So the Reviewer Told You to Use a Selection Model? Selection Models and the Study of International Relations *

Patrick T. Brandt
School of Economic, Political and Policy Sciences
University of Texas at Dallas
E-mail: pbrandt@utdallas.edu

Christina J. Schneider
Department of Politics and International Relations
University of Oxford and Max Planck Institute of Economics
E-mail: christina.schneider@politics.ox.ac.uk

*An earlier version of this paper was presented at the Annual Meeting of the Midwest Polit ical Science Association, April 15–18, 2004, Chicago, Illinois, the University of Texas at Dallas, and the University of Konstanz. We would especially like to thank Duane Gustavus of the UNT Academic Computing Center for access and help with the Beowulf/Linux cluster; John Freeman for feedback on earlier drafts; Fred Boehmke, Harold Clarke, Chetan Dave, Michael Greig, David Mason, Thomas Plümper, and Vera Troeger for useful discussions on this topic; Bill Reed for providing data; Kevin Clarke and Rob Lane for insightful comments on an earlier draft; and Kevin Quinn and Jeff Gill for feedback on the MCMC estimation of the censored probit model. Computer code in R and/or Stata is available from the first author. Brandt's work has been sponsored by the National Science Foundation under award SES-0351205 and SES-0540816. All errors remain the responsibility of the authors.

So the Reviewer Told You to Use a Selection Model?
Selection Models and the Study of International Relations

Abstract

Selection models are now widely used in political science to model conditionally observed processes (e.g., conflict onset and escalation, democratization and foreign direct investment, voter turnout and vote choices). We argue that many applications of selection models are poorly identified since the same predictors are used for predicting selection and the outcome of interest and only few additional exogenous regressors are included to predict selection. The paper shows that this can lead to biased inferences and incorrect conclusions about the presence and effects of selection. We then propose methods to evaluate the costs and consequences of poorly specified selection models and give guidance about how to identify selection models and when to abstain from estimating such models. A replication of Reed's (2000) analysis of conflict onset and escalation further illustrates our results in a substantive application.

Introduction

> We hope that future research of questions in international politics will seek not only to ameliorate the threat of selection bias but also to model the layered processes about which researchers often theorize. (Reed and Clark 2000, 393)

Sample selection models have become an important tool in political science. In his work, Reed claims "discrepancies between the theoretical expectations and the empirical results" because scholars fail to link related political processes in a unified empirical model in order to account for selection effects (Reed 2000, 84). This sample selection occurs when the presence of an observation is non-random or if the outcome of interest is observed in only some instances. In this case, the dependent variable of interest is only observed if some indicator of selection exists.[1] Sample selection creates a well known problem: sample selection or omitted variable bias. Reed (2000) illustrates the implications of estimating a selection model versus independent probit models for two major international relations theories of conflict escalation. Most importantly, accounting for selection reverses previous findings about the impact of power parity and democratic peace on conflict escalation.

Reed's paper marks a cross road in the application of selection models to political science research. Since many research questions—particularly in international relations research—can be framed as a selection problem, scholars oftentimes face the question of why they have abstained from estimating a selection model when submitting articles to a scientific journal. The increasing popularity of selection models among political scientists and the availability of software packages to estimate these models has led many to regard those models as "the" solution for the problems of censored data (e.g., Reed 2000, Lemke and Reed 2001, Meernik 2001, Fordham and McKeown 2003, Hug 2003, Jensen 2003, Sartori 2003, Sweeney and Fritz 2004).[2] A search of political science articles indexed in the Social Science Citation Index (SSCI) shows that the number of articles dealing with selection models in political science increased from a mere two in 1995 up to 17 in 2006. In the five-year period after Reed's contribution, the average number of articles increased to 14.6 publications per year compared to an average of 8 articles per year in the five years before. Until 2007, the two articles by Reed and Lemke and Reed have been cited over 87 times.[3]

This rise in the application of selection models drives more calls for their application by researchers and reviewers. In this paper, we argue that the process of reviewers' asking for a selection model poses a "flawed" selection process itself, as political scientists tend to address only one aspect of sample selection while ignoring another. In the effort to correct for possible sample selection, a simultaneity problem might arise in specifying the selection and outcome equations. Since the researcher must draw on information outside of the main outcome to account for selection, she must consider a second equation or data generation process. Determining the presence and structure of sample selection via a statistical model thus includes both an identification step and a statistical modeling step (Manski 1995, 6). The two steps are closely intertwined and must be addressed together. Most political science applications of selection models, however, do not separately deal with the identification issue. Typically, they use some of the same information to explain selection and the outcome of interest. This is problematic because a failure to properly identify the selection process can lead to biases that are as bad as the original sample selection problem that researchers are trying to correct.

In fact, many of the applied sample selection models in political science suffer from poor identification. This problem emerges because scholars typically estimate a model without sample selection and are then confronted with criticisms that there "may be sample selection." The researcher's response is to include another equation in the model to 'explain' selection. In these situations, a selection model may be appropriate, but the quality of the estimation depends on *identifying* the predictors of sample selection *separate from* the outcome of interest. Since this identification problem must be solved *a priori*, calling for sample selection models may only lead to further problems if the model is not well identified.

To clarify the problems that result from an incorrect specification of selection models, we conduct Monte Carlo simulations reflecting typical political science applications of selection models.

The simulation results illustrate how sensitive selection model inferences are to identification assumptions. Most importantly, the hypothesis tests that researchers usually rely on to determine whether the estimation of a selection model is warranted have poor power and size if the selection model is poorly identified. In this case, researchers are likely to reject the null hypothesis of no selection when in fact the two equations are independent from each other. As our results indicate, the potential costs of incorrectly specifying a selection model are severe: in this case, the coefficients of the selection model tend to be biased even worse than the results of an independent probit (i.e., a non-selection) model, even if selection is present. A replication of Reed's analysis of conflict onset and escalation further illustrates those problems.

Our contributions to the political science literature are threefold. First, while we find that selection models are widely used across all fields in the social sciences, our findings indicate that the "cure" may be worse than the "disease." In other words, scholars should not ingenuously refer to a selection model just because they face problems of missing covariates, strategic interactions, and non-random data—instances in which they (or others) think that selection models may be necessary—since the appropriateness of a selection model and the quality of the results are highly sensitive to the identification of the selection process itself. To the contrary, researchers should assess very carefully whether a selection model is the appropriate estimation method in their case. Scholars are able to avoid estimating selection models at the costs of 'good' research only if the choice of a selection model is theoretically and empirically warranted. Secondly, our findings point to strategies to assess the fragility of these models and to improve the specification and estimation of selection models. Finally and probably most importantly, we provide recommendations for scholars who consider employing selection models. These recommendations offer guidance about whether to use a selection model and how to proceed in the specification and estimation process.

