

# Mostly Harmless Econometrics

## **Session 1.1: Recitation**

Endogenous Regressors in Binary Choice & the Heckman Sample Selection Model

Kaiwen Luo

Washington University in St.Louis

October 16, 2026

## Today: two problems, one structure

Both problems are about **female labor supply**, and both share the same skeleton:

*two equations + jointly normal errors + one outcome only partially observed.*

We answer the same four questions for each, and map them onto the workflow from Session 1:

| Sub-question                          | Workflow step            | What we must show                          |
|---------------------------------------|--------------------------|--------------------------------------------|
| (1) Propose an estimator              | Estimation               | Two-step control function and full MLE     |
| (2) Assumption for identification     | Identification           | Exclusion restriction + relevance          |
| (3) Same regressors in both equations | Identification (failure) | What survives: reduced-form parameters     |
| (4) Stata command                     | Estimation & Inference   | Implementation and correct standard errors |

### Reading order

The problem asks (1) before (2). Logically, (2) comes first: an estimator is only meaningful once we know what it converges to.

## Problem 1: the setup

Female labor market participation with endogenous non-labor income:

$$y = \mathbf{1}\{x_y'\beta + w\gamma + \varepsilon > 0\} \quad (\text{participation})$$

$$w = x_w'\eta + \sigma_w v \quad (\text{non-labor income})$$

$$(\varepsilon, v) \mid x_y, x_w \sim N\left(\mathbf{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right). \quad (\text{the problem's } u \text{ is } \varepsilon)$$

### Observed

$$y \in \{0, 1\}, \quad w \in \mathbb{R}, \quad x_y, \quad x_w.$$

### Unknown

$$\beta, \gamma, \eta, \sigma_w, \rho.$$

- (1) Propose a method to estimate this model.
- (2) Which assumption identifies  $\gamma$ ? Show how.
- (3) If  $x_y = x_w$ , which reduced-form parameters involving  $\gamma$  are identified?
- (4) Which Stata command estimates this model?

## Before estimating: read the specification

**What is  $\gamma$ ?** The effect of non-labor income on the latent propensity to work. Theory (income effect  $\Rightarrow$  higher reservation wage) suggests  $\gamma < 0$ .

## Before estimating: read the specification

**What is  $\gamma$ ?** The effect of non-labor income on the latent propensity to work. Theory (income effect  $\Rightarrow$  higher reservation wage) suggests  $\gamma < 0$ .

**Why is  $w$  endogenous?** Unobserved household wealth, a spouse's earnings potential, or a taste for home production shift *both*  $w$  and the decision to work:

$$\text{Cov}(\varepsilon, w \mid x_y, x_w) = \text{Cov}(\varepsilon, \sigma_w v) = \rho \sigma_w \neq 0.$$

## Before estimating: read the specification

**What is  $\gamma$ ?** The effect of non-labor income on the latent propensity to work. Theory (income effect  $\Rightarrow$  higher reservation wage) suggests  $\gamma < 0$ .

**Why is  $w$  endogenous?** Unobserved household wealth, a spouse's earnings potential, or a taste for home production shift *both*  $w$  and the decision to work:

$$\text{Cov}(\varepsilon, w \mid x_y, x_w) = \text{Cov}(\varepsilon, \sigma_w v) = \rho \sigma_w \neq 0.$$

**Normalizations.**  $y^*$  is never observed, so its scale is not identified:  $\text{Var}(\varepsilon) = 1$ . The scale of  $w$  is observed and is carried by  $\sigma_w$ , so  $\text{Var}(v) = 1$  is a relabeling, not a restriction.

## Before estimating: read the specification

**What is  $\gamma$ ?** The effect of non-labor income on the latent propensity to work. Theory (income effect  $\Rightarrow$  higher reservation wage) suggests  $\gamma < 0$ .

**Why is  $w$  endogenous?** Unobserved household wealth, a spouse's earnings potential, or a taste for home production shift *both*  $w$  and the decision to work:

$$\text{Cov}(\varepsilon, w \mid x_y, x_w) = \text{Cov}(\varepsilon, \sigma_w v) = \rho \sigma_w \neq 0.$$

**Normalizations.**  $y^*$  is never observed, so its scale is not identified:  $\text{Var}(\varepsilon) = 1$ . The scale of  $w$  is observed and is carried by  $\sigma_w$ , so  $\text{Var}(v) = 1$  is a relabeling, not a restriction.

### Why the naive probit fails

A probit of  $y$  on  $(x_y, w)$  assumes  $\varepsilon \perp w$ . Here  $\mathbb{E}[\varepsilon \mid w, x] = \rho v \neq 0$ , so the coefficient on  $w$  mixes the causal effect  $\gamma$  with the selection term  $\rho/\sigma_w$ .

## The key trick: decompose $\varepsilon$ given $v$

Joint normality gives a linear projection with an independent remainder:

$$\varepsilon = \rho v + e, \quad e \sim N(0, 1 - \rho^2), \quad e \perp (v, x_y, x_w).$$

## The key trick: decompose $\varepsilon$ given $v$

Joint normality gives a linear projection with an independent remainder:

$$\varepsilon = \rho v + e, \quad e \sim N(0, 1 - \rho^2), \quad e \perp (v, x_y, x_w).$$

Given  $(x_y, x_w, w)$ , the residual  $v = (w - x'_w \eta) / \sigma_w$  is known. Hence

$$\mathbb{P}(y = 1 \mid x_y, x_w, w) = \mathbb{P}(e > -x'_y \beta - w\gamma - \rho v)$$

$$\mathbb{P}(y = 1 \mid x_y, x_w, w) = \Phi\left(\frac{x'_y \beta + w\gamma + \rho v}{\sqrt{1 - \rho^2}}\right)$$

## The key trick: decompose $\varepsilon$ given $v$

Joint normality gives a linear projection with an independent remainder:

$$\varepsilon = \rho v + e, \quad e \sim N(0, 1 - \rho^2), \quad e \perp (v, x_y, x_w).$$

Given  $(x_y, x_w, w)$ , the residual  $v = (w - x'_w \eta) / \sigma_w$  is known. Hence

$$\mathbb{P}(y = 1 \mid x_y, x_w, w) = \mathbb{P}(e > -x'_y \beta - w\gamma - \rho v)$$

$$\mathbb{P}(y = 1 \mid x_y, x_w, w) = \Phi\left(\frac{x'_y \beta + w\gamma + \rho v}{\sqrt{1 - \rho^2}}\right)$$

### Interpretation: a control function

Once we condition on  $v$ , the part of  $\varepsilon$  that is correlated with  $w$  is held fixed. What remains,  $e$ , is independent of  $w$ :  $v$  “controls for” the source of endogeneity.

## (1) Method A: two-step control function

Rivers and Vuong (1988); Smith and Blundell (1986) for the Tobit.

