The Simple Regression Model

1-2 Chapter

Ketevani Kapanadze

Brno, 2020

Econometrics I

- Lecturer:
- Ketevani Kapanadze,

Junior Researcher at CERGE-EI, Prague (ketevani.kapanadze@cerge-ei.cz)

- Text-book:
- **The main textbook:**
- Wooldridge, J.M. *Introductory Econometrics – A Modern Approach*. 5th ed. Michigan State University, 2013. ISBN-13: 978-1-111-53104-1.
- Hill, R.C., W.E. Griffiths and G.C. Lim. *Principles of Econometrics*. 4th ed. Hoboken: John Wiley & Sons, 2012. xxvi, 758. ISBN 9780470873724.
- **Supplementary book:**
- Heij, Ch. *Econometric methods with applications in business and economics*. 1st ed. Oxford: Oxford University Press, 2004. xxv, 787. ISBN 9780199268016.

- Lectures/Seminars: Friday 12:00-15:50

Grading

- Midterm exam (***30%***): March 27;
- 2 home assignments (projects) (***20%***): date to be confirmed;
- Final exam (***50%***): May 15.

This course is about using data to measure causal effects.

- Ideally, we would like an experiment

- But almost always we only have observational (nonexperimental) data.

- Most of the course deals with difficulties arising from using observational data to estimate causal effects
  - confounding effects (omitted factors)
  - simultaneous causality
  - “correlation does not imply causation”

In this course you will:

- Learn methods for estimating causal effects using observational data

- Learn some tools that can be used for other purposes; for example, forecasting using time series data;

- Focus on applications – theory is used only as needed to understand the whys of the methods;

- Learn to evaluate the regression analysis of others – this means you will be able to read/understand empirical economics papers in other econ courses;

Intorduction

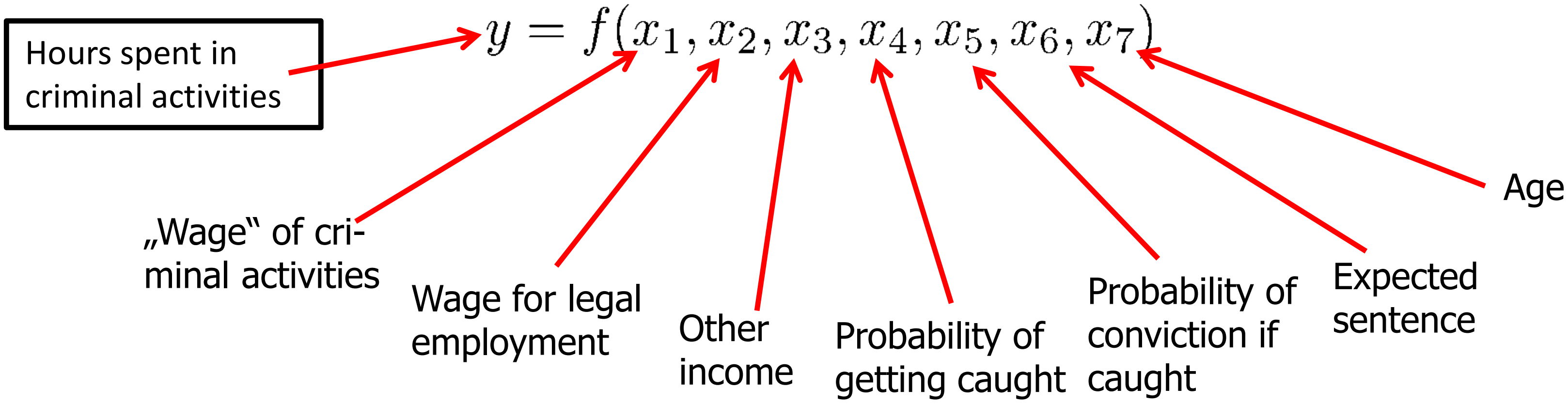
- **What is econometrics?**
  - Econometrics = use of statistical methods to analyze economic data
  - Econometricians typically analyze *nonexperimental* data
- **Typical goals of econometric analysis**
  - Estimating relationships between economic variables
  - Testing economic theories and hypotheses
  - Forecasting economic variables
  - Evaluating and implementing government and business policy

