Does Prenatal Care Result in Better Health Outcomes for Newborns?

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.

```
sample = na.exclude(data, complete.cases(sample))
sample$lbw = factor(sample$lbw)
sample$vlbw = factor(sample$vlbw)
sample$male = factor(sample$male)
sample$mwhte = factor(sample$mwhte)
sample$mblck = factor(sample$mblck)
sample$moth = factor(sample$moth)
sample$fwhte = factor(sample$fwhte)
sample$fblck= factor(sample$fblck)
sample$foth = factor(sample$foth)
```

Exploratory Data Analysis:

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.

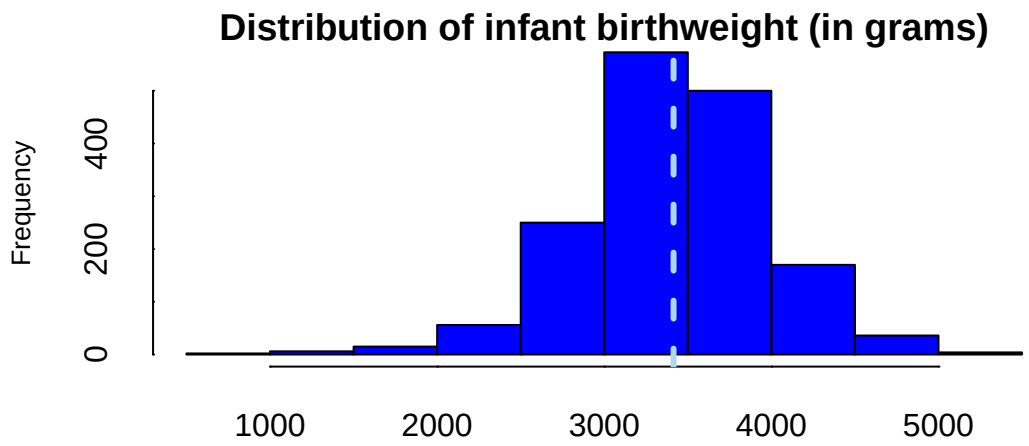
```
summary(sample)
```

```
##       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
```

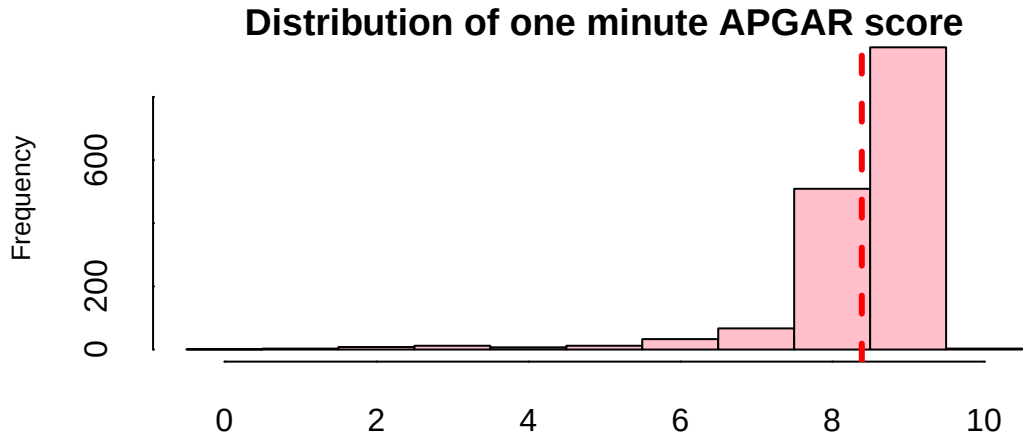
```
##  3rd Qu.:32.00   3rd Qu.:16.00   3rd Qu.:2.000   3rd Qu.:13.00
##  Max.   :44.00   Max.   :17.00   Max.   :9.000   Max.   :40.00
##       fage            feduc           bwght           omaps
##  Min.   :18.00   Min.   : 3.00   Min.   : 506   Min.   : 0.00
##  1st Qu.:28.00   1st Qu.:12.00   1st 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 initial plots show that birthweight is approximately normally distributed with a mean of 3415 grams. 'omaps' and 'fmaps,' Apgar scores are an ordinal variable as they represent a values from 0 to 10 where 10 is the best. They are a measure of the physical condition of a newborn and computed by adding points (0, 1, or 2) for heart rate, respiratory effort, muscle tone, response to stimulation, and skin coloration. 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 (9-10).

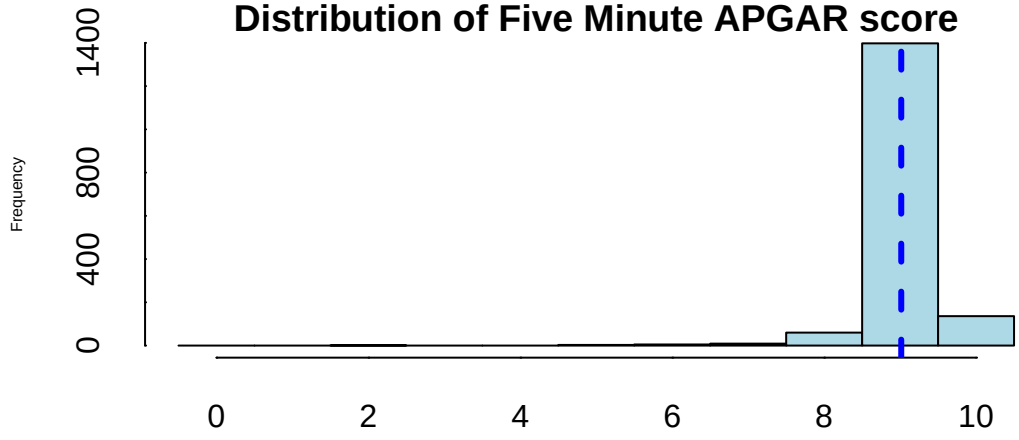
```
par(mar=c(6,6,8,3),cex.axis=1,cex.lab=0.8)
hist(sample$bwght, main ="", xlab = "", col = "blue", tcl = 0)
title("Distribution of infant birthweight (in grams)", line = 0, cex.lab =0.5)
abline(v = mean(sample$bwght, na.rm= TRUE), col="light blue", lwd=3, lty=2)
```




```
par(mar=c(6,6,8,3),cex.axis=1,cex.lab=0.8)
hist(sample$omaps, main ="", xlab = "" ,breaks = 0:11 - 0.5, col = "pink", tcl = 0)
title("Distribution of one minute APGAR score", line = 0, cex.lab =0.5)
abline(v = mean(sample$omaps, na.rm= TRUE), col="red", lwd=3, lty=2)
```




```
par(mar=c(6,6,8,3),cex.axis=1,cex.lab=0.5)
hist(sample$fmaps, main="", xlab = "", breaks = 0:11 -0.5, col = "light blue", tcl = 0)
title("Distribution of Five Minute APGAR score", line = 0, cex.lab =0.8)
abline(v = mean(sample$fmaps, na.rm= TRUE), col="blue", lwd=3, lty=2)
```

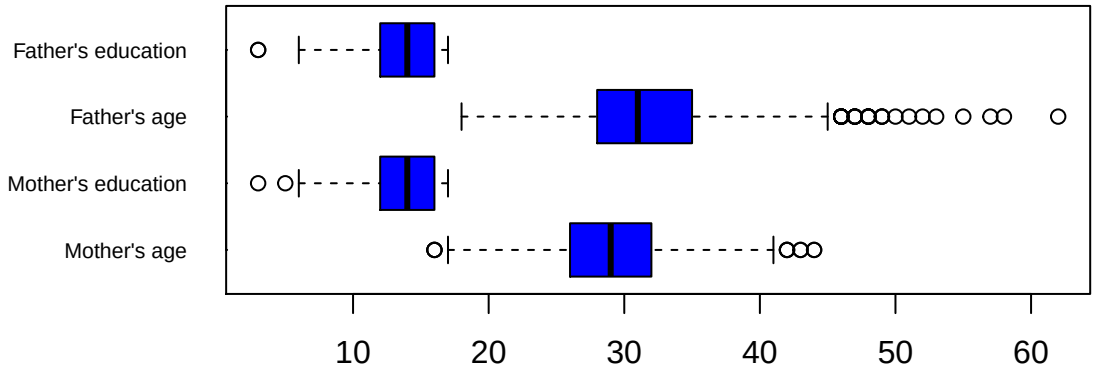


We started out by looking at the parents in our data set. Average mother and father ages are 29.5 and 31.8 years respectively. We notice the presence of several outlier variables (outside the upper whisker) for fathers' age. Histograms for both mother and father ages appear to be approximately normal. 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)
```

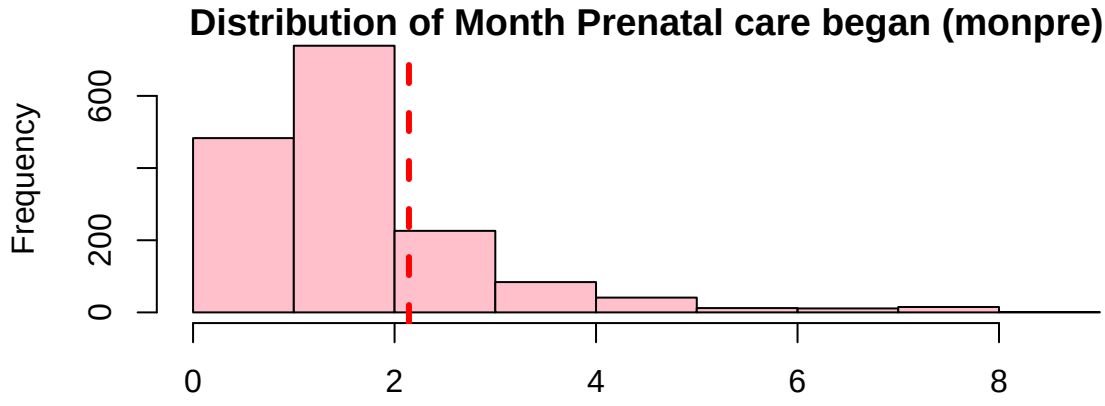
```
## No id variables; using all as measure variables
```

```
par(mar=c(7,7,8,3),cex.axis=1,cex.lab=0.7)
plot(value ~ variable, data=long, horizontal = TRUE, col = "blue", xlab ="", ylab="", yaxt ="n")
axis(2, at = c(1, 2, 3, 4), labels = colnames(parent_char), tcl = 0, las = 2, cex.axis = .7)
```

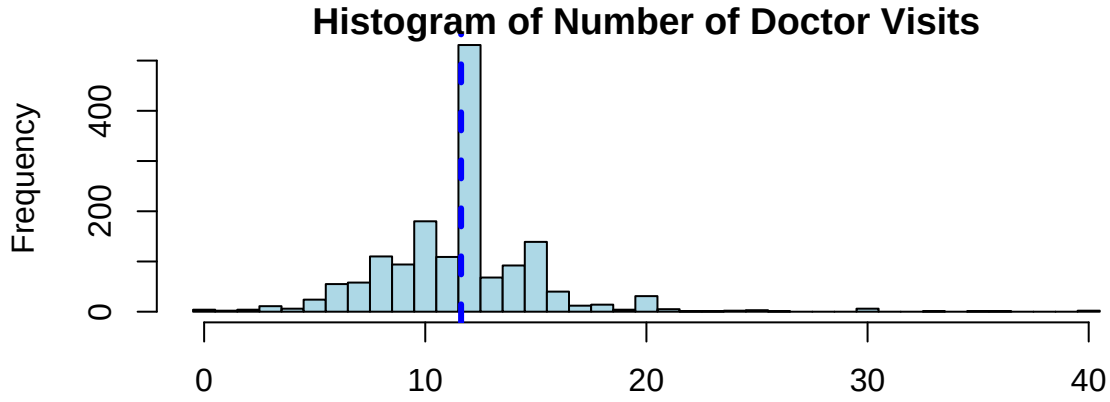


Next we analyzed some of the variables in the data set that we thought might have explanatory potential on infant health including number of prenatal visits, month prenatal care began, cigarettes smoked, and alcoholic beverages consumed. The average mother in this sample began prenatal care a little 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% (n=1460) of the values are between 5 and 15 visits. Start month of prenatal care appears to have positive skew with 90% (n=1448) of the mothers beginning care at 3 months or earlier.

```
par(mar=c(7,7,8,0))
hist(sample$monpre, main="", col = "pink", xlab ="")
title("Distribution of Month Prenatal care began (monpre)", line = 0, cex.lab =0.7)
abline(v = mean(sample$monpre, na.rm= TRUE), col="red", lwd=3, lty=2)
```



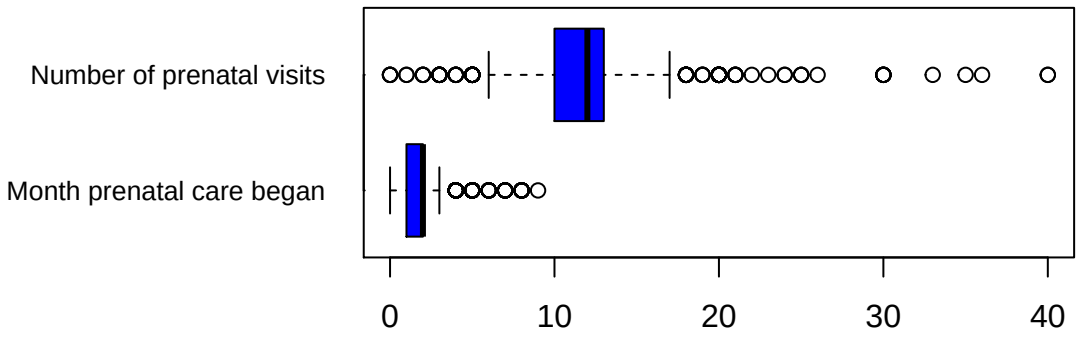
```
par(mar=c(7,7,8,0))
hist(sample$npvis, breaks = 0:41 -.5, main ="", col = "light blue", xlab = "", cex = 0.5)
title(main = "Histogram of Number of Doctor Visits", line = 0, cex.lab =0.7)
abline(v = mean(sample$npvis, na.rm= TRUE), col="blue", lwd=3, lty=2)
```




```
#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", "N
health_var = data.frame(cbind(sample$monpre, sample$npvis))
colnames(health_var)= c("Month prenatal care began", "Number of prenatal visits")

long <- melt(health_var)
```

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

```
par(mar=c(8,11,8,3),cex.axis=1,cex.lab=1)
plot(value ~ variable, data=long, horizontal = TRUE, col = "blue", xlab ="", ylab="",yaxt ="n")
#axis(2, at = c(1, 2, 3, 4), labels = colnames(health_var), tcl = 0, las = 2, cex.axis = .8)
axis(2, at = c(1, 2), labels = colnames(health_var), tcl = 0, las = 2, cex.axis = .8)
```



