Hypothesis Tests Under Separation

Carlisle Rainey*

Draft under development. It's no doubt filled with typos and errors. This is the version from April 11, 2022.

Word Count: 3,524

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 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

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

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 or penalty. Indeed, many priors that guarantee finite estimates 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). But it's not always easy to include prior information, and some scholars prefer to avoid injecting prior information into their model. How can researchers proceed in these situations? In particular, can they obtain useful p-values to test hypotheses in the usual frequentist framework without using prior information?

Below, I show that while the popular Wald test produces misleading (even nonsensical) p-values under separation, likelihood ratio tests and score tests behave in the usual manner. As such, researchers can produce meaningful p-values to test hypotheses with standard frequentist tools under separation *without the use of prior information*.

Hypothesis Tests Under Separation

Maximum likelihood provides a general and powerful framework for obtaining estimates of model parameters and testing hypotheses. 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. Then we can derive the likelihood function $L(\beta|y)$

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

and the log-likelihood function

$$\ell(\beta|y) = \log L(\beta|y) = y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i). \tag{3}$$

Researchers typically use numerical algorithms to locate the ML estimate $\hat{\beta}^{ML}$, which maximizes ℓ. Then researchers rely on certain features of ℓ to test hypotheses. To fix ideas, I focus on the simple point null hypothesis $H_0 : \beta_s = 0$. However, the intuition and conclusions generalize to more complex hypotheses.

The literature offers three common methods to assess the null hypothesis—the "holy trinity" of hypothesis tests: the Wald test, the likelihood ratio test, and the score test. For practical reasons, most regression tables in political science include p-values based on the Wald test. While usually the most convenient of the three tests, the Wald test is uniquely ill-suited for testing hypotheses under separation. The likelihood ratio and score tests, on the other hand, 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 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 the parameter near the maximum lead to large changes in the log-likelihood function, then we can treat the maximum likelihood estimate as precise. The Wald test uses the second derivative to quantify the curvature of ℓ at $\hat{\beta}^{ML}$. We can estimate the standard error $\widehat{\text{SE}}(\hat{\beta}_i^{ML})$ as

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

For large samples, the maximum likelihood estimate approximately follows a normal distribution centered at the true value of β with a standard deviation of $\widehat{\text{SE}}(\hat{\beta}_i^{ML})$. Wald (1943) advises us how compare the estimate with the standard error: the statistic $Z_w = \frac{\hat{\beta}_i^{ML}}{\widehat{\text{SE}}(\hat{\beta}_i^{ML})}$ approximately follows a standard normal distribution. (Casella and Berger 2003, pp. 492-493, Greene 2012, pp. 527-529).

Figure 1 shows this intuition for a typical, non-monotonic log-likelihood function (i.e., without separation) and for a monotonic log-likelihood function (i.e., with separation). In the absence of separation, the curvature of the log-likelihood function around the maximum speaks to the evidence against the null hypothesis. But under separation, the likelihood function, *by definition*, is flat at the maximum, regardless of the relative likelihood of the data under the null hypothesis.