**Step 1.** OLS of  $w$  on  $x_w$ :

$$\hat{\eta}, \quad \hat{\sigma}_w^2 = \frac{1}{n} \sum_i (w_i - x'_{w,i} \hat{\eta})^2, \quad \hat{v}_i = (w_i - x'_{w,i} \hat{\eta}) / \hat{\sigma}_w.$$

**Step 2.** Probit of  $y$  on  $(x_y, w, \hat{v})$ :

$$\mathbb{P}(y = 1 \mid x_y, w, \hat{v}) = \Phi(x'_y \beta_\rho + w \gamma_\rho + \hat{v} \theta_\rho), \quad (\beta_\rho, \gamma_\rho, \theta_\rho) = \frac{(\beta, \gamma, \rho)}{\sqrt{1 - \rho^2}}.$$

## (1) Method A: two-step control function

Rivers and Vuong (1988); Smith and Blundell (1986) for the Tobit.

**Step 1.** OLS of  $w$  on  $x_w$ :

$$\hat{\eta}, \quad \hat{\sigma}_w^2 = \frac{1}{n} \sum_i (w_i - x'_{w,i} \hat{\eta})^2, \quad \hat{v}_i = (w_i - x'_{w,i} \hat{\eta}) / \hat{\sigma}_w.$$

**Step 2.** Probit of  $y$  on  $(x_y, w, \hat{v})$ :

$$\mathbb{P}(y = 1 \mid x_y, w, \hat{v}) = \Phi(x'_y \beta_\rho + w \gamma_\rho + \hat{v} \theta_\rho), \quad (\beta_\rho, \gamma_\rho, \theta_\rho) = \frac{(\beta, \gamma, \rho)}{\sqrt{1 - \rho^2}}.$$

**Step 3.** Undo the scaling. Since  $1 + \theta_\rho^2 = 1/(1 - \rho^2)$ ,

$$\hat{\rho} = \frac{\hat{\theta}_\rho}{\sqrt{1 + \hat{\theta}_\rho^2}}, \quad \hat{\gamma} = \frac{\hat{\gamma}_\rho}{\sqrt{1 + \hat{\theta}_\rho^2}}, \quad \hat{\beta} = \frac{\hat{\beta}_\rho}{\sqrt{1 + \hat{\theta}_\rho^2}}.$$

## (1) Method A: two-step control function

Rivers and Vuong (1988); Smith and Blundell (1986) for the Tobit.

**Step 1.** OLS of  $w$  on  $x_w$ :

$$\hat{\eta}, \quad \hat{\sigma}_w^2 = \frac{1}{n} \sum_i (w_i - x'_{w,i} \hat{\eta})^2, \quad \hat{v}_i = (w_i - x'_{w,i} \hat{\eta}) / \hat{\sigma}_w.$$

**Step 2.** Probit of  $y$  on  $(x_y, w, \hat{v})$ :

$$\mathbb{P}(y = 1 \mid x_y, w, \hat{v}) = \Phi(x'_y \beta_\rho + w \gamma_\rho + \hat{v} \theta_\rho), \quad (\beta_\rho, \gamma_\rho, \theta_\rho) = \frac{(\beta, \gamma, \rho)}{\sqrt{1 - \rho^2}}.$$

**Step 3.** Undo the scaling. Since  $1 + \theta_\rho^2 = 1/(1 - \rho^2)$ ,

$$\hat{\rho} = \frac{\hat{\theta}_\rho}{\sqrt{1 + \hat{\theta}_\rho^2}}, \quad \hat{\gamma} = \frac{\hat{\gamma}_\rho}{\sqrt{1 + \hat{\theta}_\rho^2}}, \quad \hat{\beta} = \frac{\hat{\beta}_\rho}{\sqrt{1 + \hat{\theta}_\rho^2}}.$$

### Why it works

Step 1 makes the control  $v$  observable; Step 2 is then an ordinary probit whose error  $e/\sqrt{1 - \rho^2}$  is standard normal and independent of all regressors.

## (1) Inference for the two-step estimator

### Generated regressor

$\hat{v}_i$  is estimated, not observed. The usual probit standard errors from Step 2 ignore the sampling error in  $(\hat{\eta}, \hat{\sigma}_w)$  and are **too small**. Fix: Murphy–Topel correction, or bootstrap *both* steps together.

## (1) Inference for the two-step estimator

### Generated regressor

$\hat{v}_i$  is estimated, not observed. The usual probit standard errors from Step 2 ignore the sampling error in  $(\hat{\eta}, \hat{\sigma}_w)$  and are **too small**. Fix: Murphy–Topel correction, or bootstrap *both* steps together.

### A free test of exogeneity

Under  $H_0 : \rho = 0$  we have  $\theta_\rho = 0$ , so the first-step error does not affect Step 2. The *usual*  $t$ -statistic on  $\hat{v}$  is a valid test of  $H_0 : w$  is exogenous.

## (1) Inference for the two-step estimator

### Generated regressor

$\hat{v}_i$  is estimated, not observed. The usual probit standard errors from Step 2 ignore the sampling error in  $(\hat{\eta}, \hat{\sigma}_w)$  and are **too small**. Fix: Murphy–Topel correction, or bootstrap *both* steps together.

### A free test of exogeneity

Under  $H_0 : \rho = 0$  we have  $\theta_\rho = 0$ , so the first-step error does not affect Step 2. The *usual*  $t$ -statistic on  $\hat{v}$  is a valid test of  $H_0 : w$  is exogenous.

### Do not “plug in” $\hat{w}$

Replacing  $w$  by  $\hat{w} = x'_w \hat{\eta}$  in a probit (a “forbidden regression”) leaves the composite error  $\varepsilon + \gamma \sigma_w v$ , with variance  $1 + 2\rho\gamma\sigma_w + \gamma^2\sigma_w^2$  that depends on  $\gamma$ . It recovers  $(\beta, \gamma)$  only up to this unknown scale, relies on exact normality, and does not generalize to other nonlinear models.

