Lab 4: Does Prenatal Care Improve Infant Health?

Chris Fleisch, Victoria Baker, Frederic Suares

December 13, 2016

Introduction

It is recommended that a new mother go to the doctor on a scheduled number of prenatal visits when she becomes pregnant. There are many contributing factors that may increase the number of visits: if the mother is 35 or older or has a pre-existing health condition, which is not captured in our dataset [1]. Mothers that don't get prenatal care are 3 times more likely to have a baby with low birth weight [1]. Most babies are between 2,500 grams and 4,000 grams, which is considered healthy [2]. Other variations in weight might still be considered normal, but could require extra attention from doctors. This study will attempt to investigate whether prenatal care has an effect on birth weight for newborn infants.

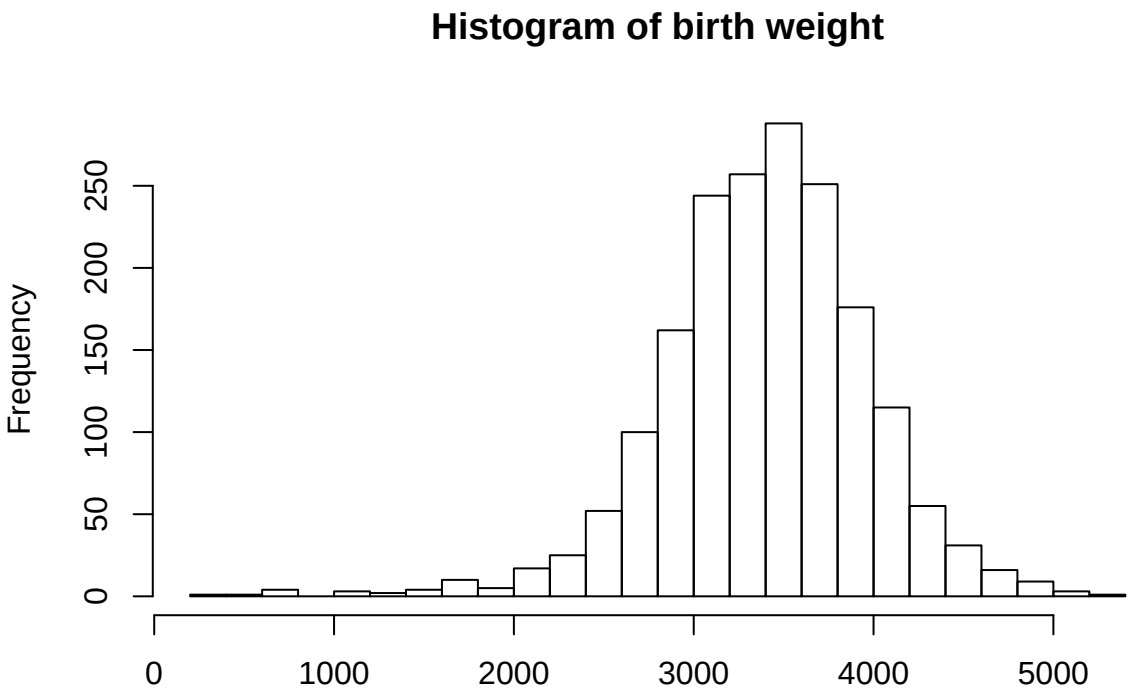
Exploratory Data Analysis

Our exploratory data analysis begins with a look at the focus of this study, birth weight.

```
summary(bdata$bwght)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     360    3076    3425    3401    3770    5204
```

```
hist(bdata$bwght, breaks = 20, main = "Histogram of birth weight", xlab = NULL)
```



Birth weight is relatively normally distributed with some left skew. We will create a new variable for birth weights based on different birth weight classes as described by the descriptions in the data. There are no null values in birth weight.

```
bdata$weight.class <- cut(bdata$bwght, c(0,1500,2000,4000,Inf),
                          labels=c('very low', 'low', 'normal', 'over weight'))
summary(bdata$weight.class)
```

```
##    very low         low      normal over weight
##          13          17        1572         230
```

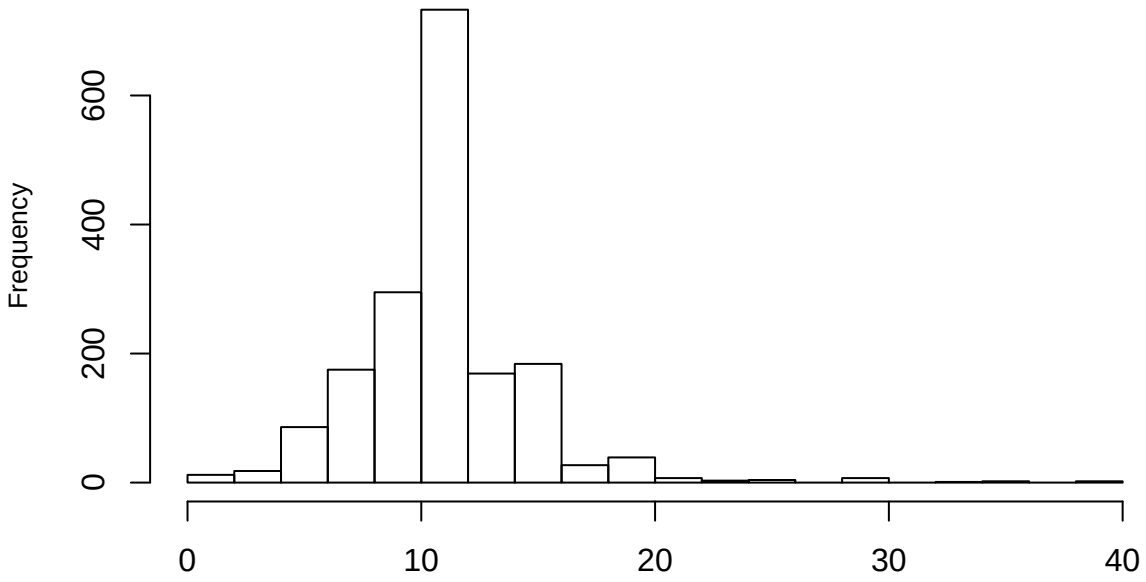
The majority of our data is in the normal weight range. Next, we want to take a look at indicators of infant health. 1 minute and 5 minute APGAR scores do not seem to be good indicators of overall infant health. Further background investigation shows that 1 minute APGAR scores are mainly to determine if the newborn needs help breathing or is having heart trouble. 5 minute APGAR scores lower than 7 are typically caused by other factors not captured with this data, such as difficulties in childbirth or fluid in the baby's airway [3]. Furthermore, most of the time a low 1 minute APGAR score will normalize by the time the 5 minute score is taken. APGAR scores are not meant to be an indication of future health. They test the physical condition of the infant and determine if emergency care is needed. For this reason, it is unlikely we will be able to show strong associations between birth weight, prenatal care, and APGAR tests since they are not good estimators of an infant's health. Birth weight is likely the better outcome variable for this data set.

Let's look at number of prenatal visits.

```
par(mfrow=c(1, 1))
summary(bdata$npvis)
```

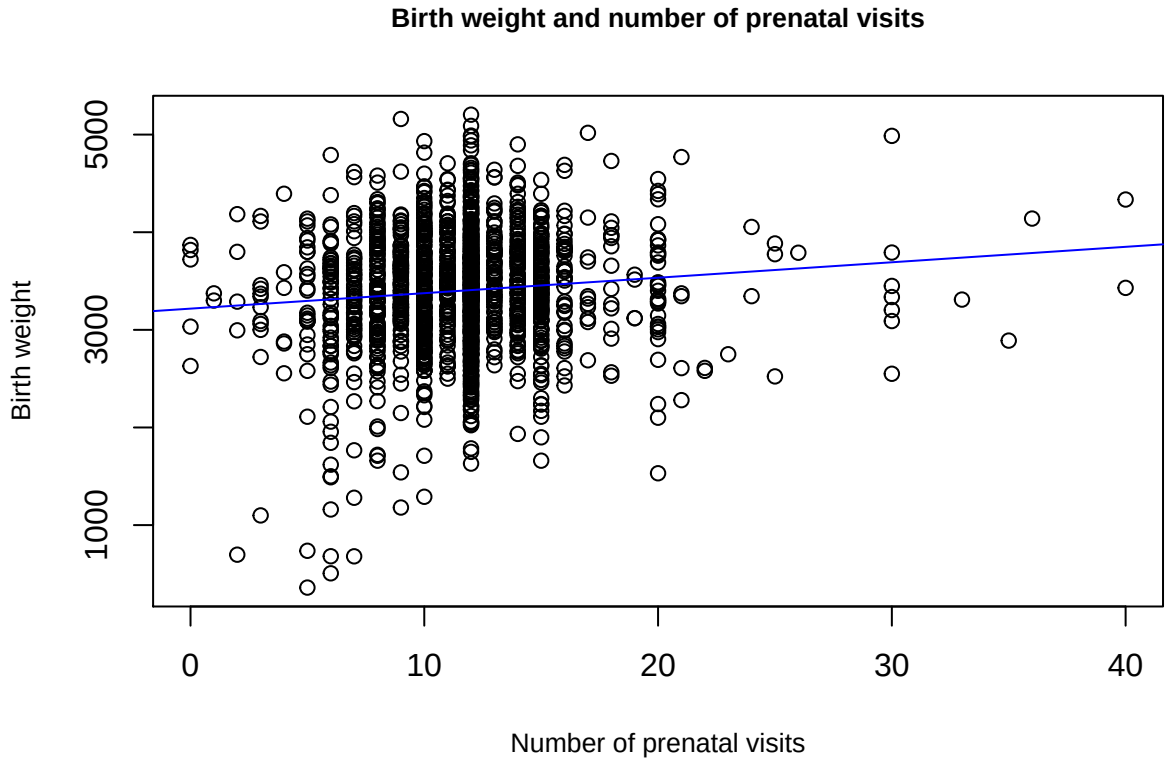
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   10.00   12.00   11.62   13.00   40.00      68
```

```
hist(bdata$npvis, breaks = 20, main = "Histogram of # prenatal visits", cex.main = .8,
     cex.lab = .8, xlab = NULL)
```




```
## # A tibble: 33 × 2
##    npvis total
##    <int> <int>
## 1     12   618
## 2     10   199
## 3     15   143
## 4      8   117
## 5     11   115
## 6     14    97
## 7      9    96
## 8     13    72
## 9     NA    68
## 10     6    59
## # ... with 23 more rows
```

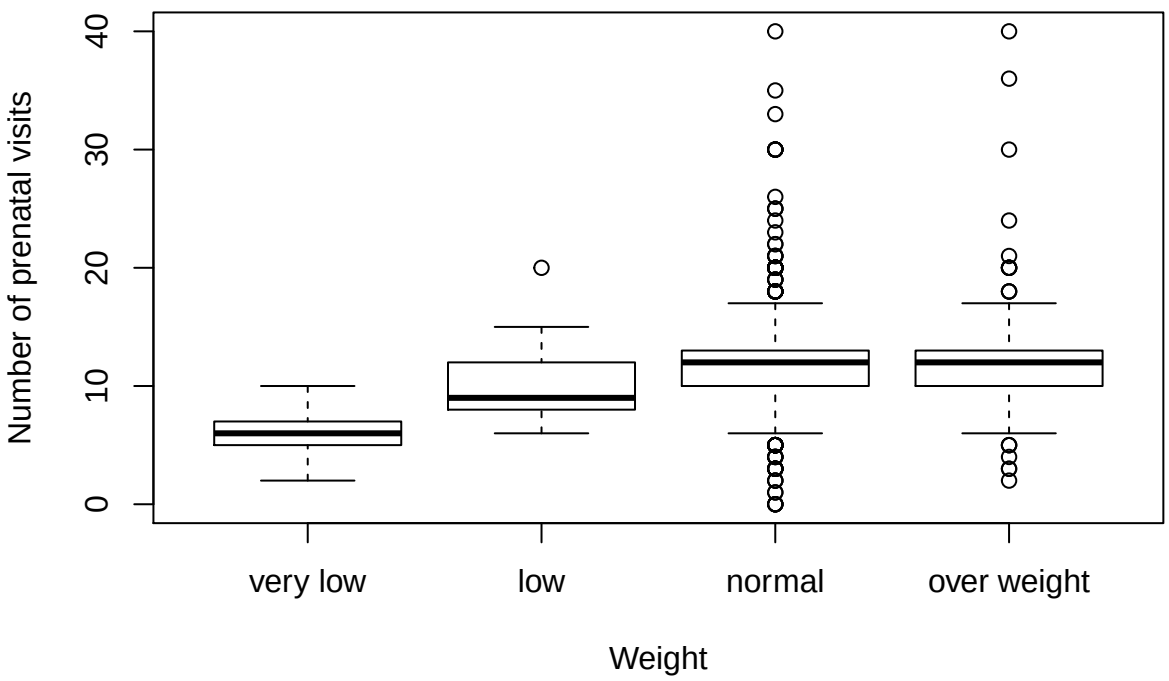
```
plot(bdata$npvis, bdata$bwght, main = "Birth weight and number of prenatal visits",
     xlab = "Number of prenatal visits", ylab = "Birth weight", cex.main = .8,
     cex.lab = .8)
abline(lm(bwght ~ npvis, data = bdata), col = "blue")
```




