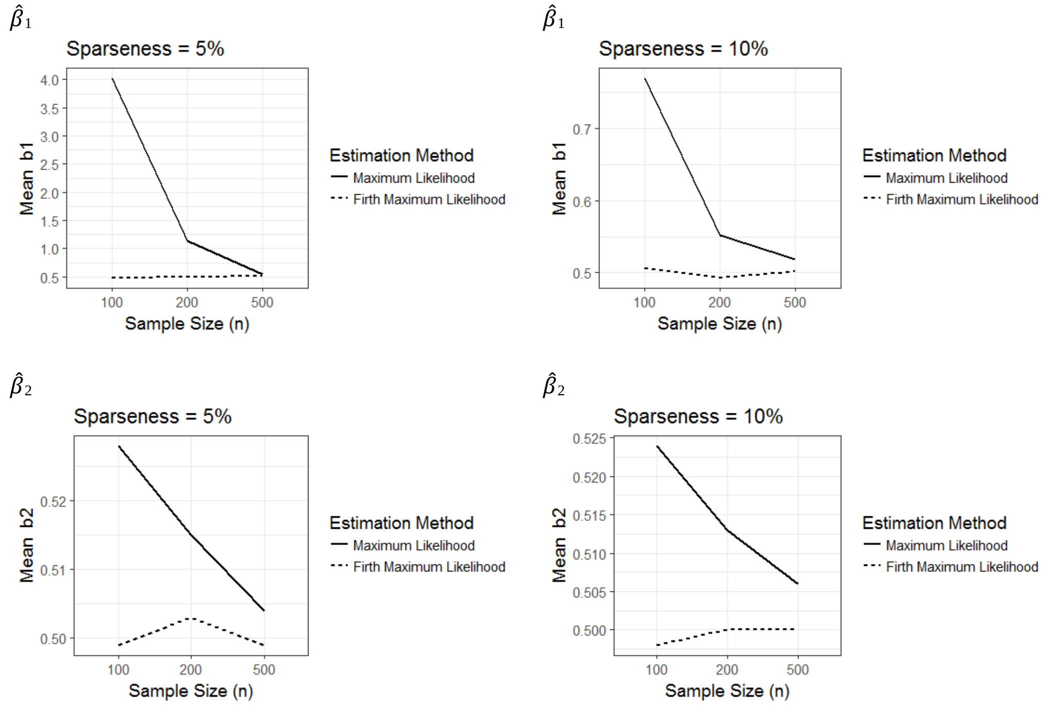

Figure 4. Estimated values of β_1 and β_2 for maximum likelihood and Firth maximum likelihood logistic regression models fitted using one binary predictor (β_1) and one continuous predictor (β_2) with either 5% sparseness or 10% sparseness; actual population values of β_1 and β_2 are 0.50

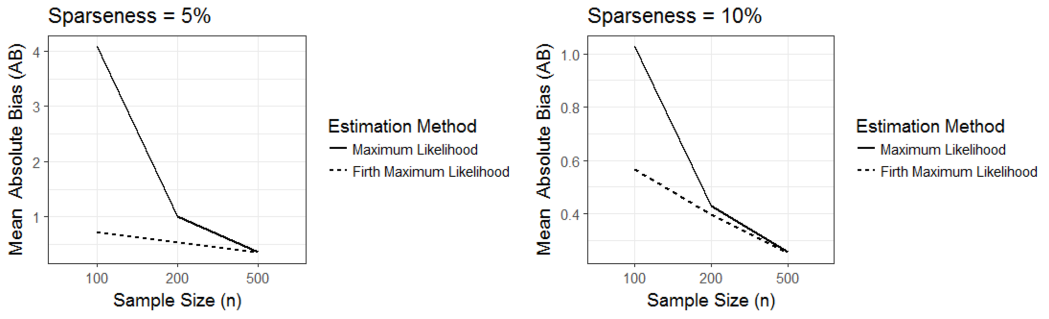

Figure 5. Estimated values of absolute bias of $\hat{\beta}_1$ for maximum likelihood and Firth maximum likelihood logistic regression models fitted using one binary predictor with either 5% sparseness or 10% sparseness

