Problem Set 2

Hyoungchul Kim

2025-02-26

First read in some necessary programming packages for analysis:

```
# Load libraries
if (!require(pacman)) install.packages("pacman")
p_load(tidyverse, data.table, fixest, modelsummary, texreg)
```

Theory questions

a.

We learned in class that 2SLS coefficients in finite sample is biased. To be exact, Hahn and Hausman (*Ecma*, 2002) showed that $\mathbb{E}(b_{2SLS} - \beta) \approx \frac{\sigma_{\eta\xi}}{\sigma_{\xi}^2}\frac{1}{F+1}$ where F is the first stagef-statistics, a proxy for the predictive strength of the instrument effect. since F is in the denominator, we can see that higher predictive power will decrease the bias. On the other hand, the predictive strength of the instrument does not affect the consistency of the IV estimator. As long as we have the exclusion condition satisfied, the probability limit of the IV estimator converges to the estimand.

b.

Since LATE is still a type of IV, we need the exclusion and relativity condition to hold. That is, the IV should be correlated with the treatment varible of interest and should only affect the outcome through the treatment variable. (Note that SUTVA, random assignment is also necessary for the IV method to hold).

One extra assumption we need is the **monotonicity condition**. This just means that there are no defier.

Now let's prove the LATE theorem under these maintained assumptions:

Using assumptions above, we can get:

$$\beta_{late} = \frac{\mathbb{E}(Y_i \mid Z_i = 1) - \mathbb{E}(Y_i \mid Z_i = 0)}{\mathbb{E}(D_i \mid Z_i = 1) - \mathbb{E}(D_i \mid Z_i = 0)} \tag{1}$$

$$= \frac{\mathbb{E}[Y_{1i} - Y_{0i} \mid \text{compliers}]P(\text{compliers})}{P(\text{compliers})} \tag{2}$$

$$= \mathbb{E}[Y_{1i} - Y_{0i} \mid \text{compliers}] \tag{3}$$

So this shows that by using the assumption we can get the ATE for the compliers in our data.

c.

You can easily check from the question b that the monotonicity is crucial for retrieving LATE. The nominator in the question b is the reduced form estimate. Without monotonicity, we cannot have the reduced form estimate to equal the change in potential outcome effect for compliers. You can also see that the denominator implies the usefulness of monotonicity in first stage. If we assume there is defiers, the first stage estimate cannot be the probability of being the complier. Thus monotonicity is crucial to pin down the LATE estimate.

d.

If there is no always takers, this means all of our data will be made up of compliers (under the assumption of monotonicity). So this implies that $\mathbb{E}[D_i \mid Z_i = 0] = 0$. Then using the fact that $\mathbb{E}[Y_i \mid Z_i = 1] = \mathbb{E}[Y_{0i}]$ and $\mathbb{E}[Y_i \mid Z_i = 1] = \mathbb{E}[Y_{0i} + (Y_{1i} - Y_{0i})D_i \mid Z_i = 1]$ from the lecture, we get:

$$\beta_{late} = \frac{\mathbb{E}[(Y_{1i} - Y_{0i})D_i \mid Z_i = 1]}{P[D_i = 1 \mid Z_i = 1]} \tag{4}$$

$$= \frac{\mathbb{E}[(Y_{1i} - Y_{0i})(1) \mid D_i = 1, Z_i = 1]P[D_i = 1 \mid Z_i = 1] + 0}{P[D_i = 1 \mid Z_i = 1]} \tag{5}$$

$$= \mathbb{E}[(Y_{1i} - Y_{0i} \mid D_i = 1, Z_i = 1)] \tag{6}$$

$$= \mathbb{E}[(Y_{1i} - Y_{0i}) \mid D_i = 1] = ATT \tag{7}$$

e.

Unlike LATE, MTC allows sorting on unobserved treatment effects. For LATE, if there is unobserved soring, it might not recover the ATE for compliers. MTA in some way alleviates this issue by allowing treatment probability to be correlated with treatment effects. It is also helpful if you want to recover population treatment effect parameters like ATT and not jus the LATE. But we need to assume some functional form restrictions. For example, we assume treatment effect is additive and separable. Also, we assume outcome Y depends on X in a linear, additive form. Using these assumptions, we can write MTC as sum of observed and unobserved components: $MTE(X = x, U_D = p) = x'(\beta_1 - \beta_0) + \mathbb{E}(u_{1i} - u_{0i} \mid p)$. Then we can express $\mathbb{E}[Y_i \mid X = x, P = p] = x'\beta_0 + x'(\beta_1 - \beta_0)p + K(p)$, where laster term on RHS is a polynomial that approximates the unobserved portion of Y. Then we can estimate the MTE by taking the derivative with respect to p.