```
cor(bdata$bwght, bdata$npvis, use = "complete")
```

```
## [1] 0.1003911
```

```
par(mfrow=c(1,1))
boxplot(bdata$npvis ~ bdata$weight.class,
        main = "Boxplot of prenatal visits",
        ylab = "Number of prenatal visits", xlab = "Weight")
```



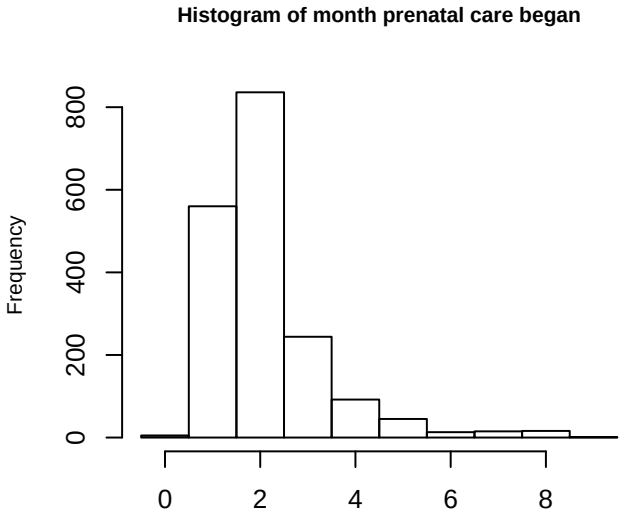
The histogram of prenatal visits shows a large spike around 12 visits and then falls off to each side with some positive skew. The plot shows that the number of visits is associated with an increase in birth weight. The box plot shows that the mean number of prenatal visits increases for each weight class of the infant. There is a positive correlation between number of visits and birth weight. Number of visits will be a good variable to have in our model.

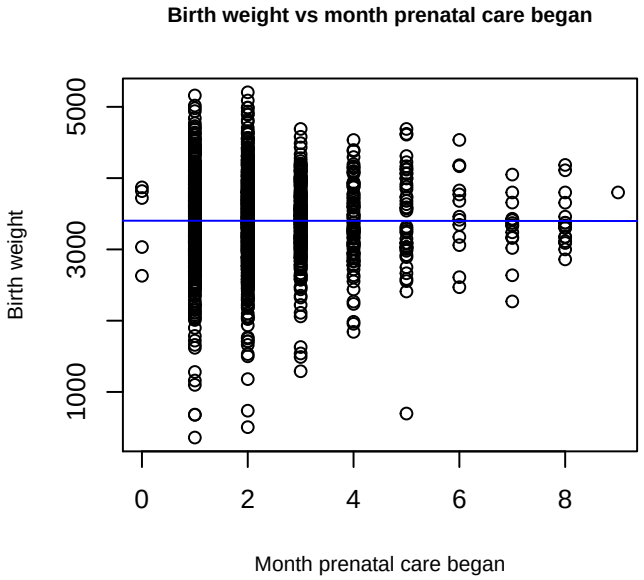
```
par(mfrow=c(2,2))
summary(bdata$monpre)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   1.000   2.000   2.122   2.000   9.000       5
```

```
hist(bdata$monpre, breaks = 0:10-.5, main = "Histogram of month prenatal care began",
     xlab = NULL, cex.main = .8, cex.lab = .8)
plot(bdata$monpre, bdata$bwght, main = "Birth weight vs month prenatal care began",
     ylab = "Birth weight", xlab = "Month prenatal care began", cex.main = .8,
     cex.lab = .8)
abline(lm(bwght ~ monpre, data = bdata), col = "blue")
cor(bdata$bwght, bdata$monpre, use = "complete")
```

```
## [1] -0.001008051
```

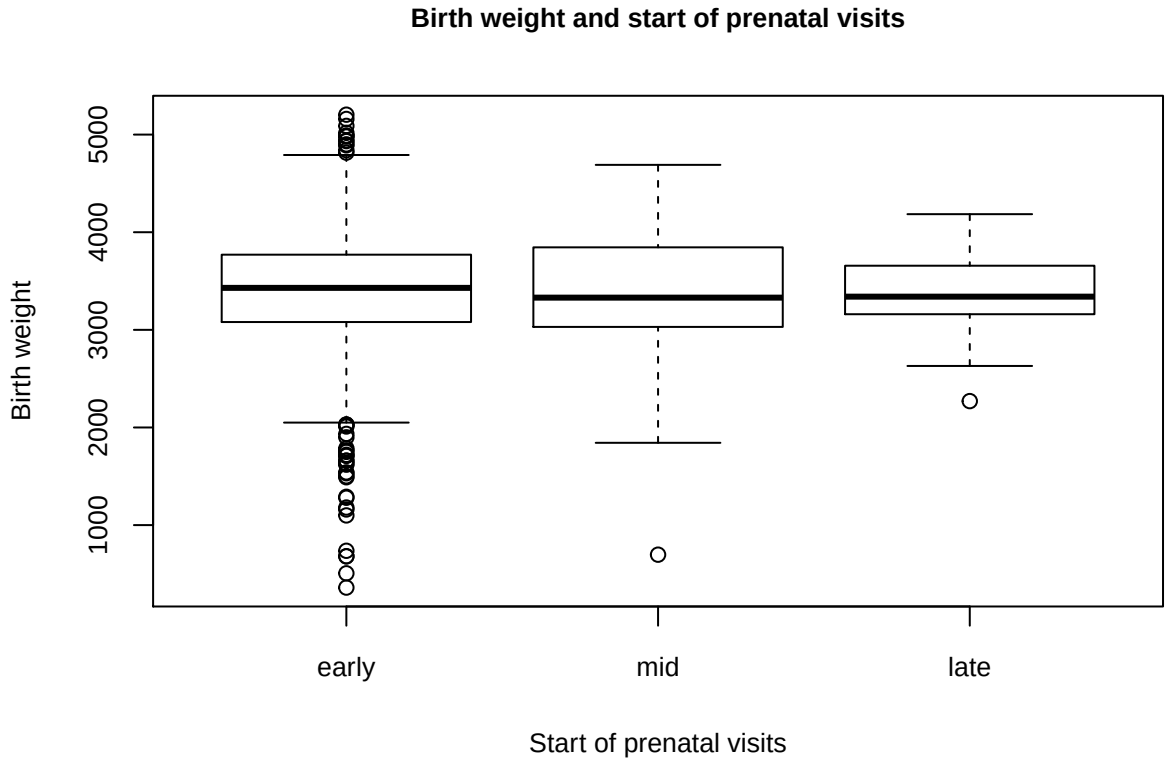





```
par(mfrow=c(1,1))
# early vs late starts
bdata$start.v[bdata$monpre > 0 & bdata$monpre <= 3] <- 'early'
bdata$start.v[bdata$monpre > 3 & bdata$monpre <= 6] <- 'mid'
bdata$start.v[bdata$monpre > 6 | bdata$monpre <= 0] <- 'late'
bdata$start.v <- factor(bdata$start.v, levels = c('early', 'mid', 'late'))
summary(bdata$start.v)
```

```
## early   mid  late  NA's
##  1640   150    37     5
```

```
boxplot(bdata$bwght~bdata$start.v, main = "Birth weight and start of prenatal visits",
        xlab = 'Start of prenatal visits', ylab = 'Birth weight',
        cex.main = .8, cex.lab = .8, cex.axis = .8)
```

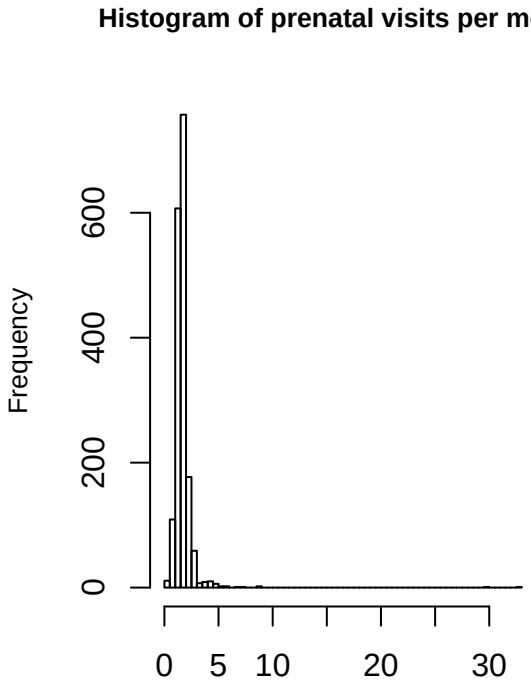


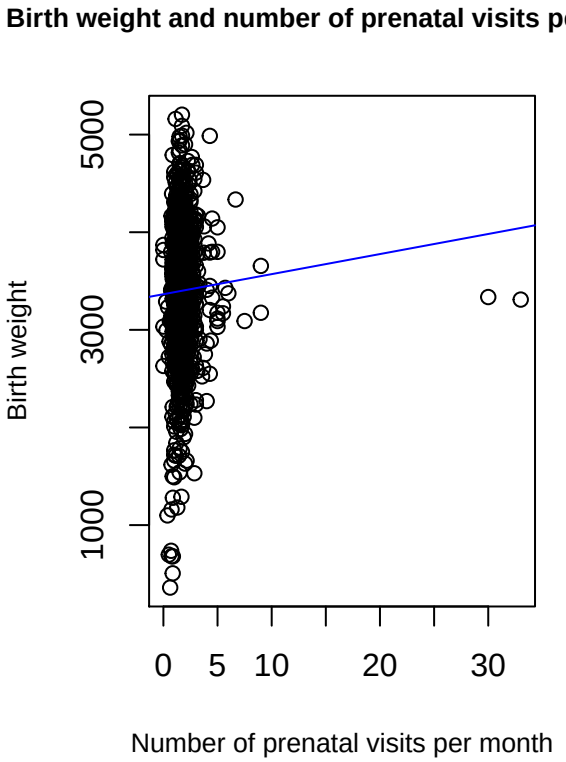
The histogram of month prenatal care began is skewed positive. This is expected because a mother may not know she is pregnant for the first couple months. The month started does not seem to be correlated with birth weight. It has an extremely small correlation value and very little slope on our line. The box plot shows that the average birth weight declines slightly as it goes to the right but on average there is a healthy birth weight for the 3 categories of start time. The ones that went earlier have the most cases of lower birth weights. This suggests that the mothers might have known there was a problem from the beginning and went to prenatal care early to try and help their situation. This doesn't appear to be a good indicator of birth weight and might also have some collinearity with number of prenatal visits.

