HW5: Instrumental Variables Simulation

Yongchao Zhao

11/3/2018

Objective

The goal of this exercise is to simulate data consistent with the assumptions of the IV estimator we discussed in class (and described in the Angrist, Imbens, Rubin article posted on the Classes site). We will also evaluate the properties of different approaches to estimating the Complier Average Causal Effect.

Part A: Generate and explore the data for the population

In this section you will simulate data consistent with the assumptions. We will generate data for a sample of 1000 individuals.

Question 1: Simulate the data as god/goddess/supreme being of your choice

(a) Simulate compliance status. Assume that 25% of individuals are compliers, 60% are never takers, and 15% are always takers. Generate D(0) and D(1) vectors to reflect this. You can also generate a vector indicating compliance type, C, if it is helpful to you.

```
# simulate compliance status
dat.full <- data.frame(
    C = c(rep("compliers", 1000*0.25), rep("never-takers", 1000*0.6),
rep("always-takers", 1000*0.15)),
    D0 = NA,
    D1 = NA
)

for(i in 1:nrow(dat.full)){
  if(dat.full$C[i] == "compliers"){
    dat.full$D0[i] = 0
    dat.full$D1[i] = 1
  } else if(dat.full$C[i] == "never-takers"){
    dat.full$D0[i] = 0
    dat.full$D1[i] = 0
  } else {
    dat.full$D0[i] = 1
    dat.full$D1[i] = 1
  }
}


#check the simulated data
table(dat.full$C, useNA = "ifany")
##
## always-takers     compliers  never-takers
##           150           250           600
```

```r
unique(dat.full)
```

```
##                 C D0 D1
## 1       compliers  0  1
## 251  never-takers  0  0
## 851 always-takers  1  1
```

(b) Which compliance group has been omitted from consideration? What assumption does that imply?

Solution.

The group of defiers (D(0) = 1 & D(1) = 0) has been omitted from consideration, which implies the assumption of monotonicity.

(c) Simulate the potential outcomes in a way that meets the following criteria:

i. The exclusion restriction is satisfied.
ii. The average treatment effect for the compliers is 4.
iii. The average Y(0) for never takers is 0; The average Y(0) for compliers is 3; The average Y(0) for always takers is 6.
iv. The residual standard deviation is 1 for everyone in the sample.

```r
# simulate the potential outcomes
set.seed(123)

dat.full$Y0 <- NA
dat.full[dat.full$C == "never-takers",]$Y0 <- rnorm(600, mean = 0, sd = 1)
dat.full[dat.full$C == "always-takers",]$Y0 <- rnorm(150, mean = 6, sd = 1)
dat.full[dat.full$C == "compliers",]$Y0 <- rnorm(250, mean = 3, sd = 1)

dat.full$Y1 <- NA
dat.full[dat.full$C == "never-takers",]$Y1 <- rnorm(600, mean = 0, sd = 1)
dat.full[dat.full$C == "always-takers",]$Y1 <- rnorm(150, mean = 6, sd = 1)
dat.full[dat.full$C == "compliers",]$Y1 <- rnorm(250, mean = 7, sd = 1)

# check the simulated data
library(dplyr)
dat.full %>% group_by(C) %>% summarise(
  count = n(),
  Y0_mean = mean(Y0),
  Y0_sd = sd(Y0),
  Y1_mean = mean(Y1),
  Y1_sd = sd(Y1)
)
```

```
## # A tibble: 3 x 6
##   C             count Y0_mean Y0_sd Y1_mean Y1_sd
##   <fct>         <int>   <dbl> <dbl>   <dbl> <dbl>
## 1 always-takers   150  5.97   1.05   6.05   1.11
## 2 compliers       250  3.03   1.02   7.07   1.03
## 3 never-takers    600  0.0218 0.967  0.0277 0.976
```

(d) Calculate the SATE for each of the compliance groups.

Solution.

Because of the exclusion restriction assumption, the SATE estimates for "always-takers" and "nerve-takers" groups are both close to 0; the estimated SATE for the "compliers" group is around 4.

```
# calculate SATE by compliance groups
dat.full %>% group_by(C) %>% summarise(
      count = n(),
      SATE = mean(Y1 - Y0)
  )
```

```
## # A tibble: 3 x 3
##   C              count    SATE
##   <fct>          <int>   <dbl>
## 1 always-takers   150   0.08379
## 2 compliers       250   4.04073
## 3 never-takers    600   0.00597
```

(e) What is another name for the SATE for the compliers?

Solution.

Another name for the SATE for the compliers is CACE (Complier Average Causal Effect).

(f) Calculate the overall SATE/ITT using your simulated data.

Solution.

From goodness' view, the overall SATE/ITT using the simulated data above is 1.03.

```
#overall SATE
mean(dat.full$Y1 - dat.full$Y0)
```

```
## [1] 1.026337
```

(g) Put D(0), D(1), Y(0), Y(1) into one dataset called dat.full. (You can also include a variable, C, indicating compliance group if you created one.)

```
head(dat.full)
```

```
##           C D0 D1       Y0       Y1
## 1 compliers  0  1 4.538430 6.298495
## 2 compliers  0  1 2.890290 7.882235
## 3 compliers  0  1 3.511471 6.866630
## 4 compliers  0  1 3.213958 5.879322
## 5 compliers  0  1 2.813879 7.461192
## 6 compliers  0  1 2.879606 8.524143
```

Question 2: Playing the role of the researcher to randomize

Now switch to the role of the researcher. Pretend that you are running the experiment that we are examining for this assignment. Generate a binary indicator for the ignorable treatment assignment. Probability of receiving the treatment should be .5.

```
set.seed(234)
ind <- rbinom(n = 1000, size = 1, prob = .5)
table(ind)
```

```
## ind
##   0   1
## 495 505
```

Question 3: Back to playing god (researcher???)

Use dat.full to create a dataset that the researcher would actually get to see given the Z generated in Question 2. It should only have D, Z, and Y in it. Call it dat.obs.

```
set.seed(234)
temp <-dat.full
temp$Z <- ind

temp$D <- ifelse(temp$Z == 1, temp$D1, temp$D0)
temp$Y <- ifelse(temp$Z == 1, temp$Y1, temp$Y0)

dat.obs <- temp %>% select(Z,D,Y)
head(dat.obs)
```

```
##   Z D        Y
## 1 1 1 6.298495
## 2 1 1 7.882235
## 3 0 0 3.511471
## 4 1 1 5.879322
## 5 0 0 2.813879
## 6 1 1 8.524143
```

Question 4: Researcher again

(a) Estimate the percent of compliers, never takers and always takers assuming that there are no defiers. Use only information in dat.obs.

Solution.

Assuming no defiers, based on the information in dat.obs, there is about 25.4% of compliers, and the remaining 74.6% are never takers and always takers combined.

```
summary(lm(D~Z, data=dat.obs))
```

```
## Call:
## lm(formula = D ~ Z, data = dat.obs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4099 -0.4099 -0.1556  0.5901  0.8444
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.15556    0.01946   7.992 3.66e-15 ***
## Z            0.25435    0.02739   9.286  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4331 on 998 degrees of freedom
## Multiple R-squared:  0.07953,    Adjusted R-squared:  0.0786
## F-statistic: 86.22 on 1 and 998 DF,  p-value: < 2.2e-16
```

```
(p.compliers <- coef(lm(D~Z, data=dat.obs))[2])
```

```
##         Z
## 0.2543454
```

(b) Estimate the naive regression estimate of the effect of the treatment on the outcome. Which estimand that we discussed in class is this equivalent to?

Solutions.

The regression shows that the estimate of the effect of the treatment on the outcome is 5.98. It is the "naïve comparison 1-as treated" analysis discussed in class, which simply makes comparisons between those who received the treatment and those who did not.

```
summary(lm(Y~D, data = dat.obs))
## Call:
## lm(formula = Y ~ D, data = dat.obs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3305 -0.9627 -0.1885  0.7128  5.0323
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.52068    0.05252   9.914   <2e-16 ***
## D            5.98091    0.09855  60.690   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.405 on 998 degrees of freedom
## Multiple R-squared:  0.7868, Adjusted R-squared:  0.7866
## F-statistic:  3683 on 1 and 998 DF,  p-value: < 2.2e-16
```

(c) Estimate the intention-to-treat effect.

Solutions.

The estimated ITT is 1.115.

```
summary(lm(Y~Z, data=dat.obs))
## Call:
## lm(formula = Y ~ Z, data = dat.obs)
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.466 -2.352 -1.103  2.817  7.045
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.6562     0.1345  12.316  < 2e-16 ***
## Z             1.1149     0.1892   5.892 5.22e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Residual standard error: 2.992 on 998 degrees of freedom
## Multiple R-squared:  0.03361,    Adjusted R-squared:  0.03264
## F-statistic: 34.71 on 1 and 998 DF,  p-value: 5.223e-09
```

```
(itt <- coef(lm(Y~Z, data=dat.obs))[2])
```

```
##        Z
## 1.114929
```

(d) Calculate an estimate of the CACE by dividing the ITT estimate by the percent of compliers in the sample.

Solutions.

The estimated CACE is 4.38.

