Heteroskedasticity and Multicollinearity

Isabel Casas
Office: V5-206a-2
icasas@sam.sdu.dk

Outline

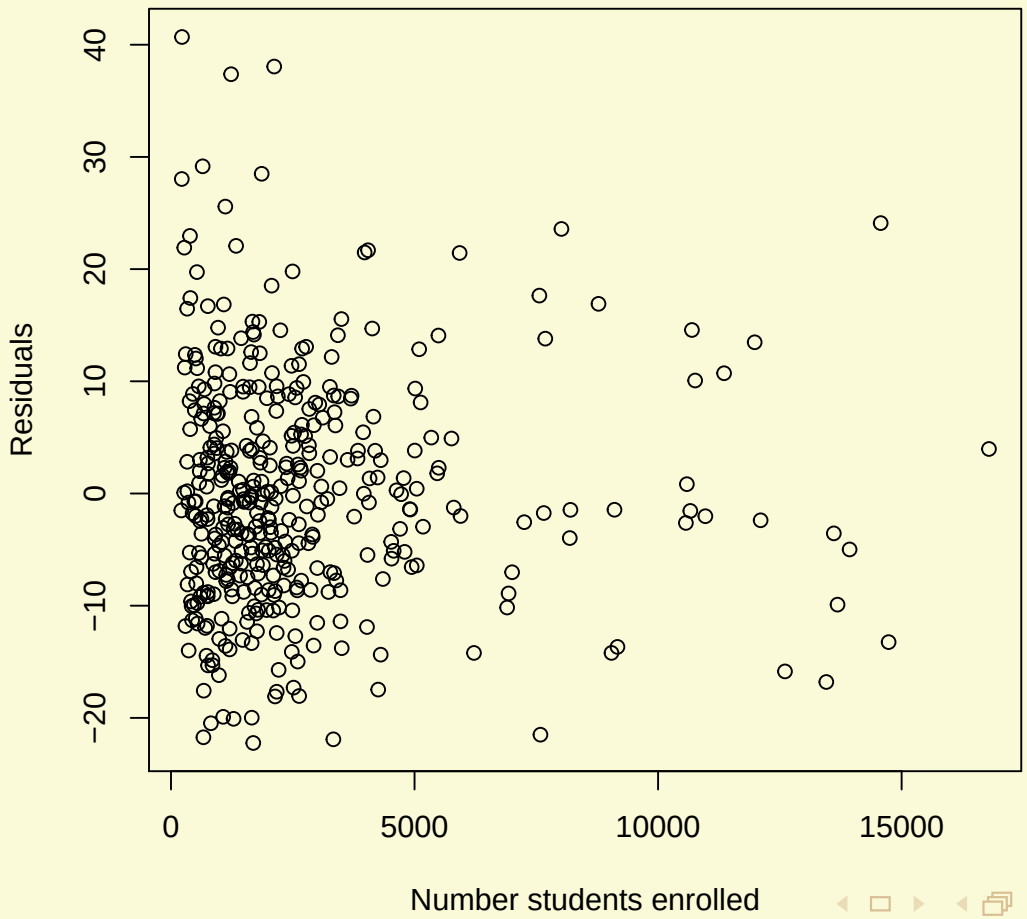
- Reminder
- Solve heteroskedasticity
  - WLS (more efficient than OLS)
  - Heteroskedasticity robust standard errors
- Multicollinearity

Example: School performance and school size (UG)

```
Call:
lm(formula = maths ~ 1 + compensation + staff + enrolled, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-22.235  -7.008  -0.807   6.097  40.689

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.2740214  6.1137938   0.372    0.710
compensation  0.0004586  0.0001004   4.570 6.49e-06 ***
staff         0.0479199  0.0398140   1.204    0.229
enrolled     -0.0001976  0.0002152  -0.918    0.359
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 10.24 on 404 degrees of freedom
Multiple R-squared: 0.05406,  Adjusted R-squared: 0.04704
F-statistic: 7.697 on 3 and 404 DF,  p-value: 5.179e-05
```

Example: School performance and school size (UG)

Figure: Plot of residuals vs enrolled



Violation of OLS.3

3. OLS. 3 $E(\boldsymbol{\epsilon}^2|X) = \sigma^2 h(X) \neq \sigma^2$. Heteroskedasticity: the variance of the errors depends on X.
   - The value of the asymptotic variance of $\hat{\beta}$ is changed.
   - Luckily, the consistency of the estimator is preserved, as well as the asymptotic normality.
   - Solution: WLS, or use robust standard errors for the tests.

Consequences of heteroskedasticity:

- The OLS $\hat{\beta}$ is consistent, linear, however inefficient.
- The $\widehat{Var}(\hat{\beta})$ is biased. So the t-test and F-test are not reliable.

Heteroskedasticity

A simple example:

$$earnings = \beta_0 + \beta_1\ male + \boldsymbol{\epsilon} \qquad male = \begin{cases} 1 & \text{if male} \\ 0 & \text{if female} \end{cases}$$

What is the average earnings of a female?

Heteroskedasticity

A simple example:

$$earnings = \beta_0 + \beta_1\ male + \boldsymbol{\epsilon} \qquad male = \begin{cases} 1 & \text{if male} \\ 0 & \text{if female} \end{cases}$$

What is the average earnings of a female? β_0

What is the average earnigns of a male?