## (1) Method B: full maximum likelihood

Factor the joint density of what we observe:

$$f(y, w \mid x) = \underbrace{\mathbb{P}(y \mid w, x)}_{\text{probit with control}} \cdot \underbrace{f(w \mid x)}_{N(x'_w \eta, \sigma_w^2)}.$$

## (1) Method B: full maximum likelihood

Factor the joint density of what we observe:

$$f(y, w | x) = \underbrace{\mathbb{P}(y | w, x)}_{\text{probit with control}} \cdot \underbrace{f(w | x)}_{N(x'_w \eta, \sigma_w^2)}.$$

With  $v_i = (w_i - x'_{w,i} \eta) / \sigma_w$  and  $m_i = \frac{x'_{y,i} \beta + w_i \gamma + \rho v_i}{\sqrt{1 - \rho^2}}$ , the log-likelihood is

$$\ell(\beta, \gamma, \eta, \sigma_w, \rho) = \sum_i \left\{ y_i \log \Phi(m_i) + (1 - y_i) \log [1 - \Phi(m_i)] + \log \phi(v_i) - \log \sigma_w \right\}$$

## (1) Method B: full maximum likelihood

Factor the joint density of what we observe:

$$f(y, w | x) = \underbrace{\mathbb{P}(y | w, x)}_{\text{probit with control}} \cdot \underbrace{f(w | x)}_{N(x'_w \eta, \sigma_w^2)}.$$

With  $v_i = (w_i - x'_{w,i} \eta) / \sigma_w$  and  $m_i = \frac{x'_{y,i} \beta + w_i \gamma + \rho v_i}{\sqrt{1 - \rho^2}}$ , the log-likelihood is

$$\ell(\beta, \gamma, \eta, \sigma_w, \rho) = \sum_i \left\{ y_i \log \Phi(m_i) + (1 - y_i) \log [1 - \Phi(m_i)] + \log \phi(v_i) - \log \sigma_w \right\}$$

- Efficient under the model; standard errors come directly from the Hessian.
- In practice, optimize over  $\text{atanh } \rho$  and  $\log \sigma_w$  so the parameter space is unrestricted.
- The two-step estimates are good starting values.

## (1) Coefficients are not the end: what is the estimand?

The model gives  $\gamma$  only *in units of*  $\text{sd}(\varepsilon)$ . A policy question is usually about probabilities.

**Average structural function** (Blundell and Powell, 2004):

$$\text{ASF}(x_y, w) = \mathbb{E}_\varepsilon [\mathbf{1}\{x_y'\beta + w\gamma + \varepsilon > 0\}] = \Phi(x_y'\beta + w\gamma).$$

## (1) Coefficients are not the end: what is the estimand?

The model gives  $\gamma$  only *in units of*  $\text{sd}(\varepsilon)$ . A policy question is usually about probabilities.

**Average structural function** (Blundell and Powell, 2004):

$$\text{ASF}(x_y, w) = \mathbb{E}_\varepsilon[\mathbf{1}\{x'_y\beta + w\gamma + \varepsilon > 0\}] = \Phi(x'_y\beta + w\gamma).$$

With the two-step estimates, average over the control function instead:

$$\widehat{\text{ASF}}(x_y, w) = \frac{1}{n} \sum_{i=1}^n \Phi(x'_y \hat{\beta}_\rho + w \hat{\gamma}_\rho + \hat{v}_i \hat{\theta}_\rho) \xrightarrow{P} \Phi(x'_y\beta + w\gamma),$$

because  $\rho v + e \sim N(0, 1)$ .

## (1) Coefficients are not the end: what is the estimand?

The model gives  $\gamma$  only *in units of*  $\text{sd}(\varepsilon)$ . A policy question is usually about probabilities.

**Average structural function** (Blundell and Powell, 2004):

$$\text{ASF}(x_y, w) = \mathbb{E}_\varepsilon[\mathbf{1}\{x'_y\beta + w\gamma + \varepsilon > 0\}] = \Phi(x'_y\beta + w\gamma).$$

With the two-step estimates, average over the control function instead:

$$\widehat{\text{ASF}}(x_y, w) = \frac{1}{n} \sum_{i=1}^n \Phi(x'_y \hat{\beta}_\rho + w \hat{\gamma}_\rho + \hat{v}_i \hat{\theta}_\rho) \xrightarrow{P} \Phi(x'_y\beta + w\gamma),$$

because  $\rho v + e \sim N(0, 1)$ .

Average partial effect of non-labor income

$$\text{APE}_w = \mathbb{E}[\gamma \phi(x'_y\beta + w\gamma)]$$

This is the object to compare with a linear probability model estimated by 2SLS.

## (2) The assumption: an exclusion restriction

Split the first-stage index into variables that do and do not enter  $y$ :

$$x_w' \eta = x_y' \pi_1 + z' \pi_2, \quad z = \text{variables in } x_w \text{ excluded from } x_y.$$

### Assumption (exclusion + relevance)

There is at least one variable  $z$  that

- ① does not enter the participation equation (exclusion)
- ② shifts non-labor income:  $\pi_2 \neq 0$  (relevance)
- ③ and  $\mathbb{E}[(x_y, z)(x_y, z)']$  has full rank (no redundancy)

## (2) The assumption: an exclusion restriction

Split the first-stage index into variables that do and do not enter  $y$ :

$$x'_w \eta = x'_y \pi_1 + z' \pi_2, \quad z = \text{variables in } x_w \text{ excluded from } x_y.$$

### Assumption (exclusion + relevance)

There is at least one variable  $z$  that

- ① does not enter the participation equation (exclusion)
- ② shifts non-labor income:  $\pi_2 \neq 0$  (relevance)
- ③ and  $\mathbb{E}[(x_y, z)(x_y, z)']$  has full rank (no redundancy)

**Example:** husband's education shifts household non-labor income, but (arguably) not the wife's own taste for work.

### Exactly as in linear IV

Normality buys us the shape of  $\mathbb{P}(y = 1 \mid \cdot)$ . It does *not* replace the instrument.

## (2) Showing that $\gamma$ is identified

**Step 1: the first stage.**  $\mathbb{E}[w \mid x] = x'_w \eta$  and  $\text{Var}(w \mid x) = \sigma_w^2$  identify  $(\eta, \sigma_w)$ . Therefore  $v = (w - x'_w \eta) / \sigma_w$  is a known function of observables.

## (2) Showing that $\gamma$ is identified

**Step 1: the first stage.**  $\mathbb{E}[w \mid x] = x'_w \eta$  and  $\text{Var}(w \mid x) = \sigma_w^2$  identify  $(\eta, \sigma_w)$ . Therefore  $v = (w - x'_w \eta) / \sigma_w$  is a known function of observables.

**Step 2: invert the probit.** Since  $\Phi$  is strictly increasing,

$$\Phi^{-1}(\mathbb{P}(y = 1 \mid x, w)) = x'_y \beta_\rho + w \gamma_\rho + v \theta_\rho.$$

The left side is a known function of the data. The coefficients are identified iff  $(x_y, w, v)$  are not perfectly collinear.

## (2) Showing that $\gamma$ is identified

**Step 1: the first stage.**  $\mathbb{E}[w \mid x] = x'_w \eta$  and  $\text{Var}(w \mid x) = \sigma_w^2$  identify  $(\eta, \sigma_w)$ . Therefore  $v = (w - x'_w \eta) / \sigma_w$  is a known function of observables.

**Step 2: invert the probit.** Since  $\Phi$  is strictly increasing,

$$\Phi^{-1}(\mathbb{P}(y = 1 \mid x, w)) = x'_y \beta_\rho + w \gamma_\rho + v \theta_\rho.$$

The left side is a known function of the data. The coefficients are identified iff  $(x_y, w, v)$  are not perfectly collinear.

