Causal statistics for treatment models

1. Rubin causal model and randomized experiments

Ecole Normale Supérieure–PSL, L3 Economics and Cogmaster, 2025

Table of contents

Starting example

Question: do hospitals make people healthier?

Survey:

- During the past 12 months, was the respondent a patient in a hospital overnight?
- Would you say your health in general is excellent, very good, good, fair, poor?

Table 1: Mean health status (1-5 scale) among subjects hospitalized/not hospitalized

Group	N	Mean Health status	s.e.
Hospital	7,774	3.21	0.014
No hospital	90,049	3.93	0.003

(Taken from Angrist & Pischke, 2009)

Causality and the notion of counterfactual

If a person eats of a particular dish, and dies in consequence, that is, would not have died if he had not eaten it, people would be apt to say that eating of that dish was the source of his death

J.S. Mill (*A System of Logic*, 1843), quoted by Imbens and Rubin (2015)

Neyman, 1923

First to formalize the concept of counterfactual in the context of the yield of plant varieties

"Potential yield":

Fig. 1. *Jerzy Neyman in Poland, not long after 1923.*

In the ith urn, let us put m balls (as many balls as plots of the field), with labels indicating the unknown potential yield of the ith variety on the respective plot, along with the label of the plot. Thus on each ball we have one of the expressions

$$U_{i1},\ U_{i2},\ \cdots,\ U_{ik},\ \cdots,\ U_{im} \tag{13}$$

[29]

where i denotes the number of the urn (variety) and k denotes the plot number, while U_{ik} is the yield of the ith variety on the kth plot.

Rubin, 1974

Journal of Educational Psychology
1974, Vol. 66, No. 5, 688-701

ESTIMATING CAUSAL EFFECTS OF TREATMENTS IN RANDOMIZED AND NONRANDOMIZED STUDIES[1]

DONALD B. RUBIN[2]

Educational Testing Service, Princeton, New Jersey

A discussion of matching, randomization, random sampling, and other methods of controlling extraneous variation is presented. The objective is to specify the benefits of randomization in estimating causal effects of treatments. The basic conclusion is that randomization should be employed whenever possible but that the use of carefully controlled nonrandomized data to estimate causal effects is a reasonable and necessary procedure in many cases.

Now define the causal effect of the E versus C treatment on Y for a particular trial (i.e., a particular unit and associated times t_1, t_2) as follows:

Let y(E) be the value of Y measured[5] at t_2 on the unit, given that the unit received the experimental Treatment E initiated at t_1;

Let y(C) be the value of Y measured at t_2 on the unit given that the unit received the control Treatment C initiated at t_1;

Then y(E) − y(C) is the causal effect of the E versus C treatment on Y for that trial, that is, for that particular unit and the times t_1, t_2.

Counterfactuals

Notations

- $Y_i(0)$: outcome if not treated
- $Y_i(1)$: outcome if treated

Counterfactuals

Notations

- $Y_i(0)$: outcome if not treated
- $Y_i(1)$: outcome if treated

Discussion

- Which is observed, for whom?
- If not observed/not realized, is it still a relevant object?
- Give examples

Counterfactuals: Semantics

Before any allocation to treatment: both $Y_i(0)$ and $Y_i(1)$ are counterfactuals
So we can reasonably use "counterfactual" for either one

Counterfactuals: Semantics

Before any allocation to treatment: both $Y_i(0)$ and $Y_i(1)$ are counterfactuals
So we can reasonably use "counterfactual" for either one

When treatment status is revealed, one is an observed counterfactual, the other an unobserved counterfactual

Counterfactuals: Semantics

Before any allocation to treatment: both $Y_i(0)$ and $Y_i(1)$ are counterfactuals
So we can reasonably use "counterfactual" for either one

When treatment status is revealed, one is an observed counterfactual, the other an unobserved counterfactual

Sometimes people use "counterfactual" only for the unobserved counterfactual

Counterfactuals: Semantics

Before any allocation to treatment: both $Y_i(0)$ and $Y_i(1)$ are counterfactuals
So we can reasonably use "counterfactual" for either one

When treatment status is revealed, one is an observed counterfactual, the other an unobserved counterfactual

Sometimes people use "counterfactual" only for the unobserved counterfactual

Be flexible, and use the context to understand when counterfactual means unobserved counterfactual, or any counterfactual...

Causality

Causal effect

$$\Delta_i = Y_i(1) - Y_i(0)$$

Causality

Causal effect

$$\Delta_i = Y_i(1) - Y_i(0)$$

Remarks

- Specific to each individual i $\rightarrow$ Treatment effect heterogeneity
- In some lectures we will assume homogeneity to manipulate simple objects
- SUTVA: give examples where applies/does not apply

Treatment

Notations

- $T_i = 0$: if i not treated
- $T_i = 1$: if i treated

Treatment

Notations

- $T_i = 0$: if i not treated
- $T_i = 1$: if i treated

Question

- Which counterfactual is oberved if $T_i = 1$?
- Which counterfactual is oberved if $T_i = 0$?
- Etc.

The identification problem

- We observe **either** $Y_i(1)$ **or** $Y_i(0)$

  Thus Δ_i cannot be computed for any subject i

The identification problem

- We observe **either** $Y_i(1)$ **or** $Y_i(0)$

  Thus Δ_i cannot be computed for any subject i

- Still, under some conditions (object of this course), some $E(\Delta_i)$ can be computed

Parameters of interest

Among many possible parameters:

ATE and ATT

- $E(\Delta_i)$: Average treatment effect (ATE)
- $E(\Delta_i | T_i = 1)$: Average treatment effect on the treated (ATT)

Discussion

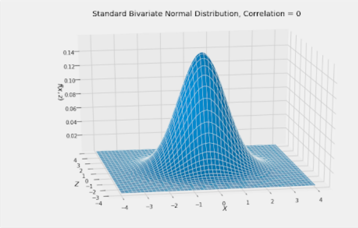
- ATE: what population?
- Why (when) would they differ?
- Think of other parameters

Digression on conditional expectations

Reminder: joint distributions, expectations

Two random variables Y, X jointly distributed:

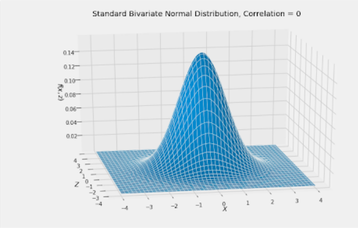
characterized by density $f(Y, X)$



Reminder: joint distributions, expectations

Two random variables Y, X jointly distributed:

characterized by density $f(Y, X)$



Marginal distribution: $f_x(X) = \int_Y f(Y, X) dY$ or $f_x(X) = \sum_Y f(Y, X)$

Expectation: $E(X) = \int_X X f_x(X) dX$ or $E(X) = \sum_X X f_x(X)$

Conditional distribution

Conditional distribution:

$$f(Y|X) = \frac{f(Y,X)}{f_x(X)}$$

Conditional expectation:

$$E(Y|X) = \int_Y Yf(Y|X)dY$$

Discrete X

We will consider mostly the case where X is discrete (for instance $T = 0/1$)

Conditional distribution:

