

PSet2

10/28/2018

ID: 12174137

1) Start by getting descriptive statistics of birthweight in the sample.

```
birthwt <- cbind(df$bweight)
options(scipen = 100)
stat.desc(birthwt)
```

```
##                               V1
## nbr.val           376408.00000000
## nbr.null           0.00000000
## nbr.na             0.00000000
## min               1350.00000000
## max               1650.00000000
## range              300.00000000
## sum               568969166.00000000
## median            1515.00000000
## mean              1511.57564664
## SE.mean            0.14509062
## CI.mean.0.95       0.28437331
## var                7923.87364219
## std.dev            89.01614259
## coef.var           0.05888964
```

From the table, we can see that:

mean = 1511.58 grams

standard deviation = 89.02 grams

min = 1350 grams

max = 1650 grams

2) Now plot one year and 28 day mortality rates against our running variable, birth weight. To do so, make bins of one ounce (28.35 grams) around the 1500 grams threshold, and get the mean mortality rate on each bin. Make a separate graph for each outcome. Describe the relationship between birth weight and mortality. Does it appear to be a discontinuity of mortality around the very low birth weight threshold. How does the number of observations in each bin affect your mean estimates?

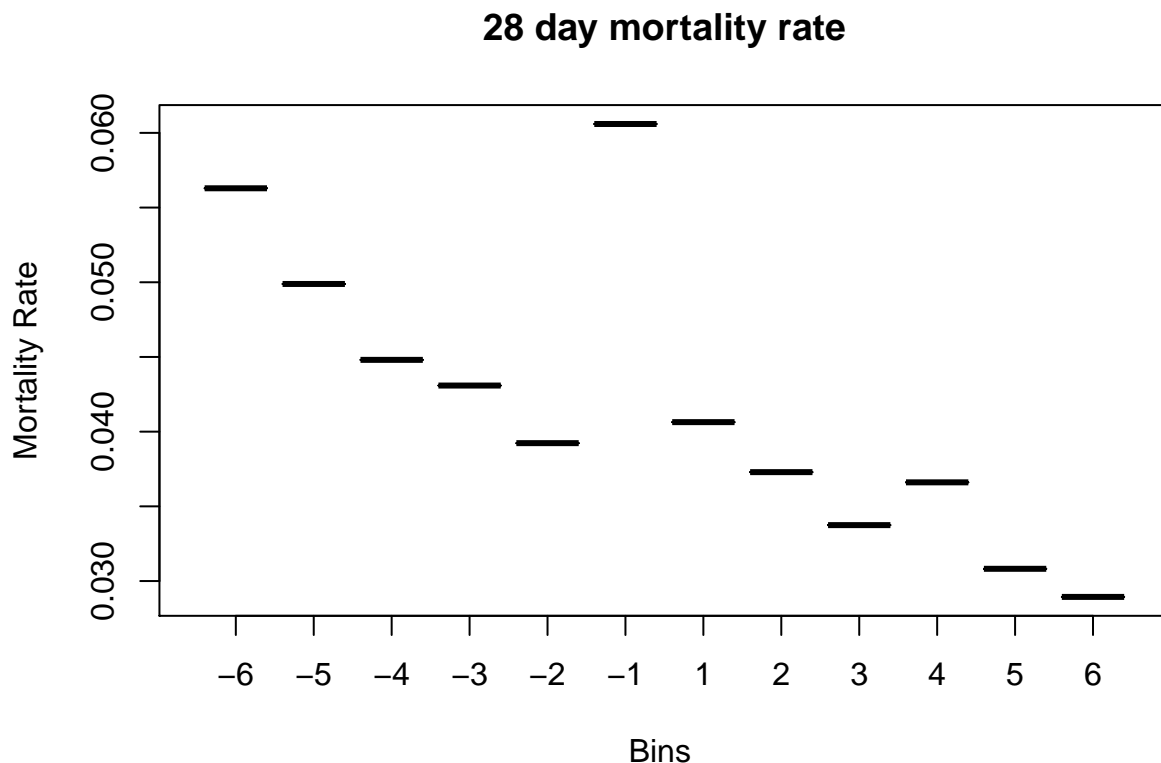
```
# I use the cut function to create a column called bbins where I store the bins for birth year. I # ma
# threshold as negative and those above as positive

df$bbins <- cut(df$bweight,
               breaks = c(1349, 1378.35, 1406.7, 1435.05, 1463.4, 1491.75, 1500, 1528.35, 1556.7),
               labels = c("-6", "-5", "-4", "-3", "-2", "-1", "1", "2", "3", "4", "5", "6"))

# This step groups by bbins and then gets the mean mortality rate and count in two new dataframes.

month_mortrate <- summarize(group_by(df, bbins), mrate=mean(agedth4, na.rm=TRUE), count=n())
year_mortrate <- summarize(group_by(df, bbins), mrate=mean(agedth5, na.rm=TRUE), count=n())

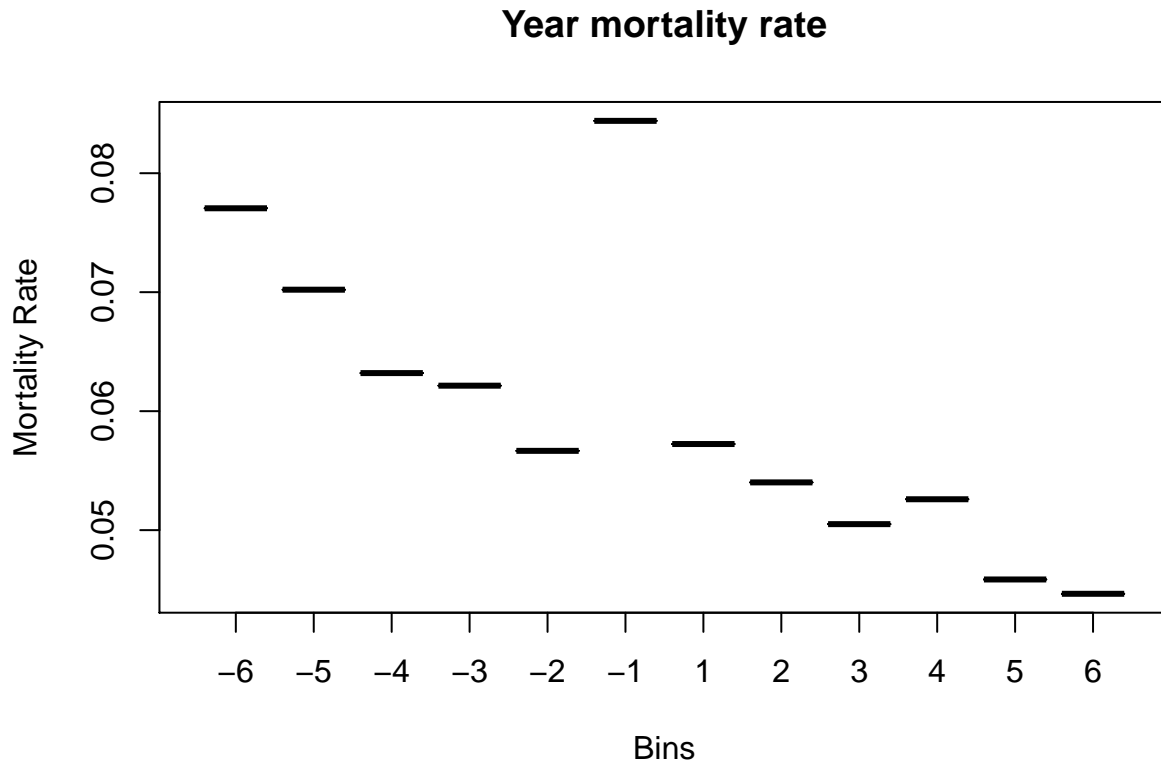
attach(month_mortrate)
plot(bbins, mrate, main="28 day mortality rate",
     xlab="Bins ", ylab="Mortality Rate")
```



```
attach(year_mortrate)

## The following objects are masked from month_mortrate:
##
##      bbins, count, mrate
```

```
plot(bbins, mrate, main='Year mortality rate', xlab='Bins', ylab='Mortality Rate')
```



