# Bucking the System: Using Simulation Methods to Estimate and Analyze Systems of Equations with Qualitative and Limited Dependent Variables*

Jude C. Hays (jchays@illinois.edu)
Assistant Professor of Political Science, University of Illinois, Urbana-Champaign

Prepared for Presentation at the St. Louis Area Methods Meeting (SLAMM)
April 24, 2009
*Preliminary Draft*

**ABSTRACT:** Quantitative empirical research in political science frequently involves the analysis of interdependent outcomes. The endogenous variables that we want to explain are dependent on other endogenous variables, either outcomes observed for other units, other variables, or both. In some instances, this interdependence is viewed as a nuisance and great effort is taken to purge the data of the connections, but much of the time, the interdependence itself is an important part of the politics that we aim to study. Unfortunately, empirical models that incorporate political and other forms of interdependence can be difficult to estimate and use for analytical purposes. This is particularly true when the endogenous variables are qualitative or limited in some way. Not surprisingly, the available methods are rarely used and remain underdeveloped. I discuss one way to use simulation methods to estimate and analyze simultaneous equation models with limited and qualitative dependent variables. I illustrate these techniques with an application to WWI participation and entry timing decisions.

*This paper is based on joint methodological work with Rob Franzese and Aya Kachi. The application to WWI participation and entry timing is based on work with Sang-hyun Chi , Paul Diehl, Colin Flint, Jürgen Scheffran, and John Vasquez. The errors are my own.

Quantitative empirical research in political science frequently involves the analysis of interdependent outcomes. The endogenous variables that we want to explain are dependent on other endogenous variables—outcomes observed for other units, other variables, or both. In some applications, this interdependence is viewed as a nuisance and great effort is taken to purge it from the data, but much of the time, the interdependence itself is an important part of the politics that we aim to study. There are numerous examples of substantively important interdependence among political actors. The strategies of candidates in electoral contests depend on the strategies of their opponents; individual legislators' votes depend on others' votes or expected votes; and states' entry decisions with respect to wars, alliances, and international organizations depend on the decisions made by other states. Globalization implies strategic interdependence in national-level macroeconomic policymaking. There are equally many examples of interdependence among endogenous variables. Perhaps, the best known involve the simultaneity of politics, policy, and institutions. The political approval of politicians, for instance, determines the likelihood that they adopt particular sets of policies and undertake institutional reforms at the same time that the policies and reforms affect the popularity of the politicians who choose to implement them. These are all examples of politically important simultaneous relationships. Unfortunately, empirical models that incorporate this kind interdependence can be difficult to estimate and use for analytical purposes. This is particularly true when the endogenous variables are qualitative or limited in some way. Not surprisingly, the available methods are rarely used and remain underdeveloped.

This paper explores one way of using simulations to estimate and analyze simultaneous equation models with limited and qualitative dependent variables. To illustrate, I model the decisions of states to enter WWI. One can treat these decisions as independent and driven by domestic and international structural factors such as regime type, trade dependence, and relative military capabilities, but this is approach unsatisfactory. In the end, each state's decision is heavily influence by the entry decisions of others, and many of the more interesting counterfactual questions address this interdependence specifically. How did America's decision to participate affect the participation of other states? To what extent did Italy's decision to enter the war in mid-1915 affect the probability that Bulgaria and Romania would be drawn into the conflict before the fighting stopped?

The paper is organized as follows. In the first section, I cover familiar ground, describing the basic approach to estimating simultaneous equation models with continuous dependent variables. I highlight the key differences between the continuous and qualitative and limited dependent variable cases in the second section, using the simultaneous probits and censored simultaneous duration models as examples. In the third section, I discuss recursive importance sampling as a method for estimating these models. I explore the possibility of using simulations for counterfactual analysis in the fourth section. I illustrate these techniques in an analysis of WWI participation and entry timing decisions in the fifth section and conclude in the sixth.

## I. Estimating Systems of Equations with Continuous Dependent Variables

I begin with simultaneous equations models for continuous dependent variables. This framework provides the basis for everything that follows. In matrix notation, the system takes the form

$$\mathbf{y} = \mathbf{A}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \quad (1)$$

where $\mathbf{y}$ is an $NP \times 1$ vector of observations on $N$ units and $P$ endogenous variables. The matrix $\mathbf{A}$ contains values that express the interdependence among units, endogenous variables, or both;

the $NP \times K$ matrix $\mathbf{X}$ contains the independent variables; and the column-vector $\boldsymbol{\beta}$ contains coefficients.[1] The $NP \times 1$ vector $\mathbf{u}$ contains i.i.d. disturbances. The reduce form of this model is

$$\begin{aligned}\mathbf{y} &= (\mathbf{I} - \mathbf{A})^{-1}\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \mathbf{A})^{-1}\mathbf{u} \\ &= \boldsymbol{\Gamma}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Gamma}\mathbf{u} \\ &= \boldsymbol{\Gamma}\mathbf{X}\boldsymbol{\beta} + \mathbf{v}\end{aligned} \tag{2}$$

The basic econometric problem is that the reduced form disturbances in $\mathbf{v}$ are interdependent and heteroscedastic. The likelihood is derived using the multivariate version of the change of variables theorem, which requires only the inverse function for $\mathbf{u}$ and the determinant of the corresponding Jacobian matrix. These are

$$\mathbf{u} = g^{-1}(\mathbf{y}) = \boldsymbol{\Gamma}^{-1}\mathbf{y} - \mathbf{X}\boldsymbol{\beta} \tag{3}$$

and

$$|\mathbf{J}| = \left| \begin{pmatrix} \frac{\partial u_{11}}{\partial y_{11}} & \cdots & \frac{\partial u_{NP}}{\partial y_{11}} \\ \vdots & \ddots & \vdots \\ \frac{\partial u_{11}}{\partial y_{NP}} & \cdots & \frac{\partial u_{NP}}{\partial y_{NP}} \end{pmatrix} \right| = |\mathbf{I} - \mathbf{A}| \tag{4}$$

respectively, and the log-likelihood is

$$\ln L = \sum_{n=1}^{N} \sum_{p=1}^{P} \ln f(u_{np}) + \ln |\mathbf{I} - \mathbf{A}|. \tag{5}$$

