Lab 4: Does Prenatal Care Improve Infant Health?

Alex Lau, Jason Hunsberger, and Rob Mulla

December 13, 2016

Introduction

In this lab we will explore data provided by the National Center for Health Statistics and from birth certificates. We will analyze this data and find linear models that help us to understand the relationship between prenatal care and the health of a newborn. Over the course of our analysis we will look at a variety of variables, use several specification techniques, and determine which models help us to understand this relationship the best. Finally, we will discuss whether or not our exploration can establish a causal relationship between prenatal visits and the health of the newborn.

Setup and Initialization

We will import our dataframe as `Wellness` and also remove the `data` dataframe to avoid confusion.

```
load('bwght_w203.RData')
Wellness <- data
data <- NULL
```

We will also load all the libraries necessary to conduct our analysis.

Exploratory Analysis

To begin our exploratory analysis, we will run `str()` command to understand the datatypes in our dataset.

```
# str(Wellness)
```

We notice that we have numeric variables as well as dummy variables for the baby's gender and the race of the parents. It's important to note the number of NA values for variables as they may introduce bias. We can see the proportion of NA values per variable by running the `apply(!is.na())` command.

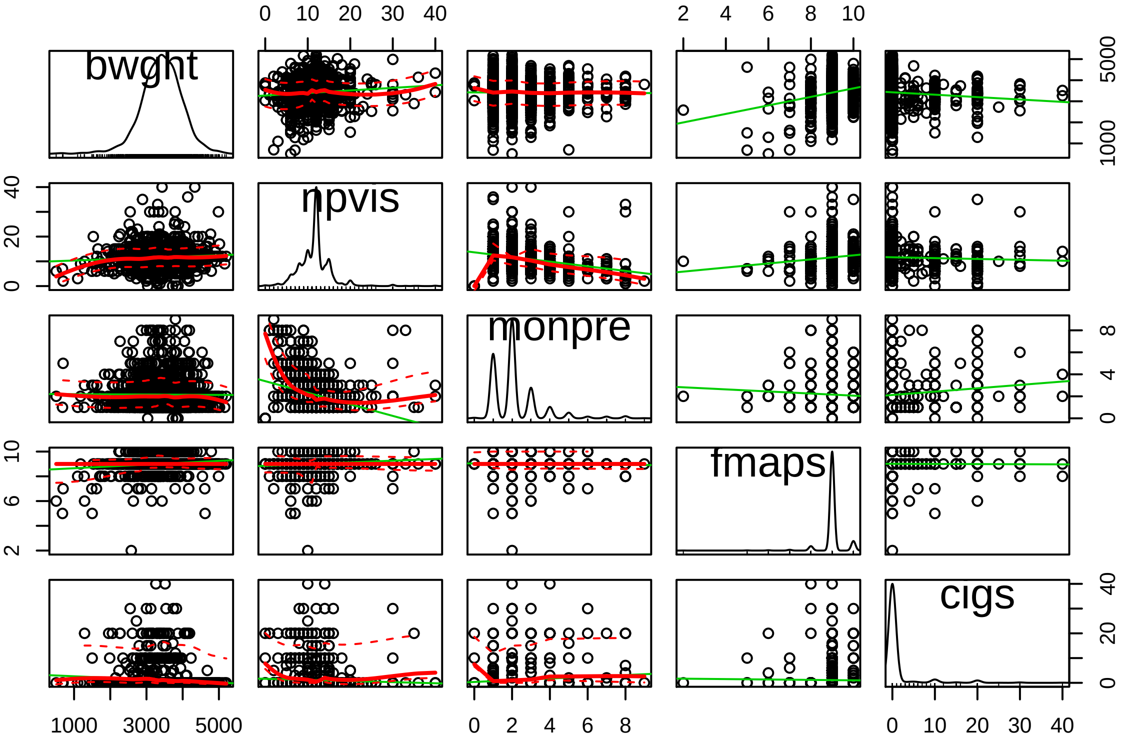
```
apply(!is.na(Wellness[,(1:20)] ) , MARGIN= 2, mean )
```

```
##      mage     meduc    monpre     npvis      fage     feduc     bwght 
## 1.0000000 0.9836245 0.9972707 0.9628821 0.9967249 0.9743450 1.0000000 
##     omaps     fmaps      cigs     drink       lbw      vlbw      male 
## 0.9983624 0.9983624 0.9399563 0.9372271 1.0000000 1.0000000 1.0000000 
##     mwhte     mblck      moth     fwhte     fblck      foth 
## 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
```

The output above shows that the variables `cigs` and `drink` seem to have the the most number of NA values. There may be a reason why some participants might not choose to answer these questions. This must be considered when creating our model.

We are trying to understand the wellness of the baby and whether prenatal visits have an influence. Our dataset contains some candidate dependent variables and independent variables. We will use a scatterplot matrix to look for any key relationships.

```
scatterplotMatrix(~ bwght + npvis + monpre + fmaps + cigs, data = Wellness)
```

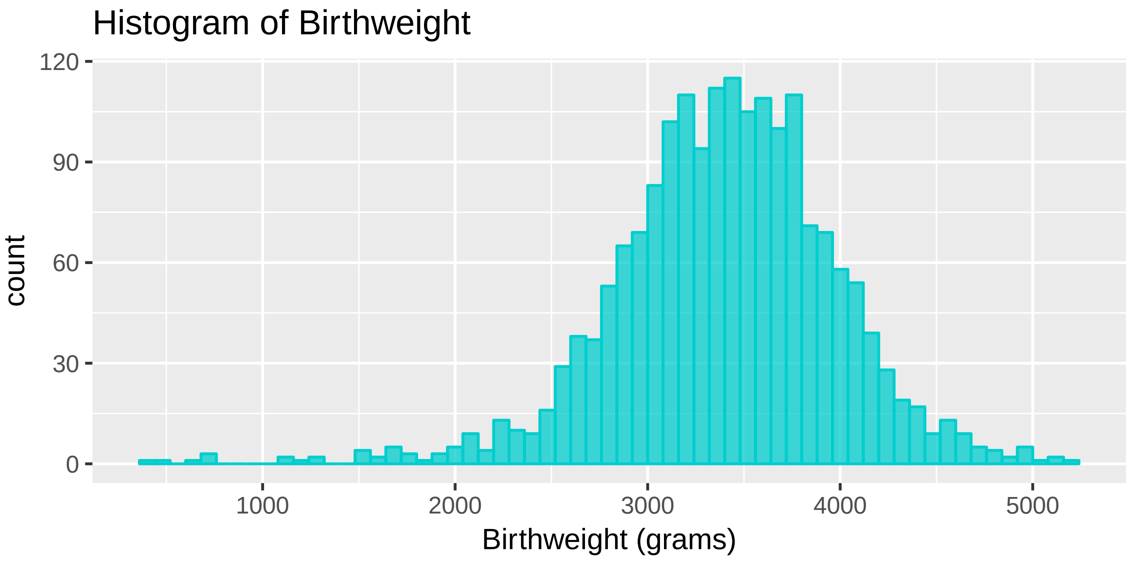


The scatterplot gives a good overall sence of the distribution for these variables, nothing jumps out at us as having a strong relationship.

Discussion of Variables and Transformations

Let us now look at each of the variables in depth. Our first will be birthweight - a clear indicator of newborn wellness.

```
qplot(Wellness$bwght, geom = "histogram", binwidth = 80,
      main = "Histogram of Birthweight", xlab = "Birthweight (grams)"
      , col = I("cyan3"), fill = I("cyan3"), alpha = I(.75))
```

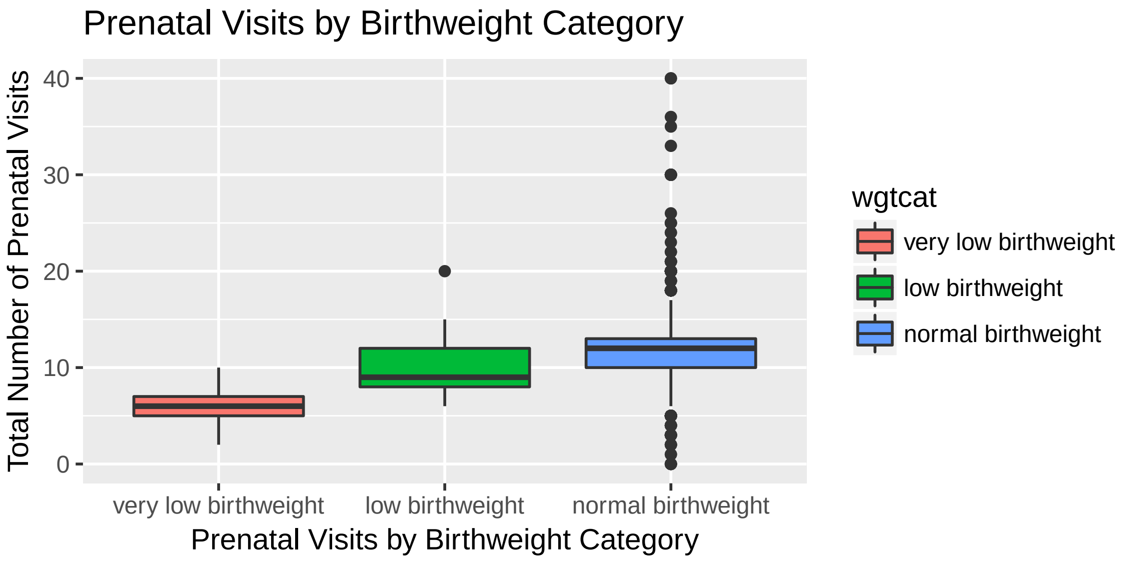


We notice that the distribution of our birthweight data looks fairly normally distributed with the exception of a few entries which skew the distribution to the left. Researching birthweights we find that a birthweight of under 1500 grams (3.3lbs) is considered very low birthweight (VLBW) and represent a population of premature infants. These infants are at increased risk for acute and chronic impairments related to their lack of development. Using the variable `vlbw` and `lbw` we can get can create a new variable for the birthweight category that divides our data into `very low birthweight`, `low birthweight` and `normal birthweight`.

```
Wellness$wgtcat = factor(ifelse(Wellness$vlbw == 1, "very low birthweight",
        ifelse(Wellness$lbw == 1, "low birthweight", "normal birthweight")))

# Force the order on the boxplot
Wellness$wgtcat <- factor(Wellness$wgtcat,
    levels = c('very low birthweight','low birthweight','normal birthweight'),
    ordered = TRUE)

ggplot(data=Wellness, aes(wgtcat, npvis, fill=wgtcat)) + geom_boxplot() +
  ggtitle("Prenatal Visits by Birthweight Category") +
  xlab("Prenatal Visits by Birthweight Category") +
  ylab("Total Number of Prenatal Visits")
```

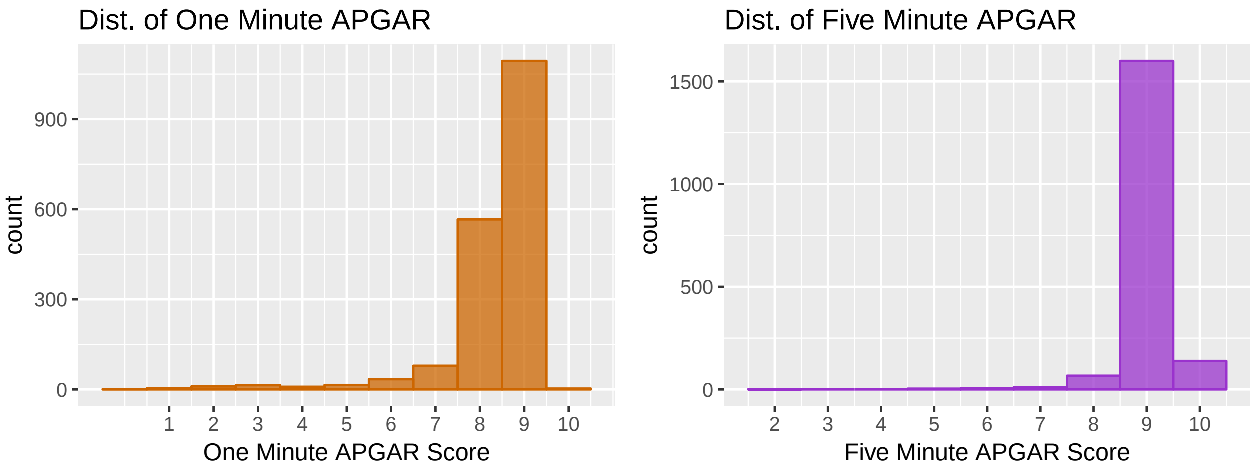


This boxplot shows generally that babies with very low and low birthweight on average had fewer prenatal visits. This finding is quite interesting for our analysis.

