Lab 4: Does Prenatal Care Improve Infant Health?

Chris Hipple, Phat Doan, Ted Pham

April 25, 2017

Introduction

Variable descriptions are provided as follows.

```
load("bwght_w203.RData")

suppressMessages(library(boot))
suppressMessages(library(car))
suppressMessages(library(gmodels))
suppressMessages(library(effsize))
suppressMessages(library(dplyr))
suppressMessages(library(sandwich))
suppressMessages(library(lmtest))
suppressMessages(library(stargazer))
```

1. A brief introduction & exploratory analysis

Get description and a quick snapshot of the dataset

```
colnames(data)
```

```
##  [1] "mage"    "meduc"   "monpre"  "npvis"   "fage"    "feduc"   "bwght"
##  [8] "omaps"   "fmaps"   "cigs"    "drink"   "lbw"     "vlbw"    "male"
## [15] "mwhte"   "mblck"   "moth"    "fwhte"   "fblck"   "foth"    "lbwght"
## [22] "magesq"  "npvissq"
```

Checking for null/na observations

```
sum(is.na(data))
```

```
## [1] 455
```

Remove null/na entirely and save to a new dataframe df_naomit

```
df_naomit = na.omit(data)
```

This data is from the National Center for Health Statistics and from birth certificates. Our team has been engaged to study the data and understand whether prenatal care improves health outcomes fo r newborn infants. There are 3 potential outcomes variables in this dataset: birthweight, and one an d five-minute APGAR scores. These are measures of the well-being of infants just after birth. The da taset also contains multiple variables that can be used as predictors such parents' demographic, pre natal care during pregnancy, etc.

2. A model building process

Outcome variables:

We identify 3 potential outcomes variables in this dataset:

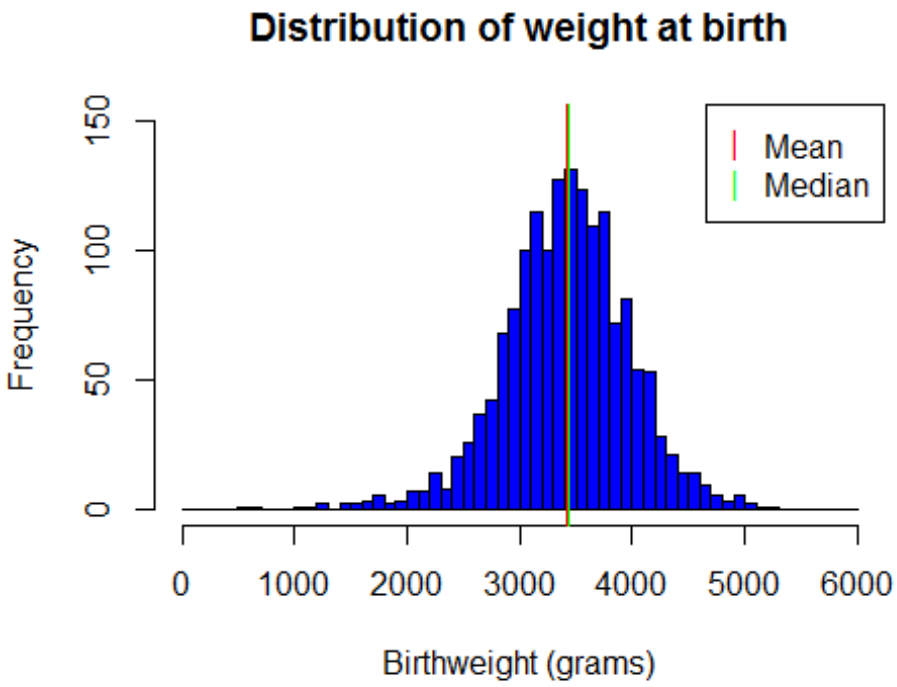
1/ Birthweight

2/ 1-minute APGAR score

3/ 5-minute APGAR score

Exploratory analysis of birthweight

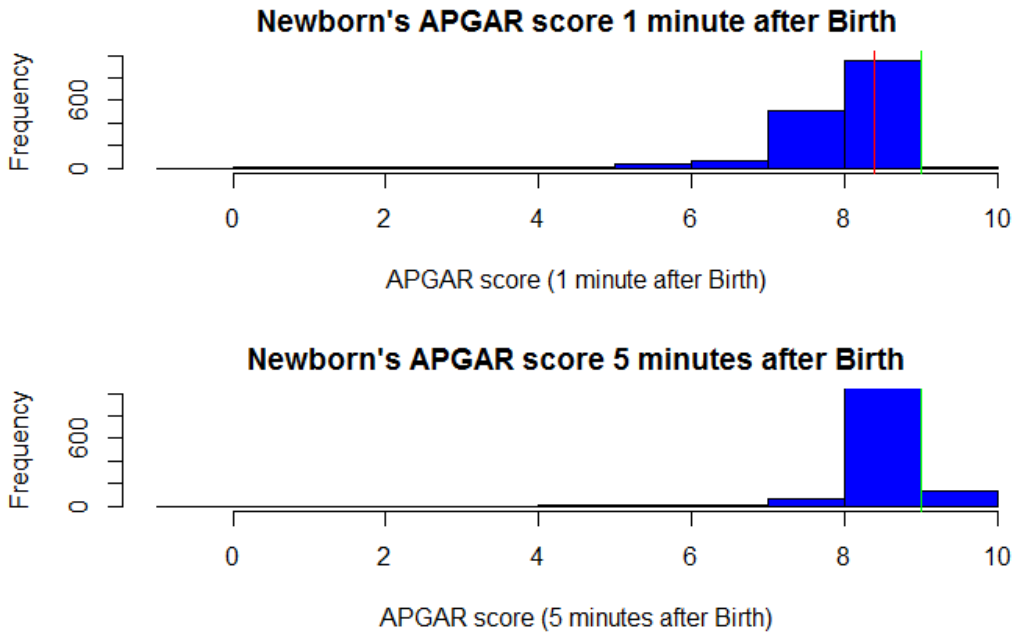
```
# descriptive stat
summary(df_naomit$bwght)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     506    3090    3430    3415    3771    5204

# histogram
hist(df_naomit$bwght, main = "Distribution of weight at birth"
               , xlab = "Birthweight (grams)"
               , breaks = seq(0, 6000, by = 100)
               , col = "blue", ylim = c(0,150))
abline(v = mean(df_naomit$bwght), col = "red")
abline(v = median(df_naomit$bwght), col = "green")

legend("topright", c("Mean", "Median"), col = c("red", "green"), pch = "|")
```



Birthweight variable shows a normal distribution. **2 outliers** with less than 1000 grams at birth. This will be investigated further.

Exploratory analysis of apgar score at 1 minute

```
# Newborn's APGAR score 1 minute after Birth
summary(df_naomit$omaps)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    8.00    9.00    8.39    9.00   10.00

# Newborn's APGAR score 5 minutes after Birth
summary(df_naomit$fmaps)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.000   9.000   9.000   9.015   9.000  10.000

# histogram
par(mfrow = c(2,1))
hist(df_naomit$omaps, main = "Newborn's APGAR score 1 minute after Birth"
               , xlab = "APGAR score (1 minute after Birth)"
               , breaks = seq(-1, 10, by = 1)
               , col = "blue", ylim = c(0,1000))
abline(v = mean(df_naomit$omaps), col = "red")
abline(v = median(df_naomit$omaps), col = "green")

hist(df_naomit$fmaps, main = "Newborn's APGAR score 5 minutes after Birth"
               , xlab = "APGAR score (5 minutes after Birth)"
               , breaks = seq(-1, 10, by = 1)
               , col = "blue", ylim = c(0,1000))
abline(v = mean(df_naomit$fmaps), col = "red")
abline(v = median(df_naomit$fmaps), col = "green")
```



APGAR Score 1 Minute after Birth is **heavily skewed right** but the **majority at 8 and 9**. **APGAR Score 5** Minute after Birth is **heavily skewed right** but the **majority at 9**. This shows a **improvement for most babies between 1 minute and 5 minutes after birth**

Exploratory analysis of Newborn weight

```
CrossTable(df_naomit$lbw, df_naomit$vlbw, format = "SPSS"
           , prop.c = FALSE, prop.r = FALSE, prop.t = TRUE
           , prop.chisq = FALSE
           , dnn = c("Low Birth Weight (<2000g)", "Very Low Birth Weight (<1500g)"))
```

```
##
##    Cell Contents
## |-------------------------|
## |                   Count |
## |           Total Percent |
## |-------------------------|
##
## Total Observations in Table:  1612
##
##                             | Very Low Birth Weight (<1500g)
## Low Birth Weight (<2000g) |        0  |        1  | Row Total |
## --------------------------|-----------|-----------|-----------|
##                         0 |     1589  |        0  |     1589  |
##                           |   98.573% |    0.000% |           |
## --------------------------|-----------|-----------|-----------|
##                         1 |       15  |        8  |       23  |
##                           |    0.931% |    0.496% |           |
## --------------------------|-----------|-----------|-----------|
##              Column Total |     1604  |        8  |     1612  |
## --------------------------|-----------|-----------|-----------|
##
##
```

Only **1.3% of the newborn** from the dataset was considered to **have low birthweight or very low birthweight**

For this exercise, we choose **birthweight as the outcome**. This variable long has been supported by the medical community as an indicator of baby's health outcome.

