Hypothesis Tests Under Separation

Carlisle Rainey*

Version from August 31, 2023.

Published at *Political Analysis*.
DOI: 10.1017/pan.2023.28

Separation commonly occurs in political science, usually when a binary explanatory variable perfectly predicts a binary outcome. In these situations, methodologists often recommend penalized maximum likelihood or Bayesian estimation. But researchers might struggle to identify an appropriate penalty or prior distribution. Fortunately, I show that researchers can easily test hypotheses about the model coefficients with standard frequentist tools. While the popular Wald test produces misleading (even nonsensical) p-values under separation, I show that likelihood ratio tests and score tests behave in the usual manner. Therefore, researchers can produce meaningful p-values with standard frequentist tools under separation *without the use of penalties or prior information*.

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

Introduction

Separation commonly occurs in political science, usually when a binary explanatory variable perfectly predicts a binary outcome (e.g., Gustafson 2020; Mehltretter 2021; Owsiak and Vasquez 2021).[1]. For example, Barrilleaux and Rainey (2014) find that being a Democrat perfectly predicts a governor supporting Medicaid expansion under the Affordable Care Act. Under separation, the usual maximum likelihood estimate is unreasonably large and the Wald test is highly misleading.

As a solution, some methodologists propose using a Bayesian prior distribution to regularize the estimates, which we can alternatively consider as a *penalized* maximum likelihood estimator. Zorn (2005; see also Heinze and Schemper 2002) points political scientists toward Firth's (1993) penalized maximum likelihood estimator, which is equivalent to Jeffreys prior distribution. Gelman et al. (2008), on the other hand, recommend a Cauchy prior distribution. Both methods ensure finite estimates in theory and usually produce reasonably-sized estimates in practice. Methodologists continue to recommend these penalized or Bayesian estimators as a solution to separation (Anderson, Bagozzi, and Koren 2021; Cook, Hays, and Franzese 2020; e.g., Cook, Niehaus, and Zuhlke 2018; Crisman-Cox, Gasparyan, and Signorino 2023).

But Rainey (2016) points out that the estimates (and especially the confidence intervals) depend largely on the chosen prior. Many priors that produce finite estimates also produce meaningfully different conclusions. He argues that the set of *a priori* "reasonable" and "implausible" parameters depends on the substantive application, so context-free

[1]Throughout this paper, I use "separation" to refer to quasicomplete separation (Albert and Anderson 1984) when a single binary explanatory variable s perfectly predicts either the outcome $y = 1$ *or* $y = 0$. Zorn (2005) notes that quasicomplete separation is "far more common" (p. 161) than complete separation in political science and focuses on the situation where "only one cell of the implied 2×2 table of s and y is 'empty' " (p. 161; *notation adjusted to match*). This is the most common form of separation identified in political science. Also, focusing here simplifies the presentation. However, separation can occur outside this scenario I consider. In particular, separation can occur due to a *linear combination* of explanatory variables (rather than a single explanatory variable) and in models other than logistic regression (e.g., ordered logistic regression, Cox proportional hazards). The arguments I present extend beyond the "single empty cell" scenario to other forms of separation (e.g., complete, linear combinations) and to other models (e.g., ordinal logistic regression, Cox proportional hazards).

defaults (like Jeffreys and Cauchy priors) might not produce reasonable results. Starkly emphasizing this point, Beiser-McGrath (2022) shows that Jeffreys prior can lead to (statistically significant) estimates in the *opposite direction* of the separation. Rainey (2016) concludes that "[w]hen facing separation, researchers must *carefully* choose a prior distribution to nearly rule out implausibly large effects" (p. 354). Given the sensitivity of the result to the chosen prior distribution, how can researchers make their analysis more compelling? In particular, can they obtain useful p-values to test hypotheses about model coefficients in the usual frequentist framework without injecting prior information into their model?

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. Thus, researchers can produce meaningful p-values with standard frequentist tools under separation *without the use of a penalty or prior*. A complete analysis of a data set with separation will usually include penalized or Bayesian estimates to obtain reasonable estimates of quantities of interest, but a hypothesis test without a penalty or prior can more convincingly establish that the most basic claim holds: the separating variable has a positive (or negative) effect.

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 of n (or that the outcome variable $y_i = 1$) as $\pi_i = \text{logit}^{-1}(X_i\beta)$ for $i = 1, 2, ..., n$, where X represents a matrix of explanatory variables and β represents a vector of coefficients. Then we have the likelihood function $L(\beta \mid y) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{(1-y_i)}$ and the log-likelihood function $\ell(\beta \mid y) = \log L(\beta \mid y) = \sum_{i=1}^{n}[y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i)]$. Researchers typically use

numerical algorithms to locate the maximum likelihood estimate $\hat{\beta}^{ML}$ that maximizes ℓ and then use certain features of ℓ to test hypotheses. To fix ideas, I focus on the 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 (also known as the Lagrange multiplier test). For practical reasons, most regression tables in political science report Wald p-values. However, the Wald test is uniquely ill-suited for testing hypotheses under separation. Because the usual Wald test fails, some researchers turn immediately penalized estimators (e.g., Bell and Miller 2015) or Bayesian inference (e.g., Barrilleaux and Rainey 2014). However, the usual likelihood ratio and score tests work as expected under separation. Thus, researchers can use the likelihood ratio or score test to evaluate the core hypothesis that the separating variable has a positive (or negative) effect before turning to penalized or Bayesian methods to estimate quantities of interest. 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

The Wald test 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. We usually estimate the standard error $\widehat{\text{SE}}(\hat{\beta}_i^{ML})$ as

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

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 2002, pp. 492-493; Greene 2012, pp. 527-529).

The Wald approach works poorly when dealing with separation. Under separation, the log-likelihood function at the numerical maximum is nearly flat. The flatness produces very large standard error estimates—much larger than the coefficient estimates. Figure 1 shows this intuition for a typical, non-monotonic log-likelihood function (i.e., without separation) and 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 monotonic likelihood function is flat at the maximum, regardless of the relative likelihood of the data under the null hypothesis.

