Introduction to Causality

Lecture 1

Carrozzo Salvatore

University of Turin and Collegio Carlo Alberto

February 28th, 2019

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

Outline

Introduction - 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*)

Univariate regression analysis

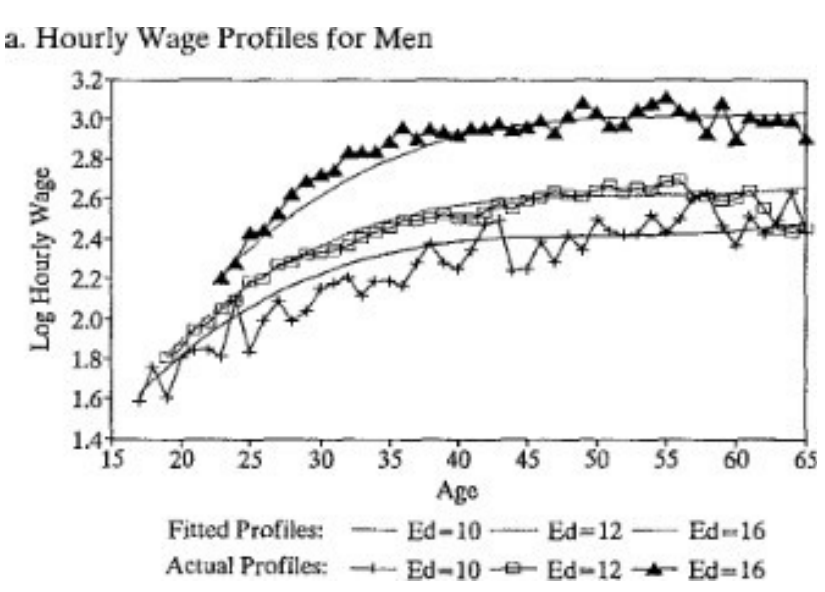
Formula

$$yi = \beta 0 + \beta 1 xi + si \quad (1)$$

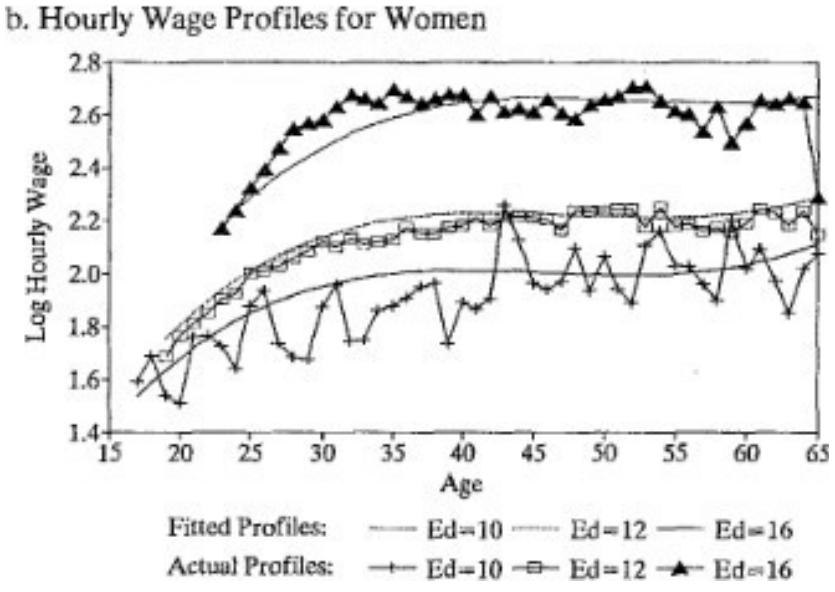
β_0: is the intercept of the line
β_1: is the slope of the line
i: indicates a person
s_i : is an error term due to the fitting
y_i : is called dependent variable
x_i : is called explanatory or independent variable

Univariate regression analysis

Cross-sectional approach



(a)label 1



(b)label 2

Figure:Age profiles of hourly wages for men (a) and women (b) (Card,1999)

Cross-sectional approach

Two graphs show the relation between the age and the earning profiles. Becoming older increases the level of experience in the labor market by providing higher wages. The solid line among different dots is the regression line. The distance between dots and line is the part of wages that age cannot explain. It is called residual component and it can be explained by different characteristics not included in the graph (e.g. cultural background, parental education, born country).

