Lab 4: Does Prenatal Care Improve Infant Health?

w203: Statistics for Data Science

April 10, 2017

Introduction

The National Center for Health Statistics published a dataset (bwght_w203.RData) about prenatal care and infant health. Our group has been hired by a health advocacy group to study the dataset and help them understand whether prenatal care improves health outcomes for newborn infants.

The data file includes a birthweight variable. Additonally, the one- and five-minute APGAR scores are included. These are measures of the well being of infants just after birth.

Variable descriptions are provided as follows.

```
library(car)
library(leaps)
library(MASS)
library(Hmisc)
library(lmtest)
library(sandwich)
library(stargazer)
library(effsize)
```

```
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
```

```r
attach(data)
dim(data)
```

```
## [1] 1832   23
```

There are 23 variables and 1832 observations. The variables can be categorized into 4 groups:

- Father's demographic characteristics (age, education, ethnicity)
- Mother's demographic characteristics (age, education, ethnicity)
- Prenatal care
- Smoking and drinking by the mother

In this report, we are interested in learning if prenatal care improves infant health. Birth weight and the two APGAR scores are common measures of health at birth.

Exploratory analysis and model Building

We try to build a multiple linear regression to study the relationship between prenatal care and infant health. Our goal is to find a model that contains fewer predictors but can explain much of the variance in the response variable.

Determine the response variable

Among the 23 variables, bwght, omaps, fmaps, and lbwght are four potential response variables. The bwght and lbwght variables measure the birth weight and log of birth weight of newborn infants. The omaps (1-minute score) determines how well the infant tolerated the birth process and fmaps (5-minute score) informs doctor how well the baby is doing after leaving monther's womb. Thus, we believe that **fmaps** is the most accurate measurement of infant health and will use it as the response variable.

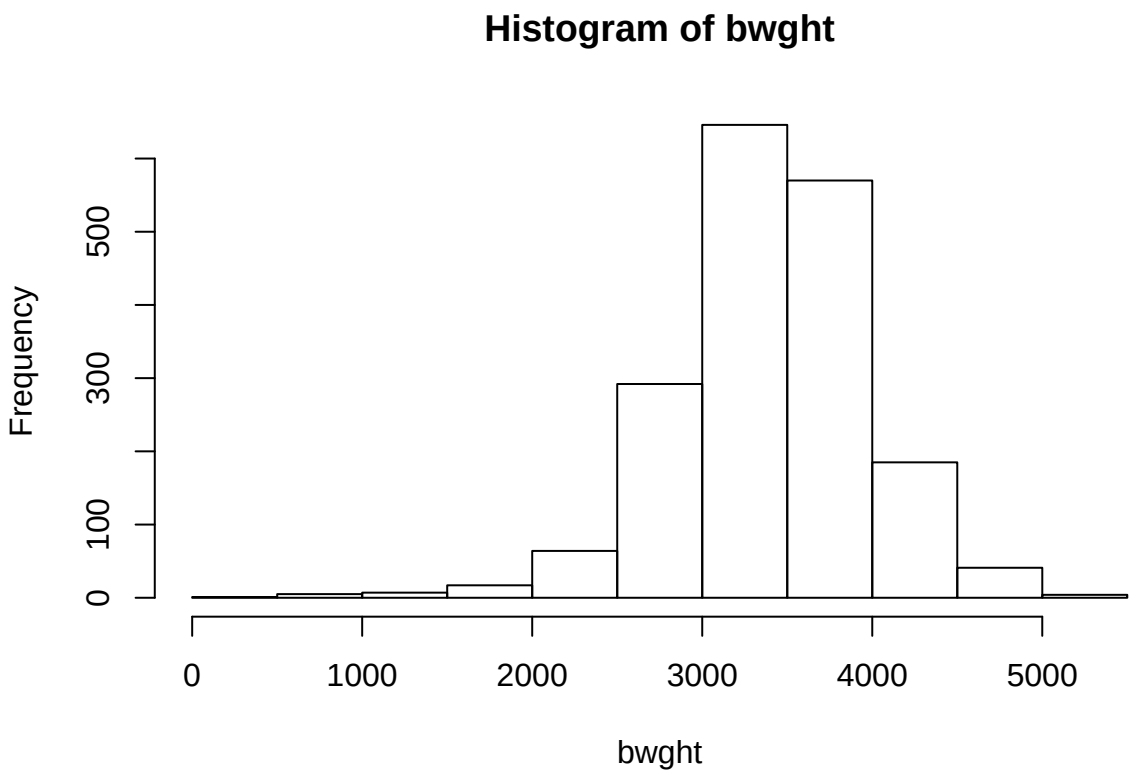
Remove irrevalent variables

To assess lbw (=1 if bwght <= 2000) and vlbw(=1 if bwght <= 1500) variables, let's take a look at the summary of bwght.

```r
summary(bwght)
```

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

```r
hist(bwght)
```




```r
table(lbw)
```

```
## lbw
##    0    1
## 1802   30
```

```r
table(vlbw)
```

```
## vlbw
##    0    1
## 1819   13
```

We see the mean of birth weight (bwght) is 3402 grams and the 1st Quantile is 3076 grams. Both lbw and vlbw variables indicate very underweight infants. lbw has 30 observations and vlbw has only 13 observations. Their sample sizes are very small. There is no practical meaning to include them in our model.

We also notice that father's demographic characteristics are not directly related to infant health. We decide to not consider father's demo in the model building process.

Create a new variable: `mrace`

Next, we look at mother's demographic characteristics. We decide to create a new variable of mother's race, called mrace, by recoding mwhte, mblck, and moth.

mrace = 1 if mother white, mrace = 2 if mother black, and mrace = 3 if mother other race.

```r
data$mrace[data$mwhte == 1] = 1 # if mother white
data$mrace[data$mblck == 1] = 2 # if mother black
data$mrace[data$moth == 1] = 3  # if mother other race
summary(data$mrace)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.000   1.168   1.000   3.000
```

```
table(data$mrace)
```

```
##
##    1    2    3
## 1624  109   99
```

We see the vast majority of mother are white (1624/1832=88.64%). Only 109+99=208 (11.35%) mothers are non-white. This contradicts to the real-life population distribution given non-white groups have much higher birth rate. We wonder this is a sampling bias but we don't look into this issue in this report.

Determine which variables to use

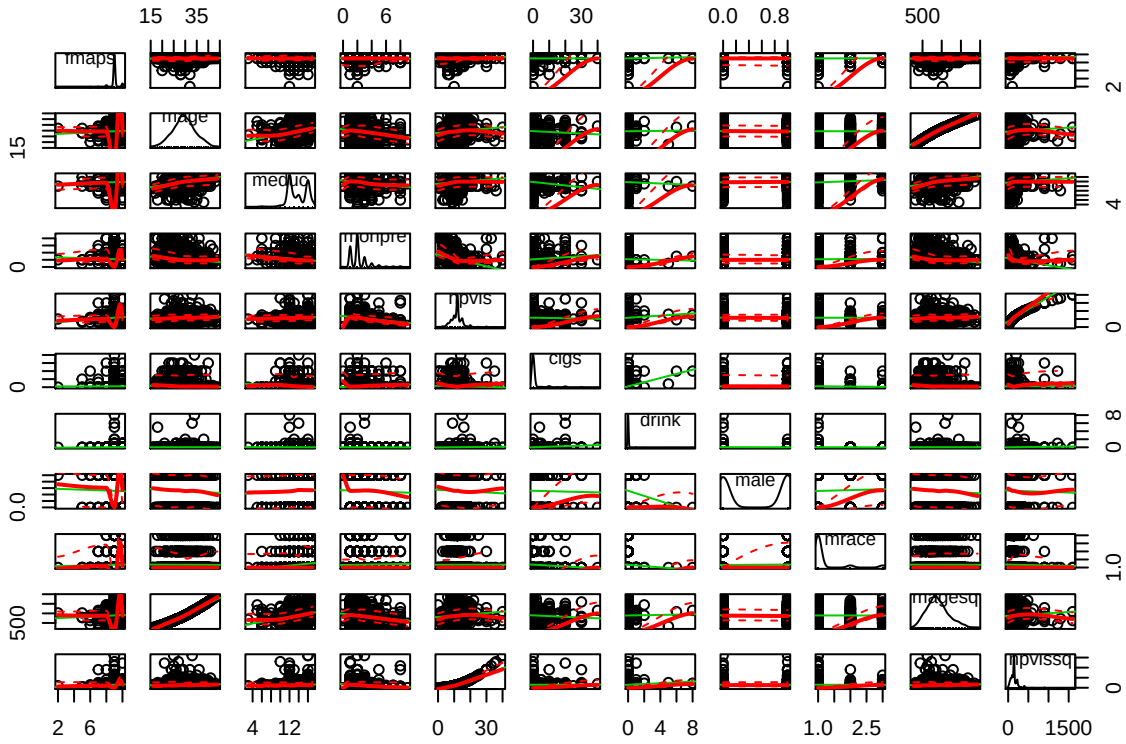
We start building our model based on the following variables:

*Response: fmaps

*Predictors: mage, meduc, monpre, npvis, cigs, drink, male, mrace, magesq, npvissq

