STAT 331 Final Project

Francis Z. Han; Catherine Wang

03/12/2018

1 Summary

This project is aimed at exploring the relationship between healthy male single-fetus birth weight and some explanatory variables based on a given dataset containing information about 1236 healthy male single-fetus births of women enrolled in the Kaiser Foundation Health Plan in the San Francisco/East Bay area between 1960-1967. The entire statistical analysis was performed in the following steps: data cleaning, imputation, pre-fitting data diagnostics, automated model selection, model diagnostics of two candidate models and discussion of the results of the chosen model. Through analysis of our final model, we concluded that the significant explanatory variables for healthy male single-fetus birth weight are: the length of the gestation period, the total number of previous pregnancies, the father's ethnicity, the mother's age, the mother's height and weight, whether the mother smoked during pregnancy and the number of cigarettes smoked per day by the mother when she was smoking.

2 Model Selection

2.1 Deal with NA's in fht, fwt and income:

To begin with, we had a look at the summary of the dataset(see A.1). It is obvious that father's height (fht), father's weight (fwt) and the family yearly income (income) contain "too many" NA's, meaning that if we include any of these three factors into our models, then it is probable that our models will be highly biased because there are too many unavailable data. Since we do not know why they are unavailable, any model including any of these three factors will not be representative of the *entire population*. Therefore, based on the given dataset, it is reasonable to remove these three variables (see A.2) and we acknowledge that we may be missing out on important covariates by doing this. After removing these three variables, we counted how many observations have NA's (see A.3). It turned out that there were only 119 observations containing NA's, which account for less than 10% of the entire dataset.

2.2 Creating the New Variable smoke_number:

We noticed that the three variables indicating 1) whether the mother smoked during pregnancy (*smoke*); 2) time since the mother quit smoking before pregnancy (*time*); 3) number of cigarettes smoked per day by the mother when she was smoking (*number*) are highly correlated with each other and are divided into too many categories. Therefore, we decided to create a new variable that could store all information from these three variables so that after we fit the models, we will get more meaningful coefficients.

First, we regrouped the variable *number* into four categories: never smoked, up to half a pack, half to full pack, and more than one pack (see A.4). Then, we regrouped the variable *smoke* in terms of the variable *time* so that the variable *smoke* can store information from both the original *smoke* and *time* (see A.5). After this, we realized that in total, there are 21 NA's in smoke and number. Thus, we removed the observations that have NA's in smoke and number (see A.6). At the end, we created a new variable named *smoke_number* which indicates when the mother quit smoking and how many cigarettes the mother smoked when she was smoking (see A.7), where we created new categories: Never smoked, Smoked more than half and pack but quit before pregnancy, Smoke more than half a pack but quit during pregnancy, Smoked up to half a pack but quit before pregnancy, Still smokes half to full pack, Still smokes more than one pack, and Still smokes up to half a pack. And now that we have a new variable, we could throw out the original *smoke*, *time* and *number* (see A.8).

2.3 Create the New Variable: BMI

We realized that Body Mass Index (BMI) combines information mother's height(mht) and mother's weight(mwt), and is a better representation of the mother's health status than just height and weight separately (see A.9). After this, we removed the original variables mht and mwt (see A.10).

2.4 Imputing the Variables mother's ethnicity (meth) and father's ethnicity(feth)

First we realized that there are subcategories in each variable that account for a very small portion of the entire observations, such as the subcategories *mixed* and *other*. Therefore, we decided to combine them to a new variable *other* for each of meth and feth (see A.11). After this, we realized that there are NA's in both feth and meth; besides, we checked that the majority of parents have the same ethnicity (see A.12). Plus, there is only a small number of observations that have NA's in meth and feth. Hence, for the missing NA's in meth and feth, it is reasonable to replace them with the corresponding ethnicity of the other parent (see A.14).

2.5 Factor all Categorical Variables

We converted all categorical variables into factors. (see A.13).

2.6 Deal with observations that contain NA's

At this point, we noted that there are still NA's in some observations. As mentioned in Section 1.1, totally there were only 119 observations containig NA's. Since they account for less than 10% of the dataset, it is acceptable to remove them from the dataset without creating much biase (see A.15).

2.7 Pair Plots

Now that we have cleaned the dataset, we did pre-fitting data diagnostics. We constructed the pair plot of birth weight(wt) with other continuous numerical variables (see A.16). We noted that there might be a quadratic relation between birth weight (wt) and the length of the gestation period (gestation), which will be checked during future model selection.

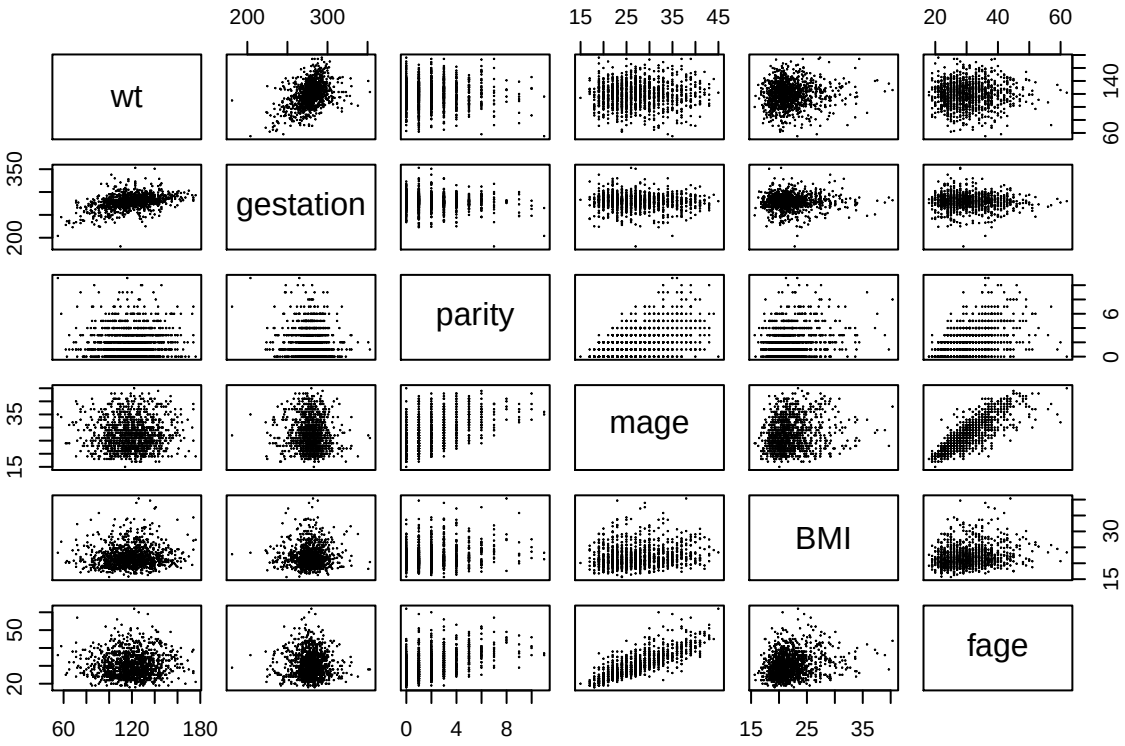

Figure 1: Pair Plot

2.8 Automated Model Selection

We constructed our minimal and maximal models for automated model selection (see A.17), where in the maximal model, we removed some interaction effect since they do not give meaningful coefficients but we added main effects. Then, we ran the forward, backward and stepwise selection (see A.18). By comparing the models generated by the above selections (see A.19), it turned out that backward and stepwise selection produced the same model; thus, we decided to keep the models generated by *forward* selection and *stepwise* selection for closer inspection.

3 Model Diagnostics

3.1 Different Residual Plots, Influences and Leverages for M1(foward selection) and M2(stepwise selection)

We plotted the ordinary residuals, standardized residuals, PRESS residuals and DIFFITS residuals against the predicted values and against the leverages for both M1 and M2. Then, we plotted the QQ-plot for both M1 and M2.

