Logistic Regression in Small Samples

(Nearly) Unbiased Estimation with Penalized Maximum Likelihood*

Kelly McCaskey†

Carlisle Rainey‡

Abstract

When used in small samples, maximum likelihood estimates of logistic regression coefficients can be substantially biased away from zero. This bias might be 25 percent or more in plausible scenarios. As a solution to this problem, we (re)introduce political scientists to Firth's (1993) penalty, which removes much of the bias from the usual estimator. We use Monte Carlo simulations to illustrate that the penalized maximum likelihood estimation eliminates most of the bias, but also reduces the variance of the estimate. We illustrate the substantive importance of the penalized estimator with a replication of Weisiger (2014).

*We thank Alex Weisiger for making his data available to us. We conducted these analyses analyses with R 3.1.0. All data and computer code necessary for replication are available at github.com/kellymccaskey/small.

†Kelly McCaskey is a Ph.D. student in the Department of Political Science, University at Buffalo, SUNY, 520 Park Hall, Buffalo, NY 14260 (kellymcc@buffalo.edu).

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

Introduction

Asymptotically, the maximum likelihood (ML) estimator for the logistic regression coefficient vector $\hat{\beta}^{mle}$ is centered at the true value β^{true}, so that $E(\hat{\beta}^{mle}) \approx \beta^{true}$ when the sample is large. 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 $\Pr(y_i) = \Pr(y_i = 1 \mid X_i) = \dfrac{1}{1 + e^{-X_i\beta}}$, 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 (PML) 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` or `brglm()` packages. See the Online Appendix Section A for an example.

Monte Carlo Simulations

To demonstrate the substantial bias of $\hat{\beta}^{mle}$ and the near unbiasedness of $\hat{\beta}^{pmle}$, we conduct a Monte Carlo simulation comparing the sampling distribution of the ML and PML estimates. These simulations demonstrate three features of the ML and PML estimators:

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

In our simulation, the true data generating process is always $\Pr(y_i = 1) = \frac{1}{1+e^{-X_i\beta}}$, where $i \in 1,2,...,n$ and $X_i\beta = \beta_{cons} + 0.5x_1 + \sum_{j=2}^{k} 0.2x_j$. We consider β_1 as the coefficient of interest. Each fixed x_j is drawn from a normal distribution with mean of zero and standard deviation of one. The simulation varies the sample size n from 40 to 150, the number of explanatory variables $k \in \{2,4,6\}$, and the the intercept $\beta_{cons} \in \{-1,-0.5,0.0\}$ (which, in turn, varies the proportion of events from

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

about 28% to 38% to 50%). We simulate 10,000 data sets for each combination of the simulation parameters and estimate the logistic regression coefficient using ML and PML using each. We calculate the expected value and variance of the estimates by computing the mean and variance of the ML and PML estimates across the 10,000 data sets.

Bias

We calculate the percent bias = $100 \times \left(\frac{E(\hat{\beta})}{\beta^{true}} - 1 \right)$ as the sample size, proportion of events, and number of explanatory variables vary. Figure 1 shows the results. The sample size varies across the x-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.

