Logistic Regression in Small Samples

(Nearly) Unbiased Estimation with Penalized Maximum Likelihood[*]

Kelly McCaskey[†]

Carlisle Rainey[‡]

Abstract

When used in small samples, maximum likelihood estimates of logistic regression coefficients can be substantially biased away from zero. This bias might be 25 percent or more in plausible scenarios. As a solution to this problem, 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 the penalized maximum likelihood estimation eliminates most of the bias, but also 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 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, University at Buffalo, SUNY, 520 Park Hall, Buffalo, NY 14260 (kellymcc@buffalo.edu).

[‡]Carlisle Rainey is Assistant Professor of Political Science, University at Buffalo, SUNY, 520 Park Hall, Buffalo, NY 14260 (rcrainey@buffalo.edu).

Introduction

Asymptotically, the maximum likelihood (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, Casella and Berger 2002, p. 470). For small samples, though, the asymptotic approximation does not work well (Long 1997, pp. 53-54). The sampling distribution of $\hat{\beta}^{mle}$ is not centered at β^{true}, so that $E(\hat{\beta}^{mle}) \not\approx \beta^{true}$. Providing a rough heuristic about appropriate sample sizes, Long (1997) writes: "It is risky to use ML with samples smaller than 100, while samples larger than 500 seem adequate" (p. 54). This presents the researcher with a problem: When dealing with small samples, how can she obtain reasonable estimates of logistic regression coefficients? Some problems do not lend themselves to gathering more data

In the typical situation, the researcher models the probability of an event as $\Pr(y_i) = \Pr(y_i = 1 \mid X_i) = \dfrac{1}{1+e^{-X_i\beta}}$, 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

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

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

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

The researcher can obtain the maximum likelihood estimate $\hat{\beta}^{mle}$ by finding the vector β that maximizes $\log L$. However, this estimate is biased in small samples, so that $E(\hat{\beta}^{mle}) \neq \beta^{true}$.

Correcting the Bias

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 yields a "penalized" likelihood function $L^*(\beta|y) = L(\beta|y)|I(\beta)|^{\frac{1}{2}}$ (see also Kosmidis 2014, and Kosmidis and Firth 2009). It turns out that this penalty is equivalent to Jeffreys' (1946) prior for the logistic regression model (Firth 1993, Poirier 1994). Taking logs yields the penalized log-likelihood function.

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

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) previously introduced political scientists to Firth's penalty in the context of solving the problem of separation, but the broader application to small sample problem seems to have gone unnoticed in political science.

A researcher can implement PML just as easily as she can implement ML, but PML estimates are both less biased (Firth 1993) and more efficient than ML estimates (Kosmidis 2007, p. 49).[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.[2]

[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.

[2]We should also note that the PML estimates also solve the problem with separation (Zorn 2005 and Rainey 2015).

Monte Carlo Simulations

To demonstrate the substantial bias of $\hat{\beta}^{mle}$ and the near unbiasedness of $\hat{\beta}^{pmle}$ in 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 three features of the ML and PML estimators:

1. The ML estimator can be quite biased in small samples, as much as 50%, though the bias quickly disappears as the sample size nears and passes 150. The PML is nearly unbiased, regardless of sample size.
2. The variance of the ML estimator can be twice as large as the PML estimator.
3. Combining the previous two features, the mean-squared error of the ML estimator is can be 400% larger than the PML estimator.

In our simulation, the true data generating process is always $\Pr(y_i = 1) = \frac{1}{1+e^{-X_i\beta}}$, where $i \in 1, 2, \ldots, n$ and $X_i\beta = \beta_{cons} + 0.5x_1 + \sum_{j=2}^{k} 0.2x_j$. We consider β_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 $k \in \{3, 6, 9\}$, and the the intercept $\beta_{cons} \in \{-1, -0.5, 0.0\}$ (which, in turn, varies the proportion of events from about 28% to 38% to 50%). We simulate 10,000 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 the expected value and variance of the estimates by computing the mean and variance of the ML and PML estimates across the 10,000 data sets.

