Logistic Regression with Small Samples

(Nearly) Unbiased Estimation with Penalized Maximum Likelihood*

Kelly McCaskey†

Carlisle Rainey‡

Abstract

In small samples, maximum likelihood estimates of logistic regression coefficients are substantially biased away from zero. This bias might be 25 percent or more in plausible scenarios. As a solution, we (re)introduce political scientists to Firth's (1993) penalty, which removes much of the bias from the usual estimator. We use Monte Carlo simulations to illustrate that penalized maximum likelihood estimation eliminates most of the bias, but also (perhaps more importantly) greatly reduces the variance of the estimate. We illustrate the substantive importance of the penalized estimator with a replication of Weisiger (2014).

*We thank Alex Weisiger for making his data available to us. We thank Chris Zorn and participants at the 2015 Annual Meeting of the Society for Political Methodology for helpful comments. We conducted these analyses analyses with R 3.1.0. All data and computer code necessary for replication are available at github.com/kellymccaskey/small.

†Kelly McCaskey is a Ph.D. student in the Department of Political Science, Texas A&M University, 2010 Allen Building, College Station, TX, 77843 (kellymccaskey@tamu.edu).

‡Carlisle Rainey is Assistant Professor of Political Science, Texas A&M University, 2010 Allen Building, College Station, TX, 77843 (crainey@tamu.edu).

The Problem

When modeling a binary outcome, the researcher might use logistic regression and model the probability of an event as $\Pr(y_i) = \Pr(y_i = 1 \mid X_i) = \frac{1}{1 + e^{-X_i\beta}}$, where y represents a vector of binary outcomes, X represents a matrix of explanatory variables and an intercept, and β represents a vector of model coefficients. Using this model, it is straightforward to derive the likelihood function

$$\Pr(y|\beta) = L(\beta|y) = \prod_{i=1}^{n} \left[\left(\frac{1}{1 + e^{-X_i\beta}} \right)^{y_i} \left(1 - \frac{1}{1 + e^{-X_i\beta}} \right)^{1-y_i} \right].$$

As usual, one can take the natural logarithm of both sides to obtain the log-likelihood function

$$\log L(\beta|y) = \sum_{i=1}^{n} \left[y_i \log \left(\frac{1}{1 + e^{-X_i\beta}} \right) + (1 - y_i) \log \left(1 - \frac{1}{1 + e^{-X_i\beta}} \right) \right].$$

The researcher can obtain the maximum likelihood (ML) estimate $\hat{\beta}^{mle}$ by finding the vector β that maximizes $\log L$ (King 1998).

Asymptotically, the ML estimator for the logistic regression coefficient vector $\hat{\beta}^{mle}$ is centered at the true value β^{true}, so that $E(\hat{\beta}^{mle}) \approx \beta^{true}$ when the sample is large (Wooldridge 2002, pp. 391-395, and Casella and Berger 2002, p. 470). For small samples, though, the asymptotic approximation does not work well–the ML estimates are biased substantially away from zero (Long 1997, pp. 53-54).

Offering a rough heuristic about appropriate sample sizes, (Long 1997, p. 54) writes: "It is risky to use ML with samples smaller than 100, while samples larger than 500 seem adequate." This presents the researcher with a problem: When dealing with small samples, how can she obtain reasonable estimates of logistic regression coefficients?

A Solution

The statistics literature offers a simple solution to the problem of bias. Firth (1993) suggests penalizing the usual likelihood function $L(\beta|y)$ by a factor equal to the square root of the determinant of the information matrix $|I(\beta)|^{\frac{1}{2}}$, which produces a "penalized" likelihood function $L^*(\beta|y) = L(\beta|y)|I(\beta)|^{\frac{1}{2}}$ (see also Kosmidis and Firth 2009 and Kosmidis 2014). It turns out that this penalty is equivalent to Jeffreys' (1946) prior for the logistic regression model (Firth 1993 and

Poirier 1994). Taking logs yields the penalized log-likelihood function.

$$\log L^*(\beta|y) = \sum_{i=1}^{n} \left[y_i \log \left(\frac{1}{1+e^{-X_i\beta}} \right) + (1-y_i) \log \left(1 - \frac{1}{1+e^{-X_i\beta}} \right) \right] + \frac{1}{2} \log |I(\beta)|.$$

Then the researcher can obtain the *penalized* maximum likelihood (PML) estimate $\hat{\beta}^{pmle}$ by finding the vector β that maximizes $\log L^*$. Zorn (2005) suggested Firth's penalty for solving the problem of separation (see also Rainey 2015), but the broader application to small sample problems seems to have gone unnoticed in political science.

A researcher can implement PML just as easily as ML, but PML estimates of logistic regression coefficients are both less biased (Firth 1993) and more efficient (Kosmidis 2007, p. 49) than ML estimates.[1] This is important to note. When choosing between estimators, researchers often face a tradeoff between bias and efficiency (Hastie, Tibshirani, and Friedman 2013, pp. 37-38). But there is no bias-variance tradeoff between ML and PML estimators. The PML estimator exhibits both lower bias *and* lower variance.

Simulations

To demonstrate the substantial improvements in both bias and variance for small sample sizes sometimes found in political science research, we conduct a Monte Carlo simulation comparing the sampling distributions of the ML and PML estimates. These simulations demonstrate two features of the ML and PML estimators:

1. The ML estimator can be quite biased in small samples, sometimes more than 50%. The PML is nearly unbiased, regardless of sample size.
2. The variance of the ML estimator can be three or more times as large as the PML estimator.

