Lab 4: Does Prenatal Care Improve Infant Health?

Matthew Holmes, Ursula Figueroa, Amy Smessaert & Cassandra Seney

April 26, 2017

Brief

The file bwght_w203.RData contains data from the National Center for Health Statistics and from birth certificates. Your team has been hired by a health advocacy group to study this data and help them understand whether prenatal care improves health outcomes for newborn infants.

Prepare a report addressing the question of whether prenatal care improves newborn health outcomes.

Executive Summary

Following an exploratory data analysis we have constructed models which demonstrate the relationship between prenatal care and health outcomes for newborn infants. In particular we have found that while prenatal care does contribute towards improving the health of newborn, there are other factors which have an impact.

Model Specification

```
# Load the data
load("bwght_w203.RData")
bwdata <- data
```

The variables are defined as follows:

```
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
```

Key explanatory variables

The variables which directly measure prenatal care are 'monpre' which is the month prenatal care began and 'npvis' which is the total number of prenatal visits.

```
# Remove NA's from the monpre data
bwdata <- bwdata[!(is.na(bwdata$monpre)),]

# Month prenatal care began
summary(bwdata$monpre)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   2.000   2.122   2.000   9.000
```

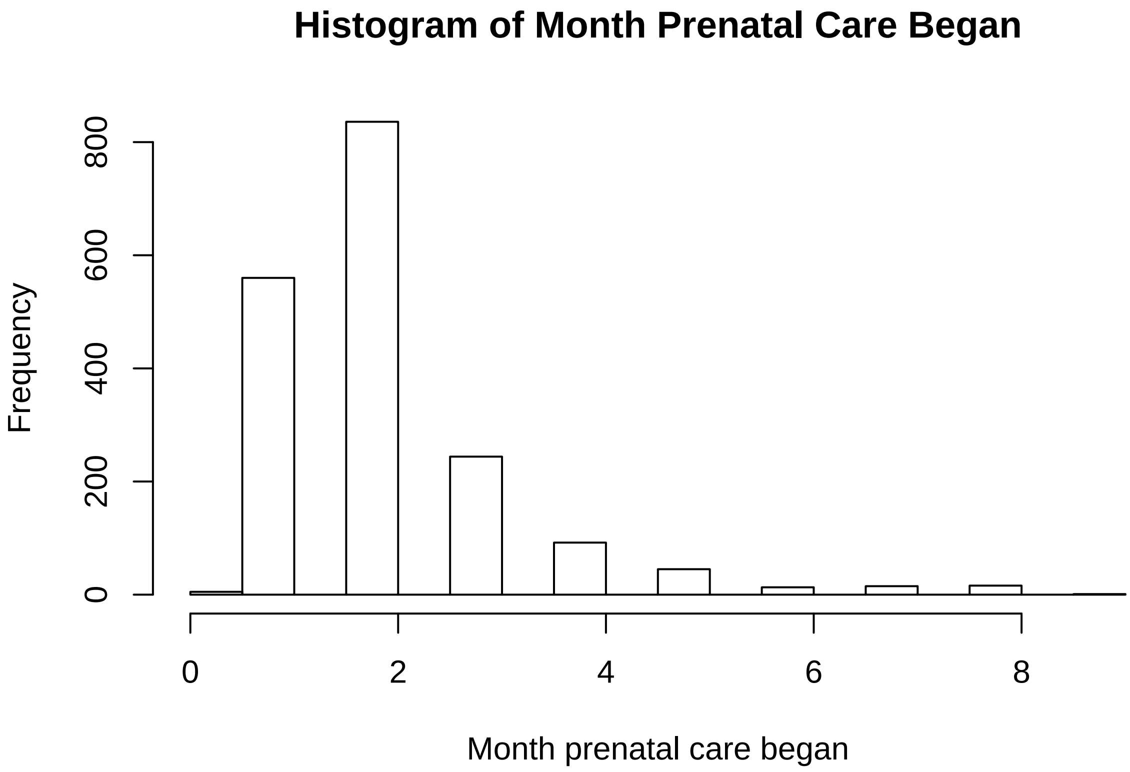
```
# Number of values below 1 and above 9
length(which(data$monpre==0))
```

```
## [1] 5
```

```
length(which(data$monpre==10))
```

```
## [1] 0
```

```
hist(data$monpre, breaks = 21, xlab="Month prenatal care began", main="Histogram of Month Prenatal Care
```



```r
#qqnorm(data$monpre); qqline(data$monpre)
```

Observations:

monpre is an integer variable with a value between 0 and 9. The zero values for monpre correspond to cases where no prenatal care was obtained and are considered valid. Although the mean and median are close together at 2 months, there are data points for the full 9 months which means there is positive skew and a log transform may be useful.

The data points to the right of the graph represent women who started receiving prenatal care later in their pregnancy. We would expect that newborn health outcomes would be better when women obtain prenatal care earlier. Normally, prenatal visits begin about 8 weeks into pregnancy, however if there have been complications with a previous pregnancy or if this is a higher risk pregnancy then they may start sooner. This suggests that starting prenatal visits very early may indicate a fertility issue or it may suggest a health problem which could negatively affect newborn health outcomes. It would useful to know why women started prenatal care earlier or later in their prenancy, since it may indicate another factor contributing to newborn health outcome such as maternal health concerns, income or access to healthcare.

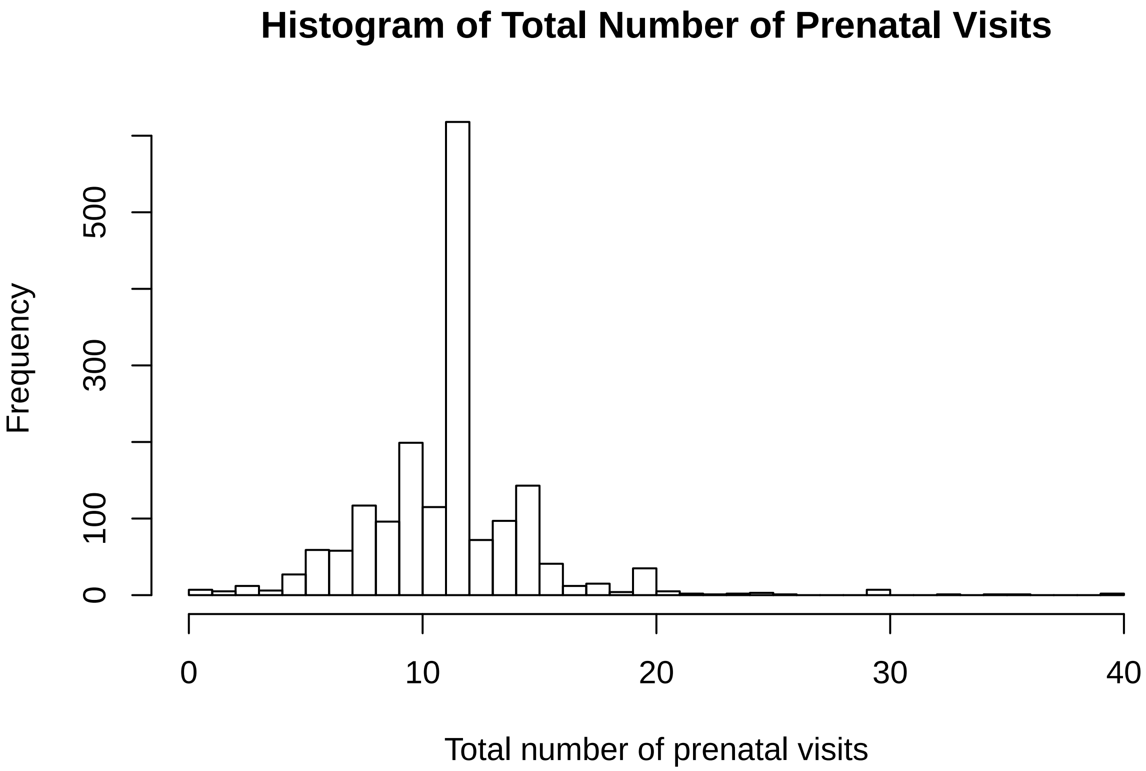
```r
# Remove NA's from the npvis data
bwdata <- bwdata[!(is.na(bwdata$npvis)),]
bwdata <- bwdata[!(is.na(bwdata$npvissq)),]

# Total number of prenatal visits
summary(bwdata$npvis)
```

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

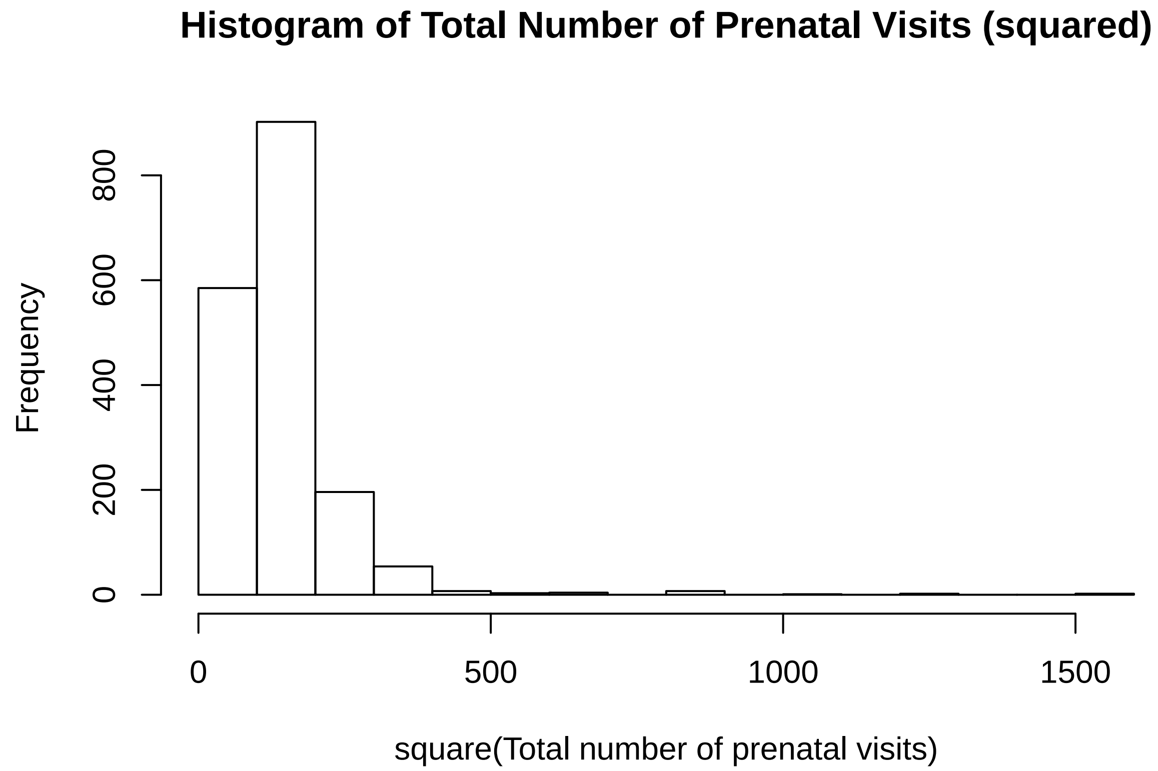
```r
# Total number of prenatal visits squared
summary(bwdata$npvissq)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   100.0   144.0   148.6   169.0  1600.0
```

```r
hist(data$npvis, breaks = 50, xlab="Total number of prenatal visits", main="Histogram of Total Number o
```



```
#qqnorm(data$npvis); qqline(data$npvis)
hist(bwdata$npvissq, xlab="square(Total number of prenatal visits)", main="Histogram of Total Number of
```



```
#qqnorm(data$npvissq); qqline(data$npvissq)
```

Observations:

npvis is an integer variable with a value between 0 and 40. Although the mean and median are close together at 12 visits, and the histogram appears normal, there are outliers causing positive skew. The data points to the right of the graph represent women who received more prenatal care. We would expect newborn health outcomes to improve with more prenatal care, so a positive coefficient would be expected for a linear relationship. However, when the amount of prenatal care required during pregnancy becomes very high this may be related to complications during pregnancy which may result in lower birth outcomes. More information about the health of the mother during pregnancy would help to better interpret the npvis values.

For a healthy pregnancy, the following schedule of prenatal visits is typical:

Weeks 8 to 28: 1 prenatal visit a month = 5 visits total Weeks 28 to 36: 1 prenatal visit every 2 weeks = 4 visits total Weeks 36 to 40: 1 prenatal visit every week = 4 visits total

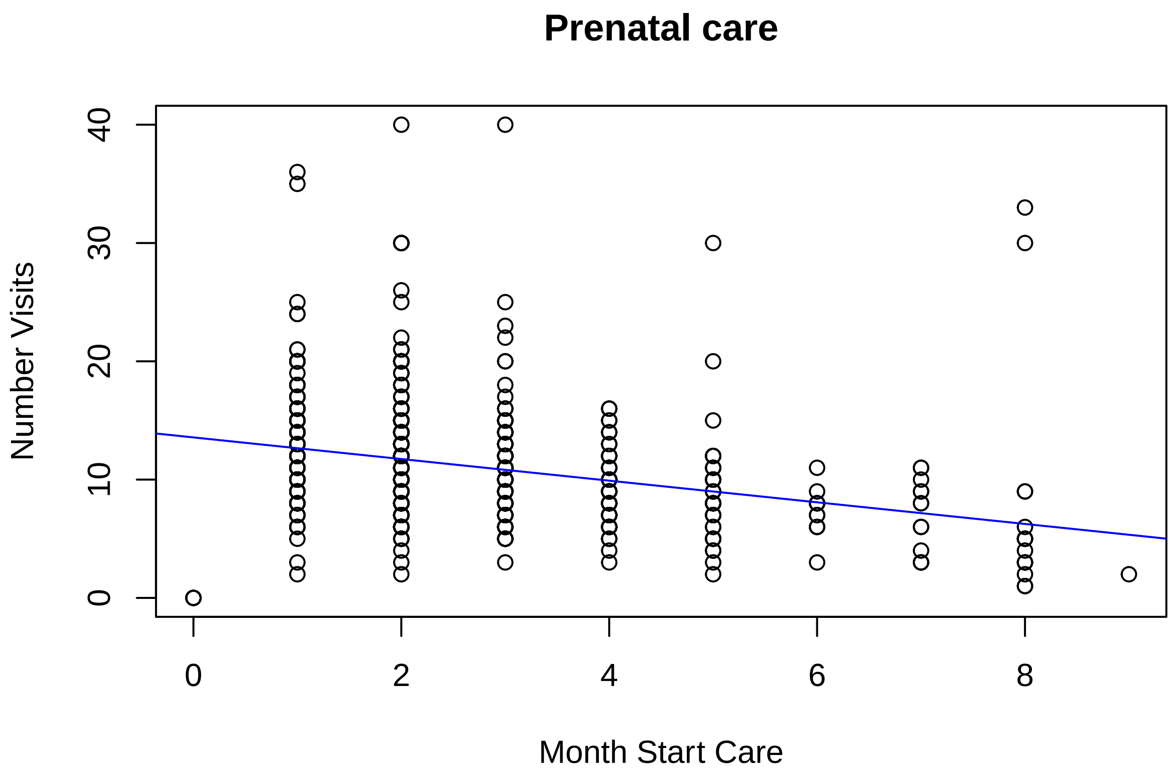
This suggests that 13 visits is typical for a normal pregnancy.

npvissq is the square of total number of prenatal visits. Using the square of npvis introduces more variation in the range of the variable which could reduce the sampling variance of the estimate. Our goal is to find the Best Linear Unbiased Estimators of the regression coefficients where Best is defined as OLS coefficients having the smallest possible variance. The npvissq variable is therefore a good independent variable candidate.

```
#Correlation of monpre and npvis
cor(bwdata$monpre,bwdata$npvis)
```

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

```
plot(bwdata$monpre, bwdata$npvis, xlab = "Month Start Care", ylab = "Number Visits",
main = "Prenatal care")
abline(lm(bwdata$npvis~bwdata$monpre, data=bwdata), col="blue")
```



##### Observations: Although the variables are not strongly correlated, the scatterplot suggests a possible relationship between the time at which prenatal care started and the total number of visits. When care started earlier, the total number of visits was higher, which also makes intuitive sense.

