Penalized Maximum Likelihood

(Nearly) Unbiased Estimation of Logistic Regression Coefficients in Small Samples[*]

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. For small samples, though, the asymptotic approximation does not work well. The sampling distribution of $\hat{\beta}^{mle}$ is not centered at β^{true}, so that $E(\hat{\beta}^{mle}) \not\approx \beta^{true}$. This presents the researcher with a problem: When dealing with small samples, how can she obtain reasonable estimates of logistic regression coefficients?

In the typical situation, the researcher models the probability of an event as $\Pr(y_i) = \Pr(y_i = 1 \mid X_i) = \frac{1}{1+e^{-X_i\beta}}$, where y is a vector of binary outcomes, X is a matrix of explanatory variables and an intercept, and β is a vector of model coefficients. Using this model, it is straightforward to calculate 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 calculate 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 $\beta \in \mathbb{R}^{k+1}$ that maximizes $\log L$. However, as noted above, this estimate is biased, so the $E(\hat{\beta}^{mle}) \neq \beta^{true}$.

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

$L^*(\beta|y) = L(\beta|y)|I(\beta)|^{\frac{1}{2}}$. [1] Taking logs yields the penalized log-likelihood function.

$$\log L^*(\beta|y) = \sum_{i=1}^{n} \left[y_i \log\left(\frac{1}{1+e^{-X_i\beta}} \right) + (1-y_i)\log\left(\frac{1}{1+e^{-X_i\beta}} \right) \right] + 0.5 \log|I(\beta)|.$$

Then the researcher can obtain the *penalized* maximum likelihood estimate $\hat{\beta}^{pmle}$ by finding the vector $\beta \in \mathbb{R}^{k+1}$ that maximizes $\log L^*$. Firth (1993) shows that $\hat{\beta}^{pmle}$ is much less biased than $\hat{\beta}^{mle}$. The penalized maximum likelihood estimate is easy to calculate in R using the `logistf()` function in the `logistf` package. See the Online Appendix for an example.

To demonstrate the substantial bias of $\hat{\beta}^{mle}$ and the near unbiasedness of $\hat{\beta}^{pmle}$, 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. The true data generating process is always $\Pr(y_i = 1) = \frac{1}{1+e^{-X_i\beta}}$, where $i \in 1, 2, ..., n$ and $X_i\beta = \beta_{cons} + 0.25x_1 + \sum_{j=2}^{k} 0.15x_j$. 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, the number of explanatory variables k, and the the intercept β_{cons} (which, in turn, varies the proportion of events). The expected values of the MLE and PMLE of the coefficient β_1 for x_1 are calculated by simulating 10,000 data sets, calculating the MLE and PLME for each data set, and then finding the average of the MLE and PMLE across the 10,000 data sets. Figure 1 shows the results. The sample size varies across the x-axes 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 40% (intercept equal to -1, 6predictors, 40 observations) to around 3% (50% events, 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 six variables, 40 observations, and an intercept of -1 (about 11 events), the percent bias in the PMLE is less than one percent–better than the best-case scenario for the MLE.

[1] It turns out that this penalty is equivalent to Jeffreys' (1946) prior for the logistic regression model (Firth 1993, Poirier 1994).

