Reasonable Measures of Uncertainty Under Separation*

Carlisle Rainey†

ABSTRACT

When facing data sets with small numbers of observations or "rare events," political scientists often encounter important explanatory variables that perfectly predict binary events or non-events. In this situation, maximum likelihood provides implausible estimates and the researcher must incorporate some form of prior information in the estimation. The most sophisticated research uses Jeffreys' invariant prior to stabilize the estimates. While Jeffreys' prior has the advantage of being automatic, I show that, in many cases, it offers too much prior information, providing confidence intervals that are much too narrow. I show that the choice of a more reasonable prior can lead to different substantive conclusions about the likely magnitude of an effect and I offer practice advice for choosing a prior distribution that represents actual prior information.

*I thank [many people]. Thanks to Mark Bell and Nicholas Miller for making their data available to me. The analyses presented here were conducted with R 3.1.0 and JAGS 3.3.0. All data and computer code necessary for replication are available at github.com/carlislerainey/priors-for-separation.

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

The Logistic Regression Model

Political scientists commonly use logistic regression to model the probability of an event of interest. In the typical situation, the researcher uses an $n \times (k+1)$ design matrix X consisting of an intercept and k covariates to model a vector n of binary outcomes y, where $y_i \in \{0, 1\}$ using the model $\Pr(y_i) = \Pr(y_i = 1|X_i) = \dfrac{1}{1+e^{-X_i\beta}}$, where β is a parameter vector of length $k+1$.

Using this model, it is straightforward to calculate the likelihood function

$$\Pr(y|\beta) = L(\beta|y) = \prod_{i=1}^{n}\left[\left(\frac{1}{1+e^{-X_i\beta}}\right)^{y_i} + \left(\frac{1}{1+e^{-X_i\beta}}\right)^{1-y_i}\right].$$

As usual, one can take the natural logarithm of both sides to calculate the log-likelihood function

$$\log L(\beta|y) = \sum_{i=1}^{n}\left[y_i \log\left(\frac{1}{1+e^{-X_i\beta}}\right) + (1-y_i)\log\left(\frac{1}{1+e^{-X_i\beta}}\right)\right]$$

and take the derivatives of the log-likelihood function to obtain the score function

$$\frac{\partial \log L(\beta|y)}{\partial \beta} = \sum_{i=1}^{n}\left(y_i - \frac{1}{1+e^{-X_i\beta}}\right)X_i.$$

Researchers routinely obtain maximum likelihood estimates $\hat{\beta}^{mld}$ of the model parameters β by setting the score function equal to zero and solving for β (i.e., maximizing the likelihood of the observed data) and estimate the standard errors are by calculating the square root of the diagonal of the inverse of Fisher's information matrix evaluated at $\hat{\beta}^{mle}$ (i.e., calculate the curvature around the maximum of the likelihood function to obtain an estimate of the uncertainty of the estimate). While this approach works quite well in most applications, it fails in a situation known as separation (Zorn 2005).

Separation

Separation occurs in models of binary outcome data when one explanatory variable (or perhaps a combination of explanatory variables, see Lesaffre and Albert (1989)) perfectly predicts zeros, ones, or both. *Complete separation* occurs when the "problematic" explanatory variable s_i (for *s*eparating explanatory variable) perfectly predicts both zeros and ones and *quasicomplete separation* occurs when s_i perfectly predicts either zeros or ones, but not both (Albert and Anderson 1984; Zorn 2005). *Overlap*, the ideal case, occurs when there is no such s_i that perfectly predicts zeros or ones.[1] In this situation, the usual maximum likelihood estimates exist and provide reasonable estimates of parameters. However, under complete or quasicomplete separation, maximum likelihood estimates do not exist (Albert and Anderson 1984; Zorn 2005).

Complete separation occurs when a covariate perfectly predicts zeros and ones. For example, suppose an explanatory variable s_i , such that $y_i = 0$ for $s_i > 0.5$ and $y_i = 0$ for $s_i \leq 0.5$. This corresponds to the middle panel of Figure 1. To maximize the likelihood of the observed data, the "S"-shaped logistic regression curve must assign probabilities of zero when $s_i \leq 0.5$ and probabilities of one when $s_i > 0.5$. Since the logistic regression curve lies strictly between zero and one, this likelihood cannot be achieved. However, it can be approached asymptotically as the coefficient for s_i approaches infinity. Thus, the likelihood function under complete separation is monotonic (has no maximum) and a finite maximum likelihood estimate does not exist.

Quasicomplete separation occurs when a covariate perfectly predicts zeros *or* ones. Figure 1 shows an example pattern in the right panel, where y_i always equals zero when s_i equals zero. This situation occurs often in applied political science research with binary inputs. For example, Gelman et al. (2008) find no African-American respondents in their data support Barry Goldwater in 1964, leading to a maximum likelihood estimate of negative infinity for the coefficients for the indicator of African-American respondents. Similarly, Rauchhaus (2009) (see

[1] [tk: cite] explain that combinations of explanatory variables can also produce separation, but I focus here on single separating explanatory variables.

Bell and Miller 2014) finds no instances of states with nuclear weapons engaging in war with each other. In this case, the estimated coefficient for the variable indicating symmetric nuclear dyads (in which both states possess nuclear weapons) equals negative infinity. To maximize the likelihood in this situations, the model must assign probabilities of zero to observations for which $s_i = 1$ (African-American respondents or symmetric nuclear dyads, in these examples). Again, because the logistic regression curve lies strictly above zero, this cannot happen, though it can be approached asymptotically as the coefficient for s_i goes to negative infinity. As with complete separation, the likelihood function under quasicomplete separation is monotonic (has no maximum) and a finite maximum likelihood estimate does not exist.

