Final Q2

Kenny Mai

12/11/2020

Question 2: Instrumental Variables (4 points).

All worlds should have one binary instrument, one binary treatment variable, and potential outcomes for one variable. You will need to generate potential outcomes both for the outcome variable and the treatment variable (just as in Assignment 5). In World A all assumptions are satisfied. In World B one key assumption is violated. In World C a different key assumption is validated. Define the typical estimand in IV analyses. Estimate a causal effect for this estimand using TSLS.

1) Description of a Hypothetical Real Life Scenario

In New York City(NYC), the Department of Education(DOE) funds a free Saturday and summer academic program called DREAM-SHSI that prepares enrolled seventh graders in the NYC public school system to take the Specialized High Schools Admissions Test (SHSAT) in the eighth grade. Students must meet the following criteria to be eligible to apply for the program: be a NYC resident, be enrolled in seventh grade in a DOE public or charter school, and score a minimum of 3.2 on their grade six English Language Arts NY State test and minimum of 4.0 on their grade six Mathematics NY State test. Students must also either meet certain income requirements or live in specific districts specified by the DOE. The sample data to be generated will be of NYC eighth graders who take the SHSAT. In real life, this number is about 27500 students, of which about 2300 were in the DREAM program. In this example, all the students in the generated sample will meet all the criteria for applying for the DREAM program, and all those who apply are accepted.

2) Data Generating Process

1000 observations will be generated, with 100 eighth graders in the treatment group and 900 eighth graders in the control group.

The instrumental variable, Z, will be an informational pamphlet given to each student by their first or second period teacher, telling the student they qualified and encouraging the student to apply and commit for the program. The treatment variable, D, is whether or not the student applies and commits to the program. The potential outcome variable is the student's SHSAT scaled composite score, ranging between 200 and 800.

A dataset for World A will be created where all assumptions are satisfied. - 20% compliers, 70% always-takers, 10% never-takers, 0% defiers - Average treatment effect for compliers will be 50 - Average Y(0) for compliers is 500 - Average Y(0) for always-takers is 600 - Average Y(0) for never-takers is 400 - Residual standard deviation for everyone in the sample is 30

A dataset for World B will be created where all monotonicity is violated. - 20% compliers, 70% always-takers, 10% never-takers, 5% defiers - Average treatment effect for compliers will be 50 - Average Y(0) for compliers is 500 - Average Y(0) for always-takers is 600 - Average Y(0) for never-takers is 400 - Average Y(0) for defiers is 300 - Residual standard deviation for everyone in the sample is 30

A dataset for World C will be created where exclusion restriction is violated. - 20% compliers, 70% always-takers, 10% never-takers, 0% defiers - Average treatment effect for compliers will be 50 - Average treatment effect for always-takers will be 20 - Average treatment effect for never-takers will be 0 - Average Y(0) for compliers is 500 - Average Y(0) for always-takers is 600 - Average Y(0) for never-takers is 400 - Residual standard deviation for everyone in the sample is 30

3) Assumptions Required

The assumptions required for an unbiased causal estimate of the complier average causal effect (CACE) are as follows.

Ignoribility: The instrument must be independent of the treatment and potential outcomes. In the example used in the problem, the instrument is randomly assigned. This satisfies the assumption.

Exclusion Restriction: For always-takers and never-takers, it must be assumed that there is no effect on their SHSAT scores depending on whether or not they received the pamphlet from school.

Monotonicity: There must be no defiers. These are individuals who, if were predisposed to committing to the DREAM program, would decide to decline upon receiving a pamphlet.

Non-zero correlation between the instrument and the treatment: So long as compliers exist, there is a non-zero correlation between the instrument and the treatment. In other words, when 20% of students are given the pamphlet, there is an effect on whether or not they decide to commit to the program.

SUTVA: In this example, this assumption means students getting into the program do not have an effect on students who do not get in. The commitment to the program of one student is assumed to not have an effect on the SHSAT score of other students.

