Econometrics - Review of Basics I

Master 1 Semestre 2 - EPOLPRO (IEDES)

Jean-Baptiste Guiffard (Telecom-Paris, CREST)

03 février 2025

Programme des séances

Date	Programme
03/02	Retour sur les bases et approfondissements (I)
04/02	Retour sur les bases et approfondissements (II)
10/02	Hétéroscédasticité (I)
17/02	Hétéroscédasticité (II)
18/02	Examen CC + Modèles de probabilité linéaire
03/03	Logit/Probit (I)
10/03	Logit/Probit (II) + Conditionnal logit
Semaine du 17/03	Examen final

Introduction

Why econometrics?

$\Rightarrow$ To answer policy relevant questions

- Effets of raising the minimum wage on unemployment ?

Why econometrics?

$\Rightarrow$ To answer policy relevant questions

- Effets of raising the minimum wage on unemployment ?
- Effets of introducing a univeral income on productivity ?

Why econometrics?

$\Rightarrow$ To answer policy relevant questions

- Effets of raising the minimum wage on unemployment ?
- Effets of introducing a univeral income on productivity ?
- Effects of income taxation on labor supply ?

Why econometrics?

$\Rightarrow$ To answer policy relevant questions

- Effets of raising the minimum wage on unemployment ?
- Effets of introducing a univeral income on productivity ?
- Effects of income taxation on labor supply ?
- ...

Why econometrics?

Econometrics is built on the development of statistical methods aimed at:

- Estimating economic relationships

$\rightarrow$ The primary focus of econometrics is to address challenges related to the collection and analysis of non-experimental/observational economic data.

Why econometrics?

Econometrics is built on the development of statistical methods aimed at:

- Estimating economic relationships
- Testing economic theories

$\rightarrow$ The primary focus of econometrics is to address challenges related to the collection and analysis of non-experimental/observational economic data.

Why econometrics?

Econometrics is built on the development of statistical methods aimed at:

- Estimating economic relationships
- Testing economic theories
- Evaluating and implementing government and business policies

$\rightarrow$ The primary focus of econometrics is to address challenges related to the collection and analysis of non-experimental/observational economic data.

Why econometrics?

The Economist's Objective $\rightarrow$ To determine whether one variable has a causal effect on another.

Identifying a dependency link between two variables does not imply a causal relationship (even if it seems likely).

Why econometrics?

The Economist's Objective $\rightarrow$ To determine whether one variable has a causal effect on another.

Identifying a dependency link between two variables does not imply a causal relationship (even if it seems likely).

$\rightarrow$ Correlation $\neq$ Causality

Why econometrics?

The Economist's Objective $\rightarrow$ To determine whether one variable has a causal effect on another.

Identifying a dependency link between two variables does not imply a causal relationship (even if it seems likely).

$\rightarrow$ Correlation $\neq$ Causality

Ceteris Paribus (All Else Equal) Analysis $\rightarrow$ Plays a crucial role in causal analysis... However... in reality, for example, the relationship between minimum wage and unemployment rates illustrates the complexity of establishing causality.

Estimation methods

- OLS most commonly used statistical method in applied economics; allows to address a wide range of questions in development
- Other methods: ML (non-linear models)

The Power of Regression Models

Regression models are versatile statistical tools capable of addressing a wide array of questions. Let's explore three key applications:

Prediction	Utilizing parents' heights to forecast the height of their children.	Through regression, we can estimate future outcomes based on known predictors.
Modeling	Establishing a simple and clear mean relationship between the heights of parents and their children.	Regression helps in identifying and quantifying the strength and form of relationships between variables.
Covariation	Examining the variation in children's heights that seems independent of parents' heights (residual variation)	Explore underlying patterns and associations, revealing influences beyond the primary variables of interest

Outline of this course

1. The classical regression model
2. Properties of the OLS estimator in large samples
3. Heteroscedasticity: the problem and correction methods
4. Non-linear models: probit, logit, tobit, poisson

The SRM

The population model

- We assume the following *linear* model to be true at the population level: $y = \beta_0 + \beta_1 x + u$

The population model

- We assume the following *linear* model to be true at the population level: $y = \beta_0 + \beta_1 x + u$
- We are interested in the effect of a 1 unit change in x on y which is measured by β_1 but β_1 is unknown

The population model

- We assume the following *linear* model to be true at the population level: $y = \beta_0 + \beta_1 x + u$
- We are interested in the effect of a 1 unit change in x on y which is measured by β_1 but β_1 is unknown
- Let's infer the value of β_1 from a random sample of the population

The population model

- We assume the following *linear* model to be true at the population level: $y = \beta_0 + \beta_1 x + u$
- We are interested in the effect of a 1 unit change in x on y which is measured by β_1 but β_1 is unknown
- Let's infer the value of β_1 from a random sample of the population
  - If inference is good, then we should expect that if we draw another random sample from the population, we will obtain 'similar' results as to the nature of the population

The population model

- We assume the following *linear* model to be true at the population level: $y = \beta_0 + \beta_1 x + u$
- We are interested in the effect of a 1 unit change in x on y which is measured by β_1 but β_1 is unknown
- Let's infer the value of β_1 from a random sample of the population
  - If inference is good, then we should expect that if we draw another random sample from the population, we will obtain 'similar' results as to the nature of the population
  - How ?

The population model

- We assume the following *linear* model to be true at the population level: $y = \beta_0 + \beta_1 x + u$
- We are interested in the effect of a 1 unit change in x on y which is measured by β_1 but β_1 is unknown
- Let's infer the value of β_1 from a random sample of the population
  - If inference is good, then we should expect that if we draw another random sample from the population, we will obtain 'similar' results as to the nature of the population
  - How ?
  - Under the zero conditional mean assumption, the inference method will only exploit information on how x and y vary and co-vary

The zero conditional mean assumption

The zero conditional mean assumption: $E(u/x) = 0$. Relies on

- Assumption 1: $E(u/x) = E(u)$ %(for any value of x, the expected value of the unobservable u is the same and therefore must equal the expected value of u in the population)

Then, $\beta_1 = \frac{Cov(y,x)}{V(x)}$. Proof :

The zero conditional mean assumption

The zero conditional mean assumption: $E(u/x) = 0$. Relies on

- Assumption 1: $E(u/x) = E(u)$ %(for any value of x, the expected value of the unobservable u is the same and therefore must equal the expected value of u in the population)
- Assumption 2: $E(u) = 0$ %(because the population model includes a constant)

Then, $\beta_1 = \frac{Cov(y,x)}{V(x)}$. Proof :

Let's re-express β_1 using $E(u/x) = 0$

- $Cov(y, x) = Cov(\beta_0 + \beta_1 x + u, x)$

Let's re-express β_1 using $E(u/x) = 0$

- $Cov(y, x) = Cov(\beta_0 + \beta_1 x + u, x)$
- $Cov(y, x) = 0 + \beta_1 V(x) + Cov(u, x)$ since $Cov(x, x) = V(x)$

Let's re-express β_1 using $E(u/x) = 0$

- $Cov(y, x) = Cov(\beta_0 + \beta_1 x + u, x)$
- $Cov(y, x) = 0 + \beta_1 V(x) + Cov(u, x)$ since $Cov(x, x) = V(x)$
- $\beta_1 = \frac{Cov(y, x)}{V(x)}$ (indeed, since $E(u/x) = 0$ then $Cov(u, x) = 0$)

Let's re-express β_1 using $E(u/x) = 0$

- $Cov(y, x) = Cov(\beta_0 + \beta_1 x + u, x)$
- $Cov(y, x) = 0 + \beta_1 V(x) + Cov(u, x)$ since $Cov(x, x) = V(x)$
- $\beta_1 = \dfrac{Cov(y, x)}{V(x)}$ (indeed, since $E(u/x) = 0$ then $Cov(u, x) = 0$)
- Note that $Cov(y, x)$ and $V(x)$ are unknown but they can be estimated

Let's re-express β_1 using $E(u/x) = 0$

- $E(y|x) = \beta_0 + \beta_1 x$ (*)

Let's re-express β_1 using $E(u/x) = 0$

- $E(y|x) = \beta_0 + \beta_1 x$ (*)
- (*) is called the population regression function (PRF)