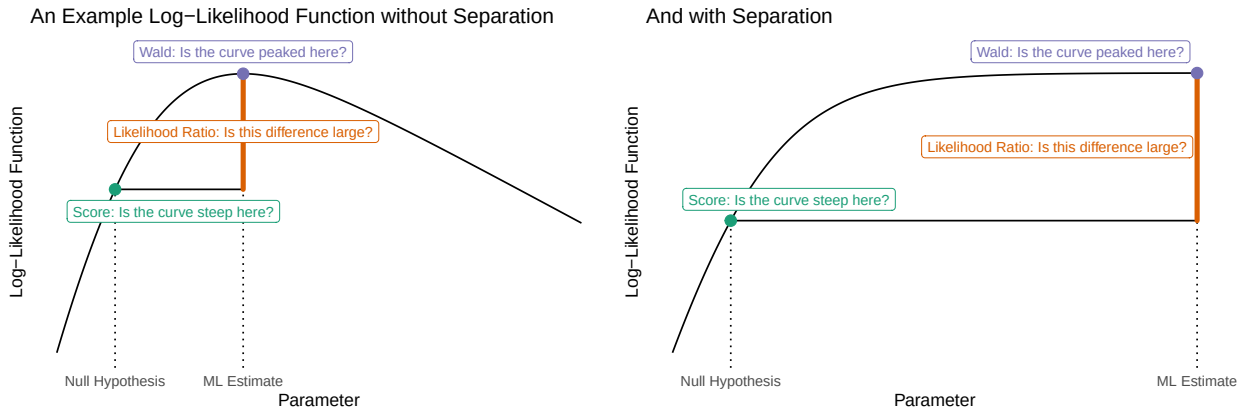

Figure 1: A figure summarizing logic of the "holy trinity" of hypothesis tests. The Wald test relies on the curvature around the maximum of the log-likelihood function, which breaks down under separation. The likelihood ratio and score test, on the other hand, rely on *other* features of the log-likelihood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

We can develop this intuition more precisely and formally. 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. Critically, though, the estimated standard error increases *faster* than the associated coefficient, because

$\lim_{\beta_s \to \infty} \left[\left(-\frac{\partial^2 \ell(\beta_s \mid y)}{\partial^2 \beta_s} \right)^{-\frac{1}{2}} - \beta_s \right] = \infty$. Thus, under separation, the estimated standard error will be much larger than the coefficient for the separating variable. This implies two conclusions. First, so long as the researcher uses a sufficiently precise algorithm, *the Wald test will never reject the null hypothesis under separation*, regardless of the data set. Second, if the Wald test can never reject the null hypothesis for any data set with separation, then the power of the test is strictly bounded by the chance of separation. In particular, *the power of the test cannot exceed* $1 - Pr(separation)$. If the data set features separation in nearly 100% of repeated samples, then the Wald test will have power near 0%.

As a final illustration, suppose an absurd example in which a binary treatment perfectly predicts 500 successes and 500 failures (i.e., $y = x$ always). Of course, this data set 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.[2]

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

[2] de Carvalho Barreto et al. (2014) make a similar point. They use a hypothetical data set to argue that "Wald statistics are inappropriate for analysis [in the context of separation], because these are affected in the presence of the phenomenon of separation of variables" (p. 725).

the region suggested by the null hypothesis. If the data set is much more likely under the unrestricted estimate than under the restricted estimate, then the researcher can reject the null hypothesis. Wilks (1938) advises us how to compare the unrestricted log-likelihood $\ell(\hat{\beta}^{ML} \mid y)$ to the restricted log-likelihood $\ell(\hat{\beta}^{ML_0} \mid y)$: $D = 2 \times \left[\ell(\hat{\beta}^{ML} \mid y) - \ell(\hat{\beta}^{ML_0} \mid y)\right]$ approximately follows a χ^2 distribution with degrees of freedom equal to the number of constrained dimensions (Casella and Berger 2002, pp. 488-492, Greene 2012, pp. 526-527).

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.[3]

Score Test

The score test (or Lagrange multiplier test) resolves the problem of the flat log-likelihood by evaluating the gradient 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 $S(\beta) = \dfrac{\partial \ell(\beta \mid y)}{\partial \beta}$ and the Fisher information $I(\beta) = -E_\beta \left(\dfrac{\partial^2 \ell(\beta \mid y)}{\partial^2 \beta} \right)$. 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. Rao (1948) advises us how to compare the score to its standard error: $Z_s = \frac{S(\beta_s^0)}{\sqrt{I(\beta_s^0)}}$ follows a standard normal distribution (Casella and Berger 2002, pp. 494-495, Greene 2012, pp. 529-530).

Figure 1 shows the intuition of the score test. The slope of the log-likelihood function

[3]As a helpful illustration, Nagashima and Sato (2017) make a similar observation in an applied data analysis. Their Wald test returns a p-value of 0.98, while the likelihood ratio test returns a p-value of less than 0.01 (p. 4323). Given their data, they find the latter more plausible. Similarly, Sun et al. (2015) recognize that the likelihood ratio test is an appropriate alternative to the Wald test under separation. They adopt a strategy preemptively: "If a separation (or monotone likelihood) problem occurs, the likelihood ratio test is used to replace Wald's test for computing the p-value" (p. 33).

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.

Table 1 summarizes the three tests. For further discussion of the connections among the tests, see Buse (1982). 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 test 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.	$Z_w = \frac{\hat{\beta}_i^{ML}}{\widehat{\mathsf{SE}}(\hat{\beta}_i^{ML})}$ follows a standard normal distribution.
Likelihood Ratio	Relative log-likelihoods of the unrestricted and restricted models.	$D = 2 \times \left[\ell(\hat{\beta}^{ML} \mid y) - \ell(\hat{\beta}^{ML_0} \mid y)\right]$ 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 three tests under separation, I use a Monte Carlo simulation to compute the power functions for a diverse collection of data-generating processes (DGPs). For each DGP, I compute the probability of rejecting the null hypothesis as the coefficient for the potentially separating explanatory variable varies from -5 to 5. For a properly functioning test, the power function should be about 5% when $\beta_s = 0$ (i.e., the "size" of the test) and grow quickly toward 100% as β_s moves away from zero (i.e., the "power" of the test).

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 or in 50% of repeated samples, etc.). To develop these 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.e., that $\beta_s = 0$).

