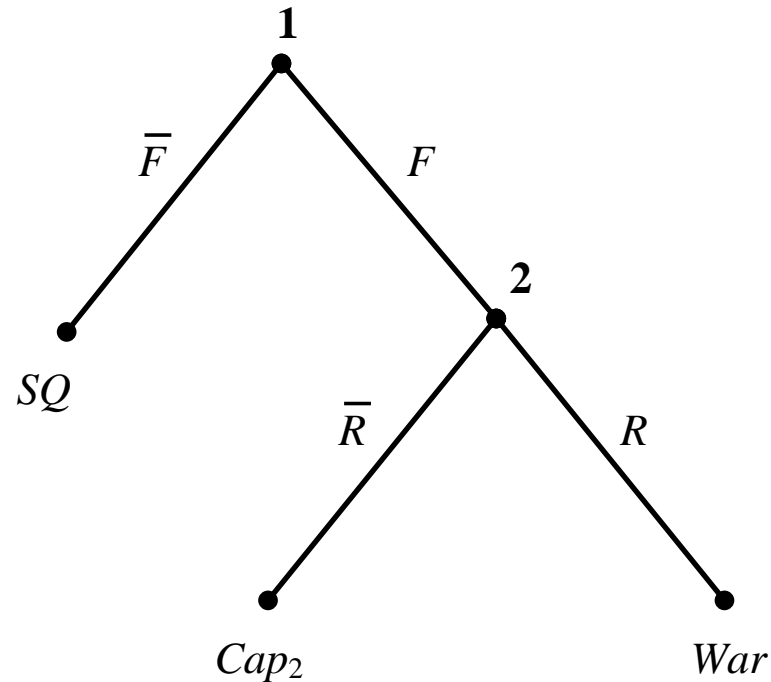

Selection Models

Curtis S. Signorino
University of Rochester

July 2, 2003

Background and Motivation



- All of international politics is selection
- IR scholars paying increased attention to selection issues (Huth, Reed, Sartori, Smith)

- Most commonly used models are Heckman 2-step (continuous) or Bivariate Probit (discrete)
- Problem? Linear selection equation, even though studying strategic behavior
- Signorino 1999, 2000 provides statistical strategic models
 - but assumes independence
- Objective of this paper:
 1. Show how strategic models are similar to or differ from traditional selection models.
 2. Derive statistical model that combines the “best” of both types of models — i.e., strategic and correlated
 3. Assess misspecification from using Bivariate Probit and Independent Strategic models

Different Forms of Sample Selection

- “selection” is often used to refer to a number of different issues concerning political science data.
- selection issues tend to be presented as one of the following:
 1. systematically selecting a sample based on an explanatory variable
 2. threats to external validity of inferences
 3. non-random selection on the dependent variable

Not so problematic

- Selecting on an explanatory variable
 - Advice often given to students studying censoring and truncation is that there is no problem if the selection mechanism is systematically related only to (or correlated only with) an explanatory variable (King et al 1994:137-149, King 1998:208).
- Threats to external validity of inferences
 - if our sample is not representative of the larger population, then even if our inferences are valid for the sample, they may not be for the larger population.
- At worst, we can say that “our results are limited to the given sample”

Selection on the dependent variable

- Heckman (1976,1979): if the sample selection is systematically related to or correlated with the dependent variable, and if this is not explicitly modeled, then resulting inferences will be biased.
- Consider

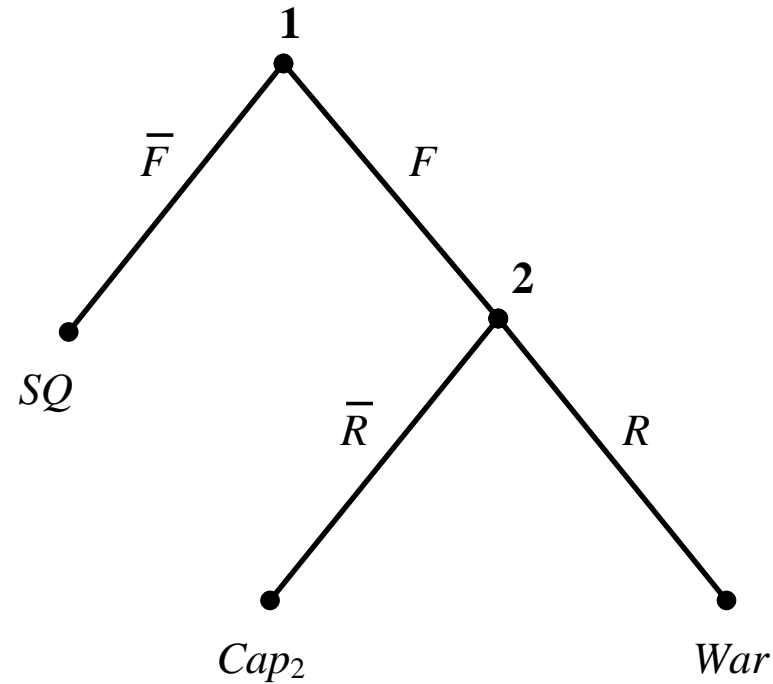
$$\begin{aligned} Y &= X\beta + \epsilon \\ E[\epsilon|X] &= 0 \end{aligned}$$

