Applications of Data Analysis (EC969)

Simonetta Longhi and Alita Nandi (ISER)
Contact: slonghi and anandi; @essex.ac.uk

Week 2 Lecture 2: Selection

Accounting for selection bias (assuming Not Missing At Random and ignoring item non-response and measurement error)

Example: What determines wages for women? How can we estimate coefficients of this wage equation that are generalizable to working as well as non-working women?

Input datasets: **`Week2Lecure2.dta`**
Do file in **`Week2Lecure2_DoFile.pdf`**

New commands

```
regress, probit, heckman
```

Accounting for selection bias (assuming Not Missing At Random and ignoring item non-response and measurement error)

What determines wages for women? How can we estimate coefficients of the wage equations that are generalizable to working as well as non-working women?

Input dataset: `Week2Lecure2.dta`

Do file in `Week2Lecure2_DoFile.pdf`

In some cases, when you want to estimate the effect of some factors on a continuous dependent variable, say wages, using OLS will give results that may be applicable to only those for whom wages are observed. Women with young children may not want to participate in the labour market. This may be because the wages that they are likely to receive will not offset the cost of childcare and their disutility from not being able to raise their children themselves. These women generally have high reservation wages than women without young children. So, women with young children who do participate in the labour market do so because the wages they are offered are higher than their reservation wages. This could be because these women have qualities such as cognitive ability, motivation, perseverance that are rewarded highly in the labour market. Suppose (as is often the case) these factors are not observed by the researcher (i.e., not asked in the survey). In such cases, if we estimate wages by OLS we will be including a random sample of women without children and a non-random sample of women with children (those with higher cognitive abilities etc). *This is a case of missing not at random and we need to take account of selection into the labour market if we want the results of the estimation to apply to all women not just those in the labour market.* We can do this by modelling the wage equation jointly with the selection into the labour market. We can use either MLE or Heckman two-step method. The advantage of using MLE is that the estimates are efficient but the disadvantage is that the model may take a long time to converge, if at all.

Wage equation: marital status, education, region of residence, age, age squared, and health status.

Selection equation (Participate or not): marital status, education, region of residence, ethnicity, age, age squared, health status and own young children.

Note the variable "own young children" is included in the selection equation but not in the wage equation. This is an instrumental variable and so, will ensure that coefficients in the wage equation are identified. This is however not necessary for identification as it is also achieved through nonlinearity in the Inverse Mill's Ratio. But as the Inverse Mill's Ratio may be linear for some ranges we may have weak identification. On the other, some claim that it is conceptually difficult to find instrumental variables as some labour market theories show that variables that determine labour market participation also determine wages.

Note if there is no sample selectivity present then the correlation coefficient between the errors of the two equations should be zero.

Before starting with wage estimation we need to create the appropriate variables.

Remember if you want to include any categorical variable then these need to be converted into 0-1 dummy variables. Also, the equation to be estimated will not be linearly independent if all dummy variables of a categorical variable are included. E.g., if you want to include the categorical variable region of residence which has 4 categories – region = 1, 2, 3 and 4. You will first need to convert this into 4 0-1 dummy variables: region1 is 1 if region=1, 0 otherwise. Similarly for region =2, 3 and 4. Now in the equation to be estimated you can include any three of the 4 region dummies (which one is omitted does not matter). The estimated coefficients of the remaining three dummies will be interpreted as the difference in effect from the omitted category. Unless you have a particular preference, generally the dummy with the highest frequency is omitted.

NB: But, Stata 11 allows you to carry out estimations using the categorical variables directly. You simply add “i.” in front of the categorical variable and Stata will understand that this is a categorical variable and transform it into four 0-1 dummies. By default Stata will make the lowest value as the omitted category. If you want to specify a different category, say a category whose value is 2, then simply write “2b.” in front of the variable. For example,

```
      regress loghhincome 5b.edu_highest i.region
```

Stata will then regress log of household income on four education dummies with the 5th category being the omitted category and the region dummies where the lowest region category is the omitted category.

