Political Analysis (2003) 11:111–138
DOI: 10.1093/pan/mpg001

An Estimator for Some Binary-Outcome Selection Models Without Exclusion Restrictions

Anne E. Sartori
Department of Politics, Princeton University,
Corwin Hall, Princeton, NJ 08544-1012
e-mail: asartori@princeton.edu

This article provides a new maximum-likelihood estimator for selection models with dichotomous dependent variables when identical factors affect the selection equation and the equation of interest. Such situations arise naturally in game-theoretic models where selection is typically nonrandom and identical explanatory variables influence all decisions under investigation. When identical explanatory variables influence selection and a subsequent outcome of interest, the commonly used Heckman-type estimators identify from distributional assumptions about the residuals alone. When its own identifying assumption is reasonable, the new estimator allows the researcher to avoid the painful choice between identifying from distributional assumptions alone and adding a theoretically unjustified variable to the selection equation in a mistaken attempt to "boost" identification. The article uses Monte Carlo methods to compare the small-sample properties of the estimator with those of the Heckman-type estimator and ordinary probit.

1 Introduction

Many of the most interesting political phenomena are ones for which the sample is nonrandomly selected. Actors' choices or circumstances determine whether or not they are observed going to war, casting a vote, or choosing a new form of government after a civil war. Standard regression techniques such as ordinary least squares and logit/probit yield inaccurate estimates if some included variable and some omitted variable affect both selection into the sample and the subsequent political outcome of interest. The problem is more serious than many researchers realize: the results obtained with standard techniques are not even accurate estimates of the effects of the independent variables, conditional on a case being in the sample. A growing number of works in political science therefore use statistical procedures designed to avoid selection bias (Brehm 1993; Mitchell et al. 1997; Berinsky 1999; Reed 2000; Boehmke 2003).

Author's note: I am grateful to Chris Achen, Larry Bartels, John Brehm, Gary King, Jeff Lewis, Elie Tamer, Jas Sekhon, anonymous reviewers, and the audience at the Society for Political Methodology 2001 Summer Meeting for helpful comments, and to Bo Honoré for important references. I also thank Shigeo Hirano for his excellent assistance in making user-friendly the computer program that implements the estimator that this article discusses. Finally, I thank CBRSS at Harvard University for support during the year in which I finished this article. Any errors are, of course, my own.

The commonly used Heckman-type selection models (Heckman 1974, 1976, 1979; Achen 1986; Van de Ven and Van Praag 1981; Dubin and Rivers 1990) are appropriate only when at least one "extra" explanatory factor influences selection but not the subsequent outcome of interest (Achen 1986, p. 99).[1] Unfortunately, this "extra" exogenous variable often does not exist. The Heckman selection models are estimable without the extra variable, but the results are then based only upon the assumptions about distributional assumptions about the residuals rather than upon variation in the explanatory variables.[2] No one who has proposed such an estimator recommends using it with identical explanatory variables in the two equations. Thus, when theory dictates identical explanatory variables in the two equations, researchers are left with an unhappy choice: to dredge up an extra explanatory variable for the selection equation (leading to specification error if the variable does not belong there) or to identify only from distributional assumptions about the residuals.

This article proposes a maximum-likelihood estimator for use with identical explanatory variables and dichotomous dependent variables that is based upon an additional identifying assumption: The error term for an observation is the same in the two equations. When the assumption is reasonable, the estimator frees researchers from the choice between Scylla and Charybdis that I discussed in the previous paragraph. The method that I propose here is easy to execute.

My identifying assumption rarely will be perfectly true, but it is likely to be reasonable exactly when the researcher believes that identical explanatory factors influence selection and the subsequent outcome of interest. More specifically, the assumption is likely to be a close match to reality when three conditions hold: (1) selection and the subsequent outcome of interest involve similar decisions or goals; (2) the decisions have the same causes; and (3) the decisions occur within a short time frame and/or are close to each other geographically.

As an example, the statistical model that I propose would be a likely candidate for modeling states' decisions about whether or not to escalate wars. The situation meets the three conditions: the decision to start a war, or select into the sample, is closely related to the subsequent decision to escalate the conflict or to continue at the current level; the same factors (e.g. the balance of forces) are posited to influence both decisions; and the decisions are likely to occur in the same locations, perhaps within a relatively short time frame.

Of course, researchers sometimes have theoretical reasons to believe that the error terms in the two equations are nearly perfectly correlated (or negatively correlated) even when one or more of the three conditions is not met. In a recent article, Lemke and Reed (2001) make such an argument about enduring rivalry and war among the major powers. They argue that states nonrandomly select into rivalries and that identical explanatory variables affect states' decisions both to become rivals and to go to war. In a companion paper, I reexamine their results using the statistical model that I propose here (Sartori 2002).[3]

Although the need for this estimator is not new, it is growing with the use of game-theoretic modeling in political science. Game-theoretic and other rational-expectations

[1] Van de Ven and Van Praag (1981) and Dubin and Rivers (1990) propose selection estimators for use when both the selection-stage dependent variable and the outcome-stage dependent variable are dichotomous. Because these estimators build closely upon Heckman's work, I refer to these estimators as "Heckman estimators" or "Heckman-type estimators" in this article.

[2] Moreover, depending upon the substantive problem, it may be impossible to recover the structural parameters of the underlying model. See Maddala (1999, pp. 231–232).

[3] Lemke and Reed state no theoretical expectations about the direction of the correlation between the error terms in the selection equation and the outcome equation, but they estimate the correlation as highly negative. In the present article, I discuss only a version of my estimator that assumes a correlation of $+1$. My software includes an option for a correlation of -1. However, as I discuss in my companion paper, I believe that theory points to a highly positive correlation in Lemke and Reed's case.

models often imply that the same factors influence both selection and a later outcome of interest. Explanatory variables usually enter a game-theoretic model through the payoff functions, so that, in many models, identical explanatory variables influence every decision made by the actors in the model. In addition, many game-theoretic models represent a series of binary choices. Thus, the statistical model that I propose here is appropriate for testing implications of many game-theoretic models, though it also is useful for testing a broader class of models that involve nonrandom selection and identical explanatory variables.

My estimator is an alternative to two procedures previously proposed by political scientists for testing game-theoretic models (Smith 1998; Signorino 1999, 2002). My method incorporates a different view of the purpose of a formal model than does Signorino's. Signorino's procedure assumes that the formal model is correct and that players "make mistakes according to some known (or assumed) distribution of errors" (Signorino 1999, p. 282). He derives the likelihood function directly from the model and from these assumptions about agent error. In contrast, I see any formal model as a representation of an (important) piece of reality, rather than a complete depiction of the complex world. For example, a crisis bargaining model may focus on the interaction between two players when in actuality many states are involved. Thus, there is no reason to believe that the appropriate statistical model can be derived directly from the formal model. Rather than give up on estimation entirely, I use the formal model only to derive hypotheses, often through comparative statics. I consider the problem of the error structure outside of the formal model and build in the simplest possible error structure that captures basic features of the substantive situation (nonrandom selection and identical explanatory factors). Regardless of the procedure that one uses to test them, the hypotheses derived from a game-theoretic model incorporate strategic interaction because the game-theoretic model itself is strategic.[4]

The article is organized as follows. The next section introduces the problem of selection bias and provides intuition about why one should not use Heckman-type models when the same variables influence selection and the subsequent phenomenon. The third section presents the new estimator. The fourth uses Monte Carlo simulations to compare the small-sample properties of the new estimator to those of probit and of the version of the Heckman estimator that is intended for dichotomous dependent variables. The fifth briefly discusses the comparison paper on enduring rivalries, and the sixth concludes. An appendix presents proofs of consistency and asymptotic normality. STATA software for implementing this technique is available on the *Political Analysis* Web site.

2 Selection Bias and the Disadvantages of the Heckman Estimator when the Independent Variables Are Identical

This section of the article briefly reviews the problem of selection bias and attempts to provide intuition about why the Heckman estimators are poor choices when the researcher believes that identical explanatory variables influence selection and the later equation of interest.[5] Although my proposed estimator is for use with dichotomous dependent variables, this section considers the simpler situation in which the dependent variable of interest

[4] Signorino's statistical procedure attempts to estimate the payoffs of the formal model directly. Such a task requires very strong identifying assumptions (Lewis and Schultz 2003), and it has not been shown that it is possible to estimate all of the parameters of interest separately. My point in the text, however, is about the desirability, rather than the feasibility, of taking the model so literally.

[5] The selection equation also is often of interest. I call the second equation the equation of interest because the purpose of selection modeling is to obtain accurate estimates of the effects of the independent variables in this equation.

is continuous because the issues are more intuitive in this simpler situation. Analogous problems exist when both dependent variables are dichotomous.

2.1 *The Problem of Selection Bias*

In Heckman's oft-cited example, the researcher would like to estimate the effect of education on women's wages (Heckman 1974).[6] The ordinary least squares (OLS) regression technique would be to estimate the effect of the program using the equation

$$y_i = \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i, \tag{1}$$

where i represents a woman in the sample, one variable in $\mathbf{x}_i$ is the woman's education, y_i measures her wages, and ε_i is a normally distributed error term. The selection problem is that the sample consists only of women who choose to work. These women may differ in important unmeasured ways from women who do not work. For example, women who are smarter may be more likely to enter the labor force than less-smart women because the smarter women may anticipate that their intelligence will be recognized on the job, resulting in higher wages.

The selection equation is

$$U_i = \boldsymbol{\gamma}'\mathbf{w}_i + u_i, \tag{2}$$

where U_i represents the utility to woman i of entering the labor force, $\mathbf{w}_i$ is a vector of factors known to influence a woman's decision to work, and u_i, which is jointly normally distributed with ε_i, contains any unmeasured characteristics in equation (2). Instead of U_i, the researcher observes a dichotomous variable Z_i with a value of 1 if the woman enters the labor force (if $U_i > 0$) and 0 otherwise. For simplicity's sake, I assume that education is the only measurable factor that influences the decision of whether or not to work and that education has a positive effect on a woman's desire to work.

Looking at Eqs. (1) and (2), one can see two kinds of selection effects. First, women with high levels of education will be more likely to enter the labor force, and so the researcher will see a sample of educated women. This nonrandom aspect of the sample is what is commonly misunderstood to be the problem of "selection bias," but it alone does not bias estimation of (1). Second, and more importantly, some uneducated women will go to work. These women decide that work is worthwhile because they have a high value on some *unmeasured* variable, part of u_i (in this example, intelligence). Thus, those observations in the sample for the second equation that have small values of the independent variables (more precisely, small values of $\boldsymbol{\gamma}'\mathbf{w}_i$) also have large error terms. Those observations in the selected sample that have larger values of the independent variables have a more usual range of errors. Whether or not education is correlated with the unmeasured intelligence in the overall population, these two variables are correlated in the selected sample. Assuming that intelligence does lead to higher wages, one will underestimate the effect of education on wages because, in the selected sample, women with little education are unusually smart.[7]

More formally, as long as the error term in the second equation is correlated with that in the first (e.g., because intelligence is an unmeasured variable in each equation), the error term for the sample used in the second-stage estimation is not of mean zero and is correlated with the explanatory variable. As Heckman and others have shown, the selection

[6] I simplify and ad lib upon Heckman's example here.

[7] The clearest discussion of this problem that I have seen is in Achen (1986, pp. 73–81), which I draw upon heavily here.

bias problem in a simple linear model is equivalent to an omitted-variable problem. Given that observation i is in the sample,

$$E(y_i) = \boldsymbol{\beta}'\mathbf{x}_i + \theta\left[\frac{\phi(\boldsymbol{\gamma}'\mathbf{w}_i)}{\Phi(\boldsymbol{\gamma}'\mathbf{w}_i)}\right], \tag{3}$$

where ϕ is a standard normal distribution and Φ is a cumulative standard normal distribution (Greene 1993, p. 709). When one estimates equation (1) using OLS, the second term is an omitted variable and the estimates of the βs are inconsistent.

