Lab 4: Healthy Momma, Healthy Baby

Krista Mar and Nikki Haas

12/1/2016

Introduction

According to the NIH, having a healthy pregancy is one of the best ways to promote a healthy birth. Getting early and regular prenatal care improves the chances of a healthy pregnancy.[1] While most low birth weight children will end up having normal outcomes, as a group they generally have more health issues including subnormal neural outcomes than healthy weight babies[2]. Birthweight is a predictor of brain development in childhood and adulthood. [2a]

Apgar scores are used as an evaluative measure to see if a newborn needs immediate attention. However, the using Apgar scores to attempt to predict long-term developmental outcomes of infants in not appropriates, so we will not be using Apgar scores in our outcome variable for newborn health. [3]

Therefore we will use birthweight as our outcome variable for our analysis based on historical research because of the limitations of our dataset.

Using data from the National Center for Health Statistics and from birth certificates, we will look at the impact of prenatal health care on health outcomes for newborn infants.

Step 1: Read in the Data

```r
setwd("/Users/krista/Desktop/Final lab 4")
load("bwght_w203.RData")
desc
```

```
##    variable                           label
## 1      mage             mother's age, years
## 2     meduc            mother's educ, years
## 3    monpre       month prenatal care began
## 4     npvis total number of prenatal visits
## 5      fage             father's age, years
## 6     feduc            father's educ, years
## 7     bwght             birth weight, grams
## 8     omaps          one minute apgar score
## 9     fmaps         five minute apgar score
## 10     cigs          avg cigarettes per day
## 11    drink             avg drinks per week
## 12      lbw             =1 if bwght <= 2000
## 13     vlbw             =1 if bwght <= 1500
## 14     male                 =1 if baby male
## 15    mwhte              =1 if mother white
## 16    mblck              =1 if mother black
## 17     moth           =1 if mother is other
## 18    fwhte              =1 if father white
## 19    fblck              =1 if father black
## 20     foth           =1 if father is other
## 21   lbwght                      log(bwght)
## 22   magesq                          mage^2
## 23  npvissq                         npvis^2
```

Step 2: Exploratory Data Analysis

First, get summary statistics on each element of the dataset:

```
nrow(data)
```

```
## [1] 1832
```

```
summary(data)
```

```
##       mage           meduc           monpre          npvis      
##  Min.   :16.00   Min.   : 3.00   Min.   :0.000   Min.   : 0.00  
##  1st Qu.:26.00   1st Qu.:12.00   1st Qu.:1.000   1st Qu.:10.00  
##  Median :29.00   Median :13.00   Median :2.000   Median :12.00  
##  Mean   :29.56   Mean   :13.72   Mean   :2.122   Mean   :11.62  
##  3rd Qu.:33.00   3rd Qu.:16.00   3rd Qu.:2.000   3rd Qu.:13.00  
##  Max.   :44.00   Max.   :17.00   Max.   :9.000   Max.   :40.00  
##                  NA's   :30      NA's   :5       NA's   :68     
##       fage           feduc           bwght          omaps       
##  Min.   :18.00   Min.   : 3.00   Min.   : 360   Min.   : 0.000  
##  1st Qu.:28.00   1st Qu.:12.00   1st Qu.:3076   1st Qu.: 8.000  
##  Median :31.00   Median :14.00   Median :3425   Median : 9.000  
##  Mean   :31.92   Mean   :13.92   Mean   :3401   Mean   : 8.386  
##  3rd Qu.:35.00   3rd Qu.:16.00   3rd Qu.:3770   3rd Qu.: 9.000  
##  Max.   :64.00   Max.   :17.00   Max.   :5204   Max.   :10.000  
##  NA's   :6       NA's   :47                     NA's   :3       
##      fmaps             cigs            drink             lbw         
##  Min.   : 2.000   Min.   : 0.000   Min.   :0.0000   Min.   :0.00000  
##  1st Qu.: 9.000   1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.:0.00000  
##  Median : 9.000   Median : 0.000   Median :0.0000   Median :0.00000  
##  Mean   : 9.004   Mean   : 1.089   Mean   :0.0198   Mean   :0.01638  
##  3rd Qu.: 9.000   3rd Qu.: 0.000   3rd Qu.:0.0000   3rd Qu.:0.00000  
##  Max.   :10.000   Max.   :40.000   Max.   :8.0000   Max.   :1.00000  
##  NA's   :3        NA's   :110      NA's   :115                        
##       vlbw               male            mwhte            mblck       
##  Min.   :0.000000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:0.0000  
##  Median :0.000000   Median :1.0000   Median :1.0000   Median :0.0000  
##  Mean   :0.007096   Mean   :0.5136   Mean   :0.8865   Mean   :0.0595  
##  3rd Qu.:0.000000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000  
##  Max.   :1.000000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
##                                                                       
##       moth             fwhte            fblck              foth        
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000  
##  1st Qu.:0.00000   1st Qu.:1.0000   1st Qu.:0.00000   1st Qu.:0.00000  
##  Median :0.00000   Median :1.0000   Median :0.00000   Median :0.00000  
##  Mean   :0.05404   Mean   :0.8897   Mean   :0.05841   Mean   :0.05186  
##  3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.00000  
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000  
##                                                                        
##      lbwght          magesq          npvissq      
##  Min.   :5.886   Min.   : 256.0   Min.   :   0.0  
##  1st Qu.:8.031   1st Qu.: 676.0   1st Qu.: 100.0  
##  Median :8.139   Median : 841.0   Median : 144.0  
##  Mean   :8.114   Mean   : 896.4   Mean   : 148.6  
##  3rd Qu.:8.235   3rd Qu.:1089.0   3rd Qu.: 169.0  
##  Max.   :8.557   Max.   :1936.0   Max.   :1600.0  
##                                   NA's   :68
```

Response Variables

The bwght, lbwght, omaps and fmaps variables are related to the health of the baby.

The first thing to check is if these variables are collinar. We will omit bwghts as that is a function of lbwghts.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

```
cor(data$omaps, data$fmaps, use = "complete.obs")
```

```
## [1] 0.5575238
```

```
cor(data$lbwght, data$fmaps, use = "complete.obs")
```

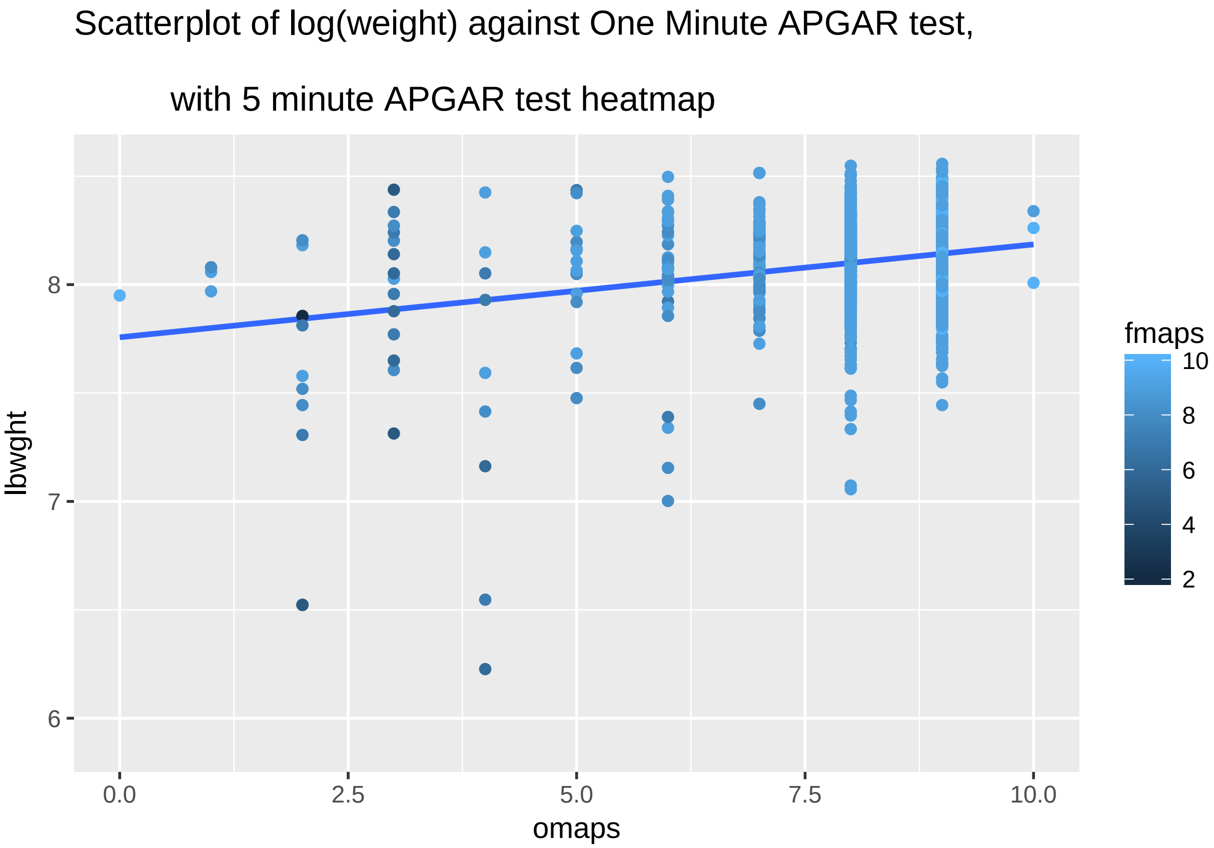
```
## [1] 0.2710456
```

```
p <- ggplot(data, aes(omaps, lbwght)) + geom_point(size = 0.25) +
  geom_smooth(method = "lm", se = FALSE) + geom_point(aes(colour = fmaps)) +
  ggtitle("Scatterplot of log(weight) against One Minute APGAR test,\n
          with 5 minute APGAR test heatmap")
p
```

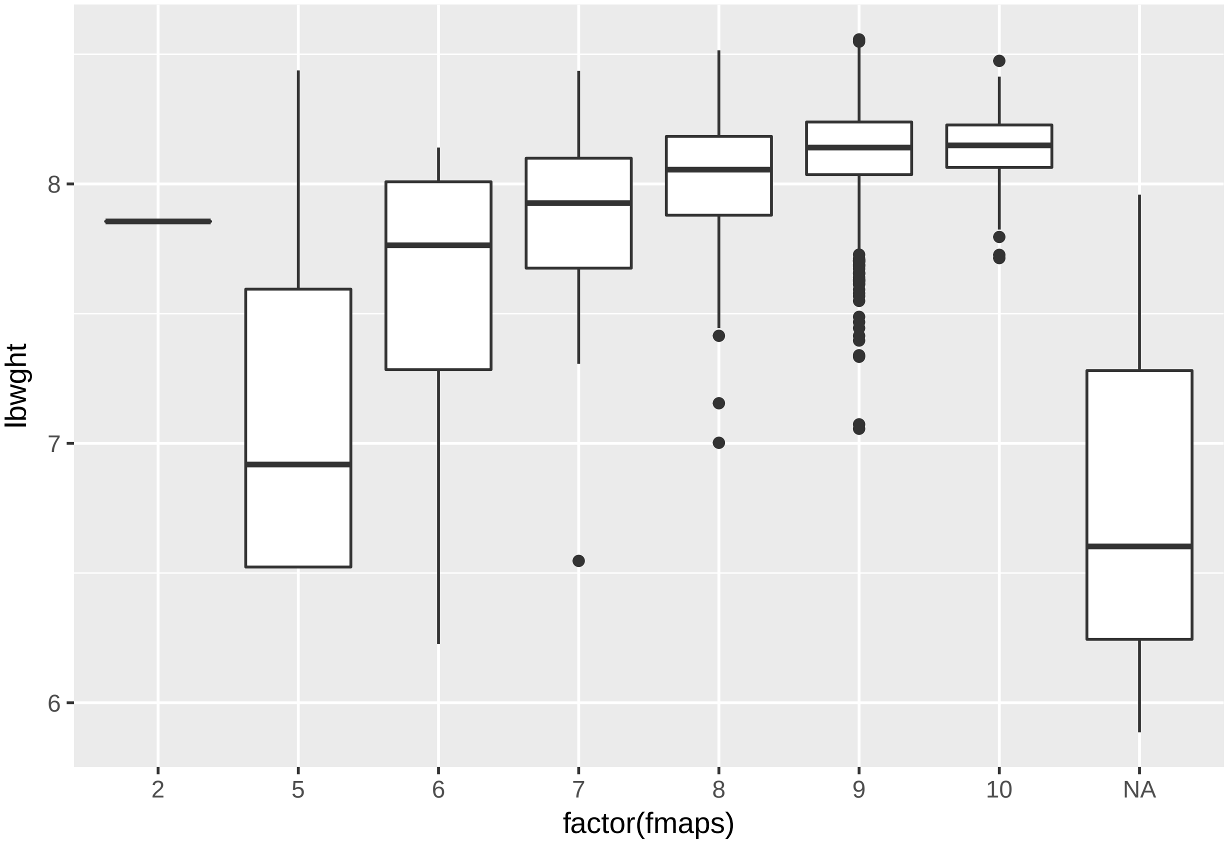
```
## Warning: Removed 3 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 3 rows containing missing values (geom_point).
```

```
## Warning: Removed 3 rows containing missing values (geom_point).
```



```
p <- ggplot(data, aes(factor(fmaps), lbwght)) + geom_boxplot()
p
```



