Introduction to Linear Regression

Salvatore Carrozzo

University of Turin

Migration in Europe
MigrEU Jean Monnet Module

Outline

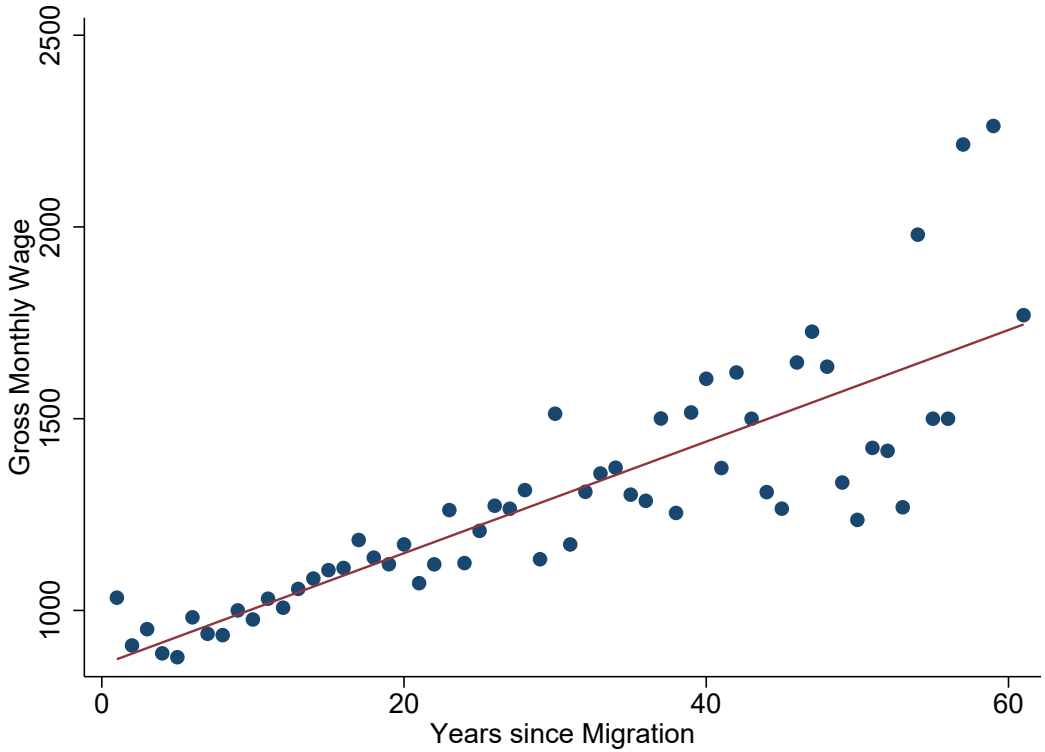
Introduction 1/2

What is a linear regression?

Linear regression aims to study the **effect**, if any, of a change in one or more independent variables (explanatory) on a dependent variable (outcome). E.g. the effect of an increase in years since migration on the individual wages.

Graphical Example

Figure: Relationship between Italian Gross Monthly Average Wage of Foreign Workers and Years Since Migration in December 2013



Source: Italian Labour Force Survey–cross-sectional quarterly data (2013)

table

Introduction 2/2

What is the linear regression design?

Linear regression fits data in a linear model to show the relationship between independent variables and dependent variable. The independent variables are multiplied by a coefficient, which shows the average effect of each independent variable on the dependent variable.

Types of Regression analysis

Univariate Regression

Univariate regression studies the relationship between an independent variable and a dependent variable.

Multivariate Regression

Multivariate regression studies the relationship between more than one independent variable and a dependent variable.

