2020 STATA ECONOMETRICS WINTER SCHOOL

Anabela Carneiro[1]   João Cerejeira[2]   Miguel Portela[2,3]   Paulo Guimarães[1,4]

[1]FEP and CEF.UP – U.Porto
[2]NIPE – UMinho
[2]IZA, Bonn
[4]Banco de Portugal
Faculdade de Economia da Universidade do Porto

January 20-24, 2020

Identifying the causal impact of some variables **X** on **y** is difficult.
Experimental research designs offer the most plausibly unbiased estimates,
but experiments are frequently infeasible due to cost or moral objections.
Although, experiments have become more common in social sciences.
Examples:
- Tennessee Project STAR
- PROGRESA's Control Randomized Experiment
- ...

Randomized experiments

Randomized experiments were first conducted in the sciences (commonly traced back to Galileo Galilei who used experiments to test his theories of falling bodies).

Randomized experiments in the social sciences in particular suffer from a major problem: the missing counterfactual - individuals or firms can usually not be observed with and without treatment at the same time.

Some notation: defining causality

Consider a population of individuals and for each there are two "potential wage levels" Y_{1i}, Y_{0i} (defined as "outcomes") depending on whether one goes to college or not ($D_i = 1$ if i goes to college; $D_i = 0$ if not; defined as "treatment"):

$$\begin{cases} Y_{1i} \text{ if } D_i = 1 \\ Y_{0i} \text{ if } D_i = 0 \end{cases} \tag{1}$$

or

$$Y_{Di} = D_i Y_{1i} + (1 - D_i) Y_{0i}. \tag{2}$$

We only observe one of these "potential wage levels" because nobody is observed in both situations.

The causal effect of college attendance on wages for individual i is defined as the difference between the two potential outcomes:

$$\tau_i = \Delta_i = Y_{1i} - Y_{0i}$$

The fundamental problem of causal inference

The identification and the measurement of the effect $\tau = Y_{1i} - Y_{0i}$ is logically impossible, because we do not oberve Y_{1i} and Y_{0i} simultaneously for each i.

It is impossible to observe for the same unit i the values $D_i = 1$ and $D_i = 0$ as well as the values Y_{1i} and Y_{0i} and, therefore, it is impossible to observe the effect of D on Y for unit i (Holland, 1986).
Another way to express this problem is to say that we cannot infer the effect of a treatment because we do not have the counterfactual evidence i.e. what would have happened in the absence of treatment.

The selection problem

If we compare averages by treatment status, we have:

$$
\begin{aligned}
E(Y_i|D_i = 1) - E(Y_i|D_i = 0) &= \\
&= E(Y_{1i}|D_i = 1) - E(Y_{0i}|D_i = 1) + \\
+E(Y_{0i}|D_i = 1) - E(Y_{0i}|D_i = 0)
\end{aligned}
$$

$E(Y_{1i}|D_i = 1) - E(Y_{0i}|D_i = 1) = E(Y_{1i} - Y_{01}|D_i = 1)$- average causal effect of college education on those who actually go to college (ATT - average treatment effect on the treated);

$E(Y_{0i}|D_i = 1) - E(Y_{0i}|D_i = 0)$ - selection bias.

The goal of causal analysis is to overcome selection bias in order to estimate consistently the causal effect

$E(Y_{1i} - Y_{01}|D_i = 1)$.

Suppose that D_i is randomly assigned, and therefore is independent of potential outcomes:

$$
\begin{aligned}
E(Y_i|D_i = 1) - E(Y_i|D_i = 0) &= \\
&= E(Y_{1i}|D_i = 1) - E(Y_{0i}|D_i = 0) = \qquad (3) \\
&= E(Y_{1i}|D_i = 1) - E(Y_{0i}|D_i = 1) = \qquad (4) \\
&= E(Y_{1i} - Y_{0i}) \qquad\qquad\qquad\qquad\quad (5)
\end{aligned}
$$

Random assignment solves many problems in causal analysis.
So, why do not use randomized trials in economics?

Example (1): Tennessee Project STAR
Krueger, AB, (1999). "Experimental Estimates of Education Production Functions", QJE.

Krueger (1999) econometrically re-analyses a randomized experiment of the effect of class size on student achievement.

The project is known as Tennessee Student/Teacher Achievement Ratio (STAR) and was run in the 1980s.

11,600 students and their teachers were randomly assigned to one of three groups:

- Small classes (13-17 students).

- Regular classes (22-25 students).

- Regular classes (22-25 students) with a full time teacher's aide.

After the assignment, the design called for students to remain in the same class type for four years.

Randomization occurred within schools.