Let's re-express β_1 using $E(u/x) = 0$

- $E(y|x) = \beta_0 + \beta_1 x$ (*)
- (*) is called the population regression function (PRF)
  - Linearity

Let's re-express β_1 using $E(u/x) = 0$

- $E(y|x) = \beta_0 + \beta_1 x$ (*)
- (*) is called the population regression function (PRF)
  - Linearity
  - 1 unit increase in x changes the expected value of y by the amount of β_1

Let's re-express β_1 using $E(u/x) = 0$

- $E(y|x) = \beta_0 + \beta_1 x$ (*)
- (*) is called the population regression function (PRF)
  - Linearity
  - 1 unit increase in x changes the expected value of y by the amount of β_1
  - β_0 : expected value of y given x$=$0

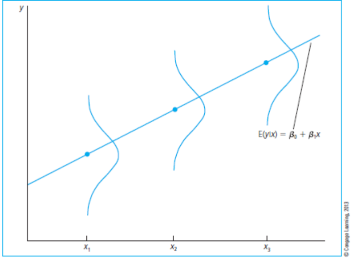
Let's re-express β_1 using $E(u/x) = 0$

- $E(y|x) = \beta_0 + \beta_1 x$ (*)
- (*) is called the population regression function (PRF)
  - Linearity
  - 1 unit increase in x changes the expected value of y by the amount of β_1
  - β_0 : expected value of y given x$=$0
- (*) is also called the conditional expectation function (CEF)

Let's re-express β_1 using $E(u/x) = 0$

- $E(y|x) = \beta_0 + \beta_1 x$ (*)
- (*) is called the population regression function (PRF)
  - Linearity
  - 1 unit increase in x changes the expected value of y by the amount of β_1
  - β_0 : expected value of y given x$=$0
- (*) is also called the conditional expectation function (CEF)
- These are population-level concepts

FIGURE 2.1 E(y|x) as a linear function of x.



Wage and Education: the case of South-Africa (1993)

- Let's assume $Log(wage) = \beta_0 + \beta_1 educ + u$ to be true at the population level.

Wage and Education: the case of South-Africa (1993)

- Let's assume $Log(wage) = \beta_0 + \beta_1 educ + u$ to be true at the population level.
- Interpret β_1 and β_0 making clear the assumption(s) made

Wage and Education: the case of South-Africa (1993)

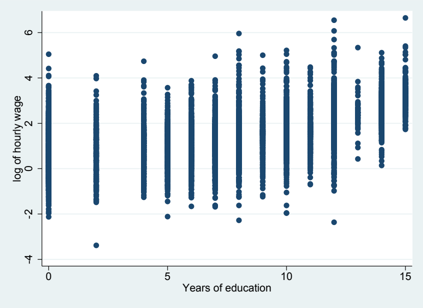
- Under the zero conditional mean assumption, we have $E(ability|educ = 5) = E(ability|educ = 15)$ (xx)

Wage and Education: the case of South-Africa (1993)

- Under the zero conditional mean assumption, we have $E(ability|educ = 5) = E(ability|educ = 15)$ (xx)
- Plausible ?

Wage and Education: the case of South-Africa (1993)

- If (xx) is not verified, what does the following observed relationship between *Log*(*wage*) and *educ* suggest ?



The OLS estimator

- OLS estimates of β_1 and β_0 are obtained by

The OLS estimator

- OLS estimates of β_1 and β_0 are obtained by
- Min $\Sigma\widehat{u}^2 = \Sigma[y - (\widehat{\beta}_0 + \widehat{\beta}_1 x)]^2$

The OLS estimator

- OLS estimates of β_1 and β_0 are obtained by
- Min $\Sigma\widehat{u}^2 = \Sigma[y - (\widehat{\beta_0} + \widehat{\beta_1}x)]^2$
  - CPO 1: $\Sigma[y - (\widehat{\beta_0} + \widehat{\beta_1}x)] = 0$

The OLS estimator

- OLS estimates of β_1 and β_0 are obtained by
- Min $\Sigma\widehat{u}^2 = \Sigma[y - (\widehat{\beta_0} + \widehat{\beta_1}x)]^2$
  - CPO 1: $\Sigma[y - (\widehat{\beta_0} + \widehat{\beta_1}x)] = 0$
  - CPO 2: $\Sigma x[y - (\widehat{\beta_0} + \widehat{\beta_1}x)] = 0 \rightarrow$ (demo p63 in Wooldridge (2013)

The OLS estimator

- OLS estimates of β_1 and β_0 are obtained by
- Min $\Sigma\widehat{u}^2 = \Sigma[y - (\widehat{\beta_0} + \widehat{\beta_1}x)]^2$
  - CPO 1: $\Sigma[y - (\widehat{\beta_0} + \widehat{\beta_1}x)] = 0$
  - CPO 2: $\Sigma x[y - (\widehat{\beta_0} + \widehat{\beta_1}x)] = 0 \rightarrow$ (demo p63 in Wooldridge (2013)
- The sample regression function : $\widehat{y} = \widehat{\beta_0} + \widehat{\beta_1}x$

The OLS estimator

- OLS estimates of β_1 and β_0 are obtained by
- Min $\Sigma\widehat{u}^2 = \Sigma[y - (\widehat{\beta_0} + \widehat{\beta_1}x)]^2$
  - CPO 1: $\Sigma[y - (\widehat{\beta_0} + \widehat{\beta_1}x)] = 0$
  - CPO 2: $\Sigma x[y - (\widehat{\beta_0} + \widehat{\beta_1}x)] = 0 \rightarrow$ (demo p63 in Wooldridge (2013)
- The sample regression function : $\widehat{y} = \widehat{\beta_0} + \widehat{\beta_1}x$
  - $\widehat{\beta_1} = \dfrac{\Sigma(x - \overline{x})(y - \overline{y})}{\Sigma(x - \overline{x})^2}$ (**)

The OLS estimator

- OLS estimates of β_1 and β_0 are obtained by
- Min $\Sigma\widehat{u}^2 = \Sigma[y - (\widehat{\beta_0} + \widehat{\beta_1}x)]^2$
  - CPO 1: $\Sigma[y - (\widehat{\beta_0} + \widehat{\beta_1}x)] = 0$
  - CPO 2: $\Sigma x[y - (\widehat{\beta_0} + \widehat{\beta_1}x)] = 0 \rightarrow$ (demo p63 in Wooldridge (2013)
- The sample regression function : $\widehat{y} = \widehat{\beta_0} + \widehat{\beta_1}x$
  - $\widehat{\beta_1} = \dfrac{\Sigma(x-\overline{x})(y-\overline{y})}{\Sigma(x-\overline{x})^2}$ (**)
  - (**) is the sample analogue of the population parameter β_1

The OLS estimator

- OLS estimates of β_1 and β_0 are obtained by
- Min $\Sigma\widehat{u}^2 = \Sigma[y - (\widehat{\beta_0} + \widehat{\beta_1}x)]^2$
  - CPO 1: $\Sigma[y - (\widehat{\beta_0} + \widehat{\beta_1}x)] = 0$
  - CPO 2: $\Sigma x[y - (\widehat{\beta_0} + \widehat{\beta_1}x)] = 0 \rightarrow$ (demo p63 in Wooldridge (2013)
- The sample regression function : $\widehat{y} = \widehat{\beta_0} + \widehat{\beta_1}x$
  - $\widehat{\beta_1} = \dfrac{\Sigma(x-\overline{x})(y-\overline{y})}{\Sigma(x-\overline{x})^2}$ (**)
  - (**) is the sample analogue of the population parameter β_1
  - Also: $\widehat{\beta_1} = \beta_1 + \dfrac{\Sigma(x-\overline{x})u}{\Sigma(x-\overline{x})^2}$