Thus, when selection is nonrandom and error terms are correlated, neglecting to model selection can be a serious mistake. The estimates are inconsistent, even for the group that one observes in the sample. In this example, estimation on a group of working women leads to an inaccurate estimate of how education affects the wages of working women.

2.2 *The Heckman Model Applied to Equations with Identical Explanatory Variables*

Heckman (1979) provides an easy method for obtaining a consistent estimate of $\boldsymbol{\beta}$ when the selection mechanism has a dichotomous outcome (the observation either is selected into the sample or not) and the subsequent equation of interest has a continuous dependent variable. The basic idea is to put the omitted variable illustrated in equation (3) back into the equation. The two-step procedure is as follows:

1. Estimate the probit equation in (2) above. Using the estimated γs, calculate $\frac{\phi(\hat{\boldsymbol{\gamma}}'\mathbf{w}_i)}{\Phi(\hat{\boldsymbol{\gamma}}'\mathbf{w}_i)}$ for each observation i in the selected sample.
2. Estimate $\boldsymbol{\beta}$ and θ by OLS of $\mathbf{y}$ on $\mathbf{x}$ and $\frac{\phi(\hat{\boldsymbol{\gamma}}'\mathbf{w}_i)}{\Phi(\hat{\boldsymbol{\gamma}}'\mathbf{w}_i)}$.

The standard errors of the estimates also must be adjusted for the selection process (Greene 1993, p. 713).

As long as at least one explanatory variable in the selection equation is not in the equation of interest, this technique is a good one. Frequently, the researcher can find such a variable, called an *exclusion restriction*.[8] However, in many substantively interesting situations, all factors that influence selection also influence the subsequent outcome of interest. For example, a variable posited to affect a state's decision about whether or not to begin an international dispute also invariably affects its decisions about whether or not to escalate the dispute toward war.

If all variables influencing selection influence the subsequent outcome of interest, then the Heckman method is of dubious value. To see the problem, consider a case with one explanatory variable. Then,

$$E(y_i) = \beta x_i + \theta\left[\frac{\phi(\gamma x_i)}{\Phi(\gamma x_i)}\right] + u_i. \tag{4}$$

Without the "extra" variable, the Heckman procedure faces the difficult task of estimating the effect of both x and a simple function of the same x on the dependent variable. It can do so, but the resulting estimates are not very good in small samples, as I show later in the article.

[8] Achen (1986, p. 99) gives the example of college admissions. The extra variable is whether or not a student's parents attended a certain college, being a "legacy" affects admissions (selection) but not performance in college (the subsequent dependent variable of interest).

Many interesting political phenomena with nonrandom selection are measured as dichotomous variables. For example, citizens may or may not vote and states may or may not go to war. Although the intuition is less straightforward, these models also are appropriate only when there is at least one exogenous variable that influences only the selection stage.

3 The New Estimator

This section of the article develops the estimator. I begin by discussing the data-generating process. Next, I discuss my identifying assumption, which determines the conditions under which this estimator is most likely to be useful. I then derive the estimator.

Maximum-likelihood estimators usually have several nice properties: consistency, asymptotic normality, and asymptotic efficiency. Four regularity conditions together are sufficient for these properties (King 1990). My estimator does not meet all of the conditions; the range of the data depends upon the unknown parameters, as I discuss later. However, the regularity conditions are not necessary for consistency or asymptotic normality. In an appendix, I prove that the estimates are consistent and asymptotically normal. Because this is a maximum-likelihood estimator, asymptotic efficiency follows from asymptotic normality (Amemiya 1985, p. 123).

3.1 *The Underlying Model*

Consider a situation with nonrandom selection, where the data-generating process is as described earlier, except that the dependent variable in the equation of interest is dichotomous. The equations in question are

$$U_{1i} = \boldsymbol{\gamma}'\mathbf{x}_i + u_{1i} \tag{5}$$

and

$$U_{2i} = \boldsymbol{\beta}'\mathbf{x}_i + u_{2i}. \tag{6}$$

In these equations, each U represents an unobserved continuous dependent variable, often an actor's utility from taking an action. Each U is the result of systematic factors $\mathbf{x}$ and a normally distributed, mean-zero error term. The explanatory variables, $\mathbf{x}$, are the same in the two stages, but the coefficients usually differ. Rather than the Us, one observes Z_{1i} and Z_{2i}, both dichotomous variables. Equation (5) is the selection equation, and Z_{1i} represents whether or not the observation is selected. Equation (6) is the equation of interest, or outcome equation, and Z_{2i} represents the observed outcome of the equation of interest. In each equation, if the relevant utility is greater than zero, the outcome of the observed dependent variable is a 1 (e.g. the actor takes the action); if not, it is a 0:

$$Z_{1i} = \begin{cases} 0 & \text{if } U_{1i} < 0 \\ 1 & \text{if } U_{1i} \geq 0; \end{cases} \tag{7}$$

$$Z_{2i} = \begin{cases} 0 & \text{if } U_{2i} < 0 \\ 1 & \text{if } U_{2i} \geq 0. \end{cases} \tag{8}$$

3.2 *Identical Errors*

To solve the identification problem discussed earlier, I make use of an additional piece of information. Under many circumstances, the researcher has theoretical reason to believe

that the error terms in the two equations, u_1 and u_2, are at least nearly identical for a given observation. In the situation that Eqs. (5) and (6) model, the observed dependent variables are dichotomous, the underlying dependent variables are on the same scale (they are both utilities, in which scale can be standardized), and both error terms are normally distributed. Thus, it is initially plausible that the two equations have similar error terms.

The error terms have two sets of components: small events that affect the relevant actor's or actors' utility, and any omitted variables. Examination of these sources of error yields three conditions under which errors are likely to be similar.

First, and most obviously, processes that involve similar decisions or goals are more likely to have similar error terms than are unrelated phenomena. If one wanted to explain warring states' decisions about child-care policy, one would not want to assume that identical random or unmeasured factors influenced that decision and the decision to go to war.

Second, errors are likely to be similar when selection and the outcome of interest have the same causes. In political science applications, the bulk of the error term often consists of omitted variables. The researcher would prefer not to omit important variables but may miss some influences and find it impossible to measure others. For example, states' resolve is notoriously hard to measure. Fortunately, when identical included explanatory variables belong in the two equations (the situation that this model is designed for), this is often because selection and the outcome of interest have the same causes. If so, then identical omitted factors also influence both dependent variables, and the two error terms are highly correlated. In most applications, the errors probably are not perfectly correlated, because omitted variables are not the only source of error and because the relationships between the unmeasured independent variables and the two dependent variables are unlikely to be identical.

Finally, errors are more likely to be similar when the decisions are close together in time and space. A part of the error term is also due to small random factors, such as the weather. These are more likely to influence two phenomena if the phenomena occur on the same day in the same place.

The ideal application for this estimator is thus one in which selection and the outcome of interest represent similar decisions, have the same causes, and are close together in time and space. However, because the unmeasured explanatory factors are often more important than the small random factors (in the example, resolve dwarfs the weather as an influence on crisis escalation), the estimator is often a good choice if the first two conditions are met, even if the two decisions are farther apart in time and/or space.

In sum, the assumption of identical errors is often a good approximation exactly when it is needed—that is, when selection and the subsequent outcome of interest are dichotomous and have the same causes. When one has reason to believe that the errors are quite similar, one can solve the identification problem by assuming that they are identical.

Occasionally, theory may suggest that selection and the subsequent outcome have identical causes but that the unmeasured causal factors have an opposite effect on the two decisions. This situation is rare; if the selection and outcome processes are similar enough to have the same causes, the effects are likely to have the same signs. When the situation does arise, however, one should use an alternative version of the estimator that rests on an assumption that the errors are opposite, rather than identical. This is an option in the software available on the *Political Analysis* Web site; a document on the site discusses the option. The simulations and proofs in this article all pertain to the version of the estimator that assumes that the errors are identical.

3.3 *One-Step Maximum-Likelihood Estimation of the Effect of x_k*

This section derives a maximum-likelihood estimator for the effect of the independent variables on the dependent variable of interest. So far, I have considered the selection mechanism and the equation of interest separately. The estimator that I develop here considers the two equations together.[9]

I now assume that the error terms for the two equations are identical for each observation, i.e., $u_{1i} = u_{2i} \forall i$.

I define three random variables Y_{ij} such that

$$Y_{i0} = \begin{cases} 1 & \text{if } Z_1 = 0 \\ 0 & \text{otherwise;} \end{cases} \tag{9}$$

$$Y_{i1} = \begin{cases} 1 & \text{if } Z_1 = 1 \text{ and } Z_2 = 0 \\ 0 & \text{otherwise;} \end{cases} \tag{10}$$

$$Y_{i2} = \begin{cases} 1 & \text{if } Z_1 = 1 \text{ and } Z_2 = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

That is, Y_{i0} has a value of 1 iff the observation is not selected; Y_{i1} has a value of 1 iff the observation is selected and the outcome of the equation of interest is 0; and Y_{i2} has a value of 1 iff the observation is selected and the outcome of the equation of interest is 1.

I also define

$$P_{ij} \equiv \text{prob}(Y_{ij} = 1). \tag{12}$$

The likelihood function is proportional to the product of the probabilities of the observations, which is $\prod_{i=1}^{n} \prod_{j=0}^{2} P_{ij}^{Y_{ij}}$. Then,

$$L^* \equiv \ln L \propto \sum_{i=1}^{n} \sum_{j=0}^{2} Y_{ij} \ln P_{ij}, \tag{13}$$

where $Y_{ij} \ln P_{ij}$ is defined as 0 if $Y_{ij} = 0$ and $P_{ij} \leq 0$ in order to simplify the notation.[10]

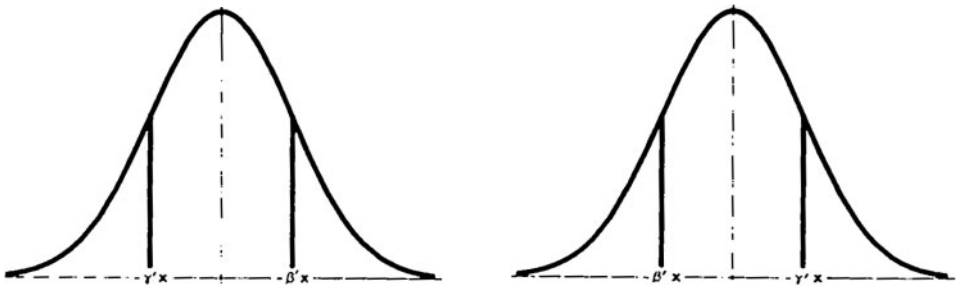
To build the likelihood function, I need the ex ante probability of seeing an observation, P_{ij}. This probability is different for each j. There are three cases:

1. $j = 0$. The observation is not selected into the second equation (i.e., $Z_1 = 0$). This outcome occurs if $\boldsymbol{\gamma}'\mathbf{x}_i + u_i < 0$. The probability of this occurrence is prob($u_i < -\boldsymbol{\gamma}'\mathbf{x}_i$). Because u_i has a standard normal distribution, this probability is just the cumulative standard normal density evaluated at $-\boldsymbol{\gamma}'\mathbf{x}_i$; this is the area under the curve to the left of the line at $-\boldsymbol{\gamma}'\mathbf{x}$ in Fig. 1A or 1B. Thus,

$$P_{i0} = \Phi(-\boldsymbol{\gamma}'\mathbf{x}_i). \tag{14}$$

[9] The analogous Heckman-type estimator also considers the two equations together.

