Applied econometrics note

Simple Linear Regression Model

Table of Contents

Preface

This set of notes originates from my Applied Econometrics I module at UNNC. Rather than adding new material, I have reorganized the logical flow of the key points presented in the lecture slides. The module is closely linked to ECON1049 Mathematical Economics and Econometrics (SPR)—now renamed ECON1057 Introductory Econometrics—so many topics will already be familiar. If you find something related to statistics or econometrics, please go back to the lecture notes in ECON1049 (/ECON 1057). This will benefit to your memory and smooth your further study.

That said, the course should not be taken lightly. Its true value lies not in memorizing the slide-deck formulas, but in learning how to apply the statistical and econometric theory covered in Year 2 to real economic-data analysis. From a practical standpoint, I strongly recommend gaining proficiency in STATA, as intensive coding will return in Applied Econometrics II and even your final year dissertation.

Because I studied this module on exchange at UNUK, I also completed a piece of coursework. That assignment—requiring the direct application of PPT concepts, extensive STATA work, and a basic research framework—proved immensely beneficial. If you know classmates heading to UNUK, do not hesitate to ask for a look at their coursework. At UNNC, Applied Econometrics I is assessed solely by a final exam; therefore, if you are curious about econometrics or research methods, consider obtaining the UNUK coursework and experimenting with it yourself (perhaps over the summer, or even with GPT's guidance). Treat it as a scaffold for practicing real applications, but ultimately try to write the code on your own.

My guiding principle is to look beyond earning marks (though I know marks are important) for isolated knowledge points and instead ask whether each step you take builds the capabilities you will need next year or even in the future. Much like the Permanent Income Hypothesis in macroeconomics, cultivating a habit of extension and forward thinking will smooth your progression to the next stage, or if you learn repeated game in game theory, you will find out that the choices in each stage will be different depending on whether the game is one-stage or multi-stage (I think that sometimes, in real life, acting in player 1 or 2 in game theory is important, because they always have rational strategy).

This is the synthesized note 1 (4 in total). If you do not have the previous notes, please visit the website: https://github.com/wang95483/notebook. And since this is just a synthesized lecture notes, its function is limited. So if you have any other ideas of how to learn this course, feel free to drop me an email first hmyhw8@nottingham.edu.cn (or if this doesn't work, use wang95483@gmail.com , my personal email).

Introduction to econometrics

Steps in econometric modelling

Definition **Ceteris paribus effect**

As long as we have included **all the important determining factors**, as the effect of a unit change in x_1 on y when all the other factors in the model are held **constant**.

1. Formulate the question(s)

2. Develop an economic model

3. Specify the econometric model

4. Set out the hypotheses that are to be tested

-- Reformulate research question(s) referring to the parameters/coefficients of the econometric model

Example:

Research question 1 Does training in machine maintenance **increase** machinists' productivity?

Question reformulated: Is the parameter β_1 **positive**? or is $\beta_1 > 0$?

Research question 2 Does experience **affect** earnings?

Question reformulated: is $\beta_2 \neq 0$?

一定要严格按照 research question 的方向来设置 hypothesis!!!

As long as we have included all the important determining factors, β_1 can be interpreted as the **ceteris paribus** effect of education on earnings of earnings in the model.

5. Estimate the econometric model

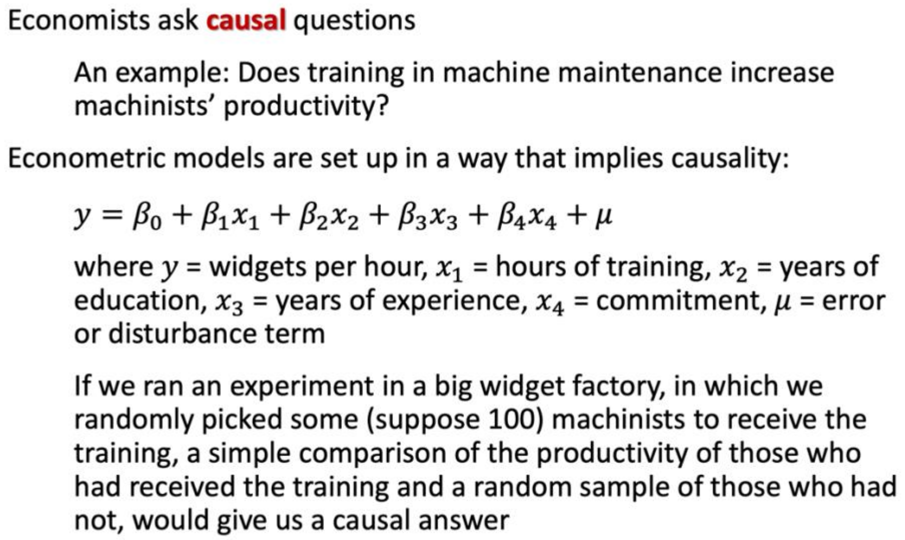
6. Conduct the hypothesis tests

7. Interpret results and draw conclusions

Causality

In the absence of an experiment, econometric analysis can give us a **causal answer**, but only if the estimated coefficients can be interpreted as **ceteris paribus effects**.

