Homework 2

Guillem_Amat

September 14, 2019

Old Faithful

Reconsider the question on the Old Faithful eruptions from the last homework. Remember the data included a variable for the day of the eruptions (called "Date" in the dataset). *For this question, use the same data in the "OldFaithful.csv" file you used for question one of the first homework.*

Exercise 1

Fit a regression of interval on duration and day (treated as a categorical/factor variable). Is there a significant difference in mean intervals for any of the days (compared to the first day)? Interpret the effects of controlling for the days (do so only for the days with significant effects, if any).

```
#Importing the Data
Eruptions_Data <- read.csv("C:/Users/Guillem/Desktop/Duke University/Modelling and Representation of Da

#Fitting dates as factors
Eruptions_Model_c <- lm(Interval ~ Duration + as.factor(Date), data = Eruptions_Data)

summary(Eruptions_Model_c)
```

```
##
## Call:
## lm(formula = Interval ~ Duration + as.factor(Date), data = Eruptions_Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.3886  -4.7332  -0.5622   3.9759  15.9639
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       32.8770     3.0672  10.719   <2e-16 ***
## Duration          10.8813     0.6622  16.431   <2e-16 ***
## as.factor(Date)2   1.3275     2.7173   0.489    0.626
## as.factor(Date)3   0.7825     2.6994   0.290    0.773
## as.factor(Date)4   0.1625     2.6461   0.061    0.951
## as.factor(Date)5   0.2463     2.6459   0.093    0.926
## as.factor(Date)6   1.9918     2.6580   0.749    0.455
## as.factor(Date)7  -0.1700     2.7020  -0.063    0.950
## as.factor(Date)8  -0.6944     2.6957  -0.258    0.797
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.866 on 98 degrees of freedom
## Multiple R-squared:  0.7408, Adjusted R-squared:  0.7196
## F-statistic:    35 on 8 and 98 DF,  p-value: < 2.2e-16
```

After fitting the linear model to the data with Dates as factors we can observe their corresponding P-values are quite high. A high P-value means that we do not have enough evidence to conclude that a non-zero correlation exists between Dates and Eruptions.

Exercise 2

Perform an F-test to compare this model to your model for this data from the last homework. In context of the question, what can you conclude from the results of the F-test?

```
Eruptions_Model_p <- lm(Interval ~ Duration, data = Eruptions_Data)

anova(Eruptions_Model_c, Eruptions_Model_p)
```

```
## Analysis of Variance Table
##
## Model 1: Interval ~ Duration + as.factor(Date)
## Model 2: Interval ~ Duration
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     98 4620.2
## 2    105 4689.0 -7   -68.853 0.2086 0.9828
```

We performed an Anova test between the two models, which is equivalent to performing an F-test. The high F-test statistics suggests that dates are not important for the predictive value of the model.

Exercise 3

Using k-fold cross validation (with k=10), compare the average RMSE for this model and the average RMSE for your model from the last homework. Which model appears to have higher predictive accuracy based on the average RMSE values?

K-fold Cross Validation on previous Model:

```
set.seed(123)
K <- 10
Eruptions_Sample <- Eruptions_Data[sample(nrow(Eruptions_Data)),]
RMSE_list <- matrix(0, nrow = K, ncol = 1)
kth_fold <- cut(seq(1, nrow(Eruptions_Sample)), breaks = K, labels = FALSE)


for(k in 1:K){
train_index <- which(kth_fold==k)
train <- Eruptions_Sample[-train_index,]
test <- Eruptions_Sample[train_index,]
train_Eruptions <- lm(Interval~Duration, data = train)
Eruptions_Prediction <- predict(train_Eruptions, test)
RMSE_list <- sqrt(mean((test$Interval - Eruptions_Prediction)^2))
}

mean(RMSE_list)
```

```
## [1] 6.029183
```