Look at the extreme fmaps case

```
data[data$fmaps< 4,]
```

```
##      mage meduc monpre npvis fage feduc bwght omaps fmaps cigs drink lbw
## NA     NA    NA     NA    NA   NA    NA    NA    NA    NA   NA    NA  NA
## 837    32    12      2    10   40    16  2580     2     2    0     0   0
## NA.1   NA    NA     NA    NA   NA    NA    NA    NA    NA   NA    NA  NA
## NA.2   NA    NA     NA    NA   NA    NA    NA    NA    NA   NA    NA  NA
##      vlbw male mwhte mblck moth fwhte fblck foth   lbwght magesq npvissq
## NA     NA   NA    NA    NA   NA    NA    NA   NA       NA     NA      NA
## 837     0    1     1     0    0     1     0    0 7.855545   1024     100
## NA.1   NA   NA    NA    NA   NA    NA    NA   NA       NA     NA      NA
## NA.2   NA   NA    NA    NA   NA    NA    NA   NA       NA     NA      NA
```

Looking at the data, we can be reasonably assured that the response variables are related, but not collinear. It may be best to make a combined variable of `fmaps` and `omaps` such as `mapscombined = fmaps + omaps`. The difference would not make much sense compared to the sum; 10 - 10 and 2 - 2 are both zero, after all.

Regressors

The variables monpre and npvis are related to the prenatal care given during pregnancy. Let us review them for collinearity:

```
cor(data$npvis, data$monpre, use = "complete.obs")
```

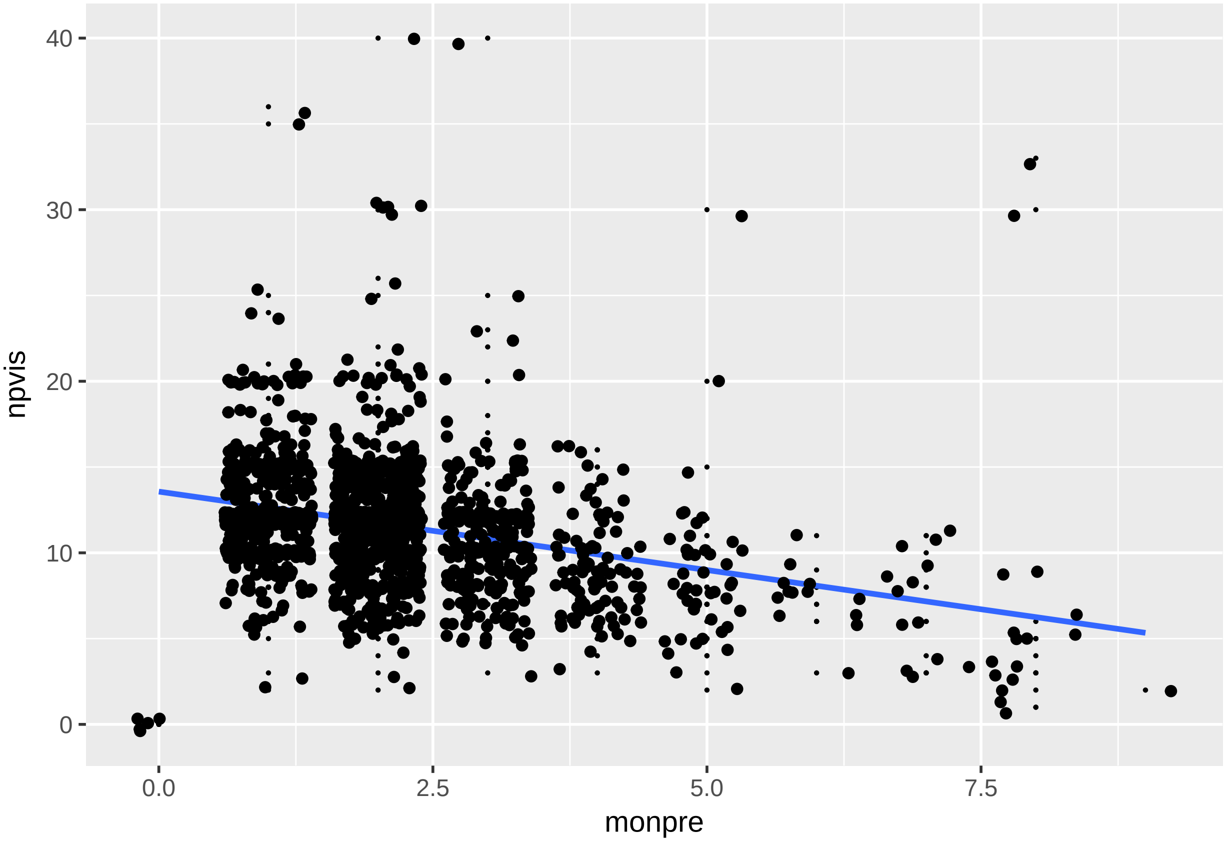
```
## [1] -0.3061006
```

```
ggplot(data, aes(monpre, npvis)) + geom_point(size = 0.25) +
  geom_smooth(method = "lm", se = FALSE) + geom_jitter()
```

```
## Warning: Removed 69 rows containing non-finite values (stat_smooth).
```

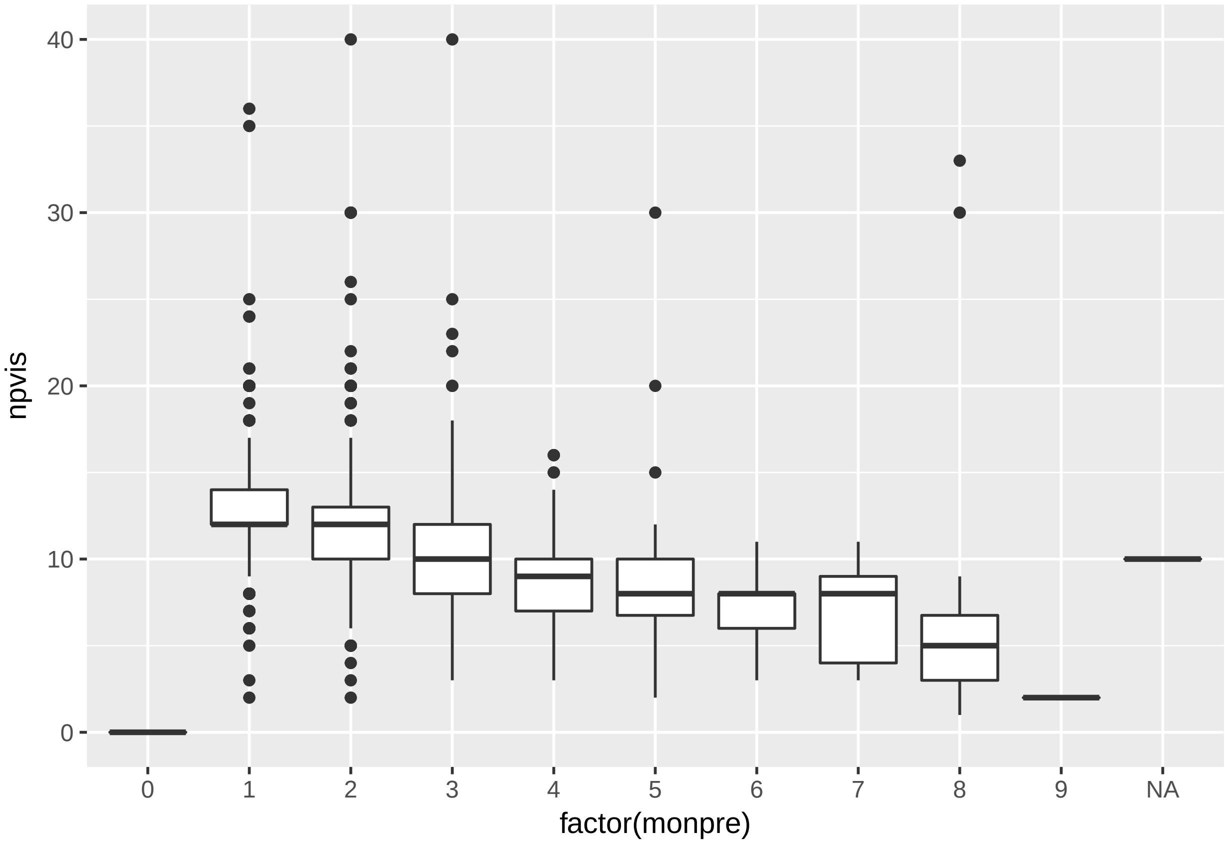
```
## Warning: Removed 69 rows containing missing values (geom_point).
```

```
## Warning: Removed 69 rows containing missing values (geom_point).
```



```
ggplot(data, aes(factor(monpre), npvis)) + geom_boxplot()
```

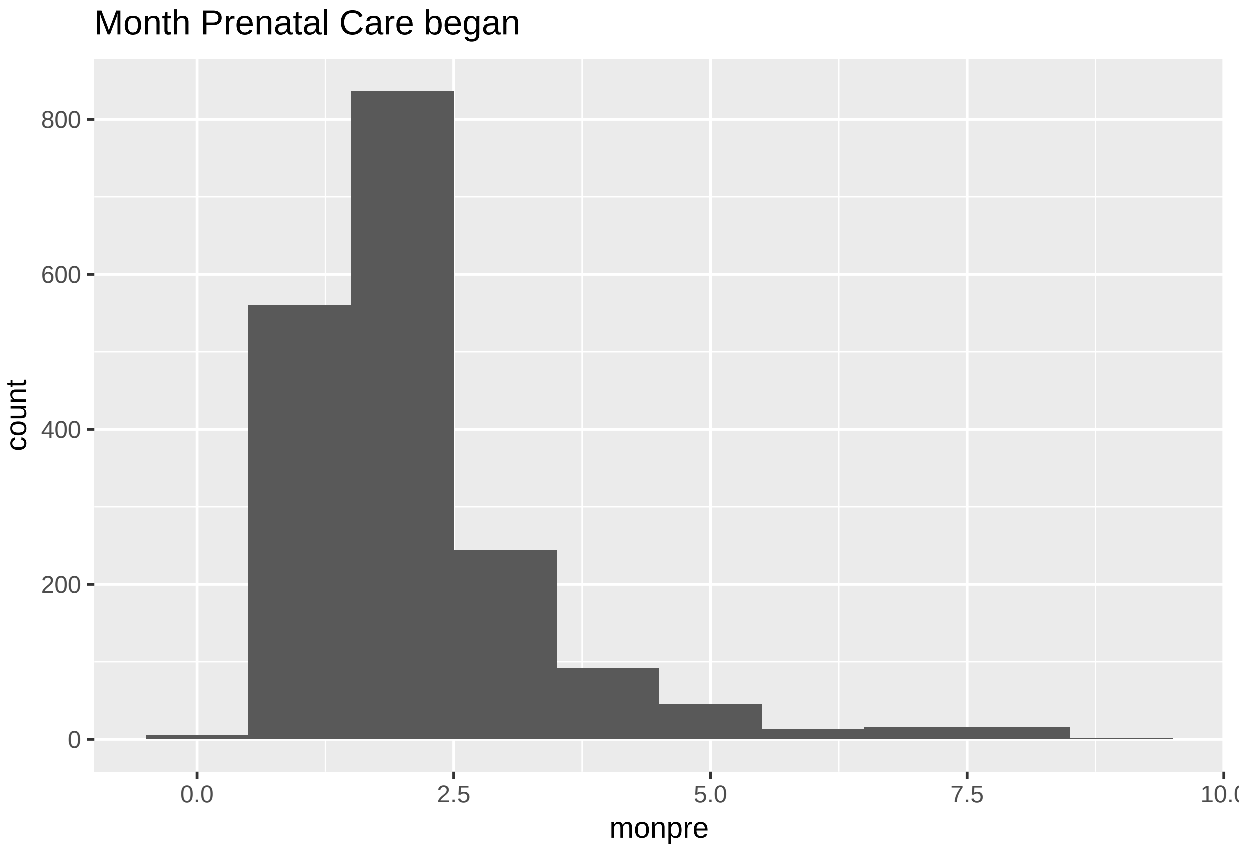
```
## Warning: Removed 68 rows containing non-finite values (stat_boxplot).
```



From this set, we can see that the data is not collinear, and indeed we can see that we might have some reporting errors. 5 mothers are listed as starting prenatal care in month 0 of their pregnancy, but they visited the doctor 0 times. These probably denote missing information or an error in reporting. Unfortunately, this data does show a definitive downward trend leading us to suspect that the number of visits is a function of month prenatal care began. This makes sense intuitively; if a mother starts prenatal care in her 2nd month of pregnancy, she has ample time for frequent doctor visits. However, if she starts her prenatal care towards the end of her pregnancy, she does not have enough time to visit the doctor as often as a woman who started in month 2.

