PE-PSet2

Drazzel Feliu - 12174100

For this assignment, provide a write–up where you answer the questions below, selectively cutting and pasting output where needed. Be concise in your write-up; excess wordiness will be penalized. Also, submit a log file that includes commands and results for your entire analysis. The assignment makes use of almond_etal_2008.dta, which you can find on Canvas.

```
# load data set
data <- read_dta("almond_etal_2008.dta")
# create data table identifying class and labels for each variable
datainfo <- data.frame(variable=colnames(data),
              class=sapply(data, class),
              label=sapply(data, function (x) attr(x, "label")))
datainfo <- `rownames<-`(datainfo, 1:44)
datainfo
```

```
##       variable    class                    label
## 1          yob  numeric             Year of birth
## 2          yod  numeric             Year of death
## 3      staters  numeric        State of Residence
## 4      mom_age  numeric  Age of mother (at birth)
## 5     mom_race labelled            Race of mother
## 6       mom_ed  numeric        Mother's education
## 7      mom_ed1  numeric               Less than HS
## 8      mom_ed2  numeric                    HS only
## 9      mom_ed3  numeric              Some college
## 10     mom_ed4  numeric           College or more
## 11     mom_ed5  numeric                Ed Missing
## 12     bweight  numeric      Birth weight (grams)
## 13        gest  numeric         Gestation (weeks)
## 14   gest_wks1  numeric       Gestation <37 weeks
## 15   gest_wks2  numeric     Gestation 37-42 weeks
## 16   gest_wks3  numeric       Gestation >42 weeks
## 17   gest_wks4  numeric         Gestation Missing
## 18      agedth labelled       Infant age at death
## 19     agedth1  numeric        One-hour mortality
## 20     agedth2  numeric         24-hour mortality
## 21     agedth3  numeric        One-week mortality
## 22     agedth4  numeric          28-day mortality
## 23     agedth5  numeric        One-year mortality
## 24   nprenatal  numeric Number of prenatal visits
## 25 nprenatal_1  numeric        Prenatal visits <9
## 26 nprenatal_2  numeric      Prenatal visits 9-14
## 27 nprenatal_3  numeric       Prenatal visits >14
## 28 nprenatal_4  numeric    Prental visits Missing
## 29      apgar5  numeric      5-minute Apgar score
## 30    apgar5_1  numeric          Apgar score <= 1
## 31    apgar5_3  numeric          Apgar score <= 3
## 32    apgar5_5  numeric          Apgar score <= 5
## 33    apgar5_7  numeric          Apgar score <= 7
## 34     stateoc  numeric            State of birth
## 35     dad_age  numeric              Father's age
```

```
## 36    dad_race labelled            Race of father
## 37         sex  numeric              Sex of child
## 38      plural  numeric                 Plurality
## 39  mom_origin  numeric           Mother's origin
## 40  dad_origin  numeric           Father's origin
## 41   tot_order  numeric         Total birth order
## 42  live_order  numeric          Live birth order
## 43       pldel labelled         Place of delivery
## 44      attend labelled        Attendant at birth
```

```
# summary statistics of variables
summary(data)
```

```
##       yob            yod            staters        mom_age
##  Min.   :1983   Min.   :  85     Min.   : 0.0   Min.   :10.00
##  1st Qu.:1987   1st Qu.:  88     1st Qu.:11.0   1st Qu.:21.00
##  Median :1995   Median :1989     Median :24.0   Median :26.00
##  Mean   :1993   Mean   :1438     Mean   :24.9   Mean   :26.43
##  3rd Qu.:1999   3rd Qu.:1997     3rd Qu.:37.0   3rd Qu.:31.00
##  Max.   :2002   Max.   :2003     Max.   :59.0   Max.   :54.00
##                 NA's   :354819
##     mom_race        mom_ed         mom_ed1          mom_ed2
##  Min.   :1.00   Min.   : 0.00   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:1.00   1st Qu.:11.00   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.00   Median :12.00   Median :0.0000   Median :0.0000
##  Mean   :1.38   Mean   :12.36   Mean   :0.2501   Mean   :0.3376
##  3rd Qu.:2.00   3rd Qu.:14.00   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :3.00   Max.   :17.00   Max.   :1.0000   Max.   :1.0000
##                 NA's   :33855
##     mom_ed3          mom_ed4          mom_ed5           bweight
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :1350
