Logistic Regression with Small Samples*

Kelly McCaskey†

Carlisle Rainey‡

Abstract

In small samples, maximum likelihood (ML) estimates of logistic regression coefficients have substantial bias away from zero. As a solution, we introduce political scientists to the penalized maximum likelihood (PML) estimator. The PML estimator eliminates most of the bias and, perhaps more importantly, greatly reduces the variance of the usual ML estimator. Thus, researchers do not face a bias-variance tradeoff when choosing between the ML and PML estimators–the PML estimator has a smaller bias *and* a smaller variance. These reductions in the bias and variance are not trivial. Monte Carlo simulations and two replication studies show significant improvements that certainly deserve the attention of substantive researchers.

Word Count: 5,283

*We thank Alex Weisiger, Tracy George, and Lee Epstein for making their data available. We conducted these analyses analyses with R 3.2.2. 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).

Maximum likelihood (ML) estimators have become ubiquitous in political science by allowing researchers to represent a wide range of theoretically-motivated data-generating processes (King 1998). Further, ML estimators have excellent large sample properties. In small samples, though, ML estimators sometimes behave quite poorly–especially for logistic regression models.

The Big Problem with Small Samples

When working with a binary outcome, the researcher might use logistic regression, modeling 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.[1] 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 find the ML estimate $\hat{\beta}^{mle}$ by finding the vector β that maximizes $\log L$ (King 1998).

The ML estimate of the logistic regression coefficient vector $\hat{\beta}^{mle}$ asymptotically unbiased, 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). Also, the ML estimate is asymptotically efficient, so that the asymptotic variance of the ML estimate equals the Cramer-Rao lower bound (Greene 2012, pp. 513-523, and Casella and Berger 2002, pp. 472, 516). For small samples, though, the ML estimator does not work well–the ML estimates have substantial bias away from zero (Long 1997, pp. 53-54). Long (1997, p. 54) offers a rough heuristic about appropriate sample sizes: "It is risky to use ML with samples smaller than 100, while samples larger than 500 seem adequate."[2] This presents the researcher with a problem: When dealing with small samples, how can she obtain reasonable estimates of logistic regression coefficients?

[1]To simplify the exposition, we focus on logistic regression models. However, all the ideas apply equally well to probit models.

[2]Making the problem worse, King and Zeng (2001) point out that ML estimates have substantial bias for much larger sample sizes if the event of interest occurs only rarely.

An Easy Solution for the Big Problem

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, but the broader and perhaps more important application to small sample problems seems to have gone unnoticed in political science.

A researcher can implement PML as easily as ML, but PML estimates of logistic regression coefficients are less biased (Firth 1993) and less variable (Kosmidis 2007, p. 49, and Copas 1988) than ML estimates.[3] This is important. When choosing among estimators, researchers often face a tradeoff between bias and variance (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.

Two concepts from statistical theory help explain the relationship between the ML and PML estimators. Suppose two estimators A and B, with squared error loss functions R^A and R^B. If $R^A \leq R^B$ for all possible true parameter values and the inequality holds strictly for at least some parameter values, then we can say that estimator B is *inadmissible* and that estimator A *dominates* estimator B. Now suppose a squared error loss function R for the logistic regression coefficients, such that $R^{mle} = E[(\hat{\beta}^{mle} - \beta^{true})^2]$ and $R^{pmle} = E[(\hat{\beta}^{pmle} - \beta^{true})^2]$. In this case, the inequality holds strictly for all β^{true} so that $R^{pmle} < R^{mle}$. Thus, we can describe the ML estimator as *inadmissible* and say that the PML estimator *dominates* the ML estimator.

We can say that the PML estimator dominates the ML estimator because the PML estimator has lower bias and variance *regardless of the sample size*. That is, the PML estimator *always* outperforms the ML estimator in terms of bias, variance, and mean-squared error. However, both

[3]The penalized maximum likelihood estimates are easy to calculate in R using the `logistf` or `brglm` packages and in Stata with the `firthlogit` module. See the Section A and Section B of the Appendix, respectively, for examples.

estimators are asymptotically unbiased and efficient, so the difference between the two estimators becomes negligible as the sample size grows large. In small samples, though, Monte Carlo simulations show substantial improvements that should grab the attention of substantive researchers.

The Big Improvements from an Easy Solution

To show that the reductions in bias, variance, and mean-squared error are large enough to concern substantive researchers, we conduct a Monte Carlo simulation comparing the sampling distributions of the ML and PML estimates for small sample sizes sometimes found in political science research. These simulations demonstrate two features of the ML and PML estimators:

1. In small samples, the ML estimator exhibits a large bias. The PML estimator is nearly unbiased, regardless of sample size.
2. In small samples, the variance of the ML estimator is much larger than the variance of the PML estimator.

Of course, the increased bias and variance of the ML estimator implies that the PML estimator will also have a 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 $X\beta = \beta_{cons} + 0.5x_1 + \sum_{j=2}^{k} 0.2x_j$. 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 data sets for each combination of the simulation parameters and estimate the logistic regression coefficients using ML and PML for each data set. We use these datasets to estimate the percent bias and variance of the ML and PML estimators, as well as the MSE inflation of the ML estimator compared to the PML estimator.

Bias

We calculate the percent bias $= 100 \times \left(\frac{E(\hat{\beta})}{\beta^{true}} - 1\right)$ as the sample size N, the intercept β_{cons} (i.e., proportion of events), and number of explanatory variables k 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% ($\beta_{cons} = -1$, $k = 9$, and $N = 30$) to around 2% ($\beta_{cons} = 0$, $k = 3$, and $N = 210$). The bias in the PMLE, on the other hand, is much smaller. For the worst-case scenario ($\beta_{cons} = -1$, $k = 9$, and $N = 30$), the percent bias in the PMLE is only about seven percent.[4]

