Lab 4: Healthy Momma, Healthy Baby

Krista Mar and Nikki Haas

12/1/2016

A Nice Introduction that Makes Us Sound Like Pros

All work and no play makes Nikki a dull girl. All work and no play makes Nikki a dull girl. All work and no play makes Nikki a dull girl. All work and no play makes Nikki a dull girl. All work and no play makes Nikki a dull girl. All work and no play makes Nikki a dull girl. All work and no play makes Nikki a dull girl. All work and no play makes Nikki a dull girl. All work and no play makes Nikki a dull girl. All work and no play makes Nikki a dull girl. All work and no play makes Nikki a dull girl. All work and no play makes Nikki a dull girl. All work and no play makes Nikki a dull girl. All work and no play makes Nikki a dull girl.

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

Step 1: Read in the Data

```
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.

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

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

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

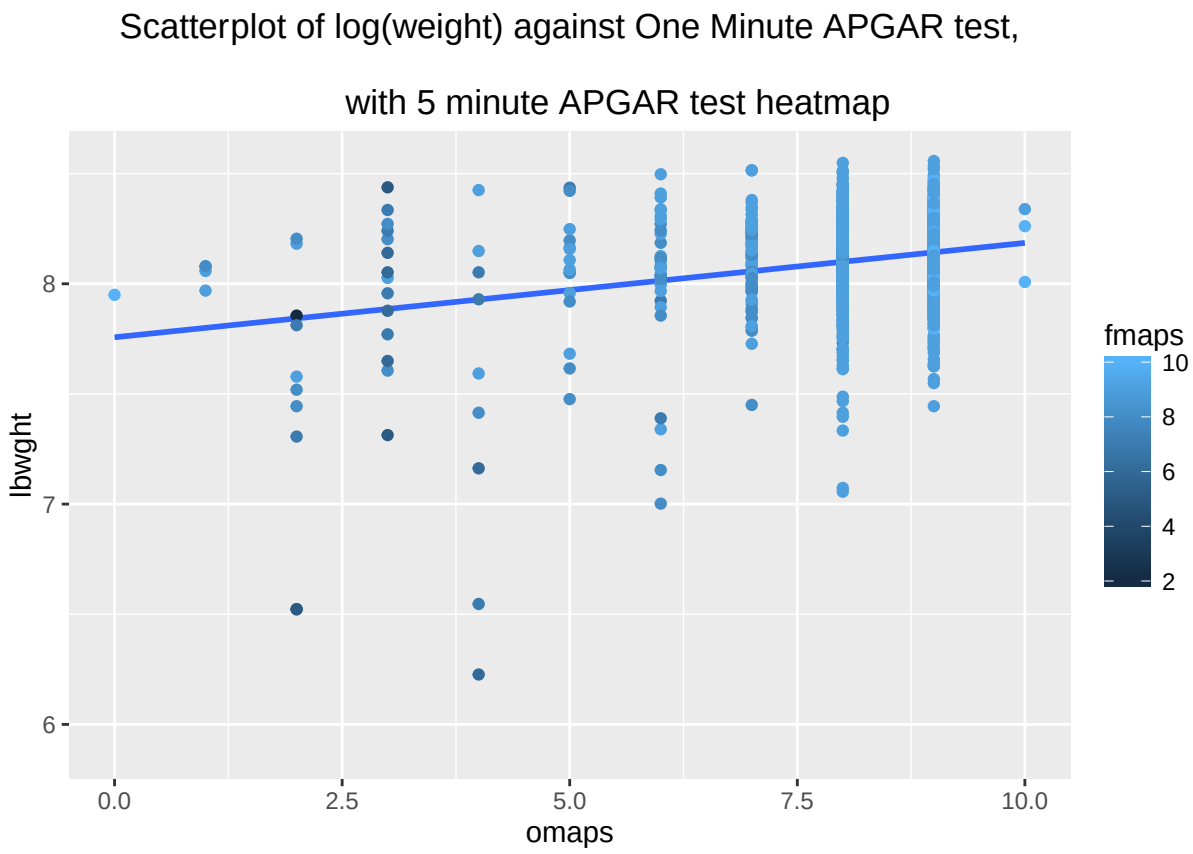
```
## [1] 0.2710456
```

```
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
```

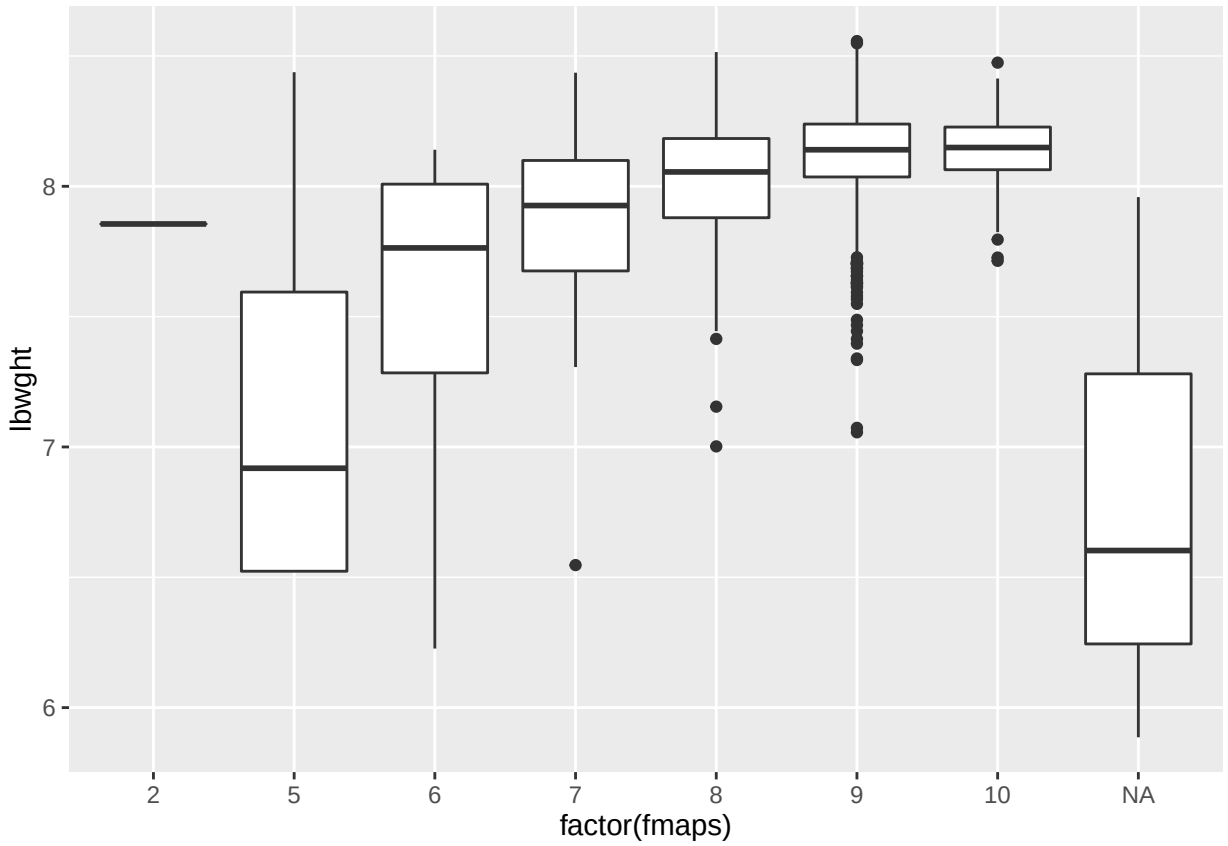
```
## Warning: Removed 3 rows containing non-finite values (stat_smooth).
```

```
## 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 fmops 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")
```

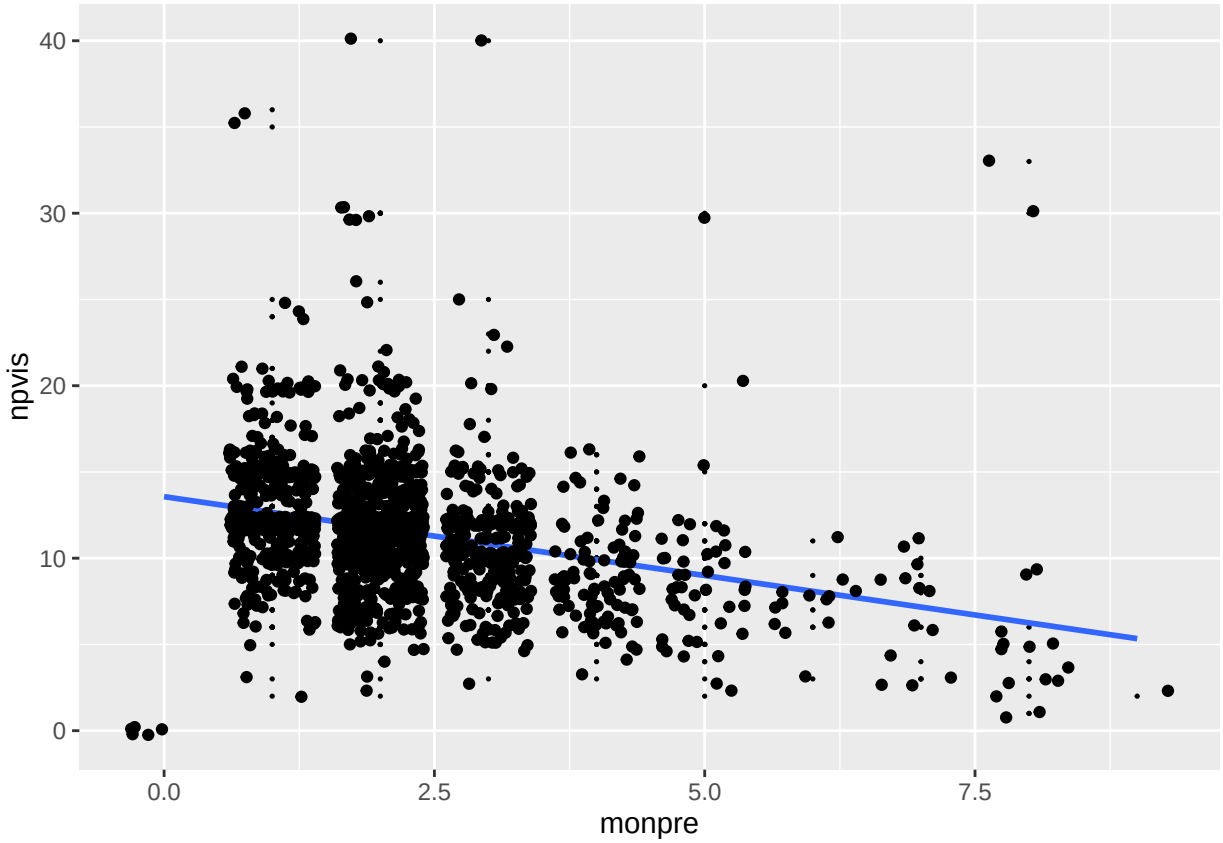
```
## [1] -0.3061006
```