The OLS estimator

- OLS estimates of β_1 and β_0 are obtained by
- Min $\Sigma\widehat{u}^2 = \Sigma[y - (\widehat{\beta_0} + \widehat{\beta_1}x)]^2$
  - CPO 1: $\Sigma[y - (\widehat{\beta_0} + \widehat{\beta_1}x)] = 0$
  - CPO 2: $\Sigma x[y - (\widehat{\beta_0} + \widehat{\beta_1}x)] = 0 \rightarrow$ (demo p63 in Wooldridge (2013)
- The sample regression function : $\widehat{y} = \widehat{\beta_0} + \widehat{\beta_1}x$
  - $\widehat{\beta_1} = \dfrac{\Sigma(x-\overline{x})(y-\overline{y})}{\Sigma(x-\overline{x})^2}$ (**)
  - (**) is the sample analogue of the population parameter β_1
  - Also: $\widehat{\beta_1} = \beta_1 + \dfrac{\Sigma(x-\overline{x})u}{\Sigma(x-\overline{x})^2}$
  - $\widehat{\beta_0} = \overline{y} - \widehat{\beta_1} * \overline{x}$

Application : Wage and Education in South-Africa (1993)

```
. reg logwphy edyrs
```

Source	SS	df	MS
Model	2368.42412	1	2368.42412
Residual	6155.34858	6,966	.883627416
Total	8523.77271	6,967	1.22344951

Number of obs	=	6,968
F(1, 6966)	=	2680.34
Prob > F	=	0.0000
R-squared	=	0.2779
Adj R-squared	=	0.2778
Root MSE	=	.94001

logwphy	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
edyrs	.1353827	.002615	51.77	0.000	.1302565	.1405088
_cons	.4581331	.0238719	19.19	0.000	.4113368	.5049294

- $\widehat{y} = 0.458 + 0.135x$

Application : Wage and Education in South-Africa (1993)

```
. reg logwphy edyrs

      Source |       SS           df       MS      Number of obs   =     6,968
-------------+----------------------------------   F(1, 6966)      =   2680.34
       Model |  2368.42412         1  2368.42412   Prob > F        =    0.0000
    Residual |  6155.34858     6,966  .883627416   R-squared       =    0.2779
-------------+----------------------------------   Adj R-squared   =    0.2778
       Total |  8523.77271     6,967  1.22344951   Root MSE        =    .94001

------------------------------------------------------------------------------
     logwphy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       edyrs |   .1353827    .002615    51.77   0.000     .1302565    .1405088
       _cons |   .4581331   .0238719    19.19   0.000     .4113368    .5049294
------------------------------------------------------------------------------
```

- $\widehat{y} = 0.458 + 0.135x$
- $\nearrow$ 1 year of education $\rightarrow$ wage $\nearrow$ $[(exp(\beta_1) - 1] * 100$ % (=14.4%)

Application : Wage and Education in South-Africa (1993)

```
. reg logwphy edyrs

      Source |       SS           df       MS      Number of obs   =     6,968
-------------+----------------------------------   F(1, 6966)      =   2680.34
       Model |  2368.42412         1  2368.42412   Prob > F        =    0.0000
    Residual |  6155.34858     6,966  .883627416   R-squared       =    0.2779
-------------+----------------------------------   Adj R-squared   =    0.2778
       Total |  8523.77271     6,967  1.22344951   Root MSE        =    .94001

------------------------------------------------------------------------------
     logwphy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       edyrs |   .1353827    .002615    51.77   0.000     .1302565    .1405088
       _cons |   .4581331   .0238719    19.19   0.000     .4113368    .5049294
------------------------------------------------------------------------------
```

- $\widehat{y} = 0.458 + 0.135x$
- $\nearrow$ 1 year of education $\rightarrow$ wage $\nearrow$ $[(exp(\beta_1) - 1] * 100$ % (=14.4%)
- More on coefficient interpretation

Properties of statistics derived from OLS estimation

- CPO (1) => $\Sigma\widehat{u} = 0$

Properties of statistics derived from OLS estimation

- CPO (1) $=> \Sigma\widehat{u} = 0$
- CP0 (2) $=> \Sigma x\widehat{u} = 0$

Properties of statistics derived from OLS estimation

- CPO (1) $=>$ $\Sigma \widehat{u} = 0$
- CP0 (2) $=>$ $\Sigma x \widehat{u} = 0$
- $(\overline{x}, \overline{y})$ belongs to the OLS equation

Properties of statistics derived from OLS estimation

- CPO (1) $=> \Sigma\widehat{u} = 0$
- CP0 (2) $=> \Sigma x\widehat{u} = 0$
- $(\overline{x}, \overline{y})$ belongs to the OLS equation

Properties of statistics derived from OLS estimation

- CPO (1) $=>\Sigma\widehat{u}=0$
- CP0 (2) $=>\Sigma x\widehat{u}=0$
- $(\overline{x},\overline{y})$ belongs to the OLS equation
- (1 & (2) $\rightarrow cov(\widehat{y},\widehat{u})=0$

Properties of statistics derived from OLS estimation

- CPO (1) $=>$ $\Sigma \widehat{u} = 0$
- CP0 (2) $=>$ $\Sigma x \widehat{u} = 0$
- $(\overline{x}, \overline{y})$ belongs to the OLS equation
- (1 & (2) $\rightarrow cov(\widehat{y}, \widehat{u}) = 0$
- Thus, SCT $=$ SCR $+$ SCE {(cf demo p74 Wooldridge, 2013)}

Why OLS estimator and why not another one (ex: min the absolute distance between y and $\widehat{y}$) ?

OLS estimator property

Validity and Precision

Validity

A measure is considered valid if it accurately captures the concept it intends to measure, meaning it exhibits low systematic error. Validity is assessed by comparing various measures of the same concept to ensure they align closely with the theoretical construct they are supposed to represent.

Precision

Precision refers to the consistency of a measure in replicating the same value across repeated observations of a phenomenon, indicating low random error. To assess precision, one can repeatedly measure a phenomenon and compare the outcomes to check for consistency (test-retest method).

Assumption for unbiasedness

Unbiasedness

Definition: an estimator is unbaised if its average value over a large number of repeated trials equals the population value ($E(\widehat{\beta_0}) = \beta_0$ and $E(\widehat{\beta_1}) = \beta_1$)

- $\widehat{\beta_k}$ is unbaised if

Assumption for unbiasedness

Unbiasedness

Definition: an estimator is unbaised if its average value over a large number of repeated trials equals the population value ($E(\widehat{\beta_0}) = \beta_0$ and $E(\widehat{\beta_1}) = \beta_1$)

- $\widehat{\beta_k}$ is unbaised if
  - (A1) The model is linear in its parameters

Assumption for unbiasedness

Unbiasedness

Definition: an estimator is unbaised if its average value over a large number of repeated trials equals the population value ($E(\widehat{\beta_0}) = \beta_0$ and $E(\widehat{\beta_1}) = \beta_1$)

- $\widehat{\beta_k}$ is unbaised if
  - (A1) The model is linear in its parameters
  - (A2) We have a random sample from the population of interest

Assumption for unbiasedness

Unbiasedness

Definition: an estimator is unbaised if its average value over a large number of repeated trials equals the population value ($E(\widehat{\beta_0}) = \beta_0$ and $E(\widehat{\beta_1}) = \beta_1$)

- $\widehat{\beta_k}$ is unbaised if
  - (A1) The model is linear in its parameters
  - (A2) We have a random sample from the population of interest
  - (A3) There is sample variation in the explanatory variables

Assumption for unbiasedness

Unbiasedness

Definition: an estimator is unbaised if its average value over a large number of repeated trials equals the population value ($E(\widehat{\beta_0}) = \beta_0$ and $E(\widehat{\beta_1}) = \beta_1$)

- $\widehat{\beta_k}$ is unbaised if
    - (A1) The model is linear in its parameters
    - (A2) We have a random sample from the population of interest
    - (A3) There is sample variation in the explanatory variables
    - (A4) The zero conditional mean assumption is verified

Assumption for unbiasedness

Unbiasedness

Definition: an estimator is unbaised if its average value over a large number of repeated trials equals the population value ($E(\widehat{\beta_0}) = \beta_0$ and $E(\widehat{\beta_1}) = \beta_1$)

- $\widehat{\beta_k}$ is unbaised if
  - (A1) The model is linear in its parameters
  - (A2) We have a random sample from the population of interest
  - (A3) There is sample variation in the explanatory variables
  - (A4) The zero conditional mean assumption is verified
- {(demo p. 88 Wooldridge (2013))

- Unbiased on average, but what about its dispersion around the true value ?