Bias

We calculate the percent bias $= 100 \times \left(\frac{E(\hat{\beta})}{\beta^{true}} - 1\right)$ as the sample size, 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, 30 observations) to around 2% (intercept equal to zero, 2 predictors, 150 observations). The bias in the PMLE, on the other hand, is barely noticeable, regardless of the simulation parameters. 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. Figures 4 and 5 in Section B of the Appendix show the expected value and (absolute) bias of these estimates.

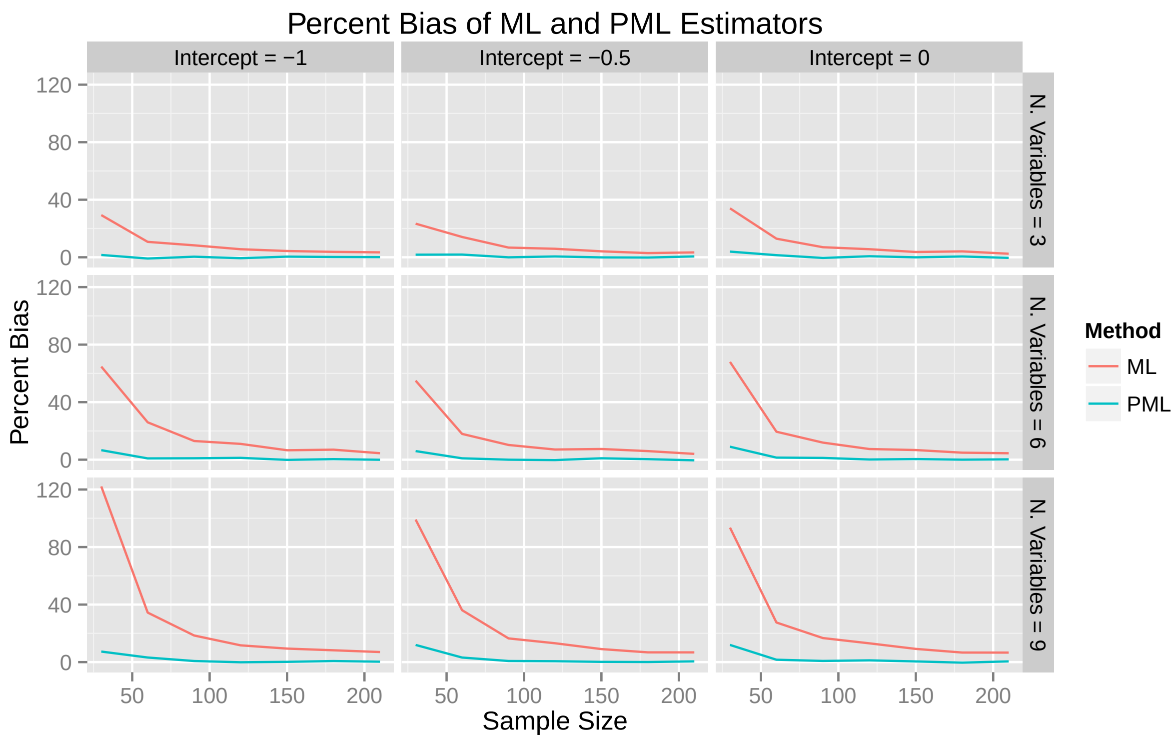

Figure 1: This figure illustrates the substantial bias of $\hat{\beta}^{mle}$ and the near unbiasedness of $\hat{\beta}^{pmle}$.

Variance

In many cases, estimation routines trade-off bias and variance, 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 (lss 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 importance 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 $= \frac{1}{n}\sum_{i=1}^{n}(\hat{\beta} - \beta^{true})^2 = \text{variance} + \text{bias}^2$. The large differences in the variance and the relatively smaller differences in bias lead to substantial differences in the MSE of the two estimators. Figures 6, 7, and 8 in Section B of the Appendix show the variance inflation, mean squared error, and mean squared error inflation of the estimates.