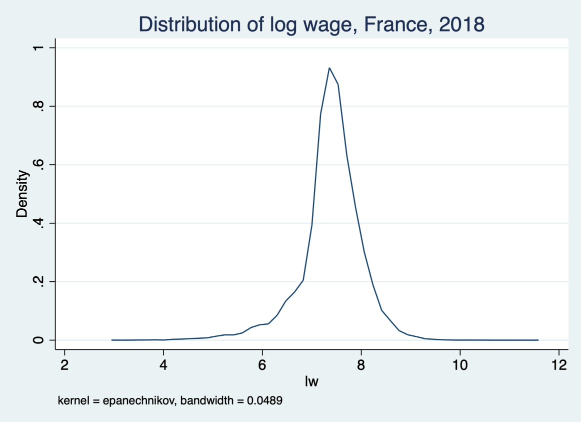
$$f(Y|T=0) = \frac{f(Y,0)}{P(T=0)}$$

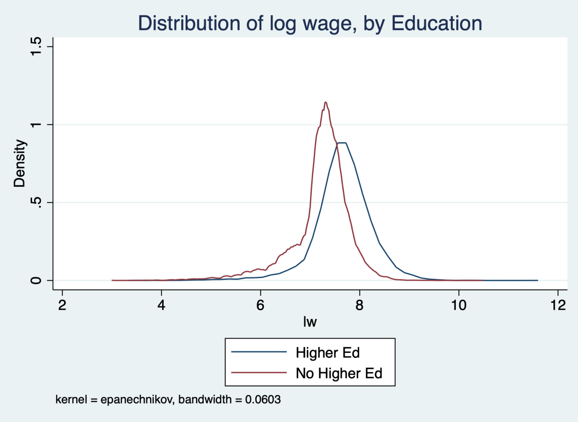
$$f(Y|T=1) = \frac{f(Y,1)}{P(T=1)}$$

Law of iterated expectations:

$$E(Y) = P(T=0)E(Y|T=0) + P(T=1)E(Y|T=1)$$

Marginal and conditional distributions





Law of iterated expectations

Table 2: Mean log-wage decomposition

Marginal mean	Prop. Low Ed.	Mean Low Ed.	Prop. High Ed.	Mean High Ed.
7.38	0.62	7.21	0.38	7.66

(Source: Enquête emploi 2018; Low Ed. = no Tertiary Education)

$$E(Y) = P(T=0)E(Y|T=0) + P(T=1)E(Y|T=1)$$

Selection bias

The naive estimator

You have data on who is treated or not (T) and some outcome Y

$$E(Y|T=1) - E(Y|T=0)$$

The naive estimator

You have data on who is treated or not (T) and some outcome Y

$$E(Y|T=1) - E(Y|T=0)$$

Compute this for T=High Education and Y=log-wage

Marginal mean	Prop. Low Ed.	Mean Low Ed.	Prop. High Ed.	Mean High Ed.
7.38	0.62	7.21	0.38	7.66

Bias of the naive estimator

Lets aim for the parameter ATT: $E(\Delta|T=1) = E(Y(1) - Y(0)|T=1)$

$$
\begin{aligned}
E(Y|T=1) \quad - & \quad E(Y|T=0) \\
= & \quad E(Y(1)|T=1) - E(Y(0)|T=0)
\end{aligned}
$$

Bias of the naive estimator

Lets aim for the parameter ATT: $E(\Delta|T=1) = E(Y(1) - Y(0)|T=1)$

$$\begin{aligned}
E(Y|T=1) &- E(Y|T=0) \\
&= E(Y(1)|T=1) - E(Y(0)|T=0) \\
&= E(Y(1)|T=1) - \overbrace{E(Y(0)|T=1) + E(Y(0)|T=1)} - E(Y(0)|T=0)
\end{aligned}$$

Bias of the naive estimator

Lets aim for the parameter ATT: $E(\Delta|T=1) = E(Y(1) - Y(0)|T=1)$

$$
\begin{aligned}
E(Y|T=1) \quad - \quad & E(Y|T=0) \\
= \quad & E(Y(1)|T=1) - E(Y(0)|T=0) \\
= \quad & E(Y(1)|T=1) - \overbrace{E(Y(0)|T=1) + E(Y(0)|T=1)} - E(Y(0)|T=0) \\
= \quad & \overbrace{E(Y(1)|T=1) - E(Y(0)|T=1)}^{\text{ATT}} + \overbrace{E(Y(0)|T=1) - E(Y(0)|T=0)}^{\text{Bias}}
\end{aligned}
$$

$\rightarrow$ Interpret bias in the case of High/Low Ed. and wage

Exercise 1

Do we have reasons to fear there is a bias?

	Average in:		Reg. of log wage **in Low Ed. group**	
	High Ed.	Low Ed.	Coef.	(s.e.)
Age	41.44	43.67	0.009	(0.0003)
Female	0.54	0.46	-0.34	(0.007)

$$\overbrace{E(Y(0)|T=1) - E(Y(0)|T=0)}^{\text{Bias}}$$

Exercise 2

Assume we *know* that the ATT of Higher Education is to increase wages by 25%

Compute the bias of the naive estimator

$$E(Y|T=1) - E(Y|T=0) = \overbrace{E(Y(1)|T=1) - E(Y(0)|T=1)}^{\text{ATT}} + \overbrace{E(Y(0)|T=1) - E(Y(0)|T=0)}^{\text{Bias}}$$

Marginal mean	Prop. Low Ed.	Mean Low Ed.	Prop. High Ed.	Mean High Ed.
7.38	0.62	7.21	0.38	7.66

Hint: Y is log wage and $\ln(1.25) = 0.22$

Randomized experiments

Identification

$$E(Y|T=1) - E(Y|T=0) \quad = \quad \overbrace{E(Y(1)|T=1) - E(Y(0)|T=1)}^{\text{ATT}}$$

$$+\overbrace{E(Y(0)|T=1) - E(Y(0)|T=0)}^{\text{Bias}}$$

If:

$$E(Y(0)|T=1) = E(Y(0)|T=0)$$

we would have:

$$E(Y|T=1) - E(Y|T=0) = ATT$$

Randomization

$$E(Y(0)|T = 1) = E(Y(0)|T = 0)$$

is a hypothesis, because $E(Y(0)|T = 1)$ can't be observed

Randomization

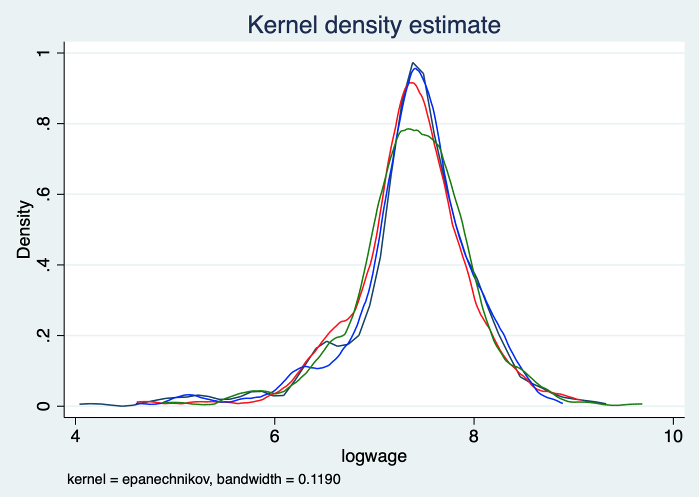
$$E(Y(0)|T = 1) = E(Y(0)|T = 0)$$

is a hypothesis, because $E(Y(0)|T = 1)$ can't be observed

Randomize $T \iff$ Make this hypothesis credible

Randomization illustration

Four $\sim 1\%$ random samples of the log wage data



	Average	N
Initial sample	7.381561	38,270
Smpl 1	7.379043	394
Smpl 2	7.363377	366
Smpl 3	7.390173	380
Smpl 4	7.392236	388
Avg. of avg.	7.3812073	.

