Selection Models

Chris Conlon

Tuesday 14th October, 2025

Applied Econometrics

Motivation: Sample Selection Bias

- Ordinary Least Squares (OLS) assumes the sample is randomly drawn from the population.
- In many cases, we only observe the outcome variable y_i for a selected subset of individuals.
- Example: Wages are only observed for those who work.

$$y_i = w_i \text{ observed only if } s_i = 1$$

- If the decision to work is correlated with unobservables affecting wages, OLS is biased.

Model Setup

Outcome Equation (latent)

$$y_i^* = x_i'\beta + \varepsilon_i,$$
$$\varepsilon_i \sim N(0, \sigma^2)$$

Selection Equation

$$s_i^* = z_i'\gamma + u_i,$$
$$s_i = 1[s_i^* > 0]$$
$$\begin{pmatrix} \varepsilon_i \\ u_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix} \right)$$

Conditional Expectation and Bias

We only observe y_i when $s_i = 1$. Then:

$$\begin{aligned}\mathbb{E}[y_i \mid x_i, s_i = 1] &= x_i'\beta + \mathbb{E}[\varepsilon_i \mid u_i > -z_i'\gamma] \\ &= x_i'\beta + \rho\sigma\lambda(z_i'\gamma)\end{aligned}$$

where

$$\lambda(z_i'\gamma) = \frac{\phi(z_i'\gamma)}{\Phi(z_i'\gamma)}.$$

Thus:

$$\mathbb{E}[y_i \mid x_i, s_i = 1] = x_i'\beta + \rho\sigma\lambda(z_i'\gamma)$$

1. **Step 1: Selection Equation (Probit)**

$$s_i = 1[z_i'\gamma + u_i > 0].$$

   Estimate $\hat{\gamma}$ via probit, compute

$$\hat{\lambda}_i = \frac{\phi(z_i'\hat{\gamma})}{\Phi(z_i'\hat{\gamma})}.$$

2. **Step 2: Outcome Equation (Corrected OLS)**

$$y_i = x_i'\beta + \rho\sigma\hat{\lambda}_i + \nu_i.$$

If $\rho \neq 0$, selection bias exists.

Note: standard errors must be corrected for the generated regressor $\hat{\lambda}_i$.

- ρ measures the correlation between the unobservables in the selection and outcome equations.
  - If $\rho \neq 0$, selection bias exists.
  - If $\rho = 0$, OLS on observed y_i is consistent.
- Identification relies on nonlinearity of $\lambda(z_i'\gamma)$, but it's better to have an **exclusion restriction**:

Some variables in z_i not in x_i.

Full Information Maximum Likelihood (FIML)

Joint likelihood for observed data:

$$\mathcal{L} = \prod_{i:s_i=1} f(y_i, s_i = 1 \mid x_i, z_i) \prod_{i:s_i=0} \text{Pr}(s_i = 0 \mid z_i)$$

Assuming joint normality:

$$\log \mathcal{L} = \sum_{s_i=1} \log \left[\frac{1}{\sigma} \phi \left(\frac{y_i - x_i' \beta}{\sigma} \right) \Phi \left(\frac{z_i' \gamma + \rho (y_i - x_i' \beta)/\sigma}{\sqrt{1 - \rho^2}} \right) \right] + \sum_{s_i=0} \log \left[1 - \Phi(z_i' \gamma) \right]$$

Estimated via MLE; asymptotically efficient.

```
n <- 2000

# regressors
educ  <- rnorm(n, 12, 2)
exper <- rnorm(n, 10, 5)

# create correlated errors (epsilon for wage, u for selection) with rho != 0
rho <- 0.6               # correlation between wage error and selection error
sigma_eps <- 1
Sigma <- matrix(c(sigma_eps^2, rho*sigma_eps,
                  rho*sigma_eps, 1), nrow = 2)
errors <- mvrnorm(n, mu = c(0,0), Sigma = Sigma)
eps <- errors[,1]   # wage disturbances
u   <- errors[,2]   # selection disturbances

# --------------------------
# Variant A: NO exclusion restriction
# selection and outcome share same regressors
# --------------------------
s_star_A <- 0.5 + 0.3*educ - 0.2*exper + u
select_A <- as.integer(s_star_A > 0)

wage_star <- 2 + 0.10*educ + 0.05*exper + eps
wage_A <- ifelse(select_A==1, wage_star, NA)

ols_A  <- lm(wage_A ~ educ + exper)
heck_A <- selection(select_A ~ educ + exper, wage_A ~ educ + exper, method = "ml")
```