Predictor variables:

We breakdown predictor variables into 4 categories:

1/ Demographics: age, race, and education

2/ Prenatal care: month prenatal care began, total number of prenatal visits

3/ Mother's behaviors: # of cigarettes per day, # of drink per day

4/ Newborn gender

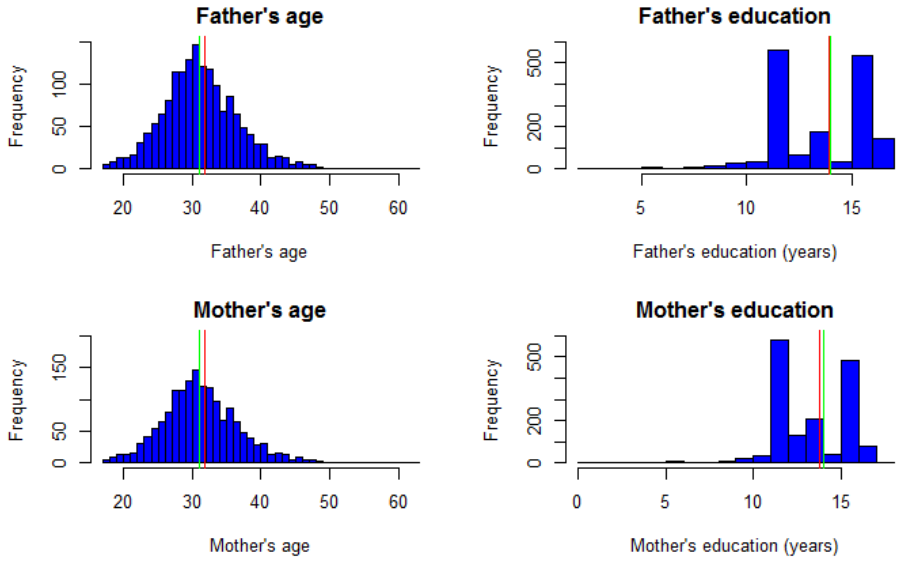
Exploratory analysis of Age & Education

```
# Father's age
summary(df_naomit$fage)
```

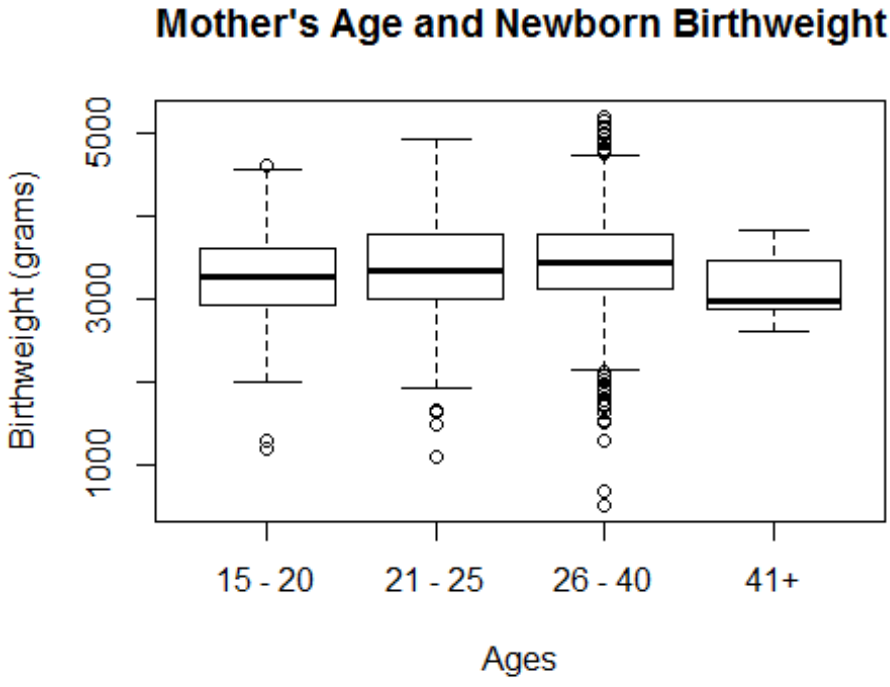
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   28.00   31.00   31.79   35.00   62.00
```

```
# Father's education
summary(df_naomit$feduc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00   12.00   14.00   13.91   16.00   17.00

# Mother's age
summary(df_naomit$mage)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   16.00   26.00   29.00   29.48   32.00   44.00

# Mother's education
summary(df_naomit$meduc)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00   12.00   14.00   13.74   16.00   17.00

# histogram
par(mfrow =c(2,2))
hist(df_naomit$fage, main = "Father's age"
               , xlab = "Father's age"
               , breaks = seq(17, 63, by = 1)
               , col = "blue", ylim = c(0,150))
abline(v = mean(df_naomit$fage), col = "red")
abline(v = median(df_naomit$fage), col = "green")

hist(df_naomit$feduc, main = "Father's education"
               , xlab = "Father's education (years)"
               , breaks = seq(2, 17,by=1)
               , col = "blue", ylim = c(0,600))
abline(v = mean(df_naomit$feduc), col = "red")
abline(v = median(df_naomit$feduc), col = "green")

hist(df_naomit$fage, main = "Mother's age"
               , xlab = "Mother's age"
               , breaks = seq(17, 63, by = 1)
               , col = "blue", ylim = c(0,200))
abline(v = mean(df_naomit$fage), col = "red")
abline(v = median(df_naomit$fage), col = "green")

hist(df_naomit$meduc, main = "Mother's education"
               , xlab = "Mother's education (years)"
               , breaks = seq(0, 18,by=1)
               , col = "blue", ylim = c(0,600))
abline(v = mean(df_naomit$meduc), col = "red")
abline(v = median(df_naomit$meduc), col = "green")
```




```
# Cutting Mother's Age
df_naomit$mage_cut <- cut(df_naomit$mage, c(-Inf, 20, 25, 40, Inf))
table(df_naomit$mage_cut)

##
## (-Inf,20]   (20,25]   (25,40] (40, Inf]
##        48       268      1285        11

boxplot(df_naomit$bwght ~ df_naomit$mage_cut
        , main = "Mother's Age and Newborn Birthweight"
        , names = c("15 - 20", "21 - 25", "26 - 40", "41+")
        , xlab = "Age"
        , ylab = "Birthweight (grams)")
```

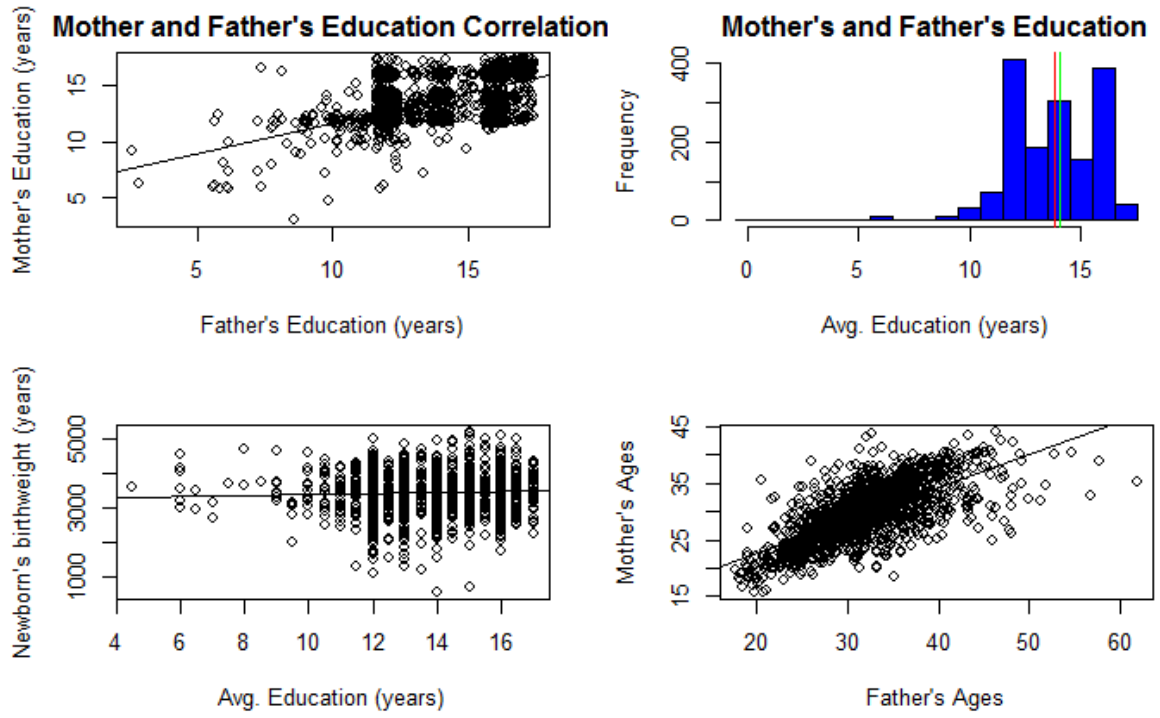


Histogram shows a normal distribution with no outliers (i.e. too young)

Mother's education variable is **highly skewed to the right** that can break down into 4 groups: **some HS, HS diploma, some college, and college degree**. Validity test shows no invalid observation for the variable.

