Practical Advice for Logistic Regression with a Small Sample

Penalized Maximum Likelihood and Metropolis Sampling[*]

Kelly McCaskey[†]

Carlisle Rainey[‡]

Abstract

When used in small samples, logistic regression models present two problems. First, the estimates are noticeably biased away from zero, as much as 25 percent in plausible scenarios. Secondly, the sampling distributions are not well-approximated by a normal distributions, which makes computing confidence intervals challenging, especially for quantities of interest, such as first-differences or risk-ratios. As a solution to these problems, we first (re)introduce political scientists to Firth's penalty, which removes much of the bias from the usual estimator. Secondly, we explain how researchers can use a random walk Metropolis algorithm to obtain Clarify-like simulations from the posterior distribution to obtain confidence intervals for quantities of interest.

[*] We thank [people] for making their data available to us. The analyses presented here were conducted with R 3.1.0. All data and computer code necessary for replication are available at [GitHub url].

[†] 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 estimator for the coefficients of a logistic regression model is centered at the true coefficient and normally distributed. However, in small samples, the sampling distribution is neither centered at the true coefficient nor normally distributed. Estimating logistic regression models with small data sets leads to estimates that are substantially biased away from zero with sampling distributions that are skewed away from zero.

Thus, logistic regression with a small sample presents the researcher with two problems:

1. How can she correct the potentially substantial bias in the coefficient estimates?
2. How can she obtain accurate confidence intervals without relying on the assumption of normality, especially for quantities of interest?

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 bias (%) as the sample size, proportion of events, and number of explanatory variables vary.[2] Figure 1 shows the results. Across the range of the parameters of our sample, the bias of the MLE varies from about 50% (19% events, 9 predictors, 50 observations) to around 2% (50% events, 3 predictors, 200 observations). Any bias in the PMLE, on the other hand, is barely noticeable.

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

[2] 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 controls the size of the sample by varying n, the number of explanatory variables by varying k, and the proportion of events by varying β_{cons}. The expected values of the MLE and PMLE of the coefficient 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. The bias (%) is calculated the average estimate divided by the true value, which is 0.25 in our simulations.

figs/bias-sims.png

Figure 1: This figure demonstrates the substantial bias of $\hat{\beta}^{mle}$ and the near unbiasedness of $\hat{\beta}^{pmle}$. Notice that when $N = 50$, the bias of $\hat{\beta}^{mle}$ away from zero is about 40% or more if events are relatively uncommon (e.g., 19% events) or the researcher uses several explanatory variables (e.g., 9). However, notice that $\hat{\beta}^{pmle}$ is essentially unbiased regardless of the small size, proportion of events, and number of explanatory variables.

Accounting for the Skewed Sampling Distribution

The usual method of obtaining standard errors for maximum likelihood estimates is to assume a multivariate normal sampling distribution for the parameter vector β and obtain the (i, j) entry of the covariance matrix $I(\hat{\beta}^{mle})^{-1} = \Sigma^{[\hat{\beta}^{mle}]}$ by calculating the curvature around the maximum likelihood using $\frac{\partial^2 \log L(\beta|Y)}{\partial\beta_i\partial\beta_j}$. This approximation also works for penalized maximum likelihood, so that $\Sigma_{i,j}^{[\hat{\beta}^{pmle}]} = \frac{\partial^2 \log L^*(\beta|Y)}{\partial\beta_i\partial\beta_j}$. However, this method only works well for large samples–with small samples, the researcher must take into account the skewness in the sampling distribution (Barndorff-Nielson and Cox 1994).

If the researcher is only interested in confidence intervals for the logistic regression coefficients, then she can obtain these through the profile likelihood. See McCullagh and Nelder (1989) for the details. The profile likelihood confidence intervals for the parameters are easily produced using the `logistf()` function in the R package `logistic`. See the Online Appendix for an example.

However, researchers are not usually directly interested in logistic regression coefficients. Instead, they are interested in *functions* of the coefficients, such as first-differences and risk-ratios (King 1989 and King, Tomz, Wittenberg 2000). However, "Clarify-like" simulation (Tomz, Wittenberg, and King 2003) or "informal Bayesian" simulation (Gelman and Hill 2006) typically relies on the asymptotic posterior, which might be a poor approximation in small samples. However, rather than relying on an inaccurate normal approximation, researchers can use a Metropolis algorithm to directly sample from the actual posterior.

Notice that Firth's penalty (Firth 1993) is equivalent to Jeffreys' prior for the parameters of the logistic regression model (Jeffreys 1946 and Poirier 1994), so the $p(\beta) = I(\beta)|^{\frac{1}{2}}$. This implies that the penalized likelihood L^* is proportional to the posterior distribution for the coefficients, so that

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

The Metropolis algorithm (Metropolis et a. 1953 and Gelman et al. 2013) allows researchers to sample directly from this posterior distribution using Markov chains. She can construct each chain as follows:

1. Choose an arbitrary starting point $\tilde{\beta}^{[0]}$ to serve as the starting point for the sampler. We recommend random, over-dispersed starting values. This proves especially helpful when monitoring convergence using multiple chains run in parallel.
2. Choose a proposal distribution from which to generate samples. We recommend the multivariate normal distribution with variance $\Sigma^{[\hat{\beta}^{mle}]}$.
3. For each subsequent iteration i of the chain:
    a. Simulate a proposed sample $\tilde{\beta}^p \sim N(\tilde{\beta}^{[i-1]}, \Sigma^{[\hat{\beta}^{mle}]})$.
    b. Calculate the acceptance ratio $\alpha = \frac{L^*(tilde\beta^p|y)}{L^*(tilde\beta^{[0]}|y)} = \frac{p(\tilde{\beta}^p|y)}{p(\tilde{\beta}^{[0]}|y)}$.
    c. Sample a value u from a uniform distribution from zero to one. Set $\tilde{\beta}^{[i]} = \tilde{\beta}^p$ if $\alpha > u$. If $\alpha \leq u$, then set $\tilde{\beta}^{[i]} = \tilde{\beta}^{[i-1]}$.