Interpretation: the effect of participating in the one-week math boot camp (versus not participating in it) for students who will actually participate in the camp when receiving the encourage email from their department chair, and will not participate if not receiving the encourage email, is about 4.4 points increase on the final math test scores.

```
# calculate CACE: ITT/Pr(compliers)
(cace <- itt/p.compliers)
```

```
##        Z
## 4.383523
```

```
# another way to compute CACE: IIT-Z-on-Y/IIT-Z-on-D
z_on_y <- dat.obs %>% group_by(Z) %>% summarise(mean(Y))
```

```
(itt_z_on_y <- z_on_y[2,2] - z_on_y[1,2])
```

```
##    mean(Y)
## 1 1.114929
```

```
z_on_d <- dat.obs %>% group_by(Z) %>% summarise(mean(D))
```

```
(itt_z_on_d <- z_on_d[2,2] - z_on_d[1,2])
```

```
##     mean(D)
## 1 0.2543454
```

```
(cace2<- itt_z_on_y/itt_z_on_d)
```

```
##      mean(Y)
## 1   4.383523
```

(e) Estimate the CACE by performing two stage least squares on your own (that is without using an IV function in the R package AER).

```
# stage 1: treatment on instrument
stage1.fit <- lm(D ~ Z, data = dat.obs)
d_hat <- fitted.values(stage1.fit)

# stage2: outcome on fitted treatment
stage2.fit <- lm(dat.obs$Y ~ d_hat)
summary(stage2.fit)
## Call:
## lm(formula = dat.obs$Y ~ d_hat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.466 -2.352 -1.103  2.817  7.045
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.9743     0.2315   4.208 2.80e-05 ***
## d_hat         4.3835     0.7440   5.892 5.22e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.992 on 998 degrees of freedom
## Multiple R-squared:  0.03361,    Adjusted R-squared:  0.03264
## F-statistic: 34.71 on 1 and 998 DF,  p-value: 5.223e-09
```

(f) Provide an estimate of CACE and it's standard error by using the ivreg command.

Solutions.

The estimated CACE is <u>4.38</u>, and the standard error is <u>0.393</u>.

```
# way 1: ivreg() in "AER" package
library(AER)
iv.fit <- ivreg(Y ~ D|Z, data = dat.obs)
coef(summary(iv.fit))
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 0.9743351  0.1222194  7.972014 4.249477e-15
## D           4.3835229  0.3927728 11.160454 2.447181e-27

#way 2: tsls() in "sem" package
library(sem)
tsls.fit <- tsls(Y ~ D, ~ Z, data = dat.obs)
coef(summary(tsls.fit))
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 0.9743351  0.1222194  7.972014 4.218847e-15
## D           4.3835229  0.3927728 11.160454 0.000000e+00
```

(g) Simulate a sampling distribution for the estimator used in (f). Is the estimator unbiased? Also report the standard deviation of the sampling distribution and compare to the standard error in (f).

Solutions.

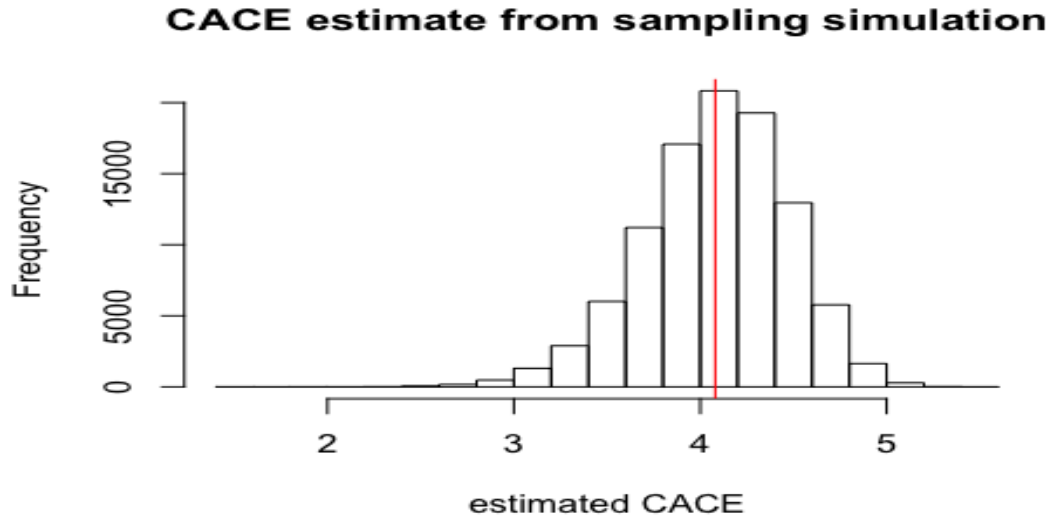
The simulation is conducted 100,000 times with random treatment assignment. According to the sampling distribution result, the mean CACE is 4.08, very close to 4 as the raw data has been generated, thus, the CACE estimator in (f) is unbiased.