The relationship between birthweight and mortality appears to be a consistently decreasing relationship – as birthweight increases, mortality decreases both by month and year. However, my graphs do show a significant discontinuity for the caliper immediately before the threshold (1491.75 to 1500) – which could be interpreted as evidence of manipulation: doctors could be falsely recording weights of newborns close to the threshold as below it in order to provide additional care to these newborns. However, it must be noted that the number of observations I have for the caliper that shows the discontinuity is significantly less than the observations for all other bins, which could help explain the apparent discontinuity. Since each caliper is calculating mean mortality, it's important to note that a mean is susceptible to an outlier effect, especially in a small sample size. The discontinuity we see here could be a result of the smaller sample size and a highly influential set of outliers that have an outsized effect given the small sample size.

3) A key assumption for an RDD to provide a causal estimate is that individuals are not able to sort according to the running variable, i.e., they should not be able to manipulate its value. Discuss in your own words whether this is a reasonable assumption in this case. In the write up include tables with the relevant info (Coefficients of interest, standard errors and sample size).

This assumption means a mother or a doctor will not be able to manipulate birthweight for it to land above or below a threshold. This is a reasonable assumption in this case – a mother will not be able to manipulate her child's weight after its born given that she has no control over the child's weight immediately after it's delivered. In fact, in order to manipulate the weight of her newborn, she'd have to bribe the attending physician and I just can't imagine a mother that's just gone through hours of labor or had a C-section being motivated enough to propose

a bribe. On the other hand, it could be realistic for an attending physician to manipulate the weight of a newborn after birth by fudging the readout from the scale. However, it's hard to understand in which direction the physician would want to manipulate the weight. If the physician wanted to make more money, she could manipulate it below the threshold in order to charge more services to the family. On the other hand, if the physician believed that the threshold should be higher, she could manipulate the weight below the threshold to make sure that babies born just over would get the care they need. Both of these mechanisms are a bit convoluted and would require a physician to make decisions that go against the hippocratic oath they've taken. That's why I don't believe the assumptions are out of line in this scenario.

In order to test this assumption I want to check if the covariates are randomly assigned around the threshold, which should give me a sense of whether or not there's manipulation happening. I'll run a series of t-tests on the following variables: mom race, mom age, mom education, gestation weeks, prenatal visits, dad age, dad race, dad age, sex of child, and attendant at birth. We would assume that these variables would not have an impact on whether or not a newborn will be above or below the threshold.

```
df$bt <- ifelse(df$bweight <= 1500, 0, 1)

mage <- lm(bt ~ mom_age, data=df)
mrace <- lm(bt ~ mom_race, data=df)
med <- lm(bt ~ mom_ed, data=df)
gest <- lm(bt ~ gest, data=df)
pren <- lm(bt ~ nprenatal, data=df)
dage <- lm(bt ~ dad_age, data=df, na.action=na.exclude)
drace <- lm(bt ~ dad_race, data=df, na.action=na.exclude)
sex <- lm(bt ~ sex, data=df)
attend <- lm(bt ~ attend, data=df)
stargazer(mage, mrace, med, type='text', title='Mother Demographic Results')
```

```
##
## Mother Demographic Results
## =====
##                               Dependent variable:
##                               -----
##                               bt
##                               (1)          (2)          (3)
## -----
## mom_age                0.0002
##                        (0.0001)
##
## mom_race                -0.009***
##                        (0.001)
##
## mom_ed                                0.002***
##                                (0.0003)
##
## Constant                0.567***          0.584***          0.553***
##                        (0.003)          (0.002)          (0.004)
## -----
## Observations            376,408          376,408          342,553
## R2                      0.00000          0.0001          0.0001
```

```

## Adjusted R2          0.00000          0.0001          0.0001
## Residual Std. Error  0.495 (df = 376406)    0.495 (df = 376406)    0.495 (df = 342551)
## F Statistic          1.622 (df = 1; 376406) 39.520*** (df = 1; 376406) 20.470*** (df = 1; 342551)
## =====
## Note:                                                         *p<0.1; **p<0.05; ***p<0.01
stargazer(gest, pren, sex, attend, type='text', title='Birth Condition Results')

##
## Birth Condition Results
## =====
##                               Dependent variable:
##                               -----
##                               bt
##                               (1)          (2)          (3)
## -----
## gest                0.023***
##                   (0.0003)
##
## nprenatal                0.003***
##                   (0.0002)
##
## sex                                0.007***
##                               (0.002)
##
## attend
##
## Constant            -0.161***
##                   (0.008)
##                   0.544***
##                   (0.002)
##                   0.562***
##                   (0.003)
## -----
## Observations          345,377          340,472          376,408
## R2                    0.023            0.001            0.00004
## Adjusted R2           0.023            0.001            0.00004
## Residual Std. Error    0.489 (df = 345375)    0.494 (df = 340470)    0.495 (df = 376406)
## F Statistic           7,954.480*** (df = 1; 345375) 487.303*** (df = 1; 340470) 16.371*** (df = 1; 376
## =====
## Note:

```

```

stargazer(dage, drace, type='text', title='Dad Demographic Results')

```

```

##
## Dad Demographic Results
## =====
##                               Dependent variable:
##                               -----
##                               bt
##                               (1)          (2)
## -----
## dad_age              0.00004
##                   (0.0001)
##
## dad_race                -0.005***
##                               (0.001)

```

```
##
## Constant                0.572***                0.580***
##                        (0.004)                (0.002)
##
## -----
## Observations            284,913                376,408
## R2                      0.00000                0.0001
## Adjusted R2             -0.00000                0.0001
## Residual Std. Error    0.495 (df = 284911)      0.495 (df = 376406)
## F Statistic            0.098 (df = 1; 284911) 26.769*** (df = 1; 376406)
## =====
## Note:                    *p<0.1; **p<0.05; ***p<0.01
```