Of course, the increased bias and variance of the ML estimator means that the PML estimator will also have a much smaller mean squared error.

In our simulation, the true data generating process is always $\Pr(y_i = 1) = \frac{1}{1+e^{-X_i\beta}}$, where $i \in 1, 2, ..., n$ and $X_i\beta = \beta_{cons} + 0.5x_1 + \sum_{j=2}^{k} 0.2x_j$. We consider the coefficient for x_1 as the coefficient of interest. Each fixed x_j is drawn from a normal distribution with mean of zero and standard deviation of one. The simulation varies the sample size n from 30 to 210, the number of explanatory variables k from 3 to 6 to 9, and the the intercept β_{cons} from -1 to -0.5 to 0 (which, in turn, varies the proportion of events from about 28% to 38% to 50%). We simulate 10,000

[1] Further, the penalized maximum likelihood estimates are easy to calculate in R using the `logistf` or `brglm` packages. See the Section A of the Appendix for an example.

data sets for each combination of the simulation parameters and estimate the logistic regression coefficients using ML and PML for each data set. We calculate mean and variance of each set of 10,000 coefficient estimates to estimate the mean and variance of the sampling distribution for each estimator.

Bias

We calculate the percent bias = $100 \times \left(\frac{E(\hat{\beta})}{\beta^{true}} - 1 \right)$ as the sample size, the intercept (i.e., proportion of events), and number of explanatory variables vary. Figure 1 shows the results. The sample size varies across the x-axis of each plot and each panel shows a distinct combination of intercept and number of variables in the model. Across the range of the parameters of our sample, the bias of the MLE varies from about 120% (intercept equal to -1, 9 predictors, and 30 observations) to around 2% (intercept equal to zero, 3 predictors, and 210 observations). The bias in the PMLE, on the other hand, is much smaller. For the worst-case scenario with nine variables, 30 observations, and an intercept of -1 (about 11 events), the percent bias in the PMLE is only about seven percent.[2]

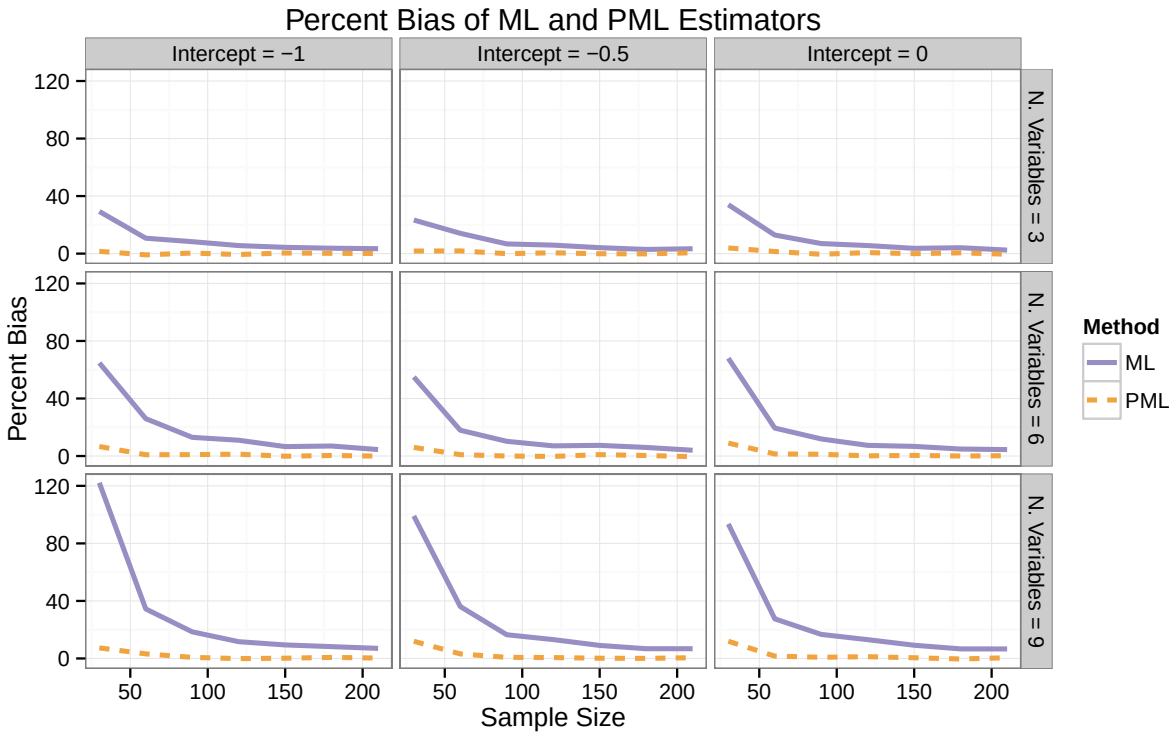

Figure 1: This figure illustrates the substantial bias of $\hat{\beta}^{mle}$ and the near unbiasedness of $\hat{\beta}^{pmle}$.

[2]Figures 5 and 6 in Section B of the Appendix show the expected value and (absolute) bias of these estimates.

