PPOL 503-03, PPOL 503-04, Fall 2016
Course Notes #14: Sample Selection Bias and Treatment Effect Models

I. Sample Selection Bias

1. This concern arises whenever non-random samples are used to estimate regression models.
2. The sampling procedure is random but some values of the dependent variable are not observed
3. Estimation based only on the sub-sample with observable values of the dependent variable introduces bias into the regression coefficients.
4. Bias occurs because the process determining when the dependent variable is observed is related to the unobservable error term in the regression equation.

II. Explanation of Selection Bias

1. In a censored sample, the original sample is random but the dependent variable is missing for censored observations.

$$Y_{1i} = \beta_{1i} X_{1i} + U_{1i} \qquad i = 1,2,3 \ldots.n \qquad (1)$$

$$Y_{2i} = \beta_{2i} X_{2i} + U_{2i} \qquad i = 1,2,3 \ldots.n \qquad (2)$$

2. For each individual in the random sample, we observe Y_{1i} only if $Y_{2i} >$ 0. The effect of the selection rule must be considered when the censored sample is used to estimate the model. When Y_{1i} is observed, the regression function is

$$E\left(Y_{1i} \,\middle|\, X_{1i} \text{, selection rule} \right) =$$

$$\beta_{1i} X_{1i} + E\left(U_{1i} \,\middle|\, U_{2i} => - \beta_{2i} X_{2i} \right) \quad (3)$$

3. Only when the sub-sample with observed values of Y_{1i} represents a randomly selected group will the conditional expectation of U_{1i} = 0. If so, the parameters of the least squares regression are unbiased.

4. Generally, this is not the case. The regression based on the non-random sub-sample omits the conditional mean of the disturbance as an explanatory variable.

III. Least Squares Correction for Selection Bias

1. This approach developed by Olsen (1980) assumes the error term in the selection equation has a uniform distribution and that the conditional expectation of U_1 given U_2 is linear in U_2.

2. Example: estimation of wages based on the sub-sample of persons who work; the original sample is random but it includes many individuals who are not working. The sub-sample of persons who are working may not be a random sample. If one estimates wages for the sub-sample with positive wages bias will occur.

3. Selectivity correction is a four step procedure:

a) estimate the probability of working using a linear probability model; generate predicted values—PR1

b) calculate the weight to correct for heteroskedasticity

$$WT1 = 1 / SqRoot (PR1 (1 – PR1)$$

For predicted probabilities that either are greater than or equal to one or that are less than or equal to zero, the calculated weights will either be imaginary number or infinity. To ensure that all weights are calculated, any predicted value that lies outside the range zero to one is set equal to .5. This minimizes the effects of such observations. Apply this weight and run weighted least squares.

c) Calculate predicted values from the weighted least squares regression—PR2. For values less than zero set to zero. For values greater than one, set to one.

d) Estimate the log of wage income as a function of relevant explanatory variables and this constructed regressor which accounts for the non-randomness of the sample.

e) Correct the log wage equation for heteroskedasticity.

IV. Example: probability of working and wage income for men with and without arthritis (Mitchell and Butler (1986)

1. To identify sample selection, one must include some variables in the probability of working equation that are excluded from wage income equation.

2. Two variables: one called "*disabling conditions*" represent the presence of health problems that are almost certain to prevent an individual from participating in the labor force. These conditions include: multiple sclerosis, mental retardation, stroke, blindness, paralysis, brain injury, spinal bifida, bed ridden or in wheelchair.

Deterring conditions controls for somewhat less disabling conditions that may also deter LFP: diabetes, missing limbs, epilepsy and deafness.

3. Table 2 shows results from the OLS wage income equation without selectivity correction and from the GLS wage income equation for men with arthritis. Table 3 shows a similar comparison for men without arthritis. The selectivity correction is highly significant in both wage income equations.

4. Table 4 compares the wage income differential between men with and without arthritis. Mean wage income of men with arthritis under OLS (no selectivity correction) is $11,861 and $14,720 under the GLS model with selectivity correction. The OLS specification underestimates the wage income of men without arthritis by almost $3,000. Mean wage income of men with arthritis is about $7,100 under OLS and close to $7,600 under GLS.

5. The actual wage gap is $4,760 under the OLS specification and $7,123 under the GLS specification.

6. Calculate the adjusted wage gap by employing the characteristics of those with arthritis in the equation of men without arthritis. The difference between this adjusted amount and the actual predicted wage income of men with arthritis is the net difference estimated to be caused by arthritis. The portion of the total wage income gap due to arthritis is $929 (19.5%) for the OLS specification and $2,320 (32.6%) under the selectivity corrected GLS specification.

V. Mill's Ratio Method to Correct for Sample Selection (aka Heckman Technique

1. Mill's ratio method is a two step estimator proposed by Heckman (1976),which first employs probit analysis to construct a regressor to account for the non-randomness of the sample. In the second step this additional regressor is included in the original model and then estimated by least squares. The estimator described by Heckman assumes that the disturbance structure has a bivariate normal density.

