# Estimation of Heckman's Selection Model using gretl

## Quantitative Microeconomics

R. Mora

Department of Economics
Universidad Carlos III de Madrid

# Outline

# Introduction

# Heckman's Selection Model

**we observe $w_i$ if $s_i = 1$**

- output equation: $w = \beta_0 + \beta x + \varepsilon$
- participation equation: $s = 1(\gamma' z + v)$
- $\begin{bmatrix} u \\ v \end{bmatrix} \sim N\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 & \rho \\ \rho & 1 \end{bmatrix} \right)$

# Estimation

- OLS is inconsistent.
- ML estimation is consistent: the actual expression for the likelihood is more complicated than that of the probit and tobit model as it requires obtaining the joint distribution of $w$ and $s$
  - In general, the likelihood function is not globally concave, and can have local maxima
- Heckman's two-stage procedure based on the conditional expectation gives consistent estimates and it is easy to implement.
  - It can be used to obtain initial conditions for MLE.
  - Usual standard errors from the second stage are not valid.

# Heckman and gretl

# Basic commands and functions for Heckit Estimation

- `heckit`: computes Heckman's selection model
- `restrict`: tests hypothesis for parameters on both equations

# heckit *output x_vars ; selection z_vars* ——two-step

- *output* represents the dependent variable in the output equation
- *x_vars* represents the list of controls in the output equation
- *selection* represents the dependent variable in the participation equation
    - *selection* must be a binary $\{0,1\}$ variable
- *z_vars* represents the list of controls in the participation equation
- ——two-step: conducts two-step Heckman's procedure, reporting correct standard errors (ML is default)

# A Simple Example

## Participation

- $U_m - U_h = -0.5 + 0.03 * educ - 1.5 * kids + v$

## Wage equation

- $wage = 5 + 0.07 educ + u$

- $cov(educ, u) = 0$
- $\begin{bmatrix} u \\ v \end{bmatrix} \sim N\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix} \right)$

# A Simple Example: ols estimation ($\beta_{educ} = .07, \rho = .9$)

## ols *wage const educ*

```
Model 1: OLS, using observations 1-5000 (n = 1112)
Missing or incomplete observations dropped: 3888
Dependent variable: wage
Heteroskedasticity-robust standard errors, variant HC1

             coefficient   std. error   t-ratio   p-value
  -------------------------------------------------------
  const       6.12441      0.0979021    62.56     0.0000   ***
  educ        0.0561435    0.00689433    8.143    1.03e-15 ***

Mean dependent var    6.904610   S.D. dependent var   0.826190
Sum squared resid     713.5680   S.E. of regression   0.801782
R-squared             0.059060   Adjusted R-squared   0.058212
F(1, 1110)            66.31550   P-value(F)           1.03e-15
Log-likelihood       -1331.197   Akaike criterion     2666.394
Schwarz criterion     2676.422   Hannan-Quinn         2670.186
```

# OLS bias

- In the example, we have the following:
  - The true returns to education are approximately 7% ($\beta_{educ} = .07$).
  - The score for participation also depends on education ($\gamma_{educ} = .03$).
  - Importantly, unobservable (for the econometrician) determinants on wages and unobservable determinants of participation are positively correlated ($\rho = .9$).
- This positive correlation implies that participants in the labor market with lower levels of education tend to have positive errors in wage equation.
- OLS under-estimates the returns to education:
  - 95% confidence interval:(4.26, 6.97)

# Setting $wage = 0$ for missing wages ($\beta_{educ} = .07, \rho = .9$)

Model 5: OLS, using observations 1–5000
Dependent variable: wage2
Heteroskedasticity-robust standard errors, variant HC1

| | Coefficient | Std. Error | $t$-ratio | p-value |
|---|---|---|---|---|
| const | 0.819650 | 0.157393 | 5.2077 | 0.0000 |
| educ | 0.0567067 | 0.0117025 | 4.8457 | 0.0000 |

| | | | |
|---|---|---|---|
| Mean dependent var | 1.581584 | S.D. dependent var | 2.925883 |
| Sum squared resid | 42586.28 | S.E. of regression | 2.919018 |
| $R^2$ | 0.004886 | Adjusted $R^2$ | 0.004687 |
| $F(1, 4998)$ | 23.48076 | P-value($F$) | 1.30e–06 |
| Log-likelihood | −12449.93 | Akaike criterion | 24903.86 |
| Schwarz criterion | 24916.89 | Hannan–Quinn | 24908.42 |

# OLS bias using the full sample

- The bias is not corrected when setting $wage = 0$ for those who do not participate.
- When we replace the true wage by 0 when $s = 0$, then the model becomes :
  - $w = \begin{cases} \beta_0 + \beta x + \varepsilon \text{ if } s = 1 \\ 0 \text{ if } s = 0 \end{cases}$
  - $s = 1(\gamma' z + v)$
  - $\begin{bmatrix} u \\ v \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 & \rho \\ \rho & 1 \end{bmatrix}\right)$
- It can be proved that $\mathrm{E}(w|x) \neq \beta_0 + \beta_1 x + \varepsilon$.

