Lab 4: Does Prenatal Care Improve Infant Health?

Bryan Moore, Ryan Delgado, Cameron Bell

April 26, 2017

Introduction

Over the last several decades, health advocacy groups have stressed the importance of prenatal healthcare. Prenatal visits consist of interventions such as dietary counseling, education on avoidance of risk factors for peripartum infections, and smoking cessation or avoidance counseling. These are meant to improve the well-being of the newborns, and to set these neonates on a path for adolescent health.

Prenatal visits are the contemporary standard-of-care in medicine, but there is not a large body of evidence as to whether or not these interventions actually improve health outcomes for newborn infants.

In the study that follows, our group conducts an exploratory and statistical data analysis of the multiple variables that may affect the birth-weight for newborn infants. Our study will seek to answer whether prenatal care improves infant health. Our data includes a birth-weight variable, and one- and five-minute APGAR scores. We'll use birth-weight as the dependent variable in our models, as it is the only indicator of newborn health on which we can run a regression. We will also discuss the five-minute APGAR scores, ignoring the 1-minute APGAR because the final effect is of more interest to us.

The study will also include an evaluation of regression assumptions for our primary models, and will also explore the benefits or problems of including additional covariates in our model.

The study will proceed as follows: Section 1 will explore the dependent variables, explanatory variables of key interest, and covariates that could either enhance the predictiveness of our model or be problematic. Section 2 will specify, fit, and discuss the regression assumptions of our proposed linear model, and analyze two additional linear models. The first additional model will seek to add covariates to our original model that increase the accuracy of our results without introducing bias. The second additional model will seek to add covariates to our original model that may be problematic for one reason or another. Section 3 concludes our analysis with high-level takeaways.

1. Initial Exploratory Analysis of Variables

A complete description of the variables in our data set:

```
options(warn=-1)  # Turn off warnings
library(ggplot2)
library(car)
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(sandwich)
library(gridExtra)
library(reshape2)
```

```
library(GGally)
library(lsr)
library(stargazer)
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

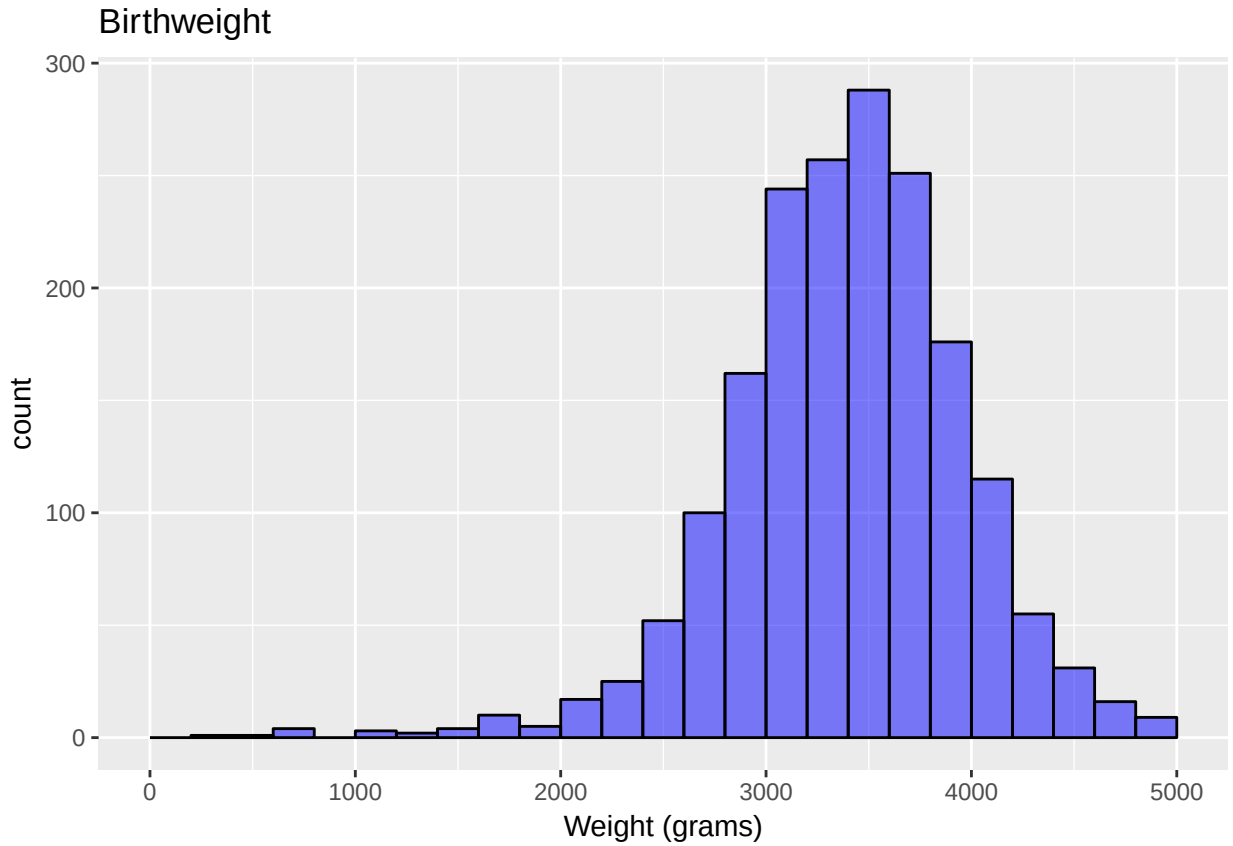
```
##  R package version 5.2. http://CRAN.R-project.org/package=stargazer
```

```
load("bwght_w203.RData")
desc
```

```
##    variable                          label
## 1      mage             mother's age, years
## 2     meduc            mother's educ, years
## 3    monpre       month prenatal care began
## 4     npvis total number of prenatal visits
## 5      fage             father's age, years
## 6     feduc            father's educ, years
## 7     bwght             birth weight, grams
## 8     omaps          one minute apgar score
## 9     fmaps         five minute apgar score
## 10     cigs          avg cigarettes per day
## 11    drink             avg drinks per week
## 12      lbw             =1 if bwght <= 2000
## 13     vlbw             =1 if bwght <= 1500
## 14     male                 =1 if baby male
## 15    mwhte              =1 if mother white
## 16    mblck              =1 if mother black
## 17     moth           =1 if mother is other
## 18    fwhte              =1 if father white
## 19    fblck              =1 if father black
## 20     foth           =1 if father is other
## 21   lbwght                     log(bwght)
## 22   magesq                         mage^2
## 23  npvissq                       npvis^2
```

Let's start by examining our outcome variable, bwght. We'll plot its histogram and analyze it:

```
bwght.hist <- qplot(data$bwght, geom='histogram',
      binwidth=1, fill=I('blue'), col=I('black'),
      alpha=I(0.5), xlab='Weight (grams)',
      breaks=seq(0, 5000, by = 200),
      main='Birthweight')

grid.arrange(bwght.hist, ncol=1)
```




```
summary(data[,c('bwght')])
```

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

Birth-weight has a slight negative skew, but is otherwise approximately normally distributed.

We see that there are some newborns with birth-weights less than 1000 grams, which may be indicative of prematurity. We'll look at the values for bwght, and other key variables of interest:

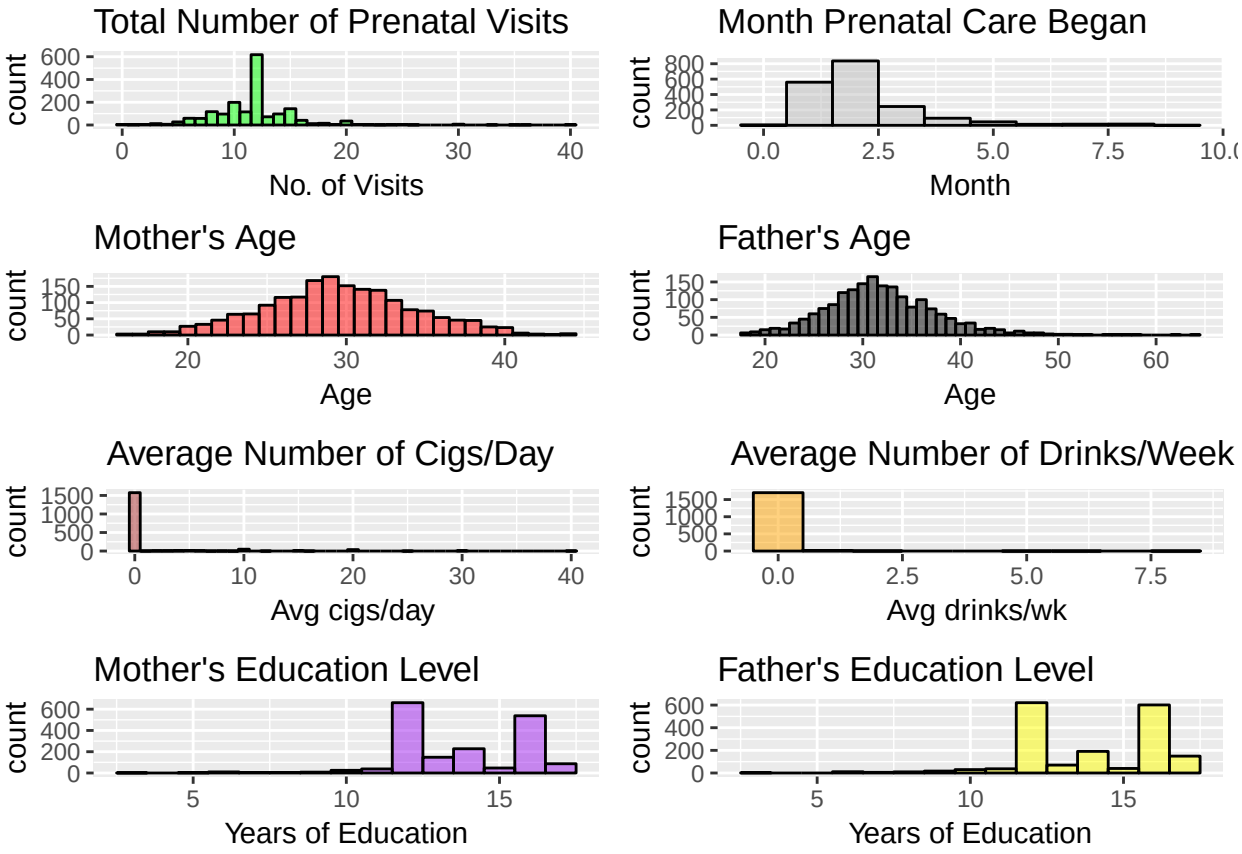
```
data[# Select only the rows from the datset where bwght < 1000
      data$bwght < 1000,