Identification of Selection Models and International Relations Research

The problem of sample selection involves the following issue: Suppose for every observation, there are three variables that characterize the observation: (y_1, y_2, x). Here, y_1 is the 'selection' variable that is 1 if the value of y_2 is observed (e.g., conflict onset) and zero, otherwise (e.g., no conflict). The variable y_2 is the main 'outcome' that we wish to predict (e.g., conflict escalation), conditional on some covariates in x (e.g., quality of democratic institutions, power parity, participation in alliances). The sample selection data generation process is a random draw of (y_1, y_2, x) for each observation where one always can observe (y_1, x), but y_2 is only observed for cases where $y_1 = 1$. The probability of observing y_2 depends on the selection probability, $Pr(y_1 = 1|x)$ and the censoring probability $Pr(y_1 = 0|x)$. The censoring problem is that the probability of the outcome for the censored cases, $Pr(y_2|x, y_1 = 0)$ is unobserved.

The nature of the selection problem becomes evident by looking at the conditional probability of the dependent variable of interest, y_2:

$$Pr(y_2|x) = Pr(y_2|x, y_1 = 1)Pr(y_1 = 1|x) + Pr(y_2|x, y_1 = 0)Pr(y_1 = 0). \quad (1)$$

The fact that neither a sample, nor even an infinite number of observations can convey any information about $Pr(y_2|x, y_1 = 0)$ poses an identification problem (Manski 1995, 26).

In what follows we discuss the trade-offs of using different non-sample information for the identification of parametric models of sample selection and highlight how identification failures—which induce bias and inefficiency in the estimates—may lead to situations where the effort to specify and estimate selection models is useless. The importance of this is almost self-evident: as the debates around Reed's publication clarify, by choosing the "wrong" model, the researcher runs the risk of incorrectly supporting or falsifying a theory. If the aim is to find strategies that avoid these problems, one must to identify their sources.

A standard solution to issues of identification and statistical issues of censored samples are two equation selection models (Heckman 1976, Heckman 1979). Meng and Schmidt (1985) and Dubin and Rivers (1989) extended the traditional Heckman selection model for the case where the second stage outcome model is based on either a logit or a probit model. Our analysis focuses on the latter models as political science applications of Heckman probit selection models are more widespread than the original Heckman model (e.g., Berinsky 1999, Boehmke 2003, Gerber, Kanthak and Morton 1999, Sartori 2003, Lemke and Reed 2001, Reed 2000, Meernik 2001).[4] The censored probit model consists of two equations. The first or selection equation defines the cases that are observed. For these observations, there is a second binary dependent variable for the outcome equation. Let y^*_{i1} be the selection variable, or the latent variable that determines whether case i is observed and y^*_{i2} be the latent variable in the second or outcome stage. We can write a system for these latent variables as functions of variables x_{ij}:

$$y^*_{i1} = x_{i1}\beta_1 + \epsilon_{i1} \tag{2}$$

$$y^*_{i2} = x_{i2}\beta_2 + \epsilon_{i2} \tag{3}$$

$$\begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right) \tag{4}$$

where x_{i1} are the covariates for the selection equation, x_{i2} are the covariates for the outcome equation, and β_i are the coefficients for the $i'th$ equation. The correlation across the equations, $\rho \in (-1, 1)$ indicates whether there is sample selection. If $\rho = 0$ then there is no sample selection present. The variables y^*_{ij} are related to the observed outcome and selection $y_{ij} \in \{0, 1\}$ by

$$y_{i2} = \begin{cases} 1 \quad if \quad y^*_{i2} > 0 \quad and \quad y^*_{i1} > 0 \\ 0 \quad if \quad y^*_{i2} \leq 0 \quad and \quad y^*_{i1} > 0 \end{cases} \tag{5}$$

Values of y_{i2} are only observed in the second stage if an observation was selected in the first stage, or $y_{i1} = 1$.

In order to estimate the model by maximum likelihood (ML), the researcher chooses variables x_{i1} and x_{i2} for the selection and outcome equation respectively. This choice is non-trivial, since the identification problem has been translated from the conditional probability relations in equation (1) to the parametric selection model. Consequentially, the selection of variables determines the identification of the unobserved probability of $Pr(y_2|x, y_1 = 0)$.

Standard identification of multiple equation models such as the selection model in equations (2) and (3) requires sufficient unique information in the model's $x'_{ij}s$ to separately identify and estimate the parameters in the selection and outcome equations. First, the variables in x_{i1} and x_{i2} may not be perfectly collinear to ensure that the covariance of the parameters is non-degenerate and the maximum likelihood estimates exist (Poirier 1980, Meng and Schmidt 1985, Sartori 2003). Second, one must have good instruments or *exclusion restrictions* that include a variable in the selection equation and not in the outcome equation.[5] To sum up, for identification it is most important to include enough information in the selection equation regressors to ensure that they are unique with respect to the other parameters in the outcome equation.[6]

The most common means of identification is to employ exclusion restrictions or unique regressors that appear only in the selection equation. Often though, a large number of the same explanatory variables appears in both the selection and outcome equations (see the citations to the literature above). In this situation, the parametric selection model—the two simultaneous equations—are subject to the similar identification requirements as a simultaneous equation or instrumental variables model. Here the instrumental variable(s) is (are) those providing information about the selection process. Yet this type of exclusion restrictions trades the identification problem in equation (1) for one of instrumental variables in a simultaneous equation models. It is well known that in such models, the instruments (in this case the covariates in the selection equation) need to be good predictors of the variable being instrumented *and* uncorrelated with the errors in the final outcome equation (Greene 2002, 196–197). The variables included in the selection equation, but excluded from the outcome equation, need to do a good job predicting sample selection, but should be uncorrelated with the errors in the outcome of interest. However, this result is theoretical and hard to know in a statistical sense.[7]

Accordingly, 'simple' exclusion restrictions are only a partial solution to the identification problem in selection models, since the excluded variables need to be good instruments for the selection process. If the excluded parameters are not good instruments for the selection process then the lack of identification and the potential misspecification of the selection process can lead to more problems than it solves. The uncertainty in the estimates from the poor identification and the estimation of a possibly incorrect model (the sample selection model instead of an independent probit model) may lead to more bias than ignoring the sample selection and estimating just the outcome equation.[8]

Our discussion has three important implications for selection models. First, if the estimates of the selection model are based on a poorly identified or specified model, they will likely be biased and inefficient. Second, hypothesis tests for selection are likely to be incorrect. Finally, estimates of the impacts of the regressors on the probabilities of the outcomes—the quantities of interest in most models—will also be biased because they are based on the incorrectly estimated coefficients. Since the primary inferences are about 1) the probability that selection is present and 2) the bias of the estimates with and without accounting for selection, knowing how robust these models are to weak identification assumptions and highly correlated regressors is of the utmost importance.