Variance

In many cases, estimators trade off bias and variance, but that is *not* the case for ML and PML. Figure 2 shows that, in addition to nearly eliminating the bias, PML also substantially reduces the variance of the estimator, especially for small sample sizes (less than about 75 observations in our simulations). For $N = 30$ and $\beta_{cons} = -1$ (about 28% events), the variance of the ML estimator is about 95%, 243%, and 610% larger than the PML estimator for 3, 6, and 9 variables, respectively. For $N = 60$, the variance is about 30%, 58% and 91% larger, respectively. For a larger sample of $N = 210$, the variance is still about 7%, 10%, and 14% larger for the ML estimator.

The smaller variance is perhaps more important than the reduced bias. Rather than focus on how close the averages of the estimates are to the true value (e.g., bias), a more important summary of estimator performance might be how far the estimates are from the true value, on average (e.g., mean squared error). The mean squared error (MSE) equals $\frac{1}{n}\sum_{i=1}^{n}(\hat{\beta} - \beta^{true})^2$ = variance + bias^2. The large differences in the variance and the relatively smaller differences in bias lead to substantial differences in the MSE of the two estimators.[3]

These simulation results show that Firth's (1993) bias correction is not trivial. In small samples, ML estimates of logistic regression coefficients are severely biased. The penalty identified by Firth nearly eliminates this bias, and, as a side effect, produces large gains in the efficiency of the estimates.

An Application

To illustrate the potential importance of using the nearly unbiased PML estimator in an application, we reanalyze a portion of the statistical analysis in Weisiger (2014). Weisiger describes how, after the official end of the war, violence sometimes continues in the form of guerrilla warfare. He argues that resistance is more likely when conditions are favorable for insurgency, such as difficult terrain, a occupying force, or a pre-war leader remains at-large in the country.

Weisiger's sample consists of 35 observations (with 14 insurgencies). We reanalyze Weisiger's data using logistic regression to show the substantial difference between the biased, high-variance ML estimates and the nearly unbiased, low-variance PML estimates. (Weisiger uses a linear probability model for in the original analysis, but that leads to probabilities that fall outside the [0, 1] interval.)[4] Prior to estimation, we standardize the continuous variables to have mean zero and

[3]Figures 7, 8, and 9 in Section B of the Appendix show the variance inflation, mean squared error, and mean squared error inflation of the estimates.

[4]Specifically, we reanalyze the Model 3 in Weisiger's Table 2 (p. 14). In the original analysis, Weisiger uses a

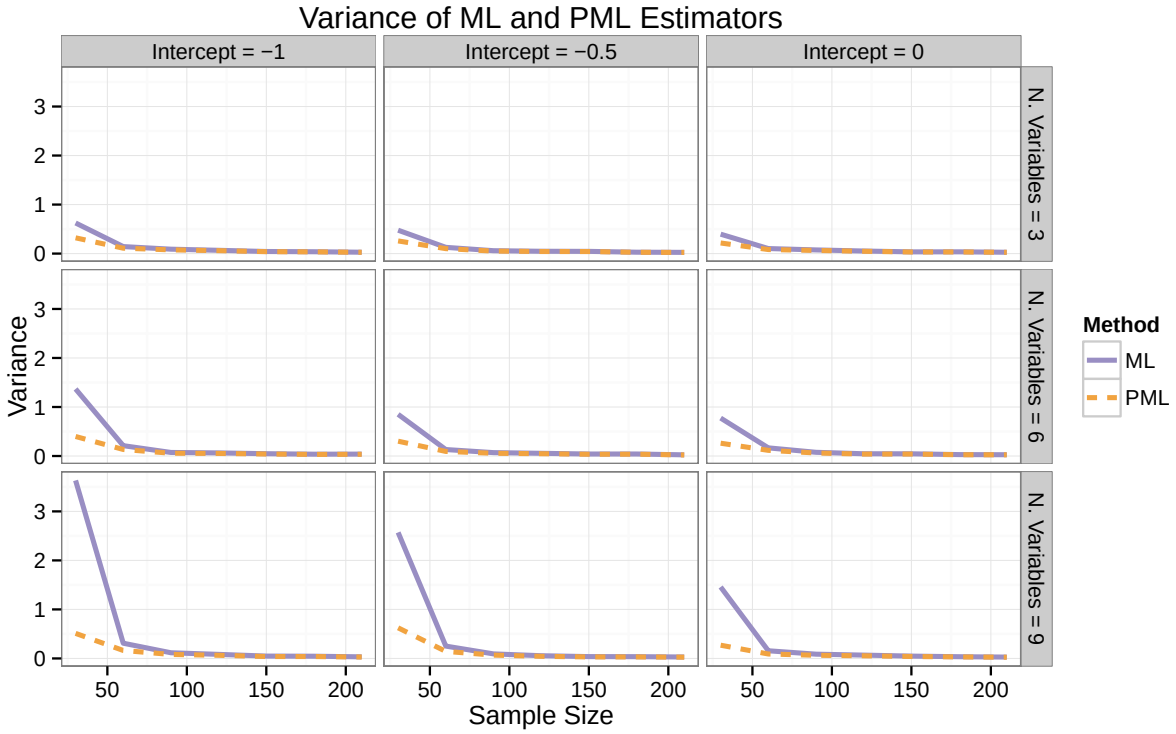

Figure 2: This figure illustrates the smaller variance of $\hat{\beta}^{pmle}$ compared to $\hat{\beta}^{mle}$.

standard deviation one-half and binary variables to have mean zero (Gelman 2008). Figure 3 shows the coefficient estimates and 90% confidence intervals using ML and PML. Notice that the PML estimates are substantially smaller in many cases. Although the coefficient for terrain only changes by 16%, each of the remaining coefficients changes by more than 45%! The coefficient for per capita GDP shrinks by more the 60% and the coefficient for occupying force density grows by nearly 100%. Also notice that the PML standard errors are much smaller–the ML estimates for the coefficients of a coordinating leader and for the intercapital distance fall outside the PML 90% confidence interval. On average, the PML confidence intervals are about half as wide as the ML confidence intervals.