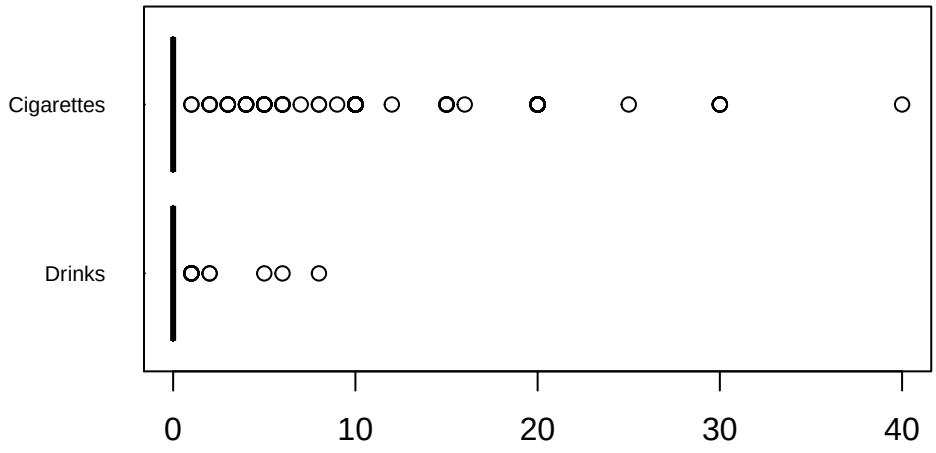
We looked for insights on the smoking and alcoholic consumption behavior of the mothers in this sample. The average mom had 1 cigarette a day but over 90% of the 1612 women we looked at had zero cigarettes a day. Unsurprisingly, analysis of histogram and box plot show that there are several outliers in the sample with one woman having reported smoking 40 cigarettes a day. On alcoholic consumption, all but 16 women had no alcohol consumption which made us question early on how useful of a variable it would be in explaining health outcomes for infant newborns and thus to include in our statistical analysis later on.

```
#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
```

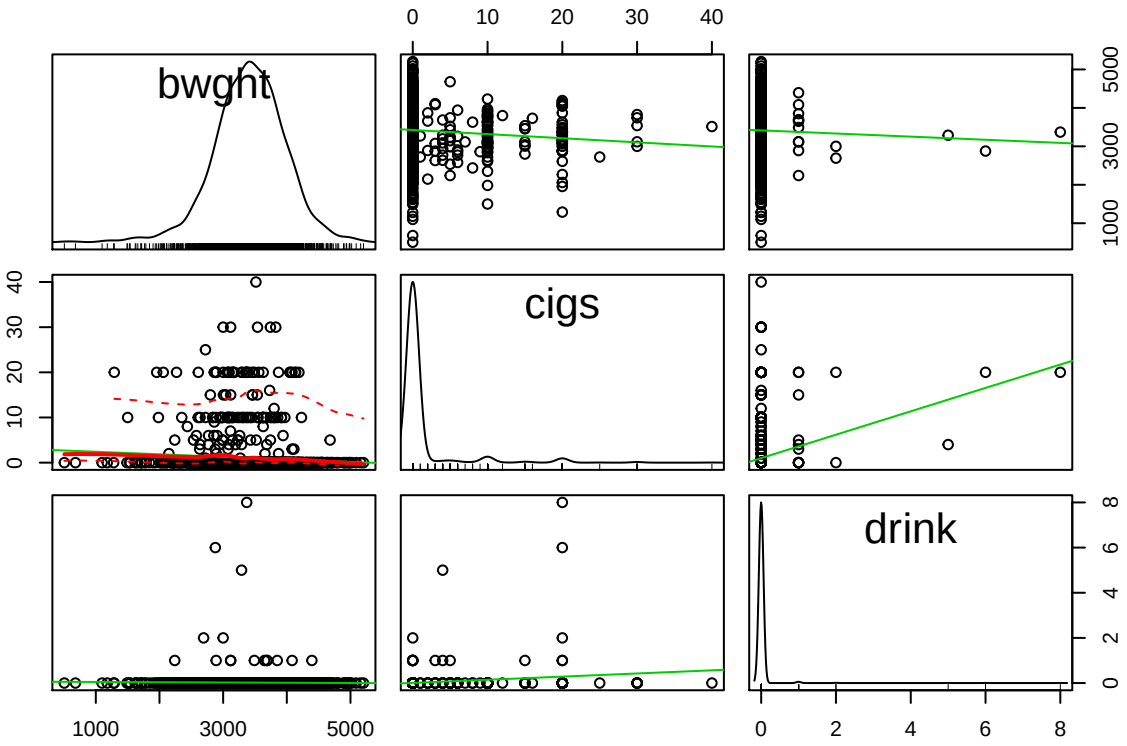
```
par(mar=c(5,5,8,7),cex.axis=1,cex.lab=1)
plot(value ~ variable, data=long, horizontal = TRUE, col = "blue", xlab ="", ylab="", yaxt ="n")
axis(2, at = c(1, 2), labels = colnames(cigs_drinks), tcl = 0, las = 2, cex.axis = .7)
```



When we plot the cigarettes and drink variables against infant birthweight, we do see a slight linear relationship between smoking and birthweight.

```
par(mar=c(7,7,8,0))
scatterplotMatrix(~bwght + cigs + drink, data = sample)
```

```
## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth
```



Lastly on race, 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.

```
sample$blackbb = factor(ifelse(sample$mblck == '1' & sample$fblck == '1', 1, 0))
sample$whitebb = factor(ifelse(sample$mwhte == '1' & sample$fwhte == '1', 1, 0))
sample$halfblk_bb = factor(ifelse((sample$mwhte == '1' & sample$fblck == '1')|(sample$mblck == '1' & sa
sample$otherbb = factor(ifelse(sample$moth == '1' & sample$foth == '1', 1, 0))
sample$halfblk_oth_bb = factor(ifelse((sample$moth == '1' & sample$fblck == '1')|(sample$mblck == '1' &
sample$halfwhte_oth_bb = factor(ifelse((sample$moth == '1' & sample$fwhte == '1')|(sample$mwhte == '1' &

num_obs = c(nrow(sample[sample$blackbb == '1', ]), nrow(sample[sample$whitebb == '1', ]),
      nrow(sample[sample$halfblk_bb == '1', ]), nrow(sample[sample$otherbb == '1', ]),
      nrow(sample[sample$halfblk_oth_bb == '1', ]), nrow(sample[sample$halfwhte_oth_bb == '1', ]))

variable = c("sample$blackbb", "sample$whitebb", "sample$halfblk_bb", "sample$otherbb",
      "sample$halfblk_oth_bb", "sample$halfwhte_oth_bb")

race = c("Black babies", "White babies", "Half black/half white babies", "Other babies",
      "Half black/half other babies", "Half white/half other babies")

race_bwght = round(c(mean(sample$bwght[sample$blackbb == '1']), mean(sample$bwght[sample$whitebb == '1'
              mean(sample$bwght[sample$halfblk_bb == '1']), mean(sample$bwght[sample$otherbb == '1']),
              mean(sample$bwght[sample$halfblk_oth_bb == '1']), mean(sample$bwght[sample$halfwhte_oth_
```

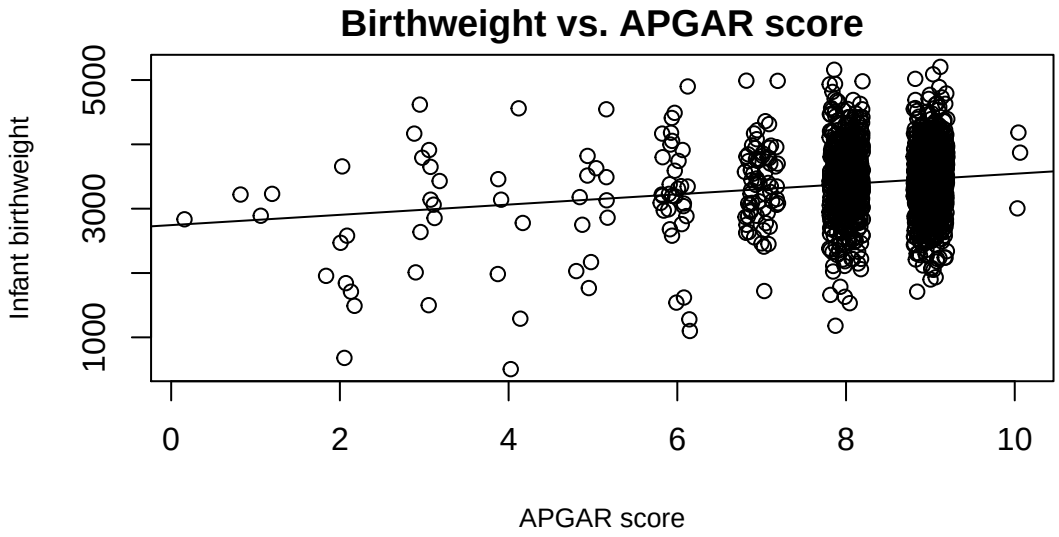
```
baby_races = data.frame(cbind(race, num_obs, race_bwght))
grid.table(baby_races)
```

	race	num_obs	race_bwght
1	Black babies	83	3413
2	White babies	1420	3425
3	Half black/half white babies	13	3320
4	Other babies	76	3184
5	Half black/half other babies	1	3600
6	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 some small positive correlation in the data set.

```
#relationship between bwght and omaps
par(mar=c(6,6,8,3),cex.axis=1,cex.lab=0.5)
z = plot(jitter(sample$omaps), jitter(sample$bwght), main = "", xlab = "APGAR score", ylab = "Infant bi
title(main = "Birthweight vs. APGAR score", line = 0.5, cex.lab = 0.6)
abline(lsfit(sample$omaps, sample$bwght))
```



Correlation between OMAPS and BWGHT variables.

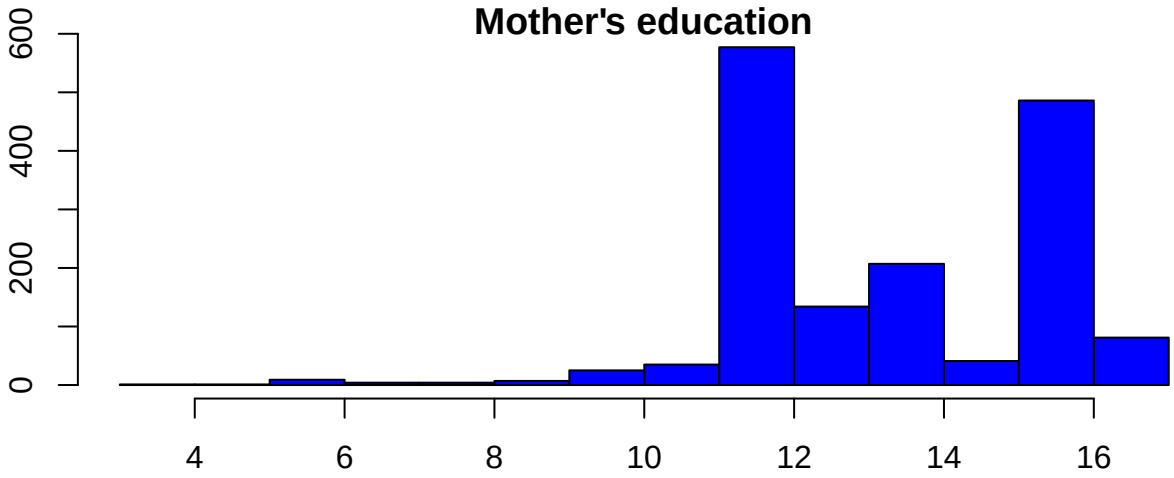
```
cor(sample$omaps, sample$bwght)
```

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

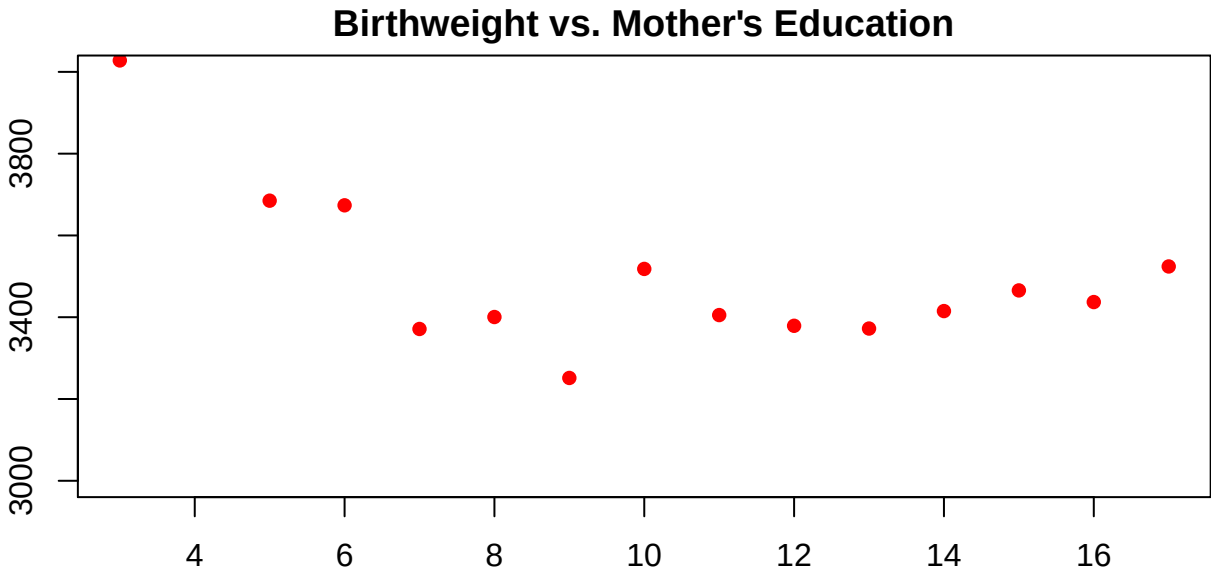
Birthweight vs. Parent education and age