Monte Carlo Experiments and Results

In this section, we present a series of Monte Carlo experiments evaluating the robustness of a sample selection models to different identifying information. The data generation process used in the Monte Carlo analysis is based on equations (2) and (3). Our simulations incorporate two different sets of experiments or data generating processes. In the first, we consider a model with separate, unique regressors in both the selection and the outcome equation:

$$y_{i1}^* = 0.75 + 0.5x_{i11} - 0.5x_{i12} + \epsilon_{i1} \quad (6)$$

$$y^*_{i2} = 0.15 + 0.5x_{i21} + 0.5x_{i22} + \epsilon_{i2} \quad (7)$$

where $x_{ijk} \sim N(0, 1)$. We assume that x_{ijk} are uncorrelated with each other in this first model. Thus, there are two unique regressors in each equation or two excluded from each equation. We refer to this model as well identified.

In the second data generation process, the model includes the same explanatory variables x_{i11} and x_{i12} in the selection and the outcome equation. The model is identified by including variable x_{i13} into the selection equation:

$$y_{i1} = 0.75 + 0.5x_{i11} - 0.5x_{i12} + 0.25x_{i13} + \epsilon_{i1} \quad (8)$$

$$y_{i2} = 0.15 + 0.5x_{i11} + 0.5x_{i12} + \epsilon_{i2}. \quad (9)$$

We refer to this model as weakly identified since only one exclusion restriction (x_{i3}) is used to identify the model.[9] This second model parallels the common practice in political science of including extra regressors in the selection equation to achieve identification. For each of these models, we generated 2000 samples of 1000 observations with values of $\rho = \{-0.8, -0.4, 0, 0.4, 0.8\}$ and estimated the censored probit model and two independent probit models (one for each equation, with the outcome equation model estimated only using the selected sample).[10]

At first, we examine the properties of the hypothesis tests for sample selection. Most researchers employ these tests to determine whether sample selection is present. If the correlation across the equations' residuals is zero, they typically fail to reject the null hypothesis that there is no sample selection and one would estimate and report two independent probit models. Alternatively, one would report the results of the sample selection or censored probit model.

Several hypothesis tests can be used to determine whether $\rho = 0$ in the sample selection model. These include a series of asymptotically equivalent χ^2 tests for maximum likelihood models: Wald, likelihood ratio, Lagrangean multiplier, and conditional moment tests. For each of these tests, we are interested in both their size and their power. The size refers to the observed p-value or level of significance of the test under the true null hypothesis (Type I error—or incorrectly rejecting a true null hypothesis), while the power is the tests' ability to discriminate alternative hypotheses (Type II error—or incorrectly failing to reject a null hypothesis). A test should have both, good size and power, since for the sample selection problem one wants to know both: whether sample selection is present (size) and the degree of correlation across the two equations (power).

To summarize the size and the power of the tests, we constructed a set of size-adjusted power curves for each experiment and test. A size-adjusted power curve plots the estimated size of the hypothesis test for sample selection against the power of the test.[11] The axes of the plot range from 0 to 1, since both the size and power are measured as probabilities.[12] The x-axis of each plot displays the *actual size* of the tests under the assumption that the null hypothesis were true based on the Monte Carlo results. The y-axis of each plot presents the power, computed for the actual size of the hypothesis test from the Monte Carlo results. A test with a good size-power curve would follow the left and upper edge of the plot. Poor size-power curves will be nearer the 45 degree line.

Figure 1 presents the plots of the size-adjusted power of the hypothesis tests for the alternative hypothesis that $\rho \neq 0$ for $N = 1000$, for various alternative values of ρ, and for the well and weakly identified models outlined above.

[Figure 1 about here.]

The results demonstrate the variation in the size and power of the test statistics for different identification assumptions. For the well identified model results in the top row of Figure 1, the size-power curves look as expected and indicate a hypothesis test with correct size and power. The size and power of the identified model only slightly decrease for lower values of ρ. But the performance of the tests for the weakly identified models in the second row of Figure 1 is poor. In the case of $|\rho| < 0.8$, the size-power curves are very close to the 45-degree line. These findings imply that unless a selection model is well identified, the tests (that $\rho = 0$) for sample selection are not robust. For weakly identified models, the tests have incorrect size and too often lead to the conclusion that the null hypothesis of no sample selection is true. Moreover, the tests will have low power, meaning that one is too likely to make a type II error and conclude that there is no sample selection, when in fact it is present. Both are problematic for data analysis because one may incorrectly choose either a sample selection or a single equation model.

If hypothesis tests for sample selection perform poorly for weakly identified models, what then is the 'cost' or consequence of using a sample selection versus a single equation model for the outcome variable? That is, even if the tests for sample selection are not robust, the estimates of the selection model may still be superior (in a mean squared error sense) to those from *any* single equation or non-sample selection model. In this case, the incorrect test results would be irrelevant. To examine the implications of choosing and specifying a selection model, we inspect the joint probability density of the coefficients in the outcome equation for the well and weakly identified models. Figure 2 plots the density contours for the coefficients of the outcome equation for the censored probit model and a probit model. Good estimates of these parameters should be centered on the true value (the solid dot).[13]

[Figure 2 about here.]

As figure 2 shows, the identification of the model plays a major role in estimating the true parameter values. For the well identified models in the first row of the figure, the censored probit model with two exclusion restrictions does a good job recovering the true parameter with a joint density that is centered over the true values. In the second row, however, the weakly identified censored probit and probit models do a very poor job since the densities are not centered over the true value.

In political science applications, these results have two substantive implications. First, the gains from using sample selection models critically depend on the identification assumptions in the model. If identification is weak, then the resulting parameters are biased and inefficient. Second, the standard evidences to justify a sample selection model—a significant estimate of ρ and the presence of changes in the coefficients from the single equation outcome model to the two equation sample selection model—may be artifacts of a weakly identified sample selection model.

Moreover, as Figure 2 shows, specifying a sample selection model may lead to problems that are worse than the actual sample selection problem, since this procedure may produce estimates that are biased in the opposite direction of the estimates of a single equation model. Suppose that the true value of one of the regressors was 0 instead of 0.5. For some of the estimates, the selection model would then produce outcome equation coefficients that were negative and the single equation model for the selected sample would produce estimates that were positive. Both are wrong and lead to the incorrect inference that the regressor is significant. When there is weaker identification, the parameter estimates become more biased *even after accounting for sample selection effects*.

Recommendations: When and How to Use Selection Models

What can be done to weigh the costs and benefits of sample selection models in political science applications? In principle, researchers should try different identification assumptions for their models and analyze the robustness of the results for the key variables. This, however, is difficult because identification involves specifying the observable relationships among the variables and parameters in the model.

In this section, we offer guidance to researchers who consider modeling selection processes. Most importantly, better theory is needed to improve the specification and estimation of selection models. As such, political scientists should outline the theory and rationale for their specification choices and identification more clearly.[14] Researchers also need to look at the specification of alternative models including detailed comparisons of the results of censored probit and independent probit models for data with possible sample selection. This consideration of the specification of the selection model should *not* be solely determined by the results of a hypothesis test for sample selection, since these tests may have poor size-power properties.

