Lab 4: Healthy Momma, Healthy Baby

Krista Mar and Nikki Haas

12/1/2016

A Nice Introduction that Makes Us Sound Like Pros

According to the NIH, having a healthy pregancy is one of the best ways to promote a healthy birth and that getting early and regular prenatal care improves the chances of a healthy pregnancy.[1] According to Hack et all, while most low birth weight children will end up having normal outcomes, as a group they generally have more health issues than healthy weight babies[2].

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.

According to Montgomery, the 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.

Something about higher birthweight that talks about neural development big babies, big brains

Step 1: Read in the Data

```r
load('/Users/nicholeh/student285/w203/w203_lab_4/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.

```r
library(ggplot2)
cor(data$omaps, data$fmaps, use = "complete.obs")
```

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

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

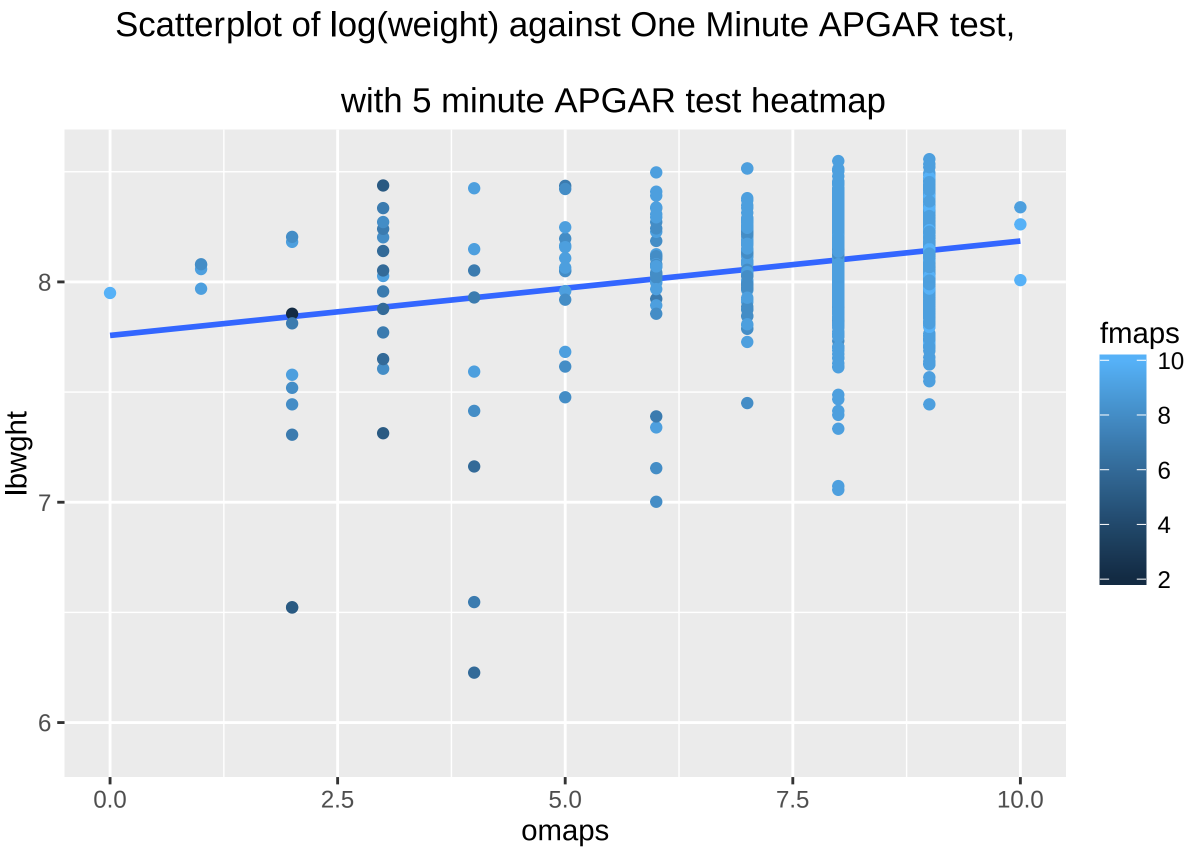
```
## [1] 0.2710456
```

```r
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
```

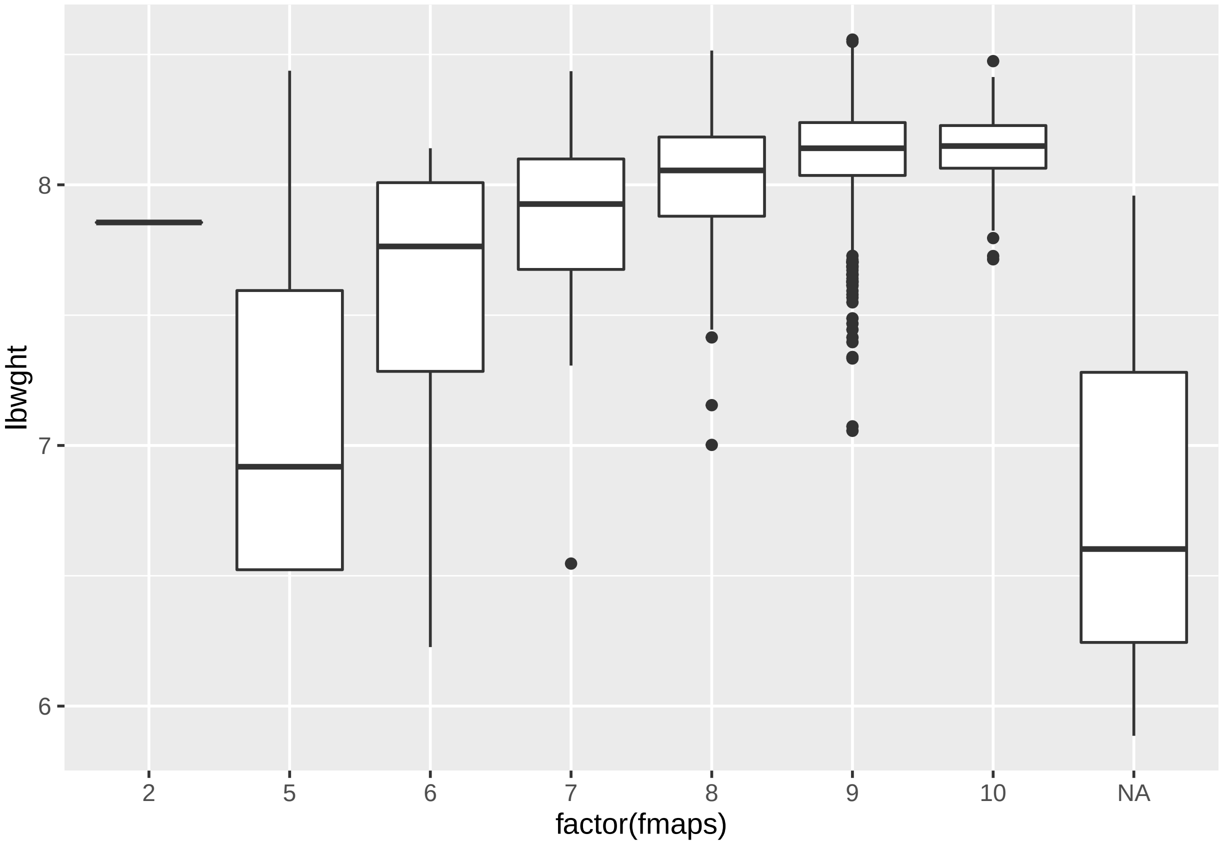
```
## Warning: Removed 3 rows containing non-finite values (stat_smooth).
```

```
## 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")
```

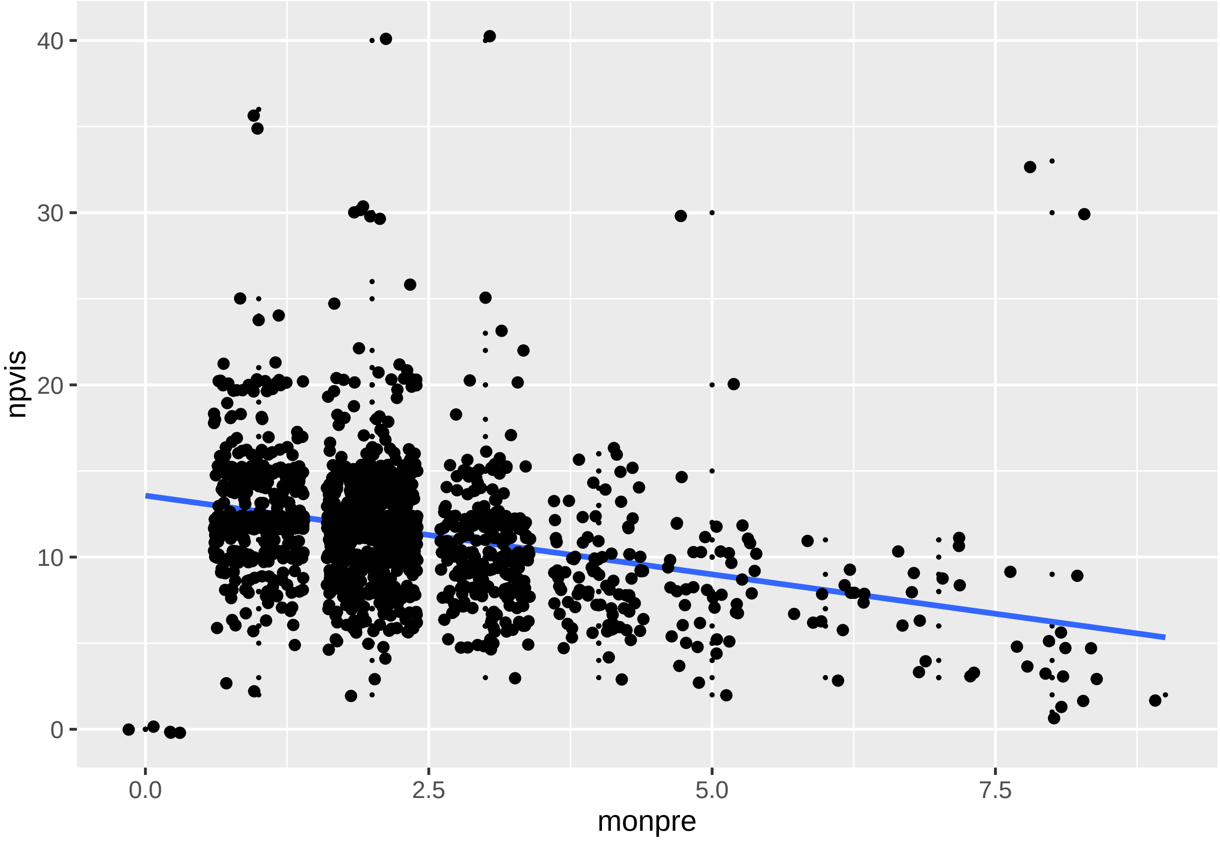
```
## [1] -0.3061006
```

