Hypothesis Tests With Separation

Carlisle Rainey*

Draft under development. It's no doubt filled with typos and errors. This is the version from January 20, 2022.

Word Count: 3,475

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). For example, Barrilleaux and Rainey (2014) find that being a Democrat perfectly predicts a governor accepting Medicaid funds under the Affordable Care Act. Under separation, the usual maximum likelihood estimates are unreasonably large and the Wald p-values are highly misleading.

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 `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 than tossing 33 consecutive heads (around 1.2 in ten billion[1]), dealing *two* consecutive five-card straight

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

[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

flushes (around 2.4 in ten billion[2]), an average golfer making two consecutive hole-in-ones (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). Of course, a governor like Deval Patrick was unlikely to oppose the Medicaid, but the chance was *much* higher than one in ten billion.

As a solution, some methodologists propose using a Bayesian prior distribution to regularize the estimates, which we can alternatively consider as *penalized* maximum likelihood estimator. Zorn (2005; see also Heinze and Schemper 2002) points political scientists toward the penalized maximum likelihood estimator proposed by Firth (1993), which is equivalent to Jeffreys' prior distribution (Jeffreys 1946). As an alternative, Gelman et al. (2008) recommend a Cauchy prior distribution. Both of these methods ensure finite estimates in theory and usually produce reasonably-sized estimates in practice.

But Rainey (2016) points out that the parameter estimates (and especially the confidence intervals) depend largely on the chosen prior distribution. Indeed, many priors guarantee finite estimates, but can lead to meaningfully different conclusions. He argues that the set of *a priori* "reasonable" and "implausible" parameters depends on the substantive application, so context-free defaults (like Jeffreys' and Cauchy priors) might not produce reasonable results. Rainey (2016) concludes that "[w]hen facing separation, researchers must *carefully* choose a prior distribution to nearly rule out implausibly large effects" (p. 354).

---

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.

But researchers cannot easily model prior information in all contexts, and some scholars prefer to avoid injecting prior information into the model. How can researchers proceed in these situations? In particular, even if they dismiss the point estimates as implausible, can they obtain useful p-values to test hypotheses in the usual frequentist framework? Below, I show that while the popular Wald test produces wildly misleading p-values under separation, likelihood ratio tests and score tests behave in the usual useful manner. As such, researchers can produce meaningful p-values to test hypotheses with standard frequentist tools under separation.

Statistial Theory

Point Estimates

Maximum likelihood provides a general and powerful framework for obtaining estimates of model parameters. 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 represents a matrix of explanatory variables and β represents a vector of coefficients.

To obtain the "likelihood function" $L(\beta|y)$, compute the product of the probabilities of each observation y_i for a given set of parameters β. 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, we obtain

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

To obtain the maximum likelihood estimates $\hat{\beta}^{ML}$, 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, $\ell(\beta|y) = \log L(\beta|y) = y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i)$.

For logistic regression (and many other models), researchers use numerical algorithms to locate the value of β that maximizes ℓ. For examples, Stata's `logit` command used a modified Newton-Raphson algorithm (link) and R's `glm()` function uses iteratively reweighted least squares.

Hypothesis Tests

With the point estimates in hand, researchers typically estimate the precision of those estimates. In some cases, researchers compare their research hypothesis H_R to a null hypothesis H_0. To conduct a hypothesis test in the context of logistic regression, the researcher 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, I focus on the simple point null hypothesis $H_0 : \beta_s = 0$, i.e., that the coefficient β_s for a variable s (that might sometimes produce separation) equals zero. However, the intuition and conclusions generalize to more complex hypotheses.