Because we do not know the true model, we cannot know which of these sets of coefficient is better. However, we can use out-of-sample prediction to help adjudicate between these two methods.

linear probability model. He writes that "I [Weisiger] make use of a linear probability model, which avoids problems with separation but introduces the possibility of non-meaningful predicted probabilities outside the [0,1] range" (p. 11). As he notes, predictions outside the [0, 1] interval pose a problem for interpreting the linear probability model. In these data for example, the linear probability model estimates a probability of 1.41 of insurgency in one case. In another, it estimates a probability of -0.22. Overall, 25% of the estimated probabilities based on the linear probability model are larger than one or less than zero. Of course, these results are nonsense. However, because of the well-known small-sample bias, methodologists discourage researchers from using logistic regression with small samples. The PML approach, though, solves the problem of bias as well as nonsense predictions.

Figure 3: This figure shows the coefficients for a logistic regression model estimated explaining post-conflict guerrilla war estimated with ML and PML. Notice that the PML estimates tend to be much smaller than the ML estimates.

We use leave-one-out cross-validation and summarize the prediction errors using Brier and log scores, for which smaller values indicate better predictive ability.[5] The ML estimates produce a Brier score of 0.14, and the PML estimates lower the Brier score by 14% to 0.12. The ML estimates produce a log score of 0.58, while the PML estimates lower the log score by 34% to 0.38. The PML estimates outperform the ML estimates for both approaches to scoring, and this provides good evidence that the PML estimates better capture the data generating process.

Because we are using a logistic regression, we might be more interested in *functions* of the coefficients than in the coefficients themselves. For an example, we focus on Weisiger's hypothesis that there will be a greater chance of resistance when the pre-conflict political leader remains at large in the conquered country. Setting all other explanatory variables at their sample medians, we calculated the predicted probabilities, the first difference, and the risk ratio for the probability of a post-conflict guerrilla war as countries gain a coordinating leader. Figure 4 shows the estimates of the quantities of interest.

[5]The Brier score is calculated as $\sum_{i=1}^{n}(y_i - p_i)^2$, where i indexes the observations, $y_i \in \{0, 1\}$ represents the actual outcome, and $p_i \in (0, 1)$ represents the estimated probability that $y_i = 1$. The log score as $-\sum_{i=1}^{n} log(r_i)$, where $r_i = y_i p_i + (1 - y_i)(1 - p_i)$. Notice that because we are logging $r_i \in [0, 1]$, $\sum_{i=1}^{n} log(r_i)$ is always negative and smaller (i.e., more negative) values indicate worse fit. We choose to take the negative of $\sum_{i=1}^{n} log(r_i)$, so that, like the Brier score, larger values indicate a worse fit.

