1. Exploring response variable

The response variable y is FEV, while the regressors x_1, x_2, x_3, x_4, x_5 are Age, Hgt, Sex, Smoke, Hgt_m respectively. Age, Hgt, Hgt_m are quantitative variables while Sex and Smoke are categorical(qualitative) variables.

Thus, x_3 is an indicator variable for Sex:

$$x_3 = \begin{cases} 0 & if\ female \\ 1 & if\ male \end{cases}$$

x_4 is an indicator variable for Smoke:

$$x_4 = \begin{cases} 0 & if\ current\ non-smoker \\ 1 & if\ current\ smoker \end{cases}$$

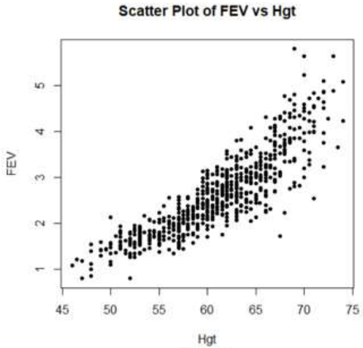


Figure 1.2

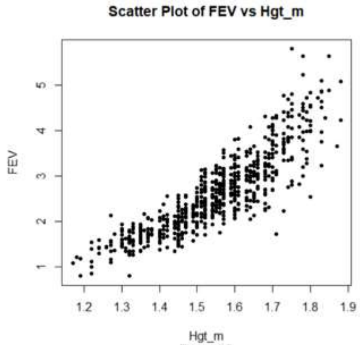


Figure 1.5

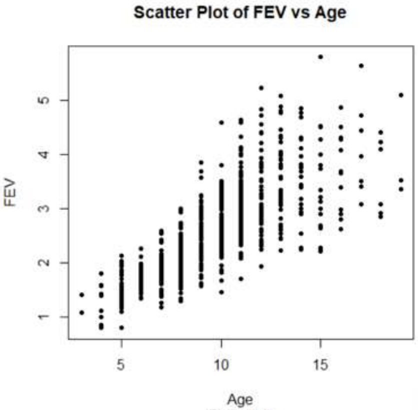


Figure 1.1

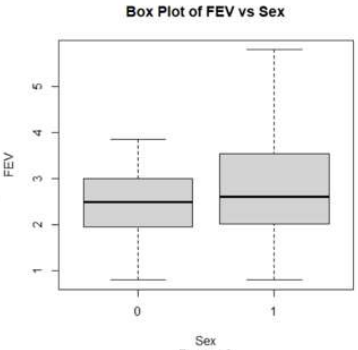


Figure 1.3

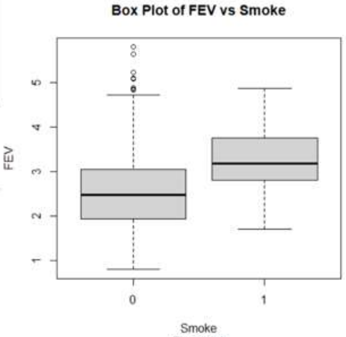


Figure 1.4

Figure 1.1: The scatter plot of FEV vs Age (y vs x_1) shows an increasing trend and positive linear relationship between them.

Figure 1.2: The scatter plot of FEV vs Hgt (y vs x_2) shows an increasing trend but it might not be a straight line-there might be a non-linear relationship between them.

Figure 1.3: The box plot of FEV vs Sex (y vs x_3) shows the median FEV for both male and female are similar (≈ 2.5) with the median for male slightly higher than female. However, the interquartile range and the overall range of FEV for male is wider than female, suggesting that the FEV for male is more spread out. The dataset for female is distributed symmetrically about the center while the distribution of the dataset for male is right-skewed. The median FEV for male is closer to the first quartile than to the third quartile. Neither of the data for male and female suggests any potential outliers as all points are within the range.

Figure 1.4: The box plot of FEV vs Smoke (y vs x_4) shows the median FEV for smoker (≈ 3) is higher than non-smoker (≈ 2.5). The interquartile range and the overall range of FEV for non-smoker is slightly wider than smoker, suggesting that the FEV for non-smoker is more spread out. The dataset for non-smoker is distributed symmetrically about the center while the distribution of the dataset for smoker is a little right-skewed. This might be due to the smaller number of smokers compared to non-smokers (there are 65 smokers and 589 non-smokers). The box plot suggests that there are potential outliers in the data for non-smoker but not for smoker.

