SAMPLE SELECTION BIAS

$$Y_i^* = \beta_1 + \beta_2 X_i + u_i$$

$$Y_i = Y_i^* \text{ if } Y_i^* > 0$$

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

In the tobit model, the values of the regressors and the disturbance term determine whether or not an observation falls into the participating category (*Y* > 0) or the non-participating category (*Y* = 0).

SAMPLE SELECTION BIAS

$$Y_i^* = \beta_1 + \beta_2 X_i + u_i$$

$$Y_i = Y_i^* \text{ if } Y_i^* > 0$$

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

However, the decision to participate may depend on factors other than those in the regression model, in which case a more general model specification with a two-stage process may be required.

SAMPLE SELECTION BIAS

$$B_i^* = \delta_1 + \sum_{j=2}^{m} \delta_j Q_{ji} + \varepsilon_i$$

The first stage, the selection process (decision to participate), depends on the net benefit of participating, B^*, a latent (unobservable) variable that depends on a set of variables Q_j and a random term ε:

SAMPLE SELECTION BIAS

$$B_i^* = \delta_1 + \sum_{j=2}^{m} \delta_j Q_{ji} + \varepsilon_i$$

$$Y_i^* = \beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} + u_i$$

$$Y_i = Y_i^* \text{ for } B_i^* > 0$$

$$Y_i \text{ unobserved if } B_i^* \leq 0$$

The second stage, the regression model, is parallel to that for the tobit model. The *X* variables may include some of the *Q* variables. A sufficient condition for identification is that at least one *Q* variable is not also an *X* variable.

SAMPLE SELECTION BIAS

$$B_i^* = \delta_1 + \sum_{j=2}^{m} \delta_j Q_{ji} + \varepsilon_i$$

$$Y_i^* = \beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} + u_i$$

$$Y_i = Y_i^* \text{ for } B_i^* > 0$$

$$Y_i \text{ unobserved if } B_i^* \leq 0$$

$$E(u_i \mid B_i^* > 0)$$

The expected value of *u* for an observation in the sample is its expected value conditional on *B > 0 (if this condition is not satisfied, the observation will not be in the sample).**

SAMPLE SELECTION BIAS

$$B_i^* = \delta_1 + \sum_{j=2}^{m} \delta_j Q_{ji} + \varepsilon_i$$

$$Y_i^* = \beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} + u_i$$

$$B_i^* > 0 \Leftrightarrow \varepsilon_i > -\delta_1 - \sum \delta_j Q_{ji}$$

$Y_i = Y_i^*$ for $B_i^* > 0$

Y_i unobserved if $B_i^* \leq 0$

$$E(u_i \mid B_i^* > 0) = E(u_i \mid \varepsilon_i > -\delta_1 - \sum \delta_j Q_{ji})$$

B* will be greater than 0 if ε satisfies the inequality shown.

SAMPLE SELECTION BIAS

$$B_i^* = \delta_1 + \sum_{j=2}^{m} \delta_j Q_{ji} + \varepsilon_i \qquad Y_i^* = \beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} + u_i$$

$$Y_i = Y_i^* \text{ for } B_i^* > 0$$

$$Y_i \text{ unobserved if } B_i^* \leq 0$$

$$E(u_i \mid B_i^* > 0) = E(u_i \mid \varepsilon_i > -\delta_1 - \sum \delta_j Q_{ji}) = \frac{\sigma_{\varepsilon u}}{\sigma_\varepsilon} \lambda_i$$

$$\lambda_i = \frac{f(v_i)}{F(v_i)} \qquad v_i = \frac{-\delta_1 - \sum \delta_j Q_{ji}}{\sigma_\varepsilon}$$

It can be shown that $E(u_i)$ is equal to $\sigma_{\varepsilon u}$, the population covariance of ε and u, divided by σ_ε, the standard deviation of ε, multiplied by λ_i, defined as shown.

SAMPLE SELECTION BIAS

$$B_i^* = \delta_1 + \sum_{j=2}^{m} \delta_j Q_{ji} + \varepsilon_i \qquad Y_i^* = \beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} + u_i$$

$$Y_i = Y_i^* \text{ for } B_i^* > 0$$

$$Y_i \text{ unobserved if } B_i^* \leq 0$$

$$E(u_i \mid B_i^* > 0) = E(u_i \mid \varepsilon_i > -\delta_1 - \sum \delta_j Q_{ji}) = \frac{\sigma_{\varepsilon u}}{\sigma_\varepsilon} \lambda_i$$

$$\lambda_i = \frac{f(v_i)}{F(v_i)} \qquad v_i = \frac{-\delta_1 - \sum \delta_j Q_{ji}}{\sigma_\varepsilon}$$

λ_i is usually described as the inverse of Mills' ratio. *f* is the density function of the standardized normal distribution and *F* is the cumulative standardized normal distribution.

SAMPLE SELECTION BIAS

$$B_i^* = \delta_1 + \sum_{j=2}^{m} \delta_j Q_{ji} + \varepsilon_i \qquad Y_i^* = \beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} + u_i$$

$$Y_i = Y_i^* \text{ for } B_i^* > 0$$

$$Y_i \text{ unobserved if } B_i^* \leq 0$$

$$E(u_i \mid B_i^* > 0) = E(u_i \mid \varepsilon_i > -\delta_1 - \sum \delta_j Q_{ji}) = \frac{\sigma_{\varepsilon u}}{\sigma_\varepsilon} \lambda_i$$

$$\lambda_i = \frac{f(v_i)}{F(v_i)} \qquad v_i = \frac{-\delta_1 - \sum \delta_j Q_{ji}}{\sigma_\varepsilon}$$

$$E(Y_i \mid \varepsilon_i > -\delta_1 - \sum \delta_j Q_{ji})$$

The nonstochastic component of *Y* in observation *i* is, as usual, its expected value. In this model the expected value must take into account the condition that the observation appears in the sample.