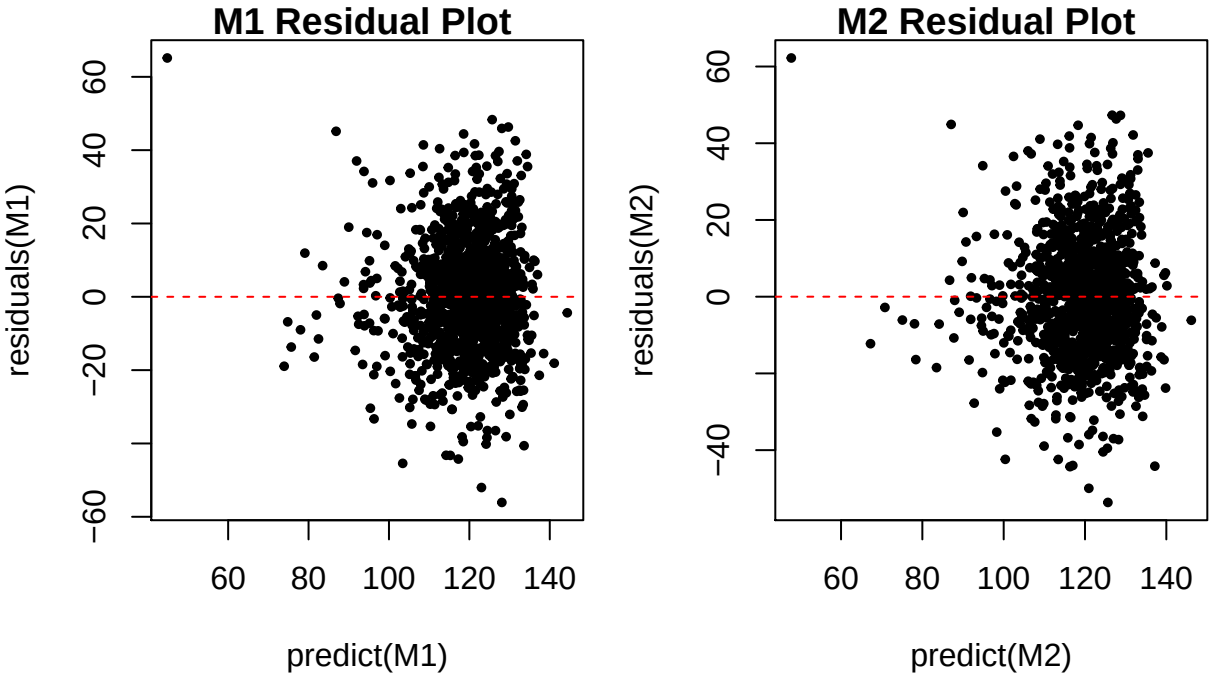

Figure 2: M1 and M2 Ordinary Residuals vs. Predicted Values

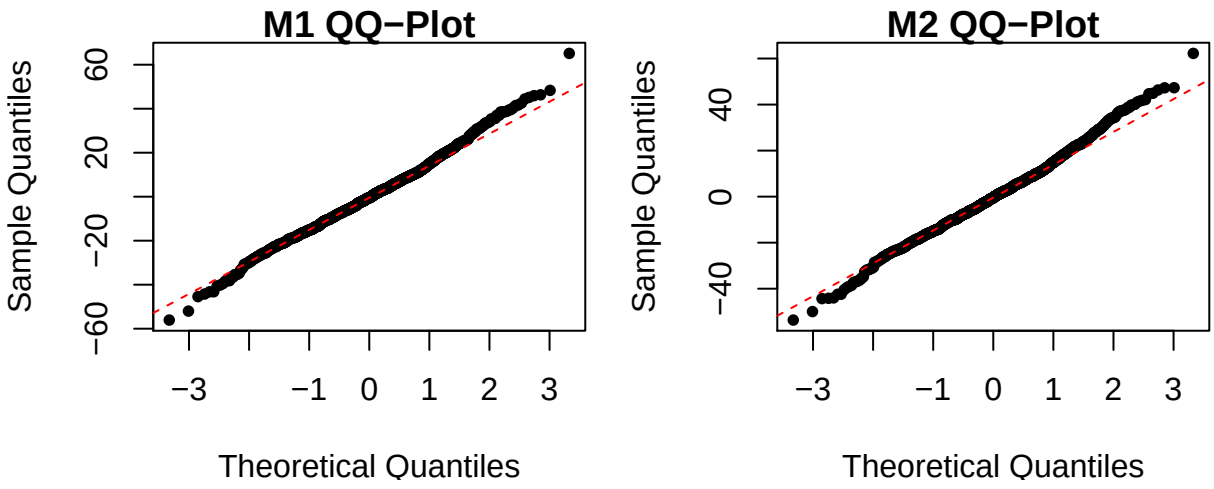

Figure 3: M1 and M2 QQ-plots

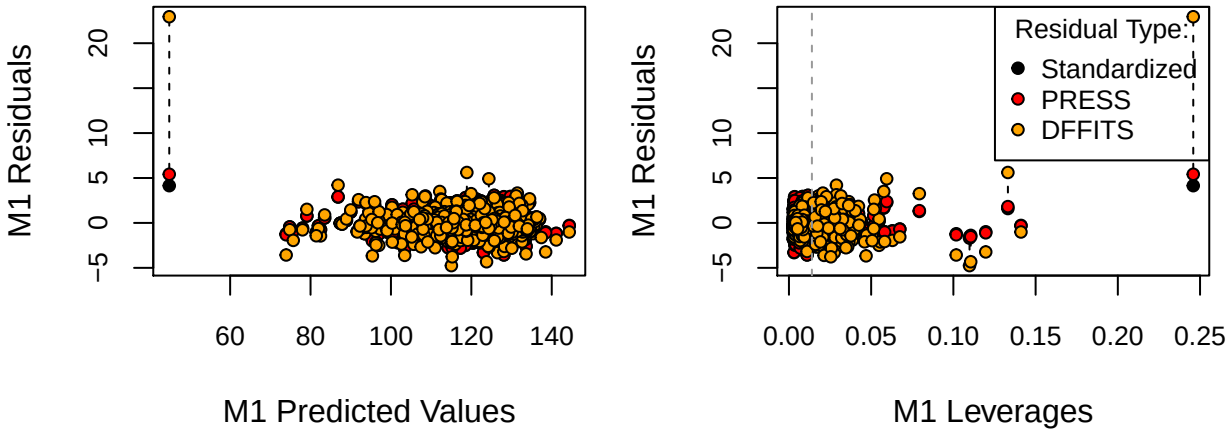

Figure 4: Predicted Value vs Residuals and Leverages for M1 and M2

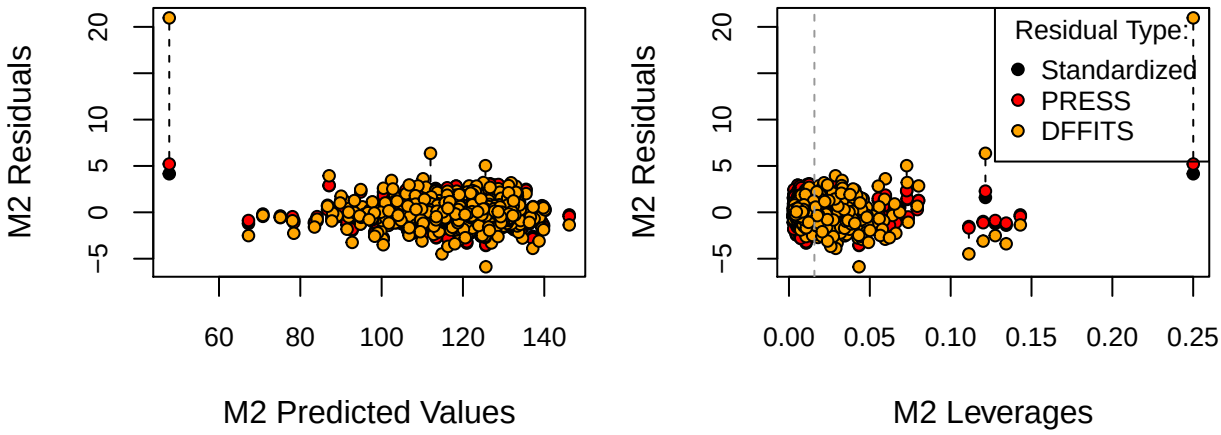

Figure 5: Predicted Value vs Residuals and Leverages for M1 and M2

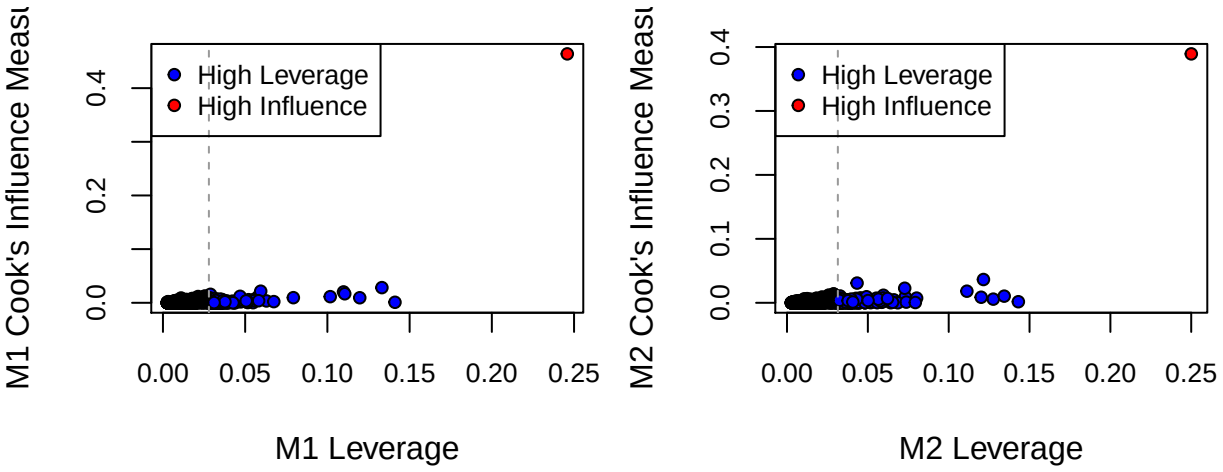

Figure 6: Cook's Distance vs Leverage

3.2 Cross-validation

We ran the cross-validation for both M1(Mfwd) and M2(Mstep) (see A.24), We produced boxplots for root mean square prediction error(rPMSE), and we calculated the Akaike Information Criterion(AIC) and produced the boxplots for PRESS statistics (see A.26)

