TA Session #7

Ken Chen & Lindsay Liebert

May 18, 2019

1. Panel Data Method

1.1 Fixed Effect Model

Fixed effects are often unobservable, but excluding them from our estimation might introduce endogeneity problems. For a basic panel data fixed effect regression, we have a generic model specification:

$$Y_{it} = X_{it}\beta + \tau D_{it} + \epsilon_{it}\epsilon_{it} = \alpha_i + \delta_t + v_{it}$$

The error term can be decomposed into three components: α_i represents the *individual-specific and time-invariant* fixed effect; δ_t represents the *time-specific and individual-invariant* fixed effect. For the sake of easy interpretation, let's restrict our discussion to individual fixed-effect model ($\delta_t = 0$). There are two popular ways to estimate:

- Gennerate a series of dummies to represent the fixed effects

$$Y_{it} = X_{it}\beta + \tau D_{it} + \sum_{j=1}^{N} 1(i = j) + \epsilon_{it}$$

- De-meaning all the variables to remove the fixed effects

$$Y_{it} - \bar{Y_i} = (X_{it} - \bar{X_i})\beta + \tau(D_{it} - \bar{D_i}) + (\alpha_i - \bar{\alpha_i}) + (\epsilon_{it} - \bar{\epsilon_i})$$
$$\tilde{Y_{it}} = \tilde{X_{it}} + \tau\tilde{D_{it}} + \tilde{\epsilon_{it}}$$

Example of running panel data fixed effect model:

Background: The Fatalities is a panel dataset that contains information on traffic deaths and alcohol taxes of various states from year 1982 to 1988. We are interested in how the policy of increasing alcohol taxes will affect the rate of traffic deaths. Let us look at how will our analysis differ when employing the cross-sectional data method vs the panel data method.

```
# import the data
library(AER)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

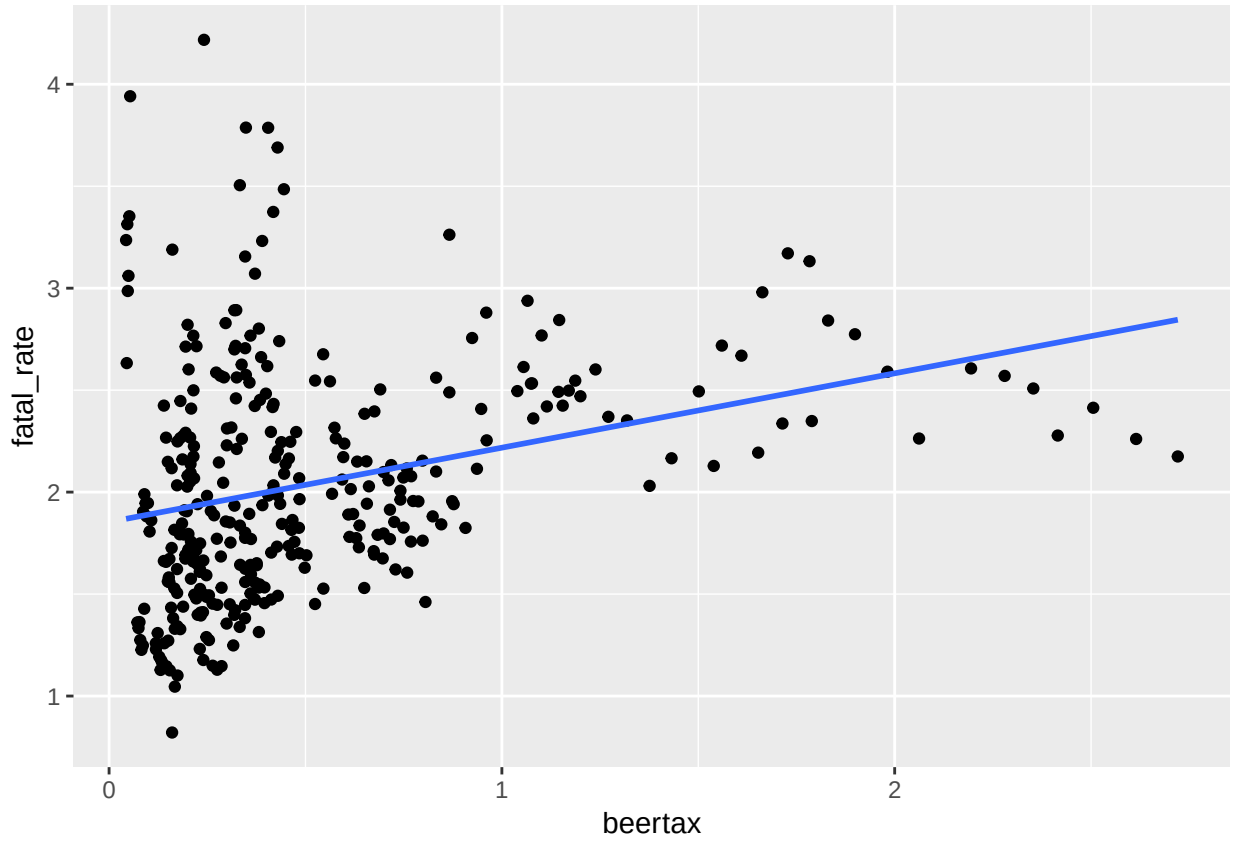
```
## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some