Next, we will explore two other variables that represent wellness of the baby: the APGAR scores at one-minute and at five-minutes. We will plot histograms of each.

```
omaps_hist <- qplot(Wellness$omaps, binwidth = 1,
                    main = "Dist. of One Minute APGAR",
                    xlab = "One Minute APGAR Score", col = I("darkorange3"),
                    fill = I("darkorange3"), alpha = I(.75)) +
              scale_x_continuous(breaks = seq(1,10,by=1))

fmaps_hist <- qplot(Wellness$fmaps, binwidth = 1,
                    main = "Dist. of Five Minute APGAR",
                    xlab = "Five Minute APGAR Score", col = I("darkorchid3"),
                    fill = I("darkorchid3"), alpha = I(.75)) +
              scale_x_continuous(breaks = seq(1,10,by=1))
grid.arrange(omaps_hist, fmaps_hist, ncol=2)
```

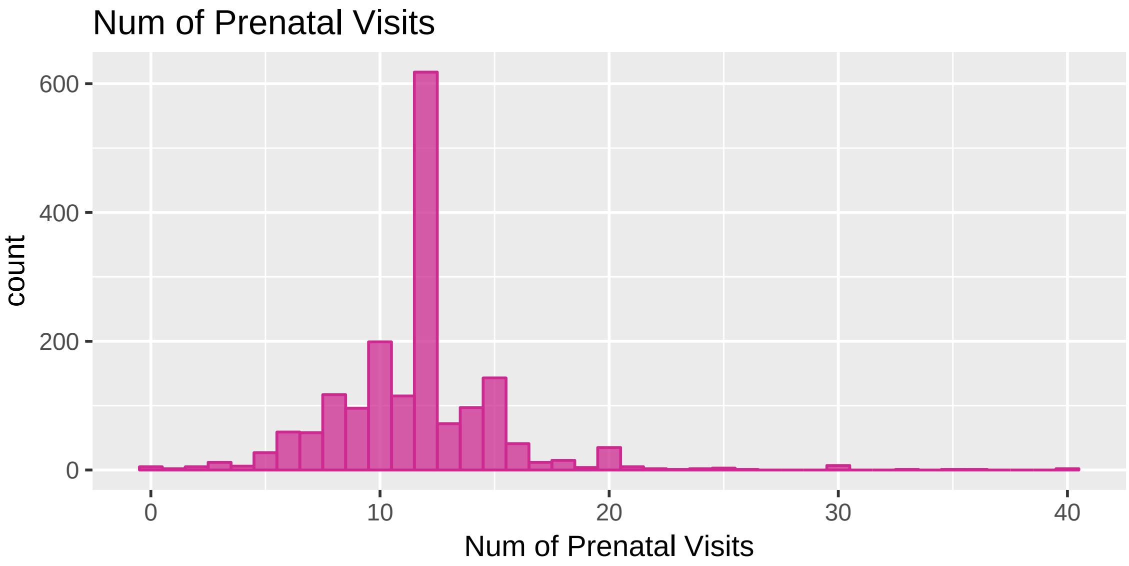


We notice that the distribution of scores seems much more varied for the one-minute test when compared to the five-minute test. We also note that both scores peak at 9 and there are very few scores of 10. In the five-minute test, there is much less veriation with fewer scores below 5.

This lack of variation in the APGAR scores concerns us. We may not have enough variation for APGAR to be a good indicator of wellness for our analysis.

We will explore the number of prenatal visits next. First we will plot the `npvis` variable

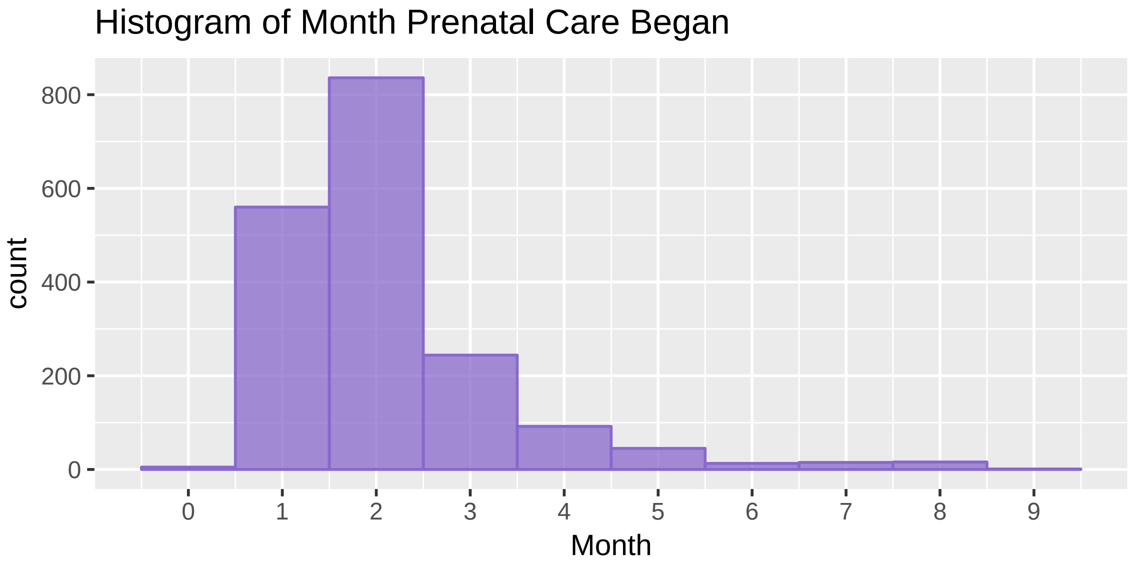
```
qplot(Wellness$npvis, geom = "histogram",
      binwidth = 1, main = "Num of Prenatal Visits",
      xlab = "Num of Prenatal Visits", col = I("maroon3"),
      fill = I("maroon3"), alpha = I(.75))
```



The prenatal visit data is not terribly normally distributed. There is a strong positive skew and a large spike at 12 visits.

A key variable for understanding the number of prenatal visits is `monpre`. This variable indicates when did prenatal care begin by month number. Here's a histogram of that data:

```
qplot(Wellness$monpre, geom = "histogram", binwidth = 1,
      main = "Histogram of Month Prenatal Care Began", xlab = "Month",
      col = I("mediumpurple3"), fill = I("mediumpurple3"), alpha = I(.75)) +
  scale_x_continuous(breaks=c(0,1,2,3,4,5,6,7,8,9))
```



```
summary(Wellness$monpre)
```

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

Both the histogram and the summary data show us that most mothers start prenatal care in month 1 or 2 of the pregnancy. The distribution is not particularly normal - it has a strong positive skew. However, given that mothers would know they are pregnant most often within a month or two of the pregnancy, this is not surprising or indicative of sampling error.

How can we analyze the number of prenatal visits in context of the month that prenatal care began? One way is to look at a boxplot of visits by month that care began. To do this we need to switch our months variable to a categorical variable and then plot the number of visits by each month. To do this, we will convert the month to a factor and inspect the data to ensure that it looks good:

```
Wellness$monpre.asfactor <- as.factor(Wellness$monpre)

str(Wellness$monpre.asfactor)
```

```
##  Factor w/ 10 levels "0","1","2","3",..: 3 3 2 6 3 2 4 7 4 3 ...
```

```
levels(Wellness$monpre.asfactor)
```

```
##  [1] "0" "1" "2" "3" "4" "5" "6" "7" "8" "9"
```

```
summary(Wellness$monpre.asfactor)
```

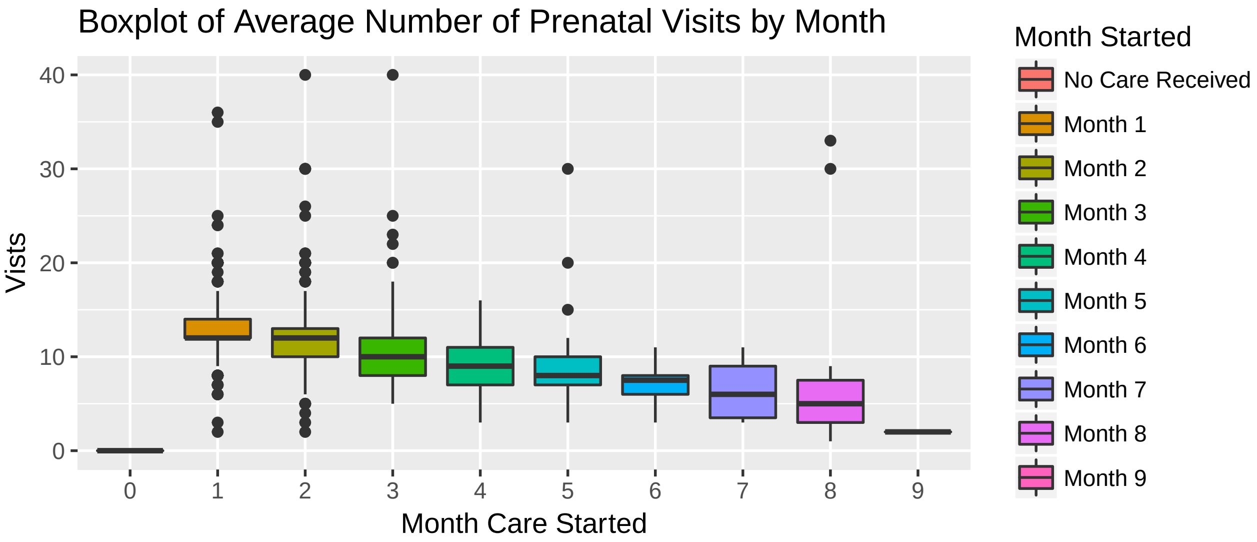
```
##    0    1    2    3    4    5    6    7    8    9 NA's 
##    5  560  836  244   92   45   13   15   16    1    5
```

The data above shows that we don’t have any unexpected factors appearing in our data. But we can also see that the number of data points that start in month 6 through 9 do not have much data associated with them. We will need to mindful not to make too many assumptions about data that comes from the third trimester.

Next, let’s plot the number of visits by month care started:

```
monthlabs <- c("No Care Received","Month 1","Month 2",
               "Month 3","Month 4","Month 5","Month 6",
               "Month 7","Month 8","Month 9")

ggplot(na.omit(Wellness), aes(x=monpre.asfactor, y=npvis, fill=monpre.asfactor)) +
  geom_boxplot() + ggtitle("Boxplot of Average Number of Prenatal Visits by Month") +
  scale_y_continuous(name = "Vists") + scale_x_discrete(name = "Month Care Started") +
  scale_fill_discrete(name = "Month Started", labels = monthlabs)
```



This boxplot is not terribly unexpected. It shows that mothers who start care in month 1 have a higher number of prenatal visits (on average) than those who start later. There are, however, some interesting outliers in this dataset. Note that most of the outliers exist for mothers who start prenatal care in the first trimester. Some mothers clearly are receiving lots more care if they start early. Is that due to troubled pregnancies? We can't tell yet. But the two outliers in month 8 do need investigation.