**Step 3: rule out collinearity.**

$$w = x'_y \pi_1 + z' \pi_2 + \sigma_w v.$$

$w$  is a linear combination of  $(x_y, v) \iff z' \pi_2 \equiv 0$ . The exclusion restriction with  $\pi_2 \neq 0$  rules this out.

## (2) Showing that $\gamma$ is identified

**Step 1: the first stage.**  $\mathbb{E}[w \mid x] = x'_w \eta$  and  $\text{Var}(w \mid x) = \sigma_w^2$  identify  $(\eta, \sigma_w)$ . Therefore  $v = (w - x'_w \eta)/\sigma_w$  is a known function of observables.

**Step 2: invert the probit.** Since  $\Phi$  is strictly increasing,

$$\Phi^{-1}(\mathbb{P}(y = 1 \mid x, w)) = x'_y \beta_\rho + w \gamma_\rho + v \theta_\rho.$$

The left side is a known function of the data. The coefficients are identified iff  $(x_y, w, v)$  are not perfectly collinear.

**Step 3: rule out collinearity.**

$$w = x'_y \pi_1 + z' \pi_2 + \sigma_w v.$$

$w$  is a linear combination of  $(x_y, v) \iff z' \pi_2 \equiv 0$ . The exclusion restriction with  $\pi_2 \neq 0$  rules this out.

**Step 4: undo the scaling.**  $\rho = \theta_\rho / \sqrt{1 + \theta_\rho^2}$  and  $\gamma = \gamma_\rho / \sqrt{1 + \theta_\rho^2}$ . ■

## (2) The intuition

Think of  $v$  as a woman's unobserved "type".

### With an instrument

Fix  $x_y$  and  $v$ . Varying  $z$  still moves  $w$ .  
Any resulting change in  $\mathbb{P}(y = 1)$  must come from  $\gamma$ , since the type is held fixed.

### Without an instrument

Fix  $x_y$  and  $v$ . Then  $w$  is fixed too.  
Every movement in  $w$  comes with a movement in  $v$ , so  $\gamma$  cannot be separated from  $\rho$ .

Identification = variation in  $w$  that is not variation in the unobservables

This is the logic of the first stage in 2SLS. Part (3) makes the failure case concrete.

**(3) When  $x_y = x_w = x$ : substitute out  $v$**

Now  $v = (w - x'\eta)/\sigma_w$ , and the probit index becomes

$$\frac{x'\beta + w\gamma + \rho(w - x'\eta)/\sigma_w}{\sqrt{1 - \rho^2}} = x' \underbrace{\frac{\beta - \rho\eta/\sigma_w}{\sqrt{1 - \rho^2}}}_{\equiv a} + w \underbrace{\frac{\gamma + \rho/\sigma_w}{\sqrt{1 - \rho^2}}}_{\equiv b}.$$

### (3) When $x_y = x_w = x$ : substitute out $v$

Now  $v = (w - x'\eta)/\sigma_w$ , and the probit index becomes

$$\frac{x'\beta + w\gamma + \rho(w - x'\eta)/\sigma_w}{\sqrt{1 - \rho^2}} = x' \underbrace{\frac{\beta - \rho\eta/\sigma_w}{\sqrt{1 - \rho^2}}}_{\equiv a} + w \underbrace{\frac{\gamma + \rho/\sigma_w}{\sqrt{1 - \rho^2}}}_{\equiv b}.$$

#### What the data identify

From  $f(w | x)$ :  $\eta, \sigma_w$ .      From  $\mathbb{P}(y = 1 | x, w) = \Phi(x'a + wb)$ :  $a, b$ .

These four objects pin down the entire distribution of  $(y, w) | x$ .

### (3) When $x_y = x_w = x$ : substitute out $v$

Now  $v = (w - x'\eta)/\sigma_w$ , and the probit index becomes

$$\frac{x'\beta + w\gamma + \rho(w - x'\eta)/\sigma_w}{\sqrt{1 - \rho^2}} = x' \underbrace{\frac{\beta - \rho\eta/\sigma_w}{\sqrt{1 - \rho^2}}}_{\equiv a} + w \underbrace{\frac{\gamma + \rho/\sigma_w}{\sqrt{1 - \rho^2}}}_{\equiv b}.$$

#### What the data identify

From  $f(w | x)$ :  $\eta, \sigma_w$ .      From  $\mathbb{P}(y = 1 | x, w) = \Phi(x'a + wb)$ :  $a, b$ .

These four objects pin down the entire distribution of  $(y, w) | x$ .

#### Equivalently: the reduced form for $y$ alone

Substituting  $w$  into the  $y$  equation,  $y = \mathbf{1}\{x'(\beta + \gamma\eta) + \varepsilon + \gamma\sigma_w v > 0\}$ , so

$$\mathbb{P}(y = 1 | x) = \Phi(x'\pi), \quad \pi = \frac{\beta + \gamma\eta}{\omega}, \quad \omega^2 = 1 + 2\rho\gamma\sigma_w + \gamma^2\sigma_w^2.$$

One can check that  $\pi = (a + b\eta)/\sqrt{1 + b^2\sigma_w^2}$ : nothing new.

### (3) Why $\gamma$ itself is lost

**Counting.** Unknowns  $(\beta, \gamma, \rho)$ :  $k + 2$ . Identified  $(a, b)$ :  $k + 1$ . One equation short.

### (3) Why $\gamma$ itself is lost

**Counting.** Unknowns  $(\beta, \gamma, \rho)$ :  $k + 2$ . Identified  $(a, b)$ :  $k + 1$ . One equation short.

**Observational equivalence.** Pick any  $\tilde{\rho} \in (-1, 1)$  and set

$$\tilde{\gamma} = b\sqrt{1 - \tilde{\rho}^2} - \tilde{\rho}/\sigma_w, \quad \tilde{\beta} = a\sqrt{1 - \tilde{\rho}^2} + \tilde{\rho}\eta/\sigma_w.$$

Then  $(\tilde{\beta}, \tilde{\gamma}, \tilde{\rho})$  produces the same  $(a, b)$ , hence the same data.

### (3) Why $\gamma$ itself is lost

**Counting.** Unknowns  $(\beta, \gamma, \rho)$ :  $k + 2$ . Identified  $(a, b)$ :  $k + 1$ . One equation short.

**Observational equivalence.** Pick any  $\tilde{\rho} \in (-1, 1)$  and set

$$\tilde{\gamma} = b\sqrt{1 - \tilde{\rho}^2} - \tilde{\rho}/\sigma_w, \quad \tilde{\beta} = a\sqrt{1 - \tilde{\rho}^2} + \tilde{\rho}\eta/\sigma_w.$$

Then  $(\tilde{\beta}, \tilde{\gamma}, \tilde{\rho})$  produces the same  $(a, b)$ , hence the same data.