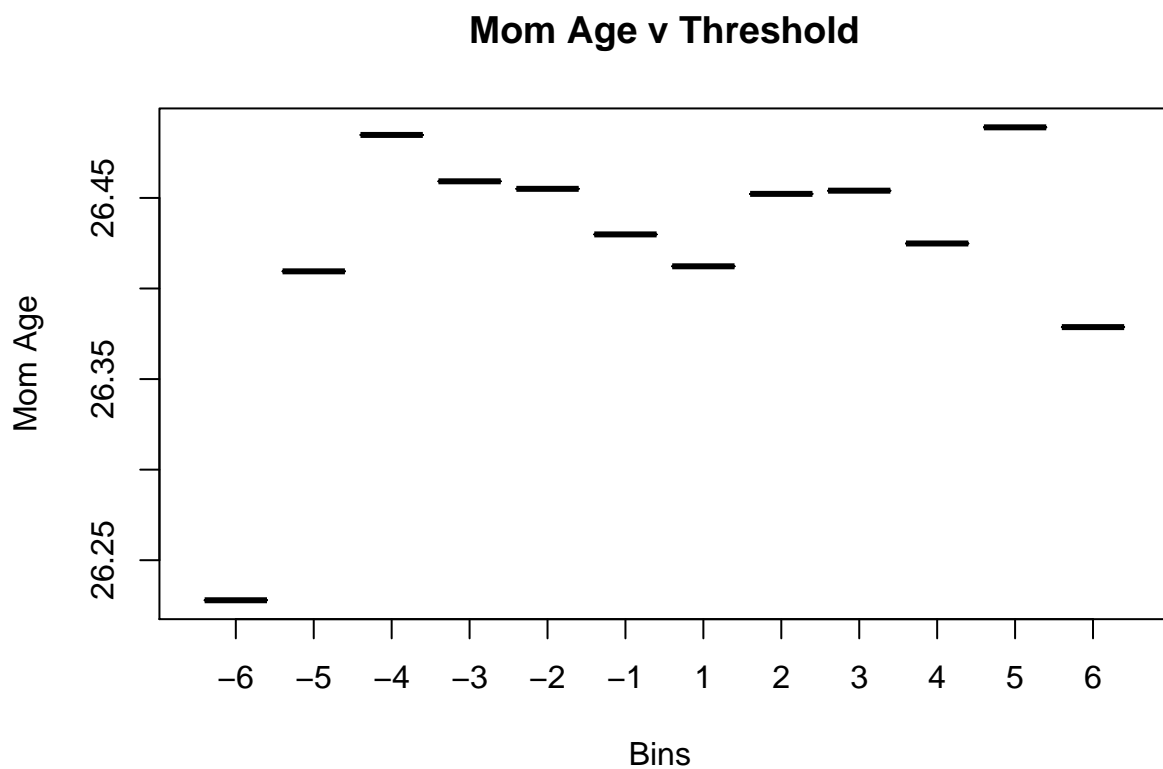
In looking at the table above, which divide the covariates into mom demographic characteristics, birth conditions, and dad demographic characteristics, it appears that none of these covariates have an effect on whether a newborn falls on either side of the threshold. All of the coefficients are pretty close to zero and all of the standard errors are as well. While there are a few coefficients that are not significant at $p < 0.01$ such as dad age and mom age, they still stick pretty closely to zero. I can say that there is no manipulation.

4) Assess informally whether the behavior of other covariates is smooth around the threshold, by plotting the mean of some covariates (mother's age, mother's education less than high school, gestational age, prenatal care visits, and year of birth) against birth weight as you did in point (2). Is there any evidence of discontinuities on other covariates around the very low birth weight threshold? If they were, how could these affect your RDD estimates?

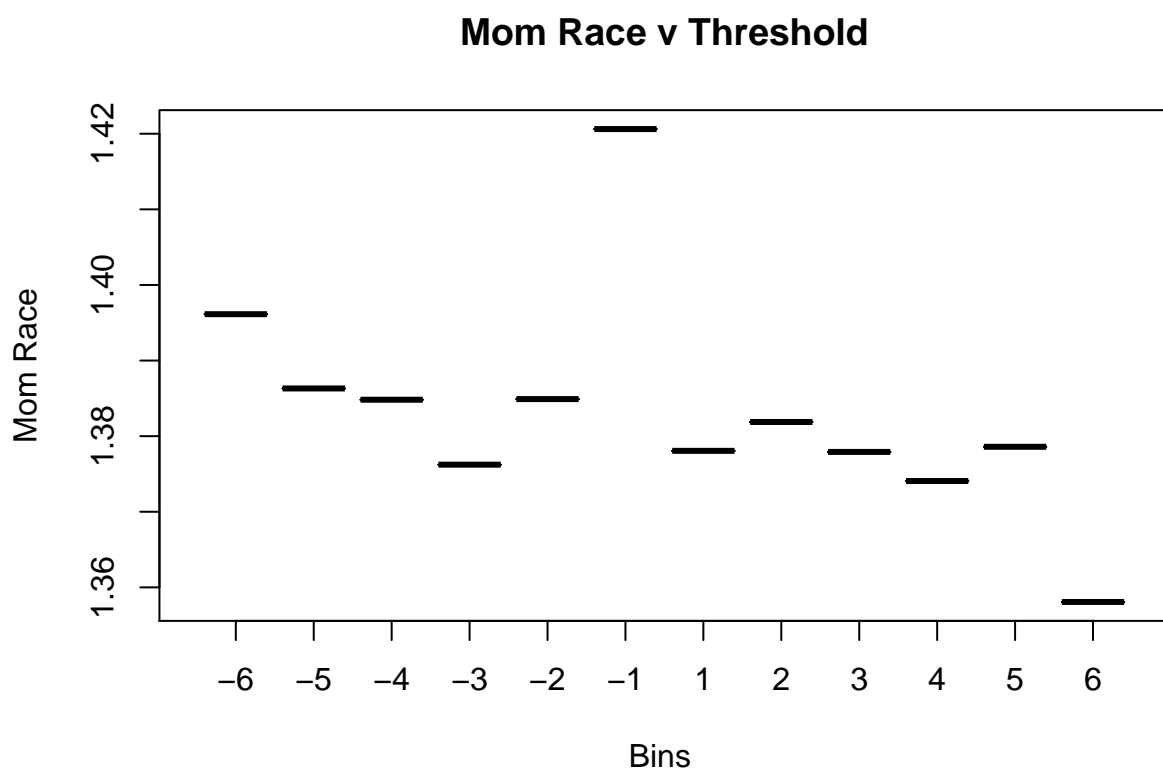
In order to conduct this informal assessment, I'll graph the covariates in the same way that I graphed birthweight around the threshold.

```
mage <- summarize(group_by(df, bbins), mage=mean(mom_age, na.rm=TRUE), count=n())
mrac <- summarize(group_by(df, bbins), mrace=mean(mom_race, na.rm=TRUE), count=n())
med <- summarize(group_by(df, bbins), med=mean(mom_ed, na.rm=TRUE), count=n())
pr <- summarize(group_by(df, bbins), prenat=mean(nprenatal, na.rm=TRUE), count=n())
g <- summarize(group_by(df, bbins), gests=mean(gest, na.rm=TRUE), count=n())
sexy <- summarize(group_by(df, bbins), sex=mean(sex, na.rm=TRUE), count=n())
atts <- summarize(group_by(df, bbins), att=mean(attend, na.rm=TRUE), count=n())
dags <- summarize(group_by(df, bbins), dage=mean(dad_age, na.rm=TRUE), count=n())
draces <- summarize(group_by(df, bbins), drace=mean(dad_race, na.rm=TRUE), count=n())

plot(mage$bbins, mage$mage, main="Mom Age v Threshold",
     xlab="Bins ", ylab="Mom Age")
```