Figure 1.5: The scatter plot of FEV vs Hgt_m (y vs x_5) shows an increasing trend but it might not be a straight line-there might be a non-linear relationship between them.

Conclusion: The scatter plots of y vs each quantitative regressor may or may not show linear patterns but the absence of linear patterns in each of them does not imply that fitting a linear model with all the regressors included is incorrect. (i.e. although the scatter plots of y vs x_2 and y vs x_5 indicate no linear relationship between (y and x_2) and (y and x_5) respectively, it does not mean that it's a bad idea to fit a linear model with x_2 and x_5 included. Need to do further analysis on this.

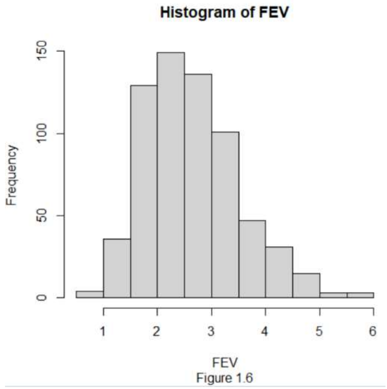


Figure 1.6

Figure 1.6: The response variable y (FEV) possibly follows a right-skewed distribution as shown in the histogram. The distribution is not symmetric, there is a slight deviation from normal as the distribution is right-skewed (right tail is longer than left tail). The QQ plot for this model will be shown in section 2.2 after the model is fitted.

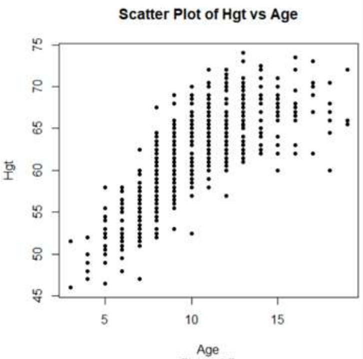


Figure 1.7

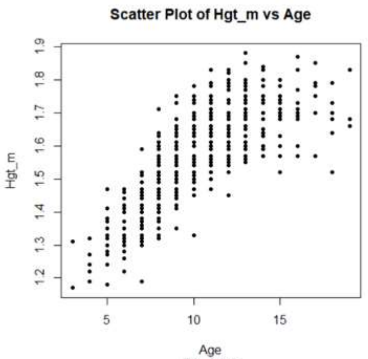


Figure 1.8

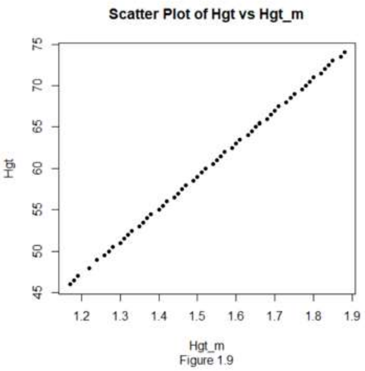


Figure 1.9

Figure 1.7: The scatter plot of Hgt vs Age (x_2 vs x_1) shows an increasing trend but it might not be a straight line-there might be a non-linear relationship between them. The correlation between them is quite high, 0.7919436.

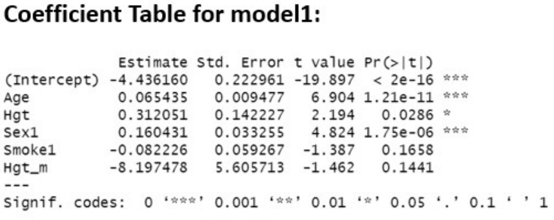
Figure 1.8: The scatter plot of Hgt_m vs Age(x_5 vs x_1) shows an increasing trend but it might not be a straight line-there might be a non-linear relationship between them. The correlation between them is quite high, 0.7917857.

Figure 1.9: The scatter plot of Hgt vs Hgt_m (x_2 vs x_5) shows an increasing trend and strong positive linear relationship between them. This is expected as Hgt and Hgt_m essentially measure the same thing but just with different units. In fact, their correlation is very high, 0.9998013.