Intorduction

- **Steps in econometric analysis**
  - 1) Economic model (this step is often skipped)
  - 2) Econometric model
- **Economic models**
  - Maybe micro- or macromodels
  - Often use optimizing behaviour, equilibrium modeling, …
  - Establish relationships between economic variables
  - Examples: demand equations, pricing equations, …

Intorduction

- **Economic model of crime (Becker (1968))**
  - Derives equation for criminal activity based on utility maximization



$$y = f(x_1, x_2, x_3, x_4, x_5, x_6, x_7)$$

Hours spent in criminal activities

„Wage“ of criminal activities

Wage for legal employment

Other income

Probability of getting caught

Probability of conviction if caught

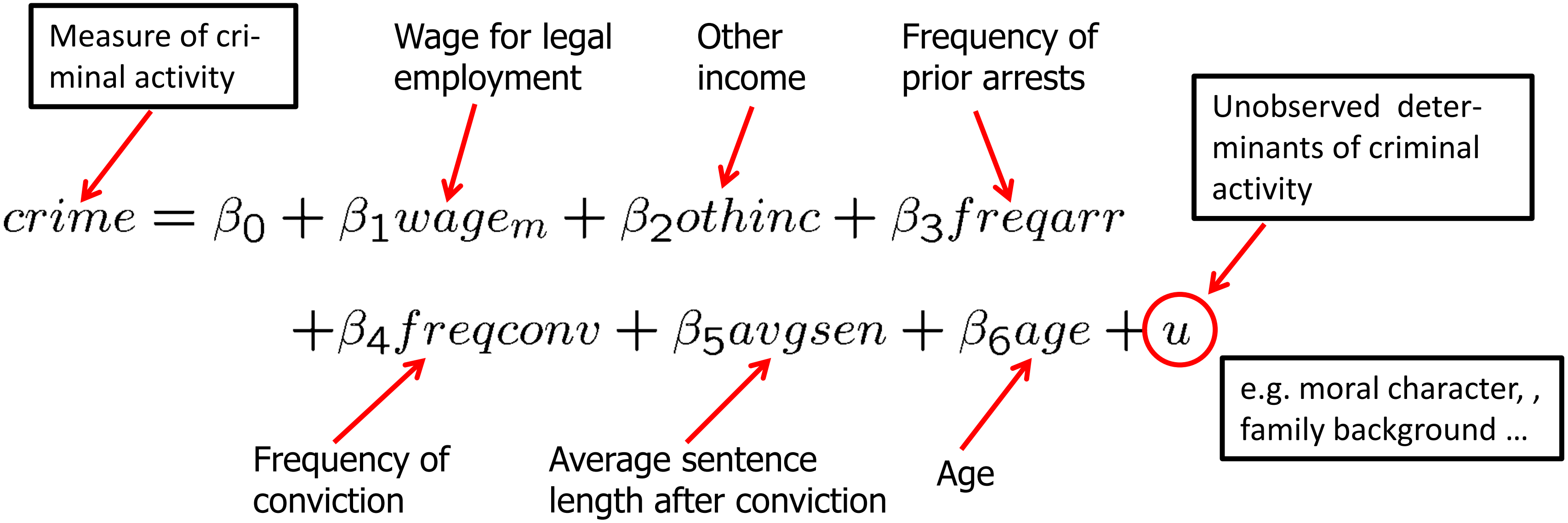
Expected sentence

Age

  - Functional form of relationship not specified
  - Equation could have been postulated without economic modeling

Intorduction

- **Econometric model of criminal activity**
  - The functional form has to be specified
  - Variables may have to be approximated by other quantities



