Inequality Pset 4

Julia Du

2021-05-27

Contents

Load necessary libraries

```
library(tidyverse)
library(lubridate)
library(tinytex)
library(stargazer)
library(lfe)

theme_set(theme_minimal())
```

Question 1

The research question is the effect of expanding access to public health insurance (Medicaid) on the health care use, financial strain, and health of low-income adults.

It's interesting because the estimates of public health insurance apply to able-bodied uninsured adults below 100% of the poverty line, which is a population of significant policy interest. It's also interesting there hasn't been a lot of research on benefits of health insurance outside health care utilization. This question lets us look at the other, "risk-spreading" benefits of health insurance, like less financial strain or less emotional stress/worry. Finally, it's interesting because the Oregon experiment is a randomized control trial (RCT), making it the gold standard for internal validity. If we find in our data something contrary to our expectations, RCTs make us confident enough in our experimental design that we don't chalk up surprises to wrong empirical specifications.

Question 2

(2a)

The identifying assumption is that, in the absence of treatment, the treatment group would've had the same distribution of outcomes as the control group.

There are 3 possible violations:

- nonrandom assignment (i.e. random assignment is done incorrectly)
  - i.e. that the assignment of the ability to apply for Medicaid (i.e. lottery results) were randomized & that the treatment/control individuals in the subsamples were not differentially selected from the full sample
  - test: covariate balance test
- differential participation in experiment across treatment groups
  - i.e. attrition rates or people's decision to drop out of experiment (i.e. not respond to survey) is a function of their assigned treatment group. in other words, the decision to drop out is NOT random
  - test: covariate balance test among those who participated, looking @ participation rates across treatment groups
- differential reporting of outcomes across groups
  - e.g. people randomized into opportunity to apply for Medicaid (win lottery, this is the treatment group) may tend to report their health care usage more than those who lose the lottery

    In this context, the line between differential participating and differential reporting is a blurry (i.e. pretty much the same thing), per office hours with Prof. Deshpande.

(2b) To ensure that the lottery indeed randomly selected people, Finkelstein et al. verified the selection process with independent computer simulations within sampling error (pg. 1074). They also demonstrated that this selection procedure created survey samples with a balance of treatment and control traits (pg. 1074) – essentially a covariate balance test.

`They also looked to address potential differential participation in experiment, as reflected by respons`

The structural equation is:

$$Y_{ipjst} = \beta_0 + \beta_1(Income_{ipjst}) + \varepsilon_{ipjst}$$

where i is the individual. The other subscripts will be discussed later.

Question 4b

If we estimate the structural equation in cross-sectional data, then our estimate of the causal parameter of interest β_1 will suffer from Omitted Variable Bias (OVB), as income is endogeneous because it is correlated with unobserved health care inputs. In general, those with better health tend to have higher income.

Recall:

$$OVB = \beta_1^{OLS} - \beta_1 = \frac{cov(Income_{ipjst}, \varepsilon_{ipjst})}{var(Income_{ipjst})}$$

So, the direction of bias depends on the sign of the covariance. In this case, it's likely our estimate is upward-biased since income tends to be positively correlated with unobserved health.

Question 4c

4c(i). I have included a graph of this a few pages earlier.

The 1st diff-in-diff is:

$$\bar{Y}_{2plus,after} - \bar{Y}_{1,after}$$

This is the difference in infant health between families with 2 or more kids and with families with just one kid after the EITC expansion occurred.

The 2nd diff-in-diff is:

$$\bar{Y}_{2plus,before} - \bar{Y}_{1,before}$$

This is the difference in infant health between families with 2 or more kids and with families with just one kid before the EITC expansion occurred.

4c(ii). The key identifying assumption is the parallel assumption, i.e. that absent the EITC expansion policy change (OBRA93), both families with one child and families with two or more children follow parallel trends over time in their infant health outcomes.

