birthweight

Sarah Cha, Andrew Kabatznick, Kalvin Kao

April 19, 2017

Introduction

The following analyses provide an investigation of the effects of prenatal care on infant health. Funding for the study is provided by an anonymous health advocacy group and the data is taken from the National Center for Health Statistics and from birth certificates. This report presents statistical models which are motivated by widely accepted claims regarding pregnancy and infant health:

Infant health can be measured by birth weight– low birth weights are associated with multiple developmental issues. Birth weight is affected by race, the duration of gestation prior to birth, and prenatal growth rate, and prenatal growth rate in turn is governed by poverty, mother's age, drug use, alcohol, smoking/nicotine, diseases, mother's diet and physical health, mother's prenatal depression, and environmental toxins. Additionally, early and regular prenatal care is known to reduce the chance of infant death and developmental problems.

This set of background information forms the basis for three linear regression models that seek to explain the effects of prenatal care on infant health.

EDA:

Our data set consists of 1612 complete observations and 23 variables that relate to characteristics of the parents (age, education, race), health of the infant (birthweight, APGAR score, and those that potential have some explanatory potential for infant health (number of prenatal hospital visits, month prenatal care began during pregnancy, average cigarettes a day, average drinks per week). Birthweight and APGAR within the data set offer insights on health outcomes for newborn infants. Naturally we care about their distribution and their relationship with other variables in the data set.

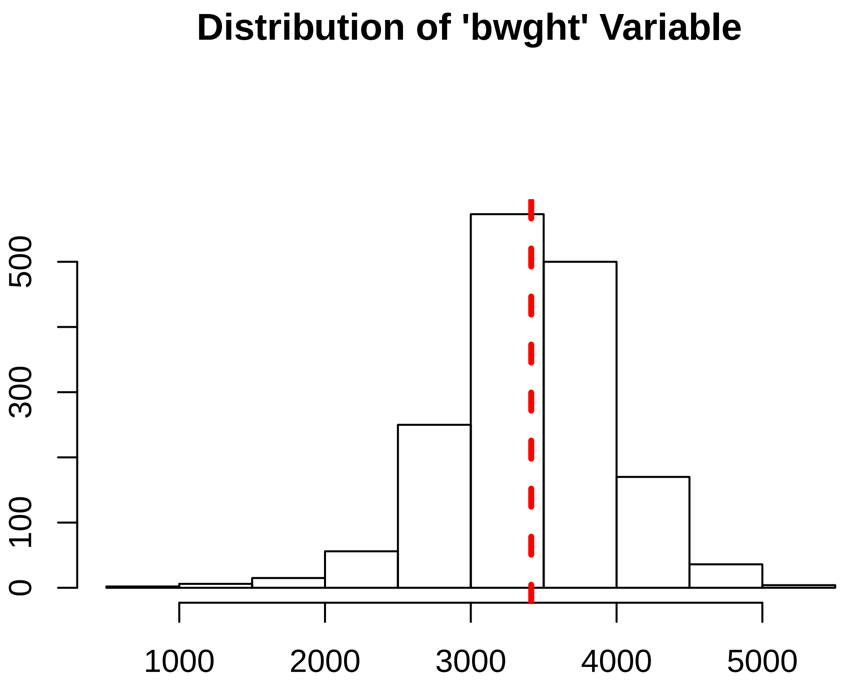
We start off with a glance at all the variables in the data set.

```
summary(sample)
```

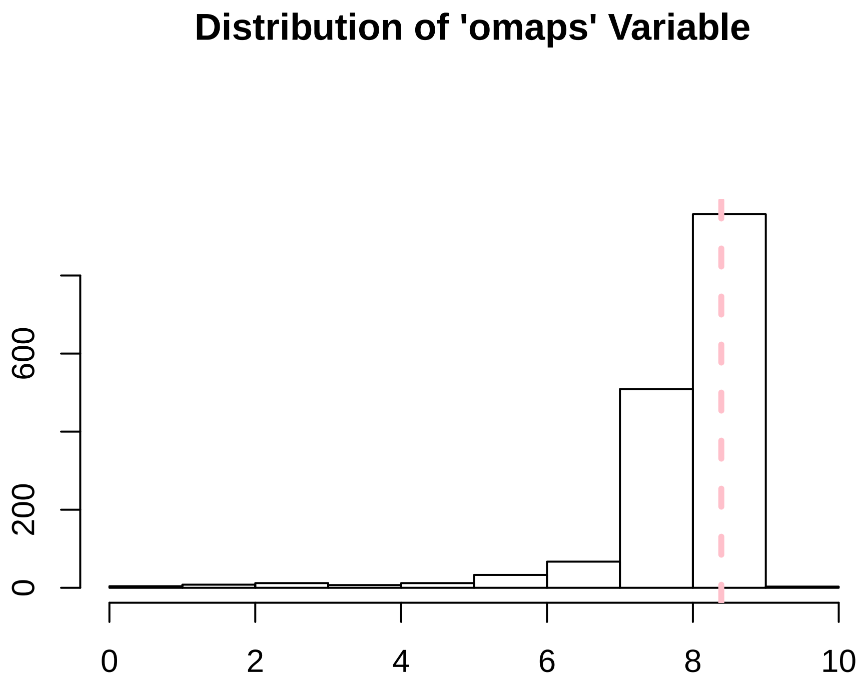
```
##       mage           meduc           monpre          npvis
##  Min.   :16.00   Min.   : 3.00   Min.   :0.000   Min.   : 0.00
##  1st Qu.:26.00   1st Qu.:12.00   1st Qu.:1.000   1st Qu.:10.00
##  Median :29.00   Median :14.00   Median :2.000   Median :12.00
##  Mean   :29.48   Mean   :13.74   Mean   :2.143   Mean   :11.62
##  3rd Qu.:32.00   3rd Qu.:16.00   3rd Qu.:2.000   3rd Qu.:13.00
##  Max.   :44.00   Max.   :17.00   Max.   :9.000   Max.   :40.00
##       fage           feduc           bwght          omaps
##  Min.   :18.00   Min.   : 3.00   Min.   : 506   Min.   : 0.00
##  1st Qu.:28.00   1st Qu.:12.00   1st Qu.:3090   1st Qu.: 8.00
##  Median :31.00   Median :14.00   Median :3430   Median : 9.00
##  Mean   :31.79   Mean   :13.91   Mean   :3415   Mean   : 8.39
##  3rd Qu.:35.00   3rd Qu.:16.00   3rd Qu.:3771   3rd Qu.: 9.00
##  Max.   :62.00   Max.   :17.00   Max.   :5204   Max.   :10.00
##      fmaps            cigs            drink          lbw      vlbw
##  Min.   : 2.000   Min.   : 0.000   Min.   :0.00000   0:1589   0:1604
##  1st Qu.: 9.000   1st Qu.: 0.000   1st Qu.:0.00000   1:  23   1:   8
##  Median : 9.000   Median : 0.000   Median :0.00000
##  Mean   : 9.015   Mean   : 1.057   Mean   :0.02109
##  3rd Qu.: 9.000   3rd Qu.: 0.000   3rd Qu.:0.00000
##  Max.   :10.000   Max.   :40.000   Max.   :8.00000
##  male     mwhte     mblck     moth      fwhte     fblck     foth
##  0:784    0: 181    0:1524    0:1519    0: 171    0:1520    0:1533
##  1:828    1:1431    1:  88    1:  93    1:1441    1:  92    1:  79
##
##
##
##
##      lbwght          magesq          npvissq
##  Min.   :6.227   Min.   : 256.0   Min.   :   0.0
##  1st Qu.:8.036   1st Qu.: 676.0   1st Qu.: 100.0
##  Median :8.140   Median : 841.0   Median : 144.0
##  Mean   :8.120   Mean   : 891.3   Mean   : 148.9
##  3rd Qu.:8.235   3rd Qu.:1024.0   3rd Qu.: 169.0
##  Max.   :8.557   Max.   :1936.0   Max.   :1600.0
```

Our first plots show that birthweight is approximately normally distributed with a mean of 3415 grams. ‘omaps’ and ‘fmaps,’ the one minute and five minute APGAR scores respectively both have distributions with negative skew and medians of 9 though ‘fmaps’ has a slightly more pronounced skew suggesting that more 5-min APGAR scores are bunched up toward the high end of the scale (~9).

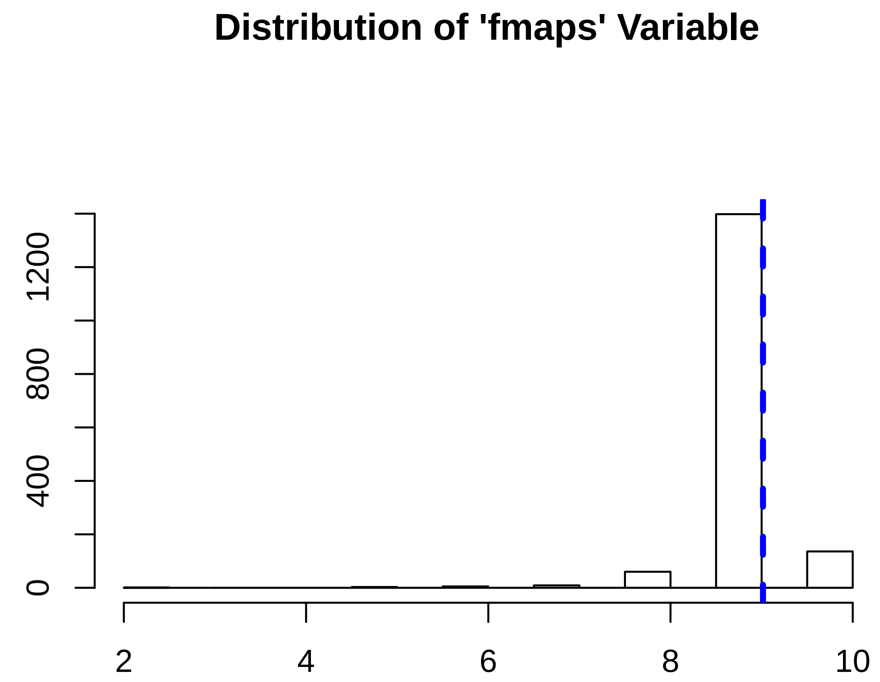
```
#birthweight-- almost normal with a negative skew
par(mar=c(3,3,9,9),cex.axis=1,cex.lab=1)
hist(sample$bwght, main = "Distribution of 'bwght' Variable",
     xlab = "Infant Birthweight (grams)")
abline(v = mean(sample$bwght, na.rm= TRUE), col="red", lwd=3, lty=2)
```



```
#one minute apgar score-- exponential, extreme negative skew with spikes at 8 and 9
hist(sample$omaps, main = "Distribution of 'omaps' Variable",
     xlab = "One Minute Apgar Score")
abline(v = mean(sample$omaps, na.rm= TRUE), col="pink", lwd=3, lty=2)
```



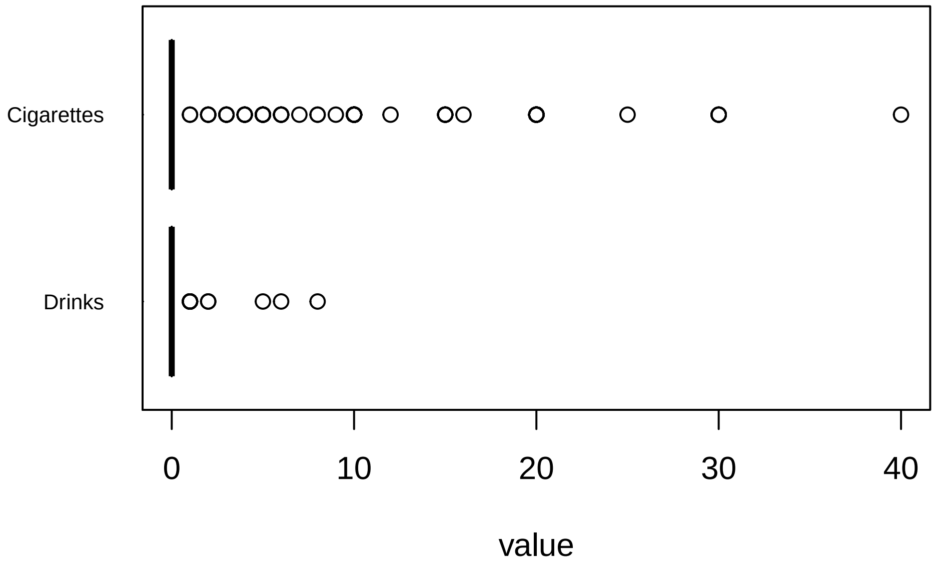
```
#five minute apgar score-- extreme negative skew, mainly values of 9
hist(sample$fmaps, main = "Distribution of 'fmaps' Variable",
     xlab = "Five Minute Apgar Score")
abline(v = mean(sample$fmaps, na.rm= TRUE), col="blue", lwd=3, lty=2)
```



```
#Plot 1
cigs_drinks = data.frame(cbind(sample$drink, sample$cigs))
colnames(cigs_drinks) = c("Drinks", "Cigarettes")
long <- melt(cigs_drinks)
```

```
## No id variables; using all as measure variables
```

```
par(mar=c(5,5,7,7),cex.axis=1,cex.lab=1)
plot(value ~ variable, data=long, horizontal = TRUE, xlab ="", yaxt ="n")
axis(2, at = c(1, 2), labels = colnames(cigs_drinks), tcl = 0, las = 2, cex.axis = .7)
```

