p-Values Without Penalties With Perfect Predictions

Carlisle Rainey*

Draft under developement. It's no doubt filled with typos and errors. This is the version from September 9, 2019.

Separation commonly occurs in political science, usually when the presence (or absence) of a binary explanatory variable perfectly predicts the presence or absence of a binary outcome (e.g., Bell and Miller 2015; Mares 2015; Vining, Wilhelm, and Collens 2015). Under separation, maximum likelihood estimation leads to infinite coefficient estimates and standards errors. In practice, though, optimization routines converge before reaching infinite estimates and return implausibly large finite estimates and standard errors.

As an example of an implausible estimate, consider the model fit by Barrilleaux and Rainey (2014). For their application, the maximum likelihood estimates produced by the default `glm()` routine in R suggest that a governor like Deval Patrick, the Democratic governor of Massachusetts, had about a one in ten *billion* chance of opposing the Medicaid expansion under the Affordable Care Act. To give some perspective, this is *less* likely

*Carlisle Rainey is Associate Professor of Political Science, Florida State University, 540 Bellamy, Tallahassee, FL, 32306. (crainey@fsu.edu).

than you tossing 33 consecutive heads (around 1.2 in ten billion[1], you dealing a new poker player's first two hands as five-card straight flushes (around 2.4 in ten billion[2]), an average golfer making aces on their next two attempts on a par-3 hole (around 64 in ten billion[3]). It would take about ten billion years before a similarly situated Democratic governor would oppose the ACA–a little less than the age about the universe (about 13 billion years), but more than 30,000 *times* longer than *Homo sapiens* have existed (about 315,000 years) and about two million *times* longer than taxes have existed (about 5,000 years).

As a solution, Zorn (2005; see also Heinze and Schemper 2002) points political scientists toward the penalized maximum likelihood estimator proposed by Firth (1993). Even under separation, penalized maximum likelihood ensures finite estimates in theory and usually produces reasonably-sized estimates in practice. Conceptually, penalized maximum likelihood uses Jeffreys prior (Jeffreys 1946) to shrink the maximum likelihood estimates toward zero (Firth 1993).

Rainey (2016) points out that the parameter estimates (and especially the confidence intervals) depend largely on the chosen penalty. Indeed, other priors also guarantee finite, but different, estimates. For example, Gelman et al. (2008) recommend a Cauchy prior distribution. Rainey (2016) argues that the set of "reasonable" and "implausible" estimates depends on the substantive application, so context-free defaults (like Jeffreys and Cauchy) might not produce reasonable results. Rainey (2016) concludes that "[w]hen facing separation, researchers must *carefully* choose a prior distribution to nearly rule out

[1]The probability of tossing 33 consecutive heads equals $\left(\frac{1}{2}\right)^{33} \approx 1.16 \times 10^{-10}$. If you tossed one coin per second for 24 hours, then you could complete 2,618 33-toss trials in one day. To obtain an all-head 33-toss sequence, you would need to continue this routine daily for about 10,000 years.]

[2]There are $\binom{52}{5} = 2,598,960$ ways to draw five cards from a 52-deck. There are ten straight flushes per suit and four suits, so there are $10 \times 4 = 40$ total royal flush possibilities. The chance of a straight flush is then $\frac{40}{2,598,960} \approx 1.54 \times 10^{-5}$. The chance of two consecutive straight flushes is twice this chance. If two poker players sat down with a dealer that shuffled and dealt two five-card hands per minute, those players would need to play for about 8,000 years before the dealer dealt both a straight flush.]

[3]The chance that an average golfer makes an ace is about one in 12,500. The chance of two consecutive aces is twice that. If an average golfer played a typical nine-hole course every day, it would take about 430,000 years before they would ace both par-3 holes.

implausibly large effects" (p. 354).

But researchers cannot access useful prior information in all contexts, and some scholars prefer to avoid injecting prior information into the model. How can researchers proceed in these situations? Below, I show that while maximum likelihood produces implausibly large estimates under separation and standard errors, standard likelihood ratio tests behave in the usual manner. As such, researchers can produce meaningful p-values with a standard, well-known tool even while eyeing the coefficient estimates with suspicion.