- Assume now that we select only those observations of Y such that $Y > 0$. Then

$$\begin{aligned} E[Y|X, Y > 0] &= E[X\beta|X, Y > 0] + E[\epsilon|X, Y > 0] \\ &= X\beta + E[\epsilon|Y > 0] \\ &= X\beta + E[\epsilon|\epsilon > -X\beta] \end{aligned}$$

- So, if the selection mechanism is systematically related to or correlated with the dependent variable of the observed sample, failure to account for that selection mechanism will lead to biased inferences.
- This is where the “selection” issue differs so dramatically from the previous cases. In those, one could sustain internal validity and avoid the external validity critique simply by claiming that one’s inferences were limited to the given sample.
- However, when the selection mechanism is correlated with the dependent variable and the researcher does not correctly model it, then the inferences are biased, even for the given sample.

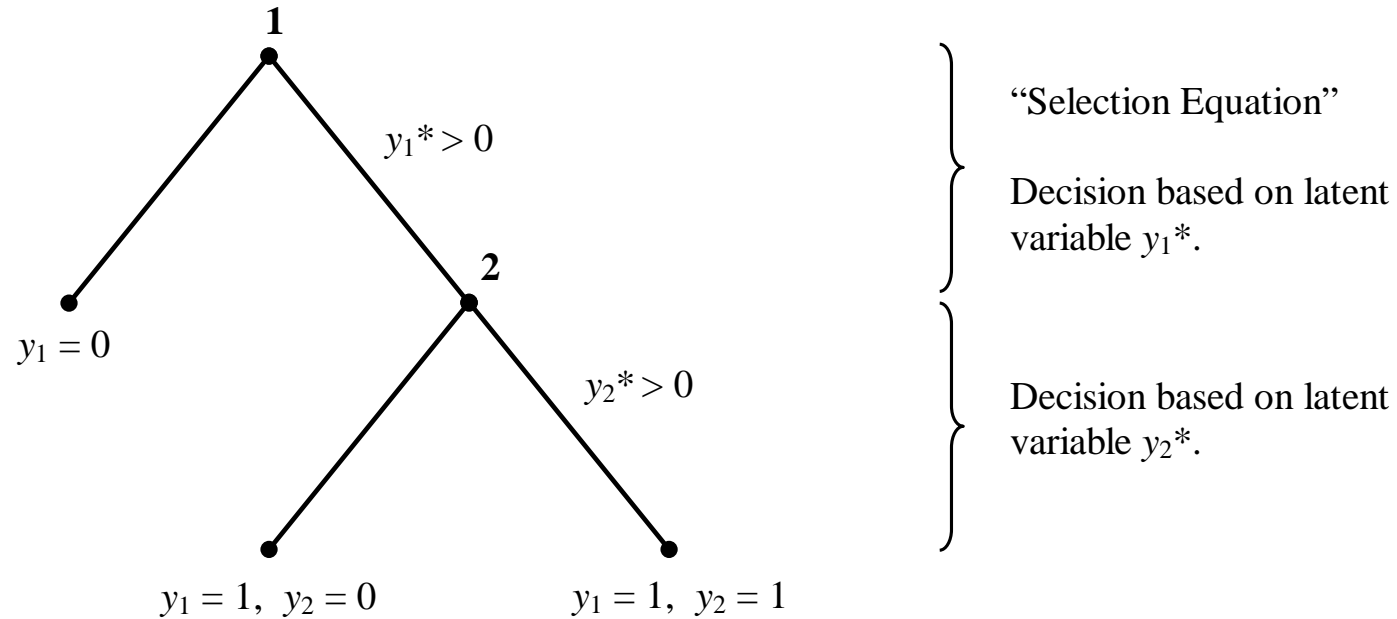
Example of Selection Problem



Suppose

1. we had data only on whether state 2 resisted or not
2. we used standard logit or probit to analyze that data

