Heckman Selection Model: SW 683

This example is taken from http://www.gseis.ucla.edu/courses/ed231c/notes3/selection.html

Consider a model, in which, we try to predict women's wages from their education and age. We have an artificially constructed example of a sample of 2,000 women but we only have wage data for 1,343 of them. The remaining 657 women were not working and so did not receive wages. We will start off with a simple-minded model in which we estimate the regression model using only the observations that have wage data.

First Try

```
use http://www.gseis.ucla.edu/courses/data/wages

univar wage education age
                                         -------------- Quantiles -------------
-
Variable         n      Mean      S.D.       Min       .25       Mdn       .75
Max
-----------------------------------------------------------------------------
    wage      1343     23.69      6.31      5.88     19.31     23.51     28.05     45.81
education     2000     13.08      3.05     10.00     10.00     12.00     16.00     20.00
     age      2000     36.21      8.29     20.00     30.00     36.00     42.00     59.00
-----------------------------------------------------------------------------

regress wage education age

      Source |       SS       df       MS              Number of obs =    1343
-------------+------------------------------           F(  2,  1340) =  227.49
       Model |  13524.0337     2  6762.01687           Prob > F      =  0.0000
    Residual |  39830.8609  1340  29.7245231           R-squared     =  0.2535
-------------+------------------------------           Adj R-squared =  0.2524
       Total |  53354.8946  1342  39.7577456           Root MSE      =   5.452

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   education |   .8965829   .0498061    18.00   0.000     .7988765    .9942893
         age |   .1465739   .0187135     7.83   0.000      .109863    .1832848
       _cons |   6.084875   .8896182     6.84   0.000     4.339679    7.830071
------------------------------------------------------------------------------

predict pwage
```

This analysis would be fine if, in fact, the missing wage data were missing completely at random. However, the decision to work or not work was made by the individual woman. Thus, those who were not working constitute a self-selected sample and not a random sample. It is likely some of the women that would earn low wages choose not to work and this would account for much of the missing wage data. Thus, it is likely that we will over estimate the wages of the women in the population. So, somehow, we need to account for information that we have on the non-working women. Maybe, we could replace the missing values with zeros. The variable **wage0** does the trick.

Second Try

```
univar wage0

Variable        n      Mean      S.D.       Min       .25       Mdn       .75       Max
-------------------------------------------------------------------------------
   wage0     2000     15.91     12.27      0.00      0.00     19.39     25.77     45.81
-------------------------------------------------------------------------------
```

```
regress wage0 education age

      Source |       SS       df       MS              Number of obs =    2000
-------------+------------------------------           F(  2,  1997) =  208.32
       Model |  51956.6949     2  25978.3475           Prob > F      =  0.0000
    Residual |  249038.262  1997   124.70619           R-squared     =  0.1726
-------------+------------------------------           Adj R-squared =  0.1718
       Total |  300994.957  1999  150.572765           Root MSE      =  11.167

------------------------------------------------------------------------------
       wage0 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   education |   1.064572   .0844208    12.61   0.000     .8990101    1.230134
         age |   .3907662   .0310308    12.59   0.000     .3299101    .4516223
       _cons |  -12.16843   1.398146    -8.70   0.000    -14.91041   -9.426456
------------------------------------------------------------------------------
```

```
predict pwage0
```

This analysis is also troubling. Its true that we are using data from all 2,000 women but using zero is not a fair estimate of what the women would have earned if they had chose to work. It is likely that this model will under estimate the wages of women in the population. The solution to our quandary is to use the Heckman selection model (Gronau 1974, Lewis 1974, Heckman 1976).

The Heckman selection model is a two equation model. First, there is the regression model,

$$\mathbf{y = v\beta + u_1}$$

And second, there is the selection model,

$$\mathbf{z\gamma + u_2 > 0}$$

Where the following holds,

$$\mathbf{u_1 \sim N(0,\sigma)}$$
$$\mathbf{u_2 \sim N(0, 1)}$$
$$\mathbf{corr(u_1, u_2) = \rho}$$

When $\rho = 0$ OLS regression provides unbiased estimates, when ρ ~= 0 the OLS estimates are biased. The Heckman selection model allows us to use information from non-working women to improve the estimates of the parameters in the regression model. The Heckman selection model provides consistent, asymptotically efficient estimates for all parameters in the model.