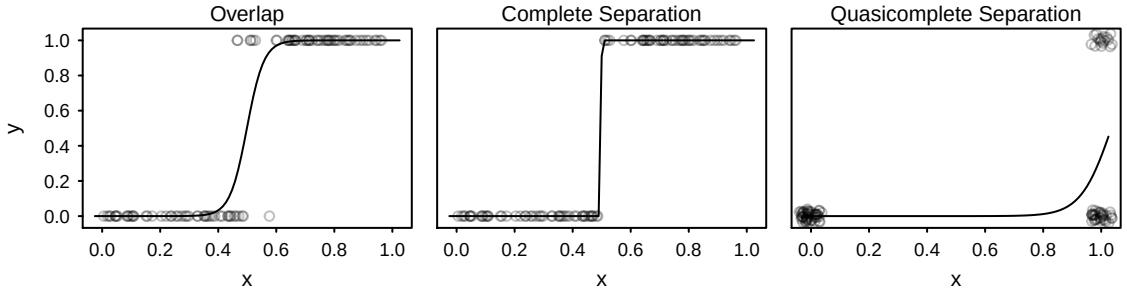

FIGURE 1: This figure illustrates overlap, complete separation, and quasicomplete separation as defined by Albert and Anderson (1984). The maximum likelihood estimates only exist under overlap. Under complete and quasicomplete separation, maximum likelihood fails, returning infinite estimates and standard errors.

Along coefficient estimates under separation are infinite in theory, the hill-climbing algorithms approximate the infinite estimates with large, finite values that increase with the precision of the algorithm. Table 1 shows the estimates from R's `glm()` function for each of the hypothetical data sets in Figure 1. To illustrate the problem, I vary the convergence tolerance between $\epsilon = 10^{-8}$ (the default) and $\epsilon = 10^{-16}$. With the default tolerance, `glm()` returns very large estimates and standard errors. There is no finite maximum, but the likelihood is flat enough around the large estimates to satisfy the algorithm, given the default tolerance. With the more sensitive tolerance of $\epsilon = 10^{-16}$, the algorithm converges even closer to infinity,

but still returns finite estimates. This is a failure of maximum likelihood, because estimates of infinity are usually implausible (i.e., there is always some, though perhaps tiny, probability of a zero or one in any given situation, see Heinze and Schemper 2002).

	Overlap		Complete Separation		Quasicomplete Separation	
	$\epsilon = 10^{-8}$	$\epsilon = 10^{-16}$	$\epsilon = 10^{-8}$	$\epsilon = 10^{-16}$	$\epsilon = 10^{-8}$	$\epsilon = 10^{-16}$
constant	−17.05	−17.05	−739.27	−739.27	−19.57	−26.57
	(5.79)	(5.79)	(139744.58)	(139744.58)	(1520.85)	(50363.70)
x	34.19	34.19	1483.21	1483.21	18.90	25.90
	(11.92)	(11.92)	(280801.40)	(280801.40)	(1520.85)	(50363.70)
Log Likelihood	-9.74	-9.74	0.00	0.00	-32.05	-32.05
Num. obs.	100	100	100	100	100	100

Standard errors in parentheses.

TABLE 1: A table providing estimates based on the data shown in Figure 1 using the R function `glm()` varying the convergence tolerance under overlap, complete separation, and quasicomplete separation. Notice that the estimation algorithm returns estimates and standard errors arbitrarily close to infinity under both types of separation as the tolerances shrinks to zero.

Perhaps more starkly, notice that the strong pattern in the middle panel of Figure 1 does not produce statistically significant result. This is because the likelihood is essentially flat around the "maximum" found by the hill-climbing algorithm. As the region around the maximum flattens, the estimates of the standard errors increases. Thus, separation leads to implausible large estimates *and* standard errors. Notice, for example, that while the data in the middle and right panel would almost never occur under the null hypothesis of no relationship, none of the estimates are statistically significant.

Solutions to Separation

The maximum likelihood framework (ML) requires the researcher to find the parameter vector that "maximizes the likelihood of the observed data." Of course, infinite coefficients *always* generate separated data, while finite coefficients only generate separation *sometimes*. Thus, under separation, the researcher can only obtain infinite estimates using ML.

Before addressing potential solutions, let me mention two unsatisfactory "solutions" found in applied work. In some cases, researchers simply ignore the problem of separation and interpret the large estimates and standard errors as though these are reasonable. However, this approach leads researchers to overstate the magnitude of the effect and the uncertainty of those estimates. A second approach used in practice is to drop the separating variable from the model. Zorn (2005) correctly dismisses this approach:

> As a practical matter, separation forces the analyst to choose from a number of problematic alternatives for dealing with the problem. The most widely used "solution" is simply to omit the offending variable or variables from the analysis. In political science, this is the approach taken in a number of studies in international relations, comparative politics, and American politics. It is also the dominant approach in sociology, economics, and the other social sciences, and it is the recommended method in a few prominent texts in statistics and econometrics. Of course, this alternative is a particularly unattractive one; omitting a covariate that clearly bears a strong relationship to the phenomenon of interest is nothing more than deliberate specification bias.

As an alternative, Zorn (2005) recommends building prior information $p(\beta)$ into the model using Bayes' rule, so that

$$p(\beta|y) = \frac{\overbrace{p(y|\beta)}^{likelihood}\overbrace{p(\beta)}^{prior}}{\int p(y|\beta)p(\beta)d\beta}.$$

In this case, the estimate switches to from the maximum likelihood estimate to a summary of the location of the posterior distribution, such as the posterior mode or mean. The current literature on dealing with separation suggests researcher take an automatic approach, adopting Jeffrey's invariant prior distribution.

