ST3131 Assignment (Koo Yong Jie A0201805U)

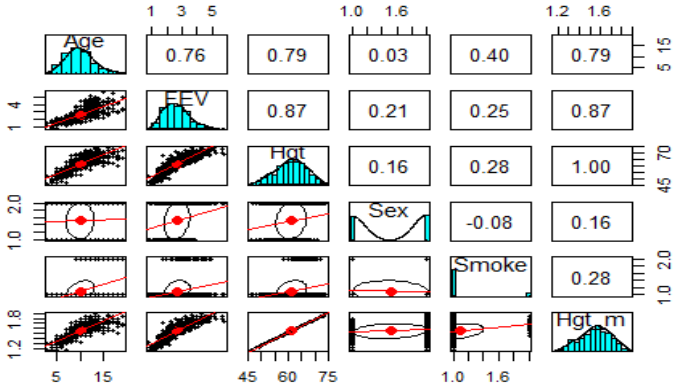
1. Introduction

In this report, I will be investigating the differences in Forced Expiratory Volume (FEV), an index of pulmonary function that measures the volume of air expelled after 1 second of constant effort, among children aged 3 to 19 years. I have developed a linear regression model to predict FEV for the different ages using various variables in question.

Using R, I first come begin exploring dataset of 654 entries, in particular, the response variable FEV and its relation with each of the regressors. Next, I explored the regressors themselves by considering any possible relationship between them. Then, I come up with a basic model and begin testing its appropriateness and adequacy. Finally, after considering the results from the regression output, a final model was created that I deemed most appropriate to represent the relationship between the regressors and FEV.

2. Data Exploration

Figure 1: Correlation Matrix



2.1. Response FEV

Figure 2: Boxplot at 1.5 IQR

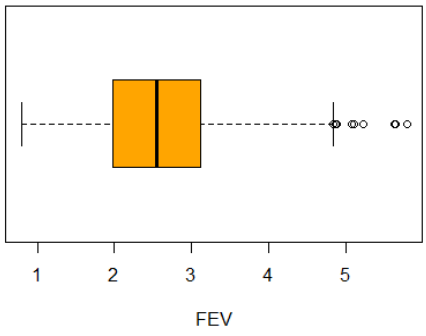
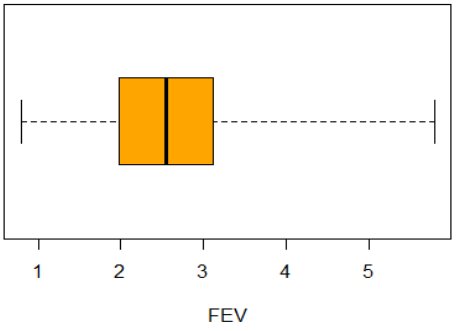


Figure 3: Boxplot at 3 IQR



The response FEV, regressors Hgt, Hgt_m and Age are continuous variables, while regressors Sex and Smoke are categorical variables. I took Hgt_m as the variable to be compared and use in modelling later as Hgt_m and Hgt are just measurements of the same thing, differing by a fixed factor, with correlation coefficient 1.00 as depicted in Figure 1.

From the box plots (Figures 2 and 3), there are 9 outliers found from boxplot of 1.5 IQR and and none from the boxplot of 3 IQR for response variable FEV. In this case, the outliers are mild, so I choose to leave them intact.

Figure 4: Histogram of response variable FEV

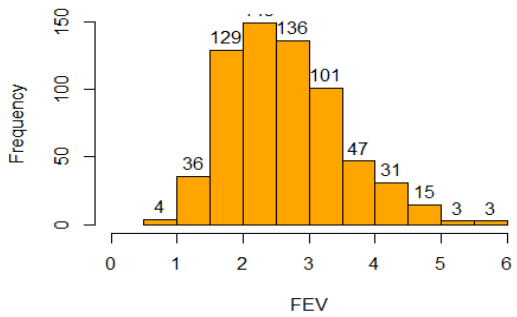


Figure 5: QQ-plot of FEV

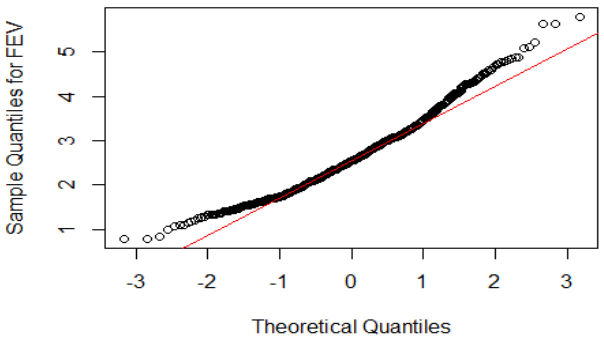


Table 1: Descriptive Statistics for FEV, Age, Height(in), Height(m)