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 histogram
par(mar=c(3,3,10,0))
z = hist(sample$meduc, xlab ="Years of education", main = "", col = "blue")
title(main = "Mother's education", line = 0, cex.lab = 0.7)
```

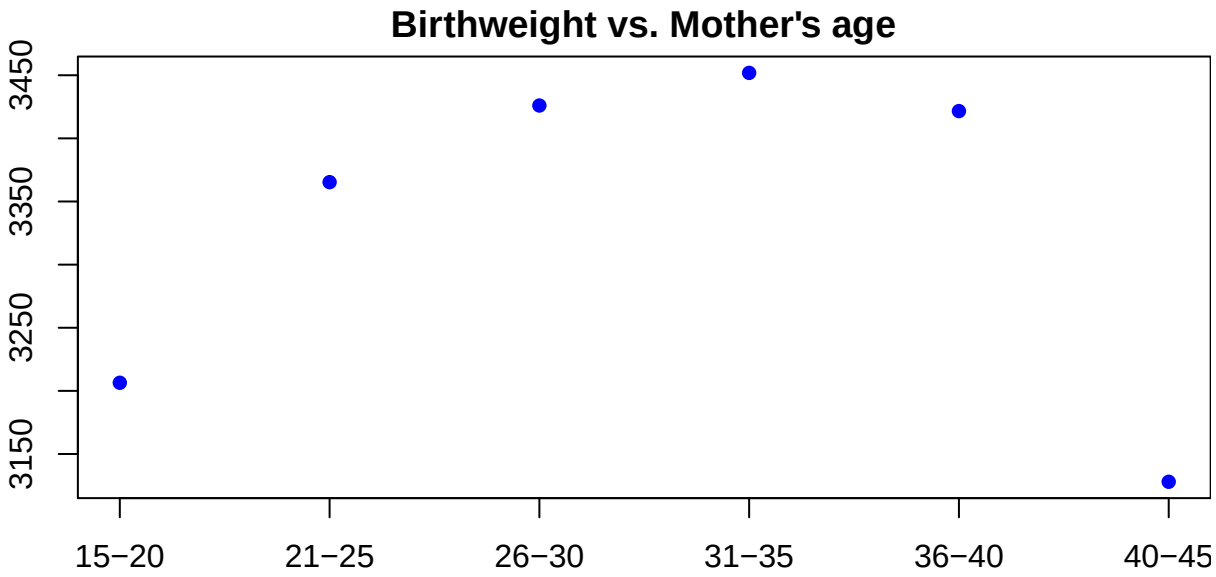



```r
b = data.frame(cbind(z$breaks))
colnames(b) = "education_years"

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(3,3,8,0))
plot(a$education_years, a$average_birthweight, ylim = seq(3000,4100,1000), xlab = "Years of Mothers' Edu
     ylab = "Average Birthweight", main="", col = "red", pch = 16)
title(main ="Birthweight vs. Mother's Education", line = 0.5, cex =0.5)
```




```
#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")
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"))
```

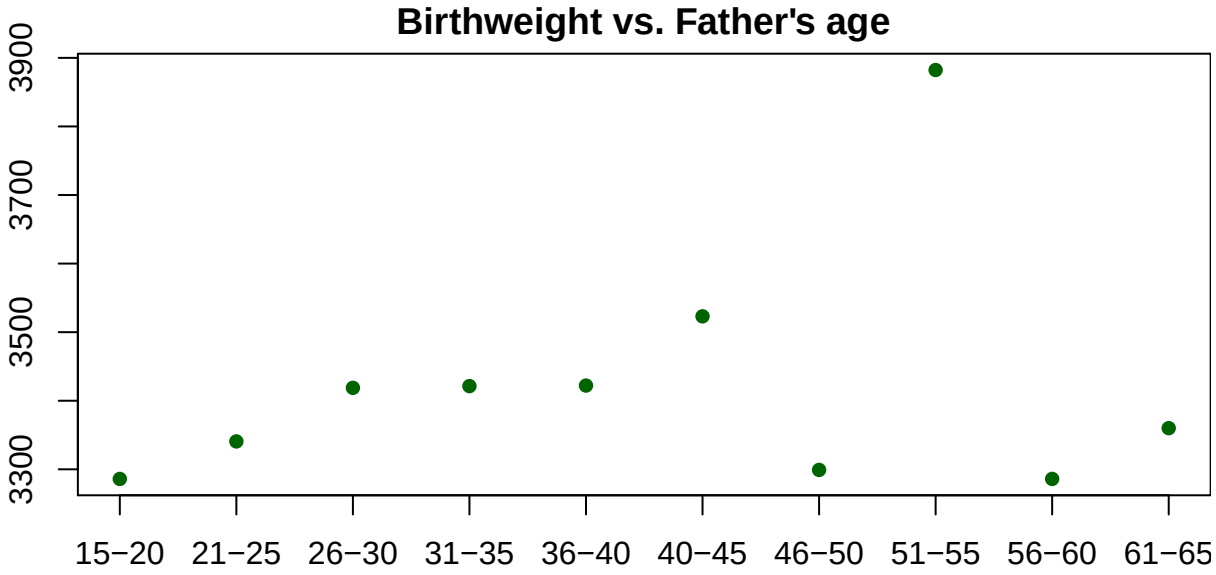
```
par(mar=c(3,3,8,0))
plot(by(mage_exp$birthweight,mage_exp$agebin, mean), main = "", xaxt = "n",pch = 16, col = "blue")
title(main ="Birthweight vs. Mother's age", line = 0.5, cex =0.5)
axis(1, at =seq(1,6,1), labels = c("15-20","21-25","26-30","31-35","36-40","40-45"),xlab = "Mother's ag
```



Interestingly, we don't see the same pattern between father's age and infant birthweight.

```
fage_exp = data.frame(cbind(sample$fage, sample$bwght, sample$omaps, sample$fmaps))
colnames(fage_exp) = c("father_age", "birthweight", "OMAPS", "FMAPS")
fage_exp$agebin <- cut(fage_exp$father_age, breaks = seq(15, 65, by = 5),
                       labels = c("15-20","21-25","26-30","31-35","36-40","40-45",
                                  "46-50", "51-55", "56-60", "61-65"))
```

```
par(mar=c(3,3,8,0))
plot(by(fage_exp$birthweight,fage_exp$agebin, mean), main ="", xaxt = "n", pch = 16, col = "dark green"
title(main ="Birthweight vs. Father's age", line = 0.5, cex =0.5)
axis(1, at =seq(1,10,1), labels = c("15-20","21-25","26-30","31-35","36-40","40-45", "46-50", "51-55",
```



Birthweight and Apgar scores 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 at all 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.

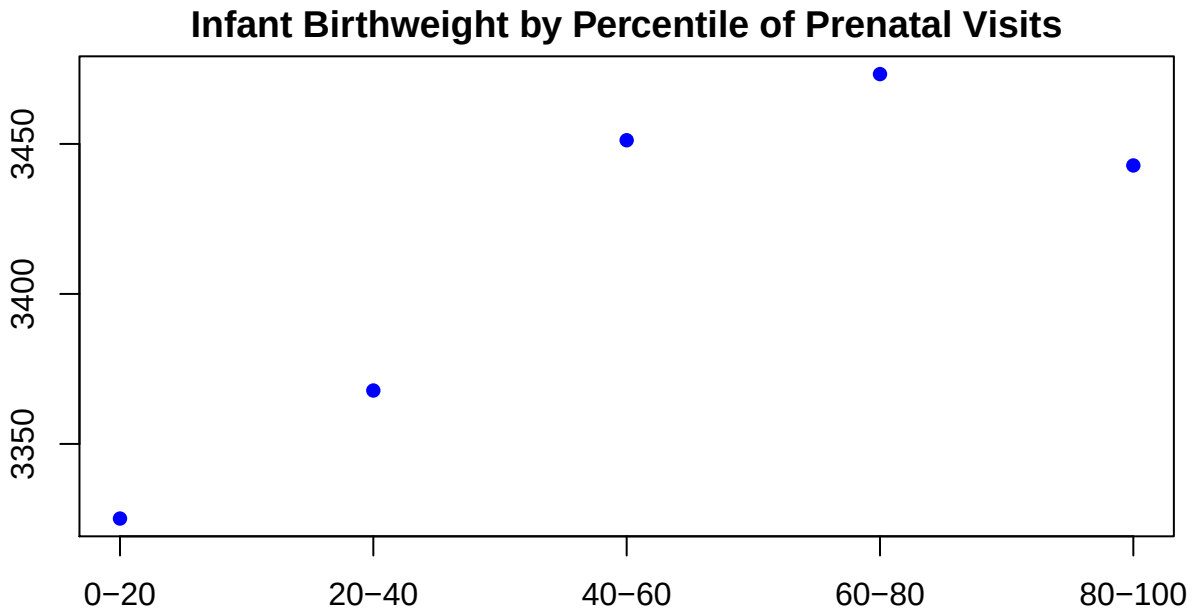
As we looked deeper into both variables, both variables had outlier values (with very few observations) and seemed to be skewing the summary statistics. For example, the median number of visits by the baby mother was 12 but there was an observation where a mother had 40 visits. So we attempted cut up a couple of different ways in an effort to minimize the skew. Namely we 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)
```

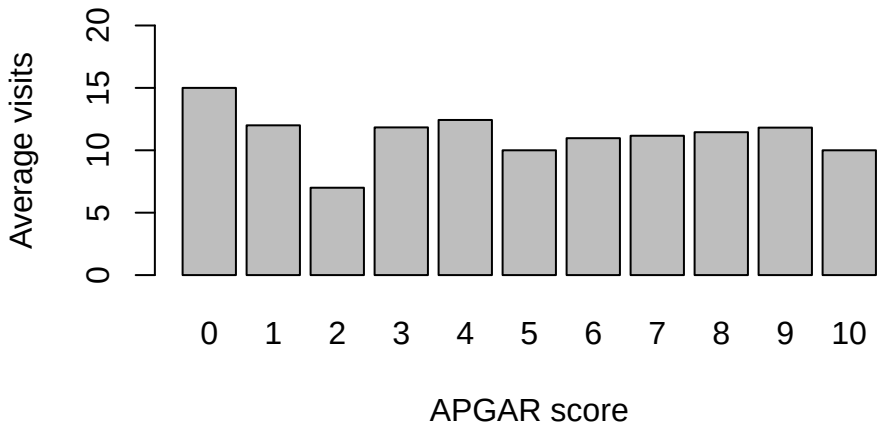
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.429   1.714   1.769   1.875  33.000
```

```r
sample$visitsbin <- cut(sample$visits_pr_mo, breaks=c(quantile(sample$visits_pr_mo, probs = seq(0, 1, by
par(mar=c(2,2,8,2))
plot(by(sample$bwght, sample$visitsbin, mean), xaxt ="n", xlab ="", ylab ="Average infant birthweight",
title(main = "Infant Birthweight by Percentile of Prenatal Visits", line = 0.5, cex.lab = 0.7)
axis(1, at =seq(1,5,1), labels = c("0-20","20-40","40-60","60-80","80-100"),xlab = "Percentile of prenat
```



As it relates to the Apgar scores, we didn't find through exploratory data analysis a reason to believe that they might be highly correlated with number of prenatal visits or month prenatal care began.

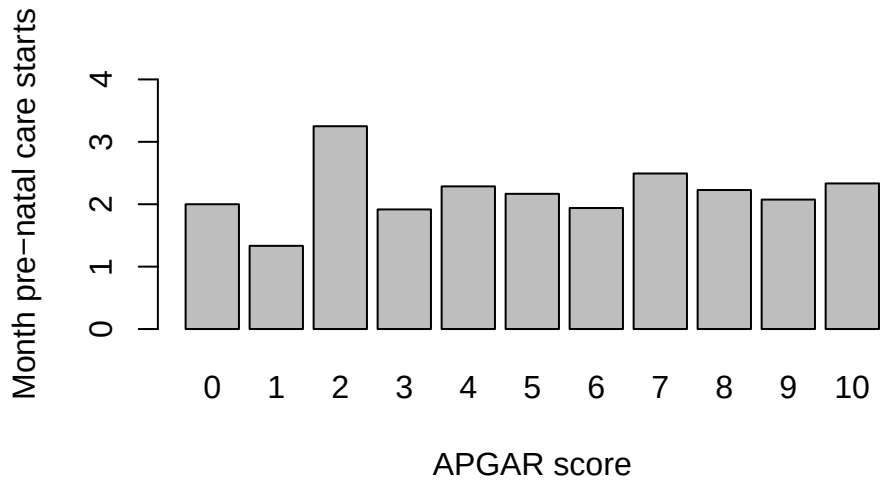
```r
# looking at omaps vs # of visits - no strong trend
avg_visits<- sapply(split(sample,sample$omaps), function(x) {
  colMeans(x["npvis"],na.rm=TRUE)
})
par(mar=c(8,8,8,5),cex.axis=1,cex.lab=1)
barplot(avg_visits, names.arg= c(0,1,2,3,4,5,6,7,8,9,10), ylim = c(0, 20), xlab = "APGAR score", ylab =
```



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

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