Finally, scholars should diagnose the possible fragility of their censored probit model estimates. That is, they need to assess the impact of sample selection and model specification when the regressors in the model may be (highly) correlated across the equations. We recommend researchers to characterize the densities of the parameters in the model and see how much they differ from a model with sample selection to a model without sample selection. The standard method of reporting results for sample selection models is based on reporting the mean coefficient estimates and their standard errors. However, this does not fully account for the difference of the estimates across the two models or the costs of the two estimation methods.[15] Instead, characterizing the parameter densities ensures a better understanding of how the parameters differ across the two model specifications and how the identification assumptions affect the inferences.

Two methods appear appropriate for this task: one frequentist and the other Bayesian. From a frequentist perspective, a parametric bootstrap can assess the relationship between the theoretical density of the parameters and the actual density in a given sample. Maximum likelihood estimates are (asymptotically) normally distributed and equal to their population values. This may not hold if the selection model is incorrectly specified. A bootstrap of the censored probit model evaluates whether this is the case. The properties of the estimates are simply determined by re-sampling additional data sets (Efron and Tibshirani 1994). The bootstrap re-sampling of the data treats the exogenous variables in the model as fixed and then re-samples the residuals to generate new values of the dependent variable(s). For each of these re-sampled data sets new estimates are then computed and summarized graphically. This re-sampling method depicts how the parameter variation is a function of the underlying data variation. The bootstrap procedure can be easily implemented in existing software.[16] Unfortunately, the approach depends on the specification of the censored probit model. Any specification errors, such as weak identification in the model, are reflected in the bootstrapped sample. In the bootstrapped results, if there is poor identification one should expect large variances or flat densities for the estimated parameters.

From a Bayesian perspective, Markov chain Monte Carlo (MCMC) estimation methods produce samples of the posterior distribution of the parameters. Accordingly, the researcher can examine the posterior densities of the censored probit model coefficients accounting for the uncertainty of the model specification and the parameter uncertainty. Models that are weakly identified will exhibit parameter densities that are flat or have large variances. A Bayesian censored probit model requires a proper (but diffuse) prior to ensure that the posterior is well defined (since we may have a likelihood that is weakly identified). The Markov chain Monte Carlo method employs a marginalization of the model and samples from each of the conditional distributions of the model parameters to generate the joint posterior distribution of the parameters.[17] Both approaches allow us to assess the sensitivity of the censored probit estimates. In addition, it is possible to compute empirical confidence intervals (or posterior density regions in the Bayesian case) that will measure the 'true' sample uncertainty of the model estimates. The sensitivity of the model specifications may then be assessed.

As a summary, we propose four steps researchers may follow if they aim at specifying a censored probit model:

1. *Theory and Identification*: Theory matters! If the theory predicts that only a sub-sample of observations will be observed, then the theory must differentiate selection from the outcomes.

2. *Choice of Specifications*: Outline explicitly the theoretical and methodological rationale for the choice of specifying a selection model.

    (a) If selection is expected theoretically and empirically, specify a censored probit model.

    (b) If the underlying theoretical model incorporates a strategic interaction, specify a strategic model such as proposed by Signorino and Tarar (2006).

    (c) If no selection is expected theoretically and empirically, specify an independent probit model.

3. *Estimation*: Estimate multiple identified models. If estimating a sample selection model, also estimate an independent equations version of the model and compare the estimates.

4. *Diagnosis and Testing*: Employ a bootstrap or the Bayesian method to evaluate the estimated model against the alternative specification. Characterize the densities of the parameters in the model and evaluate how much the results differ from a selection model to a non-selection model. This incorporates not only detection of contradictory coefficients, but also large changes in the confidence regions for both the censored and non-censored models.

Example: Conflict Onset and Escalation

This section replicates and extends Reed's (2000) analysis of conflict onset and escalation in the international system, illustrating the fragility of his selection model inferences from weak identification. Reed claims that the onset of international conflict is endogenously related to the escalation of conflicts to war. Conflict must onset (selection) before it can escalate to war (the final outcome). Therefore, the causes of conflict escalation are largely unclear—only those nation states where conflict has onset can then escalate to war. A failure to model the selection of states into conflict onset accordingly leads to incorrect inferences about the likelihood of conflict escalation.

Reed's analysis includes a total of 20990 dyad-years from 1945-1985 of which 947 experience conflict onset. The binary outcomes onset and escalation are functions of the same six variables: power parity, joint democracy, joint satisfaction, alliance, development, and interdependence. His onset (selection) equation model also includes 33 peace year dummies to account for the temporal dependency of onset.[18]

Formally, the set of exclusion restrictions that identify the model—variables included in the conflict onset (selection) equation that are excluded from the conflict escalation (outcome) equation—are the variables measuring the number of years that each of the dyads in the analysis was at peace. All other variables are included in both stages. However, while controlling for the duration of peace in a dyad seems theoretically appealing, it does not really serve as identifying information in the model. Knowing that a dyad was at peace for some number of years is the same information coded into the dependent variable for the onset (selection) equation. If there has been onset, then the peace-year dummy will be the same as the year in which onset occurs in the dyad. Thus, the peace-year dummies are perfect predictors of onset and may have little possible variation that can be used to predict onset and therefore escalation.[19]

Accordingly, we expect that the selection equation is weakly identified and the results should be—consistent with our earlier Monte Carlo findings—relatively fragile.[20] Table 1 presents the relevant coefficient estimates and confidence regions for each model. For the maximum likelihood (bootstrapped and Bayesian) models, we report the mean (median) estimate of the coefficient and the 95% (posterior) confidence regions.[21]

First note, the single equation models in columns 1-3 differ greatly from those in columns 4-6. That is, correcting for sample selection would appear to be the correct route to pursue. Substantively, the power parity variable is insignificant in the independent probit models using the selected sample. Joint democracy depresses escalation in the single equation probit models using the censored sample.

More importantly, the columns 4-6 present the censored probit results that replicate and extend Reed's analysis. The bootstrap and the Bayesian censored probit results (columns 5 and 6) are quite similar and show that the estimates have larger 95% confidence regions than the maximum likelihood results in column 4. Of interest are the different findings for the two key variables: power parity and joint democracy. The bootstrap and Bayesian censored probit estimates show more variance and are less certain than the maximum likelihood estimates. Moreover, the Bayesian estimate of ρ for the censored probit model has a confidence region that includes zero. Finally, the inferences about the effects of power parity and joint democracy on *onset* are not the same as in the maximum likelihood model and the Bayesian model—the confidence regions of both coefficients include zero. Substantively, this means that the effects of these variables on onset may be zero and that the earlier results on the effects of these variables on onset may be wrong.

As we noted earlier, point estimates and confidence regions are only one way of summarizing the coefficient estimates. Figure 3 presents the densities for the Bayesian and bootstrapped censored probit models and the Bayesian single equation probit model for the censored sample. The left (right) column presents the posterior densities for the onset (escalation) equation.

[Figure 3 about here.]

The bootstrapped and Bayesian results differ dramatically from the maximum likelihood estimates reported by Reed. First, Reed hypothesizes that power parity increases the likelihood of onset, but may not affect escalation to war. The Bayesian and bootstrapped censored probit results are to the contrary: power parity has a null effect on both the likelihood of onset and escalation.

