Political Science Research and Methods (2022), **10**, 428–437
doi:10.1017/psrm.2020.46

RESEARCH NOTE

Separation and Rare Events*

Liam F. Beiser-McGrath[1,2,3]

[1]Department of Politics, International Relations, and Philosophy, Royal Holloway, University of London, Egham, United Kingdom, [2]Department of Politics and Public Administration, Universität Konstanz, Konstanz, Germany and [3]D-GESS, ETH Zürich, Zürich, Switzerland
Corresponding author. Email: liam@liambeisermcgrath.com; liam.beiser-mcgrath@rhul.ac.uk

(Received 18 December 2018; revised 2 March 2020; accepted 16 March 2020; first published online 11 December 2020)

Abstract
When separation is a problem in binary dependent variable models, many researchers use Firth's penalized maximum likelihood in order to obtain finite estimates (Firth, 1993; Zorn, 2005; Rainey, 2016). In this paper, I show that this approach can lead to inferences in the opposite direction of the separation when the number of observations are sufficiently large and both the dependent and independent variables are rare events. As large datasets with rare events are frequently used in political science, such as dyadic data measuring interstate relations, a lack of awareness of this problem may lead to inferential issues. Simulations and an empirical illustration show that the use of independent "weakly-informative" prior distributions centered at zero, for example, the Cauchy prior suggested by Gelman *et al.* (2008), can avoid this issue. More generally, the results caution researchers to be aware of how the choice of prior interacts with the structure of their data, when estimating models in the presence of separation.

Keywords: Bayesian; categorical data analysis; discrete choice models

1. Introduction

The existence of separation in binary choice models, where an independent variable perfectly predicts a binary dependent variable, is a problem within political science. The default response to this problem suggested by Zorn (2005) and further evaluated by Rainey (2016), is the use of penalized maximum likelihood (PMLE), equivalent to the use of a Jeffreys prior (Jeffreys, 1946), developed by Firth (1993). Such an approach has also been used to provide finite estimates of fixed effects for units that never experience the outcome event (Cook *et al.*, 2020).

In this paper, I demonstrate that PMLE can lead to statistically significant point estimates in the *opposite* direction to that of the separation, when the number of observations are sufficiently large and the binary dependent and independent variables responsible for separation are rare events. Dyadic data, with tens to hundreds of thousands of observations, on countries over time is a common example of such data. In these cases, researchers often focus on rare events both as dependent and independent variables, such as interstate war, the onset of sanctions, possession of nuclear weapons, and the signing of preferential trade agreements.

This reversal in sign is a result of the Jeffreys prior resulting in non-independent prior densities for parameters, which can lead to high prior density for parameters opposite the direction of separation when including a binary rare event independent variable. To demonstrate how this occurs and its effect on inferences, I use simulated data and an empirical illustration. As an alternative to Jeffreys prior/Firth's PMLE, independent "weakly-informative" priors such as the

*Thanks to Carlisle Rainey, Jonah Gabry, and Ben Goodrich for comments and suggestions.

© The Author(s), 2020. Published by Cambridge University Press on behalf of the European Political Science Association. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike licence (http://creativecommons.org/licenses/by-nc-sa/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the same Creative Commons licence is included and the original work is properly cited. The written permission of Cambridge University Press must be obtained for commercial re-use.

Konstanzer Online-Publikations-System (KOPS)
URL: http://nbn-resolving.de/urn:nbn:de:bsz:352-2-xoos3thk4kg10

Cauchy prior suggested by Gelman *et al.* (2008) ensure that the point estimate remains in the direction of separation for this specification.

Researchers should be mindful that Firth's method can lead to statistically significant point estimates in the opposite direction of separation. While there may be occasions researchers believe this is an appropriate result, it should be made clear that this is due to the specific choice of the prior. Researchers who wish to ensure point estimates are always in the same direction of the separation can instead use the Cauchy prior approach of Gelman *et al.* (2008). Even so, and echoing Rainey (2016), researchers should be aware that any default solution to separation necessarily has inbuilt assumptions about researcher's prior information that may not be universally applicable.

2. How penalized maximum likelihood and Jeffreys prior lead to opposite point estimates

Suppose our estimating equation is $y = \alpha + \beta x + \epsilon$ and we are faced with negative quasi-complete separation. That is, there exist no observations in the 2×2 table, displayed in Table 1a such that $x = 1 \wedge y = 1$.

In this single covariate case, Firth's Jeffreys prior approach is equivalent to adding 0.5 to each cell in Table 1a (Zorn, 2005), resulting in Table 1b.