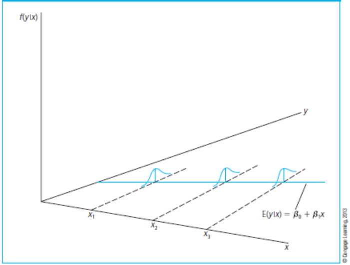
- Unbiased on average, but what about its dispersion around the true value ?
- Ceteris paribus, we clearly prefer an estimator with minimum variance

Assumption for minimum variance

Minimum variance

- $\widehat{\beta_k}$ is with minimum variance if

FIGURE 2.8 The simple regression model under homoskedasticity.

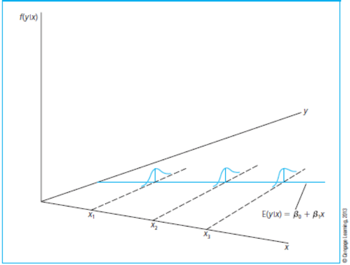


Assumption for minimum variance

Minimum variance

- $\widehat{\beta_k}$ is with minimum variance if
  - (A5) $Var(u/x) = \sigma^2$ is verified (homoscedasticity)

FIGURE 2.8 The simple regression model under homoskedasticity.

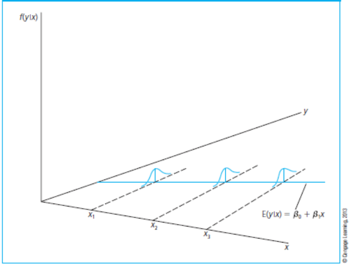


Assumption for minimum variance

Minimum variance

- $\widehat{\beta_k}$ is with minimum variance if
  - (A5) $Var(u/x) = \sigma^2$ is verified (homoscedasticity)

FIGURE 2.8 The simple regression model under homoskedasticity.



- If (A1) to (A5) are verified, then $\widehat{\beta_k}$ is BLUE ({not demonstrated})

Variance expression

- $V(\widehat{\beta}_1) = \frac{\sigma^2}{SCT_x}$ {(cf demo)} $\rightarrow$ This formula highlights two critical components influencing the estimator's precision:

Variance expression

- $V(\widehat{\beta}_1) = \frac{\sigma^2}{SCT_x}$ {(cf demo)} $\rightarrow$ This formula highlights two critical components influencing the estimator's precision:
  - Error Variance : Reflects the variability in the observed values around the regression line. A higher error variance means more uncertainty in our estimation (σ^2).

Variance expression

- $V(\widehat{\beta_1}) = \frac{\sigma^2}{SCT_x}$ {(cf demo)} → This formula highlights two critical components influencing the estimator's precision:
  - Error Variance : Reflects the variability in the observed values around the regression line. A higher error variance means more uncertainty in our estimation (σ^2).
  - Total variation in x: Denotes the sum of squared deviations of x values from their mean. Greater variation in x provides a stronger base for estimating the slope, reducing the variance of $\widehat{\beta_1}$.

Variance expression

- $V(\widehat{\beta_1}) = \dfrac{\sigma^2}{SCT_x}$ {(cf demo)} $\rightarrow$ This formula highlights two critical components influencing the estimator's precision:
  - Error Variance : Reflects the variability in the observed values around the regression line. A higher error variance means more uncertainty in our estimation (σ^2).
  - Total variation in x: Denotes the sum of squared deviations of x values from their mean. Greater variation in x provides a stronger base for estimating the slope, reducing the variance of $\widehat{\beta_1}$.
- The problem is that we do not know σ^2...

What is the difference between errors and residuals?

Population Model: $y_i = \beta_0 + \beta_1 x_i + u_i \rightarrow u_i$ represents the error for observation i.

Expressing y_i in terms of its fitted value and residual: $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i$ $\rightarrow$ Residuals are part of the estimated equation.

Key Differences:

- Errors (u_i) can never be directly observed since they represent the deviation of observed values from the true population parameters.
- Residuals ($\hat{u}_i$) are calculated from the data and represent the difference between observed values and those predicted by the model.

Residuals in Relation to Errors:

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

$$\hat{u}_i = (\beta_0 + \beta_1 x_i + u_i) - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

$$\hat{u}_i = u_i - (\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1) x_i$$

Variance expression

- $\sigma^2 = E(u^2)$. An unbiased estimate is $\frac{\Sigma \widehat{u}^2}{n-2} = \frac{SCR}{n-2}$ %({p97 in Wooldridge (2013)})

Variance expression

- $\sigma^2 = E(u^2)$. An unbiased estimate is $\frac{\Sigma \widehat{u}^2}{n-2} = \frac{SCR}{n-2}$ %({p97 in Wooldridge (2013)})
  - $\frac{SCR}{n}$ is biased because it omits the two conditions residuals should verify in an OLS model (if we know the value of $n-2$ residuals, the two last are constrained by the conditions)

Variance expression

- Where do we read $V(\widehat{\beta}_1)$ (and $V(\widehat{\beta}_0)$) in the regression results ?

```
. reg logwphy edyrs

      Source |       SS           df       MS      Number of obs   =     6,968
-------------+----------------------------------   F(1, 6966)      =   2680.34
       Model |  2368.42412         1  2368.42412   Prob > F        =    0.0000
    Residual |  6155.34858     6,966  .883627416   R-squared       =    0.2779
-------------+----------------------------------   Adj R-squared   =    0.2778
       Total |  8523.77271     6,967  1.22344951   Root MSE        =    .94001

------------------------------------------------------------------------------
     logwphy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       edyrs |   .1353827    .002615    51.77   0.000     .1302565    .1405088
       _cons |   .4581331   .0238719    19.19   0.000     .4113368    .5049294
------------------------------------------------------------------------------
```

Assumption violation

- What if (A5) is violated ? Then OLS estimator has not minimum variance

Assumption violation

- What if (A5) is violated ? Then OLS estimator has not minimum variance
  - More on heteroscedasticity later (what can we say based on the graph?)

Assumption violation

- What if (A5) is violated ? Then OLS estimator has not minimum variance
  - More on heteroscedasticity later (what can we say based on the graph?)
- What if (A2) is violated ? Then OLS estimator is *biased* and its causal interpretation is challenged

Assumption violation

- What if (A5) is violated ? Then OLS estimator has not minimum variance
  - More on heteroscedasticity later (what can we say based on the graph?)
- What if (A2) is violated ? Then OLS estimator is *biased* and its causal interpretation is challenged
  - Random sampling ($\rightarrow$ errors are uncorrelated)

Assumption violation

- What if (A5) is violated ? Then OLS estimator has not minimum variance
  - More on heteroscedasticity later (what can we say based on the graph?)
- What if (A2) is violated ? Then OLS estimator is *biased* and its causal interpretation is challenged
  - Random sampling ($\rightarrow$ errors are uncorrelated)
  - Stratified random sampling ($\rightarrow$ less sure that errors are uncorrelated ...)