```
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).
```

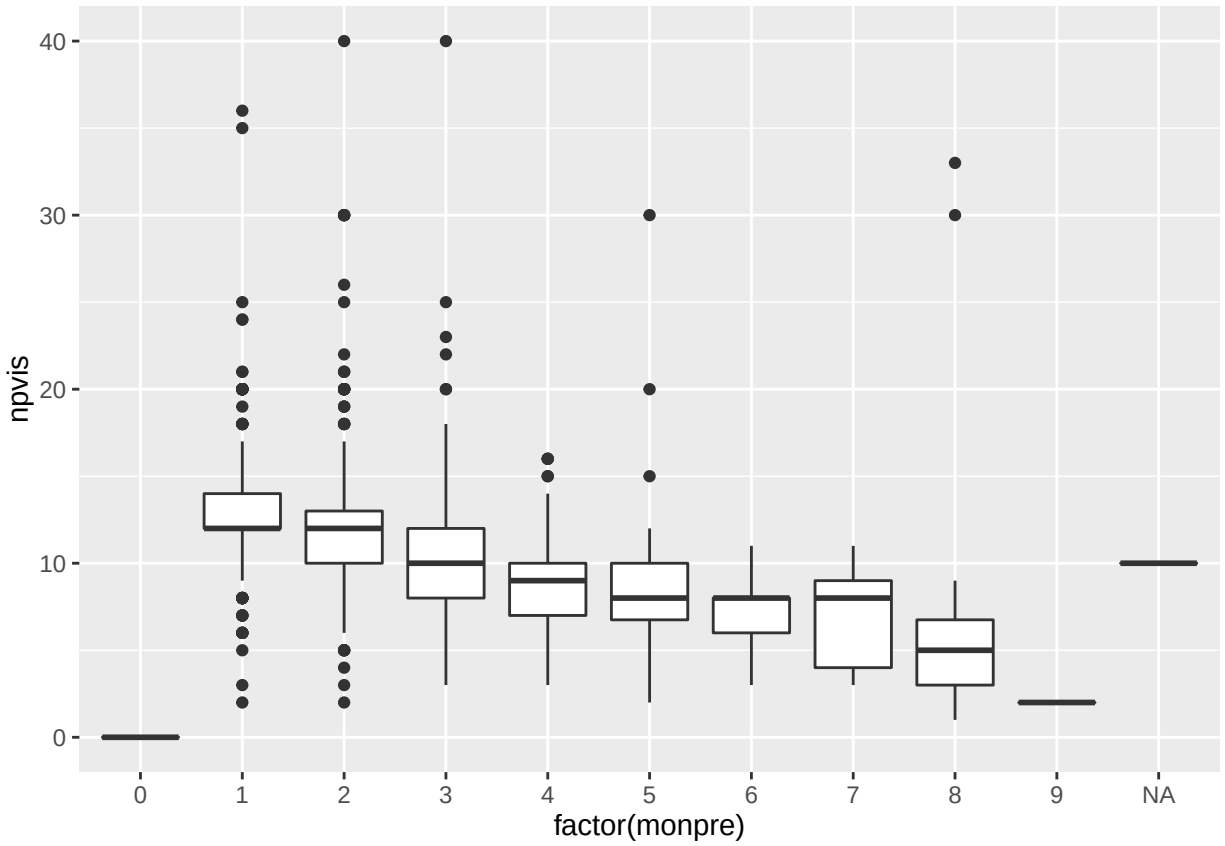
```
## Warning: Removed 69 rows containing missing values (geom_point).
```

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




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

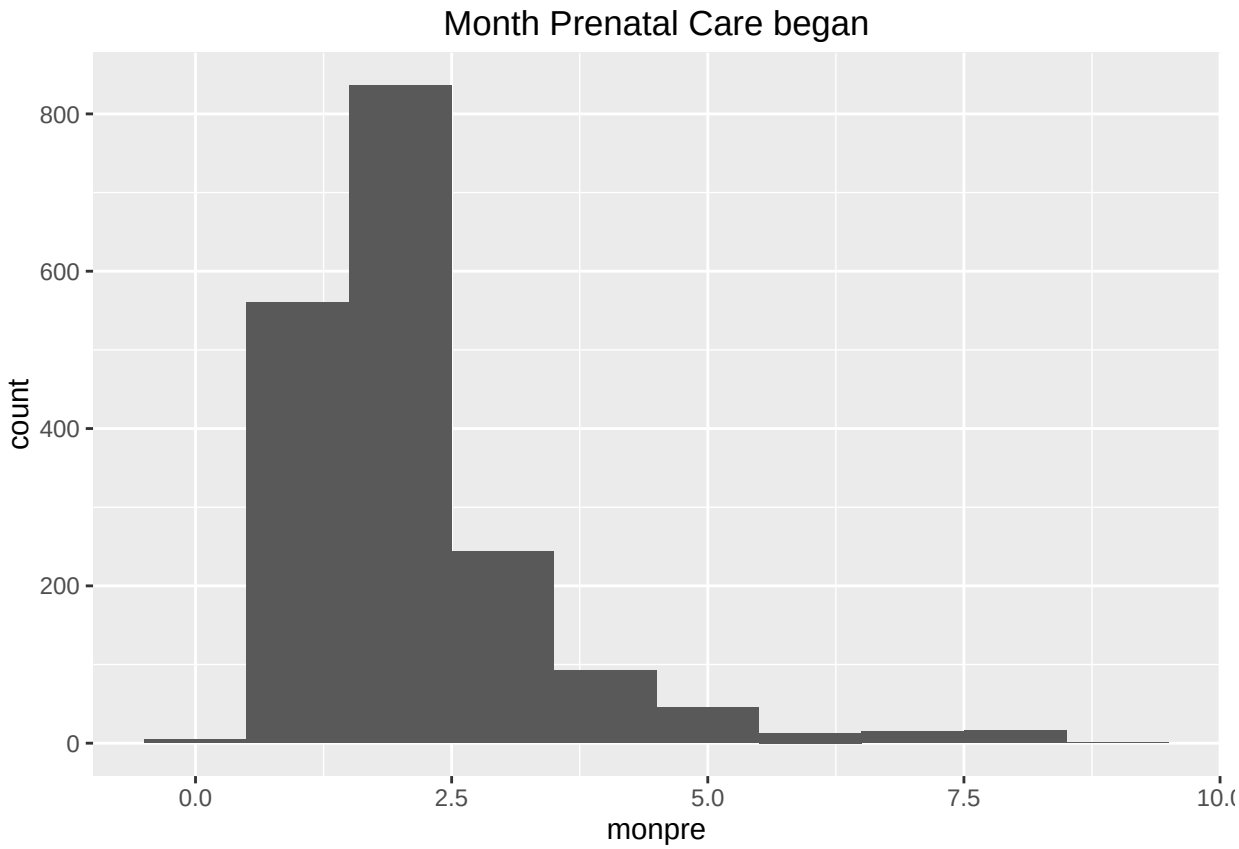
```
## Warning: Removed 68 rows containing non-finite values (stat_boxplot).
```



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.

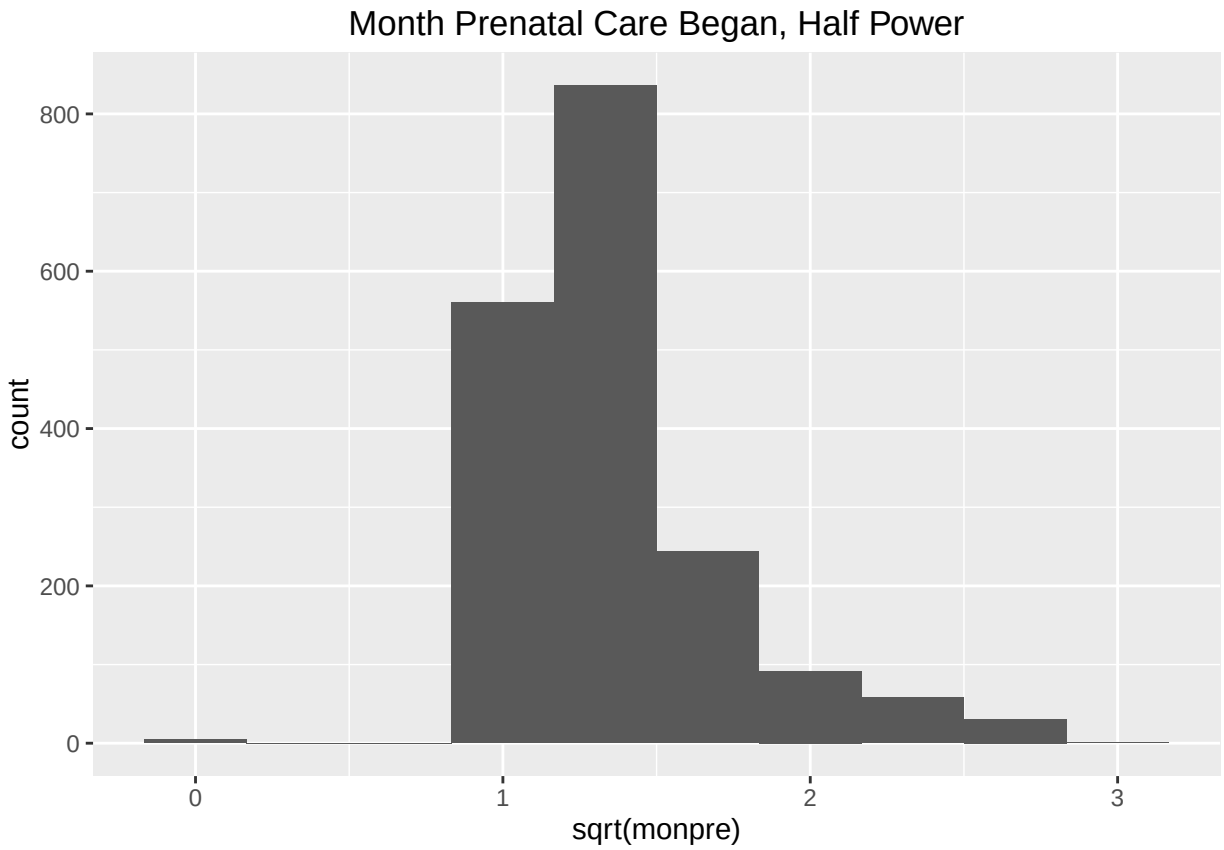
```
ggplot(data, aes(x=monpre)) + geom_histogram(aes(y = ..count..),bins = 10) +
  ggtitle("Month Prenatal Care began")
```

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



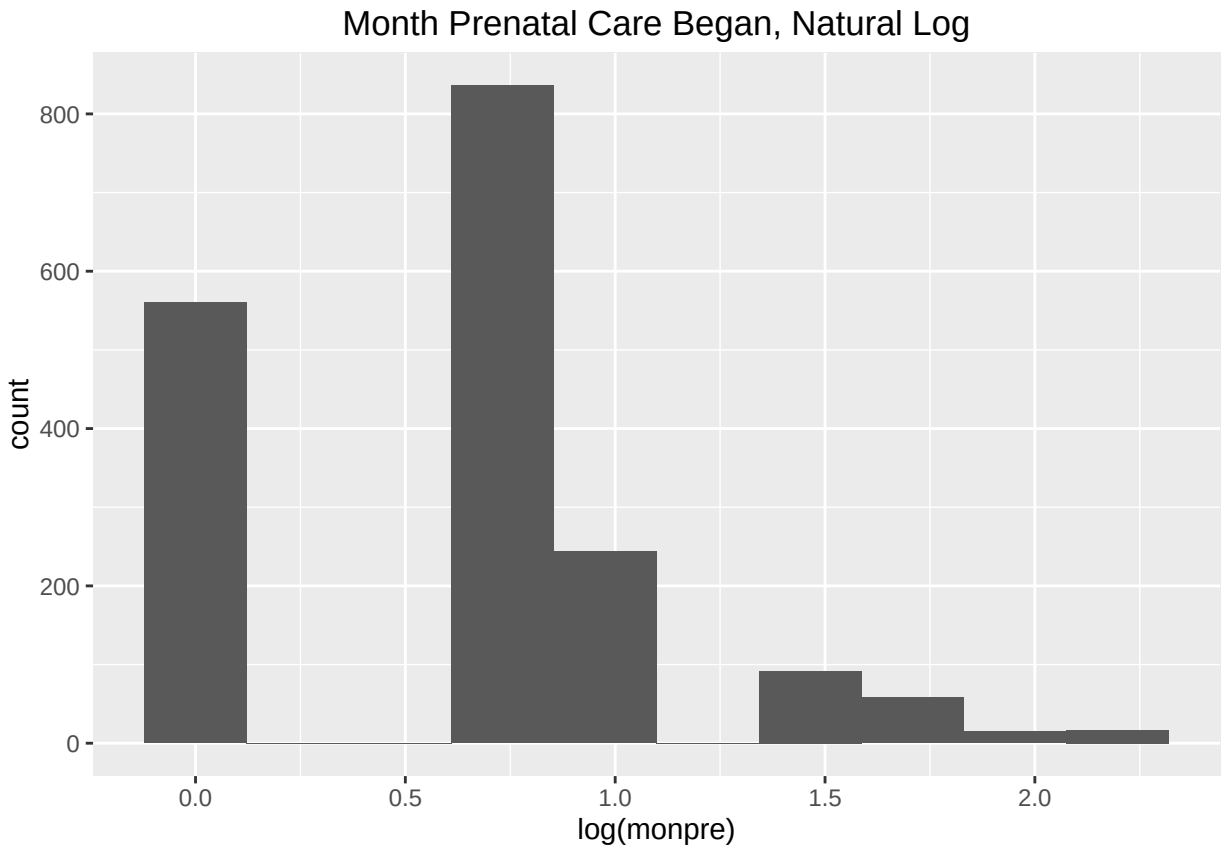