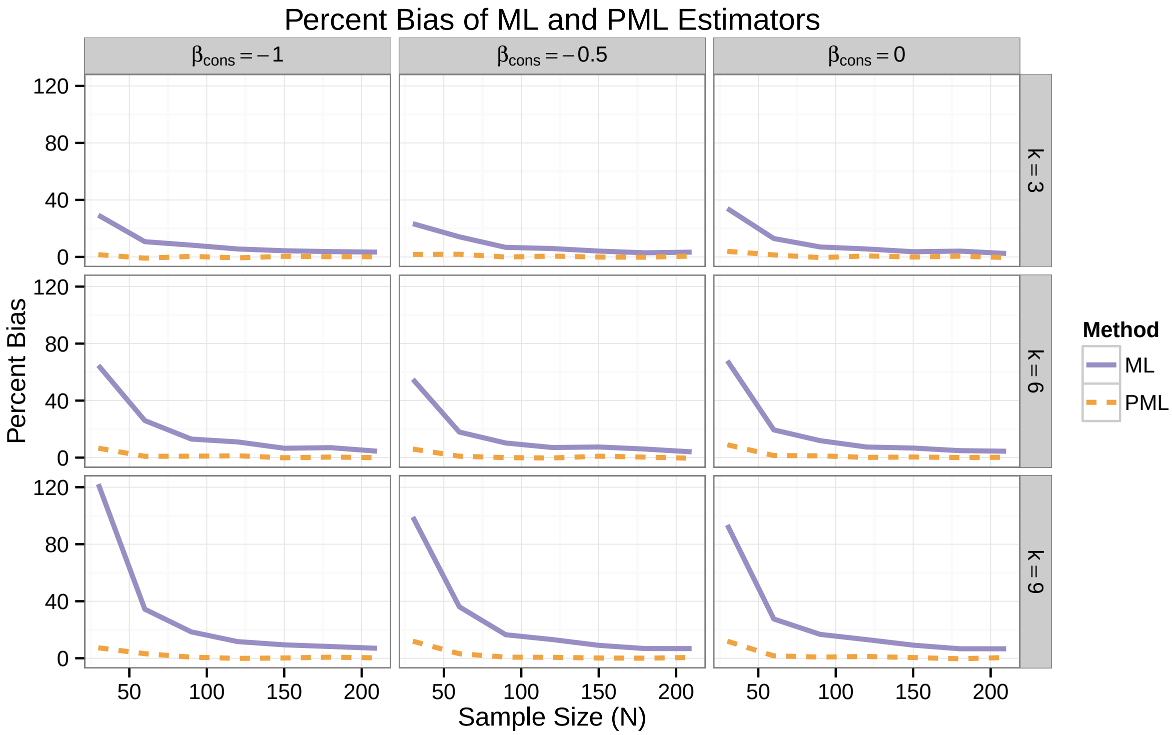

Figure 1: This figure illustrates the substantial bias of $\hat{\beta}^{mle}$ and the near unbiasedness of $\hat{\beta}^{pmle}$.

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.[5]

[4]Figures 9 and 10 in Section C of the Appendix show the expected value and (absolute) bias of these estimates.

[5]Figure 11 in the Appendix shows the variance inflation $= 100 \times \left(\frac{Var(\hat{\beta}^{mle})}{Var(\hat{\beta}^{pmle})} - 1 \right)$.

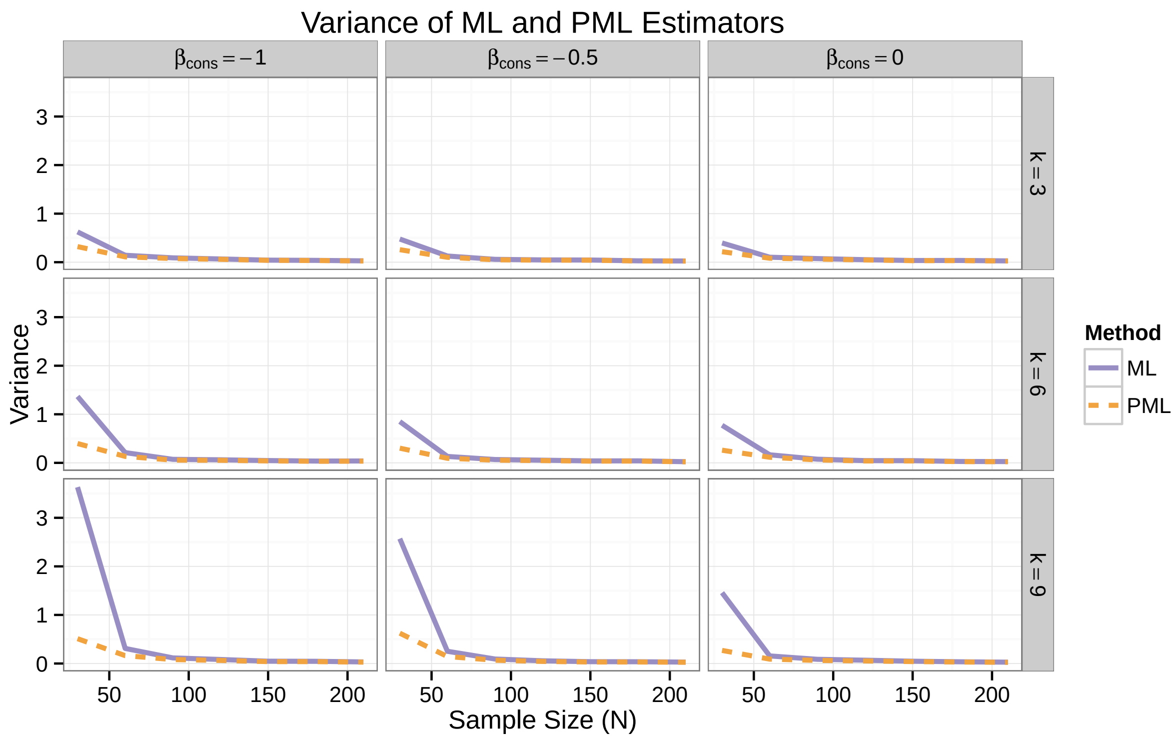

Figure 2: This figure illustrates the smaller variance of $\hat{\beta}^{pmle}$ compared to $\hat{\beta}^{mle}$.

Mean-Squared Error

Although both bias and variance alone are important, the mean-squared error (MSE) combines the two into an overall measure of the accuracy of the estimator, where

$$\begin{aligned} MSE(\hat{\beta}) &= E[(\hat{\beta} - \beta^{true})^2] \\ &= Var(\hat{\beta}) + [E(\hat{\beta})]^2. \end{aligned} \tag{1}$$

Since the bias and the variance of ML estimators are greater than the PML estimator, it follows the the ML estimator must have a larger MSE, so that $MSE(\hat{\beta}^{mle}) - MSE(\hat{\beta}^{pmle}) > 0$.

