Introduction to Linear Regression Analysis

Interpretation of Results

Samuel Nocito

Collegio Carlo Alberto

UNIVERSITÀ DEGLI STUDI DI TORINO

Lecture 2

March 8th, 2018

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

Migration in Europe

MigrEU Jean Monnet Module

Lecture 1 Summary

- Why and how we use econometric tools in empirical research.
- Ordinary Least Square (OLS) estimation method
  - simple theoretical framework;
  - graphical representation;
  - coe cient estimation in the simple case with one regressor (little algebra!);
  - practical example using NLS data on wages.

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

OLS: Dependent and Explanatory Variables

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

where:

- y_i dependent variable (explained, response or predicted variable);
- x_i independent variable (explanatory, control or predictor variable).
- ε_i is the error term.

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

OLS: Definition of the Variables

Either dependent or independent variables can be:

- CONTINUOUS y_i^c (or x_i^c) taking any real value;
- DUMMY y_i^d (or x_i^d) taking values 1 (if yes) and 0 (if no) (e.g., variable *Male* of the wage example);
- LOGARITHMIC $ln(y_i)$ (or $ln(x_i)$) simply the natural logarithm of a continuous variable.

The interpretation of the coefficient estimates changes according to the combination of these types of variables.

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

OLS Coe cient Interpretation: Continuous Dep. Variable

Model A: continuous dependent variable.

$$y_i^c = \beta_0 + \beta_1 x^c_{1i} + \beta_2 ln(x_{2i}) + \beta_3 x^d_{3i} + \varepsilon_i$$

- β_1 = a one unit change in x^c_1 generates a β_1 unit change in y_i^c.
- β_2 = a 100% change in x_{2i} generates a β_2 change in y^c_i.
- β_3 = the movement of x^d_{3i} from 0 to 1 produces a β_3 unit change in y_i^c.

OLS Coe cient Interpretation: Dummy Dep. Variable

Model B: dummy dependent variable.

$$y_i^d = \beta_0 + \beta_1 x_{1i}^c + \beta_2 ln(x_{2i}) + \beta_3 x_{3i}^d + \varepsilon_i$$

- β_1 = a one unit change in x_{1i}^c generates a $100\beta_1$ percentage points change in the probability y_i^d occurs.
- β_2 = a 100% change in x_{2i} generates a $100\beta_2$ percentage points change in the probability y_i^d occurs.
- β_3 = the movement of x_{3i}^d from 0 to 1 produces a $100\beta_3$ percentage points change in the probability y_i^d occurs.

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

OLS Coe cient Interpretation: log Dep. Variable

Model C: logarithm dependent variable.

$$ln(y_i) = \beta_0 + \beta_1 x^c_{1i} + \beta_2 ln(x_{2i}) + \beta_3 x^d_{3i} + \varepsilon_i$$

- β_1 = a one unit change in x^c_1 generates a $100\beta_1$ percent change in y_i.
- β_2 = a 100% change in x_{2i} generates a $100\beta_2$ percent change in y_i.
- β_3 = the movement of x^d_{3i} from 0 to 1 produces a $100\beta_3$ percent change in y_i.

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

OLS Coe cient Interpretation: Wage Example

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

This is a model of type A ⇒ continuous dep. variable and β_1 refers to a DUMMY explanatory variable (*Male*).

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

$$Wage_i = 5.15 + 1.17 Male_i$$

- β_1= the movement of *Male* from 0 to 1 produces a β_1 (1.17) unit change in $Wage_i$.

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

Types of Data

There are four di erent types of data:

- **Cross-sectional**: sample of observations taken at a given point in time.
- **Time series**: observations on a variable or several variables over time.
- **Pooled cross-sectional**: di erent random samples are asked the same questions over time.
- **Panel (or longitudinal)**: consists of a time series on same individuals (i.e., ask to Sarah the same question in two di erent years).

Coe cient Interpretation in the Literature: Example 1

- Does foreign language pro ciency foster migration of young individual within the European Union? (Aparicio Fenoll and Kuehn, 2016)

Model equation (of type A):

$$M_{a,o,d,t} = \beta_0 + \beta_1 L_{a,o,d,t} + \ldots + \varepsilon_{a,o,d,t}$$

- M: number of immigrants of age a from country o to d in year t.
- L: exposure to compulsory language courses in the o cial language of country d.
- Other controls (i.e., dummies and predetermined controls as unemployment rate).