Measure of criminal activity

Wage for legal employment

Other income

Frequency of prior arrests

Unobserved determinants of criminal activity

$$crime = \beta_0 + \beta_1 wage_m + \beta_2 othinc + \beta_3 freqarr$$

$$+ \beta_4 freqconv + \beta_5 avgsen + \beta_6 age + u$$

e.g. moral character, , family background …

Frequency of conviction

Average sentence length after conviction

Age

Intorduction

- **Econometric analysis requires data**
- **Different kinds of economic data sets**
  - Cross-sectional data
  - Time series data
  - Panel/Longitudinal data
- **Econometric methods depend on the nature of the data used**
  - Use of inappropriate methods may lead to misleading results

Intorduction

- <u>**Causality and the notion of ceteris paribus**</u>

Definition of causal effect of x on y:

> „How does variable y changes if variable x is changed <u>but all other relevant factors are held constant</u>“

- **Most economic questions are ceteris paribus questions**
- **It is important to define which causal effect one is interested in**
- **It is useful to describe how an experiment would have to be designed to infer the causal effect in question**

Intordution

- **Effect of law enforcement on city crime level**
  - „If a city is randomly chosen and given ten additional police officers, by how much would its crime rate fall? "
  - Alternatively: „If two cities are the same in all respects, except that city A has ten more police officers, by how much would the two cities crime rates differ?"
- **Experiment:**
  - Randomly assign number of police officers to a large number of cities
  - In reality, number of police officers will be determined by crime rate (simultaneous determination of crime and number of police)

Outline today

1. The population linear regression model
2. The ordinary least squares (OLS) estimator and the sample regression line
3. Measures of fit of the sample regression
4. The least squares assumptions

Linear regression lets us estimate the slope of the population regression line.

- The slope of the population regression line is the expected effect on *Y* of a unit change in *X*.

- Ultimately our aim is to estimate the causal effect on *Y* of a unit change in *X* – but for now, just think of the problem of fitting a straight line to data on two variables, *Y* and *X*.

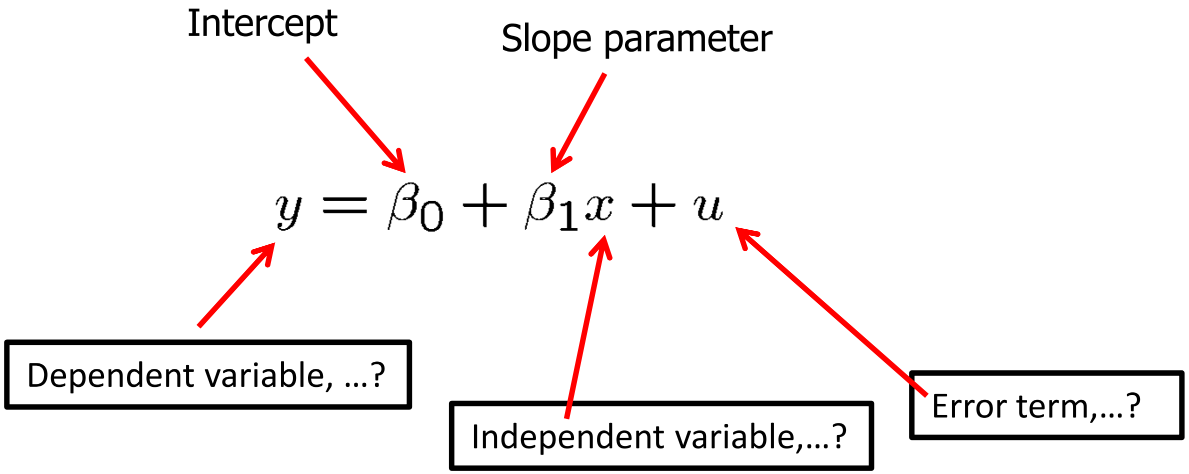
Statistical, or econometric, inference about the slope entails:

- Estimation:
  - How should we draw a line through the data to estimate the population slope?
    - Answer: *ordinary least squares* (OLS).
  - What are advantages and disadvantages of OLS?
- Hypothesis testing:
  - How to test if the slope is zero?
- Confidence intervals:
  - How to construct a confidence interval for the slope?