## Loading required package: lmtest

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich

## Loading required package: survival
```

```
data(Fatalities)
Fatalities <- Fatalities %>% mutate(fatal_rate = fatal/pop*10000)
reg <- lm(fatal_rate ~ beertax, data = Fatalities)
stargazer(reg, type = 'text')
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                              fatal_rate
## ------------------------------------------------
## beertax                       0.365***
##                               (0.062)
##
## Constant                      1.853***
##                               (0.044)
##
## ------------------------------------------------
## Observations                    336
## R2                             0.093
## Adjusted R2                    0.091
## Residual Std. Error       0.544 (df = 334)
## F Statistic           34.394*** (df = 1; 334)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

```
ggplot(data = Fatalities) +
  geom_point(mapping = aes(x = beertax, y = fatal_rate)) +
  geom_smooth(mapping = aes(x = beertax, y = fitted(reg)))
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



If the we omit the individual fixed effects, which could be viewed as the baseline traffic condition specific to a certain state, we would arrive at the fatal causal inference and conclude that the more we tax on alcohol, the more traffic fatalities there would be. We may want to reconsider the specification of our model by adding to it the individual fixed effects:

$$FatalRate_{it} = \beta BeerTax_{it} + StateFixedEffect_{it} + \epsilon_{it}$$

```
library(plm)
```

```
##
## Attaching package: 'plm'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, lag, lead
```

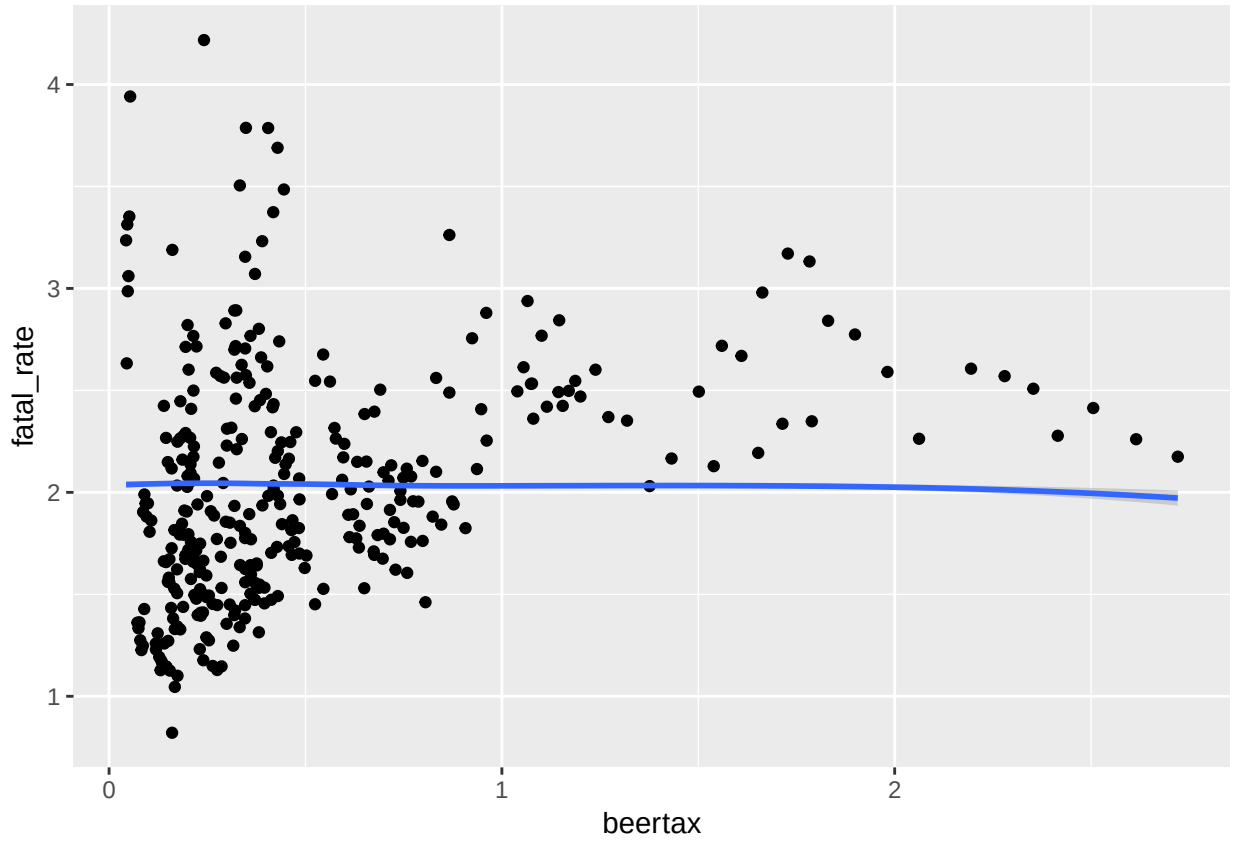
```
fe <- plm(fatal_rate ~ beertax,
          data = Fatalities,
```

```
          model = "within", # setting for de-meaining within individual fixed effect model
          index = c('state', 'year')) # specify the individual and time index
coeftest(fe, vcov=vcovHC(fe, type="HC0", cluster="group")) # Get the clustered std.Error
```

```
##
## t test of coefficients:
##
##         Estimate Std. Error t value Pr(>|t|)
## beertax -0.65587    0.28837 -2.2744  0.02368 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ggplot(data = Fatalities) +
  geom_point(mapping = aes(x = beertax, y = fatal_rate)) +
  geom_smooth(mapping = aes(x = beertax, y = fitted(fe) + mean(Fatalities$fatal_rate)))
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



By running the individual fixed effect model, we are now able to obtain a more resonalble result. The coefficient is *-0.65587*, which is negative and significant at the 99 percent confidence level.