Coe cient Interpretation in the Literature: Example 1

Figure: Results (Aparicio Fenoll and Kuehn, Tab 4.3)

	(1)	(2)	(3)	(4)
treated	813.91	521.079	523.899	544.316
	(339.438)**	(236.434)**	(260.825)**	(273.013)**
Destination by age		X	X	X
Destination by year		X	X	X
Origin by year		X	X	X
Origin by age		X	X	X
Destination by origin by year			X	X
Destination by age by year				X
Obs.	6784	6784	6784	6784
R^2	0.762	0.843	0.868	0.872

The dependent variable is the number of immigrants, the variable treated identifies the cohorts from the country of origin who were exposed to learning the language of the country of destination during compulsory schooling. The coefficients are marked with * if the level of significance is between 5% and 10%, ** if the level of significance is between 1% and 5% and *** if the level of significance is less than 1%. All regressions contain year-fixed effects, age indicators, binary variables for each pair of origin and destination countries, dummies for each combination of age and year, a variable for differences in lagged unemployment rate between origin and destination countries and the stock of co-nationals from each cohort in the destination country in the previous period. Errors are clustered by origin-destination-age.

"Exposure to language learning during compulsory education increases the number of individuals of a cohort that migrate to the country where the language is spoken by 544 per year, a magnitude similar to the standard deviation of the number of immigrants in the sample."

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

Migration in Europe
MigrEU Jean Monnet Module

OLS Endogeneity Issues

Endogeneity occurs whenever the explanatory variable (regressor) is correlated with the error term.

Endogeneity conditions:

- Measurement error: error made in measuring the dependent or the explanatory variable.

  Example: wages is an information that people not always want to provide. Di cult to measure the sample information $\Rightarrow$ data itself correlated with the error.

OLS Endogeneity Issues

Endogeneity conditions:

- Reverse causality: $x \Rightarrow y$(what we look for), $y \Rightarrow x$ (reverse causality), or $y \Leftrightarrow x$ (simultaneity).

  Example (police and crime): increased police force might cause a reduction in crime, however an increase/decrease in crime might cause an increase/decrease in policeman number.

- Omitted variable: some unobservable variables a ecting both $y$and x.

  Example: ability a ects both education and wages $\Rightarrow$ return on education is a di cult question.

OLS results often a ected by endogeneity.
Infer causality with OLS is hard and rare.

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

Correlation vs Causality

- Correlation is a statistical measure describing the size and the direction of a relationship between two or more variables.
- Causality indicates that one event is the result of the occurrence of the other event.[1]

Example 1: Smoking might be correlated with alcoholism but it is not a cause of it.
Example 2: Immigration might be correlated to the total level of crime in a speci c region or province, however it is not a direct cause of it (see next example).

- Causality is compromised by endogeneity
  ⇒ other driven factors a ecting the choice.

[1] Australian Bureau of Statistics.

Instrumental Variable (IV): basic concept

$$Crime_p = \beta_0 + \beta_1 Immigrants_p + \varepsilon_p$$

Suppose we want to measure the impact of immigrants on crime at province (p) level.

- The choice of migrating in a particular province is endogenous. $\Rightarrow$ we can see only correlation.
- We can use an Instrumental Variable to investigate causality.

The Instrument must be:

- Assumption 1: (strongly) correlated with the endogenous variable.
- Assumption 2: independent of y (exogenous).
- Assumption 3: built to a ect all the treated in the same way.

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

Coe cient Interpretation in the Literature: Example 2

- Do immigrants cause crime?
  (Bianchi, M., Buonanno, P. and Pinotti, P., 2008)

Endonegeity: e.g., lower housing prices, improvements in labour market conditions as driven factors for migration (endogenous at provincial level).

OLS provides only correlation.

Instrument: (exogenous) supply-push component of migration (i.e., economic crisis, political turmoil, wars and natural disaster in the country of origin).

- The instrument satis es all the assumptions.