SAMPLE SELECTION BIAS

$$B_i^* = \delta_1 + \sum_{j=2}^{m} \delta_j Q_{ji} + \varepsilon_i \qquad Y_i^* = \beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} + u_i$$

$$Y_i = Y_i^* \text{ for } B_i^* > 0$$

$$Y_i \text{ unobserved if } B_i^* \leq 0$$

$$E(u_i \mid B_i^* > 0) = E(u_i \mid \varepsilon_i > -\delta_1 - \sum \delta_j Q_{ji}) = \frac{\sigma_{\varepsilon u}}{\sigma_\varepsilon} \lambda_i$$

$$\lambda_i = \frac{f(v_i)}{F(v_i)} \qquad v_i = \frac{-\delta_1 - \sum \delta_j Q_{ji}}{\sigma_\varepsilon}$$

$$E(Y_i \mid \varepsilon_i > -\delta_1 - \sum \delta_j Q_{ji}) = E(\beta_1 + \sum \beta_j X_{ji} + u_i \mid \varepsilon_i > -\delta_1 - \sum \delta_j Q_{ji})$$

Substituting for Y_i, we obtain the expression shown.

SAMPLE SELECTION BIAS

$$B_i^* = \delta_1 + \sum_{j=2}^{m} \delta_j Q_{ji} + \varepsilon_i \qquad Y_i^* = \beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} + u_i$$

$$Y_i = Y_i^* \text{ for } B_i^* > 0$$

$$Y_i \text{ unobserved if } B_i^* \leq 0$$

$$E(u_i \mid B_i^* > 0) = E(u_i \mid \varepsilon_i > -\delta_1 - \sum \delta_j Q_{ji}) = \frac{\sigma_{\varepsilon u}}{\sigma_\varepsilon} \lambda_i$$

$$\lambda_i = \frac{f(v_i)}{F(v_i)} \qquad v_i = \frac{-\delta_1 - \sum \delta_j Q_{ji}}{\sigma_\varepsilon}$$

$$E(Y_i \mid \varepsilon_i > -\delta_1 - \sum \delta_j Q_{ji}) = E(\beta_1 + \sum \beta_j X_{ji} + u_i \mid \varepsilon_i > -\delta_1 - \sum \delta_j Q_{ji})$$

$$= \beta_1 + \sum \beta_j X_{ji} + \frac{\sigma_{\varepsilon u}}{\sigma_\varepsilon} \lambda_i$$

The first two components of Y_i are not affected by taking expectations. The expected value of u_i is as shown above.

SAMPLE SELECTION BIAS

$$B_i^* = \delta_1 + \sum_{j=2}^{m} \delta_j Q_{ji} + \varepsilon_i \qquad Y_i^* = \beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} + u_i$$

$$Y_i = Y_i^* \text{ for } B_i^* > 0$$

$$Y_i \text{ unobserved if } B_i^* \leq 0$$

$$E(u_i \mid B_i^* > 0) = E(u_i \mid \varepsilon_i > -\delta_1 - \sum \delta_j Q_{ji}) = \frac{\sigma_{\varepsilon u}}{\sigma_\varepsilon} \lambda_i$$

$$\lambda_i = \frac{f(v_i)}{F(v_i)} \qquad v_i = \frac{-\delta_1 - \sum \delta_j Q_{ji}}{\sigma_\varepsilon}$$

$$E(Y_i \mid \varepsilon_i > -\delta_1 - \sum \delta_j Q_{ji}) = E(\beta_1 + \sum \beta_j X_{ji} + u_i \mid \varepsilon_i > -\delta_1 - \sum \delta_j Q_{ji})$$

$$= \beta_1 + \sum \beta_j X_{ji} + \frac{\sigma_{\varepsilon u}}{\sigma_\varepsilon} \lambda_i$$

If the random component of the selection process is distributed independently of the random component of the function for *Y*, the population covariance of ε and *u* will be 0 and the term involving the inverse of Mills' ratio drops out.

SAMPLE SELECTION BIAS

$$B_i^* = \delta_1 + \sum_{j=2}^{m} \delta_j Q_{ji} + \varepsilon_i \qquad Y_i^* = \beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} + u_i$$

$$Y_i = Y_i^* \text{ for } B_i^* > 0$$

$$Y_i \text{ unobserved if } B_i^* \leq 0$$

$$E(u_i \mid B_i^* > 0) = E(u_i \mid \varepsilon_i > -\delta_1 - \sum \delta_j Q_{ji}) = \frac{\sigma_{\varepsilon u}}{\sigma_\varepsilon} \lambda_i$$

$$\lambda_i = \frac{f(v_i)}{F(v_i)} \qquad v_i = \frac{-\delta_1 - \sum \delta_j Q_{ji}}{\sigma_\varepsilon}$$

$$E(Y_i \mid \varepsilon_i > -\delta_1 - \sum \delta_j Q_{ji}) = E(\beta_1 + \sum \beta_j X_{ji} + u_i \mid \varepsilon_i > -\delta_1 - \sum \delta_j Q_{ji})$$

$$= \beta_1 + \sum \beta_j X_{ji}$$

In that case we could use least squares to fit the model as usual.