```
par(mfrow=c(1,2))
# make a new variable visits/per month
bdata$pnvpm <- bdata$npvis/(9 - bdata$monpre)
# remove the one Inf value
bdata$pnvpm[is.infinite(bdata$pnvpm)] <- NA
summary(bdata$pnvpm)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   1.429   1.714   1.766   1.875  33.000      70
```

```
hist(bdata$pnvpm, breaks = 50, main = "Histogram of prenatal visits per month",
     xlab = NULL, cex.main = .8, cex.lab = .8)
plot(bdata$pnvpm, bdata$bwght,
     main = "Birth weight and number of prenatal visits per month",
     cex.main = .8, cex.lab = .8, xlab = "Number of prenatal visits per month",
     ylab = "Birth weight")
abline(lm(bwght ~ pnvpm, data = bdata), col = "blue")
```






```
cor(bdata$bwght, bdata$pnvpm, use = 'complete')
```

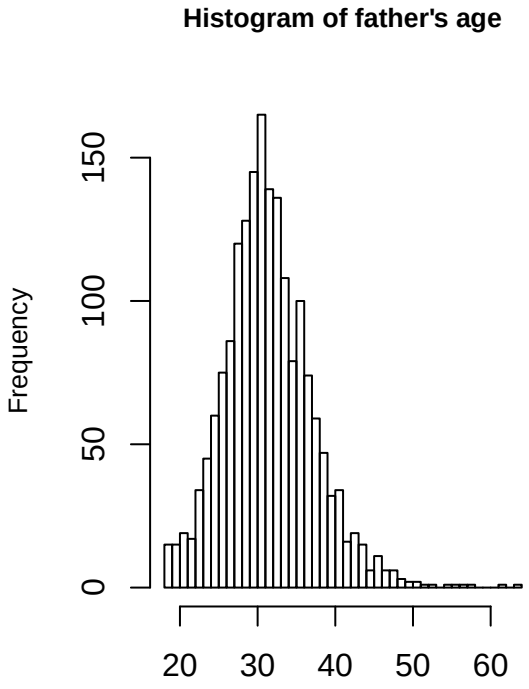
```
## [1] 0.0422944
```

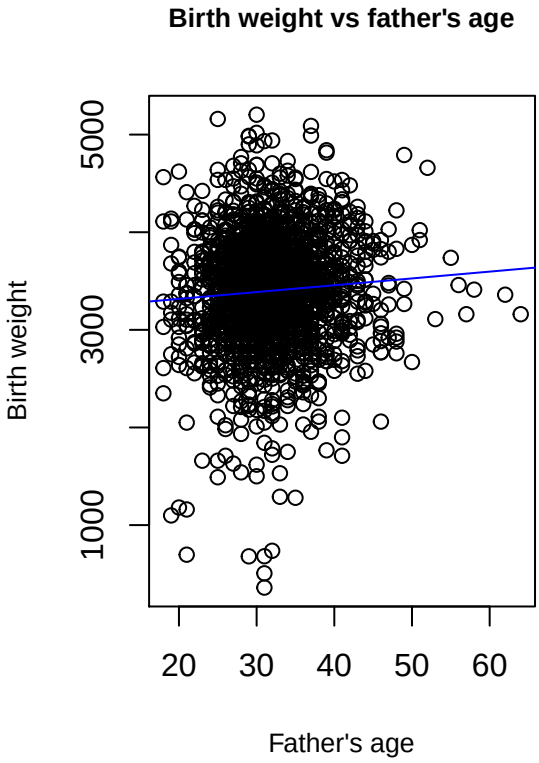
This new variable captures the average number of prenatal visits per month. The recommended schedule is that the number of visits should increase as the pregnancy progresses [4]. This variable attempts to examine the association between number of visits per month and infant health. The histogram shows a concentration around the mean and median with a few extreme outliers. The plot shows a positive slope between birth weight and prenatal visits per month. There's a lot of data concentrated around the mean, so this may not be a good variable to include in the model. It also has multicollinearity with number of prenatal visits and month prenatal care began, which are probably stronger variables to include.

```
par(mfrow=c(1, 2))
summary(bdata$fage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   18.00   28.00   31.00   31.92   35.00   64.00       6
```

```
hist(bdata$fage, breaks = 50, main = "Histogram of father's age", cex.main = .8,
     xlab = NULL, cex.lab = .8)
plot(bdata$fage, bdata$bwght, main = "Birth weight vs father's age",
     xlab = "Father's age", ylab = "Birth weight", cex.main = .8,
     cex.lab = .8)
abline(lm(bwght ~ fage, data = bdata), col = "blue")
```






```
cor(bdata$fage, bdata$bwght, use = "complete")
```

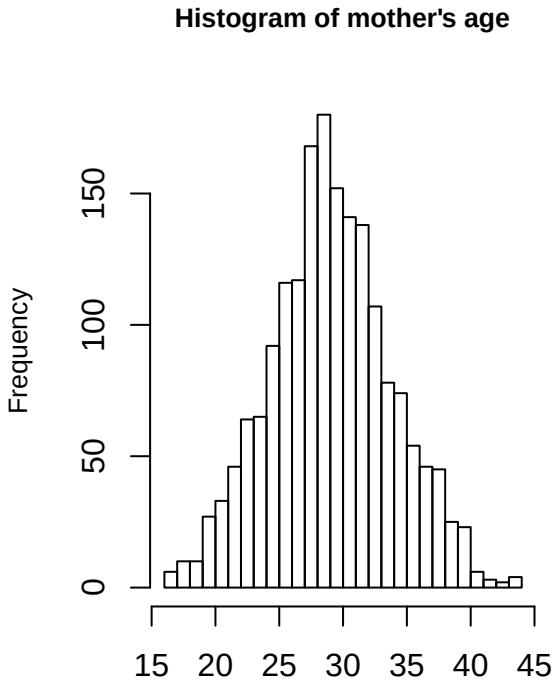
```
## [1] 0.06929068
```

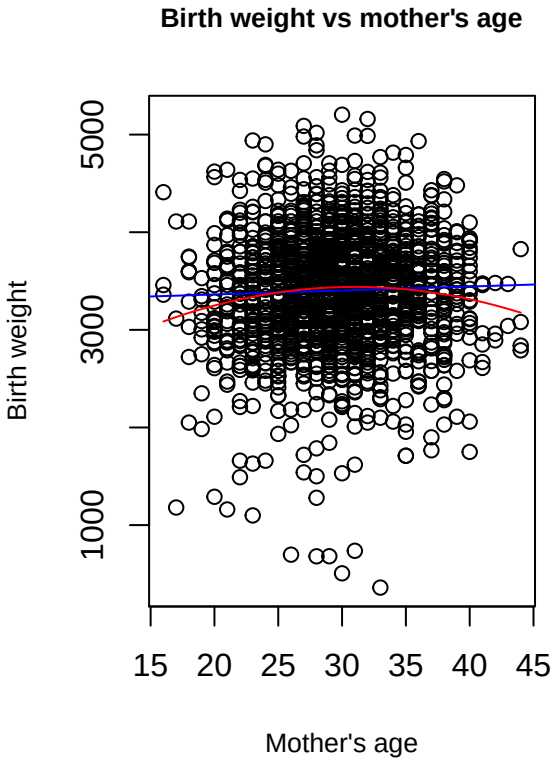
The father's age has a distribution that is approaching normal with some positive skew and attenuation around 18. There is a slight positive correlation between father's age and birth weight. It is unexpected to have any influence on the birth weight, which might make it a weak variable to include. It may also have some multicollinearity with the mother's age.

```
par(mfrow = c(1, 2))
summary(bdata$mage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   16.00   26.00   29.00   29.56   33.00   44.00
```

```
hist(bdata$mage, breaks = 30, main = "Histogram of mother's age", cex.main = .8,
     cex.lab = .8, xlab = NULL)
plot(bdata$mage, bdata$bwght, main = "Birth weight vs mother's age",
     xlab = "Mother's age", ylab = "Birth weight", cex.main = .8,
     cex.lab = .8)
abline(lm(bwght ~ mage, data = bdata), col = "blue")
my.data <- bdata[complete.cases(bdata[, c('bwght', 'mage')]),]
lines(sort(my.data$mage),
      predict(lm(bwght ~ mage + I(mage^2), data = my.data))[order(my.data$mage)],
      col="red")
```





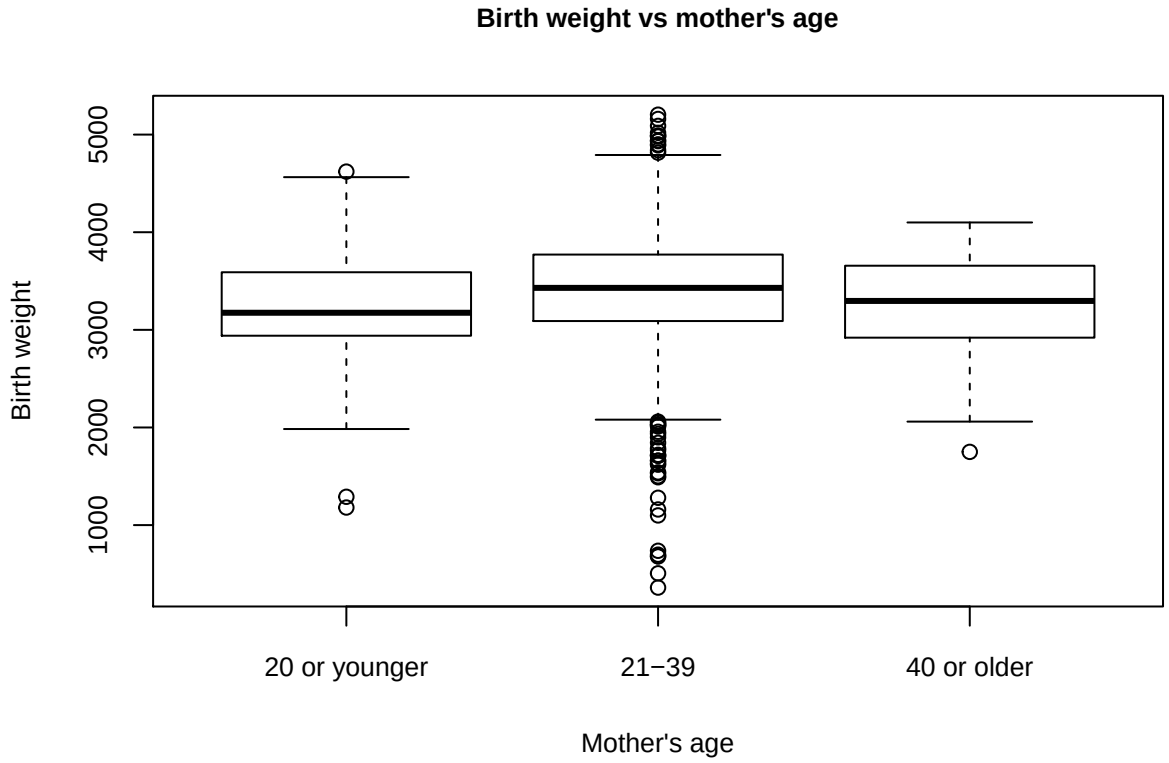

```
cor(bdata$mage, bdata$bwght, use = "complete")
```

```
## [1] 0.03368266
```

```
par(mfrow = c(1, 1))
bdata$mage.class <- cut(bdata$mage, c(0,21,40,Inf),
                        labels=c('20 or younger', '21-39', '40 or older'),
                        right=FALSE)
summary(bdata$mage.class)
```

```
## 20 or younger         21-39   40 or older
##            53          1741            38
```

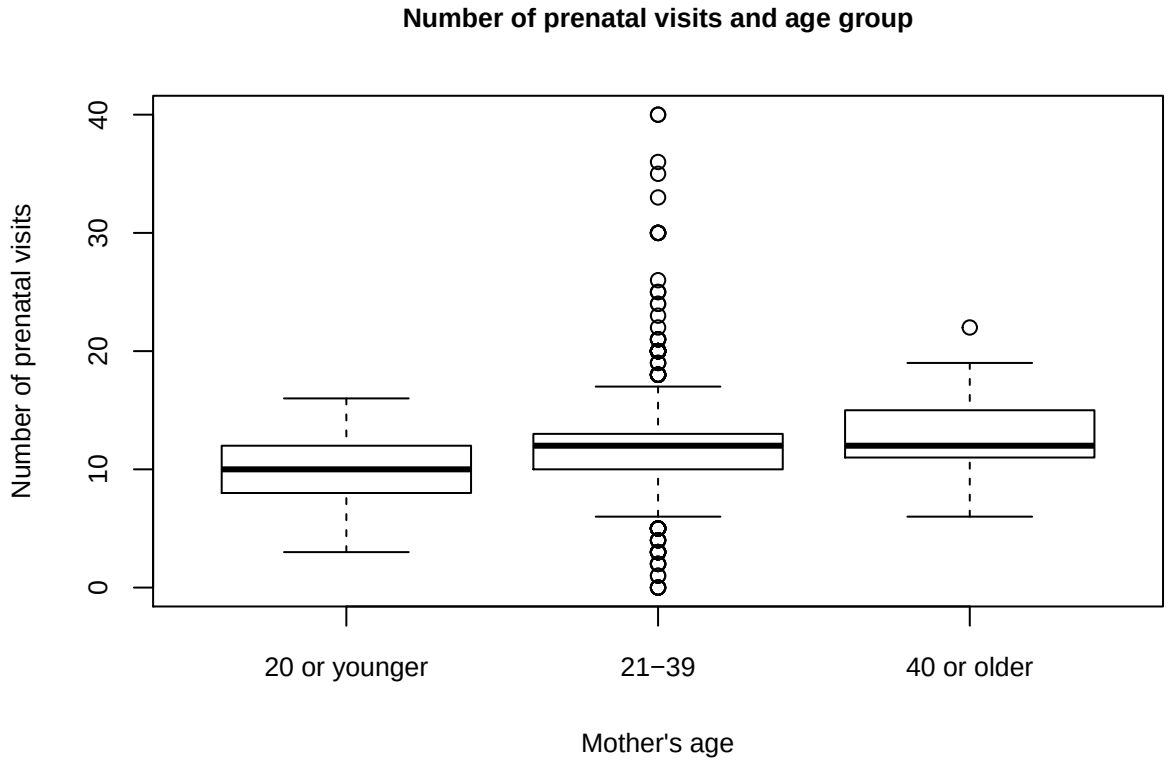
```
plot(bdata$mage.class, bdata$bwght, main = "Birth weight vs mother's age",
     xlab = "Mother's age", ylab = "Birth weight", cex.main = .8,
     cex.lab = .8, cex.axis = .8)
```




```
cor(bdata$mage, bdata$npvis, use = "complete")
```

```
## [1] 0.1020855
```

```
boxplot(bdata$npvis ~ bdata$mage.class, main = "Number of prenatal visits and age group",
        xlab = "Mother's age", ylab = "Number of prenatal visits", cex.main = .8,
        cex.lab = .8, cex.axis = .8)
```



The mother's age has a nice normal distribution. The plot shows a fairly flat line and only a slight positive correlation. When we add a squared term to the mother's age we get a line that is lower at the extremes. This might help us capture differences in birth weight that are associated with younger and older mothers seen in the box plot. Teen mothers and older mothers are more at risk for lower birth weights [6], but the mean ages for birth weights are pretty even across the different weight categories. Mother's age^2 will be a good variable to include our model. Mother's age also shows correlation with the number of visits. This is expected; as earlier stated, older mothers on average go to more prenatal visits and there could be multicollinearity between the mother's age and number of visits.