Univariate Regression

Formula

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{1}$$

β_0: is the intercept of the line
β_1: is the slope of the line
i: indicates a person
ε_i: is an error term due to the fitting
y_i: is the dependent variable
x_i: is the explanatory or independent variable

How to interpret the coefficients

Coefficient formulas

$$\beta_1 = \frac{Cov(y_i, x_i)}{Var(x_i)} \tag{2}$$

$$\beta_0 = E[y_i] - \beta_1 E[x_i] \tag{3}$$

β_1: provides the effect of one unit increase in all x_is on all y_is.
β_0: provides the value of y_is when the independent variable is equal to zero.

Table Example (1/2)

Table: The Effect Of An Increase In Years Since Migration On The Foreigners' Gross Monthly Wage

	Wage
Years since migration	13.67***
	(0.787)
N	3331
R^2	0.118

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

graph

Table Example (2/2)

Interpretation:

We claim that an unit change of the years since migration increases the gross monthly wage by 13.67 euro (β_1) on average. If the change is equal to two, the increase in the gross monthly wage is 2*13.67 euro=27.34.

Multivariate Regression

Formula

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i \tag{4}$$

β_0: is the intercept
β_1, β_2: are the coefficients
i: indicates a person
ε_i: is an error term due to the fitting
y_i: is the dependent variable
x_{i1}, x_{i2}: are the explanatory or independent variables

How to interpret the coefficients

β_2: provides the effect of one unit increase in all x_{i2} on all y_i from averaging out the effect of x_i2 on x_i1.
β_1: provides the effect of one unit increase in all x_{i1} on all y_i from averaging out the effect of x_i1 on x_i2.
β_0: provides the value of y_i when both independent variables are equal to zero.

Table Example (1/2)

Table: The Effect Of Both An Increase In Years Since Migration And Total Hours Worked On The Foreigners' Gross Monthly Wage

	Wage
Years since	13.93***
migration	(0.769)
Total hours	9.87***
worked	(0.618)
N	3326
R^2	0.209

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

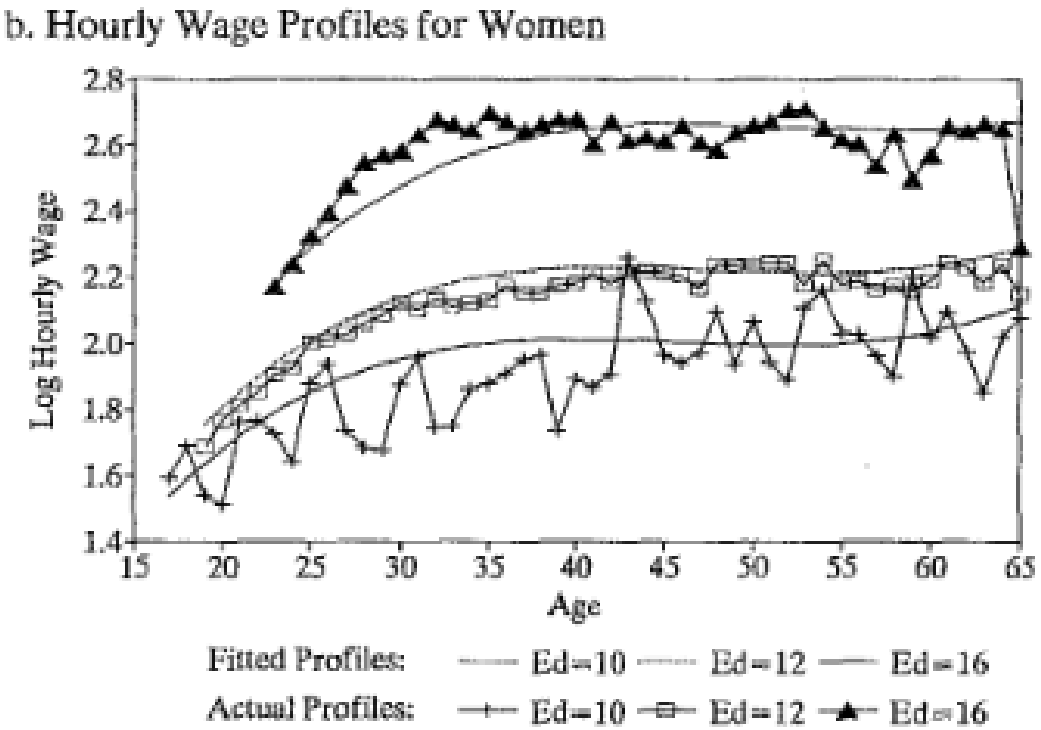
Table Example (2/2)

