Part 2: Simultaneous Equation Models (SEM)

Chapter 17. Endogeneity and Instrumental Variables

Chapter 18. Why Should We Concern SEM ?

Chapter 19. What is the Identification Problem ?

Chapter 20. How to Estimate SEM ?

Chapter 17. Endogeneity and Instumental Variables

17.1 Definition and source of endogeneity

17.2 Estimation problem with endogeneity

17.3 IV and choices

17.4 Two-stage least squares method

17.5 Testing instrument validity

17.6 Testing regressor endogeneity

17.1 Definition and sources of endogeneity

Review: the CLRM assumptions

Let's revise the classic linear regression model assumptions(CLRM):

- **A1**: The true model is $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.
- **A2**: $\forall i, E(\epsilon_i|X) = 0$ (conditional zero mean) more about this later.
- **A3**: $Var(\epsilon|X) = E(\epsilon\epsilon'|X) = \sigma^2 I$ (identical conditional variance).
- **A4**: $\boldsymbol{X}$ has full column rank.
- **A5**: (for inference purposes, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\boldsymbol{I})$).

Under **A1-A4**（also namely **CLRM**）, the OLS is **Gauss-Markov** efficient.

Under **A1-A5**, we donote **N-CLRM**.

Review: the A2 assumptions

For the population regression model(PRM):

$$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad \text{(PRM)}$$

- **CLRM A2** assumes that X is fixed (given) or independent of the error term. The regressors X **are not** random variables. At this point, we can use the OLS method and get **BLUE**(Best Linear Unbiased Estimator).

$$Cov(X_i, u_i) = 0; \quad E(X_i u_i) = 0$$

- If the above A2 assumption is violated, the independent variable X is related to the random disturbance term. In this case, OLS estimation will no longer get **BLUE**, and **instrumental variable method** (IV) should be used for estimation.

$$Cov(X_i, u_i) \neq 0; \quad E(X_i u_i) \neq 0$$

Good model: random experiment

Randomized controlled experiment : Ideally, the value of the independent variable X is randomly changed (refer to the **reason**), and then we look at the change in the dependent variable Y (refer to the **result**).

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$$

- If Y_i and X_i does exist systematic relationship (linear relationship), then change X_i causes the corresponding change of Y_i.
- Any other random factors will be added to the random disturbance u_i. The effect of the random disturbance on the change of Y_i should be **independent** to the effect of X_i on the change of Y_i.

Good model: exogenous regressors

Exogenous regressors: If independent variables X_i is **perfectly random** (randomly assigned) as mentioned above , then they are called **exogenous regressor**. More precisely, they can be defined as:

Strictly Exogeneity:

$$E(u_i \mid X_1, \ldots, X_N) = E(u_i \mid \mathbf{x}) = 0$$

Endogeneity: definition

We use the term **endogeneity** frequently in econometrics.

Also this concept is used broadly to describe any situation where a regressor is **correlated** with the error term.

- Assume that we have the bivariate linear model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- The explanatory variable X is said to be **Endogenous** if it is correlated with ϵ.

$$Cov(X_i, \epsilon_i) \neq 0; \quad E(X_i \epsilon_i) \neq 0$$

- And if X is **uncorrelated** with ϵ, it is said to be **Exogenous**.

$$Cov(X_i, \epsilon_i) = 0; \quad E(X_i \epsilon_i) = 0$$

Endogeneity: sources

In applied econometrics, endogeneity usually arises in one of four ways:

- **Omitted variables**: when the model is set incorrectly.
- **Measurement errors** in the regressors.
- **Autocorrelation** of the error term in autoregressive models.
- **Simultaneity**: when $\boldsymbol{Y}$ and $\boldsymbol{X}$ are simultaneously determined, as in the supply/demand model (we will go to explain it in the next three chapter).

Source 1: Omitted variables

Suppose that the **"assumed true model"** for wage determination is:

$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i \quad \text{(the assumed true model)}$$

However, because the individual's **ability variable**(Abl) is often not directly observed, so we often can't put it into the model, and build a **mis-specified model**.

$$Wage_i = \alpha_1 + \alpha_2 Edu_i + v_i \quad \text{(the mis-specified model)}$$

- Where **ability variable**(Abl) is included in the new disturbance v_i, and $v_i = \beta_3 abl_i + u_i$
- Obviously, in the mis-specified model, we ignore the **ability variable**(Abl), while variable **years of education**(Edu) is actually related to it.
- So in the mis-specified model, $cov(Edu_i, v_i) \neq 0$, thus **years of education**(Edu) may cause the endogenous problem.

Source A: Omitted variables (demo1) 1/4

Here we show a visual demonstration on this situation:

Assumed "TRUE MODEL":

$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i$$

Source A: Omitted variables (demo1) 2/4

Here we show a visual demonstration on this situation:

Assumed "TRUE MODEL":

$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i$$

"A's mis-specification MODEL":

$$Wage_i = \alpha_1 + \alpha_2 Edu_i + v_i$$

Source A: Omitted variables (demo1) 3/4

Here we show a visual demonstration on this situation:

Assumed "TRUE MODEL":

$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i$$

"A's mis-specification MODEL":

$$Wage_i = \alpha_1 + \alpha_2 Edu_i + v_i$$

Source A: Omitted variables (demo1) 4/4

Here we show a visual demonstration on this situation:

Assumed "TRUE MODEL":

$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i$$

"A's mis-specification MODEL":

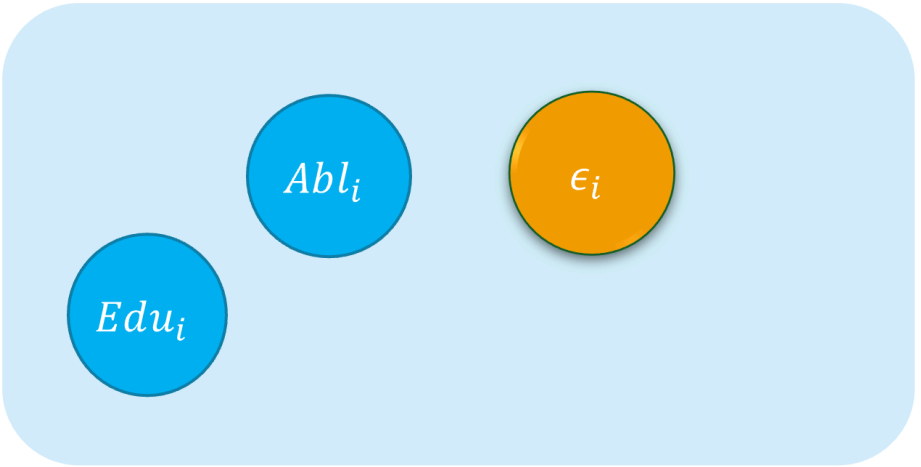
$$Wage_i = \alpha_1 + \alpha_2 Edu_i + v_i$$

Abl_i

Omitted `≠` Disappeared

Source 1： Omitted variables (demo2)

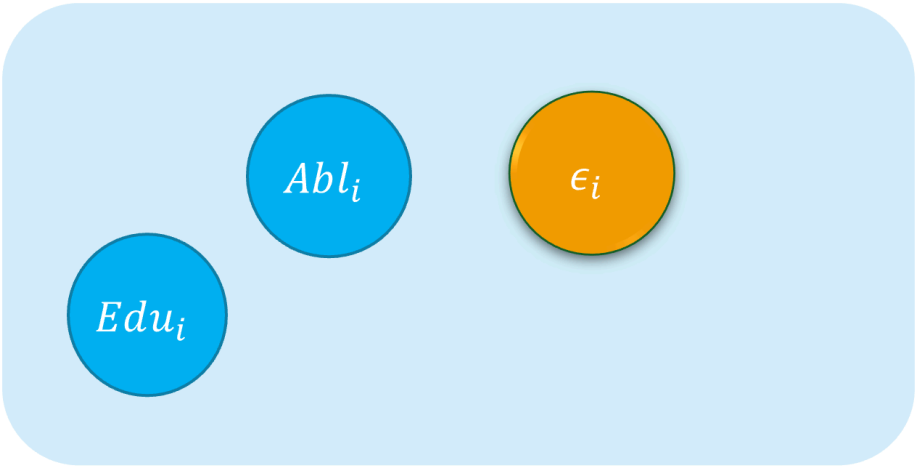
An intuitive demonstration is show as follows:

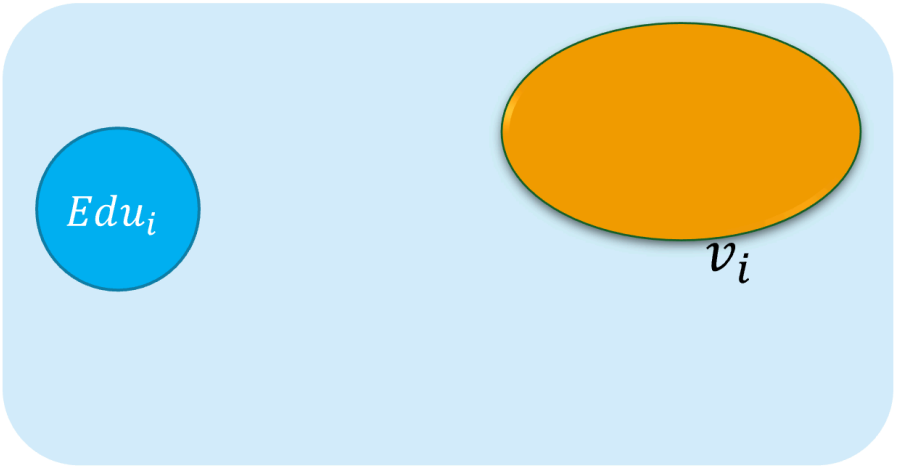


$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i$$

Source 1： Omitted variables (demo2)

An intuitive demonstration is show as follows:



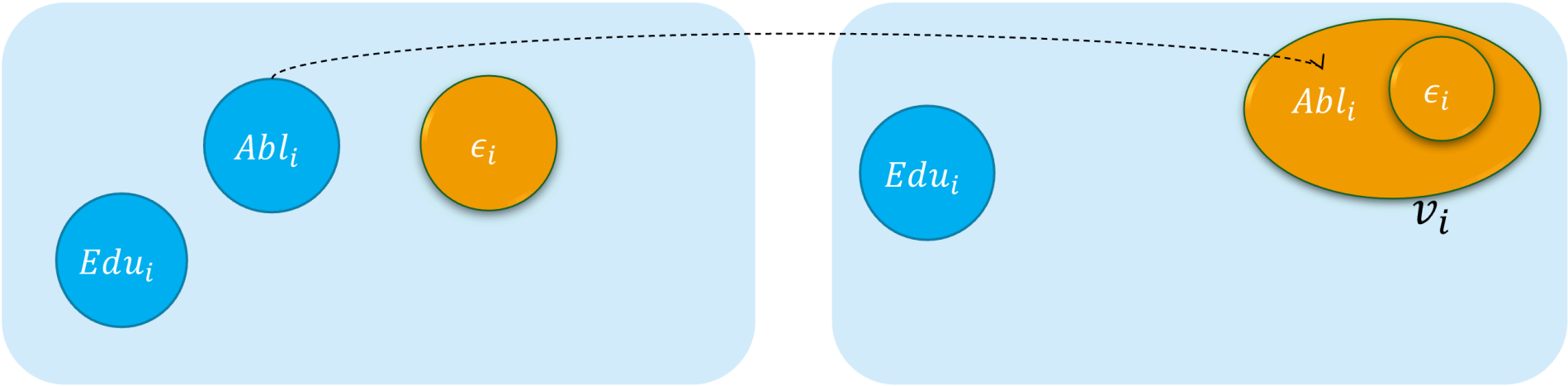


$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i$$

$$Wage_i = \alpha_1 + \alpha_2 Edu_i + v_i$$

Source 1： Omitted variables (demo2)

An intuitive demonstration is show as follows:



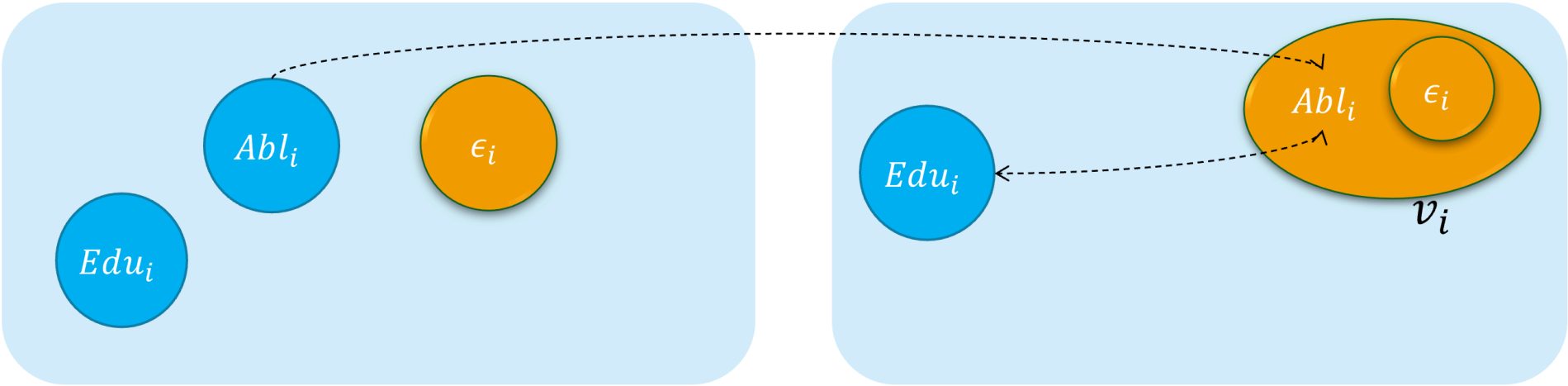
$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i$$

$$Wage_i = \alpha_1 + \alpha_2 Edu_i + v_i$$

$$v_i = \beta_3 abl_i + \epsilon_i$$

Source 1： Omitted variables (demo2)

An intuitive demonstration is show as follows:



$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i$$

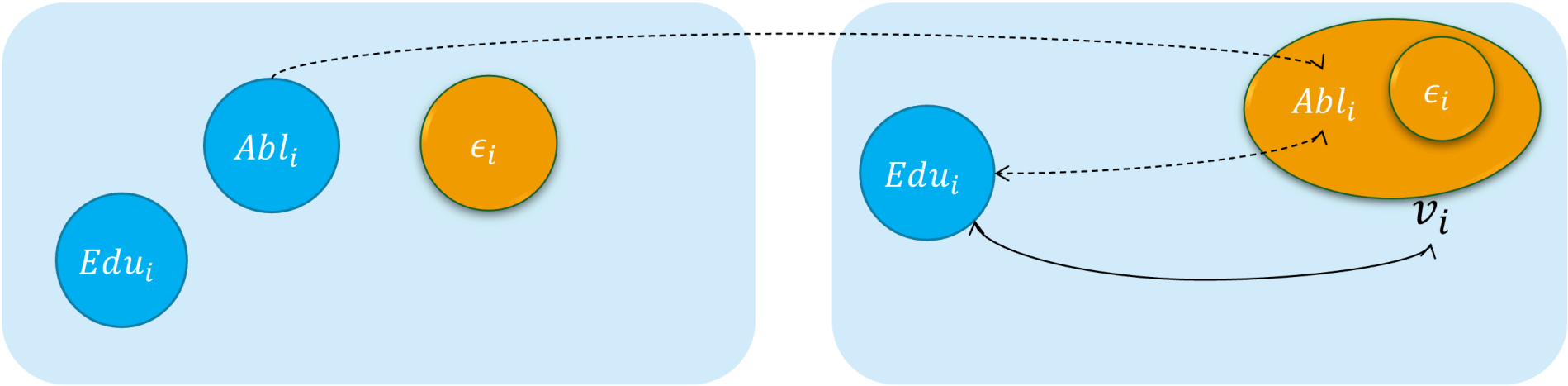
$$Wage_i = \alpha_1 + \alpha_2 Edu_i + v_i$$

$$v_i = \beta_3 abl_i + \epsilon_i$$

$$Cov(Edu_i, Abl_i) \neq 0$$

Source 1： Omitted variables (demo2)

An intuitive demonstration is show as follows:



$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i$$

$$Wage_i = \alpha_1 + \alpha_2 Edu_i + v_i$$

$$\left.\begin{array}{l} v_i = \beta_3 abl_i + \epsilon_i \\ Cov(Edu_i, Abl_i) \neq 0 \end{array}\right\} \Rightarrow \quad Cov(Edu_i, v_i) \neq 0$$

Source B: Measurement errors

Again, let's consider the **"assumed true model"**:

$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i \quad \text{(the assumed true model)}$$

It is hard to observe individual's **ability variable**(Abl), and somebody will instead to use the variable **IQ score**(IQ_i), and construct the mis-specified **"proxy variable" model**:

$$Wage_i = \alpha_1 + \alpha_2 Edu_i + \alpha_3 IQ_i + v_i \quad \text{(the mis-specified model)}$$

- It should exist stuffs (Abl_other_i) which the model does not include (due to the measurement error). So the measurement errors (Abl_other_i) will go to the disturbance term v_i in the mis-specified model.
- And we know that measurement errors (Abl_other_i) will be correlated with the education variable. Thus $cov(Edu_i, v_i) \neq 0$, and the **education variable**(Edu) may cause the endogenouse problem.