SAMPLE SELECTION BIAS

$$B_i^* = \delta_1 + \sum_{j=2}^{m} \delta_j Q_{ji} + \varepsilon_i \qquad Y_i^* = \beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} + u_i$$

$$Y_i = Y_i^* \text{ for } B_i^* > 0$$

$$Y_i \text{ unobserved if } B_i^* \leq 0$$

$$E(u_i \mid B_i^* > 0) = E(u_i \mid \varepsilon_i > -\delta_1 - \sum \delta_j Q_{ji}) = \frac{\sigma_{\varepsilon u}}{\sigma_\varepsilon} \lambda_i$$

$$\lambda_i = \frac{f(v_i)}{F(v_i)} \qquad v_i = \frac{-\delta_1 - \sum \delta_j Q_{ji}}{\sigma_\varepsilon}$$

$$E(Y_i \mid \varepsilon_i > -\delta_1 - \sum \delta_j Q_{ji}) = E(\beta_1 + \sum \beta_j X_{ji} + u_i \mid \varepsilon_i > -\delta_1 - \sum \delta_j Q_{ji})$$

$$= \beta_1 + \sum \beta_j X_{ji} + \frac{\sigma_{\varepsilon u}}{\sigma_\varepsilon} \lambda_i$$

However, in some situations it is reasonable to hypothesize that the random components are not distributed independently because some of the unobserved characteristics affecting *Y also affect the selection process.**

SAMPLE SELECTION BIAS

$$B_i^* = \delta_1 + \sum_{j=2}^{m} \delta_j Q_{ji} + \varepsilon_i \qquad Y_i^* = \beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} + u_i$$

$$Y_i = Y_i^* \text{ for } B_i^* > 0$$

$$Y_i \text{ unobserved if } B_i^* \leq 0$$

$$E(u_i \mid B_i^* > 0) = E(u_i \mid \varepsilon_i > -\delta_1 - \sum \delta_j Q_{ji}) = \frac{\sigma_{\varepsilon u}}{\sigma_\varepsilon} \lambda_i$$

$$\lambda_i = \frac{f(v_i)}{F(v_i)} \qquad v_i = \frac{-\delta_1 - \sum \delta_j Q_{ji}}{\sigma_\varepsilon}$$

$$E(Y_i \mid \varepsilon_i > -\delta_1 - \sum \delta_j Q_{ji}) = E(\beta_1 + \sum \beta_j X_{ji} + u_i \mid \varepsilon_i > -\delta_1 - \sum \delta_j Q_{ji})$$

$$= \beta_1 + \sum \beta_j X_{ji} + \frac{\sigma_{\varepsilon u}}{\sigma_\varepsilon} \lambda_i$$

Observations in the sample will then be systematically different from those not in the sample and the model is said to be subject to sample selection bias.

SAMPLE SELECTION BIAS

$$B_i^* = \delta_1 + \sum_{j=2}^{m} \delta_j Q_{ji} + \varepsilon_i \qquad Y_i^* = \beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} + u_i$$

$$Y_i = Y_i^* \text{ for } B_i^* > 0$$

$$Y_i \text{ unobserved if } B_i^* \leq 0$$

$$E(u_i \mid B_i^* > 0) = E(u_i \mid \varepsilon_i > -\delta_1 - \sum \delta_j Q_{ji}) = \frac{\sigma_{\varepsilon u}}{\sigma_\varepsilon} \lambda_i$$

$$\lambda_i = \frac{f(v_i)}{F(v_i)} \qquad v_i = \frac{-\delta_1 - \sum \delta_j Q_{ji}}{\sigma_\varepsilon}$$

$$E(Y_i \mid \varepsilon_i > -\delta_1 - \sum \delta_j Q_{ji}) = E(\beta_1 + \sum \beta_j X_{ji} + u_i \mid \varepsilon_i > -\delta_1 - \sum \delta_j Q_{ji})$$

$$= \beta_1 + \sum \beta_j X_{ji} + \frac{\sigma_{\varepsilon u}}{\sigma_\varepsilon} \lambda_i$$

Effectively, this is a special type of omitted variable bias. A regression of *Y* on the *X* variables will yield inconsistent estimates caused by the omission of the λ_i term. Note that λ_i has different values in different observations and is therefore a special type of variable.

SAMPLE SELECTION BIAS

$$B_i^* = \delta_1 + \sum_{j=2}^{m} \delta_j Q_{ji} + \varepsilon_i \qquad Y_i^* = \beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} + u_i$$

$$Y_i = Y_i^* \text{ for } B_i^* > 0$$

$$Y_i \text{ unobserved if } B_i^* \leq 0$$

$$E(u_i \mid B_i^* > 0) = E(u_i \mid \varepsilon_i > -\delta_1 - \sum \delta_j Q_{ji}) = \frac{\sigma_{\varepsilon u}}{\sigma_\varepsilon} \lambda_i$$

$$\lambda_i = \frac{f(v_i)}{F(v_i)} \qquad v_i = \frac{-\delta_1 - \sum \delta_j Q_{ji}}{\sigma_\varepsilon}$$

$$E(Y_i \mid \varepsilon_i > -\delta_1 - \sum \delta_j Q_{ji}) = E(\beta_1 + \sum \beta_j X_{ji} + u_i \mid \varepsilon_i > -\delta_1 - \sum \delta_j Q_{ji})$$

$$= \beta_1 + \sum \beta_j X_{ji} + \frac{\sigma_{\varepsilon u}}{\sigma_\varepsilon} \lambda_i$$

However, since its components depend only on the selection process, λ_i can be estimated from the results of probit analysis of the selection process. This is the first step of the Heckman two-step procedure.