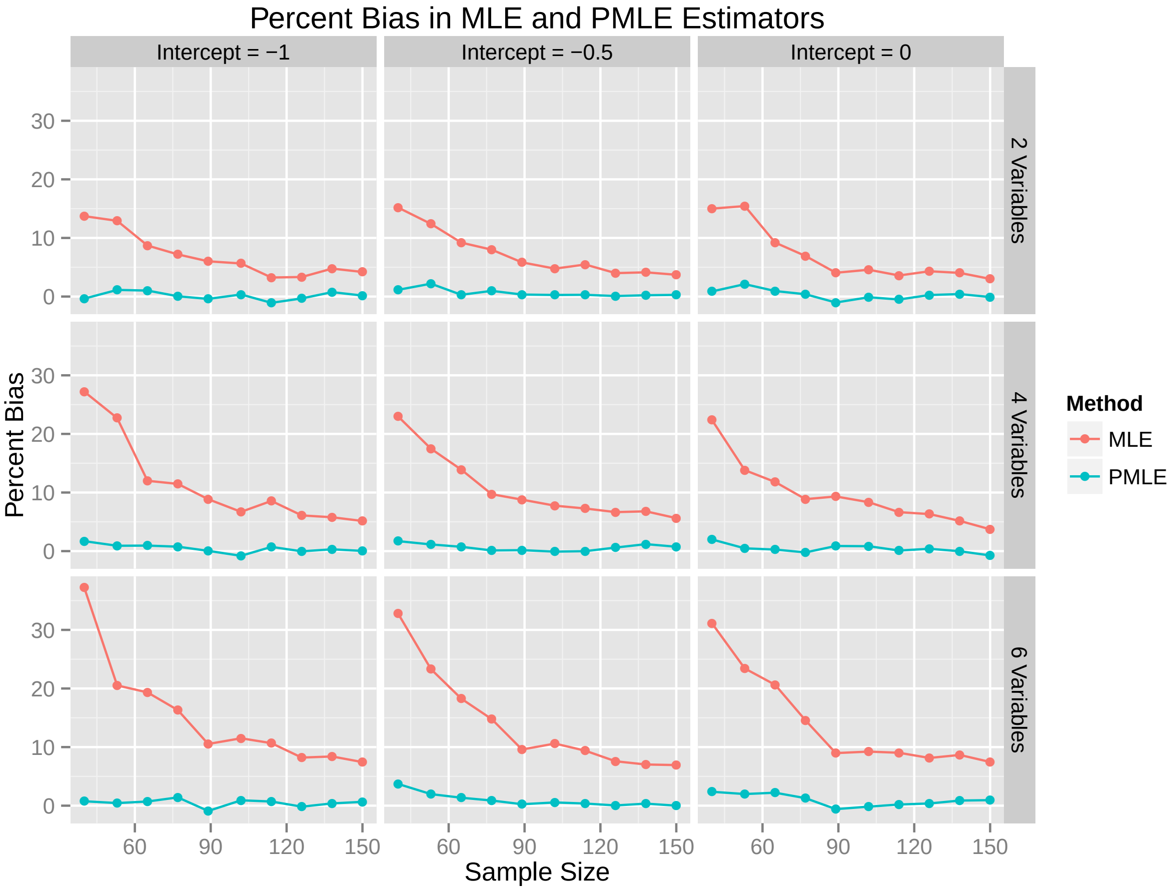

FIGURE 1: This figure illustrates the substantial bias of $\hat{\beta}^{mle}$ and the near unbiasedness of $\hat{\beta}^{pmle}$. Notice that when $N = 40$, the bias of $\hat{\beta}^{mle}$ away from zero is about 40% or more if events are relatively uncommon (e.g., $\beta_0 = -1$, which leads to about 28% events) or the researcher uses several explanatory variables (e.g., 6 or more). However, notice that $\hat{\beta}^{pmle}$ is essentially unbiased regardless of the sample size, frequency of events, or number of explanatory variables.

Replication of Weisiger (2014)

Weisiger (2014) explains that conquerors in conventional wars cannot expect to win over the defeated population. Additionally, he describes how sometimes violence continues after the official end of the war in the form of guerrilla warfare instead. Weisiger argues that if resistance occurs, it will be where conditions are favorable for insurgency, such as, difficult terrain, the

	MLE	PMLE	Percent Change
Constant	-8.20	-6.54	-20
	(5.63)	(4.73)	
Conqueror's Polity Score	-0.10	-0.07	-30
	(0.10)	(0.09)	
Distance	1.06	0.80	-24
	(0.71)	(0.60)	
Terrain	2.11	1.74	-18
	(2.49)	(2.23)	
Occupying Army Density	-0.07	-0.09	31
	(0.22)	(0.20)	
GDP per capita	-1.02	-0.58	-43
	(0.73)	(0.46)	
Coordination	2.19^*	1.72^*	-21
	(1.07)	(0.94)	

Notes: $^*p < 0.1$, standard errors in parentheses.