## II. Systems of Equations with Limited and Qualitative Dependent Variables

Analytically transforming an assumed joint density for independent disturbances into a joint

[1] When the system contains multiple endogenous variables (P>1) and multiple sets of independent variables, the observations in $\mathbf{y}$ are stacked by the left-hand-side variable (the first *N* observations are for the first endogenous variable, the second *N* observations are for the second endogenous variable, etc.), $\mathbf{X}$ is block diagonal, and *K* is the sum of independent variables in the system as a whole.

density for interdependent disturbances is tractable, making estimation relatively simple. The problem with systems of equations that have limited and qualitative dependent variables is that we do not always observe the continuous realizations of the endogenous variables. Given this censoring problem, it should not come as a surprise that it is more difficult to estimate the interdependence among the endogenous variables. We need to integrate over multivariate distributions, and so the likelihoods contain high-order integrals on the right-hand-side. Simulation can be a useful way to calculate the likelihood in these cases.[2]

Consider the simultaneous probits model. The structural model for the latent endogenous variables takes the form:

$$\mathbf{y}^* = \mathbf{A}\mathbf{y}^* + \mathbf{X}\boldsymbol{\beta} + \mathbf{u}\,, \tag{6),}$$

which reduces to:

$$\mathbf{y}^* = \boldsymbol{\Gamma}\mathbf{X}\boldsymbol{\beta} + \mathbf{v}\;\;,\;\text{with}\;\mathbf{v} = \boldsymbol{\Gamma}\mathbf{u}\,. \tag{7),}$$

The latent-variable $\mathbf{y}^*$ links to the observed binary-outcome, $\mathbf{y}$, through the measurement equation:

$$y_i = \{1 \text{ if } y_i^* > 0\;\;;\;\;0 \text{ if } y_i^* \le 0\}\,. \tag{8).}$$

The probabilities that the $i^{th}$ observations are one are calculated as follows:

$$\begin{aligned} p(y_i = 1 | \mathbf{X}) &= p\left(\left[\boldsymbol{\Gamma}\mathbf{X}\boldsymbol{\beta}\right]_i + \left[\boldsymbol{\Gamma}\mathbf{u}\right]_i > 0\right) \\ &= p\left(v_i < \left[\boldsymbol{\Gamma}\mathbf{X}\boldsymbol{\beta}\right]_i / \sigma_i\right) = \boldsymbol{\Phi}\left\{\left[\boldsymbol{\Gamma}\mathbf{X}\boldsymbol{\beta}\right]_i / \sigma_i\right\} \end{aligned} \tag{9).[3]}$$

---

[2] See Gill's (2002) chapter 8 and Stern's (1997) for excellent introductions to simulation techniques. This approach is typically called maximum simulated likelihood (MSL) estimation.

[3] In the middle step, note that the symmetry about zero of $\mathbf{u}$, and so of $\mathbf{v}$, implies that $\text{pr}\left(-v_i < \text{x}\right) = \text{pr}\left(v_i < \text{x}\right)$ for any x.

As in the standard probit, a cumulative-normal distribution, $\mathbf{\Phi}\{\bullet\}$, gives the marginal probability (integrating over the other $v_{\sim i}$) that the systematic component, $[\mathbf{\Gamma X \beta}]_i / \sigma_i$, exceeds the stochastic component, $v_i$. The disturbances in $\mathbf{u}$ are distributed *multivariate* normal with mean **0** and spherical variance-covariance $\sigma^2\mathbf{I}$, with $\sigma^2$ normalized to 1 as usual for a probit model. However, in the simultaneous equations probit, the interdependence of the $y_i^*$ induces a non-sphericity of the stochastic components; specifically, $\mathbf{v}$ is distributed *multivariate* normal with variance-covariance matrix $\mathbf{\Omega} = \mathbf{\Gamma}'\mathbf{\Gamma}$ (and mean **0**).

To simplify the expression for the likelihood, define $\mathbf{v}^{**} = \mathbf{Q}\mathbf{v}$, where $q_{ij} = 1 - 2y_i$ for $i = j$ and 0 otherwise, and $\mathbf{Z} = \mathbf{Q\Gamma X \beta}$.[4] This allows us to write (8) and (9) in terms of upper bounds only. The likelihood for this model is

$$\ln L = \ln \int_{-\infty}^{\mathbf{Z}} f_n(\mathbf{v}^{**}; 0, \mathbf{\Omega}) dv \tag{10}$$

Next, consider the simultaneous durations model. Without censoring, the log-linear representation of several of the standard models can take the same form as (1), although it is useful to add a shape parameter so that we have

$$\mathbf{y} = \mathbf{A}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \sigma\mathbf{u} \,. \tag{11}$$

If we assume that **u** is distributed type-1 extreme value, we have a Weibull model. If we assume that **u** is distributed as a standard normal, we have a log-normal model. To estimate, we simply transform an assumed joint density for the independent $\mathbf{u}$ into a joint density for the interdependent $\mathbf{v}$.[5] Formally, the log-likelihood without censoring is

[4] Multiplying by **Q** selects the right sign on the systematic component up to which to integrate the distribution of the stochastic component **v**.