Heteroskedasticity

A simple example:

$$earnings = \beta_0 + \beta_1\ male + \boldsymbol{\epsilon} \qquad male = \begin{cases} 1 & \text{if male} \\ 0 & \text{if female} \end{cases}$$

What is the average earnings of a female? β_0

What is the average earnigns of a male? $\beta_0 + \beta_1$

A priori, could there be heteroskedasticity in this example?

Heteroskedasticity

A simple example:

$$earnings = \beta_0 + \beta_1\ male + \boldsymbol{\epsilon} \qquad male = \begin{cases} 1 & \text{if male} \\ 0 & \text{if female} \end{cases}$$

What is the average earnings of a female? β_0

What is the average earnigns of a male? $\beta_0 + \beta_1$

A priori, could there be heteroskedasticity in this example? It is common to find both men and women earning the low salaries but there are not as many women earning high salaries. So one would expect less variance in the earnings of females $\Rightarrow$ Heteroskedasticity

Heteroskedasticity

Because we assume $E(\boldsymbol{\epsilon}|X) = 0$

$$Var(\boldsymbol{\epsilon}|X) = E(\boldsymbol{\epsilon}'\boldsymbol{\epsilon}|X) = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \dots & & & \dots \\ 0 & 0 & \dots & \sigma_n^2 \end{pmatrix} = \boldsymbol{\Omega}$$

Those individuals j that are females will have a different income variance than those who are males.

So $Var(\boldsymbol{\epsilon}|X)$ depends on the variable X.

Heteroskedasticity

From the proof of asymptotic normality:

$$\sqrt{n}(\hat{\beta} - \beta) = (\frac{1}{n}X'X)^{-1}(\frac{\sqrt{n}}{n}X'\boldsymbol{\epsilon})$$

We are interested in the asymptotic behaviour and we know that plim $\frac{1}{n}X'X = E(X'X) = \Sigma$.

Similarly, plim $\frac{1}{n}X'\boldsymbol{\epsilon}\boldsymbol{\epsilon}'X = \boldsymbol{\Omega}$

We know that $E(\hat{\beta}) = \beta +$ something asymptotically zero.

So the asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta)$ is:

$$\text{plim } \left(\frac{1}{n}X'X\right)^{-1}\left(\frac{1}{\sqrt{n}}X'\boldsymbol{\epsilon}\frac{1}{\sqrt{n}}\boldsymbol{\epsilon}'X\right)\left(\frac{1}{n}X'X\right)^{-1} =$$

$$\Sigma^{-1}\Sigma\boldsymbol{\Omega}\Sigma^{-1} = \boldsymbol{\Omega}\Sigma^{-1}$$

Heteroskedasticity

- If the error term is homokedastic

$$\mathbf{\Omega} = \sigma^2 \mathbf{I}$$

  so we find the standard errors by estimating σ^2 by s^2.
- Otherwise, we have to estimate each element of $\mathbf{\Omega}$
- File Cov_ HC.pdf describes the most common estimators.

Solving heteroskedasticity: solution I, robust se

The variance of the OLS estimator is

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}) = & \quad E\left[(\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}}))(\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}}))'\right] = E\left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\right] \\ & = E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right] \end{aligned}$$

- We have something difference from σ^2 in the blue term and therefore the variance is not correctly estimated by OLS.
- The robust standard errors are the square roots of the diagonal of the matrix above
- We can the perform a heteroskedasticity-robust t-test $H_0 : \beta_i = 0$:

$$T = \frac{\hat{\beta}_i}{\text{robust standard error of } \hat{\beta}_i} \sim t_{n-k-1}$$

Solving heteroskedasticity: solution I, robust se

- The F-test is not accurate in presence of heteroskedasticity.
- Instead we can use other tests like the Wald and the LM tests.

Matrix form notation of the null hypothesis:

$$H_0: \ \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$$

$$H_1: \ H_0 \text{ is not satisfied}$$

- $\mathbf{R}$ is a $q \times (k+1)$ matrix
- $\boldsymbol{\beta}$ is a $(k+1) \times 1$ vector
- $\mathbf{r}$ is a $q \times 1$ vector.

Solving heteroskedasticity: solution I, robust se

Example:

$$y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + \beta_4 X_{4j} + \beta_5 X_{5j} + \epsilon_j$$

and we want to test the restrictions $H_0: \ \beta_2 = 0, \ \beta_5 = 0$. There are $q = 2$ restriction, the hypothesis:

$$\underbrace{\begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}}_{\mathbf{R}} \overbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix}}^{\boldsymbol{\beta}} = \underbrace{\begin{pmatrix} 0 \\ 0 \end{pmatrix}}_{\mathbf{r}}$$

$$H_0: \ \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$$

Solving heteroskedasticity: solution I, robust se

The test statistics:

$$(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{R}\boldsymbol{\beta})'(\mathbf{R}\hat{\mathbf{V}}\mathbf{R}')(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{R}\boldsymbol{\beta}) \sim^{a} \chi^2_q$$

where $\hat{\mathbf{V}}$ is the estimated asymptotic variance of $\hat{\boldsymbol{\beta}}$.
Intuition proof:
The distribution of the OLS estimator:

