Introduction to Linear Regression Analysis

Samuel Nocito

Collegio Carlo Alberto
UNIVERSITÀ DEGLI STUDI DI TORINO

Lecture 1

March 2nd, 2018

Co-funded by the Erasmus+ Programme of the European Union

Migration in Europe
MigrEU Jean Monnet Module

Econometrics: What is it?

- Interaction of economic theory, observed data and statistical methods.
- The science of testing economic theory.
- The application of statistical techniques for solving empirical problems.
- The set of tools used either for predicting future variables (prices, demographic trends, etc.) or for phenomenon estimation.
- The science of using data to make quantitative inference for policy recommendations.

Econometrics: Why do we need it?

- Is there gender discrimination in the labor market (wage gender gap)?
- How much can "carbon tax" reduce the use of fossil fuels?
- Is there racial discrimination in the market for home loans?
- What is the economic return of education?
- What will the life expectancy at birth be in the next 20 years?

Co-funded by the Erasmus+ Programme of the European Union

Migration in Europe
MigrEU Jean Monnet Module

Migration Topics Addressed by Econometrics

Broad questions:

(A) Who chooses to migrate?

- Impact of personal characteristics.

(B) Why do people migrate to different countries?

- Push and pull factors.

(C) What is the impact of emigration?

- Effect on the country of origin.

(D) What is the impact of immigration?

- Effect on the host country.

Co-funded by the Erasmus+ Programme of the European Union

Migration in Europe
MigrEU Jean Monnet Module

Migration Topics Addressed by Econometrics

Specific questions (examples):

- Does foreign language proficiency foster migration of young individual within the European Union? (Aparicio Fenoll and Kuehn, 2016)
  ⇒ Point (A) "broad questions".
- Do immigrants cause crime? (Bianchi et al., 2012)
  ⇒ Point (D) "broad questions".

Co-funded by the Erasmus+ Programme of the European Union

Migration in Europe MigrEU Jean Monnet Module

Principal Econometrics Methods

- Linear Regression model: Ordinary Least Squares (OLS)
- Non Linear Regression Models:
  - Maximum Likelihood Estimation (MLE)
  - Probit, Logit, Tobit
- Differences-in-Differences
- Instrumental Variable Estimation (IV)

Co-funded by the
Erasmus+ Programme
of the European Union

Migration in Europe
MigrEU Jean Monnet Module

Principal Econometrics Methods in the Literature

	1995-1999	2000-2004	2005-2009
Number of papers	31	40	51
By empirical technique			
OLS	14	11	20
MLE, Probit, Logit, Tobit	3	9	9
Differences-in-Differences	1	2	0
Instrumental Variable	4	12	8
Others	9	6	14
By topic			
Assimilation	14	17	14
Immigrants selection	6	7	8
Native outcome	8	9	12
Others	3	7	12

American Economic Review, Quarterly Journal of Economics, Journal of Political Economy, Journal of Labour Economics, and others top journals.
Source: Sona Kalataryan, Methodological Workshop, MPC (EUI) 2016.

Principal Econometrics Methods: We focus on

- Ordinary Least Squares (OLS)
  - Simple mathematical and graphical explanation
  - Practical examples
  - Interpretation of results
- Instrumental Variable (IV)
  - Very short introduction on the topic
  - Correlation vs causality
  - Interpretation of results (OLS vs IV)
  - Tackled in lecture 2

Co-funded by the Erasmus+ Programme of the European Union

Migration in Europe
MigrEU Jean Monnet Module

Ordinary Least Squares (OLS)

Suppose we have a sample of N observations on individual wages and personal characteristics:

	y	X	
i	Wage	Age	Gender
1	6	18	M
2	5	18	F
3	5.8	20	F
.	.	.	.
N	6.9	22	M

US National Longitudinal Survey (NLS) of 1987 (Example).

N=3294 young working individuals, 1569 females.

Hourly wage rates. Males average 6.31, females 5.15.