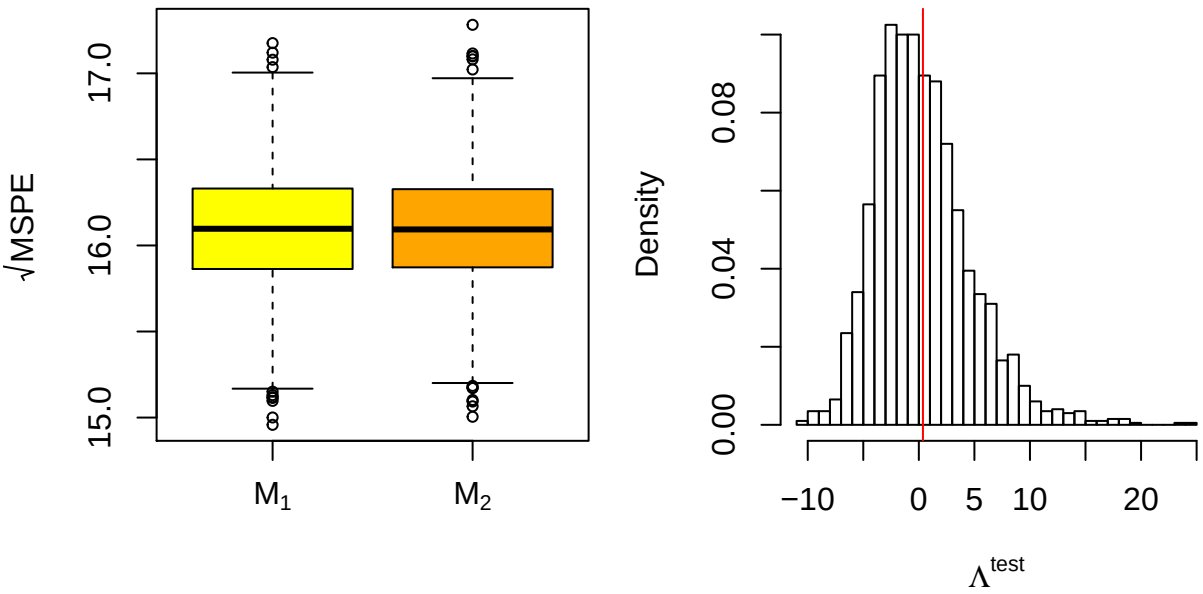

Figure 7: Cross-validation model comparison results. Left: Root mean square prediction error (rMSPE). Right: Out-of-sample likelihood ratio statistic.

Based on the graphs, it could be seen that, for the two candidate models, different residual plots, influence and leverages (Figures 2-6) did not show a big difference; however, it was the boxplots for root mean square prediction error(rPMSE) and the Akaike Information of the two models that showed us M2(stepwise model) is the better one (Figure 7 and 8). Hence, we decided to choose the stepwise model, Mstep, as our final model (see A.28)

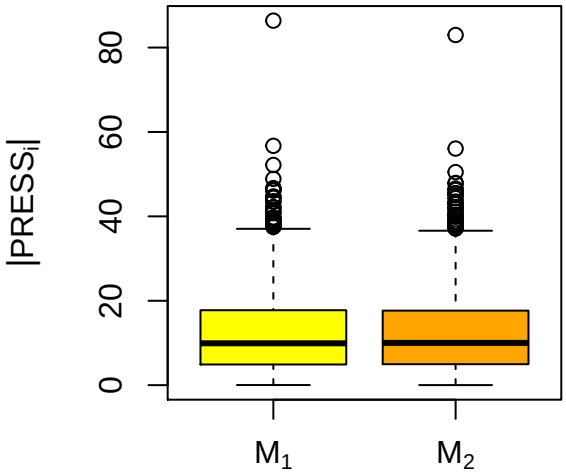

Figure 8: Boxplot of squared PRESS statistics for models M1 and M2.

```
##    parameter estimates standard errors    t-value     Pr(>|t|)
## 1        -3.267585e+02    9.325777e+01 -3.5038209 4.766252e-04
## 2         2.605378e+00    5.246369e-01  4.9660585 7.888145e-07
## 3         1.053998e+00    3.078874e-01  3.4233237 6.408301e-04
## 4        -4.301856e+00    1.474451e+00 -2.9175988 3.597365e-03
## 5         7.309917e+00    2.497848e+00  2.9264857 3.496932e-03
## 6        -3.748702e+00    2.873201e+00 -1.3047127 1.922571e-01
## 7        -6.489434e+00    3.006977e+00 -2.1581254 3.112878e-02
## 8         2.399658e+00    2.137877e+00  1.1224491 2.619106e-01
## 9        -1.266005e+01    1.899566e+00 -6.6647072 4.140941e-11
## 10       -9.227012e+00    1.365056e+00 -6.7594387 2.217798e-11
## 11       -4.517099e+00    1.239694e+00 -3.6437227 2.809560e-04
## 12        1.380406e+00    2.816424e+00  0.4901272 6.241393e-01
## 13       -7.505960e+00    1.287695e+00 -5.8289902 7.273708e-09
## 14       -9.340931e+00    2.799532e+00 -3.3366040 8.757549e-04
## 15       -1.098299e+00    3.099624e+00 -0.3543331 7.231556e-01
## 16       -3.631931e-03    8.296689e-04 -4.3775667 1.311593e-05
## 17        1.501133e-02    5.260930e-03  2.8533613 4.405072e-03
## 18       -2.468958e-02    8.925108e-03 -2.7663060 5.762201e-03
```

```
##  [1] "(Intercept)"
##  [2] "gestation"
##  [3] "parity"
##  [4] "mage"
##  [5] "BMI"
##  [6] "smoke_numberSmoked more than half a pack but quit before pregnancy"
##  [7] "smoke_numberSmoked more than half a pack but quit during pregnancy"
##  [8] "smoke_numberSmoked up to half a pack but quit before pregnancy"
##  [9] "smoke_numberStill smokes half to full pack"
```

```
## [10] "smoke_numberStill smokes more than one pack"
## [11] "smoke_numberStill smokes up to half a pack"
## [12] "fethMexican"
## [13] "fethAfrican-American"
## [14] "fethAsian"
## [15] "fethOther"
## [16] "I(gestation^2)"
## [17] "gestation:mage"
## [18] "gestation:BMI"
```

4 Discussion

4.1 Results

4.1.1 Significant Factors:

The most significant factors influencing birth weight of healthy single fetus males in this CHDS sample are: the length of the gestation period, the total number of previous pregnancies, the father's ethnicity, the mother's age, the mother's height and weight, whether the mother smoked during pregnancy and the number of cigarettes smoked per day by the mother when she was smoking.

4.1.2 Coefficients with High p-values:

There are some coefficients with *high p-values* retained in our model, such as fethMexican, fethOther and Smoked up to half a pack but quit before pregnancy.

For the coefficients related to ethnicities (fethMexican and fethOther), note that the majority observations in the given dataset have ethnicity *Caucasian*, thus, other ethnicities only account for a small portion of the entire dataset. So, even if in reality these coefficients are significant to the birth weight of a single-fetus male, they might not be revealed by this given dataset. Hence, we decided to retain them in our model.

For the coefficient about smoking (Smoked up to half a pack but quit *before pregnancy*), note that all the other coefficients about smoking are shown to be significant in our model, and they are all about *smoking during pregnancy*. We decided to retain the coefficient Smoked up to half a pack but quit before pregnancy in our model so that it shows a *contrast* with other coefficients about smoking, and it would be more obvious to see that *smoking during pregnancy* has a really significant influence on the birth weight of a single-fetus male.

4.2 Recommendation

Based on our result, it is shown that, compared to the birth weight of a single-fetus male whose mother never smoked, the expected birth weight of a single-fetus male whose mother

smoked but quit during pregnancy or smoked throughout pregnancy is less. Thus, we recommend that mothers should refrain from smoking during pregnancy and maintain a normal or higher than normal BMI (i.e. BMI above 18.5) to reduce the likelihood of low birth weight if they are expecting to give birth to a healthy single-fetus male.

4.3 Deficiency

At the beginning, we threw out three columns containing too many NA's; next, we threw out close to 10% of the entire observations which contain NA's in any columns. After this, we imputed missing mother's ethnicity (meth) and father's ethnicity (feth) by matching misssing ethnicity in one parent with the one of the other parent. Every step above was well justified, but there is still possibility that our model could be biased.

Another deficiency worth mentioning is that our model is based on this given dataset which dates back to 1960s, and "an algorithm is only as good as its inputs". Even though our model reveals all significant factors, when given another dataset obtained in the 21st century, we might be able to obtain a different model with more or less significant factors.

