Sample Selection and Heckman Models

Applied Econometrics for Researchers, PhD
Vera Rocha, CBS-SI, vr.si@cbs.dk
9th November 2022

Current state of affairs…

- Linear regression
- Non-linear (binary) models
- Sample selection

- Matching models
- Instrumental Variables
- Panel Data
- Policy evaluation

Agenda for today

1. Forms of sample selection
2. Implications of sample selection bias
3. Econometric models addressing sample selection (Heckman 2-step model)
4. Sample selection in research (example)
5. Heckman models in Stata

Key readings:

– Certo, S. T., Busenbark, J. R., Woo, Y., Semadeni, M. (2016), ”Sample selection bias and Heckman models in strategic management research”, *Strategic Management Journal*, 37, 2639-57.

– **Cameron & Trivedi**, Section **16.5** & **Wooldridge**, Sections **17.1-17.4.1**.

– Naldi, L. Davidsson, P. (2014), ”Entrepreneurial growth: The role of international knowledge acquisition as moderated by firm age”, *Journal of Business Venturing*, 29, 687-703. **(optional; only an applied example)**

Assumptions of a Linear Regression Model

1. The regression model is linear in β and additive in u

2. The observations have been obtained as a random sample

3. No x variable is a linear function of (one or more) of the other x's (i.e., no exact multicollinearity)

4. u has a zero population mean and all x's are uncorrelated with u (zero correlation)

5. u has a constant variance (no heteroskedasticity)

6. u is normally distributed

OLS:
1-4: consistency
1-5: efficiency
6. Not needed in ”large samples”

Assumptions of a Linear Regression Model

1. The regression model is linear in β and additive in u

2. The observations have been obtained as a random sample

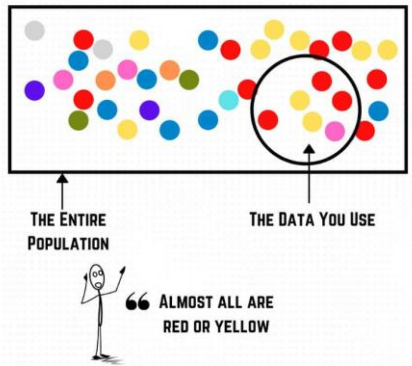
3. No x variable is a linear function of (one or more) of the other x's (i.e., no exact multicollinearity)

4. u has a zero population mean and all x's are uncorrelated with u (zero correlation)

5. u has a constant variance (no heteroskedasticity)

6. u is normally distributed

Today's focus! What if #2 fails?



Forms of Sample Selection (1/2)

I. Missing data

- Units could simply be ”missing at random”: we randomly *miss observations on one or more of the variables* that enter our model (testable!)
- *Reduced sample, but Assumption 2 is not violated.*

II. Deliberate selective sampling: e.g., stratified sampling

- Divide population into strata: exhaustive & non-overlapping groups
- Possible risk of over/undersampling of certain groups
- *Still not problematic for Assumption 2*

Forms of Sample Selection (2/2)

III. Non-random sampling

➢ **A) Exogenous sample selection:**

- Sample selection based on the **independent variables**; or, more generally, on factors that are **independent** of the error term of the equation we are considering

➢ **B) Endogenous sample selection:**

- Sample selection based on the **dependent variable**; or, more generally, on factors that are **related** to the error term of the equation we are considering

Sample selection is a problem if it causes the error term to be correlated with an explanatory variable!