```
ggplot(data, aes(x=sqrt(monpre))) + geom_histogram(aes(y = ..count..), bins = 10) +
  ggtitle("Month Prenatal Care Began, Half Power")
```

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



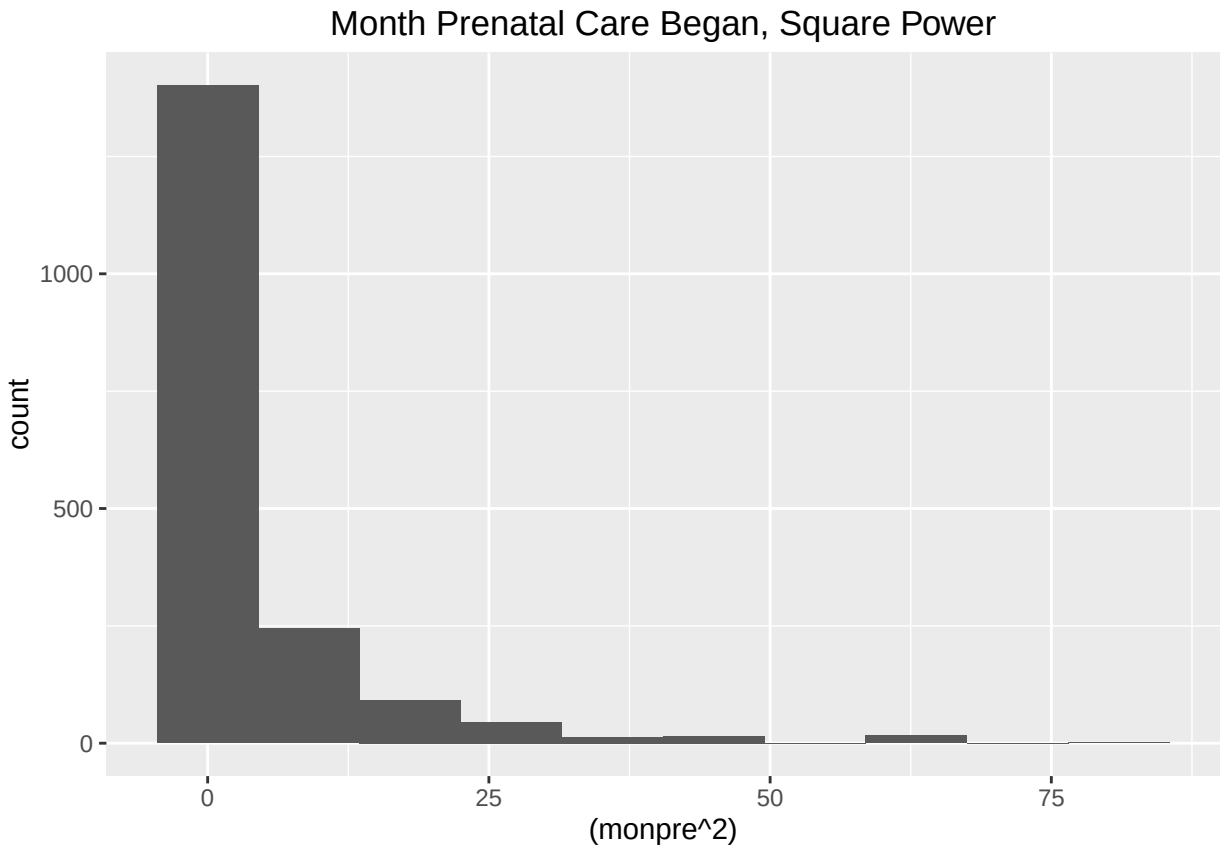