The Simple Regression Model: Definition

- **Definition of the simple linear regression model**

„Explains variable y in terms of variable x"



Intercept

Slope parameter

$$y = \beta_0 + \beta_1 x + u$$

Dependent variable, …?

Independent variable,…?

Error term,…?

The Simple Regression Model: Definition

- **Interpretation of the simple linear regression model**

„Studies how y varies with changes in x :"

$$\frac{\partial y}{\partial x} = \beta_1 \qquad \text{as long as} \qquad \frac{\partial u}{\partial x} = 0$$

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

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

- **The simple linear regression model is rarely applicable in practice but its discussion is useful for pedagogical reasons**

The Simple Regression Model: Definition

- **Example: Soybean yield and fertilizer**

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

Rainfall, …?

Measures the effect of fertilizer on yield, holding all other factors fixed

- **Example: A simple wage equation**

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

Labor force experience, …?

Measures the change in hourly wage given another year of education, holding all other factors fixed

The Simple Regression Model: Definition

- **When is there a causal interpretation?**
- **Conditional mean independence assumption**

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

The explanatory variable must not contain information about the mean of the unobserved factors

- **Example: wage equation**

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

e.g. intelligence …

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

The Simple Regression Model: Definition

- **Population regression function (PFR)**
  - The conditional mean independence assumption implies that

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

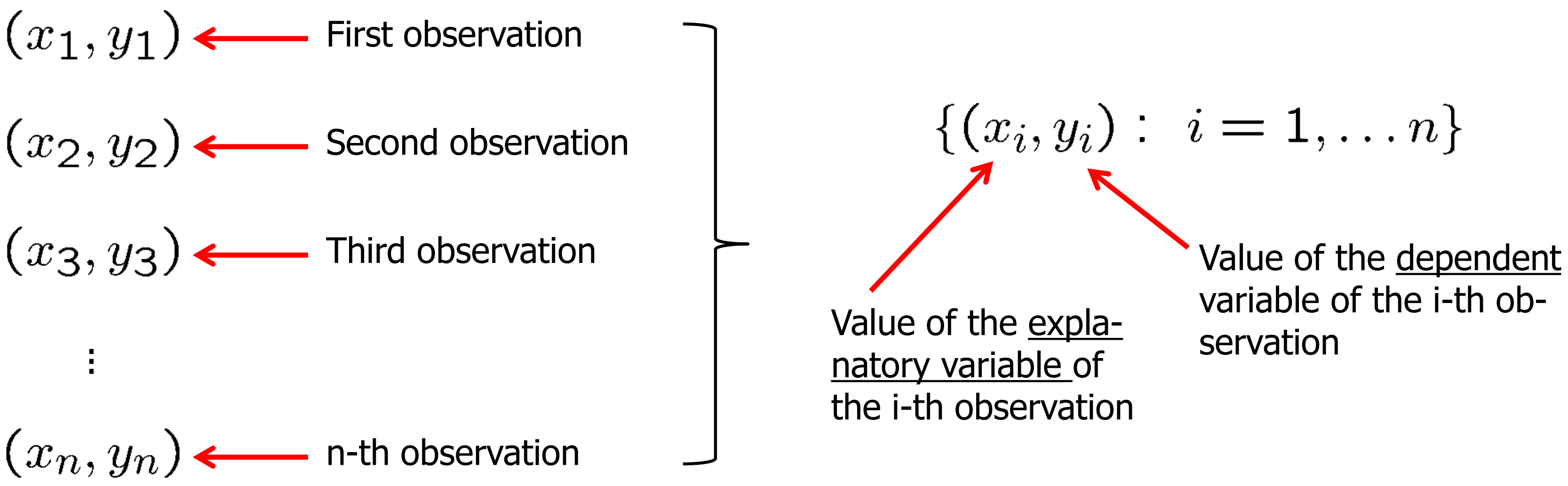
…………….