We will filter the data to only show the rows for mothers with over 30 prenatal visits that start receiving care in month 8.

```
na.omit(Wellness[Wellness$npvis >= 30 & Wellness$monpre == 8,])
```

```
##      mage meduc monpre npvis fage feduc bwght omaps fmaps cigs drink lbw
## 52     29    11      8    30   31    12  3336     8     9    0     0   0
## 1168   25    12      8    33   26    12  3310     9     9    0     0   0
##      vlbw male mwhte mblck moth fwhte fblck foth   lbwght magesq npvissq
## 52      0    0     1     0    0     1     0    0 8.112528    841     900
## 1168    0    0     1     0    0     1     0    0 8.104704    625    1089
##                  wgtcat monpre.asfactor
## 52   normal birthweight               8
## 1168 normal birthweight               8
```

This number of prenatal visits with less than two months to go in the pregnancy seem extremely unlikely.

One member of our team has had two children and finds this frequency inconsistent with what would actually happen. At eight months the baby is fully viable. If a mother came in at eight months and doctors found there to be an issue, the doctors would not have them come in every day or every other day for check-ups. Instead, they likely do two things for these mothers: have them put on forced bedrest in the hospital under constant monitoring and at some point schedule a c-section.

If we look at the data on these two mothers, there are no other indicators of trouble in the birth: both mothers are under 35, don't smoke or drink, both babies are at a healthy 8+ pounds, their APGAR scores are 8 and 9. After a careful analysis, we believe these are likely a coding or measurement error and should be excluded from the dataset.

```
# remove the rows that contain the outliers
Wellness <- Wellness[-c(52, 1168),]
```

Since the number of prenatal visits varies depending upon which month the mother started receiving prenatal care, we can get a clearer picture of normal versus abnormal amounts of visits by taking an average of the visits per month. Granted, an average may not truly reflect the nature of the data on an individual case. It could well be that there are mothers who start out with a large number of visits and then they drop their visits over the pregnancy. Or vice versa. Absent data that plots the number of visits per month, a simple average is the best approximation method available to us to gain insight into the number of prenatal visits.

To calculate this we need to take `npvis` and divide it by the number of months that care was received. We don't have a direct measure of the number of months that care was received. Instead, we know when prenatal care started. We can see that with a simple summary command:

```
summary(Wellness$monpre)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   1.000   2.000   2.116   2.000   9.000       5
```

What this shows is that the months range from 0 to 9 months. The 0 represents mothers who received no care. How many rows of data does this represent? Let's look.

```
nrow(Wellness[Wellness$monpre < 1,])
```

```
## [1] 10
```

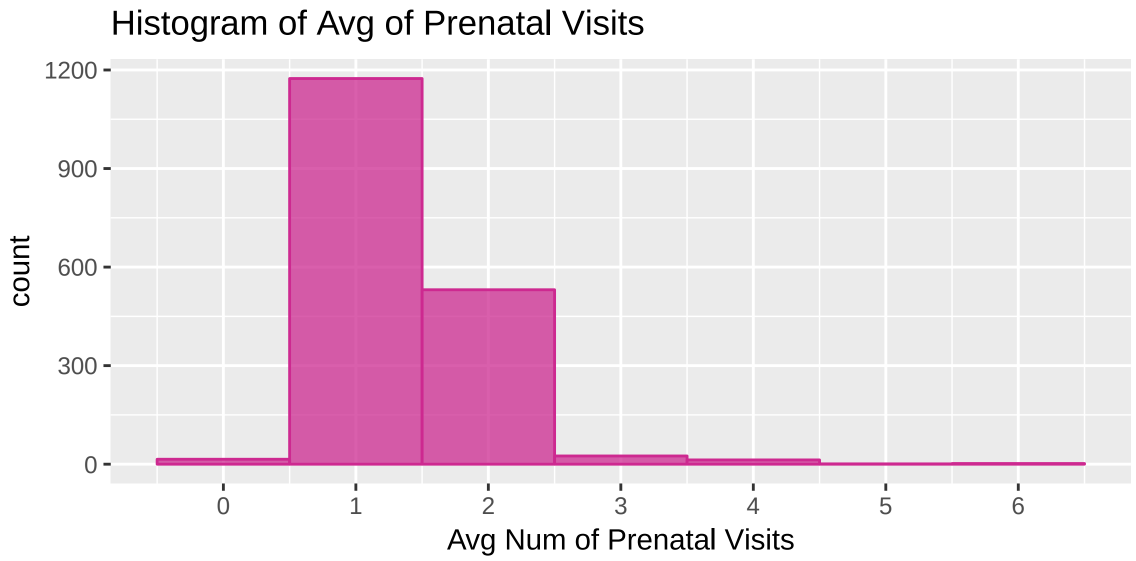
Only 10 mothers in our dataset received no prenatal care.

With the presence of mothers that received no care in our dataset, we need to recognize that there are actually 10 different values that can be selected for this measure. A 9 represents month 9 and a 1 represents month 1. As such, to get the number of months, we need to subtract from 10. Putting it together, we can calculate the `Average Number of Prenatal Visits per Month` as follows:

```
Wellness$avg.npvis <- Wellness$npvis / (10 - Wellness$monpre)
```

Let's take a look at a histogram of this data:

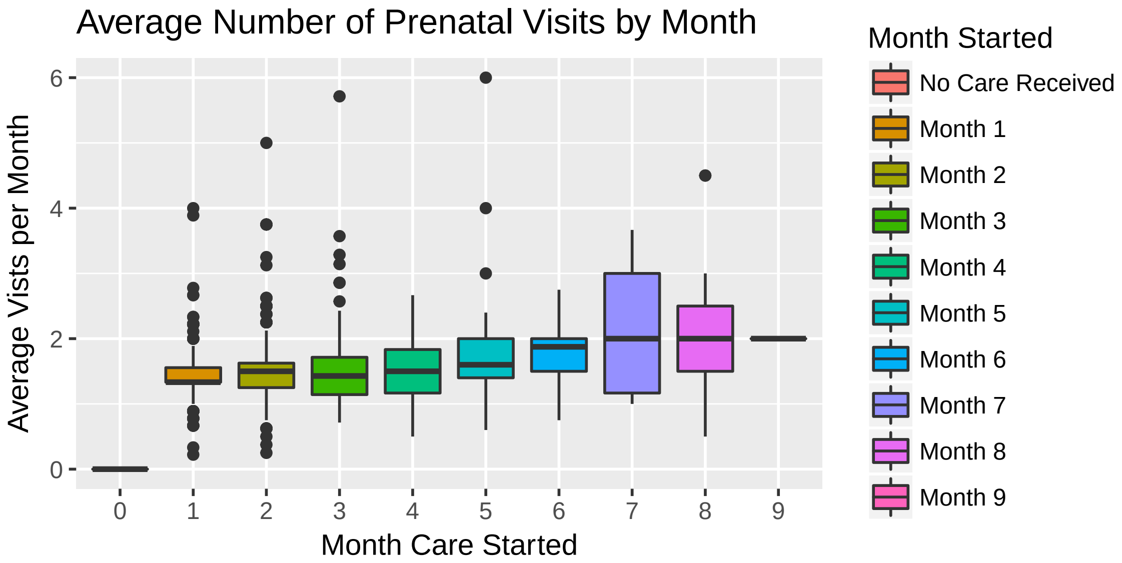
```
qplot(Wellness$avg.npvis, geom = "histogram", binwidth = 1,
      main = "Histogram of Avg of Prenatal Visits",
      xlab = "Avg Num of Prenatal Visits", col = I("maroon3"),
      fill = I("maroon3"), alpha = I(.75)) +
  scale_x_continuous(breaks=c(0,1,2,3,4,5,6))
```



This histogram shows that in our sample, most mothers visit once or twice a month. We note that while this variable provides benefeficial insight about the frequency of visits for each mother, it does come at the expense of losing detail about when the mother started their care. We will consider the benefits and drawbacks of using this variable in our model.

Now, let's see if the average number of prenatal visits changes by the month that care started. Let's return to the boxplot:

```
ggplot(na.omit(Wellness), aes(x=monpre.asfactor, y=avg.npvis, fill=monpre.asfactor)) +
  geom_boxplot() + ggtitle("Average Number of Prenatal Visits by Month") +
  scale_y_continuous(name = "Average Vists per Month") +
  scale_x_discrete(name = "Month Care Started") +
  scale_fill_discrete(name = "Month Started", labels = monthlabs)
```



This boxplot is both a little hard to understand and fascinating. What is shows is how the *average number of prenatal visits* changes based upon the *month in which care is started*. We can see that the average is quite tight - from the left, most months the average is around 1.5 visits per month and increases to approximately 2 per month 8.

Another feature of this boxplot is the outliers. Again, most of the outliers exist in the first trimester, where some women receive between 4 and 6 prenatal visits per month. We understand that it is not unheard of that doctors might ask women with certain risk factors to visit once a week. In short, at this point this data all seems within what is reasonable.

Models

With our exploration of our key variables complete, we will now turn to building models to better understand the relationships between the variarbles.

Our first priority in the model process is to help understand whether prenatal care improves health outcomes for newborn infants. We will add additional variables to our models to gain a greater understanding of what variables may impact a newborn. We will explore transformations of variables, discuss which variables should be included in each model, and whether model assumptions are met.

Model 1 - Only the explanatory variables of key interest.

For our first model, we are going use birthweight as our indicator of baby wellness. We were troubled by the lack of variation in the APGAR scores during our exploratory phase and feel birthweight is a sufficient proxy for overall health. Additionally, we are going to use the average number of prenatal visits as our predictor variable. We felt this was a better way to understand the impact of prenatal care and will be easier to communicate our findings with. As noted above we do understand we lose some detail about the prenadal care, specifically the impact of starting care late in pregnancy, but we believe the benefits of using average monthly vists is worth losing this detail.

$$bwght = \beta_0 + \beta_1 avg.npvis$$

```
model1 = lm(bwght ~ avg.npvis, data = Wellness)
```

Let's analyze this model to see if it satisfies the classical linear model assumptions.

MLR.1 The population is able to be represented by a linear model

We assert that it is reasonable to consider the relationship between the number of prenatal visits and birthweight to be linear in nature. Regular check-ups would create an opportunity for diagnostics to measure the health of both the mother and the baby. In circumstances where trouble is encountered, doctors would have the opportunity to intervene at a stage earlier than birth to help improve the potential outcome. This prenatal care doesn't need non-linear models in order for the primary benefit of prenatal care to be experienced. As such, we feel a linear model is sufficient for our purposes.

MLR.2 The population was sampled randomly

The data we are analyzing is from the National Center for Health Statistics (NCHS) and from birth certificates. How these statistics were gathered and which birth certificates were selected are unknown to us. We will have to trust the professionalism of the NCHS and their statisticians that a random sample was obtained.