Correlated Errors and Selection



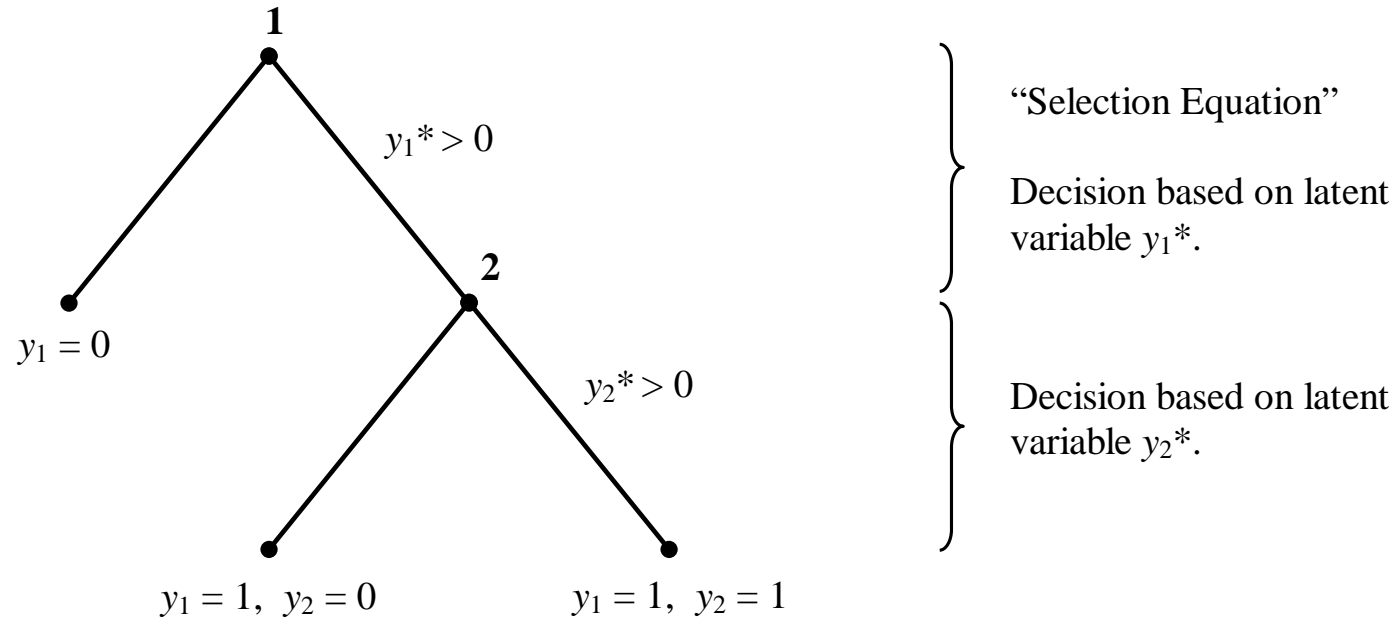
Typical Bivariate Probit Selection Model

$$y_1^* = X_1\beta_1 + \epsilon_1$$

$$y_2^* = X_2\beta_2 + \epsilon_2$$

$$\epsilon_1, \epsilon_2 \sim BN \left[0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right]$$

Correlated Errors and Selection



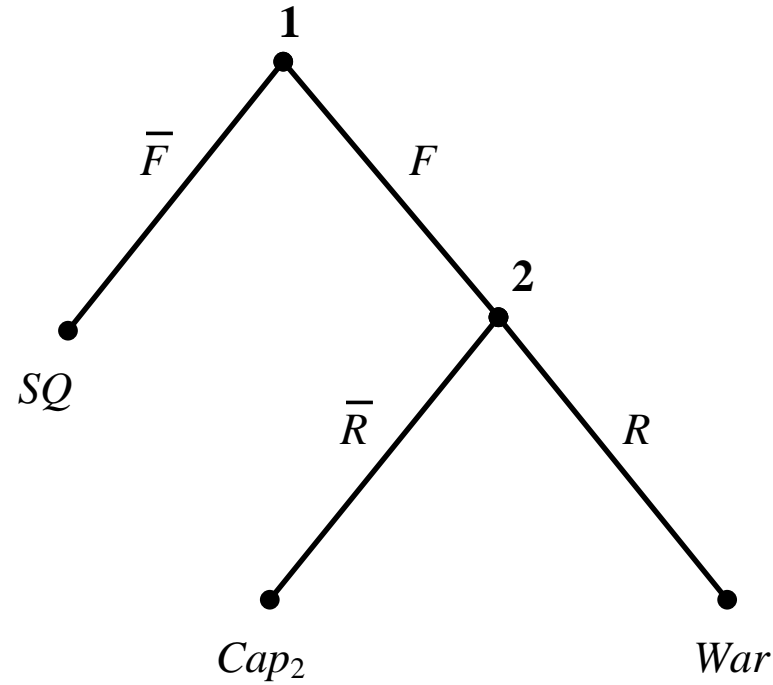
Bivariate probit probabilities:

$$\Pr(y_1 = 0) = \Phi_n(-X_1\beta_1) \quad (1)$$

$$\Pr(y_1 = 1, y_2 = 0) = \Phi_{bn}(X_1\beta_1, -X_2\beta_2, -\rho) \quad (2)$$

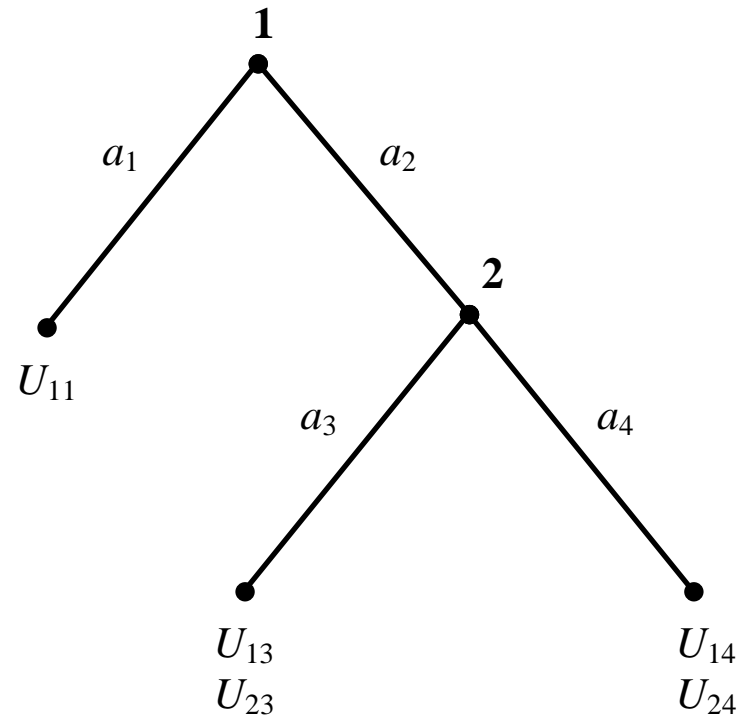
$$\Pr(y_1 = 1, y_2 = 1) = \Phi_{bn}(X_1\beta_1, X_2\beta_2, \rho) \quad (3)$$

Strategic Models Are Selection Models



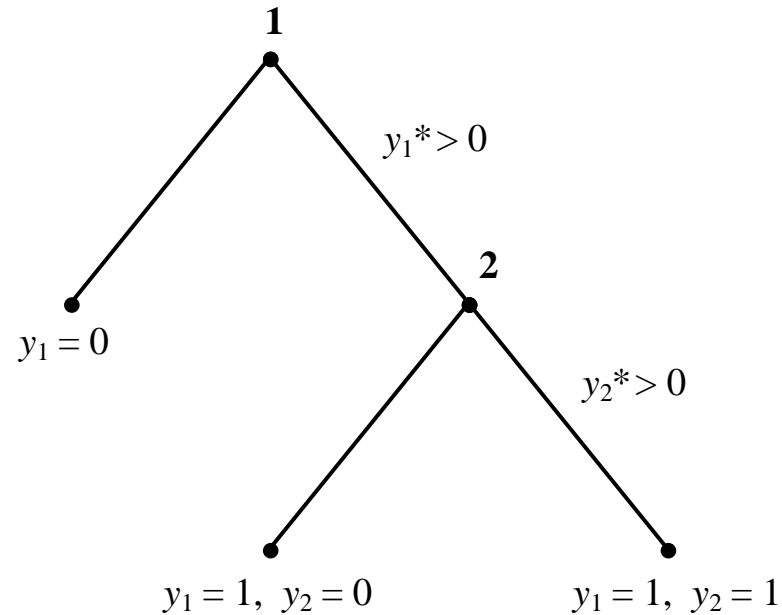
Players select themselves into “subsamples” or subgames via utility maximization

Strategic Models Are Selection Models



- Can reformulate Strategic Probit models as selection models

Strategic Selection Models



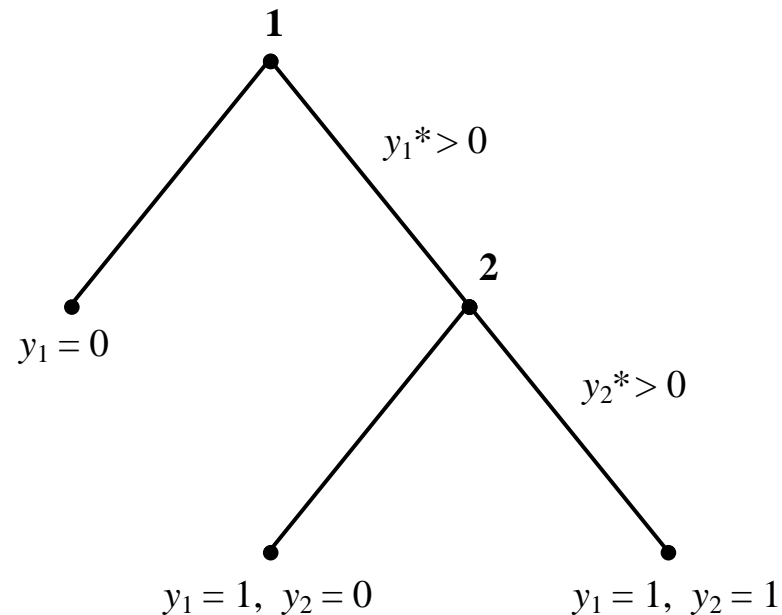
- Independent Strategic Probit

$$y_1^* = \Pr(y_2^* \leq 0) U_{13} + \Pr(y_2^* > 0) U_{14} - U_{11} + \epsilon_1$$