	Probit Full Sample	Probit Censored Sample	Bayesian Probit Censored Sample	ML Censored Probit	Bootstrap Censored Probit	Bayesian Censored Probit
Outcome: Escalation						
Intercept	-2.15	-0.54	-0.55	0.65	0.69	-0.38
	(-2.22, -2.09)	(-0.65, -0.43)	(-0.66, -0.44)	(0.49, 0.80)	(0.49, 1.03)	(-2.02, 0.50)
Power Parity	0.46	-0.09	-0.14	-0.33	-0.37	-0.24
	(0.17, 0.74)	(-0.51, 0.34)	(-0.57, 0.26)	(-0.67, 0.00)	(-0.83, -0.03)	(-2.26, 1.59)
Joint Democracy	-1.16	-1.27	-1.35	-0.30	-0.36	-2.48
	(-1.71, -0.61)	(-2.14, -0.42)	(-2.42, -0.58)	(-0.86, 0.26)	(-21.77, 0.33)	(-7.54, 1.16)
Joint Satisfaction	-0.65	-0.58	-0.65	-0.05	-0.11	-3.33
	(-1.08, -0.22)	(-1.20, 0.04)	(-1.35, -0.07)	(-0.49, 0.39)	(-14.49, 0.43)	(-13.87, 0.87)
Alliance	-0.48	-0.86	-0.88	-0.64	-0.63	-1.44
	(-0.70, -0.25)	(-1.19, -0.54)	(-1.21, -0.56)	(-0.90, -0.37)	(-1.02, -0.37)	(-5.40, 1.23)
Development	0.04	0.06	0.06	0.05	0.05	0.07
	(0.02, 0.06)	(0.03, 0.08)	(0.03, 0.08)	(0.03, 0.07)	(0.02, 0.08)	(-0.18, 0.33)
Interdependence	-34.80	-34.50	-6.24	-3.88	-9.12	-2.66
	(-74.52, 4.92)	(-91.23, 22.23)	(-23.50, 9.93)	(-26.88, 19.11)	(-2115.78, 49.89)	(-26.82, 15.71)
Selection: Onset						
Intercept				-0.48	-0.48	-0.44
				(-0.55, -0.42)	(-0.56, -0.42)	(-0.84, -0.04)
Power Parity				0.36	0.36	0.37
				(0.19, 0.52)	(0.18, 0.51)	(-0.10, 0.86)
Joint Democracy				-0.61	-0.61	-1.01
				(-0.74, -0.48)	(-0.76, -0.48)	(-3.75, 0.18)
Joint Satisfaction				-0.16	-0.17	-0.24
				(-0.29, -0.04)	(-0.30, -0.04)	(-0.92, 0.13)
Alliance				0.04	0.04	0.03
				(-0.06, 0.14)	(-0.06, 0.15)	(-0.43, 0.43)
Development				-0.01	-0.01	-0.01
				(-0.02, -0.00)	(-0.02, 0.00)	(-0.20, 0.17)
Interdependence				-1.43	-1.59	-3.56
				(-8.12, 5.26)	(-9.72, 3.74)	(-18.87, 2.18)
ρ				-0.77	-0.80	-0.20
				(-0.84, -0.68)	(-0.91, -0.70)	(-0.69, 0.14)

Table 1: Maximum Likelihood Results, Bootstrap Estimates, and MCMC Parameter Densities for the Censored Probit Model for Reed's conflict onset and escalation model. Bootstrap and Bayesian coefficients are median estimates. Intervals in parentheses are 95% confidence intervals for ML and bootstrap results (computed by percentile method) and 95% posterior regions for Bayesian results. Uncensored (censored) sample size is 20990 (947).

Furthermore, Reed argues and finds that joint democracy lowers the probability of onset but has a minimal effect on escalation. The results are similar, but in the bootstrap and Bayesian estimates joint democracy depresses the probability of escalation to a greater degree. These results are also consistent with those from the independent probit models. It would appear that part of the democratic peace reduces escalation as well. Furthermore, in Reed's results, joint satisfaction with the status quo and joint democracy are not significant predictors of escalation. In contrast, the Bayesian findings offer evidence that both factors lower the probability of escalation.[22]

In summary, the replication and extension of Reed's analysis of war onset and escalation highlights two of our earlier claims about the application of the censored probit model (and sample selection models in general). First, regardless of the sample size, the correlation in the exogenous variables across equations matters. The more variables the two equations have in common, the less information exists for the identification of the system of equations. This means that maximum likelihood coefficient estimates—particularly for the outcome equation—will be too confident. In some cases the increased uncertainty from the poor instruments used for identification of the sample selection process may even swamp the correction of bias by the sample selection model. Secondly, users of selection models need to analyze the differences across models with and without sample selection. One should generally be skeptical about hypothesis tests reported by standard statistical software. These selection tests assume that the model specification is correct, the model is well identified, and there is sufficient information to estimate the selection and outcome effects. If any of these assumptions is violated or compromised, one should suspect that the hypothesis test for selection may be too good to be true.

Conclusion

Political scientists have become increasingly aware of the problems that arise if their sample is censored. Since Reed's work on conflict onset and escalation, scholars frequently apply selection models to various fields in political science, and especially in international relations. In this paper, we have argued that those models are often very poorly identified. This eventually leads to wrong inferences about the social phenomena of interest. The higher the correlation in the exogenous variables across the equations the more likely researchers may draw incorrect inferences not only about whether selection is present, but also about the predominant effects of their key explanatory variables.

To evaluate the properties of the tests for sample selection in censored probit models and the costs of employing the "wrong" model, we utilized a Monte Carlo experiment comparing the performance of different identification decisions. Our findings demonstrate that the tests for sample selection have poor size and power if identification is weak and if the correlation of regressors across equations is high. The higher the correlation of the errors across the two equations and the regressors, the more biased the estimated marginal effects and the more likely one is to incorrectly conclude that selection is present. The costs that come along with this misconception are high. As the Monte Carlo results indicate, the estimates of a weakly identified selection model could be even more biased than the estimates of a single equation probit model. Those problems were further elucidated in our replication of Reed's analysis of conflict onset and war.

Based on those findings, we provided recommendations for researchers who estimate selection models. Scholars should consider two questions. First, is selection likely to be present? Second, is there sufficient structure and information to strongly or weakly identify the selection and outcome processes, separately? If selection is believed to be present in the data, then the quality of the inferences depends on the identification of the selection and outcome equations. One cannot just say that there "may be selection", since sample selection is not just a data problem. It poses a theoretical problem that affects the specification of the empirical model.

Researchers employing and reviewers calling for selection models should then be prepared to substantiate why there may be selection based on theory. At the same time, calls for selection models should also include a clear specification and identification of the selection process. Our paper does not call for a general abandonment of selection models in political science. However, if the structure of the specified selection model is fragile—as determined by the suggested bootstrap and Bayesian methods—then selection effects should be interpreted skeptically if at all.