Examples of Non-Random Sampling (1/3)

A) <u>Exogenous</u> sample selection (sampling based on <u>independent</u> variables)

E.g.: Estimating a saving function

- $saving = \beta_0 + \beta_1 income + \beta_2 age + u$.
- A survey of adults with age > 45 → Nonrandom sample based on ***age***
- But ***age*** is assumed to be exogenous (not correlated with the error term)
- The factor that determines the selection into the sample is <u>independent</u> of the error term
- **OLS estimator is unbiased**

Examples of Non-Random Sampling (2/3)

B1) <u>Endogenous</u> sample selection (sampling based on <u>dependent</u> variables)

E.g.: Determinants of income when income is top-coded at k

- $y = \beta_0 + \beta_1 x + u$.
- Actual y can only be observed if $y \leq k$, where k is the top-coded value
- Often done to increase reporting rates
- Selection based on y causes a **selection bias** on $\widehat{\beta_1}$ **if we use OLS**
- Use **Tobit (censored regression)* model instead**, <u>whenever k (cut-off point) is known/observed</u>

** Out of our scope in this course, but you can get a flavor of it [here]*

Examples of Non-Random Sampling (3/3)

B2) <u>Endogenous</u> sample selection (sampling based on <u>dependent</u> variables)

E.g.: Effect of human capital (education & experience) on wages

- $y = \beta_0 + \beta_1 educ + \beta_2 exp + \cdots + u$
- y is **only observed for individuals who are actually working**; not all do!
- Potential **non-random (i.e., *selected*) sample** ⇔ decision to work is likely related to **unobserved factors** that also influence wages (e.g., ability; reservation wages; education quality; network)
- Error term becomes correlated with one or several x's.

Incidental truncation problem: today's focus!

Non-Random Sampling: More Formally

Participation equation (e.g., labor market participation)

$$s = \begin{cases} 1 & if & s^* > 0 \\ 0 & if & s^* \leq 0 \end{cases}$$

where $s^* = X_1'\beta_1 + \varepsilon_1$

Likely correlated! Unobserved factors $\boldsymbol{W}$ may affect both y and s

Outcome equation (e.g., wages)

$$y = \begin{cases} y^* & if & s^* > 0 \\ - & if & s^* \leq 0 \end{cases}$$

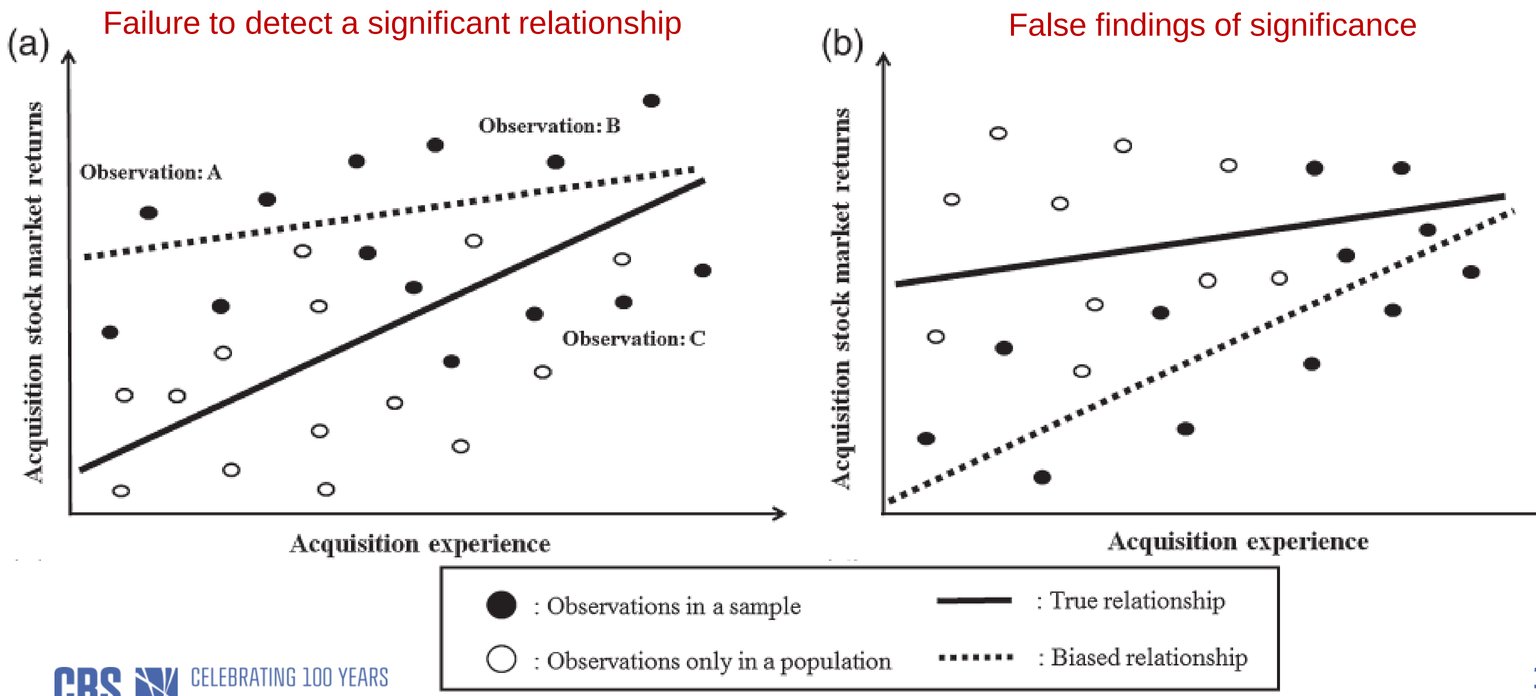
where $y^* = X_2'\beta_2 + \varepsilon_2$