What the identified parameter means

$$b\sqrt{1 - \rho^2} = \gamma + \frac{\rho}{\sigma_w} = \underbrace{\gamma}_{\text{causal}} + \underbrace{\frac{\text{Cov}(\varepsilon, w \mid x)}{\text{Var}(w \mid x)}}_{\text{selection on unobservables}}$$

The familiar omitted-variable-bias formula, inside a probit.

### (3) Why $\gamma$ itself is lost

**Counting.** Unknowns  $(\beta, \gamma, \rho)$ :  $k + 2$ . Identified  $(a, b)$ :  $k + 1$ . One equation short.

**Observational equivalence.** Pick any  $\tilde{\rho} \in (-1, 1)$  and set

$$\tilde{\gamma} = b\sqrt{1 - \tilde{\rho}^2} - \tilde{\rho}/\sigma_w, \quad \tilde{\beta} = a\sqrt{1 - \tilde{\rho}^2} + \tilde{\rho}\eta/\sigma_w.$$

Then  $(\tilde{\beta}, \tilde{\gamma}, \tilde{\rho})$  produces the same  $(a, b)$ , hence the same data.

#### What the identified parameter means

$$b\sqrt{1 - \rho^2} = \gamma + \frac{\rho}{\sigma_w} = \underbrace{\gamma}_{\text{causal}} + \underbrace{\frac{\text{Cov}(\varepsilon, w \mid x)}{\text{Var}(w \mid x)}}_{\text{selection on unobservables}}$$

The familiar omitted-variable-bias formula, inside a probit.

#### Normality does not rescue us here

The control  $\rho v$  enters the index *linearly*, so it is exactly collinear with  $(x, w)$ . Only outside information on  $\rho$  (e.g.  $\rho = 0$ , or its sign) turns  $b$  into a statement about  $\gamma$ .

## (4) Stata

Wooldridge's mroz data:  $\text{inlf} = y$ ,  $\text{nwifeinc} = w$ ,  $\text{huseduc} = z$ .

```
bcuse mroz, clear                // ssc install bcuse

// (a) Full MLE (default)
ivprobit inlf educ exper expersq age kidslt6 kidsge6 (nwifeinc = huseduc)
margins, dydx(nwifeinc) predict(pr)    // APE from the structural probability

// (b) Two-step estimator (Newey 1987), corrected standard errors
ivprobit inlf educ exper expersq age kidslt6 kidsge6 (nwifeinc = huseduc), twostep

// (c) Control function by hand: point estimates only, bootstrap the SEs
regress nwifeinc educ exper expersq age kidslt6 kidsge6 huseduc
predict vhat, residuals
probit inlf educ exper expersq age kidslt6 kidsge6 nwifeinc vhat

// (d) Extended regression models (Stata 15+)
eprobit inlf educ exper expersq age kidslt6 kidsge6, ///
    endogenous(nwifeinc = educ exper expersq age kidslt6 kidsge6 huseduc)
```

- ivprobit reports a Wald test of exogeneity ( $H_0 : \rho = 0$ ); in (c), the  $t$ -test on vhat is the same idea.
- Stata's first stage regresses  $w$  on *all* exogenous variables,  $x_y \cup x_w$ .

## Problem 2: the setup

Gronau (1974); Heckman (1979). We want the **wage offer** equation for *all* women, but wages are observed only for those who work:

$$\begin{aligned}y &= x'\beta + u_1 && \text{(log wage offer, observed iff } d = 1\text{)} \\d &= \mathbf{1}\{z'\delta + u_2 > 0\} && \text{(participation)}\end{aligned}$$

$$(u_1, u_2) \mid x, z \sim N\left(\mathbf{0}, \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix}\right).$$

### Observed

$d$  and  $(x, z)$  for everyone;  $y$  only if  $d = 1$ .

### Unknown

$\beta, \delta, \sigma, \rho$ .

- (1) Propose a method to estimate this model.
- (2) Which assumption identifies  $\beta$ ? Show how.
- (3) If  $x = z$ , what can still be identified?
- (4) Which Stata command estimates this model?

## Before estimating: read the specification

**What is  $\beta$ ?** E.g. the return to education in the *offered* wage, for the whole population of women. It is **not** the slope of observed wages among workers.

## Before estimating: read the specification

**What is  $\beta$ ?** E.g. the return to education in the *offered* wage, for the whole population of women. It is **not** the slope of observed wages among workers.

**Why is OLS on workers biased?** Women who work may be those with unusually good wage offers:

$$\mathbb{E}[y \mid x, z, d = 1] = x'\beta + \mathbb{E}[u_1 \mid u_2 > -z'\delta] \neq x'\beta \quad \text{when } \rho \neq 0.$$

The bias term varies with  $z$ , which overlaps with  $x$ , so the slopes are biased, not just the intercept.

## Before estimating: read the specification

**What is  $\beta$ ?** E.g. the return to education in the *offered* wage, for the whole population of women. It is **not** the slope of observed wages among workers.

**Why is OLS on workers biased?** Women who work may be those with unusually good wage offers:

$$\mathbb{E}[y \mid x, z, d = 1] = x'\beta + \mathbb{E}[u_1 \mid u_2 > -z'\delta] \neq x'\beta \quad \text{when } \rho \neq 0.$$

The bias term varies with  $z$ , which overlaps with  $x$ , so the slopes are biased, not just the intercept.

**Normalizations.**  $d^*$  is latent:  $\text{Var}(u_2) = 1$ .  $y$  is observed in levels, so  $\sigma$  is a free parameter.

### Compare with Part I

Part I: a *regressor* is endogenous. Part II: the *sample* is endogenous. Once written down, both are omitted-variable problems.

The key trick again: decompose  $u_1$  given  $u_2$

Joint normality:

$$u_1 = \rho\sigma u_2 + e, \quad e \sim N(0, \sigma^2(1 - \rho^2)), \quad e \perp (u_2, x, z).$$

## The key trick again: decompose $u_1$ given $u_2$

Joint normality:

$$u_1 = \rho\sigma u_2 + e, \quad e \sim N(0, \sigma^2(1 - \rho^2)), \quad e \perp (u_2, x, z).$$

Truncated normal mean: for  $c = z'\delta$ ,

$$\mathbb{E}[u_2 \mid u_2 > -c] = \frac{\phi(c)}{\Phi(c)} \equiv \lambda(c) \quad (\text{inverse Mills ratio}).$$

## The key trick again: decompose $u_1$ given $u_2$

Joint normality:

$$u_1 = \rho\sigma u_2 + e, \quad e \sim N(0, \sigma^2(1 - \rho^2)), \quad e \perp (u_2, x, z).$$

Truncated normal mean: for  $c = z'\delta$ ,

$$\mathbb{E}[u_2 \mid u_2 > -c] = \frac{\phi(c)}{\Phi(c)} \equiv \lambda(c) \quad (\text{inverse Mills ratio}).$$

Therefore

$$\mathbb{E}[y \mid x, z, d = 1] = x'\beta + \underbrace{\rho\sigma}_{\equiv \theta} \lambda(z'\delta)$$

## The key trick again: decompose $u_1$ given $u_2$

Joint normality:

$$u_1 = \rho\sigma u_2 + e, \quad e \sim N(0, \sigma^2(1 - \rho^2)), \quad e \perp (u_2, x, z).$$

Truncated normal mean: for  $c = z'\delta$ ,

$$\mathbb{E}[u_2 \mid u_2 > -c] = \frac{\phi(c)}{\Phi(c)} \equiv \lambda(c) \quad (\text{inverse Mills ratio}).$$

Therefore

$$\mathbb{E}[y \mid x, z, d = 1] = x'\beta + \underbrace{\rho\sigma}_{\equiv \theta} \lambda(z'\delta)$$

### Interpretation: a control function

Selection bias is an **omitted variable**,  $\lambda(z'\delta)$ . Part I controlled for  $v$ , the residual of a *continuous* first stage; here we control for  $\lambda$ , the generalized residual of a *binary* first stage.