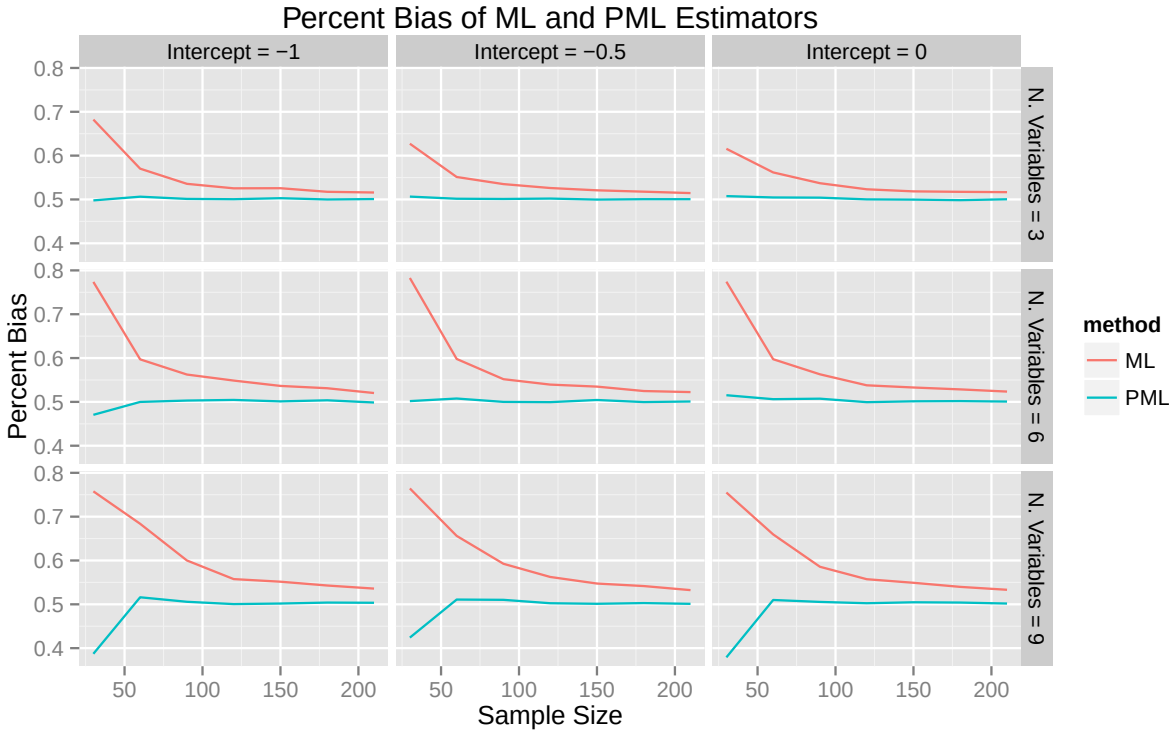

Figure 1: This figure illustrates the substantial bias of $\hat{\beta}^{mle}$ and the near unbiasedness of $\hat{\beta}^{pmle}$. Notice that when $N = 40$, the bias of $\hat{\beta}^{mle}$ away from zero is about 40% or more if events are relatively uncommon (e.g., $\beta_0 = -1$, which leads to about 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.

Variance

Replication of Weisiger (2014)

To illustrate the potential importance of using the nearly unbiased PML estimator, we replicate a portion of the statistical analysis in Weisiger (2014). Weisiger explains that conquerors in conventional wars cannot expect to win over the defeated population and he describes how sometimes violence continues after the official end of the war in the form of guerrilla warfare instead. Weisiger argues that resistance is more likely when 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. We focus on his hypothesis 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 35 observations (with 14 insurgencies). In the orginal analysis, Weisiger uses a linear probability model. He writes:

> For multivariate analysis I 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 Weisiger notes, predictions outside the $[0, 1]$ interval pose a problem for interpreting the linear probability model.[2] For example, in one case the linear probability model estimates a probability of 1.41 of insurgency. In another, it estimates a probability of -0.22. 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.

Coefficient Estimates

We re-analyze Weisiger's data using logistic regression to show the substantial difference between the biased ML estimates and the unbiased PML estimates. Figure 2 shows the coefficient estimates and 90% confidence intervals using ML and PML. Notice that the unbiased PML estimates are substantially smaller in many cases.

[2]We should also note that the PML estimates also solve the problem with separation. See **?** and **?**.

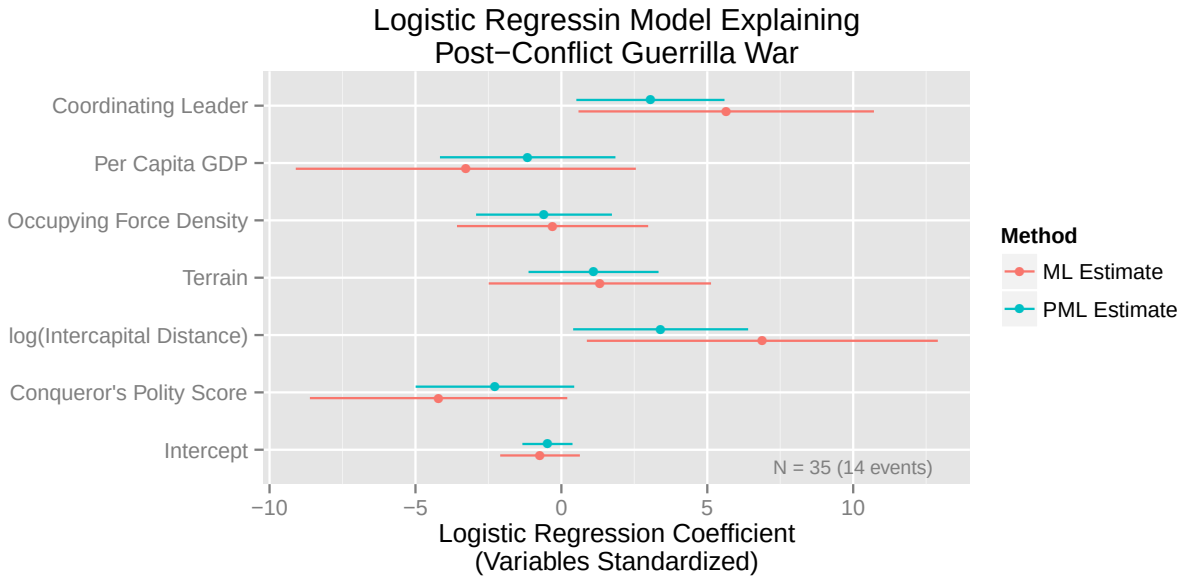

Figure 2: This figure shows the coefficients for a logistic regression model estimated explaining post-conflict guerrilla war estimated with ML and PML. Notice that the PML estimates tend to be much smaller than the ML estimates.

