Detecting and Correcting for Separation in Strategic Choice Models

Casey Crisman-Cox* Olga Gasparyan† Curtis S. Signorino‡

July 29, 2022

Abstract

Separation or "perfect prediction" is a common problem in discrete choice models that, in practice, leads to inflated point estimates and standard errors. Standard statistical packages do not provide clear advice on how to correct these problems. Furthermore, separation can go completely undiagnosed in fitting advanced models that optimize a user-supplied log-likelihood rather than relying on pre-programmed estimation procedures. In this paper, we both describe the problems that separation can cause and address the issue of detecting it in empirical models of strategic interaction. We then consider several solutions based on penalized maximum likelihood estimation. Using Monte Carlo experiments and a replication study we demonstrate that when separation is detected in the data, the penalized methods we consider are superior to ordinary maximum likelihood estimators.

WORD COUNT: 7,232

*Texas A&M University. Email: c.crisman-cox@tamu.edu

†Hertie School. Email: o.gasparyan@hertie-school.org

‡University of Rochester. Email: curt.signorino@rochester.edu

1 Introduction

Separation is a common problem in modeling categorical dependent variables wherein a linear combination of one or more explanatory variables perfectly predicts values of the outcome variable. It presents theoretical and practical problems. Theoretically, under data generating processes (DGPs) and sample sizes where separation is plausible, the statistical properties of an estimator are poorly defined (e.g., unidentified point estimates with infinite expected values). Practically, in datasets where separation appears, the magnitudes of numerically calculated point estimates and standard errors tend to inflate, sometimes heavily, toward positive or negative infinity. In binary-outcome models, solutions to the separation problem have been proposed and examined by Beiser-McGrath (2020), Gelman et al. (2008), Zorn (2005), and others. This line of inquiry has been invaluable for applied researchers. However, a binary outcome is only one type of categorical choice model used by political scientists; separation problems also plague more advanced or complicated models.[1]

Specifically, no one has approached the separation problem within the context of discrete-choice strategic models (e.g., Signorino 1999). By considering this issue, we make three specific contributions. First, we derive bias-reduced (BR) strategic estimators based on penalized likelihood (PL) estimation and demonstrate these estimators using Monte Carlo simulations and a replication of Signorino and Tarar (2006). Second, we introduce political scientists to a tool for diagnosing separation from Konis (2007) and demonstrate how it applies to strategic models. Third, we provide software for researchers to easily fit the BR strategic estimators.

Throughout, we focus on separation problems in a two-player, extensive-form deterrence game, which is a standard workhorse model for political scientists interested in the empirical implications of theoretical models (EITM). This model and extensions to it are used to study key questions across political science. In many cases, scholars derive an empirical model from

[1]For example, Cook et al. (2018) discuss separation problems in multinomial logit models.

a formal theory and then supply a self-coded, log-likelihood function to a numeric optimizer to find maximum likelihood estimates (MLE). This approach is extremely useful for fitting advanced models to data. However, separation becomes more difficult to diagnose in these settings, as optimization software will issue successful convergence codes without raising any warnings about the numerical instability caused by separation. Additionally, because these models often endogenize one or more choices, separation-induced inflation in one estimate can corrupt other estimates.

Before proceeding, it is worth pointing out that while BR estimators are the primary tool for addressing separation, they were initially proposed to combat small sample bias in binary choice models (Firth 1993; Rainey and McCaskey 2021).[2] As such, we expect that these approaches may also reduce bias in the coefficient estimates even when separation is not necessarily a concern. Indeed, the entire enterprise of fitting strategic models may be improved by considering the wider application of BR estimators, especially given that these models sometimes involve large numbers of interrelated parameters with moderately sized samples. However, bias reduction is not costless; as Rahman and Sultana (2017) point out, bias reduction in point estimates does not always translate into bias reduction in predicted probabilities and in some cases PL estimation can *increase* this bias. Future work should analyze the trade-off between bias reduction in the estimates and possible bias increases in the choice probabilities in finite samples without separation. However, given the relatively complicated nature of strategic modeling, it seems likely that BR estimators have more to offer this family of models than just a solution to separation problems.

2 Separation problems

Separation occurs in discrete choice models when a linear combination of one or more independent variables perfectly predicts a category of the outcome variable (Albert and Anderson 1984). There are two main reasons why separation occurs: (1) at least one of the

[2]This origin is why these penalized estimators are called BR estimators.

parameters is infinite; or (2) the true parameters are finite, but perfect prediction occurs as an artifact of a particular DGP and realized sample. We focus only on the latter case. Here, separation can be thought of as a finite-sample problem: if enough additional data is collected, the problem disappears.

In cases like these, where the true parameters are finite, separation creates theoretical and practical problems. To understand these problems, consider a sample where a single predictor perfectly predicts a category of the outcome variable. In such a situation, the sample log-likelihood function is monotonic in the estimate of that predictor's parameter (i.e., better fit can always be found by moving the estimate further from zero). As Albert and Anderson (1984) show, because of the monotonicity, there is no unique MLE that solves the first order conditions. Instead, the log-likelihood converges to an asymptote as the estimate goes to $\pm\infty$, depending on the true parameter's sign.

Regarding the estimator's theoretical finite-sample properties, recall that bias is defined based on the expected value of the MLE (i.e., the average MLE over possible samples), and consider a DGP where separation is plausible in any given realized sample. In these situations, the expected value of the MLE includes samples where the estimate is $\pm\infty$. Therefore, the estimator's moments are undefined.

Concerning practical problems in estimation, separation leads to numerically computed estimates and standard errors that are much larger than the truth.[3] Because of the monotone log-likelihood, the numerically obtained MLE will tend to be (i) much larger in magnitude than the true parameter and (ii) a function of the optimization software's numeric tolerance (Zorn 2005). To put this another way, while the true MLE is infinite, numerical optimizers will return a finite estimate that is typically much larger than the true parameter. Additionally, because a unique MLE does not exist, tests based on asymptotic results are likely

[3]Defining "true" standard errors is difficult given the infinite expectations. We use the curvature of the likelihood at the true parameters to reflect this quantity despite the violation of standard regularity conditions.

misleading as a unique MLE is a standard regularity condition for these results.

