Estimating Logit Models with Small Samples*

Kelly McCaskey†

Carlisle Rainey‡

Abstract

In small samples, maximum likelihood (ML) estimates of logit model coefficients have substantial bias away from zero. As a solution, we introduce political scientists to the penalized maximum likelihood (PML) estimator (Firth 1993). 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 and the reductions are substantial in small samples. Monte Carlo simulations and a re-analysis of George and Epstein (1992) show that PML estimates significantly improve upon ML estimates.

Word Count: 8,011

*We thank Tracy George, Lee Epstein, and Alex Weisiger 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).

Logit and probit models have become a staple in quantitative political and social science–nearly as common as linear regression (Krueger and Lewis-Beck 2008). Political scientists typically rely on maximum likelihood (ML) to estimate these models. While ML estimates have excellent large-sample properties, the ML estimates of logit and probit model coefficients behave quite poorly in small samples.

The Big Problem with Small Samples

When working with a binary outcome y_i, the researcher typically models probability of an event, so that

$$\Pr(y_i) = \Pr(y_i = 1 \mid X_i) = g^{-1}(X_i\beta)\,, \tag{1}$$

where y represents a vector of binary outcomes, X represents a matrix of explanatory variables and a constant, β represents a vector of model coefficients, and g^{-1} represents some inverse-link function that maps $\mathbb{R}$ into $[0, 1]$. When g^{-1} represents the inverse-logit function $\text{logit}^{-1}(\alpha) = \dfrac{1}{1+e^{-\alpha}}$ or the cumulative normal distribution function $\Phi(\alpha) = \int_{-\infty}^{\alpha} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$, then we refer to Equation 1 as a logit or probit model, respectively. To simplify the exposition, we focus on logit models, because the canonical logit link function induces nicer theoretical properties (McCullagh and Nelder 1989, pp. 31-32). In practice, though, Kosmidis and Firth (2009) shows that the ideas we discuss applies equally well to probit models.

To develop the ML estimator of the logit model, we can 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, we 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 logit model coefficient vector $\hat{\beta}^{mle}$ is 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), and 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 of logit model coefficients 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."[1] This presents the researcher with a problem: When dealing with small samples, how can she obtain reasonable estimates of logit model coefficients?

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).[2] It turns out that this penalty is equivalent to Jeffreys' (1946) prior for the logit model (Firth 1993 and Poirier 1994). We take the natural logarithm of both sides to obtain the *penalized* log-likelihood function.

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

[1] 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.

[2] The statistics literature offers other approaches to bias reduction and correction as well. See Kosmidis (2014) for a useful overview.

Then the researcher can find the penalized maximum likelihood (PML) estimate $\hat{\beta}^{pmle}$ by finding the vector β that maximizes $\log L^*$. Zorn (2005) suggested PML for solving the problem of separation, but the broader and perhaps more important application to small sample problems seems to remain unnoticed in political science.

A researcher can implement PML as easily as ML, but PML estimates of logit model coefficients have a smaller bias (Firth 1993) and a smaller variance (Kosmidis 2007, p. 49, and Copas 1988).[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 illuminate the relationship between the ML and PML estimators. Suppose two estimators A and B, with a quadratic loss function so that the risk functions R^A and R^B (i.e., the expected loss) correspond to the mean-squared error (MSE). If $R^A \leq R^B$ for all parameter values and the inequality holds strictly for at least some parameter values, then we can refer to estimator B as *inadmissible* and say that estimator A *dominates* estimator B (DeGroot and Schervish 2012, p. 458, and Leonard and Hsu 1999, pp. 143-146). Now suppose a quadratic loss function for the logit model 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.

The intuition of the bias reduction is subtle. First, consider the source of the bias. Calculate the score function s as the gradient (or first-derivative) of the log-likelihood with respect to β so that $s(y, \beta) = \nabla \log L(\beta|y)$. (Note that solving $s(y, \hat{\beta}^{mle}) = 0$ is equivalent to finding $\hat{\beta}^{mle}$ that maximizes $\log L(\beta|y)$.) Now recall that at the true parameter vector β^{true}, the expected value of

[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.

the score function is zero so that $E\left[s(y,\beta^{true})\right] = 0$ (Greene 2012, p. 517). This implies that $E\left[s(y,\beta^{true})|s(y,\beta^{true}) > 0\right] = E\left[s(y,\beta^{true})|s(y,\beta^{true}) < 0\right]$–that high and low misses cancel exactly. However, if the score function s is curved in the area around β_j^{true} so that $s_j'' = \frac{\partial^2 s(y,\beta)}{\partial^2 \beta_j} > 0$, then $s(y,\beta^{true}) > 0$ generates large misses to the right of β_j^{true} and $s(y,\beta^{true}) < 0$ generates small misses to the left of β_j^{true}. Similarly, if $s_j'' < 0$, then $s(y,\beta^{true}) > 0$ generates small misses to the right of β_j^{true} and $s(y,\beta^{true}) < 0$ generates large misses to the left of β_j^{true}. The large and small misses do not cancel out, so that $E(\hat{\beta}_j^{mle}) > \beta_j^{true}$ when $s_j'' > 0$ and $E(\hat{\beta}^{mle}) < \beta_j^{true}$ when $s_j'' < 0$. Cox and Snell (1968, pp. 251-252) derive a formal statement of this bias of order n^{-1}, which we denote as $\text{bias}_{n^{-1}}(\beta^{true})$.

Now consider the bias reduction strategy. At first glance, one may simply decide to subtract the bias $\text{bias}_{n^{-1}}(\beta^{true})$ from the estimate $\hat{\beta}^{mle}$. However, note that the bias depends on the true parameter. Because researchers do not know the true parameter, this is not a viable strategy, though Anderson and Richardson (1979) explore the option of correcting the bias by using $\hat{\beta}^{mle} - \text{bias}_{n^{-1}}(\hat{\beta}^{mle})$. However, Firth (1993) suggests modifying the score function, so that $s^*(y,\beta) = s(y,\beta) - \gamma(\beta)$, where γ shifts the score function upward or downward. Firth (1993) shows that one good choice γ takes $\gamma_j = \frac{1}{2}\text{trace}\left[I^{-1}\left(\frac{\partial I}{\partial \beta_j}\right)\right] = \frac{\partial}{\partial \beta_j}\left(\log|I(\beta)|\right)$. Integrating, we can see that solving $s^*(y,\hat{\beta}^{pmle}) = 0$ is equivalent to finding $\hat{\beta}^{pmle}$ that maximizes $\log L^*(\beta|y)$ with respect to β.

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*–the PML estimator always outperforms the ML estimator, and least in terms of the bias, variance, and MSE. However, both 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 size of reductions in bias, variance, and MSE should draw the attention of substantive researchers, 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. 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.

3. The increased bias and variance of the ML estimator implies that the PML estimator will also have a smaller MSE, but importantly, the variance make a much greater contribution to the MSE than the bias.