4c(iii). One possible violation could be that there is a macro shock occurs sometime before the EITC expansion, and the shock affects the two family types differently. For example, maybe there's another policy that seeks to discourage families from having more children and thus gives HHs with only 1 kid bigger subsidies. In that case, families with 1 kid and those with 2 or more would be on different trends.

Question 4d

The **"reduced-form" equation** is:

$$Y_{pjst} = \alpha + \delta After_t \times Parity2plus_p + \beta X_{st} + \gamma_p + \eta_s + \delta_t + \phi_j + \varepsilon_{pjst}$$

where Y_{pjst} is a measure of infant health (specifically, the fraction of low birth weight infants multiplied by 100) for the cell defined by parity p, demographic group j, in state s for effective tax year t. γ_p is a set of dummy variables for birth order, η_s is a set of dummy variables for state of residence, δ_t is a set of dummy variables for effective tax year. We also include fixed effects for demographic group ϕ_j. X_{st} includes controls for unemployment rate, welfare reform and Medicaid or SCHIP eligibility. a is the intercept, representing the baseline infant health (i.e. for families with a first-order birth before the policy expansion). ε_{pjst} represents the unobserved variation.

$After$ is a dummy variable equaling one for effective tax years 1994 through 1998, $Parity2plus_p$ is a dummy variable indicating if a birth is second or higher order. Their interaction lets us make use of DD strategy in trying to suss out the difference in infant health outcomes before and after the EITC policy change, while

also factoring in the difference between the treated and control groups (families with 2nd-order or higher births and families with 1st-order birth, respectively)

δ is our coefficient of interest, i.e. the DD estimate. It shows the effect of the treatment (i.e. policy expansion) on the treated's (in this case, families whose birth is 2nd-order or higher) infant birth weight.

Question 4e

Alternative specification to equation:

$$Y_{pjst} = \alpha + \delta After_t \times Parity2plus_p + \varphi_1 After_t + \varphi_2 Parity2plus_p + \beta \tilde{X}_{pjst} + \varepsilon_{pjst}$$

In this case, φ_1 represents the difference in the outcome before & after the EITC policy change, while φ_2 represents the difference between the treated and control groups.

Question 5: Data exercise

Just so you know, the dummy syntax for the `felm` command is as follows:

```
felm(causal relation of interest | fixed effects | IVs | clusters, data = your_data)
```

Question 5a & 5b

```
q4 <- read_csv("./dataexercise_pset4/pset_experiment_data.csv")

# 5b
q4 <- q4 %>%
  drop_na(treatment) %>%
  mutate(
    treatment = if_else(treatment == "Selected", 1, 0),
    returned_12m = if_else(returned_12m == "Yes", 1, 0))

# calculate avg survey response rate
q4 %>%
  drop_na(returned_12m) %>% # CAN I DROP NA
  group_by(treatment) %>%
  summarize(response_rate = mean(returned_12m))
```

```
## # A tibble: 2 x 2
##   treatment response_rate
##       <dbl>         <dbl>
## 1         0         0.415
## 2         1         0.399
```

```
# t-test to see if difference in rate is sig
q4_treatonly <- q4 %>%
  filter(treatment == 1)
q4_controlonly <- q4 %>%
  filter(treatment == 0)

t.test(q4_treatonly$returned_12m, q4_controlonly$returned_12m)
```

```
##
##  Welch Two Sample t-test
##
## data:  q4_treatonly$returned_12m and q4_controlonly$returned_12m
## t = -3.9564, df = 58341, p-value = 7.618e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.024056231 -0.008117376
## sample estimates:
## mean of x mean of y
## 0.3991686 0.4152554
```

(5b) The response rates are pretty similar across the control and treatment groups (41.5% and 39.9%, respectively). Given this closeness, I'd say the difference in survey response rates isn't concerning.

This is supported by the t-test. We see that although we reject the null (i.e. there is a significant difference between the means), the 95% confidence interval contains values quite close to 0. Thus, though there is evidence of differential response rates, the response rates are close enough (much smaller than in the RAND experiment) that it doesn't seem like an existential threat to our experiment. Discussion of Table II in the paper on pg. 1075-76 lends further credence to the idea that this small difference in response rates doesn't disqualify the experiment's results.

Question 5c

```
q4 <- q4 %>%
  mutate(
    ohp_all_mo_survey = parse_number(ohp_all_mo_survey),
    ohp_all_ever_survey = if_else(ohp_all_ever_survey == "Enrolled", 1, 0))

dummies_q4 <- q4 %>%
  select(starts_with ("ddd")) %>%
  colnames()

# drop NAs for weights to
q4_test <- q4 %>%
  drop_na(weight_12m) %>%
  filter(sample_12m_resp == "12m mail survey responder")

q4_1ststage_ols_ever <- felm(as.formula(
  paste("ohp_all_ever_survey ~  treatment", "+",
        paste(dummies_q4, collapse = " + "),
        "| 0 | 0 | household_id", sep = "")),
  data = q4_test, weights = q4_test$weight_12m)

q4_1ststage_ols_mo <- felm(as.formula(
  paste("ohp_all_mo_survey ~  treatment", "+",
        paste(dummies_q4, collapse = " + "),
        "| 0 | 0 | household_id", sep = "")),
  data = q4_test, weights = q4_test$weight_12m)

q4_1ststage_ols_end <- felm(as.formula(
  paste("ohp_all_end_survey ~  treatment", "+",
```

```
        paste(dummies_q4, collapse = " + "),
        "| 0 | 0 | household_id", sep = "")),
  data = q4_test, weights = q4_test$weight_12m)

#q4_1ststage_ols_end %>%
 # summary("robust")
```

Question 5c(ii): 1st stage OLS results

```
stargazer(q4_1ststage_ols_ever, q4_1ststage_ols_mo, q4_1ststage_ols_end,
          type = "latex",
          keep = "treatment",
          title = "Question 5c(ii)",
          dep.var.labels =c("Ever on Medicaid", "Number of months on Medicaid",
                            "On Medicaid at end"),
          omit.stat =c("f","rsq","adj.rsq","ser"),
          dep.var.caption = "OLS 1st stage",
          digits = 4)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Thu, May 27, 2021 - 12:15:04 AM

Table 1: Question 5c(ii)

	OLS 1st stage		
	Ever on Medicaid	Number of months on Medicaid	On Medicaid at end
	(1)	(2)	(3)
treatment	0.2902***	3.9427***	0.1890***
	(0.0066)	(0.0896)	(0.0061)
Observations	23,741	23,741	23,741
Note:			*p<0.1; **p<0.05; ***p<0.01

Question 5d(i)

```
#healthnotpoor has 1, 0, & NA. notbaddays is dbl w/ NAs - don't need if_else for them
# note: if_else command preserves NAs

q4 <- q4 %>%
  mutate(
    rx_any_12m = if_else(rx_any_12m == "Yes", 1, 0),
    doc_any_12m = if_else(doc_any_12m == "Yes", 1, 0),
    er_any_12m = if_else(er_any_12m == "Yes", 1, 0),
    hosp_any_12m = if_else(hosp_any_12m == "Yes", 1, 0),
    cost_any_oop_12m = if_else(cost_any_oop_12m == "Yes", 1, 0),
    cost_any_owe_12m = if_else(cost_any_owe_12m == "Yes", 1, 0),
```

```
    )

# drop NAs for weights too & select only survey data
# (have to rerun same command as before since we changed q4 above)
q4_test_1 <- q4 %>%
  #drop_na(weight_12m) %>%
  filter(sample_12m_resp == "12m mail survey responder")

q4_struct_ols_a <- felm(as.formula(
  paste("rx_any_12m ~ ",
        paste(dummies_q4, collapse = " + "),
        "| 0 | (ohp_all_ever_survey ~ treatment) | household_id", sep = "")),
  data = q4_test_1, weights = q4_test_1$weight_12m)

q4_struct_ols_a <- felm(as.formula(
  paste("rx_any_12m ~ ",
        paste(dummies_q4, collapse = " + "),
        "| 0 | (ohp_all_ever_survey ~ treatment) | household_id", sep = "")),
  data = q4_test_1, weights = q4_test_1$weight_12m)