```
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).
```

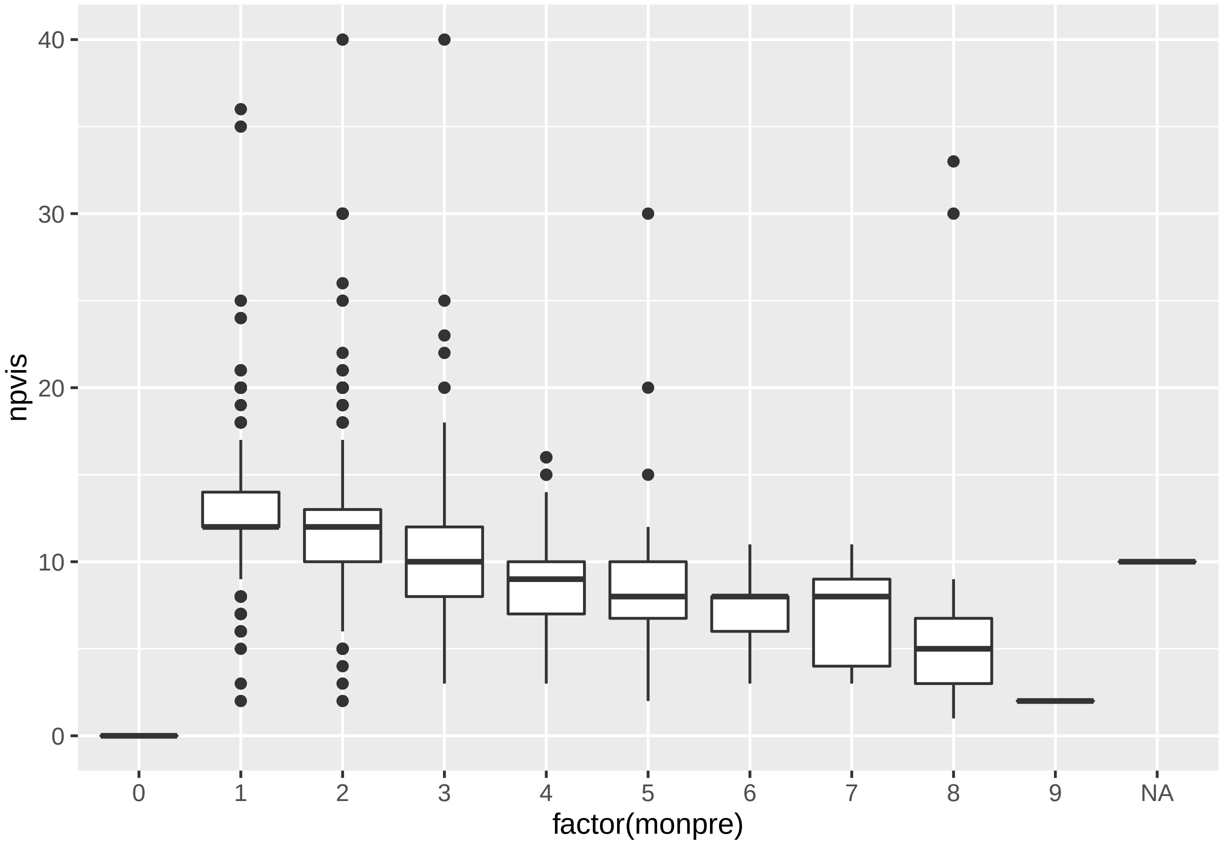
```
## Warning: Removed 69 rows containing missing values (geom_point).
```

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



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

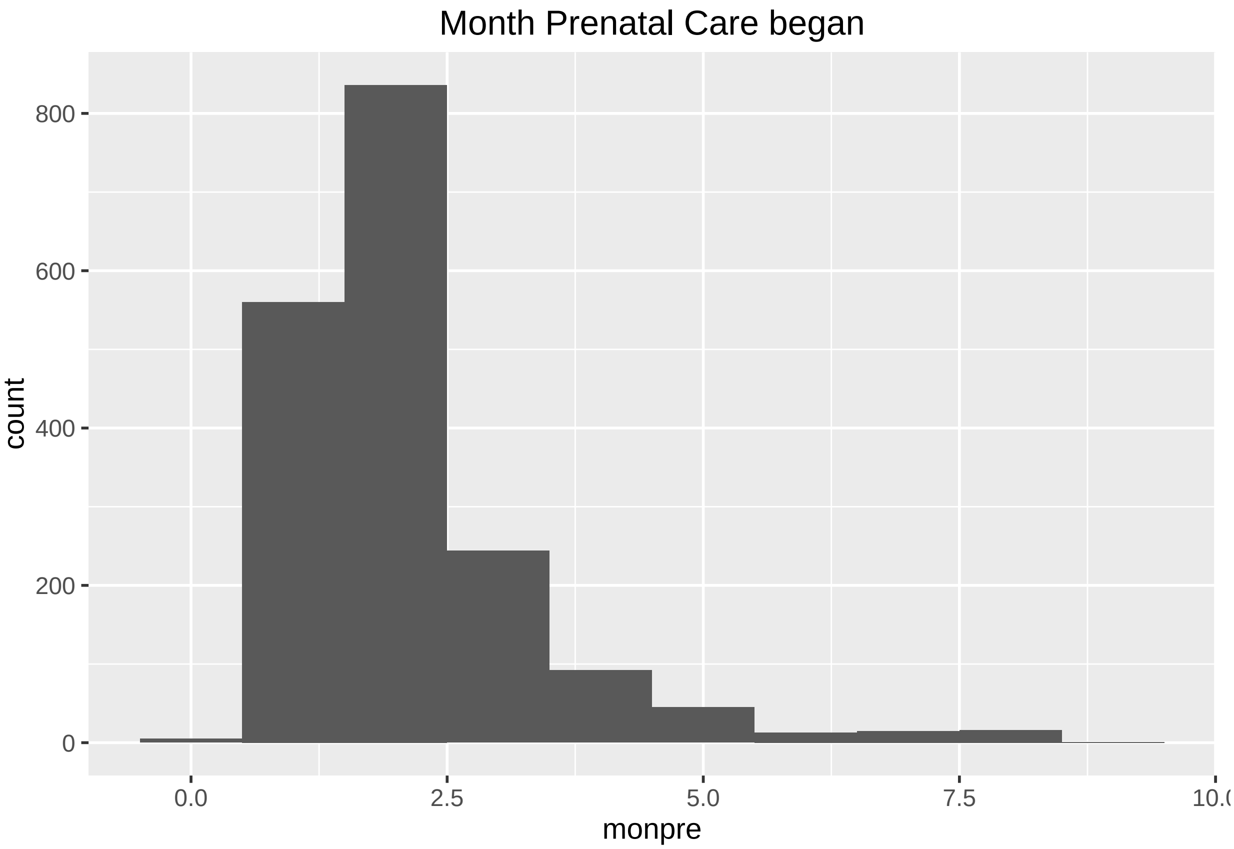
```
## Warning: Removed 68 rows containing non-finite values (stat_boxplot).
```



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.

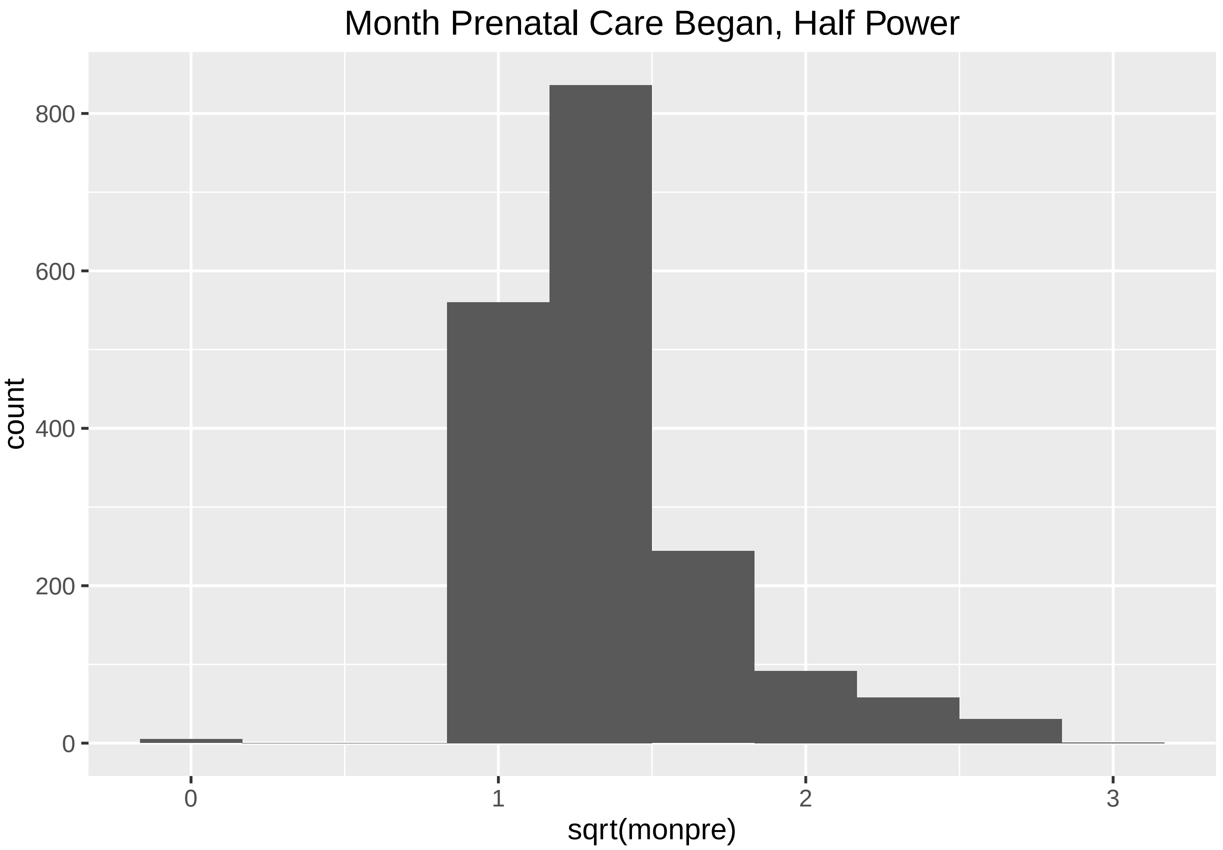
```
ggplot(data, aes(x=monpre)) + geom_histogram(aes(y = ..count..),bins = 10) +
  ggtitle("Month Prenatal Care began")
```

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