This solution becomes problematic when it does not maintain the feature that there are relatively more observations where $y = 1$ when $x = 0$ than when $x = 1$. For inferences about the effect of x on y to have the same sign as the separation, it must hold that:

$$\frac{n_3 + 0.5}{n_1 + n_3 + 1} > \frac{0.5}{n_2 + 1} \tag{1}$$

Rare events in large datasets can violate this inequality. When y is a rare event, n_3 is small. If x is also a rare event, n_1 will be very large while n_2 remains small. This is important as if n_2 is small, then the addition of 0.5 to all cells will lead the relative frequency of $y = 1$ when $x = 1$ to be larger than when $x = 0$, in spite of the data suggesting the opposite. This is due to this relevant proportion being strongly affected due to n_2 being small, while the proportion for when $x = 0$ is mostly unaffected due to n_1 being so large.

As an illustration, consider the case of war between two nuclear powers. Table 2 displays the appropriate frequencies using data from Rauchhaus (2009). The relationship between these two variables shows quasi-complete separation, with no observations where both countries had nuclear weapons and went to war. The relevant relative frequencies in this case are $\mathrm{p}_{10} = \frac{62}{62+454,\,752} \approx 0.0001$ and $\mathrm{p}_{11} = \frac{0}{805} = 0$.

Suppose we were to use the Firth's PMLE estimator to examine this bivariate relationship. As noted previously, this is akin to adding 0.5 to each cell. This makes the new relative frequencies: $\mathrm{p}'_{10} = \frac{62.5}{62+454,\,752+1} \approx 0.0001$ and $\mathrm{p}'_{11} = \frac{0.5}{805+1} \approx 0.0006$. Thus this approach to dealing with separation results in the relative frequency of war in nuclear dyads being six times *larger* than in dyads where both states do not simultaneously have nuclear weapons. Jeffreys prior leads to inferences about the relationship between nuclear weapons and interstate war opposite to the direction seen in the data.

More concretely, while Jeffreys prior is non-informative with regard to the baseline probabilities, it is ultimately informative about the parameters. Jeffreys prior results in prior distributions for the constant term and parameter for x that are not independent. Jeffreys prior gives more (less) prior mass for large positive coefficients for x than equally large negative coefficients for x, when evaluated at a large negative (positive) value for the constant term, ensuring that the joint prior distribution remains uninformative with regard to the baseline probability.

This is problematic when y is a rare event. The likelihood is high at large negative values for the constant, the point where Jeffreys prior assigns more mass to large positive values for the coefficient of the variable that leads to separation. Therefore, the prior leads to the estimate for the coefficient on x to be positive, even though this is in the direction opposite to that of the separation.

Table 1. Negative quasi-complete separation and Jeffreys prior

(a) The observed data		
	$x = 0$	$x = 1$
$y = 0$	n_1	n_2
$y = 1$	n_3	0
(b) Using Jeffreys prior		
	$x = 0$	$x = 1$
$y = 0$	$n_1 + 0.5$	$n_2 + 0.5$
$y = 1$	$n_3 + 0.5$	0.5

Table 2. Nuclear dyads and war

	Not both nuclear	Both nuclear
No war	454,752	805
War	62	0

Panels (a) and (b) in Figure 1 illustrate this for the case where $n_1 = 50,000$, $n_2 = 100$, $n_3 = 100$. Panel (a) shows that Jeffreys prior gives more mass to positive values for the coefficient on x relative to equally negative values when the constant is negative. As the likelihood has high density when the constant has a large negative value, the Jeffreys prior pulls the posterior toward being positive in spite of negative separation.

Under the same specification, estimates in the direction of the prior can be retained if researchers instead use independent "weakly-informative" prior distributions centered at zero.[1] This retains the property that the posterior point estimate of the coefficient for the variable that leads to separation is never in a direction opposite to the separation.[2]

The Cauchy prior suggested by Gelman *et al.* (2008) is one such prior.[3] They advocate using independent Cauchy distributions as priors for parameters, with a location of zero and scale of 2.5 for independent variables and 10 for the constant. These default priors are weakly informative, based upon plausible baseline probabilities and effect sizes for covariates that are rescaled to have mean zero and standard deviation 0.5 (if continuous or a symmetric binary variable). For the independent variables, the prior corresponds to the idea that absolute changes of less than 5 in logit probability (e.g. moving from 0.5 to 0.99) are plausible when increasing a variable from one standard deviation below to above its mean. The scale is widened for the prior for the constant, which due to rescaling reflects the expected success probability when all variables are at their means, to correspond to the range of plausible success probabilities to be from 10^{-9} to $1 - 10^{-9}$.