```
ggplot(data, aes(x=log(monpre))) + geom_histogram(aes(y = ..count..), bins = 10) +
  ggtitle("Month Prenatal Care Began, Natural Log")
```

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



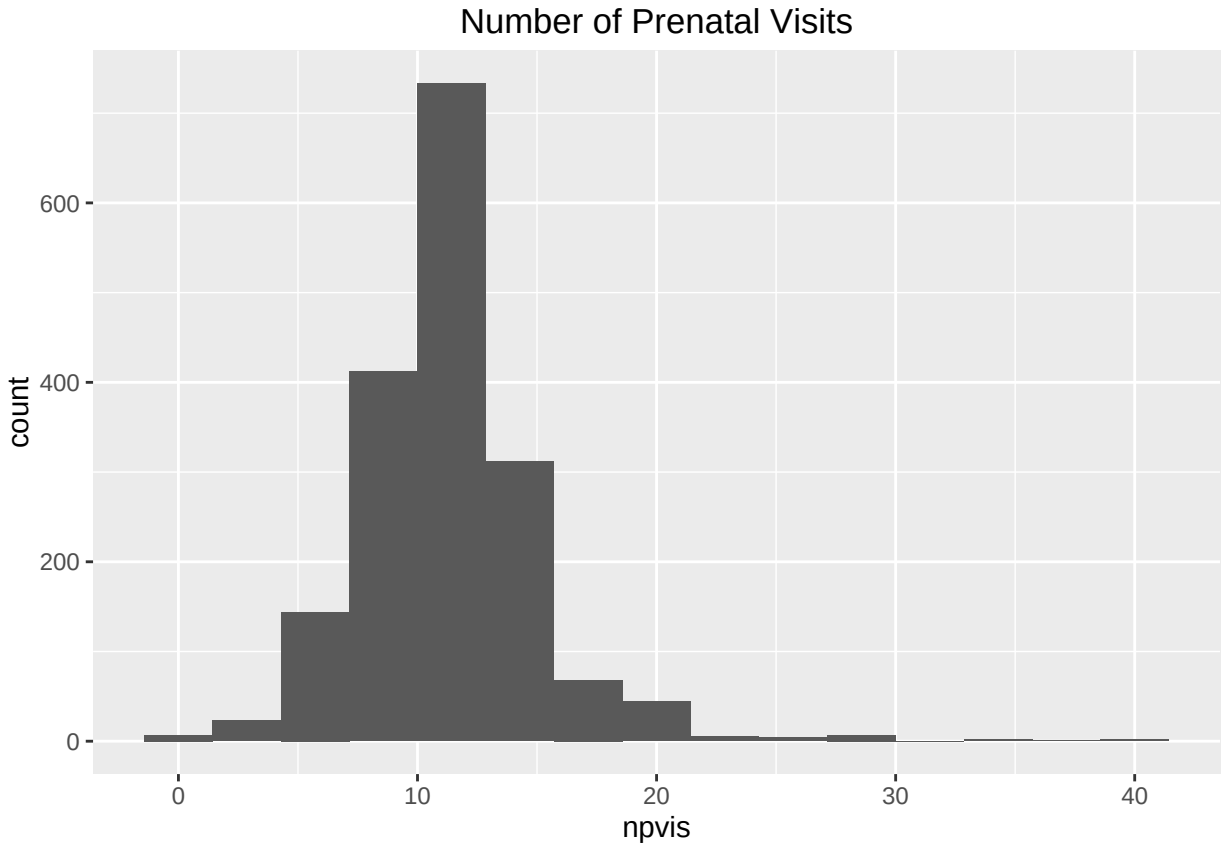

```
ggplot(data, aes(x=(monpre^2))) + geom_histogram(aes(y = ..count..), bins = 10) +
  ggtitle("Month Prenatal Care Began, Square Power")
```

```
## 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).
```



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. Then we say smart things about how that will all relate to each other.

Step 3: Modeling

Model 1: Basic Linear Model

First we will check variants on the requested model; baby health as a function of the mother's prenatal care. We will not

```
model1 = lm(data$bwght ~ data$monpre + data$npvis)
model1
```

```
##
## Call:
## lm(formula = data$bwght ~ data$monpre + data$npvis)
##
## Coefficients:
## (Intercept)  data$monpre   data$npvis
##     3161.27        17.06        17.55
```

```
model2 = lm(data$lbwght ~ data$monpre + data$npvis)
model2
```

```
##
## Call:
## lm(formula = data$lbwght ~ data$monpre + data$npvis)
##
## Coefficients:
```

```
## (Intercept)  data$monpre   data$npvis
##    8.008629     0.008570     0.007503
```

```
model3 = lm(data$bwght ~ sqrt(data$monpre) + data$npvis)
model3
```

```
##
## Call:
## lm(formula = data$bwght ~ sqrt(data$monpre) + data$npvis)
##
## Coefficients:
##       (Intercept)  sqrt(data$monpre)         data$npvis
##           3122.10              54.93              17.39
```

```
model4 = lm(data$lbwght ~ sqrt(data$monpre) + data$npvis)
model4
```

```
##
## Call:
## lm(formula = data$lbwght ~ sqrt(data$monpre) + data$npvis)
##
## Coefficients:
##       (Intercept)  sqrt(data$monpre)         data$npvis
##          7.988854           0.027646           0.007423
```

```
AIC(model1)
```

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

```
AIC(model2)
```

```
## [1] -603.5921
```

```
AIC(model3)
```

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

```
AIC(model4)
```

```
## [1] -603.6182
```

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 sum up `omaps` and `fmaps`. 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.

```
data$combinded_apgarscores = data$omaps + data$fmaps
```

Now that we have a calculated field that sums up the APGAR tests, we can try an alternate linear model:

```r
a1 = lm(data$combinded_apgarscores ~ data$monpre + data$npvis)
a3 = lm(data$combinded_apgarscores ~  data$npvis)
summary(a1)
```

```
##
## Call:
## lm(formula = data$combinded_apgarscores ~ data$monpre + data$npvis)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.343  -0.409   0.555   0.624   2.792
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.084596   0.149953 113.933  < 2e-16 ***
## data$monpre -0.036001   0.029285  -1.229 0.219117
## data$npvis   0.033032   0.009818   3.364 0.000783 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.441 on 1757 degrees of freedom
##   (72 observations deleted due to missingness)
## Multiple R-squared:  0.009575,   Adjusted R-squared:  0.008448
## F-statistic: 8.493 on 2 and 1757 DF,  p-value: 0.0002134
```

```r
summary(a3)
```

```
##
## Call:
## lm(formula = data$combinded_apgarscores ~ data$npvis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.3313  -0.4050   0.5582   0.6319   2.7792
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.962919   0.113922 148.900  < 2e-16 ***
## data$npvis   0.036838   0.009342   3.943 8.35e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.441 on 1759 degrees of freedom
##   (71 observations deleted due to missingness)
## Multiple R-squared:  0.008763,   Adjusted R-squared:  0.008199
## F-statistic: 15.55 on 1 and 1759 DF,  p-value: 8.35e-05
```