In our simulation, the true data generating process corresponds to $\Pr(y_i = 1) = \frac{1}{1+e^{-X_i\beta}}$, where $i \in 1, 2, ..., n$ and $X\beta = \beta_{cons} + 0.5x_1 + \sum_{j=2}^{k} 0.2x_j$, and we focus on the coefficient for x_1 as the coefficient of interest. We draw each fixed x_j from a normal distribution with mean of zero and standard deviation of one and vary 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 P_{cons} from about 0.28 to 0.38 to 0.50). The biostatistics literature uses the number of events per explanatory variable $\frac{1}{k}\sum y_i$ as a measure of the information in the data set (e.g., Peduzzi et al. 1996 and Vittinghoff and McCulloch 2007), and each parameter of our simulation varies this information, where $\frac{N \times P_{cons}}{k} \approx \frac{1}{k}\sum y_i$. For each combination of the simulation parameters, we draw 10,000 data sets and use each data set to estimate the logit model coefficients using ML and PML. From these estimates, we compute 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 intercept β_{cons}, the number of explanatory variables k, and the sample size N vary. Figure 1 shows the results. The sample size varies across the horizontal-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 ML estimate has an upward bias of about 120%, while the PML estimate has an upward bias of only about 7%.[4]

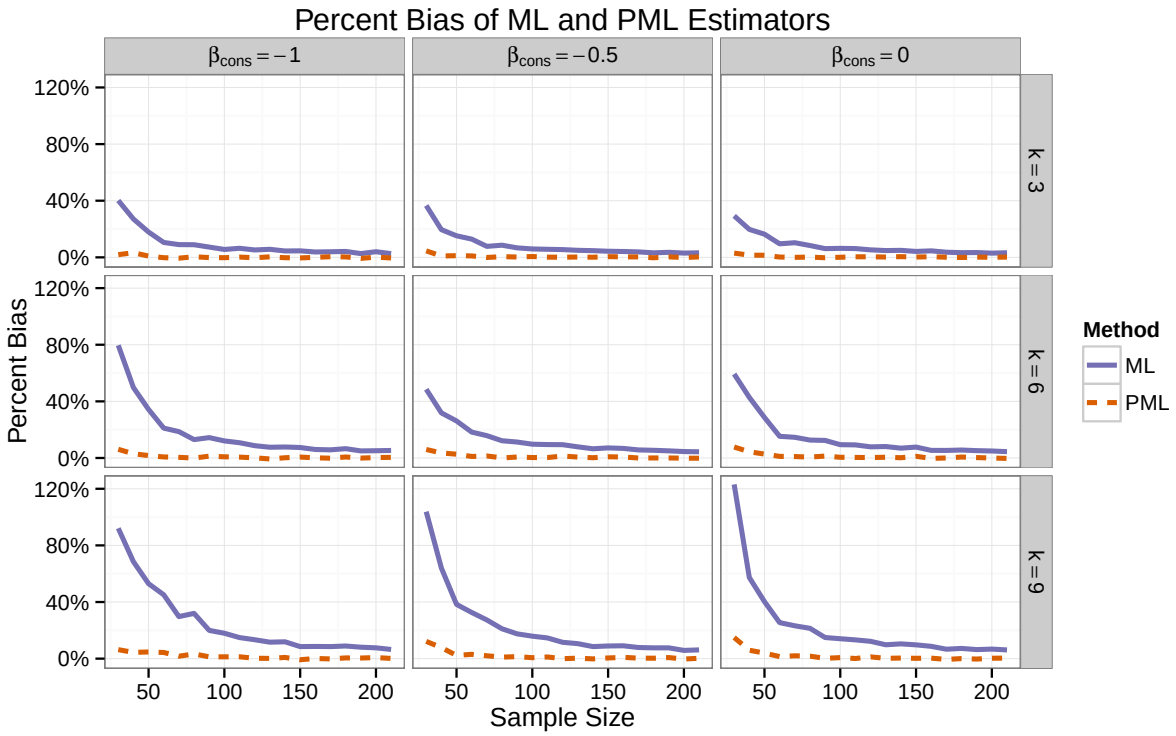

Figure 1: This figure illustrates the substantial bias of $\hat{\beta}^{mle}$ and the near unbiasedness of $\hat{\beta}^{pmle}$.

[4]Figures 8 and 9 in Section C of the Appendix show the expected value and (absolute) bias of these estimates.

Variance

In many cases, estimators trade off bias and variance, but the PML estimator reduces both. In addition to nearly eliminating the bias, Figure 2 shows that the PML estimator also substantially reduces the variance, especially for the smaller sample sizes. For $\beta_{cons} = -1$ and $N = 30$, the variance of the ML estimator is about 95%, 243%, and 610% larger than the PML estimator for 3, 6, and 9 variables, respectively. Doubling the sample size to $N = 60$, the variance remains about 30%, 58% and 91% larger, respectively. Even for a larger sample of $N = 210$, the variance of the ML estimator is about 7%, 10%, and 14% larger than the PML estimator.[5]

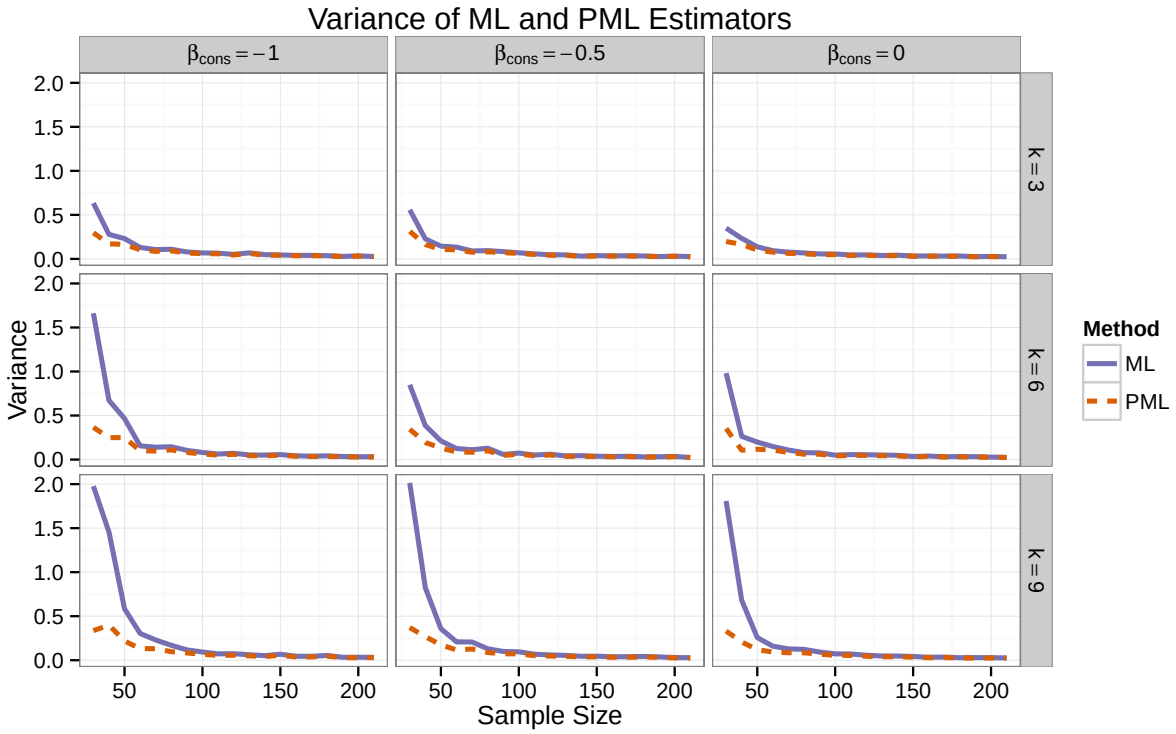

Figure 2: This figure illustrates the smaller variance of $\hat{\beta}^{pmle}$ compared to $\hat{\beta}^{mle}$.

[5]Figure 10 in the Appendix shows the variance inflation = $100 \times \left(\frac{Var(\hat{\beta}^{mle})}{Var(\hat{\beta}^{pmle})} - 1\right)$.

Mean-Squared Error

However, neither the bias nor the variance serves as a complete summary of the performance of an estimator. The MSE, though, combines the bias and variance into an overall measure of the accuracy, where

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

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

We care about the magnitude of this difference, though, not the sign. To summarize the magnitude of the difference, we compute the percent increase in the MSE for the ML estimator compared to the PML estimator. We refer to this quantity as the "MSE inflation," where

$$\text{MSE inflation} = 100 \times \frac{MSE(\hat{\beta}^{mle}) - MSE(\hat{\beta}^{pmle})}{MSE(\hat{\beta}^{pmle})} \,. \tag{3}$$

An MSE inflation of zero indicates that the ML and PML estimators perform equally well, but because the PML estimator dominates the ML estimator, the MSE inflation is strictly greater than zero. Figure 3 shows the MSE inflation for each combination of the parameter simulations on the $\log_{10}$ scale. Notice that for 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., $\beta_{cons} = 0$, $k = 3$, and $N = 210$). The MSE inflation exceeds 100% for about 13% of the simulation parameter combinations, 50% for about 24% of the combinations, and 25% for about 49% of the combinations. These large sacrifices in MSE should command the attention of researchers working with binary outcomes and small data sets.

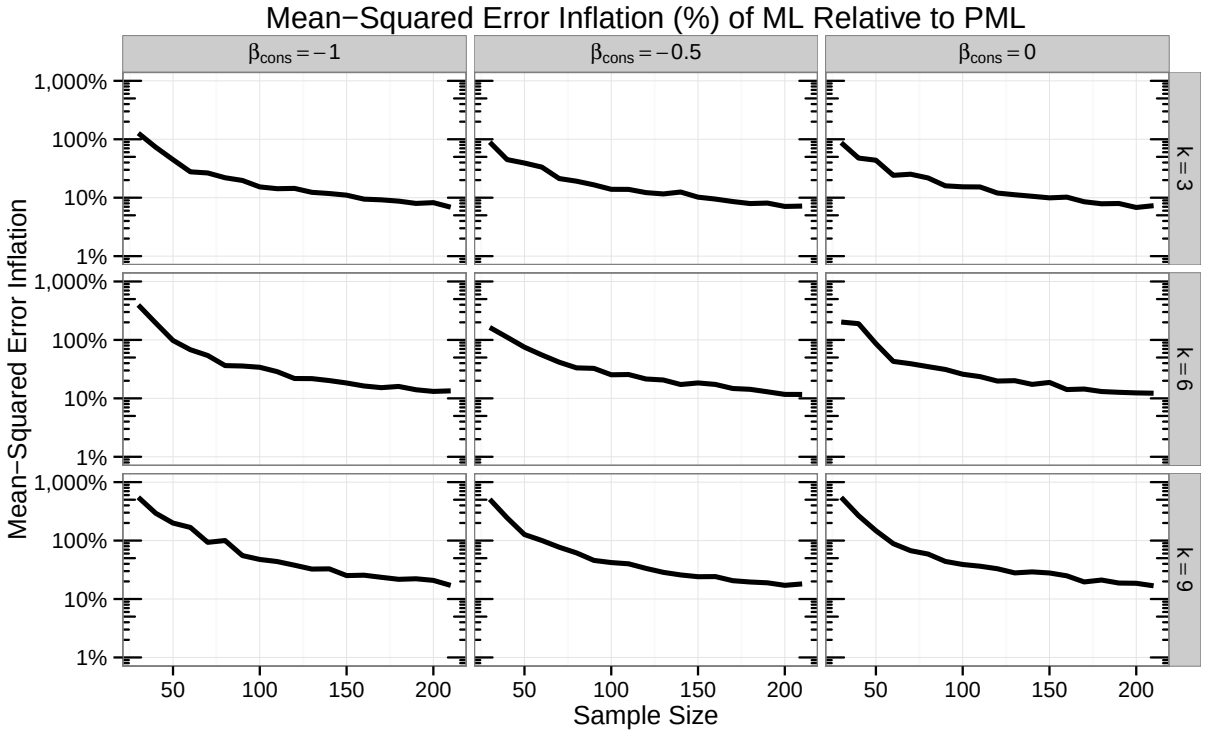

Figure 3: This figure shows the percent increase in the mean-squared error of $\hat{\beta}^{mle}$ compared to $\hat{\beta}^{pmle}$.

However, the larger bias and variance of the ML estimator do not contribute equally to the MSE inflation. Substituting Equation 2 into Equation 3 for $MSE(\hat{\beta}^{mle})$ and $MSE(\hat{\beta}^{pmle})$ and rearranging, we obtain

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

which additively separates the contribution of the bias and variance to the MSE inflation. If we wanted, we could 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

$$\text{relative contribution of variance} = \frac{\text{contribution of variance}}{\text{contribution of bias}} . \tag{4}$$

Figure 4 shows the relative contribution of the variance. 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 relative contribution of 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 logit model coefficients, the small sample *variance* is a more important problem to address, at least in terms of the accuracy of the estimator. Fortunately, the PML estimator greatly reduces the bias *and* variance, resulting in a much smaller MSE, especially for small samples.

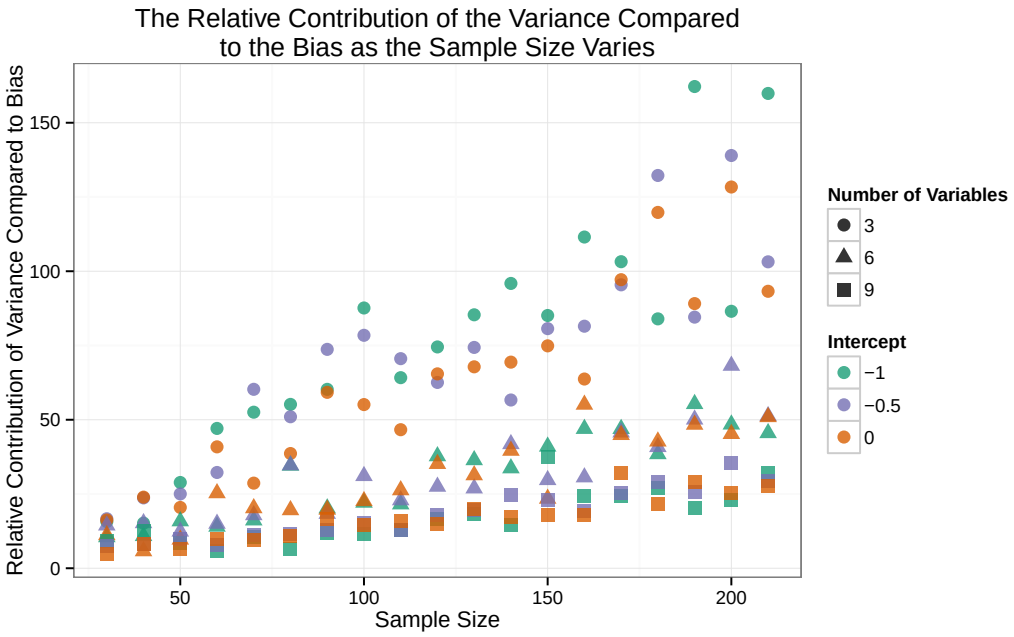

Figure 4: This figure shows the relative contribution of the variance and bias to the MSE inflation. The relative contribution is defined in Equation 4.

These simulation results show that the bias, variance, and MSE of the ML estimates of logit model coefficients are not trivial in small samples. Researchers cannot safely ignore these problems. Fortunately, researchers can implement the PML estimator with little to no added effort and obtain substantial improvements over the usual ML estimator. And these improvements are not limited to Monte Carlo studies. In the example application that follows, 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

To illustrate the substantive importance of using the PML estimator, we reanalyze a portion of the statistical analysis in George and Epstein (1992).[6] We re-estimate the integrated model of U.S. Supreme Court decisions developed by George and Epstein (1992) and find substantial differences in the ML and PML coefficient estimates and the confidence intervals.

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 posits *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. George and Epstein use a small sample of 64 Court decisions involving capital punishment from 1971 to 1988. The data set has only with 29 events (i.e., conservative decisions). They originally use ML to estimate their model and we reproduce their ML estimates exactly. For comparison, we also estimate the model with the PML estimator that we recommend. Figure 5 shows the coefficient estimates for each method. In all

[6] In Section D of the Appendix, we include a second re-analysis with similar results.

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 at 41%. Additionally, the PML estimator substantially reduces the width of all the confidence intervals. Three of the 11 coefficients lose statistical significance.

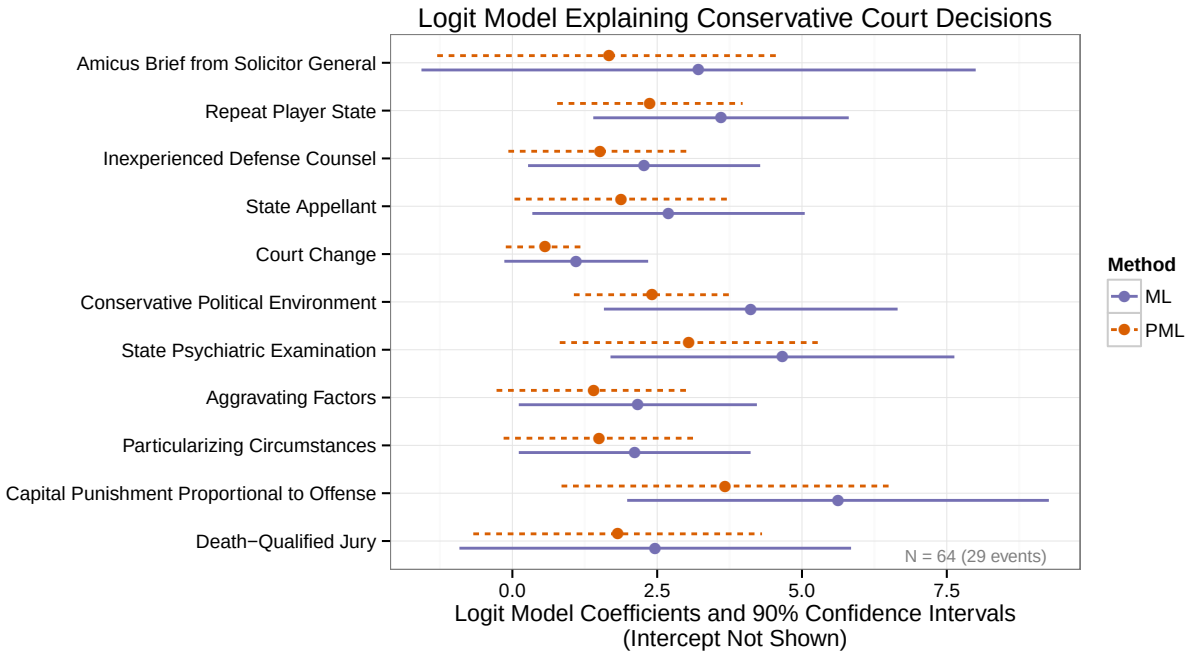

Figure 5: This figure shows the coefficients for a logit model model estimating U.S. Supreme Court Decisions by both ML and PML.

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 0.17, 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%

[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.

to 0.53. 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 estimate a logit model, we are likely more interested in the functions of the coefficients rather than the coefficients themselves (King, Tomz, and Wittenberg 2000). For an example, 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 medians. George and Epstein hypothesize that repeat players have greater expertise and are more likely to win the case. Figure 7 shows the estimates of the quantities of interest. The PML

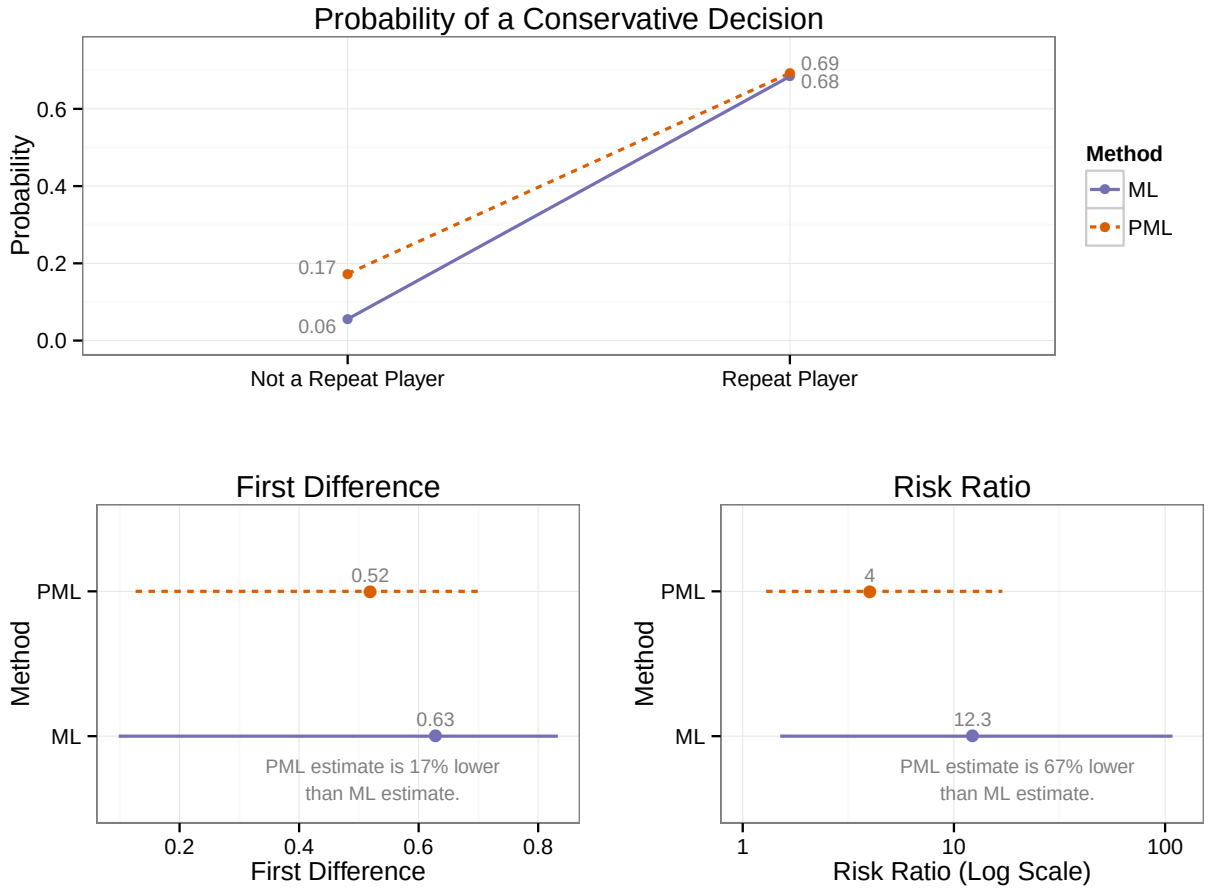

Figure 6: This figure shows the quantities of interest for the effect of the solicitor general filing a brief amicus curiae on the probability of a decision in favor of capital punishment.

estimator pools the estimated probabilities toward one-half. When the state is not a repeat player,

the PML estimates suggest a 17% chance of a conservative decision while ML estimates suggest only 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 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.

This example application clearly highlights the differences between the ML and PML estimators. The PML estimator shrinks the coefficient estimates and confidence intervals substantial. Theoretically, we know that these estimates have a smaller bias, variance, and MSE. Practically, though, this shrinkage manifests 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.

Recommendations to Substantive Researchers

Throughout this paper, we emphasize one key point–when using small samples to estimate logit and probit models, the PML estimator offers a substantial improvement over the usual maximum likelihood estimator. But what actions should substantive researchers take in response to our methodological point? In particular, at what sample sizes should researchers consider switching from the ML estimator to the PML estimator?

Concrete Advice About Sample Sizes

Prior research suggests two rules of thumb about sample sizes. First, Peduzzi et al. (1996) recommend about 10 events per explanatory variable, though Vittinghoff and McCulloch (2007) suggest relaxing this rule. Second, Long (1997, p. 54) suggests that "it is risky to use ML with samples smaller than 100, while samples larger than 500 seem adequate." In both of these cases, the al-

ternative to a logit or probit model seems to be no regression at all. Here though, we present the PML estimator as an alternative, so we have room to make more conservative recommendations. On the grounds that the PML estimator dominates the ML estimator, we might recommend that researchers always use the PML estimator. But we do not want or expect researchers to switch from the common and well-understood ML estimator to the PML estimator without a clear, meaningful improvement in the estimates.

We measure the cost of using ML rather than PML as the MSE inflation defined in Equation 3– the percent increase in the MSE when using ML rather than PML. The MSE inflation summarizes the relative inaccuracy of the ML estimator compared to the PML estimator.

To measure the information in a data set, we follow the biostatistics literature and use the number of events per explanatory variable $\frac{1}{k}\sum y_i$ (e.g., Peduzzi et al. 1996 and Vittinghoff and McCulloch 2007). However, this approach assumes that events occur less often than non-events. To generalize this idea to a statistic more suitable to political science data, in which the number of events commonly exceeds the number of non-events, we use a measure of information ξ that takes the minimum of the events and non-events per explanatory variable, so that

$$\xi = \frac{1}{k} \min \left[\sum_{i=1}^{n} y_i, \sum_{i=1}^{n} (1 - y_i) \right] . \tag{5}$$

We use a Monte Carlo simulation to develop rules of thumb that link the amount of information ξ to the cost of using ML rather than PML. The cost of using ML rather than PML decreases continuously with the amount of information in the data set, but to make concrete suggestions, we break the costs into three categories: substantial, noticeable, and negligible. We use the following cutoffs.

1. Negligible: If the MSE inflation probably falls below 3%, then we refer to the cost as *negligible*.

2. Noticeable: If the MSE inflation of ML probably falls below 10%, but not probably below

3%, then we refer to the cost as *noticeable*.

3. Substantial: If the MSE inflation of ML might rise above 10%, then we refer then the cost as *substantial*.

To develop our recommendations, we estimate the MSE inflation for a wide range of hypothetical analyses across which the true coefficients, the number of explanatory variables, and the sample size varies. To create each hypothetical analysis, we do the following:

1. Choose the number of covariates k randomly from a uniform distribution from 3 to 12.

2. Choose the sample size n randomly from a uniform distribution from 200 to 3,000.

3. Choose the intercept β_{cons} randomly from a uniform distribution from -4 to 4.

4. Choose the slope coefficients $\beta_1, ..., \beta_k$ randomly from a normal distribution with mean 0 and standard deviation 0.5.

5. Choose a covariance matrix Σ for the explanatory variables randomly using the method developed by Joe (2006) such that the variances along the diagonal range from from 0.25 to 2.

6. Choose the explanatory variables $x_1, x_2, ..., x_k$ randomly from a multivariate normal distribution with mean 0 and covariance matrix Σ.

For each hypothetical analysis, we simulate 1,000 data sets and compute the MSE inflation of the ML estimator relative to the PML estimator using Equation 3. We then use quantile regression to model the 90th percentile as a function of the information ξ in the data set. This quantile regression allows us to estimate the amount of information that researchers need to before the MSE inflation "probably" (i.e., about a 90% chance) falls below some threshold. We then calculate the thresholds at which the MSE inflation probably falls below 10% and 3%.

Interestingly, ML requires more information to estimate the intercept β_{cons} accurately relative to PML than the slope coefficients $\beta_1, ..., \beta_k$ (see King and Zeng 2001). Because of this, we calculate the cutoffs separately from the intercept and slope coefficients.

If the researcher simply wants accurate estimates of the slope coefficients, then she risks substantial costs when using ML with $\xi \leq 12$ and noticeable costs when using ML with $\xi \leq 51$. If the researcher also wants an accurate estimate of the intercept, then she risks substantial costs when using ML with $\xi \leq 33$ and noticeable costs when using ML when $\xi \leq 96$.

Importantly, the cost of ML only becomes negligible for all model coefficients when $\xi \geq 96$–this threshold diverges quite a bit from the prior rules of thumb. For simplicity, assume the researcher wants to include eight explanatory variables in her model. In the best case scenario of 50% events, she should definitely use the PML estimator with fewer than $\frac{8\times12}{0.5} = 192$ observations and ideally use the PML estimator with fewer than $\frac{8\times51}{0.5} = 816$ observations. But if she would also like accurate estimates of the intercept, then these thresholds increase to $\frac{8\times33}{0.5} = 528$ and $\frac{8\times115}{0.5} = 1,536$ observations. Many logit and probit models estimated using survey data have fewer than 1,500 observations and these studies risk a noticeable cost by using a ML estimator rather than the PML estimator. Further, these estimates assume 50% events. As the number of events drifts toward 0% or 100% or the number of variables increases, then the researcher needs even more observations.

Acceptable Inaccuracy	Slopes Coefficients	Intercept
Substantial	$\xi < 12$	$\xi < 33$
Noticeable	$12 \leq \xi < 51$	$33 \leq \xi < 96$
Negligible	$\xi \geq 51$	$\xi \geq 96$

Table 1: This table shows the thresholds at which the cost of ML relative to PML becomes substantial, noticeable, and negligible when estimating the slope coefficients and the intercept.

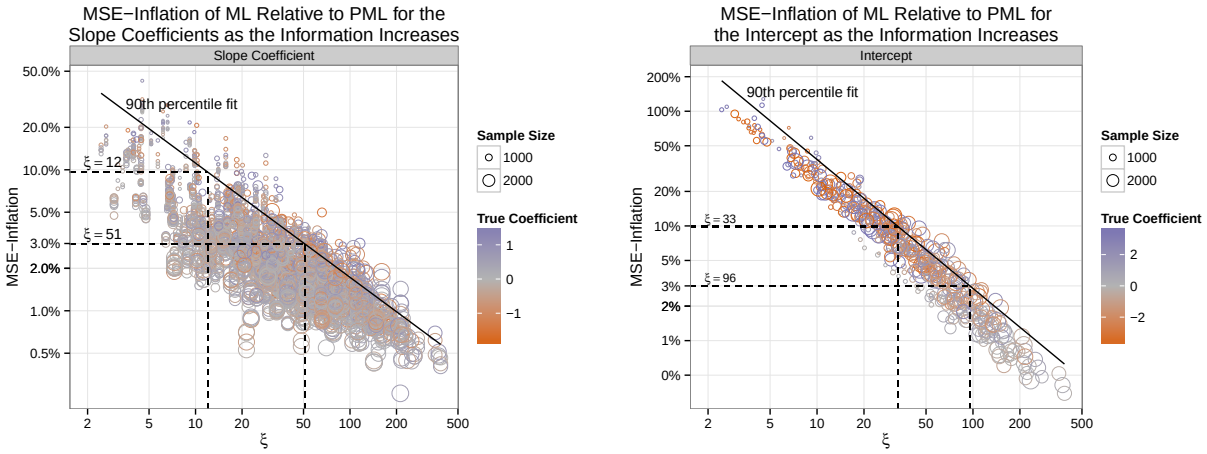

Figure 7: This figure shows the MSE inflation as the events per explanatory variable increases. The left panel shows the MSE inflation for the slope coefficients and the right panel shows the MSE inflation for the intercept.

Concrete Advice About Estimators

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 logit model 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 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 logit or probit 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

Anderson, J. A., and S. C. Richardson. 1979. "Logistic Discrimination and Bias Correction in Maximum Likelihood Estimation." *Technometrics* 21(1):71–78.

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.

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

Cox, D. R., and E. J. Snell. 1968. "A General Definition of Residuals." *Journal of the Royal Statistical Society, Series B* 30(2):248–275.

DeGroot, M. H., and M. J. Schervish. 2012. *Probability and Statistics*. 4th edition ed. Boston, MA: Wiley.

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.

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.

Joe, Harry. 2006. "Generating Random Correlation Matrices Based on Partial Correlations." *Journal of Multivariate Analysis* 97(10):2177–2189.

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.

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44(2):341–355.

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.

Krueger, James S., and Michael S. Lewis-Beck. 2008. "Is OLS Dead?" *The Political Methodologist* 15(2):2–4.

Leonard, Thomas, and John S. J. Hsu. 1999. *Bayesian Methods*. Cambridge Series in Statistical and Probabilistic Mathematics Cambridge: Cambridge University Press.

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

McCullagh, Peter, and John A. Nelder. 1989. *Generalized Linear Models*. Second ed. Boca Raton, FL: Chapman and Hall.

Peduzzi, Peter, John Concato, Elizabeth Kemper, Theodore R Holford, and Alvan R Feinstein. 1996. "A Simulation Study of the Number of Events per Variable in Logistic Regression Analysis." *Journal of Clinical Epidemiology* 49(12):1373–1379.

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

Vittinghoff, Eric, and Charles E. McCulloch. 2007. "Relaxing the Rule of Ten Events per Variable in Logistic and Cox Regression." *American Journal of Epidemiology* 165(6):710–718.

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

Estimating Logit Models 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 logit model 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 logit model 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 logit model 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

