Dealing with Separation in Logistic Regression Models*

Carlisle Rainey†

Working paper: Comments welcome!

July 16, 2015

ABSTRACT

When facing small numbers of observations or "rare events," political scientists often encounter separation, in which important explanatory variables 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 into the model. The most sophisticated research uses Jeffreys' invariant prior to stabilize the estimates. While Jeffreys' prior has the advantage of being automatic, I show that it often provides too much prior information, producing point estimates that are too small and confidence intervals that are too narrow. To help researchers choose a more reasonable prior distribution, I introduce the concept of a partial prior predictive distribution and develop the tools required to compute the partial prior predictive distribution, estimate the subsequent model, and summarize the results.

*I thank Mark Bell and Nicholas Miller for making their data available. I thank Mark Bell, David Firth, and Nicholas Miller for providing helpful comments. The analyses presented here were conducted with R 3.1.0. All data and computer code necessary for replication are available at github.com/carlislerainey/priors-for-separation. The (working) R package `separation` implements the procedures discussed in the paper and is available at github.com/carlislerainey/separation.

†Carlisle Rainey is Assistant Professor of Political Science, Texas A&M University, 2010 Allen Building, College Station, TX, 77843 (crainey@tamu.edu).

Separation, in which an explanatory variable perfectly predicts some binary observations, remains a problem in political science research (e.g., Barrilleaux and Rainey 2014, Bell and Miller 2014, Leeman and Mares 2014, Reiter 2014, and Weisiger 2014). Zorn (2005) offers the most principled solution to the problem of separation by suggesting that researchers maximize a penalized version of the usual likelihood function. Zorn's approach has the advantage of being automatic and easy for researchers to use.

However, when implementing Zorn's approach, substantive researchers face two major problems. First, the usual asymptotic confidence intervals and p-values do not work well. While a good method exists for finding confidence intervals and p-values for the model coefficients, this method does not extend to the typical quantities of interest, such as first differences. Researchers must still rely on the (poor) asymptotic approximation to simulate these quantities (King, Tomz, and Wittenberg 2000). Secondly, and perhaps most importantly, the penalty suggested by Zorn is designed for bias reduction in logistic regression models and not for handling separation (Firth 1993). Whether the suggested penalty approximates actual prior information in particular substantive settings remains an open and problem-specific question. To address these two problems, I suggest that researchers use a range of priors, focusing on an informative prior, and use MCMC to simulate directly from the posterior.

In this paper, I introduce conceptual and computational tools that help researchers understand the information provided by a given prior distribution and use that prior distribution to obtain meaningful point estimates and confidence intervals. I make three specific contributions. First, I use statistical theory and two applied examples to demonstrate the importance of choosing a prior distribution that represents actual prior information (and conducting robustness checks using a variety of prior distributions). Second, I introduce the concept of a partial prior predictive distribution, a powerful tool in understanding and choosing a prior when facing separation. Third, I introduce new software that make it easy for researchers to choose an informative prior distribution, simulate directly from the posterior distribution, and summarize

the inferences.[1]

I begin with a basic overview of the logistic regression model and summary of the impact of separation on the maximum likelihood estimates. I then describe two default prior distributions that researchers might use to handle separation. Next, I use a theoretical result and an applied example to demonstrate the importance of choosing an informative prior. I then introduce researchers to the concept of a partial prior predictive distribution, which enables researchers to understand complex prior distributions in terms of the key quantities of interest. To illustrate how these ideas work in practice, I conclude with a replication of Rauchhaus (2009) and Bell and Miller (2014), whose disagreement about the effect of nuclear weapons on war hinges, in part, on how to deal with separation.

The Logistic Regression Model

Political scientists commonly use logistic regression to model the probability of events such as war (e.g., Fearon 1994), policy adoption (e.g., Berry and Berry 1990), turning out to vote (e.g., Wolfinger and Rosenstone 1980), and government formation (e.g., Martin and Stevenson 2001). In the typical situation, the researcher uses an $n \times (k+1)$ design matrix X consisting of an intercept and k explanatory variables to model a vector of n binary outcomes y, where $y_i \in \{0,1\}$, using the model $\Pr(y_i) = \Pr(y_i = 1 \mid X_i) = \dfrac{1}{1+e^{-X_i\beta}}$, where β is a parameter vector of length $k+1$. Using this model, it is straightforward to calculate the likelihood function

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

Researchers routinely obtain the maximum likelihood estimate $\hat{\beta}^{mle}$ of the model parameters β by finding the vector β that maximizes L (i.e., maximizing the likelihood of the observed

[1] A working version of the R package `separation`, which implements the ideas in this paper, is available at github.com/carlislerainey/separation.

data). 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 perfectly predicts zeros, ones, or both.[2] *Complete separation* occurs when the "problematic" explanatory variable s_i (for separating explanatory variable) perfectly predicts both zeros *and* ones. *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. When the data overlap, the usual maximum likelihood estimates exist and provide reasonable estimates of parameters. However, under complete or quasicomplete separation, finite maximum likelihood estimates do not exist and the usual method of calculating standard errors fails (Albert and Anderson 1984; Zorn 2005).

Complete separation occurs when s_i perfectly predicts *both* zeros *and* ones. For example, suppose a dichotomous explanatory variable s_i , such that $y_i = 0$ for $s_i = 0$ and $y_i = 1$ for $s_i = 1$. To maximize the likelihood of the observed data, the "S"-shaped logistic regression curve must assign $\Pr(y_i) = \frac{1}{1+e^{-X_i\beta}} = 0$ when $s_i = 0$ and $\Pr(y_i) = \frac{1}{1+e^{-X_i\beta}} = 1$ when $s_i = 1$. Since the logistic regression curve lies strictly between zero and one, this likelihood cannot be achieved, only approached asymptotically as the coefficient β_s for s_i approaches infinity. Thus, the likelihood function under complete separation is monotonic, which implies that a finite maximum likelihood estimate does not exist.[3]

Quasicomplete separation occurs when s_i perfectly predicts *either* zeros *or* ones. For example,

[2] Separation can also occur when a *combination* of explanatory variables perfectly predicts zeros, ones, or both, see Lesaffre and Albert (1989).

[3] Although coefficient estimates under separation are infinite *in theory*, the hill-climbing algorithms approximate the infinite estimates with large, finite values *in practice*. These approximations increase with the precision of the algorithm. See Zorn (2005) for an illustration using software in Stata and R.

suppose that, when $s_i = 0$, sometimes $y_i = 1$ and other times $y_i = 0$, but $y_i = 1$ *always* when $s_i = 1$. To maximize the likelihood of the observed data, the "S"-shaped logistic regression curve must assign $\Pr(y_i) = \frac{1}{1+e^{-X_i\beta}} \in (0,1)$ when $s_i = 0$ and $\Pr(y_i) = \frac{1}{1+e^{-X_i\beta}} = 1$ when $s_i = 1$. Again, since the logistic regression curve lies strictly between zero and one, this likelihood cannot be achieved, only approached asymptotically. Thus, the likelihood function under quasicomplete separation also monotonically increases as the coefficient for s_i increases, which again implies that the maximum likelihood estimate does not exist.