Panels (c) and (d) in Figure 1 illustrate the use of independent Cauchy distributions. As can be seen from the prior, equally large positive and negative values for the coefficient on x have equal prior density independent of the value of the constant. As a result at the point where the likelihood has high density for the constant, the posterior remains in the direction of the separation as movements toward positive values for the coefficient on x lead to lower prior density. Thus the use of an independent prior density centered at zero ensures that inferences in terms of point estimates remain in the direction of the separation, unlike those of the Jeffreys prior/PMLE.

[1]If researchers instead use a specification where dummy variables for both values of x are included and the intercept is dropped, then the opposite inference can still occur with independent "weakly-informative" prior distributions centered at zero. In this case, independence is not sufficient, suggesting that researchers should keep in mind how their estimating equation is specified when choosing priors to deal with separation.

[2]The most extreme case would be as $n_1 \to \infty$, while n_2 and n_3 remain fixed and $n_4 = 0$. In this case, the likelihood increasingly provides little information. Thus the posterior is approximately the prior, which is centered at zero.

[3]While suggested by Gelman *et al.* (2008) as a default prior, other research suggests either assessing the sensitivity of inferences to prior choices (Rainey, 2016) or instead the use of a different prior distribution as a default prior (Ghosh *et al.*, 2015).

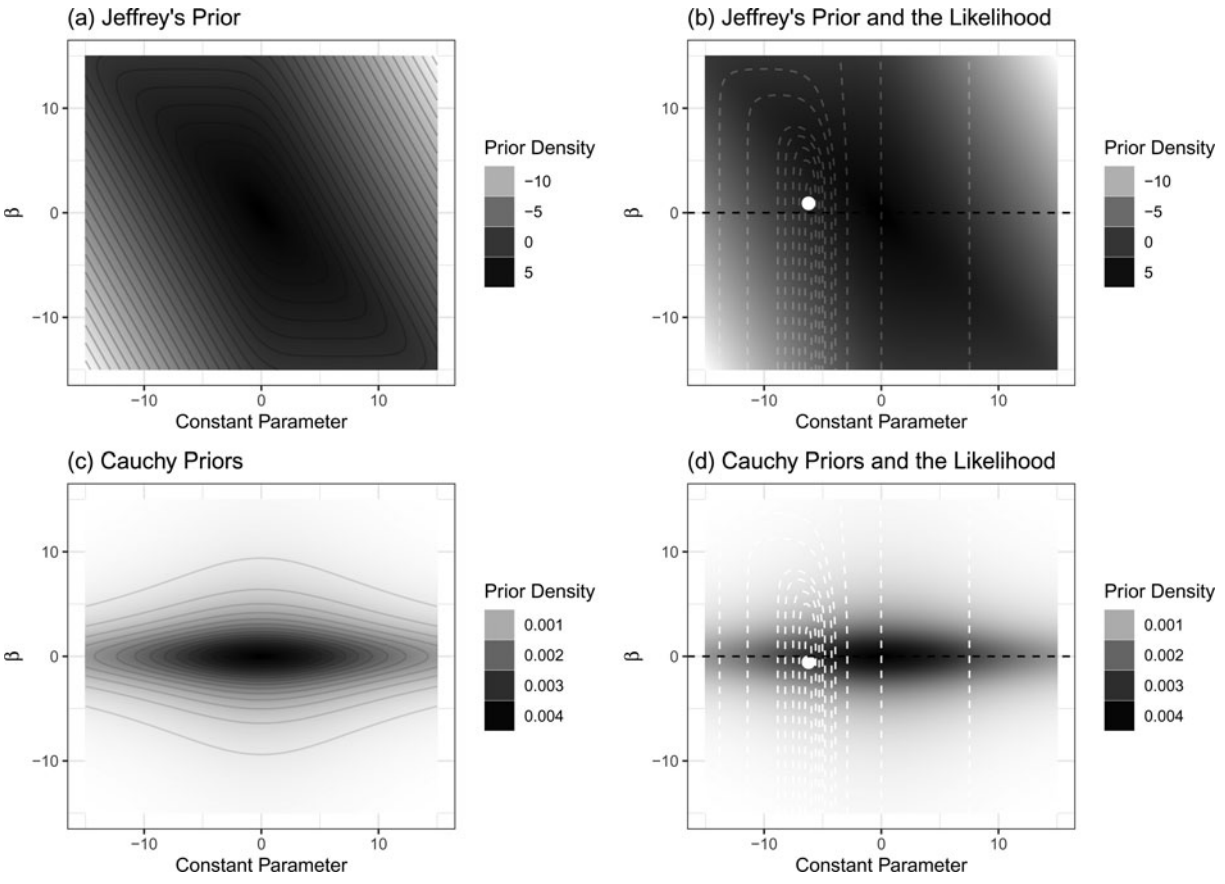

Figure 1. The use of priors for separation. (a) The joint prior density for Jeffreys prior overlaid with contours of the prior. (b) The prior density for Jeffreys prior overlaid with the contours of the likelihood. The point indicates the resulting estimate from the penalized maximum likelihood. (c) The joint prior density for the Cauchy priors suggested by Gelman *et al.* (2008), overlaid with contours of the prior. (d) The joint prior density for the Cauchy priors overlaid with the contours of the likelihood. The point indicates the resulting posterior mode.

To further illustrate this result, I compare the use of Jeffreys prior/Firth's PMLE estimator to the use of independent Cauchy prior distributions on hypothetical data.[4,5] I examine how the performance of these estimators changes with different values of n_1, n_2,, and n_3 while keeping $n_4 = 0$ to maintain negative quasi-complete separation.[6]

Figure 2 displays the results of estimating these models on the simulated data. We can see that when $y = 1 \wedge x = 0$ and $y = 0 \wedge x = 1$ are rare events relative to $y = 0 \wedge x = 0$, the PMLE estimator can lead to coefficients in the opposite (positive) direction to that of the separation. In contrast, using the Cauchy prior ensures that the coefficient on x remains in the same direction of the separation although this parameter asymptotically approaches zero.

Figure 3 displays how the features of the data and choice of prior impact uncertainty around the parameter estimate. I focus on two cases, where the number of observations with $x = 0 \wedge y = 0$ (n1) changes from 1000 to 100,000, while the number of observations with $x = 1 \wedge y = 0$ (n2) and $x = 0 \wedge y = 1$ (n3) remaining fixed at 50.