These inflated estimates may not be of major concern if the standard errors also inflate enough to prevent type-1 errors; however, there is no guarantee that this will be the case. In our replication study below, where separation is detected, some null hypotheses are rejected only when the separation problem is ignored but not once it is corrected. While it is impossible to say which decision is correct, the presence of separation suggests that the former is more suspect than the latter. Additionally, inflated standard errors raise the prospect of type-2 errors and under-powered studies. In our simulations, we find that separation can severely affect power, and in Appendix B.5 we show an example where both type-1 and type-2 errors can increase when separation is present but goes uncorrected.

Two further complications emerge in moving from the binary to multinomial outcomes. First, because there are more categories in the outcome, samples need to be larger in order for the threat of separation by chance to disappear. For example, with one binary regressor and a binary outcome we just need enough observations for every box in the cross tabulation to be filled. As the number of outcomes increases, this task requires more observations. Second, common implementations of multinomial models (e.g., Stata or R) provide neither warnings of possible separation nor make any attempt to identify problematic regressors.

Moving to the strategic setting introduces two more complications. First, standard visual diagnostics are less informative. Specifically, common rules-of-thumb ask analysts to look for estimates that are implausibly large, while this can be an important red flag, it is often difficult to know exactly how big is too big. This determination is clouded in the strategic context where the scale parameter is not always fixed to 1 like it is in ordinary logits and probits. In fact, the scale parameters sometimes contain another player's estimated choice probabilities (e.g., Signorino and Tarar 2006) or are estimated as free parameters (e.g., Whang et al. 2013), making the context of "too big" difficult to pin down. Second, strategic models contain interdependent and endogenous parameters by construction. When separation leads to inflated estimates in one player's utility function, this inflation can spill over into

estimates of that player's conditional choice probability, which then affects the estimation of other players' utility functions. Analyzing strategic interdependence is a main motivator of structural modeling, but care must be taken to minimize biases that may cascade up a game tree.

2.1 Separation corrections

With logits and probits, the primary existing solutions to the separation problem involve PL estimation (Zorn 2005). Penalization requires the analyst to impart some extra-empirical information (i.e., information from outside the data) to induce numerical stability in the optimization routine. We want to choose information that encapsulates our belief that the coefficient estimates should not be too large. From a Bayesian perspective, penalization is a type of prior belief where the true parameters are unlikely to be huge for any particular variable. As Gelman et al. (2008) put it, the key idea is that large changes on the logit/probit scale (typically 5 or more) are very rare and the penalty/prior should reflect this understanding (2008, 1361). In most cases, this information takes the form of a Jeffreys prior penalty term that is maximized when the parameters are all zero, although others propose penalty terms based on the Cauchy with median 0 and scale 2.5 (Gelman et al. 2008) or log-$F(1,1)$ (Greenland and Mansournia 2015).[4] All of these penalties pull the estimates away from $\pm\infty$ and towards 0.

Before deriving the BR strategic estimators, we first describe the model. Consider the extensive-form deterrence game in Figure 1. There are two actors, A and B, each of whom has two actions $y_i \in \{0,1\}$ for $i \in \{A,B\}$. At the start of the game, each player receives private information in the form of an action-specific shock $\varepsilon_i(y_i)$. Each shock reflects private information that i has regarding her payoff for taking action y_i.

After receiving her information, A acts. If A chooses $y_A = 0$, the game ends at the

[4] As Greenland and Mansournia (2015) note, the degrees of freedom in the log-F can be increased to log-$F(m,m)$ where larger m lead to more severe shrinkage. We find that $m = 1$ works well and stick to that throughout, but analysts may consider adjusting this to their own needs.

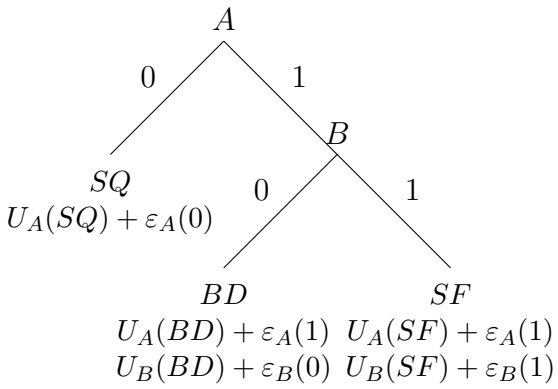

Figure 1: Standard two-player deterrence game

status quo (SQ). However, if A challenges B by taking action $y_A = 1$, then B responds by either backing down to A's challenge by taking action $y_B = 0$ (ending the game at BD) or standing firm against A by taking action $y_B = 1$ (ending the game at SF). When the game ends at outcome $o \in \{SQ, BD, SF\}$, players receive a payoff equal to $U_i(o) + \varepsilon_i(y_i)$. This payoff contains a deterministic component: $U_i(o)$ representing a commonly known and observable payoff to each player and a stochastic component: $\varepsilon_i(y_i)$, which is the privately known cost/benefit to player i for taking action y_i.

The solution concept for this game is quantal response equilibrium (QRE). At the QRE, B chooses 1 if $U_B(SF) + \varepsilon_B(1) > U_B(BD) + \varepsilon_B(0)$, which can be described as

$$y_B = \mathbb{I}\left[U_B(SF) - U_B(BD) + \varepsilon_B(1) - \varepsilon_B(0) > 0\right],$$

where $\mathbb{I}[\cdot]$ is the indicator function. Likewise, A chooses 1 if

$$y_A = \mathbb{I}\left[(1 - \Pr(y_B = 1))U_A(BD) + \Pr(y_B = 1)U_A(SF) - U_A(SQ) + \varepsilon_A(1) - \varepsilon_A(0) > 0\right].$$

To transform this game into an empirical model we need to (i) specify the deterministic portion of the utilities in terms of observed data and (ii) assume a distribution for the

action-specific shocks. For exposition, consider the following specification:

$$
\begin{aligned}
U_B(SF) &= X_B\beta \\
U_B(BD) &= 0 \\
U_A(SQ) &= X_{SQ}\ \alpha_{SQ} \\
U_A(BD) &= X_{BD}\ \alpha_{BD} \\
U_A(SF) &= X_{SF}\ \alpha_{SF} \\
\Pr(y_B = 1) &= p_B = F_B\left(U_B(SF)\right) \\
\Pr(y_A = 1) &= p_A = F_A\big((1-p_B)U_A(BD) + p_B U_A(SF) - U_A(SQ)\big),
\end{aligned}
$$

where F_i is the distribution that describes $\varepsilon_i(1) - \varepsilon_i(0)$. Our goal is to estimate the parameters $\theta = (\alpha, \beta)$ using D observations of actors playing this game. Standard practices estimate θ in one of two ways: a full information maximum likelihood (FIML) estimator or a two-step from Bas et al. (2008) called statistical backwards induction (SBI).