```
par(mfrow = c(2, 2))
summary(bdata$meduc)
```

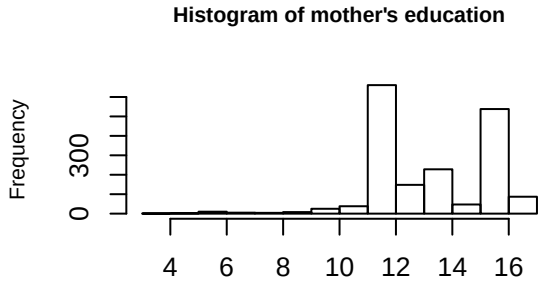
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    3.00   12.00   13.00   13.72   16.00   17.00      30
```

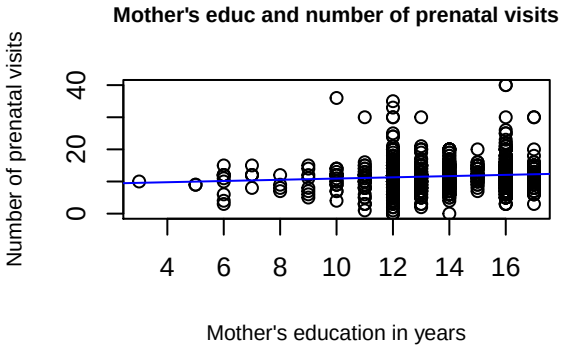
```
hist(bdata$meduc, main = "Histogram of mother's education", xlab = NULL,
     cex.main = .8, cex.lab = .8)
plot(bdata$meduc, bdata$npvis, main = "Mother's educ and number of prenatal visits",
     xlab = "Mother's education in years", ylab = "Number of prenatal visits",
     cex.main = .8, cex.lab = .8)
abline(lm(npvis ~ meduc, data = bdata), col = "blue")
cor(bdata$meduc, bdata$npvis, use = "complete")
```

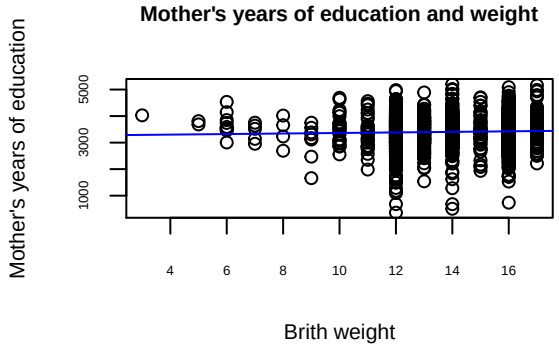
```
## [1] 0.1086247
```

```
plot(bdata$meduc, bdata$bwght, main = "Mother's years of education and weight",
        xlab = "Brith weight", ylab = "Mother's years of education", cex.main = .8,
        cex.lab = .8, cex.axis = .5)
abline(lm(bwght ~ meduc, data = bdata), col = "blue")
cor(bdata$meduc, bdata$bwght, use = "complete")
```

```
## [1] 0.0377613
```





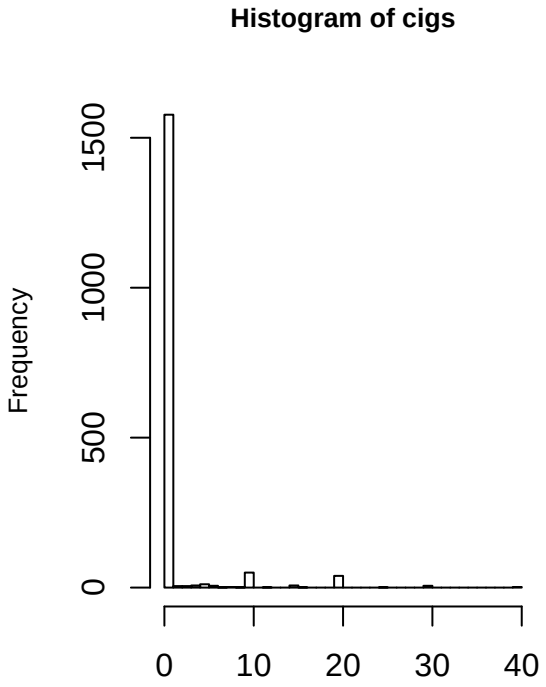


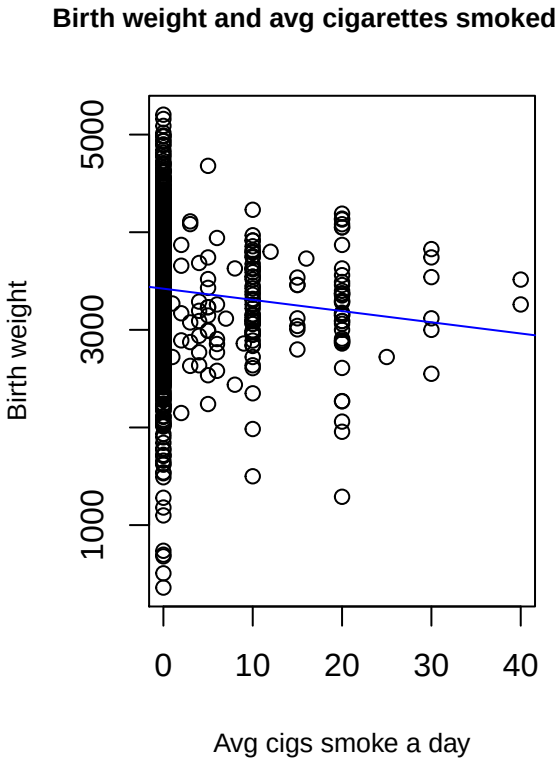
The histogram of mother's education shows two spikes at 12 and 16 years of age. Those are most likely graduation effects. There's not a lot of sample data for lower years of education. The plot shows that the more educated a mother is, the higher the number of prenatal visits. We might want to include mother's education in our model to account for differences in prenatal visits. The plot of mother's education and birth weight does not indicate that there's much correlation in mother's education and birth weight. So, this would not be a good variable to include in our model.

```
par(mfrow = c(1, 2))
summary(bdata$cigs)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   0.000   0.000   1.089   0.000  40.000     110
```

```
hist(bdata$cigs, breaks = 30, cex.main = .8, cex.lab = .8,
     main = "Histogram of cigs", xlab = NULL)
plot(bdata$cigs, bdata$bwght, main = "Birth weight and avg cigarettes smoked a day",
     xlab = "Avg cigs smoke a day", ylab = "Birth weight",
     cex.main = .8, cex.lab = .8)
abline(lm(bwght ~ cigs, data = bdata), col = "blue")
```





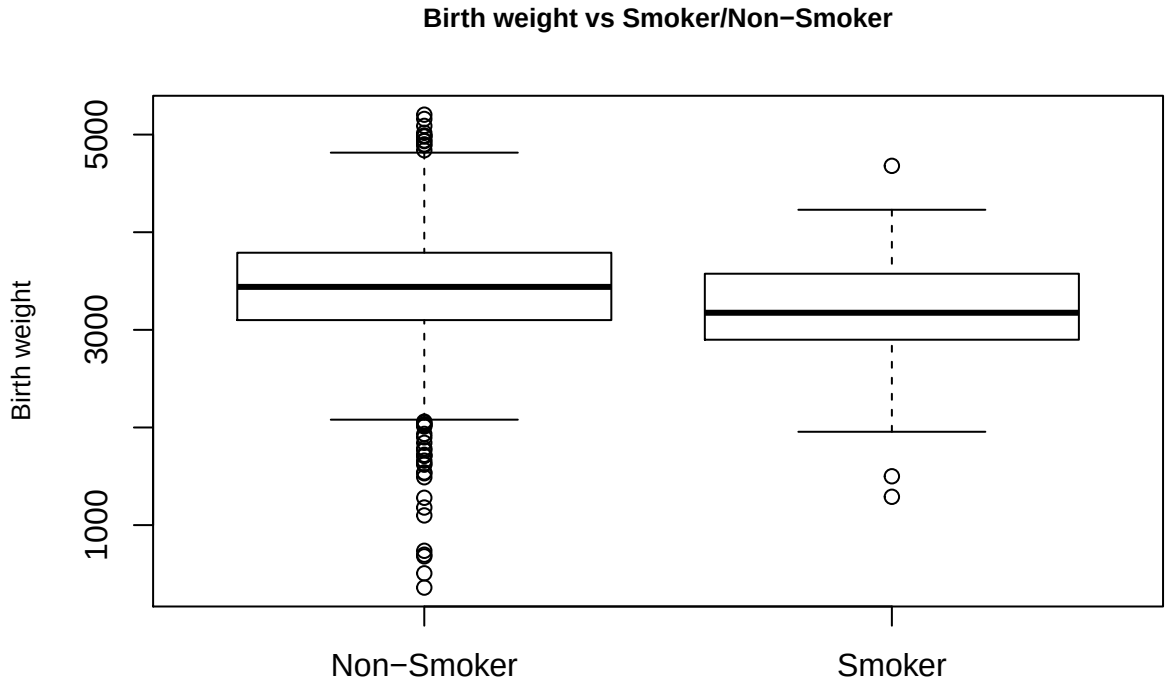

```
cor(bdata$bwght, bdata$cigs, use = 'complete')
```

```
## [1] -0.08499059
```

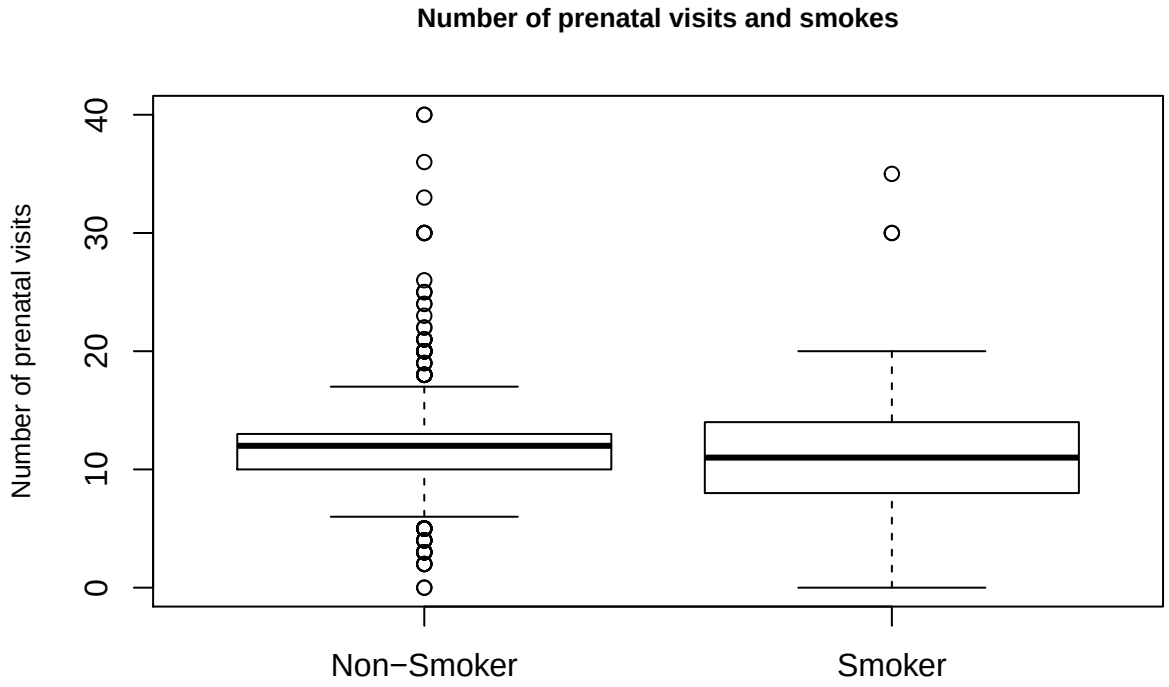
```
bdata$smokes <- cut(bdata$cigs, c(0,1,Inf), labels=c('Non-Smoker', 'Smoker'), right=FALSE)
summary(bdata$smokes)
```

```
## Non-Smoker     Smoker       NA's
##       1575        147        110
```

```
par(mfrow = c(1, 1))
boxplot(bdata$bwght~bdata$smokes, main = "Birth weight vs Smoker/Non-Smoker",
     ylab = "Birth weight", cex.main = .8, cex.lab = .8)
```




```
boxplot(bdata$npvis~bdata$smokes, main = "Number of prenatal visits and smokes",
        ylab = "Number of prenatal visits", cex.main = .8, cex.lab = .8)
```