Example (2): PROGRESA's Control Randomized Experiment

Gertler, P., (2004). "Do Conditional Cash Transfers Improve Child Health? Evidence from PROGRESA's Control Randomized Experiment", AER.

The Mexican anti-poverty program PROGRESA combines a traditional cash-transfer program with financial incentives for positive behaviour in health and nutrition, namely for child care.

Due to budgetary and logistical constraints, the government was unable to enroll all eligible families simultaneously.

The government decided to randomly choose which villages would receive benefits first.

320 treatment and 185 control villages in seven states for a total of 505 experimental villages.

Eligible houseolds in treatment villages receive benefits immediatly (Aug-Set 1998).

Eligible households in control villages receive benefits two years later.

Regression and causality

Substituting 1 into 2, the casual relationship between Y and D is:

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i$$
$$Y_i = E(Y_{0i}) + (Y_{1i} - Y_{0i})D_i + Y_{0i} - E(Y_{0i})$$

if $E(Y_{0i}) = \beta_0$, and $(Y_{1i} - Y_{0i}) = \beta_1$ (assuming the tratment effect is homogeneous), then:

$$Y_i = \beta_0 + \beta_1 D_i + v_i$$

where $v_i = Y_{0i} - E(Y_{0i})$ is the random part of Y_{0i}.

Evaluating the conditional expectation of this equation with different treatment status gives:

$$E(Y_i|D_i = 1) = \beta_0 + \beta_1 + E(v|D = 1)$$
$$E(Y_i|D_i = 0) = \beta_0 + E(v|D = 0)$$

and therefore:

$$E(Y_i|D_i = 1) - E(Y_i|D_i = 0) = \beta_1 + E(v|D = 1) - E(v|D = 0)$$
$$E(Y_i|D_i = 1) - E(Y_i|D_i = 0) = \text{treatment effect} + \text{selection bias}$$

If $E(v|D = 1) - E(v|D = 0) = E(Y_{0i}|D = 1) - E(Y_{0i}|D = 0) \neq 0$, then there are correlation between D and the error term and therefore we cannot get consistent estimates of β_1 by OLS. The selection bias is the difference in potencial outcomes, with no treatment between those who get treated and those who don't.

Example

Group	$E(Y_{1i}|D_i = ...)$	$E(Y_{0i}|D_i = ...)$
Treatment $D_i = 1$	10 observed	6 not observed
Control $D_i = 0$	8 not observed	4 observed

How large is the selection bias? What is the causal effect of the treatment on the treated?

Note that in a randomized trial:

$$
\begin{aligned}
E(Y_{0i}|i &= \ Control) = E(Y_{0i}|i = Treatement) \text{ or} \\
E(Y_{0i}|D &= \ 1) = E(Y_{0i}|D = 0), \text{ no selection bias.}
\end{aligned}
$$

Include additional controls?

To evaluate experimental or non experimental data one may want to add additional controls in the regression. Instead of estimating equation

$$Y_i = \beta_0 + \beta_1 D_i + v_i$$

one would estimate:

$$Y_i = \beta_0 + \beta_1 D_i + \mathbf{X}_i'\gamma + v_i$$

There are 2 main reasons for including additional controls in the regression model.

- Conditional random assignment: sometimes randomization is done conditional on some observables;

- Additional controls increase precision: Although the control variables $\mathbf{X}_i$ are uncorrelated with D_i they may have substantial explanatory power for Y_i . Including controls thus reduces residual variance and therefore lowers the standard errors of the regression estimates.

Sources of Bias and Inconsistency

The selection bias (or omitted variable bias) in an ordinary regression arises from endogeneity (a regressor is said to be endogenous if it is correlated with the error).

This also occurs if the explanatory variable is measured with error, or in a system of "simultaneous equations" (e.g. suppose work also has a causal impact on mental health, or higher earnings cause increases in education; in this case, it is not clear what impact, if any, our single-equation regressions identify).

Omitted variables

Often a suspected type of endogeneity can be reformulated as a case of omitted variables:

$$y = X\beta + W\gamma + \varepsilon,$$

if W is unobserved, then the formula for omitted variable bias in linear regression is:

$$p\lim \widehat{\beta} = \beta + \gamma \frac{Cov(W, X)}{Var(X)}.$$

So the bias is proportional to the correlation of X and W and to the effect of W (the omitted variable) on y.

Alternatives to experimental designs

Three types of research designs offering approaches to causal inference
using observational data will be discussed next days:
- differences-in-differences (DiD) estimators;
- instrumental variables (IV);
- propensity scores matching
- panel methods,
Each method has strengths and weaknesses but in practice, the data and
the assumptions that allow causal inference with these models determine
the method.