(Law of large numbers)

Randomization

On average, randomization balances all characteristics between Treatment and Control

Including those that don't exist and will never happen

Therefore,

$$E(Y(0)|T=1) = E(Y(0)|T=0)$$

is plausible

The $Y(0)$ of the Control is a good estimate for the $Y(0)$ of the Treatment

ATE and ATT

The naive estimator estimates ATT when assignment to treatment is random

It also estimates ATE: why ?

Let’s be more precise

ATE is an estimand (the parameter of interest)

Let's be more precise

ATE is an estimand (the parameter of interest)

Because $ATE = E(Y|T = 1) - E(Y|T = 0)$ (under randomization) we have a natural estimator: the difference in empirical means in a sample

Let's be more precise

ATE is an estimand (the parameter of interest)

Because $ATE = E(Y|T = 1) - E(Y|T = 0)$ (under randomization) we have a natural estimator: the difference in empirical means in a sample

$$\frac{1}{N_1} \sum_{i/T=1} y_i - \frac{1}{N_0} \sum_{i/T=0} y_i$$

Let's be more precise

$$\frac{1}{N_1}\sum_{i/T=1} y_i - \frac{1}{N_0}\sum_{i/T=0} y_i \text{ is an unbiased estimator of ATE :}$$

$$E[\frac{1}{N_1}\sum_{i/T=1} y_i] - E[\frac{1}{N_0}\sum_{i/T=0} y_i]$$

Let's be more precise

$$\frac{1}{N_1}\sum_{i/T=1} y_i - \frac{1}{N_0}\sum_{i/T=0} y_i \text{ is an unbiased estimator of ATE :}$$

$$\begin{aligned} & E[\frac{1}{N_1}\sum_{i/T=1} y_i] - E[\frac{1}{N_0}\sum_{i/T=0} y_i] \\ = \quad & \frac{1}{N_1}\sum_{i/T=1} E[y_i|T_i = 1] - \frac{1}{N_0}\sum_{i/T=0} E[y_i|T_i = 0] \end{aligned}$$

Let's be more precise

$$\frac{1}{N_1}\sum_{i/T=1} y_i - \frac{1}{N_0}\sum_{i/T=0} y_i \text{ is an unbiased estimator of ATE :}$$

$$\begin{aligned}
& E[\frac{1}{N_1}\sum_{i/T=1} y_i] - E[\frac{1}{N_0}\sum_{i/T=0} y_i] \\
= \quad & \frac{1}{N_1}\sum_{i/T=1} E[y_i | T_i = 1] - \frac{1}{N_0}\sum_{i/T=0} E[y_i | T_i = 0] \\
= \quad & \frac{1}{N_1}\sum_{i/T=1} E[y_i(1) | T_i = 1] - \frac{1}{N_0}\sum_{i/T=0} E[y_i(0) | T_i = 0]
\end{aligned}$$

Let's be more precise

$$\frac{1}{N_1}\sum_{i/T=1} y_i - \frac{1}{N_0}\sum_{i/T=0} y_i \text{ is an unbiased estimator of ATE :}$$

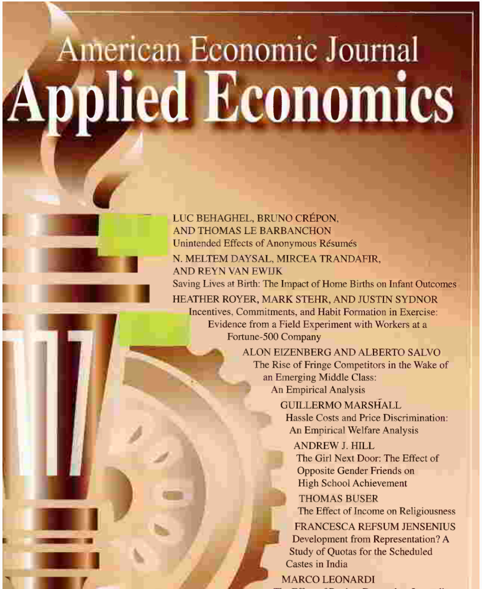
$$\begin{aligned}
& E[\frac{1}{N_1}\sum_{i/T=1} y_i] - E[\frac{1}{N_0}\sum_{i/T=0} y_i] \\
= \quad & \frac{1}{N_1}\sum_{i/T=1} E[y_i|T_i=1] - \frac{1}{N_0}\sum_{i/T=0} E[y_i|T_i=0] \\
= \quad & \frac{1}{N_1}\sum_{i/T=1} E[y_i(1)|T_i=1] - \frac{1}{N_0}\sum_{i/T=0} E[y_i(0)|T_i=0] \\
= \quad & \frac{1}{N_1}\sum_{i/T=1} E[y_i(1)] - \frac{1}{N_0}\sum_{i/T=0} E[y_i(0)]
\end{aligned}$$

Let's be more precise

$$\frac{1}{N_1}\sum_{i/T=1} y_i - \frac{1}{N_0}\sum_{i/T=0} y_i \text{ is an unbiased estimator of ATE :}$$

$$\begin{aligned}
& E[\frac{1}{N_1}\sum_{i/T=1} y_i] - E[\frac{1}{N_0}\sum_{i/T=0} y_i] \\
= \quad & \frac{1}{N_1}\sum_{i/T=1} E[y_i|T_i = 1] - \frac{1}{N_0}\sum_{i/T=0} E[y_i|T_i = 0] \\
= \quad & \frac{1}{N_1}\sum_{i/T=1} E[y_i(1)|T_i = 1] - \frac{1}{N_0}\sum_{i/T=0} E[y_i(0)|T_i = 0] \\
= \quad & \frac{1}{N_1}\sum_{i/T=1} E[y_i(1)] - \frac{1}{N_0}\sum_{i/T=0} E[y_i(0)] \\
= \quad & E[y(1) - y(0)]
\end{aligned}$$

RCTs in Economics



“No causation without manipulation”, Holland 1986

Statistics and Causal Inference

PAUL W. HOLLAND*

Problems involving causal inference have dogged at the heels of statistics since its earliest days. Correlation does not imply causation, and yet causal conclusions drawn from a carefully designed experiment are often valid. What can a statistical model say about causation? This question is addressed by using a particular model for causal inference (Holland and Rubin 1983; Rubin 1974) to critique the discussions of other writers on causation and causal inference. These include selected philosophers, medical researchers, statisticians, econometricians, and proponents of causal modeling.