2.1.1 Statistical Backwards Induction

The SBI procedure is as follows:

1. Using only observations where $y_A = 1$, regress y_B on X_B using a logit or probit (depending on F_B) to produce $\hat{\beta}^{SBI}$. Estimate $\hat{p}_B^{SBI} = F_B(X_B\hat{\beta}^{SBI})$.
2. Regress y_A on $Z^{SBI} = \begin{bmatrix} -X_{SQ} & X_{BD}(1-\hat{p}_B^{SBI}) & X_{SF}(\hat{p}_B^{SBI}) \end{bmatrix}$ using a logit or probit (depending on F_A) to produce $\hat{\alpha}^{SBI}$.

Note that because each step is a binary choice model, the MLE for $\hat{\theta}$ solves

$$
\hat{\beta}^{SBI} = \underset{\beta}{\operatorname{argmax}} \sum_{d:y_{A,d}=1} \left\{ \mathbb{I}(y_{B,d}=1) \log\left[F_B(x'_{B,d}\beta)\right] + \mathbb{I}(y_{B,d}=0) \log\left[1 - F_B(x'_{B,d}\beta)\right] \right\} \quad (1)
$$

$$
\hat{\alpha}^{SBI} = \underset{\alpha}{\operatorname{argmax}} \sum_{d=1}^{D} \left\{ \mathbb{I}(y_{A,d}=1) \log\left[F_A(z'_d\alpha)\right] + \mathbb{I}(y_{A,d}=0) \log\left[1 - F_A(z'_d\alpha)\right] \right\},
$$

where $d = 1, \ldots, D$ indexes each observed play of this game.

Because this approach relies on two distinct binary outcome models, standard PL-based solutions apply. Let $L_B(\beta \mid y)$ and $L_A(\alpha \mid y)$ be the objective functions in Eq. 1, then the bias-reduced SBI (BR-SBI) estimates are

$$\begin{aligned} \hat{\beta}^{BR-SBI} &= \underset{\beta}{\operatorname{argmax}}\, L_B(\beta \mid y) + g(\beta) \\ \hat{\alpha}^{BR-SBI} &= \underset{\alpha}{\operatorname{argmax}}\, L_A(\alpha \mid y) + g(\alpha), \end{aligned} \tag{2}$$

where g is the logged penalty function. If the penalty is a density function (e.g., Cauchy or log-F) then g is the logged density function, while if g is the Jeffreys prior penalty then

$$g(\cdot) = \frac{1}{2}\log(\det(I(\cdot))),$$

where I is the estimated Fisher-information matrix calculated using the Hessians of the uncorrected log-likelihoods. Firth (1993, 36) suggests that standard errors for $\hat{\beta}^{BR-SBI}$ can be estimated using $I(\hat{\beta}^{BR-SBI})^{-1}$. This means that standard errors for $\hat{\alpha}^{BR-SBI}$ can be estimated using common two-step maximum likelihood results.

2.1.2 Full Information ML

The SBI estimator is easily implemented, but this ease comes at the cost of statistical efficiency. The FIML maximizes a single log-likelihood function that re-computes the choice probabilities at every step in the optimization process. Because the theoretical model has a unique equilibrium, the FIML is consistent and asymptotically efficient.

Using the above parameterization, the FIML estimates maximize the log-likelihood:

$$\begin{aligned} L(\theta \mid y) = \sum_{d=1}^{D} \{ & \mathbb{I}(y_{A,d} = 0)\log(1 - p_{A,d}) \\ & + \mathbb{I}(y_{A,d} = 1)\mathbb{I}(y_{B,d} = 0)\log(p_{A,d} \cdot (1 - p_{B,d})) \\ & + \mathbb{I}(y_{A,d} = 1)\mathbb{I}(y_{B,d} = 1)\log(p_{A,d} \cdot p_{B,d})\}, \end{aligned} \tag{3}$$

and the bias-reduced FIML estimates are given as

$$\hat{\theta}^{BR-FIML} = \underset{\theta}{\text{argmax}}\ L(\theta \mid y) + g(\theta). \tag{4}$$

If g is the logged Jeffreys prior, then the Hessian of Eq. 3 needs to be computed at each step in the numeric optimization process: a non-trivial task. Alternatively, Cauchy or log-F penalties can also be used. We provide an extension to R's `games` package called `games2` that allows analysts to fit the BR-FIML with Jeffreys prior, Cauchy(0, 2.5) or log-F(1, 1) penalties.

In choosing among these three penalties, we point out some pros and cons. The main advantages of the Jeffreys prior are that it is widely used and implemented for binary outcome models; as such, the BR-SBI with Jeffreys prior can be easily fit using existing software. For the FIML, however, the Jeffreys prior requires that the Hessian be negative definite at every guess of the parameter values. This requirement always holds with logits and probits but can fail in more complicated likelihoods. When the logged Jeffreys prior does not exist, density-based penalties based on the Cauchy or log-F distributions provide easy-to-use alternatives. Additionally, the density-based penalties perform best in simulations. In particular, the log-F penalty performs very well, although all three offer vast improvements over the uncorrected methods. Further, Beiser-McGrath (2020) finds that the Firth correction can be problematic in the kind of large-N, rare-events data that dominate international relations. Specifically, he finds that the Jeffreys prior penalty can produce estimates that are in different directions from the original results, implying that this penalty may do more than just shrink the estimates. Separation-induced inflation is always away from zero, so sign changes are concerning. Given this finding, the density-based penalties may be preferred, but we recommend that analysts consider multiple penalties where possible to ensure that the corrections are not dependent on the specific penalty.