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:1445
##  Median :0.0000   Median :0.0000   Median :0.00000   Median :1515
##  Mean   :0.1733   Mean   :0.1491   Mean   :0.08994   Mean   :1512
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:1588
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1650
##
##       gest         gest_wks1        gest_wks2         gest_wks3
##  Min.   :17.00   Min.   :0.0000   Min.   :0.00000   Min.   :0.000000
##  1st Qu.:30.00   1st Qu.:1.0000   1st Qu.:0.00000   1st Qu.:0.000000
##  Median :32.00   Median :1.0000   Median :0.00000   Median :0.000000
##  Mean   :32.14   Mean   :0.8331   Mean   :0.07871   Mean   :0.005752
##  3rd Qu.:34.00   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.000000
##  Max.   :52.00   Max.   :1.0000   Max.   :1.00000   Max.   :1.000000
##  NA's   :31031
##    gest_wks4           agedth          agedth1            agedth2
##  Min.   :0.00000   Min.   :1.0      Min.   :0.000000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:2.0      1st Qu.:0.000000   1st Qu.:0.00000
##  Median :0.00000   Median :3.0      Median :0.000000   Median :0.00000
##  Mean   :0.08244   Mean   :3.3      Mean   :0.005823   Mean   :0.01971
##  3rd Qu.:0.00000   3rd Qu.:5.0      3rd Qu.:0.000000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :5.0      Max.   :1.000000   Max.   :1.00000
##                    NA's   :354819
##     agedth3           agedth4           agedth5          nprenatal
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   : 0.00
```

```
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.: 6.00
##  Median :0.00000   Median :0.00000   Median :0.00000   Median : 8.00
##  Mean   :0.03127   Mean   :0.03988   Mean   :0.05736   Mean   : 8.94
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:12.00
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :49.00
##                                                        NA's   :35936
##   nprenatal_1       nprenatal_2       nprenatal_3       nprenatal_4
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.00000
##  Mean   :0.4564   Mean   :0.3326   Mean   :0.1155   Mean   :0.09547
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000
##
##      apgar5          apgar5_1         apgar5_3         apgar5_5
##  Min.   : 0.00   Min.   :0.00    Min.   :0.00    Min.   :0.00
##  1st Qu.: 7.00   1st Qu.:0.00    1st Qu.:0.00    1st Qu.:0.00
##  Median : 8.00   Median :0.00    Median :0.00    Median :0.00
##  Mean   : 7.91   Mean   :0.01    Mean   :0.03    Mean   :0.07
##  3rd Qu.: 9.00   3rd Qu.:0.00    3rd Qu.:0.00    3rd Qu.:0.00
##  Max.   :10.00   Max.   :1.00    Max.   :1.00    Max.   :1.00
##  NA's   :80056   NA's   :80056   NA's   :80056   NA's   :80056
##     apgar5_7         stateoc         dad_age          dad_race
##  Min.   :0.00    Min.   : 1.0   Min.   :10.00   Min.   :1.000
##  1st Qu.:0.00    1st Qu.:12.0   1st Qu.:25.00   1st Qu.:1.000
##  Median :0.00    Median :27.0   Median :29.00   Median :1.000
##  Mean   :0.26    Mean   :26.9   Mean   :29.85   Mean   :1.702
##  3rd Qu.:1.00    3rd Qu.:39.0   3rd Qu.:34.00   3rd Qu.:3.000
##  Max.   :1.00    Max.   :56.0   Max.   :89.00   Max.   :3.000
##  NA's   :80056                  NA's   :91495
##       sex             plural         mom_origin      dad_origin
##  Min.   :1.000   Min.   :1.000   Min.   : 0.00   Min.   : 0.00
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.: 0.00   1st Qu.: 0.00
##  Median :1.000   Median :1.000   Median : 0.00   Median : 1.00
##  Mean   :1.497   Mean   :1.289   Mean   :11.92   Mean   :17.07
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.: 4.00   3rd Qu.: 9.00
##  Max.   :2.000   Max.   :5.000   Max.   :99.00   Max.   :99.00
##
##    tot_order        live_order         pldel           attend
##  Min.   : 1.000   Min.   : 1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.:1.000   1st Qu.:1.000
##  Median : 2.000   Median : 2.000   Median :1.000   Median :1.000
##  Mean   : 2.594   Mean   : 2.129   Mean   :1.021   Mean   :1.071
##  3rd Qu.: 3.000   3rd Qu.: 3.000   3rd Qu.:1.000   3rd Qu.:1.000
##  Max.   :31.000   Max.   :31.000   Max.   :9.000   Max.   :9.000
##  NA's   :3307     NA's   :2722
```