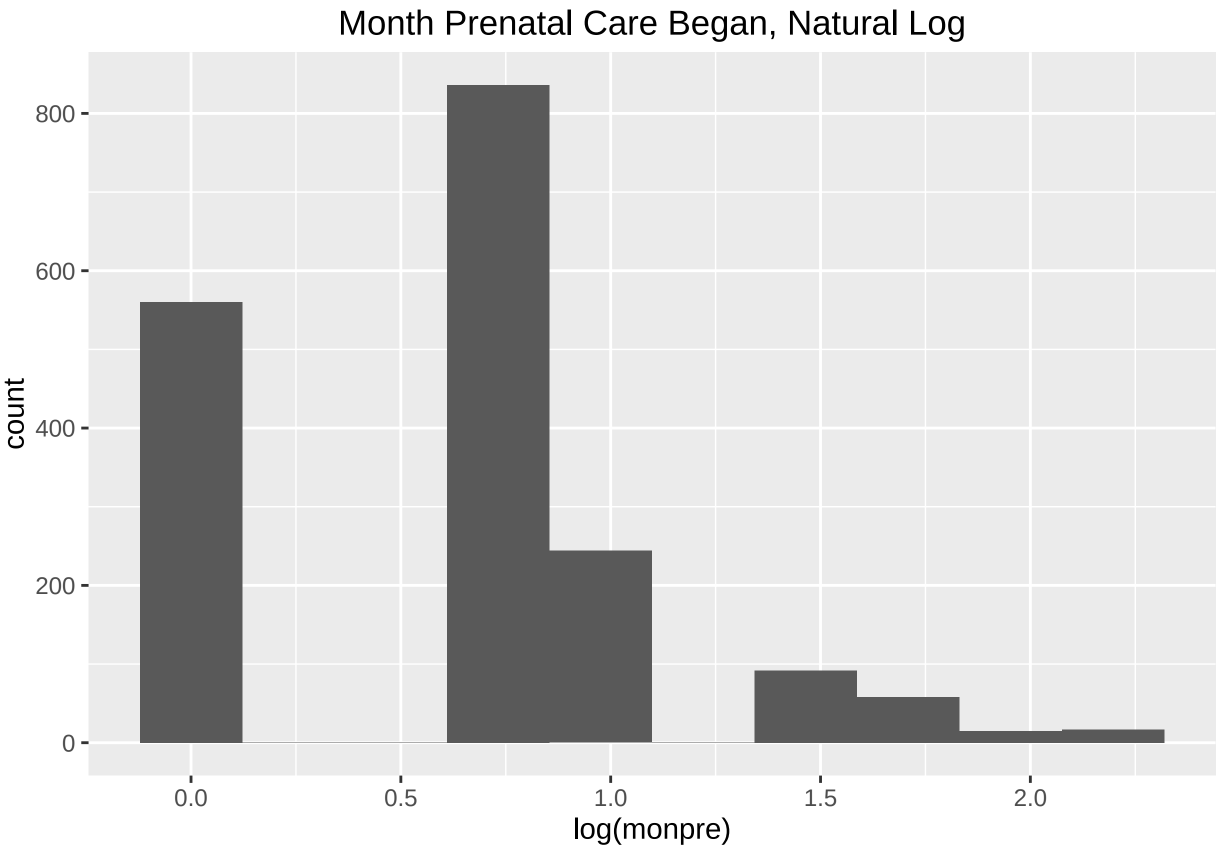
```
ggplot(data, aes(x=sqrt(monpre))) + geom_histogram(aes(y = ..count..), bins = 10) +
  ggtitle("Month Prenatal Care Began, Half Power")
```

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



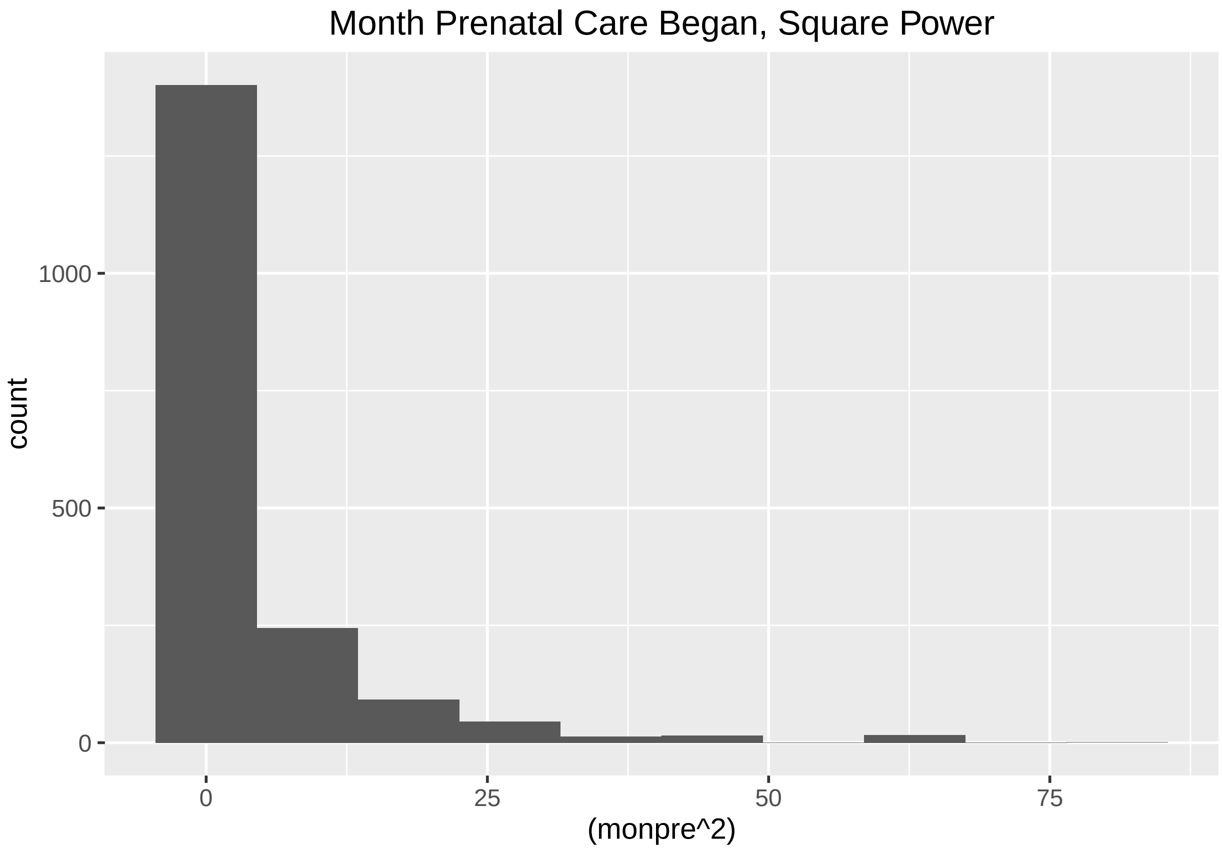
```
ggplot(data, aes(x=log(monpre))) + geom_histogram(aes(y = ..count..), bins = 10) +
  ggtitle("Month Prenatal Care Began, Natural Log")
```

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



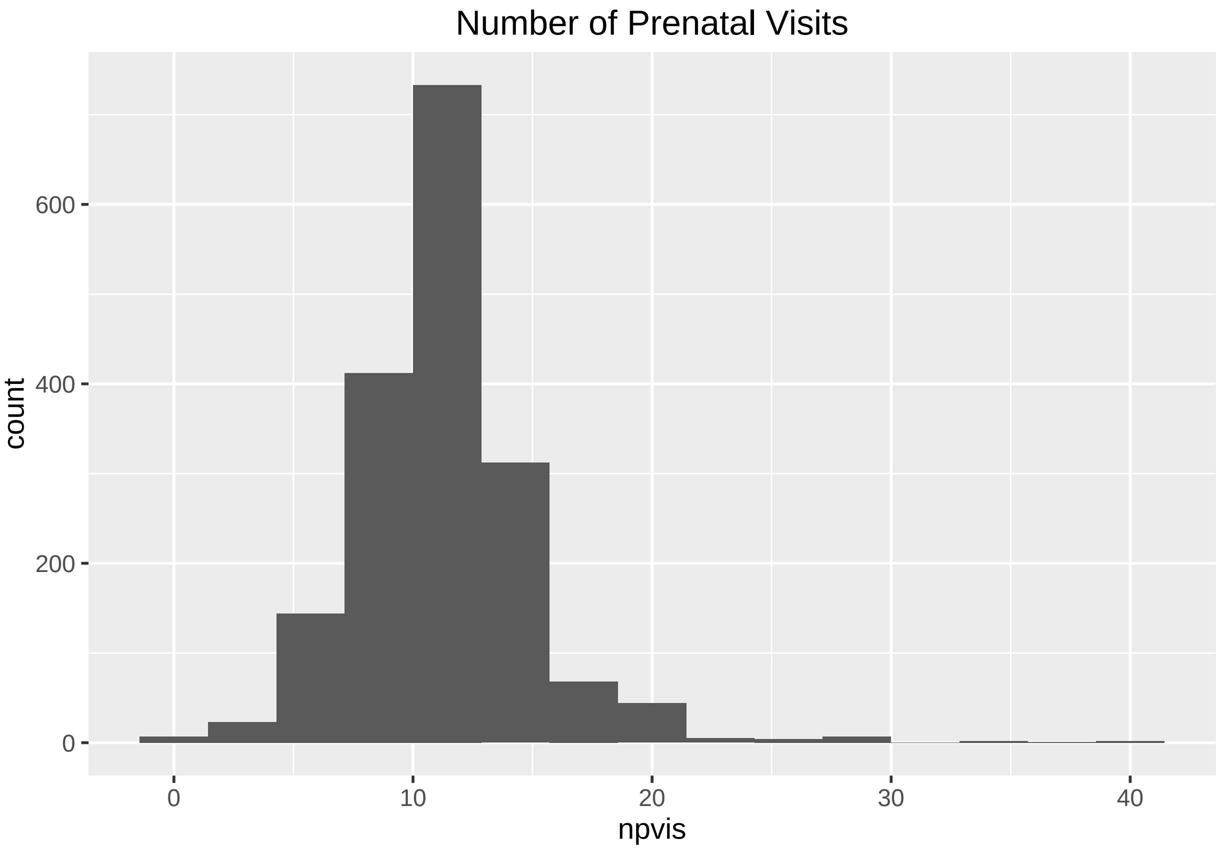
```
ggplot(data, aes(x=(monpre^2))) + geom_histogram(aes(y = ..count..), bins = 10) +
  ggtitle("Month Prenatal Care Began, Square Power")
```

```
## 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")
```

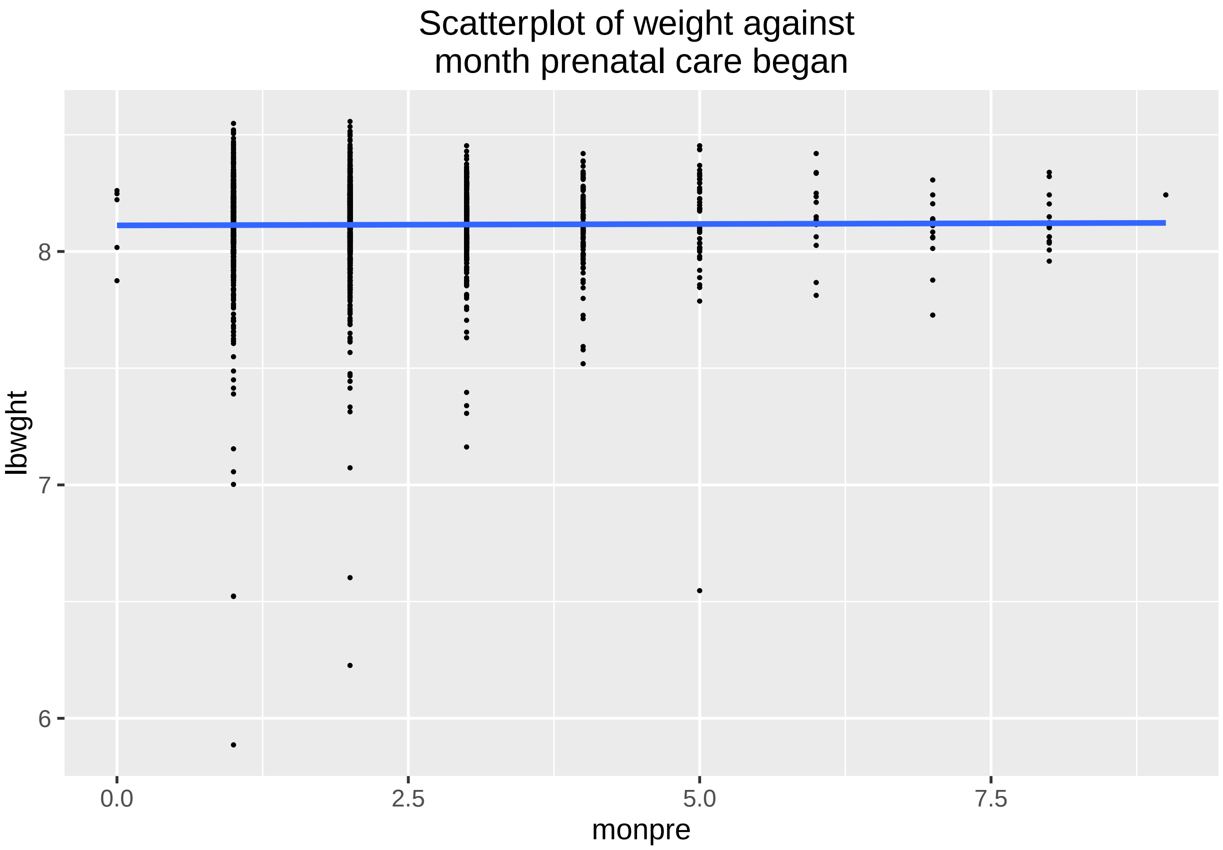
```
## Warning: Removed 68 rows containing non-finite values (stat_bin).
```



```r
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

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 or miscarriages.
3) No perfect co-linearity

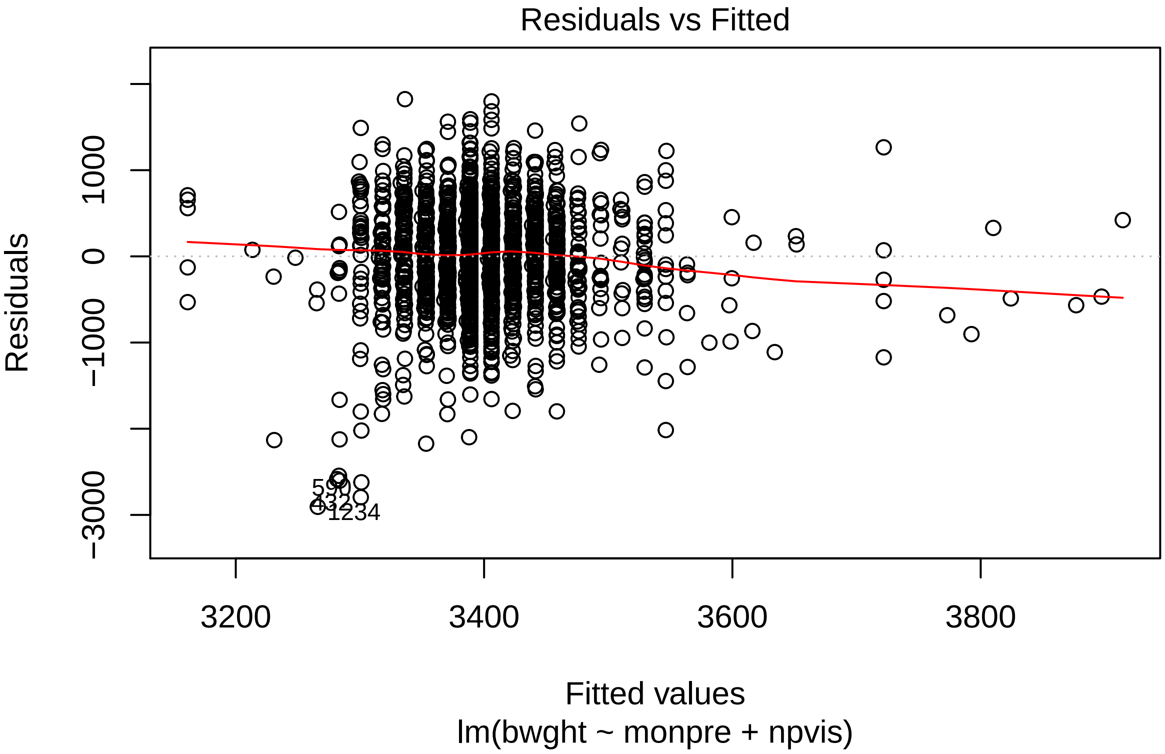
```
cor(data$monpre, data$npvis, use="complete.obs")
```

```
## [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 whcih prenatal care started.

4) Zero conditional mean

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



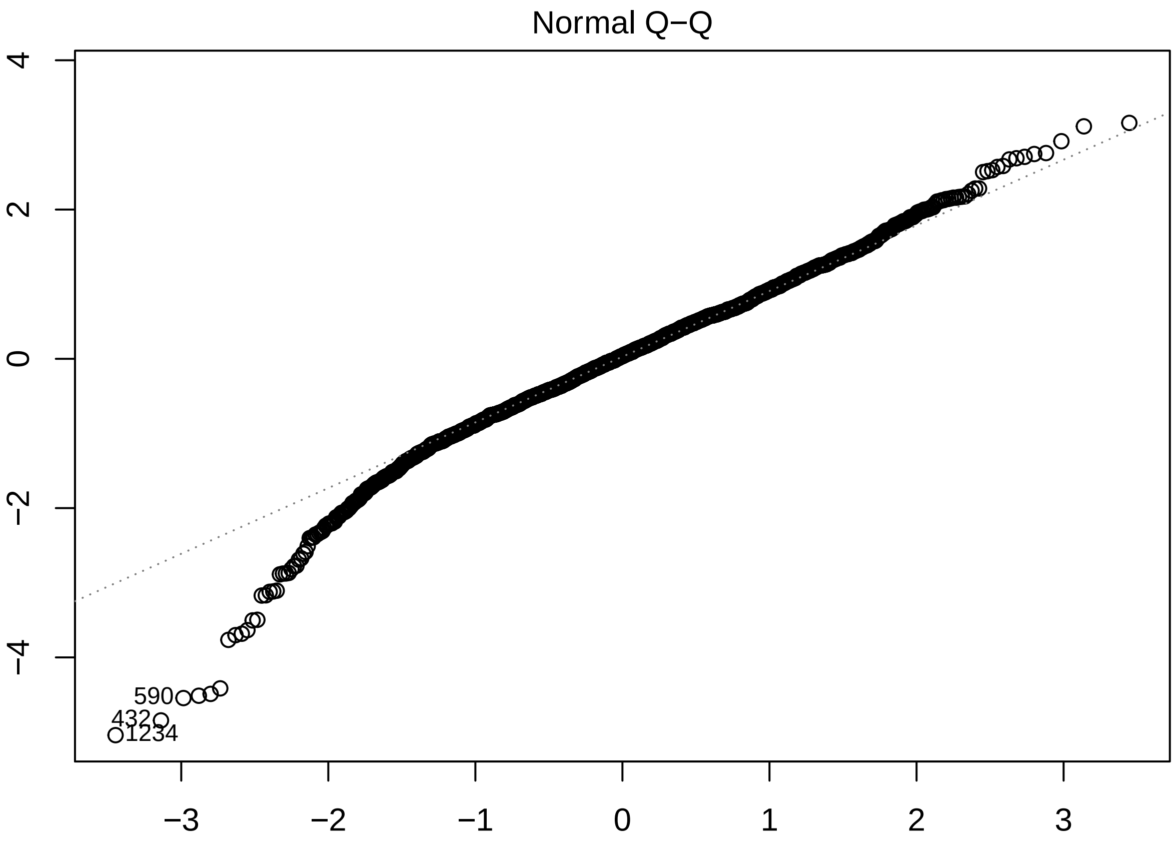
Looking at the Residuals vs. Fitted plot shows that the zero conditional mean is met because the red line is approximately at 0.

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.

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
```