ε_2 correlated with some variable(s) in X_2 (through the unobserved factors W)
[when run on the observed subsample]

The cause of selection bias

Implications of Sample Selection Bias

Example: The role of acquisition experience (x) in stock market reactions to acquisition announcements (y). y **only available for firms that actually complete acquisitions**.



Omitted variables create a positive/ negative correlation between the errors in a/b (e.g., unobserved integration capabilities/ CEO overcon-fidence).

Source: Certo et al. (2016), SMJ

Heckman Two-Step Estimation Method

The two-step model (recall slide 11)

- The outcome equation: $y = X_2'\beta_2 + \varepsilon_2, \quad \varepsilon_2 | X_2 \sim N(0, \sigma^2)$
- The selection equation: $s = 1\,[X_1'\beta_1 + \varepsilon_1 \geq 0]$, where $s = 1$ if we observe y, and zero otherwise; $\varepsilon_1 | X_1 \sim N(0,1)$

Assumptions

- Elements of X_1 and X_2 are always observed and are exogenous: $E(\varepsilon_1 | X_1, X_2) = 0$, $E(\varepsilon_2 | X_1, X_2) = 0$
- **$\boldsymbol{X_2}$ is a subset of $\boldsymbol{X_1}$**: any x_{2j} is an element of X_1, but some elements of X_1 are not in X_2 **(exclusion restrictions)**

Heckman Two-Step Model: Underlying Logic

The cause of bias

- $cov(\varepsilon_1, \varepsilon_2) \neq 0$ in the selected sample, so:
- $E(y|X_2, \varepsilon_2, s) = X_2'\beta_2 + \gamma_1 E(\varepsilon_2|\varepsilon_1)$ (instead of $X_2'\beta_2$ only)

The Heckman solution

- obtain a term to correct for the $E(\varepsilon_2|\varepsilon_1)$ term [the *first* step]
- use the term as an extra explanatory variable in the outcome equation [the *second* step]
- this (in principle!*) removes the part of the error that is correlated with X_2

How to estimate $\boldsymbol{E(\varepsilon_2|\varepsilon_1)}$?

- If ε_1 and ε_2 are jointly normal, $E(\varepsilon_2|\varepsilon_1) = \rho\varepsilon_1$, where ρ is the correlation between ε_1 and ε_2.
- We go from estimating $E(\varepsilon_2|\varepsilon_1)$ to estimating $E(\varepsilon_1|X_1, s = 1)$ → probit model in the first step

**dependent on strong exclusion restrictions*

Heckman Two-Step Model: Underlying Logic

…estimating $\boldsymbol{E(\varepsilon_1|X_1, s=1)}$ with a probit model in the first step:

When $s = 1$ $(selection\ equation)$, $E(\varepsilon_1|X_1, s = 1) = \boldsymbol{\phi(X_1'\beta_1)/\Phi(X_1'\beta_1)}$

$\downarrow$

The Inverse Mills Ratio, $\boldsymbol{\lambda(X_1'\beta_1)}$:
a ratio beween the std normal pdf and std normal cdf, evaluated at $\boldsymbol{X_1'\beta_1}$.