Source B: Measurement errors (demo 1) 1/4

Here we show a visual demonstration on this situation:

Assumed TRUE MODEL:

$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i$$

Source B: Measurement errors (demo 1) 2/4

Here we show a visual demonstration on this situation:

Assumed TRUE MODEL:

$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i$$

B's mis-specification MODEL:

$$Wage_i = \alpha_1 + \alpha_2 Edu_i + \alpha_3 IQ_i + v_i$$

Source B: Measurement errors (demo 1) 3/4

Here we show a visual demonstration on this situation:

Assumed TRUE MODEL:

$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i$$

B's mis-specification MODEL:

$$Wage_i = \alpha_1 + \alpha_2 Edu_i + \alpha_3 IQ_i + v_i$$

$$Abl_i = \{\boldsymbol{IQ_i}, Abl\_other_i\}$$

Source B: Measurement errors (demo 1) 4/4

Here we show a visual demonstration on this situation:

Assumed TRUE MODEL:

$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i$$

B's mis-specification MODEL:

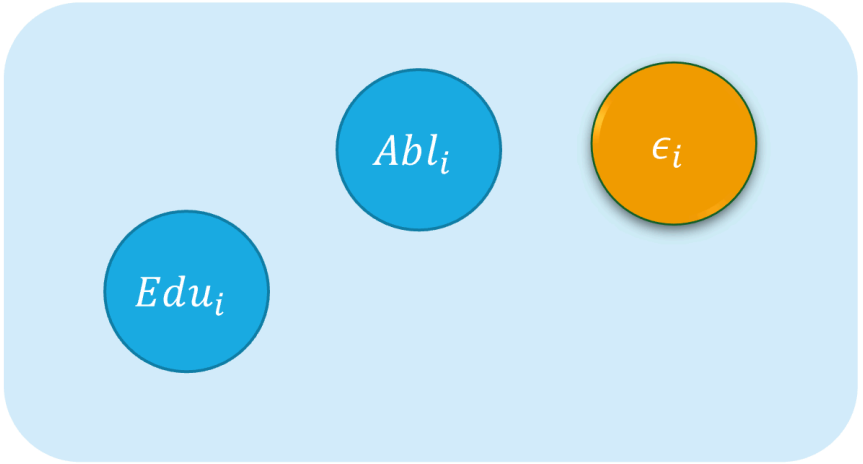
$$Wage_i = \alpha_1 + \alpha_2 Edu_i + \alpha_3 IQ_i + v_i$$

$$Abl_i = \{\boldsymbol{IQ_i}, Abl\_other_i\}$$

Measure Error `≠` Disappeared

Source 2: Measurement errors (demo 2)

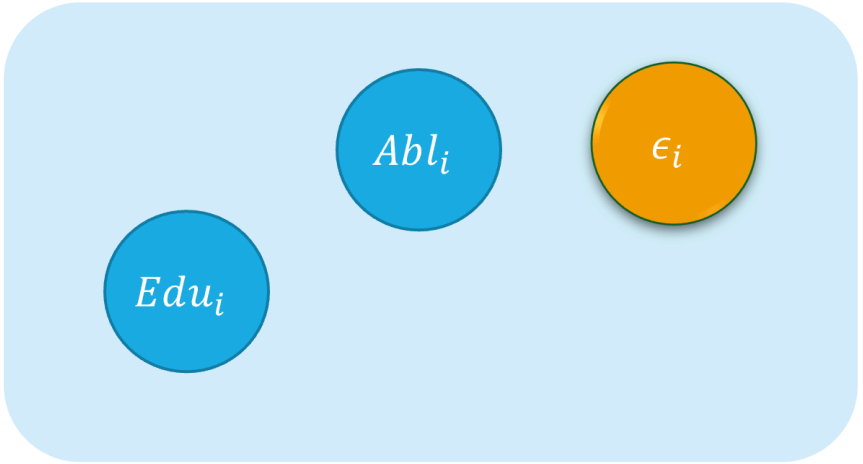
An intuitive demonstration is show as follows:

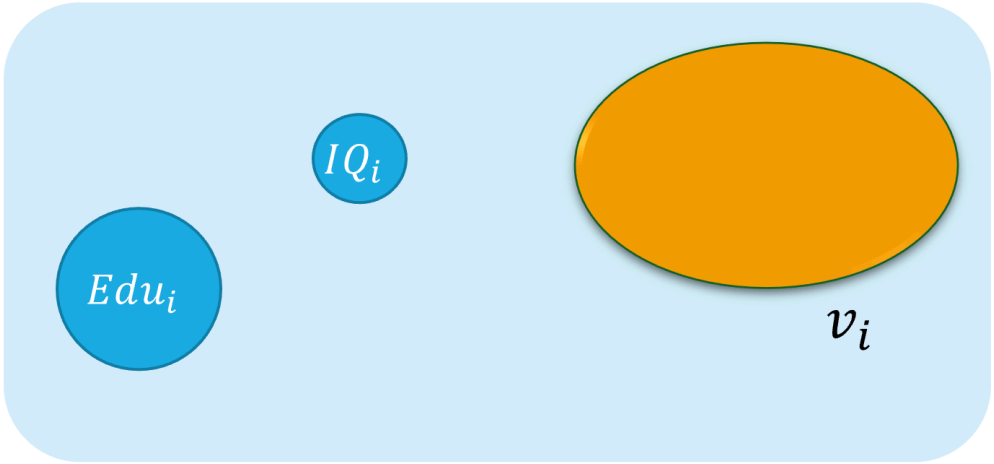


$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i$$

Source 2: Measurement errors (demo 2)

An intuitive demonstration is show as follows:



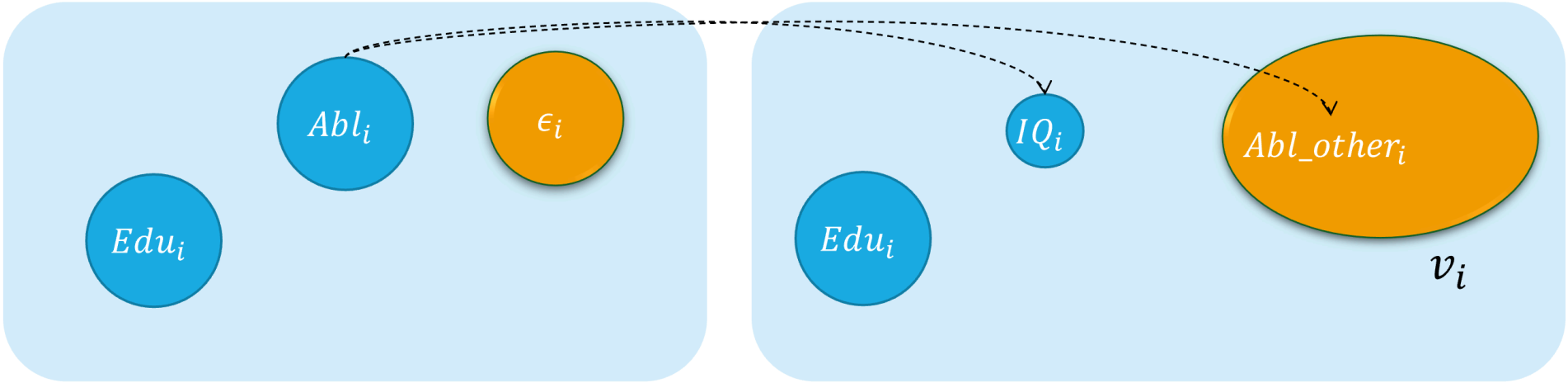


$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i$$

$$Wage_i = \alpha_1 + \alpha_2 Edu_i + \alpha_3 IQ_i + v_i$$

Source 2: Measurement errors (demo 2)

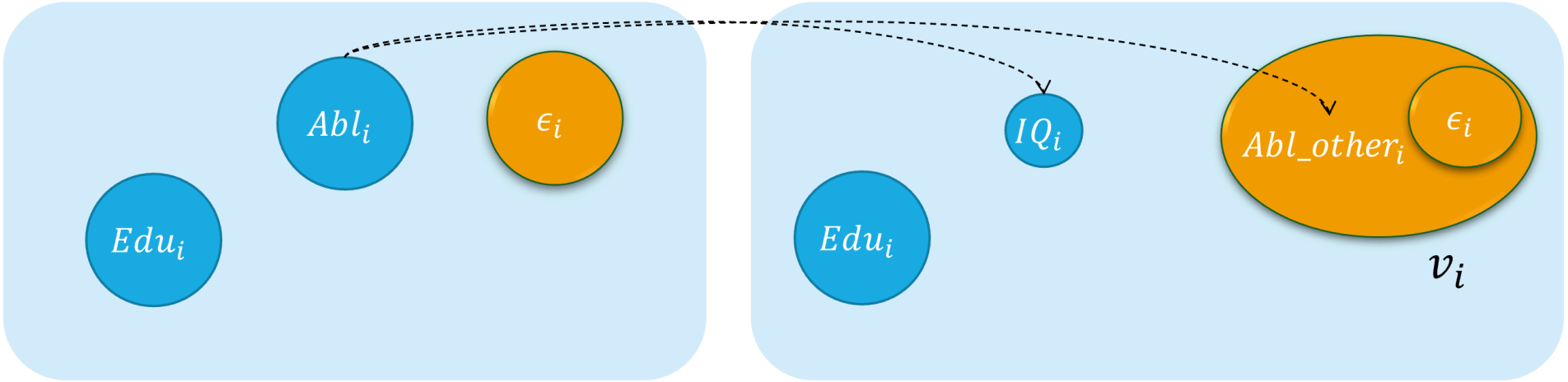
An intuitive demonstration is show as follows:



$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i \qquad Wage_i = \alpha_1 + \alpha_2 Edu_i + \alpha_3 IQ_i + v_i$$

Source 2: Measurement errors (demo 2)

An intuitive demonstration is show as follows:

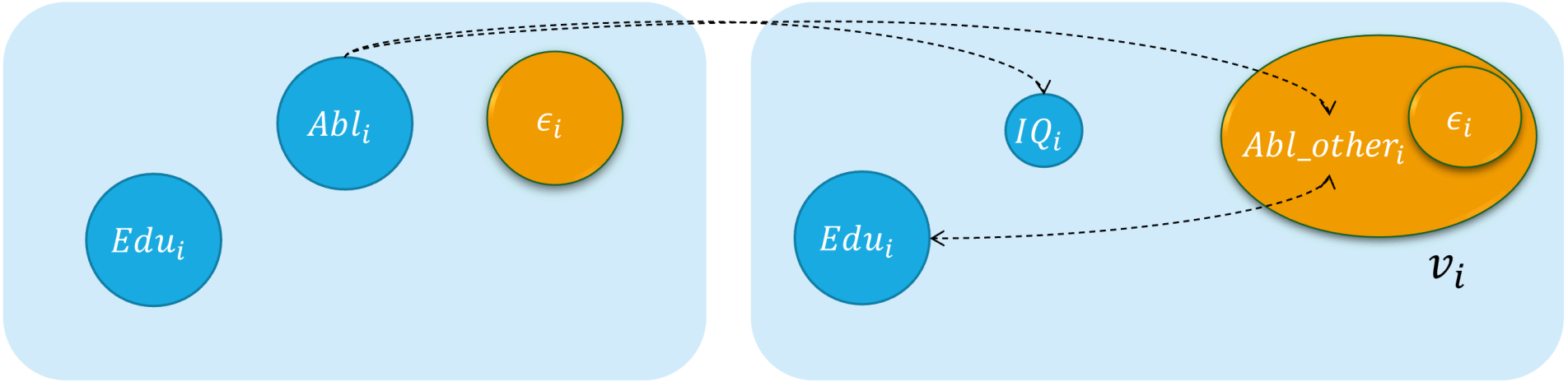


$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i \qquad Wage_i = \alpha_1 + \alpha_2 Edu_i + \alpha_3 IQ_i + v_i$$

$$v_i = \beta_3 Abl\_other_i + \epsilon_i$$

Source 2: Measurement errors (demo 2)

An intuitive demonstration is show as follows:



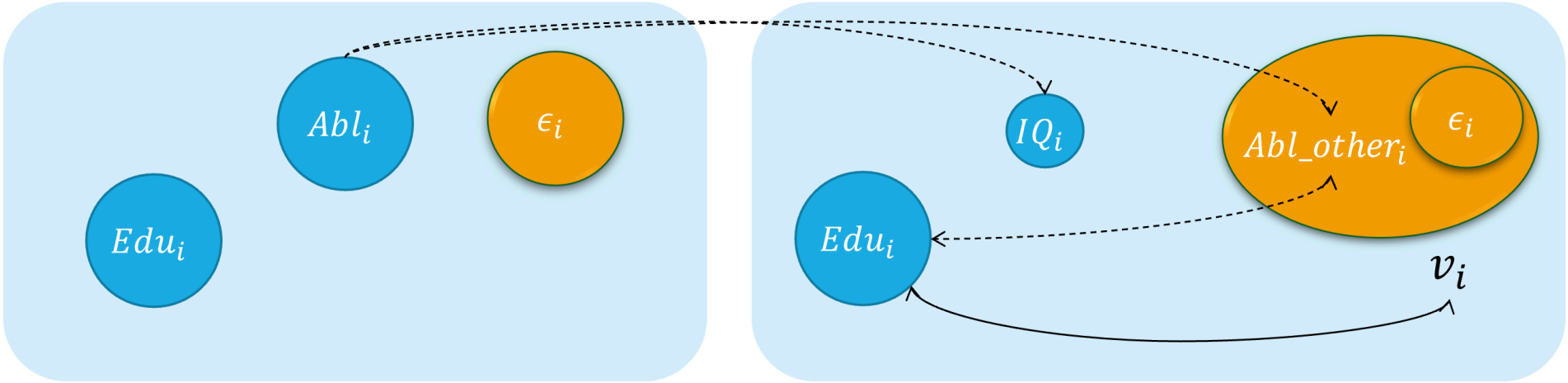
$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i \qquad Wage_i = \alpha_1 + \alpha_2 Edu_i + \alpha_3 IQ_i + v_i$$

$$v_i = \beta_3 Abl\_other_i + \epsilon_i$$

$$Cov(Edu_i, Abl\_other_i) \neq 0$$

Source 2: Measurement errors (demo 2)

An intuitive demonstration is show as follows:



$$Wage_i = \beta_1 + \beta_2 Edu_i + \beta_3 Abl_i + \epsilon_i$$

$$Wage_i = \alpha_1 + \alpha_2 Edu_i + \alpha_3 IQ_i + v_i$$

$$\left.\begin{array}{r} v_i = \beta_3 Abl\_other_i + \epsilon_i \\ \\ Cov(Edu_i, Abl\_other_i) \neq 0 \end{array}\right\} \Rightarrow \quad Cov(Edu_i, v_i) \neq 0$$

Source C: Autocorrelation

Autoregressive model: Lag variable of dependent variable($Y_{t-1}, \ldots, Y_{t-p}, \ldots$) appears in the model as regressors.

$$Y_t = \beta_1 + \beta_2 Y_{t-1} + \beta_3 X_t + u_t$$

If the disturbance term determined following a first-order autocorrelation AR(1):

$$u_t = \rho u_{t-1} + \epsilon_t$$

Then, it is obvious that $cov(Y_{t-1}, u_{t-1}) \neq 0$ and $cov(Y_{t-1}, u_t) \neq 0$.

