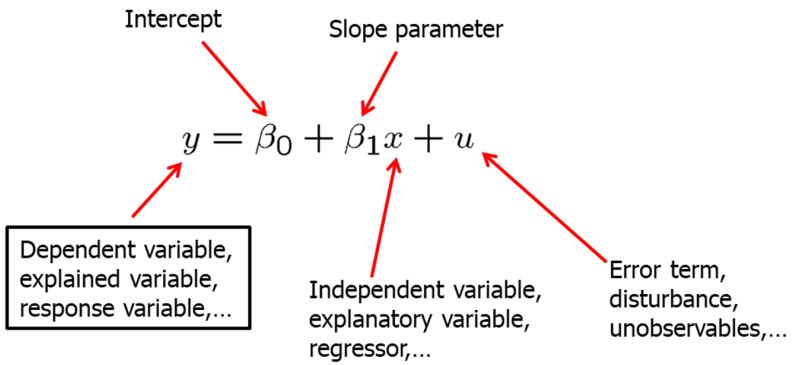
Chapter 2: The Simple Regression Model

Introductory Econometrics: A Modern Approach

2.1 Definition of the Simple Regression Model

"Explains variable y in terms of variable x"



The Simple Regression Model

By how much does the dependent variable change if the independent variable is increased by one unit?

$$\frac{\Delta y}{\Delta x} = \beta_1$$

Interpretation only correct if all other things remain equal when the independent variable is increased by one unit.

$$\frac{\Delta u}{\Delta x} = 0$$

The Simple Regression Model

- Example: Soybean yield and fertilizer

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

Rainfall,
land quality,
presence of parasites, ...

Measures the effect of fertilizer on yield

- Example: A simple wage equation

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

Labor force experience,
tenure with current employer,
work ethic, intelligence, ...

Measures the change in hourly wage
given another year of education

Does β_1 measure a CAUSAL effect of x on y?

- It depends on the relationship with x and u

The Simple Regression Model

Assumption 1: $E(u) = 0$

As long as β_0 is included in the regression, nothing is lost assuming that the average value of u in the population is zero. Why?

- When is there a causal interpretation?

Assumption 2: $E(u|x) = E(u)$

- The independent variable does not contain information about the mean of the unobserved factors
- Average values of the unobservables is the same across all values of x, and that the common average is equal to the average of u over the entire population

The Simple Regression Model

- Assumption 1: $E(u) = 0$
- Assumption 2: $E(u|x) = E(u)$

Combining the assumption 1 and 2:

Zero Conditional Mean Assumption:

$$E(u|x) = 0$$

If zero conditional mean assumption holds, then we have a causal interpretation of our coefficient.

Example:

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

Does $\widehat{\beta_1}$ have a causal interpretation (e.g. does education causes higher wages)? Why or why not?

e.g. Intelligence is part of the error term u

The conditional mean independence assumption is unlikely to hold because individuals with more education will also be more intelligent on average.

