

LECTURE #1

Introductory Econometrics

INTRODUCTION TO THE COURSE & RECAP OF STATISTICAL BACKGROUND

Jiri Kukacka, Ph.D.

Institute of Economic Studies, Faculty of Social Sciences, Charles University

Winter semester 2021, September 29

What is econometrics?

"To beginning students, it may seem as if econometrics is an overly complex obstacle to an otherwise useful education. (...)

To professionals in the field, econometrics is a fascinating set of techniques that allows the measurement and analysis of economic phenomena and the prediction of future economic trends."

(Studenmund, 2016, *Using Econometrics: A Practical Guide*)

What is econometrics?

- ▶ Econometrics is the quantitative measurement of actual economic and business phenomena
- ▶ It attempts to:
 - ▶ quantify economic reality
 - ▶ bridge the gap between the abstract world of economic theory and the real world of human activity
- ▶ It has three major uses:
 1. describing economic reality
 2. testing hypotheses about economic theory
 3. forecasting future economic activity



"Are you just pissing and moaning, or can you verify what you're saying with data?"

Why do we need econometrics?

"We need a special field called econometrics, and textbooks about it, because it is generally accepted that economic data possess certain properties that are not considered in standard statistics texts or are not sufficiently emphasized there for use by economists."

(Studenmund, 2016, *Using Econometrics: A Practical Guide*)

Example

- ▶ Consumer demand for a particular commodity can be thought of as a relationship between:
 - ▶ quantity demanded (Q)
 - ▶ commodity's price (P)
 - ▶ price of substitute good (P_s)
 - ▶ disposable income (Yd)
- ▶ Theoretical functional relationship:

$$Q = f(P, P_s, Yd)$$

- ▶ Econometrics allows us to specify:

$$Q = 27.7 - 0.11P + 0.03P_s + 0.23Yd$$

Introductory Econometrics course

- ▶ **Lecturer:** Jiri Kukacka, Ph.D.
 - ▶ email: jiri.kukacka@fsv.cuni.cz
 - ▶ web: ies.fsv.cuni.cz/en/staff/kukacka
- ▶ **Lectures:** Wednesday, 11:00, lecture hall 314, Opletalova 26
- ▶ **Office hours:** Wednesday, 13:30–15:00, room 406, by appointment via email, please, or we can Meet/Zoom
- ▶ **Seminars:** Thursday, 15:30 and 17:00, room 016
- ▶ **Teaching assistants:**
 - ▶ Jan Sila (jan.sila@fsv.cuni.cz)
 - ▶ Periklis Brakatsoulas (peribrak@gmail.com)
- ▶ **Hybrid form of teaching:**
 - ▶ simultaneous online broadcast via Zoom
 - ▶ links and a passcode in SIS

Introductory Econometrics course

- ▶ **Course requirements:**

- ▶ 3 homework assignments (0-40 points in total: 13+13+14)
- ▶ Final exam online (Moodle part: 0-30 points; oral part: 0-30 points; to pass each part, a student has to achieve at least 15 points; to pass the exam, a student has to pass both parts)
- ▶ Detailed info in SIS: [JEM062](#)

- ▶ **Grading policy (Dean's Measure no. 20/2019):**

- ▶ 90+ to 100 points result in 'A' ('Excellent')
- ▶ 80+ to 90 points result in 'B' ('Very good')
- ▶ 70+ to 80 points result in 'C' ('Good')
- ▶ 60+ to 70 points result in 'D' ('Satisfactory')
- ▶ 50+ to 60 points result in 'E' ('Sufficient')
- ▶ 50 or less points result in 'F' (not passed)

Introductory Econometrics course

- ▶ **Homework assignments:**

- ▶ Announced via SIS on Wednesdays, due in 8 days via SIS
- ▶ Teams of two
- ▶ ‘Academic integrity’ required
- ▶ Results and feedback via SIS

- ▶ **Final exam online:**

- ▶ Moodle part (open-book) —> oral part (closed-book)
- ▶ Good and stable internet connection and a webcam required
- ▶ Term 1: January 12 (Wed), 2022, 11:00
- ▶ Term 2: January 19 (Wed), 2022, 18:00
- ▶ Term 3: January 26 (Wed), 2022, 11:00
- ▶ Term 4: February 2 (Wed), 2022, 18:00
- ▶ Enrolment and results via SIS
- ▶ Specific details will be provided at the beginning of December

Introductory Econometrics course

- ▶ **Course materials:**

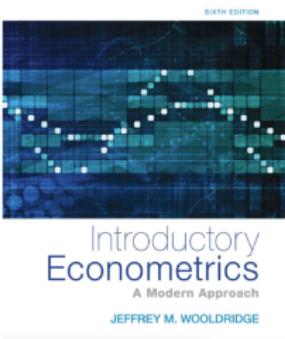
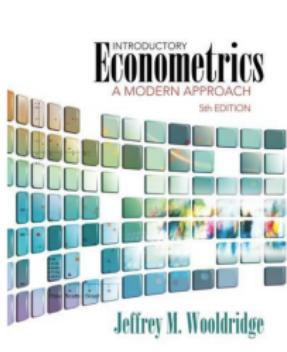
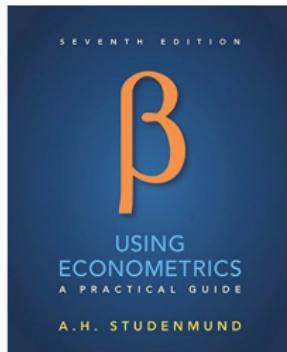
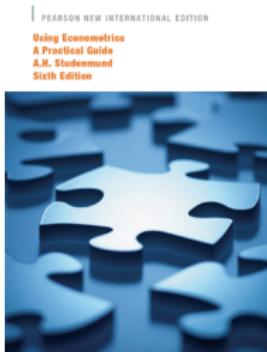
- ▶ Lecture handouts (Mo)
- ▶ Seminar handouts and datasets (Mo)
- ▶ A sketch of solutions and answers to exercises (Thu or Fri)
- ▶ Solutions to home assignments (Fri)
- ▶ A specimen final exam (beginning of December)

- ▶ **Software:**

- ▶ [Gretl](#) (free, open-source, all platforms)
- ▶ Excel ([provided to students](#) by the Faculty)

Core textbooks

- ▶ Studenmund, A. H. (2016). *Using Econometrics: A Practical Guide*. Pearson Educ., 7th Edition (pdf)
- ▶ Wooldridge, J. M. (2016). *Introductory Econometrics: A Modern Approach*. Cengage, 6th Edition (pdf)



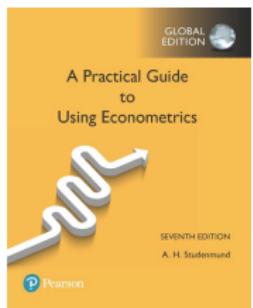
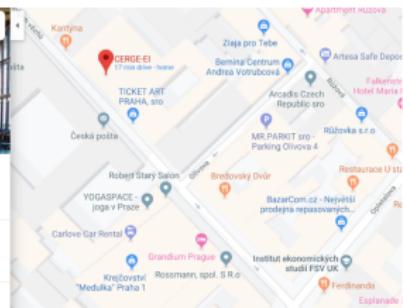
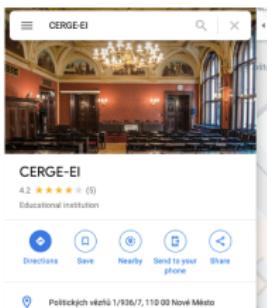
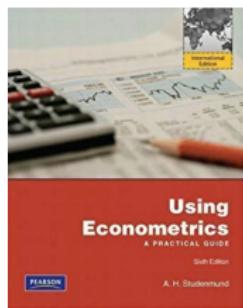
Libraries

IES library: ies.fsv.cuni.cz/en/node/189

CERGE-EI library: olib.cerge.cuni.cz



e-book (7th Ed. 2017)



Course content

- ▶ **Lectures and HAs (tentative):**

- ▶ Lecture 1: Intro & Recap of statistical background
- ▶ Lectures 2–6: Linear regression model & Hypotheses testing, HA#1 & HA#2
Lecture 5, October 27: **ONLINE ONLY**
- ▶ Lectures 7–10: Violations of model assumptions, HA#3
(no lecture on November 17)
- ▶ Lecture 11: Introduction to qualitative dependent variables
- ▶ Lecture 12: Revision, Questions & Answers
- ▶ Detailed info in SIS: [JEM062](#)

- ▶ **Seminars (exercise sessions, practicals):**

- ▶ Serve to clarify and apply concepts presented in lectures
- ▶ Both 'pen and paper' and software exercises
- ▶ Discussion
- ▶ October 28: **ONLINE ONLY**; November 18: no seminars

Lecture #1

- ▶ **Recap of statistical background**
 - ▶ Probability theory
 - ▶ Statistical inference
- ▶ Readings:
 - ▶ Studenmund (2016 & 17, [2014]): Chapters 1.1, [15/17 Statistical Principles]
 - ▶ Wooldridge (2016, 2012): Appendix B, C-1–C-3

Random variables

- ▶ A **random variable** X is a variable whose numerical value is determined by chance. It is a quantification of the outcome of a random phenomenon
 - ▶ A sum of random variables is a random variable too...
- ▶ **Discrete random variable:** has a countable number of possible values
 - ▶ Example: the number of times that a coin will be flipped before a head is obtained, gender, outcome of rolling dice
- ▶ **Continuous random variable:** can take on any value in an interval
 - ▶ Example: height, temperature, speed, individual wealth, time until a breakdown of an engine

Discrete random variables

- ▶ Described by listing the possible values and the associated probability that it takes on each value
- ▶ **Probability distribution** of a variable X that can take values x_1, x_2, x_3, \dots :

$$P(X = x_1) = p_1$$

$$P(X = x_2) = p_2$$

$$P(X = x_3) = p_3$$

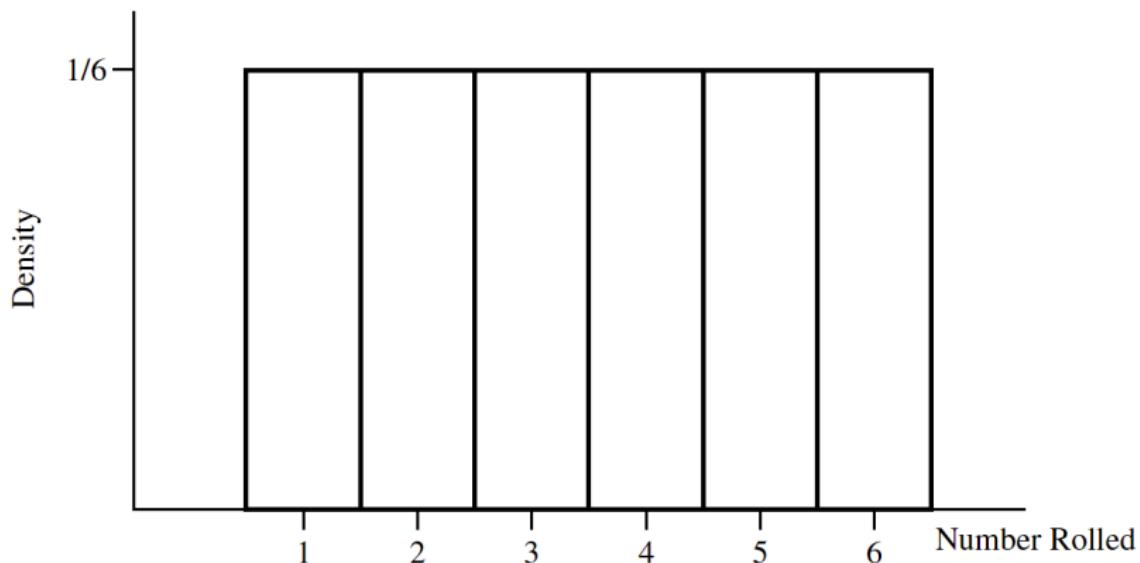
⋮

$$\sum_{i=1} P(X = x_i) = 1$$

- ▶ **Cumulative distribution function (CDF):**

$$F_X(x) = P(X \leq x) = \sum_{i=1, x_i \leq x} P(X = x_i)$$

Six-sided die: probability distribution

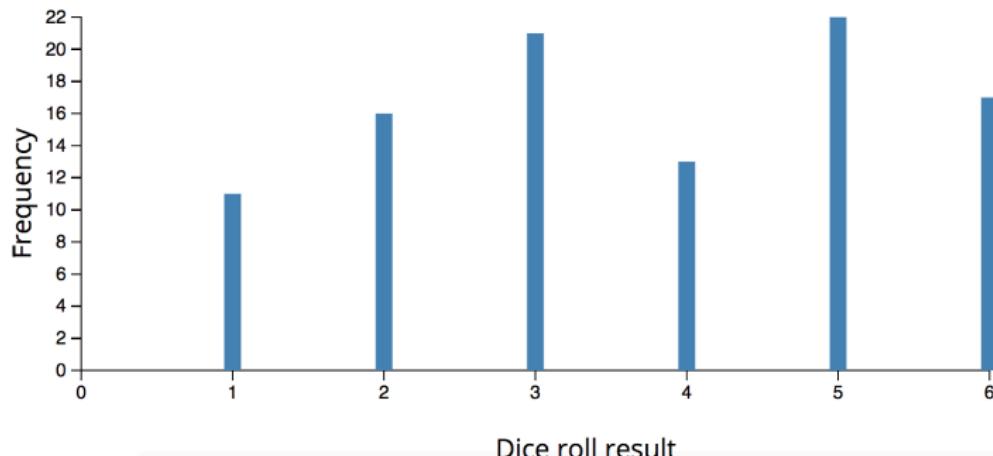


Source: Studenmund (2014, pg. 509)

Six-sided die: histogram of data (100 rolls)



Number of rolls: 100

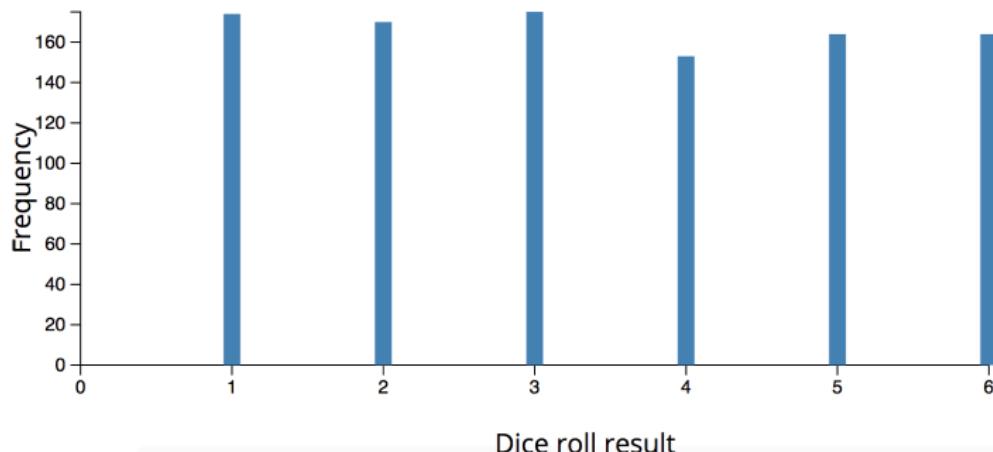


Source: academo.org/demos/dice-roll-statistics

Six-sided die: histogram of data (1000 rolls)



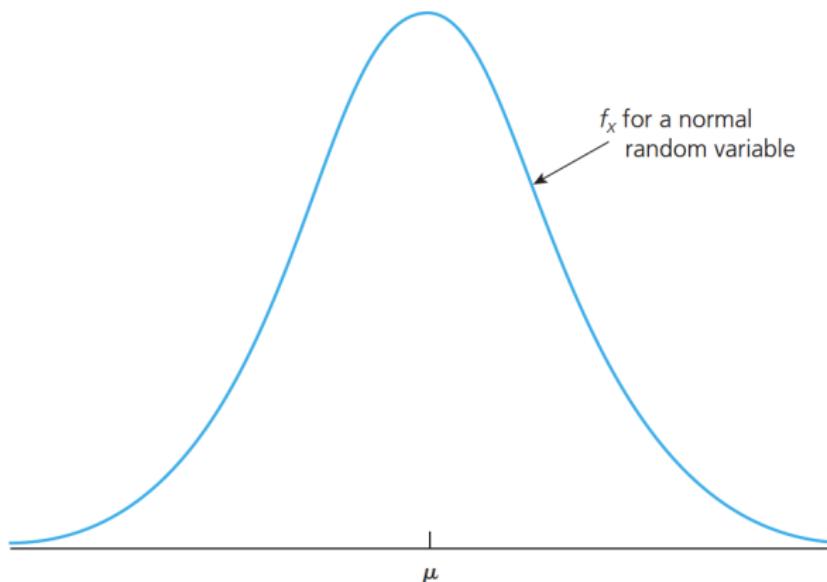
Number of rolls: 1000



Source: academo.org/demos/dice-roll-statistics

Continuous random variables

Probability density function (PDF) $f_X(x)$ describes the relative likelihood for the random variable X to take on a particular value x

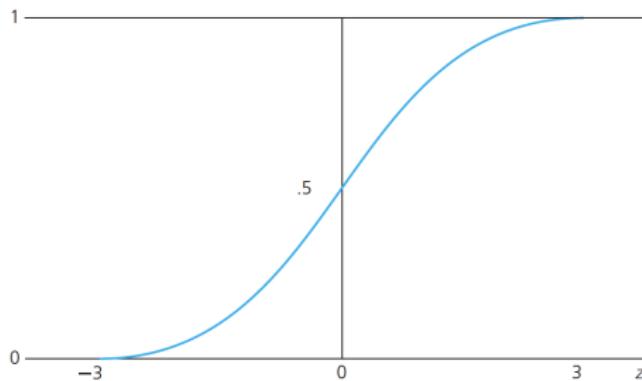


Source: Wooldridge (2016, pg. 666)

Continuous random variables

Cumulative distribution function (CDF):

$$F_X(x) = P(X \leq x) = \int_{-\infty}^x f_X(t)dt$$



Source: The standard normal CDF, Wooldridge (2016, pg. 667)

Computational rule: $P(X > x) = 1 - P(X \leq x)$

Expected value vs. median

► Expected value (mean):

- Mean is the (long-run) average value of a random variable
- It is a **weighted** average of all its possible values
- The weights are determined by the PDF

Discrete variable:

$$E[X] = \sum_{i=1} x_i P(X = x_i)$$

Continuous variable:

$$E[X] = \int_{-\infty}^{+\infty} x f_X(x) dx$$

- Example: calculating mean of six-sided die (seminar #1)
- **Median:** 'the value in the middle'

17 rolls D6 (ordered): 1 1 1 1 2 2 3 3 **3** 3 4 4 5 5 5 6 6

Variance and standard deviation

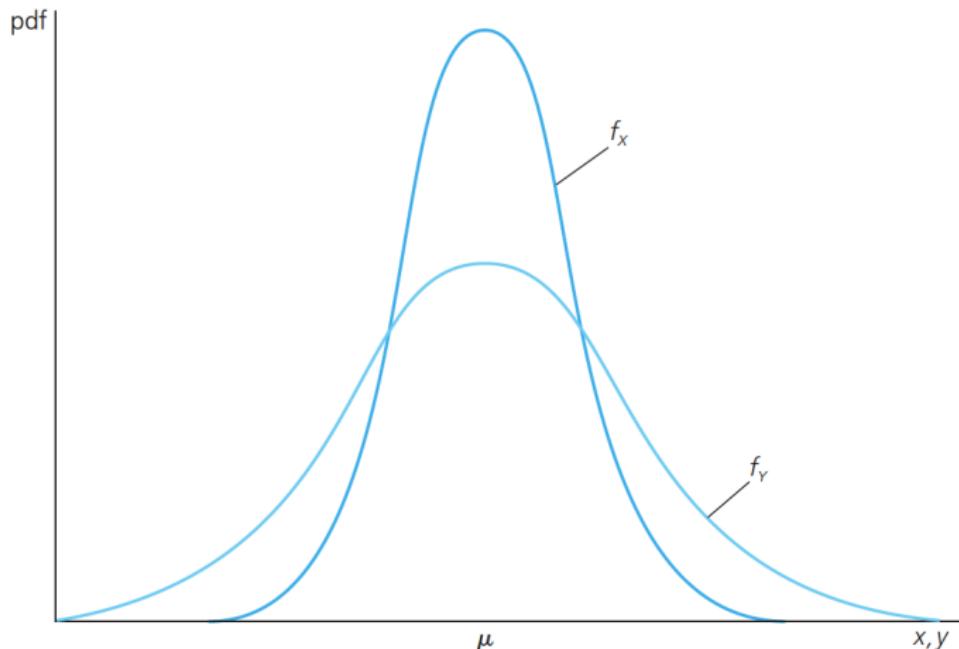
- ▶ **Variance:**

- ▶ Measures the extent to which the values of a random variable are dispersed from its expected value (mean)
- ▶ If values (outcomes) are far away from the mean, variance is high. If they are close to the mean, variance is low

$$\text{Var}[X] = E \left[(X - E[X])^2 \right] = E[X^2] - (E[X])^2$$

- ▶ **Standard deviation:** $\sigma_X = \sqrt{\text{Var}[X]}$

RVs with the same mean but different variances



Source: Wooldridge (2016, pg. 656)

Covariance and correlation

► Covariance:

- ▶ How, on average, two random variables vary with one another
- ▶ Do the two variables move in the same or opposite direction?
- ▶ Measures the amount of linear dependence between two RVs

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

► Correlation:

- ▶ Similar concept to covariance, but easier to interpret
- ▶ It has values between -1 and 1
- ▶ Does not depend on the units of measurement
- ▶ $\text{Corr}(X, Y) = 0 \Rightarrow$ no linear relationship between X and Y

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Independence of variables

- ▶ **Independence:** X and Y are independent if the conditional probability distribution of X given the observed value of Y is the same as if the value of Y had not been observed
- ▶ If X and Y are independent, then $\text{Cov}(X, Y) = 0$ (not the other way round in general)

Sample moments

- ▶ Counterparts of theoretical moments of the distribution of X , computed based on observations X_1, \dots, X_n drawn from this distribution

- ▶ **Sample mean:**

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

- ▶ **Sample variance and standard error:**

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

- ▶ **Sample covariance:**

$$Cov_n(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)$$

Computational rules

$$E[aX + Y + b] = aE[X] + E[Y] + b$$

$$\text{Var}[aX + b] = a^2 \text{Var}[X]$$

$$\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}(X, Y)$$

$$\text{Cov}(aX, bY) = \text{Cov}(bY, aX) = ab\text{Cov}(X, Y)$$

$$\text{Cov}(X + Z, Y) = \text{Cov}(X, Y) + \text{Cov}(Z, Y)$$

$$\text{Cov}(X, X) = \text{Var}[X]$$

where X, Y, Z are random variables and a, b are scalars (constants)

Random vectors

- Sometimes, we deal with vectors of random variables

- Example: $\mathbf{x} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$

- Expected value: $E[\mathbf{x}] = \begin{pmatrix} E[X_1] \\ E[X_2] \\ E[X_3] \end{pmatrix}$

- Variance/covariance matrix:

$$Var[\mathbf{x}] = \begin{pmatrix} Var[X_1] & Cov(X_1, X_2) & Cov(X_1, X_3) \\ Cov(X_2, X_1) & Var[X_2] & Cov(X_2, X_3) \\ Cov(X_3, X_1) & Cov(X_3, X_2) & Var[X_3] \end{pmatrix}$$

- Comp. rule: $Var[\mathbf{Ax}] = \mathbf{A}Var[\mathbf{x}]\mathbf{A}'$, \mathbf{A} is a nonrandom matrix

Selected properties of matrix operations (transpose, inverse; Wooldridge (2016): Appendix D)

$$a' = a$$

$$(\mathbf{A}')' = \mathbf{A}^*$$

$$(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'$$

$$(\mathbf{AB})' = \mathbf{B}'\mathbf{A}^{**}$$

$$(\mathbf{ABC})' = \mathbf{C}'\mathbf{B}'\mathbf{A}'$$

Def.: $\mathbf{D}' = \mathbf{D} \Leftrightarrow \mathbf{D}$ symmetric

$\mathbf{A}'\mathbf{A}$... symmetric matrix: $(\mathbf{A}'\mathbf{A})'^{**} \stackrel{*}{=} \mathbf{A}'(\mathbf{A}')' \stackrel{*}{=} \mathbf{A}'\mathbf{A}$

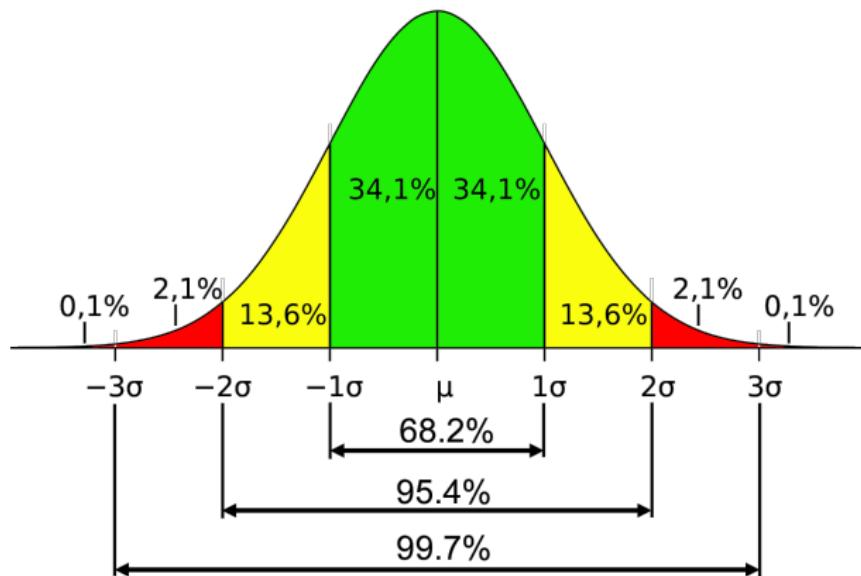
$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$$

$$(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$$

where $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ are matrices and a is a scalar (constant)

Normal (Gaussian) distribution

- Notation: $X \sim N(\mu, \sigma^2)$
- $E[X] = \mu$
- $Var[X] = \sigma^2$



Source: www.muelaner.com

Normal (Gaussian) distribution

- ▶ The most widely used distribution in statistics and econometrics
- ▶ Certain random variables appear to roughly follow a normal distribution: human heights and weights, test scores (IQ, grades), leaves of trees, country unemployment rates...
- ▶ **Central Limit Theorem:** the sum (or the mean) of a number of independent, identically distributed random variables will tend to be normally distributed, **regardless of their distribution**, if their number is large enough

Standardized random variable

- ▶ Standardization is used for better comparison of different variables
- ▶ Define Z to be the standardized variable of X :

$$Z = \frac{X - E[X]}{\sigma_X}$$

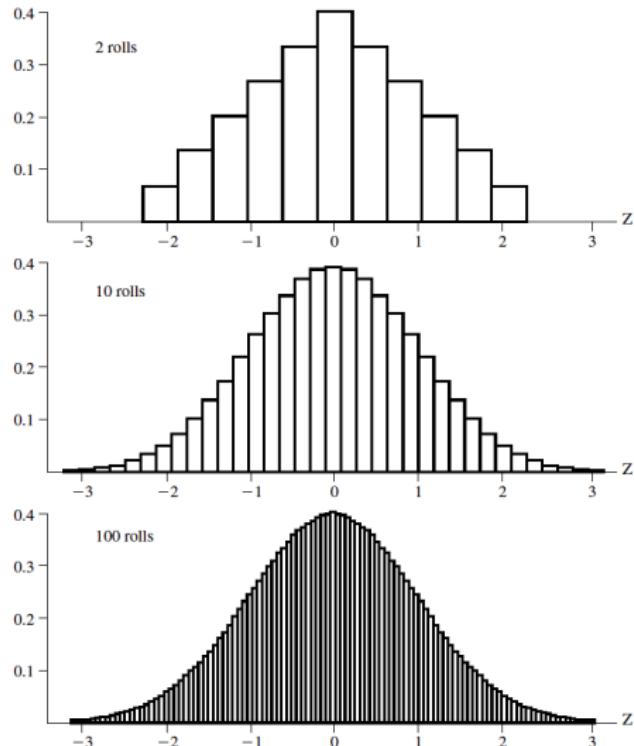
- ▶ No matter what are the expected value and variance of X , it always holds that

$$E[Z] = 0 \quad \text{and} \quad \text{Var}[Z] = \sigma_Z^2 = 1$$

- ▶ Standard normal distribution:

$$X \sim N(\mu, \sigma^2) \quad \rightarrow \quad Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

PDF of sums of six-sided dice rolls (standardized)

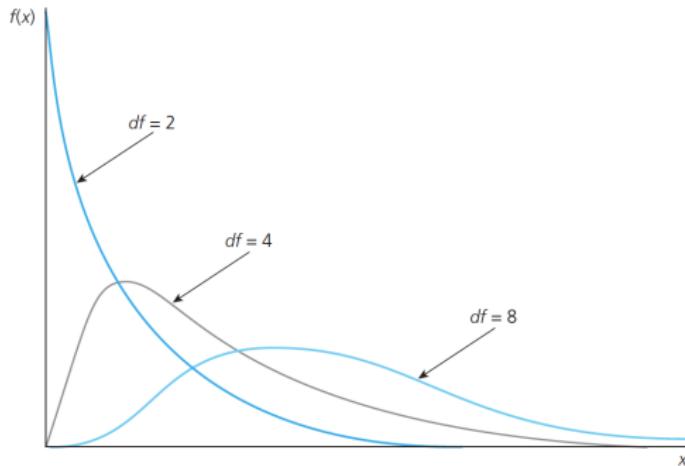


Source: Studenmund (2014, pg. 516)

Chi-squared distribution

- ▶ **Chi-squared distribution** with m degrees of freedom: χ_m^2
- ▶ Let $Z_i \sim N(0, 1)$ for each i and independent, then

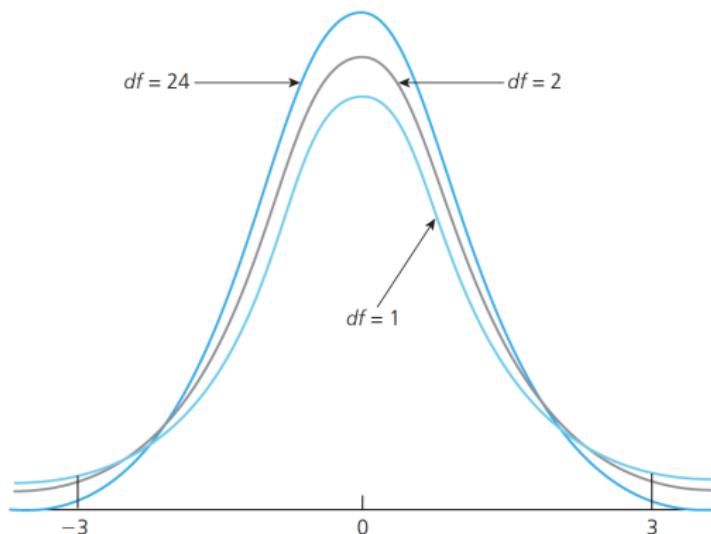
$$X = \sum_{i=1}^m Z_i^2 \quad \longrightarrow \quad X \sim \chi_m^2$$



Source: Wooldridge (2016, pg. 670)

t distribution

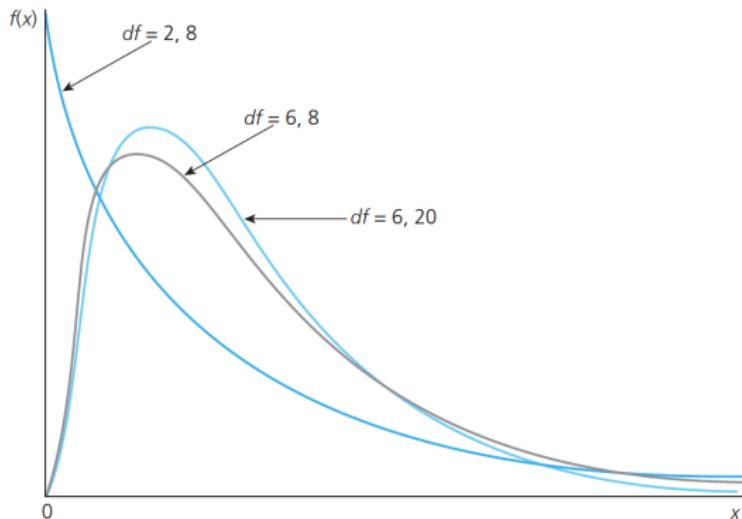
- ▶ (Student's) **t distribution** with m degrees of freedom: t_m
- ▶ Let $Z \sim N(0, 1)$, $X \sim \chi^2_m$, independent: $T = \frac{Z}{\sqrt{X/m}} \sim t_m$
- ▶ Note also that as m grows, t_m distribution approaches $N(0, 1)$



Source: Wooldridge (2016, pg. 671)

F distribution

- (Fisher-Snedecor) **F distribution** with m and o degrees of freedom: $F_{m,o}$
- Let $X \sim \chi^2_m$, $Y \sim \chi^2_o$, independent: $F = \frac{X/m}{Y/o} \sim F_{m,o}$
- Why all important: hypotheses testing, confidence intervals



Source: Wooldridge (2016, pg. 671)

Some terminology

- ▶ **Population:** the entire group of items of our research interest
- ▶ **Sample:** a part of the population that we actually observe
- ▶ **Statistical inference:** use of a sample to draw conclusion about the characteristics of the population from which the sample came
- ▶ Examples: opinion polls, medical experiments

Random sampling

- ▶ Statistical inference can be performed correctly only on a **random sample**, i.e. a sample that reflects the true distribution of the population
- ▶ Each member of the population is equally likely to be included in a random sample
- ▶ Each observation in a random sample is an **independent** random variable drawn from the same population
- ▶ **Biased sample:** any sample that differs systematically from the population that it is intended to represent

Selection biases

- ▶ **Selection bias:** occurs when the selection of the sample systematically excludes or under represents certain groups
 - ▶ Example: opinion poll about tuition payments among undergraduate students vs all citizens
- ▶ **Self-selection bias:** occurs when we examine data for a group of people who have chosen to be in that group
 - ▶ Example: accident statistics of people who buy collision insurance
- ▶ **Survivor bias:** occurs when we choose a sample from a current population (survivors) in order to draw inferences about a past population
 - ▶ Example: S&P500 companies, medical records of old people
- ▶ **Nonresponse bias:** the systematic refusal of some groups to participate in an experiment or to respond to a poll

Some more terminology

- ▶ **Parameter:** a true characteristic of the distribution of a variable, whose value is unknown, but can be estimated
 - ▶ Example: population mean $E[X]$
- ▶ **Estimator:** a sample statistic that is used to estimate the value of the parameter
 - ▶ Example: sample mean \bar{X}_n
 - ▶ Note that the estimator is a random variable (it has a probability distribution, mean, variance,...)
- ▶ **Estimate:** the specific value of the estimator that is obtained using an estimation technique and a particular sample

Linearity and linear combination

- ▶ In mathematics, a linear function $f(x)$ is a function that satisfies the following two properties:
 - ▶ Additivity: $f(x + y) = f(x) + f(y)$
 - ▶ Homogeneity of degree 1: $f(\alpha x) = \alpha f(x) \quad \forall \alpha$
- ▶ **Linear combination** is an expression constructed from a set of terms by multiplying each term by a constant and adding the results
- ▶ E.g. a linear combination of x and y would be any expression of the form $ax + by$, where a and b are constants

Source: Wikipedia [here](#) (linearity) and [here](#) [accessed 2018-09-18].

Properties of an estimator

- ▶ An estimator is **unbiased** if the mean of its distribution is equal to the true value of the parameter it is estimating
- ▶ An estimator is **consistent** if it converges to the true value of the parameter as the sample size increases
- ▶ An estimator is **efficient** if the variance of its sampling distribution is the smallest possible

Properties of an estimator: Example

- ▶ Let X_i be observations sampled from a distribution with mean μ and variance σ^2
- ▶ Let us consider the sample mean $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ as an estimator of μ
- ▶ It can be shown that:
 1. $E[\bar{X}_n] = \mu$
 2. $\bar{X}_n \rightarrow \mu$ as n increases
 3. \bar{X}_n has the smallest variance of all possible estimators of μ
- ▶ Hence, the sample mean \bar{X}_n is an unbiased, consistent, and efficient estimator of μ

Summary

- ▶ Today, we have revised some important concepts from statistics that we will use throughout our econometrics classes
- ▶ It was a very brief and quick overview, serving only for information what students are expected to know already
- ▶ The focus was on distributions and their moments, on sampling and estimation terminology

Seminars and the next lecture #2

- ▶ In the upcoming **seminars**, we will practice some of the concepts mentioned today:
 - ▶ basic statistical concepts (mean, median, variance)
 - ▶ work with the standard normal distribution (statistical tables)
 - ▶ properties of the sample mean
- ▶ In the next **lecture**, we will start with regression analysis and introduce the Ordinary Least Squares (OLS) estimator
- ▶ Readings for lecture #2:
 - ▶ Studenmund (2016 & 17, [2014]): Chapters 1.2–1.5, 2.1–2.3, 3.1–3.2 [3], 7.1
 - ▶ Wooldridge (2016, 2012): Chapters 1, 2-1–2-2, 2-3a, 2-4c

LECTURE #2

Introductory Econometrics

INTRODUCTION TO THE LINEAR REGRESSION MODEL

Jiri Kukacka, Ph.D.