In our example, we have one model predicting wages and one model predicting whether a women will be working. We will use **married**, **children**, **education** and **age** to predict selection. Checkout this probit example.

```
generate s=wage~=.

tab s

          s |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        657       32.85       32.85
          1 |       1343       67.15      100.00
------------+-----------------------------------
      Total |       2000      100.00

probit s married children education age

Probit estimates                                  Number of obs   =       2000
                                                  LR chi2(4)      =     478.32
                                                  Prob > chi2     =     0.0000
Log likelihood = -1027.0616                       Pseudo R2       =     0.1889

------------------------------------------------------------------------------
           s |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     married |   .4308575    .074208     5.81   0.000     .2854125    .5763025
    children |   .4473249   .0287417    15.56   0.000     .3909922    .5036576
   education |   .0583645   .0109742     5.32   0.000     .0368555    .0798735
         age |   .0347211   .0042293     8.21   0.000     .0264318    .0430105
       _cons |  -2.467365   .1925635   -12.81   0.000    -2.844782   -2.089948
------------------------------------------------------------------------------
```

Now we are ready to try the full Heckman selection model.

Third Time's a Charm

```
heckman wage education age, select(married children education age)
/* can also be written as
  heckman wage education age, select(s=married children education age)  */

Heckman selection model                         Number of obs      =      2000
(regression model with sample selection)        Censored obs       =       657
                                                Uncensored obs     =      1343

                                                Wald chi2(2)       =    508.44
Log likelihood = -5178.304                      Prob > chi2        =    0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
wage         |
   education |   .9899537   .0532565    18.59   0.000     .8855729    1.094334
         age |   .2131294   .0206031    10.34   0.000     .1727481    .2535108
       _cons |   .4857752   1.077037     0.45   0.652    -1.625179     2.59673
-------------+----------------------------------------------------------------
select       |
```

```
     married |   .4451721   .0673954     6.61   0.000      .3130794    .5772647
    children |   .4387068   .0277828    15.79   0.000      .3842534    .4931601
   education |   .0557318   .0107349     5.19   0.000      .0346917    .0767718
         age |   .0365098   .0041533     8.79   0.000      .0283694    .0446502
       _cons |  -2.491015   .1893402   -13.16   0.000     -2.862115   -2.119915
-------------+----------------------------------------------------------------
     /athrho |   .8742086   .1014225     8.62   0.000      .6754241    1.072993
    /lnsigma |   1.792559    .027598    64.95   0.000      1.738468     1.84665
-------------+----------------------------------------------------------------
         rho |   .7035061   .0512264                      .5885365    .7905862
       sigma |   6.004797   .1657202                       5.68862    6.338548
      lambda |   4.224412   .3992265                      3.441942    5.006881
------------------------------------------------------------------------------
LR test of indep. eqns. (rho = 0):   chi2(1) =    61.20   Prob > chi2 = 0.0000
------------------------------------------------------------------------------
```

predict pheckman

In addition to the two equations, **heckman** estimates rho (actually the inverse hyperbolic tangent of rho) the correlation of the residuals in the two equations and sigma (actually the log of sigma) the standard error of the residuals of the wage equation. Lambda is rho*sigma. The output also includes a likelihood ratio test of rho = 0.

Recall that it was stated at the beginning that this dataset was constructed. As it turns out, we do have full wage information on all 2,000 women. The variable **wagefull** has the complete wage data. We can therefore run a regression using the full wage information to use as a comarison.

regress wagefull education age

```
      Source |       SS       df       MS              Number of obs =    2000
-------------+------------------------------           F(  2,  1997) =  398.82
       Model |   28053.371     2  14026.6855           Prob > F      =  0.0000
    Residual |  70234.8124  1997  35.1701614           R-squared     =  0.2854
-------------+------------------------------           Adj R-squared =  0.2847
       Total |  98288.1834  1999   49.168676           Root MSE      =  5.9304

------------------------------------------------------------------------------
    wagefull |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   education |   1.004456   .0448325    22.40   0.000      .9165328    1.092379
         age |   .1874822   .0164792    11.38   0.000       .155164    .2198004
       _cons |   1.381099   .7424989     1.86   0.063     -.0750544    2.837253
------------------------------------------------------------------------------
```

predict pfull

If we compare (see below) the predicted wages from the first model (omit missing), the second model (substitute zero for missing) and the heckman model to the complete wage and predicted full wage values, we note the following:
1) The first model tends to over predict wages;
2) the second model tends to way underestimate wages;
3) the heckman model does the best job in predicting wages.

univar pwage pwage0 pheckman wagefull pfull

```
                                   -------------- Quantiles --------------
Variable        n     Mean     S.D.      Min      .25      Mdn      .75      Max
-------------------------------------------------------------------------------
   pwage     2000    23.12     3.24    17.98    20.36    22.56    25.71    32.66
  pwage0     2000    15.91     5.10     6.29    11.76    15.95    19.36    32.18
pheckman     2000    21.16     3.84    14.65    18.06    20.83    24.00    32.86
wagefull     2000    21.31     7.01    -1.68    16.46    21.18    26.14    45.81
   pfull     2000    21.31     3.75    15.18    18.18    20.77    24.20    32.53
-------------------------------------------------------------------------------
```