K-fold Cross Validation on current Model:

```
set.seed(123)
K <- 10
Eruptions_Sample <- Eruptions_Data[sample(nrow(Eruptions_Data)),]
RMSE_list <- matrix(0, nrow = K, ncol = 1)
```

```r
kth_fold <- cut(seq(1, nrow(Eruptions_Sample)), breaks = K, labels = FALSE)


for(k in 1:K){
train_index <- which(kth_fold==k)
train <- Eruptions_Sample[-train_index,]
test <- Eruptions_Sample[train_index,]
train_Eruptions <- lm(Interval ~ Duration + as.factor(Date), data = train)
Eruptions_Prediction <- predict(train_Eruptions, test)
RMSE_list <- sqrt(mean((test$Interval - Eruptions_Prediction)^2))
}

mean(RMSE_list)
```

```
## [1] 6.700356
```

After performing K-fold Cross Validation on both models we can see that our previous model has higher predictive accuracy than the current one. The current model has significantly higher RMSE(+0.7), which means that residuals tend to be further away from their fitted values.

Maternal Smoking and Birth Weights

Summary

The study identified a non-zero relationship between mothers that smoke and the weight of their babies. A likely range of values for the weight difference would be between 6 and 11 units. Race and mother's smoking habits were analyzed to check whether they had a significant interaction. The analysis yielded a negative result. The other variables were also analyzed for significant interactions but none were identified.

Introduction

The study concerned investigating whether there were significant differences on baby weights based on whether their mothers smoked or not during their pregnancy. Other factors such as mother's weight, height or race were also evaluated based on their predictive ability on babies weight.

Data

The original dataset contained 21 variables that included parent's and pregnancy information for babies being born at the Kaiser Foundation Hospital. Many values where missing for the fathers, as it is sometimes hard to record information on them, so a first step in the analysis was to filter columns with this issue in order to get a dataset with complete observations.

A next step included filtering down the dataset to only include observations for mothers who smoked during their pregnancy and those that never did. Cases where mothers smoked until the pregnancy and mothers who smoked in the past were excluded from the analysis.

Furthermore some variables were collapsed: 0-5 values in race to white and 6-7 in education to trade school. The Id and date variables were also excluded from the analysis as they did not add any specific information. [Appendix 1]

The variable weight was plotted as a histogram to observe its distribution. As the distribution was completely normal, no transformations were considered. [Appendix 2]

A linear model predicting weight based on smoke was fitted to the data. The null hypothesis of no correlation between the two variables was rejected given the extremely low P-value: We have enough evidence to say that mothers who smoke have on average babies with lower weight. A 95% confidence interval for the model yielded a likely range for the difference between mothers that smoked against those that did not of 6 to 11 absolute units. [Appendix 2]

Another model was fitted, this time including race and the interaction between race and smoking. Although race appeared to be a good predictor of baby weight, the interaction between race and smoke was not meaningful. [Appendix 3]

A correlation table was used to identify if there was any specific relationship between variables and a potential interaction between Age and Parity was identified. This interaction was not significant enough and it was not included in the final model when performing Stepwise Selection. [Appendix 4]

Model

The final model selected was *bwt.oz ~ parity + mrace + mht + mpregwt + smoke*.

The model was selected based on the results of the EDA and using a Forward Stepwise Selection Method by maximizing AIC. AIC provides a means of model selection by comparing models. AIC penalizes complexity of models and since there were a lot of variables in the dataset, it was determined that it would be a good criteria for selection. [Appendix 5]

Results and Conclusions

The final model had an Intercept of 49, which would be the predicted weight of baby from a non-smoking, white mother with 0 parity, height and weight (difficult to interpret given that no one weighs 0 kg or is 0 cm tall). The assumptions of Linearity, Normality, Equal Variance and Independence were all met by the model. Adjusted R-Squared, which determines how much of the variance is explained by the model, is quite low at 0.15. My hypothesis would be that the weight of babies is mostly determined by genetics, although I would need to further explore this claim. We have grounds nevertheless to claim that mother's smoking habits, their race, height, weight and parity influence babies weight to a certain extent. [Appendix 5 and 6]