```r
bracket <- c(0,11.5,12.5,15.5, Inf)
labels = c("some HS", "HS", "some college", "college")
c1 <- cut(df_naomit$meduc, breaks = bracket)
# table(c1)
levels(c1) <- labels
#table(c1)
df_naomit$meduc_level <- c1

by_educ <- group_by(df_naomit, meduc_level)
meduc_influence <- summarise(by_educ,
                             avg_bwght = mean(bwght),
                             n = n(),
                             avg_npvis = mean(npvis),
                             avg_omaps = mean(omaps),
                             avg_fmaps = mean(fmaps),
                             avg_meduc = mean(meduc),
                             avg_feduc = mean(feduc))

par(mfrow=c(2,2))
plot(x = jitter(df_naomit$feduc, 2), y = jitter(df_naomit$meduc, 2),
     main = "Mother and Father's Education Correlation",
     xlab = "Father's Education (years)",
     ylab = "Mother's Education (years)")

abline(lm(meduc ~ feduc, data = df_naomit))

print("Correlation between Father's and Mother's Education")

## [1] "Correlation between Father's and Mother's Education"

cor(df_naomit$feduc, df_naomit$meduc)

## [1] 0.5944859

# Create a variable of average education for parents
df_naomit<- df_naomit %>% rowwise() %>% mutate(avg_educ = mean(c(meduc, feduc)))

#Do the same for age
df_naomit<- df_naomit %>% rowwise() %>% mutate(avg_age = mean(c(mage, fage)))

hist(df_naomit$avg_educ
     , breaks = seq(-0.5, 18, 1)
     , col = "blue"
     , main = "Avg. education between Mother and Father"
     , xlab = "Avg. Education (years)")
abline(v = mean(df_naomit$avg_educ), col = "red")
abline(v = median(df_naomit$avg_educ), col = "green")

edu_model = lm(bwght ~ avg_educ, data = df_naomit)

plot(x = df_naomit$avg_educ, y = df_naomit$bwght
    , main = "Avg. education between Mother and Father"
    , xlab = "Avg. Education (years)"
    , ylab = "Newborn's birthweight (years)")
```

```
abline(edu_model)

plot(x = jitter(df_naomit$fage, 2), y = jitter(df_naomit$mage, 2),
     main = "Mother and Father's Age Correlation",
     xlab = "Father's Age",
     ylab = "Mother's Age")
abline(lm(mage ~ fage, data = df_naomit))
```




```
# Correlation between Father's and Mother's Age
cor(df_naomit$mage, df_naomit$fage)

## [1] 0.689369
```

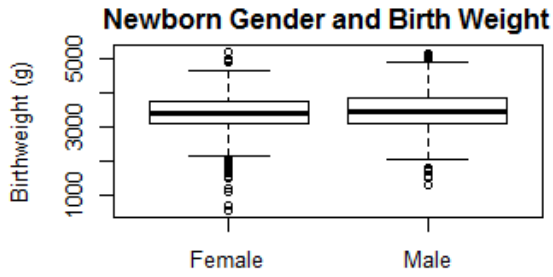
High correlation between Mother's and Father's age that could lead to multi-collinearity between the two variables.

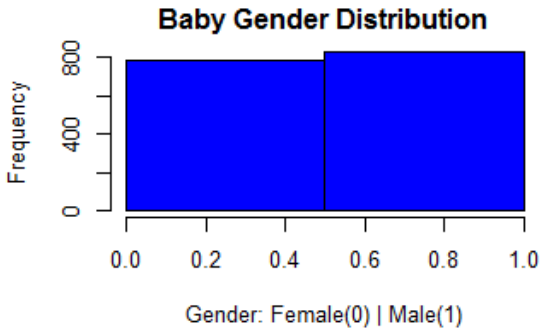
Exploratory analysis of Newborn gender

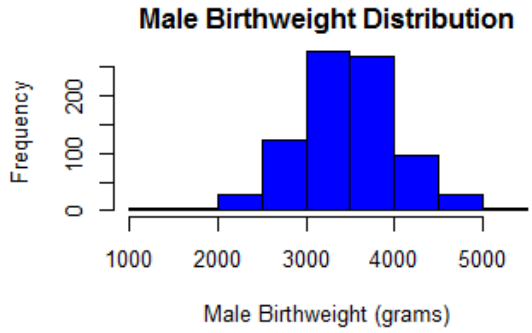
There is a close split between male and female newborn in the dataset

```
par(mfrow=c(2,2))
boxplot(bwght ~ male, data = df_naomit, main = "Newborn Gender and Birth Weight"
        , names = c("Female","Male")
        , ylab = "Birthweight (g)")

# conduct 2-sample t-test to see if there is a statistical difference between male an
d female birthweight
male_bwght = sort(df_naomit$bwght[which(df_naomit$male == 1)])

female_bwght = sort(df_naomit$bwght[which(df_naomit$male == 0)])

hist(df_naomit$male
     , breaks = seq(0, 1, 0.5)
```

```
     , xlab = "Gender: Female(0) | Male(1)"
     , col = "blue"
     , main = "Baby Gender Distribution")

hist(male_bwght
    , col = "blue"
    , main = "Male Birthweight Distribution"
    , xlab = "Male Birthweight (grams)")
abline(v = mean(df_naomit$male_bwght), col = "red")

## Warning: Unknown or uninitialised column: 'male_bwght'.

## Warning in mean.default(df_naomit$male_bwght): argument is not numeric or
## logical: returning NA

abline(v = median(df_naomit$male_bwght), col = "green")

## Warning: Unknown or uninitialised column: 'male_bwght'.

## Warning in is.na(x): is.na() applied to non-(list or vector) of type 'NULL'

hist(female_bwght
    , col = "blue"
    , main = "Female Birthweight Distribution"
    , xlab = "Female Birthweight (grams)")
abline(v = mean(df_naomit$female_bwght), col = "red")

## Warning: Unknown or uninitialised column: 'female_bwght'.

## Warning in mean.default(df_naomit$female_bwght): argument is not numeric or
## logical: returning NA

abline(v = median(df_naomit$female_bwght), col = "green")

## Warning: Unknown or uninitialised column: 'female_bwght'.

## Warning in is.na(x): is.na() applied to non-(list or vector) of type 'NULL'
```








```
# variance test to verify homoskestacity
var.test(male_bwght, female_bwght)

##
##  F test to compare two variances
##
## data:  male_bwght and female_bwght
## F = 1.0219, num df = 827, denom df = 783, p-value = 0.7595
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8897628 1.1732969
## sample estimates:
## ratio of variances
##           1.021877

# 2-sample t-test
t.test(male_bwght, female_bwght)

##
##  Welch Two Sample t-test
##
## data:  male_bwght and female_bwght
## t = 3.1239, df = 1606.9, p-value = 0.001816
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   32.47702 142.07362
## sample estimates:
## mean of x mean of y
##  3456.981  3369.705
```

From visual graph, there is no significant difference between Male and Female birthweight. However, we conducted the 2-sample t-test to verify the statistical significance. First, the variance

test has a p-value > 0.05, therefore we failed to reject the null hypothesis and **variance is homogeneous**. The **2-sample t-test has p-value = 0.001**, thus we reject null hypothesis, meaning there is a **statistically significant difference between the mean of male and female birthweight**

Exploratory analysis of Mother's race

```
CrossTable(df_naomit$mwhte, df_naomit$mblck, format = "SPSS"
           , prop.c = FALSE, prop.r = FALSE, prop.t = TRUE
           , prop.chisq = FALSE
           , dnn = c("Mother is White","Mother is Black"))

##
##    Cell Contents
## |-------------------------|
## |                   Count |
## |           Total Percent |
## |-------------------------|
##
## Total Observations in Table:  1612
##
##                  | Mother is Black
## Mother is White |         0  |         1  | Row Total |
## ----------------|------------|------------|-----------|
##               0 |        93  |        88  |       181 |
##                 |     5.769% |     5.459% |           |
## ----------------|------------|------------|-----------|
##               1 |      1431  |         0  |      1431 |
##                 |    88.772% |     0.000% |           |
## ----------------|------------|------------|-----------|
##    Column Total |      1524  |        88  |      1612 |
## ----------------|------------|------------|-----------|
##
##
```

Heavily skewed white with only ~5% black and ~5% other

Exploratory analysis of Father's race

```
CrossTable(df_naomit$fwhte, df_naomit$fblck, format = "SPSS"
           , prop.c = FALSE, prop.r = FALSE, prop.t = TRUE
           , prop.chisq = FALSE
           , dnn = c("Father is White","Father is Black"))

##
##    Cell Contents
## |-------------------------|
## |                   Count |
## |           Total Percent |
## |-------------------------|
##
## Total Observations in Table:  1612
##
##                  | Father is Black
## Father is White |         0  |         1  | Row Total |
## ----------------|------------|------------|-----------|
##               0 |        79  |        92  |       171 |
##                 |     4.901% |     5.707% |           |
## ----------------|------------|------------|-----------|
```

```
##               1 |     1441  |        0  |     1441  |
##                 |   89.392% |    0.000% |           |
## ----------------|-----------|-----------|-----------|
##    Column Total |     1520  |       92  |     1612  |
## ----------------|-----------|-----------|-----------|
##
##
```

Similar to **Mother's race variable**, **father's race is heavily skewed white with only ~5% black and ~5% other**

combine race into white or others

```
CrossTable(df_naomit$fwhte, df_naomit$mwhte, format = "SPSS"
          , prop.c = FALSE, prop.r = FALSE, prop.t = T
          , prop.chisq = FALSE
          , dnn = c("Father is White","Mother is White"))
```