Statistial Theory

Maximum likelihood provides a general and powerful framework for obtaining estimates of regression models. In our case of logistic regression, we write the probability π_i that an event occurs for observation i (or that the outcome variable $y_i = 1$) as

$$\pi_i = \text{logit}^{-1}(X_i\beta) \text{ for } i = 1, 2, ..., n, \tag{1}$$

where X is a matrix of covariates and β is a vector of regression coefficients. To obtain the likelihood function, simply compute the product of the probabilities of each y_i. If $y_i = 1$, then this probability equals π_i. If $y_i = 0$, then this probability equals $1 - \pi_i$ Using some clever algebra, the probability of each y_i is $p_i^{y_i}(1 - p_i)^{(1-y_i)}$. We refer to this function as the "likelihood function," so that

$$L(\beta|y) = p_i^{y_i}(1 - p_i)^{(1-y_i)}, \text{ where } \pi_i = \text{logit}^{-1}(X_i\beta) \tag{2}$$

To obtain the maximum likelihood estimates $\hat{\beta}^{ML}$, we simply find the maximum of the likelihood function with respect to β. Thus, we use as our estimate of β the values that would most likely generate the observed data.

In practice, though, we typically work with the log-likelihood function. For convenience, I denote the log-likelihood function as ℓ. In this case, $\ell(\beta|y) = \log L(\beta|y) = y_i \log(p_i) + (1 - y_i) \log(1 - p_i)$.

To obtain the maximum likelihood estimates, we use numerical algorithms to locate the value of β that maximizes ℓ.

But with the maximum likelihood estimates in hand, researchers typically want information about the precision of those estimates. In some cases, researchers want to compare their research hypothesis H_R to a null hypothesis H_0. To conduct a hypothesis test in the context of logistic regression, the research composes a null hypothesis $H_0 : \beta \in B_0 \subset R^n$, which leaves the research hypothesis $H_R : \beta \in B_0^C$. Depending on the data,the researcher may then choose to reject H_0 in favor of H_R or fail to distinguish between the two.

To fix ideas, suppose the simple point null hypothesis $H_0 : \beta_1 = 0$.

In order to assess the plausiblitity of the null hypothesis, we must compare the null hypothesis with the maximum likelihood estimates, accounting for the precision of the estimates.

The precision follows from the shape of the (log-)likelihood function. If small changes in β lead to large changes in the likelihood function, then we can take the maximum likelihood estimates as precise. However, if large changes in β lead to small changes in the likelihood function, then we must treat the estimates as imprecise.

The methodology literature offers two common tools to formally compare the the null hypothesis to the maximum likelihood estimates.

Wald Test

First, the Wald test quantifies the curvature of ℓ at $\hat{\beta}^{ML}$. If ℓ descends rapidly away from $\hat{\beta}^{ML}$, then we take $\hat{\beta}^{ML}$ as a precise estimate. A second derivative intuitively quantifies the notion of "curvature," and it turns out that

$$\widehat{\text{Var}}(\beta) = \left(-\frac{\partial^2 \ell(\hat{\beta}^{ML}|y)}{\partial \hat{\beta}^{ML} \partial \left[\hat{\beta}^{ML}\right]'} \right)^{-1}, \tag{3}$$

so that

$$\widehat{\text{SE}}(\hat{\beta}_i^{ML}) = \left(-\frac{\partial^2 \ell(\hat{\beta}_i^{ML}|y)}{\partial^2 \hat{\beta}_i^{ML}} \right)^{-\frac{1}{2}}. \tag{4}$$

The curvature of the log-likelihood functions provides a direct method to estimate the standard error of the maximum likelihood estimates.

For large (repeated) samples, the maximum likelihood estimates follow a normal distribution centered at the true value of β with a standard deviation of $\widehat{\text{SE}}(\hat{\beta}_i^{ML})$ from Equation 4.

Using this large-sample approximation, we can perform a z-test for our H_0.

$$\text{Wald } p\text{-value} = \Pr(|z| > 1.65) = 2\Phi(|z|), \text{ where } z = \frac{\hat{\beta}_i^{ML}}{\widehat{\text{SE}}(\hat{\beta}_i^{ML})}. \tag{5}$$

Following the usual procedure in political science researcher, if the p-value is less than 0.05, the researher rejects the null hypothesis (that $\beta_i = 0$, in this case) in favor of the research hypothesis (that $\beta_i \neq 0$, in this case). if the p-value is greater than 0.05, then the research cannot distinguish between the two hypotheses.