Appendix

Appendix 1. Importing and Cleaning Data

```
#Installing dplyr
#install.packages("dplyr")
#install.packages("plyr")
library(plyr)
```

```
## Warning: package 'plyr' was built under R version 3.5.3
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.5.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
#rm(list =ls())
#setwd("C:/Users/Guillem/Desktop/Duke University/Modelling and Representation of Data/1_Homework Exerci

#We import the data
Smoking <- read.csv("C:/Users/Guillem/Desktop/Duke University/Modelling and Representation of Data/1_Ho

#We simplify the data by filtering down to mothers who smoke and mothers that have never smoked
Smoking_Simple <- filter(Smoking, smoke == 1 | smoke == 0)
Smoking_Simple$mrace <- as.factor(Smoking_Simple$mrace)
Smoking_Simple$med <- as.factor(Smoking_Simple$med)
Smoking_Simple$date <- NULL
Smoking_Simple$id <- NULL

#Renaming Values
Smoking_Simple <- transform(Smoking_Simple, mrace = revalue(mrace, c("0" = "white", "1"="white", "2" = 

Smoking_Simple <- transform(Smoking_Simple, med = revalue(med, c("0" = "< 8th", "1" = "8th to 12th", "2
```

```
## The following `from` values were not present in `x`: 6, 9
```

Appendix 2: Fitting a linear model based on Mothers who smoke

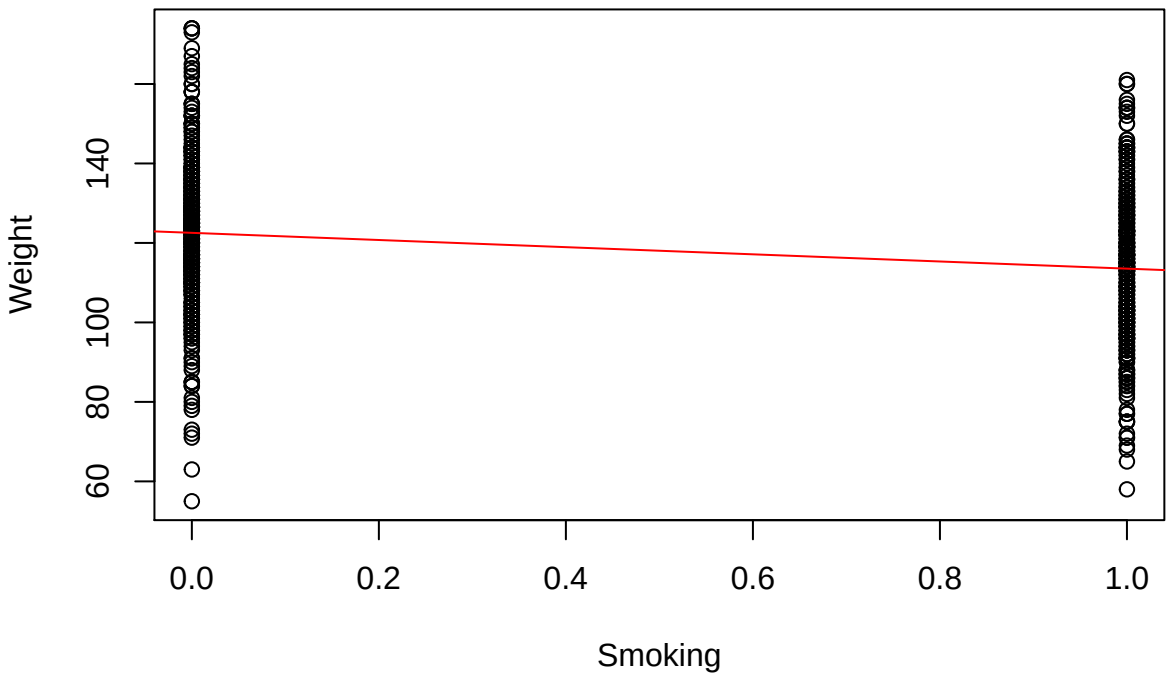
```r
#Trying to fit a line to predict weight based on whether a mother smokes or not
Smoke_model <- lm(bwt.oz ~ as.factor(smoke), data = Smoking_Simple)
summary(Smoke_model)
```

```
##
## Call:
## lm(formula = bwt.oz ~ as.factor(smoke), data = Smoking_Simple)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.539 -10.539   0.471  10.461  51.461
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       122.5386     0.8103 151.226  < 2e-16 ***
## as.factor(smoke)1  -9.0101     1.1899  -7.572 9.39e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.49 on 867 degrees of freedom
## Multiple R-squared:  0.06203,    Adjusted R-squared:  0.06095
## F-statistic: 57.34 on 1 and 867 DF,  p-value: 9.39e-14
```