[10] The likelihood is proportional to the probability of the data (King 1990, p. 22). The maximum of the likelihood is attained at the same parameter values as the maximum of the probability of the data, and so I replace the proportionality sign with an "equals" sign for the remainder of the paper.



(A) (B)

Fig. 1 The distribution of u_i when (A) $-\gamma'\mathbf{x}_i < -\beta'\mathbf{x}_i$ and (B) $-\beta'\mathbf{x}_i < -\gamma'\mathbf{x}_i$.

2. $j = 1$. The observation is selected into the second equation and has an observed outcome of 0 in that equation. In the equation of interest (6), the outcome is 0 if $\beta'\mathbf{x}_i + u_i < 0$, or $u_i < -\beta'\mathbf{x}_i$. The derivation of P_{i1} requires two steps:
   1. Consider an observation and true model for which $-\beta'\mathbf{x}_i > -\gamma'\mathbf{x}_i$ for the true parameter values and the observation in question (Fig. 1A). In the usual probit setup, the probability of a 0 would be $\Phi(-\beta'\mathbf{x}_i)$, represented by the area to the left of the line at $-\beta'\mathbf{x}$ in Fig. 1A. However, the observations for which $u_i < -\gamma'\mathbf{x}_i$ did not make it into the sample. Thus, the ex ante probability that a 0 is in the equation of interest is $\Phi(-\beta'\mathbf{x}_i) - \Phi(-\gamma'\mathbf{x}_i)$. This is the probability represented by the region of Fig. 1A between the two vertical lines.
   2. Now consider an observation and model for which $-\beta'\mathbf{x}_i \leq -\gamma'\mathbf{x}_i$ (Fig. 1B). In this case, one cannot observe $Y_{i1} = 1$. To see this, suppose that the error term for the observation is such that $u_i < -\beta'\mathbf{x}_i$. Then $u_i < -\gamma'\mathbf{x}_i$ also. Thus, any observation that would have an outcome of 0 in the second stage is not selected into the sample for that stage. Thus,

$$P_{i1} = \begin{cases} \Phi(-\beta'\mathbf{x}_i) - \Phi(-\gamma'\mathbf{x}_i) & \text{if } (\gamma' - \beta')\mathbf{x}_i > 0 \\ 0 & \text{if } (\gamma' - \beta')\mathbf{x}_i \leq 0. \end{cases} \tag{15}$$

   When the observed Y_{i1} equals 1, $-\beta'\mathbf{x}_i \leq -\gamma'\mathbf{x}_i$ leads to a contradiction; the probability of an existing observation cannot be 0. If an observation is selected into the second equation and has an outcome of 0 in that equation, it therefore must be that $-\beta'\mathbf{x}_i > -\gamma'\mathbf{x}_i$ for that observation at the true values of the parameters.
3. $j = 2$. The observation is selected into the second equation and has an observed outcome of 1 in that equation. In the equation of interest (6), the outcome is 1 if $\beta'\mathbf{x}_i + u_i > 0$, or $u_i > -\beta'\mathbf{x}_i$. The derivation of P_{i2} also takes two steps:
   1. If $-\beta'\mathbf{x}_i > -\gamma'\mathbf{x}_i$ (Fig. 1A), then any observation for which $\beta'\mathbf{x}_i + u_i > 0$ has a value of 1 on the observed dependent variable. The probability of such an occurrence is $1 - \Phi(-\beta'\mathbf{x}_i)$, or $\Phi(\beta'\mathbf{x}_i)$.
   2. If $-\beta'\mathbf{x}_i \leq -\gamma'\mathbf{x}_i$ (Fig. 1B), then not all observations for which $u_i > -\beta'\mathbf{x}_i$ are in the sample for the second stage. In this situation, only when $u_i > -\gamma'\mathbf{x}_i$ does one observe a 1 in the dependent variable of interest. Thus, the probability of a 1 in the second-stage is $1 - \Phi(-\gamma'\mathbf{x}_i)$, or $\Phi(\gamma'\mathbf{x}_i)$.

Thus,

$$P_{i2} = \begin{cases} \Phi(\boldsymbol{\beta}'\mathbf{x}_i) & \text{if } (\boldsymbol{\gamma}' - \boldsymbol{\beta}')\mathbf{x}_i > 0 \\ \Phi(\boldsymbol{\gamma}'\mathbf{x}_i) & \text{if } (\boldsymbol{\gamma}' - \boldsymbol{\beta}')\mathbf{x}_i \leq 0. \end{cases} \tag{16}$$

If $-\boldsymbol{\beta}'\mathbf{x}_i < -\boldsymbol{\gamma}'\mathbf{x}_i$, then the second version of P_{i2} will be smaller; if the reverse is true, the first version will be smaller, and so in practice,

$$P_{i2} = \min[\Phi(\boldsymbol{\gamma}'\mathbf{x}_i), \Phi(\boldsymbol{\beta}'\mathbf{x}_i)]. \tag{17}$$

I distinguish the two probabilities by subscript because I use them separately in the proofs:

$$\begin{aligned} P_{i21} &= \Phi(\boldsymbol{\beta}'\mathbf{x}_i); \\ P_{i22} &= \Phi(\boldsymbol{\gamma}'\mathbf{x}_i). \end{aligned} \tag{18}$$

Having calculated the probabilities that enter into the likelihood function, I now define the estimator:

$$\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}} := \max_{\boldsymbol{\beta}, \boldsymbol{\gamma} \in \boldsymbol{\Theta}} L^*, \tag{19}$$

where $\boldsymbol{\theta}$ is the vector of all parameters and $\boldsymbol{\Theta}$ is the parameter space.[11] The variance–covariance matrix of the estimators is estimated by:

$$\left(-\frac{\partial^2 L(\hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}} \partial \hat{\boldsymbol{\theta}}'} \right)^{-1} \tag{20}$$

(Greene 1993, p. 115).

The estimates never imply that an observed outcome is impossible. However, depending upon the data, they may imply that an outcome that never is observed is impossible. In other words, they may imply that if an observation with certain values of the independent variables $\mathbf{x}_i$ selects in, then that observation always obtains a value of 1 for the observed dependent variable in the outcome equation (never has the outcome $Y_{i1} = 1$). This is most likely for values of $\mathbf{x}_i$ outside the defined or usual ranges of the independent variables.[12]

4 Experimental Comparison of Probit, Heckman, and Sartori Estimators

In this section, I use Monte Carlo simulations to investigate the conditions under which my estimator is an improvement over the Heckman estimator and over ordinary probit. The simulations vary the "real" situation in order to evaluate the performance of each estimator under different conditions, all of which include nonrandom selection.

[11] The possible presence of a zero or negative probability in the likelihood (from $\hat{P}_{i1}$) when the parameters are not equal to their true values can pose a problem for estimation using a numerical search routine since the log likelihood is undefined for these probabilities. The software surmounts this problem by replacing the zero or negative number in the likelihood function with a "punishment" variable that moves the optimization routine away from guesses that are impossible. It also confirms that the final estimates are in the feasible range.

[12] It might be possible to modify the computer program to restrict the estimates to values that imply that all outcomes are possible within the plausible ranges of the independent variables.

As I discussed earlier, scholars have two main reactions when theory implies that identical explanatory factors influence selection and the subsequent outcome of interest. Some estimate with identical explanatory variables, whereas others add an exclusion restriction. I thus perform two sets of simulations. In the first and primary set of simulations, I evaluate the three estimators, assuming that identical explanatory factors influence selection and the subsequent outcome of interest (the situation for which my estimator is intended). In the second set of simulations, I evaluate them assuming that there is a valid but weak exclusion restriction. Note that if identical explanatory factors truly do influence selection and the subsequent outcome of interest, a valid exclusion restriction simply does not exist. My second set of simulations is based on the idea that the researcher, through additional thought, might come to the conclusion that his/her original theory was incomplete, and that selection and the subsequent outcome of interest do not after all have identical causes.

In each experiment, I estimate 1000 times using each of the three estimators.[13] To create the data sets, I begin by creating an *explanatory variable*, x. I use two samples, one with 100 and one with 1000 observations. I do so by drawing 1000 observations from a normal distribution with mean 0 and variance 0.64. I set the variance of the explanatory variable at 0.64 so that the systematic component of the selection mechanism will have the same variance as the error term; in other words, the systematic component and the error are equally important in the selection equation. (Later, I report results from simulations that vary the ratio of the systematic component to the error.) I take the first 100 observations as my sample of 100 and the entire data set as my sample of 1000. The realized mean for the first 100 observations of the independent variable is 0.135 and the standard deviation is 0.6464. The realized mean for the whole sample of 1000 is 0.0385 and the standard deviation is 0.63264.

Next, for each experiment, I generate 1000 samples of the dependent variable. As I describe later, in each sample, the dependent variable is a function of the explanatory variable (the 100 or 1000 observations of x) and of new draws of the error terms. In other words, the data-generating process has a systematic component and a random component.

Because ordinary probit gives consistent estimates of the selection-equation parameters, I discuss only the estimates for the outcome equation. In the interests of brevity, I discuss only the slope estimates; I also consider only one criterion, the root mean square error, for evaluating the simulations with a weak exclusion restriction.

4.1 *Simulations with Identical Explanatory Variables*

In the simulations with identical explanatory variables, I create the underlying utilities and the dichotomous dependent variables as follows:

$$U_1 = 1.25 * x + u_1 \tag{21}$$

$$U_2 = -0.7 + 1.5 * x + u_2. \tag{22}$$

Here, U_1 represents the actor's utility from "selecting in" and U_2 represents its utility from "going on," or getting a 1 in the outcome equation. As is usual in probit-style models, one observes a dichotomous dependent variable instead of the underlying utility. One observes

[13] I conduct all of the simulations in STATA using STATA's heckprob and probit commands and the program that implements my estimator.

Table 1 Mean bias ($\beta_2 = 1.5$) (numbers rounded)

ρ	*Probit*	*Heckman*	*Sartori*
	100 observations		
.9	−.181	−.212	.102
.5	−.0657	−.164	.159
.1	.105	−.0647	.294
	1000 observations		
.9	−.314	−.217	−.00693
.5	−.227	−.157	.0478
.1	−.0437	−.0859	.176

$Y_1 = 1$ if $U_1 > 0$; $Y_1 = 0$ otherwise, and $Y_2 = 1$ if $U_1 > 0$; $Y_2 = 0$ otherwise. In the analyses that follow, I refer to the intercept and slope in the selection equation (21) as α_1 and β_1, respectively, and those from the outcome equation (22) as α_2 and β_2, respectively.

The Heckman estimator's key identifying assumption is that the errors in the selection equation and the outcome equation are jointly normally distributed. In practice, the researcher often does not know to what extent this assumption accurately represents the data-generating process. For this reason, one purpose of the simulations was to evaluate the performance of each estimator with a variety of error-term distributions. However, it turns out that the Heckman estimates are quite poor when there is no valid exclusion restriction, even when the estimator's key identifying assumption is met.[14] Thus, this article discusses only the simulations with error terms (us) that are drawn from a bivariate normal distribution.

My estimator's key identifying assumption is that the errors from the two equations have a correlation of 1 (or of −1, depending upon the choice of the user).[15] In practice, the researcher may know from theory that the errors are highly correlated, but is unlikely to know that the correlation is 1. Thus, a second goal of the simulations was to vary the true correlation between the errors and to evaluate the performance of each estimator. I consider situations in which the true correlation (ρ) between u_1 and u_2 is .9 (my estimator's assumption is close to true). I then consider situations in which the true ρ is only .5 (the assumption is fairly far from true) and .1 (the assumption is quite far from true). Later, I report results from simulations in which the true ρ is −.1, −.5, and −.9.