I generate a diverse collection of 150 DGPs using the following process. First, I choose the total number of observations randomly from $\{50, 100, 1000\}$. Then I choose the frequency that $s = 1$ ($\sum s$) from $\{5, 10, 25, 50, 100\}$ (subject to the constraint that $\sum s$ must be less than the total number of observations). Next, I draw the value of the constant term (β_{cons}) from a continuous uniform distribution from -5 to 0, the number of control variables (k) from a uniform distribution from 0 to 6, the correlation among the the explanatory variables (ρ) from a continuous uniform distribution from 0 to 0.5.[4] I simulate many of these DGPs and keep 150 that feature (1) separation in a least 30% of repeated samples for some $\beta_s \in [-5, 5]$ and (2) variation in the outcome variable in at least 99.9% of repeated samples. For each of the 150 DGPs. I use Monte Carlo simulation to compute the power function for each of the three tests discussed above.[5] For comparison, I also compute the power function for Wald tests using Firth's (1993) penalty and Gelman et al.'s (2008) Cauchy penalty.

A Close Look at a Single DGP

First, I describe the results for a single DGP. For this particular DGP, there are 1,000 total observations, $s = 1$ for only five of the observations and $s = 0$ for the other 995 observations, the constant term β_{cons} equals -4.1, there are three control variables, and the latent correlation ρ among the explanatory variables is 0.06. Table 2 shows the power function for each test and the chance of separation as β_s varies. Separation is relatively rare when β_s—the coefficient for the potentially separating variable—is between -0.5 and

[4] I simulate each control variable z_i from a normal distribution with a standard deviation of 0.5 and set the coefficient for each control variable β_{z_i} to one. To create the dependence between s and the continuous control variables, I draw a latent random variable s^* from a multivariate normal distribution with the control variables and code the $\sum s$ highest values as 1 all others as 0.

[5] I use 2,500 simulations per estimate, for a worst-case Monte Carlo standard error of $\frac{\sqrt{0.5 \times 0.5}}{\sqrt{2{,}500}} = \frac{0.5}{50} = 0.01$ or 1 percentage point.

and 2.0. But for β_s above 2.0 or below -0.50, separation becomes more common. For β_s larger than about 4.0 and smaller than about -2.0, though, a majority of the data sets feature separation.

Table 2:

β_s	Ideal Power	Chance of Separation	ML w/ Wald	ML w/ LR	ML w/ Score	PML (Firth) w/ Wald	PML (Cauchy) w/ Wald
5.00		62%	34%	98%	98%	92%	87%
4.00		35%	53%	92%	92%	85%	76%
3.00		14%	57%	75%	75%	76%	57%
2.00		4%	39%	45%	45%	70%	33%
1.00		1%	14%	16%	16%	74%	12%
0.75	As high as possible.	1%	10%	12%	12%	76%	8%
0.50		1%	7%	8%	8%	80%	6%
0.40		1%	6%	7%	7%	79%	5%
0.30		2%	5%	7%	7%	83%	5%
0.20		2%	4%	6%	5%	85%	3%
0.10		3%	4%	7%	7%	83%	4%
0.00	5%	3%	2%	6%	6%	86%	3%
-0.10		4%	2%	6%	6%	88%	3%
-0.20		5%	2%	6%	6%	88%	2%
-0.30		5%	2%	7%	6%	90%	2%
-0.40		6%	2%	8%	8%	91%	2%
-0.50		8%	1%	8%	8%	92%	2%
-0.75	As high as possible.	11%	1%	11%	11%	94%	2%
-1.00		16%	0%	16%	15%	96%	2%
-2.00		42%	0%	43%	41%	100%	7%
-3.00		70%	0%	69%	66%	100%	11%
-4.00		87%	0%	87%	83%	100%	14%
-5.00		95%	0%	95%	91%	100%	16%

This table shows the power for the Wald, likelihood ratio, and score tests for a data-generating process that often features separation, as well as the power for the Wald tests using Firth's and the Cauchy penalty. I selected this particular DGP to highlight tendencies in the larger collection, but this particular DGP is not necessarily representative in all respects. See Figure 3 for a more diverse collection.

These power functions clearly demonstrate the poor performance of the Wald test. Even though the data sets with separation should allow the researcher to reject the null hypothesis, at least occasionally, the power of the Wald test is low even for very large effects. This happens because the Wald test cannot reject the null hypothesis under separation. The cases where $\beta_s = 4.0$ and $\beta_s = 5.0$ show this clearly. The Wald test *fails to reject* when separation exists, but *does reject* the null hypothesis when separation does not exist (i.e., when the sample effects are *smaller*).

The likelihood ratio and score tests, on the other hand, perform as expected. For both, 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. This table also shows the Wald tests for Firth's and the Cauchy penalty. Compared to the likelihood ratio and score tests,

the Wald test using the Cauchy penalty is under-powered, especially (but not only) for negative values of β_s, and the Wald test using Firth's penalty rejects the null hypotheses far too often under the null. I emphasize that I selected this particular DGP to highlight tendencies in the larger collection, but this particular DGP is not necessarily representative in all respects. See Figure 3 for a more diverse collection.

A Broad Look at Many DGPs

Using the algorithm I describe above, I create a diverse collection of 150 DGPs. Figure 2 shows the power (i.e., the probability of rejecting the null hypothesis) of each test as the chance of separation varies across the many scenarios. Each point shows the power for a particular scenario (where $\beta_s \neq 0$, though some β_s are small). 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 a properly-functioning test should increase as well, because separation is evidence of a large coefficient. But because a large coefficient makes separation more likely, a large coefficient *decreases* the power of the Wald test. The likelihood ratio test, the score test, and the two Wald tests using penalized estimates do not exhibit this pattern.

Figure 3 shows the power function for each of the 150 DGPs in the diverse collection. Each of the many lines shows the power function for a particular DGP as β_s varies. First, the power functions for the Wald tests show its consistently poor properties. For most of the Wald power functions, as the true coefficient grows larger in magnitude from about two or 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 tests 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. Third, the Wald tests using the penalized estimates exhibit some troubling patterns. For the Cauchy penalty, the tests