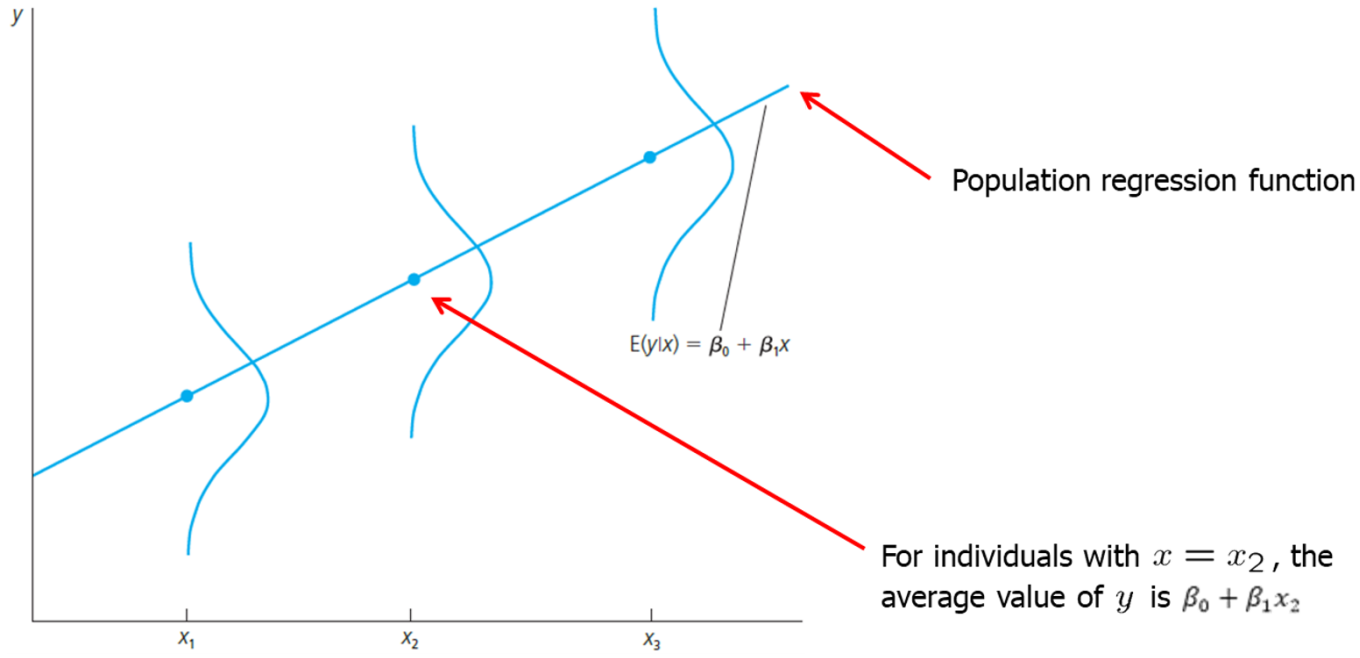
Population regression function (PFR)

- The zero conditional mean independence assumption $E(u|x) = 0$ implies that:

$$\begin{aligned} E(y \mid x) &= E\left(\beta_0 + \beta_1 x + u \mid x\right) \\ &= \beta_0 + \beta_1 x + E(u \mid x) \\ &= \beta_0 + \beta_1 x \end{aligned}$$

- It means that the average value of the dependent variable can be expressed as a linear function of the explanatory variable. This is the *unknown* population function.
- $E(y \mid x)$ tells us how the average value if y changes with x; it does not say that y equals $\beta_0 + \beta_1 x$ for all units in the population
  - Example $E(colGPA|hsGPA) = 1.5 + 0.5 hsGPA$
- Note: One unit increase in x changes the *expected values* of y by the amount of β_1

Population regression function (PFR)



Causation VS Correlation

What is correlation?

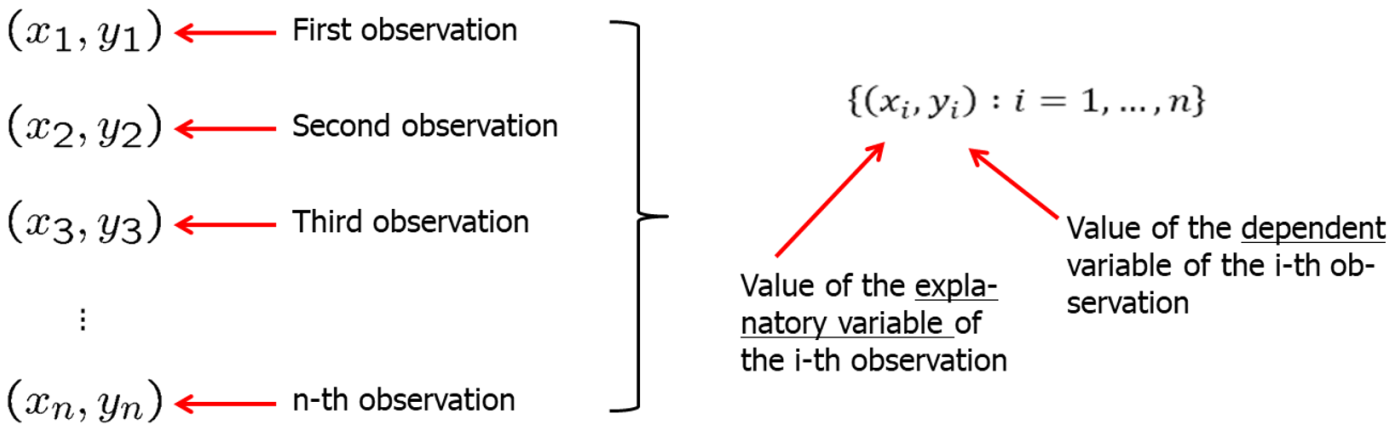
- Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate).
- Spurious correlations LINK

Correlation **DOES NOT IMPLY** causation !!!