First, we take a look at the scatterplotMatrix to roughly determine if we have a linear correlation between multiple variables. We also compute the correlation matrix to investigate the dependence between multiple variables.

```
scatterplotMatrix(data[, c("fmaps","mage","meduc","monpre","npvis","cigs","drink","male","mrace","mages
```




```
round(rcorr(as.matrix(data[, c("fmaps","mage","meduc","monpre","npvis","cigs","drink","male","mrace","ma
```

```
##           fmaps    mage   meduc  monpre   npvis    cigs   drink    male
## fmaps    1.0000  0.0299  0.0276 -0.0266  0.1022 -0.0092  0.0239 -0.0196
## mage     0.0299  1.0000  0.3210 -0.1836  0.1021 -0.0566  0.0041 -0.0421
## meduc    0.0276  0.3210  1.0000 -0.1830  0.1086 -0.1451 -0.0197  0.0304
## monpre  -0.0266 -0.1836 -0.1830  1.0000 -0.3061  0.0979 -0.0097 -0.0079
```

```
## npvis    0.1022  0.1021  0.1086 -0.3061  1.0000 -0.0387  0.0527 -0.0264
## cigs    -0.0092 -0.0566 -0.1451  0.0979 -0.0387  1.0000  0.1820 -0.0075
## drink    0.0239  0.0041 -0.0197 -0.0097  0.0527  0.1820  1.0000 -0.0466
## male    -0.0196 -0.0421  0.0304 -0.0079 -0.0264 -0.0075 -0.0466  1.0000
## mrace    0.0042 -0.0077  0.1217  0.0192 -0.0057 -0.0458 -0.0231  0.0402
## magesq   0.0306  0.9940  0.3075 -0.1654  0.0965 -0.0521  0.0071 -0.0435
## npvissq  0.0666  0.0607  0.0732 -0.1799  0.9341  0.0052  0.0504 -0.0124
##           mrace  magesq npvissq
## fmaps    0.0042  0.0306  0.0666
## mage    -0.0077  0.9940  0.0607
## meduc    0.1217  0.3075  0.0732
## monpre   0.0192 -0.1654 -0.1799
## npvis   -0.0057  0.0965  0.9341
## cigs    -0.0458 -0.0521  0.0052
## drink   -0.0231  0.0071  0.0504
## male     0.0402 -0.0435 -0.0124
## mrace    1.0000 -0.0055 -0.0145
## magesq  -0.0055  1.0000  0.0569
## npvissq -0.0145  0.0569  1.0000
```

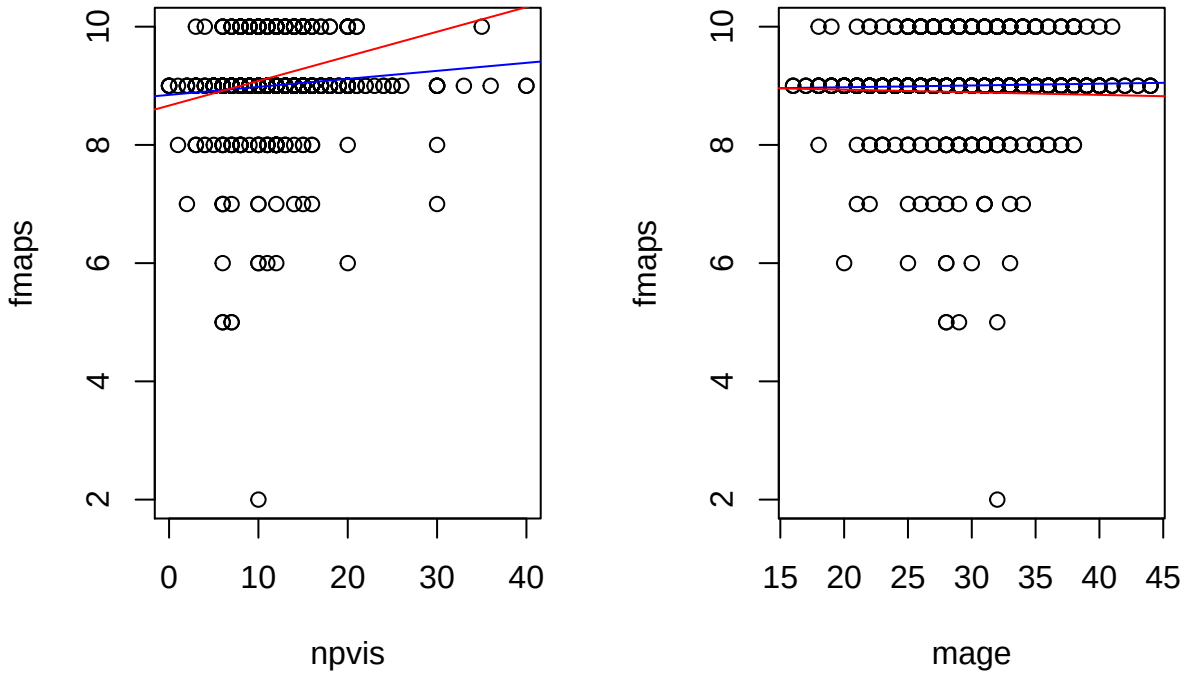
Based on the correlation matrix, monpre, cigs, and male have slightly negative correlation with fmaps. npvis and npvissq have significant positive correlation with fmaps. mage, meduc, drink, mrace, and magesq have slightly positive correlation with fmaps but the correlation coefficients are insignificant.

Note that magesq and npvissq are mage^2 and npvis^2. As discussed in class before, the linear regression assumes a linear relationship between the response and predictors. But in some cases, the true relationship between the response and the predictor may be non-linear. In this case we have to extend the linear model to accommodate non-linear relationships, using polynomial regression.

In fact, we do intend to include the **quadratic transformation** in our model because mother's age (mage) and total number of prenatal visits (npvis) have non-linear relationship with fmaps. Very young (<20) and very old mother (>35) are possible to have poor health infants, while middle aged mother are more likely to give birth to healty infants. Similarly, mothers who need prenatal care frequently are more likely to have poor health infants. So the relationship between fmaps and npvis is not perfectly linear.

Consider the following scatter plots of npvis vs. fmaps and mage vs. fmaps. The number of npvis, mage, and fmaps are shown.

```
par(mfrow=c(1,2))
plot(npvis,fmaps)
abline(lm(fmaps~npvis, data=data), col="blue")
abline(lm(fmaps~npvis+npvissq, data=data), col="red")

plot(mage,fmaps)
abline(lm(fmaps~mage, data=data), col="blue")
abline(lm(fmaps~mage+magesq, data=data), col="red")
```




```
par(mfrow=c(1,1))
```

For each plot, the blue line represents the linear regression fit. Although it appears to have a linear relationship, we think the data suggest a curved relationship. The red curve contains the quadratic term and may provide a better fit. After all, it is still a linear model.

Build our models

We start this process by testing multiple models. Our three model specifications: `fit1`, `fit2` and `fit3` will be presented at the end of this section, and we will do a thorough review of the three models performance.

First, We fit a model called `null1` with zero predictor and full model called `full1` containing all possible predictors.

```
# null model
null1 = lm(fmaps ~ 1, data = data)
summary(null1)
```

```
##
## Call:
## lm(formula = fmaps ~ 1, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.0038 -0.0038 -0.0038 -0.0038  0.9962
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.00383    0.01122   802.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4799 on 1828 degrees of freedom
```

```
##   (3 observations deleted due to missingness)
```

```
# full model
full1 = lm(fmaps ~mage+meduc+monpre+npvis+cigs+drink+male+mrace+magesq+npvissq, data=data)
summary(full1)
```

```
##
## Call:
## lm(formula = fmaps ~ mage + meduc + monpre + npvis + cigs + drink +
##     male + mrace + magesq + npvissq, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9855 -0.0504 -0.0196  0.0285  1.1769
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.589e+00  3.609e-01  23.803  < 2e-16 ***
## mage        -1.301e-03  2.341e-02  -0.056  0.95569
## meduc        2.627e-03  6.183e-03   0.425  0.67095
## monpre       1.328e-02  1.088e-02   1.220  0.22273
## npvis        4.126e-02  9.853e-03   4.187 2.98e-05 ***
## cigs         1.470e-03  2.983e-03   0.493  0.62219
## drink        2.721e-02  4.094e-02   0.665  0.50637
## male        -1.938e-03  2.388e-02  -0.081  0.93534
## mrace        7.027e-03  2.373e-02   0.296  0.76718
## magesq       5.597e-05  3.900e-04   0.143  0.88592
## npvissq     -9.529e-04  3.074e-04  -3.100  0.00197 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4799 on 1621 degrees of freedom
##   (200 observations deleted due to missingness)
## Multiple R-squared:  0.01695,    Adjusted R-squared:  0.01089
## F-statistic: 2.795 on 10 and 1621 DF,  p-value: 0.001956
```

The model p-value: 0.001956 shows that the full model is statistical significant different than the null model. Besides, variables npvis and npvissq are statistically significant at 95%. The coefficients' p-values indicate that npvis and npvissq are related to fmaps. Note that the coefficient estimate of npvis is positive but the coefficient estimate of npvissq is negative. This means that if a mother seeks a large number of prenantal care, it's likely to have a poor health infant. The adjusted $R^2 = 0.01089$. The model explains 1.089% of the variance in the response varialbe fmaps.