Assumption violation

- What if (A5) is violated ? Then OLS estimator has not minimum variance
  - More on heteroscedasticity later (what can we say based on the graph?)
- What if (A2) is violated ? Then OLS estimator is *biased* and its causal interpretation is challenged
  - Random sampling ($\rightarrow$ errors are uncorrelated)
  - Stratified random sampling ($\rightarrow$ less sure that errors are uncorrelated . . .)
- What if (A4) is violated ? Then OLS estimator is *biased* and its causal interpretation is challenged

Assumption violation

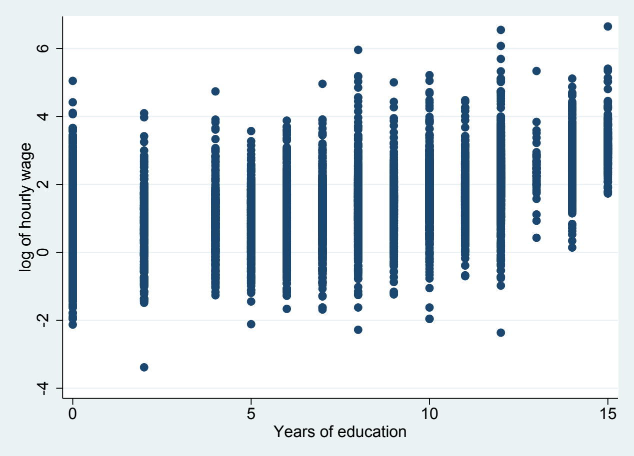
- What if (A5) is violated ? Then OLS estimator has not minimum variance
  - More on heteroscedasticity later (what can we say based on the graph?)
- What if (A2) is violated ? Then OLS estimator is *biased* and its causal interpretation is challenged
  - Random sampling ($\rightarrow$ errors are uncorrelated)
  - Stratified random sampling ($\rightarrow$ less sure that errors are uncorrelated ...)
- What if (A4) is violated ? Then OLS estimator is *biased* and its causal interpretation is challenged
  - On the benefit of adding controls correlated with the x of interest

Assumption violation

- What if (A5) is violated ? Then OLS estimator has not minimum variance
  - More on heteroscedasticity later (what can we say based on the graph?)
- What if (A2) is violated ? Then OLS estimator is *biased* and its causal interpretation is challenged
  - Random sampling ($\rightarrow$ errors are uncorrelated)
  - Stratified random sampling ($\rightarrow$ less sure that errors are uncorrelated . . .)
- What if (A4) is violated ? Then OLS estimator is *biased* and its causal interpretation is challenged
  - On the benefit of adding controls correlated with the x of interest
  - On the benefit of conducting experiments (whenever possible)% [more on this next year!]

Wage and Education: the case of South-Africa (1993)

- Intuition regarding the risk of heteroscedasticity



Extending the SRM

A dummy as explanatory variable

```
. reg logwphy prim_com

      Source |       SS           df       MS      Number of obs   =     6,968
-------------+----------------------------------   F(1, 6966)      =   1121.68
       Model |  1182.15867         1  1182.15867   Prob > F        =    0.0000
    Residual |  7341.61404     6,966  1.05392105   R-squared       =    0.1387
-------------+----------------------------------   Adj R-squared   =    0.1386
       Total |  8523.77271     6,967  1.22344951   Root MSE        =    1.0266

------------------------------------------------------------------------------
     logwphy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  prim_compl |   .9492352   .0283426    33.49   0.000      .893675    1.004795
       _cons |   .8374504   .0245196    34.15   0.000     .7893846    .8855163
------------------------------------------------------------------------------
```

- $E(Log(wage)/prim = 1) = \beta_0 + \beta_1 = ?$

A dummy as explanatory variable

```
. reg logwphy prim_com

      Source |       SS           df       MS      Number of obs   =     6,968
-------------+----------------------------------   F(1, 6966)      =   1121.68
       Model |  1182.15867         1  1182.15867   Prob > F        =    0.0000
    Residual |  7341.61404     6,966  1.05392105   R-squared       =    0.1387
-------------+----------------------------------   Adj R-squared   =    0.1386
       Total |  8523.77271     6,967  1.22344951   Root MSE        =    1.0266

------------------------------------------------------------------------------
     logwphy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  prim_compl |   .9492352   .0283426    33.49   0.000      .893675    1.004795
       _cons |   .8374504   .0245196    34.15   0.000     .7893846    .8855163
------------------------------------------------------------------------------
```

- $E(Log(wage)/prim = 1) = \beta_0 + \beta_1 = ?$
- $E(Log(wage)/prim = 0) = \beta_0 = ?$

A dummy as explanatory variable

```
. reg logwphy prim_com

      Source |       SS           df       MS      Number of obs   =     6,968
-------------+----------------------------------   F(1, 6966)      =   1121.68
       Model |  1182.15867         1  1182.15867   Prob > F        =    0.0000
    Residual |  7341.61404     6,966  1.05392105   R-squared       =    0.1387
-------------+----------------------------------   Adj R-squared   =    0.1386
       Total |  8523.77271     6,967  1.22344951   Root MSE        =    1.0266

------------------------------------------------------------------------------
     logwphy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  prim_compl |   .9492352   .0283426    33.49   0.000      .893675    1.004795
       _cons |   .8374504   .0245196    34.15   0.000     .7893846    .8855163
------------------------------------------------------------------------------
```

- $E(Log(wage)/prim = 1) = \beta_0 + \beta_1 = ?$
- $E(Log(wage)/prim = 0) = \beta_0 = ?$
- Completing primary education $\rightarrow$ wage increases by $[(exp(\beta_1) - 1] * 100\ \%$ (=158 %)

Wage and primary education in South Africa (1993) (2)

```
. bysort prim_com: sum logwphy

-> prim_compl = 0

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     logwphy |      1,753    .8374504    1.059755  -3.383063   5.047622

-> prim_compl = 1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     logwphy |      5,215    1.786686    1.015225  -2.364309    6.64859
```

Multiple Linear Regression (MLR)

Multiple Linear Regression (MLR) is an extension of simple linear regression that allows for the prediction of a dependent variable based on the values of two or more independent variables. By incorporating multiple predictors, MLR facilitates a more nuanced analysis, enabling researchers and analysts to understand the complex relationships between variables.

- *Ceteris Paribus Reasoning*: MLR is well-suited for 'ceteris paribus' analysis, allowing for the explicit consideration of many factors that simultaneously affect the dependent variable.
- *Incorporating Multiple Predictors*: MLR allows for the inclusion of numerous explanatory variables, providing a framework to add useful factors for explaining variations in the dependent variable.
- *Enhanced Predictive Power*: By accounting for multiple influencing factors, MLR can lead to improved predictions of the dependent variable, offering deeper insights into how different variables interact.

Advantages of Multiple Linear Regression

- *Comprehensive Analysis*: MLR enables a more comprehensive examination of the data by considering multiple factors at once, which is more reflective of real-world complexities.
- *Improved Prediction Accuracy*: The inclusion of multiple relevant variables can improve the model's accuracy in predicting the outcome.
- *Diverse Functional Forms*: MLR accommodates various functional forms, allowing for the modeling of complex relationships among variables.
- *Control for Confounding Variables*: By including multiple predictors, MLR helps to control for the potential confounding effects of variables, leading to more reliable and valid conclusions.

Multiple regression model: The OLS estimator

- Let's assume the population model is :
  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + (...) + u$

Multiple regression model: The OLS estimator

- Let's assume the population model is :
  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + (...) + u$
- How compute $\widehat{\beta}_j$?

Multiple regression model: The OLS estimator

- Let's assume the population model is :
  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + (...) + u$
- How compute $\widehat{\beta_j}$?
- $\widehat{\beta_j}$ is obtained by minimizing Σu^2 which amounts to solving the following set of equations

Multiple regression model: The OLS estimator

- Let's assume the population model is :
  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + (...) + u$
- How compute $\widehat{\beta_j}$?
- $\widehat{\beta_j}$ is obtained by minimizing Σu^2 which amounts to solving the following set of equations
  - $\Sigma[y - (\widehat{\beta_0} + \widehat{\beta_1} x_1 + ... + \widehat{\beta_k} x_k)] = 0$

Multiple regression model: The OLS estimator

- Let's assume the population model is :
  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + (...) + u$
- How compute $\widehat{\beta_j}$?
- $\widehat{\beta_j}$ is obtained by minimizing Σu^2 which amounts to solving the following set of equations
  - $\Sigma[y - (\widehat{\beta_0} + \widehat{\beta_1} x_1 + ... + \widehat{\beta_k} x_k)] = 0$
  - $\Sigma x_1 [y - (\widehat{\beta_0} + \widehat{\beta_1} x_1 + ... + \widehat{\beta_k} x_k)] = 0$