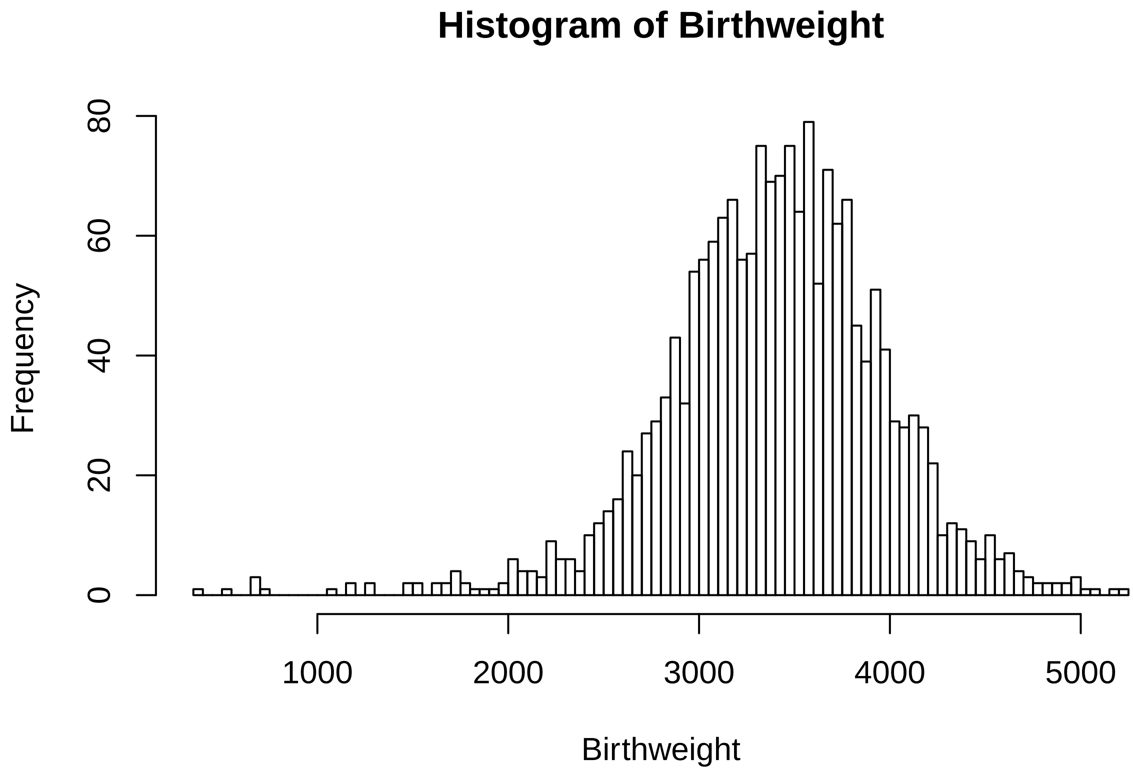
```
bwdata$monpre[bwdata$monpre == 0] <- 0.0001
bwdata$npvis[bwdata$npvis == 0] <- 0.0001
bwdata$log_monpre <- log(bwdata$monpre)
```

Explained variables

```
# Remove NA's from the bwght data
bwdata <- bwdata[!(is.na(bwdata$bwght)),]

# Birth weight in grams
summary(bwdata$bwght)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     360    3080    3430    3402    3770    5204
```

```
hist(data$bwght, breaks = 100, xlab="Birthweight", main="Histogram of Birthweight")
```



```
#qqnorm(data$bwght); qqline(data$bwght)
```

Observations

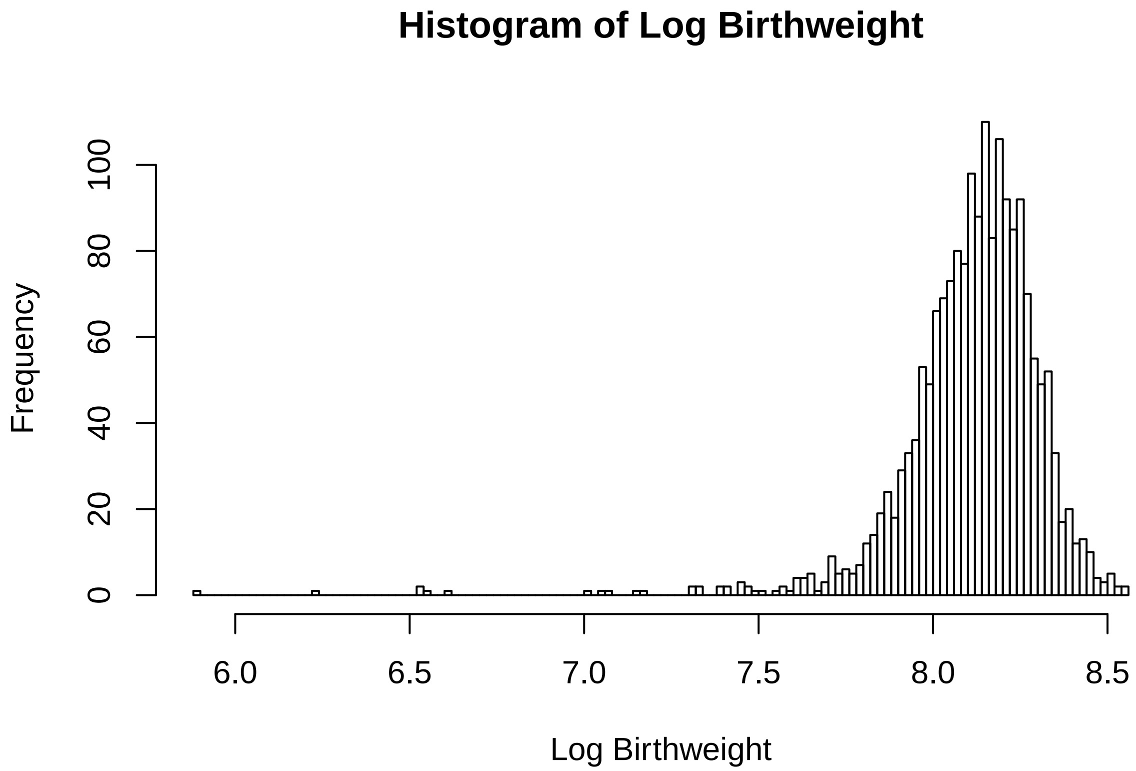
bwght is an integer value between 360 and 5204 grams. The mean and median are close together and the values are normally distributed.

Babies who have very low weight at birth (less than 1,500 grams, or 3.3 pounds) face infant mortality rates more than 100 times that of their heavier peers (more than 2,500 grams, or 5.5 pounds). Mortality among slightly heavier, but still low birthweight babies (between 1,500 and 2,499 grams) is much lower (13 per 1,000), though still 15 times higher than the mortality of babies who are born above that weight (2 per 1,000). Risk factors for low and very low birthweight include multiple births (more than one fetus carried to term), maternal smoking, low maternal weight gain or low pre-pregnancy weight, maternal or fetal stress, infections, and violence toward the pregnant woman.

```
# Log Birth weight
summary(bwdata$lbwght)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.886   8.033   8.140   8.114   8.235   8.557
```

```
hist(data$lbwght, breaks = 100, xlab="Log Birthweight", main="Histogram of Log Birthweight")
```



#### Observations: The data set contains an lbwght variable which is the log transform of bwght. The log transformed variable has a less normal distribution than the variable without the transform. Although we did experiment with including this variable in our models we did not find it useful.

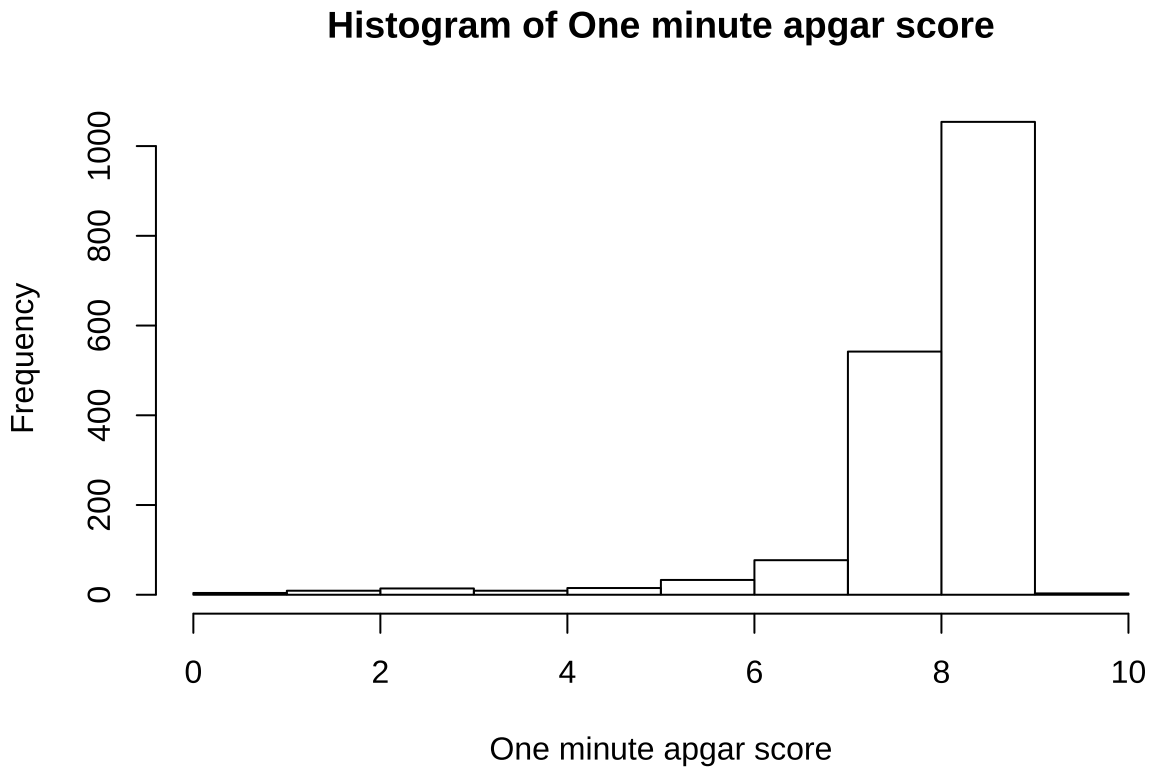
```
# One minute apgar score
round(summary(bwdata$omaps),0)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##       0       8       9       8       9      10       3 
```

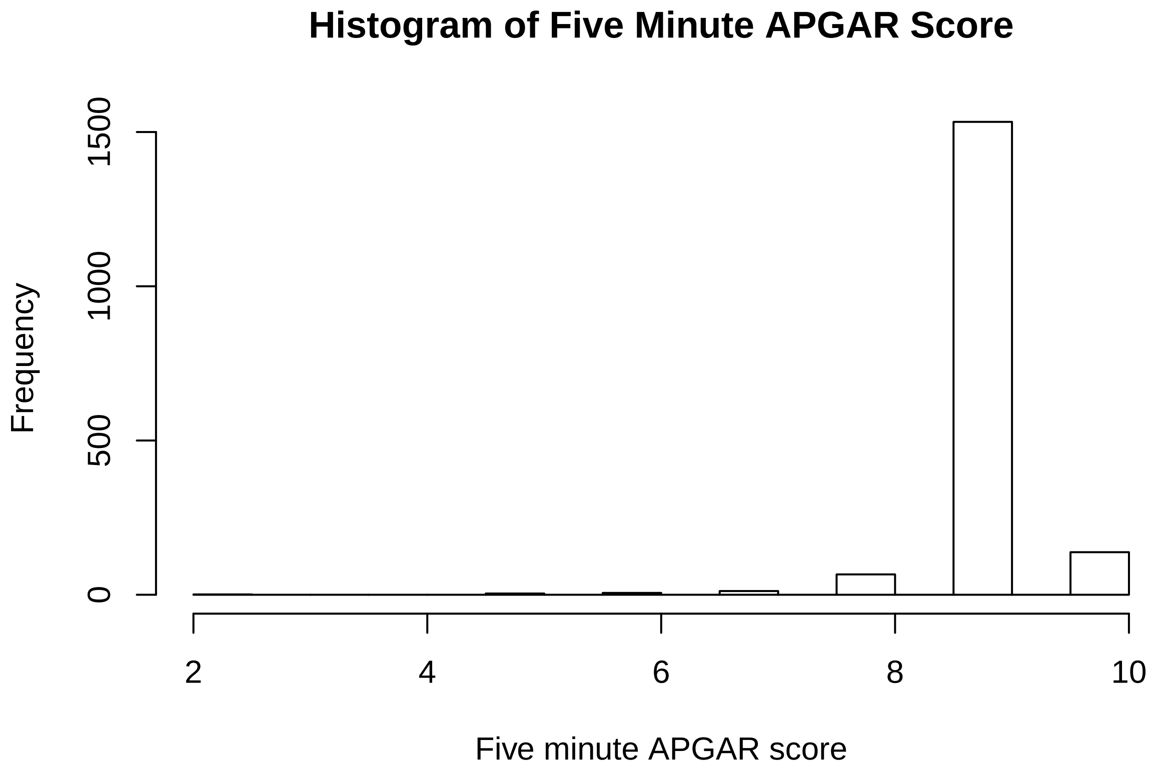
```
# Five minute apgar score
round(summary(bwdata$fmaps),0)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##       2       9       9       9       9      10       3 
```

```
hist(bwdata$omaps, xlab="One minute apgar score", main="Histogram of One minute apgar score")
```



```r
hist(bwdata$fmaps, xlab="Five minute APGAR score", main="Histogram of Five Minute APGAR Score")
```



Observations

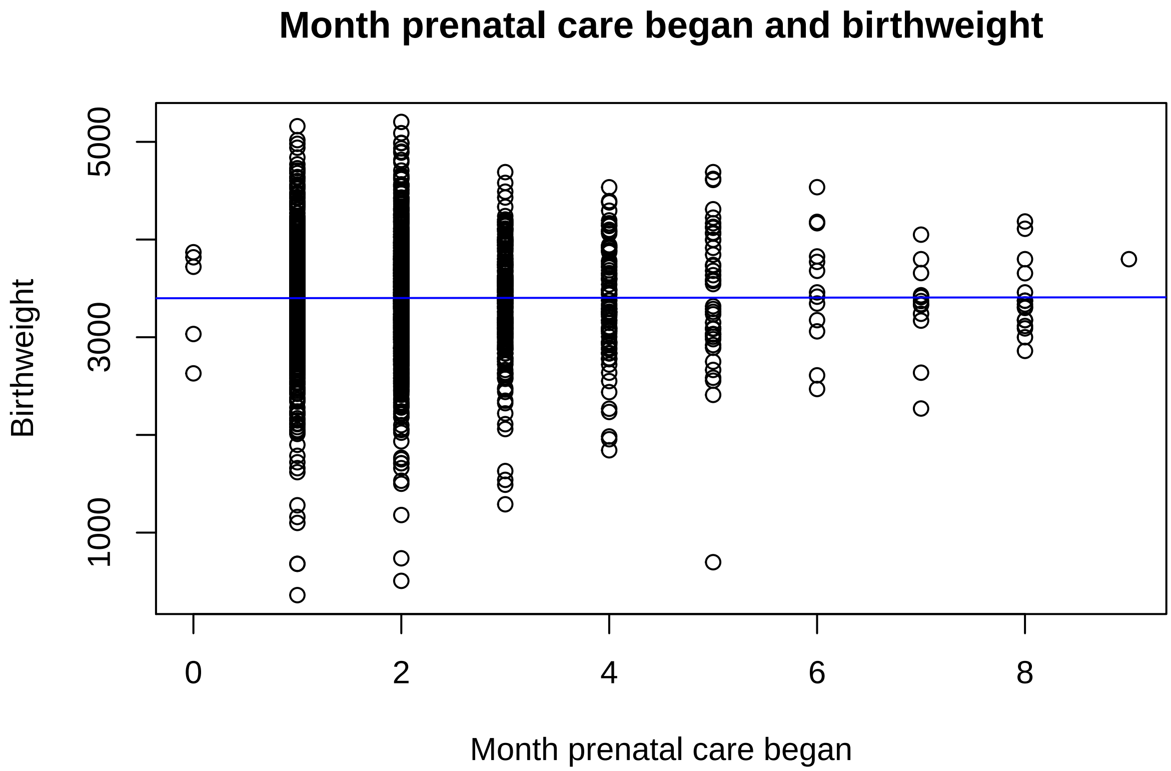
The APGAR scores are both integer values on a scale of 0 to 10. A value of 0 should be rare but not impossible.