q4_struct_ols_b <- felm(as.formula(
  paste("doc_any_12m ~ ",
        paste(dummies_q4, collapse = " + "),
        "| 0 | (ohp_all_ever_survey ~ treatment) | household_id", sep = "")),
  data = q4_test_1, weights = q4_test_1$weight_12m)

q4_struct_ols_c <- felm(as.formula(
  paste("er_any_12m ~ ",
        paste(dummies_q4, collapse = " + "),
        "| 0 | (ohp_all_ever_survey ~ treatment) | household_id", sep = "")),
  data = q4_test_1, weights = q4_test_1$weight_12m)

q4_struct_ols_d <- felm(as.formula(
  paste("hosp_any_12m ~ ",
        paste(dummies_q4, collapse = " + "),
        "| 0 | (ohp_all_ever_survey ~ treatment) | household_id", sep = "")),
  data = q4_test_1, weights = q4_test_1$weight_12m)

q4_struct_ols_e <- felm(as.formula(
  paste("cost_any_oop_12m ~ ",
        paste(dummies_q4, collapse = " + "),
        "| 0 | (ohp_all_ever_survey ~ treatment) | household_id", sep = "")),
  data = q4_test_1, weights = q4_test_1$weight_12m)

q4_struct_ols_f <- felm(as.formula(
  paste("cost_any_owe_12m ~ ",
        paste(dummies_q4, collapse = " + "),
        "| 0 | (ohp_all_ever_survey ~ treatment) | household_id", sep = "")),
  data = q4_test_1, weights = q4_test_1$weight_12m)

q4_struct_ols_g <- felm(as.formula(
  paste("health_notpoor_12m ~ ",
        paste(dummies_q4, collapse = " + "),
```

```
        "| 0 | (ohp_all_ever_survey ~ treatment) | household_id", sep = "")),
  data = q4_test_1, weights = q4_test_1$weight_12m)

q4_struct_ols_h <- felm(as.formula(
  paste("notbaddays_tot_12m ~ ",
        paste(dummies_q4, collapse = " + "),
        "| 0 | (ohp_all_ever_survey ~ treatment) | household_id", sep = "")),
  data = q4_test_1, weights = q4_test_1$weight_12m)

#q4_struct_ols_f %>%
 # summary("robust")
```

Question 5d(ii)

```
q4_itt_ols_a <- felm(as.formula(
  paste("rx_any_12m ~ treatment + ",
        paste(dummies_q4, collapse = " + "),
        "| 0 | 0 | household_id", sep = "")),
  data = q4_test_1, weights = q4_test_1$weight_12m)

q4_itt_ols_b <- felm(as.formula(
  paste("doc_any_12m ~ treatment + ",
        paste(dummies_q4, collapse = " + "),
        "| 0 | 0 | household_id", sep = "")),
  data = q4_test_1, weights = q4_test_1$weight_12m)

q4_itt_ols_c <- felm(as.formula(
  paste("er_any_12m ~ treatment + ",
        paste(dummies_q4, collapse = " + "),
        "| 0 | 0 | household_id", sep = "")),
  data = q4_test_1, weights = q4_test_1$weight_12m)

q4_itt_ols_d <- felm(as.formula(
  paste("hosp_any_12m ~ treatment + ",
        paste(dummies_q4, collapse = " + "),
        "| 0 | 0 | household_id", sep = "")),
  data = q4_test_1, weights = q4_test_1$weight_12m)

q4_itt_ols_e <- felm(as.formula(
  paste("cost_any_oop_12m ~ treatment + ",
        paste(dummies_q4, collapse = " + "),
        "| 0 | 0 | household_id", sep = "")),
  data = q4_test_1, weights = q4_test_1$weight_12m)

q4_itt_ols_f <- felm(as.formula(
  paste("cost_any_owe_12m ~ treatment + ",
        paste(dummies_q4, collapse = " + "),
        "| 0 | 0 | household_id", sep = "")),
  data = q4_test_1, weights = q4_test_1$weight_12m)