[5] Hays and Kachi (2009) focus on the Weibull model. In this paper, I consider the log-normal model

$$\ln L = \sum_{n=1}^{N} \ln f(v_i) \tag{12}$$

With censoring, the likelihood becomes

$$\ln L = \sum_{n=1}^{N} \delta_i \ln\left\{ f(v_i)\right\} + \left(1-\delta_i\right)\ln\left\{S(v_i)\right\} \tag{13}$$

where $S(v)$, the survival function, is $S(v) = 1 - F(v) = 1 - \int_0^v f(v)dv = \Pr(V \geq v)$ and $\delta_i$ is an indicator variable that takes a value of 1 when a failure is observed and 0 otherwise. The integral makes estimation, the topic I turn to next, difficult. I start with a general discussion of sampling methods and then describe two specific estimators for the spatial-lag probit and duration models.

## III. Recursive Importance Sampling

How should we estimate these models? In theory, we could draw vectors of reduced-form disturbances and calculate the joint probabilities (i.e., relative frequencies) of observing any combination of 1's and 0's conditional on the model parameters and data, but this "brute force" approach is not a feasible strategy for estimation. In the case of a binary outcome for *N* units, for example, there are $2^N$ possible combinations. To estimate the underlying probabilities using sample frequencies, we would need considerably more than $2^N$ draws from the distribution of reduced-form disturbances, and to maximize this simulated likelihood, we would need to recalculate the probabilities every time we updated the parameter vector. The computational costs are simply too high to proceed in this way. Moreover, this simulator is not continuous nor is it differentiable with respect to the parameters.[6] The RIS estimator provides a feasible

---

only.

[6] See Lerman and Manski (1981) for this kind of "brute force" strategy.

alternative.

Recursive Importance-Sampling (RIS) uses simulation and decomposition to estimate multivariate integrals that are difficult to calculate analytically, and can be employed to estimate models like those described above. I introduce RIS following Vijverberg's (1997) notation. To approximate an *n*-dimensional cumulative multivariate-normal distribution, e.g.,

$$p = \int_{-\infty}^{\mathbf{x}_0} f_n(\mathbf{x})\, d\mathbf{x}\,, \qquad (14),$$

where $f_n(\mathbf{x})$ is the density and $[-\infty, \mathbf{x}_0]$ the interval over which we want to integrate, we first choose a *n*-dimensional sampling-distribution with well-known properties. We work with a truncated sampling distribution with support over $[-\infty, \mathbf{x}_0]$. Defining $g_n^c(\mathbf{x})$ as the density for this *n*-dimensional truncated sampling distribution, we then multiply and divide the right-hand-side of the integral we wish to calculate, (14), by this density, which simply restates (14) as:

$$p = \int_{-\infty}^{\mathbf{x}_0} \frac{f_n(\mathbf{x})}{g_n^c(\mathbf{x})}\; g_n^c(\mathbf{x})\, d\mathbf{x} \qquad (15).$$

By definition, the solution to this integral is a mean because $g_n^c(\mathbf{x})$ is a valid *pdf* over the integral's range, so (15) gives the probability sought, *p*, as the mean of $f_n(\mathbf{x})/g_n^c(\mathbf{x})$, which we can estimate using a sample of *R* draws of the $n \times 1$ vector $\mathbf{x}$ from the importance distribution. Formally:

$$p = E\left[\frac{f_n(\mathbf{x})}{g_n^c(\mathbf{x})}\right] \approx \frac{1}{R}\sum_{r=1}^{R} \frac{f_n(\tilde{\mathbf{x}}_r)}{g_n^c(\tilde{\mathbf{x}}_r)} \equiv \hat{p} \qquad (16).$$

To implement the RIS estimator, we simply draw $\mathbf{x}$ from the importance-distribution, for which we will use a truncated multivariate (independent) normal,[7] and calculate $f_n(\mathbf{x})/g_n^c(\mathbf{x})$. Since $g_n^c(\mathbf{x})$ is a density for a truncated normal, the ratio $f_n(\mathbf{x})/g_n^c(\mathbf{x})$ simplifies to $G(\mathbf{x}) = \mathbf{\Phi}(\mathbf{x})$.

To illustrate this strategy I describe how to use importance sampling to estimate the spatial-lag duration and probit models. These models substitute $\rho\mathbf{Wy}$ and $\rho\mathbf{Wy}^*$ for $\mathbf{Ay}$ and $\mathbf{Ay}^*$ in equations (1) and (6) respectively. The parameter $\rho$ is the spatial autoregressive coefficient and $\mathbf{W}$ is an $N\times N$ spatial-weighting matrix, with elements $w_{ij}$ reflecting the relative degree of connection from unit *j* to *i*. $\mathbf{Wy}$ is thus the spatial lag. For each observation $y_{it}$, $\mathbf{Wy}$ gives a weighted sum of the $y_{jt}$, with weights, $w_{ij}$, given by the relative connectivity from *j* to *i*. Notice how $\mathbf{Wy}$ provides a way to directly model the interdependence of outcomes and behavior in our data.

I begin with the spatial-lag probit. In the standard probit-model with independent errors, the numerator would simply sum *n* univariate cumulative standard-normal distributions, which is manageable. In spatial-lag probit, with its interdependent errors, however, the numerator is a single *n*-dimensional cumulative-normal:

$$p(\mathbf{v} < \mathbf{Z}) \qquad (17),$$

with $\mathbf{v}$ the $n\times 1$ vector of errors distributed $MVN(\mathbf{0}, \mathbf{\Omega})$ and $\mathbf{\Omega} = \left[(\mathbf{I}-\rho\mathbf{W})'(\mathbf{I}-\rho\mathbf{W})\right]^{-1}$ and $\mathbf{Z}$**,** the $n\times 1$ vector defined above.