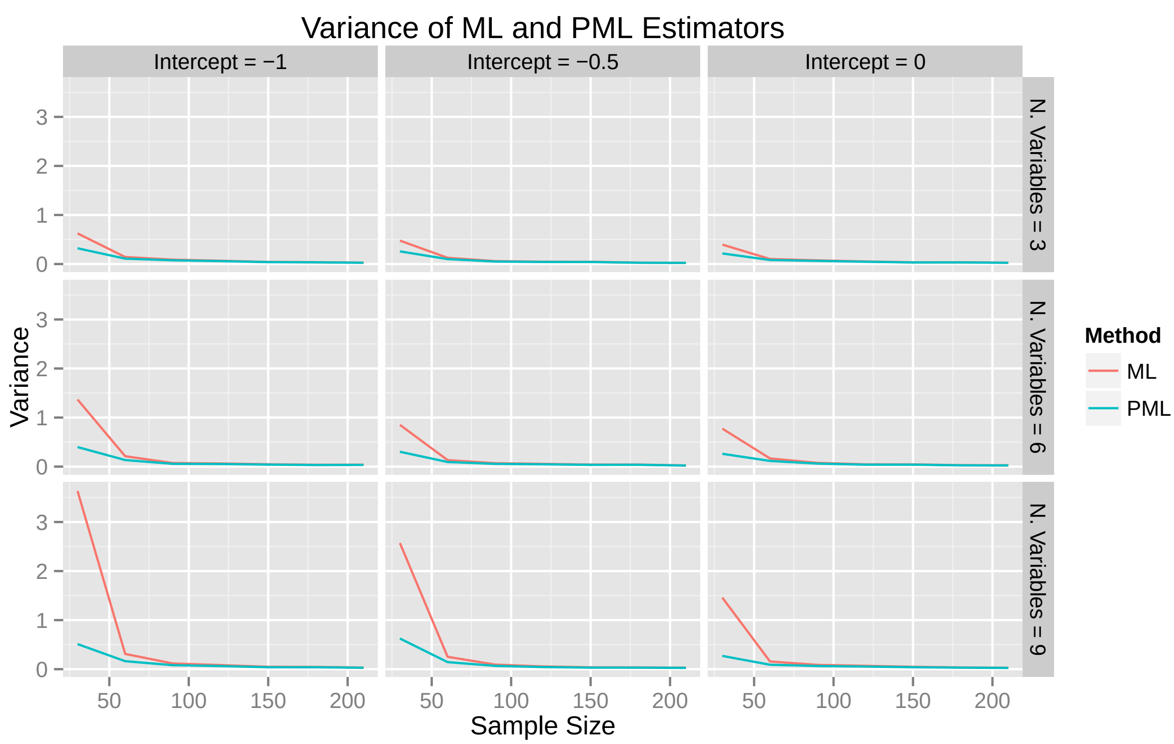

Figure 2: This figure illustrates the smaller variance of $\hat{\beta}^{pmle}$ compared to $\hat{\beta}^{mle}$.

These simulation results illustrate that Firth's (1993) bias correction is not trivial. In small samples, logistic regression coefficients estimated with ML 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.

Replication of Weisiger (2014)

To illustrate the potential importance of using the nearly unbiased PML estimator in an application, we replicate 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 re-analyze Weisiger's data using logistic regression to show the substantial difference between the biased ML estimates and the nearly unbiased PML estimates. Weisiger uses a linear probability model for this analysis, but that leads to the probabilities that fall outside the $[0, 1]$ interval.[3] Figure 3 shows the coefficient estimates and 90% confidence intervals using ML and PML. Notice that the unbiased 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%.