1.2 Cumulative effects

By generalizing a bit the fixed-effect regression model, we are able to capture the treatment effects at different moments. This is made possible by including a series of indicators of time dummies.

$$Y_{it} = \sum_{s=0}^{S} \tau_s D_{i,t-s} + \alpha_i + \delta_t + X_{it}\beta + v_{it}$$

, where s indicates the time when the program is implemented. This models allows us to investigate the lasting effect of a program over different time horizons. The interpretation of the coefficients τ_s should be partial, which means the effect at a given moment holding the effects at other times as constant. Therefore, the cumulative effect should be the summation of all these coefficients.

$$T_q = \sum_{s=0}^{S} \tau_s$$

, where q is the length of post-treatment periods. And more generally, we can even include indicators of pre-treatment periods to test for confounding factors, if any. The estimated coefficients for the pre-treatment time indicators should be **centered around zero** and **suggest no trending**

$$Y_{it} = \sum_{s=-R}^{S} \tau_s D_i 1(t = s) + \alpha_i + \delta_t + X_{it}\beta + v_{it}$$

Example of studying cumulative effects:

Background: we want to examine the effect of "Carrying a concealed weapon" law on the number of violent crimes.

```
data(Guns)
Guns <- Guns %>% na.omit()
Guns$year <- as.numeric(Guns$year)
Guns$law <- ifelse(Guns$law=='yes', 1, 0)
# Find the starting year of the law for each state
start_year <- Guns %>%
  filter(law==1) %>%
  group_by(state) %>%
  summarise(start_year = min(year))
year_dict <- start_year$start_year
names(year_dict) <- start_year$state
Guns$strt_year = 0
for (i in 1:nrow(Guns)) {
  if (Guns[i, 'state'] %in% names(year_dict)) {
    Guns[i, 'strt_year'] = year_dict[as.character(Guns[i, 'state'])]
  }
}

# For the sake of easy interpretation, let us constrain
# our example to states who initiated the law in 1990
# And we kept the time window to be year7 to year 13
Guns_new <- Guns %>%
  filter(state %in% names(year_dict)[year_dict==10] | !(state %in% names(year_dict))) %>%
  filter(year %in% 7:13)
Guns_new$yr_ind <- Guns_new$year-10
Guns_new$law <- ifelse(Guns_new$state %in% names(year_dict)[year_dict==10], 1, 0)
```

Our model for the question is:

$$log(Violent_{it}) = \sum_{s=-3}^{3} \tau_s law_i 1(t = s) + \alpha_i + \delta_t + X_{it}\beta + v_{it}$$

```
# Fixed effect regression design
fe.gun <- plm(log(violent) ~ as.factor(yr_ind)*law + afam + cauc +
                male + income + population + density,
          data = Guns_new,
          model = "within",
          effect = 'twoway', #include both time and individual fixed effects
          index = c('state', 'year'))
coeftest(fe.gun, vcov=vcovHC(fe.gun, type="HC0", cluster="group")) # Get the clustered std.Error
```

```
##
## t test of coefficients:
##
##                              Estimate  Std. Error t value  Pr(>|t|)
## afam                        5.2053e-01  1.3113e-01  3.9695 0.0001203 ***
## cauc                        2.1686e-01  6.1820e-02  3.5079 0.0006267 ***
## male                       -1.8752e-01  9.8883e-02 -1.8964 0.0602023 .
## income                      4.9050e-05  2.4679e-05  1.9875 0.0490308 *
## population                  4.0059e-02  1.7443e-02  2.2965 0.0232968 *
## density                    -1.3954e+00  2.5778e-01 -5.4132  3.01e-07 ***
## as.factor(yr_ind)-2:law     1.0456e-01  6.7918e-02  1.5396 0.1261746
## as.factor(yr_ind)-1:law    -1.7041e-02  9.5114e-02 -0.1792 0.8580924
## as.factor(yr_ind)0:law     -9.6221e-02  4.4653e-02 -2.1549 0.0330742 *
## as.factor(yr_ind)1:law     -5.2475e-02  5.0487e-02 -1.0394 0.3006249
## as.factor(yr_ind)2:law     -8.4068e-02  8.5989e-02 -0.9777 0.3301187
## as.factor(yr_ind)3:law     -5.4674e-02  5.0104e-02 -1.0912 0.2772556
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we can see, the pre-treatment effects were insignificant. The post-treatment effect is significant at only period 0 and 1. The effect of the first periods is about *-0.096*, which implies that implementing the law would bring about 9.6% of violent crimes at the end of the first year of implementation. The cummulative effect for later years is not significant for this question.

2. Regression Discontinuity Design

2.1 When to use RD

We can use RD when we have a continuous or discrete variable that includes a cut off or threshold for determining treatment and control (ex; birthweight for determining additional hospital care, pollution index for waste cleanup requirements). RD allows us to mimick random assignment by looking at unit outcomes just above and just below a cut off point.

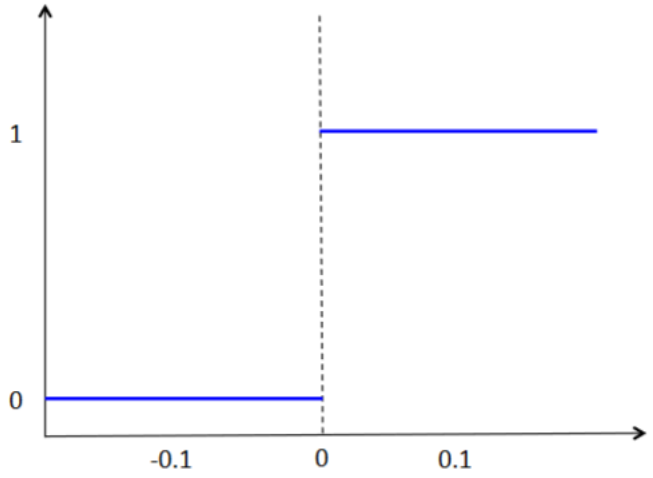
2.2 Types of RD

Sharp RD

Sharp RD is when we have perfect compliance across the cut off point. That is, everyone below the cut off is not treated and everyone above the cut off is treated. In math this is equivalent to:

$$Pr(D_i = 1|X_i \geq c) = 1 \quad and \quad Pr(D_i = 1|X_i < c) = 0$$

In pictures this is equivalent to:

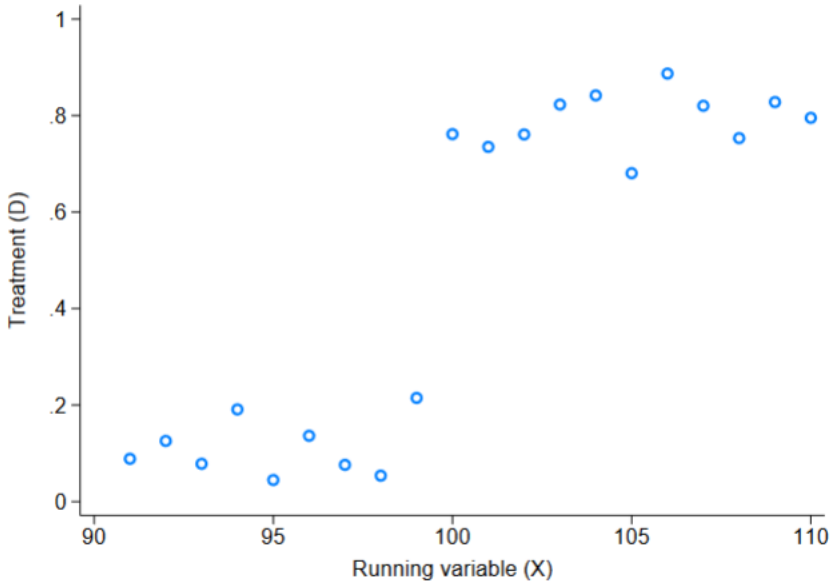


Fuzzy RD

Fuzzy RD is when we have imperfect compliance across the cut off point. That is, not everyone below the cut off is not treated. Some units somehow end up getting treated. Similarly, not everyone above the cut off is treated. Some units somehow end up not getting treated. In math this is equivalent to:

$$Pr(D_i = 1|X_i \geq c) - Pr(D_i = 1|X_i < c) = k \quad where \quad 0 < k < 1$$

In pictures this is equivalent to:



Fuzzy RD will be covered more next week. We will stick to sharp RD for the remainder of these notes.

2.3 Identifying Assumption

For RD we only need one assumption: Continuity in Y across the cut off. Putting this into math, we want to see $E(Y_i(1)|X_i = x)$ and $E(Y_i(0)|X_i = x)$ continuous in x. We cannot observe this since we can never observe the counterfactual outcome. However, there are proxy tests for arguing this assumption holds.

2.4 Graphical Support for RD

When running an RD design, there are 4 graphical depictions of data that are important to include in your analysis:

- **Density of Running Variable** - This helps with the argument that there is no manipulability in the running variable. When we see smooth/continous density across the threshold, this provides evidence that the cutoff is essentially random.
- **Continuity in Covariates** - This serves as a proxy for continuity in potential outcomes. By observing continuity in the observables across the cut off, we can argue that the discontinuity in Y for treated and untreated is solely from difference in treatment status.
- **Outcome across Running Variable** - This is a good first check to see if there is potential for using an RD design. This will give a visual idea of whether there is a discontinuous jump between treatment and control groups.
- **Proportion of Treatment across Running Variable** - This helps determine if our RD design is a fuzzy or sharp RD.

2.5 Regression Model for Sharp RD

Our regression model for estimating the treatment effect within some bandwidth is as follows:

$$Y_i = \alpha + \tau D_i + \beta_1(X_i - c) + \beta_2(X_i - c)D_i + \epsilon_i$$

$\hat{\tau}^{SRD}$ estimates our LATE (local average treatment effect for units at the cutoff) $\hat{\beta}_1$ provides the slope for values below the cutoff $\hat{\beta}_2$ provides the slope for values above the cutoff

2.6 Case Study

This case study comes from Almond et. al (2008) paper on mortality rates for newborns. When newborns are below a certain weight, they receive additional care to increase chance of survival. The RDD uses birthweight as a running variable and examines the effect of additional care on mortality rate of newborns. Here we have below the cutoff being treated and above the cutoff being untreated.