Jeffrey's Invariant Prior

Zorn (2005) introduces political scientists to Firth's 1993 modified score function (see Heinze and Schemper 2002 as well). Firth suggests replacing the usual log-likelihood function $\log L(\beta|y)$ with a "penalized" likelihood function $\log L(\beta|y)$, so that $\log L^*(\beta|y) = \log L(\beta|y)|I(\beta)|^{\frac{1}{2}}$. It turns out that this penalty is equivalent to Jeffreys' prior for the logistic regression model. The posterior distribution can be obtained by applying Jeffreys' (1946) rule, which requires setting the prior $p(\beta)$ to be proportional to the square root of the determinant of the information matrix, so that $p(\beta) = |I(\beta)|^{\frac{1}{2}}$. Then, of course, applying Bayes' rule to obtain the posterior distribution $p(\beta|y) \propto L(\beta|y)|I(\beta)|$, so that Firth's penalty approach is equivalent to a Bayesian approach with Jeffreys' prior.

The usual method of obtaining standard errors is to assume a multivariate normal sampling distribution for the parameter vector β and obtain the (i, j) entry of the covariance matrix Σ_β by calculating the curvature around the maximum likelihood using $\frac{\partial^2 \log L(\beta|Y)}{\partial \beta_i \partial \beta_j}$ or posterior mode $\frac{\partial^2 \log p(\beta|Y)}{\partial \beta_i \partial \beta_j}$.

But (Heinze and Schemper 2002) and Zorn (2005) point out that this asymptotic approximation of the sampling distribution as normally (symmetrically) distributed might not be appropriate under separation. As an alternative, the suggest using likelihood profiling to obtain

the desired confidence interval. They suggest that analysts obtain a $(1-\alpha)100\%$ confidence interval for the model parameter β_i by calculating the (continuous) set of values for which the likelihood-ratio falls below the $(1-\alpha)100$ percentile of the χ^2_1 distribution (Heinze and Schemper 2002).

However, Firth (1993) did not propose his prior to solve the separation problem. Instead, his purpose was to reduce the well-known small sample bias in logistic regression models. And while it is true that Firth's correction does provide finite estimates under separation, it remains an open question as to whether this automatic prior, designed for other purposes, provides a reasonable estimate of the uncertain of the estimates.

The Cauchy Prior

Indeed, Gelman et al. (2008) note that Firth's application of Jeffreys' prior is not easily interpretable as an actual application of prior information because the prior $p(\beta) = |I(\beta)|^{\frac{1}{2}}$. Instead, they suggest an informative prior that, like Jeffreys' prior, bounds the estimates away from positive and negative infinity but has a scale parameter σ that allows researchers to choose the amount of prior information provided to the model. They suggest using the Cauchy distribution as priors for the model coefficients.[2]

The Cauchy distribution resembles a normal distribution, but has much heavier tails (it is equivalent to a t distribution with one degree of freedom), which captures the prior belief that the effect is probably smaller, but has some chance of being quite large.

[2]Though Gelman et al. (2008) allow for various user-defined prior distributions, they suggest a Cauchy prior centered at zero with scale of 2.5 as a default for appropriately rescaled variables. They suggest rescaling binary inputs to have mean zero and rescaling continuous inputs to have a mean of zero and a standard deviation of one half before estimating the model.

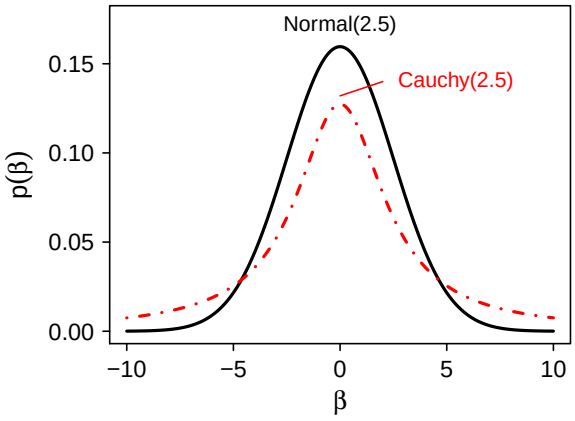

FIGURE 2: This figure compares the Normal(2.5) and Cauchy(2.5) prior distributions. Notice that the Cauchy distribution has a similar shape to a Normal distribution, but has much heavier tails. Substantively, this prior notes that the coefficient is likely close to zero (e.g. $|\beta| < 2$), but might be quite large (e.g., $|\beta| > 5$).

The posterior distribution is not easily available analytically, but one can easily use MCMC to simulate from the posterior distribution. [3] To facilitate this estimation, I have written an R function, `cauchy`, to obtain posterior simulations.[4] Once a researcher has the MCMC simulations, she can obtain the $(1-\alpha)100\%$ Bayesian credible interval for parameters by summarizing the simulations.

[3] (Gelman et al. 2008) obtain standard errors of their estimates by simply calculating the curvature around the posterior mode. They note that although the entire posterior distribution can be computed using MCMC methods, "it is desirable to have quick calculations that returns a point estimate of the regression coefficients and the standard errors" (p. 1366). They note that this approximation "works well in statistical practice and, in addition, recognizes the approximate nature of the model itself." However, our purpose is to accurately characterize the uncertainty of the estimates, and asymptotic approximations can underestimate credible intervals by quite a bit.