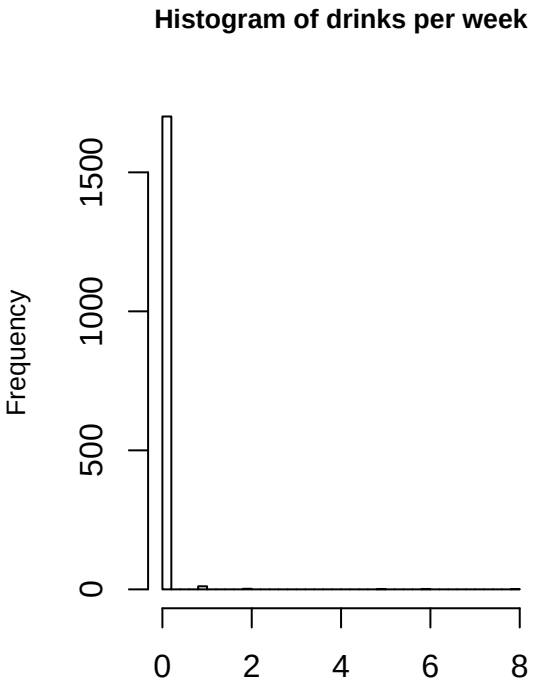


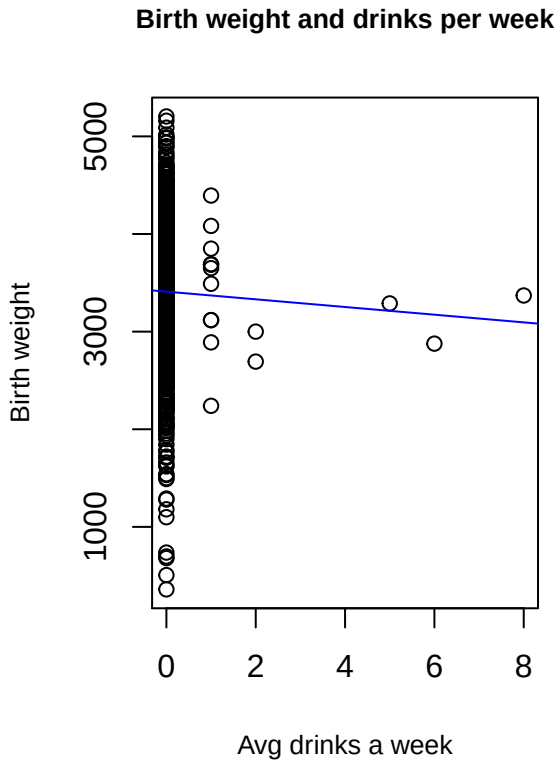
The histogram for cigarettes is heavily skewed right with most of the data at 0. The plot shows bands of data with some noise in between. The regression line shows that birth weight decreases as number of cigarettes smoked per day increases. There could be some measurement error in this data (likely asked people to guess how many they smoked on average) that may increase the variance of our response variable and introduce bias that might reduce the real effects of smoking. We converted the variable to an indicator for smoking to reduce the measurement errors. Smoker birth weight is much lower as displayed in the first box plot. Smokers on average also have slightly lower number of prenatal visits on average seen in the second box plot. This would be a good variable to include since we have a large number of observations and smoking has been associated with lower birth weights in other studies.

```
par(mfrow = c(1, 2))
summary(bdata$drink)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.0000  0.0000  0.0198  0.0000  8.0000     115
```

```
hist(bdata$drink, breaks = 30, cex.main = .8, cex.lab = .8,
     main = "Histogram of drinks per week", xlab = NULL)
plot(bdata$drink, bdata$bwght, main = "Birth weight and drinks per week",
     xlab = "Avg drinks a week", ylab = "Birth weight",
     cex.main = .8, cex.lab = .8)
abline(lm(bwght ~ drink, data = bdata), col = "blue")
```





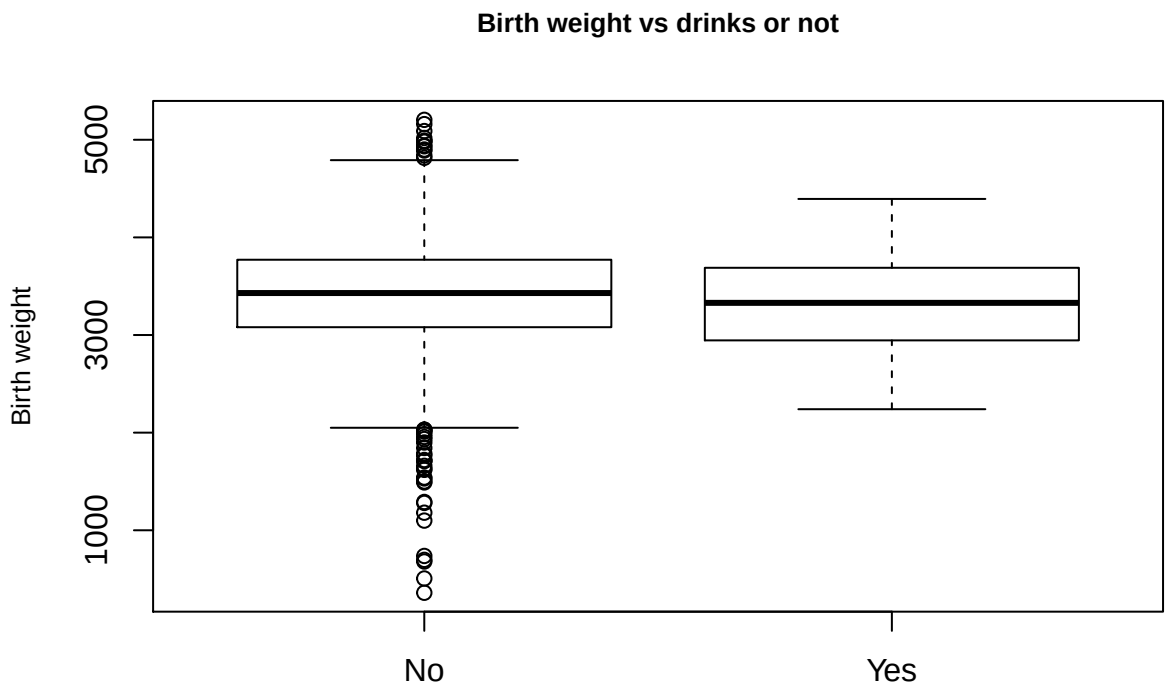

```
cor(bdata$bwght, bdata$drink, use = 'complete')
```

```
## [1] -0.01990582
```

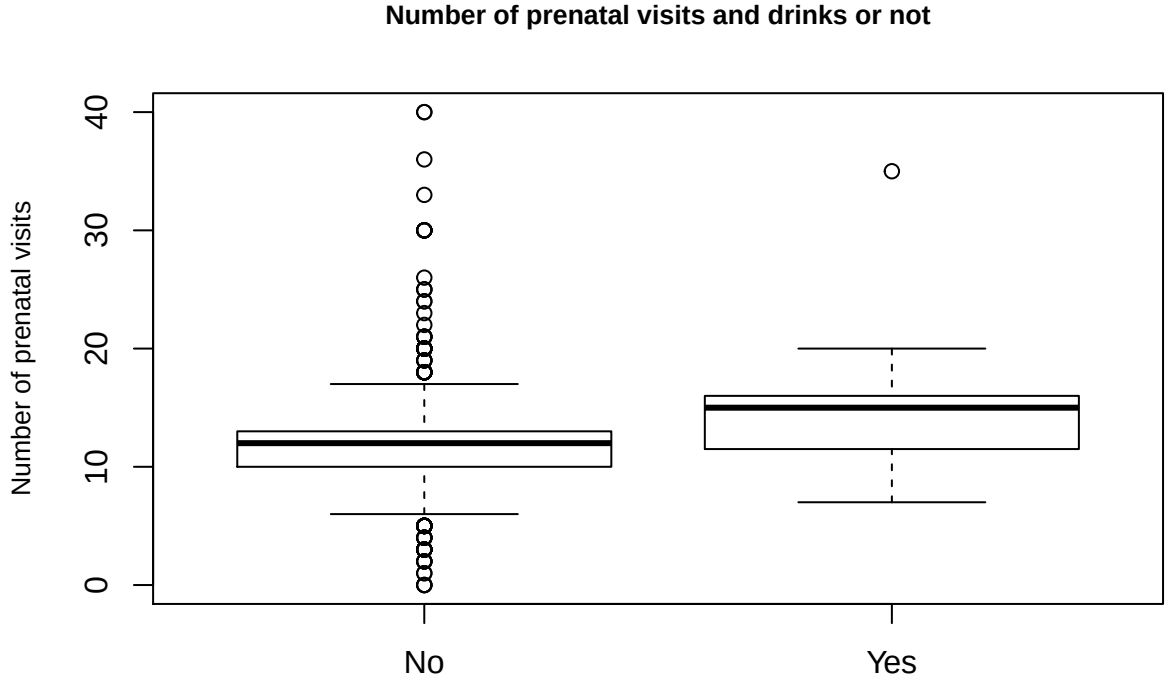
```
bdata$drinks <- cut(bdata$drink, c(0,1,Inf), labels=c('No', 'Yes'), right=FALSE)
summary(bdata$drinks)
```

```
##   No  Yes NA's
## 1701   16  115
```

```
par(mfrow = c(1, 1))
boxplot(bdata$bwght~bdata$drinks, main = "Birth weight vs drinks or not",
     ylab = "Birth weight", cex.main = .8, cex.lab = .8)
```




```
boxplot(bdata$npvis~bdata$drinks, main = "Number of prenatal visits and drinks or not",
        ylab = "Number of prenatal visits", cex.main = .8, cex.lab = .8)
```



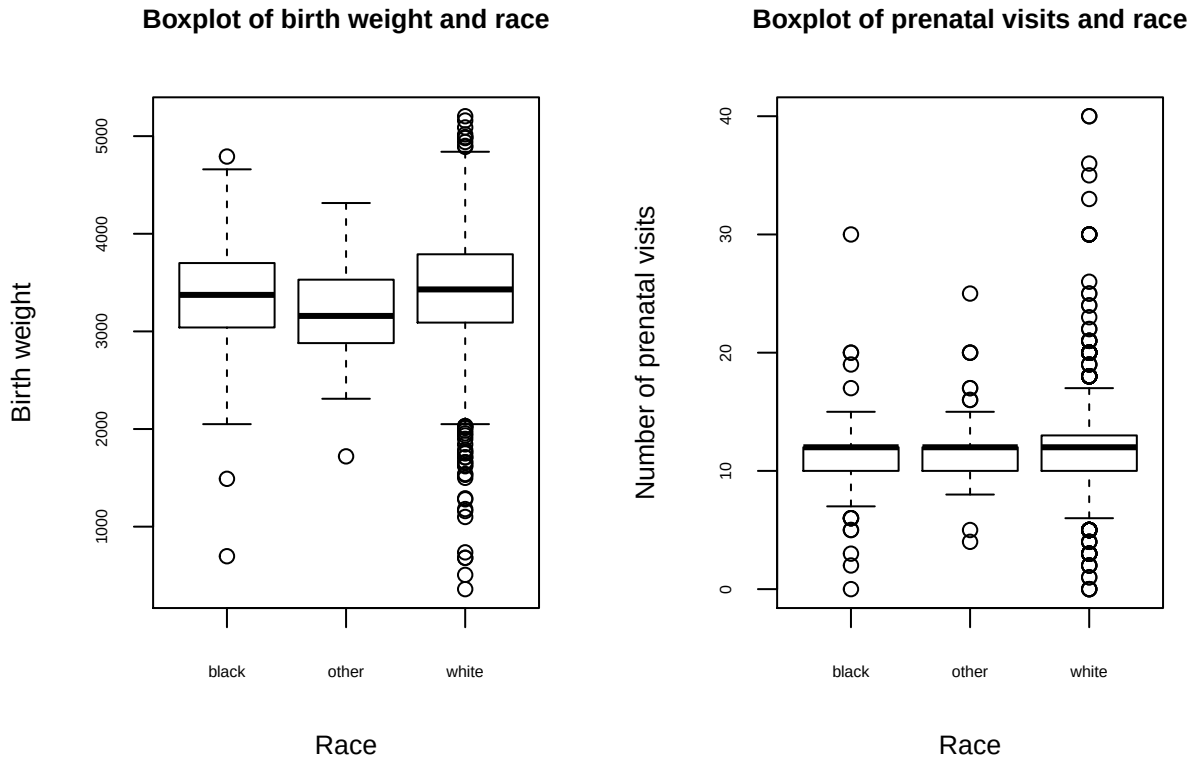
The histogram for drinks is heavily skewed right with most of the data at 0. The plot shows not a lot of data to the right. The regression line shows that birth weight decreases as number of drinks per week increases. There could be some measurement error in this data like the smoking variable that may increase the variance of our response variable and introduce bias that might reduce the real effects of drinking.