```r
AIC(a1)
```

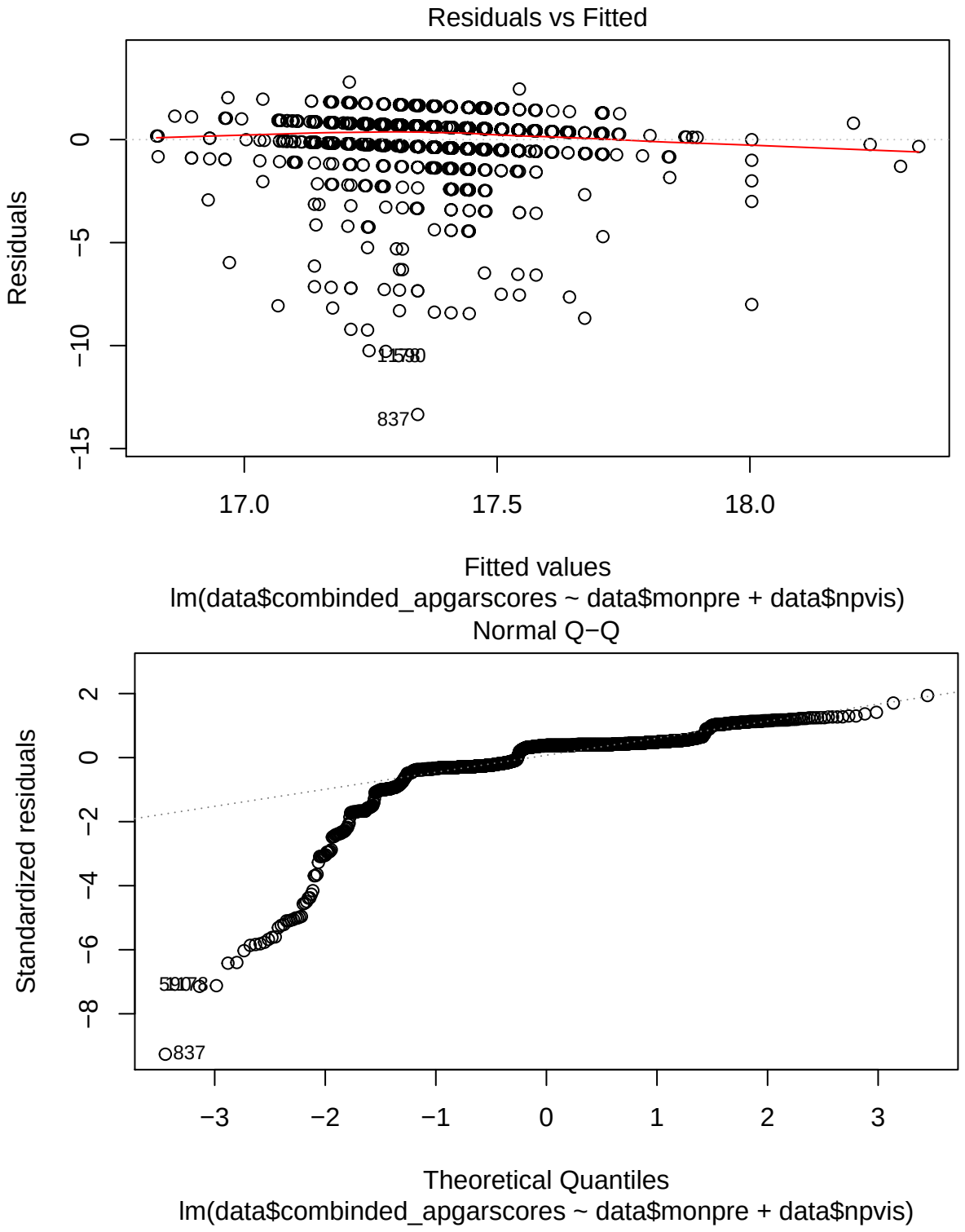
```
## [1] 6285.572
```

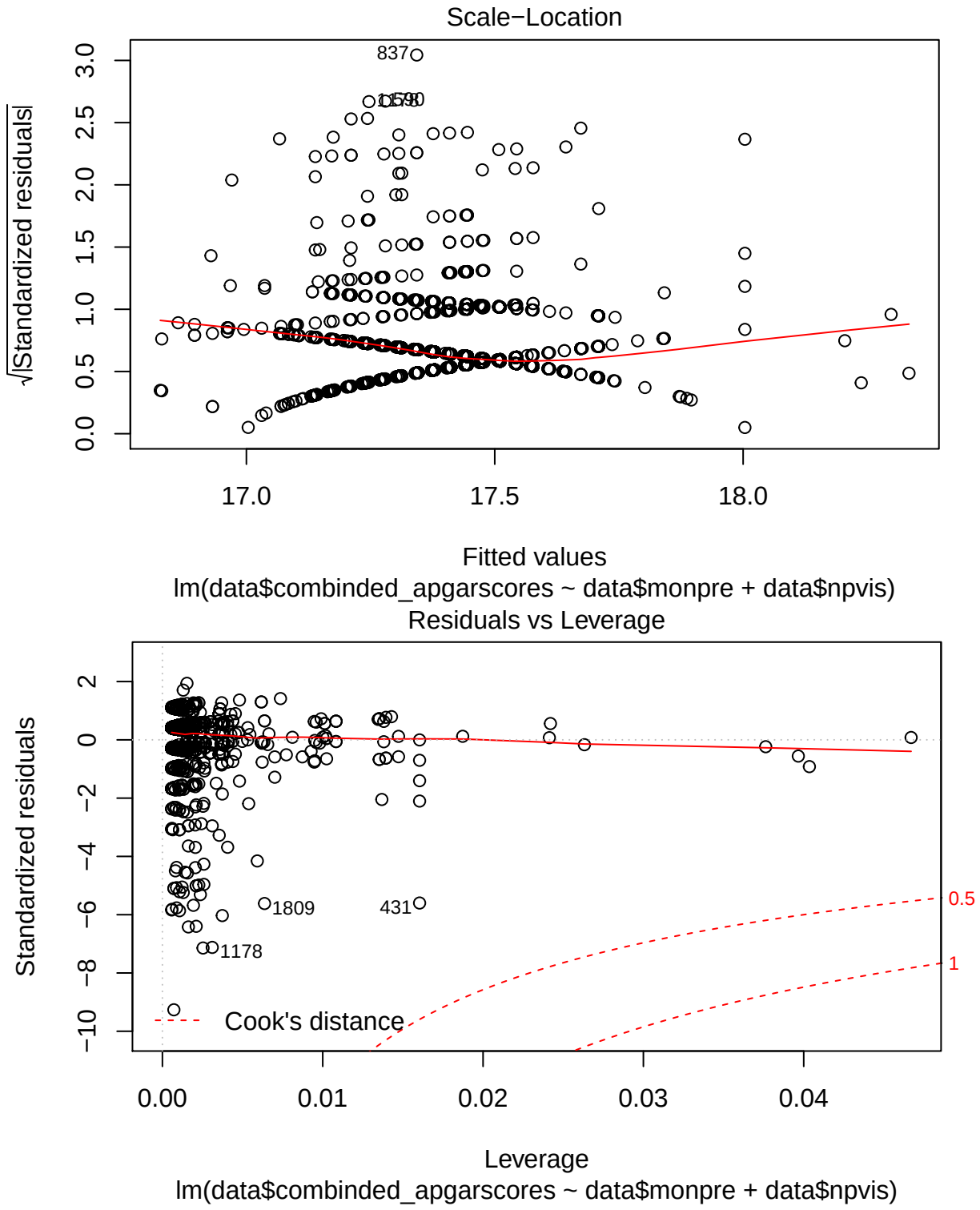
```r
AIC(a3)
```

```
## [1] 6288.508
```

Our AIC scores tell us that a1 more efficient, so let us exlore a1 further:

```
plot(a1)
```


Residuals vs Fitted
Residuals
0
−5
−10
−15
837
17.0
17.5
18.0
Fitted values
lm(data$combinded_apgarscores ~ data$monpre + data$npvis)
Normal Q−Q
Standardized residuals
2
0
−2
−4
−6
−8
837
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(data$combinded_apgarscores ~ data$monpre + data$npvis)