KEY WORDS: Causal model; Philosophy; Association; Experiments; Mill's methods; Causal effect; Koch's postulates; Hill's nine factors; Granger causality; Path diagrams; Probabilistic causality.

1. INTRODUCTION

The reaction of many statisticians when confronted with the possibility that their profession might contribute to a discussion of causation is immediately to deny that there

where statistics, which is concerned with measurement, has contributions to make. It is my opinion that an emphasis on the effects of causes rather than on the causes of effects is, in itself, an important consequence of bringing statistical reasoning to bear on the analysis of causation and directly opposes more traditional analyses of causation.

2. MODEL FOR ASSOCIATIONAL INFERENCE

The model appropriate for associational inference is simply the standard statistical model that relates two variables over a population. For clarity and for comparison with the model for causal inference described in the next section, however, I will briefly review association here. If I seem overly explicit in describing the model it is only because I wish to be absolutely clear on the fundamental elements of the theory presented here.

Put as bluntly and as contentiously as possible, in this article I take the position that causes are only those things that could, in principle, be treatments in experiments. The qualification “in principle” is important because practical, ethical, and other considerations might make some experiments infeasible, that is, limit us to contemplating *hypothetical experiments*. For example, in the medical and social

I believe that the notion of cause that operates in an experiment *and* in an observational study *is the same*. The difference is in the degree of control an *experimenter* has over the phenomena under investigation compared with that which an *observer* has. In Rubin's model this is ex-

An attribute cannot be a cause in an experiment, because the notion of *potential exposability* does not apply to it. The only way for an attribute to change its value is for the unit to change in some way and no longer be the same unit. Statements of “causation” that involve attributes as “causes” are always statements of association between the values of an attribute and a response variable across the units in a population. In (A) all that is meant is that the performance of women on the exam exceeds, in some sense, that of men.

Ethics

Don't mix "Ethics" and "Politics"

When people say that randomizing a school or labor market intervention is not ethical, they mean that it is politically complicated

If it were non ethical, we would not be running clinical trials

Ethics: Belmont report, USA 1978

1. **Respect for persons**

   Treat people as autonomous agents (or protect those with limited autonomy): capable of deliberation about personal goals

$\rightarrow$ implies information on research and no coercion into research: informed consent

2. **Beneficience**

   Make effort to secure the subject's well-being; maximize benefits, minimize harms. Benefits include social benefits of research; in that sense a bad research is unethical if it requires something from subjects. Beneficience implies Equipoise (do not experiment a treatment that is known to be efficient, or dangerous).

3. **Justice**

   Do not put the burden of risks to a category of people and the benefits to another

Ethics: Institutional Review Boards (IRB)

Defined by the US Department of Health and Human Services under the Code of Federal Regulations, Title 45, Part 46

- Board within academic institutions made of members of the institutions + at least one from outside + at least one non-scientific
- Review all projects (in practice in France: non-medical experimental research) and grant authorization
- Important for most social science research: no more than minimal risk

> (c) An IRB may waive the requirement for the investigator to obtain a signed consent form for some or all subjects if it finds either:
>
> (1) That the only record linking the subject and the research would be the consent document and the principal risk would be potential harm resulting from a breach of confidentiality. Each subject will be asked whether the subject wants documentation linking the subject with the research, and the subject's wishes will govern; or
>
> (2) That the research presents no more than minimal risk of harm to subjects and involves no procedures for which written consent is normally required outside of the research context.

PSE IRB, Trainings on human research

Ethics: GDPR

Handling personal data is closely regulated by law

Ethics: GDPR

Handling personal data is closely regulated by law

Data manipulation must have a legal ground ("léicité" in French)

Ethics: GDPR

Handling personal data is closely regulated by law

Data manipulation must have a legal ground ("léicité" in French)

One is explicit consent; another is public service: the latter grounds most of our work

Ethics: GDPR

Handling personal data is closely regulated by law

Data manipulation must have a legal ground ("léicité" in French)

One is explicit consent; another is public service: the latter grounds most of our work

When you do not need consent, you still need to inform people (very specific wording)

Ethics: GDPR

Handling personal data is closely regulated by law

Data manipulation must have a legal ground ("léicité" in French)

One is explicit consent; another is public service: the latter grounds most of our work

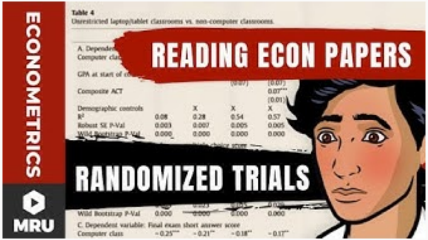
When you do not need consent, you still need to inform people (very specific wording)

You must declare every data manipulation to a Data Protection Officer (DPO): PSE has one, ENS has one, CNRS has one, etc.

RCT Examples

The impact of computer usage on academic performance

Example used in "Mastering Econometrics" series:



Susan Payne Carter, Kyle Greenberg, Michael S. Walker, "The impact of computer usage on academic performance: Evidence from a randomized trial at the United States Military Academy", *Economics of Education Review*, 2017.

Online learning during COVID

AER: Insights 2024, 6(3): 324–340
https://doi.org/10.1257/aeri.20230077

Zooming to Class? Experimental Evidence on College Students' Online Learning during COVID-19†

By Michael S. Kofoed, Lucas Gebhart, Dallas Gilmore, and Ryan Moschitto*

One persistent question in higher education is the efficacy of online education. In the fall of 2020, we randomized 551 West Point students in a required introductory economics course across 12 instructors to either an online or in-person class as a response to the COVID-19 pandemic. Final grades for online students dropped by 0.215 standard deviations, a result apparent in both assignments and exams and largest for academically at-risk students. A postcourse survey finds that online students struggled to concentrate in class and felt less connected to their instructors and peers. Our results show detrimental effects for online learning. (*JEL* A22, I12, I23, I26)

Michael S. Kofoed, Lucas Gebhart, Dallas Gilmore, Ryan Moschitto, "Zooming to Class? Experimental Evidence on College Students' Online Learning during COVID-19", *American Economics Review: Insights*, 2024.

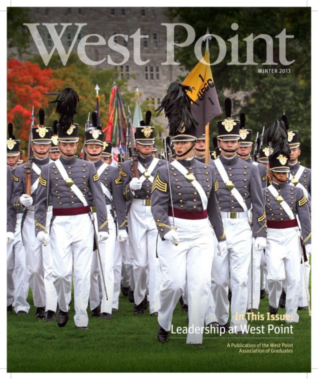
Online learning during COVID

During fall 2020, US Military Academy at West Point switched some students into online instruction

Randomized online vs. in-person to measure impact of type of instruction

Did not assign students to online vs in-person generally, but randomized students to a condition independently for each class-hour

NB: instructors are doing both types, so we do not confound with instructor effects
All the experiment limited to the "Principle of Economics" class...