We need to create the following variables for the analysis:

Dependent variable
1. Log of wage

Independent variables
2. age and age-squared

3. Using the highest educational qualification variable, QFEDHI, we have created four 0-1 dummy variables which
– take on the value 1 if no educational qualification, 0 otherwise
– take on the value 1 if highest educational qualification is vocational or training only, 0 otherwise
– take on the value 1 if highest educational qualification is GCSE or o-levels, 0 otherwise
– take on the value 1 if highest educational qualification is A-levels or equivalent, 0 otherwise
– take on the value 1 if highest educational qualification is college or university, 0 otherwise

4. Using the current marital status variable, MASTAT, we have created three 0-1 dummy variables which
– take on the value 1 if never married, 0 otherwise
– take on the value 1 if married or cohabiting, 0 otherwise
– take on the value 1 if separated, divorced or widowed, 0 otherwise

5. Using the series of questions asking respondents if they have had any disability or health problem, HLBRBA-HLBRBL we have created a single 0-1 dummy variable which takes on the value 1 if the person reports any of the 12 health problem or disability (HLBRBA-HLBRBL).

Some independent variables have been provided in the dataset, so we don’t need to create those. These are the region dummies and whether there are young children in the household.

Stata code for this estimation is as follows:

```
        heckman depvar indepvars, select(depvar_s=indepvars_s) first
        heckman depvar indepvars, select(depvar_s=indepvars_s) twostep first
```

`select(...)` specifies the variables and options for the selection equation. If *`depvar_s`* is specified, it should be coded as 0 or 1, with 0 indicating an observation not selected and 1 indicating a selected observation. If *`depvar_s`* is not specified, observations for which *`depvar_s`* is not missing are assumed selected, and those for which *`depvar_s`* is missing are assumed not selected.

Stata can estimate this model using maximum likelihood or Heckman two-step. The default is to produce full ML estimates. **`twostep`** specifies that Heckman's two-step efficient estimates of the parameters, standard errors, and covariance matrix be produced.

`first` specifies that the first-step probit estimates of the selection equation be displayed before estimation.

To get the Inverse Mill's Ratio specify either **`nshazard(`***`newvar`***`)`** or **`mills(`***`newvar`***`)`** in the options. Here *`newvar`* is the name of the new variable containing the Inverse Mill's Ratio

Sometimes when estimating by ML method, the model may not converge if Stata cannot find feasible initial values. In that case it is a good idea to do a basic regression and use those coefficients as the initial values. **`from()`**specifies the initial values for the coefficients

In this case first run:

```
        regress depvar indepvars
```

Save the cofficients vector as:

```
        matrix intB = e(b)
```

Then run the selection model by asking Stata to take the initial values from the vector specified in **`from()`**

```
        heckman depvar indepvars, select(depvar_s indepvars_s) from(intB)
```

When you take a look at the Stata output you will see that in addition to the estimated coefficients, their standard errors and p-values Stata also produces the following values and their standard errors.

rho in the Stata output represents the estimated correlation coefficient between the error terms in the two equations, and

/athrho: Inverse hyperbolic tangent of **rho**

sigma in the Stata output represents the estimated standard deviation of the error term in the wage equation

lambda in the Stata output is the the estimated coefficient of the Inverse Mill's Ratio (it is also the product of **sigma** and **rho)**.

When ML method is used Stata also reports the likelihood ratio test for **rho=0** and the t-test for **/athrho**=0. Both are equivalent.

When two-step method is used Stata reports the test for **lambda=0** which is equivalent to testing for **rho=0**.

Estimate women's wages by using Heckman two-step. Test if sample selectivity is present.

Estimate women's wages by using MLE and if it does not converge then take the coefficients estimated by OLS as initial values. Test if sample selectivity is present.

Estimate the wage model using OLS and see if the estimated coefficients in the wage model estimate using OLS is different from those estimated after correcting for sample selectivity.

[Optional] How to create datasets **`Week2Lecure2.dta`**

We have provided these datasets, but if you wanted to create these yourself, here is a guide to do that. See **Week2_dataprep_DoFile.pdf** which contains the corresponding do file for this. The dataset **`Week2Lecure2.dta`** is the same as **`Week2Lecure1.dta`** with a few additional sample selection restrictions:

(i) Keep only those who were interviewed face-to-face

(ii) Drop all those cases for which any of these variables have a missing value (HLPRBA, QFEDHI, SEX, AGE, RACEL, REGION, MASTAT, EMPLOYED OR YOUNGCHILDREN)

Reference:

Heckman, James. 1979. "Sample Selection Bias as a Specification Error" *Econometrica*, 47(1): 153-162.

Vella, Francis. 1998. "Estimating Models with Sample Selection Bias: A Survey" *The Journal of Human Resources*. 33(1):127-169.