Institute of Economic Studies, Faculty of Social Sciences, Charles University

Winter semester 2021, October 6

In the previous lecture #1

- ▶ We revised basic statistical concepts:
 - ▶ random variable, PDF, CDF, independence...
- ▶ We recalled basic characteristics of RVs:
 - ▶ mean, variance, correlation...
- ▶ We briefly recapitulated important distributions:
 - ▶ normal/Gaussian, standard normal, Chi-squared, t , F
- ▶ We discussed sampling and estimation terminology:
 - ▶ random sampling and selection biases
 - ▶ population → sample → statistical inference
 - ▶ parameter → estimator → estimate
 - ▶ properties of an estimator: unbiasedness, consistency, efficiency

In today's lecture #2, we will...

- ▶ Introduce an 'econometric model'
- ▶ Discuss various types of data
- ▶ Summarize important steps of an econometric analysis
- ▶ Explain the OLS estimation method of the linear regression model
- ▶ Derive the OLS estimator
- ▶ Learn how to interpret the estimated coefficients
- ▶ Readings for lecture #2:
 - ▶ Studenmund (2016 & 17, [2014]): Chapters 1.2–1.5, 2.1–2.3, 3.1–3.2 [3], 7.1
 - ▶ Wooldridge (2016, 2012): Chapters 1, 2-1–2-2, 2-3a, 2-4c

Econometric models

- ▶ Econometric model is an estimable formulation of a theoretical relationship
- ▶ Theory says:
$$Q = f(P, P_s, Yd)$$
 - ▶ Q ... quantity demanded
 - ▶ P ... commodity's price
 - ▶ P_s ... price of substitute good
 - ▶ Yd ... disposable income
- ▶ We simplify:
$$Q = \beta_0 + \beta_1 P + \beta_2 P_s + \beta_3 Yd$$
- ▶ We estimate:

$$Q = 27.7 - 0.11P + 0.03P_s + 0.23Yd$$

Econometric models

- ▶ Today's econometrics deals also with different, even very general models
- ▶ During the course we will cover just linear regression models
- ▶ We will see how these models are estimated by:
 - ▶ Ordinary Least Squares (OLS)
 - ▶ Generalized Least Squares (GLS)
 - ▶ Two-Stage Least Squares (2SLS)
- ▶ We will perform estimation on different types of data

Data used in econometrics

Cross-section

sample of units

(e.g. firms, individuals)

taken at a given point in time

Pooled/repeated cross-section

several independent

samples of units

(e.g. firms, individuals)

taken at different points in time

Time-series

observations of a variable(s)

in different points in time

Panel data

time series for each

cross-sectional unit

in the data set

Data used in econometrics: Cross-section

TABLE 1.1 A Cross-Sectional Data Set on Wages and Other Individual Characteristics

obsno	wage	educ	exper	female	married
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
.
.
.
525	11.56	16	5	0	1
526	3.50	14	5	1	0

Source: Wooldridge (2016, pg. 6)

Data used in econometrics: Pooled/repeated cross-section

TABLE 1.4 Pooled Cross Sections: Two Years of Housing Prices						
obsno	year	hprice	proptax	sqrf	bdrms	bthrms
1	1993	85,500	42	1600	3	2.0
2	1993	67,300	36	1440	3	2.5
3	1993	134,000	38	2000	4	2.5
.
.
.
250	1993	243,600	41	2600	4	3.0
251	1995	65,000	16	1250	2	1.0
252	1995	182,400	20	2200	4	2.0
253	1995	97,500	15	1540	3	2.0
.
.
.
520	1995	57,200	16	1100	2	1.5

Source: Wooldridge (2016, pg. 8)

Data used in econometrics: Time-series

TABLE 1.3 Minimum Wage, Unemployment, and Related Data for Puerto Rico

obsno	year	avgmin	avgcov	prunemp	prgnp
1	1950	0.20	20.1	15.4	878.7
2	1951	0.21	20.7	16.0	925.0
3	1952	0.23	22.6	14.8	1015.9
.
.
.
37	1986	3.35	58.1	18.9	4281.6
38	1987	3.35	58.2	16.8	4496.7

Source: Wooldridge (2016, pg. 7)

Data used in econometrics: Panel data

TABLE 1.5 A Two-Year Panel Data Set on City Crime Statistics

obsno	city	year	murders	population	unem	police
1	1	1986	5	350,000	8.7	440
2	1	1990	8	359,200	7.2	471
3	2	1986	2	64,300	5.4	75
4	2	1990	1	65,100	5.5	75
.
.
.
297	149	1986	10	260,700	9.6	286
298	149	1990	6	245,000	9.8	334
299	150	1986	25	543,000	4.3	520
300	150	1990	32	546,200	5.2	493

Source: Wooldridge (2016, pg. 9)

Steps of an econometric analysis

1. Formulation of an economic model (based on a review of the literature, rigorous or intuitive)
2. Formulation of an econometric model based on the economic model & hypotheses
3. Collection of data ⇒ dataset
4. Estimation of the econometric model
5. Evaluation of 1.–4.
6. Interpretation and documentation of results
7. Predictions (cross-sections) or forecasting (time-series data)

Example: Economic model

- Denote:

- p ... price of an ordinary good
- c ... firm's average cost per one unit of output
- $q(p)$... demand for firm's output

Firm's profit:

$$\pi = q(p) \cdot (p - c)$$

Demand for the good:

$$q(p) = a - b \cdot p$$

- Derive:

$$q = \frac{a}{2} - \frac{b}{2} \cdot c$$

- We call q dependent (or explained) variable and c independent (or explanatory) variable

Example: Econometric model

- ▶ Write the relationship in a simple linear form:

$$q = \beta_0 + \beta_1 c$$

(have in mind that $\beta_0 = \frac{a}{2}$ and $\beta_1 = -\frac{b}{2}$)

- ▶ There are other (unpredictable) effects that influence firms' sales \Rightarrow add the error/disturbance term:

$$q = \beta_0 + \beta_1 c + \varepsilon$$

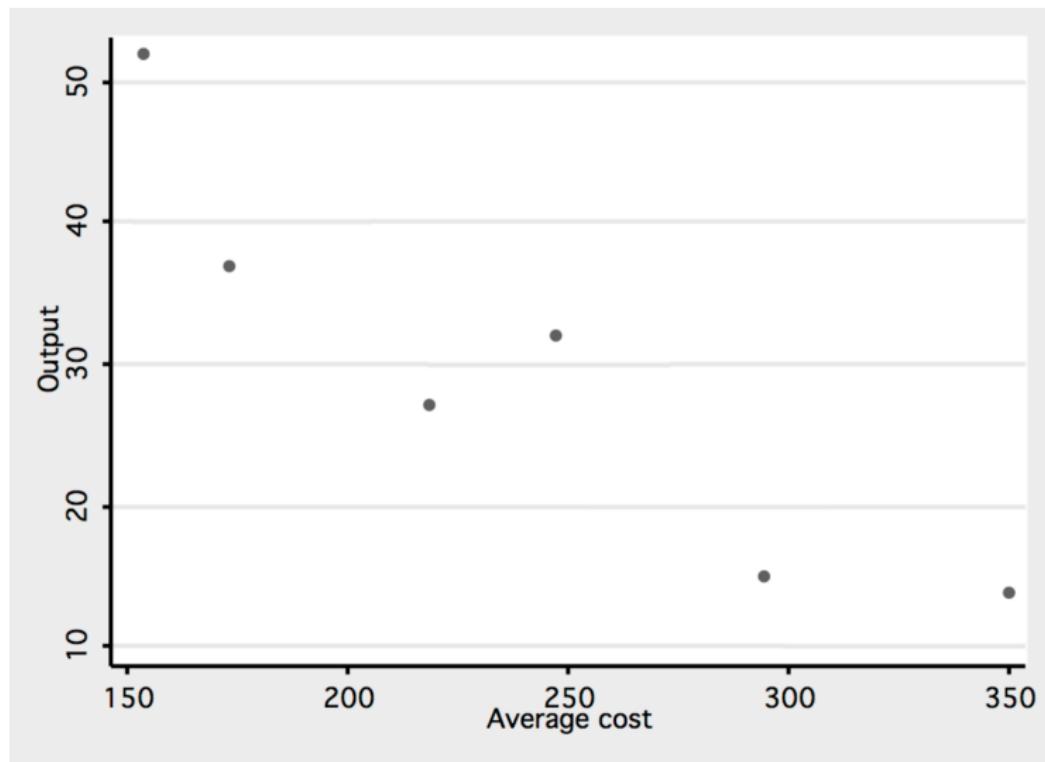
- ▶ Find the value of parameters β_1 (slope) and β_0 (intercept)
- ▶ Hypothesis about the expected sign: $\beta_1 < 0$

Example: Data

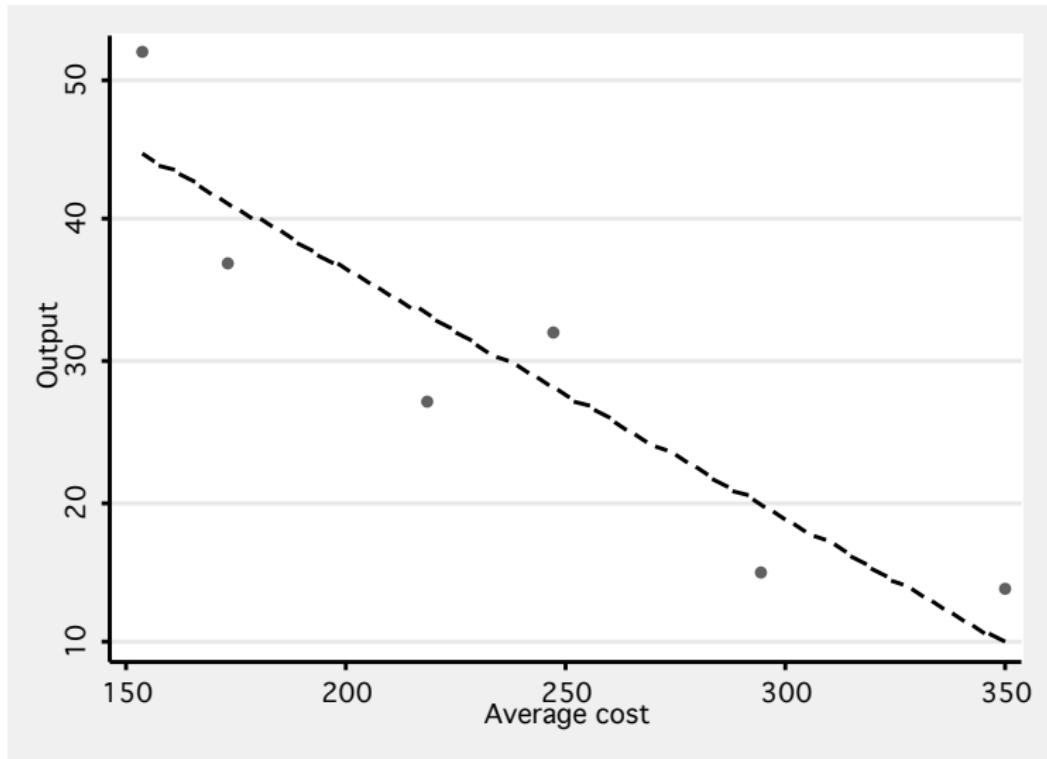
- ▶ Ideally: investigate all firms in the economy
- ▶ Reality: investigate a sample of firms
 - ▶ we need a random sample (w/o a bias) of firms (lecture #1)
- ▶ Collect data:

Firm	1	2	3	4	5	6
q	15	32	52	14	37	27
c	294	247	153	350	173	218

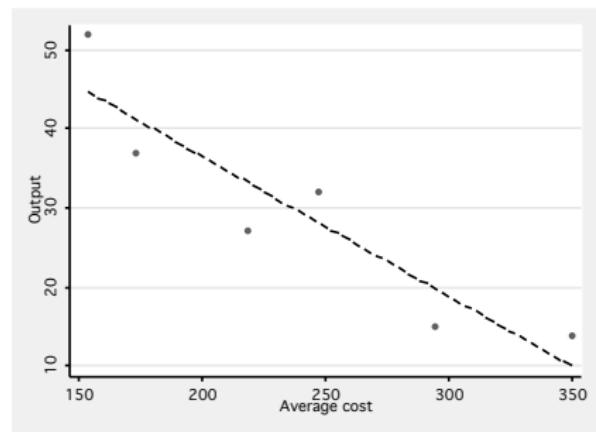
Example: Data



Example: Estimation



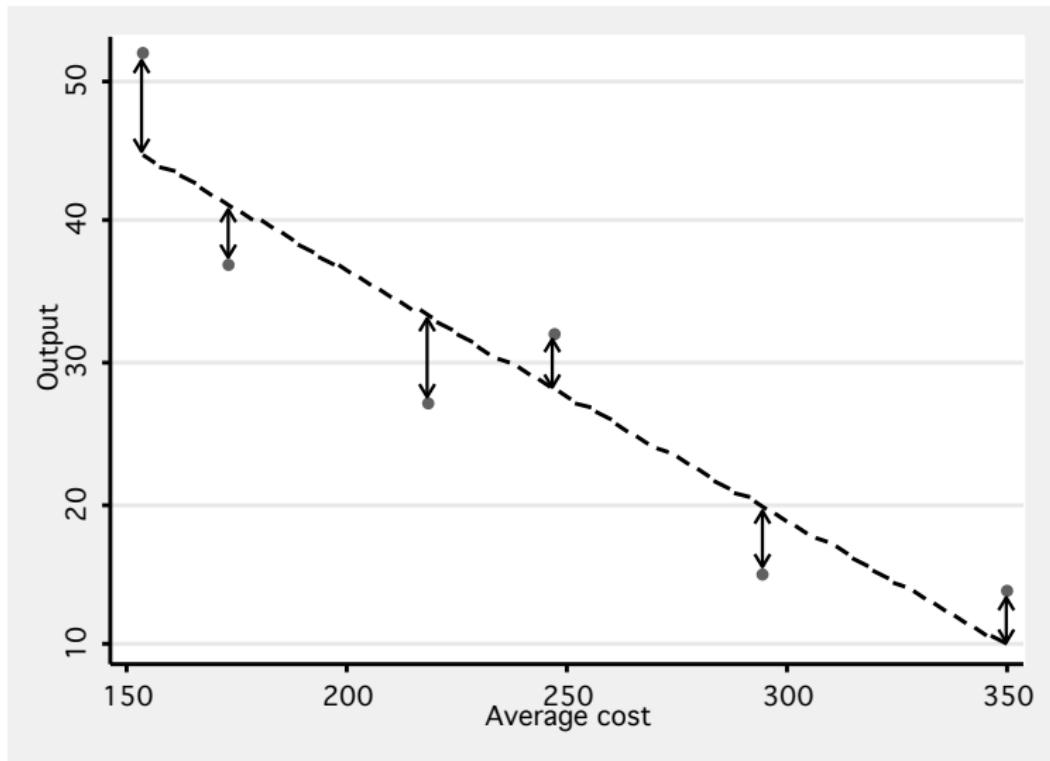
Example: Estimation



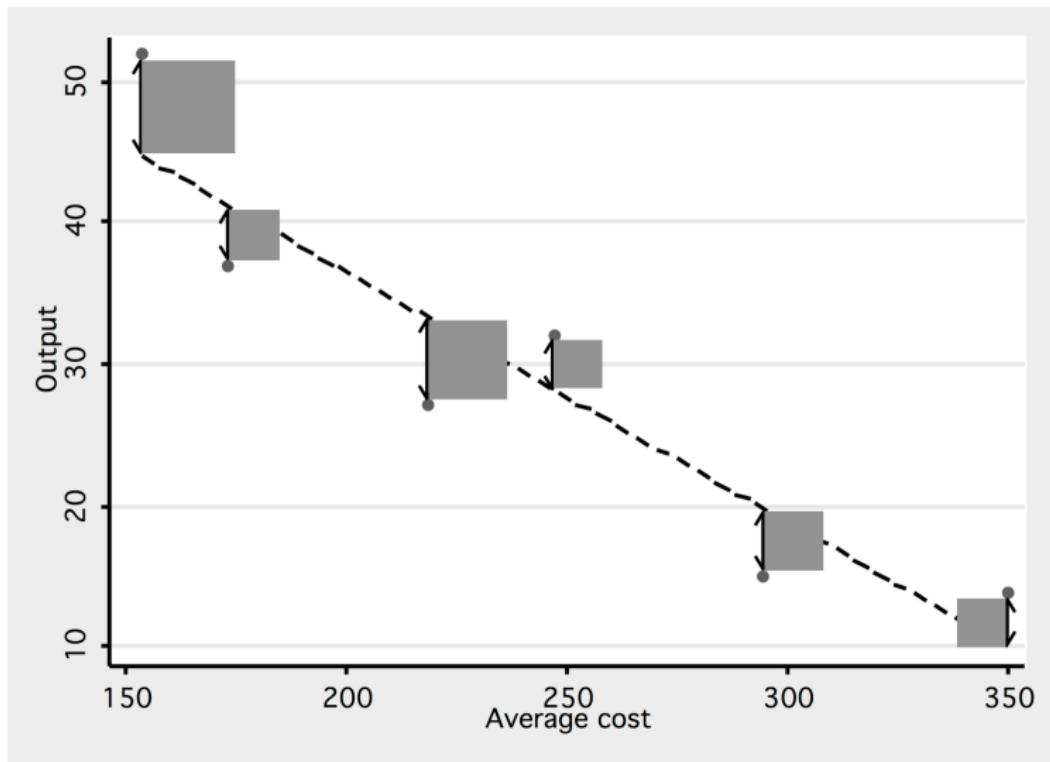
OLS method:

Make the fit as good as possible
↓
Make the misfit as low as possible
↓
Minimize the (vertical) distance
between data points and the
regression line
↓
Minimize the sum of squared
deviations

Example: Estimation



Example: Estimation



Terminology

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$... regression model (line)

y_i ... dependent/explained variable (i -th observation)

x_i ... independent/explanatory variable (i -th observation)

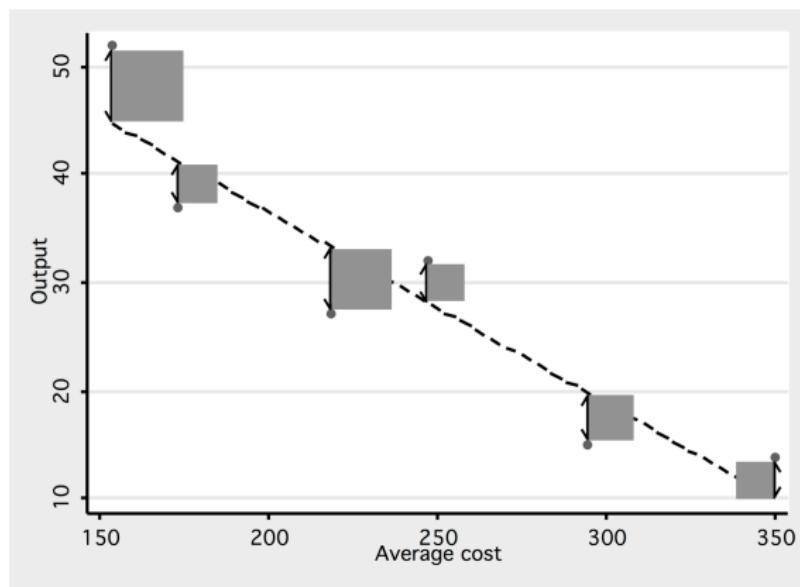
ε_i ... random error/disturbance term (of i -th observation)

β_0 ... intercept parameter ($\hat{\beta}_0$... estimate of this parameter)

β_1 ... slope parameter ($\hat{\beta}_1$... estimate of this parameter)

Ordinary Least Squares

- ▶ OLS = fitting the regression line by minimizing the sum of squared vertical distances between the observed points and the regression line



Ordinary Least Squares: Principle

- ▶ Take the squared differences between observed point y_i and the regression line $\beta_0 + \beta_1 x_i$:

$$(y_i - [\beta_0 + \beta_1 x_i])^2$$

- ▶ Sum them over all n observations:

$$\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

- ▶ Find $\hat{\beta}_0$ and $\hat{\beta}_1$ such that they minimize this sum:

$$[\hat{\beta}_0, \hat{\beta}_1] = \operatorname{argmin}_{\beta_0, \beta_1} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

Ordinary Least Squares: Derivation

$$\left[\hat{\beta}_0, \hat{\beta}_1 \right] = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

► FOC:

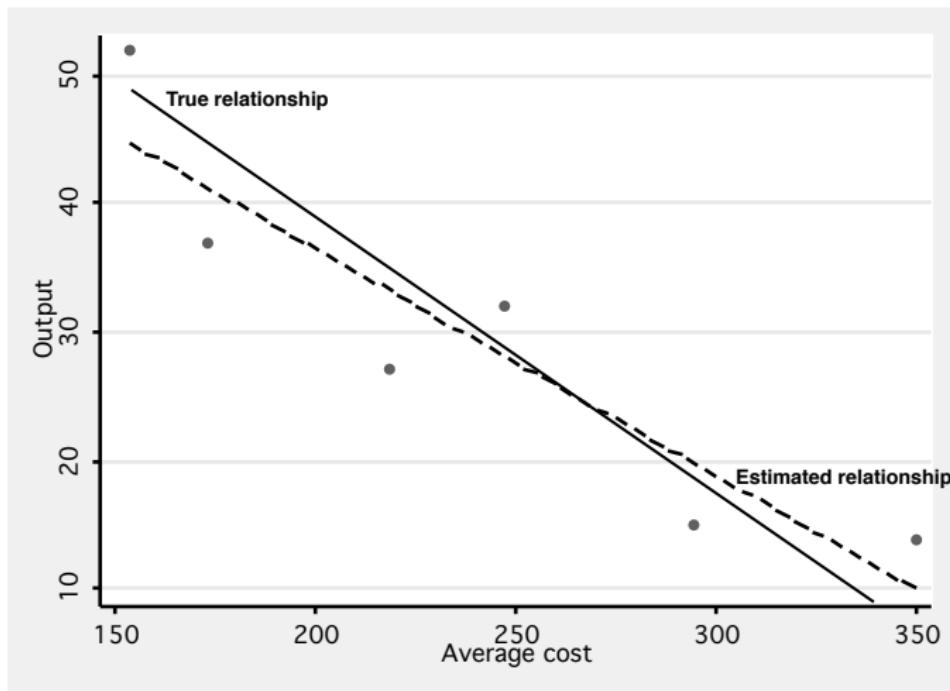
$$\begin{aligned}\frac{\partial}{\partial \beta_0} & : & -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) &= 0 \\ \frac{\partial}{\partial \beta_1} & : & -2 \sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) &= 0\end{aligned}$$

► We express ($\hat{\beta}_0$ in the lecture, $\hat{\beta}_1$ in seminars):

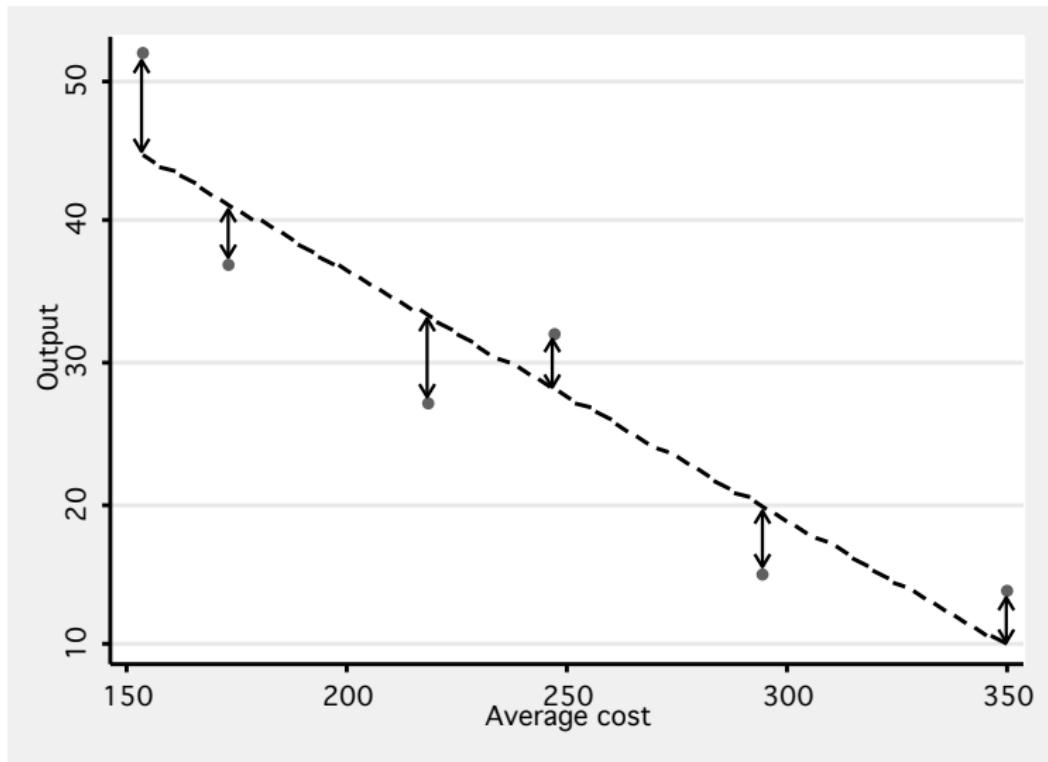
$$\hat{\beta}_0 = \bar{y}_n - \hat{\beta}_1 \bar{x}_n \quad \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} = \frac{Cov(x, y)}{Var(x)}$$

Example

- Estimated regression line: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$



Residual



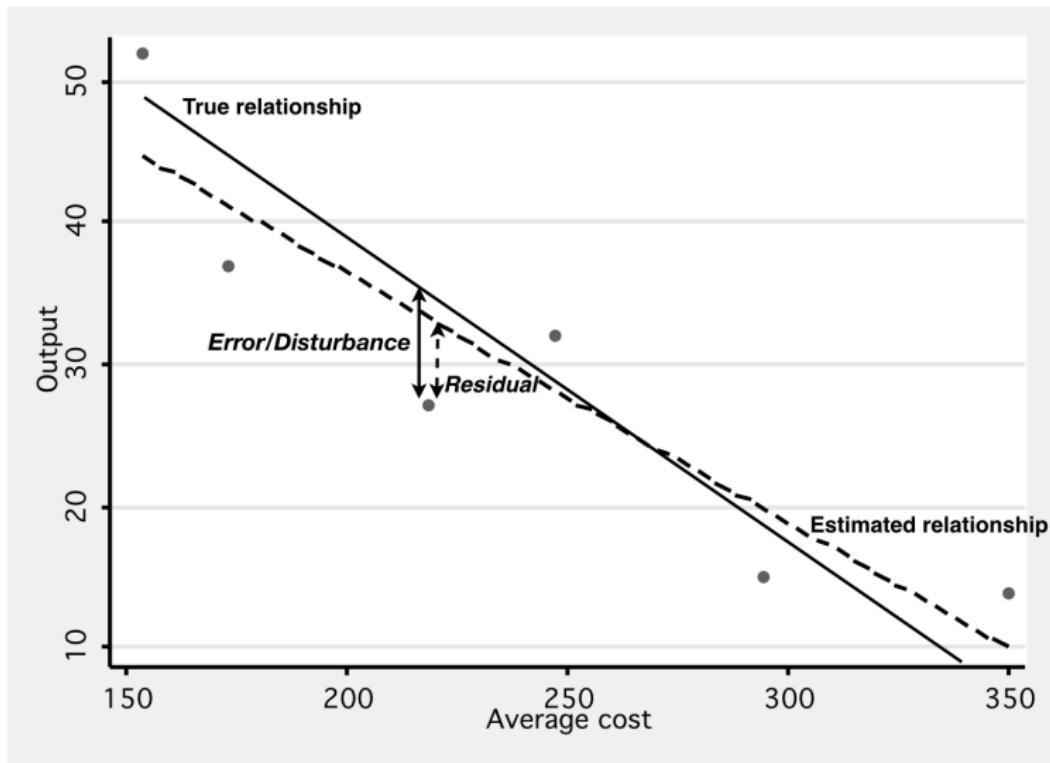
Residual

- ▶ Residual is the vertical distance between the observation points and the estimated regression line
- ▶ It is the difference between the true value y_i and the estimated/fitted value:

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

- ▶ OLS thus minimize the sum of squares of all residuals
- ▶ Residual e_i (observed) is **not the same** as the error ε_i (unobserved)
- ▶ Residual is an estimate of the error: $\hat{\varepsilon}_i = e_i$

Example



Getting back to the example

- We have the economic model:

$$q = \frac{a}{2} - \frac{b}{2} \cdot c$$

- We estimate:

$$q_i = \beta_0 + \beta_1 c_i + \varepsilon_i$$

(having in mind that $\beta_0 = \frac{a}{2}$ and $\beta_1 = -\frac{b}{2}$)

- Over data:

Firm	1	2	3	4	5	6
q	15	32	52	14	37	27
c	294	247	153	350	173	218

Getting back to the example

- When we plug in the formula:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^6 (c_i - \bar{c})(q_i - \bar{q})}{\sum_{i=1}^6 (c_i - \bar{c})^2} = -0.177$$
$$\hat{\beta}_0 = \bar{q} - \hat{\beta}_1 \bar{c} = 71.74$$

- The estimated equation is:

$$\hat{q} = 71.74 - 0.177c$$

and so:

$$\hat{a} = 2\hat{\beta}_0 = 143.48 \quad \text{and} \quad \hat{b} = -2\hat{\beta}_1 = 0.354$$

Example: Evaluation & interpretation

1. Economic model
2. Econometric model & the hypothesis: $\beta_1 < 0$
3. Dataset
4. Estimation $\Rightarrow \hat{q} = 71.74 - 0.177c$
5. Evaluation
6. Interpretation: when the average cost increases by 1 unit, quantity demanded decreases by 0.177 units

Several explanatory variables

- ▶ One explanatory variable \Rightarrow a **simple** linear regression model
- ▶ Usually, there are more than one explanatory variables in regression models
- ▶ **Multivariate/multiple** regression model with k explanatory variables:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i$$

- ▶ For observations $1, 2, \dots, n$, we have:

$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \dots + \beta_k x_{1k} + \varepsilon_1$$

$$y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \dots + \beta_k x_{2k} + \varepsilon_2$$

$$\vdots \quad \vdots$$

$$y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \dots + \beta_k x_{nk} + \varepsilon_n$$

Matrix notation

- We can write in a matrix form:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

or in a simplified notation:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Recap slides 22, 23: Ordinary Least Squares: Derivation

- ▶ Find $\hat{\beta}_0$ and $\hat{\beta}_1$ such that they minimize the sum:

$$[\hat{\beta}_0, \hat{\beta}_1] = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

- ▶ FOC: ...
- ▶ We express $\hat{\beta}_0$, $\hat{\beta}_1$

Simplified OLS derivation under ‘matrix notation’

- We need to minimize $\varepsilon^2 = \varepsilon' \varepsilon$ again:

$$\begin{aligned}\hat{\beta} &= \underset{\beta}{\operatorname{argmin}} \varepsilon^2 = \underset{\beta}{\operatorname{argmin}} (\mathbf{y} - \mathbf{X}\beta)^2 \\ &= \underset{\beta}{\operatorname{argmin}} \mathbf{y}^2 - 2\beta\mathbf{X}'\mathbf{y} + \mathbf{X}^2\beta^2\end{aligned}$$

- FOC:

$$\begin{aligned}\frac{\partial}{\partial \beta} : \quad -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}^2\hat{\beta} &= 0 \\ \mathbf{X}'\mathbf{X}\hat{\beta} &= \mathbf{X}'\mathbf{y}\end{aligned}$$

- This gives us:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$$

► Complete OLS derivation under matrix notation

Linearity of the model

- ▶ We talk about linear regression model—what do we mean by linearity?
- ▶ Regression models are **linear in parameters**, but they do not need to be linear in variables:
 - ▶ $y = \beta_0 + \beta_1 x + \varepsilon \dots$ is a linear model
 - ▶ $\ln y = \beta_0 + \beta_1 \ln x + \beta_2 \sqrt{z} + \varepsilon \dots$ is a linear model
 - ▶ $y = \beta_0 + x^{\beta_1} + \varepsilon \dots$ is **NOT** a linear model

Meaning of regression parameters

- ▶ Consider the multivariate regression model:

$$Q = \beta_0 + \beta_1 P + \beta_2 P_s + \beta_3 Yd + \varepsilon$$

estimated as: $\hat{Q} = 27.7 - 0.11P + 0.03P_s + 0.23Yd$

Q ... quantity demanded

P ... commodity's price

P_s ... price of substitute

Yd ... disposable income

- ▶ Meaning of β_1 is the impact of a one unit increase in P on the dependent variable Q , holding the other included independent variables P_s and Yd constant (*ceteris paribus*)
- ▶ When price P increases by 1 unit (and P_s and Yd remain the same), quantity demanded Q decreases by 0.11 units

Interpretation of the estimated intercept?

- ▶ We almost always include the intercept β_0 to the regression model but mostly, we do not interpret $\hat{\beta}_0$
- ▶ $\hat{\beta}_0$ has at least three components:
 - ▶ the true β_0
 - ▶ the constant impact of any specification errors
 - ▶ the mean of ε (if $\neq 0$) \Rightarrow theoretical importance (lecture #3)
 - ▶ unfortunately, we only observe their sum
- ▶ Next, the origin often lies outside the range of sample observations (a mathematical extreme)
- ▶ In specific econometric models, $\hat{\beta}_0$ might have a good economic meaning
- ▶ In rare(!) cases, we can suppress β_0 based on economic theory

Summary

- ▶ We have learned that an econometric analysis consists of:
 1. definition of the model
 2. estimation
 3. interpretation
- ▶ We have explained the principle of OLS: minimizing the sum of squared differences between the observations and the regression line
- ▶ We have derived the formula for the OLS estimator:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Seminars and the next lecture #3

- ▶ In the upcoming **seminars**, we will:
 - ▶ derive the OLS formula for $\hat{\beta}_1$
 - ▶ estimate simple regression models using summation and matrix formulas in Excel (BYOD?)
 - ▶ evaluate and interpret these models
 - ▶ demonstrate the difference between the residual and the error term
- ▶ In the **next lecture**, we will:
 - ▶ learn the Classical Assumptions of regression models
 - ▶ study properties of the OLS estimator
- ▶ Readings for lecture #3:
 - ▶ Studenmund (2016 & 17, 2014): Chapter 4
 - ▶ Wooldridge (2016, [2012]): Chapters 2-3b (pg. 32–33 [36]), 2-5-2-6, 3-1-3-2a-i\h, (3-3-3-4), 3-5-3-6, 4-1, (5)

Appendix: OLS derivation under matrix notation

- We need to minimize $\varepsilon^2 = \varepsilon' \varepsilon$ again:

$$\begin{aligned}\hat{\beta} &= \underset{\beta}{\operatorname{argmin}} (\mathbf{y} - \mathbf{X}\beta)' (\mathbf{y} - \mathbf{X}\beta) \\ &= \underset{\beta}{\operatorname{argmin}} \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\beta - \beta'\mathbf{X}'\mathbf{y} + \beta'\mathbf{X}'\mathbf{X}\beta \\ &= \underset{\beta}{\operatorname{argmin}} \mathbf{y}'\mathbf{y} - 2\beta'\mathbf{X}'\mathbf{y} + \beta'\mathbf{X}'\mathbf{X}\beta\end{aligned}$$

- FOC:

$$\frac{\partial \varepsilon' \varepsilon}{\partial \beta} : \quad -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\beta} = 0$$
$$\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y}$$

- This gives us:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$$

▶ Back

Appendix: Selected properties of matrix differentiation

$$\begin{aligned}\frac{\partial \mathbf{b}' \mathbf{a}}{\partial \mathbf{b}} &= \frac{\partial \mathbf{a}' \mathbf{b}}{\partial \mathbf{b}} = \mathbf{a} \\ \frac{\partial \mathbf{b}' \mathbf{D} \mathbf{b}}{\partial \mathbf{b}} &= 2\mathbf{D}\mathbf{b}\end{aligned}$$

where \mathbf{a}, \mathbf{b} are $k \times 1$ vectors and \mathbf{D} is a symmetric matrix

LECTURE #3

Introductory Econometrics

LINEAR REGRESSION MODEL ASSUMPTIONS & OLS PROPERTIES

Jiri Kukacka, Ph.D.

Institute of Economic Studies, Faculty of Social Sciences, Charles University

Winter semester 2021, October 13

In the previous lecture #2

- ▶ We discussed various dataset structures and important steps of an econometric analysis
- ▶ We explained the principle of the OLS estimator: minimizing the sum of squared differences between the observation and the regression line $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- ▶ We found the formulae for the OLS estimator:

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

- ▶ We also expressed the general (multivariate) model in matrix notation $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and we derived the formula:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$$

In today's lecture #3, we will...

- ▶ List the assumptions about the error term and the explanatory variables that are required in classical regression models
- ▶ Derive the properties of the OLS estimator for the case when Classical Assumptions hold
- ▶ Show that under these assumption, OLS is the best estimator available for regression models
- ▶ Readings for lecture #3:
 - ▶ Studenmund (2016 & 17, 2014): Chapter 4
 - ▶ Wooldridge (2016, [2012]): Chapters 2-3b (pg. 32–33 [36]), 2-5-2-6, 3-1-3-2a-i\h, (3-3-3-4), 3-5-3-6, 4-1, (5)

Home assignment #1

- ▶ Assigned today via SIS
- ▶ Teams of two, one report
- ▶ Submit electronically in the .pdf format [5 MB max, .xls(x) can be attached in .zip] via the **Study group roster** app in SIS
- ▶ Deadline: Thursday, October 21, 2021, 23:59:59

HA#1 submission

SIS Student Information System (core version: 1636)

Faculty of Social Sciences

59:52 Text mode News

There is a new subject available for erasmus students in winter semester called "Introduction to Journalistic Ethics" (JJB0111) taught by professor Neuzil visiting our faculty via Fulbright Scholar Program.