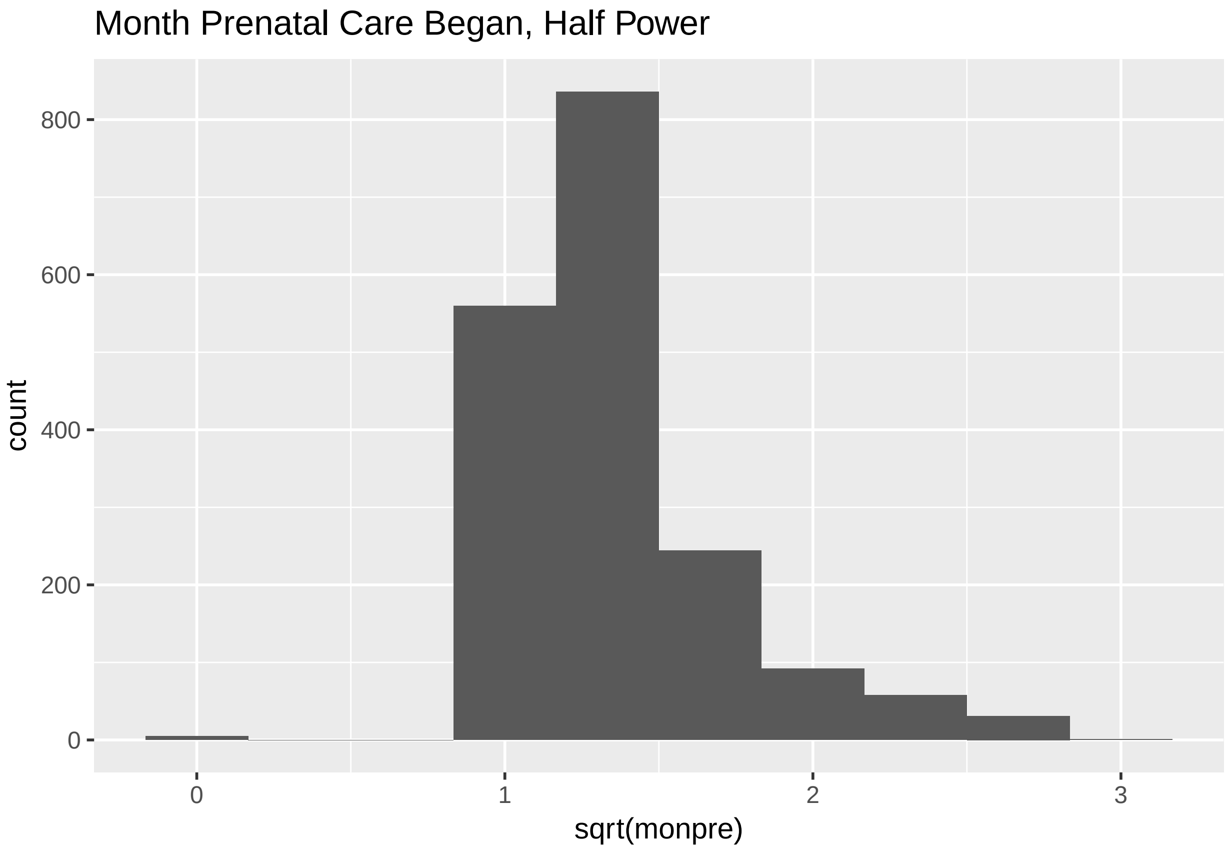
```
ggplot(data, aes(x=monpre)) + geom_histogram(aes(y = ..count..),bins = 10) +
  ggtitle("Month Prenatal Care began")
```

```
## Warning: Removed 5 rows containing non-finite values (stat_bin).
```



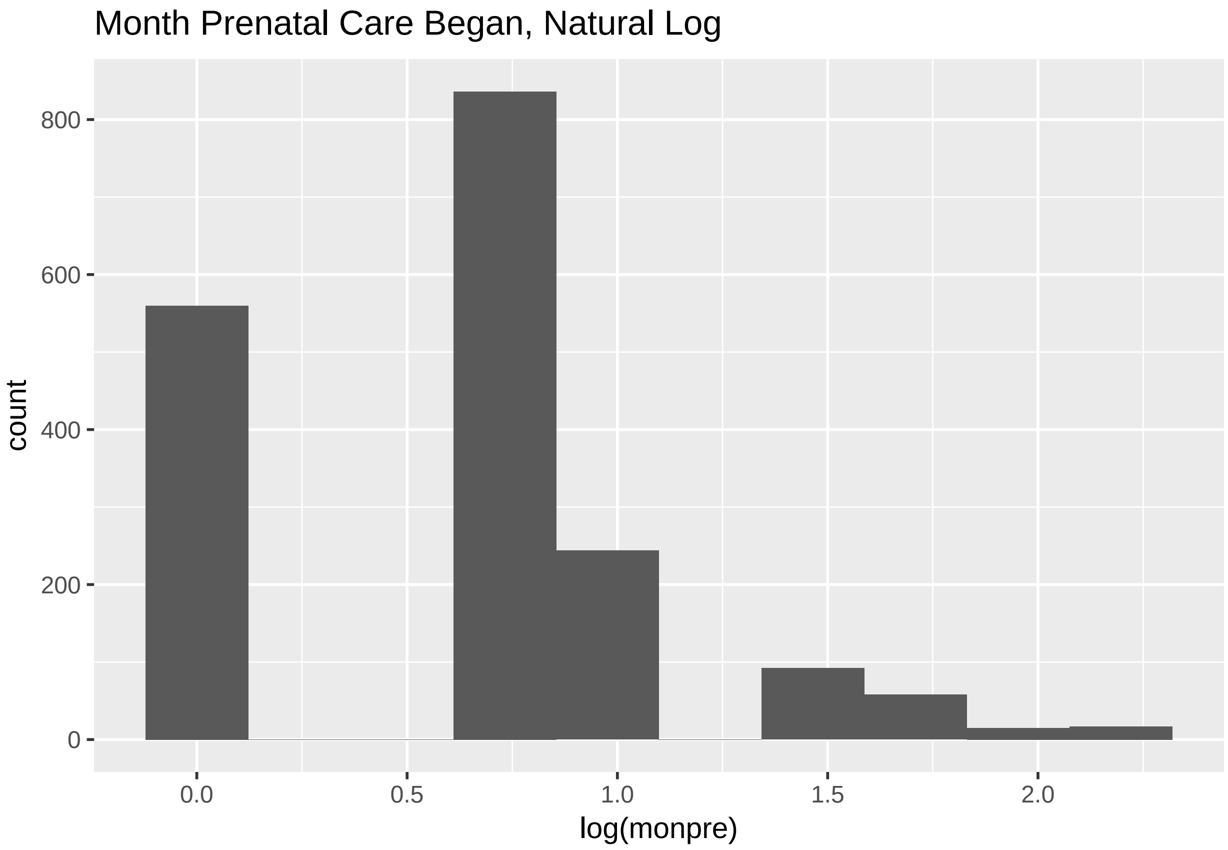
```
ggplot(data, aes(x=sqrt(monpre))) + geom_histogram(aes(y = ..count..), bins = 10) +
  ggtitle("Month Prenatal Care Began, Half Power")
```

```
## Warning: Removed 5 rows containing non-finite values (stat_bin).
```



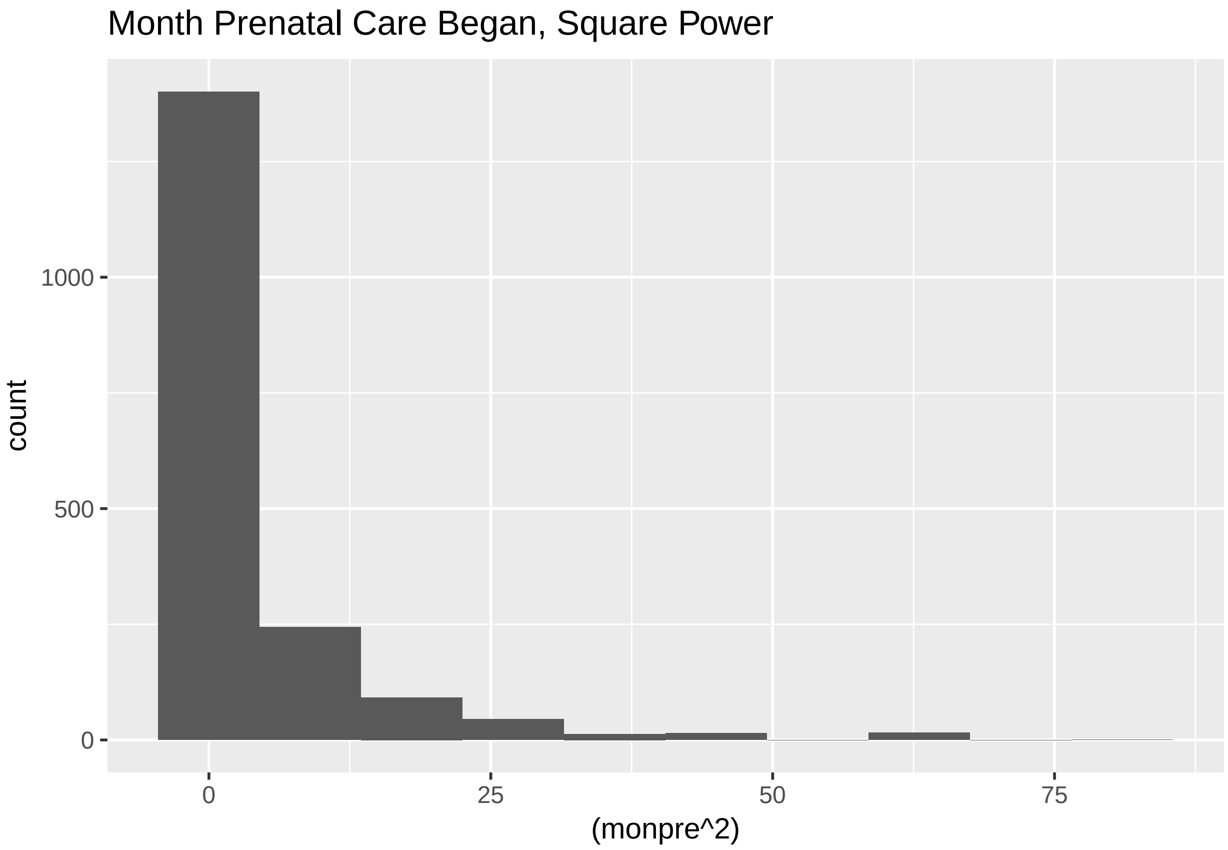
```
ggplot(data, aes(x=log(monpre))) + geom_histogram(aes(y = ..count..), bins = 10) +
  ggtitle("Month Prenatal Care Began, Natural Log")
```

```
## Warning: Removed 10 rows containing non-finite values (stat_bin).
```