```
#set your working directory and then load the data
library(foreign) # allows R to read .dta files from STATA
setwd("C:/Users/thama/Desktop/Lindsay")
data = read.dta("almond_etal_2008.dta")
var.labels = attr(data, "var.labels")
data.key = data.frame(var.name = names(data), var.labels)
```

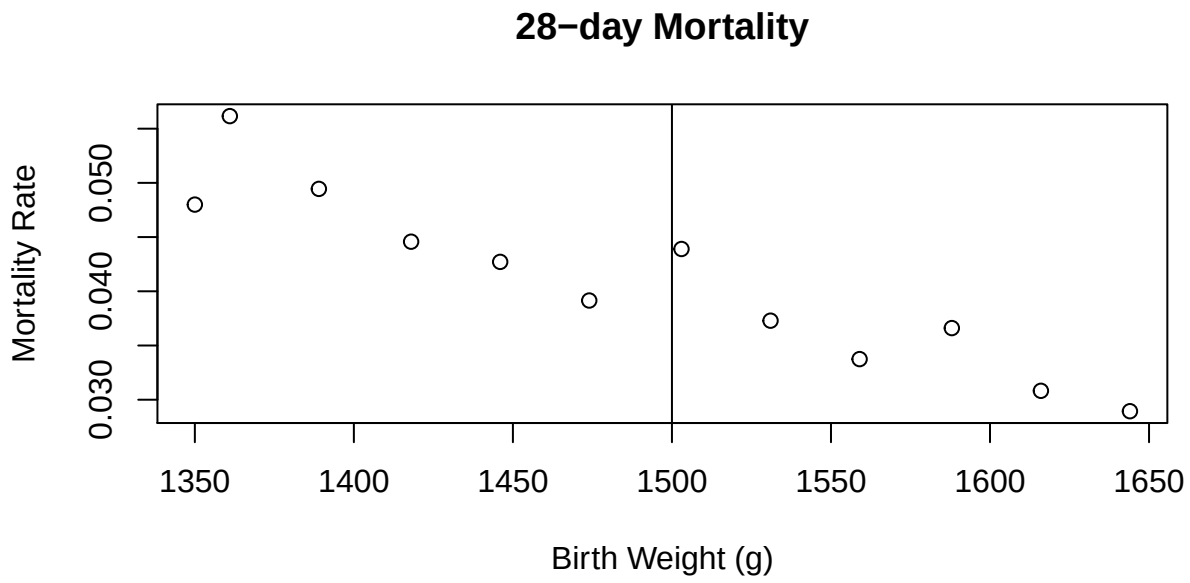
For this case study, we will look at 28 day mortality for newborns based on a cutoff of 1500g. Let's start off by examining how 28 day mortality looks across the running variable, birthweight.

```
data$bin = floor((data$bweight - 1500)/28.35)
mort = data %>% group_by(bin) %>% summarise(m_mort28 =
mean(agedth4), med_weight = median(bweight))
```

Here, we're binning our data by every 28.35 grams (or in one ounce increments) and taking the median weight within each bin.

Plotting this data provides the following graph:

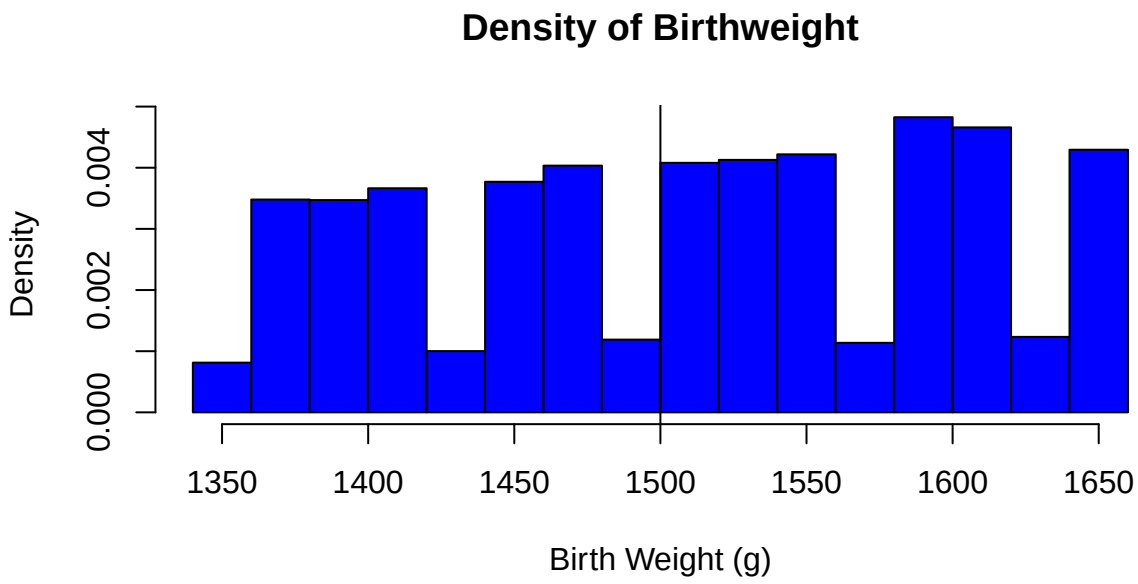
```
plot(mort$med_weight, mort$m_mort28, main="28-day Mortality", xlab="Birth Weight (g)",
     ylab = "Mortality Rate")
abline(v=1500)
```



From the above we can see that mortality is declining prior to the cutoff and then jumps up once we cross the 1500g threshold. This suggests that there is an increase in mortality rate for the untreated units (> 1500g)

Now let's look at the density of the running variable to assess manipulability