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

Coe cient Interpretation in the Literature: Example 2 (OLS)

Figure: OLS Results (Bianchi et al., Tab 3)

	(1) *total*	(2) *violent*	(3) *property*	(4) *drug*	(5) *robbery*	(6) *theft*	(7) *car theft*
migr	0.102***	0.003	0.084***	−.103	0.092*	0.093***	0.057
	(0.039)	(0.084)	(0.028)	(0.074)	(0.05)	(0.03)	(0.041)
pop	0.028	-0.338	0.96	-2.550	4.285***	1.155*	0.365
	(0.641)	(1.660)	(0.718)	(1.552)	(1.026)	(0.686)	(0.958)
urban	0.003*	−0.003	0.003	−0.010***	0.0007	0.004	0.004**
	(0.002)	(0.003)	(0.003)	(0.002)	(0.004)	(0.002)	(0.002)
*male*1539	0.131***	0.236**	0.041	0.325***	-0.145*	0.052	0.1
	(0.045)	(0.11)	(0.053)	(0.108)	(0.084)	(0.053)	(0.072)
gdp	0.15	-0.116	0.171	0.423	-0.155	0.113	0.611***
	(0.14)	(0.319)	(0.166)	(0.378)	(0.267)	(0.164)	(0.232)
unemp	-0.004	0.011	-0.007*	0.019*	-0.022***	-0.006*	−.003
	(0.007)	(0.003)	(0.005)	(0.004)	(0.01)	(0.003)	(0.01)
clear-up	-0.004	-0.008***	-0.030***	0.0003	-0.005***	-0.030***	-0.005**
	(0.003)	(0.002)	(0.006)	(0.003)	(0.001)	(0.006)	(0.003)
partisan	0.007	0.045**	0.007	0.023	0.006	0.007	-0.003
	(0.01)	(0.019)	(0.009)	(0.015)	(0.013)	(0.009)	(0.011)
Obs.	1,045	1,045	1,045	1,045	1,045	1,045	1,045
Provinces	95	95	95	95	95	95	95
Prov. FE	yes	yes	yes	yes	yes	yes	yes
Year FE	yes	yes	yes	yes	yes	yes	yes
R^2	0.220	0.321	0.302	0.189	0.241	0.28	0.323
F-stat.	14.81	7.37	11.68	17.26	14.17	9.77	14.72

Notes: This table presents the results of OLS estimates on a panel of yearly observations for all 95 Italian provinces during the period 1991–2003. The dependent variable is the log of crimes reported by the police over the total population, for each category of criminal offense. The variable *migr* is the log of immigrants (i.e. residence permits) over province population. The sources of data for residence permits and reported crimes are ISTAT and the Italian Ministry of the Interior, respectively. All other variables are defined in Appendix A. Province and year fixed effects are included in all specifications. Robust standard errors are presented in parentheses.
*, ** and *** denote rejection of the null hypothesis of the coefficient being equal to 0 at 10%, 5% and 1% significance level, respectively.

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

Coe cient Interpretation in the Literature: Example 2 (IV)

Figure: OLS vs IV Results (Bianchi et al., Tab 4)

TABLE 4. Ten-year difference regressions: total crimes.

	(1) OLS	(2) OLS	(3) IV	(4) IV	(5) OLS	(6) IV	(7) IV
$\Delta migr$	.156*** (.049)		.105 (.187)	.029 (.125)		.055 (.244)	−.029 (.175)
$\widehat{\Delta migr}$		.055 (.104)					
$\Delta \widetilde{migr}$					.137*** (.039)		
Obs.	95	95	95	95	95	95	95
F statistic	5.401	3.095	3.395	3.243	6.399	3.283	3.001
R^2	.241	.182			.249		

- Total crime is not related to the size of immigrants (IV).
- NO statistically signi cant result in the IV.
- POSITIVE and statistically signi cant correlation. NO causality e ect.

Summary

- OLS as a tool to answer economic questions.
- OLS implies correlation but not always causality.
- IV can infer causality under certain assumptions.
- The variable types (log, dummy, etc.) determine the coe cient interpretation.
- Standard errors show the magnitude of the estimation error (the smaller the better!).
- Statistic signi cance (stars!) to see if the estimated coe cient is statistically signi cantly di erent from 0.
- R^2 is the fraction of the sample variation in y that is explained by x.

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

Migration in Europe
MigrEU Jean Monnet Module

References

- APARICIO FENOLL, Ainhoa; KUEHN, Zoë. Does foreign language pro ciency foster migration of young individuals within the European Union. The economics of language policy, 2016, 331-355.
- BIANCHI, Milo; BUONANNO, Paolo; PINOTTI, Paolo. Do immigrants cause crime?. Journal of the European Economic Association, 2012, 10.6: 1318-1347.

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