For example, Barrilleaux and Rainey (2014) find that no Democratic governors opposed the Medicaid expansion under the Affordable Care Act, leading to a maximum likelihood estimate of positive infinity for the coefficient for the indicator of Republican governors. Similarly, Rauchhaus (2009) (see Bell and Miller 2014) finds no instances of states with nuclear weapons engaging in war with each other, leading to an estimated coefficient of negative infinity for the coefficient for the variable indicating nuclear dyads. To maximize the likelihood in these situations, the model must assign zero probability of opposition to states with Democratic governors and zero probability of war to nuclear dyads. 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.

For simplicity, this paper focuses the more common situation of quasicomplete separation. However, the ideas apply equally well to the less common situation of complete separation. For convenience, I say that the "direction of the separation" is positive if and only if $s_i = 1 \implies y_i = 1$ $\underline{\text{or}}$ $s_i = 0 \implies y_i = 0$ and that the direction of separation is negative if and only if $s_i = 0 \implies y_i = 1$ $\underline{\text{or}}$ $s_i = 1 \implies y_i = 0$. Thus, $\hat{\beta}^{mle} = +\infty$ when the direction of the estimate is positive, and $\hat{\beta}^{mle} = -\infty$ when the direction of the estimate is negative.

Solutions to Separation

The maximum likelihood (ML) framework 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 *sometimes* generate separation. Thus, under separation, the ML can only produce infinite estimates.

Before addressing potential solutions to this problem, 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 the estimates. Secondly, researchers sometimes "solve" the problem of separation by dropping the separating variable from the model. Zorn (2005, pp. 161-162) 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.

One principled solution is to build prior information $p(\beta)$ (the same prior information that leads researchers to deem infinite coefficients "implausibly large") into the model using Bayes' rule, so that

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

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 median. The current literature on dealing with separation suggests researcher take an automatic approach by using a default prior distribution, such as Jeffreys' invariant prior distribution (Jeffreys 1946, Zorn 2005) or a heavy-tailed Cauchy(0, 2.5) prior distribution (Gelman et al. 2008).

Jeffreys' Invariant Prior

Zorn (2005) suggests that political scientists deal with separation by maximizing a penalized likelihood rather than the likelihood (see Heinze and Schemper 2002 as well). Zorn suggests replacing the usual likelihood function $L(\beta|y)$ with a "penalized" likelihood function $L^*(\beta|y)$, so that $L^*(\beta|y) = L(\beta|y)|I(\beta)|^{\frac{1}{2}}$. It turns out that the penalty $|I(\beta)|^{\frac{1}{2}}$ is equivalent to Jeffreys' (1946) prior for the logistic regression model (Firth 1993, Poirier 1994). The posterior distribution can be obtained by applying Jeffreys' rule (Jeffreys 1946), 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 yields the posterior distribution $p(\beta|y) \propto L(\beta|y)|I(\beta)|^{\frac{1}{2}}$, so that Firth's penalized likelihood is equivalent to a Bayesian approach with Jeffreys' prior. The researcher can then sample from this posterior distribution using MCMC to obtain the features of interest, such as the mean and standard deviation.

However, Firth (1993) did not propose this 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 uncertainty of the estimates for particular research problems.

The Cauchy(0, 2.5) 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 $p(\beta) = |I(\beta)|^{\frac{1}{2}}$ lacks an interpretable scale and depends on the data in complex ways. Instead, they suggest standardizing continuous inputs to have mean zero and standard deviation one-half and simply centering binary inputs (Gelman 2008). Then, they suggest placing a weakly informative Cauchy(0, 2.5) prior on the coefficients for these rescaled variables that, like Jeffreys' prior, bounds the estimates away from positive and negative infinity but can also be interpreted as actual prior information.[4] Gelman et al. (2008, p. 1363) write:

> Our key idea is that actual effects tend to fall within a limited range. For logistic regression, a change of 5 moves a probability from 0.01 to 0.5, or from 0.5 to 0.99. We rarely encounter situations where a shift in input x corresponds to the probability of outcome y changing from 0.01 to 0.99, hence, we are willing to assign a prior distribution that assigns low probabilities to changes of 10 on the logistic scale.

As before, the posterior distribution is not easily available analytically, but one can easily use MCMC to simulate from the posterior distribution. Once a researcher has the MCMC simulations, she can obtain the point estimates and credible intervals for the parameters by summarizing the simulations.

Gelman et al. (2008) design their prior distribution to be reflective of prior informative for a range of situations. In many cases, their weakly informative prior supplies too little prior information. In a few cases, it might supply too much. In any either case, it remains an open question as to whether this general prior provides a reasonable estimate of the uncertainty of the estimates for *particular* research problems.

[4] Gelman et al. (2008) use a Cauchy(0, 2.5) prior for the coefficients but a Cauchy(0, 10) prior for the *intercept*. This allows the intercept to take on a *much* larger range of values (e.g., from 10^{-9} to $1 - 10^{-9}$)

The Importance of the Prior

While default priors, such as Zorn's suggested Jeffreys' prior or Gelman et al.'s suggested Cauchy(0, 2.5) prior are often useful as starting points, choosing an informative prior distribution is crucial for dealing with separation in a substantively meaningful manner. Further, whether a particular prior is reasonable depends on the particular application.

In most data analyses, the data swamp the contribution of the prior, so that the choice of prior has little effect on the posterior. However, in the case of separation, the prior essentially determines the shape of the posterior in the direction of the separation. The likelihood has an "S"-shape that approaches a limit 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 an important choice that affects the inferences. We can see the importance in both theory and practice.

The Impact of the Prior in Theory

Although it is intuitive that the prior drives the inferences in the direction of the separation, it is also easy to characterize the impact of the prior on a monotonic increasing likelihood in a general way. Suppose quasicomplete separation, such that whenever an explanatory variable $s_i = 1$, a binary outcome $y_i = 1$, but when $s_i = 0$, y_i might equal zero or one. Suppose further that the analyst wishes to obtain plausible estimates of coefficients for the model

$$Pr(y_i = 1) = \text{logit}^{-1}(\beta_{cons} + \beta_s s_i + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}).$$

It is easy to find plausible estimates for the coefficients of $x1, x2, ..., x_k$ using maximum likelihood, but finding a plausible estimate of β_s proves more difficult because maximum likelihood

suggests an estimate of $+\infty$. In order to obtain a plausible estimate of β_s, 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 β_s given the data. Using Bayes' Rule, the posterior distribution of $\beta = \langle \beta_{cons}, \beta_s, \beta_1, \beta_2, ..., \beta_k \rangle$ depends on the likelihood and the prior, so that $p(\beta|y) \propto p(y|\beta)p(\beta)$. 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 $p(\beta_s) = p(\beta_s|\sigma)$, though the results below simply depend on having any proper prior distribution. The informativeness of the prior distribution depends on σ, which is chosen by the researcher and "flattens" the prior $p(\beta_s) = p(\beta_s|\sigma)$, such that as σ increases, the rate at which the prior descends to zero decreases.