Because we do not know the true model, we cannot know which of these sets of coefficient is

[3]Specifically, we re-analyze the Model 3 in Weisiger's Table 2 (p. 14). In the original analysis, Weisiger uses a 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. We should also note that the PML estimates also solve the problem with separation (Zorn 2005 and Rainey 2015).

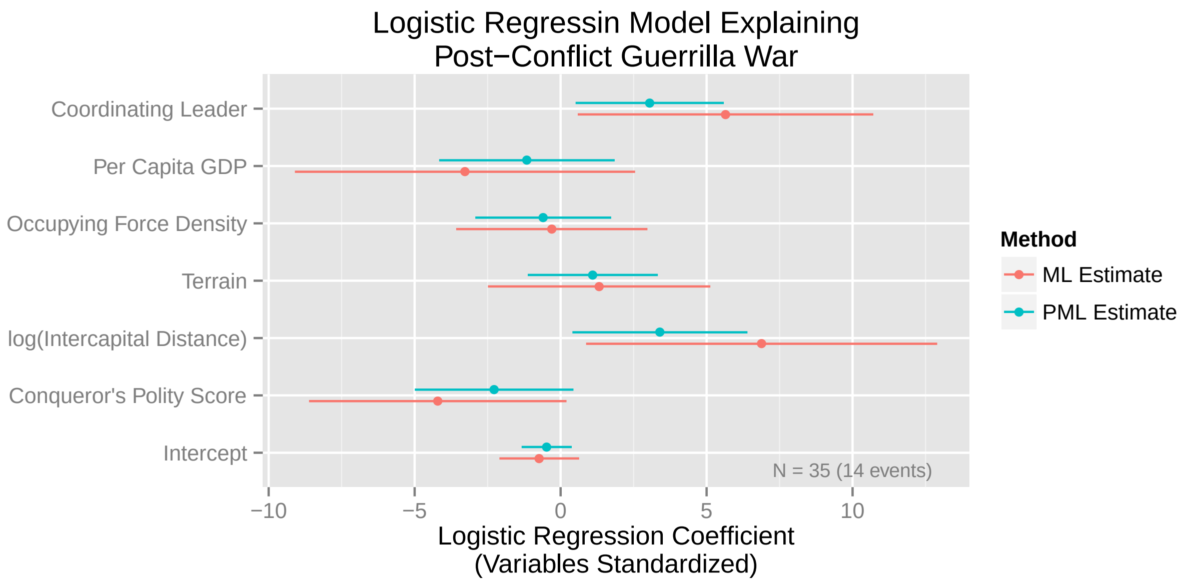

Figure 3: This figure shows the coefficients for a logistic regression model estimated explaining post-conflict guerrilla war estimated with ML and PML. Notice that the PML estimates tend to be much smaller than the ML estimates.

best. However, we can use out-of-sample prediction to help adjudicate between these two methods. We use leave-one-out cross-validation and summarize the prediction errors using Brier and log scores, for which larger values indicate worse predictive ability.[4] 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. 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 the coefficients themselves. In this example, we focus on Weisiger's hypothesis that there will be a greater chance of resistance when the pre-conflict political leader remains at

[4]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.

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.

PML pools the estimated probabilities toward zero. 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.

This substantial difference in these estiamtes demonstrates the potential importance of using PML whne 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.

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.