[4] The R package `MCMCpack` offers some potential to sample from the posterior distribution via a Metropolis-Hastings algorithm with little custom software via the `MCMClogit()` function. However, the default proposal distribution relies on the large-sample covariance matrix for the parameter estimates. This covariance matrix, of course, is not useful since the variances for the separated variables are much too large. As far as I can tell, the contribution of the large-sample covariance matrix to the proposal distribution is not easily modified. For this reason, I have written custom samplers that use a Metropolis-Hastings algorithm to obtain posterior simulations.

The Importance of the Prior

Choosing a reasonable prior distribution is crucial for dealing with separation in a substantively meaningful manner. In many cases, the data (though the likelihood) swamp the contribution of the prior. However, in the case of separation such that s_i perfectly predicts events, the likelihood determines the shape of the left-hand side of the posterior distribution and the prior (symmetric about zero) determines the shape of the right hand side of the posterior.

The likelihood has an "S"-shape that approaches a limit of one as the parameter coefficient for the separating variable s_i approaches infinity. Thus, for large values of the coefficient, the likelihood is essentially flat, which allows the prior distribution to drive the inferences. Thus the prior distribution is not an arbitrary choice made for computational convenience–but a choice that affects the inferences.

The Impact of the Prior in Theory

Suppose that an explanatory variable s_i perfectly predicts a binary outcome variable $y_i = 1$, such that whenever $s_i = 1$, $y_i = 1$, but when $s_i = 0$, y_i might equal zero or one. Albert and Anderson (1984) refer to this situation as quasicomplete separation. Suppose further an additional set of covariates X_i and the analyst wishes to obtain plausible estimates of coefficients the model $Pr(y_i = 1) = \text{logit}^{-1}(\alpha + \delta s_i + X_i\beta)$. It is easy to find plausible estimates of β using the techniques discussed above (even maximum likelihood usually provides reasonable estimates of these parameters), but finding plausible estimates of α and δ proves more difficult because maximum likelihood suggests estimates of $-\infty$ and $+\infty$, respectively. In order to obtain a plausible estimate of δ (which will, in turn, provide a plausible estimate of α), the researcher must introduce prior information into the model. My purpose here is to characterize how this prior information impacts the posterior distribution.

In the general situation, the analyst is interested in computing and characterizing the

posterior distribution of the coefficient for s_i given the data. Using Bayes' Rule, this posterior depends on the likelihood and the prior, so that $p(\delta, \beta|y) = p(y|\beta, \delta)p(\beta, \delta)$.[5] In particular, the analyst might have in mind a family of priors centered at and monotonically decreasing away from zero with varying scale σ, so that $p(\delta) = p(\delta|\sigma)$. Suppose that for a particular $\delta^* \geq 0$ the prior distribution is decreasing in δ at a decreasing rate. Intuitively, this assumption of a δ^* allows the result to generalize to many common distributions.[6] Finally, suppose that the informativeness of the prior distribution depends on scale parameter σ that is chosen by the researcher and "flattens" the prior $p(\delta) = p(\delta|\sigma)$, such that as σ increases, the rate a which the prior descends to zero decreases.

Theorem 1. *The impact of the researchers choice of σ on the posterior distribution $p(\delta|y)$ is increasing in δ for $\delta > \delta^*$.*

In many cases, researchers summarize the posterior distribution by providing the 5th and 95th percentiles and a measure of centrality, such as the median.

> PRACTICAL IMPLICATION OF THEOREM 1: Under quasicomplete separation where x_i perfectly predicts $y_i = 1$, the prior has a small impact on the lower bound of the 90% credible interval, a moderate impact on the measures of the location of the posterior (i.e., mean, median, and mode), and a large impact on the upper-bound of the credible interval.

The Impact of the Prior in Practice

To illustrate the impact of the prior on inferences when facing separation, I replicate a results from Barrilleaux and Rainey (2014), who are interesting in the effect of partisanship on

[5] If we wanted only the posterior distribution of the scalar δ, we could integrate out the parameter vector β, giving $p(\delta|y) = \int_{-\infty}^{\infty} p(y|\beta, \delta)p(\beta, \delta)d\beta$.

[6] In particular, if the prior distribution is in the form of a double-exponential, which lacks "shoulders," then $\delta^* = 0$. However, the most common prior distributions used in applied work, such as the normal, t, and Jeffreys', have "shoulders" such that $\delta^* > 0$. In this case, the exact curvature of the distribution in the region $[0, \delta^*]$ affects the relative impact of the prior.

governors' decisions to oppose the Medicaid expansion in their states under the Patient Protection and Affordable Care Act (ACA).[7] As the authors note, no Democrats opposed the expansion leading to a problem of separation. I use MCMC to simulate from the posterior using several different prior distributions, including Jeffreys' prior (Zorn 2005) and the Cauchy prior with scales of 1, 2.5, and 5 (Gelman et al. 2008). While the choice of prior does not affect the conclusion about the *direction* of the effect, it has a large impact on the conclusion about the *magnitude* of the effect. This can be especially important when researchers are making claims about the substantive importance of their estimated effects (see Rainey 2014, Gross 2014, and McCaskey and Rainey 2014).

Figure 3 shows the posterior distribution for the coefficient for the indicator of Republican governors. Notice that the different priors lead to different posterior distributions. Notice, in particular, that the choice of the prior has a large impact on the right-hand side of the posterior. More informative priors (e.g., Jeffrey's prior) lead to a more peaked posterior distribution that rules out very large effects. Less informative priors (e.g., Cauchy(2.5)) lead to the conclusion that even large effects are plausible. These differences affect the conclusions that the researchers draw about the likely magnitude of the effect.