Theorem 1. *For a monotonic likelihood $p(y|\beta)$ increasing [decreasing] in β_s, proper prior distribution $p(\beta|\sigma)$, and large positive [negative] β_s, the posterior distribution of β_s is proportional to the prior distribution for β_s, so that $p(\beta_s|y) \propto p(\beta_s|\sigma)$.*

Proof: See the Technical Appendix.

Theorem 1 simply implies that for large values of β_s the posterior distribution depends almost entirely on the researcher's *choice* of prior distribution. Thus, the choice of prior matters. While the choice of prior might not affect the conclusion about the *direction* of the effect, it has a large impact on the conclusion about the *magnitude* of the effect. Credible intervals are crucial when discussing effect magnitudes (see Rainey 2014, Gross 2014, and McCaskey and Rainey 2014), and the choice of prior essentially drives the width 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 results from Barrilleaux and Rainey (2014), who are interested in the effect of partisanship on gover-

nors' decisions to oppose the Medicaid expansion in their states under the Patient Protection and Affordable Care Act (ACA).[5] As the authors note, no Democratic governors opposed the expansion, which leads to a problem of separation. To see whether the choice of prior matters, I use MCMC to simulate from the posterior using Zorn's (2005) and Gelman et al.'s (2008) suggested *default* prior distributions.

Figure 1 shows the posterior medians and 90% credible interval for the two default priors.[6] 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). For example, the Cauchy(0, 2.5) prior leads to a posterior median that is over 40% larger than the posterior median from Jeffreys' prior (4.9 compared to 3.5). The posterior mean is more than 80% larger using the Cauchy(0, 2.5) prior (7.1 compared to 3.9). Further, the 90% credible interval is more than twice as wide for the Cauchy(0, 2.5) prior. The choice between two *default* priors leads to a large change in inferences.

[5] 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.

[6] The credible intervals I use throughout this paper are 90% HDP credible intervals. Obviously, we could define many intervals that have 90% chance of containing the true parameter. However, the HPD interval is theoretically appealing because it is the *shortest* of these intervals. See Gill (2008, esp. pp. 48-51) and Casella and Berger (2002, esp. p. 448). The alternative, equal-tailed intervals, tends to exacerbate the differences between the priors.

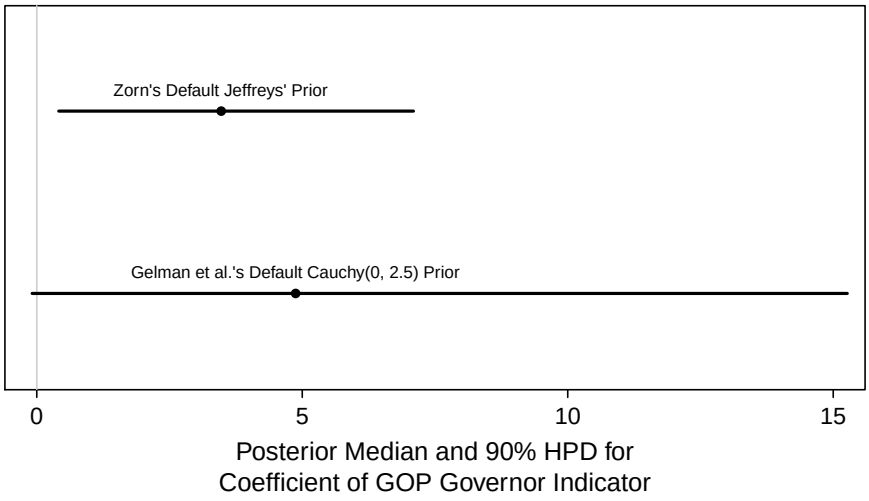

FIGURE 1: This figure provides the posterior medians and 90% credible intervals for the coefficient of the indicator for GOP governors in the model offered by Barrilleaux and Rainey (2014). Notice that Jeffreys' prior, suggested by Zorn (2005), is the more informative of these priors, suggesting that a coefficient larger than about seven is quite unlikely. On the other hand, the credible interval using the Cauchy(0, 2.5) prior, as suggested by Gelman et al. (2008), is about *twice* as wide as the credible interval from Jeffreys' prior, suggesting that effects as large as about 15 are plausible. Further, the posterior median from the Cauchy(0, 2.5) prior is about 40% larger than the posterior median from Jeffreys' prior.

Figure 2 shows the posterior distribution for the coefficient for the indicator of Republican governors. Notice that these two *default* 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, as suggested by Theorem 1. The more informative Jeffreys' prior leads to a more peaked posterior distribution that rules out coefficients larger than about seven. The less informative Cauchy(0, 2.5) prior leads to the conclusion that much larger coefficients, such as 15, are plausible. These differences are not trivial–there are large differences in the posterior distributions, and these differences can affect the conclusions that the researchers draw about the likely magnitude of the effect.

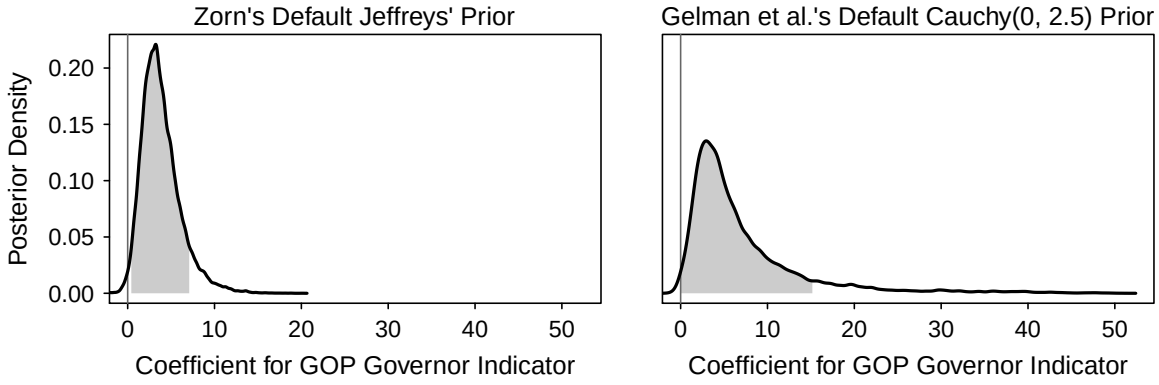

FIGURE 2: This figure shows the posterior distribution for the coefficient of the indicator for GOP governors in the model offered by Barrilleaux and Rainey (2014) for different default prior distributions. The grey shading indicates the 90% credible interval. Notice that the location and the spread of the posterior depend on the prior chosen, especially the right-hand side of the distribution, as suggested by Theorem 1.