- Ceteris Paribus VIDEO
- **The required assumption to establish causality is the zero conditional mean independence** $E(u|x) = 0$

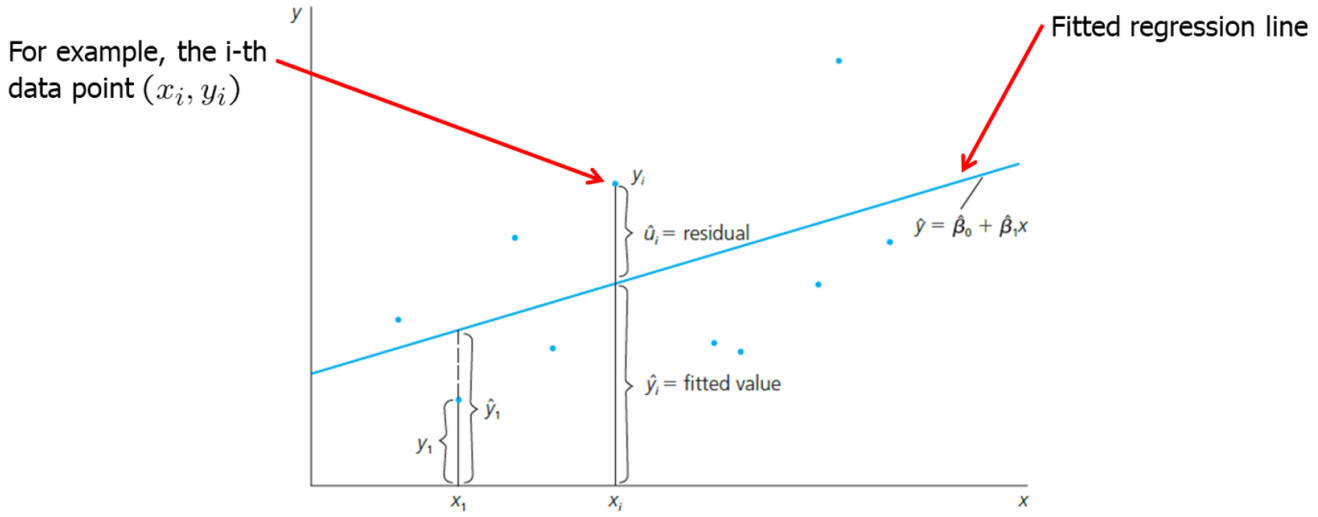
2.2 Deriving the Ordinary Least Square (OLS) Estimates

- In order to estimate the regression model one needs data
- A random sample of n observations



Deriving the Ordinary Least Square (OLS) Estimates

The Ordinary Least Squares method estimates the intercept and slope of a line that “best fits” the observed data by minimizing the sum of the squared distances between the points and the line.



Guess the regression line- SIMULATION

Deriving the Ordinary Least Square (OLS) Estimates (Math)

Let's derive OLS estimates mathematically.

- Define regression **residuals**:

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

- Minimize the sum of the squared regression residuals

$$\min \sum_{i=1}^{n} \widehat{u}_i^2 \rightarrow \widehat{\beta}_0, \widehat{\beta}_1$$

- OLS estimators

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

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

Example 2.3 CEO Salary and Return to Equity

To study the relationship between this measure of firm performance and CEO compensation, we postulate the simple model

$$salary = \beta_0 + \beta_1 roe + u$$

The slope parameter β_1 measures the average change in annual salary, in thousands of dollars, when return on equity increases by one percentage point.