Notes

[1]This is the case when one can observe the covariates that influence selection and the outcome of interest, but not necessarily the outcome and stands in contrast to a *truncation* problem where the covariates and the outcome of interest are unobserved.

[2]See also Schultz (1999) for a more critical treatment and Hansen (1990) for an early application.

[3]These counts are not constrained to an analysis of war onset and escalation. Most articles are published in the *Journal of Conflict Resolution*, the *American Journal of Political Science*, *International Organization*, *International Interactions*, and the *Journal of Politics*.

[4]Note, however, that the substantial results also hold for the traditional Heckman selection models.

[5]As noted by Meng and Schmidt (1985) and Poirier (1980), the standard identification conditions for ML models are only necessary, but not sufficient in this case. Depending on the values of the parameters in the two equations further restrictions on the coefficients and parameters may need to be satisfied. For instance, it must be the case that $\beta_1 \neq \rho\beta_2$, so that the labeling of the two equations is unique, and that there is enough variation in some of the variables.

[6]Note, when ρ is known, the censored probit model is identified even if the regressors are not perfectly collinear across the two equations—this is the genesis of the Sartori estimator (Sartori 2003).

[7]The distinction here is that theoretical identification of the equations differs from the formal statistical identification of the equations. The latter may be possible, but not the former, since one may be able to estimate a model, but it may not be behaviorally interpretable in the manner in which it is estimated (e.g., Leeper, Sims and Zha 1996).

[8]Moreover, if a variable is included in the model "only to ensure that there is some identification", and if it is a poor instrument, then the resulting estimates are likely weakly or not identified.

[9]Note, the results presented in this section are relatively robust to different data generating processes. We have, for example, investigated additional models and experiments that look at different degrees of correlation among the regressors in the selection and outcome equations and different sample sizes (from 50 to 10000). The findings are consistent with those reported here. Results are available upon request.

[10]The uncensored sample has 1000 observations. Based on the chosen parameters, the selected sample has approximately 770 observations. A 'successful' outcome is observed in about 56% of the observations. The discussion below does not directly address the impact of the degree of selection in the first stage on the properties of the tests and estimates. However, standard results on sample selection indicate that these problems worsen if sample selection is worse. Our results thus serve as a benchmark case.

[11]The presented curves are size-adjusted in that one must specify the (nominal) size of a test to compute its power. Rather than assuming that these nominal sizes are the same as the actual sizes of the tests in the Monte Carlo results

(they are not), we estimate the actual sizes of the hypothesis test statistics and use these observed or actual sizes.

[12]The curves are computed using the method described in Davidson and MacKinnon (1998).

[13]One could criticize these Monte Carlo experiment results as an artifact of the data generation process and the fact that we estimate the exact model we used to generate the data. This is incorrect because the alternative would be to generate data from a identified model and estimate an unidentified model—which is not possible. Alternatively, we could generate data from a model with additional exclusion restrictions (say two per equation) and estimate a model with only one exclusion restriction. This would not tell us much, since we know that in this case the omission of a true regressor or exclusion restriction in the model would produce omitted variable biases. What we are most interested in here is the robustness of the sample selection model to recover weakly identified data generation processes like those specified here, since this is the case we suspect that most researchers face.

[14]In this sense, the work of Signorino (1999, 2002) and Sartori (2003) is an attempt to marry better theory and identification assumptions. See also the discussion in Huth and Allee (2004) where these sample selection issues are discussed in international relations. We would however caution that their recommendation to use sample selection models liberally should be tempered by the results presented here.

[15]For a Bayesian presentation of this idea see Gill (2002).

[16]Stata for instance has a bootstrap command that can be used for the censored probit model.

[17]See Appendix for the Gibbs sampling algorithm for sampling the posterior distribution of the parameters of interest in a censored probit model.

[18]Details of these variables and their coding can be found in Reed (2000, 88–90). We do not report the coefficients for the dummy variables. They are available upon request.

[19]This is a well known result: binary outcomes over time can be coded as events or as count processes. While typically seen with dependent variables (logit versus event history modeling), here is appears in the coding of the independent and dependent variables.

[20]Note that this model is not identical to the one discussed by Sartori (2003) since the system of equations for the selection and outcome equations is identified by the inclusion of temporal dummy variables, which are not included in the escalation equation.

[21]The estimates are based on the MCMC algorithm described in the Appendix for the censored probit model. The results are based on a final set of 200000 iterations, with a diffuse prior centered on zero and a variance of 100 for each parameter. The bootstrap results are based on 10000 samples.

[22]Like Reed, we find that alliances have a significant negative effect on escalation, but no effect on onset and we also see that economic interdependence does not affect escalation. Yet, the densities and confidence regions for the effect of interdependence on onset are different. Most of the density estimate is skewed negative, indicating that dependence may reduce the likelihood of onset, thus supporting the idea of a "liberal peace."

Appendix: A Bayesian Sampling Algorithm for the Censored Probit Model

The following procedure describes the basic Gibbs sampling algorithm with a Metropolis-Hastings step to generate a sample from the posterior distribution of the parameters of the censored probit model.

1. Initialize the starting values for the sampling with parameters for $(\beta_1, \beta_2, \rho) = (\beta_1^0, \beta_2^0, \rho^0)$. We use the estimates from two independent probit models as the initial values.

2. Sample $(y_{i1}^*, y_{i2}^*|\beta_1, \beta_2, \rho)$ using data augmentation. This sample is from a truncated bivariate normal distribution, $TN(x_i\beta_1, x_2\beta_2, \rho)$ with truncation points defined by the orthants of R^2 corresponding to the signs of the latent variables for each observation (For details, see Chib and Greenberg 1993, Chib and Greenberg 1998).[23]

3. Sample $\beta_1, \beta_2|y_{i1}^*, y_{i2}^*, \rho$ from a multivariate normal distribution with a mean and variance computed by a multivariate generalized least squares regression for the latent variables y^*. Note that ρ is assumed fixed.

4. Sample $\rho|y_{i1}^*, y_{i2}^*, \beta_1, \beta_2$ using a Metropolis-Hastings step. This is done by sampling ρ from a candidate distribution that is truncated normal $TN_{(-1,1)}(\hat{\rho}, (1 - \hat{\rho^2})^2/n_1)$, where $\hat{\rho}$ is the estimated correlation in the residuals at the $i'th$ iteration for the n_1 observations in the selected sample for the outcome equation and the selection equation. This proposal density for ρ is suggested in Chib and Greenberg (1998).

5. Repeat steps 2-4 $B + G$ times. Discard the first B iterations, the burn-in of the sampling algorithm to eliminate dependence on the initial values.

References

Berinsky, Adam J. 1999. "The Two Faces of Public Opinion." *American Journal of Political Science* 43(4):1209–1230.

Boehmke, Frederick J. 2003. "Using Auxiliary Data to Estimate Selection Bias Models, with an Application to Interest Group Use of the Direct Initiative Process." *Political Analysis* 11(3):234.