     # Select the bwght and variables that indicate prenatal care
      c('bwght', 'fmaps', 'npvis', 'monpre')]
```

```
##      bwght fmaps npvis monpre
## 88     697     7     2      5
## 432    506     6     6      2
## 590    680     5     7      1
## 1178   681     5     6      1
## 1234   360    NA     5      1
## 1558   737    NA     5      2
```

We don't see any inconsistencies here, and no reason to remove any of these observations. It looks like these are just the sickest / least healthy babies, representing the far-left of the normal distribution. The NA values for fmaps (which are also NA for the 1-minute APGAR) suggests a stillbirth. Those 5-minute APGARS line-up with those birth-weights, and represent babies born prematurely that are likely heading straight to the Neonatal ICU for closer observation or acute care, instead of to the regular Neonatal floor unit.

Let's next examine key metric variables of interest, and other covariates that could be predictive of the outcome variables: npvis, monpre, mage, meduc, fage, feduc, cig, and drink. We'll plot histograms of each variables and discuss the results:

```
npvis.hist <- qplot(data$npvis, geom='histogram',
      binwidth=1, fill=I('green'), col=I('black'),
      alpha=I(0.5), xlab='No. of Visits',
      main='Total Number of Prenatal Visits')

monpre.hist <- qplot(data$monpre, geom='histogram',
      binwidth=1, fill=I('grey'), col=I('black'),
      alpha=I(0.5), xlab='Month',
      main='Month Prenatal Care Began')

mage.hist <- qplot(data$mage, geom='histogram',
      binwidth=1, fill=I('red'), col=I('black'),
      alpha=I(0.5), xlab='Age',
      main='Mother\'s Age')

cig.hist <- qplot(data$cigs, geom='histogram',
      binwidth=1, fill=I('brown'), col=I('black'),
      alpha=I(0.5), xlab='Avg cigs/day',
      main='Average Number of Cigs/Day')

drink.hist <- qplot(data$drink, geom='histogram',
      binwidth=1, fill=I('orange'), col=I('black'),
      alpha=I(0.5), xlab='Avg drinks/wk',
      main='Average Number of Drinks/Week')

meduc.hist <- qplot(data$meduc, geom='histogram',
      binwidth=1, fill=I('purple'), col=I('black'),
      alpha=I(0.5), xlab='Years of Education',
      main='Mother\'s Education Level')

fage.hist <- qplot(data$fage, geom='histogram',
      binwidth=1, fill=I('black'), col=I('black'),
      alpha=I(0.5), xlab='Age',
      main='Father\'s Age')

feduc.hist <- qplot(data$feduc, geom='histogram',
      binwidth=1, fill=I('yellow'), col=I('black'),
      alpha=I(0.5), xlab='Years of Education',
      main='Father\'s Education Level')


grid.arrange(npvis.hist, monpre.hist, mage.hist, fage.hist, cig.hist, drink.hist, meduc.hist, feduc.his
```




```
summary(data[,c('npvis', 'monpre', 'mage', 'meduc', 'cigs', 'drink', 'fage', 'feduc')])
```

```
##      npvis           monpre           mage           meduc      
##  Min.   : 0.00   Min.   :0.000   Min.   :16.00   Min.   : 3.00  
##  1st Qu.:10.00   1st Qu.:1.000   1st Qu.:26.00   1st Qu.:12.00  
##  Median :12.00   Median :2.000   Median :29.00   Median :13.00  
##  Mean   :11.62   Mean   :2.122   Mean   :29.56   Mean   :13.72  
##  3rd Qu.:13.00   3rd Qu.:2.000   3rd Qu.:33.00   3rd Qu.:16.00  
##  Max.   :40.00   Max.   :9.000   Max.   :44.00   Max.   :17.00  
##  NA's   :68      NA's   :5                       NA's   :30     
##       cigs            drink            fage           feduc      
##  Min.   : 0.000   Min.   :0.0000   Min.   :18.00   Min.   : 3.00  
##  1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.:28.00   1st Qu.:12.00  
##  Median : 0.000   Median :0.0000   Median :31.00   Median :14.00  
##  Mean   : 1.089   Mean   :0.0198   Mean   :31.92   Mean   :13.92  
##  3rd Qu.: 0.000   3rd Qu.:0.0000   3rd Qu.:35.00   3rd Qu.:16.00  
##  Max.   :40.000   Max.   :8.0000   Max.   :64.00   Max.   :17.00  
##  NA's   :110      NA's   :115      NA's   :6       NA's   :47
```

Observations: * Mother's age look approximately normally distributed, and doesn't have a significant amount of skew. * Number of Prenatal visits is not normally distributed, is high concentrated at 12, but also has a non-trivial number of outliers. * Month prenatal care began is reported in gestational month. We see that it has a significant positive skew. This is an ordinal variable. * Cigarettes and alcohol consumption both have a significant positive skew and are not normally distributed. These are ordinal variables. * Mother's education is not normally distributed, and there seem to be "graduation effects" at the twelve-year (high-school) and sixteen-year (college) levels. This is an ordinal variable. * Father's age looks approximately normally distributed, and does not have a significant amount of skew. There is more variance, however, in

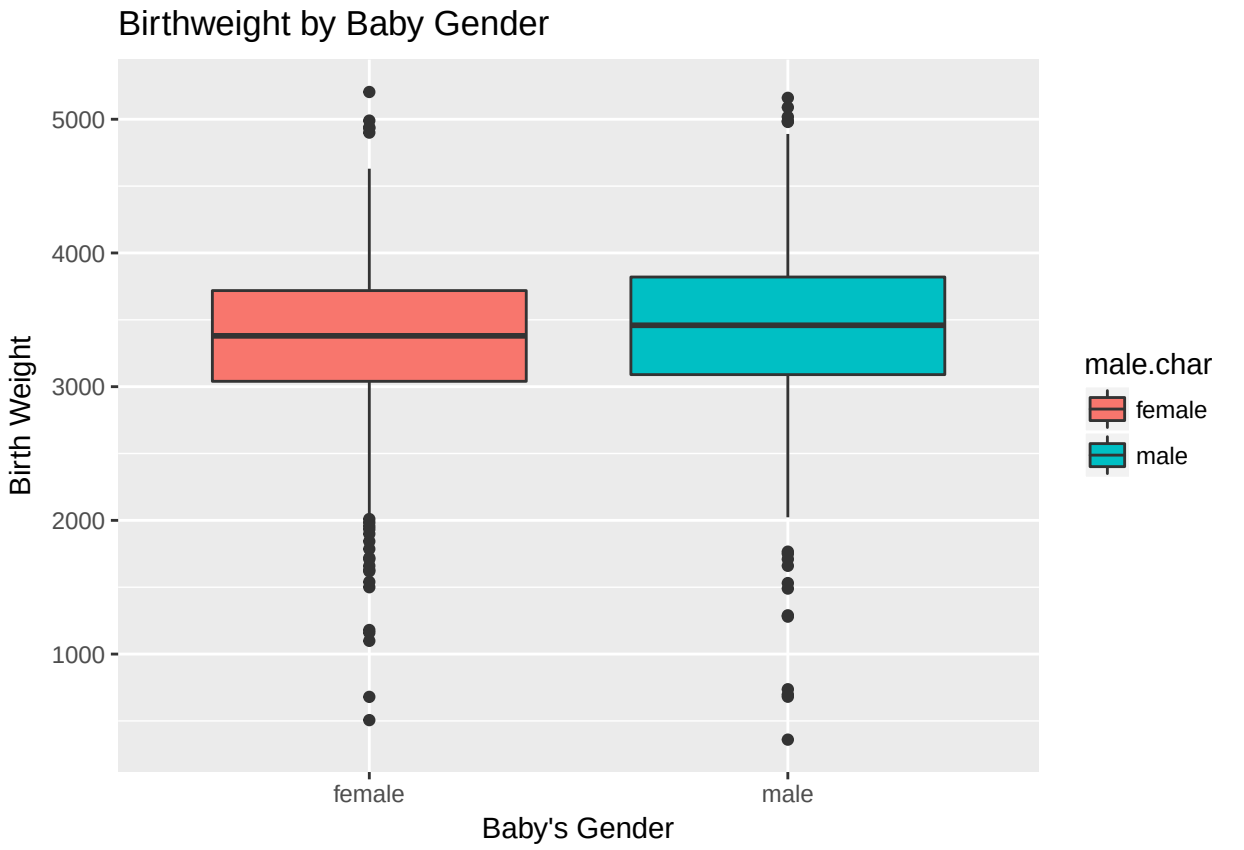
father's age when compared to mother's age. * Father's education is not normally distributed, and there seem to be "graduation effects" at the twelve-year (high-school) and sixteen-year (college) levels. This is an ordinal variable.

Intuitively, we think that neonates born to mothers with average numbers of prenatal visits would be the healthiest, as no visits is associated with poor neonatal health, and a high number of prenatal visits is likely due to health problems discovered in the developing baby during the gestational period.

Several of the ordinal variables have a significant amount of skew, so we might be tempted to perform log transformations on these variables for our model. However, the interpretability of the coefficients is problematic. A percent change in an ordinal value does not make sense.

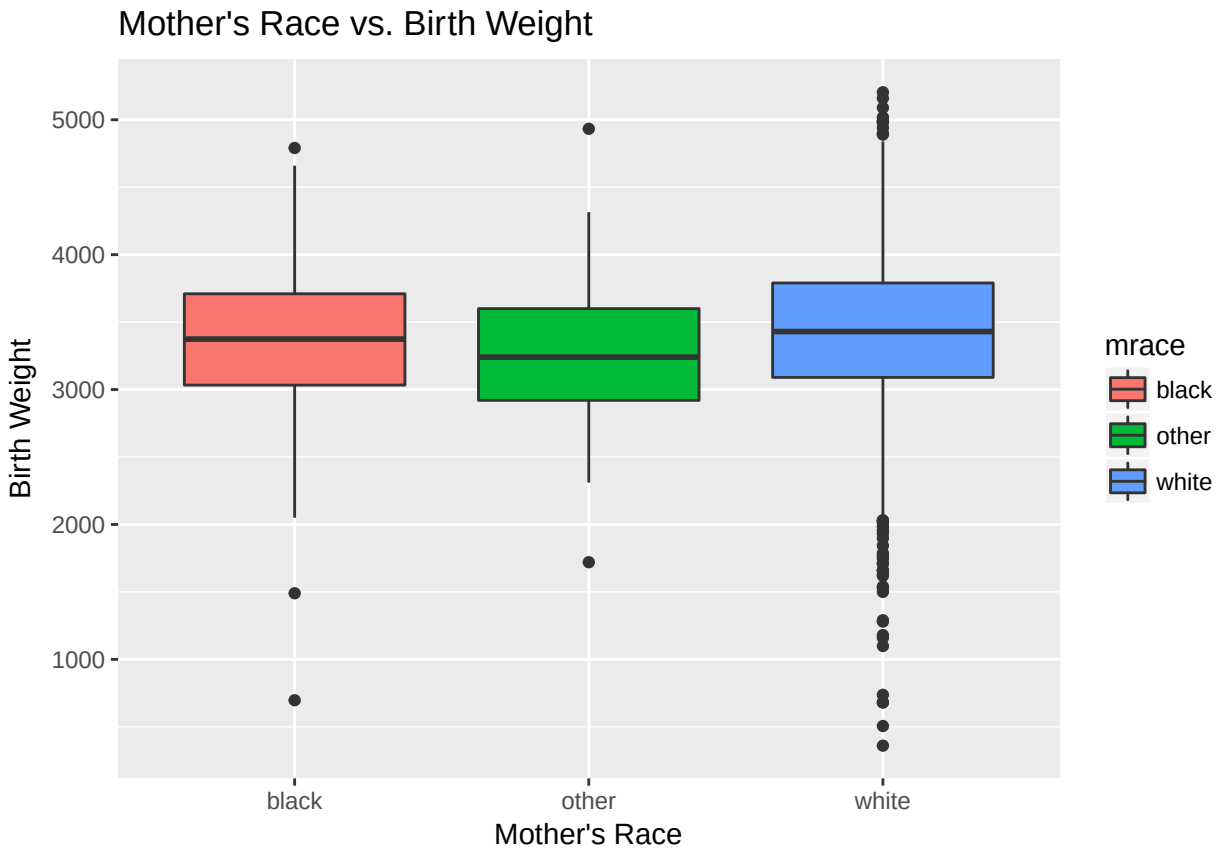
Let's examine the categorical variables in our data set: male, and the mother's race indicator variables. We'll make box plots of bwght fmaps divided by the different categories and discuss the results:

```
# Create character variables for male/female babies so ggplot2 can automatically
# make the legend for us.
data[,'male.char']                <- 'male'
data[data$male == 0, 'male.char'] <- 'female'

male.bwght.plot <- ggplot(data=data, aes(x=male.char, y=bwght)) +
    geom_boxplot(aes(fill=male.char)) +
        xlab('Baby\'s Gender') +
        ylab('Birth Weight') +
        ggtitle('Birthweight by Baby Gender')

grid.arrange(male.bwght.plot,
             ncol=1)
```



It looks like both the mean and variance of birth-weight are higher for males. The higher average birth-weight for males is reasonable (Van Vliet, 2009 - see "Decreasing Sex Difference in Birth Weight", *Epidemiology*, Vol 20, Issue 4, p 622). Because of this disparity, including gender as one of the explanatory variables will likely improve our model. The higher variance in males can be explained by a higher number of observations for males than females.

```
# Create character variable for the mother's race variables.
data[,'mrace']                  <- 'white'
data[data$mblck == 1, 'mrace'] <- 'black'
data[data$moth  == 1, 'mrace'] <- 'other'

mrace.bwght.plot <- ggplot(data=data, aes(x=mrace, y=bwght)) +
        geom_boxplot(aes(fill=mrace)) +
        xlab('Mother\'s Race') +
        ylab('Birth Weight') +
        ggtitle('Mother\'s Race vs. Birth Weight')

grid.arrange(mrace.bwght.plot,
             ncol=1)
```

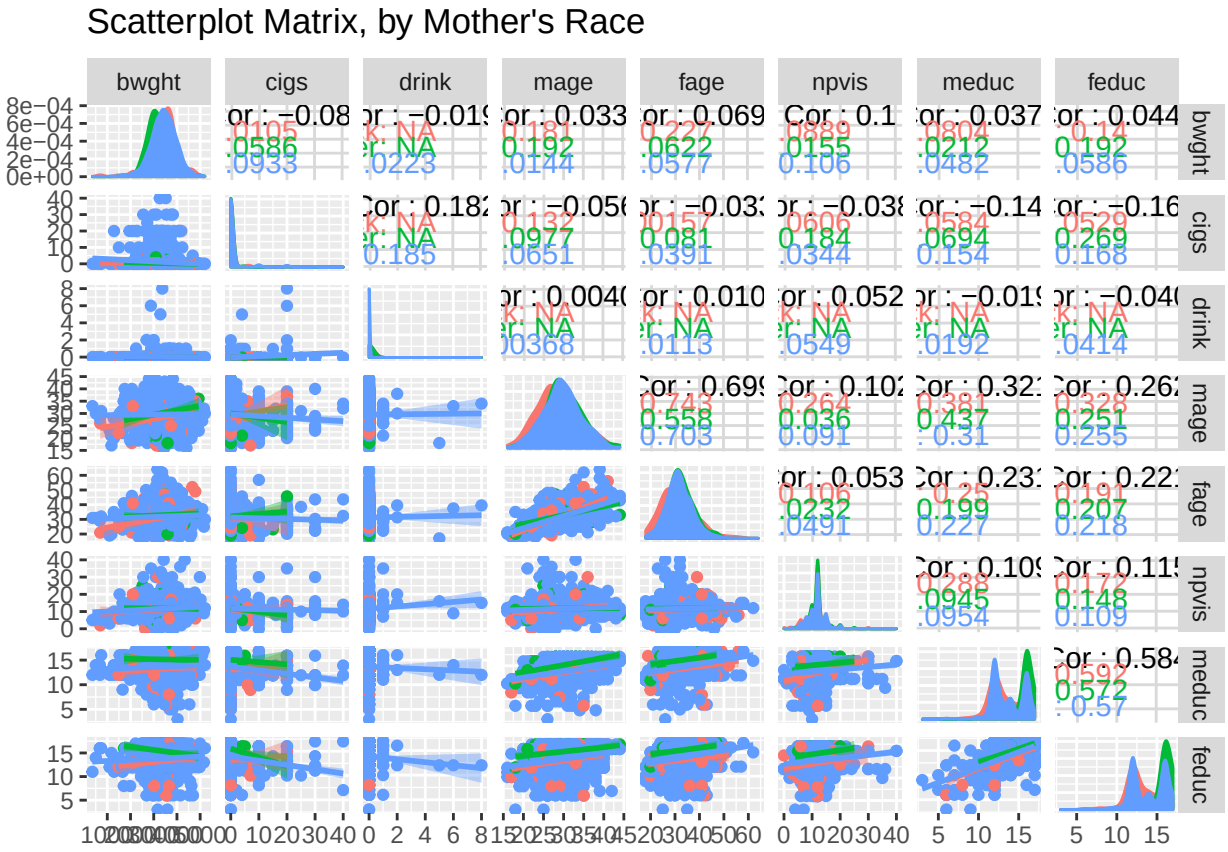


It looks like whites have the highest average and variance of birth-weight among the races. Again, the higher variance can be explained by having more observations of white babies than the other races.

Let's explore the correlations between the predictor variables, and the potential effects that the race variables have on the correlations. We'll create a scatterplot matrix of all of the predictor variables, and distinguish the observations by race:

```
ggpairs(data = data,
        columns = c('bwght', 'cigs', 'drink', 'mage',
```

```
                'fage', 'npvis', 'meduc', 'feduc'),
        title = "Scatterplot Matrix, by Mother's Race",
        lower = list(continuous="smooth"),
        ggplot2::aes(fill = mrace, colour = mrace))
```



Observations: * The correlations between the variables all seem to be low, aside from the correlations between mother & father age. The correlation coefficient of mother & father age is 0.7, so we likely don't need to worry about bias from multicollinearity if we choose to include both variables in a regression. * The linear relationship between bwght and mage seems to differ between the mrace options. Slope dummy variables for mrace could make a regression model more predictive. * The linear relationship between bwght and meduc seems to shift between the mraces options. Intercept dummy variables may be helpful to us in a regression. * Mother's and father's age and education appear to be strongly correlated, which could cause problems in our model if both are included.

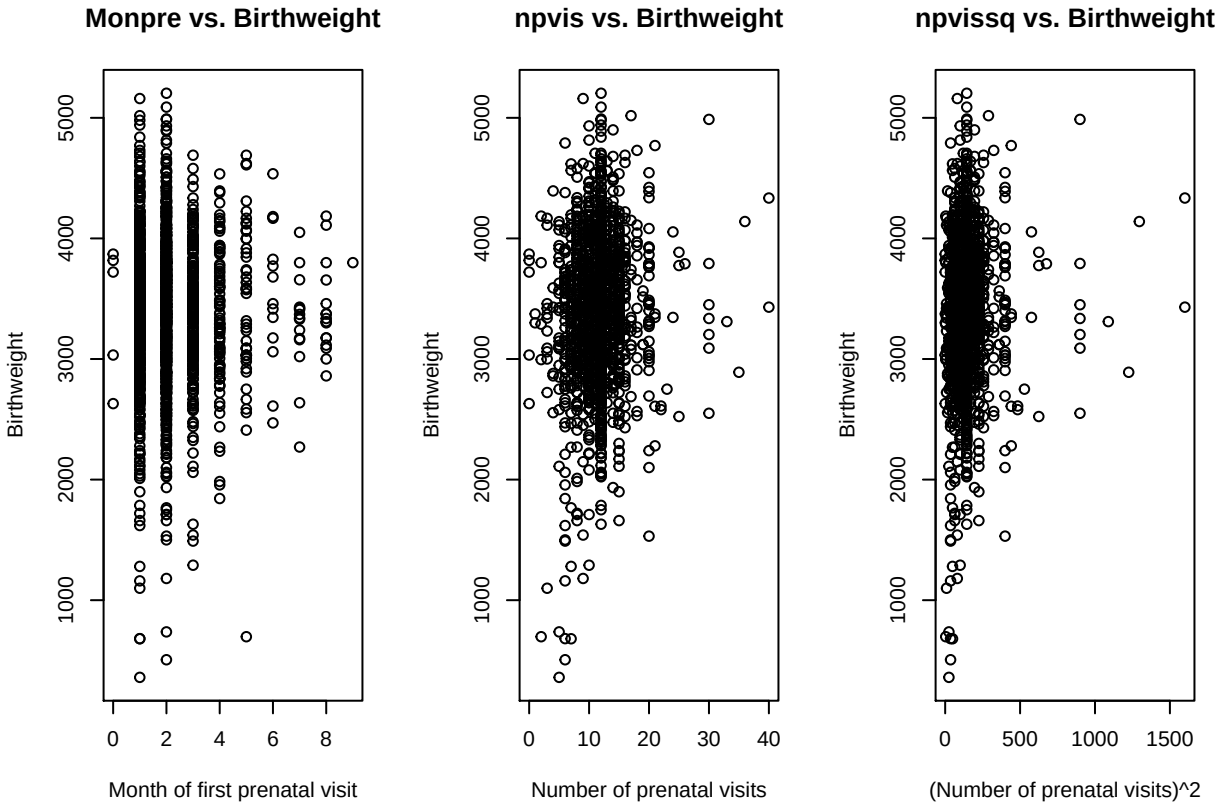
2. Model Building Process

Transformations to apply to variables / Creation of new variables

As stated in the introduction, our predictor variable is "prenatal care". We must find a way to operationalize this concept from some of our variables. For our model, we chose month prenatal care began (monpre) and number of prenatal visits (npvis) as our predictor variables. Our predicted variables, as dictated by the assignment, will be birth weight.

Though an exponential transformation of npvis (npvissq) is already included in the data set, we have found no compelling reason to use it. We didn't want to weight the outliers too heavily.