Conclusion: From the plots, there are positive relationships/correlations between any 2 quantitative regressors. It is also obvious that the scatter plot between x_2 and x_5 shows a very strong positive linear pattern, but we still cannot conclude that we should remove any of these variables as it is possible that when the full model with all regressors included is fitted, the model does not indicate any serious problem, in that case I would choose to include all the regressors in the model. To find out if any regressors need to be removed, multicollinearity diagnostics will be performed in section 2.1. For now, I choose not to remove any of the regressors yet.

2. The Model

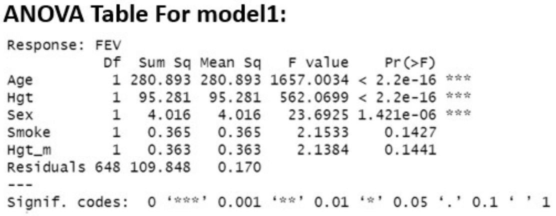
y, x_1, x_2, x_3, x_4, x_5 are defined in section 1. Firstly, I fit the full model (called model1) with all regressors x_1, x_2, x_3, x_4, x_5 included, without any higher order terms or interaction terms as I have no expert advice whether to add interaction term or not. i.e FEV ~ Age + Hgt + Sex + Smoke + Hgt_m.



Coefficient Table for model1:

```
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -4.436160   0.222961 -19.897  < 2e-16 ***
Age          0.065435   0.009477   6.904 1.21e-11 ***
Hgt          0.312051   0.142227   2.194   0.0286 *  
Sex1         0.160431   0.033255   4.824 1.75e-06 ***
Smoke1      -0.082226   0.059267  -1.387   0.1658    
Hgt_m       -8.197478   5.605713  -1.462   0.1441    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 1.1



ANOVA Table For model1:

```
Response: FEV
           Df  Sum Sq Mean Sq   F value    Pr(>F)    
Age         1 280.893 280.893 1657.0034 < 2.2e-16 ***
Hgt         1  95.281  95.281  562.0699 < 2.2e-16 ***
Sex         1   4.016   4.016   23.6925 1.421e-06 ***
Smoke       1   0.365   0.365    2.1533    0.1427    
Hgt_m       1   0.363   0.363    2.1384    0.1441    
Residuals 648 109.848   0.170                        
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 1.2

Hence based on Table 1.1, the fitted equation of **model1**:

$$\hat{y} = -4.436 + 0.0654x_1 + 0.312x_2 + 0.160x_3 – 0.0822x_4 – 8.197x_5$$

From coefficient table in Table 1.1, Age, Hgt, Sex (x_1, x_2, x_3) are significant at level 0.05 since the p values for t-test of the parameters for these 3 regressors are < 0.05; while Smoke and Hgt_m(x_4 and x_5) are insignificant as the p-values for these regressors are > 0.05.

From the ANOVA table for model1 in Table 1.2,

$$F_0 = \frac{(280.893 + 95.281 + 4.016 + 0.365 + 0.363)/5}{109.848/648} = 449.4 > F_{5,\,648}(0.05) = 2.227931$$

Hence, ≥ 1 coefficients for regressors are not 0, the model is significant at 0.05 significance level.

Notice that the standard error of the coefficient estimate for Hgt_m is significantly large compared to others, implying that its variance is high also. This is one of the effects of multicollinearity (large variance of coefficient estimate). This is consistent with my previous analysis that the variables Hgt and Hgt_m are greatly correlated to each other. Therefore, I decided to check for multicollinearity and which regressors cause it before plotting the residual plots to further check for model adequacy.

2.1 Model Adequacy Checking for model1

Multicollinearity Diagnostics - Variance Inflation Factors (VIF)

Before removing Hgt variable:

Age (x_1)	Hgt (x_2)	Hgt_m (x_5)
2.682221	2517.811867	2516.124454

Table 2.1.1 VIFs for FEV data (uncentered)

After removing Hgt variable:

Age (x_1)	Hgt_m (x_5)
2.680424	2.680424

Table 2.1.2 VIFs for FEV data (uncentered)

I used VIF to detect the multicollinearity. Based on Table 2.1.1, the VIF values for Hgt and Hgt_m exceeds 10, which indicates that the regression coefficients for Hgt and Hgt_m are poorly estimated which is also reflected in the large variance in the coefficient of Hgt_m. Hence, this is consistent with what I suspected before, that Hgt and Hgt_m both mean the same thing but just with different unit. So, I decided to remove one of them in my model. I have decided to remove Hgt from the model as this new model will give me a higher R^2_{Adj} compared to the model with Hgt_m removed.