Motivation

A key policy question in health economics is whether the benefits of additional medical expenditures exceed their cost. The question is particularly relevant since medical expenditures in the United States have been on the rise for a long time. To analyze this question Almond et al (2008), use a RDD design and compare

health outcomes of newborns around the threshold of very low birth weight (1500 grams).They argue that the threshold is commonly used as a rule of thumb to prescribe medical treatment, which is followed mainly by convention, and does not reflect biological criteria. In this problem set we will reproduce some of their basic results,so start by reading their paper, which you can find in Canvas.

Questions:

1

Start by getting the descriptive statistics of birth weight in the sample, what is the mean, standard deviation, minimum, and maximum?

```
mean(data$bweight)
```

```
## [1] 1511.576
```

```
sd(data$bweight)
```

```
## [1] 89.01614
```

```
min(data$bweight)
```

```
## [1] 1350
```

```
max(data$bweight)
```

```
## [1] 1650
```

Answer

The mean birth weight is 1511.58 grams. The standard deviation is 89.02. The minimum birth weight is 1350 grams. The maximum birth weight is 1650 grams.

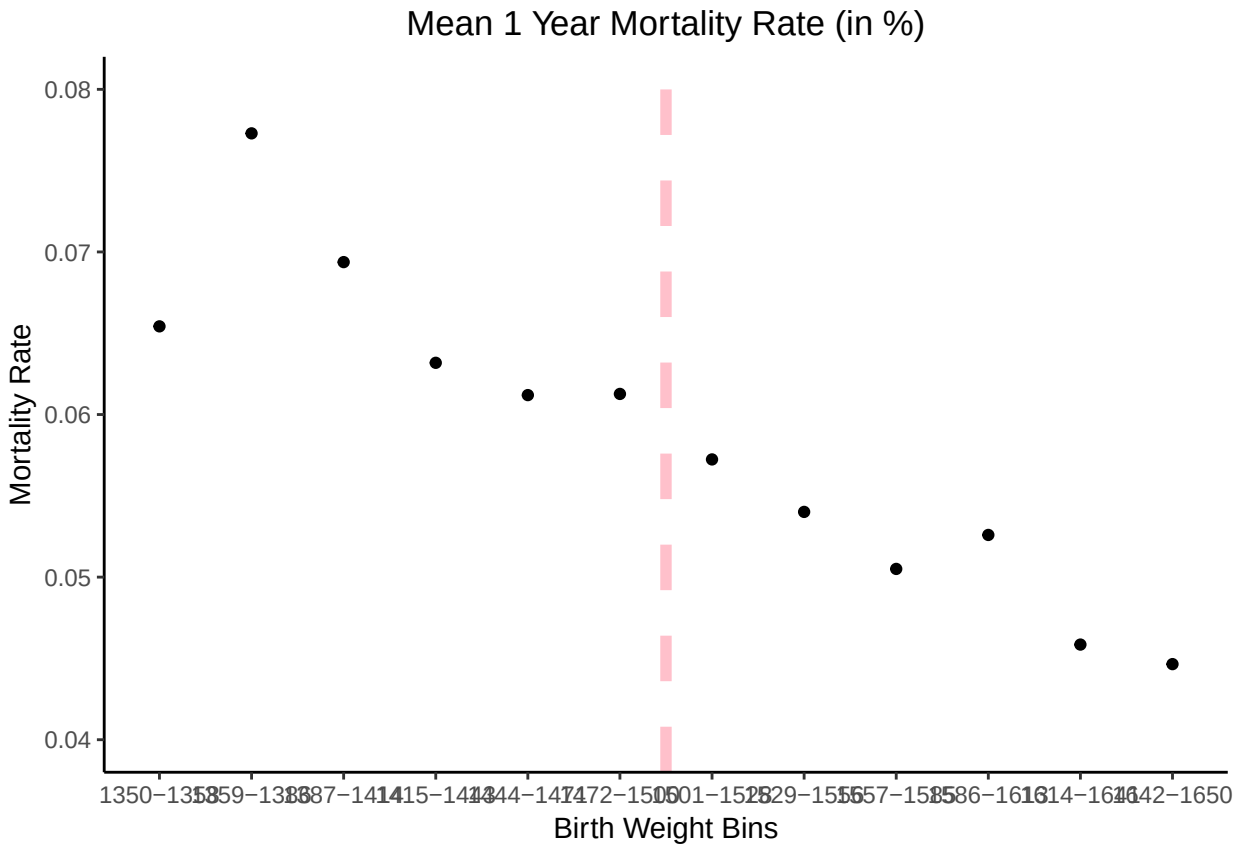
2

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?