Based on the full model, we use `setp()` function to perform **variable selection** by running backward elimination and stepwise regression.

```
data1 <- na.omit(data[, c("fmaps","mage","meduc","monpre","npvis","cigs","drink","male","mrace","magesq
full1.1 = lm(fmaps ~mage+meduc+monpre+npvis+cigs+drink+male+mrace+magesq+npvissq, data=data1)
null1.1 = lm(fmaps ~1, data=data1)
step(null1.1, scope = list(upper=full1.1), data=data1, direction="both")
```

```
## Start:  AIC=-2377.51
## fmaps ~ 1
##
##           Df Sum of Sq    RSS     AIC
## + npvis    1    3.7263 376.03 -2391.6
```

```
## + npvissq  1    1.7003 378.05 -2382.8
## <none>                 379.75 -2377.5
## + magesq   1    0.4420 379.31 -2377.4
## + mage     1    0.4408 379.31 -2377.4
## + monpre   1    0.3100 379.44 -2376.8
## + meduc    1    0.2852 379.47 -2376.7
## + drink    1    0.2176 379.54 -2376.4
## + mrace    1    0.0496 379.71 -2375.7
## + male     1    0.0268 379.73 -2375.6
## + cigs     1    0.0015 379.75 -2375.5
##
## Step:  AIC=-2391.61
## fmaps ~ npvis
##
##           Df Sum of Sq    RSS     AIC
## + npvissq  1    1.9719 374.06 -2398.2
## <none>                 376.03 -2391.6
## + magesq   1    0.2389 375.79 -2390.6
## + mage     1    0.2291 375.80 -2390.6
## + drink    1    0.1334 375.90 -2390.2
## + meduc    1    0.1112 375.92 -2390.1
## + mrace    1    0.0487 375.98 -2389.8
## + male     1    0.0154 376.01 -2389.7
## + cigs     1    0.0128 376.02 -2389.7
## + monpre   1    0.0030 376.03 -2389.6
## - npvis    1    3.7263 379.75 -2377.5
##
## Step:  AIC=-2398.19
## fmaps ~ npvis + npvissq
##
##           Df Sum of Sq    RSS     AIC
## <none>                 374.06 -2398.2
## + monpre   1    0.2954 373.76 -2397.5
## + drink    1    0.1364 373.92 -2396.8
## + magesq   1    0.1327 373.92 -2396.8
## + mage     1    0.1216 373.94 -2396.7
## + cigs     1    0.0825 373.97 -2396.6
## + meduc    1    0.0488 374.01 -2396.4
## + mrace    1    0.0292 374.03 -2396.3
## + male     1    0.0072 374.05 -2396.2
## - npvissq  1    1.9719 376.03 -2391.6
## - npvis    1    3.9978 378.05 -2382.8
##
## Call:
## lm(formula = fmaps ~ npvis + npvissq, data = data1)
##
## Coefficients:
## (Intercept)        npvis      npvissq
##   8.7065925    0.0371461   -0.0008438
```

```
step(full1.1, data=data1, direction = "backward")
```

```
## Start:  AIC=-2385.42
## fmaps ~ mage + meduc + monpre + npvis + cigs + drink + male +
```

```
##     mrace + magesq + npvissq
##
##           Df Sum of Sq    RSS     AIC
## - mage     1    0.0007 373.32 -2387.4
## - male     1    0.0015 373.32 -2387.4
## - magesq   1    0.0047 373.32 -2387.4
## - mrace    1    0.0202 373.34 -2387.3
## - meduc    1    0.0416 373.36 -2387.2
## - cigs     1    0.0559 373.37 -2387.2
## - drink    1    0.1017 373.42 -2387.0
## - monpre   1    0.3426 373.66 -2385.9
## <none>                 373.32 -2385.4
## - npvissq  1    2.2129 375.53 -2377.8
## - npvis    1    4.0377 377.36 -2369.9
##
## Step:  AIC=-2387.41
## fmaps ~ meduc + monpre + npvis + cigs + drink + male + mrace +
##     magesq + npvissq
##
##           Df Sum of Sq    RSS     AIC
## - male     1    0.0015 373.32 -2389.4
## - mrace    1    0.0205 373.34 -2389.3
## - meduc    1    0.0409 373.36 -2389.2
## - cigs     1    0.0562 373.37 -2389.2
## - drink    1    0.1022 373.42 -2389.0
## - magesq   1    0.1362 373.45 -2388.8
## - monpre   1    0.3553 373.67 -2387.9
## <none>                 373.32 -2387.4
## - npvissq  1    2.2148 375.53 -2379.8
## - npvis    1    4.0410 377.36 -2371.8
##
## Step:  AIC=-2389.41
## fmaps ~ meduc + monpre + npvis + cigs + drink + mrace + magesq +
##     npvissq
##
##           Df Sum of Sq    RSS     AIC
## - mrace    1    0.0203 373.34 -2391.3
## - meduc    1    0.0403 373.36 -2391.2
## - cigs     1    0.0562 373.38 -2391.2
## - drink    1    0.1035 373.42 -2390.9
## - magesq   1    0.1381 373.46 -2390.8
## - monpre   1    0.3578 373.68 -2389.8
## <none>                 373.32 -2389.4
## - npvissq  1    2.2229 375.54 -2381.7
## - npvis    1    4.0569 377.38 -2373.8
##
## Step:  AIC=-2391.32
## fmaps ~ meduc + monpre + npvis + cigs + drink + magesq + npvissq
##
##           Df Sum of Sq    RSS     AIC
## - meduc    1    0.0487 373.39 -2393.1
## - cigs     1    0.0546 373.39 -2393.1
## - drink    1    0.1021 373.44 -2392.9
## - magesq   1    0.1341 373.47 -2392.7
```

```
## - monpre   1    0.3674 373.71 -2391.7
## <none>                 373.34 -2391.3
## - npvissq  1    2.2439 375.58 -2383.5
## - npvis    1    4.0860 377.43 -2375.6
##
## Step:  AIC=-2393.1
## fmaps ~ monpre + npvis + cigs + drink + magesq + npvissq
##
##           Df Sum of Sq    RSS     AIC
## - cigs     1    0.0434 373.43 -2394.9
## - drink    1    0.1014 373.49 -2394.7
## - magesq   1    0.2016 373.59 -2394.2
## - monpre   1    0.3434 373.73 -2393.6
## <none>                 373.39 -2393.1
## - npvissq  1    2.2477 375.64 -2385.3
## - npvis    1    4.1039 377.49 -2377.3
##
## Step:  AIC=-2394.92
## fmaps ~ monpre + npvis + drink + magesq + npvissq
##
##           Df Sum of Sq    RSS     AIC
## - drink    1    0.1329 373.56 -2396.3
## - magesq   1    0.1952 373.63 -2396.1
## - monpre   1    0.3564 373.79 -2395.4
## <none>                 373.43 -2394.9
## - npvissq  1    2.2080 375.64 -2387.3
## - npvis    1    4.0606 377.49 -2379.3
##
## Step:  AIC=-2396.33
## fmaps ~ monpre + npvis + magesq + npvissq
##
##           Df Sum of Sq    RSS     AIC
## - magesq   1    0.1964 373.76 -2397.5
## - monpre   1    0.3591 373.92 -2396.8
## <none>                 373.56 -2396.3
## - npvissq  1    2.2068 375.77 -2388.7
## - npvis    1    4.0864 377.65 -2380.6
##
## Step:  AIC=-2397.48
## fmaps ~ monpre + npvis + npvissq
##
##           Df Sum of Sq    RSS     AIC
## - monpre   1    0.2954 374.06 -2398.2
## <none>                 373.76 -2397.5
## - npvissq  1    2.2643 376.03 -2389.6
## - npvis    1    4.1897 377.95 -2381.3
##
## Step:  AIC=-2398.19
## fmaps ~ npvis + npvissq
##
##           Df Sum of Sq    RSS     AIC
## <none>                 374.06 -2398.2
## - npvissq  1    1.9719 376.03 -2391.6
## - npvis    1    3.9978 378.05 -2382.8
```

```
##
## Call:
## lm(formula = fmaps ~ npvis + npvissq, data = data1)
##
## Coefficients:
## (Intercept)        npvis      npvissq
##   8.7065925    0.0371461   -0.0008438
```

Both backward elimination and stepwise regression give us the same results. Variables npvis and npvissq should be included in our best model.

Recall the correlation matrix, the correlation between fmaps and meduc is 0.03 and the correlation between fmaps and mrace is approximately 0.00. We believe that meduc and mrace have very slim impact on infant health. The scatter plots with abline are shown below.

We also suspect that cigarettes and drinks have very slim impact on infant health because the correlation between fmaps and cigs is -0.01 and the correlation between fmaps and drink is 0.02. The correlation coefficients are very small.

In addition, the correlation between fmaps and male is -0.02. Male infants have slightly worse fmaps score. The correlation is insignificant at 95%. We perform a **two-sample independent t-test** to compare the mean fmaps value for male and femle infants.

Null hypothesis: mean fmaps value for male is equal to mean fmaps value for female

Alternative hypothesis: mean fmaps value for male is NOT equal to mean fmaps value for female

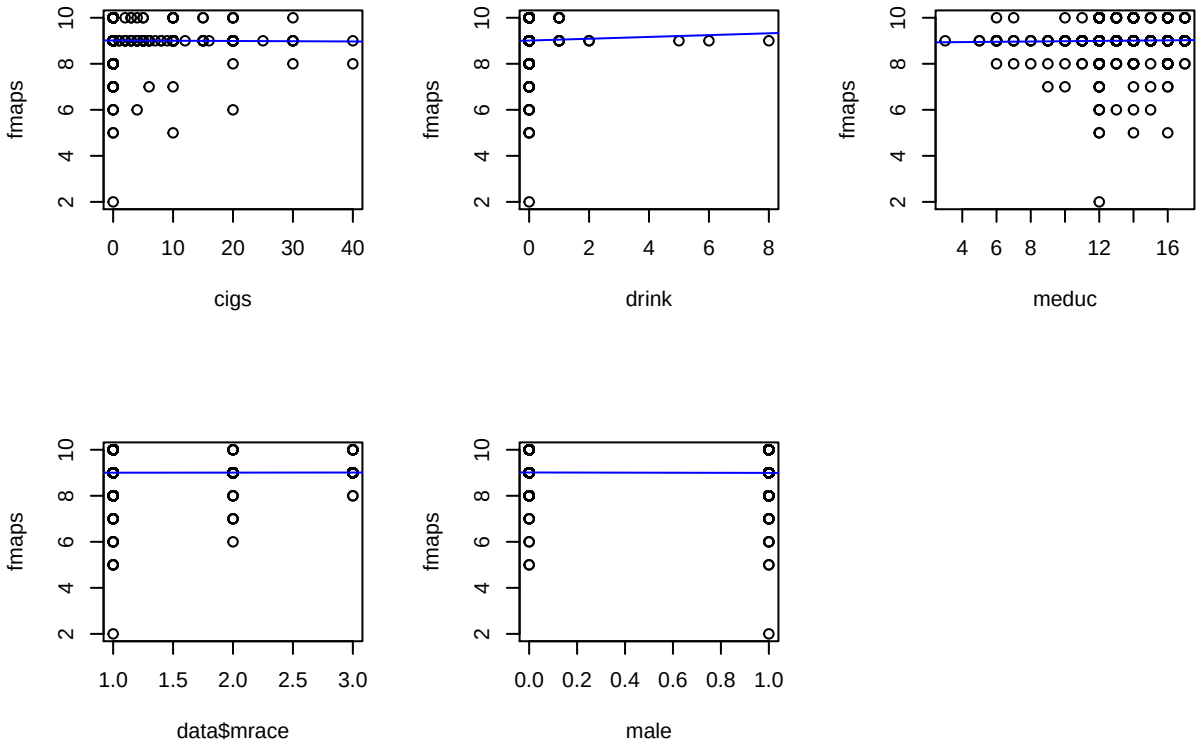
```
# t-test comparision of male and female fmaps score.
male.fmaps <- data$fmaps[data$male == 1]
female.fmaps <- data$fmaps[data$male == 0]
t.test(male.fmaps, female.fmaps)
```

```
##
##  Welch Two Sample t-test
##
## data:  male.fmaps and female.fmaps
## t = -0.84238, df = 1766.5, p-value = 0.3997
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.06256692  0.02496991
## sample estimates:
## mean of x mean of y
##  8.994670  9.013468
```

The mean fmaps score for male infants is 8.995, and the mean fmaps score for female infants is 9.01. Since the p-value = 0.3997, we do not reject the null hypothesis and conclude that there is NO difference in mean fmaps between male infants and female infants.

```
par(mfrow=c(2,3))
plot(cigs,fmaps)
abline(lm(fmaps~cigs, data = data), col="blue")

plot(drink,fmaps)
abline(lm(fmaps~drink, data = data), col="blue")

plot(meduc, fmaps)
abline(lm(fmaps~meduc, data=data), col="blue")

plot(data$mrace, fmaps)
```

```
abline(lm(fmaps~mrace, data=data), col="blue")

plot(male, fmaps)
abline(lm(fmaps~male, data=data), col="blue")
```



For each scatter plot, the regression abline appears to be a constant line. This means that cigs, drink, meduc, mrace, and male have slim or virtually no impact on fmaps. Thus we can exclude those five variables and fit our second model `full2` with mage, monpre, npvis, magesq and npvissq.

```
full2=lm(fmaps ~mage+monpre+npvis+magesq+npvissq, data=data)
summary(full2)
```

```
##
## Call:
## lm(formula = fmaps ~ mage + monpre + npvis + magesq + npvissq,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9815 -0.0442 -0.0169  0.0358  1.1955
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.6766994  0.3451582  25.138  < 2e-16 ***
## mage        -0.0067800  0.0224093  -0.303 0.762268
## monpre       0.0136330  0.0105922   1.287 0.198236
## npvis        0.0460754  0.0096050   4.797 1.75e-06 ***
## magesq       0.0001394  0.0003739   0.373 0.709396
## npvissq     -0.0010895  0.0003023  -3.604 0.000323 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.4843 on 1754 degrees of freedom
##   (72 observations deleted due to missingness)
## Multiple R-squared:  0.01818,    Adjusted R-squared:  0.01538
## F-statistic: 6.494 on 5 and 1754 DF,  p-value: 5.42e-06
```

The p-value: 5.42e-06 shows that the `full2` model is statistically significant different than the null model. Variables npvis and npvissq are statistically significant at 95%. The coefficients' p-values indicate that both npvis and npvissq are related to fmaps. The adjusted $R^2 = 0.01538$. The model explains 1.538% of the variance in the response varialbe fmaps.

To determine whether to include both mage and magesq in the same model. We use `anova()` function to perform a hypothesis test comparing the two models.

Null hypothesis: the two models fit the data equally well.

Alternative hypothesis: the full model fits data better than the reduced model.

```
m0 <- lm(fmaps~mage, data=data)
m1 <- lm(fmaps~mage+magesq, data=data)
anova(m0, m1)
```

```
## Analysis of Variance Table
##
## Model 1: fmaps ~ mage
## Model 2: fmaps ~ mage + magesq
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   1827 420.60
## 2   1826 420.57  1  0.027676 0.1202 0.7289
```

The F-statistic is 0.1202 and the associated p-value is $0.7289 > 0.05$. So we do not reject the null and conclude that the two models fit the data equally well and we don't have evidence to include both mage and magesq. We decide to include only mage in the model.

Similarly, we compare the two models for npvis and npvissq.

```
m2 <- lm(fmaps~npvis, data=data)
m3 <- lm(fmaps~npvis+npvissq, data=data)
anova(m2,m3)
```

```
## Analysis of Variance Table
##
## Model 1: fmaps ~ npvis
## Model 2: fmaps ~ npvis + npvissq
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   1759 414.60
## 2   1758 411.86  1    2.7345 11.672 0.0006489 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-statistic is 11.672 and the associated p-value is $0.0006489 < 0.05$. So we reject the null and conclude that the model containing both npvis and npvissq is far superior to the model that only contains npvis.

Finally, we have four predictors (mage, monpre, npvis and npvissq) to construct the model with only the explanatory variables of key interest.

Three model specifications

Model with only the explanatory variables of key interest

Our first model, `fit1`, contains predictors: mage, monpre, npvis and npvissq. They are the explanatory variables of key interest.

```
fit1 = lm(fmaps ~mage+monpre+npvis+npvissq, data=data)
summary(fit1)
```

```
##
## Call:
## lm(formula = fmaps ~ mage + monpre + npvis + npvissq, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9840 -0.0435 -0.0145  0.0348  1.1949
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.5547324  0.1097723  77.932  < 2e-16 ***
## mage         0.0015219  0.0024663   0.617 0.537267
## monpre       0.0142141  0.0104743   1.357 0.174941
## npvis        0.0461193  0.0096019   4.803 1.69e-06 ***
## npvissq     -0.0010910  0.0003022  -3.610 0.000315 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4842 on 1755 degrees of freedom
##   (72 observations deleted due to missingness)
## Multiple R-squared:  0.0181, Adjusted R-squared:  0.01586
## F-statistic: 8.087 on 4 and 1755 DF,  p-value: 1.85e-06
```

The model p-value: 1.85e-06 shows that the `fit1` model is statistically significant different than the null model. Variables npvis and npvissq are statistically significant at 95%. The coefficients' p-values indicate that npvis and npvissq are related to fmaps. The coefficient estmate for npvis is 0.046, which means if there is one unit increase in npvis, fmaps value will increase by 0.046. Note that the coefficient estimate of npvis is positive but the coefficient estimate of npvissq is negative. This means that if a mother seeks a large number of prenantal care, it's likely to have a poor health infant. The adjusted $R^2 = 0.01586$ shows that `fit1` fits the data better than full1 and full2. The model explains 1.586% of the variance in the response varialbe fmaps.

Model that includes only covariates that you believe increase the accuracy of the results without introducing bias.

Our second model, `fit2`, contains an interaction term mage×npvis. We believe that npvis has a different effect on fmaps depending on the values of mage.

```
fit2=lm(fmaps ~mage+monpre+npvis+npvissq+mage*npvis, data=data)
summary(fit2)
```

```
##
## Call:
## lm(formula = fmaps ~ mage + monpre + npvis + npvissq + mage *
##     npvis, data = data)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9921 -0.0437 -0.0148  0.0328  1.2844
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.0960919  0.2526650  32.043  < 2e-16 ***
## mage         0.0169247  0.0080317   2.107 0.035237 *
## monpre       0.0169426  0.0105524   1.606 0.108549
## npvis        0.0861126  0.0220452   3.906 9.73e-05 ***
## npvissq     -0.0011302  0.0003026  -3.735 0.000194 ***
## mage:npvis  -0.0013361  0.0006631  -2.015 0.044063 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4837 on 1754 degrees of freedom
##   (72 observations deleted due to missingness)
## Multiple R-squared:  0.02037,    Adjusted R-squared:  0.01757
## F-statistic: 7.293 on 5 and 1754 DF,  p-value: 8.983e-07
```

```
#compare two models
anova(fit1, fit2)
```

```
## Analysis of Variance Table
##
## Model 1: fmaps ~ mage + monpre + npvis + npvissq
## Model 2: fmaps ~ mage + monpre + npvis + npvissq + mage * npvis
##   Res.Df    RSS Df Sum of Sq    F  Pr(>F)
## 1   1755 411.39
## 2   1754 410.44  1   0.95005 4.06 0.04406 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here, the Adjusted R-squared=0.01757, which has increased from 0.01586 in `fit1`. Based on the coefficients' p-value, variables mage, npvis, npvissq, and mage×npvis are all significant at 95%. The `anova()` function compares two models `fit1` and `fit2` and produces p-value = 0.04406 < 0.05. This provides evidence that the model `fit2` is superior to `fit1`. Thus, adding the interaction term has increased the accuracy of our results without introducing bias. The model explains 1.757% of the variance in the response varialbe fmaps.

Model that includes the previous covariates, but also covariates that may be problematic for one reason or another

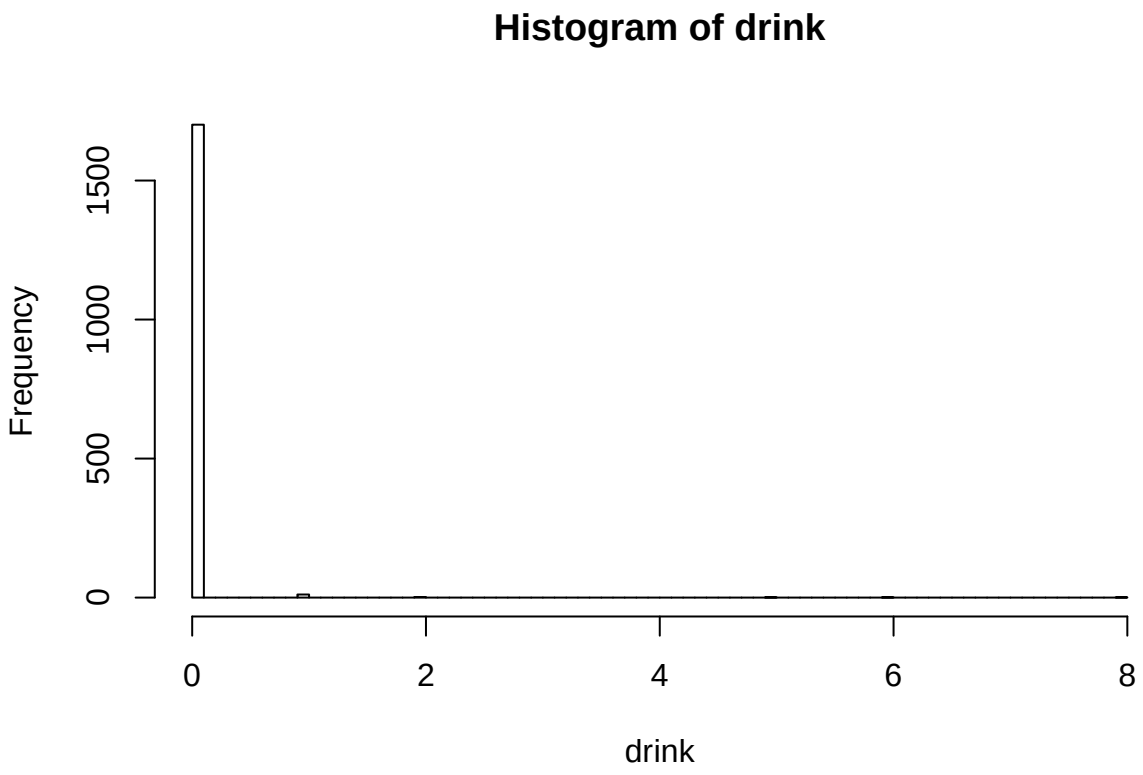
We find that variable drink may be problematic. Recall the correlation matrix, the correlation between fmaps and drink is 0.024, which is positive. However, we normally believe drink should have a negative relationship with infant health. The distribution of drink is actually very skewed. 1701 observations have zero drinks. Only 16 observations have positive drink values. The sample size of drink is too small so it's problematic if we include drink in our model.

The model `fit3` contains a problematic variable drink.

```
table(drink)
```

```
## drink
##    0    1    2    5    6    8
## 1701   11    2    1    1    1
```

```r
hist(drink, breaks = 100)
```




```r
fit3=lm(fmaps ~mage+monpre+npvis+npvissq+mage*npvis+drink, data=data)
summary(fit3)
```

```
##
## Call:
## lm(formula = fmaps ~ mage + monpre + npvis + npvissq + mage *
##     npvis + drink, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.0013 -0.0516 -0.0230  0.0309  1.2771
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.1267055  0.2594470  31.323  < 2e-16 ***
## mage         0.0166197  0.0081682   2.035 0.042043 *
## monpre       0.0153515  0.0107435   1.429 0.153222
## npvis        0.0817586  0.0225706   3.622 0.000301 ***
## npvissq     -0.0010759  0.0003051  -3.527 0.000433 ***
## drink        0.0282059  0.0401106   0.703 0.482029
## mage:npvis  -0.0012434  0.0006724  -1.849 0.064608 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4793 on 1640 degrees of freedom
##   (185 observations deleted due to missingness)
## Multiple R-squared:  0.02091,    Adjusted R-squared:  0.01733
## F-statistic: 5.837 on 6 and 1640 DF,  p-value: 4.932e-06
```

The Adjusted R-squared= 0.01733, which is slight less than model `fit2`.

CLASSICAL LINEAR MODEL (CLM)

The following 6 assumptions will be explained based on our `fit1` model (`fit1=lm(fmaps ~mage+npvis+npvissq, data=data)`).

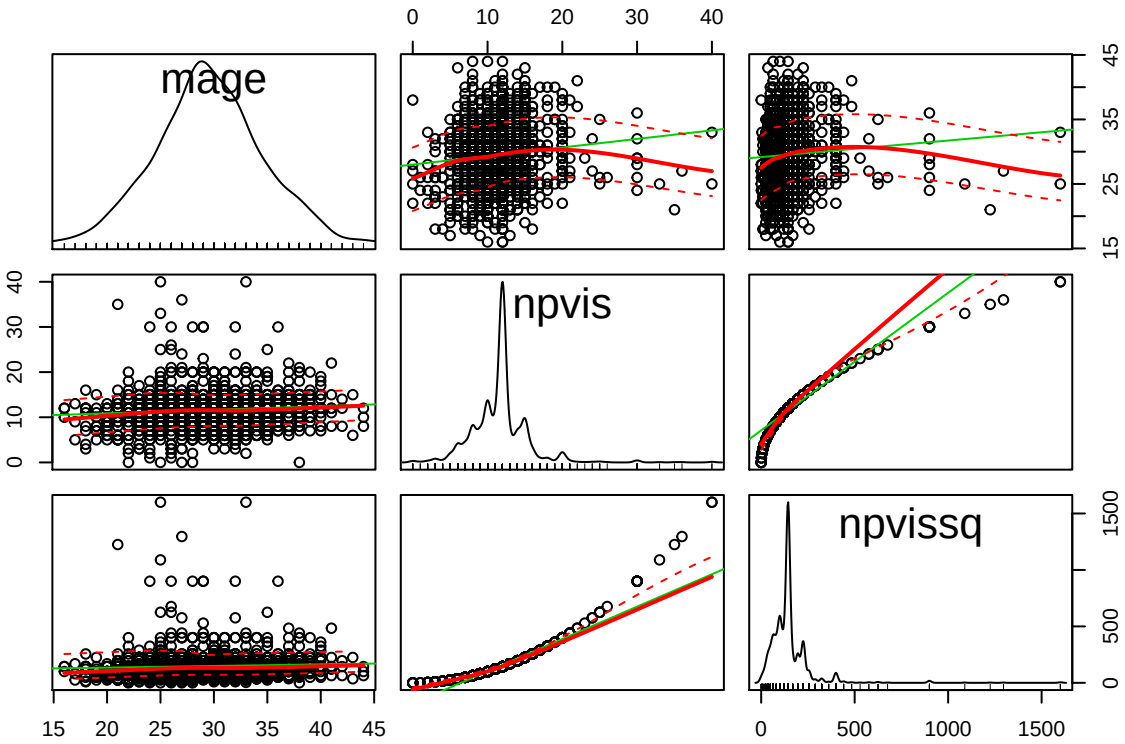
**NOTE: Explanations below are for fit1 and fit2. Fit3 represents a model where problem causing variables are induced so we did not waste time analyzing this model as we know it will violate these assumpitons further and rather explored our important models in better detail.

1. CLM 1 - Linearity Though there is no test or rigorous proof that displays why the model is linear, we can look at the actual model and note that our explanatory variables themselves take the linear and polynomial form (squared value of number of visits), and the coefficients are left linear in nature. (Same applies for fit2).
2. CLM 2 - Random Sampling It is important to note that no background was given as to how this data was sampled. We do not know if this is in fact a random sample that represents the true population sampled from. As we have described earlier in this report, examination of one of the variables, race, allows us to see the disparity in the number of white versus black and other races of the mothers. Is this an accurate representation of the population or is this the result of a sampling bias? We cannot determine this. However, to fit and analyze a linear model, we mus assume that random sample (but we are weary of it). (Same applies for fit2)
3. CLM 3 - Multi-Collinearity?

```
vif(fit1)
```

```
##     mage   monpre    npvis  npvissq
## 1.042291 1.251724 9.358841 8.744732
```

```
scatterplotMatrix(data[ , c("mage", "npvis", "npvissq")])
```



When we run vif on our fit1 model, we see here that the VIF (1.04, 1.25) for 'mage' and 'monpre' are relatively low so they do not cause too much concern. However, the VIF for both 'npvis' and 'npvissq' are 9.35 and 8.74, which are prett high VIF values. However, we note that 'npvis' and 'npvissq' are transformations of each other, allowing to expect some correlation. As explained in the introduction, we are aware and intend on including both the original and the transformed variable in our model. (We note the correlation from our scatterplot matrix as well). But none of our VIF are 10 or greater, so there are likely no LARGE standard errors that would affect our model.