```r
par(mfrow = c(1, 3))
plot(data$bwght ~ data$monpre, main = "Monpre vs. Birthweight",
     xlab = "Month of first prenatal visit", ylab = "Birthweight")
plot(data$bwght ~ data$npvis, main = "npvis vs. Birthweight",
     xlab = "Number of prenatal visits", ylab = "Birthweight")
plot(data$bwght ~ data$npvissq, main = "npvissq vs. Birthweight",
     xlab = "(Number of prenatal visits)^2", ylab = "Birthweight")
```



It was dictated to us in the analysis assignment that our operationalized outcome variables for "prenatal health" were birth-weight, 1-minute APGAR, and 5-minute APGAR, as these are the "measures of the well-being of infants just after birth". However, we run into logistical issues when attempting to use the 1-minute APGAR or the 5-minute APGAR as an outcome variable, because these are both ordinal variables. Therefore, we will use the numeric variable of birth-weight as our operationalized outcome variable for the operationalized predictor variables that we choose.

For our initial model, we will choose our predictor variables to be the month that prenatal care began (monpre) and the number of prenatal visits (npvis). Intuitively, we would expect that earlier initiation of prenatal care would lead to better health outcomes in newborns, as much of the crucial period of embryonic development occurs in the first trimester. Mothers receiving prenatal care earlier may initiate healthy behavioral, dietary, and pharmacologic interventions during this period.

Intuitively, the relationship with birth-weight and number of prenatal visits may be more complex, and possibly not linear. It may be possible that the relationship is one where the birth-weight is best modeled by the number of prenatal visits with a negative-square relationship. This would mean that the women with the average number of prenatal visits have the healthiest babies with the highest birth-weights, whereas mothers that did not have a lot of prenatal care (low number of npvis) or mothers with a high number of prenatal care visits have the least healthy babies with the lowest birth-weights. It is intuitive as to why mothers that do not get prenatal care would have unhealthy babies. However, we theorize that mothers with very high number of prenatal visits may be receiving extra visits because they are deemed "At-risk pregnancies",

where a health condition is identified in the baby that may put the baby at risk for low birth-weight and lower level of health. To check this assumption quickly, we refer to the previous plots of birth-weight against npvis and npvissq (see above).

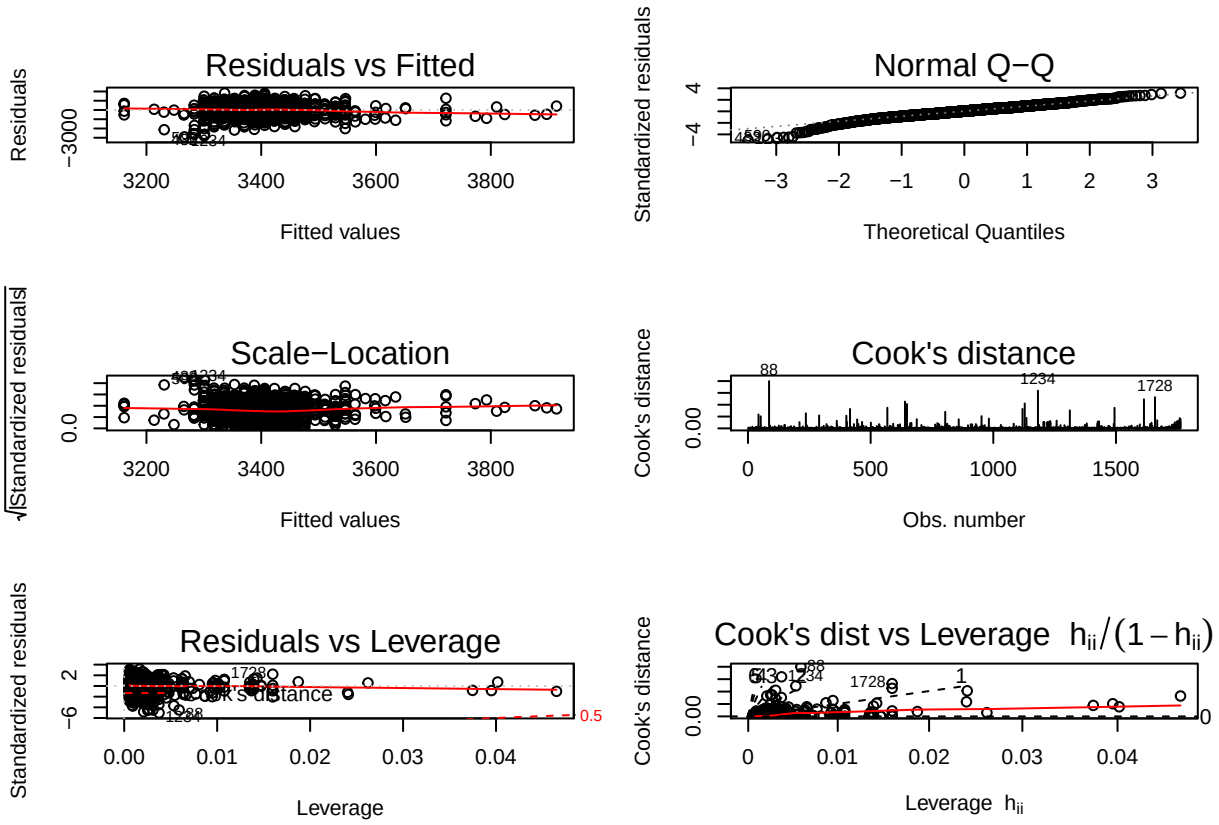
We are surprised that we do not see the negative-square relationship that we predicted. We even performed a log transform and the plot looked practically similar.

Now, we proceed with our initial model using the aforementioned predictor and predicted variables:

```
model_bwght <- lm(bwght ~ monpre + npvis, data = data)
summary(model_bwght)
```

```
##
## Call:
## lm(formula = bwght ~ monpre + npvis, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2906.08  -326.21    21.36   359.01  1823.72
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3161.271     59.917  52.761  < 2e-16 ***
## monpre        17.062     11.727   1.455    0.146
## npvis         17.549      3.925   4.471 8.28e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 577.5 on 1760 degrees of freedom
##   (69 observations deleted due to missingness)
## Multiple R-squared:  0.01124,    Adjusted R-squared:  0.01011
## F-statistic: 9.999 on 2 and 1760 DF,  p-value: 4.806e-05
```

```
par(mfrow = c(3, 2))
for (i in 1:6) {
  plot(model_bwght, which=i)
  }
```



We will check the classical linear regression model assumptions of our model:

a. Linear population model

By the definition of the lm command, the model is linear:

$$bwght_i = \beta_0 + \beta_1 monpre + \beta_2 npvis + u_i$$

b. Random Sampling

In order to know about this assumption, we would need to know specifics on the data collection methods for the bwght_w203 Rdata, specifically about the sampling methods. We are told that the data comes from the National Center for Health Statistics and from birth certificates. Since this is a national database and most babies are born in hospitals, it is reasonable to assume that all neonates born in the U.S. have an equal chance of being represented in the database. Therefore, we will proceed as if the random sampling assumption is met here.

c. No perfect multicollinearity

Here we need to know that there is no perfect collinearity between our individual independent variables. Thus, we need to see that the correlation between each of our independent variables is not close to one or negative one.

```
cor(data$monpre,data$npvis, use = "complete.obs")
```

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

```
vif(model_bwght)
```

```
##   monpre    npvis
## 1.103384 1.103384
```

```
r_sqrd <- cor(data$monpre,data$npvis, use = "complete.obs") * cor(data$monpre,data$npvis, use = "comple
r_sqrd
```

```
## [1] 0.09369755
```

We see that the correlation between month prenatal care was initiated and the number of prenatal visits is -0.31, and that the variance inflation factors are small. Our variance inflation factors are < 10, which is where we start to consider serious issues with multicollinearity. We also see that the r-squared value is 0.094, which is not even close to the value of 1.0, where we worry about problems associated with multicollinearity.

The assumption of no perfect multicollinearity is met.

d. Zero-conditional mean

The Zero-conditional mean assumption requires that for any possible values of our predictors, our error is zero in expectation. We can visualize this by again plotting the residuals vs fitted plot (refer to the diagnostic plots of the model above), where we expect the data to be evenly distributed above and below the horizontal line at zero. This means that the data at each fitted value should be centered at zero.

The red line, which is a smoothing curve that R displays to help visualize the mean of the residuals as you move along the fitted values axis. Again, this red line is slightly positive up to a fitted value of approximately 3450, and then becomes slightly negative. This may lead one to think there is a violation of zero-conditional mean here, as the average value of the residuals at each value of the fitted value should be around zero. However, we do not see this as a major violation; as for the majority of the plot, the red line is nearly horizontal and straddles zero.

We therefore state that we meet the zero-conditional mean assumption with the model. Additionally, even if there was evidence of violation of the zero-conditional mean assumption, given our large sample size we are confident that the coefficients are consistent due to OLS asymptotics.

e. Homoskedasticity:

We will next test for the assumption of homoskedascticity (the "homogeneity of variance"), which assumes that the variance of the error term is constant across all value of the independent variable(s). When the assumption of homoskedascticity fails, the standard errors from our OLS regression estimates will be inconsistent.

To visualize if homoskedasticity is present, we can again view the residuals plotted against the fitted values (refer to the diagnostic plots of the model above) to see whether the variance of residuals is constant across the fitted values. If homoskedascticity is present, then there should not be areas where the range in the vertical deviation is significantly larger than other regions, with a nice, smooth range in vertical height along the residuals axis. We can see here that there seems to be an increase in vertical deviation around fitted values of 3300. This make us suspect there is heteroskedascticity in the data. To follow-up, we perform a Score-test for non-constant error variance.

```
ncvTest(model_bwght)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 28.74387    Df = 1     p = 8.261083e-08
```

The resulting p-value of 8.261083e-08 leads us to reject the null hypothesis, which states that the error variance is constant. This tells us that the homoskedascticity assumption is violated and that there is heteroskedascticity in the data. Heteroscedasticity can be addressed by calculating heteroskedasticity-robust standard errors.

Therefore, we state that the homoskedascticity assumption is violated.

f. Normality of Errors

To test the assumption of normality of errors, we can simply visualize a histogram of our residuals, which represents our estimator for the parameter of errors.

```
hist(resid(model_bwght))
```

We see here that the residuals appear approximately normally distributed with the distribution centered at zero, but that there is a slight negative skew to the residuals.

As this negative skew is only minor, we therefore state that the normality of errors assumption is met here.

Improved Model

We seek to add covariates to our model that increase accuracy without introducing bias. This means that the new model would provide a better prediction of our outcome variable without causing the expected value of our estimator to deviate further from the value of our parameter.

It makes intuitive sense to assume that the number of cigarettes smoked per day and the number of alcoholic beverages consumed per week by the Mother is an intermediate outcome for our predictor of "prenatal care". There is extensive medical literature showing that cigarette smoking and alcohol consumption by mothers during the gestational period is linked with low birth weight and worse health outcomes in neonates. Therefore, we might reason that our operationalized predictor variable of "prenatal care" predicts cigarette smoking and alcohol consumption during the gestational period, which in-turn predicts neonatal birth-weight and thus neonatal health. Many medical providers specifically focus their prenatal health visits on avoiding tobacco and alcohol in their pregnant patients. Since we believe that much of the variance in cigarette smoking and alcohol consumption is predicted by prenatal health, it makes sense to operationalize the cigs and drink variables as predictor variables for neonatal health, which is operationalized by birth-weight.

To start, we will look at the difference in birth weights between babies born to mothers that smoke vs mothers that do not smoke, to see if there is a statistically and practically significant difference between these two groups. This may provide some insight into whether or not the cigs variable can achieve our goal of increasing accuracy of our model without introducing bias.