We should care about the size of this difference, though, not the sign. Take the percent increase in the MSE for the ML estimator compared to the PML estimator as the quantity of interest. Refer to this quantity as the "MSE inflation" and denote it as $MSE^{\Uparrow}$, where

$$MSE^{\Uparrow} = 100 \times \frac{MSE(\hat{\beta}^{mle}) - MSE(\hat{\beta}^{pmle})}{MSE(\hat{\beta}^{pmle})} \ . \tag{2}$$

Figure 3 shows the MSE inflation for each combination of the parameter simulations on the $\log_{10}$ scale. Notice that in the worst-case scenario ($\beta_{cons} = -1$, $k = 9$, and $N = 30$), the MSE of the ML estimates is about 681% larger than the MSE of the PML estimates. The MSE inflation only barely drops below 10% for the most information-rich parameter combinations (e.g., $\beta_{cons} = 0$, $k = 2$, and $N = 210$). The MSE for the ML estimator is more than 100% larger than the PML estimator for about 13% of the simulation parameter combinations, more than 50% larger for about 24% of the combinations, and more than 25% larger for 49% of the combinations. These differences are certainly large enough to command the attention of researchers working with small data sets.

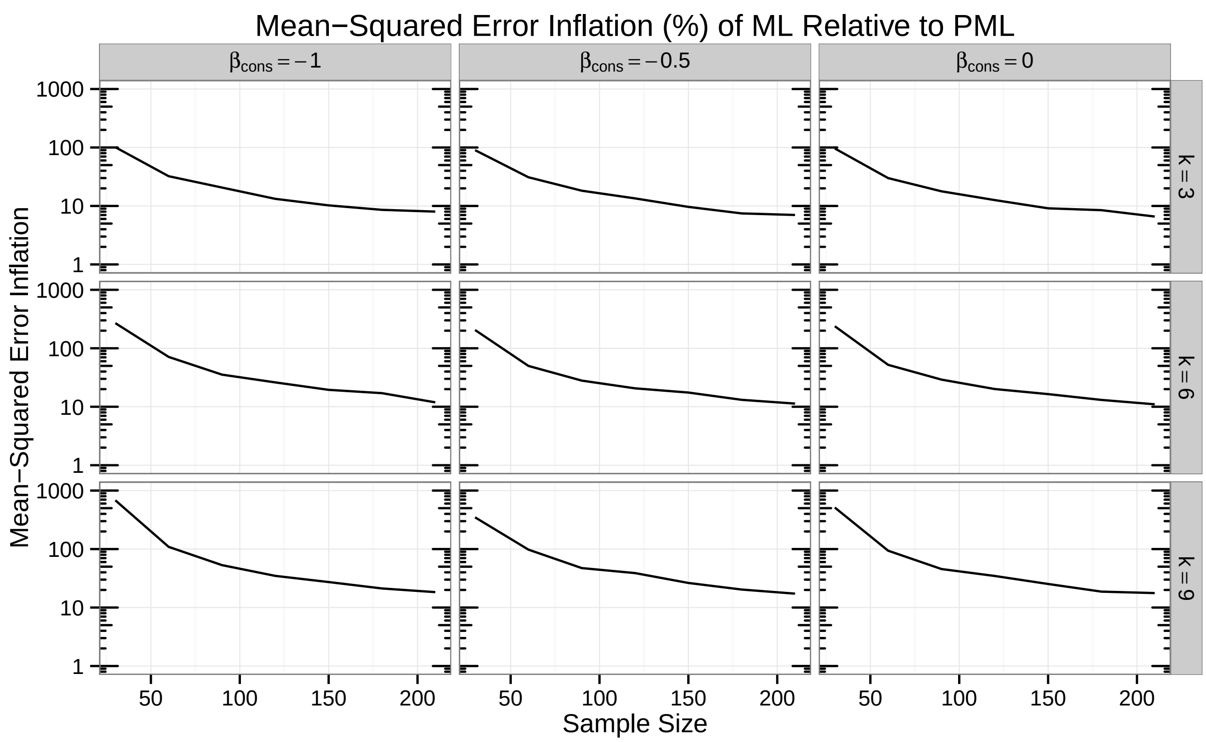

Figure 3: This figure shows the percent increase in the mean square error of $\hat{\beta}^{mle}$ compared to $\hat{\beta}^{pmle}$.

We might, ask though, about the relative contributions of the bias and variance to this difference. Substituting Equation 1 into Equation 2 for $MSE(\hat{\beta}^{mle})$ and $MSE(\hat{\beta}^{pmle})$ and rearranging, we obtain

$$MSE^{\Uparrow} = \overbrace{100 \times \frac{Var(\hat{\beta})}{Var(\hat{\beta}^{pmle}) + [E(\hat{\beta}^{pmle})]^2}}^{\text{contribution of variance}} + \underbrace{100 \times \frac{[E(\hat{\beta})]^2}{Var(\hat{\beta}^{pmle}) + [E(\hat{\beta}^{pmle})]^2}}_{\text{contribution of bias}} + 100 \,,$$

which allows us to easily see the contribution of the bias and variance to the total MSE. We can simply plug in the simulation estimates of the bias and variance of each estimator to obtain the contribution of each. But notice that we can easily compare the *relative* contributions of the bias and variance using the ratio

$$\text{relative contribution of variance} = \frac{\text{contribution of variance}}{\text{contribution of bias}}. \tag{3}$$

Figure 4 shows the relative contribution of the variance as the sample size increases. Values less than one indicate that the bias makes a greater contribution and values greater than one indicate that the variance makes a greater contribution. In each case, the variance is much larger than one. For $N = 30$, the contribution of the variance is between 7 and 36 times larger than the contribution of the bias. For $N = 210$, the contribution of the variance is between 25 and 180 times larger than the contribution of the bias. In spite of the attention paid to the small sample *bias* in ML estimates of logistic regression coefficients, the small sample *variance* is a more important problem to address. Fortunately, the PML estimator greatly reduces the variance (in addition to shrinking the bias), resulting in a much smaller MSE, especially for smaller samples.