The RIS estimator for spatial-lag probit exploits the fact that, as a variance-covariance matrix, $\mathbf{\Omega}$ is positive definite, and so a Cholesky decomposition exists such that $\mathbf{\Omega}^{-1} = \mathbf{A}'\mathbf{A}$,

[7] Other importance distributions, such as a *t* or a uniform may be used. With a normal importance-distribution, RIS is equivalent to the better-known GHK (Geweke-Hajivassiliou-Keane) simulation estimator.

with $\mathbf{A}$ being an upper-triangular matrix and $\boldsymbol{\eta} = \mathbf{A}\mathbf{v}$ giving $n$ independent standard-normal variables, $\boldsymbol{\eta}$. (This exploitation is familiar as the same one applied in GLS.) Let $\mathbf{B} \equiv \mathbf{A}^{-1}$; substituting $\mathbf{v} = \mathbf{A}^{-1}\boldsymbol{\eta} \equiv \mathbf{B}\boldsymbol{\eta}$ into (17) then gives:

$$\Pr\left(\mathbf{B}\boldsymbol{\eta} < \mathbf{Z}\right) = \Pr\left(\begin{bmatrix} b_{1,1} & b_{1,2} & \cdots & \cdots & b_{1,n} \\ 0 & b_{2,2} & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & b_{n\text{-}1,n\text{-}1} & b_{n\text{-}1,n} \\ 0 & \cdots & 0 & 0 & b_{n,n} \end{bmatrix} \begin{bmatrix} \eta_1 \\ \vdots \\ \vdots \\ \eta_{n\text{-}1} \\ \eta_n \end{bmatrix} < \begin{bmatrix} z_1 \\ \vdots \\ \vdots \\ z_{n\text{-}1} \\ z_n \end{bmatrix}\right) \quad (18).$$

The elements of the $n\times 1$ vector $\boldsymbol{\eta}$ are independent, so the probability in (18) can be calculated by first evaluating the cumulative-normal distribution function at the implied upper bounds, which are determined recursively starting with the last observation, and then multiplying these probabilities. To determine these upper bounds, start by solving the inequalities in (18) for the vector $\boldsymbol{\eta}$:

$$\Pr\left(\begin{bmatrix} \sum_{i=1}^{n} b_{1,i}\eta_i \\ \vdots \\ b_{n\text{-}1,n\text{-}1}\eta_{n\text{-}1} + b_{n\text{-}1,n}\eta_n \\ b_{n,n}\eta_n \end{bmatrix} < \begin{bmatrix} z_1 \\ \vdots \\ z_{n\text{-}1} \\ z_n \end{bmatrix}\right) = \Pr\left(\begin{bmatrix} \eta_1 \\ \vdots \\ \eta_{n\text{-}1} \\ \eta_n \end{bmatrix} < \begin{bmatrix} b_{1,1}^{-1}\left(v_1 - \sum_{i=2}^{n} b_{1,i}\eta_i\right) \\ \vdots \\ \vdots \\ b_{n\text{-}1,n\text{-}1}^{-1}\left(v_{n\text{-}1} - b_{n\text{-}1,n}\eta_n\right) \\ b_{n,n}^{-1}v_n \end{bmatrix}\right) \quad (19)$$

Next, calculate the upper bound for the truncated-normal distribution of the $n^{th}$ observation, which is $b_{n,n}^{-1}z_n$. Call the cumulative standard-normal evaluated at this upper bound $p_n$. Then take a draw from the standard-normal distribution truncated at $b_{n,n}^{-1}z_n$; call that draw $\tilde{\eta}_n$ and use it to calculate the upper bound for the truncated-normal distribution for the $(n\text{-}1)^{th}$ observation conditional on the $n^{th}$ as $b_{n\text{-}1,n\text{-}1}^{-1}\left[v_{n\text{-}1} - b_{n\text{-}1,n}\tilde{\eta}_n\right]$. Evaluate the cumulative standard-normal at this upper bound and call it $p_{n\text{-}1}$. Then use the first two draws to calculate the $(n\text{-}2)^{th}$

upper bound and calculate $p_{n-2}$ analogously, and so on through all $n$ observations. Formally, this recursive process for calculating the upper bounds is:

$$\left.\begin{array}{c}\eta_n < b_{n,n}^{-1} z_n \equiv \zeta_n \\ \eta_{n\text{-}1} < b_{n\text{-}1,n\text{-}1}^{-1}\left[z_{n\text{-}1} - b_{n\text{-}1,n}\tilde{\eta}_n\right] \equiv \zeta_{n\text{-}1} \\ \eta_{n\text{-}2} < b_{n\text{-}2,n\text{-}1}^{-1}\left[z_{n\text{-}2} - b_{n\text{-}2,n\text{-}1}\tilde{\eta}_{n\text{-}1} - b_{n\text{-}2,n}\tilde{\eta}_n\right] \equiv \zeta_{n\text{-}2} \\ \vdots \end{array}\right\} \Rightarrow \eta_j < b_{j,j}^{-1}\left[z_j - \sum_{i=j+1}^{n} b_{j,i}\tilde{\eta}_i\right] \equiv \zeta_j \qquad (20).$$

The probability of observing a given sample of ones and zeros can now be found by evaluating the *univariate* cumulative-normal distribution function at each of these bounds, $p_i$, and then multiplying those probabilities: $\prod_{j=1}^{n} p_j = \prod_{j=1}^{n} \Phi(\zeta_j)$. Repeating the entire process $R$ times and averaging gives the RIS estimate for the joint probability, i.e., the simulated likelihood, as this mean:

$$\hat{l} = (1/R)\sum_{r=1}^{R}\left[\prod_{j=1}^{n}\Phi(\zeta_{j,r})\right] \qquad (21).$$

One can then maximize this simulated likelihood by any standard optimization routine to estimate parameters and apply standard ML estimators for the variance-covariance (e.g., the observed information matrix, $-[\mathbf{H}(\hat{l})]^{-1}$).