Choosing an Informative Prior

While 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, in the replication of Barrilleaux and Rainey (2014) above, I show that the overall posterior distribution, the width of the 90% credible interval, and the posterior median largely depend on the prior one chooses. This implies that researchers relying on default priors alone risk under- or over-representing their confidence in the magnitude of the effect.

When facing separation, I suggest researchers use a prior distribution that satisfies three properties:

1. *Pools toward zero.* While the ultimate goal is to choose a prior distribution based on actual prior information, the prior distribution should also be appropriately conservative. As mentioned before, the prior distribution largely drives the inferences in the direction of the separation. 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 toward zero (Gelman and Jakulin 2007).

2. *Allows plausible effects.* The prior distribution should assign realistic prior probabilities to estimates that are *a priori* plausible according to the researcher's own prior beliefs.

3. *Rules out implausible effects.* The prior distribution should assign essentially no prior probability to estimates that are *a priori* implausible according to the researcher's own prior beliefs.

Different researchers will inevitably have different prior beliefs. For example there is substantial disagreement among international relations theorists about the likely effects of nuclear weapons on conflict. Some optimists believes that nuclear weapons make peace much more likely. Mearsheimer (1993, p. 57) argues that "nuclear weapons are a powerful force for peace" and observes:

> In the pre-nuclear world of industrialized great powers, there were two world wars between 1900 and 1945 in which some 50 million Europeans died. In the nuclear age, the story is very different. Only some 15,000 Europeans were killed in minor wars between 1945 and 1990, and there was a stable peace between the superpowers that became increasingly robust over time. A principal cause of this "long peace" was nuclear weapons.

Bueno de Mesquita and Riker (1982, p. 283) even theorize that the probability of conflict "decreases to zero when all nations are nuclearly armed." On the other hand, some pessimists (e.g., Sagan 1994) believe that nuclear weapons do not deter conflict, only make it more catastrophic. Mueller (1988, pp. 68-69) writes:

> Nuclear weapons may well have enhanced this stability–they are certainly dramatic reminders of how horrible a big war could be. But it seems highly unlikely that, in

> their absence, the leaders of the major powers would be so unimaginative as to need such reminding. Wars are not begun out of casual caprice or idle fancy, but because one country or another decides that it can profit from (not simply win) the war–the combination of risk, gain, and cost appears to be preferable to peace. Even allow considerably for stupidity, ineptness, miscalculation, and self-deception in these considerations, it does not appear that a large war, nuclear or otherwise, has been remotely in the interest of essentially-contended, risk-averse, escalation-anticipating powers that have dominated world affairs since 1945.

The optimists and the pessimists have different prior beliefs about the likely effects nuclear weapons. These different beliefs must lead to different interpretations of the evidence because the prior distribution has such a strong impact on the posterior distribution in the direction of the separation. However, researchers must clearly communicate the dependence of the inferences on the choice of prior by transparently developing an informative prior distribution and providing the inferences for alternative prior beliefs.

However, choosing a prior distribution is quite difficult, especially for multidimensional problems. Gill and Walker (2005) provide an overview of methods of choosing a prior appropriate to social science research. However, the most sensible approach for choosing a prior distribution depends on the nature of the statistical model and the prior information.

In general, the researcher might assess the reasonableness of the prior distribution by examining the prior predictive distribution and asking herself whether the prior and model produce 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 hypothetical 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 a researcher can use simulation to interpret the coefficient estimates of nonlinear models, she can use simulation to interpret the prior distribution.

Definition 1 (Prior Predictive Distribution)**.** *The prior predictive distribution, denoted as* $p(y_{new})$*, is the prior distribution of hypothetical data, so that* $p(y_{new}) = \int_{-\infty}^{\infty} p(y_{new}|\beta)p(\beta)d(\beta)$.

However, it is difficult to work with more than one dimension of the prior distribution. Specifying the full prior distribution requires simultaneously choosing prior distributions for the $k+1$ explanatory variables, as well as the relationships among these variables (e.g., family, location, scale, and correlations of all the parameters). This process is intractably tedious, as the researcher must evaluate the prior for each combination of each parameter set at a range of values. Even if the researcher considers only independent normal priors and ten values for each parameter, then the researcher must examine 10^{k+1} prior predictive distributions. If the researcher has eight control variables, so that $k = 8$, (i.e., Barrilleaux and Rainey 2014), then the researcher must evaluate one *billion* prior predictive distributions.

But only specific regions of the $k+1$ dimensional prior distribution are practically important when addressing separation. this allows the researcher to dramatically simplify the choice of prior. In particular, the researcher can simplify the focusing in two specific ways.

1. *Focus only on the separated coefficient.* Since the data swamp the prior for all the model coefficients except β_s, the only relevant "slices" of the prior distribution are those in which all other coefficients are near their maximum likelihood estimates.

2. *Focus in the direction of the separation.* The likelihood also swamps the prior in the direction opposite the separation. Unless the researcher has an extremely small data set (i.e., smaller than Barrilleaux and Rainey (2014), who have $N = 50$), the the likelihood essentially rules out values less [greater] than zero when the direction of separation is positive [negative].

I refer to this simplified focus as the *partial* prior predictive distribution.

Definition 2 (Partial Prior Predictive Distribution)**.** *The partial prior predictive distribution, denoted as* $p^*(y_{new})$*, is the prior distribution of* y_{new} *given that the separated coefficient lies in the direction of the separation and all other coefficients equal their maximum likelihood estimates, so that* $p^*(y_{new}) = \int_0^\infty p(y_{new}|\beta_s, \hat{\beta}_{-s}^{mle})p(\beta_s|\beta_s \geq 0)d(\beta_s)$ *when* $\hat{\beta}_s^{mle} = +\infty$ *and* $p^*(y_{new}) = \int_{-\infty}^0 p(y_{new}|\beta_s, \hat{\beta}_{-s}^{mle})p(\beta_s|\beta_s \leq 0)d(\beta_s)$ *when* $\hat{\beta}_s^{mle} = -\infty$.

For example, in Barrilleaux and Rainey (2014), we do not need to use prior information to obtain reasonable estimates for our measures of need and public opinion. Further, because no Democratic governors opposed the Medicaid expansion, then we do not need the prior to rule out large *positive* effects for Democratic partisanship. In both cases, the likelihood is sufficiently informative.