Empirical Example in R: Heckman Results

```
Tobit 2 model (sample selection model)
Maximum Likelihood estimation
Newton-Raphson maximisation, 3 iterations
Return code 8: successive function values within relative tolerance limit (reltol)
Log-Likelihood: -2940.355
2000 observations (173 censored and 1827 observed)
8 free parameters (df = 1992)
Probit selection equation:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.19856    0.30870   0.643     0.52
educ         0.35155    0.03057  11.501   <2e-16 ***
exper       -0.21976    0.01460 -15.055   <2e-16 ***
Outcome equation:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.956873   0.153297  12.765  < 2e-16 ***
educ        0.095521   0.012536   7.619 3.91e-14 ***
exper       0.057020   0.005744   9.927  < 2e-16 ***
   Error terms:
      Estimate Std. Error t value Pr(>|t|)
sigma  1.01697    0.01851  54.935  < 2e-16 ***
rho    0.54797    0.09150   5.989  2.5e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Add excluded variable kids to selection equation:

```
kids <- rbinom(n, 1, 0.3)   # 0/1 indicator: has young children

s_star_B <- 0.5 + 0.3*educ - 0.2*exper - 0.8*kids + u   # kids reduces labor supply
select_B <- as.integer(s_star_B > 0)

wage_B <- ifelse(select_B==1, wage_star, NA)