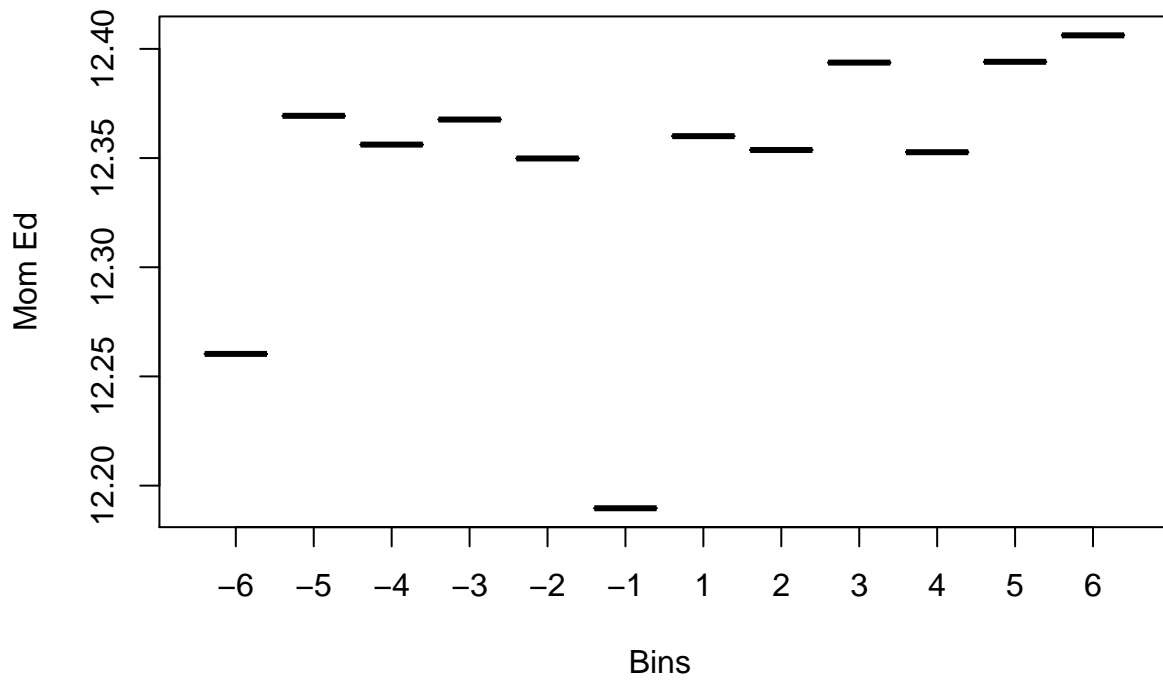


```
plot(mrac$bbins, mrac$mrace, main="Mom Race v Threshold",  
     xlab="Bins ", ylab="Mom Race")
```



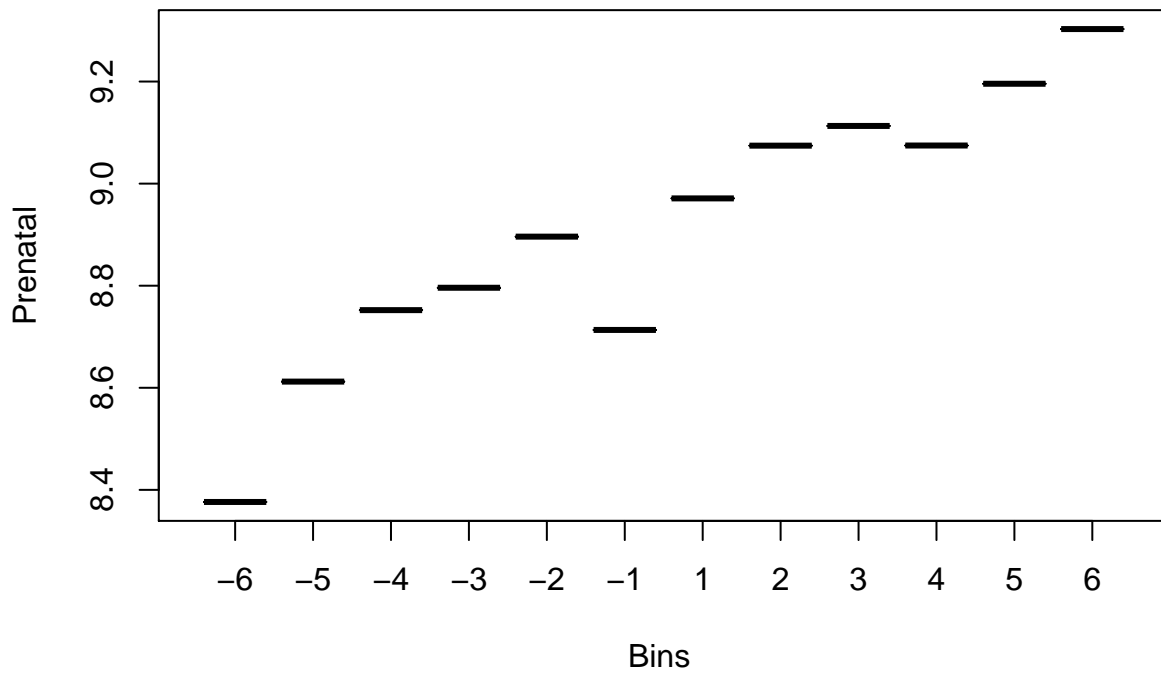
```
plot(med$bbins, med$med, main="Mom Ed v Threshold",  
     xlab="Bins ", ylab="Mom Ed")
```

Mom Ed v Threshold



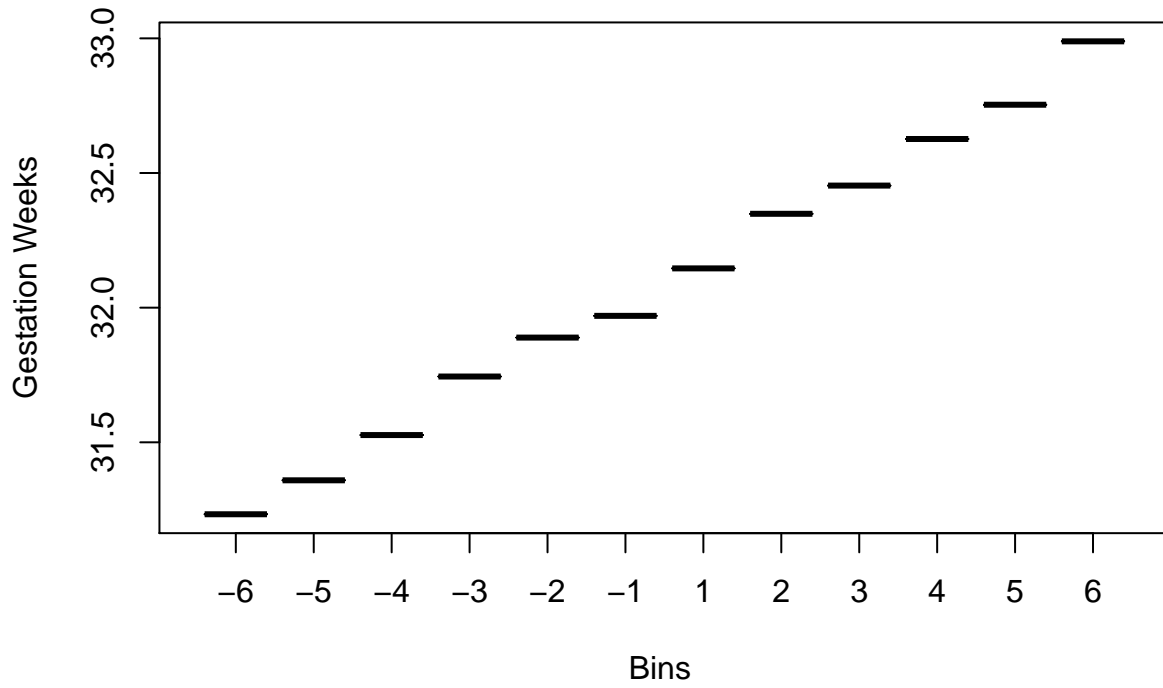
```
plot(pr$bbins, pr$prenat, main="Prenatal v Threshold",  
     xlab="Bins ", ylab="Prenatal")
```