Probit with Selection

Stata also includes another selection model the **heckprob** which works in a manner very similar to **heckman** except that the response variable is binary. **heckprob** stands for heckman probit estimation. We can illustrate **heckprob** using a dataset schvote that we also used in a bivariate probit example. This time we will predict going to private school (**priv**) with selection determined on whether the individual voted to increase property taxes (**vote**). Admittedly, this example is more than a bit contrived.

```
use http://www.gseis.ucla.edu/courses/data/schvote

tab1 priv vote

-> tabulation of priv

    private |
     school |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         70       87.50       87.50
          1 |         10       12.50      100.00
------------+-----------------------------------
      Total |         80      100.00

-> tabulation of vote

  voted for |
        tax |
   increase |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         29       36.25       36.25
          1 |         51       63.75      100.00
------------+-----------------------------------
      Total |         80      100.00

univar years inc ptax
                                   -------------- Quantiles --------------
Variable        n     Mean     S.D.      Min      .25      Mdn      .75      Max
-------------------------------------------------------------------------------
   years       80     8.78     9.91     1.00     3.00     5.00    11.00    49.00
     inc       80     9.97     0.42     8.29     9.77    10.02    10.22    10.82
    ptax       80     6.94     0.33     5.99     6.75     7.05     7.05     7.50
-------------------------------------------------------------------------------


heckprob priv years ptax, select(vote=years inc ptax)
```

```
Probit model with sample selection               Number of obs      =         80
                                                  Censored obs       =         29
                                                  Uncensored obs     =         51

                                                  Wald chi2(2)       =       1.10
Log likelihood = -60.49573                        Prob > chi2        =     0.5771

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
priv         |
       years |  -.1508048   .1602441    -0.94   0.347    -.4648774    .1632678
        ptax |   .2507531   1.228139     0.20   0.838    -2.156356    2.657862
       _cons |  -2.127264     8.3273    -0.26   0.798    -18.44847    14.19394
-------------+----------------------------------------------------------------
vote         |
       years |  -.0082359   .0159395    -0.52   0.605    -.0394767     .023005
         inc |   1.572097   .5672177     2.77   0.006     .4603703    2.683823
        ptax |  -2.019357   .7200663    -2.80   0.005    -3.430661   -.6080533
       _cons |  -1.203783   4.465327    -0.27   0.787    -9.955663    7.548096
-------------+----------------------------------------------------------------
     /athrho |  -.4722769   1.254446    -0.38   0.707    -2.930946    1.986392
-------------+----------------------------------------------------------------
         rho |  -.4400372   1.011544                     -.9943244    .9630535
------------------------------------------------------------------------------
LR test of indep. eqns. (rho = 0):   chi2(1) =     0.11   Prob > chi2 = 0.7392
------------------------------------------------------------------------------
```

The **heckprob** command shares a number of features with **biprobit** models. Both involve two equations, both of which are probit models. Both have correlated residuals from the two equations. Here is a similar **biprobit** model using **priv** and **vote** as response variables looks like.

```
biprobit priv vote years ptax inc

Bivariate probit regression                         Number of obs   =         80
                                                    Wald chi2(6)    =      11.91
Log likelihood = -74.171253                         Prob > chi2     =     0.0640

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
priv         |
       years |  -.0146627   .0264275    -0.55   0.579    -.0664596    .0371342
        ptax |  -.0923143   .6922562    -0.13   0.894    -1.449112    1.264483
         inc |   .3644544   .5588324     0.65   0.514    -.7308371    1.459746
       _cons |  -4.040363   4.872994    -0.83   0.407    -13.59126    5.510529
-------------+----------------------------------------------------------------
vote         |
       years |   -.008866   .0159739    -0.56   0.579    -.0401742    .0224422
        ptax |  -2.054462   .7310168    -2.81   0.005    -3.487229   -.6216959
         inc |   1.574388   .5638432     2.79   0.005      .469276    2.679501
       _cons |  -.9732729   4.487075    -0.22   0.828    -9.767779    7.821233
-------------+----------------------------------------------------------------
     /athrho |  -.3425239   .2536544    -1.35   0.177    -.8396774    .1546297
-------------+----------------------------------------------------------------
         rho |  -.3297287   .2260769                     -.6856382    .1534089
------------------------------------------------------------------------------
```

```
Likelihood ratio test of rho=0:     chi2(1) =  1.95532    Prob > chi2 = 0.1620
```

http://www.gseis.ucla.edu/courses/ed231c/notes3/selection.html