Slides for the course

Statistics and econometrics

Part 6: Causality and Regression

European University Institute

Andrea Ichino

September 11, 2014

Outline

A brief introduction to the problem of causality

What is needed for a “causal" interpretation of the PRF

Section 1

A brief introduction to the problem of causality

Causality, the PRF and the CEF

So far we have characterized the Population Regression Function as a linear approximation to the Conditional Expectation Function.

OLS-MM is an estimator of the PRF with some desirable properties.

Given a specific sample, the Sample Regression Function estimated with OLS-MM is a “good" estimate of the PRF-CEF.

It is not an estimate of the causal effect of x on y unless the CEF-PRF itself can be interpreted in a causal sense.

We want to briefly introduce what it means to give a causal interpretation to the PRF-CEF and what this implies for the regression.

Examples of causal question

- Does smoking cause lung cancer?
- Does aspirin reduce the risk of heart attacks?
- Does an additional year of schooling increase future earnings?
- Are temporary jobs a stepping stone to permanent employment?
- Does EPL increase unemployment?

The answers to these questions (and to many others which affect our daily life) involve the identification and measurement of causal links: an old problem in philosophy and statistics.

The counterfactual definition of causality

Consider the simple case of a binary variable X and an outcome Y.

To give a precise meaning to the sentence

$$X \text{ causes } Y$$

we need to define counterfactuals (or potential ouccomes).

This requires assuming that:

- the outcome Y_1 that occurs when $X = 1$ and
- the outcome Y_0 that occurs when $X = 0$

are both well defined even if we can observe only one of them.

$$Y_{obs} = Y_1 X + Y_0(1 - X) \tag{1}$$

Within this framework, the causal effect of X on Y, is

$$\tau = Y_1 - Y_0 \tag{2}$$

Fundamental problem of causal inference

It is impossible to observe for the same unit i the values $x_i = 1$ and $x_i = 0$ as well as the values Y_1 and Y_0 and, therefore, it is impossible to observe *the causal effect of X on Y for a specific unit i.*

Causal analysis, in different disciplines, tries to solve this problem.

Statistics looks for solutions based on

- randomized experiments when possible;
- alternative strategies that try to approximate randomized experiments.

to identify causal effects for groups of units in the population.

See Holland (1986) for a discussion of alternative approaches to causality.

Statistical causal effects

- *A*verage treatment effect

$$ATE = E(Y_1) - E(Y_0) \tag{3}$$

- *A*verage effect of treatment on the treated

$$ATT = E(Y_1|X = 1) - E(Y_0|X = 1) \tag{4}$$

- *A*verage effect of treatment on the non treated

$$ATNT = E(Y_1|X = 0) - E(Y_0|X = 0) \tag{5}$$

Why randomized experiments solve the problem

Given two random samples C and T from the population:

$$E(Y_0|C) = E(Y_0|T) = E(Y_0) \tag{6}$$

and

$$E(Y_1|C) = E(Y_1|T) = E(Y_1). \tag{7}$$

The:

$$E(\tau) = E(Y_1) - E(Y_0) = E(Y_1|T) - E(Y_0|C) \tag{8}$$

Randomization solves the Fundamental Problem of Causal Inference because it allows to use the *control* units C as an image of what would happen to the *treated* units T in the counterfactual situation of no treatment, and vice-versa.

Section 2

What is needed for a “causal" interpretation of the PRF

Definition of counterfactual wages and education

For each subject there exist two “potential wage levels" depending on going to college (high education) or not (low education):

$$\begin{aligned} y_h &= \mu_h + \nu \\ y_l &= \mu_l + \nu \end{aligned} \tag{9}$$

where $E(\nu) = 0$.

The “causal effect" of college attendance on earnings for a subject

$$\tau = y_h - y_l = \mu_h - \mu_l \tag{10}$$

is not identified because only one potential outcome is observable.

Let $x = 1$ denote college attendance while $x = 0$ indicates lower education. The observed wage level y is given by:

$$y = y_l(1 - x) + y_h x \tag{11}$$

The Population Regression Function

We want to know if and under what conditions the parameter β_1

$$y = \beta_0 + \beta_1 x + u \tag{12}$$

is the average causal effect of college on wages in the population.

Substituting 9 in 11 the causal relationship between x and y is:

$$y = \mu_l + (\mu_h - \mu_l)x + \nu \tag{13}$$

which looks promising, but we need to show that, given how we defined β_1 in the PRF,

$$(\beta_0, \beta_1) = arg \min_{b_0, b_1} E\left[(y - b_0 - b_1 x)^2\right] \tag{14}$$

then

$$\beta_1 = \mu_h - \mu_l \tag{15}$$

A useful result: regression when x is a dummy

We have seen that the solution to problem 14 is

$$\beta_1 = \frac{Cov(y,x)}{V(x)} = \frac{E(yx) - E(y)E(x)}{E(x^2) - (E(x))^2} \tag{16}$$

Note that β_1 is a population parameter (not an estimator).

