Homework 12: Selection Models

Use the data set called gssheckman.dta. This is data from the General Social Survey. Data range from 1975 through 1989. If you type 'desc', you'll get a description of the variables. We're going to try and model a respondent's liberal-conservative self-identification. However, we might be worried that selection issues are at play since not everyone answers this question.

Exercise 1: Heckman Selection Model

- Let's start by renaming (and generating) some variables. rename polvmiss missingview; rename educ education; recode sex 2=0; gen male=sex; recode race 2=0; recode race 3=.; gen white = race; rename relig religious; rename polviews conservative;
- Estimate an OLS model where the dependent variable is CONSERVATIVE. The independent variables should be: PARTYID REGION UNION AGE EDUCATION MALE INCOME RELIGIOUS. This model ignores any selection issues. Put the results in the first column of Table 1.

Table 1: Determinants of Political Views

Outcome Equation Dependent Variable: Conservatism

Selection Equation Dependent Variable: Did they respond? 1, 0

Regressor	OLS	By Hand		Two-Step		MLE	
		Selection	Outcome	Selection	Outcome	Selection	Outcome
Education							
Male							
White							
Income							
Religious							
PartyID							
Region							
Union							
Age							
$\hat{\lambda}$							
ρ							
σ_ϵ							
Constant							
Log Likelihood							
Observations							

* $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$ (two-tailed)
(Standard errors are given in parentheses)

- Now, we are going to estimate a Heckman selection model by hand. Let's assume that EDUCATION MALE WHITE INCOME RELIGIOUS influence whether a respondent answers the liberal-conservative question or not. So run a probit model with these independent variables and MISSINGVIEW as the dependent variable. Put the results in the second column of Table 1.

- Now create the inverse Mills ratio and include it as an independent variable in the outcome equation where the dependent variable is CONSERVATIVE and the independent variables are PARTYID REGION UNION AGE EDUCATION MALE INCOME RELIGIOUS. Put the results in the third column of Table 1.

- Write down the overall model that you have estimated i.e. write down the selection and outcome equations as you would report them in a paper.

- Now estimate the exact same model using STATA's two-step version of the 'heckman' command. Put the results in columns 4 and 5 of Table 1.

- Now estimate the exact same model using STATA's MLE version of the 'heckman' command. Put the results in columns 6 and 7 of Table 1.

- Compare the results across all four models - what is the same and what is different? Is there anything weird about the number of observations and the log-likelihood scores? How would you interpret the coefficients on EDUCATION? Do we have selection bias? How do you know? Do all the models indicate selection bias? Should we be wary about the results from any of the models - why?

- Using the results from the MLE model, what are the marginal effects of PARTYID REGION UNION and AGE on a respondent's liberal-conservative score? Provide confidence intervals.

- Still using the results from the MLE model, we're going to look at some other marginal effects. Calculate the marginal effect of EDUCATION, INCOME, and RELIGION on a respondent's liberal-conservative score use STATA's 'mfx compute' command. You should set EDUCATION=12 MALE=0 WHITE=0 INCOME=9 RELIGIOUS=2. What is different about these marginal effects compared to those in the previous bullet point question? Now try to calculate these marginal effects by hand.