However, we do need to use the prior to rule out large *negative* effects for Democratic partisanship, because the likelihood cannot effectively rule out implausibly large negative effects. Indeed, the likelihood is monotonically decreasing in the coefficient for the indicator of Democratic governors. That is, the likelihood increases as the coefficient for Democratic partisanship becomes more negative. The larger the negative effect, the more likely separation would occur. The usual maximum likelihood, therefore, provides implausibly large negative estimates and unreasonable standard errors. Theorem 1 provides a more formal treatment of this intuition, but prior information is essential to obtain reasonable estimates and measures of uncertainty.

Choosing a prior, though, requires thoughtful effort. As I show above, default priors can lead to much different conclusions, so it is essential to build actual prior information into the model. In order to choose a reasonable, informative prior distribution, researchers need to obtain the partial prior predictive distribution defined in Definition 2. The following steps describe how researchers can simulate from the partial prior predictive distribution and use the simulations to check the reasonableness of the choice.

1. Estimate the model coefficients using maximum likelihood, given 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 the purpose is to choose reasonable values at which to fix the *other* parameters in order to focus on a single slice of the full prior.

2. Choose a prior distribution $p(\beta_s)$ for the separating variable s that is centered at zero. The most common choice is the scaled t distribution, which has the normal and Cauchy families as special cases ($\text{df} = \infty$ and $\text{df} = 1$, respectively).

3. Choose a large number of simulations n_{sims} to perform (e.g., $n_{sims} \geq 10,000$) and for i in 1 to n_{sims}, 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 and store the quantity of interest $\tilde{q}^{[i]} = q\left(\tilde{\beta}^{[i]}\right)$. This quantity of interest might be a first-difference or risk-ratio, for example.

4. Keep only those simulations in the direction of the separation (e.g., $\tilde{\beta}_s^{[i]} \geq 0$ when $\hat{\beta}_s^{mle} = +\infty$ and $\tilde{\beta}_s^{[i]} \leq 0$ when $\hat{\beta}_s^{mle} = -\infty$).

5. Summarize the simulations $\tilde{q}$ using quantiles, histograms, or density plots. If the prior is inadequate, then update the prior distribution $p(\beta_s)$.

Given that the inference can be highly dependent on the choice of prior, I recommend that the researcher choose at least three prior distributions: (1) an *informative* prior distribution that represents her actual beliefs, (2) a highly *skeptical* prior distribution that suggests the effect is likely small, and (3) a highly *enthusiastic* prior that represents the suggests the effect might be very large. Combined with Zorn's (2005) and Gelman et al.'s (2008) suggested defaults, these provide a range of prior distributions that researchers can use to evaluate their inferences.

Estimating the Full Model

Once the researcher obtains a reasonable prior distribution as well as several to use for robustness checks, she must use MCMC to obtain simulations from the posterior. Zorn (2005) and Gelman et al. (2008) suggest variations on maximum likelihood to quickly obtain estimates and confidence intervals. However, the normal approximation typically used to simulate the parameters and calculate quantities of interest (King, Tomz, and Wittenberg 2000) is particularly inaccurate under separation. As an alternative, I recommend the researcher use MCMC to simulate directly from the posterior distribution. The researcher can then use these simulations to calculate point estimates and confidence intervals for any desired quantity of interest. For the informative $p_{\text{inf}}(\beta_s)$, skeptical $p_{\text{skep}}(\beta_s)$, and enthusiastic $p_{\text{opt}}(\beta_s)$ priors , I suggest the model:

$$Pr(y_i = 1) = \text{logit}^{-1}(\beta_{cons} + \beta_s s_i + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik})$$

$$\beta_s \sim p_k(\beta_s), \text{ for } k \in \{\text{inf, skep, enth}\},$$

with improper, constant priors on the other model coefficients.

Application: Nuclear Proliferation and War

A recent debate emerged in the conflict literature between Rauchhaus (2009) and Bell and Miller (2014) that revolves around the issue of separation. Rauchhaus (2009, p. 262) hypothesizes that "[t]he probability of major war between two states will decrease if both states possess nuclear weapons." 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, p. 9) claim that "nonnuclear dyads are in fact no more likely to fight wars than nonnuclear dyads." Their disagreement hinges, in part, on whether and how to handle separation, because no nuclear dyad in Rauchhaus data engages in war.[7] Rauchhaus (2009) ignores the separation and estimates 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' (1946) invariant prior, as suggested by Zorn (2005), and estimate that nonnuclear dyads are only about 1.6 times more likely to engage in war. Because these authors use very different prior distributions, they reach very different conclusions. This raises important questions. First, would a reasonable, informative prior distribution support Rauchhaus' position of a meaningful effect or Bell and Miller's position of essentially no effect? Second, how robust is the conclusion to a range of more and less informative prior distributions?

Prior

The first step in dealing with the separation in a principled manner is to choose a prior distribution that represents actual prior information. To choose a reasonable prior, I follow the process above to generate a partial prior predictive distribution for the risk-ratio that Bell and Miller (2014) emphasize. I experimented with a range of prior distributions, from a variety of families, but settled on the normal family. After some experimentation, I selected a normal distribution with mean zero and standard deviation 4.5 to serve as an informative prior and represent my own prior beliefs. I chose this prior distribution because it essentially rules out risk-ratios larger than 1,000–effects that I find implausibly large–and treats risk-ratios smaller than 1,000 as plausible. Figure 3 and Table 1 summarize the partial prior predictive distributions for the this

[7] Bell and Miller (2014) also disagree with Rauchhaus's (2009) coding of the 1999 conflict in Kargil between India and Pakistan, which both possessed nuclear weapons. This conflict is excluded from Rauchhaus's data set, but Bell and Miller argue that it should be included as a war between two nuclear-armed states. However, 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.

normal distribution with standard deviation 4.5.

To evaluate the robustness of any statistical claims to the choice of prior, I also selected a highly skeptical and highly enthusiastic prior. I chose a normal distribution with mean zero and standard deviation two to serve as a skeptical prior that represents the belief that any pacifying effect of nuclear weapons is small (e.g., Mueller 1988). This skeptical prior distribution essentially rules out risk-ratios larger than 25 as implausibly large. Finally, I selected a normal distribution with mean zero and standard deviation eight to serve as an enthusiastic prior that represents the belief that the pacifying effects of nuclear weapons might be quite large (e.g., Mearsheimer 1993). This enthusiastic prior, on the other hand, treats risk-ratios as large as 500,000 as plausible. Figure 3 shows the partial prior predictive distributions for the informative, skeptical, and enthusiastic prior distributions. For convenience, Table 1 provides the deciles of the PPPDs shown in Figure 3.