MLR.3 There is no perfect multicollinearity between predictor variables

In analyzing our model, we need to look for possible issues with multicollinearity. Since we only have one predictive variable (`avg.npvis`) in our model, we cannot use the standard Variance Inflation Factor (VIF) test. Instead, we will need to look at the r^2 value for this model.

```r
summary(model1)$adj.r.squared
```

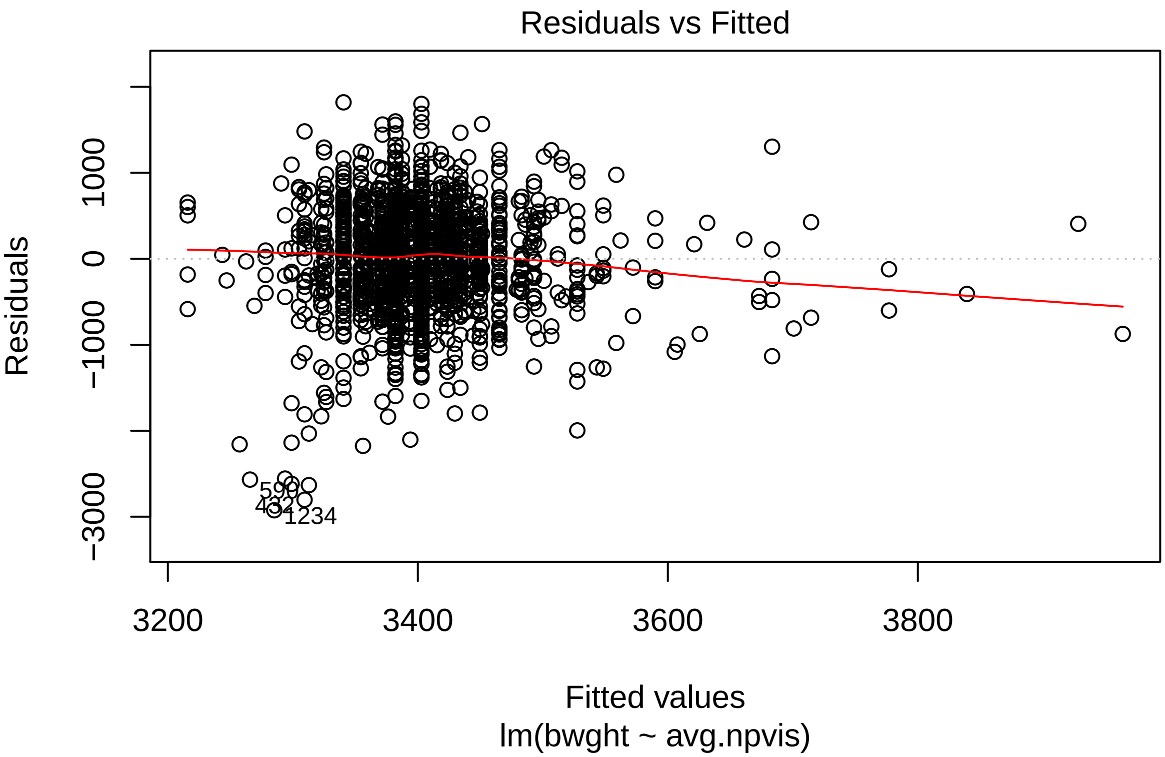
```
## [1] 0.009867269
```

We can see that we have a very low adjusted r^2 value of 0.010. This is well below 1 and indicates that multicollinearity is not an issue for our model.

MLR.4 Zero-Conditional Mean

This assumption asserts that the expectation of our model's error term conditional on our variables is zero. We can see this by running a residuals vs. fitted plot and inspecting the graph for a flat fitted line.

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



In the above graph we can see that for the bulk of our data our fitted line is almost flat. The only issue we see with it is that as babies reach the high-end of the weight scale, we are no longer able to maintain the zero conditional mean - some factors of our error begin to take over and we under-predict birthweight. This could simply be attributed to the lack of samples on the right side of the graph.

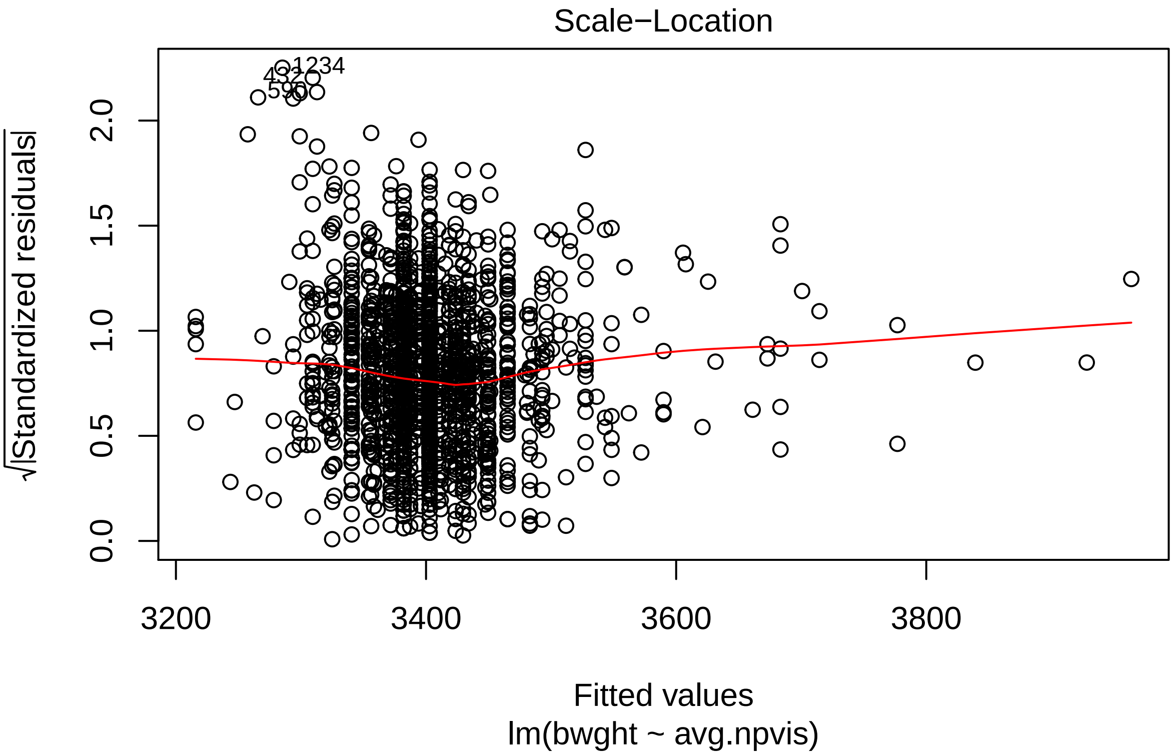
MLR.4' Exogeneity

Since we cannot fully assert zero conditional mean, we cannot assert that our model is fully causal. Instead, we can say that the model is associative in nature and that the model is exogenous. Exogeneity means that in our population the error terms are consistent and have no covariance on our selected variables. If we think about birthweight and the average number of prenatal visits in the population, we struggle to come up with a scenario that wouldn't apply consistently. We believe it is safe to assert exogeneity.

MLR.5 The variance of the error term is constant - homoskedasticity

Next, we need to understand whether we have evenly distributed error throughout our model. We can inspect the scale-location plot and look again for a straight line as well as the overall shape and distribution of the residuals.

```
plot(model1, which = 3)
```



Our scale-location plot shows us that we aren't quite able to fulfill the homoskedaticity requirements of the CLM. We have a dip around the first third of the plot, and as the weight of the babies increases, an overall rise in errors occurs. We note that the variance of the error appears to become more narrow as we approach the right side of the plot.

We are also going to look at the results of a Breusch-Pagan test to see if it provides any additional insight on whether we have homosketasticity or not. NOTE: this test is sensitive to large sample sizes, which we have here.

```
bptest(model1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 13.052, df = 1, p-value = 0.000303
```

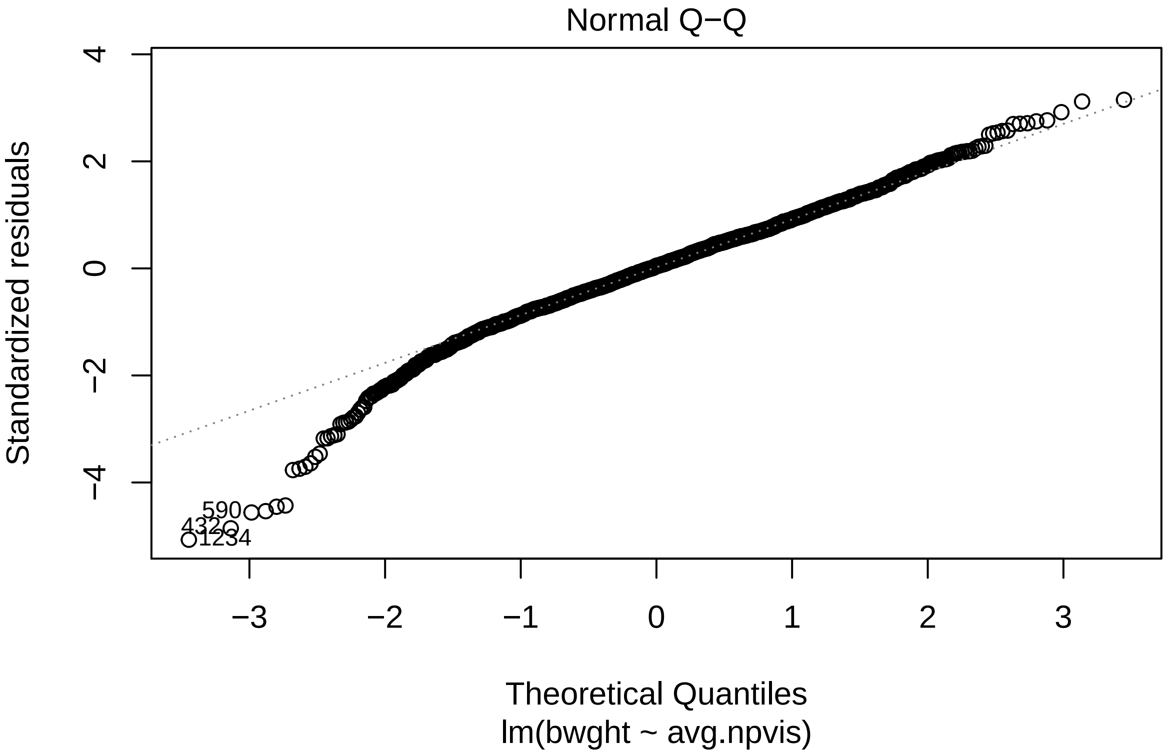
The Breusch-Pagan test confirms that we have a statistically significant reason for rejecting the null hypothesis that the data is homoskedastic. In other words, our data *is heteroskedastic*.

With that said, we can account for this model's heteroskedasticity by applying heterskedastic-sensitive methods of calculation to compensate.

MLR.6 Normality of the error terms

Finally, we need to evaluate the residuals and check to see if they are normal in distribution. First, we will look at a Q-Q plot.

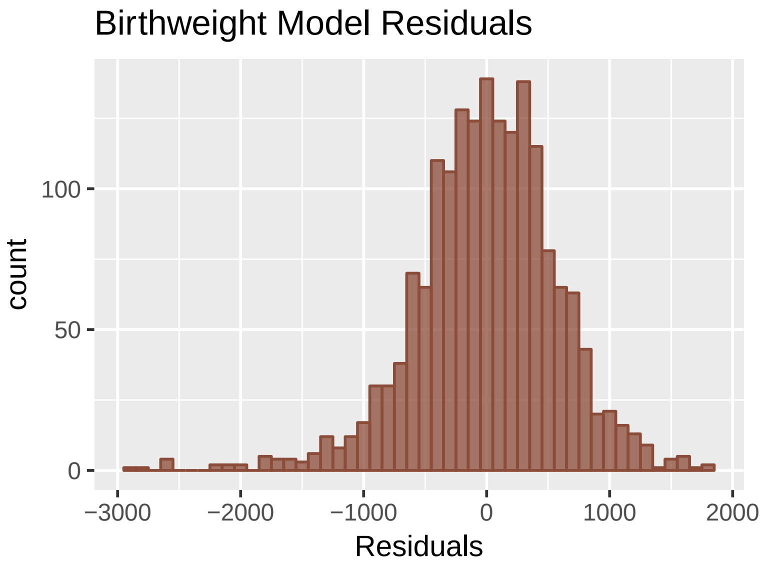
```
plot(model1, which = 2)
```



The Q-Q plot for our residuals has some deviations from normality. While it is largely fine for the upper 5/7ths of the graph, the bottom 2/7ths fall away. This means that our model may heavily underestimate birthweight for cases of low birthweight.

This can also be seen by plotting a histogram of our model residuals.

```
qplot(model1$residuals, geom = "histogram", binwidth = 100,
      main = "Birthweight Model Residuals", xlab = "Residuals",
      col = I("salmon4"), fill = I("salmon4"), alpha = I(.75))
```



Our histogram shows that while the residuals are mostly normally distributed there is a slight negative skew. Fortunately, we have a very large sample size and can assert the central limit theorem that the residuals are normally distributed.