SAMPLE SELECTION BIAS

$$B_i^* = \delta_1 + \sum_{j=2}^{m} \delta_j Q_{ji} + \varepsilon_i \qquad Y_i^* = \beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} + u_i$$

$$Y_i = Y_i^* \text{ for } B_i^* > 0$$

$$Y_i \text{ unobserved if } B_i^* \leq 0$$

$$E(u_i \mid B_i^* > 0) = E(u_i \mid \varepsilon_i > -\delta_1 - \sum \delta_j Q_{ji}) = \frac{\sigma_{\varepsilon u}}{\sigma_\varepsilon} \lambda_i$$

$$\lambda_i = \frac{f(v_i)}{F(v_i)} \qquad v_i = \frac{-\delta_1 - \sum \delta_j Q_{ji}}{\sigma_\varepsilon}$$

$$E(Y_i \mid \varepsilon_i > -\delta_1 - \sum \delta_j Q_{ji}) = E(\beta_1 + \sum \beta_j X_{ji} + u_i \mid \varepsilon_i > -\delta_1 - \sum \delta_j Q_{ji})$$

$$= \beta_1 + \sum \beta_j X_{ji} + \frac{\sigma_{\varepsilon u}}{\sigma_\varepsilon} \lambda_i$$

The second step is to regress *Y* on the *X* variables and the estimated λ.

SAMPLE SELECTION BIAS

```
. heckman LGEARN S ASVABC ETHBLACK ETHHISP if MALE==0, select(S AGE CHILDL06
CHILDL16 MARRIED ETHBLACK ETHHISP)
```

We will illustrate the heckman procedure by fitting an earnings function for females using the LFP data set on the website. The includes 2,661 females, of whom 2,021 had earnings, in 1994.

SAMPLE SELECTION BIAS

```
. heckman LGEARN S ASVABC ETHBLACK ETHHISP if MALE==0, select(S AGE CHILDL06
CHILDL16 MARRIED ETHBLACK ETHHISP)
```

In Stata the regression command is '`heckman`'.

SAMPLE SELECTION BIAS

```
. heckman LGEARN S ASVABC ETHBLACK ETHHISP if MALE==0, select(S AGE CHILDL06
CHILDL16 MARRIED ETHBLACK ETHHISP)
```

It is followed by the dependent variable and the explanatory variables in the regression. Note that we are restricting the sample to females.

SAMPLE SELECTION BIAS

```
. heckman LGEARN S ASVABC ETHBLACK ETHHISP if MALE==0, select(S AGE CHILDL06
CHILDL16 MARRIED ETHBLACK ETHHISP)
```

After a comma, the selection process is specified using '`select`' followed by the selection variables in parentheses.

SAMPLE SELECTION BIAS

```
. heckman LGEARN S ASVABC ETHBLACK ETHHISP if MALE==0, select(S AGE CHILDL06
CHILDL16 MARRIED ETHBLACK ETHHISP)
```

CHILDL06 **is a dummy equal to 1 if there is a child aged less than 6.** ***CHILDL16*** **is 1 if there is a child aged less than 15 and** ***CHILDL06*** **is 0.** ***MARRIED*** **is equal to 1 if the respondent is married with spouse present. Otherwise the variables are as defined in the** ***EAEF*** **data sets.**

SAMPLE SELECTION BIAS

```
. heckman LGEARN S ASVABC ETHBLACK ETHHISP if MALE==0, select(S AGE CHILDL06
CHILDL16 MARRIED ETHBLACK ETHHISP)

Iteration 0:   log likelihood = -2683.5848  (not concave)
Iteration 1:   log likelihood = -2681.9013  (not concave)
Iteration 2:   log likelihood = -2679.8394  (not concave)
Iteration 3:   log likelihood =  -2677.646  (not concave)
Iteration 4:   log likelihood =   -2674.93
Iteration 5:   log likelihood = -2670.7334
Iteration 6:   log likelihood = -2668.8143
Iteration 7:   log likelihood = -2668.8105
Iteration 8:   log likelihood = -2668.8105
```

Since the model involves probit analysis, it is fitted using maximum likelihood estimation.

SAMPLE SELECTION BIAS

```
Heckman selection model                          Number of obs      =      2661
(regression model with sample selection)         Censored obs       =       640
                                                 Uncensored obs     =      2021
                                                 Wald chi2(4)       =    714.73
Log likelihood =  -2668.81                       Prob > chi2        =    0.0000
------------------------------------------------------------------------------
             |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
LGEARN       |
           S |    .095949   .0056438    17.001   0.000        .0848874    .1070106
      ASVABC |   .0110391   .0014658     7.531   0.000        .0081663    .0139119
    ETHBLACK |   -.066425   .0381626    -1.741   0.082       -.1412223    .0083722
     ETHHISP |   .0744607   .0450095     1.654   0.098       -.0137563    .1626777
       _cons |   4.901626   .0768254    63.802   0.000        4.751051    5.052202
-------------+----------------------------------------------------------------
select       |
           S |   .1041415   .0119836     8.690   0.000        .0806541    .1276288
         AGE |  -.0357225    .011105    -3.217   0.001       -.0574879   -.0139572
    CHILDL06 |  -.3982738   .0703418    -5.662   0.000       -.5361412   -.2604064
    CHILDL16 |   .0254818   .0709693     0.359   0.720       -.1136155     .164579
     MARRIED |   .0121171   .0546561     0.222   0.825       -.0950069    .1192412
    ETHBLACK |  -.2941378   .0787339    -3.736   0.000       -.4484535   -.1398222
     ETHHISP |  -.0178776   .1034237    -0.173   0.863       -.2205843    .1848292
       _cons |   .1682515   .2606523     0.646   0.519       -.3426176    .6791206
------------------------------------------------------------------------------
```

The numbers of participating and non-participating respondents are given at the top of the output.