I present some preliminary Monte Carlo results in Table 1 and compare these with earlier experimental results provided in Beron and Vijverberg (2004). The data generating process is (7)-(8). The sample size is 48, and I use the row-standardized binary-contiguity weights matrix for the lower 48 US states. My experiments use slightly different parameter values and a different **X** than Beron and Vijverberg. They also use antithetical sampling and a much larger $R$ than I do (2000 vs. 100). Nevertheless, the results are similar. Like Beron and Vijverberg, I find that $\rho$ is underestimated, although their RIS estimates are closer to the truth on average than mine. The also provide results for a simple (spatial) linear probability model (LPM), which I report as well. Beron and Vijverberg do not report results for standard-error estimation. However, I find that the observed information matrix provides good estimates for the variance of the sampling distribution.

[Table 1 Here]

Can we use this technique to simulate the likelihood in (13)? With the log-normal spatial-lag duration model, we have interdependent failures with probabilities determined by the multivariate normal distribution. In some cases we observe the failures and in other cases we do not. One strategy is to sample from the multivariate normal distribution using the recursive approach described above. When we observe a failure, we calculate the density at that point. When we are censored, we calculate the survival function at that point, and then sample from the truncated distribution. See McCarty and Rothenberg (2000) for a similar approach to estimating a Tobit model with interdependent disturbances.

I provide preliminary Monte Carlo results in Table 2 that compare the standard ML (using the likelihood in (12)) and MSL estimators with and without censoring. The data-generating process is mostly the same as that used previously *mutatis mutandis*. In the no-censoring case, the standard ML likelihood is properly specified, and so the estimator should give the correct estimates on average. This experiment simply compares the Jacobian transformation and Cholesky decomposition approaches to modeling the reduced-form disturbances. With censoring only the maximum simulated likelihood estimator should give the correct estimates on average. This is exactly what I find. Both approaches give reasonably good estimates on average when there is no censoring. But the ML estimator is biased for all the parameters in the model when there is censoring. The MSL estimator is unbiased and relatively efficient on $\beta$ and the spatial lag coefficient $\rho$. The standard ML estimator weakly dominates the MSL estimator in MSE terms when there is no censoring, whereas the MSL estimator dominates the ML estimator in MSE when censoring is present.

[Table 2 Here]

## IV. Calculating Effects

Estimating parameters is an important first step in any analysis, but ultimately we are interested in calculating the effects implied by our estimates. Returning to the spatial probit model, Beron and Vijverberg (2004) give the marginal effect of $X_i$ on the probability that $y_i$ equals one as

$$\frac{d\,\Pr[y_i = 1 \mid X, \mathbf{W}]}{dX_i} = \phi\left(\Omega_{\alpha,ii}^{-1/2}[\Gamma_\alpha X\beta]_i\right)\Omega_{\alpha,ii}^{-1/2}\Gamma_\alpha\beta \tag{22}$$

where $\phi$ is the univariate density function for the standard normal distribution , $\Gamma_\alpha = \dfrac{dy^*}{d(X_i\beta)}$, and $[\Gamma_\alpha X\beta]_i$ is the $i^{\text{th}}$ element of the vector $\Gamma_\alpha X\beta$. Note that changes in $X_j$ also effect $y_i^*$, so the quantity $\dfrac{d\,\Pr[y_i = 1 \mid X, \mathbf{W}]}{dX_j}$ is of interest. Beron and Vijverberg argue that it is inappropriate to condition the effects of changes in *X* on the probability $y_i$ equals one on the other $y_j$ because the $y_j$ are responding endogenously to the changes in *X*.

This seems unnecessarily restrictive given that, after we estimate the model, we can easily sample from the distribution of disturbances using the reduced-form, generate y's according to the measurement equation, and calculate conditional frequencies. In other words, we can use the model to generate counterfactual values of the dependent variable for a given set of **X** and **W**, and then estimate the probabilities $\Pr[y_i = 1 \mid X, \mathbf{W}, y_j = 1]$ and

$\Pr[y_i = 1 \mid X, \mathbf{W}, y_j = 0]$ using relatively frequencies. While this sampling strategy was infeasible for estimating the model's parameters, once we have the parameters and a relatively specific counterfactual, the computation costs are relatively low. I give an example of this approach to counterfactual analysis in the illustration below.

## V. World War I Participation and Entry Timing Decisions

In this section, I model the decisions of states to enter WWI. One can treat these decisions as independent and driven purely by domestic and international structural factors such as regime type, trade exposure, and relative military capabilities, but this is approach unsatisfactory. Ultimately, each state's decision was heavily influence by the entry decision of others, and any empirical analysis should take this interdependence into account. I incorporate four forms of interdependence into my models: contiguity, territorial dispute, rivalry, and targeted alliances. These sources of interdependence suggest that a state will be influenced by the participation and entry timing decisions of those states with which it shares a border or a pre-existing territorial dispute or those states that are rivals or the targets of alliances. One might expect the interdependence to be positive, but negative interdependence, suggesting free-riding behavior, is plausible as well. Take the case of rivals as an example. A participating state's rivals may stay out of the conflict hoping for a favorable outcome—that the participant will lose the war and suffer a decrease in power—at no cost to the non-participating rival.

Unfortunately, due to a lack of variation in WWI participation among states involved in disputes, rivalries, and alliances, I am unable to estimate the effects of these forms of interdependence on the binary participation choice variable, but I am able to estimate the full set of duration models, and I am able to estimate a contiguity spatial-lag probit model as well. The results are presented in Table 3. For controls I include national capabilities, democracy, trade, and a Europe dummy variable. (These variables are described in the notes to Table 3.) I find statistically significant positive interdependence in the contiguity probit model. In the duration models, the contiguity spatial lag has a positive coefficient (z-stat > 1.6) while the territorial disputes lag coefficient is negative, although the size of the latter estimate is outside of traditionally accepted bounds, making interpretation difficult.[8] The contiguity results are consistent with theories of war as a contagion. The territorial dispute result (negative interdependence) is consistent with free-riding behavior. The set of territorial disputes prior to the onset of WWI were clustered in the Balkans. Two of the countries involved these disputes, Turkey and Serbia, joined the war early. The others—Greece, Bulgaria, and Romania—were late joiners. This pattern is the source of the negative spatial-lag coefficient estimate, but it is too early to conclude that this reflects true negative interdependence. To give a concrete example, the free-riding argument would say that Greece stayed out of the war as long as possible hoping that the Central Powers would be defeated and its territorial dispute with the Ottoman Empire would be resolved favorably at no cost.