4) R Code for Data Generating

```r
library(AER)
```

```r
# Generate data for World A
set.seed(0)
fullA <- data.frame(
  C  = c(
        rep("compliers",1000*0.2),
        rep("always-takers",1000*0.7),
        rep("never-takers",1000*0.1)),
  Z  = rbinom(1000,1,0.5),
  D0 = c(
        rep(0,1000*0.2),
        rep(1,1000*0.7),
        rep(0,1000*0.1)),
  D1 = c(
        rep(1,1000*0.2),
        rep(1,1000*0.7),
        rep(0,1000*0.1)),
  Y0 = c(
        round(rnorm(1000*0.2,500,30)),
        round(rnorm(1000*0.7,600,30)),
        round(rnorm(1000*0.1,400,30))),
  Y1 = c(
        round(rnorm(1000*0.2,550,30)),
        round(rnorm(1000*0.7,600,30)),
        round(rnorm(1000*0.1,400,30)))
  )
# Create observed dataset
obsA <- data.frame(
  D = ifelse(fullA$Z==1,fullA$D1,fullA$D0),
  Z = fullA$Z,
  Y = ifelse(fullA$Z==1,fullA$Y1,fullA$Y0)
)
```

```
# Generate data for World B
set.seed(1)
fullB <- data.frame(
  C  = c(
        rep("compliers",1000*0.2),
        rep("always-takers",1000*0.7),
        rep("never-takers",1000*0.05),
        rep("defiers",1000*0.05)),
  Z  = rbinom(1000,1,0.5),
  D0 = c(
        rep(0,1000*0.2),
        rep(1,1000*0.7),
        rep(0,1000*0.05),
        rep(1,1000*0.05)),
  D1 = c(
        rep(1,1000*0.2),
        rep(1,1000*0.7),
        rep(0,1000*0.05),
        rep(0,1000*0.05)),
  Y0 = c(
        round(rnorm(1000*0.2,500,30)),
        round(rnorm(1000*0.7,600,30)),
        round(rnorm(1000*0.05,400,30)),
        round(rnorm(1000*0.05,300,30))),
  Y1 = c(
        round(rnorm(1000*0.2,550,30)),
        round(rnorm(1000*0.7,600,30)),
        round(rnorm(1000*0.05,400,30)),
        round(rnorm(1000*0.05,300,30)))
  )
# Create observed dataset
obsB <- data.frame(
  D = ifelse(fullB$Z==1,fullB$D1,fullB$D0),
  Z = fullA$Z,
  Y = ifelse(fullB$Z==1,fullA$Y1,fullB$Y0)
)
```

```
# Generate data for World C
set.seed(2)
fullA <- data.frame(
  C  = c(
        rep("compliers",1000*0.2),
        rep("always-takers",1000*0.7),
        rep("never-takers",1000*0.1)),
  Z  = rbinom(1000,1,0.5),
  D0 = c(
        rep(0,1000*0.2),
        rep(1,1000*0.7),
        rep(0,1000*0.1)),
  D1 = c(
        rep(1,1000*0.2),
        rep(1,1000*0.7),
        rep(0,1000*0.1)),
  Y0 = c(
        round(rnorm(1000*0.2,500,30)),
        round(rnorm(1000*0.7,600,30)),
        round(rnorm(1000*0.1,400,30))),
  Y1 = c(
        round(rnorm(1000*0.2,550,30)),
        round(rnorm(1000*0.7,620,30)),
        round(rnorm(1000*0.1,400,30)))
  )
# Create observed dataset
obsC <- data.frame(
  D = ifelse(fullA$Z==1,fullA$D1,fullA$D0),
  Z = fullA$Z,
  Y = ifelse(fullA$Z==1,fullA$Y1,fullA$Y0)
)
```

5) Methods and Estimand

The estimand of interest is the complier average causal effect (CACE). This estimand will be estimated using two-stage least squares (TSLS) analysis.

The casual effect can be estimated by using the instrument variable to predict the treatment. Primarily this method targets the "compliers", and the method allows a causal inference to be made for those targetted in the sample, hence the complier part of CACE.

TSLS uses regression to estimate the CACE. The treatment variable is regressed onto the instrument, then predicted values are piped back into the regression model in order to predict the outcome variable.

```
ivregA        <- ivreg(Y ~ D|Z, data = obsA)
ivregB        <- ivreg(Y ~ D|Z, data = obsB)
ivregC        <- ivreg(Y ~ D|Z, data = obsC)
result        <- data.frame(World = c("A","B","C"),CACE = NA,SD = NA)
result$CACE   <- c(summary(ivregA)$coefficients["D","Estimate"],
                   summary(ivregB)$coefficients["D","Estimate"],
                   summary(ivregC)$coefficients["D","Estimate"])
result$SD     <- c(summary(ivregA)$coefficients["D","Std. Error"],
                   summary(ivregB)$coefficients["D","Std. Error"],
                   summary(ivregC)$coefficients["D","Std. Error"])
```

6) Results

The TSLS estimate of the CACE are shown below:

```
result
```

```
##   World      CACE        SD
## 1     A  53.55831  17.70956
## 2     B 100.23605 270.92567
## 3     C 123.32313  12.35493
```

7) Discuss the Bias

As shown by the results, failure to meet all 5 assumptions results in terribly inaccurate CACE estimates.

In the case of violating monotonicity in World B, the proportion of defiers will increase the bias of the causal estimate greatly, and the standard deviation also increases dramatically.

In the case of violating exclusion restriction in World C, bias is increased if the treatment effect on the instrument for the always-takers vs compliers.

8) Conclusion

Instrumental variable analysis as a method for causal inference is ideally used when the ignoribility of the treatment assignment is tenuous, as is the case for many intervention-type programs that use some kind of advertising or promotion. It can also be a good tool when the treatment is conditionally assigned, as opposed to randomly assigned. IV can also be a preferred tool if the target population that the research question asks about is a group that can fall under the definition of a "complier" in regards to a treatment or intervention. Although important in any causal estimation strategy, the satisfaction of the 5 primary assumptions is paramount to estimating a valid CACE. Any violation, as shown in the above example is likely to be magnified dramatically in the estimate.