I evaluate the coefficient estimates and the standard-error estimates produced by each of the estimators using three criteria: the mean bias of the estimates (MB), the root mean square error (RMSE) of the estimates around the true parameter, and the coverage.[16] The mean bias is information about the average result: insofar as the estimates differ from the truth, to what extent do positive errors and negative errors cancel each other out? The RMSE is the average variability of the point estimate around the true parameter. The RMSE differs from the sampling variability, which is variation around the sample mean of the estimate, rather than around the true parameter. For example, the RMSE of the probit estimate of

[14] Some results from simulations with modified t and chi-square errors are available on the *Political Analysis* Web site.

[15] In the simulations, I use the version of my estimator that assumes that the correlation between the errors is +1. My estimator also assumes that the errors are normally distributed.

[16] The mean bias is the bias divided by the number of replications in which the estimates in question converged.

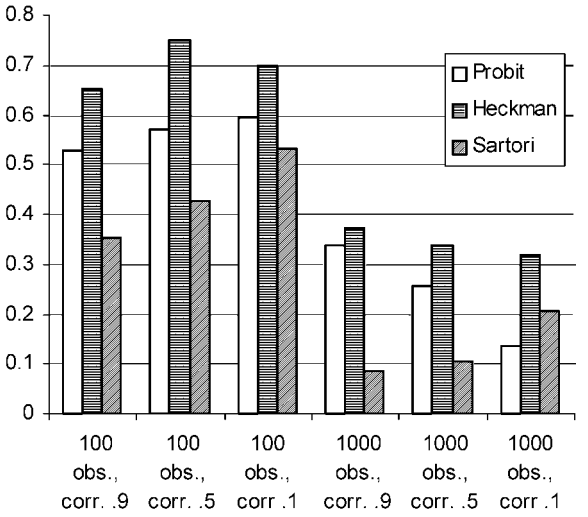


Fig. 2 RMSEs of estimates of β_2, 1000 observations.

β_2 is $\sqrt{\frac{\sum_i(\hat{\beta}_{i2}^{\mathrm{P}}-\beta_2)^2}{n}}$, where β_2 is the true value, $\hat{\beta}_{i2}^{\mathrm{P}}$ is the probit estimate of β_2 in the ith replication, and n is the number of estimates.[17]

The coverage is information about the combined accuracy of the coefficient estimates and the standard error estimates. I compute the 95% confidence interval for each replication of the simulation using the estimate of β_2 and the estimate of its standard error. The coverage is the frequency with which these confidence intervals contain the true value of β_2. Ideally, one would like the 95% confidence interval to contain the true value 95% of the time.

In calculating these summary statistics, the issue arises of how to treat replications in which one or more estimators did not converge. Only probit converged in every replication. For example, with 100 observations, Heckman converged 912/1000 times with a true ρ of .9 and 907/1000 times with a true ρ of .5. With the same sample, my estimator converged 922/1000 times with a true ρ of .9 and 943/1000 times with a true ρ of .5. In the statistics that follow, the numbers reported are averages over the replications that did converge. I would speculate that the Heckman estimator and my estimator are failing to converge when estimation is difficult, and that the nonconvergence results should be considered bad estimates. (If the estimates had converged in replication j, they would have been far from the truth.) If so, my estimator probably performs better relative to the Heckman estimator than I report below in most of the simulations (because it converges more frequently), but both selection estimators perform worse in relation to ordinary probit.[18]

4.1.1 Coefficient Estimates: Mean Bias and RMSEs

The striking result of the simulations is that my estimator performs better than Heckman's, *even when Heckman estimator's key assumption about functional form is met, and my estimator's assumption of identical errors is fairly far from the truth*. Table 1 and Fig. 2

[17] The number of estimates is not 1000 for every estimator and simulation because the Heckman estimator and my new estimator did not always converge.

[18] However, one might prefer a result of nonconvergence to an inaccurate estimate, because one has no way of knowing which estimates are inaccurate.

Table 2 Mean bias, new sample of 100 observations ($\beta_2 = 1.5$) (numbers rounded)

ρ	*Probit*	*Heckman*	*Sartori*
.9	−.218	−.320	.0743
.5	−.120	−.287	.123
.1	.114	−.229	.305

show the mean bias and the RMSE of the slope coefficient in the outcome equation. In all the simulations that the table and the figure summarize, Heckman's key assumption is met. In none of these simulations is my estimator's key assumption met—the highest correlation in the simulations is .9, whereas my estimator assumes a correlation of 1.0.

Despite the Heckman estimator having this advantage, the Heckman estimates of the slope coefficient have greater mean bias than my estimates, even when my estimator's assumption is fairly far from the truth—the true ρ is only .5. Only when there is almost no selection bias (the true ρ is .1), does Heckman estimator have less bias than my estimator. Moreover, when the true correlation is this low, ordinary probit outperforms the Heckman estimator in the simulation with 1000 observations.[19] The magnitudes of the biases are small but not negligible; for example, the Heckman estimator's bias of $-.217$ when the true ρ is .9 and the sample size is 1000 corresponds to about 15% of the magnitude of the coefficient. Of course, the direction of the bias also matters. With the sample of 100 observations, my estimator's average bias would overstate the effect of the independent variable, whereas Heckman's would understate it. With the sample of 1000 observations and correlations of .5 and .9, my estimator has only trivial bias, making the direction of the bias almost inconsequential.

The results from the simulations with 100 observations are somewhat peculiar. Because ordinary probit is consistent when the errors in the selection equation and the outcome equation are uncorrelated, I would expect probit to do worse at high (in absolute value) values of ρ. In this set of simulations, however, the ordinary probit estimates have less average bias when the true correlation between the errors is .5 than they do when the true correlation is .1. This result is probably an artifact of the particular small sample of the explanatory variable used in the simulations. Table 2 shows results of simulations that use a different sample of 100 observations of the explanatory variable.[20] Using this second set of data, probit has lower mean bias when ρ is lower.

My estimator has a smaller RMSE for the outcome-equation estimates than either of the other estimators in every simulation shown in the figure—regardless of whether the sample has 100 or 1000 observations, and, more surprisingly, of whether the true ρ is .9, .5, or .1. The impact of the Heckman estimator's RMSE when the independent variables are identical is again not trivial. Even with samples of 1000 observations, the Heckman estimator's RMSE is more than a fifth of the true slope. With 1000 observations, my estimator's RMSE for $\hat{\beta}_2$ is only 23% or 31% that of Heckman's, depending upon the true correlation between the error terms.

[19] With this sample of 100 and a ρ of .5, ordinary probit has lower mean bias than either of the other two estimators. However, my estimator has lower RMSE and more-accurate coverage.

[20] The new sample of the explanatory variable is drawn from the same distribution as the previous one; the realized mean of the explanatory variable is -0.069. I create 1000 samples on which to estimate the parameters in the same way as before, by taking new draws of the error terms.

Table 3 Coverage, 95% confidence interval (numbers rounded)

ρ	*Probit*	*Heckman*	*Sartori*
	100 observations		
.9	.873	.828	.933
.5	.905	.856	.934
.1	.952	.836	.931
	1000 observations		
.9	.253	.868	.948
.5	.499	.889	.919
.1	.930	.797	.599

In sum, the coefficient estimates in the simulations suggest that my estimator is better suited to the problem of identical explanatory variables than is the Heckman one. My estimates have lower mean bias and RMSE around the true coefficient than Heckman's, even when the assumption of identical errors is fairly far from true and Heckman's assumption of normally distributed errors is true. In very small samples, however, my estimator, on average, leads to overstating the impact of the explanatory variable whereas the Heckman estimator leads to understating them. I next investigate the coverage, because this may affect the conclusions the researcher draws from the estimates of the coefficients.

4.1.2 Inference: Coverage

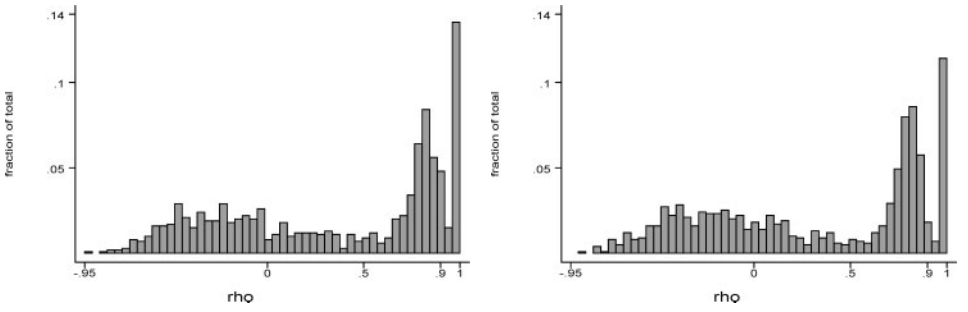
When identical explanatory variables influence selection and the outcome of interest, my new estimator also results in more accurate inference than the Heckman estimator, even when the Heckman estimator's key assumption is met and my estimator's assumption of identical errors is fairly far from the truth. As Table 3 shows, my estimator's 95% confidence interval is closer to accurate when the true ρ between the errors is .9 or .5. When the true correlation is .1 (there is almost no selection bias) and the sample size is 1000, Heckman's confidence interval is more accurate than mine, but probit's is the most accurate of all.

4.1.3 The Heckman Estimator's Estimates of ρ

The Heckman estimator is the only one of the three procedures that yields an estimate of the correlation between the error terms in the two equations, ρ. This is an advantage—in general, one would rather estimate a parameter than assume its value. However, when identical explanatory variables belong in the selection equation and the outcome equation, Heckman estimator's estimate of ρ is often extremely misleading, as Table 4 and Figs. 3A and B show. In some of the simulations, such as the simulation with 1000 observations shown in the figure, there is a relatively high frequency of estimates that are 1 or very close to 1, regardless of whether the true ρ is .9 or .5. The distribution of ρ has a spike just shy of 1;

Table 4 RMSEs of Heckprob's estimates of ρ

True ρ	*100 observations*	*1000 observations*
.9	1.02	.784
.5	.826	.582
.1	.788	.586



(A) (B)

Fig. 3 Histogram of Heckman's estimates of ρ: true ρ is (A) .9 and (B) .5.

this spike need not occur around the truth, since it also appears when the true ρ is .5. Aside from those two areas of higher probability, the distribution is fairly flat. Thus, the fact that the Heckman procedure provides estimates of ρ is not an advantage when identical explanatory factors influence selection and the outcome and one has strong prior beliefs about ρ.

4.1.4 At What True Values of ρ is Each Estimator Best?

The figures and tables above show that my estimator is a better choice for this problem than the Heckman-type ones, even when the true ρ between the errors is as low as .5. My estimator allows one to assume that the true correlation between the errors is $+1$ or -1. In the simulations, I use the version of the estimator that assumes a correlation of positive one. As one would expect, the tables and figures mentioned above show that this version of my estimator performs better when the true correlation is higher.

Figure 4 draws out this pattern further. The figure examines the RMSE of the slope coefficient in the outcome equation for a greater range of true correlations of the error term. In each of these simulations, the errors are normally distributed and the sample size is 1000.

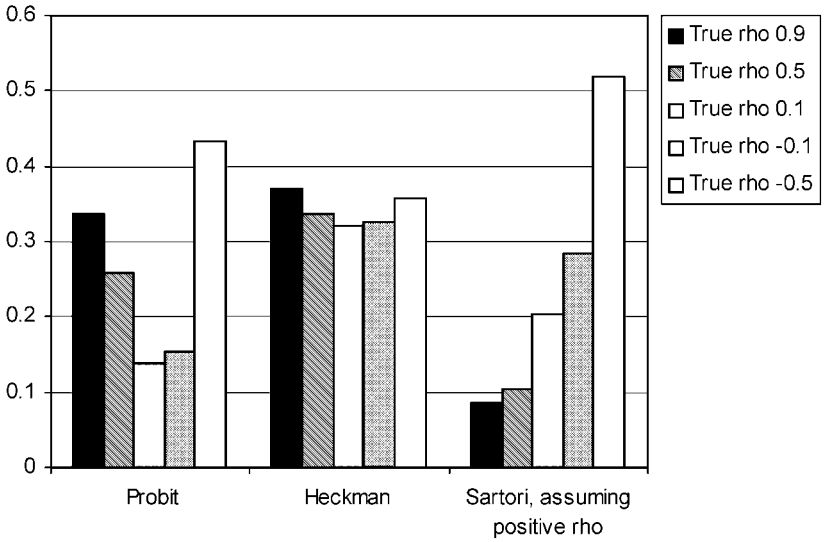


Fig. 4 RMSEs of estimates of β_2.