```
data$product_apgarscores = data$omaps * data$fmaps
```

```
a5 = lm(data$product_apgarscores ~ data$monpre + data$npvis)
summary(a5)
```

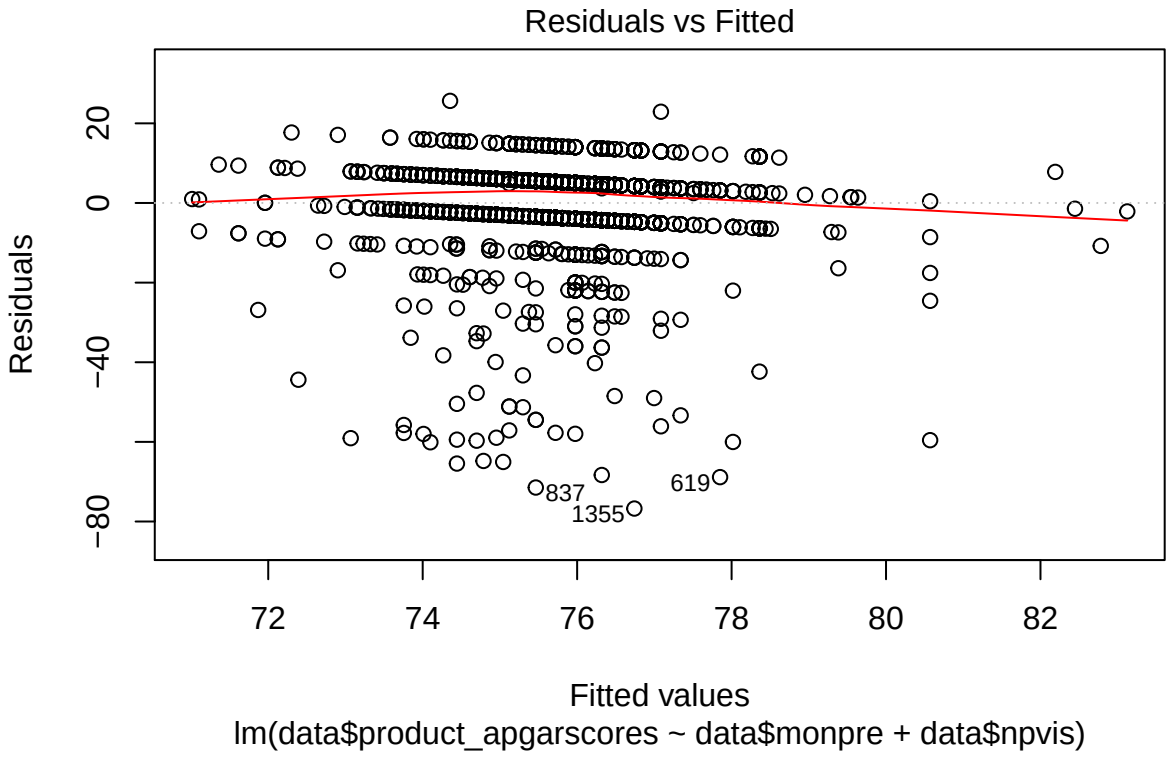
```
##
## Call:
```

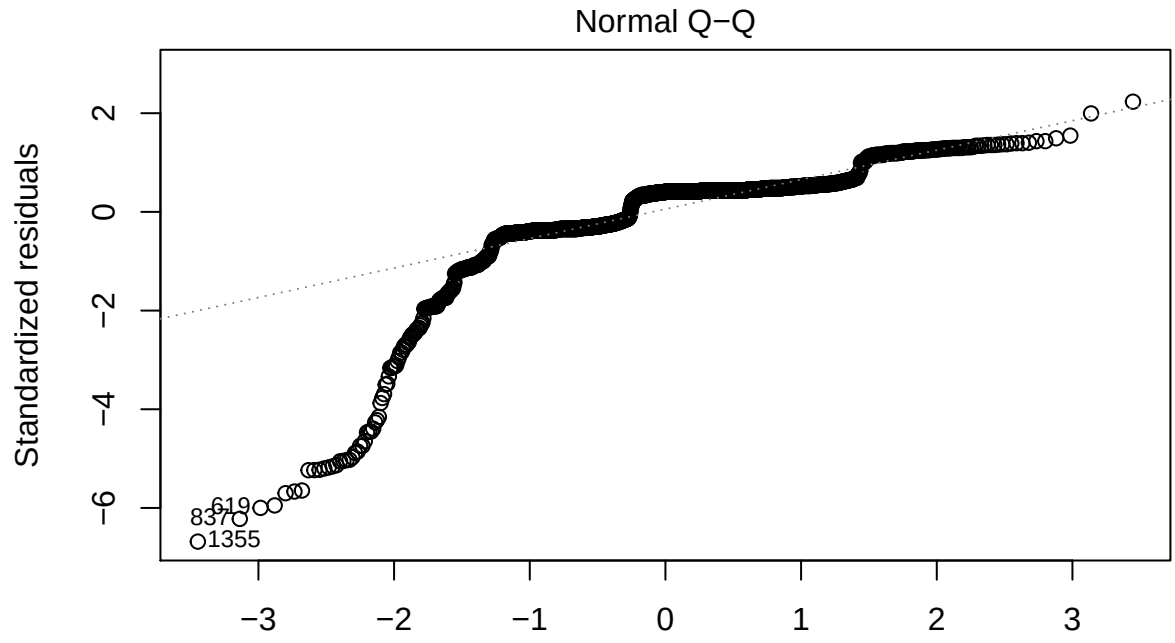
```
## lm(formula = data$product_apgarscores ~ data$monpre + data$npvis)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -76.741  -3.975   4.681   5.280  25.645
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 73.59915    1.19580  61.548  < 2e-16 ***
## data$monpre -0.34388    0.23353  -1.472  0.14107
## data$npvis   0.25532    0.07829   3.261  0.00113 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.49 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
```

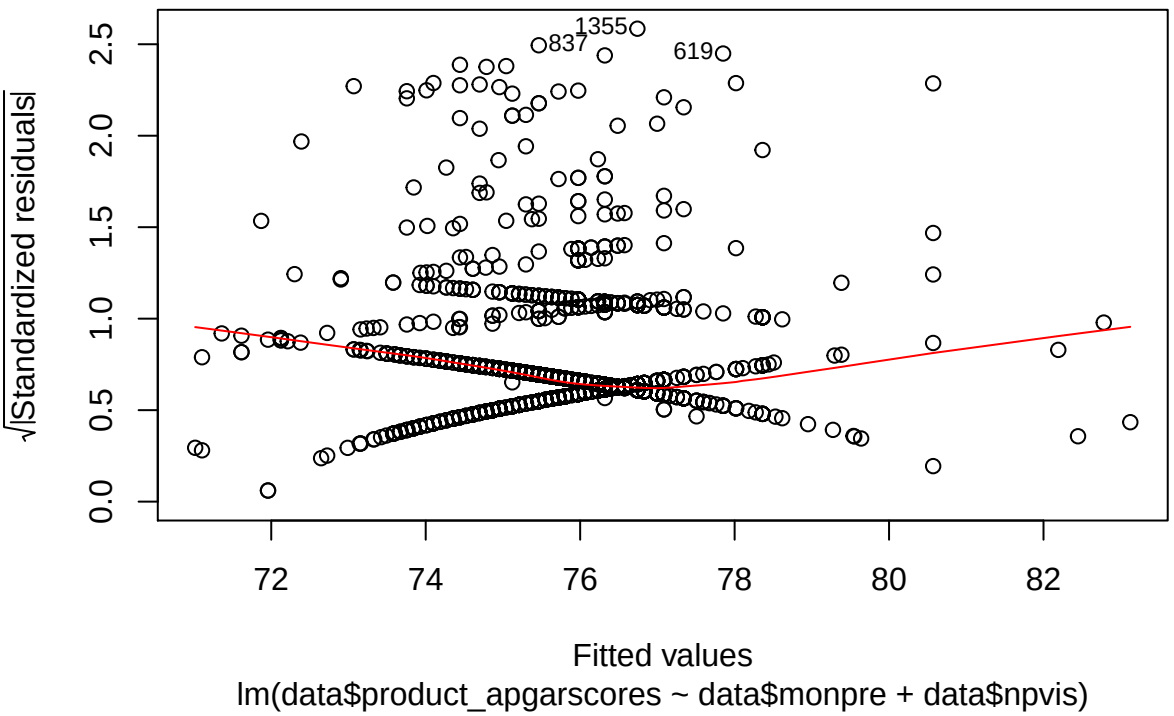
```
AIC(a5)
```

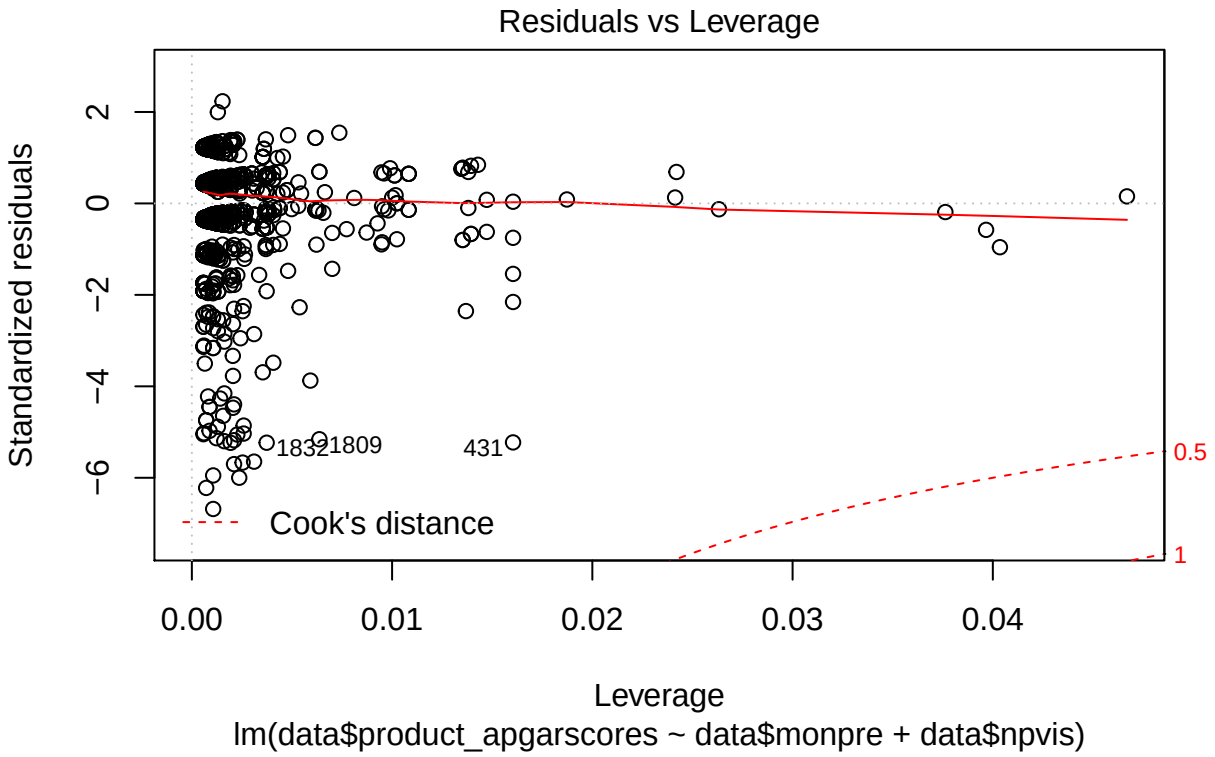
```
## [1] 13593.95
```