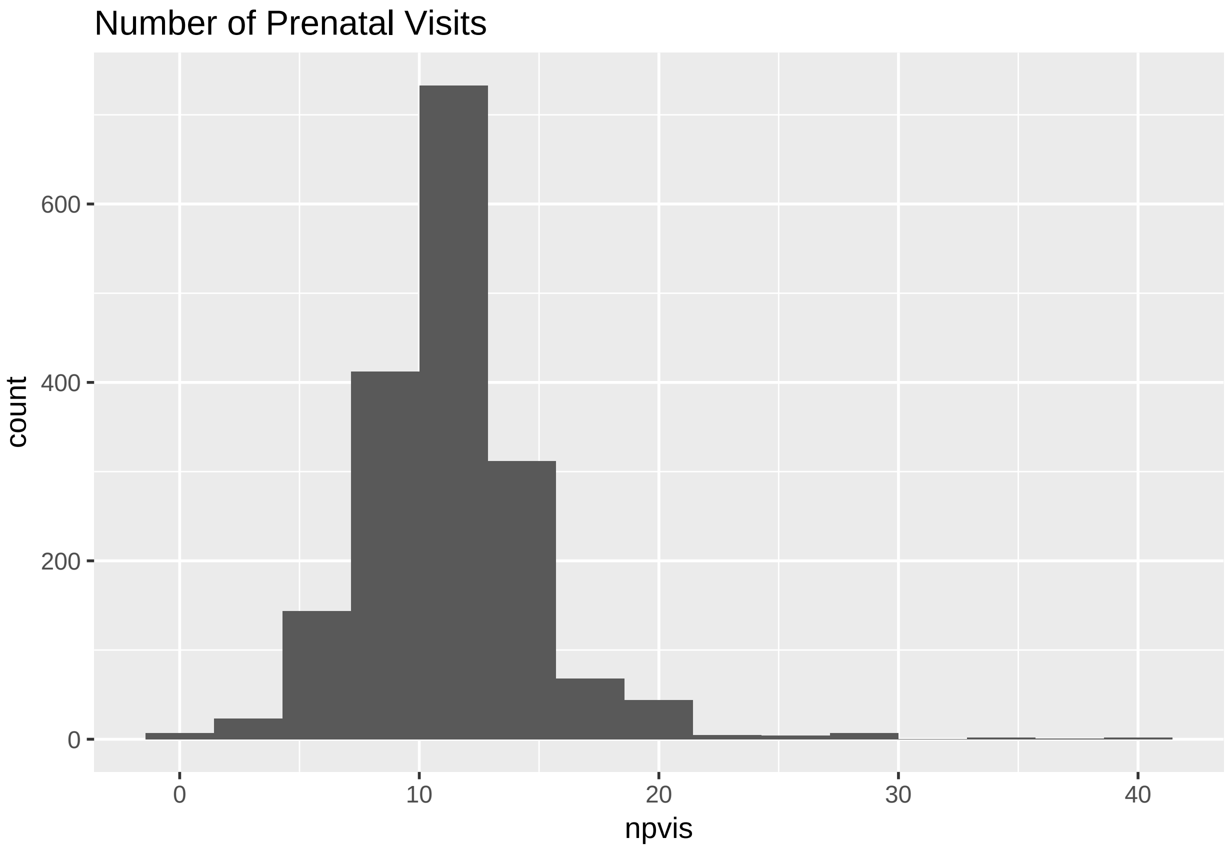
```
ggplot(data, aes(x=(monpre^2))) + geom_histogram(aes(y = ..count..), bins = 10) +
  ggtitle("Month Prenatal Care Began, Square Power")
```

```
## Warning: Removed 5 rows containing non-finite values (stat_bin).
```



```
ggplot(data, aes(x=npvis)) + geom_histogram(aes(y = ..count..), bins = 15) +
  ggtitle("Number of Prenatal Visits")
```

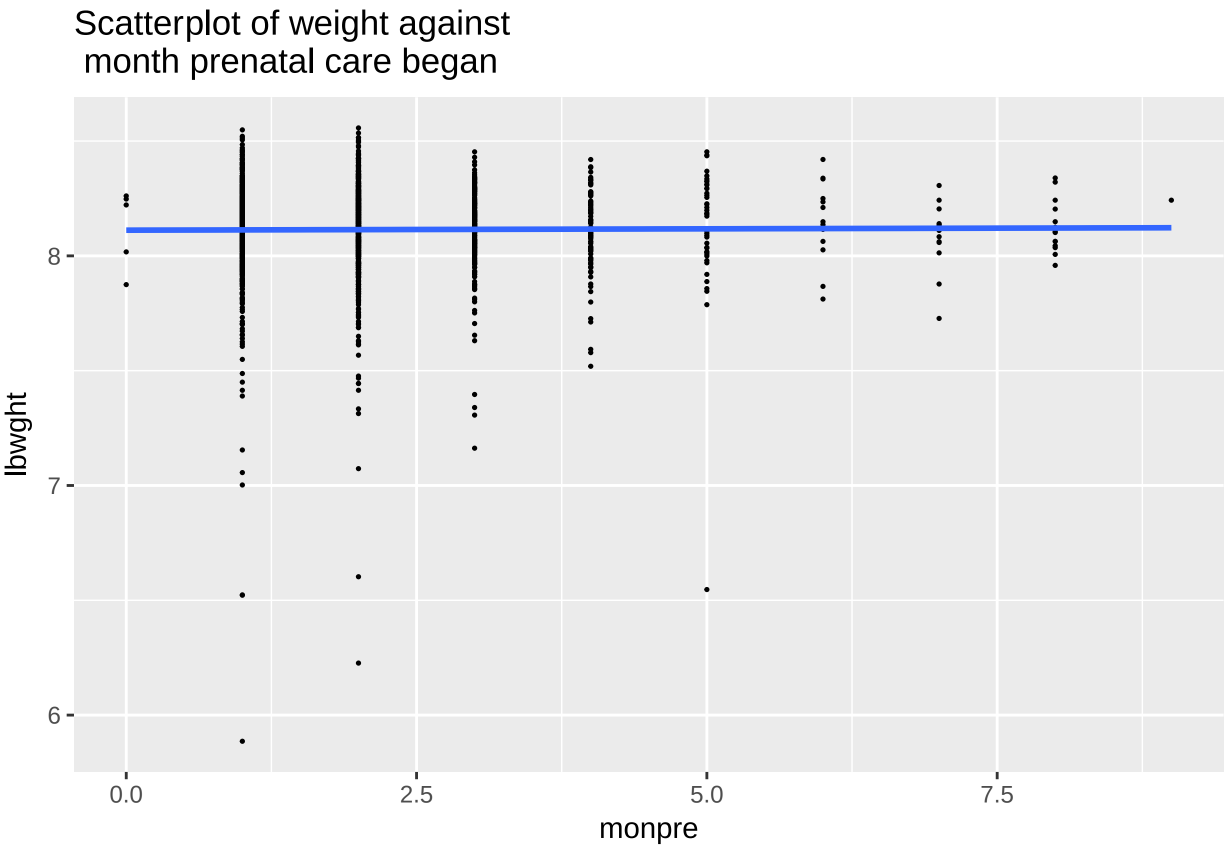
```
## Warning: Removed 68 rows containing non-finite values (stat_bin).
```



```
ggplot(data, aes(monpre, lbwght)) + geom_point(size = 0.25) +
  geom_smooth(method = "lm", se = FALSE)  +
  ggtitle("Scatterplot of weight against \n month prenatal care began ")
```

```
## Warning: Removed 5 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 5 rows containing missing values (geom_point).
```



All in all, the number of visits follows a mostly normal curve, and the square root of the month prenatal care began follow a mostly normal curve. However, we can tell right now that `monpre` does not have much practical significance with respect to the baby's weight from looking at the graph.

Step 3: Modeling and checking CLM

Model 1: Basic Linear Model

```
model1<-lm(bwght ~ monpre + npvis, data = data)
summary(model1)$r.squared
```

```
## [1] 0.01123524
```

6 CLM assumptions:

1) Linearity in parameters: We can assume this.
2) Random sampling of data: Not random because are not including still births.
3) No perfect co-linearity.

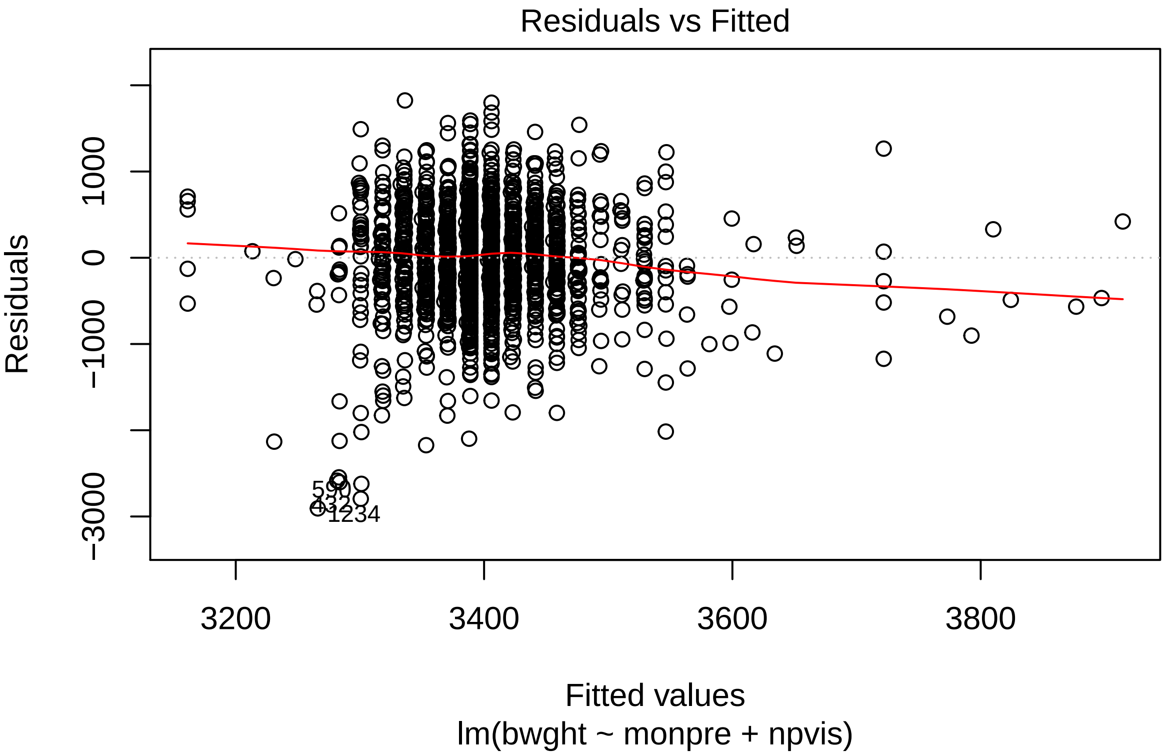
```
cor(data$monpre, data$npvis, use="complete.obs")
```

```
## [1] -0.3061006
```

There is no perfect multicolineraity between our variables. With a correlation of -0.3061006, this shows that the number of prenatal visits is moderately negatively correlated to the month in which prenatal care started. This makes sense because in the scale for prenatal care visits, being lower (e.g. starting in month 0) is better. If you started prenatal care visits in month 0, you are likely to have visited the doctor more times.