Each chain eventually converges to the posterior distribution. However, this convergence might require a large number of iterations. In practice, we recommend that researchers (1) run multiple chains (either in parallel or sequentially), (2) discard the first half of the chain as a "burn-in" period, and (3) monitor convergence of the second-half using Gelman and Rubin's (1992) $\hat{R}$ statistic. If the starting values for each chain are over-dispersed, then an $\hat{R}$ near one (e.g., 1.02) indicates convergence. Once the researcher is satisfied that the chains have converged, she may combine the chains and treat the simulations as a large number of independent simulations from the posterior distribution and proceed as recommended by King, Tomz, and Wittenberg (2000). To make this simulation easy, we have written an R function `logit.jeffreys()` in the R package pppd available on GitHub. For an example, see the Online Appendix.

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 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-conquest political leader “remains at large in the conquered country" (p. 8).

Weisiger’s sample consists of 35 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 and account for the skewed sampling distribution by using the Metropolis algorithm to sample directly from the posterior distribution and using those simulations in penalized maximum likelihood estimation (PMLE).

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 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.[3] We did this for the standard MLE with asymptotic approximation, PMLE with asymptotic approximation, and PMLE with MCMC sampling. The resulting first differences and 90% confidence intervals, shown in Figure 2, demonstrate how PMLE with MCMC sampling provides an unbiased estimate and an appropriately skewed distribution. Using PMLE with MCMC sampling provides us with a p-value of 0.01, indicating that we can be more confident in our rejection of the null hypothesis than had we used a multivariate linear probability analysis.[4]

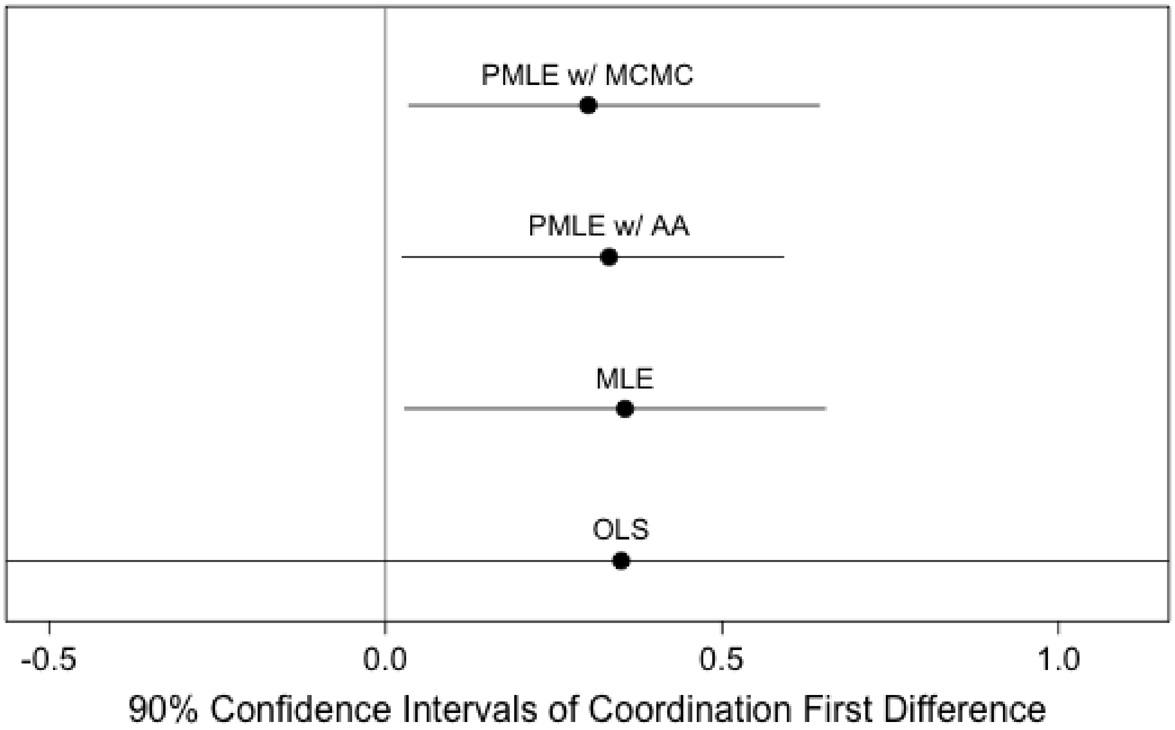

FIGURE 2: This figure demonstrates the first difference when we move from a scenario in which no pre-conquest leader remains to a situation in which the pre-conquest leader serves as a post-conquest rallying point, using a multivariate linear probability model, MLE, PMLE with asymptotic approximation, and PMLE with MCMC sampling.

[3] 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.

[4] PMLE with asymptotic approximation yields a p-value of 0.03, MLE with a p-value of 0.03, and OLS with a p-value of 0.38.

By replicating Weisiger (2014), we demonstrate an appropriate way to use logistic regression with small samples. His original study uses multivariate linear probability regressions on his sample of 35 observations to avoid problems with separation and bias. Using MCMC sampling directly from the posterior, 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 and the skew of the sampling distributions.

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