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)
```

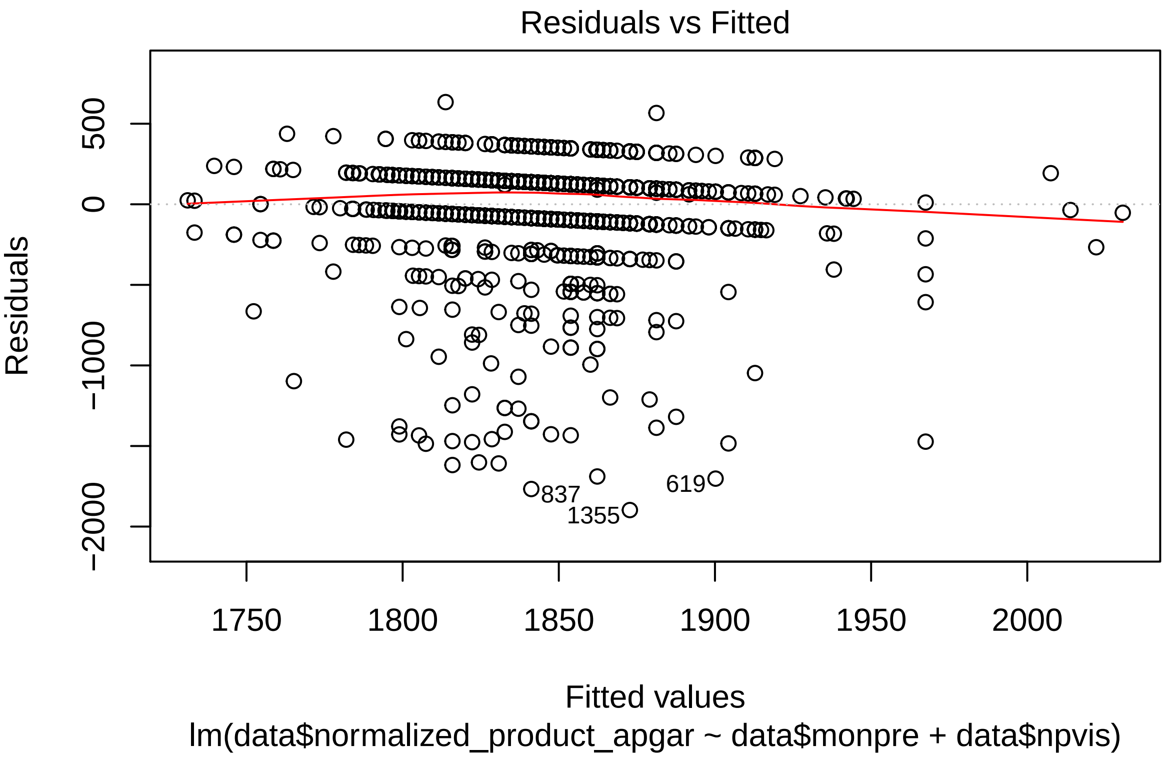
```
## [1] 24899.94
```

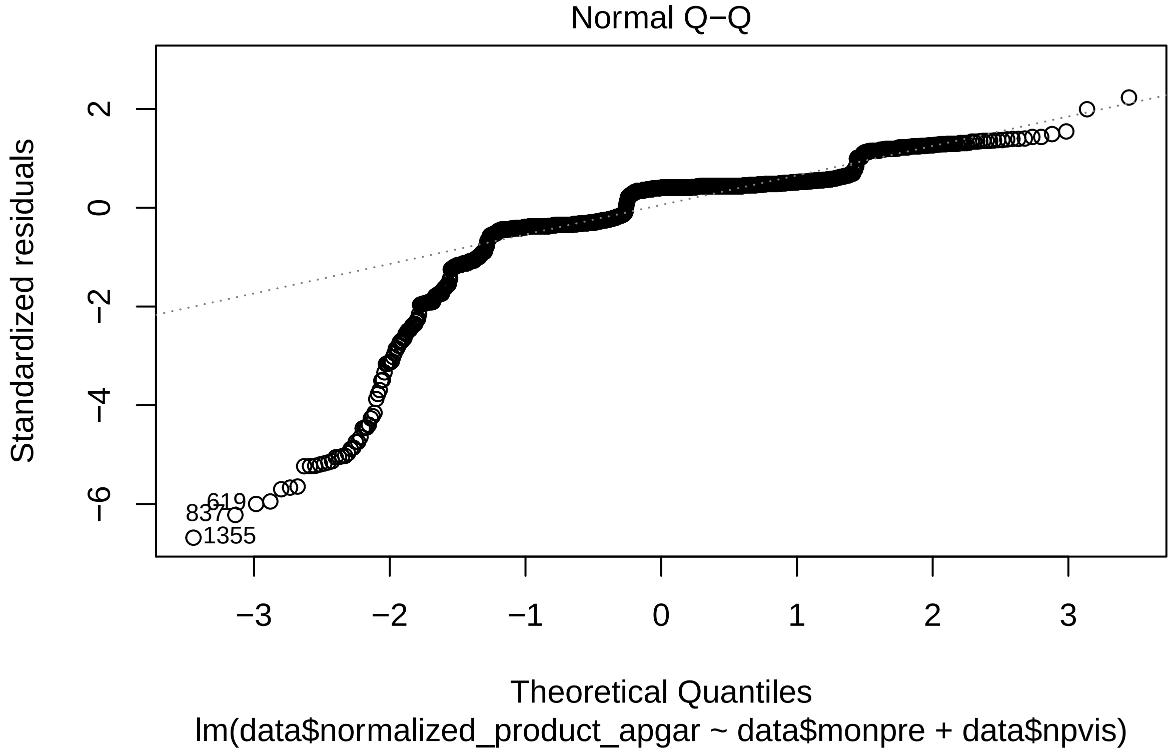
Model a8 has a nominally lower AIC score, so let's continue on with that one.

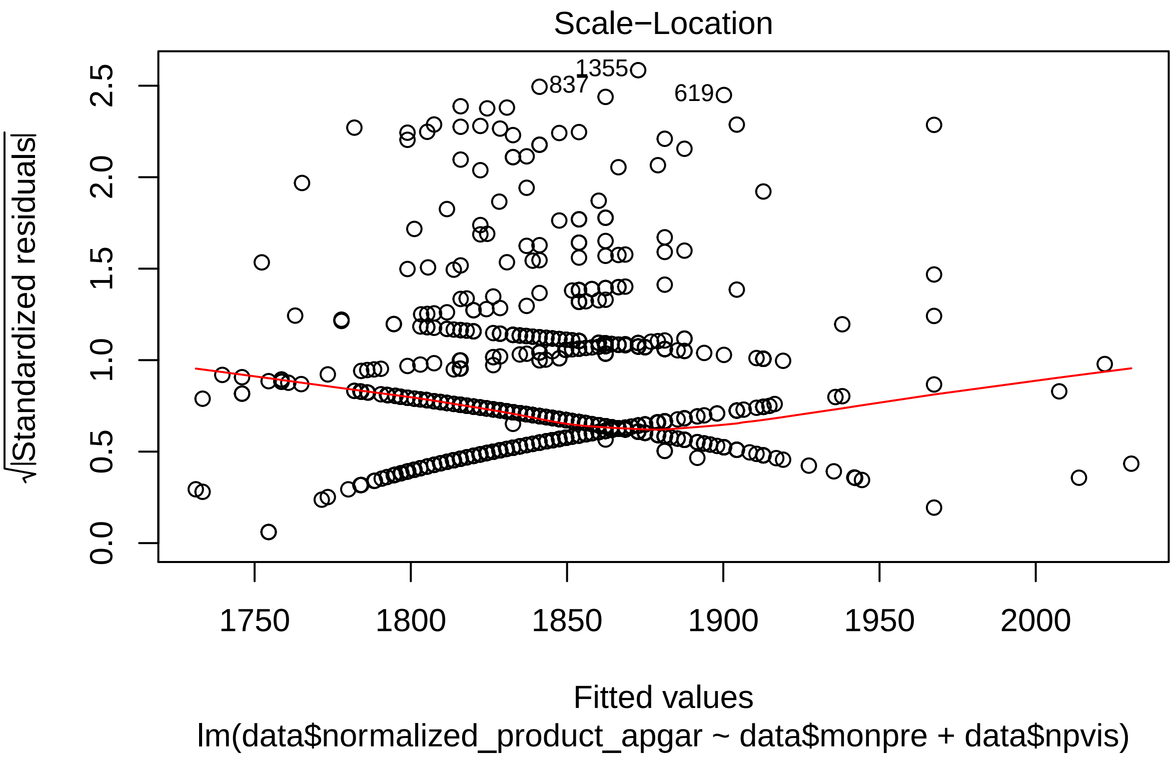
```
summary(a8)
```

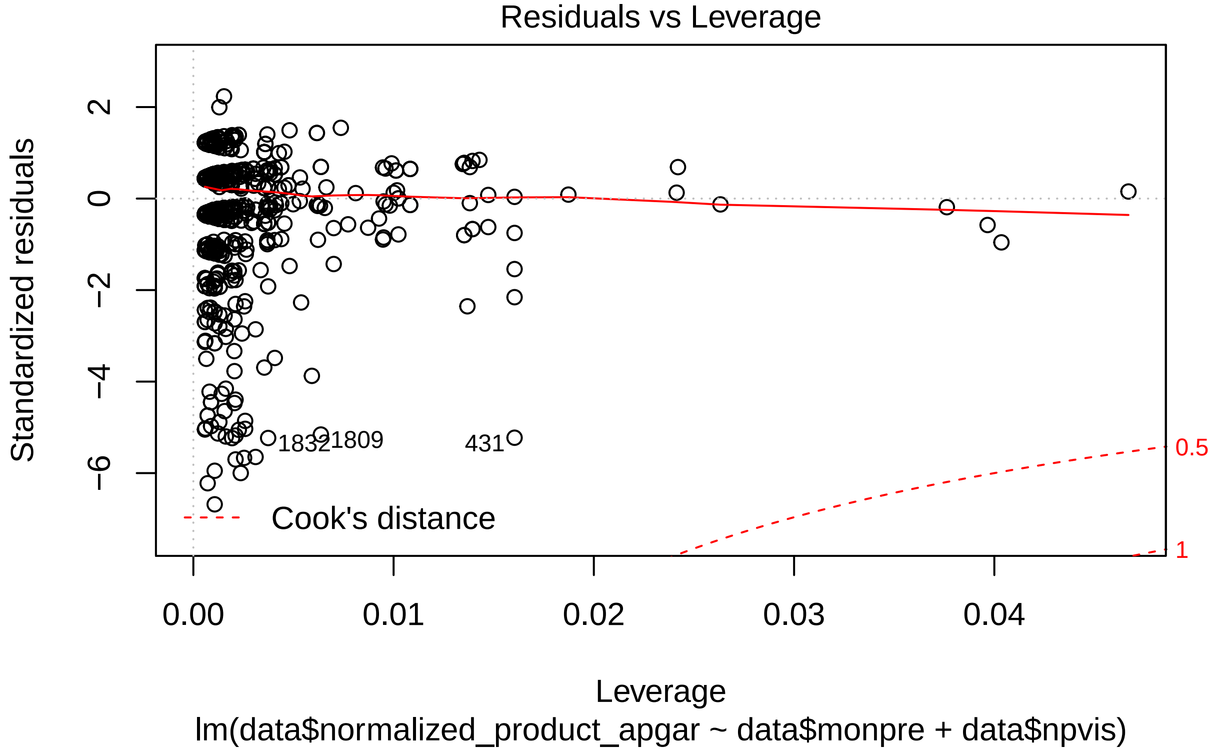
```
##
## Call:
## lm(formula = data$normalized_product_apgar ~ data$monpre + data$npvis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1897.44   -98.29   115.74   130.55   634.08
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1795.067     29.566  60.713  < 2e-16 ***
## data$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
```

```r
plot(a8)
```









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

```
model3<-lm(bwght ~ monpre + npvis + cigs + drink + mage + 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 stillbirths are omitted.