Answer

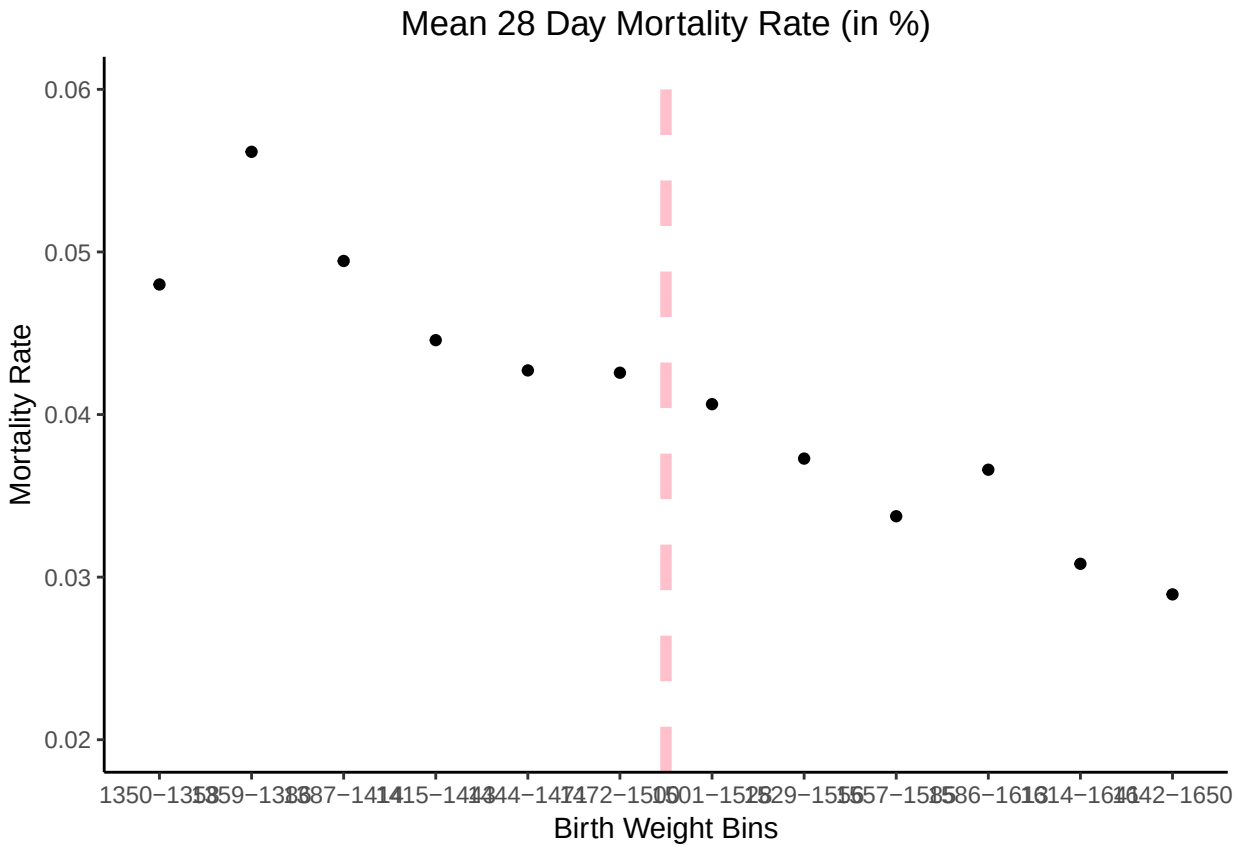
Mean 1 Year Mortality Rate by Birth Weight

```
plot1
```



Mean 28 Day Mortality Rate by Birth Weight

```
plot2
```



The relationship between birth weight and both of the mortality measures we analyzed in our plots are negatively correlated and highly statistically significant, indicating that higher birth weights are correlated with a minimized probability of mortality, both 28 days out and 1 year out from date of birth.

```
cor.test(data$agedth4,data$bweight)
```

```
##
##  Pearson's product-moment correlation
##
## data:  data$agedth4 and data$bweight
## t = -22.761, df = 376410, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.04026415 -0.03388369
## sample estimates:
##        cor
## -0.0370743
```

```
cor.test(data$agedth5,data$bweight)
```

```
##
##  Pearson's product-moment correlation
##
## data:  data$agedth5 and data$bweight
## t = -23.633, df = 376410, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.04168074 -0.03530097
```

```
## sample estimates:
##         cor
## -0.03849125
```

The regression discontinuity at 1500 grams seems to be fuzzy. It does not indicate a marked increase in the probability of a premature death at either of the date cutoffs specified previously (28 days or 1 year).