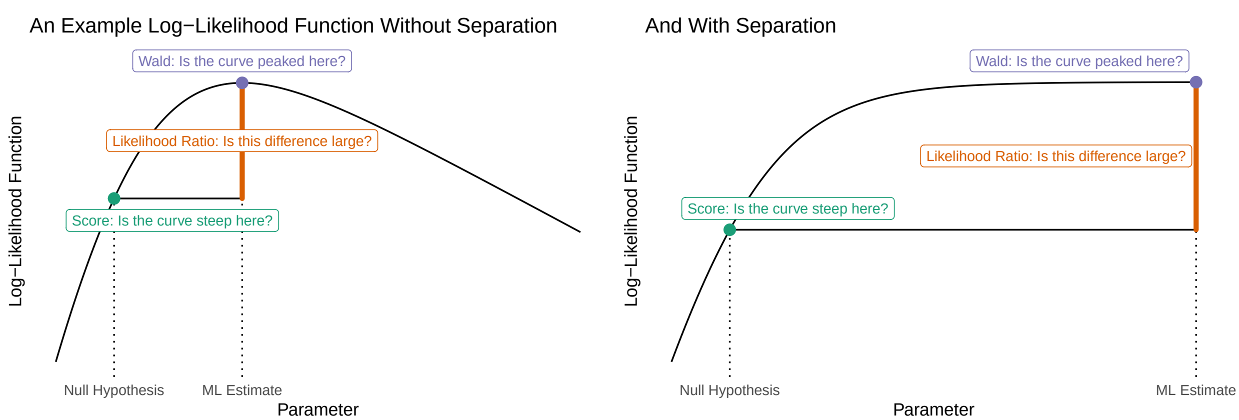

Figure 1: A figure summarizing logic of the "holy trinity" of hypothesis tests. The Wald test relies on the curvature around the maximum of the log-likelihood function, which breaks down under separation. But the likelihood ratio and score test rely on *other* features of the log-likelihood function, that are not meaningfully impacted by separation.

This approach works poorly when dealing with separation. Under separation, the log-likelihood function at the numerical maximum is nearly. The flatness produces very large standard error estimates.

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. 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. The standard error estimate associated with each β_s increases as well. But critically, the estimated standard error increases *faster* than the the associated coefficient, because $\lim_{\beta_s \to \infty} \left[\left(-\frac{\partial^2 \ell(\beta_s|y)}{\partial^2 \beta_s} \right)^{-\frac{1}{2}} - \beta_s \right] = \infty$. 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. So long as the researcher uses a sufficiently precise algorithm, the Wald test will *never* reject the null hypothesis, regardless of the dataset.

If the Wald test can never reject the null hypothesis for any dataset with separation, then the power of the dataset is strictly bounded by the chance of separation. In particular, the power of the test cannot exceed $1 - Pr(\text{separation})$.

This result highlights the poor behavior of the Wald test under separation. Increasing the coefficient of a potentially separating variable typically increases the chance of separation, which might *decrease* the power of the test. This leads to a pathological result: moving the coefficient further from zero might *decrease* the power of the test. Also, if the dataset features separation in nearly 100% of repeated samples, then the test will have power near 0%.

To see the importance of this result, suppose an absurd example in which a binary treatment perfectly predicts 500 successes and 500 failures (i.e., $y = x$ always). Of course, this dataset is *extremely* unlikely under the null hypothesis that the coefficient for the treatment indicator equals zero. The exact p-value for the 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.) Yet, 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).

When dealing with separation, the Wald test 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. If the data are much more likely under the unrestricted estimate than under the restricted estimate, then the researcher can reject the null hypothesis.

Figure 1 shows the intuition of the likelihood ratio test. The gap between the unrestricted and restricted maximum summarizes the evidence against the null hypothesis. Importantly, the logic does not break down under separation. Unlike the Wald test, the likelihood ratio test can reject the null hypothesis under separation.

Wilks' theorem (1938) advises us how to compare the unrestricted log-likelihood $\ell(\hat{\beta}^{ML}|y)$ to the restricted log-likelihood $\ell(\hat{\beta}^{ML_0}|y)$: $D = 2 \times \left[\ell(\hat{\beta}^{ML}|y) - \ell(\hat{\beta}^{ML_0}|y)\right]$ approximately follows a χ^2 distribution with degrees of freedom equal to the number of constrained dimensions (Casella and Berger 2003, pp. 488-492, Greene 2012, pp. 526-527).

Score Test

The score test resolves the problem of the flat log-likelihood by evaluating the slope of the log-likelihood function at the null hypothesis. If the log-likelihood function is increasing rapidly at the null hypothesis, this casts doubt on the null hypothesis. The score test uses the score function $S(\beta) = \dfrac{\partial \ell(\beta|y)}{\partial \beta}$ and the 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 the

Fisher information quantifies the variance of that slope in repeated samples. If the score at the null hypothesis is large, then the researcher can reject the null hypothesis.

Figure 1 shows the intuition of the score test. The slope of the log-likelihood function under the null hypothesis summarizes the evidence against the null hypothesis. As with the likelihood ratio test, the logic works even under separation, and the score test can reject the null hypothesis under separation.