Interpretation:

We can claim that an unit change on average of the years since migration increases the gross monthly wage by 13.93 euro (β_1). Further, we. claim that an unit change on average of the total hours worked increases the gross monthly wage by 9.87 euro (β_2).

The Relationship between Education and Wage

Figure: Age profiles of hourly wages for women with different level of education (Card, 1999)



Comments:

Graph shows the relationship between age the earning profiles for different level of education. Age is a good proxy for the experience and, hence, has a positive effect on hourly wages. Further, higher education level leads to higher earnings per se. The solid lines among dots are the regression lines.

Estimated education coefficients from standard human capital earnings function fit to hourly wages, annual earnings, and various measures of hours for men and women in March 1994–1996 Current Population Survey[a]

	Dependent variable				
	Log hourly earnings	Log hours per week	Log weeks per year	Log annual hours	Log annual earnings
	(1)	(2)	(3)	(4)	(5)
A. Men					
Education	0.100	0.018	0.025	0.042	0.142
coefficient	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
R-squared	0.328	0.182	0.136	0.222	0.403
B. Women					
Education	0.109	0.022	0.034	0.056	0.165
coefficient	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
R-squared	0.247	0.071	0.074	0.105	0.247

[a] Notes: Table reports estimated coefficient of linear education term in model that also includes cubic in potential experience and an indicator for non-white race. Samples include men and women age 16–66 who report positive wage and salary earnings in the previous year. Hourly wage is constructed by dividing wage and salary earnings by the product of weeks worked and usual hours per week. Data for individuals whose wage is under \$2.00 or over \$150.00 (in 1995 dollars) are dropped. Sample sizes are: 102,639 men and 95,309 women.

Comments:

This regression analysis's table shows the effect of an increase in the education level on labor outcomes. The red circled shows the increase of one more year in the education level on wage, which is 0.1 log points (around 10% increase).