The mean estimate for the lowest bin (1350-1358) is sensitive because the number of observations is ~10% the size of the other bins. Any outliers in either direction are likely to bias the mean estimate of that bin.

```
data %>% count(bins)
```

```
## # A tibble: 12 x 2
##    bins          n
##    <fct>     <int>
##  1 1350-1358  3271
##  2 1359-1386 30236
##  3 1387-1414 29002
##  4 1415-1443 31386
##  5 1444-1471 31653
##  6 1472-1500 35729
##  7 1501-1528 32286
##  8 1529-1556 35344
##  9 1557-1585 35680
## 10 1586-1613 39966
## 11 1614-1641 39649
## 12 1642-1650 32206
```

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. (Include tables with the relevant info (Coefficients of interest, standard errors and sample size).)

Answer

This assumption is reasonable. While there may be some financial incentive to induce early births on behalf of the practicing obstetrician (in order to charge more in future services), there seems likely to be enough practical resistance to inducing early births around the cutoff. Especially when considering that attempting to fall below the cutoff only endangers the life of the baby even more. No parent, under this scenario, would attempt to induce an early birth. The cutoff of 1500 grams is also so low that no parent would reasonably attempt to have a baby born at so low a weight (3.3 lbs).

I have conducted a balance test between the two bins immediately above and below the 1500 gram threshold. Any statistically significant differences amongst the two would indicate that there was a likelihood of some manipulation around the boundary of the regression discontinuity.

```
databal <- data
databal$bins <- NULL
databal$bins <- as.numeric(cut(data$bweight, breaks = binbreaks[6:8], labels = 0:1))
databal <- drop_na(databal, bins)

baltest <- lm(data = databal,bins ~ bweight)
summary(baltest)
```

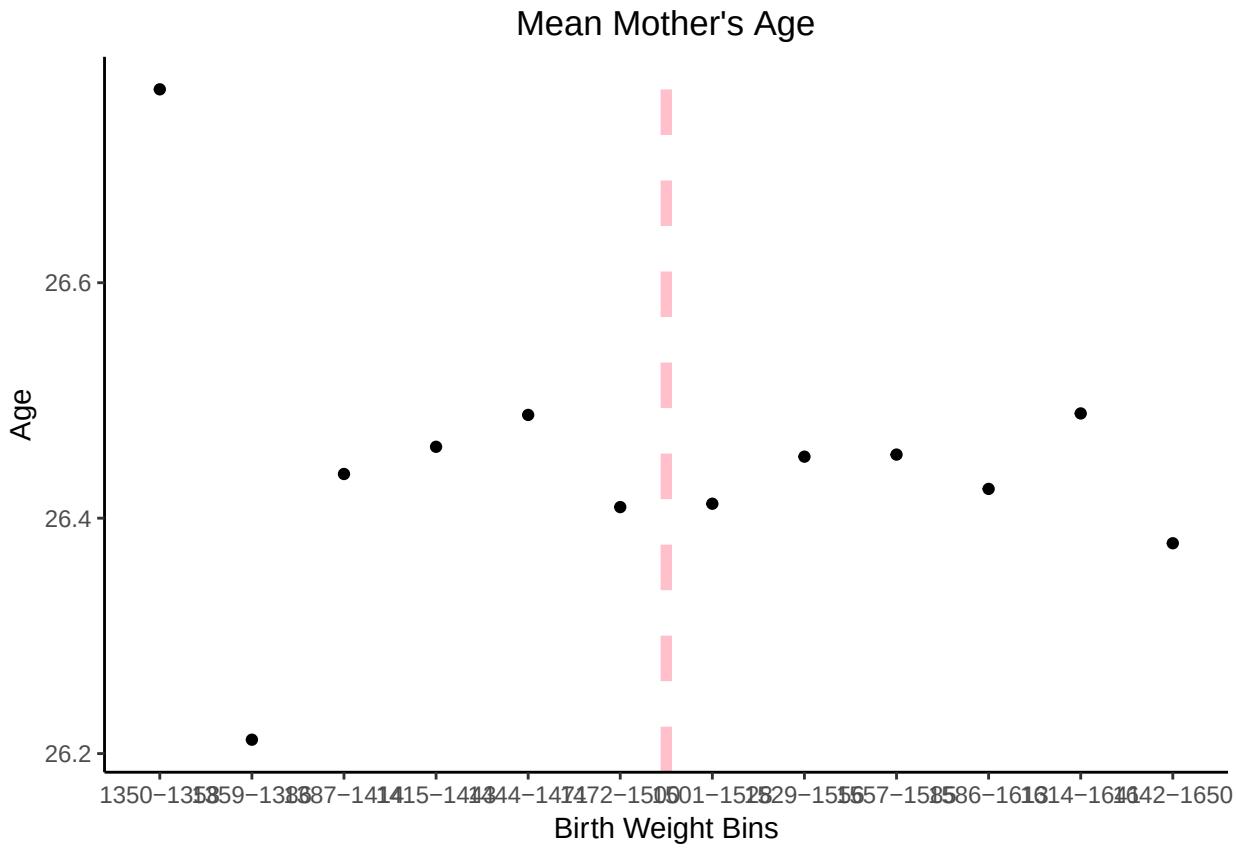
```
##
## Call:
## lm(formula = bins ~ bweight, data = databal)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68424 -0.09413  0.02622  0.23378  0.28843
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.930e+01  9.231e-02  -425.8   <2e-16 ***
## bweight      2.733e-02  6.185e-05   441.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2538 on 68013 degrees of freedom
## Multiple R-squared:  0.7416, Adjusted R-squared:  0.7416
## F-statistic: 1.952e+05 on 1 and 68013 DF,  p-value: < 2.2e-16
```