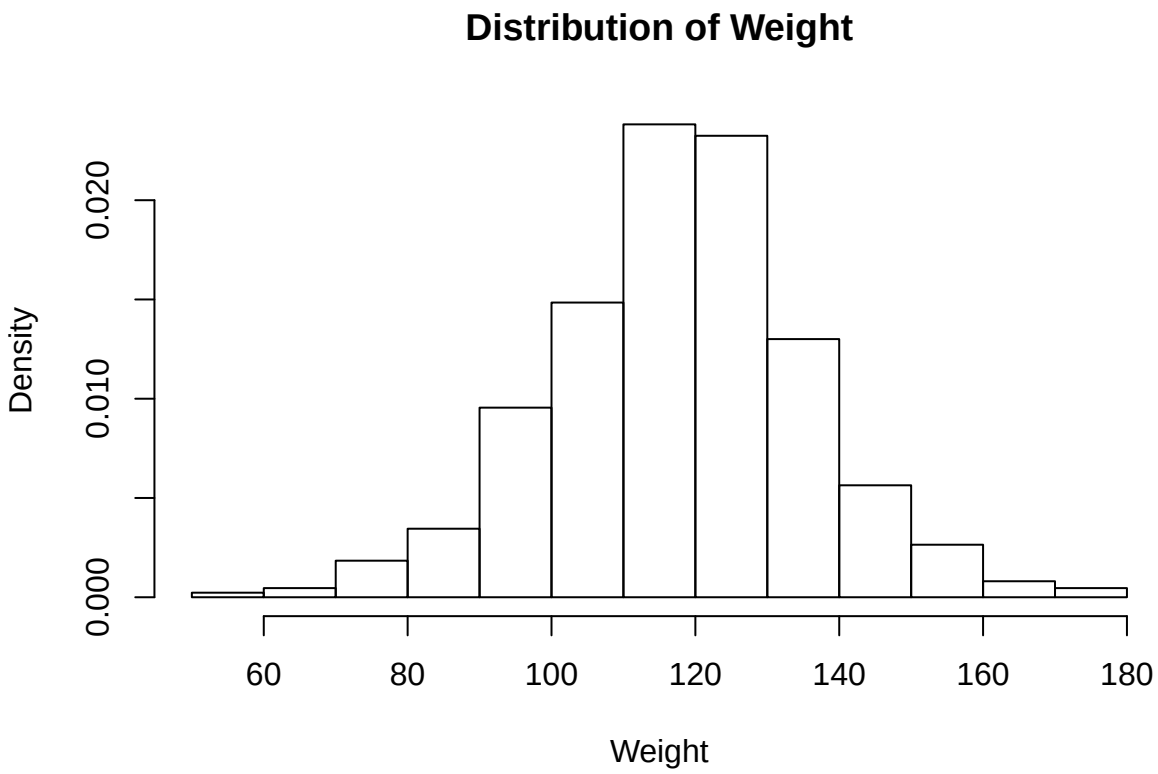
```r
#Confidence Interval
confint(Smoke_model, level = 0.95)
```

```
##                       2.5 %     97.5 %
## (Intercept)       120.94824 124.129009
## as.factor(smoke)1 -11.34548  -6.674704
```

```r
#Slope is not 0, very low p-value, but further relation needed
plot(Smoking$smoke, Smoking$bwt.oz, xlab = "Smoking", ylab = "Weight", main="")
abline(Smoke_model, col = "red")
```




```
#weight is normally distributed
hist(Smoking_Simple$bwt.oz, xlab = "Weight", main = "Distribution of Weight", freq = FALSE)
```