[Table 3 Here]

There are many interesting counterfactual questions related to the interdependence of states' participation and entry-timing decisions. How did America's decision to participate affect the participation of others? To what extent did Italy's decision to enter the war in mid-1915 affect the probability that Bulgaria and Romania would be drawn into the conflict before the fighting stopped? I address the latter question using the spatial probit estimates from Table 3. In

[8] Specifically, the estimate falls outside the "bounds of singularity" for the $(\mathbf{I}-\rho\mathbf{W})$ matrix. The lower bound is negative one over the absolute value of the minimum eigenvalue for the weights matrix. The multiplier can be calculated, but the interpretation of effects is not straightforward.

terms of the model, the question becomes, given that Italy's reduced-form disturbance is above or below its reduced-form cutpoint, what is the probability that Romania's reduced-form disturbance will be above or below its reduced-form cutpoint? To answer this question, I sample from the reduced form disturbances. More specifically, I draw 1,000 times from a $N(0,1)$ for each of the 44 states in the sample. This gives a $44 \times 1000$ matrix of i.i.d. standard normal disturbances. Then I pre-multiply this disturbance matrix by the $44 \times 44$ spatial multiplier, which gives $\mathbf{V} = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{U}$. Since the counterfactual question involves the participation of Italy and Romania, I take a bivariate slice of the resulting 44-dimensional multivariate distribution.

The vector of reduced-form cupoints is calculated as $\mathbf{\Gamma X \beta}$. A country participates if its reduced-form disturbance is greater than the negative value of its reduced-form cuptoint. Given their covariates, Italy's cutpoint is .505 and Romania's is .844. In the simulation, Romania's disturbances are above the critical level 75% of the time when Italy's disturbances are below its critical level and above 80% of the time when Italy's disturbances are above its critical level. The reduced form disturbances are plotted in Figure 1. The linear correlation between them is about .15. The model suggests that Italy's participation in the War is associated with a 5% increase in Romania's participation. Or, viewed the other way, had Italy not entered the War, the probability that Romania would have stayed out increases by 5%. It should be noted that Italy and Romania did not share a border, and therefore, these are second-order (neighborhood) effects.

## VI. Conclusion

Quantitative empirical research in political science frequently involves the analysis of interdependent data. In some applications, this interdependence is viewed as a nuisance and great effort is taken to purge the data of these connections, but much of the time, the interdependence itself is an important part of the politics that we aim to study. Unfortunately, empirical models that incorporate political and other forms of interdependence can be difficult to estimate and use for analytical purposes. Not surprisingly, the available methods are rarely used and underdeveloped. In this paper, I discuss how to use simulation methods to estimate and analyze simultaneous equation models with limited and qualitative dependent variables. To illustrate, I modeled the decisions of states to enter WWI using both a spatial-probit and spatial-duration model with right-censoring. The patterns of participation provide support for both the contagion and free-rider hypotheses.

## VII. References


Franzese, Robert J. and Jude C. Hays. 2008. "The Spatial Probit Model of Interdependent Binary Outcomes: Estimation, Interpretation, and Presentation." Available at SSRN: http://ssrn.com/abstract=1116393

Gill, Jeff. 2002. *Bayesian Methods: A Social and Behavioral Sciences Approach*. Boca Raton, Florida: Chapman & Hall/CRC.

Hays, Jude C. and Aya Kachi. 2009. "Full Information Maximum Likelihood Estimation of Interdependent Weibull Duration Models with an Application to Government Formation and Survival." Available at SSRN: http://ssrn.com/abstract=1372937

Lawrence, Eric. 1996. "Simulated Maximum Likelihood via the GHK Simulator: An Application to the 1988 Democratic Super Tuesday Primary." Paper presented at the 1996 Midwest Political Science Associations Annual Conference.

Lerman, Steven and Charles Manksi. 1981. "On the Use of Simulated Frequencies to Approximate Choice Probabilities." In *Structural Analysis of Discrete Data with Econometric Applications*. Eds. Charles Manski and Daniel McFadden. Cambridge: MIT Press, pp. 305-319.

McCarty, Nolan and Lawrence S. Rothenberg. 2000. "The Time to Give: PAC Motivations and

Electoral Timing." *Political Analysis* 8(3): 239-259.

Stern, Steven. 1997. "Simulation-Based Estimation." *Journal of Economic Literature* 35(Dec.): 2006-2039.

**Table 1. Spatial Probit Monte Carlo Results (100 Samples)**

| Parameter | Result | B&V(RIS) | B&V(LPM) | Hays(RIS) | Hays(Wy) |
|---|---|---|---|---|---|
| $B\&V: \hat{\beta} = 3$ <br> $Hays: \hat{\beta} = 1$ | mean | 3.29 / .85* | .89* | .95 | .91 |
| | s.d. | 1.08 | .17 | .27 | .26 |
| | rmse | 1.12 | N.A. | .28 | .28 |
| | mean $\widehat{s.e.}$ | N.A. | N.A. | .28 | .27 |
| | overconfidence | N.A. | N.A. | .97 | .96 |
| $\hat{\rho} = .5$ | mean | .41 | .25 | .35 | .40 |
| | s.d. | .22 | .16 | .20 | .36 |
| | rmse | .24 | .30 | .25 | .37 |
| | mean $\widehat{s.e.}$ | N.A. | N.A. | .20 | .36 |
| | overconfidence | N.A. | N.A. | 1.00 | .99 |