New Outcome Equation: $E(y|X_2, \boldsymbol{s = 1}) = X_2'\beta_2 + \gamma_1\lambda(X_1'\beta_1)$

- If $\rho = 0$, no selection bias (The correlation between ε_1 and ε_2 does not cause a sample selection problem);
- If $\rho \neq 0$, a consistent estimate for β_1 using the *selected* sample can only be obtained if we include the IMR – $\lambda(X_1'\beta_1)$ – as an additional regressor.

Heckman Model in Steps: Wrap-up

The first step

- Use the entire sample (all *N* observations) to estimate a probit model of s_i on X_{1i}
  $P(s=1|X_1) = \Phi(X_1'\beta_1)$
- Compute the Inverse Mills Ratio, $\widehat{\lambda_i} = \lambda(X_1'\widehat{\beta_1})$ for each i.

  (Actually, we only need these for the observations with $s_i = 1$)

The second step

- Using the selected sample ($s_i = 1$), run the regression $y = X_2'\beta_2 + \gamma_1\lambda\left(X_1'\widehat{\beta_1}\right) + \varepsilon_2$

to obtain $\widehat{\beta_2}$, which is consistent and approximately normally distributed.

A test of selection bias: use the t statistic on $\hat{\lambda}$ (i.e., test significance of γ_1) as a test of H_0: $\rho = 0$. (There is no sample selection problem)

Heckman 2-Step vs. ML Estimation

Note that: $\boldsymbol{X_2}$ must be a subset of $\boldsymbol{X_1}$

- If X_1 = X_2, $\hat{\lambda}$ can be highly correlated with X_2 → high std errors for $\widehat{\beta_2}$ and identification hinges on functional form only: Not credible.
- Excluding one or more X_2 variables from X_1 helps identifying the $\widehat{\beta_2}$ parameters (but needs a theoretical argument!) – **exclusion restrictions**
- Excluding *x*-variables from X_1 can lead to inconsistency (and would be hard to defend)

Heckman two-step estimation is a *second-best* alternative to ML

- Consistent but inefficient
- Introducing a measurement error problem

Sample Selection in Research: Example

Hypothesis 1. Acquisition of knowledge from international markets has a positive effect on a firm's subsequent entrepreneurial growth as reflected in an increase of

a) sales generated in geographically new domestic markets,
b) sales generated in geographically new international markets,
c) sales from new products in domestic markets; and
d) sales from new products in international markets.

Naldi & Davidsson (2014), JBV

- Sample of Swedish SMEs – initially 2,455 >>> 1,633 interviewed by phone, out of which **885 had international activity (eligible!)**
- **218/885 suspended operations** during the 6-y follow-up period
- Final sample of 138 firms after all follow-ups

Sample selection (bias)

Sample attrition

Based on the DV!

Using Heckman 2-step to address both issues

Sample selection bias

- Probit model for the likelihood of internationalization
- <u>Exclusion restriction:</u> whether or not the firm had a Swedish name

Sample attrition bias

- Probit model where the DV measures whether the firm participated in all survey rounds
- <u>Exclusion restriction:</u> firm's location in a metropolitan area

Respective IMRs obtained and used as a regressor in the 2nd step.

(How much do we trust these exclusion restrictions?)