Time series approach

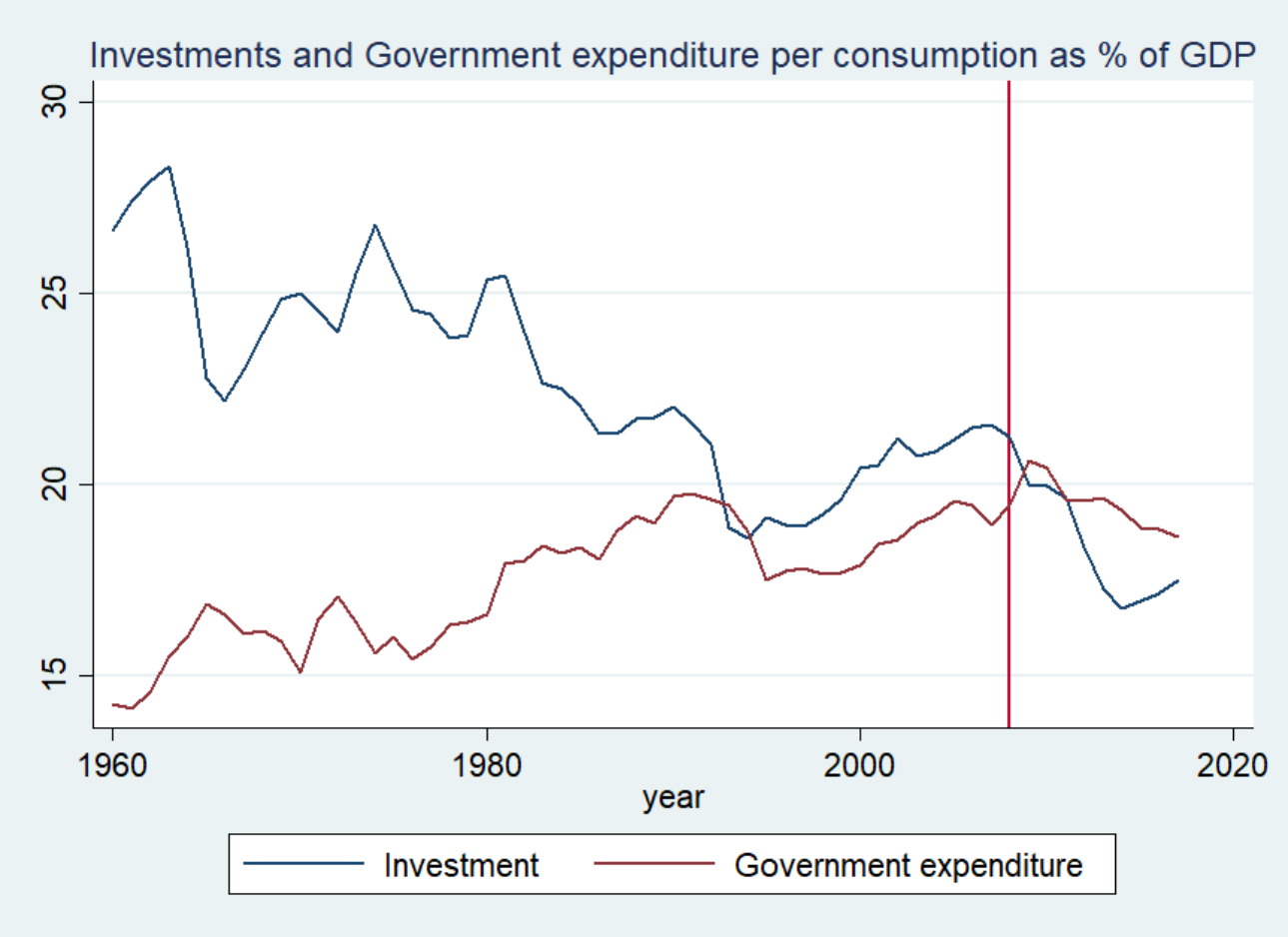


Figure: World Bank national accounts data, and OECD National Accounts data files

A time series approach is different with respect to a cross-sectional one. In the last you look at correlation at the same time, while here you look at the movement of the two series over the time. Usually they are macroeconomic variable that follow the same path with the same trend or an inverse trend. In the graph above government expenditure per consumption and the Investments have the opposite trend but they move together or with a lag.

Source	SS	df	MS		
Model	299.278796	1	299.278796	Prob > F	0.0000
Residual	195.999186	56	3.49998547	R-squared	0.6043
Total	495.277982	57	8.68908741	Root MSE	1.8708

	Number of obs	=	58
	F(1, 56)	=	85.51
	Prob > F	=	0.0000
	R-squared	=	0.6043
	Adj R-squared	=	0.5972
	Root MSE	=	1.8708

FI_GDP	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
G_GDP	-1.388732	.1501805	-9.25	0.000	-1.68958	-1.087884
_cons	46.74079	2.67872	17.45	0.000	41.37466	52.10691

Figure: World Bank national accounts data, and OECD National Accounts data files

The red and blue circled have the same interpretation as before. The main difference is the R squared, now it is much higher than before by taking into account only one variable. In time series analysis having high R squared is common. Macro series move together and for this reason each of them is a good predictor of the others. The green circled number is the ratio between the coefficient and the Std. Err.. If that number is larger than 1.96 the variable has a good predictive power.