Interpretation of Coefficients

Based upon our analysis of model1, we can say that our model meets the classic linear model assumptions except as an exogenous model that needs to be adjusted for heterskedastic errors. Let's take a look now at our coefficients and interpret their significance.

```
# calculate the coefficients with heterskedastic-sensitive methods
coeftest(model1, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  3215.86      56.19 57.2321 < 2.2e-16 ***
## avg.npvis     124.68      35.31  3.5311 0.0004245 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output above tells us several things about our model. First, that our variable `avg.npvis` and the y-intercept are highly statistically significant, with very low p-values. Second, it says that our heterskedastic standard errors are relatively small. Third, that if the average number of prenatal visits per month is increased by one value, we are likely to see an increase of $125g$ in birthweight. Our mean birthweight in this sample is 3401.

```
mean(Wellness$bwght, na.rm = T)
```

```
## [1] 3401.208
```

This means that increasing the average number of prenatal visits per month by one visit per month during pregnancy is associated with increases the birthweight of the baby by approximately 4%.

This is helpful to understand as the practical significance of this finding is not what we might want it to be. Potentially, it would take a significant increase in the average number of prenatal visits per month to have a meaningful effect on the wellness of the baby.

Model 2 - Covariates we believe increase the accuracy of your results without introducing bias.

In our second model, we've introduced variables relevant to the mother which we believe may be predictors of birthweight, our proxy for newborn wellness. We contiously did not include variables we thought may be subject to bias.

$$bwght = \beta_0 + \beta_1 avg.npvis + \beta_2 mage + \beta_3 magesq + \beta_4 mwhte + \beta_5 fwhte + \beta_6 mfwhte$$

Mother's Age ($mage$ and $magesq$)

Mothers in the teens or over 40 years old may be at higher risk for delivering low weight babies (http://www.amhsjournal.org/article.asp?issn=2321-4848;year=2013;volume=1;issue=1;spage=33;epage=37;aulast=Aras). As such, we included both the mother's age and its squared term to try and represent a more realistic model where age can be positively correlated with birthweight, but then have a negative correlation after a certain point.

Parent Ethnicity ($fwhte$, $mwhte$, and $mfwhte$)

Since access to healthcare can be correlated to income and ethnicity, we felt it prudent to include the indicator values for the parents' ethnicities with our base case being that both parents are white. In our sample we noted that nearly 88% of couples were both white. We included the $mfwhte$ indicator variable to capture any interaction when both parents were white. Given our available variables, ethnicity may be our few proxies to roughly gauge factors such as income or access to healthcare and quality of healthcare, which lie outside our dataset.

```r
# Create mother & father ethnicity interaction indicator
Wellness$mfwhte = ifelse(Wellness$mwhte == 1 & Wellness$fwhte == 1, 1, 0)
model2 = lm(bwght ~ avg.npvis + mage + magesq + fwhte + mwhte + mfwhte, data = Wellness)
```

MLR.1 & MLR.2

None of our earlier Model 1 assumptions in regards to linearity or random sampling have changed with the addition of the age and ethnicity variables.

MLR.3 No Perfect Multicollinearity Between Predictor Variables

With our added variables we can take advantage of the Variance Inflation Factor (VIF) test.

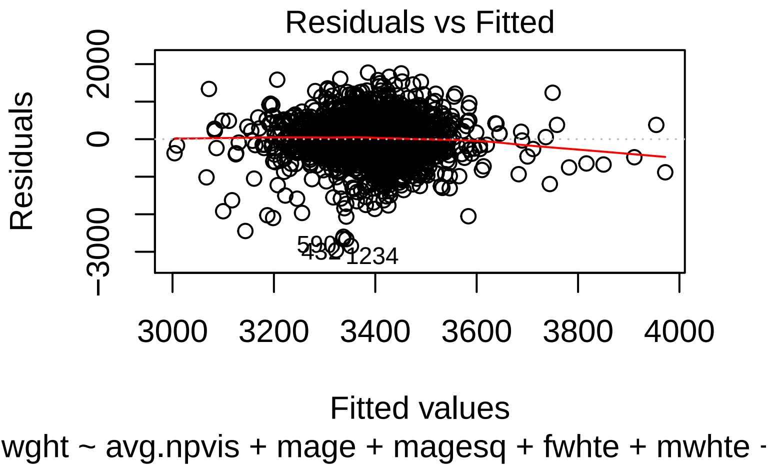
```r
vif(model2)
```

```
## avg.npvis      mage    magesq     fwhte     mwhte    mfwhte
##  1.003547 83.657026 83.650394  8.513103 13.881496 23.148021
```

Unsurprisingly, multicollinearity is high because we have included `magesq` which is a squared term of `mage`. Also, we have indicator variables for ethnicity with an indicator term which introduces multicollinearity. As such, we feel comfortable with the high VIF values. We would not expect much collinearity between ethnicity and age.

MLR.4 & 4' Zero-Conditional Mean and Exogeneity

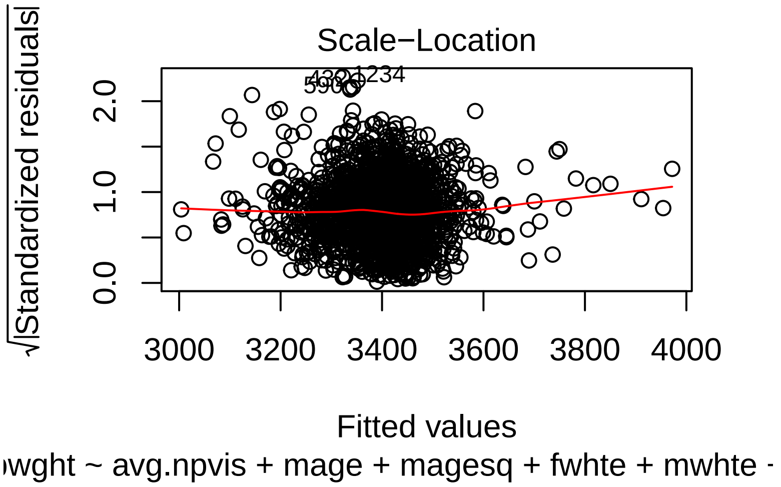
```
plot(model2, which = 1)
```



There is some improvement in the plot compared to Model 1, although we are still not in a situation to accept the proposition of zero-conditional mean. Again, we continue to rely on the assumption of exogeneity and would view our model as one of association versus causation.

MLR.5 Homoskedasticity

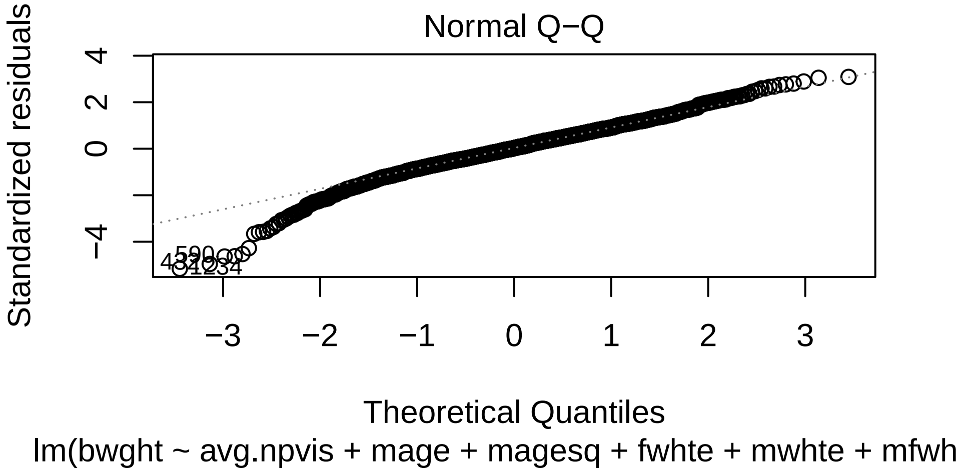
```
plot(model2, which = 3)
```



The results of our plot are very similar to that in the first model, and we will need to rely on robust estimators to heteroskedasticity.

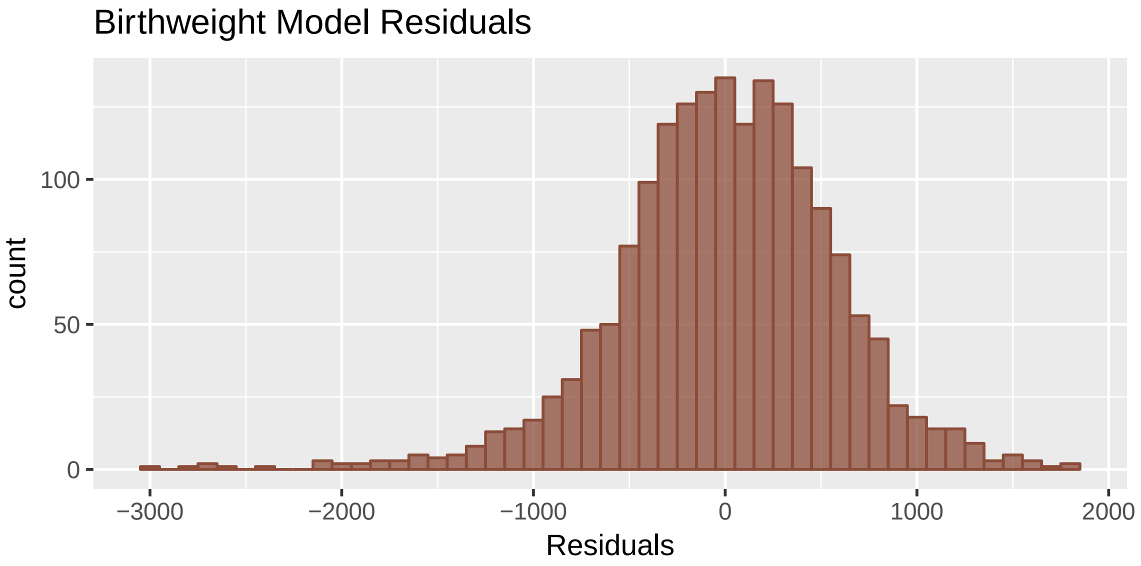
MLR.6 Normality of the Error Terms

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



Our second model's Q-Q plot exhibits the same deviations at the margins as the first model did.

```r
qplot(model2$residuals,
      geom = "histogram", binwidth = 100,
      main = "Birthweight Model Residuals",
      xlab = "Residuals",
      col = I("salmon4"), fill = I("salmon4"), alpha = I(.75))
```



The histogram of Model 2 residuals is also very similar to the Model 1 histogram. Again, with the large sample size, we can assert the central limit theorem and claim that normality for our residual values.

Interpretation of Coefficients

```r
coeftest(model2, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 1653.69177  433.40883  3.8155 0.0001406 ***
## avg.npvis    132.41593   35.30975  3.7501 0.0001825 ***
## mage          94.96550   28.50123  3.3320 0.0008802 ***
## magesq        -1.52675    0.46835 -3.2598 0.0011362 **
## fwhte        215.23059  123.58764  1.7415 0.0817674 .
## mwhte        117.16278  141.79229  0.8263 0.4087471
## mfwhte      -208.34894  184.36453 -1.1301 0.2585918
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Our avg.npvis variable continues to be statistically significant in Model 2, with a p-value that has dropped from Model 1. The coefficient is slightly higher and predicts an increase in $132.4g$ in birthweight for each unit increase in the average number of monthly prenatal visits. As this relates to a 4% change in birthweight, we feel this also has valid practical significance.

The `mage` and `magesq` coefficients need to be taken together in context. Because of the squared term, newborn weight will increase with mother's age to a point, at which further increases in age will predict a lower birthweight. Given the two coefficients here, we expect to see a negative correlation between birthweight and mother's age after the age of 31 years old. We can check the joint significance of both terms using the linear hypothesis test:

```
linearHypothesis(model2, c("mage = 0", "magesq = 0"), vcov = vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## mage = 0
## magesq = 0
##
## Model 1: restricted model
## Model 2: bwght ~ avg.npvis + mage + magesq + fwhte + mwhte + mfwhte
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F   Pr(>F)
## 1   1756
## 2   1754  2 5.6901 0.003442 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Our p-value shows that together the mother's age is statistically significant. Depending on the mother's age, the practical effect can change, but still remains significant at reasonable age estimates.

Finally, whether one or both parents are white did not individually appear significant in the previous coefficient output, but we can check for joint significance like we did for mother's age:

```
linearHypothesis(model2, c("fwhte = 0", "mwhte = 0", "mfwhte = 0"), vcov = vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## fwhte = 0
## mwhte = 0
## mfwhte = 0
##
## Model 1: restricted model
## Model 2: bwght ~ avg.npvis + mage + magesq + fwhte + mwhte + mfwhte
##
## Note: Coefficient covariance matrix supplied.
```

```
##
##   Res.Df Df     F  Pr(>F)
## 1   1757
## 2   1754  3 3.021 0.02873 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From this test we see that taken together, having one or both parents white is statistically significant to birthweight. The coefficients from earlier also require all three terms to be taken into context. Having just a white father alone predicts a $215g$ increase in birthweight, while having a white mother alone predicts a $117g$ increase. The negative coefficient for the interaction term needs to be netted against the earlier two individual mother and father ethnicity terms if both parents are white. This results in a net $124g$ predicted increase in birthweight for two white parents, a practically significant result as well.

