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 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.

Now we get the balance table by treatment status.

Variable	Control_Mean	Treatment_Mean	Difference	Control_SD	Treatment_SD	SE
Birth Year	1968.0367979	1968.2898936	0.2530956	0.0934849	0.0926367	0.1316151
Female	0.5563614	0.5391469	-0.0172144	0.0038371	0.0038168	0.0054126
English Materials Requested	0.9085177	0.9020445	-0.0064732	0.0022266	0.0022760	0.0031853
Signed Self Up	0.8763579	0.8324037	-0.0439542	0.0025423	0.0028599	0.0038320
Signed Up First Day	0.0900322	0.0953294	0.0052972	0.0022107	0.0022486	0.0031543
Gave Phone Number	0.8706070	0.8686516	-0.0019555	0.0025923	0.0025863	0.0036623
Gave PO Box as Address	0.1145806	0.1168259	0.0022453	0.0024600	0.0024595	0.0034791
ZIP is an MSA	0.7693568	0.7619855	-0.0073713	0.0032535	0.0032608	0.0046071

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:

Variable	ITT_Estimate	SE
Any Prescription in 12 Months	0.0121	0.0071
Any Doctor Visit in 12 Months	0.0555	0.0064
Any ER Visit in 12 Months	-0.0007	0.0057
Any Hospitalization in 12 Months	0.0000	0.0034

Somehow I am getting bit weird coefficient. This should be all positive as winning the lottery will make people more likely to get a health care insurance.

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 1: 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 mediciad as the variable.

```
=============================================================================
                        Current prescription  Any primary care  Any ER
-----------------------------------------------------------------------------
Ever on mediciad             0.12 ***              0.16 ***          0.05 ***
                            (0.01)                (0.01)            (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 ***              0.32 ***          0.06 ***
                          (0.01)                (0.01)            (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 fist stage regrssion result using two possible measures of medicaid coverage.

Variable	Control_Mean	First_Stage	SE
Ever on Medicaid	0.141	0.2565	0.0046
Currently on Medicaid	0.053	0.0907	0.0032

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 table below shows the complier share for two medicaid measures.

We can also use background info on people to characterize the composition of the complier group. This can be thought of as first stage coefficient of people with $X_i = 1$ (people with certain characteristics) divided by the coefficient of the full sample.I wrote down some values for few demographics. It seems that complier group is more older less likely to be female, and signed up on the first day more often.

Demographic	Ever on Medicaid	Currently on Medicaid
Share of Compliers	0.257	0.189
old (Born before 1968)	1.144	1.160
female	0.975	0.947
signed self up	1.067	1.089

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.

Variable	LATE	SE
Any Prescription in 12 Months	0.0422	0.0246
Any Doctor Visit in 12 Months	0.1908	0.0216
Any ER Visit in 12 Months	-0.0023	0.0197
Any Hospitalization in 12 Months	0.0001	0.0116

R codes

```
# downloaded data from "https://data.nber.org/oregon/4.data.html" and
 ↪  unzipped it.

# Using "oregon_hie_qje_replication.do" file, I copy data files from input
 ↪  folder to the data folder under name "repl_code".
```

```r
################################################################################
# AUTHORS: Hyoungchul Kim
# CREATED: 2023-02-24
# PURPOSE: Solve coding portions of HCMG 901 Problem Set 2 using tidyverse
################################################################################

library(tidyverse)
library(haven)
library(fixest)
library(modelsummary)
library(kableExtra)

# Load Data
analysis_data <- read_dta("repl_data/data_for_analysis.dta")


# Define variable lists with labels
variable_labels <- c(
  birthyear_list = "Birth Year",
  female_list = "Female",
  english_list = "English Materials Requested",
  self_list = "Signed Self Up",
  first_day_list = "Signed Up First Day",
  have_phone_list = "Gave Phone Number",
  pobox_list = "Gave PO Box as Address",
  zip_msa = "ZIP is an MSA",
  rx_any_12m = "Any Prescription in 12 Months",
  doc_any_12m = "Any Doctor Visit in 12 Months",
  er_any_12m = "Any ER Visit in 12 Months",
  hosp_any_12m = "Any Hospitalization in 12 Months",
  ohp_all_ever_survey = "Ever on Medicaid",
  ohp_all_at_12m = "Currently on Medicaid"
)

baseline_list <- names(variable_labels)[1:8]
survey_useext_list <- names(variable_labels)[9:12]
mdcd_covg_vars <- names(variable_labels)[13:14]

# Function to save results as LaTeX tables with labels
save_tex <- function(data, filename) {
  data <- data %>%
    mutate(Variable = recode(Variable, !!!variable_labels)) %>%
```