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/wei

# 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

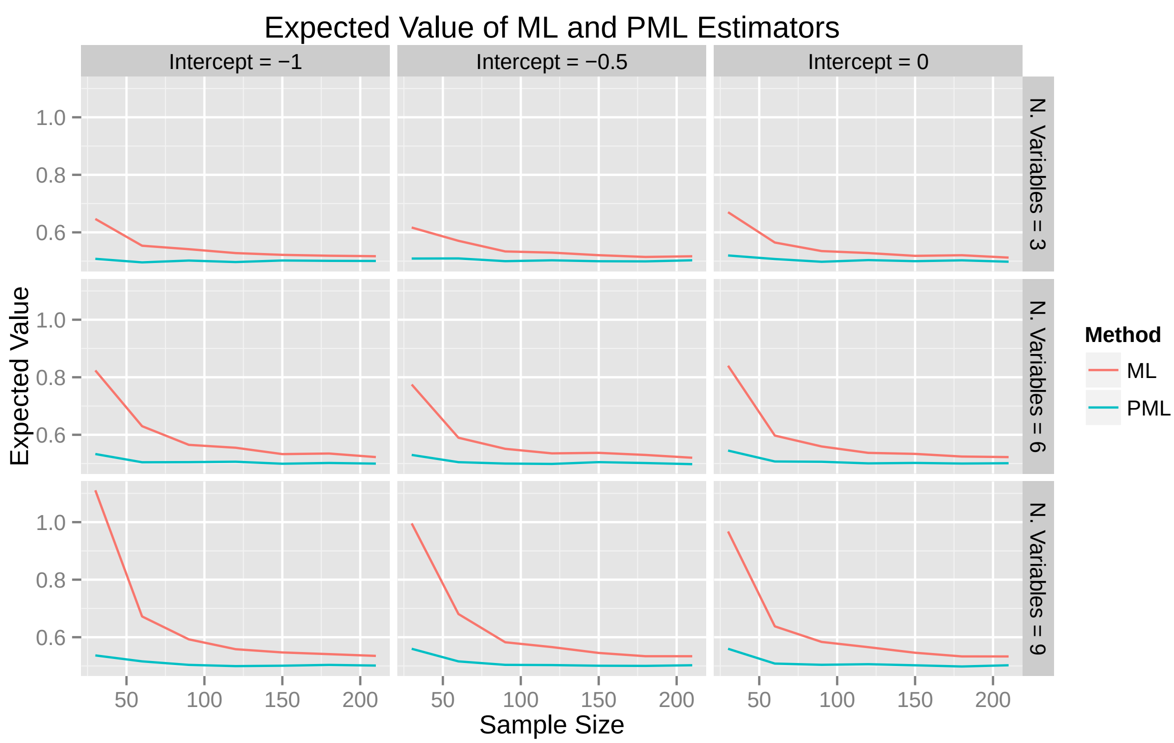

Figure 4: This figure shows the expected value of $\hat{\beta}^{mle}$ and $\hat{\beta}^{pmle}$. The true value is $\beta = 0.5$.