Rao (1948) advises us how assess the score: $Z_s = \frac{S(\beta_s^0)}{\sqrt{I(\beta_s^0)}}$ follows a standard normal distribution (Casella and Berger 2003, pp. 494-495, Greene 2012, pp. 529-530).

Table 1 summarizes the three tests. Most importantly, the likelihood ratio and score tests rely on features of the log-likelihood function that are not meaningfully affected by a monotonic log-likelihood function. The Wald test, on the other hand, cannot provide a reasonable tests under separation.

Table 1: A table summarizing the "holy trinity" of hypothesis tests.

Test	Feature	Statistic and Distribution
Wald	Curvature of the log-likelihood function around the maximum.	$Z_w = \frac{\hat{\beta}_i^{ML}}{\widehat{\text{SE}}(\hat{\beta}_i^{ML})}$ 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.

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 feature separation in more than 10% of repeated samples. 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 15% 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 as β_s varies for each of the three tests: Wald test, likelihood ratio test, and score test.

A Close Look at a Single DGP

Following the structure discussed above, Table **??** shows the parameters for this single DGP with 50 observations of which five have $s = 1$, the intercept $\beta_{cons} = 0$, and the number of control variables $k = 2$. Table 2 shows the power function for each of the three tests, as well as the ideal power, and the percent of the data sets that featured separation for the parameter combination and the value of β_s. For this particular DGP, separation is relatively rare when β_s—the coefficient for the potentially separating variable—is near zero. But for $\beta_s = \pm 2$, about 50% of the data sets feature separation. And for β_s larger/smaller than ± 4, more than 90% of the data sets feature separation.

This DGP clearly demonstrates the poor performance of the Wald test. Even though the datasets with separation should allow the researcher to reject the null hypothesis, at occasionally, the power of the Wald test is near 0%, even for very large effects. The likelihood ratio and score tests, on the other hand, perform as expected. For both alternatives, the power of the test when $\beta_s = 0$ is about 5%, as designed, and the power approaches 100% relatively quickly as β_s moves away from zero.

A Broad Look at Many DGPs

Table 3 shows the broad collection of parameters I used in the simulations. All combinations of the values below yield 150 unique combinations. I exclude combinations

Table 2

β_s	Ideal Power	Percent with Separation	Wald Test Power	Likehood Ratio Test Power	Score Test Power
-10		100%	0%	99%	85%
-7		99%	0%	99%	85%
-4	As high as possible.	89%	0%	90%	77%
-2		45%	2%	49%	42%
-1		13%	2%	18%	15%
0	5%	4%	1%	6%	5%
1		19%	0%	20%	16%
2		50%	0%	51%	41%
4	As high as possible.	90%	0%	89%	70%
7		99%	0%	97%	76%
10		100%	0%	98%	78%

This table shows the power for the Wald, likelihood ratio, and score tests for a DGP that often features separation.

in which the frequency that $s = 1$ exceeded the number of observations, which leaves 110 combinations. For each of these 110 combinations, I use Monte Carlo simulations to estimate the probability that each method rejects the null hypothesis as the coefficient β_s varies from -10 to 10 (across the particular values shown in Table 2). I also exclude any scenario in which the parameter and the coefficient β_s would sometimes produce data sets with no variation in the outcome variable. As such, one might consider this a diverse collection of DGPs that sometimes generate separation but almost always have variation in the outcome.

Table 3

Parameter	Value
Total Number of Observations	50, 250, 500
Frequency that $s = 1$	5, 25, 50, 100, 250
The Value of the Intercept β_{cons}	-8, -5, -2.5, -1, 0
The Number of Control Variables k	2, 6

Figure 2 shows the statistical power of each of the three tests as the chance of separation varies across the many DGPs. Most starkly, the power of the Wald test is bounded above by $1 - Pr(\text{separation})$, and many scenarios achieve the boundary. Intuitively, as the chance of separation increases, the power of the test should increases as well, because separation is evidence of a large coefficient. The likelihood ratio and score test, on the other hand, behave as expected.

