MQE: Economic Inference from Data: Module 4: Randomized Control Trials

Claire Duquennois

6/9/2020

Causal Inference with non-random assignment:

Randomizing treatment is not always possible:

- the program or policy has already happened
- randomization in unfeasible
- randomizing treatment would be unethical (ex: randomizing exposure to pollutants)

Causal Inference with non-random assignment:

With no randomized control trial we have to assume that the treatment was not randomly assigned:

- treatment will depend on observable and/or unobservable characteristics
- their are important differences between our treated and untreated units that we cannot control for
- Leaving these variables out in the error term will cause OVB.

Differences-in-differences is a way of getting around non-random assignment of a treatment.

DID: Example and intuition[1]

2002: Craigslist opens a new section called "erotic services" in San Francisco:

- mostly used by sex workers to advertise and solicit clients.
- Sex workers claimed it made them safer, because they could solicit indoors from their computers and learn more about the men contacting them.
- Activists and law enforcement worried that it was facilitating sex trafficking and increasing violence against women.

Which was it? Was erotic services (ERS) making women safer, or was it placing them in harm's way?

[1] This section is borrowed from scunning.com

SOS Empiricist!

This is an empirical question: What is the effect of ERS on female safety?

- The fundamental problem of causal inference strikes again!
- We can't know what effect it had because we are missing the data for the counterfactual:

$$E[\tau] = E[\underbrace{Y_{SF,2003}(D_{SF,2003} = 1)}_{\text{observed}}] - E[\underbrace{Y_{SF,2003}(D_{SF,2003} = 0)}_{\text{unobserved}}]$$

In 2003 only the first occurs, and the second is a counterfactual. So how do we proceed?

Diff-in-Diff to the rescue:

The standard differences-in-differences strategy (DiD):

- Define the intervention, $D=$ the availability of the Craigslist ERS webpage.
- We want to know the causal effect, τ of D on $Y=$ female murders.

Can we just compare SF murders in, say, 2003 with some other city, like Pittsburgh?

Differencing A:

....City....	.....Outcome.....
San Francisco	$Y_{SF,2003} = \alpha_{SF} + \tau$
Pittsburgh	$Y_{P,2003} = \alpha_P$

- α_{SF} is a San Francisco fixed effect
- α_P is a Pittsburgh fixed effect.

If make a simple comparison between Pittsburgh and SF:

$$\tilde{\tau} = Y_{SF,2003} - Y_{P,2003} = \alpha_{SF} + \tau - \alpha_P.$$

Differencing A:

The simple difference is biased because of the difference in the underlying murder rates between the two cities:

$$\tilde{\tau} - \tau = \alpha_{SF} - \alpha_{P}.$$

Differencing B:

What if we compare SF to itself? Say in 2003 to 2001?

.....City....	..Time..	.....Outcome.....
San Francisco	Before	$Y_{SF,2001} = \alpha_{SF}$
	After	$Y_{SF,2003} = \alpha_{SF} + \lambda_{03} + \tau$

Again, this doesn't lead to an unbiased estimate of τ since:

$$\tilde{\tau} = \alpha_{SF} + \lambda_{03} + \tau - \alpha_{SF} = \lambda_{03} + \tau$$

We eliminated the city fixed effect but not the changes in the murder rate over time which will bias my estimate:

$$\tilde{\tau} - \tau = \lambda_{03}$$

How can I identify and control for these time effects?

Differencing A+B= Diff-in-Diff

Combining the two approaches to eliminate both the city effects and the time effects:

...City...	..Time..	.....Outcome.....	1st Diff	2nd Diff
SF	Before	$Y_{SF,2001} = \alpha_{SF}$		
	After	$Y_{SF,2003} = \alpha_{SF} + \lambda_{03} + \tau$	$\lambda_{03} + \tau$	
				τ
Pittsburgh	Before	$Y_{P,2001} = \alpha_P$		
	After	$Y_{P,2003} = \alpha_P + \lambda_{03}$	λ_{03}	

The idea

Sometimes treatment and control group outcomes move in parallel in the absence of treatment.

When they do, the divergence of a post-treatment path from the trend established by a comparison group may signal a treatment effect.

The mechanics

Difference-in-differences can be implemented as follows:

1) Compute the difference in the mean outcome variable Y in the post treatment period ($t = 1$) and the before treatment period ($t = 0$) for the control group C:

$$\bar{Y}_{C,1} - \bar{Y}_{C,0} = \Delta \bar{Y}_C$$

$\Rightarrow$ allows us to cancel out the control group fixed effect and identify the time fixed effect since

$$\bar{Y}_{C,1} - \bar{Y}_{C,0} = \alpha_C + \lambda_1 - \alpha_C = \lambda_1 = \Delta \bar{Y}_C$$

The mechanics

2) Compute the difference in the mean outcome variable Y in the post treatment period ($t = 1$) and the before treatment period ($t = 0$) for the treated group T:

$$\bar{Y}_{T,1} - \bar{Y}_{T,0} = \Delta \bar{Y}_T$$

which allows us to cancel out the treated group fixed effect

$$\bar{Y}_{T,1} - \bar{Y}_{T,0} = \alpha_T + \lambda_1 + \tau - \alpha_T = \lambda_1 + \tau = \Delta \bar{Y}_T$$