2.2 Detecting Separation

Having considered the nature of and solutions to the separation problem, we are left

with the task of diagnosing it within specific samples. Current advice in political science is to look for point estimates and standard errors that are so large as to strain credibility. However, the different and sometimes endogenous scale parameters used in strategic models makes defining "too big" potentially ambiguous. As an alternative, we introduce an easy-to-use linear programming (lp) diagnostic from Konis (2007) to political scientists. We defer technical and implementation details to Appendix A and instead describe its application to strategic models.[5]

The lp-diagnostic is designed for binary outcome data and can be applied to the SBI without change. Directly generalizing this diagnostic to the full information strategic setting is infeasible, because the full design matrix contains the endogenous quantity p_B. As a result of this endogenous quantity, we cannot know *a priori* if separation exists between the covariates describing A's decision-making and the three outcomes of the strategic model. However, the lp-diagnostic can be applied both before and after estimation. We recommend the following work flow:

1. Using the observations where $y_A = 1$, check for separation in X_B and y_B .
2. Generate $\hat{p}_B^{SBI}$ and Z^{SBI}. Check for separation in Z and y_A.
3. Post-estimation, use the lp-diagnostic to search for separation in $[Z^{SBI}, X_B]$ or $[Z^{FIML}, X_B]$ against each of the three outcomes (SQ, BD, SF), individually.

If separation is detected at any point, a BR estimator should be considered.

3 Performance

We now consider Monte Carlo experiments to compare the BR-SBI and BR-FIML estimators given by Eq. 2 and Eq. 4, respectively, to their unpenalized counterparts. The experimental setup is presented in Figure 2, where we consider four parameters. The β parameters and the variable X_B characterize B's payoffs, while the α parameters and X_A

[5]The diagnostic is implemented in the R function `detectseparation::detect_separation`.

form A's payoffs. Regressors X_A and X_B are i.i.d. Bernoulli(0.5), while values of α and β are chosen to induce separation. In the interest of space, we present the simplest experiment here, while additional and more realistic simulations are deferred to the online appendix.

Our main simulation considers a sparse model where separation is likely to emerge in the data recording B's choice of 0 or 1. Let B's choice be given by

$$\begin{aligned} y_B &= \mathbb{I}\left[-1 + 4X_B + \varepsilon_B(1) > 0 + \varepsilon_B(0)\right] \\ &= \mathbb{I}\left[-1 + 4X_B + \varepsilon_B(1) - \varepsilon_B(0) > 0\right]. \end{aligned}$$

Each error term is i.i.d. standard normal, such that $p_B = \Phi\left(\frac{-1+4X_B}{\sqrt{2}}\right)$. Note that a large, but not unreasonable, coefficient on X_B will ensure that in most samples $y_B = 0$ only when $X_B = 0$.

The DGP for player A is

$$\begin{aligned} y_A &= \mathbb{I}\left[-2(X_A p_B) + \varepsilon_A(1) > 1.5 + \varepsilon_A(0)\right] \\ &= \mathbb{I}\left[-1.5 - 2(X_A p_B) + \varepsilon_A(1) - \varepsilon_A(0) > 0\right]. \end{aligned}$$

In terms of Figure 2, the parameters of interest are $\alpha_0 = 1.5$, $\alpha_1 = -2$, $\beta_0 = -1$ and $\beta_1 = 4$. We repeat the Monte Carlo experiment 5,000 times with samples of size $D = 500$ and keep the results where the lp-diagnostic detects separation between X_B and ending the game at outcome BD. In cases where the lp-diagnostic does not detect separation, the results are nearly identical across estimators. As with many applications of strategic probits, the status quo is the most common outcome (about 90% of observations), while BD and SF each emerge about 5% of the time. This means that the first step of the SBI typically has about 50 observations to use.

Before considering the simulation results, one additional point is worth mentioning. Recall that the expected value for $\hat{\beta}_1$ is undefined for the uncorrected estimators. As such, the observed estimates are whatever values get "close enough," such that the optimization software issues a successful convergence code. In other words, the numeric estimates pro-

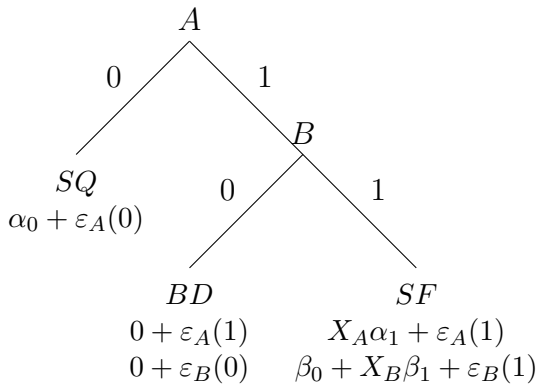

Figure 2: Monte Carlo version of the two-player deterrence game

duced by the ordinary SBI and FIML estimators reflect a type of regularization: They will be closer to the zero (and the truth) than the true MLE of $\pm\infty$, but in ways that are highly dependent on algorithm and tolerance choices.

3.1 Parameter Estimates

The Monte Carlo results are reported in Table 1. The first thing to note is that the BR techniques makes a noticeable and positive impact on both the point estimates and their precision. This translates into substantial decreases in the multivariate root-mean-squared error (RMSE). For both the SBI and the FIML, the PL approach helps when separation is present. The BR-FIML (log-F) has the smallest RMSE of all the estimators considered, while also having the least bias in estimating β_1.

Second, we see that the FIML estimators tend to outperform their SBI counterparts. One reason for this is that the FIML is a system estimator and will be more efficient by construction. However, it is also worth noting that the separation-induced inflation is worst in the unpenalized SBI and that while the FIML still exhibits bias, its RMSE is about 3/4 that of the SBI. These differences emerge in part because the SBI is less efficient by construction, but they are mostly due to differences in their default fitting algorithms.[6]

[6]See Appendix B.2 more discussion on these algorithmic differences.