q4_itt_ols_g <- felm(as.formula(
```

```
  paste("health_notpoor_12m ~ treatment + ",
        paste(dummies_q4, collapse = " + "),
        "| 0 | 0 | household_id", sep = "")),
  data = q4_test_1, weights = q4_test_1$weight_12m)

q4_itt_ols_h <- felm(as.formula(
  paste("notbaddays_tot_12m ~ treatment + ",
        paste(dummies_q4, collapse = " + "),
        "| 0 | 0 | household_id", sep = "")),
  data = q4_test_1, weights = q4_test_1$weight_12m)

#q4_itt_ols_a %>%
 # summary("robust")
```

For Question 5d(iii), I created 2 separate tables (one of $\hat{\pi}$ estimates for the structural equation, one of $\hat{\beta}$ estimates for the ITT equation) since combining all 16 regressions into 1 table produces something illegible.

Question 5d(iii): $\hat{\pi}$ Coefficients

```
stargazer(q4_struct_ols_a, q4_struct_ols_b, q4_struct_ols_c, q4_struct_ols_d,
          q4_struct_ols_e, q4_struct_ols_f, q4_struct_ols_g, q4_struct_ols_h,
          type = "latex",
          omit = c(dummies_q4, "Constant"),
          title = "Question 5d(iii): Pi Estimates (Structural Eqn)",
          covariate.labels = "Ever on Medicaid",
          dep.var.labels = c("rx", "doc", "er", "hosp", "cost oop", "cost owe",
                             "health", "notbad"),
          omit.stat =c("f","rsq","adj.rsq","ser"),
          dep.var.caption = "",
          digits = 4,
         font.size = "small",
         column.sep.width = "0pt")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Thu, May 27, 2021 - 12:15:10 AM

Table 2: Question 5d(iii): Pi Estimates (Structural Eqn)

	rx	doc	er	hosp	cost oop	cost owe	health	notbad
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Ever on Medicaid	0.0878^{***}	0.2124^{***}	0.0223	0.0077	-0.1995^{***}	-0.1798^{***}	0.0990^{***}	1.3171^{**}
	(0.0288)	(0.0252)	(0.0231)	(0.0136)	(0.0262)	(0.0265)	(0.0176)	(0.5629)
Observations	18,308	23,492	23,514	23,573	23,426	23,451	23,361	21,881

Note: *p<0.1; **p<0.05; ***p<0.01

Question 5d(iii): $\hat{\beta}$ coefficients

```
stargazer(q4_itt_ols_a, q4_itt_ols_b, q4_itt_ols_c, q4_itt_ols_d, q4_itt_ols_e,
          q4_itt_ols_f, q4_itt_ols_g, q4_itt_ols_h,
          type = "latex",
          omit = c(dummies_q4, "Constant"),
          title = "Question 5d(iii): Beta Estimates (ITT Equation)",
          covariate.labels = "Ever on Medicaid",
          dep.var.labels = c("rx", "doc", "er", "hosp", "cost oop", "cost owe",
                             "health", "notbad"),
          omit.stat =c("f","rsq","adj.rsq","ser"),
          dep.var.caption = "",
          digits = 4,
         font.size = "small",
         column.sep.width = "0pt")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Thu, May 27, 2021 - 12:15:10 AM

Table 3: Question 5d(iii): Beta Estimates (ITT Equation)

	rx	doc	er	hosp	cost oop	cost owe	health	notbad
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Ever on Medicaid	0.0252***	0.0617***	0.0065	0.0022	−0.0580***	−0.0523***	0.0288***	0.3810**
	(0.0083)	(0.0074)	(0.0067)	(0.0040)	(0.0077)	(0.0076)	(0.0051)	(0.1618)
Observations	18,308	23,492	23,514	23,573	23,426	23,451	23,361	21,881

Note: *p<0.1; **p<0.05; ***p<0.01