Prenatal v Threshold



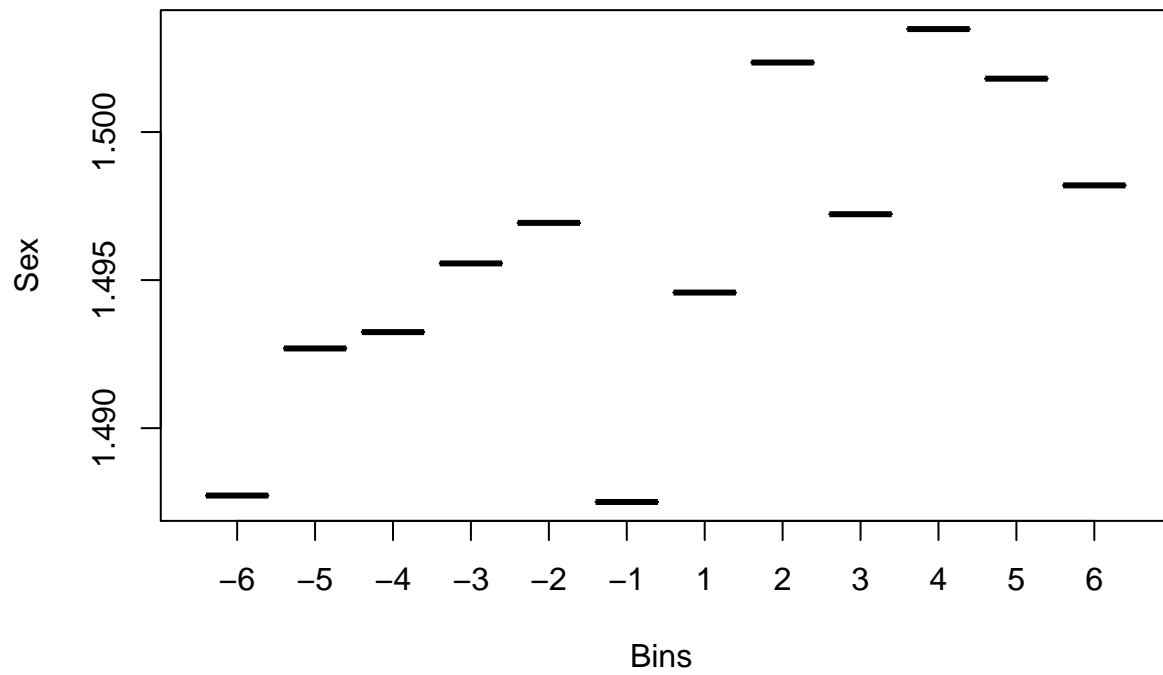
```
plot(g$bbins, g$gests, main="Gestation v Threshold",  
     xlab="Bins ", ylab="Gestation Weeks")
```


Gestation v Threshold



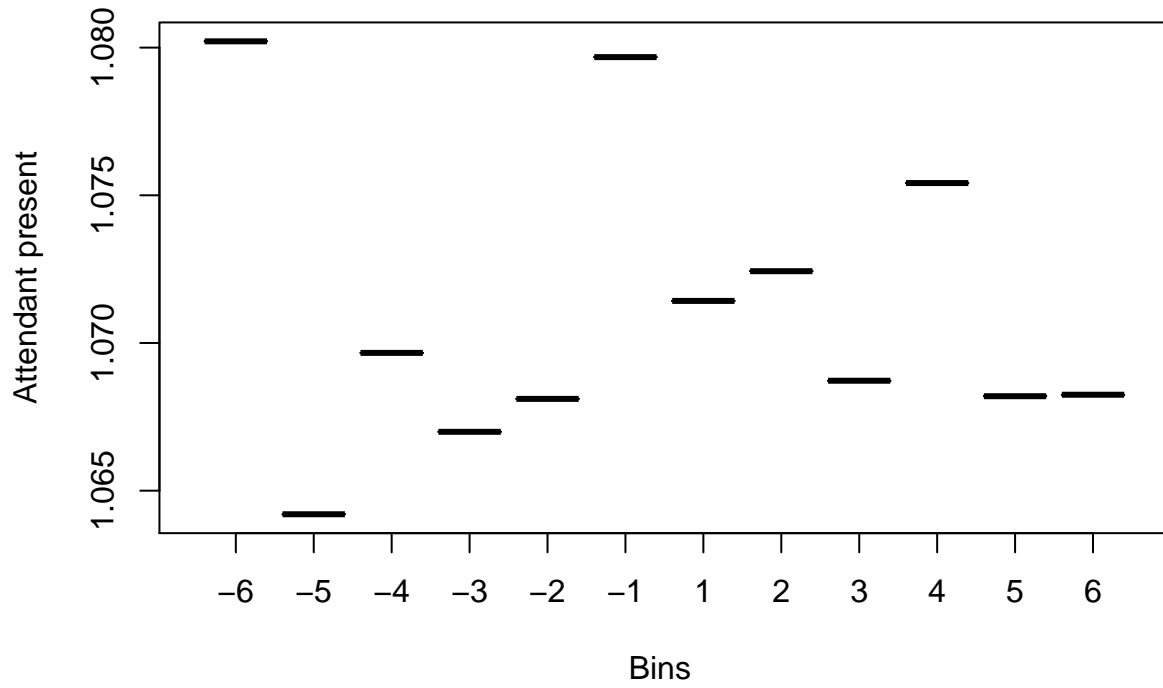
```
plot(sexy$bbins, sexy$sex, main="Sex v Threshold",  
     xlab="Bins ", ylab="Sex")
```

