Inequality PSet 2 - Question #5 (data exercise)

Julia Du

2021-04-27

Load necessary libraries

```
library(tidyverse)
library(lubridate)
library(broom)
library(stats)
library(ivreg)
library(lmtest)
library(sandwich)
library(moderndive)
library(kableExtra)
library(tinytex)
library(stargazer)

theme_set(theme_minimal())
```

Question 5a

The causal relationship of interest is the returns to additional medical spending, i.e. the effect of more medical expenditures on health outcomes. In this case, that's the effect of additional infant health care on one-year mortality rate for infants.

The structural equation can be seen as:

$$infant\ mortality = \beta_0 + \beta_1(health\ inputs) + U_i$$

Question 5b

Cross-sectional data looks at a defined population at a specific point in time. This could lead to an underestimate of the effect of health inputs on health outcomes.

For example, a cross-section could miss how medical technologies improve over time, which could lead to an increase in return on medical spending (e.g. as tech improves with time, spending another dollar on health care likely has a bigger positive effect on patient health outcomes). If we didn't focus only on one specific point in time (as we do with cross-sectional data), we might find that spending another dollar would reduce infant mortality by an even greater amount than initially calculated.

Question 5c

The instrument here is the Very Low Birthweight (VLBW) indicator, which is a dummy that indicates if the newborn was classified as weighing strictly less than 1500 grams. VLBW allows us to get an unbiased estimate of the causal relationship because it is correlated with the variable of interest (health inputs/spending - if a baby is classified underweight, there should be more money spent on their health) and because it should only affect the dependent variable (mortality outcome) through VLBW's effect on health inputs/spending. That is, it satisfies the relevance assumption & exclusion restriction.

More specifically, it isn't really possible to predict or control infants' birth weights, so birth weight should be as good as random assignment to the treatment of additional health care (and thus increased health spending) - and so infants on either side of the 1500g cutoff should see the same distribution of characteristics. Then, any difference seen in mortality outcomes around the cutoff can be chalked up to the effect of additional health spending on mortality.

Question 5d

The **"first-stage" equation** is:

$$Y_i = b_0 + b_1 VLBW_i + b_2 VLBW_i \times (g_i - 1500) + b_3(1 - VLBW_I) \times (g_i - 1500) + b_t + b_s + \delta X_i' + e_i$$

where Y_i is the health inputs (aka costs).

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

$$Y_i = a_0 + a_1 VLBW_i + a_2 VLBW_i \times (g_i - 1500) + a_3(1 - VLBW_I) \times (g_i - 1500) + a_t + a_s + \delta X_i' + \epsilon_i$$

where Y_i is one-year mortality.

As you can see, there's a great deal of similarity between these two equations. The variables are:

- $VLBW_i$ is the indicator instrumental variable on whether the newborn was classified as VLBW.
- $VLBW_i \times (g_i - 1500)$ is how far below the cutoff of 1500 g the infant is (if they are indeed below the cutoff).
- $(1 - VLBW_I) \times (g_i - 1500)$ is how far above the cutoff the infant is (if they are indeed above the cutoff).
- g_i is infant i's weight in grams
- b_t and a_t are indicators for each year of birth t (essentially controls)
- b_s and a_s are indicators for each state of birth s (essentially controls)
- X_i' are newborn characteristics (essentially control variables)

So similarly, the coefficients are (I will be writing just the a coefficients here, but they describe the same effect on mortality as the b coefficients have on health inputs):

- a_1 and b_1 are the effect of VLBW on outcome Y_i (i.e. mortality outcome for a_1 & health input for b_1)
- a_2 is the effect of VLBW on Y_i if the infant is below the cutoff of 1500 grams
- a_3 is the effect of VLBW on Y_i if the infant is above the cutoff of 1500 grams. If the trends for infants on either side of the cutoff is the same, $a_2 = a_3$
- δ is the effect of the infant characteristics X_i' on outcome Y_i