```
hist(data$bweight, main = "Density of Birthweight", breaks = 15, freq = FALSE,
     xlab = "Birth Weight (g)", col = "blue")
abline(v = 1500)
```

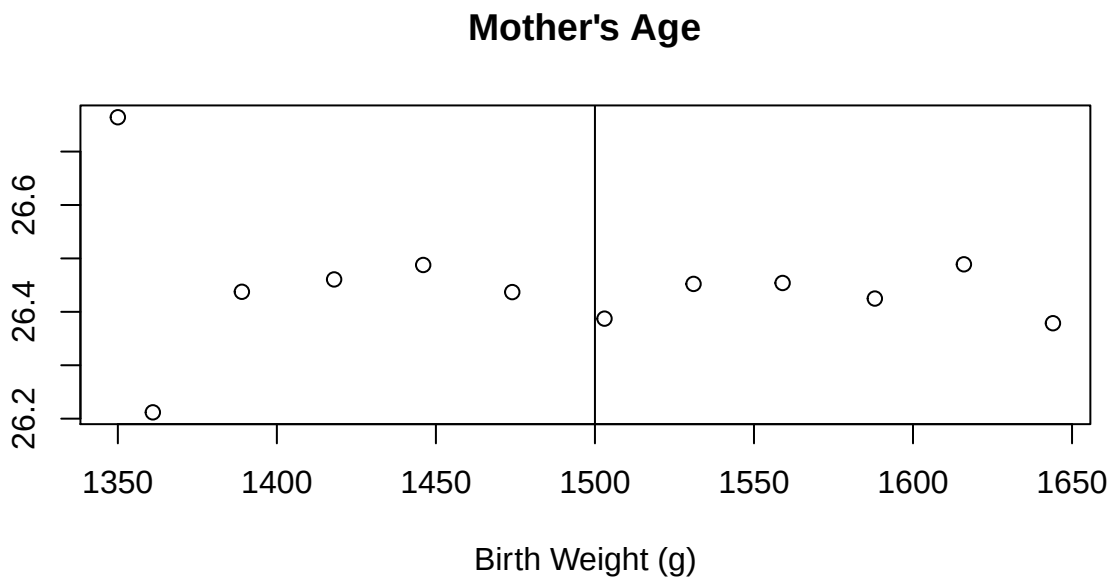


There is a clear drop off in birthweight just before the cutoff. Though notice there are additional drop offs happening across the data. This may indicate some underlying mechanism of reporting birthweight that we are not aware of, rather than manipulability. Additionally, recall that being below the cutoff receives treatment in this case. If doctors were falsely reporting birthweight to move newborns into treatment we would expect opposite bunching than what we see above. Therefore, it is safe to assume no manipulability.

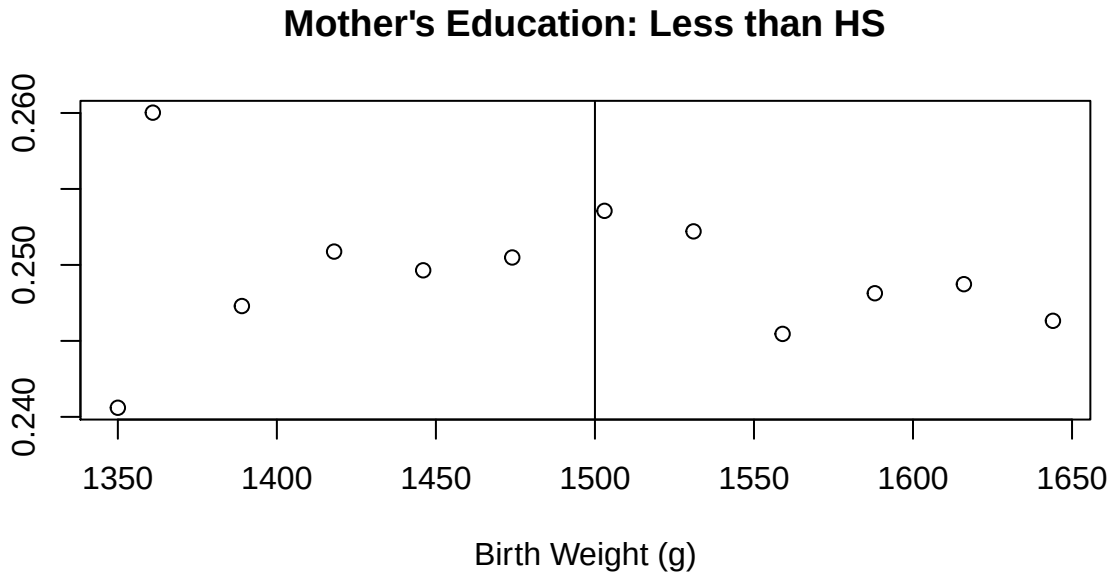
We are going to assume a sharp RD design for this. So in our data, everyone below 1500g will be treated and everyone above 1500g will be untreated.

Lastly, we need to test our assumption for RD, that there is continuity in potential outcomes. Since we cannot directly test this we will assess continuity for some observable covariates as a proxy.

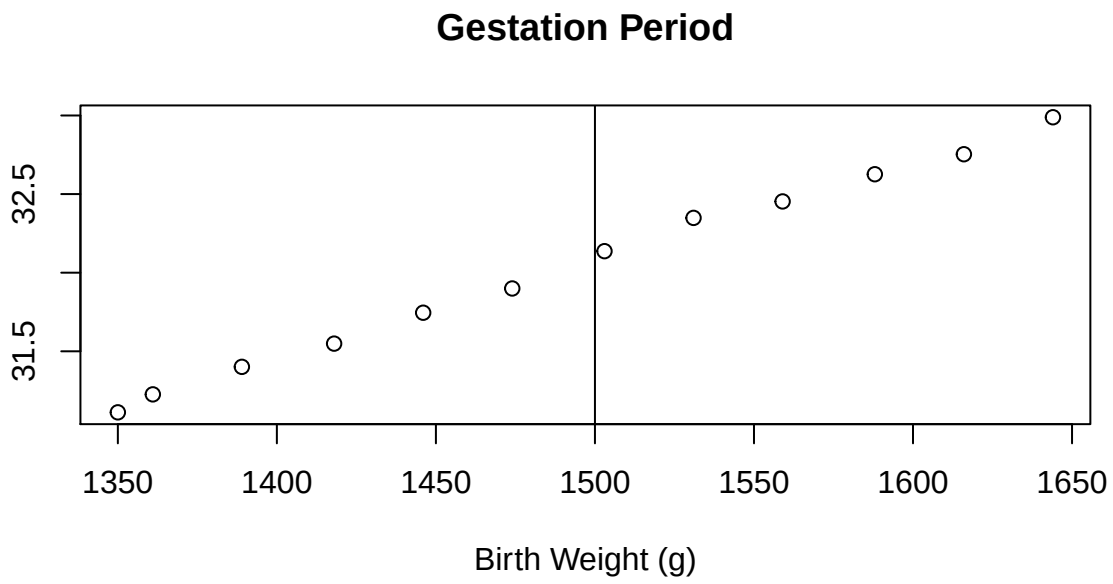
```
# we'll plot using our bins again
covar = data %>% group_by(bin) %>% summarise(m_mom_age = mean(mom_age),
      m_mom_ed1 = mean(mom_ed1),
      m_gest = mean(gest, na.rm=TRUE),
      m_nprenatal = mean(nprenatal, na.rm=TRUE),
      m_yob = mean(yob),
      med_weight = median(bweight))