Table 1: Monte Carlo results when separation is present in Player B's decision

Estimator	Quantity	α_0	α_1	β_0	β_1	RMSE
Ordinary SBI	Est	1.50	−3.05	−1.04	9.28	5.73
	St.Dev.	0.13	1.99	0.49	0.60	
	St.Err.	0.13	28.84	22.30	1140.57	
	Power	1.00	0.98	0.78	0.00	
	Coverage	0.95	0.91	0.96	1.00	
BR-SBI (Firth)	Est	1.51	−2.59	−1.00	3.90	1.37
	St.Dev.	0.13	1.23	0.39	0.41	
	St.Err.	0.13	0.75	0.39	1.05	
	Power	1.00	1.00	0.77	1.00	
	Coverage	0.95	0.89	0.96	1.00	
Ordinary FIML	Est	1.50	−3.19	−1.08	7.04	4.33
	St.Dev.	0.13	2.80	0.39	0.99	
	St.Err.	0.13	1.53	0.37	9902.47	
	Power	1.00	0.85	0.89	0.00	
	Coverage	0.95	0.96	0.96	1.00	
BR-FIML (Firth)	Est	1.50	−2.46	−0.96	3.90	1.01
	St.Dev.	0.13	0.88	0.32	0.33	
	St.Err.	0.13	0.79	0.33	1.08	
	Power	1.00	0.99	0.86	1.00	
	Coverage	0.95	0.94	0.96	1.00	
BR-FIML (Cauchy)	Est	1.52	−2.49	−0.97	4.27	0.93
	St.Dev.	0.13	0.72	0.33	0.37	
	St.Err.	0.13	0.83	0.34	1.35	
	Power	1.00	0.99	0.85	1.00	
	Coverage	0.96	0.94	0.96	1.00	
BR-FIML (log-F)	Est	1.51	−2.43	−0.92	4.03	0.76
	St.Dev.	0.13	0.60	0.32	0.32	
	St.Err.	0.13	0.77	0.33	1.19	
	Power	1.00	1.00	0.83	1.00	
	Coverage	0.96	0.94	0.96	1.00	
Truth	Parameters	1.50	−2.50	−1.00	4.00	
	St. Err. (SBI)	0.13	0.71	0.38	1.11	
	St. Err. (FIML)	0.13	0.77	0.33	1.17	

Note: St. Dev. refers to the standard deviation of estimates produced by the simulation. St. Err. refers to the standard errors produced by each estimator averaged over simulations. True standard errors are estimated using Hessian curvature at the true parameter values and the data within each simulation and then averaging over simulations. Power refers to the proportion of simulations where the null hypothesis is correctly rejected, and coverage refers to the proportion of simulations where the 95% confidence interval contain the true value.

3.2 Uncertainty

The next thing we want to consider is the uncertainty around these estimates. There are three quantities we consider here. First, we calculate the standard deviation of the estimates over the Monte Carlo iterations. These values are simulation estimates of the standard deviation of the sampling distribution for each parameter, making it an estimate of the "true" standard error (St. Dev. rows in Table 1). Second, we compute the true standard errors within each simulation by evaluating the relevant derivatives at the true parameter values and simulated data. Averaging over simulations gives us another estimate of the true standard errors (Truth rows in Table 1). Third, we will compare these values to the average computed standard errors at the estimates (St. Err. rows in Table 1). Absent separation, these three values should be nearly identical; with the numerical issues induced by separation they will diverge.

The ordinary SBI estimator does poorly here, only estimating the uncertainty around $\hat{\alpha}_0$ correctly. This status quo payoff is the only parameter not directly affected by p_B. The BR-SBI estimator does notably better, more closely approximating the standard errors obtained by evaluating the relevant derivatives at the true parameters. Interestingly, while the average standard error on α_1 is very close to what we expect the true standard error to be, this value is overconfident given the simulation results. Further analysis shows that BR-probit estimates of α_1 have a long tail in the direction of the separation which is why the simulation standard deviation is notably larger.[7]

Once again, the ordinary FIML tends to perform a bit better than the ordinary SBI.

[7]In the online appendix, we further consider the effect of p_B by asking: How much of the bias and variance in the SBI estimates of α can be attributed to estimating p_B? To answer this question, we rerun the main simulation where we fit only the second stage of each SBI, but with $\hat{p}_B$ fixed to its true value. Any differences these values and the SBI result in Table 1 can thus be attributed to estimating p_B in the presence of separation. We find that nearly all the problems in estimating α_1 go away when the estimate of p_B improves, which suggests that the SBI problems in estimating α_1 are second-order problems driven by the bias in $\hat{\beta}_1$ and $\hat{p}_B$.

Here, the three standard errors quantities closely match for both constant terms. As with the ordinary SBI, we see huge standard errors for $\hat{\beta}_1$ despite there being little actual variation across simulations. We also see some overconfidence in the average standard error of $\hat{\alpha}_1$ relative to the simulation standard deviation.

Overall, the BR-FIML standard errors closely match the true standard errors produced by evaluating the Hessian at the truth, providing some confidence in the procedure. However, like the BR-SBI, we observe that the standard error on $\hat{\beta}_1$ is notably larger than the simulated sampling distribution. As previously mentioned, we follow standard practices by using the Hessians from the uncorrected likelihoods when computing standard errors for all the BR procedures. Ignoring the extra-empirical information from the penalty produces, on average, conservative standard errors. Analysts who want this information included in their uncertainty measures may be better off adopting a Bayesian approach, as standard errors based on the BR-Hessian can be difficult to derive.

3.3 Coverage and Power

Another relevant measure here is coverage. Here, we report the proportion of 95% confidence intervals, calculated within each iteration using the estimated standard errors, contain the true parameter value. Ideally, this value will be 0.95. Larger values reflect conservative standard errors (over-wide intervals), while smaller values tend to reflect over-confidence with narrower intervals around a poor point estimate. In many cases, we see that coverage for everything but $\hat{\beta}_1$ is about 0.95. The most notable exceptions are α_1 for the SBI and BR-SBI where the coverage is too small. For the latter, this poor coverage again reflects a skewed simulated sampling distribution with a tail that pulls in the direction of $-\infty$. Interestingly, for $\hat{\beta}_1$ all six estimators have 100% coverage across the simulations.[8] In the uncorrected case, this is not surprising as the standard errors are orders of magnitude larger than the estimates and covering the true value is easy but not particularly meaningful. In the corrected case, high coverage reflects the conservative standard errors mentioned above.

[8] Analysts should take care to note that 100% coverage and/or power are definitely not general results.