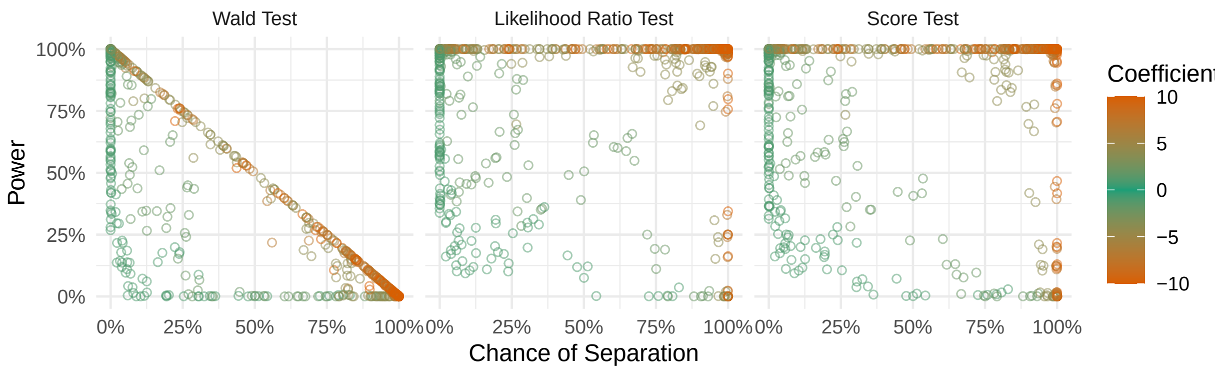

Figure 2: The power of tests across a range of scenarios as the chance of separation varies.

Because a large coefficient makes separation more likely, a large coefficient *decreases* the power of the test.

Figure 3 plots the power function for each of the 110 DGPs in the diverse collection.[1] The fine lines show the power function for each DGP and the color of the points indicates the chance of separation at that point. The heavy, solid lines show the average of the collection of power functions. The heavy, dashed lines show the power functions for the other two tests, for comparison. The power functions for the Wald tests show its poor properties. For most of the power functions, as the true coefficient grows larger in magnitude from about three, the test becomes less powerful. This occurs because separation becomes more likely and the test cannot reject the null hypothesis when separation occurs. Second, the likelihood ratio and score test behave reasonably well. Most importantly, the size of the likelihood ratio and score tests is about 5% when the coefficient equals zero and grows as the coefficient moves away from zero. The likelihood ratio test is slightly preferred, at least of this collection of DGPs, because it is slightly more powerful when the coefficient does not equal zero, especially for large, negative

[1] I exclude parameter combinations that might not produce variation in the outcome variable, so I do not compute the power function across the entire range of the coefficient for every power function. For 47 of the 110 DGPs, I exclude a portion of the range. For 63 of the 110, I compute the power function across the entire range from -10 to 10.

coefficients.[2]