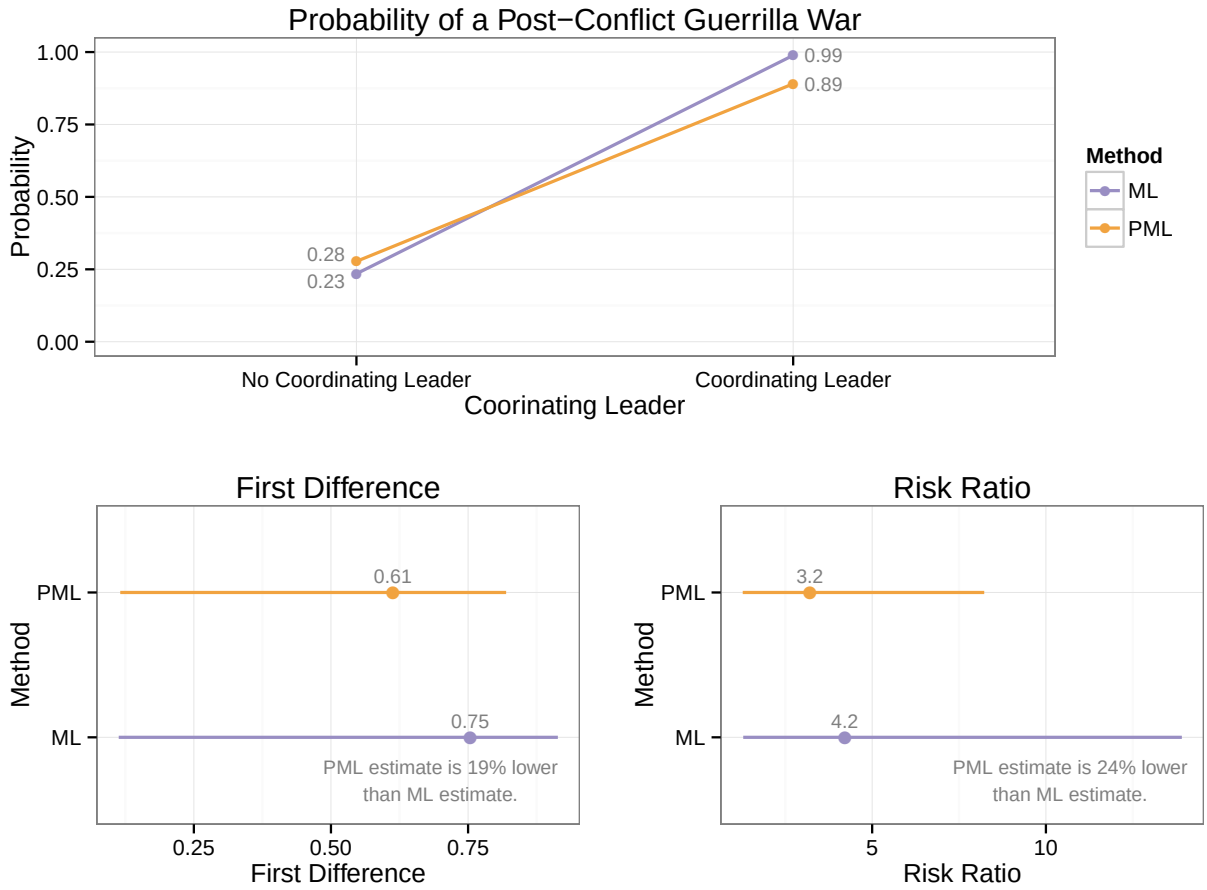

Figure 4: This figure shows the quantities of interest for the effect of a coordinating leader on the probability of a post-conflict guerrilla war.

PML pools the estimated probabilities toward zero, so that when a country lacks a coordinating leader, ML suggests a 23% chance of rebellion while PML suggests a 28% chance. On the other hand, when country *does have* a coordinating leader, ML suggests a 99% chance of rebellion of 0.99, but PML lowers this to 89%. Accordingly, PML suggests smaller effect sizes, whether using a first difference or risk ratio. PML shrinks the estimated first difference by 19% from 0.75 to 0.61 and the risk ratio by 24% from 4.2 to 3.2.

Conclusion

The substantial difference in these estimates demonstrates the potential importance of using PML when dealing with small samples and binary outcomes. In small samples, ML leads to large biases

in logistic regression coefficients. Further, these estimates are biased *away* from zero, leading researchers to conclude that variables have larger effects than they actually do. However, PML is nearly unbiased, regardless of sample size. And these reductions in bias do not come at a large cost. Indeed, the PML estimates exhibit less bias, smaller variance, and lower mean squared error than the ML estimates.

References

Casella, George, and Roger L. Berger. 2002. *Statistical Inference*. 2nd ed. Pacific Grove, California: Duxbury.

Firth, David. 1993. "Bias Reduction of Maximum Likelihood Estimates." *Biometrika* 80(1):27–38.

Gelman, Andrew. 2008. "Scaling Regression Inputs by Dividing by Two Standard Deviations." *Statistics in Medicine* 27(15):2865–2873.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2013. *The Elements of Statistical Learning*. Springer Series in Statistics second ed. New York: Springer.

Jeffreys, H. 1946. "An Invariant Form of the Prior Probability in Estimation Problems." *Proceedings of the Royal Society of London, Series A* 186(1007):453–461.

King, Gary. 1998. *Unifying Political Methodology: The Likelihood Theory of Statistical Inference*. Ann Arbor: Michigan University Press.

Kosmidis, Ioannis. 2007. Bias Reduction in Exponential Family Nonlinear Models PhD thesis University of Warwick.

Kosmidis, Ioannis. 2014. "Bias in Parametric Estimation: Reduction and Useful Side-Effects." *WIREs Computational Statistics* 6(3):185–196.

Kosmidis, Ioannis, and David Firth. 2009. "Bias Reduction in Exponential Family Nonlinear Models." *Biometrika* 96(4):793–804.

Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Advanced Quantitative Techniques in the Social Sciences Thousand Oaks, CA: Sage.

Poirier, Dale. 1994. "Jeffreys' Prior for Logit Models." *Journal of Econometrics* 63(2):327–339.

Rainey, Carlisle. 2015. "Dealing with Separation in Logistic Regression Models." Working paper. Latest version at https://github.com/carlislerainey/priors-for-separation.

Weisiger, Alex. 2014. "Victory Without Peace: Conquest, Insurgency, and War Termination." *Conflict Management and Peace Science* 31(4):357–382.

Wooldridge, Jeffrey M. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge: MIT Press.

Zorn, Christopher. 2005. "A Solution to Separation in Binary Response Models." *Political Analysis* 13(2):157–170.