The APGAR score is a measure of the need for resuscitation and a predictor of the infant's chances of surviving the first year of life. It is a summary measure of the infant's condition based on heart rate, respiratory effort, muscle tone, reflex irritability, and color. Each of these factors is given a score of 0, 1, or 2; the sum of these 5 values is the APGAR score. A score of 0 to 3 indicates an infant in need of resuscitation; a score of 4 to 6 is considered intermediate; a score of 7 or greater indicates that the neonate is in good to excellent physical condition. The APGAR score is measured at 1 and 5 minutes after delivery.

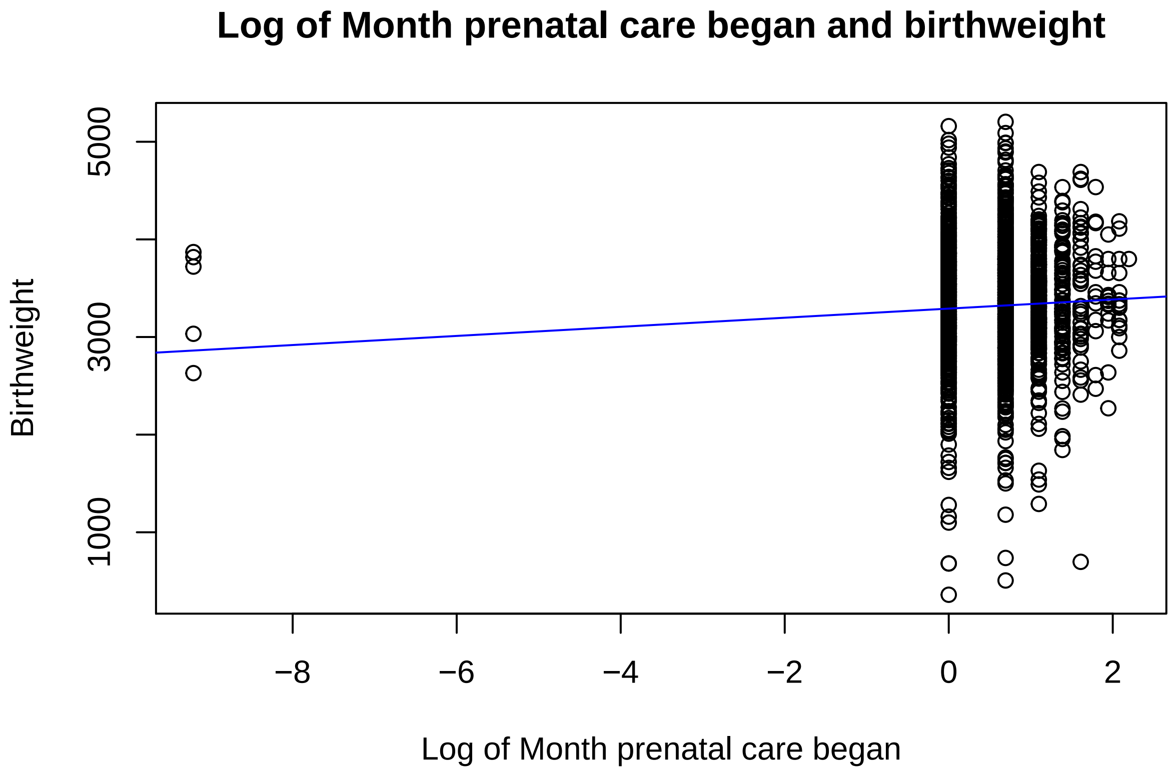
"The Apgar score is an expression of the infant's physiologic condition at one point in time, which includes subjective components. There are numerous factors that can influence the Apgar score, including maternal sedation or anesthesia, congenital malformations, gestational age, trauma, and interobserver variability." [1]

Dependent and Explanatory Variable Exploration

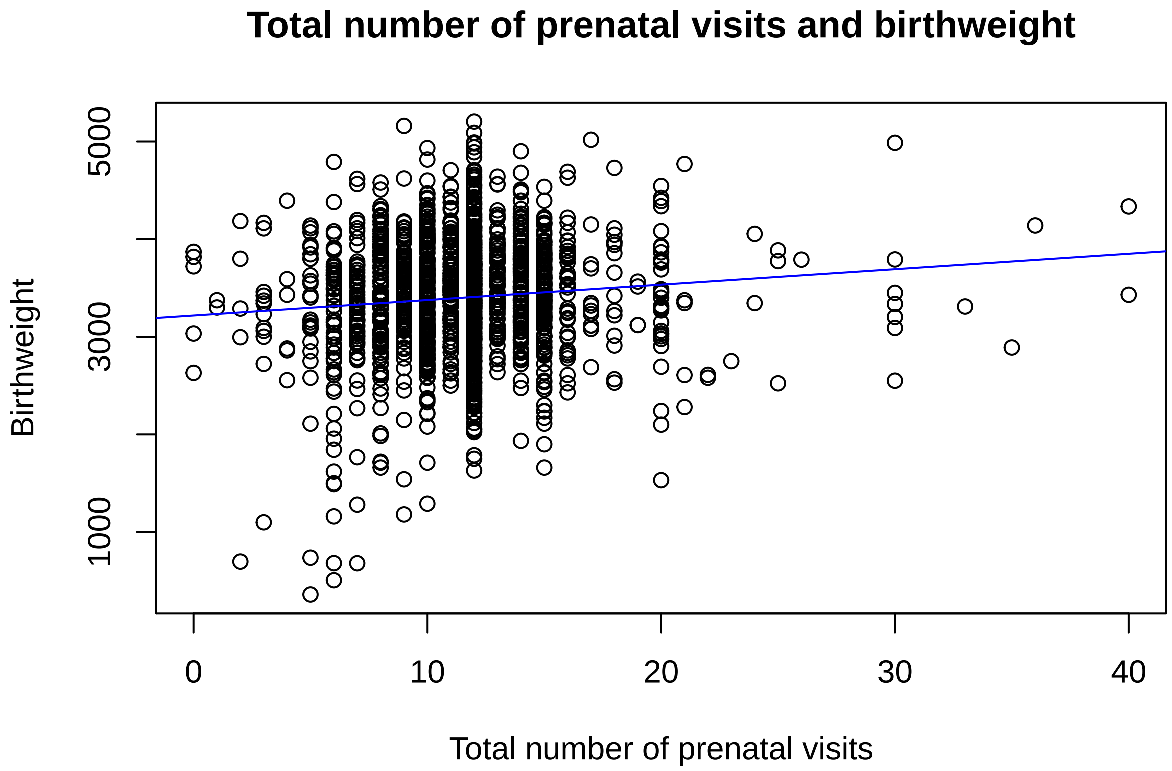
```
# monpre
plot(bwdata$monpre, bwdata$bwght, xlab = "Month prenatal care began", ylab = "Birthweight",
main = "Month prenatal care began and birthweight")
abline(lm(bwdata$bwght ~ bwdata$monpre), col="blue")
```



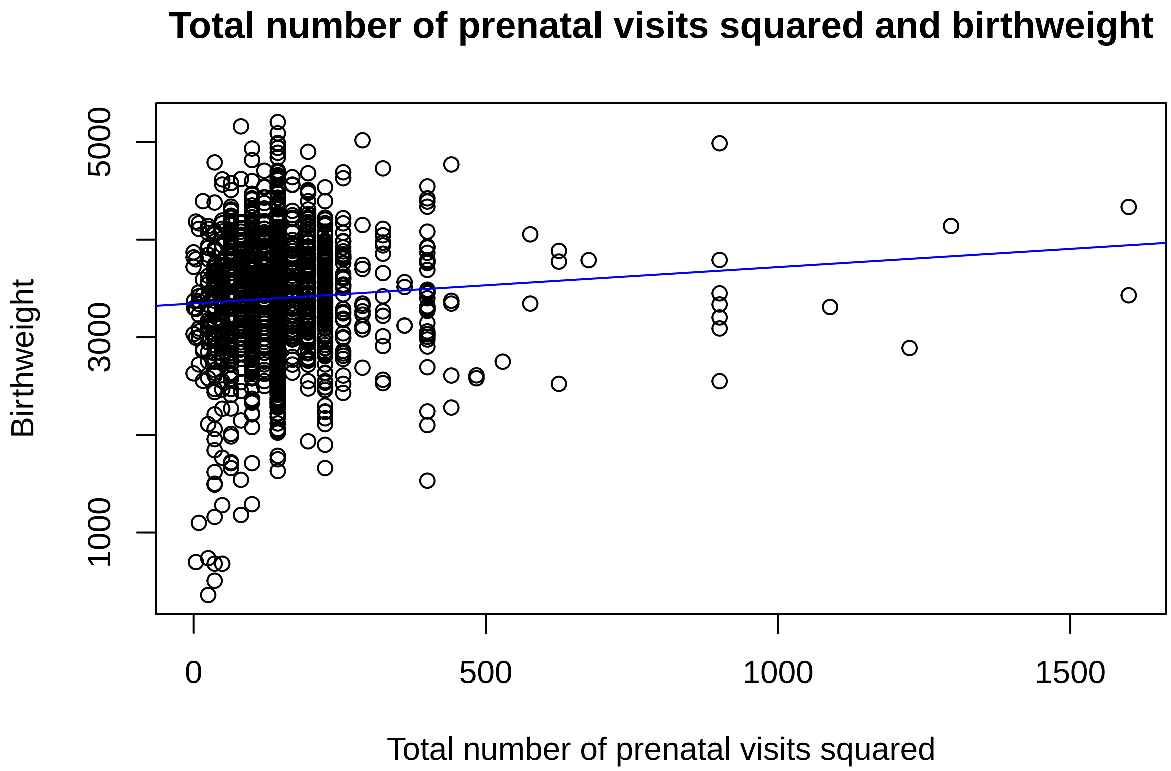
```
# log(monpre)
plot(log(bwdata$monpre), bwdata$bwght, xlab = "Log of Month prenatal care began", ylab = "Birthweight",
main = "Log of Month prenatal care began and birthweight")
abline(lm(bwdata$bwght ~ log(bwdata$npvis)), col="blue")
```



```r
# npvis
plot(bwdata$npvis, bwdata$bwght, xlab = "Total number of prenatal visits", ylab = "Birthweight",
main = "Total number of prenatal visits and birthweight")
abline(lm(bwdata$bwght ~ bwdata$npvis), col="blue")
```



```
# npvis^2
plot(bwdata$npvissq, bwdata$bwght, xlab = "Total number of prenatal visits squared", ylab = "Birthweigh
main = "Total number of prenatal visits squared and birthweight")
abline(lm(bwdata$bwght ~ bwdata$npvissq), col="blue")
```



```
# npvis without outliers
#npvis_mid_data <- subset(data, data$npvis < 20 & data$npvis > 5)
#plot(npvis_mid_data$npvissq, npvis_mid_data$bwght, xlab = "Total number of prenatal visits squared", y
#main = "Total number of prenatal visits squared and birthweight")
#abline(lm(npvis_mid_data$bwght ~ npvis_mid_data$npvissq), col="blue")
```

Model 1

To construct a model we need to select an outcome variable.

We were concerned about the subjectivity of the APGAR scores. We also noted that they can be impacted by factors such as maternal sedation during delivery which has no bearing on prenatal care. "Like birth weight, the Apgar score is an indicator of the overall health of the baby in utero and at birth, but unlike the birth weight, it is not well correlated with some key dimensions of well-being or with future health indicators (CDC, 1981)." [2]

We chose birthweight as our outcome variable.

```
#(model1 <- lm(bwght ~ log(monpre) + npvis, data = bwdata))
(model1 <- lm(bwght ~ log(monpre) + npvissq, data = bwdata))
```

```
##
## Call:
## lm(formula = bwght ~ log(monpre) + npvissq, data = bwdata)
##
## Coefficients:
## (Intercept)  log(monpre)      npvissq
##   3339.9235       9.2201       0.3774
```

```
#(model1 <- lm(lbwght ~ log(monpre) + npvissq, data = bwdata))
#(model1 <- lm(bwght ~ monpre + npvissq, data = bwdata))
#(model1 = lm(bwght ~ poly(monpre,2) + poly(npvis,2), data = bwdata))
```

MLR.1 Linearity

Any population model can be represented as a linear model plus some error. This assumption is valid.

MLR.2 Random sampling

Data must be a random sample drawn from the population. This data set was provided by the client who sourced it from the National Center for Health Statistics (CDC) and from birth certificates. Examining some of the data shows a reasonable population distribution. This assumption is valid.

```
# Male and female newborns are equally well represented in the data
mean(bwdata$male)
```

```
## [1] 0.514464
```

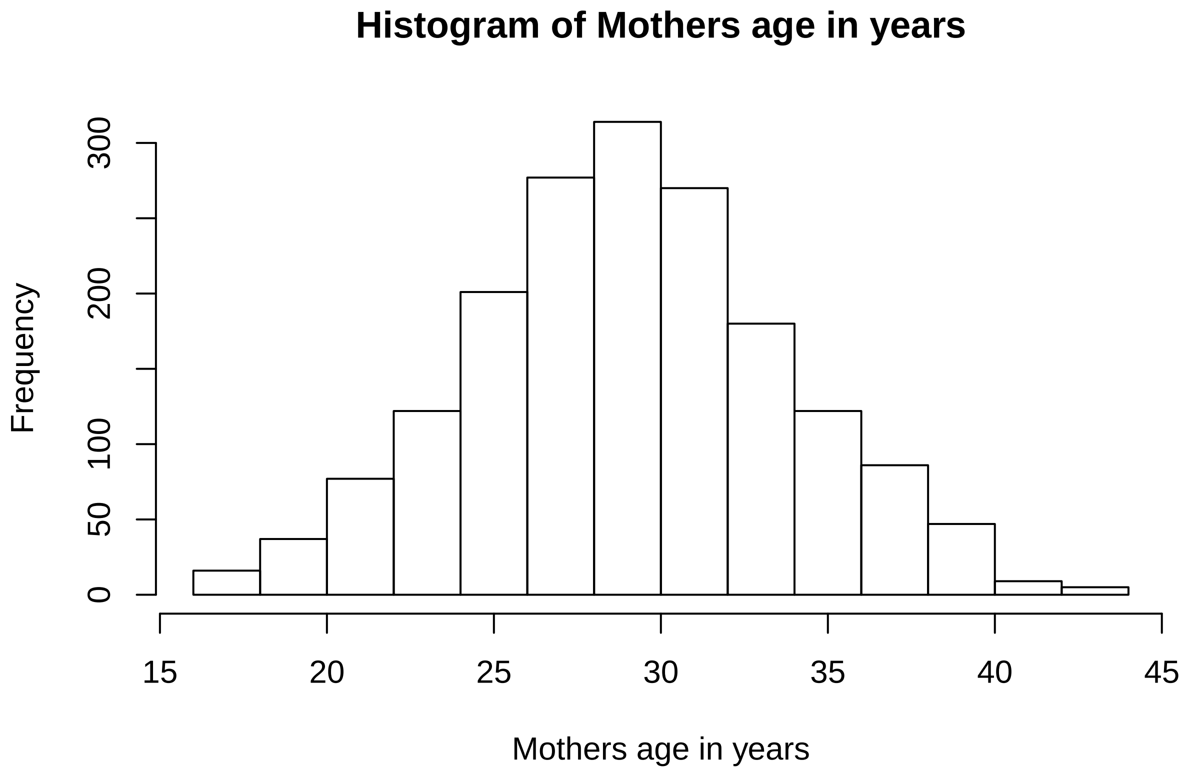
```
# The data set includes white/black/other mothers
(mrace <- c(nrow(subset(bwdata, bwdata$mwhte == 1)), nrow(subset(bwdata, bwdata$mblck == 1)),
            nrow(subset(bwdata, bwdata$moth == 1))))
```

```
## [1] 1559  107   97
```