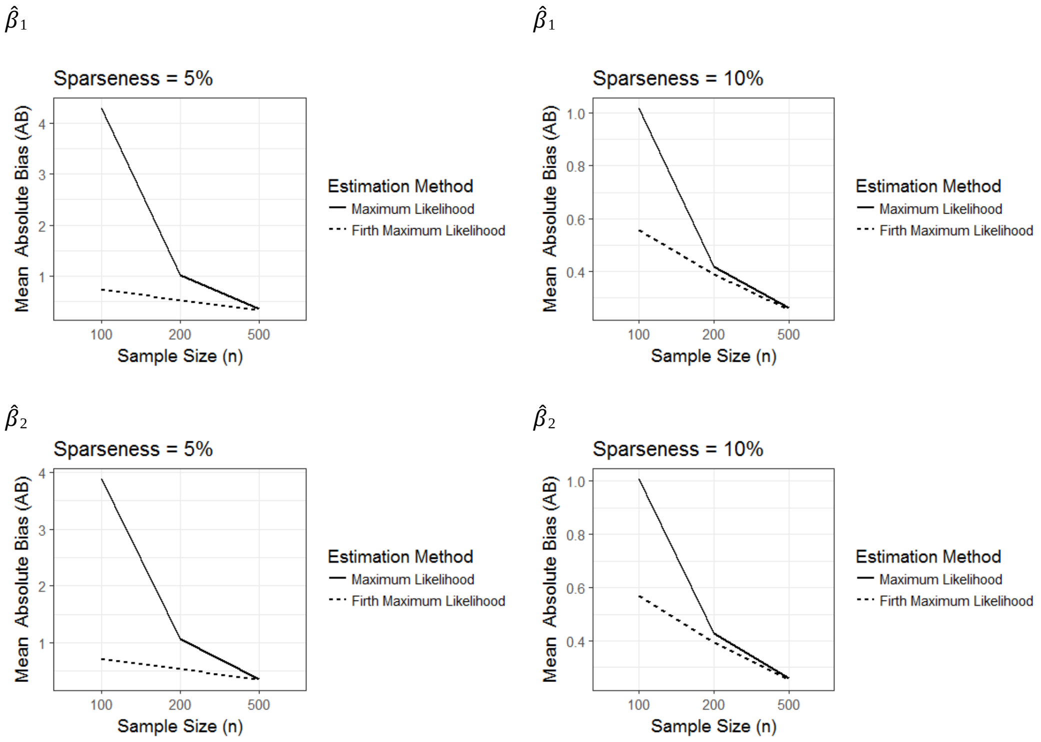

Figure 6. Estimated values of absolute bias of $\hat{\beta}_1$ and $\hat{\beta}_2$ for maximum likelihood and Firth maximum likelihood logistic regression models fitted using two binary predictors with either 5% sparseness or 10% sparseness