## (1) Method A: Heckman's two-step ("Heckit")

**Step 1.** Probit of  $d$  on  $z$ , using the full sample:

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

**Step 2.** OLS of  $y$  on  $(x, \hat{\lambda})$ , using workers only ( $d_i = 1$ ):

$$y_i = x_i' \beta + \theta \hat{\lambda}_i + \text{error}_i.$$

## (1) Method A: Heckman's two-step ("Heckit")

**Step 1.** Probit of  $d$  on  $z$ , using the full sample:

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

**Step 2.** OLS of  $y$  on  $(x, \hat{\lambda})$ , using workers only ( $d_i = 1$ ):

$$y_i = x_i' \beta + \theta \hat{\lambda}_i + \text{error}_i.$$

### Inference

- The error is **heteroskedastic**:  $\text{Var}(y \mid x, z, d = 1) = \sigma^2 - \theta^2 \lambda(c) [\lambda(c) + c]$ ,  $c = z' \delta$ .
- $\hat{\lambda}$  is a **generated regressor**: OLS standard errors are wrong.
- Free test: under  $H_0 : \rho = 0$ , the usual  $t$ -test on  $\hat{\lambda}$  is valid. (Same as  $\hat{v}$  in Part I.)

## (1) Method A: Heckman's two-step ("Heckit")

**Step 1.** Probit of  $d$  on  $z$ , using the full sample:

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

**Step 2.** OLS of  $y$  on  $(x, \hat{\lambda})$ , using workers only ( $d_i = 1$ ):

$$y_i = x_i' \beta + \theta \hat{\lambda}_i + \text{error}_i.$$

### Inference

- The error is **heteroskedastic**:  $\text{Var}(y \mid x, z, d = 1) = \sigma^2 - \theta^2 \lambda(c) [\lambda(c) + c]$ ,  $c = z' \delta$ .
- $\hat{\lambda}$  is a **generated regressor**: OLS standard errors are wrong.
- Free test: under  $H_0 : \rho = 0$ , the usual  $t$ -test on  $\hat{\lambda}$  is valid. (Same as  $\hat{v}$  in Part I.)

**Recover**  $\sigma, \rho$ :  $\hat{\sigma}^2 = \frac{1}{n_1} \sum_{d_i=1} [\hat{e}_i^2 + \hat{\theta}^2 \hat{\lambda}_i (\hat{\lambda}_i + z_i' \hat{\delta})]$ ,  $\hat{\rho} = \hat{\theta} / \hat{\sigma}$ .

## (1) Method B: full maximum likelihood

The two kinds of observations contribute differently:

$$d_i = 0 : \quad \mathbb{P}(d = 0 \mid z) = 1 - \Phi(z_i' \delta)$$

$$d_i = 1 : \quad f(y \mid x) \cdot \mathbb{P}(d = 1 \mid y, x, z) = \frac{1}{\sigma} \phi\left(\frac{y_i - x_i' \beta}{\sigma}\right) \Phi\left(\frac{z_i' \delta + \frac{\rho}{\sigma}(y_i - x_i' \beta)}{\sqrt{1 - \rho^2}}\right)$$

## (1) Method B: full maximum likelihood

The two kinds of observations contribute differently:

$$d_i = 0 : \quad \mathbb{P}(d = 0 \mid z) = 1 - \Phi(z'_i \delta)$$

$$d_i = 1 : \quad f(y \mid x) \cdot \mathbb{P}(d = 1 \mid y, x, z) = \frac{1}{\sigma} \phi\left(\frac{y_i - x'_i \beta}{\sigma}\right) \Phi\left(\frac{z'_i \delta + \frac{\rho}{\sigma}(y_i - x'_i \beta)}{\sqrt{1 - \rho^2}}\right)$$

$$\ell = \sum_{d_i=0} \log [1 - \Phi(z'_i \delta)] + \sum_{d_i=1} \left\{ \log \Phi\left(\frac{z'_i \delta + \frac{\rho}{\sigma}(y_i - x'_i \beta)}{\sqrt{1 - \rho^2}}\right) - \log \sigma + \log \phi\left(\frac{y_i - x'_i \beta}{\sigma}\right) \right\}$$

## (1) Method B: full maximum likelihood

The two kinds of observations contribute differently:

$$d_i = 0 : \quad \mathbb{P}(d = 0 \mid z) = 1 - \Phi(z'_i \delta)$$

$$d_i = 1 : \quad f(y \mid x) \cdot \mathbb{P}(d = 1 \mid y, x, z) = \frac{1}{\sigma} \phi\left(\frac{y_i - x'_i \beta}{\sigma}\right) \Phi\left(\frac{z'_i \delta + \frac{\rho}{\sigma}(y_i - x'_i \beta)}{\sqrt{1 - \rho^2}}\right)$$

$$\ell = \sum_{d_i=0} \log [1 - \Phi(z'_i \delta)] + \sum_{d_i=1} \left\{ \log \Phi\left(\frac{z'_i \delta + \frac{\rho}{\sigma}(y_i - x'_i \beta)}{\sqrt{1 - \rho^2}}\right) - \log \sigma + \log \phi\left(\frac{y_i - x'_i \beta}{\sigma}\right) \right\}$$

### MLE

Efficient if the model is right; joint normality is used everywhere.

### Two-step

Needs only normal  $u_2$  and linear  $\mathbb{E}[u_1 \mid u_2]$ .  
Simpler and more robust, but less efficient.

Parallel with Part I: the  $d = 1$  term is a *probit with a control* times a *normal density*.

## (2) The assumption: an exclusion restriction

### Assumption (exclusion + relevance)

$z$  contains at least one variable that

- |                                                                |             |
|----------------------------------------------------------------|-------------|
| ① does not affect the wage offer: it is excluded from $x$      | (exclusion) |
| ② shifts participation: its coefficient in $\delta$ is nonzero | (relevance) |

## (2) The assumption: an exclusion restriction

### Assumption (exclusion + relevance)

$z$  contains at least one variable that

- ① does not affect the wage offer: it is excluded from  $x$  (exclusion)
- ② shifts participation: its coefficient in  $\delta$  is nonzero (relevance)

**Classic choices:** number of young children, and **non-labor income**. Non-labor income raises the reservation wage, so it shifts participation; an employer's wage offer does not depend on it.