$$y_2^* = U_{24} - U_{23} + \epsilon_2$$

$$\epsilon_1, \epsilon_2 \sim N(0, 1)$$

Strategic Selection Models



Strategic probit probabilities

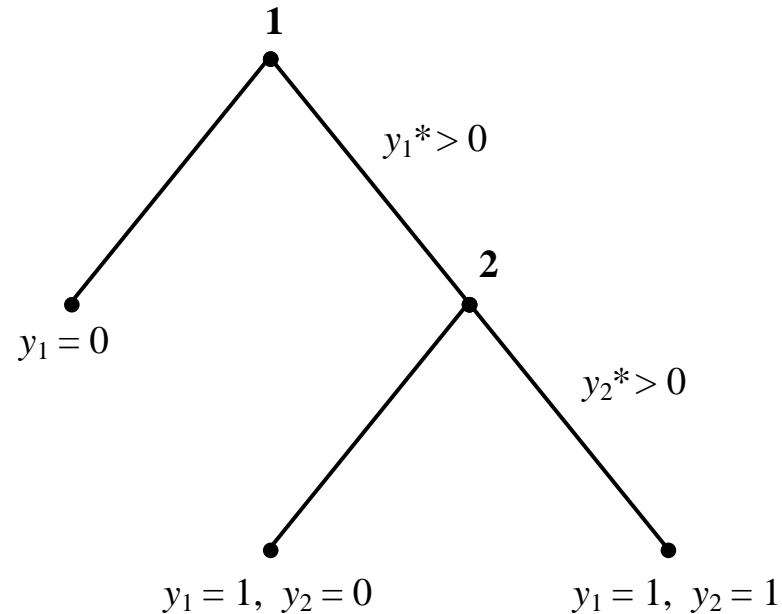
$$\begin{aligned}\Pr(y_1 = 0) &= \Phi_n [U_{11} - (p_3 U_{13} + p_4 U_{14})] \\ \Pr(y_1 = 1, y_2 = 0) &= \Phi_n [p_3 U_{13} + p_4 U_{14} - U_{11}] \Phi_n [U_{23} - U_{24}] \\ \Pr(y_1 = 1, y_2 = 1) &= \Phi_n [p_3 U_{13} + p_4 U_{14} - U_{11}] \Phi_n [U_{24} - U_{23}]\end{aligned}$$

where

$$p_3 = \Phi_n [U_{23} - U_{24}]$$

$$p_4 = \Phi_n [U_{24} - U_{23}]$$

Strategic Selection Models



- Correlated Strategic Probit

$$y_1^* = \Pr(y_2^* \leq 0 | \epsilon_1) U_{13} + \Pr(y_2^* > 0 | \epsilon_1) U_{14} - U_{11} + \epsilon_1$$

$$y_2^* = U_{24} - U_{23} + \epsilon_2$$

$$\epsilon_1, \epsilon_2 \sim BN \left[0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right]$$

State 1's private component tells it something about state 2's random component, which 1 uses in assessing the probability that 2 will resist or not.

However, the analyst does not observe state 1's disturbance.

Because (ϵ_1, ϵ_2) are distributed bivariate normal, the resulting outcome probabilities will take the same bivariate probit form as for the typical selection model, but with the expected utility calculation reflected in the arguments.

$$\Pr(y_1 = 0) = \Phi_n [-\epsilon_1^\circ] \quad (4)$$

$$\Pr(y_1 = 1, y_2 = 0) = \Phi_{bn} [\epsilon_1^\circ, U_{23} - U_{24}, -\rho] \quad (5)$$

$$\Pr(y_1 = 1, y_2 = 1) = \Phi_{bn} [\epsilon_1^\circ, U_{24} - U_{23}, \rho] \quad (6)$$

where ϵ_1° solves

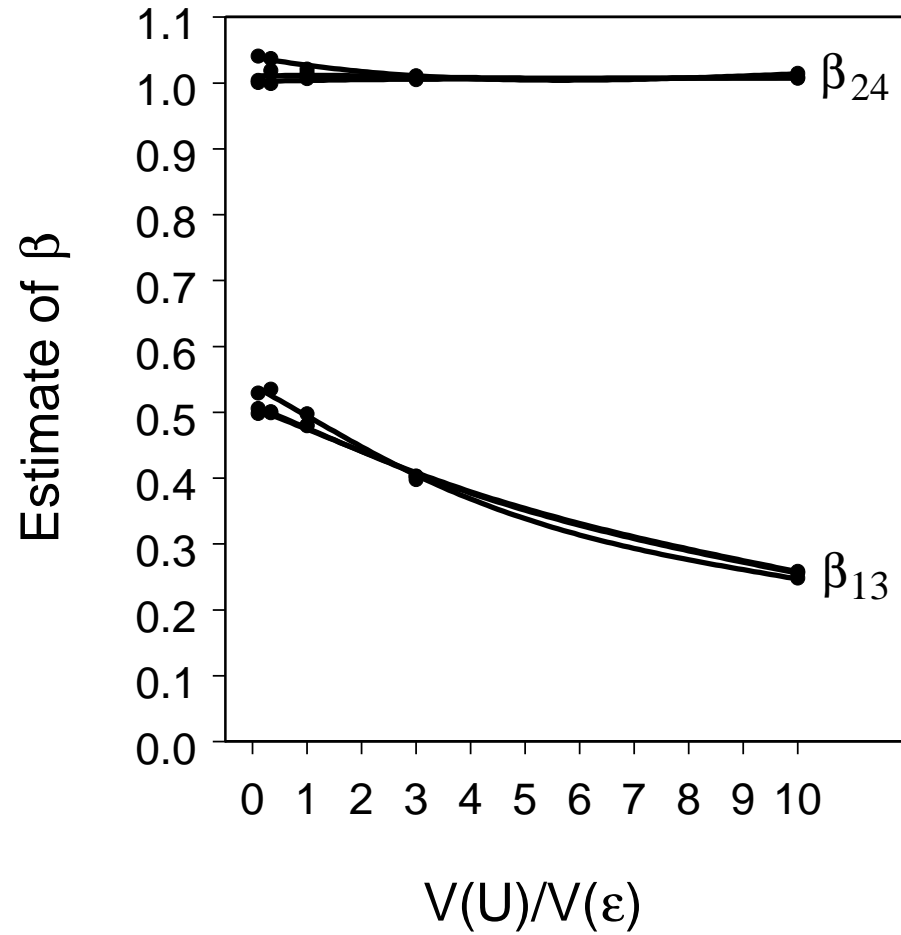
$$\Pr(y_2^* \leq 0 | \epsilon_1) U_{13} + \Pr(y_2^* > 0 | \epsilon_1) U_{14} + \epsilon_1 = U_{11} \quad (7)$$

for ϵ_1 .

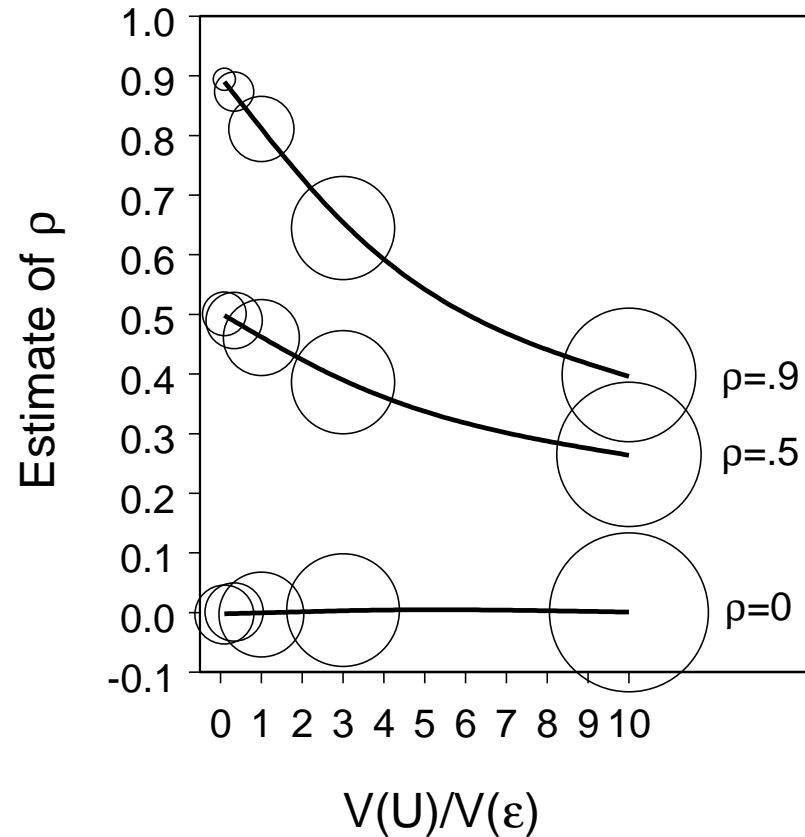
Monte Carlo Analysis

- Assumed behavior underlying Correlated Strategic Model
- $\rho = \{0, .5, .9\}$
- $V(U)/V(\epsilon) = \{\frac{1}{10}, \frac{1}{3}, 1, 3, 10\}$
 - Sense of how much error term matters vs observed utilities
- Results that follow are mean of monte carlo densities

When Bivariate Probit Goes Bad...