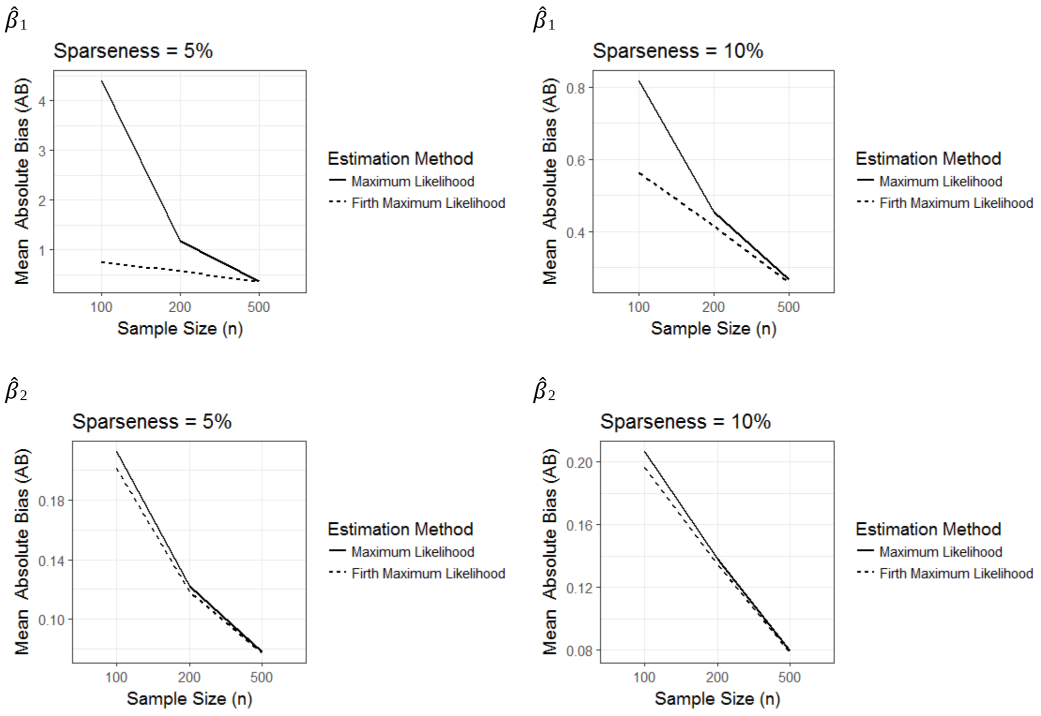

Figure 7. Estimated values of absolute bias of $\hat{\beta}_1$ and $\hat{\beta}_2$ for maximum likelihood and Firth maximum likelihood logistic regression models fitted using one binary predictor $\left(\hat{\beta}_1\right)$ and one continuous predictor $\left(\hat{\beta}_2\right)$ with either 5% sparseness or 10% sparseness

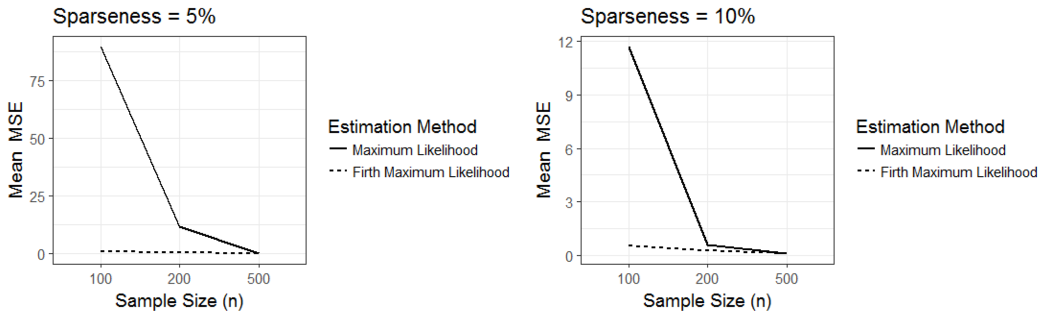

Figure 8. Estimated values of mean square error (*MSE*) of $\hat{\beta}_1$ for maximum likelihood and Firth maximum likelihood logistic regression models fitted using one binary predictor with either 5% sparseness or 10% sparseness