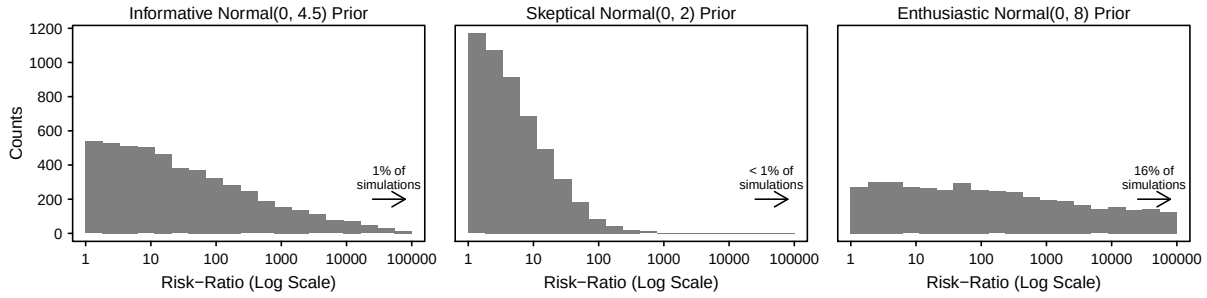

FIGURE 3: This figure shows the partial prior predictive distribution for the risk-ratio of war in nonnuclear to nuclear dyads. The risk-ratio is tells us how many times more likely war is in non-nuclear dyads compared to nuclear dyads. Notice that the informative prior treats effects smaller than 100 as plausible, but essentially rules out effects larger than 10,000. 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	1.7	3.1	5.6	10.3	19.5	44.1	98.8	296.9	1,577.7
Skeptical Normal(0, 2) Prior	1.3	1.7	2.2	2.9	3.9	5.4	8.3	13.2	26.7
Enthusiastic Normal(0, 8) Prior	2.9	8.1	24.7	76.5	256	987.6	5,300.9	42,466.7	643,954.6

TABLE 1: 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. The skeptical prior suggests a median ratio of about 4 and the enthusiastic prior suggests a median ratio of over 200.

Notice that the skeptical prior suggests that risk-ratios above and below 4 are equally likely (i.e., 50th percentile of the PPPD is 3.9), while the enthusiastic prior suggests that effects above and below 220 are equally likely. The informative prior, on the other hand, suggests (more reasonably, in my view) that the effect is equally likely to fall above and below 20. These three prior distributions, along with the defaults suggested by Zorn (2005) and Gelman et al. (2008), provide a range of distributions to represent a range of prior beliefs.

Posterior

Figure 4 shows the posterior distributions for the coefficient of the indictor of nuclear dyads from the informative, skeptical, enthusiastic, and two default prior distributions. The areas shaded grey indicate the 90% credible intervals. Notice that the location (e.g., peak or mode), shape, scale, and credible interval depend on the choice of prior. While the magnitude of this coefficient is not easily interpretable, notice that Gelman et al.'s (2008) suggested default prior is somewhat similar to the informative prior, but Zorn's (2005) suggested default is quite similar to the *skeptical* prior. Thus, these distributions illustrate that the prior is important when dealing with separation. Indeed, it is a critical step in obtaining reasonable inferences.

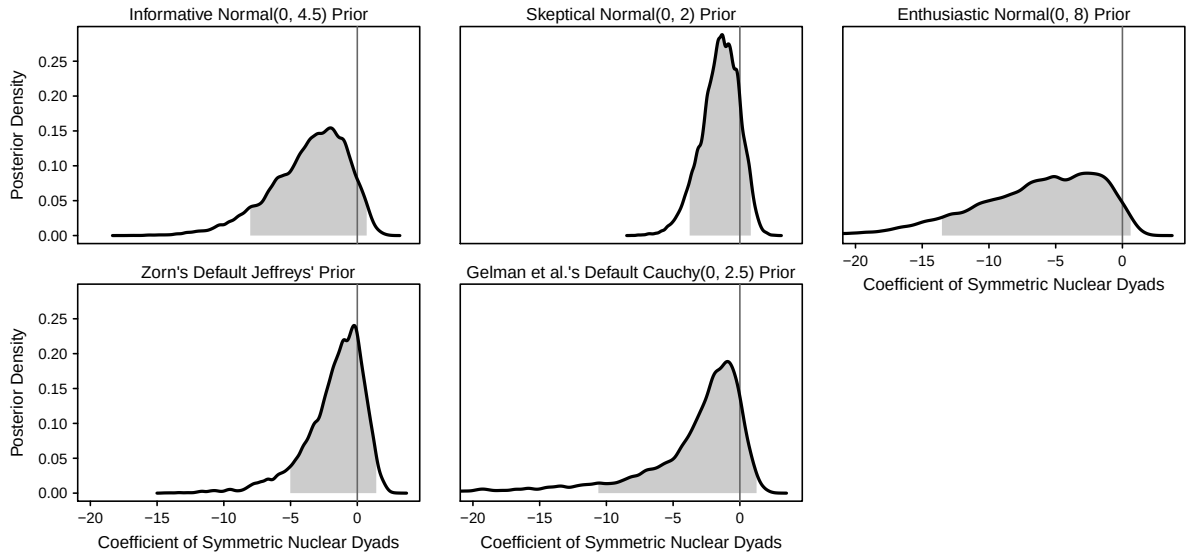

FIGURE 4: This figure shows the posterior distribution for the logit coefficient using each of the five prior distributions. The grey shading indicates the 90% credible interval . Notice that the choice of prior has a large effect on the inferences. For example, the enthusiastic prior suggests the ratio might be as large as -13, while the skeptical prior suggests the ratio might be as large as -4. Importantly, notice the similarity between the posterior from Zorn's (2005) suggested default and the skeptical prior, in terms of their peak (i.e., mode), shape, and highest posterior density.

I now turn to the posterior distribution of the risk-ratios–the key quantity of interest in the debate between Bell and Miller (2014) and Rauchhaus (2009). Figure 5 presents the posterior medians and the 90% credible intervals for each prior. While an initial glance at the figure shows substantial variation in the point estimates and the width of the intervals, notice that these are plotted on the log scale (otherwise the wider intervals dominate the plot). Notice that the informative prior suggests the true risk-ratio has about a 90% chance of falling between about 0.1 and about 2000, with a posterior median of about 25.[8] The skeptical prior, on the other hand, suggests the risk-ratio has about a 90% chance of falling between 0.1 and 30, with a posterior median of about 4. The enthusiastic prior suggests the risk-ratio has about a 90% chance of falling between 0.1 and 500,000. The inferences from these priors are *very* different.

[8]It is worth pointing out that these wide confidence interval suggests that, although the credible intervals overlap zero, these data do not warrant a conclusion of "no effect" (see Rainey 2014).