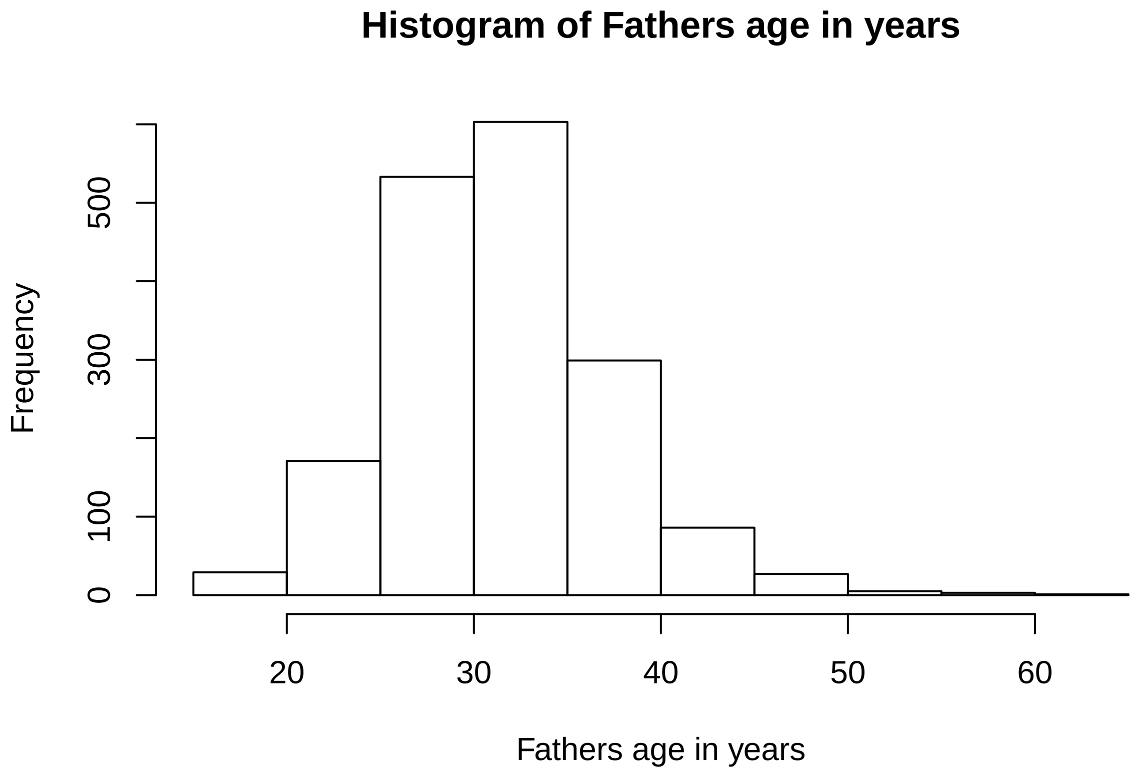
```
# The data set includes white/black/other fathers
(frace <- c(nrow(subset(bwdata, bwdata$fwhte == 1)), nrow(subset(bwdata, bwdata$fblck == 1)),
            nrow(subset(bwdata, bwdata$foth == 1))))
```

```
## [1] 1568  105   90
```

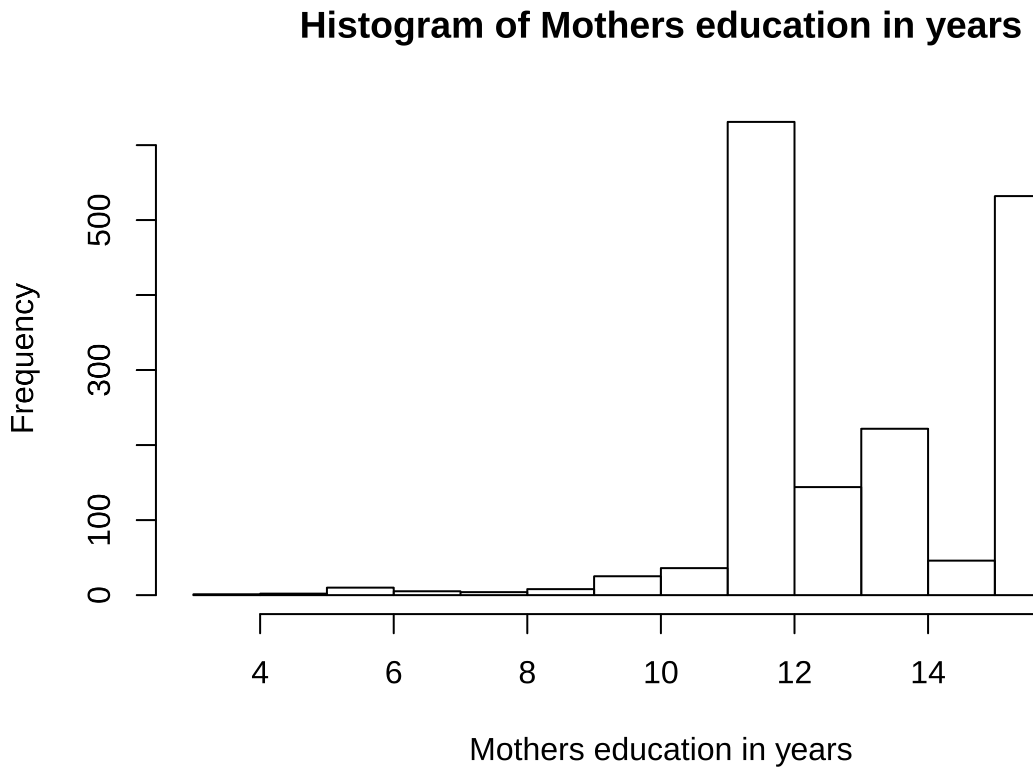
```
# There is a good distribution of maternal and paternal ages
hist(bwdata$mage, xlab="Mothers age in years", main="Histogram of Mothers age in years")
```



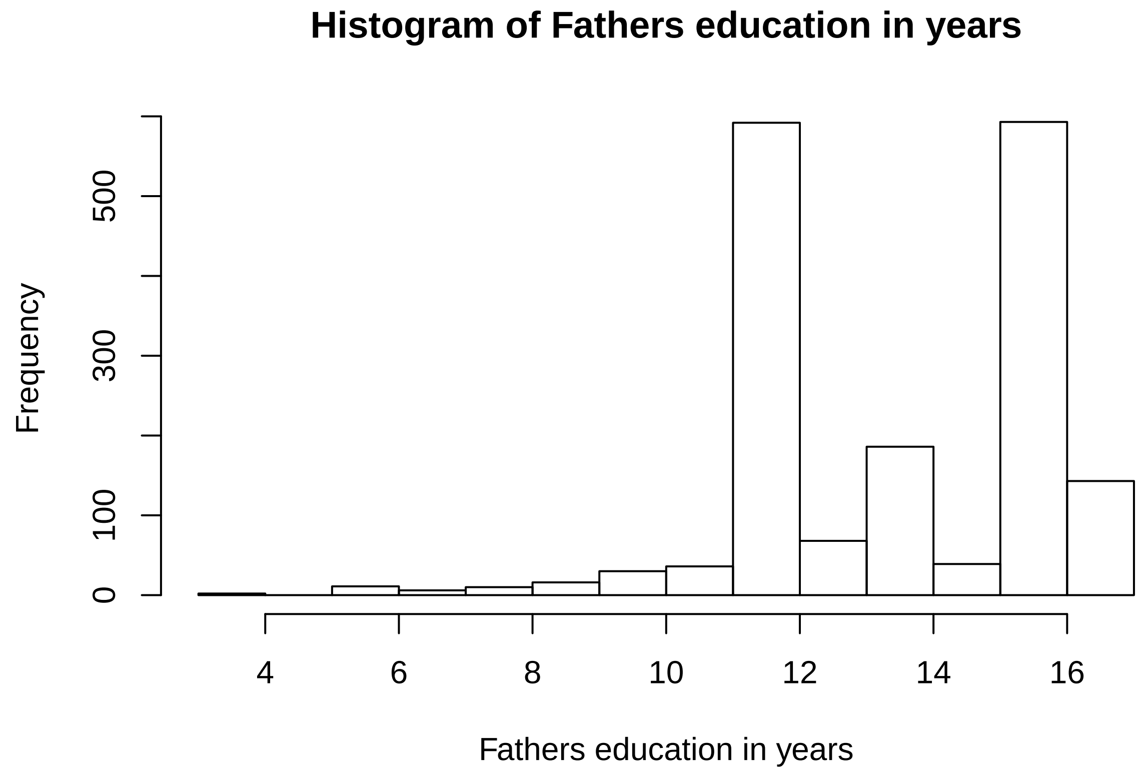
```r
hist(bwdata$fage, xlab="Fathers age in years", main="Histogram of Fathers age in years")
```



```
# There is a good distribution of maternal and paternal education
hist(bwdata$meduc, xlab="Mothers education in years", main="Histogram of Mothers education in years")
```



```
hist(bwdata$feduc, xlab="Fathers education in years", main="Histogram of Fathers education in years")
```



MLR.3 Multicollinearity.

Verify that none of the independent variables are constant (collinear with the intercept), and there are no exact relationships among the independent variables. When variables are highly correlated but not perfectly collinear, OLS will still work but estimates will be much less precise. Check the correlation of the explanatory variables and their Variance Inflation Factors (VIF).

```
# Check the correlation of the independent variables.
cor(bwdata[,c("monpre","npvissq")])
```

```
##              monpre    npvissq
## monpre    1.0000000 -0.1799157
## npvissq -0.1799157  1.0000000
```

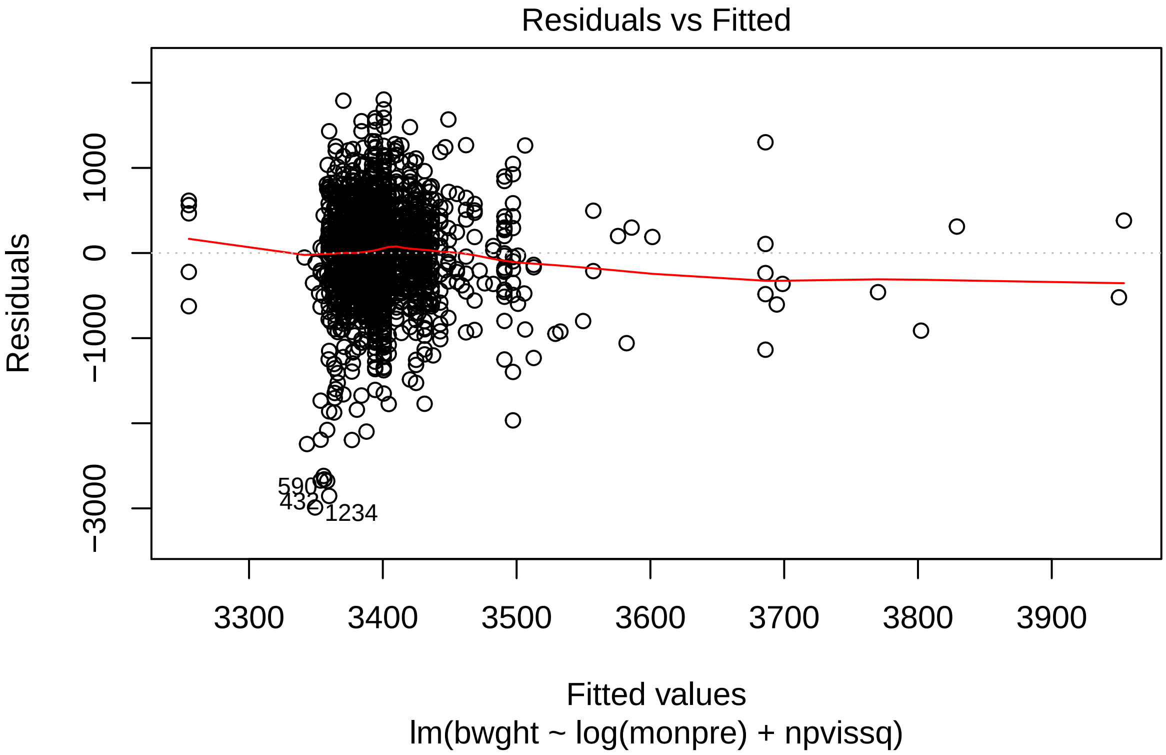
```
#  Variance Inflation Factors
vif(model1)
```

```
## log(monpre)     npvissq
##    1.006548    1.006548
```

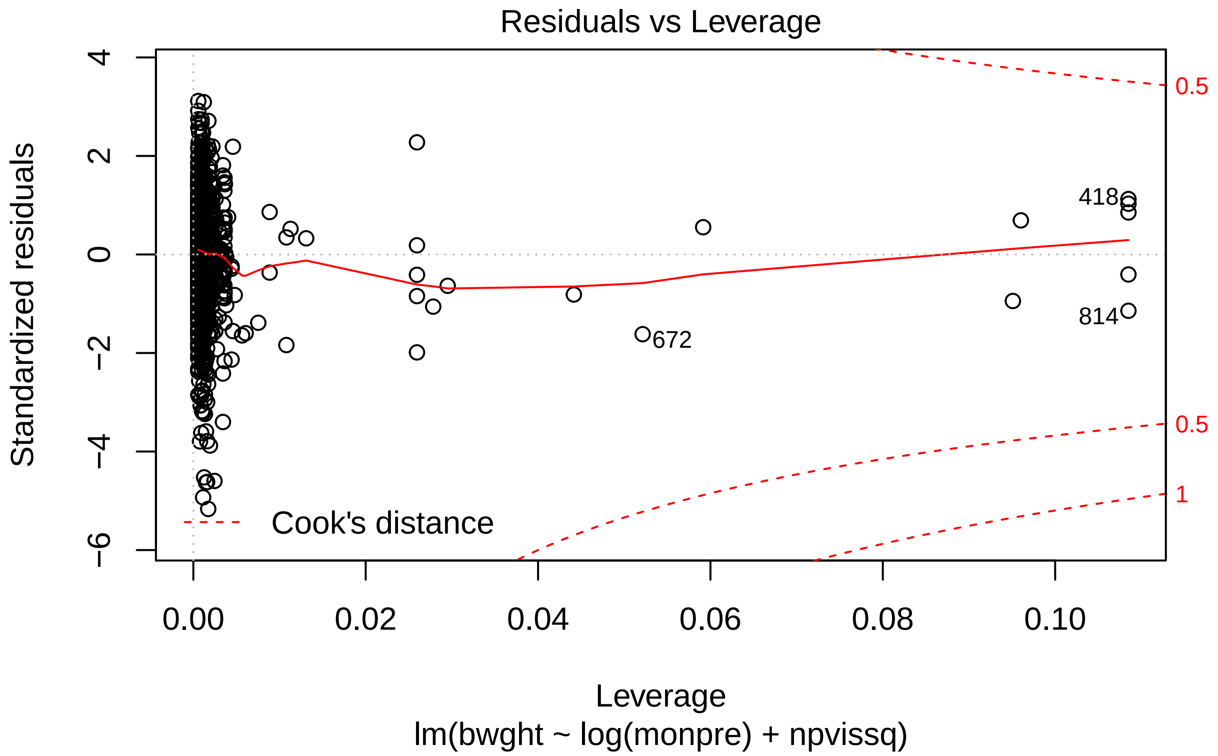
The explanatory variables (monpre, npvis) are not perfectly correlated and the VIFs are low (i.e. less than 10), so there is no perfect multicollinearity of the independent variables. This assumption is valid.

MLR.4 Zero-conditional mean or MLR.4' Exogeneity

```
# Residuals vs fitted to assess zero conditional mean
plot(model1, which = 1)
```



```
plot(model1, which = 5)
```

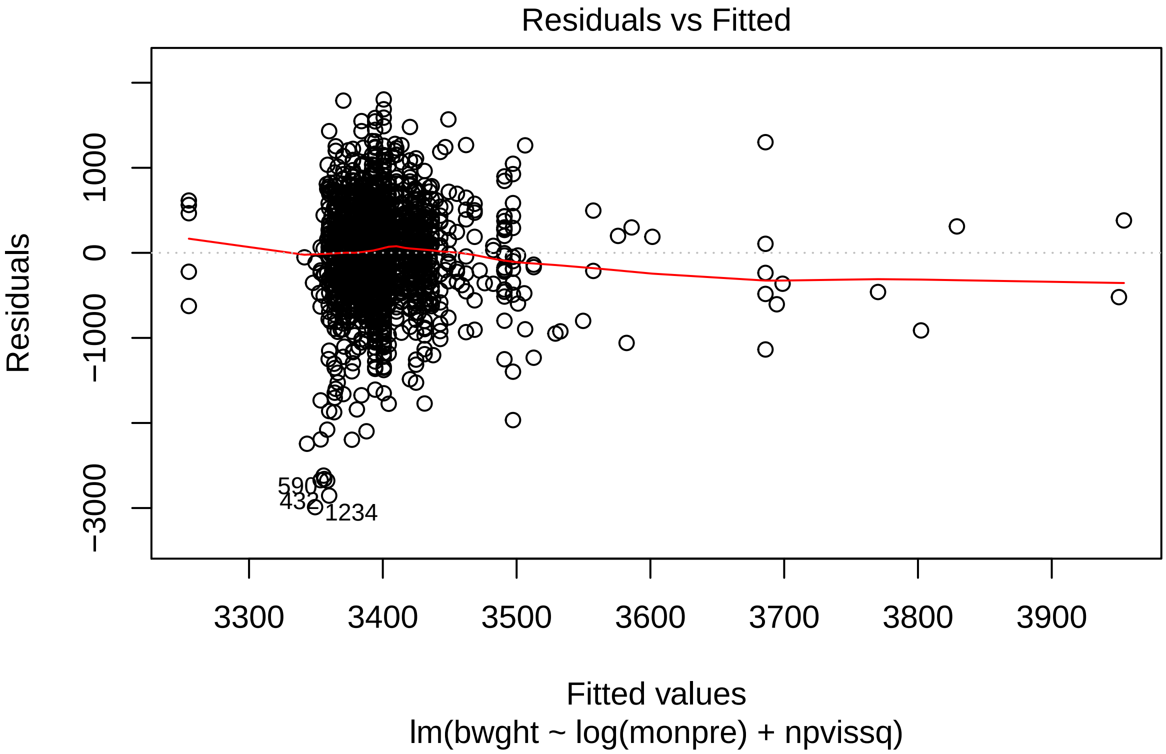


The residuals vs fitted graph looks fairly flat through the data with some variation where there is less observed data. No data point has a large Cook's distance, which means that the outliers do not seem to have significant leverage. The zero-conditional mean assumption is valid.