4

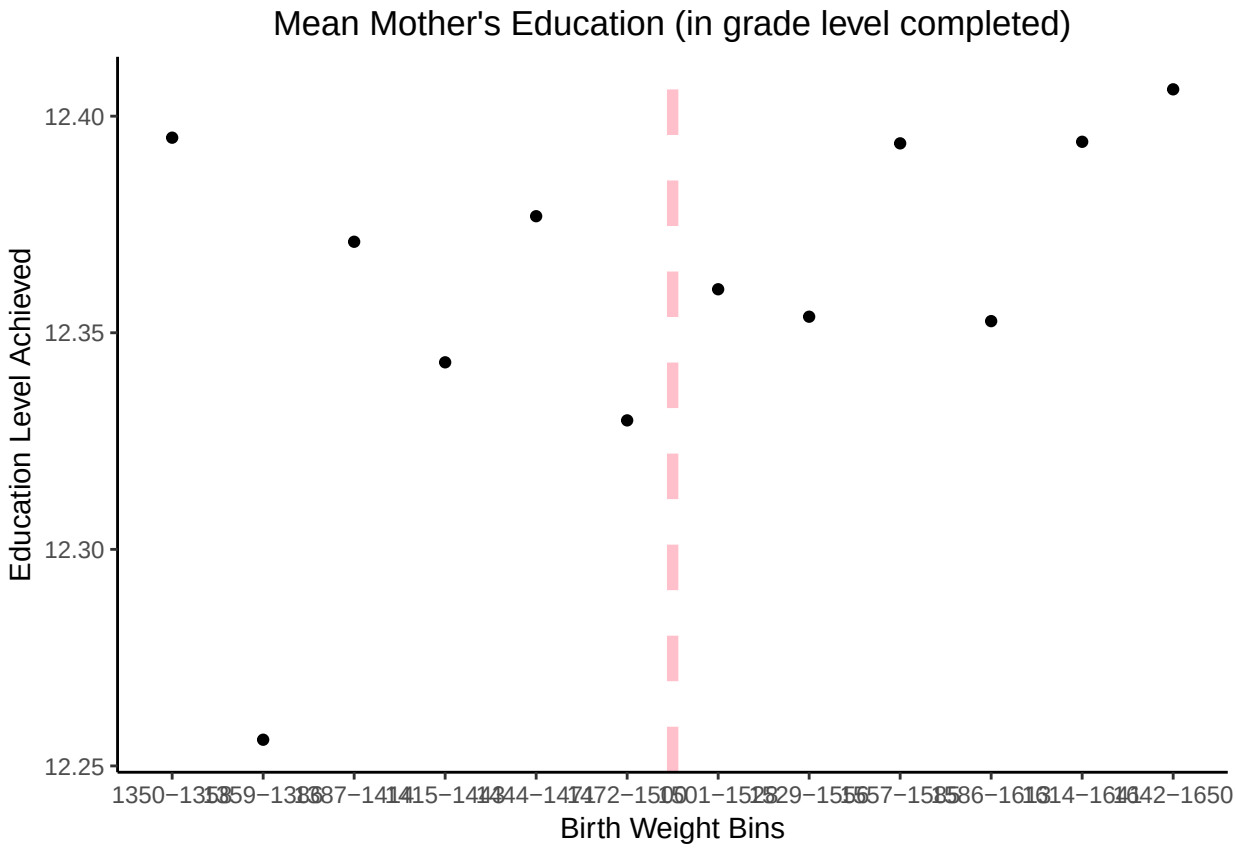
Assess informally whether the behavior of other covariates is smooth around the threshold, by plotting the mean of some covariates (mother's age, mother's education less than high school, gestational age, prenatal care visits, and year of birth) against birth weight as you did in point (2). Is there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

Answer

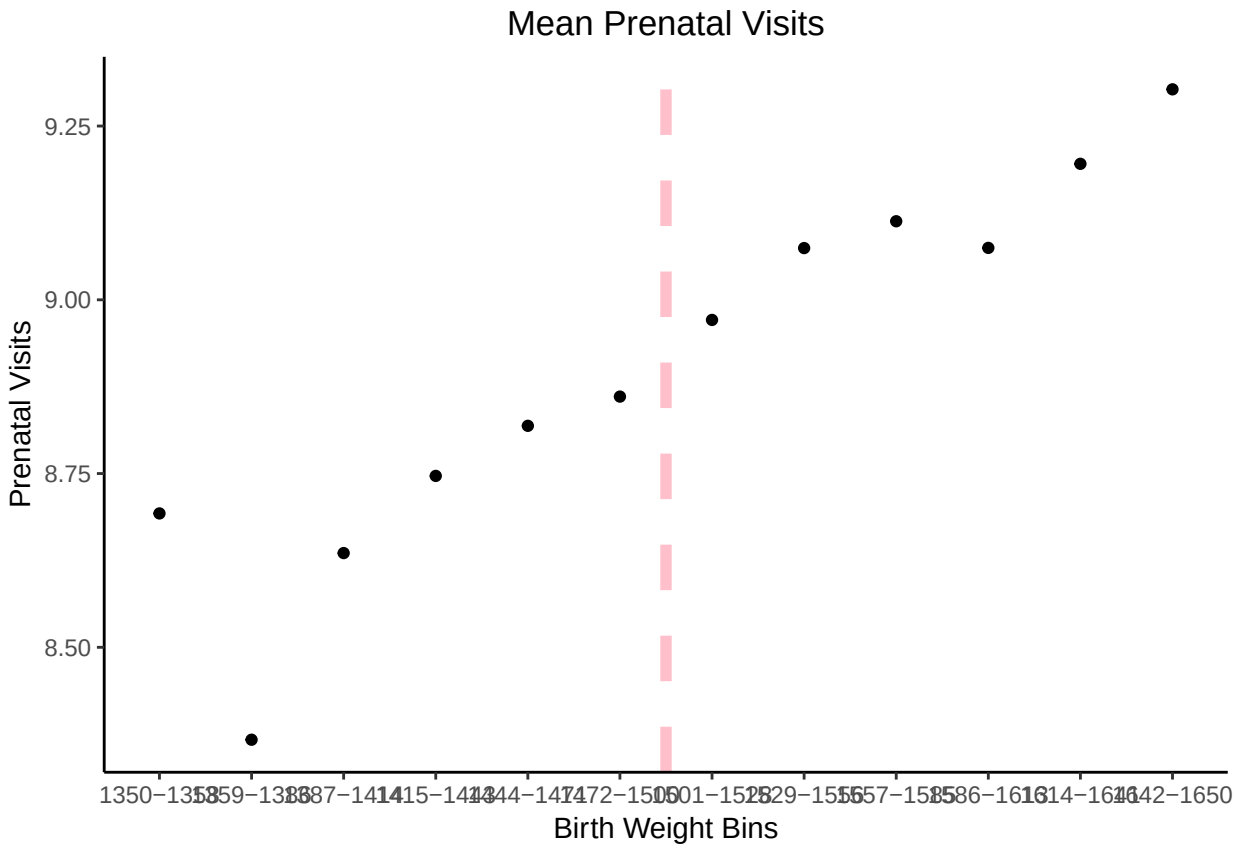
```
plot3 <- data %>% group_by(bins) %>% summarize(mean = mean(mom_age)) %>% ggplot(aes(bins,mean)) + geom_
plot3
```