An Example in Economics of Immigration

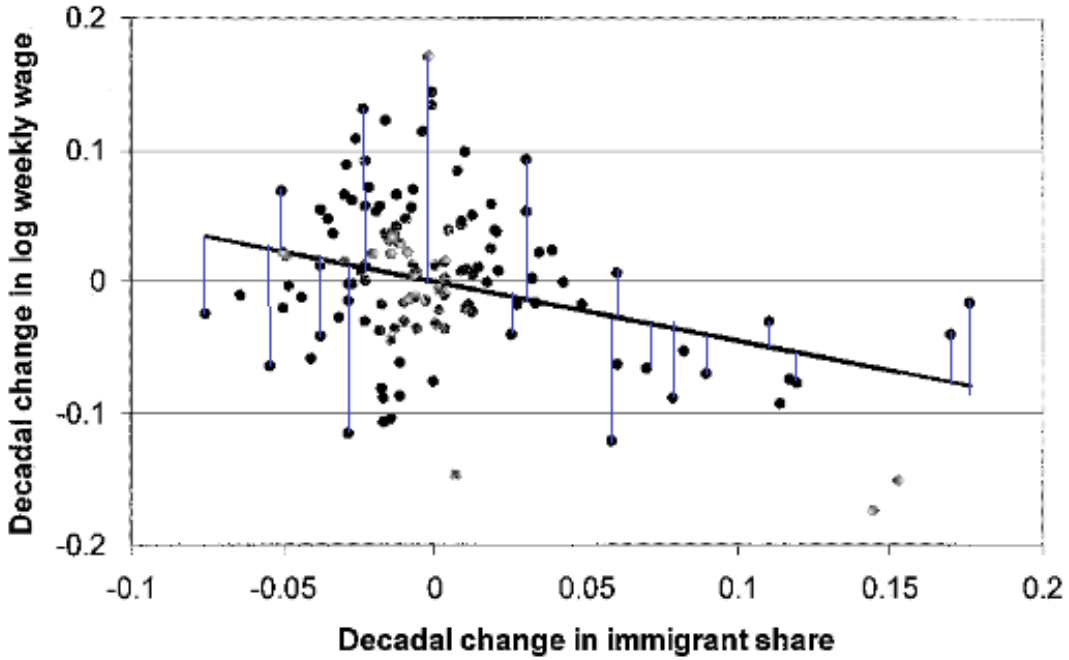


Figure: Scatter Diagram Relating Wages and Immigration, 1960–2000, Borjas (2003)

How to interpret the coefficients

Coefficient formulas

$$\beta_1 = \frac{Cov(Immshare, weekwage)}{Var(Immshare)} \tag{5}$$

$$\beta_0 = E[Weeklywage] - \beta_1 E[Immshare] \tag{6}$$

β_1: provides an indication of how many \$ a weekly wage increases or decreases when there is an increase in migrant share.
β_0: provides an indication of a weekly wage when there is not immigration.

How to interpret the regression

Assuming that the error term has zero mean. We can say that on average if the immigration share is equal to one, salary is $\beta_0 + \beta_1$, if it is equal to two the salary is $\beta_0 + \beta_1 * 2$ and so on and so forth. Hence the predicted values of weekly wages are:

Predicted values formula

$$Weekwage = \beta_0 + \beta_1 Immshare \tag{7}$$

How to interpret the Standard Errors - 1

How to interpret the Standard Errors - 2

The graph above shows a 95% confidence interval of regression estimate. The confidence interval (CI) is the distance between the error mean (zero by assumption) plus 0.025 standard deviation and minus 0.025 standard deviation when we assign a level of 95%. ($\text{CI}=0 \pm 0.025 * SE$). The SE formula is given by :

Standard error formula

$$SE = \sqrt{\frac{(\Sigma_i Y_i - \beta_0 - \beta_1 X_i)^2}{N}} \tag{8}$$

The standard error is computed as the square root of the squared sum of the difference between the y and the predicted values divided by the number of observations.

What are the main issues with linear regression?