Thus the **lag dependent variable**(Y_{t-1}) will cause the endogeneity problem in the **Autoregressive model**.

Source C: Autocorrelation (demo 1) 1/4

Here we show a visual demonstration on this situation:

Assumed TRUE MODEL: AR(1)

Source C: Autocorrelation (demo 1) 2/4

Here we show a visual demonstration on this situation:

Assumed TRUE MODEL: AR(1)

$$Y_t = \beta_1 + \beta_2 Y_{t-1} + \beta_3 X_t + u_t \text{ (main model)}$$

$$u_t = \rho u_{t-1} + \epsilon_t \text{ (auxiliary model)}$$

Source C: Autocorrelation (demo 1) 3/4

Here we show a visual demonstration on this situation:

Assumed TRUE MODEL: AR(1)

$$\begin{cases} Y_t = \beta_1 + \beta_2 Y_{t-1} + \beta_3 X_t + u_t \text{ (main model)} \\ u_t = \rho u_{t-1} + \epsilon_t \text{ (auxiliary model)} \end{cases}$$

Source C: Autocorrelation (demo 1) 4/4

Here we show a visual demonstration on this situation:

Assumed TRUE MODEL: AR(1)

$$\begin{cases} Y_t = \beta_1 + \beta_2 Y_{t-1} + \beta_3 X_t + u_t \text{ (main model)} \\ u_t = \rho u_{t-1} + \epsilon_t \text{ (auxiliary model)} \end{cases}$$

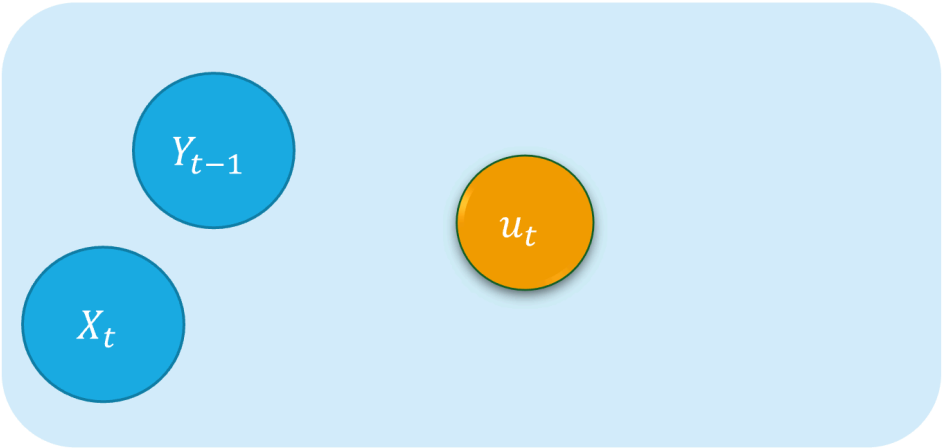
Hidden `≠` Disappeared

Source C: Autocorrelation (demo 2) 1/5

An intuitive demonstration is show as follows:

(main model)

$$Y_t = \beta_1 + \beta_2 Y_{t-1} + \beta_3 X_t + u_t$$



Source C: Autocorrelation (demo 2) 2/5

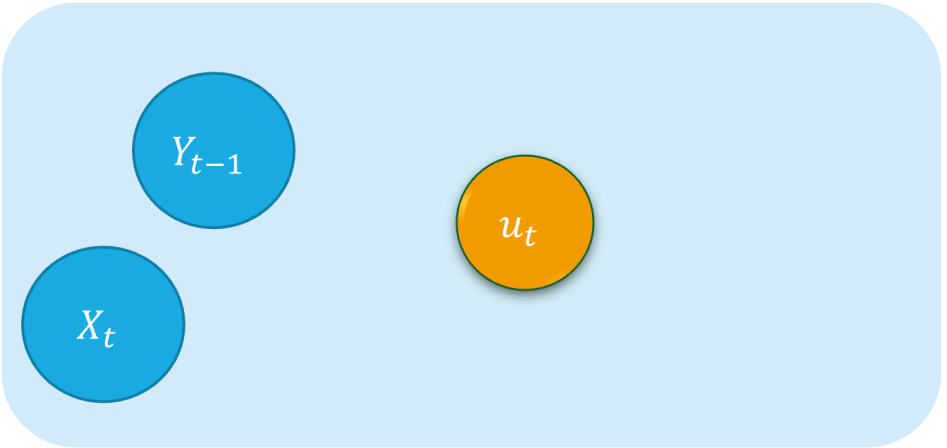
An intuitive demonstration is show as follows:

(main model)

$$Y_t = \beta_1 + \beta_2 Y_{t-1} + \beta_3 X_t + u_t$$

(derived model)

$$Y_{t-1} = \beta_1 + \beta_2 Y_{t-2} + \beta_3 X_{t-1} + u_{t-1}$$



Source C: Autocorrelation (demo 2) 3/5

An intuitive demonstration is show as follows:

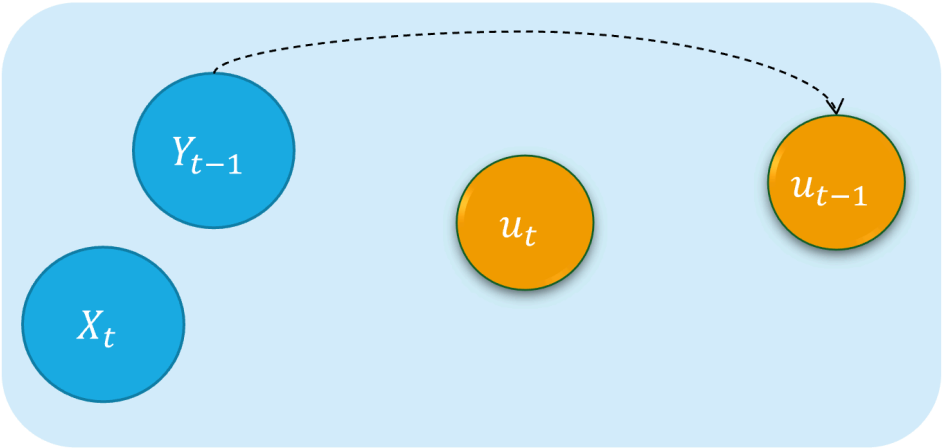
(main model)

$$Y_t = \beta_1 + \beta_2 Y_{t-1} + \beta_3 X_t + u_t$$

(derived model)

$$Y_{t-1} = \beta_1 + \beta_2 Y_{t-2} + \beta_3 X_{t-1} + u_{t-1}$$

$$Cov(Y_{t-1}, u_{t-1}) \neq 0$$



Source C: Autocorrelation (demo 2) 4/5

An intuitive demonstration is show as follows:

(main model)

$$Y_t = \beta_1 + \beta_2 Y_{t-1} + \beta_3 X_t + u_t$$

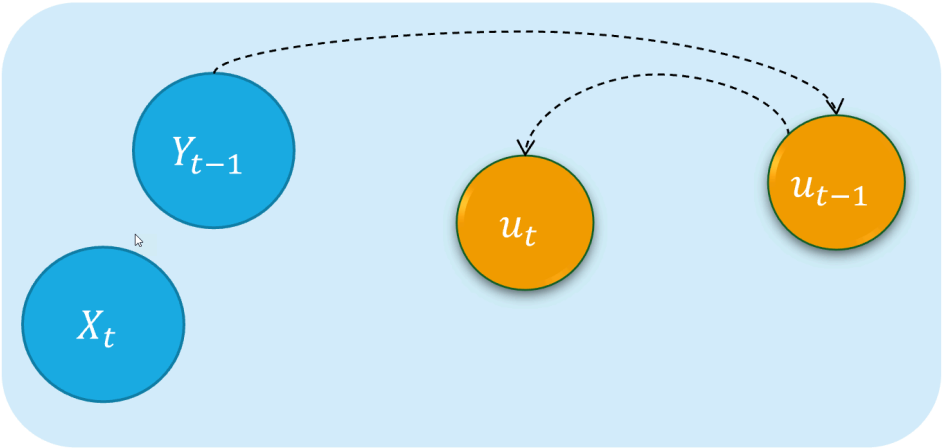
(derived model)

$$Y_{t-1} = \beta_1 + \beta_2 Y_{t-2} + \beta_3 X_{t-1} + u_{t-1}$$

$$Cov(Y_{t-1}, u_{t-1}) \neq 0$$

$$u_t = \rho u_{t-1} + \epsilon_t$$

(auxiliary model)



Source C: Autocorrelation (demo 2) 5/5

An intuitive demonstration is show as follows:

(main model)

$$Y_t = \beta_1 + \beta_2 Y_{t-1} + \beta_3 X_t + u_t$$

(derived model)

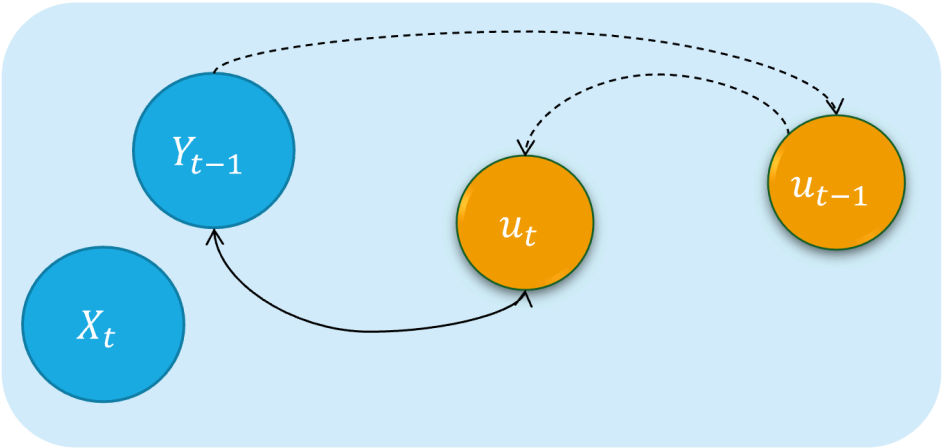
$$Y_{t-1} = \beta_1 + \beta_2 Y_{t-2} + \beta_3 X_{t-1} + u_{t-1}$$

$$Cov(Y_{t-1}, u_{t-1}) \neq 0$$

$$u_t = \rho u_{t-1} + \epsilon_t$$

(auxiliary model)

$$\Rightarrow Cov(Y_{t-1}, u_t) \neq 0$$



Source D: Simultaneity

For the equations system of supply and demand:

$$\begin{cases} \text{Demand: } Q_i = \alpha_1 + \alpha_2 P_i + u_{di} \\ \text{Supply: } Q_i = \beta_1 + \beta_2 P_i + u_{si} \end{cases}$$

As we all know, because of the price P_i will both affect supply and the demand Q_i, And vice versa. There is a feedback cycle mechanism in this system.

So, we can get $cov(P_i, u_{di}) \neq 0$, and $cov(P_i, u_{si}) \neq 0$, which will cause the endogenous problem finally.

Source D: Simultaneity (demo 1) 1/4

Here we show a visual demonstration on this situation:

$$\begin{cases} Q_i = \alpha_1 + \alpha_2 P_i + \alpha_3 \mathrm{I}_i + u_{di} \text{ (Demand } \alpha_2 < 0, \alpha_3 > 0) \end{cases}$$

Source D: Simultaneity (demo 1) 2/4

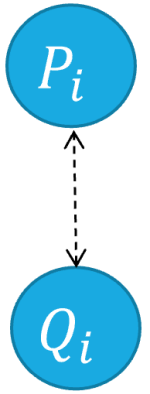
Here we show a visual demonstration on this situation:

$$\begin{cases} Q_i = \alpha_1 + \alpha_2 P_i + \alpha_3 \mathrm{I}_i + u_{di} & \text{(Demand } \alpha_2 < 0, \alpha_3 > 0\text{)} \\ Q_i = \beta_1 + \beta_2 P_i + u_{si} & \text{(Supply } \beta_2 > 0\text{)} \end{cases}$$

Source D: Simultaneity (demo 1) 3/4

Here we show a visual demonstration on this situation:

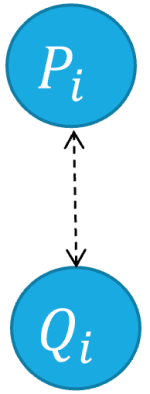
$$\begin{cases} Q_i = \alpha_1 + \alpha_2 P_i + \alpha_3 \mathrm{I}_i + u_{di} & \text{(Demand } \alpha_2 < 0, \alpha_3 > 0) \\ Q_i = \beta_1 + \beta_2 P_i + u_{si} & \text{(Supply } \beta_2 > 0) \end{cases}$$



Source D: Simultaneity (demo 1) 4/4

Here we show a visual demonstration on this situation:

$$\begin{cases} Q_i = \alpha_1 + \alpha_2 P_i + \alpha_3 \mathrm{I}_i + u_{di} & \text{(Demand } \alpha_2 < 0, \alpha_3 > 0) \\ Q_i = \beta_1 + \beta_2 P_i + u_{si} & \text{(Supply } \beta_2 > 0) \end{cases}$$

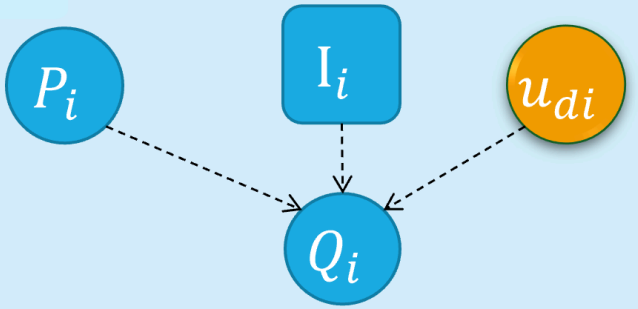


Hidden `≠` Disappeared

Source D: Simultaneity (demo 2) 1/5

An intuitive demonstration is show as follows:

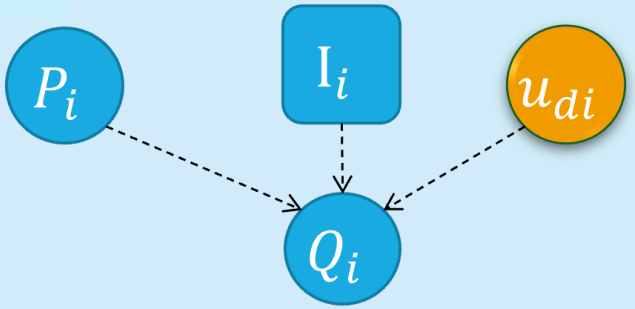
D: $Q_i = \alpha_1 + \alpha_2 P_i + \alpha_3 \mathrm{I}_i + u_{di}$



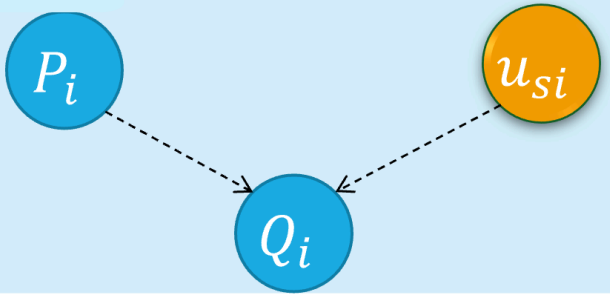
Source D: Simultaneity (demo 2) 2/5

An intuitive demonstration is show as follows:

D: $Q_i = \alpha_1 + \alpha_2 P_i + \alpha_3 \mathrm{I}_i + u_{di}$

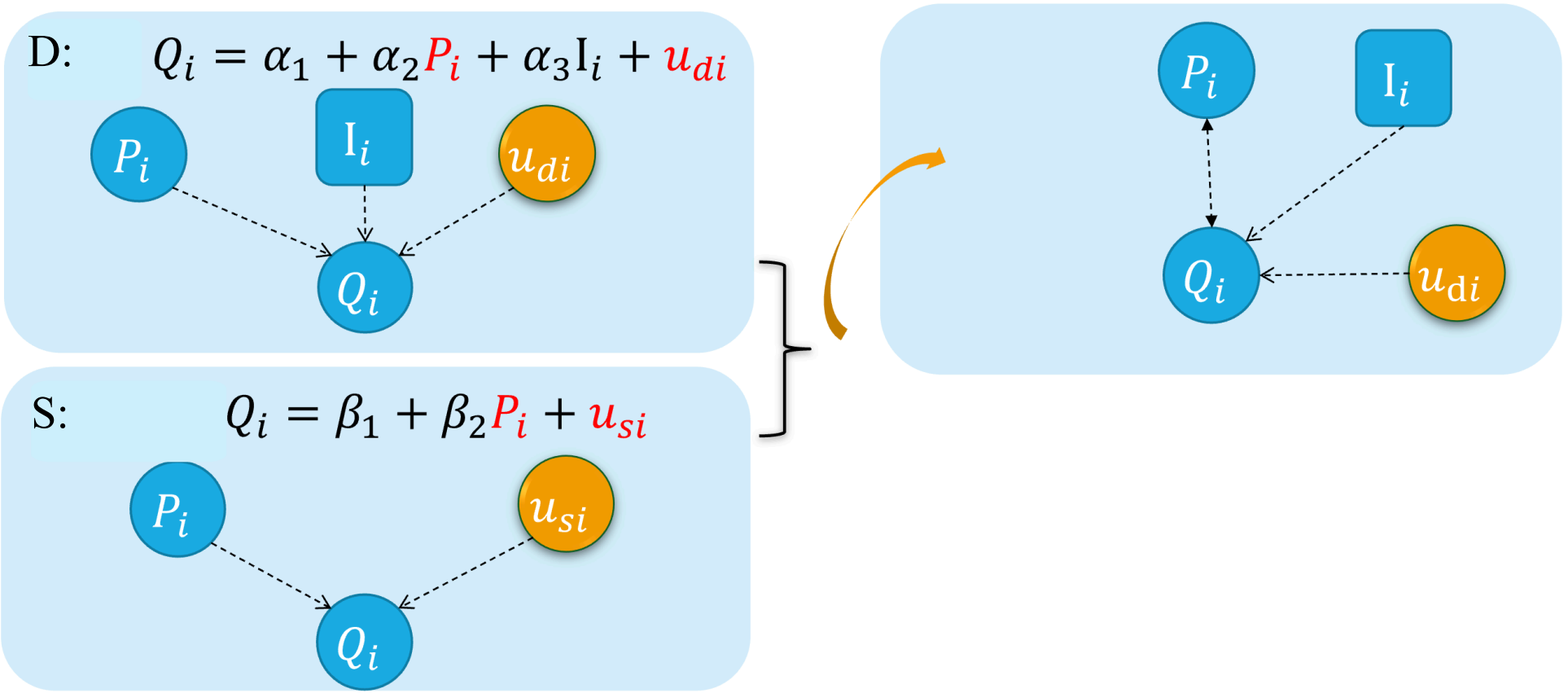


S: $Q_i = \beta_1 + \beta_2 P_i + u_{si}$



Source D: Simultaneity (demo 2) 3/5

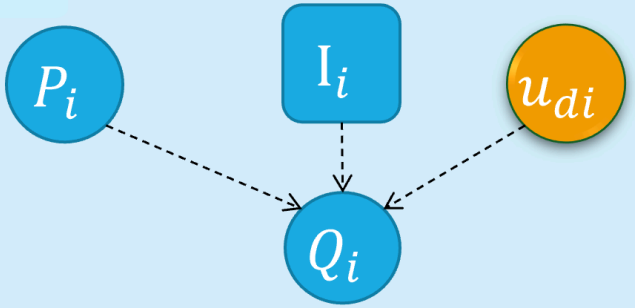
An intuitive demonstration is show as follows:



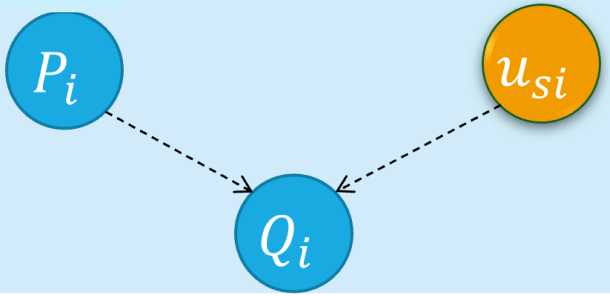
Source D: Simultaneity (demo 2) 4/5

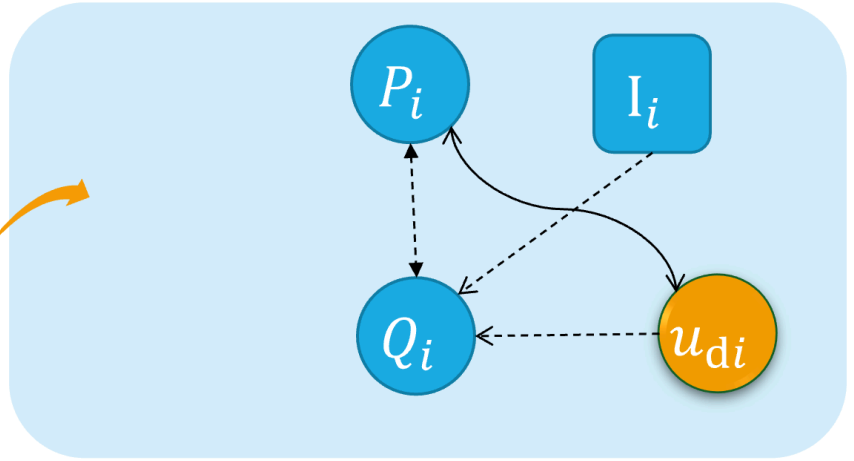
An intuitive demonstration is show as follows:

D: $Q_i = \alpha_1 + \alpha_2 P_i + \alpha_3 \mathrm{I}_i + u_{di}$



S: $Q_i = \beta_1 + \beta_2 P_i + u_{si}$



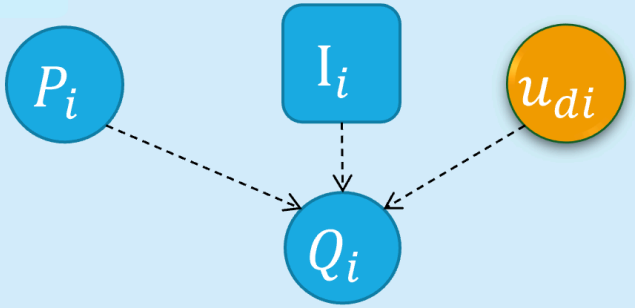


$$\Rightarrow \begin{cases} Cov(P_i, u_{di}) \neq 0 \end{cases}$$

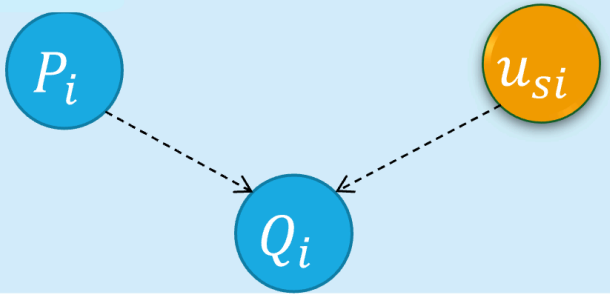
Source D: Simultaneity (demo 2) 5/5

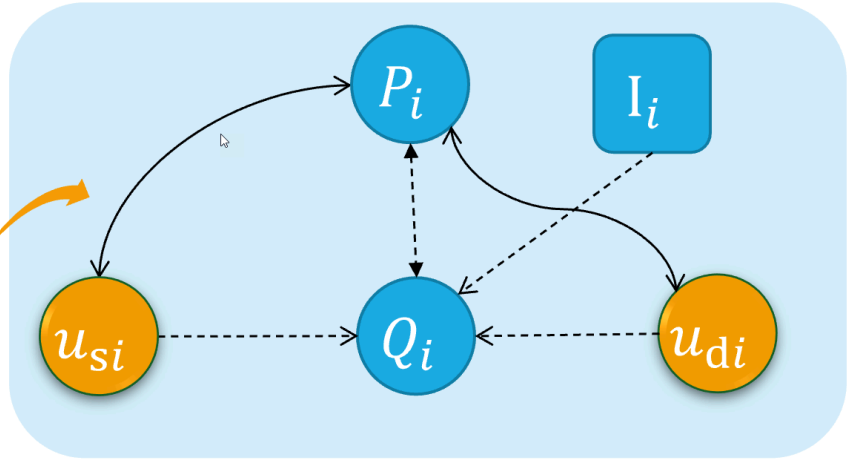
An intuitive demonstration is show as follows:

D: $Q_i = \alpha_1 + \alpha_2 P_i + \alpha_3 \mathrm{I}_i + u_{di}$



S: $Q_i = \beta_1 + \beta_2 P_i + u_{si}$





$$\Rightarrow \begin{cases} Cov(P_i, u_{di}) \neq 0 \\ Cov(P_i, u_{si}) \neq 0 \end{cases}$$

17.2 Estimation problem with endogeneity

Inconsistent estimates with measurement error

Consider the simple regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad (1)$$

We would like to measure the effect of the variable X on Y, but we can observe only an imperfect measure of it (i.e. a **proxy variable**), which is

$$X_i^* = X_i - v_i \quad (2)$$

Where v_i is a random disturbance with mean 0 and variance σ_v^2.

Further, let's assume that X_i, ϵ_i and v_i are **pairwise independent**.

Inconsistent estimates with measurement error

Given the assumed true model (1):

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad \text{eq}(1) \text{ assumed true model}$$

with the proxy variable X_i^*, we may use the error specified model (4):

$$\begin{aligned} X_i^* &= X_i - v_i && \text{eq}(2) \text{ proxy variable} \\ X_i &= X_i^* + v_i && \text{eq}(3) \\ Y_i &= \beta_0 + \beta_1 X_i^* + u_i && \text{eq}(4) \text{ error specified model} \end{aligned}$$

We can substitute eq (3) into the model (1) to obtain eq(5)

$$Y_i = \beta_0 + \beta_1 X_i^* + \epsilon_i = \beta_0 + \beta_1 \left(X_i^* + v_i\right) + \epsilon_i = \beta_0 + \beta_1 X_i^* + (\epsilon_i + \beta_1 v_i) \qquad \text{eq}(5)$$

which means $u_i = (\epsilon_i + \beta_1 v_i)$ in the error specified model. As we know, the OLS consistent estimator of β_1 in the last equation **requires** $\text{Cov}\left(X_i^*, u_i\right) = 0$.

Inconsistent estimates with measurement error

Note that $E(u_i) = E(\epsilon_i + \beta_1 v_i) = E(\epsilon_i) + \beta_1 E(v_i) = 0$

However,

$$
\begin{aligned}
\operatorname{Cov}\left(X_i^*, u_i\right) &= E\left[\left(X_i^* - E\left(X_i^*\right)\right)\left(u_i - E\left(u_i\right)\right)\right] \\
&= E\left(X_i^* u_i\right) \\
&= E\left[\left(X_i - v_i\right)\left(\epsilon_i + \beta_1 v_i\right)\right] \\
&= E\left[X_i \epsilon_i + \beta_1 X_i v_i - v_i \epsilon_i - \beta_1 v_i^2\right] \leftarrow \quad \text{(pairwise independent)} \\
&= -E\left(\beta_1 v_i^2\right) \\
&= -\beta_1 \operatorname{Var}(v_i) \\
&= -\beta_1 \sigma_{v_i}^2 \neq 0
\end{aligned}
$$

Thus, X in model (4) is **endogenous** and we expect the OLS estimator of β_1 to be **inconsistent**.

OLS estimation: Violation of A2

In general, when **A2** is violated, we expect OLS estimates to be **biased**:

The OLS estimators of $\hat{\beta}$ is

$$\widehat{\beta} = \beta + (X'X)^{-1}X'\epsilon \qquad (6)$$

and we can take expectation on both sides.

$$\begin{aligned} E(\widehat{\beta}) &= \beta + E\left((X'X)^{-1}X'\epsilon\right) \\ &= \beta + E\left(E\left((X'X)^{-1}X'\epsilon|X\right)\right) \\ &= \beta + E\left((X'X)^{-1}X'E(\epsilon|X)\right) \neq \beta \end{aligned}$$

If **A2** $E(\epsilon|X) = 0$ is violated, which means $E(\epsilon|X) \neq 0$, the OLS estimator is **biased**.

OLS estimation: consistency

Let's see under what conditions we can establish consistency.

$$p\lim \widehat{\beta} = \beta + p\lim \left((X'X)^{-1} X'\epsilon \right) = \beta + p\lim \left(\left(\frac{1}{n} X'X \right)^{-1} \frac{1}{n} X'\epsilon \right)$$

$$= \beta + p\lim \left(\frac{1}{n} X'X \right)^{-1} \times p\lim \left(\frac{1}{n} X'\epsilon \right)$$

By the WLLN (Weak Law of Large Numbers)

$$\frac{1}{n} X'\epsilon = \frac{1}{n} \sum_{i=1}^{n} X_i \epsilon_i \xrightarrow{p} E\left(X_i \epsilon_i\right)$$

Hence $\widehat{\beta}$ is consistent if $E\left(X_i \epsilon_i\right) = 0$ for all i. The condition $E\left(X_i \epsilon_i\right) = 0$ is more likely to be satisfied than A2 $E(\epsilon|X) = 0$. Thus, a large class of estimators that cannot be proved to be **unbiased** are **consistent**.

Wage example: the origin model

Consider the following "error specified" wage model:

$$lwage_i = \beta_1 + \beta_2 educ_i + \beta_3 exper_i + \beta_4 expersq_i + e_i$$

The difficulty with this model is that the error term may include some unobserved attributes, such as personal **ability**, that determine both wage and education.

In other words, the independent variable **educ** is correlated with the error term. And it is endogenous variable.

> **Note**:
> We will use **years of schooling** as the proxy variable of **educ** in practice, and it surely bring in error measurement issues as we have mentioned.

Wage example: All variables in dataset

With data set `wooldridge::mroz`, researchers were interest in the return to education for married women.

Copy CSV Excel

Variables and Labels

index	vars	labels
1	educ	years of schooling
2	exper	actual labor mkt exper
3	expersq	exper^2
4	fatheduc	father's years of schooling
5	lwage	log(wage)
6	motheduc	mother's years of schooling

Showing 1 to 6 of 22 entries

Previous 1 2 3 4 Next

Wage example: Raw dataset

Copy CSV Excel

dataset n=(428)

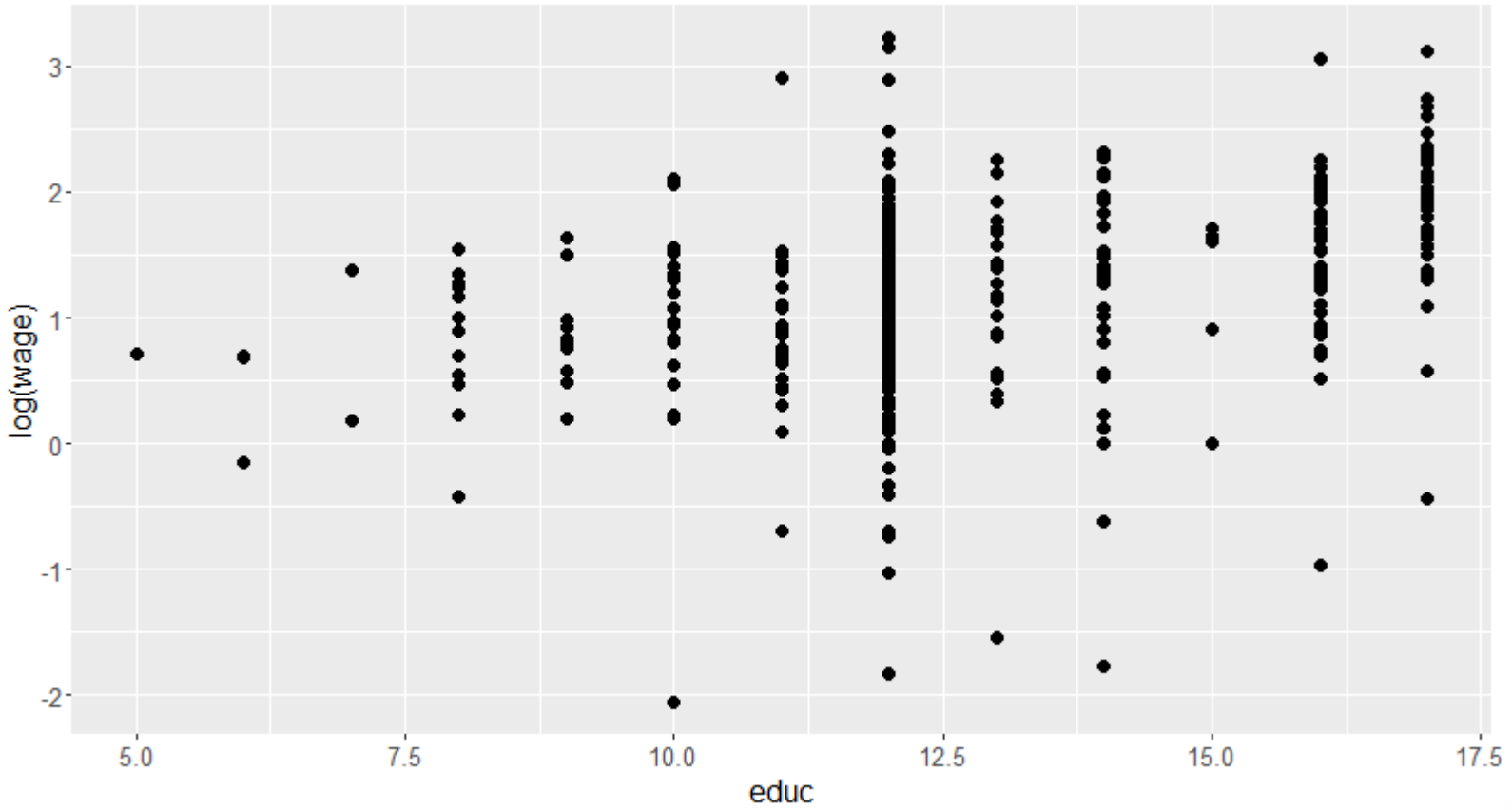
id	lwage	educ	exper	expersq	fatheduc	motheduc
1	1.21	12	14	196	7	12
2	0.33	12	5	25	7	7
3	1.51	12	15	225	7	12
4	0.09	12	6	36	7	7
5	1.52	14	7	49	14	12
6	1.56	12	33	1089	7	14
7	2.12	16	11	121	7	14
8	2.06	12	35	1225	3	3

Showing 1 to 8 of 428 entries

Previous 1 2 3 4 5 … 54 Next

Try yourself! Click and download the `.csv` file

Wage example: the scatter



Wage example: the mis-specified model

The Assumed TRUE model is

$$log(wage_i) = \beta_0 + \beta_1 edu_i + \beta_2 exper_i + \beta_3 exper_i^2 + \beta_4 ability_i + \epsilon_i$$

Now, let's consider the mis-specified model

$$log(wage_i) = \alpha_0 + \alpha_1 edu_i + \alpha_2 exper_i + \alpha_3 exper_i^2 + u_i \quad \text{(omitted ability)}$$

- In the theory analysis, we have known that this model is mis-specified due to the important omitted variables(`ability`) and it will be dropped to the disturbance term u_i.
- While the regressor `edu` is correlated with the disturbance term u_i (formally the omitted variables `ability`), thus the regressor `edu` is endogenous variable!

Wage example: use OLS method directly

Of course, you can conduct the OLS regression directly without considering problems due to endogeneity, and may obtain the inconsistent estimators (as we have proved).

```
mod_origin <- formula(lwage ~ educ +exper+expersq)
ols_origin <- lm(formula = mod_origin, data = mroz)
```

The OLS regression resuts is

$$
\begin{aligned}
\widehat{lwage} = &-0.52 &&+0.11educ &&+0.04exper &&-0.00expersq \\
(\text{t}) \quad &(-2.6282) &&(7.5983) &&(3.1549) &&(-2.0628) \\
(\text{se}) \quad &(0.1986) &&(0.0141) &&(0.0132) &&(0.0004) \\
(\text{fitness}) &R^2 = 0.1568; \bar{R}^2 = 0.1509 \\
&F^* = 26.29;\ p = 0.0000
\end{aligned}
$$

This looks good, but we know it is not reliable due to endogeneity behind this "error specified" model.

R Supplements: simple OLS

```
mod_origin <- formula(lwage ~ educ +exper+expersq)
ols_origin <- lm(formula = mod_origin, data = mroz)
summary(ols_origin)
```

```
Call:
lm(formula = mod_origin, data = mroz)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0840 -0.3063  0.0495  0.3750  2.3712

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.522041   0.198632   -2.63   0.0089 **
educ         0.107490   0.014146    7.60  1.9e-13 ***
exper        0.041567   0.013175    3.15   0.0017 **
expersq     -0.000811   0.000393   -2.06   0.0397 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

17.3 IV and the choices

IV: Motivation

We have seen that OLS estimator of β is inconsistent when one or more regressors is endogenous .

The **problems** of OLS arise because we imposed $E(X_i\epsilon_i) = 0$, which means we believe the sample data with

$$\boldsymbol{X}'\boldsymbol{e} = \mathbf{0}$$

When in fact error terms and regressors are correlated $E(X_i\epsilon_i) \neq 0$.

IV: Motivation

Suppose we can find a set of **explanatory variables** $\boldsymbol{Z}$ satisfying two conditions:

- **Relevance**: $\boldsymbol{Z}$ is correlated with $\boldsymbol{X}$
- **Exogeneity**: $\boldsymbol{Z}$ is not correlated with $\boldsymbol{\epsilon}$

These variables ($\boldsymbol{Z}$, in matrix form) can be used for consistent estimation and are known as **Instrumental Variables (IV)** .

IV: Estimators

Our instrumental variable estimator, $\hat{\beta}_{IV}$ is defined in terms of the following "**normal equation**" (**moment condition**, to be more precise)

$$\boldsymbol{Z}'\hat{\boldsymbol{\epsilon}} = \boldsymbol{Z}'\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{IV}}\right) = \mathbf{0}$$

and thus, provided that $Z'X$ is **square** and **non singular**,

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{IV}} = \left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{Z}'\boldsymbol{y}$$

The condition that $\boldsymbol{Z}'\boldsymbol{X}$ is square and non singular, intuitively, is satisfied when we have as many instruments as regressors (a situation that is called **exact identification**).

However $\hat{\boldsymbol{\beta}}_{\boldsymbol{IV}}$ is generally **biased** in **finite sample**, but we can show that it is still **consistent**.

IV: Consistency

$\hat{\boldsymbol{\beta}}_{\boldsymbol{IV}}$ is consistent. Since:

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{IV}} = (\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{y} = (\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}) = \boldsymbol{\beta} + (\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{\epsilon}$$

$$\begin{aligned} p\lim \hat{\boldsymbol{\beta}}_{\boldsymbol{IV}} &= \boldsymbol{\beta} + \boldsymbol{p}\boldsymbol{\lim}\left((\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{\epsilon}\right) \\ &= \boldsymbol{\beta} + \left(\boldsymbol{p}\boldsymbol{\lim}\left(\frac{\boldsymbol{1}}{\boldsymbol{n}}\boldsymbol{Z}'\boldsymbol{X}\right)\right)^{-1}\boldsymbol{\mathrm{plim}}\left(\frac{\boldsymbol{1}}{\boldsymbol{n}}\boldsymbol{Z}'\boldsymbol{\epsilon}\right) = \boldsymbol{\beta} \end{aligned}$$

- Relevance guarantees

$$\begin{aligned} p\lim\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right) &= p\lim\left(\frac{1}{n}\sum z_i X_i'\right) \\ &= E\left(Z_i X_i'\right) \neq 0 \end{aligned}$$

- Exogeneity ensures

$$\begin{aligned} p\lim\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{\epsilon}\right) &= p\lim\left(\frac{1}{n}\sum Z_i\epsilon_i\right) \\ &= E\left(Z_i\epsilon_i\right) = 0 \end{aligned}$$

IV: Inference

The natural estimator for σ^2 is

$$\hat{\sigma}^2_{IV} = \frac{\sum e_i^2}{n-k} = \frac{\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{IV}}\right)'\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{IV}}\right)}{n-k}$$

can be shown to be consistent (not proved here).

Thus, we can perform hypothesis testing based on IV estimators $\hat{\boldsymbol{\beta}}_{\boldsymbol{IV}}$.

Choice of instruments

However, finding **valid instruments** is the most difficult part of IV estimation in practice.

> Good instruments need not only to be exogenous, but also need be highly correlative with the regressors.
>
> **Joke**: If you can find a valid instrumental variable, you can get PhD from MIT.

Without a proof, we say that the **asymptotic variance** of $\hat{\beta}_{IV}$ is

$$\text{Var}\left(\hat{\boldsymbol{\beta}}_{\boldsymbol{IV}}\right) = \sigma^2 \left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1} \left(\boldsymbol{Z}'\boldsymbol{Z}\right) \left(\boldsymbol{X}'\boldsymbol{Z}\right)^{-1}$$

Where $\boldsymbol{X}'\boldsymbol{Z}$ is the matrix of covariances between instruments and regressors.

If such correlation is low, $\boldsymbol{X}'\boldsymbol{Z}$ will have elements close to zero and hence $\left(\boldsymbol{X}'\boldsymbol{Z}\right)^{-1}$ will have huge elements. Thus, $\text{Var}\left(\hat{\boldsymbol{\beta}}_{\boldsymbol{IV}}\right)$ will be very large.

Choice of instruments

The **common strategy** is to construct $\boldsymbol{Z} = (X_{ex}, X^*)$ generally.

- Variables X_{ex} in $\boldsymbol{X} = (X_{ex}, X_{en})$ are the assumed **exogenous** variables included in model.
- Other exogenous variables $\boldsymbol{X}^*$ are "close" to the model, but do not enter the model explicitly.

If X can be shown to be exogenous, then $X = Z$ and Gauss-Markov efficiency is recovered.

> **Instrumental Variable Estimators** do not have any efficiency properties .
> We can only talk about **relative efficiency**. It means that we can only choose the optimal set of instruments. Such that our estimator is the best we can obtain within all the class of possible instrumental variable estimators.

Too many available instruments

In case there are more **instruments** than **endogenous variables** (**over-identification**), we want to choose those instruments that have the highest correlation with X and hence give the lowest possible variance.

The best choice is obtained by using the fitted values of an OLS regression of each column of $\boldsymbol{X}$ on all instruments $\boldsymbol{Z}$, that is (after running k regressions, one for each column of $\boldsymbol{X}$)

$$\hat{\boldsymbol{X}} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{F}$$

We now use $\hat{\boldsymbol{X}}$ as instrument, which is $\hat{\boldsymbol{\beta}}_{\boldsymbol{IV}} = \left(\hat{\boldsymbol{X}}'\boldsymbol{X}\right)^{-1}\hat{\boldsymbol{X}}'\boldsymbol{y}$

We notice that (try to prove this):

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{IV}} = \left(\hat{\boldsymbol{X}}'\boldsymbol{X}\right)^{-1}\hat{\boldsymbol{X}}'\boldsymbol{y} = \left(\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y}$$

IV solution with omitted variables

Let's go back to our example of the wage equation. Assume we are modeling wage as a function of **education** and **ability**.

$$Wage_i = \beta_0 + \beta_1 Edu_i + \beta_2 Abl_i + \epsilon_i$$

However, individual's ability is clearly something that is not observed or measured and hence cannot be included in the model. since, ability is not included in the model it is included in the error term.

$$Wage_i = \beta_0 + \beta_1 Edu_i + e_i, \quad \text{where} \quad e_i = \beta_2 Abl_i + \epsilon_i$$

The problem is that ability not only affects wages but the more able individuals may spend more years in school, causing a positive correlation between the error term and education, $\text{cov}(Edu_i, e_i) > 0$.

Thus, $Educ$ is an **endogenous variable**.

IV solution to omitted variables

If we can find a valid instrument for Educ we can estimate β_1 using IV method. Suppose that we have a variable Z that satisfies the following conditions

- Z does not directly affect wages
- Z is uncorrelated with e (exogeneity), i.e.

$$\text{Cov}(e, z) = 0 \quad (4)$$

> since $e_i = \beta_2 Abl_i + \epsilon_i$, Z must be uncorrelated with ability.

- Z is (at least partially) correlated with the endogenous variable,i.e. Education (relevance),

$$\text{Cov}(Z, Edu) \neq 0 \quad (5)$$

> Such condition can be tested(α_2) by using a simple regression$:

$$Edu_i = \alpha_1 + \alpha_2 Z_i + u_i$$

Then, Z is a valid instrument for $Educ_i$. We showed earlier that the IV estimator of β_1 is consistent.

IV solution to omitted variables

Several economists have used **family background variables** as IVs for education.

- For example, **mother's education** is positively correlated with child's education, so it may satisfies condition of **Relevance**.
- Also, **father's education** is positively correlated with child's education, and it may satisfies condition of **Relevance**.

> The problem is that mother's or father 's education might also be correlated with child's ability(genetic inherited), in which case the condition of **Exogeneity** fails.

- Another IV for education that has been used by economists is the number of **siblings** while growing up.

> Typically, having more siblings is associated with lower average levels of education and it should be uncorrelated with innate ability.

17.4 Two-stage least squares method

Two-stage least squares: glance

When we have **more** instruments than endogenous variables, $\hat{\boldsymbol{\beta}}_{\boldsymbol{IV}}$ can be computed in 2 steps:

- **Step 1**: Regress each column of X on all the instruments (Z ,in matrix form). For each column of X, get the fitted values and combine them into the matrix $\hat{X}$.
- **Step 2**: Regress Y on $\hat{X}$

And, this procedure is named **two-stage least squares** or **2SLS** or **TSLS**.

Two-stage least squares: indentification

Consider the model setting

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ji} + \sum_{s=1}^{r} \beta_{k+s} W_{ri} + \epsilon_i$$

where $(X_{1i}, \ldots, X_{ki})$ are **endogenous regressors**, $(W_{1i}, \ldots, W_{ri})$ are **exogenous regressors** and there are m **instrumental variables** $(Z_{1i}, \ldots, Z_{mi})$ satisfying instrument relevance and instrument exogeneity conditions.

- When $m = k$,the coefficients are **exactly identified**.
- When $m > k$,the coefficients are **overidentified**.
- When $m < k$, the coefficients are **underidentified**.
- Finnaly, coefficients can be identified only when $m \geq k$.

Two-stage least squares: the procedure

- **Stage 1**: Regress X_{1i} on constant, all the instruments $(Z_{1i}, \ldots, Z_{mi})$ and all exogenous regressors $(W_{1i}, \ldots, W_{ri})$ using OLS and obtain the fitted values $\hat{X}_{1i}$. Repeat this to get $\left(\hat{X}_{1i}, \ldots, \hat{X}_{ki}\right)$
- **Stage 2**: Regress Y_i on constant, $\left(\hat{X}_{1i}, \ldots, \hat{X}_{ki}\right)$ and $(W_{1i}, \ldots, W_{ri})$ using OLS to obtain $\left(\hat{\beta}_0^{IV}, \hat{\beta}_1^{IV}, \ldots, \hat{\beta}_{k+r}^{IV}\right)$

Two-stage least squares: the solutions

We can conduct the **2SLS** procedure with following two solutions:

- use the **"Step-by-Step solution"** methods without variance correction.
- use the **"Integrated solution"** with variance correction.

Notice:

DO NOT use **"Step-by-Step solution"** solution in your paper! It is only for teaching purpose here.

In `R` ecosystem, we have two packages to execute the **Integrated solution**:

- We can use `systemfit` package function `systemfit::systemfit()`.
- Or we may use `ARE` package function `ARE::ivreg()`.

Step-by-step solution: stage 1 model

First, let's try to use $matheduc$ as instrument of endogenous variable $educ$.

Stage 1 of 2SLS: with mother education as instrument

we can obtain the fitted variable $\widehat{educ}$ by conduct the following **step 1** OLS regression

$$\widehat{educ} = \hat{\gamma}_1 + \hat{\gamma}_2 exper + \hat{\gamma}_3 expersq + \hat{\gamma}_4 mothereduc$$

Step-by-step solution: stage 1 OLS estimate

Here we obtain the OLS results of **Stage 1 of 2SLS**:

```
mod_step1 <- formula(educ~exper + expersq + motheduc)  # modle setting
ols_step1 <- lm(formula = mod_step1, data = mroz)  # OLS estimation
```

$$
\begin{aligned}
\widehat{educ} = &+ 9.78 &&+ 0.05 exper_i - 0.00 expersq_i + 0.27 motheduc_i \\
(s) \quad &(0.4239) &&(0.0417) \quad (0.0012) \quad (0.0311) \\
(t) \quad &(+23.06) &&(+1.17) \quad (-1.03) \quad (+8.60) \\
(fit) \quad &R^2 = 0.1527 \bar{R}^2 = 0.1467 \\
(Ftest) &F^* = 25.47 \ p = 0.0000
\end{aligned}
$$

The t -value for coefficient of *mothereduc* is so large (larger than 2), indicating a strong correlation between this instrument and the endogenous variable *educ* even after controlling for other variables.

Step-by-step solution: stage 1 OLS predicted values

Along with the regression of **Stage 1 of 2SLS**, we will extract the fitted value $\widehat{educ}$ and add them into new data set.

```
mroz_add <- mroz %>% mutate(educHat = fitted(ols_step1)) # add fitted educ to data
```

id	lwage	educ	exper	expersq	fatheduc	motheduc	educHat
1	1.21	12	14	196	7	12	13.42
2	0.33	12	5	25	7	7	11.86
3	1.51	12	15	225	7	12	13.43
4	0.09	12	6	36	7	7	11.90
5	1.52	14	7	49	14	12	13.27

Showing 1 to 5 of 428 entries

Previous 1 2 3 4 5 … 86 Next

Step-by-step solution: stage 2 model

Stage 2 of 2SLS: with mother education as instrument

In the second stage, we will regress log(wage) on the $\widehat{educ}$ from stage 1 and experience and its quadratic term exp square.

$$lwage = \hat{\beta}_1 + \hat{\beta}_2 \widehat{educ} + \hat{\beta}_3 exper + \hat{\beta}_4 expersq + \hat{\epsilon}$$

```
mod_step2 <- formula(lwage~educHat + exper + expersq)
ols_step2 <- lm(formula = mod_step2, data = mroz_add)
```

Step-by-step solution: stage 2 OLS estimate

By using the new data set (`moroz_add`), the result of the explicit 2SLS procedure are shown as below.

$$
\begin{array}{lllll}
\widehat{lwage} = +0.20 & & +0.05 educHat_i + 0.04 exper_i - 0.00 expersq_i & & \\
(s) \quad (0.4933) & & (0.0391) & (0.0142) & (0.0004) \\
(t) \quad (+0.40) & & (+1.26) & (+3.17) & (-2.17) \\
(fit) \quad R^2 = 0.0456 & \bar{R}^2 = 0.0388 & & & \\
(Ftest) \quad F^* = 6.75 & p = 0.0002 & & &
\end{array}
$$

Keep in mind, however, that the **standard errors** calculated in this way are incorrect (Why?).

Integrated solution: the whole story

We need a **Integrated solution** for following reasons:

- We should obtain the correct estimated error for test and inference.
- We should avoid tedious steps in the former step-by-step routine. When the model contains more than one endogenous regressors and there are lots available instruments, then the step-by-step solution will get extremely tedious.

Integrated solution: the `R` toolbox

In `R` ecosystem, we have two packages to execute the integrated solution:

- We can use `systemfit` package function `systemfit::systemfit()`.
- Or we may use `ARE` package function `ARE::ivreg()`.

Both of these tools can conduct the integrated solution, and will adjust the variance of estimators automatically.

Integrated solution: `motheduc` IV model

In order to get the correct estimated error, we need use the **"integrated solution"** for 2SLS. And we will process the estimation with proper software and tools.

Firstly, let's consider using *matheduc* as the only instrument for *educ*.

$$\begin{cases} \widehat{educ} = \hat{\gamma}_1 + \hat{\gamma}_2 exper + \hat{\gamma}_3 expersq + \hat{\gamma}_4 motheduc & \text{(stage 1)} \\ lwage = \hat{\beta}_1 + \hat{\beta}_2 \widehat{educ} + \hat{\beta}_3 exper + \hat{\beta}_4 expersq + \hat{\epsilon} & \text{(stage 2)} \end{cases}$$

Integrated solution: `motheduc` IV results

2SLS result(*motheduc* as instrument)

eq	vars	Estimate	Std. Error	t value	Pr(>|t|)
eq1	(Intercept)	9.7751	0.4239	23.0605	0.0000
eq1	exper	0.0489	0.0417	1.1726	0.2416
eq1	expersq	-0.0013	0.0012	-1.0290	0.3040
eq1	motheduc	0.2677	0.0311	8.5992	0.0000
eq2	(Intercept)	0.1982	0.4729	0.4191	0.6754
eq2	educ	0.0493	0.0374	1.3159	0.1889
eq2	exper	0.0449	0.0136	3.3039	0.0010
eq2	expersq	-0.0009	0.0004	-2.2690	0.0238

- The t-test for variable *educ* is significant (p-value less than 0.05).

Note : The corresponding code of `R` programming is in the following slides. The table results use the report from the `systemfit::systemfit()` function.

(Supplements) R code (m): `systemfit::systemfit()`

The R code using `systemfit::systemfit()` as follows:

```r
# load pkg
require(systemfit)
# set two models
eq_1 <- educ ~  exper + expersq + motheduc
eq_2 <- lwage ~ educ + exper + expersq
sys <- list(eq1 = eq_1, eq2 = eq_2)
# specify the instruments
instr <- ~  exper + expersq + motheduc
# fit models
fit.sys <- systemfit(
  sys, inst=instr,
  method="2SLS", data = mroz)
# summary of model fit
smry.system_m <- summary(fit.sys)
```

(Supplements) R report (m): `systemfit::systemfit()`

The following is the 2SLS analysis report using `systemfit::systemfit()`:

```
smry.system_m
```

```
systemfit results
method: 2SLS

         N  DF  SSR detRCov OLS-R2 McElroy-R2
system 856 848 2085    1.97   0.15      0.112

      N  DF  SSR   MSE RMSE    R2 Adj R2
eq1 428 424 1890 4.457 2.11 0.153  0.147
eq2 428 424  196 0.462 0.68 0.123  0.117

The covariance matrix of the residuals
      eq1   eq2
eq1 4.457 0.305
eq2 0.305 0.462

The correlations of the residuals
```

NOTE : `systemfit::systemfit()` simultaneously reports the analysis results of two equations in 2SLS!

(Supplements) R code (m): `ARE::ivreg()`

The R code using `ARE::ivreg()` as follows:

```r
# load pkg
require(AER)
# specify model
mod_iv_m <- formula(lwage ~ educ + exper + expersq
                    | motheduc + exper + expersq)
# fit model
lm_iv_m <- ivreg(formula = mod_iv_m, data = mroz)
# summary of model fit
smry.ivm <- summary(lm_iv_m)
```

(Supplements) R report (m): `ARE::ivreg()`

The following is the 2SLS analysis report using `ARE::ivreg()`:

```
smry.ivm
```

```
Call:
ivreg(formula = mod_iv_m, data = mroz)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1080 -0.3263  0.0602  0.3677  2.3435

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.198186   0.472877    0.42    0.675
educ         0.049263   0.037436    1.32    0.189
exper        0.044856   0.013577    3.30    0.001 **
expersq     -0.000922   0.000406   -2.27    0.024 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note: `ARE::ivreg()` Only reports the result of the last equation of 2SLS, not include the first equation!

Integrated solution: `fatheduc` IV model

Now let's consider using $fatheduc$ as the only instrument for $educ$.

$$\begin{cases} \widehat{educ} = \hat{\gamma}_1 + \hat{\gamma}_2 exper + \hat{\gamma}_3 expersq + \hat{\gamma}_4 fatheduc & \text{(stage 1)} \\ lwage = \hat{\beta}_1 + \hat{\beta}_2 \widehat{educ} + \hat{\beta}_3 exper + \hat{\beta}_4 expersq + \hat{\epsilon} & \text{(stage 2)} \end{cases}$$

We will repeat the whole procedure with `R`.

Integrated solution: `fatheduc` IV results

2SLS result(*fatheduc* as instrument)

eq	vars	Estimate	Std. Error	t value	Pr(>|t|)
eq1	(Intercept)	9.8870	0.3956	24.9920	0.0000
eq1	exper	0.0468	0.0411	1.1391	0.2553
eq1	expersq	-0.0012	0.0012	-0.9364	0.3496
eq1	fatheduc	0.2705	0.0289	9.3670	0.0000
eq2	(Intercept)	-0.0611	0.4364	-0.1400	0.8887
eq2	educ	0.0702	0.0344	2.0389	0.0421
eq2	exper	0.0437	0.0134	3.2590	0.0012
eq2	expersq	-0.0009	0.0004	-2.2003	0.0283

- The t-test for variable *educ* is significant (p-value less than 0.05).

Note : The corresponding code of `R` programming is in the following slides. The table results use the report from the `systemfit::systemfit()` function.

(Supplements) R code (f): `systemfit::systemfit()`

The R code using `systemfit::systemfit()` as follows:

```r
# load pkg
require(systemfit)
# set two models
eq_1 <- educ ~  exper + expersq + fatheduc
eq_2 <- lwage ~ educ + exper + expersq
sys <- list(eq1 = eq_1, eq2 = eq_2)
# specify the instruments
instr <- ~ exper + expersq + fatheduc
# fit models
fit.sys <- systemfit(
  sys, inst=instr,
  method="2SLS", data = mroz)
# summary of model fit
smry.system_f <- summary(fit.sys)
```

(Supplements) R report (f): `systemfit::systemfit()`

The following is the 2SLS analysis report using `systemfit::systemfit()`:

```
smry.system_f
```

```
systemfit results
method: 2SLS

         N  DF  SSR detRCov OLS-R2 McElroy-R2
system 856 848 2030    1.92  0.173      0.135

      N  DF  SSR   MSE  RMSE    R2 Adj R2
eq1 428 424 1839 4.337 2.082 0.176  0.170
eq2 428 424  191 0.451 0.672 0.143  0.137

The covariance matrix of the residuals
      eq1   eq2
eq1 4.337 0.195
eq2 0.195 0.451

The correlations of the residuals
```

NOTE : `systemfit::systemfit()` simultaneously reports the analysis results of two equations in 2SLS!

(Supplements) R code (f): `ARE::ivreg()`

The R code using `ARE::ivreg()` as follows:

```
# load pkg
require(AER)
# specify model
mod_iv_f <- formula(lwage ~ educ + exper + expersq
                    | fatheduc + exper + expersq)
# fit model
lm_iv_f <- ivreg(formula = mod_iv_f, data = mroz)
# summary of model fit
smry.ivf <- summary(lm_iv_f)
```

(Supplements) R report (f): `ARE::ivreg()`

The following is the 2SLS analysis report using `ARE::ivreg()`:

```
smry.ivf
```

```
Call:
ivreg(formula = mod_iv_f, data = mroz)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0917 -0.3278  0.0501  0.3736  2.3535

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.061117   0.436446   -0.14   0.8887
educ         0.070226   0.034443    2.04   0.0421 *
exper        0.043672   0.013400    3.26   0.0012 **
expersq     -0.000882   0.000401   -2.20   0.0283 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note: `ARE::ivreg()` Only reports the result of the last equation of 2SLS, not include the first equation!

Integrated solution: `mothedu` and `fatheduc` IV model

Also, we can use both $matheduc$ and $fatheduc$ as instruments for $educ$.

$$\begin{cases} \widehat{educ} = \hat{\gamma}_1 + \hat{\gamma}_2 exper + \hat{\beta}_3 expersq + \hat{\beta}_4 motheduc + \hat{\beta}_5 fatheduc & \text{(stage 1)} \\ lwage = \hat{\beta}_1 + \hat{\beta}_2 \widehat{educ} + \hat{\beta}_3 exper + \hat{\beta}_4 expersq + \hat{\epsilon} & \text{(stage 2)} \end{cases}$$

Integrated solution: `mothedu` and `fatheduc` IV results

2SLS result(*motheduc* and *fatheduc* as instruments)

eq	vars	Estimate	Std. Error	t value	Pr(>|t|)
eq1	(Intercept)	9.1026	0.4266	21.3396	0.0000
eq1	exper	0.0452	0.0403	1.1236	0.2618
eq1	expersq	-0.0010	0.0012	-0.8386	0.4022
eq1	motheduc	0.1576	0.0359	4.3906	0.0000
eq1	fatheduc	0.1895	0.0338	5.6152	0.0000
eq2	(Intercept)	0.0481	0.4003	0.1202	0.9044
eq2	educ	0.0614	0.0314	1.9530	0.0515
eq2	exper	0.0442	0.0134	3.2883	0.0011
eq2	expersq	-0.0009	0.0004	-2.2380	0.0257

Note : The corresponding code of `R` programming is in the following slides. The table results use the report from the `systemfit::systemfit()` function.

(Supplements) R code (mf): `systemfit::systemfit()`

The R code using `systemfit::systemfit()` as follows:

```r
# load pkg
require(systemfit)
# set two models
eq_1 <- educ ~ exper + expersq + motheduc + fatheduc
eq_2 <- lwage ~ educ + exper + expersq
sys <- list(eq1 = eq_1, eq2 = eq_2)
# specify the instruments
instr <- ~ exper + expersq + motheduc + fatheduc
# fit models
fit.sys <- systemfit(
  sys, inst=instr,
  method="2SLS", data = mroz)
# summary of model fit
smry.system_mf <- summary(fit.sys)
```

(Supplements) R report (mf): `systemfit::systemfit()`

The following is the 2SLS analysis report using `systemfit::systemfit()`:

```
smry.system_mf
```

```
systemfit results
method: 2SLS

         N  DF  SSR detRCov OLS-R2 McElroy-R2
system 856 847 1952    1.83  0.205      0.149

      N  DF  SSR   MSE  RMSE    R2 Adj R2
eq1 428 423 1759 4.157 2.039 0.211  0.204
eq2 428 424  193 0.455 0.675 0.136  0.130

The covariance matrix of the residuals
      eq1   eq2
eq1 4.157 0.242
eq2 0.242 0.455

The correlations of the residuals
```

NOTE : `systemfit::systemfit()` simultaneously reports the analysis results of two equations in 2SLS!

(Supplements) R code (mf): `ARE::ivreg()`

The R code using `ARE::ivreg()` as follows:

```r
# load pkg
require(AER)
# specify model
mod_iv_mf <- formula(
  lwage ~ educ + exper + expersq
        | motheduc + fatheduc + exper + expersq)
# fit model
lm_iv_mf <- ivreg(formula = mod_iv_mf, data = mroz)
# summary of model fit
smry.ivmf <- summary(lm_iv_mf)
```

(Supplements) R report (mf): `ARE::ivreg()`

The following is the 2SLS analysis report using `ARE::ivreg()`:

```
smry.ivmf
```

```
Call:
ivreg(formula = mod_iv_mf, data = mroz)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0986 -0.3196  0.0551  0.3689  2.3493

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.048100   0.400328    0.12   0.9044
educ         0.061397   0.031437    1.95   0.0515 .
exper        0.044170   0.013432    3.29   0.0011 **
expersq     -0.000899   0.000402   -2.24   0.0257 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note: `ARE::ivreg()` Only reports the result of the last equation of 2SLS, not include the first equation!