Online learning during COVID

TABLE 1—SUMMARY STATISTICS COMPARING FULL SAMPLE AND IN-PERSON AND ONLINE CLASSROOMS

	Full sample mean/SD (1)	In-person mean/SD (2)	Online mean/SD (3)	Difference (columns 3–2) b/SE (4)	*F*-stat *F*-stat/*p*-value (5)
Panel A. Student characteristics					
Online	0.612	0.000	1.000		
	(0.488)	(0.000)	(0.000)		
Female	0.230	0.238	0.226	−0.013	0.06
	(0.422)	(0.427)	(0.419)	[0.037]	(0.800)
Black	0.140	0.140	0.139	−0.001	0.00
	(0.347)	(0.348)	(0.347)	[0.030]	(0.956)
Hispanic	0.033	0.042	0.027	−0.015	1.54
	(0.178)	(0.201)	(0.161)	[0.016]	(0.215)
Asian	0.056	0.065	0.050	−0.015	0.02
	(0.231)	(0.248)	(0.219)	[0.020]	(0.899)
NCAA athlete	0.290	0.327	0.267	−0.060	1.15
	(0.454)	(0.470)	(0.443)	[0.040]	(0.284)
Prep school	0.142	0.140	0.142	0.002	0.04
	(0.349)	(0.348)	(0.350)	[0.031]	(0.848)
Prior enlisted	0.169	0.182	0.160	−0.022	0.53
	(0.375)	(0.387)	(0.367)	[0.033]	(0.466)
CEER score	623.806	624.944	623.083	−1.861	0.41
	(73.377)	(77.155)	(70.981)	[6.419]	(0.522)
F-stat for joint significance for all covariates					0.95
					(0.473)
Panel B. Outcomes					
Final grade (standardized)	0.000	0.144	−0.092	−0.236	
	(1.000)	(0.980)	(1.003)	[0.087]	
Final grade (percent)	83.367	84.38	82.73	−1.650	
	(6.998)	(6,860)	(7.018)	[0.608]	
Observations	551	337	214	551	

Link to regression

Observed outcome/counterfactuals

$$Y_i = Y_i(0)(1 - T_i) + Y_i(1)T_i$$

Regression representation

Set:

$$
\begin{aligned}
Y_i(0) &= a + \varepsilon_i(0), \quad E(Y_i(0)) = a, E(\varepsilon_i(0)) = 0 \\
Y_i(1) &= b + \varepsilon_i(1), \quad E(Y_i(1)) = b, E(\varepsilon_i(1)) = 0
\end{aligned}
$$

Regression representation

Set:

$$\begin{aligned} Y_i(0) &= a + \varepsilon_i(0), \quad E(Y_i(0)) = a, E(\varepsilon_i(0)) = 0 \\ Y_i(1) &= b + \varepsilon_i(1), \quad E(Y_i(1)) = b, E(\varepsilon_i(1)) = 0 \end{aligned}$$

$$ATE = E(Y_i(1) - Y_i(0)) =$$

Regression representation

Set:

$$\begin{aligned} Y_i(0) &= a + \varepsilon_i(0), \quad E(Y_i(0)) = a, E(\varepsilon_i(0)) = 0 \\ Y_i(1) &= b + \varepsilon_i(1), \quad E(Y_i(1)) = b, E(\varepsilon_i(1)) = 0 \end{aligned}$$

$$ATE = E(Y_i(1) - Y_i(0)) = b - a + E(\varepsilon_i(1) - \varepsilon_i(0)) = b - a$$

Regression representation

$$Y_i \quad = \quad Y_i(0)(1 - T_i) + Y_i(1)T_i$$

Regression representation

$$\begin{aligned} Y_i &= Y_i(0)(1 - T_i) + Y_i(1)T_i \\ &= Y_i(0) + [Y_i(1) - Y_i(0)]T_i \end{aligned}$$

Regression representation

$$\begin{aligned} Y_i &= Y_i(0)(1 - T_i) + Y_i(1)T_i \\ &= Y_i(0) + [Y_i(1) - Y_i(0)]T_i \\ &= a + \varepsilon_i(0) + [b - a]T_i + [\varepsilon_i(1) - \varepsilon_i(0)]T_i \end{aligned}$$

Regression representation

$$
\begin{aligned}
Y_i &= Y_i(0)(1-T_i) + Y_i(1)T_i \\
&= Y_i(0) + [Y_i(1) - Y_i(0)]T_i \\
&= a + \varepsilon_i(0) + [b-a]T_i + [\varepsilon_i(1) - \varepsilon_i(0)]T_i \\
&= a + [b-a]T_i + \underbrace{[\varepsilon_i(1)T_i + \varepsilon_i(0)(1-T_i)]}_{\text{residual}}
\end{aligned}
$$

$$
\text{with } E[\varepsilon_i(1)T_i + \varepsilon_i(0)(1-T_i)|T_i] = 0
$$

Regression representation

Why $E[\varepsilon_i(1)T_i + \varepsilon_i(0)(1 - T_i)|T_i] = 0$?

Regression representation

Why $E[\varepsilon_i(1)T_i + \varepsilon_i(0)(1-T_i)|T_i] = 0$?

$$E(\varepsilon_i(1)T_i|T_i) = T_i E(\varepsilon_i(1)|T_i)$$

Regression representation

Why $E[\varepsilon_i(1)T_i + \varepsilon_i(0)(1 - T_i)|T_i] = 0$?

$$E(\varepsilon_i(1)T_i|T_i) = T_iE(\varepsilon_i(1)|T_i)$$

Because of randomization $E(\varepsilon_i(1)|T_i = 0) = E(\varepsilon_i(1)|T_i = 1) = 0$

Regression representation

Why $E[\varepsilon_i(1)T_i + \varepsilon_i(0)(1 - T_i)|T_i] = 0$?

$$E(\varepsilon_i(1)T_i|T_i) = T_i E(\varepsilon_i(1)|T_i)$$

Because of randomization $E(\varepsilon_i(1)|T_i = 0) = E(\varepsilon_i(1)|T_i = 1) = 0$

Same for $E(\varepsilon_i(0)(1 - T_i)|T_i)$

Regression representation

You can equivalently:

- Compute the sample analog of $E(Y|T=1) - E(Y|T=0)$
  i.e. the comparison of sample means
- or regress Y on constant and T

Regression representation

You can equivalently:

- Compute the sample analog of $E(Y|T = 1) - E(Y|T = 0)$
  i.e. the comparison of sample means

- or regress Y on constant and T

Are algebraically equivalent *in all cases* Proof

Because no selectivity (in Rubin model terms) or conditional independence (in regression terms), it estimates ATE

Gives a clear interpretation to the coefficient that OLS estimates when "T is independent from the residual" (which is the most obscure statement because the residual depends on the coefficient...)

Using regression?

- The command is ready
- Always use standard errors robust to heteroskedasticity
- Routinely runs tests (and R2, but we tend not to care)
- Can add control variables $\rightarrow$ more on this below

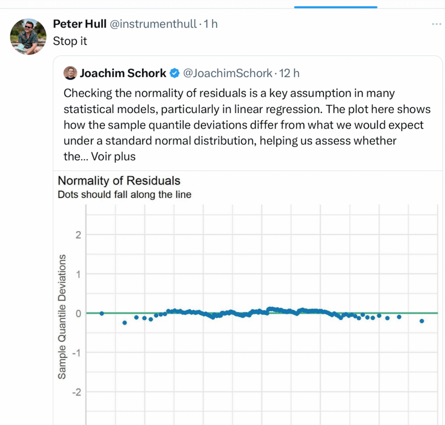
What about normality of the residuals?