We converted the variable to an indicator for drinks to reduce the measurement errors. The birth weight for drinkers is lower as displayed in the first box plot. Drinkers on average also have a higher number of prenatal visits on average as seen in the second box plot. This could be a good variable to include since drinking is associated with lower birth weights, but we don't have a lot data here.

```
# ethnic differences
bdata$race[bdata$mblck == 1 & bdata$fblck == 1] <- 'black'
bdata$race[bdata$mwhte == 1 & bdata$fwhte == 1] <- 'white'
bdata$race[bdata$moth == 1 & bdata$foth == 1] <- 'other'
bdata$race <- as.factor(bdata$race)
summary(bdata$race)
```

```
## black other white  NA's
##    97    82  1607    46
```

```
par(mfrow = c(1, 2))
boxplot(bdata$bwght ~ bdata$race, main = "Boxplot of birth weight and race",
        xlab = "Race", ylab = "Birth weight", cex.main = .8, cex.lab = .8, cex.axis = .5)
boxplot(bdata$npvis ~ bdata$race, main = "Boxplot of prenatal visits and race",
        xlab = "Race", ylab = "Number of prenatal visits",
        cex.main = .8, cex.lab = .8, cex.axis = .5)
```



There's a lot more data for white parents in this data set compared to black and other. The box plot shows that parents falling under the other race have lower mean birth weights than black and white parents. However, the box plot of prenatal visits and race shows that the mean number of visits for each race is about the same. Background research yielded a study found that other races had lower birth weights [5]. Because of this, race would be a good variable to include in the model to account for the ethnic differences in infant sizes.

Modelling

For this study, we have chosen three model specifications. The first model has only the explanatory variables of key interest, birth weight and number of prenatal visits.

```
# model with number of prenatal visits
m1 <- lm(bwght ~ npvis, data = bdata)
```

The second group of models includes only covariates that we believe increase the accuracy of our results without introducing bias, as covered in our exploratory data analysis.

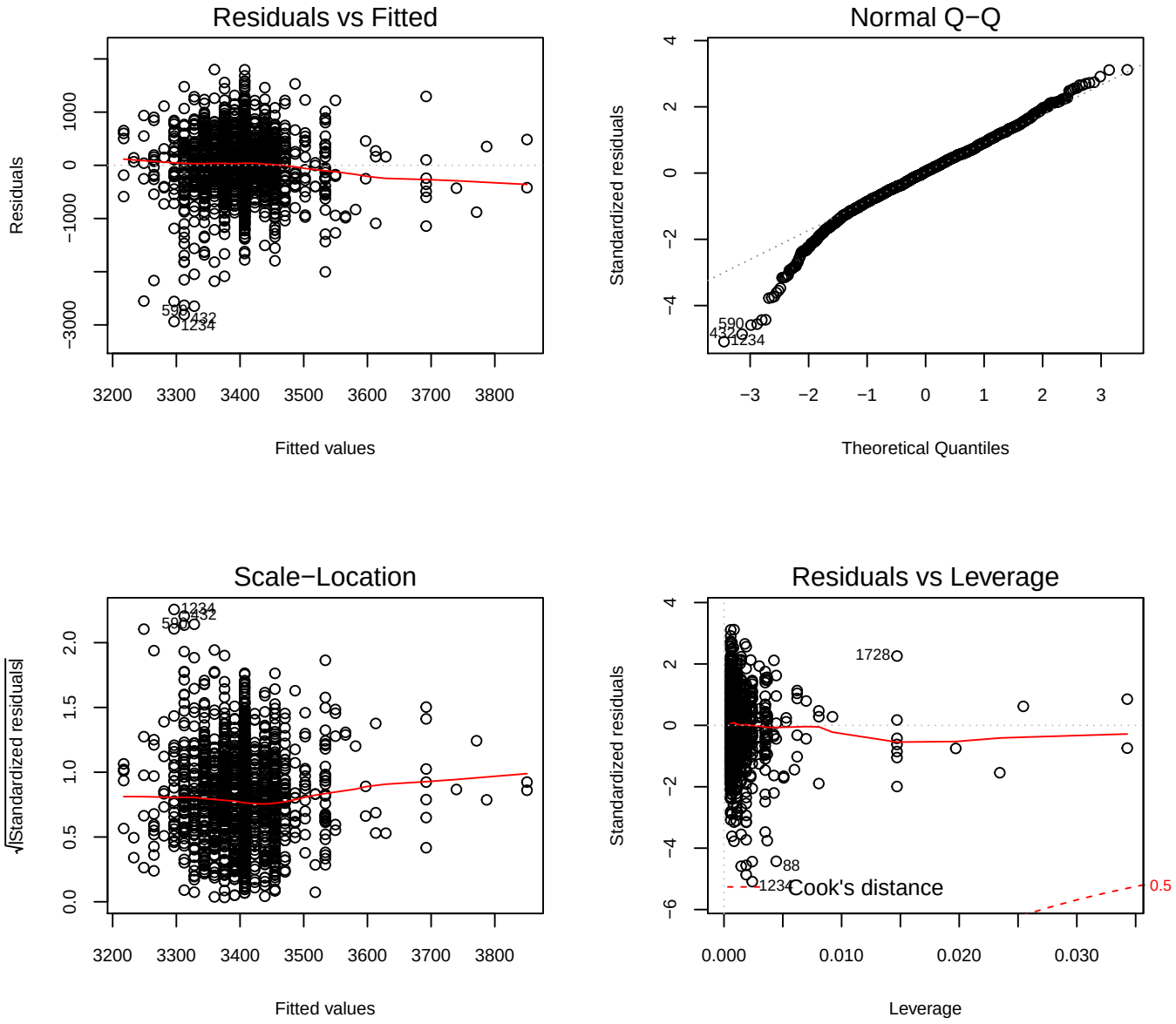
```
m2 <- lm(bwght ~ npvis + race, data = bdata)
m3 <- lm(bwght ~ npvis + race + smokes, data = bdata)
m4 <- lm(bwght ~ npvis + smokes + mage + I(mage^2) + race, data = bdata)
```

The third model includes the previous covariates, but also covariates that may be problematic–mother's education, father's age, and drinks. Instead of splitting the variables more logically across multiple models, all of these estimators are together in one model.

```
m5 <- lm(bwght ~ npvis + pnvpm + race + fage + mage + I(mage^2) + smokes + meduc + drinks,
         data = bdata)
```

Let's examine the CLM assumptions of the first model:

```
par(mfrow = c(2,2))
plot(m1, cex.main = .8, cex.lab = .8, cex.axis = .8)
```



CLM.1 - Our coefficients are assumed to be linear, but our independent variables can take on different transformations like polynomials or log. This is not a strong assumption and it is not restrictive, because we can represent any populalation plus some error.

CLM.2 - Our data is assumed to be random and iid. We have no indication as to where the data is from or how it was collected. We didn't see any indications that the data had any clustering or groupings. If the data was not random, then our estimates will not be as precise and we would need to account for this using clustered standard errors or another technique. For our purposes we'll assume it is random and iid.

CLM.3 - We did not find any of variables to be linear combinations of each other.

CLM.4 - The residuals vs fitted plot shows a the red line above zero on the left and moves more negative as it goes to the right. We do not meet the zero conditional mean for this model. We can not assume that our estimators are unbiased.

CLM.4' - Since we did not meet the zero conditional mean, we can assume we're now making an associative model and get exogeneity. This is a weaker assumption than zero condiational mean. Our estimators are no longer unbiased, but because we have a large sample they will be consistent. As our n goes to infinity, the bias should go to zero. We have over 1800 observations which should be sufficent to get consistency.

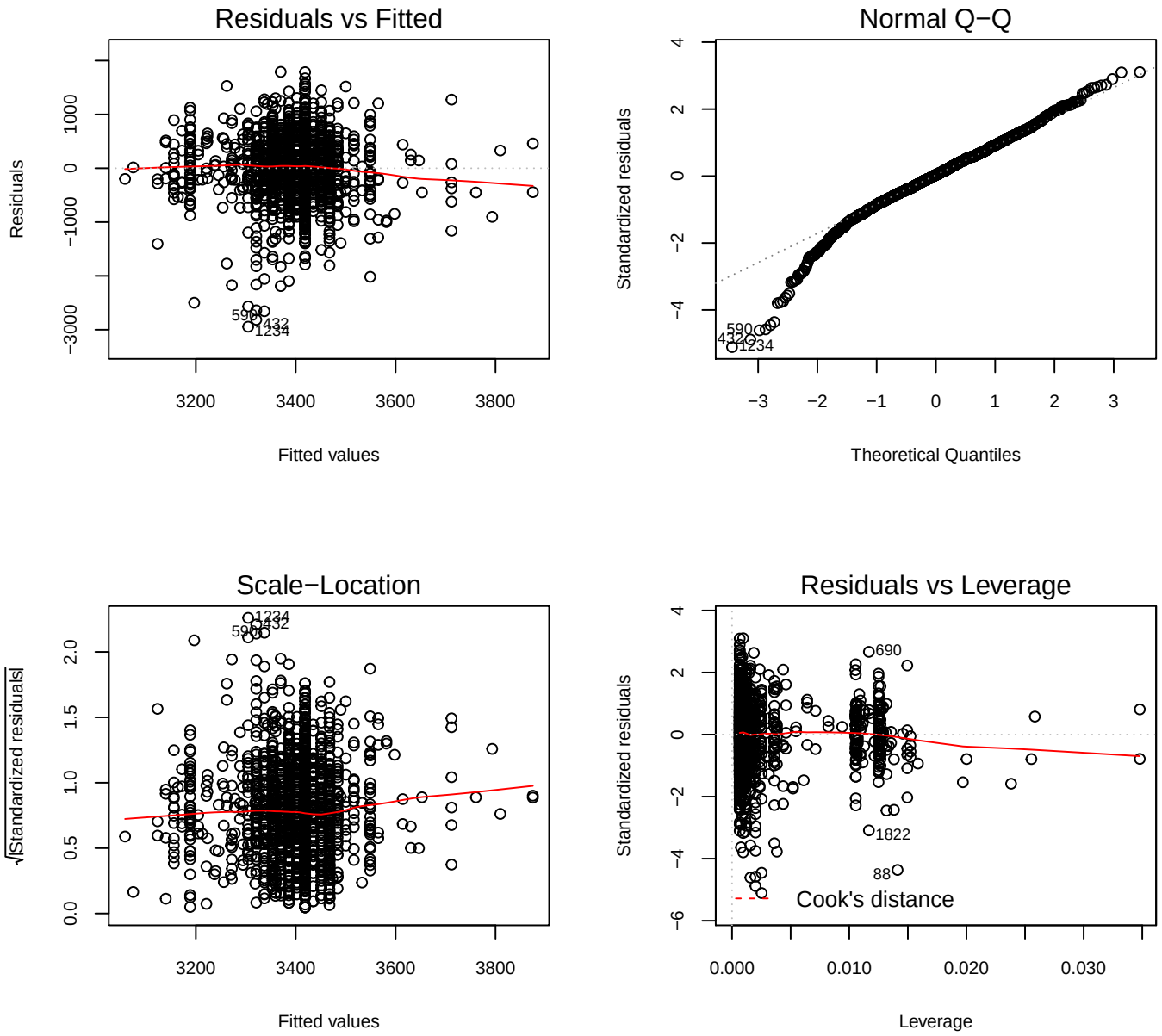
CLM.5 - The scale-location plot does not show a straight line. And the band of data on the residuals

vs. fitted plot is not even all the way across. Both plots indicate that we have heteroskedasticity. We can correct for this using robust standard errors which will be more conservative.