The standard deviation of estimated CACE is 0.387, also very close to the standard error obtained in (f) as 0.393.

```
# sampling simulation with 100,000 repetitions
n <- 100000
sim.cace<-rep(NA, n)

for(i in 1:n){
  set.seed(i)
  new.dat.full <- dat.full
  new.dat.full$Z <- rbinom(n = 1000, size = 1, prob = .5)
  new.dat.full$D <- ifelse(new.dat.full$Z == 1, new.dat.full$D1,
new.dat.full$D0)
  new.dat.full$Y <- ifelse(new.dat.full$Z == 1, new.dat.full$Y1,
new.dat.full$Y0)
  new.dat.obs <- new.dat.full %>% select(Z,D,Y)
  fit <- ivreg(Y ~ D|Z, data = new.dat.obs)
  sim.cace[i] = coef(summary(fit))[2]}

summary(sim.cace)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.547   3.839   4.105   4.080   4.349   5.521
sd(sim.cace)
## [1] 0.3869928
hist(sim.cace, main = "CACE estimate from sampling simulation", xlab =
"estimated CACE"); abline(v = mean(sim.cace), col="red")
```



Question 5: Exploring the connection between the DGP and the assumptions.

Now we're back to the role of god of Statistics.

(a) Describe the assumptions required to obtain and unbiased estimate of the treatment effect, as described in AIR. We have generated data that satisfy these assumptions. Suppose instead you were handed data from the study described above. Comment on the plausibility of each of the required assumptions in that setting.

Solutions.

The assumptions required to obtain and unbiasedly estimate the CACE effect are:

- **Ignorability of the instrument:** the instrument (*receiving the encourage email from department chair*) needs to be randomly assigned to the students in this experiment;
- **Exclusion restriction:** for students who would/would not have participated in the math boot camp regardless of receiving their department chair's encourage email, their final math test scores on the final would stay the same;
- **Monotonicity:** assume that there were no students who would participate in the math boot camp if they didn't receive the encourage email, and who would not participate if they received the encourage email;
- **Non-zero correlation between instrument and treatment:** the instrument (*receiving the encourage email from department chair*) would affect a student's decision on participating in the one-week math boot camp;
- **SUTVA:** one student's participation in the math boot camp would not affect another student's math test score on the final.

(b) Suppose that the data generating process above included a covariate that predicted both Z and Y. Which of the assumptions described in (a) would that change and how?

Solutions.

If there is a covariate that predictes both Z and Y, the "ignorability of the instrument" assumption would be affected.

The instrument Z is assumed to be a randomly assigned variable, meaning it is uncorrelated with any unobserved characteristics of sample subjects, with pre-study levels of outcome, and with any other pre-study variables that could predict the outcome Y. Since the covariate is predicting Z and Y, the "ignorability of the instrument" assumption is thus violated.

(c) Suppose that the directions for Q1.c.iii was amended as follows " (iii) The average Y(0) for never takers is 0; The average Y(0) for compliers is 3; The average Y(0) for always takers is 6. The average Y(1) for never takers is 2." Which of the assumptions described in (a) would that violate?

Solutions.

The monotonicity assumptions would be violated, because outcomes for *never takers* would be different in this case: Y(0, never-takers) = 0 vs. Y(1, never-takers) = 2.

(d) Redo one of the commands from Question 1 (just provide the code – you don't have to run it) such that the monotonicy assumption is violated.

```
dat.full[dat.full$C == "never-takers",]$Y0 <- rnorm(600, mean = 0, sd = 1)
# the potential outcome Y1 has mean of 2 instead of 0
dat.full[dat.full$C == "never-takers",]$Y1 <- rnorm(600, mean = 2, sd = 1)
```

(e) How could we alter the study design to preclude the existence of always takers? Would this be ethical?

Solutions.

Before conducting the study, a pre-survey could be sent to all entering undergraduate students in the potential target sample, asking for their attitude towards participating in an optional one-week math boot camp, the suggested answer options could include: (1) Will definitely attend, (2) Might attend, and (3) Not attending. Students who respond that they will attend definitely are considered as "always takers", and then be precluded from the following study with randomly assigned instrument variable (receiving the encourage email) as described above.

However, removing the self-identified "always takers" from the experiment might be ethically questionable, because attending the math boot camp could be indeed a rather beneficial treatment for student's math study, and the treatment would be withheld from those "always takers" who might need or deserve it.