

Problem Set 2

Hyoungchul Kim

2025-03-05

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 stage F -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 relevance condition to hold. That is, the IV should be correlated with the treatment variable 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 defiers.

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

Using assumptions above, we can get:

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

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

$$= \mathbb{E}[Y_{1i} - Y_{0i} | compliers] \quad (3)$$

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 $\mathbb{E}[D_i | Z_i = 0] = 0$. Then using the fact that $\mathbb{E}[Y_i | Z_i = 1] = \mathbb{E}[Y_{0i}]$ and $\mathbb{E}[Y_i | Z_i = 1] = \mathbb{E}[Y_{0i} + (Y_{1i} - Y_{0i})D_i | Z_i = 1]$ from the lecture, we get:

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

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

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

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

e.

Unlike LATE, MTC allows sorting on unobserved treatment effects. For LATE, if there is unobserved sorting, 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 just 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: $MTE(X = x, U_D = p) = x'(\beta_1 - \beta_0) + \mathbb{E}(u_{1i} - u_{0i} | p)$. Then we can express $\mathbb{E}[Y_i | X = x, P = p] = x'\beta_0 + x'(\beta_1 - \beta_0)p + K(p)$, 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.

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

Full sample

```
data_des <-  
  ↪ haven::read_dta("../data/OHIE_Public_Use_Files/OHIE_Data/oregonhie_descriptive_vars.dta")  
data_state <-  
  ↪ haven::read_dta("../data/OHIE_Public_Use_Files/OHIE_Data/oregonhie_stateprograms_vars.dta")  
data_0 <-  
  ↪ haven::read_dta("../data/OHIE_Public_Use_Files/OHIE_Data/oregonhie_survey0m_vars.dta")  
data_12 <-  
  ↪ haven::read_dta("../data/OHIE_Public_Use_Files/OHIE_Data/oregonhie_survey12m_vars.dta")  
  
data <- data_des %>%  
  left_join(data_state, by="person_id") %>%  
  left_join(data_0, by = "person_id") %>%  
  left_join(data_12, by = "person_id")  
  
# Convert treatment to a factor and then assign new labels
```

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

```
data <- as.data.table(data)
data[, age := 2009 - birthyear_list]
data[, treatment := as_factor(treatment)]
data[, treatment := fcase(
  treatment == "Selected", "treat",
  treatment == "Not selected", "control"
)]

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

datasummary(age + female_list + english_list + zip_msa_list ~ treatment *
  ↪ mean_na,
  data = data,
  title = "Balance table (full sample)",
  fmt = 3)
```

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

Survey responder

```
datasummary(age + female_list + english_list + zip_msa_list + employ_hrs_12m
  ↪ + edu_12m ~ treatment * mean_na,
  data = data,
  title = "Balance table (survey responders only)",
  fmt = 3)
```

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	2.059	2.106
Highest level of education completed	2.280	2.274