Further, the 90% credible interval using Zorn's (2005) default is *much* narrower than the informative prior, and the posterior median of Zorn's suggested default prior is even smaller than the *skeptical* prior. For this particular application, Gelman's suggested default more closely matches the informative prior, though notice that the point estimate is less than half and the upper-bound of the credible interval is ten times larger than than the point estimate and upper-bound of the credible interval from the informative prior.

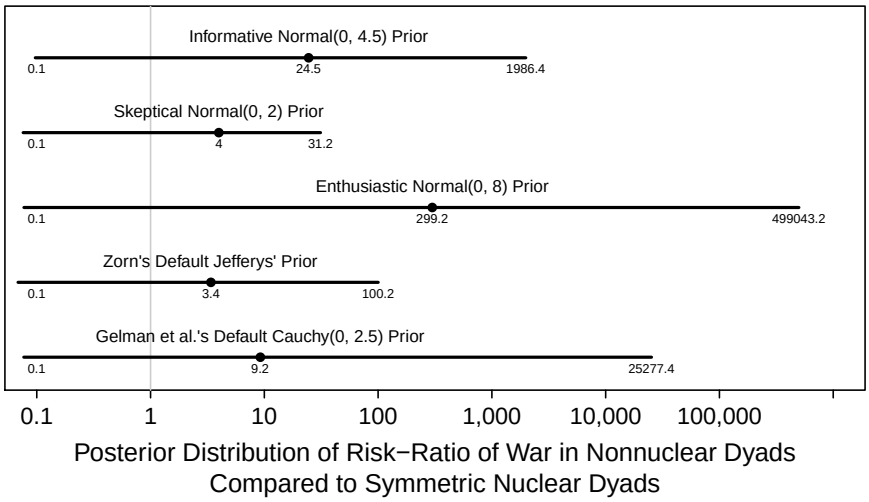

FIGURE 5: This figure shows the posterior median and 90% credible intervals for the risk-ratio using each of the five prior distributions *on the log scale*. Notice that the choice of prior has a huge effect on the inferences about the risk-ratio. For example, the skeptical prior suggests the ratio might be as large as 31, while the enthusiastic prior suggests the ratio might be as large as 500,000. Also, notice that the posterior median from Zorn's proposed default prior is *smaller* than the posterior median from the skeptical prior.

Finally, I use the posterior distributions from each prior to calculate the probability that the presence of nuclear weapons make war less likely (i.e., the probability that the risk-ratios shown in Figure 5 are greater than one). Recall Rauchhaus' hypothesis that nuclear weapons decrease the chance of war. These probabilities can be thought of as the probability that Rauchhaus' hypothesis is correct. Following the standard of $p \leq 0.05$ as offering strong evidence against the null hypothesis, it is reasonable to take $Pr(RR > 1) \geq 0.95$ as strong evidence for the

research hypothesis. Figure 6 shows the probability that the hypothesis is correct for each prior distribution. Notice that while only the enthusiastic prior falls above the 0.95 standard, the evidence for the claim is at least suggestive. Perhaps most importantly for my purposes, Zorn's suggested default leads to the *least* evidence in favor of Rauchhaus' hypothesis–even the skeptical prior provides more evidence for Rauchhaus' claim.

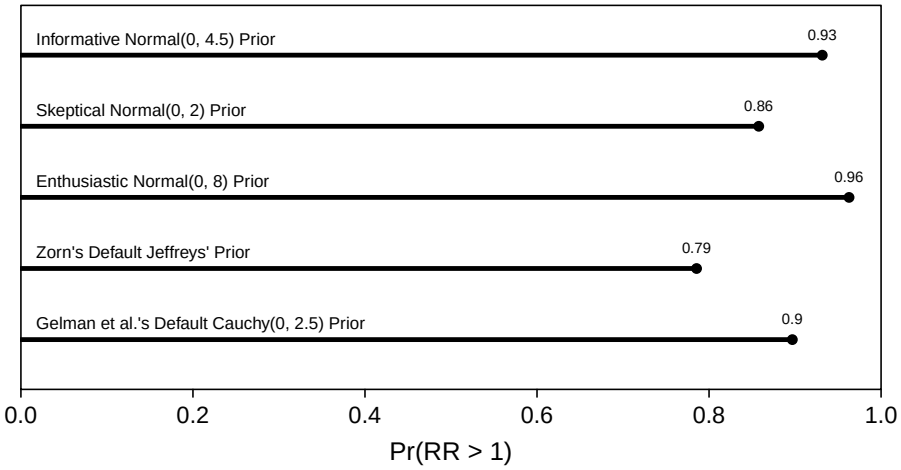

FIGURE 6: This figure shows the posterior probability of the hypothesis that nonnuclear dyads are *more* likely to engage in war than symmetric nuclear dyads for each of the five prior distribution. From a hypothesis testing perspective, the evidence for the hypothesis is borderline or suggestive for each prior. However, notice that the skeptical prior, perhaps held by a researcher who believes the pacifying effect of nuclear weapons is small or nil, yields *greater* evidence for the hypothesis than Jeffreys' invariant prior suggested as a default by Zorn (2005).

Conclusion

Separation is a relatively common situation in political science. It is also an unusual "problem" because the effects in the sample are "too big" for maximum likelihood. In this situation, dropping the separating variables (i.e., deliberate specification bias) or interpreting the implausible coefficients and standard errors are particularly unattractive options. But even the most principled solution to date, the incorporation of prior information via default priors (Zorn 2005,

Gelman et al. 2008), has shortcomings.

First, the normal approximations necessary to simulate from these default prior distributions perform poorly. While it is possible to use profile likelihood methods to obtain more accurate confidence intervals for the coefficients (Zorn 2005, Heinze and Schemper (2002), McCullagh and Nelder (1989)), it is difficult to translate these intervals into confidence intervals for quantities of interest. I provide the computational tools to simulate directly from the posterior using both Zorn's (2005) and Gelman et al.'s Zorn's (2008) suggested default priors.

Second, the applications examining the effect of nuclear weapons and the effect of governors' partisanship illustrate what Theorem 1 proves–under separation, the choice of prior affects substantive conclusions. Even the predominant default priors used to deal with separation can provide very different inferences. A carefully-chosen, informative prior is an essential step in the process of obtaining reasonable inferences under separation. But what does this mean for applied researchers? I suggest two implications:

1. When facing separation, the choice of prior matters. Researchers must carefully choose a prior that represents actual prior information. Otherwise, the point and interval estimates will be too small or too large.

2. In addition to carefully choosing an informative prior that represents her own beliefs, the researcher should show how the inferences change for a range of prior distributions. In the debate between Bell and Miller (2014) and Rauchhaus (2009), the choice of prior almost completely drives the inferences about the likely magnitude of the risk-ratio. Thus, to the extent that there is disagreement about the prior, there will be disagreement about the results. In particular, I suggest that researchers report the key quantities of interest for a skeptical prior, an enthusiastic prior, and the two default priors suggested by Zorn (2005) and Gelman et al. (2008).