```r
# 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
```



Model 1

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.

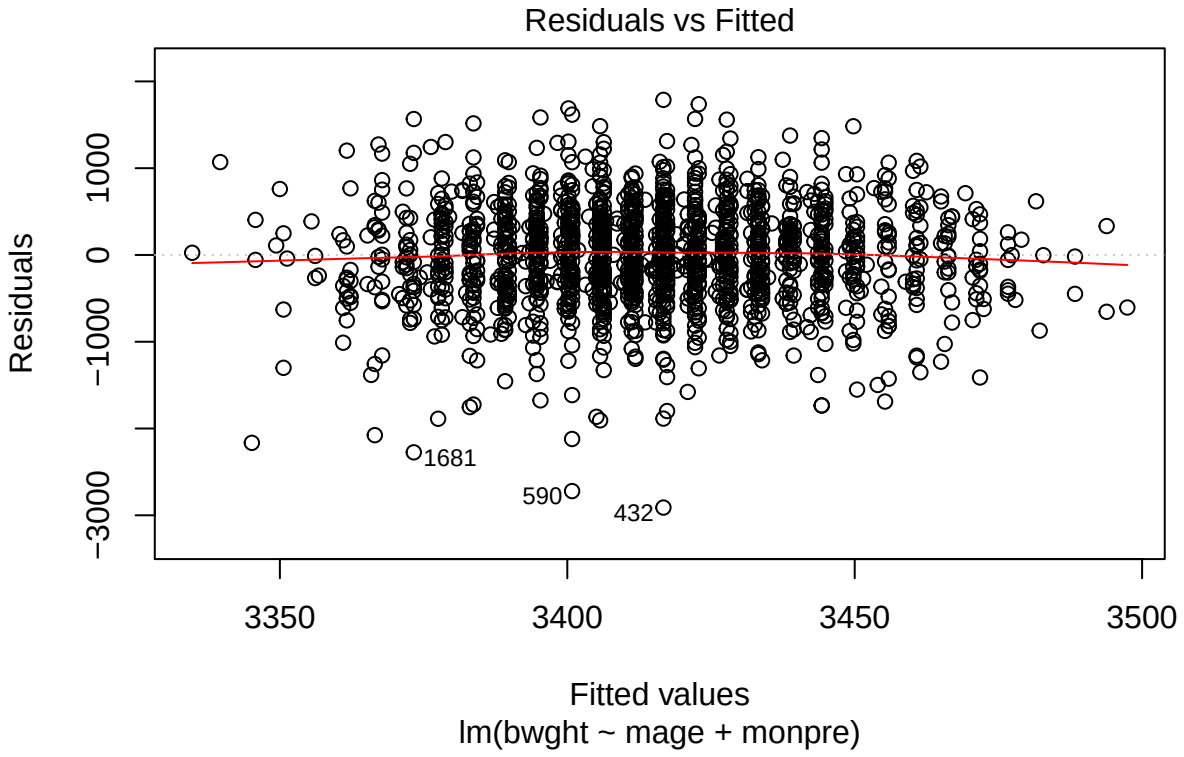
This study begins its analysis by characterizing a simple foundational model upon which deeper analysis can then be performed: $bwght = \beta_0 + \beta_1 mage + \beta_2 monpre + u$

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

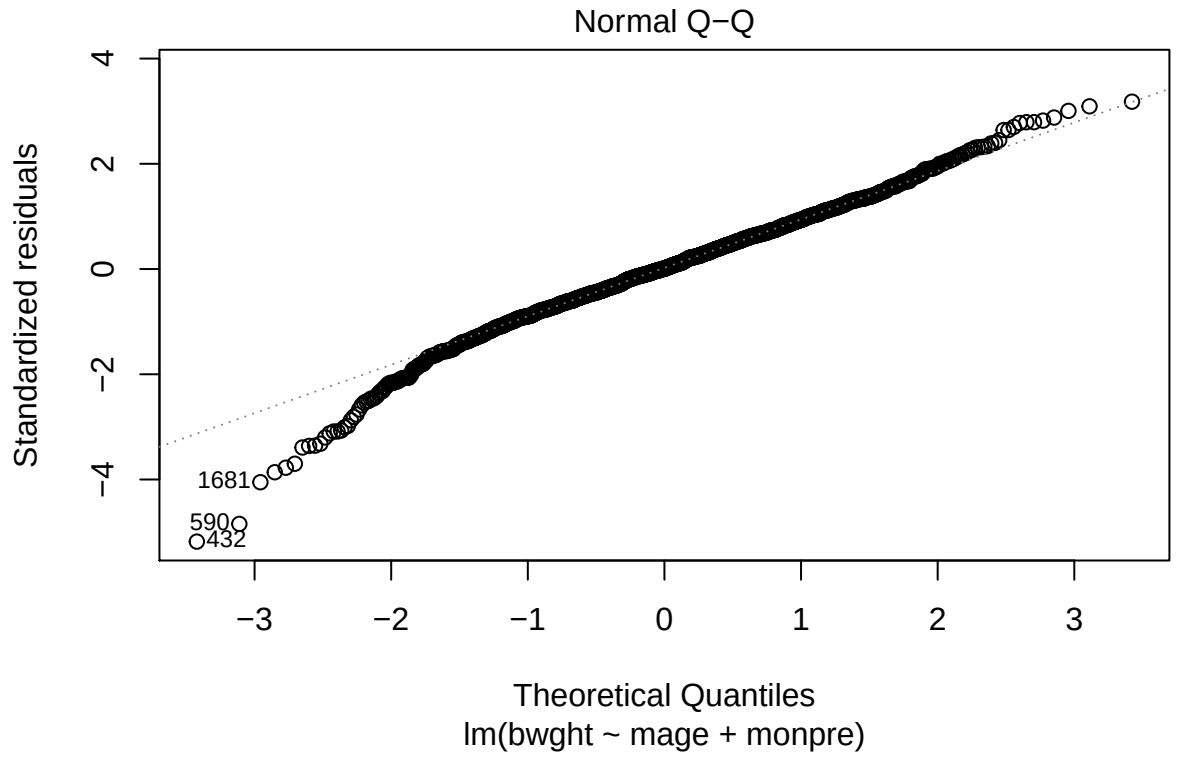
```
##
## Call:
## lm(formula = bwght ~ mage + monpre, data = sample)
##
```

```
## 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
```

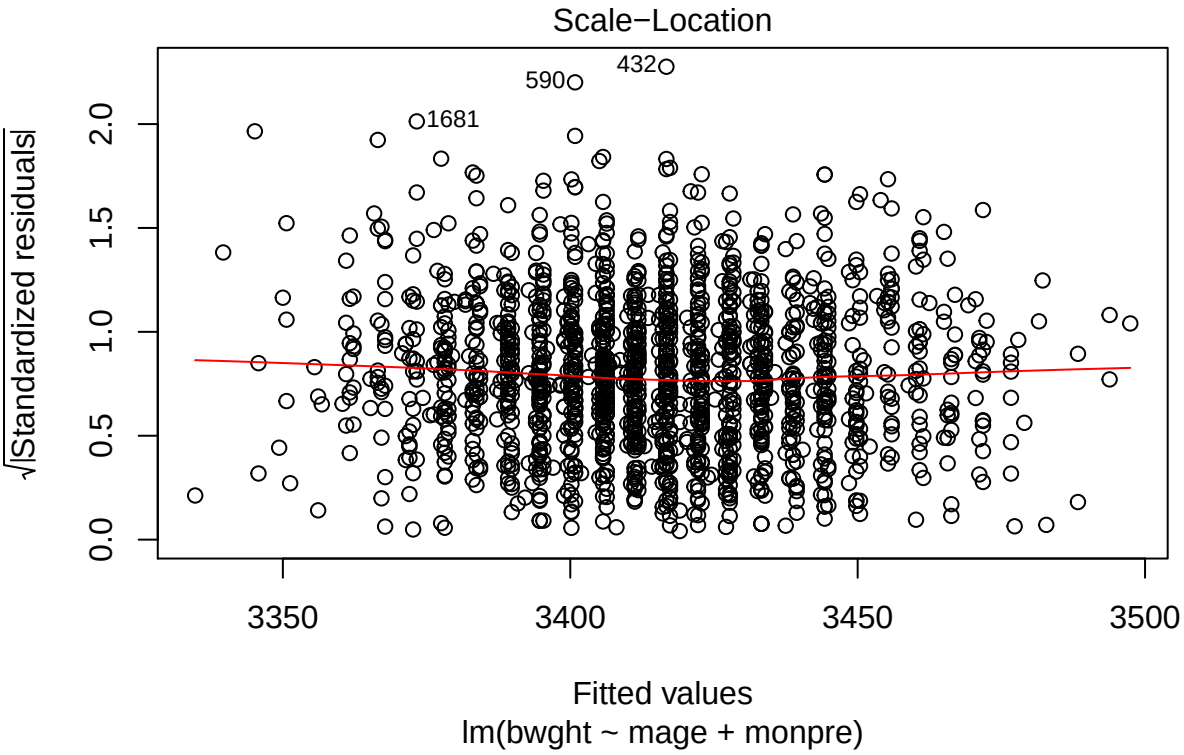
```
plot(model_1, which = 1)
```



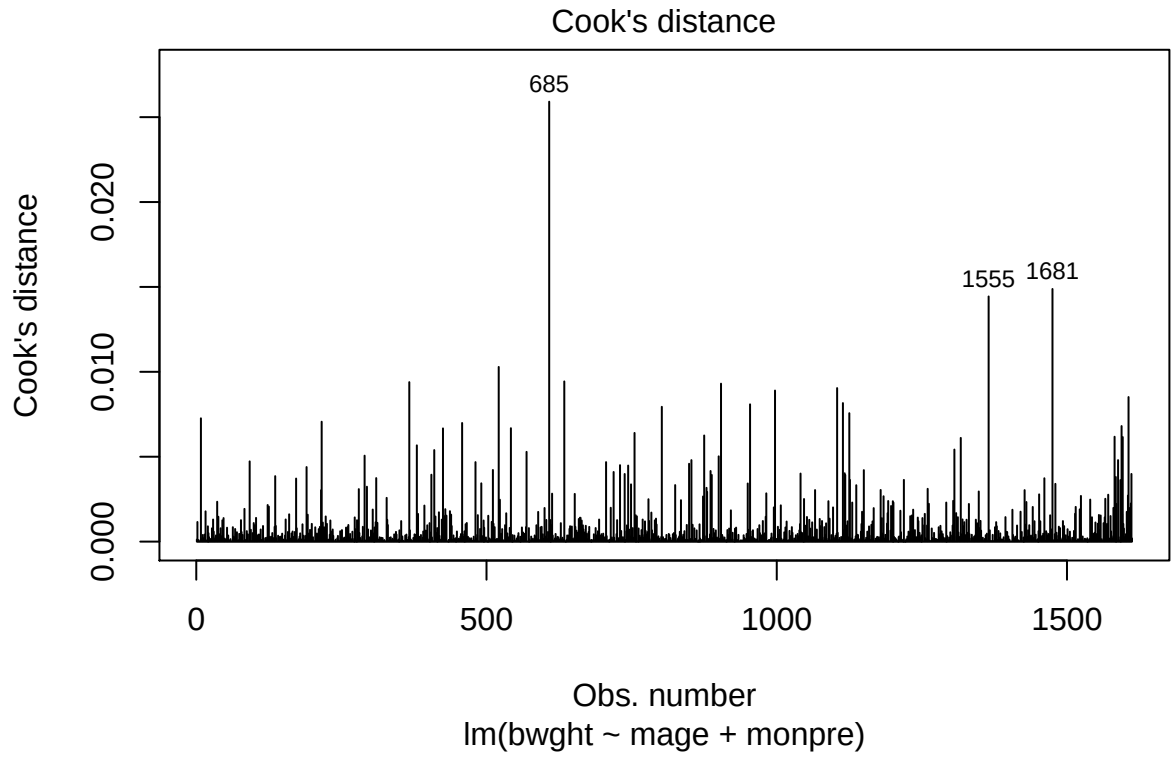

```
plot(model_1, which = 2)
```




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

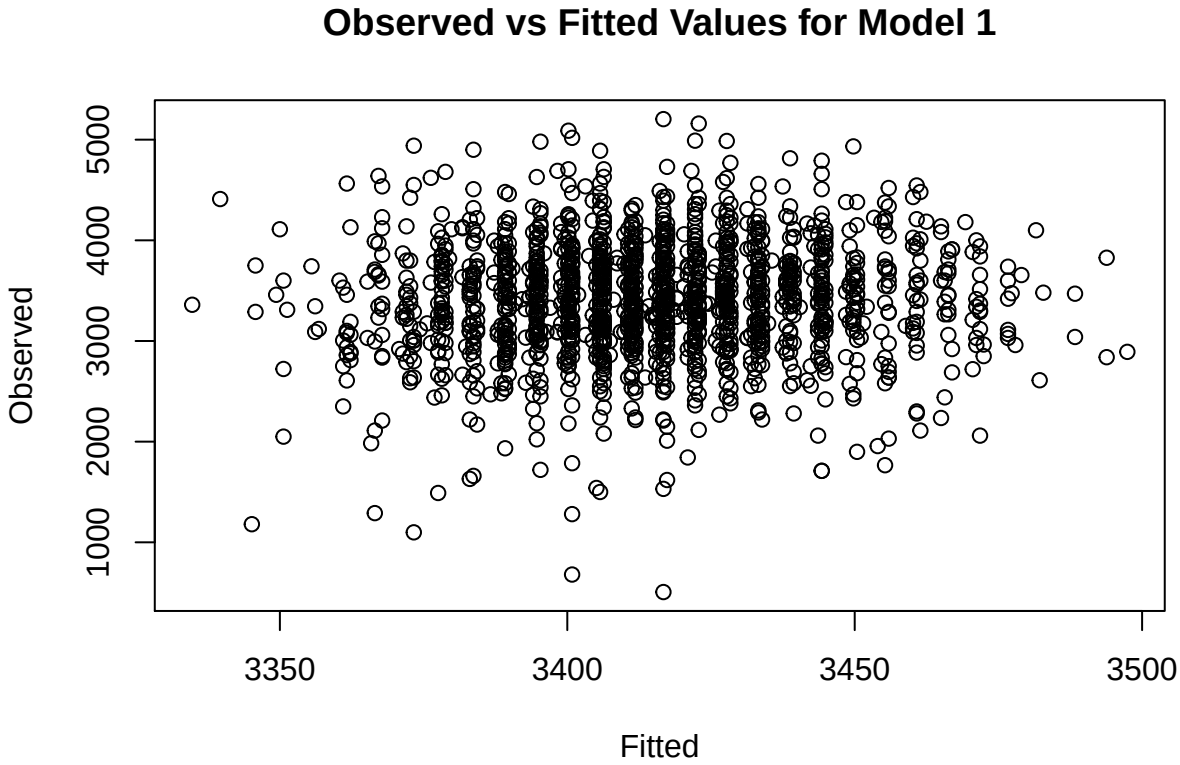



```
plot(model_1, which = 4)
```



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.

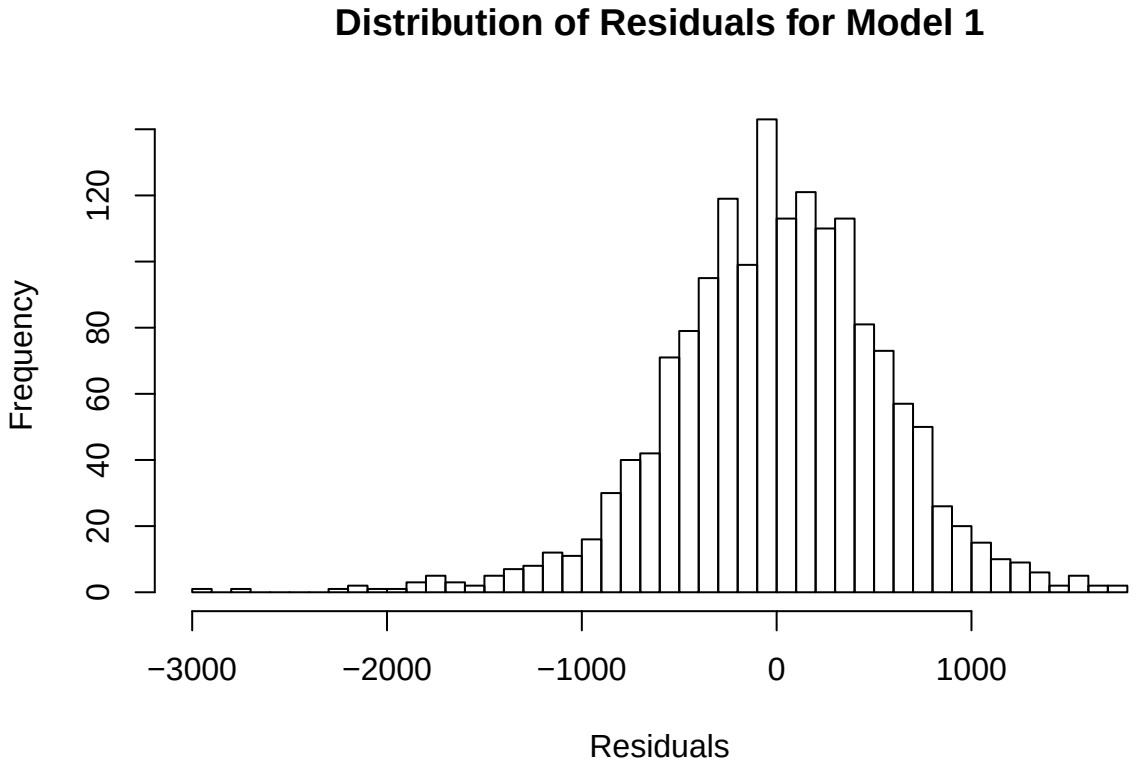
```
bptest(model_1)
```