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.

The table shows the mathematical results of a regression analysis. The red circled number is the coefficient and it is the multiplier of an increase in the education level.(E.g. one more year in the education level increases the wage by 0.1 log points). The blue circled number is an index of precision of the estimated coefficient. The black circled number is the part of income explained by the education level and other variables.

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!!!!

Warning

Time series

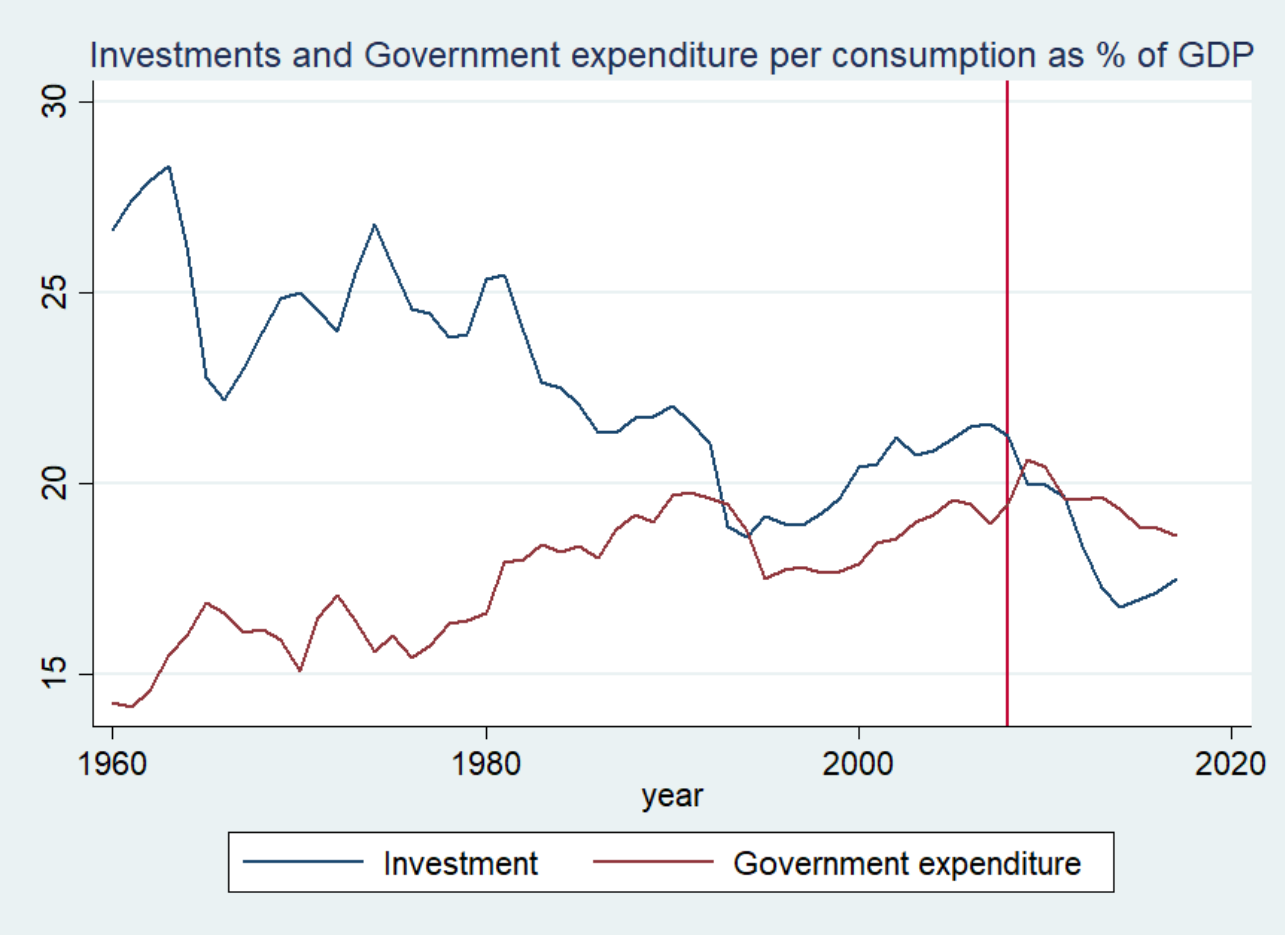


Figure:World Bank national accounts data, and OECD National Accounts data files

Migration in Europe
MigrEU Jean Monnet Module

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
				Adj R-squared	0.5972
Total	495.277982	57	8.68908741	Root MSE	1.8708

Number of obs = 58
F(1, 56) = 85.51

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.