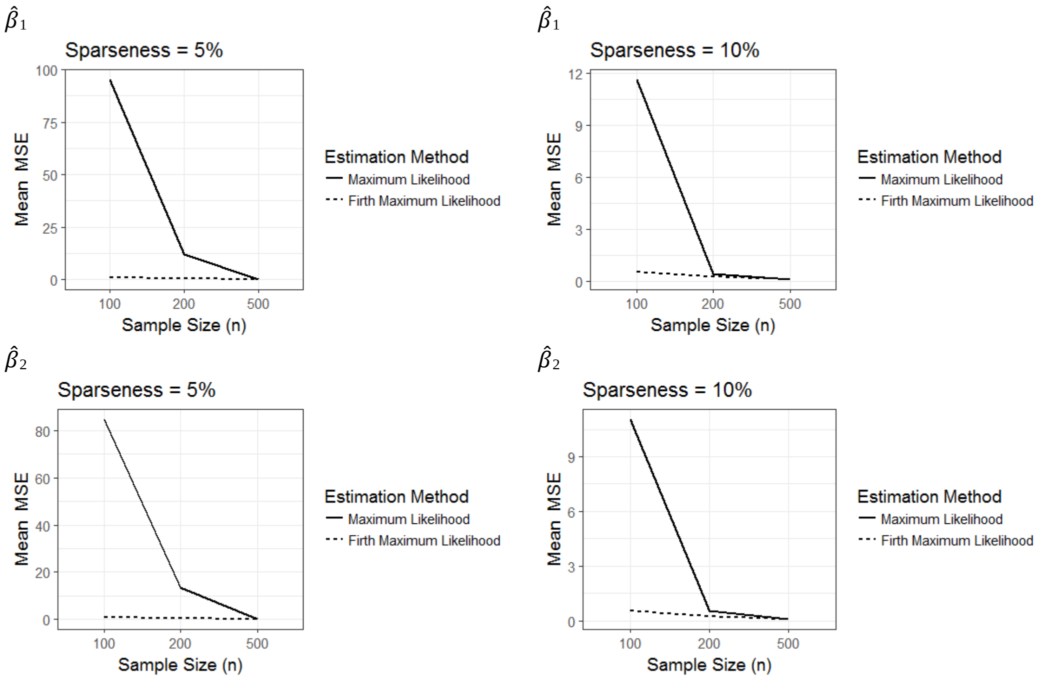

Figure 9. Estimated values of mean square error (*MSE*) of $\hat{\beta}_1$ and $\hat{\beta}_2$ for maximum likelihood and Firth maximum likelihood logistic regression models fitted using two binary predictors with either 5% sparseness or 10% sparseness.

In the experimental conditions involving one sparse binary predictor and one continuous, normally distributed predictor, estimates for the effect of the continuous predictor were more biased using ML_F estimation than when using *ML* estimation under the small (n = 100) and medium (n = 200) sample size conditions, but the difference was slight (see Figures 4, 7, and 10) and much less than the ML_F vs *ML* bias distinction in the estimate of the effect of the binary predictor.

In all experimental conditions, the level of sparseness had some effect on the bias of estimates. With a binary predictor in the model that occurred less frequently (5% of cases), the effect of sample size on reducing bias of the *ML* estimator was more immediate than in a data condition where the binary predictor appeared more frequently (10% of cases), with the difference in bias between the two estimation methods decreasing more rapidly as the sample size increased.