```
##
##  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

```
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 = 0.9844, 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 'male' variable in model 1 and built on from there with these additions and transformations:

- **Prenatal care:** Instead of including 'monpre' in its current form we transformed the variable to visits per month which we calculated as: npvis/(9 months - monpre). This is in-line with our working theory that starting prenatal care earlier on in the pregnancy and more regular visits to the doctor over the course of the pregnancy is positively correlated with good health outcomes for newborn infants. We saw in the exploratory data analysis that visits per month has a concave, parabolic relationship with birthweight when binned into quintiles. This makes intuitive sense as an increase in visits per month beyond a certain point could be the result of health complications. Further we know that this variable is susceptible to outliers. As a result we made the following additional 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.
- **Baby race:** 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.
- **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)
- **Mother's age:** 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)
```

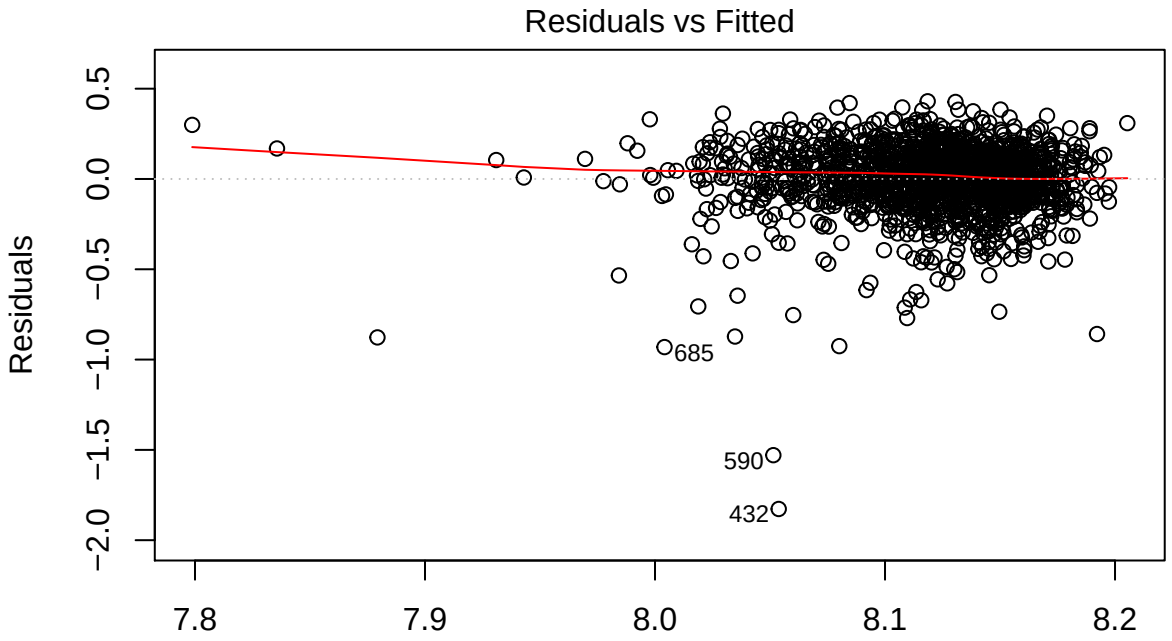
```
model2=  lm(lbwght ~ cigs + logvis_mo + logvis_mo_sq + otherbb+ male + mage + magesq, data = sample)
summary(model2)
```

```
##
## Call:
## lm(formula = lbwght ~ cigs + logvis_mo + logvis_mo_sq + otherbb +
##     male + mage + magesq, data = sample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.82725 -0.08891  0.01970  0.11220  0.43015
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.6114440  0.1301516  58.481  < 2e-16 ***
## cigs         -0.0029819  0.0011235  -2.654 0.008032 **
## logvis_mo     0.1258063  0.0218903   5.747 1.08e-08 ***
## logvis_mo_sq -0.0387198  0.0122877  -3.151 0.001657 **
## otherbb1     -0.0806592  0.0215176  -3.749 0.000184 ***
## male1         0.0286724  0.0091250   3.142 0.001708 **
## mage          0.0291786  0.0087842   3.322 0.000915 ***
## magesq       -0.0004586  0.0001474  -3.110 0.001904 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1825 on 1604 degrees of freedom
## Multiple R-squared:  0.04612,    Adjusted R-squared:  0.04196
## F-statistic: 11.08 on 7 and 1604 DF,  p-value: 9.624e-14
```

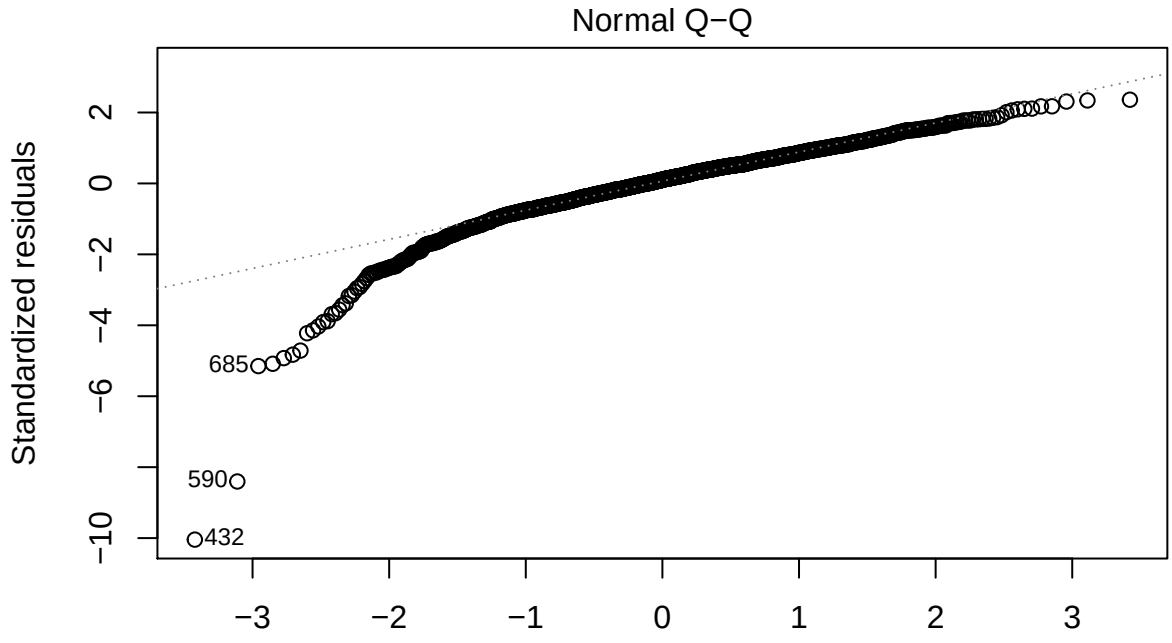
```
vif(model2)
```

```
##         cigs    logvis_mo logvis_mo_sq      otherbb         male
##     1.009451     2.419816     2.418890     1.007113     1.007161
##         mage       magesq
##    83.994334    83.898013
```

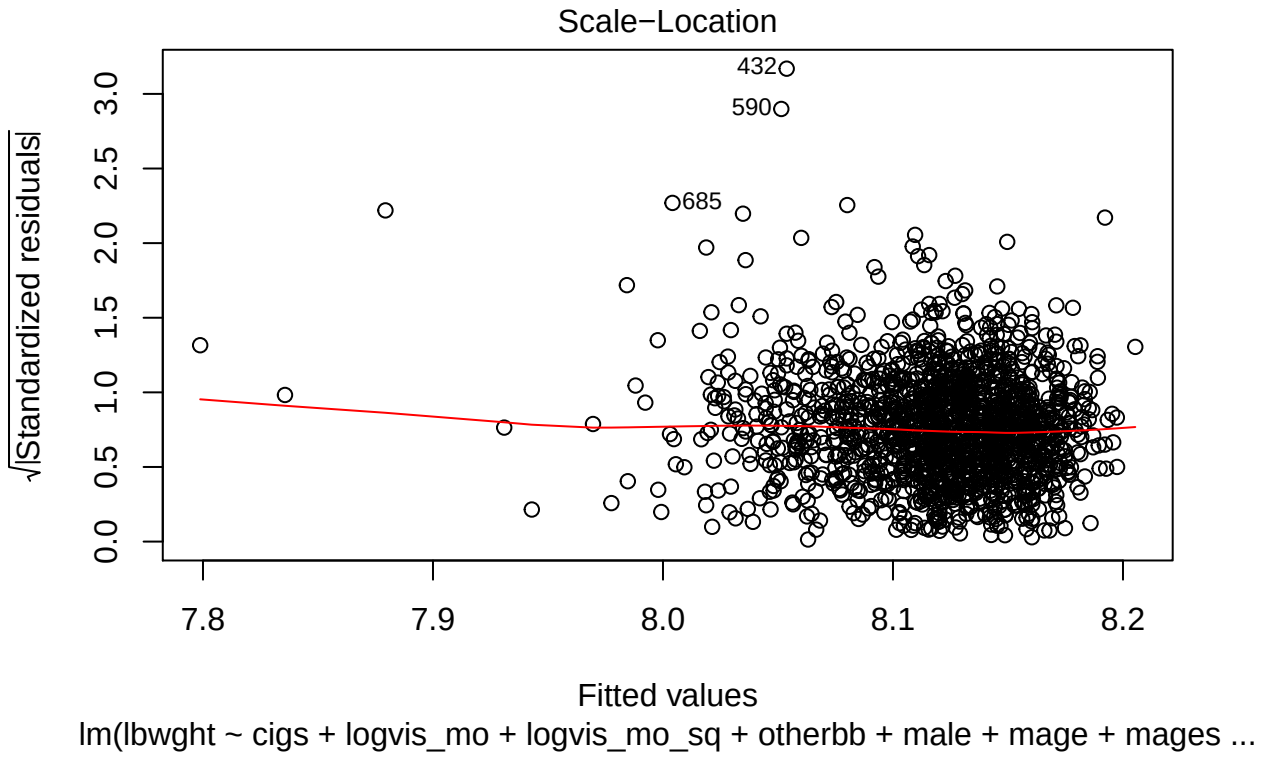
```
plot(model2, which = 1)
```



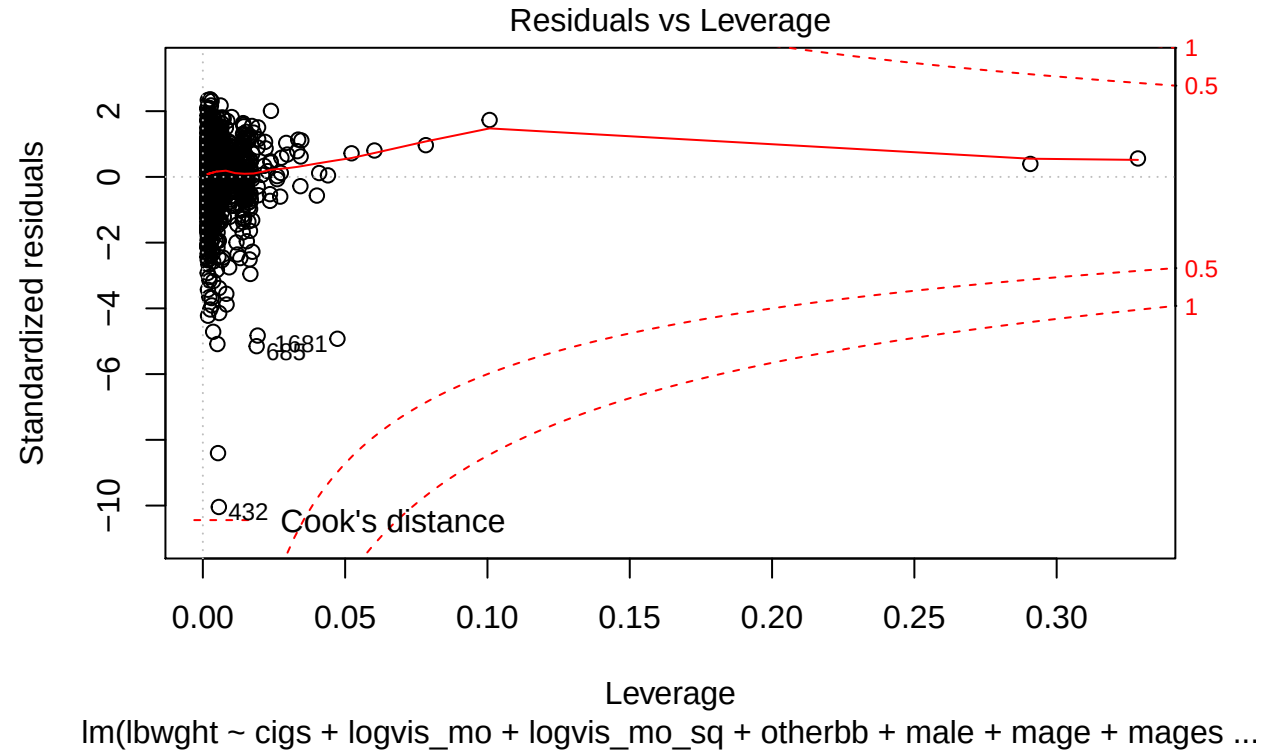

```
plot(model2, which = 2)
```




```
plot(model2, which = 3)
```