To assess the plausibility of the null hypothesis, we compare the null hypothesis $\beta_s = 0$ to the maximum likelihood estimate $\hat{\beta}_s^{ML}$, accounting for the precision of the estimate. The literature offers three methods—the "holy trinity"—to test hypotheses: the Wald test, the likelihood ratio test, and the score test. For practical reasons, most tables of logistic regression coefficient estimates in political science include p-values based on the Wald test. However, the Wald test is uniquely ill-suited for testing hypotheses under separation. While the Wald test perform poorly, the likelihood ratio and score tests work

as expected. Below I briefly describe each test, explain why the Wald test works poorly under separation, and describe why the likelihood ratio and score tests perform better.

Wald Test

Of the three methods, researchers usually report the Wald test most often because it requires fitting only the full model. The Wald procedure uses the shape of the log-likelihood function around the maximum to estimate the precision of the point estimate. If small changes in β lead to large changes in the likelihood function, then we can treat the maximum likelihood estimate 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 Wald test uses the second derivative to quantify the curvature of ℓ at $\hat{\beta}^{ML}$. The second derivative intuitively quantifies the notion of "curvature," and it turns out that we can estimate the sampling variance $\widehat{\text{Var}}(\beta)$ as

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

so that an estimate of the standard error $\widehat{\text{SE}}(\hat{\beta}_i^{ML})$ is

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

For large (repeated) samples, the maximum likelihood estimate follows 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 H_0. For the simple null hypothesis H_0 that $\beta_s = 0$ (constraint in only one dimension), we have

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

Following the common convention, if the p-value is less than 0.05, the researcher rejects the null hypothesis (that $\beta_i = 0$, typically) in favor of the research hypothesis (that $\beta_i \neq 0$). If the p-value is greater than 0.05, then the researcher cannot distinguish between the two hypotheses.

However, *this approach works poorly when dealing with separation*. Under separation, the log-likelihood function at the found maximum is numerically flat or nearly so. The flatness produces very large standard error estimates (that have a troubling dependence on the error tolerance of the algorithm).

For a logistic regression model with a separating variable, we can state a precise result that relates the size of the coefficient of the separating variable to the Wald estimate of its standard error.

Theorem 1 *Suppose that a binary explanatory variable s with coefficient β_s perfectly predicts the outcome y_i such that when $s_i = 1$ then $y_i = 1$. Then the log-likelihood function increases in β_s and the quantity* $\lim_{\beta_s \to \infty} \left[\left(-\frac{\partial^2 \ell(\beta_s | y)}{\partial^2 \beta_s} \right)^{-\frac{1}{2}} - \beta_s \right] = \infty$. *Thus, under separation, the estimated standard error will be much larger than the coefficient for the separating variable for any algorithm that obtains a sufficiently large coefficient.*

Theorem 1 shows that, so long as the researcher uses a sufficiently precise algorithm, the Wald test will *never* reject the null hypothesis.

Suppose an absurd example in which a binary treatment perfectly predicts 500 successes and 500 failures (i.e., $y = x$ always). The default `glm()` routine in R calculates a Wald p-value of 0.998 with the default precision and 1.000 with the maximum precision. But of course, these data are *extremely* unlikely under the null hypothesis that the coefficient for the treatment indicator equals zero. The exact p-value for the (equivalent) null hypothesis that successes and failures are equally likely under both treatment and control equals $2 \times \left(\frac{1}{2}\right)^{500} \times \left(\frac{1}{2}\right)^{500} = \frac{2}{2^{1000}} \approx \frac{2}{10^{301}}$. (For comparison, there are about 10^{80} atoms in the known universe.)

When dealing with separation, the approach of using curvature around the maximum

to estimate the relative likelihood of the restricted model breaks down; researchers cannot use the Wald test to obtain reasonable p-values for the coefficient of a separating variable.

Likelihood Ratio Test

The likelihood ratio test resolves the problem of the flat log-likelihood by comparing the maximum log-likelihood of two models: an "unrestricted" model ML that imposes no bounds on the estimates and a "restricted" model ML_0 that constrains the estimates to the region suggested by the null hypothesis. The likelihood ratio test compares the value of the unrestricted log-likelihood $\ell(\hat{\beta}^{ML}|y)$ to the restricted log-likelihood $\ell(\hat{\beta}^{ML_0}|y)$. For the simple null hypothesis $H_0 : \beta_s = 0$, we can understand ML_0 as a separate model fit without the explanatory variable of interest s.

If the data are much more likely under the unrestricted estimates $\hat{\beta}^{ML}$ than under the restricted estimates $\hat{\beta}^{ML_0}$, then the researcher can reject the null hypothesis. Wilks' theorem (1938) advises us how to compare the two likelihoods: $D = 2 \times \left[\ell(\hat{\beta}^{ML}|y) - \ell(\hat{\beta}^{ML_0}|y)\right]$ follows a χ^2 distribution with degrees of freedom equal to the number of constrained dimensions (Wilks 1938, Casella and Berger 2003, pp. 488-492, Greene 2003, pp. 484-486).

Score Test

The score test resolves the problem of the flat log-likelihood by focusing on the slope or gradient of the log-likelihood function at the null hypothesis. If the null hypothesis is correct, then the log-likelihood should not be increasing much around that point. But if the log-likelihood function is increasing rapidly at the null hypothesis, this indicates that the data are inconsistent with the null hypothesis. The score test uses two quantities: the score function $S(\beta) = \dfrac{\partial \ell(\beta|y)}{\partial \beta}$ and the information number (i.e., Fisher information) $I(\beta) = -E_\beta \left(\dfrac{\partial^2 \ell(\beta|y)}{\partial^2 \beta} \right)$. When evaluated at the null hypothesis, the score function quantifies the slope and that information number is the variance of that slope in repeated

samples.

If the score at the null hypothesis is large relative to its standard deviation, then researchers can reject the null hypothesis. Rao (1948) advises us how evaluate the assess the statistic: $Z_s = \frac{S(\beta_s^0)}{\sqrt{I(\beta_s^0)}}$ follows a standard normal distribution. (Rao 1948, Casella and Berger 2003, pp. 494-495, Greene 2003, pp. 489-490 (lagrange multiplier framework)).

Penalized Maximum Likelihood

Some scholars suggest using *penalized* maximum likelihood (PML) to ensure a non-monotonic likelihood function when dealing with separation. Zorn (2005) points political scientists to Firth's PML, which we can understand as ???? inveriant prior distribution. Gelman et al (YYYY) suggest an alternative penality that we can understand as a Cauchy prior distribution. These approach can be quite helpful in obtaining reasonable point estimates and hypothesis tests. While both method have a reasonable default penalty, Rainey (2016) shows that the inferences can be quite sensitive to the chosen penalty (i.e., the values that the research is willing to rule out as implausible).

For completeness, I include Wald tests of the point estimate from these two ML approaches. Because both guarantee non-monotonic likelihood, the critique of the Wald test under separation does not apply. The tests should work well. However, Rainey's (2016) critique does apply—the reasonableness of the inferences will depend on the reasonableness of the penalty. When researchers are unable or unwilling assess the plausibility and implausibility of certain parameter values *a priori*, PML is not a viable approach.

Tests Under Separation

Above, I describe why the Wald test fails under separation: the monotonic log-likelihood functions flattens quickly, which essentially guarantees a standard error much larger than the coefficient estimate. The likelihood ratio and score tests avoid this problem by using

Test	Feature	Statistic and Distribution
Wald	Curvature of the log-likelihood function around the maximum.	$Z_w = \frac{\hat{\beta}_i^{ML}}{\widehat{\mathsf{SE}}(\hat{\beta}_i^{ML})}$ follows a standard normal distribution.
Likelihood Ratio	Relative log-likelihoods of the unrestricted and restricted models.	$D = 2 \times \left[\ell(\hat{\beta}^{ML}\|y) - \ell(\hat{\beta}^{ML_0}\|y)\right]$ follows a χ^2 distribution with degrees of freedom equal to the number of constrained dimensions.
Score	Slope of the log-likelihood function *at the null hypothesis*.	$Z_s = \frac{S(\beta_s^0)}{\sqrt{I(\beta_s^0)}}$ follows a standard normal distribution.

alternative features of the log-likelihood function to test the hypothesis.