SAMPLE SELECTION BIAS

```
Heckman selection model                         Number of obs      =      2661
(regression model with sample selection)        Censored obs       =       640
                                                Uncensored obs     =      2021
                                                Wald chi2(4)       =    714.73
Log likelihood =  -2668.81                      Prob > chi2        =    0.0000
------------------------------------------------------------------------------
             |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
LGEARN       |
           S |    .095949   .0056438     17.001   0.000        .0848874    .1070106
      ASVABC |   .0110391   .0014658      7.531   0.000        .0081663    .0139119
    ETHBLACK |   -.066425   .0381626     -1.741   0.082       -.1412223    .0083722
     ETHHISP |   .0744607   .0450095      1.654   0.098       -.0137563    .1626777
       _cons |   4.901626   .0768254     63.802   0.000        4.751051    5.052202
-------------+----------------------------------------------------------------
select       |
           S |   .1041415   .0119836      8.690   0.000        .0806541    .1276288
         AGE |  -.0357225    .011105     -3.217   0.001       -.0574879   -.0139572
    CHILDL06 |  -.3982738   .0703418     -5.662   0.000       -.5361412   -.2604064
    CHILDL16 |   .0254818   .0709693      0.359   0.720       -.1136155     .164579
     MARRIED |   .0121171   .0546561      0.222   0.825       -.0950069    .1192412
    ETHBLACK |  -.2941378   .0787339     -3.736   0.000       -.4484535   -.1398222
     ETHHISP |  -.0178776   .1034237     -0.173   0.863       -.2205843    .1848292
       _cons |   .1682515   .2606523      0.646   0.519       -.3426176    .6791206
-------------+----------------------------------------------------------------
```

Next comes the regression output.

SAMPLE SELECTION BIAS

```
Heckman selection model                         Number of obs      =      2661
(regression model with sample selection)        Censored obs       =       640
                                                Uncensored obs     =      2021
                                                Wald chi2(4)       =    714.73
Log likelihood =  -2668.81                      Prob > chi2        =    0.0000
------------------------------------------------------------------------------
             |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
LGEARN       |
           S |    .095949   .0056438    17.001   0.000        .0848874    .1070106
      ASVABC |   .0110391   .0014658     7.531   0.000        .0081663    .0139119
    ETHBLACK |   -.066425   .0381626    -1.741   0.082       -.1412223    .0083722
     ETHHISP |   .0744607   .0450095     1.654   0.098       -.0137563    .1626777
       _cons |   4.901626   .0768254    63.802   0.000        4.751051    5.052202
-------------+----------------------------------------------------------------
select       |
           S |   .1041415   .0119836     8.690   0.000        .0806541    .1276288
         AGE |  -.0357225    .011105    -3.217   0.001       -.0574879   -.0139572
    CHILDL06 |  -.3982738   .0703418    -5.662   0.000       -.5361412   -.2604064
    CHILDL16 |   .0254818   .0709693     0.359   0.720       -.1136155     .164579
     MARRIED |   .0121171   .0546561     0.222   0.825       -.0950069    .1192412
    ETHBLACK |  -.2941378   .0787339    -3.736   0.000       -.4484535   -.1398222
     ETHHISP |  -.0178776   .1034237    -0.173   0.863       -.2205843    .1848292
       _cons |   .1682515   .2606523     0.646   0.519       -.3426176    .6791206
-------------+----------------------------------------------------------------
```

The results of the probit analysis of the selection process follow.

SAMPLE SELECTION BIAS

```
Heckman selection model                         Number of obs      =      2661
          |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
----------+------------------------------------------------------------------
select    |
        S |   .1041415   .0119836      8.690   0.000        .0806541    .1276288
      AGE |  -.0357225    .011105     -3.217   0.001       -.0574879   -.0139572
 CHILDL06 |  -.3982738   .0703418     -5.662   0.000       -.5361412   -.2604064
 CHILDL16 |   .0254818   .0709693      0.359   0.720       -.1136155     .164579
  MARRIED |   .0121171   .0546561      0.222   0.825       -.0950069    .1192412
 ETHBLACK |  -.2941378   .0787339     -3.736   0.000       -.4484535   -.1398222
  ETHHISP |  -.0178776   .1034237     -0.173   0.863       -.2205843    .1848292
    _cons |   .1682515   .2606523      0.646   0.519       -.3426176    .6791206
----------+------------------------------------------------------------------
  /athrho |    1.01804   .0932533     10.917   0.000        .8352669    1.200813
 /lnsigma |  -.6349788   .0247858    -25.619   0.000       -.6835582   -.5863994
----------+------------------------------------------------------------------
      rho |    .769067   .0380973                           .683294    .8339024
    sigma |   .5299467   .0131352                          .5048176    .5563268
   lambda |   .4075645     .02867                          .3513724    .4637567
------------------------------------------------------------------------------
LR test of indep. eqns. (rho = 0):   chi2(1) =    32.90   Prob > chi2 = 0.0000
------------------------------------------------------------------------------
```

The final part of the output gives the selection bias statistics. rho gives an estimate of the correlation between ε and *u*, here 0.77.