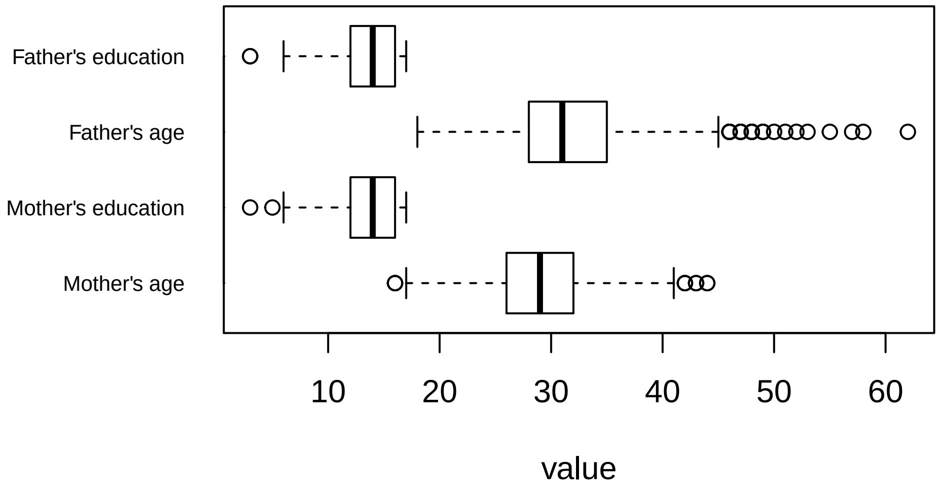


Next we look at our parents in the data set. Average mother and father ages are 29.5 and 31.8 years respectively. We notice the presence of several outlier variables (outside the upper whisker) for fathers' age. The quartile averages for parent education years appear similiar. Average education years for both mothers and fathers in this sample are just under 14 years.

```
#Plot 2
parent_char = data.frame(cbind(sample$mage, sample$meduc, sample$fage, sample$feduc))
colnames(parent_char)= c("Mother's age", "Mother's education", "Father's age", "Father's education")
long <- melt(parent_char)
```

```
## No id variables; using all as measure variables
```

```
par(mar=c(7,7,7,7),cex.axis=1,cex.lab=1)
plot(value ~ variable, data=long, horizontal = TRUE, xlab ="", yaxt ="n")
axis(2, at = c(1, 2, 3, 4), labels = colnames(parent_char), tcl = 0, las = 2, cex.axis = .7)
```



Next we analyzed some of the health variables in the data set including number of prenatal visits, month prenatal care began, and the one and five minute APGAR scores. The average mother in this sample began prenatal care a little more than 2 months into their pregnancy (~2.14 months) while number of prenatal visits averaged 11.62. We can see that the distribution of pre-natal care visits is wide spanning anywhere from 0 and 40 while 90% of the values are between 5 and 15 visits. Start month of prenatal care appears to have positive skew with 90% of the mothers beginning care at 3 months or earlier.

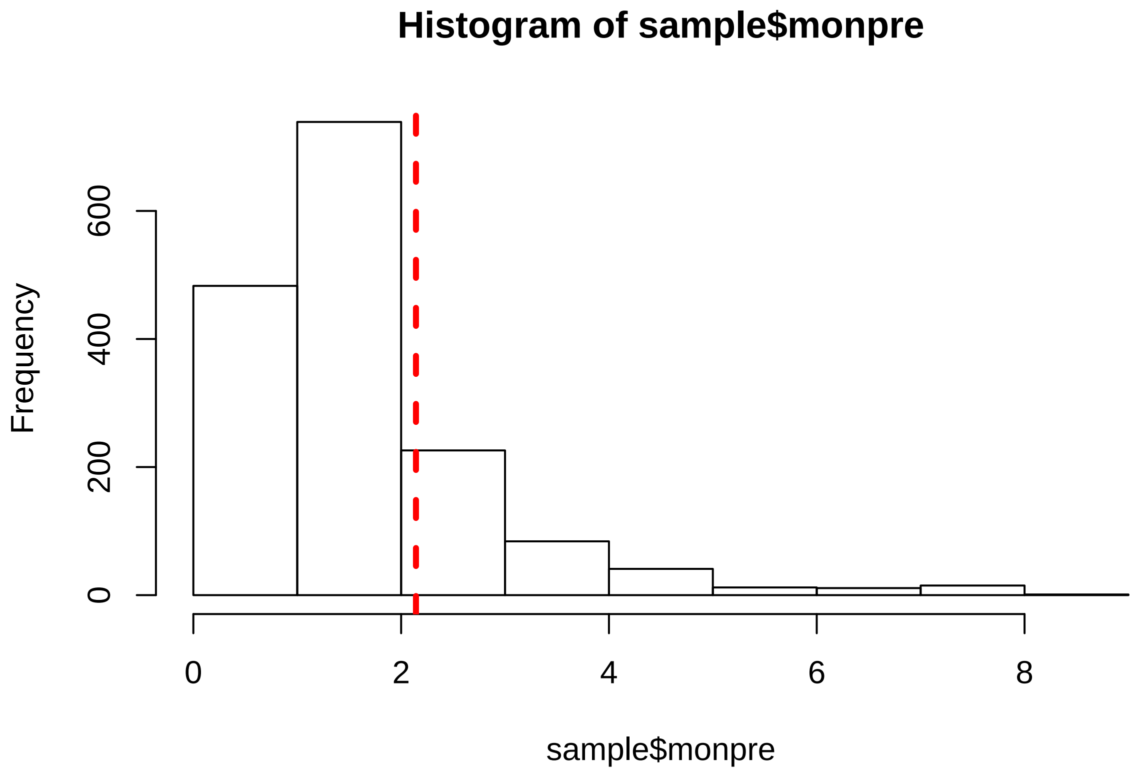
```
nrow(sample[(sample$npvis <= 15) & (sample$npvis >= 5), ])
```

```
## [1] 1460
```

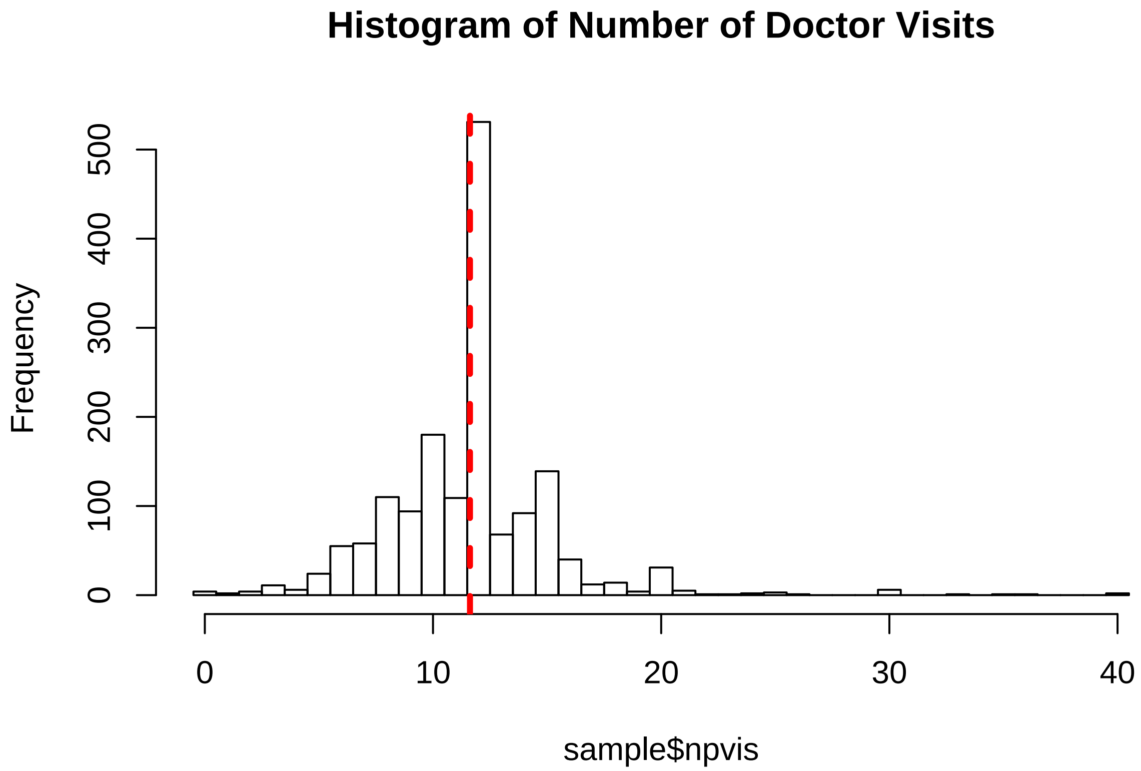
```
nrow(sample[(sample$monpre <= 3), ])
```

```
## [1] 1448
```

```
hist(sample$monpre)
abline(v = mean(sample$monpre, na.rm= TRUE), col="red", lwd=3, lty=2)
```



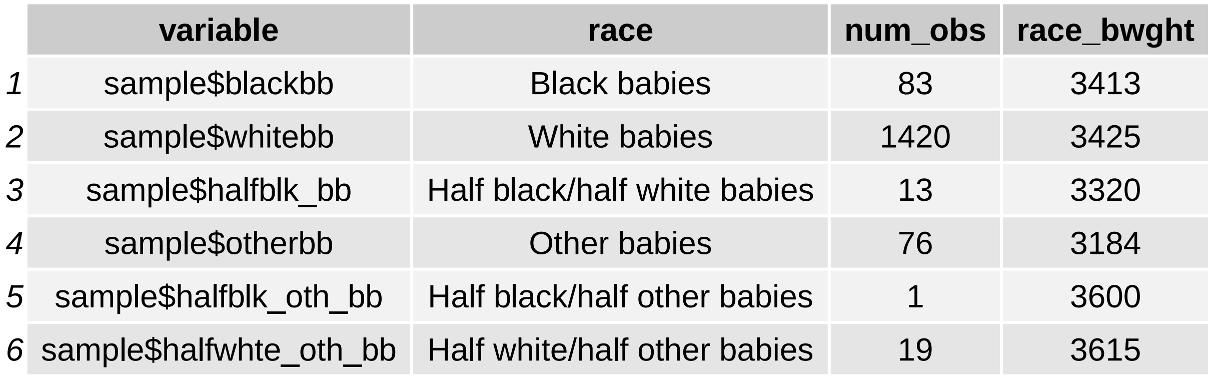
```
hist(sample$npvis, breaks = 0:41 -.5, main = "Histogram of Number of Doctor Visits")
abline(v = mean(sample$npvis, na.rm= TRUE), col="red", lwd=3, lty=2)
```



```
#Plot 3
# health_var = data.frame(cbind(sample$omaps, sample$fmaps, sample$monpre, sample$npvis))
# colnames(health_var)= c("One min APGAR score", "Five min APGAR score", "Month prenatal care began", "
# long <- melt(health_var)
# par(mar=c(10,7,6,6),cex.axis=1,cex.lab=1)
# plot(value ~ variable, data=long, horizontal = TRUE, xlab ="", yaxt ="n")
# axis(2, at = c(1, 2, 3, 4), labels = colnames(health_var), tcl = 0, las = 2, cex.axis = .7)
```

Almost 90% of the babies in the sample were white babies (n = 1420) while 5% were black (n = 83), and a little less than 5% other (n = 76). With the skew of the data in mind, race does seem to have some effect on baby birthweight at first glance of the data. In particular average birthweight gaps are the largest between “other” babies and “half white/half other” babies though admittedly the sample size of “half white/half other” babies is much smaller (n = 19). Further “other” babies appear to have the smallest birthweights of all the groupings.

```
baby_races = data.frame(cbind(variable, race, num_obs, race_bwght))
grid.table(baby_races)
```