Mathematical Background

Definition

Linear regression, or ordinary least squares method (OLS), is a method that tries to fit the data l

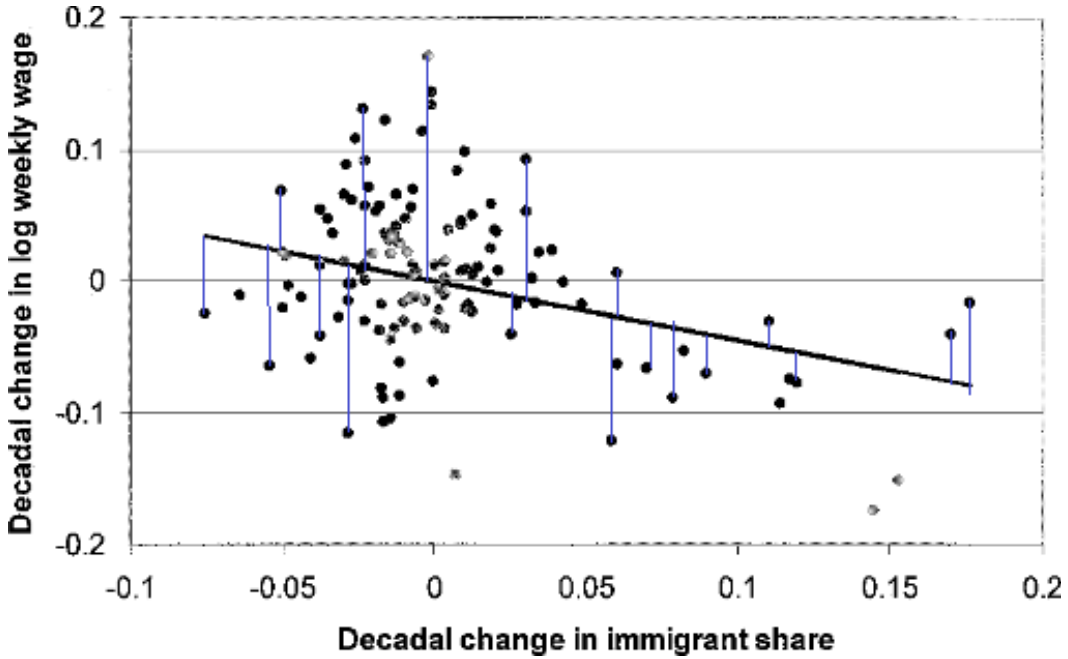


Figure:Scatter Diagram Relating Wages and Immigration, 1960–2000 Borjas (2003)

How to interpret the coefficients

Coefficient formulas

$$\beta 1 = \frac{\text{Cov (Immshare, weekwage)}}{\text{Var (Immshare)}} \quad (2)$$

$$\beta 0 = E\,[\text{Weeklywage}] - \beta 1 E\,[\text{Immshare}] \quad (3)$$

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 thaton 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

$$Weekwage = \beta 0 + \beta 1 Immshare \quad (4)$$

How to interpret the Standard Errors -

How to interpret the Standard Errors -

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%. ($CI = 0 \pm 0.025 * SE$). The SE formula is given by :

Standard error formula

$$SE = \sqrt{\frac{(\Sigma_i Y - \beta - \beta_1 X_i)^2}{N}} \quad (5)$$

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.

How to interpret the Standard Errors -

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

$$wage_i = \beta_0 + \beta_1 experience_i + s_i \quad (6)$$

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

$$wage_i = \beta_0 + \beta_1 experience_i + \beta_2 experience^2 + s_i \qquad (7)$$

Full rank bias

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

Example

$$wage_i = \beta_0 + \beta_1 experience_i + \beta_2 experience_i^2 + \beta_2 age_i + \beta_3 yearsofschooling_i + s_i \quad (8)$$

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.