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \hat{\mathbf{V}}) \Rightarrow \mathbf{R}\hat{\boldsymbol{\beta}} \sim N(\mathbf{R}\boldsymbol{\beta}, \mathbf{R}\hat{\mathbf{V}}\mathbf{R}')$$

The test statistics:

$$\mathbf{W} = (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{R}\boldsymbol{\beta}) = (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}) \sim N(\mathbf{0}, \mathbf{R}\hat{\mathbf{V}}\mathbf{R}')$$

Exercise: Dividing the Wald test statistics by its variance, obtain the $F - test$ expression when you assume homokedasticity

Exercise (30 minutes)

Example_HypTesting.pdf

Possible violations of the OLS assumptions

2. OLS.2 Rank $E(\mathbf{X}'\mathbf{X}) < k+1$ fails: the number of normal equations is less than the number of parameters β
   - Consequence: the parameters cannot be identified, we cannot find an unique solution
   - Solution: identify the dependent variables in $\mathbf{X}$ and remove them from the model

Multicollinearity

- Perfect multicollinearity appears when one regressor is a perfect linear combination of the other regressors.
  - It is impossible to identify all the parameters
  - The software will not give us an answer.
- Imperfect multicollinearity appears when one regressor is very highly correlated with the other regressors
  - We can estimate the parameters but some might be imprecise.

Multicollinearity: Examples

- We include a variable fraction and the same values in percentage. For example: the fraction of foreigners in class $5/50 = 1/10$ and the percentage 10%.
  - Solution: remove one of them.
- The dummy variable trap. We include a dummy variable to indicate male and another one to indicate female.
  - Solution: if we have G categories of a variable, only creat $G - 1$ dummy variables
- We assume our data set has male and female individuals and we would like to categorise them diffeincomely. We include only one dummy variable to avoid de dummy variable trap. If by chance our set does not have any females, we are creating multicollinearity

Sources of endogeneity

- We are interested in dealing with the problem of endogeneity.
- A variable is *endogenous* when it depends of other variables, and therefore it can be explained by then, or when it is correlated with the error $\boldsymbol{\epsilon}$.
- So if $E(X_j\boldsymbol{\epsilon}) = \text{cov}(X_j, \boldsymbol{\epsilon}) \neq 0 \Rightarrow$ Endogeneity.
- If the model cannot be changed to remove the endogeneity, then the OLS main condition for consistency is not satisfied.
- Result: inconsistent OLS estimators

Sources of endogeneity

- Three common causes of endogeneity:
  - *Omission of relevant regressors*.
  - *Measurement errors*. It occurs when Y or certain variables X_j are erratically observed.
  - *Simultaneous causality*. This occurs when one or more regressors are determined simultaneously with the dependent variable Y.

Outline

- Omitted variables
  - Consequences: Biased and inconsistent OLS estimator
  - Solution: Proxy variables or instrumental variables (IV)
- Measument errors
  - Consequences: Biased and inconsistent estimators
  - Solution: IV
- Simultaneity
  - Consequences: Biased and inconsistent estimators
  - Solution: IV
- A more general solution:
  - 2SLS estimator and its properties

Source of Endogeneity: Omitted Variables

- Solution I: Proxy variables
- Solution II: If we cannot find good proxy variables $\Rightarrow$ IV
- Definition of instrument
- IV estimator (there is only one instrument per endogenous variable)

Do not just too quickly

Omitted variables

Let assume that the *correctly* specified model looks like:

$$Y = X\beta + \gamma\ q + \boldsymbol{\epsilon} = \beta_0 + \beta_1 X_1 + \ldots \beta_k X_k + \gamma q + \boldsymbol{\epsilon}$$

However as q is unknown and we estimate the *incorrect* model:

$$Y = \beta_0 + \beta_1 X_1 + \ldots \beta_k X_k + \eta$$

where

$$\eta = q\gamma + \boldsymbol{\epsilon}$$

Let assume X_k is correlated with q, then $\text{cov}(X_k, \eta) \neq 0$
OLS condition is not satisfied.

Omitted variables

The OLS estimator of the *incorrect* model is:

$$\begin{aligned}\hat{\beta}^{inc} =& (X'X)^{-1}X'Y = (X'X)^{-1}X'\left(X\beta + \gamma q + \boldsymbol{\epsilon}\right)\\ =& \beta + (X'X)^{-1}X'\ q\ \gamma + (X'X)^{-1}X'\boldsymbol{\epsilon}\end{aligned}$$

Therefore, unless the regressors X_k and q are orthogonal and therefore $X'q = 0$, we have that $\hat{\beta}^{inc}$ is biased:

$$E(\hat{\beta}^{inc}) = \beta + (X'X)^{-1}X'\ q\ \gamma$$

and inconsistent

$$plim(\hat{\beta}_k^{inc}) = \beta_k + \gamma\frac{\text{cov}(X_k, q)}{Var(X_k)}$$

Omitted variables: solutions

- First and more intuitively, if q is unknown, find a variable that explains it but it is known: *proxy* variable (z).
- If this does not work, use an *instrument* to scoop the endogeneity out of the endogenous variable.

Omitted variables: proxy variables

There is an economic theory that says that the expected wage depends on the education, experience and ability. So we should have something like

$$\log(wage) = \beta_0 + \beta_1\ exper + \beta_2\ educ + \gamma\ ability + \boldsymbol{\epsilon} \qquad (1)$$

but $ability$ is unobservable.

So if we estimate

$$\log(wage) = \beta_0 + \beta_1\ exper + \beta_2\ educ + \eta \qquad (2)$$

because $E(educ\ ability) \neq 0$ then $E(educ\ \eta) \neq 0$ and

$\hat{\beta}$ is inconsistent and biased.

Omitted variables: proxy variables

variable name	variable type	variable label
wage	int	monthly earnings
hours	byte	average weekly hours
iq	int	IQ score
kww	byte	knowledge of world work score
educ	byte	years of education
exper	byte	years of work experience
tenure	byte	years with curincome employer
age	byte	age in years
married	byte	=1 if married
black	byte	=1 if black
south	byte	=1 if live in south
urban	byte	=1 if live in SMSA
sibs	byte	number of siblings
brthord	byte	birth order
meduc	byte	mother's education
feduc	byte	father's education
lwage	float	log(wage)

Omitted variables: proxy variables

Let us look at Example 4.3 of Wooldridge (page 65) and run the regression using IQ as the proxy variable of $ability$.

```
> data<-read.table("nls80.txt", header=T, na=".")
> names(data)
 [1] "wage"    "hours"   "iq"      "kww"     "educ"    "exper"
 "tenure"  "age"     "married" "black"   "south"   "urban"
 "sibs"    "brthord" "meduc"   "feduc"   "lwage"
> lwage.1<-lm(lwage~exper+tenure+married+south+urban+black+educ,data=data)
```

Omitted variables: proxy variables

```
> summary(lwage.1)
Call:
lm(formula = lwage ~ exper + tenure + married + south + urban +
    black + educ, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.98069 -0.21996  0.00707  0.24288  1.22822

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.395497   0.113225  47.653  < 2e-16 ***
exper        0.014043   0.003185   4.409 1.16e-05 ***
tenure       0.011747   0.002453   4.789 1.95e-06 ***
married      0.199417   0.039050   5.107 3.98e-07 ***
south       -0.090904   0.026249  -3.463 0.000558 ***
urban        0.183912   0.026958   6.822 1.62e-11 ***
black       -0.188350   0.037667  -5.000 6.84e-07 ***
educ         0.065431   0.006250  10.468  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.3655 on 927 degrees of freedom
Multiple R-squared: 0.2526,     Adjusted R-squared: 0.2469
F-statistic: 44.75 on 7 and 927 DF,  p-value: < 2.2e-16
```

Omitted variables: proxy variables

```
> lwage.2<-lm(lwage~exper+tenure+married+south+urban+black+educ+iq,data=data)
> summary(lwage.2)

Call:
lm(formula = lwage ~ exper + tenure + married + south + urban +
    black + educ + iq, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.01203 -0.22244  0.01017  0.22951  1.27478

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.1764390  0.1280006  40.441  < 2e-16 ***
exper        0.0141458  0.0031651   4.469 8.82e-06 ***
tenure       0.0113951  0.0024394   4.671 3.44e-06 ***
married      0.1997644  0.0388025   5.148 3.21e-07 ***
south       -0.0801695  0.0262529  -3.054 0.002325 **
urban        0.1819463  0.0267929   6.791 1.99e-11 ***
black       -0.1431253  0.0394925  -3.624 0.000306 ***
educ         0.0544106  0.0069285   7.853 1.12e-14 ***
iq           0.0035591  0.0009918   3.589 0.000350 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.3632 on 926 degrees of freedom
Multiple R-squared: 0.2628,     Adjusted R-squared: 0.2564
F-statistic: 41.27 on 8 and 926 DF,  p-value: < 2.2e-16
```

Omitted variables: proxy variables

model	β_0	exper	tenure	married	south	urban	black	educ	iq
No IQ	5.40	0.014	0.012	0.199	-0.091	0.184	-0.188	0.065	
W IQ	5.18	0.014	0.011	0.2	-0.08	0.182	-0.143	0.054	0.0036

Questions:

- What variables is IQ correlated with?
- Why is the effect of education smaller when IQ is included?
- How much of the variation of $\log(wage)$ is explained by IQ?
- Is IQ a good proxy variable?

Omitted variables: proxy variables

z is a proxy variable of the omitted q variable, if the following two requirements are satisfied:

1. The proxy variable (z) should be redundant
   - $E(\log(wage)|X, q, z) = E(\log(wage)|X, q)$
   - Example: The wage does not depend on the IQ. The variable IQ controls for $ability$ and $educ$.
2. It should explain all the effect of $ability$ over the other regressors.
   - $ability = \delta_0 + \delta_1 X_1 + \ldots + \delta_k X_k + \rho\ z + \nu$, then $\delta_1 = \delta_2 = \ldots = \delta_k = 0, \quad \rho \neq 0$. Once z is taken into account then X is not related to q.
   - Example: Once IQ is accounted for, then $educ$ is not correlated with $ability$

Omitted variables: proxy variables

If conditions 1 and 2 are satisfied : Including z in the regression (endogeneity is gone) $\Rightarrow$ OLS estimators are **consistent** and **asymptotically normal**.

If condition 2 is not satisfied: z is an imperfect proxy (still endogenous model) $\Rightarrow$ the OLS estimators are **inconsistent**.
Solution: Instrumental Variables.

Even if the proxy is imperfect, including it might still:

- Reduce asymptotic bias of estimators
- Reduce variance of estimators

Also there could be more than one proxy for each variable.

Exercise: Run the regression including iq and kww.

IV for Omitted Variables

We have the MLR

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_{k-1} X_{k-1} + \beta_k X_k + \epsilon$$

such that

- $E(\epsilon) = 0$,
- $\text{cov}(X_j, \epsilon) = 0$ for $j = 1, \ldots, k-1$ but
- $\text{cov}(X_k, \epsilon) \neq 0 \Rightarrow$ Endogenous variable

It might not be possible to remove the endogeneity of X_k with proxy variables, then so we use an instrument z for X_k in our model.

What is an instrument?

An *instrument* or an *instrumental variable* for X_k is an observable variable z not included in the model that satisfies:

1. Relevant: X_k is partially correlated with z once the effects of the other exogenous variables is removed.
   - $X_k = \delta_0 + \delta_1 X_1 + \ldots + \delta_{k-1} X_{k-1} + \pi_1 \ z + \eta$ (reduced eq)
   - $H_0 : \ \pi_1 \neq 0$
2. Exogenous: $\text{cov}(z, \epsilon) = 0$: z is exogenous and uncorrelated with the omitted variables
   - Difficult to prove. It is assumed by definition of the economical model or tested with other proxy variable of the omitted variable.

IV for Omitted Variables

Substitute the reduced equation into the main equation:

$$Y = \alpha_0 + \alpha_1 X_1 + \ldots + \alpha_{k-1} X_{k-1} + \lambda_1 \; z + \nu$$

where $\nu = \epsilon + \beta_k \eta$ and $\alpha_j = \beta_j + \beta_k \delta_j$ and $\lambda_1 = \beta_k \pi_1$

All the variables of this model are uncorrelated with ν and therefore the OLS estimates $\hat{\alpha}_j, \hat{\lambda}_1$ are consistent.

The IV estimator

The initial model

$$Y = X\beta + \epsilon \quad \text{with } X_k \text{ endogenous}$$

Let

$$Z = \begin{pmatrix} 1 & X_{1,1} & \dots & X_{k-1,1} & z_1 \\ 1 & X_{1,2} & \dots & X_{k-1,2} & z_2 \\ \vdots & & & \vdots & \\ 1 & X_{1,n} & \dots & X_{k-1,n} & z_n \end{pmatrix}$$

The consistent IV estimator is:

$$\hat{\beta}^{IV} = (Z'X)^{-1}Z'Y$$

Education example

In the example:

$$\log(wage) = \beta_0 + \beta_1\ educ + \epsilon,$$

where $\text{cov}(educ, \epsilon) \neq 0$ because $educ$ is correlated with $ability$.

We should find an instrumental variable z for edu that is:

- Relevant: correlated with $educ$
- Exogenous: uncorrelated with $ability$

Question: Is IQ a good instrumental variable for $educ$?

IV for edu

Possible instrumental variables z for *educ* may be:

- $z =$ last digit of the social security number?
- $z =$ IQ?
- $z =$ mother's education?
- $z =$ number of siblings?

Hint: Is z correlated with $educ$? And with $ability$?

IV for edu

Possible instrumental variables z for *educ* may be:

- $z =$ last digit of the social security number?
- $z =$ IQ?
- $z =$ mother's education?
- $z =$ number of siblings?

IV for edu

Possible instrumental variables z for *educ* may be:

- $z =$ last digit of the social security number? it satisfies $\text{cov}(z, abil) = 0$. However, $\text{cov}(z, educ) = 0$. Therefore, it cannot be an instrumental variable.
- $z =$ IQ?
- $z =$ mother's education?
- $z =$ number of siblings?

IV for edu

Possible instrumental variables z for *educ* may be:

- $z =$ last digit of the social security number? it satisfies $\text{cov}(z, abil) = 0$. However, $\text{cov}(z, educ) = 0$. Therefore, it cannot be an instrumental variable.
- $z =$ IQ? it is correlated with $ability$ so it cannot be an instrumental variable.
- $z =$ mother's education?
- $z =$ number of siblings?

IV for edu

Possible instrumental variables z for *educ* may be:

- $z =$ last digit of the social security number? it satisfies $\text{cov}(z, abil) = 0$. However, $\text{cov}(z, educ) = 0$. Therefore, it cannot be an instrumental variable.
- $z =$ IQ? it is correlated with $ability$ so it cannot be an instrumental variable.
- $z =$ mother's education? $\text{cov}(z, educ) > 0$, however it might happen that the mother's education influences the ability of the son in his early education: $\text{cov}(z, ability) \neq 0$.
- $z =$ number of siblings?

IV for edu

Possible instrumental variables z for *educ* may be:

- $z =$ last digit of the social security number? it satisfies $\text{cov}(z, abil) = 0$. However, $\text{cov}(z, educ) = 0$. Therefore, it cannot be an instrumental variable.
- $z =$ IQ? it is correlated with $ability$ so it cannot be an instrumental variable.
- $z =$ mother's education? $\text{cov}(z, educ) > 0$, however it might happen that the mother's education influences the ability of the son in his early education: $\text{cov}(z, ability) \neq 0$.
- $z =$ number of siblings? The academic education decreases as the number of siblings increases: $\text{cov}(z, educ) < 0$. If we assume no correlation between n. of siblings and $ability$, then z could be used as an IV.