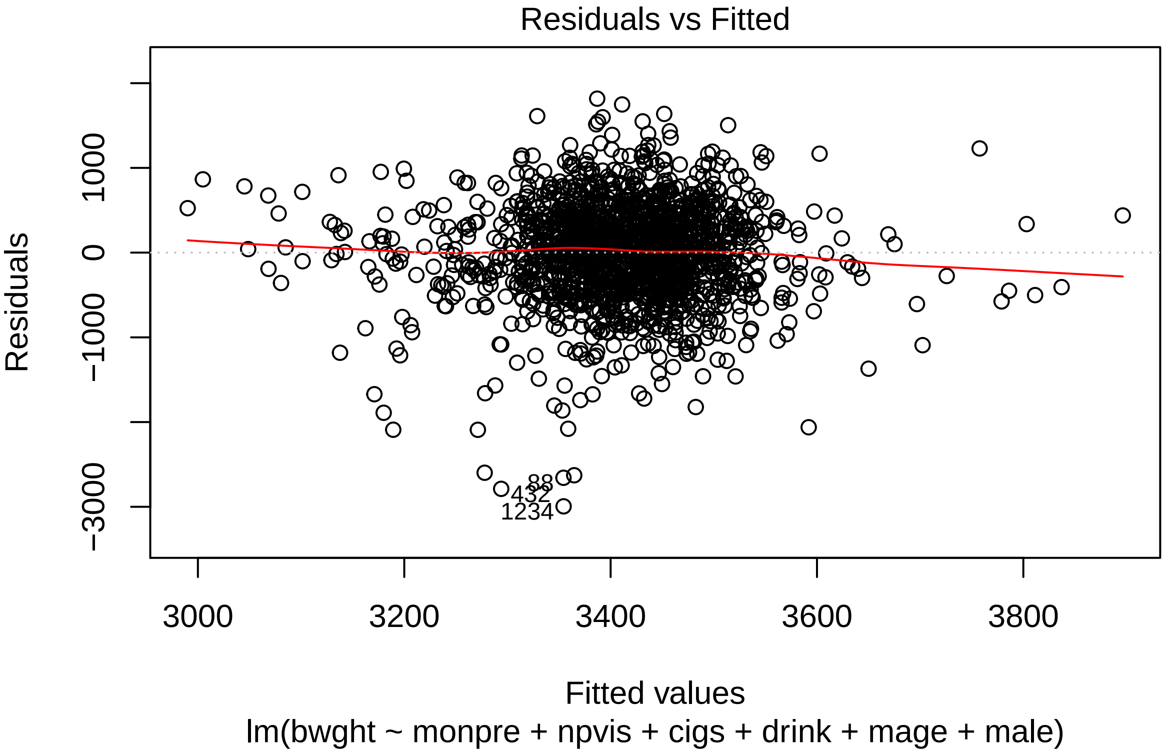
3) No perfect co-linearity: As previously stated, our regressors do not have perfect collinearity.

```
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

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



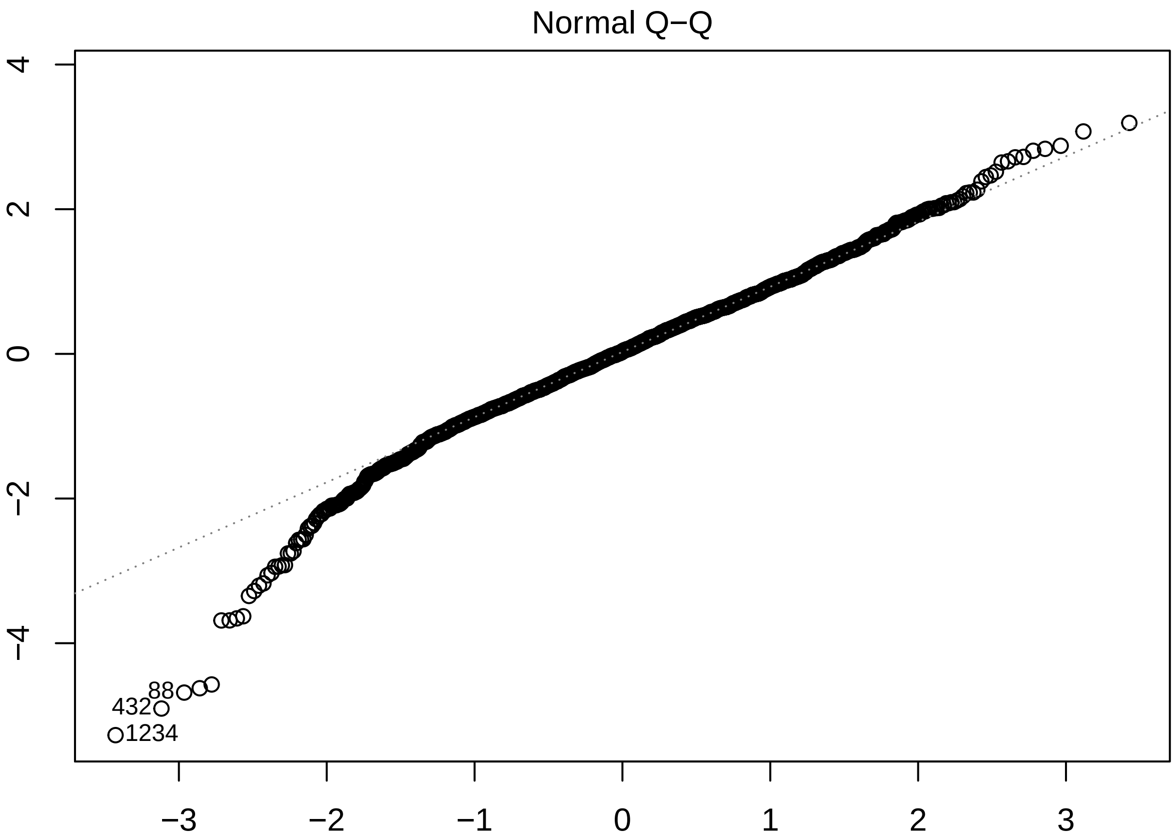
Looking at the Residuals vs. Fitted plot shows that the zero conditional mean is met because the red line is approximately at 0.

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 roboust to heteroskadacity.

6) Errors are normally distributed

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



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

```
##
##  Shapiro-Wilk normality test
##
## data:  model3$residuals
## W = 0.97835, p-value = 4.598e-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(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
```

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

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 2999.6248   124.7351 24.0480 < 2.2e-16 ***
## monpre        20.9010    12.0531  1.7341  0.083091 .
## npvis         15.5046     4.6619  3.3258  0.000901 ***
## cigs         -11.2291     3.6793 -3.0520  0.002310 **
## drink        -14.0495    33.0106 -0.4256  0.670451
## mage           5.3168     3.1399  1.6933  0.090592 .
## male          80.9374    28.2671  2.8633  0.004246 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AIC(model1)
```

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

```
AIC(model3)
```

```
## [1] 25582.07
```