```r
smoker <- data[data$cigs > 0, ]
non_smoker <- data[data$cigs == 0, ]
summary(smoker$bwght)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1290    2899    3175    3214    3575    4680     110
```

```r
summary(non_smoker$bwght)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     360    3100    3440    3427    3790    5204     110
```

```r
t.test(smoker$bwght, non_smoker$bwght)
```

```
##
##  Welch Two Sample t-test
##
## data:  smoker$bwght and non_smoker$bwght
## t = -4.5636, df = 177.86, p-value = 9.35e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -305.4487 -121.0302
## sample estimates:
## mean of x mean of y
##   3214.17   3427.41
```

```r
cohensD(smoker$bwght,non_smoker$bwght)
```

```
## [1] 0.3758818
```

We see that there is a statistically significant difference in the mean birth weights of the children of smokers vs non-smokers. The mean birth weight of babies for mothers that smoke is 3214, whereas the mean birth weight of babies for mothers that do not smoke is 3427. This is statistically significant at an alpha of 0.0001. Additionally, there is practical significance to this finding (the mean birth-weight of the children of non-smokers is 7% higher). The Cohen's d value is 0.38, indicative of a small practical effect, but approaching a moderate practical effect.

In our two-sample independent t-test between smokers vs non-smokers, we simply looked at the difference between any cigarette use and no use. This does not necessarily guarantee a linear, predictive relationship between number of cigarettes smoked and our operationalized outcome for neonatal health (birth weight). But we must explore this further.

Next we look at the difference in birth weights between babies born to mothers that drink vs mothers that do not drink, to see if there is a statistically significant or practically significant difference between these two groups. This may provide some insight into whether or not the drink variable can achieve our goal of increasing accuracy of our model without introducing bias.

```r
drinker <- data[data$drink > 0, ]
non_drinker <- data[data$drink == 0, ]
summary(drinker$bwght)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    2240    2972    3330    3340    3687    4394     115
```

```r
summary(non_drinker$bwght)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     360    3080    3430    3410    3771    5204     115
```

```
t.test(drinker$bwght, non_drinker$bwght)
```

```
##
##  Welch Two Sample t-test
##
## data:  drinker$bwght and non_drinker$bwght
## t = -0.50336, df = 15.306, p-value = 0.6219
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -364.0168  224.7360
## sample estimates:
## mean of x mean of y
##  3339.938  3409.578
```

```
cohensD(drinker$bwght,non_drinker$bwght)
```

```
## [1] 0.1218548
```

We see that there is not a statistically significant difference in the mean birth weights of the mothers that drink alcohol during the gestational period vs mothers that do not drink alcohol during the gestational period. The mean birth weight of babies for mothers that drink is 3340, whereas the mean birth weight of babies for mothers that do not drink is 3410. This is not statistically significant at even an alpha of 0.1. Additionally, there is not practical significance to this finding. The Cohen's d value is 0.12, indicating that there is not even a small practically significant effect. We can note, however, that the mean birth-weight for mothers that drink is lower than for mothers that do not drink. However, this difference in means is not statistically or practically significant per our analysis.

We will now look at sequentially adding both the cigs and the drink variable to our regression, to see how if affects our prediction accuracy and efficiency. One may be tempted to add cigs and not drink, as there was significant difference between birth-weights of mothers that smoked vs mothers that did not, whereas this was not the case for mothers that drank vs mothers that did not. We know from the principles of regression that any time we add an additional predictor variable to a regression model, we increase the value of R-squared, which is one metric by which we evaluate the explanatory power of our model. However, the R-squared value does not tell us whether or not the coefficient estimates or predictors are biased.

If some of the variance of the outcome variable (birth-weight) is explained by cigs and/or drink, and these variables are not included in the original regression model, then our original model was afflicted by omitted-variable bias (OVB). Omitted-variable bias occurs when a regression model omits an important variable that explains part of the variance in the outcome variable, leading to the introduction of bias into the estimates of the parameters in the regression analysis.

If our goal is to increase the accuracy of our results without introducing bias, then the removal of bias with the addition of a predictor variable is even better.

To show that OVB was present in our original regression model, we must show that the omitted variables' (cigs and drink) true regression coefficient are not zero. This means that the omitted variables are determinants of the dependent variable. This will be shown after adding cigs and drink to our "improved" regression model. We must also show that the omitted variables are correlated with one of the independent variables specified in the regression (npvis and/or monpre). This means that we must show that the covariance between our omitted variables and the predictor variables in our original regression are not zero. We will accomplish this here.

```
cov(data$monpre, data$cigs, use = "complete.obs")
```

```
## [1] 0.5182971
```

```
cov(data$monpre, data$drink, use = "complete.obs")
```

```
## [1] -0.003528883
```

```
cov(data$npvis, data$cigs, use = "complete.obs")
```

```
## [1] -0.6127706
```

```
cov(data$npvis, data$drink, use = "complete.obs")
```

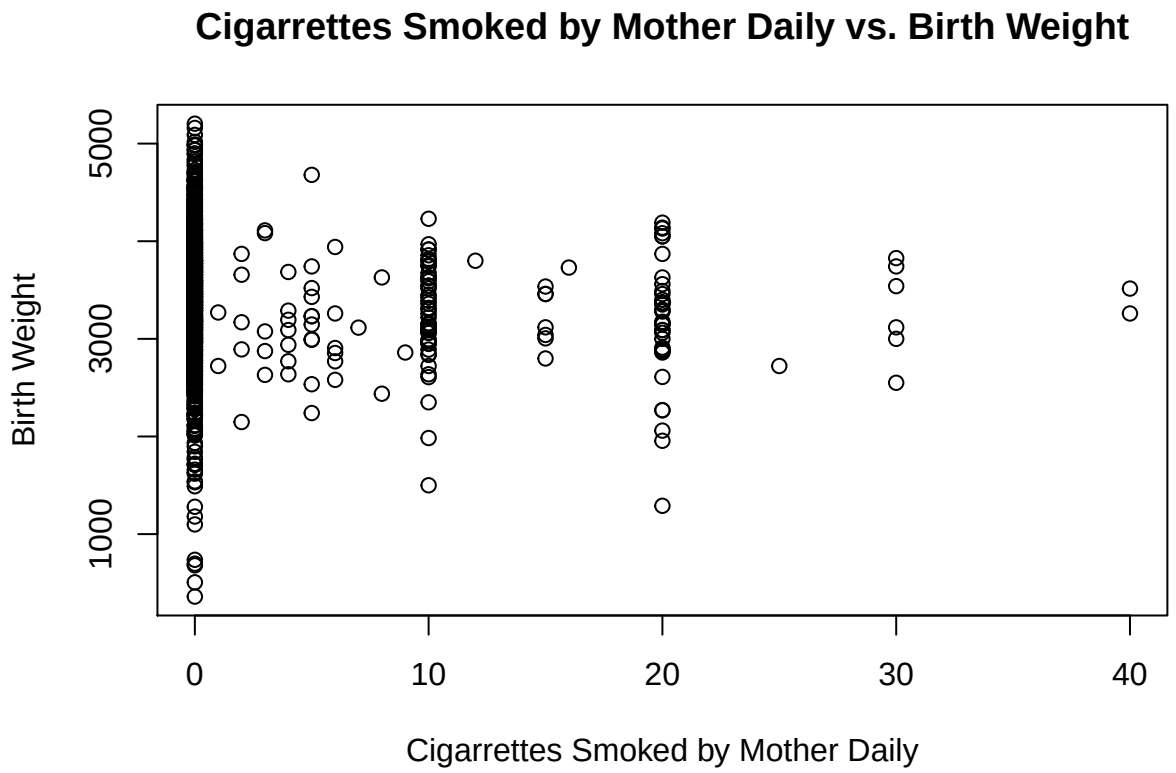
```
## [1] 0.05817154
```

We find that the covariance between cigs and both monpre and npvis is not close to zero. However, we find that the covariance between drink and both monpre and npvis approaches zero. This may tell us that the omission of drinks from the regression model does not introduce bias.

Moving on, in our two-sample independent t-tests between smokers vs non-smokers and between drinkers vs non-drinkers, we simply looked at the difference between any cigarette use vs no use and any alcohol consumption vs no consumption. This does not necessarily guarantee a linear, predictive relationship between number of cigarettes smoked or the number of drinks consumed and our operationalized outcome for neonatal health (birth weight). But we must explore this further. We will start by adding cigs to our regression model.

We also add the variable "male" to our model. As discussed previously, there is a significant difference between the birth-weights of males and females, so this will only help the model's accuracy, without introducing bias:

```
plot(data$bwght~data$cigs, main = "Cigarettes Smoked by Mother Daily vs. Birth Weight",
     xlab = "Cigarettes Smoked by Mother Daily", ylab = "Birth Weight")
```

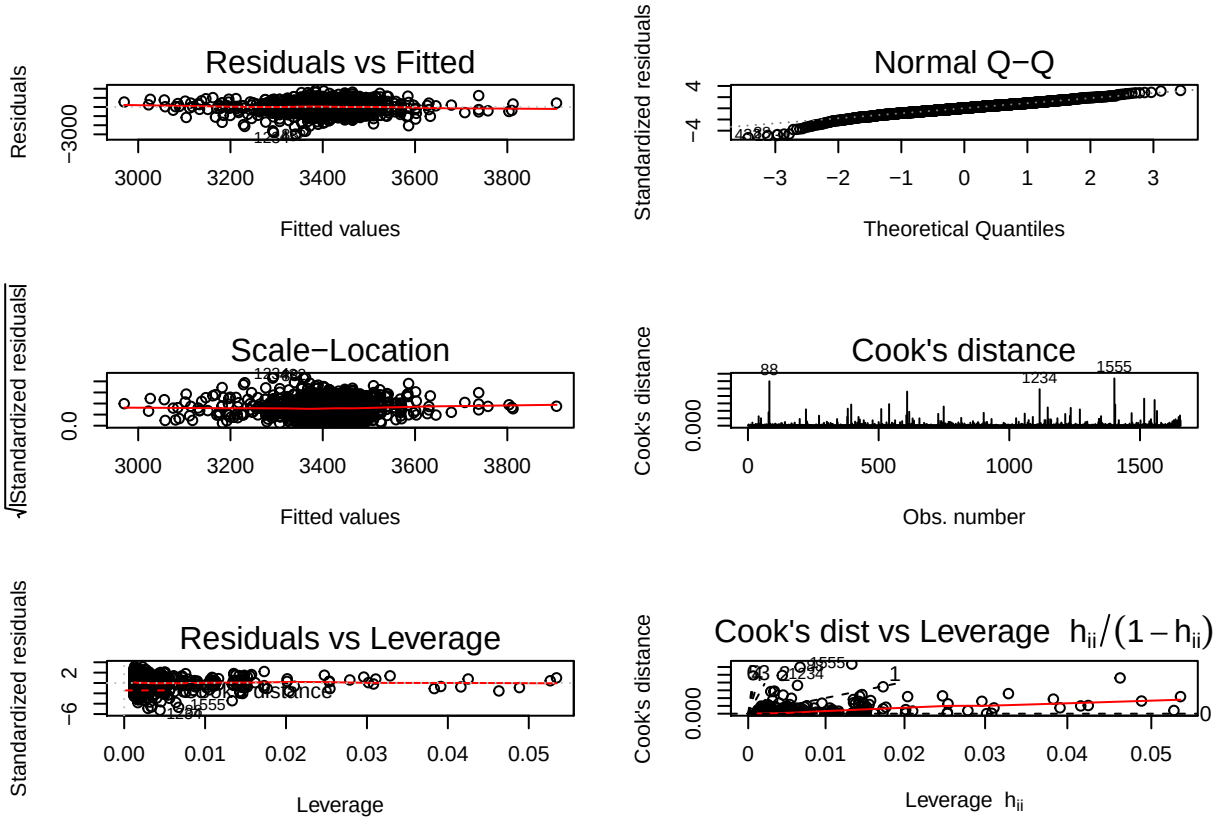



```
model_bwght_better <- lm(bwght ~ monpre + npvis + cigs + male, data = data)
summary(model_bwght_better)
```

```
##
## Call:
## lm(formula = bwght ~ monpre + npvis + cigs + male, data = data)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2981.43  -334.48    22.52   357.68  1819.68
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3168.609     62.580  50.633  < 2e-16 ***
## monpre        16.708     11.820   1.414 0.157677
## npvis         15.191      3.933   3.863 0.000116 ***
## cigs         -11.126      3.328  -3.343 0.000847 ***
## male          80.158     28.001   2.863 0.004254 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 568.9 on 1650 degrees of freedom
##   (177 observations deleted due to missingness)
## Multiple R-squared:  0.02014,    Adjusted R-squared:  0.01776
## F-statistic: 8.478 on 4 and 1650 DF,  p-value: 9.048e-07
```

```
par(mfrow = c(3, 2))
for (i in 1:6) {
  plot(model_bwght_better, which=i)
  }
```




```
AIC(model_bwght)
```

```
## [1] 27428.8
```

```
AIC(model_bwght_better)
```

```
## [1] 25701.3
```

We next look at an efficiency-adjusted measure of fit: The Akaike Information Criterion (AIC). The AIC penalizes a model as the number of variables increases if those variables do not commensurately increase the predictive ability of the model. When looking at different models for the same data, a larger AIC is indicative of a worse fit.

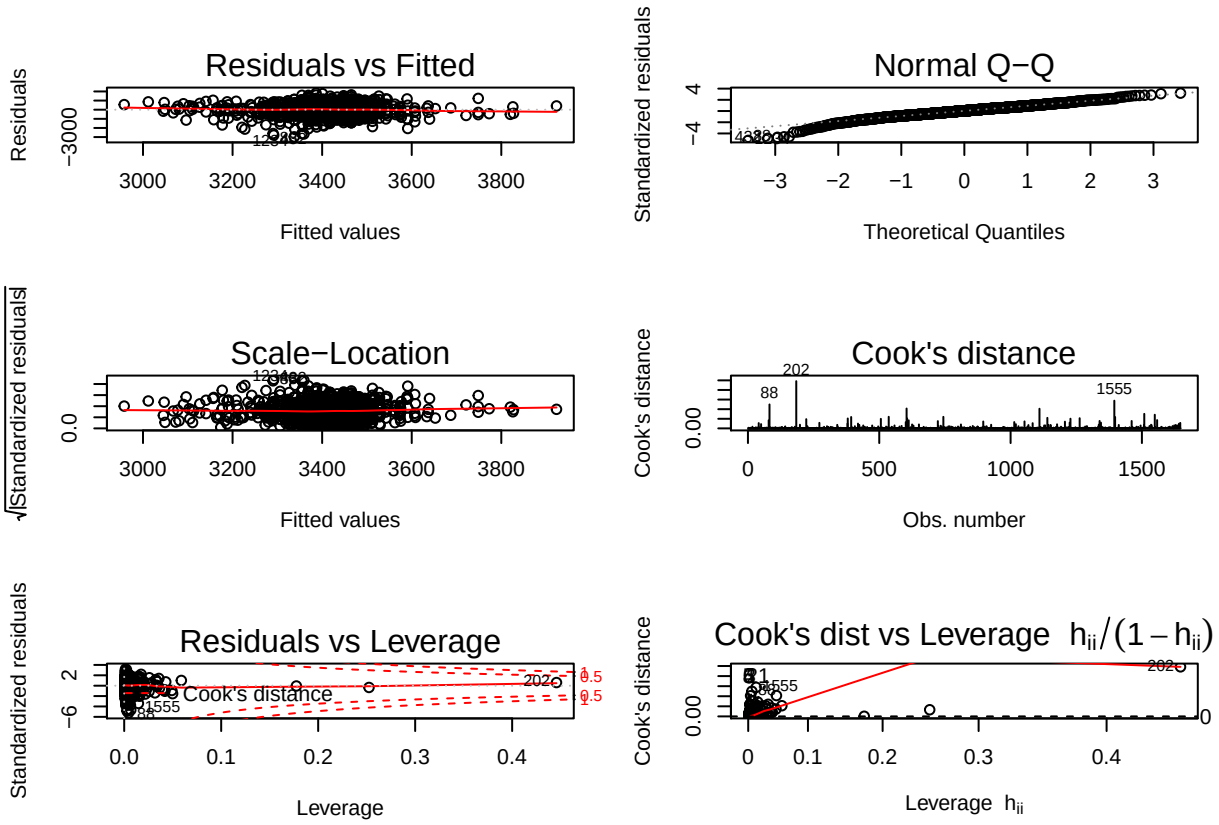
We see that the AIC value for our original model (predictor variables of month prenatal care began and number of prenatal visits) is 27428.8. This is higher than the AIC value for our improved model (where cigarettes smoked per day on average by the mother and the baby's gender were added as predictor variables), which is 25701.3. This indicates that our improved model is a better fit, thus achieving our goal of increasing our model's accuracy in an efficient manner.

Next, we add drink to the initially improved regression model:

```
model_bwght_better_with_etoh <- lm(bwght ~ monpre + npvis + cigs + male + drink, data = data)
summary(model_bwght_better_with_etoh)
```

```
##
## Call:
## lm(formula = bwght ~ monpre + npvis + cigs + male + drink, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2977.72  -335.54    23.64   360.83  1817.64
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3163.013     62.809  50.359  < 2e-16 ***
## monpre        17.199     11.848   1.452  0.14678
## npvis         15.746      3.955   3.981 7.15e-05 ***
## cigs         -11.495      3.493  -3.291  0.00102 **
## male          78.783     28.142   2.800  0.00518 **
## drink        -13.482     48.581  -0.278  0.78141
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 569.8 on 1641 degrees of freedom
##   (185 observations deleted due to missingness)
## Multiple R-squared:  0.02095,    Adjusted R-squared:  0.01797
## F-statistic: 7.024 on 5 and 1641 DF,  p-value: 1.664e-06
```

```
par(mfrow = c(3, 2))
for (i in 1:6) {
  plot(model_bwght_better_with_etoh, which=i)
  }
```




```
AIC(model_bwght)
```

```
## [1] 27428.8
```

```
AIC(model_bwght_better)
```

```
## [1] 25701.3
```

```
AIC(model_bwght_better_with_etoh)
```

```
## [1] 25583.16
```

We see that by adding drink to the regression model, our AIC has decreased further, from 25701 to 25583. However, is this a significant difference. We could run a Vuong test on nonnested models, which is a paired z-test of individual log-likelihoods, in order to determine whether or not the model with drink added to it is statistically equivalent to the model without drink added to it. However, this is beyond the scope of this analysis.

Instead, we will perform an F-test for exclusion restriction on our model with drink added (model_bwght_better_with_etoh). This will tell us whether or not the drink variable can be excluded from the model because its coefficient equals zero, indicating it has no partial effect on the predicted value of the dependent variable (birth-weight). With the F-test, we measure our model fit by the SSR (Sum of Squared Residuals). If the SSR goes down, then there is evidence of a better fit. By the principles of regression, anytime you take variables out of a regression, then the SSR will go up. But the F-test will tell whether or not the change is statistically significant.

```
data_no_na <- data[data$bwght != "NA" &  data$monpre != "NA" & data$npvis != "NA" & data$cigs != "NA" &
model_bwght_better_no_na <- lm(bwght ~ monpre + npvis + cigs + male, data = data_no_na)
model_bwght_better_with_etoh_no_na <- lm(bwght ~ monpre + npvis + cigs + male + drink, data = data_no_na
anova(model_bwght_better_no_na, model_bwght_better_with_etoh_no_na)
```

```
## Analysis of Variance Table
##
## Model 1: bwght ~ monpre + npvis + cigs + male
## Model 2: bwght ~ monpre + npvis + cigs + male + drink
##   Res.Df       RSS Df Sum of Sq     F Pr(>F)
## 1   1642 532751977
## 2   1641 532726974  1     25003 0.077 0.7814
```

We see here that our F-statistic value is 0.077, with a corresponding p-value of 0.7814. Therefore, we fail to reject the null hypothesis (that the coefficient for drink in the regression model is zero). The results of the F-test tell us that drinks does not contribute significant explanation of the birth-weight once month of prenatal care initiation, number of prenatal visits, and cigarettes smoked by the mother have been taken into consideration in the model. Thus we choose our initial improved model "model_bwght_better"" as the model that will increase our accuracy without introducing bias. This is the model where we added cigs, but had not yet added drink.

We will now quickly check the 6 assumptions of the classical linear model within our improved regression model (model_bwght_better). Because most of the assumptions are very similar to our detailed analysis of the original model, we will only discuss notable differences here.

a. Linear population model, d. Zero-conditional mean, e. Homoskedasticity

These assumptions are met (or not met) equally with the original model.

b. Random Sampling

The number is cigarettes smoked per day appears to be somewhat discrete, broken into units of 5 cigarettes per day. However, this is not indicative of a lack of random sampling. We do not see any major change in meeting this assumption in our improved model when compared to our original model.

The gender attribute can be safely assumed to be randomly sampled. To check, we make sure that the number of boys born is 51% of the total births, as is consistent with general birth statistics.