A R code

A.1 Summary of the dataset

```
getwd() # make sure that the file is in the working directory
```

```
## [1] "/Users/FrancisHan/Documents/STAT 331/Project"
```

```
# load the data as a data.frame
birth <- read.csv(file = "chds_births.csv")

# summary of data
summary(birth)
```

```
##        wt          gestation         parity            meth
##  Min.   : 55.0   Min.   :148.0   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:108.8   1st Qu.:272.0   1st Qu.: 0.000   1st Qu.: 0.000
##  Median :120.0   Median :280.0   Median : 1.000   Median : 3.000
##  Mean   :119.6   Mean   :279.3   Mean   : 1.932   Mean   : 3.129
##  3rd Qu.:131.0   3rd Qu.:288.0   3rd Qu.: 3.000   3rd Qu.: 7.000
##  Max.   :176.0   Max.   :353.0   Max.   :13.000   Max.   :10.000
##                  NA's   :13                       NA's   :1
##       mage            med             mht             mwt
##  Min.   :15.00   Min.   :0.000   Min.   :53.00   Min.   : 87.0
##  1st Qu.:23.00   1st Qu.:2.000   1st Qu.:62.00   1st Qu.:114.8
##  Median :26.00   Median :2.000   Median :64.00   Median :125.0
##  Mean   :27.26   Mean   :2.917   Mean   :64.05   Mean   :128.6
##  3rd Qu.:31.00   3rd Qu.:4.000   3rd Qu.:66.00   3rd Qu.:139.0
##  Max.   :45.00   Max.   :7.000   Max.   :72.00   Max.   :250.0
##  NA's   :2       NA's   :1       NA's   :22      NA's   :36
##       feth            fage            fed             fht
##  Min.   : 0.000   Min.   :18.00   Min.   :0.000   Min.   :60.0
##  1st Qu.: 0.000   1st Qu.:25.00   1st Qu.:2.000   1st Qu.:68.0
##  Median : 3.000   Median :29.00   Median :4.000   Median :71.0
##  Mean   : 3.154   Mean   :30.35   Mean   :3.127   Mean   :70.2
##  3rd Qu.: 7.000   3rd Qu.:34.00   3rd Qu.:5.000   3rd Qu.:72.0
##  Max.   :10.000   Max.   :62.00   Max.   :7.000   Max.   :78.0
##  NA's   :31       NA's   :7       NA's   :13      NA's   :492
##       fwt            marital         income          smoke
##  Min.   :110.0   Min.   :0.000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:155.0   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:0.0000
##  Median :170.0   Median :1.000   Median :3.000   Median :1.0000
```

```
##  Mean   :171.2   Mean   :1.038   Mean   :3.701   Mean   :0.8018
##  3rd Qu.:185.0   3rd Qu.:1.000   3rd Qu.:5.000   3rd Qu.:1.0000
##  Max.   :260.0   Max.   :5.000   Max.   :9.000   Max.   :3.0000
##  NA's   :499                     NA's   :124     NA's   :10
##       time             number
##  Min.   :0.0000   Min.   :0.00
##  1st Qu.:0.0000   1st Qu.:0.00
##  Median :1.0000   Median :1.00
##  Mean   :0.9625   Mean   :1.76
##  3rd Qu.:1.0000   3rd Qu.:3.00
##  Max.   :9.0000   Max.   :8.00
##  NA's   :10       NA's   :21
```

A.2

```
# throw out fht, fwt and income
birth <- birth[,-c(12, 13, 15)]
```

A.3 Check how many observations contain NA's

```
n <- length(birth$wt) # number of rows
count <- 0 # initialize count of rows with at least one NA
for (ii in 1:n) {
  if(sum(is.na(birth[ii,])) >= 1) {
    count = count + 1
  }
}
total_NA_rows <- count
total_NA_rows
```

```
## [1] 119
```

A.4 Regroup the variable number

```
# create new descriptive groups for number of cigarettes smoked per day

num2 <- birth$number
# new categories for variable number
```

```
numnames <- c("never smoked", "up to half a pack", "half to full pack", "more than one p

for(ii in 0:9) {
  if(ii==0) {
    num2[num2 == ii] <- numnames[1]
  }
  if(0 < ii & ii <= 2) {
    num2[num2 == ii] <- numnames[2]
  }
  if(2 < ii & ii <= 4) {
    num2[num2 == ii] <- numnames[3]
  } else {
    num2[num2 == ii] <- numnames[4]
  }
}

birth$number <- num2 # replace in dataset
```

A.5 Regroup the variable smoke in terms of time

```
# combine smoke and time into a single variable smoke (relabelling smoke):
# create categories for new variables
newnames <- c("Never Smoked", "Still Smokes",
              "Smoked but quit during pregnancy",
              "Smoked but quit before pregnancy")

# extract the corresponding columns
smoke <- birth$smoke

n <- length(smoke) # number of observations
newsmoke <- rep(NA, n)

for (ii in 1:n) {
  if (is.na(smoke[ii])) {
    newsmoke[ii] <- NA
  }
  else {
    for (jj in 0:3) {
      if (smoke[ii] == jj) {
        newsmoke[ii] <- newnames[jj+1]
      }
    }
```

```
  }
}
birth$smoke <- newsmoke
```

A.6 Remove observations containing NA's in smoke and number

```
# locate the observations that have NA's in smoke
ind_smoke <- NA
for (ii in 1:length(birth$smoke)) {
  if (is.na(birth$smoke[ii])) {
    ind_smoke <- c(ind_smoke, ii)
  }
}
ind_smoke <- ind_smoke[2:length(ind_smoke)]

# locate the observations that have NA's in number
ind_number <- NA
for (ii in 1:length(birth$number)) {
  if (is.na(birth$number[ii])) {
    ind_number <- c(ind_number, ii)
  }
}
ind_number <- ind_number[2:length(ind_number)]

# Throw out the observations that have NA's in smoke and number
# note that observations that have NA's in smoke also have NA's in number
remove_smoke <- ind_number
birth <- birth[-remove_smoke,]
```

A.7 Create the new variable smoke_number

```
# create a new variable to combine smoke and number
# note that nobody is in the category "Smoked up to half a pack but quit during pregna

smoke_number <- birth$smoke
for (ii in 1:(length(birth$smoke))) {
  if (smoke_number[ii] != "Never Smoked") {
    if (birth$number[ii] == "up to half a pack") {
      if (smoke_number[ii] == "Smoked but quit before pregnancy") {
        smoke_number[ii] <- "Smoked up to half a pack but quit before pregnancy"
```

```
      }
      else {smoke_number[ii] <- "Still smokes up to half a pack"}
    }
    else if (birth$number[ii] == "half to full pack") {
      if (smoke_number[ii] == "Smoked but quit before pregnancy") {
        smoke_number[ii] <- "Smoked more than half a pack but quit before pregnancy"
      }
      else if (smoke_number[ii] == "Smoked but quit during pregnancy") {
        smoke_number[ii] <- "Smoked more than half a pack but quit during pregnancy"
      }
      else {smoke_number[ii] <- "Still smokes half to full pack"}
    }
    else {
      if (smoke_number[ii] == "Smoked but quit before pregnancy") {
        smoke_number[ii] <- "Smoked more than half a pack but quit before pregnancy"
      }
      else if (smoke_number[ii] == "Smoked but quit during pregnancy") {
        smoke_number[ii] <- "Smoked more than half a pack but quit during pregnancy"
      }
      else {smoke_number[ii] <- "Still smokes more than one pack"}
    }
  }
}

# factor the new variable smoke_number
birth$smoke_number <- factor(smoke_number)
```

A.8 Remove the original smoke, time and number

```
# Remove the original smoke, time and number
birth <- birth[,-c(13, 14, 15)]
```

A.9 BMI

```
# Combine mht and mwt to create a new variable BMI
# BMI is called Body Mass Index, which is calculated by weight in kg / height^2 in m
# Convert mht and mwt to standard metric
lbs_to_kg <- 0.453592
birth$mwt <- lbs_to_kg * birth$mwt
inch_to_meter <- 0.0254
```

```r
birth$mht <- inch_to_meter * birth$mht

# Create BMI variable
mht <- birth$mht
mwt <- birth$mwt
BMI <- mwt / mht^2
birth$BMI <- BMI
```

A.10

```r
# throw out mht and mwt
birth <- birth[,-c(7, 8)]
```

A.11

```r
# Combine other and mixed in feth and call it other
for (ii in 1:length(birth$feth)) {
  if (!is.na(birth$feth[ii])) {
    if(birth$feth[ii] == 10) {
      birth$feth[ii] <- 9
    }
  }
}

# Combine other and mixed in meth and call it other
for (ii in 1:length(birth$meth)) {
  if (!is.na(birth$meth[ii])) {
    if(birth$meth[ii] == 10) {
      birth$meth[ii] <- 9
    }
  }
}
```