Online Appendix

Logistic Regression in Small Samples

A Penalized Maximum Likelihood Estimation of Logistic Regression Models in R

This example code is available at https://github.com/kellymccaskey/small/blob/master/R/example.R.

```
# load data from web
library(readr)  # for read_csv()
weisiger <- read_csv("https://raw.githubusercontent.com/kellymccaskey/small/master/weisiger-replication,

# quick look at data
library(dplyr)  # for glimpse()
glimpse(weisiger)

# model formula
f <- resist ~ polity_conq + lndist + terrain +
  soldperterr + gdppc2 + coord

# ----------------------------- #
# pmle with the logistf package #
# ----------------------------- #

# estimate logistic regression with pmle
library(logistf)  # for logistf()
m1 <- logistf(f, data = weisiger)

# see coefficient estimates, confidence intervals, p-values, etc.
summary(m1)

# logistf does **NOT** work with texreg package
library(texreg)
screenreg(m1)

# see help file for more
help(logistf)

# --------------------------- #
# pmle with the brglm package #
# --------------------------- #

# estimate logistic regression with pmle
library(brglm)  # for brglm()
m2 <- brglm(f, family = binomial, data = weisiger)

# see coefficient estimates, standard errors, p-values, etc.
```

```
summary(m2)

# brglm works with texreg package
screenreg(m2)

# see help file for more
help(brglm)
```

B Additional Simulation Results

B.1 Expected Value

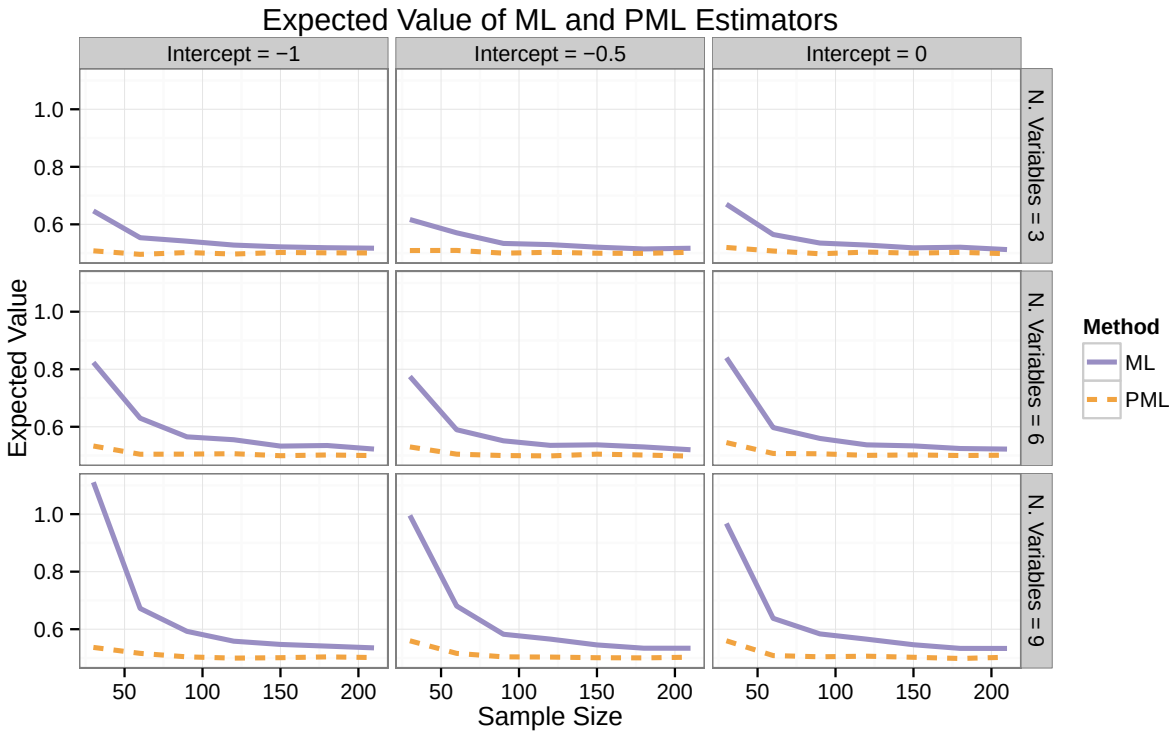

Figure 5: This figure shows the expected value of $\hat{\beta}^{mle}$ and $\hat{\beta}^{pmle}$. The true value is $\beta = 0.5$.

B.2 Bias

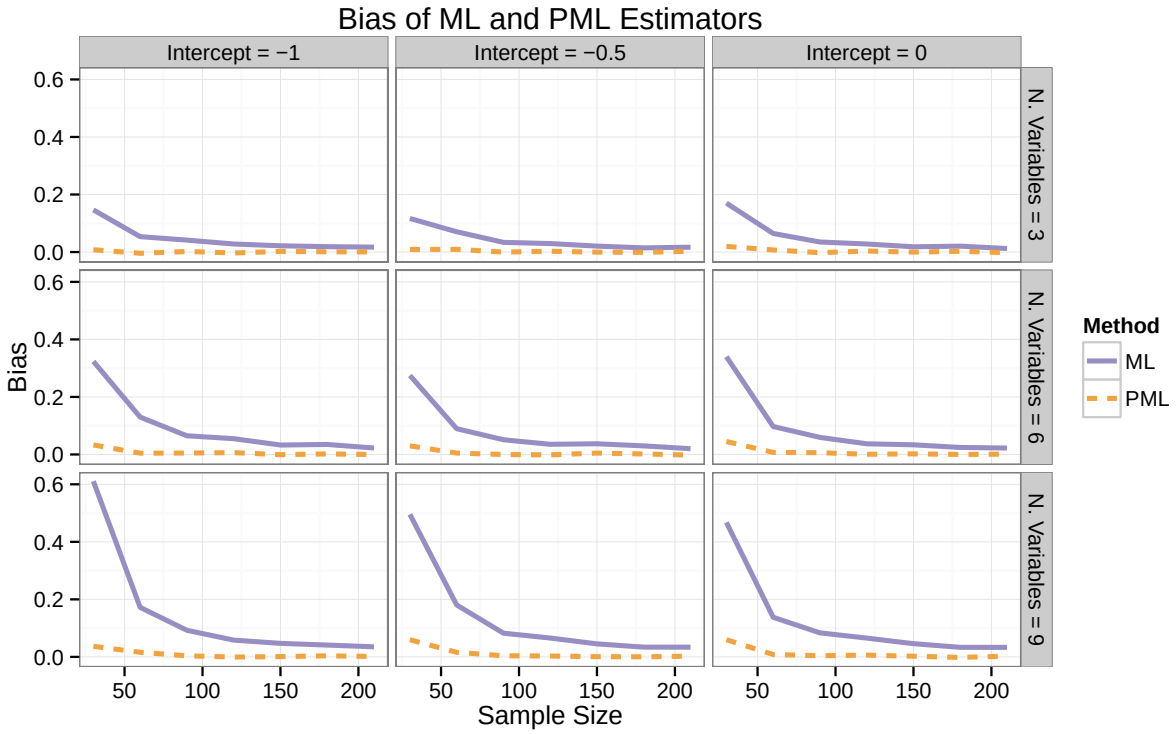

Figure 6: This figure shows the bias of $\hat{\beta}^{mle}$ and $\hat{\beta}^{pmle}$.

B.3 Variance Inflation

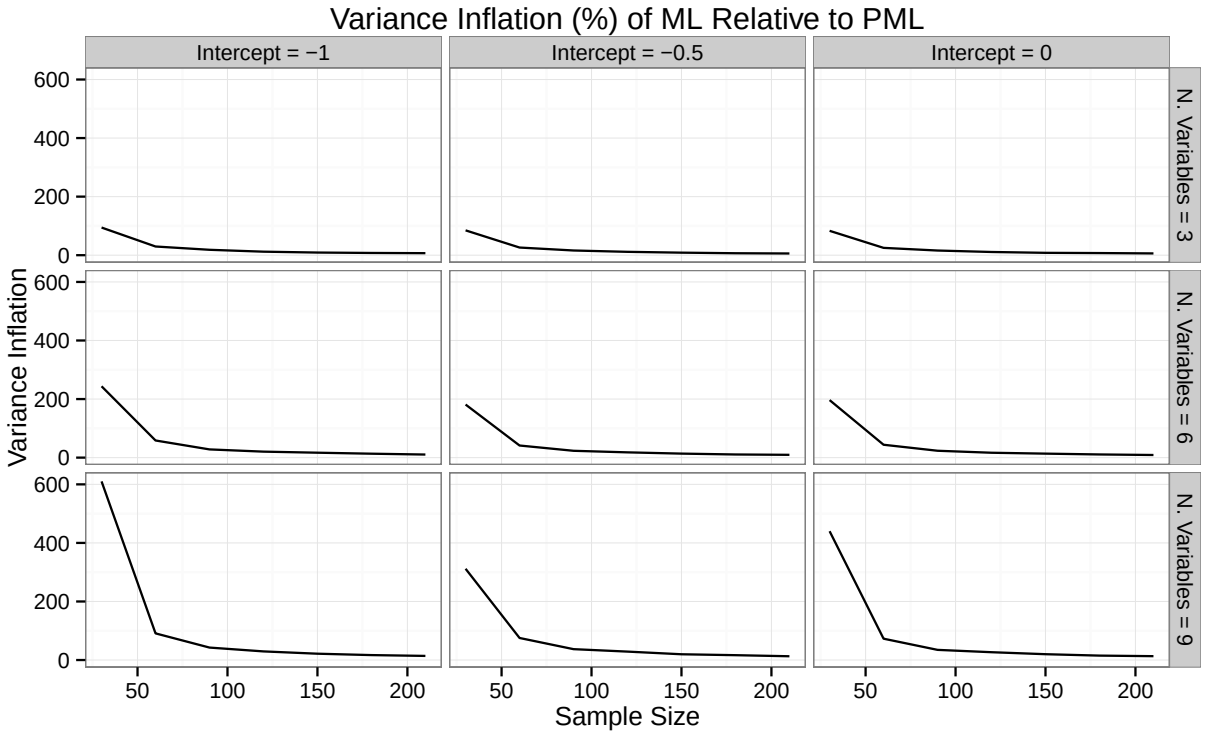

Figure 7: This figure shows the percent inflation in the variance of $\hat{\beta}^{mle}$ compared to $\hat{\beta}^{pmle}$.