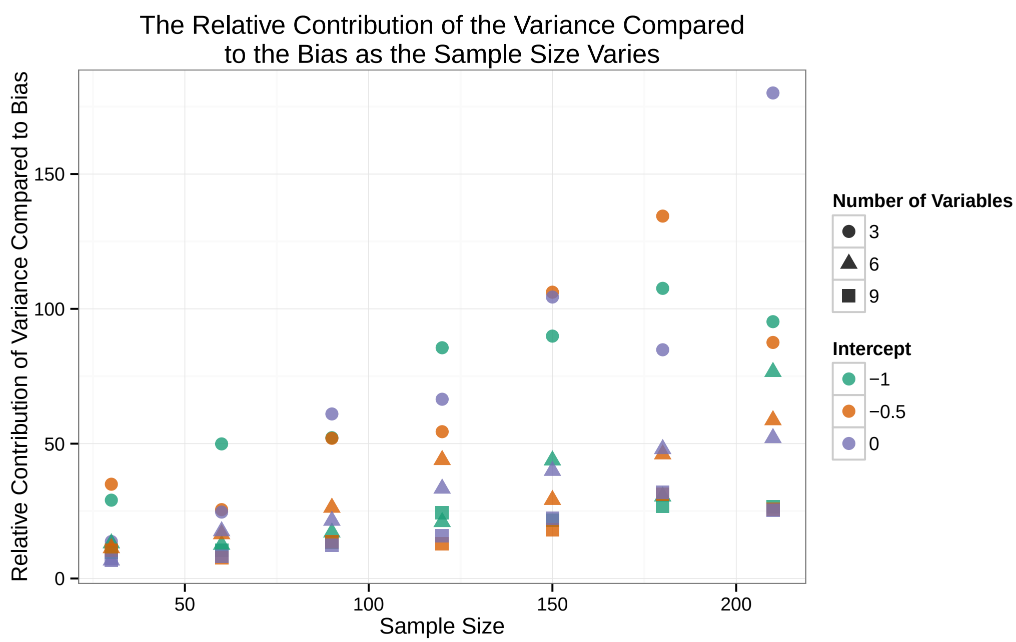

Figure 4: This figure shows the relative contribution of the variance and bias to the MSE inflation. The relative contribution is defined in Equation 3.

These simulation results show that the bias, variance, and mean-squared error of the ML estimates of logistic regression coefficients are not trivial in small samples. These problems cannot be safely ignored. Fortunately, Firth's (1993) bias correction is nearly trivial to implement and offers substantial improvements. And these improvements are not limited to Monte Carlo studies. In the two applications that follow, we show that the PML estimator leads to substantial reductions in the magnitude of the coefficient estimates and in the width of the confidence intervals.

The Substantive Importance of the Big Improvements

Weisiger (2014)

To illustrate the substantive importance of using the PML estimator, we reanalyze a portion of the statistical analysis in Weisiger (2014) and show the substantial differences in the estimates and the confidence intervals from the ML and PML estimators.

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.[6] Prior to estimation, we standardize the continuous variables to have mean zero and standard deviation one-half and binary variables to have mean zero (Gelman 2008). Figure 5 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,

[6]Specifically, we reanalyze 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.

the PML confidence intervals are about half as wide as the ML confidence intervals.

Figure 5: 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 and confidence intervals tend to be much smaller than the ML estimates and confidence intervals.

Because we do not know the true model, we cannot know which of these sets of coefficients is better. 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 smaller values indicate better predictive ability.[7] 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

[7]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.

post-conflict guerrilla war as countries gain a coordinating leader. Figure 6 shows the estimates of the quantities of interest.

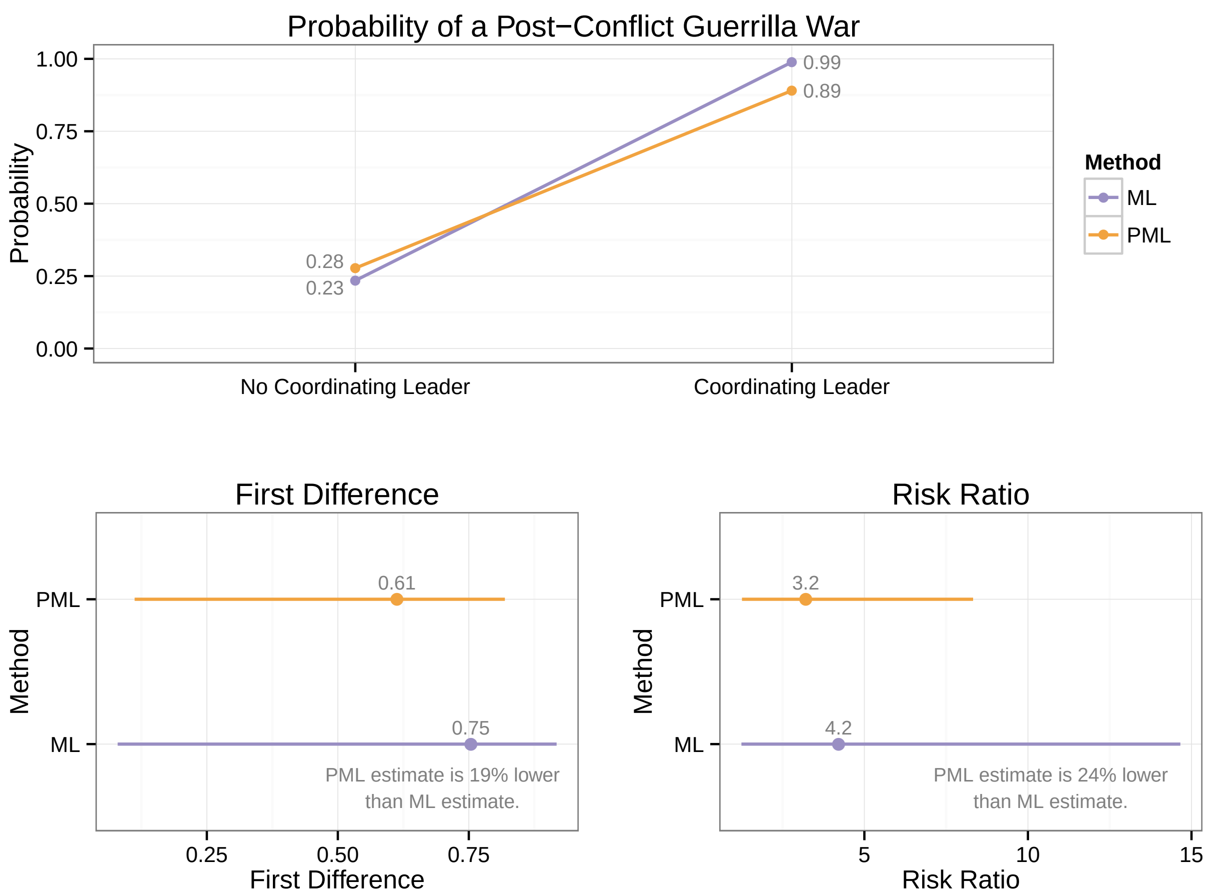

Figure 6: This figure shows the quantities of interest for the effect of a coordinating leader on the probability of a post-conflict guerrilla war.