B.2 Bias

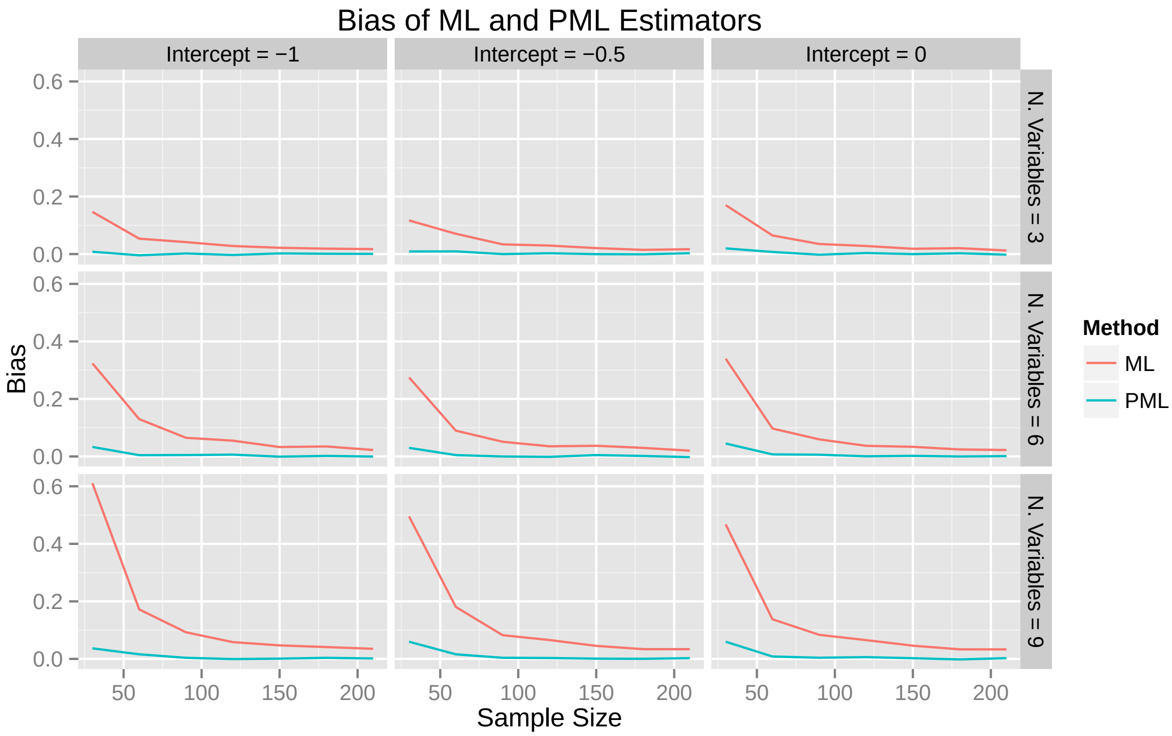

Figure 5: This figure shows the bias of $\hat{\beta}^{mle}$ and $\hat{\beta}^{pmle}$.

