Problem Set 4

Chris Conlon

Spring 2023

Econometrics I
NYU Stern

Professor Chris Conlon
Email: ctc5@stern.nyu.edu

1 Describing Complier Groups

This problem asks you to analyze Abadie's (2003) χ-function for describing complier groups. Consider a potential outcomes model for treatment effects with imperfect compliance where Z_i a binary instrumental variable which is 1 if individual i is selected as part of the control group and 0 else, and unit i's treatment status is described by

$$D_i = (1 - Z_i)D_{0i} + Z_i D_{1i},$$

where D_{0i}, D_{1i} are the potential values of the treatment indicator for the respective values of Z_i. We also observe the exogenous characteristics X_i for each unit and an outcome variable $Y_i = (1 - D_i)Y_{0i} + D_i Y_{1i}$.

a) Tabulate the possible combinations for values of the potential treatments (D_{0i}, D_{1i}), and assign the labels of "never-takers", "always-takers", "compliers", and "defiers." What is the role of those treatment groups in the interpretation of the 2SLS estimator for a regression of Y_i on D_i using Z_i as an instrumental variable?

b) Now we are interested in characterizing the treatment groups in terms of socio-economic background X_i (e.g. parents' education). Show that $E[X_i]$ is equal to a sum of the conditional expectations

$$\begin{aligned} h_{00} = &\quad E[X_i|D_{0i} = 0, D_{1i} = 0] \\ h_{01} = &\quad E[X_i|D_{0i} = 0, D_{1i} = 1] \\ h_{10} = &\quad E[X_i|D_{0i} = 1, D_{1i} = 0] \\ h_{11} = &\quad E[X_i|D_{0i} = 1, D_{1i} = 1] \end{aligned}$$

weighted by the probabilities $\pi_{01} := P(D_{0i} = 0, D_{1i} = 1)$, etc..

c) Suppose the monotonicity condition holds, i.e. $D_{1i} \geq D_{0i}$ for all individuals, and that (X_i, D_i) is independent of Z_i. Show that

$$E[D_i(1 - Z_i)X_i] = E[X_i|D_{0i} = 1, Z_i = 0]P(D_{0i} = 1|Z_i = 0)P(Z_i = 0).$$

d) Show that under these conditions

$$E[D_i(1 - Z_i)X_i] = E[X_i|D_{0i} = 1, D_{1i} = 1]P(D_{0i} = 1, D_{1i} = 1)P(Z_i = 0).$$

e) Show that under the same conditions

$$E[(1 - D_i)Z_iX_i] = E[X_i|D_{0i} = 0, D_{1i} = 0]P(D_{0i} = 0, D_{1i} = 0)P(Z_i = 1).$$

f) Use your results from parts (b), (d), and (e) to show that for the function

$$w_i := 1 - \frac{D_i(1 - Z_i)}{P(Z_i = 0)} - \frac{(1 - D_i)Z_i}{P(Z_i = 1)}$$

we have

$$E[w_iX_i] = E[X_i|D_{0i} = 0, D_{1i} = 1]P(D_{0i} = 0, D_{1i} = 1)$$

How is that result useful to describe the complier group with respect to the instrumental variable Z_i in terms of other observable characteristics?

2 LATE empirically

Download the provided dataset. It is simulated data based on the STAR class size experiment in Tennessee which attempted to estimate the effect of class size on test scores. Treatment (randomly assigned) is initially being put in a small class but students did not necessarily stay in their assigned classroom.

a) Who are the compliers, always takers, and never takers and who are defiers in this experiment? What would be the concern with the always takers in this case and the expected impact on results?

b) Estimate the LATE in your dataset. Estimate the ATE. Estimate the OLS equation. Why do the above differ?

c) What is the TOT in this case? When if ever will it differ from LATE? Interpret.

d) When might it be useful to look at the LATE and when might it be useful to consider TOT? What is more useful in this case?

e) What is the (theoretically & in this data) is the difference between ITT and ATE?

3 Regression Discontinuity

Consider the RDD model with a structural equation of interest

$$y_i = \beta_0 + x_i\beta_1 + w_i\beta_2 + h^1(z_i) + \epsilon_i$$

Where x is the treatment variable and w is a vector of covariates. The first stage regression (how does treatment change at the discontinuity) is of the form

$$x_i = \pi_0 + D_i\pi_1 + w_i\pi_2 + h^2(z_i) + \nu_i$$

Where $D_i = I(z_i \geq z_0)$ and π_1 is the coefficient of interest. Assume also that h^m is defined as

$$h^m(z_i) = \sum_{j=1}^{\rho} \left[D_i\delta_j^{m+}(z_i - z_0)^j + (1 - D_i)\delta_j^{m-}(z_i - z_0)^j \right]$$

a) Show that if we do not subtract off z_0 in the term $z_i - z_0$ we cannot treat π_1 as the impact of crossing the threshold on treatment. To make your life easier, assume linear terms in $z_i - z_0$ only.

4 Regression Discontinuity Replication

The data is simulated to look similar to the data used by Carpenter and Dobkin (2009), (we can't use the actual data since it is proprietary) they are estimating the effect of alcohol consumption on mortality by utilising the minimum drinking age within a regression discontinuity design.

a) Create a well-labeled scatterplot with mortality on the y-axis and age on the x-axis, mark the minimum drinking age (21)

b) Create a dummy variable indicating whether an individual is above or below the cutoff, also create a variable that indicates how far (in years) each individual is from the threshold (for both 20 year olds and 22 year olds, this variable should be one). Regress with OLS all deaths per 100,000 on your dummy, and your distance to the cutoff. How to interpret your two variables? What constraint does this specification put on the slopes before and after the cutoff?

c) Now add an interaction between your threshold variable and your distance from threshold variable. How do you interpret the coefficient on this interaction?

d) Add your regression lines to your scatterplot in part 1.

e) Now lets use the package "rdd" in R. Adapting the command below, perform an rdd.

```
rdd_model <- RDestimate(formula,
data,
cutpoint = NULL,
kernel = "triangular")
```

f) What is the difference between the regression you ran above and the rdd you performed?

5 Synthetic Controls

We are going to estimate a synthetic control model here using the package "gsynth" by Xiu (2017) using data from Abadie (2021) and Abadie, Diamond, and Hainmueller (2015) (hint: to import you will need to use the package "haven" to convert stata files to R). The intervention in this case is the reunification of Germany, the "treated" unit is the former West Germany and the "donor pool" is a set of industrialized countries and your dependent variable is

a) Graph GDP over time for West Germany and a simple average of GDP for the control countries.

b) Estimate the change in GDP from reunification using a differences in differences method and the synthetic control model. Your code should look something like this:

```
gsynth.out <- gsynth(Y ~ T + X1 + X2,
                     data = df,
                     index = c("unit","time"),
                     force = "two-way",
                     se = TRUE,
                     inference = "parametric",
                     nboots = 1000,
                     parallel = TRUE)
```

Where index specifies the structure of your data, "unit" and "time" should be replaced with the relevant variables in your dataset. The formula is constructed as usual in R. If you want to know more about the other parameters check out the package tutorial available here.

c) Graph GDP over time for West Germany and the GDP of the synthetic control, including confidence intervals for your estimates. (The package will do this for you, check out the tutorial linked above for the different plotting options)

d) What 5 countries are weighted the highest in the synthetic control and what are the weights? One threat to validity is "interference" that is the treatment can spread from the treated unit to the control unit. Why is this a threat in this identification strategy? Is there potential for concern here? Why or why not?