Example 1



Economists ask **causal** questions

An example: Does training in machine maintenance increase machinists' productivity?

Econometric models are set up in a way that implies causality:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \mu$$

where y = widgets per hour, x_1 = hours of training, x_2 = years of education, x_3 = years of experience, x_4 = commitment, μ = error or disturbance term

If we ran an experiment in a big widget factory, in which we randomly picked some (suppose 100) machinists to receive the training, a simple comparison of the productivity of those who had received the training and a random sample of those who had not, would give us a causal answer

Example 2

Question: Does education increase earnings in the UK?

People with more innate ability get more education and people who get more education inevitably have less work experience at any given age and innate ability and work experience are also likely to affect earnings, i.e., to be important determining factors.

Model

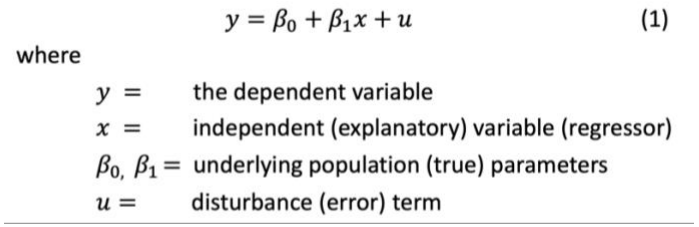
$$y = \beta_0 + \beta_1 Ed + \mu$$

where y = monthly earning, Ed = year of education, and μ = error term

An estimate of β_1 would not be a ceteris paribus effect and could not be used to answer our causal question. Our model also needs to include experience (easy to measure) and innate ability (not easy to measure)

Simple linear regression model

Introduction



$$y = \beta_0 + \beta_1 x + u \qquad (1)$$

where

y =	the dependent variable
x =	independent (explanatory) variable (regressor)
$\beta_{0,}\ \beta_1$ =	underlying population (true) parameters
u =	disturbance (error) term

<u>Valuation</u>

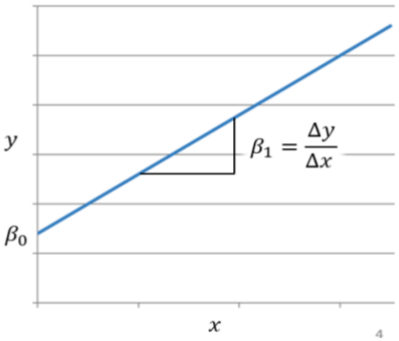
Also referred to as the bivariate (two-variable) linear regression model

Usually we need to include more variables

Often we need to consider non-linear relationships

But we can learn a lot from it that remains relevant when we move on to more complicated multi-variate and non-linear models.

(I) β0 and β1



Graphically, β_0 is the intercept (constant term) and β_1 is the slope parameter of a linear relationship between x and y

If all the other factors (contained in u) are held constant, so that $\Delta u = 0$, then $\Delta y = \beta_1 \Delta x$

Intercept term β_0 may have a useful interpretation, but it is usually just the β_1 slope that we are interested in

(II) u

(i) Definition

u represents the **factors other than x that affect y**, i.e., factors that we have not explicitly accounted for in the model

- there may be many such factors
- some of these factors may be **unobservable** (not possible to measure)
- also accounts for (random**) measurement error** in variables

(ii) Two key assumptions about the disturbance or error term

-reason for the assumption

Obtain an estimate of $\beta 1$, the ceteris paribus effect of x on y

when we are simply ignoring these other factors, by relegating them to u?

• We are not ignoring them: we are going to make a several important

assumptions about them and, it is only if these assumptions are valid that

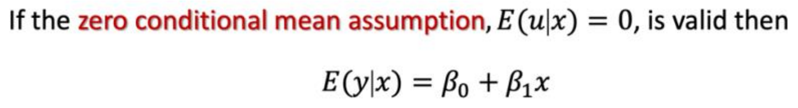
we can interpret our estimate of $\beta 1$ as the ceteris paribus effect of x on y