[7] Barrilleaux and Rainey (2014) use a logistic regression modeling the probability that a governor opposes the expansion using the following explanatory variables: the partisanship of the governor, the percent of the state's residents who are favorable toward the ACA, whether Republicans control the state legislature, the percent of the state that is uninsured, a measure of the fiscal health of the state, the Medicaid multiplier for the state, the percent of the state that is nonwhite, and the percent of the state that resides in a metropolitan area. See their paper for more details.

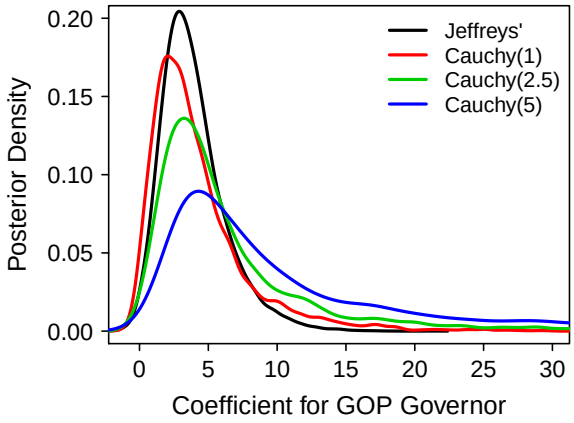

FIGURE 3: This figure provides the posterior distribution for the coefficient of the indicator for GOP governors in the model offered by Barrilleaux and Rainey (2014). Notice that the location and the spread of the posterior depend on the prior chosen, especially the right-hand side of the distribution.

Figure 4 shows how the choice of prior impacts the 90% credible interval. Notice that different prior distributions lead to different conclusions about the plausible values of the effect. In particular, different priors lead to different conclusions about the upper-bound on the plausible effect sizes. For example, Jeffreys' prior, the default proposed by Zorn (2005) and Heinze and Schemper (2002), suggests the effect lies in the range $\beta_{\text{GOP Gov.}} \in [0.9, 8.4]$, with a posterior mean of 3.9. On the other hand, the less informative Cauchy(2.5) prior, the default proposed by Gelman et al. (2008), suggests the effect lies in the range $\beta_{\text{GOP Gov.}} \in [1.0, 22.5]$, with a posterior mean of 7.3. A simple change from one proposed default to another more than doubles the upper bound on the 90% credible interval and almost doubles the posterior mean. Further, the Cauchy(5) prior, a plausible prior if one believes the effect might be large, produces the upper-bound on the 90% credible interval from is more than four times larger than the upper-bound produced by Jeffrey's prior. The posterior mean from the Cauchy(5) prior is larger falls above the upper-bound from the 90% credible interval from Jeffrey's prior.

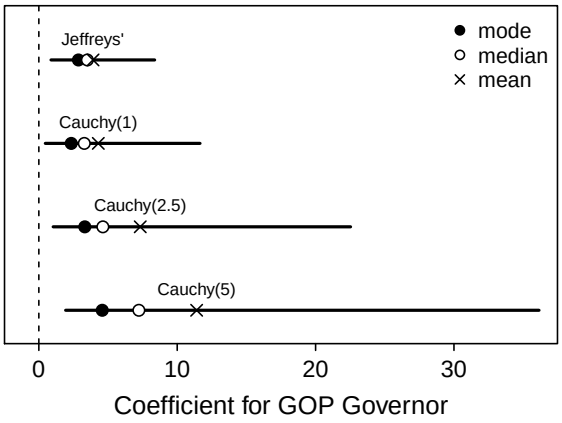

FIGURE 4: This figure provides the (equal-tailed) 90% credible intervals for the coefficient of the indicator for GOP governors in the model offered by Barrilleaux and Rainey (2014). Notice that the location and the spread of the posterior depend on the prior chosen, especially the right-hand side of the distribution. Note that Jeffrey's prior, suggested by Zorn (2005), is the most informative of these priors, suggesting that a coefficient smaller than about 10 is quite unlikely. On the other hand, credible interval using the Cauchy(2.5) prior, as suggested by Gelman et al. (2008), is about *twice* as wide as the credible interval from Jeffreys' prior. Finally, notice that the Cauchy(5) prior–a plausible prior if the researcher believes the effect might be large–produces a posterior mean larger than the upper bound of the 90% credible interval using Jeffrey's prior.

This leads us to conclude that the choice of prior matters–it affects the inferences that we draw from the data. It is not sufficient to rely on the prior distribution designed as a default for other purposes. Instead, we must rely on prior distributions that represent actual prior information about the likely magnitude of the coefficients.

Choosing an Appropriate Prior Distribution

Whilet is often sufficient to rely on default priors, this is not the case if one is interested in obtaining reasonable measures of uncertainty under separation. Indeed, I show above that the width of the 90% credible interval and the posterior mean depend largely on the prior one chooses in the replication of Barrilleaux and Rainey (2014). This implies that researchers relying on default priors alone riskunder- or over-representing their confidence in the magnitude

of the effect.

When facing separation, our goal is to choose a prior distribution that satisfies three properties:

1. Appropriately skeptical. As mentioned before, the prior distribution largely drives the right-hand side of the posterior distribution when s_i perfectly predicts $y_i = 1$ and the left-hand side of the distribution when s_i perfectly predicts $y_i = 0$. In this case, a non-central prior distribution in the direction of the separation has an especially large impact on the inferences. For this reason, I focus on prior distributions centered at zero to conservatively pool coefficients downward toward zero.