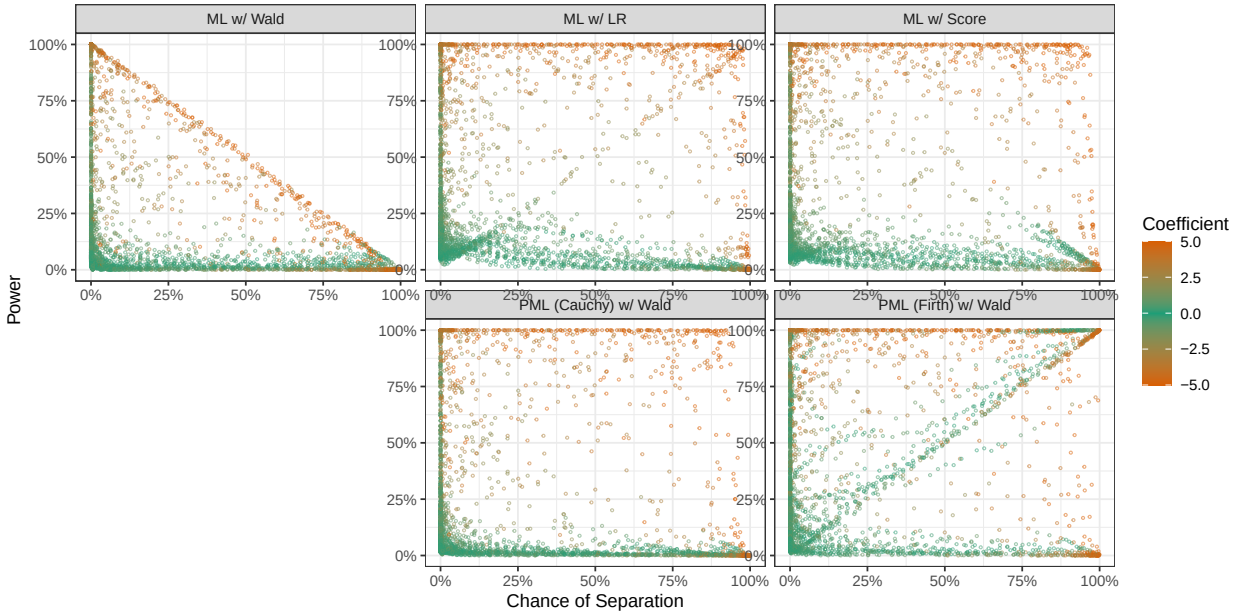

Figure 2: This figure shows the power of tests across a range of scenarios as the chance of separation varies.

seem under-powered relative to the likelihood ratio and score tests. For Firth's penalty, the chances of rejection when $\beta_s = 0$ seem high (e.g., 25% or more) for many DGPs.

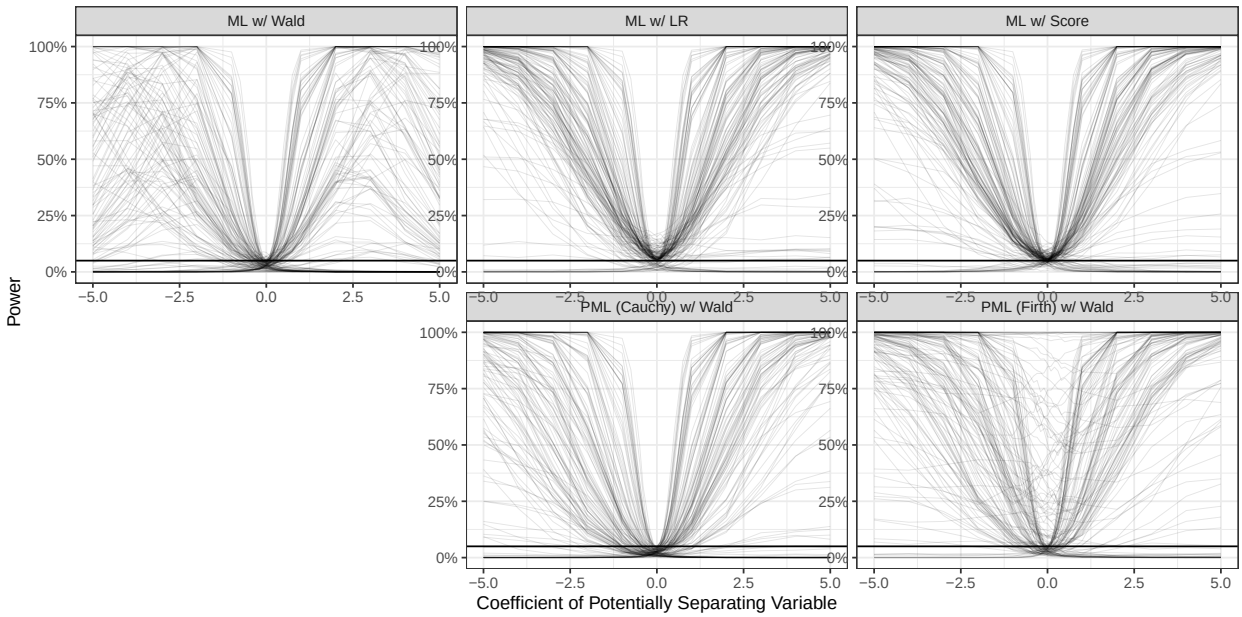

Figure 3: This figure shows the power for the Wald, likelihood ratio, and score tests for a diverse collection of data-generating processes, as well as the power for the Wald tests using Firth's and the Cauchy penalty.

Figure 4 summarizes the many power functions in Figure 3 using the median power across all 150 DGPs. The solid, dark line shows the median power and the two dashed lines show the 25th and 75th percentiles. This figure clearly shows the unusual behavior of the Wald test—the power *decreases* when the magnitude of the coefficient grows larger than about two or three. Further, it shows that both the likelihood ratio and score tests work well. For both tests, the chance of rejection is about 5% when the coefficient equals zero and grows quickly as the coefficient moves away from zero. This figure also shows that the likelihood ratio tests tend to be slightly more powerful than the score tests. The problematic patterns for the penalized estimators appear here as well. For the Cauchy penalty, the Wald tests can have relatively low power. For Firth's penalty, the Wald tests can reject the null hypothesis far too often when the null hypothesis is true.