```
##
##    Cell Contents
## |-------------------------|
## |                   Count |
## |           Total Percent |
## |-------------------------|
##
## Total Observations in Table:  1612
##
##                  | Mother is White
## Father is White |        0  |        1  | Row Total |
## ----------------|-----------|-----------|-----------|
##               0 |      160  |       11  |      171  |
##                 |    9.926% |    0.682% |           |
## ----------------|-----------|-----------|-----------|
##               1 |       21  |     1420  |     1441  |
##                 |    1.303% |   88.089% |           |
## ----------------|-----------|-----------|-----------|
##    Column Total |      181  |     1431  |     1612  |
## ----------------|-----------|-----------|-----------|
##
##
```

```
cor(df_naomit$fwhte, df_naomit$mwhte)
```

```
## [1] 0.8984158
```

Applying Transformations

```
males = df_naomit[df_naomit$male == 1,]
females = df_naomit[df_naomit$male == 0,]
t.test(males$bwght, females$bwght, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  males$bwght and females$bwght
## t = 3.1239, df = 1606.9, p-value = 0.0009082
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  41.29526      Inf
## sample estimates:
```

```
## mean of x mean of y
##  3456.981  3369.705

df_naomit <- df_naomit %>% mutate(whte = mwhte & fwhte)
table(df_naomit$whte)

##
## FALSE  TRUE
##   192  1420

t.test(df_naomit$bwght[df_naomit$whte], df_naomit$bwght[!df_naomit$whte])

##
##  Welch Two Sample t-test
##
## data:  df_naomit$bwght[df_naomit$whte] and df_naomit$bwght[!df_naomit$whte]
## t = 2.2019, df = 258.47, p-value = 0.02855
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##    9.310805 166.866307
## sample estimates:
## mean of x mean of y
##  3425.026  3336.938
```

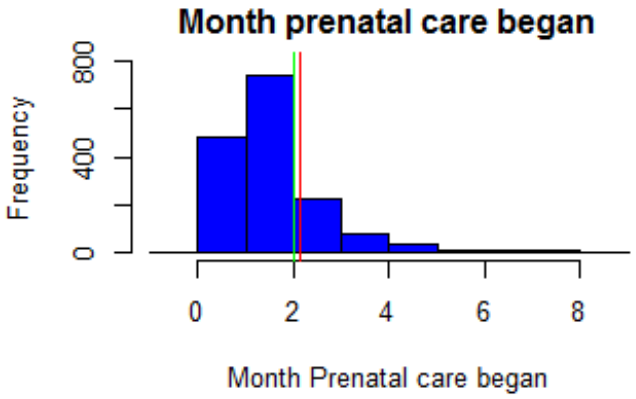
There is a **statistically significant difference between the birth weights of babies with two-white parents vs babies which do not have two white parents**.

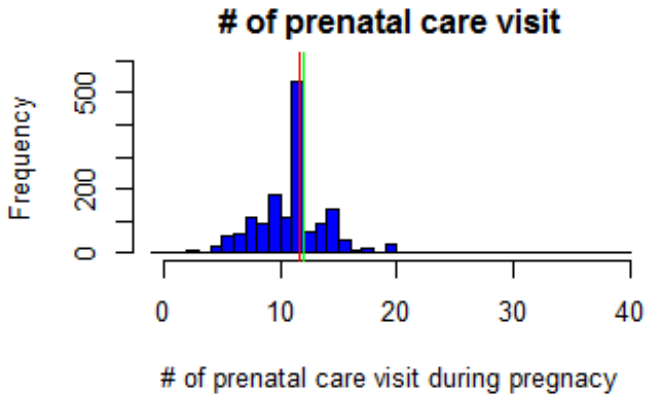
Exploratory analysis of month prenatal care began

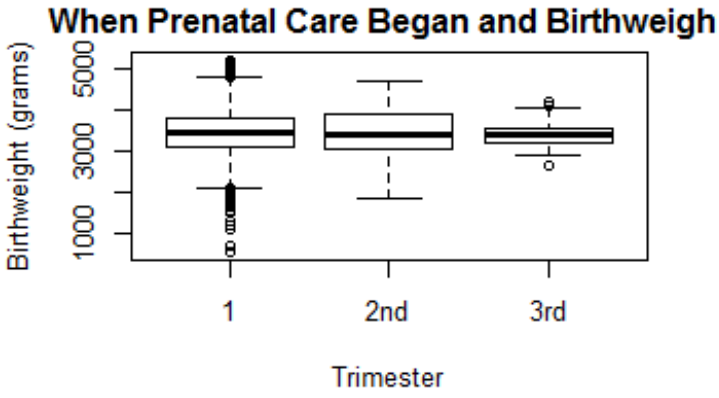
Cutting monpre variable

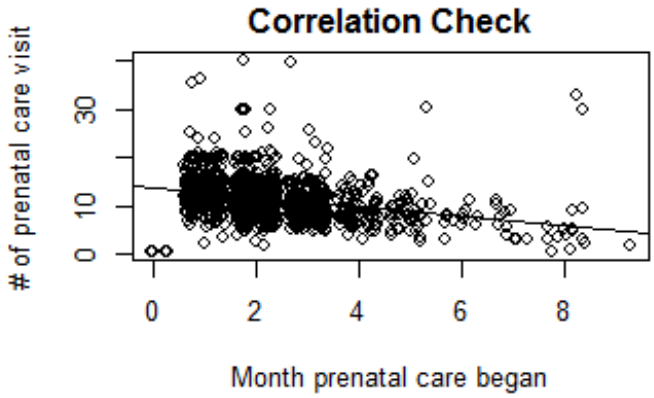
```
table(df_naomit$monpre)

##
##   0   1   2   3   4   5   6   7   8   9
##   4 479 739 226  84  41  12  11  15   1

# Month prenatal care began
summary(df_naomit$monpre)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   2.000   2.143   2.000   9.000

# Number of Prenatal Care Visits
summary(df_naomit$npvis)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   12.00   11.62   13.00   40.00

# histogram
par(mfrow = c(2,2))

hist(df_naomit$monpre, main = "Month prenatal care began"
               , xlab = "Month Prenatal care began"
               , breaks = seq(-1, 9, by = 1)
               , col = "blue", ylim = c(0,800))
abline(v = mean(df_naomit$monpre), col = "red")
abline(v = median(df_naomit$monpre), col = "green")

hist(df_naomit$npvis, main = "# of prenatal care visit"
```

```r
               , xlab = "# of prenatal care visit during pregnacy"
               , breaks = seq(-1, 40, by = 1)
               , col = "blue", ylim = c(0,600))
abline(v = mean(df_naomit$npvis), col = "red")
abline(v = median(df_naomit$npvis), col = "green")

df_naomit$monpre_cut <- cut(df_naomit$monpre, c(-Inf, 3, 6, Inf))

boxplot(df_naomit$bwght ~ df_naomit$monpre_cut
        , main = "When Prenatal Care Began and Birthweight"
        , names = c("1", "2nd", "3rd")
        , xlab = "Trimester"
        , ylab = "Birthweight (grams)")

# checking for multicollinerarity between when prenatal care started and numbers of p
renatal care visit
plot(x = jitter(df_naomit$monpre, 2), y = jitter(df_naomit$npvis, 2),
     main = "Correlation Check",
     xlab = "Month prenatal care began",
     ylab = "# of prenatal care visit")
abline(lm(npvis ~ monpre, data = df_naomit))
```









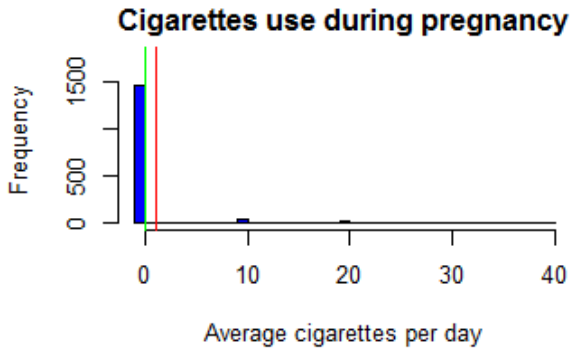

```r
# Correlation between when prenatal care started and numbers of prenatal care visit
cor(df_naomit$monpre, df_naomit$npvis)

## [1] -0.3134265
```

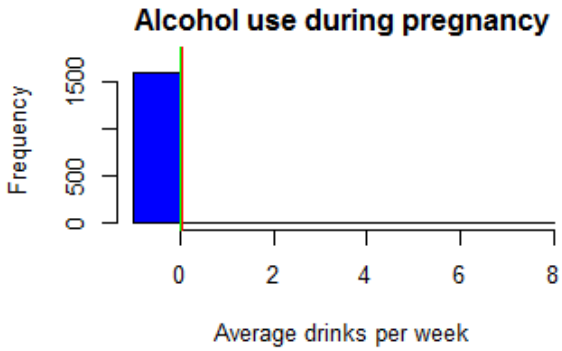
Variable is heavily skewed to the left with **majority of mothers start prenatal care at 2 months** (first tri-semester). No outlier was observed.

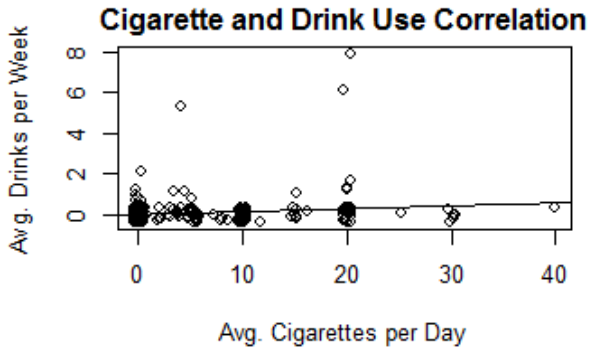
Normal distribution with most mothers have **12 prenatal care visits during their pregnancy**. There are a few outliers with greater than 20 visits. This might be an indicator of high risk pregnancy

Exploratory analysis of cigarette and alcohol use during pregnancy