Model 1 vs Model 2: R-Squared & AIC

Number	Adj. R-Squared	AIC
Model 1	0.0098673	27399.09
Model 1	0.0200819	27385.81

Our second model's adjusted r-squared value increased by 0.01 from 0.01 to 0.02 showing that the second model is able to account for a greater change in birthweight compared to our first model. The AIC value also dropped from $27,399$ to $27,386$ under our second model indicating that even with the added complexity in model 2, we are not losing goodness of fit.

Model 3 - Problematic covariates (measurement error/bias)

In our third model, we will add additional covariates to our model knowing that they will cause problems and look specifically at how issues with our data can lead to measurement error and increased bias from potential underreporting.

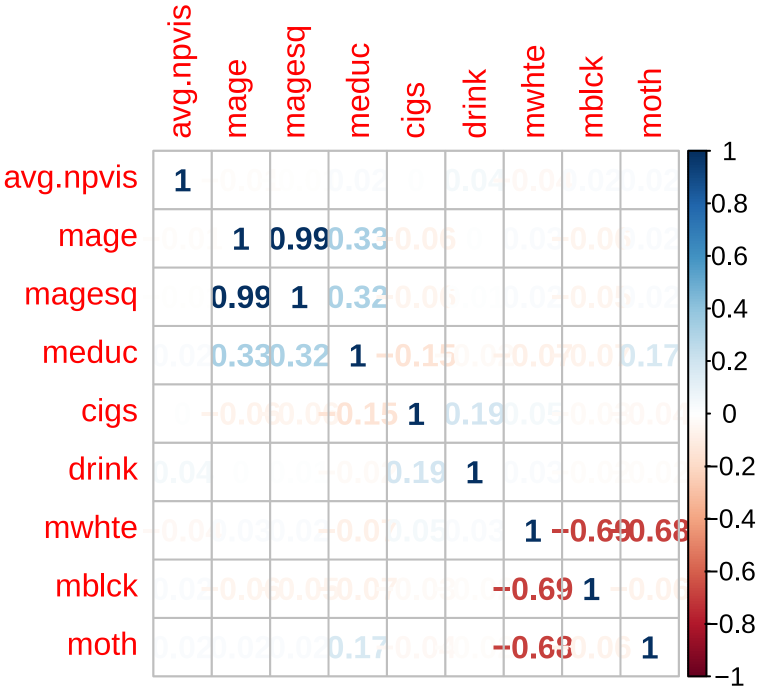
$$bwght = \beta_0 + \beta_1 avg.npvis + \beta_2 mage + \beta_3 magesq + \beta_4 mwhte + \beta_5 fwhte + \beta_6 mfwhte + \beta_7 meduc + \beta_8 cigs + \beta_9 drink + \beta_9 moth$$

Cigarettes & Drinking ($cigs$ & $drink$)

Cigarettes have been known to be correlated with low birthweight (https://www.ncbi.nlm.nih.gov/pubmed/14976791) and seem like a very likely predictor for birthweight in our model. Although there are 109 NA values for *cigs*, this amount seems immaterial given our sample size of 1,832. Of the remaining observations, 147 mothers reported that they had smoked. There is potentially an under-reporting bias if women are less-inclined to admit to smoking during pregnancy. If we assume this is true and that there is a negative correlation between increased smoking and birthweight, we may need to concede that our coefficient for *cigs* may be lower than if no reporting bias existed.

Drinking appears to be less correlated with birthweight, but may cause other types of problems such as birth defects or mental defects (http://www.upmc.com/patients-visitors/education/pregnancy/Pages/smoking-alcohol-and-drugs-can-harm-your-baby.aspx). These may not be well-captured in looking at birthweight, however including the variable in our model may help to reduce error variance if the variable is not highly correlated with our other explanatory variables.

```
par(mfrow=c(1,1))
m = cor(Wellness[ , c("avg.npvis", "mage", "magesq", "meduc", "cigs",
                             "drink", "mwhte", "mblck", "moth")], use = "complete.obs")
corrplot(m, method = "number")
```

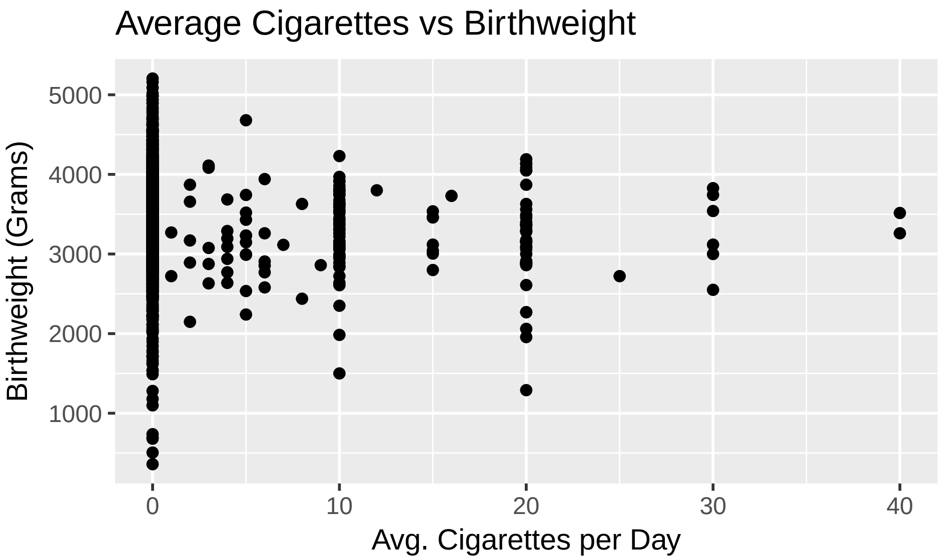


Drinking is not highly correlated with any of our other explanatory variables, and the decreased reduced error variance in our model may be advantageous. We also noted 114 missing values, although 106 of these overlap those with missing cigarette data. The number of missing values may be of concern.

Measurement error

We should also note the potential for measurement error in our `cigs` variable. We will plot `cigs` and `bwght`.

```
qplot(Wellness$cigs, Wellness$bwght,
     main = "Average Cigarettes vs Birthweight",
     xlab = "Avg. Cigarettes per Day",
     ylab = "Birthweight (Grams)")
```



The graph above shows a significant amount of clustering. This could be due to two factors: an aspect of the data with a strong *graduation effect* or *attenuation bias*. In this case, we see a significant number of respondents chose round numbers $(0, 10, 20)$ as well as numbers ending in $5(15, 25)$. This could be the result of a *graduation effect* around how many cigarettes there are in a pack. A pack of cigarettes contains 20 cigarettes. If a mother smokes a pack a day, the data could cluster at that point. Similar logic can hold for 10 (half-pack), 30 (a pack and a half) and 40 (two-packs).

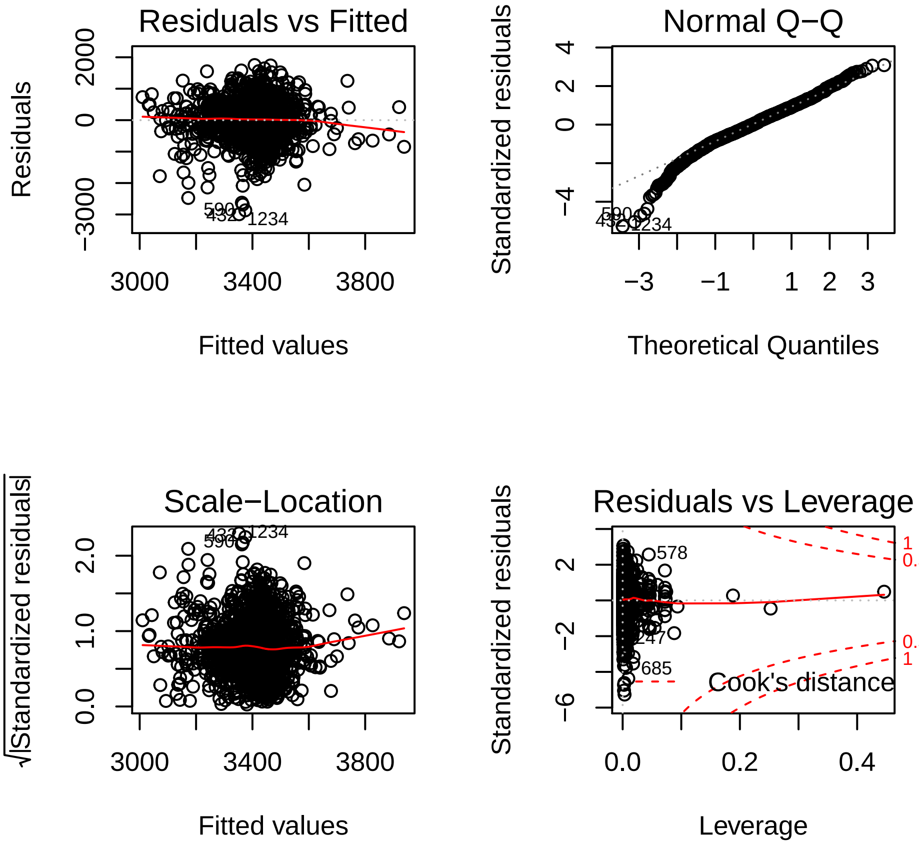
Another explanation is that these numbers represent easy to remember rounded figures. This can introduce *attenuation bias* and may lead to biased OLS regression.

Regardless of the source of the clustering, we felt that it represented a form of measurement bias. OLS regression is not robust to measurement bias and so we excluded cigs from our primary model.

