Modeling Error: Heckman Selection Models

Zoe Ang

Washington University in St. Louis

December 6, 2018

Overview

Motivation

Heckman Selection Model
- The Model
- Assumptions

Application
- Selection Model
- Treatment Model

Motivating Examples

Selection models can address phenomena such as

1. Workforce participation
2. Compulsory school attendance laws and academic or other outcomes
3. General election candidates in systems with primaries
4. Supreme Court case selection

Sample Selection Decision Tree

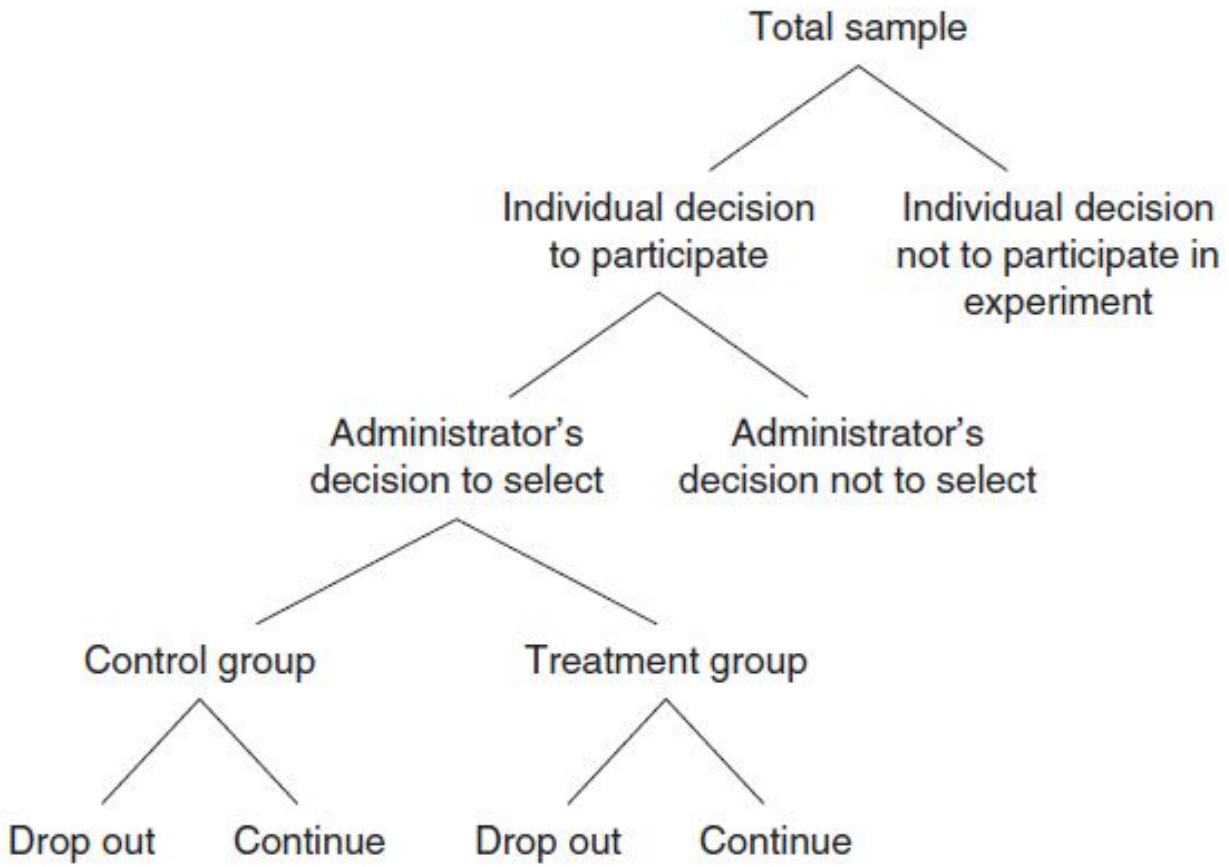

Figure 4.1 Decision Tree for Evaluation of Social Experiments

When to use Selection Models

Use a sample selection model when only selected cases are observed (selection variable=w=1)[1] and:

1. The sample being inferred was not generated randomly;
2. the binary selection explanatory variable, w, was endogenous rather than exogenous; and
3. sample selection or incidental truncation must be considered in the evaluation of the impact of the selection variable.
   - When there is truncation, the cases that are not observed will have an outcome that is systematically different than the observed cases

[1] If both 0s and 1s are present for the selection variable, use a treatment effect model.