Appendix 3: Race Analysis

```
Smoking_Race <- lm(bwt.oz ~ smoke + mrace + smoke:mrace,data = Smoking_Simple)
summary(Smoking_Race)
```

```
##
## Call:
## lm(formula = bwt.oz ~ smoke + mrace + smoke:mrace, data = Smoking_Simple)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -62.37 -10.71   0.04  10.63  56.63
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        124.70886    0.96527 129.196  < 2e-16 ***
## smoke               -9.49918    1.37169  -6.925 8.53e-12 ***
## mracemexican        -2.18254    4.05317  -0.538  0.59039
## mraceblack          -7.33652    2.01594  -3.639  0.00029 ***
## mraceasian         -11.10886    3.56498  -3.116  0.00189 **
## mracemix            -0.20886    5.04656  -0.041  0.96700
## smoke:mracemexican  11.13953    8.15169   1.367  0.17213
## smoke:mraceblack     0.08684    2.98991   0.029  0.97684
## smoke:mraceasian    -6.21193    6.80981  -0.912  0.36192
## smoke:mracemix     -10.33415   11.16072  -0.926  0.35474
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 17.16 on 859 degrees of freedom
## Multiple R-squared:  0.1057, Adjusted R-squared:  0.09636
## F-statistic: 11.28 on 9 and 859 DF,  p-value: < 2.2e-16
```