ols_B  <- lm(wage_B ~ educ + exper)
heck_B <- selection(select_B ~ educ + exper + kids, wage_B ~ educ + exper, method = "ml")
```

OLS vs. Heckman Selection Model (with exclusion restriction)

	OLS (observed)	Heckman (MLE)	Heckman (Exclusion)
educ	0.071^{***} (0.012)	0.096^{***} (0.013)	0.100^{***} (0.013)
exper	0.078^{***} (0.005)	0.057^{***} (0.006)	0.057^{***} (0.006)
Constant	2.154^{***} (0.150)	1.957^{***} (0.153)	1.898^{***} (0.156)
Observations	1,754	2,000	2,000
R^2	0.138		
Adjusted R^2	0.137		
Log Likelihood		$-2,940.355$	$-2,903.073$
ρ		0.548^{***} (0.091)	0.647^{***} (0.068)
Note:			$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

- **Extensions:**
  - Multinomial or dynamic selection.
  - Semiparametric or nonparametric corrections.
  - Panel data with selection.
- **Applications:**
  - Wage equations (Heckman 1979)
  - Labor force participation
  - Credit approval models
  - Censored health outcomes

Example: Borjas (1987)

- Consider two countries $(0/1)$ (source and host).

$$\ln w_0 = \alpha_0 + u_0 \quad \text{with } u_0 \sim N(0, \sigma_0^2) \text{ source country}$$
$$\ln w_1 = \alpha_1 + u_1 \quad \text{with } u_1 \sim N(0, \sigma_1^2) \text{ host country}$$

- Now we allow for migration cost of C which he writes in hours: $\pi = \frac{C}{w_0}$.
- Assume workers know everything; you only see u_0 OR u_1 depending on country.
- Correlation in earnings is $\rho = \frac{\sigma_{01}}{\sigma_0 \sigma_1}$.

- Workers will migrate if:

$$(\alpha_1 - \alpha_0 - \pi) + (u_1 - u_0) > 0$$

- Who migrates? Probability of migration. Define $\nu = u_1 - u_0$.

$$P = \text{Pr}\left[\nu > (\alpha_0 - \alpha_1 + \pi)\right] = \text{Pr}\left[\frac{\nu}{\sigma_\nu} > \frac{(\alpha_0 - \alpha_1 + \pi)}{\sigma_\nu}\right]$$
$$= 1 - \Phi\left(\frac{(\alpha_0 - \alpha_1 + \pi)}{\sigma_\nu}\right) \equiv 1 - \Phi(z)$$

- Higher $z \rightarrow$ less migration.

Example: Borjas (1987): How does selection work?

Construct counterfactual wages for workers in source country for those who immigrate:

- For now ignore mean differences $\alpha_0 = \alpha_1 = \alpha$.

$$\mathbb{E}\left(w_0 | \text{ Immigrate }\right) = \alpha + \mathbb{E}\left(u_0 | \frac{\nu}{\sigma_\nu} > z\right)$$
$$= \alpha + \sigma_0 \cdot \mathbb{E}\left(\frac{u_0}{\sigma_0} | \frac{\nu}{\sigma_\nu} > z\right)$$

- Wages depend on:
  1. Mean earnings in the source country
  2. Both error terms (u_0, u_1) through ν
  3. Implicitly, it also depends on the correlation between the error terms.

Example: Borjas (1987): How does selection work?

- If everything is normal, we just run univariate regression $\mathbb{E}\left(u_0 | \nu\right) = \frac{\sigma_{0\nu}}{\sigma_\nu^2}\nu$:

$$\mathbb{E}\left(\frac{u_0}{\sigma_0} | \frac{\nu}{\sigma_\nu}\right) = \frac{1}{\sigma_0} \cdot \frac{\sigma_{0\nu}}{\sigma_\nu^2} \cdot \frac{\sigma_\nu^2}{\sigma_\nu^2} \cdot \nu = \frac{\sigma_{0\nu}}{\sigma_0 \sigma_\nu} \frac{\nu}{\sigma_\nu} = \rho_{0\nu} \frac{\nu}{\sigma_\nu}$$

$$\begin{aligned}
\mathbb{E}\left(w_0 | \text{ Immigrate }\right) &= \alpha_0 + \sigma_0 \cdot \mathbb{E}\left(\frac{u_0}{\sigma_0} | \frac{\nu}{\sigma_\nu} > z\right) \\
&= \alpha_0 + \rho_{0\nu} \cdot \sigma_0 \cdot \mathbb{E}\left(\frac{\nu}{\sigma_\nu} | \frac{\nu}{\sigma_\nu} > z\right) \\
&= \alpha_0 + \rho_{0\nu} \cdot \sigma_0 \left(\frac{\phi(z)}{1 - \Phi(z)}\right)
\end{aligned}$$

- This hazard rate of the standard normal has a special name Inverse Mills Ratio $\mathbb{E}[x | x > z]$.

Example: Borjas (1987): How does selection work?

- A similar expression for those who do immigrate:

$$\begin{aligned}\mathbb{E}\left(w_1 \middle| \text{ Immigrate }\right) &= \alpha_1 + \mathbb{E}\left(u_1 \middle| \frac{\nu}{\sigma_\nu} > z\right) \\ &= \alpha_1 + \rho_{1\nu}\sigma_1\left(\frac{\phi(z)}{\Phi(-z)}\right)\end{aligned}$$

- We can re-write both expressions in terms of the Inverse Mills Ratio

$$\begin{aligned}
\mathbb{E}\left(w_0 \middle| \text{ Immigrate }\right) &= \alpha_0 + \rho_{0\nu}\sigma_0 \left(\frac{\phi(z)}{1-\Phi(z)}\right) \\
&= \alpha_0 + \frac{\sigma_0\sigma_1}{\sigma_\nu}\left(\rho - \frac{\sigma_0}{\sigma_1}\right)\left(\frac{\phi(z)}{1-\Phi(z)}\right) \\
\mathbb{E}\left(w_1 \middle| \text{ Immigrate }\right) &= \alpha_1 + \rho_{1\nu}\sigma_1 \left(\frac{\phi(z)}{1-\Phi(z)}\right) \\
&= \alpha_1 + \frac{\sigma_0\sigma_1}{\sigma_\nu}\left(\frac{\sigma_1}{\sigma_0} - \rho\right)\left(\frac{\phi(z)}{1-\Phi(z)}\right)
\end{aligned}$$

Where $\rho = \sigma_{01}/\sigma_0\sigma_1$.

Positive Hierarchical Sorting

Let $Q_0 = \mathbb{E}\left(u_0 | I = 1\right), Q_1 = \mathbb{E}\left(u_1 | I = 1\right)$ (expected skill of immigrants).

- Immigrants are positively selected and above average $(Q_0, Q_1) > 0$ and $\frac{\sigma_1}{\sigma_0} > 1$ and $\rho > \frac{\sigma_0}{\sigma_1}$
  - $\frac{\sigma_1}{\sigma_0} > 1$ returns to "skill" are higher in host country.
  - $\rho > \frac{\sigma_0}{\sigma_1}$ correlation between valued skills in both counties is high (similar skills valued in both countries).
- Best and brightest leave because returns to skill are too low in home country.

We swap the standard deviations:

- Immigrants are negatively selected and below average $(Q_0, Q_1) < 0$ and $\frac{\sigma_1}{\sigma_0} > 1$ and $\rho > \frac{\sigma_0}{\sigma_1}$
    - $\frac{\sigma_0}{\sigma_1} > 1$ returns to “skill” are lower in host country.
    - $\rho > \frac{\sigma_1}{\sigma_0}$ correlation between valued skills in both counties is high (similar skills valued in both countries).
- Compressed wage structure attracts the low skill types because it provides “insurance” or “subsidizes” low wage workers.

- Immigrants are below average at home and above average in host $(Q_0 < 0, Q_1 > 1)$ and $\frac{\sigma_1}{\sigma_0} > 1$:
  - $\rho <$ min $\left(\frac{\sigma_1}{\sigma_0}, \frac{\sigma_0}{\sigma_1}\right)$ being below average in source country makes you above average in host country.
- You are a nerdy intellectual in a country that values physical labor, or are otherwise discriminated against in the labor market.

The missing (fourth) case:

- Mathematically impossible $\rho >$ max $\left(\frac{\sigma_1}{\sigma_0}, \frac{\sigma_0}{\sigma_1}\right)$

What can we learn here?

- Heckman won a Nobel Prize for his work on selection...
- You need to know what an inverse Mills ratio is
- But today it is hard to get away with strong parametric assumptions (bivariate normal) on error terms.
- Doing MLE with a fully normal model is not a terrible place to start sometimes
  - Sometimes helpful to know how bad the selection problem might be.
  - But you probably need exclusion restrictions!
- R package is sampleSelection and see https://rpubs.com/hacamvan/316839 and https://cran.r-project.org/web/packages/sampleSelection/vignettes/selection.pdf.