Hence, I refit the model (called model2) with only $x_{1,}$ $x_{3,}$ $x_{4,}$ x_5 (FEV ~ Age + Sex + Smoke + Hgt_m)

The fitted equation of **model2**:

$$\hat{y} = -4.449 + 0.0661x_1 + 0.156x_3 – 0.0894x_4 + 4.0948x_5$$

From coefficient table, Age, Sex, Hgt_m (x_1, $x_{3,}$ x_5) are significant at level 0.05 since the p values for t-test of the parameters for these 3 regressors are < 0.05; while Smoke(x_4) is insignificant as the p-value for its parameter is > 0.05. From summary table, $R^2_{Adj} = 0.7731$, 77.31% of the total variability in response y is explained by the model2. After excluding Hgt variable, multicollinearity is not detected in this model anymore, the VIF values for each regressor in model2 is <10 (Table 2.1.2)

2.2 Model Adequacy Checking for model2

Residual Plots for model2

Plots of Standardized Residuals vs Response and Regressors for model2

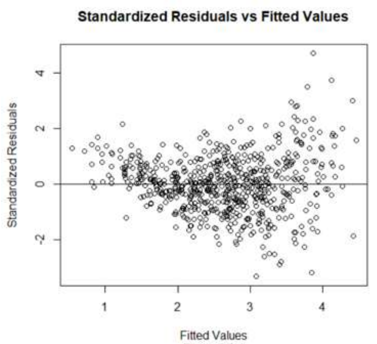


Figure 2.2.1

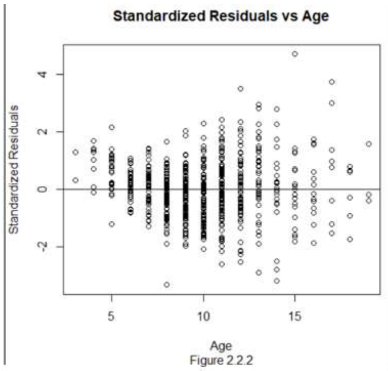


Figure 2.2.2

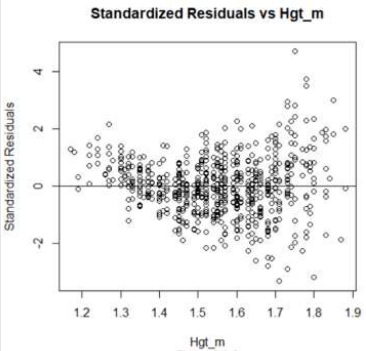


Figure 2.2.3

Figure 2.2.1: The plot of S.R vs fitted values has some dots not within the band from -3 to 3, these large residuals might be the outliers, further tests conducted in Section 2.5. The plot shows a clearly trend of quadratic curve, which indicates nonlinearity. It also shows an outward-opening funnel shape which implies that the variance of the errors is not constant, that is, the variance is an increasing function of y. Hence transformation of data is needed to fix this error and stabilize the variance.

Figure 2.2.2 & Figure 2.2.3: For both plot of S.R vs Age (S.R vs x_1) and S.R vs Hgt_m (S.R vs x_5) , the pattern is similar to the plot of standardized vs fitted values in Figure 2.2.1 where the constant variance assumption and the linearity assumption of the model are violated due to the quadratic curve trend and the funnel shape trend. They also have dots exceeding band from -3 to 3, indicating outliers.

Normal probability Plot of Standardized Residuals for model2

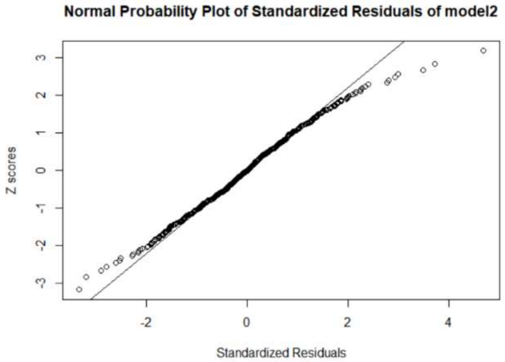


Figure 2.2.4

Figure 2.2.4: The normal probability plot of S.R shows a slight deviation from normal since both the right tail and left tail are thicker than normal. It also indicates that there may be outliers in the data.