```
# Cigarette Use During Pregnancy
summary(df_naomit$cigs)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   1.057   0.000  40.000

# Drink Use During Pregnancy
summary(df_naomit$drink)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.00000 0.00000 0.02109 0.00000 8.00000

# histogram
par(mfrow=c(2,2))
hist(df_naomit$cigs, main = "Cigarettes use during pregnancy"
                , xlab = "Average cigarettes per day"
                , breaks = seq(-1, 40, by = 1)
                , col = "blue", ylim = c(0,1800))
abline(v = mean(df_naomit$cigs), col = "red")
abline(v = median(df_naomit$cigs), col = "green")

# histogram
hist(df_naomit$drink, main = "Alcohol use during pregnancy"
                , xlab = "Average drinks per week"
                , breaks = seq(-1, 8, by = 1)
                , col = "blue", ylim = c(0,1800))
abline(v = mean(df_naomit$drink), col = "red")
abline(v = median(df_naomit$drink), col = "green")

# checking for multicollinerarity between cigarett use started and number of drinks p
er day
plot(x = jitter(df_naomit$cigs, 2), y = jitter(df_naomit$drink, 2),
     main = "Cigarette and Drink Use Correlation",
     xlab = "Avg. Cigarettes per Day",
     ylab = "Avg. Drinks per Week")

abline(lm(drink ~ cigs, data = df_naomit))

print("Correlation between Cigarette and Drink Use")

## [1] "Correlation between Cigarette and Drink Use"

cor(df_naomit$cigs, df_naomit$drink)

## [1] 0.1899853
```







Cutting into Smoker's and Non Smokers

```
df_naomit$is_smoker = cut(df_naomit$cigs, c(-1, 0.5, Inf), labels = c(0, 1))
t.test(df_naomit$bwght[df_naomit$is_smoker == 0], df_naomit$bwght[df_naomit$is_smoker
== 1])

##
##  Welch Two Sample t-test
##
## data:  df_naomit$bwght[df_naomit$is_smoker == 0] and df_naomit$bwght[df_naomit$is_
smoker == 1]
## t = 4.2614, df = 165.41, p-value = 3.402e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  111.1656 303.1068
## sample estimates:
## mean of x mean of y
##  3432.267  3225.130

mean_cigs <- mean(df_naomit$cigs[df_naomit$is_smoker == 1])
```

There is a highly statistically significant difference between the birthweight of babies whose mothers smoke vs those whose mothers did not smoke.

However, there are many more mothers in the sample who do not smoke than those who do.

Cutting into Drinkers and Non Drinkers

The drink variable was very highly skewed, with almost no mother's drinking, and some drinking a couple of drinks per day. With the skew as is, we can't meaningfully analyze the effect drinking has on birthweight, but due to prevailing wisdom of drinking during pregnancy being bad, we chose to look to see if we could cut the group into drinkers and non-drinkers and see if there was a meaningful difference.

```
df_naomit$is_drinker= cut(df_naomit$drink, c(-1, 0.5, Inf), labels = c(0, 1))
t.test(df_naomit$bwght[df_naomit$is_drinker == 0], df_naomit$bwght[df_naomit$is_drink
er == 1])

##
##  Welch Two Sample t-test
##
## data:  df_naomit$bwght[df_naomit$is_drinker == 0] and df_naomit$bwght[df_naomit$is
_drinker == 1]
## t = 0.5445, df = 15.316, p-value = 0.5939
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -219.0632  369.7521
## sample estimates:
## mean of x mean of y
##  3415.282  3339.938

table(df_naomit$is_drinker)

##
##    0    1
## 1596   16
```

With this dataset, there are almost 1600 mothers who do not drink and only 16 who do, soeven though there is a difference of about 80 grams of birthweight for the baby between the two groups, we fail to reject the null hypothesis that there is a difference between the two groups.

3 & 4. Model1 Building Process and Checking all Assumptions

For the model building process, we're going to start with looking directly at the proposed question, which is does prenatal care improve infant birth health. For this analysis, we're going to use weight at birth as our measure of infant health, with a heavier baby being healthier. This is a farily common accepted metric. Other possible metrics included in the dataset are the APGAR scores for the baby at 1 and 5 minutes. These ordinal values have potential to be a good metric, but our dataset does not have a wide variety of APGAR scores, with most scores being in the 7-9 range.

The first model we'll look at is birthweight as a function of just number of visits to a prenatal physician and which month the family began prenatal care.

Model 1

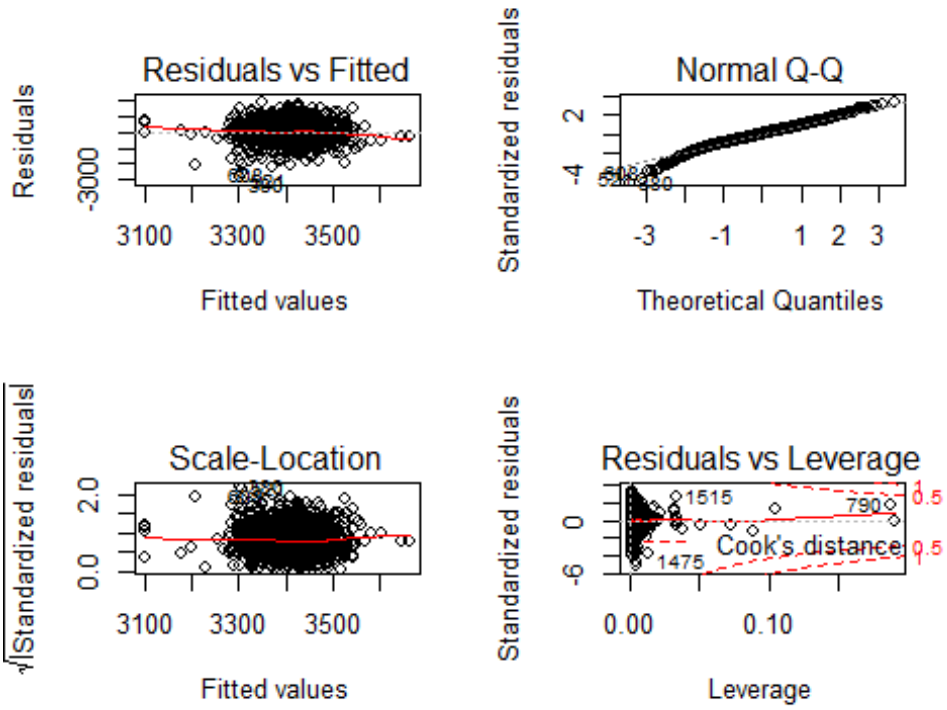
- One model with only the explanatory variables of key interest: birthweight, npvis, monpre, cigs, and drink

```
model1_v1 <- lm(bwght ~ npvis + monpre + cigs + drink, data = df_naomit)

model1_v2 <- lm(bwght ~ npvis + cigs + drink, data = df_naomit)

model1_v3 <- lm(bwght ~ npvis + cigs, data = df_naomit)
```

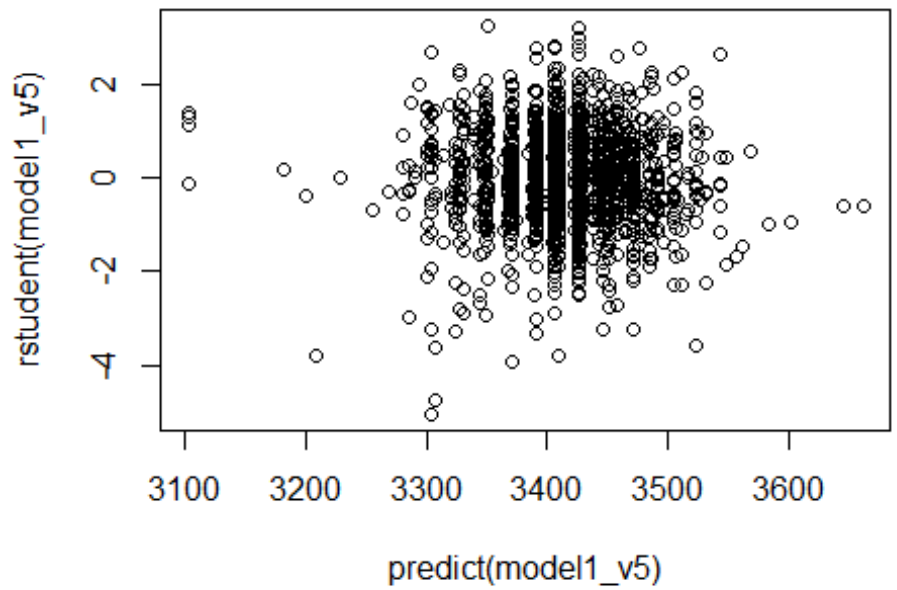
```r
model1_v4 <- lm(bwght ~ npvis + monpre, data = df_naomit)

model1_v5 <- lm(bwght ~ monpre + npvis + npvissq, data = df_naomit)

stargazer(model1_v1, model1_v2, model1_v3, model1_v4, model1_v5, type = "text"
          , star.cutoffs = c(0.05, 0.01, 0.005)
          , title = "Model 1: Explantory Variables of Key Interst"
          , align = T, no.space = T
          , omit.stat = c("ser", "f"))
```

