Inequality Pset 3

Julia Du

2021-05-12

Load necessary libraries

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

theme_set(theme_minimal())
```

Question 4a

The causal relationship of interest is the effect of income (where your income is either increased by the changed EITC) on infant health outcomes, specifically infant birth weight.

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 (maybe just write this out)

4c(i). on paper

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 5d

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 5e

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 4f-4g(i): Model 1

Note: cannot upload csv to Github as file too big

```
q3 <- read_csv("./dataexercise_pset3/pset_dd_data.csv")

# 4g(i).A & .B: restrict sample, create dummies
q3 <- q3 %>%
  filter(dmeducgrp %in% c("drop out", "high school"),
```

```
        marital == "not married") %>%
  filter(between(effective, 1991, 1998)) %>%
  mutate(y_lowbirth = 100 * lowbirth) %>%
  mutate(after = if_else(between(effective, 1994, 1998), 1, 0)) %>%
  mutate(
    treat1 = if_else(parvar == "1st live birth", 0, 1),
    treat2 = if_else(parvar %in% c("2nd live birth"), 1, 0),
    treat3 = if_else(parvar %in% c("3rd  birth", "4th  birth"), 1, 0)) %>%
  mutate(treat1_after = after*treat1)
# just fyi: there are 2 spaces in between the words for "3rd birth" & "4th birth"

# process data for controls
q3 <- q3 %>%
  mutate(stateres = as_factor(stateres))

fe_varlist <- q3 %>%
  select(starts_with("fe_I")) %>%
  colnames()

independ_var <- c("other", "black", "age2", "age3", "high", "racemiss",
                  "hispanic", "hispanicmiss", "reform", "a_urate_st",
                  "threshpreg", "as_factor(effective)", "as_factor(parvar)")

# 4g(i).C: estimate model(1) by OLS
q3_ols1 <- felm(as.formula(
  paste("y_lowbirth ~ treat1_after +",
        paste(independ_var, collapse = " + "), "+",
        paste(fe_varlist, collapse = " + "),
        "| stateres | 0 | stateres", sep = "")),
  data = q3, weights = q3$cellnum)

# commenting out the summary of reg results as we have a stargazer table comin up
#q3_ols1 %>%
 # summary("robust")
```

Question 4g(ii): Model 1'

```
q3 <- q3 %>%
  mutate(treat2_after = after*treat2,
         treat3_after = after*treat3)

q3_ols2 <- felm(as.formula(
  paste("y_lowbirth ~ treat2_after + treat3_after + ",
        paste(independ_var, collapse = " + "), "+",
        paste(fe_varlist, collapse = " + "),
        "| stateres | 0 | stateres", sep = "")),
  data = q3, weights = q3$cellnum)

#q3_ols2 %>%
 # summary("robust")
```

4g(ii).B. It makes sense that the EITC expansion affects families with more kids (3 or more) on a bigger

scale. Families with more kids (i.e. 3rd or higher order births) likely face more costs in raising their kids than a family with just 2 kids (i.e. 2nd-order birth); accordingly, those 3 or more kids families benefit more from expanding the EITC.

Question 4g(iii): Model 1''

```
q3_data3 <- q3 %>%
  filter(parvar != "1st live birth")

q3_ols3 <- felm(as.formula(
  paste("y_lowbirth ~ treat3_after + ",
        paste(independ_var, collapse = " + "), "+",
        paste(fe_varlist, collapse = " + "),
        "| stateres | 0 | stateres", sep = "")),
  data = q3_data3, weights = q3_data3$cellnum)

#q3_ols3 %>%
 # summary("robust")
```

4g(iii).B* The treatment group is families with 3rd or higher order births, and the control group is families with 2nd order births.

Question 4g(iv): Summary table

```
stargazer(q3_ols1, q3_ols2, q3_ols3, type = "latex",
          keep = c("treat1_after", "treat2_after", "treat3_after"),
          title = "Question 4g(iv)",
          dep.var.labels =c("Model 1", "Model 1'", "Model 1''"),
          omit.stat =c("f","rsq","adj.rsq","ser"),
          covariate.labels =c("Parity2+ x After", "Parity=2 x After",
                              "Parity3+ x After"),
          dep.var.caption = "OLS",
          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: Wed, May 12, 2021 - 9:03:29 PM

Question 4h

```
q3_yr <- q3 %>%
  mutate(eventyr = effective - 1993) %>%
  #mutate(eventyr0 = if_else(eventyr == -2, 1, 0)) %>% # can also do this method
  mutate(eventyr0 = as.numeric(eventyr == -2),
         eventyr1 = as.numeric(eventyr == -1),
         eventyr2 = as.numeric(eventyr == 0),
         eventyr3 = as.numeric(eventyr == 1),
         eventyr4 = as.numeric(eventyr == 2),
```