2. Allow plausible effects. The prior distribution should assign realistic prior probabilities to estimates that are *a priori* plausible.

3. Rule out implausible effects. The prior distribution should assigned essentially know prior probability to estimates that are *a priori* implausible.

Prior Predictive Distribution

One method of assessing the reasonableness of the prior distribution is "prior predictive checks," in which the researcher asks herself whether the prior distribution (combined with the model) produces a distribution for the data that matches her prior beliefs. Under the Bayesian framework, the researcher has a fully specified model $p(y_{new}|\theta)p(\theta)$ and can thus simulate data y_{new} from the model prior to observing the data. The distribution of the unobserved outcome y_{new} is given by $p(y_{new}) = \int p(y_{new}|\theta)p(\theta)d\theta$ (Box 1980). In practice, this process involves Clarify-like simulation (King, Tomz, and Wittenberg 2000), but rather than using the asymptotic posterior (e.g., $\beta_{sim} \sim N\left[\hat{\beta}^{mle}, I(\hat{\beta}^{mle})^{-1}\right]$), researchers simulate the model parameters from the prior distribution. Just as when using simulation to interpret the estimates, researchers can use simulation to interpret the prior distribution.

However, it is difficult to work with more than one dimension of the prior distribution at once and choosing the full prior distribution requires simultaneously choosing prior distributions for the $k+1$ explanatory variables, as well as the relationships among these variables. But not all of this $k+1$ dimensional distribution is of practical importance. Since the data swamp the prior for all but the separated coefficient, only the "slice" of the prior distribution with all other coefficients near their maximum likelihood estimates are relevant. I refer to this simplified focus as the *partial* prior predictive distribution.

Partial Prior Predictive Distribution

In many applications, the data swamp the prior so that changes in the prior distribution lead to only small changes in the inferences. In the case of separation, this is also the case, except for the separating variable in the direction of the separation. The likelihood makes a large contribution to the posterior for the non-separated coefficients and to the separated coefficient against the direction of the separation. However, the choice of the prior distribution has a large impact on the prior distribution for the separated coefficient in the direction of the separation.

Thus, in handling separation, we need to focusing on choosing reasonable prior distribution for the coefficient of the separating variable that assigned higher probabilities to more plausible estimates and lower probabilities to less plausible estimates.

For example, partisanship might have only a negligible effect on governors' decisions to oppose expansion. This is inconsistent with the hypothesis proposed by Barrilleaux and Rainey (2014), but it offers a *plausible* alternative. On the other hand, we can *a priori* rule out Democrats being 100% likely to support the expansion and/or Republicans being 100% likely to oppose–there is always some chance of Democratic governors opposing the expansion and Republican governors supporting it. However, the trick is to use substantive knowledge of the process to decide which effects are implausible large and which are not.

the we need to focus on using the prior distribution to rule out only implausible values in the

direction of the separation [ctk: I need to define the direction of the separation]. For example, if no Democratic governors oppose the expansion, then we do not need to worry about ruling out large *positive* effects for Democratic partisanship, since the likelihood component of the model effectively rules out these implausible effects (e.g., large positive effects of Democratic partisanship on the probability of opposing the expansion are unlikely to generate data in which no Democrat opposes the expansion). We do, on the other hand, need to worry about ruling out the large *negative* effects, since the likelihood cannot effectively rule out the implausibly large negative effects (e.g., the larger the negative effect, the more likely the observed separation will occur). See Theorem 1 for a more formal treatment of this intuition.

Steps for choosing a reasonable prior for β_s.

1. Estimate the model coefficient using maximum likelihood, giving the coefficient vector $\hat{\beta}^{mle}$. Include the separating variable s_i in the model. Of course, this leads to implausible estimates for β_s, but our goal here is to choose reasonable values at which to fix the other parameters.

2. Choose a prior distribution $p(\beta_s)$ for the separating variable s.

3. Choose a number of simulations n_{sims} to perform (e.g., $n_{sims} \geq 1,000$) and for i in 1 to n_{sims} (e.g., $n_{sims} \geq 1,000$), do the following:

    a. Simulate $\tilde{\beta}_s^{[i]} \sim p(\beta_s)$.

    b. Replace $\hat{\beta}_s^{mle}$ in $\hat{\beta}^{mle}$ with $\tilde{\beta}_s^{[i]}$, yielding the vector $\tilde{\beta}^{[i]}$.

    c. Calculate the store the quantity of interest $q^{[i]} = q\left(\tilde{\beta}^{[i]}\right)$. This quantity of interest might be a first-difference or risk-ratio, for example.

4. Summarize the simulations $q^{[i]}$ using quantiles, histograms, or density plots. If the prior is inadequate, then update the prior distribution $p(\beta_s)$.

Application: Nuclear Proliferation and War

I recent debate emerged in the conflict literature around the effect of nuclear weapons on the probability of war in a dyad. Rauchhaus (2009) hypothesizes that "[t]he probability of major war between two states will decrease if both states possess nuclear weapons (p. 262)." Summarizing his empirical results, Rauchhaus writes:

> The hypotheses on nuclear symmetry find strong empirical support. The probability of a major war between two states is found to decrease when both states possess nuclear weapons (p. 269).

Despite using the same data, Bell and Miller (2014) claim that "nonnuclear dyads are in fact no more likely to fight wars than nonnuclear dyads (p. 9)." The disagreement hinges, in part, on whether and how to handle separation.[8] Rauchhaus (2009) ignores the separation and estimate that nonnuclear dyads are about 2.7 million times more likely to go to war than symmetric nuclear dyads. Bell and Miller (2014), on the other hand, use Jeffreys' (**?**) invariant prior, as suggested by Zorn (2005), and estimate that nonnuclear dyads are only about 1.6 times more likely to engage in war.