vars	n	mean	sd	median	min	max	range	skew	kurtosis
FEV	654	2.637	0.867	2.55	0.79	5.79	5.00	0.659	0.285
Age	654	9.931	2.954	10.00	3.00	19.00	16.00	0.412	0.094
Height(in)	654	61.144	5.704	61.50	46.00	74.00	28.00	-0.213	-0.501
Height(m)	654	1.553	0.145	1.56	1.17	1.88	0.71	-0.209	-0.496

From the histogram of FEV (Figure 4), it appears that the distribution of data for FEV is single-peaked and slightly right-skewed, from Table 1, FEV ranges from 0.79 to 5.79.

The normal probability plot of FEV (Figure 5) also shows a slightly right-skewed distribution with the points forming an upward sloping curve.

Overall, FEV is largely normal, hence no transformation of the response variable is required at this stage.

Figure 6: Scatterplot of FEV to Age

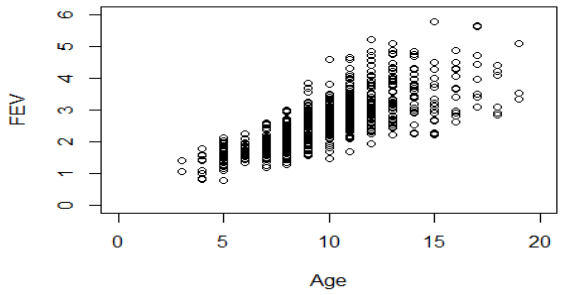
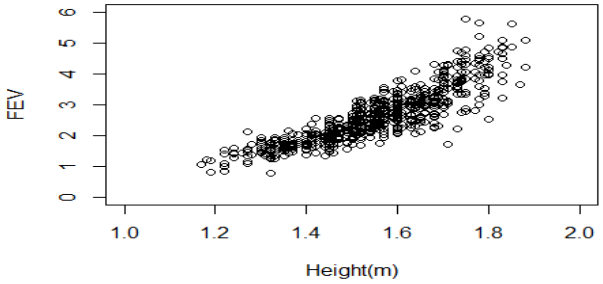


Figure 7: Scatterplot of FEV to Height(m)



The scatterplot of FEV to Age (Figure 6) seems to show a strong positive linear relationship between Age and FEV in general, with a correlation coefficient of 0.76 as depicted in Figure 1.

The scatterplot of FEV to Height(m) (Figure 7) both show a similar strong positive linear relationship between Hgt_m and FEV in general, with a correlation coefficient of 0.87 from Figure 1.

Figure 8: Boxplot of FEV to Sex

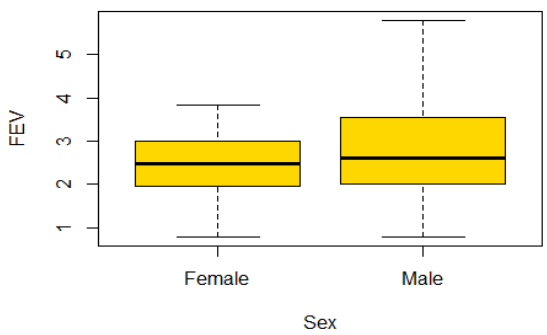
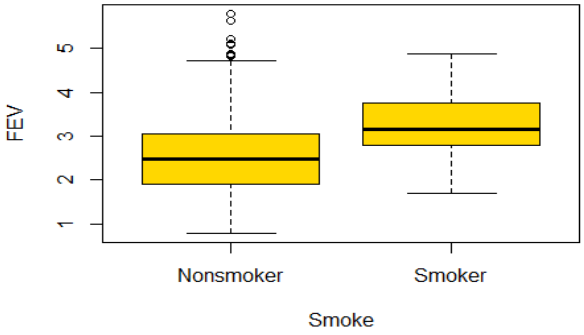


Figure 9: Boxplot of FEV to Smoke



The boxplot of FEV to Sex (Figure 8) does not show any significant relationship between FEV and Sex as the median FEV are roughly the same for both sexes.