At this point, it is worth reconsidering the practical consequences and whether the numerical and statistical issues with separation are worth worrying about. After all, if both the point estimates and standard errors inflate the way they do in Table 1, then a likely outcome is that researchers will fail to reject the null hypothesis for the numerically problematic parameters and the harm done is perhaps minimized. However, there is no guarantee that inflation will always be more pronounced in the standard errors. As we see with the Signorino and Tarar (2006) replication, below, and as Cook et al. (2018) show in their analysis of the multinomial logit, there are cases where separation appears present, based on visual or lp diagnostics, and the null hypothesis is rejected only when the issue goes unnoticed/uncorrected. As such, it is not obvious that separation is relatively harmless from a type-1 error perspective. Additionally, type-1 errors are not the only mistakes that matter. With inflated standard errors, type-2 errors may become more pronounced as well. Unsurprisingly given their variances, the uncorrected estimators have extremely low (zero) power with respect to the hypothesis $\beta_1 = 0$. In contrast, the BR estimators correctly reject the null hypotheses at high rates. The high power *and* coverage of the BR estimators highlights their usefulness at producing both reasonable estimates and inference when separation is present.

3.4 Choice Probabilities

Moving beyond the point estimates, $\hat{p}_B$ plays a key part in fitting the model, particularly for the SBI. As such, we want to know if any of these corrections have negative consequences on estimating p_B. In Table 2, we consider the statistical properties of $\hat{p}_B$. Because X_B is binary, there are only two values that p_B can take on, making it easy to break down this analysis by X_B. There are three important takeaways from these results. First, the BR-FIMLs are more biased when estimating p_B when $X_B = 0$, this result matches Rahman and Sultana (2017) who finds that BR correction in the parameters can sometimes make bias in predicted probabilities worse. Second, despite this bias when $X_B = 0$ the BR estimators offer modest improvements in RMSE when $X_B = 0$ and substantial improvements in both bias and RMSE when $X_B = 1$. These latter results are unsurprising given the inflation in

$\hat{\beta}_1$. Third, when combining the results we see that the three BR-FIMLs are most preferred from a RMSE perspective, despite having more bias when $X_B = 0$. The bias and RMSE improvements they offer when $X_B = 1$ offset these concerns in this experiment.

Table 2: Bias and RMSE in estimating p_B

	$X_B = 0$		$X_B = 1$		Combined	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
Ordinary SBI	0.003	0.084	0.017	0.017	0.010	0.061
BR-SBI	0.008	0.082	-0.003	0.006	0.002	0.058
Ordinary FIML	-0.009	0.076	0.017	0.017	0.004	0.055
BR-FIML (Firth)	0.015	0.072	-0.002	0.005	0.006	0.051
BR-FIML (Cauchy)	0.013	0.074	0.007	0.008	0.010	0.052
BR-FIML (log-F)	0.022	0.074	0.003	0.005	0.012	0.053

4 Application: Deterrence in interstate disputes

We now reexamine results from Signorino and Tarar 2006 who study deterrence in interstate disputes using data on 58 crises between 1885 and 1983. The players in this game are an aggressor and defender state. The aggressor (A) decides between attacking a protégé state of the defender (B) or preserving the status quo. If A chooses the latter, the game ends, but if A chooses the former, then the defender can either protect its protégé or back down. The dependent variable takes on three values: status quo, attack-war, attack-back down. Appendix C contains descriptions of the independent variables and the model specification. We start by applying the lp-diagnostic to the data. The diagnostic results are reported in Table 3, where four of five checks provide evidence of separation.[9]

Compounding the separation problem is the issue of fitting a complicated strategic model to a relatively small sample. In replicating these results, we found that the determinant of the FIML information matrix is negative at many steps in the optimization process, making the logged Jeffreys prior penalty term undefined. As a result, we use the log-F penalty as it does not rely on the curvature of the baseline log-likelihood and performed well in simulations. The BR-SBI continues to use the Jeffreys prior penalty here as the probit objective function

[9]We also consider an application where we believe separation is not present in Appendix D.

Table 3: Checking for separation in Signorino and Tarar (2006)

Regressors	Outcome	Result
X_B	$y_B \mid y_A = 1$	Yes
Z^{SBI}	y_A	No
$[Z^{FIML} \quad X_B]$	$\mathbb{I}(y_A = 0)$	Yes
$[Z^{FIML} \quad X_B]$	$\mathbb{I}(y_A = 1)\mathbb{I}(y_B = 0)$	Yes
$[Z^{FIML} \quad X_B]$	$\mathbb{I}(y_A = 1)\mathbb{I}(y_B = 1)$	Yes

Note: The Z variables are transformed using estimates of p_B from the unpenalized estimators.

does not have the same complexity as the FIML, the penalty always exists, and it remains the most common choice for binary-outcome models. Beyond these difficulties, we also note that fitting a 21 parameter strategic model with 58 observations is a demanding proposition. Nonetheless, this example provides us with a clear case where separation is present.

The results are presented in Table 4. Fitting the ordinary SBI produced severe numerical instability; as such, the estimates and standard errors are the means and standard deviations from a non-parametric bootstrap where we discard results beyond ± 50 to keep everything on roughly the same scale across the estimators. The fact that we even had to consider this approach with the SBI is a warning against using an uncorrected model. There are slight differences between the replicated FIML and published results, which we attribute to slight differences in software implementation.

What is most striking about the results in Table 4 is that while many of the point estimates have the same sign across all four estimators, some results that were significant in the Signorino and Tarar (2006) analysis are no longer significant at traditional levels. Additionally, we note that the estimates and standard errors on the uncorrected SBI are incredibly large despite the precautions we took to make the estimates appear more reasonable. Combining this observation with the lp-diagnostic results provides us with good reason to suspect that a BR estimator may be more appropriate. Indeed, the two BR estimators largely agree with each other in terms of magnitude and sign in 18 out of 21 estimates, although in the BR-SBI case fewer estimates are statistically significant at standard levels. This difference may result from the relative inefficiency of the two-step estimator.

The fact that a few estimates change signs across the estimators is an interesting puzzle. Specifically, cases where signs are different across the ordinary-SBI and BR-SBI are unexpected. Correcting for separation is not supposed to change the direction of an estimate, although Beiser-McGrath (2020) notes that this can happen in some binary-outcome cases with the Jeffreys prior penalty. He finds this to be the case in large, rare-events data. Here, however, we see sign flips in small, even-event data, and it also occurs with the density-based penalties. These unexplained sign flips may suggest that there may be some heretofore unknown issues with BR estimation in (very) small samples. In some exploratory simulations, we find that sign flips can happen in small, highly colinear samples like this one, but we cannot be certain that colinearity is causing the sign flips here. Future work should spend more time on this puzzle as it is very unusual to see signs change when applying PL methods.