The figure shows more clearly that my estimator does best when its assumption is closest to true. My estimate has lower RMSE than Heckman's in each simulation in this set except the simulation with a true ρ of $-.5$.[21] However, when the true ρ between the errors is small in magnitude (.1 or $-.1$ in these simulations), ordinary probit has lower RMSE than either of the selection estimators.

The mean bias of my estimates is lower than that of Heckman's at high correlations and becomes higher than Heckman's as the correlation moves toward zero. In the sample with 1000 observations, the switch occurs somewhere between a true ρ of .3, where my estimates have a smaller mean bias, and one of .2, where Heckman's have a smaller mean bias. Similarly, my estimator's coverage is more accurate at high correlations and becomes less accurate at low correlations. In the sample with 1000 observations, the switch occurs somewhere between a true ρ of .4, where my estimator's coverage is more accurate, and .3, where Heckman's coverage is more accurate. At very low correlations, probit has the lowest bias and most-accurate coverage.

The right choice when identical explanatory variables belong in the two equations thus depends upon the guidance that theory provides about the sign and size of the correlation. Probit is the right choice if the errors are largely unrelated. My estimator is the right choice if the correlation between the errors is high and of a specific sign, either positive or negative. Finally, the Heckman estimator is the best choice if the errors are highly correlated (making probit undesirable), but the researcher is quite uncertain about the true sign of the correlation (making my estimator undesirable). Even under these very limited circumstances, Heckman estimates are not good, they are just better than those of the other estimators. Since Heckman estimates of ρ are often very far from the truth (see above), the choice of estimators must be made largely on theory rather than on the basis of those estimates.

4.1.5 Do the Results Change if the Systematic Component Is More/Less Important Relative to the Error Term?

In the simulations I discussed earlier, the systematic component and the error term in the selection equation are equally important. One might expect ordinary probit to perform better in relation to both Heckman and my estimator when the error term is less important. This is because the inconsistency of probit ("selection bias") is due, in part, to the fact that the observations' error terms affect their probabilities of selecting into the sample. To assess the sensitivity of my conclusions to the relative importance of the error term, I perform additional simulations. These are identical to those I described earlier, except that I vary the ratio of the variance of the systematic component to that of the error term in the selection equation [$\text{var}(\beta_1 x)/\text{var}(u)$].[22] The RMSEs reveal the same general pattern as before: My estimator performs better than Heckman's and probit when the errors are highly correlated. Probit performs better than both selection estimators when the correlation between the errors is very low.

As expected, probit tends to perform better in absolute terms when the systematic component is more important relative to the error term. However, it does not perform better *relative* to the other two estimators. This is because all the estimators tend to perform better when the systematic component is more important; a more important systematic component

[21] I did not investigate correlations between $-.1$ and $-.5$. The Heckman estimator probably begins to outperform mine when the correlation is between those values.

[22] I do this by changing the variance of the explanatory variable. The values of the ratio $\frac{\text{var}(\beta_1 x)}{\text{var}(u)}$ in the table are based on the theoretical variance of x rather than its sample variance.

Table 5 RMSEs of β_2, $\rho = .5$, varying the relative importance of the systematic component and the error

$\frac{\text{var}(\beta_1 x)}{\text{var}(u)}$	*Probit*	*Heckman*	*Sartori*
0.1	.374	.910	.210
0.5	.283	.461	.115
1	.258	.337	.105
2	.241	.275	.00994
10	.234	.260	.119

means more information and less noise for the estimators to work with. As an example, Table 5 shows the results of varying $\frac{\text{var}(\beta_1 x)}{\text{var}(u)}$ when the true ρ is .5 and the sample size is 1000. There are two surprising results. First, the Heckman estimator performs especially badly in comparison to the others when there is very little information [$\frac{\text{var}(\beta_1 x)}{\text{var}(u)} = .1$]. Second, although all of the estimators improve in RMSE as the importance of the systematic component increases, my estimator does worse when the ratio of the variance of the systematic component to that of the error is very high (10:1), than when it is moderately high (2:1). At even higher ratios of $\frac{\text{var}(\beta_1 x)}{\text{var}(u)}$ (not shown), the RMSEs of probit and the Heckman-type estimator also begin to increase.

4.2 *Simulations with an Exclusion Restriction*

When theory points to identical explanatory factors, a common response is a mad search for an exclusion restriction.[23] This practice is dangerous because including the "extra" variable may lead to specification error. For example, if the extra variable does not affect selection in the population but is correlated with a true explanatory variable and with selection in the sample, including it in the equation can bias the estimates of the effect of the explanatory variable.

This section of the article assumes a best-case scenario for a weak exclusion restriction: the dredged up explanatory variable (z) actually does influence selection and is uncorrelated with the explanatory variable of interest. I investigate the RMSEs of the slope coefficient (β_2) in the outcome equation for each of the estimators under this scenario, using a sample size of 1000 and errors from a bivariate standard normal distribution with correlation .9, .5, or .1. I create the "extra" variable that enters the selection equation, z, by taking 1000 draws from a uniform normal distribution, independent of the distribution of the other explanatory variable, x. I examine two cases: one in which the true coefficient on z is small (.05), and one in which the true coefficient is larger (.25).[24] The small coefficient corresponds to a "weak" exclusion restriction, whereas the larger one corresponds to a stronger restriction. When I estimate with ordinary probit or with the Heckman estimator, I include z in the model. However, when I use my estimator, I estimate only with x.

For the weaker exclusion restriction, I choose the magnitude of the coefficient (c) on the extra variable (z) to obtain a t-statistic of just over 1, on average, for the Heckman estimates. My reasoning is that a researcher finding a t-statistic of that size might believe

[23] If there is a valid and significant exclusion restriction, then the Heckman estimator is the best choice. By definition, however, there is no valid exclusion restriction if selection and the subsequent outcome truly have identical causes.

[24] Note that the extra variable z has higher variance than the original explanatory variable x, so that its coefficient is not immediately comparable to the coefficient on x.

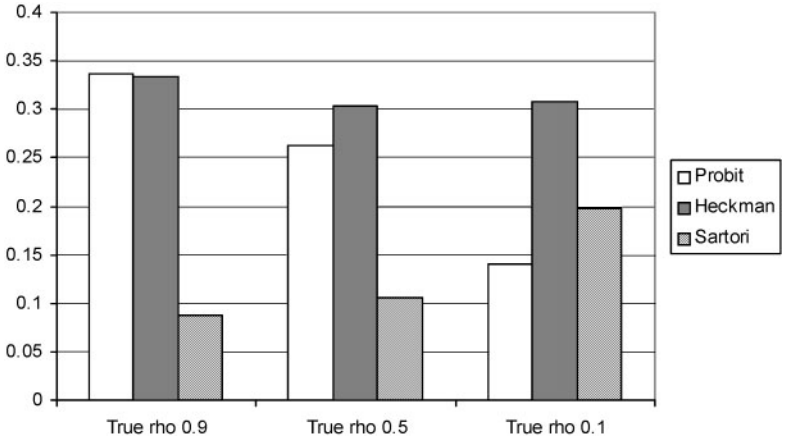


Fig. 5 RMSEs of $\hat{\beta}_2$ when a very weak exclusion restriction is valid.

himself or herself to have a valid, though weak, exclusion restriction.[25] With the weak exclusion restriction, moving from one standard deviation below the mean of *z* to one standard deviation above (with *x* at its mean) leads to a true increase of about 4 percentage points in the probability of selecting into the sample.

For the stronger exclusion restriction, I choose the magnitude of the coefficient on the extra variable to obtain a *t*-statistic of over 5, on average, for the Heckman estimates. Moving from one standard deviation below the mean of *z* to one standard deviation above, again with *x* at its mean, leads to an increase of almost 20 percentage points in the probability of selecting into the sample.

Figure 5 shows the RMSEs of the slope coefficients in the outcome equation when the weaker exclusion restriction is valid. The figure demonstrates that the weaker exclusion restriction is a poor assistant to the Heckman model. Heckman estimator's RMSEs are very similar to what they were for the model without a valid exclusion restriction. My estimator's RMSEs are also very similar, and remain much smaller than Heckman's in all of the simulations.[26] As in the previous simulations, my estimator outperforms probit in the simulations where the true ρ between the errors is .5 or .9, and underperforms probit when the true ρ is .1.

The stronger exclusion restriction is more of a help to the Heckman model, but it is not a savior. When the true ρ between the errors is .5 or .9 and the extra variable really is explanatory, the Heckman estimator now does considerably better than ordinary probit, which it did not before (see Fig. 6). However, my estimator still has lower RMSEs than the Heckman estimator for all three correlations (.9, .5, and .1), even though I used an incorrect specification for my estimates, because the extra variable does influence selection in this set of simulations.

Clearly, the Heckman estimator will outperform mine when the extra variable has a large enough effect (and truly is explanatory). However, as these simulations show, *large enough* is not a trivial requirement. In this example, varying the more-important "extra" explanatory variable from a standard deviation below its mean to a standard deviation above

[25] Of course, *t*-statistics vary with sample size.

[26] I do not examine here situations in which the true correlation is highly negative and I assume it to be highly positive. The Heckman estimator would surely outperform mine in such a simulation, as it did without the exclusion restriction.

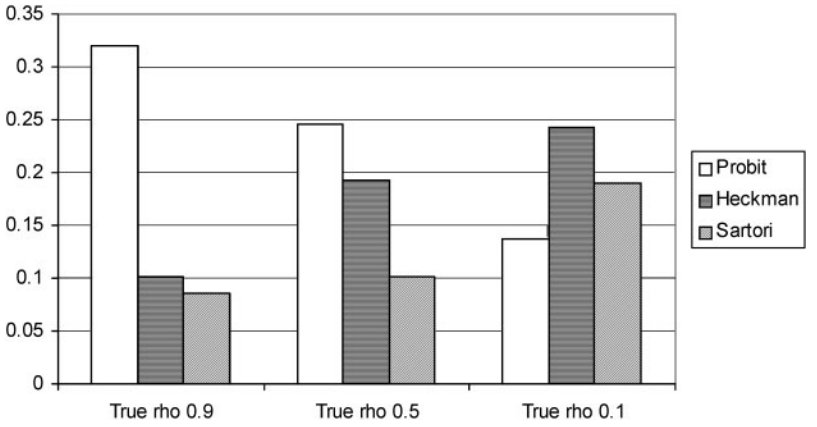


Fig. 6 RMSEs of $\hat{\beta}_2$ when a stronger (but still weak) exclusion restriction is vaild.

results in a 20 percentage point increase in the probability of selection, but the Heckman estimator's RMSEs are still larger than mine. If a researcher grubs up a variable simply to get an exclusion restriction, chances are that this variable will not be important enough to make Heckman estimates the better choice, even if the variable is actually related to selection. If theory truly indicates that at least one variable has a reasonably large effect on selection but not on the subsequent dependent variable, then the Heckman estimates indeed are likely to be better.