*Notes*: R=2000 vs. R=100, N=50 vs. N=48 ; * marginal effects

**Table 2. Spatial Duration Monte Carlo Results (100 Samples)**

| | | No Censoring | | Censoring* | |
|---|---|---|---|---|---|
| Parameter | Result | ML | MSL | ML | MSL |
| $\beta = 1$ | mean | 1.03 | .950 | .819 | 1.03 |
| | s.d. | .138 | .157 | .246 | .226 |
| | rmse | .142 | .165 | .305 | .227 |
| | mean $\widehat{s.e.}$ | .147 | .154 | .132 | .197 |
| | overconfidence | .939 | 1.02 | 1.86 | 1.14 |
| $\rho = .5$ | mean | .484 | .532 | .614 | .473 |
| | s.d. | .071 | .080 | .160 | .095 |
| | rmse | .073 | .086 | .197 | .099 |
| | mean $\widehat{s.e.}$ | .076 | .081 | .072 | .090 |
| | overconfidence | .934 | .994 | 2.24 | 1.05 |
| $\sigma = 1$ | mean | .984 | .980 | .708 | 1.02 |
| | s.d. | .096 | .095 | .103 | .153 |
| | rmse | .097 | .097 | .310 | .154 |
| | mean $\widehat{s.e.}$ | .101 | .100 | .077 | .157 |
| | overconfidence | .950 | .950 | 1.34 | .97 |

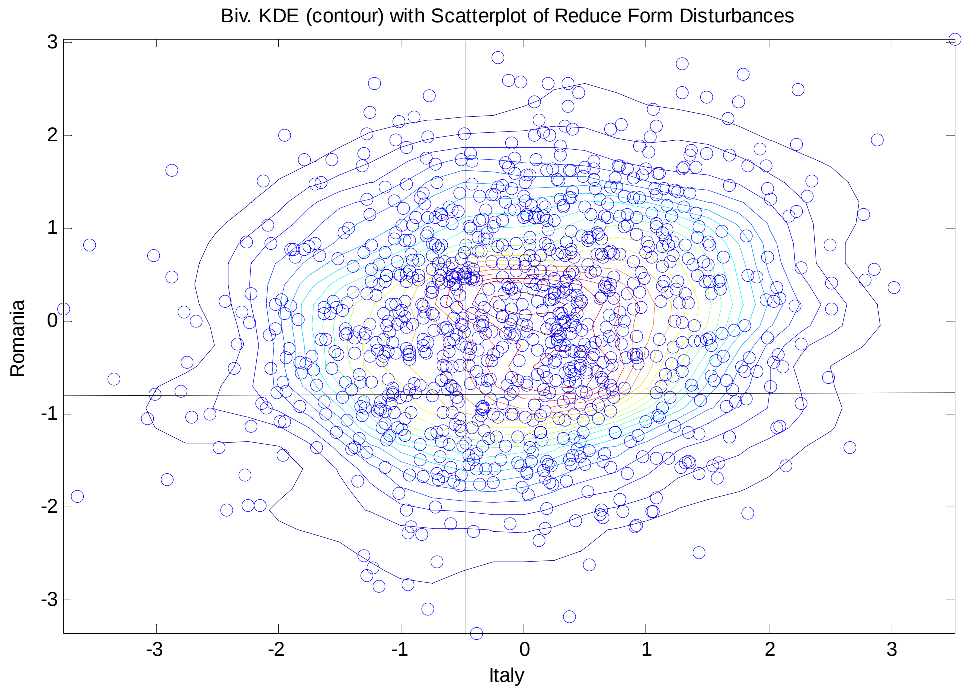
*Notes*: N=48, R=100. *Durations larger than 4 were censored. In the experiments, on average, this amounted to 50% of the durations being observed and 50% being censored.

**Table 3. Spatial Models of WWI Participation and Entry Timing**

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| Constant | -.806 | 8.19*** | 9.87*** | 9.81*** | 11.58*** |
| | (.611) | (2.21) | (2.19) | (2.14) | (2.54) |
| Shape Parameter | | 3.33*** | 3.42*** | 3.40*** | 3.32*** |
| | | (.650) | (.685) | (.676) | (.656) |
| Contiguity | .482** | .289 | | | |
| Spatial Lag | (.217) | (.180) | | | |
| Targeted Alliance | | | .180 | | |
| Spatial Lag | | | (1.32) | | |
| Rivalry | | | | .288 | |
| Spatial Lag | | | | (.491) | |
| Territorial Dispute | | | | | -2.42** |
| Spatial Lag | | | | | (1.21) |
| National Capabilities | 18.48*** | -41.45*** | -38.03** | -39.68** | -38.18** |
| | (6.61) | (15.62) | (17.98) | (15.69) | (15.37) |
| Democracy | -.055 | .045 | .081 | .09 | .040 |
| | (.137) | (.118) | (.129) | (.116) | (.119) |
| Trade | -.156* | -.088 | -.178 | -.149 | -.469 |
| | (.088) | (.331) | (.336) | (.338) | (.370) |
| Europe | 1.54** | -4.52*** | -5.48*** | -5.47*** | -4.51*** |
| | (.648) | (1.65) | (1.77) | (1.65) | (1.66) |
| Model | Probit | Weibull | Weibull | Weibull | Weibull |
| Observations | 44 | 44 | 44 | 44 | 44 |
| Log-Likelihood | -14.616 | -48.151 | -49.332 | -49.190 | -47.079 |

*Notes*: In the probit case, the dependent variable reflects participation in WWI (0=No, 1=Yes). For the duration models, the dependent variable is the number of months before entering WWI. Of the 44 sample countries, 15 enter the War. All the spatial weights matrices are row-standardized. *National Capabilities* are the COW CINC index scores. *Democracy* is Polity measures of regime type. *Trade* is the value of total trade in current US dollars (Source: Barbieri 2002). Europe is a dummy variable that takes a value of 1 for countries located on the continent, including the United Kingdom. Parentheses contain standard error estimates. ***significant at 1%; **significant at 5%; *significant at 10%.