A.12

```r
# Want to count how many rows have meth=feth
count.eth <- 0
for (ii in 1:length(birth$meth)) {
```

```
  if ((is.na(birth$feth[ii])==FALSE) & is.na(birth$meth[ii])==FALSE){
    if (birth$meth[ii]<=5) {
      if (birth$feth[ii]<=5){
        count.eth <- count.eth+1
      }
    }
    else {
      if (birth$meth[ii]==birth$feth[ii]){
        count.eth <- count.eth+1
      }
    }
  }
}
count.eth
```

```
## [1] 1125
```

A.13 Factor all categorical variables

A.13.1 meth

```
# convert categorical variables to factors

# convert meth to non-numeric categorical variable
meth2 <- birth$meth
length(meth2)
```

```
## [1] 1215
```

```
# "levels" of variable meth
methnames <- c("Caucasian","Mexican", "African-American","Asian", "Other")
# assign "Caucasian" level to values less than or equal to 5
# otherwise assign levels based on order in methnames vector
for(ii in 0:9) {
  if(ii<=5){
    meth2[meth2 == ii] <- "Caucasian"
  } else {
    meth2[meth2 == ii] <- methnames[ii-4]
  }
}
meth2 <- factor(meth2, levels = methnames) # order levels based on methnames
levels(meth2)
```

```
## [1] "Caucasian"        "Mexican"          "African-American" "Asian"
## [5] "Other"
```

```
birth$meth <- meth2 # replace in dataset
##########################################
```

A.13.2 med

```
# convert med to categorical variables
med2 <- birth$med
# "levels" of variable med
mednames <- c("elementary school", "middle school", "high school", "high school + trade"
med2 <- mednames[birth$med+1]
med2 <- factor(med2, levels = mednames) # order levels based on mednames
levels(med2)
```

```
## [1] "elementary school"          "middle school"
## [3] "high school"                "high school + trade"
## [5] "high school + some college" "college grad"
## [7] "trade school"               "high school unclear"
```

```
birth$med <- med2 # replace in dataset
```

A.13.3 feth

```
# convert feth to categorical variables
feth2 <- birth$feth
# "levels" of variable feth
fethnames <- c("Caucasian","Mexican", "African-American","Asian", "Other")
# assign "Caucasian" level to values less than or equal to 5
# otherwise assign levels based on order in fethnames vector
for(ii in 0:9) {
  if(ii<=5){
    feth2[feth2 == ii] <- "Caucasian"
  } else {
    feth2[feth2 == ii] <- fethnames[ii-4]
  }
}
feth2 <- factor(feth2, levels = fethnames) # order levels based on methnames
levels(feth2)
```

```
## [1] "Caucasian"        "Mexican"          "African-American" "Asian"
## [5] "Other"
```

```
birth$feth <- feth2 # replace in dataset
```

A.13.4 fed

```
# convert fed to categorical variables
fed2 <- birth$fed
# "levels" of variable fed
fednames <- c("elementary school", "middle school", "high school", "high school + trade"
fed2 <- fednames[birth$fed+1]
fed2 <- factor(fed2, levels = fednames) # order levels based on fednames
levels(fed2)
```

```
## [1] "elementary school"          "middle school"
## [3] "high school"                "high school + trade"
## [5] "high school + some college" "college grad"
## [7] "trade school"               "high school unclear"
```

```
birth$fed <- fed2 # replace in dataset
```

A.13.5 marital

```
marital <- birth$marital
maritalnames <- c("married", "legally separated", "divorced", "widowed",
                  "never married") # new values
# encode the factor
marital2 <- rep(NA, length(marital))
for (ii in 1:5) {
  marital2[marital == ii] <- maritalnames[ii]
}
marital2 <- factor(marital2, levels = maritalnames)
birth$marital <- marital2
```

A.14 Imputing meth and feth

```r
# we assume fathers and mothers have the same ethnicity background
# based on the majority of the dataset
# deal with the NA's in meth
for (ii in 1:length(birth$meth)) {
  if(is.na(birth$meth)[ii]) {
    birth$meth[ii] <- birth$feth[ii]
  }
}

# deal with the NA's in feth
for (ii in 1:length(birth$feth)) {
  if(is.na(birth$feth)[ii]) {
    birth$feth[ii] <- birth$meth[ii]
  }
}
```

A.15

```r
# locate the observations that have NA's in gestation
ind_gestation <- NA
for (ii in 1:length(birth$gestation)) {
  if (is.na(birth$gestation[ii])) {
    ind_gestation <- c(ind_gestation, ii)
  }
}
ind_gestation <- ind_gestation[2:length(ind_gestation)]

# locate the observations that have NA's in mage
ind_mage <- NA
for (ii in 1:length(birth$mage)) {
  if (is.na(birth$mage[ii])) {
    ind_mage <- c(ind_mage, ii)
  }
}
ind_mage <- ind_mage[2:length(ind_mage)]

# locate the observations that have NA's in med
ind_med <- NA
for (ii in 1:length(birth$med)) {
  if (is.na(birth$med[ii])) {
    ind_med <- c(ind_med, ii)
  }
```

```
}
ind_med <- ind_med[2:length(ind_med)]

# locate the observations that have NA's in fage
ind_fage <- NA
for (ii in 1:length(birth$fage)) {
  if (is.na(birth$fage[ii])) {
    ind_fage <- c(ind_fage, ii)
  }
}
ind_fage <- ind_fage[2:length(ind_fage)]

# locate the observations that have NA's in fed
ind_fed <- NA
for (ii in 1:length(birth$fed)) {
  if (is.na(birth$fed[ii])) {
    ind_fed <- c(ind_fed, ii)
  }
}
ind_fed <- ind_fed[2:length(ind_fed)]

# locate the observations that have NA's in marital
ind_marital <- NA
for (ii in 1:length(birth$marital)) {
  if (is.na(birth$marital[ii])) {
    ind_marital <- c(ind_marital, ii)
  }
}
ind_marital <- ind_marital[2:length(ind_marital)]


# locate the observations that have NA's in BMI
ind_BMI <- NA
for (ii in 1:length(birth$BMI)) {
  if (is.na(birth$BMI[ii])) {
    ind_BMI <- c(ind_BMI, ii)
  }
}
ind_BMI <- ind_BMI[2:length(ind_BMI)]

missing <- c(ind_gestation, ind_mage, ind_BMI, ind_marital, ind_fage, ind_fed, ind_med)
missing
```

```
##  [1]    4   89   93   98  239  636  690  723  861  944  952 1172  390  419   39
```

```
## [16]   42   85  102  110  113  152  157  183  191  202  227  304  328  343  351
## [31]  354  390  429  433  466  467  470  496  513  634  639  682  737  746  830
## [46]  856  864  994 1025 1157 1170 1195    9  255  285  529  692  788  835  923
## [61] 1060   42  207  247  434  470  521  549  600  652  887 1060 1146 1204  652
```

```
# remove missing observations

# first remove duplicated numbers in the missing list
m <- length(missing)
new_missing <- NA
for (ii in 1:m) {
  if (missing[ii] %in% new_missing == FALSE) {
    new_missing <- c(new_missing, missing[ii])
  }
}
new_missing <- new_missing[2:length(new_missing)]

# display the indexes of all observations with NA's
new_missing
```

```
##  [1]    4   89   93   98  239  636  690  723  861  944  952 1172  390  419   39
## [16]   42   85  102  110  113  152  157  183  191  202  227  304  328  343  351
## [31]  354  429  433  466  467  470  496  513  634  639  682  737  746  830  856
## [46]  864  994 1025 1157 1170 1195    9  255  285  529  692  788  835  923 1060
## [61]  207  247  434  521  549  600  652  887 1146 1204
```

```
# remove observations containing NA's
birth <- birth[-new_missing,]
```

A.16 Pair Plot

```
# Figure 1. Pair plots of wt and other continuous numerical variables
pairs(~ wt + gestation + parity + mage + BMI + fage,
      pch = 19,
      cex = 0.05,
      data = birth)
```