Conclusion: Model2 is inadequate from the residual plots, hence a transformation is needed.

2.3 Correct Model Inadequacy for model2

Selecting a Transformation for model2 (The Box-Cox method)

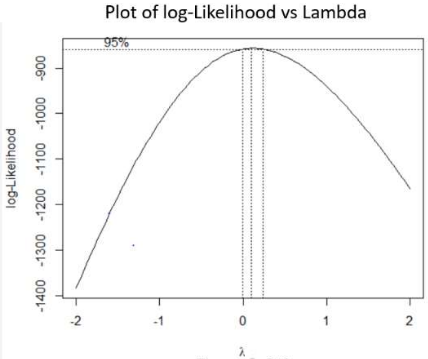


Figure2 .3.1

After performing Box-Cox method to see what type of transformation on y is needed for model2, the plot in Figure 2.3.1 suggests that lambda should be within the interval about 0 to 0.3. For easy interpretation, the lambda = 0 is chosen. Hence the plot proposes that lambda = 0 should be used, which means that a transformation on y: log(y) should be used.

Hence, I fit another new model (called model3), with the response be log(y). The new fitted equation of **model3** is:

$$\widehat{\log(y)} = -1.940 + 0.0236x_1 + 0.0287x_3 - 0.0471x_4 + 1.681x_5$$

From the coefficient table of model3, all the regressors Age, Hgt_m, Sex, Smoke(x_1, x_3, x_4, x_5) are significant at level 0.05 since the p values for t-test of the parameters for these regressors are < 0.05. From summary table, $R^2_{Adj} = 0.8084$, which has improved compared to the R^2_{Adj} in model2. 80.84% of the total variability in response y is explained by the model3.

2.4 Model Adequacy Checking for model3

Now we check for model adequacy again to see if after performing the transformation, the model is adequate or not.

Residual Plots for model3

Plots of Standardized Residuals vs Response and Regressors for model3

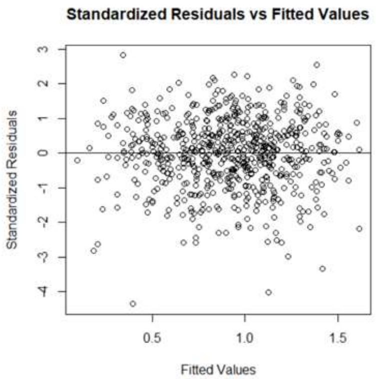


Figure 2.4.1

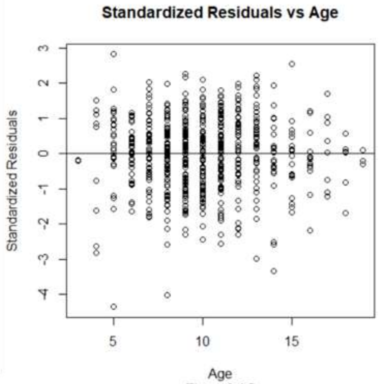


Figure 2.4.2

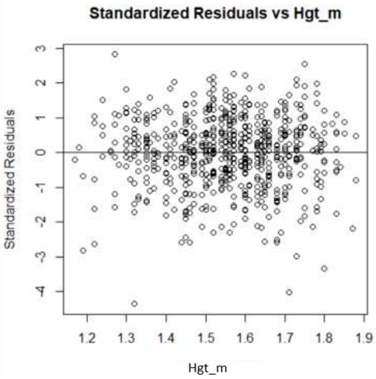


Figure 2.4.3

Figure 2.4.1 to Figure 2.4.3: For all the SR plots, i.e plot of standardized residuals (S.R) vs fitted values, plot of S.R vs Age (S.R vs x_1) and plot of S.R vs Hgt_m (S.R vs x_1), all of them still have some dots not within the band from -3 to 3, these large residuals might be the outliers, further testing are conducted at section 2.5. However, all 3 plots look much better now, there is no more pattern or trend in all the 3 S.R plots, the dots fluctuate randomly around 0 in all residual plots. Hence the model assumptions are not violated.