5 Does it Matter?

Does the choice of estimators for this problem affect the conclusions one reaches when analyzing real data? In a companion paper, I show that the answer is "yes" for one important substantive example (Sartori 2002). Lemke and Reed (2001) use the Heckman estimator to examine the causes of war among enduring rivals, arguing that identical explanatory variables influence the decision to become rivals and the decision to go to war. Their results contradict all of the findings in the existing literature on the causes of war among rivals that they consider.[27] I argue that there are good theoretical reasons to believe that ρ in their case is large and positive. Using my estimator, I obtain estimates that are almost entirely contrary to those of Lemke and Reed on the causes of war among rivals and are consistent with the previous literature.

6 Conclusion

This article provides a consistent, asymptotically normal, maximum-likelihood estimator for use in selection models when the same set of independent variables affect both the selection equation and the equation of interest and when both dependent variables are binary. The estimator identifies from an assumption that the error terms in the selection equation and the outcome equation are identical for a given observation. It fills a need in quantitative political science research because nonrandom selection is ubiquitous and because the existing, Heckman-type models are best used only with an exclusion restriction—when the

[27] Lemke and Reed's results do not contradict the findings in the existing literature they consider on the causes of rivalry.

researcher believes that at least one variable that influences selection does not influence the subsequent process of interest. In many substantively interesting problems, the same factors affect both processes. The new estimator's identifying assumption is likely to be reasonable under exactly this circumstance. When the assumption is reasonable, as long as both dependent variables are binary, this model avoids the unfortunate choice among identifying from functional form alone, adding a theoretically unjustified variable in a mistaken attempt to "boost" identification, and giving up on estimation entirely.

The simulations confirm that this estimator is a better solution to the problem of identical explanatory factors than either of the common alternatives for estimation—identifying from functional form alone or adding an "extra" variable to the selection equation when one is skeptical that it belongs there. First, this estimator is a better choice than the Heckman estimator when identical explanatory variables affect selection and the dependent variable of interest and the sample size is small. The proposed estimator has lower mean bias and lower mean square error for the slope estimates in the equation of interest—not only when its own identifying assumption is close to being met, but also when its assumption is fairly far from the truth (the true ρ is only .5) and Heckman's identifying assumption is met. It also provides more accurate 95% confidence intervals, thus decreasing the chances of incorrect inference. Second, dredging up an exclusion restriction is a poor solution to this problem. Even if the added variable actually does belong in the equation, that variable must have quite a large effect on selection for the Heckman estimates to have lower RMSE than mine.

As sample size approaches infinity, *and if the errors truly are normally distributed*, the Heckman estimator is the best choice. It is consistent and identified, though its identification is weak. In small samples, it does not have the information needed to distinguish between the effects it is trying to estimate; very large samples provide the needed information.

The choice of estimators with small-to-medium samples should depend on political theory. When theory points to identical explanatory variables but very little selection bias, ordinary probit is the best choice of the three estimators. When theory suggests identical explanatory variables and a high degree of selection bias, but not the sign of the correlation between the errors (a situation likely to be rare), the Heckman estimator is the best choice, although still a poor option. However, when theory points to identical explanatory variables and implies that the errors are correlated in a particular direction, the new estimator gives considerably better results than either probit or the Heckman-type estimator. The new estimator is therefore a badly needed tool for applied researchers who believe that identical explanatory variables influence selection and a later binary dependent variable of interest.

Appendix: Proofs

Maximum-likelihood estimators usually have several nice properties: consistency, asymptotic normality, and asymptotic efficiency. However, not every maximum-likelihood estimator has these properties (Amemiya 1985, ch. 4). In this appendix, I prove that the estimator derived in this article is consistent and asymptotically normal. Asymptotic efficiency follows from asymptotic normality (Amemiya 1985, p. 123). My proofs build on proofs of the consistency and asymptotic normality of the parameter estimates in a probit model given by Amemiya (1985, pp. 271–273).

Notation: Let $\boldsymbol{\gamma}_0$ and $\boldsymbol{\beta}_0$ be the true parameter values, P_{ij} or P_{ijk} be the probability that $Y_{ijk} = 1$ (with k used only when necessary to clarify the distinction between P_{i21} and P_{i22}), P_{ijk0} be the value of P_{ijk} at the true values of the parameters, $\boldsymbol{\theta}$ be the set $\{\boldsymbol{\gamma}, \boldsymbol{\beta}\}$, $\boldsymbol{\theta}_0$ be the set $\{\boldsymbol{\gamma}_0, \boldsymbol{\beta}_0\}$, and $\mathbf{x}_i$ be a row vector containing the values of the explanatory variables for observation i.

A.1 *Consistency*

I make three assumptions about the model:

Assumption 1. The parameter space Θ is an open-bounded subset of Euclidian K-space.

Assumption 2. $\{\mathbf{x}_i\}$ are uniformly bounded in i and $\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\mathbf{x}_i'\mathbf{x}_i$ is a finite non-singular matrix. The empirical distribution of $\{\mathbf{x}_i\}$ converges to a distribution function.

Assumption 3. $\text{prob}(\boldsymbol{\beta}_0'\mathbf{x}_i = \boldsymbol{\gamma}_0'\mathbf{x}_i) = 0$.

The statement in Assumption 3 is true as long as at least one independent variable is continuous without any mass points and $\boldsymbol{\beta}_0 \neq \boldsymbol{\gamma}_0$. Alternatively, it may be true due to bounds on the independent variables. If all independent variables are dichotomous, the statement restricts the applicability of the model to data-generating processes that meet the assumption.

That the statement is ever true without assumption is not intuitive, because the equation $\boldsymbol{\beta}_0'\mathbf{x}_i = \boldsymbol{\gamma}_0'\mathbf{x}_i$ has an infinite number of solutions as long as it contains more than one independent variable. It is true for continuous $\mathbf{x}$ without mass points for the following reason. The equation in the assumption defines a hyperplane. If there are k independent variables ($\mathbf{x}$s), the hyperplane has dimension $k-1$. The space of $\mathbf{x}$s over which solutions might be found has k dimensions. Thus, the probability of finding a solution is zero even though the number of solutions is infinite.

Theorem 1. *Given Assumptions* 1–3, $\hat{\boldsymbol{\beta}}$ *and* $\hat{\boldsymbol{\gamma}}$ *as defined in* (19) *are consistent estimates of* $\boldsymbol{\gamma}_0$ *and* $\boldsymbol{\beta}_0$.[28]

Proof: I prove three lemmas that lead to the theorem.

Without loss of generality, I divide the data into two parts. The first part, observations $i = 1, \ldots, m$, consists of observations such that $\Phi(-\boldsymbol{\beta}'\mathbf{x}_i) > \Phi(-\boldsymbol{\gamma}'\mathbf{x}_i)$. The second part, observations $i = m+1, \ldots, n$, consists of observations such that $\Phi(-\boldsymbol{\beta}'\mathbf{x}_i) \leq \Phi(-\boldsymbol{\gamma}'\mathbf{x}_i)$. I now make a fourth assumption about the model.

Assumption 4. The number of observations m and the number $n-m$ go to infinity at the same rate as n. The fractions $\frac{m}{n}$ and $\frac{n-m}{n}$ converge to fractions μ_1 and μ_2, where $0 < \mu_1 \leq 1$ and $0 \leq \mu_2 < 1$.

Let $Q_n(\mathbf{y}, \theta) \equiv L^*$. Then, from the text,

$$Q_n(\mathbf{y}, \theta) = \sum_{i=1}^{n} Y_{i0} \ln \Phi(-\boldsymbol{\gamma}'\mathbf{x}_i) + \sum_{i=1}^{m} Y_{i1} \ln[\Phi(\boldsymbol{\gamma}'\mathbf{x}_i) - \Phi(\boldsymbol{\beta}'\mathbf{x}_i)] + \sum_{i=1}^{m} Y_{i2} \ln \Phi(\boldsymbol{\beta}'\mathbf{x}_i) + \sum_{i=m+1}^{n} Y_{i2} \ln \Phi(\boldsymbol{\gamma}'\mathbf{x}_i).$$

[28] The theorem that I use (4.1.2 in Amemiya 1985) is a standard theorem for proving consistency; for example, Amemiya uses it to show consistency of the estimates obtained through probit estimation (Amemiya 1985, pp. 270–273). Technically, the theorem states that one of the local maxima is consistent and does not specify which one if there are multiple local maxima (Amemiya 1985, p. 111). However, as long as $Q(\theta)$ (the function to which $n^{-1}L^*$ converges in probability in the proofs below) has a unique global maximum, the conditions of Theorem 4.1.1 (Amemiya 1985, pp. 106–107) also are met and $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ as defined in (19) are consistent estimates of $\boldsymbol{\gamma}_0$ and $\boldsymbol{\beta}_0$.

Lemma 1. *$\partial Q_n(\mathbf{y}, \theta)/\partial\theta$ exists and is continuous in an open neighborhood $N_1(\theta_0)$ of θ_0.*

Proof: This lemma follows from the normality of the cumulative densities in P_{ij}. The partial derivatives of the likelihood are not continuous at $\Phi(\boldsymbol{\beta}'\mathbf{x}_i) = \Phi(\boldsymbol{\gamma}'\mathbf{x}_i)$. Such an event has probability of 0 in an open neighborhood of the true parameters by Assumption 3 and thus does not affect consistency (Huber 1967; Pakes and Pollard 1989). □

Lemma 2. *Let $g_{ijk}(\mathbf{y}, \theta) \equiv (Y_{ij} - P_{ijk0}) \ln P_{ijk}$. Then $n^{-1}\sum_{i=1}^{n} g_{i0}(\mathbf{y}, \theta)$ converges to 0; $m^{-1}\sum_{i=1}^{m} g_{i1}(\mathbf{y}, \theta)$ converges to 0; $m^{-1}\sum_{i=1}^{m} g_{i21}(\mathbf{y}, \theta)$ converges to 0; and $(n - m)^{-1}\sum_{i=m+1}^{n} g_{i22}(\mathbf{y}, \theta)$ converges to 0. In addition, there exists an open neighborhood of the two parameter vectors, $N_2(\theta_0)$, such that $n^{-1}Q_n(\mathbf{y}, \theta)$ converges to a nonstochastic function $Q(\theta)$, where*

$$Q(\theta) = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} P_{i00}\ln P_{i0} + \lim_{m\to\infty} \frac{m}{n} m^{-1}\sum_{i=1}^{m}(P_{i10}\ln P_{i1} + P_{i210}\ln P_{i21}) + \lim_{(n-m)\to\infty} \frac{n-m}{n}(n-m)^{-1}\sum_{i=m+1}^{n} P_{i220}\ln P_{i22}. \tag{A1}$$

Proof:

1. The normality of the cumulative densities in P_{ij} and Assumptions 1–3 imply that $g_{ijk}(\mathbf{y}, \theta)$ satisfies the conditions for g_t in Theorem 4.2.2 in Amemiya (1985, p. 117) in a compact neighborhood of θ_0.
2. By Theorem 4.2.3 in Amemiya (1985, p. 118), the limits in $Q(\theta)$ exist.
3. By Theorem 4.2.2, $n^{-1}\sum_{i=1}^{n} g_{i0}(\mathbf{y}, \theta)$, $\mathrm{m}^{-1}\sum_{i=1}^{m} g_{i1}(\mathbf{y}, \theta)$, $\mathrm{m}^{-1}\sum_{i=1}^{m} g_{i21}(\mathbf{y}, \theta)$, and $(n-m)^{-1}\sum_{i=m+1}^{n} g_{i22}(\mathbf{y}, \theta)$ converge uniformly to 0 in probability when $\theta \in \boldsymbol{\Theta}$.
4. By the previous step, $n^{-1}\sum_{i=1}^{n} Y_{i0}\ln P_{i0} - n^{-1}\sum_{i=1}^{n} P_{i00}\ln P_{i0}$ converges to 0, so that $n^{-1}\sum_{i=1}^{n} Y_{i0}\ln P_{i0}$ converges uniformly to $n^{-1}\sum_{i=1}^{n} P_{i00}\ln P_{i0}$ in probability when $(\boldsymbol{\beta}, \boldsymbol{\gamma}) \in N(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)$, an open neighborhood of θ_0.