4) Zero conditional mean

```r
plot(model1, which=1)
```



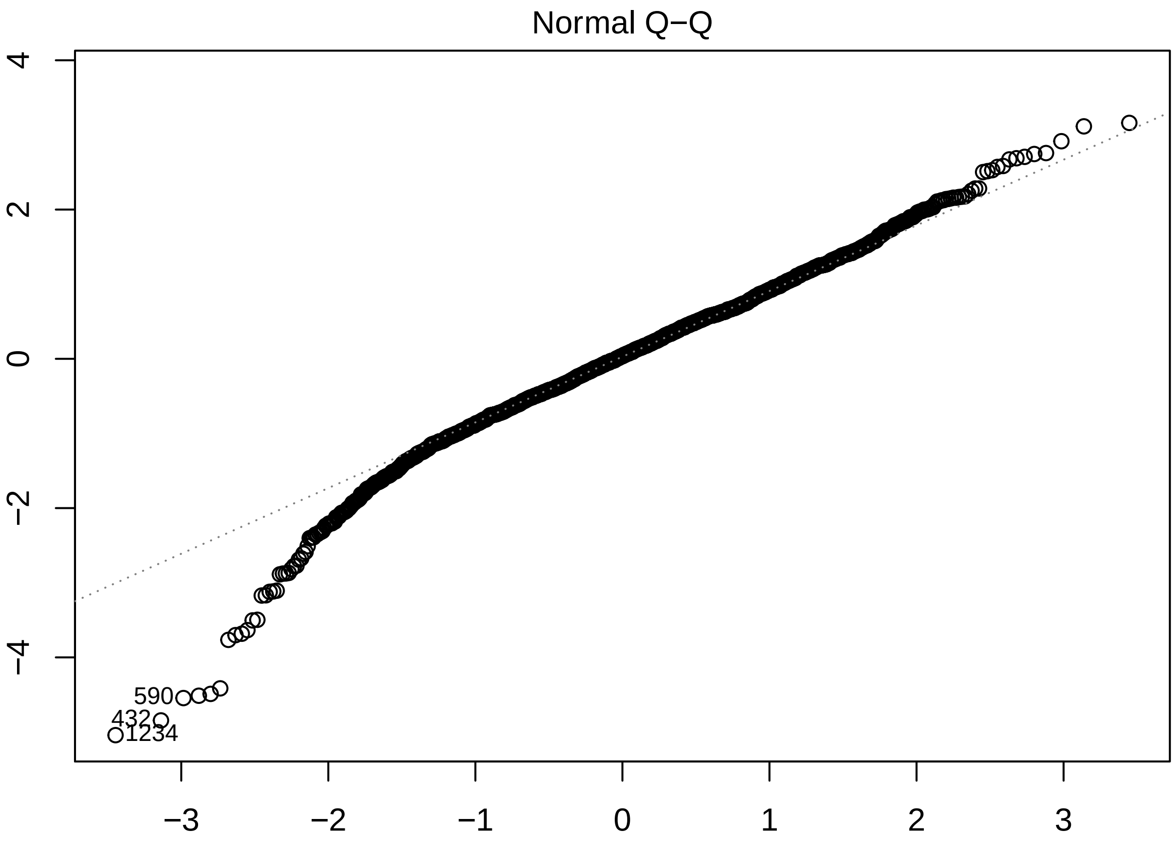
Looking at the Residuals vs. Fitted plot shows that the zero conditional mean is met because the red line is approximately at 0.

5) Homoskedacity of errors

From the residuals vs. fitted plot, we can see that we do not have homoskedacity of erorrs because the data is not in an even band across the plot. This means that we'll have to white standard errors, which are roboust to heteroskadacity.

6) Errors are normally distributed

```r
par(mar = rep(2, 4))
plot(model1, which=2)
```



```
shapiro.test(model1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model1$residuals
## W = 0.97715, p-value = 3.714e-16
```

Checking the normal Q-Q plot, it looks like our errors are roughly normally distributed. The Q-Q plot isn't perfectly normal.

Using the shapiro wilke test, we can reject the null hypothesis that the population has a normal distribution.

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

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

```
library(sandwich)
coeftest(model1, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 3161.2707    74.6049 42.3735 < 2.2e-16 ***
## monpre        17.0622    12.0277  1.4186 0.1561984
## npvis         17.5494     4.8342  3.6302 0.0002913 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Running our coeftest, we see that the number of prenatal care visits is statistically significant. We also notice taht month prenatal care began is not statistically significant.

Model 2: An Alternate Main Model

The 1 minute and 5 minute APGAR scores on their own do not tell us much. As we can see from the heatmap on the first scatterplot, a baby who has a low one minute score tends to have a higher five minute score. There are very few examples of a baby having a worse five minute score than a one minute score:

```
nrow(data[!is.na(data$fmaps) < !is.na(data$omaps),])
```

```
## [1] 3
```

However, we can get some information if we take the product of `omaps` and `fmaps` and then normalize it. A baby that goes from 0 to 10 then would have an overal low score compared to a baby who started with a score of 10 and was still at 10 5 minutes later, so the difference doesn't make sense.

```
data$product_apgarscores = data$omaps * data$fmaps
data$normalized_product_apgar =
  (data$product_apgarscores -
     mean(!is.na(data$product_apgarscores)))/sd(!is.na(data$product_apgarscores))

a8 = lm(data$normalized_product_apgar~data$monpre + data$npvis)
a9 = lm(data$normalized_product_apgar~ data$npvis)

AIC(a8)
```

```
## [1] 24885.48
```

```
AIC(a9)
```

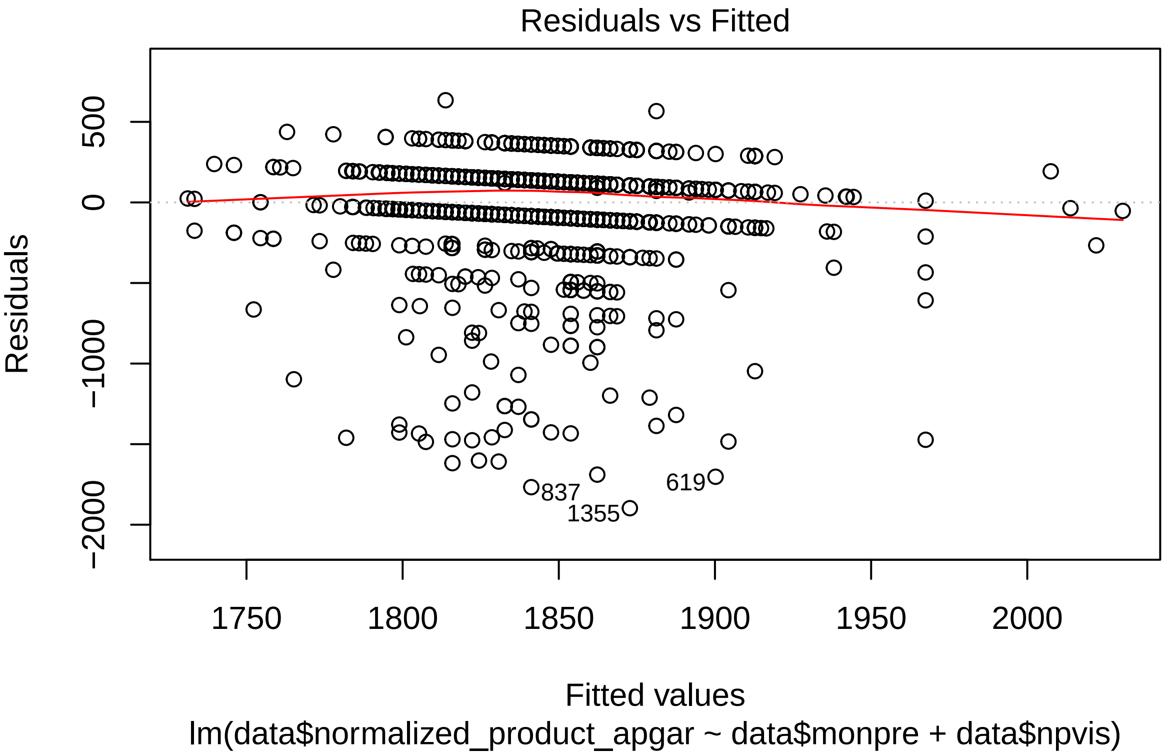
```
## [1] 24899.94
```

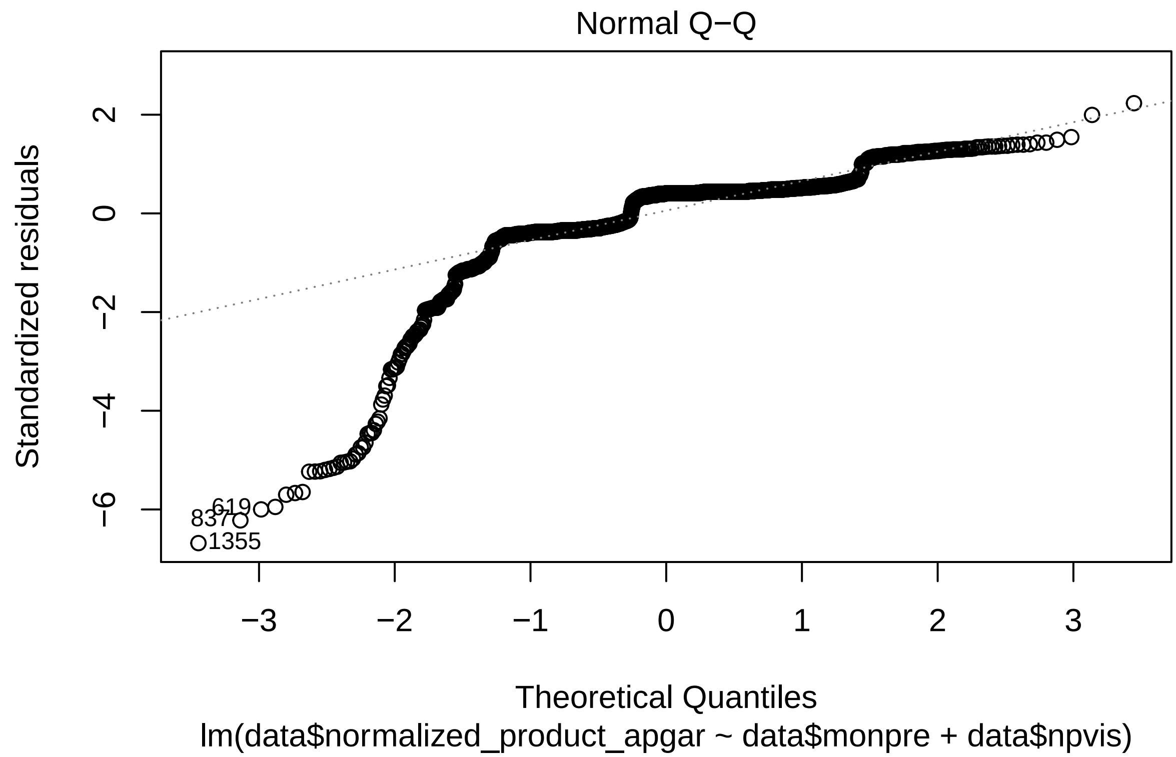
Model a8 has a nominally lower AIC score, so let's continue on with that one.

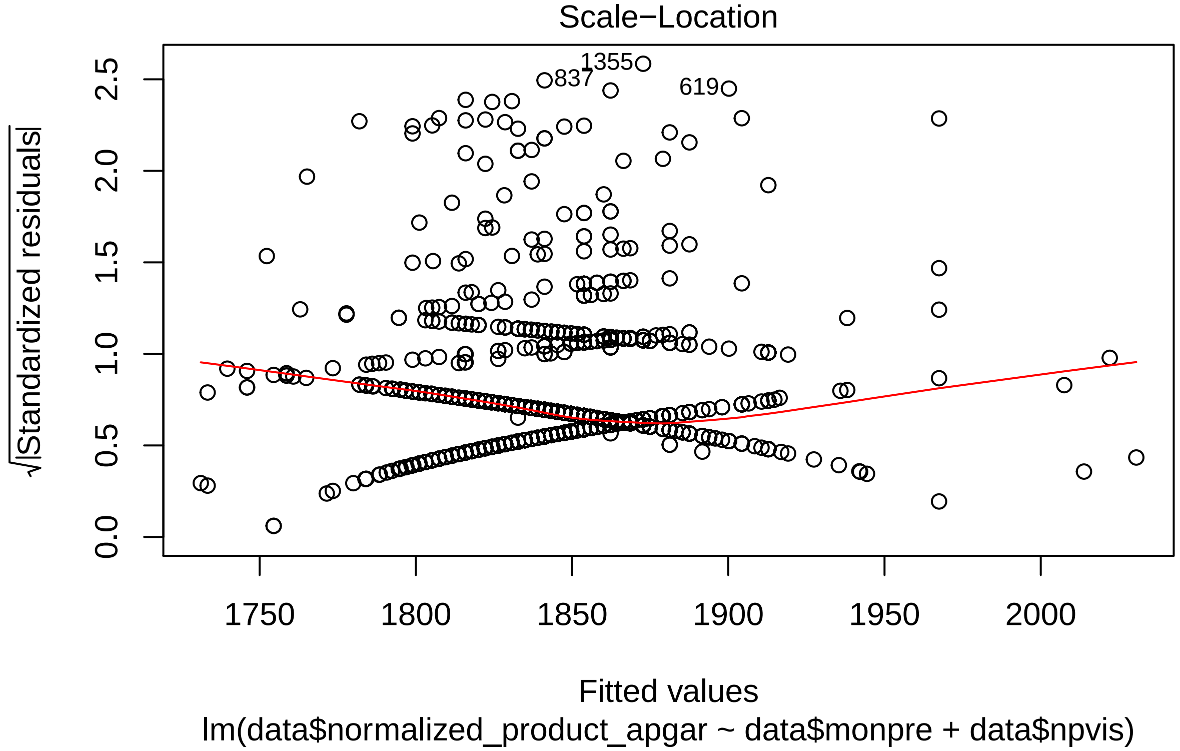
```
summary(a8)
```

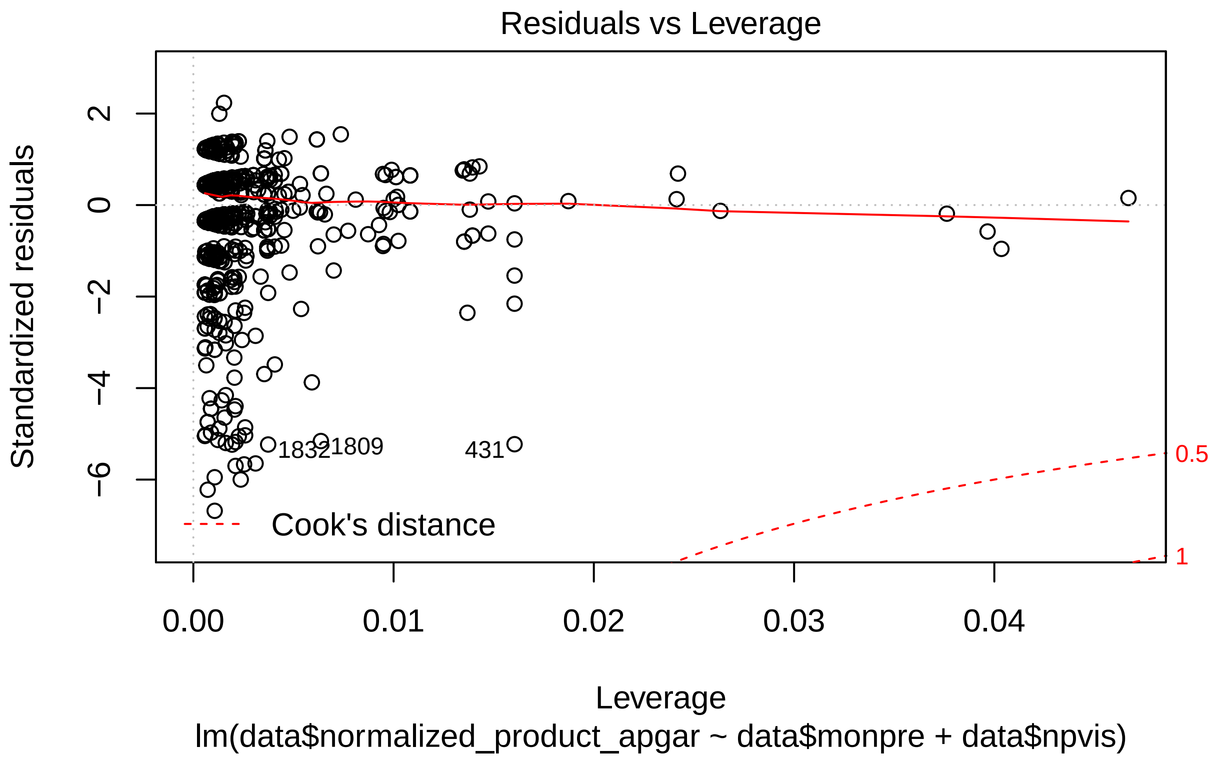
```
##
## Call:
## lm(formula = data$normalized_product_apgar ~ data$monpre + data$npvis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1897.44   -98.29   115.74   130.55   634.08
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1795.067     29.566  60.713  < 2e-16 ***
## data$monpre   -8.502      5.774  -1.472  0.14107
## data$npvis     6.313      1.936   3.261  0.00113 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 284.1 on 1757 degrees of freedom
##   (72 observations deleted due to missingness)
## Multiple R-squared:  0.00981,    Adjusted R-squared:  0.008683
## F-statistic: 8.704 on 2 and 1757 DF,  p-value: 0.0001732
```

```
plot(a8)
```




Normal Q−Q
Standardized residuals
2
0
−2
−4
−6
619
837
1355
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(data$normalized_product_apgar ~ data$monpre + data$npvis)


Scale−Location
√|Standardized residuals|
2.5
2.0
1.5
1.0
0.5
0.0
1355
837
619
1750
1800
1850
1900
1950
2000
Fitted values
lm(data$normalized_product_apgar ~ data$monpre + data$npvis)



We did not see very good results with the APGAR score variations, but as discussed in the introduction, we were expecting the baby's birth weight would have a better indication.

6 CLM assumptions:

1) Linearity in parameters: We can assume this.
2) Random sampling of data: This data is not random because stillbirths are omitted.
3) No perfect co-linearity

As previously stated, our regressors do not have perfect collinearity.

4) Zero conditional mean

Looking at the Residuals vs. Fitted plot above shows that the zero conditional mean is met because the red line is approximately at 0 and has very little curvature.

5) Homoskedacity of errors

From the residuals vs. fitted plot, we can see that we do not have homoskedacity of erorrs because the data is not in an even band across the plot. This means that we'll have to use white standard errors, which are roboust to heteroskadacity.

6) Errors are normally distributed

```
par(mar = rep(2, 4))
shapiro.test(a8$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  a8$residuals
## W = 0.71096, p-value < 2.2e-16
```

From normal Q-Q plot, it looks like our errors are roughly normally distributed except at the very highest and very lowest percentiles. This is to be expected in a dataset such as this.

Using the shapiro wilke test, we can reject the null hypothesis that the population has a normal distribution.

```
library(lmtest)
library(sandwich)
coeftest(a8, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1795.0673    36.0086 49.8510  < 2e-16 ***
## data$monpre   -8.5024     5.7237 -1.4855  0.13760
## data$npvis     6.3128     2.5166  2.5084  0.01222 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model 3: Unbiased Covariants

For our third model, we added in additional features that we think will improve our model, but not introduce bias. We included cigarette consumption and alcohol consumption, which we think would have a negative impact on birthweight. Male babies tend to be heavier than female babies. We've also included mother's and father's age in our model due to scientific research that has found an effect of these variables.

```
model3<-lm(bwght ~ monpre + npvis + cigs + drink + mage + fage +  male, data = data)
```

6 CLM assumptions:

1) Linearity in parameters: We can assume this.

2) Random sampling of data: This data is not random because still births are omitted.

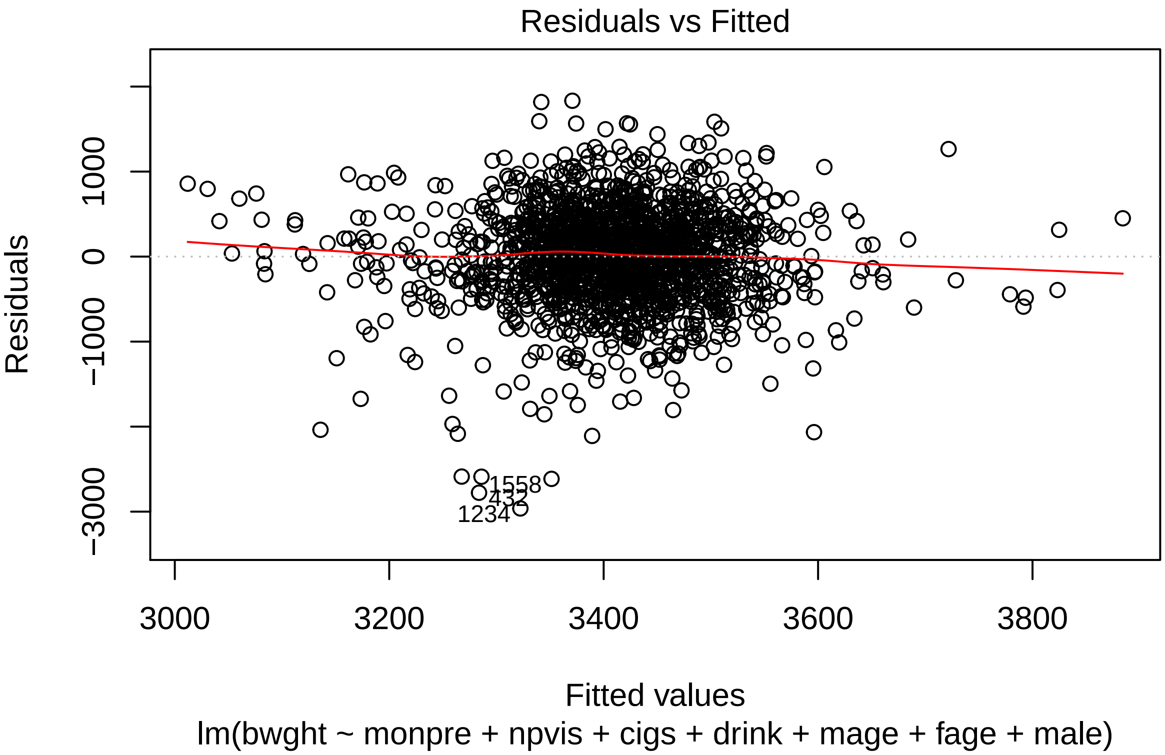
3) No perfect co-linearity: As previously stated, our regressors do not have perfect collinearity.

```
cor(data[,c('monpre', 'npvis', 'cigs', 'drink', 'mage', 'male')], use="complete.obs")
```

```
##             monpre       npvis        cigs        drink         mage
## monpre  1.00000000 -0.31315406  0.09905318 -0.010319741 -0.199115953
## npvis  -0.31315406  1.00000000 -0.03736714  0.052639350  0.096492503
## cigs    0.09905318 -0.03736714  1.00000000  0.185567975 -0.061323113
## drink  -0.01031974  0.05263935  0.18556797  1.000000000  0.004413966
## mage   -0.19911595  0.09649250 -0.06132311  0.004413966  1.000000000
## male   -0.01868132 -0.02185506 -0.01102578 -0.047648827 -0.039928312
##               male
## monpre -0.01868132
## npvis  -0.02185506
## cigs   -0.01102578
## drink  -0.04764883
## mage   -0.03992831
## male    1.00000000
```

4) Zero conditional mean

```r
plot(model3, which=1)
```



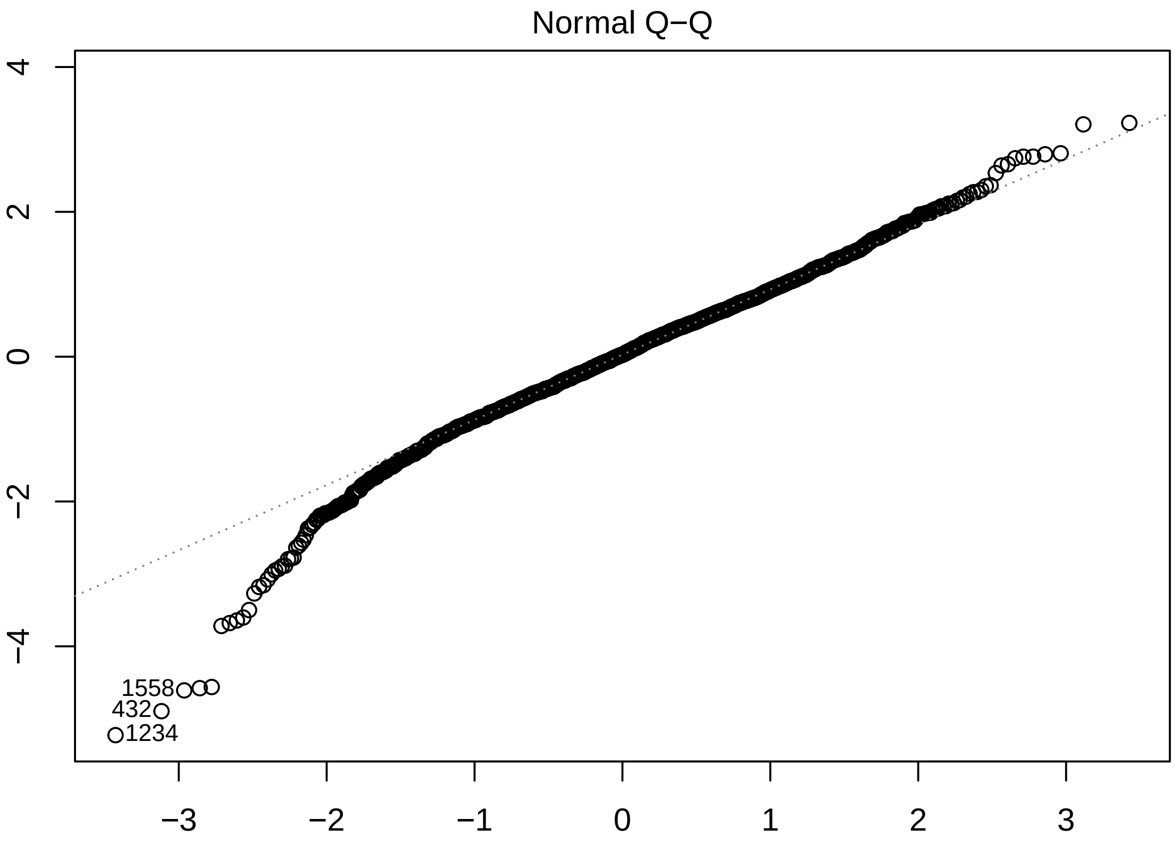
Looking at the Residuals vs. Fitted plot shows that the zero conditional mean is met because the red line is approximately at 0.