plot(covar$med_weight, covar$m_mom_age, main = "Mother's Age", xlab = "Birth Weight (g)",
     ylab = "")
abline(v = 1500)
```



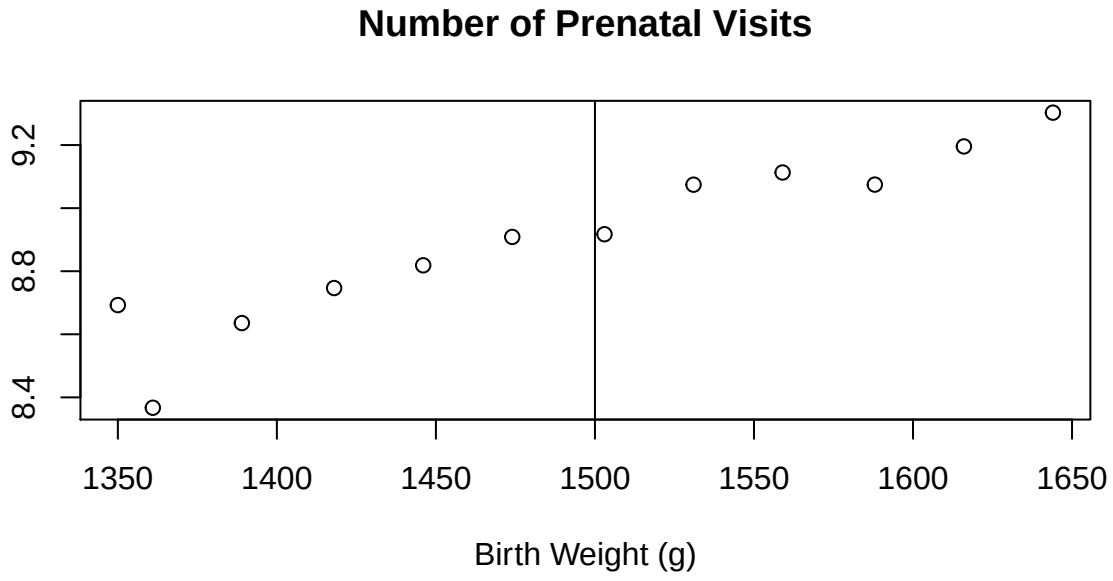

```
plot(covar$med_weight, covar$m_mom_ed1, main = "Mother's Education: Less than HS",
     xlab = "Birth Weight (g)", ylab = "")
abline(v = 1500)
```



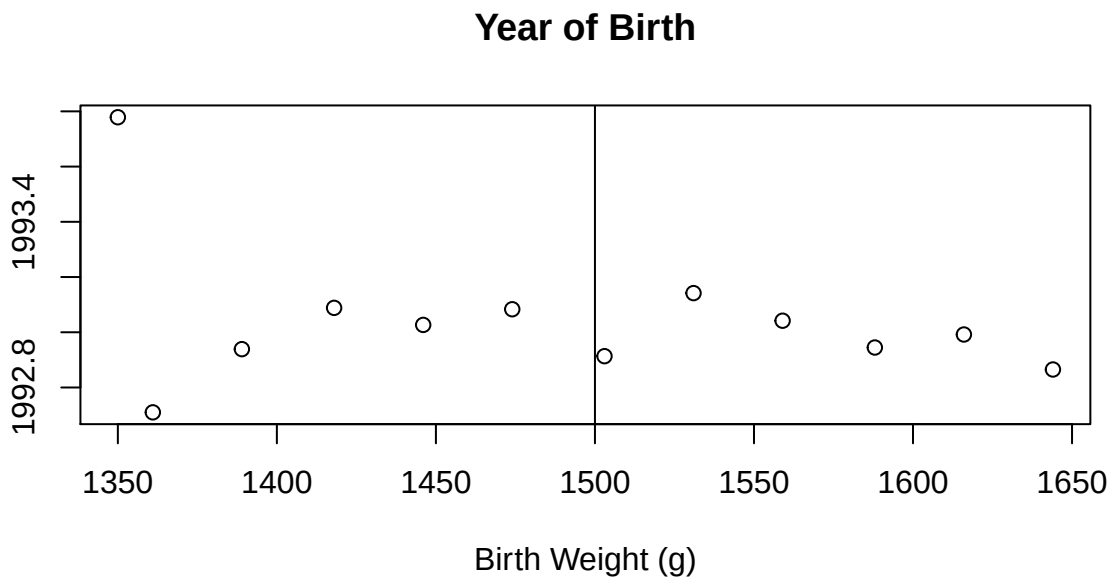