A Two Stage Model

The goal is to use the observed variables to estimate the regression coefficients that are applicable to sample participants whose values of w equal either 0 or 1.

Stage 1: Selection

1. Selection equation: considers a portion of the sample whose outcome is observed and mechanisms determining the selection process (pre-selection variables that predict propensity)
   - $w_i^* = z_i\gamma + u_i = \begin{cases} w_i = 1 & w_i^* > 0 \\ w_i = 0 & \text{otherwise} \end{cases}$
   - $\text{Prob}(w_i = 1|z_i) = \Phi(z_i, \gamma)$
   - $\text{Prob}(w_i = 0|z_i) = 1 - \Phi(z_i, \gamma)$
2. The first stage is typically a probit regression (like a logit, but with a normal CDF); tobit regression is used when the dependent variable is censored

A Two Stage Model

Stage 2: Regression

1. Regression equation: considers mechanisms that influence the outcome variable
   - $y_i = x_i\beta + \epsilon_i$ iff $w_i = 1$
2. x_i is a vector of exogenous variables determining the outcome y_i

Inverse Mills Ratio

- The inverse Mills ratio, λ is used to estimate the outcome regression
  - assumes a normal distribution of the population
  - $\lambda(c_z) = \frac{\phi(c_z)}{1-\phi(c_z)}$
  - $c_z = (a - \mu_z)$; a is the cutoff threshold; $\phi$is the standard normal CDF
- For each observation, there is an inverse Mills Ratio δ, which is used to correct for sample selection bias.
- At each observation, the true conditional variance of the disturbance is:

$$\sigma_i^2 = \sigma_\epsilon^2(1 - \rho^2\delta_i)$$

Model Assumptions

1. The errors from the selection equation, μ_i, and the regression equation, ϵ_i, are correlated, notated ρ
2. Both error terms are normally distributed with mean 0.
3. Both error terms are independent from their respective sets of explanatory variables.

The selection equation sets a minimum bound on μ_i. Because μ_i and ϵ_i are correlated, ϵ_i is also bounded. This correlation is notated as ρ.

Application

What is the effect of unions on wages?

- Naive OLS: Regress personal characteristics and a dummy variable for unions on wage
- OLS assumes that participation in a union is exogenous.

Union participation is endogenous and should be modeled directly.

- Observed skill: Low observed skill workers will self-select into union job; high observed skill, non-union jobs
- Unobserved skill: Employers of union jobs will hire low observed and high unobserved skill workers and high observed and low unobserved skill workers.

Application: Balance of Covariates between Groups

Table: Balance of Control and Treatment Groups: Mean and Standard Deviation

Covariate	union=0	union=1
Age	49.1	49.62
	(6.86)	(7.51)
Race	1.69	1.6
	(1.10)	(1.01)
Sex	1.46	1.46
	(0.498)	(0.499)
Education	12.4	12.3
	(3.08)	(3.08)
Experience	29.2	29.4
	(6.78)	(6.99)
Skill	1.49	1.19
	(0.50)	(0.39)
N	500	500

Application: Selection Model

Table: Estimates of the Effect of Unions on Wage

	Dependent variable: Wage	
	OLS	*Heckman Selection*
Age	8.244	8.244
	(18.769)	(18.600)
Black	−808.883**	−808.882**
	(391.526)	(387.986)
Asian	−154.007	−153.996
	(667.005)	(660.970)
Latino	−58.266	−58.262
	(421.048)	(417.240)
Sex	−214.691	−214.686
	(267.704)	(265.283)
Education	−63.740	−63.719
	(42.366)	(41.972)
Experience	9.953	9.953
	(18.961)	(18.789)
Skill	−48.202	
	(336.388)	
Constant	52,313.160***	52,365.840***
	(1,180.508)	(1,261.962)
Observations	500	1,000
Note:	*p<0.1; **p<0.05; ***p<0.01	