```
##
## Model 1: Explantory Variables of Key Interst
## ================================================================================
##                                      Dependent variable:
##              ----------------------------------------------------------------
##                                            bwght
##                  (1)          (2)          (3)          (4)          (5)
## --------------------------------------------------------------------------------
## npvis         13.094***    11.450***    11.331***    13.374***     30.570**
##                (3.955)      (3.763)      (3.754)      (3.955)      (11.536)
## npvissq                                                             -0.573
##                                                                    (0.361)
## monpre          16.089                                13.251       19.816
##                (11.941)                              (11.931)     (12.622)
## cigs          -10.159***    -9.783**    -10.109***
##                (3.512)      (3.502)      (3.435)
## drink          -22.404      -22.947
##                (47.704)     (47.715)
## Constant     3,239.109*** 3,292.307*** 3,293.553*** 3,230.727*** 3,102.165***
##                (60.784)     (46.227)     (46.144)     (60.804)    (101.284)
## --------------------------------------------------------------------------------
## Observations    1,612        1,612        1,612        1,612        1,612
## R2              0.013        0.012        0.012        0.007        0.009
## Adjusted R2     0.010        0.010        0.010        0.006        0.007
## ================================================================================
## Note:                                          *p<0.05; **p<0.01; ***p<0.005
```

```r
par(mfrow=c(2,2))
plot(model1_v5)
```




```
par(mfrow=c(1,1))
plot(predict(model1_v5),rstudent(model1_v5))
```



Adjusted R_squared is 0.014011. There are polynomial relationship between birthweight and number of visit. More visits doesn't necessarily incidate good outcome. There are Evidence for outliers in the data in the rstudent plot.

Checking the Classical Linear Model Assumptions for Ordinary Least Squares Regression.

Assumptions 1 and 2:

The first of the Classical Linear Model Assumptions state that there is a linear population model. This is a very weak assumption and we can accept it knowing that we'll only use linear parameters for our coefficients.

Assumption 2 states that the data comes from a random sample with the population. Looking at this data, we can see that the predominant race in the data is white, which we know is not representative of the population as a whole, but for populations of many communities across America it is. We'll choose to accept this assumption for this analysis.

Assumption 3 - No Perfect Multicollinearity

Assumption 3 states that no two of the independent variables are perfectly correlated. Naturally we would expect some correlation between number of visits and how early the parents began prenatal care, as an early start date would allow for more time for more visits, but they will not be perfectly correlated.

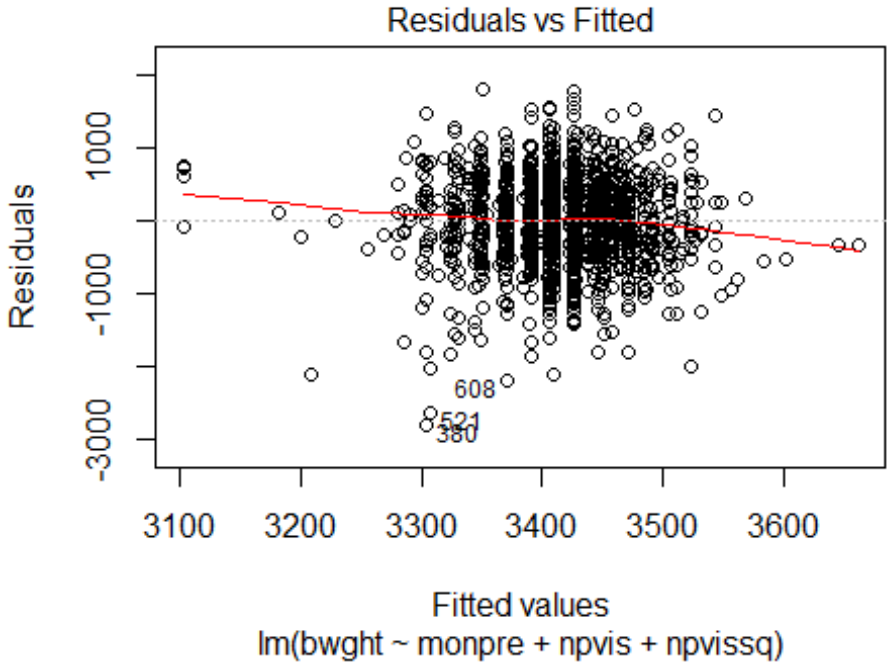
```
vif(model1_v5)

##   monpre    npvis  npvissq
## 1.242415 9.443689 8.807055
```

The variance inflation factors between the two variables are just over 1, which is not a significant cause for worry. This allows us to accept Assumption 3, No Perfect Multicollinearity.

Assumption 4 - Zero Conditional Mean

Assumption 4 states that none of the independent variables provide any information of the value of the residuals. For a given value of the independent variable, the expected value of the error term should not change, and should still be equal to 0. To check this assumption, we look at the residuals vs fitted values plot, and make sure there aren't any trends.

```
plot(model1_v5, which = 1)
```



There are a few slight bumps in the fitted curve, but across the dataset it does seem to be very flat and smooth. On the right side of the chart where the fitted values get very high, there is a noticeable downward slope on the residual, but there is significantly less data over there and it's not a major shift.

We'll choose to accept assumption 4 given this information.

Having accepted Assumption #4, we can now say that our estimators are unbiased and consistent.

Assumption 5 - Homoskedasticity of Errors.

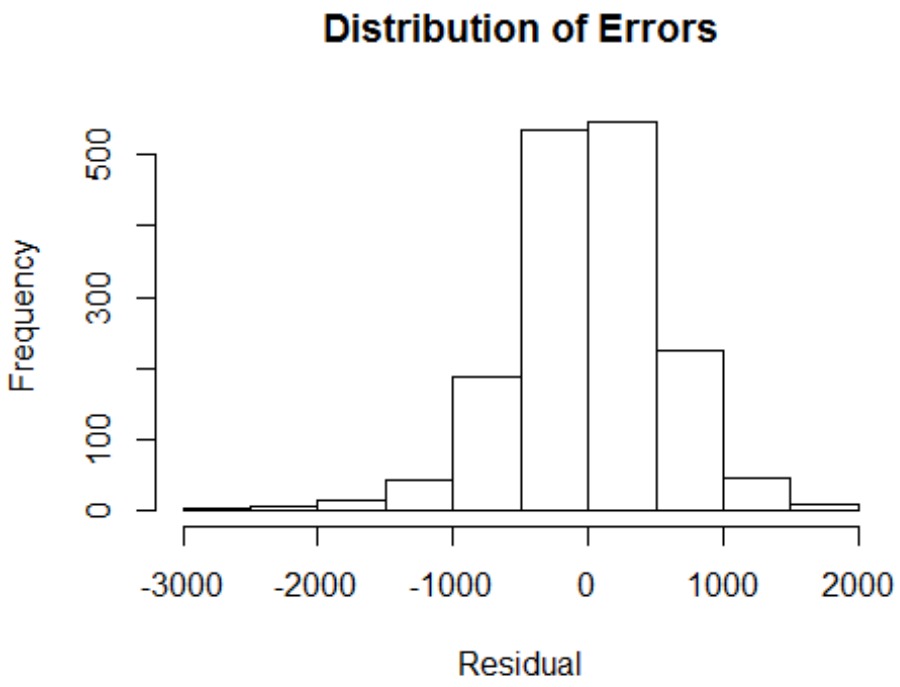
Assumption 5 states that the variance of the error term should be consistent across the entire range of fitted values. To test this, we can look at the plot above and see if the thickness of the band of points changes throughout the graph. Again, at the higher ranges of fitted values, we seem to have found some changes from the rest of the data. It appears there is less variance among the residuals when the fitted value is above 3500 grams.

To account for this we'll make sure to use heteroskedastic robust standard errors in our tests.

Assumption 6 - Normal Distribution of Errors

Assumption 6 requires that the errors are normally distributed. Looking at the histogram of the model's errors will show us whether we have a problem.

```
hist(model1_v5$residuals, main = "Distribution of Errors",
     xlab = "Residual")
```



We can see a strong concentration in center with a slight right skew, indicating not a perfect normal distribution, but the overall shape is okay, so we'll choose to accept this assumption.

Notes on this model: Omitted Variable Bias and Accuracy

This model is an extremely simplified approximation of a statistic that is known to have very many factors. Not including the other variables costs us a significant amount of accuracy and likely introduces omitted variable bias.

Model 2: Improving on previous model

From the exploratory analysis, we know that there are several more factors which are important in predicting infant birth weight. We saw that gender of the baby, cigarette use, and age of the father, education, and race were all important factors.

It is known that the age of the father is a factor in baby's health, but typically a mother's age is thought of to be more important. In this dataset, the father's age is a more important factor in the baby's weight, so we'll model using that.