The mechanics

3) Treatment impact is then measured by the difference-in-differences:

$$(\bar{Y}_{T,1} - \bar{Y}_{T,0}) - (\bar{Y}_{C,1} - \bar{Y}_{C,0}) = (\Delta\bar{Y}_T - \Delta\bar{Y}_C)$$

since by comparing the differences we can cancel out the time fixed effect and isolate the treatment effect of interest:

$$\Delta\bar{Y}_T - \Delta\bar{Y}_C = \lambda_1 + \tau - \lambda_1 = \tau$$

DID Regressions:

This can be done in a regression framework:

$$Y_{it} = \beta_0 + \beta_1 Post_t + \beta_2 GetsTreat_i + \beta_3 Post_t \times GetsTreat_i + u_{it}$$

- $Post_t$ is an indicator for the post treatment period,
- $GetsTreat_i$ is an indicator for observations in the treatment group that eventually gets treated.

DID Regressions:

β_3 is the difference-in-differences estimator:

- estimates the differential impact of being in the post treatment period if you are in the treated group since

$$E[Y_{C,1}] - E[Y_{C,0}] = (\beta_0 + \beta_1) - (\beta_0) = \beta_1$$
$$E[Y_{T,1}] - E[Y_{T,0}] = (\beta_0 + \beta_1 + \beta_2 + \beta_3) - (\beta_0 + \beta_2) = \beta_1 + \beta_3$$

$$(E[Y_{T,1}] - E[Y_{T,0}]) - (E[Y_{C,1}] - E[Y_{C,0}]) = \beta_3$$

A simulation:

Suppose you are a principal of a school:

- ten 4th grade classrooms of 30 students each.
- Starting in 2001 school year, teachers can enroll their class in the scholastic book club
- 4 of your fourth grade teachers opted to enroll.

You are interested in estimating the effect of participation in the book club on 4th grade reading scores.

A simulation:

```
set.seed(6000)
scores<-as.data.frame(rep(c(1,2,3,4,5,6,7,8,9,10),times=30))
names(scores)<-c("class")
scores <- fastDummies::dummy_cols(scores, select_columns = "class")

scores$error<-rnorm(300, mean=0, sd=10)

#suppose teachers in the better performing classes (classes, 7,8,9,10) select to participate in the book cl
scores$treat<-0
scores$treat[scores$class%in%c(7,8,9,10)]<-1

tau<-10

#the data generating process
scores$read4<-(85+tau*scores$treat
               +(-10)*scores$class_1+(-15)*scores$class_2+(-5)*scores$class_3
               +(-8)*scores$class_4+(-3)*scores$class_5+(3)*scores$class_6
               +(5)*scores$class_7+(8)*scores$class_8+(10)*scores$class_9+(12)*scores$class_10
               +scores$error)

scores$year<-"2001"
scores01<-scores
rm(scores)
```

A simulation:

```
scores<-as.data.frame(rep(c(1,2,3,4,5,6,7,8,9,10),times=30))
names(scores)<-c("class")
scores <- fastDummies::dummy_cols(scores, select_columns = "class")

scores$error<-rnorm(300, mean=0, sd=10)

scores$treat<-0
scores$treat[scores$class%in%c(7,8,9,10)]<-1

#the data generating process
scores$read4<-(78
               +(-10)*scores$class_1+(-15)*scores$class_2+(-5)*scores$class_3
               +(-8)*scores$class_4+(-3)*scores$class_5+(3)*scores$class_6
               +(5)*scores$class_7+(8)*scores$class_8+(10)*scores$class_9+(12)*scores$class_10
               +scores$error)

scores$year<-"2000"
scores00<-scores
rm(scores)

scores<-rbind(scores01, scores00)
```

A simulation:

```
regnodid<-felm(read4~treat,scores[scores$year=="2001",])

scores$post<-0
scores$post[scores$year=="2001"]<-1
regdid<-felm(read4~post+treat+post*treat,scores)

regdidfe<-felm(read4~post+treat+post*treat|class,scores)
```

```
## Warning in chol.default(mat, pivot = TRUE, tol = tol): the matrix is either
## rank-deficient or indefinite
```

A simulation:

```
stargazer(regnodid, regdid,regdidfe, type="latex", header=FALSE,
          add.lines = list(c("Class FE", "No", "No", "Yes")), omit.stat = c("ser","rsq", "adj.rsq"))
```

Table 4

	Dependent variable:		
	read4		
	(1)	(2)	(3)
post		7.612***	7.612***
		(1.127)	(1.023)
treat	22.648***	13.553***	
	(1.211)	(1.260)	
post:treat		9.095***	9.095***
		(1.782)	(1.617)
Constant	80.109***	72.496***	
	(0.766)	(0.797)	
Class FE	No	No	Yes
Observations	300	600	600
Note:		*p<0.1; **p<0.05; ***p<0.01	

Identifying Assumption:

Key assumption:

the difference between before and after in the comparison group is a good counterfactual for the treatment group.

- the trend in outcomes of the comparison group is what we would have observed in the treatment group absent the treatment

Identifying Assumption:

This assumption is best illustrated graphically.