Education

- Exam dates
- Final Exams
- Subjects and schedule registration
- Subjects
- Study group roster

Time-table

- Schedule NG

Admission process

- Admission

Utils

- Committees
- Invitations for state exams

Noneducational agenda

- Central catalogue
- E-resources Portal
- Discovery system
- Moodle (E-learning)

Others

- Bookmarks
- Life-Long Education programs
- Harmonogram
- Who is Who
- Login searching

- Notice-board
- Personal data
- Study charges and petitions
- Graduation
- List of advisors

HA#1 submission

59/53 My Study Group Roster

Study group roster (version: 229)
Student details

Faculty of Social Sciences

Filter:
Show : actual year only actual and last year

Year	Code	Instruction type	Course title	Time and place	Schedule item	Note
2017/18 winter	JMM615	Lecture	Democracy Promotion: history, theories, practice	Mon 17:00 R201	17aJMM615p1	
2017/18 winter	JMM615	Practicals	Democracy Promotion: history, theories, practice	Mon 17:40 R201	17aJMM615x01	
2017/18 winter	JEM123	Lecture	Economics of Least Developed Countries	Tue 09:30 O314	17aJEM123p1	
2017/18 winter	JEM123	Practicals	Economics of Least Developed Countries	Tue 11:00 O206	17aJEM123x01	
2017/18 winter	JMM143	Lecture	Economy and Politics in the 20th Century Eastern Europe	Thu 12:30 J3015	17aJMM143p1	
2017/18 winter	JMM143	Practicals	Economy and Politics in the 20th Century Eastern Europe	Thu 13:10 J3015	17aJMM143x01	
2017/18 winter	JEM162	Lecture	Energy Markets & Economics	Fri 12:30 O109	17aJEM162p1	
2017/18 winter	JEM162	Practicals	Energy Markets & Economics	Fri 14:00 O109	17aJEM162x01	
2017/18 winter	JEB022	Lecture	Institutional Economics	Wed 09:30 O109	17aJEB022p1	
2017/18 winter	JEB039	Practicals	International Trade	Fri 11:00 O314	17aJEB039x01	
2017/18 winter	JEM082	Lecture	Introductory Econometrics	Wed 11:00 O109	17aJEM082p1	
2017/18 winter	JEM082	Practicals	Introductory Econometrics	Thu 15:30 O016	17aJEM082x01	
2017/18 winter	JEM027	Lecture	Monetary Economics	Mon 17:00 O314	17aJEM027p1	
2017/18 winter	JEM027	Practicals	Monetary Economics	Mon 18:30 O314	17aJEM027x01	

Click on the Course title or the Detail icon for more details.
 Click on the List icon to see the list of all corresponding study groups.
 Click on the Course code for more information about the course.

HA#1 submission

 Study group roster (version: 229)
Student's results

59:04     My Study Group Roster

Group

Course: JEM062 Introductory Econometrics
Year: 2017/18 winter
Instruction type: Lecture
Teacher: PhDr. Mgr. Jiří Kukačka, Ph.D. (jiri.kukacka@fsv.cuni.cz)
Schedule: Wed 11:00 O109
Schedule item: 17aJEM062p1

Results

test HW submission 

Upload file:	<input type="button" value="Choose file"/>	No file chosen	(max. 5000 kB)
File upload deadline:	5.10.2017		
Final version:	<input type="checkbox"/>		

test HW points

test HW comments

test sum 0

 Please check the validity of the uploaded file or seminar paper - by checking the approximate size of the file and
The file name should not exceed 60 characters. If it does exceed 60 characters, please rename the file before up

HA#1 submission

 **Study group roster** (version: 229)
Student's results

59:39     My Study Group Roster

Group

Course: JEM062 Introductory Econometrics
Year: 2017/18 winter
Instruction type: Lecture
Teacher: PhDr. Mgr. Jiří Kukačka, Ph.D. (jiri.kukacka@fsv.cuni.cz)
Schedule: Wed 11:00 O109
Schedule item: 17aJEM062p1

Results

test HW submission 

Uploaded file:	 size: 81kB, inserted: 04.10.2017 12:58:36 
Upload file:	<input type="file"/> Choose file No file chosen (max. 5000 kB)
File upload deadline:	5.10.2017
Final version:	<input checked="" type="checkbox"/> 

test HW points

test HW comments

test sum 0

 Save

 Please check the validity of the uploaded file or seminar paper - by checking the approximate size of the file and
The file name should not exceed 60 characters. If it does exceed 60 characters, please rename the file before up

HA#1 evaluation

 **Study group roster** (version: 229)
Student's results

59:39     My Study Group Roster

Group

Course: JEM062 Introductory Econometrics
Year: 2017/18 winter
Instruction type: Lecture
Teacher: PhDr. Mgr. Jiří Kukačka, Ph.D. (jiri.kukacka@fsv.cuni.cz)
Schedule: Wed 11:00 O109
Schedule item: 17aJEM062p1

Results

test HW submission 	Uploaded file:  size: 81kB, inserted: 04.10.2017 12:58:36 
	Upload file: <input type="file"/> Choose file No file chosen (max. 5000 kB)
	File upload deadline: 5.10.2017
	Final version: <input checked="" type="checkbox"/> 
test HW points	10
test HW comments	well done!
test sum	10

 Save

 Please check the validity of the uploaded file or seminar paper - by checking the approximate size of the file and.
The file name should not exceed 60 characters. If it does exceed 60 characters, please rename the file before up

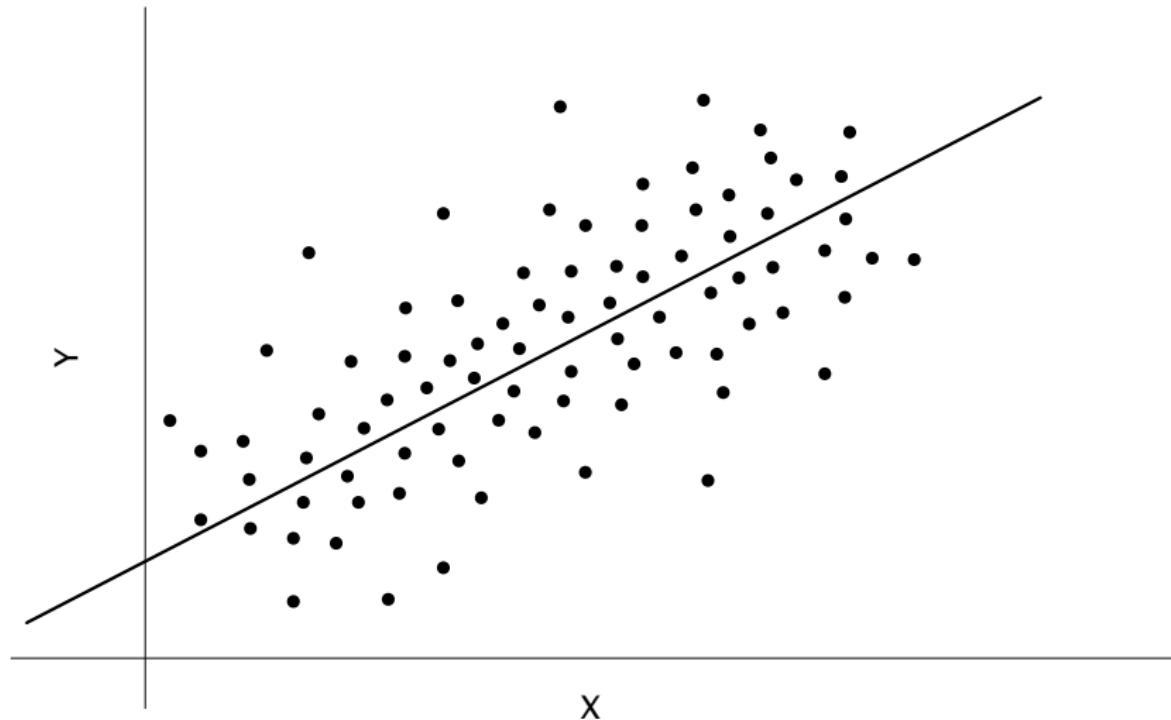
Importance of the error/disturbance term

- ▶ The stochastic error term must be **always** present in a regression equation because of:
 1. omission of many minor influences (unavailable data)
 2. measurement error
 3. possibly incorrect functional form
 4. stochastic character of unpredictable human behavior
- ▶ Understand that all of these factors are included in the error term and may influence its properties
- ▶ **The properties of the error term determine the properties of the OLS estimator**

The Classical Assumptions

1. The regression model is linear in parameters, is correctly specified, and has an additive error term
2. The error term has a zero population mean
3. All explanatory variables are uncorrelated with the error term
4. Realizations of the error term are uncorrelated with each other
5. The error term has a constant variance
6. No explanatory variable is a perfect linear function of any other explanatory variable(s)
7. (The error term is normally distributed)

Graphical representation



1. Linearity in parameters

The regression model is linear in parameters, is correctly specified, and has an additive error term.

- ▶ Was discussed in the previous lecture #2
- ▶ Linearity in variables is not required
- ▶ Example: Cobb-Douglas production function $Y = AL^{\beta_1}K^{\beta_2}$ for which we suppose $A = \exp^{\beta_0+\varepsilon}$ can be transformed so that:

$$\ln Y = \beta_0 + \beta_1 \ln L + \beta_2 \ln K + \varepsilon$$

and the linearity in parameters/coefficients is restored

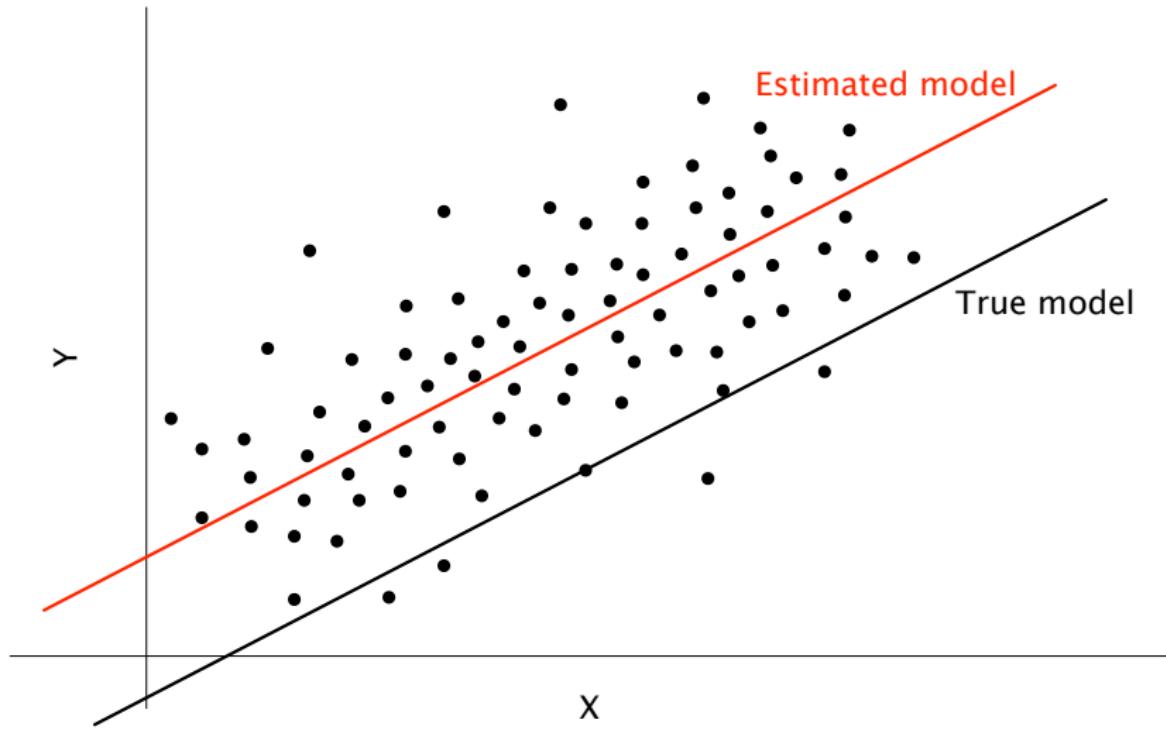
- ▶ Note that it is the linearity in parameters that allows us to rewrite the general regression model in the matrix form

2. Zero mean of the error term

The error term has a zero population mean.

- ▶ Notation: $E[\varepsilon_i] = 0$ or $E[\varepsilon] = \mathbf{0}$
- ▶ Idea: observations are evenly distributed along the true relationship, the average of deviations is zero
- ▶ In fact, the mean of ε_i is forced to be zero by the existence of **the intercept**/constant term (β_0) in the equation
- ▶ Hence, this assumption is satisfied as long as there is an intercept included in the equation

Graphical representation

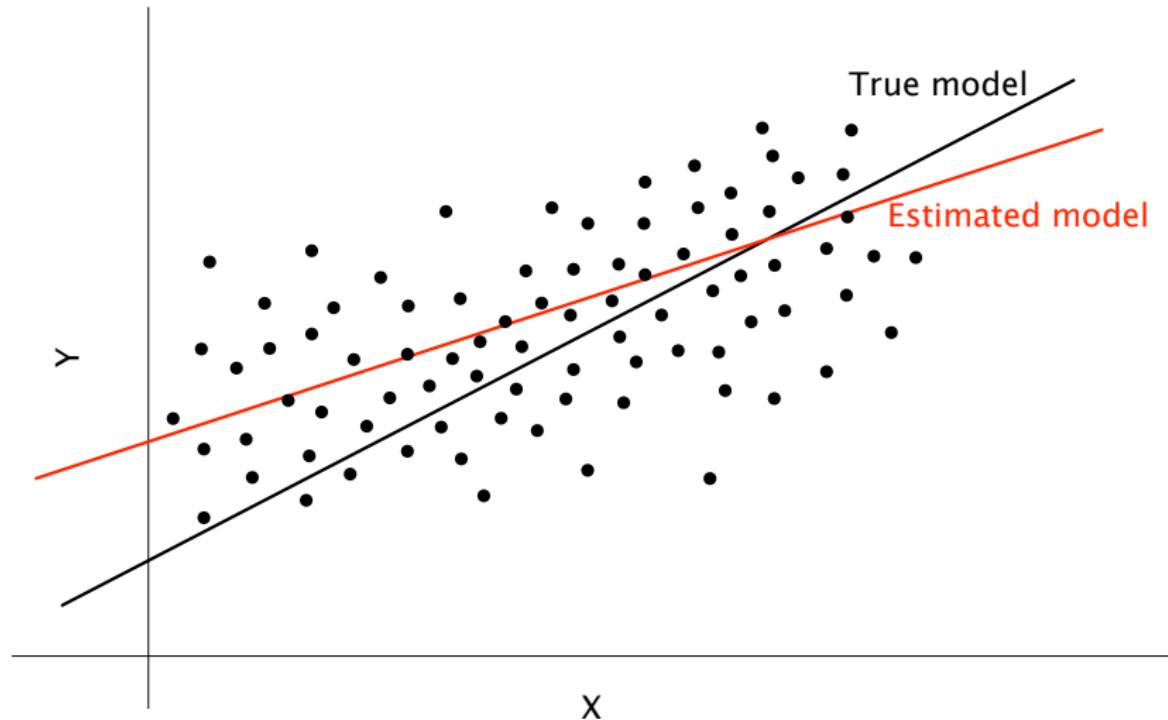


3. Variables uncorrelated with the error term

All explanatory variables are uncorrelated with the error term.

- ▶ Notation: $E[x_i \varepsilon_i] = 0$ or $E[\mathbf{X}' \boldsymbol{\varepsilon}] = \mathbf{0}$
- ▶ If an explanatory variable and the error term were correlated with each other, the OLS estimator would be likely to attribute to the x some of the variation in y that actually came from the error term
- ▶ Example: analysis of household consumption patterns
 - ▶ households with lower incomes may indicate higher consumption (because of shame)
 - ▶ negative correlation between x and ε (measurement error higher for lower incomes)
- ▶ Leads to a **biased and inconsistent** OLS estimator
- ▶ We will solve this problem using the Two-Stage Least Squares (2SLS) approach

Graphical representation

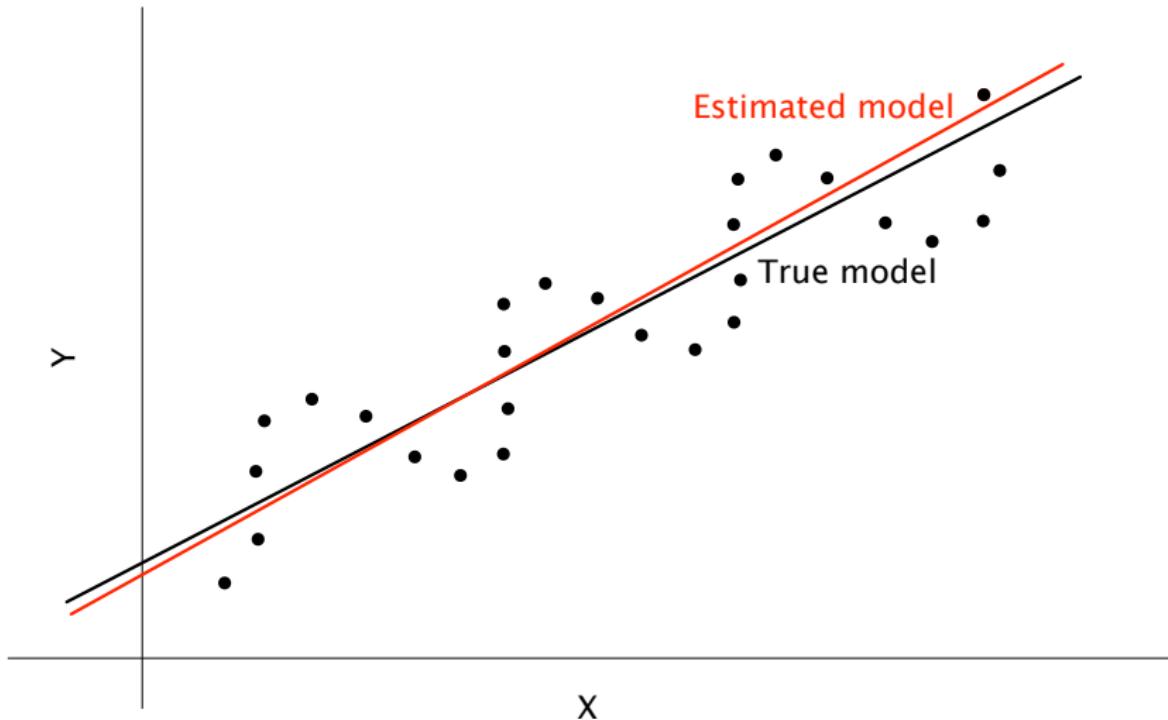


4. Errors uncorrelated with each other

Realizations of the error term are uncorrelated with each other.

- ▶ If there is a systematic correlation between one realization of the error term and another (autocorrelation/serial correlation), it is more difficult for OLS to get precise estimates of the coefficients of the explanatory variables
- ▶ Technically: the OLS estimator remains unbiased and consistent, but **not efficient**
- ▶ Often happens in time series data, where a random shock in one time period affects the random shock in another time period
- ▶ We will solve this problem using the Generalized Least Squares (GLS) estimator

Graphical representation



5. Constant variance of the error term

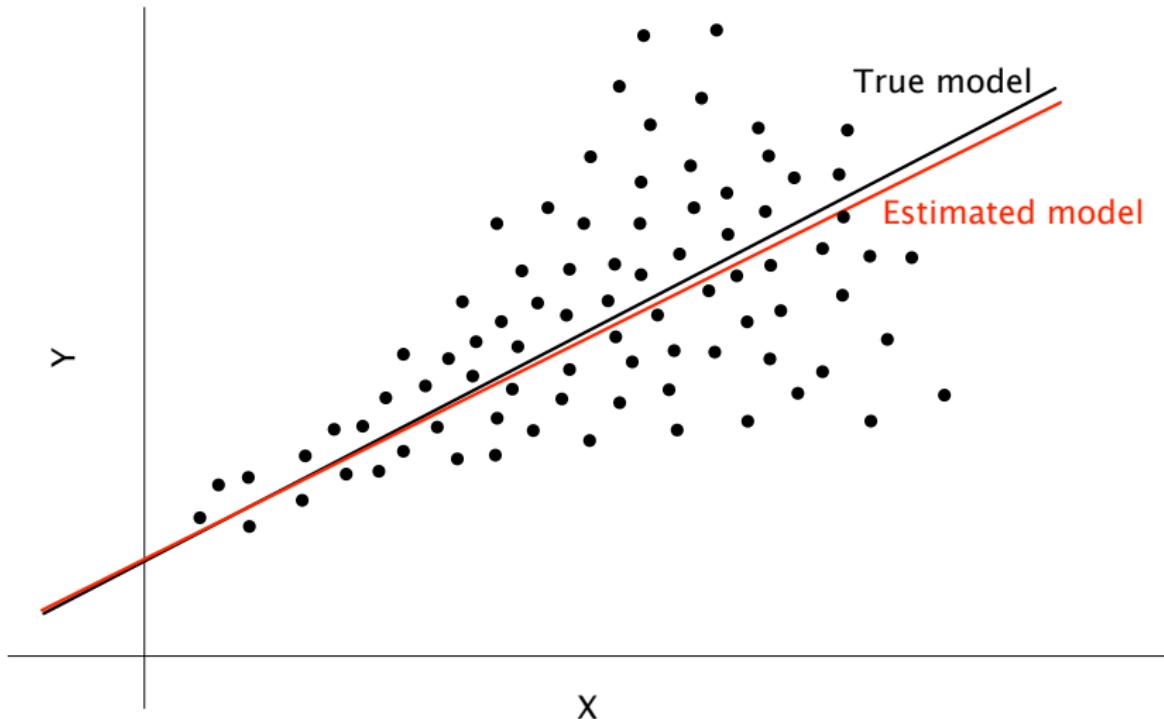
The error term has a constant variance.

- ▶ This property is called *homoskedasticity*
- ▶ If it is not satisfied, we talk about *heteroskedasticity*
- ▶ It states that each realization of the error is drawn from a distribution with the same variance and thus varies in the same manner along the true relationship
- ▶ If the error term is heteroskedastic, it is more difficult for OLS to get precise estimates of the coefficients of the explanatory variables
- ▶ Technically: the OLS estimator remains unbiased and consistent, but **not efficient**

5. Constant variance of the error term

- ▶ Heteroskedasticity is often present in cross-sectional data
- ▶ Example: analysis of household consumption patterns
 - ▶ variance of the consumption of certain goods might be greater for higher-income households
 - ▶ these have more discretionary income than do lower-income households
- ▶ We will solve this problem using the White heteroskedasticity-corrected standard errors

Graphical representation



4. No autocorrelation + 5. Homoskedasticity

- ▶ Notation:

- ▶ no autocorrelation:

$$\text{Corr}(\varepsilon_i, \varepsilon_j) = 0 \Rightarrow E[\varepsilon_i \varepsilon_j] = 0 \text{ for each } i, j; i \neq j$$

- ▶ homoskedasticity:

$$\text{Var}(\varepsilon_i) = \sigma^2 \Rightarrow E[\varepsilon_i^2] = \sigma^2 \text{ for each } i$$

- ▶ Matrix notation:

$$\text{Var}[\varepsilon] = \begin{pmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma^2 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & \sigma^2 \end{pmatrix} = \sigma^2 \mathbf{I}$$

6. Linearly independent variables

No explanatory variable is a perfect linear function of any other explanatory variable(s).

- ▶ If this condition does not hold, we talk about **perfect (multi)collinearity**
- ▶ (Multi)collinearity can also be **imperfect**
- ▶ **Perfect multicollinearity:** one explanatory variable is an exact linear combination of one or more other explanatory variables
 - ▶ in this case, the OLS method is incapable to distinguish one variable from the other
 - ▶ technical consequence: $(\mathbf{X}'\mathbf{X})^{-1}$ does not exist
 - ▶ OLS estimation cannot be conducted
 - ▶ example: we include dummy variables for men and women together with the intercept

6. Linearly independent variables

- ▶ **Imperfect multicollinearity:**

- ▶ there is a linear relationship between the variables, but there is some error in that relationship
- ▶ example: we include two different variables that proxy for individual physical performance

- ▶ Consequences of the imperfect multicollinearity:

- ▶ the OLS estimator remains unbiased and consistent
- ▶ but the standard errors are inflated, often making the impacted parameters 'insignificant' even though they might be 'significant' individually (topic of lecture #4)

- ▶ Solution: drop some of the variables

7. Normality of the error term

The error term is normally distributed.

- ▶ This assumption is optional, but usually it is invoked
- ▶ Normality of the error term is inherited by the estimator $\hat{\beta}$
- ▶ Knowing the distribution of the estimator allows us to find its confidence intervals and to test hypotheses about parameters (as we will see in the next lecture #4)

Properties of the OLS estimator

- ▶ OLS estimator is defined by the formula:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

where $\mathbf{y} = \mathbf{X}\beta + \varepsilon$

- ▶ Hence, it is dependent on the random variable ε and thus $\hat{\beta}$ is a random variable
- ▶ The properties of $\hat{\beta}$ are based on the properties of ε
- ▶ The probability distribution of $\hat{\beta}$ is called a **sampling distribution**

Gauss-Markov Theorem

Given Classical Assumptions 1. - 6., the OLS estimator of β is the minimum variance estimator from among the set of all linear unbiased estimators of β .

- ▶ The theorem is also known as a stating: '**OLS is BLUE**', where BLUE stands for '**Best Linear Unbiased Estimator**'
- ▶ It means that:
 - ▶ OLS has the minimum variance of all unbiased linear estimators (it is **efficient**), i.e. it is the best
 - ▶ OLS is linear: $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{L}\mathbf{y}$
 - ▶ OLS is **unbiased** (see a proof on the next slide)
- ▶ Assumption 7., normality, is not needed for this theorem but if it also holds, OLS even becomes '**BUE**'!

Expected value of the OLS estimator

- We show:

▶ Var OLS

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \varepsilon) = \\ &= \underbrace{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}}_I\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon\end{aligned}$$

$$\begin{aligned}E[\hat{\beta}] &= E[\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon] = E[\beta] + \underbrace{E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon]}_! = \\ &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\underbrace{E[\varepsilon]}_0 = \beta\end{aligned}$$

- Since $E[\hat{\beta}] = \beta$, OLS is **unbiased**

Consistency of the OLS estimator

- When no explanatory variables are correlated with the error term (Assumption 3.), the OLS estimator is **consistent**:

$$E [\mathbf{X}' \boldsymbol{\varepsilon}] = \mathbf{0} \quad \Rightarrow \quad \hat{\boldsymbol{\beta}} \xrightarrow{n \rightarrow \infty} \boldsymbol{\beta}$$

- In other words: as the number of observations increases, the estimator converges to the true value of the parameter
- Consistency is the most important property of any estimator**

▶ Idea of the proof

Consistency of the OLS estimator

- ▶ As long as the OLS estimator $\hat{\beta}$ is consistent, the residual is a consistent estimator of the error term: $\hat{\varepsilon} = e$;
- ▶ If we have a consistent estimator of the error term, we can test if it satisfies the Classical Assumptions
- ▶ Moreover, possible deviations from the Classical Assumptions can be corrected
- ▶ As a consequence, the Assumption 3. of zero correlation between explanatory variables and the error term:

$$E [\mathbf{X}' \varepsilon] = \mathbf{0}$$

is **the most important** one to satisfy in regression models

Variance of the OLS estimator

- We show:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon$$

► EV OLS

$$\begin{aligned}Var[\hat{\beta}] &= Var[\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon] = \\&= Var(\beta) + Var[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon] = \\&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \cdot Var[\varepsilon] \cdot [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' = \\&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \cdot \underbrace{Var[\varepsilon]}_{\sigma^2\mathbf{I}} \cdot \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \\&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

Variance for the simple linear regression model case

- ▶ In the special case of a ‘regression line’:

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

we have:

$$\text{Var} [\hat{\beta}_1] = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}$$

Normality of the OLS estimator

- When we assume that $\varepsilon_i \sim N(0, \sigma^2)$, we can see that:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon = \beta + \mathbf{L}\varepsilon$$

is also normally distributed (it is a linear combination of normally distributed variables)

- Hence, we say that $\hat{\beta}$ is jointly normal:

$$\hat{\beta} \sim N\left(\beta, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}\right)$$

- This will help us to test hypotheses about regression parameters (next lecture #4)
- Note that the normality of the error term is not required for large samples, because then $\hat{\beta}$ is asymptotically normal (CLT)

Summary

- ▶ We have listed the Classical Assumptions of regression models:
 - ▶ model linear in parameters, explanatory variables linearly independent
 - ▶ (normally distributed) error term with zero mean and constant variance, no autocorrelation
 - ▶ no correlation between the error term and explanatory variables
- ▶ If these assumptions hold, the OLS estimator is:
 - ▶ unbiased (if $E[\varepsilon] = \mathbf{0}$ & no correlation between \mathbf{X} and ε)
 - ▶ consistent (if no correlation between \mathbf{X} and ε)
 - ▶ efficient (if homoskedasticity and no autocorrelation of ε)
 - ▶ normally distributed (if ε normally distributed)

Seminars and the next lecture #4

- ▶ **Seminars** tomorrow:
 - ▶ interpretation of estimated coefficients
 - ▶ specification and assumptions of regression models
 - ▶ prediction/forecasting using estimated models
- ▶ **Next lecture:**
 - ▶ testing hypotheses about parameters/coefficients (t -test)
 - ▶ software output of a regression
- ▶ **Readings for lecture #4:**
 - ▶ Studenmund (2016 & 17, [2014]): Chapter 5.1-5.5 [5]
 - ▶ Wooldridge (2016, 2012): Chapters 2-4a, 4-2-4-4, (6-1)

Appendix: Idea of the proof (not mandatory)

- When no explanatory variables are correlated with the error term (Assumption 3.), the OLS estimator is **consistent**:

$$E [\mathbf{X}' \boldsymbol{\varepsilon}] = \mathbf{0} \quad \Rightarrow \quad \hat{\boldsymbol{\beta}} \xrightarrow{n \rightarrow \infty} \boldsymbol{\beta}$$

- We can express:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y} = \boldsymbol{\beta} + (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \boldsymbol{\varepsilon} \\ &= \boldsymbol{\beta} + \left(\frac{1}{n} \mathbf{X}' \mathbf{X} \right)^{-1} \frac{1}{n} \mathbf{X}' \boldsymbol{\varepsilon}\end{aligned}$$

- We assume that there exists a finite matrix \mathbf{Q} so that:

$$\frac{1}{n} \mathbf{X}' \mathbf{X} \xrightarrow{n \rightarrow \infty} \mathbf{Q}$$

- It can be shown that: $\frac{1}{n} \mathbf{X}' \boldsymbol{\varepsilon} \xrightarrow{n \rightarrow \infty} E[\mathbf{X}' \boldsymbol{\varepsilon}] \stackrel{\text{by assumption}}{=} \mathbf{0}$

- This implies the consistency of $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}} \xrightarrow{n \rightarrow \infty} \boldsymbol{\beta} + \mathbf{Q}^{-1} \cdot E[\mathbf{X}' \boldsymbol{\varepsilon}] = \boldsymbol{\beta} + \mathbf{Q}^{-1} \cdot \mathbf{0} = \boldsymbol{\beta}$$

▶ Back

LECTURE #4

Introductory Econometrics

HYPOTHESES TESTING AFTER OLS ESTIMATION

Jiri Kukacka, Ph.D.

Institute of Economic Studies, Faculty of Social Sciences, Charles University

Winter semester 2021, October 20

In the previous lecture #3

- ▶ We have listed the Classical Assumptions of regression models:
 1. The regression model is linear in parameters, is correctly specified, and has an additive error term
 2. The error term has a zero population mean
 3. All explanatory variables are uncorrelated with the error term
 4. Observations of the error term are uncorrelated with each other
 5. The error term has a constant variance
 6. No explanatory variable is a perfect linear function of any other explanatory variable(s)
 7. (The error term is normally distributed)

In the previous lecture #3

- ▶ If all these assumptions hold, OLS estimator is:
 - ▶ unbiased
 - ▶ consistent
 - ▶ efficient
 - ▶ (normally distributed)
- ▶ Therefore, under Assumptions 1. - 6., OLS is the best linear unbiased estimator for linear regression models ('**OLS is BLUE**')
- ▶ And, under additional Assumption 7., the OLS estimator is normally distributed and it becomes '**BUE**'
- ▶ Unbiasedness of the OLS estimator: $E [\hat{\beta}] = \beta$
- ▶ Variance of the OLS estimator: $\text{Var} [\hat{\beta}] = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$

In today's lecture #4, we will...

- ▶ Discuss how hypotheses about parameters/coefficients can be tested in regression models
- ▶ Explain what significance of parameters/coefficients means
- ▶ Learn how to read software regression output
- ▶ Readings for this week:
 - ▶ Studenmund (2016 & 17, [2014]): Chapter 5.1-5.5 [5]
 - ▶ Wooldridge (2016, 2012): Chapters 2-4a, 4-2-4-4, (6-1)

Questions we ask

- ▶ What conclusions can we draw from our regression?
- ▶ What can we learn about the real world from a sample?
- ▶ Is it likely that our results could have been obtained by chance?
- ▶ If our theory is correct, what are the odds that this particular outcome would have been observed?

Hypothesis testing

- ▶ We **cannot prove** that a given hypothesis is '**correct**' using hypothesis testing
- ▶ All that can be done is to state that a particular sample conforms to a particular hypothesis
- ▶ We can often **reject** a given hypothesis with a certain degree of confidence
- ▶ In such a case, we conclude that it is very unlikely the sample result would have been observed if the hypothesized theory were correct

Null and alternative hypotheses

- ▶ First step in hypothesis testing: state explicitly the hypothesis to be tested
- ▶ *Null hypothesis*: specification of the range of values of the regression parameter/coefficient that would be expected to occur if the researcher's theory were **not correct**
- ▶ *Alternative hypothesis*: specification of the range of values of the parameter/coefficient that would be expected to occur if the researcher's theory were **correct**
- ▶ In other words: we define the **null hypothesis** as the result **we do not expect**

Null and alternative hypotheses

- ▶ Consider the model:

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

- ▶ Notation:

- ▶ H_0 ... null hypothesis
- ▶ H_A ... alternative hypothesis

- ▶ Examples:

- ▶ *One-sided test:*

$$H_0 : \beta_1 \leq 0$$

$$H_A : \beta_1 > 0$$

- ▶ *Two-sided test:*

$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 \neq 0$$

Type I and Type II errors

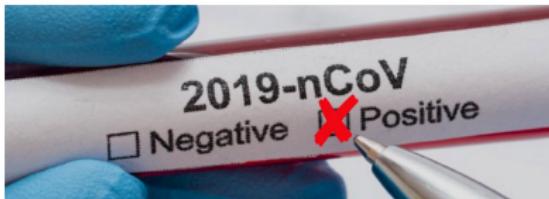
- ▶ It would be unrealistic to think that conclusions drawn from regression analysis will always be right
- ▶ There are two types of errors we can make:
 - ▶ Type I: we reject a true null hypothesis
 - ▶ Type II: we do not reject a false null hypothesis
- ▶ Example:
 - ▶ $H_0 : \beta_1 = 0$
 - ▶ $H_A : \beta_1 \neq 0$
 - ▶ Type I error: it holds that $\beta_1 = 0$, but we conclude $\beta_1 \neq 0$
 - ▶ Type II error: it holds that $\beta_1 \neq 0$, but we conclude $\beta_1 = 0$

Type I and Type II errors

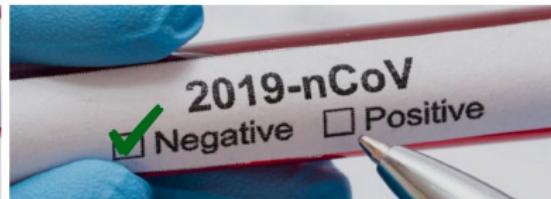
- ▶ Example:
 - ▶ H_0 : The defendant is innocent
 - ▶ H_A : The defendant is guilty
 - ▶ Type I error = sending an innocent person to jail
 - ▶ Type II error = freeing a guilty person
- ▶ Obviously, lowering the probability of Type I error means increasing the probability of Type II error
- ▶ In hypothesis testing, we focus on Type I error and we ensure that its probability is not unreasonably large

An alternative way to remember

Type I error (false positive)



Type II error (false negative)



Null:

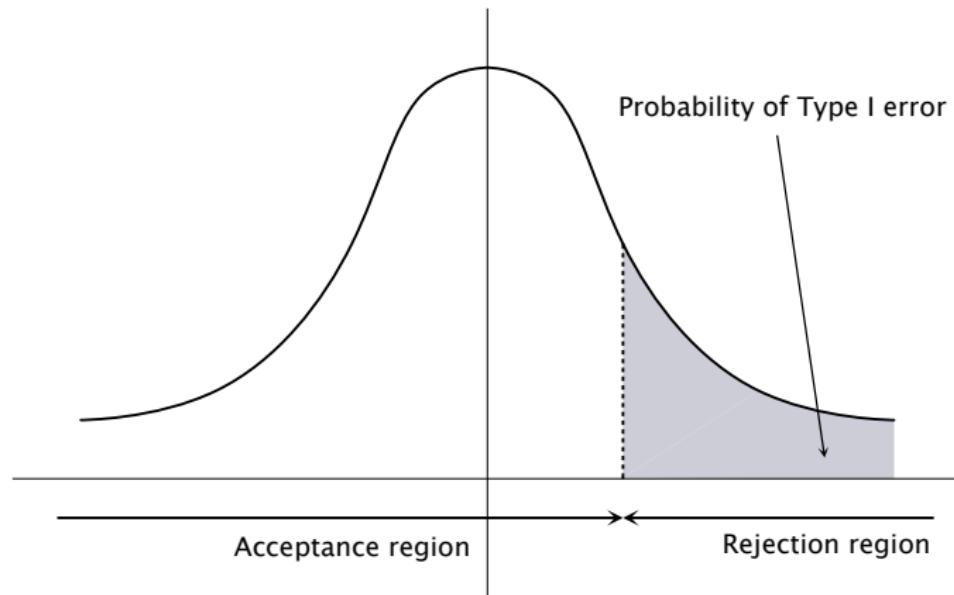


Decision rule

1. Calculate the **sample statistic**
2. Compare the sample statistic with the **critical value** (from the statistical tables)
 - ▶ The critical value divides the range of possible values of the statistic into two regions: “*acceptance*” region and *rejection region*:
 - ▶ if the sample statistic falls into the rejection region, **reject H_0**
 - ▶ if it falls into the “*acceptance*” region, **do not reject H_0**
 - ▶ The idea is that if the value of the parameter estimate does not support H_0 , the sample statistic should fall into the rejection region

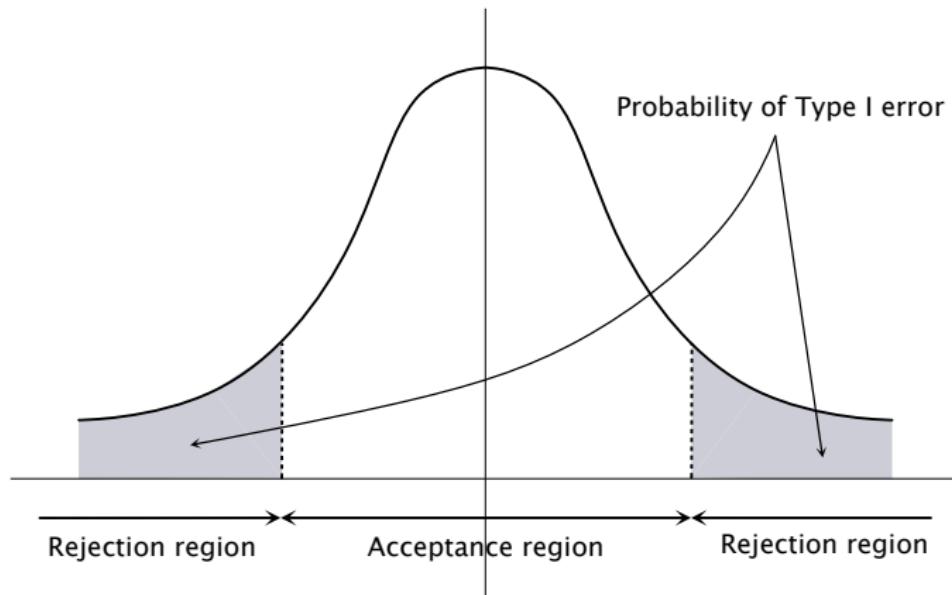
One-sided rejection region

- $H_0 : \beta_1 \leq 0$ vs $H_A : \beta_1 > 0$
- Distribution of $\hat{\beta}_1$:



Two-sided rejection region

- $H_0 : \beta_1 = 0$ vs $H_A : \beta_1 \neq 0$
- Distribution of $\hat{\beta}_1$:



The t -test

- ▶ We use t -test to test a hypothesis about an individual regression parameter (after controlling for all other independent variables)
- ▶ Tests of more than one parameter at a time (joint hypotheses) are typically done with the F -test (next lecture #5)
- ▶ The t -test is appropriate to use when the stochastic error term is **normally** distributed (CA 7.) and when the variance of that distribution is unknown
 - ▶ these are usual assumptions in the regression analysis
- ▶ The t -test accounts for differences in the units of measurement of the variables

The t -test

- ▶ Consider the model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

- ▶ Suppose we want to test (b is some constant):

$$H_0 : \beta_1 = b \quad \text{vs} \quad H_A : \beta_1 \neq b$$

- ▶ We know from the last lecture that:

$$\hat{\beta}_1 \sim N\left(\beta_1, Var(\hat{\beta}_1)\right) \Rightarrow \frac{\hat{\beta}_1 - \beta_1}{\sqrt{Var(\hat{\beta}_1)}} \sim N(0, 1),$$

where $Var(\hat{\beta}_1)$ is an element of the var-cov matrix of $\hat{\beta}$

The t -test

$$\begin{aligned}Var(\hat{\beta}) &= Var \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} \\&= \begin{pmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) & Cov(\hat{\beta}_0, \hat{\beta}_2) \\ Cov(\hat{\beta}_1, \hat{\beta}_0) & Var(\hat{\beta}_1) & Cov(\hat{\beta}_1, \hat{\beta}_2) \\ Cov(\hat{\beta}_2, \hat{\beta}_0) & Cov(\hat{\beta}_2, \hat{\beta}_1) & Var(\hat{\beta}_2) \end{pmatrix} \\&= \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

$$Var(\hat{\beta}_1) = \sigma^2 (\mathbf{X}'\mathbf{X})_{22}^{-1} \Rightarrow \sigma_{\hat{\beta}_1} = \sqrt{Var(\hat{\beta}_1)} = \sqrt{\sigma^2 (\mathbf{X}'\mathbf{X})_{22}^{-1}}$$

The t -test

- ▶ Problem: we do not know the value of the parameter σ^2 (the variance of the error term)
- ▶ It has to be estimated as:

$$\hat{\sigma}^2 := s^2 = \frac{\mathbf{e}'\mathbf{e}}{n - k - 1},$$

where k is the number of independent/explanatory variables (here $k = 2$ and we need to consider additional -1 for the intercept), and \mathbf{e} is the vector of residuals

- ▶ We denote **standard error** of $\hat{\beta}_1$ (sample counterpart of standard deviation $\sigma_{\hat{\beta}_1}$):

$$s.e.(\hat{\beta}_1) = \sqrt{s^2 (\mathbf{X}'\mathbf{X})_{22}^{-1}}$$

The t -test

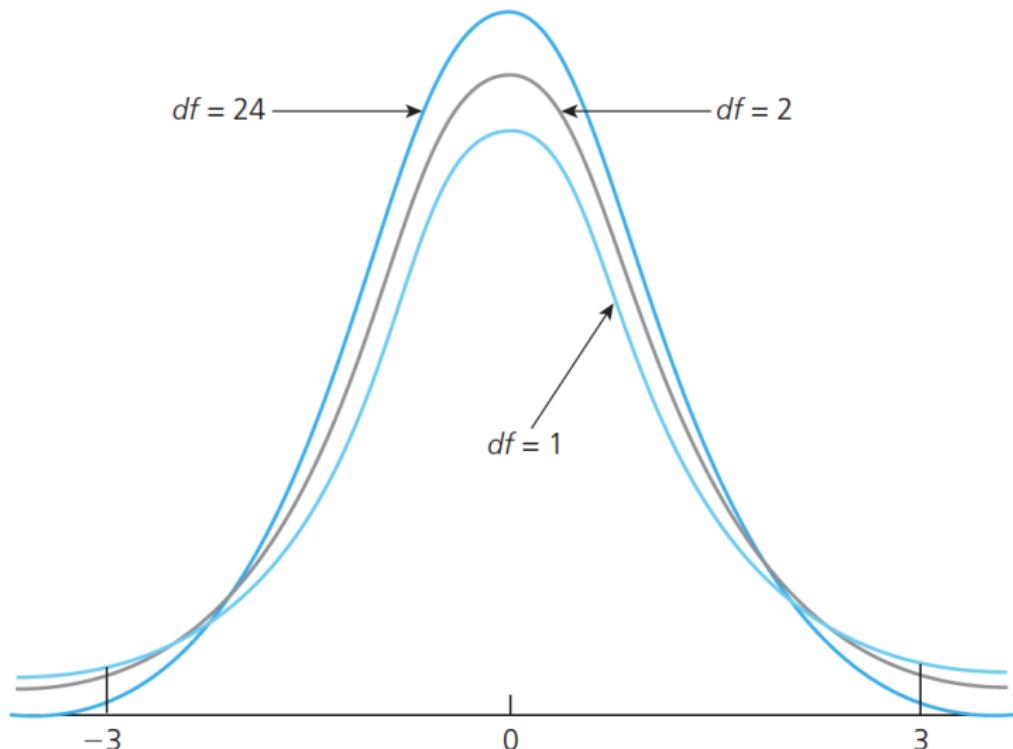
- We define the t -statistic:

$$t := \frac{\widehat{\beta}_1 - \beta_1}{\text{s.e.}(\widehat{\beta}_1)} = \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{s^2 (\mathbf{X}'\mathbf{X})_{22}^{-1}}} \sim t_{n-k-1},$$

where $\widehat{\beta}_1$ is the estimated parameter and β_1 is the value of the parameter that is stated in our null hypothesis

- This statistic depends on the estimate $\widehat{\beta}_1$, our null hypothesis about β_1 , $\text{s.e.}(\widehat{\beta}_1)$, and it has a known distribution

t_{n-k-1} distribution with various dfs



Source: Wooldridge (2016, pg. 671)

One-sided t -test

- ▶ Suppose our hypothesis is:

$$H_0 : \beta_1 \leq b \quad \text{vs} \quad H_A : \beta_1 > b$$

- ▶ Our t -statistic is:

$$t = \frac{\hat{\beta}_1 - b}{s.e.(\hat{\beta}_1)}$$

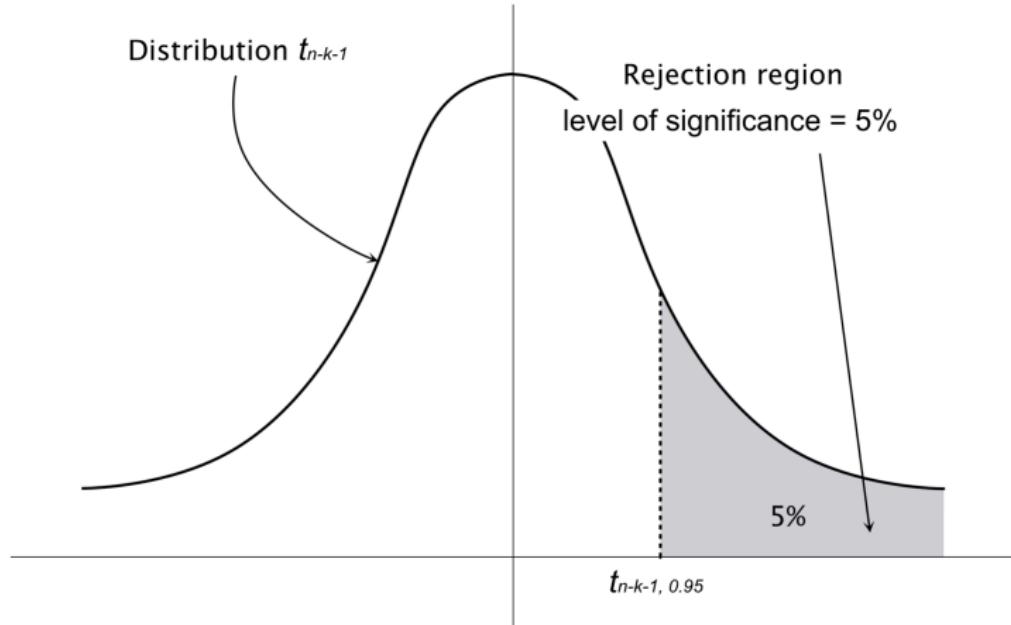
- ▶ $\hat{\beta}_1$ is the estimated regression parameter β_1
- ▶ b is the constant from our null hypothesis
- ▶ $s.e.(\hat{\beta}_1)$ is the estimated standard deviation of $\hat{\beta}_1$

One-sided t -test

How to determine the **critical value** for this test statistic?

1. We set the probability of Type I error:
 - ▶ keep in mind that extremely low probability of Type I error dramatically increases the probability of Type II error
 - ▶ typical practice is thus 5%, denoted as **level of significance** α
 - ▶ $1 - \alpha = 95\%$ is then called **level of confidence**
2. We find the critical value in the statistical tables: $t_{n-k-1, 0.95}$, the critical value depends on:
 - ▶ the chosen level of significance/confidence
 - ▶ degrees of freedom $n - k - 1$
 - ▶ the type of hypothesis: 1-sided or 2-sided

One-sided t -test



- We **reject** H_0 if $|t| > t_{n-k-1, 0.95}$ and t also has the sign implied by H_A , otherwise we **do not** reject

Two-sided t -test

- Our hypothesis is:

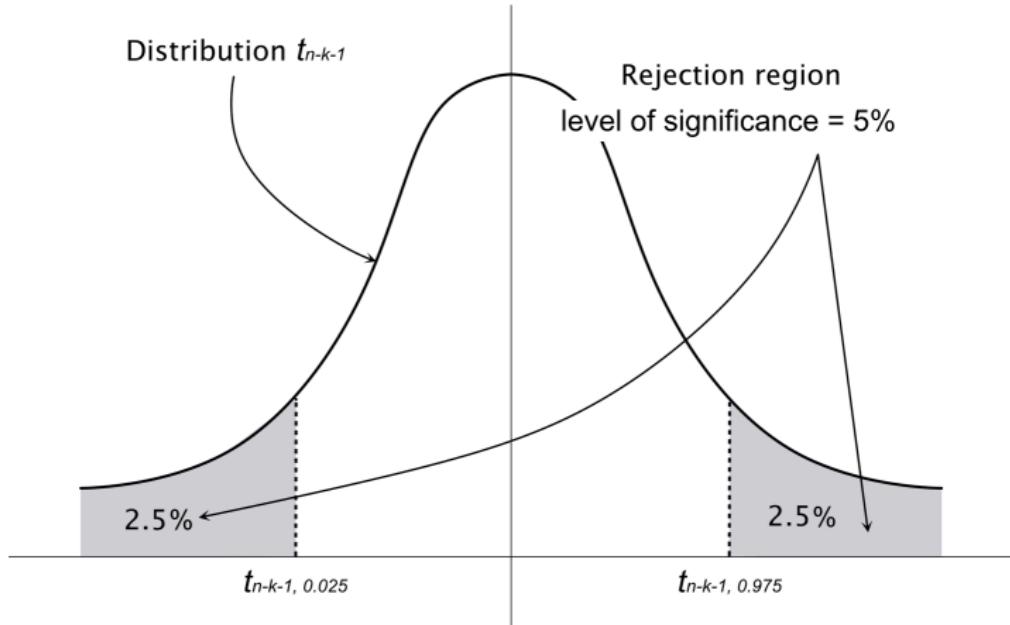
$$H_0 : \beta_1 = b \quad \text{vs} \quad H_A : \beta_1 \neq b$$

- Hence, our t -statistic still is:

$$t = \frac{\widehat{\beta}_1 - b}{s.e.(\widehat{\beta}_1)}$$

- We set the level of significance $\alpha = 5\%$
- We compare our statistic to the critical values $t_{n-k-1, 0.975}$ and $t_{n-k-1, 0.025}$

Two-sided t -test



- ▶ Note that $t_{n-k-1, 0.975} = -t_{n-k-1, 0.025}$
- ▶ We **reject** H_0 if $|t| > t_{n-k-1, 0.975}$

Statistical significance of a parameter

- The most common test performed in a regression is:

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_A : \beta_1 \neq 0$$

with the t -statistic:

$$t = \frac{\widehat{\beta}_1}{\text{s.e.}(\widehat{\beta}_1)} \sim t_{n-k-1}$$

- If we reject $H_0 : \beta_1 = 0$, we say the parameter/coefficient β_1 is **statistically significantly different from 0** (or briefly that it is '**significant**')
- This t -statistic is displayed in most regression outputs

The p -value

- ▶ Classical approach to hypothesis testing:
 - ▶ first choose the level of significance (e.g. 5%)
 - ▶ then test the null hypothesis at that level of significance
- ▶ However, there is no ‘correct’ level of significance!
- ▶ Or we can ask a more informative question: **What is the smallest level of significance at which the null hypothesis would still be rejected?**
 - ▶ this level of significance is called **p -value**
 - ▶ the smaller the p -value → the smaller the probability of rejecting the true H_0 (i.e. of Type I error) → the bigger the level of confidence that H_0 is indeed correctly rejected
 - ▶ p -value for $H_0 : \beta_j = 0$ is displayed in most regression outputs

Example

- Let us study the impact of years of education on wages:

$$wage = \beta_0 + \beta_1 education + \beta_2 experience + \varepsilon$$

- Output from Gretl:

Model 1: OLS, using observations 1-526

Dependent variable: wage

	coefficient	std. error	t-ratio	p-value
<hr/>				
const	-3.39054	0.766566	-4.423	1.18e-05 ***
educ	0.644272	0.0538061	11.97	2.28e-29 ***
exper	0.0700954	0.0109776	6.385	3.78e-10 ***

Mean dependent var	5.896103	S.D. dependent var	3.693086
Sum squared resid	5548.160	S.E. of regression	3.257044
R-squared	0.225162	Adjusted R-squared	0.222199
F(2, 523)	75.98998	P-value(F)	1.07e-29
Log-likelihood	-1365.969	Akaike criterion	2737.937
Schwarz criterion	2750.733	Hannan-Quinn	2742.948

Limitations of the *t*-test

- ▶ It does not test importance:

$$\widehat{sales}_t = 300 + \underset{(1)}{10}adv_exp_{1,t} + \underset{(25)}{200}adv_exp_{2,t}$$

- ▶ It does not test the theoretical validity:

$$\widehat{P}_{Amazon,2017,t} = 757.92 + \underset{(0.14)}{0.52} CR_{UK,2017,t}$$

Confidence interval

- ▶ A 95% confidence interval of β is an interval centred around $\hat{\beta}$ such that β falls into this interval with probability 95%:

$$P(\hat{\beta} - c < \beta < \hat{\beta} + c) =$$

$$= P\left(-\frac{c}{s.e.(\hat{\beta})} < \frac{\hat{\beta} - \beta}{s.e.(\hat{\beta})} < \frac{c}{s.e.(\hat{\beta})}\right) = 0.95$$

- ▶ Since $\frac{\hat{\beta} - \beta}{s.e.(\hat{\beta})} \sim t_{n-k-1}$, we derive the confidence interval:

$$\hat{\beta} \pm t_{n-k-1, 1-\frac{\alpha}{2}} \cdot s.e.(\hat{\beta})$$

Confidence interval

- ▶ Output from Gretl (wage regression):

Model 1: OLS, using observations 1-526

Dependent variable: wage

	coefficient	std. error	t-ratio	p-value

const	-3.39054	0.766566	-4.423	1.18e-05 ***
educ	0.644272	0.0538061	11.97	2.28e-29 ***
exper	0.0700954	0.0109776	6.385	3.78e-10 ***

...

- ▶ A 95% confidence interval for the coefficient on education:

$$\widehat{\beta}_1 \pm t_{n-k-1, 0.975} \cdot s.e.(\widehat{\beta}_1) = 0.644 \pm 1.96 \cdot 0.054 = \\ \langle 0.538; 0.750 \rangle$$

$\Rightarrow \beta_1 \in \langle 0.538; 0.750 \rangle$ with 95% probability

Summary

- ▶ We discussed the principle of hypothesis testing
- ▶ We derived the t -statistic
- ▶ We defined the concept of the p -value
- ▶ We explained what significance of a parameter/coefficient means
- ▶ We observed a regression output on an example
- ▶ We learnt how to construct confidence intervals

Seminars and the next lecture #5

- ▶ Home assignment #1 is due tomorrow!
- ▶ **Seminars:**
 - ▶ testing hypotheses in regression models and practical implications of the t -test
 - ▶ linear transformation of variables
 - ▶ computing confidence intervals
 - ▶ interpreting output from statistical software
- ▶ Next **lecture & seminars** (October 27 & 28):
 - ▶ **ONLINE ONLY**, delivered via **pre-recorded videos**
 - ▶ testing of multiple hypotheses (F -test)
 - ▶ assessing the goodness of fit (R^2)
- ▶ Readings for lecture #5:
 - ▶ Studenmund (2016 & 17, [2014]): Chapters 2.2 (pg. 47–49 & 65–67 [48–50]), 2.4–2.5, 5.5 [Chapter 5 Appendix]
 - ▶ Wooldridge (2016, 2012): Chapters 2-3b–c, 3-2h, 4-5–4-6, 6-1, 6-3

LECTURE #5

Introductory Econometrics

MULTIPLE HYPOTHESES TESTING & GOODNESS OF FIT

Jiri Kukacka, Ph.D.

Institute of Economic Studies, Faculty of Social Sciences, Charles University

Winter semester 2021, October 27

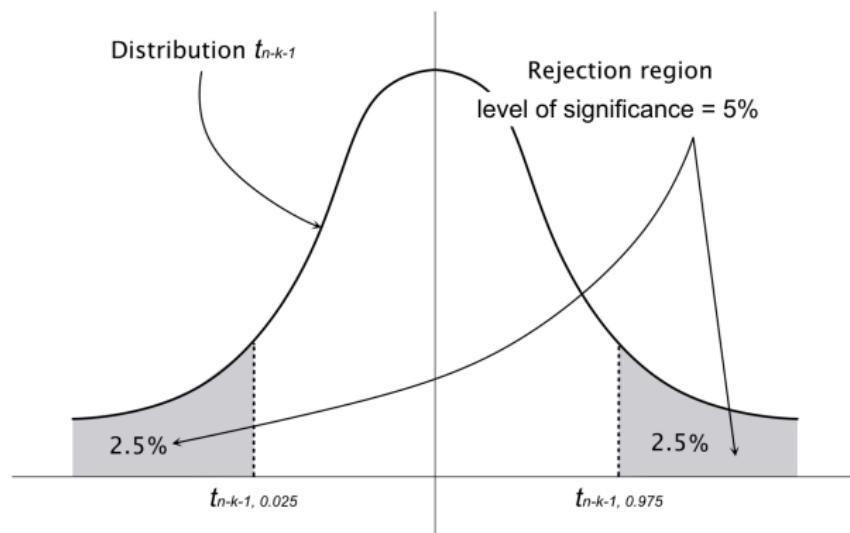
In the previous lecture #4

- ▶ We discussed the principle of hypotheses testing:
 - ▶ Type I and Type II errors
 - ▶ critical value and rejection region
- ▶ We derived the t -statistic: $t = \frac{\hat{\beta} - \beta}{s.e.(\hat{\beta})}$
- ▶ We defined the concept of the p -value

Recap: Two-sided t -test

- We test:

$$H_0 : \beta = 0 \quad \text{vs} \quad H_A : \beta \neq 0$$



- We **reject** H_0 if $|t| > t_{n-k-1, 0.975}$, otherwise we **do not** reject

In the previous lecture #4

- ▶ We explained what significance of a parameter/coefficient means
- ▶ We derived the confidence interval: $\hat{\beta} \pm t_{n-k-1, 1-\frac{\alpha}{2}} \cdot s.e.(\hat{\beta})$
- ▶ We observed the regression output on an example:

Model 1: OLS, using observations 1-526

Dependent variable: wage

	coefficient	std. error	t-ratio	p-value

const	-3.39054	0.766566	-4.423	1.18e-05 ***
educ	0.644272	0.0538061	11.97	2.28e-29 ***
exper	0.0700954	0.0109776	6.385	3.78e-10 ***

Mean dependent var	5.896103	S.D. dependent var	3.693086
Sum squared resid	5548.160	S.E. of regression	3.257044
R-squared	0.225162	Adjusted R-squared	0.222199
F(2, 523)	75.98998	P-value(F)	1.07e-29
Log-likelihood	-1365.969	Akaike criterion	2737.937
Schwarz criterion	2750.733	Hannan-Quinn	2742.948

In today's lecture #5, we will...

- ▶ Explain how multiple hypotheses are tested in a regression model
- ▶ Define the notion of the overall significance of a regression
- ▶ Introduce a measure of the goodness of fit of a regression (R^2)
- ▶ Readings for this week:
 - ▶ Studenmund (2016 & 17, [2014]): Chapters 2.2 (pg. 48–50 [47–49]), 2.4–2.5, 5.5 [Chapter 5 Appendix]
 - ▶ Wooldridge (2016, 2012): Chapters 2-3b–c, 3-2h, 4-5–4-6, 6-1, 6-3

Testing multiple hypotheses

- ▶ Suppose we have a model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

- ▶ We want to test multiple linear hypotheses/restrictions in this model
- ▶ For example, we want to see if the following restrictions on coefficients hold jointly:

$$\beta_1 + \beta_2 = 1 \quad \text{and} \quad \beta_3 = 0$$

- ▶ We cannot use the t -test in this case (t -test can be used only for one hypothesis about an individual regression parameter)
- ▶ We will use an F -test

Restricted vs. unrestricted model

- ▶ We can reformulate the model by plugging the restrictions as if they were true (model under H_0)
- ▶ We call this model *restricted model* as opposed to the *unrestricted model*
- ▶ The unrestricted model is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

- ▶ We derive (in the lecture) the restricted model for:

$$H_0 : \beta_1 + \beta_2 = 1 \quad \text{and} \quad \beta_3 = 0$$

- ▶ It becomes: $y_i^* = \beta_0 + \beta_1 x_i^* + \varepsilon_i$
with $y_i^* = y_i - x_{i2}$ and $x_i^* = x_{i1} - x_{i2}$

Idea of the F -test

- ▶ If the restrictions are true, then the restricted model fits the data in the same way as the unrestricted model
 - ▶ residuals are nearly the same
- ▶ If the restrictions are false, then the restricted model fits the data poorly
 - ▶ residuals from the restricted model are much larger than those from the unrestricted model
- ▶ The idea is thus to compare the residuals from the two models

Idea of the F -test

- ▶ How to compare residuals in the two models?
 - ▶ calculate the sum of squared residuals in the two models
 - ▶ test if the difference between the two sums is equal to zero (statistically):
 - ▶ H_0 : the difference is zero (residuals in the two models are the same, restrictions hold)
 - ▶ H_A : the difference is positive (residuals in the restricted model are bigger, restrictions do not hold)
- ▶ Sum of squared residuals (also called **Residual Sum of Squares**):

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2$$

F -test

- The test statistic is defined as:

$$F = \frac{(RSS_R - RSS_U)/J}{RSS_U/(n - k - 1)} \sim F_{J, n-k-1}$$

where:

RSS_R ... Residual Sum of Squares from the restricted model

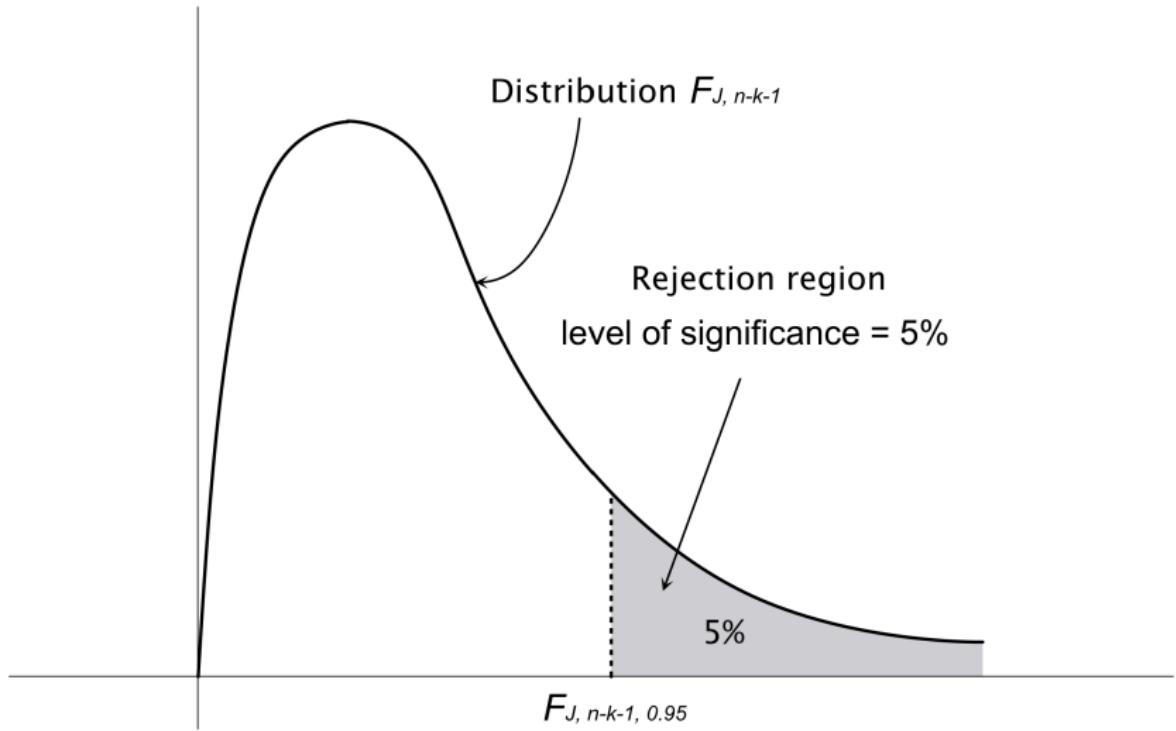
RSS_U ... Residual Sum of Squares from the unrestricted model

J ... number of restrictions

n ... number of observations

k ... number of independent/explanatory variables and we need to consider additional -1 for the intercept

F -test



Back to our example

- We had the model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

- We wanted to test:

$$H_0 : \begin{cases} \beta_1 + \beta_2 = 1 \\ \beta_3 = 0 \end{cases} \quad \text{vs} \quad H_A : \begin{cases} \beta_1 + \beta_2 \neq 1 \\ \text{or} \\ \beta_3 \neq 0 \end{cases}$$

- Under H_0 , we obtained the restricted model:

$$y_i^* = \beta_0 + \beta_1 x_i^* + \varepsilon_i,$$

where $y_i^* = y_i - x_{i2}$ and $x_i^* = x_{i1} - x_{i2}$

Back to our example

- ▶ From the regression on the unrestricted model we get RSS_U
- ▶ From the regression on the restricted model we get RSS_R
- ▶ We have $J = 2$ and $k = 3$
- ▶ We construct the F -statistic:
$$F = \frac{(RSS_R - RSS_U)/2}{RSS_U/(n-3-1)}$$
- ▶ We find the critical value of the F distribution with 2 and $n - 4$ degrees of freedom at the 95% confidence level
- ▶ If $F > F_{2,n-4,0.95}$, we reject the null hypothesis at the 5% significance level
 - ▶ we reject that both 2 restrictions hold jointly

Application 1: Overall significance of the regression

- ▶ Usually, we are interested in knowing if the model has some explanatory power, i.e. if the independent variables indeed 'explain' the dependent variable
- ▶ We test this using the F -test of the joint significance of all k slope coefficients, i.e. $J = k$:

$$H_0 : \left\{ \begin{array}{l} \beta_1 = 0 \\ \beta_2 = 0 \\ \vdots \\ \beta_k = 0 \end{array} \right. \quad \text{vs} \quad H_A : \left\{ \begin{array}{l} \beta_j \neq 0 \\ \text{for at least one } j = 1, \dots, k \end{array} \right.$$

Application 1: Overall significance of the regression

- Unrestricted model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i$$

- Restricted model:

$$y_i = \beta_0 + \varepsilon_i$$

- F -statistic:

$$F = \frac{(RSS_R - RSS_U)/k}{RSS_U/(n - k - 1)} \sim F_{k,n-k-1}$$

- number of restrictions: $J = k$ (in this specific case only)
- this F -statistic and the corresponding p -values are part of the regression output

Application 1: Overall significance of the regression

Output from Gretl:

Model 1: OLS, using observations 1-526

Dependent variable: wage

	coefficient	std. error	t-ratio	p-value	

const	-3.39054	0.766566	-4.423	1.18e-05	***
educ	0.644272	0.0538061	11.97	2.28e-29	***
exper	0.0700954	0.0109776	6.385	3.78e-10	***
Mean dependent var	5.896103	S.D. dependent var	3.693086		
Sum squared resid	5548.160	S.E. of regression	3.257044		
R-squared	0.225162	Adjusted R-squared	0.222199		
F(2, 523)	75.98998	P-value(F)	1.07e-29		
Log-likelihood	-1365.969	Akaike criterion	2737.937		
Schwarz criterion	2750.733	Hannan-Quinn	2742.948		

Application 2: Test for seasonality

A test for quarterly seasonality, in which 3 variables equal to 1 for given season and 0 otherwise represent Q_1 , Q_2 , and Q_3 , is equivalent to:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \quad \text{vs} \quad H_A : \text{otherwise},$$

where β_1 , β_2 , β_3 are the respective coefficients on Q_1 , Q_2 , Q_3

Application 3: Test for constant returns to scale

- ▶ Assume linearized Cobb-Douglas production function:

$$\ln Y = \beta_0 + \beta_1 \ln L + \beta_2 \ln K + \varepsilon$$

- ▶ A test for constant returns to scale is equivalent to:

$$H_0 : \beta_1 + \beta_2 = 1 \quad \text{vs} \quad H_A : \beta_1 + \beta_2 \neq 1$$

Application 4: Chow test

A so-called Chow test of whether the two sets of data result in significantly different regression coefficients assuming the same theoretical model

- ▶ estimate the model on individual samples to get RSS_1 and RSS_2
- ▶ estimate the model on a pooled sample to get RSS_P
- ▶ test whether $RSS_P - (RSS_1 + RSS_2)$ is significantly different from 0
- ▶ a special application of the general F -test

Goodness of fit measure

- ▶ We know that education and experience have a significant effect on wages
- ▶ But how important are they in determining wages?
- ▶ How much of difference in wages between people is explained by differences in education and in experience?
- ▶ How well variation in the independent variable(s) explains variation in the dependent variable?
- ▶ These are the questions answered by the goodness of fit measure: R^2

Total and explained variation

- ▶ **Total variation** in the dependent variable:

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

- ▶ Predicted value of the dependent variable: a part that is explained by independent variables:

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

(case of a simple regression model for simplicity of notation)

- ▶ **Explained variation** in the dependent variable:

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

Goodness of fit: R^2

- ▶ Denote:

- ▶ $TSS = \sum_{i=1}^n (y_i - \bar{y}_n)^2 \dots \text{Total Sum of Squares}$
- ▶ $ESS = \sum_{i=1}^n (\hat{y}_i - \bar{y}_n)^2 \dots \text{Explained Sum of Squares}$

- ▶ Define the measure of the goodness of fit:

$$R^2 = \frac{ESS}{TSS} = \frac{\text{Explained variation in } y}{\text{Total variation in } y}$$

Goodness of fit: R^2

- ▶ For all models: $0 \leq R^2 \leq 1$
- ▶ R^2 tells us what percentage of the total variation in the dependent variable is explained by the variation in the independent variable(s)
 - ▶ $R^2 = 0.3$ means that the independent variables can explain 30% of the variation in the dependent variable
- ▶ Higher R^2 means better fit of the regression model but not necessarily a better model!

Decomposing the variance

- ▶ For models with intercept, R^2 can be rewritten using the decomposition of variance
- ▶ Variance decomposition:

$$\sum_{i=1}^n (y_i - \bar{y}_n)^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y}_n)^2 + \sum_{i=1}^n e_i^2$$

- ▶ $TSS = \sum_{i=1}^n (y_i - \bar{y}_n)^2 \dots$ Total Sum of Squares
- ▶ $ESS = \sum_{i=1}^n (\hat{y}_i - \bar{y}_n)^2 \dots$ Explained Sum of Squares
- ▶ $RSS = \sum_{i=1}^n e_i^2 \dots$ Residual Sum of Squares

Variance decomposition and R^2

- ▶ Variance decomposition: $TSS = ESS + RSS$
- ▶ Intuition: total variation can be divided between the explained variation and the unexplained variation
 - ▶ the observed value y_i is a sum of estimated (explained) \hat{y}_i and the residual e_i (unexplained part): $y_i = \hat{y}_i + e_i$
- ▶ We can rewrite R^2 :

$$R^2 = \frac{ESS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

Adjusted R^2

- ▶ The Residual Sum of Squares (RSS) decreases when additional explanatory variables are introduced in the model, whereas Total Sum of Squares (TSS) remains the same
 - ▶ $R^2 = 1 - \frac{RSS}{TSS}$ increases if we add explanatory variables
 - ▶ models with more variables automatically have better fit
- ▶ To deal with this problem, we define the **adjusted** R^2 :

$$R_{adj}^2 = 1 - \frac{\frac{RSS}{n-k-1}}{\frac{TSS}{n-1}} \quad (\leq R^2),$$

where k is the number of independent/explanatory variables

- ▶ Adjusted R_{adj}^2 thus introduces a ‘punishment’ for including more explanatory variables

Regression output example

Output from Gretl:

Model 1: OLS, using observations 1-526

Dependent variable: wage

	coefficient	std. error	t-ratio	p-value	

const	-3.39054	0.766566	-4.423	1.18e-05	***
educ	0.644272	0.0538061	11.97	2.28e-29	***
exper	0.0700954	0.0109776	6.385	3.78e-10	***
Mean dependent var	5.896103	S.D. dependent var	3.693086		
Sum squared resid	5548.160	S.E. of regression	3.257044		
R-squared	0.225162	Adjusted R-squared	0.222199		
F(2, 523)	75.98998	P-value(F)	1.07e-29		
Log-likelihood	-1365.969	Akaike criterion	2737.937		
Schwarz criterion	2750.733	Hannan-Quinn	2742.948		

An example of the misuse of R_{adj}^2

► Model 1:

$$\hat{Q}_{mozzarella,t} = -0.85 + \frac{0.38}{(0.045)} Yd_t$$

$$R^2 = 0.897 \quad R_{adj}^2 = 0.885 \quad n = 10 \text{ (2000 – 2009)}$$

$Q_{mozzarella,t}$... per capita demand for mozzarella (in pounds)

Yd_t ... per capita disposable income (in USD thousands)

► Model 2:

$$\hat{Q}_{mozzarella,t} = -3.33 + \frac{0.25}{(0.034)} Yd_t - \frac{0.046}{(0.009)} DRO_t$$

$$R^2 = 0.978 \quad R_{adj}^2 = 0.972 \quad n = 10$$

DRO_t ... ?