We now can setup our third (biased) model and explore is our OLS model assumptions are met.

```
model3 = lm(bwght ~ avg.npvis + mage + magesq + fwhte + mwhte + mfwhte + cigs + drink, data = Wellness)
```

```
par(mfrow=c(2,2)) # Print 4 plots in one
plot(model3)
```



Again, we assume linear population model and random sampling. The newly added `cigs` and `drink` variables do not show a VIF over 4 indicating no additional issues with our multicollinearity assumption. We note that the residuals vs. fitted plot shows a fairly flat smoothing line, indicating no obvious violation of zero-conditional mean. There does, however, look to be a strong indication of heteroskedasticity, so we will be sure to use heteroskedastic-robust standard errors. Lastly, our Q-Q plot indicates deviation from normality in our error terms. As noted in our previous models we have a large sample size, so the central limit theorem (CLT) implies that OLS coefficients have a normal sampling distribution. Lastly, we note that we have some observations with high leverage, but none of them with a Cook's distance larger than 1, therefore not causing major concern.

```r
summary(model3)$adj.r.squared #Adjusted R-squared term
```

```
## [1] 0.02194281
```

```r
AIC(model3)
```

```
## [1] 25550.37
```

```r
coeftest(model3, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 1998.00927  441.02038  4.5304 6.313e-06 ***
## avg.npvis    119.71322   34.67799  3.4521 0.0005703 ***
## mage          73.46256   29.11328  2.5233 0.0117188 *
## magesq        -1.15940    0.48014 -2.4147 0.0158561 *
## fwhte        213.84313  121.47879  1.7603 0.0785382 .
## mwhte        121.84900  132.93712  0.9166 0.3594920
## mfwhte      -209.18032  175.59758 -1.1912 0.2337289
## cigs         -11.48063    3.67298 -3.1257 0.0018050 **
## drink        -18.23196   32.40267 -0.5627 0.5737378
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
vif(model3)
```

```
## avg.npvis      mage    magesq     fwhte     mwhte    mfwhte      cigs
##  1.006590 84.017433 83.950655  7.998643 13.032119 21.795905  1.046037
##     drink
##  1.038783
```

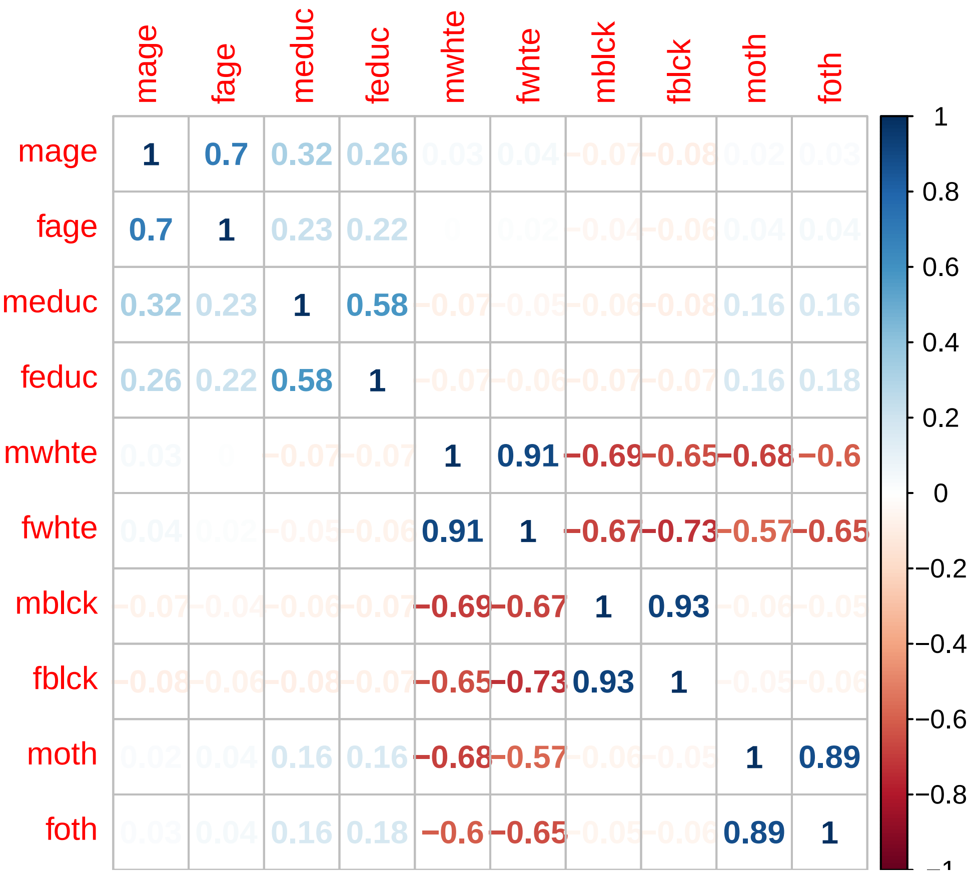
Although our diagnostics appear good, we still felt the earlier *graduation effect* and *attenuation bias* issues with `cigs` increase the risks of problems for our model. Additionally, the numbers of cigarettes and drinks may be underreported by respondents who may be unwilling to admit they had smoked or drank during pregnancy, or perhaps had toned down their response. The risk that a person would overestimate or claim to have smoked or drank more is probably very low. Underreporting would create a bias in our data and where some lower birthweights may not be properly associated with actual smoking and drinking. In such a case we would expect our coefficients for `cigs` especially to be too high and there biased.

Model 4 - Problematic covariates (multicollinear variables)

Next, we will introduce variables into our model which may cause issues with multicollinearity. We know that R will throw an error if we are using variables where perfect multicollinearity exists. We must also be mindful of imperfect collinearity. If it exists, it increases the standard error of our model.

We will look for variables highly correlated with variables already in our model, add them to a new model and assess the Variance Inflation Factor (VIF), which would indicate how much the standard error of each coefficient is inflated due to collinearity with other variables. We will look to see if VIF increases over 4 and 10 (the rules of thumb VIF for when to worry about multicollinearity.)

```
par(mfrow=c(1,1))
multicollcor = cor(Wellness[ , c("mage", "fage", "meduc", "feduc", "mwhte", "fwhte", "mblck", "fblck",
corrplot(multicollcor, method = "number")
```



It is clear that we have strong correlations between the race of the mother and father (dummy) variables. It is also important to note a positive correlation between mother and father's age and education. There is also correlation between age and education. Considering we are attempting to create issues in our model we will add pairs of variables showing high corrleation.

First we should note the VIF of our other models.

```
vif(model2)
```

```
## avg.npvis      mage    magesq     fwhte     mwhte    mfwhte
##  1.003547 83.657026 83.650394  8.513103 13.881496 23.148021
```

```
vif(model3)
```

```
## avg.npvis      mage    magesq     fwhte     mwhte    mfwhte      cigs
##  1.006590 84.017433 83.950655  7.998643 13.032119 21.795905  1.046037
##     drink
##  1.038783
```

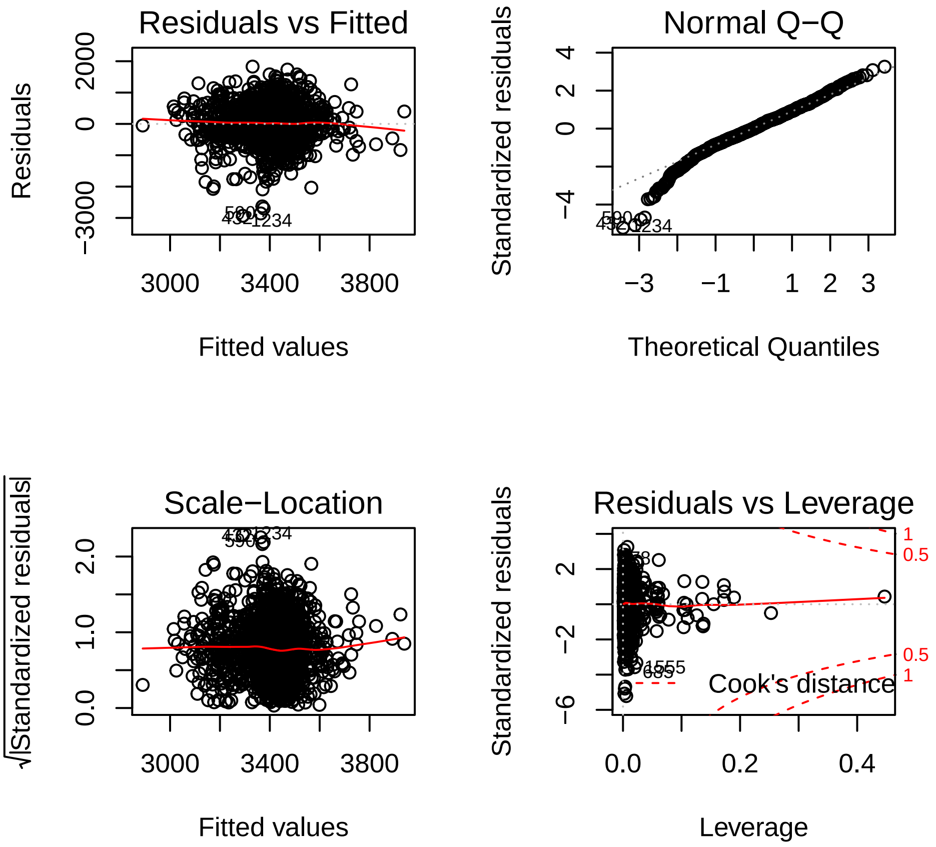
We note that all the values are within normal bounds except for the variables we discussed in model2. We are not concerned about these values as it is expected.

We will now setup the fourth model, including the correlated variables.

$$bwght = \beta_0+\beta_1 avg.npvis+\beta_2 mage+\beta_3 magesq+\beta_4 mwhte+\beta_5 fwhte+\beta_6 mfwhte+\beta_7 meduc+\beta_8 cigs+\beta_9 drink+\beta_9 moth+\beta_{10} f$$

```
model4 = lm(bwght ~ avg.npvis + mage + magesq + fwhte + mwhte
            + mfwhte + cigs + drink + moth + fage + feduc
            + meduc + foth, data = Wellness)
```

```
par(mfrow=c(2,2)) #Set up plots to print in single image
plot(model4)
```



```
summary(model4)$adj.r.squared #Adjusted R-squared term
```

```
## [1] 0.02789615
```

```
AIC(model4)
```

```
## [1] 25026.32
```

```
coeftest(model4, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 1901.63014  455.61717  4.1737 3.158e-05 ***
## avg.npvis    122.49375   34.42144  3.5586 0.0003837 ***
## mage          71.07590   30.52700  2.3283 0.0200205 *
## magesq        -1.22360    0.49703 -2.4618 0.0139289 *
## fwhte       -119.77559  214.52334 -0.5583 0.5766948
## mwhte        221.89493  157.52262  1.4087 0.1591318
## mfwhte       -96.20036  222.42105 -0.4325 0.6654257
## cigs          -9.43517    3.63255 -2.5974 0.0094796 **
## drink        -21.05836   31.99533 -0.6582 0.5105239
## moth         345.64158  206.52359  1.6736 0.0944013 .
## fage           6.46098    3.52780  1.8314 0.0672197 .
## feduc         10.82696    8.22438  1.3164 0.1882127
## meduc         -1.34580    8.69219 -0.1548 0.8769762
## foth        -613.76499  206.11434 -2.9778 0.0029470 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Many of our observations from the plots are similar to model3. The linear population model and random sampling are assumed. We note that the residuals vs. fitted plot shows a fairly flat smoothing line, indicating no obvious violation of zero-conditional mean. There does however look to be strong indication of heteroskedasticity, so we will be sure to use heteroskedasticity-robust standard errors. Our Q-Q plot indicates deviation from normality, but with a large sample size the CLT implies that OLS coefficients have a normal sampling distribution. Again, we note that we have some observations with high leverage, but none of them with a Cook's distance larger than 1, therefore not causing major concern.

Lastly, since we have added in variables with high correlation to this model, we will look at the VIF for this model to test if collinearity exists.