The boxplot of FEV to Smoke (Figure 9) seems to show that smokers tends to have higher FEV as compared to nonsmokers as the median FEV for smokers is higher than that for nonsmokers.

2.2. Regressors

Figure 10: Scatterplot of Height(m) to Age

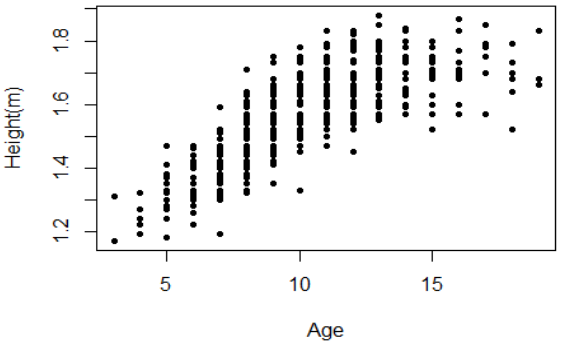
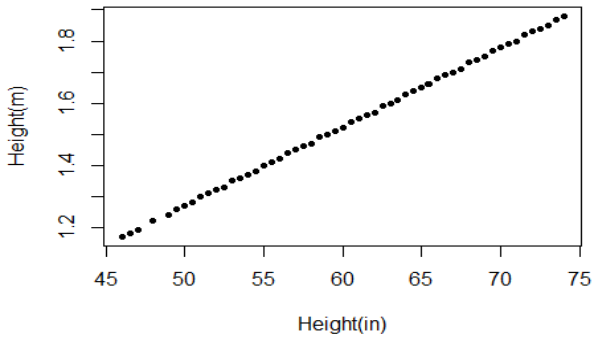


Figure 11: Scatterplot of Height(m) to Height(in)



From the scatterplot of Hgt_m to Age (Figure 10), there appears to be a strong positive relationship between Age and Height, which is expected since children in this age range (of 3 - 19) tend to grow taller with age, supported by the large correlation positive coefficient of 0.79 from Figure 1. As such, Hgt_m would be a physical constraint and hence source of natural multicollinearity when building the model.

As mentioned previously, the regressors Hgt and Hgt_m differs by just a fixed factor, hence Figure 11 clearly depicts their logical strong positive linear relationship, further supported by the perfect positive correlation coefficient of 1 from Figure 1.

Figure 12: Boxplot of Age to Sex

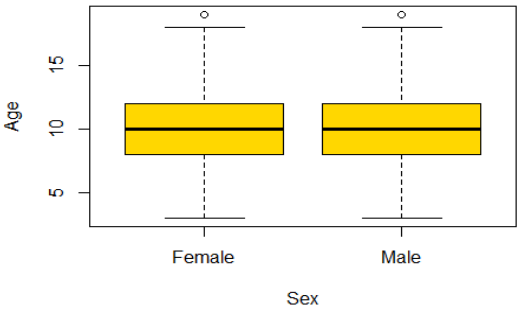


Figure 13: Boxplot of Age to Smoking Status

From the boxplots of Age to Sex (Figure 12), there does not appear to be a relationship between Age and Sex as for both sexes, the median, interquartile range and range are the same.

From the boxplots of Age to Smoke (Figure 13), it seems that the larger the age, the more likely the child is a smoker, as the interquartile range of age as well as the median for smokers is entirely above that of non-smokers.

Figure 14: Boxplot of Height(m) to Sex

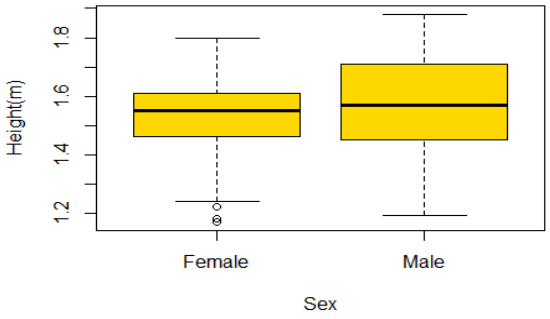
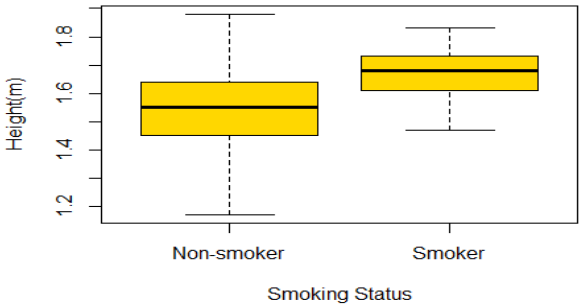


Figure 15: Boxplot of Height(m) to Smoking Statu



From the boxplot of Hgt_m to Sex (Figure 14), the median height for males is roughly the same as that for females, suggesting that there might not be a relationship between Hgt_m and Sex.

From the boxplot of Hgt_m to Smoke (Figure 15), the median height for smokers is higher than that for non-smokers but the range of heights is smaller, suggesting that there could be a relationship between Hgt_m and Smoke.

```
##              sex
## smoke         Female Male
##   Non-Smoker    279  310
##   Smoker         39   26
```

Table 2: Smoke VS Sex

From Table 2, there does not seem to be a relationship between Smoke and Sex as the number of non-smokers is larger for both sexes as compared to smokers.

3. Choosing A Model

3.1. Basic Model

First model (results attached in R-codes as Figures 16) with all the regressors: hat{FEV} = -4.449226 + 0.066086 * Age + 4.094782 * Hgt_m - 0.089377 * I(Smoke = 1) + 0.155678 * I(Sex = 1) with $R^2 = 0.7745$ which means that 77.45% of variability of response data around its mean is explained the model and a strong positive linear relationship between the response and the regressors.
Hypotheses to test the significance of m1: H_0: all coefficients = 0 and H_1: at least a coefficient is non-zero.

Test statistic: $F = 557.3 \sim F_{4,649}$ which has p-value < 2.2e-16. Hence, data provide strong evidence that model is significant.
From the model summary statistics table, Smoke is less significant compared to the other regressors, with its coefficient having a p-value of 0.133.

From the Anova table (Figure 17), it appears that Smoke is less significant in the model with its low SS_R values.

3.2. Stepwise regression model

Employing stepwise regression (Figure 18) based on the metric AIC using both forward and backward elimination so as not to underfit or overfit the data, we end up with the same model as the basic model with AIC value of -1151.90.

3.3. Model Evaluation

3.3.1. Residual Plot

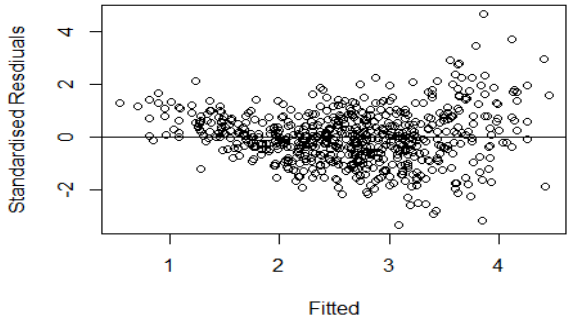


Figure 19: Residual Plot from m1

The residual plots of standardised residuals vs fitted values (Figure 19) seems to show an outward opening funnel and they cover a large range from -4 to 4.5, violating constant variance assumption, suggesting that a transformation should be done to the model.

3.3.2. Normal Probability Plot

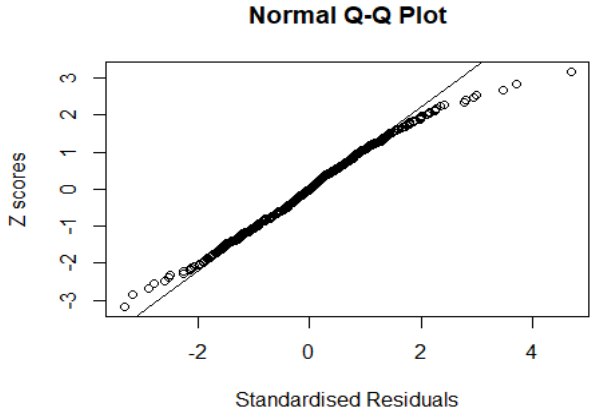


Figure 20: QQ-Plot of residuals from m1

The normal probability plot of the standardised residuals (Figure 20) looks normal, with a large majority of point closely aligned to the line, though the right tail is slightly thinner.