-Assumptions

zero conditional mean assumption $\boldsymbol{E(u|x) = E(u)}$

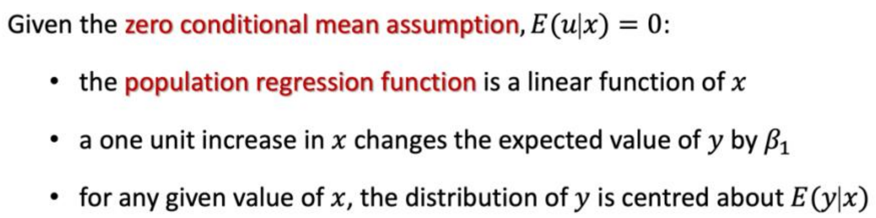
Formula expression	$\boldsymbol{E(u) = 0}$	$\boldsymbol{E(u|x) = E(u)}$
Explanation	On average, the disturbance term is zero	The disturbances are unrelated to the explanatory variable = the average or expected value of u does not depend on the value of x
Valuation	for any single observation it will be positive or negative	knowing something about x does not tell us anything about u

Valuation



If the zero conditional mean assumption, $E(u|x) = 0$, is valid then

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

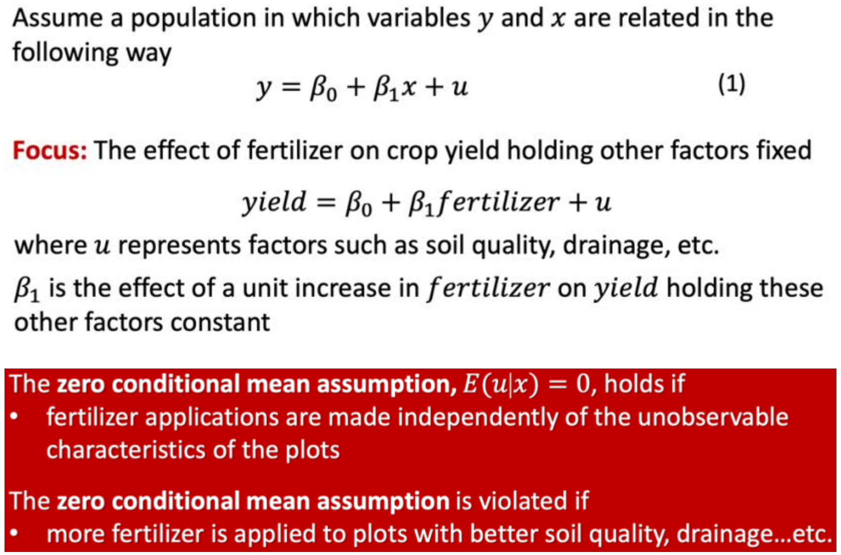


Given the zero conditional mean assumption, $E(u|x) = 0$:

- the population regression function is a linear function of x
- a one unit increase in x changes the expected value of y by β_1
- for any given value of x, the distribution of y is centred about $E(y|x)$

Examine some of the contents later.

Example of simple linear regression model



Assume a population in which variables y and x are related in the following way

$$y = \beta_0 + \beta_1 x + u \qquad (1)$$

Focus: The effect of fertilizer on crop yield holding other factors fixed

$$yield = \beta_0 + \beta_1 fertilizer + u$$

where u represents factors such as soil quality, drainage, etc.

β_1 is the effect of a unit increase in $fertilizer$ on $yield$ holding these other factors constant

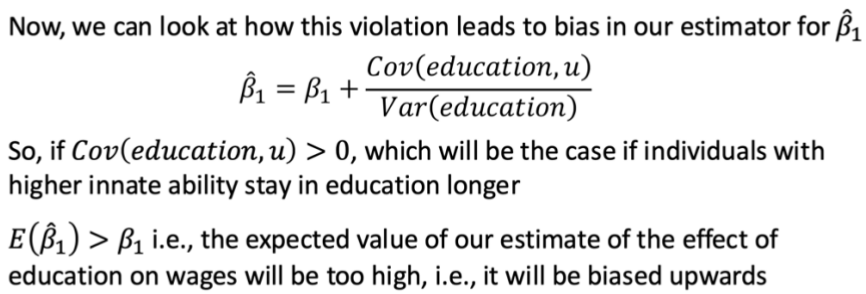
The **zero conditional mean assumption,** $E(u|x) = 0$, holds if

- fertilizer applications are made independently of the unobservable characteristics of the plots

The **zero conditional mean assumption** is violated if

- more fertilizer is applied to plots with better soil quality, drainage…etc.