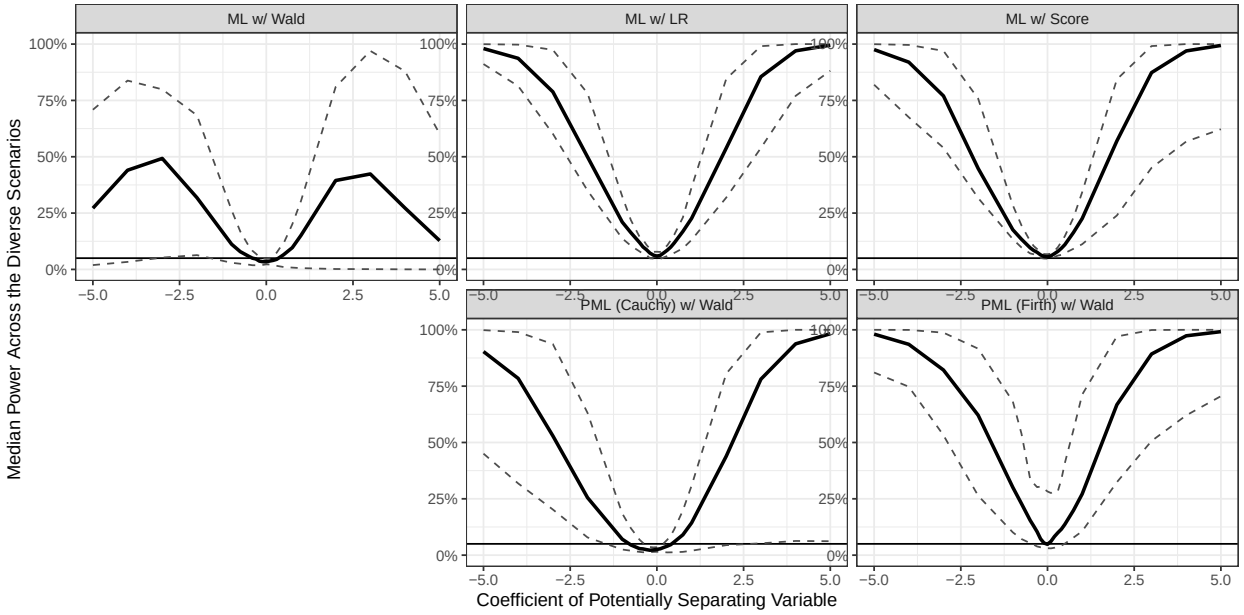

Figure 4: This figure shows the median, 25th percentile, and 75th percentile power for the Wald, likelihood ratio, and score tests for a diverse collection of data-generating processes, as well as for the Wald tests using Firth's and the Cauchy penalty.

To further see the behavior *under separation*, I divide the scenarios into three categories: low chance of separation, where the chance of separation is less than 10%; moderate chance of separation, between 10% and 30%; and high chance of separation, greater than

30%. Figure 5 shows the power for the various scenarios.

The bottom panel of Figure 5 is particularly helpful. When the chance of separation is high, the Wald tests rarely reject the null hypothesis. The likelihood ratio and score tests, on the other hand, still function as expected. Again, the likelihood ratio tests tend to exhibit *slightly* greater power than the score tests. The penalized estimates perform noticeably worse. Using the Cauchy penalty, the Wald test is under-powered compared to both the likelihood ratio and score tests. Using Firth's penalty, the results are even worse. Many of these tests reject the null in 50% or more repeated samples *when the null hypothesis that* $\beta_s = 0$ *is correct.*

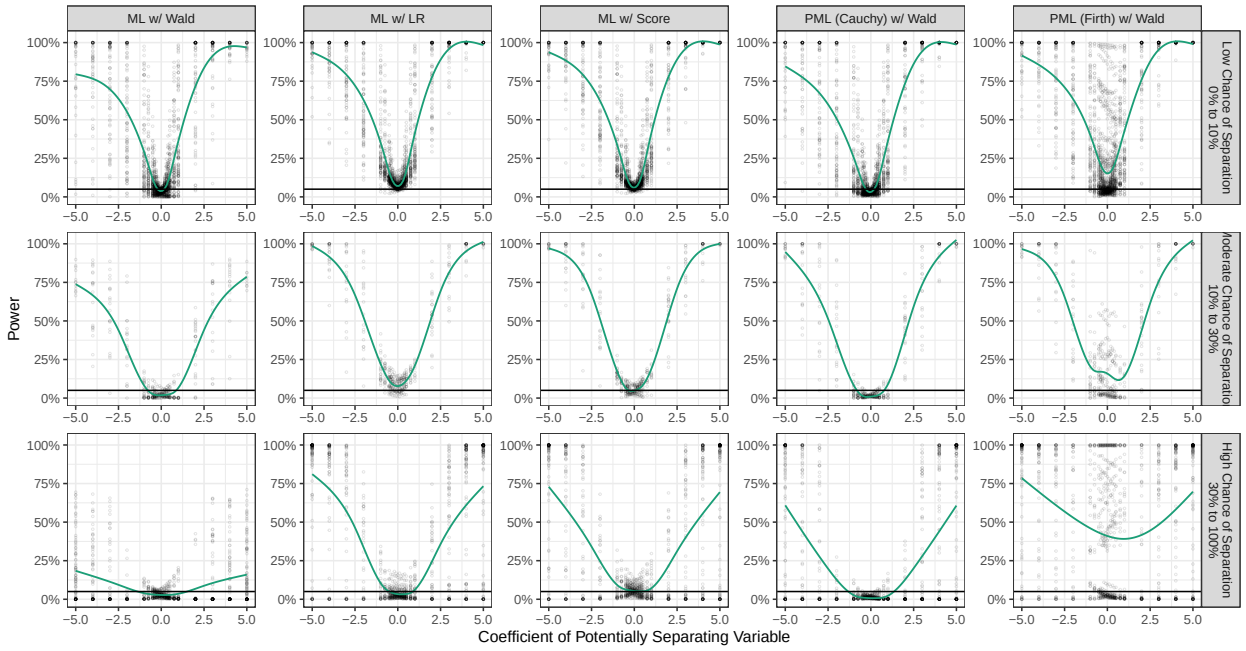

Figure 5: This figure shows the power for the Wald, likelihood ratio, and score tests for three levels of chance of separation, as well as for the Wald tests using Firth's and the Cauchy penalty. The smoothed lines are an additive model.

Finally, Figure 6 plots the size of the test (i.e., the chance of rejecting the null hypothesis that $\beta_s = 0$ when the null hypothesis is true) as the chance of separation varies for each of the 150 DGPs. Ideally, the size should be about 5%. The Wald, likelihood ratio, and score tests all have reasonable size. The size of the Wald tests falls between 2% and 5%, depending on the chance of separation. (The problem with the Wald tests is power, not

size.) The size of likelihood ratio tests falls between about 2.5% and about 10%. Notably, the likelihood ratio tests are over-sized when separation is relatively *un*likely. The size of the score tests falls around 5% regardless of the chance of separation. The Wald tests using the penalized estimates perform worse. Using the Cauchy penalty, the Wald test is under-sized, around 2% regardless of the chance of separation. Using Firth's penalty, the Wald test performs surprisingly poorly. For some DGPs, the Wald tests using Firth's penalized estimates *always* reject the null hypothesis when separation occurs, *even when separation is common under the null hypothesis*.[6] This underscores the advice of Rainey (2016) and Beiser-McGrath (2022) to treat default penalties with care.