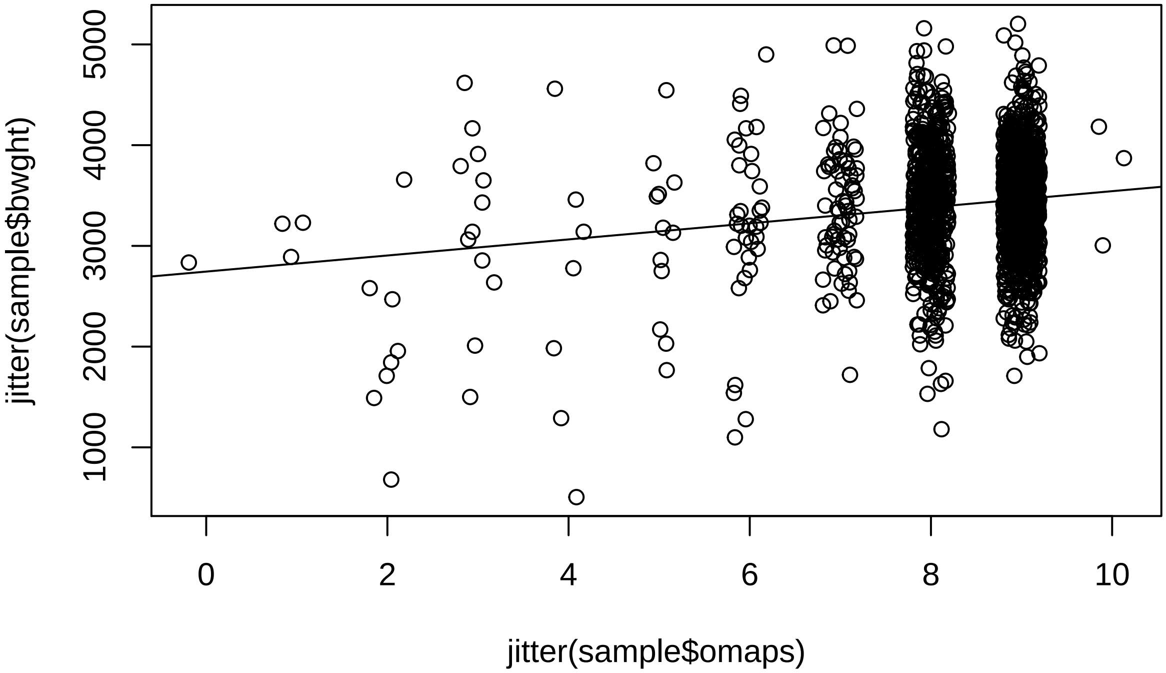


	variable	race	num_obs	race_bwght
1	sample$blackbb	Black babies	83	3413
2	sample$whitebb	White babies	1420	3425
3	sample$halfblk_bb	Half black/half white babies	13	3320
4	sample$otherbb	Other babies	76	3184
5	sample$halfblk_oth_bb	Half black/half other babies	1	3600
6	sample$halfwhte_oth_bb	Half white/half other babies	19	3615

Next, we looked at relationships between key variables in the data set, particularly relationship with variables in the data set and the potential outcome variables, birthweight and APGAR scores.

Birthweight and APGAR scores do show positive correlation in the data set:

```r
#relationship between bwght and omaps
z = plot(jitter(sample$omaps), jitter(sample$bwght))
abline(lsfit(sample$omaps, sample$bwght))
```

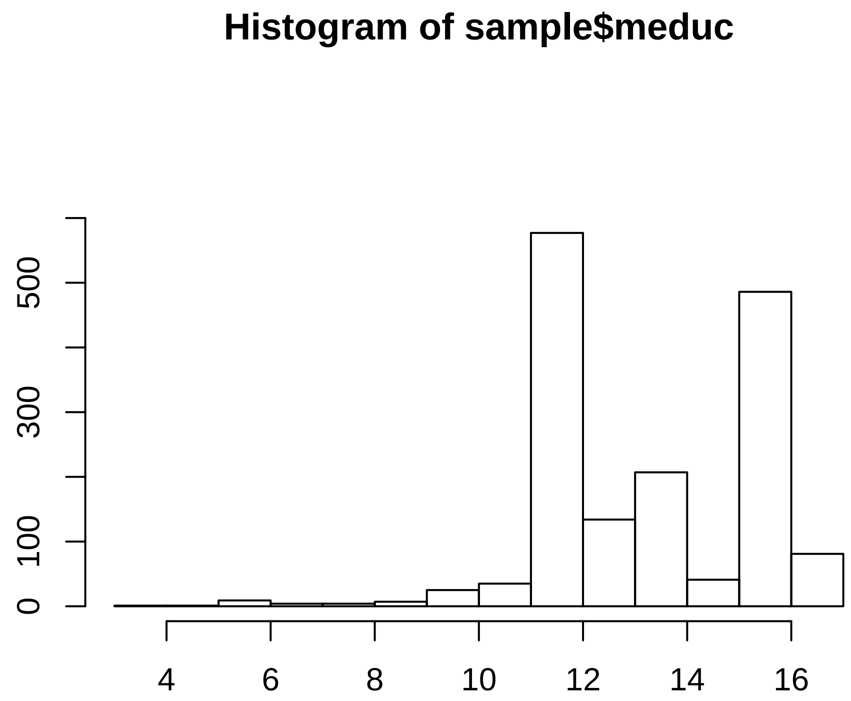


```r
cor(sample$omaps, sample$bwght)
```

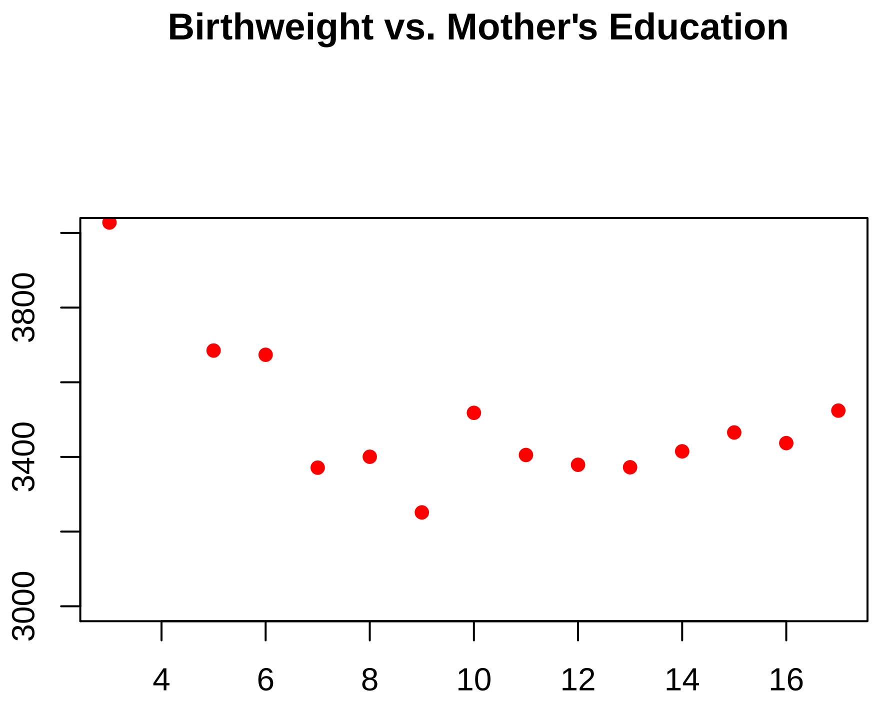
```
## [1] 0.1558288
```

At initial glance mother's education appears negative correlated with birthweight but we notice that this is being skewed by very few data points for mothers with years of education less than 9 years. We do notice however that birthweight seems to have diminishing, concave exponential relationship with mother's age.

```r
#education vs avg bwght
par(mar=c(2,2,10,10),cex.axis=1,cex.lab=1)
z = hist(sample$meduc)
```



```
b = data.frame(cbind(z$breaks))
colnames(b) = "education_years"

sorting <- sapply(split(sample,sample$meduc), function(x) {
  colMeans(x["bwght"],na.rm=TRUE)
})
bwght_by_meduc = data.frame(cbind(sorting))
bwght_by_meduc = cbind(bwght_by_meduc, c(3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17))
colnames(bwght_by_meduc)= c("average_birthweight", "education_years")
a = merge(b, bwght_by_meduc, by.x = 1, by.y = 2, all.x = TRUE)
a[is.na(a)] <-0

par(mar=c(2,2,10,10),cex.axis=1,cex.lab=1)
plot(a$education_years, a$average_birthweight, ylim = seq(3000,4100,1000), xlab = "Years of Mothers' Edu
     ylab = "Average Birthweight", main =" Birthweight vs. Mother's Education", col = "red", pch = 16)
```

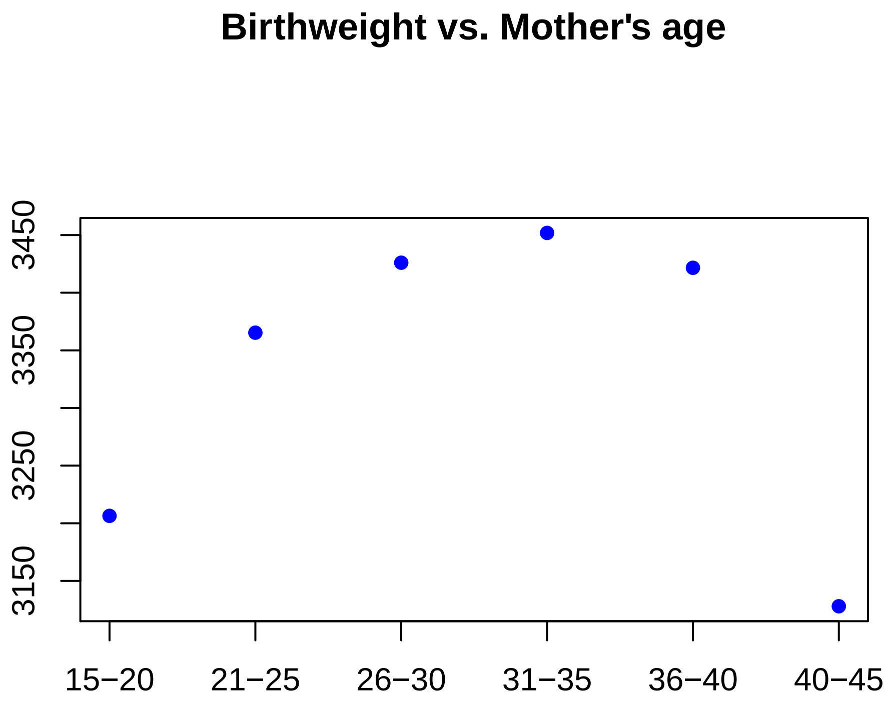


Birthweight vs. Mother's Education

```
#mother's age vs bwght - looks like there is a concave exponential relationship
mage_exp = data.frame(cbind(sample$mage, sample$bwght, sample$omaps, sample$fmaps))
colnames(mage_exp) = c("Mother_age", "birthweight", "OMAPS", "FMAPS")
summary(sample$mage)
```

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

```
mage_exp$agebin <- cut(mage_exp$Mother_age, breaks = seq(15, 45, by = 5),
                       labels = c("15-20","21-25","26-30","31-35","36-40","41-45"))

plot(by(mage_exp$birthweight,mage_exp$agebin, mean), main = "Birthweight vs. Mother's age", xaxt = "n",
axis(1, at =seq(1,6,1), labels = c("15-20","21-25","26-30","31-35","36-40","40-45"),xlab = "Mother's ag
```



Birthweight vs. pre-natal care:

We compared birthweight other pre-natal care factors such as number of visits and month prenatal care began. It wasn't clear from first glance at the data that there was a notable trend.