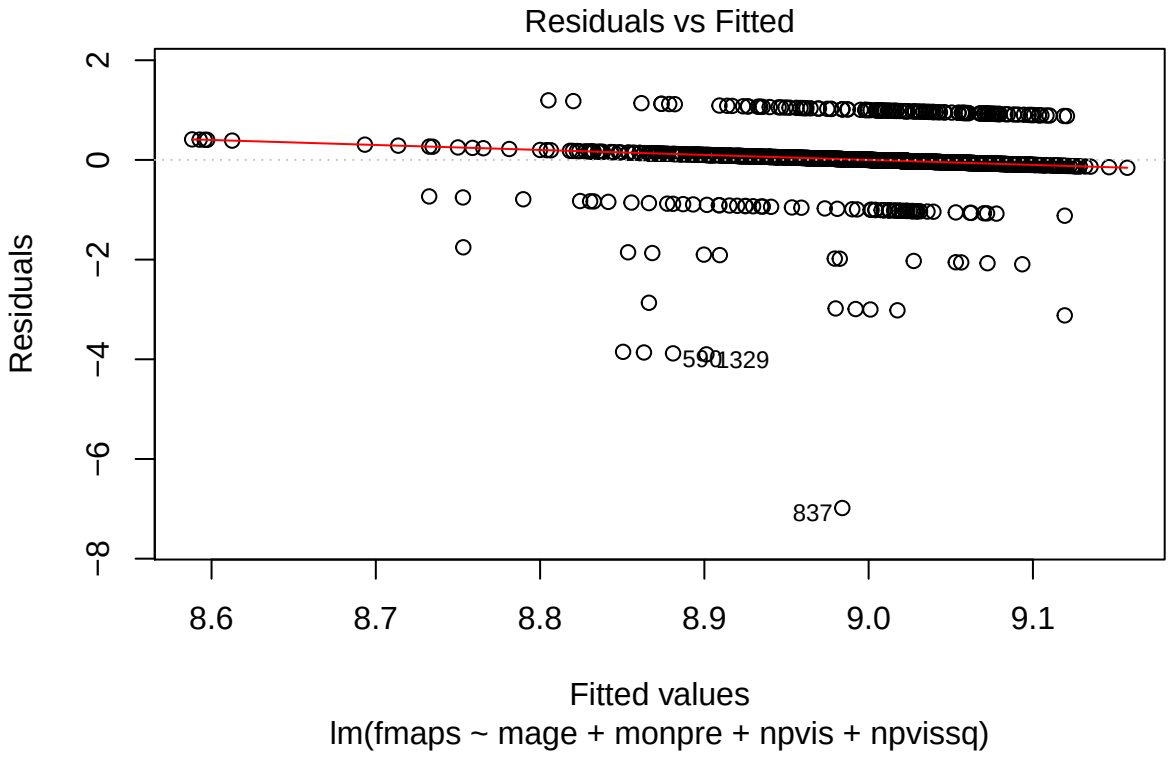
```
vif(fit2)
```

```
##       mage     monpre      npvis    npvissq mage:npvis
##  11.073409   1.272682  49.419115   8.781042  52.246988
```

Looking at the VIF values for model for our fit2 causes a little more concern because of the high value for mage:npvis, but again this makes sense as it is a variable transformation from two of our variables already in the model, so correlation is expected. Once again, we refer to our introduction as to our explanation for keeping the variables.

4. CLM 4 - Zero Conditional Mean (CLM 4' - Exogeneity)

```
plot(fit1, which = 1)
```



Looking at our Residual vs. Fitted values plot for our fit1 model, right off the bat we can see an interesting result. It seems like we do not have a collective group of data, but rather we have groupings (layers) of data points in a stepwise manner. We need to remember that our outcome variable, 'fmaps' is an ordinal variable, an integer value form 1-10, not a continous variable. Our plot reflects the ordinal nature of the variable.

But looking at the model overall, it seems to relatively keep with the zero conditional mean assumption, wiht a slight skew upward near the left side of the plot. The slight skew could cause us to assume a less stringent version of the zero conditional mean: exogeneity, where our variables are uncorrelated with the error term.

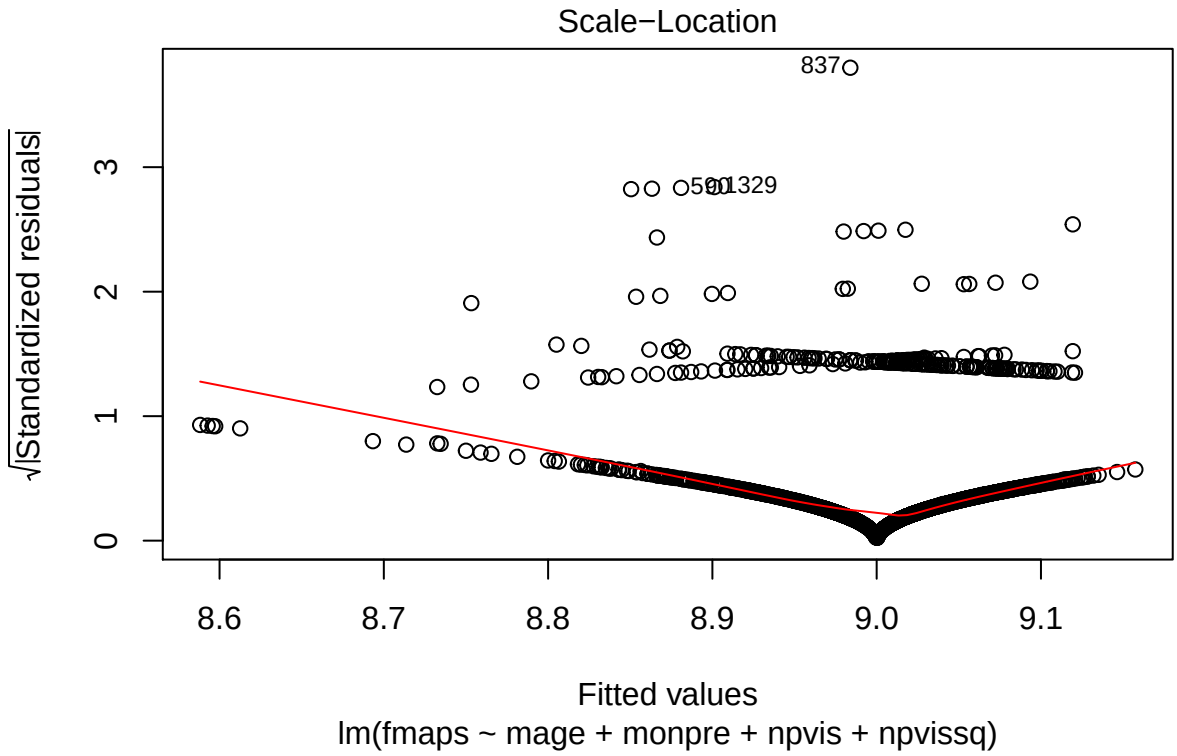
However, let us for one second consider our model and think about the possibility of endogeneity within our variables for fit1 (mage, monpre, npvis, npvissq). Mother's age is likely to not have any other hidden

factor that could effect our error, likewise for starting month of prenatal care. However with number of visits, we might be able to deduce another factor: wealth from this. If a mother can afford more visits, it is not unreasonable to assume that a baby's prenatal care in a wealthy family differs in treatment than that of a low income family. This is a bit of a stretch and an extrapolation that we cannot necessarily deduce from the information given, but we wanted to explore an explanation for the possibility of endogeneity.