Question 5e

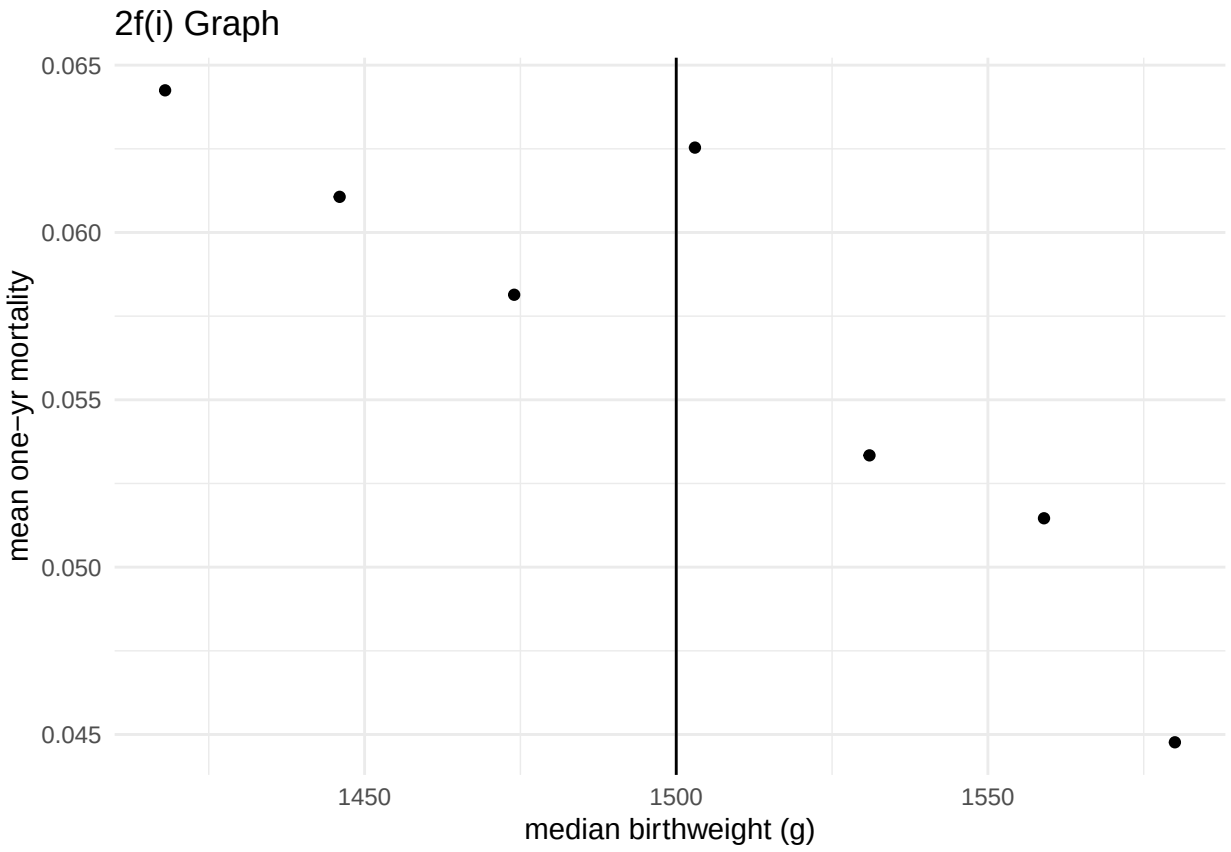
In using fuzzy RD, we assume that the first-stage/relevance assumption and the exclusion restriction assumption is met.

To satisfy the exclusion restriction, we need to first check if the assignment to treatment is as good as random. In other words, we need to make sure that infants' classified weights are indeed being recorded accurately so that their assignment to the treatment of increased health inputs/spending is as good as random. Doctors may sometimes manipulate the running variable of birth weight, marking babies as underweight because their families can then get more government support. To check this, we should perform the McCrary test to see if the density is different above and below the 1500g cutoff. We can also check this by seeing if covariates like gestation age are continuous with regard to birth weight. If treatment assignment is truly good as random, then the baseline characteristics should have the same distribution below and above the cutoff.