A.17 Max and min modes

```r
# Minimal model
M0 <- lm(wt ~ 1, data = birth)
summary(M0)
```

```
##
## Call:
## lm(formula = wt ~ 1, data = birth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -64.431 -11.431   0.569  11.569  56.569
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 119.4306     0.5437   219.7   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.4 on 1144 degrees of freedom
```

```r
# new maximal model
Mmaxnew <- lm(wt ~ (.-fed - med - feth - meth - marital - smoke_number)^2
              + smoke_number + feth + meth + fed + med + marital
              + I(gestation^2), data = birth)
summary(Mmaxnew)
```

```
##
## Call:
## lm(formula = wt ~ (. - fed - med - feth - meth - marital - smoke_number)^2 +
##     smoke_number + feth + meth + fed + med + marital + I(gestation^2),
##     data = birth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.156  -9.645  -0.550   8.983  54.341
##
## Coefficients:
##                                                                  Estimate
## (Intercept)                                                  -3.236e+02
## gestation                                                     2.726e+00
## parity                                                        1.055e+01
## mage                                                         -4.468e+00
## fage                                                         -1.549e+00
```

```
## BMI                                                                  6.619e+00
## smoke_numberSmoked more than half a pack but quit before pregnancy -3.545e+00
## smoke_numberSmoked more than half a pack but quit during pregnancy -6.201e+00
## smoke_numberSmoked up to half a pack but quit before pregnancy      2.377e+00
## smoke_numberStill smokes half to full pack                         -1.278e+01
## smoke_numberStill smokes more than one pack                        -8.921e+00
## smoke_numberStill smokes up to half a pack                         -4.125e+00
## fethMexican                                                          1.589e+00
## fethAfrican-American                                               -6.140e-01
## fethAsian                                                          -8.330e+00
## fethOther                                                            2.850e+00
## methMexican                                                          1.211e+00
## methAfrican-American                                               -6.731e+00
## methAsian                                                          -2.062e-01
## methOther                                                          -3.822e+00
## fedmiddle school                                                     1.861e+00
## fedhigh school                                                       2.444e+00
## fedhigh school + trade                                               3.959e+00
## fedhigh school + some college                                        1.182e+00
## fedcollege grad                                                      1.825e+00
## fedtrade school                                                      2.514e+00
## fedhigh school unclear                                             -1.219e+01
## medmiddle school                                                   -8.535e-01
## medhigh school                                                       1.697e+00
## medhigh school + trade                                               1.987e+00
## medhigh school + some college                                        2.090e+00
## medcollege grad                                                      9.685e-01
## medhigh school unclear                                             -1.246e+00
## maritallegally separated                                           -1.286e+00
## maritaldivorced                                                      1.498e+01
## maritalnever married                                               -6.737e+00
## I(gestation^2)                                                     -3.980e-03
## gestation:parity                                                   -2.202e-02
## gestation:mage                                                       1.093e-02
## gestation:fage                                                       7.967e-03
## gestation:BMI                                                      -2.522e-02
## parity:mage                                                        -3.202e-02
## parity:fage                                                          7.605e-03
## parity:BMI                                                         -1.118e-01
## mage:fage                                                          -5.154e-03
## mage:BMI                                                             6.586e-02
## fage:BMI                                                           -2.228e-02
##                                                                   Std. Error
## (Intercept)                                                          9.833e+01
## gestation                                                            5.459e-01
```

```
## parity                                                                 5.108e+00
## mage                                                                   2.960e+00
## fage                                                                   2.746e+00
## BMI                                                                    2.838e+00
## smoke_numberSmoked more than half a pack but quit before pregnancy     2.912e+00
## smoke_numberSmoked more than half a pack but quit during pregnancy     3.061e+00
## smoke_numberSmoked up to half a pack but quit before pregnancy         2.176e+00
## smoke_numberStill smokes half to full pack                             1.951e+00
## smoke_numberStill smokes more than one pack                            1.410e+00
## smoke_numberStill smokes up to half a pack                             1.268e+00
## fethMexican                                                            4.545e+00
## fethAfrican-American                                                   4.826e+00
## fethAsian                                                              5.645e+00
## fethOther                                                              4.092e+00
## methMexican                                                            4.486e+00
## methAfrican-American                                                   4.915e+00
## methAsian                                                              5.486e+00
## methOther                                                              3.706e+00
## fedmiddle school                                                       3.341e+00
## fedhigh school                                                         3.307e+00
## fedhigh school + trade                                                 4.207e+00
## fedhigh school + some college                                          3.426e+00
## fedcollege grad                                                        3.510e+00
## fedtrade school                                                        1.613e+01
## fedhigh school unclear                                                 7.798e+00
## medmiddle school                                                       4.372e+00
## medhigh school                                                         4.296e+00
## medhigh school + trade                                                 4.706e+00
## medhigh school + some college                                          4.387e+00
## medcollege grad                                                        4.488e+00
## medhigh school unclear                                                 7.757e+00
## maritallegally separated                                               4.323e+00
## maritaldivorced                                                        1.121e+01
## maritalnever married                                                   8.137e+00
## I(gestation^2)                                                         9.052e-04
## gestation:parity                                                       1.554e-02
## gestation:mage                                                         1.006e-02
## gestation:fage                                                         9.157e-03
## gestation:BMI                                                          9.582e-03
## parity:mage                                                            7.773e-02
## parity:fage                                                            6.571e-02
## parity:BMI                                                             8.975e-02
## mage:fage                                                              1.327e-02
## mage:BMI                                                               4.599e-02
## fage:BMI                                                               4.030e-02
```

```
##                                                                  t value
## (Intercept)                                                       -3.291
## gestation                                                          4.994
## parity                                                             2.065
## mage                                                              -1.509
## fage                                                              -0.564
## BMI                                                                2.332
## smoke_numberSmoked more than half a pack but quit before pregnancy -1.217
## smoke_numberSmoked more than half a pack but quit during pregnancy -2.026
## smoke_numberSmoked up to half a pack but quit before pregnancy     1.092
## smoke_numberStill smokes half to full pack                        -6.552
## smoke_numberStill smokes more than one pack                       -6.327
## smoke_numberStill smokes up to half a pack                        -3.254
## fethMexican                                                        0.350
## fethAfrican-American                                              -0.127
## fethAsian                                                         -1.476
## fethOther                                                          0.696
## methMexican                                                        0.270
## methAfrican-American                                              -1.369
## methAsian                                                         -0.038
## methOther                                                         -1.031
## fedmiddle school                                                   0.557
## fedhigh school                                                     0.739
## fedhigh school + trade                                             0.941
## fedhigh school + some college                                      0.345
## fedcollege grad                                                    0.520
## fedtrade school                                                    0.156
## fedhigh school unclear                                            -1.563
## medmiddle school                                                  -0.195
## medhigh school                                                     0.395
## medhigh school + trade                                             0.422
## medhigh school + some college                                      0.476
## medcollege grad                                                    0.216
## medhigh school unclear                                            -0.161
## maritallegally separated                                          -0.297
## maritaldivorced                                                    1.336
## maritalnever married                                              -0.828
## I(gestation^2)                                                    -4.397
## gestation:parity                                                  -1.417
## gestation:mage                                                     1.086
## gestation:fage                                                     0.870
## gestation:BMI                                                     -2.632
## parity:mage                                                       -0.412
## parity:fage                                                        0.116
## parity:BMI                                                        -1.245
```

```
## mage:fage                                                          -0.388
## mage:BMI                                                            1.432
## fage:BMI                                                           -0.553
##                                                                  Pr(>|t|)
## (Intercept)                                                       0.00103 **
## gestation                                                        6.87e-07 ***
## parity                                                            0.03920 *
## mage                                                              0.13148
## fage                                                              0.57295
## BMI                                                               0.01988 *
## smoke_numberSmoked more than half a pack but quit before pregnancy  0.22382
## smoke_numberSmoked more than half a pack but quit during pregnancy  0.04304 *
## smoke_numberSmoked up to half a pack but quit before pregnancy      0.27495
## smoke_numberStill smokes half to full pack                       8.70e-11 ***
## smoke_numberStill smokes more than one pack                      3.63e-10 ***
## smoke_numberStill smokes up to half a pack                        0.00117 **
## fethMexican                                                       0.72674
## fethAfrican-American                                              0.89879
## fethAsian                                                         0.14030
## fethOther                                                         0.48637
## methMexican                                                       0.78726
## methAfrican-American                                              0.17112
## methAsian                                                         0.97003
## methOther                                                         0.30272
## fedmiddle school                                                  0.57768
## fedhigh school                                                    0.46001
## fedhigh school + trade                                            0.34697
## fedhigh school + some college                                     0.73015
## fedcollege grad                                                   0.60330
## fedtrade school                                                   0.87618
## fedhigh school unclear                                            0.11843
## medmiddle school                                                  0.84526
## medhigh school                                                    0.69284
## medhigh school + trade                                            0.67297
## medhigh school + some college                                     0.63387
## medcollege grad                                                   0.82918
## medhigh school unclear                                            0.87242
## maritallegally separated                                          0.76616
## maritaldivorced                                                   0.18193
## maritalnever married                                              0.40792
## I(gestation^2)                                                   1.20e-05 ***
## gestation:parity                                                  0.15685
## gestation:mage                                                    0.27753
## gestation:fage                                                    0.38446
## gestation:BMI                                                     0.00861 **
```

```
## parity:mage                                                    0.68048
## parity:fage                                                    0.90788
## parity:BMI                                                     0.21331
## mage:fage                                                      0.69773
## mage:BMI                                                       0.15236
## fage:BMI                                                       0.58048
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.72 on 1098 degrees of freedom
## Multiple R-squared:  0.2992, Adjusted R-squared:  0.2698
## F-statistic: 10.19 on 46 and 1098 DF,  p-value: < 2.2e-16
```

```
beta.maxnew <- coef(Mmaxnew)
names(beta.maxnew)[is.na(beta.maxnew)]
```

```
## character(0)
```

```
anyNA(coef(Mmaxnew))
```

```
## [1] FALSE
```