5) Homoskedacity of errors

From the residuals vs. fitted plot, we can see that we do not have homoskedacity of erorrs because the data is not in an even band across the plot. This means that we'll have to white standard errors, which are robust to heteroskadacity.

6) Errors are normally distributed

```r
par(mar = rep(2, 4))
plot(model3, which=2)
```



```
shapiro.test(model3$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model3$residuals
## W = 0.97866, p-value = 6.507e-15
```

Checking the normal Q-Q plot, it looks like our errors are roughly normally distributed.

Using the shapiro wilke test, we can reject the null hypothesis that the population has a normal distribution.

```
coeftest(model3, vcov=vcovHC)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 2954.1379   126.7383 23.3090 < 2.2e-16 ***
## monpre        19.3509    12.0294  1.6086 0.1078904
## npvis         15.7507     4.6663  3.3754 0.0007544 ***
## cigs         -10.5228     3.7003 -2.8438 0.0045133 **
## drink        -17.3425    32.2848 -0.5372 0.5912209
## mage          -1.1386     4.4974 -0.2532 0.8001678
## fage           7.4386     3.6162  2.0570 0.0398433 *
## male          77.0481    28.2816  2.7243 0.0065121 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AIC(model1)
```

```
## [1] 27428.8
```

```
AIC(model3)
```

```
## [1] 25498.84
```

Not all of the covariates that we thought would have statistical significance do. Number of prenatal visists, father's age, and the baby being male male all have positive betas. Mother's cigarette consumption has a negative impact on birthweight. Our other covariates including number of prenatal visists, mother's alcohol consumption, and mother's age do not have a statistically signifcant effect on birthweight.

Our AIC for model 3 has gone down from 27428.8 in model 1 to 25498.84 in model 3, which shows that model 3 is a better fit.

Model 4: Problematic Covariants

We will select the attributes of baby's gender and parent's race as well. In the United States, it is a sad fact that minorities such as African Americans do not have adequate access to proper health care as often as non-minorities. Their babies might not fare as well, and their mothers may not get the proper prenatal care.

From all of the summaries, we can tell that the t-statistic for the `monpre` variable is not significant. Thus, we cannot trust this particular regressor, and will omit it from this test.

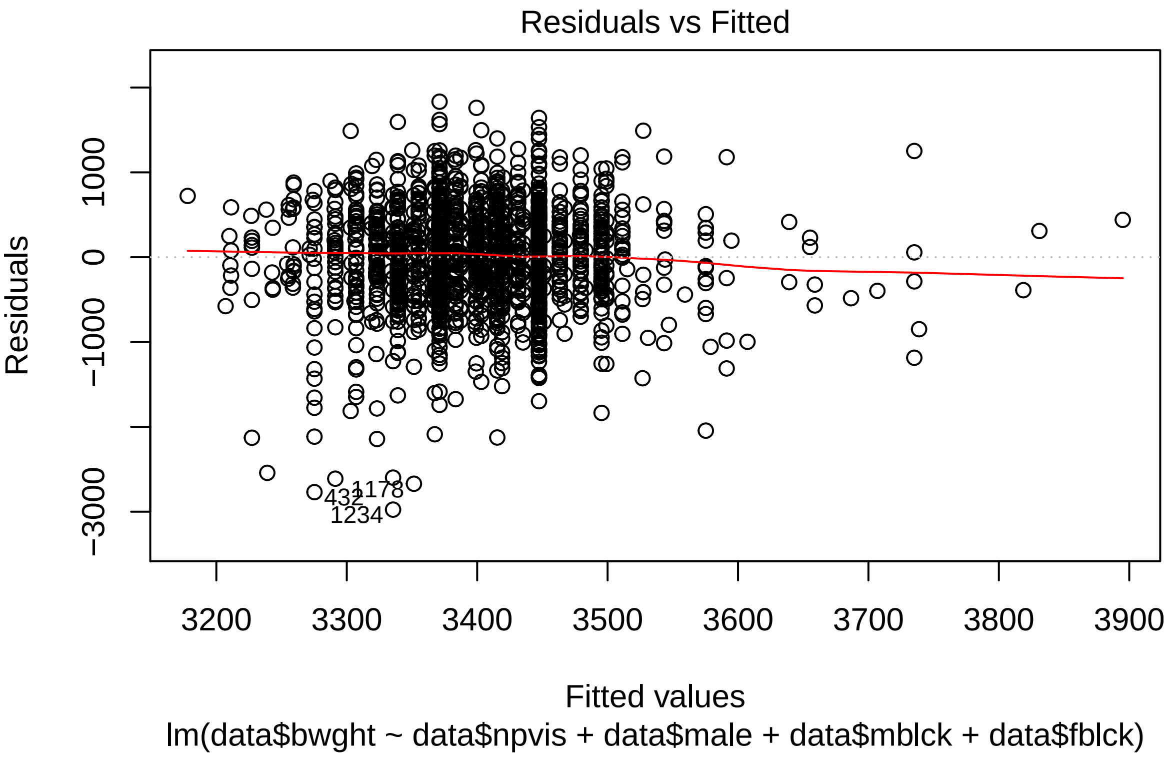
```
c1 = lm(data$bwght ~ data$npvis + data$male +
          data$mblck + data$fblck)

summary(c1)
```

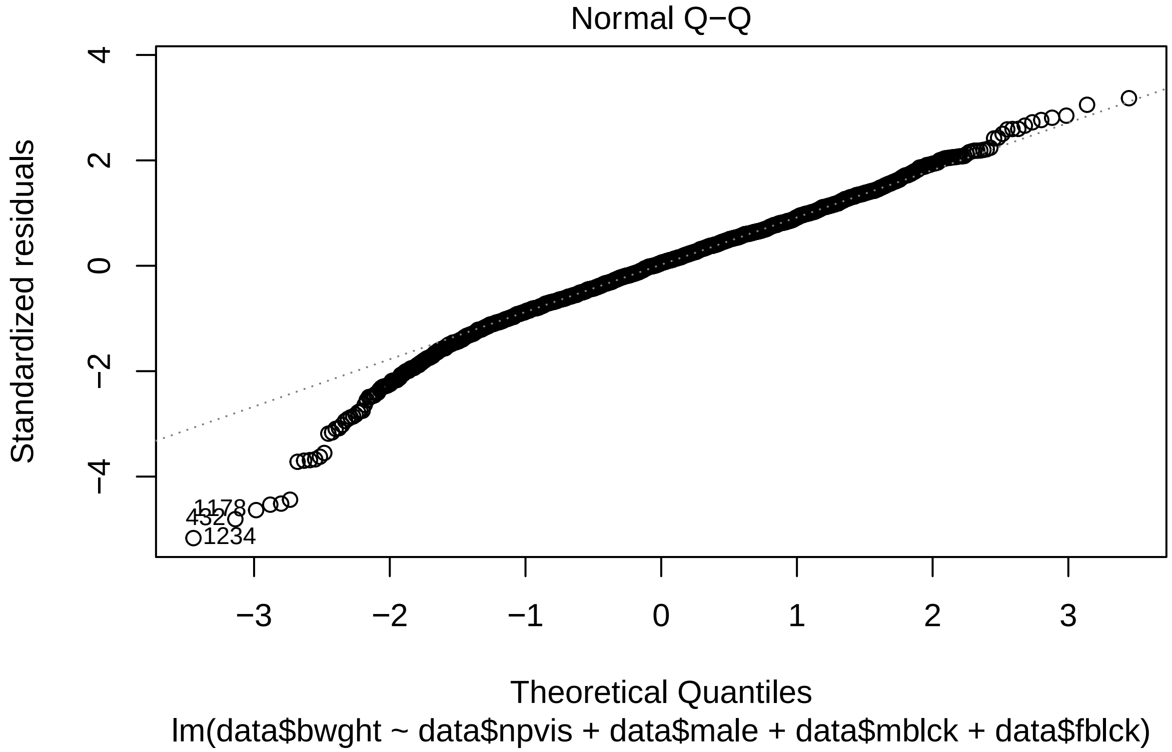
```
##
## Call:
## lm(formula = data$bwght ~ data$npvis + data$male + data$mblck +
##     data$fblck)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2975.51  -336.55    31.69   360.92  1832.85
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3179.315     48.188  65.977  < 2e-16 ***
## data$npvis    15.986      3.735   4.280 1.97e-05 ***
## data$male     76.262     27.534   2.770  0.00567 **
## data$mblck   -97.221    126.174  -0.771  0.44109
## data$fblck    48.729    127.179   0.383  0.70166
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 576.7 on 1759 degrees of freedom
##   (68 observations deleted due to missingness)
## Multiple R-squared:  0.01479,    Adjusted R-squared:  0.01255
## F-statistic:   6.6 on 4 and 1759 DF,  p-value: 2.857e-05
```