   Analogous reasoning applies to the other terms in the empirical likelihood. Along with Assumption 4, Steps 1–4 imply that $n^{-1}Q_n(\mathbf{y}, \theta)$ converges to $Q(\theta)$. □

Lemma 3. *$Q(\theta)$ attains a strict local maximum at $\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0$.*

Proof: I first differentiate $Q(\theta)$ in an open neighborhood of the true parameter values to show that a critical point occurs at $\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0$. The assumptions allow differentiation inside the limit operation (Amemiya 1985, p. 271). Without loss of generality, I again divide the data into two parts. The first part, observations $i = 1, \ldots, m$, consists of observations such that $\Phi(-\boldsymbol{\beta}_0'\mathbf{x}_i) > \Phi(-\boldsymbol{\gamma}_0'\mathbf{x}_i)$. The second part, observations $i = m + 1, \ldots, n$, consists of observations such that $\Phi(-\boldsymbol{\beta}_0'\mathbf{x}_i) \leq \Phi(-\boldsymbol{\gamma}_0'\mathbf{x}_i)$. The proof is easier to follow with abbreviated notation. For example, let $F_{i\beta 0}$ represent the cumulative density of the normal distribution evaluated at $\boldsymbol{\beta}_0'\mathbf{x}_i$ and $f_{i\beta 0}$ be its first derivative. $F_{i\beta}$ is the cumulative normal evaluated at $\boldsymbol{\beta}'\mathbf{x}_i$, where $\boldsymbol{\beta}$ need not be the true value, and $f_{i\beta}$ is its first derivative.

Then,

$$\frac{\partial Q}{\partial \boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}} \lim_{n\to\infty} n^{-1}\left(\sum_{i=1}^{m}[(1-F_{i\gamma 0})\ln(1-F_{i\gamma}) + (F_{i\gamma 0}-F_{i\beta 0})\ln(F_{i\gamma}-F_{i\beta}) + F_{i\beta 0}\ln F_{i\beta}] + \sum_{i=m+1}^{n}[(1-F_{i\gamma 0})\ln(1-F_{i\gamma}) + F_{i\gamma 0}\ln F_{i\gamma}]\right)$$
$$= \lim_{n\to\infty} n^{-1}\sum_{i=1}^{m}\left[\frac{-(F_{i\gamma 0}-F_{i\beta 0})}{F_{i\gamma}-F_{i\beta}} f_{i\beta}\mathbf{x} + \frac{F_{i\beta 0}}{F_{i\beta}} f_{i\beta}\mathbf{x}\right].$$

Note that for observations $i = 1, \ldots, m$, $\ln(F_{i\gamma} - F_{i\beta})$ is always strictly positive and thus is always defined. At $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$, the numerator and denominator of each fraction in the summation sign cancel to 1. Thus, $\frac{\partial Q}{\partial \boldsymbol{\beta}} = 0$ at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$.

A similar exercise shows that $\frac{\partial Q}{\partial \boldsymbol{\gamma}} = 0$ at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$. Thus, $\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0$ is a critical point of Q.

Next, I show that Q attains a strict local maximum at $\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0$. The Hessian of Q is

$$\mathbf{H} = \begin{pmatrix} \frac{\partial^2 Q}{\partial \boldsymbol{\beta}_1^2} & \cdots & \frac{\partial^2 Q}{\partial \boldsymbol{\beta}_c \partial \boldsymbol{\beta}_1} & \frac{\partial^2 Q}{\partial \boldsymbol{\gamma}_1 \partial \boldsymbol{\beta}_1} & \cdots & \frac{\partial^2 Q}{\partial \boldsymbol{\gamma}_c \partial \boldsymbol{\beta}_1} \\ \vdots & \ddots & & & & \vdots \\ \frac{\partial^2 Q}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_c} & & \frac{\partial^2 Q}{\partial \boldsymbol{\beta}_c^2} & & & \\ \frac{\partial^2 Q}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\gamma}_1} & & & \frac{\partial^2 Q}{\partial \boldsymbol{\gamma}_1^2} & & \\ \vdots & & & & \ddots & \\ \frac{\partial^2 Q}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\gamma}_c} & \cdots & & \frac{\partial^2 Q}{\partial \boldsymbol{\gamma}_1 \partial \boldsymbol{\gamma}_c} & \cdots & \frac{\partial^2 Q}{\partial \boldsymbol{\gamma}_c^2} \end{pmatrix} \tag{A2}$$

where c is the number of independent variables. At the true values of the parameters, each of the second partials with respect to β_j, β_k or γ_j, γ_k can be written as the same constant (one for the βs and one for the γs) multiplied by $x_{ij}x_{ik}$. For example,

$$\frac{\partial^2 Q}{\partial \beta_j^2} = -\lim_{n\to\infty} n^{-1}\sum_{i=1}^{m}\frac{F_{i\gamma 0}}{F_{i\beta 0}(F_{i\gamma 0}-F_{i\beta 0})} f_{\beta 0}^2 x_{ij}^2$$

and

$$\frac{\partial^2 Q}{\partial \beta_j \partial \beta_k} = -\lim_{n\to\infty} n^{-1}\sum_{i=1}^{m}\frac{F_{i\gamma 0}}{F_{i\beta 0}(F_{i\gamma 0}-F_{i\beta 0})} f_{\beta 0}^2 x_{ij}x_{ik},$$

where j, k denote arbitrary parameters in the vector $\boldsymbol{\beta}$.

For a given observation i, I could rewrite the Hessian for an observation at the true values of the parameters as the Kronecker product of two matrices:

$$\begin{pmatrix} a & b \\ b & d \end{pmatrix} \otimes \mathbf{x}_i'\mathbf{x}_i. \tag{A3}$$

The problem with doing so is that the terms in $\frac{\partial^2 Q}{\partial \boldsymbol{\beta}^2}$ and $\frac{\partial^2 Q}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\beta}}$ (a and b above) rely only upon observations $1, \ldots, m$. Some terms in $\frac{\partial^2 Q}{\partial \boldsymbol{\gamma}^2}$ rely on observations $1, \ldots, m$ and others

rely upon $m+1,\ldots,n$. Thus, I divide the Hessian into two parts. $\tilde{\mathbf{H}}$ contains the terms involving the first set of observations, and $\mathbf{H}''$ the others. Then,

$$\mathbf{H} = \tilde{\mathbf{H}} + \mathbf{H}''. \tag{A4}$$

To examine the definiteness of $\tilde{\mathbf{H}}$ and $\mathbf{H}''$, let $\tilde{\mathbf{H}}_i$ be the Hessian for one arbitrary observation i, where $1 \le i \le m$. Let $\mathbf{H}_i''$ be the Hessian for one arbitrary observation i, where $m+1 \le i \le n$. Then

$$\tilde{\mathbf{H}}_i = \begin{pmatrix} a & b \\ b & d' \end{pmatrix} \otimes \mathbf{x}_i'\mathbf{x}_i, \quad i \in \{1,\ldots,m\} \tag{A5}$$

and

$$\mathbf{H}_i'' = \begin{pmatrix} 0 & 0 \\ 0 & d'' \end{pmatrix} \otimes \mathbf{x}_i'\mathbf{x}_i, \quad i \in \{m+1,\ldots,n\}. \tag{A6}$$

Evaluated at $\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0$,

$$a = -\frac{F_{i\gamma 0}}{F_{i\beta 0}(F_{i\gamma 0} - F_{i\beta 0})} f_{i\beta 0}^2, \tag{A7}$$

$$b = \frac{1}{F_{i\gamma 0} - F_{i\beta 0}} f_{i\beta 0} f_{i\gamma 0}, \tag{A8}$$

$$d' = -\frac{1}{1 - F_{i\gamma 0}} f_{i\gamma 0}^2 - \frac{1}{F_{i\gamma 0} - F_{i\beta 0}} f_{i\gamma 0}^2, \tag{A9}$$

and

$$d'' = -\frac{1}{1 - F_{i\gamma 0}} f_{i\gamma 0}^2 - \frac{1}{F_{i\gamma 0}} f_{i\gamma 0}^2. \tag{A10}$$

The equations above express $\tilde{\mathbf{H}}_i$ and $\mathbf{H}_i''$ each as the Kronecker product of two matrices. To determine the definiteness of $\mathbf{H}$, I first examine $\tilde{\mathbf{H}}$ and $\mathbf{H}''$ separately. The left-hand matrix in $\tilde{\mathbf{H}}_i$ is negative definite because $a < 0$ and $ad' - b^2 > 0$. The data matrix is positive definite.

The definiteness of the two Kronecker-multiplied matrices determines the definiteness of the product matrix. The characteristic roots of $\mathbf{A} \otimes \mathbf{B}$ are $\lambda_i \mu_i$, where λ_i are the characteristic roots of $\mathbf{A}$ and μ_i are the roots of $\mathbf{B}$ (Bellman 1960, p. 227). By definition, a matrix is positive definite if it has all positive characteristic roots, whereas a matrix is negative definite if it has all negative characteristic roots (Greene 1993, pp. 35–36). Thus, the Kronecker product of a negative definite matrix and a positive definite matrix is negative definite. Thus, $\tilde{\mathbf{H}}_i$ is negative definite for all i. By similar reasoning, $\mathbf{H}_i''$ is negative semidefinite for all i. (The first component has one zero root and one negative root, while the data matrix again has all positive roots.)

Note that $\tilde{\mathbf{H}} = \lim_{m\to\infty} \frac{m}{n}\frac{1}{m}\sum_{i=1}^{m} \tilde{\mathbf{H}}_i$ and $\mathbf{H}'' = \lim_{(n-m)\to\infty} \frac{(n-m)}{n}\frac{1}{(n-m)}\sum_{i=m+1}^{n} \mathbf{H}_i''$. $\tilde{\mathbf{H}}$ is negative definite for the following reason: We have seen that $\tilde{\mathbf{H}}_i$ is negative definite for an arbitrary observation, i. Let $\mathbf{Z}_\eta = \frac{1}{\eta}\sum_{i=1}^{\eta} \tilde{\mathbf{H}}_i$. Then $\mathbf{Z}_{\eta+1} = \frac{\eta}{\eta+1}\mathbf{Z}_\eta + \frac{1}{\eta+1}\tilde{\mathbf{H}}_{\eta+1}$. Multiplying a negative definite matrix by a positive constant results in a negative definite

matrix, and adding two negative definite matrices results in a negative definite matrix. Thus $\tilde{\mathbf{H}}$ is negative definite. It is also straightforward to show that $\mathbf{H}''$ is negative semidefinite.

When we add together $\tilde{\mathbf{H}}$ and $\mathbf{H}''$, a negative definite and a negative semidefinite matrix, to form the Hessian, we get a matrix that is negative definite. Because the Hessian of $Q(\theta)$ is negative definite, $Q(\theta)$ attains a strict local maximum at β_0, γ_0 (Simon and Blume 1994, p. 399).[29]

Together with Theorem 4.1.2 in Amemiya (1985, p. 110), Lemmas 1–3 imply Theorem 1. □

A.2 *Asymptotic Normality*

Theorem 2. *Given Assumptions* 1–4, $\hat{\beta}$ *and* $\hat{\gamma}$ *as defined in* (19) *are asymptotically normally distributed with respective means* β_0 *and* γ_0 *and variances* $\mathbf{A}^{-1}$ *and* $\mathbf{C}^{-1}$, *where*

$$\mathbf{A} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{m} \frac{F_{i\gamma 0}}{F_{i\beta 0}(F_{i\gamma 0} - F_{i\beta 0})} f_{i\beta 0}^2 \mathbf{x}_i' \mathbf{x}_i, \tag{A11}$$

$$\mathbf{C} = \lim_{n\to\infty} \frac{1}{n} \left(\sum_{i=1}^{n} \frac{1}{1 - F_{i\gamma 0}} f_{i\gamma 0}^2 \mathbf{x}_i' \mathbf{x}_i + \sum_{i=1}^{m} \frac{1}{F_{i\gamma 0} - F_{i\beta 0}} f_{i\gamma 0}^2 \mathbf{x}_i' \mathbf{x}_i + \sum_{i=m+1}^{n} + \frac{1}{F_{i\gamma 0}} f_{i\gamma 0}^2 \mathbf{x}_i' \mathbf{x}_i \right). \tag{A12}$$

Proof: I prove four additional lemmas which lead to the theorem.