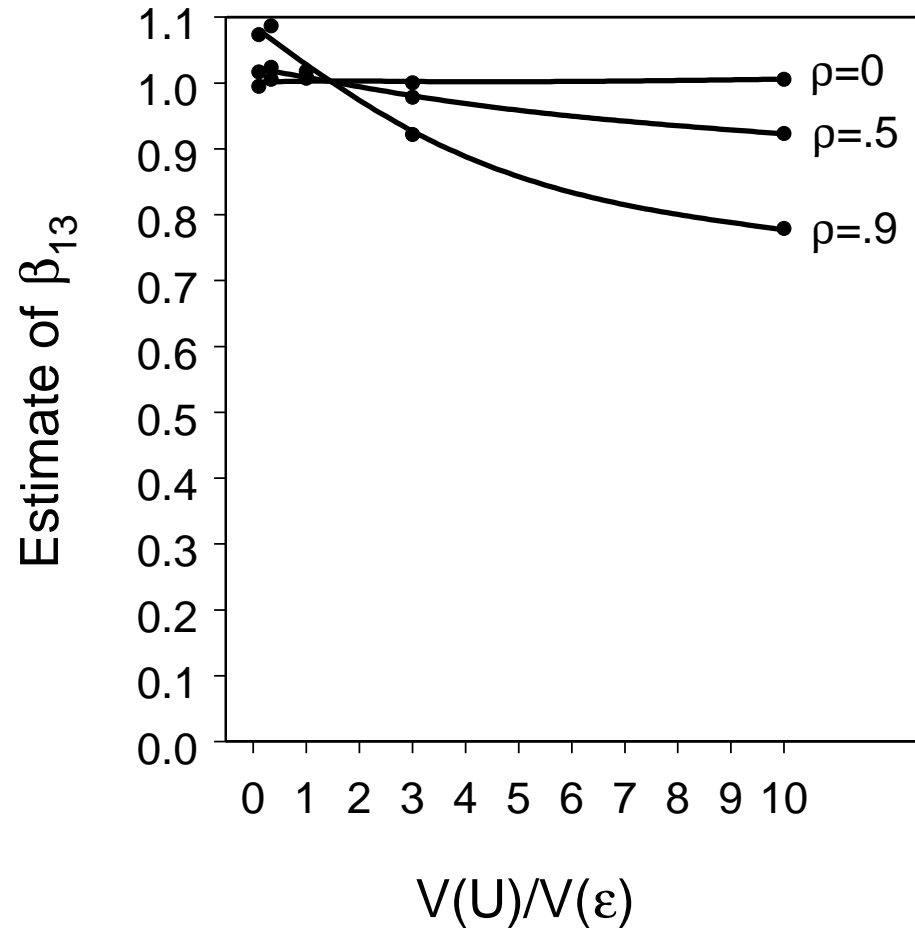


When Bivariate Probit Goes Bad...

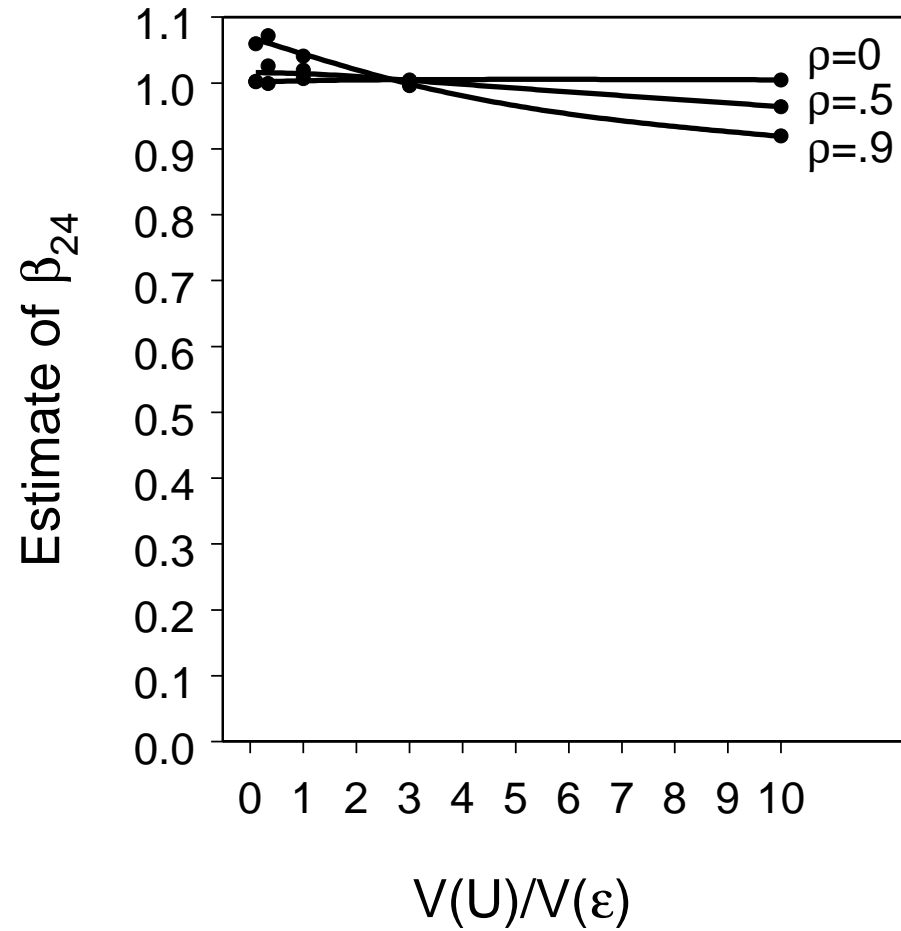


- Attenuation as error matters less — or structure and utilities matter more
- Less precise estimate of ρ as error matters less