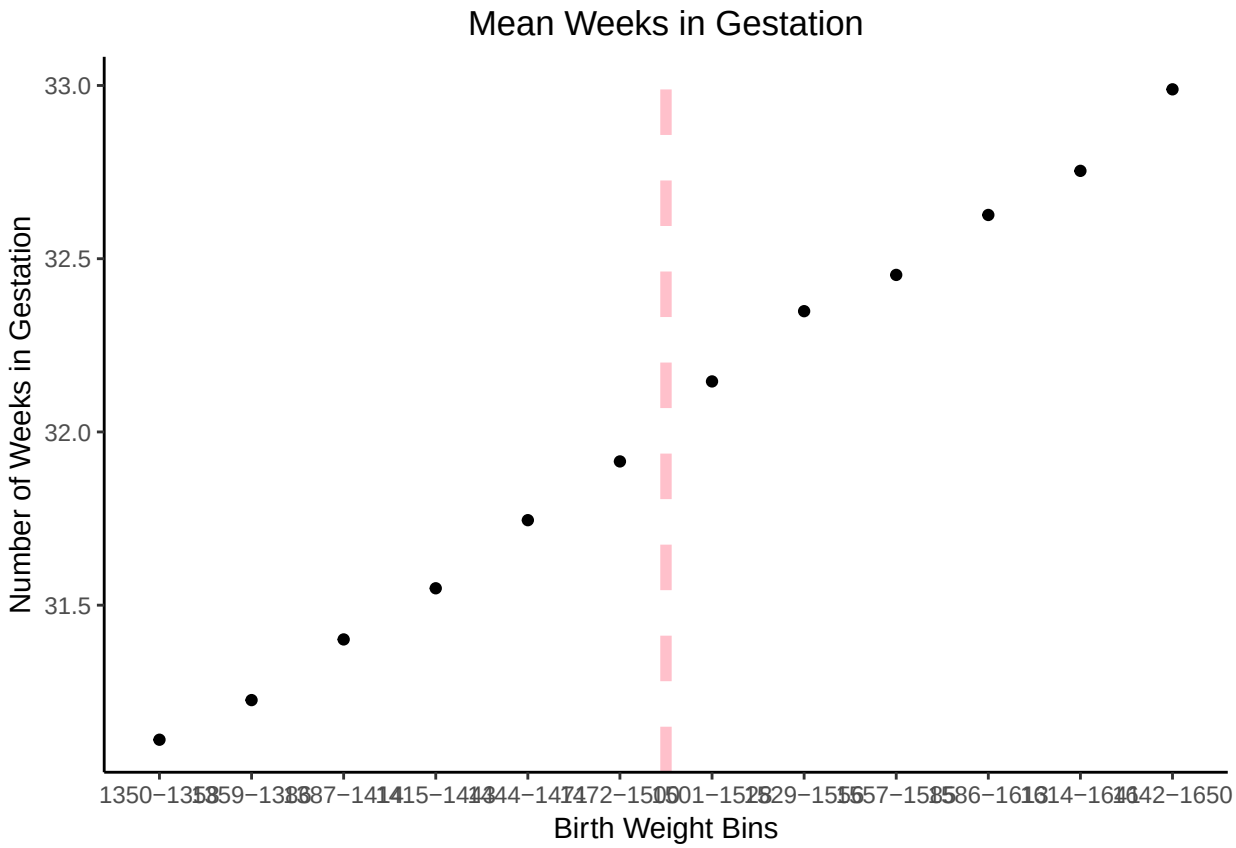

```
plot4 <- data %>% group_by(bins) %>% summarize(mean = mean(mom_ed, na.rm = TRUE)) %>% ggplot(aes(bins,m
plot4
```




```
plot5 <- data %>% group_by(bins) %>% summarize(mean = mean(nprenatal, na.rm = TRUE)) %>% ggplot(aes(bins
plot5
```




```
plot6 <- data %>% group_by(bins) %>% summarize(mean = mean(gest, na.rm = TRUE)) %>% ggplot(aes(bins,mea
plot6
```



There is evidence of discontinuities on other covariates (particularly number of prenatal visits). The problem this could likely introduce in our RDD estimates is that the supposition that our populations are similar around the threshold will not hold and thereby undermining the assumption necessary for us to estimate the RDD accurately.

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). To do so, use the following model:

$$Y_i = \alpha_0 + \alpha_1 VLBW_i + \alpha_2 VLBW_i * (g_i - 1500) + \alpha_3 * (1 - VLBW_i) * (g_i - 1500) + \epsilon_i$$

where Y_i is the outcome of interest, $VLBW_i$ indicates that a newborn had very low birth weight (<1500 grams), g_i is birth weight, and ϵ_i a disturbance term. Interpret the coefficients α_1, α_2, and α_3.

Answer

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.

Answer

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?

Answer

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.

Answer