We wanted to first understand if birthweight was correlated with whether the mothers received prenatal care or not? There is some difference in mean between the two groups - namely that babies that received no prenatal care had higher birthweights vs. those who did.

```
mean(sample$bwght[sample$npvis > 0])
```

```
## [1] 3414.047
```

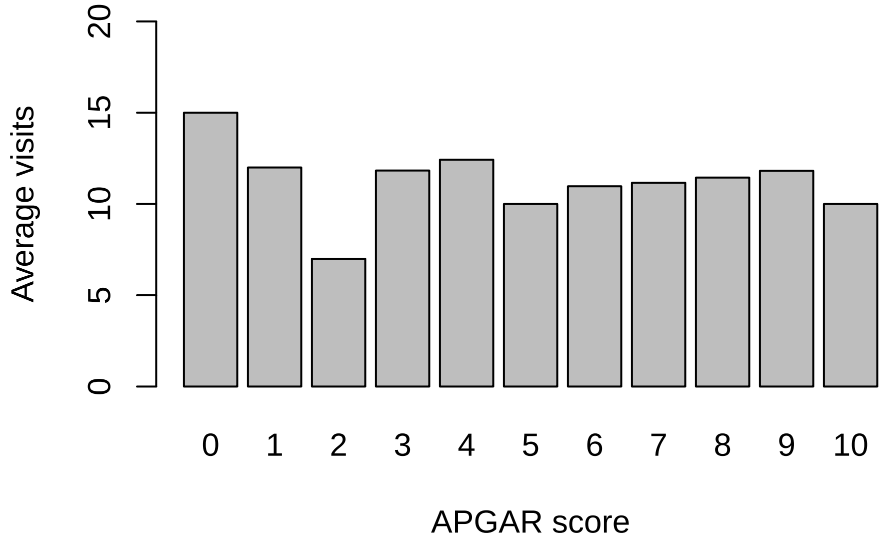
```
mean(sample$bwght[sample$npvis == 0])
```

```
## [1] 3610.5
```

However this data set actually has very few moms who received no prenatal care (n = 4) making this metric a less valuable one.

```
# looking at omaps vs # of visits - no strong trend
avg_visits<- sapply(split(sample,sample$omaps), function(x) {
  colMeans(x["npvis"],na.rm=TRUE)
})
```

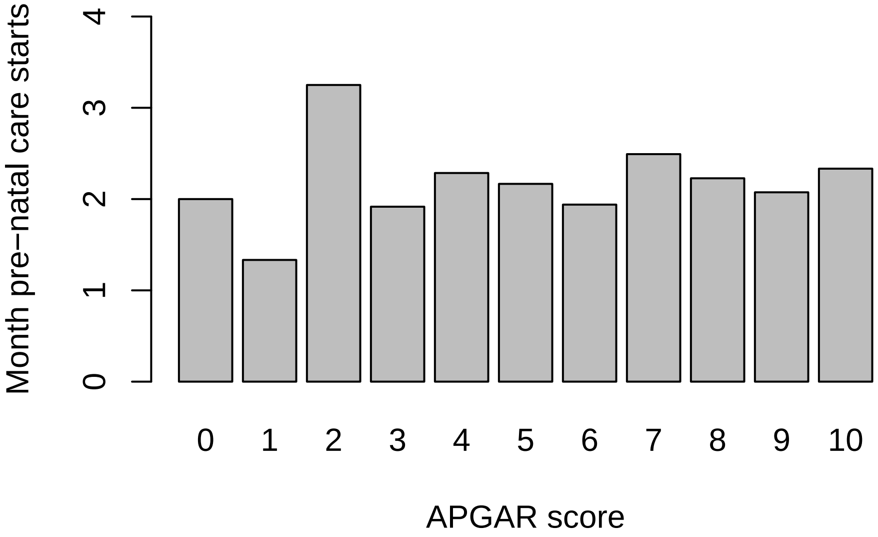
```
par(mar=c(8,8,5,5),cex.axis=1,cex.lab=1)
barplot(avg_visits, names.arg= c(0,1,2,3,4,5,6,7,8,9,10), ylim = c(0, 20), xlab = "APGAR score", ylab =
```



```
cor(sample$npvis, sample$omaps, use = "complete.obs")
```

```
## [1] 0.07020163
```

```
# looking at omaps vs # monpre- no strong trend
avg_mon<- sapply(split(sample,sample$omaps), function(x) {
  colMeans(x["monpre"],na.rm=TRUE)
})
barplot(avg_mon, names.arg= c(0,1,2,3,4,5,6,7,8,9,10), ylim = c(0, 4), xlab = "APGAR score", ylab = "Mo
```

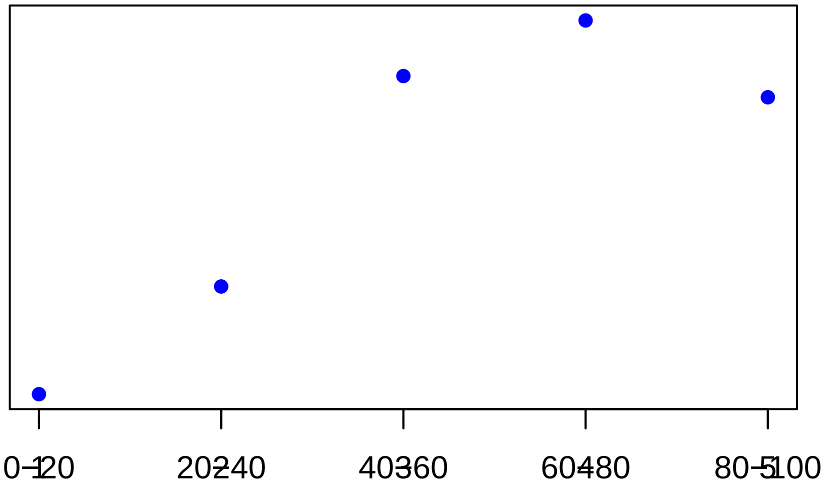


As we looked deepper into both variables, both variables had outlier values (with very few observations) and seemed to be skewing the summary statistics. For example, the median number of visits by the baby mother was 12 but there was an observation where a mother had 40 visits. So i attempted cut up a couple of different ways in an effort to minimize the skew. Namely I looked at visits per week and visits per month. Binning visits per month into quintiles showed signs that lower quintile visits (ie less frequent visits), could be correlated with lower birthweights and we know that this representation is less skewed by outliers. Further we see signs that there is an concave, exponential relationship between birthweight and monthly visits. Notably at higher visits, there is a diminishing relationship with infant birthweight.

```
sample$visits_pr_mo = sample$npvis/(9 - (sample$monpre))
sample$visits_pr_mo[sample$visits_pr_mo == Inf ] = 0
summary(sample$visits_pr_mo)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.429   1.714   1.769   1.875  33.000
```

```
sample$visitsbin <- cut(sample$visits_pr_mo, breaks=c(quantile(sample$visits_pr_mo, probs = seq(0, 1, b
                              labels=c("0-20","20-40","40-60","60-80", "80-100"), include.lowest=TRUE)
par(mar=c(2,2,10,10))
plot(by(sample$bwght, sample$visitsbin, mean), xaxt ="n", yaxt ="n", xlab ="", ylab ="", pch = 16, col
axis(1,axis(1, at =seq(1,5,1), labels = c("0-20","20-40","40-60","60-80","80-100"),xlab = "Percentile o
```



```
#mother's age-- nice almost normal distribution
hist(sample$mage, main = "Distribution of 'mage' Variable",
     xlab = "Mother's Age (years)")

#mother's education-- spikes at 12 and 16 years (HS and college), with some extreme low values
#30 NAs
hist(sample$meduc, main = "Distribution of 'meduc' Variable",
     xlab = "Mother's Education (years)")

#month prenatal care began-- nearly exponential distribution with spikes at 1 and 2 months
#assuming this is the month of pregnancy.  given that there are fewer values above 3 months, we may wan
#5 NAs
hist(sample$monpre, main = "Distribution of 'monpre' Variable",
     xlab = "Month Prenatal Care Began")

#number of prenatal visits-- normal with a spike at 12 visits and a long right tail
#68 NAs
hist(sample$npvis, main = "Distribution of 'npvis' Variable",
     xlab = "Total Number of Prenatal Visits")

#father's age-- almost normal with a positive skew
#6 NAs
hist(sample$fage, main = "Distribution of 'fage' Variable",
     xlab = "Father's Age (years)")

#father's education-- very similar to meduc (spikes at 12 and 16 years, with some extreme low values)
```

```r
hist(sample$feduc, main = "Distribution of 'feduc' Variable",
     xlab = "Father's Education (years)")

#cigs-- extreme positive skew, with mainly values of 0
hist(sample$cigs, main = "Distribution of 'cigs' Variable",
     xlab = "Average Cigarettes Per Day")

#drinks-- pretty much all values of 0
hist(sample$drink, main = "Distribution of 'drinks' Variable",
     xlab = "Average Drinks Per Week")

#lbw =1 if bwght <= 2000
sum(sample$lbw == 0)#1802
sum(sample$lbw == 1)#30

#vlbw =1 if bwght <= 1500
sum(sample$vlbw == 0)#1819
sum(sample$vlbw == 1)#13

#male =1 if baby male
sum(sample$male == 0)#891
sum(sample$male == 1)#941

#mwhte =1 if mother white
sum(sample$mwhte == 0)#208
sum(sample$mwhte == 1)#1624

#mblck =1 if mother black
sum(sample$mblck == 0)#1723
sum(sample$mblck == 1)#109

#moth =1 if mother is other
sum(sample$moth == 0)#1733
sum(sample$moth == 1)#99

#fwhte =1 if father white
sum(sample$fwhte == 0)#202
sum(sample$fwhte == 1)#1630

#fblck =1 if father black
sum(sample$fblck == 0)#1725
sum(sample$fblck == 1)#107

#foth =1 if father is other
sum(sample$foth == 0)#1737
sum(sample$foth == 1)#95

#lbwght log(bwght)-- similar distribution as level bwght... why would we use a log transform
hist(sample$lbwght, main = "Distribution of 'lbwght' Variable",
     xlab = "log(bwght)")

#magesq mage~2-- again, an almost normal distribution, but mage was already almost normal so why would
hist(sample$magesq, main = "Distribution of 'magesq' Variable",
     xlab = "mage^2")

#npvissq npvis^2-- exponential with positive skew... but npvis was already almost normal, so why would
hist(sample$npvissq, main = "Distribution of 'npvissq' Variable",
     xlab = "npvis^2")
```

Model 1

3. A minimum of three model specifications. In particular, you should include

- One model with only the explanatory variables of key interest.

The sample contains multiple variables related to infant health, including birth weight (‘bwght’), one-minute APGAR score (omaps), five-minute APGAR score (fmaps), low birth weight (‘lbw’), and very-low birth weight (‘vlbw’). The purpose of the APGAR score is to determine if a newborn requires immediate medical attention, and background knowledge indicates that infant birth weight is highly indicative of future infant health, so the ‘bwght’ variable has thus been selected to operationalize the concept of ‘infant health’ in this preliminary model. The ‘lbw’ and ‘vlbw’ variables are indicators that focus only on a small subset of infants, and a model that uses ‘lbw’ or ‘vlbw’ as its dependent variable requires an advanced form of analysis that will not be used in this study.

The given premises regarding infant health identify multiple other variables in the sample that have strong explanatory potential for infant health, namely number of prenatal visits (‘npvis’), the month of pregnancy prenatal care began (‘monpre’), mother’s age (‘mage’), drinks per week (‘drink’), and cigarettes per day (‘cigs’). The sample additionally contains multiple indicator variables representing the race of the parents, including ‘mwhte’, ‘mblck’, ‘moth’, ‘fwhte’, ‘fblck’, and ‘foth’. These variables may also increase the explanatory ability of the model, given that a baby’s race is related to its birth weight.

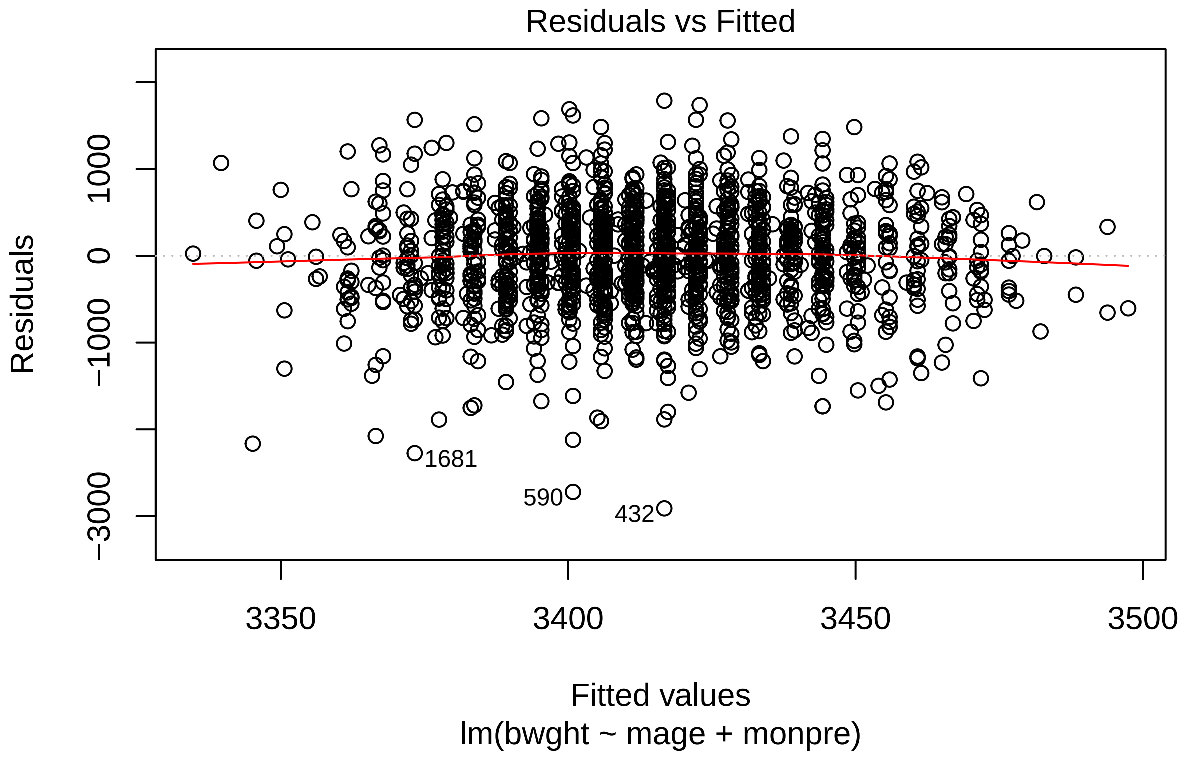
This study begins its analysis by characterizing a simple foundational model upon which deeper analysis can then be performed: $bwght = \beta_0 + \beta_1 mage + \beta_2 monpre + u$

```
# #maybe save the race stuff for later
# blackBaby = sample$mblck*sample$fblck
# whiteBaby = sample$mwhte*sample$fwhte
# bBabyWeight = sample$bwght[blackBaby == 1]
# wBabyWeight = sample$bwght[whiteBaby == 1]
# mean(bBabyWeight)
# mean(wBabyWeight)
# #hist(bBabyWeight)
# #hist(wBabyWeight)

# #data cleaning in progress
# cleanData_1 = sample[(!is.na(sample$monpre)) & (!is.na(sample$cigs)) & (!is.na(sample$drink)),]
# #the following are indicators we might want to use instead
# latePre = factor(ifelse(cleanData_1$monpre>6, 1, 0))
# earlyPre = factor(ifelse(cleanData_1$monpre<3, 1, 0))
# #prenatal1 = factor(ifelse(cleanData_1$monpre==1, 1, 0))
# prenatal2 = factor(ifelse(cleanData_1$monpre==2, 1, 0))
# prenatal3 = factor(ifelse(cleanData_1$monpre==3, 1, 0))
# prenatal4 = factor(ifelse(cleanData_1$monpre==4, 1, 0))
# prenatal5 = factor(ifelse(cleanData_1$monpre==5, 1, 0))
# prenatal6 = factor(ifelse(cleanData_1$monpre==6, 1, 0))
# prenatal7 = factor(ifelse(cleanData_1$monpre==7, 1, 0))
# prenatal8 = factor(ifelse(cleanData_1$monpre==8, 1, 0))
```

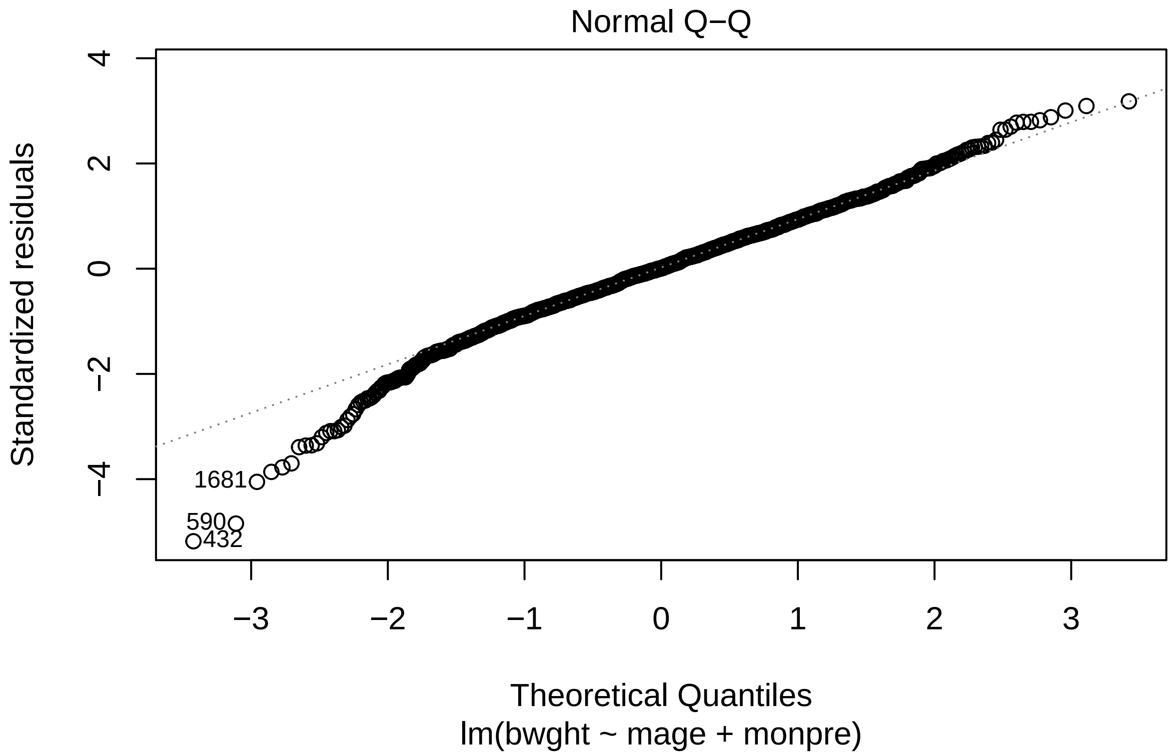
```r
# prenatal9 = factor(ifelse(cleanData_1$monpre==9, 1, 0))
# #prenatalTri1 = factor(ifelse(cleanData_1$monpre < 4, 1, 0))
# prenatalTri2 = factor(ifelse((cleanData_1$monpre > 3) & (cleanData_1$monpre < 7), 1, 0))
# prenatalTri3 = factor(ifelse(cleanData_1$monpre > 6, 1, 0))
# yesCigs = factor(ifelse(cleanData_1$cigs>0, 1, 0))
# yesDrinks = factor(ifelse(cleanData_1$drink>0, 1, 0))
#
# visitFreq = cleanData_1$npvis/(10-cleanData_1$monpre)

#model_1 = lm(bwght ~ mage + magesq + cigs + drink + visitFreq, data = cleanData_1)
#model_1_alt = lm(bwght ~ mage + cigs + monpre, data=sample)
#model_1 = lm(bwght ~ mage + cigs + npvis, data=cleanData_1)
#model_1 = lm(bwght ~ cigs + monpre, data = cleanData_1)
#model_1 = lm(bwght ~ drink + npvis, data = cleanData_1)
#model_1 = lm(bwght ~ cigs + npvis, data = cleanData_1)
#model_1_alt2 = lm(bwght ~ mage, data=cleanData_1)
#model_1 = lm(bwght ~ mage + npvis, data = cleanData_1)
#model_1_alt3 = lm(bwght ~ mage + cigs, data=cleanData_1)
#model_1 = lm(bwght ~ cigs, data=cleanData_1)
#model_1 = lm(bwght ~ cigs + monpre, data=cleanData_1)
#model_1 = lm(bwght ~ monpre, data=cleanData_1)#heteroskedasticity
#model_1 = lm(bwght ~ npvis, data=cleanData_1)#heteroskedasticity
#model_1 = lm

model_1 = lm(bwght ~ mage + monpre, data=sample)
summary(model_1)
```

```
##
## Call:
## lm(formula = bwght ~ mage + monpre, data = sample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2910.72  -335.35     3.53   361.38  1787.28
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3241.745     97.963  33.091   <2e-16 ***
## mage           5.508      3.014   1.827   0.0679 .
## monpre         4.871     11.595   0.420   0.6745
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 562.1 on 1609 degrees of freedom
## Multiple R-squared:  0.002072,   Adjusted R-squared:  0.000832
## F-statistic: 1.671 on 2 and 1609 DF,  p-value: 0.1884
```

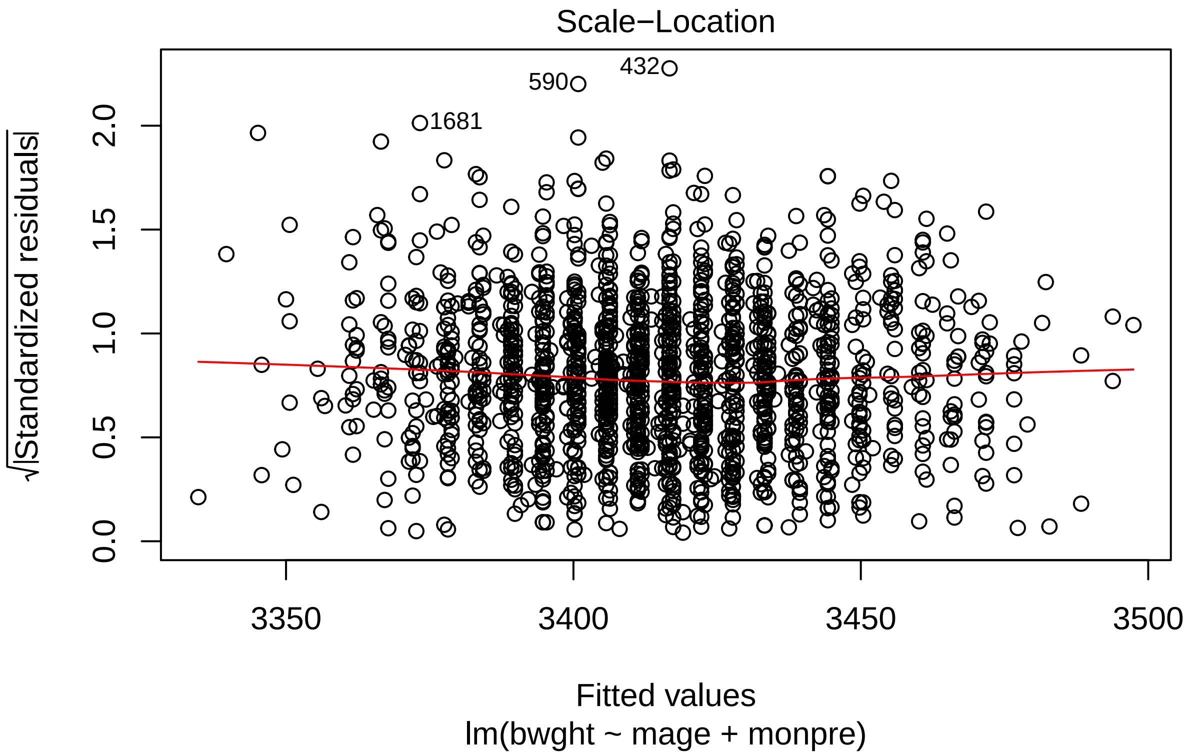
```r
plot(model_1, which = 1)
```



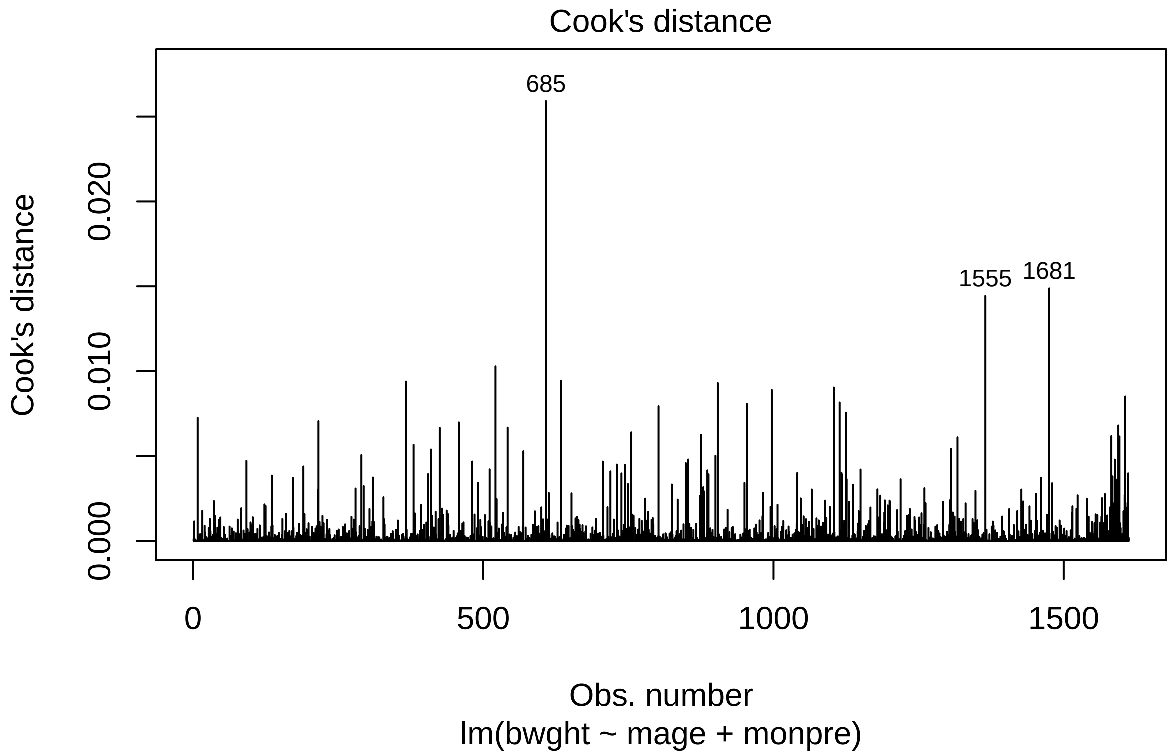
```
plot(model_1, which = 2)
```



```r
plot(model_1, which = 3)
```