F -test revisited

- ▶ Let us recall the F -statistic:

$$F = \frac{(RSS_R - RSS_U)/J}{RSS_U/(n - k - 1)} \sim F_{J,n-k-1}$$

- ▶ We can use the formula $R^2 = 1 - \frac{RSS}{TSS}$ to rewrite the F -statistic in R^2 form:

$$F = \frac{(R_U^2 - R_R^2)/J}{(1 - R_U^2)/(n - k - 1)} \sim F_{J,n-k-1}$$

- ▶ we can use this R^2 form of the F -statistic under the condition that $TSS_U = TSS_R$ (the dependent variables in restricted and unrestricted models are the same)

Summary

- ▶ We showed how linear restrictions are incorporated in regression models
- ▶ We explained the idea of the F -test
- ▶ We defined the notion of the overall significance of a regression
- ▶ We introduced the measure of the goodness of fit: R^2
- ▶ We learned how total variation in the dependent variable can be decomposed
- ▶ We showed how the F -test and the R^2 are related

Seminars and the next lecture #6

- ▶ **Seminars:**
 - ▶ testing single and multiple hypotheses in regression models
 - ▶ assessing goodness of fit of regression models
 - ▶ working with Gretl (BYOD?)
- ▶ **Next lecture:**
 - ▶ nonlinear specifications (double-log, semi-log, polynomial, inverse, lags)
 - ▶ dummy independent/explanatory variables
 - ▶ Chow test
- ▶ **Readings for lecture #6:**
 - ▶ Studenmund (2016 & 17, [2014]): Chapter[s 3.3,] 7
 - ▶ Wooldridge (2016, 2012): Chapters 2-4b, 6-2, 7-1–7-4

LECTURE #6

Introductory Econometrics

NONLINEAR SPECIFICATIONS & DUMMY VARIABLES

Jiri Kukacka, Ph.D.

Institute of Economic Studies, Faculty of Social Sciences, Charles University

Winter semester 2021, November 3

In the previous lecture #5, we have learnt...

- ▶ How linear restrictions are incorporated in regression models
- ▶ The idea of the F -test
- ▶ The notion of the overall significance of a regression
- ▶ Measure or the goodness of fit: R^2
- ▶ And how the F -test and the R^2 are related

In today's lecture, we will. . .

- ▶ Discuss different nonlinear specifications in dependent and independent variables and their interpretation
- ▶ Define the notion of a dummy variable and show its different uses in linear regression models
- ▶ Revisit the Chow test of the stability of coefficients
- ▶ Readings for this week:
 - ▶ Studenmund (2016 & 17, [2014]): Chapters 3.3, 7 [Chapter 7 only]
 - ▶ Wooldridge (2016, 2012): Chapters 2-4b, 6-2, 7-1–7-4

Nonlinear specification

- ▶ There is not always a linear relationship between dependent variable and explanatory variables:
 - ▶ the use of OLS requires that the equation be linear in parameters/coefficients
 - ▶ however, there is a wide variety of functional forms that are linear in parameters while being nonlinear in variables!
- ▶ We have to choose carefully the functional form of the relationship between the dependent variable and each explanatory variable:
 - ▶ the choice of a functional form should be based on the underlying economic theory and/or intuition
 - ▶ do we expect a curve instead of a straight line? Does the effect of a variable peak at some point and then start to decline?

Linear form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

- ▶ Assumes that the effect of the explanatory variable on the dependent variable is constant:

$$\frac{\partial y}{\partial x_k} = \beta_k \quad k = 1, 2$$

- ▶ Interpretation: if x_k increases by 1 **unit** (in which x_k is measured), then y will change by β_k **units** (in which y is measured)
- ▶ Linear form is used as default functional form unless we find a strong evidence that it is inappropriate

Double-log form

$$\ln y = \beta_0 + \beta_1 \ln x_1 + \beta_2 \ln x_2 + \varepsilon$$

- ▶ Assumes that the elasticity of the dependent variable with respect to the explanatory variable is constant:

$$\frac{\partial \ln y}{\partial \ln x_k} = \frac{\partial y/y}{\partial x_k/x_k} = \beta_k \quad k = 1, 2$$

- ▶ Interpretation: if x_k increases by 1 **percent**, then y will change by β_k **percent**
- ▶ Before using a double-log model, make sure that there are no negative or zero observations in the data set
 - ▶ we can use $\ln(1 + x_k)$ if x_k contains zeros

Example

- ▶ Estimating the production function of Indian sugar industry:

$$Y = AL^{\beta_1}K^{\beta_2}$$

Y ... output

L ... labour

K ... capital employed

A ... $\exp^{\beta_0 + \varepsilon}$

$$\widehat{\ln Y} = 2.7 + \begin{matrix} 0.59 \\ (0.14) \end{matrix} \ln L + \begin{matrix} 0.33 \\ (0.17) \end{matrix} \ln K$$

- ▶ interpretation: if we increase the amount of labour by 1%, the production of sugar will increase by 0.59%, holding the other included independent/explanatory variables constant (*ceteris paribus*)

Semi-log forms

- Log-linear form:

$$\ln y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

- interpretation: if x_k increases by 1 **unit**, then y will change by approx. $(\beta_k \cdot 100)$ **percent** ($k = 1, 2$)

- Linear-log form:

$$y = \beta_0 + \beta_1 \ln x_1 + \beta_2 \ln x_2 + \varepsilon$$

- interpretation: if x_k increases by 1 **percent**, then y will change by $(\beta_k / 100)$ **units** ($k = 1, 2$)

Example of log-linear form

- ▶ Estimating the impact of education and experience on wages:

$$\widehat{\ln wage} = 0.217 + 0.098 \text{ educ} + 0.010 \text{ exper}$$

(0.008) (0.002)

wage ... hourly wage (in USD)

educ ... years of education

exper ... years of experience

- ▶ Interpretation: an increase in education by one year increases hourly wage by (approx.) 9.8%, ceteris paribus; an increase in experience by one year increases hourly wage by (approx.) 1%, ceteris paribus

► Mathematically precise interpretation

Example of linear-log form

- ▶ Estimating consumption of beef:

$$\hat{C}_{beef} = -71.8 - 0.87 P + 98.9 \ln Yd$$

(0.13) (11.16)

$$R^2 = 0.750 \quad n = 28$$

C_{beef} ... annual per capita beef consumption (in pounds)

P ... price of beef (in USD cents)

Yd ... annual per capita disposable income (in USD thousands)

- ▶ Interpretation: an increase in the disposable income by 1% increases beef consumption by 0.989 pound per year, cet. par.
- ▶ Compare to linear form:

$$\hat{C}_{beef} = 37.5 - 0.88 P + 11.9 Yd$$

(0.16) (1.76)

$$R^2 = 0.631 \quad n = 28$$

Polynomial form

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

- ▶ To determine the effect of x on y , we need to calculate the first derivative:

$$\frac{\partial y}{\partial x} = \beta_1 + 2\beta_2 x$$

- ▶ Effect of x on y changes with the level of x
- ▶ We might also have higher order polynomials, e.g.:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \varepsilon$$

Example of polynomial form

- ▶ The impact of the number of hours of studying on the grade from Introductory Econometrics:

$$\widehat{\text{grade}} = 30 + 1.4\text{hours} - 0.009\text{hours}^2$$

- ▶ To determine the effect of hours on grade, calculate the first derivative:

$$\frac{\partial y}{\partial x} = \frac{\partial \text{grade}}{\partial \text{hours}} = 1.4 - 2 \cdot 0.009 \cdot \text{hours} = 1.4 - 0.018\text{hours}$$

- ▶ decreasing returns to hours of studying: more hours imply higher grade, but the positive effect of additional hour of studying decreases with more hours

(Lagged independent variables)

$$y_t = \beta_0 + \beta_1 x_{1,t-1} + \beta_2 x_{2,t} + \varepsilon_t$$

- ▶ In many real world situations time elapses between a change in the independent variable and the resulting change in the dependent variable
- ▶ Interpretation: be aware of the time structure!
- ▶ Example: see seminar #3/exercise 2, Stock Prices and Wall Street Weather:

$$\widehat{DJ}_t = \widehat{\beta}_0 + 0.1R_{t-1} + 0.001J_t - 0.017M_t + 0.0005C_t$$

DJ_t ... the percentage change in the DJIA on day t

R_{t-1} ... the daily index capital gain or loss for day $t - 1$

...

▶ Back

Choice of a correct functional form

- ▶ The functional form has to be correctly specified in order to avoid biased and inconsistent estimator
 - ▶ remember that one of the OLS assumptions is that the model is correctly specified
- ▶ Ideally: the specification is given by underlying theory of the equation
- ▶ In reality: underlying theory does not give precise functional form
- ▶ In most cases, either linear form is adequate, or common sense will point out an easy choice from among the alternatives

Choice of correct functional form

- ▶ Nonlinearity of explanatory variables
 - ▶ often approximated by polynomial form
 - ▶ missing higher powers of a variable can be detected as omitted variables (will be a topic of lecture #7)
- ▶ Nonlinearity of dependent variable
 - ▶ harder to detect based on statistical fit of the regression
 - ▶ R^2 is incomparable across models where the y is transformed
 - ▶ dependent variables are often transformed to log-form in order to make their distribution closer to the normal distribution

Dummy variables

- ▶ **Dummy variable:** takes on the values of 0 or 1, depending on a qualitative attribute
- ▶ Examples of dummy variables:

$$Male = \begin{cases} 1 & \text{if the person is male} \\ 0 & \text{if the person is female} \end{cases}$$

$$Weekend = \begin{cases} 1 & \text{if the day is on weekend} \\ 0 & \text{if the day is a work day} \end{cases}$$

$$NewStadium = \begin{cases} 1 & \text{if the team plays on new stadium} \\ 0 & \text{if the team plays on old stadium} \end{cases}$$

Intercept dummy

- ▶ Dummy variable included in a regression alone (not interacted with other variables) is an intercept dummy
- ▶ It changes the intercept for the subset of data defined by a dummy variable condition:

$$y_i = \beta_0 + \beta_1 D_i + \beta_2 x_i + \varepsilon_i,$$

where:

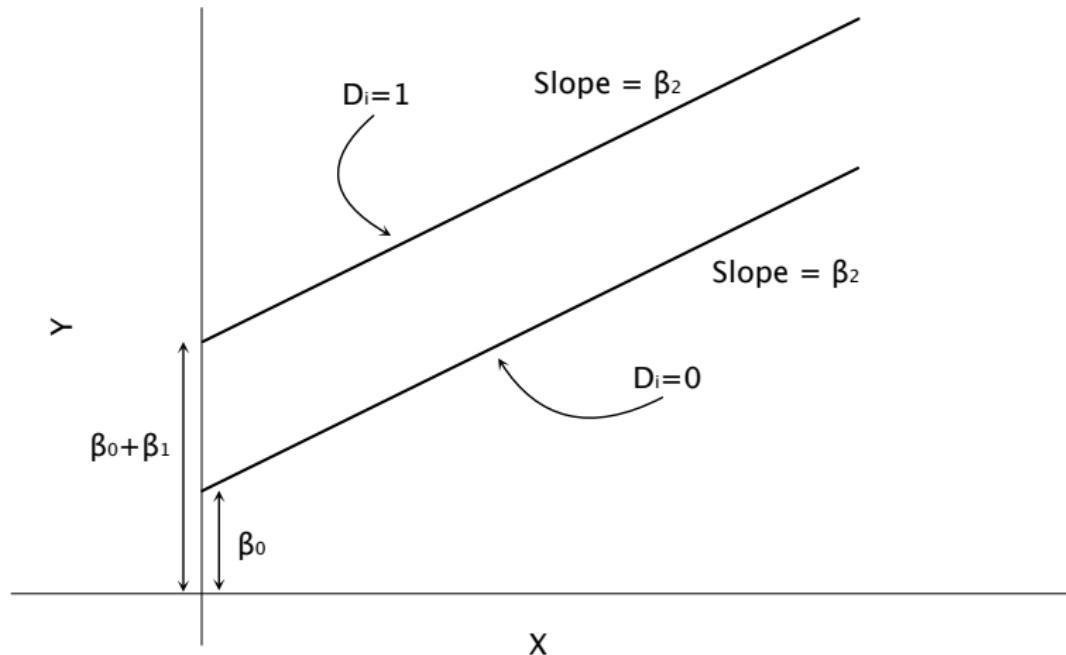
$$D_i = \begin{cases} 1 & \text{if the } i\text{-th observation meets a particular condition} \\ 0 & \text{otherwise} \end{cases}$$

- ▶ We have:

$$y_i = (\beta_0 + \beta_1) + \beta_2 x_i + \varepsilon_i \quad \text{if } D_i = 1$$

$$y_i = \beta_0 + \beta_2 x_i + \varepsilon_i \quad \text{if } D_i = 0$$

Intercept dummy



Example: Wage equation revisited

- ▶ Estimating the determinants of hourly wage (in USD):

$$\widehat{\ln \text{wage}_i} = 0.053 + 0.337 M_i + 0.084 \text{ educ}_i \\ (0.036) \qquad (0.007) \\ + 0.039 \text{ exper}_i - 0.0007 \text{ exper}_i^2, \\ (0.005) \qquad (0.0001)$$

where:

$$M_i = \begin{cases} 1 & \text{if the } i\text{-th person is male} \\ 0 & \text{if the } i\text{-th person is female} \end{cases}$$

- ▶ Compared to previous incomplete versions (L#4–6, S#5), this is an example of a well functionally specified ‘wage equation’
- ▶ Interpretation: men earn on average by 33.7% per hour more than women, ceteris paribus

Slope dummy

- ▶ If a dummy variable is interacted with another variable (x), it is a slope dummy
- ▶ It changes the relationship between x and y for a subset of data defined by a dummy variable condition:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 (x_i \cdot D_i) + \varepsilon_i,$$

where:

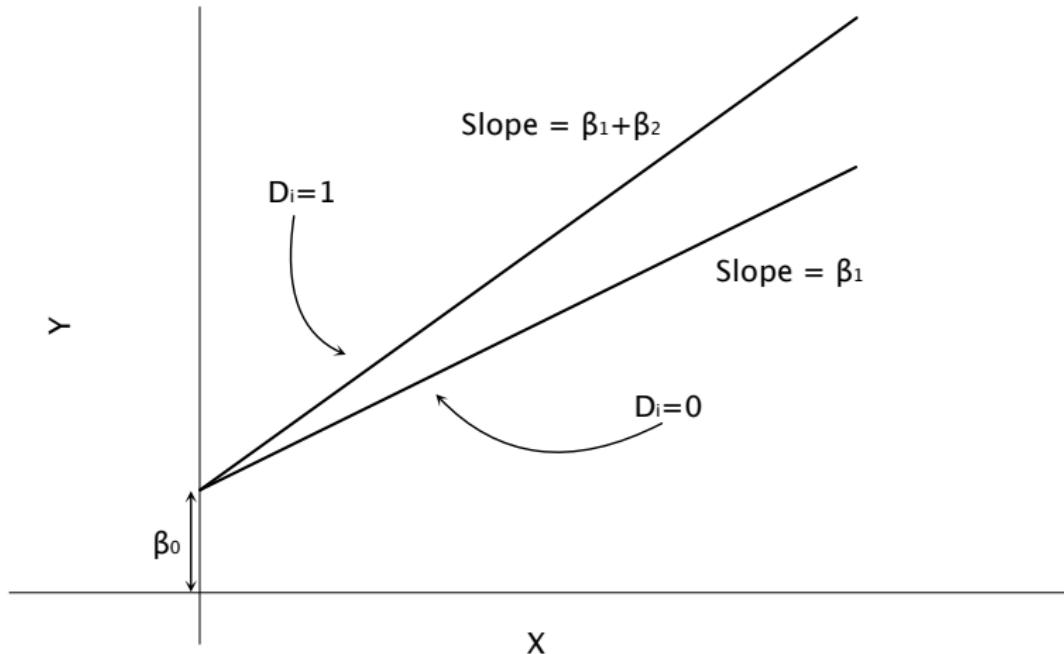
$$D_i = \begin{cases} 1 & \text{if the } i\text{-th observation meets a particular condition} \\ 0 & \text{otherwise} \end{cases}$$

- ▶ We have:

$$y_i = \beta_0 + (\beta_1 + \beta_2)x_i + \varepsilon_i \quad \text{if } D_i = 1$$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{if } D_i = 0$$

Slope dummy



Example

- ▶ Estimating the determinants of hourly wage (in USD):

$$\widehat{\ln \text{wage}_i} = 0.249 + 0.069 \text{ educ}_i + 0.026 (\text{educ}_i \cdot M_i) \\ (0.007) \qquad \qquad \qquad (0.003) \\ + 0.039 \text{ exper}_i - 0.0007 \text{ exper}_i^2, \\ (0.005) \qquad \qquad \qquad (0.0001)$$

where:

$$M_i = \begin{cases} 1 & \text{if the } i\text{-th person is male} \\ 0 & \text{if the } i\text{-th person is female} \end{cases}$$

- ▶ Interpretation: men gain on average by 2.6% per hour more than women for each additional year of education, ceteris paribus

Slope and intercept dummies

- ▶ Allow both for different slope and intercept for two subsets of data distinguished by a qualitative condition:

$$y_i = \beta_0 + \beta_1 D_i + \beta_2 x_i + \beta_3 (x_i \cdot D_i) + \varepsilon_i,$$

where:

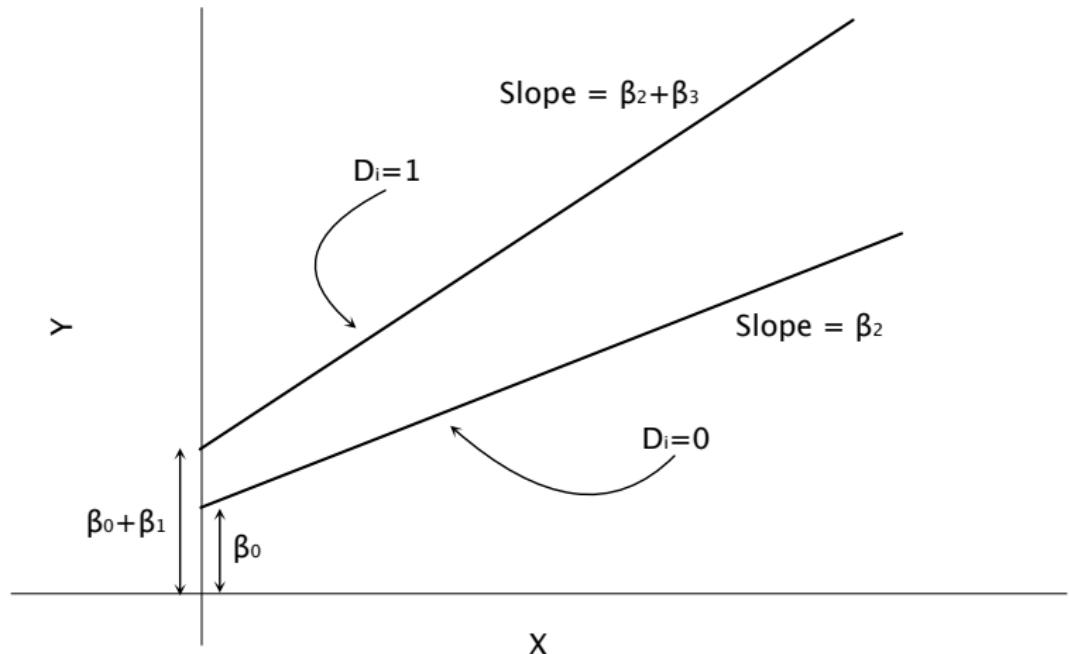
$$D_i = \begin{cases} 1 & \text{if the } i\text{-th observation meets a particular condition} \\ 0 & \text{otherwise} \end{cases}$$

- ▶ We have:

$$y_i = (\beta_0 + \beta_1) + (\beta_2 + \beta_3)x_i + \varepsilon_i \quad \text{if } D_i = 1$$

$$y_i = \beta_0 + \beta_2 x_i + \varepsilon_i \quad \text{if } D_i = 0$$

Slope and intercept dummies



Dummy variables: Extension

- ▶ What if a variable defines three or more qualitative attributes?
- ▶ Example: level of education (elementary school, high school, and college)
- ▶ Define and use a set of dummy variables:

$$H = \begin{cases} 1 & \text{if high school} \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad C = \begin{cases} 1 & \text{if college} \\ 0 & \text{otherwise} \end{cases}$$

- ▶ Should we include also a third dummy in the regression, which is equal to 1 for people with elementary education?
 - ▶ no, unless we exclude the intercept!
 - ▶ using a full set of dummies leads to perfect multicollinearity ('Dummy variable trap', will be a topic of lecture #7)

Stability of coefficients

- ▶ Suppose we have two subsamples (sub-periods) and we suspect that coefficients differ for each of them:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{for the first subsample}$$

$$y_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_i + \varepsilon_i \quad \text{for the second subsample}$$

- ▶ We want to test if $\beta_0 = \tilde{\beta}_0$ and $\beta_1 = \tilde{\beta}_1$
- ▶ Define a dummy:

$$D_i = \begin{cases} 1 & \text{if the } i\text{-th observation is from the second subsample} \\ 0 & \text{if the } i\text{-th observation is from the first subsample} \end{cases}$$

- ▶ Redefine the model:

$$y_i = \beta_0 + \alpha_0 D_i + \beta_1 x_i + \alpha_1 (x_i \cdot D_i) + \varepsilon_i$$

Testing for the stability of coefficients

- This gives:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{for the first subsample}$$

$$y_i = (\beta_0 + \alpha_0) + (\beta_1 + \alpha_1)x_i + \varepsilon_i \quad \text{for the second subsample}$$

- We test:

$$H_0 : \alpha_0 = 0 \quad \& \quad \alpha_1 = 0 \quad \text{vs.} \quad H_A : \alpha_0 \neq 0 \quad \text{or} \quad \alpha_1 \neq 0$$

in the model:

$$y_i = \beta_0 + \alpha_0 D_i + \beta_1 x_i + \alpha_1 (x_i \cdot D_i) + \varepsilon_i$$

using the F -test and the pooled data (both subsamples together)

Testing for the stability of coefficients

- Unrestricted model:

$$y_i = \beta_0 + \alpha_0 D_i + \beta_1 x_i + \alpha_1 (x_i \cdot D_i) + \varepsilon_i$$

- Restricted model:

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

- F -test:

$$F = \frac{(RSS_R - RSS_U)/J}{RSS_U/(n - k - 1)} \sim F_{J, n-k-1},$$

where $J = 2$, $k = 3$ in this case

Chow test of the stability of coefficients

- ▶ It can be shown that if we run the models separately over the two subsamples as:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{for the first subsample} \quad (1)$$

$$y_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_i + \varepsilon_i \quad \text{for the second subsample} \quad (2)$$

or together for pooled data as:

$$y_i = \beta_0 + \alpha_0 D_i + \beta_1 x_i + \alpha_1 (x_i \cdot D_i) + \varepsilon, \quad (3)$$

it holds for the sums of squared residuals:

$$RSS_{(1)} + RSS_{(2)} = RSS_{(3)}$$

- ▶ Note that $RSS_{(3)} = RSS_U$ in the previous notation

Chow test of the stability of coefficients

- ▶ Note that if we run the model:

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

on the pooled data (both subsamples together) and we denote RSS_P the residual sum of squares in this pooled model, we obtain $RSS_P = RSS_R$ in the previous notation

- ▶ The F -test now becomes:

$$F = \frac{(RSS_P - (RSS_{(1)} + RSS_{(2)})) / 2}{(RSS_{(1)} + RSS_{(2)}) / (n - 4)} \sim F_{2,n-4}$$

- ▶ Under this form, the test is known as the *Chow test of the stability of coefficients* (can be generalized for models with more coefficients and also for more potential subsamples)

Chow test for two subsamples: Steps

1. Run identically specified regressions on the two subsamples and save the residual sums of squares from the two ($RSS_{(1)}$ and $RSS_{(2)}$)
2. Pool the data from the two subsamples, run identically specified regression on the combined sample, and save the residual sum of squares (RSS_P) from this one
3. Calculate the F -statistic:

$$F = \frac{(RSS_P - (RSS_{(1)} + RSS_{(2)})) / (k + 1)}{(RSS_{(1)} + RSS_{(2)}) / (n_1 + n_2 - 2(k + 1))} \sim F_{k+1, n_1+n_2-2(k+1)},$$

where k is the number of explanatory/independent variables in each regression, n_1 and n_2 is the number of observations in the first and the second subsample, respectively

4. If $F > F_{k+1, n_1+n_2-2(k+1)}$ at a given significance level, reject the null hypothesis that the coefficients are the same over the two subsamples

Summary

- ▶ We discussed different nonlinear specifications of a regression equation and their interpretation
- ▶ We defined the concept of a dummy variable and we have shown its use
- ▶ We derived the Chow test of the stability of coefficients

Seminars and the next lecture #7

- ▶ **Seminars:**
 - ▶ use and interpretation of nonlinear specifications
 - ▶ dummy variables and the Chow test
- ▶ **Next lecture:**
 - ▶ omitted and irrelevant variables
 - ▶ selection of explanatory variables
- ▶ **Practical information:**
 - ▶ home assignment #2 will be assigned today (see SIS)
 - ▶ deadline: Thursday, November 11, 2021, 23:59:59
- ▶ **Readings for lecture #7:**
 - ▶ Studenmund (2016 & 17, 2014): Chapter 6 + Appendix: RESET only
 - ▶ Wooldridge (2016, 2012): Chapters 3-3, 3-4intro+b, 9-1a, (9-1b, 9-2)

Appendix: Precise interpretation (not mandatory)

For the **log-linear functional form**, due to the logarithmic nature of the dependent variable, the mathematically precise interpretation of the estimated coefficient of the non-logarithmized independent variable follows: if x_k increases by **1 unit**, then y will change by:

$$(\exp(\hat{\beta}_k) - 1) \cdot 100 \text{ percent},$$

ceteris paribus. This is often approximated by $\hat{\beta}_k \cdot 100$ percent. However, such simplification holds well only for ‘very small’ units of x_k .

▶ Back

LECTURE #7

Introductory Econometrics

CHOOSING EXPLANATORY VARIABLES

Jiri Kukacka, Ph.D.

Institute of Economic Studies, Faculty of Social Sciences, Charles University

Winter semester 2021, November 10

What we have learnt so far...

- ▶ We know what a linear regression model is and how its parameters are estimated by OLS
- ▶ We know what the properties of the OLS estimator are
- ▶ We know how to test single and multiple hypotheses in linear regression models
- ▶ We know how to assess the goodness of fit using R^2
- ▶ We started to talk about the specification of a regression equation

Specification of a regression equation

- ▶ **Specification** consists of choosing:
 1. correct functional form
 2. correct set of independent variables
 3. correct form of the stochastic error term
- ▶ Specification mistake \Rightarrow **specification error**
- ▶ We discussed the choice of the functional form in the previous lecture #6
- ▶ We will discuss the choice of independent variables today
- ▶ We will study the form of the error term in the next two lectures

Revision of nonlinear functional forms

1. Log-log form:

$1\% \text{ increase in } x \text{ implies } \beta\% \text{ change in } y$

$$\widehat{\ln \text{production_sugar}} = 2.70 + 0.59 \ln \text{labor} + 0.33 \ln \text{capital}$$

2. Log-linear form:

$1\text{-unit increase in } x \text{ implies } (\beta \cdot 100)\% \text{ change in } y$

$$\widehat{\ln \text{wage}} \quad = 0.217 + 0.098 \text{ education}_{(\text{in years})} + 0.010 \text{ experience}_{(\text{in years})}$$

Revision of function forms

3. Linear-log form:

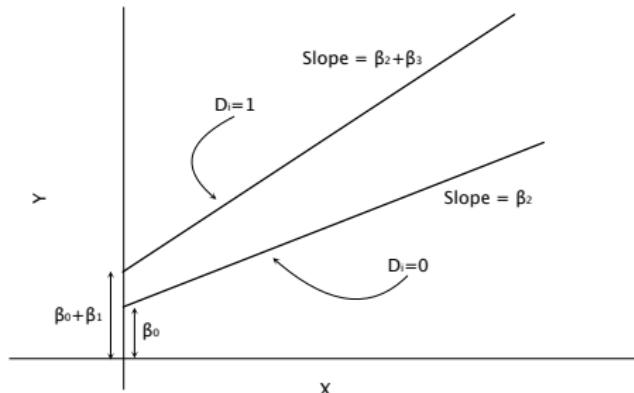
1% increase in x implies $(\beta/100)$ -unit change in y

4. Polynomial form:

$$\widehat{\text{grade}} = 30 + 1.4\text{hours} - 0.009\text{hours}^2$$

In the previous lecture #6, we also discussed...

- ▶ Dummy variables: intercept and slope dummy



- ▶ Chow test of the stability of coefficients:

$$F = \frac{(RSS_P - (RSS_{(1)} + RSS_{(2)})) / (k + 1)}{(RSS_{(1)} + RSS_{(2)}) (n_1 + n_2 - 2(k + 1))} \sim F_{k+1, n_1 + n_2 - 2(k+1)}$$

In today's lecture, we will...

- ▶ Focus on the choice of independent variables
- ▶ Learn that:
 - ▶ omitting a relevant variable from an equation is likely to bias the OLS estimator of remaining coefficients
 - ▶ including an irrelevant variable in an equation leads to higher variance of the OLS estimator
 - ▶ our choice should be led by the economic theory and confirmed by a set of statistical tools
- ▶ Readings for this week:
 - ▶ Studenmund (2016 & 17, 2014): Chapter 6 + Appendix: RESET only
 - ▶ Wooldridge (2016, 2012): Chapters 3-3, 3-4intro+b, 9-1a, (9-1b, 9-2)

Omitted variables

- ▶ We omit a variable when we:
 - ▶ forget to include it
 - ▶ do not have data for it
- ▶ This misspecification results in:
 - ▶ not having the coefficient for this variable
 - ▶ biasing the OLS estimator of the coefficients of other variables in the equation ⇒ **omitted variable bias**

Omitted variables

- ▶ Where does the omitted variable bias come from?
- ▶ Assume true model:

$$y_i = \beta_0 + \beta_1 x_i + \gamma z_i + \varepsilon_i$$

- ▶ When we omit variable z , the model becomes:

$$y_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_i + \tilde{\varepsilon}_i,$$

implying:

$$\tilde{\varepsilon}_i = \gamma z_i + \varepsilon_i$$

- ▶ We assume that $\text{Cov}(x_i, \varepsilon_i) = 0$, but it does not have to be true that $\text{Cov}(x_i, \tilde{\varepsilon}_i) = 0$
- ▶ In general: $\text{Cov}(x_i, \tilde{\varepsilon}_i) = \gamma \cdot \text{Cov}(x_i, z_i)$
- ▶ CA 3. is violated \Rightarrow **biased and inconsistent OLS estimator**

Omitted variables

- ▶ For the model with omitted variable:

$$E(\tilde{\beta}_1) = \beta_1 + \text{bias}$$

$$\text{bias} = \gamma \cdot \alpha_1$$

- ▶ coefficients β_1 and γ are from the true model:

$$y_i = \beta_0 + \beta_1 x_i + \gamma z_i + \varepsilon_i$$

- ▶ coefficient α_1 is from a regression of z on x , i.e.:

$$z_i = \alpha_0 + \alpha_1 x_i + u_i$$

- ▶ The bias is zero if $\gamma = 0$ or $\alpha_1 = 0$ (not likely to happen)

Omitted variables

- ▶ Intuitive explanation:
 - ▶ if we leave out an important variable ($\gamma \neq 0$) from the regression, the OLS estimator of coefficients of other variables is biased unless the omitted variable is uncorrelated with all included independent variables (i.e. unless $\alpha_1 = 0$)
 - ▶ this is because the included variables pick up some of the effect of the omitted variable (to the extent they are correlated) and the coefficients of included variables thus change
- ▶ Example: what would happen if you estimated a production function with capital only and omitted labor?

Omitted variables: Example

- ▶ Assume for now that this is the true model estimating demand for chicken meat in the US:

$$\hat{Q}_{chicken,t} = 27.7 - \frac{0.11}{(0.03)} P_t + \frac{0.03}{(0.017)} P_{b,t} + \frac{0.23}{(0.01)} Yd_t$$

$$R^2 = 0.9914 \quad R_{adj}^2 = 0.9904 \quad n = 29 \text{ (1974 – 2002)}$$

$Q_{chicken,t}$... per capita demand for chicken (in pounds)

P_t ... price of chicken (in USD cents)

$P_{b,t}$... price of beef (in USD cents)

Yd_t ... per capita disposable income (in USD hundreds)

Omitted variables: Example

- When we omit price of beef:

$$\hat{Q}_{chicken,t} = 30.7 - \frac{0.09}{(0.03)} P_t + \frac{0.25}{(0.005)} Yd_t$$

$$R^2 = 0.9902 \quad R_{adj}^2 = .9895 \quad n = 29$$

- Compare to the true model:

$$\hat{Q}_{chicken,t} = 27.7 - \frac{0.11}{(0.03)} P_t + \frac{0.03}{(0.017)} P_{b,t} + \frac{0.23}{(0.01)} Yd_t$$

$$R^2 = 0.9914 \quad R_{adj}^2 = 0.9904 \quad n = 29$$

- We observe a positive shift in the estimated coefficients of P (as well as of Yd). But was it expected?

Omitted variables: Direction of the bias

- ▶ Determining the bias direction: $bias = \gamma \cdot \alpha_1$
 - ▶ where γ has the same sign as correlation between the omitted variable and the dependent variable (price of beef and demand for chicken)
 - ▶ γ is likely to be positive
 - ▶ where α_1 has the same sign as correlation between the omitted variable and the included independent variable (price of beef and price of chicken)
 - ▶ α_1 is likely to be positive
- ▶ Conclusion: bias of the OLS estimator of the coefficient of price of chicken (P) is likely to be positive if we omit price of beef (P_b) from the equation

Omitted variables

- ▶ In reality, we usually do not have the true model to compare with, because:
 - ▶ we do not know what the true model is
 - ▶ we do not have data for some important variable
- ▶ We can often recognize symptoms of the bias if we obtain some unexpected results
- ▶ We can prevent omitting variables by relying on the theory
- ▶ If we cannot prevent omitting variables, we can at least determine in what direction this biases our OLS estimator
- ▶ For key but unobserved explanatory variables we can use a proxy (*IQ* as a proxy for *ability* in the wage equation)

Irrelevant variables

- ▶ A second type of specification error is including a variable that does not belong to the (true) model
- ▶ This misspecification:
 - ▶ does not cause bias
 - ▶ but it increases the standard errors of the estimated coefficients of the included variables

Irrelevant variables

- ▶ True model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (1)$$

- ▶ Model as it looks when we add irrelevant z :

$$y_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_i + \tilde{\gamma} z_i + \tilde{\varepsilon}_i \quad (2)$$

- ▶ We can represent the error term as $\tilde{\varepsilon}_i = \varepsilon_i - \tilde{\gamma} z_i$
- ▶ But since from the true model $\tilde{\gamma} = 0$, we have $\tilde{\varepsilon}_i = \varepsilon_i$ and there is no bias

Irrelevant variables

- ▶ However, irrelevant variables cause problems with standard errors of the estimated coefficients:
 - ▶ in the model with an irrelevant variable, the variance of the estimator of the coefficient of the other explanatory variable is:

$$\text{Var} \left[\hat{\beta}_1^{(2)} \right] = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2} \cdot \frac{1}{(1 - \text{Corr}_{x,z}^2)},$$

where $\text{Corr}_{x,z}$ is the correlation coefficient between the other explanatory variable and the irrelevant variable

- ▶ and this variance is generally bigger than variance in the true simple linear regression model:

$$\text{Var} \left[\hat{\beta}_1^{(1)} \right] = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}$$

- ▶ **Irrelevant variables increase variance of the OLS estimator of coefficients of other independent variables**

Irrelevant variables

- True model:

$$\hat{Q}_{chicken,t} = 27.7 - \frac{0.11}{(0.03)} P_t + \frac{0.03}{(0.017)} P_{b,t} + \frac{0.23}{(0.01)} Yd_t$$

$$R^2 = 0.9914 \quad R_{adj}^2 = 0.9904 \quad n = 29$$

- If we include the average annual change in temperature $TEMP$ (an irrelevant variable):

$$\hat{Q}_{chick,t} = 26.9 - \frac{0.11}{(0.03)} P_t + \frac{0.04}{(0.018)} P_{b,t} + \frac{0.23}{(0.015)} Yd_t - \frac{0.02}{(0.02)} TEMP_t$$

$$R^2 = 0.9916 \quad R_{adj}^2 = 0.9903 \quad n = 29$$

- We observe that $TEMP_t$ is statistically insignificant and standard errors of some other explanatory variables increase

Summary of the theory

- ▶ Potential dilemma: leaving a relevant variable out of a model is likely to bias the OLS estimator of remaining parameters, but including an irrelevant variable leads to its higher variance
- ▶ Bias-efficiency trade-off:

	Omitted variable	Irrelevant variable	
Bias	Yes*	No	
Variance	Decreases*	Increases*	

*Unless no correlation between explanatory x and omitted/irrelevant z

Summary of the example

- True model:

$$\hat{Q}_{chicken,t} = 27.7 - \frac{0.11}{(0.03)} P_t + \frac{0.03}{(0.017)} P_{b,t} + \frac{0.23}{(0.01)} Yd_t$$

$$R^2 = 0.9914 \quad R_{adj}^2 = 0.9904 \quad n = 29$$

- Model with omitted variable:

$$\hat{Q}_{chicken,t} = 30.7 - \frac{0.09}{(0.03)} P_t + \frac{0.25}{(0.005)} Yd_t$$

$$R^2 = 0.9902 \quad R_{adj}^2 = .9895 \quad n = 29$$

- Model with irrelevant variable:

$$\hat{Q}_{chick,t} = 26.9 - \frac{0.11}{(0.03)} P_t + \frac{0.04}{(0.018)} P_{b,t} + \frac{0.23}{(0.015)} Yd_t - \frac{0.02}{(0.02)} TEMP_t$$

$$R^2 = 0.9916 \quad R_{adj}^2 = 0.9903 \quad n = 29$$

Four important selection criteria

Does an explanatory variable belong to the model?

1. *Theory*: Is the variable's place in the equation unambiguous and theoretically sound? Does intuition suggest you it should be included?
2. *t-test*: Is the variable's estimated coefficient significant in the expected direction? (*F-test* can help us when thinking of excluding multiple variables.)
3. R_{adj}^2 : Does the overall fit of the equation improve (enough) when the variable is added to the model?
4. *Bias*: Do estimated coefficients of other variables change considerably when the variable is added to the model?

Four important selection criteria

- ▶ If all conditions hold, the variable belongs to the model equation
- ▶ If none of them holds, the variable is irrelevant and can be safely excluded
- ▶ If the criteria give contradictory answers, most importance should be attributed to theoretical justification
- ▶ Therefore, if theory (or intuition) says that a variable belongs to the model equation, we include it (even though its coefficient might be insignificant!)

Estimating demand for Brazilian coffee

$$\hat{Q}_{coffee} = 9.1 + \frac{7.8}{(15.6)} P + \frac{2.4}{(1.2)} P_{tea} + \frac{0.0035}{(0.0010)} Yd$$

$$t = 0.5 \quad 2.0 \quad 3.5$$

$$R^2_{adj} = 0.60 \quad n = 25$$

- Q_{coffee} ... demand for Brazilian coffee (in the US)
 P ... price of Brazilian coffee
 P_{tea} ... price of tea
 Yd ... disposable income (in the US)

Estimating demand for Brazilian coffee

- ▶ Should we include price of Brazilian coffee into the equation?

$$\begin{aligned}\hat{Q}_{coffee} &= 9.3 & + 2.6 P_{tea} &+ 0.0036 Yd \\ && (1.0) & (0.0009) \\ t &= & 2.6 & 4.0 \\ R^2_{adj} &= 0.61 & n &= 25\end{aligned}$$

$$\begin{aligned}\hat{Q}_{coffee} &= 9.1 + 7.8 P + 2.4 P_{tea} + 0.0035 Yd \\ & (15.6) & (1.2) & (0.0010) \\ t &= 0.5 & 2.0 & 3.5 \\ R^2_{adj} &= 0.60 & n &= 25\end{aligned}$$

- ▶ The three criteria do not hold (and theory seems inconclusive)
⇒ price of Brazilian coffee does not belong to the model
(perhaps Brazilian coffee is price inelastic)

Estimating demand for Brazilian coffee

- ▶ Really???
- ▶ What if we add price of Colombian coffee (P_{cc})?

$$\begin{aligned}\hat{Q}_{coffee} &= 10.0 + \frac{8.0}{(4.0)} P_{cc} - \frac{5.6}{(2.0)} P + \frac{2.6}{(1.3)} P_{tea} + \frac{0.0030}{(0.0010)} Yd \\ t &= 2.0 \quad -2.8 \quad 2.0 \quad 3.0 \\ R^2_{adj} &= 0.65 \quad n = 25\end{aligned}$$

$$\begin{aligned}\hat{Q}_{coffee} &= 9.1 + \frac{7.8}{(15.6)} P + \frac{2.4}{(1.2)} P_{tea} + \frac{0.0035}{(0.0010)} Yd \\ t &= \quad 0.5 \quad 2.0 \quad 3.5 \\ R^2_{adj} &= 0.60 \quad n = 25\end{aligned}$$

- ▶ All four criteria hold \Rightarrow price of Brazilian coffee belongs to the model (Brazilian coffee is price elastic)!

The danger of specification searches

- ▶ “*If you torture the data long enough, it will confess.*”
(Ronald Coase)
- ▶ If too many specifications are tried:
 - ▶ the final result has desired properties only by chance
 - ▶ the statistical significance of the results is overestimated because the estimations of the previous regressions are ignored
⇒ ‘Multiple Testing Problem’*
- ▶ How to proceed:
 - ▶ keep the number of estimated regressions low
 - ▶ focus on theoretical considerations: keep the insignificant variables in the equation if the theory suggests they should be included
 - ▶ document all specifications investigated

* If you test 20 colours of jelly beans at the 5% level, how many null hypotheses do you expect to reject?

A formal specification test

- ▶ Ramsey's Regression Specification Error Test (RESET)
 - ▶ allows to detect general **functional form misspecification**
 - ▶ additionally: **might** also suggest that an important variable is omitted (but generally poor performance in this respect)
 - ▶ unfortunately it does not allow to detect the source of the misspecification
- ▶ There are two forms of this test, both based on similar intuition:
 - ▶ if the model is correctly specified, nothing is missing in the equation and thus no nonlinear functions of the explanatory variables should be significant when added to equation and the residuals are a white noise
- ▶ We will derive the test for the model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

RESET I

1. We run the regression: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$
2. We save the predicted values: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}$
3. We run an augmented regression:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \gamma_1 \hat{y}_i^2 + \gamma_2 \hat{y}_i^3 + \varepsilon_t$$

[more powers of \hat{y} can be included, these additional polynomial terms act as proxies for a correct functional form (or possibly an omitted variable)]

4. We test $H_0 : \gamma_1 = \gamma_2 = 0$ using a standard F -test
 5. If we reject H_0 , there is a misspecification problem in our model
- Intuition: if the model is correct, y is well explained by x_1 and x_2 and any nonlinear function of the predicted/fitted values of y (i.e. \hat{y} raised to higher powers) should not be significant

RESET II

1. We run the regression: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$
2. We save the predicted values: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}$
and the residuals: $e_i = y_i - \hat{y}_i$
3. We run the regression:

$$e_i = \alpha_0 + \alpha_1 \hat{y}_i + \alpha_2 \hat{y}_i^2 + \varepsilon_i$$

[more powers of \hat{y} can be included]

4. We test $H_0 : \alpha_1 = \alpha_2 = 0$ using a standard F -test
 5. If we reject H_0 , there is a misspecification problem in our model
- Intuition: if the model is correct, residuals should not display any pattern depending on the explanatory variables

Summary

- ▶ Omitted variable causes bias (and decreases variance) of the OLS estimator
 - ▶ direction of this bias can be predicted
- ▶ Included irrelevant variable increases variance (but does not cause bias)
 - ▶ such variable is insignificant in the regression
 - ▶ it does not contribute to the overall fit of the regression
- ▶ There is a set of selection criteria that help us to recognize correct specification (in terms of explanatory variables)
 - ▶ these criteria have to be applied with caution
 - ▶ theoretical justification has always priority over statistical properties

Seminars and the next lecture #8

- ▶ **Seminars:**

- ▶ a discussion on several models to decide if they are correctly specified
- ▶ omitted and irrelevant variables practice
- ▶ RESET test (in Gretl)

- ▶ **Next lecture:**

- ▶ we will study the issues of multicollinearity and heteroskedasticity
- ▶ we will start to talk about the form of the error term

- ▶ **Readings for lecture #8:**

- ▶ Studenmund (2016 & 17, [2014]): Chapters 8, 10 [Chapter 8 only]
- ▶ Wooldridge (2016, 2012): Chapters 3-4a, 8-1-8-4

LECTURE #8

Introductory Econometrics

MULTICOLLINEARITY & HETEROSKEDASTICITY

Jiri Kukacka, Ph.D.

Institute of Economic Studies, Faculty of Social Sciences, Charles University

Winter semester 2021, November 24

In previous lectures...

- ▶ We discussed the specification of a linear regression model
- ▶ **Specification** consists of choosing:
 1. correct functional form
 2. correct independent variables
 3. correct form of the stochastic error term
- ▶ In lecture #6, we talked about the choice of a correct functional form:
 - ▶ what are the most common function forms?

Short revision of lecture #7

- ▶ We studied what happens if:
 - ▶ we omit a relevant variable
 - ▶ does omitting a relevant variable cause a bias?
 - ▶ we include an irrelevant variable
 - ▶ does including an irrelevant variable cause a bias?
- ▶ We defined the four selection criteria that determine if a variable belongs to the equation:
 - ▶ can you list some of these selection criteria?
- ▶ Finally, we have learnt a formal test: RESET

In today's lecture #8, we will...

- ▶ Finish the discussion of the choice of independent variables by talking about **multicollinearity**
- ▶ Start the discussion of the correct form of the error term by talking about **heteroskedasticity**
- ▶ For both of these issues, we will learn:
 - ▶ what is the nature of the problem
 - ▶ what are its consequences
 - ▶ how it is diagnosed
 - ▶ what are the remedies available
- ▶ Readings for this week:
 - ▶ Studenmund (2016 & 17, [2014]): Chapters 8, 10 [Chapter 8 only]
 - ▶ Wooldridge (2016, 2012): Chapters 3-4a, 8-1-8-4

Multicollinearity

Perfect multicollinearity

- ▶ Some explanatory variable is a perfect linear function of one or more other explanatory variables
- ▶ Violation of the Classical Assumptions 6.
- ▶ OLS estimation cannot be conducted:
 - ▶ intuitively: the estimator cannot distinguish which of the explanatory variables causes the change of the dependent variable if they move together
 - ▶ technically: the matrix $\mathbf{X}'\mathbf{X}$ is singular (noninvertible)
- ▶ Rare and easy to detect

Example of perfect multicollinearity

Dummy variable trap:

- ▶ Inclusion of a dummy variable for each category in the model with intercept
- ▶ Example: wage equation for sample of individuals who have high-school education or higher:

$$\ln \text{wage}_i = \beta_0 + \beta_1 \text{high_school}_i + \beta_2 \text{university}_i + \beta_3 \text{phd}_i + \varepsilon_i$$

- ▶ Automatically detected by most statistical softwares

Imperfect multicollinearity

- ▶ Two or more explanatory variables are highly correlated in the particular data set
- ▶ OLS estimator can be found, but it may be very imprecise:
 - ▶ intuitively: the estimator can hardly distinguish the effects of the explanatory variables if they are highly correlated
 - ▶ technically: the matrix $\mathbf{X}'\mathbf{X}$ is nearly singular which causes the variance of the OLS estimator $\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$ to be very large
- ▶ Usually referred as to ‘multicollinearity’ for simplicity

Consequences of multicollinearity

1. OLS estimator remains unbiased, consistent, and keeps BLUE
2. Large standard errors of estimated coefficients:
 - ▶ t -statistics are smaller... relevant variables may be insignificant
 - ▶ confidence intervals are very large... estimates are less reliable
 - ▶ contradiction between t and F -statistics (jointly significant + high R^2)
3. Estimates often very sensitive to changes in specification
4. Overall fit of the equation and estimated coefficients of non-multicollinear variables are usually unaffected

Detection of multicollinearity

- ▶ Some multicollinearity exists in every model, the aim is to recognize when it causes a severe problem
- ▶ Multicollinearity can be signalled by the underlying theory, but it is very sample dependent
- ▶ We judge the severity of multicollinearity based on the properties of our sample and on the results we obtain:
 - ▶ a simple method: examine **simple correlation coefficients** between explanatory variables ($> 75\text{--}80\%$, say)
 - ▶ **Variance Inflation Factor** for each β : $VIF(\hat{\beta}_j) = \frac{1}{1-R_j^2} > 5$ (or 10 or 2.5), where R_j^2 is R^2 from regressing x_j on all other x
 - ▶ If some of them is too high, we may suspect that the coefficients of these variables are affected by multicollinearity
 - ▶ Problem: neither of above is necessary, only sufficient (low scores do not mean no MC), also *How high is too high?*

Remedies for multicollinearity

- ▶ Do nothing:
 - ▶ when multicollinearity does not cause insignificance or unreliable estimated coefficients
 - ▶ keep in mind that deletion of multicollinear variable that belongs to the equation will cause a specification bias
- ▶ Drop a redundant variable:
 - ▶ when the variable is not needed to represent the effect on the dependent variable
 - ▶ in case of severe multicollinearity, it makes no statistical difference which variable is dropped
 - ▶ theoretical underpinnings of the model should be the basis for such a decision

Remedies for multicollinearity

- ▶ Transform the multicollinear variables
 - ▶ in case when all variables are extremely important on theoretical grounds
 - ▶ possible transformations:
 1. combination of multicollinear variables
 2. first differences (for time series)
 3. dummy variables (e.g. a dummy for urban highway miles above certain threshold in the following example)
- ▶ Increase the size of the sample
 - ▶ the confidence intervals are narrower when we have more observations

Example

- ▶ Estimating the demand for gasoline in the US:

$$\widehat{Q}_{petrol,i} = 389.6 - \frac{36.5}{(13.2)} TAX_i + \frac{60.8}{(10.3)} UHM_i - \frac{0.061}{(0.043)} REG_i$$

$$t = -2.77 \qquad \qquad \qquad 5.92 \qquad \qquad \qquad -1.43$$

$$R_{adj}^2 = 0.919 \qquad n = 50 \qquad \text{Corr}(UHM, REG) = 0.978$$

$Q_{petrol,i}$... petroleum consumption in the i -th state

TAX_i ... the gasoline tax rate in the i -th state

UHM_i ... urban highway miles within the i -th state

REG_i ... motor vehicle registrations in the i -th state

Example

- ▶ We suspect a multicollinearity between urban highway miles (*UHM*) and motor vehicle registration (*REG*) across states, because those states that have a lot of highways might also have a lot of motor vehicles. Therefore, we might run into multicollinearity problems.
- ▶ How we detect multicollinearity:
 - ▶ look at the correlation coefficient. It is indeed huge (0.978).
 - ▶ look at the coefficients of the two variables. Are they both individually significant? *UHM* is significant, but *REG* is not. This further suggests a presence of multicollinearity.
- ▶ Remedy: try dropping one of the correlated variables

Example

$$\hat{Q}_{petrol,i} = 551.7 - 53.6 \text{ } TAX_i + 0.186 \text{ } REG_i$$
$$(16.9) \qquad \qquad (0.012)$$
$$t = -3.18 \qquad \qquad 15.88$$

$$R_{adj}^2 = 0.861 \qquad n = 50$$

$$\hat{Q}_{petrol,i} = 410.0 - 39.6 \text{ } TAX_i + 46.4 \text{ } UHM_i$$
$$(13.1) \qquad \qquad (2.16)$$
$$t = -3.02 \qquad \qquad 21.40$$

$$R_{adj}^2 = 0.916 \qquad n = 50$$

Example

- ▶ Estimating the demand equation for fish meat:

$$\widehat{Q}_{fish,t} = 7.96 + 0.03 P_t + 0.0047 P_{B,t} + 0.36 \ln Yd_t$$

(0.03) (0.019) (1.15)

$$t = \quad \quad \quad 0.98 \quad \quad \quad 0.24 \quad \quad \quad 0.31$$

$$R^2_{adj} = 0.667 \quad n = 25$$

- $Q_{fish,t}$... demand for fish
 P_t ... price of fish
 $P_{B,t}$... price of beef
 Yd_t ... disposable income

Example

- ▶ Let us transform: $RP_t = \frac{P_t}{P_{B,t}}$

- ▶ We obtain:

$$\widehat{Q}_{fish,t} = -5.17 - 1.93 RP_t + 2.71 \ln Yd_t$$

(1.43) (0.66)

$$t = \quad \quad \quad -1.35 \quad \quad \quad 4.13$$

$$R^2_{adj} = 0.588 \quad \quad \quad n = 25$$

- ▶ This is better, but still not perfect
- ▶ We need to increase the number of observations

Heteroskedasticity

(Pure) heteroskedasticity

- ▶ Observations of the error term are drawn from a distribution that has no longer a constant variance:

$$\text{Var}(\varepsilon_i) = \sigma_i^2 \quad i = 1, 2, \dots, n$$

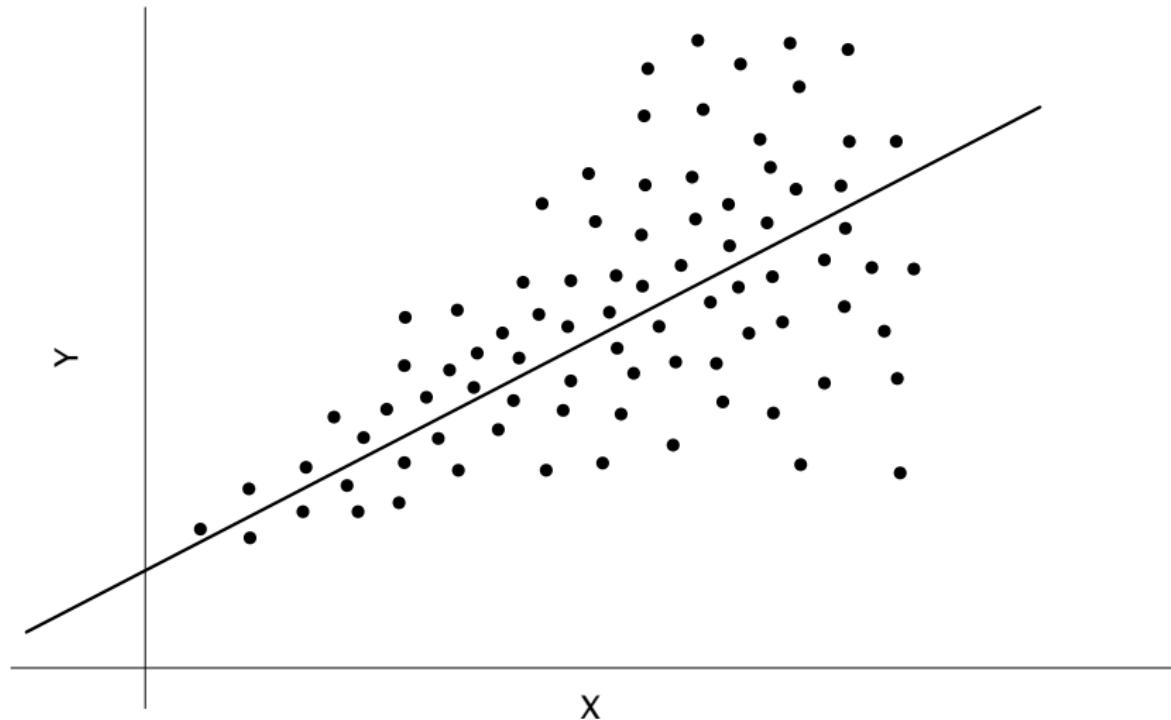
[note: constant variance means: $\text{Var}(\varepsilon_i) = \sigma^2 \quad i = 1, 2, \dots, n$]

- ▶ Often occurs in data sets in which there is a wide disparity between the largest and smallest observed values
 - ▶ smaller values often connected to smaller variance and larger values to larger variance (e.g. consumption of households based on their income level)
- ▶ One particular form of heteroskedasticity (variance of the error term is a function of some observable variable):

$$\text{Var}(\varepsilon_i) = h(x_i) \quad i = 1, 2, \dots, n$$

$$\text{Var}(\varepsilon_i) = \sigma^2 x_i^2 \quad \Rightarrow \quad \text{sd}(\varepsilon_i) = \sigma x_i$$

Heteroskedasticity



Impure heteroskedasticity

- ▶ Heteroskedasticity caused by a specification error in the equation:
 - ▶ mostly an omitted variable
- ▶ Intuition:
 - ▶ recall that the error term includes the omitted variables, nonlinearities, measurement errors, and the classical error term
 - ▶ if the omitted variable has a heteroskedastic component, the error term of the misspecified equation might be heteroskedastic even if the true error is not
- ▶ Impure heteroskedasticity can be corrected by better choice of specification (as opposed to pure heteroskedasticity)

Consequences of heteroskedasticity

1. OLS estimator remains unbiased and consistent
2. Variance of the $\hat{\beta}^{OLS}$ distribution increases:
 - ▶ because the heteroskedastic error term explains a larger proportion of fluctuations of the dependent variable
 - ▶ violation of the Classical Assumptions 5. \Rightarrow OLS not BLUE
 \Rightarrow OLS more likely to misestimate the true β
3. But OLS tends to underestimate the variance of $\hat{\beta}_k^{OLS}$
[as long as positive correlation between $(x_{ki} - \bar{x}_k)^2$ and σ_i^2]:
 - ▶ increase of the (true) variance is, however, masked by OLS because it assumes a homoskedastic error
 - ▶ OLS thus attributes the impact of the heteroskedastic error to the independent variables
 \Rightarrow t -statistics tend to be larger under heteroskedasticity, statistical inference becomes unreliable and incorrect

Detection of heteroskedasticity

- ▶ There is a battery of tests for heteroskedasticity
- ▶ We will derive two tests, both for the model:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i$$

- ▶ Both are based on an analysis of residuals:

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2})$$

Detection of heteroskedasticity

- ▶ OLS estimator is consistent even under heteroskedasticity
- ▶ This allows us to employ residuals that consistently estimate the stochastic error term
- ▶ The null hypothesis for the tests is 'no heteroskedasticity':

$$E(e^2) = \sigma^2$$

- ▶ therefore, we will analyse the relationship between e^2 and explanatory variables
- ▶ Sometimes, simple visual analysis of residuals is sufficient to detect heteroskedasticity

White test for heteroskedasticity

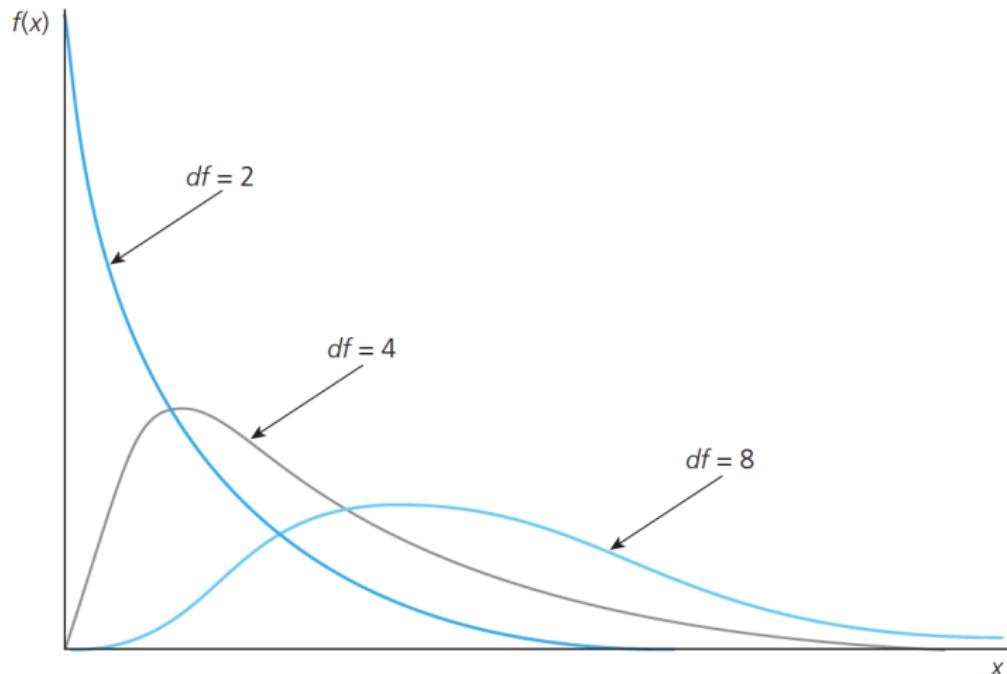
1. Estimate the model, get the residuals e_i ;
2. Regress the residuals squared on all explanatory variables and on squares and cross-products of all explanatory variables:

$$e_i^2 = \alpha_0 + \alpha_1 x_{i,1} + \alpha_2 x_{i,2} + \alpha_3 x_{i,1}^2 + \alpha_4 x_{i,2}^2 + \alpha_5 x_{i,1} x_{i,2} + v_i \quad (1)$$

3. Get the R^2 of this regression and the sample size n
4. Test the overall significance of (1):
 LM test statistic* $= nR^2 \sim \chi_k^2$,
where k is the number of slope coefficients in (1)
5. If nR^2 is larger than the χ_k^2 critical value, then we reject H_0 of 'no heteroskedasticity'

* F -statistic is valid only if the errors are normally distributed but with squared residuals this is not a 'safe' assumption—a χ^2 test is a proper procedure then

Chi-squared distribution with various dfs



Source: Wooldridge (2016, pg. 670)

Testing in Gretl

Model 1: OLS, using observations 1-526

Dependent variable: l_wage

	coefficient	std. error	t-ratio	p-value

const	0.127997	0.105932	1.208	0.2275
educ	0.0903658	0.00746800	12.10	6.98e-30 ***
exper	0.0410089	0.00519652	7.892	1.77e-14 ***
sq_exper	-0.000713558	0.000115764	-6.164	1.42e-09 ***
 Mean dependent var				
Mean dependent var	1.623268	S.D. dependent var	0.531538	
Sum squared resid	103.7904	S.E. of regression	0.445906	
R-squared	0.300273	Adjusted R-squared	0.296251	
F(3, 522)	74.66828	P-value(F)	3.38e-40	
Log-likelihood	-319.5317	Akaike criterion	647.0633	
Schwarz criterion	664.1245	Hannan-Quinn	653.7435	

Testing in Gretl

gretl: model 3

File Edit Tests Save Graphs Analysis LaTeX

Model 3:
Dependent

const
educ
exper
sq_expe

Mean depe
Sum squar
R-squared
F(3, 522)
Log-likel
Schwarz c

Log-likel

Omit variables
Add variables
Sum of coefficients
Linear restrictions
Non-linearity (squares)
Non-linearity (logs)
Ramsey's RESET
Heteroskedasticity
Normality of residual
Influential observations
Collinearity
Chow test
Autocorrelation
Durbin-Watson p-value

26

or	t-ratio	p-value
00	12.10	6.98e-30 ***
52	7.892	1.77e-14 ***
764	-6.164	1.42e-09 ***

White's test

White's test (squares only)
Breusch-Pagan
Koenker

The screenshot shows the Gretl software interface with the title bar "gretl: model 3". The menu bar includes File, Edit, Tests (which is currently selected and highlighted in blue), Save, Graphs, Analysis, and LaTeX. A toolbar icon for saving is also present. The main window displays a list of model statistics on the left and a detailed test output on the right. The test output for the "White's test" is shown in a expanded submenu, listing three options: "White's test (squares only)", "Breusch-Pagan", and "Koenker".

White test

White's test for heteroskedasticity

OLS, using observations 1-526

Dependent variable: uhat^2

	coefficient	std. error	t-ratio	p-value

const	0.309100	0.270676	1.142	0.2540
educ	-0.0271963	0.0349493	-0.7782	0.4368
exper	-0.0144795	0.0231748	-0.6248	0.5324
sq_exper	0.00110799	0.00151740	0.7302	0.4656
sq_educ	0.00100090	0.00120809	0.8285	0.4078
X2_X3	0.000879437	0.00140299	0.6268	0.5310
X2_X4	-1.32202e-05	3.02447e-05	-0.4371	0.6622
X3_X4	-2.62071e-05	5.00494e-05	-0.5236	0.6008
sq_sq_exper	1.66594e-07	5.29612e-07	0.3146	0.7532

Unadjusted R-squared = 0.044471

Test statistic: TR^2 = 23.391852,

with p-value = P(Chi-square(8) > 23.391852) = 0.002896

Breusch-Pagan test for heteroskedasticity

1. Estimate the model, get the residuals e_i
2. (Scale the squares of residuals: $e_{i,scaled}^2 = \frac{e_i^2}{\frac{1}{n} \sum_{i=1}^n e_i^2}$)
3. Regress:

$$e_{i,(scaled)}^2 = \alpha_0 + \alpha_1 x_{i,1} + \alpha_2 x_{i,2} + u_i \quad (2)$$

4. Get the R^2 of this regression and the sample size n
5. Test the overall significance of (2):
 LM test statistic $= nR^2 \sim \chi_k^2$,
where k is the number of slope coefficients in (2)
6. If nR^2 is larger than the χ_k^2 critical value, then we reject H_0 of 'no heteroskedasticity'

Breusch-Pagan test automated

```
Breusch-Pagan test for heteroskedasticity
OLS, using observations 1-526
Dependent variable: scaled uhat^2
```

	coefficient	std. error	t-ratio	p-value

const	0.00899884	0.374915	0.02400	0.9809
educ	0.0296505	0.0264307	1.122	0.2625
exper	0.0720829	0.0183915	3.919	0.0001 ***
sq_exper	-0.00128451	0.000409712	-3.135	0.0018 ***

Explained sum of squares = 50.9586

Test statistic: LM = 25.479276,
with p-value = P(Chi-square(3) > 25.479276) = 0.000012

Remedies for (pure) heteroskedasticity

1. Redefine the variables:

- ▶ in order to reduce the variance of observations with extreme values
- ▶ e.g. by taking logarithms or rescaling (e.g. per capita rather than aggregate values)

2. Weighted Least Squares (WLS):

- ▶ consider the model: $y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i$
- ▶ suppose: $\text{Var}(\varepsilon_i) = \sigma^2 x_{i,2}^2$
- ▶ we can prove (not mandatory) that if we redefine the model as:

$$\frac{y_i}{x_{i,2}} = \beta_0 \frac{1}{x_{i,2}} + \beta_1 \frac{x_{i,1}}{x_{i,2}} + \beta_2 + \frac{\varepsilon_i}{x_{i,2}},$$

it becomes homoskedastic

3. White heteroskedasticity-consistent robust standard errors

White HC robust standard errors

- ▶ The logic behind: since heteroskedasticity causes problems with the OLS standard errors but not with the coefficients' estimates, it makes sense to improve the estimation of the standard errors in a way that does not alter the estimates of the coefficients (White, 1980)
- ▶ In matrix notation (for multiple regression model), we still have:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$$

- ▶ We estimate the variance as:

$$\widehat{\text{Var}}(\hat{\beta}) = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\hat{\Sigma}\mathbf{X} (\mathbf{X}'\mathbf{X})^{-1},$$

where $\mathbf{X}'\hat{\Sigma}\mathbf{X}$ is a $(k+1) \times (k+1)$ matrix that can be estimated out of the n observation

White HC robust standard errors

- More specifically:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 \\ 0 & 0 & \sigma_3^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sigma_n^2 \end{pmatrix}$$

and

$$\hat{\Sigma} = \begin{pmatrix} e_1^2 & 0 & 0 & \dots & 0 \\ 0 & e_2^2 & 0 & \dots & 0 \\ 0 & 0 & e_3^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & e_n^2 \end{pmatrix}$$

- We use this $\widehat{\text{Var}}(\hat{\beta})$ for statistical inference in our model

Robust standard errors in Gretl

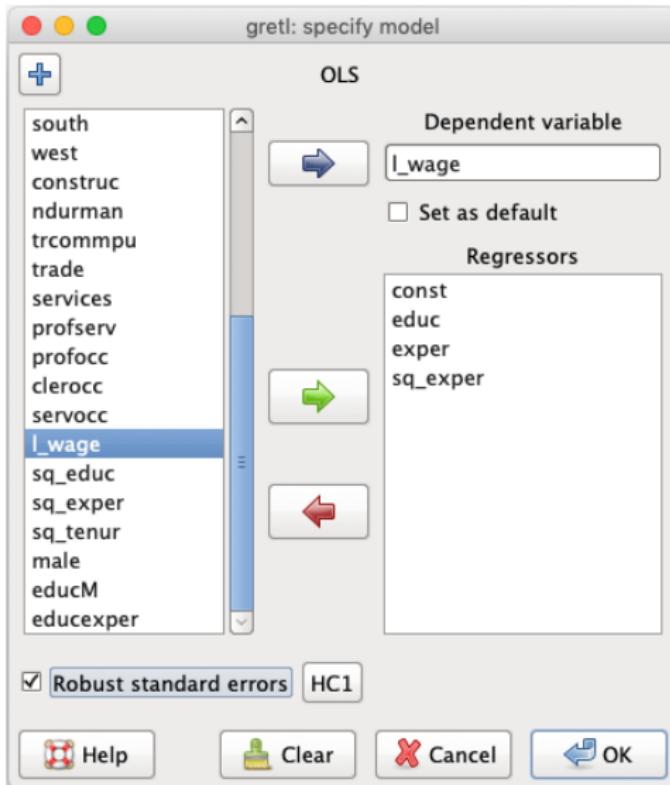
Model 1: OLS, using observations 1-526

Dependent variable: l_wage

	coefficient	std. error	t-ratio	p-value

const	0.127997	0.105932	1.208	0.2275
educ	0.0903658	0.00746800	12.10	6.98e-30 ***
exper	0.0410089	0.00519652	7.892	1.77e-14 ***
sq_exper	-0.000713558	0.000115764	-6.164	1.42e-09 ***
 Mean dependent var				
Sum squared resid	1.623268	S.D. dependent var	0.531538	
R-squared	103.7904	S.E. of regression	0.445906	
F(3, 522)	0.300273	Adjusted R-squared	0.296251	
Log-likelihood	74.66828	P-value(F)	3.38e-40	
Schwarz criterion	-319.5317	Akaike criterion	647.0633	
	664.1245	Hannan-Quinn	653.7435	

Robust standard errors in Gretl



Robust standard errors in Gretl

Model 2: OLS, using observations 1-526

Dependent variable: l_wage

Heteroskedasticity-robust standard errors, variant HC1

	coefficient	std. error	t-ratio	p-value

const	0.127997	0.107126	1.195	0.2327
educ	0.0903658	0.00778267	11.61	6.95e-28 ***
exper	0.0410089	0.00502368	8.163	2.46e-15 ***
sq_exper	-0.000713558	0.000109776	-6.500	1.88e-10 ***
Mean dependent var	1.623268	S.D. dependent var	0.531538	
Sum squared resid	103.7904	S.E. of regression	0.445906	
R-squared	0.300273	Adjusted R-squared	0.296251	
F(3, 522)	71.03088	P-value(F)	1.55e-38	
Log-likelihood	-319.5317	Akaike criterion	647.0633	
Schwarz criterion	664.1245	Hannan-Quinn	653.7435	

Summary

- ▶ Multicollinearity:
 - ▶ does not lead to a biased and consistent OLS estimator, but it makes the explanatory variables less significant
 - ▶ if really necessary, it can be remedied by dropping or transforming explanatory variables, or by getting more data
- ▶ Heteroskedasticity:
 - ▶ does not lead to a biased and consistent OLS estimator, but it makes the inference wrong
 - ▶ can be simply remedied by the use of robust standard errors

Seminars and the next lecture #9

- ▶ **Seminars:**
 - ▶ correct specifications of linear regression equations
 - ▶ practice of the tests presented today
- ▶ **Next lecture:**
 - ▶ we will continue the discussion about the form of the error
 - ▶ we will discuss the autocorrelation issue: testing and remedy
- ▶ **Readings for lecture #9:**
 - ▶ Studenmund (2016 & 17, 2014): Chapter 9
 - ▶ Wooldridge (2016, 2012): Chapter 12-1-5

LECTURE #9

Introductory Econometrics

AUTOCORRELATION & GLS

Jiri Kukacka, Ph.D.

Institute of Economic Studies, Faculty of Social Sciences, Charles University

Winter semester 2021, December 1

In previous lectures...

- ▶ We discussed the **specification** of regression equations
- ▶ We talked about the functional forms and the choice of independent variables
- ▶ In the last lecture #8, we continued the topic discussing **multicollinearity**, its consequences, detection (VIF), and remedies
- ▶ We also started to talk about the form of the error term: we discussed **heteroskedasticity**, its consequences, detection (White test, Breusch-Pagan test), and remedies (WLS, White HC robust SE)

In today's lecture #9, we will...