For the simple null hypothesis H_0 that $\beta_1 = 0$ (contraint in only one dimmension), we have

$$\text{likelihood ratio } p\text{-value} = \Pr(D > 1.65) = 2\Phi(|z|), \text{ where } z = \frac{\hat{\beta}_i^{ML}}{\widehat{\text{SE}}(\hat{\beta}_i^{ML})}. \tag{6}$$

Likelihood Ratio Test

Rather than use the precision of the maximum likelihood estimates to test the null hypothesis against the research hypothesis, the Likelihood Ratio test compares the value of $\ell(\hat{\beta}^{ML}|y)$ to $\ell(\hat{\beta}^{ML_0}|y)$, where $\hat{\beta}^{ML_0}$ represents the maximum likelihood estimates contrained to be consistent with the null hypothesis. For the simple null hypothesis H_0 that $\beta_1 = 0$, we can simple fit a separate model without the explanatory variable x_1.

If the data are much more likely under maximum likelihood estimates $\hat{\beta}^{ML}$ than under the contrained maximum likelihood estiamtes $\hat{\beta}^{ML_0}$, the researcher can reject the null hypothesis. Wilk's theorem advises us how to compare the two likelihoods. Wilk's theorem notes that $D = 2 \times \left[\ell(\hat{\beta}^{ML}|y) - \ell(\hat{\beta}^{ML_0}|y)\right]$ follows a χ^2 distribution with degrees of freedom equal to the number of contrained dimmensions.

Illustrations

To illustrate the simplicity of hypothesis testing cunder separation compared to estimation, I reanalyze data from Barrilleaux and Rainey (2014) and Bell and Miller (2015) the that Rainey (2016) considers in great detail. For Barrilleaux and Rainey (2014), the likelihood ratio test provides a useful evaluation of their substantive claims. For Bell and Miller (2015) the likelihood ratio test proves less useful.

Barrilleaux and Rainey (2014)

Barrilleaux and Rainey (2014) examine U.S. state governors decisions to support or oppose the Medicaid expansion under the 2010 Affordable Care Act. But because all Democratic governors supported the expansion, separation occurs–Democratic governors perfectly predict support for Medicaid expansion.

I focus on their first hypothesis:

> Republican governors are more likely to oppose the Medicaid expansion funds

than Democratic governors.

In part to address separation, Barrilleaux and Rainey adopt a fully Bayesian approach. Here, I re-estiamte Barrilleaux and Rainey's (2014) logistic regression model using several frequentist procedures. Table 1 presents these results.

Table 1

Estimator	Coef. Est.	SE Est.	Wald *p*-Value	LR *p*-Value
ML with Default Precision	-20.35	3,224	0.99	0.00
ML with Maximum Precision	-35.22	15 million	1.00	0.00
PML with Jeffreys Penalty	-2.68	1.42	0.06	
PML with Cauchy Penalty	-3.38	1.63	0.04	

This table shows the p-values from several procedures that researcers might use when dealing with separation in logistic regression models. The Wald test relies on unreasonable standard errors that depend heavily on the precision of the algorithm and, as a consequence, produces unrealistic p-values. However, the p-values from the likelihood ratio test seem reasonable and resemble the p-vales from the more conservative penalized maximum likelihood approaches.

References

Barrilleaux, Charles, and Carlisle Rainey. 2014. "The Politics of Need: Examining Governors' Decisions to Oppose the 'Obamacare' Medicaid Expansion." *State Politics and Policy Quarterly* 14(4): 437–60.

Bell, Mark S., and Nicholas L. Miller. 2015. "Questioning the Effect of Nuclear Weapons on Conflict." *Journal of Conflict Resolution* 59(1): 74–92.

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

Gelman, Andrew, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. 2008. "A Weakly Informative Prior Distribution for Logistic and Other Regression Models." *The Annals of Applied Statistics* 2(4): 1360–83.

Heinze, Georg, and Michael Schemper. 2002. "A Solution to the Problem of Separation in Logistic Regression." *Statistics in Medicine* 21(16): 2409–19.

Jeffreys, H. 1946. "An Invariant Form of the Prior Probability in Estimation Problems." *Proceedings of the Royal Society of London, Series A* 186(1007): 453–61.

Mares, Isabela. 2015. *From Open Secrets to Secret Voting: Democratic Electoral Reforms and Voter Autonomy*. Cambridge: Cambridge University Press.

Rainey, Carlisle. 2016. "Dealing with Separation in Logistic Regression Models." *Political Analysis* 24(3): 339–55.

Vining, Richard L., Jr., Teena Wilhelm, and Jack D. Collens. 2015. "A Market-Based Model of State Supreme Court News: Lessons from Captial Cases." *State Politics and Policy Quarterly* 15(1): 3–23.

Zorn, Christopher. 2005. "A Solution to Separation in Binary Response Models." *Political Analysis* 13(2): 157–70.