详细见 unbiased

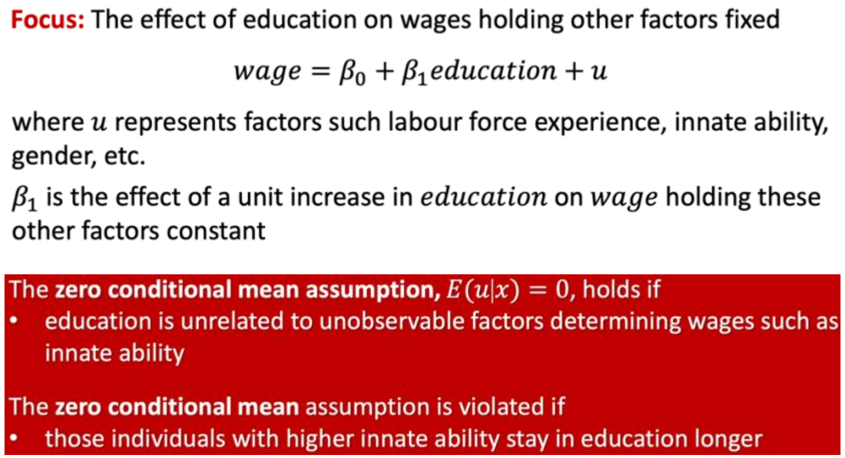


Now, we can look at how this violation leads to bias in our estimator for $\hat{\beta}_1$

$$\hat{\beta}_1 = \beta_1 + \frac{Cov(education, u)}{Var(education)}$$

So, if $Cov(education, u) > 0$, which will be the case if individuals with higher innate ability stay in education longer

$E(\hat{\beta}_1) > \beta_1$ i.e., the expected value of our estimate of the effect of education on wages will be too high, i.e., it will be biased upwards



Focus: The effect of education on wages holding other factors fixed

$$wage = \beta_0 + \beta_1 education + u$$

where u represents factors such labour force experience, innate ability, gender, etc.

β_1 is the effect of a unit increase in $education$ on $wage$ holding these other factors constant

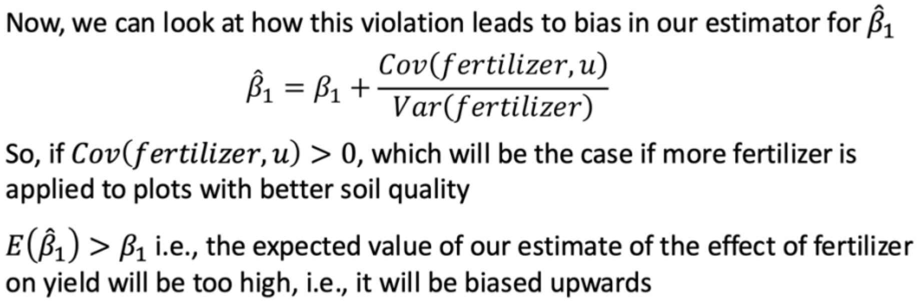
The **zero conditional mean assumption,** $E(u|x) = 0$, holds if

- education is unrelated to unobservable factors determining wages such as innate ability

The **zero conditional mean** assumption is violated if

- those individuals with higher innate ability stay in education longer

详细见 unbiased



Now, we can look at how this violation leads to bias in our estimator for $\hat{\beta}_1$

$$\hat{\beta}_1 = \beta_1 + \frac{Cov(fertilizer, u)}{Var(fertilizer)}$$

So, if $Cov(fertilizer, u) > 0$, which will be the case if more fertilizer is applied to plots with better soil quality

$E(\hat{\beta}_1) > \beta_1$ i.e., the expected value of our estimate of the effect of fertilizer on yield will be too high, i.e., it will be biased upwards

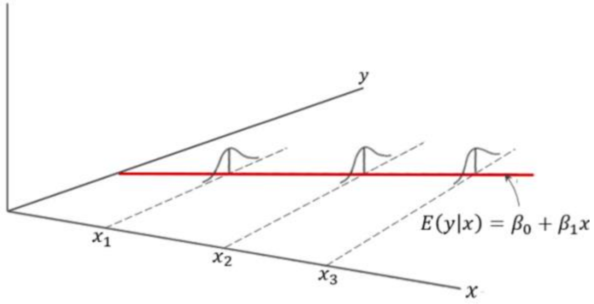
Population regression function

Introduction

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

Given the zero conditional mean assumption, $E(u|x) = 0$:

- the population regression function is a linear function of x
- a one unit increase in x changes the expected value of y by β_1
- for any given value of x, the distribution of y is centred about $E(y|x)$



- It does not tell us that $y = \beta_0 + \beta_1 x$ for every observation in the population
- Some observations will have y values below $E(y|x)$, some will have y values above $E(y|x)$, which depends on u in each specific case
- Because $E(u|x) = 0$ we can **decompose** the data generating process into:
  - A **systematic part** (the PRF): $\beta_0 + \beta_1 x$
    that is attributable to the effect of x in the model
  - A **stochastic (random) part**: u
    that is due to unobservable factors