CLM.6 - The QQ-plot does not follow a nice diagonal line in the lower left likley due to the skewness in the birth weight. This means we have a violation of normality of errors. We have a large data set so we can rely on asymptoics and we don't need to correct for this.

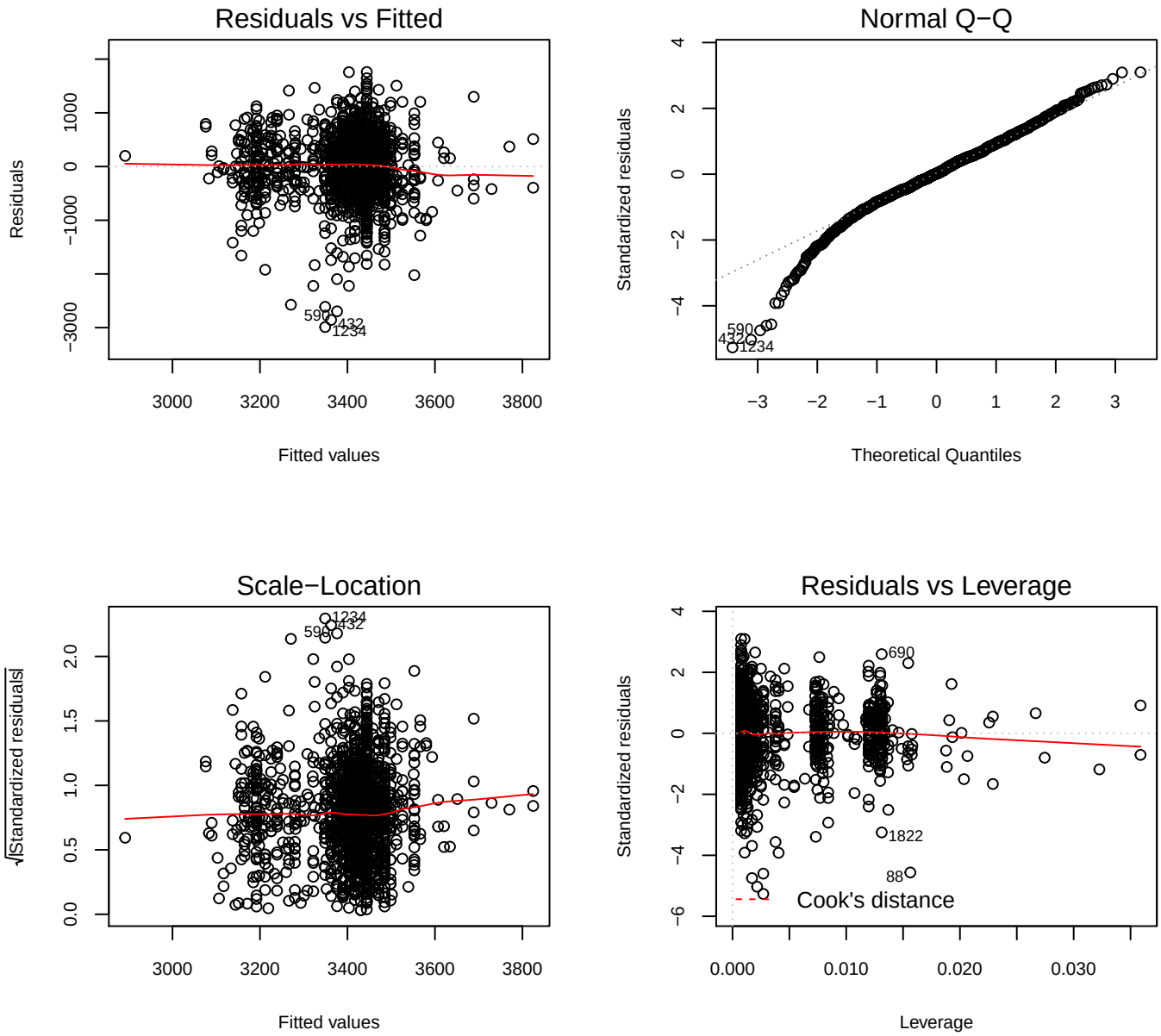
There are also no data points beyond Cook's distance so we will not remove any of the data points.

```
par(mfrow = c(2,2))
plot(m2, cex.main = .8, cex.lab = .8, cex.axis = .8)
```



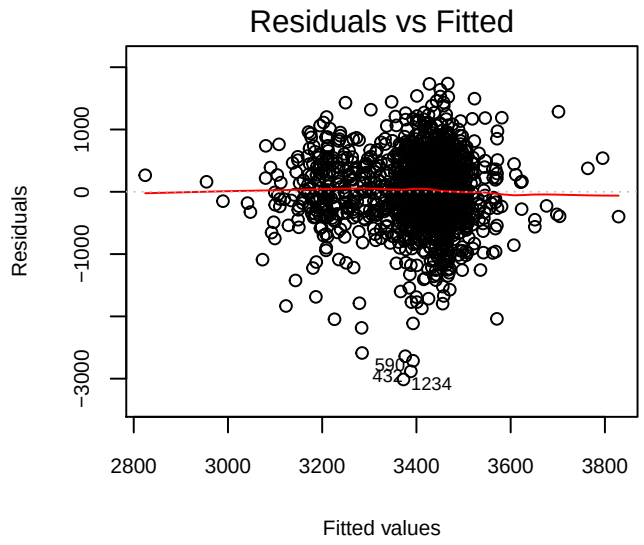
The m2 model has all of the same issues and resolutions as m1.

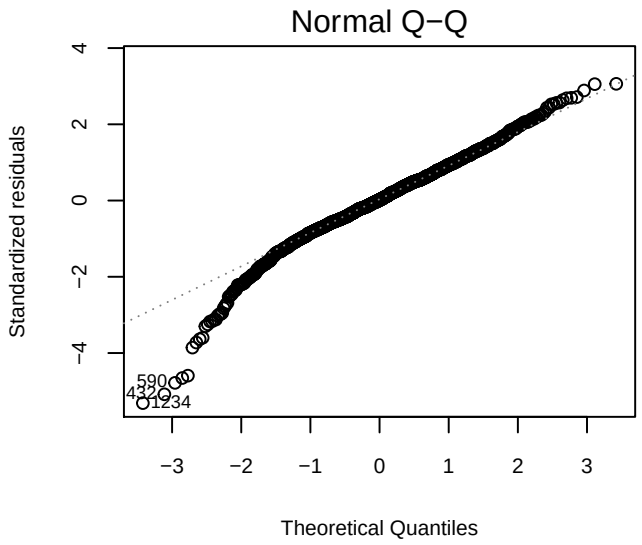
```
par(mfrow = c(2,2))
plot(m3, cex.main = .8, cex.lab = .8, cex.axis = .8)
```

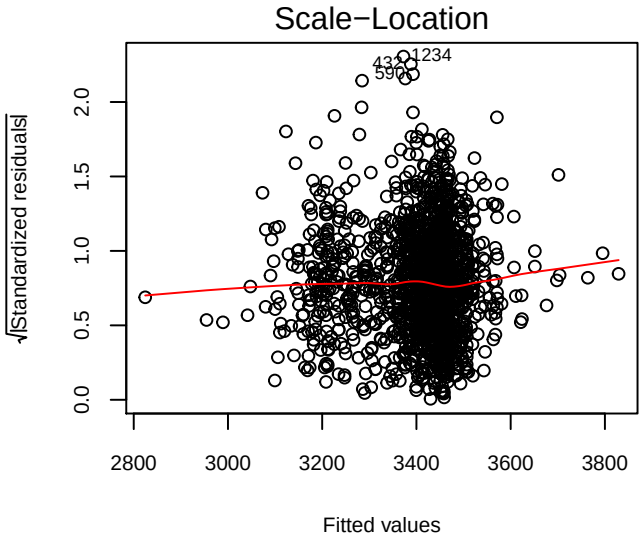


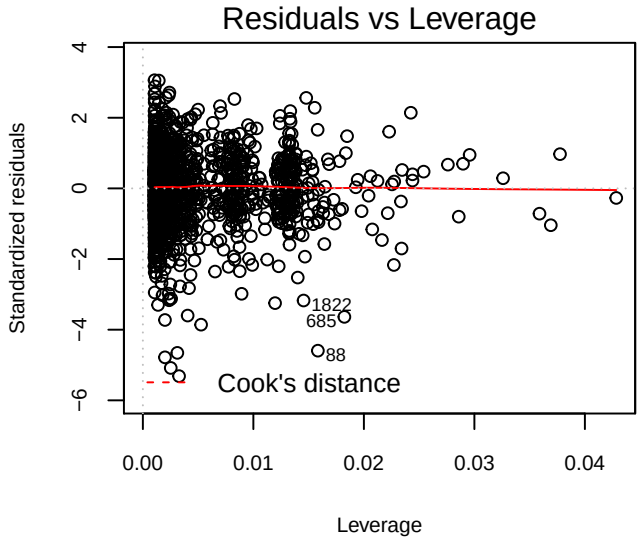
The plots for m3 have similar output as m2.

```
par(mfrow = c(2,2))
plot(m4, cex.main = .8, cex.lab = .8, cex.axis = .8)
```










```
cor.test(bdata$bwght, I(bdata$mage^2), method="pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  bdata$bwght and I(bdata$mage^2)
## t = 1.0344, df = 1830, p-value = 0.3011
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.02164774  0.06989291
## sample estimates:
##        cor
## 0.02417325
```

```r
cor.test(bdata$bwght, bdata$mage, method="pearson")
```

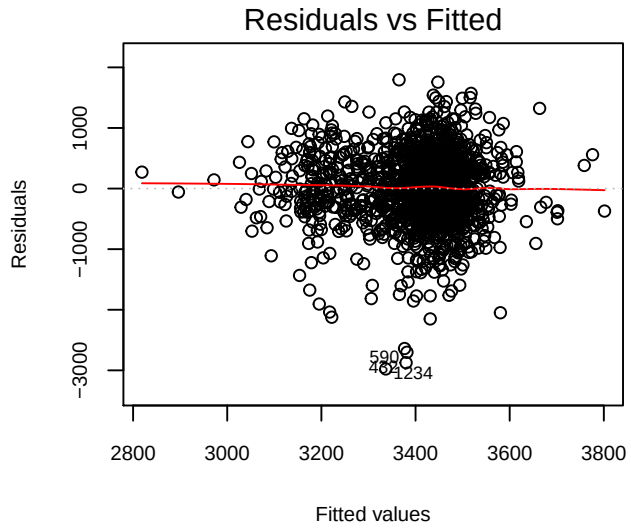
```
##
##  Pearson's product-moment correlation
##
## data:  bdata$bwght and bdata$mage
## t = 1.4417, df = 1830, p-value = 0.1496
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.01213309  0.07935728
## sample estimates:
##        cor
## 0.03368266
```