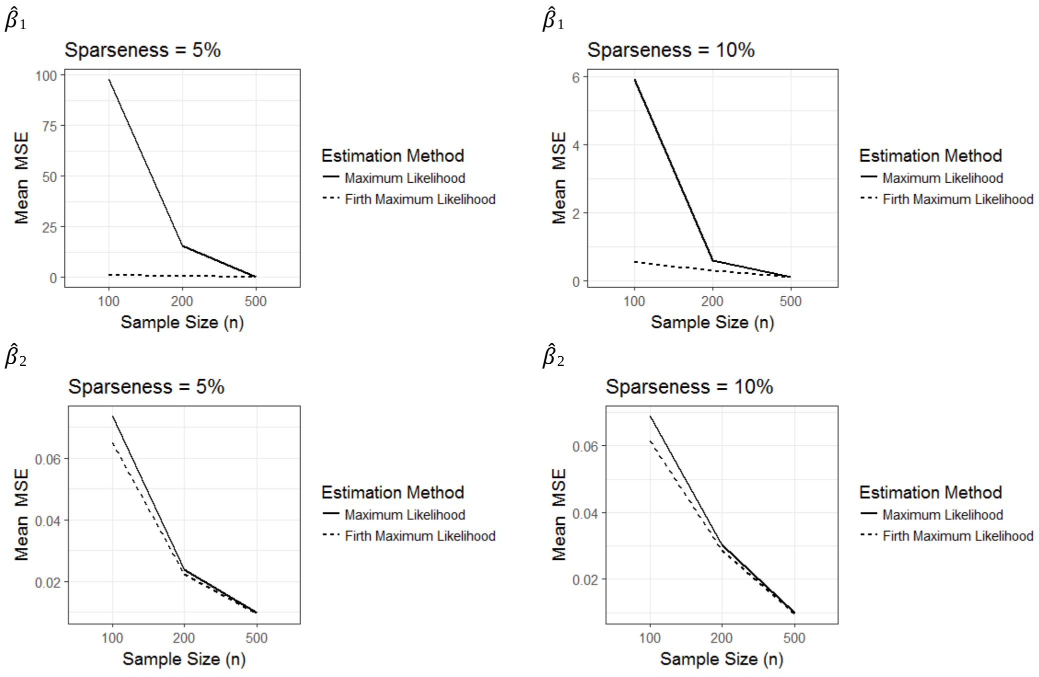

Figure 10. Estimated values of mean square error (*MSE*) of $\hat{\beta}_1$ and $\hat{\beta}_2$ for maximum likelihood and Firth maximum likelihood logistic regression models fitted using one binary predictor $\left(\hat{\beta}_1\right)$ and one continuous predictor $\left(\hat{\beta}_2\right)$ with either 5% sparseness or 10% sparseness

When coverage probabilities for the 95% confidence intervals of the regression slopes were examined, the results (Figures 11-13) showed that, for both *ML* and *ML*F estimation, smaller sample sizes resulted in coverage probabilities that were larger than the expected 95%. That is, in these situations, the standard errors of the regression coefficients appear to have been overestimated. Additionally, in each experimental condition, coverage probabilities using *ML*F estimation were slightly higher than the coverage probabilities that resulted using *ML* estimation.

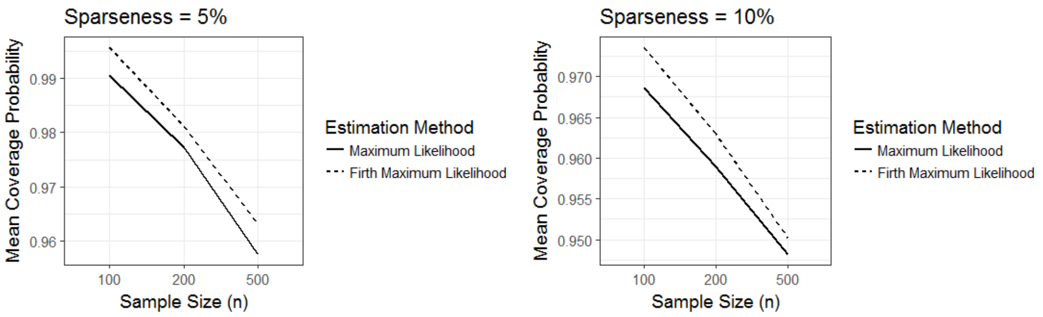

Figure 11. Estimated values of mean coverage probability (*CP*) for 95% confidence intervals for β_1 for maximum likelihood and Firth maximum likelihood logistic regression models fitted using one binary predictor with either 5% sparseness or 10% sparseness