# ML estimation $(\beta_{educ} = .07, \rho = .9)$

```
heckit wage const educ ; work const educ kids
```

```
Model 3: ML Heckit, using observations 1-5000
Dependent variable: wage
Selection variable: work

             coefficient   std. error   t-ratio    p-value
  --------------------------------------------------------
  const       4.97764      0.110538     45.03     0.0000    ***
  educ        0.0700091    0.00705706    9.920    3.39e-23  ***
  lambda      0.933032     0.0374084    24.94     2.62e-137 ***

                      Selection equation

  const      -0.473329     0.0877324    -5.395    6.85e-08  ***
  educ        0.0257985    0.00617236    4.180    2.92e-05  ***
  kids       -1.46115      0.0438994   -33.28     6.57e-243 ***

Mean dependent var   6.923428   S.D. dependent var   0.815325
sigma                1.036613   rho                  0.900076
Log-likelihood      -3142.936   Akaike criterion     6291.873
Schwarz criterion    6306.835   Hannan-Quinn         6297.538

Total observations: 5000
Censored observations: 3917 (78.3%)
```

# Two-stage estimation ($\beta_{educ} = .07, \rho = .9$)

**heckit *wage const educ ; work const educ kids* --two-step**

```
Model 2: Two-step Heckit, using observations 1-5000
Dependent variable: wage
Selection variable: work

             coefficient   std. error   t-ratio    p-value
  -------------------------------------------------------
  const        5.05706      0.121918     41.48    0.0000    ***
  educ         0.0690590    0.00713587    9.678   3.75e-22  ***
  lambda       0.874918     0.0541856    16.15    1.20e-58  ***

                       Selection equation

  const       -0.473118     0.0894129    -5.291   1.21e-07  ***
  educ         0.0261450    0.00629308    4.155   3.26e-05  ***
  kids        -1.48345      0.0470196   -31.55    1.82e-218 ***

Mean dependent var    6.923428   S.D. dependent var    0.815325
sigma                 1.009690   rho                   0.866521

Total observations: 5000
Censored observations: 3917 (78.3%)
```

- Both the ML and the two-step procedure give consistent estimates.
- The ML estimator is a bit more precise. This is true for large samples.
- Among the results, we also get estimates for the correlation of the errors.
  - Recall that if $\rho = 0$, then there is no sample selection bias.
- Inference for the significance of the parameters in the output and participation equations can be carried out as usual.
- Prediction cannot be implemented using `fcast`

# restrict

- The restrict command allows the simultaneous test of several restrictions.
  - The test is performed on the estimates of the last model estimated before the command is invoked.
  - After heckit, the numbering of the parameters follows the order of the display of the output.
  - The –quiet option hides the output of the restricted model estimation.

- To implement it, we create a block:

  restrict

  *here we insert as many lines as restrictions to be tested*

  end restrict

# Example of testing

```
? restrict --quiet
? b[lambda]=0
? end restrict
Restriction:
 b[lambda] = 0

Test statistic: chi^2(1) = 732.782, with p-value = 2.22441e-161

? restrict --quiet
? b[5]=0.03
? b[6]=-1.5
? end restrict
Restriction set
 1: b[educ] = 0.03
 2: b[kids] = -1.5

Test statistic: chi^2(2) = 0.635692, with p-value = 0.727715
```

# Testing the significance of $\lambda$

- We can test the significance of the parameter associated to $\lambda$ in the conditional expectation of the output equation.
- This test is a test of random sample selection.
- If the parameter is not significant, then we do not reject the null of random selection (and OLS is consistent).
- In the previous example, the null hypothesis is strongly rejected: we find evidence of nonrandom sample selection.

# Marginal Effects

```
# marginal effects of another year of education
genr coeff=$coeff
genr beta=coeff[1:2]
genr gamma=coeff[4:6]
series educ0=educ
matrix x0={const,educ0}
series educ1=educ+1
matrix x1={const,educ1}
series x1b = x1*beta
series x0b = x0*beta
genr Mg_educ = mean(x1b-x0b)
```

```
Generated scalar Mg_educ = 0.0700091
```

# Effects of the observed wages

- The previous marginal effect is on the unconditional expectation of wages.
- This is the relevant notion if what we want is the effect on the wage offers.
- If we want to learn the effect on average observed wages, we need to restrict the sample to those observed.
- The best way to do this is by using the analytical expression of the conditional expectation:

$$E[w|x,z,s=1] = x\beta + \rho\lambda(z\gamma)$$

## Summary

- gretl allows for estimation of Heckman's Selection Model.
- Both two-stage and ML estimation.
- Testing and prediction is computed as usual.