For the first case, while both y and x are still rare-events, estimates of the effect of x remain in the same direction as the separation due to the relatively low overall number of observations. The magnitude of the point estimates are similar across models. Examining the distribution of the estimates, we can see that both have a similar amount of mass in the direction of the separation.

Things change considerably in the second case. The point estimate obtained using Firth's PMLE/Jeffreys prior is now in the opposite direction of the separation. Furthermore, the majority of the estimate's distribution is also in the opposite direction of the separation, with approximately 92 percent of the mass of the profile likelihood being positive. Estimates obtained using the Cauchy prior do not exhibit this behavior. However, there are differences in the distribution of the posterior dependent upon the choice of estimation algorithm.[7]

In summary, a seemingly innocuous choice of estimator to deal with separation has an important consequence. Namely, coefficient estimates in the opposite direction of the separation can be obtained. While researchers prior beliefs may suggest this is an appropriate inference to make, for example that given enough time it is likely we would observe a case in the empty cell thus leading to this relationship, reading of the prior literature suggests this is not the case.

3. Empirical illustration: sanctions and interstate war

In this section, I illustrate the occurrence of this phenomenon using data on economic sanctions and interstate war. Economic sanctions are a tool used by states to facilitate and improve outcomes in bargaining situations (Morgan and Schwebach, 1997). Yet historical cases suggest that sanctions may lead to war. For example, the imposition of extensive sanctions by the United States upon Japan in 1941 is a common explanation for Japan initiated war with the United States in that same year.[8]

Understanding such interstate relations, through the use of dyadic data, often confronts rare events for both the dependent and independent variables. As displayed in Table 3, both the onset of sanctions and war between states are rare events. Furthermore, there is quasi-complete negative separation. There are no observations where a sanction onset and war occur.

[4]For estimating Firth's PMLE, I use the function logistf from the logistf package in R (Heinze *et al.*, 2013). To estimate models with Cauchy prior distributions, I use the function bayesglm from the arm package in R (Gelman and Su, 2015), which applies an Expectation-Maximization (EM) algorithm, and the function stan_glm from the rstanarm package in R (Stan Development Team, 2014) applying an optimization algorithm (L-BFGS). Due to computational constraints, estimation by MCMC is reserved for the empirical illustration.

[5]Additional results showing the same behavior for estimation in Stata are included in the appendix.

[6]The exact values used are: $n1 \in \{1000, 1389, 1931, 2683, 3728, 5179, 7197, 10, 000, 13, 895, 19, 307, 26, 827, 37, 276, 51, 795, 71, 969, 100, 000\}$, $n_2 \in \{50, 100, 200\}$ and $n_3 \in \{50, 100, 200\}$.