Because a higher roe is good for the company, we think $\beta_1 > 0$.

$$\widehat{\text{salary}} = 963.191 + 18.501 roe$$

- We use “salary hat” to indicate that this is an estimated equation

Example 2.3 CEO Salary and Return to Equity

Deriving the Ordinary Least Square (OLS) Estimates in R

Install *wooldridge* package to have access to the data sets-link

Introductory Econometrics Examples-link

Example 2.4 Wage and Education

```
#install.packages("wooldridge")
library (wooldridge) # need to load the package before using it
library(fixest) # needed to run the regression feols
library(modelsummary)

data ("wage1") # load the data
?wage1 #check out the documentation in the Help panel
# we could use lm package, but feols is useful for future chapters
model<- feols(wage~educ, data=wage1)
#summary(model)
modelsummary(model)
```

Example 2.4 Wage and Education

	Model 1
(Intercept)	-0.905
	(0.685)
educ	0.541
	(0.053)
Num.Obs.	526
R2	0.165
R2 Adj.	0.163
R2 Within	
R2 Pseudo	
AIC	2775.4
BIC	2784.0
Log.Lik.	-1385.712
Std.Errors	IID

```
library(DT)
DT::datatable(head(wage1, 10),
  fillContainer = FALSE, options = list(pageLength = 5))
```

Show 5 entries Search:

	wage	educ	exper	tenure	nonwhite	female	married	numdep	smsa	north
1	3.09999990463257	11	2	0	0	1	0	2	1	
2	3.24000000953674	12	22	2	0	1	1	3	1	
3	3	11	2	0	0	0	0	2	0	
4	6	8	44	28	0	0	1	0	1	
5	5.30000019073486	12	7	2	0	0	1	1	0	

Showing 1 to 5 of 10 entries Previous 1 2 Next

2.3 Properties of OLS on Any Sample of Data

Example: CEO data

Properties of OLS on Any Sample of Data

Fitted (or predicted) values: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_0 x_i$

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

Algebraic properties:

- Deviation from the regression (residuals) line sum up to zero

$$\sum_{i=1}^{n} \widehat{u}_i = 0$$

- Covariance between residuals and independent variables is zero

$$\sum_{i=1}^{n} x_i \widehat{u}_i = 0$$

- Sample averages of y and x lie on the regression line

$$\bar{y} = \widehat{\beta}_0 + \widehat{\beta}_1 \bar{x}$$

Goodness of fit

How well does an independent variable explain the dependent variable?

Total sum of squares (SST)- represents the total variation in the dependent variable

$$SST \equiv \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Explained sum of squares (SSE)- represents variation explained by regression

$$SSE \equiv \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Residual sum of squares (SSR)- represents variation **NOT** explained by regression

$$SSR \equiv \sum_{i=1}^{n} \hat{u}_i{}^2$$

Goodness of fit

Decomposition of total variation:

Total variation= Explained part + Unexplained part

$$SST = SSE + SSR$$

- R-squared measure the fraction of the total variation that is explained by the regression

$$R^2 \equiv \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

Goodness-of-fit measure (R-squared)

	salary_hat	vote_hat
(Intercept)	963.191	26.812
	(213.240)	(0.887)
roe	18.501	
	(11.123)	
shareA		0.464
		(0.015)
Num.Obs.	209	173
R2	0.013	0.856
R2 Adj.	0.008	0.855
R2 Within		
R2 Pseudo		
AIC	3613.1	1134.4
BIC	3619.8	1140.7

Goodness-of-fit measure (R-squared)

Interpretation: $100 * R^2$ is the percentage of sample variation in Y that is explained by X.

- The regression explains only 1.3% of the total variation in salaries
- The regression explains 85.6% of the total variation in election outcomes

What would R^2 be if all data points lie on the same line (perfect fit)?

Note: Low R^2 does not necesarilly mean that the OLS regression is useless !