Figure 3 shows the percentage change from the ML estimator to the PML estimator. 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%.

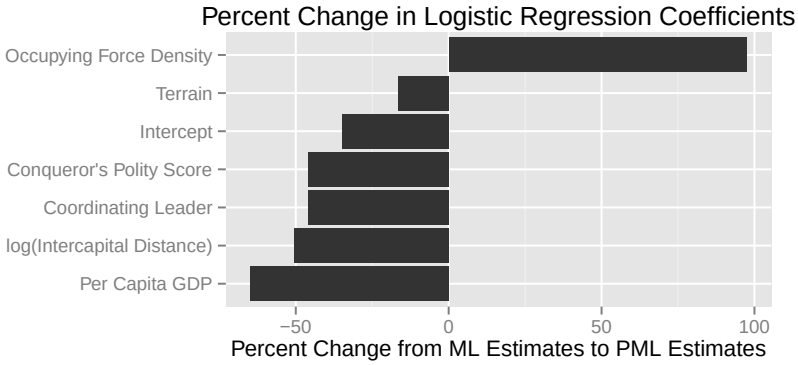

Figure 3: This figure shows the percentage change in the logistic regression coefficients in the model of post-conflict guerrilla war from ML to PML. Notice that these changes are substantial–the coefficient for per capita GDP shrinks by over 60% and the coefficient for occupying force density grows by nearly 100%. All the coefficients change by more than 45%, with the exception of terrain, which decreases by 16%.

Out-of-Sample Fit

Because we do not know the true model, we cannot know which of these sets of coefficient is best. However, we can use out-of-sample prediction to help adjudicate between these two methods. We use leave-one-out cross-validation and summarize the prediction errors using Brier and log scores. The procedure works as follows:

1. Choose an observation i to predict.
    a. Create a training data set ($N = 34$) by dropping observation i from the full data set ($N = 35$).
    b. Estimate the the logistic regression model using both ML and PML using the training data.
    c. Estimate the probability of an event (i.e., insurgency) for observation i. Store this estimate as p_i.
2. Repeat Step 1 for each observation in the full data set to obtain a full vector of out-of-sample predictions p.
3. Calculate the Frier score B 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 $y_i = 1$. This corresponds to the mean squared errors, so larger values indicate a worse fit.

4. Calculate 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.

Figure 4 shows the Brier and log scores for the two sets of estimates. PML estimates outperform the ML estimates with both approach to scoring.

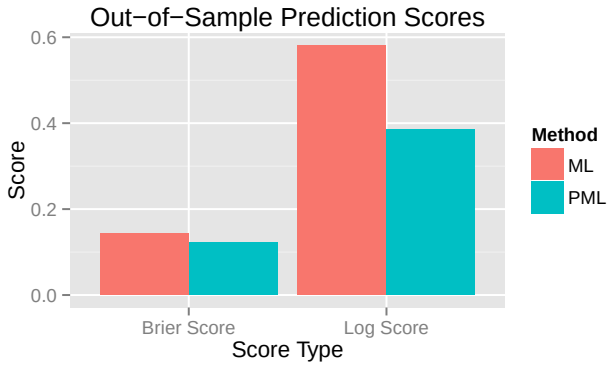

Figure 4: This figure shows the out-of-sample predictive ability (leave-on-out cross-validation) of the ML and PML estimators using both Brier and log scores to measure performance. Higher scores indicate better performance. Notice that the PML estimator out-performs the ML estimator according to both scores. This suggests that the ML estimator is overfitting the data.

Quantities of Interest

Because we are using a logistic regression, we might be more interested in *functions* of the coefficients rather than the coefficients themselves. 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 5 shows the estimates of the quantities of interest.