Chib, Albert and Edward Greenberg. 1993. "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association* 88:669–679.

Chib, Siddartha and Edward Greenberg. 1998. "Analysis of Multivariate Probit Models." *Biometrika* 85(2):347–361.

Davidson, Russell and James G. MacKinnon. 1998. "Graphical methods for investigating the size and power of test statistics." *The Manchester School* 66:1–26.

Dubin, James and Douglas Rivers. 1989. "Selection Bias in Linear Regression, Logit and Probit Models." *Sociological Methods and Research* 18(2 and 3):360–390.

Efron, Bradley and Robert J. Tibshirani. 1994. *An Introduction to the Bootstrap*. Chapman and Hall/CRC.

Fordham, Benjamin O. and Timothy J. McKeown. 2003. "Selection and Influence: Interest Groups and Congressional Voting on Trade Policy." *International Organization* 57(3):519–549.

Gerber, Elisabeth R., Kristin Kanthak and Rebecca Morton. 1999. "Selection Bias in a Model of Candidate Entry Decision." Society for Political Methodology Working Paper.

Gill, Jeffrey. 2002. *Bayesian Methods: A Social and Behavioral Sciences Approach*. Boca Raton: Chapman and Hall.

Greene, William H. 2002. *Econometric Analysis*. Prentice Hall.

Hansen, John M. 1990. "Taxation and the Political Economy of the Tariff." *International Organization* 44(4):527–551.

Heckman, James J. 1976. "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and A Simple Estimator for Such Models." *Annals of Economic and Social Measurement* 5:475–492.

Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47:153–161.

Hug, Simon. 2003. "Selection Bias in Comparative Research: The Case of Incomplete Datasets." *Political Analysis* 11:255–274.

Huth, Paul and Todd Allee. 2004. Research Design in Testing Theories of International Conflict. In *Models, Numbers & Cases: Methods for Studying International Relations*, ed. Detlef Sprinz and Yael Wolinsky-Nahmias. Ann Arbor: University of Michigan Press chapter 9, pp. 193–226.

Jensen, Nathan M. 2003. "Democratic Governance and Multinational Corporations: Political Regimes and Inflows of Foreign Direct Investment." *International Organization* 57(3):587–616.

Leeper, Eric M., Christopher A. Sims and Tao Zha. 1996. "What Does Monetary Policy Do?" *Brookings Papers on Economic Activity* 1996(2):1–63.

Lemke, Douglas and William Reed. 2001. "The Relevance of Politically Relevant Dyads." *Journal of Conflict Resolution* 45(1):126–144.

Manski, Charles. 1995. *Identification Problems in the Social Sciences*. Cambridge: Harvard University Press.

Meernik, James. 2001. "Domestic Politics and the Political Use of Military Force by the United States." *Political Research Quarterly* 54(4):889–904.

Meng, Chun-Lo and Peter Schmidt. 1985. "On the Costs of Partial Observability in the Bivariate Probit Model." *International Economic Review* 26:71–86.

Poirier, Dale J. 1980. "Partial Observability in Bivariate Probit Models." *Journal of Econometrics* 12:209–217.

Reed, William. 2000. "A Unified Statistical Model of Conflict Onset and Escalation." *American Journal of Political Science* 44(1):84–93.

Reed, William and David H. Clark. 2000. "War Initiators and War Winners: The Consequences of Linking Theories of Democratic War Success." *Journal of Conflict Resolution* 44(3):378–395.

Sartori, Anne. 2003. "An Estimator for Some Binary-Outcome Selection Models Without Exclusion Restrictions." *Political Analysis* 11(2):111–138.

Schultz, Kenneth A. 1999. "Do Democratic Institutions Constrain or Inform? Contrasting Two Institutional Perspectives on Democracy and War." *International Organization* 53(2):233–266.

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

Sweeney, Kevin and Paul Fritz. 2004. "Jumping on the Bandwagon: An Interest-Based Explanation for Great Power Alliances." *The Journal of Politics* 66(2):428–449.

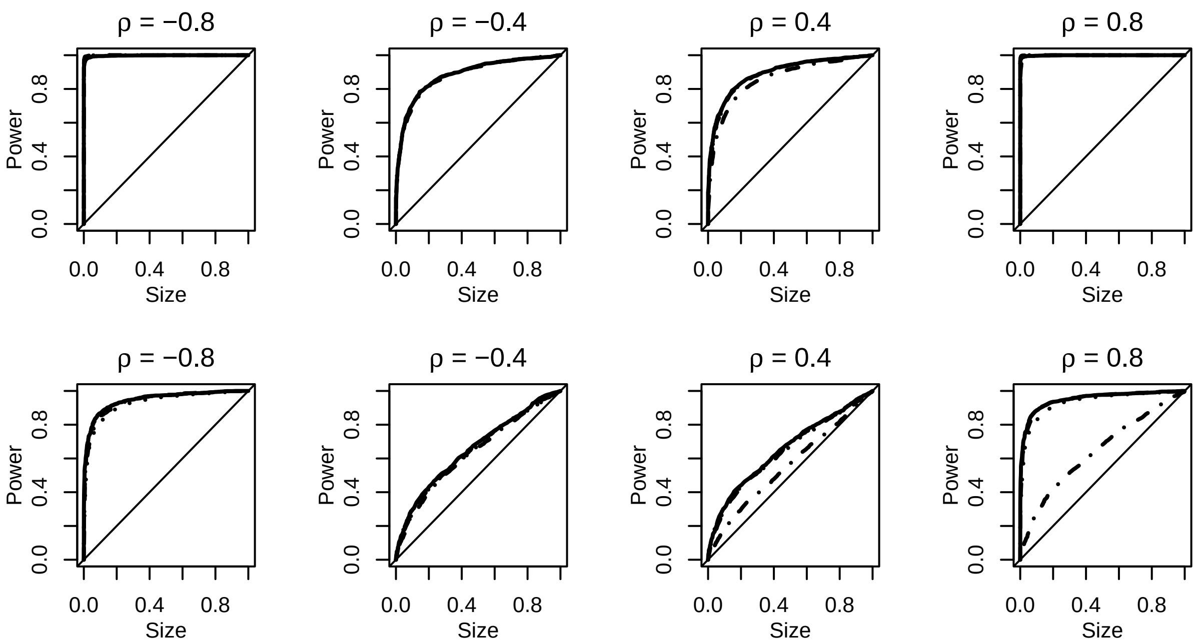

Figure 1: Empirical Size-Adjusted Power for $N = 1000$: Row 1 are the size-power curves for the test statistics for the well identified model. Row 2 are the size-power curves for the tests statistics for the weakly identified model. Wald test is the solid line, Likelihood Ratio test is the dashed line, Lagrangean Multiplier test is the dotted line, and the Conditional Moment test is the dot-dash line.