Numerical questions

a.

First, we read in the data.

```
data <-
 ↪  haven::read_dta("../data/OHIE_Public_Use_Files/OHIE_Data/oregonhie_descriptive_vars.dta")
```

Now we get the balance table by treatment status. We do it for full sample:

Full sample

```
# Convert treatment to a factor and then assign new labels
data <- as.data.table(data)
balance_full <- data[, .(person_id, treatment, birth_year = birthyear_list,
 ↪  female = female_list, english = english_list, msa = zip_msa_list)]
balance_full[, age := 2009 - birth_year]
balance_full[, treatment := as_factor(treatment)]
balance_full[, treatment := fcase(
  treatment == "Selected", "treat",
  treatment == "Not selected", "control"
)]

mean_na <- function(x) mean(x, na.rm=T)
```

Table 1: Balance table (full sample)

	control	treat
age	41.002	40.804
Female: lottery list data	0.557	0.541
Individual requested english-language materials: lottery list data	0.922	0.902
Zip code from lottery list is a metropolitan statistical area	0.773	0.764

```
datasummary(age + female + english + msa ~ treatment * mean_na,
  data = balance_full,
  title = "Balance table (full sample)",
  fmt = 3)
```

Next we do it for people who responded to the survey:

Survery responder

```
data2 <-
 ↪ haven::read_dta("../data/OHIE_Public_Use_Files/OHIE_Data/oregonhie_survey0m_vars.dta")

data2 <- data2 %>%
  select(person_id, employ = employ_hrs_0m, education = edu_0m,
   ↪ starts_with("race"), starts_with("ins"), language = surv_lang_0m, )

data2 <- data2 %>%
  left_join(balance_full, by="person_id")


datasummary(age + female + english + msa + employ + education ~ treatment *
 ↪ mean_na,
  data = data2,
  title = "Balance table (survey responders only)",
  fmt = 3)
```

We can see that there is not much significant differences between control group and treatment group. This seems to imply that randomization is quite well done.

Now let's check ITT:

Table 2: Balance table (survey responders only)

	control	treat
age	41.002	40.804
Female: lottery list data	0.557	0.541
Individual requested english-language materials: lottery list data	0.922	0.902
Zip code from lottery list is a metropolitan statistical area	0.773	0.764
Average hrs worked/week	2.117	2.146
Highest level of education completed	2.231	2.220

```
data3 <-
 ↪ haven::read_dta("../data/OHIE_Public_Use_Files/OHIE_Data/oregonhie_stateprograms_vars.dt

data3 <- data3 %>%
  select(person_id, medicaid = ohp_all_ever_inperson, ohp =
   ↪ ohp_std_ever_inperson)
data3 <- data3 %>% mutate(health = if_else(medicaid > 0 | ohp > 0, 1, 0))
data2 <- data2 %>%
left_join(data3, by="person_id")

itt <- feols(health ~ treatment, data = data2)
```

```
NOTE: 54,177 observations removed because of NA values (LHS: 54,177).
```

```
screenreg(itt, digits=3, custom.model.names = c("Health Care") ,
 ↪ custom.coef.map = list("treatmenttreat" = "Lottery"), include.rsquared =
 ↪ FALSE,
       include.adjrs = FALSE)
```

```
========================
           Health Care
------------------------
Lottery        0.156 ***
              (0.005)
------------------------
Num. obs.  20745
========================
*** p < 0.001; ** p < 0.01; * p < 0.05
```

You can see that the estimate is positive. This makes sense as winning the lottery makes you more likely to get a health care insurance. In a way, this could be understood as policy relevant as this shows you the compliance rate of people the government wants to treat. This estimate implies the magnitude of how likely people will respond to the treatment (lottery) policy.

b.

Using OLS is problematic as people receive medicaid and people who don't are different set of people. That is, the potential outcome (health) will be correlated with the treatment of having medicaid. Thus running OLS is give us a biased result.

Let's look at a balance table between those who received medicaid and the remaining people.

Table 3: Balance table (survey responders only)

	No medicaid	Yes medicaid
age	40.089	40.184
Female: lottery list data	0.522	0.642
Individual requested english-language materials: lottery list data	0.903	0.909
Zip code from lottery list is a metropolitan statistical area	0.997	0.997
Average hrs worked/week	2.203	1.722
Highest level of education completed	2.335	2.227

While there is too big difference, you can see some characteristics differ between two groups than the previous randomization (gender ratio, average hours worked, etc).

c.

d.

e.

f.

g.