```
model2 <- lm(bwght ~ npvis + monpre_cut + male + cigs + fage + avg_educ + whte, data
= df_naomit)
summary(model2)

##
## Call:
## lm(formula = bwght ~ npvis + monpre_cut + male + cigs + fage +
##     avg_educ + whte, data = df_naomit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2802.54  -327.46    13.13   362.61  1818.95
##
```

```
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2872.835    133.890  21.457  < 2e-16 ***
## npvis                 12.506      3.894   3.212 0.001346 **
## monpre_cut(3,6]       64.685     52.297   1.237 0.216315
## monpre_cut(6, Inf]   119.732    110.466   1.084 0.278583
## male                  92.036     27.811   3.309 0.000956 ***
## cigs                  -9.929      3.470  -2.862 0.004270 **
## fage                   5.384      2.557   2.105 0.035445 *
## avg_educ               6.847      7.628   0.898 0.369514
## whteTRUE              97.925     42.993   2.278 0.022878 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 556.4 on 1603 degrees of freedom
## Multiple R-squared:  0.02567,    Adjusted R-squared:  0.02081
## F-statistic: 5.279 on 8 and 1603 DF,  p-value: 1.498e-06
```

```
coeftest(model2, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                      Estimate Std. Error t value  Pr(>|t|)
## (Intercept)        2872.8346   133.2649 21.5573 < 2.2e-16 ***
## npvis                12.5063     4.2485  2.9437  0.003289 **
## monpre_cut(3,6]      64.6846    53.4465  1.2103  0.226354
## monpre_cut(6, Inf]  119.7318    76.0267  1.5749  0.115485
## male                 92.0361    28.0414  3.2821  0.001052 **
## cigs                 -9.9290     3.5461 -2.7999  0.005172 **
## fage                  5.3835     2.6263  2.0499  0.040538 *
## avg_educ              6.8470     7.1767  0.9541  0.340199
## whteTRUE             97.9245    40.0368  2.4459  0.014558 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

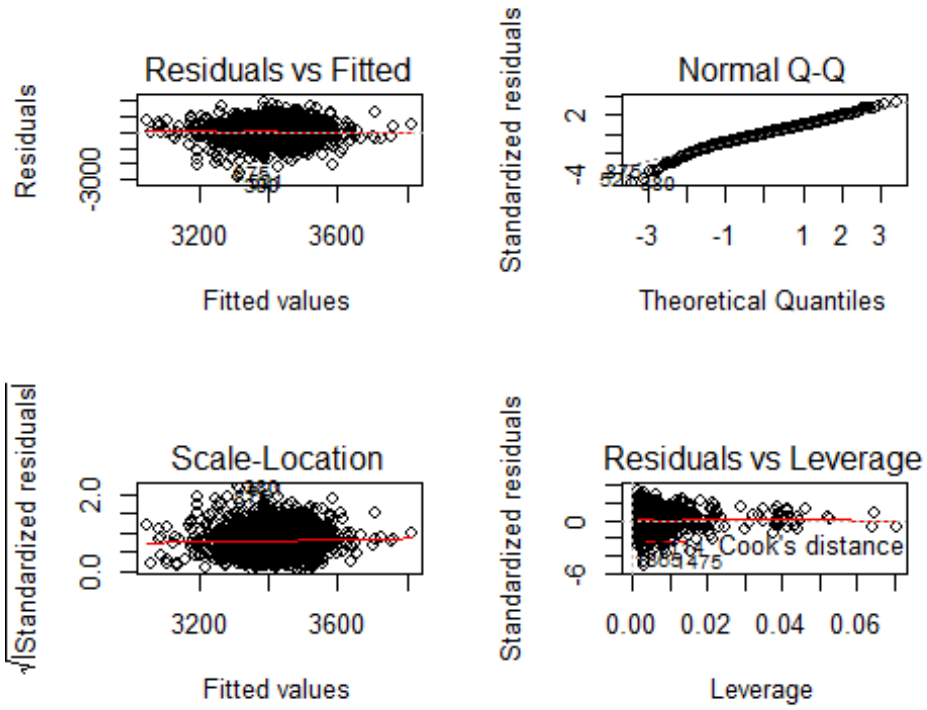
Checking CLM Assumptions for updated model.

After changing the model, we must reexamine the CLM assumptions to see if we still have BLUE OLS estimators and that we have normally distributed error terms. Normally distributed error terms will allow us to perform hypothesis tests on the coefficients, which will help us further refine the model.

```
vif(model2, vcov = vcovHC)
```

```
##                GVIF Df GVIF^(1/(2*Df))
## npvis      1.091466  1        1.044732
## monpre_cut 1.147313  2        1.034953
## male       1.005936  1        1.002964
## cigs       1.035158  1        1.017427
## fage       1.074198  1        1.036435
## avg_educ   1.151275  1        1.072975
## whte       1.009682  1        1.004829
```

```
par(mfrow = c(2,2))
plot(model2)
```



None of the coefficients have a variance inflation factor much greater than one, suggesting we have no issues with multicollinearity. The residuals vs fitted values for plot one show that we have a zero conditional mean for our errors and that heteroskedasticity isn't a significant problem.

Our Q-Q plot has some tails similar to the previous model, but the majority of the data falls nicely onto the straight line, so we don't foresee any issues with the distribution of our error terms.

In the summary output of the model, we notice that average years of education between the parents is does not have a significant coefficient value. We can run an F-test on the model and a reduced model without education to see if that would be a cause for improvement.

```
linearHypothesis(model2, c("avg_educ = 0"), vcov = vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## avg_educ = 0
##
## Model 1: restricted model
## Model 2: bwght ~ npvis + monpre_cut + male + cigs + fage + avg_educ +
##     whte
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F Pr(>F)
## 1   1604
## 2   1603  1 0.9102 0.3402
```

Running an F test for a reduced model, we see that we fail to reject the null that the restricted model is better by setting the coefficient for average education to zero. While not statistically significant on its own, it does improve the model.

Model 3: Problematic covariates

Mother's age has many non-linear interactions with the other explanatory variables when the mother is young. This makes sense that a mother under the age of 23 will not have had as much time to get more education. There also may be a social stigma which prevents them telling anyone they are pregnant until it begins to show, which would explain the later start to getting prenatal care.

To remedy this, we'll create a young mom indicator variable which will help account for the non linear effects of being a young mother. With this, we'll also keep the average education between the parents in the model because we may now be able to account for it properly.

```
df_naomit$young_mom <- cut(df_naomit$mage, c(-Inf, 23, Inf), labels = c(1, 0))
t.test(df_naomit$bwght[df_naomit$young_mom == 1], df_naomit$bwght[df_naomit$young_mom
== 0])

##
##  Welch Two Sample t-test
##
## data:  df_naomit$bwght[df_naomit$young_mom == 1] and df_naomit$bwght[df_naomit$you
ng_mom == 0]
## t = -2.7547, df = 210.11, p-value = 0.006391
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -235.51116  -39.03611
## sample estimates:
## mean of x mean of y
##  3292.333  3429.607

model3 <- lm(bwght ~ npvis + monpre_cut + male + cigs + fage + avg_educ + whte + youn
g_mom,
             data = df_naomit)
coeftest(model3, vcov = vcovHC)

##
## t test of coefficients:
##
##                      Estimate Std. Error t value  Pr(>|t|)
## (Intercept)        2889.9937   132.1460 21.8697 < 2.2e-16 ***
## npvis                12.1532     4.2359  2.8691  0.004170 **
## monpre_cut(3,6]      69.6594    53.7148  1.2968  0.194874
## monpre_cut(6, Inf]  129.6335    77.3354  1.6762  0.093884 .
## male                 92.0213    28.0339  3.2825  0.001051 **
## cigs                 -9.7350     3.5217 -2.7642  0.005771 **
## fage                  3.4466     2.7977  1.2319  0.218149
## avg_educ              4.6308     7.2097  0.6423  0.520766
## whteTRUE             96.4462    39.8794  2.4184  0.015698 *
## young_mom0           89.5119    54.6669  1.6374  0.101742
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(model3)

##
## Call:
## lm(formula = bwght ~ npvis + monpre_cut + male + cigs + fage +
##     avg_educ + whte + young_mom, data = df_naomit)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2814.55  -329.56    16.15   359.71  1809.35
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2889.994    134.150  21.543  < 2e-16 ***
## npvis                 12.153      3.896   3.119  0.00185 **
## monpre_cut(3,6]       69.659     52.337   1.331  0.18339
## monpre_cut(6, Inf]   129.634    110.534   1.173  0.24105
## male                  92.021     27.792   3.311  0.00095 ***
## cigs                  -9.735      3.469  -2.806  0.00507 **
## fage                   3.447      2.780   1.240  0.21517
## avg_educ               4.631      7.725   0.599  0.54893
## whteTRUE              96.446     42.973   2.244  0.02495 *
## young_mom0            89.512     50.500   1.773  0.07650 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 556.1 on 1602 degrees of freedom
## Multiple R-squared:  0.02758,    Adjusted R-squared:  0.02212
## F-statistic: 5.048 on 9 and 1602 DF,  p-value: 9.632e-07
```

```
linearHypothesis(model3, c("fage = 0", "avg_educ = 0", "young_mom0 = 0"), vcov = vcov
HC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## fage = 0
## avg_educ = 0
## young_mom0 = 0
##
## Model 1: restricted model
## Model 2: bwght ~ npvis + monpre_cut + male + cigs + fage + avg_educ +
##     whte + young_mom
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F  Pr(>F)
## 1   1605
## 2   1602  3 2.7634 0.04072 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

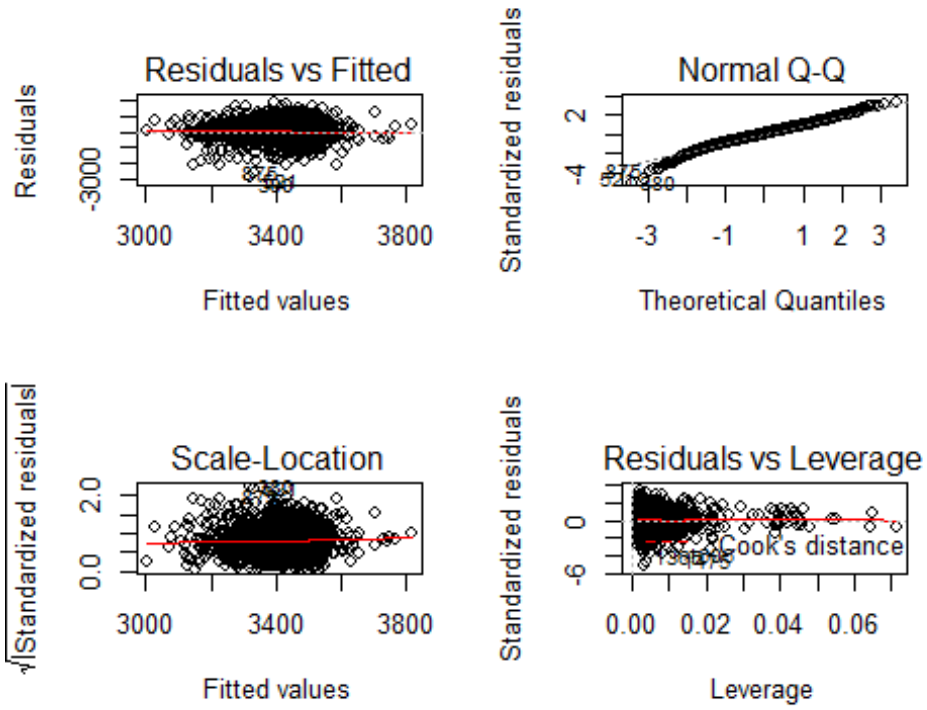
In model 3, On our linear hypothesis tests, we reject the restricted model which would remove the three variables accounting for age and education in our model. Of the models we built, model3 also contains the highest adjusted R-squared, of 0.02291.

The young_mom0 coefficient means that the adjustment for being a young mother is a reduction in

Checking the CLM Assumptions for the new model

After adding the indicator for young mothers, let's check whether our CLM assumptions still hold.