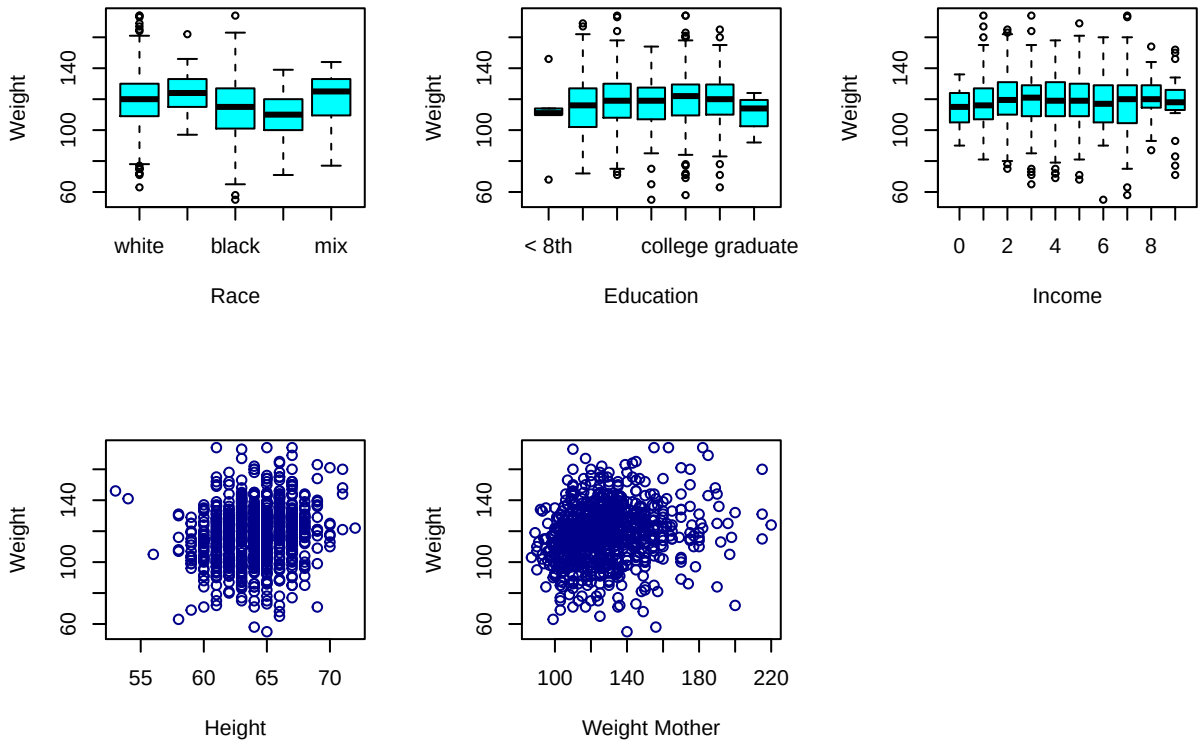
Appendix 4: Correlation between variables

```
cor(Smoking_Simple[,c("bwt.oz", "parity", "mage", "mht", "mpregwt")])
```

```
##             bwt.oz      parity         mage          mht   mpregwt
## bwt.oz   1.00000000  0.04106952  0.044343562  0.187758035 0.1821158
## parity   0.04106952  1.00000000  0.523690421 -0.042815618 0.1505379
## mage     0.04434356  0.52369042  1.000000000 -0.005470885 0.1461368
## mht      0.18775804 -0.04281562 -0.005470885  1.000000000 0.4604463
## mpregwt 0.18211578  0.15053789  0.146136818  0.460446304 1.0000000
```

```
par(mfrow = c(2, 3))

#boxplot(bwt.oz ~ parity,data=Smoking_Simple, ylab="Weight", xlab="Parity", col=rainbow(15))
boxplot(bwt.oz ~ mrace,data=Smoking_Simple, ylab="Weight", xlab="Race", col="cyan")
#boxplot(bwt.oz ~ mage,data=Smoking_Simple, ylab="Weight", xlab="Age", col=rainbow(15))
boxplot(bwt.oz ~ med,data=Smoking_Simple, ylab="Weight", xlab="Education", col="cyan")
boxplot(bwt.oz ~ inc,data=Smoking_Simple, ylab="Weight", xlab="Income", col="cyan")
plot(bwt.oz ~ mht,data=Smoking_Simple, ylab="Weight", xlab="Height", col="darkblue")
plot(bwt.oz ~ mpregwt,data=Smoking_Simple, ylab="Weight", xlab="Weight Mother", col="darkblue")
```



Appendix 5: Stepwise Model Selection

```r
#Defining Null and Full Model
NullModel <- lm(bwt.oz~1,data= Smoking_Simple)
FullModel <- lm(bwt.oz ~ parity + mrace + mage + med + mht + mpregwt + inc + smoke + parity:mage + mrac

#Steping through the models, backwards and forwards
Model_Forward <- step(NullModel, scope = formula(FullModel), direction = "forward", trace = 0)
Model_Backward <- step(FullModel, scope = formula(NullModel), direction = "backward", trace = 0)

#Checking the generated models
Model_Backward$call
```

```
## lm(formula = bwt.oz ~ parity + mrace + mht + mpregwt + smoke,
##     data = Smoking_Simple)
```

```r
Model_Forward$call
```

```
## lm(formula = bwt.oz ~ smoke + mht + mrace + mpregwt + parity,
##     data = Smoking_Simple)
```