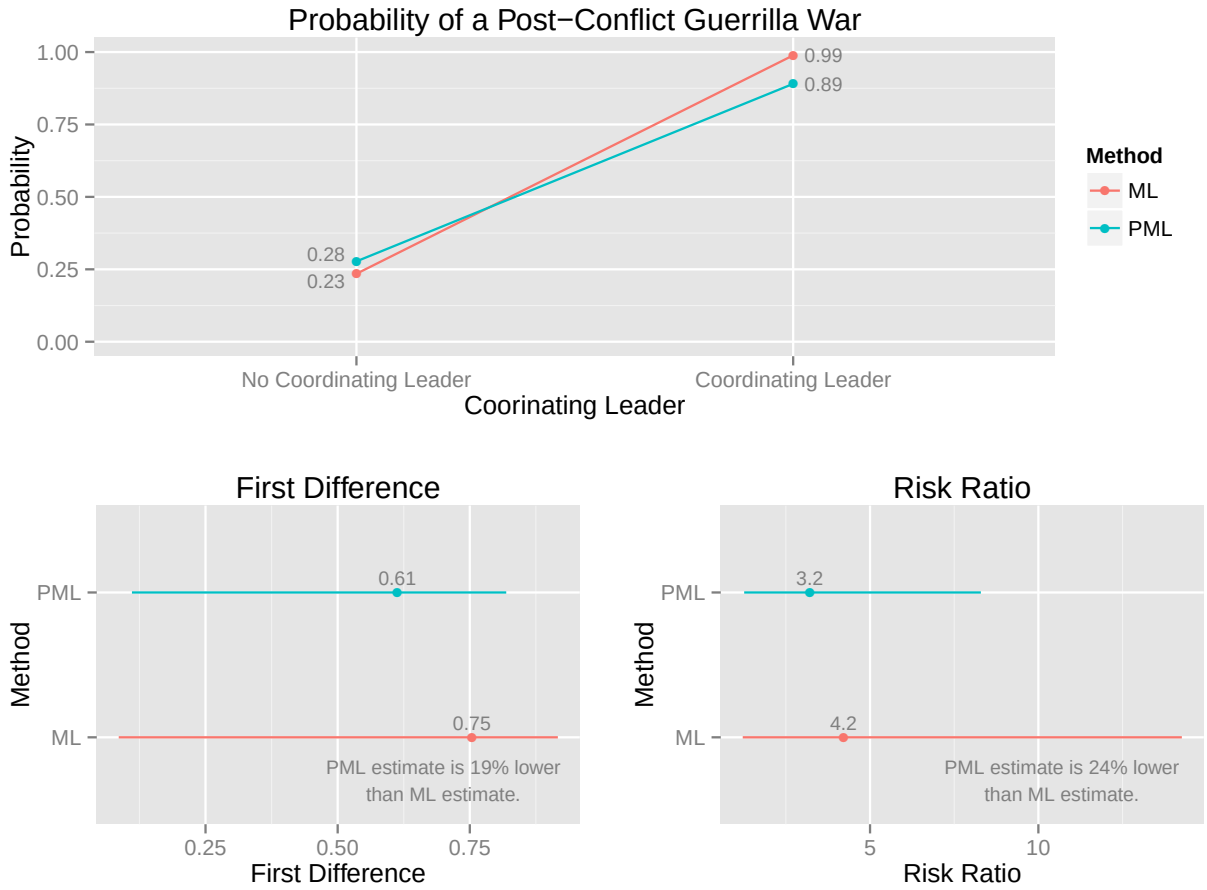

Figure 5: This figure shows the quantities of interest for the effect of a coordinating leader on the probability of a post-conflict guerrilla war. Notice that the PML estimates suggests 19% smaller first difference and a 24% smaller risk ratio.

PML pools the estimated probabilities toward zero. When a country lacks a coordinating leader, ML suggests a 23% chance of rebellion while PML suggests a 28% chance. On the other hand, when country *does have* a coordinating leader, ML suggests a 99% chance of rebellion of 0.99, but PML lowers this to 89%. Accordingly, PML suggests smaller effect sizes, whether using a first difference or risk ratio. PML shrinks the estimated first difference by 19% from 0.75 to 0.61. It shrinks the risk ration by 24% from 4.2 to 3.2.

By re-analyzing data from Weisiger (2014), we demonstrate an improved method of estimating logistic regression coefficients, especially with small samples. In small samples, maximum

likelihood is substantially biased upward, but *penalized* maximum likelihood is nearly unbiased. These reductions in bias do not come at a large cost. Though perhaps more subtle and theoretically difficult that maximum likelihood, penalized maximum likelihood is just as easy to use in practice.

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
Logistic Regression in Small Samples

A Penalized Maximum Likelihood Estimation of Logistic Regression Models in R

```
# load data
library(readr)  # for read_csv()
weisiger <- read_csv("weisiger-replication/data/weisiger.csv")

# quick look at data
library(dplyr)  # for glimpse()
glimpse(weisiger)

# model formula
f <- resist ~ polity_conq + lndist + terrain_alt +
  soldperterr + gdppc2_alt + 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)
```