- ▶ Finish the discussion of the form of the error term by talking about **autocorrelation (serial correlation)**
- ▶ Learn:
 - ▶ what is the nature of the problem
 - ▶ what are its consequences
 - ▶ how it is diagnosed
 - ▶ what are the remedies available
- ▶ Readings for this week:
 - ▶ Studenmund (2016 & 17, 2014): Chapter 9
 - ▶ Wooldridge (2016, 2012): Chapter 12-1-5

Nature of autocorrelation

- ▶ Observations of the error term are correlated with each other:

$$\text{Cov}(\varepsilon_{t_1}, \varepsilon_{t_2}) \neq 0, \quad t_1 \neq t_2$$

- ▶ Violation of one of the Classical Assumptions 4.
- ▶ Can exist in any data in which the order of the observations has some meaning, most frequently in **time-series data**
- ▶ Particular form of autocorrelation: $AR(p)$ process:

$$\varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + \dots + \rho_p \varepsilon_{t-p} + u_t$$

- ▶ u_t is a classical (not autocorrelated) error term
- ▶ ρ_1 to ρ_p are autocorrelation coefficients (between -1 and 1)

Examples of pure autocorrelation

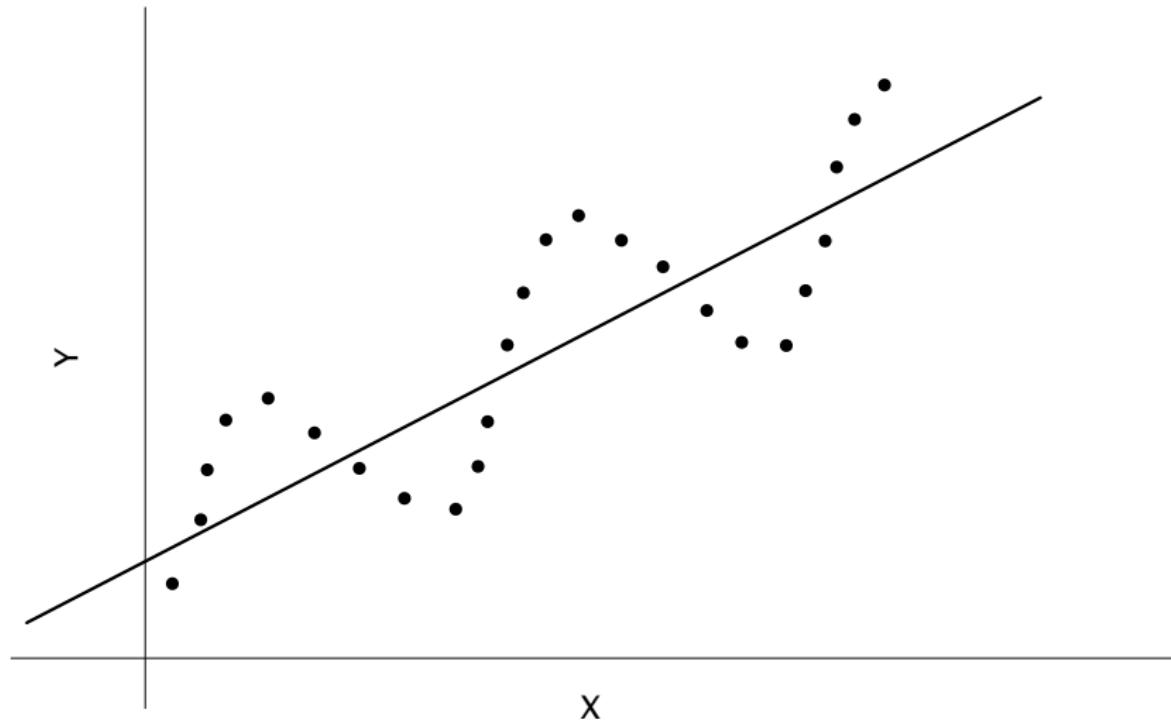
- ▶ Distribution of the error term has an autocorrelation nature
- ▶ First-order autocorrelation: $AR(1)$ process:

$$\varepsilon_t = \rho_1 \varepsilon_{t-1} + u_t$$

- ▶ positive serial correlation: ρ_1 is positive
- ▶ negative serial correlation: ρ_1 is negative
- ▶ no serial correlation: ρ_1 is zero
- ▶ positive autocorrelation very common in time series data
- ▶ e.g.: a shock to GDP persists for more than one period
- ▶ Seasonal autocorrelation (in quarterly data):

$$\varepsilon_t = \rho_4 \varepsilon_{t-4} + u_t$$

Autocorrelation



Examples of impure autocorrelation

- ▶ Autocorrelation caused by a specification error in the equation:
 - ▶ incorrect functional form
 - ▶ omitted variable
- ▶ Intuition:
 - ▶ recall that the error term includes the omitted variables, nonlinearities, measurement errors, and the classical error term
 - ▶ if we omit a serially correlated variable, it is included in the error term, causing the autocorrelation problem:
 - ▶ assume the true model: $y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t$
 - ▶ but we estimate: $y_t = \beta_0 + \beta_1 x_{1t} + v_t$
 - ▶ then the error term v_t in the misspecified model contains the impact of $\beta_2 x_{2t}$ together with ε_t
 $\Rightarrow v_t$ can be serially correlated even if ε_t is not
 - ▶ Impure autocorrelation can be corrected by better choice of specification (as opposed to pure autocorrelation)

Consequences of autocorrelation

1. OLS estimator remains unbiased and consistent (for pure autocorrelation)
 2. Variance of the $\hat{\beta}^{OLS}$ distribution increases:
 - ▶ because the autocorrelated error term explains a larger proportion of fluctuations of the dependent variable
 - ▶ violation of the Classical Assumptions 4. \Rightarrow OLS not BLUE
 - \Rightarrow OLS more likely to misestimate the true β
 3. But OLS tends to underestimate the variance of $\hat{\beta}^{OLS}$:
 - ▶ increase of the (true) variance is, however, masked by OLS because it assumes an unautocorrelated error
 - ▶ OLS thus attributes the impact of the autocorrelated error to the independent variables
 - \Rightarrow t -statistics tend to be larger under autocorrelation, statistical inference becomes unreliable and incorrect
- \Rightarrow The same consequences as for heteroskedasticity!

Variance of OLS under autocorrelation

- ▶ Consider the model: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
- ▶ Suppose that: $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$
- ▶ OLS estimator is: $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$
- ▶ Variance of the OLS estimator:

$$\text{Var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Omega\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1},$$

where (simplified):

$$\Omega = \text{Var}(\boldsymbol{\varepsilon}) = \sigma_u^2 \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{T-1} \\ \rho & 1 & \rho & \dots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \dots & 1 \end{pmatrix}$$

Durbin-Watson test for autocorrelation, AR(1)

- ▶ Used to determine if there is a first-order serial correlation by examining the residuals of the equation
- ▶ Assumptions (criteria for using this test):
 - ▶ the regression includes the intercept
 - ▶ if autocorrelation is present, it is of $AR(1)$ type:

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t$$

- ▶ the regression does not include a lagged dependent variable

Durbin-Watson test for autocorrelation, AR(1)

- Durbin-Watson d statistic (for T observations):

$$d = \frac{\sum_{t=2}^T (e_t - e_{t-1})^2}{\sum_{t=1}^T e_t^2} \approx 2(1 - \hat{\rho}),$$

where $\hat{\rho}$ is the estimated autocorrelation coefficient

- Values:
 1. extreme positive serial correlation: $d \approx 0$
 2. extreme negative serial correlation: $d \approx 4$
 3. no serial correlation: $d \approx 2$

Using the Durbin-Watson test

1. Estimate the equation by OLS, save the residuals
2. Calculate the d statistic
3. Determine the sample size T and the number of explanatory variables k (intercept not considered)
4. Find the upper critical value d_U and the lower critical value d_L for T and k in statistical tables [e.g. in Studenmund (2014: pg. 547–549; 2016: pg. 525; 2017: pg. 542-543), or in Gretl]
5. Evaluate the test as one-sided or two-sided (see next slides)

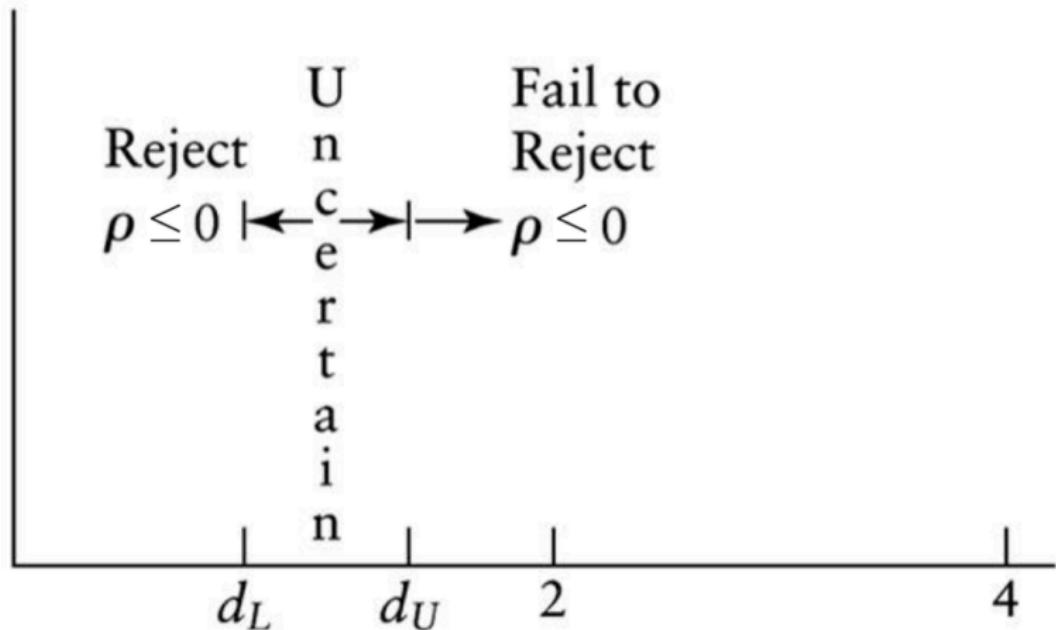
One-sided Durbin-Watson test

- ▶ For cases when we consider only positive serial correlation as an option
- ▶ Hypotheses:

$$H_0 : \rho \leq 0 \quad (\text{no positive serial correlation})$$
$$H_A : \rho > 0 \quad (\text{positive serial correlation})$$

- ▶ Decision rule:
 - ▶ if $d < d_L$: reject H_0
 - ▶ if $d > d_U$: do not reject H_0
 - ▶ if $d_L \leq d \leq d_U$: inconclusive

Durbin-Watson critical values for one-sided test



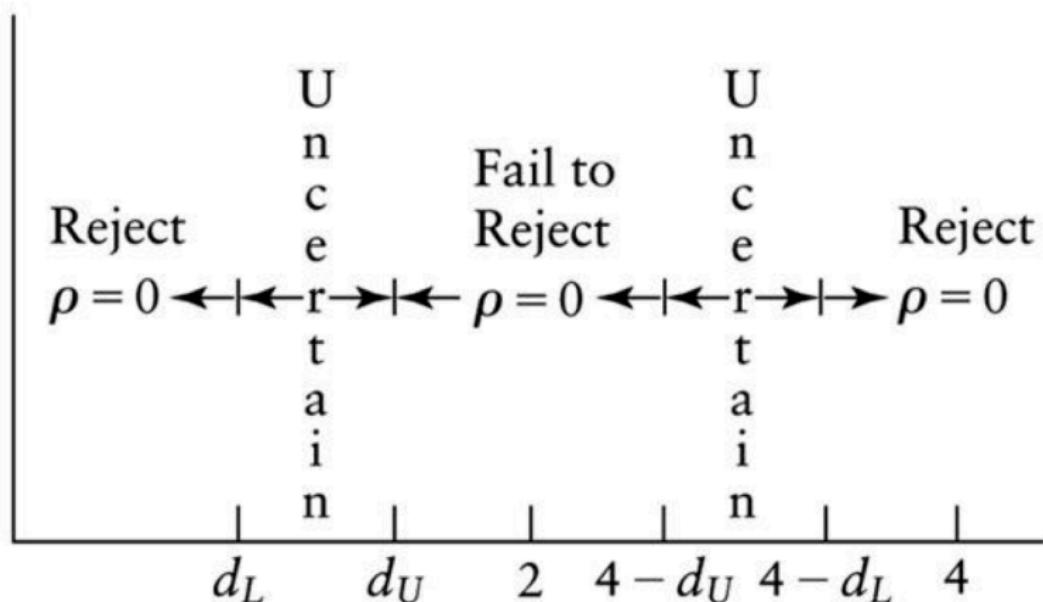
Two-sided Durbin-Watson test

- ▶ For cases when we consider both signs of serial correlation
- ▶ Hypotheses:

$$H_0 : \rho = 0 \quad (\text{no serial correlation})$$
$$H_A : \rho \neq 0 \quad (\text{serial correlation})$$

- ▶ Decision rule:
 - ▶ if $d < d_L$: reject H_0
 - ▶ if $d > 4 - d_L$: reject H_0
 - ▶ if $d > d_U$: do not reject H_0
 - ▶ if $d < 4 - d_U$: do not reject H_0
 - ▶ otherwise: inconclusive

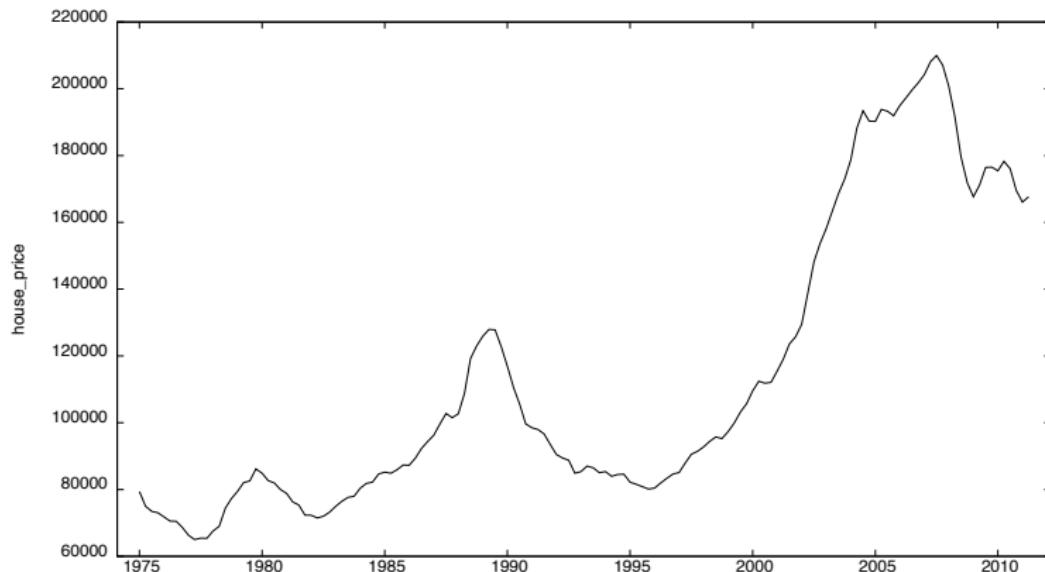
Durbin-Watson critical values for two-sided test



Example

- ▶ Estimating housing prices in the UK
- ▶ Quarterly time series data on prices of a representative house in the UK (in \$)
- ▶ Explanatory variable: GDP (in billions of \$)
- ▶ Time span: 1975:Q1–2011:Q2
- ▶ The time series is in real prices (i.e. adjusted for inflation)

Example



Example in Gretl

Model 1: OLS, using observations 1975:1-2011:2 (T = 146)
Dependent variable: house_price

	coefficient	std. error	t-ratio	p-value

const	-38409.8	6675.01	-5.754	5.04e-08 ***
gdp	737.065	31.4846	23.41	6.09e-51 ***
Mean dependent var	113072.8	S.D. dependent var	43254.80	
Sum squared resid	5.65e+10	S.E. of regression	19799.38	
R-squared	0.791921	Adjusted R-squared	0.790476	
F(1, 144)	548.0434	P-value(F)	6.09e-51	
Log-likelihood	-1650.595	Akaike criterion	3305.191	
Schwarz criterion	3311.158	Hannan-Quinn	3307.615	
rho	0.984890	Durbin-Watson	0.023930	

Example

- We test for positive serial correlation:

$$H_0 : \rho \leq 0 \quad (\text{no positive serial correlation})$$

$$H_A : \rho > 0 \quad (\text{positive serial correlation})$$

- One-sided DW critical values at the 95% confidence level for $T = 146$ and $k = 1$ are:

$$d_L = 1.72 \quad \text{and} \quad d_U = 1.74$$

- Decision rule:

- if $d < 1.72$ reject H_0
- if $d > 1.74$ do not reject H_0
- if $1.72 \leq d \leq 1.74$ inconclusive

- Since $d \doteq 0.024 < 1.72$, we reject the null hypothesis of no positive serial correlation

Breusch-Godfrey test for autocorrelation

- ▶ Suppose we suspect the stochastic error term to be $AR(p)$:

$$\varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + \dots + \rho_p \varepsilon_{t-p} + u_t$$

- ▶ Since $\hat{\beta}^{OLS}$ is consistent under autocorrelation, the residuals serve as a consistent estimator of the stochastic error term
- ▶ Hence, it is sufficient to:
 1. estimate the original model by OLS, save the residuals e_t
 2. regress:
$$e_t = \alpha_0 + \alpha_1 x_1 + \dots + \alpha_k x_k + \rho_1 e_{t-1} + \rho_2 e_{t-2} + \dots + \rho_p e_{t-p} + v_t$$
 3. test if $\rho_1 = \rho_2 = \dots = \rho_p = 0$ using a standard $F(t)$ -test
 4. or use the LM test statistic $= (T - p)R^2 \sim \chi_p^2$

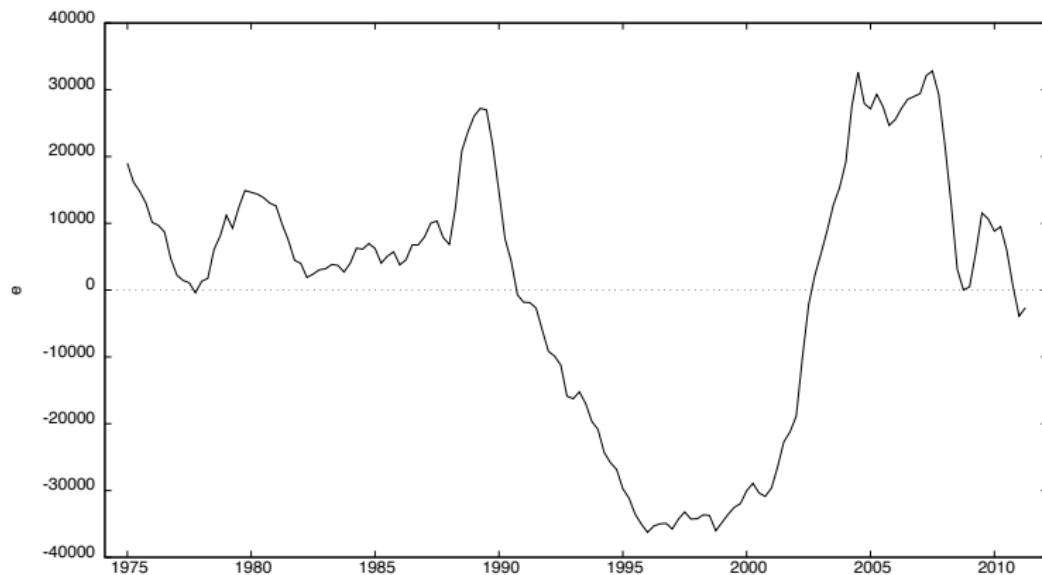
Back to example in Gretl

Model 1: OLS, using observations 1975:1-2011:2 (T = 146)
Dependent variable: house_price

	coefficient	std. error	t-ratio	p-value

const	-38409.8	6675.01	-5.754	5.04e-08 ***
gdp	737.065	31.4846	23.41	6.09e-51 ***
Mean dependent var	113072.8	S.D. dependent var	43254.80	
Sum squared resid	5.65e+10	S.E. of regression	19799.38	
R-squared	0.791921	Adjusted R-squared	0.790476	
F(1, 144)	548.0434	P-value(F)	6.09e-51	
Log-likelihood	-1650.595	Akaike criterion	3305.191	
Schwarz criterion	3311.158	Hannan-Quinn	3307.615	
rho	0.984890	Durbin-Watson	0.023930	

Example



Back to example in Gretl

Breusch-Godfrey test for autocorrelation up to order 4

OLS, using observations 1975:1-2011:2 (T = 146)

Dependent variable: e

	coefficient	std. error	t-ratio	p-value

const	-47.4964	1026.50	-0.04627	0.9632
gdp	0.189795	4.84365	0.03918	0.9688
e_1	1.44176	0.0846476	17.03	6.18e-36 ***
e_2	-0.466424	0.149103	-3.128	0.0021 ***
e_3	0.0791018	0.149400	0.5295	0.5973
e_4	-0.0795904	0.0856819	-0.9289	0.3545

Unadjusted R-squared = 0.977134

Test statistic: LMF = 1495.639866,

with p-value = P(F(4,140) > 1495.64) = 9.66e-114

Alternative statistic: TR^2 = 142.661527,

with p-value = P(Chi-square(4) > 142.662) = 7.6e-30

Available remedies

What to do after a test detects serial correlation:

- ▶ First of all, reconsider your specification, any possible source of impure autocorrelation?
- ▶ Generalized Least Squares (GLS) estimation
- ▶ Feasible Generalized Least Squares (FGLS) estimation
- ▶ Newey-West robust standard errors

Generalized Least Squares (GLS)

- When the variance-covariance matrix of the error term is not diagonal, but it has some known structure
- It solves both heteroskedasticity and/or autocorrelation issue
- Suppose we have the model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where the variance-covariance matrix of the error term is:

$$Var(\boldsymbol{\varepsilon}) = \boldsymbol{\Omega} \neq \sigma^2 \mathbf{I}$$

- Generalized Least Squares (GLS) estimator:

$$\hat{\boldsymbol{\beta}}^{GLS} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1} \mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}$$

- Variance of the GLS estimator is:

$$Var\left(\hat{\boldsymbol{\beta}}^{GLS}\right) = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}$$

▶ Derivation

Example of the GSL estimator for AR(1) error term

Assume (1) $Y_t = \beta_0 + \beta_1 x_{1t} + \varepsilon_t$, where $\varepsilon_t = \rho \varepsilon_{t-1} + u_t$

⇓

$$(2) \quad Y_t = \beta_0 + \beta_1 x_{1t} + \rho \varepsilon_{t-1} + u_t$$

&

$$(3) \quad \rho Y_{t-1} = \rho \beta_0 + \rho \beta_1 x_{1t-1} + \rho \varepsilon_{t-1}$$

⇓ $(2) - (3)$

$$(4) \quad Y_t - \rho Y_{t-1} = \beta_0(1 - \rho) + \beta_1(x_{1t} - \rho x_{1t-1}) + u_t$$

⇓

$$Y_t^* = \beta_0^* + \beta_1 x_{1t}^* + u_t$$

Feasible Generalized Least Squares (FGLS)

- ▶ Usually, we do not know the matrix Ω
- ▶ We can estimate it from our data and use the estimated $\hat{\Omega}$:

$$\hat{\beta}^{FGLS} = \left(\mathbf{x}' \hat{\Omega}^{-1} \mathbf{x} \right)^{-1} \mathbf{x}' \hat{\Omega}^{-1} \mathbf{y}$$

- ▶ Procedure:
 1. estimate the original model (with autocorrelated errors) by OLS (consistent)
 2. save the residuals and use them to estimate $\hat{\Omega}$
 3. use $\hat{\Omega}$ to find FGLS estimates and SEs
 4. (use them in 1. again)
- ▶ Cochrane-Orcutt method (in practice iterative, it repeats 1.–4. until $\hat{\Omega}$ stabilizes), or Prais-Winsten method

Back to example in Gretl

Model 1: OLS, using observations 1975:1-2011:2 (T = 146)
Dependent variable: house_price

	coefficient	std. error	t-ratio	p-value

const	-38409.8	6675.01	-5.754	5.04e-08 ***
gdp	737.065	31.4846	23.41	6.09e-51 ***
Mean dependent var	113072.8	S.D. dependent var	43254.80	
Sum squared resid	5.65e+10	S.E. of regression	19799.38	
R-squared	0.791921	Adjusted R-squared	0.790476	
F(1, 144)	548.0434	P-value(F)	6.09e-51	
Log-likelihood	-1650.595	Akaike criterion	3305.191	
Schwarz criterion	3311.158	Hannan-Quinn	3307.615	
rho	0.984890	Durbin-Watson	0.023930	

Back to example in Gretl

Performing iterative calculation of rho...

ITER	RHO	ESS
1	0.98489	1.31572e+09
... 8	0.98529	1.31571e+09

Model 3: Cochrane-Orcutt, using observations 1975:2-2011:2 (T = 145)

Dependent variable: house_price

rho = 0.985292

	coefficient	std. error	t-ratio	p-value	
<hr/>					
const	-103719	41986.6	-2.470	0.0147	**
gdp	937.741	139.437	6.725	3.90e-10	***

Statistics based on the rho-differenced data:

Mean dependent var	113305.1	S.D. dependent var	43313.28
Sum squared resid	1.32e+09	S.E. of regression	3033.273
R-squared	0.995134	Adjusted R-squared	0.995100
F(1, 143)	45.22821	P-value(F)	3.90e-10
rho	0.603212	Durbin-Watson	0.791914

Alternative remedy: Robust standard errors

- ▶ Note that autocorrelation does not lead to an inconsistent OLS estimator, only to incorrect inference (similar to heteroskedasticity problem)
- ⇒ We can keep the estimated coefficients, and only adjust the standard errors
- ▶ The general idea is similar to White heteroskedasticity-consistent robust SEs, we can use estimated var-cov matrix $\widehat{\Omega}$ to compute SEs of $\widehat{\beta}$:

$$\widehat{Var}(\widehat{\beta}) = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\widehat{\Omega}\mathbf{X} (\mathbf{X}'\mathbf{X})^{-1},$$

- ▶ Keep in mind that this Ω is not the same as Σ in White SEs
- ▶ Note that consistent estimator of Ω would solve not only autocorrelation, but also heteroskedasticity (HAC robust SEs)

Newey-West robust standard errors

- ▶ In Gretl use (Newey-West) robust standard errors (HAC)
- ▶ Newey-West robust SEs work for fairly arbitrary forms of autocorrelation
- ▶ N-W robust SEs are larger than OLS SEs \Rightarrow lower t -statistics
- ▶ Also note that all derived results hold iff the 3. Classical Assumption $\text{Cov}(x, \varepsilon) = 0$ is not violated \Rightarrow **first make sure the specification of the model is correct**, only then try to correct for the form of the error term!

Summary

- ▶ Autocorrelation does not lead to a biased and inconsistent OLS estimator, but it makes the inference wrong (estimated coefficients are correct, but their standard errors tend to be underestimated)
- ▶ It can be diagnosed using:
 - ▶ Durbin-Watson test
 - ▶ Breusch-Godfrey test
 - ▶ analysis of residuals
- ▶ It can be remedied by:
 - ▶ FGLS method
 - ▶ Newey-West robust standard errors

Seminars and the next lecture #10

- ▶ **Seminars:**
 - ▶ autocorrelation practice in Gretl: estimation, inference, and testing
- ▶ **Next lecture:**
 - ▶ we will discuss the issue of endogeneity
- ▶ **Practical information:**
 - ▶ home assignment #3 will be assigned today (see SIS)
 - ▶ deadline: Thursday, December 9, 2021, 23:59:59
 - ▶ see 'Requirements to the exam' in SIS
 - ▶ see *Final_exam_ONLINE_info_JEM062_2021.pdf* in SIS
- ▶ **Readings for lecture #10:**
 - ▶ Studenmund (2016 & 17, [2014]): Chapter 14 [13]
 - ▶ Wooldridge (2016, 2012): Chapter 15, Chapter 16-1–16-3

Appendix: Derivation of GLS (mandatory?)

- We find the Cholesky decomposition of Ω :

$$\Omega^{-1} = \mathbf{D}'\mathbf{D}$$

- We multiply the model by \mathbf{D} :

$$\begin{aligned}\mathbf{y} &= \mathbf{X}\beta + \varepsilon \\ \mathbf{Dy} &= \mathbf{DX}\beta + \mathbf{D}\varepsilon \\ \mathbf{y}^* &= \mathbf{X}^*\beta + \varepsilon^*\end{aligned}$$

- Now, we have:

$$\begin{aligned}Var(\varepsilon^*) &= Var(\mathbf{D}\varepsilon) = \mathbf{D}Var(\varepsilon)\mathbf{D}' = \mathbf{D}\Omega\mathbf{D}' \\ &= \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}' = \mathbf{DD}^{-1}(\mathbf{D}')^{-1}\mathbf{D}' = \mathbf{I}\end{aligned}$$

- Hence, the transformed error term satisfies the Classical Assumption 4. and 5. \Rightarrow we can apply OLS on the transformed model:

$$\begin{aligned}\hat{\beta}^{GLS} &= (\mathbf{X}^*\mathbf{X})^{-1}\mathbf{X}^*\mathbf{y}^* = ((\mathbf{DX})'\mathbf{DX})^{-1}(\mathbf{DX})'\mathbf{Dy} \\ &= (\mathbf{X}'\mathbf{D}'\mathbf{DX})^{-1}\mathbf{X}'\mathbf{D}'\mathbf{Dy} = (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}'\Omega^{-1}\mathbf{y}\end{aligned}$$

LECTURE #10

Introductory Econometrics

ENDOGENEITY

Jiri Kukacka, Ph.D.

Institute of Economic Studies, Faculty of Social Sciences, Charles University

Winter semester 2021, December 8

In the previous lectures #6, 7, 8, 9...

- ▶ We discussed what happens if some of the CAs are violated
- ▶ Linearity in parameters and no perfect multicollinearity are essential for the definition of the OLS estimator
- ▶ Zero mean of the error term is always ensured by the inclusion of the intercept
- ▶ Normality of the stochastic error term is needed for statistical inference but if the number of observations is sufficiently high, the OLS estimator will have asymptotically standard normal distribution even if the error term is not normally distributed
- ▶ Heteroskedasticity and serial correlation lead to incorrect statistical inference, but we have studied a set of techniques to overcome this problem (detection → remedy)

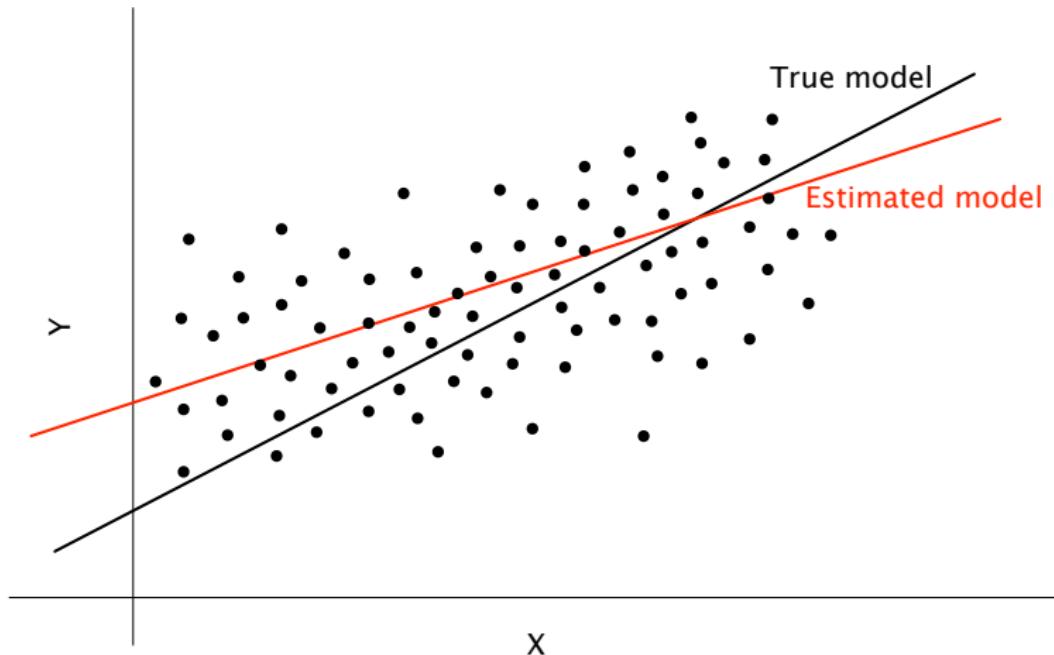
On today's lecture, we will...

- ▶ Study the crucial CA 3. of no correlation between explanatory variables and the error term
- ▶ Learn that variables correlated with the error term are called **endogenous variables** (as opposed to **exogenous variables**)
- ▶ Show that the OLS estimator of endogenous variables is biased and inconsistent
- ▶ Define the concept of Instrumental variables (IV)
- ▶ Derive the 2SLS technique to deal with endogeneity
- ▶ Describe the Hausman test for endogeneity
- ▶ Readings for this week:
 - ▶ Studenmund (2016 & 17, [2014]): Chapter 14 [13]
 - ▶ Wooldridge (2016, 2012): Chapter 15, Chapter 16-1-16-3

Endogenous variables

- ▶ Notation: $E[x_i \varepsilon_i] = \text{Cov}(x_i, \varepsilon_i) \neq 0$ or $E[\mathbf{X}' \boldsymbol{\varepsilon}] \neq \mathbf{0}$
- ▶ Example: Analysis of household consumption patterns
 - ▶ households with lower income may indicate higher consumption (because of shame)
- ▶ Intuition behind the bias:
 - ▶ if an explanatory variable x and the error term ε are correlated with each other, the OLS estimator attributes to x some of the variation in y that actually comes from the error term ε
- ▶ Lead to biased and inconsistent OLS estimator

Graphical representation



Inconsistency of the OLS estimator (not mandatory)

- We can express:

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon \\ &= \beta + \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\frac{1}{n}\mathbf{X}'\varepsilon\end{aligned}$$

- We assume that there exists a finite matrix \mathbf{Q} so that:

$$\frac{1}{n}\mathbf{X}'\mathbf{X} \xrightarrow{n \rightarrow \infty} \mathbf{Q}$$

- It can be shown that: $\frac{1}{n}\mathbf{X}'\varepsilon \xrightarrow{n \rightarrow \infty} E[\mathbf{X}'\varepsilon] \stackrel{\text{endogeneity}}{\neq} \mathbf{0}$

- This implies:

$$\hat{\beta} \xrightarrow{n \rightarrow \infty} \beta + \mathbf{Q}^{-1} \cdot E[\mathbf{X}'\varepsilon] = \beta + bias$$

Typical cases of endogeneity

1. Omitted variable/influence bias

- ▶ an explanatory variable is omitted from the equation and makes part of the error term
- ▶ an unobservable characteristic has influence on both dependent variable y and explanatory variable x (*abil?*)

2. Simultaneity

- ▶ the causal relationship between the dependent variable and the explanatory variable goes in both directions

3. Measurement error

- ▶ some of the variables are measured with error
- ▶ In all 3 cases, the direction of the bias is given by the sign of $\text{Cov}(x_i, \tilde{\varepsilon}_i)$, but...