```
vif(model4)
```

```
## avg.npvis      mage    magesq     fwhte     mwhte    mfwhte      cigs
##  1.010620 88.115686 85.601206 21.065384 17.890005 30.479107  1.072706
##     drink      moth      fage     feduc     meduc      foth
##  1.041768 12.926568  1.920630  1.595143  1.676858 11.593286
```

Because we have squared terms and dummy indicators, there are higher VIF numbers than otherwise might be normal. Although it can be difficult to test multicollinearity directly, the increasing standard errors for our variables help to show the increased variance we've added to our model by throwing in new variables.

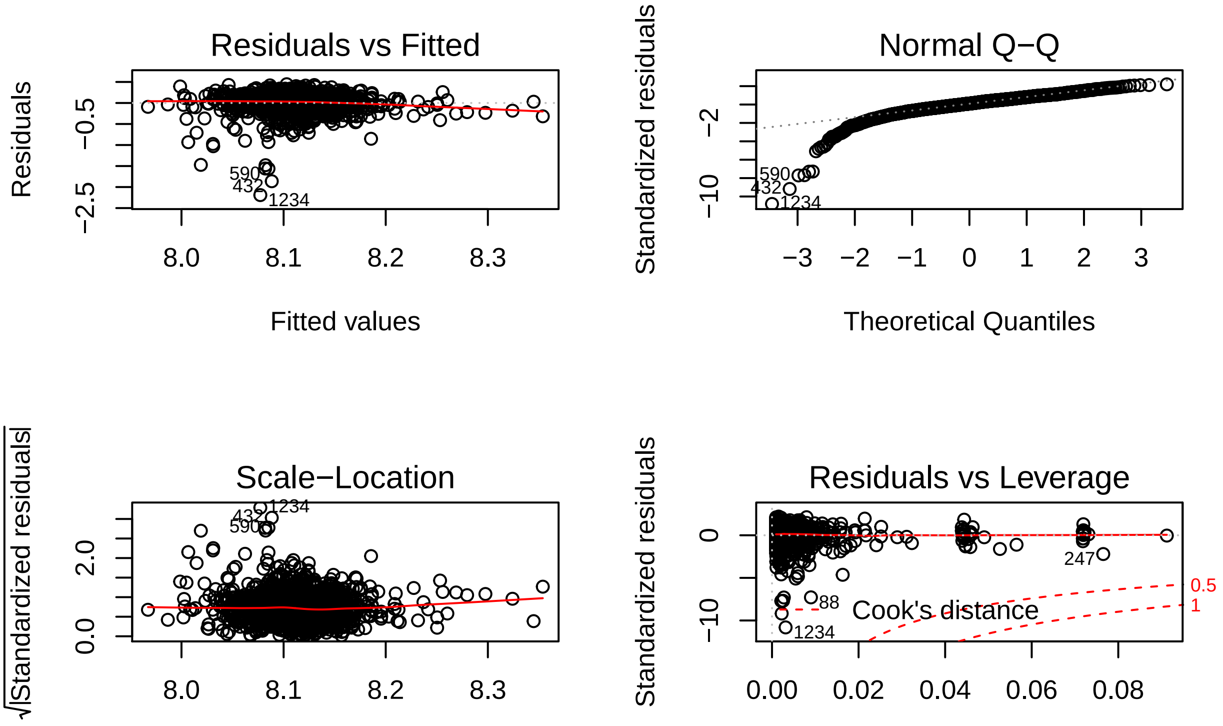
Model 5 - Changing the Dependant Variable

Until this point we've only focused on the birthweight of the child as an indicator of health. We know this is actually just one part of many variables which can be used to determine health. The only other potnetial dependent variables available to us are are APGAR scores. We noted this in our exploritory section. We can take a different approach here by transforming our dependent variable and evaluating the log of birthweight. This makes logical sense as the new model will show changes in our independent variables in terms of the percent increase in the dependent variable, birthweight.

```
model5 = lm(lbwght ~ avg.npvis + mage + magesq + mwhte + fwhte + mfwhte, data = Wellness)
```

We will then take a look at our plots for this model.

```
par(mfrow=c(2,2)) #Set up plots to print in single image
plot(model5)
```



With this new model we see some changes between this model and Model 2 which regressed `bwght`. We still have non-normal errors for low birthweight, which can be addressed by applying the central limit theorem. In general, this new model has tighter best-fit lines and less heterskedasticity than before. We noted a 5.5% increase in birthweight under this new model for an increase of one prenatal visit on average per month, quite a large change from our earlier prediction of 4%. We note that since we are looking at the log of weight this % change can only be evaluated for small changes in our independent variables.

```
summary(model5)$adj.r.squared
```

```
## [1] 0.02282618
```

```
AIC(model5)
```

```
## [1] -610.2278
```

```
coeftest(model5, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  7.51759749  0.15763459 47.6900 < 2.2e-16 ***
## avg.npvis    0.05555724  0.01420907  3.9100 9.582e-05 ***
## mage         0.03147388  0.01032073  3.0496  0.002326 **
## magesq      -0.00050174  0.00016826 -2.9819  0.002904 **
## mwhte        0.04163318  0.04585753  0.9079  0.364066
## fwhte        0.07047782  0.03672774  1.9189  0.055156 .
## mfwhte      -0.07562100  0.05712562 -1.3238  0.185753
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The change due to the mother's age varies due to the squared term, and more extreme values that may be unreasonable are seen at the low and high margins for age. Our linear hypothesis test on `mage` and `magesq` also shows the that the two terms continue to exhibit joint significance:

```
linearHypothesis(model5, c("mage = 0", "magesq = 0"), vcov = vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## mage = 0
## magesq = 0
##
## Model 1: restricted model
## Model 2: lbwght ~ avg.npvis + mage + magesq + mwhte + fwhte + mfwhte
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F   Pr(>F)
## 1   1756
## 2   1754  2 4.8044 0.008302 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The coefficients for `fwhte`, `mwhte`, and `mfwhte` appear very high, especially the predicted 7% increase in birthweight for having only a white father. A linear hypothesis test shows that these terms actually do not exhibit joint significance compared to our earlier Model 2. Along with the large coefficient for `fwhte`, this may give us pause about including these variables in our model.

```
linearHypothesis(model5, c("mwhte = 0", "fwhte = 0", "mfwhte = 0"), vcov = vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## mwhte = 0
## fwhte = 0
## mfwhte = 0
##
## Model 1: restricted model
## Model 2: lbwght ~ avg.npvis + mage + magesq + mwhte + fwhte + mfwhte
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F  Pr(>F)
## 1   1757
## 2   1754  3 2.2328 0.08256 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model Summary (regression table)

Finally, to compare our five models we will create a regression table using stargazer. We will use the heteroskedasticity-robust (Huber-White) standard errors as all of our models seemed to indicate some form of heteroskedasticity.

```
# Setup robust standard errors
se.model1 = sqrt(diag(vcovHC(model1)))
se.model2 = sqrt(diag(vcovHC(model2)))
```

```r
se.model3 = sqrt(diag(vcovHC(model3)))
se.model4 = sqrt(diag(vcovHC(model4)))
se.model5 = sqrt(diag(vcovHC(model5)))

stargazer(model1, model2, model3, type="latex",
          se = list(se.model1, se.model2, se.model3),
          star.cutoffs = c(0.05, 0.01, 0.001),
          object.names = TRUE,
          model.numbers = FALSE)
```

% 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 - 21:23:11

Table 1:

	Dependent variable:		
	bwght		
	model1	model2	model3
avg.npvis	124.682***	132.416***	119.713***
	(35.310)	(35.310)	(34.678)
mage		94.965***	73.463*
		(28.501)	(29.113)
magesq		−1.527**	−1.159*
		(0.468)	(0.480)
fwhte		215.231	213.843
		(123.588)	(121.479)
mwhte		117.163	121.849
		(141.792)	(132.937)
mfwhte		−208.349	−209.180
		(184.365)	(175.598)
cigs			−11.481**
			(3.673)
drink			−18.232
			(32.403)
Constant	3,215.863***	1,653.692***	1,998.009***
	(56.190)	(433.409)	(441.020)
Observations	1,761	1,761	1,645
R^2	0.010	0.023	0.027
Adjusted R^2	0.010	0.020	0.022
Residual Std. Error	577.863 (df = 1759)	574.875 (df = 1754)	568.952 (df = 1636)
F Statistic	18.539*** (df = 1; 1759)	7.011*** (df = 6; 1754)	5.610*** (df = 8; 1636)
Note:			*p<0.05; **p<0.01; ***p<0.001

See table 2 at the end of the report.

```r
stargazer(model2, model4, model5, type="latex",
          se = list(se.model2,se.model4, se.model5),
          star.cutoffs = c(0.05, 0.01, 0.001),
          object.names = TRUE,
          model.numbers = FALSE)
```

% 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 - 21:23:12

See table 3 at the end of the report.

Causal Inference

It is important to note that a well-fit model does not necessarily indicate that our independent variables in any way cause the dependent variable. In order to test for a causal relationship we must account for all variables that may impact the health of an infant. This is known as the *ceteris paribus* assumption which say that all other factors go into our error term u. We can think of many other variables not represented in our dataset: mother's stress, exposure to sun, fitness level, sleep, allergies. The list could continue but the important thing to note is that in order to find the *true* causal inference of prenatal care on a newborns health we would, in theory, need all of these variables in our model. Only then would our $\beta_prenatalcare$ represent a causal relationship.

We must also consider omitted variable bias on our analysis. In our analysis the β value for our key variable `avg.npvis` has a positive sign in relationship to our dependent variable of birthweight. In order to examine sign of omitted variable bias we would also need to consider the sign of γ_1. γ we understand to be the the regression of our omitted variable on our estimated model. This is easy to examine in our first model, `model1`, as `avg.npvis` is our only independent variable. In our multiple regression examples we would need to consider the entire regression of the omitted variable on the other variables.

One way to more accurately determine a causal relationship would be to run a *true experiment* where mothers are randomly selected to receive prenatal care while others receive no (or less) prenatal care. If conducting this experiment were possible we would assume to have exogeneity. Omitted variable bias would no longer be of concern. Conducting a study like this obviously would raise some ethical concerns and would not be practical.

Conclusion

Our main objective of this lab was explore relationships between prenatal care and the health of a newborn. Our main indicator for prenatal health was `avg.npvis`, the average visits per month, starting at the month of the first visit. Our main independent variable for health of a newborn was birthweight. Over the course of our analysis we explored covariates which improved the accuracy of our model and some that hurt model accuracy. We also discussed any causal inference that could be taken from our analysis.

Our first model, $model1$, was relatively simple using our computed `avg.npvis` variable as the only indicator for birthweight. This model met all of our OLS assumptions with the exception of homoskedasticity, causing us to use heteroskedastic-sensitive standard errors. The first model showed a relationship where an increase of one average number of prenatal visits per month after first visit correlated to an increase of 125g in birthweight. Our second model added various other variables including mother's age, and the parent's ethnicity. In this second model we found a slightly higher coefficient with increase in 132.4g in birthweight for each unit increase in the average number of monthly prenatal visits.

Table 2:

	Dependent variable:		
	bwght		lbwght
	model2	model4	model5
avg.npvis	132.416***	122.494***	0.056***
	(35.310)	(34.421)	(0.014)
mage	94.965***	71.076*	0.031**
	(28.501)	(30.527)	(0.010)
magesq	-1.527^{**}	-1.224^{*}	-0.001^{**}
	(0.468)	(0.497)	(0.0002)
fwhte	215.231	−119.776	0.070
	(123.588)	(214.523)	(0.037)
mwhte	117.163	221.895	0.042
	(141.792)	(157.523)	(0.046)
mfwhte	−208.349	−96.200	−0.076
	(184.365)	(222.421)	(0.057)
cigs		-9.435^{**}	
		(3.633)	
drink		−21.058	
		(31.995)	
moth		345.642	
		(206.524)	
fage		6.461	
		(3.528)	
feduc		10.827	
		(8.224)	
meduc		−1.346	
		(8.692)	
foth		-613.765^{**}	
		(206.114)	
Constant	1,653.692***	1,901.630***	7.518***
	(433.409)	(455.617)	(0.158)
Observations	1,761	1,613	1,761
R^2	0.023	0.036	0.026
Adjusted R^2	0.020	0.028	0.023
Residual Std. Error	574.875 (df = 1754)	563.302 (df = 1599)	0.203 (df = 1754)
F Statistic	7.011*** (df = 6; 1754)	4.558*** (df = 13; 1599)	7.852*** (df = 6; 1754)
Note:			*p<0.05; **p<0.01; ***p<0.001

We introduced problem variables into our `model3` and `model4` knowing the impacts of bias and multicollinearity would cause issues for the accuracy of our model. Finally, in our `model5` we evaluated the impacts on the log of birthweight. What we found was a 5.5% increase in birthweight under this new model for an increase of one prenatal visit on average per month, which was quite a large change from our earlier prediction of 4%.

In summary, we did find a statistically significant relationship between the average number of prenatal visits after first receiving care, and the birthweight of the child. This relationship is somewhere between 4% and 5.5% increase in birthweight per average visit. We believe this to be a practically significant finding.

Recommendation

Our primary task for this lab was to investigate whether there was any predictive relationship between prenatal care and the wellness of a newborn child. As discussed throughout this report, there are numerous reasons this relationship is difficult to establish with the given dataset. We would recommend that further pursuit of this research goal include an expanded dataset with more variables as those suggested earlier, and also detailed documentation on the methodology behind how the sample is taken. Although these recommendations are non-trivial, these two items would prove critical to a more thorough future analysis.