PML pools the estimated probabilities toward one-half, 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.

George and Epstein (1992)

In another application, we reanalyze the statistical analysis of the integrated model of U.S. Supreme Court decisions developed by George and Epstein (1992). As before, we find substantial differences in the estimates and the confidence intervals from the ML and PML estimators.

George and Epstein (1992) combine the legal and extralegal models of Court decision making in order to overcome the complementary idiosyncratic shortcomings of each. The legal model claims *stare decisis*, or the rule of law, as the key determinant of future decisions while the extralegal model takes a behavioralist approach containing an array of sociological, psychological, and political factors.

The authors model the probability of a conservative decision in favor of the death penalty as a function of a variety of legal and extralegal factors. Like the previous example of the replication of Weisiger (2014), George and Epstein's sample is relatively small, consisting of 64 Court decisions involving capital punishment from 1971 to 1988, with 29 events (i.e., conservative decisions). They originally use maximum likelihood estimation so we reanalyze their data using both ML and PML to demonstrate the difference in the estimates with such a small sample. Figure 7 shows the shift in coefficient estimates. In all cases, the PML estimate is smaller than the ML estimate. Each coefficient decreases by at least 25% with three decreasing by more than 40%: court change, the largest, at 49%, solicitor general at 48%, and political environment with 41%. Additionally, the PML standard errors are all smaller than their ML counterparts. Three of the 11 coefficients lose their statistical significance.

Again, we use out-of-sample prediction to evaluate the performance of each model, and, again, PML out-performs ML. The ML estimates produce a Brier score of 0.17, and the PML estimates lower the Brier score by 8% to 0.16. Similarly, the ML estimates produce a log score of 0.89, while the PML estimates lower the log score by 41% to 0.53.

Again, because we are using maximum likelihood, we are likely more interested in the functions of the coefficients rather than the coefficients themselves. For instance, we take George and Epstein's integrated model of Court decisions and calculate a first difference and risk ratio as the repeat-player status of the state varies, setting all other explanatory variables at their medias. George and Epstein hypothesize that repeat players have greater expertise and are more likely to win the case. Figure 8 shows the estimates of the quantities of interest.

As we show, PML pools the estimated probabilities toward one-half. When the state is not a repeat player, the PML estimates suggest 17% chance of a conservative decision while ML estimates suggest a 6% chance. However, when the state is a repeat player, the PML estimates suggest that the Court has a 53% chance of a conservative decision compared to the 60% chance suggested by

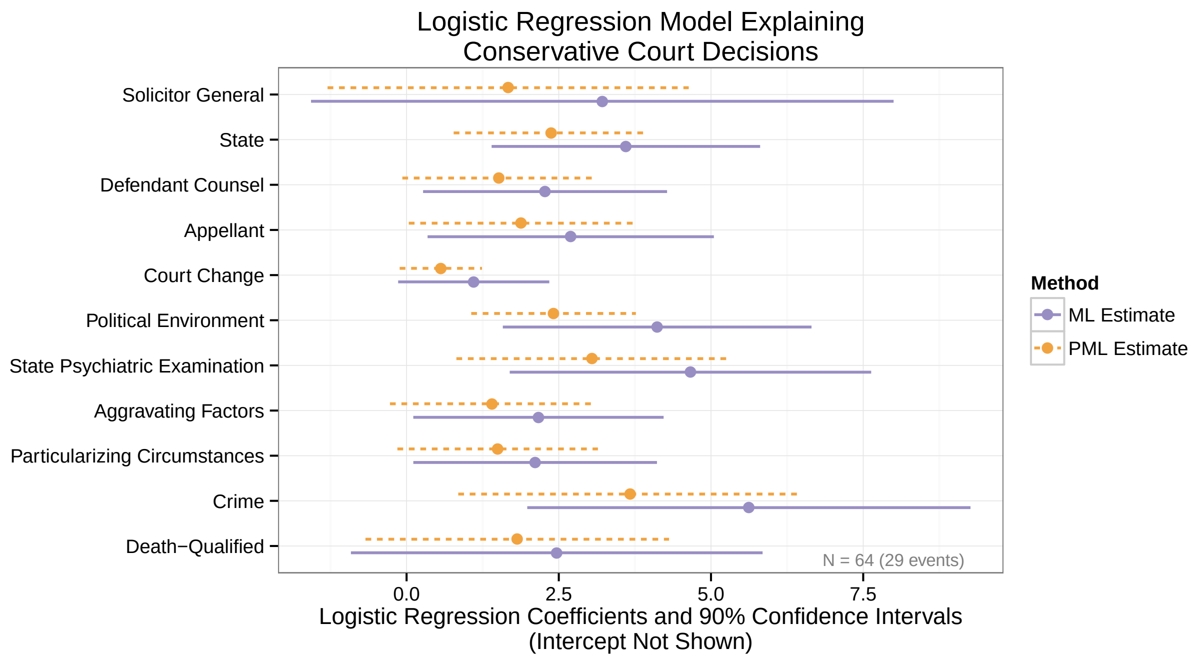

Figure 7: This figure shows the coefficients for a logistic regression model estimating U.S. Supreme Court Decisions by both ML and PML. Note that the PML estimates are smaller with narrower variances than those of ML

ML. Thus, PML also provides smaller effect sizes for both the first difference and the risk ratio. PML decreases the estimated first difference by 17% from 0.63 to 0.52 and the risk ratio by 67% from 12.3 to 4.0.

Across these two applications, the difference between the estimators is clear. The PML coefficient estimates and confidence intervals are consistently and meaningfully smaller. Theoretically, we know that these estimates are less biased and less variable. Importantly, though, these smaller estimates manifest practically in better out-of-sample predictions. And these improvements come at almost *no cost* to researchers. The PML estimator is nearly trivial to implement but *dominates* the ML estimator–the PML estimator always has lower bias, lower variance, and lower MSE.

Conclusion

When estimating a model of a binary outcome with a small sample, a researcher faces several options. First, she might avoid analyzing the data altogether because she realizes that maximum likelihood estimates of logistic regression coefficients have significant bias. We see this as the least attractive option. Even small data sets contain information and avoiding these data sets leads to a

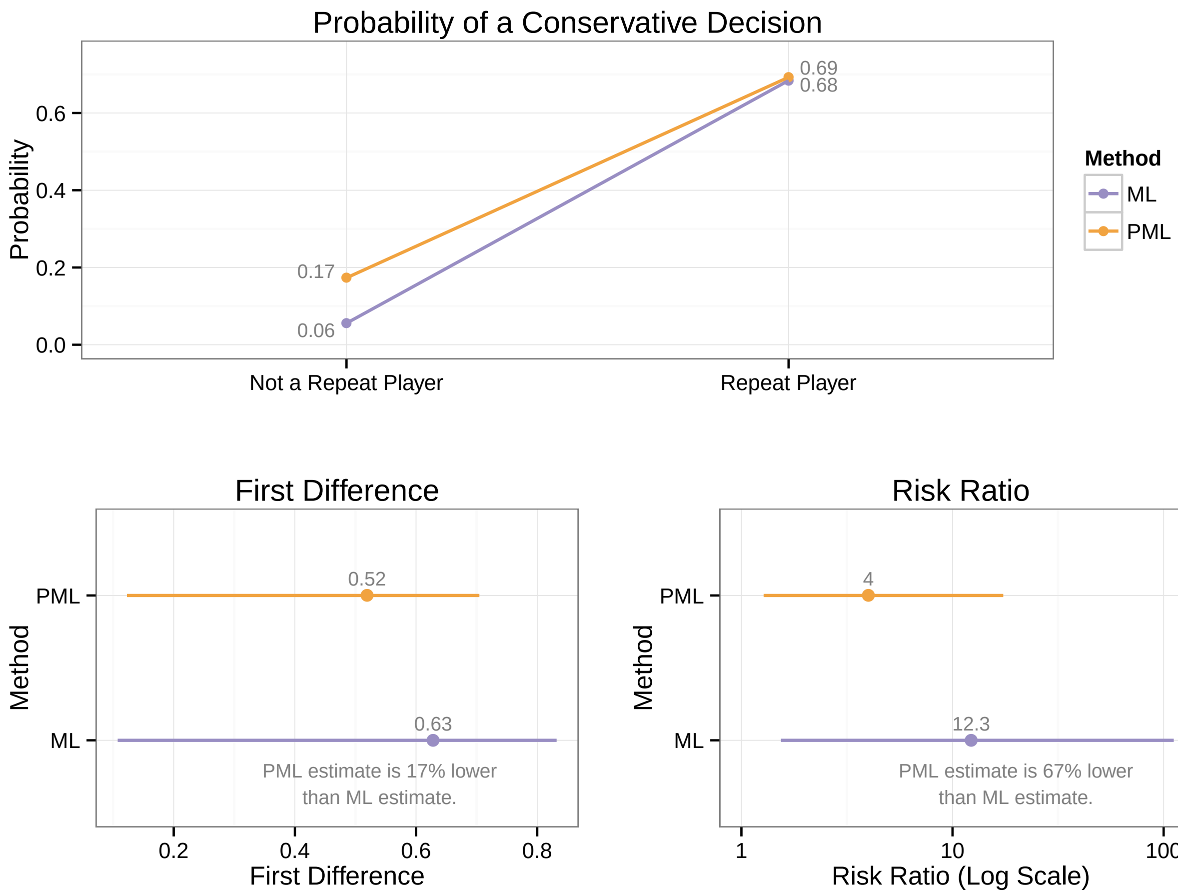

Figure 8: This figure shows the quantities of interest for the effect of the solicitor general filing a brief amicus curiae on the probability of a decision in favor of capital punishment.

lost opportunity.

Second, the researcher might proceed with the biased and inaccurate estimation using maximum likelihood. We also see this option as unattractive, because simple improvements can dramatically shrink the bias and variance of the estimates.

Third, the researcher might use least squares to estimate a linear probability model (LPM). If the probability of an event is a linear function of the explanatory variables, then this approach is reasonable, as long as the researcher takes steps to correct the standard errors. However, in most cases, using an "S"-shaped inverse-link function (i.e., logit or probit) makes the most theoretical sense, so that marginal effects shrink toward zero as the probability of an event approaches zero or one (e.g., Berry, DeMeritt, and Esarey 2010 and Long 1997, pp. 34-47). Long (1997, p. 40) writes: "In my opinion, the most serious problem with the LPM is its functional form." Additionally, the

LPM sometimes produces nonsense probabilities that fall outside the $[0, 1]$ interval and nonsense risk ratios that fall below zero. If the researcher is willing to accept these nonsense quantities and assume that the functional form is linear, then the LPM offers a reasonable choice. However, we agree with Long (1997) that without evidence to the contrary, the logistic regression model offers a more plausible functional form.

Finally, the researcher might simply use penalized maximum likelihood, which allows the theoretically-appealing "S"-shaped functional form while greatly reducing the bias *and* variance. Indeed, the penalized maximum likelihood estimates always have a smaller bias and variance than the maximum likelihood estimates. These substantial improvements come at almost no cost to the researcher in learning new concepts or software beyond maximum likelihood and simple commands in R and/or Stata.[8] We see this as the most attractive option. Whenever researchers have concerns about bias and variance due to a small sample, a simple change to a penalized maximum likelihood estimator can easily ameliorate any concerns with little to no added difficulty for researchers or their readers.

References

Berry, William D., Jacqueline H. R. DeMeritt, and Justin Esarey. 2010. "Testing for Interaction in Binary Logit and Probit Models: Is a Product Term Essential." *American Journal of Political Science* 54(1):105–119.

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

Copas, John B. 1988. "Binary Regression Models for Contaminated Data." *Journal of the Royal Statistical Society, Series B* 50(2):225–265.

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.

George, Tracey E., and Lee Epstein. 1992. "On the Nature of Supreme Court Decision Making." *American Political Science Review* 86(2):323–337.

Greene, William H. 2012. *Econometric Analysis*. 7th ed. Upper Saddle River, New Jersey: Prentice Hall.

[8] Appendices A and B offer a quick overview of computing penalized maximum likelihood estimates in R and Stata, respectively.

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.

King, Gary, and Langche Zeng. 2001. "Logistic Regression in Rare Events Data." *Political Analysis* 9(2):137–163.

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.

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.