B.4 Mean Squared Error

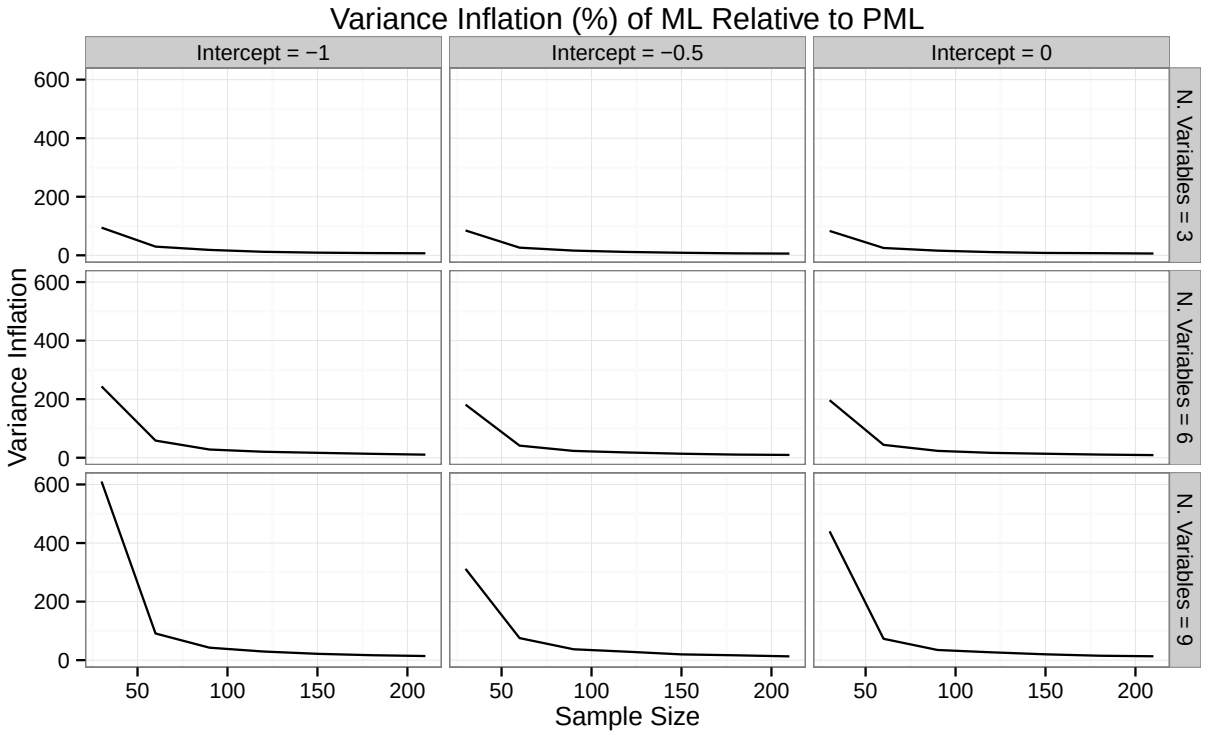

Figure 8: This figure shows the mean squared error of $\hat{\beta}^{mle}$ and $\hat{\beta}^{pmle}$.

B.5 Mean Squared Error Inflation

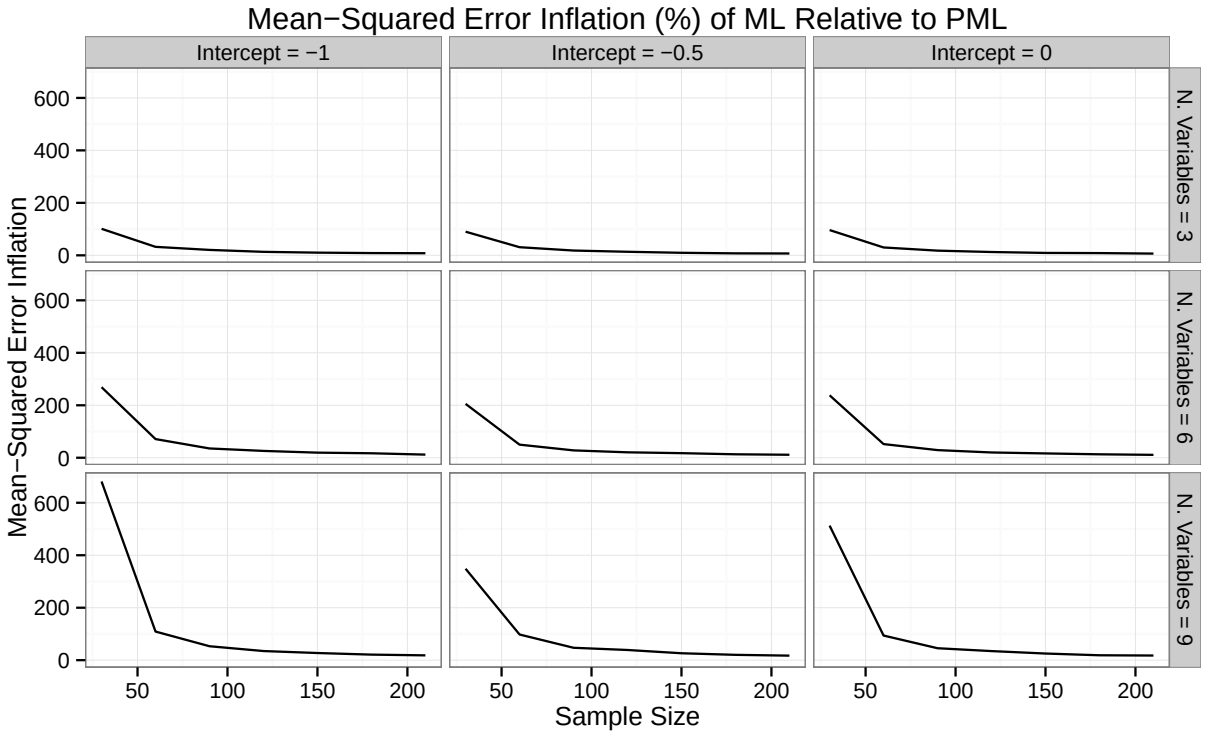

Figure 9: This figure shows the percent increase in the mean square error of $\hat{\beta}^{mle}$ compared to $\hat{\beta}^{pmle}$.