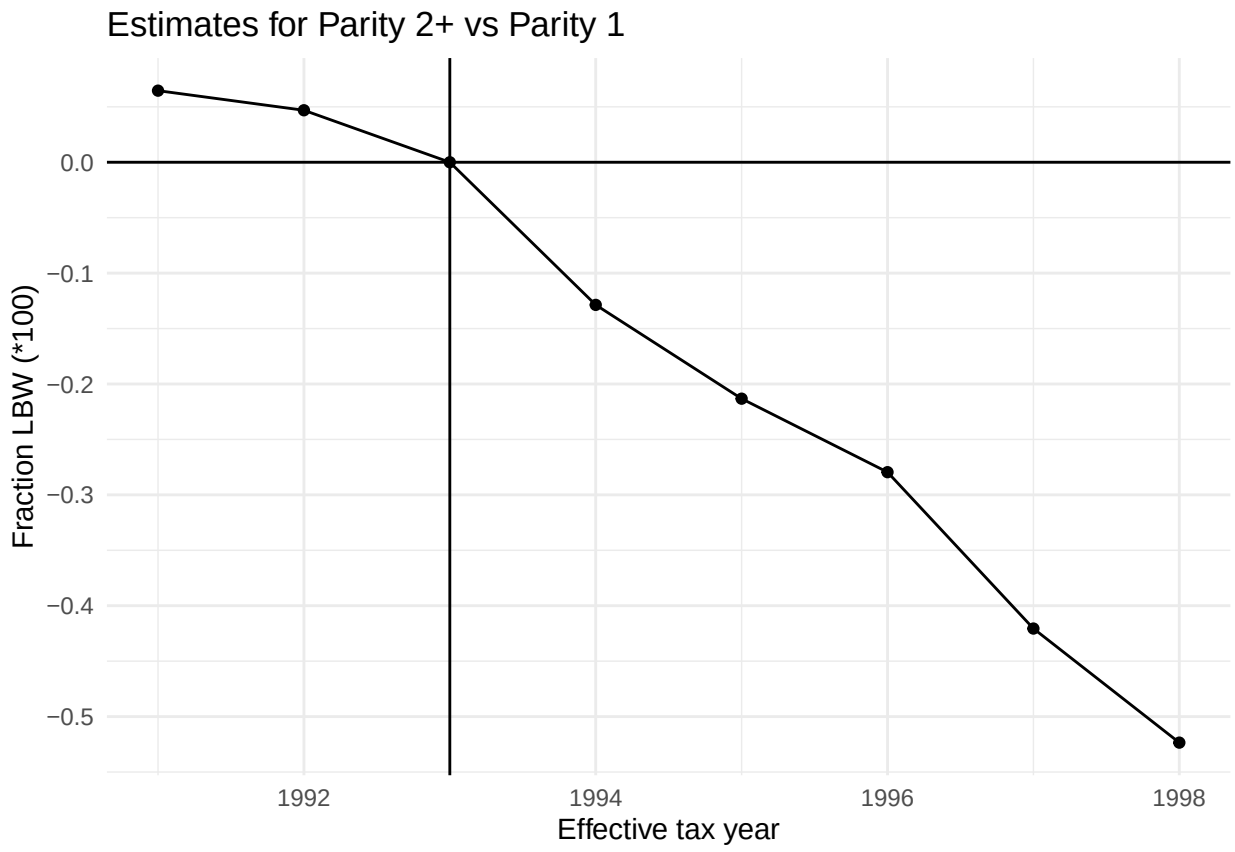
Table 1: Question 4g(iv)

	OLS		
	Model 1	Model 1'	Model 1"
	(1)	(2)	(3)
Parity2+ x After	−0.3538***		
	(0.0742)		
Parity=2 x After		−0.1637**	
		(0.0719)	
Parity3+ x After		−0.5277***	−0.3404***
		(0.0902)	(0.0684)
Observations	47,687	47,687	35,467
Note:			*p<0.1; **p<0.05; ***p<0.01

```
         eventyr5 = as.numeric(eventyr == 3),
         eventyr6 = as.numeric(eventyr == 4),
         eventyr7 = as.numeric(eventyr == 5)) %>%
  #only need dummies for treat 1 since estimating model 1
  mutate(treat1_eventyr0 = treat1*eventyr0,
         treat1_eventyr1 = treat1*eventyr1,
         treat1_eventyr2 = treat1*eventyr2,
         treat1_eventyr3 = treat1*eventyr3,
         treat1_eventyr4 = treat1*eventyr4,
         treat1_eventyr5 = treat1*eventyr5,
         treat1_eventyr6 = treat1*eventyr6,
         treat1_eventyr7 = treat1*eventyr7)

evt1_var <- names(q3_yr)[grep("treat1_e", names(q3_yr))]
  #q3_yr %>%
  #select(starts_with("treat1_e")) %>%
  #colnames()
#unsure why i'm doing this
evt1_var <- setdiff(evt1_var, "treat1_eventyr2")

evt_control <- c("other", "black", "age2", "age3", "high", "racemiss",
                 "hispanic", "hispanicmiss", "reform", "a_urate_st",
                 "threshpreg", "as_factor(eventyr)", "as_factor(parvar)")

q3_yr_reg <- felm(as.formula(
  paste("y_lowbirth ~ ",
        paste(evt1_var, collapse = " + "), "+",
        paste(evt_control, collapse = " + "), "+",
        paste(fe_varlist, collapse = " + "),
        "| stateres | 0 | stateres", sep = "")),
  data = q3_yr, weights = q3_yr$cellnum)
```

Plot Figure 3a

```
df.plot <- as.data.frame(matrix(ncol = 2, nrow = 8))
names(df.plot) <- c("eyear", "lowbirth2_a")
df.plot$eyear <- 1991:1998
df.plot$lowbirth2_a[1:2] <- summary(q3_yr_reg)$coefficients[1:2,1]
df.plot$lowbirth2_a[3] <- 0
df.plot$lowbirth2_a[4:8] <- summary(q3_yr_reg)$coefficients[3:7,1]

df.plot %>%
  ggplot()+
  geom_line(mapping = aes(x = eyear, y = lowbirth2_a)) +
  geom_point(mapping = aes(x = eyear, y = lowbirth2_a)) +
  geom_vline(xintercept = 1993) +
  geom_hline(yintercept = 0) +
  labs(x = "Effective tax year", y = "Fraction LBW (*100)",
       title = "Estimates for Parity 2+ vs Parity 1")
```



4h(ii). The plot shows that there no pretrend (i.e. quite flat), indicating that there is no violation of the parallel trends assumption for model 1.