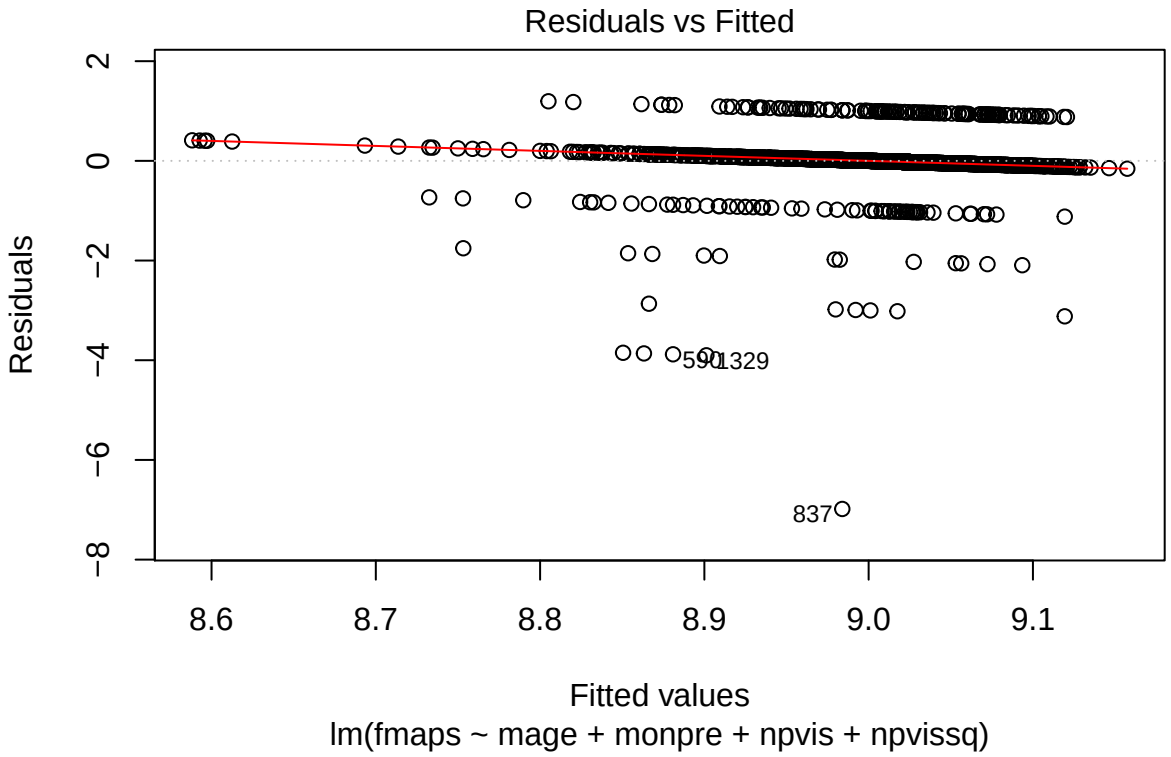
Regardless, it is convenient that we can rely on assymptotics because of the large sample size of our data. Though a couple of the other assumptions of our 6 CLM statements are a little wishy washy, this statement is an important one because endogeneity (violation of zero conditional mean) would imply that our OLS coefficients are biased and inconsistent.

5. CLM 5 - Homoskedasticity

```
plot(fit1, which = 3)
```




```
plot(fit1, which = 1)
```



To assess the if our data is in keeping with homoskedasticity, we look at a scale-location plot. The unlinear shape of in data points shows us that the errors are very likely heteroskedastic.

We can also assess homoskedasticity from our Residual vs. Fitted plot, looking at band thickness across the x's and we can very clearly see that the variance in the error is in fact not constant. So we are working with a heteroskedastic model.

To account for the heterskedasticity, we will use be using the White standard errors which are robust to heterskedasticity.

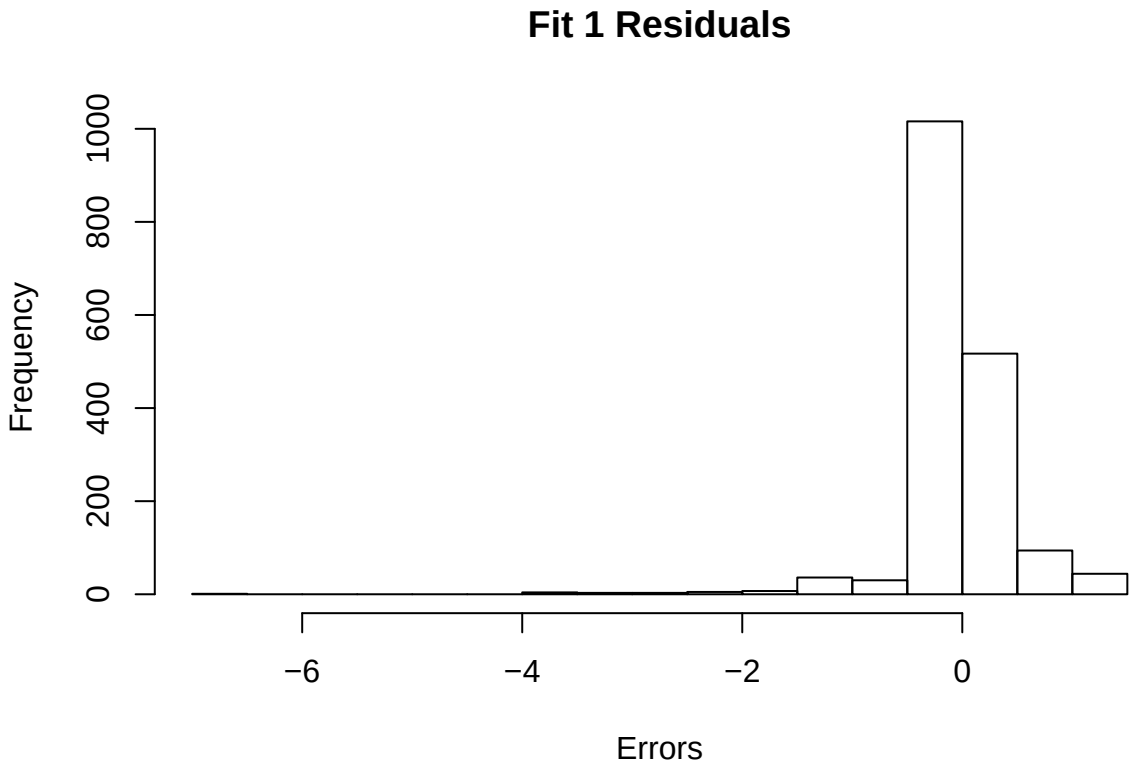
```
coeftest(fit1, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  8.5547324  0.1286889 66.4761 < 2.2e-16 ***
## mage         0.0015219  0.0019549  0.7785 0.4363796
## monpre       0.0142141  0.0105740  1.3442 0.1790441
## npvis        0.0461193  0.0126344  3.6503 0.0002696 ***
## npvissq     -0.0010910  0.0003911 -2.7896 0.0053342 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

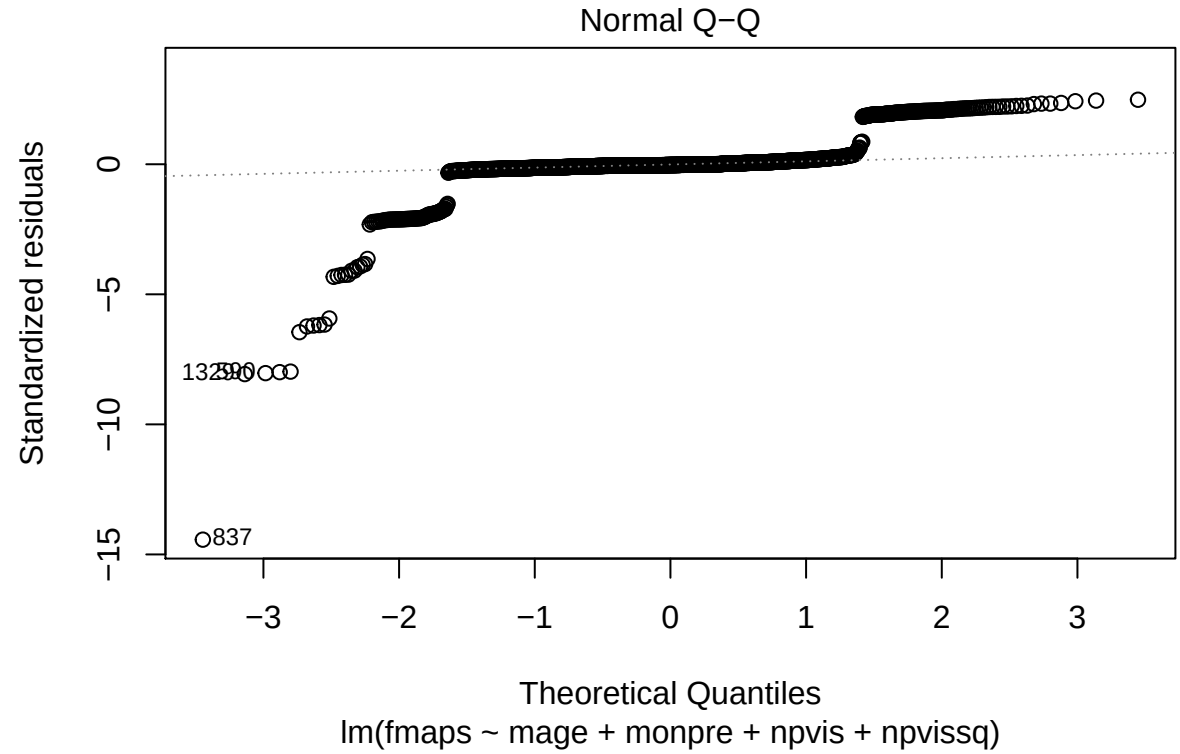
Looking at the Residual vs. Fitted and Scale Location plot for our fit2 (our better model), we see barely any change (not shown because change in graphs from fit1 to fit2 are minimal), indicating heterdoskedasticity for even our best model, so we shall use White standard errors for that model as well.

6. CLM 6 - Normal Distribution of Errors

```
hist(fit1$residuals, main = "Fit 1 Residuals", xlab = "Errors")
```




```
#Q-Q Plot
plot(fit1, which = 2)
```



From just our histogram of the residuals of our model, we can tell that our data is somewhat normally distributed, but there is a slight deviation from normality. There is a slight skew to to the left.

We then look at our Q-Q plot, we can see the plot reflect the ordinal nature of our outcome variable, so we will take that into account. Otherwise, the plot is somewhat diagnol in nature. In any case, we can rely on asympototics of the our large sample set of data to reduce the effect of any error from our model.

Regression Table

5. A well-formatted regression table summarizing your model results. Make sure that standard errors presented in this table are valid. Also be sure to comment on both statistical and practical significance.

We take the vectors of robust standard errors

```
se.fit1 = sqrt(diag(vcovHC(fit1)))
se.fit2 = sqrt(diag(vcovHC(fit2)))
se.fit3 = sqrt(diag(vcovHC(fit3)))
```

We use the robust standard errors in the summary of the model results