[8] Bell and Miller (2014) also disagree with a coding decision of Rauchhaus (2009), but that portion of their argument is less relevant to my purpose. Instead, my goal is to illustrate how one might choose a reasonable prior distribution and highlight the importance of the choice of prior.

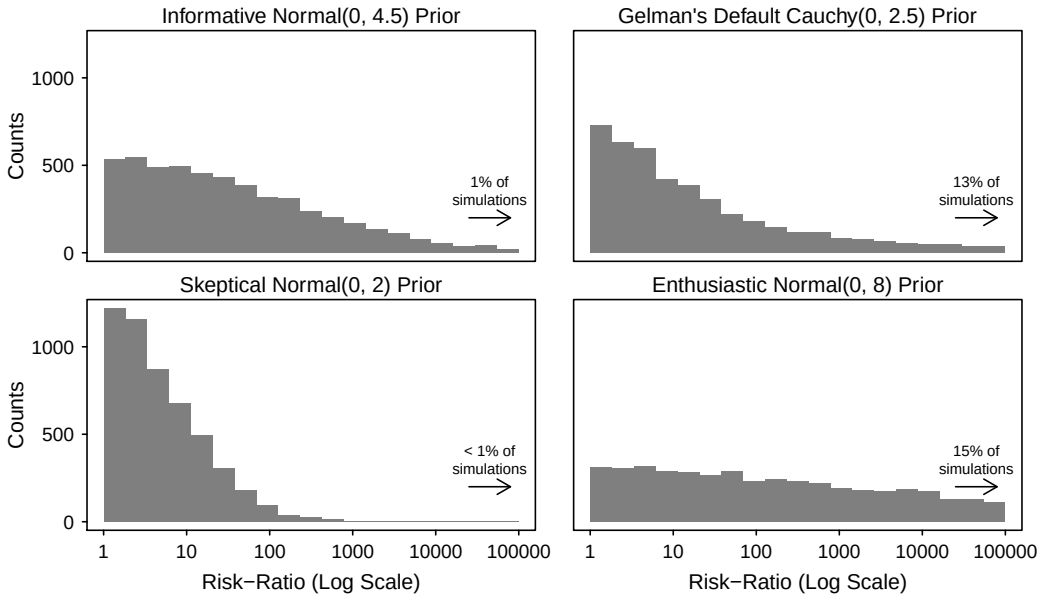

FIGURE 5: This figure shows the partial prior predictive distribution for the risk-ratio of war in nonnuclear to nuclear dyads. The risk-ratio is tells us how many times more likely war is in non-nuclear dyads compared to nuclear dyads. Notice that the informative prior treats effects smaller than 100 as plausible, but essentially rules out effects larger than 10,000. Gelman's suggested default places more weight closer to zero and more weight above 10,000. The skeptical prior essentially rules out effects larger than 100, while the enthusiastic prior treats effects between 1 and 100,000 as essentially equally likely.

	10%	20%	30%	40%	50%	60%	70%	80%	90%
Informative Normal(0, 4.5) Prior	1.80	3.20	5.90	10.90	21.90	44.20	107.20	305.10	1,493.50
Gelman's Default Cauchy(0, 2.5) Prior	1.50	2.30	3.70	6.30	13.10	30.30	120.80	1,586.10	1,846,124.50
Skeptical Normal(0, 2) Prior	1.30	1.60	2.10	2.80	3.70	5.20	8.00	12.70	26.90
Enthusiastic Normal(0, 8) Prior	2.70	7.10	20.60	62.70	218.10	833.60	4,323.20	27,202.40	560,885.80

TABLE 2: This table provides the deciles prior predictive distribution for the risk-ratio of war in nonnuclear and nuclear dyads. The risk-ratio is tells us how many times more likely war is in non-nuclear dyads compared to nuclear dyads. Notice that that the, informative prior suggests a median risk-ratio of about 20, which is a large, but plausible effect. Gelman's proposed default prior suggests a slightly smaller median ratio of about 13, but also allows very large effects. The skeptical prior suggests a median ratio of about 4 and the enthusiastic prior suggests a median ratio of over 200.

References

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

Barrilleaux, Charles, and Carlisle Rainey. 2014. "The Politics of Need: Examining Governors' Decisions to Oppose the 'Obamacare' Medicaid Expansion." Forthcoming in *State Politics and Policy Quarterly*. Manuscript at http://www.carlislerainey.com/files/need.pdf.

Bell, Mark S., and Nicholas L. Miller. 2014. "Questioning the Effect of Nuclear Weapons on Conflict." *Journal of Conflict Resolution* .

Box, George E. P. 1980. "Sampling and Bayes' Inference in Scientific Modelling and Robustness." *Journal of the Royal Statistical Society A* 143(4):383–430.

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):1360–1383.

Gross, Justin H. 2014. "Testing What Matters (If You Must Test at All): A Context-Driven Approach to Substantive and Statistical Significance." Forthcoming in *American Journal of Political Science*.

Heinze, Georg, and Michael Schemper. 2002. "A Solution to the Problem of Separation in Logistic Regression." *Statistics in Medicine* 21(16):2409–2419.

Jeffreys, H. 1946. "An Invariant Form of the Prior Probability in Estimation Problems." *Proceedings of the Royal Society of London, Series A* 186(1007):453–461.