Estimating $\beta 0$ and $\beta 1$ by Ordinary Least Squares (OLS)

Definition

Suppose we have a random sample of observations from the population

For each sampled observation, i = 1, 2, … , n, we know that

$yi = \beta 0 + \beta 1 xi + ui$

The OLS estimator "finds" the values for $\beta 0$ and $\beta 1$ that **minimise the sum of the squared vertical distances** between each observation (point) and the line

= the OLS estimator minimises the sum of the squared **residuals** (see it in SRF)

Valuation

Intuitive and works amazingly well in practice

simple to use

generalises to the multiple regressor case

it relies on the existence of a population model and corresponding PRF

it only works well under certain conditions, i.e., when our assumptions are valid

→

OLS regression line / sample regression function (SRF)

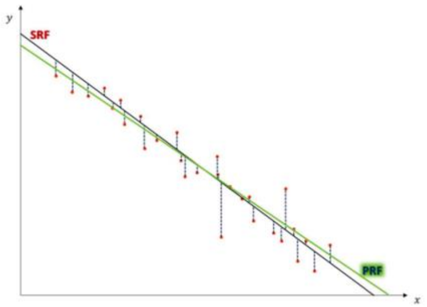
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

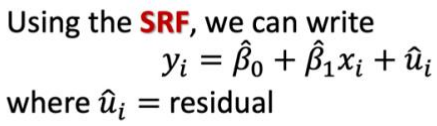
Definition

an estimate of the PRF, $E(y|x) = \beta 0 + \beta 1 x$

distinct from the PRF

specific to the random sample used in its estimation





Using the **SRF**, we can write

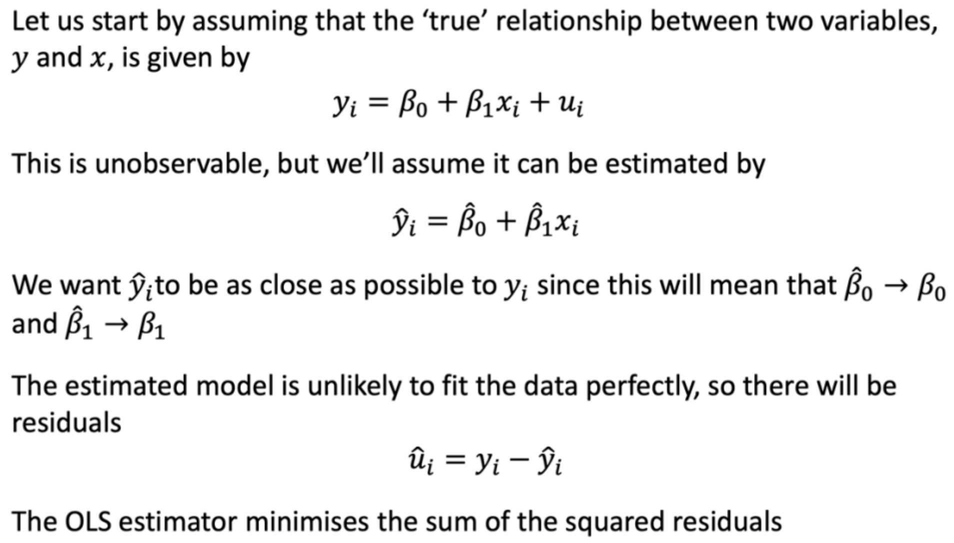
$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i$$

where $\hat{u}_i$ = residual

The residuals are not the same as the disturbances, ui, in the population relationship.

Deriving OLS estimator

Reasons for OLS estimator derivation



Let us start by assuming that the ‘true’ relationship between two variables, y and x, is given by

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

This is unobservable, but we’ll assume it can be estimated by

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

We want $\hat{y}_i$to be as close as possible to y_i since this will mean that $\hat{\beta}_0 \rightarrow \beta_0$ and $\hat{\beta}_1 \rightarrow \beta_1$

The estimated model is unlikely to fit the data perfectly, so there will be residuals

$$\hat{u}_i = y_i - \hat{y}_i$$

The OLS estimator minimises the sum of the squared residuals

详细过程见 MEE

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{Cov(x,y)}{Var(x)}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Desirable Properties of the OLS estimator

These properties relate to the estimator not the estimates themselves:

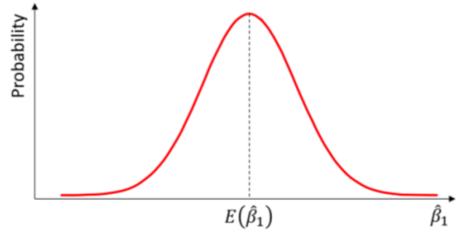
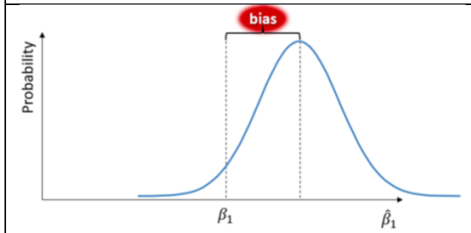
The properties of an estimate also depend on the sample used to generate it. But, without any knowledge of the properties of the sample, we’d always prefer to use an estimator that is unbiased and efficient.