## (2) The assumption: an exclusion restriction

### Assumption (exclusion + relevance)

$z$  contains at least one variable that

- ① does not affect the wage offer: it is excluded from  $x$  (exclusion)
- ② shifts participation: its coefficient in  $\delta$  is nonzero (relevance)

**Classic choices:** number of young children, and **non-labor income**. Non-labor income raises the reservation wage, so it shifts participation; an employer's wage offer does not depend on it.

### Notice the reversal

In Part I, non-labor income was the *endogenous regressor* that needed an instrument. In Part II, the same variable *is* the instrument.

Exogeneity is a property of a variable *in an equation*, relative to a question, not of the variable itself.

## (2) Showing that $\beta$ is identified

**Step 1: the selection equation.**  $\mathbb{P}(d = 1 \mid z) = \Phi(z'\delta)$  identifies  $\delta$ , using all observations. Hence  $\lambda(z'\delta)$  is a known function of observables.

## (2) Showing that $\beta$ is identified

**Step 1: the selection equation.**  $\mathbb{P}(d = 1 \mid z) = \Phi(z'\delta)$  identifies  $\delta$ , using all observations. Hence  $\lambda(z'\delta)$  is a known function of observables.

**Step 2: the selected sample.**  $\mathbb{E}[y \mid x, z, d = 1] = x'\beta + \theta \lambda(z'\delta)$ , so  $(\beta, \theta)$  are identified iff  $\mathbb{E}[(x, \lambda)(x, \lambda)' \mid d = 1]$  has full rank.

## (2) Showing that $\beta$ is identified

**Step 1: the selection equation.**  $\mathbb{P}(d = 1 \mid z) = \Phi(z'\delta)$  identifies  $\delta$ , using all observations. Hence  $\lambda(z'\delta)$  is a known function of observables.

**Step 2: the selected sample.**  $\mathbb{E}[y \mid x, z, d = 1] = x'\beta + \theta \lambda(z'\delta)$ , so  $(\beta, \theta)$  are identified iff  $\mathbb{E}[(x, \lambda)(x, \lambda)' \mid d = 1]$  has full rank.

**Step 3: rule out collinearity.** With an excluded  $z_k$  and  $\delta_k \neq 0$ ,  $\lambda(z'\delta)$  moves while  $x$  is held fixed, so  $\lambda$  is not a linear function of  $x$ .

## (2) Showing that $\beta$ is identified

**Step 1: the selection equation.**  $\mathbb{P}(d = 1 \mid z) = \Phi(z'\delta)$  identifies  $\delta$ , using all observations. Hence  $\lambda(z'\delta)$  is a known function of observables.

**Step 2: the selected sample.**  $\mathbb{E}[y \mid x, z, d = 1] = x'\beta + \theta \lambda(z'\delta)$ , so  $(\beta, \theta)$  are identified iff  $\mathbb{E}[(x, \lambda)(x, \lambda)' \mid d = 1]$  has full rank.

**Step 3: rule out collinearity.** With an excluded  $z_k$  and  $\delta_k \neq 0$ ,  $\lambda(z'\delta)$  moves while  $x$  is held fixed, so  $\lambda$  is not a linear function of  $x$ .

**Step 4: recover  $(\sigma, \rho)$ .**

$$\text{Var}(y \mid x, z, d = 1) = \sigma^2 - \theta^2 \lambda(z'\delta) [\lambda(z'\delta) + z'\delta]$$

identifies  $\sigma^2$ , and then  $\rho = \theta/\sigma$ . ■

## (2) Showing that $\beta$ is identified

**Step 1: the selection equation.**  $\mathbb{P}(d = 1 \mid z) = \Phi(z'\delta)$  identifies  $\delta$ , using all observations. Hence  $\lambda(z'\delta)$  is a known function of observables.

**Step 2: the selected sample.**  $\mathbb{E}[y \mid x, z, d = 1] = x'\beta + \theta \lambda(z'\delta)$ , so  $(\beta, \theta)$  are identified iff  $\mathbb{E}[(x, \lambda)(x, \lambda)' \mid d = 1]$  has full rank.

**Step 3: rule out collinearity.** With an excluded  $z_k$  and  $\delta_k \neq 0$ ,  $\lambda(z'\delta)$  moves while  $x$  is held fixed, so  $\lambda$  is not a linear function of  $x$ .

**Step 4: recover  $(\sigma, \rho)$ .**

$$\text{Var}(y \mid x, z, d = 1) = \sigma^2 - \theta^2 \lambda(z'\delta) [\lambda(z'\delta) + z'\delta]$$

identifies  $\sigma^2$ , and then  $\rho = \theta/\sigma$ . ■

### Intuition

The instrument changes *who* works without changing *what they are offered*. How observed wages respond reveals the selection term.

### (3) When $x = z$ : identification by functional form

Now  $\mathbb{E}[y \mid x, d = 1] = x'\beta + \theta \lambda(x'\delta)$ .

**Unlike Part I**,  $\lambda$  is *nonlinear*, so  $(x, \lambda(x'\delta))$  is not exactly collinear. Under exact normality,  $\beta$  is still identified.

### (3) When $x = z$ : identification by functional form

Now  $\mathbb{E}[y \mid x, d = 1] = x'\beta + \theta \lambda(x'\delta)$ .

**Unlike Part I**,  $\lambda$  is *nonlinear*, so  $(x, \lambda(x'\delta))$  is not exactly collinear. Under exact normality,  $\beta$  is still identified.

**But** over the range where most data live,  $\lambda$  is nearly linear. For  $c \in [-1, 1.5]$  (participation rates 16% to 93%),

$$\text{corr}(c, \lambda(c)) = -0.99.$$

### (3) When $x = z$ : identification by functional form

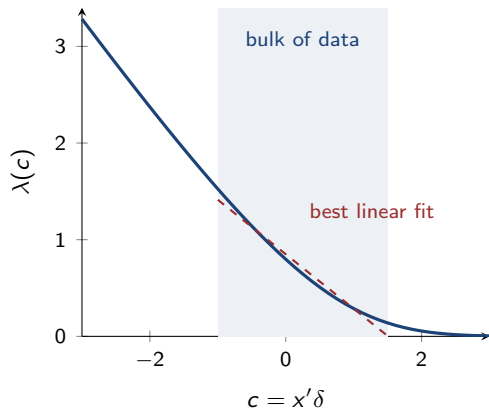
Now  $\mathbb{E}[y \mid x, d = 1] = x'\beta + \theta \lambda(x'\delta)$ .