Using regression?

- The command is ready
- Always use standard errors robust to heteroskedasticity
- Routinely runs tests (and R2, but we tend not to care)
- Can add control variables → more on this below

What about normality of the residuals?

R: We don't care about normality



Control or not control?

Control or not control?

$$Y_i = a + (b - a)T_i + \varepsilon_i$$

or

$$Y_i = a + (b - a)T_i + \beta x_i + \varepsilon_i'$$

where x would be **predetermined** variables such as age, gender, area dummies, etc.

Control or not control?

$$Y_i = a + (b - a)T_i + \varepsilon_i$$

or

$$Y_i = a + (b - a)T_i + \beta x_i + \varepsilon_i'$$

where x would be **predetermined** variables such as age, gender, area dummies, etc.

Intuitively, $E(x_i|T_i = 1) = E(x_i|T_i = 0)$ so x and T are uncorrelated: x is not the source of a missing variable bias, and there is no need to control for x

Control or not control?

Reasons for controlling

- There can be precision gains (smaller variance of the estimator of the ATE)
- We are unlucky and in a given sample, a variable in x is correlated with treatment: treatement effect would be counfunded

Foundations for control

Set:

$$
\begin{aligned}
Y_i(0) &= a + \beta_0(x - E(x)) + \varepsilon_i(0), \quad E(\varepsilon_i(0)) = 0 \\
Y_i(1) &= b + \beta_1(x - E(x)) + \varepsilon_i(1), \quad E(\varepsilon_i(1)) = 0
\end{aligned}
$$

so that

$$
\begin{aligned}
E(Y_i(0)) &= a + \beta_0 E(x - E(x)) = a \\
E(Y_i(1)) &= b + \beta_1 E(x - E(x)) = b
\end{aligned}
$$

(This is of course a restrictive model in x unless saturated)

Foundations for control

$$Y_i \quad = \quad Y_i(0)(1 - T_i) + Y_i(1)\,T_i$$

Foundations for control

$$
\begin{aligned}
Y_i &= Y_i(0)(1 - T_i) + Y_i(1)T_i \\
&= Y_i(0) + [Y_i(1) - Y_i(0)]T_i
\end{aligned}
$$

Foundations for control

$$\begin{aligned} Y_i &= Y_i(0)(1-T_i) + Y_i(1)T_i \\ &= Y_i(0) + [Y_i(1) - Y_i(0)]T_i \\ &= a + \beta_0(x_i - E(x_i)) + [b-a]T_i + [\beta_1 - \beta_0]\boldsymbol{(x_i - E(x_i)) \times T_i} + \varepsilon_i' \end{aligned}$$

Foundations for control

$$
\begin{aligned}
Y_i &= Y_i(0)(1 - T_i) + Y_i(1)T_i \\
&= Y_i(0) + [Y_i(1) - Y_i(0)]T_i \\
&= a + \beta_0(x_i - E(x_i)) + [b - a]T_i + [\beta_1 - \beta_0]\boldsymbol{(x_i - E(x_i)) \times T_i} + \varepsilon_i'
\end{aligned}
$$

Justifies a model with a control for x (normalized as $x - \bar{x}$) *and* interactions with T

Foundations for control

$$\begin{aligned} Y_i &= Y_i(0)(1 - T_i) + Y_i(1)T_i \\ &= Y_i(0) + [Y_i(1) - Y_i(0)]T_i \\ &= a + \beta_0(x_i - E(x_i)) + [b - a]T_i + [\beta_1 - \beta_0]\boldsymbol{(x_i - E(x_i)) \times T_i} + \varepsilon_i' \end{aligned}$$

Justifies a model with a control for x (normalized as $x - \bar{x}$) *and* interactions with T

1. The standard (non-interacted) model imposes the restriction $\beta_1 = \beta_0$
2. Interestingly, under randomization, the non-interacted model, even if wrong, does identify $b - a = ATE$ ▸ Why?
3. The interacted model improves the precision of the estimator of $(b - a)$, compared to the uncontroled regression of y on T (Lin, 2013)
4. Whereas the non-interacted model does not always improve precision (Freedman, 2008)

In practice

In practice

- Always show the simple comparison (with no controls)

In practice

- Always show the simple comparison (with no controls)
- In finite samples, x and T are never perfectly independent; if you have a long list of possible variables for x you can play with it and change the value of the coefficient on T: data mining !

In practice

- Always show the simple comparison (with no controls)
- In finite samples, x and T are never perfectly independent; if you have a long list of possible variables for x you can play with it and change the value of the coefficient on T: data mining !
- (Exception to the above, if randomization is stratified: discuss this later)
- Thus your list of x must be either imposed by theory, pre-registered, or determined by some data driven algorithm (e.g. Lasso): to be credible, tie your hands

In practice

- Always show the simple comparison (with no controls)
- In finite samples, x and T are never perfectly independent; if you have a long list of possible variables for x you can play with it and change the value of the coefficient on T: data mining !
- (Exception to the above, if randomization is stratified: discuss this later)
- Thus your list of x must be either imposed by theory, pre-registered, or determined by some data driven algorithm (e.g. Lasso): to be credible, tie your hands
- Adding controls is relevant if they *do* predict Y ($\beta_0 \neq 0$) (otherwise no gain, but no cost when properly interacted with T)

In practice

- Always show the simple comparison (with no controls)
- In finite samples, x and T are never perfectly independent; if you have a long list of possible variables for x you can play with it and change the value of the coefficient on T: data mining !
- (Exception to the above, if randomization is stratified: discuss this later)
- Thus your list of x must be either imposed by theory, pre-registered, or determined by some data driven algorithm (e.g. Lasso): to be credible, tie your hands
- Adding controls is relevant if they *do* predict Y ($\beta_0 \neq 0$) (otherwise no gain, but no cost when properly interacted with T)
- If unlucky and some variable is correlated with T then...

In practice

- Always show the simple comparison (with no controls)
- In finite samples, x and T are never perfectly independent; if you have a long list of possible variables for x you can play with it and change the value of the coefficient on T: data mining !
- (Exception to the above, if randomization is stratified: discuss this later)
- Thus your list of x must be either imposed by theory, pre-registered, or determined by some data driven algorithm (e.g. Lasso): to be credible, tie your hands
- Adding controls is relevant if they *do* predict Y ($\beta_0 \neq 0$) (otherwise no gain, but no cost when properly interacted with T)
- If unlucky and some variable is correlated with T then... you're unlucky. No good solution, but practice is to control for that variable

Assume you are interested in the effect on some schooling program on student's test score.

Which baseline controls should you collect to increase the statistical precision of your estimation?

Bad controls

Bad controls

One thing you should **never** do: control for variables that can be affected by the treatment !

$$X(0), X(1), \text{ and } X = TX(1) + (1 - T)X(0)$$

Bad controls

One thing you should **never** do: control for variables that can be affected by the treatment !

$$X(0), X(1), \text{ and } X = TX(1) + (1-T)X(0)$$

Think of controlling as estimating the causal effect conditional on some value of x.