```
plot(a5)
```




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


Scale−Location
√|Standardized residuals|
1355
837
619
0.0
0.5
1.0
1.5
2.0
2.5
72
74
76
78
80
82
Fitted values
lm(data$product_apgarscores ~ data$monpre + data$npvis)



This data is not looking good either. Let's try removing the non-normally distributed `data$monpre` field

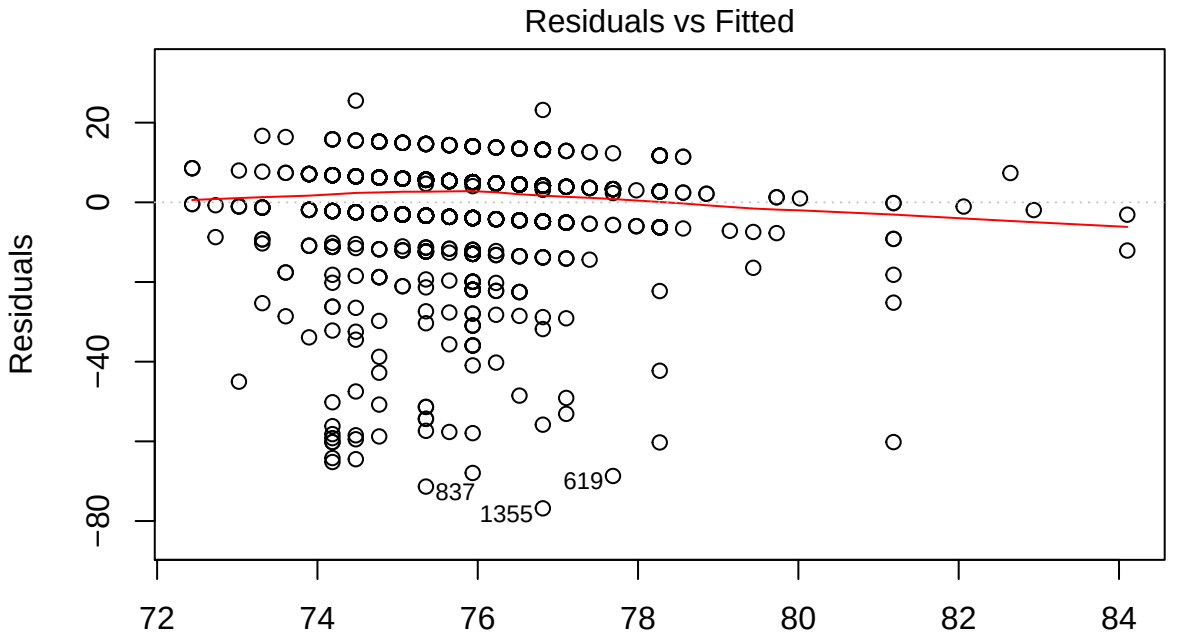
```
a6 = lm(data$product_apgarscores ~ data$npvis)
summary(a6)
```

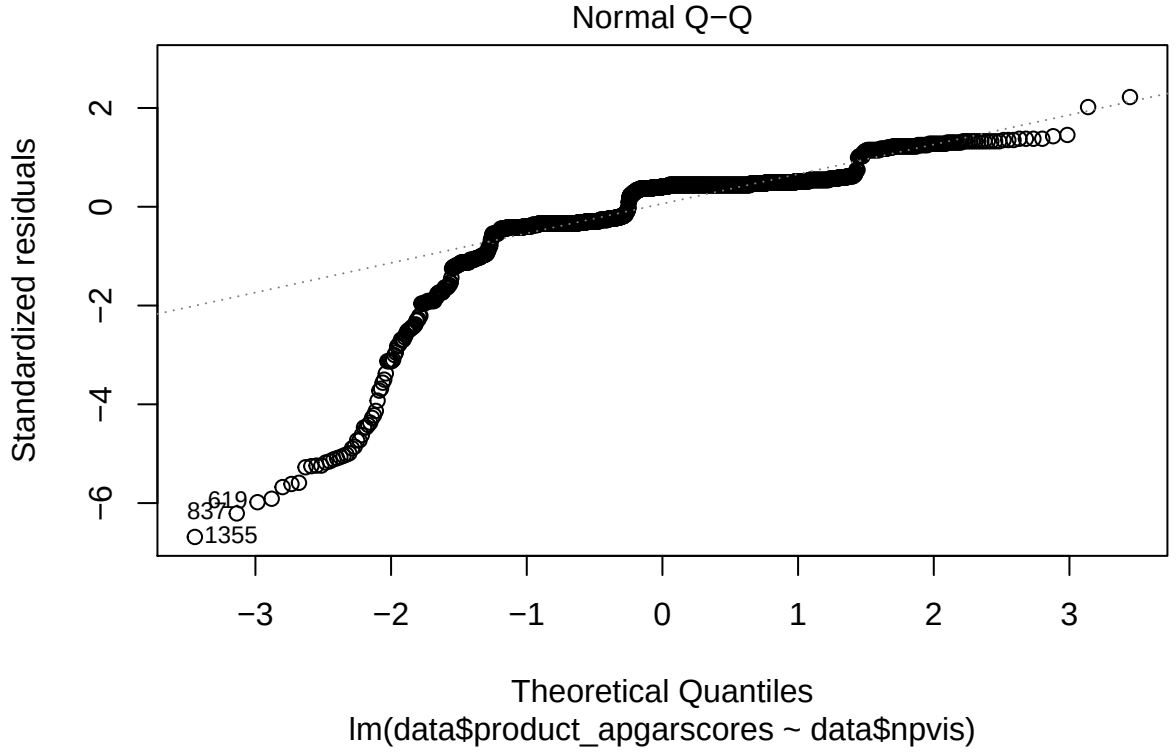
```
##
## Call:
## lm(formula = data$product_apgarscores ~ data$npvis)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -76.812  -3.937   4.771   5.354  25.521
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 72.43739    0.90871  79.715  < 2e-16 ***
## data$npvis   0.29165    0.07452   3.914 9.43e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.5 on 1759 degrees of freedom
##   (71 observations deleted due to missingness)
## Multiple R-squared:  0.008633,   Adjusted R-squared:  0.00807
## F-statistic: 15.32 on 1 and 1759 DF,  p-value: 9.429e-05
```

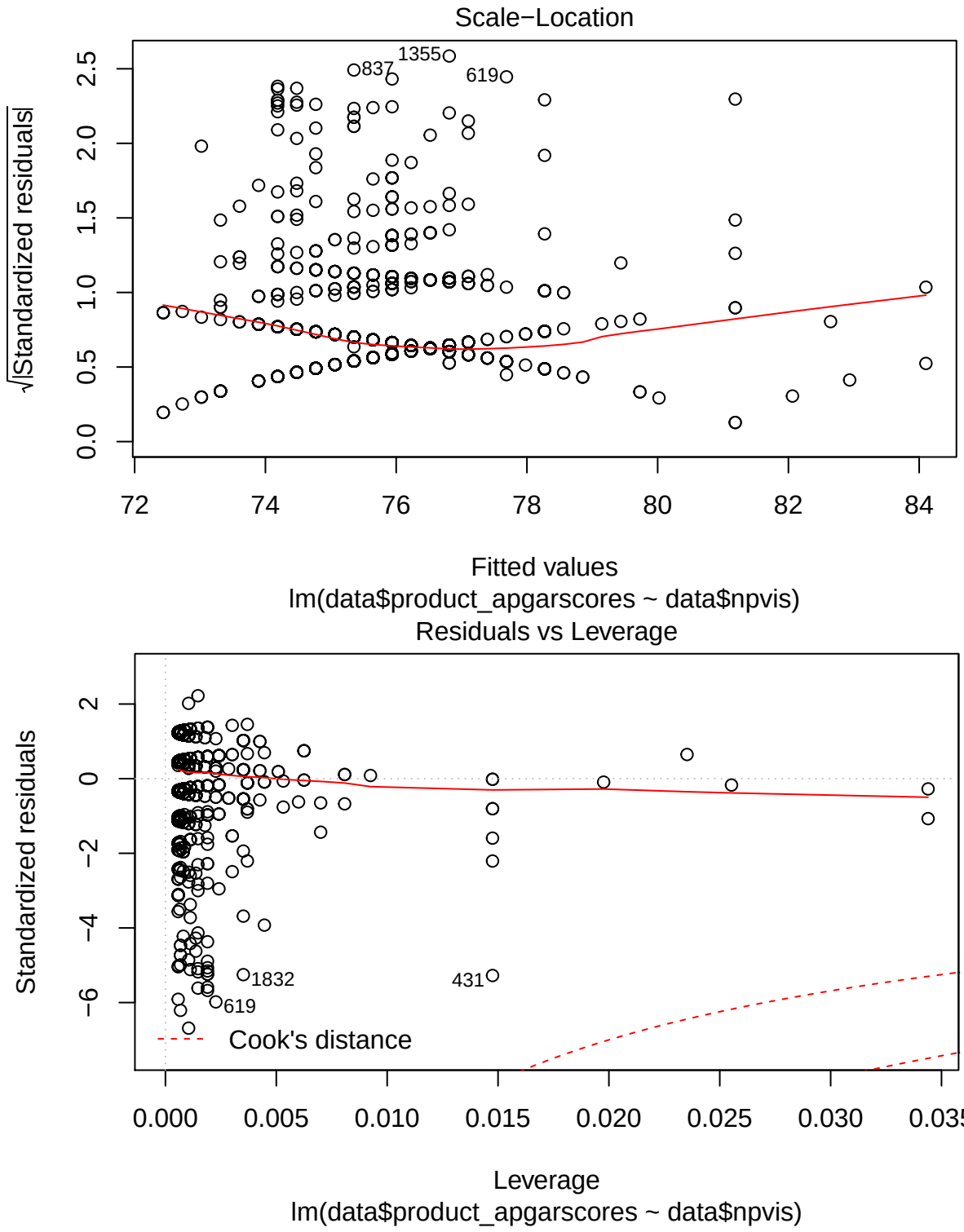
```
AIC(a6)
```

```
## [1] 13601.99
```

```
plot(a6)
```


Residuals vs Fitted
Residuals
20
0
−40
−80
837
619
1355
72
74
76
78
80
82
84
Fitted values
lm(data$product_apgarscores ~ data$npvis)


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



Unfortunately that too is a little worse. Let's try normalizing these calculated values:

```
data$normalized_combined_apgar = (data$combinded_apgarscores - mean(!is.na(data$combinded_apgarscores))
data$normalized_product_apgar =  (data$product_apgarscores- mean(!is.na(data$product_apgarscores)))/sd(

a7 = lm(data$normalized_combined_apgar~data$monpre + data$npvis)
a8 =lm(data$normalized_product_apgar~data$monpre + data$npvis)
```

```r
summary(a7)
```

```
##
## Call:
## lm(formula = data$normalized_combined_apgar ~ data$monpre + data$npvis)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -329.91  -10.11   13.72   15.43   69.04
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 397.7350     3.7076 107.275  < 2e-16 ***
## data$monpre  -0.8901     0.7241  -1.229 0.219117
## data$npvis    0.8167     0.2428   3.364 0.000783 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.63 on 1757 degrees of freedom
##   (72 observations deleted due to missingness)
## Multiple R-squared:  0.009575,   Adjusted R-squared:  0.008448
## F-statistic: 8.493 on 2 and 1757 DF,  p-value: 0.0002134
```