2.4 Units of measurement and functional form

- **Level- level regression:** $y = \beta_0 + \beta_1 x + \epsilon$

Interpretation: the coefficient β_1 gives us directly the average change in Y for a one-unit change in X

- **Log- level regression:**

$$\ln(y) = \beta_0 + \beta_1 x + \epsilon$$

$$\ln(wage) = \beta_0 + \beta_1 educ + \epsilon$$

$$\beta_1 = \frac{\Delta \ln(\text{ wage })}{\Delta educ} = \frac{1}{wage} \cdot \frac{\Delta \text{ wage}}{\Delta educ} = \frac{\frac{\Delta wage}{\text{wage}}}{\Delta educ}$$

$$\widehat{\ln(wage)} = 0,584 + 0.083\ educ + \epsilon$$

Interpretation: The wage increases by 8.3% for every additional year of education.

Incorporating nonlinearities

- **Level- log regression:**

$$y = \beta_0 + \beta_1 ln(x) + \epsilon$$

Interpretation: The $\frac{\widehat{\beta_1}}{100}$ *can be interpreted as the expected increase in* Y *from a 1% increase in* X

- **Log- log regression:**

$$\ln(y) = \beta_0 + \beta_1 ln(x) + \epsilon$$

Interpretation: The $\widehat{\beta_1}$ *is the expected percentage change in* Y *when* X *increases by some percentage (elasticity)*

Incorporating nonlinearities (summary)

Table 2.3. Summary of Functional Forms Involving Logarithms

Model	Dependent Variable	Independent Variable	Interpretation of β_1
Level-level	y	x	$\Delta y = \beta_1 \Delta x$
Level-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
Log-level	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
Log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

2.5 Expected values and variances of the OLS estimators

Population model: $y = \beta_0 + \beta_1 x + \epsilon$

- $\widehat{\beta_0}, \widehat{\beta_1}$ are the estimators for the parameters β_0, β_1
- $\widehat{\beta_0}, \widehat{\beta_1}$ are random variables with **sampling distributions**
- Sampling distribution is everything we could have got: all the possible results from all the possible samples. It allows us to measure the uncertainty.

Sample Distribution of OLS Estimators- Simulations

KEY IDEA: The greater the sample size, the more confidence you have that your result is close to the true one.

Central Limit Theorem (CLT)

Central Limit Theorem (CLT) states that if you have *any* population and take sufficiently large random samples from the population, then the distribution of the sample means will be approximately normally distributed.

Central Limit Theorem- Simulations

- This simulation demonstrates the shape of the sampling distribution of the sample mean.

Expected values and variances of the OLS estimators

The estimated regression coefficients are random variables because they are calculated from a random sample:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}, \quad \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$$

Data is random and depends on particular sample that has been drawn.

The question is what the estimators will estimate on average and how large will their variability be in repeated samples

$$E\left(\widehat{\beta}_0\right) = ?, E\left(\widehat{\beta}_1\right) = ?$$

$$\mathbf{Var}\left(\widehat{\beta}_0\right) = ?, \mathbf{Var}\left(\widehat{\beta}_1\right) = ?$$

Unbiasedness

$$E(\widehat{\beta_1}) = \beta_1$$

Unbiasedness means that if we could take as many random samples on Y as we want from the population and compute an estimate each time, the average of the estimates would be equal to β_1

There are several assumptions required for OLS estimates to be unbiased

Standard Assumptions for the Linear Regression Model (SLR- Simple Linear Regression)

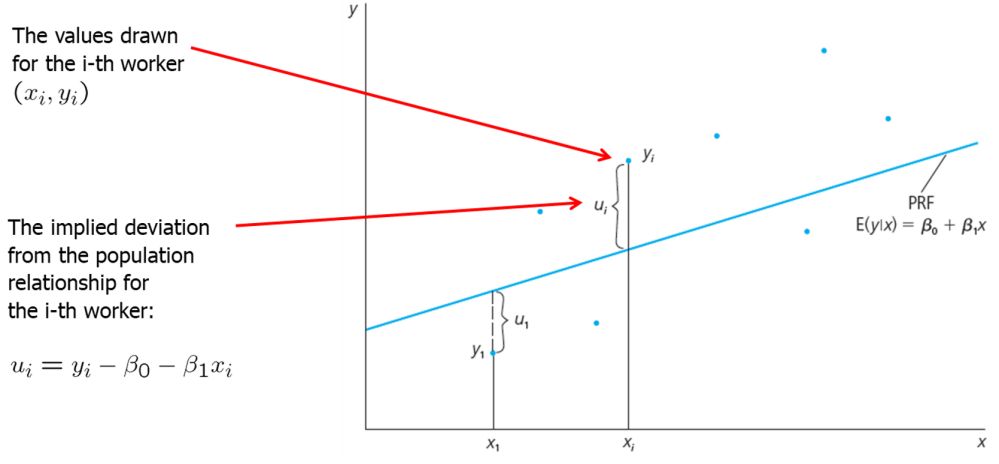
- Assumption SLR 1: **Linear in parameters**

$y = \beta_0 + \beta_1 x + u \longleftarrow$ In the population, the relationship between y and x is linear

- Assumption SLR.2 **Random sampling**

$\{(x_i, y_i) : i = 1, \ldots, n\} \leftarrow$ The data is a random sample drawn from the population

$y_i = \beta_0 + \beta_1 x_i + u_i \longleftarrow$ Each data point therefore follows the population equation



Standard Assumptions for the Linear Regression Model (SLR- Simple Linear Regression)

- Assumption SLR.3 **Sample variation in the explanatory variable**

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 > 0$$

The values of the explanatory variables are not all the same (otherwise it would be impossible to study how different values of the explanatory variable aligned lead to different values of the dependent variable)

- Assumption SLR.4 **Zero conditional mean**

$$E(u_i \mid x_i) = 0$$

The value of the explanatory variable must contain no information about the mean of the unobserved factors

Unbiasedness of the OLS estimates

$$SLR.1 - SLR.4 \Rightarrow E\left(\widehat{\beta}_0\right) = \beta_0, E\left(\widehat{\beta}_1\right) = \beta_1$$

Interpretation of unbiasedness:

- The estimated coefficients may be smaller or larger, depending on the sample that is the result of a random draw.
- However, on average, they will be equal to the values that characterize the true relationship between y and x in the population.
- "On average" means if sampling was repeated, i.e. if drawing the random sample and doing the estimation was repeated many times.
- In a given sample, estimates may differ considerably from true values.

Variances of the OLS estimators

Depending on the sample, the estimates will be nearer or farther away from the true population values.

How far can we expect our estimates to be away from the true population values on average (= sampling variability)?

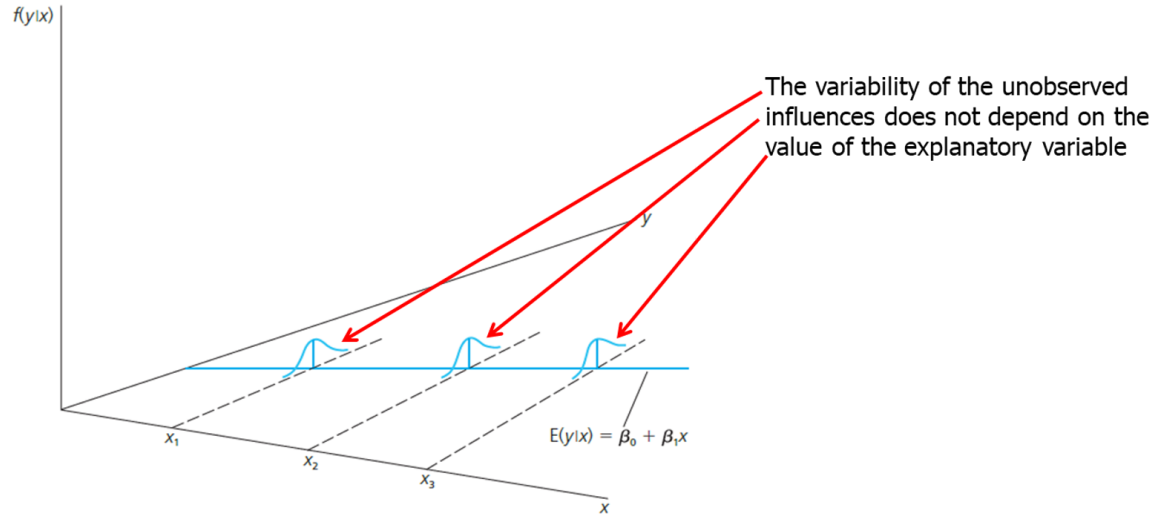
Sampling variability is measured by the estimator's variances

- Assumption SLR5: **Homoskedasticity**

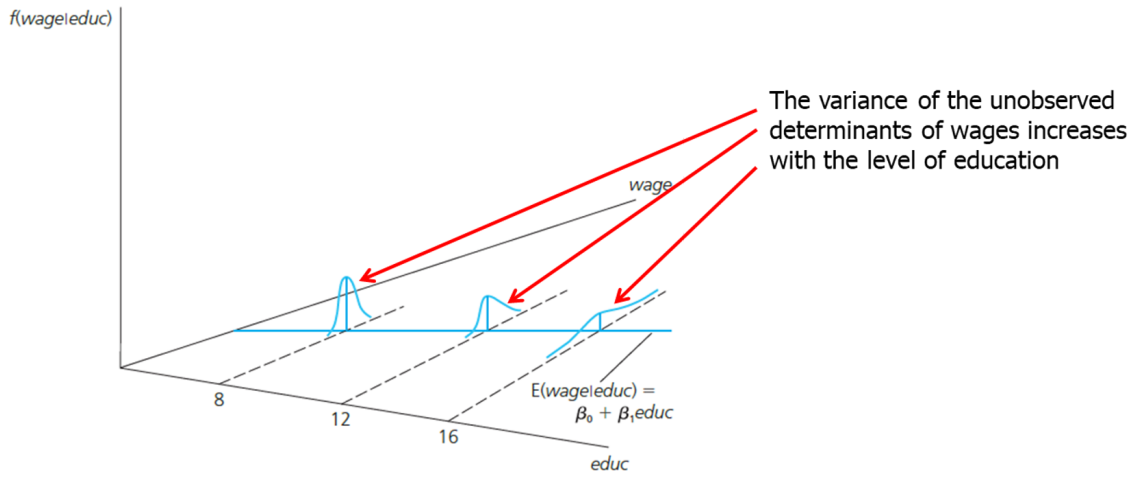
$$\mathrm{Var}(u_i \mid x_i) = \sigma^2$$

- The error term u_i has the **same variance** given an value of the independent variables.

Homoskedasticity VS Heteroskedasticity



The variability of the unobserved influences does not depend on the value of the explanatory variable



The variance of the unobserved determinants of wages increases with the level of education

Variances of the OLS estimators

$$\operatorname{Var}\left(\widehat{\beta}_1\right) = \frac{\sigma^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = \frac{\sigma^2}{SST_x}$$

Factors which affect the variance:

- σ
- SST_x

$$\operatorname{Var}\left(\widehat{\beta}_0\right) = \frac{\sigma^2 n^{-1}\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = \frac{\sigma^2 n^{-1}\sum_{i=1}^{n} x_i^2}{SST_x}$$

Estimating the error variance

The variance of u does not depend on x i.e. equal to the unconditional variance

$$\mathrm{Var}(u_i \mid x_i) = \sigma^2 = \mathrm{Var}(u_i)$$

One could estimate the variance of the errors by calculating the variance of the residuals in the sample; unfortunately this estimate would be biased

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{u}_i - \bar{\hat{u}}_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2$$

An unbiased estimate of the error variance can be obtained by substracting the number of estimated regression coefficients from the number of observations:

$$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\widehat{u}_i^2$$

Unbiasedness of the error variance

$$SLR.1 - SLR.5 \quad \Rightarrow \quad E\left(\widehat{\sigma}^2\right) = \sigma^2$$

- Calculation of standard errors for regression coefficients

$$\operatorname{se}\left(\widehat{\beta}_1\right) = \sqrt{\widehat{\operatorname{Var}}\left(\widehat{\beta}_1\right)} = \sqrt{\hat{\sigma}^2 / SST_x}$$

$$\operatorname{se}\left(\hat{\beta}_0\right) = \sqrt{\widehat{\operatorname{Var}}\left(\hat{\beta}_0\right)} = \sqrt{\hat{\sigma}^2 n^{-1} \sum_{i=1}^{n} x_i^2 / SST_x}$$

The estimated **standard deviations** of the regression coefficients are called **"standard errors."** They measure how precisely the regression coefficients are estimated.