Normal probability Plot of Standardized Residuals for model3

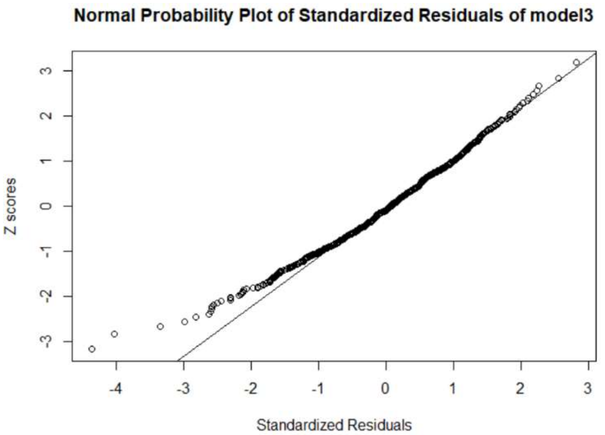


Figure 2.4.4

Figure 2.4.4: The normal probability plot of S.R of model3 only shows a slight deviation from normal as the left tail is slightly thicker than normal. It also indicates possible outliers in the data.

Conclusion: The residual plots shows the model is adequate, neither of the residual plots indicate any problem like misspecification of regressors or inequality of variance, except for some outliers which are shown in the large residuals in the residual plots and the normal probability plot. Hence now we proceed to investigate these possible outliers, i.e. check if they are influential points.

2.5 Detection of Outliers

There are all together 3 possible outliers based on the residual plots (where the $|SR|>3$). The index value of these 3 outliers are 2, 140, 473. To determine whether they are influential points, I perform Cook's Distance. It turns out that none of the D_i is greater than 1. Hence, I conclude that there are no influential points among the outliers. Besides, since we do not know the cause of the outliers, discarding the observation is inappropriate. Now, I can proceed to variable selection.

2.6 Variable Selection (Stepwise Regression Method)

Now that we have an adequate model after performing all the model adequacy checking, so we can proceed to perform computational techniques for variable selection. Here I used Stepwise Regression Method on model3 and the result shows that the current model3 has indeed all the appropriate regressors in it, no removal of regressors are needed.

Conclusion: The final model that I think is the best among all the possible models is model3:

$$\log(FEV) \sim Age + Sex + Smoke + Hgt\_m$$

Fitted equation of model3: $\widehat{\log(y)} = -1.940 + 0.0236x_1 + 0.0287x_3 – 0.0471x_4 + 1.681x_5$