2. Selection equation:

$$Z^* = \alpha_i W_i + u_i$$

$Z_i = 1$ if $Z^* > 0$ and $= 0$ otherwise

$$\text{Prob}(Z_i = 1) = \Phi(\alpha_1 W_i)$$

$$\text{Prob}(Z_i = 0) = 1 - \Phi(\alpha_1 W_i)$$

3. Regression Model:

$Y_i = \beta_i X_i + e_i$ observed only if $Z_i = 1$

$(u_i, e_i) \sim$ bivariate normal $(0, 0, 1, \sigma_e, \rho)$

$$E[Y_i \mid Z_i = 1] = \beta_i X_i + \sigma_e \rho \lambda (\alpha_i W_i)$$

a. Estimate the probit equation predicting the probability of being in the labor force to obtain estimates of (α_i).
b. Compute $\lambda = \varphi(\alpha_1 W_i) / \Phi(\alpha_1 W_i)$
c. Estimate β_i and $\beta_\lambda = \sigma_e \rho$ by the least squares regression of Y on X and λ.

VI. Example: Gender Differences in Wages Losses from Impairments (Baldwin et al. 1994)

1. The objective of this study was to estimate the effect of disabilities on the wages of men and women. Analysis is based on a sample of 10,382 men, of whom 95% (9,910) work and have positive wages; and a sample of 11,055 women, 73.6% (8,142) work and have positive wages.

2. The estimation of wages based on the sub-samples of those who work may result in sample selection bias. This study uses the Heckman two-step procedure to account for possible selection bias.

3. Step 1: estimate a probit model to predict the probability of employment stratified by gender.

$$Y_i = 1 \text{ if employed}, = 0 \text{ if not employed}$$

Step 2: construct the selectivity correction λ (called lambda) from the first stage probit equation.

$$\lambda = \varphi(\alpha_1 W_i) / \Phi(\alpha_1 W_i)$$

Step 3: include λ in the log wage equation to correct for selection bias that may arise from estimating a log wage equation based on the sub-sample of those who are employed.

$$\text{Log (wages)} = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \beta_\lambda \lambda + u_i$$

4. If λ in the log wage equation is statistically significant, this implies that failure to acknowledge that the sub-sample with positive wages is not random will lead to biased results.

5. Identifying variables in the employment equations include various types of non-wage income: property income, veterans' compensation, unemployment/disability income, retirement income, child support, spouse's earnings and other non-wage income. For both men and women, most of the non-wage income variables have significant negative effects on employment participation as predicted by labor theory (See Table 6).

6. The log wage equation results are reported in Table 8. The lambda variable is not statistically significant in the log wage equation for men. This is not surprising because 95% of the sample is working. Thus, if one estimated the log wage equation and failed to account for the potential selection bias associated with the sub-sample of men with positive wages, one would still obtain unbiased estimates.

7. This is not the case for women. The lambda coefficient = .118 (se = .023) is highly significant (1%) level. The significance of lambda in the log wage equation

is not surprising because less than 75% of the sample of women is working and has positive wages. Because the lambda term is highly significant, this implies that if we fail to account for the selection bias associated with the non-random sub-sample of women who work, the parameter estimates will be biased.

VII. Treatment Effects Model

1. This model is appropriate when program participation is voluntary. In this case we know participation status for each person and we also have complete information on the outcome of interest (expenditures, test scores). The problem arises because program participation is voluntary and those individuals who have the most to gain from participation may self-select into the program.

2. Example: To measure the effect of participation in a home and community-based waiver program on monthly expenditures (Anderson and Mitchell, 1997)

$$P^* = \alpha_i X_i + u_i \quad (1)$$

Where $P = 1 \text{ if } P^* > 0 \quad (2)$

$$P = 0 \text{ if } P^* \leq 0$$

P = 1 if the Medicaid recipient chooses to enroll in the Medicaid home and community-based waiver program; = 0 if the Medicaid recipient opts for standard care.

Monthly expenditures are estimated as:

$$E_i = \beta_i Z_i + \delta_i P_i + e \quad (3)$$

The coefficient on “P” measures the expenditure differences between waiver and non-waiver participation controlling for the non-random selection of the waiver.

The model also estimates “ρ” (rho). A positive significant “ρ” implies that patients with a greater propensity to join the waiver also tend to incur higher treatment costs. This could occur, for example, if sicker patients prefer home and community-based services relative to standard care.

3. Results:

a. Probit model predicting waiver participation is highly significant. Men are more likely than women to join the waiver. Whites are more likely than either blacks or Hispanics to enroll. Also find significant regional variation in waiver participation.

b. Chi-square test for instrument validity shows that the three identifying variables (home health/10000 pop, hospital beds/ 10000, case managers/AIDS pop) are uncorrelated with the residuals in the expenditure equation. This implies they are valid instruments.

c. Expenditure equation in which waiver participation is treated as an exogenous RHS variable shows that waiver participants incur 22% lower monthly expenditures compared to non-participants.

d. Expenditure equation controlling for non-random selection (column 3) shows that program participation reduces monthly expenditures by almost 27% ($p<.01$).

e. The estimate of "ρ" = .063 but is not statistically significant. This finding implies that there is no evidence of non-random selection through unobservables not controlled for in the model (morbidity, income, education).

f. In this case, failure to control for non-random program selection in evaluating the waiver initiative does not seriously bias the impact of the program on expenditures per beneficiary.