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:

```
# itt1 <- feols(hosp_any_12m ~ treatment + hhsize_12m + draw_lottery +
  ↪ hosp_any_0m + er_any_0m + not_er_noner_0m, data = data, weights =
  ↪ ~weight_12m, cluster = ~household_id)
itt1 <- feols(hosp_any_12m ~ treatment, data = data)

itt2 <- feols(er_any_12m ~ treatment, weights = ~ weight_12m, data = data,
  ↪ cluster = ~ household_id)

library(texreg)
screenreg(list(itt1, itt2))
```

```
=====
                        Model 1      Model 2
-----
(Intercept)              0.07 ***      0.26 ***
                        (0.00)      (0.00)
treatmenttreat           0.00          -0.00
                        (0.00)      (0.01)
-----
Num. obs.                23609        23514
R^2 (full model)          0.00          0.00
R^2 (proj model)
Adj. R^2 (full model)     -0.00        -0.00
Adj. R^2 (proj model)
=====
*** 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
birthyear_list	1968.322	1965.767
Female	0.539	0.630
English materials requested	0.905	0.931
ZIP is an MSA	0.767	0.753
Average hrs worked/week	2.208	1.624

You can see some characteristics differ between two groups than the previous randomization (gender ratio, average hours worked, etc).

c.

Let's run the result. For indicator of medicaid, we will use currently on medicaid, ever on medicaid as the variable. The coefficients in this case would be ATE but as there is selection problem, they would be biased. First result used Every on medicaid as the variable.

	Current prescription	Any primary care	Any ER
Ever on medicaid	0.12 *** (0.01)	0.16 *** (0.01)	0.05 *** (0.01)
Num. obs.	18308	23492	23514
R ² (full model)	0.01	0.02	0.00
R ² (proj model)			
Adj. R ² (full model)	0.01	0.02	0.00
Adj. R ² (proj model)			

*** p < 0.001; ** p < 0.01; * p < 0.05

Now I will use currently on medicaid as the variable.

	Current prescription	Any primary care	Any ER
--	----------------------	------------------	--------

Currently on medicaid	0.20 *** (0.01)	0.32 *** (0.01)	0.06 *** (0.01)

Num. obs.	18047	23157	23177
R ² (full model)	0.04	0.10	0.00
R ² (proj model)			
Adj. R ² (full model)	0.04	0.10	0.00
Adj. R ² (proj model)			
=====			
*** p < 0.001; ** p < 0.01; * p < 0.05			

d.

For monotonicity, I guess it would be assuming that people who got into lottery will not have tendency to not get a health insurance. I think monotonicity would hold in this case as the lottery is just giving you chance to apply for health care (OHS Standard). Thus it will at least weakly make people apply for health care and get one.

Also, exclusion in this case would mean the lottery will only affect the outcome (health and hospital related outcome) only through the treatment (getting a health care). This also seems reasonable as lottery is just like an experiment. Thus it is very unlikely it will directly affect the outcome.

e.

We run the first stage regression result using two possible measures of medicaid coverage.

=====		
	Lottery	Lottery

Ever on medicaid	0.49 *** (0.01)	
Currently on medicaid		0.18 *** (0.01)

Num. obs.	33823	23350
R ² (full model)	0.13	0.03
R ² (proj model)		
Adj. R ² (full model)	0.13	0.03
Adj. R ² (proj model)		

```
=====
*** p < 0.001; ** p < 0.01; * p < 0.05
```

f.

Compliers in our case are people who actually receive medicaid after they win the lottery, but would not have gotten it if they had not won the lottery. To characterize the complier group, we can try to difference the means of medicaid variables for lottery winners and losers. This is somewhat like a first stage estimates. The estimates for “curenly on medicaid” is:

```
# A tibble: 2 x 2
  Lottery      Mean
  <dbl+lbl>   <dbl>
1 0 [Not selected] 0.319
2 1 [Selected]    0.508
```

So difference in mean is about 0.2. We can also do it for “every on medicaid”:

```
# A tibble: 2 x 2
  Lottery      Mean
  <dbl+lbl>   <dbl>
1 0 [Not selected] 0.0276
2 1 [Selected]    0.285
```

g.

We will get the LATE estimates using 2SLS regression. This is ATE result for the compliers only. If we assume there is heterogeneous effects, we should not think of this estimate as the ATT.

	Curenly taking med	Any primary care	Any ER
Currently on medicaid	0.12 * (0.05)	0.34 *** (0.04)	0.02 (0.04)
Num. obs.	17215	21979	21992
Num. groups: hhsize_12m^draw_survey_12m	102	108	109
R^2 (full model)	0.06	0.11	0.01

R ² (proj model)	0.04	0.10	0.00
Adj. R ² (full model)	0.06	0.11	0.01
Adj. R ² (proj model)	0.04	0.10	0.00

*** p < 0.001; ** p < 0.01; * p < 0.05

	Curenly taking med	Any primary care	Any ER
Ever on medicaid	0.07 * (0.03)	0.19 *** (0.03)	0.01 (0.02)
Num. obs.	17451	22277	22294
Num. groups: hhsize_12m^draw_survey_12m	103	108	109
R ² (full model)	0.04	0.03	0.01
R ² (proj model)	0.01	0.02	0.00
Adj. R ² (full model)	0.03	0.02	0.00
Adj. R ² (proj model)	0.01	0.02	0.00

*** p < 0.001; ** p < 0.01; * p < 0.05