```
plot(covar$med_weight, covar$m_gest, main = "Gestation Period",
     xlab = "Birth Weight (g)", ylab = "")
abline(v = 1500)
```




```
plot(covar$med_weight, covar$m_nprenatal, main = "Number of Prenatal Visits",
     xlab = "Birth Weight (g)", ylab = "")
abline(v = 1500)
```




```
plot(covar$med_weight, covar$m_yob, main = "Year of Birth", xlab = "Birth Weight (g)",
     ylab = "")
abline(v = 1500)
```



Based on the graphs above, there are no egregious discontinuities across the covariates so we assume that our identifying assumption holds.

Now we can run the RD regression and assess our treatment effect.

We will run the following regression:

$$Y_i = \alpha + \tau D_i + \beta_1(X_i - 1500) + \beta_2(X_i - 1500)D_i + \epsilon_i$$

```
# Create treatment dummy for sharp RD
data$vlbw[data$bweight < 1500] = 1
data$vlbw[data$bweight >= 1500] = 0

# Create centered bweight variable (x-c) and interaction term (x-c)D
data$bweight_c = data$bweight - 1500
data$inter = data$bweight_c*data$vlbw

# Run RD regression above
rdd = lm(agedth4 ~ vlbw + bweight_c + inter, data)
stargazer(rdd, type = "text", title = "Effect of Additional Care on 28 Day Mortality")
```

```
##
## Effect of Additional Care on 28 Day Mortality
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                 agedth4
## ------------------------------------------------
## vlbw                           -0.008***
##                                 (0.001)
##
## bweight_c                     -0.0001***
##                                 (0.00001)
```

```
##
## inter                        -0.00005***
##                               (0.00001)
##
## Constant                      0.043***
##                               (0.001)
##
## ------------------------------------------------
## Observations                    376,408
## R2                               0.001
## Adjusted R2                      0.001
## Residual Std. Error     0.196 (df = 376404)
## F Statistic         184.749*** (df = 3; 376404)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

From the regression output above we see that by receiving more additional hospital care, the 28 day mortality rate decreases by .008. Note that the rate across all the data is .0398. This is a 20% reduction from the full sample mortality rate.

In RD it is a best practice to check for robustiness by using different bandwidths and seeing if results hold. We will run this again for weights between 1415g and 1585g (120g bandwidth) and also between 1470g and 1530g (30g bandwidth).

```
rdd2 = lm(agedth4 ~ vlbw + bweight_c + inter, data[data$bweight >= 1415 & data$bweight <=1585,])
rdd3 = lm(agedth4 ~ vlbw + bweight_c + inter, data[data$bweight >= 1470 & data$bweight <=1530,])
stargazer(rdd, rdd2, rdd3, type = "text", title = "Effect of Additional Care on 28 Day Mortality - Robus
```

```
##
## Effect of Additional Care on 28 Day Mortality - Robustiness Check
## ================================================================================================
##                                              Dependent variable:
##                     ---------------------------------------------------------------------------
##                                                    agedth4
##                             Full Sample              120g                    30g
##                                 (1)                  (2)                     (3)
## ------------------------------------------------------------------------------------------------
## vlbw                         -0.008***            -0.009***               -0.020***
##                               (0.001)              (0.002)                 (0.004)
##
## bweight_c                    -0.0001***           -0.0002***              -0.001***
##                              (0.00001)            (0.00002)               (0.0001)
##
## inter                        -0.00005***           0.0001**                0.0003
##                              (0.00001)            (0.00004)               (0.0002)
##
## Constant                      0.043***             0.045***                0.049***
##                               (0.001)              (0.001)                 (0.001)
##
## ------------------------------------------------------------------------------------------------
## Observations                    376,408              202,078                 72,941
## R2                               0.001                0.001                   0.001
## Adjusted R2                      0.001                0.0005                  0.001
## Residual Std. Error     0.196 (df = 376404)  0.196 (df = 202074)     0.199 (df = 72937)
```

```
## F Statistic         184.749*** (df = 3; 376404) 34.513*** (df = 3; 202074) 18.367*** (df = 3; 72937)
## ================================================================================================
## Note:                                                                *p<0.1; **p<0.05; ***p<0.01
```

Notice that the treatment effect for the 120g bandwidth is similar to our original findings, but the 30g bandwidth estimate drastically changes. When we look very close to the cutoff the effect is larger at -.020. This makes sense based on our original graph of outcomes vs. the running variable. Mortality rates for very very low birthweights are high indicating that treatment is not much help for severely underweight newborns. Similarly, mortality rates are low for higher birthweights so lack of additional care is not effecting mortality for those units. Those around the cutoff see the most benefits.