[7]This is a result of the algorithm used for bayesglm being an approximation to the posterior, with posterior uncertainty simulated by assuming multivariate normality.

[8]E.g. https://visitpearlharbor.org/american-sanctions-spur-pearl-harbor/

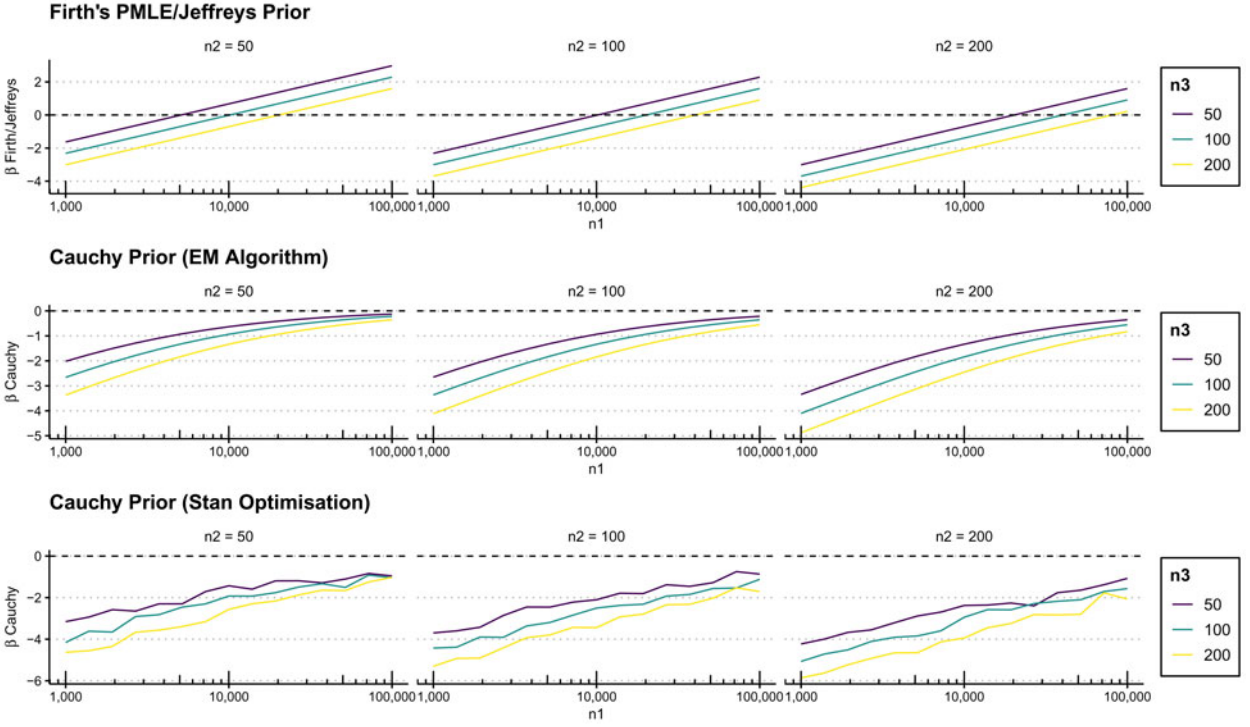

Figure 2. Estimated coefficients using Firth's penalized maximum likelihood and Gelman et al.'s Cauchy prior when x and y are rare events and there is negative quasi-complete separation. Lines display the estimated coefficient for x, β, for different scenarios. As all scenarios have negative quasi-complete separation, negative estimated coefficients are in the same direction as the separation, while positive coefficients are in the opposite direction of the separation. The x-axis, displayed on a log scale, denotes the number of observations where $x = 0 \wedge y = 0$ (n1), the columns indicate the number of observations where $x = 1 \wedge y = 0$ (n2), and lines denote the number of observations where $x = 0 \wedge y = 1$ (n3).

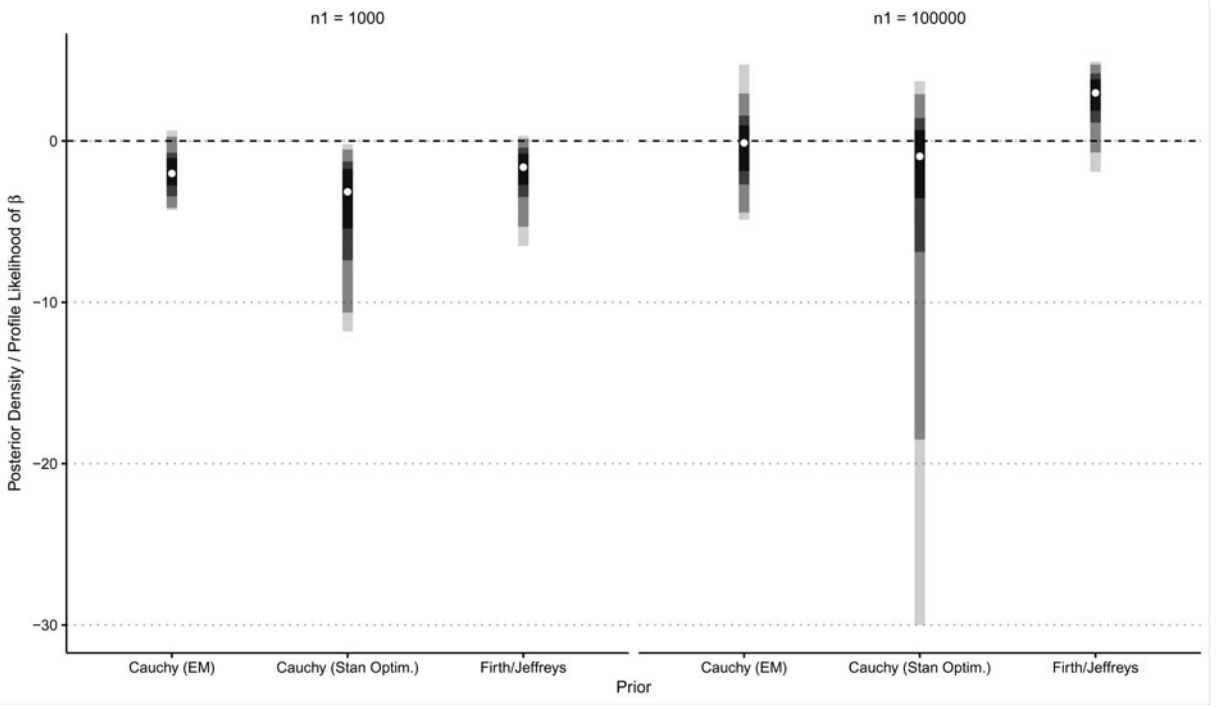

Figure 3. Estimated coefficients, with measures of uncertainty, for two scenarios using Firth's penalized maximum likelihood and Gelman et al.'s Cauchy prior when x and y are rare events and there is negative quasi-complete separation. The shaded areas indicate the central 50, 68, 90, and 95 percent areas of the posterior density (Cauchy prior) and profile likelihood (Firth's PMLE) for estimated effect of x (β). Each panel indicates how these distributions change when increasing the number of observations, where $x = 0 \wedge y = 0$ (n1), from 1000 to 100,000. The number of observations where $x = 1 \wedge y = 0$ (n2) and $x = 0 \wedge y = 1$ (n3) are held constant at 50.

Table 3. Sanction onsets and war

	No sanction onset	Sanction onset
No war	368,481	26
War	32	0

Table 4. How choice of prior affects inferences about the effect of sanction onsets upon interstate war

	Jeffreys/PMLE	Cauchy (EM)	Cauchy (Full Bayes)	Jeffreys/PMLE	Cauchy (EM)	Cauchy (Full Bayes)
Sanction	5.37**	− 0.01	− 0.1	4.07*	− 0.03	− 0.11
	[0.52, 7.37]	[− 4.06, 3.56]	[− 4.9, 4.32]	[− 2.69, 4.94]	[− 4.06, 3.56]	[− 4.98, 4.11]
Controls				✓	✓	✓
n	368,539	368,539	368,539	368,539	368,539	368,539

*** $p < 0.005$, ** $p < 0.05$, * $p < 0.1$
Based upon p-value, or whether zero falls into the associated credible interval. For models using Cauchy priors, all variables are centered. These are further divided by two times their standard deviation if continuous.

To illustrate how the choice of prior would affect inferences about the effect of sanctions, I combined dyadic data on economic sanctions collected by Hufbauer *et al.* (1990); Elliott *et al.* (2007) and collated by Hafner-Burton and Montgomery (2008) with dyadic data on interstate war collected by Bell and Miller (2015). After removing missing values, this results in a dataset with approximately 370,000 observations for 12,268 dyads from 1951 to 1999.