Omitted variable bias (lecture #7: EBA)

- ▶ True model: $y_i = \beta_0 + \beta_1 x_i + \gamma z_i + \varepsilon_i$
- ▶ Model as it looks when we omit explanatory variable z :

$$y_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_i + \tilde{\varepsilon}_i \quad \Rightarrow \quad \tilde{\varepsilon}_i = \gamma z_i + \varepsilon_i$$

- ▶ This gives:

$$\text{Cov}(x_i, \tilde{\varepsilon}_i) = \text{Cov}(x_i, \gamma z_i + \varepsilon_i) = \gamma \text{Cov}(x_i, z_i) \neq 0$$

- ▶ It can be remedied by including the variable in question, but often we do not have data for it... we can include a proxy for such variable, but this may not reduce the bias completely

Actually:

$$\plim \tilde{\beta}_1^{OLS} = \beta_1 + \text{Corr}(x, \tilde{\varepsilon}) \frac{\sigma_{\tilde{\varepsilon}}}{\sigma_x} = \beta_1 + \frac{\text{Cov}(x, \tilde{\varepsilon})}{\sigma_x \sigma_{\tilde{\varepsilon}}} \frac{\sigma_{\tilde{\varepsilon}}}{\sigma_x} = \beta_1 + \frac{\gamma \text{Cov}(x, z)}{\text{Var}(x)}$$

Simultaneity

- ▶ Occurs in models where variables are jointly determined:

$$y_{1i} = \alpha_0 + \alpha_1 y_{2i} + \varepsilon_{1i}$$

$$y_{2i} = \beta_0 + \beta_1 y_{1i} + \varepsilon_{2i}$$

- ▶ Intuitively: change in y_{1i} will cause a change in y_{2i} , which will in turn cause y_{1i} to change again...
- ▶ Technically:

$$\begin{aligned} \text{Cov}(y_{2i}, \varepsilon_{1i}) &= \text{Cov}(\beta_0 + \beta_1 y_{1i} + \varepsilon_{2i}, \varepsilon_{1i}) \\ &= \beta_1 \text{Cov}(y_{1i}, \varepsilon_{1i}) \\ &= \beta_1 \text{Cov}(\alpha_0 + \alpha_1 y_{2i} + \varepsilon_{1i}, \varepsilon_{1i}) \\ &= \beta_1 [\alpha_1 \text{Cov}(y_{2i}, \varepsilon_{1i}) + \text{Var}(\varepsilon_{1i})] \dots \\ \text{Cov}(y_{2i}, \varepsilon_{1i}) &= \frac{\beta_1}{1 - \alpha_1 \beta_1} \text{Var}(\varepsilon_{1i}) \neq 0 \end{aligned}$$

Simultaneity

- ▶ A simplified example:

$$Q_{Di} = \alpha_0 + \alpha_1 P_i + \alpha_2 Yd_i + \varepsilon_{1i}$$

$$Q_{Si} = \beta_0 + \beta_1 P_i + \varepsilon_{2i}$$

$$Q_{Di} = Q_{Si}$$

$$\Rightarrow P_i = \gamma_0 + \gamma_1 Q_{Di} + \varepsilon_{3i},$$

where:

Q_D ... quantity demanded

Q_S ... quantity supplied

P ... commodity price

Yd ... disposable income

- ▶ Endogeneity of price: it is determined from the interaction of supply and demand

Measurement error

- ▶ E.g. measurement error in the dependent variable (in the explanatory variable also possible)
- ▶ Measurement error is correlated with an explanatory variable:

$$y_i = y_i^* + \nu_i \quad \text{where} \quad \text{Cov}(x_i, \nu_i) \neq 0$$

- ▶ True regression model: $y_i^* = \beta_0 + \beta_1 x_i + \varepsilon_i$

- ▶ Estimated regression: $y_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_i + \tilde{\varepsilon}_i$,

where $\tilde{\varepsilon}_i = \varepsilon_i + \nu_i$ and so:

$$\text{Cov}(x_i, \tilde{\varepsilon}_i) = \text{Cov}(x_i, \varepsilon_i + \nu_i) = \text{Cov}(x_i, \nu_i) \neq 0$$

- ▶ Example: analysis of household consumption patterns (above)

Instrumental variables (IV)

- ▶ A remedy for the situation when $\text{Cov}(x, \varepsilon) \neq 0$
- ▶ Instrument should be a variable z such that:
 1. z is uncorrelated with the error term: $\text{Cov}(z, \varepsilon) = 0$
 2. z is correlated with the explanatory variable x : $\text{Cov}(z, x) \neq 0$
- ▶ Idea behind Instrumental variables approach:
 - ▶ regress the endogenous variable x on the instrument $z \Rightarrow$ IV
 - ▶ this IV is uncorrelated with the error term and can be used as an explanatory variable instead of x

Instrumental variables (IV)

- ▶ Suppose the equation we want to estimate is:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (1)$$

- ▶ we can have several instruments for several endogenous variables \Rightarrow we will use the matrix notation \mathbf{Z} and \mathbf{X}
- ▶ \mathbf{X} denotes endogenous variable(s) in (1)
- ▶ \mathbf{Z} denotes instrument(s)
- ▶ assume that $\text{rank}(\mathbf{Z}) \geq \text{rank}(\mathbf{X})$, i.e. we have at least as many instruments as endogenous variables

Two-Stage Least Squares (2SLS, TSLS)

- ▶ 2SLS is a method of implementing Instrumental variables approach
- ▶ Consist of two steps:
 1. regress the endogenous variables on the instruments:

$$\mathbf{X} = \mathbf{Z}\boldsymbol{\pi} + \boldsymbol{u}$$

(for one instrument: $x = \pi_0 + \pi_1 z + u$)

and get fitted/predicted values:

$$\hat{\mathbf{X}} = \mathbf{Z}\hat{\boldsymbol{\pi}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X},$$

(for one instrument: $\hat{x} = \hat{\pi}_0 + \hat{\pi}_1 z$)

2. use these predicted values instead of \mathbf{X} in the original equation:

$$\mathbf{y} = \hat{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

(for one instrument: $y = \beta_0 + \beta_1 \hat{x} + \tilde{\varepsilon}$)

Two-Stage Least Squares

- The estimator is:

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

- This estimator is consistent (what about the bias?), but it is not efficient as it has higher variance than OLS
- Intuitively:
 - only part of the variation in \mathbf{X} that is 'almost' uncorrelated with the error term is used for the estimation ($\hat{\mathbf{X}}$ is 'almost' uncorrelated with error term)
⇒ this ensures consistency
 - but it makes $\hat{\beta}^{2SLS}$ less precise (higher variance of the estimator) because not all variation in \mathbf{X} is used

2SLS: Multiple regression model

- ▶ Assume structural equation:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 x + \varepsilon,$$

where y_2 is endogenous and x is exogenous

- ▶ We look for an instrument z such that:

1. z is related to y_2 , i.e. $\pi_1 \neq 0$ in $y_2 = \pi_0 + \pi_1 z + \pi_2 x + u$
(reduced-form equation)
2. $Cov(z, \varepsilon) = Cov(x, \varepsilon) = 0$
3. $E(u) = 0$ and $Cov(z, u) = Cov(x, u) = 0$

- ▶ 2SLS:

1. estimate reduced-form equation to get \hat{y}_2
2. estimate the structural equation using \hat{y}_2 instead of y_2

Example: OLS

- ▶ Estimating the impact of education on the number of children for a sample of women in Botswana (data from 1988) by OLS:

Model 1: OLS, using observations 1-4361

Dependent variable: children

	coefficient	std. error	t-ratio	p-value

const	-4.13831	0.240594	-17.20	3.33e-64 ***
educ	-0.0905755	0.00592069	-15.30	1.68e-51 ***
age	0.332449	0.0165495	20.09	6.52e-86 ***
sq_age	-0.00263082	0.000272592	-9.651	8.03e-22 ***
 Mean dependent var				
Sum squared resid	2.267828	S.D. dependent var	9284.147	2.222032
R-squared	0.568724	S.E. of regression	0.568427	
F(3, 4357)	1915.196	Adjusted R-squared	P-value(F)	0.000000

Example: Instrument

- ▶ Education may be endogenous \Rightarrow both education and number of children may be influenced by some unobserved socio-cultural-economic factors:
 - ▶ family background is an unobserved factor that influences both the number of children and years of education
- ▶ Finding a possible instrument:
 - ▶ a variable that explains education
 - ▶ but is not correlated with the family background (in the error)
- ▶ A dummy variable:

$$frsthalf = \begin{cases} 1 & \text{if the woman was born in the first} \\ & \text{six months of a year} \\ 0 & \text{otherwise} \end{cases}$$

Example: Intuition behind the instrument

- ▶ The first condition: instrument explains education:
 - ▶ school year in Botswana starts in January & after 6 of age
⇒ thus, women born in the first half of the year start school when they are at least six and a half
 - ▶ schooling is compulsory till the age of 15
⇒ thus, women born in the first half of the year get less education if they leave school at the age of 15
- ▶ The second condition: instrument is uncorrelated with the family background, thus also with the error term:
 - ▶ being born in the first half of the year is uncorrelated with the unobserved socio-cultural-economic factors that influence education and number of children (family background, etc.)

Example: First stage regression

Model 2: OLS, using observations 1-4361

Dependent variable: educ

	coefficient	std. error	t-ratio	p-value	
const	9.69286	0.598069	16.21	2.09e-57	***
frsthalf	-0.852285	0.112830	-7.554	5.12e-14	***
age	-0.107950	0.0420402	-2.568	0.0103	**
sq_age	-0.000505567	0.000692940	-0.7296	0.4657	
Mean dependent var	5.855996	S.D. dependent var	3.927075		
Sum squared resid	60001.14	S.E. of regression	3.710957		
R-squared	0.107651	Adjusted R-squared	0.107037		
F(3, 4357)	175.2068	P-value(F)	3.0e-107		

- ▶ Save fitted values as *educ_hat2* (fitted values from model 2)

Example: Second stage regression

Model 3: OLS, using observations 1-4361

Dependent variable: children

	coefficient	std. error	t-ratio	p-value

const	-3.38781	0.550340	-6.156	8.14e-10 ***
educ_hat2	-0.171499	0.0533921	-3.212	0.0013 ***
age	0.323605	0.0179310	18.05	2.84e-70 ***
sq_age	-0.00267228	0.000280805	-9.516	2.88e-21 ***
Mean dependent var	2.267828	S.D. dependent var	2.222032	
Sum squared resid	9759.726	S.E. of regression	1.496667	
R-squared	0.546632	Adjusted R-squared	0.546320	
F(3, 4357)	1751.100	P-value(F)	0.000000	

2SLS

- ▶ Note that the endogenous variable has to be instrumented by the instrument and by all other exogenous variables included in the regression
- ▶ Including all exogenous variables in the first stage makes them orthogonal to the residual \hat{u} and hence uncorrelated to the error term $\tilde{\varepsilon} = \hat{u}\beta + \varepsilon$ in the second stage

Example: 2SLS (jointly)

Model 4: TSLS, using observations 1-4361

Dependent variable: children

Instrumented: educ

Instruments: const frsthalf age sq_age

	coefficient	std. error	z	p-value	

const	-3.38781	0.548150	-6.180	6.39e-10	***
educ	-0.171499	0.0531796	-3.225	0.0013	***
age	0.323605	0.0178596	18.12	2.24e-73	***
sq_age	-0.00267228	0.000279687	-9.555	1.24e-21	***
Mean dependent var	2.267828	S.D. dependent var	2.222032		
Sum squared resid	9682.216	S.E. of regression	1.490712		
R-squared	0.552676	Adjusted R-squared	0.552368		
F(3, 4357)	1765.119	P-value(F)	0.000000		

Back to the example

- ▶ Compare the estimates from OLS and 2SLS
- ▶ OLS:

	coefficient	std. error	t-ratio	p-value

educ	-0.0905755	0.00592069	-15.30	1.68e-51 ***

- ▶ 2SLS (first manually, then jointly):

	coefficient	std. error	t/z-ratio	p-value

educ_hat2	-0.171499	0.0533921	-3.212	0.0013 ***

educ	-0.171499	0.0531796	-3.225	0.0013 ***

- ▶ Is the bias reduced by the IV?
- ▶ Are these results statistically different?

Hausman test

- ▶ Compares estimates from two different methods (1 and 2) which have the following properties:

	Method 1	Method 2
Under H_0	consistent efficient	consistents inefficient
Under H_A	inconsistent	consistent

- ▶ In the context of IV estimation (comparison of OLS and 2SLS results):

	OLS	2SLS
H_0 : no endogeneity	unbiased consistent efficient	biased consistent inefficient
H_A : endogeneity	biased inconsistent	some bias treated consistent

Hausman test

- ▶ Intuition:
 - ▶ under H_0 (no endogeneity), OLS and 2SLS should give statistically equal results as both consistent
 - ▶ under H_A (endogeneity), 2SLS is consistent whereas OLS is not, and so the two methods should give different results
- ▶ Tests whether the difference between the two (sets of) estimates is systematic (i.e. statistically different from 0)
- ▶ Hausman (Wald) test statistic:

$$H \text{ (or } W) = \left(\hat{\beta}^{2SLS} - \hat{\beta}^{OLS} \right)' \left(\text{Var} \left(\hat{\beta}^{2SLS} - \hat{\beta}^{OLS} \right) \right)^{-1} \left(\hat{\beta}^{2SLS} - \hat{\beta}^{OLS} \right)$$

- ▶ $H \sim \chi^2_{k+1}$, where k is the number of explanatory variables (+1 for the intercept)

Example: Hausman test in Gretl

Model 4: TSLS, using observations 1-4361

Dependent variable: children

Instrumented: educ

Instruments: const frsthalf age sq_age

	coefficient	std. error	z	p-value	
<hr/>					
const	-3.38781	0.548150	-6.180	6.39e-10	***
educ	-0.171499	0.0531796	-3.225	0.0013	***
age	0.323605	0.0178596	18.12	2.24e-73	***
sq_age	-0.00267228	0.000279687	-9.555	1.24e-21	***

...

Hausman test -

Null hypothesis: OLS estimates are consistent

Asymptotic test statistic: Chi-square(1) = 2.4501

with p-value = 0.117517

⇒ Hausman test would reject the H_0 at the 12% signif. level

[note: Gretl considers df equal to the number of endogenous variables by default]

Summary

- ▶ The OLS estimator of coefficients of endogenous variables is biased and inconsistent
- ▶ In which situations we may encounter endogenous variables:
 - ▶ omitted variable/influence bias
 - ▶ omitting important variable which is correlated with an included independent variable
 - ▶ an unobserved factor influencing both dependent and independent variable
 - ▶ simultaneity (joint determination \Rightarrow causality goes both directions)
 - ▶ measurement error (in either dependent or independent variable)
- ▶ We can deal with endogeneity by using IV (2SLS technique) and test for endogeneity by the Hausman test

Seminars and the next lecture #11

- ▶ **Seminars:**
 - ▶ IV and 2SLS practice in Gretl
- ▶ **Next lecture:**
 - ▶ we will discuss models with qualitative (binary) dependent variables
- ▶ Readings for lecture #11:
 - ▶ Studenmund (2016, [2014]): Chapter 13 [12]
 - ▶ Wooldridge (2016, 2014): Chapter 7-5, Chapter 17-1

Final exam reminder

- ▶ Online: Moodle part ⇒ oral part
- ▶ Registration in SIS begins on Monday, Jan 3, 2022, 8:00 a.m.
- ▶ See 'Requirements to the exam' in SIS
- ▶ See *Final_exam_ONLINE_info_JEM062_2021.pdf* in SIS
- ▶ See a 'Specimen final exam' in [Moodle](#) to have a better idea of what to expect

LECTURE #11

Introductory Econometrics

INTRODUCTION TO QUALITATIVE DEPENDENT VARIABLES

Jiri Kukacka, Ph.D.

Institute of Economic Studies, Faculty of Social Sciences, Charles University

Winter semester 2021, December 15

The last Q&A lecture #11 and seminars

- ▶ Next **lecture**:
 - ▶ revision: **Questions & Answers** ⇒ please **send** your questions or unclear topics in advance via **email**, **deadline Saturday 19**
 - ▶ I will try to aggregate them into 'topic blocks' to discuss
 - ▶ lecture slides will be distributed via SIS on Tuesday (evening)
- ▶ **Seminars** tomorrow:
 - ▶ **ONLINE ONLY**, delivered via Zoom (find the link in SIS) according to the standard schedule
 - ▶ LPM, Logit, and Probit practice in Gretl
- ▶ SIS: *Final_exam_ONLINE_info_JEM062_2021_UPDATED.pdf*
- ▶ Readings for lecture #11:
 - ▶ your favourite book :-) or selected chapters from Studenmund (2016 & 17) and Wooldridge (2016)

A small advert

- ▶ Did you like Introductory Econometrics?
- ▶ Then continue with us and attend [Applied Econometrics](#) (JEM116) during the Summer semester!
- ▶ Teachers: Horvath, Barunik, Baxa | SW: R, Jupyter Notebook
- ▶ Topics: IVs; time series analysis: ARIMA & GARCH; nonlinear models; cointegration; VAR; filters; QDV_s in finance

A kind request

- ▶ Please fill in the electronic evaluation (anonymous) of our course Introductory Econometrics (JEM062)



- ▶ You will be informed via multiple channels (email, web, etc.) when and where to log in

On previous lecture #10...

- ▶ We studied the topic of **endogeneity** (endogenous vs exogenous variables in the linear regression model) and discussed three main situations leading to this issue
- ▶ We showed that the OLS estimator of endogenous variables is biased and **inconsistent** (CA 3. violated)
- ▶ We defined the concept of **instrumental variables (IV)** and derive the **2SLS technique** to deal with endogeneity
- ▶ We described and practised the **Hausman test** for endogeneity

Motivation for today's lecture #11

- ▶ Until now, our discussion of dummy variables has been restricted to dummy explanatory variables
- ▶ However, there are many important research topics for which the dependent variable is appropriately treated as a dummy
- ▶ For any topic that involves discrete choice of some sort, the dependent variable is typically a dummy variable:
 - ▶ Will a woman participate in the labor market if she has kids?
YES/NO
 - ▶ Will the borrower be able to repay the debt in full on time?
YES/NO

On today's lecture, we will...

- ▶ Study **binary models** for which the outcome variable is:

$$Y_i = \begin{cases} 1 \\ 0 \end{cases},$$

depending on some qualitative choice

- ▶ Look at 3 most commonly used approaches:
 - ▶ Linear Probability Model
 - ▶ Logit Model
 - ▶ Probit Model
- ▶ Readings for this week:
 - ▶ Studenmund (2016, [2014]): Chapter 13 [12]
 - ▶ Wooldridge (2016, 2012): Chapter 7-5, Chapter 17-1

Probability distribution of Y_i

- Y_i is a discrete random variable with Bernoulli distribution:

$$Y_i = \begin{cases} 1 & \text{with probability } p_i \\ 0 & \text{with probability } 1 - p_i \end{cases}$$

- We can find the expected value:

$$E[Y_i] = 1 \cdot p_i + 0 \cdot (1 - p_i) = p_i$$

and the variance:

$$\begin{aligned} Var[Y_i] &= E[Y_i^2] - (E[Y_i])^2 = 1^2 \cdot p_i + 0^2 \cdot (1 - p_i) - p_i^2 \\ &= p_i - p_i^2 = p_i(1 - p_i) \end{aligned}$$

Linear Probability Model

- ▶ Running the usual OLS on an equation with dummy dependent variable:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i$$

- ▶ Why do we call it 'linear probability' model?
- ▶ Let us take the expected value:

$$\begin{aligned} E[Y_i] &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + E[\varepsilon_i] \\ p_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} \end{aligned}$$

- ▶ Hence, $p_i = \text{Prob}(Y_i = 1)$ is a linear function of explanatory variables

Problems with LPM

1. The error term is not normally distributed:

- ▶ because Y_i takes on only two values, the error term:

$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik})$$

also takes on only two possible values for given X_i , i.e., the error term is **binomial**

2. The error term is inherently heteroskedastic:

- ▶ we have :

$$\text{Var} [\varepsilon_i] = \text{Var} [Y_i] = p_i(1 - p_i),$$

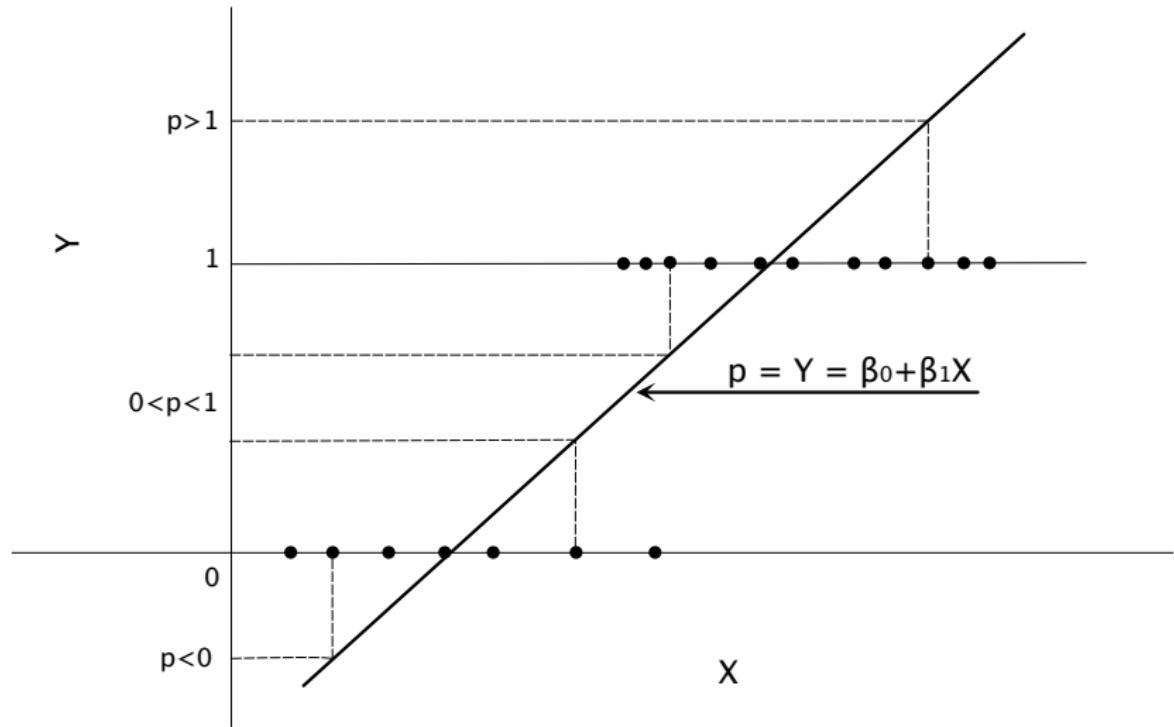
where $p_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}$

3. Constant marginal effect of Δx_i :

4. The predicted probability is not bounded by 0 and 1:

$$\hat{p}_i = \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_k x_{ik}$$

Problems with LPM



Example

- ▶ Based on *Instrumental Variables Methods in Experimental Criminological Research: What, Why, and How?* by Joshua Angrist (2006), in which the results from Minneapolis Domestic Violence Experiment (MDVE) are reexamined
- ▶ Will there be re-offense in cases of domestic violence if the offender is arrested on the spot? YES/NO?
- ▶ The author estimates the determinants of the re-offense status y for cases of domestic violence (y is dummy variable for cases when re-offense occurred)
- ▶ Main explanatory variable:

$$d_coddled = \begin{cases} 1 & \text{if the offender was not arrested} \\ 0 & \text{if the offender was arrested} \end{cases}$$

- ▶ Other controls: dummies indicating the presence drugs and weapons, race dummies

Example: OLS (with White HC robust standard errors)

Model 1: OLS, using observations 1-330

Dependent variable: y

Heteroskedasticity-robust standard errors, variant HC1

	coefficient	std. error	t-ratio	p-value	
<hr/>					
const	0.0901995	0.0511667	1.763	0.0789	*
d_coddled	0.0873254	0.0410044	2.130	0.0340	**
drugs	0.0479707	0.0437274	1.097	0.2734	
weapon	0.0113562	0.0480876	0.2362	0.8135	
nonwhite	-0.0274346	0.0425991	-0.6440	0.5200	
mixed	0.0740200	0.0518510	1.428	0.1544	
<hr/>					
Mean dependent var	0.181818	S.D. dependent var	0.386280		
Sum squared resid	47.91695	S.E. of regression	0.384567		
R-squared	0.023914	Adjusted R-squared	0.008851		
F(5, 324)	1.542048	P-value(F)	0.176298		

Latent (hidden) variable approach

- ▶ Leads to derivation of Logit and Probit models
- ▶ Suppose we have a continuous variable y_i^* (called *latent variable*), following:

$$y_i^* = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i \quad (1)$$

and the relationship (*indicator function*):

$$Y_i = \begin{cases} 1 & \text{for } y_i^* > 0 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

- ▶ Equations (1) and (2) together define the binary model
- ▶ Underlying heuristic: the value of the qualitative dependent variable depends on a choice based on a latent continuous utility and a simple decision rule

Latent variable approach

- Let us express the probability that $Y_i = 1$ under this approach:

$$\begin{aligned} p_i &= \text{Prob}(Y_i = 1) = \text{Prob}(y_i^* > 0) \\ &= \text{Prob}(\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \varepsilon_i > 0) \\ &= \text{Prob}(\varepsilon_i > -\beta_0 - \beta_1 x_{i1} - \dots - \beta_k x_{ik}) \\ &= 1 - \text{Prob}(\varepsilon_i \leq -\beta_0 - \beta_1 x_{i1} - \dots - \beta_k x_{ik}) \\ &= 1 - F(-\beta_0 - \beta_1 x_{i1} - \dots - \beta_k x_{ik}), \end{aligned}$$

where $F(\cdot)$ denotes the cumulative distribution function of the error term ε_i .

Possible probability distributions of the error term

- Standard normal:

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}$$

$$F(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{t^2}{2}\right\} dt$$

- Logistic:

$$f(x) = \frac{\exp(-x)}{(1 + \exp(-x))^2}$$

$$F(x) = \frac{1}{1 + \exp(-x)}$$

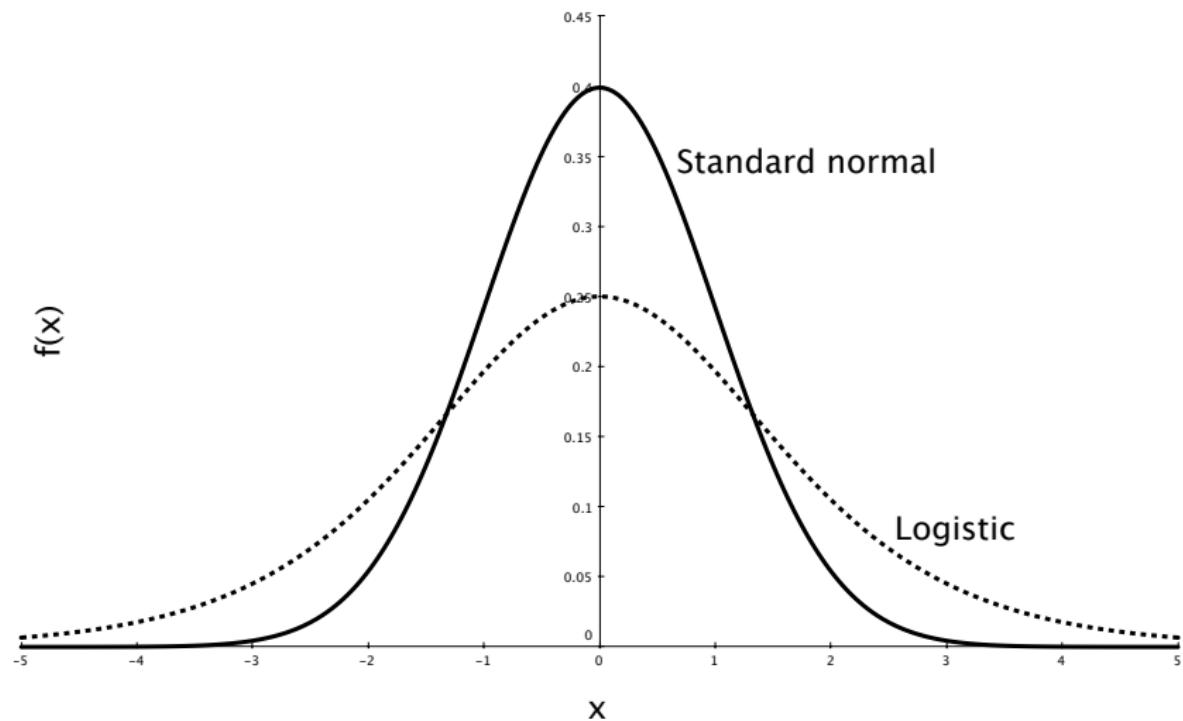
- Both distributions have the following property:

$$1 - F(-x) = F(x)$$

- This allows us to write:

$$\begin{aligned} p_i = \text{Prob}(Y_i = 1) &= 1 - F(-\beta_0 - \beta_1 x_{i1} - \dots - \beta_k x_{ik}) \\ &= F(\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}) \end{aligned}$$

Possible probability distributions



Probit and Logit models

- Both models define the probability of $Y_i = 1$ as a function of explanatory variables:

$$p_i = \text{Prob}(Y_i = 1) = F(\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}),$$

where $F(\cdot)$ denotes the cumulative distribution function (CDF) of the error term ε_i

- Probit model uses the standard normal CDF
- Logit model uses the logistic CDF
- Parameters $\beta_0, \beta_1, \dots, \beta_k$ are estimated by the Maximum Likelihood Estimator (MLE)

▶ MLE derivation

Comparison of the models

- ▶ In the LPM model, we had:

$$\hat{p}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_k x_{ik},$$

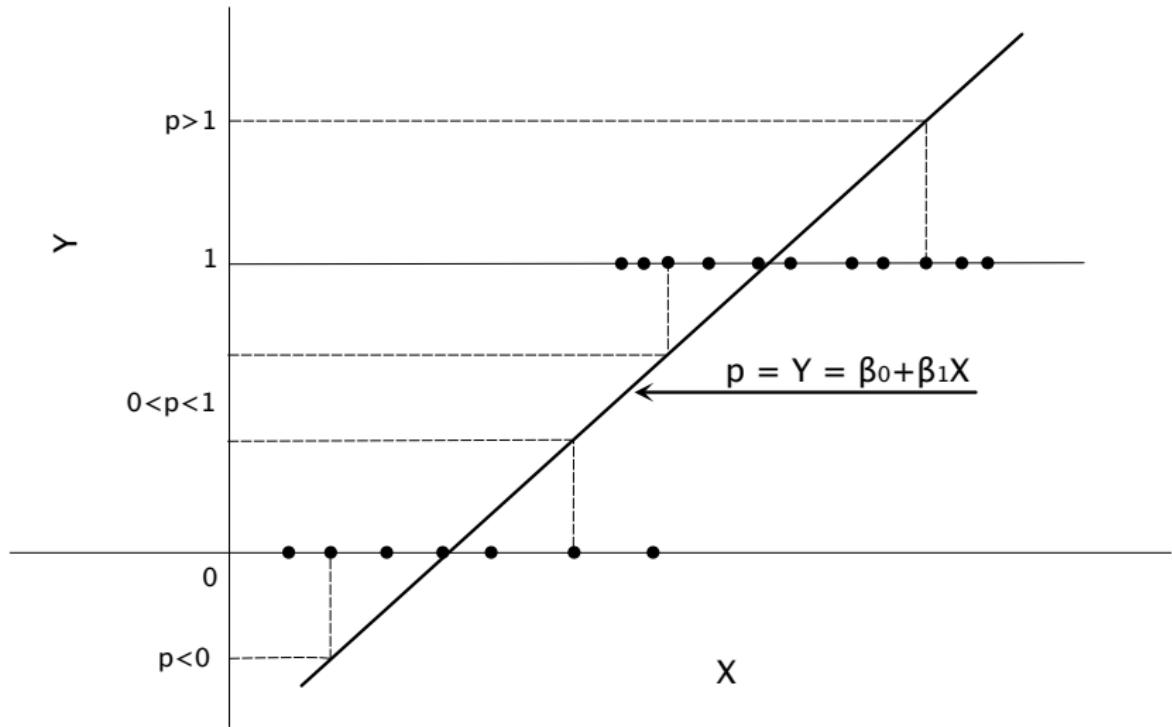
which was not bounded by 0 and 1

- ▶ In the Logit and Probit models, we have:

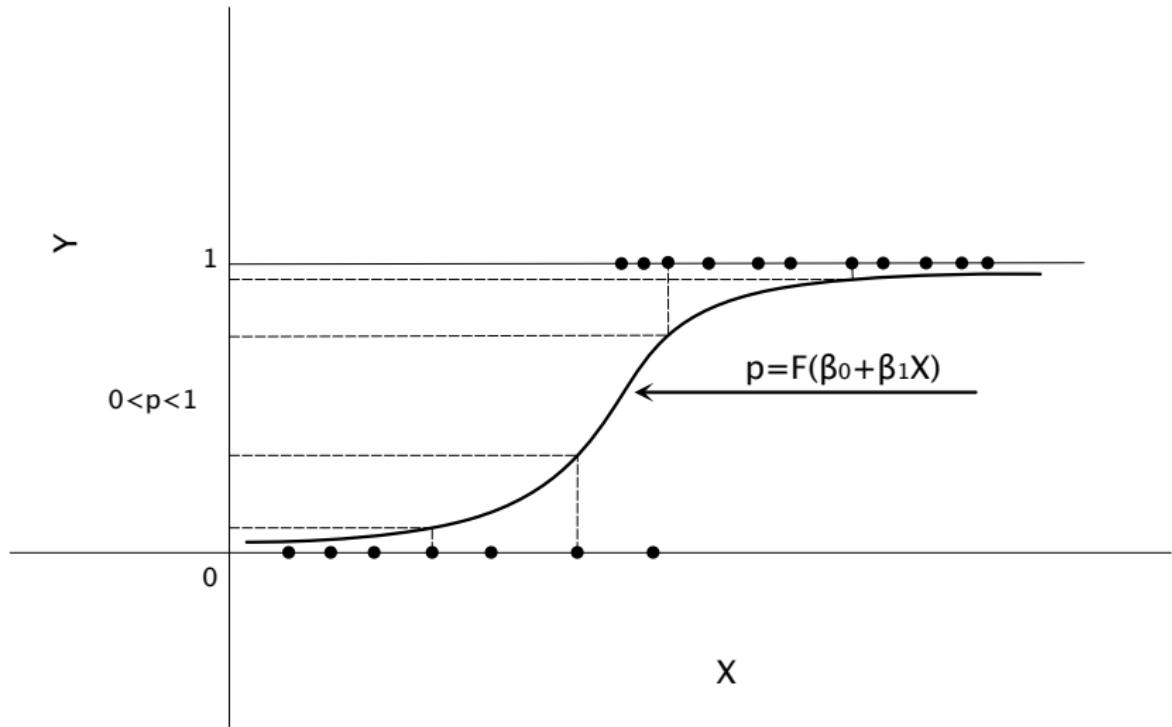
$$\hat{p}_i = F(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_k x_{ik}),$$

which is bounded by 0 and 1 thanks to the properties of a cumulative distribution function

Linear Probability Model



Logit/Probit



Interpretation

- In the LPM model, we had:

$$\hat{p}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_k x_{ik},$$

which gave a simple interpretation of the coefficients:

$$\frac{\partial \hat{p}_i}{\partial x_{ij}} = \hat{\beta}_j$$

- In the Logit and Probit models, we have:

$$\hat{p}_i = F(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_k x_{ik}),$$

which gives:

$$\frac{\partial \hat{p}_i}{\partial x_{ij}} = f(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_k x_{ik}) \cdot \hat{\beta}_j$$

Marginal effects

- More than in estimated coefficients $\hat{\beta}_j$, we are interested in marginal effects of the explanatory variables on the probability of $Y_i = 1$:

$$\frac{\partial \hat{p}_i}{\partial x_{ij}} = f(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_k x_{ik}) \cdot \hat{\beta}_j$$

- In order to obtain an average marginal effect, the function $f(\cdot)$ in this expression is usually evaluated at the mean of observations:

$$\frac{\partial \hat{p}}{\partial x_j} = f(\hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \dots + \hat{\beta}_k \bar{x}_k) \cdot \hat{\beta}_j$$

Back to the example: Logit

Model 2: Logit, using observations 1-330

Dependent variable: y

Standard errors based on Hessian

	coefficient	std. error	z	p-value	

const	-2.18996	0.399820	-5.477	4.32e-08	***
d_coddled	0.623532	0.315123	1.979	0.0479	**
drugs	0.333920	0.307037	1.088	0.2768	
weapon	0.0745484	0.332376	0.2243	0.8225	
nonwhite	-0.194676	0.301318	-0.6461	0.5182	
mixed	0.473299	0.315932	1.498	0.1341	

Mean dependent var 0.181818 S.D. dependent var 0.386280

McFadden R-squared 0.025463 Adjusted R-squared -0.012884

Log-likelihood -152.4819

...

Likelihood ratio test: Chi-square(5) = 7.96819 [0.1580]

Back to the example: AVG marginal effect after Logit

Model 3: Logit, using observations 1-330

Dependent variable: y

Standard errors based on Hessian

	coefficient	std. error	z	slope

const	-2.18996	0.399820	-5.477	
d_coddled	0.623532	0.315123	1.979	0.0867072
drugs	0.333920	0.307037	1.088	0.0468831
weapon	0.0745484	0.332376	0.2243	0.0108449
nonwhite	-0.194676	0.301318	-0.6461	-0.0277203
mixed	0.473299	0.315932	1.498	0.0731195

Mean dependent var 0.181818 S.D. dependent var 0.386280
McFadden R-squared 0.025463 Adjusted R-squared -0.012884
Log-likelihood -152.4819

...

Likelihood ratio test: Chi-square(5) = 7.96819 [0.1580]

Daniel McFadden, 2000 Nobel Laureate...

- ▶ ...in Economic Sciences “for his development of theory and methods for analyzing discrete choice”
- ▶ In 1974 he introduced the Conditional logit analysis

Back to the example: Probit

Model 4: Probit, using observations 1-330

Dependent variable: y

Standard errors based on Hessian

	coefficient	std. error	z	p-value	

const	-1.28198	0.218259	-5.874	4.26e-09	***
d_coddled	0.349425	0.174955	1.997	0.0458	**
drugs	0.183915	0.171969	1.069	0.2849	
weapon	0.0445090	0.188513	0.2361	0.8134	
nonwhite	-0.110622	0.168797	-0.6554	0.5122	
mixed	0.256326	0.182690	1.403	0.1606	

Mean dependent var 0.181818 S.D. dependent var 0.386280

McFadden R-squared 0.025050 Adjusted R-squared -0.013297

Log-likelihood -152.5465

...

Likelihood ratio test: Chi-square(5) = 7.839 [0.1653]

Back to the example: AVG marginal effects after Probit

Model 5: Probit, using observations 1-330

Dependent variable: y

Standard errors based on Hessian

	coefficient	std. error	z	slope

const	-1.28198	0.218259	-5.874	
d_coddled	0.349425	0.174955	1.997	0.0877187
drugs	0.183915	0.171969	1.069	0.0466293
weapon	0.0445090	0.188513	0.2361	0.0116215
nonwhite	-0.110622	0.168797	-0.6554	-0.0283666
mixed	0.256326	0.182690	1.403	0.0699435

Mean dependent var	0.181818	S.D. dependent var	0.386280
McFadden R-squared	0.025050	Adjusted R-squared	-0.013297
Log-likelihood	-152.5465		

...

Likelihood ratio test: Chi-square(5) = 7.839 [0.1653]

Back to the example: Comparison

- ▶ LPM:

	coefficient	HC std. error	t-ratio	p-value	
<hr/>					
d_coddled	0.0873254	0.0410044	2.130	0.0340	**

- ▶ Logit (marginal effect):

	coefficient	std. error	z	slope
<hr/>				
d_coddled	0.623532	0.315123	1.979	0.0867072

- ▶ Probit (marginal effect):

	coefficient	std. error	z	slope
<hr/>				
d_coddled	0.349425	0.174955	1.997	0.0877187

Summary

- ▶ We presented models with binary dependent variables
- ▶ We discussed the Linear Probability Model (usual OLS):
 - ▶ advantages: simple linear interpretation, often works well
 - ▶ weaknesses: **unbounded probability**, non-normal and heteroskedastic error term
- ▶ We derived the Logit and Probit models under the latent variable approach:
 - ▶ advantages: **probability bounded by 0 and 1**, non-constant marginal effect
 - ▶ weaknesses: interpretation not straightforward
- ▶ We showed that the models **can** give analogous results:
 - ▶ but mostly only around the average values of **X**
 - ▶ this is why some authors prefer LMP over Logit or Probit

Appendix: MLE (not mandatory)

- ▶ The principle of the Maximum Likelihood Estimator (MLE) is to maximize the likelihood function L as a function of the parameters which are to be estimated
- ▶ The likelihood function represents the probability of the sample as we observe it
- ▶ For binary models with n observations, it looks as:

$$L = \prod_{i=1}^n p_i^{Y_i} (1 - p_i)^{(1 - Y_i)}$$

with:

$$p_i = \text{Prob}(Y_i = 1) = F(\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik})$$

Appendix: MLE (not mandatory)

- The MLE estimates of $\beta_0, \beta_1, \dots, \beta_k$ are such that they maximize the logarithm of the likelihood function:

$$\ln L = \sum_{i=1}^n Y_i \ln p_i + (1 - Y_i) \ln(1 - p_i)$$

with:

$$p_i = \text{Prob}(Y_i = 1) = F(\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik})$$

- The choice of $F(\cdot)$ depends on whether we use Probit or Logit model
- MLE estimator for both models is consistent and efficient under the condition that the choice of $F(x)$ is correct (very limiting!)
- Readings about MLE in Wooldridge (2016): Appendix 17A, Appendix C-4b

▶ Back