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

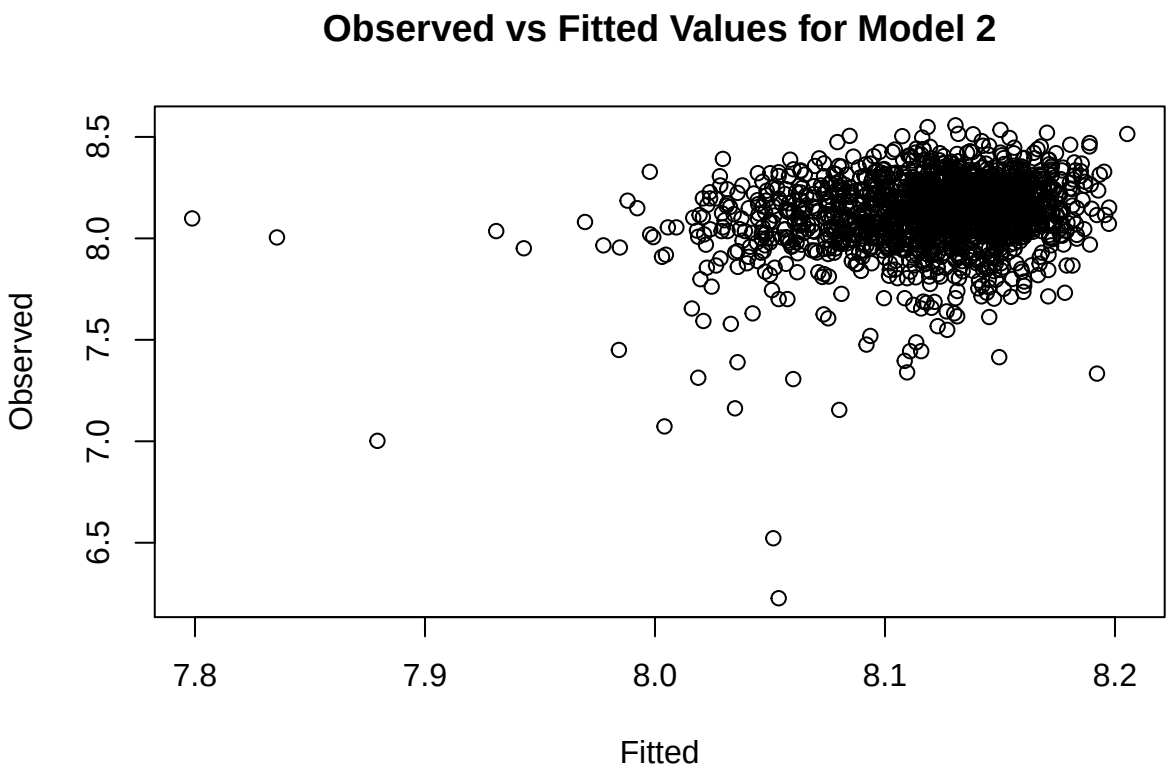


Changes to the 6 Classical Linear Model Assumptions for Model 2:
We see signs of potential bias in Model 2 relative to Model 1 in looking at the diagnostic plots. Linearity and random sampling are still met while there are potential violations of multicollinearity, heteroskedascity, and normal distribution though we are least concerned with the latter given the large sample. 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.

The assumption of a linear population is still met since this model has been defined to be linear in its parameters.

```
plot(model2$fitted.values, (model2$fitted.values + model2$residuals), main = "Observed vs Fitted Values
```



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

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

```
##                       cigs   logvis_mo  logvis_mo_sq     otherbb
## cigs          1.0000000000 -0.02260892 -0.0008940864 -0.05497831
## logvis_mo    -0.0226089166  1.00000000  0.7644971751  0.05109556
## logvis_mo_sq -0.0008940864  0.76449718  1.0000000000  0.03051319
## otherbb      -0.0549783143  0.05109556  0.0305131890  1.00000000
## male         -0.0113782175 -0.05154273 -0.0688795938 -0.01778549
## mage         -0.0568818752 -0.03143181 -0.0542498230  0.02888371
## magesq       -0.0515798587 -0.02371570 -0.0484624548  0.02650980
##                      male        mage      magesq
## cigs         -0.01137822 -0.05688188 -0.05157986
## logvis_mo    -0.05154273 -0.03143181 -0.02371570
## logvis_mo_sq -0.06887959 -0.05424982 -0.04846245
## otherbb      -0.01778549  0.02888371  0.02650980
## male          1.00000000 -0.04017634 -0.04045312
## mage         -0.04017634  1.00000000  0.99397772
## magesq       -0.04045312  0.99397772  1.00000000
```

The VIF is low (<<4) for all variables in model 2 except for 'mage' and 'magesq' whose VIF was north of 80.

The correlation between these two variables is extremely high (0.99), which results in a large standard error for their coefficients. This suggests multicollinearity is present in this model. We can correct for this by either dropping the variables in question or increasing the sample size.

To test whether inclusion of the 'magesq' term significantly improves fit, we used an F-test that is generalized for the usual F-test of overall significance but allows for heteroskedascity-robust errors. The F-stat is 7.35 implying a high level of significance in the magesq term. This suggests that dropping the variable would increase residuals (SSR) and worsen fit.

```
model2_rest = lm(lbwght ~ cigs + logvis_mo + logvis_mo_sq + otherbb+ male + mage, data = sample)
linearHypothesis(model2, "magesq = 0", vcov = vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## magesq = 0
##
## Model 1: restricted model
## Model 2: lbwght ~ cigs + logvis_mo + logvis_mo_sq + otherbb + male + mage +
##     magesq
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F   Pr(>F)
## 1   1605
## 2   1604  1 7.3536 0.006764 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

When we compare the residual standard errors of the two models, there is an uptick in the restricted model but only modestly so (from .1825 to .183) which would seem in favor to removing the 'magesq' term.

Finally we highight that both under the restricted and original model 2 that includes the 'magesq' term the interpretation of the coefficients for the main independent variables of interest, 'logvis_mo' and 'logvis_mo_sq', is unchanged. All variables in both models are significant (p-values < 0.05). Having said that, removing the magesq term will result in a downward bias on the variable 'mage' given the negative beta coefficient on 'magesq' which we can see observe in the change in the coefficient of the 'mage' variable (0.002 vs 0.029 in the original model).

```
summary(model2_rest)
```

```
##
## Call:
## lm(formula = lbwght ~ cigs + logvis_mo + logvis_mo_sq + otherbb +
##     male + mage, data = sample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.81926 -0.08918  0.01929  0.11157  0.43682
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.0048614  0.0306989 260.754  < 2e-16 ***
```

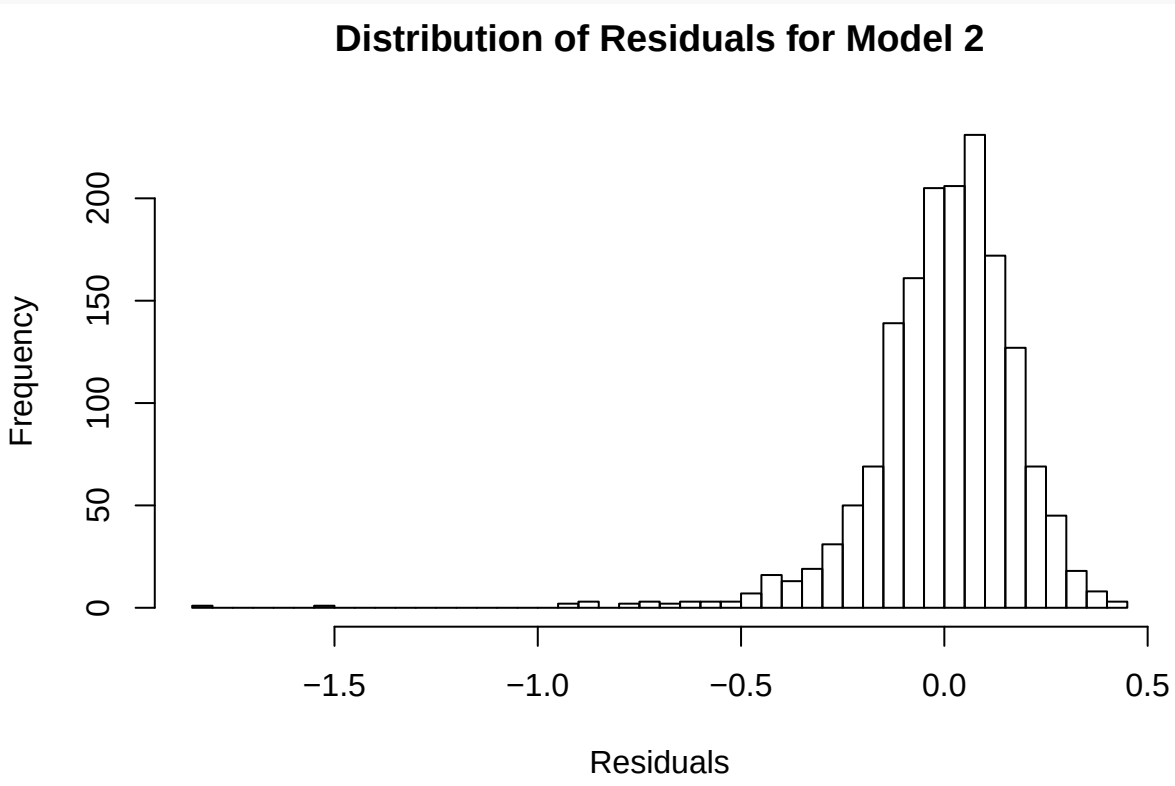
```
## cigs         -0.0031429  0.0011254  -2.793 0.005289 **
## logvis_mo     0.1224082  0.0219220   5.584 2.76e-08 ***
## logvis_mo_sq -0.0385127  0.0123206  -3.126 0.001805 **
## otherbb1     -0.0792265  0.0215707  -3.673 0.000248 ***
## male1         0.0287001  0.0091496   3.137 0.001739 **
## mage          0.0020239  0.0009654   2.096 0.036208 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.183 on 1605 degrees of freedom
## Multiple R-squared:  0.04037,    Adjusted R-squared:  0.03678
## F-statistic: 11.25 on 6 and 1605 DF,  p-value: 2.436e-12
```

```
bptest(model2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model2
## BP = 40.1361, df = 7, p-value = 1.185e-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$residuals, breaks = "fd", main = "Distribution of Residuals for Model 2", xlab = "Residuals
```




```
shapiro.test(model2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model2$residuals
## W = 0.8983, p-value < 2.2e-16
```

As mentioned above, 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.

The Akaike Information Criterion (AIC) for Model 2 is -900.062215. This is much lower than the AIC for Model 1, indicating a substantially improved fit for Model 2.

```
residualsSquared2 = (model2$residuals)^2
model2_unrestricted=  lm(lbwght ~ cigs + logvis_mo + logvis_mo_sq + otherbb+ male + mage + magesq + resi
summary(model2_unrestricted)
```

```
##
## Call:
## lm(formula = lbwght ~ cigs + logvis_mo + logvis_mo_sq + otherbb +
##     male + mage + magesq + residualsSquared2, data = sample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41065 -0.10604 -0.00732  0.09725  0.79074
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        7.7712290  0.1106347  70.242  < 2e-16 ***
## cigs              -0.0030328  0.0009535  -3.181  0.00150 **
## logvis_mo          0.0574185  0.0187771   3.058  0.00227 **
## logvis_mo_sq      -0.0154051  0.0104693  -1.471  0.14136
## otherbb1          -0.0854580  0.0182613  -4.680 3.11e-06 ***
## male1              0.0227299  0.0077473   2.934  0.00339 **
## mage               0.0226278  0.0074591   3.034  0.00246 **
## magesq            -0.0003582  0.0001252  -2.862  0.00427 **
## residualsSquared2 -0.8034644  0.0321572 -24.986  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1548 on 1603 degrees of freedom
## Multiple R-squared:  0.3135, Adjusted R-squared:  0.3101
## F-statistic:  91.5 on 8 and 1603 DF,  p-value: < 2.2e-16
```

```
#should also do secondary test for zero conditional mean assumption
```

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.

Model 3

A problematic model may also arise in the attempt to grapple with the complexities in the given data. Motivations might include:

1. Infant race– following the finding that the 'male' variable was highly statistically significant in Model 2, a sloppy researcher may add an indicator variable, 'female' (=1 if baby is female), in an attempt to increase the model accuracy further.
2. A theory that more educated mothers seek better prenatal care– indicator variables for 1) mothers who only graduated high school ('mHS'), and 2) mothers who graduated from college ('mCollege') can be included in the model to allow for different effects based on education level.
3. Indicator variables that seem to predict low birthweights– 'lbw' (=1 if birth weight <2000 grams) and 'vlbw' (=1 if birth weight <1500 grams).
4. The minor correlation between five-minute Apgar score, 'fmaps', and birthweight, which was observed in the exploratory data analysis.
5. The fact that mothers 35 years and older are considered high-risk and are recommended to seek more frequent prenatal care. This requires a more complex term– an interaction that allows the effect of prenatal care to differ for older mothers. A careless research may include an interaction between an indicator variable representing mothers 35 and older and a metric variable representing the level form of prenatal visits per month, instead of the log form as is currently used. In the following example, this interaction is represented by the variable 'vis_mo_35'.

The model is now presented, its assumptions will be assessed, and a discussion of its problematic variables will follow.

The model: $log(bwght) = \beta_0 + \beta_1 cigs + \beta_2 log(vis_m o) + \beta_3 (log(vis_m o))^2 + \beta_3 vis\_mo\_35 + \beta_4 otherbb + \beta_5 female + \beta_6 male + \beta_7 mage + \beta_8 magesq + \beta_9 lbw + \beta_{10} vlbw + \beta_{11} fmaps + \beta_{12} mHS + \beta_{13} mCollege$

```
fHS = (factor(ifelse(((sample$feduc)>=12)&((sample$feduc<=16)), 1, 0)))
fCollege = (factor(ifelse((sample$feduc>=16), 1, 0)))
mHS = (factor(ifelse((sample$meduc>=12)&(sample$meduc<=16), 1, 0)))
mCollege = (factor(ifelse((sample$meduc>=16), 1, 0)))

#indicator term for mothers 35+
mOver35 = as.numeric(factor(ifelse((sample$mage>=35), 1, 0))) - 1

#interaction-- mothers over 35 are high risk and should have more frequent visits
#vis_mo_35 = mOver35*(sample$logvis_mo)
vis_mo_35 = mOver35*(sample$visits_pr_mo)