Problems start when we are not consistent with the assumptions.
The assumptions are three:

- not omitted variables;
- full rank;
- homoskedasticity.

Omitted variable bias

Omitted variable bias arises when we do not take into account all the possible variables linked to both dependent and independent variables.

Example

$$wage_i = \beta_0 + \beta_1 experience_i + \epsilon_i \quad (9)$$

Omitted variable bias

Experience cannot fit well the wage, because the shape is not linear but it is more similar to a quadratic form. We have to add a quadratic term to our regression, in this case the square of experience.

Example

$$wage_i = \beta_0 + \beta_1 experience_i + \beta_2 experience_i^2 + \epsilon_i \quad (10)$$

Full rank bias

Full rank bias arises when we include too many variables in the regression that are similar among them.

Example

$$wage_i = \beta_0 + \beta_1 experience_i + \beta_2 experience_i^2 + \beta_2 age_i + \beta_3 yearsofschooling_i + \quad (11)$$

Given that *experience* is computed as a proxy of *age* minus the years of school, we cannot identify the parameter. When there is a variable that is a very similar to another variable, most of the time is impossible to compute all the coefficients.

What is causality?

"...it is of consequence to know the principle whence any phenomenon arises, and to distinguish between a cause and a concomitant effect. Besides that the speculation is curious, it may frequently be of use in the conduct of public affairs. At least, it must be owned, that nothing can be of more use than to improve, by practice, the method of reasoning on these subjects, which of all others are the most important; though they are commonly treated in the loosest and most careless manner." *On Interest*, Hume (1742, p. 304).

Definition

Relation that holds between two temporally simultaneous or successive events when the first event (the cause) brings about the other (the effect). According to David Hume, when we say of two types of object or event that "X causes Y" (e.g., fire causes smoke), we mean that:

- Xs are "constantly conjoined" with Ys;
- Ys follow Xs and not vice versa;
- there is a "necessary connection" between Xs and Ys such that whenever an X occurs, a Y must follow.

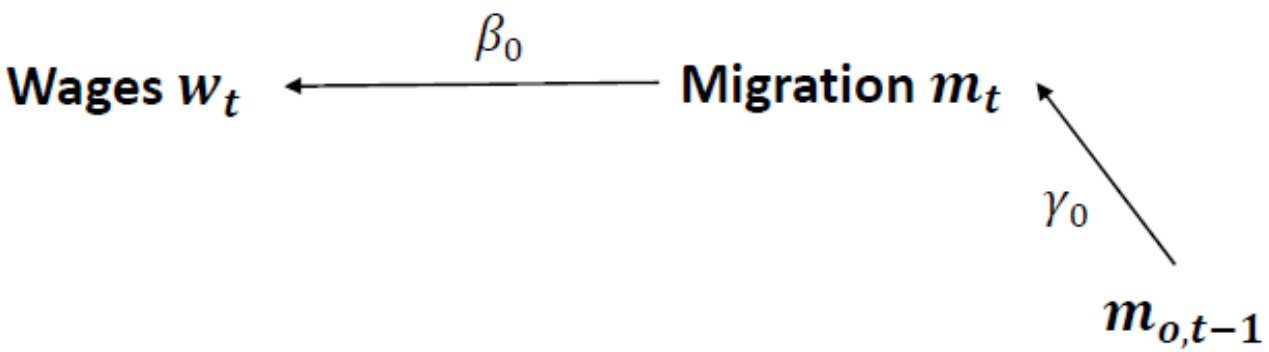
(*The Editors of Encyclopaedia Britannica*)

Warning

Remember to have a casual relation between the two characteristics it is used to have on the x-axis the cause and on the y-axis the effect. If you shift the characteristics on the axes you have the same plot but there is not a casual relation because earning more doesn't mean increasing the education level or becoming older!!!!

Causality

Figure: A third explains migration but not wages



Correlation

Figure: A third variable explains both migration and wages