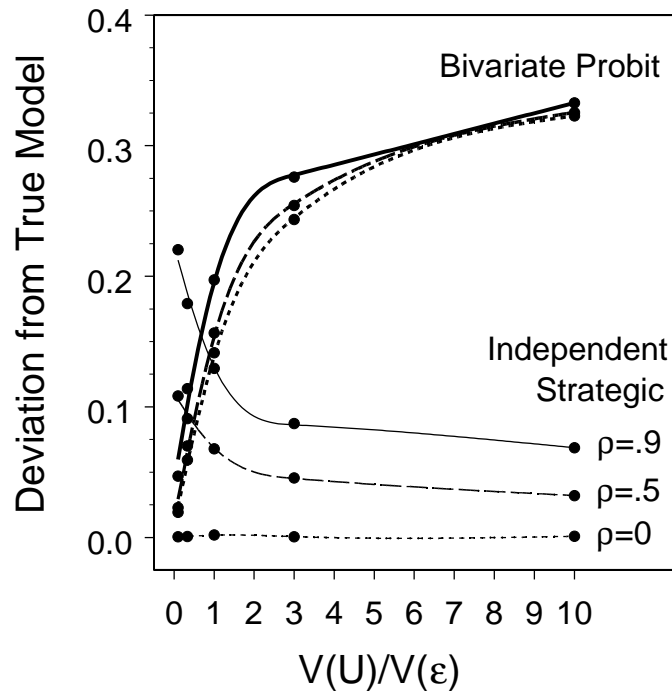
When Strategic Probit Goes Bad?...



When Strategic Probit Goes Bad?...



How Well Do They Approximate the True Model?



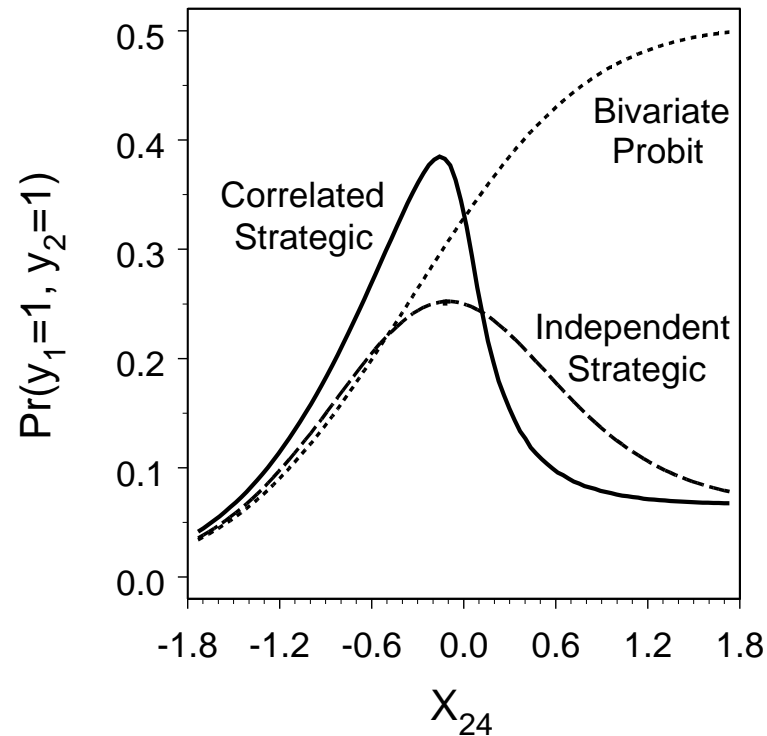
Bivariate Probit

- Best when error dominates. Worst when error matters little.
- Worst of the worst
- Can never be equivalent to true model here

Independent Strategic

- Larger $\rho \rightarrow$ larger misspecification
- Best when utilities and structure dominate
- Can be equivalent to true model here, since special case
- Over entire range, seems to be the lesser of the evils

How Well Do They Approximate the True Model?



- Bivariate Probit does not capture nonmonotonic relationship due to expected utility calculations
- Independent Strategic model does, even though misspecified

Conclusion

- Three tasks in this paper:
 - to clarify that international relations scholars cannot shield themselves from selection bias simply by assuming their results are limited to a given sample
 - to show how recent statistical strategic models relate to traditional selection models and to generalize the two sets of models by deriving a correlated strategic model
 - to examine the effects of misspecifying either the correlated errors or the strategic functional form in this class of selection models.

- Misspecifying either correlated errors or the strategic functional form will result in biased inferences. The analysis conducted here suggests that if one has to “get by” with one canned method versus the other, the better option would be to use a method that models the strategic interaction but forgoes the correlated errors.
- Most international relations researchers employing selection models are usually interested in finding general “processes” by which states escalate crises or enter into war. Indeed, the rationale for using the selection models is to better model the process itself.

Ironically, the results here suggest that the typical selection model is most appropriate when the “process” essentially does not matter.