Example 5.1 of Wooldridge

Angrist and Krueger (1991) choose the dummy variable, birth in first quarter of the year, as the IV for education.

- It is clearly uncorrelated with ability
- Is it related with education?
- The t-test is not very convincing

The main idea behind this example is to show that it can be difficult to find an IV too.

Are instruments of education useful?

- Carneiro and Heckman (2002) argue that they are either not exogenous or weak instruments for education (CH_table).
- First column shows the estimated correlation between the instrument and education. If low (compared to the standard error in paincomeheis), then the instrument is week
- Second column shows the correlation between the instrument and (unobserved) ability. If high (compared to standard error), instrument is not exogenous.
- Implications? Estimates of the education parameter can be severely biased and highly imprecise.

Source of Endogeneity: Measurement Error

- The dependent variable is erratically measured.
  - Because this error is uncorrelated with the regressors $\Rightarrow$ OLS is fine
- One of the regressors is erratically measured.
  - If the error is uncorrelated with true variable $\Rightarrow$ OLS is fine
  - If the error is correlated with the true variable, then we need an IV

Measurement error in the dependent variable

Let define the *correctly* specified model:

$$y^{\star} = \beta_0 + \beta_1 X_1 + \ldots \beta_k X_k + \epsilon$$

We only observe $y = y^{\star} + e_0$, then the *incorrect* model:

$$y = \beta_0 + \beta_1 X_1 + \ldots \beta_k X_k + \underbrace{\epsilon + e_0}_{\eta}$$

If $E(X_j \eta) = 0$ for $j = 1, \ldots, k$
(No correlation between regressors and committed error, e_0)
$\Rightarrow$ the OLS estimators are consistent.

Measurement error in the dependent variable

Example:

$$Beer\_consumption = \beta_0 + \beta_1\ income + \beta_2\ student + \beta_3\ price + \epsilon$$

We do a self-answered survey. We might get a measurement error in the variable $Beer\_consumption^{\star}$, the "bragging" effect, underestimation, etc.

- People who never drinks will often report right results. So $e_0 = 0$.
- People who drink, might not report right results: correlation between $Beer\_consumption^{\star}$ and e_0

Measurement error in a regressor

Let us assume that $X_k^\star$ is the true value of this regressor. However, we observe $X_k = X_k^\star + e_k$ (for example beer price).

The *correctly* specified model:

$$y = \beta_0 + \beta_1 X_1 + \ldots \beta_k X_k^\star + \epsilon$$

The *incorrect* model:

$$y = \beta_0 + \beta_1 X_1 + \ldots \beta_k X_k + \underbrace{\epsilon + e_k}_{\eta}$$

Measurement error in a regressor

Assume $E(e_k) = 0$. Otherwise $\beta_k e_k$ is added to the intercept.

Case 1 : $\text{cov}(X_k, e_k) = 0$, **no endogeneity in our model**.

- The OLS estimators are consistent with greater variance due to the error

Case 2 : $\text{cov}(X_k^\star, e_k) = 0$

- The true variable is uncorrelated with the error, but the variable we use it is:

$$\text{cov}(X_k, e_k) = E(X_k e_k) = E(X_k^\star e_k) + E(e_k^2) = Var(e_k)$$

- Therefore, there is **endogeneity** in our model.
- The OLS estimators are biased and inconsistent
- Solution: Instrumental Variables

IV when there are measurement errors

- $\text{cov}(X_j, \epsilon) = 0$ for $j = 1, \ldots, k-1$
- $\text{cov}(e_k, X_k^\star) = 0 \Rightarrow \text{cov}(e_k, X_k) \neq 0$.
- Therefore we need an IV for X_k.
- Find an instrument z and construct:

$$Z = \begin{pmatrix} 1 & X_{1,1} & \ldots & X_{k-1,1} & z_1 \\ 1 & X_{1,2} & \ldots & X_{k-1,2} & z_2 \\ \vdots & & & \vdots & \\ 1 & X_{1,n} & \ldots & X_{k-1,n} & z_n \end{pmatrix}$$

- The consistent IV estimator is:

$$\hat{\beta}^{IV} = (Z'X)^{-1}Z'Y$$

Source of Endogeneity: Simultaneous Equations

- The dependent and independent variables are entwined
- Solution: Instrumental variables

Simultaneity

$$homocides = \alpha_1 \ police + \beta_{1,0} + \beta_{1,1} \ family\_income + \epsilon_1$$
$$police = \alpha_2 \ homocides + \beta_{2,0} + \text{ other factors} + \epsilon_2$$

- An increase in the number of $homocides$ will affect the number of $police$ agents in the streets
- Therefore, there is correlation between ϵ_1 and $police \Rightarrow$ Endogeneity
- Solution: Instrumental Variables

IV with simultaneous equations

We have the structural model of two equations:

$$y_1 = \alpha_1 y_2 + \beta_1 w_1 + \epsilon_1$$
$$y_2 = \alpha_2 y_1 + \beta_2 w_2 + \epsilon_2$$

- w_1, w_2 are exogenous
- We consider the intercept zero for simplicity
- Parameters $\alpha_1, \alpha_2, \beta_1, \beta_2$ are called *structural parameters*
- ϵ_1, ϵ_2 are called *structural errors*