In examining player B's (the defender's) utility function, Signorino and Tarar (2006) find that the defender is more likely to support its protégé if B has nuclear weapons, if the protégé imports a lot of arms from B, and if there was a past, but unresolved, crisis between the defender and the aggressor (2006, 593). Our analysis concurs with these results in terms of sign, but only the effect of nuclear weapons remains significant at the 5% level. The overall decrease in coefficient magnitudes is consistent with a separation problem. The changes in significance suggest that some original findings resulted from separation-induced inflation in the point estimates that exceeded the inflation in the standard errors. Many of these findings may, of course, still be true, but we cannot reject these null hypotheses with these data once we correct for separation.

The uncorrected SBI is the most conservative model here: it rejects no hypotheses and, as such, makes no type-1 errors. In contrast, we may suspect that the uncorrected FIML is guilty of some type-1 errors, making the SBI, and its extreme results, a safe choice for cautious analysts. However, this protection against type-1 errors comes at the cost of power. Based on the simulations in Table 1 and in Appendix B, we find that the uncorrected SBI has almost no power to identify effects on coefficients where separation is a concern. Analysts can

Table 4: Signorino and Tarar Replication

	FMLE	SBI	BR-FMLE	BR-SBI
U_A(SQ): Const.	−5.04	−5.69	−1.30	−3.06
	(2.39)	(6.70)	(1.65)	(1.75)
U_A(SQ): Tit-for-Tat	17.27	2.13	2.47	1.40
	(7.22)	(2.24)	(1.04)	(0.62)
U_A(SQ): Firm-Flex	6.59	1.05	1.26	0.61
	(3.26)	(2.44)	(0.88)	(0.59)
U_A(SQ): Democratic Attacker	15.75	−0.40	0.52	−0.65
	(8.60)	(3.96)	(1.57)	(0.93)
U_A(SQ): Year	−0.35	−0.03	−0.03	−0.01
	(0.18)	(0.05)	(0.03)	(0.01)
U_A(BD): Const.	13.51	−7.34	1.94	−3.03
	(12.76)	(6.75)	(2.83)	(2.18)
U_A(War): Nuclear	−9.13	−0.50	−0.99	−0.90
	(5.00)	(10.55)	(1.47)	(1.95)
U_A(War): Immediate Balance	−12.51	−2.93	−1.43	−1.15
	(5.26)	(6.28)	(1.10)	(0.64)
U_A(War): Short-term Balance	−6.22	−3.84	−1.89	−3.18
	(3.26)	(7.34)	(1.62)	(1.80)
U_A(War): Long-term Balance	3.35	0.83	0.45	0.69
	(1.57)	(2.57)	(0.50)	(0.53)
U_A(War): Military Alliance	12.62	1.92	2.53	1.31
	(5.23)	(3.49)	(1.37)	(1.10)
U_A(War): Arms Transfers	−0.86	−0.23	−0.16	−0.06
	(0.49)	(0.35)	(0.16)	(0.13)
U_B(War): Const.	−10.93	−20.30	−2.71	−1.07
	(5.88)	(15.84)	(1.32)	(1.43)
U_B(War): Nuclear	6.64	−3.52	2.41	0.13
	(2.62)	(20.04)	(1.10)	(1.74)
U_B(War): Immediate Balance	5.46	16.46	1.22	0.66
	(2.90)	(14.73)	(0.74)	(0.88)
U_B(War): Short-term Balance	4.16	3.80	1.24	0.25
	(2.37)	(15.06)	(0.79)	(1.09)
U_B(War): Military Alliance	13.39	11.69	1.64	2.39
	(7.61)	(19.37)	(1.59)	(1.86)
U_B(War): Arms Transfers	−1.75	−1.47	−0.29	−0.38
	(0.86)	(2.09)	(0.23)	(0.24)
U_B(War): Foreign Trade	4.85	5.96	0.90	0.71
	(2.55)	(2.71)	(0.54)	(0.41)
U_B(War): Stalemate	8.40	16.37	1.38	1.79
	(4.21)	(14.65)	(1.12)	(1.18)
U_B(War): Democratic Defender	5.93	1.72	1.08	−0.02
	(2.86)	(11.34)	(0.87)	(1.05)
Observations	58	58	58	58

Notes: Standard errors in parentheses (Model 6 is bootstrapped)

weigh their own acceptance for type-1 and type-2 errors, but we find that the BR estimators present a good balance between these two concerns.

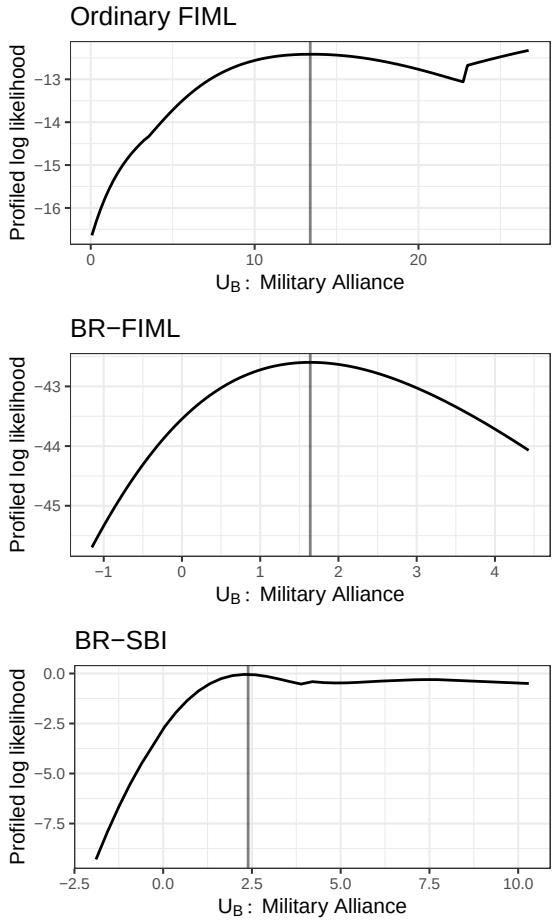

Figure 3: Profiled log-likelihood on the coefficient associated with how a military alliance affects B's decision to intervene.