Example

- **Selection bias** due to the fact that y is only observed for firms with international activity is **significant**, **but does not seem to bias the coefficient of interest** (acquisition of knowledge)
- **Positive coefficient for *Mills (selection bias)*** could suggest that unobserved factors that make firms more likely to internationalize also increase their growth through international sales. (though not a major concern here)

CBS CELEBRATING 100 YEARS
COPENHAGEN BUSINESS SCHOOL

	Sales from new international markets			
	Model 5a	Model 6a	Model 7a	Model 8a
CEO age	−0.11**	−0.11**	−0.11**	−0.11**
	(0.039)	(0.037)	(0.036)	(0.036)
CEO gender	3.18*	3.17*	3.02	2.10*
	(1.51)	(1.50)	(1.54)	(0.96)
CEO education	−0.26	−0.31	−0.33	−0.50
	(0.47)	(0.45)	(0.45)	(0.46)
CEO prior leadership experience	0.017	0.058	0.050	0.060
	(0.46)	(0.48)	(0.48)	(0.48)
CEO prior experience-same industry	−0.097	−0.091	−0.10	0.50
	(0.45)	(0.46)	(0.47)	(0.44)
CEO prior experience-other industries	0.22	0.19	0.17	0.0070
	(0.44)	(0.45)	(0.47)	(0.40)
Manufacturer	0.23	0.13	−0.034	−1.43*
	(0.51)	(0.54)	(0.85)	(0.73)
Service	0.70	0.54	0.52	−0.47
	(0.53)	(0.56)	(0.56)	(0.61)
Retailer	−12.7***	−12.8***	−11.6***	−12.2***
	(0.70)	(0.69)	(0.87)	(0.70)
Past performance	−0.27	−0.27	−0.23	−0.65*
	(0.42)	(0.41)	(0.39)	(0.29)
Firm size	−1.38	−1.37	−1.40	−0.96
	(0.74)	(0.75)	(0.72)	(0.86)
Acquisition of knowledge	**1.09*** **	**1.10*** **	**1.09*** **	**1.20*** **
	(0.29)	(0.26)	(0.25)	(0.26)
Firm age	−0.0044	0.0015	0.0011	−0.0080
	(0.0081)	(0.0072)	(0.0074)	(0.0078)
Acquisition of knowledge * Firm age		−0.010	−0.01	−0.01
		(0.0087)	(0.0088)	(0.0084)
Mills (attrition)			−0.31	
			(1.07)	
Mills (selection bias)				7.96***
				(2.30)
Constant	−4.04*	−3.87*	−3.47	−6.72***
	(1.90)	(1.95)	(2.59)	(1.83)
LL	−3.04	−3.03	−3.03	−2.97

When can selection bias be really serious?

- Inverse Mills Ratio has a significant coefficient ⇔ the error terms of the selection and outcome equations are significantly correlated (i.e., $\rho \neq 0$).

AND

- The independent variable(s) of interest must be a significant predictor in the selection equation
  - (test the significance of their coefficients in the first stage; also check the correlation between the x of interest and the IMR)

All in all, the significance of the IMR alone may not indicate (serious) sample selection bias. *(Certo et al., 2016)*

Heckman Models in STATA

dummy = 1 if wage observed
(selection indicator)

Key STATA commands:

- Heckman selection model (Two-Step)

```
heckman wage edu exp, select (in = married children edu exp)
twostep
```

outcome of interest
(only observed for a selected sample: `in = 1`)

possible exclusion restrictions
(variables from X_1 not in X_2)

Note that `edu` and `exp` are present in both equations!

- Heckman selection model (Maximum Likelihood)

```
heckman wage edu exp, select (in = married children edu exp)
```

STATA example

Female wage offers in a sample of American women

```
. tab inlf

  =1 if in |
 lab frce, |
      1975 |      Freq.     Percent        Cum.
-----------+-----------------------------------
         0 |        325       44.16       44.16
         1 |        411       55.84      100.00
-----------+-----------------------------------
     Total |        736      100.00
```

```
. tabstat lwage, by(inlf)

Summary for variables: lwage
     by categories of: inlf (=1 if in lab frce, 1975)

    inlf |      mean
---------+----------
       0 |         .
       1 |  1.273238
---------+----------
   Total |  1.273238
--------------------
```

Only 411 women in the labor force. lwage is only observed for this sub-sample.

We are interested in the effect of **education** and **experience** on women's wage.

Any potential problem?