IV with simultaneous equations

We want to estimate y_2. We substitute y_1 in the second equation and regress on y_2.

$$\begin{aligned} y_2 =& \alpha_2(\alpha_1\ y_2 + \beta_1\ w_1 + \epsilon_1) + \beta_2\ w_2 + \epsilon_2 \\ & \Rightarrow \\ (1 - \alpha_2\alpha_1)y_2 =& \alpha_2\beta_1\ w_1 + \beta_2 w_2 + \alpha_2\epsilon_1 + \epsilon_2 \end{aligned}$$

We have to assume that $\alpha_1\alpha_2 \neq 1$, so:

$$y_2 = \pi_{21}\ w_1 + \pi_{22}\ w_2 + \nu_2$$

IV with simultaneous equations

$$y_2 = \pi_{21}\ w_1 + \pi_{22}\ w_2 + \nu_2 \quad \text{reduced eq}$$

where

- $\pi_{21} = \alpha_2\beta_1/(1-\alpha_2\alpha_1)$
- $\pi_{22} = \beta_2/(1-\alpha_2\alpha_1)$
- $\nu_2 = (\alpha_2\epsilon_1 + \epsilon_2)/(1-\alpha_2\alpha_1)$
- ν_2 is a linear function of ϵ_1 and ϵ_2, then ν_2 is uncorrelated with w_1 and w_2.
- We can estimate π_{12} y π_{22} by OLS.
- If $\alpha_2 = 0$ then there is no simultaneity (test this).
- If $\alpha_2 \neq 0$, then there is endogeneity and we need IV for y_1.

2 Stages Least Squares

- When each endogenous variable has more than one IV
- Statistical properties of the 2SLS
- Example with simultaneous equations

Two steps least squares (2SLS)

$$Y = X\beta + \epsilon \quad \text{with } X_k \text{ endogenous}$$

Assuming that we have valid instruments: $z_1, z_2, \ldots, z_m$ for X_k, i.e:

- $\text{cov}(z_j, \epsilon) = 0$ for $j = 1, \ldots, m$
- Each z_j is partially correlated with X_k
- Out of all linear combinations of z_j the 2SLS method used the most highly correlated with X_k
- $X_k = \delta_0 + \delta_1 X_1 + \ldots + \delta_{k-1} X_{k-1} + \pi_1 \ z_1 + \ldots + \pi_m \ z_m + \eta$
- As z is uncorrelated with ϵ then $\hat{X}_k = \hat{\delta}_0 + \hat{\delta}_1 X_1 + \ldots + \hat{\delta}_{k-1} X_{k-1} + \hat{\pi}_1 \ z_1 + \ldots + \hat{\pi}_m \ z_m$ isn't either.
- So $\hat{X}_k$ can be used as an instrument of X_k

Two steps least squares (2SLS)

We could estimate the parameters of interest with two regressions:

[Stage 1] Estimate X_k by OLS:

- $X_k = \delta_0 + \delta_1 X_1 + \ldots + \delta_{k-1} X_{k-1} + \pi_1 \ z_1 + \ldots + \pi_m \ z_m + \eta$
- As z is uncorrelated with ϵ then $\hat{X}_k = \hat{\delta}_0 + \hat{\delta}_1 X_1 + \hat{\delta}_{k-1} X_{k-1} \hat{\pi}_1 \ z_1 + \ldots + \hat{\pi}_m \ z_m$ isn't either.

[Stage 2] Substitute X by $\hat{X}$ in the original equation:

- $\hat{\beta}^{2SLS} = (\hat{X}'X)^{-1}\hat{X}'Y$

The 2SLS estimator is the same than the IV estimator if we have only one IV.

If there are more than one endogenous variable, we have more regressions in Stage 1.

2SLS in R

- First time, install the package sem.

  ```
  > install.packages("sem")
  ```

- Include this library with

  ```
  > library(sem)
  ```

- Look at the help of function tsls

  ```
  > ?tsls
  ```

Statistical properties of 2SLS

Let have the model

$$Y = X\beta + \epsilon$$

where $X = (\mathbf{1}, X_1, \ldots, X_k)'$.

- There might be several endogenous variables amongst the regressors (correlated with ϵ).
- We have one or more IV for each endogenous variable
- There exists $Z = (Z_1, Z_2, \ldots, Z_l)'$
- Any exogenous elements of X are included in Z, plus the instrumental variables of the endogenous variables.

Consistency of 2SLS

Assumption 2SLS.1: IV is exogenous

- $E(Z'\epsilon) = 0$.

Assumption 2SLS.2: Multicollinearity

- rank $E(Z'Z) = l$: this is automatically satisfy because all the variables in Z are lineally independent
- rank $E(Z'X) = k + 1$: It is necessary $l \geq k + 1$ and Z and X are appropriately correlated

Asymptotic normality of 2SLS

Assumption 2SLS.3: Homokedasticity, $E(\epsilon^2|Z) = \sigma^2$

- $E(\epsilon^2 Z'Z) = \sigma^2 E(Z'Z)$

Theorem

Under Assumptions 2SLS.1–2SLS.3,

$$\sqrt{n}(\hat{\beta}^{2SLS} - \beta) \to^d N\left(0, \frac{\sigma^2}{E(X'Z)[E(Z'Z)]^{-1}E(Z'X)}\right)$$

as $n \to \infty$

Residuals of 2SLS

The 2SLS residuals are $\hat{\epsilon}_i = y_i - X_i\hat{\beta}^{2SLS}$ for $i = 1, 2, \ldots, n$.