```r
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
```

We did not see very good results with the APGAR score variations, so now let's try to normalize the baby's birth weight by APGAR.

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.

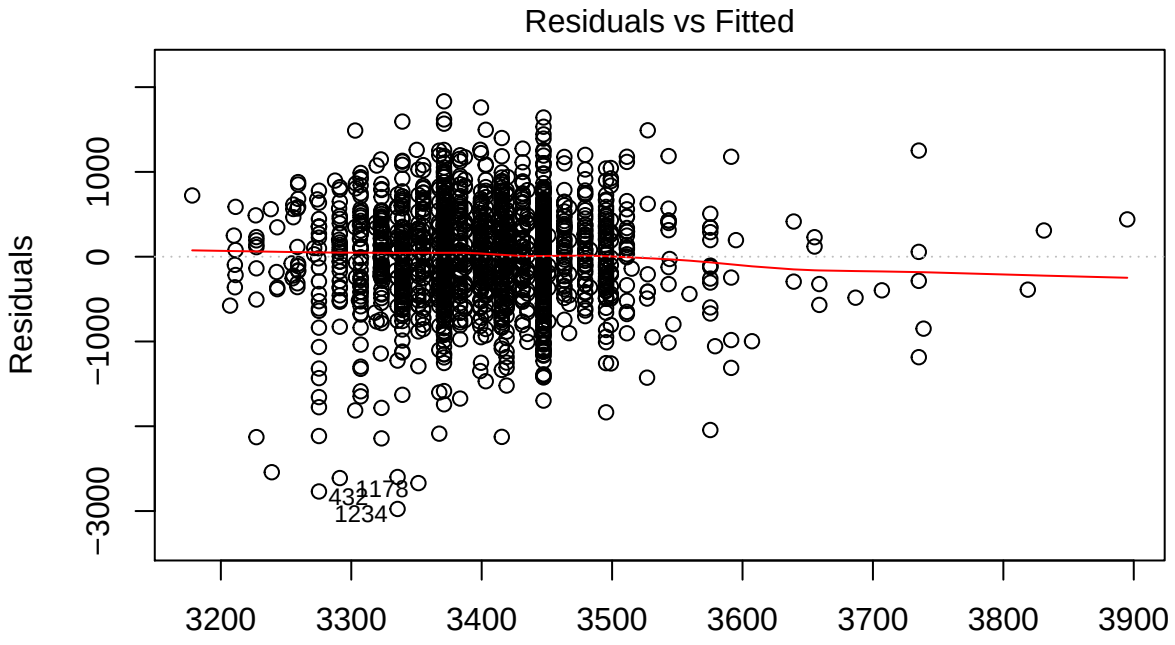
```r
c1 = lm(data$bwght ~ data$npvis + data$male +
          data$mblck + data$fblck)
summary(c1)
```

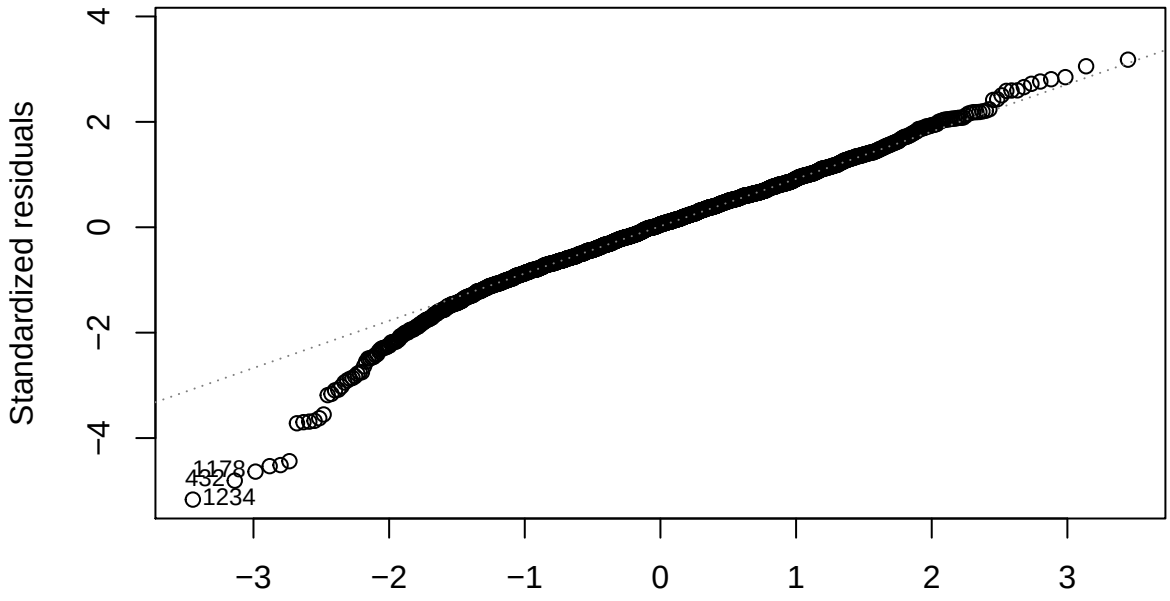
```
##
## Call:
## lm(formula = data$bwght ~ data$npvis + data$male + data$mblck +
##     data$fblck)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2975.51  -336.55    31.69   360.92  1832.85
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)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

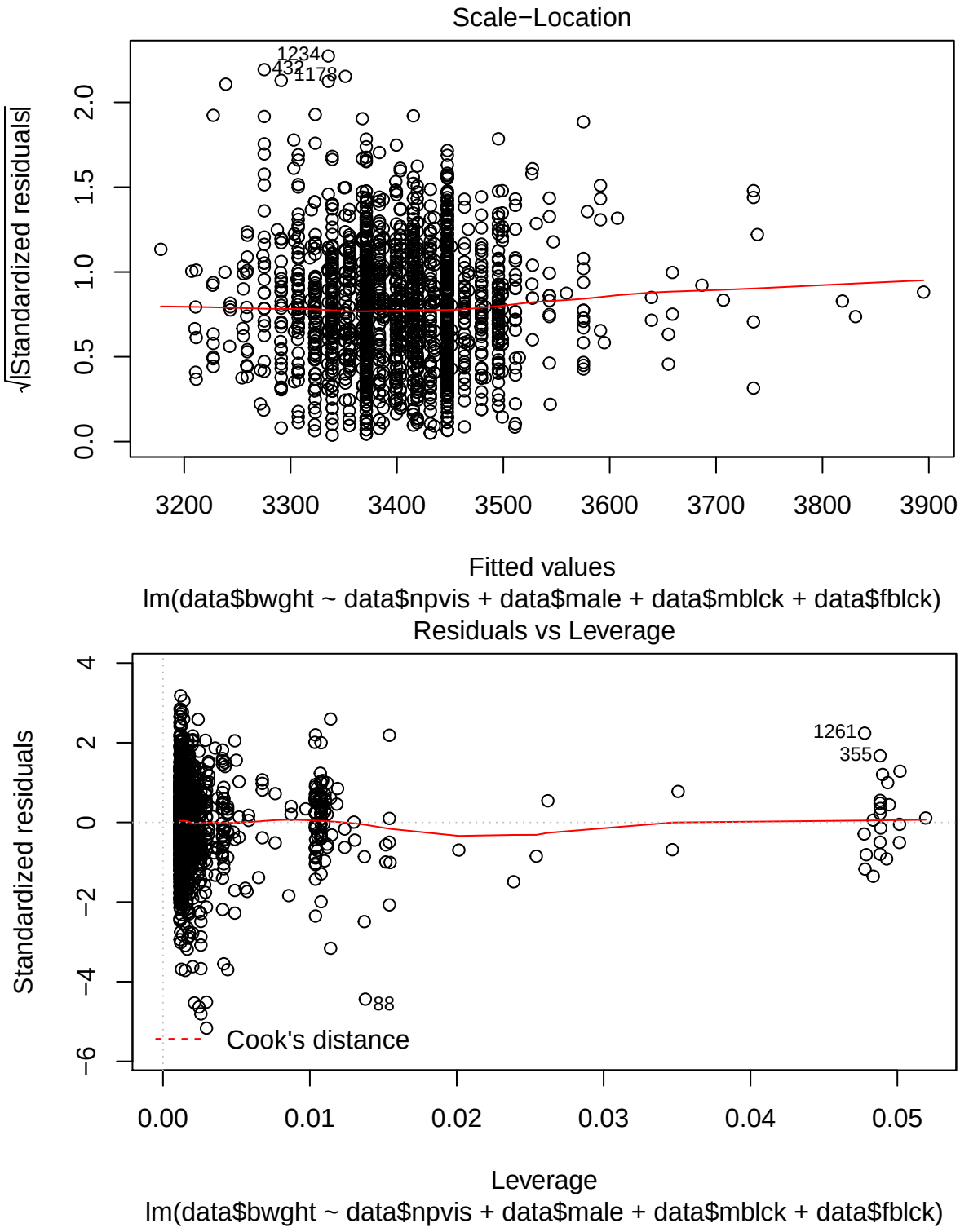
```r
AIC(c1)
```

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

```r
plot(c1)
```


Residuals vs Fitted
Residuals
1000
0
−1000
−3000
3200
3300
3400
3500
3600
3700
3800
3900
1178
1234
Fitted values
lm(data$bwght ~ data$npvis + data$male + data$mblck + data$fblck)


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



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, as the summary statistics on this regressor suggested.

Step 4: CLM and the Models

Step 5: Regression Tables and Model Analysis

Step 6: Causality

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. 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.

Step 7: Conclusion