$$= \beta_0 + \beta_1 x$$

  - This means that the average value of the dependent variable can be expressed as a linear function of the explanatory variable

Deriving OLS Estimates

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



(x_1, y_1) First observation

(x_2, y_2) Second observation

(x_3, y_3) Third observation

$\vdots$

(x_n, y_n) n-th observation

$$\{(x_i, y_i) : \; i = 1, \ldots n\}$$

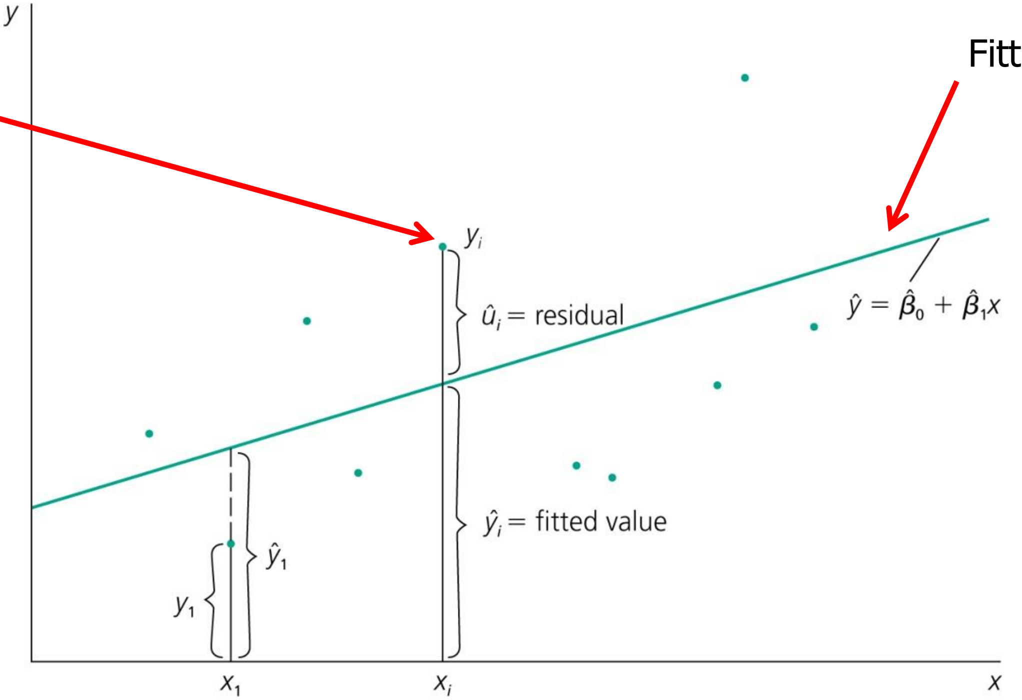
Value of the explanatory variable of the i-th observation

Value of the dependent variable of the i-th observation

Deriving OLS Estimates

- **Fit as good as possible a regression line through the data points:**

For example, the i-th data point


y
y_i
$\hat{u}_i =$ residual
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
$\hat{y}_i =$ fitted value
$\hat{y}_1$
y_1
x_1
x_i
x

Fitted regression line

Deriving OLS Estimates

- **What does „as good as possible“ mean?**
- **Regression residuals**

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

- **Minimize sum of squared regression residuals**

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

- **Ordinary Least Squares (OLS) estimates**

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

Deriving OLS Estimates: Examples Interpretation

- **CEO Salary and return on equity**

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

Salary in thousands of dollars

Return on equity of the CEO's firm

- **Fitted regression**

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

Intercept

- **Causal interpretation?**

Deriving OLS Estimates: Examples Interpretation

- **CEO Salary and return on equity**

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

Salary in thousands of dollars

Return on equity of the CEO‘s firm

- **Fitted regression**

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

Intercept

If the return on equity increases by 1 percent, then salary is predicted to change by 18,501 $