From the summary table, $R^2 = 0.8096$; $R^2_{Adj} = 0.8084$ and MSres = 0.1458. All the regressors are also significant to model3 (from summary table). Hence the goodness of fit is quite good. Overall, I am satisfied with this model3 as the R^2 value and adjusted R^2 value are also large enough, the MSres is small enough and the model is not too complicated.

```
setwd("C:/Users/sr_te/OneDrive/Desktop/ST3131")
library(MASS)

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

#relationship of the response and the regressors
pairs(FEV ~ Age + Hgt + Hgt_m)
plot(Age, FEV, main = "Scatter Plot of FEV vs Age", sub = "Figure 1.1", pch =
20)
plot(Hgt, FEV, main = "Scatter Plot of FEV vs Hgt", sub = "Figure 1.2", pch =
20)
boxplot(FEV ~ Sex, main = "Box Plot of FEV vs Sex", sub = "Figure 1.3")
boxplot(FEV ~ Smoke, main = "Box Plot of FEV vs Smoke", sub = "Figure 1.4")
length(which(data$Smoke == 1))
length(which(data$Smoke == 0))
plot(Hgt_m, FEV, main = "Scatter Plot of FEV vs Hgt_m", sub = "Figure 1.5",
pch = 20)

# Histogram of FEV data
hist(FEV, main = "Histogram of FEV", sub = "Figure 1.6")

#relationship between the regressors
plot(Age, Hgt, main = "Scatter Plot of Hgt vs Age", sub = "Figure 1.7", pch =
20)
plot(Age, Hgt_m, main = "Scatter Plot of Hgt_m vs Age", sub = "Figure 1.8",
pch = 20)
plot(Hgt_m, Hgt, main = "Scatter Plot of Hgt vs Hgt_m", sub = "Figure 1.9",
pch = 20)
cor(Age, Hgt)
cor(Age, Hgt_m)
cor(Hgt_m, Hgt)

########################Fitting model1####################################
model1 = lm(FEV ~ Age + Hgt + Sex + Smoke + Hgt_m)
summary(model1)

anova(model1)

################################Find VIF###################################
x <- cbind(Age, Hgt, Hgt_m)
#then finding the correlation:
x<- cor(x) #this is X'X (correlation form)
C<-solve(x)  #this is (X'X)^(-1) where X'X is in correlation form

VIF <- diag(C) # VIF for Hgt and Hgt_m >10, multicollinearity indicated,
caused by Hgt and Hgt_m
VIF

# Exclude variable Hgt
new_x <- cbind(Age, Hgt_m)
new_x <- cor(new_x) #this is X'X (correlation form)
new_C <- solve(new_x)  #this is (X'X)^(-1) where X'X is in correlation form
```

```
new_VIF <- diag(new_C) # multicollinearity not indicated.
new_VIF

###################Fitting model2######################
model2 <- lm(FEV ~ Age + Sex + Smoke + Hgt_m)
summary(model2)

anova(model2)

##################Model Adequacy Checking for model2######################
#standardized residuals vs Fitted Values:
plot(model2$fitted.values,rstandard(model2), xlab="Fitted Values", ylab=
"Standardized Residuals", main = "Standardized Residuals vs Fitted Values",
sub = "Figure 2.2.1")
abline(h=0)
#standardized residuals vs Age:
plot(Age,rstandard(model2), xlab="Age", ylab= "Standardized Residuals", main =
"Standardized Residuals vs Age", sub = "Figure 2.2.2")
abline(h=0)
#standardized residuals vs Hgt_m:
plot(Hgt_m,rstandard(model2), xlab="Hgt_m", ylab= "Standardized Residuals",
main = "Standardized Residuals vs Hgt_m", sub = "Figure 2.2.3")
abline(h=0)
#Normal probability plot/QQ plot
qqnorm(rstandard(model2),datax = TRUE, ylab = "Standardized Residuals", xlab =
"Z scores", main = "Normal Probability Plot of Standardized Residuals of
model2", sub = "Figure 2.2.4")
qqline(rstandard(model2),datax = TRUE)

########################## Transformation(Box-cox method) ##################
boxcox(model2, lambda=seq(-2, 2, by=0.5), optimize = TRUE, plotit = TRUE)
# lambda should be within the interval from about 0 to 0.3, for easy
interpretation, lambda chosen as 0
# the plot proposes that lambda = 0 should be used, that means a
transformation on
# y: log(y) should be used.

############################Fitting new model3#############################
model3 <- lm(log(FEV) ~ Age + Sex + Smoke + Hgt_m)

summary(model3)

#################Model Adequacy Checking for model3######################
#standardized residuals vs Fitted values:
plot(model3$fitted.values,rstandard(model3), xlab ="Fitted Values", ylab =
"Standardized Residuals",main = "Standardized Residuals vs Fitted Values", sub
= "Figure 2.4.1")
abline(h=0)
#standardized residuals vs Age:
plot(Age,rstandard(model3), xlab ="Age", ylab = "Standardized Residuals", main
= "Standardized Residuals vs Age", sub = "Figure 2.4.2")
abline(h=0)
#standardized residuals vs Hgt_m:
plot(Hgt_m,rstandard(model3), xlab = "Hgt_m", ylab = "Standardized Residuals",
main = "Standardized Residuals vs Hgt_m", sub = "Figure 2.4.3")
abline(h=0)
```

```
#Normal probability plot =  QQ plot
qqnorm(rstandard(model3),datax = TRUE, ylab = "Standardized Residuals", xlab =
"Z scores", main = "Normal Probability Plot of Standardized Residuals of
model3", sub = "Figure 2.4.4")
qqline(rstandard(model3),datax = TRUE)


################# Detect outliers(influential points) ##################
SR <- rstandard(model3)
which(SR < -3)#outliers
which(SR > 3)#outliers
cook_dist <- cooks.distance(model3)
which(cook_dist > 1) # No influential points

######################## Stepwise Selection ####################
sw<-step(model3, direction = c("both"))

summary(sw)

anova(sw)
```