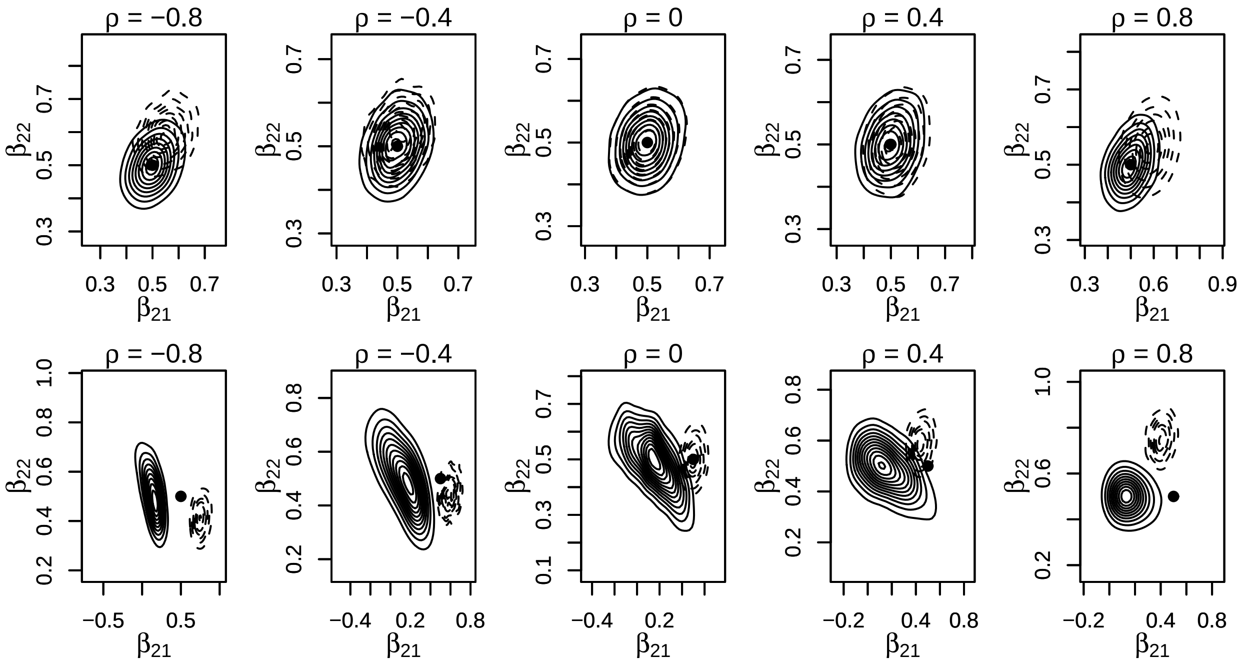

Figure 2: Density contours for well and weakly identified censored probit models. Solid contours are the densities for the censored probit estimates; dashed contours are the densities for the independent probit model coefficients. Top row are the results for the well identified models and the bottom row are the weakly identified model.

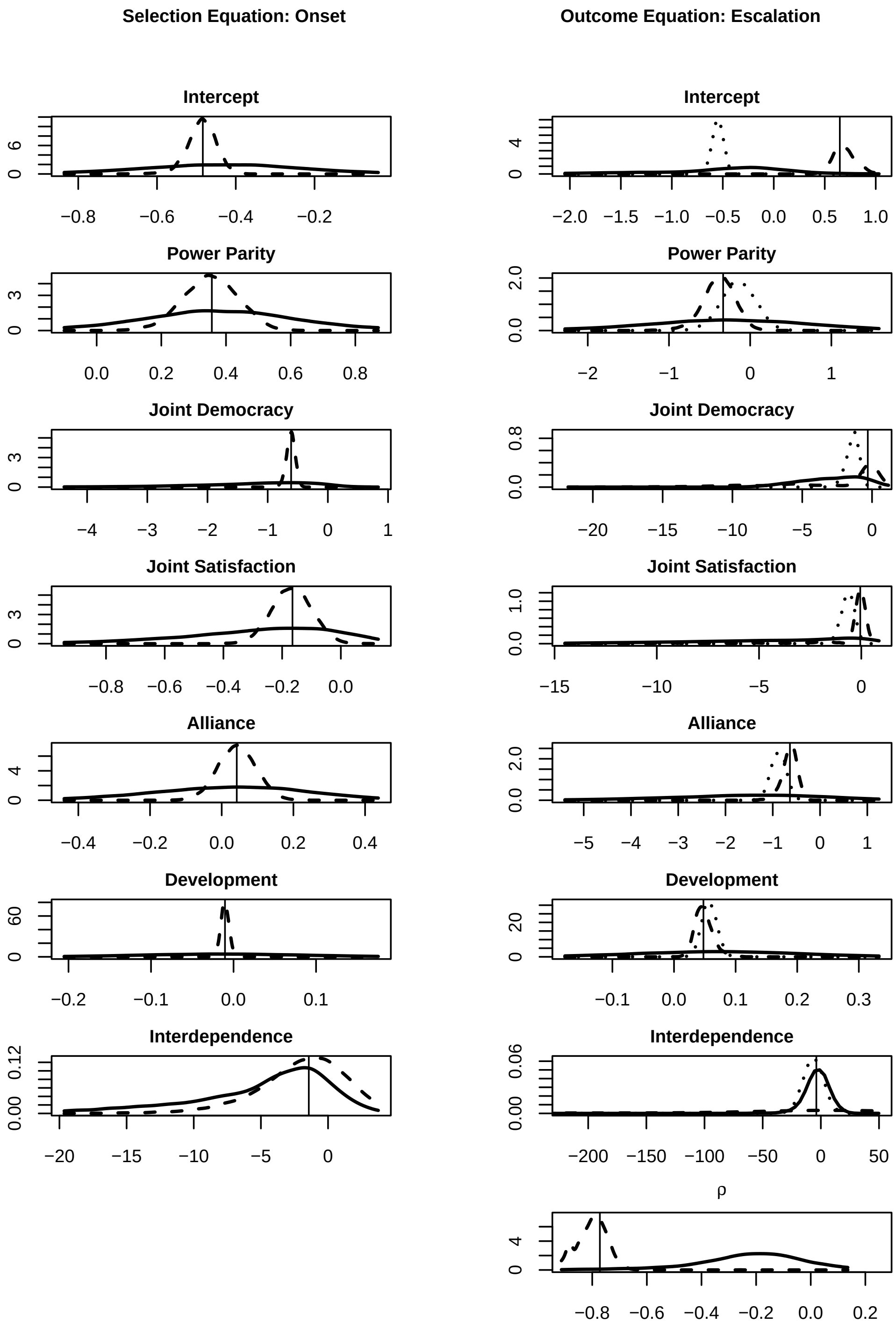

Figure 3: Bayesian Censored Probit, Bootstrapped Censored Probit, and Bayesian Probit Results for Reed's (2000) Model. Solid densities are based on a Bayesian censored probit model. Dashed densities are from the bootstrapped censored probit model and dotted densities are from the Bayesian probit model. Coefficients for peace-years dummies in the selection equation are not presented. Vertical lines are the maximum likelihood estimates.