Solutions comparison: a glance

Until now, we obtain totally **Five** estimation results with different model settings or solutions:

a. Error specification model with OLS regression directly.

b. (**Step-by-Step solution**) Explicit 2SLS estimation **without** variance correction (IV regression step by step with only $matheduc$ as instrument).

c. (**Integrated solution**) Dedicated IV estimation **with** variance correction (using `R` tools of `systemfit::systemfit()` or `ARE::ivreg()`).

- The IV model with only $motheduc$ as instrument for endogenous variable edu
- The IV model with only $fatheduc$ as instrument for endogenous variable edu
- The IV model with both $motheduc$ and $fatheduc$ as instruments

For the purpose of comparison, all results will show in next slide.

Solutions comparison: tidy reports (png)

lwage equation: OLS, 2SLS, and IV models compared

	Dependent variable: lwage				
	OLS	explicit 2SLS	IV mothereduc	IV fathereduc	IV mothereduc and fathereduc
	(1)	(2)	(3)	(4)	(5)
Constant	-0.5200***	0.2000	0.2000	-0.0610	0.0480
	(0.2000)	(0.4900)	(0.4700)	(0.4400)	(0.4000)
educ	0.1100***		0.0490	0.0700**	0.0610*
	(0.0140)		(0.0370)	(0.0340)	(0.0310)
educHat		0.0490			
		(0.0390)			
exper	0.0420***	0.0450***	0.0450***	0.0440***	0.0440***
	(0.0130)	(0.0140)	(0.0140)	(0.0130)	(0.0130)
expersq	-0.0008**	-0.0009**	-0.0009**	-0.0009**	-0.0009**
	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)
Observations	428	428	428	428	428
R^2	0.1600	0.0460	0.1200	0.1400	0.1400
Adjusted R^2	0.1500	0.0390	0.1200	0.1400	0.1300
Residual Std. Error (df = 424)	0.6700	0.7100	0.6800	0.6700	0.6700
F Statistic (df = 3; 424)	26.0000***	6.8000***			

Solutions comparison: tidy reports (html)

	Dependent variable: lwage				
	OLS	explicit 2SLS	IV mothereduc	IV fathereduc	IV mothereduc and fathereduc
	(1)	(2)	(3)	(4)	(5)
Constant	-0.5200***	0.2000	0.2000	-0.0610	0.0480
	(0.2000)	(0.4900)	(0.4700)	(0.4400)	(0.4000)
educ	0.1100***		0.0490	0.0700**	0.0610*
	(0.0140)		(0.0370)	(0.0340)	(0.0310)
educHat		0.0490			
		(0.0390)			
exper	0.0420***	0.0450***	0.0450***	0.0440***	0.0440***
	(0.0130)	(0.0140)	(0.0140)	(0.0130)	(0.0130)
	****	****	****	****	****

Solutions comparison: report tips

- The second column shows the result of the direct OLS estimation, and the third column shows the result of explicit 2SLS estimation without variance correction.
- While the last three column shows the results of IV solution with variance correction.
- And we should also remind that the *educ* in the IV model is equivalent to the *educHat* in 2SLS.
- The value within the bracket is the standard error of the estimator.

Solutions comparison: report insights

So the key points of this comparison including:

- Firstly, the table shows that the importance of education in determining wage decreases in the IV model (3) (4) and (5) with the coefficients 0.049, 0.07, 0.061 respectively. And the standard error also decrease along IV estimation (3) , (4) and (5).
- Secondly, It also shows that the explicit 2SLS model (2) and the IV model with only *motheduc* instrument yield the same coefficients, but the **standard errors** are different. The standard error in explicit 2SLS is 0.039, which is little large than the standard error 0.037 in IV estimation.
- Thirdly, the t-test of the coefficient on education shows no significance when we use `motheduc` as the only instrument for education. You can compare this under the explicit 2SLS estimation or IV estimation.
- Fourthly, we can fully feel and understand the **relative estimated efficiency** of 2SLS!

Solutions comparison: further thinking

After the empirical comparison, we will be even more confused with these results.

While, new question will arise inside our mind.

- Which estimation is the best?
- How to judge and evaluate different instrument choices?

We will discuss these topics in the next section.

17.5 Testing Instrument validity

Instrument vality: the concept

Consider the general model

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ji} + \sum_{s=1}^{r} \beta_{k+s} W_{ri} + \epsilon_i$$

- Y_i is the dependent variable
- $\beta_0, \ldots, \beta_{k+1}$ are $1 + k + r$ unknown regression coefficients
- $X_{1i}, \ldots, X_{ki}$ are k endogenous regressors
- $W_{1i}, \ldots, W_{ri}$ are r exogenous regressors which are uncorrelated with u_i
- u_i is the error term
- $Z_{1i}, \ldots, Z_{mi}$ are m instrumental variables

Instrument valid means satisfy both Relevance and Exogeneity conditions.

$$E\left(Z_i X_i'\right) \neq 0 \qquad E\left(Z_i \epsilon_i\right) = 0$$

Instrument Relevance: relax condition

In practice, **Instrument Relevance** also means that:

If there are k endogenous variables and m instruments Z, and $m \geq k$, it must hold that the exogenous vector

$$\left(\hat{X}_{1i}^{*}, \ldots, \hat{X}_{ki}^{*}, W_{1i}, \ldots, W_{ri}, 1\right)$$

should not be **perfectly multicollinear**.

> **Where**:
>
> - $\hat{X}_{1i}^{*}, \ldots, \hat{X}_{ki}^{*}$ are the predicted values from the k first stage regressions.
> - 1 denotes the constant regressor which equals 1 for all observations.

Instrument Relevance: Weak instrument

Instruments that explain little variation in the endogenous regressor X are called **weak instruments**.

Formally, When $\text{corr}(Z_i, X_i)$ is close to zero, z_i is called a weak instrument.

- Consider a simple one regressor model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
- The IV estimator of β_1 is $\widehat{\beta}_1^{IV} = \frac{\sum_{i=1}^{n}(Z_i - \bar{Z})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(Z_i - \bar{Z})(X_i - \bar{X})}$

> Note that $\frac{1}{n}\sum_{i=1}^{n}\left(Z_i - \bar{Z}\right)\left(Y_i - \bar{Y}\right) \xrightarrow{p} \text{Cov}(Z_i, Y_i)$
> and $\frac{1}{n}\sum_{i=1}^{n}\left(Z_i - \bar{Z}\right)\left(X_i - \bar{X}\right) \xrightarrow{p} \text{Cov}(Z_i, X_i)$.

- Thus,if $\text{Cov}(Z_i, X_i) \approx 0$, then $\widehat{\beta}_1^{IV}$ is useless.

Example: Weak instrument

We want to run a simple regression to assess the effect of smoking on child birth weight. The model we run is

$$\log(\text{bwght}) = \beta_0 + \beta_1 \text{packs} + \epsilon_i$$

where packs is the number of packs of cigarettes mother smokes per day. We suspect that packs might be endogenous. (So why?) so we use average price of cigarettes in the state of residence, as an instrument. We assume that cigprice is uncorrelated with ϵ.

Example: Weak instrument

However, by regressing *packs* on *cigprice* in stage 1, we find basically no effect.

$$\begin{aligned}\widehat{packs} &= 0.067 + 0.0003 \text{ cigprice} \\ &\quad (0.103)(0.0008)\end{aligned}$$

If we insist to use *cigprice* as an instrument, and conduct the stage 2 regression, we will find

$$\begin{aligned}\log(\widehat{bwght}) &= 4.45 + 2.99\text{packs} \\ &\quad (0.91)(8.70)\end{aligned}$$

Obviously, this estimate is meaningless (Why?).

The *cigprice* behaves as a **weak instrument**, and the problem was already exposed in stage 1 regression.

Weak instrument: the strategy

There are two ways to proceed if instruments are weak:

- Discard the **weak instruments** and/or find **stronger instruments**.

> While the former is only an option if the unknown coefficients remain identified when the weak instruments are discarded, the latter can be difficult and even may require a redesign of the whole study.

- Stick with the weak instruments but use methods that improve upon TSLS.

> Such as **limited information maximum likelihood estimation (LIML)**.

Weak instrument: restricted F-test (idea)

In case with a **single** endogenous regressor, we can take the **F-test** to check the **Weak instrument**.

The basic idea of the F-test is very simple:

If the estimated coefficients of **all instruments** in the **first-stage** of a 2SLS estimation are **zero**, the instruments do not explain any of the variation in the X which clearly violates the relevance assumption.

Weak instrument: restricted F-test (procudure)

We may use the following rule of thumb:

- Conduct the **first-stage regression** of a 2SLS estimation

$$X_i = \hat{\gamma}_0 + \hat{\gamma}_1 W_{1i} + \ldots + \hat{\gamma}_p W_{pi} + \hat{\theta}_1 Z_{1i} + \ldots + \hat{\theta}_q Z_{qi} + v_i \quad (3)$$

- Test the restricted joint hypothesis $H_0 : \hat{\theta}_1 = \ldots = \hat{\theta}_q = 0$ by compute the F-statistic.
- If the F-statistic is less than critical value, the instruments are **weak**.

The rule of thumb is easily implemented in `R`. Run the first-stage regression using `lm()` and subsequently compute the restricted F-statistic by `R` function of `car::linearHypothesis()`.

Wage example: restricted F-test (models)

For all three IV model, we can test instrument(s) relevance respectively.

$$educ = \gamma_1 + \gamma_2 exper + \gamma_2 expersq + \theta_1 motheduc + v \quad \text{(relevance test 1)}$$
$$educ = \gamma_1 + \gamma_2 exper + \gamma_2 expersq + \theta_2 fatheduc + v \quad \text{(relevance test 2)}$$
$$educ = \gamma_1 + \gamma_2 exper + \gamma_2 expersq + \theta_1 motheduc + \theta_2 fatheduc + v \quad \text{(relevance test 3)}$$

Wage example: restricted F-test (model 1)

Consider model 1:

$$educ = \gamma_1 + \gamma_2 exper + \gamma_3 expersq + \theta_1 motheduc + v$$

The restricted F-test' null hypothesis: $H_0 : \theta_1 = 0$.

We will test whether `motheduc` are week instruments.

Wage example: restricted F-test (model 1)

The result show that the p-value of F^* is much smaller than 0.01. Null hypothesis H_0 was rejected. `motheduc` is **instruments relevance** (exogeneity valid).

R Code R result

```
# restricted F-test
constrain_test1 <- linearHypothesis(ols_relevance1, c("motheduc=0"))
# obtain F statistics
F_r1 <- constrain_test1$F[[2]]
```

Wage example (compare): classic F-test (model 1)

> Note: **Restriced F test** (73.95) is different with the **classical OLS F test**(show bellow 25.47).

$$educ = \gamma_1 + \gamma_2 exper + \gamma_2 expersq + \theta_1 motheduc + v$$

The classic OLS F-test' null hypothesis: $H_0 : \gamma_2 = \gamma_3 = \theta_1 = 0$.

The OLS estimation results are:

$$\begin{array}{llllll}
\widehat{educ} = & +9.78 & +0.05 exper_i & -0.00 expersq_i & +0.27 motheduc_i \\
(s) & (0.4239) & (0.0417) & (0.0012) & (0.0311) \\
(t) & (+23.06) & (+1.17) & (-1.03) & (+8.60) \\
(fit) & R^2 = 0.1527 & \bar{R}^2 = 0.1467 & & \\
(Ftest) & F^* = 25.47 & p = 0.0000 & &
\end{array}$$

Wage example: restricted F-test (model 2)

Consider model 2:

$$educ = \gamma_1 + \gamma_2 exper + \gamma_3 expersq + \theta_1 fatheduc + v \qquad \text{(relevance test 2)}$$

The restricted F-test' null hypothesis: $H_0 : \theta_1 = 0$.

We will test whether `fatheduc` are week instruments.

Wage example: restricted F-test (model 2)

The result show that the p-value of F^* is much smaller than 0.01. Null hypothesis H_0 was rejected. `fatheduc` is **instruments relevance** (exogeneity valid).

R Code R result

```
constrain_test2 <- linearHypothesis(ols_relevance2, c("fatheduc=0"))
# obtain F statistics
F_r2 <- constrain_test2$F[[2]]
```

Wage example: restricted F-test (model 3)

Consider model 3:

$$educ = \gamma_1 + \gamma_2 exper + \gamma_3 expersq + \theta_1 motheduc + \theta_2 fatheduc + v \qquad \text{(relevance test 3)}$$

The restricted F-test' null hypothesis: $H_0 : \theta_1 = \theta_2 = 0$.

We will test whether `motheduc` and `fatheduc` are week instruments.

Wage example: restricted F-test (model 3)

The result show that the p-value of F^* is much smaller than 0.01. Null hypothesis H_0 was rejected. `fatheduc` and `motheduc` are **instruments relevance** (exogeneity valid).

R Code R result

```
constrain_test3 <- linearHypothesis(ols_relevance3, c("motheduc=0", "fatheduc=0"))
# obtain F statistics
F_r3 <- constrain_test3$F[[2]]
```

Weak instrument: Cragg-Donald test

The former test for weak instruments might be unreliable with **more than** one endogenous regressor, though, because there is indeed one F-statistic for each endogenous regressor.

An alternative is the **Cragg-Donald test** based on the following statistic:

$$F = \frac{N - G - B}{L} \frac{r_B^2}{1 - r_B^2}$$

- where: G is the number of exogenous regressors; B is the number of endogenous regressors; L is the number of external instruments; r_B is the lowest canonical correlation.

> **Canonical correlation** is a measure of the correlation between the endogenous and the exogenous variables, which can be calculated by the function `cancor()` in `R`.

Hour example: backgound

Let us construct another IV model with two endogenous regressors. We assumed the following work hours determination model:

$$hushrs = \beta_1 + \beta_2 mtr + \beta_3 educ + \beta_4 kidsl6 + \beta_5 nwifeinc + e$$

- $hushrs$: work hours of husband, 1975
- mtr: federal marriage tax rate on woman
- $kidslt6$: have kids < 6 years (dummy variable)
- $nwifeinc$: wife's net income

There are:

- Two **endogenous variables**: $educ$ and mtr
- Two **exogenous regressors**: $nwifeinc$ and $kidslt6$
- And two external **instruments**: $motheduc$ and $fatheduc$.

Hour example: Cragg-Donald test (R code)

The data set is still `mroz`, restricted to women that are in the labor force($inlf = 1$).

```
# filter samples
mroz1 <- wooldridge::mroz %>%
  filter(wage>0, inlf==1)
# set parameters
N <- nrow(mroz1); G <- 2; B <- 2; L <- 2
# for endogenous variables
x1 <- resid(lm( mtr ~ kidslt6 + nwifeinc, data = mroz1))
x2 <- resid(lm( educ ~ kidslt6 + nwifeinc, data = mroz1))
# for instruments
z1 <-resid(lm(motheduc ~ kidslt6 + nwifeinc, data = mroz1))
z2 <-resid(lm(fatheduc ~ kidslt6 + nwifeinc, data=mroz1))
# column bind
X <- cbind(x1,x2)
Y <- cbind(z1,z2)
# calculate Canonical correlation
rB <- min(cancor(X,Y)$cor)
# obtain the F statistics
CraggDonaldF <- ((N-G-L)/L)/((1-rB^2)/rB^2)
```

Hour example: Cragg-Donald test (result)

Run these code lines, we can obtain the results:

Cragg-Donald test results

G	L	B	N	rb	CraggDonaldF
2	2	2	428	0.0218	0.1008

The result show the Cragg-Donald $F = 0.1008$, which is much smaller than **the critical value** `4.58`[1].