Sex v Threshold



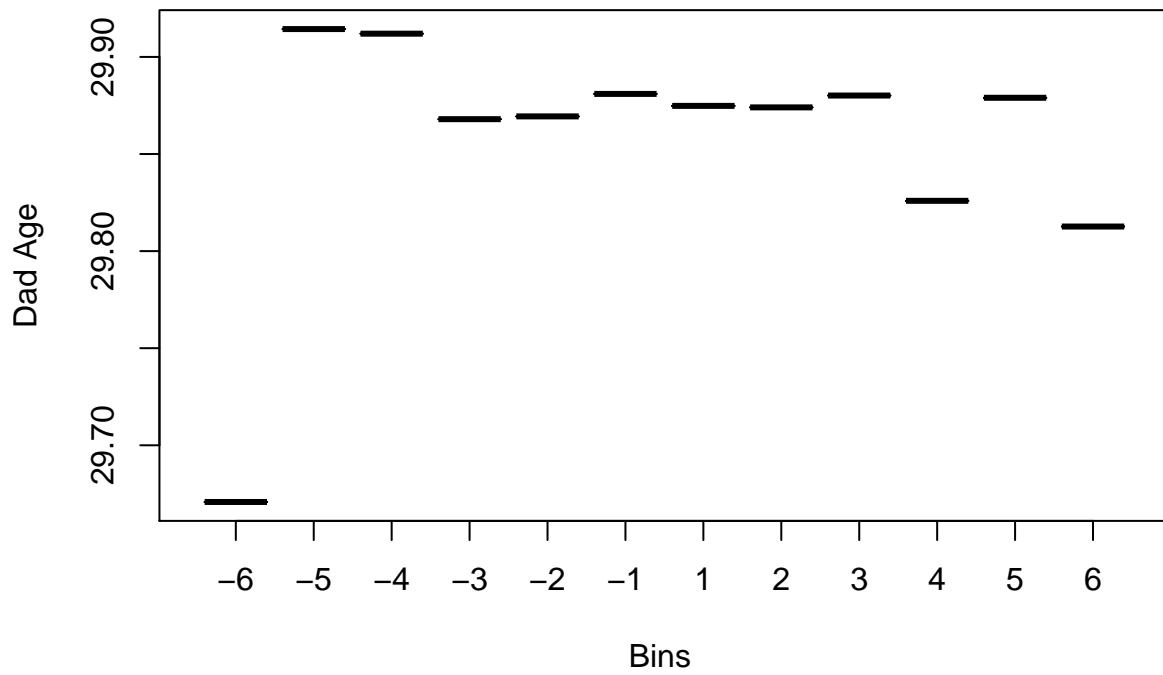
```
plot(atts$bbins, atts$att, main="Attendant v Threshold",  
     xlab="Bins ", ylab="Attendant present")
```

Attendant v Threshold

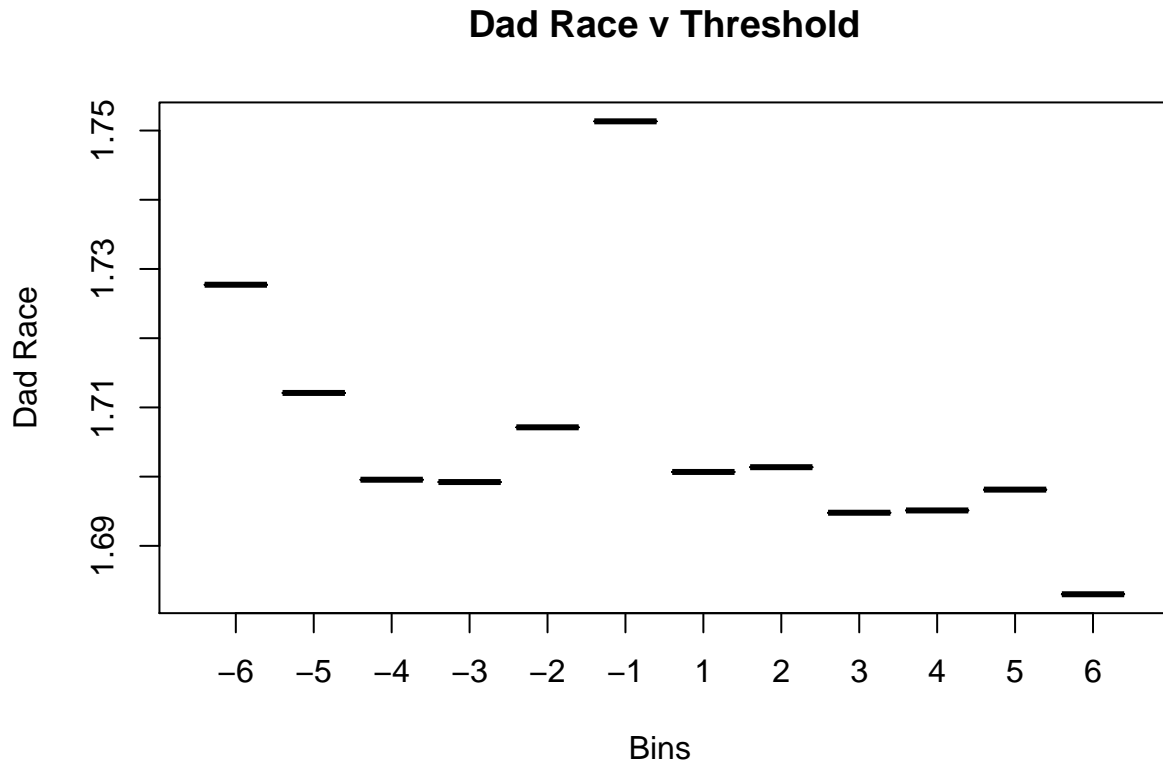


```
plot(dags$bbins, dags$dage, main="Dad Age v Threshold",
     xlab="Bins ", ylab="Dad Age")
```

Dad Age v Threshold



```
plot(draces$bbins, draces$drace, main="Dad Race v Threshold",
     xlab="Bins ", ylab="Dad Race")
```



From the graphs above, it's initially apparent that there might be some discontinuities around the threshold in covariates like dad race, mom age, mom education, among others. However, in all of those instances, looking at the y axis is instructive, as it shows that the discontinuity is relatively small (in mom race for example, the discontinuity is 0.02 for instance). This indicates to me that any discontinuity within these covariates is almost insignificantly small. If these discontinuities did exist, it would make our regression discontinuity analysis less accurate.

5) Now get an estimate of the size of the discontinuity in one-year and 28-day mortality, around the 1500 grams threshold using a caliper of 85 grams (above and below the threshold).

The first step here is to take a subset of the dataframe for birthweight values between 1415g and 1585g, then I'll assign 0 to the values below the threshold and 1 to those above.

```
subs <- subset(df, bweight >= 1415 & bweight <= 1585)
subs$newbins <- ifelse(subs$bweight>1500, 0, 1)
```

Now I'll run a regression for 28 day mortality and return the table.

```
var1 <- subs$newbins*(subs$bweight-1500)
var2 <- (1 - subs$newbins)*(subs$bweight - 1500)
month_regr <- lm(agedth4 ~ newbins + var1 + var2, data=subs)
stargazer(month_regr, type='text', title='28 Day Regression')
```

```
##
## 28 Day Regression
## =====
```

```
##                               Dependent variable:
##                               -----
##                               agedth4
## -----
## newbins                        0.001
##                               (0.002)
##
## var1                          -0.00001
##                               (0.00002)
##
## var2                          -0.0001***
##                               (0.00003)
##
## Constant                      0.042***
##                               (0.001)
## -----
## Observations                  202,078
## R2                            0.0004
## Adjusted R2                   0.0004
## Residual Std. Error          0.196 (df = 202074)
## F Statistic                   26.719*** (df = 3; 202074)
## =====
## Note:                         *p<0.1; **p<0.05; ***p<0.01
```

The table above indicates that there is little discontinuity around the threshold for 28 day mortality, as evidenced by the coefficients, standard errors, and residual standard error which all hover around zero. The coefficients are as follows

$a_1 = 0.001$, this is a non-significant result, with a standard error of 0.002. This means that this result likely doesn't tell us much about the coefficient itself. If it did, however, it would tell us that a newborn with a weight lower than the threshold is marginally (0.001) more likely to die within 28 days.