The results displayed in Table 4 show that using Firth's PMLE/Jeffreys prior leads to the coefficient for the onset of sanctions to be opposite to the direction of the separation. Furthermore, this effect is classified as statistically significant at conventional levels, for models both with and without controls (models 4 and 1, respectively). Figure 4 displays the appropriate measures of uncertainty for estimates of the effect of sanction onsets upon war. We can see that the majority of the profile likelihood's mass is in the opposite direction of the separation when using Firth's PMLE. Interpretation of the effect of sanction onsets, estimated using Firth's PMLE, would be that sanctions *increase* the probability of conflict, even though no such cases occur in the data. The use of the Cauchy prior avoids this problem and the posterior distribution resembles the Cauchy prior, reflecting that the data has little information to inform us about the effect of sanctions upon war.

Therefore, researchers should be wary of using Firth's PMLE as a default solution to separation in binary choice models. In cases where researchers are dealing with rare events and large amounts of data, such an approach can result in misleading inferences. Rather, researchers should be mindful of the (implicit) choice of prior when they are attempting to deal with separation and if there has to be a default then they should use independent "weakly-informative" prior distributions centered at zero, such as the Cauchy prior suggested by Gelman *et al.* (2008).

4. Conclusion

In this paper, I have shown that the commonly suggested PMLE approach to separation (Zorn, 2005) can lead to statistically significant point estimates opposite to the direction of separation. This occurs when confronted with separation in datasets with a large number of observations where the dependent and independent variables of interest are rare events.

Therefore, Firth's PMLE/Jeffreys prior is not necessarily a suitable default choice for dealing with separation when dealing with rare events. Independent "weakly-informative" prior distributions centered at zero, such as the Cauchy prior suggested by Gelman *et al.* (2008), are one such means to ensure that point estimates remain in the same direction of separation. Even so, and

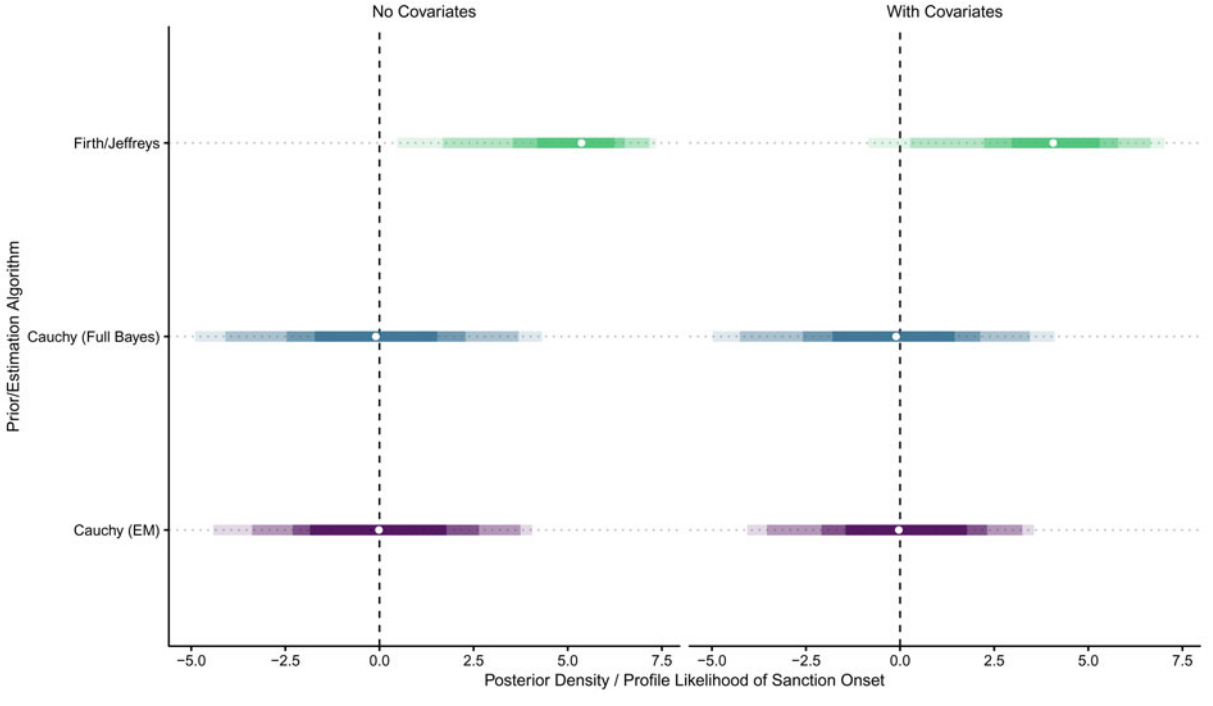

Figure 4. Estimated coefficients, with measures of uncertainty, for the onset of sanctions using Firth's penalized maximum likelihood and Gelman et al.'s Cauchy prior, estimated via E–M or with MCMC. The shaded areas indicate the central 50, 68, 90, and 95 percent areas of the posterior density (Cauchy prior) and profile likelihood (Firth's PMLE) for the estimates. The left panel displays the results for models with no covariates (models 1–3) and the right panel displays the results for models including covariates (models 4–6).

echoing Rainey (2016), researchers should be mindful of how prior choices shape uncertainty estimates in such cases.