To satisfy the exclusion restriction, we also need to check if there's continuity in potential outcomes with regards to birth weight. If there is no treatment, then the outcome (e.g. mortality) shouldn't be discontinuously changed around the cutoff. If this condition isn't met, that indicates there's another factor influencing the mortality outcome here, invalidating our current RD design.

Question 5f(i)

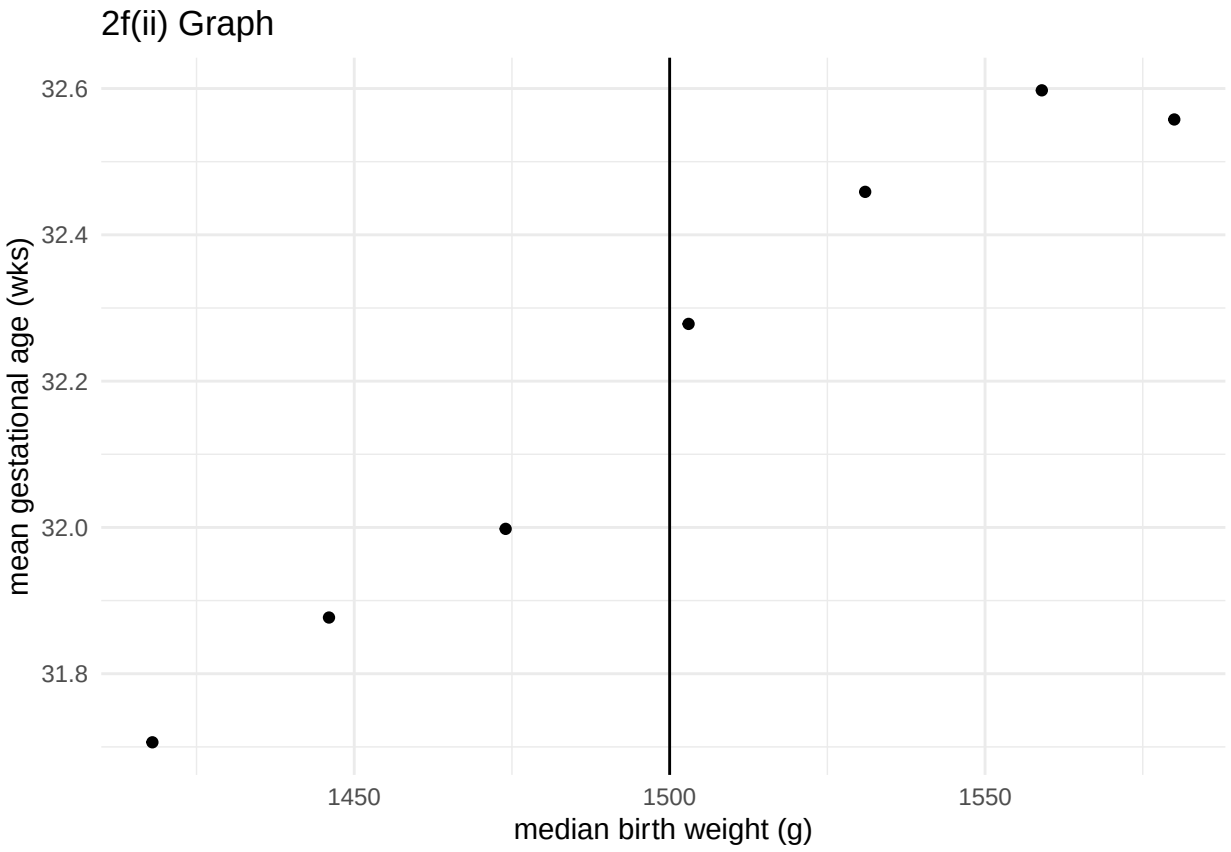
```
#import the dataset
q2 <- read_csv("./DataExercise_RD/adkw.csv")

#graph Figure 2A
q2 %>%
  mutate(bin = cut_width(x = dbirwt, width = 28.3495, center = 1500)) %>%
  drop_na(death1year) %>%
  group_by(bin) %>%
  summarize(med_bin = median(dbirwt), mean_death1year = mean(death1year)) %>%
  ggplot()+
  geom_point(mapping = aes(x = med_bin, y = mean_death1year)) +
  geom_vline(xintercept = 1500) +
  labs(x = "median birthweight (g)", y = "mean one-yr mortality", title = "2f(i) Graph")
```