- **Causal interpretation?**

Deriving OLS Estimates: Examples Interpretation

- **Wage and education**

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

Hourly wage in dollars

Years of education

- **Fitted regression**

$$\widehat{wage} = -0.90 + 0.54\ educ$$

Intercept

- **Causal interpretation?**

Deriving OLS Estimates: Examples Interpretation

- **Wage and education**

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

Hourly wage in dollars

Years of education

- **Fitted regression**

$$\widehat{wage} = -0.90 + 0.54 \ educ$$

Intercept

In the sample, one more year of education was associated with an increase in hourly wage by 0.54 $

- **Causal interpretation?**

Deriving OLS Estimates: Examples Interpretation

- **Voting outcomes and campaign expenditures (two parties)**

$$voteA = \beta_0 + \beta_1 shareA + u$$

Percentage of vote for candidate A

Percentage of campaign expenditures candidate A

- **Fitted regression**

$$\widehat{voteA} = 26.81 + 0.464\ shareA$$

Intercept

- **Causal interpretation?**

Deriving OLS Estimates: Examples Interpretation

- **Voting outcomes and campaign expenditures (two parties)**

$$voteA = \beta_0 + \beta_1 shareA + u$$

Percentage of vote for candidate A

Percentage of campaign expenditures candidate A

- **Fitted regression**

$$\widehat{voteA} = 26.81 + 0.464\ shareA$$

Intercept

If candidate A‘s share of spending increases by one percentage point, he or she receives 0.464 percentage points more of the total vote

- **Causal interpretation?**

Properties of OLS

- <u>**Properties of OLS on any sample of data**</u>
- **Fitted values and residuals**

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

Fitted or predicted values

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

Deviations from regression line (= residuals)

- **Algebraic properties of OLS regression**

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

Deviations from regression line sum up to zero

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

Correlation between deviations and regressors is zero

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

Sample averages of y and x lie on regression line

Properties of OLS

- **Goodness-of-Fit**

„How well does the explanatory variable explain the dependent variable?‟

- **Measures of Variation**

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \qquad SSR = \sum_{i=1}^{n} \hat{u}_i^2$$

Total sum of squares, represents total variation in dependent variable

Explained sum of squares, represents variation explained by regression

Residual sum of squares, represents variation not explained by regression

Properties of OLS

- **Decomposition of total variation**

$$SST = SSE + SSR$$

Total variation Explained part Unexplained part

- **Goodness-of-fit measure (R-squared)**

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

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

Properties of OLS: Examples

- **CEO Salary and return on equity**

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

$$n = 209, \quad R^2 = 0.0132$$

The regression explains only 1.3 % of the total variation in salaries

- **Voting outcomes and campaign expenditures**

$$\widehat{voteA} = 26.81 + 0.464\ shareA$$

$$n = 173, \quad R^2 = 0.856$$

The regression explains 85.6 % of the total variation in election outcomes

- **<u>Caution:</u> A high R-squared does not necessarily mean that the regression has a causal interpretation!**

Expected Values and Variance of the OLS

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

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

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 their variability in repeated samples is**

$$E(\hat{\beta}_0) = ?,\ E(\hat{\beta}_1) = ? \qquad Var(\hat{\beta}_0) = ?,\ Var(\hat{\beta}_1) = ?$$

Expected Values and Variance of the OLS

- **Standard assumptions for the linear regression model**
- **Assumption SLR.1 (Linear in parameters)**

$$y = \beta_0 + \beta_1 x + u$$

In the population, the relationship between y and x is linear

- **Assumption SLR.2 (Random sampling)**

$$\{(x_i, y_i) : \ i = 1, \dots n\}$$

The data is a random sample drawn from the population

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

Each data point therefore follows the population equation

Expected Values and Variance of the OLS

- **Assumptions for the linear regression model (cont.)**

- **Assumption SLR.3 (Sample variation in 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 lead to different values of the dependent variable)