**Unlike Part I**,  $\lambda$  is *nonlinear*, so  $(x, \lambda(x'\delta))$  is not exactly collinear. Under exact normality,  $\beta$  is still identified.

**But** over the range where most data live,  $\lambda$  is nearly linear. For  $c \in [-1, 1.5]$  (participation rates 16% to 93%),

$$\text{corr}(c, \lambda(c)) = -0.99.$$

Result: severe multicollinearity, huge standard errors, and estimates driven by the tails, which is exactly where normality is least credible.



### (3) What is identified without leaning on normality?

Reduced-form objects that are always identified

- $\mathbb{P}(d = 1 \mid x) = \Phi(x'\delta)$ , so  $\delta$  (participation equation)
- $\mathbb{E}[y \mid x, d = 1]$  (the wage regression *among workers*)

### (3) What is identified without leaning on normality?

#### Reduced-form objects that are always identified

- $\mathbb{P}(d = 1 \mid x) = \Phi(x'\delta)$ , so  $\delta$  (participation equation)
- $\mathbb{E}[y \mid x, d = 1]$  (the wage regression *among workers*)

Drop normality and let the selection term be an unknown function  $g$ :

$$\mathbb{E}[y \mid x, d = 1] = x'\beta + g(x'\delta).$$

For any constant  $c$ , the pair  $(\beta + c\delta, g(t) - c t)$  fits the data equally well:

$\beta$  is identified only up to  $\beta + c\delta$

Only the part of  $\beta$  not in the direction of  $\delta$  (e.g.  $\beta_k - \beta_1\delta_k/\delta_1$ ) is learned.

### (3) What is identified without leaning on normality?

#### Reduced-form objects that are always identified

- $\mathbb{P}(d = 1 \mid x) = \Phi(x'\delta)$ , so  $\delta$  (participation equation)
- $\mathbb{E}[y \mid x, d = 1]$  (the wage regression *among workers*)

Drop normality and let the selection term be an unknown function  $g$ :

$$\mathbb{E}[y \mid x, d = 1] = x'\beta + g(x'\delta).$$

For any constant  $c$ , the pair  $(\beta + c\delta, g(t) - c t)$  fits the data equally well:

$\beta \text{ is identified only up to } \beta + c\delta$

Only the part of  $\beta$  not in the direction of  $\delta$  (e.g.  $\beta_k - \beta_1\delta_k/\delta_1$ ) is learned.

#### The parallel with Part I

Part I without exclusion: only  $\gamma + \rho/\sigma_w$  is identified. Part II without exclusion: only  $\beta$  modulo the direction  $\delta$ . In both, the causal and the selection components cannot be separated.

## (4) Stata

Same mroz data:  $lwage = y$ ,  $inlf = d$ ; excluded:  $nwifeinc$ ,  $age$ ,  $kidslt6$ ,  $kidsge6$ .

```
// (a) Full MLE (default): reports rho, sigma, lambda = rho*sigma, LR test of rho = 0
heckman lwage educ exper expersq, ///
        select(inlf = educ exper expersq nwifeinc age kidslt6 kidsge6)

// (b) Heckman two-step with corrected standard errors
heckman lwage educ exper expersq, ///
        select(inlf = educ exper expersq nwifeinc age kidslt6 kidsge6) twostep

// (c) By hand: point estimates only
probit inlf educ exper expersq nwifeinc age kidslt6 kidsge6
predict zd, xb
generate mills = normalden(zd)/normal(zd)
regress lwage educ exper expersq mills if inlf == 1

// Postestimation after (a) or (b)
predict yhat, xb          // E[y|x]: the wage offer
predict ycond, ycond      // E[y|x, d=1]: observed wages
```

Related: `heckprobit` (binary outcome with selection), `eregress ...`, `select()` (Stata 15+).

## Two problems, one structure

|                             | Part I: endogenous probit                             | Part II: Heckman selection                 |
|-----------------------------|-------------------------------------------------------|--------------------------------------------|
| <b>What is endogenous?</b>  | A regressor, $w$                                      | The sample, $d = 1$                        |
| <b>First stage</b>          | Linear regression of $w$                              | Probit of $d$                              |
| <b>Control function</b>     | $\rho v$                                              | $\rho\sigma\lambda(z'\delta)$              |
| <b>Second stage</b>         | Probit with $\hat{v}$                                 | OLS with $\hat{\lambda}$                   |
| <b>Exclusion</b>            | $z \in x_w, z \notin x_y$                             | $z$ in selection, not in $x$               |
| <b>Without exclusion</b>    | Exactly not identified; only $\gamma + \rho/\sigma_w$ | Only through the nonlinearity of $\lambda$ |
| <b>Free exogeneity test</b> | $t$ -test on $\hat{v}$                                | $t$ -test on $\hat{\lambda}$               |
| <b>Stata</b>                | ivprobit                                              | heckman                                    |

### The common recipe

**Specification & Estimand:** write down the joint normal model. **Identification:** condition on the first-stage residual, then find variation that moves the endogenous object but not the outcome.

**Estimation:** two-step or MLE. **Inference:** correct for generated regressors.

## Takeaways

- ① Under joint normality, endogeneity and selection both become **omitted-variable problems**, and the omitted variable is a control function:  $v$  or  $\lambda$ .
- ② Identification comes from an **exclusion restriction**, not from the probit or the normal distribution.
- ③ Without an exclusion restriction, Part I fails *exactly*, while Part II hangs on the curvature of  $\lambda$ : a thread, not an identification strategy.
- ④ Two-step estimators are easy and transparent but need corrected standard errors; MLE is efficient but leans harder on normality.
- ⑤ The same variable can be endogenous in one question and an instrument in another.

**A control function  
is only as good as  
its excluded variable.**

## Selected references

- Blundell, R., and J. L. Powell (2004). "Endogeneity in Semiparametric Binary Response Models." *Review of Economic Studies*, 71(3), 655–679.
- Gronau, R. (1974). "Wage Comparisons—A Selectivity Bias." *Journal of Political Economy*, 82(6), 1119–1143.
- Heckman, J. J. (1979). "Sample Selection Bias as a Specification Error." *Econometrica*, 47(1), 153–161.
- Lewbel, A. (2019). "The Identification Zoo: Meanings of Identification in Econometrics." *Journal of Economic Literature*, 57(4), 835–903.
- Mroz, T. A. (1987). "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions." *Econometrica*, 55(4), 765–799.
- Newey, W. K. (1987). "Efficient Estimation of Limited Dependent Variable Models with Endogenous Explanatory Variables." *Journal of Econometrics*, 36(3), 231–250.
- Rivers, D., and Q. H. Vuong (1988). "Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models." *Journal of Econometrics*, 39(3), 347–366.
- Smith, R. J., and R. W. Blundell (1986). "An Exogeneity Test for a Simultaneous Equation Tobit Model with an Application to Labor Supply." *Econometrica*, 54(3), 679–685.
- Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*, 2nd ed. MIT Press, Chapters 15 and 19.