Appendix

Logistic Regression with Small Samples

A PML Estimation 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/data/weisiger.csv")

# 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 PML Estimation in Stata

This example code is available at https://github.com/kellymccaskey/small/blob/master/stata/example.do. The example data used is available at https://github.com/kellymccaskey/small/blob/master/stata/ge.csv.

```
* set working directory and load data
* data can be found at https://github.com/kellymccaskey/small/blob/master/stata/ge.csv
cd "your working directory"
insheet using "ge.csv", clear

* install firthlogit
ssc install firthlogit

* estimate logistic regression with pmle
* see coefficient values, standard errors, p-values, etc.
firthlogit court dq cr pc ag sp pe cc ap dc st sg

* see help file for more
help firthlogit
```

C Additional Simulation Results

C.1 Expected Value

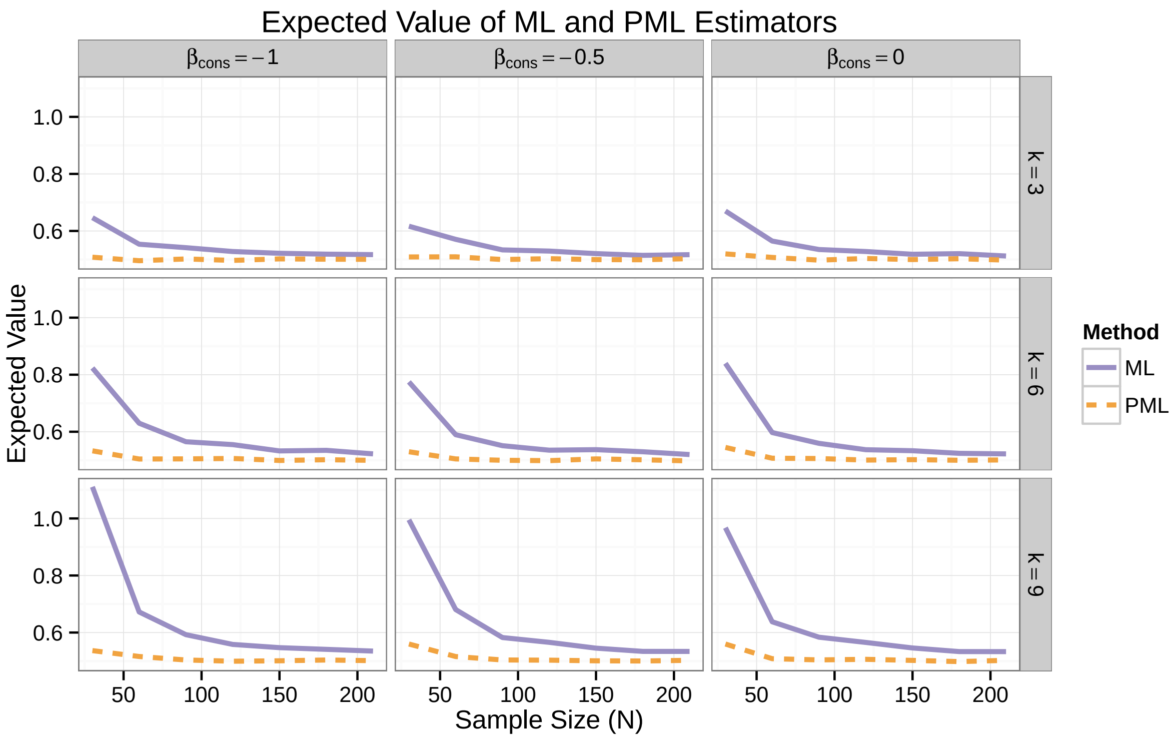

Figure 9: This figure shows the expected value of $\hat{\beta}^{mle}$ and $\hat{\beta}^{pmle}$. The true value is $\beta = 0.5$.

C.2 Bias

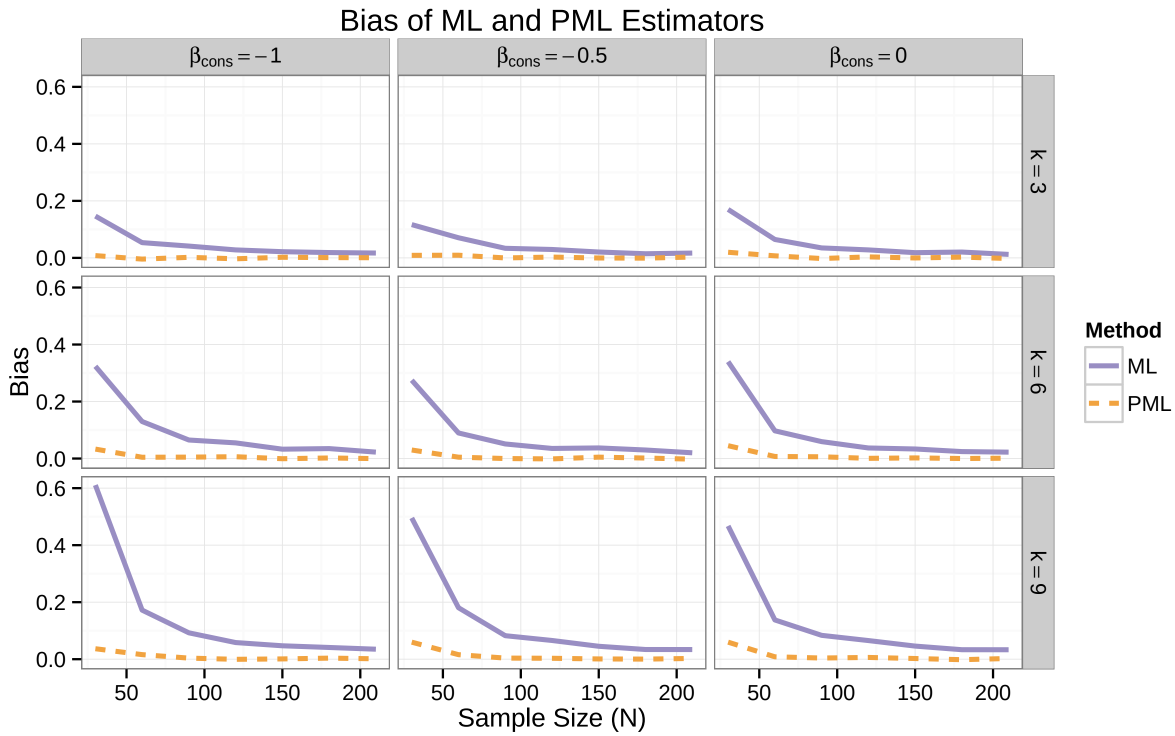

Figure 10: This figure shows the bias of $\hat{\beta}^{mle}$ and $\hat{\beta}^{pmle}$.