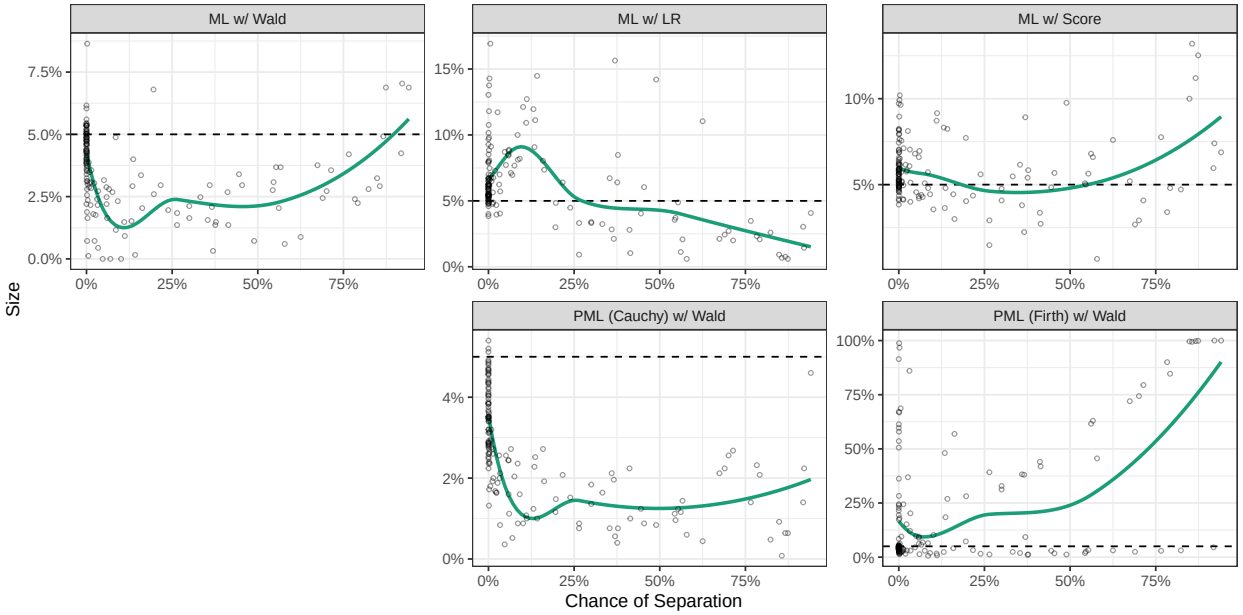

Figure 6: This figure shows the size for the Wald, likelihood ratio, and score tests for a diverse collection of data-generating processes, as well as the size for the Wald tests using Firth's and the Cauchy penalty. The smoothed lines are an additive model.

Concrete Recommendations

Given the arguments above, how should researchers proceed when facing separation? I offer the following four suggestions, which incorporate the arguments above with the

[6] Beiser-McGrath (2022) discusses the details of this unusual property of Firth's penalty in greater detail.

larger literature. Importantly, I view the likelihood ratio and/or score tests as *a supplement to* (not a replacement for) penalized estimation (e.g., Bell and Miller 2015) or Bayesian inference with informative priors (e.g., Barrilleaux and Rainey 2014).

1. *Identify separation.* Software varies in how and whether it detects and reports the presence of separation. Become familiar with your preferred software.[7]
2. *Do not drop the separating variable.* If a variable creates separation, then researchers might be tempted to omit the offending variable from the model. This is poor practice. See Zorn (2005, pp. 161-162) for more details.
3. *Test the hypothesis about the coefficient of the separating variable*. While the maximum likelihood estimate of the coefficient might be implausible (see Rainey 2016) and the Wald p-values nonsensical (see above), researchers can still use a likelihood ratio and/or score test to obtain a useful p-value and test the null hypothesis that the coefficient for the separating variable equals zero.[8] The researcher can report this test in the text of the paper and/or in a regression table, carefully distinguishing the likelihood ratio and/or score tests from the Wald test readers expect. In particular, I recommend the following three changes to the standard regression table:
    a. Replace the finite numerical maximum likelihood estimate with the theoretical maximum likelihood estimate of ∞ or $-\infty$. Researchers can code the binary separating variable so that values of 1 perfectly predict the outcome. This makes the ML estimate of the intercept finite and interpretable.[9]

[7] As of this writing, the `glm()` function in R *sometimes* reports a warning that "fitted probabilities numerically 0 or 1 occurred." The researcher must detect separation by noticing the unusually large coefficient estimate and standard error. Stata's `logit` and `probit` commands print a "note" that describes the separation. It then drops the problematic variable and the observations it perfectly predicts. See Konis (2007) for a thorough discussion of identifying separation.

[8] Even if the coefficient of the separating variable is not of substantive interest (i.e., the researcher does not have a specific hypothesis about its value; the researcher includes the separating variable only as a control), then I still recommend reporting this hypothesis test. By convention, researchers report p-values for model coefficients, but researchers should *not* report the Wald p-values because these are highly misleading. Researchers can either supply the p-value from the likelihood ratio and/or score test for completeness or omit the p-value for this variable.

[9] For simplicity, imagine that s perfectly predicts $y = 1$. To borrow the language of dummy variables for regression models, researchers might want to choose the "reference category" of the dummy variable (i.e., $s = 0$) so that $\hat{\beta}_{cons}$ is finite and $\hat{\beta}_s$ pushes the probability up to one, rather than choosing the

b. Omit the standard error estimate for the separating variable.

c. Replace the Wald p-value for the coefficient of the separating variable with the likelihood ratio p-value.[10] Clearly indicate this change. The simulations above suggest that the likelihood ratio test works *marginally* better than the score test in scenarios that commonly feature separation, so I suggest that researcher report the likelihood ratio test by default. In the table note, clearly explain the rationale for using the alternative test and supply the p-value from the score test as additional information. There is no need to replace the Wald p-values for variables that do not create separation with likelihood ratio or score p-values. The usual standard errors are meaningful and the Wald p-values work well for these variables that do not create separation, even when another variable in the model does create separation.[11]