As with the paper, my graph shows that there is a general reduction in mortality as birth weight increases, reflecting the health benefits associated with higher birth weight. There is a noticeable increase in mortality in the ounce bin just above the 1500g cutoff (about 6.25%), compared to the ounce bin right below the cutoff (about 5.75%) - so it seems that treatment (i.e. increased medical spending for infants just barely classified as underweight) does positively affect health outcomes (i.e. it decreases mortality for those infants).

Question 5f(ii)

```
q2 %>%
  mutate(bin = cut_width(x = dbirwt, width = 28.3495, center = 1500)) %>%
  drop_na(death1year, gestat) %>%
  group_by(bin) %>%
  summarize(med_bin = median(dbirwt), mean_death1year = mean(death1year),
            mean_gestat = mean(gestat))%>%
  ggplot()+
  geom_point(mapping = aes(x = med_bin, y = mean_gestat)) +
  geom_vline(xintercept = 1500) +
  labs(x = "median birth weight (g)", y = "mean gestational age (wks)",
       title = "2f(ii) Graph")
```



This figure supports the RD identifying assumption that covariates of interest (in this case, gestational age) are continuous with regard to the running variable (mortality). This helps us conclude that the treatment assignment is as good as random since the baseline characteristics for infants below and above the cutoff of 1500 grams have similar distributions.

Question 5f(iii)

```
q2_reg_data <- q2 %>%
  mutate(bin = cut_width(x = dbirwt, width = 28.3495, center = 1500)) %>%
  group_by(bin) %>%
  mutate(VLBW = if_else(dbirwt < 1500, 1, 0)) %>%
  mutate(VLBW2 = VLBW*(dbirwt-1500)/100,
         VLBW3 = (1-VLBW)*(dbirwt-1500)/100)
# divide by 100 since paper has values in (100s) for those variables

# estimate reduced form eqn by OLS
q2_ols <- lm(death1year ~ VLBW + VLBW2 + VLBW3, data = q2_reg_data)

# heteroskedastic error
coeftest(q2_ols, vcov = vcovHC(q2_ols, type = "HC1")) %>%
  tidy() %>%
  select(term, estimate, std.error)
```

```
## # A tibble: 4 x 3
```

```
##   term        estimate std.error
##   <chr>          <dbl>     <dbl>
## 1 (Intercept)  0.0631    0.00127
## 2 VLBW        -0.00954   0.00217
## 3 VLBW2       -0.0136    0.00324
## 4 VLBW3       -0.0224    0.00288
```

```
# helpful link on printing stargazer table:
# https://stackoverflow.com/questions/45724432/stargazer-output-is-code-not-a-table
```

```
stargazer(q2_ols, type = "latex", title = "5f(iii).C", 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: Tue, Apr 27, 2021 - 1:16:58 AM

Table 1: 5f(iii).C

	Dependent variable:
	death1year
VLBW	-0.0095^{***}
	(0.0022)
VLBW2	-0.0136^{***}
	(0.0032)
VLBW3	-0.0224^{***}
	(0.0029)
Constant	0.0631^{***}
	(0.0012)
Observations	202,071
R^2	0.0005
Adjusted R^2	0.0005
Residual Std. Error	0.2331 (df = 202067)
F Statistic	33.7525^{***} (df = 3; 202067)
Note:	*p<0.1; **p<0.05; ***p<0.01