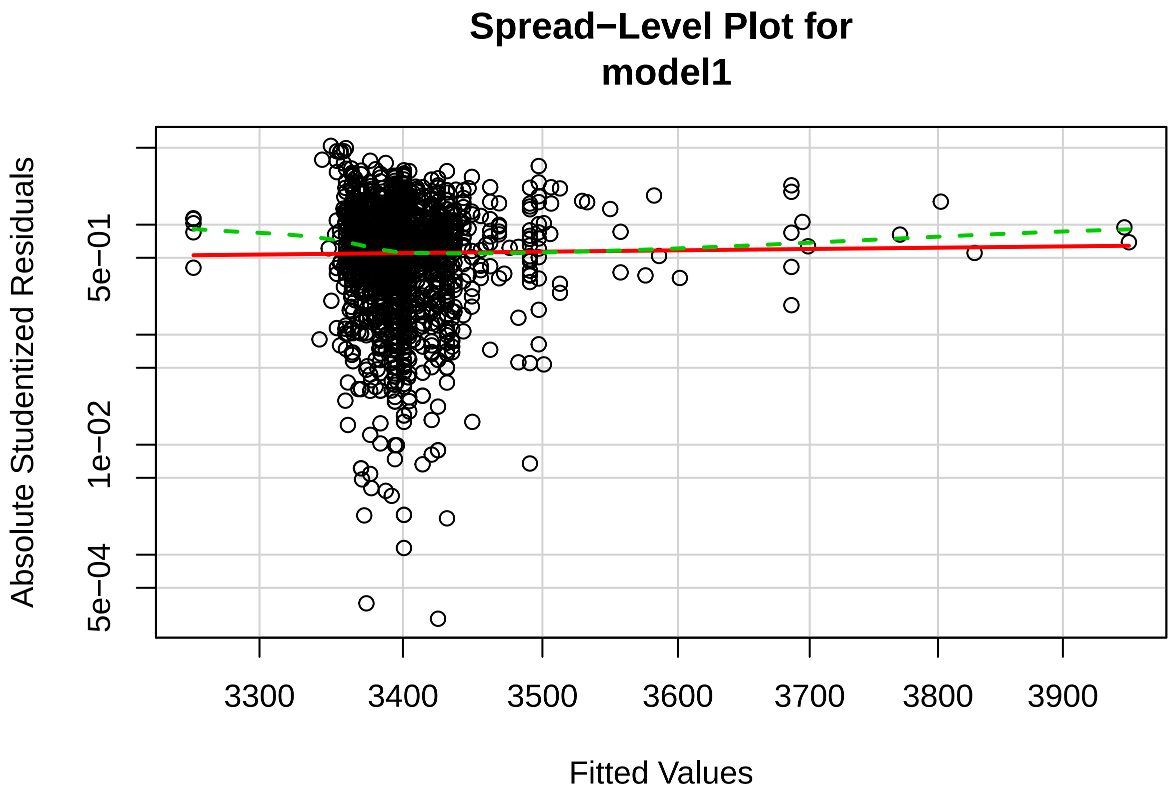
```
# Exogeneity: Explanatory variables that are uncorrelated with the error term are called exogenous.
#(cov(log(bwdata$monpre),model1$residuals))
#(cov((bwdata$npvis)^2,model1$residuals))
#The covariances of the independent variables with the residuals are very
#close to zero indicating they are likely exogenous.
```

MLR.5 Homoskedasticity: The variance of the error term is constant.

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



```r
spreadLevelPlot(model1)
```



```
##
## Suggested power transformation:  -0.01698665
```

```
# Check for the presense of heteroskedasticity with a Breusch-Pagan test.
# Null hypothesis is Homoskedasticity
bptest(model1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 4.4113, df = 2, p-value = 0.1102
```

```
# Non-constant error variance test
ncvTest(model1)
```

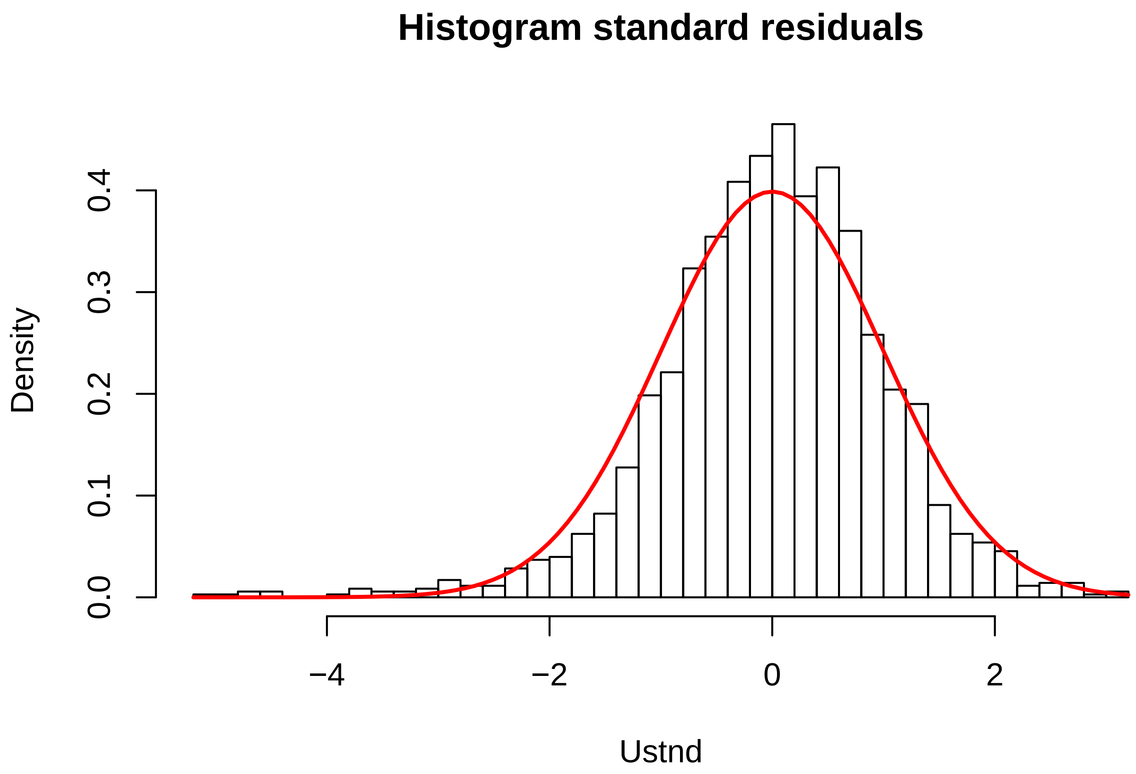
```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 7.143927    Df = 1     p = 0.007521826
```

The Breusch-Pagan test is not significant, which would allow us to accept the hypothesis of homoskedasticity. However the Non-Constant Variance test result was significant, so the tests are giving mixed results. This assumption might hold, but we cannot be completely certain, so we will use robust standard errors.

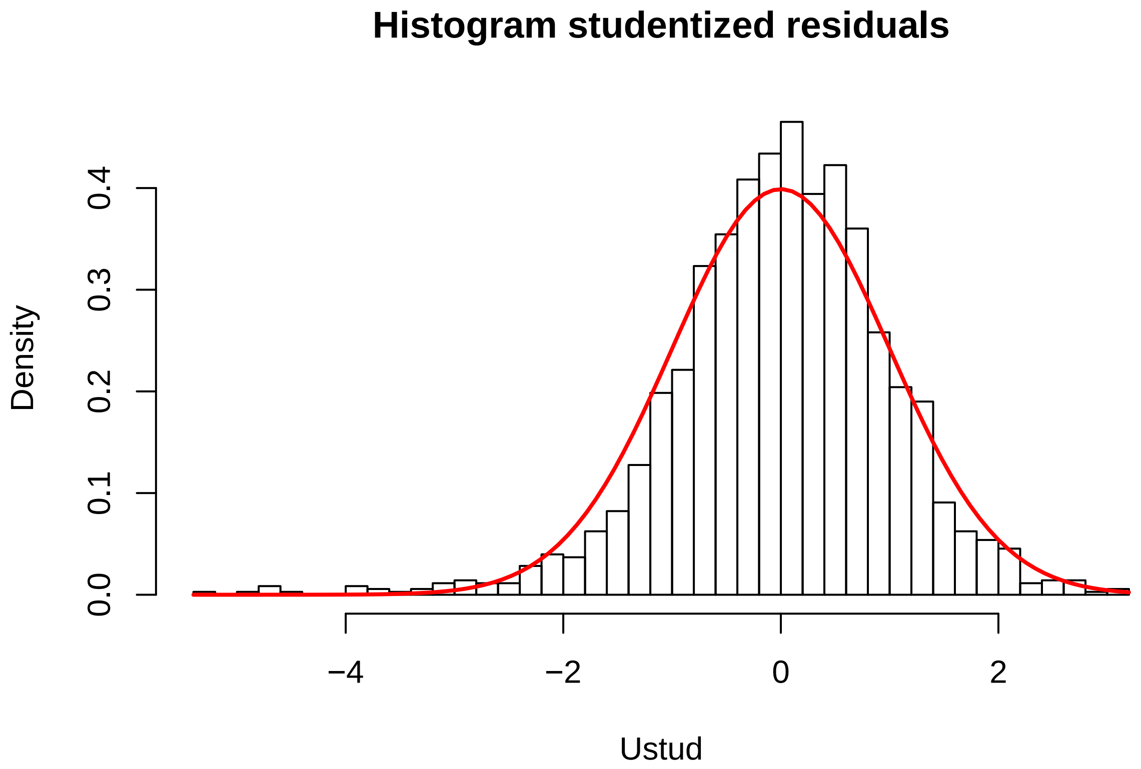
MLR.6 Normality of error terms

```
# normality of standard residuals
Ustnd = rstandard(model1)
```

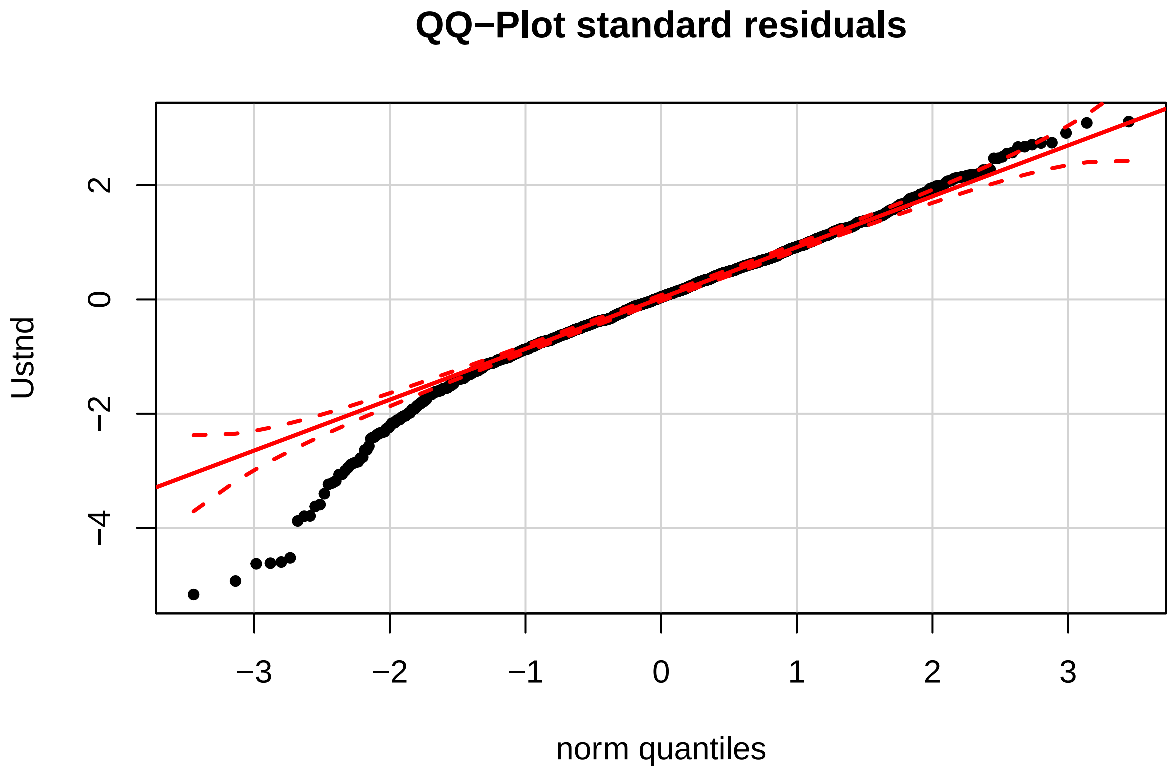
```
hist(Ustnd, main="Histogram standard residuals", breaks = 50, freq=FALSE)
curve(dnorm(x, mean=0, sd=sd(Ustnd)), col="red", lwd=2, add=TRUE)
```



```
# normality of studentized residuals
Ustud = rstudent(model1)
hist(Ustud, main="Histogram studentized residuals", breaks = 50, freq=FALSE)
curve(dnorm(x, mean=0, sd=1), col="red", lwd=2, add=TRUE)
```



```r
# Q-Q plot standard residuals
qqPlot(Ustnd, distribution="norm", pch=20, main="QQ-Plot standard residuals")
qqline(Ustnd, col="red", lwd=2)
```



The histograms are normally distributed and the QQ-Plot also does not deviate from normality significantly either. The normality of error term assumption holds.

Model 2

Adding the gender of the newborn and the race of the mother should increase the accuracy of the results without introducing bias. We expect that newborn gender impact birthweight as male newborn weight is typically higher. We believed that the mother's race may also impact birthweight. These variables should be uncorrelated with variables measuring newborn prenatal visits.

```
#(model2 <- lm(bwght ~ log(monpre) + npvissq + mage, data = bwdata))
#(model2 <- lm(bwght ~ log(monpre) + npvissq + male, data = bwdata))
(model2 <- lm(bwght ~ log(monpre) + npvissq + moth + male, data = bwdata))
```

```
##
## Call:
## lm(formula = bwght ~ log(monpre) + npvissq + moth + male, data = bwdata)
##
## Coefficients:
## (Intercept)  log(monpre)      npvissq         moth         male
##   3308.1633      10.8985       0.3815    -140.5682      73.6274
```

MLR.3 Multicollinearity.

```
# Check the correlation of the independent variables.
cor(bwdata[,c("monpre","npvissq", "male","moth")])
```

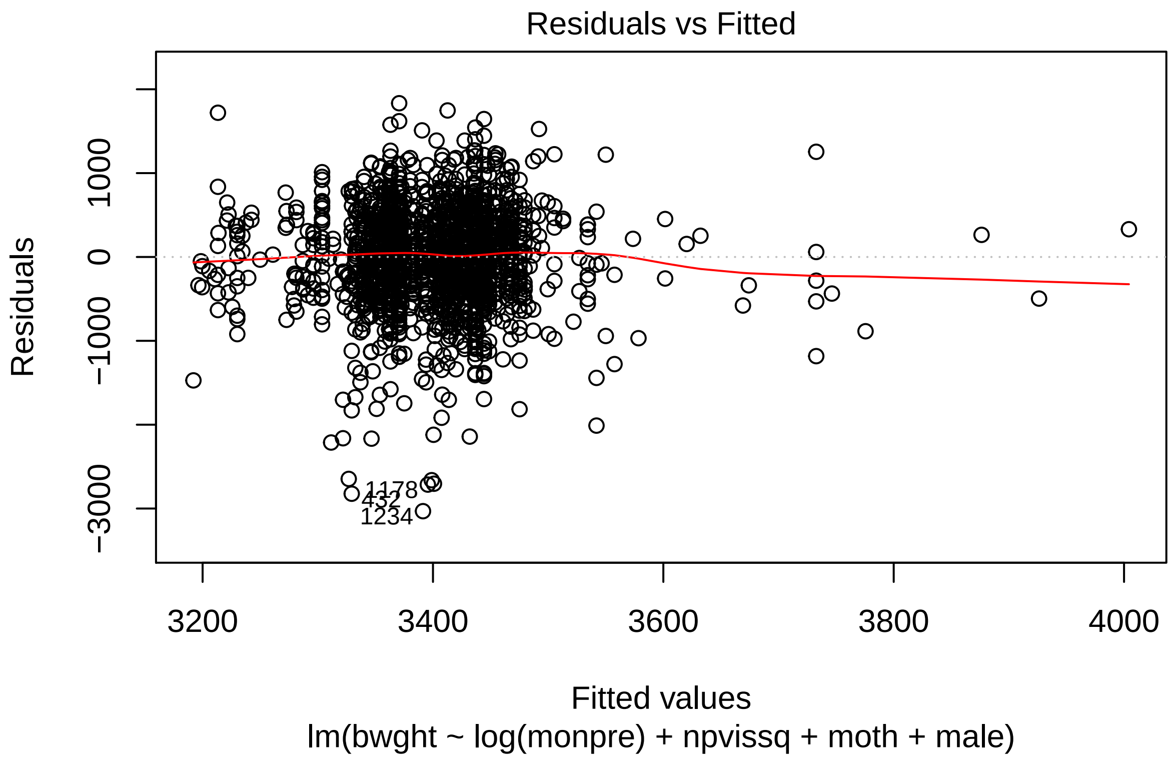
```
##                 monpre      npvissq         male         moth
## monpre   1.000000000 -0.179915685 -0.008091111 -0.013103639
## npvissq -0.179915685  1.000000000 -0.012658376 -0.003126214
## male    -0.008091111 -0.012658376  1.000000000  0.015414475
## moth    -0.013103639 -0.003126214  0.015414475  1.000000000
```

```r
#Variance Inflation Factors
vif(model2)
```

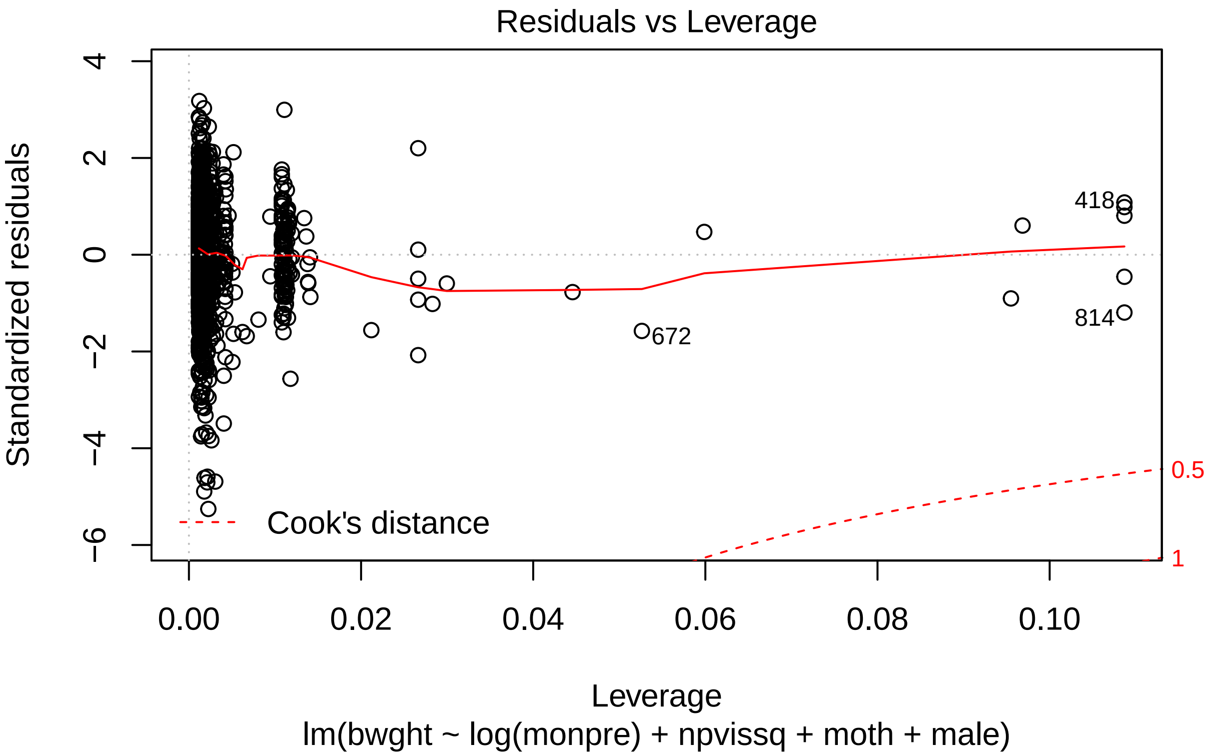
```
## log(monpre)     npvissq        moth        male
##    1.007594    1.006789    1.000248    1.001436
```

MLR.4 Zero-conditional mean or MLR.4' Exogeneity

```r
# Residuals vs fitted to assess zero conditional mean
plot(model2, which = 1)
```

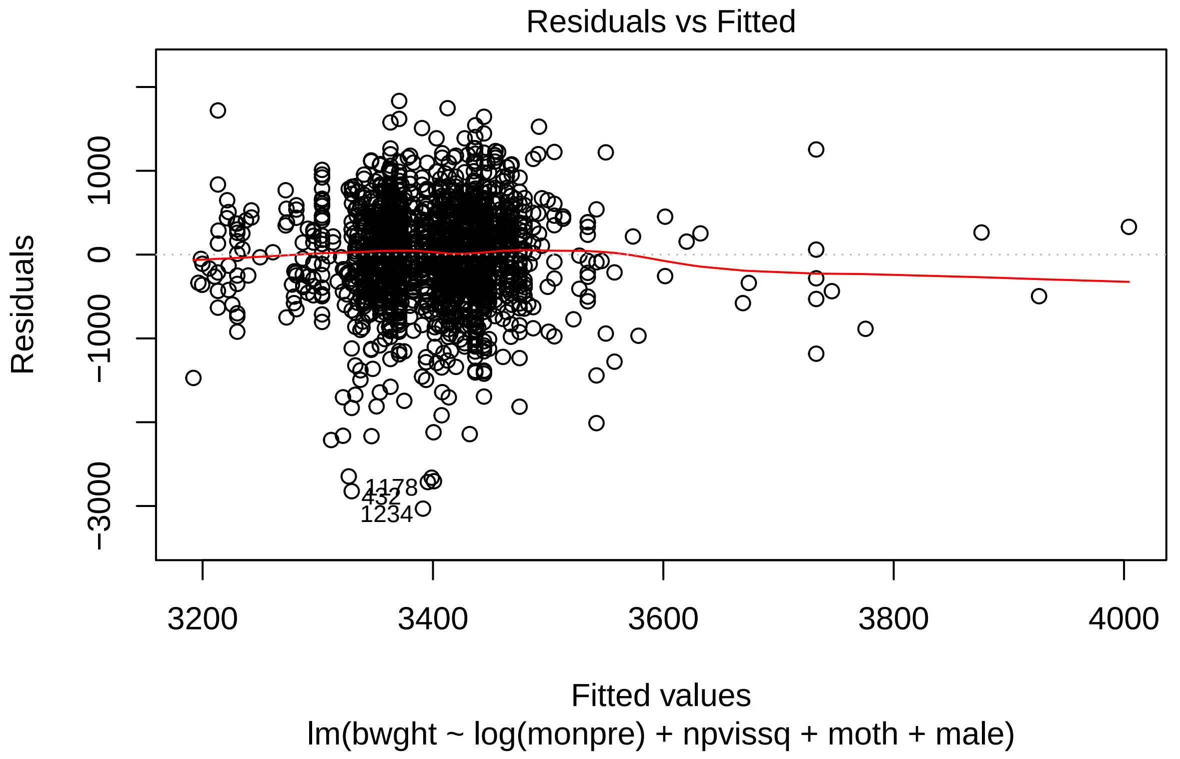


```r
plot(model2, which = 5)
```

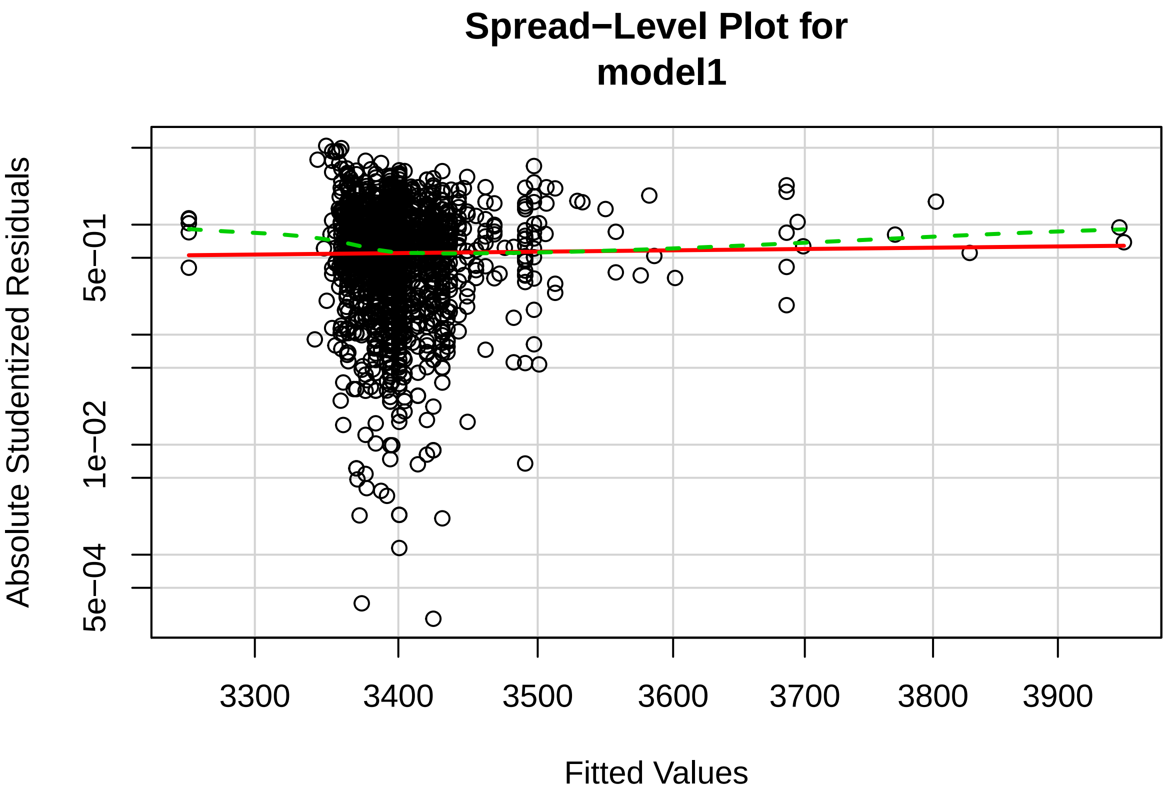


MLR.5 Homoskedasticity: The variance of the error term is constant.

```
plot(model2, which = 1)
```



```r
spreadLevelPlot(model1)
```



```
##
## Suggested power transformation:  -0.01698665
```

```
# Check for the presense of heteroskedasticity with a Breusch-Pagan test.  Null hypothesis is Homoskeda
bptest(model2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model2
## BP = 7.6523, df = 4, p-value = 0.1052
```

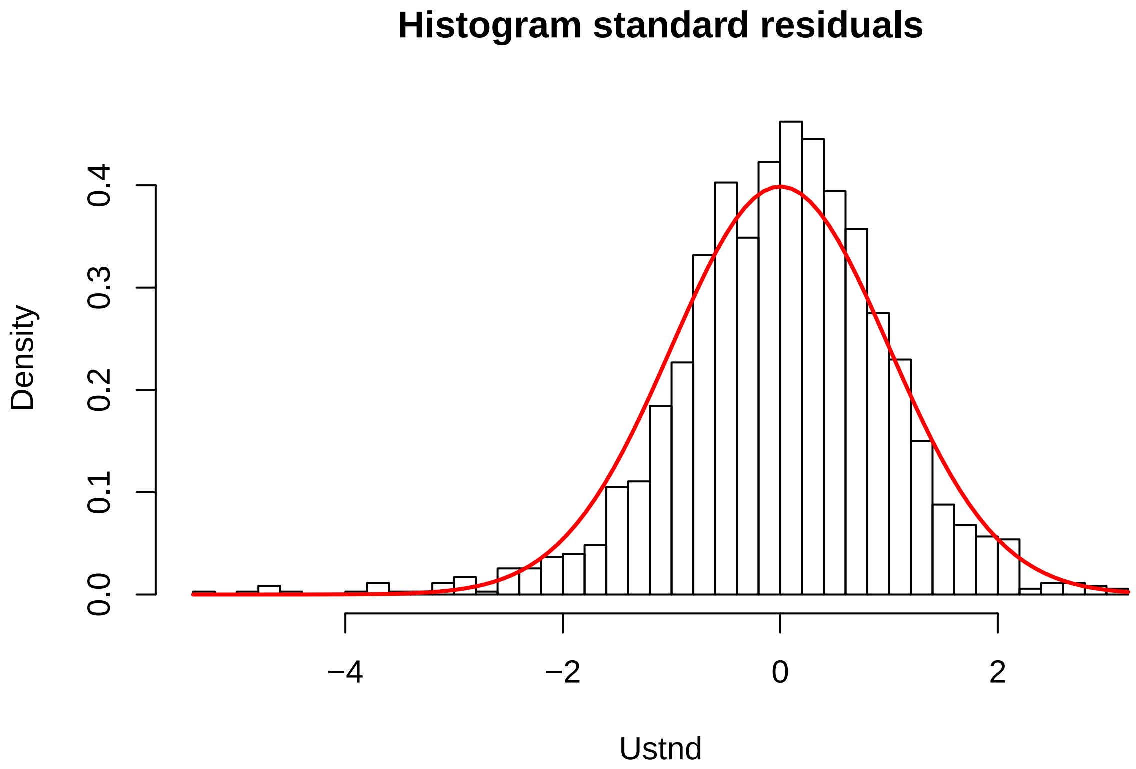
```
# Non-constant error variance test
ncvTest(model2)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.001592595    Df = 1     p = 0.968167
```

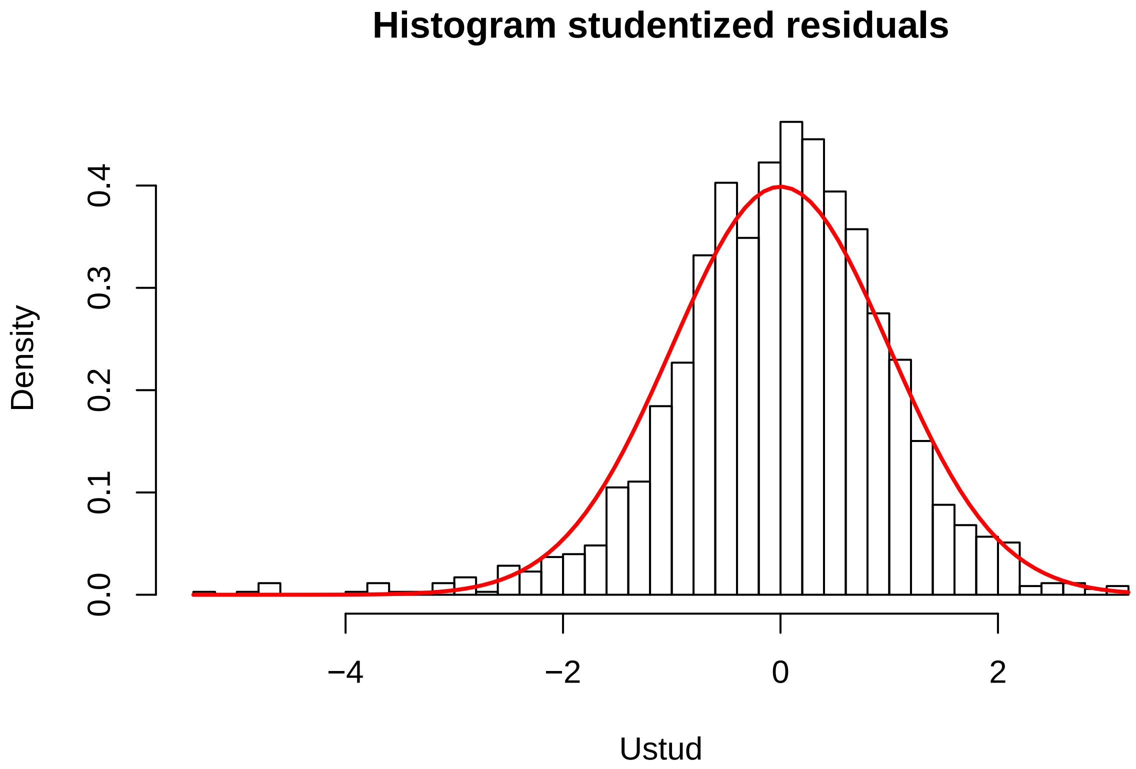
This model demonstrates Homoskedasticity more conclusively, with both of the tests allowing us to accept the null hypothesis.

MLR.6 Normality of error terms

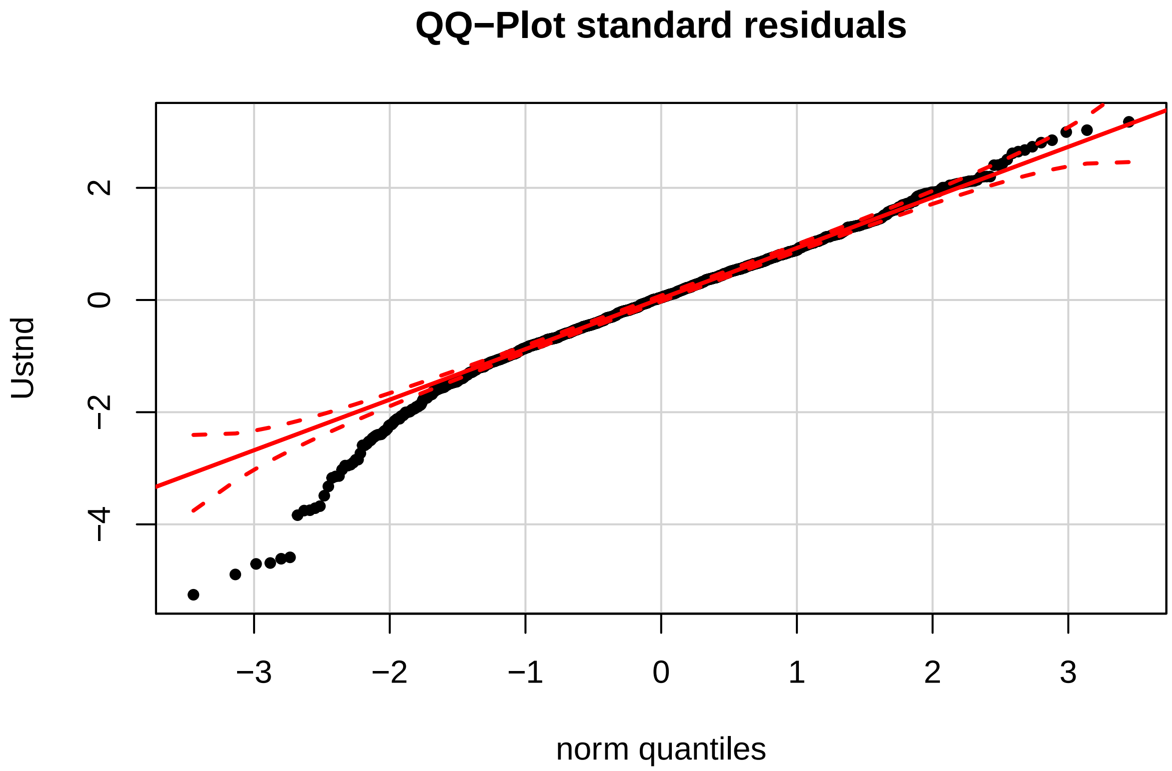
```
# normality of standard residuals
Ustnd = rstandard(model2)
hist(Ustnd, main="Histogram standard residuals", breaks = 50, freq=FALSE)
curve(dnorm(x, mean=0, sd=sd(Ustnd)), col="red", lwd=2, add=TRUE)
```



```
# normality of studentized residuals
Ustud = rstudent(model2)
hist(Ustud, main="Histogram studentized residuals", breaks = 50, freq=FALSE)
curve(dnorm(x, mean=0, sd=1), col="red", lwd=2, add=TRUE)
```



```
# Q-Q plot standard residuals
qqPlot(Ustnd, distribution="norm", pch=20, main="QQ-Plot standard residuals")
qqline(Ustnd, col="red", lwd=2)
```



Model 3

For the third model we decided to add variables mother's age, daily cigarette consumption, weekly drinks and mother's education.

```
bwdata <- bwdata[!(is.na(bwdata$mage)),]

bwdata <- bwdata[!(is.na(bwdata$cigs)),]

bwdata <- bwdata[!(is.na(bwdata$drink)),]

bwdata <- bwdata[!(is.na(bwdata$meduc)),]

#plot(bwdata$mage, bwdata$npvis, xlab = "Mother Age", ylab = "Number Visits", main = "Visits by Age")

(model3 <- lm(bwght ~ log(monpre) + npvissq + moth + male + mage +cigs + drink+ meduc, data = bwdata))
```

```
##
## Call:
## lm(formula = bwght ~ log(monpre) + npvissq + moth + male + mage +
##     cigs + drink + meduc, data = bwdata)
##
## Coefficients:
## (Intercept)  log(monpre)      npvissq         moth         male
##    3151.817        9.452        0.309     -165.500       82.599
##        mage         cigs        drink        meduc
##       4.462      -11.365      -11.731        4.016
```

MLR.3 Multicollinearity.

Check the correlation of the explanatory variables and their Variance Inflation Factors (VIF).

```
# Check the correlation of the independent variables.
cor(bwdata[,c("monpre","npvissq","moth","male","mage", "cigs","drink", "meduc")])
```

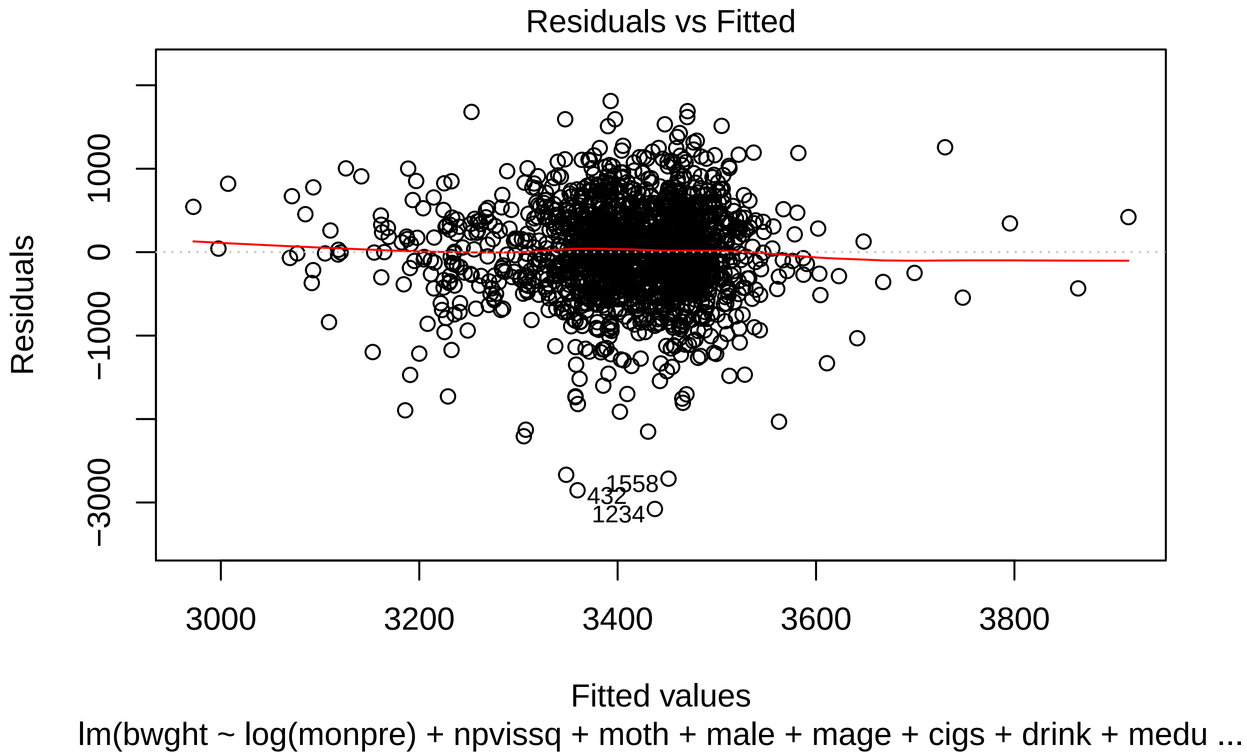
```
##              monpre       npvissq          moth         male         mage
## monpre   1.00000000 -0.1835297449 -0.0143086394 -0.021699022 -0.197188536
## npvissq -0.18352974  1.0000000000  0.0001830714 -0.010440596  0.056899560
## moth    -0.01430864  0.0001830714  1.0000000000  0.005245361  0.022820671
## male    -0.02169902 -0.0104405965  0.0052453611  1.000000000 -0.039802890
## mage    -0.19718854  0.0568995599  0.0228206713 -0.039802890  1.000000000
## cigs     0.09874793  0.0080019577 -0.0357602883 -0.010813825 -0.060800074
## drink   -0.01056430  0.0503304039 -0.0172507066 -0.047753940  0.004593838
## meduc   -0.18802956  0.0691869266  0.1656752493  0.030045983  0.329836911
##                 cigs        drink       meduc
## monpre   0.098747934 -0.010564303 -0.18802956
## npvissq  0.008001958  0.050330404  0.06918693
## moth    -0.035760288 -0.017250707  0.16567525
## male    -0.010813825 -0.047753940  0.03004598
## mage    -0.060800074  0.004593838  0.32983691
## cigs     1.000000000  0.185483896 -0.14750616
## drink    0.185483896  1.000000000 -0.02061634
## meduc   -0.147506160 -0.020616343  1.00000000
```

```
vif(model3)
```

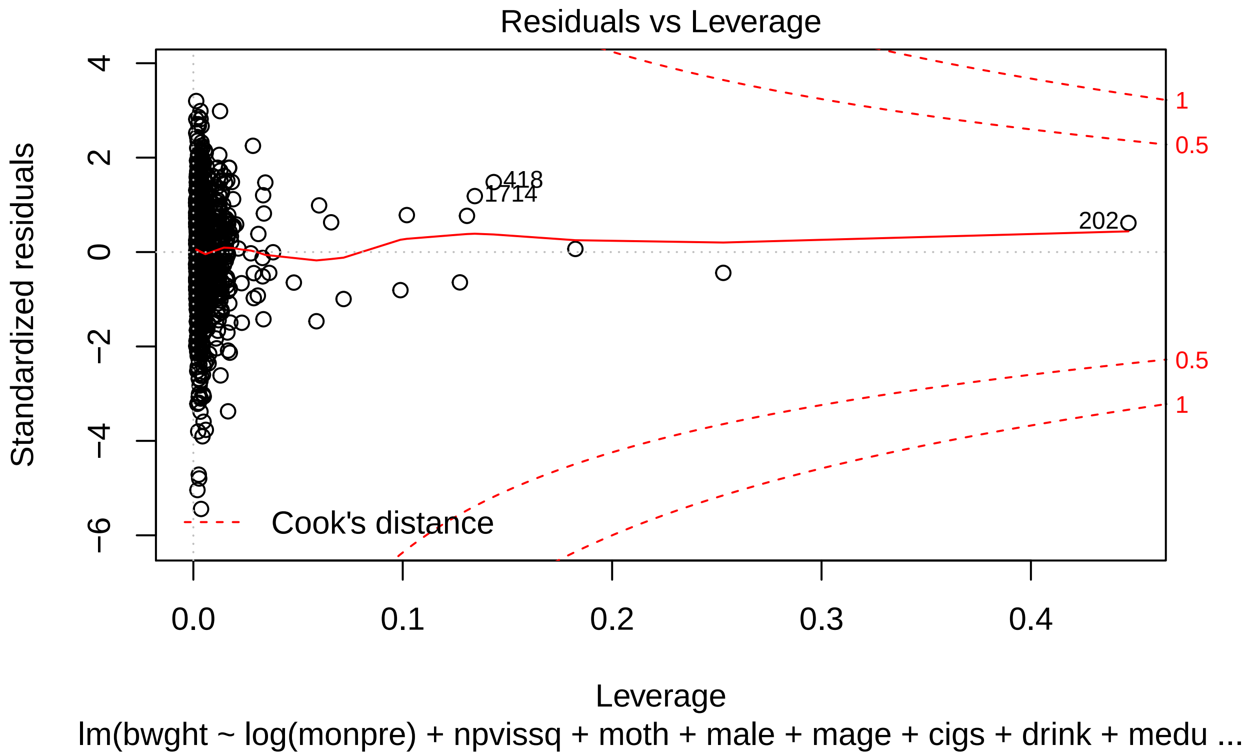
```
## log(monpre)     npvissq        moth        male        mage        cigs
##    1.031069    1.016705    1.029989    1.007702    1.138763    1.058724
##       drink       meduc
##    1.040703    1.185788
```

MLR.4 Zero-conditional mean or MLR.4’ Exogeneity

```
# Residuals vs fitted to assess zero conditional mean
plot(model3, which = 1)
```