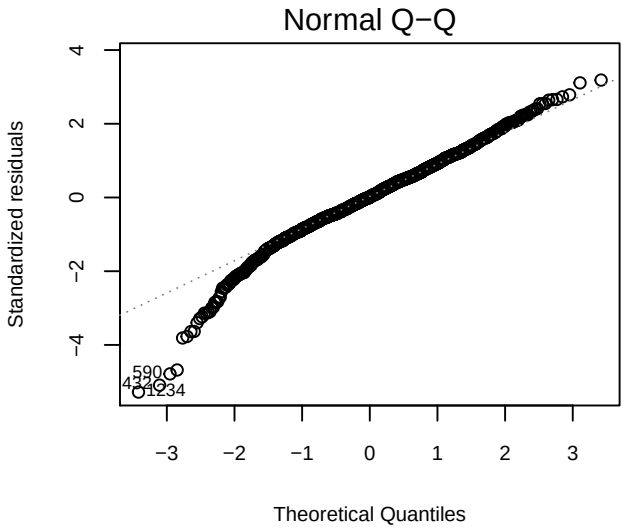
```r
vif(m4)
```

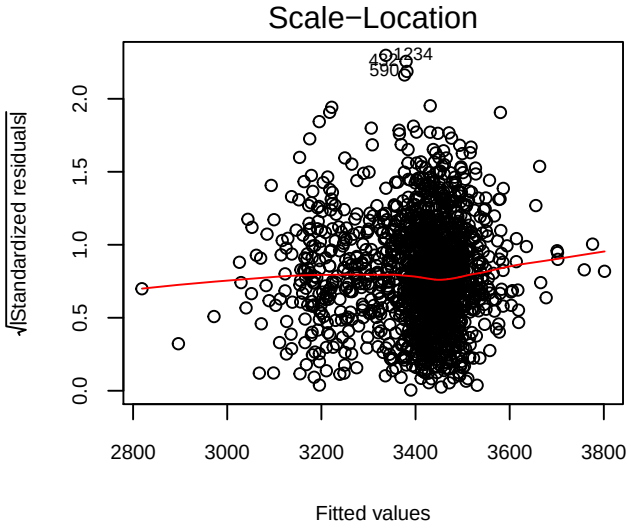
```
##                GVIF Df GVIF^(1/(2*Df))
## npvis      1.012919  1        1.006439
## smokes     1.016362  1        1.008148
## mage      85.245342  1        9.232840
## I(mage^2) 85.008531  1        9.220007
## race       1.013864  2        1.003448
```

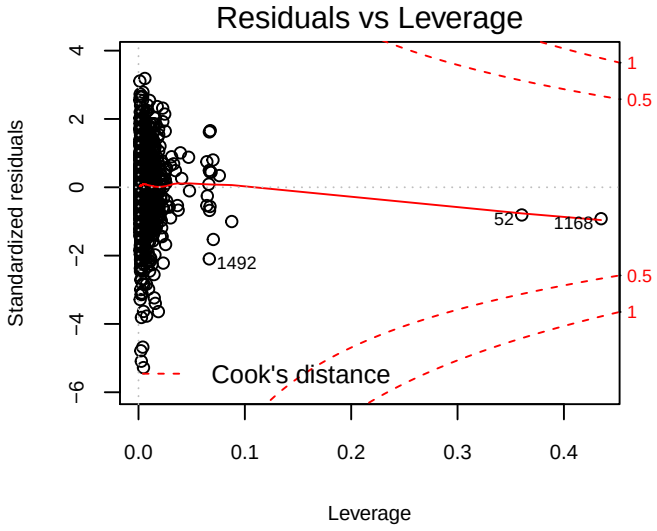
This model (m4) meets the zero conditional mean assumption. We now have unbiased estimators and no longer need to rely on our assumption of exogeneity. However, there is some multicollinearity between mother's age/mother's age squared and birth weight. We've added in the age^2 variable to account for the non-linear relationship between age and birth weight. We don't need to worry about this. The p-value for age^2 will not affected by the multicollinearity.

```r
par(mfrow = c(2,2))
plot(m5, cex.main = .8, cex.lab = .8, cex.axis = .8)
```










```
bptest(m5)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  m5
## BP = 13.784, df = 10, p-value = 0.1831
```

The fifth model, m5, has all the same issues as m4; however, a Breusch-Pagan test indicates that there isn't enough evidence to confirm the presence of heteroskedasticity, unlike the previous models. Since we have a large sample size we would need to see very large deviations before getting a statistically significant result. Since it's good practice and the evidence of our plots shows heteroskedasticity we'll use the robust standard errors.

```r
# generate the robust standard errors
se.m1 = sqrt(diag(vcovHC(m1)))
se.m2 = sqrt(diag(vcovHC(m2)))
se.m3 = sqrt(diag(vcovHC(m3)))
se.m4 = sqrt(diag(vcovHC(m4)))
se.m5 = sqrt(diag(vcovHC(m5)))
```

```r
# results='asis'
stargazer(m1, m2, m3, m4, m5,
          type = "latex",
          se = list(se.m1, se.m2, se.m3, se.m4, se.m5),
          df = FALSE,
          star.cutoffs = c(.05, .01, .001), title = "Results",
          table.placement = '!h')
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Dec 12, 2016 - 11:19:43

Assessments for statistical and practical significance for each model are as follows:

m1 - We see statistical significance for the number of prenatal visits at the .001 level. This model predicts that mothers should see an increase in birth weight of ~16 grams for each prenatal visit. For 10 visits, the mother could expect to see the weight of the infant increase by 160 grams. Considering that the infant should be betwen 2500 and 4000 grams, a difference of 160 doesn't seem to make that much of a difference. If the mother has reason to believe that the infant will have a very low birth weight, then the difference could be more significant for her.

m2 - We still have high statistical significance on our prenatal visits variable and a practical significance of ~16 grams per visit. We've added the categorical race variable. For race other we have statistical significance at the .05 level. Race other has a practical significance of reducing the predicted weight by 171 grams. Race white is not statistically signficant but has some practical effect of increasing weight by 59 grams.

m3 - We added the smoker indicator which has very high statistical significance and a practical significance of reducing weight by 205 grams.

m4 - We added mother's age and mother's age^2. These both are statistically significant at the .05 level. Age has a practical significance of increasing birth weight by 64 grams for each year of age. The age^2 variable has some practical significance of reducing the birth weight for different ages of mothers, since birth weight doesn't just increase linearly. This helps to take into account the lower birth weights at the extremes of the age range.

m5 - This model includes the prenatal visits per month variable which has no practical or statistical significance. It has multicollinearity with number of visits. The better variable to include is number of prenatal visits.

Overall, the number of prenatal visits is statistically and practically significant. Father's age is statistically significant, but has very little practical significance. This variable has some collinearity with mother's age, so it could be absorbing some of the effects of mother's age. Mother's education was not statistically significant or practically significant. The smoking variable was statistically significant, but the drink variable was not statistically significant. It had very little practical significance by reducing weight by 10 grams per 1 increase in drink. Mother's age has now lost statistical significance, but maintains some practical significance.

Adding these variables seems to have reduced the effects of our variable of interest and alone they don't have much practical or statistical significance.

Table 1: Results

	Dependent variable:				
	bwght				
	(1)	(2)	(3)	(4)	(5)
npvis	15.828***	16.288***	13.585**	13.045**	12.495**
	(4.394)	(4.413)	(4.325)	(4.337)	(4.736)
pnvpm					0.994
					(14.004)
raceother		−171.011*	−214.601*	−228.890**	−278.609***
		(80.699)	(84.895)	(84.586)	(80.773)
racewhite		59.151	37.523	26.258	−6.285
		(62.039)	(65.787)	(65.380)	(57.435)
fage					8.247*
					(3.588)
smokesSmoker			−205.094***	−193.146***	−195.272***
			(49.403)	(49.431)	(52.228)
meduc					3.487
					(7.114)
drinksYes					−10.673
					(151.022)
mage				63.533*	53.957
				(30.321)	(31.028)
I(mageˆ2)				−1.022*	−0.980
				(0.499)	(0.505)
Constant	3,217.367***	3,164.125***	3,243.619***	2,297.885***	2,269.377***
	(54.623)	(82.069)	(86.324)	(453.782)	(460.002)
Observations	1,764	1,721	1,613	1,613	1,590
R^2	0.010	0.018	0.026	0.029	0.033
Adjusted R^2	0.010	0.016	0.024	0.026	0.027
Residual Std. Error	577.568	577.008	568.460	567.797	565.196
F Statistic	17.939***	10.372***	10.731***	8.130***	5.468***
Note:					*p<0.05; **p<0.01; ***p<0.001

```
linearHypothesis(m5, c("pnvpm = 0", "fage = 0", "meduc = 0", "drinksYes = 0"), vcov = vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## pnvpm = 0
## fage = 0
## meduc = 0
## drinksYes = 0
##
## Model 1: restricted model
## Model 2: bwght ~ npvis + pnvpm + race + fage + mage + I(mage^2) + smokes +
##     meduc + drinks
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F Pr(>F)
## 1   1583
## 2   1579  4 1.3814  0.238
```

Together, the additional variables don't have statistical significance and should be removed from the model.

Causality

These results that prenatal care increases health cannot be interpretted causally. We don't have any background on how this study was conducted, but it does not look like a randomized study that assigned prenatal care to mothers. It also does not seem likely that prenatal care was assigned in a functionally random process, which would not make it an instrumental variable. Most theories behind birth weight do not relate to the number of prenatal visits. We also don't know what happened at the prenatal visits. Ideally every mother should follow the doctor's advice, but we don't know how many ignored the doctor's orders. It's difficult for us to say that an extra visit would see an increase in 15 grams of birth weight, for example.

Birth weight has more to do with other factors like mother's age, race, and smoking. It could be affected by variables not captured by our data. Other factors like genetics, mother's weight, income, proximity to health care, proximity of birth to due date, and pre-existing medical problems would be good variables to include in our analysis.

$bwght = \beta_0 + \beta_1 pnvis + u$, omitted: genetics

beta_2 + (genetics), gamma_1 - (correlation between genetics, pnvis is negative) = -
negative omitted bias, towards zero

$bwght = \beta_0 + \beta_1 pnvis + u$, omitted: mother's weight

beta_2 + (weight), gamma_1 - (correlation between pnvis, mother's weight is negative) = -
negative omitted bias, towards zero

$bwght = \beta_0 + \beta_1 pnvis + u$, omitted: income

beta_2 + (income), gamma_1 + (correlation between pnvis, income is positive) = +
positive omitted bias, away from zero

$bwght = \beta_0 + \beta_1 pnvis + u$, omitted: proximity to health care

beta_2 + (proximity), gamma_1 + (correlation between pnvis, proximity is positive) = +
positive omitted bias, away from zero

$bwght = \beta_0 + \beta_1 pnvis + u$, omitted: pre-existing medical condition

beta_2 - (condition), gamma_1 + (correlation of pnvis, pre-exist. medical condition is positive) = - negative omitted bias, towards zero

$bwght = \beta_0 + \beta_1 pnvis + u$, omitted: proximity of birth to due date

beta_2 + (proximity), gamma_1 + (correlation between pnvis, proximity is positive) = + positive omitted bias, away from zero

Including the smoking and drinks variables may bias the causal effects of prenatal care visits. By going to prenatal care you would expect to see some mothers reduce or stop their smoking and drinking. By including these variables, you take away the greater effect prenatal visits may have had in playing a role in a healthy birth weight.

Conclusion

This analysis examined the effects of several variables on birth weight using 5 different linear models. Each model was tested for CLM assumptions and adjusted accordingly. Overall, the number of prenatal visits was statistically and practically significant. We also found strong evidence that race other and smoking have high statistical significance and very high practical significance. Mother's age was transformed to account for the different ages and associated birth weights and found to be statistically significant and have practical significance.

We concluded the results that prenatal care increases health cannot be interpreted causally. It's unlikely that prenatal care was assigned in a functionally random process that would make it an instrumental variable. Most theories behind birth weight do not relate to the number of prenatal visits and we do not know what happened as a result of the prenatal visits.

Several factors were examined for omitted variable bias such as genetics, mother's weight, income, proximity to health care, proximity of birth to due date, and pre-existing medical problems. Including certain variables may bias the causal effects of prenatal care visits. Ultimately, by including too many variables we risk taking away the greater effect prenatal visits may have had in playing a role in a healthy birth weight.

Works Cited

[1] "How often do I need prenatal visits?," in WebMD, WebMD, 2016. [Online]. Available: http://www.webmd.com/baby/how-often-do-i-need-prenatal-visits. Accessed: Dec. 11, 2016.

[2] andytsh, "What affects a baby's birth weight?," Pregnancy & Baby, 2014. [Online]. Available: http://www.pregnancyandbaby.com/baby/articles/940601/what-affects-a-babys-birth-weight. Accessed: Dec. 11, 2016.

[3] "Apgar score," 2016. [Online]. Available: https://medlineplus.gov/ency/article/003402.htm. Accessed: Dec. 11, 2016.

[4] "How often do I need prenatal visits?," in WebMD, WebMD, 2016. [Online]. Available: http://www.webmd.com/baby/how-often-do-i-need-prenatal-visits. Accessed: Dec. 11, 2016.

[5] "Racial differences in birth weight of term infants in a northern California population," vol. 22, no. 3, Apr. 2002. [Online]. Available: http://www.nature.com/jp/journal/v22/n3/full/7210703a.html. Accessed: Dec. 11, 2016.

[6] Reichman, Nancy E., and Julien O. Teitler. "Paternal Age as a Risk Factor for Low Birthweight." American Journal of Public Health 96.5 (2006): 862–866. PMC. Web. 11 Dec. 2016. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1470584/