Multiple regression model: The OLS estimator

- Let's assume the population model is :
  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + (...) + u$
- How compute $\widehat{\beta_j}$?
- $\widehat{\beta_j}$ is obtained by minimizing Σu^2 which amounts to solving the following set of equations
  - $\Sigma[y - (\widehat{\beta_0} + \widehat{\beta_1} x_1 + ... + \widehat{\beta_k} x_k)] = 0$
  - $\Sigma x_1[y - (\widehat{\beta_0} + \widehat{\beta_1} x_1 + ... + \widehat{\beta_k} x_k)] = 0$
  - $\Sigma x_2[y - (\widehat{\beta_0} + \widehat{\beta_1} x_1 + ... + \widehat{\beta_k} x_k)] = 0$

Multiple regression model: The OLS estimator

- Let's assume the population model is :
  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + (...) + u$
- How compute $\widehat{\beta_j}$?
- $\widehat{\beta_j}$ is obtained by minimizing Σu^2 which amounts to solving the following set of equations
  - $\Sigma[y - (\widehat{\beta_0} + \widehat{\beta_1} x_1 + ... + \widehat{\beta_k} x_k)] = 0$
  - $\Sigma x_1[y - (\widehat{\beta_0} + \widehat{\beta_1} x_1 + ... + \widehat{\beta_k} x_k)] = 0$
  - $\Sigma x_2[y - (\widehat{\beta_0} + \widehat{\beta_1} x_1 + ... + \widehat{\beta_k} x_k)] = 0$
  - $(\ldots)$

Multiple regression model: The OLS estimator

- Let's assume the population model is :
  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + (...) + u$
- How compute $\widehat{\beta_j}$?
- $\widehat{\beta_j}$ is obtained by minimizing Σu^2 which amounts to solving the following set of equations
  - $\Sigma[y - (\widehat{\beta_0} + \widehat{\beta_1} x_1 + ... + \widehat{\beta_k} x_k)] = 0$
  - $\Sigma x_1[y - (\widehat{\beta_0} + \widehat{\beta_1} x_1 + ... + \widehat{\beta_k} x_k)] = 0$
  - $\Sigma x_2[y - (\widehat{\beta_0} + \widehat{\beta_1} x_1 + ... + \widehat{\beta_k} x_k)] = 0$
  - $(\ldots)$
  - $\Sigma x_k[y - (\widehat{\beta_0} + \widehat{\beta_1} x_1 + ... + \widehat{\beta_k} x_k)] = 0$

Multiple regression model: The OLS estimator

- Let's assume the population model is :
  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + (...) + u$
- How compute $\widehat{\beta}_j$?
- $\widehat{\beta}_j$ is obtained by minimizing Σu^2 which amounts to solving the following set of equations
  - $\Sigma[y - (\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + ... + \widehat{\beta}_k x_k)] = 0$
  - $\Sigma x_1[y - (\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + ... + \widehat{\beta}_k x_k)] = 0$
  - $\Sigma x_2[y - (\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + ... + \widehat{\beta}_k x_k)] = 0$
  - $(\dots)$
  - $\Sigma x_k[y - (\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + ... + \widehat{\beta}_k x_k)] = 0$
- $\rightarrow$ better handled using the matrix form {see application }; computers solve the problem

Expression of coefficient estimate

$$\widehat{\beta} = (X'X)^{-1}X'Y$$

Property of OLS estimator

- $\widehat{\beta}_j$ estimated by OLS is unbaised if

Property of OLS estimator

- $\widehat{\beta}_j$ estimated by OLS is unbaised if
  - (A1)' The model is linear in its parameters

Property of OLS estimator

- $\widehat{\beta}_j$ estimated by OLS is unbaised if
  - (A1)' The model is linear in its parameters
  - (A2)' We have a random sample from the population of interest

Property of OLS estimator

- $\widehat{\beta}_j$ estimated by OLS is unbaised if
  - (A1)' The model is linear in its parameters
  - (A2)' We have a random sample from the population of interest
  - (A3)' There is sample variation in all explanatory variables and no explanatory variable is collinear with other explanatory variables (no variable is derived from the combination of other variables)

Property of OLS estimator

- $\widehat{\beta}_j$ estimated by OLS is unbaised if
  - (A1)' The model is linear in its parameters
  - (A2)' We have a random sample from the population of interest
  - (A3)' There is sample variation in all explanatory variables and no explanatory variable is collinear with other explanatory variables (no variable is derived from the combination of other variables)
  - (A4)' The zero conditional mean assumption is verified : $E[u/x1, x2, ..., xk] = 0$

Property of OLS estimator

- $\widehat{\beta}_j$ estimated by OLS is unbaised if
  - (A1)' The model is linear in its parameters
  - (A2)' We have a random sample from the population of interest
  - (A3)' There is sample variation in all explanatory variables and no explanatory variable is collinear with other explanatory variables (no variable is derived from the combination of other variables)
  - (A4)' The zero conditional mean assumption is verified : $E[u/x1, x2, ..., xk] = 0$
  - {(not demonstrated)}

Property of OLS estimator

- $\widehat{\beta}_j$ estimated by OLS is unbaised if
  - (A1)' The model is linear in its parameters
  - (A2)' We have a random sample from the population of interest
  - (A3)' There is sample variation in all explanatory variables and no explanatory variable is collinear with other explanatory variables (no variable is derived from the combination of other variables)
  - (A4)' The zero conditional mean assumption is verified : $E[u/x1, x2, ..., xk] = 0$
  - {(not demonstrated)}
- $\widehat{\beta}_j$ estimated by OLS is BLUE if, in addition,

Property of OLS estimator

- $\widehat{\beta}_j$ estimated by OLS is unbaised if
  - (A1)' The model is linear in its parameters
  - (A2)' We have a random sample from the population of interest
  - (A3)' There is sample variation in all explanatory variables and no explanatory variable is collinear with other explanatory variables (no variable is derived from the combination of other variables)
  - (A4)' The zero conditional mean assumption is verified : $E[u/x1, x2, ..., xk] = 0$
  - {(not demonstrated)}
- $\widehat{\beta}_j$ estimated by OLS is BLUE if, in addition,
  - (A5)' $V[u/x1, x2, ..., xk] = \sigma^2$ {(not demonstrated)}

Property of OLS estimator

- $\widehat{\beta}_j$ estimated by OLS is unbaised if
  - (A1)' The model is linear in its parameters
  - (A2)' We have a random sample from the population of interest
  - (A3)' There is sample variation in all explanatory variables and no explanatory variable is collinear with other explanatory variables (no variable is derived from the combination of other variables)
  - (A4)' The zero conditional mean assumption is verified : $E[u/x1, x2, ..., xk] = 0$
  - {(not demonstrated)}
- $\widehat{\beta}_j$ estimated by OLS is BLUE if, in addition,
  - (A5)' $V[u/x1, x2, ..., xk] = \sigma^2$ {(not demonstrated)}
- These hypothesis are called hypothesis of Gauss Markov

Expression of coefficient variance

Under Gauss-Markov hypothesis,

- $V(\widehat{\beta}_j) = \dfrac{\sigma^2}{SCT_{xj}(1 - R^2_{xj})}$ {(not demonstrated)}

Expression of coefficient variance

Under Gauss-Markov hypothesis,

- $V(\widehat{\beta}_j) = \dfrac{\sigma^2}{SCT_{xj}(1 - R^2_{xj})}$ {(not demonstrated)}
  - with σ^2 estimated by $\dfrac{SCR}{n-(1+k)}$

Expression of coefficient variance

Under Gauss-Markov hypothesis,

- $V(\widehat{\beta}_j) = \dfrac{\sigma^2}{SCT_{xj}(1 - R^2_{xj})}$ {(not demonstrated)}
  - with σ^2 estimated by $\dfrac{SCR}{n-(1+k)}$
  - with R^2_{xj} the R^2 of a model where x_j is regressed on all other x (and measure how strongly the other explanatory variables in the model correlate with x_j)

Expression of coefficient variance

Under Gauss-Markov hypothesis,

- $V(\widehat{\beta}_j) = \dfrac{\sigma^2}{SCT_{xj}(1 - R^2_{xj})}$ {(not demonstrated)}
  - with σ^2 estimated by $\dfrac{SCR}{n - (1 + k)}$
  - with R^2_{xj} the R^2 of a model where x_j is regressed on all other x (and measure how strongly the other explanatory variables in the model correlate with x_j)
- {(see application)}

Expression of coefficient variance

Under Gauss-Markov hypothesis,

- $V(\widehat{\beta}_j) = \dfrac{\sigma^2}{SCT_{xj}(1 - R^2_{xj})}$ {(not demonstrated)}
  - with σ^2 estimated by $\dfrac{SCR}{n-(1+k)}$
  - with R^2_{xj} the R^2 of a model where x_j is regressed on all other x (and measure how strongly the other explanatory variables in the model correlate with x_j)
- {(see application)}
- NB: the matrix form of $V(\widehat{\beta}) = \sigma^2(X'X)^{-1}$

Issue if collinearity

- Application: let's compare the two formula of $V(\widehat{\beta_1})$

Issue if collinearity

- Application: let's compare the two formula of $V(\widehat{\beta_1})$
  - Coefficient on x1 if SRM: $V(\widehat{\beta_1}) = \frac{\sigma^2}{SCT_x}$

Issue if collinearity

- Application: let's compare the two formula of $V(\widehat{\beta_1})$
  - Coefficient on x1 if SRM: $V(\widehat{\beta_1}) = \dfrac{\sigma^2}{SCT_x}$
  - Coefficient on x1 if MRM: $V(\widehat{\beta_1}) = \dfrac{\sigma^2}{SCT_{x1}(1 - R^2_{x1})}$

Issue if collinearity

- Application: let's compare the two formula of $V(\widehat{\beta_1})$
  - Coefficient on x1 if SRM: $V(\widehat{\beta_1}) = \dfrac{\sigma^2}{SCT_x}$
  - Coefficient on x1 if MRM: $V(\widehat{\beta_1}) = \dfrac{\sigma^2}{SCT_{x1}(1 - R^2_{x1})}$
  - $V(\widehat{\beta_1})$ is higher if collinearity between the x

Issue if omitted variables (1)

(particular case with 2 independant variables)

- $Log(wage) = \beta_0 + \beta_1 Educ + \beta_2 InAbility + u$

Issue if omitted variables (1)

(particular case with 2 independant variables)

- $Log(wage) = \beta_0 + \beta_1 Educ + \beta_2 InAbility + u$
  - Assume that the zero mean assumption is verified here

Issue if omitted variables (1)

(particular case with 2 independant variables)

- $Log(wage) = \beta_0 + \beta_1 Educ + \beta_2 InAbility + u$
  - Assume that the zero mean assumption is verified here
- What if we estimate instead

Issue if omitted variables (1)

(particular case with 2 independant variables)

- $Log(wage) = \beta_0 + \beta_1 Educ + \beta_2 InAbility + u$
  - Assume that the zero mean assumption is verified here
- What if we estimate instead
  - $Log(wage) = \beta_0' + \beta_1' Educ + u'$?

Issue if omitted variables (1)

(particular case with 2 independant variables)

- $Log(wage) = \beta_0 + \beta_1 Educ + \beta_2 InAbility + u$
  - Assume that the zero mean assumption is verified here
- What if we estimate instead
  - $Log(wage) = \beta'_0 + \beta'_1 Educ + u'$?
- Unless $\beta_2 = 0$ or $Cov(InAbility, Educ)$=0, then β_1 is biased

Issue if omitted variables (2)

- Let's write : $InAbility = \delta_0 + \delta_1 Educ + e$

Issue if omitted variables (2)

- Let's write : $InAbility = \delta_0 + \delta_1 Educ + e$
  - with $\delta_1 = \dfrac{Cov(Educ, InAbility)}{V(Educ)}$

Issue if omitted variables (2)

- Let's write : $InAbility = \delta_0 + \delta_1 Educ + e$
  - with $\delta_1 = \dfrac{Cov(Educ, InAbility)}{V(Educ)}$
- Let's replace *Inability* in the first model, the wage equation can be re-written

Issue if omitted variables (2)

- Let's write : $InAbility = \delta_0 + \delta_1 Educ + e$
  - with $\delta_1 = \dfrac{Cov(Educ, InAbility)}{V(Educ)}$
- Let's replace *Inability* in the first model, the wage equation can be re-written
- $Log(wage) = (\beta_0 + \beta_2 * \delta_0) + (\beta_1 + \beta_2 * \delta_1)Educ + (u + \beta_2 * e)$

Issue if omitted variables (2)

- Let's write : $InAbility = \delta_0 + \delta_1 Educ + e$
  - with $\delta_1 = \dfrac{Cov(Educ, InAbility)}{V(Educ)}$
- Let's replace *Inability* in the first model, the wage equation can be re-written
- $Log(wage) = (\beta_0 + \beta_2 * \delta_0) + (\beta_1 + \beta_2 * \delta_1)Educ + (u + \beta_2 * e)$
  - with $E[(u + \beta_2 * e)/educ)] = 0$

Issue if omitted variables (2)

- Let's write : $InAbility = \delta_0 + \delta_1 Educ + e$
  - with $\delta_1 = \frac{Cov(Educ, InAbility)}{V(Educ)}$
- Let's replace *Inability* in the first model, the wage equation can be re-written
- $Log(wage) = (\beta_0 + \beta_2 * \delta_0) + (\beta_1 + \beta_2 * \delta_1)Educ + (u + \beta_2 * e)$
  - with $E[(u + \beta_2 * e)/educ)] = 0$
- $\beta_1 + \beta_2 * \delta_1 \neq \beta_1$ unless $\beta_2 = 0$ or $Cov(Educ, InAbility)$=0

Issue if omitted variables (2)

- Let's write : $InAbility = \delta_0 + \delta_1 Educ + e$
  - with $\delta_1 = \frac{Cov(Educ, InAbility)}{V(Educ)}$
- Let's replace *Inability* in the first model, the wage equation can be re-written
- $Log(wage) = (\beta_0 + \beta_2 * \delta_0) + (\beta_1 + \beta_2 * \delta_1)Educ + (u + \beta_2 * e)$
  - with $E[(u + \beta_2 * e)/educ)] = 0$
- $\beta_1 + \beta_2 * \delta_1 \neq \beta_1$ unless $\beta_2 = 0$ or $Cov(Educ, InAbility)$=0
- Omitting *InAbility* will lead to a biased estimate of the effect of *Educ* on *Logwage* unless $\beta_2 = 0$ or $Cov(Educ, InAbility)$ =0

Issue if omitted variables (2)

- Let's write : $InAbility = \delta_0 + \delta_1 Educ + e$
  - with $\delta_1 = \dfrac{Cov(Educ, InAbility)}{V(Educ)}$
- Let's replace *Inability* in the first model, the wage equation can be re-written
- $Log(wage) = (\beta_0 + \beta_2 * \delta_0) + (\beta_1 + \beta_2 * \delta_1)Educ + (u + \beta_2 * e)$
  - with $E[(u + \beta_2 * e)/educ)] = 0$
- $\beta_1 + \beta_2 * \delta_1 \neq \beta_1$ unless $\beta_2 = 0$ or $Cov(Educ, InAbility)$=0
- Omitting *InAbility* will lead to a biased estimate of the effect of *Educ* on *Logwage* unless $\beta_2 = 0$ or $Cov(Educ, InAbility)$ =0
- The sign of the bias depends on the sign of $\beta_2 * \delta_1$

Coefficient interpretation

- β_1: effect of x_1 on y, ceteris paribus or effect of x_1 on y, net of the influence of x_k

Coefficient interpretation

- β_1: effect of x_1 on y, ceteris paribus or effect of x_1 on y, net of the influence of x_k
- Proof : Frisch-Vaugh theorem (case with two independant variables)

Application : accounting for returns to work experience

- Let's estimate: $Log(wage) = \beta_0'' + \beta_1'' Educ + \beta_2'' WorkExp + u''$

Application : accounting for returns to work experience

- Let's estimate: $Log(wage) = \beta_0'' + \beta_1'' Educ + \beta_2'' WorkExp + u''$
- Interpret β_1''

Application : accounting for returns to work experience

- Let's estimate: $Log(wage) = \beta_0'' + \beta_1'' Educ + \beta_2'' WorkExp + u''$
- Interpret β_1''
- How β_1'' is expected to vary compared to β_1' ?