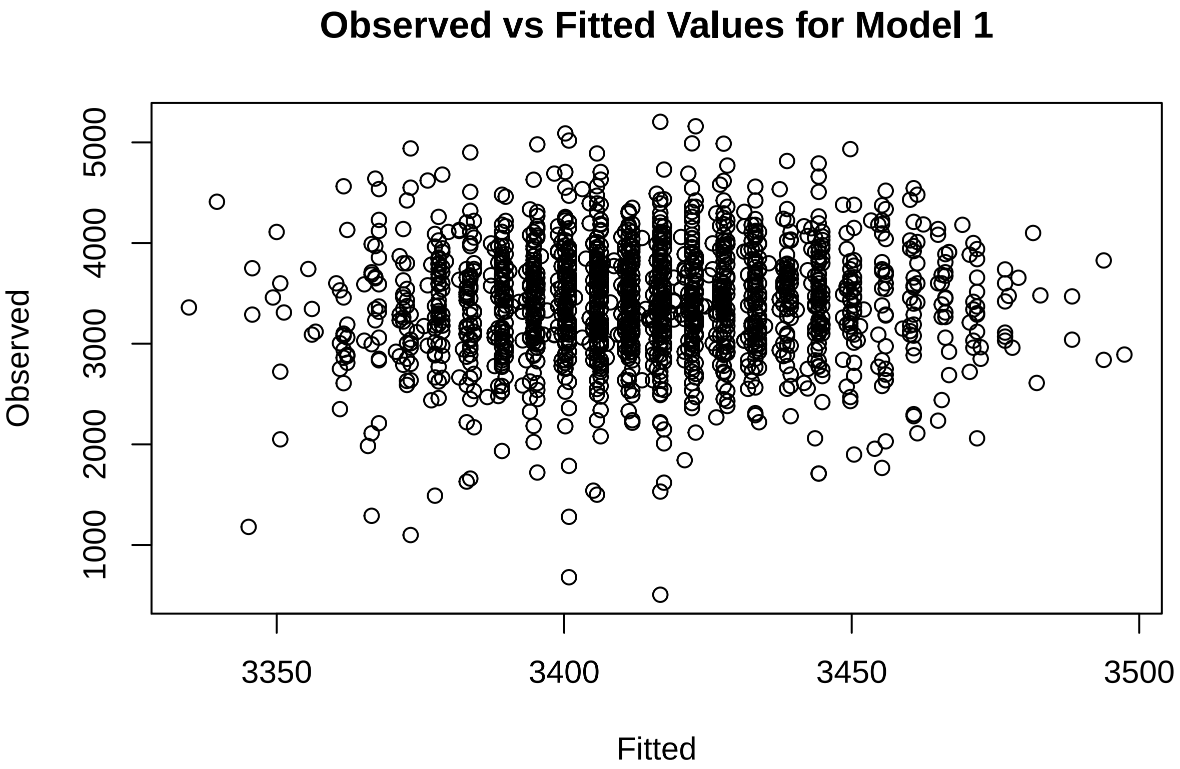


```
plot(model_1, which = 4)
```



Assumption 1: Linear Population Model
The assumption of a linear population model is met since this model has been defined to be linear in its parameters.

```
plot(model_1$fitted.values, (model_1$fitted.values + model_1$residuals), main = "Observed vs Fitted Valu
```



Additionally, the observed vs predicted values plot for this model does not provide a strong indication of non-linearity.

Assumption 2: Random Sampling
The sample used is expected to be random by design, and unfortunately, the data collection process cannot be evaluated. However, a closer look at the 'fblck', 'fwhte', 'mblck', and 'mwhte' variables shows that less than 100 babies (5%) in the data are black (defined as having both black parents), while the vast majority are white (defined by having both white parents). Given that the black population in the US is approximately 12% (or 'over 10%') of the total population, the sample used in this study is not completely representative of the population and may have a minor grouping issue with respect to the race of babies. Examination of the 'monpre', 'npvis', 'omaps', 'fmaps', 'cigs', and 'drink' variables shows that most mothers in the data had early and frequent prenatal care, most infants had high APGAR scores (especially by the five-minute mark), and that most mothers refrained from cigarettes and drinking during pregnancy. These variables lack data with respect to poor prenatal care and poor infant health, but this is acceptable since such distributions are roughly representative of the population.

The under-representation of black babies is not large in this sample– this study thus assumes that the sample is random.

Assumption 3: No Perfect Multicollinearity
The variance inflation factor explains how much the standard error of each coefficient is inflated due to collinearity with other variables:

```
vif(model_1)
```

```
##     mage   monpre
## 1.042237 1.042237
```

The VIF is low enough (<<4) to allay concerns about multicollinearity in this model. In fact, the VIF being close to 1 demonstrates almost no multicollinearity in the model.

Assumption 4: Zero-Conditional Mean
The smoothing curve in the residuals vs fitted plot for this model (which tracks the conditional mean of the residuals) shows nearly no curvature, especially in the bulk of data points, which indicates that the zero-conditional mean assumption holds.

Assumption 5: Homoskedasticity
The same residuals vs fitted plot analyzed previously also demonstrates a band of approximately equal width in the residuals, across all fitted values, suggesting that the assumption of homoskedasticity holds for this model. This is also apparent in the scale-location plot, in which its horizontal smoothing curve is expected when homoskedasticity holds.

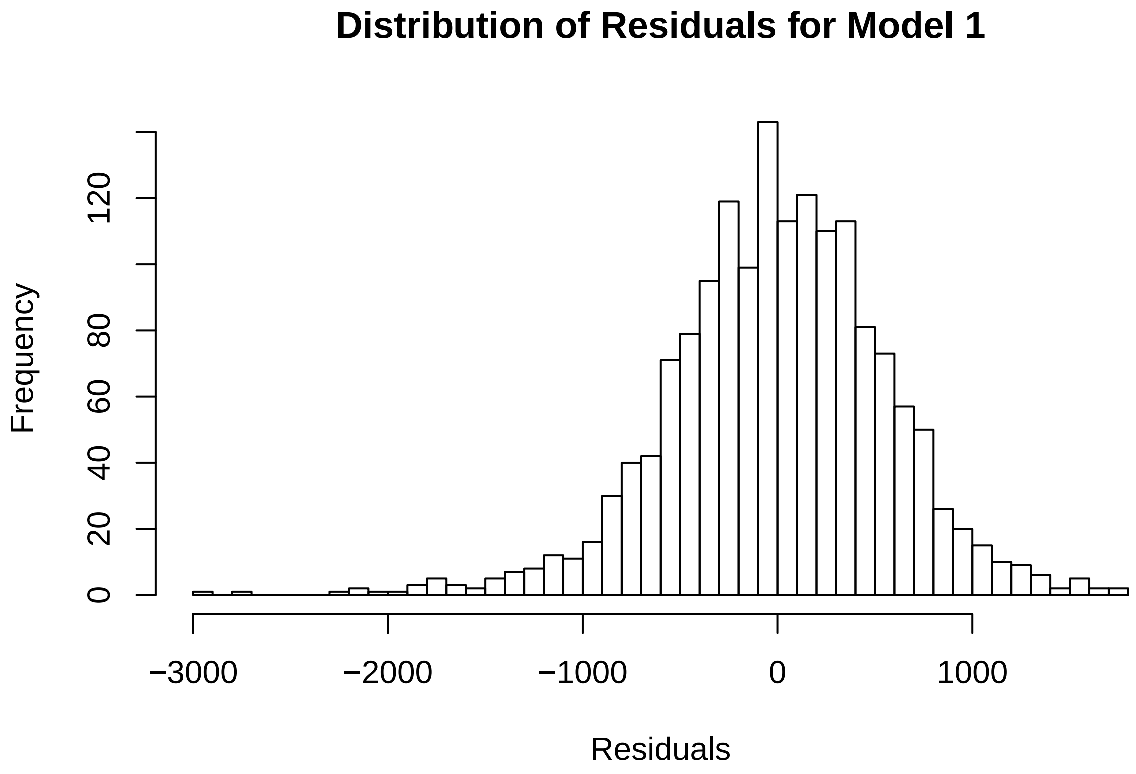
```r
bptest(model_1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_1
## BP = 5.8613, df = 2, p-value = 0.05336
```

The Breush-Pagan test is not significant enough to reject the null hypothesis of constant variance at the 5% significance level, which is consistent with the homoskedasticity demonstrated in the diagnostic plots. Its p-value of 0.06 means this test does have borderline significance, but this low p-value may simply be due to the large sample size. For safe measure, any hypothesis testing using this model should still use the heteroskedasticity-robust Huber-White standard errors.

Assumption 6: Normality of Errors

```r
hist(model_1$residuals, breaks = "fd", main = "Distribution of Residuals for Model 1", xlab = "Residuals
```



```
shapiro.test(model_1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model_1$residuals
## W = 0.98442, p-value = 3.335e-12
```

Both the normal q-q plot and the histogram of residuals show a minor departure from normality in the residuals, suggesting a violation of the assumption of normality of errors. Additionally, the Shapiro-Wilk test is significant at any significance level, which further indicates that this analysis should reject the null hypothesis that the residuals come from a normal distribution. However, due to the large sample size, the Central Limit Theorem allows the OLS coefficients in this model to still be treated as normal.

Additional Notes:
This model is not concerned with outliers since the diagnostic plots show that Cook's distance (a measure of influence) is small for every observation.

```
AIC(model_1)
```

```
## [1] 24992.94
```

The Akaike Information Criterion (AIC), a parsimony-adjusted measure of fit, for this model is 28420.36.

```
residualsSquared = (model_1$residuals)^2
model_1_unrestricted = lm(bwght ~ mage + monpre + residualsSquared, data = sample)
summary(model_1_unrestricted)
```

```
##
```

```
## Call:
## lm(formula = bwght ~ mage + monpre + residualsSquared, data = sample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1276.60  -376.12   -57.13   323.13  2386.52
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.353e+03  9.643e+01  34.775   <2e-16 ***
## mage              4.303e+00  2.945e+00   1.461    0.144
## monpre           -6.493e-03  1.133e+01  -0.001    1.000
## residualsSquared -2.082e-04  2.318e-05  -8.980   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 548.7 on 1608 degrees of freedom
## Multiple R-squared:  0.04973,    Adjusted R-squared:  0.04796
## F-statistic: 28.05 on 3 and 1608 DF,  p-value: < 2.2e-16
```

The regression specification error test (RESET) shows that when adding the squared residuals to the model as an independent variable, the coefficient for that term is highly significant. This suggests that the model is actually misspecified.

Model 2

For model 2, we took the variables in model 1 and made a few additions. Namely, given the exploratory data analysis showed that "other" babies had far lower birth weights than other races, we added an indicator variable for race which = 1 if baby's race was "other" and =0 otherwise.

We saw in the exploratory data analysis that visits per month has a concave, parabolic relationship with birthweight when binned into quintiles. Further we know that this variable is susceptible to outliers. As a result we made the following transformations:

log: all positive values and known zero point and there are outliers in the data.

quadratic: from binning the data we know that there is an decreasing effect on birthweight – ie as number of visits reaches high levels (5th quintile), there is a decreasing relationship with birthweight.