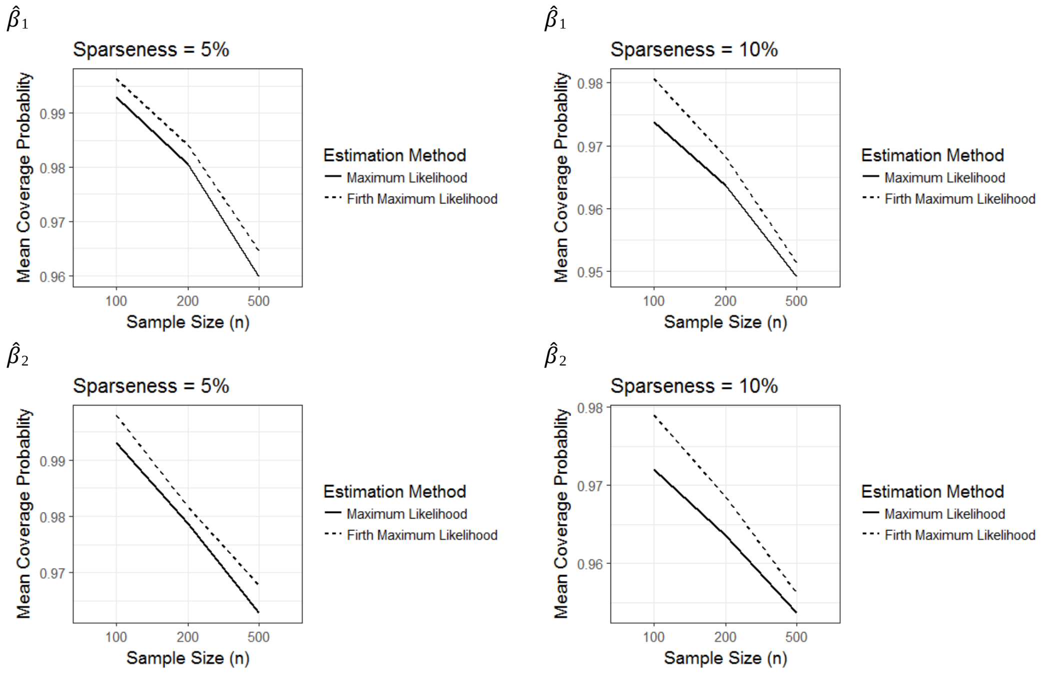

Figure 12. Estimated values of mean coverage probability (*CP*) for 95% confidence intervals for β_1 and β_2 for maximum likelihood and Firth maximum likelihood logistic regression models fitted using two binary predictors with either 5% sparseness or 10% sparseness

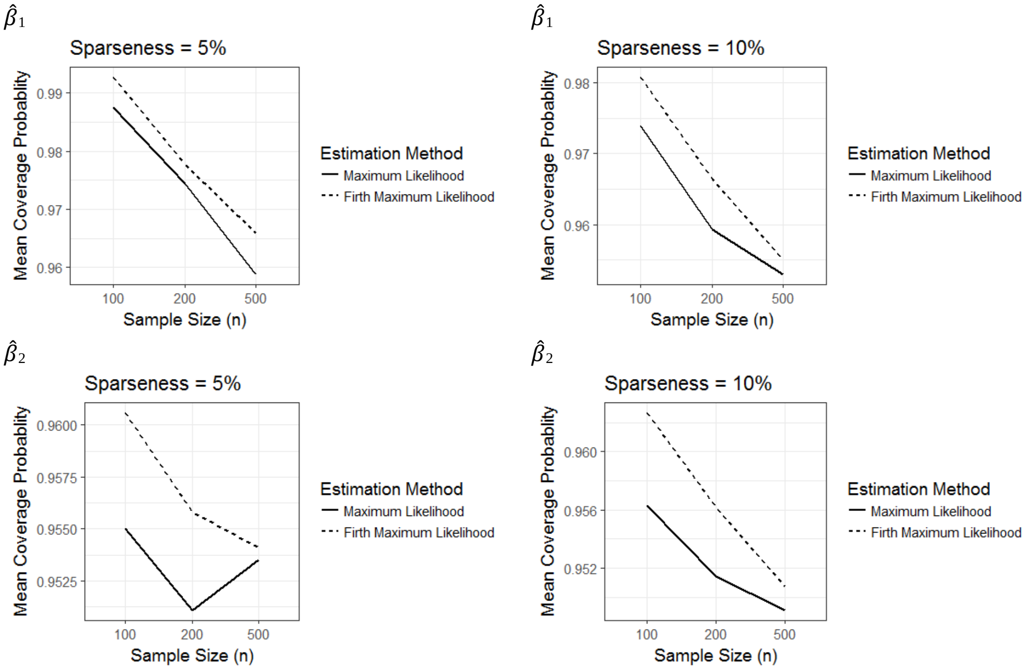

Figure 13. Estimated values of mean coverage probability (*CP*) for 95% confidence intervals for β_1 and β_2 for maximum likelihood and Firth maximum likelihood logistic regression models fitted using one binary predictor (β_1) and one continuous predictor (β_2) with either 5% sparseness or 10% sparseness

Conclusion

The use of binary logistic regression is ubiquitous in education and the social sciences. As it occurs, researchers carrying out cross-sectional, observational studies have little, if any control, over the distributional characteristics of the data they collect. As such, sparse data situations can arise in many instances. The present research seeks to provide insight into how such data sparseness among predictor variables might affect inferences made from logistic regression, as well as to evaluate an estimation technique that might address potential biases resulting from these data situations. The results from the simulations carried out in this study suggest that, when a sparse binary predictor is used with a relatively small sample size (n = 100), large bias occurs in the typically-employed *ML* estimates of slope parameters. However, in these situations the ML_F estimator of these parameters markedly reduces bias. Reductions in bias, although on a smaller scale, are evident when using ML_F estimation with somewhat larger sample sizes (n = 200). The advantages of ML_F estimation become minimal with large sample sizes ($n = 500$). Thus, it appears that bias in these conditions is affected by the absolute frequency of the sparse event(s), more so than by the relative frequency. A corresponding recommendation to researchers who encounter sparseness of binary predictors is to use ML_F estimation rather than ML estimation with sample sizes less than or equal to 200.