```
length(which(data$male == 1))/length(data$male)
```

```
## [1] 0.5136463
```

c. No perfect multicollinearity

```
cor(data$monpre,data$cigs, use = "complete.obs")
```

```
## [1] 0.09791768
```

```
cor(data$npvis,data$cigs, use = "complete.obs")
```

```
## [1] -0.03872346
```

```
vif(model_bwght_better)
```

```
##   monpre    npvis     cigs     male
## 1.119402 1.109507 1.010800 1.001249
```

```
r_sqrd_monpre_cigs <- cor(data$monpre,data$cigs, use = "complete.obs") * cor(data$monpre,data$cigs, use
r_sqrd_npvis_cigs <- cor(data$npvis,data$cigs, use = "complete.obs") * cor(data$npvis,data$cigs, use =
r_sqrd_monpre_cigs
```

```
## [1] 0.009587873
```

```
r_sqrd_npvis_cigs
```

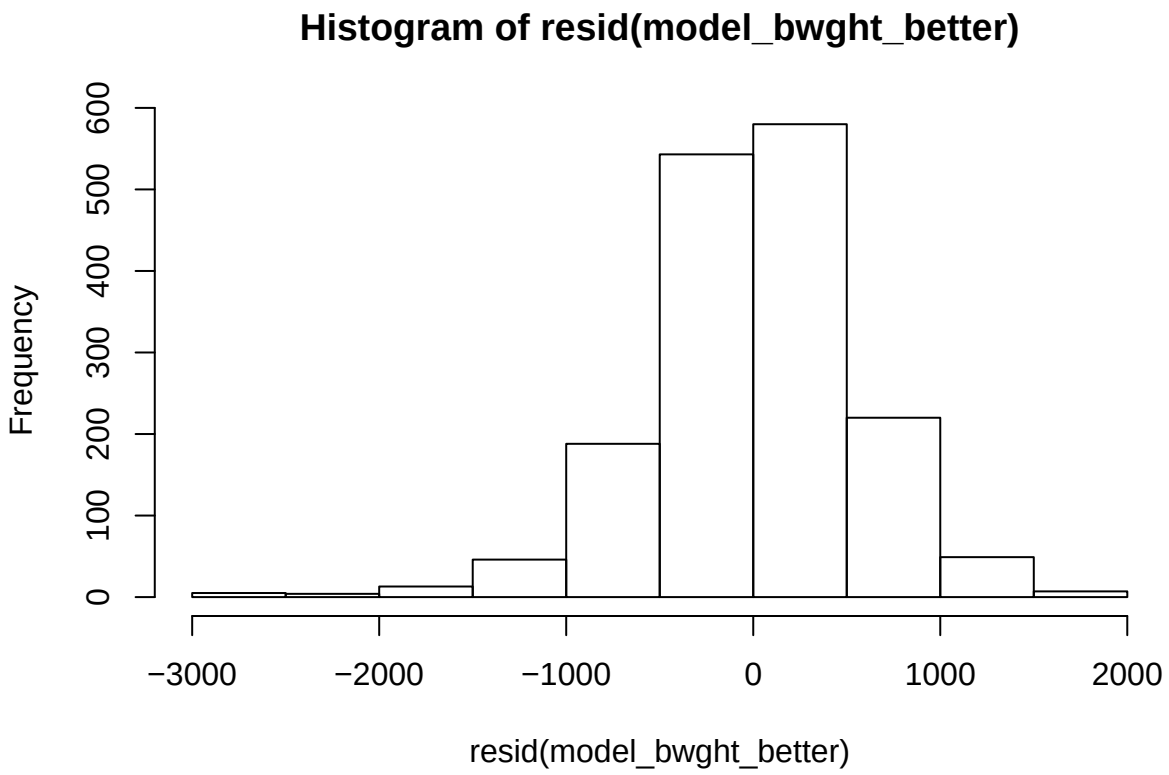
```
## [1] 0.001499506
```

We see that the correlation between month prenatal care was initiated and cigarettes smoked is 0.09791768, the correlation between number of prenatal visits and cigarettes smoked is -0.03872346,and that the variance inflation factors are small. Our variance inflation factors are < 10. We also see that the r-squared values are not even close to the value of 1.0, where we worry about problems associated with multicollinearity.

The assumption of no perfect multicollinearity is met in the improved model, as it was in the original model.

f. Normality of Errors

```
hist(resid(model_bwght_better))
```



We see here that the residuals appear approximately normally distributed with the distribution centered at zero, but that there is a slight negative skew to the residuals. This is not a major change from what we saw in the original model. As this negative skew is only minor, we therefore state that the normality of errors assumption is met here.

Making the original model worse

Next, we will seek to add covariates to our original model that may be problematic for one reason or another. One reason that adding a covariate to an existing model may be problematic is that it does not add much in the way of predictive value with respect to efficiency. If we add covariates that do not provide much partial explanatory value of our outcome variable, then this may be especially apparent.

There are four variables that we will evaluate in our attempt to make our model worse, as they all have intuitively complex relationships with neonatal health: mother age, father age, mother education, and father education. We will evaluate their addition to our regression by looking at their AIC vs the AIC for the original model, then perform a series of F-tests to see if we can accept the null hypothesis that the coefficient on each variable is zero when added into the "worse" model as compared to the original model..

In our data, we theorize that father's age may be a variable that does not provide much partial explanatory value for the variance in the neonate's birth-weight. One could theorize that older fathers may have more

financial resources, as individuals accumulate wealth as they age. This would lead to higher socioeconomic status, which has been linked to better neonatal health. One could theorize a similar mechanism for mother age. However, there is medical literature suggesting that older mothers have higher incidence of fetal abnormalities such as Trisomy 21, with the incidence increasing with maternal age. This makes the mage variable more complex. Both mother education and father education would theoretically have a positive relationship with neonatal health, as better-educated individuals may have more financial resources or may be more inclined to understand and appreciate prenatal care recommendations. However, these are all theorized relationships, and we will look at the data for support in how to make our original model worse.

For the purposes of F-testing in R, we will have to slightly restructure our original model here to ensure there are no NAs.

```
data_no_no_na <- data[data$bwght != "NA" &  data$monpre != "NA" & data$npvis != "NA" & data$cigs != "NA
model_bwght_no_na <- lm(bwght ~ monpre + npvis, data = data_no_no_na)
model_bwght_worse_with_feduc <- lm(bwght ~ monpre + npvis + feduc, data = data_no_no_na)
model_bwght_worse_with_meduc <- lm(bwght ~ monpre + npvis + meduc, data = data_no_no_na)
model_bwght_worse_with_fage <- lm(bwght ~ monpre + npvis + fage, data = data_no_no_na)
model_bwght_worse_with_mage <- lm(bwght ~ monpre + npvis + mage, data = data_no_no_na)
AIC(model_bwght_no_na)
```

```
## [1] 25077.41
```

```
AIC(model_bwght_worse_with_feduc)
```

```
## [1] 25075.67
```

```
AIC(model_bwght_worse_with_meduc)
```

```
## [1] 25077.76
```

```
AIC(model_bwght_worse_with_fage)
```

```
## [1] 25073.67
```

```
AIC(model_bwght_worse_with_mage)
```

```
## [1] 25076.92
```

```
anova(model_bwght_no_na, model_bwght_worse_with_feduc)
```

```
## Analysis of Variance Table
##
## Model 1: bwght ~ monpre + npvis
## Model 2: bwght ~ monpre + npvis + feduc
##   Res.Df       RSS Df Sum of Sq      F  Pr(>F)
## 1   1612 521421131
## 2   1611 520212984  1   1208147 3.7414 0.05325 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(model_bwght_no_na, model_bwght_worse_with_meduc)
```

```
## Analysis of Variance Table
##
## Model 1: bwght ~ monpre + npvis
## Model 2: bwght ~ monpre + npvis + meduc
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1   1612 521421131
## 2   1611 520887446  1    533686 1.6506 0.1991
```

```
anova(model_bwght_no_na, model_bwght_worse_with_fage)
```

```
## Analysis of Variance Table
##
## Model 1: bwght ~ monpre + npvis
## Model 2: bwght ~ monpre + npvis + fage
##   Res.Df       RSS Df Sum of Sq      F  Pr(>F)
## 1   1612 521421131
## 2   1611 519571086  1   1850045 5.7363 0.01673 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(model_bwght_no_na, model_bwght_worse_with_mage)
```

```
## Analysis of Variance Table
##
## Model 1: bwght ~ monpre + npvis
## Model 2: bwght ~ monpre + npvis + mage
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1   1612 521421131
## 2   1611 520617251  1    803880 2.4875 0.1149
```

We know that an increase in AIC from our original model to our worse model indicates a less efficient model with respect to explanatory ability. Out of the four variables we looked at, only mother's education causes an increase in the AIC value relative to the original model. Additionally, the results of our F-tests show that we fail to reject the null hypothesis that the coefficient in the regression on the mother's education variable is zero. Thus we feel confident stating that the addition of mother's age to a regression model does not add any additional explanatory ability, since the coefficient on mother's age is not statistically different from zero. We see that we can also make this statement about the mother's age variable. However, the mother's age variable is much closer to approaching statistical significance than the mother's education variable.

Interestingly, we see that the father's age variable does have some explanatory ability, with a F-test p-value of 0.017. We will not add it to a model to make it worse.

Per our reasoning, our worse model will add the variable of mother's education, as it will add an additional variable without improving the predictive efficiency of our original model:

Worse model:

```
model_bwght_worse_with_meduc_final <- lm(bwght ~ monpre + npvis + meduc, data = data)
```

We will quickly check the 6 CLM assumptions for our worse model.

a. Linear population model: The model is specified such that the dependent variable is a linear function of the explanatory variables, so we have met this assumption.

b. Random Sampling: There is no major difference here between the worse model and the original model, as they used the exact same data set and data gathering methods.

c. No perfect multicollinearity:

```
cor(data$monpre,data$meduc, use = "complete.obs")
```

```
## [1] -0.1829704
```

```
cor(data$npvis,data$meduc, use = "complete.obs")
```

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

```
vif(model_bwght_worse_with_meduc)
```

```
##   monpre    npvis    meduc
## 1.139110 1.110349 1.040295
```

```
r_sqrd_monpre_meduc <- cor(data$monpre,data$meduc, use = "complete.obs") * cor(data$monpre,data$meduc, u
r_sqrd_npvis_meduc <- cor(data$npvis,data$meduc, use = "complete.obs") * cor(data$npvis,data$meduc, use
r_sqrd_monpre_meduc
```