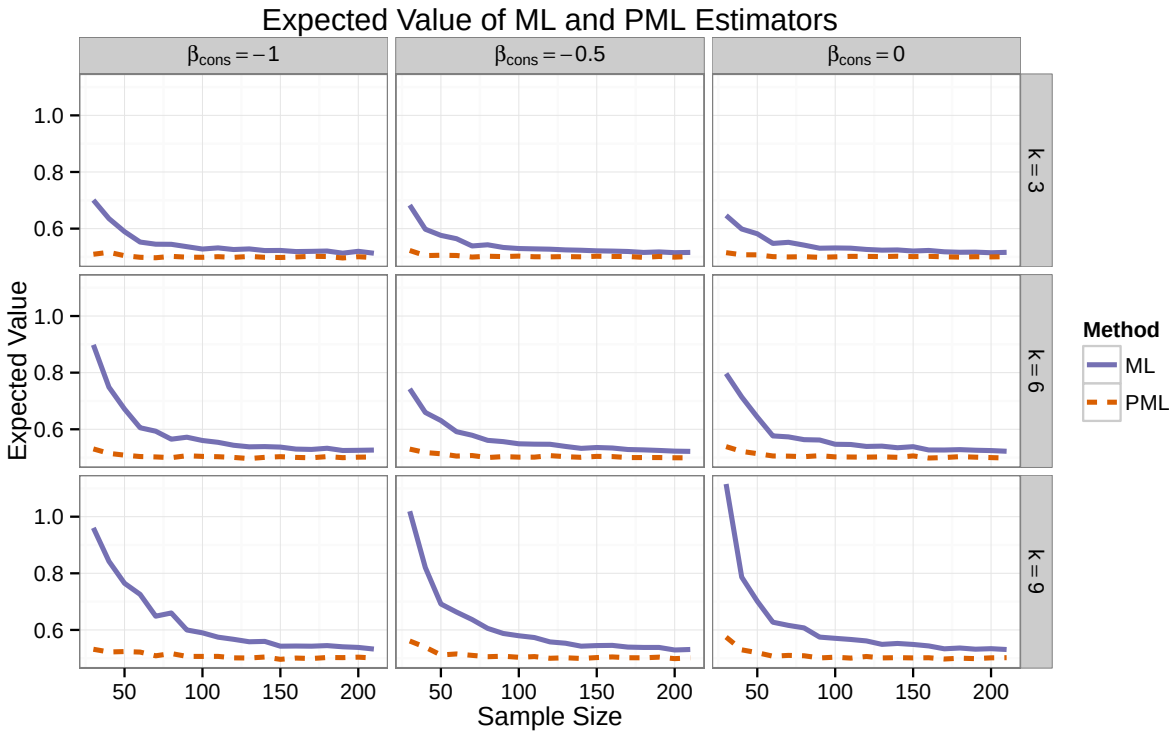

Figure 8: This figure shows the expected value of $\hat{\beta}^{mle}$ and $\hat{\beta}^{pmle}$. The true value is $\beta = 0.5$.

C.2 Bias

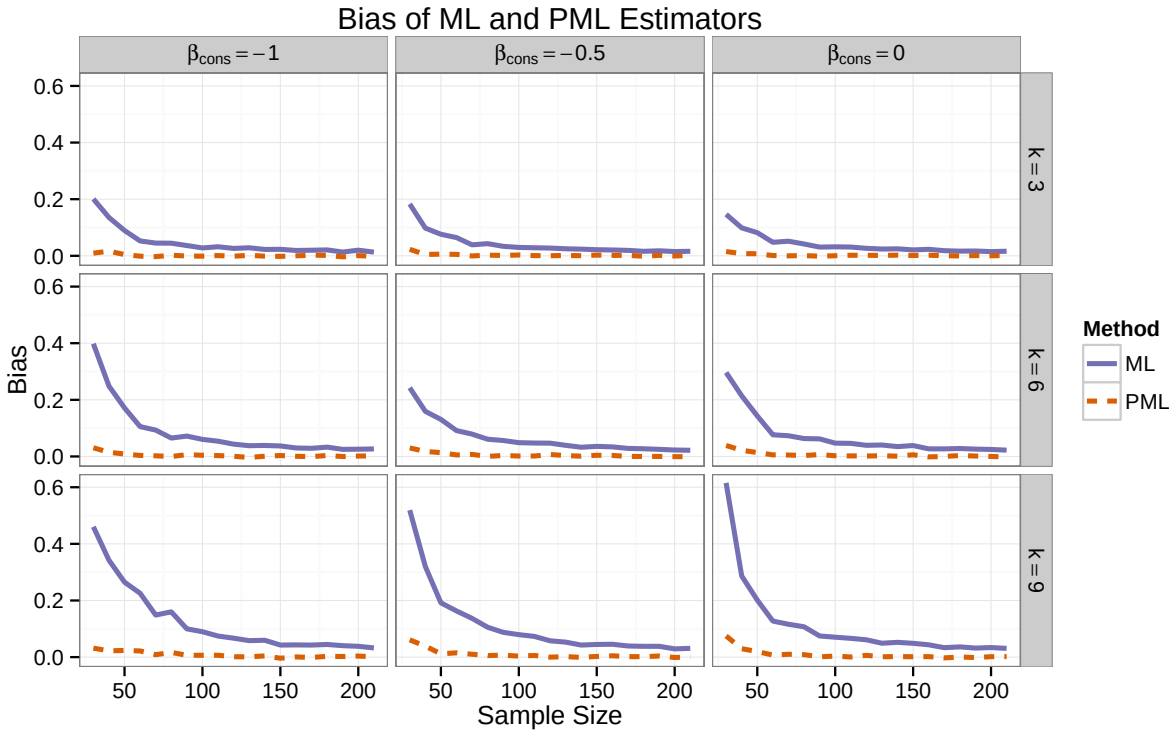

Figure 9: This figure shows the bias of $\hat{\beta}^{mle}$ and $\hat{\beta}^{pmle}$.

C.3 Variance Inflation

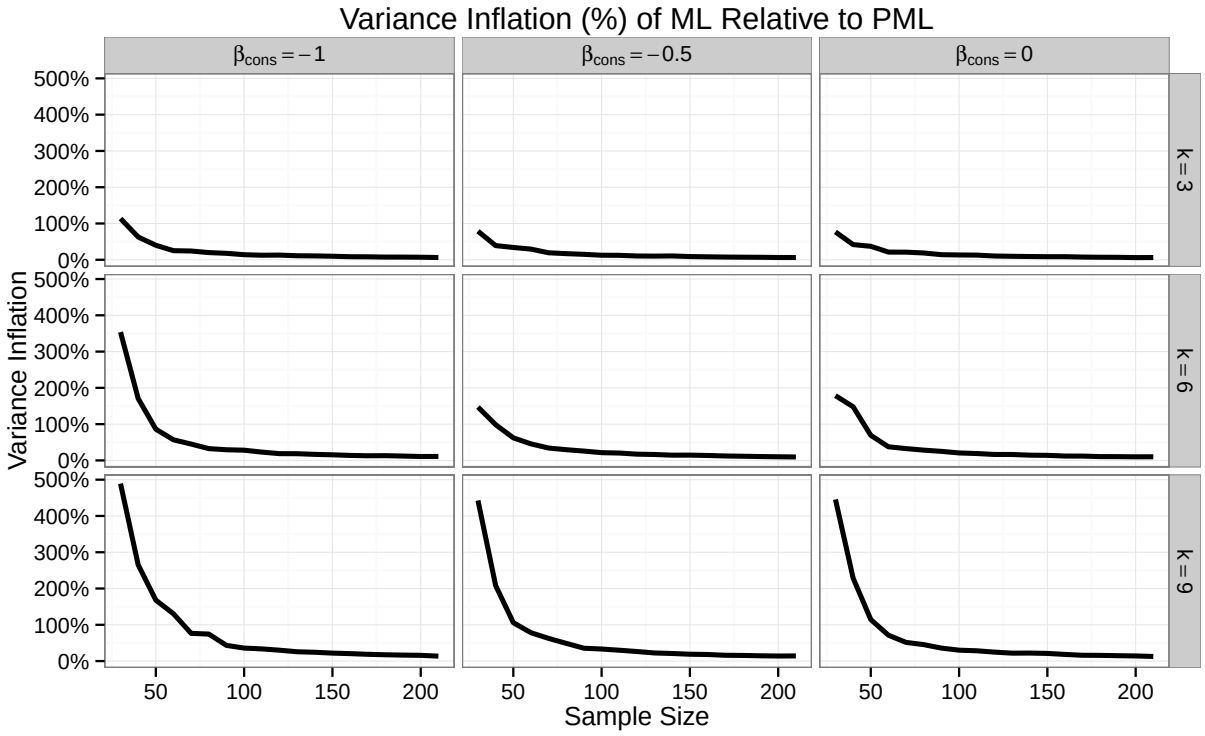

Figure 10: This figure shows the percent inflation in the variance of $\hat{\beta}^{mle}$ compared to $\hat{\beta}^{pmle}$.