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)
```

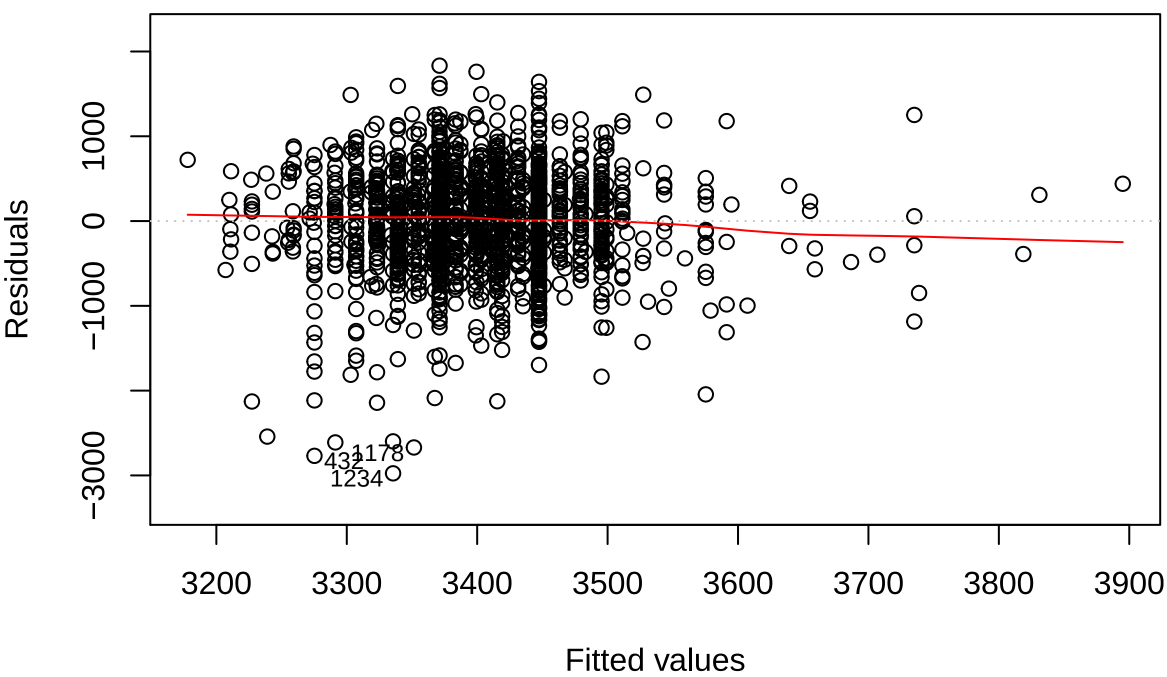
```
##
## Call:
## lm(formula = data$bwght ~ data$npvis + data$male + data$mblck +
##     data$fblck)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2975.51  -336.55    31.69   360.92  1832.85
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3179.315     48.188  65.977  < 2e-16 ***
## data$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
```

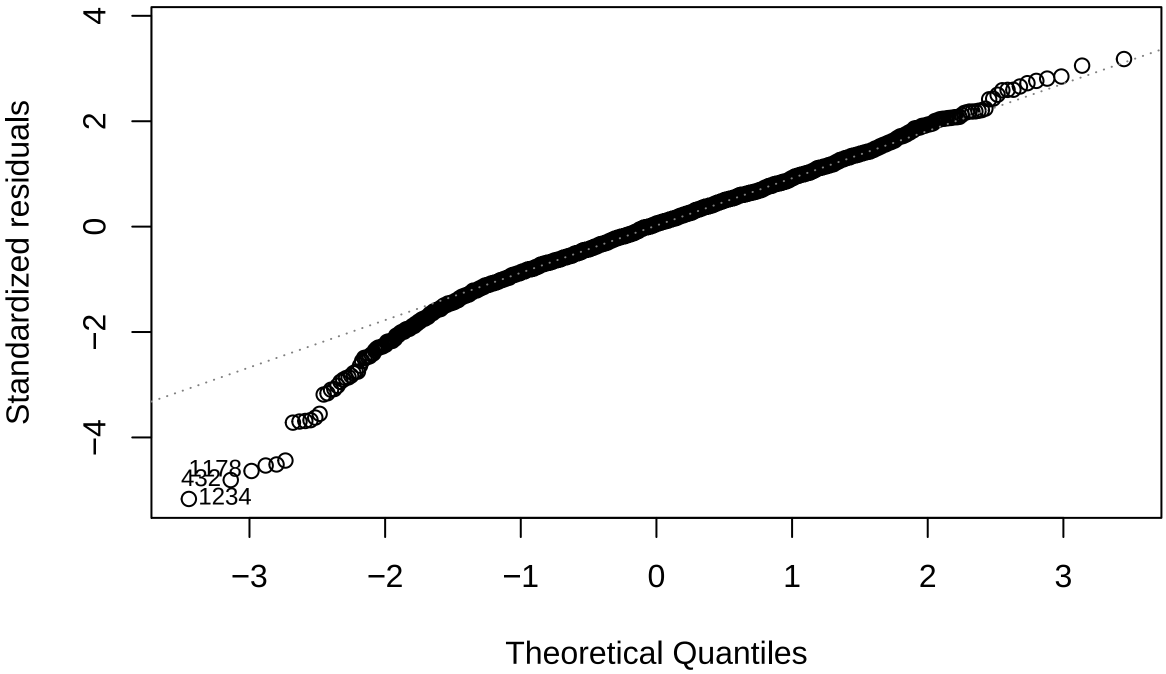
```r
plot(c1)
```

Residuals vs Fitted

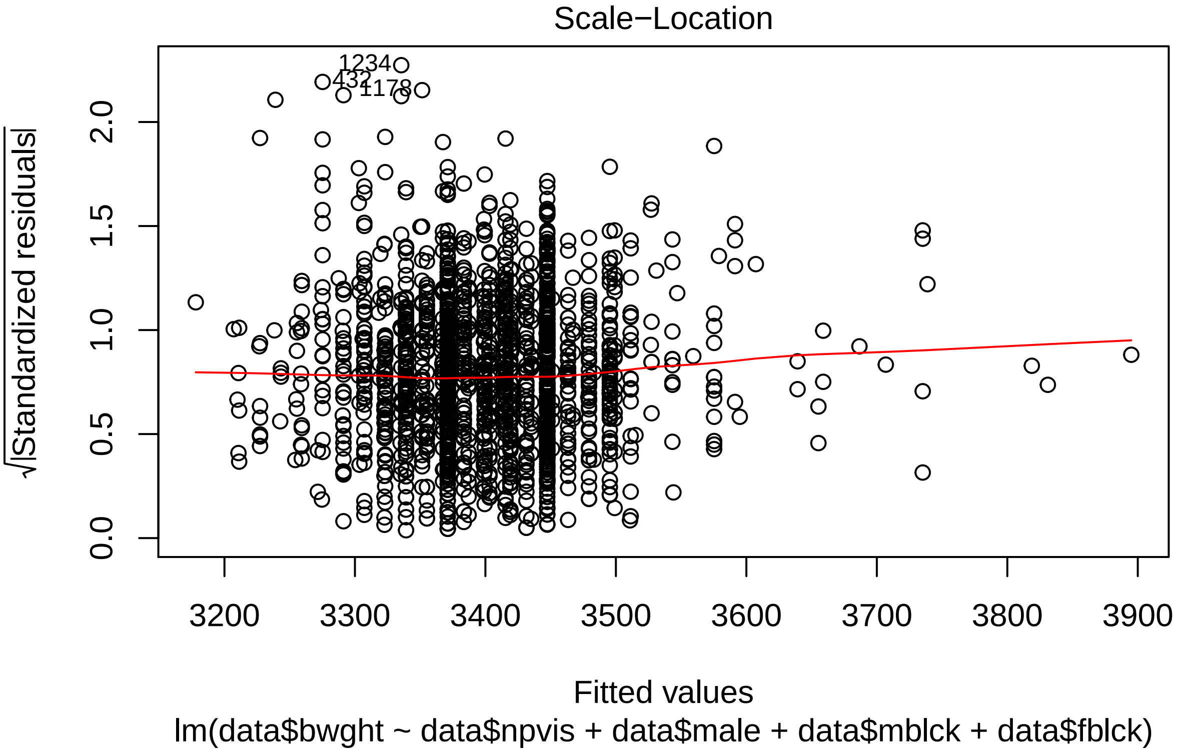


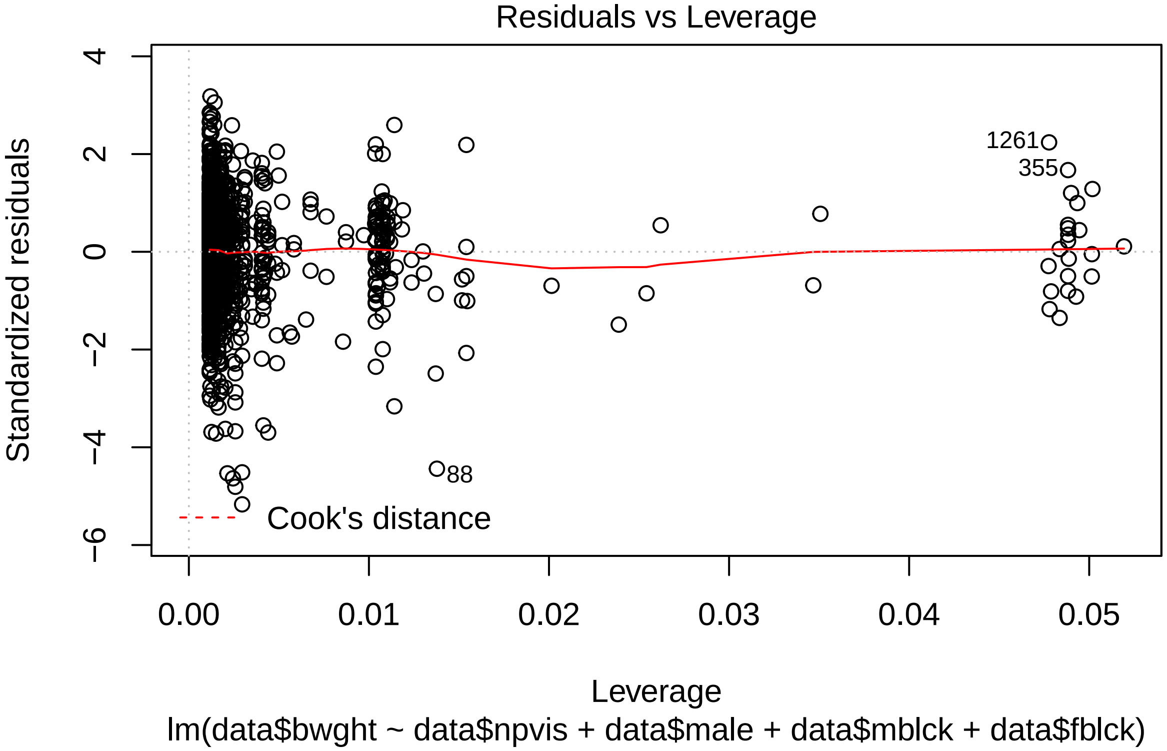
lm(data$bwght ~ data$npvis + data$male + data$mblck + data$fblck)

Normal Q−Q



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:

```r
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: CLM and the Models

Step 5: Regression Tables and Model Analysis

```
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: Fri, Dec 09, 2016 - 19:37:41

See table 1 on the next page.

```
AIC(model1)
```

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

```
AIC(model3)
```

```
## [1] 25582.07
```

```
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 R^2, 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 words, 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 like that gives us good grounds to completely ignore the `monpre` variable.

Table 1:

	Dependent variable:			
	bwght	normalized_product_apgar	bwght	bwght
	(1)	(2)	(3)	(4)
monpre	17.062 (12.028)		20.901 (12.053)	
npvis	17.549*** (4.834)		15.505*** (4.662)	
monpre		−8.502 (5.724)		
npvis		6.313* (2.517)		15.986*** (4.352)
cigs			−11.229** (3.679)	
drink			−14.050 (33.011)	
mage			5.317 (3.140)	
male			80.937** (28.267)	
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)	2,999.625*** (124.735)	3,179.315*** (56.248)
Observations	1,763	1,760	1,647	1,764
R^2	0.011	0.010	0.023	0.015
Adjusted R^2	0.010	0.009	0.019	0.013
Residual Std. Error	577.470 (df = 1760)	284.115 (df = 1757)	569.408 (df = 1640)	576.683 (df = 1759)
Note:				*p<0.05; **p<0.01; ***p<0.001

Step 6: Causality

We choose to operationalize infant health by birthweight. There are many other factors that influence birthweight that are not captured in this data set, which leads to omitted variable bias.

1) Mother's weight is a strong predictor for newborn weight.

2) Socioeconomic status of mother.

3) Having more than one baby at a time reduces the weight of each baby. (E.g. twins will be smaller)

4)

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.

Step 7: Conclusion

Prenatal care, as shown by number of prenatal care visits has a positive impact on birthweight. Other explanatory factors are mother's cig consumption, which has a negative impact on birthweigth. Being male has a positive impact on birthweight.

References

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

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

[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