3.3.3. Collinearity Analysis

From early analysis for Figure 10 with Figure 1 and now Table 3 (attached in R-codes), there is a high correlation between Age and Hgt_m since $x_{12} = 0.7917857$.

Proceeding to check the Variation Inflation Factors (VIF) and Condition Number of the correlation matrix, none of the VIF values in Table 4 are > 10, and the condition number = 10.73429 < 100, hence there does not seem to be multicollinearity in the data.

4. Improved Model

4.1. Applying Boxcox Method

From 3.3.1, the residual plot is not satisfactory, hence I choose to apply the Boxcox method to transform the response in hope of achieving a better residual plot.

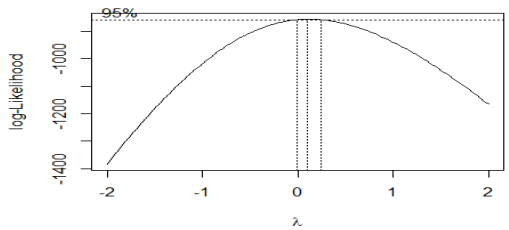


Figure 21: BoxCox plot

The BoxCox plot suggests that lambda = 0 can be used, i.e natural logarithm transformation.

New fitted model (results in Figure 22): hat{ln(FEV)} = -1.940303 + 0.023628*Age + 1.681433*Hgt_m - 0.047056*I(Smoke = 1) + 0.028735*I(Sex = 1) with a greater R^2 value of 0.8096.

Hypotheses to test the significance of m2: H_0: all coefficients = 0 and H_1: at least a coefficient is non-zero.
Test statistic: F = 689.7 ~ $F_{4,649}$ which has p-value < 2.2e-16. Hence, data provide strong evidence that model is significant.
Furthermore, all the coefficient before the regressors are significant with their low p-values.
From the Anova table (Figure 23), though Smoke and Sex seem less significant compared to the other variables, all of them have significant p-values from the F-test.

4.2. Model Evaluation

4.2.1. Residual Plot

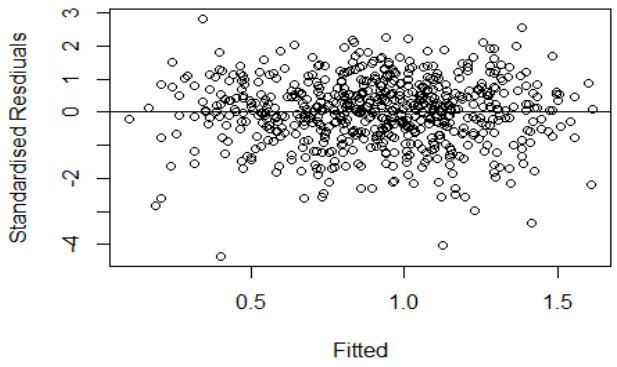


Figure 24: Residual Plot from m2

The residual plot of standardised residuals vs fitted values (Figure 24) is now satisfactory, with all the points randomly scattered about 0, between -4.5 and 3.

4.2.2. Normal Probability Plot

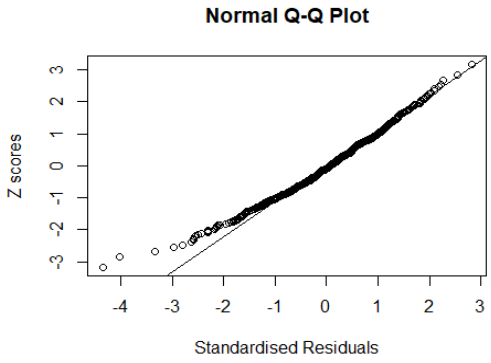


Figure 25: QQ-Plot of residuals from m2

The normal probability plot of the standardised residuals (Figure 25) looks normal, with a large majority of point closely aligned to the line, though the left tail is slightly thinner.