#
female = (factor(ifelse((sample$male==0), 1, 0)))
model_3 =  lm(lbwght ~ cigs + logvis_mo + logvis_mo_sq + vis_mo_35 + otherbb + female + male + mage + ma
model_3b =  lm(lbwght ~ cigs + logvis_mo + logvis_mo_sq + vis_mo_35 + otherbb + male + mage + magesq + l

summary(model_3)
```

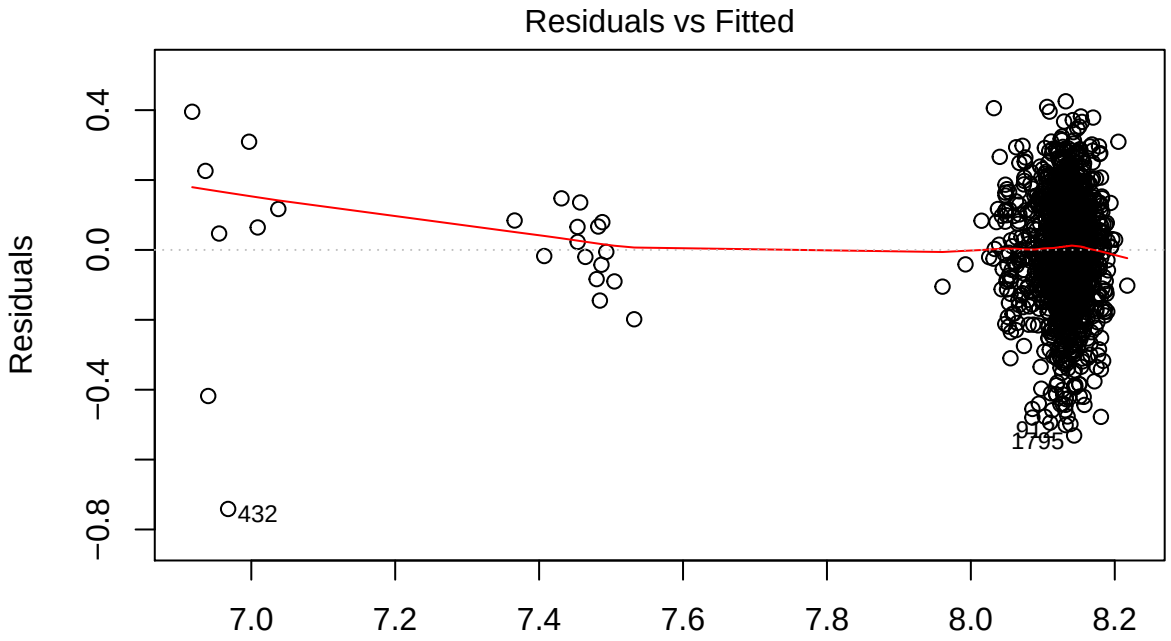
```
##
## Call:
## lm(formula = lbwght ~ cigs + logvis_mo + logvis_mo_sq + vis_mo_35 +
##     otherbb + female + male + mage + magesq + lbw + vlbw + fmaps +
##     mHS + mCollege, data = sample)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.74128 -0.09063  0.00854  0.10262  0.42532
##
## Coefficients: (1 not defined because of singularities)
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.6531167  0.1466785  52.176  < 2e-16 ***
## cigs         -0.0021322  0.0009520  -2.240  0.02525 *
## logvis_mo     0.0440427  0.0188327   2.339  0.01948 *
## logvis_mo_sq -0.0113240  0.0104177  -1.087  0.27720
## vis_mo_35    -0.0025398  0.0092128  -0.276  0.78283
## otherbb1     -0.0831076  0.0183361  -4.532 6.26e-06 ***
## female1      -0.0186227  0.0076760  -2.426  0.01537 *
## male1                NA         NA      NA       NA
## mage          0.0172561  0.0089311   1.932  0.05352 .
## magesq       -0.0002770  0.0001577  -1.756  0.07927 .
## lbw1         -0.6385479  0.0400241 -15.954  < 2e-16 ***
## vlbw1        -0.4192022  0.0694123  -6.039 1.92e-09 ***
## fmaps         0.0262640  0.0084635   3.103  0.00195 **
## mHS1         -0.0278538  0.0128206  -2.173  0.02996 *
## mCollege1     0.0068640  0.0086662   0.792  0.42845
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1531 on 1598 degrees of freedom
## Multiple R-squared:  0.3309, Adjusted R-squared:  0.3255
## F-statistic: 60.79 on 13 and 1598 DF,  p-value: < 2.2e-16
```

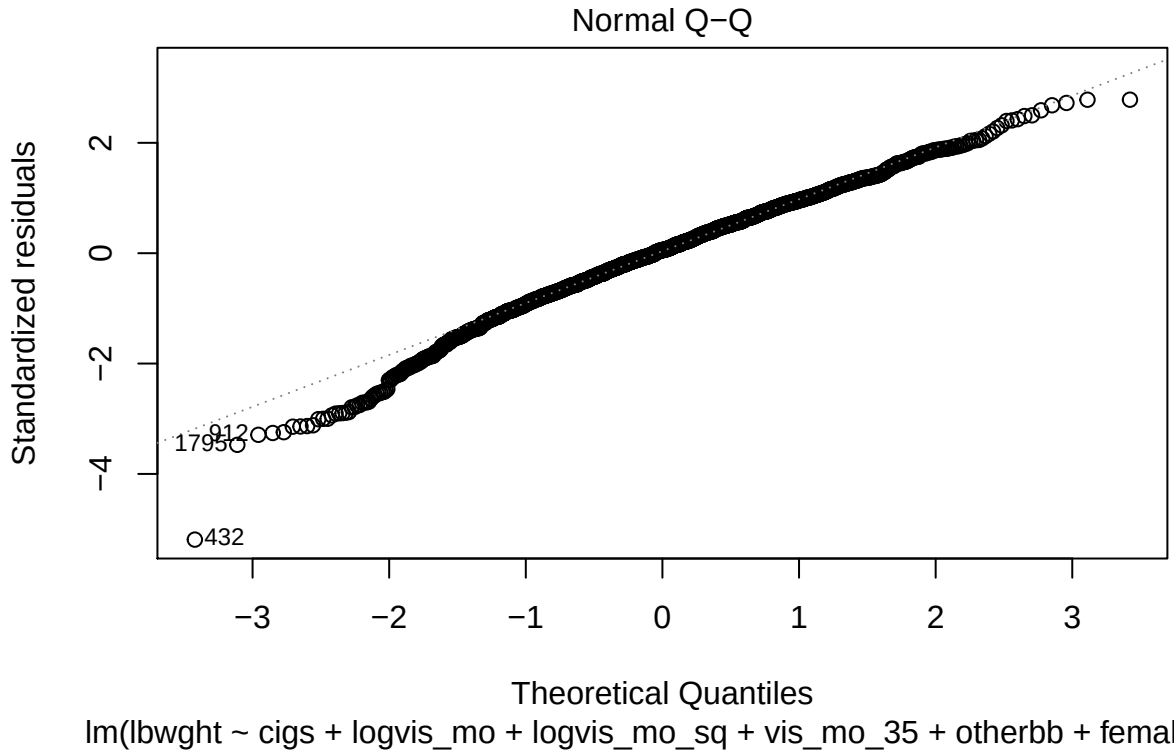
Changes to the 6 Classical Linear Model Assumptions for Model 3

Diagnostic Plots for Model 3:

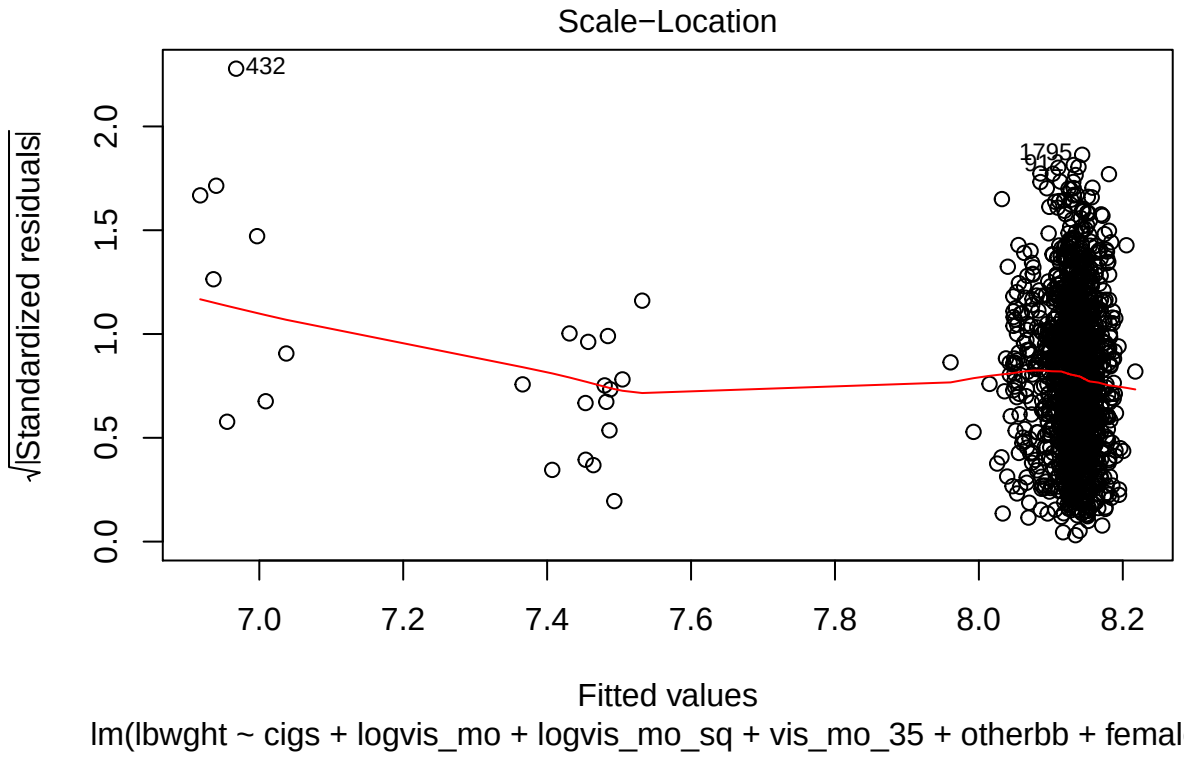
```
plot(model_3, which = 1)
```



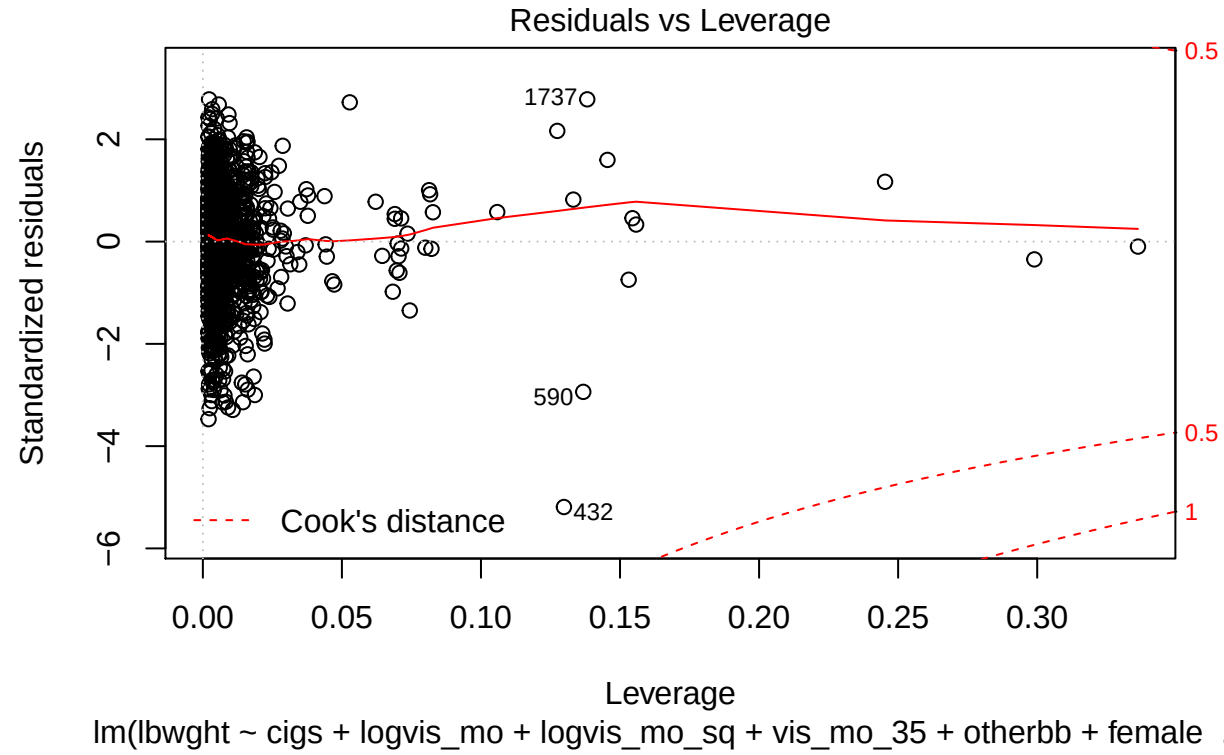

```
plot(model_3, which = 2)
```




```
plot(model_3, which = 3)
```

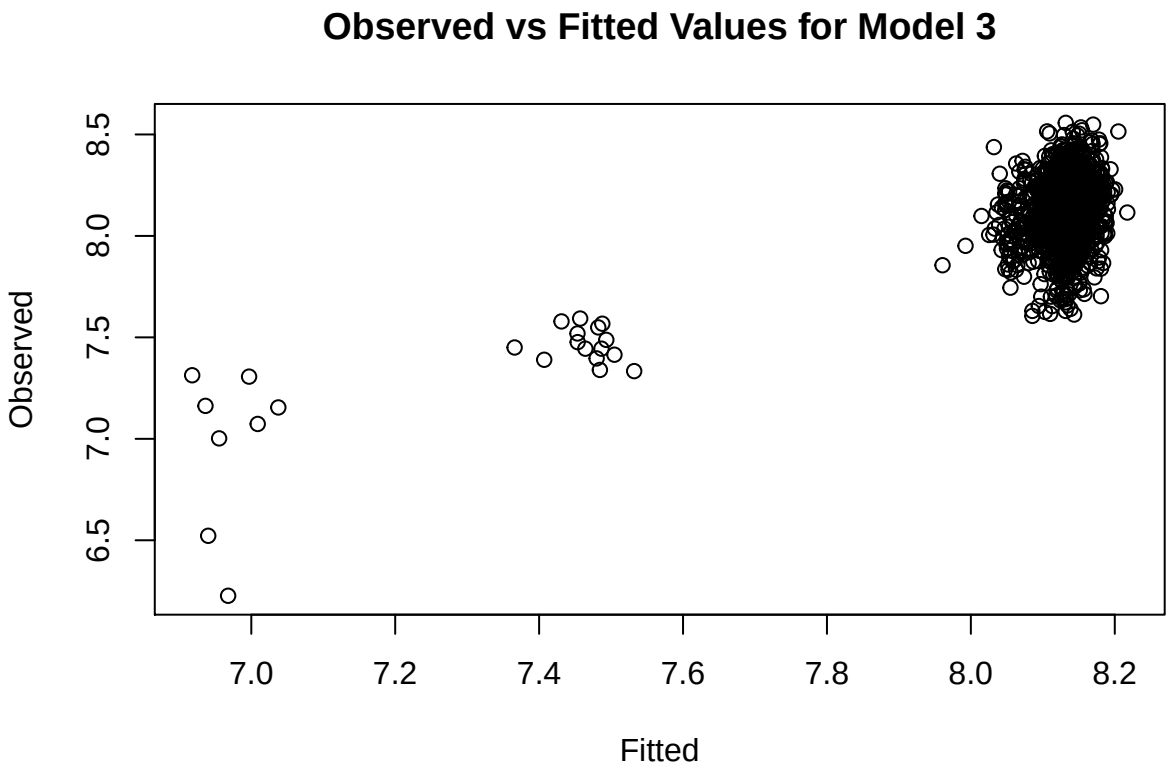



```
plot(model_3, which = 5)
```



The assumption of a linear population is still met since this model has been defined to be linear in its parameters.