```r
    mutate(across(where(is.numeric), ~ round(.x, 4)))  # Round all numeric
     ↪ values to 2 decimal places
  tex_output <- data %>%
    kbl(format = "latex", booktabs = TRUE, caption = filename) %>%
    kable_styling(latex_options = c("hold_position"))

  writeLines(tex_output, filename)
}



# Balance Table by Lottery Outcome
balance_results <- map_dfr(baseline_list, function(var) {
  filtered_data <- analysis_data %>%
    filter(!is.na(weight_12m), !is.na(.data[[var]]))

  if (nrow(filtered_data) == 0) return(NULL)

  control_stats <- feols(as.formula(paste(var, "~ 1")),
                         data = filter(filtered_data, treatment == 0),
                         weights = ~ weight_12m)

  treatment_stats <- feols(as.formula(paste(var, "~ 1")),
                           data = filter(filtered_data, treatment == 1),
                           weights = ~ weight_12m)

  diff_stats <- feols(as.formula(paste(var, "~ treatment")),
                      data = filtered_data,
                      weights = ~ weight_12m)

  tibble(
    Variable = var,
    Control_Mean = coef(control_stats)[1],
    Treatment_Mean = coef(treatment_stats)[1],
    Difference = coef(diff_stats)[2],
    Control_SD = sqrt(vcov(control_stats)[1,1]),
    Treatment_SD = sqrt(vcov(treatment_stats)[1,1]),
    SE = sqrt(vcov(diff_stats)[2,2])
  )
})
save_tex(balance_results, "tab_a_balance.tex")

# ITT Effects of Lottery on Healthcare Usage
```

```r
itt_results <- map_dfr(survey_useext_list, function(var) {
  filtered_data <- analysis_data %>% filter(!is.na(weight_12m),
 ↪  !is.na(.data[[var]]))
  if (nrow(filtered_data) == 0) return(NULL)

  model <- feols(as.formula(paste(var, "~ treatment")),
                 data = filtered_data,
                 weights = ~ weight_12m)

  tibble(
    Variable = var,
    ITT_Estimate = coef(model)[2],
    SE = sqrt(vcov(model)[2,2])
  )
})
save_tex(itt_results, "tab_a_itt.tex")

# First-Stage Estimates
first_stage_results <- map_dfr(mdcd_covg_vars, function(var) {
  filtered_data <- analysis_data %>% filter(!is.na(weight_12m),
 ↪  !is.na(.data[[var]]))
  if (nrow(filtered_data) == 0) return(NULL)

  model <- feols(as.formula(paste(var, "~ treatment")),
                 data = filtered_data,
                 weights = ~ weight_12m)

  tibble(
    Variable = var,
    Control_Mean = coef(model)[1],
    First_Stage = coef(model)[2],
    SE = sqrt(vcov(model)[2,2])
  )
})
save_tex(first_stage_results, "tab_e.tex")

# 2SLS Estimation for LATE
late_results <- map_dfr(survey_useext_list, function(var) {
  filtered_data <- analysis_data %>% filter(!is.na(weight_12m),
 ↪  !is.na(.data[[var]]))
  if (nrow(filtered_data) == 0) return(NULL)

  model <- feols(as.formula(paste(var, "~ 1 | ohp_all_ever_survey ~
 ↪  treatment")),
```

```
                  data = filtered_data,
                  weights = ~ weight_12m)

  tibble(
    Variable = var,
    LATE = coef(model)[2],
    SE = sqrt(vcov(model)[2,2])
  )
})
save_tex(late_results, "tab_g.tex")




# Compliers Analysis
analysis_data <- analysis_data %>%
  mutate(
    old = birthyear_list < 1968,
    fem = female_list,
    self = self_list
  )

dem_vars <- c("old", "fem", "self")
RHS_vars <- c("ohp_std_ever_survey", "ins_any_12m")

complier_results <- map_dfr(RHS_vars, function(RHS_var) {
  treated <- analysis_data %>% filter(treatment == 1) %>% summarise(mean =
↪  mean(.data[[RHS_var]], na.rm = TRUE)) %>% pull(mean)
  untreated <- analysis_data %>% filter(treatment == 0) %>% summarise(mean =
↪  mean(.data[[RHS_var]], na.rm = TRUE)) %>% pull(mean)
  csize <- round(treated - untreated, 3)

  frac_results <- map_dfr(dem_vars, function(var) {
    treated_var <- analysis_data %>% filter(treatment == 1, .data[[var]] ==
↪  1) %>% summarise(mean = mean(.data[[RHS_var]], na.rm = TRUE)) %>%
↪  pull(mean)
    untreated_var <- analysis_data %>% filter(treatment == 0, .data[[var]] ==
↪  1) %>% summarise(mean = mean(.data[[RHS_var]], na.rm = TRUE)) %>%
↪  pull(mean)
    frac <- round((treated_var - untreated_var) / csize, 3)
    tibble(Demographic = var, Outcome = RHS_var, Fraction = frac)
  })
```

```r
  bind_rows(tibble(Demographic = "Share Compliers", Outcome = RHS_var,
   ↪  Fraction = csize), frac_results)
})

# Save as LaTeX
complier_tex <- complier_results %>%
  pivot_wider(names_from = Outcome, values_from = Fraction) %>%
  mutate(Demographic = recode(Demographic, !!!variable_labels)) %>%
  kbl(format = "latex", booktabs = TRUE, caption = "Meet The Compliers") %>%
  kable_styling(latex_options = "hold_position")

writeLines(complier_tex, "QDc_compliers.tex")
```