Application : accounting for returns to work experience (2)

Table 3: Model comparison

	(1)	(2)	(3)	(4)
Years of education	0.14***	0.16***	0.16***	-0.03***
	(0.00)	(0.00)	(0.00)	(0.01)
Potential experience		0.02***	0.04***	0.06***
		(0.00)	(0.00)	(0.00)
Potential experience squarred			-0.00***	-0.00***
			(0.00)	(0.00)
Years of education squarred				0.01***
				(0.00)
Constant	0.46***	-0.15***	-0.41***	-0.14***
	(0.02)	(0.04)	(0.05)	(0.05)
N	6,968	6,968	6,968	6,968
R	0.28	0.31	0.31	0.36

Application : accounting for returns to work experience (3)

Account for non-linearities in the effect of education and of work experience

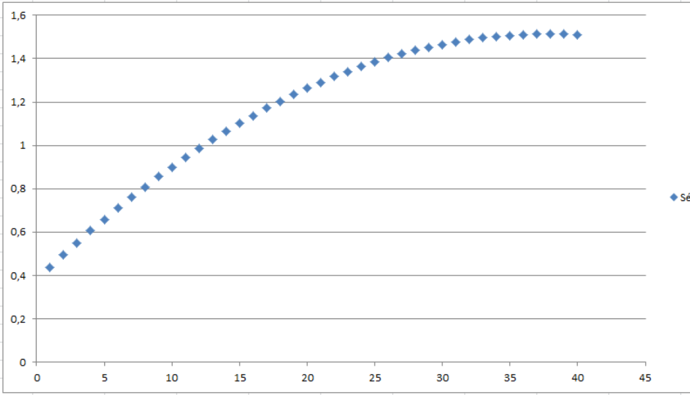
Application : accounting for returns to work experience (4)

Table 4: Model comparison

	(1)	(2)	(3)	(4)
Years of education	0.14***	0.16***	0.16***	-0.03***
	(0.00)	(0.00)	(0.00)	(0.01)
Potential experience		0.02***	0.04***	0.06***
		(0.00)	(0.00)	(0.00)
Potential experience squarred			-0.00***	-0.00***
			(0.00)	(0.00)
Years of education squarred				0.01***
				(0.00)
Constant	0.46***	-0.15***	-0.41***	-0.14***
	(0.02)	(0.04)	(0.05)	(0.05)
N	6,968	6,968	6,968	6,968
R	0.28	0.31	0.31	0.36

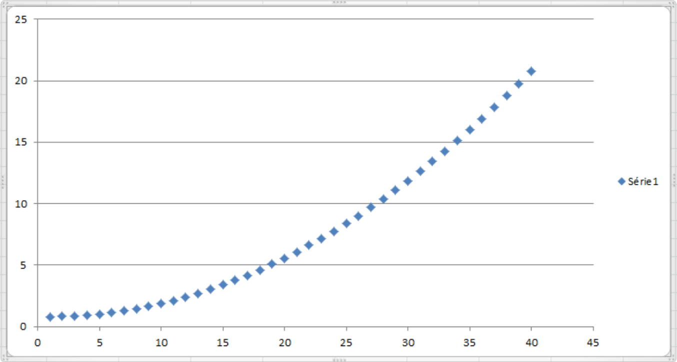
Relationship between $Log(wage)$ and experience (for average value of education)

	LOGWAGE with avg educ level (7,31)
1	0,43662191
2	0,49452191
3	0,55082191
4	0,60552191
5	0,65862191
6	0,71012191
7	0,76002191
8	0,80832191
9	0,85502191
10	0,90012191
11	0,94362191
12	0,98552191
13	1,02582191
14	1,06452191
15	1,10162191
16	1,13712191
17	1,17102191
18	1,20332191
19	1,23402191
20	1,26312191
21	1,29062191
22	1,31652191
23	1,34082191
24	1,36352191
25	1,38462191



Relationship between $Log(wage)$ and education (for average value of experience)

	LOGWAGE with avg work exp (22,52)
1	0,80082568
2	0,81502568
3	0,85542568
4	0,92202568
5	1,01482568
6	1,13382568
7	1,27902568
8	1,45042568
9	1,64802568
10	1,87182568
11	2,12182568
12	2,39802568
13	2,70042568
14	3,02902568
15	3,38382568
16	3,76482568
17	4,17202568
18	4,60542568
19	5,06502568
20	5,55082568
21	6,06282568
22	6,60102568



Interaction term

Understanding Variable Interactions in Multiple Linear Regression

Until now, we've assumed that the effect of each independent variable remains constant, regardless of the values taken by other independent variables in the model. However, it's possible for the effect of a variable, say x_1 or x_2, to vary depending on the values of another variable in the model.

- For instance, the effect of x_1 might change based on the value of x_2.
- This scenario is referred to as an interaction between x_1 and x_2.

Interaction term - Key Points

- **Variable Interaction**: Occurs when the effect of one independent variable on the dependent variable changes depending on the level of another independent variable.
- **Modeling Interactions**: It's crucial to include interaction terms in the regression model when hypothesizing that such dynamics exist between variables, to accurately capture the complexity of their relationships.
- **Implication for Analysis**: Recognizing and modeling interactions allow for a more nuanced understanding of how variables collectively influence the dependent variable, providing insights that would be missed by assuming constant effects.

Interaction between two quantitative variables

```
##
## Regression Results
## ===================================================================
##                                  Dependent variable:
##                     -----------------------------------------------
##                                         lwage
##                               (1)                     (2)
## -------------------------------------------------------------------
## interexpereduc                                      0.003**
##                                                     (0.002)
##
## educ                        0.078***                0.044**
##                             (0.007)                 (0.017)
##
## exper                       0.020***                -0.021
##                             (0.003)                 (0.020)
##
## Constant                    5.503***                5.949***
##                             (0.112)                 (0.241)
##
## -------------------------------------------------------------------
## Observations                  935                     935
## R2                           0.131                   0.135
## Adjusted R2                  0.129                   0.132
## Residual Std. Error    0.393 (df = 932)        0.392 (df = 931)
## F Statistic         70.162*** (df = 2; 932) 48.407*** (df = 3; 931)
## ===================================================================
## Note:                                  *p<0.1; **p<0.05; ***p<0.01
```

Interaction between a quantitative variable and a categorical variable

```
##
## Regression Results
## ==================================================================
##                                   Dependent variable:
##                      ---------------------------------------------
##                                          lwage
##                              (1)                    (2)
## ------------------------------------------------------------------
## intereducsupexper                                -0.022*
##                                                  (0.012)
##
## exper                      0.006*                0.026**
##                           (0.003)                (0.011)
##
## educ_sup                  0.238***               0.558***
##                           (0.048)                (0.181)
##
## Constant                  6.493***               6.193***
##                           (0.066)                (0.177)
##
## ------------------------------------------------------------------
## Observations                935                    935
## R2                         0.026                  0.029
## Adjusted R2                0.024                  0.026
## Residual Std. Error    0.416 (df = 932)       0.416 (df = 931)
## F Statistic         12.379*** (df = 2; 932) 9.392*** (df = 3; 931)
## ==================================================================
## Note:                                  *p<0.1; **p<0.05; ***p<0.01
```

Interaction between 2 categorical variables

```
##
## Regression Results
## ====================================================================
##                                    Dependent variable:
##                     ------------------------------------------------
##                                           lwage
##                                (1)                     (2)
## --------------------------------------------------------------------
## intersouthblack                                      -0.141*
##                                                      (0.083)
##
## black                       -0.248***               -0.165***
##                              (0.041)                 (0.063)
##
## south                       -0.132***               -0.112***
##                              (0.029)                 (0.031)
##
## Constant                    6.856***                6.850***
##                              (0.017)                 (0.017)
##
## --------------------------------------------------------------------
## Observations                   935                     935
## R2                            0.075                   0.077
## Adjusted R2                   0.073                   0.075
## Residual Std. Error     0.406 (df = 932)        0.405 (df = 931)
## F Statistic         37.565*** (df = 2; 932) 26.064*** (df = 3; 931)
## ====================================================================
## Note:                                  *p<0.1; **p<0.05; ***p<0.01
```