```
## [1] 0.03347815
```

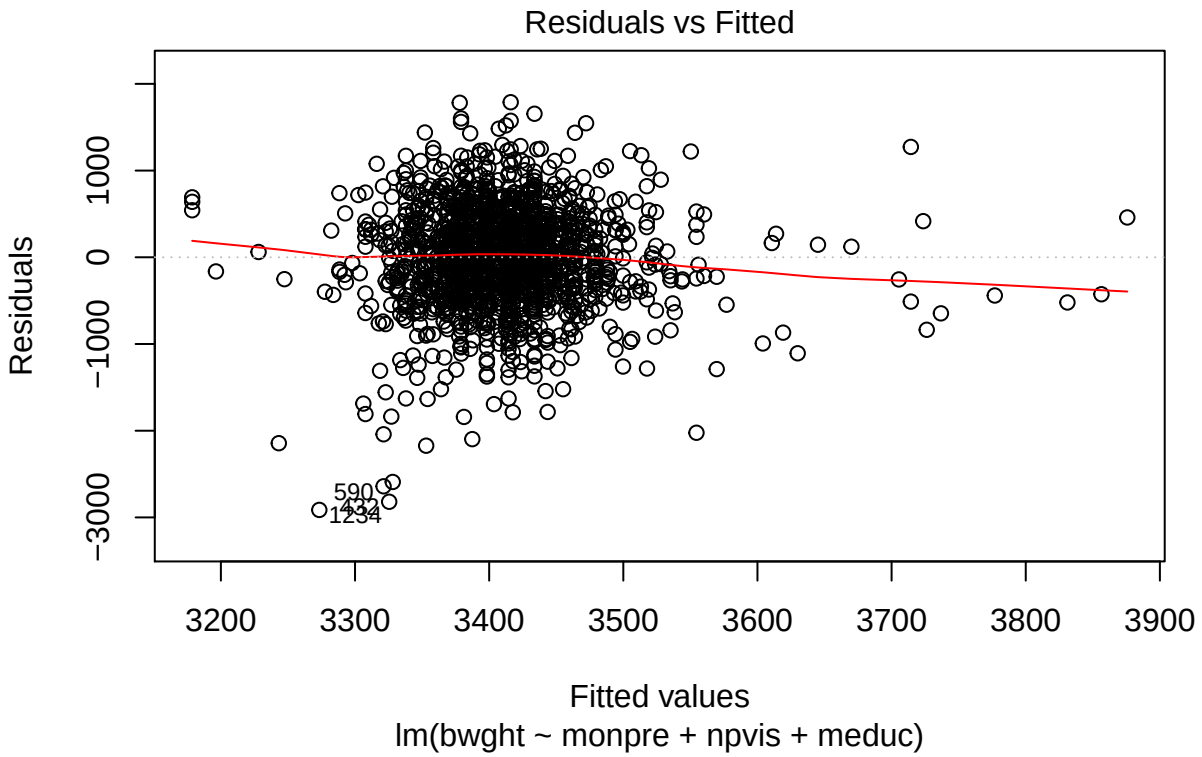
```
r_sqrd_npvis_meduc
```

```
## [1] 0.01179933
```

There is no major difference in meeting this assumption in our worse model relative to our original model. Our variance inflation factors are still < 10 and the r-squared values are still less than 1.0.

d. Zero-conditional mean

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



There is no major change in adhering to the zero-conditional mean assumption in our worse model relative to our original model.

e. Homoskedasticity

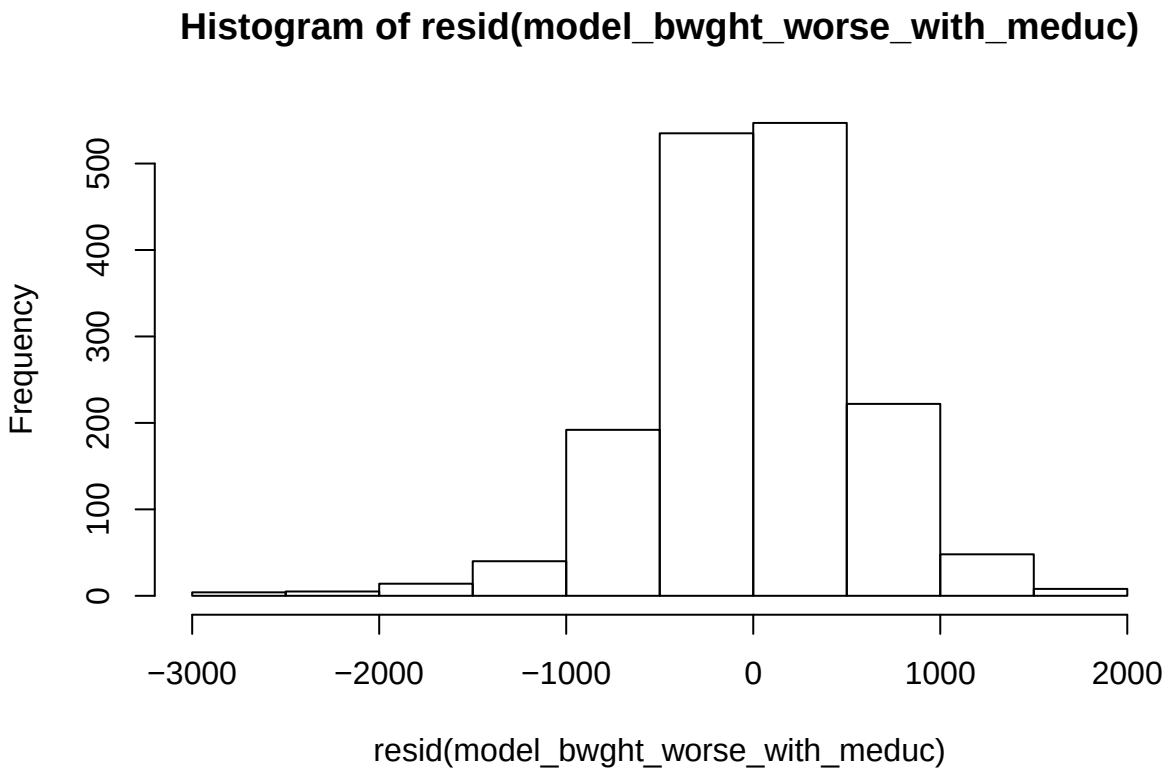
```
ncvTest(model_bwght_worse_with_meduc)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 18.19489    Df = 1     p = 1.994132e-05
```

The Score-test for non-constant error variance for our worse model shows that there is a violation of homoskedasticity, which is not changed when compared to our original model.

f. Normality of Errors

```
hist(resid(model_bwght_worse_with_meduc))
```



There is no major chance in the normality of our errors in our worse model when compared to our original model.

Regression Table Summarizing Model Results

```
se.model_bwght  <- coeftest(model_bwght, vcov=vcovHC)
se.model_bwght_better  <- coeftest(model_bwght_better, vcov=vcovHC)
se.model_bwght_worse_with_meduc_final  <- coeftest(model_bwght_worse_with_meduc_final, vcov=vcovHC)
```

```
stargazer(model_bwght, model_bwght_better, model_bwght_worse_with_meduc_final, title = "Regression Table
```

```
##
## Regression Table for Original, Better, and Worse Models
## ================================================================================================
##                                                  Dependent variable:
##                     ---------------------------------------------------------------------------
##                                                      Birthweight
##                                  (1)                    (2)                    (3)
## ------------------------------------------------------------------------------------------------
## Month prenatal care began        17.062                 16.708                 20.150
##
## Number of prenatal visits        17.549                 15.191                 15.580
##
## Cigs smoked by mother daily                             -11.126
##
```

```
## Dummy variable for Male                                    80.158
##
## Mother's education in yrs                                              8.923
##
## Constant                          3,161.271             3,168.609             3,056.212
##
## ------------------------------------------------------------------------------------------------
## Observations                        1,763                 1,655                 1,750
## R2                                  0.011                 0.020                 0.010
## Adjusted R2                         0.010                 0.018                 0.009
## Residual Std. Error       577.470 (df = 1760)   568.893 (df = 1650)   572.498 (df = 1746)
## F Statistic           9.999*** (df = 2; 1760) 8.478*** (df = 4; 1650) 6.126*** (df = 3; 1746)
## ================================================================================================
## Note:                                                *p<0.05; **p<0.01; ***p<0.001
```

In our regression table for our three models, we see that all of our models have F-statistics indicating statistical significance for the explanatory ability of each model as a whole, with all having significance at an alpha of 0.001. The p-value here is the probability of observing a value of F at least as large as we did, given that the null hypothesis is true. We can state that there is "overall significance" for our original model, our "better" model, and our "worse" model.

With respect to practical significance, we look at the Cohen's f-squared value for each of our models, as the Cohen's f-squared is one effect-size measure for F-tests in the context of regression. Cohen's f-squared = R-squared / (1-R-squared):

```
cfsqr_original <- 0.011/(1-0.011)
cfsqr_better <- 0.020/(1-0.020)
cfsqr_worse <- 0.010/(1-0.010)
cfsqr_original
```

```
## [1] 0.01112235
```

```
cfsqr_better
```

```
## [1] 0.02040816
```

```
cfsqr_worse
```

```
## [1] 0.01010101
```

The accepted standard cut-offs for small, medium, and large effect size with Cohen's f-squared are 0.02, 0.15, and 0.35 respectively. Our Cohen's f-squared values tell us that only our improved model actually achieves practical significance as measured by the effect size here. It achieves this with a small effect size. The original model and the worse model do not come close to practical significance.

Causality of Results

We see that our regression analysis has yielded important correlation between our predictor variables and our predicted variable. However, correlation does not imply or prove causation. There is no mathematical or statistical method by which to prove causation. In contemporary medical research, the research method that is seen to best-approach an argument for causation is a randomized, controlled, international, multi-centered, double-blinded experiment. Even an experiment with this level of engineering falls short of proving causation, as to prove causation one must control for all possible predictor variables that can affect the predicted variable. This is not possible.

Our data set does not even come from an experiment, but simply from a registration database. Therefore, it falls well-short of the ability to show causation. At most we can state that we have shown important

correlations, but not causation.

We discussed in-detail how our original model was most certainly affected by omitted variable bias when it did not include the cigs variable. Again, omitted variable bias (OVB) is evident when a regression model omits an important variable that explains part of the variance in the outcome variable. This leads to the introduction of bias into the estimates of the parameters in the regression analysis. We provided a thorough argument as to why there was evidence of OVB when the cigs variable was omitted from the model. What other variables are not included in our data set and analysis that likely cause OVB? One variable that is not included is maternal cocaine use in the gestational period. Maternal cocaine use is associated with conditions such as abruptio placenta, prematurity, and low birth weight.

The direction of the OVB depends on both the estimators and the covariance between the regressors and the omitted variables. A positive covariance of the omitted variable with both a regressor and the dependent variable will lead the OLS estimate of the included regressor's coefficient to be greater than the true value of that coefficient.

The absence of a variable representing cocaine use in the gestational period could cause omitted variable bias in a positive or a negative direction, depending on the prevalence of cocaine use amongst the mothers sampled. If the mothers sampled in this data set were heavy cocaine users, then the birth-weight in our original model is likely biased in a positive direction. The opposite would be true if none of the mothers sampled in our data set used cocaine during the gestational period.

Certain variables may bias results by absorbing some of the causal effect of prenatal care. We discussed this earlier when deciding whether or not to include cigs and drink in our regression. Number of cigarettes smoked per day and number of drinks consumed per week are likely intermediate outcomes in the relationship between "prenatal care" and "neonatal health". The presence or absence of the cigs variable was shown to likely bias results, as previously demonstrated.

High-level Takeaways

Through our analysis, we found that there was statistically significant predictive ability for our operationalized outcome of neonatal health (birth-weight) by a model including the predictor variables of month prenatal care began and the number of prenatal visits. Predictive ability was improved when we added the number of cigarettes smoked per day and male sex as predictor variables in the model, and made worse when we added mother's education in years as a predictor variable in the model. All of these models showed statistically significant predictive ability, whereas only our improved model showed practically significant predictive ability as measured by effect size.

As stated earlier, we cannot make any causal statements from the results of this exploratory and statistical analysis. One important finding that we noted was that there is a negative correlation between the number of cigarettes smoked per day and the neonatal birth-weight. This is in-line with previously published clinical research on this association. It would be unethical to attempt a randomized trial on this matter, as there is no clinical equipoise, in order to try and show more evidence towards causation. However, this analysis provides even more evidence that maternal smoking is associated with reduced neonatal health, and should therefore be avoided.