We want to answer:

how in this sample wages are related to other observables?

Ordinary Least Squares (OLS)

OLS general equation:

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

In our empirical case:

$$Wage_i = \beta_0 + \beta_1 ttender_i + \varepsilon_i$$

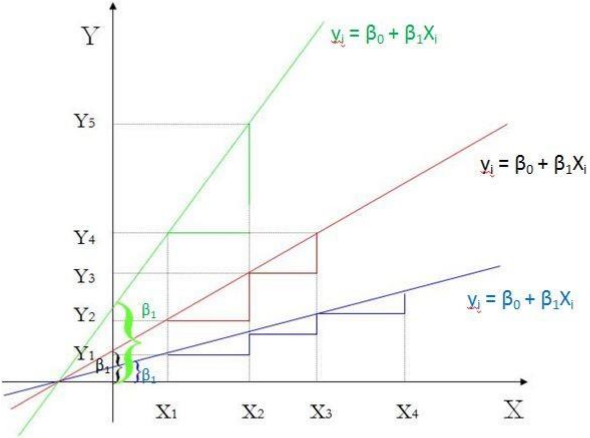
Where:

- y_i (individual wage): dependent variable (explained)
- x_i (gender): independent variable (explanatory)
- ε_i: is the error term

Ordinary Least Squares (OLS)

$y_i = \beta_0 + \beta_1 X_i$ is a linear equation model where

- β_0 is the intercept of the curve
- β_1 is the slope of the curve



Y
$y_i = \beta_0 + \beta_1 X_i$
Y5
$y_i = \beta_0 + \beta_1 X_i$
Y4
Y3
$y_i = \beta_0 + \beta_1 X_i$
Y2
β_1
Y1
β_1
β_1
X1
X2
X3
X4
X

Co-funded by the Erasmus+ Programme of the European Union

Migration in Europe
MigrEU Jean Monnet Module

Ordinary Least Squares (OLS)

In the empirical case:

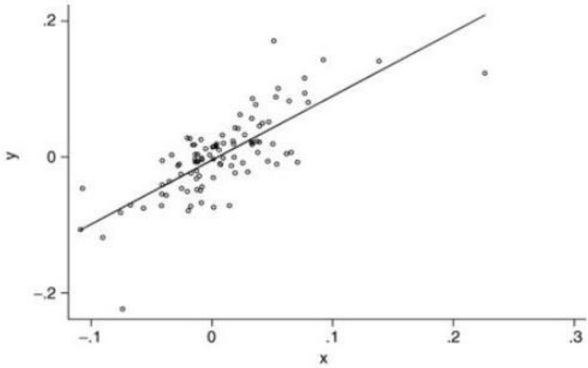


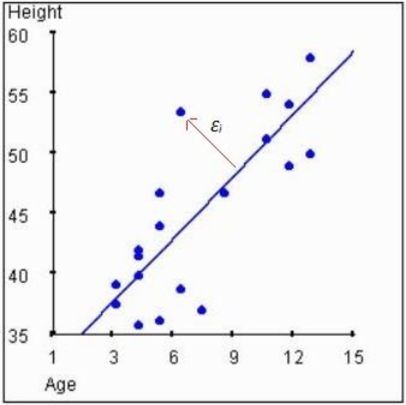
Figure: Fitted line and observation points (Verbeek, Fig. 2.1)

Co-funded by the Erasmus+ Programme of the European Union

Migration in Europe
MigrEU Jean Monnet Module

Ordinary Least Squares (OLS)

Figure: Linear Regression Example: Height and Age (months)



- blue dots: observed data (combinations of height and age).
- blue line: OLS linear equation.
- red arrow: error term ε_i.

Co-funded by the Erasmus+ Programme of the European Union

Migration in Europe
MigrEU Jean Monnet Module

Ordinary Least Squares (OLS)

- We observe x and y.
- We want to estimate β_0 and β_1 to understand the relation between x and y.
- The distance between the dot and the line is the error term ε_i of the OLS.
- We want to minimize the error term.