B.3 Variance Inflation

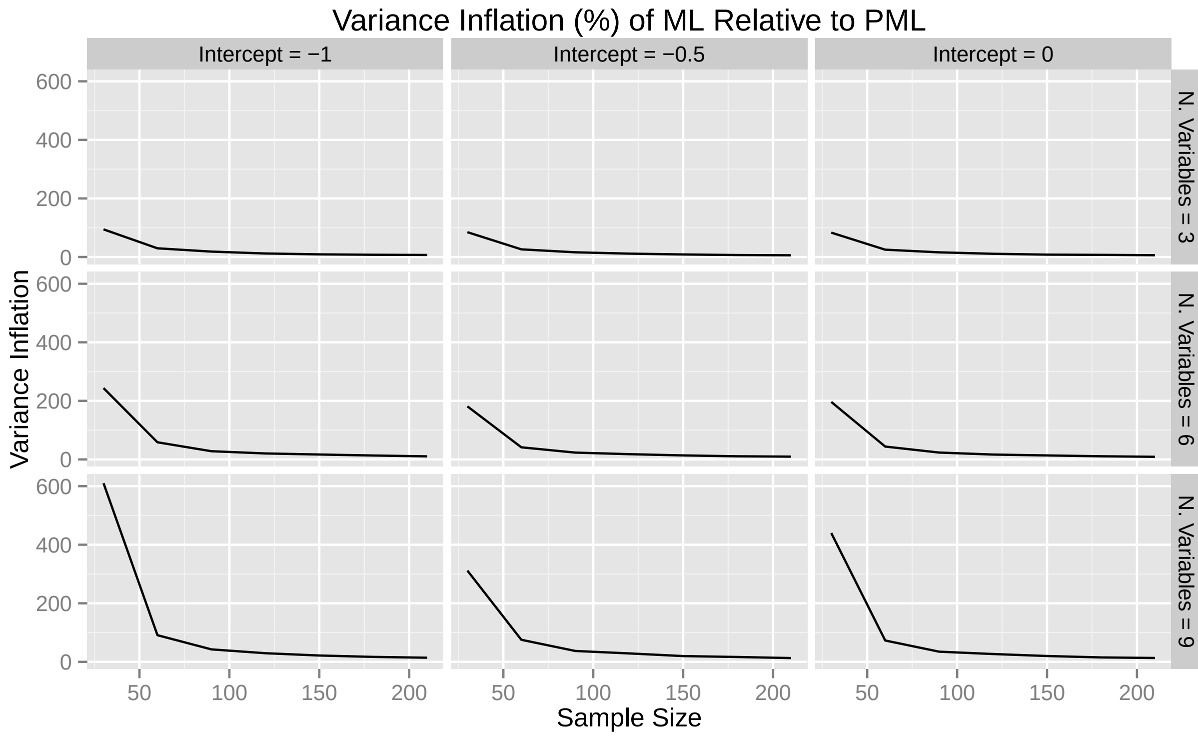

Figure 6: This figure shows the percent inflation in the variance of $\hat{\beta}^{mle}$ compared to $\hat{\beta}^{pmle}$.

B.4 Mean Squared Error

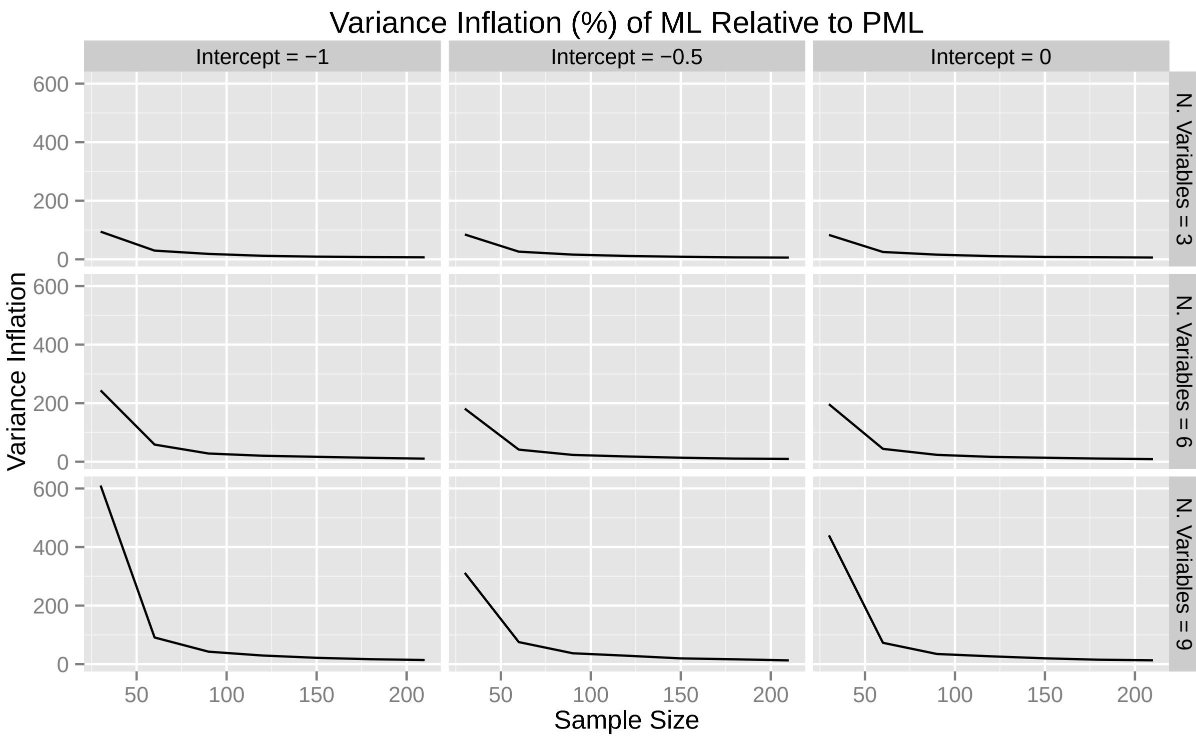

Figure 7: This figure shows the mean squared error of $\hat{\beta}^{mle}$ and $\hat{\beta}^{pmle}$.

B.5 Mean Squared Error Inflation

Figure 8: This figure shows the percent increase in the mean square error of $\hat{\beta}^{mle}$ compared to $\hat{\beta}^{pmle}$.