Final exam

Jorge Daniel Guevara Acevedo

Saturday, June 20^{th}, 2020

1 Regression discontinuity

Suppose there is a health-based intervention in the United States where if a person's income fell below $\$20,000$ US dollars, the person and their family are enrolled in a program that provides free food. Assume you are interested in the effect that the program has on child health outcomes.

a. If you used a RD design for this project, what is the running variable and what is the cutoff?

The running variable is the person's income and the cut off is $20.000 US dollars. Bellow the cut off are the treated, above the non-treated.

b. What does sorting on the running variable mean and how would you evaluate it in this project?

Sorting on the runnning variable means that the individuals ”choosing” sites above or under the cut off. That problem could be evaluated visually by the graph of the distribution of the running variable or the McCrary Test.

c. What does Barrecca, Guldi, Lindo and Waddell suggest if you find large heaps of observations at regular intervals, including the cutoff?

They suggest drop the problematic units and re-estimate a model they call *donut RDD* using the units around the dropped ones.

d. Explain what we mean by the smoothness assumption. How would you express that assumption in this application?

The smoothness assumption lays in $E[Y_i^0|X = C_0]$ and $E[Y_i^1|X = C_0]$ to be continuous. This implies that all other unobserved determinants of Y are continuously related with the running variable. In this application the distribution of income should be smooth so there is no jump.

e. Describe two auxiliary tests you would perform to evaluate the credibility of the smoothness assumption in this project.

The main test would be a placebo test, which consist in evaluate different parts of the distribution searching for jumps or discontinuities where should not be. The other one, would be the test proposed by Imbens and Lemieux(2008) that suggest to take one side of the discontinuity and take the median value of the running variable. Also in the analysis I would perform a McCrary test over the cut off and observational analysis of the distribution of my running variable.

2 Instrumental variables

Suppose there is a health-based intervention in a small US city which distributes free food to poor families. You are interested in the effect of the food subsidies on child health outcomes. The program is expensive and underfunded. The city runs a lottery and individuals who win the lottery have the option to enroll. No one can participate unless they win the lottery, but conditional on winning, the person still has to choose to participate.

a. Assume that treatment effects are homogenous. Under what conditions will the bias of 2SLS be centered on the bias of OLS?

The bias of 2SLS will be centered on the bias of OLS if we use a weak instrument.

b. How can we provide evidence that the bias of 2SLS is not centered on the bias of OLS?

Knowing the instrument Z, the regression over the instrument will be $X = Z'\pi + \eta$. If there is a bias on OLS:

$$E[\beta_{OLS} - \beta] = \frac{Cov(v, X)}{Var(X)} \tag{1}$$

The expression of the bias caused by the correlation of v_i and η_i is $\frac{\sigma_{v\eta}}{\sigma_\eta^2}$.

After derivating the bias of 2SLS, we obtain:

$$E[\hat{\beta}_{2SLS} - \beta] \approx \frac{\sigma_{v\eta}}{\sigma_\eta^2}\frac{1}{1+F} \tag{2}$$

Where F is the F-statistical for joint significance for the first stage. So, if the First stage is jointly weak then $F \to 0$ thus the bias of 2SLS gets closer to $\frac{\sigma_{v\eta}}{\sigma_\eta^2}$.

c. What are the identifying assumptions for obtaining a consistent estimate of the program's causal effect on child health outcomes under homogeneous treatment effects?

There are two conditions needed:

- $Cov(S, Z) \neq 0$ (Relevance).
- $Cov(A, Z) = Cov(v, Z) = 0$ (Exclusion)

For Y dependent variable, S is the independent one, A and v are unobservable and Z the instrument:

$$Cov(Y, Z) = \delta Cov(S, Z) + Cov(A, Z) + Cov(v, Z) \tag{3}$$

We obtain $\delta_{IV} = \frac{Cov(Y,Z)}{Cov(S,Z)}$

d. Now assume that treatment effects are heterogenous. What is meant by the terms complier, defier, always taker and never taker sub-populations? Interpret each in the context of this example.

- **Compliers:** their status is affected by the treatment in the correct way. That is, $D_i^1 = 1$ and $D_i^0 = 0$. In this case, are those families which win the lottery and enroll the program just because they won, otherwise they wouldn't.
- **Always takers:** they always take the treatment independently of the instrument. In this case, are those families that winning or losing the lottery, will always enroll.
- **Never takers:** they never take the treatment independently of the instrument. In this case, are those families that winning or losing the lottery, will never enroll the program.
- **Defiers:** their status is affected by the treatment in the wrong way. In this case, would be the families that enroll when they lose the lottery and don't enroll when they win it.

e. Are there any always-takers in this program? Why/why not?