SAMPLE SELECTION BIAS

Heckman selection model — Number of obs = 2661

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
select						
S	.1041415	.0119836	8.690	0.000	.0806541	.1276288
AGE	-.0357225	.011105	-3.217	0.001	-.0574879	-.0139572
CHILDL06	-.3982738	.0703418	-5.662	0.000	-.5361412	-.2604064
CHILDL16	.0254818	.0709693	0.359	0.720	-.1136155	.164579
MARRIED	.0121171	.0546561	0.222	0.825	-.0950069	.1192412
ETHBLACK	-.2941378	.0787339	-3.736	0.000	-.4484535	-.1398222
ETHHISP	-.0178776	.1034237	-0.173	0.863	-.2205843	.1848292
_cons	.1682515	.2606523	0.646	0.519	-.3426176	.6791206
/athrho	1.01804	.0932533	10.917	0.000	.8352669	1.200813
/lnsigma	-.6349788	.0247858	-25.619	0.000	-.6835582	-.5863994
rho	.769067	.0380973			.683294	.8339024
sigma	.5299467	.0131352			.5048176	.5563268
lambda	.4075645	.02867			.3513724	.4637567

LR test of indep. eqns. (rho = 0): chi2(1) = 32.90 Prob > chi2 = 0.0000

For technical reasons, ρ is estimated indirectly via atanh ρ. However, a test of the null hypothesis H_0: atanh $\rho = 0$ is equivalent to a test of the null hypothesis of H_0: $\rho = 0$.

SAMPLE SELECTION BIAS

```
Heckman selection model                         Number of obs      =      2661
          |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
----------+------------------------------------------------------------------
select    |
        S |   .1041415   .0119836     8.690   0.000      .0806541    .1276288
      AGE |  -.0357225    .011105    -3.217   0.001     -.0574879   -.0139572
 CHILDL06 |  -.3982738   .0703418    -5.662   0.000     -.5361412   -.2604064
 CHILDL16 |   .0254818   .0709693     0.359   0.720     -.1136155     .164579
  MARRIED |   .0121171   .0546561     0.222   0.825     -.0950069    .1192412
 ETHBLACK |  -.2941378   .0787339    -3.736   0.000     -.4484535   -.1398222
  ETHHISP |  -.0178776   .1034237    -0.173   0.863     -.2205843    .1848292
    _cons |   .1682515   .2606523     0.646   0.519     -.3426176    .6791206
----------+------------------------------------------------------------------
  /athrho |    1.01804   .0932533    10.917   0.000      .8352669    1.200813
/lnsigma  |  -.6349788   .0247858   -25.619   0.000     -.6835582   -.5863994
----------+------------------------------------------------------------------
      rho |    .769067   .0380973                       .683294    .8339024
    sigma |   .5299467   .0131352                      .5048176    .5563268
   lambda |   .4075645     .02867                      .3513724    .4637567
------------------------------------------------------------------------------
LR test of indep. eqns. (rho = 0):   chi2(1) =    32.90   Prob > chi2 = 0.0000
------------------------------------------------------------------------------
```

The asymptotic *t* statistic is 10.92 and so the null hypothesis is rejected.

SAMPLE SELECTION BIAS

```
Heckman selection model                         Number of obs      =      2661
(regression model with sample selection)        Censored obs       =       640
                                                Uncensored obs     =      2021
                                                Wald chi2(4)       =    714.73
Log likelihood =  -2668.81                      Prob > chi2        =    0.0000
------------------------------------------------------------------------------
             |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
LGEARN       |
           S |    .095949   .0056438    17.001   0.000        .0848874    .1070106
      ASVABC |   .0110391   .0014658     7.531   0.000        .0081663    .0139119
    ETHBLACK |   -.066425   .0381626    -1.741   0.082       -.1412223    .0083722
     ETHHISP |   .0744607   .0450095     1.654   0.098       -.0137563    .1626777
       _cons |   4.901626   .0768254    63.802   0.000        4.751051    5.052202
-------------+----------------------------------------------------------------
select       |
           S |   .1041415   .0119836     8.690   0.000        .0806541    .1276288
         AGE |  -.0357225    .011105    -3.217   0.001       -.0574879   -.0139572
    CHILDL06 |  -.3982738   .0703418    -5.662   0.000       -.5361412   -.2604064
    CHILDL16 |   .0254818   .0709693     0.359   0.720       -.1136155     .164579
     MARRIED |   .0121171   .0546561     0.222   0.825       -.0950069    .1192412
    ETHBLACK |  -.2941378   .0787339    -3.736   0.000       -.4484535   -.1398222
     ETHHISP |  -.0178776   .1034237    -0.173   0.863       -.2205843    .1848292
       _cons |   .1682515   .2606523     0.646   0.519       -.3426176    .6791206
-------------+----------------------------------------------------------------
```

An alternative test involves a comparison of the log likelihood for this model with that for a restricted version where ρ is assumed to be 0.