Unbiased

The **most likely estimates** that the estimators produce are the actual population parameters

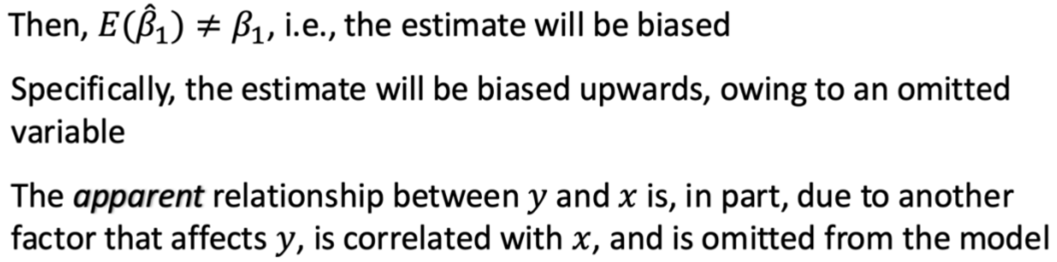
The unbiasedness property relates to the sampling distribution of $\hat{\beta}_1$, not to a specific$\hat{\beta}_1$

	$E(\hat{\beta}_1) = \beta_1$
Explanation	If $E(\hat{\beta}_1) = \beta_1$ the estimator is unbiased
Estimator	$\hat{\beta}_1 = \frac{Cov(x,y)}{Var(x)}$ is a **random variable**
Procedure	见 ppt 和 MEE 书本
Parts	$\hat{\beta}_1 = \beta_1 + \frac{Cov(x,u)}{Var(x)}$
	a **non-random** (deterministic) part that captures the true underlying relationship, i.e., that is equal to the population parameter, $\beta1$
	a **random** (stochastic) part responsible for variations around the population parameter
sampling distribution of $\hat{\beta}_1$	Like any random variable, the set of values it can take is represented by a probability distribution If you drew lots of random samples from the population, used each to calculate an estimate of $\beta1$, and then looked at the distribution of these estimates this is what you would get

	
	If $E(\hat{\beta}_1) \neq \beta_1$ the estimator is biased
	

biased upwards

$$\hat{\beta}_1 = \frac{Cov(x,y)}{Var(x)} = \beta_1 + \frac{Cov(x,u)}{Var(x)}$$



Then, $E(\hat{\beta}_1) \neq \beta_1$, i.e., the estimate will be biased

Specifically, the estimate will be biased upwards, owing to an omitted variable

The ***apparent*** relationship between y and x is, in part, due to another factor that affects y, is correlated with x, and is omitted from the model

Assumptions for unbiased

The OLS estimators are unbiased only under certain conditions, i.e., only if certain assumptions are valid.

	Assumption 1	**Assumption 2**	**Assumption 3**	**Assumption 4**
Content	Linear in Parameters	Random Sampling	Sample variation in the explanatory variable	Zero conditional mean
Explanation	The population model that describes how the data is generated is linear in parameters $y = \beta_0 + \beta_1 x + u$	The sample of size n, $\{(x_i, y_i)$: i = 1, 2, … , $n\}$, taken from the population and used to generate the estimates, is random		$E(u\|x)$ = 0 and that u contains all unobservable factors that affect y
Related formula			$\hat{\beta}_1 = \frac{Cov(x,y)}{Var(x)} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$	$\hat{\beta}_1 = \frac{Cov(x,y)}{Var(x)} = \beta_1 + \frac{Cov(x,u)}{Var(x)}$
Violation			When there is no sample variation in the explanatory variable it is simply not possible to estimate the relationship	$E(u\|x)$ ≠ 0 and, Cov (x, u)> 0 = one of the factors included in u is positively correlated with x →$E(\hat{\beta}_1) \neq \beta_1$ = the estimate will be biased upwards, owing to an omitted Variable which is correlated with x and also affect y
Valuation	(1) this is a good first approximation (2) non-linear relationships are not uncommon, they can be accommodated	(1) this assumption is not always valid (2) important to understand how the data we are using was **generated** (3) throughout the remainder of this module, we will assume random sampling	almost always valid sometimes x does not vary much, this affects the accuracy of our estimates	this assumption is tricky it is often not valid, usually because factors affecting y and correlated with x have been omitted from the model this is why it is important to include all relevant factors in the model and to

				think about what might have been omitted

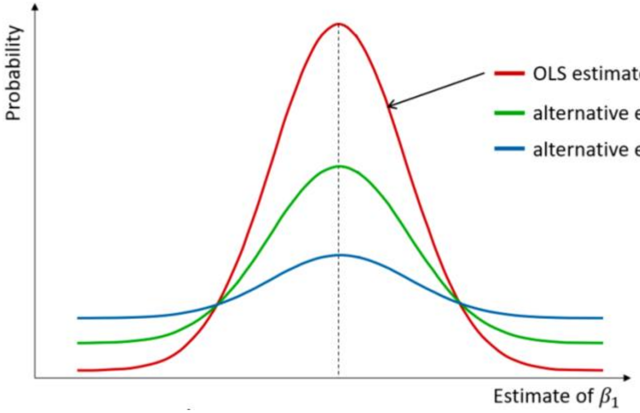
Efficient (‘best’) / minimum variance property

Definition

The **smallest dispersion** (minimum variance) of any linear unbiased estimator.

The **minimum variance** of any linear unbiased estimators.

Measures of the dispersion of the OLS estimators in the SLR model.



Explanation

Compared to other estimators, it achieves the same degree of accuracy with a **smaller sample** of data.

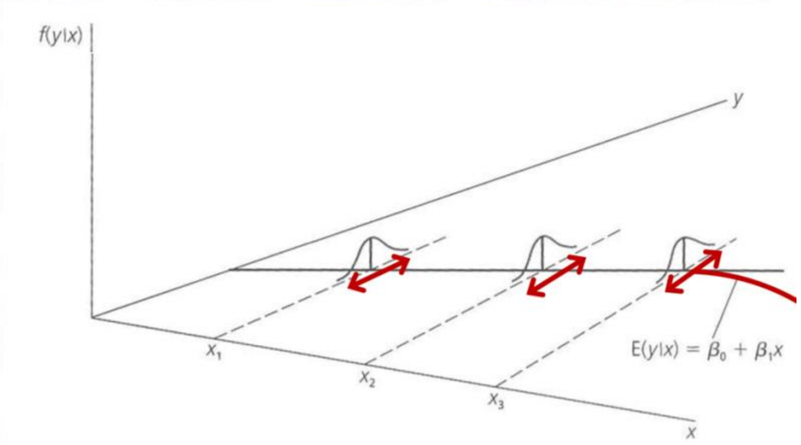
Based on the 4 assumptions, add the

Fifth assumption: Homoskedasticity

Definition

The error u has the **same variance** given any value of the explanatory variable $\mathrm{Var}(u|x) = \sigma^2$.

FIGURE 2.8 The simple regression model under homoskedasticity.



y has a deterministic component, E($y|x$) and a stochastic component, u

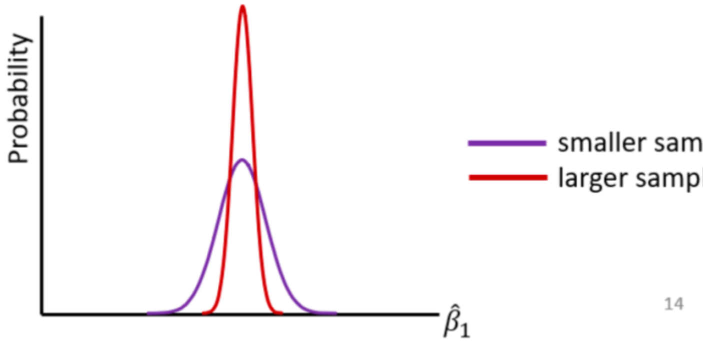
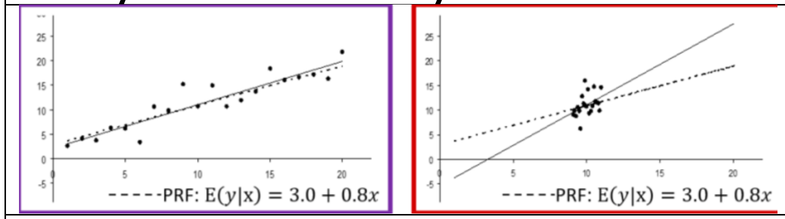
At any given value of x, the distribution of y:

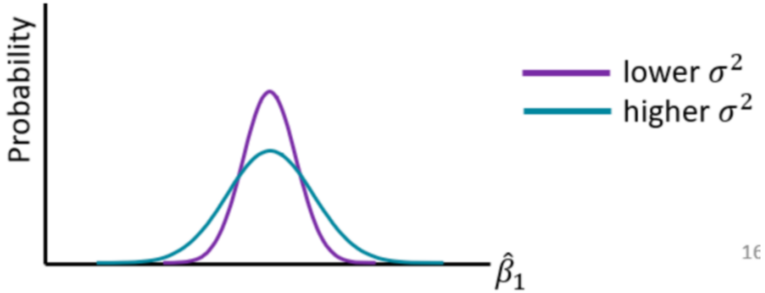
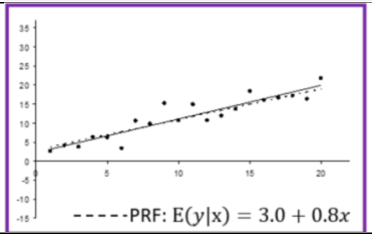
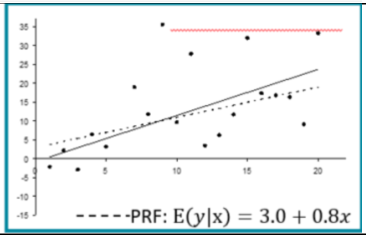
- is centred around E($y|x$) (which varies with x)
- has the same dispersion, i.e., has **constant variance**

Valuation

Probably not valid in realistic.

Under assumptions SLR1: Linear in Parameters and SLR5: homoskedasticity

	$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\sigma^2}{SST_x}$	$Var(\hat{\beta}_0) = \frac{\sigma^2\, n^{-1}\sum_{i=1}^{n}x_i^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$
	It is the relative size (not the absolute size) of these three factors that is important.	
$\sum_{i=1}^{n}(x_i-\bar{x})^2$	**(1)** $Var(\hat{\beta}_1)$ **is inversely related to sample size,** n As n (sample size) increases, $Var(\hat{\beta}_1)$ decreases	
	The more information (data points) we have, the more accurate our estimates are likely to be	
		
	(2) If x **doesn't vary very much, it is difficult to reliably detect its effect on y.**	
		
	on RHS: It is less obvious where the line should be	

	OLS estimators likely to give a less accurate approximation to the true model.	
σ^2	**(3) It is proportional to the variance in the error or disturbance, σ^2** The variance in the stochastic component of y increases, $Var(\hat{\beta}_1)$ increases. The less 'noisy' the relationship is, the more accurate our estimates are likely to be.	
		
	 	
	on RHS: It is less obvious where the line should be OLS. Estimators likely to give a less accurate approximation to the true model.	

The standard errors of OLS estimators

Since $Var(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\hat{\sigma}^2}{SST_x}$ and $Var(\hat{\beta}_0) = \frac{\hat{\sigma}^2 \, n^{-1}\sum_{i=1}^{n} {x_i}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ are measures of the dispersion of the OLS estimators in the SLR model, but they are usual to report the square roots of the variances, i.e., the standard errors, so we have the definition:

Definition of the standard errors of OLS estimators

Estimates of the standard deviations of the sampling distributions of the OLS estimators.

random variables = take different values for each sample of data

Formula

$$\mathrm{se}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad \mathrm{se}(\hat{\beta}_0) = \sqrt{\frac{\hat{\sigma}^2 \, n^{-1}\sum_{i=1}^{n} {x_i}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

The standard error of the regression / the root mean square error

Definition

It is an estimate of the standard deviation of the unobserved factors affecting y.

a measure of the accuracy of the regression as a whole

Intuition

σ^2 is part of the true, but unknown model for both $Var(\hat{\beta}_1) = \dfrac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \dfrac{\sigma^2}{SST_x}$ and $Var(\hat{\beta}_0) = \dfrac{\sigma^2\, n^{-1}\sum_{i=1}^{n} {x_i}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

So, use observed residuals $\hat{u}_i$ in $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i$, now denoted as $\hat{\sigma}^2$, which we can calculate.

Valuation

under reasonable assumptions, the estimator of σ2 is unbiased, i.e., $E(\hat{\sigma}^2) = \sigma^2$

→ then the standard errors of OLS estimators $\mathrm{se}(\hat{\beta}_1) = \sqrt{\dfrac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$ $\mathrm{se}(\hat{\beta}_0) = \sqrt{\dfrac{\hat{\sigma}^2\, n^{-1}\sum_{i=1}^{n} {x_i}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$ are also unbiased.

Multivariable linear regression model

“Partialling out” interpretation

You're interested in how one variable (say, X1) affects another variable (say, Y), but there may be other variables (like X2) that could also influence Y. Partialling out allows you to control for X2, so you can see the "pure" effect of X1 on Y, independent of X2's influence.