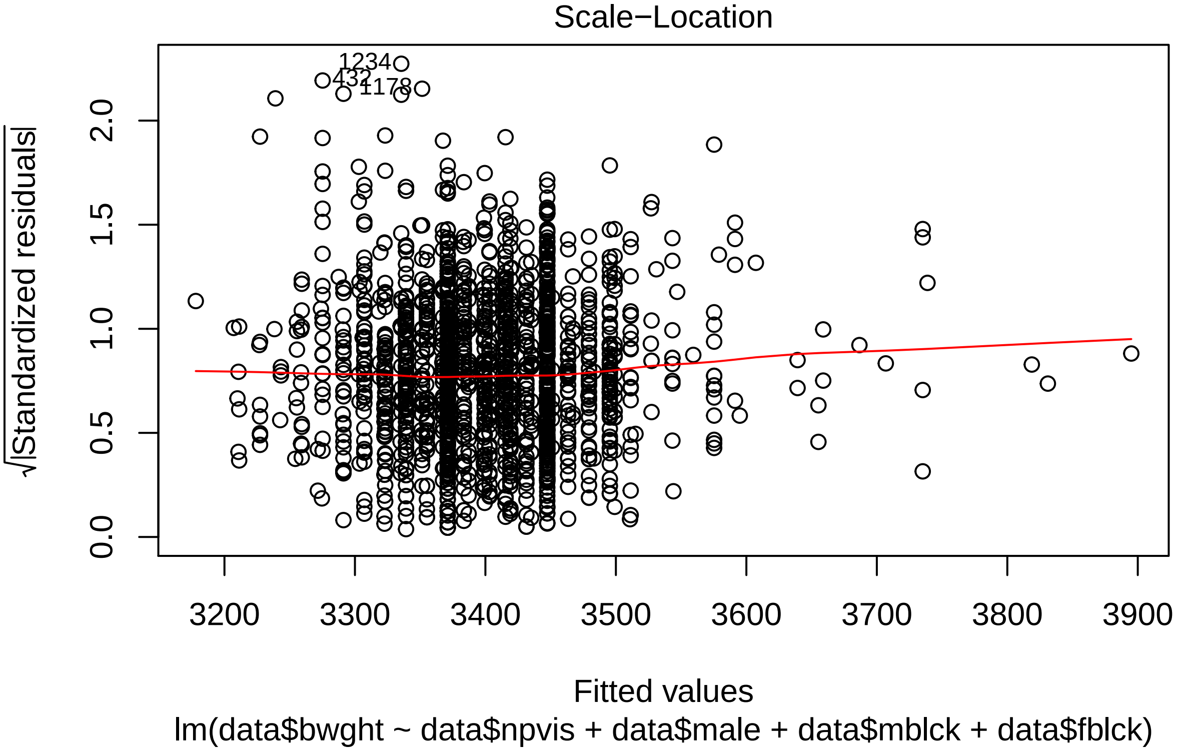
```
AIC(c1)
```

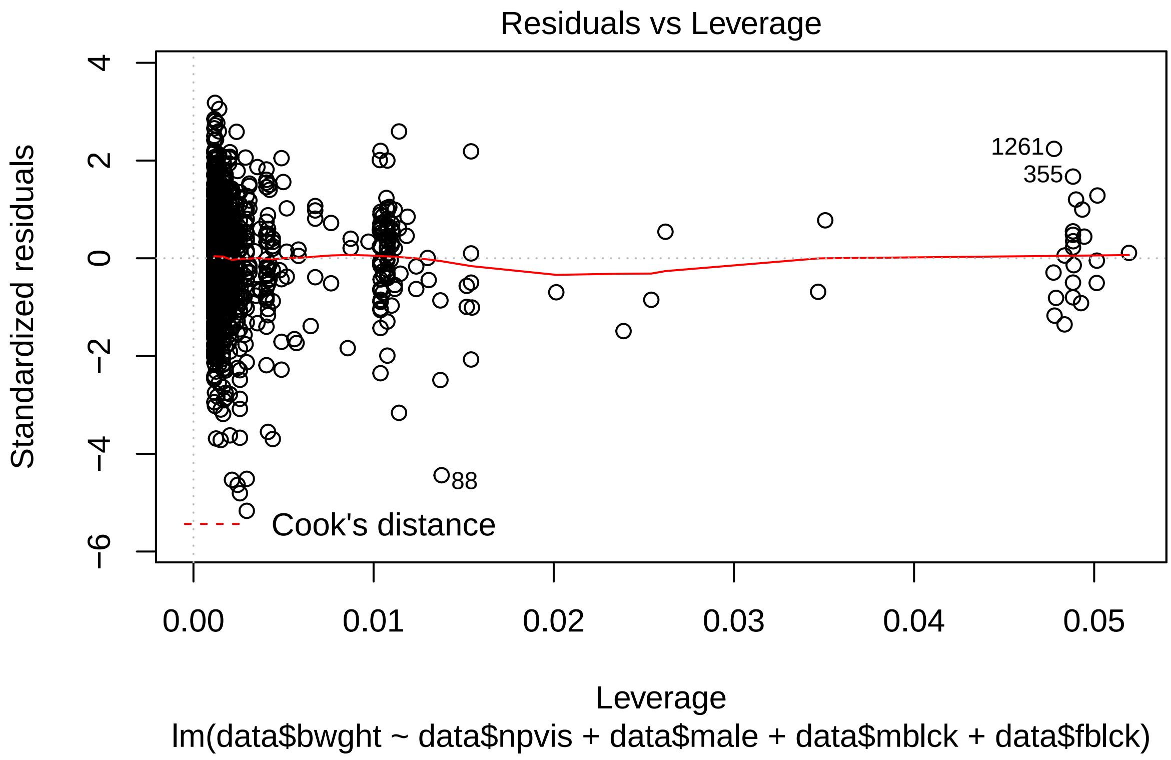
```
## [1] 27441.54
```

```
plot(c1)
```




Normal Q−Q
Standardized residuals
4
2
0
−2
−4
1178
432
1234
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(data$bwght ~ data$npvis + data$male + data$mblck + data$fblck)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
1234
432
1178
3200
3300
3400
3500
3600
3700
3800
3900
Fitted values
lm(data$bwght ~ data$npvis + data$male + data$mblck + data$fblck)



6 CLM assumptions:

1) Linearity in parameters: We can assume this.
2) Random sampling of data: This data is not random because stillbirths are omitted.
3) No perfect co-linearity in regressors:

```
cor(data[,c('npvis', 'mblck', 'fblck', 'male')], use="complete.obs")
```

```
##             npvis       mblck       fblck        male
## npvis  1.00000000 -0.03379275 -0.03133149 -0.02635585
## mblck -0.03379275  1.00000000  0.88963736  0.04743914
## fblck -0.03133149  0.88963736  1.00000000  0.02402644
## male  -0.02635585  0.04743914  0.02402644  1.00000000
```

As previously stated, our regressors do not have perfect collinearity.

4) Zero conditional mean

Looking at the Residuals vs. Fitted plot above shows that the zero conditional mean has not been met because the red line shows curvature for larger babies.

5) Homoskedacity of errors

From the residuals vs. fitted plot, we can see that we do not have homoskedacity of erorrs because the data is not in an even band across the plot. This means that we'll have to use white standard errors, which are roboust to heteroskadacity.

6) Errors are normally distributed

```r
par(mar = rep(2, 4))
shapiro.test(c1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  c1$residuals
## W = 0.97639, p-value < 2.2e-16
```

From normal Q-Q plot, it looks like our errors are roughly normally distributed except at the very lowest percentiles. This is to be expected in a dataset such as this.

Using the shapiro wilke test, we can reject the null hypothesis that the population has a normal distribution.

```r
library(lmtest)
library(sandwich)
coeftest(c1, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 3179.3155    56.2484 56.5228 < 2.2e-16 ***
## data$npvis    15.9863     4.3518  3.6735 0.0002464 ***
## data$male     76.2618    27.4392  2.7793 0.0055056 **
## data$mblck   -97.2213   121.8744 -0.7977 0.4251425
## data$fblck    48.7286   118.4882  0.4113 0.6809371
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we were hoping with such biased data, we can see that the race of the parents is not statistically significant so it is inappropriate to include it in our model.

Step 4: Regression Tables and Model Analysis

```r
library(stargazer)
```

```
##
## Please cite as:

##  Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2. http://CRAN.R-project.org/package=stargazer
```

```
se.model1 = sqrt(diag(vcovHC(model1)))
se.a8 = sqrt(diag(vcovHC(a8)))
se.model3 = sqrt(diag(vcovHC(model3)))
se.c1 = sqrt(diag(vcovHC(c1)))

stargazer(model1,a8,model3,c1, type = "latex", omit.stat = "f",
          se = list(se.model1, se.a8, se.model3, se.c1),
          star.cutoffs = c(0.05, 0.01, 0.001),
          table.placement = '!h')
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sat, Dec 10, 2016 - 20:19:35

See table 1 on the next page.

```
AIC(model1)
```

```
## [1] 27428.8
```

```
AIC(model3)
```

```
## [1] 25498.84
```

```
AIC(c1)
```

```
## [1] 27441.54
```

From the Akaike Information Criterion test, we see that `model3` is the best option for a linear model predicting the health of the baby. Model3 has the highest adjusted Rˆ2, showing that virutally 2% of all variability in the baby's health indicators can be determined by the months prenatal visits started, number of prenatal visits, the mother's smoking and driking habits, the mother's age, and the baby's gender. As always, `monpre` was not a statistically significant regressor, and neither was the mother's age or drinking habits. In other words for practical significance, we can say if the baby is a boy we can expect he will weigh 80.937 grams more than if he is a girl, for every year older his mother is, he will weight 5.317 grams more, for every alcoholic drink his mother inbibes per week he will weigh 14.050 grams less, for every cigarette his mother smokes per day, he will weigh 11.229 grams less, for each prenatal visit, he will weight 15.505 more, and for each month the mother waits to to start her prenatal care, the baby weight 20.901 grams more. Just writing it out what the model means stresses even more that we should ignore the `monpre` variable in modeling a baby's health.

Table 1:

	Dependent variable:			
	bwght	normalized_product_apgar	bwght	bwght
	(1)	(2)	(3)	(4)
monpre	17.062 (12.028)		19.351 (12.029)	
npvis	17.549*** (4.834)		15.751*** (4.666)	
monpre		$-$8.502 (5.724)		
npvis		6.313* (2.517)		15.986*** (4.352)
cigs			$-$10.523** (3.700)	
drink			$-$17.343 (32.285)	
mage			$-$1.139 (4.497)	
fage			7.439* (3.616)	
male			77.048** (28.282)	
male				76.262** (27.439)
mblck				$-$97.221 (121.874)
fblck				48.729 (118.488)
Constant	3,161.271*** (74.605)	1,795.067*** (36.009)	2,954.138*** (126.738)	3,179.315*** (56.248)
Observations	1,763	1,760	1,642	1,764
R^2	0.011	0.010	0.024	0.015
Adjusted R^2	0.010	0.009	0.020	0.013
Residual Std. Error	577.470 (df = 1760)	284.115 (df = 1757)	568.265 (df = 1634)	576.683 (df = 1759)
Note:				*p<0.05; **p<0.01; ***p<0.001

Step 5: Causality

We cannot claim causality in our model. We have an strong correlation, but we are hesitant to call our models causal because of omitted variables. There are many other factors that influence birthweight that are not captured in this data set, which leads to omitted variable bias. We know we have ommitted variable bias because our R-squared value of all of our models is very low at around 1%- 2%. We will provide a few examples of potential omitted variables with analysis and then list a few more.

1) Nutrition Mother's nutrition is likely to have an impact on birthweight. If a mother has good nutrition she will also likely have good prenatal care. Nutrition would have a positive influence on birthweight. This means that our model is overstating the value of prenatal care on birthweight. Our beta in our model for prenatal care in this case is higher than it should be.
2) Birth order Birth weight increases with increasing birth order. [6] However, prenatal care decreases with birth order. Since we do not have birth order in our model, our model is likely underpredicting the effect of prenatal care on birthweight.
3) Socioeconomic status Higher socioeconomic status is likely to result in higher birthweight. Higher socioeconomic status is likely to be correlated to having better access to prenatal care. Therefore our model is likely overpredicting the effect of prenatal care on birthweight because we are not including socioeconomic status in our model.

Other potential omitted variables include genetics, weight of mother, age of mother's first child, number of children (e.g. twins, etc.). We will not do a full analysis of all of these other potential omitted variables, but want to highlight the fact that there are many potential omitted variables. Because of all of the omitted variables, we cannot claim that our model is causal.

Biases and Limitation

This data is extremely biased in that no still births were included in our dataset. It is a sad fact in the United States that over 2 in 1,000 births are stillbirths[5]. Since we do not know the prenatal care data for stillbirths, we cannot completely guage how much prenatal care contributes to a child's health at birth.

In addition, it appears that there is little correlation between the Apgar score and the later health of the baby. The Apar is only meant to be used in the context of emergency situations. In this manner, looking at a baby's weight will give us deeper insight into the baby's overall health.

No miscarriages were included in the data, so this further biases our data.

Using birthweight as a proxy for infant health was the best that we could do given our data set, but is by no means a comprehensive view on an infants' health.

In section 4, we talked about variables that could absorb the effect of prenatal care in our model. These variables could include race.

Step 7: Conclusion

Through our data analysis, we saw that prenatal care has a statistically significant positive effect on infant health. This analysis was operationalized in our data set with prenatal care as number of prenatal care visits and birthweight as representing infant health. While these ways of operationalizing infant health and prenatal care aren't ideal, they were the best options given the data set.

Other factors influence birthweight in a positive direction–including being male and in a negative direction–including mother's cigarette consumption.

Even with our best model, we had a very low r-squared value. This indicated to us that there were omitted variables.

While doing a randomized controlled trial would be unethical in assigning some mothers as not receiving prenatal care, we think using different statistical techniques and with further research that we can potentially establish a causal link between prenatal health and infant health outcomes. This research design is beyond the scope of this paper, but may include things like phasing in free prenatal care to different communities that do not have access in order to measure the effect.

In conclusion, while in our model number of prenatal care visits was a statistically signifcant regressor, we cannot prove causality of our models. We think there may be significant omitted variables that prevent our model from being causal. There is a strong association, so we would still recommend prenatal care for mothers.

References

[1]https://www.nichd.nih.gov/health/topics/pregnancy/conditioninfo/pages/prenatal-care.aspx

[2]https://www.ncbi.nlm.nih.gov/pubmed/7543353

[2a]http://sciencenordic.com/birth-weight-predicts-brain-development

[3]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1595023/

[4]http://ije.oxfordjournals.org/content/30/6/1233.long

[5]https://www.washingtonpost.com/news/wonk/wp/2014/09/29/our-infant-mortality-rate-is-a-national-embarrassment/?utm_term=.58dedfd178fd

[6]https://www.ncbi.nlm.nih.gov/pubmed/3260664

[7]https://www.ncbi.nlm.nih.gov/pubmed/25108692