A.18 AMS

```
# Automated Model Selection
# Forward selection
system.time({ # time the calculation
  Mfwd <- step(object = M0, # starting point model
               scope = list(lower = M0, upper = Mmaxnew), # smallest and largest model
               direction = "forward",
               trace = FALSE) # trace prints out information
})
```

```
##    user  system elapsed
##   0.127   0.009   0.136
```

```
# Backward elimination
system.time({
  Mback <- step(object = Mmaxnew, # starting point model
                scope = list(lower = M0, upper = Mmaxnew),
                direction = "backward", trace = FALSE)
})
```

```
##    user  system elapsed
##   0.322   0.028   0.350
```

```
# Stepwise selection
Mstart <- lm(wt ~ ., data = birth)
system.time({
  Mstep <- step(object = Mstart,
                scope = list(lower = M0, upper = Mmaxnew),
                direction = "both", trace = FALSE)
})
```

```
##    user  system elapsed
##   0.409   0.038   0.448
```

A.19 Mfwd, Mback and Mstep

```
Mfwd$call
```

```
## lm(formula = wt ~ gestation + smoke_number + meth + I(gestation^2) +
##     parity + BMI + gestation:BMI, data = birth)
```

```
Mback$call
```

```
## lm(formula = wt ~ gestation + parity + mage + BMI + smoke_number +
##     feth + I(gestation^2) + gestation:mage + gestation:BMI, data = birth)
```

```
Mstep$call
```

```
## lm(formula = wt ~ gestation + parity + mage + feth + smoke_number +
##     BMI + I(gestation^2) + gestation:mage + gestation:BMI, data = birth)
```

A.20 M1 and M2 Ordinary Residuals vs. Prediected Values

```
# Figure 2
# plot setup
cex <- .8
pch <- 20
par(mfrow = c(1,2), mar = c(4,4,1,1))
```

```r
# assign new name to Mfwd and Mstep
M1 <- Mfwd
M2 <- Mback

# residuals for M1 and M2
res1 <- residuals(M1)
res2 <- residuals(M2)

# extract sigma.hat for both M1 and M2
sigma.hat1 <- sigma(M1)
sigma.hat2 <- sigma(M2)

# plot ordinary residuals for M1
plot(predict(M1), residuals(M1), pch = pch, cex = cex, col = "black",
     main = "M1 Residual Plot ")
abline(h = 0, col = "red", lty = 2) # horizontal line

# plot ordinary residuals for M2
plot(predict(M2), residuals(M2), pch = pch, cex = cex, col = "black",
     main = "M2 Residual Plot ")
abline(h = 0, col = "red", lty = 2) # horizontal line
```

A.21 M1 and M2 QQ-plots

```r
# Figure 3
# plot setup
cex <- .8
pch <- 16
par(mfrow = c(1,2), mar = c(4,4,1,1))

# QQ-plot for M1
qqnorm(res1, main = "M1 QQ-Plot", cex = cex, pch = pch)
qqline(res1, col = "red", lty = 2)

# QQ-plot for M2
qqnorm(res2, main = "M2 QQ-Plot", cex = cex, pch = pch)
qqline(res2, col = "red", lty = 2)
```

A.22 M1 and M2 Predicted Value vs Residuals and Leverages

```r
# Figure 4 and Figure 5
# plot setup
cex <- .8
pch <- 16
par(mfrow = c(1,2), mar = c(4,4,1,1))

# Predicted Value vs Residuals and Leverages for M1

h1 <- hatvalues(M1)  # HAT Matrix
stan.res1 <- res1/sigma.hat1 # Standardized residuals for M1

# PRESS residuals
press1 <- res1/(1-h1)

# DFFITS residuals
dfts1 <- dffits(M1) # the R way
# standardize each of these such that they are identical at the average leverage value
p1 <- length(coef(M1))
n1 <- nobs(M1)
hbar1 <- p1/n1 # average leverage
press1 <- press1*(1-hbar1)/sigma.hat1 # at h1 = hbar1, press1 = stan.res1
dfts1 <- dfts1*(1-hbar1)/sqrt(hbar1) # at h1 = hbar1, dfts1 = stan.res1

# plot all residuals
# against predicted values
plot(predict(M1), rep(0, length(predict(M1))), type = "n", # empty plot to get the axis
     ylim = range(stan.res1, press1, dfts1), cex.axis = cex,
     xlab = "M1 Predicted Values", ylab = "M1 Residuals")
# dotted line connecting each observations residuals for better visibility
segments(x0 = predict(M1),
         y0 = pmin(stan.res1, press1, dfts1),
         y1 = pmax(stan.res1, press1, dfts1),
         lty = 2)
points(predict(M1), stan.res1, pch = 21, bg = "black", cex = cex)
points(predict(M1), press1, pch = 21, bg = "red", cex = cex)
points(predict(M1), dfts1, pch = 21, bg = "orange", cex = cex)
# against leverages
plot(h1, rep(0, length(predict(M1))), type = "n", cex.axis = cex,
     ylim = range(stan.res1, press1, dfts1),
     xlab = "M1 Leverages", ylab = "M1 Residuals")
segments(x0 = h1,
         y0 = pmin(stan.res1, press1, dfts1),
```

```
         y1 = pmax(stan.res1, press1, dfts1),
         lty = 2)
points(h1, stan.res1, pch = 21, bg = "black", cex = cex)
points(h1, press1, pch = 21, bg = "red", cex = cex)
points(h1, dfts1, pch = 21, bg = "orange", cex = cex)
abline(v = hbar1, col = "grey60", lty = 2)
legend("topright", legend = c("Standardized", "PRESS", "DFFITS"),
       pch = 21, pt.bg = c("black", "red", "orange"), title = "Residual Type:",
       cex = cex, pt.cex = cex)


# Figure 5
# Predicted Value vs Residuals and Leverages for M2
h2 <- hatvalues(M2)  # HAT Matrix for M2
stan.res2 <- res1/sigma.hat2 # Standardized residuals for M2

# PRESS residuals
press2 <- res2/(1-h2)

# DFFITS residuals
dfts2 <- dffits(M2)
# standardize each of these such that they are identical at the average leverage value
p2 <- length(coef(M2))
n2 <- nobs(M2)
hbar2 <- p2/n2 # average leverage
press2 <- press2*(1-hbar2)/sigma.hat2 # at h = hbar2, press2 = stan.res
dfts2 <- dfts2*(1-hbar2)/sqrt(hbar2) # at h = hbar2, dfts2 = stan.res

# plot all residuals
par(mfrow = c(1,2), mar = c(4,4,1,1))
# against predicted values
plot(predict(M2), rep(0, length(predict(M2))), type = "n", # empty plot to get the axis
     ylim = range(stan.res2, press2, dfts2), cex.axis = cex,
     xlab = "M2 Predicted Values", ylab = "M2 Residuals")
# dotted line connecting each observations residuals for better visibility
segments(x0 = predict(M2),
         y0 = pmin(stan.res2, press2, dfts2),
         y1 = pmax(stan.res2, press2, dfts2),
         lty = 2)
points(predict(M2), stan.res2, pch = 21, bg = "black", cex = cex)
points(predict(M2), press2, pch = 21, bg = "red", cex = cex)
points(predict(M2), dfts2, pch = 21, bg = "orange", cex = cex)
# against leverages
plot(h2, rep(0, length(predict(M2))), type = "n", cex.axis = cex,
```

```
     ylim = range(stan.res2, press2, dfts2),
     xlab = "M2 Leverages", ylab = "M2 Residuals")
segments(x0 = h2,
         y0 = pmin(stan.res2, press2, dfts2),
         y1 = pmax(stan.res2, press2, dfts2),
         lty = 2)
points(h2, stan.res2, pch = 21, bg = "black", cex = cex)
points(h2, press2, pch = 21, bg = "red", cex = cex)
points(h2, dfts2, pch = 21, bg = "orange", cex = cex)
abline(v = hbar2, col = "grey60", lty = 2)
legend("topright", legend = c("Standardized", "PRESS", "DFFITS"),
       pch = 21, pt.bg = c("black", "red", "orange"), title = "Residual Type:",
       cex = cex, pt.cex = cex)
```

A.23 M1 and M2 Cook's Distance vs Leverage

```
# Figure 6
# plot setup
cex <- .8
pch <- 20
par(mfrow = c(1,2), mar = c(4,4,1,1))

# cook's distance vs. leverage for M1
D1 <- cooks.distance(M1)
# flag some of the points
infl.ind1 <- which.max(D1) # top influence point
lev.ind1 <- h1 > 2*hbar1 # leverage more than 2x the average
clrs1 <- rep("black", len = n1)
clrs1[lev.ind1] <- "blue"
clrs1[infl.ind1] <- "red"
plot(h1, D1, xlab = "M1 Leverage", ylab = "M1 Cook's Influence Measure",
     pch = 21, bg = clrs1, cex = cex, cex.axis = cex)
p1 <- length(coef(M1))
n1 <- nrow(birth)
hbar1 <- p1/n1 # average leverage
abline(v = 2*hbar1, col = "grey60", lty = 2) # 2x average leverage
legend("topleft", legend = c("High Leverage", "High Influence"), pch = 21,
       pt.bg = c("blue", "red"), cex = cex, pt.cex = cex)

# cook's distance vs. leverage for M2
D2 <- cooks.distance(M2)
# flag some of the points
```

```r
infl.ind2 <- which.max(D2) # top influence point
lev.ind2 <- h2 > 2*hbar2 # leverage more than 2x the average
clrs2 <- rep("black", len = n1)
clrs2[lev.ind2] <- "blue"
clrs2[infl.ind2] <- "red"
plot(h2, D2, xlab = "M2 Leverage", ylab = "M2 Cook's Influence Measure",
     pch = 21, bg = clrs2, cex = cex, cex.axis = cex)
p2 <- length(coef(M2))
n2 <- nrow(birth)
hbar2 <- p2/n2 # average leverage
abline(v = 2*hbar2, col = "grey60", lty = 2) # 2x average leverage
legend("topleft", legend = c("High Leverage", "High Influence"), pch = 21,
       pt.bg = c("blue", "red"), cex = cex, pt.cex = cex)
```