The likelihood ratio test does not rely on the curvature of the likelihood function. Instead, the likelihood ratio test compares the relative likelihood of the unrestricted and restricted model. If excluding an explanatory variable from the model decreases the likelihood substantially, the we reject the null hypothesis that its coefficient equals zero. The score test uses the slope (or gradient) of the log-likelihood function *at the null hypothesis*. Both approaches approach avoid the issue of the flat log-likelihood at the maximum. While all three tests assumes large samples, the likelihood ratio and score tests offer a better framework in situations where separations might occur. The simulations below demonstrate this is the case.

Simulations

To evaluate the performance of the various methods for testing hypothesis under separation, I use a diverse collection of data-generating processes (DGPs) that sometimes feature separation. Importantly, I cannot focus on data sets *with separation* because separation is a feature of a particular sample. Instead, I focus on DGPs that *sometimes* feature separation (e.g., in 10% of repeated samples, in 50% of repeated samples, etc.).

To create the collection of DGPs, I imagine the logistic regression model $\Pr(y = 1) = \text{logit}^{-1}(\beta_{\text{cons}} + \beta_s s + \beta_{z_1} z_1 + ... + \beta_{z_k} z_k)$ and a researcher testing the null hypothesis that the binary explanatory variable s (that might produce separation) has no effect on a binary outcome variable y. I vary the frequency that $s = 1$, the value of β_{cons}, the number

of control variables (k), and the total number of observations. For each DGP, I use Monte Carlo simulation to compute the power function for each of five methods: Wald test, likelihood ratio test, and score test for ML estiamtes, and the Wald test for the two PML alternatives.

I use all 944 (logically possible) combinations of the values below. When the null hypothesis true (i.e., $\beta_s = 0$), the percent of repeated samples with separation ranges from 98% (ctk) to 14% (ctk) in this collection.

A Careful Look at a Single DGP

To understand the simulation results, I

Examples

To illustrate frequentist hypothesis testing under separation, I reanalyze data from Barrilleaux and Rainey (2014) that Rainey (2016) considers in great detail. 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–being a Democratic governor perfectly predicts support for Medicaid expansion.

I focus on their first hypothesis: Republican governors are more likely to oppose the Medicaid expansion funds than Democratic governors. Barrilleaux and Rainey adopt a fully Bayesian approach, modeling the probability that a state's governor opposes the Medicaid expansion as a function of the governor's partisanship and several other covariates. Here, I re-estimate their logistic regression model using several frequentist procedures. The appendix provides the full details. Table 1 presents these estimates and p-values for the indicator for Democratic governors.

Because no Democratic governors oppose the expansion, being a Democrat perfectly

Table 1

Estimator	Coef. Est.	SE Est.	Wald *p*-Value	LR *p*-Value	Score *p*-Value
ML with Default Precision	-20.35	3,224	0.99	0.00	0.01
ML with Maximum Precision	-35.22	15 million	1.00	0.00	0.01
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.

predicts non-opposition. Therefore, the coefficient estimates are implausibly large. Barrilleaux and Rainey use a Bayesian approach. Other authors suggest penalized maximum likelihood, but Rainey (2016) shows that the choice of penalty can impact the results. Here I compute Wald p-values for four approaches: maximum likelihood with default precision, maximum likelihood with maximum precision, penalized maximum likelihood with a Jeffries penalty (Firth 1993), and penalized maximum likelihood with a Cauchy penalty (Gelman ????). The status quo practice treats the former two as unreliable and the latter two options as suitable. But Rainey (2016) argues that the appropriate penalty is subjective and application-specific. However, but the likelihood ratio and score p-values return reasonable p-values without selecting a suitable penalty (Rainey 2016) or using prior information (Barrilleaux and Rainey 2014).

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.