```
stargazer(fit1, fit2, fit3,
          title = "Linear models to predict Infant health score",
          type = "text", omit.stat = "f",
          se = list(se.fit1, se.fit2, se.fit3))
```

```
##
## Linear models to predict Infant health score
## ================================================================================
##                                        Dependent variable:
##                          ---------------------------------------------------
##                                                fmaps
##                                (1)                (2)                (3)
## --------------------------------------------------------------------------------
## mage                          0.002             0.017**            0.017**
##                              (0.002)            (0.008)            (0.008)
##
## monpre                        0.014              0.017              0.015
##                              (0.011)            (0.011)            (0.010)
##
## npvis                       0.046***           0.086***           0.082***
##                              (0.013)            (0.024)            (0.024)
##
## npvissq                     -0.001***          -0.001***          -0.001***
##                             (0.0004)           (0.0004)           (0.0004)
##
## drink                                                              0.028
##                                                                   (0.038)
##
## mage:npvis                                      -0.001*            -0.001*
##                                                 (0.001)            (0.001)
##
## Constant                    8.555***           8.096***           8.127***
##                              (0.129)            (0.276)            (0.277)
##
## --------------------------------------------------------------------------------
## Observations                  1,760              1,760              1,647
## R2                            0.018              0.020              0.021
## Adjusted R2                   0.016              0.018              0.017
## Residual Std. Error 0.484 (df = 1755) 0.484 (df = 1754) 0.479 (df = 1640)
## ================================================================================
## Note:                                          *p<0.1; **p<0.05; ***p<0.01
```

Testing models for goodness of fit

```
cat("The AIC values for Model 1 - ", AIC(fit1), ",
    Model 2 - ", AIC(fit2), ", Model 3 - ", AIC(fit3))
```

```
## The AIC values for Model 1 -  2448.454 ,
##     Model 2 -  2446.384 , Model 3 -  2260.382
```

AIC values indicate that fit3 is a more reliable model in terms of variables than fit1 and fit2.

Statistical and practical Significance

Regression table we see that both npvis and npvissq are highly statistically significant in all the three models. Mother's age is significant at the 0.05 significance level in models 2 and 3. Practically this means that every prenatal visit(npvis) contributes to an 8.6% increase in the five minute apgar score and npvissq contributes to a 0.1% decrease in the score. The mother's age contributes to a 1.7% increase in the score. The interaction between mother's age and prenatal visit is marginally significant at 0.1 significance level. This contributes to a 0.1% decrease in the score

Causality

6. A discussion of whether your results can be interpretted causally. In particular, include a discussion of what variables are not included in your analysis and the likely direction of omitted variable bias. Also include a discussion of which included variables may bias your results by absorbing some of the causal effect of prenatal care.

Interpretation of causality

From the model we can infer a causal relationship between prenatal visits, mother's age and the five minute apgar score but we have not accounted for unobserved variables like family income, medical history which also influence the score. To establish the causality we need to take into account factors such as income, mother's medical history. If we take into account all the relevant variables we may be able to establish causality but the coefficient will be smaller(since there are other factors contributing to the score as well).

Omitted variable bias

cigs(no of cigarettes smoked) - This variable is negatively correlated with both the outcome variable fmaps(-0.01) and independent variable npvis(-0.04) from the correlation matrix.

```
fit4=lm(fmaps ~mage+monpre+npvis+npvissq+mage*npvis+cigs, data=data)

summary(fit2)
```

```
##
## Call:
## lm(formula = fmaps ~ mage + monpre + npvis + npvissq + mage *
##     npvis, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9921 -0.0437 -0.0148  0.0328  1.2844
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.0960919  0.2526650  32.043  < 2e-16 ***
## mage         0.0169247  0.0080317   2.107 0.035237 *
## monpre       0.0169426  0.0105524   1.606 0.108549
## npvis        0.0861126  0.0220452   3.906 9.73e-05 ***
## npvissq     -0.0011302  0.0003026  -3.735 0.000194 ***
## mage:npvis  -0.0013361  0.0006631  -2.015 0.044063 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4837 on 1754 degrees of freedom
##   (72 observations deleted due to missingness)
## Multiple R-squared:  0.02037,    Adjusted R-squared:  0.01757
## F-statistic: 7.293 on 5 and 1754 DF,  p-value: 8.983e-07
```

```
summary(fit4)
```

```
##
## Call:
## lm(formula = fmaps ~ mage + monpre + npvis + npvissq + mage *
```

```
##     npvis + cigs, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9994 -0.0519 -0.0237  0.0311  1.2774
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.1401273  0.2589292  31.438  < 2e-16 ***
## mage         0.0160635  0.0081437   1.972 0.048721 *
## monpre       0.0144560  0.0107891   1.340 0.180471
## npvis        0.0816507  0.0224969   3.629 0.000293 ***
## npvissq     -0.0010791  0.0003068  -3.517 0.000448 ***
## cigs         0.0002514  0.0028253   0.089 0.929093
## mage:npvis  -0.0012257  0.0006695  -1.831 0.067312 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4805 on 1645 degrees of freedom
##   (180 observations deleted due to missingness)
## Multiple R-squared:  0.02061,    Adjusted R-squared:  0.01704
## F-statistic: 5.771 on 6 and 1645 DF,  p-value: 5.871e-06
```

In the above comparison of the two models fit2(without cigs) and fit4(with cigs) we see that the coefficient of npvis is positively biased in fit2 by omitting the cigs variable(both the correlations are negative leading to a positive/upward bias)

Included variable bias

Our model includes both npvis and npvissq. Our hypothesis is that the relationship between fmaps and npvis is linear. Inclusion of npvissq introduces a bias in the coefficients

```
fit6 = lm(fmaps ~mage+npvis+mage*npvis, data=data)

summary(fit2)
```

```
##
## Call:
## lm(formula = fmaps ~ mage + monpre + npvis + npvissq + mage *
##     npvis, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9921 -0.0437 -0.0148  0.0328  1.2844
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.0960919  0.2526650  32.043  < 2e-16 ***
## mage         0.0169247  0.0080317   2.107 0.035237 *
## monpre       0.0169426  0.0105524   1.606 0.108549
## npvis        0.0861126  0.0220452   3.906 9.73e-05 ***
## npvissq     -0.0011302  0.0003026  -3.735 0.000194 ***
## mage:npvis  -0.0013361  0.0006631  -2.015 0.044063 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4837 on 1754 degrees of freedom
##   (72 observations deleted due to missingness)
## Multiple R-squared:  0.02037,    Adjusted R-squared:  0.01757
## F-statistic: 7.293 on 5 and 1754 DF,  p-value: 8.983e-07
```

```
summary(fit6)
```

```
##
## Call:
## lm(formula = fmaps ~ mage + npvis + mage * npvis, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9943 -0.0232 -0.0088  0.0248  1.2096
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.4185529  0.2308406  36.469   <2e-16 ***
## mage         0.0149429  0.0079446   1.881   0.0602 .
## npvis        0.0462683  0.0193111   2.396   0.0167 *
## mage:npvis  -0.0011410  0.0006595  -1.730   0.0838 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4853 on 1757 degrees of freedom
##   (71 observations deleted due to missingness)
## Multiple R-squared:  0.01245,    Adjusted R-squared:  0.01076
## F-statistic: 7.383 on 3 and 1757 DF,  p-value: 6.475e-05
```

From the comparison of the models above we see the the coefficient of npvis shows a positive/upward bias in fit2 due to the inclusion of the npvissq variable

Conclusion and takeaways

Our exploratory analysis and model building show that fit2 is the best model in representing how prenatal care affects new born health conditions. With the limited background information we have about the data, we have determined that fmaps, or 5 minute APGAR scores, is the best singular variable representation of infant health.

We note here that our fit2 has the best adjusted R^2 value. However, our fit3 model has the best (lowest) AIC score. Though a lower AIC score is preferable, it does not necessarily indicate a better model. Thus we go with fit2, our better R^2 model.

Though we cannot deduce a causual relationship, we can infer the explanatory variables of our choice, namely: Mothers age, starting month of prenatal care, number of visits, and our created transformations: number of visits squared and mother's age * number of visits, do have an affect on the baby's health.

Looking at our general model, we can see that number of visits play an important role in the outcome of the baby's health. In terms of practical significance, a mother can look at this model and infer that every additional visit could potentially increase her baby's APGAR score by 8.6%. In addition, many mothers are often worried about having children later in life. Looking at our model alone, ceteris peribus, all other factors constant, it seems that age is not as practically significant as number of visits. Getting pregnant a year or two later, according to the model, might not cause dramatic differences in health, each year resulting in only a 0.1% decrease in the 5 min APGAR score.

Let's also take a second to note an interestingly funny take away from our model as well. One would hypothesize that more alcoholic beverages, termed as the variable "drinks" in our model would result in a negative effect on a baby's health. However, our model shows a slightly positive correlation. So more alcohol means a healthier baby? Ah, not so fast. We did explain that the data for the drinks is heavily skewed to the left (lower amounts of drinks in general). So that would not be a wise point to take away from the model. Looks like the mothers' whose drinking data were collected were being responsible after all.

In our analysis, we did exclude certain variables from being part of the model as explanatory variables because they were deemed unfit or problematic as shown in our report., example: drinks. However, if we could expand our analysis, it would be nice to also look at how some excluded variables such as race and paternal characteristics (father's characterisitics) play a role, as they are more related to genetic influence on the baby's health outcome.