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

$$E(u_i|x_i) = 0$$

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

Expected Values and Variance of the OLS

- **Theorem 2.1 (Unbiasedness of OLS)**

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

- **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

Expected Values and Variance of the OLS

- **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

$$Var(\widehat{\beta}_0),\ Var(\widehat{\beta}_1)$$

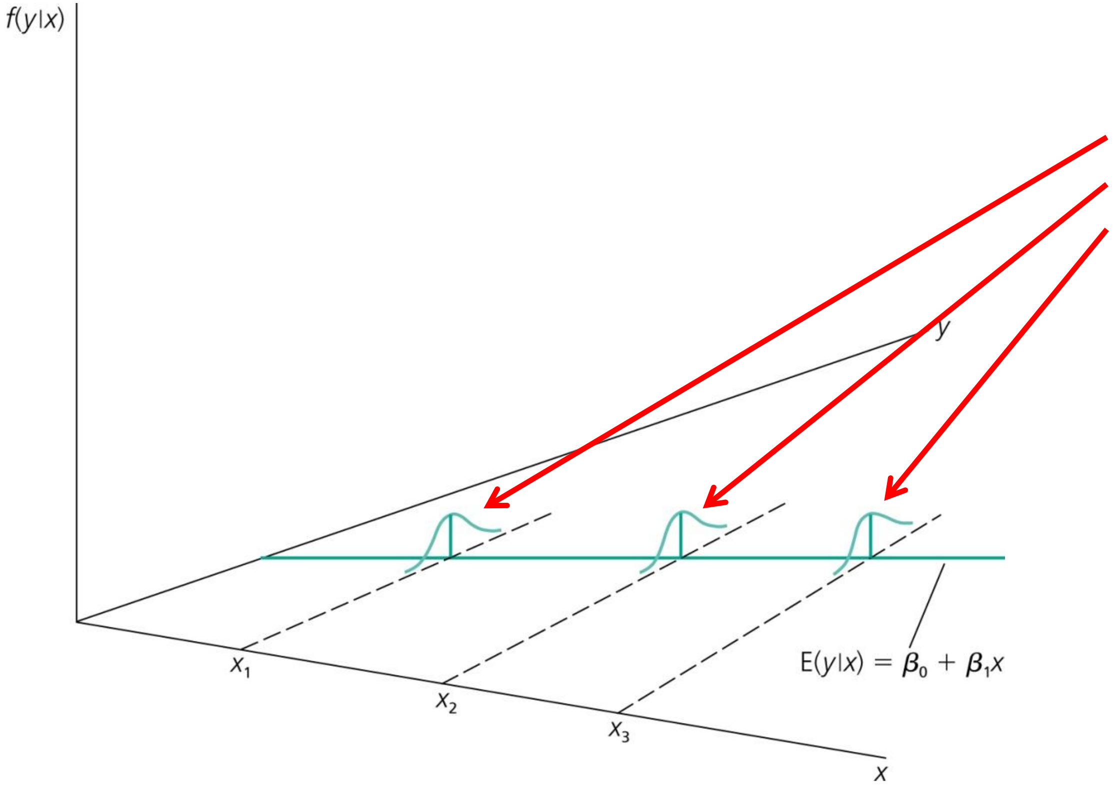
- **Assumption SLR.5 (Homoskedasticity)**