5. Conclusion

With no constant variance assumption or normality assumption violated, a greater R^2 value and a greater overall significance of the model with a large F value, it appears that model m2, transformed from model m1, is a significant linear model that could be used to predict FEV. As such my final proposed model is:

$$\ln(FEV) = -1.940303 + 0.023628*Age + 1.681433*Hgt\_m - 0.047056*I(Smoke = 1) + 0.028735*I(Sex = 1) + \varepsilon$$

6. Appendix (R-codes and some figures mentioned in the analysis above)

```
data <- read.csv("FEV.csv", header = TRUE, sep = ',')
data$Sex = as.factor(data$Sex)
data$Smoke = as.factor(data$Smoke)
attach(data)
```

```
pairs.panels(data[colnames(data)!="ID"], lm=TRUE, main = "Figure 1: Correlation Matrix")

# response

b1 <- boxplot(data$FEV, horizontal = T, col = "orange", xlab = "FEV", main = "Figure 2: Boxplot at 1.5 IQR")
b2 <- boxplot(data$FEV, horizontal = T, range = 3, col = "orange", xlab = "FEV", main = "Figure 3: Boxplot at 3 IQR")
b1$out

h1 <- hist(data$FEV, main = "Figure 4: Histogram of response variable FEV", xlab = "FEV", col = "orange", labels = T,
xlim = c(0, 6))

qqnorm(data$FEV,ylab="Sample Quantiles for FEV", main = "Figure 5: QQ-plot of FEV")
qqline(data$FEV,col="red")
```

```
index <- describe(data$FEV)
age <- describe(data$Age)
ht <- describe(data$Hgt)
ht_m <- describe(data$Hgt_m)
tab<-rbind(index, age, ht, ht_m)
tab$trimmed <- tab$mad <- tab$se <- NULL
tab$vars[1]<-"FEV"
tab$vars[2]<-"Age"
tab$vars[3]<-"Height(in)"
tab$vars[4]<-"Height(m)"
kable(tab,row.names = F,caption = "Descriptive Statistics for FEV, Age, Height(in), Height(m)",digits = 3)

#FEV to age
plot(data$Age, data$FEV, pch=1, xlab = "Age", ylab = "FEV", main = "Figure 6: Scatterplot of FEV to Age", xlim = c(0,20),
ylim = c(0,6))

#FEV to Hgt_m
plot(data$Hgt_m, data$FEV, xlab = "Height(m)", ylab = "FEV", main = "Figure 7: Scatterplot of FEV to Height(m)", xlim =
c(1,2), ylim = c(0,6))

boxplot(FEV~Sex, xlab = "Sex", names = c("Female", "Male"), ylab = "FEV", main = "Figure 8: Boxplot of FEV to Sex", col =
"gold")
boxplot(FEV~Smoke, xlab = "Smoke", names = c("Nonsmoker", "Smoker"), ylab = "FEV", main = "Figure 9: Boxplot of FEV to
Smoke", col = "gold")
```

```
#regressors

plot(data$Age, data$Hgt_m, xlab = "Age", ylab = "Height(m)", main = "Figure 10: Scatterplot of Height(m) to Age", pch =
20)
plot(data$Hgt, data$Hgt_m, xlab = "Height(in)", ylab = "Height(m)", main = "Figure 11: Scatterplot of Height(m) to
Height(in)", pch = 20)

boxplot(Age~Sex, xlab = "Sex", ylab = "Age", main = "Figure 12: Boxplot of Age to Sex", names = c("Female", "Male"), col
= "gold")

boxplot(Age~Smoke, xlab = "Smoking Status", ylab = "Age", main = "Figure 13: Boxplot of Age to Smoking Status", names =
c("Non-smoker", "Smoker"), col = "gold")

boxplot(Hgt_m~Sex, xlab = "Sex", ylab = "Height(m)", main = "Figure 14: Boxplot of Height(m) to Sex", names = c("Female",
"Male"), col = "gold")

boxplot(Hgt_m~Smoke, xlab = "Smoking Status", ylab = "Height(m)", main = "Figure 15: Boxplot of Height(m) to Smoking
Status", names = c("Non-smoker", "Smoker"), col = "gold")

smoke = factor(data$Smoke, labels = c('Non-Smoker', 'Smoker'))
sex = factor(data$Sex, labels = c('Female', 'Male'))
table(smoke,sex)
```

```r
#basic model
m1 = lm(FEV~Age + Hgt_m + Smoke + Sex, data=data)
summary(m1)
anova(m1)

#stepwise regression
sw <- step(m1,  direction = c("both"))
summary(sw)

#model adequacy check
plot(m1$fitted.values, rstandard(m1), xlab = "Fitted", ylab = "Standardised Resdiuals")
abline(h=0)

qqnorm(rstandard(m1), datax = T, ylab = "Standardised Residuals", xlab = "Z scores")
qqline(rstandard(m1), datax = T)

x<-cbind(Age, Hgt_m, Sex, Smoke)
x<-cor(x)
x

C<-solve(x)  #this is (X'X)^(-1) where X'X is in correlation form
VIF <- diag(C)
VIF
```

```r
#condition number is:
cond1 <- max(eigen(x)$values)/min(eigen(x)$values)
cond1 #condition number > 1000 --> strong multicollinearity
```

```
             Age      Hgt_m         Sex        Smoke
Age    1.0000000 0.7917857  0.02914420  0.40425248
Hgt_m  0.7917857 1.0000000  0.16032276  0.28119350
Sex    0.0291442 0.1603228  1.00000000 -0.07561166
Smoke  0.4042525 0.2811935 -0.07561166  1.00000000
     Age    Hgt_m      Sex    Smoke
3.016139 2.830049 1.061296 1.209343
[1] 10.73439
```