OLS in a selected sample

```
. reg lwage educ exper expersq

      Source |       SS           df       MS      Number of obs   =       411
-------------+----------------------------------   F(3, 407)       =     30.53
       Model |   26.761859         3  8.92061966   Prob > F        =    0.0000
    Residual |  118.924353       407  .292197428   R-squared       =    0.1837
-------------+----------------------------------   Adj R-squared   =    0.1777
       Total |  145.686212       410  .355332225   Root MSE        =    .54055

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1056312   .0117748     8.97   0.000     .0824841    .1287783
       exper |       .024   .0110258     2.18   0.030     .0023254    .0456746
     expersq |  -.0004584   .0003257    -1.41   0.160    -.0010986    .0001819
       _cons |  -.2742848    .166226    -1.65   0.100    -.6010535     .052484
------------------------------------------------------------------------------
```

$$lwage = \beta_0 + \beta_1 educ + \beta_2\ exper + \beta_3 exper^2 + \varepsilon_2$$

Only 411 women in the labor force. lwage is only observed for this sub-sample.

What if ε_2 is correlated with educ or exper?

Digging into the potential problem

```
educ            byte    %9.0g                 years of schooling

Two-sample t test with equal variances
```

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
0	325	11.79692	.1210353	2.181995	11.55881	12.03504
1	411	12.6837	.1120819	2.272251	12.46337	12.90403
combined	736	12.29212	.0838368	2.274435	12.12753	12.45671
diff		-.8867752	.165744		-1.212164	-.5613864

```
    diff = mean(0) - mean(1)                              t =  -5.3503
Ho: diff = 0                             degrees of freedom =      734

    Ha: diff < 0             Ha: diff != 0               Ha: diff > 0
 Pr(T < t) = 0.0000    Pr(|T| > |t|) = 0.0000       Pr(T > t) = 1.0000
```

Women currently in the labor force are more educated and experienced than women outside of the labor force.

```
exper           byte    %9.0g                 actual labor mkt exper

Two-sample t test with equal variances
```

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
0	325	7.461538	.3837731	6.918567	6.706537	8.21654
1	411	13.23844	.3972984	8.054484	12.45745	14.01944
combined	736	10.6875	.2983866	8.095025	10.10171	11.27329
diff		-5.776904	.5622216		-6.880658	-4.67315

```
    diff = mean(0) - mean(1)                              t = -10.2751
Ho: diff = 0                             degrees of freedom =      734

    Ha: diff < 0             Ha: diff != 0               Ha: diff > 0
 Pr(T < t) = 0.0000    Pr(|T| > |t|) = 0.0000       Pr(T > t) = 1.0000
```

Any potential unobserved factors that can be correlated with both variables (and lwage)?
What if we are observing a ***positive selection*** of women?

Probit model of labor force participation

```
. ** Probit model of labor force participation <=> Heckman 1st step
. probit inlf educ exper expersq nwifeinc age kidslt6 kidsge6

Iteration 0:   log likelihood = -505.12037
Iteration 1:   log likelihood =  -389.9303
Iteration 2:   log likelihood = -389.31823
Iteration 3:   log likelihood =  -389.3179
Iteration 4:   log likelihood =  -389.3179

Probit regression                               Number of obs     =        736
                                                LR chi2(7)        =     231.60
                                                Prob > chi2       =     0.0000
Log likelihood =  -389.3179                     Pseudo R2         =     0.2293
```

inlf	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
educ	.1350915	.0258422	5.23	0.000	.0844417	.1857413
exper	.1308743	.0190863	6.86	0.000	.0934659	.1682828
expersq	-.0020464	.0006063	-3.38	0.001	-.0032348	-.0008581
nwifeinc	-.0115492	.0048753	-2.37	0.018	-.0211045	-.0019939
age	-.0534212	.0086097	-6.20	0.000	-.0702958	-.0365466
kidslt6	-.8738286	.1205009	-7.25	0.000	-1.110006	-.6376511
kidsge6	.036665	.0439668	0.83	0.404	-.0495083	.1228383
_cons	.1505762	.5179195	0.29	0.771	-.8645273	1.16568

4 new variables included in this probit (1st step) model: reasoning?