Application: Selection Model Output

```
--------------------------------------------
Tobit 2 model (sample selection model)
2-step Heckman / heckit estimation
1000 observations (500 censored and 500 observed)
14 free parameters (df = 987)
Probit selection equation:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.290620   0.177558   1.637    0.102
skill1      -0.860998   0.087963  -9.788   <2e-16 ***
education   -0.000452   0.013363  -0.034    0.973
Outcome equation:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 52365.841   1261.962  41.496   <2e-16 ***
age             8.244     18.600   0.443   0.6577
race1        -808.882    387.986  -2.085   0.0373 *
race2        -153.996    660.970  -0.233   0.8158
race3         -58.262    417.240  -0.140   0.8890
sex1         -214.686    265.283  -0.809   0.4186
education     -63.719     41.972  -1.518   0.1293
experience      9.953     18.789   0.530   0.5964
Multiple R-Squared:0.0167,      Adjusted R-Squared:7e-04
   Error terms:
               Estimate Std. Error t value Pr(>|t|)
invMillsRatio  -84.57697  583.45660  -0.145    0.885
sigma         2908.56952         NA      NA       NA
rho             -0.02908         NA      NA       NA
--------------------------------------------
```

Application: Selection Model

- The substantive and statistical differences between the OLS and selection model are small/relatively nonexistent.
- This implies that the OLS estimate may not have been too bias.
- A low value of ρ is consistent with the little difference between the OLS and selection model
- The insignificant inverse Mills Ratio means that selection bias was not a major concern.

Application: Treatment Model

Table: Estimates of the Effect of Unions on Wage

	Dependent variable: Wage	
	OLS	*Heckman Treatment*
Age	14.505	9.312
	(15.419)	(17.689)
Black	−548.181*	−1,043.947***
	(323.327)	(367.358)
Asian	−138.754	−217.090
	(606.332)	(705.150)
Latino	−41.277	−391.647
	(338.685)	(390.789)
Sex	−34.844	−113.486
	(220.305)	(252.428)
Education	14.055	63.013
	(35.505)	(64.427)
Experience	−5.431	−16.696
	(15.642)	(17.609)
Union Member	10,442.550***	
	(217.823)	
Constant	40,897.330***	46,239.170***
	(1,000.190)	(1,290.391)
Observations	1,000	1,000
Note:		*p<0.1; **p<0.05; ***p<0.01

Application: Treatment Model Output

```
--------------------------------------------
Tobit treatment model (switching regression model)
Maximum Likelihood estimation
Newton-Raphson maximisation, 10 iterations
Return code 2: successive function values within tolerance limit
Log-Likelihood: -10299.6
1000 observations: 500 non-participants (selection 0) and 500 participants (selection 1)

13 free parameters (df = 987)
Probit selection equation:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.12131    0.16655  -0.728    0.467
skill1      -0.32183    0.05289  -6.085 1.67e-09 ***
education    0.00963    0.01272   0.757    0.449
Outcome equation:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 46239.173   1290.391  35.833  < 2e-16 ***
age             9.312     17.689   0.526  0.59869
race1       -1043.947    367.358  -2.842  0.00458 **
race2        -217.090    705.150  -0.308  0.75825
race3        -391.647    390.789  -1.002  0.31649
sex1         -113.486    252.428  -0.450  0.65311
education      63.013     64.427   0.978  0.32829
experience    -16.696     17.609  -0.948  0.34328
   Error terms:
       Estimate Std. Error t value Pr(>|t|)
sigma 6.243e+03  1.397e+02   44.68   <2e-16 ***
rho   9.574e-01  6.046e-03  158.36   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
--------------------------------------------
```