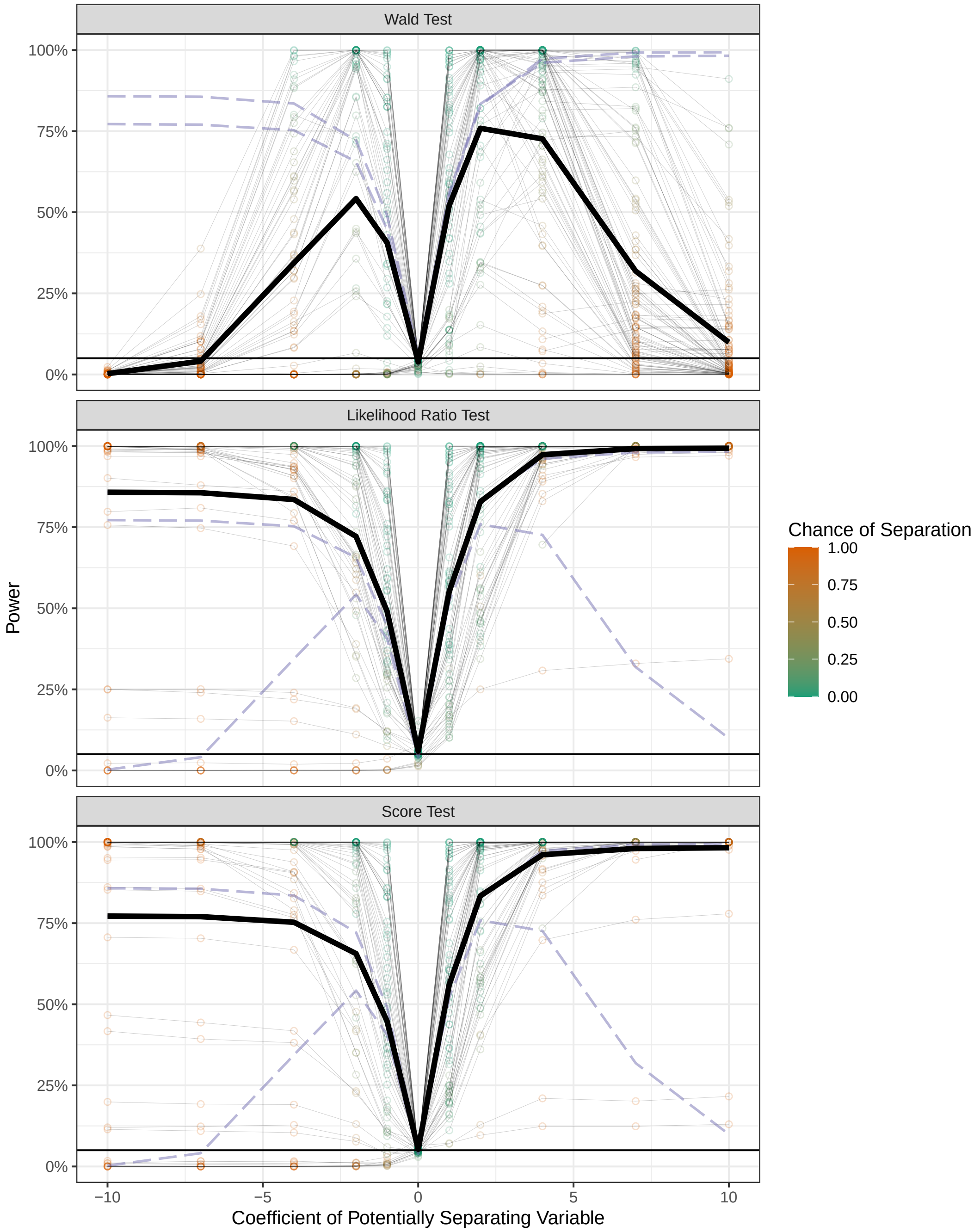

Figure 3: The power functions for the Wald, likelihood ratio, and score tests for a diverse collection of data-generating processes that sometimes generate separation.

[2]This asymmetry occurs because I only consider negative intercepts.

Re-Analysis of Barrilleaux and Rainey (2014)

To illustrate frequentist hypothesis testing under separation, I reanalyze data from Barrilleaux and Rainey (2014), who 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 4 presents the estimates and p-values for the coefficient of the binary Democratic governor variable.

Table 4

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

The p-values from several procedures that researchers might use when dealing with separation in logistic regression models. The Wald test for maximum likelihood estimates relies on unreasonable standard errors that depend heavily on the precision of the algorithm and, as a consequence, produces unrealistic p-values. However, the likelihood ratio and score tests produce reasonable p-values.

Because no Democratic governors oppose the expansion, being a Democrat perfectly predicts non-opposition. Therefore, the coefficient estimates and standard errors are implausibly large. To address this, Barrilleaux and Rainey use a Bayesian approach. This example shows that the likelihood ratio and score tests offer reasonable p-values without using prior information.

Conclusion

Separation commonly occurs in political science. When this happens, I shows that the usual Wald p-values are highly misleading. But researchers cannot always use suitable prior information to address a monotonic likelihood function. Even without a suitable prior or penalty, I show that the standard likelihood ratio and score tests behave in the usual way. As such, researchers can use the likelihood ratio and score tests to produce meaningful p-values under separation *without the use of prior information*.

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.