SAMPLE SELECTION BIAS

Heckman selection model — Number of obs = 2661

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
select						
S	.1041415	.0119836	8.690	0.000	.0806541	.1276288
AGE	-.0357225	.011105	-3.217	0.001	-.0574879	-.0139572
CHILDL06	-.3982738	.0703418	-5.662	0.000	-.5361412	-.2604064
CHILDL16	.0254818	.0709693	0.359	0.720	-.1136155	.164579
MARRIED	.0121171	.0546561	0.222	0.825	-.0950069	.1192412
ETHBLACK	-.2941378	.0787339	-3.736	0.000	-.4484535	-.1398222
ETHHISP	-.0178776	.1034237	-0.173	0.863	-.2205843	.1848292
_cons	.1682515	.2606523	0.646	0.519	-.3426176	.6791206
/athrho	1.01804	.0932533	10.917	0.000	.8352669	1.200813
/lnsigma	-.6349788	.0247858	-25.619	0.000	-.6835582	-.5863994
rho	.769067	.0380973			.683294	.8339024
sigma	.5299467	.0131352			.5048176	.5563268
lambda	.4075645	.02867			.3513724	.4637567

LR test of indep. eqns. (rho = 0): chi2(1) = 32.90 Prob > chi2 = 0.0000

The test statistic 2 (log L_U – log L_R), where log L_U and log L_R are the log-likelihoods for the unrestricted and restricted versions, is distributed as a chi-squared statistic with 1 degree of freedom under the null hypothesis that the restriction $\rho = 0$ is valid.

SAMPLE SELECTION BIAS

```
Heckman selection model                         Number of obs      =      2661
          |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
----------+------------------------------------------------------------------
select    |
        S |   .1041415   .0119836     8.690   0.000      .0806541    .1276288
      AGE |  -.0357225    .011105    -3.217   0.001     -.0574879   -.0139572
 CHILDL06 |  -.3982738   .0703418    -5.662   0.000     -.5361412   -.2604064
 CHILDL16 |   .0254818   .0709693     0.359   0.720     -.1136155    .164579
  MARRIED |   .0121171   .0546561     0.222   0.825     -.0950069    .1192412
 ETHBLACK |  -.2941378   .0787339    -3.736   0.000     -.4484535   -.1398222
  ETHHISP |  -.0178776   .1034237    -0.173   0.863     -.2205843    .1848292
    _cons |   .1682515   .2606523     0.646   0.519     -.3426176    .6791206
----------+------------------------------------------------------------------
  /athrho |    1.01804   .0932533    10.917   0.000      .8352669    1.200813
 /lnsigma |  -.6349788   .0247858   -25.619   0.000     -.6835582   -.5863994
----------+------------------------------------------------------------------
      rho |    .769067   .0380973                       .683294    .8339024
    sigma |   .5299467   .0131352                      .5048176    .5563268
   lambda |   .4075645     .02867                      .3513724    .4637567
------------------------------------------------------------------------------
LR test of indep. eqns. (rho = 0):   chi2(1) =    32.90   Prob > chi2 = 0.0000
------------------------------------------------------------------------------
```

In this example the test statistic is 32.90. The critical value of chi-squared with one degree of freedom at the 0.1 percent level is 10.83, so the null hypothesis is rejected.

SAMPLE SELECTION BIAS

```
. heckman LGEARN S ASVABC ETHBLACK ETHHISP if MALE==0, select(S AGE CHILDL06
CHILDL16 MARRIED ETHBLACK ETHHISP)
------------------------------------------------------------------------------
             |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
LGEARN       |
           S |    .095949   .0056438    17.001   0.000        .0848874    .1070106
      ASVABC |   .0110391   .0014658     7.531   0.000        .0081663    .0139119
  ETHBLACK   |   -.066425   .0381626    -1.741   0.082       -.1412223    .0083722
   ETHHISP   |   .0744607   .0450095     1.654   0.098       -.0137563    .1626777
       _cons |   4.901626   .0768254    63.802   0.000        4.751051    5.052202
-------------+----------------------------------------------------------------

. reg LGEARN S ASVABC ETHBLACK ETHHISP if MALE==0
------------------------------------------------------------------------------
      LGEARN |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |   .0807836    .005244    15.405   0.000        .0704994    .0910677
      ASVABC |   .0117377   .0014886     7.885   0.000        .0088184     .014657
  ETHBLACK   |  -.0148782   .0356868    -0.417   0.677       -.0848649    .0551086
   ETHHISP   |   .0802266    .041333     1.941   0.052       -.0008333    .1612865
       _cons |   5.223712   .0703534    74.250   0.000        5.085739    5.361685
------------------------------------------------------------------------------
```

It is instructive to compare the fitted earnings functions for the heckman and least squares models. The coefficients are fairly similar, despite the inconsistency of the least squares estimates.

SAMPLE SELECTION BIAS

```
. heckman LGEARN S ASVABC ETHBLACK ETHHISP if MALE==0, select(S AGE CHILDL06
CHILDL16 MARRIED ETHBLACK ETHHISP)
------------------------------------------------------------------------------
             |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
LGEARN       |
           S |    .095949   .0056438    17.001   0.000     .0848874    .1070106
      ASVABC |   .0110391   .0014658     7.531   0.000     .0081663    .0139119
   ETHBLACK |   -.066425   .0381626    -1.741   0.082    -.1412223    .0083722
    ETHHISP |   .0744607   .0450095     1.654   0.098    -.0137563    .1626777
       _cons |   4.901626   .0768254    63.802   0.000     4.751051    5.052202
-------------+----------------------------------------------------------------

. reg LGEARN S ASVABC ETHBLACK ETHHISP if MALE==0
------------------------------------------------------------------------------
      LGEARN |      Coef.   Std. Err.       t     P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |   .0807836    .005244    15.405   0.000     .0704994    .0910677
      ASVABC |   .0117377   .0014886     7.885   0.000     .0088184     .014657
   ETHBLACK |  -.0148782   .0356868    -0.417   0.677    -.0848649    .0551086
    ETHHISP |   .0802266    .041333     1.941   0.052    -.0008333    .1612865
       _cons |   5.223712   .0703534    74.250   0.000     5.085739    5.361685
------------------------------------------------------------------------------
```

The coefficient of schooling is a little higher in the heckman regression.