4. *Estimate the coefficients and uncertainty using penalized maximum likelihood or Bayesian estimation and compute substantively meaningful quantities of interest*. Firth (1993; Zorn 2005) and Gelman et al. (2008) offer reasonable default penalties or prior distributions that might work well for a given application. However, Rainey (2016) and Beiser-McGrath (2022) show that the inferences can meaningfully depend on the

reference category so that $\hat{\beta}_{cons}$ is infinite and $\hat{\beta}_s$ pulls the probability down from one. Suppose a simple logistic regression model where $\Pr(y \mid s) = \beta_{cons} + \beta_s s$. Let $\widehat{\Pr}(\cdot)$ denote the maximum likelihood estimate of the probability of an event for a given s. If $\widehat{\Pr}(y \mid s = 0) = \text{logit}^{-1}(\hat{\beta}_{cons}) \in (0, 1)$ and $\widehat{\Pr}(y \mid s = 1) = \text{logit}^{-1}(\hat{\beta}_{cons} + \hat{\beta}_s) = 1$, then $\hat{\beta}_{cons}$ is finite and $\hat{\beta}_s = \infty$. But if we recode s so that $\widehat{\Pr}(y \mid s = 1) = \text{logit}^{-1}(\hat{\beta}_{cons} + \hat{\beta}_s) \in (0, 1)$ and $\widehat{\Pr}(y \mid s = 0) = \text{logit}^{-1}(\hat{\beta}_{cons}) = 1$ then $\hat{\beta}_{cons} + \hat{\beta}_s$ is finite and $\beta_{cons} = \infty$. For $\hat{\beta}_{cons} + \hat{\beta}_s$ to be finite while $\beta_{cons} = \infty$, $\hat{\beta}_s$ must equal $-\infty$. Thus, coding the separating variable s so that $s = 1$ perfectly predicts y produces the most interpretable table of regression coefficients. Importantly, this logic does not affect my main argument about testing the null hypothesis that $\beta_s = 0$—the likelihood ratio and score tests work well and the Wald test works poorly *regardless of the coding*.

[10] While I focus on how researchers can compute a reasonable p-value *without a penalty*, the simulations above support the claim of Rainey (2016) that default penalities do not necessarily produce reasonable tests. The simulations suggest that researchers should be especially skeptical of the p-values from the Wald test using Firth's penalized estimates—they show that these p-values can reject the null hypothesis incorrectly at well above the nominal rate for a variable that often creates separation when the null hypothesis is true. While the Cauchy penalty performs better than the Firth penalty, it is under-sized and under-powered relative to the likelihood ratio and score tests.

[11] Among my DGPs, the p-values produced by the likelihood ratio and Wald test for these other variables are nearly identical in almost all cases. The correlation between the two is 0.99 and the average absolute difference is less than 0.01. This does not necessarily apply to the intercept term, see footnote 9.

chosen penalty or prior. With this sensitivity in mind, the researcher should choose the penalty or prior *carefully* and demonstrate the robustness of their conclusions to alternative prior specifications. Researchers using penalized maximum likelihood can use the informal posterior simulation procedure suggested by King, Tomz, and Wittenberg (2000; see also Gelman and Hill 2006) to compute point estimates and confidence intervals for the quantities of interest. See Bell and Miller (2015) for an example. Researchers using full posterior simulation can transform the simulations of the model coefficients to obtain posterior simulations of the quantities of interest. See Barrilleaux and Rainey (2014) for an example. While researchers should rely primarily on a model with a thoughtful penalty or prior, it can be helpful to also report estimates using both Firth's (1993) and Gelman et al.'s (2008) default priors so that readers have a common benchmark.

Re-Analysis of Barrilleaux and Rainey (2014)

To illustrate the power and simplicity of frequentist hypothesis tests 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. Because no Democratic governors oppose the expansion, separation occurs—being a Democratic governor perfectly predicts non-opposition.

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 and test their hypothesis using the likelihood ratio and score tests.

Table 3 illustrates how a researcher can implement the third concrete suggestion above

(i.e., "test the hypothesis about the coefficient of the separating variable"). Table 3 does the following: (1) replaces the finite estimates returned by R's `glm()` function with the correct estimate of $-\infty$ and describes this change in footnote (a); (2) omits the problematic standard error estimate; (3) replaces the usual Wald p-value with the likelihood ratio p-value, clearly indicates this change, and explains the reason in footnote (b) (and leaves the remaining Wald p-values unchanged).

Substantively, Table 3 shows that the likelihood ratio test *unambiguously rejects* the null hypothesis that the coefficient for Democratic governors equals zero. That is, Democratic governors are less likely to oppose the Medicaid expansion than their Republican counterparts. The likelihood ratio and score p-values are 0.003 and 0.009 respectively.[12] This contrasts with the default penalized estimators, which produce a less-convincing pair of results. Firth's penalty gives a p-value of 0.060 and Gelman et al.'s (2008) suggested Cauchy penalty gives a p-value of 0.038.

Table 3:

	ML			PML w/ Firth's Penalty			PML w/ Cauchy Penalty		
	Coef. Est.	SE Est.	p-value	Coef. Est.	SE Est.	p-value	Coef. Est.	SE Est.	p-value
Democratic Governor	$-\infty$[b]	—	0.003 (LR)[b]	-2.677	1.421	0.060	-3.377	1.629	0.038
Percent Uninsured	0.923	2.234	0.680	0.180	1.127	0.873	0.599	1.078	0.578
Percent Favorable to ACA	0.128	1.549	0.934	-0.138	1.313	0.916	-0.209	1.035	0.840
Republican-Controlled Legislature	2.429	1.480	0.101	1.618	1.174	0.168	1.695	1.061	0.110
Fiscal Health	-0.054	0.854	0.950	-0.123	0.725	0.865	0.155	0.751	0.837
Medicaid Multiplier	-0.355	1.193	0.766	-0.326	1.018	0.748	-0.162	0.877	0.853
Percent Non-White	1.434	2.616	0.584	1.562	1.208	0.196	0.934	1.245	0.453
Percent Metropolitan	-2.759	1.687	0.102	-1.820	1.188	0.126	-1.459	1.044	0.162
Intercept	-0.715	0.667	0.283	-0.425	0.513	0.408	-0.561	0.524	0.284

[a] Being a Democratic governor perfectly predicts non-opposition, so these data feature separation. While the numerical maximum likelihood algorithm returns a finite estimate (about -20 for default precision and -36 for maximum precision), the maximum likelihood estimate is actually $-\infty$.
[b] Following the advice I develop above, I replace the default Wald p-value with a likelihood ratio p-value for this particular coefficient. The Wald test for maximum likelihood estimates relies on unreasonable standard errors that produce nonsensical p-values. However, the likelihood ratio and score tests produce reasonable p-values. The score test is another suitable alternative and produces a p-values of 0.009. The remainder of the p-values for all three models are from Wald tests.