```
par(mfrow = c(2,2))
plot(model3)
```

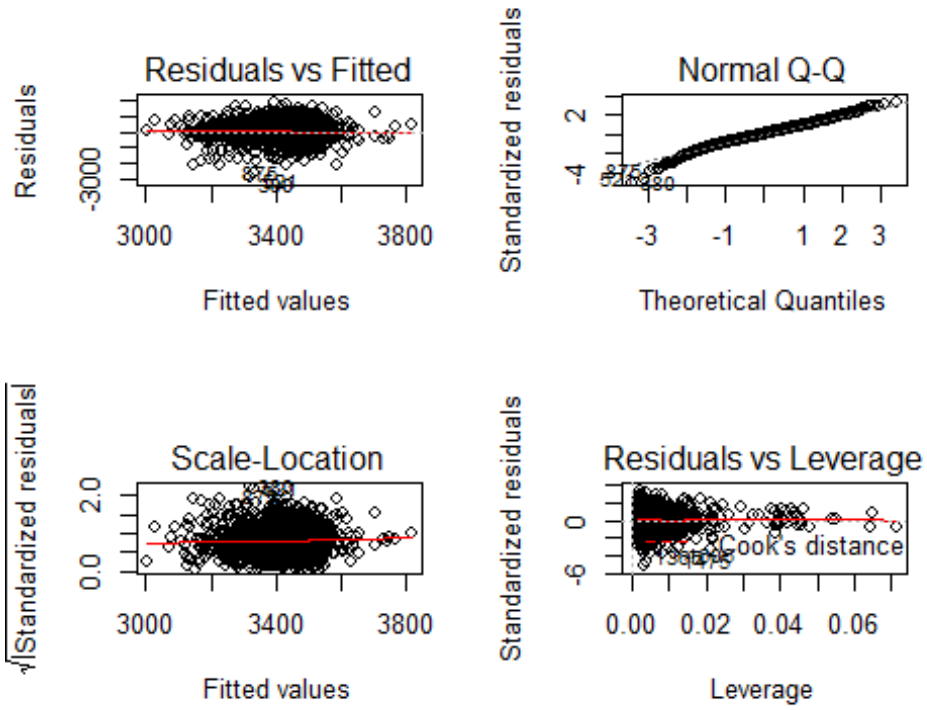


The plot of residuals vs fitted values shows that we maintain a zero-conditional mean and have actually reduced our heteroskedasticity.

The normality of the errors is still clumped towards the center with tails on either side. Again, we'll choose to accept the normality of errors assumption, CLM #6.

5. Regression Table

```
par(mfrow = c(2,2))
plot(model3)
```




```
stargazer::stargazer(model1_v5, model2, model3
                     , type = "text"
                     , star.cutoffs = c(0.05, 0.01, 0.005)
                     , title = "Regression analysis of baby health outcome and prenata
l care"
                     , covariate.labels = c("Prenatal Care (Month)", "Prenatal Visits"
                                            , "Prenatal Visits (Squared)"
                                            , "Prenatal Care (1st/2nd Tri)"
                                            , "Prenatal Care (3rd Tri)"
                                            , "Newborn is Male", "Cigarettes per Day"
                                            , "Father's Age", "Parents' Education"
                                            , "Parents are white", "Mother a young pare
nt?")
                     , dep.var.labels   = "Birthweight (in grams)"
                     , single.row = F, column.sep.width = "2pt"
                     , align = T, no.space = T)
```

```
## Regression analysis of baby health outcome and prenatal care
## ================================================================================================
##                                              Dependent variable:
##                                  ------------------------------------------------------------
##                                              Birthweight (in grams)
##                                      (1)                 (2)                 (3)
## ------------------------------------------------------------------------------------------------
## Prenatal Care (Month)              19.816
##                                   (12.622)
## Prenatal Visit                    30.570**            12.506***           12.153***
##                                   (11.536)             (3.894)             (3.896)
## Prenatal Visit (Squared)          -0.573
##                                   (0.361)
## Prenatal Care (1st/2nd Tri)                           64.685              69.659
##                                                      (52.297)            (52.337)
## Prenatal Care (3rd Tri)                              119.732             129.634
##                                                      (110.466)           (110.534)
## Newborn is Male                                      92.036***           92.021***
##                                                      (27.811)            (27.792)
## Cigarettes per Day                                   -9.929***           -9.735**
##                                                      (3.470)             (3.469)
## Father's Age                                         5.384*              3.447
##                                                      (2.557)             (2.780)
## Parents' Education                                   6.847               4.631
##                                                      (7.628)             (7.725)
## Parents are white                                    97.925*             96.446*
##                                                      (42.993)            (42.973)
## Mother a young parent?                                                   89.512
##                                                                          (50.500)
## Constant                        3,102.165***        2,872.835***        2,889.994***
##                                   (101.284)           (133.890)           (134.150)
## ------------------------------------------------------------------------------------------------
## Observations                        1,612               1,612               1,612
## R2                                  0.009               0.026               0.028
## Adjusted R2                         0.007               0.021               0.022
## Residual Std. Error           560.421 (df = 1608) 556.444 (df = 1603) 556.072 (df = 1602)
## F Statistic                4.656*** (df = 3; 1608) 5.279*** (df = 8; 1603) 5.048*** (df = 9; 1602)
## ================================================================================================
## Note:                                                          *p<0.05; **p<0.01; ***p<0.005
```

Statistical significance

We observed multiple instances with statistical significance:

1/ Male and female birthweight are different

2/ The birth weights of babies with two-white parents vs babies which do not have two white parents

3/ Babies whose mothers smoke vs those whose mothers did not smoke

4/ With this dataset, there are almost 1600 mothers who do not drink and only 16 who do, even though there is a difference of about 80 grams of birthweight for the baby between the two groups

Practical significance

From the practical view of the statistically significant observations, we draw these conclusions:

1/ Male is born, on average, ~3% heavier than female

2/ Babies that are born to two white parents tends to be heavier than babies that do not have two white parents.

3/ Mothers who do not smoke, on average, give birth to babies that are ~6% heavier than babies of mothers who smoke during pregnancy

4/ Prenatal care visit and cigarette usage during the pregnancy can affect newborn weight

6. Discussions

We cannot draw causality from the model for two reasons:

1/ The fits are not ideal, at on 2.29%

2/ There's too many other omitted variables that could be causing bias. For examples

a/ Parents' income

b/ Parents' health metrics such as BMI

c/ If newborn was premature. This is from observing mothers with more than 20 prenatal care visits. This is often an indication of high risk pregnancy.

d/ Only 1.3% of the newborn from the dataset was considered to have low birthweight or very low birthweight

At the same time, dataset is potentially bias and might not represent a truly random sample of American population:

1/ Most variables are heavily skewed to predominantly white, highly educated parents that are also non-smoker, non-drinker

2/ Majority of mothers start prenatal care at 2 months (first tri-semester), indicating a planned pregnancy

7. High-level takeaways.

In summary, our model explains a week relationship between the outcome measured in birthweight and the number prenatal care visits. Father age was a surprising contributing variable in our model. This seems counter-intuitive especially when mother age did not have the same effect; therefore, we suspect that the data was potentially non-random. Mothers' supposedly detrimental behaviors as cigarettes and drink consumptions were considered but only # of cigarette was statistically significant in explaining some of the outcome variability. Only 16/1612 mothers included in the analysis consumed alcohol during pregnancy while 147 smoked. Further studies on the effect of cigarette and birthweight might prove beneficial in advising mother's behaviors.