We need them to estimate σ^2:

$$\hat{\sigma}^2 = \frac{1}{n-k-1}\sum \hat{\epsilon}_i^2$$

The variance–covariance matrix is

$$Var(\hat{\beta}^{2SLS}) = \frac{\hat{\sigma}^2}{\hat{X}'\hat{X}}$$

where $\hat{X}$ is estimated in Stage 2.

The standard error is the standard deviation of the diagonal of the variance–covariance matrix.

Covariance wit heteroskedasticity

If $E(\epsilon^2) \neq \sigma^2)$, the robust variance–covariance matrix is

$$Var(\hat{\beta}^{2SLS}) = (\hat{X}'\hat{X})^{-1} \sum \epsilon_i^2 \hat{X}_i' \hat{X}_i (\hat{X}'\hat{X})^{-1}$$

Hypothesis test of 2SLS estimates

- Confidence intervals
- t-statistics on single variables
- They are obtained as usual, using the standard errors or robust standard errors as necessary
- Multiple restrictions of the form $H_0: \ R\beta = r$ are tested with the Wald statistics
- LM test (page 99–101 in Wooldridge).

Pitfalls with 2SLS

- The 2SLS estimator is never unbiased
- For example, in a simple model with only one explanatory variable X_1 whose instrument is z, the asymptotic bias is:

$$\text{plim } \hat{\beta}_1^{2SLS} = \beta_1 + \frac{\text{cov}(z, \epsilon)}{\text{cov}(z, X_1)}$$

- If $\text{cov}(z, \epsilon) = 0 \Rightarrow$ consistent estimator
- Otherwise, if $\text{cov}(z, X_1)$ is small (z is a weak instrument of X_1) then the inconsistency can be large

Pitfalls with 2SLS

- The standard errors tend to be large (imprecise estimator), especially if the instrument is weak. See (AngristKrueger_table.pdf).
- Bias in small sample is going to be large we have a weak instrument $\Rightarrow$ it is important to test for the strength of the instrument. $\Rightarrow$ important to test for strength of instrument in the first stage of 2SLS:

$$H_0 : \ \pi_1 = \ldots = \pi_m = 0$$

- Rule-of-thumb: F-statistics should exceed 10, otherwise a weak instrument.

How do we solve it?

1. Ignore it $\Rightarrow$ biased and inconsistent parameter estimates
2. Use proxy (only works with omitted variables).
   - If imperfect $\Rightarrow$ still biased and inconsistent estimates
   - But may reduce bias and lower the variance
3. IV: Often weak instruments or not exogenous $\Rightarrow$ biased and imprecise estimates
4. What to do? Try it all.

TABLE VI

OLS AND TSLS ESTIMATES OF THE RETURN TO EDUCATION FOR MEN BORN 1940–1949: 1980 CENSUS[a]

Independent variable	(1) OLS	(2) TSLS	(3) OLS	(4) TSLS	(5) OLS	(6) TSLS	(7) OLS
Years of education	0.0573	0.0553	0.0573	0.0948	0.0520	0.0393	0.052
	(0.0003)	(0.0138)	(0.0003)	(0.0223)	(0.0003)	(0.0145)	(0.000
Race (1 = black)	—	—	—	—	−0.2107	−0.2266	−0.210
					(0.0032)	(0.0183)	(0.003
SMSA (1 = center city)	—	—	—	—	0.1418	0.1535	0.141
					(0.0023)	(0.0135)	(0.002
Married (1 = married)	—	—	—	—	0.2445	0.2442	0.244
					(0.0022)	(0.0022)	(0.002
9 Year-of-birth dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes
8 Region-of-residence dummies	No	No	No	No	Yes	Yes	Yes
Age	—	—	0.1800	0.1325	—	—	0.151
			(0.0389)	(0.0486)			(0.037
Age-squared	—	—	0.0023	0.0016	—	—	0.001
			(0.0006)	(0.0007)			(0.000
χ^2 [dof]	—	101.6 [29]	—	49.1 [27]	—	93.6 [29]	—

a. Standard errors are in parentheses. Sample size is 486,926. Instruments are a full set of quarter-of-birth times year-of-birth interactions. Sample consists of ma
States. The sample is drawn from the 5 percent samples of the 1980 Census. The dependent variable is the log of weekly earnings. Age and age-squared are measure
Each equation also includes an intercept.

2SLS with simultaneous equations

To make sure that Assumption 2SLS.2 is satisfied:

- At least one of the exogenous variables of the second equation is not in the first equation.
- At least one of the exogenous variables of the first equation should have a nonzero coefficient.

Counter example: House expenses and savings

Let us assume that house $expenses$ and $savings$ of a random family are determined simultaneously by:

$$expenses = \alpha_1\ savings + \beta_{10} + \beta_{11}\ income + \beta_{12}\ edu + \beta_{13}\ age + \epsilon_1$$
$$savings = \alpha_2\ expenses + \beta_{20} + \beta_{21}\ income + \beta_{12}\ edu + \beta_{13}\ age + \epsilon_2$$