D Re-Analysis of 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 a logit model to show the substantial difference between the biased, high-variance ML estimates and the nearly unbiased, low-variance PML estimates.[9] Prior to estimation, we stan-

[9]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 a logit model with small samples. The PML approach, though, solves the problem of bias as well as nonsense predictions.

dardize the continuous variables to have mean zero and standard deviation one-half and binary variables to have mean zero (Gelman 2008). Figure 11 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.

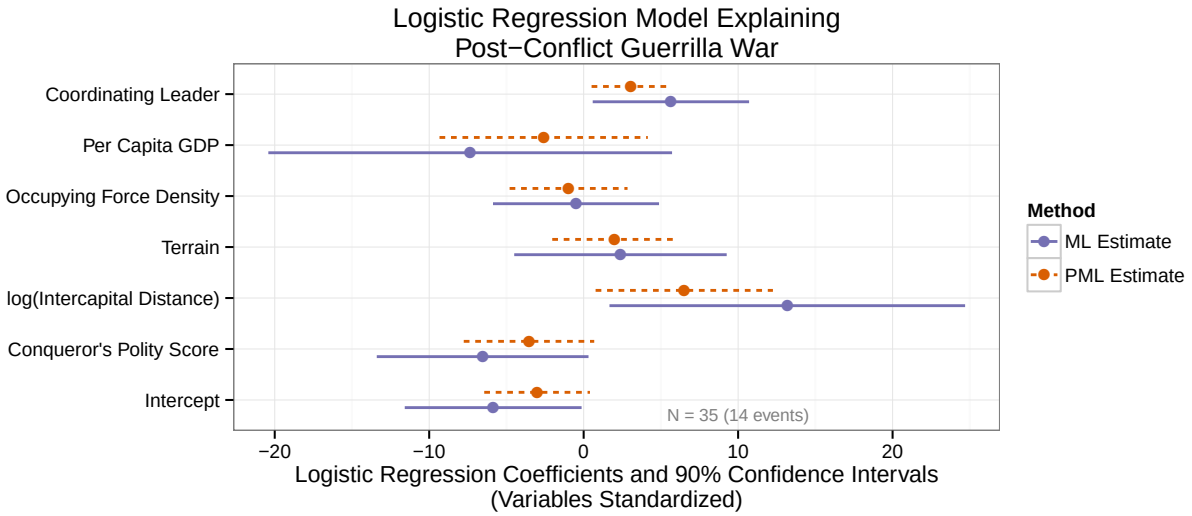

Figure 11: This figure shows the coefficients for a logit model model estimated explaining post-conflict guerrilla war estimated with ML and PML. Notice that the PML estimates and confidence intervals tend to be much smaller than the ML estimates and confidence intervals.

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.[10] 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 logit model, we might be more interested in *functions* of the coefficients

[10]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.

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 12 shows the estimates of the quantities of interest.

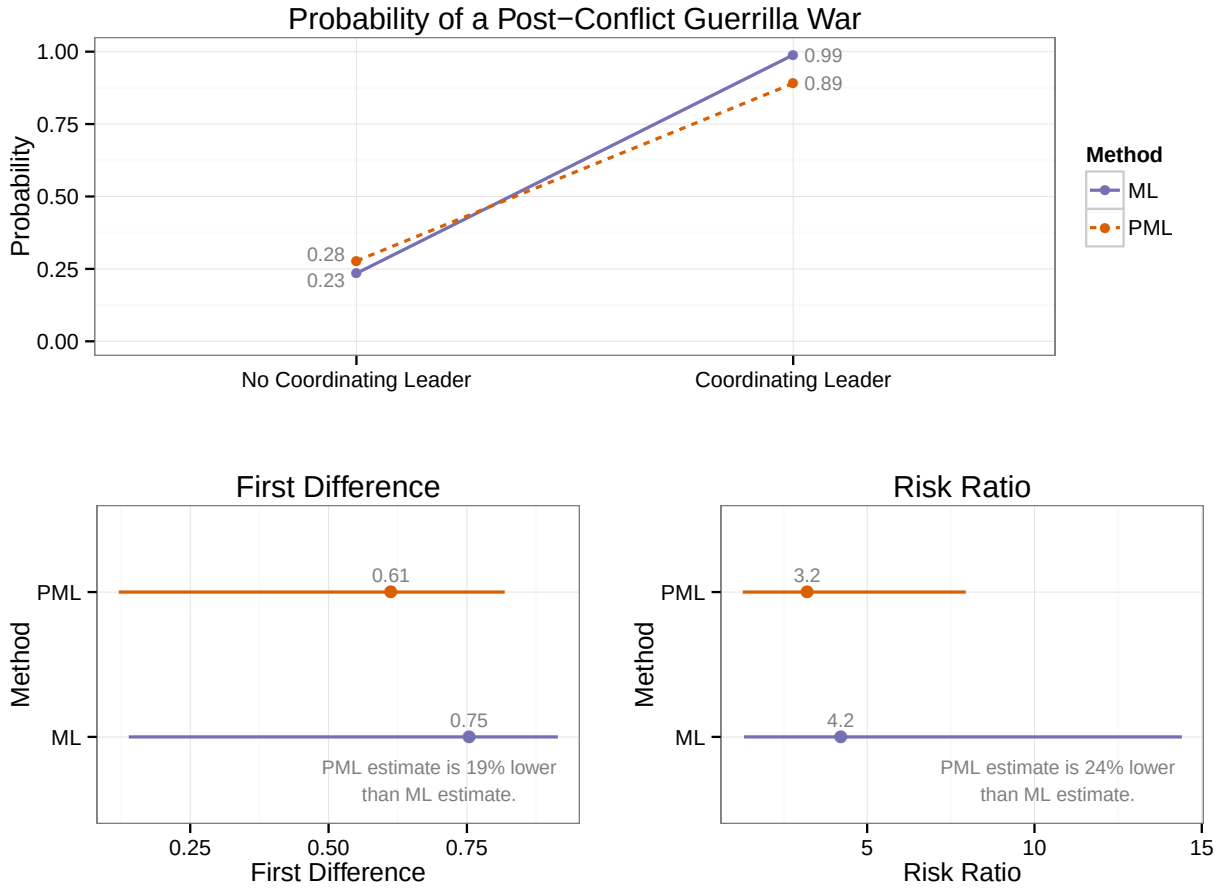

Figure 12: This figure shows the quantities of interest for the effect of a coordinating leader on the probability of a post-conflict guerrilla war.

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, 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.