No, there is not always takers. If we assume that the program is clean and there is no trick then can't exist always takers. The always takers, always get the treatment no matter of the instrument, but in this case if they lose the lottery, there will be a scenario where they no receive the treatment.

f. Which of the four sub-populations in part (d) contribute to the parameter estimate when using an IV design?

The sub-population that contribute to the parameter estimate are the ones classified as compliers. That is the reason, defiers are a problem, because they partly cancell the effects of compliers.

g. Which parameter is estimated using an IV design with heterogenous treatment effects?

The parameter estimated is the Local average treatment effect (LATE) of D on Y.

h. Why should you use IV to estimate the causal effect of the health intervention on child health outcomes as opposed to merely comparing people on the program to those not on the program?

There could be a lot of reasons, but the main reason is because of the omitted variable bias. That is due to the health outcome of the children depends on many variables we do not possess (or mention) in the example.

3 Difference-in-differences

Suppose there was a health-based intervention in twenty USA cities in 1985 (i.e., no differential timing) that provided free food to poor families with children (based on 1984 incomes) and that you are interested in estimating the effects of this intervention on earnings. The data runs from 1980 to 1990.

a. If you estimated the causal effect of the program on earnings using twoway fixed effects, what are the identifying assumptions needed to obtain an unbiased estimate of the program's average treatment effect on the treated?

In order to obtain an unbiased estimate is necessary to have these identifying assumptions:
i) The regressors are strictly exogenous conditional on the unobserved effect.

$$E[\varepsilon_{it}|D_{i1}, D_{i2}, ..., D_{iT}, u_i] = 0; t = 1, 2, ..., T \tag{4}$$

ii) Regressors must vary over time for at least some i and not be collinear in order that $\hat{\delta}\approx\delta$

$$rank(\sum_{i=1}^{T} E[\ddot{D}'_{it}\ddot{D}_{it}]) = K \tag{5}$$

If the estimator satisfies assumptions i and ii then $\hat{\delta}$ is consistent and unbiased conditional on D.

On the other hand, in the specific method of difference in differences it is necessary to consider two additional assumptions.

i) There is no time-variant city specific unobservables. Nothing unobserved in the twenty USA cities evaluated is changing over time that also determines earnings.
ii) It is assumed that T (free food) is the same for all units.This second assumption is called the parallel trends assumption.

b. If there are dynamic treatment effects, will twoway fixed effects yield an unbiased estimate of the ATT? Why/why not?

Time varying treatment effects, even if they are identical across units, generate cross-group heterogeneity because of the differing post-treatment windows. If the dynamic treatment effects make time-variant Unobserved Heterogeneity then even with fixed effects we are going to have a endogenous problem that makes our estimator biased. Fixed effects is only appropriate if u_i is unchanging

Now suppose there is a health-based intervention in twenty US cities but the programs are introduced to cities at different points in time. The first group of five cities gets treated in 1985, the second group of five in 1990, the third group of five in 1995 and the fourth group of five in 2000. Assume your dataset runs from 1980 to 2005.

c. Write out Goodman-Bacon's 2018 decomposition of twoway fixed effects theorem and explain what each element of the decomposition means.

They explain this TWEF decomposition with the theorem:

$$\hat{\delta}_{kU}^{2x2} = (\bar{y}_k^{post(k)} - \bar{y}_k^{pre(k)}) - (\bar{y}_U^{post(k)} - \bar{y}_U^{pre(k)}) \tag{6}$$

A timing group compared to untreated.

$$\hat{\delta}_{kl}^{2x2} = (\bar{y}_k^{mid(k,l)} - \bar{y}_k^{pre(k)}) - (\bar{y}_l^{mid(k,l)} - \bar{y}_l^{pre(k)}) \tag{7}$$

Group compared to yet-to-be-treated timing group.

$$\hat{\delta}_{lk}^{2x2} = (\bar{y}_l^{post(l)} - \bar{y}_l^{mid(k,l)}) - (\bar{y}_k^{post(l)} - \bar{y}_k^{mid(k,l)}) \tag{8}$$

eventually-treated compared to the already-treated.

So the least squares estimate is a weighted combination of each groups.

$$\hat{\delta}^{DD} = \sum_{k \neq U} s_{kU} \hat{\delta}_{kU}^{2x2} + \sum_{k \neq U} \sum_{I \geq k} s_{kl} [\mu_{kl} \hat{\delta}_{kl}^{2x2,k} + (1 - \mu_{kl}) \hat{\delta}_{lk}^{2x2,l}] \tag{9}$$

d. How many individual 2x2 estimates are there if you estimate this program's effect using twoway fixed effects?