C.3 Variance Inflation

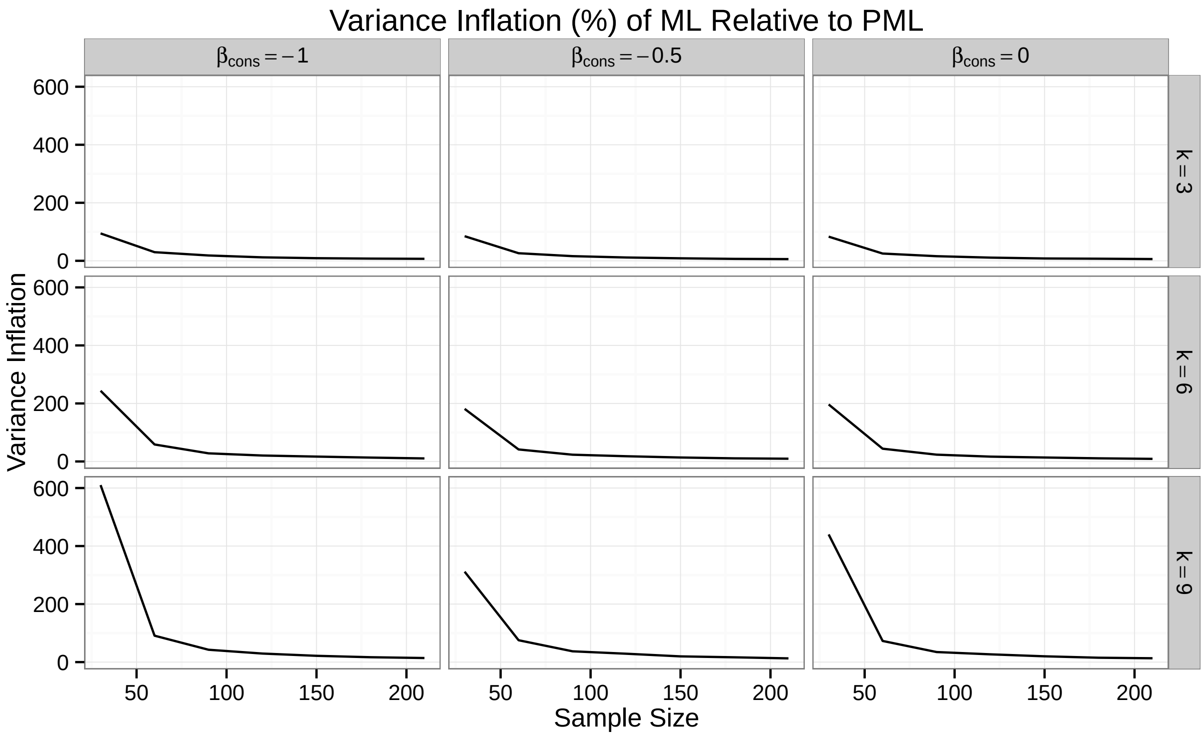

Figure 11: This figure shows the percent inflation in the variance of $\hat{\beta}^{mle}$ compared to $\hat{\beta}^{pmle}$.

C.4 Mean-Squared Error

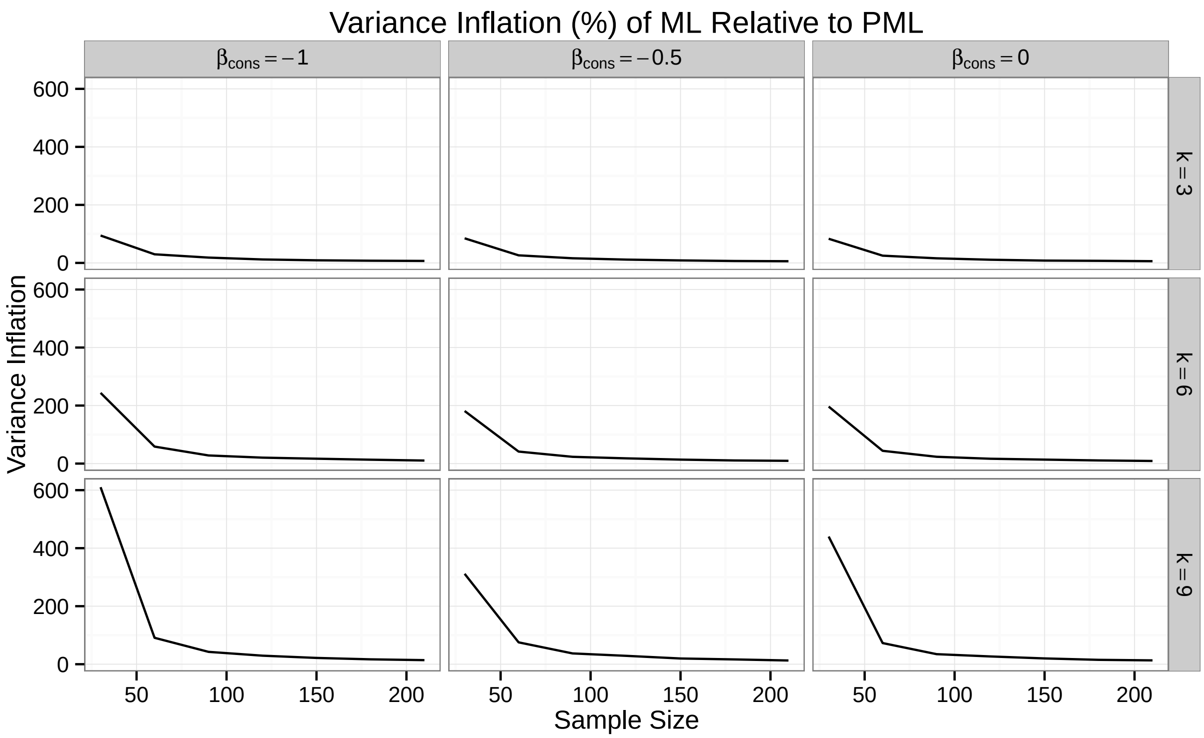

Figure 12: This figure shows the mean-squared error of $\hat{\beta}^{mle}$ and $\hat{\beta}^{pmle}$.