When facing separation, researchers must *carefully* choose a prior distribution to rule out

implausibly large effects. This paper introduces the concept of a partial prior predictive distribution and the associated computational tools to help researchers choose a prior distribution that represents actual prior information for their particular research problem. By presenting results using several prior distributions, including an informative prior, researchers can increase the transparency, credibility, and accuracy of their inferences when dealing with separation.

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* .

Berry, Frances Stokes, and William D. Berry. 1990. "State Lottery Adoptions as Policy Innovations: An Event History Analysis." *American Political Science Review* 84(2):395–415.

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.

Bueno de Mesquita, Bruce, and William Riker. 1982. "An Assessment of the Merits of Selec- An Assessment of the Merits of Selective Nuclear Proliferation." *Journal of Conflict Resolution* 26(2):283–306.

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

Fearon, James D. 1994. "Signaling versus the Balance of Power and Interests: An Empirical Test of a Crisis Bargaining Model." *Journal of Conflict Resolution* 38(2):236–269.

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

Gelman, Andrew. 2008. "Scaling Regression Inputs by Dividing by Two Standard Deviations." *Statistics in Medicine* 27(15):2865–2873.

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.

Gelman, Andrew, and Aleks Jakulin. 2007. "Bayes: Radical, Liberal, or Conservative?" *Statistica Sinica* 17(2):422–426.

Gill, Jeff. 2008. *Bayesian Methods: A Social and Behavioral Science Approach*. 2nd ed. Boca Raton, FL: Chapman and Hall.

Gill, Jeff, and Lee D. Walker. 2005. "Elicited Priors for Bayesian Model Specifications in Political Science Research." *Journal of Politics* 67(3):841–872.

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.

Leeman, Lucas, and Isabella Mares. 2014. "The Adoption of Proportional Representation." *Journal of Politics* 76(2):461–478.

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

Martin, Lanny W., and Randolph T. Stevenson. 2001. "Government Formation in Parliamentary Democracies." *American Journal of Political Science* 45(1):33–50.

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.

McCullagh, Peter, and John A. Nelder. 1989. *Generalized Linear Models*. Second ed. Boca Raton, FL: Chapman and Hall.

Mearsheimer, John J. 1993. "The Case for a Ukrainian Nuclear Deterrent." *Foreign Affairs* 72(3):50–66.

Mueller, John. 1988. "The Essential Irrelevance of Nuclear Weapons: Stability in the Postwar World." *International Security* 13(2):55–79.

Poirier, Dale. 1994. "Jeffreys' Prior for Logit Models." *Journal of Econometrics* 63(2):327–339.

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.

Reiter, Dan. 2014. "Security Commitments and Nuclear Proliferation." *Foreign Policy Analysis* 10(1):61–80.

Sagan, Scott D. 1994. "The Perils of Proliferation." *International Security* 18(4):66–107.

Weisiger, Alex. 2014. "Victory Without Peace: Conquest, Insurgency, and War Termination." *Conflict Management and Peace Science* 31(4):357–382.

Wolfinger, Raymond E., and Steven J. Rosenstone. 1980. *Who Votes?* Yale University Press.

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

Technical Appendix

Proof of Theorem 1

Recall Theorem 1:

Theorem 1. *For a monotonic likelihood $p(y|\beta)$ increasing [decreasing] in β_s, proper prior distribution $p(\beta|\sigma)$, and large positive [negative] β_s, the posterior distribution of β_s is proportional to the prior distribution for β_s, so that $p(\beta_s|y) \propto p(\beta_s|\sigma)$.*

Proof. Due to separation, $p(y|\beta)$ is monotonic increasing in β_s to a limit $\underline{\mathscr{L}}$, so that $\lim_{\beta_s \to \infty} p(y|\beta_s) = \underline{\mathscr{L}}$. By Bayes' rule,

$$p(\beta|y) = \frac{p(y|\beta)p(\beta|\sigma)}{\int_{-\infty}^{\infty} p(y|\beta)p(\beta|\sigma)d\beta} = \frac{p(y|\beta)p(\beta|\sigma)}{\underbrace{p(y|\sigma)}_{\text{constant w.r.t. } \beta}}.$$

Integrating out the other parameters $\beta_{-s} = \langle \beta_{cons}, \beta_1, \beta_2, ..., \beta_k \rangle$ to obtain the posterior distribution of β_s,

$$p(\beta_s|y) = \frac{\int_{-\infty}^{\infty} p(y|\beta)p(\beta|\sigma)d\beta_{-s}}{p(y|\sigma)}, \tag{1}$$

and the prior distribution of β_s,

$$p(\beta_s|\sigma) = \int_{-\infty}^{\infty} p(\beta|\sigma)d\beta_{-s}.$$

Notice that $p(\beta_s|y) \propto p(\beta_s|\sigma)$ iff $\frac{p(\beta_s|y)}{p(\beta_s|\sigma)} = k$, where the constant $k \neq 0$.Thus, Theorem 1 implies that

$$\lim_{\beta_s \to \infty} \frac{p(\beta_s|y)}{p(\beta_s|\sigma)} = k$$

Substituting in Equation 1,

$$\lim_{\beta_s \to \infty} \frac{\frac{\int_{-\infty}^{\infty} p(y|\beta)p(\beta|\sigma)d\beta_{-s}}{p(y|\sigma)}}{p(\beta_s|\sigma)} = k.$$

Multiplying both sides by $p(y|\sigma)$, which is constant with respect to β,

$$\lim_{\beta_s \to \infty} \frac{\int_{-\infty}^{\infty} p(y|\beta)p(\beta|\sigma)d\beta_{-s}}{p(\beta_s|\sigma)} = kp(y|\sigma).$$

Setting $\int_{-\infty}^{\infty} p(y|\beta)p(\beta|\sigma)d\beta_{-s} = p(y|\beta_s)p(\beta_s|\sigma)$,

$$\lim_{\beta_s \to \infty} \frac{p(y|\beta_s)p(\beta_s|\sigma)}{p(\beta_s|\sigma)} = kp(y|\sigma).$$

Canceling $p(\beta_s|\sigma)$ in the numerator and denominator,

$$\lim_{\beta_s \to \infty} p(y|\beta_s) = kp(y|\sigma).$$

Recalling that $\lim_{\beta_s \to \infty} p(y|\beta) = \underline{\mathscr{L}}$ and substituting,

$$\underline{\mathscr{L}} = kp(y|\sigma),$$

which implies that $k = \frac{p(y|\sigma)}{\underline{\mathscr{L}}}$, which is a positive constant. □