SAMPLE SELECTION BIAS

```
Heckman selection model                         Number of obs      =      2661
(regression model with sample selection)        Censored obs       =       640
                                                Uncensored obs     =      2021
                                                Wald chi2(4)       =    714.73
Log likelihood =  -2668.81                      Prob > chi2        =    0.0000
------------------------------------------------------------------------------
             |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
LGEARN       |
           S |    .095949   .0056438    17.001   0.000        .0848874    .1070106
      ASVABC |   .0110391   .0014658     7.531   0.000        .0081663    .0139119
    ETHBLACK |   -.066425   .0381626    -1.741   0.082       -.1412223    .0083722
     ETHHISP |   .0744607   .0450095     1.654   0.098       -.0137563    .1626777
       _cons |   4.901626   .0768254    63.802   0.000        4.751051    5.052202
-------------+----------------------------------------------------------------
select       |
           S |   .1041415   .0119836     8.690   0.000        .0806541    .1276288
         AGE |  -.0357225    .011105    -3.217   0.001       -.0574879   -.0139572
    CHILDL06 |  -.3982738   .0703418    -5.662   0.000       -.5361412   -.2604064
    CHILDL16 |   .0254818   .0709693     0.359   0.720       -.1136155     .164579
     MARRIED |   .0121171   .0546561     0.222   0.825       -.0950069    .1192412
    ETHBLACK |  -.2941378   .0787339    -3.736   0.000       -.4484535   -.1398222
     ETHHISP |  -.0178776   .1034237    -0.173   0.863       -.2205843    .1848292
       _cons |   .1682515   .2606523     0.646   0.519       -.3426176    .6791206
-------------+----------------------------------------------------------------
```

The probit analysis showed that schooling has a highly significant positive effect on labor force participation, controlling for other characteristics such as number of children of school age.

SAMPLE SELECTION BIAS

```
. heckman LGEARN S ASVABC ETHBLACK ETHHISP if MALE==0, select(S AGE CHILDL06
CHILDL16 MARRIED ETHBLACK ETHHISP)
------------------------------------------------------------------------------
             |      Coef.   Std. Err.       z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
LGEARN       |
           S |    .095949   .0056438    17.001   0.000     .0848874    .1070106
      ASVABC |   .0110391   .0014658     7.531   0.000     .0081663    .0139119
  ETHBLACK   |   -.066425   .0381626    -1.741   0.082    -.1412223    .0083722
   ETHHISP   |   .0744607   .0450095     1.654   0.098    -.0137563    .1626777
       _cons |   4.901626   .0768254    63.802   0.000     4.751051    5.052202
-------------+----------------------------------------------------------------
```

```
. reg LGEARN S ASVABC ETHBLACK ETHHISP if MALE==0
------------------------------------------------------------------------------
      LGEARN |      Coef.   Std. Err.       t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |   .0807836    .005244    15.405   0.000     .0704994    .0910677
      ASVABC |   .0117377   .0014886     7.885   0.000     .0088184     .014657
  ETHBLACK   |  -.0148782   .0356868    -0.417   0.677    -.0848649    .0551086
   ETHHISP   |   .0802266    .041333     1.941   0.052    -.0008333    .1612865
       _cons |   5.223712   .0703534    74.250   0.000     5.085739    5.361685
------------------------------------------------------------------------------
```

If females with higher levels of schooling are relatively keen to work, they will tend to be willing to accept lower wages, controlling for other factors including education, than those who are reluctant to work.

SAMPLE SELECTION BIAS

```
. heckman LGEARN S ASVABC ETHBLACK ETHHISP if MALE==0, select(S AGE CHILDL06
CHILDL16 MARRIED ETHBLACK ETHHISP)
------------------------------------------------------------------------------
             |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
LGEARN       |
           S |    .095949   .0056438    17.001   0.000     .0848874    .1070106
      ASVABC |   .0110391   .0014658     7.531   0.000     .0081663    .0139119
  ETHBLACK   |   -.066425   .0381626    -1.741   0.082    -.1412223    .0083722
   ETHHISP   |   .0744607   .0450095     1.654   0.098    -.0137563    .1626777
       _cons |   4.901626   .0768254    63.802   0.000     4.751051    5.052202
-------------+----------------------------------------------------------------

. reg LGEARN S ASVABC ETHBLACK ETHHISP if MALE==0
------------------------------------------------------------------------------
      LGEARN |      Coef.   Std. Err.       t     P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |   .0807836    .005244    15.405   0.000     .0704994    .0910677
      ASVABC |   .0117377   .0014886     7.885   0.000     .0088184     .014657
  ETHBLACK   |  -.0148782   .0356868    -0.417   0.677    -.0848649    .0551086
   ETHHISP   |   .0802266    .041333     1.941   0.052    -.0008333    .1612865
       _cons |   5.223712   .0703534    74.250   0.000     5.085739    5.361685
------------------------------------------------------------------------------
```

Hence the wages of more-educated females will tend not to reflect the full value of education in the market place. The least squares regression does not take account of this, and hence the estimate of the return to schooling is lower.