More generally, researchers should be mindful of how the prior used to deal with separation contains specific information that has consequences for the possible direction of parameter estimates, that needs to be justified *a priori*. In particular, the use of data-dependent priors, such as Jeffreys prior, require researchers to be mindful about the particular features of their data and how this translates into the prior and thus influences the parameters of interest. In many circumstances, researchers would benefit from instead formulating priors based upon theory and previous evidence. For example, researchers may face cases where separation occurs due to it being impossible for the effect of a variable to be a certain sign. In such cases, researchers may be comfortable assigning zero prior, and thus posterior, mass for a given effect size direction or to functions of the posterior distribution.[9] At a minimum, however, researchers should be cognizant of the implications about prior information that default solutions to separation make, how these influence their estimates, and whether it is appropriate to the problem at hand.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/psrm.2020.46.

References

Bell MS and Miller NL (2015) Questioning the effect of nuclear weapons on conflict. *Journal of Conflict Resolution* **59**(1), 74–92. Available at http://jcr.sagepub.com/content/early/2013/08/19/0022002713499718.abstract.

Cook SJ, Hays JC and Franzese RJ (2020) Fixed effects in rare events data: a penalized maximum likelihood solution. *Political Science Research and Methods* **8**(1), 92–105.

Elliott KA, Schott JJ, Hufbauer GC and Oegg B (2007) *Economic Sanctions Reconsidered*, 3rd Edn. Washington, DC: Institute for International Economics.

Firth D (1993) Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38. Available at http://www.jstor.org/stable/2336755.

Gelman A and Y-S Su (2015) arm: data analysis using regression and multilevel/hierarchical models. R package version 1.8-5. Available at http://CRAN.R-project.org/package=arm.

Gelman A, Jakulin A, Pittau MG and Su Y-S (2008) A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics* **2**, 1360–1383.

Ghosh J, Li Y and Mitra R (2015) On the use of cauchy prior distributions for bayesian logistic regression. *ArXiv e-prints*.

Hafner-Burton EM and AH Montgomery (2008) Power or plenty—how do international trade institutions affect economic sanctions? *Journal of Conflict Resolution* **52**, 213–242. Workshop on New Science of International Organizations, Philadelphia, PA, August 30, 2006.

Heinze G, Ploner M, Dunkler D and Southworth H (2013) logistf: Firth's bias reduced logistic regression. R package version 1.21. Available at http://CRAN.R-project.org/package=logistf.

Hufbauer GC, Schott JJ, Elliott KA and Institute for International Economics (U.S.) (1990) *Economic Sanctions Reconsidered*. Washington, DC: Institute for International Economics.

Jeffreys GH (1946) An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* **186**, 453–461. Available at http://www.jstor.org/stable/97883.

Morgan TC and Schwebach VL (1997) Fools suffer gladly: the use of eco- nomic sanctions in international crises. *International Studies Quarterly* **41**, 27–50.

Rainey C (2016) Dealing with separation in logistic regression models. *Political Analysis* **24**, 339–355.

Rauchhaus R (2009) Evaluating the nuclear peace hypothesis: a quantitative approach. *Journal of Conflict Resolution* **53**, 258–277. Available at http://jcr.sagepub.com/content/53/2/258.abstract.

Stan Development Team (2014) RStan: the R interface to Stan, Version 2.5.0. Available at http://mc-stan.org/rstan.html.

Zorn C (2005) A solution to separation in binary response models. *Political Analysis* **13**, 157–170. Available at http://pan.oxfordjournals.org/content/13/2/157.abstract.

[9]For example, step 4 of the simulation of quantities of interest from a partial posterior distribution outlined by Rainey (2016, 348) states to not use parameter values that are in the direction opposite to the separation.

Cite this article: Beiser-McGrath LF (2022). Separation and Rare Events. *Political Science Research and Methods* **10**, 428–437. https://doi.org/10.1017/psrm.2020.46