TABLE 1: MLE and PMLE Comparison

size and concentration of the occupying force, or if there remains a pre-war leader for potential insurgents to rally around. He hypothesizes that there will be a greater chance of resistance when the pre-conflict political leader "remains at large in the conquered country" (p. 8).

Weisiger's sample consists of 49 observations so the original analysis uses multivariate linear probability regressions to test his hypotheses. He does so in order to avoid problems with separation and biases that may occur with his small sample if he were to use the standard logit or probit analysis. He acknowledges the limitations to using multivariate linear probability regressions in this situation by noting that the method introduces the possibility of estimating non-meaningful predicted probabilities outside the [0, 1] range. We replicate his results and are able to correct the bias of the coefficient estimates using PMLE. Table 1 shows the coefficient estimates using MLE, PMLE, and the percent change between the two. Notice that MLE substantially over-estimates many of the coefficients 30 percent or more.

The use of a logit analysis implies that we are more interested in the functions of the coefficients rather than the coefficients themselves, so we calculated the first differences and the risk ratios for his hypothesis about the presence of the pre-conquest political leader. To do so, we evaluated the difference in the probability of resistance from a situation where there a pre-conquest leader did not remain to a situation where one did remain. We then set the Polity score of the Conqueror to 10, a full democracy, and the rest of the variables at their medians.[2] We did then calculate the first difference and the risk ratio for both the standard MLE and the PMLE.

The values of the quantities of interest of the PMLE decrease in comparison to those of the standard MLE. Using the standard MLE, resistance in a situation where the pre-conquest leader remains is 4.96 times more likely than where the pre-conquest leader does not remain. Using the PMLE, we see the risk ration decrease to 3.25. Similarly, we see the first differences for the same situational change also decrease when using the PMLE instead of the standard MLE. Using the PMLE, the likelihood of resistance in going from a scenario in which the pre-conquest leader remains to one in which none remain decreases by 35 percentage points. With the standard MLE, the likelihood decreases by 40 percentage points. We are then able to calculate our confidence intervals using Clarify or Clarify-like simulations. As with the change in coefficients previously demonstrated in Table 1, the changes in the quantities of interest when one uses the PMLE as opposed to the standard MLE, the PMLE corrects for the tendency for coefficients produced by the standard MLE to be biased away from zero.

Because we do not know the true model, we can use out-of-sample predictions as the best way to judge. Table 2 displays both the Brier and the log scores for the standard MLE and for the PMLE. Both scoring methods indicate that the PMLE is a better predictor than the standard MLE. The Brier score B is calculated as $B = \sum_{i=1}^{n}(y_i - p_i)^2$, 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 is calculated as $\sum_{i=1}^{n} log(r_i)$, where $r_i = y_i p_i + (1 - y_i)(1 - p_i)$.

[2] These include: the natural log of the intercapital distance, the type of terrain, the density of the occupying force and the conquered country's per capita GDP.

Estimation Technique	Brier Score	Log Score
MLE	0.173	-0.621
PMLE	0.168	-0.547

TABLE 2: This table compares the Brier and log scores of out-of-sample predictions using MLE and PMLE. Notice that PMLE out-performs MLE according to both measures.

By replicating Weisiger (2014), we demonstrate an appropriate way to use logistic regression with small samples. The original study uses multivariate linear probability regressions on a sample of 35 observations to avoid problems with separation and bias. Using a method of penalized maximum likelihood, with the associated risk ratios, first differences, and the Brier and log scores of out-of-sample predictions, we show that the author would be able to make a stronger case for his argument by obtaining more precise estimates by correcting the bias on the coefficients.

References

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

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.

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.

Online Appendix

Practical Advice for Logistic Regression with a Small Sample

A Using `logistf()` to Calculate the PMLE and Confidence Interval

B Using `logit.jeffreys()` to Obtain Posterior Simulations Using Jeffreys' Prior