Table 3

and output from VIF and cond1

```
Call:
lm(formula = FEV ~ Age + Hgt_m + Smoke + Sex, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.36154 -0.25026  0.00473  0.25268  1.92639

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.449226   0.223535 -19.904  < 2e-16 ***
Age          0.066086   0.009501   6.956 8.57e-12 ***
Hgt_m        4.094782   0.187847  21.799  < 2e-16 ***
Smoke1      -0.089377   0.059351  -1.506    0.133
Sex1         0.155678   0.033282   4.678 3.53e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4129 on 649 degrees of freedom
Multiple R-squared:  0.7745,    Adjusted R-squared:  0.7731
F-statistic: 557.3 on 4 and 649 DF,  p-value: < 2.2e-16
```

Figure 16: summary of model m1

```
Analysis of Variance Table

Response: FEV
           Df  Sum Sq Mean Sq   F value    Pr(>F)
Age         1 280.893 280.893 1647.3231 < 2.2e-16 ***
Hgt_m       1  94.865  94.865  556.3428 < 2.2e-16 ***
Smoke       1   0.614   0.614    3.5986   0.05827 .
Sex         1   3.731   3.731   21.8799 3.533e-06 ***
Residuals 649 110.664   0.171
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 17: anova of model m1

```
Start:  AIC=-1151.9
FEV ~ Age + Hgt_m + Smoke + Sex

        Df Sum of Sq    RSS      AIC
<none>               110.66 -1151.90
- Smoke  1     0.387 111.05 -1151.62
- Sex    1     3.731 114.39 -1132.22
- Age    1     8.251 118.92 -1106.87
- Hgt_m  1    81.025 191.69  -794.61
```

Figure 18: stepwise regression

```
boxcox(m1, lambda=seq(-2, 2, by=0.5), optimize=TRUE, plotit = TRUE)
#The BoxCox plot suggest that lambda = 0 can be used.

m2<-lm(log(FEV) ~ Age + Hgt_m + Smoke + Sex, data=data)
summary(m2)
anova(m2)
```

```
Call:
lm(formula = log(FEV) ~ Age + Hgt_m + Smoke + Sex, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.63305 -0.08571  0.00991  0.09277  0.40943

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.940303   0.078950 -24.576  < 2e-16 ***
Age          0.023628   0.003356   7.041 4.86e-12 ***
Hgt_m        1.681433   0.066346  25.344  < 2e-16 ***
Smoke1      -0.047056   0.020962  -2.245   0.0251 *
Sex1         0.028735   0.011755   2.445   0.0148 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1458 on 649 degrees of freedom
Multiple R-squared:  0.8096,    Adjusted R-squared:  0.8084
F-statistic: 689.7 on 4 and 649 DF,  p-value: < 2.2e-16
```

Figure 22: summary of model m2 after applying BoxCox method, giving lambda = 0

```
Analysis of Variance Table

Response: log(FEV)
           Df Sum Sq Mean Sq   F value  Pr(>F)
Age         1 43.192  43.192 2030.5938 < 2e-16 ***
Hgt_m       1 15.237  15.237  716.3319 < 2e-16 ***
Smoke       1  0.128   0.128    6.0229 0.01438 *
Sex         1  0.127   0.127    5.9758 0.01477 *
Residuals 649 13.805   0.021
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 23: anova of model m2