Take $X = 0/1$:

$$E(X_i|X_i = 1, T_i = 1) - E(X_i|X_i = 1, T_i = 0)$$

Bad controls

One thing you should **never** do: control for variables that can be affected by the treatment !

$$X(0), X(1), \text{ and } X = TX(1) + (1-T)X(0)$$

Think of controlling as estimating the causal effect conditional on some value of x.

Take $X = 0/1$:

$$\begin{aligned} & E(X_i|X_i = 1, T_i = 1) - E(X_i|X_i = 1, T_i = 0) \\ = \quad & E(Y_i(1)|X_i(1) = 1, T_i = 1) - E(Y_i(0)|X_i(0) = 1, T_i = 0) \end{aligned}$$

Bad controls

One thing you should **never** do: control for variables that can be affected by the treatment !

$$X(0), X(1), \text{ and } X = TX(1) + (1-T)X(0)$$

Think of controlling as estimating the causal effect conditional on some value of x.

Take $X = 0/1$:

$$\begin{aligned} & E(X_i|X_i = 1, T_i = 1) - E(X_i|X_i = 1, T_i = 0) \\ = \quad & E(Y_i(1)|X_i(1) = 1, T_i = 1) - E(Y_i(0)|X_i(0) = 1, T_i = 0) \\ = \quad & E(Y_i(1)|X_i(1) = 1) - E(Y_i(0)|X_i(0) = 1) \quad \text{(tell me why)} \end{aligned}$$

Bad controls

One thing you should **never** do: control for variables that can be affected by the treatment !

$$X(0), X(1), \text{ and } X = TX(1) + (1-T)X(0)$$

Think of controlling as estimating the causal effect conditional on some value of x.

Take $X = 0/1$:

$$\begin{aligned}
& E(X_i|X_i = 1, T_i = 1) - E(X_i|X_i = 1, T_i = 0) \\
= \quad & E(Y_i(1)|X_i(1) = 1, T_i = 1) - E(Y_i(0)|X_i(0) = 1, T_i = 0) \\
= \quad & E(Y_i(1)|X_i(1) = 1) - E(Y_i(0)|X_i(0) = 1) \quad \text{(tell me why)} \\
= \quad & \underbrace{E(Y_i(1) - Y_i(0)|X_i(1) = 1)}_{\text{causal effect}} + \underbrace{E(Y_i(0)|X_i(1) = 1) - E(Y_i(0)|X_i(0) = 1)}_{\text{selection bias}}
\end{aligned}$$

Take an example to interpret this.

RCT Examples

Online learning during COVID

AER: Insights 2024, 6(3): 324–340
https://doi.org/10.1257/aeri.20230077

Zooming to Class? Experimental Evidence on College Students' Online Learning during COVID-19†

By Michael S. Kofoed, Lucas Gebhart, Dallas Gilmore, and Ryan Moschitto*

One persistent question in higher education is the efficacy of online education. In the fall of 2020, we randomized 551 West Point students in a required introductory economics course across 12 instructors to either an online or in-person class as a response to the COVID-19 pandemic. Final grades for online students dropped by 0.215 standard deviations, a result apparent in both assignments and exams and largest for academically at-risk students. A postcourse survey finds that online students struggled to concentrate in class and felt less connected to their instructors and peers. Our results show detrimental effects for online learning. (*JEL* A22, I12, I23, I26)

Michael S. Kofoed, Lucas Gebhart, Dallas Gilmore, Ryan Moschitto, "Zooming to Class? Experimental Evidence on College Students' Online Learning during COVID-19", *American Economics Review: Insights*, 2024.

Online learning during COVID

During fall 2020, US Military Academy at West Point switched some students into online instruction

Randomized online vs. in-person to measure impact of type of instruction

Did not assign students to online vs in-person generally, but randomized students to a condition independently for each class-hour

NB: instructors are doing both types, so we do not confound with instructor effects
All the experiment limited to the "Principle of Economics" class...

Online learning during COVID

Table 1—Summary Statistics Comparing Full Sample and In-Person and Online Classrooms

	Full sample mean/SD (1)	In-person mean/SD (2)	Online mean/SD (3)	Difference (columns 3–2) b/SE (4)	F-stat F-stat/p-value (5)
Panel A. Student characteristics					
Online	0.612	0.000	1.000		
	(0.488)	(0.000)	(0.000)		
Female	0.230	0.238	0.226	−0.013	0.06
	(0.422)	(0.427)	(0.419)	[0.037]	(0.800)
Black	0.140	0.140	0.139	−0.001	0.00
	(0.347)	(0.348)	(0.347)	[0.030]	(0.956)
Hispanic	0.033	0.042	0.027	−0.015	1.54
	(0.178)	(0.201)	(0.161)	[0.016]	(0.215)
Asian	0.056	0.065	0.050	−0.015	0.02
	(0.231)	(0.248)	(0.219)	[0.020]	(0.899)
NCAA athlete	0.290	0.327	0.267	−0.060	1.15
	(0.454)	(0.470)	(0.443)	[0.040]	(0.284)
Prep school	0.142	0.140	0.142	0.002	0.04
	(0.349)	(0.348)	(0.350)	[0.031]	(0.848)
Prior enlisted	0.169	0.182	0.160	−0.022	0.53
	(0.375)	(0.387)	(0.367)	[0.033]	(0.466)
CEER score	623.806	624.944	623.083	−1.861	0.41
	(73.377)	(77.155)	(70.981)	[6.419]	(0.522)
F-stat for joint significance for all covariates					0.95
					(0.473)
Panel B. Outcomes					
Final grade (standardized)	0.000	0.144	−0.092	−0.236	
	(1.000)	(0.980)	(1.003)	[0.087]	
Final grade (percent)	83.367	84.38	82.73	−1.650	
	(6.998)	(6,860)	(7.018)	[0.608]	
Observations	551	337	214	551	

Online learning during COVID

$$y_{ijdt} = \beta_0 + \beta_1 online_{ijdt} + \gamma_j + \xi_d + \phi_t + X_{ijdt} + \epsilon_{ijdt}, \tag{1}$$

- Can be estimated without controls
- Because rand. within hour (instructor j, day d, time t) these variables can be controls because we have treated and untreated within each slot; in theory, we could control for $(j \times d \times t)$ rather than additive
- We could even have student effects (??)
- Interacted controls x online would in principle be better

Online learning during COVID

TABLE 2—MAIN EFFECTS FOR ONLINE INSTRUCTION

	Final grade (1)	Final grade (2)	Final grade (3)	Final grade (4)	Final grade (5)
Online	−0.236 (0.086)	−0.220 (0.087)	−0.223 (0.086)	−0.218 (0.089)	−0.215 (0.084)
Instructor FEs	No	Yes	Yes	Yes	Yes
Class day FEs	No	No	Yes	Yes	Yes
Time of day FEs	No	No	No	Yes	Yes
Covariates	No	No	No	No	Yes
Observations	551	551	551	551	551
R^2	0.013	0.026	0.026	0.034	0.173
Robust SEs p-values	0.007	0.011	0.010	0.014	0.011
Wild bootstrapped SEs p-values	0.007	0.010	0.009	0.012	0.013