Similiarly birthweight appears ripe for log transformation as well given its right tail, all positive values, and a known zero point. As a result we transformed the dependent variable, birthweight, to log(birthweight)

Finally we also included a quadratic term for mother's age given early diagnostic plots which suggest a parabolic relationship there as well.

```
sample$logvis_mo = ifelse(sample$visits_pr_mo > 0, log(sample$visits_pr_mo), 0)
sample$logvis_mo_sq = ifelse(sample$visits_pr_mo > 0, (log(sample$visits_pr_mo))^2, 0)
```

```
model2_3=  lm(lbwght ~ cigs + drink + logvis_mo + logvis_mo_sq + otherbb+ male + mage + magesq, data = s
summary(model2_3)
```

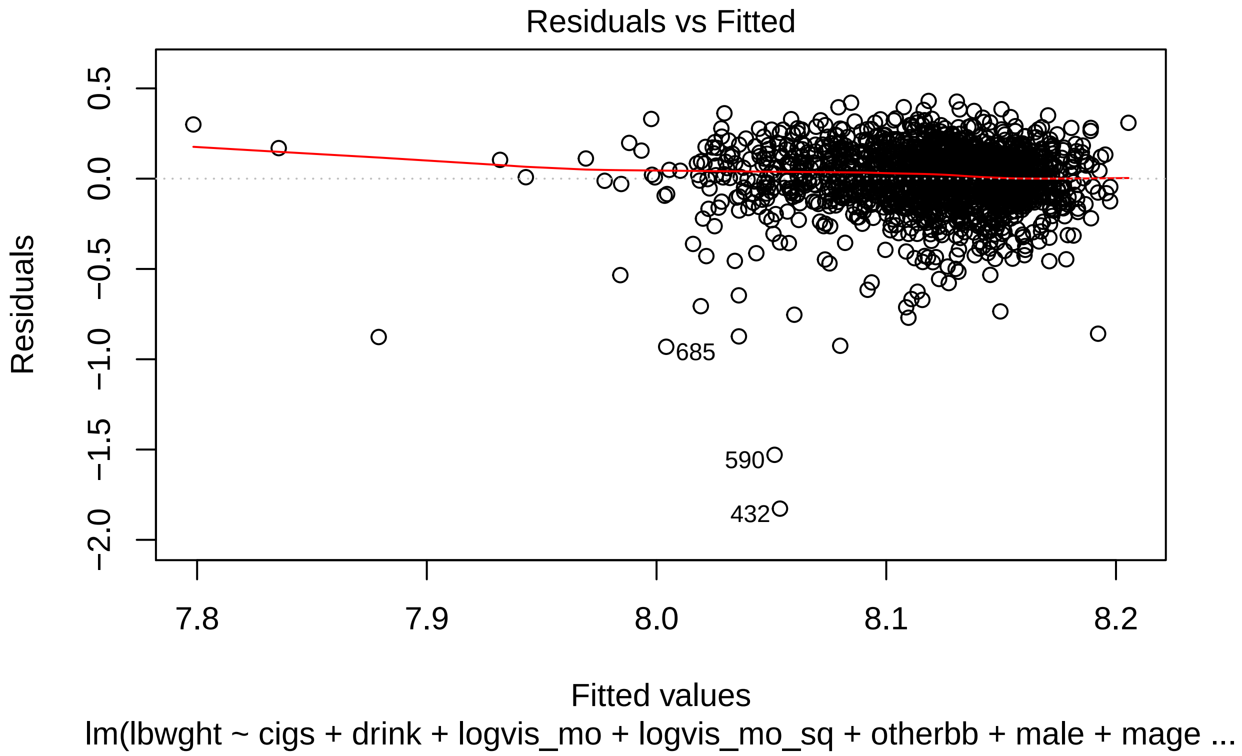
```
##
```

```
## Call:
## lm(formula = lbwght ~ cigs + drink + logvis_mo + logvis_mo_sq + 
##     otherbb + male + mage + magesq, data = sample)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.82719 -0.08878  0.01914  0.11214  0.43024 
## 
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   7.6118655  0.1302010  58.462  < 2e-16 ***
## cigs         -0.0029285  0.0011447  -2.558 0.010612 *  
## drink        -0.0038256  0.0155674  -0.246 0.805910    
## logvis_mo     0.1260193  0.0219138   5.751 1.06e-08 ***
## logvis_mo_sq -0.0387735  0.0122932  -3.154 0.001640 ** 
## otherbb1     -0.0807043  0.0215247  -3.749 0.000184 ***
## male1         0.0285725  0.0091367   3.127 0.001796 ** 
## mage          0.0291450  0.0087879   3.316 0.000932 ***
## magesq       -0.0004579  0.0001475  -3.104 0.001940 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.1825 on 1603 degrees of freedom
## Multiple R-squared:  0.04616,    Adjusted R-squared:  0.04139 
## F-statistic: 9.696 on 8 and 1603 DF,  p-value: 3.272e-13
```

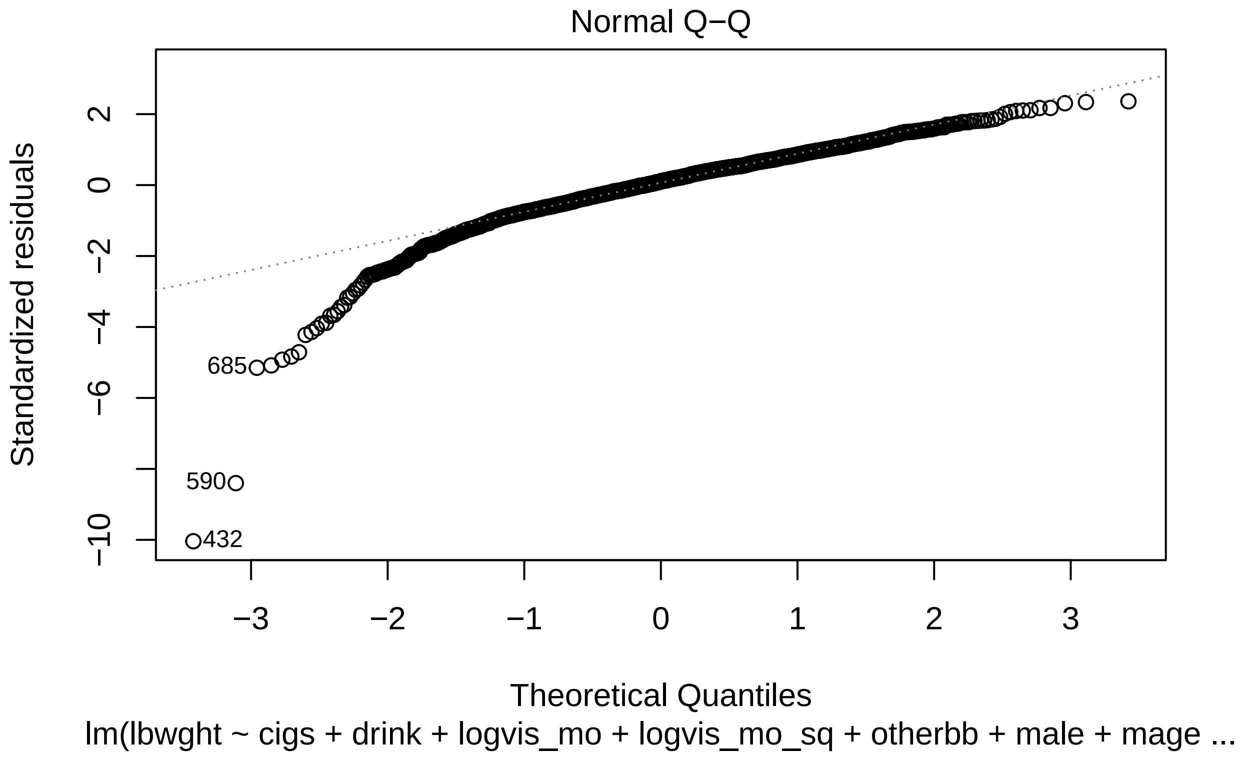
```
vif(model2_3)
```

```
##         cigs        drink    logvis_mo logvis_mo_sq      otherbb 
##     1.047248     1.042378     2.423610     2.419653     1.007186 
##         male         mage       magesq 
##     1.009160    84.014700    83.923120 
```

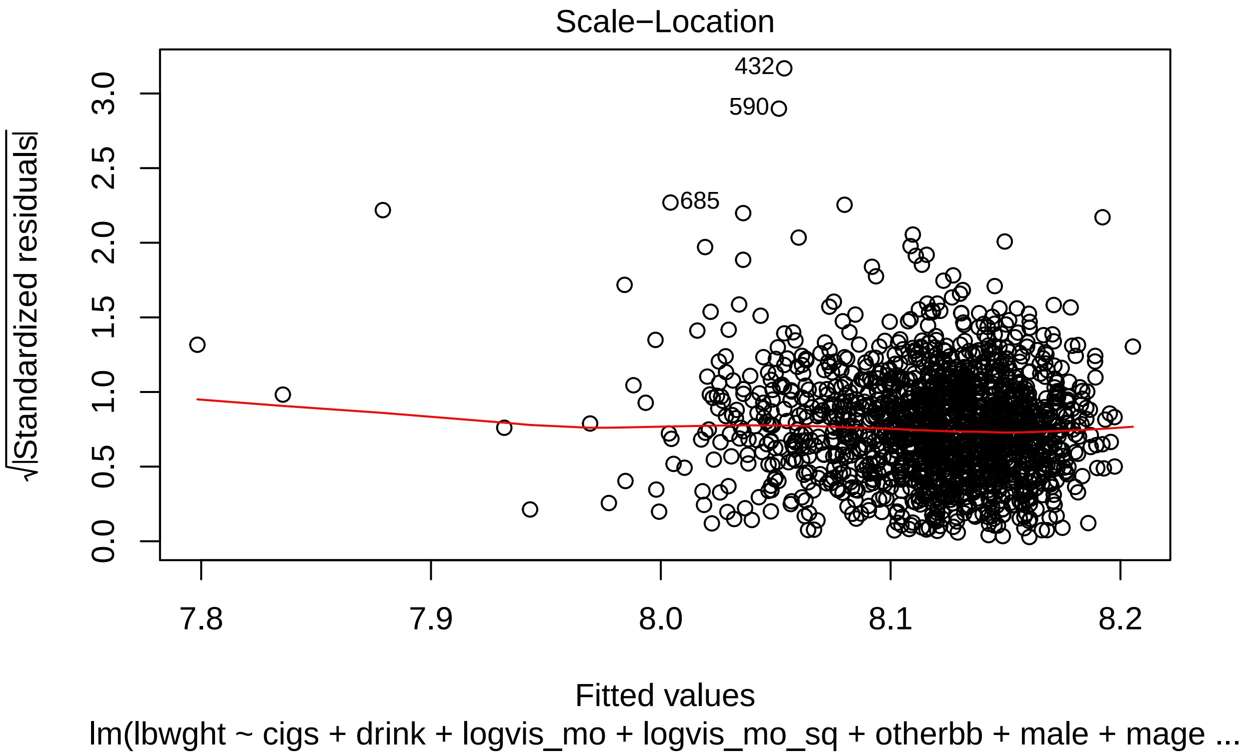
```
plot(model2_3, which = 1)
```



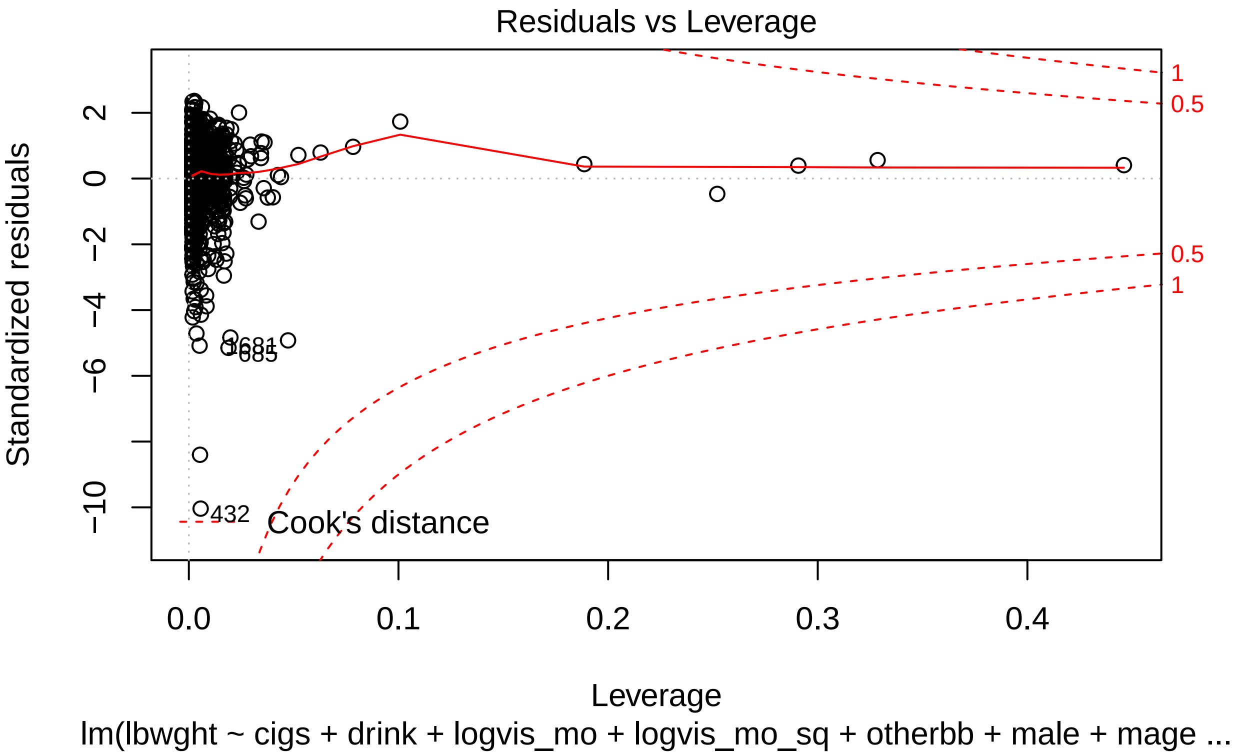
```r
plot(model2_3, which = 2)
```



```
plot(model2_3, which = 3)
```