$a_2 = -.00001$, this is a significant result with a standard error of 0.00002. This interaction coefficient tells us that a newborn that weights marginally less than the threshold and a newborn that weighs significantly less than the threshold die in the first month at similar rates – with the newborn weighting significantly less having a 0.001% greater likelihood of dying in that first month.

$a_3 = -0.0001$, this is a significant result with a standard error of 0.00003. This interaction coefficient measures how much more likely to die is a newborn that weighs above the threshold than one below – 0.003%.

And now I'll run a regression for one year mortality and return the table

```
year_regr <- lm(agedth5 ~ newbins + var1 + var2, data=subs)
stargazer(year_regr, type='text', title='Year Regression')
```

```
##
## Year Regression
```

```

## =====
##                               Dependent variable:
##                               -----
##                               agedth5
## -----
## newbins                      0.003
##                               (0.002)
##
## var1                        -0.00001
##                               (0.00003)
##
## var2                        -0.0001***
##                               (0.00003)
##
## Constant                    0.059***
##                               (0.001)
## -----
## Observations                202,078
## R2                          0.0004
## Adjusted R2                 0.0004
## Residual Std. Error        0.233 (df = 202074)
## F Statistic                27.827*** (df = 3; 202074)
## =====
## Note:                       *p<0.1; **p<0.05; ***p<0.01

```

The table above indicates that there is little discontinuity around the threshold for year mortality, as evidenced by the coefficients, standard errors, and residual standard error which all hover around zero. The coefficients are as follows

$a_1 = 0.003$, this is a non-significant result, with a standard error of 0.002. This means that this result likely doesn't tell us much about the coefficient itself. If it did, however, it would tell us that a newborn with a weight lower than the threshold is marginally (0.003) more likely to die within a year.

$a_2 = -0.00001$, this is a significant result with a standard error of 0.00003. This interaction coefficient tells us that a newborn that weights marginally less than the threshold and a newborn that weighs significantly less than the threshold die in the first year at similar rates – with the newborn weighting significantly less having a 0.001% greater likelihood of dying in that first year.

$a_3 = -0.0001$, this is a significant result with a standard error of 0.00003. This interaction coefficient measures how much more likely to die is a newborn that weighs above the threshold than one below – 0.003%.

6) Now add covariates to the model in (5). Include mother's age, indicators for mother's education and race, indicators for year of birth, indicators for gestational age and prenatal care visits. Use the dummies provided in the data for gestational age and prenatal care visits. Compare your estimates to those obtained in (5) and explain the difference if any.

```
month_regr_cov <- lm(agedth4 ~ newbins + var1 + var2 + mom_age + mom_ed + mom_race + yob + gest_wks1 + gest_wks2 + gest_wks3 + gest_wks4 + nprenatal_1 + nprenatal_2, data=stargazer_data)
stargazer(month_regr_cov, type='text', title='28 Day Regression')
```

```
##
## 28 Day Regression
## =====
##                               Dependent variable:
##                               -----
##                               agedth4
## -----
## newbins                      0.001
##                               (0.002)
##
## var1                         -0.00001
##                               (0.00003)
##
## var2                         -0.0001***
##                               (0.00003)
##
## mom_age                      0.00003
##                               (0.0001)
##
## mom_ed                       -0.0002
##                               (0.0002)
##
## mom_race                     -0.013***
##                               (0.001)
##
## yob                          -0.002***
##                               (0.0001)
##
## gest_wks1                    -0.009***
##                               (0.002)
##
## gest_wks2                    0.014***
##                               (0.002)
##
## gest_wks3                    0.0005
##                               (0.006)
##
## gest_wks4
##
## nprenatal_1                  -0.001
##                               (0.002)
##
## nprenatal_2                  -0.007***
```

```
## (0.002)
##
## nprenatal_3 -0.012***
## (0.002)
##
## nprenatal_4
##
##
## Constant 3.763***
## (0.150)
##
## -----
## Observations 184,057
## R2 0.007
## Adjusted R2 0.007
## Residual Std. Error 0.190 (df = 184043)
## F Statistic 102.463*** (df = 13; 184043)
## =====
## Note: *p<0.1; **p<0.05; ***p<0.01
```

There are no differences in the coefficient and standard error estimates with the covariates added, which means that our covariates are fully independent from the threshold.

```
year_regr_cov <- lm(agedth5 ~ newbins + var1 + var2 + mom_age + mom_ed + mom_race + yob + gest_wks1 + g
stargazer(year_regr_cov, type='text', title='Year Regression')
```

```
##
## Year Regression
## =====
## Dependent variable:
## -----
## agedth5
## -----
## newbins 0.003
## (0.002)
##
## var1 -0.00001
## (0.00003)
##
## var2 -0.0001***
## (0.00003)
##
## mom_age -0.0002**
## (0.0001)
##
## mom_ed -0.001***
## (0.0002)
##
## mom_race -0.010***
## (0.001)
##
## yob -0.002***
## (0.0001)
##
```

```

## gest_wks1                -0.011***
##                          (0.002)
##
## gest_wks2                0.019***
##                          (0.003)
##
## gest_wks3                0.003
##                          (0.008)
##
## gest_wks4
##
##
## nprenatal_1             -0.0004
##                          (0.002)
##
## nprenatal_2             -0.010***
##                          (0.002)
##
## nprenatal_3             -0.013***
##                          (0.003)
##
## nprenatal_4
##
##
## Constant                4.988***
##                          (0.179)
##
## -----
## Observations            184,057
## R2                      0.009
## Adjusted R2             0.008
## Residual Std. Error    0.226 (df = 184043)
## F Statistic            122.059*** (df = 13; 184043)
## =====
## Note:                   *p<0.1; **p<0.05; ***p<0.01

```

There are no differences in the coefficient and standard error estimates with the covariates added, which means that our covariates are fully independent from the threshold.

7) Use the model in (6) to assess the sensitivity of the estimates to the use of different calipers. Use calipers of 30 and 120 grams (above and below the 1500 threshold). Are the estimates any different to those obtained in (6)? What is the tradeoff that we face when increasing/decreasing the caliper?

First I'll do the 30g caliper

```

subs30 <- subset(df, bweight >= 1470 & bweight <= 1530)
subs30$newbins <- ifelse(subs30$bweight>1500, 0, 1)

var11 <- subs30$newbins*(subs30$bweight-1500)
var22 <- (1 - subs30$newbins)*(subs30$bweight - 1500)