```
plot(model_3$fitted.values, (model_3$fitted.values + model_3$residuals), main = "Observed vs Fitted Valu
```



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.

```
vif(model_3b)
```

```
##         cigs    logvis_mo logvis_mo_sq    vis_mo_35      otherbb
##     1.029330     2.543770     2.469408     2.712299     1.038677
##         male         mage       magesq          lbw          vlbw
##     1.012213   123.316986   136.359060     1.549373     1.636168
##        fmaps          mHS     mCollege
##     1.141786     1.049684     1.177655
```

The VIF shown is for a variation of model 3 that excludes the 'female' indicator variable since R is unable to perform a VIF calculation due to the perfect multicollinearity between 'female' and 'male', which will be discussed later. The VIF is still low (<<4) for all variables in model 3 except for 'mage' and 'magesq', as observed in model 2. Please refer to the discussion from model 2 for an explanation surrounding multicollinearity in the 'mage' and 'magesq' variables.

The residuals vs fitted plot for Model 3 is difficult to interpret since data is clustered in three areas. The curve tracking conditional mean shows some curvature, especially at lower fitted values– it is likely that the zero-conditional mean assumption has been violated for model 3. The width of residual values also seems to change across the fitted values, suggesting heteroskedasticity. The scale-location plot confirms this observation since there is substantial curvature in its smoothing line.

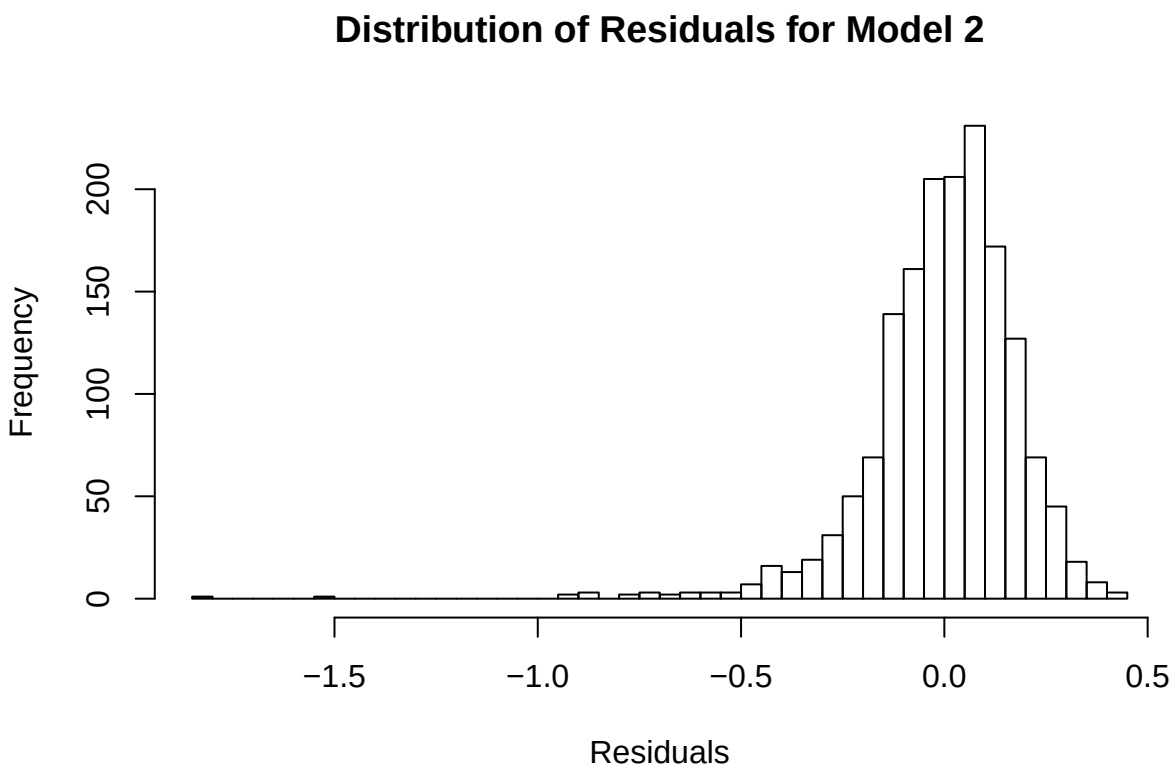
```
bptest(model_3)
```

```
##
##  studentized Breusch-Pagan test
##
```

```
## data:  model_3
## BP = 89.4129, df = 14, p-value = 4.899e-13
```

The Breusch-Pagan test is significant at any level, allowing the rejection of the null hypothesis of constant variance, and further supporting the observation of heteroskedasticity. While the large sample size may have contributed to the low p-value, the combined evidence indicates that the assumption of homoskedasticity is violated for model 3, and that the heteroskedasticity-robust Huber-White standard errors should again be used.

```
hist(model2$residuals, breaks = "fd", main = "Distribution of Residuals for Model 2", xlab = "Residuals
```




```
shapiro.test(model2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model2$residuals
## W = 0.8983, p-value < 2.2e-16
```

The normal q-q plot, histogram of residuals, and highly significant Shapiro-Wilk test for model 3 again demonstrate a departure from normality, but as with the previous models, the large sample size and the Central Limit Theorem again allow Model 3 to rely on OLS asymptotics.

Cook's distance remains small for every observation, so outliers are again of no concern in model 3.

The AIC for Model 3 is -1459.6786959, which is the lowest of the 3 models thus far, indicating a further improvement of fit over the prior 2 models.

```
#really not sure if the RESET should be done here
residualsSquared3 = (model_3$residuals)^2
model_3_unrestricted=  lm(lbwght ~ cigs + logvis_mo + logvis_mo_sq + vis_mo_35 + otherbb + female + mal
summary(model_3_unrestricted)
```

```
##
```

```
## Call:
## lm(formula = lbwght ~ cigs + logvis_mo + logvis_mo_sq + vis_mo_35 +
##     otherbb + female + male + mage + magesq + lbw + vlbw + fmaps +
##     mHS + mCollege + residualsSquared3, data = sample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.38393 -0.10841 -0.01775  0.08901  0.61410
##
## Coefficients: (1 not defined because of singularities)
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         7.783e+00  1.409e-01  55.252  < 2e-16 ***
## cigs               -2.247e-03  9.116e-04  -2.465  0.01380 *
## logvis_mo           4.489e-02  1.803e-02   2.490  0.01289 *
## logvis_mo_sq       -1.378e-02  9.977e-03  -1.381  0.16755
## vis_mo_35           3.468e-06  8.824e-03   0.000  0.99969
## otherbb1           -8.648e-02  1.756e-02  -4.925 9.30e-07 ***
## female1            -2.176e-02  7.354e-03  -2.959  0.00313 **
## male1                      NA         NA      NA       NA
## mage                1.687e-02  8.552e-03   1.973  0.04872 *
## magesq             -2.714e-04  1.510e-04  -1.797  0.07256 .
## lbw1               -6.587e-01  3.836e-02 -17.172  < 2e-16 ***
## vlbw1              -2.944e-01  6.726e-02  -4.376 1.29e-05 ***
## fmaps               1.564e-02  8.152e-03   1.918  0.05524 .
## mHS1               -2.632e-02  1.228e-02  -2.144  0.03221 *
## mCollege1           7.254e-03  8.298e-03   0.874  0.38218
## residualsSquared3 -1.184e+00  9.801e-02 -12.080  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1466 on 1597 degrees of freedom
## Multiple R-squared:  0.3869, Adjusted R-squared:  0.3815
## F-statistic: 71.99 on 14 and 1597 DF,  p-value: < 2.2e-16
#should also do secondary test for zero conditional mean assumption
```

The regression specification error test continues to show that when adding the squared residuals to the model as an independent variable, the coefficient for that term is highly significant, suggesting that model 3 is still misspecified.

The problems with Model 3:

1. The 'Female' indicator variable has perfect collinearity with the 'Male' variable because the two terms add to 1. As a result, statistical software packages like R won't even estimate the model as intended since it is a clear violation of an MLR assumption– this is apparent in the result output of 'NA' for the coefficient on the 'male' variable. This is called a 'dummy variable trap'.
2. There is statistical significance in the indicator variable for mothers with a high school education. While this helps increase the fit of the model, it is very possible that this term absorbs some of the causal effect of prenatal care, since it is likely that mothers with adequate education have better knowledge regarding prenatal care. This will be discussed further in the 'Causal Interpretation' section.
3. Adding two indicator variables for low birthweights is problematic for multiple reasons:

- The two terms 'lbw' and 'vlbw' introduce high collinearity because the two are highly related and overlapping.
- Since lbw and vlbw are calculated from the dependent variable, their use as explanatory variables ruins

the ceteris paribus interpretation of the model coefficients. Together, they set a different intercept for each weight range (0-1500, 1500-2000, and 2000+). It is nonsensical to have separate intercepts for different weight ranges since the model is intended to explain what causes birth weight to go from one range to the next. Furthermore, using independent variables that are calculated from the dependent variable is a violation of the linearity assumption.

- The separation of fitted values into three clusters, as shown in the diagnostic plots, demonstrates the problem of 'lbw' and 'vlbw' resulting in a separate estimation for each weight range, which is incorrect.

Regression summary

Find below a regression table summarizing the results of our 3 models including the coefficients, heteroskedascity-robust standard errors as well as significance (stars represent p-values of 0.05, 0.01, and 0.001 respectively). This shows that almost every single regressor in the later models, 2 and 3, is significant with the exception of 'college.' Having said that, statistical significance is not the same as practical significance.

Practical significance tells us the magnitude of the given effect and whether the observed effect is meaningful. For example interpretation of the 'mage' variable suggests that each 2.9 years of aging for the mother results in a 1% change in infant birthweight. This would seem to be practically significant. On the other hand, we see fmaps as being statistically significant in model 3 but is interpreted as a 2.6% change fmaps score equates to a 1% change in birthweight. It's unclear what the practical interpretation and significance of this is given fmaps scores can only be integers between 0 and 10 for a given newborn infant.

```
##
## ================================================================================
##                                        Dependent variable:
##                          ------------------------------------------------------
##                                  bwght                          lbwght
##                                   (1)                  (2)                (3)
## --------------------------------------------------------------------------------
## cigs                                                -0.003*            -0.002*
##                                                     (0.001)            (0.001)
##
## logvis_mo                                           0.126***            0.044**
##                                                     (0.038)            (0.016)
##
## logvis_mo_sq                                        -0.039*            -0.011
##                                                     (0.017)            (0.006)
##
## vis_mo_35                                                              -0.003
##                                                                        (0.010)
##
## otherbb1                                           -0.081***          -0.083***
##                                                     (0.018)            (0.017)
##
## female1                                                                -0.019*
##                                                                        (0.008)
##
## male1                                               0.029**
##                                                     (0.009)
##
## mage                                 5.508          0.029**             0.017
##                                     (3.201)         (0.010)            (0.009)
```

```
##
## monpre                      4.871
##                           (10.727)
##
## magesq                                      -0.0005**          -0.0003
##                                             (0.0002)          (0.0002)
##
## lbw1                                                          -0.639***
##                                                               (0.028)
##
## vlbw1                                                         -0.419**
##                                                               (0.147)
##
## fmaps                                                          0.026*
##                                                               (0.010)
##
## mHS1                                                          -0.028*
##                                                               (0.013)
##
## mCollege1                                                      0.007
##                                                               (0.009)
##
## Constant                3,241.745***         7.611***          7.653***
##                           (104.704)           (0.158)           (0.159)
##
## ------------------------------------------------------------------------
## Observations                1,612               1,612             1,612
## R2                          0.002               0.046             0.331
## Adjusted R2                 0.001               0.042             0.325
## Residual Std. Error 562.091 (df = 1609) 0.182 (df = 1604) 0.153 (df = 1598)
## ========================================================================
## Note:                                        *p<0.05; **p<0.01; ***p<0.001
```

Causal interpretation

We have been interpreting this regression predictively, ie. for given values of the independent variables what does the fitted model tell us about the baby's birthweight? That said, causal interpretations of regression coefficients in our best-fit model can only be justified by relying on much stricter assumptions than are needed for predictive inference.

Further to be a causal model relies on the assumption that birthweight has causes (which we think it does) but also that our model includes and measures all of the causal variables. Our model does not provide or include:

1. The health of the mother (though we do look at things like cigarettes smoked/drinks consumed which may shed some light). Heart defects, diabetes, kidney disease, and high levels of stress have all been linked to low baby birthweights. We could measure this by including a variable like mother's blood pressure. The omitted variable bias is negative given the beta coefficient is most likely negative as higher levels of mother's blood pressure is likely to be negatively correlated with infant's birthweight while blood pressure is positively correlated with other variables in the model including number of visits per month and cigarettes smoked.

2. Premature births. We do not know at how many weeks the infants were born. Typically babies born ahead of the full 37 week of pregnancy have smaller birthweights than those carried to full term. We would expect the omitted variable bias in this case to be positive (upward) as the beta coefficient is likely positive.

3. More than one baby at a time. Carrying more than 1 baby at a time has been shown to increase stress

on the mother's body but is also typically associated with lower newborn birthweights. Excluding an indicator variable for more than 1 baby at a time likely presents an negative bias on the existing OLS estimates.

If we had a way to measure these terms for our model, we might then be able to assume that our error term is truely random. It is also worth pointing out that as it relates to our existing model and causal interpretation, some of the effect of prenatal care could be hidden by variables like mother's age and the indicator variable for high school and college education we included in Model 3. We know that pregnant women over the age of 35 are labeled as "higher risk" and as such may be asked to come in for doctor's visits earlier and more frequently. Further, mothers who are more educated are more likely to be knowledgeable of the benefits of pre-natal care and opt in. We see that the addition of both of these variables into Model 3 has resulted in better fit but decreased the statistical significance of the pre-natal care variable 'visits_pr_mo'.

Finally in our discussion around causality, it is worth highlighting though that when we plot our independent variables against the model residuals, we find that the residuals appear unrelated to the values of the independent variables suggesting that we may be able to assume exogeneity. That said we know that exogeneity does not imply causality. Exogeneity is about whether OLS can correctly identify the beta coefficients while causality has stricter assumptions and is about whether manipulations to the regressors do not infludence the error term.
Conclusion Our findings support the findings that prenatal care improves certain health outcomes of newborn infants, namely birthweight. Prenatal hospital visits per month was highly statistically significant (<0.001) in our regresional analysis. That said we know from our analysis that baby race and mother's age are also key significant predictors of birthweight.