Use this probit model to estimate the IMR:

$$\frac{\phi(X_1'\beta_1)}{\Phi(X_1'\beta_1)}$$

OLS with *mills* as a regressor

```
. reg lwage educ exper expersq mills

      Source |       SS       df       MS              Number of obs =     411
-------------+------------------------------           F(4, 406)     =   22.86
       Model |  26.7758022     4  6.69395054           Prob > F      =  0.0000
    Residual |   118.91041   406  .292882783           R-squared     =  0.1838
-------------+------------------------------           Adj R-squared =  0.1757
       Total |  145.686212   410  .355332225           Root MSE      =  .54119

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1068455   .0130363     8.20   0.000     .0812184    .1324727
       exper |   .0258102   .0138088     1.87   0.062    -.0013354    .0529558
     expersq |  -.0004956   .0003681    -1.35   0.179    -.0012193     .000228
       mills |   .0241741   .1107941     0.22   0.827    -.1936276    .2419759
       _cons |   -.317652   .2592319    -1.23   0.221    -.8272563    .1919523
------------------------------------------------------------------------------
```

The t-test for the coefficient of ***mills*** suggests that selection bias is not a problem in the current data.

Heckman 2-step instead

```
. * Automatically instead, using Heckman 2step
. heckman lwage educ exper expersq, select(inlf = educ exper expersq nwifeinc age kidslt6 kidsge6) twostep

Heckman selection model -- two-step estimates   Number of obs      =       736
(regression model with sample selection)        Censored obs       =       325
                                                Uncensored obs     =       411

                                                Wald chi2(3)       =     68.00
                                                Prob > chi2        =    0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lwage        |
        educ |   .1068455   .0129602     8.24   0.000      .081444    .1322471
       exper |   .0258102    .013727     1.88   0.060    -.0010942    .0527145
     expersq |  -.0004956   .0003659    -1.35   0.176    -.0012129    .0002216
       _cons |   -.317652   .2577073    -1.23   0.218    -.8227489     .187445
-------------+----------------------------------------------------------------
inlf         |
        educ |   .1350915   .0258422     5.23   0.000     .0844417    .1857413
       exper |   .1308743   .0190863     6.86   0.000     .0934659    .1682828
     expersq |  -.0020464   .0006063    -3.38   0.001    -.0032348   -.0008581
    nwifeinc |  -.0115492   .0048753    -2.37   0.018    -.0211045   -.0019939
         age |  -.0534212   .0086097    -6.20   0.000    -.0702958   -.0365466
     kidslt6 |  -.8738286   .1205009    -7.25   0.000    -1.110006   -.6376511
     kidsge6 |    .036665   .0439668     0.83   0.404    -.0495083    .1228383
       _cons |   .1505762   .5179195     0.29   0.771    -.8645273     1.16568
-------------+----------------------------------------------------------------
mills        |
      lambda |   .0241741    .110135     0.22   0.826    -.1916865    .2400347
-------------+----------------------------------------------------------------
         rho |    0.04492
       sigma |  .53814503
------------------------------------------------------------------------------
```

Same results as an OLS including ***mills*** as a regressor.