This test can not rejects the null hypothesis, thus we may conclude that some of these instruments are **weak**.

[1] The critical value can be found in table 10E.1 at: Hill C, Griffiths W, Lim G. Principles of econometrics[M]. John Wiley & Sons, 2018.

Instrument Exogeneity: the difficulty

Instrument Exogeneity means all m instruments must be uncorrelated with the error term,

$$Cov(Z_{1i}, \epsilon_i) = 0; \quad \ldots; \quad Cov(Z_{mi}, \epsilon_i) = 0.$$

- In the context of the simple IV estimator, we will find that the exogeneity requirement **can not** be tested. (Why?)
- However, if we have more instruments than we need, we can effectively test whether **some of** them are uncorrelated with the structural error.

Instrument Exogeneity: over-identification case

Under **over-identification** $(m > k)$, consistent IV estimation with (multiple) different combinations of instruments is possible.

> If instruments are exogenous, the obtained estimates should be **similar**.
>
> If estimates are very **different**, some or all instruments may not be exogenous.

The **Overidentifying Restrictions Test** (**J test**) formally check this.

- The null hypothesis is Instrument Exogeneity.

$$H_0 : E\left(Z_{hi}\epsilon_i\right) = 0, \text{ for all } h = 1, 2, \ldots, m$$

Instrument Exogeneity: J-test (procedure)

The **overidentifying restrictions test** (also called the J-test, or **Sargan test**) is an approach to test the hypothesis that the additional instruments are exogenous.

Procedure of overidentifying restrictions test is:

- **Step 1**: Compute the **IV regression residuals** :

$$\hat{\epsilon}_i^{IV} = Y_i - \left(\hat{\beta}_0^{IV} + \sum_{j=1}^{k} \hat{\beta}_j^{IV} X_{ji} + \sum_{s=1}^{r} \hat{\beta}_{k+s}^{IV} W_{si} \right)$$

- **Step 2**: Run the **auxiliary regression**: regress the IV residuals on instruments and exogenous regressors. And test the joint hypothesis $H_0 : \alpha_1 = 0, \ldots, \alpha_m = 0$

$$\hat{\epsilon}_i^{IV} = \theta_0 + \sum_{h=1}^{m} \theta_h Z_{hi} + \sum_{s=1}^{r} \gamma_s W_{si} + v_i \quad (2)$$

Instrument Exogeneity: J-test (procedure)

- **Step3**: Compute the **J statistic**: $J = mF$

> where F is the F-statistic of the m restrictions $H_0 : \theta_1 = \ldots = \theta_m = 0$ in eq(2)

Under the **null hypothesis**, J statistic is distributed as $\chi^2(m-k)$ approximately for large samples.

$$\boldsymbol{J} \sim \chi^2(m-k)$$

> IF J is **less** than **critical value**, it means that all instruments are exogenous.
>
> IF J is **larger** than **critical value**, it mean that some of the instruments are enogenous.

- We can apply the J-test by using `R` function `linearHypothesis()`.

Wage example: J-test (models)

Again, we can use both $matheduc$ and $fatheduc$ as instruments for $educ$.

Thus, the IV model is over-identification, and we can test the exogeneity of both these two instruments by using **J-test**.

The 2SLS model will be set as below.

$$\begin{cases} \widehat{educ} = \hat{\gamma}_1 + \hat{\gamma}_2 exper + \hat{\beta}_3 expersq + \hat{\beta}_4 motheduc + \hat{\beta}_5 fatheduc & \text{(stage 1)} \\ lwage = \hat{\beta}_1 + \hat{\beta}_2 \widehat{educ} + \hat{\beta}_3 exper + \hat{\beta}_4 expersq + \hat{\epsilon} & \text{(stage 2)} \end{cases}$$

And the auxiliary regression should be

$$\hat{\epsilon}^{IV} = \hat{\alpha}_1 + \hat{\alpha}_2 exper + \hat{\alpha}_3 expersq + \hat{\theta}_1 motheduc + \hat{\theta}_2 fatheduc + v \qquad \text{(auxiliary model)}$$

Wage example: J-test (R code for 2SLS residuals)

We have done the 2SLS estimation before, here is the `R` code (by using `ivreg::ivreg()` function):

```r
# load pkg
require(AER)
# specify model
mod_iv_mf <- formula(
  lwage ~ educ + exper + expersq
        | motheduc + fatheduc + exper + expersq)
# fit model
lm_iv_mf <- ivreg(formula = mod_iv_mf, data = mroz)
# summary of model fit
smry.ivmf <- summary(lm_iv_mf)
```

After the 2SLS estimation, we can obtain the IV residuals of the second stage:

```r
# obtain residual of IV regression, add to data set
mroz_resid <- mroz %>%
  mutate(resid_iv_mf = residuals(lm_iv_mf))
```

Wage example: J-test (new data set)

Data set with the 2SLS residuals

id	lwage	educ	exper	expersq	fatheduc	motheduc	resid_iv_mf
1	1.21	12	14	196	7	12	-0.0169
2	0.33	12	5	25	7	7	-0.6547
3	1.51	12	15	225	7	12	0.2690
4	0.09	12	6	36	7	7	-0.9254
5	1.52	14	7	49	14	12	0.3515
6	1.56	12	33	1089	7	14	0.2930
7	2.12	16	11	121	7	14	0.7127
8	2.06	12	35	1225	3	3	0.8300
9	0.75	12	24	576	7	7	-0.5728
10	1.54	12	21	441	7	7	0.2289

Showing 1 to 10 of 428 entries

Previous 1 2 3 4 5 … 43 Next

Wage example: J-test (run auxiliary regression)

R Code R result

We run the auxiliary regression with `R` code lines:

```
# set model formula
mod_jtest <- formula(resid_iv_mf ~ exper +expersq +motheduc +fatheduc)
# OLS estimate
lm_jtest <- lm(formula = mod_jtest, data = mroz_resid)
```

Then we can obtain the OLS estimation results.

Wage example: J-test (Restricted F-test)

As what we have done before, We conduct the restrict F-test for the auxiliary regression.

R Code R result

We will restrict jointly with $\theta_1 = \theta_2 = 0$, and using the R function `linearHypothesis()`:

```
# restricted F-test
restricted_ftest <- linearHypothesis(lm_jtest, c("motheduc = 0", "fatheduc = 0"),
# obtain the F statistics
restricted_f <- restricted_ftest$F[[2]]
```

The restricted F-statistics is 0.1870 (with round digits 4 here).

Wage example: J-test (calculate J-statistic by hand)

Finally, We can calculate J-statistic by hand or obtain it by using special tools.

- Calculate J-statistic by hand

```
# numbers of instruments
m <- 2
# calculate J statistics
(jtest_calc <- m*restricted_f)
```

```
[1] 0.37
```

- The calculated J-statistic is 0.3740 (with round digits 4 here).

Wage example: J-test (obtain J-statistic with tools)

Also, We can obtain J-statistic by using special tools.

R Code R result

- using tools of `linearHypothesis(., test = "Chisq")`

```
# chi square test directly
jtest_chitest <- linearHypothesis(
  lm_jtest,  c("motheduc = 0", "fatheduc = 0"),
  test = "Chisq")
# obtain the chi square value
jtest_chi <- jtest_chitest$Chisq[2]
```

- We obtain the J-statistic 0.3740 (with round digits 4 here). It's the same as what we have calculated by hand!

Wage example: J-test (adjust the freedoms)

Caution: In this case the p-Value reported by `linearHypothesis(., test = "Chisq")` is wrong because the degrees of freedom are set to 2, and the correct freedom should be $(m - k) = 1$.

- We have obtain the J statistics $\chi^{2^*} = 0.3740$, and its correct freedom is $(m - k) = 1$.
- Then we may compute the correct p-Value of this the J statistics (by using function `pchisq()` in `R`).

```
# correct freedoms
f <- m -1
# compute correct p-value for J-statistic
(pchi <- pchisq(jtest_chi, df = f, lower.tail = FALSE))
```

```
[1] 0.54
```

Wage example: J-test (the conclutions)

Now we can get the conclusions of J-test.

Since the p-value of J-test(0.5408)is larger than the criteria value 0.1, we can't reject the null hypothesis that both instruments are exogenous.

This means both instruments(`motheduc` and `fatheduc`) are **exogenous**.

17.6 Testing Regressor endogeneity

Regressor Endogeneity: the concepts

How can we test the regressor endogeneity?

Since OLS is in general more efficient than IV (recall that if Gauss-Markov assumptions hold OLS is BLUE), we don't want to use IV when we don't need to get the consistent estimators.

Of course, if we really want to get a consistent estimator, we also need to check whether the endogenous regressors are really **endogenous** in the model.

So we should test following hypothesis:

$$H_0 : \text{Cov}(X, \epsilon) = 0 \text{ vs. } H_1 : \text{Cov}(X, \epsilon) \neq 0$$

Regressor Endogeneity: Hausman test

`Hausman` tells us that we should use OLS if we fail to reject H_0. And we should use IV estimation if we reject H_0

Let's see how to construct a `Hausman test`. While the idea is very simple.

- If X is **exogenous** in fact, then both OLS and IV are consistent, but OLS estimates are more efficient than IV estimates.
- If X is **endogenous** in fact, then the results from OLS estimators are different, while results obtained by IV (eg. 2SLS) are consistent.

Hausman test: the idea

We can compare the difference between estimates computed using both OLS and IV.

- If the difference is **small**, we can conjecture that both OLS and IV are consistent and the small difference between the estimates is not systematic.
- If the difference is **large** this is due to the fact that OLS estimates are not consistent. We should use IV in this case.

Hausman test: the statistics

The **Hausman test** takes the following statistics form

$$\hat{H} = n\left[\hat{\boldsymbol{\beta}}_{\boldsymbol{IV}} - \hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right]' \left[\mathbf{Var}\left(\hat{\boldsymbol{\beta}}_{\boldsymbol{IV}} - \hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right)\right]^{-1} \left[\hat{\boldsymbol{\beta}}_{\boldsymbol{IV}} - \hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right] \xrightarrow{d} \chi^2(k)$$

- If $\hat{H}$ is less than the critical χ^2 value, we can not reject the null hypothesis, and the regressor should **not be endogenous**.
- If $\hat{H}$ is **larger** than the critical χ^2 value, the null hypothesis is rejected , and the regressor should **be endogenous**.

Wage example: Hausman test (the origin IV model)

Again, we use both $matheduc$ and $fatheduc$ as instruments for $educ$ in our IV model setting.

$$\begin{cases} \widehat{educ} = \hat{\gamma}_1 + \hat{\gamma}_2 exper + \hat{\beta}_3 expersq + \hat{\beta}_4 motheduc + \hat{\beta}_5 fatheduc & \text{(stage 1)} \\ lwage = \hat{\alpha}_1 + \hat{\alpha}_2 \widehat{educ} + \hat{\alpha}_3 exper + \hat{\alpha}_4 expersq + \hat{\epsilon} & \text{(stage 2)} \end{cases}$$

in `R`, we can use IV model diagnose tool to check the Hausman test results.

In fact, `R` function `summary(lm_iv_mf, diagnostics = TRUE)` by setting `diagnostics = TRUE` will give you these results.

Wage example: Hausman test (full model diagnose)

```
summary(lm_iv_mf, diagnostics = TRUE)
```

```
Call:
ivreg(formula = mod_iv_mf, data = mroz)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0986 -0.3196  0.0551  0.3689  2.3493

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.048100   0.400328    0.12   0.9044
educ         0.061397   0.031437    1.95   0.0515 .
exper        0.044170   0.013432    3.29   0.0011 **
expersq     -0.000899   0.000402   -2.24   0.0257 *

Diagnostic tests:
                 df1 df2 statistic p-value
Weak instruments   2 423     55.40  <2e-16 ***
Wu-Hausman         1 423      2.79   0.095 .
Sargan             1  NA      0.38   0.539
---
Signif. codes:
```

Wage example: the diagnosed conclusions

The results for the lwage equation are as follows:

- **(Wu-)Hausman test** for endogeneity: **barely reject** the null that the variable of concern is uncorrelated with the error term, indicating that `educ` is marginally endogenous. The Hausman statistics $\hat{H} = \chi^{2^*} = 2.79$, and its p-value is 0.095.
- **Weak instruments test**: **rejects** the null hypothesis(Weak instruments). At least one of these instruments(`motheduc` or `fatheduc`) is strong. The **restricted F-test** statistics $F_R^* = 55.4$, and its p-value is 0.0000.
- **Sargan overidentifying restrictions**(Instruments exogeneity J-test): **does not** reject the null. The extra instruments (`motheduc` and `fatheduc`) are valid (both are exogenous, and are uncorrelated with the error term).

Summary

- An **instrumental variable** must have two properties:
  - (1) it must be exogenous, that is, uncorrelated with the error term of the structural equation;
  - (2) it must be partially correlated with the endogenous explanatory variable.

> Finding a variable with these two properties is usually challenging.

- Though we can **never** test whether all IVs are **exogenous**, we can test that at least some of them are.
- When we have valid instrumental variables, we can test whether an explanatory variable is **endogenous**.
- The method of **two stage least squares** is used routinely in the empirical social sciences.

> But when instruments are poor, then 2SLS can be **worse** than OLS.

Exercise example 1: Card (1995)

In Card (1995) education is assumed to be endogenous due to omitted **ability** or **measurement error**. The standard wage function

$$\ln(wage_i) = \beta_0 + \beta_1 Educ_i + \sum_{m=1}^{M} \gamma_m W_{mi} + \varepsilon_i$$

is estimated by **Two Stage Least Squares** using a **binary instrument**, which takes value 1 if there is an **accredited 4-year public college in the neighborhood** (in the "local labour market"), 0 otherwise.

> It is argued that the presence of a local college decreases the cost of further education (transportation and accommodation costs) and particularly affects the schooling decisions of individuals with poor family backgrounds.

The set of exogenous explanatory regressors W includes variables like race, years of potential labour market experience, region of residence and some family background

Exercise example 1: Card (1995)

The dataset is available online at http://davidcard.berkeley.edu/data_sets.html and consists of 3010 observations from the National Longitudinal Survey of Young Men.

- **Education** is measured by the years of completed schooling and varies we between 2 and 18 years.

> To overcome the small sample problem, you might group the years of education into four educational levels: less than high school, high school graduate, some college and post-college education (a modified version of Acemoglu and Autor (2010) education grouping).

- Since the **actual labour market experience** is not available in the dataset, Card (1995) constructs a potential experience as **age-education-6**.

> Since all individuals in the sample are of similar age (24-34), people with the same years of schooling have similar levels of potential experience.

Exercise example 2: Angrist and Krueger (1991)

The data is available online at http://economics.mit.edu/faculty/angrist/data1/data/angkru1991 and consists of observations from 1980 Census documented in Census of Population and Housing, 1980: Public Use Microdata Samples.

The sample consists of men born in the United States between 1930-1949 divided into two cohorts: those born in the 30's (329509 observations) and those born in the 40's (486926 observations).

Angrist and Krueger (1991) estimate the conventional linear earnings function

$$\ln(wage_i) = \beta Educ_i + \sum_c \delta_c Y_{ci} + \sum_{s=1}^{S} \gamma_s W_{si} + \varepsilon_i$$

for each cohort separately, by 2SLS using the **quarter of birth** as an instrument for (assumed) endogenous **education**.

Exercise example 2: Angrist and Krueger (1991)

- They observe that individuals born earlier in the year (first two quarters) have less schooling than those born later in the year.

> It is a consequence of **the compulsory schooling laws**, as individuals born in the first quarters of the year reach ***the minimum school leaving age*** at the lower grade and might legally leave school with less education.

- The main criticism of Angrist and Krueger (1991) analysis, pointed out by Bound, Jaeger and Baker (1995) is that the quarter of birth is a **weak instrument**.
- A second criticism of Angrist and Krueger (1991) results, discussed by Bound and Jaeger (1996) is that quarter of birth might be **correlated** with unobserved ability and hence does not satisfy the **instrumental exogeneity condition**.

End Of This Chapter