It depends on the timing groups and the untreated groups. If there are K timing groups plus untreated groups, there are K^2 distinct 2x2 estimates.

e. Calculate the variance of treatment for each group. Which group's variance of treatment is largest?

f. Which individual group's 2x2 weight will be largest and why?

g. Goodman-Bacon (2018; 2019) warns against using early groups as controls for later groups. What is the risk of doing so and is it possible to avoid this using twoway fixed effects?

h. Under what assumptions must twoway fixed effects yield the ATT?

i. What are the implications of treatment dynamics when estimating this difference-in-differences model with twoway fixed effects?

Attenuates the bias.

j. What is the main parameter of interest in the Callaway and Sant'anna framework?

The main parameter of interest for Callaway and Santana is:

$$ATT(g,t) = E[Y_t^1 - Y_t^0 | G_g = 1] \tag{10}$$

The Average treatment effect for individuals who are members of a particular group called g in time t.

k. Write down the identifying assumptions of the Callaway and Sant'anna estimator.

The Assumptions are:

1. Sampling of the variables is independent and identically distributed.

2. Conditional parallel trends.

3. Irreversible treatment.

4. Common support (propensity score).

l. How many individual group-time ATT parameters are there in this dataset?

In this data set are four group-time ATT parameters to calculate:

- ATT(1985,1985).

- ATT(1985,1990).

- ATT(1990,1995).

- ATT(1995,2000).

m. Write down the Callaway and Sant'anna estimator. Explain what each individual element of the estimator means.

$$ATT(g,t) = E[(\frac{G_g}{E[G_g]} - \frac{\frac{\hat{p}(X)C}{1-\hat{p}(X)}}{E[\frac{\hat{p}(X)C}{1-\hat{p}(X)}]})(Y_t - Y_{g-1})] \tag{11}$$

Where

- g and t represents the group and time, respectively.

- $\hat{p}_g(X)$ is the generalized propensity score also indicates the probability that an unit is treated conditional on being a member of group g or the control group. (g generalized to 1).

- C is a binary variable that is 1 for units in the control group.

- define G_g to be a binary variable that is 1 for an unit that is first treated in period t.

n. Since there is no "never treated" in this dataset, what do Callaway and Sant'anna suggest you use as controls?

In those cases they propose that we can consider the "not yet treated"$(D_t = 0)$ as a control group insted of "never treated".

o. Can you estimate, using Callaway and Santa'anna, the last group's group-time ATT? Why/why not?

4 Synthetic control

Suppose there was a health-based intervention in Houston Texas in 1985 that provided nutritious food to poor families with children (based on 1984 incomes) and that

we are interested in estimating the effects of this intervention on earnings. The data runs from 1970 to 1995.

a. What are the advantages of using synthetic control to estimate the effect of this program versus a traditional quantitative case study approach?

In some cases, is difficult to address a good control for some experiments. Sometimes the selection could turn ambiguous and arbitrary, in this case the controls would be other families of an aggregate population of Houston. With a synthetic control we can model a control group by weighting average of units that would be part of the control group.

b. In your own words, explain the synthetic control estimator.

A Synthetic control estimate let us make a better approach by comparing several control groups. It models through linear combinations of chosen units (by a process of optimization) to be the synthetic control from a pool of observation groups that were not treated.

c. What is the minimization problem, what are the constraints, what values can the vector of weights take on?

The function to minimize by the synthetic control weights is:

$$\sum_{m=1}^{k} vm(X_{1m} - \sum_{j=2}^{j+1} w_j X_{jm})^{2} \tag{12}$$

And the constrains are this ones, which are related to the weights:

1. $W = (w_2, ..., w_{j+1})', w_j \geq 0, for j = 2, ..., j+1.$

2. $\sum W = w_2 + ... + w_{j+1} = 1$

d. Explain precisely how Abadie, Diamond and Hainmueller (2011) recommend calculating p-values with this estimator.

The procedure they explain is made trough randomization of the treatment to each unit, re-estimating the model, and calculating a set of root mean squared prediction error (RMSPE) values for the pre- and post-treatment period. As follows:

1. Iteratively apply the synthetic control method to each unit in the donor pool and obtain a distribution of placebo effects.

2. Calculate the RMSPE for each placebo for the pre-treatment period:

$$RMSPE = (\frac{1}{T - Y_0} \sum_{t=T_0+t}^{T} (Y_{1t} - \sum_{j=2}^{j+1} w_j^{*} Y_{jt})^{2})^{\frac{1}{2}} \tag{13}$$

3. Calculate the RMSPE for each placebo for the post-treatment period (similar equation but for the post-treatment period).

4. Compute the ratio of the post-to-pre-treatment RMSPE.

5. Sort this ratio in descending order from greatest to highest.

6. Calculate the treatment unit's ratio in the distribution as $p = \frac{RANK}{TOTAL}$

The responses are based on *Causal Inference: The Mixtape.*, *Difference-in-Differences with Multiple Time Periods* by Brantly Callaway and Pedro H. C. Sant'Anna, and the slides of the class.