```



```
month_regr_cov30 <- lm(agedth4 ~ newbins + var11 + var22 + mom_age + mom_ed + mom_race + yob + gest_wks1 + gest_wks2 + gest_wks3 + gest_wks4 + nprenatal_1 + nprenatal_2 + nprenatal_3 + nprenatal_4, data=month_data, type='text', title='28 Day Regression')
```

```
##
## 28 Day Regression
## =====
##                               Dependent variable:
##                               -----
##                               agedth4
## -----
## newbins                      0.010***
##                               (0.003)
##
## var11                        0.0005***
##                               (0.0001)
##
## var22                       -0.0002**
##                               (0.0001)
##
## mom_age                     -0.0001
##                               (0.0001)
##
## mom_ed                      -0.0001
##                               (0.0003)
##
## mom_race                    -0.012***
##                               (0.001)
##
## yob                         -0.002***
##                               (0.0001)
##
## gest_wks1                   -0.013***
##                               (0.003)
##
## gest_wks2                   0.007*
##                               (0.004)
##
## gest_wks3                   -0.003
##                               (0.011)
##
## gest_wks4
##
##
## nprenatal_1                 -0.006*
##                               (0.003)
##
## nprenatal_2                 -0.012***
##                               (0.003)
##
## nprenatal_3                 -0.019***
##                               (0.004)
##
## nprenatal_4
##
```

```
##
## Constant                3.670***
##                        (0.253)
## -----
## Observations            66,370
## R2                      0.007
## Adjusted R2             0.007
## Residual Std. Error    0.193 (df = 66356)
## F Statistic            35.369*** (df = 13; 66356)
## =====
## Note:                   *p<0.1; **p<0.05; ***p<0.01
year_regr_cov30 <- lm(agedth5 ~ newbins + var11 + var22 + mom_age + mom_ed + mom_race + yob + gest_wks1
stargazer(year_regr_cov30, type='text', title='Year Regression')

##
## Year Regression
## =====
##                        Dependent variable:
##                        -----
##                        agedth5
## -----
## newbins                0.015***
##                        (0.004)
##
## var11                  0.001***
##                        (0.0001)
##
## var22                  -0.0003*
##                        (0.0001)
##
## mom_age                -0.0002
##                        (0.0001)
##
## mom_ed                 -0.002***
##                        (0.0004)
##
## mom_race               -0.010***
##                        (0.002)
##
## yob                    -0.002***
##                        (0.0002)
##
## gest_wks1              -0.016***
##                        (0.003)
##
## gest_wks2              0.014***
##                        (0.005)
##
## gest_wks3              -0.005
##                        (0.013)
##
## gest_wks4
```

```
##
## nprenatal_1          -0.007*
##                      (0.004)
##
## nprenatal_2          -0.015***
##                      (0.004)
##
## nprenatal_3          -0.021***
##                      (0.004)
##
## nprenatal_4
##
##
## Constant             4.755***
##                      (0.301)
##
## -----
## Observations          66,370
## R2                    0.008
## Adjusted R2           0.008
## Residual Std. Error   0.229 (df = 66356)
## F Statistic           43.404*** (df = 13; 66356)
## =====
## Note:                  *p<0.1; **p<0.05; ***p<0.01
```

The estimates with a smaller caliper are different than those from the 85g caliper. Generally speaking, with a smaller caliper, these estimates point to a greater discontinuity around the threshold than those of the 85g caliper.

Next I'll do the 120g caliper

```
subs120 <- subset(df, bweight >= 1380 & bweight <= 1620)
subs120$newbins <- ifelse(subs120$bweight>1500, 0, 1)

var111 <- subs120$newbins*(subs120$bweight-1500)
var222 <- (1 - subs120$newbins)*(subs120$bweight - 1500)

month_regr_cov120 <- lm(agedth4 ~ newbins + var111 + var222 + mom_age + mom_ed + mom_race + yob + gest_
stargazer(month_regr_cov120, type='text', title='28 Day Regression')
```

```
##
## 28 Day Regression
## =====
##                      Dependent variable:
##                      -----
##                      agedth4
## -----
## newbins              0.001
##                      (0.001)
##
## var111               -0.0001***
##                      (0.00002)
##
```

```

## var222                -0.0001***
##                      (0.00001)
##
## mom_age               0.0001
##                      (0.0001)
##
## mom_ed               -0.0002
##                      (0.0002)
##
## mom_race             -0.012***
##                      (0.001)
##
## yob                  -0.002***
##                      (0.0001)
##
## gest_wks1            -0.008***
##                      (0.001)
##
## gest_wks2            0.013***
##                      (0.002)
##
## gest_wks3            0.002
##                      (0.005)
##
## gest_wks4
##
## nprenatal_1          -0.001
##                      (0.002)
##
## nprenatal_2          -0.007***
##                      (0.002)
##
## nprenatal_3          -0.011***
##                      (0.002)
##
## nprenatal_4
##
## Constant             3.829***
##                      (0.121)
##
## -----
## Observations          277,373
## R2                    0.007
## Adjusted R2           0.007
## Residual Std. Error   0.189 (df = 277359)
## F Statistic           156.290*** (df = 13; 277359)
## =====
## Note:                  *p<0.1; **p<0.05; ***p<0.01
year_regr_cov120 <- lm(agedth5 ~ newbins + var111 + var222 + mom_age + mom_ed + mom_race + yob + gest_w
stargazer(year_regr_cov120, type='text', title='Year Regression')
##

```

```

## Year Regression
## =====
##                               Dependent variable:
##                               -----
##                               agedth5
## -----
## newbins                      0.002
##                               (0.002)
##
## var111                       -0.0001***
##                               (0.00002)
##
## var222                       -0.0001***
##                               (0.00001)
##
## mom_age                      -0.0002**
##                               (0.0001)
##
## mom_ed                       -0.001***
##                               (0.0002)
##
## mom_race                     -0.010***
##                               (0.001)
##
## yob                          -0.003***
##                               (0.0001)
##
## gest_wks1                    -0.010***
##                               (0.002)
##
## gest_wks2                    0.019***
##                               (0.002)
##
## gest_wks3                    0.006
##                               (0.006)
##
## gest_wks4
##
## nprenatal_1                  -0.0002
##                               (0.002)
##
## nprenatal_2                  -0.009***
##                               (0.002)
##
## nprenatal_3                  -0.013***
##                               (0.002)
##
## nprenatal_4
##
## Constant                     5.107***
##                               (0.145)
##

```

```
## -----
## Observations          277,373
## R2                    0.009
## Adjusted R2           0.009
## Residual Std. Error   0.226 (df = 277359)
## F Statistic           192.998*** (df = 13; 277359)
## =====
## Note:                  *p<0.1; **p<0.05; ***p<0.01
```

The estimates with a larger caliper are relatively similar, if a little smaller, to the estimates obtained with the 85g caliper.

8) Synthetize your findings and discuss what kind of supplementary information would you need to make a cost–benefit analysis of treatment received by newborns close to the very low birth weight threshold.

Through this analysis, I’ve determined that it’s highly unlikely that there is manipulation of birthweight around the threshold – there’s no evidence of it from running regressions on it, or from graphing it both without covariates and with covariates. While there is some evidence of slight discontinuities with smaller calipers versus bigger ones, this can be explained by the fact that newborns immediately below the threshold are very similar to newborns immediately above it, so the fact that they get extra care from being born below the threshold does have a discernible effect on mortality rates. However, for larger calipers, it appears that a newborn with a birthweight above the threshold has similar mortality outcomes to a newborn with a birthweight below the threshold, thus suggesting that the threshold is largely arbitrary and has no outcomes on mortality rates.

In terms of additional information to do a cost-benefit analysis for newborns just around the threshold, I’d want to take into account other outcomes beyond mortality – does providing a low weight newborn baby with extra care reduce follow-up doctor visits? Does not providing a low weight newborn baby with extra care reduce length of hospital stay and by how much compared to those who receive extra care? Does receiving treatment affect other outcomes, such as whether the newborns meet critical milestones in their first year of life: crawling, walking talking. This broader picture would let me decide whether or not extra care improves outcomes for newborns and for the hospital, and is therefore cost effective.

Mario Moreno