This table provides the maximum likelihood estimates and penalized maximum liklihood estimates for Barrilleaux and Rainey's (2014) model explaining governors' opposition to Medicaid expansion. It illustrates how researchers might design their regression tables to include reasonable hypothesis tests for variables that create separation (i.e., likelihood ratio or score test, not Wald test).

In a complete analysis, the researcher should also compute substantively meaningful quantities of interest. While this (usually) requires a penalty or a prior, these estimates

[12]The nonsensical p-value from the Wald test is 0.995 using `glm()`'s default precision and 1.000 using the maximum precision.

are a critical part of a complete analysis of a logistic regression model with separation. Barrilleaux and Rainey (2014), Bell and Miller (2015), and Rainey (2016) offer examples of this important component. As Rainey (2016) emphasizes, though, estimates and confidence intervals for quantities of interest can depend heavily on the penalty or prior, so the researcher must choose their prior carefully and explore the robustness of results to other prior specifications.

Conclusion

Separation commonly occurs in political science. When this happens, I show that the usual p-values based on a Wald test are highly misleading. Zorn (2005) and Gelman et al. (2008) suggest that substantive researchers use penalized maximum likelihood to obtain reasonable point estimates and standard errors. However, Rainey (2016) and Beiser-McGrath (2022) urge substantive scholars to apply these default penalties cautiously. In this paper, I show that substantive researchers can use the usual likelihood ratio and score tests to test hypotheses about the coefficients, even under separation. While estimating quantities of interest (usually) requires a penalty or prior, researchers can use likelihood ratio or score tests to produce meaningful p-values under separation *without using penalties or prior information*.

Data Availability Statement

All data and code for the paper are available on the Open Science Framework (OSF) at https://doi.org/10.17605/OSF.IO/WN2S4 (Rainey 2023a) and Dataverse at https://doi.org/10.7910/DVN/6EYRJG (Rainey 2023b). A computational companion that illustrates how one can compute the quantities I discuss in the paper is available in the Supplementary Material on the publisher's website and in the OSF and Dataverse

repositories.

Acknowledgements

I am grateful to an especially thoughtful and careful pool of peer reviewers that helped me make this paper better.

References

Albert, Adelin, and John A Anderson. 1984. "On the Existence of Maximum Likelihood Estimates in Logistic Regression Models." *Biometrika* 71(1): 1–10.

Anderson, Noel, Benjamin E Bagozzi, and Ore Koren. 2021. "Addressing Monotone Likelihood in Duration Modelling of Political Events." *British Journal of Political Science* 51(4): 1654–71.

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.

Beiser-McGrath, Liam F. 2022. "Separation and Rare Events." *Political Science Research and Methods* 10(2): 428–37.

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

Buse, Adolf. 1982. "The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An Expository Note." *The American Statistician* 36(3a): 153–57.

Casella, George, and Roger L. Berger. 2002. *Statistical Inference*. 2nd ed. Pacific Grove, CA: Duxbury.

Cook, Scott J, Jude C Hays, and Robert J Franzese. 2020. "Fixed Effects in Rare Events Data: A Penalized Maximum Likelihood Solution." *Political Science Research and Methods* 8(1): 92–105.

Cook, Scott J, John Niehaus, and Samantha Zuhlke. 2018. "A Warning on Separation in Multinomial Logistic Models." *Research & Politics* 5(2): 1–5.

Crisman-Cox, Casey, Olga Gasparyan, and Curtis S. Signorino. 2023. "Detecting and Correcting for Separation in Strategic Choice Models." *Political Analysis* 31(3): 414–29.

de Carvalho Barreto, Ikaro Daniel, Suzana Leitão Russo, Gutemberg Hespanha Brasil, and Vitor Hugo Simon. 2014. "Separation Phenomena Logistic Regression." *REVISTA GEINTEC-GESTAO INOVACAO E TECNOLOGIAS* 4(4): 716–28.

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

Gelman, Andrew, and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

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.

Greene, William H. 2012. *Econometric Analysis*. 7th ed. Upper Saddle River, New Jersey: Prentice Hall.

Gustafson, Daniel. 2020. "Hunger to Violence: Explaining the Violent Escalation of Nonviolent Demonstrations." *Journal of Conflict Resolution* 64(6): 1121–45.

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

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44(2): 341–55.

Konis, Kjell. 2007. "Linear Programming Algorithms for Detecting Separated Data in Binary Logistic Regression Models." PhD thesis. University of Oxford.

Mehltretter, Andreas. 2021. "Arming for Conflict, Arming for Peace? How Small Arms Imports Affect Intrastate Conflict Risk." *Conflict Management and Peace Science*: Forthcoming.

Nagashima, Kengo, and Yasunori Sato. 2017. "Information Criteria for Firth's Penalized Partial Likelihood Approach in Cox Regression Models." *Statistics in medicine* 36(21): 3422–36.

Owsiak, Andrew P, and John A Vasquez. 2021. "Peaceful Dyads: A Territorial Perspective." *International Interactions* 47(6): 1040–68.

Rainey, Carlisle. 2016. "Dealing with Separation in Logistic Regression Models." *Political*

Analysis 24(3): 339–55.

———. 2023a. "OSF Project for: 'Hypothesis Tests Under Separation'." https://doi.org/10.17605/OSF.IO/WN2S4.

———. 2023b. "Replication Data for: 'Hypothesis Tests Under Separation'." https://doi.org/10.7910/DVN/6EYRJG.

Rao, C Radhakrishna. 1948. "Large Sample Tests of Statistical Hypotheses Concerning Several Parameters with Applications to Problems of Estimation." In *Mathematical Proceedings of the Cambridge Philosophical Society*, Cambridge University Press, 50–57.

Sun, Chi et al. 2015. "Juvenile Dermatomyositis: A 20-Year Retrospective Analysis of Treatment and Clinical Outcomes." *Pediatrics & Neonatology* 56(1): 31–39.

Wald, Abraham. 1943. "Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations Is Large." *Transactions of the American Mathematical society* 54(3): 426–82.

Wilks, Samuel S. 1938. "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses." *The annals of mathematical statistics* 9(1): 60–62.

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