Ordinary Least Squares (OLS)

Formally:

$$y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad \Leftrightarrow \quad \varepsilon_i = y_i - \beta_0 - \beta_1 X_i$$

where ε_i is the error term.

In particular we want to minimize:

$$\sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 X_i)^2$$

Remark: we use the quadratic transformation to avoid issues with the sign of the error term.

Co-funded by the Erasmus+ Programme of the European Union

Migration in Europe
MigrEU Jean Monnet Module

Ordinary Least Squares (OLS)

In the case with one regressor (i.e., gender) and a constant., the solutions of β_0 and β_1 that minimize the error are:

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\beta_1 = \frac{Cov(x, y)}{Var(x)}$$

Where:

- $\bar{y}$ is the sample average of the y_i.
- $\bar{x}$ is the sample average of the x_i.
- $Cov(x, y)$ is the sample covariance between x and y.
- $Var(x)$ is the sample variance of x.

The intercept (β_0) is determined to make the average error equal to zero.

OLS: Application to the Wage Example

We create the variable *Male* using the information of gender (dummy variable).

	y	X		
i	Wage	Age	Gender	Male
1	6	18	M	1
2	5	18	F	0
3	5.8	20	F	0
.	.	.	.	.
N	6.9	22	M	1

We use OLS to estimate:

$$Wage_i = \beta_0 + \beta_1 Male_i + \varepsilon_i$$

OLS: Application to the Wage Example

Table: OLS results wage equation (Verbeek, tab. 2.1)

Dependent variable: wage		
Variable	Estimate	Standard Error
Constant	5.1469	0.0812
Male	1.1661	0.1122
	$R^2 = 0.0317$	F=107.93

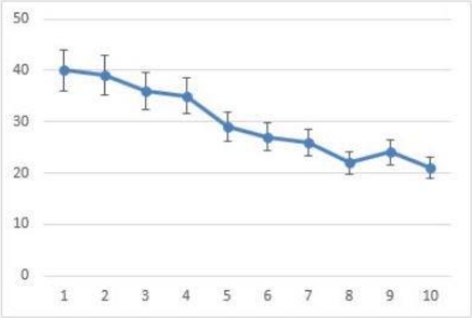
$$Wage_i = 5.15 + 1.17 Male_i$$

$$\beta_0 = 5.15 \text{ and } \beta_1 = 1.17$$

- $\beta_1 = 1.17$ means that males receive 1.17 dollar per hour more than females.
- Standard errors show the error in the estimate of the coefficient (the smaller the better!).
- $R^2 = 0.0317$ means that approximately 3.2% of the variation in individual wages is given to gender differences.

OLS: Application to the Wage Example

Figure: Graphical Representation of the Standard Errors (example)



Suppose each dot is a coefficient estimate:

- The standard error shows the interval in which the coefficient lies.
- The smaller is the interval the higher is the precision of the estimate.

Co-funded by the Erasmus+ Programme of the European Union
Migration in Europe
MigrEU Jean Monnet Module

Lecture 2 in Sketches

- Dependent Variable and Explanatory variables
  - How to interpret coefficient estimates with different variable definitions.
  - Analysis of an empirical paper results.
  - OLS issues.
- Correlation vs causality
  - Short introduction to IV estimates (conceptual).
  - Comparison of results (OLS vs IV) of an empirical paper.

Co-funded by the Erasmus+ Programme of the European Union

Migration in Europe

MigrEU Jean Monnet Module

References

- Marno Verbeek, A Guide to Modern Econometrics, 3^{rd} Ed., Wiley, 2008, Chapter 2, pp. 6-31.

- Suggested (not used in class):
  - Stock, James H., and Mark W. Watson, Introduction to Econometrics, Global Edition, MA: Pearson Education, 2012.

Co-funded by the Erasmus+ Programme of the European Union

Migration in Europe

MigrEU Jean Monnet Module