To better demonstrate these numeric issues and illustrate how the BR corrections work we consider the profiled log-likelihood of the FIML, the BR-FIML, and the BR-SBI for the coefficient on military alliance in B's utility function. We focus on this variable as the uncorrected coefficient estimate of about 13 (against a scale of $\sqrt{2}$) is suggestive of a separation problem. The profiling procedure fixes the value of a single coefficient and refits the model. Repeating this procedure at many values demonstrates the model's sensitivity to changes in this estimate. For a well-behaved problem, we would expect a classic upside-down U shape with a maximum at the estimated parameter value. The profiled results are

shown in Figure 3. Specifically, for the ordinary FIML (top pane) there appears to be a local maxima at the estimate, but model fit can be improved by increasing this estimate past positive 20. Put another way, while the estimate is a local maximum, it is not the global maximum; "better" fit can be found at estimates further toward ∞. This push towards $\pm\infty$ is the classic sign of the separation problem. Looking at the two BR profiles we see that, at least in the range considered, the estimates are at the maximum. Note that the BR-SBI has a flattish section at the right-hand end of the plot, however, this drops off quickly if we explore past this region, and we find no reason to suspect that there are better log-likelihood values beyond the range presented here.

5 Conclusions and Recommendations

Penalized likelihood methods provide a useful technique for addressing separation in discrete choice modeling. In this paper, we adapt PL methods to estimate the parameters from extensive form games of incomplete information. Using Monte Carlo experiments and replication analysis we find that the BR estimators offer substantial gains in bias, RMSE, and numerical stability. We offer two strategies (BR-SBI and BR-FIML) that provide analysts with options for fitting games to data where separation problems exists. The BR-SBI is easily implemented using the existing R package `brglm`, while we offer our own R package, `games2`, for fitting the BR-FIML. Additionally, we describe tools to diagnose separation in situations where software does not issue warnings and standard visual inspections are less clear because of differences in the scale parameters. Our recommendation uses the linear programming diagnostic from Konis (2007). We detail five ways to use this tool with strategic models that are fast and easy for analysts.

Additionally, the simulations and application allow us to note some limitations in PL methods for fitting strategic models. Notably, fitting strategic models to small samples can be very demanding of the data and lead to numeric concerns beyond just separation. For example, in the Signorino and Tarar (2006) application we found the Jeffreys prior approach

to be unreliable as the Hessian of log-likelihood function was not negative definite at many guesses of the parameters. This experience leads to our first piece of advice: When the Jeffreys prior struggles, analysts should consider one of the density based penalizations. While we observe that log-$F(1,1)$ tends to be the best choice, we found almost no cases where the differences between the log-F and Cauchy penalties are pronounced. As such, analysts should feel comfortable with either of these approaches, even with small samples. That said, sensitivity to the exact penalty may indicate that there is not enough information to provide meaningful analysis. At this point, analysts may want to consider using a less demanding model. This leads to our second piece of advice: To the extent that various penalties might produce difference results, analysts should note any differences and consider additional analysis to assess the sensitivity of their results to the penalty choice. This analysis may require additional programming as analysts may want to try a range of (non-standardized) t, log-$F(m,m)$, or other distributions in assessing this sensitivity.

Several avenues of future work present themselves. First, researchers should consider extending the BR framework even further into the empirical analysis of discrete choice games. For example, extensive-form signaling models are also common in EITM studies of international relations (e.g., Crisman-Cox and Gibilisco 2021). Extending the BR framework could be helpful for scholars interested in empirical models of strategic interactions.

Second, more work should be done on the benefits that BR estimation can bring to small-sample strategic models even absent separation concerns. As mentioned, the original contribution from Firth (1993) was to reduce finite-sample bias in logit models. It is likely that BR estimation can be helpful to strategic models in this context, however, more analysis needs to be considered regarding the trade-off between improved point estimation and potentially worse estimation of choice probabilities. This analysis is particularly important with strategic models given their endogenous construction. With separation, we find some evidence that this trade-off exists in strategic case, but that the benefits outweigh the costs in the cases we considered. More work should assess this trade-off in finite samples absent

separation.

Finally, there are many discrete choice models that may be vulnerable to separation and where scholars may benefit from knowing more about how well standard corrections work. For example, bivariate, multinomial, and spatial probits along with various tobit models (e.g., selection models) all involve categorical outcome that can be affected by separation, but it remains an open question as to how well different penalization solutions perform in these cases. Given recent concerns about the Jeffreys prior approach in international relations data (e.g., Beiser-McGrath 2020) and our own problems with Jeffreys in the Signorino and Tarar (2006) example, more analysis of density based solutions in these more complicated models will be highly useful.

References

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

Bas, M. A., C. S. Signorino, and R. W. Walker. 2008. "Statistical Backwards Induction: A Simple Method for Estimating Recursive Strategic Models." *Political Analysis* 16(1): 21–40.

Beiser-McGrath, L. F. 2020. "Separation and Rare Events." *Political Science Research and Methods* 10(2): 428–437.

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

Crisman-Cox, C. and M. Gibilisco. 2021. "Estimating Signaling Games in International Relations: Problems and Solutions." *Political Science Research and Methods* 9(3): 565–582.

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

Gelman, A., A. Jakulin, M. G. Pittau, and Y.-S. Su. 2008. "A Weakly Informative Default Prior Distribution for Logisitic and Other Regression Models." *The Annals of Applied Statistics* 2(4): 1360–1383.

Greenland, S. and M. A. Mansournia. 2015. "Penalization, Bias Reduction, and Default Priors in Logistic and Related Categorical and Survival Regressions." *Statistics in medicine* 34(23): 3133–3143.

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

Rahman, M. S. and M. Sultana. 2017. "Performance of Firth- and Log f-type Penalized Methods in Risk Prediction for Small or Sparse Binary Data." *BMC Medical Research Methodology* 17(1): 1–15.

Rainey, C. and K. McCaskey. 2021. "Estimating Logit Models with Small Samples." *Political Science Research and Methods* 9(3): 549–564.

Signorino, C. S. 1999. "Strategic Interaction and the Statistical Analysis of International Conflict." *American Political Science Review* 93(2): 279–297.

Signorino, C. S. and A. Tarar. 2006. "A Unified Theory and Test of Extended Immediate Deterrence." *American Journal of Political Science* 50(3): 586–605.

Whang, T., E. V. McLean, and D. W. Kuberski. 2013. "Coercion, Information, and the Success of Sanction Threats." *American Journal of Political Science* 57(1): 65–81.

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