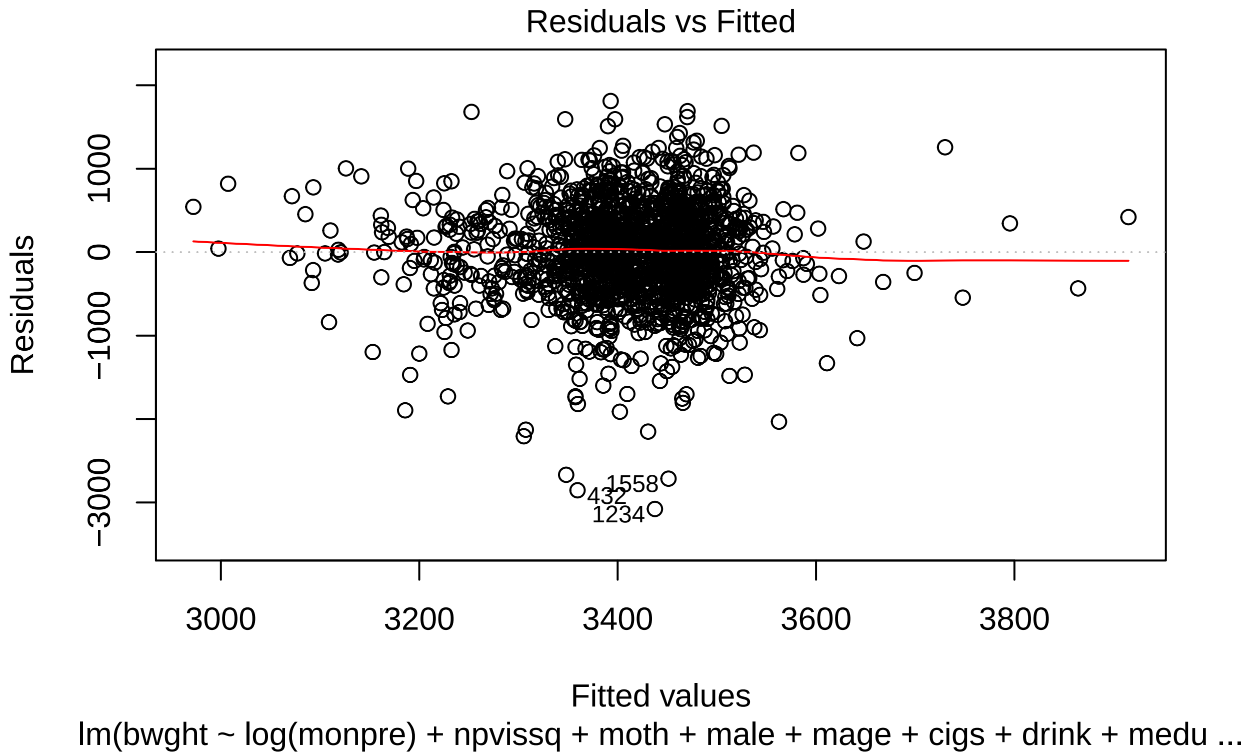


```
plot(model3, which = 5)
```

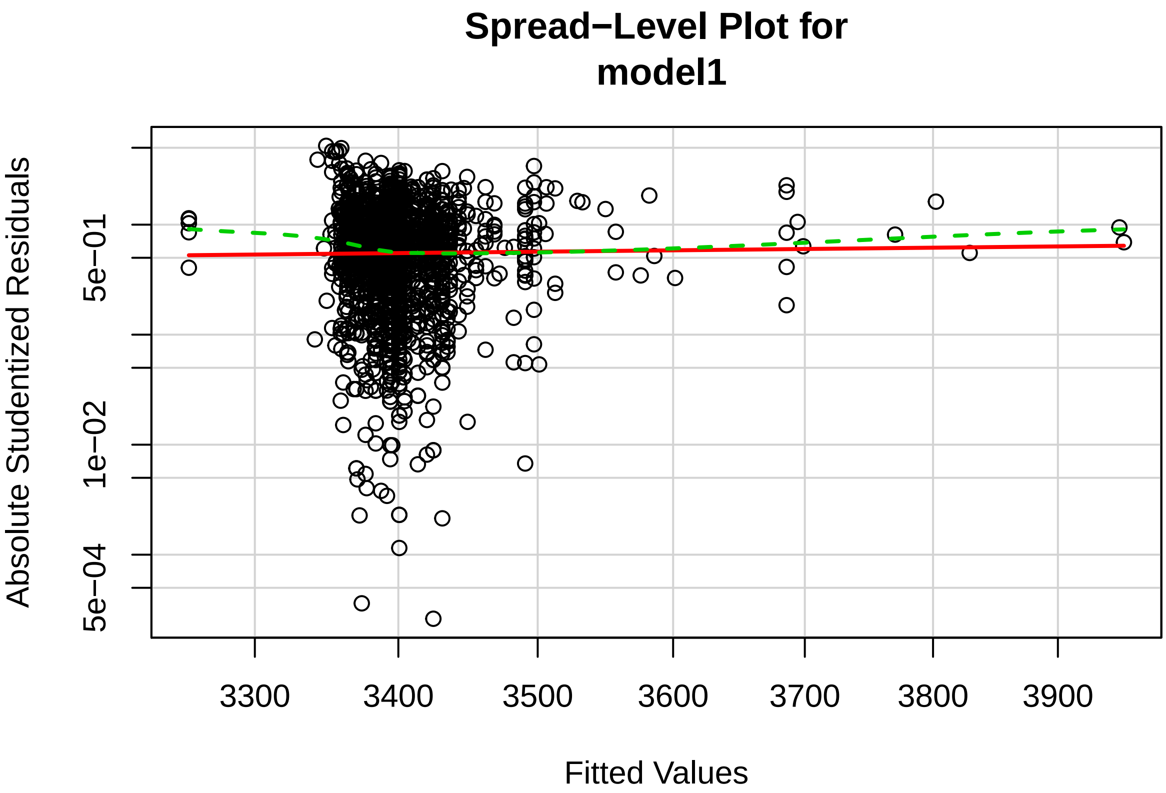


MLR.5 Homoskedasticity: The variance of the error term is constant.

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



```
spreadLevelPlot(model1)
```



```
##
## Suggested power transformation:  -0.01698665
```

```
# Check for the presense of heteroskedasticity with a Breusch-Pagan test.  Null hypothesis is Homoskeda
bptest(model3)
```

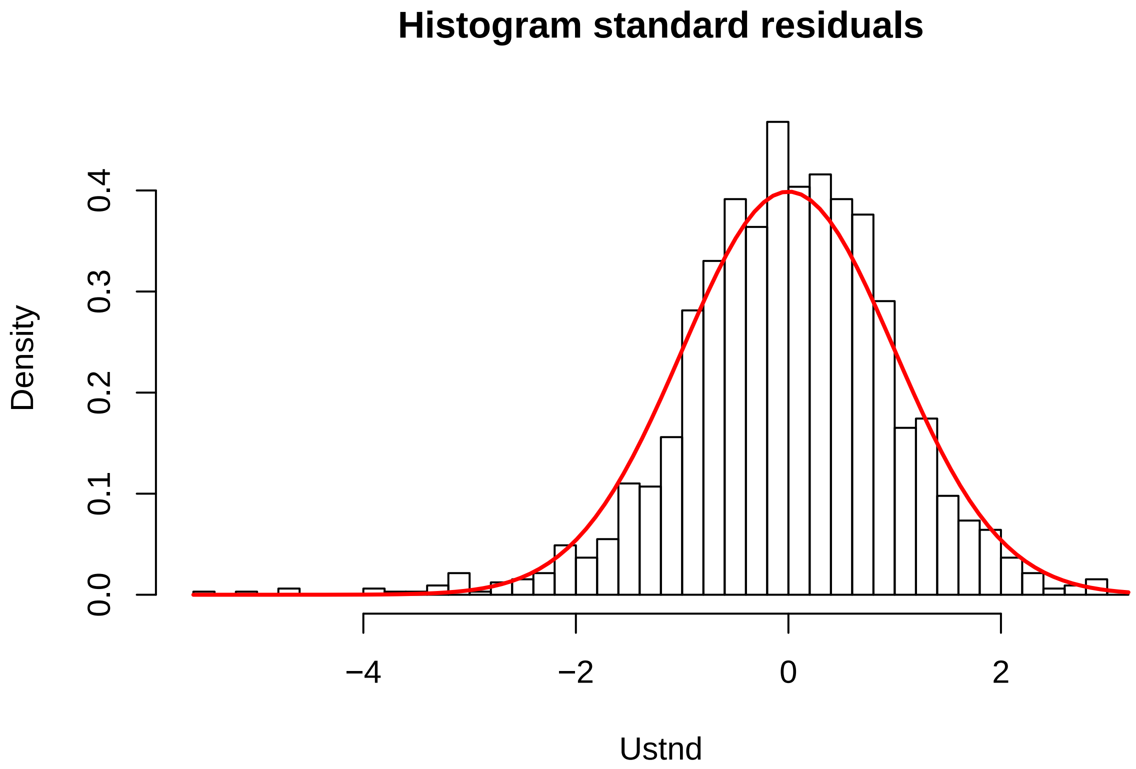
```
##
##  studentized Breusch-Pagan test
##
## data:  model3
## BP = 8.172, df = 8, p-value = 0.4168
```

```
# Non-constant error variance test
ncvTest(model3)
```

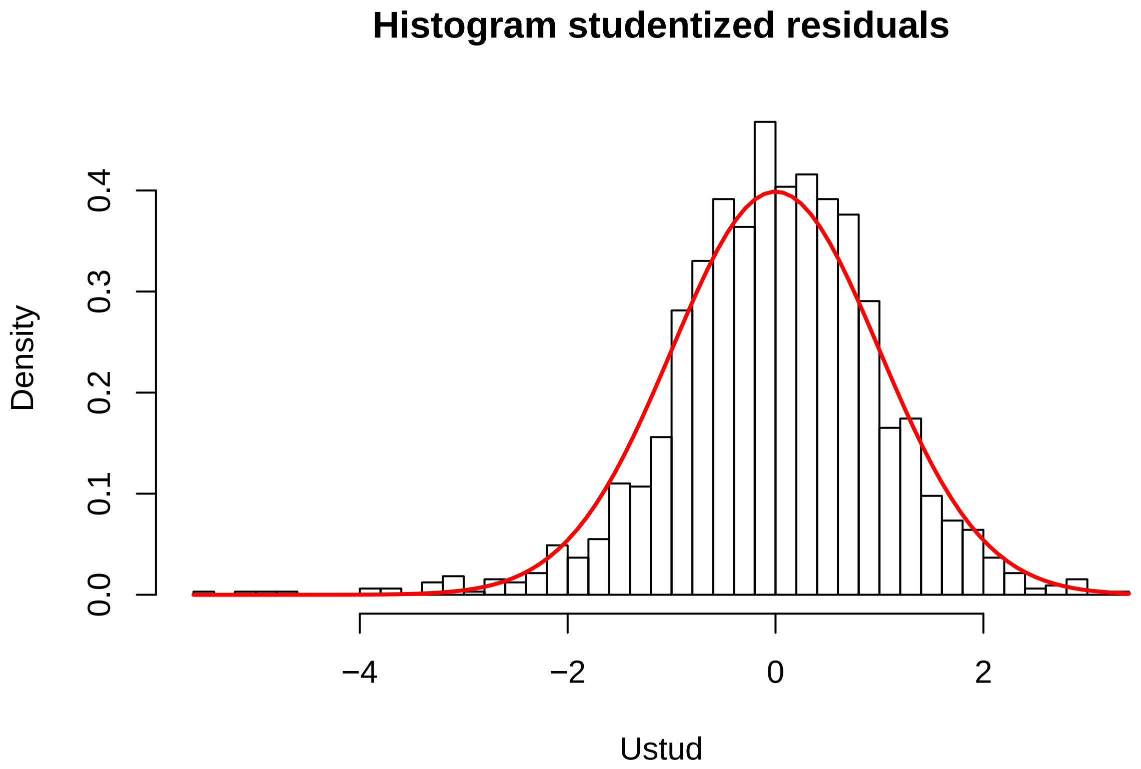
```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.009629436    Df = 1     p = 0.9218293
```

MLR.6 Normality of error terms

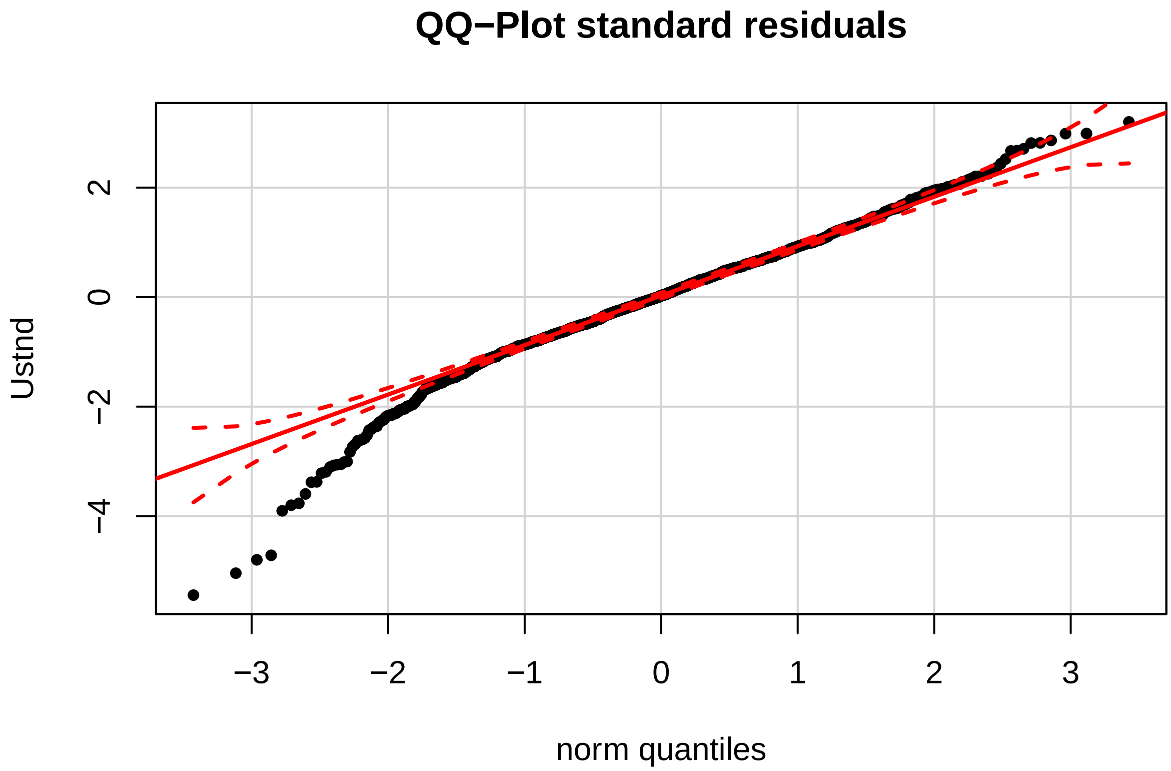
```
# normality of standard residuals
Ustnd = rstandard(model3)
hist(Ustnd, main="Histogram standard residuals", breaks = 50, freq=FALSE)
curve(dnorm(x, mean=0, sd=sd(Ustnd)), col="red", lwd=2, add=TRUE)
```



```
# normality of studentized residuals
Ustud = rstudent(model3)
hist(Ustud, main="Histogram studentized residuals", breaks = 50, freq=FALSE)
curve(dnorm(x, mean=0, sd=1), col="red", lwd=2, add=TRUE)
```



```r
# Q-Q plot standard residuals
qqPlot(Ustnd, distribution="norm", pch=20, main="QQ-Plot standard residuals")
qqline(Ustnd, col="red", lwd=2)
```



Regression table

The results of the models are reported in the table below:

```
(se.m1 = coef(summary(model1))[, "Std. Error"])
```

(Intercept) log(monpre) npvissq 26.2214812 19.1591869 0.1225799

```
(se.m2 = coef(summary(model2))[, "Std. Error"])
```

(Intercept) log(monpre) npvissq moth male 30.2114563 19.1128079 0.1222343 60.3237503 27.5377782

```
(se.m3 = coef(summary(model3))[, "Std. Error"])
```

(Intercept) log(monpre) npvissq moth male mage 117.7013843 20.3754029 0.1221258 61.3960467 28.1458506 3.1510631 cigs drink meduc 3.4937322 48.2836185 7.2661922

```
stargazer(model3, model2, model1, type = "latex", model.numbers = FALSE,
title = "Linear Models Predicting Birthweight",
se = list(se.m3, se.m2, se.m1), omit.stat=c("f","ser"),
          star.cutoffs = c(0.05, 0.01, 0.001))
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu % Date and time: Wed, Apr 26, 2017 - 9:22:11 AM

```
#stargazer(model1, model2, model3, type = "latex",
#title = "Linear Models Predicting Birthweight",
#se = list(se.m1, se.m2, se.m3), omit.stat=c("f","ser"),
```

Table 1: Linear Models Predicting Birthweight

	Dependent variable:		
	bwght		
log(monpre)	9.452 (20.375)	10.899 (19.113)	9.220 (19.159)
npvissq	0.309* (0.122)	0.382** (0.122)	0.377** (0.123)
moth	−165.500** (61.396)	−140.568* (60.324)	
male	82.599** (28.146)	73.627** (27.538)	
mage	4.462 (3.151)		
cigs	−11.365** (3.494)		
drink	−11.731 (48.284)		
meduc	4.016 (7.266)		
Constant	3,151.817*** (117.701)	3,308.163*** (30.211)	3,339.924*** (26.221)
Observations	1,635	1,763	1,763
R^2	0.022	0.012	0.005
Adjusted R^2	0.018	0.010	0.004
Note:		*p<0.05; **p<0.01; ***p<0.001	

```
#          star.cutoffs = c(0.05, 0.01, 0.001))
```

Interpretation

The results describe a relationship between prenatal care and newborn health. The causes of newborn health include a variety of other factors, such as genetic problems impacting the baby, maternal medical conditions such as gestational diabetes or thyroid conditions which can impact the health of the baby, and the weight and diet of the mother during pregnancy. Prenatal care can help to mitigate some of these problems by providing advice to the mother, but ultimately much of the outcome for the newborn depends on the mother.

Certain of the included variables such as maternal age may bias the results by absorbing some of the causal effect of prenatal care.

Conclusion

Our analysis found that it is possible to build a linear model of birthweight using variables representing prenatal care.

However, the model becomes more accurate when other variables are added.

We also had concerns about some data that was not included in the data set and which may have improved the model.

References

1. American Academy of Pediatrics-Committee on Fetus and Newborn (October, 2015). “The Apgar Score”. http://www.acog.org/Resources-And-Publications/Committee-Opinions/Committee-on-Obstetric-Practice/The-Apgar-Score
2. Almond et al. (2005). “The Costs of Low Birth Weight”. https://www.princeton.edu/~davidlee/wp/birthweight.pdf