A.24 Cross-Validation

```r
# models to compare
M1 <- Mfwd
M2 <- Mstep
Mnames <- expression(M[FWD], M[STEP])
# Cross-validation setup
nreps <- 2e3 # number of replications
ntot <- nrow(birth) # total number of observations
ntrain <- 500 # size of training set
ntest <- ntot-ntrain # size of test set
mspe1 <- rep(NA, nreps) # sum-of-square errors for each CV replication
mspe2 <- rep(NA, nreps)
logLambda <- rep(NA, nreps) # log-likelihod ratio statistic for each replication
system.time({
  for(ii in 1:nreps) {
    if(ii%%400 == 0) message("ii = ", ii)
    # randomly select training observations
    train.ind <- sample(ntot, ntrain) # training observations
    # refit the models on the subset of training data; ?update for details!
    M1.cv <- update(M1, subset = train.ind)
    M2.cv <- update(M2, subset = train.ind)
    # out-of-sample residuals for both models
    # that is, testing data - predictions with training parameters
    M1.res <- birth$wt[-train.ind] -
      predict(M1.cv, newdata = birth[-train.ind,])
    M2.res <- birth$wt[-train.ind] -
      predict(M2.cv, newdata = birth[-train.ind,])
```

```
    # mean-square prediction errors
    mspe1[ii] <- mean(M1.res^2)
    mspe2[ii] <- mean(M2.res^2)
    # out-of-sample likelihood ratio
    M1.sigma <- sqrt(sum(resid(M1.cv)^2)/ntrain) # MLE of sigma
    M2.sigma <- sqrt(sum(resid(M2.cv)^2)/ntrain)
    # since res = y - pred, dnorm(y, pred, sd) = dnorm(res, 0, sd)
    logLambda[ii] <- sum(dnorm(M1.res, mean = 0, sd = M1.sigma, log = TRUE))
    logLambda[ii] <- logLambda[ii] -
      sum(dnorm(M2.res, mean = 0, sd = M2.sigma, log = TRUE))
  }
})
```

```
## ii = 400

## ii = 800

## ii = 1200

## ii = 1600

## ii = 2000

##    user  system elapsed
##  13.926   0.540  14.471
```

A.25 M1 and M2 rPMSE

```
# Figure 7
# plot setup
cex <- .8
pch <- 16
par(mfrow = c(1,2), mar = c(4,4,1,1))

# plot rMSPE and out-of-sample log(Lambda)
boxplot(x = list(sqrt(mspe1), sqrt(mspe2)), names = Mnames, cex = .7,
        ylab = expression(sqrt(MSPE)), col = c("yellow", "orange"))

hist(logLambda, breaks = 50, freq = FALSE,
     xlab = expression(Lambda^{test}),
     main = "", cex = .7)
abline(v = mean(logLambda), col = "red") # average value
```

A.26 AIC for M1 and M2

```
# Akaike Information Criterion
# calculate the values AIC1 and AIC2
AIC1 <- AIC(M1)
AIC2 <- AIC(M2)
# display the results
AIC <- c(AIC1, AIC2)
AIC
```

```
## [1] 9576.488 9571.468
```

```
# display results for both AIC and PRESS
disp <- rbind(AIC = c(AIC1, AIC2),
              PRESS = c(sum(press1^2), sum(press2^2)))
colnames(disp) <- Mnames
disp # both metrics favor M2
```

```
##           M[FWD]    M[STEP]
## AIC     9576.488   9571.468
## PRESS 290492.304 289392.399
```

A.27 PRESS statistics

```
# Figure 8
# plot setup
cex <- .8
pch <- 16
par(mfrow = c(1,2), mar = c(5,5,1,1))

# PRESS Statistics
# PRESS statistics
press1 <- res1/(1-hatvalues(M1)) # M1
press2 <- res2/(1-hatvalues(M2)) # M2

# plot PRESS statistics
boxplot(x = list(abs(press1), abs(press2)), names = Mnames,
        ylab = expression(group("|", PRESS[i], "|")),
        col = c("yellow", "orange"))
```

A.28 Final candidate model: Mstep

```
summary(Mstep)
```

```
##
## Call:
## lm(formula = wt ~ gestation + parity + mage + feth + smoke_number +
##     BMI + I(gestation^2) + gestation:mage + gestation:BMI, data = birth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.636 -10.070  -0.418   9.184  62.229
##
## Coefficients:
##                                                                          Estimate
## (Intercept)                                                            -3.268e+02
## gestation                                                               2.605e+00
## parity                                                                  1.054e+00
## mage                                                                   -4.302e+00
## fethMexican                                                             1.380e+00
## fethAfrican-American                                                   -7.506e+00
## fethAsian                                                              -9.341e+00
## fethOther                                                              -1.098e+00
## smoke_numberSmoked more than half a pack but quit before pregnancy     -3.749e+00
## smoke_numberSmoked more than half a pack but quit during pregnancy     -6.489e+00
## smoke_numberSmoked up to half a pack but quit before pregnancy          2.400e+00
## smoke_numberStill smokes half to full pack                             -1.266e+01
## smoke_numberStill smokes more than one pack                            -9.227e+00
## smoke_numberStill smokes up to half a pack                             -4.517e+00
## BMI                                                                     7.310e+00
## I(gestation^2)                                                         -3.632e-03
## gestation:mage                                                          1.501e-02
## gestation:BMI                                                          -2.469e-02
##                                                                        Std. Error
## (Intercept)                                                             9.326e+01
## gestation                                                               5.246e-01
## parity                                                                  3.079e-01
## mage                                                                    1.474e+00
## fethMexican                                                             2.816e+00
## fethAfrican-American                                                    1.288e+00
## fethAsian                                                               2.800e+00
## fethOther                                                               3.100e+00
## smoke_numberSmoked more than half a pack but quit before pregnancy      2.873e+00
```

```
## smoke_numberSmoked more than half a pack but quit during pregnancy  3.007e+00
## smoke_numberSmoked up to half a pack but quit before pregnancy      2.138e+00
## smoke_numberStill smokes half to full pack                          1.900e+00
## smoke_numberStill smokes more than one pack                         1.365e+00
## smoke_numberStill smokes up to half a pack                          1.240e+00
## BMI                                                                 2.498e+00
## I(gestation^2)                                                      8.297e-04
## gestation:mage                                                      5.261e-03
## gestation:BMI                                                       8.925e-03
##                                                                    t value
## (Intercept)                                                         -3.504
## gestation                                                            4.966
## parity                                                               3.423
## mage                                                                -2.918
## fethMexican                                                          0.490
## fethAfrican-American                                                -5.829
## fethAsian                                                           -3.337
## fethOther                                                           -0.354
## smoke_numberSmoked more than half a pack but quit before pregnancy  -1.305
## smoke_numberSmoked more than half a pack but quit during pregnancy  -2.158
## smoke_numberSmoked up to half a pack but quit before pregnancy       1.122
## smoke_numberStill smokes half to full pack                          -6.665
## smoke_numberStill smokes more than one pack                         -6.759
## smoke_numberStill smokes up to half a pack                          -3.644
## BMI                                                                  2.926
## I(gestation^2)                                                      -4.378
## gestation:mage                                                       2.853
## gestation:BMI                                                       -2.766
##                                                                    Pr(>|t|)
## (Intercept)                                                        0.000477 ***
## gestation                                                          7.89e-07 ***
## parity                                                             0.000641 ***
## mage                                                               0.003597 **
## fethMexican                                                        0.624139
## fethAfrican-American                                               7.27e-09 ***
## fethAsian                                                          0.000876 ***
## fethOther                                                          0.723156
## smoke_numberSmoked more than half a pack but quit before pregnancy 0.192257
## smoke_numberSmoked more than half a pack but quit during pregnancy 0.031129 *
## smoke_numberSmoked up to half a pack but quit before pregnancy     0.261911
## smoke_numberStill smokes half to full pack                         4.14e-11 ***
## smoke_numberStill smokes more than one pack                        2.22e-11 ***
## smoke_numberStill smokes up to half a pack                         0.000281 ***
## BMI                                                                0.003497 **
## I(gestation^2)                                                     1.31e-05 ***
```

```
## gestation:mage                                                  0.004405 **
## gestation:BMI                                                   0.005762 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.68 on 1127 degrees of freedom
## Multiple R-squared:  0.2849, Adjusted R-squared:  0.2741
## F-statistic: 26.42 on 17 and 1127 DF,  p-value: < 2.2e-16
```