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):341–355.

Lesaffre, E., and A. Albert. 1989. "Partial Separation in Logistic Discrimination." *Journal of the Royal Statistical Society. Series B (Methodological)*. 51(1):109–116.

McCaskey, Kelly, and Carlisle Rainey. 2014. "Meaningful Inferences: The Importance of Explicit Statistical Arguments for Substantive Significance." Working paper. Latest version at https://github.com/carlislerainey/meaningful-inferences.

Rainey, Carlisle. 2014. "Aruging for a Negligible Effect." Forthcoming in *American Journal of Political Science*.

Rauchhaus, Robert. 2009. "Evaluating the Nuclear Peace Hypothesis: A Quantitative Approach." *Journal of Conflict Resolution* 53(2):258–278.

Zorn, Christopher. 2005. "A Solution to Separation in Binary Response Models." *Political Analysis* 13(2):157–170.

Online Appendix

Proof of Theorem 1

Assumption 1 (Separation). *Suppose quasicomplete separation such that s_i perfectly predicts $y_i = 1$.*

Assumption 2 (Prior Shape). *Suppose that the researcher computes the posterior distribution $p(\beta|y) = p(y|\beta)p(\beta)$ such that for a particular $\beta^* \geq 0$ the prior distribution is decreasing at a decreasing rate.*

Intuitively, this assumption of a β^* allows the result to generalize to a range of common distributions. In particular, if the prior distribution is in the form of a double-exponential, which lacks "shoulders," then $\beta^* = 0$. However, the most common prior distributions used in applied work, such as the normal, t, and Jeffreys', have "shoulders" such that $\beta^* > 0$. In this case, the exact curvature of the distribution in the region $[0, \beta^*]$ affects the relative impact of the prior.

Assumption 3 (Scale Parameter). *Suppose finally the that informativeness of the prior distribution depends on scale parameter σ "flattens" the prior $p(\beta) = p(\beta|\sigma)$, such that as σ increases, the rate a which the prior descends to zero decreases.*

σ is chosen by the researcher based on prior information about the likely values of the coefficients.

Before proving Theorem 1, it is helpful to show several initial results.

Lemma 1. $\dfrac{\partial p(y|\beta)}{\partial \beta} > 0$ *for all* β.

Proof of Lemma 1. The quantity $p(y|\beta)$ is the probability of observing y (i.e., an outcome variable separated by s). Increasing values of β make this separation increasingly likely. Thus, $p(y|\beta)$ is increasing in β so that $\dfrac{\partial p(y|\beta)}{\partial \beta} > 0$. □

Lemma 2. $p(\beta|\sigma) > 0$ *for all* β.

Proof of Lemma 2. The quantity $p(\beta|\sigma)$ is a probability distribution defined to have support over the real line and thus $p(\beta|\sigma) > 0$ for all β. □

Lemma 3. $p(y|\beta) > 0$ *for all* β.

Proof of Lemma 3. The quantity $p(y|\beta)$ is a probability and thus bounded between zero and one. As long as data lie within the support of the probability model, this quantity lies strictly above zero. Since the theorem defines the data as such, $p(y|\beta) > 0$. □

Lemma 4. $\frac{\partial^2 p(\beta|\sigma)}{\partial\beta\partial\sigma}$ *for* $\beta > \beta^*$.

Proof of Lemma 4. By assumption, the prior density is decreasing at a decreasing rate in β for all $\beta > \beta^*$. Also by assumption, the scale parameter σ controls the rate at which β decreases such that increasing σ leads to a slower rate of decrease. These two assumptions together imply that $\frac{\partial^2 p(\beta|\sigma)}{\partial\beta\partial\sigma}$ for $\beta > \beta^*$. □

Now recall Theorem 1:

Theorem 1. *The impact of the researchers choice of* σ *on the posterior distribution* $p(\delta|y)$ *is increasing in* δ *for* $\delta > \delta^*$.

Proof of Theorem 1. To show that the effect of σ is increasing in β, I simply need to show that $\frac{\partial^2 p(\beta|y)}{\partial\beta\partial\sigma} > 0$ for $\beta > \beta^*$.

Recall that the posterior $p(\beta|y)$ is proportional to the likelihood $p(y|\beta)$ times the prior $\beta(\beta|\sigma)$, so that $p(\beta|y) \propto p(y|\beta)p(\beta|\sigma)$. First, we can use the product rule to obtain the derivative of $p(\beta|y)$ so that

$$\frac{\partial p(\beta|y)}{\partial\beta} \propto \frac{\partial p(y|\beta)}{\partial\beta}p(\beta|\sigma) + p(y|\beta)\frac{\partial p(\beta|\sigma)}{\partial\beta\partial}.$$

Only the final term involves σ, so we can easily obtain the desired derivative

$$\frac{\partial^2 p(\beta|y)}{\partial\beta\partial\sigma} \propto \overbrace{\frac{\partial p(y|\beta)}{\partial\beta}}^{\text{Lemma 1: +}} \overbrace{p(\beta|\sigma)}^{\text{Lemma 2: +}} + \overbrace{p(y|\beta)}^{\text{Lemma 3: +}} \overbrace{\frac{\partial^2 p(\beta|\sigma)}{\partial\beta\partial\sigma}}^{\text{Lemma 4: +}}. \tag{1}$$

Each term on the right-hand side of Equation 1 is positive for $\beta > \beta^*$ (Lemmas 1-4), so that $\frac{\partial^2 p(\beta|y)}{\partial\beta\partial\sigma} > 0$ for $\beta > \beta^*$.

□