Since x is a dummy, $V(x) = p(1-p)$ where $p = Pr(x = 1)$, while the numerator of 16 is:

$$\begin{aligned} E(yx) - E(y)E(x) &= E(y|x=1)p - pE(y) \qquad (17) \\ &= E(y|x=1)p - p[E(y|x=1)p + E(y|x=0)(1-p)] \\ &= E(y|x=1)p(1-p) - E(y|x=0)p(1-p) \end{aligned}$$

and therefore

$$\beta_1 = \frac{Cov(y,x)}{V(x)} = E(y|x=1) - E(y|x=0) \tag{18}$$

Sample averages in the RHS of 18 give the “Wald estimator".

The Selection Bias

Substituting 13 in 18, we get

$$\begin{aligned} \beta_1 &= E(y|x=1) - E(y|x=0) \\ &= E(\mu_h + \nu|x=1) - E(\mu_l + \nu|x=0) \\ &= \mu_h - \mu_l + [E(\nu|x=1) - E(\nu|x=0)] \\ &= \tau + [E(\nu|x=1) - E(\nu|x=0)] \end{aligned} \tag{19}$$

where the term in brackets is called *Selection Bias* (SB)

SB captures the (pre-treatment) unobservable differences between college graduates and other subjects, which are not attributable to college attendance.

Is β_1 a causal parameter?

The PRF and β_1 have a causal interpretation only if the Selection Bias is zero,

$$[E(\nu|x = 1) - E(\nu|x = 0)] = 0 \tag{20}$$

SB is zero only if the “treated" and the “non-treated" have on average the same unobservables in the hypothetical case in which both were treated or not.

This can happen only :

- in a randomized controlled experiment;
- when for other reasons not controlled by the researcher, exposure to treatment is random in the population.

Generalization when x is not dummy

In the more general situation in which x is not a dummy

$$\begin{aligned} \beta_1 &= \frac{Cov(y,x)}{V(x)} = \frac{Cov[(\mu_I + \tau x + \nu), x]}{V(x)} \\ &= \tau + \frac{Cov(\nu, x)}{V(x)} \end{aligned} \tag{21}$$

and the PRF is causal when, in the population, the treatment x is uncorrelated with unobservable pre-treatment characteristics ν.

The interpretation is the same as in the “binary x" case.

Causality is a feature of the relationship between x and y, and can be identified only when subjects are randomly exposed to x.

When random exposure of subjects to x occurs in the population of interest, we can interpret the PRF as a causal relationship.

Another way to put it

Compare:

$$y = \beta_0 + \beta_1 x + u \tag{22}$$

$$y = \mu_I + \tau x + \nu \tag{23}$$

We know that by definition β_0 and β_1 in 22 imply

$$Cov(x, u) = E(xu) = 0 \tag{24}$$

but nothing guarantees that the u which derive from the definition of the PRF parameters and that satisfies 24, coincide with ν.

Only when x and ν are such that

$$Cov(x, \nu) = E(x\nu) = 0 \tag{25}$$

i.e. we have random exposure of subjects to x in the population, then

$$\nu = u \qquad \text{and} \qquad \beta_1 = \tau \tag{26}$$

and the PRF can be interpreted causally.

Consistency and causality

Following the Angrist Pischke approach, the OLS estimator is consistent for the PRF by definition of the population parameters it aims to estimate because

$$Cov(x, u) = E(xu) = 0 \tag{27}$$

follows from the definition of β_1 and β_0 and is not an assumption.

But “consistency" simply means that the SRF can be made arbitrarily close to the PRF by increasing the sample size.

Consistency and causality (cont)

Thus, consistency of OLS implies nothing about causality. Only if

$$Cov(x, \nu) = E(x\nu) = 0 \tag{28}$$

the PRF is a causal relationship, in which case the OLS is consistent for the causal effect of x on y in the population.

Note however that even when the PRF has no causal interpretation:

- the PRF it is still the best linear approximation to the CEF
- the PRF it is still the best linear approximation to the unkown relationship between x and y.

Regression is therefore a useful statistical tool also when it cannot be given a causal interpretation.

To summmarize this brief introduction to causality

- The causal effect of x and y requires comparing counterfactuals and cannot be identified for a specific subject.
- If we have a population in which exposure to x is random, then the PRF identifies the average causal effect of x on y.
- But even if exposure to x is not random, we are still interested in the PRF, which is the MMSE approximation to the unknown CEF.
- The PRF defines its parameters in a way such that the population residuals are uncorrelated with x, but this does not ensure a causal interpretation.
- However this definition of the PRF guarantees that we can say something about the PRF (and the CEF) with a random sample.
- The OLS estimator is the BLUE for the PRF parameters if the SLR 1 - SLR 5 assumptions of Gauss Markov hold.
- SLR 1 - SLR 3 are enough for OLS to be consistent for the PRF.