Heckman Model by Maximum Likelihood

```
Heckman selection model                         Number of obs      =       736
(regression model with sample selection)        Censored obs       =       325
                                                Uncensored obs     =       411

                                                Wald chi2(3)       =     74.40
Log likelihood = -717.6465                      Prob > chi2        =    0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lwage        |
        educ |   .1063089   .0124263     8.56   0.000     .0819539    .1306639
       exper |   .0250121   .0125961     1.99   0.047     .0003242       .0497
     expersq |  -.0004792   .0003482    -1.38   0.169    -.0011617    .0002033
       _cons |  -.2985193   .2219991    -1.34   0.179    -.7336295     .136591
-------------+----------------------------------------------------------------
inlf         |
        educ |   .1355116   .0259857     5.21   0.000     .0845806    .1864425
       exper |   .1309137   .0190936     6.86   0.000      .093491    .1683364
     expersq |  -.0020482   .0006068    -3.38   0.001    -.0032375    -.000859
    nwifeinc |  -.0116317   .0049004    -2.37   0.018    -.0212364    -.002027
         age |  -.0534266   .0086108    -6.20   0.000    -.0703035   -.0365497
     kidslt6 |  -.8729929    .120636    -7.24   0.000    -1.109435   -.6365507
     kidsge6 |   .0364668   .0439742     0.83   0.407     -.049721    .1226546
       _cons |   .1474565     .51828     0.28   0.776    -.8683536    1.163267
-------------+----------------------------------------------------------------
     /athrho |   .0251354   .1536882     0.16   0.870     -.276088    .3263589
    /lnsigma |  -.6199196   .0349263   -17.75   0.000    -.6883739   -.5514654
-------------+----------------------------------------------------------------
         rho |   .0251301   .1535912                     -.2692806    .3152453
       sigma |   .5379877   .0187899                      .5023924     .576105
      lambda |   .0135197   .0826562                     -.1484835    .1755229
------------------------------------------------------------------------------
LR test of indep. eqns. (rho = 0):   chi2(1) =     0.03   Prob > chi2 = 0.8698
```

- Conclusions about **educ** and **exper**? An additional year of educ (exp) is associated with a 10.6% (2.5%*) increase in wage offered *(note that wage is in logs).*
- OLS would be consistent in this case, since selection bias is not a problem.

**on average, but at a sligthly decreasing rate*

Can we trust our exclusion restrictions?

```
. reg lwage educ exper expersq nwifeinc age kidslt6 kidsge6

      Source |       SS           df       MS      Number of obs   =       411
-------------+----------------------------------   F(7, 403)       =     13.62
       Model |  27.8667467         7  3.98096382   Prob > F        =    0.0000
    Residual |  117.819465       403  .292355994   R-squared       =    0.1913
-------------+----------------------------------   Adj R-squared   =    0.1772
       Total |  145.686212       410  .355332225   Root MSE        =     .5407

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1006009   .0125667     8.01   0.000     .0758964    .1253054
       exper |   .0230481   .0111876     2.06   0.040     .0010547    .0450414
     expersq |  -.0003973   .0003322    -1.20   0.232    -.0010504    .0002558
    nwifeinc |   .0042206   .0027117     1.56   0.120    -.0011103    .0095516
         age |  -.0035799   .0044829    -0.80   0.425    -.0123926    .0052329
     kidslt6 |  -.0746493   .0730267    -1.02   0.307    -.2182103    .0689116
     kidsge6 |  -.0141123   .0230222    -0.61   0.540     -.059371    .0311463
       _cons |  -.1127322   .2647116    -0.43   0.670    -.6331203    .4076559
------------------------------------------------------------------------------
```

```
. test nwifeinc age kidslt6 kidsge6

 ( 1)  nwifeinc = 0
 ( 2)  age = 0
 ( 3)  kidslt6 = 0
 ( 4)  kidsge6 = 0

       F(  4,   403) =    0.94
            Prob > F =    0.4379
```

By using these 4 variables as **exclusion restrictions**, we are assuming they are **jointly insignificant in the main (wage) equation.** They are indeed irrelevant for wages.

Yet, remember that selection bias may be serious in many other settings

"I'm pathetic, uninformed and don't give a damn.
Do you still want to continue with the poll?"

Wrap-up of today and next sessions

	Heckman models (11/09)	Matching Models (PSM) (11/16)	Instrumental Variables (11/23)
When	Y is missing in some cases (for a non-random reason)	…	…
Problem	The missings in Y are driven by a "selection process"	…	…
Stata commands	*heckman, (twostep)*	…	…
Key tests	Significance of the IMR or of the *rho*	…	…
Attention!	Need for valid exclusion restrictions; selection bias important when *IMR/rho* significant and *X* predicts selection (1^{st} stage)	…	…
First stage	Probit predicting selection into the sample (Y ≠ missing)	…	…