**Figure 1. Scatterplot of Reduced-Form Cutpoints and Disturbances for Italy and Romania**



## Appendix: MATLAB Code (Likelihood Evaluators)

```
function LL_normal = sprobit09_llf_finaledit(p)
global k rand_mat n Q W X i_n r

%dbstop in sprobit09_llf_finaledit at 72
A_Like = (i_n-p(1,k)*W);
mult = inv(A_Like);
vcov = Q'*inv(A_Like'*A_Like)*Q;
ACH = chol(inv(vcov));
BCH = inv(ACH);
mx = mult*X(:,1:k-1);
V = -Q*mx*p(1,1:k-1)';

for j = 1:r

    nu0(n,j) = (1/BCH(n,n))*V(n,1);
    if nu0(n,j) < -8
        nu0(n,j) = -8;
    else
        nu0(n,j) = nu0(n,j);
    end
    if nu0(n,j) > 8
        nu0(n,j) = 8;
    else
        nu0(n,j) = nu0(n,j);
    end
    ln_prob(n,j) = log(norm_cdf(nu0(n,j)));
    nu(n,j) = norm_inv(rand_mat(n,j)*(norm_cdf(nu0(n,j))));

    if nu(n,j) < -8
        nu(n,j) = -8;
    else
        nu(n,j) = nu(n,j);
    end
    if nu(n,j) > 8
        nu(n,j) = 8;
    else
        nu(n,j) = nu(n,j);
    end
        for z = 1:n-1
        sumterm = 0;
            for m=1:z
                sumterm1 = BCH(n-z,n-z+m)*nu(n-z+m,j);
                sumterm = sumterm+sumterm1;
            end
                nu0(n-z,j) = (1/BCH(n-z,n-z))*(V(n-z,1)-sumterm);
                if nu0(n-z,j) < -8
                    nu0(n-z,j) = -8;
                else
                    nu0(n-z,j) = nu0(n-z,j);
                end
                if nu0(n-z,j) > 8
                    nu0(n-z,j) = 8;
                else
```

```
                nu0(n-z,j) = nu0(n-z,j);
            end

            ln_prob(n-z,j) = log(norm_cdf(nu0(n-z,j)));
            nu(n-z,j) = norm_inv(rand_mat(n-z,j)*(norm_cdf(nu0(n-z,j))));

            if nu(n-z,j) < -8
                nu(n-z,j) = -8;
            else
                nu(n-z,j) = nu(n-z,j);
            end
            if nu(n-z,j) > 8
                nu(n-z,j) = 8;
            else
                nu(n-z,j) = nu(n-z,j);
            end
        end
end
jnt_prob = sum(ln_prob);
LL_normal = -(mean(jnt_prob'));
```

```
function LL_normal = spdur09_llf2_desktop2(p)
global k rand_mat n W X r i_n durs censor;

A_Like = (1/(p(1,k+2)))*(i_n - p(1,k+1)*W);
mult = inv(i_n - p(1,k+1)*W);
ACH = chol(A_Like'*A_Like);
BCH = inv(ACH);
mx = mult*X(:,1:k);
v = durs - mx*p(1,1:k)';

for j = 1:r
    nu0(n,j) = (1/BCH(n,n))*v(n,1);

    if nu0(n,j) > 10
          nu0(n,j) = 10;
    else
          nu0(n,j) = nu0(n,j);
    end

    if nu0(n,j) < -10
          nu0(n,j) = -10;
    else
          nu0(n,j) = nu0(n,j);
    end

        if censor(n,1) == 1
            prob(n,j) = 1 - norm_cdf(nu0(n,j));
        else
            prob(n,j) = (1/(p(1,k+2)))*norm_pdf(nu0(n,j));
        end
            if prob(n,j) < .00000000001;
                prob(n,j) = .00000000001;
            else
                prob(n,j) = prob(n,j);
            end

    ln_prob(n,j) = log(prob(n,j));

        if censor(n,1) == 1
            nu(n,j) = -norm_inv(rand_mat(n,j)*(norm_cdf(-nu0(n,j))));

        else
            nu(n,j) = nu0(n,j);
        end

        for z = 1:n-1
        sumterm = 0;
            for m=0:z
                sumterm1 = BCH(n-z,n-z+m)*nu(n-z+m,j);
                sumterm = sumterm+sumterm1;
            end
                nu0(n-z,j) = (1/BCH(n-z,n-z))*(v(n-z,1)-sumterm);

                if nu0(n-z,j) > 10
                    nu0(n-z,j) = 10;
```

```
            else
                nu0(n-z,j) = nu0(n-z,j);
            end
            if nu0(n-z,j) < -10
                nu0(n-z,j) = -10;
            else
                nu0(n-z,j) = nu0(n-z,j);
            end

            if censor(n-z,1) == 1
                prob(n-z,j) = 1 - norm_cdf(nu0(n-z,j));
            else
                prob(n-z,j) = (1/(p(1,k+2)))*norm_pdf(nu0(n-z,j));
            end

            if prob(n-z,j) < .00000000001;
                prob(n-z,j) = .00000000001;
            else
                prob(n-z,j) = prob(n-z,j);
            end

          ln_prob(n-z,j) = log(prob(n-z,j));

            if censor(n-z,1) == 1
                nu(n-z,j) = - norm_inv(rand_mat(n,j)*(norm_cdf(-nu0(n-
z,j))));
            else
                nu(n-z,j) = nu0(n-z,j);
            end

        end
end
LL_normal= -(mean(sum(ln_prob)));
```