$$Var(u_i|x_i) = \sigma^2$$

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

Expected Values and Variance of the OLS

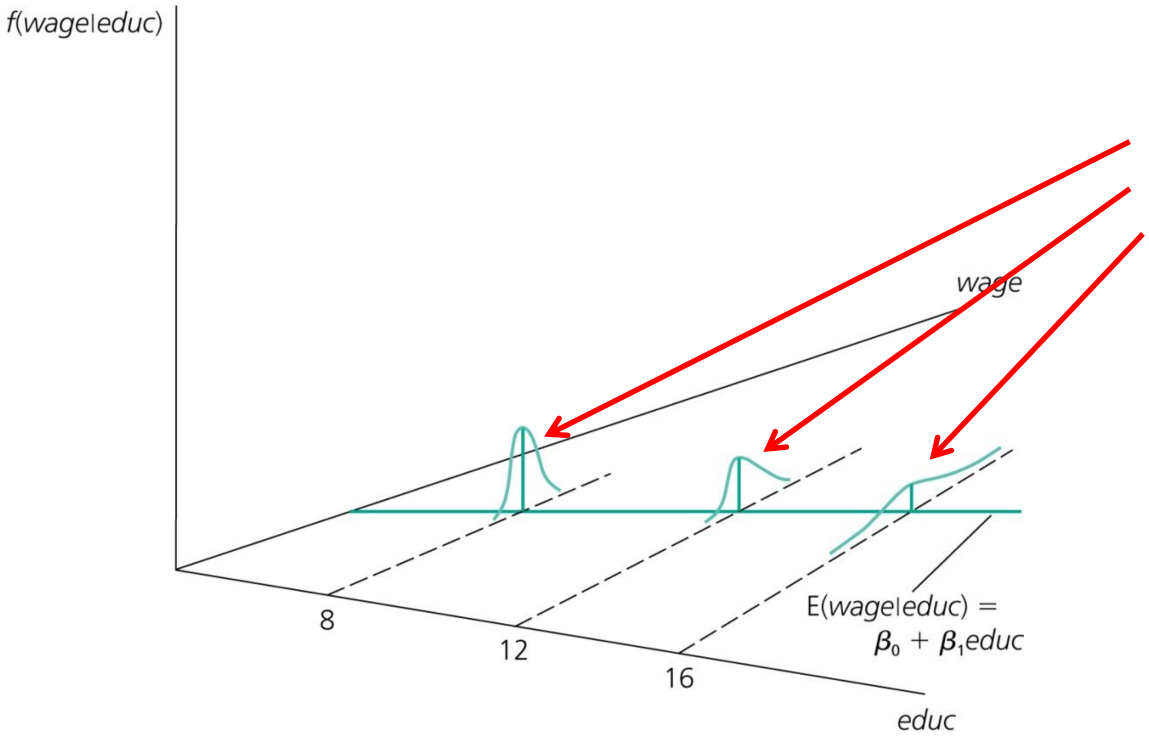
- **Graphical illustration of homoskedasticity**



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

Expected Values and Variance of the OLS

- **An example for heteroskedasticity: Wage and education**



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

Expected Values and Variance of the OLS

- **Theorem 2.2 (Variances of OLS estimators)**

  Under assumptions SLR.1 – SLR.5:

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

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

- **Conclusion:**
  - The sampling variability of the estimated regression coefficients will be the higher the larger the variability of the unobserved factors, and the lower, the higher the variation in the explanatory variable

Expected Values and Variance of the OLS

- **Estimating the error variance**

$$Var(u_i|x_i) = \sigma^2 = Var(u_i)$$

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

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

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

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

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

Expected Values and Variance of the OLS

- **<u>Theorem 2.3 (Unbiasedness of the error variance)</u>**

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

- **Calculation of standard errors for regression coefficients**

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

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

Plug in $\widehat{\sigma}^2$ for the unknown σ^2

The estimated standard deviations of the regression coefficients are called „standard errors“. <u>They measure how precisely the regression coefficients are estimated.</u>

Next Class

- SEMINAR

- Introduction to STATA and some exercises in it.

- **The sampling distribution of the OLS estimator**

- **Modelling issues and further inference in the multiple regression model**