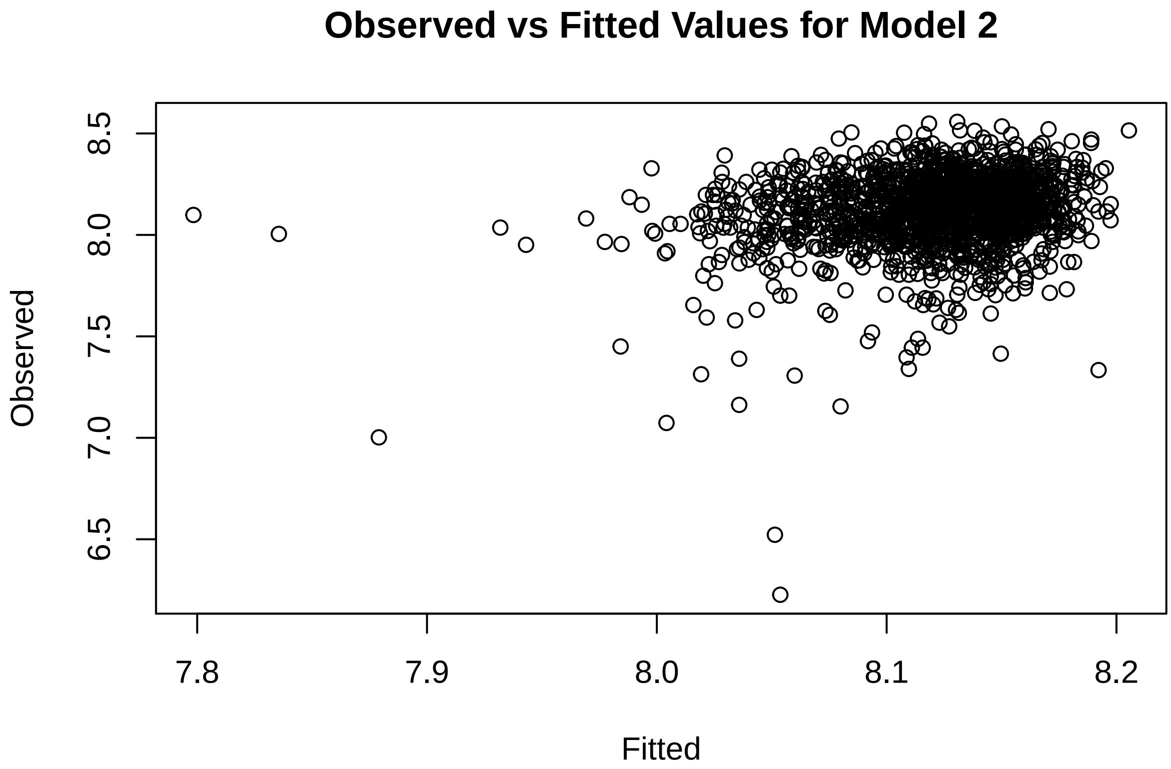


```r
plot(model2_3, which = 5)
```



Changes to the 6 Classical Linear Model Assumptions for Model 2: The assumption of a linear population is still met since this model has been defined to be linear in its parameters.

```
plot(model2_3$fitted.values, (model2_3$fitted.values + model2_3$residuals), main = "Observed vs Fitted
```



Additionally, the observed vs predicted values plot for this model does not provide a clear indication of non-linearity.

The sample has not changed for Model 2, so it is still considered to be random.

```
X2 = data.matrix(subset(sample, select=c("cigs", "drink", "logvis_mo", "logvis_mo_sq", "otherbb", "male
(Cor = cor(X2))
```

```
##                       cigs        drink   logvis_mo  logvis_mo_sq
## cigs          1.0000000000  0.189985308 -0.02260892 -0.0008940864
## drink         0.1899853078  1.000000000  0.03787852  0.0217150444
## logvis_mo    -0.0226089166  0.037878518  1.00000000  0.7644971751
## logvis_mo_sq -0.0008940864  0.021715044  0.76449718  1.0000000000
## otherbb      -0.0549783143 -0.015736583  0.05109556  0.0305131890
## male         -0.0113782175 -0.047725348 -0.05154273 -0.0688795938
## mage         -0.0568818752  0.004740146 -0.03143181 -0.0542498230
## magesq       -0.0515798587  0.007867081 -0.02371570 -0.0484624548
##                   otherbb        male         mage       magesq
## cigs         -0.05497831 -0.01137822 -0.056881875 -0.051579859
## drink        -0.01573658 -0.04772535  0.004740146  0.007867081
## logvis_mo     0.05109556 -0.05154273 -0.031431810 -0.023715695
## logvis_mo_sq  0.03051319 -0.06887959 -0.054249823 -0.048462455
## otherbb       1.00000000 -0.01778549  0.028883708  0.026509797
## male         -0.01778549  1.00000000 -0.040176336 -0.040453119
## mage          0.02888371 -0.04017634  1.000000000  0.993977724
## magesq        0.02650980 -0.04045312  0.993977724  1.000000000
```

The VIF is low (<<4) for all variables in model 2 except for ‘mage’ and ‘magesq’. The correlation between these two variables is extremely high (0.99), which results in a large standard error for their coefficients. However, this does not bias results, and does not affect interpretation of the coefficients for the main independent variables of interest, 'logvis_mo' and 'logvis_mo_sq'.

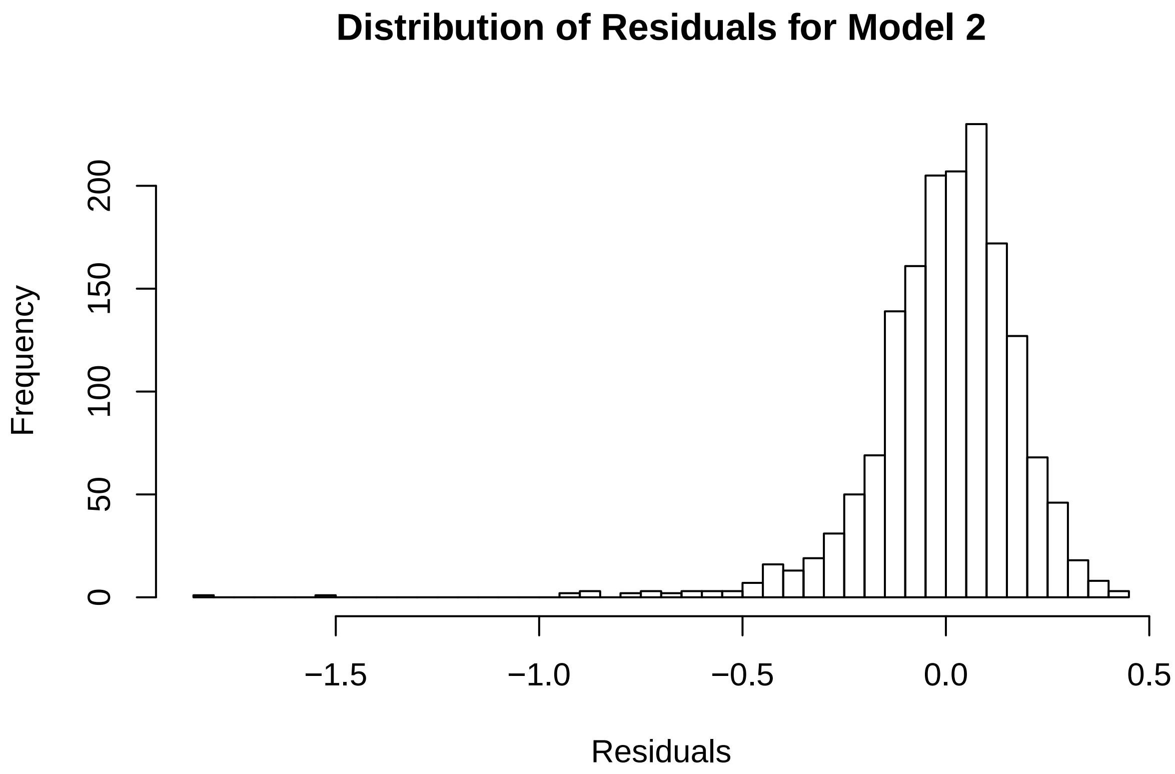
The residuals vs fitted plot for Model 2 shows some curvature, but is reasonably flat in the bulk of the data, so the zero-conditional mean assumption still holds.

```
bptest(model2_3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model2_3
## BP = 40.223, df = 8, p-value = 2.912e-06
```

The Breush-Pagan test for Model 2 is highly significant, which indicates that heteroskedasticity is present. However, it is unclear that this will be an issue since the very low p-value may again be due to the large sample size, and since the scale-location plot shows a flat smoothing curve in the bulk of data. The heteroskedasticity-robust Huber-White standard errors will be used for this model due to the uncertainty surrounding violoation of the homoskedasticity assumption.

```
hist(model2_3$residuals, breaks = "fd", main = "Distribution of Residuals for Model 2", xlab = "Residual
```



```
shapiro.test(model2_3$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model2_3$residuals
## W = 0.89826, p-value < 2.2e-16
```

The large sample size and the Central Limit Theorem again allow Model 2 to rely on OLS asymptotics despite the departure from normality demonstrated by its normal q-q plot, the histogram of residuals, and the Shapiro-Wilk test.

Cook's distance remains small for every observation, so outliers are again of no concern.

```
AIC(model2_3)
```

```
## [1] -898.1229
```

The Akaike Information Criterion (AIC) for Model 2 is -898.1228. This is much lower than the AIC for Model 1, indicating a substantially improved fit for Model 2.

```
residualsSquared2 = (model2_3$residuals)^2
model2_3_unrestricted=  lm(lbwght ~ cigs + drink + logvis_mo + logvis_mo_sq + otherbb+ male + mage + ma
summary(model2_3_unrestricted)
```

```
##
## Call:
## lm(formula = lbwght ~ cigs + drink + logvis_mo + logvis_mo_sq +
##     otherbb + male + mage + magesq + residualsSquared2, data = sample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41054 -0.10511 -0.00729  0.09680  0.79131
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        7.7720928  0.1106666  70.230  < 2e-16 ***
## cigs              -0.0029538  0.0009713  -3.041  0.00240 **
## drink             -0.0058976  0.0132098  -0.446  0.65533
## logvis_mo          0.0577123  0.0187946   3.071  0.00217 **
## logvis_mo_sq      -0.0154807  0.0104728  -1.478  0.13956
## otherbb1          -0.0855296  0.0182655  -4.683 3.07e-06 ***
## male1              0.0225804  0.0077566   2.911  0.00365 **
## mage               0.0225643  0.0074615   3.024  0.00253 **
## magesq            -0.0003571  0.0001252  -2.852  0.00441 **
## residualsSquared2 -0.8036660  0.0321639 -24.987  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1549 on 1602 degrees of freedom
## Multiple R-squared:  0.3136, Adjusted R-squared:  0.3098
## F-statistic: 81.34 on 9 and 1602 DF,  p-value: < 2.2e-16
```

The regression specification error test again shows that when adding the squared residuals to the model as an independent variable, the coefficient for that term is highly significant, suggesting that model 2 is still misspecified, and moreover that there are still some important explanatory variables that have not been observed in the data.