Online learning during COVID

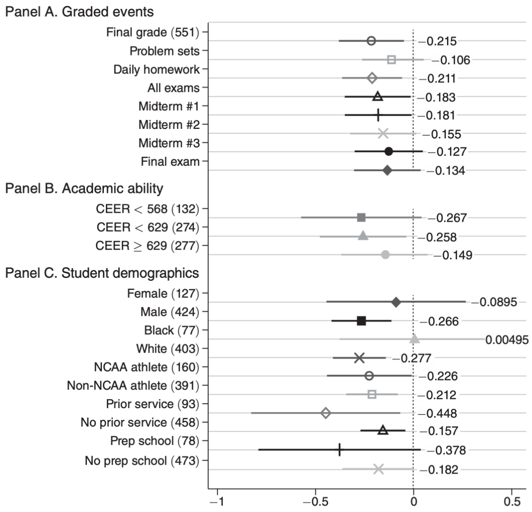


Figure 1. Heterogenous Treatment Effects of Online Instruction

Online learning during COVID

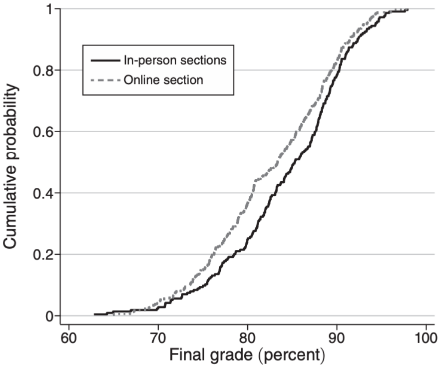


FIGURE 2. CUMULATIVE DISTRIBUTION FUNCTION FOR FINAL COURSE GRADE BY TEACHING MODALITY

Online learning during COVID

TABLE 3—EFFECTS OF ONLINE INSTRUCTION ON POSTCOURSE SURVEY

	Study time (mins.)	Student concentration		Connected to instructor		Instructor cares		Connected to peers	
	OLS	OLS	Ordered probit	OLS	Ordered probit	OLS	Ordered probit	OLS	Ordered probit
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Online	2.352	−0.557	−0.692	−0.342	−0.500	−0.126	−0.200	−0.500	−0.534
	(1.699)	(0.100)	(0.124)	(0.084)	(0.121)	(0.084)	(0.127)	(0.119)	(0.129)
Observations	402	402	402	402	402	402	402	402	402
R^2	0.104	0.146	0.079	0.179	0.079	0.100	0.046	0.123	0.045
Robust SEs *p*-values	0.167	0.000	0.000	0.000	0.000	0.134	0.116	0.000	0.000
Wild bootstrapped SEs *p*-values	0.161	0.000	0.000	0.000	0.000	0.121	0.121	0.001	0.000

Today

We have explored the modelling of a causal model. At the end of this lecture, you should:

- Understand and manipulate formally the notions of counterfactual and causality
- Understand what is a selectivity bias
- Understand why randomization solves "the identification problem"
- Connect the counterfactual representation with the regression representation
- Know how to use regression in practice to analyze RCTs
- Have a notion of Ethics principles for experimental research

Have a nice day

Regression and difference in means

First get rid of constant to simplify:

$$y = \alpha + \beta T + u$$

Regression and difference in means

First get rid of constant to simplify:

$$\begin{aligned} y &= \alpha + \beta T + u \\ \bar{y} &= \alpha + \beta \bar{T} + \bar{u} \end{aligned}$$

Regression and difference in means

First get rid of constant to simplify:

$$
\begin{aligned}
y &= \alpha + \beta T + u \\
\bar{y} &= \alpha + \beta \bar{T} + \bar{u} \\
y - \bar{y} &= \beta(T - \bar{T}) + (u - \bar{u})
\end{aligned}
$$

Regression and difference in means

First get rid of constant to simplify:

$$\begin{aligned} y &= \alpha + \beta T + u \\ \bar{y} &= \alpha + \beta \bar{T} + \bar{u} \\ y - \bar{y} &= \beta(T - \bar{T}) + (u - \bar{u}) \end{aligned}$$

OLS is the empirical counterpart to $Cov(y - \bar{y}, T - \bar{T})/V(T - \bar{T})$

Regression and difference in means

$Cov(y - \bar{y}, T - \bar{T})/V(T - \bar{T})$:
(take $\bar{T} = 1/2$)

$$\frac{\sum_i (y_i - \bar{y})(T_i - \bar{T})}{\sum_i (T_i - \bar{T})^2}$$

Regression and difference in means

$Cov(y - \bar{y}, T - \bar{T})/V(T - \bar{T})$:
(take $\bar{T} = 1/2$)

$$\frac{\sum_i (y_i - \bar{y})(T_i - \bar{T})}{\sum_i (T_i - \bar{T})^2}$$

$$= \frac{\sum_{i/T=1} (y_i - \bar{y})(1/2) + \sum_{i/T=0} (y_i - \bar{y})(-1/2)}{N(1/2)^2}$$

Regression and difference in means

$Cov(y - \bar{y}, T - \bar{T})/V(T - \bar{T})$:
(take $\bar{T} = 1/2$)

$$
\begin{aligned}
& \frac{\sum_i (y_i - \bar{y})(T_i - \bar{T})}{\sum_i (T_i - \bar{T})^2} \\
= & \frac{\sum_{i/T=1} (y_i - \bar{y})(1/2) + \sum_{i/T=0} (y_i - \bar{y})(-1/2)}{N(1/2)^2} \\
= & \frac{1}{N/2} \sum_{i/T=1} (y_i - \bar{y}) - \frac{1}{N/2} \sum_{i/T=0} (y_i - \bar{y})
\end{aligned}
$$

Regression and difference in means

$Cov(y - \bar{y}, T - \bar{T})/V(T - \bar{T})$:
(take $\bar{T} = 1/2$)

$$
\begin{aligned}
& \frac{\sum_i (y_i - \bar{y})(T_i - \bar{T})}{\sum_i (T_i - \bar{T})^2} \\
= & \frac{\sum_{i/T=1} (y_i - \bar{y})(1/2) + \sum_{i/T=0} (y_i - \bar{y})(-1/2)}{N(1/2)^2} \\
= & \frac{1}{N/2} \sum_{i/T=1} (y_i - \bar{y}) - \frac{1}{N/2} \sum_{i/T=0} (y_i - \bar{y}) \\
= & \frac{1}{N/2} \sum_{i/T=1} y_i - \frac{1}{N/2} \sum_{i/T=0} y_i
\end{aligned}
$$

Back

Robustness

You don't include the term $(x_i - E(x_i)) \times T_i$ that belongs to the regression. But,

$$E[(x_i - E(x_i))T_i|T_i] = T_i E[x_i - E(x_i)|T_i] = T_i(E(x_i|T_i) - E(x_i)) = 0$$

So in the regression, $(x_i - E(x_i))T_i$ is not correlated with T_i and thus not a source of omitted variable bias

Back