Lemma 4. $\partial^2 Q_n / \partial\theta\partial\theta'$ *exists and is continuous in an open, convex neighborhood of* θ_0.

Proof: This lemma follows from the normality of the cumulative densities in P_{ijk} and from Assumption 3. The partial derivatives are not continuous at $\Phi(\beta'\mathbf{x}_i) = \Phi(\gamma'\mathbf{x}_i)$, which occurs with probability zero by Assumption 3 and is therefore not a problem (Huber 1967; Pakes and Polland 1989). □

Lemma 5. $n^{-1}(\partial^2 Q_n / \partial\theta\partial\theta')_{\theta_*}$ *converges to a finite nonsingular matrix* $\mathbf{G}(\theta_0) = \lim E n^{-1}(\partial^2 Q_n / \partial\theta\partial\theta')_{\theta_0}$ *in probability for any sequence* θ_n *such that* $\operatorname{plim} \theta_n = \theta_0$.

Proof: This result follows from the normality of the terms in L^*, Assumptions 1–4, and Theorems 4.1.5 and 4.2.2 in Amemiya (1985, pp. 113, 117). □

Lemma 6. $n^{-1/2}(\partial Q_n / \partial \beta_j)_{\beta_{j_0}} \to N[\mathbf{0}, \mathbf{A}]$ *and* $n^{-1/2}(\partial Q_n / \partial \gamma_j)_{\gamma_{j_0}} \to N[\mathbf{0}, \mathbf{C}]$.

Proof: I divide $Q_n = L^*$ into observations $i = 1, \ldots, m$ and $i = m+1, \ldots, n$ as in the proof of consistency. Then,

$$\frac{1}{\sqrt{n}} \frac{\partial L^*}{\partial \beta}\bigg|_{\beta=\beta_0, \gamma=\gamma_0} = \frac{1}{\sqrt{n}} \sum_{i=1}^{m} \left[\frac{Y_{i21}(F_{i\gamma 0} - F_{i\beta 0}) - Y_{i1} F_{i\beta 0}}{F_{i\beta 0}(F_{i\gamma 0} - F_{i\beta 0})} \right] f_{i\beta 0} \mathbf{x}_i \tag{A13}$$

[29] If there are no observations such that $i = m+1, \ldots, n$, then $\mathbf{H} = \tilde{\mathbf{H}}$, which is negative definite.

and

$$\frac{1}{\sqrt{n}}\frac{\partial L^*}{\partial \boldsymbol{\gamma}}\bigg|_{\beta=\beta_0,\gamma=\gamma_0} = \frac{1}{\sqrt{n}}\left[\sum_{i=1}^{n}\frac{-Y_{i0}}{(1-F_{i\gamma 0})} + \sum_{i=1}^{m}\frac{Y_{i1}}{(F_{i\gamma 0}-F_{i\beta 0})} + \sum_{i=m+1}^{n}\frac{Y_{i22}}{F_{i\gamma 0}}\right] f_{i\gamma 0}\mathbf{x}_i. \tag{A14}$$

The Liapounov Central Limit Theorem (Amemiya 1985, p. 92) and the four assumptions together imply the asymptotic normality of the terms in (A13) and (A14) and thus the lemma, where

$$\mathbf{A} = \lim_{n\to\infty} \operatorname{var}\left(\frac{1}{\sqrt{n}}\frac{\partial L^*}{\partial \boldsymbol{\beta}}\bigg|_{\beta=\beta_0,\gamma=\gamma_0}\right) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{m}\frac{F_{i\gamma 0}}{F_{i\beta 0}(F_{i\gamma 0}-F_{i\beta 0})} f_{i\beta 0}^2\mathbf{x}_i'\mathbf{x}_i, \tag{A15}$$

$$\begin{aligned}\mathbf{C} &= \lim_{n\to\infty} \operatorname{var}\left(\frac{1}{\sqrt{n}}\frac{\partial L^*}{\partial \boldsymbol{\gamma}}\bigg|_{\beta=\beta_0,\gamma=\gamma_0}\right) \\ &= \lim_{n\to\infty}\frac{1}{n}\left(\sum_{i=1}^{n}\frac{1}{1-F_{i\gamma 0}} f_{\gamma 0}^2\mathbf{x}_i'\mathbf{x}_i + \sum_{i=1}^{m}\frac{1}{F_{i\gamma 0}-F_{i\beta 0}} f_{\gamma 0}^2\mathbf{x}_i'\mathbf{x}_i + \sum_{i=m+1}^{n} + \frac{1}{F_{i\gamma 0}} f_{\gamma 0}^2\mathbf{x}_i'\mathbf{x}_i\right).\end{aligned} \tag{A16}$$

□

Lemma 7. $\mathbf{G}(\boldsymbol{\beta}_{j0}) = -\mathbf{A}$ *and* $\mathbf{G}(\boldsymbol{\gamma}_{j0}) = -\mathbf{C}$.

Remember that $\mathbf{G}(\theta_0) = \lim En^{-1}(\partial^2 Q_n/\partial\theta\partial\theta')_{\theta_0}$. The second derivatives of the likelihood function in an open neighborhood of the true parameters are

$$\frac{\partial^2 L^*}{\partial^2 \boldsymbol{\beta}} = \sum_{i=1}^{m}\left\{-\frac{Y_{i21}(F_{i\gamma}-F_{i\beta})^2 + Y_{i1}F_{i\beta}^2}{[F_{i\beta}(F_{i\gamma}-F_{i\beta})]^2} f_\beta^2\mathbf{x}_i'\mathbf{x}_i + \frac{Y_{i21}(F_{i\gamma}-F_{i\beta}) - Y_{i1}F_{i\beta}}{F_{i\beta}(F_{i\gamma}-F_{i\beta})} f_\beta'\mathbf{x}_i'\mathbf{x}_i\right\}; \tag{A17}$$

$$\begin{aligned}\frac{\partial^2 L^*}{\partial^2 \boldsymbol{\gamma}} = \sum_{i=1}^{m}&\left\{\left[\frac{-Y_{i1}}{(F_{i\gamma}-F_{i\beta})^2} + \frac{Y_{i0}}{(1-F_{i\gamma})^2}\right] f_\gamma^2\mathbf{x}_i'\mathbf{x}_i + \left[\frac{Y_{i1}}{(F_{i\gamma}-F_{i\beta})} - \frac{Y_{i0}}{(1-F_{i\gamma})}\right] f_\gamma'\mathbf{x}_i'\mathbf{x}_i\right\} \\ &+ \sum_{i=m+1}^{n}\left\{\left[\frac{-Y_{i2}}{F_{i\gamma}^2} + \frac{Y_{i0}}{(1-F_{i\gamma})^2}\right] f_\gamma^2\mathbf{x}_i'\mathbf{x}_i + \left[\frac{Y_{i2}}{F_{i\gamma}} - \frac{Y_{i0}}{(1-F_{i\gamma})}\right] f_\gamma'\mathbf{x}_i'\mathbf{x}_i\right\}.\end{aligned} \tag{A18}$$

From Eqs. (A17) and (A18), one can calculate that

$$\lim_{n\to\infty}\frac{1}{n}E\frac{\partial^2 L^*}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}}\bigg|_{\beta=\beta_0,\gamma=\gamma_0} = -\mathbf{A}, \tag{A19}$$

$$\lim_{n\to\infty}\frac{1}{n}E\frac{\partial^2 L^*}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}}\bigg|_{\beta=\beta_0,\gamma=\gamma_0} = -\mathbf{C}. \tag{A20}$$

Using Theorem 4.2.4 in Amemiya (1985, p. 121), the four new lemmas imply Theorem 2:

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to \mathbf{N}(\mathbf{0}, \mathbf{A}^{-1})$$

and

$$\sqrt{n}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \rightarrow \mathbf{N}(\mathbf{0}, \mathbf{C}^{-1}). \qquad \square$$

References

Achen, C. H. 1986. *The Statistical Analysis of Quasi-Experiments*. Berkeley: University of California Press.

Amemiya, T. 1985. *Advanced Econometrics*. Cambridge, MA: Harvard University Press.

Bellman, R. 1960. *Introduction to Matrix Analysis*. New York: McGraw-Hill.

Berinsky, A. 1999. "The Two Faces of Public Opinion." *American Journal of Political Science* 43:1209–1230.

Boehmke, F. J. 2003. "Using Auxiliary Data to Estimate Selection Bias Models, with an Application to Interest Groups' Use of the Direct Initiative Process." *Political Analysis* in press.

Brehm, J. 1993. *The Phantom Respondents: Opinion Surveys and Political Representation*. Ann Arbor: University of Michigan Press.

Dubin, J. A., and D. Rivers. 1990. "Selection Bias in Linear Regression, Logit and Probit Models." In *Modern Methods of Data Analysis*, eds. J. Fox, and J. S. Long. Newbury Park, CA: Sage, pp. 410–443.

Greene, W. H. 1993. *Econometric Analysis*, 2nd ed. New York: Macmillan.

Heckman, J. J. 1974. "Shadow Prices, Market Wages, and Labor Supply." *Econometrica* 42:679–694.

Heckman, J. J. 1976. "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models." *The Annals of Economic and Social Measurement* 5(4):475–492.

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

Huber, P. J. 1967. "The Behavior of Maximum-Likelihood Estimates under Nonstandard Conditions." In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. eds. L. M. Le Cam, and J. Neyman. Berkeley: University of California Press, pp. 221–233.

King, G. 1990. *Unifying Political Methodology*. Cambridge: Cambridge University Press.

Lemke, D., and W. Reed. 2001. "War and Rivalry among Great Powers." *American Journal of Political Science* 45:457–469.

Lewis, J. B., and K. A. Schultz. 2003. "Revealing Preferences: Empirical Modeling of a Crisis Bargaining with Incomplete Information." Manuscript.

Maddala, G. 1999. *Limited-dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

Mitchell, N. J., W. L. Hansen, and E. M. Jepsen. 1997. "The Determinants of Domestic and Foreign Corporate Political Activity." *The Journal of Politics* 59:1096–1113.

Pakes, A., and D. Pollard. 1989. "Simulation and the Asymptotics of Optimization Estimators." *Econometrica* 57:1027–1057.

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

Sartori, A. E. 2002. "Enduring Facts about Enduring Rivals." Presented at the Annual Meeting of the American Political Science Association, Boston.

Signorino, C. S. 1999. "Strategic Interaction and the Statistical Analysis of International Conflict." *American Political Science Review* 92:279–297.

Signorino, C. S. 2002. "Strategy and Selection in International Relations." *International Interactions* 28:93–115.

Simon, C. P., and L. Blume. 1994. *Mathematics for Economists*. New York: W. W. Norton.

Smith, A. 1998. "A Summary of Political Selection: The Effect of Strategic Choice on the Escalation of International Crises." *American Journal of Political Science* 42:698–702.

Van de Ven, W. P., and B. Van Praag. 1981. "The Demand for Deductibles in Private Health Insurance." *Journal of Econometrics* 17:229–252.