Interestingly, when a normally-distributed, continuous predictor was included in a model together with a sparse binary predictor, bias in the effect of the continuous predictor also was apparent when using ML estimation with small sample sizes, and this bias was reduced slightly when using the ML_F estimator. Thus, it appears that the biasing effects of sparse binary predictors may extend to the effects of other non-binary predictors in the model. Future research might consider examining situations with polytomous categorical predictors with sparseness in one or more categories, how this affects the estimated parameters, and how potential bias might be addressed. Perhaps similar approaches also might be proposed for continuous predictors that are badly skewed (e.g., zero-inflated) and are producing problems in estimation.

Although the effects of sparseness on parameter estimates are well-known when sparseness of the outcome variables is considered, very little research has considered the effects of sparseness among predictor variables. The present research begins this inquiry. Additional research might explore a wider variety of data conditions, including other sparseness levels, more varied sample sizes, and larger numbers of predictors. Another avenue of research could explore the effects of sparse predictors on other regression models such as ordinal regression. Lipsitz et al. (2013), for example, propose a bias-correction procedure that can be employed in proportional odds logistic regression for ordinal outcomes. Perhaps an estimation technique such as this might address potential biases introduced by sparse predictors.

Future research also might examine how joint sparseness in both the predictors and the outcome may impact inferences, and how techniques such as the Firth procedure might be used to address these situations. Additionally, at a practical level, it is recommended that researchers employing logistic regression screen their data for sparseness—both in the outcome variable(s) as well as the predictors. If sparseness is evident, the Firth procedure may be effective in alleviating either source of bias.

References

Carsey, T. M., & Harden, J. J. (2013). *Monte Carlo simulation and resampling methods for social science*. Thousand Oaks, CA: Sage Publications. doi: 10.4135/9781483319605

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika, 80*(1), 27-38. doi: 10.2307/2336755

Hair, J. F., Ringle, C. M., & Sarstedt, M. (2011). PLS-SEM: Indeed a silver bullet. *Journal of Marketing Theory and Practice, 19*(2), 139-152. doi: 10.2753/MTP1069-6679190202

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis, 9*(2), 137-163. doi: 10.1093/oxfordjournals.pan.a004868

Lipsitz, S. R., Fitzmaurice, G. M., Regenbogen, S. E., Sinha, D., Ibrahim, J. G., & Gawande, A. A. (2013). Bias correction for the proportional odds logistic regression model with application to a study of surgical complications. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 62*(2), 233-250. doi: 10.1111/j.1467-9876.2012.01057.x

Manski, C. F., & Lerman, S. R. (1977). The estimation of choice probabilities from choice based samples. *Econometrica, 45*(8), 1977-1988. doi: 10.2307/1914121

Prentice, R. L., & Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika, 66*(3), 403-411. doi: 10.1093/biomet/66.3.403

UCLA Statistical Consulting Group. (2017). *FAQ: What is complete or quasi-complete separation in logistic/probit regression and how do we deal with them?* Retrieved from https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faqwhat-is-complete-or-quasi-complete-separation-in-logisticprobit-regression-and-how-do-we-deal-with-them/

van Smeden, M., de Groot, J. A. H., Moons, K. G. M., Collins, G. S., Altman, D. G., Eijkemans, M. J. C., & Reitsma, J. B. (2016). No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Medical Research Methodology, 16*, 163. doi: 10.1186/s12874-016-0267-3

Vittinghoff, E., & McCulloch, C. E., (2007). Relaxing the rule of ten events per variable in logistic and Cox regression. *American Journal of Epidemiology, 165*(6), 710-718. 10.1093/aje/kwk052