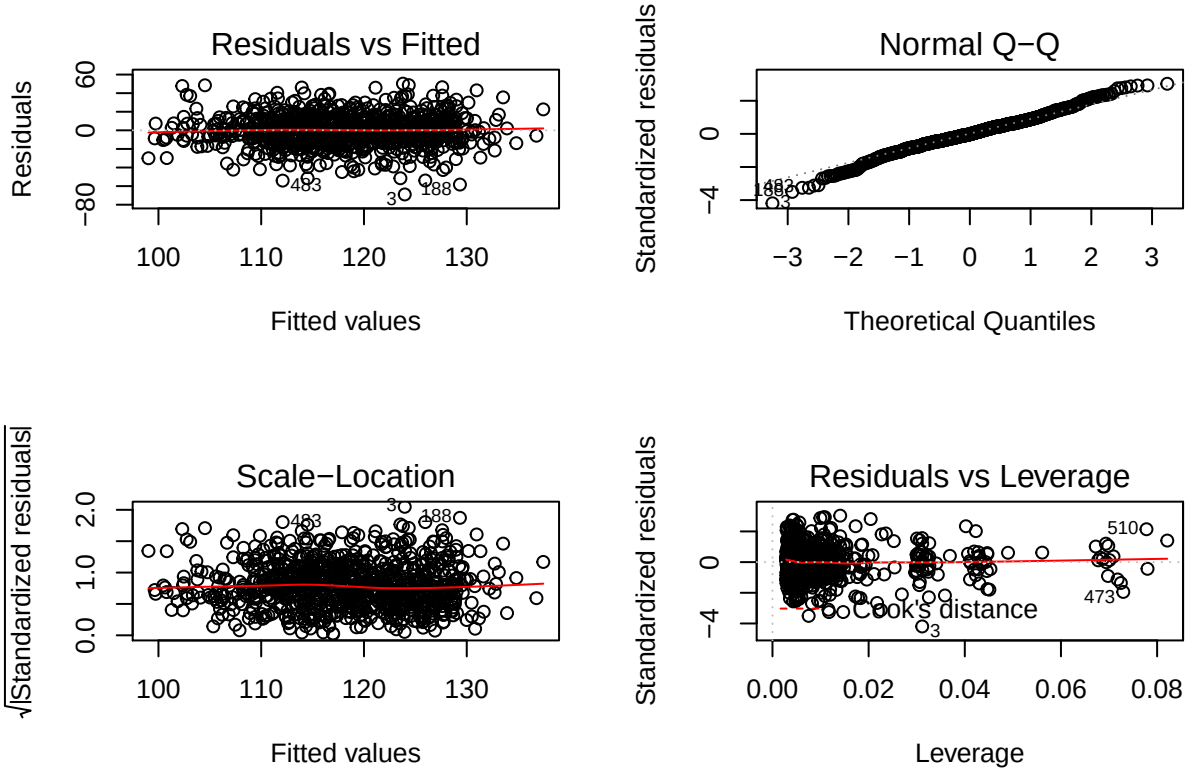
```r
summary(Model_Forward)
```

```
##
## Call:
## lm(formula = bwt.oz ~ smoke + mht + mrace + mpregwt + parity,
##     data = Smoking_Simple)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.969  -9.525  -0.336  10.131  50.206
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  49.64712   15.38492   3.227 0.001298 **
## smoke        -9.35194    1.15218  -8.117 1.65e-15 ***
## mht           0.93387    0.26070   3.582 0.000360 ***
## mracemexican  3.29715    3.46725   0.951 0.341902
## mraceblack   -8.82690    1.51623  -5.822 8.22e-09 ***
## mraceasian   -7.93888    3.03506  -2.616 0.009060 **
## mracemix     -1.98421    4.38639  -0.452 0.651126
## mpregwt       0.10808    0.03217   3.360 0.000814 ***
## parity        0.66507    0.31422   2.117 0.034584 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.71 on 860 degrees of freedom
## Multiple R-squared:  0.1514, Adjusted R-squared:  0.1435
## F-statistic: 19.17 on 8 and 860 DF,  p-value: < 2.2e-16
```

Appendix 6: Model assessment

The model seems to meet the Normality, Equal Variance, Independence and Linearity assumptions.

```r
par(mfrow = c(2,2))
plot(Model_Forward, which=1)
plot(Model_Forward, which=2)
plot(Model_Forward, which=3)
plot(Model_Forward, which=5)
```


Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage