Contents

1. Explore Data 2

1.1. Summary statistics 2

1.2. Scatterplot matrix 3

1.3. Boxplot 4

1.4. Histogram 5

2. Model Building 7

2.1. Standardization 7

2.2. Basic model 8

2.3. Model with interaction term 11

2.4. Transformation on response 13

2.5. Removal of interaction term 16

2.6. Transformation on regressors 18

2.7. Multicollinearity 23

1. Explore Data

1.1. Summary statistics

After loading the data, the summary statistics are shown in Figure 1. For this data, we will be ignoring the variables ID and Hgt. ID does not contribute any significant meaning to the model while Hgt can simply be replaced with Hgt_m. From the table, we can see:

1. The average age of the sample is around 9 years old, with the youngest at 3 and oldest at 19 years old.
2. The height of the sample ranges from 1.17m to 1.88m, with the average around 1.55m.
3. The summary statistics of Smoke and Sex does not provide meaningful information. As such, we will have to visualise our data to extract meaningful information.

```
##        ID             Age              FEV             Hgt
##  Min.   :  201   Min.   : 3.000   Min.   :0.790   Min.   :46.00
##  1st Qu.:15811   1st Qu.: 8.000   1st Qu.:1.982   1st Qu.:57.00
##  Median :36071   Median :10.000   Median :2.550   Median :61.50
##  Mean   :37170   Mean   : 9.931   Mean   :2.637   Mean   :61.14
##  3rd Qu.:53639   3rd Qu.:12.000   3rd Qu.:3.118   3rd Qu.:65.50
##  Max.   :90001   Max.   :19.000   Max.   :5.790   Max.   :74.00
##       Sex             Smoke              Hgt_m
##  Min.   :0.0000   Min.   :0.00000   Min.   :1.170
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:1.450
##  Median :1.0000   Median :0.00000   Median :1.560
##  Mean   :0.5138   Mean   :0.09939   Mean   :1.553
##  3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:1.660
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.880
```

Figure 1: Summary statistics of data

1.2. Scatterplot matrix

From the scatterplot matrix, we can see that all remaining numerical variables are positively correlated. FEV (Force Expiratory Volume) measures the volume of air blown out in a second. As such, it makes sense for:

1. FEV to be strongly positively correlated with Age – The older you are, the larger your lungs are.
2. FEV to be strongly positively correlated with Height – The taller you are, the larger your lungs are.
3. Age to be strongly positively correlated with Height – The older you are, the taller you are.

However, upon careful observation, the relationship is not necessary a straight line. There seems to be a slight curvature in all 3 scatter plot matrices. This will be further explored in figures 20 and 21. A relationship between different variables, also known as multicollinearity, is not ideal. However, due to the constraints and nature of the data, we are not able to control this (it is natural for an older person to be taller).

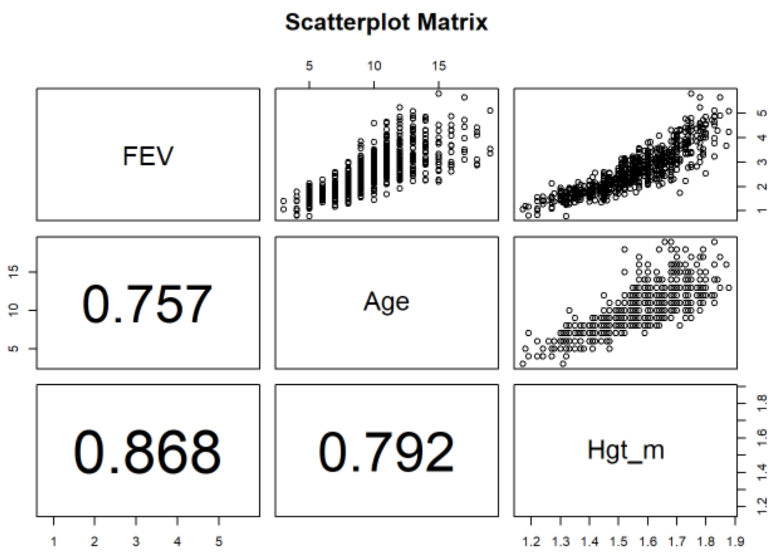

Figure 2: Scatterplot matrix

1.3. Boxplot

The boxplot provides us with a visualisation of the summary statics in 1.1. Summary statistics. Through the boxplots in figure 3, we observe that FEV seem to have many outliers while there is an outlier for age, which should be the person at 19 years old.

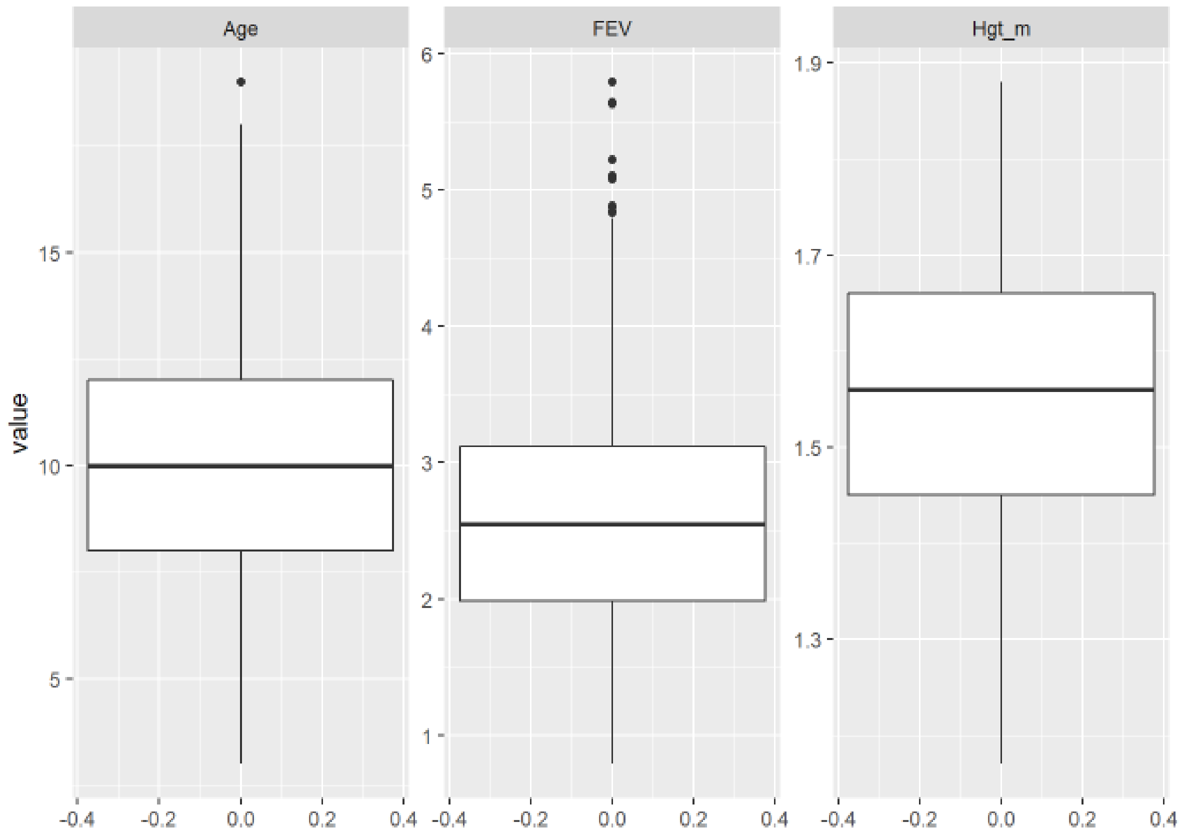

Figure 3: Boxplot

1.4. Histogram

Through the histogram in figure 4, we are also able to visualise the categorical variables – Smoke and Sex which we were previously not able to. We observe that there is quite an equal number of males and females. There are much more non-smokers than smokers, which makes sense as most of the people in the samples are young.

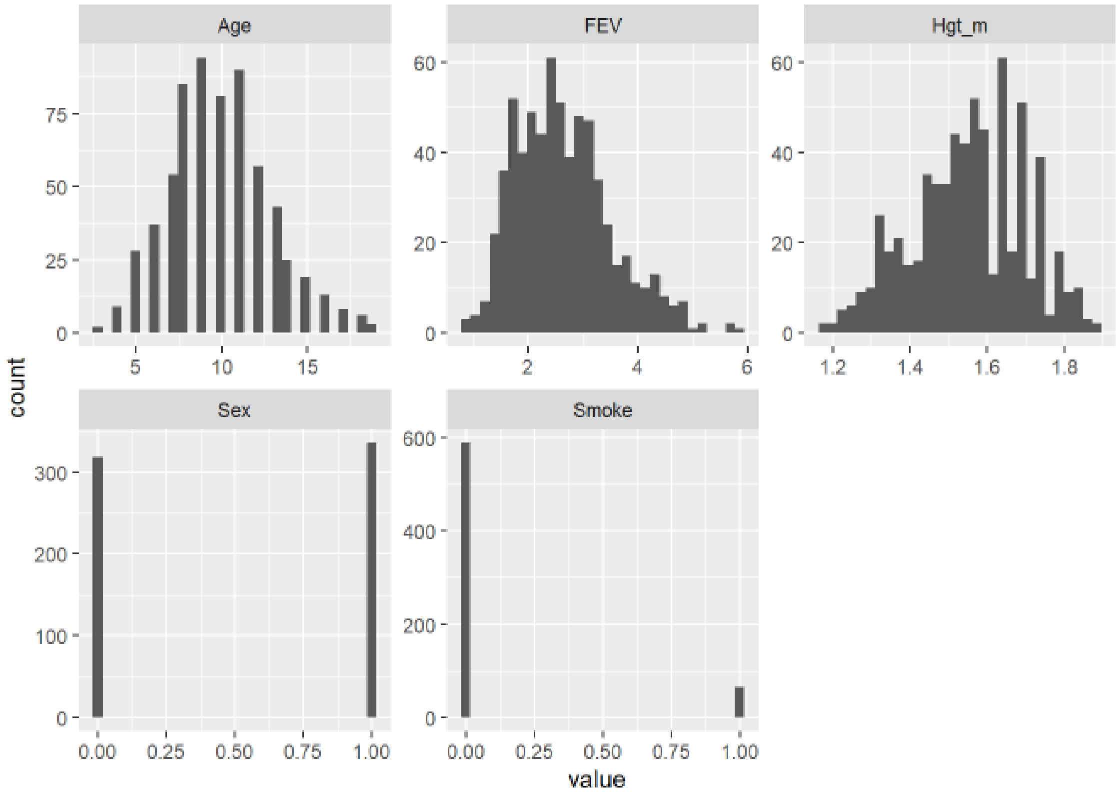

Figure 4: Histogram

We are also able to have a rough gauge of the distribution of our numerical variables – Age, FEV, Height. The histogram seems to show that there is not a huge deviation from a normal distribution. We can try to confirm this with normal probability plots and Shapiro-Wilk test. However as seen in figure 5, there is deviation from normality. Shapiro-Wilk test in figure 6 also returned very small p-values, which indicates that the variables all do not follow a normal distribution.

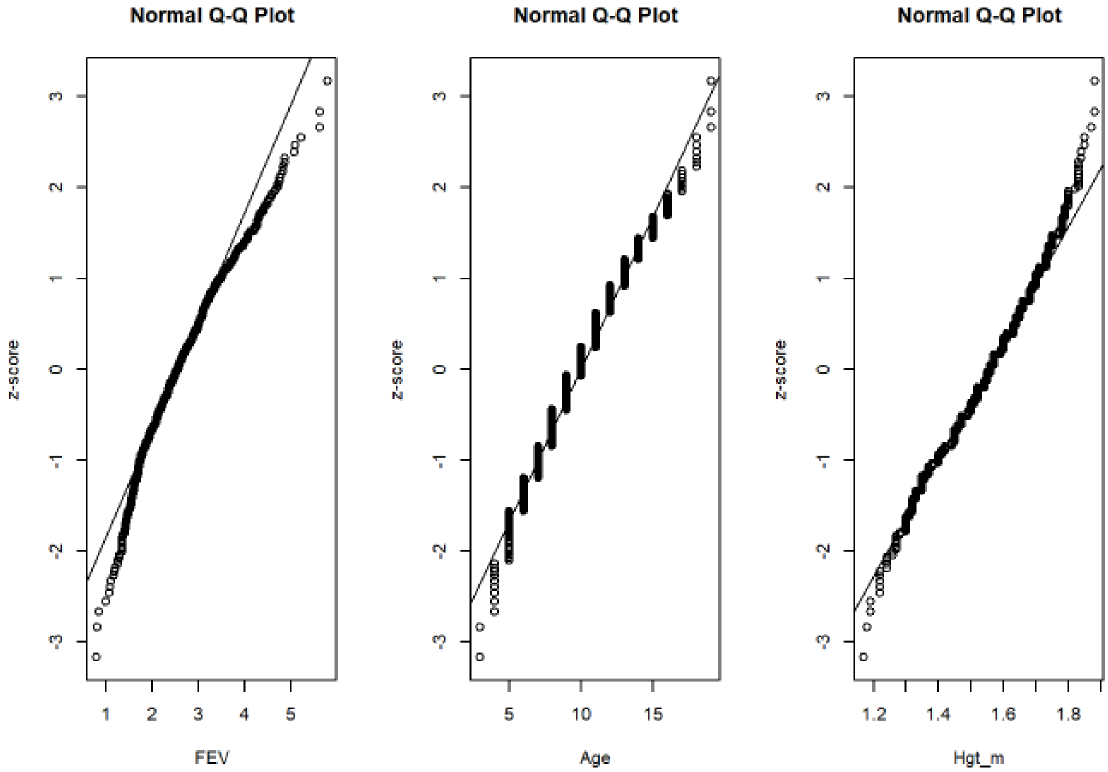

Figure 5: Normal probability test for variables and FEV

```
##
##  Shapiro-Wilk normality test
##
## data:  new_fev_data$FEV
## W = 0.97059, p-value = 3.521e-10
```

```
##
##  Shapiro-Wilk normality test
##
## data:  new_fev_data$Age
## W = 0.97801, p-value = 2.404e-08
```

```
##
##  Shapiro-Wilk normality test
##
## data:  new_fev_data$Hgt_m
## W = 0.98938, p-value = 0.0001134
```

Figure 6: Shapiro-Wilk tests for variables and FEV

2. Model Building

2.1. Standardization

To ensure that the units of all our predictors are the same and to narrow the magnitude of our predictors, we conduct unit normal scaling.

We subtract every value with the mean and divide the result by the standard deviation.

$$\boxed{z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}}$$

Figure 7: Unit normal scaling

2.2. Basic model

The first step would be to build a basic model where all our relevant predictors are included.

1. The response (FEV) is not pre-processed.
2. Numerical predictors (Age and Height) have been standardized.
3. Categorical variables (Sex and Smoke) have been transformed into factors.

Model

$$FEV = 0.195 * Age + 0.593 * Height + 0.156 * Sex(type = 1) - 0.0894 * Smoke\ (type = 1) + 2.57$$

Model Interpretation: As every other predictor remains fixed,

1. As age increases by 1 unit, the mean FEV increases by 0.195
2. As height increases by 1 unit, the mean FEV increases by 0.593
3. Males have a higher mean FEV as compared to females by 0.156
4. Smokers have a lower mean FEV as compared to non-smokers by 0.0894

The interpretation of age and height is in line with what was mentioned in 1.2. Scatterplot matrix. According to the model, males have a larger lung capacity than females. Smokers have a lower lung capacity than non-smokers.

```
Call:
lm(formula = new_fev_data$FEV ~ Age_standardized + Hgt_m_standardized +
    sex_factor + smoke_factor)

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)         2.56621    0.02459 104.376  < 2e-16 ***
Age_standardized    0.19521    0.02806   6.956 8.57e-12 ***
Hgt_m_standardized  0.59258    0.02718  21.799  < 2e-16 ***
sex_factor1         0.15568    0.03328   4.678 3.53e-06 ***
smoke_factor1      -0.08938    0.05935  -1.506    0.133
---
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

Analysis of Variance Table

Response: new_fev_data$FEV
                    Df  Sum Sq Mean Sq   F value    Pr(>F)
Age_standardized     1 280.893 280.893 1647.3231 < 2.2e-16 ***
Hgt_m_standardized   1  94.865  94.865  556.3428 < 2.2e-16 ***
sex_factor           1   3.958   3.958   23.2108 1.809e-06 ***
smoke_factor         1   0.387   0.387    2.2677    0.1326
Residuals          649 110.664   0.171
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 8: Summary and ANOVA (Basic model)

Model Adequacy

From figure 8, the p-value of t-test of smoke is 0.133. This suggests that the predictor smoke is not significant in the basic model. Likewise, with the ANOVA table, given all other predictors that are present, the p-value of f-test of smoke is 0.1326, which also signifies that it is not relevant in basic model. Adjusted R-squared value is 0.7731, which is not bad but is also not the best. We should seek to improve this.

Other than the numerical values, we should also analyse the residuals. From figure 9, we observe that there is a slight quadratic shape in the graph. This is not ideal as there should be no relationship between the fitted values and the standardized residuals. We can also observe a slight quadratic shape for residuals against the numerical predictors. The normal probability plot also tells us that the residuals have a deviation from normality. Both the tails are heavier than normal. As such, we should improve our basic model.

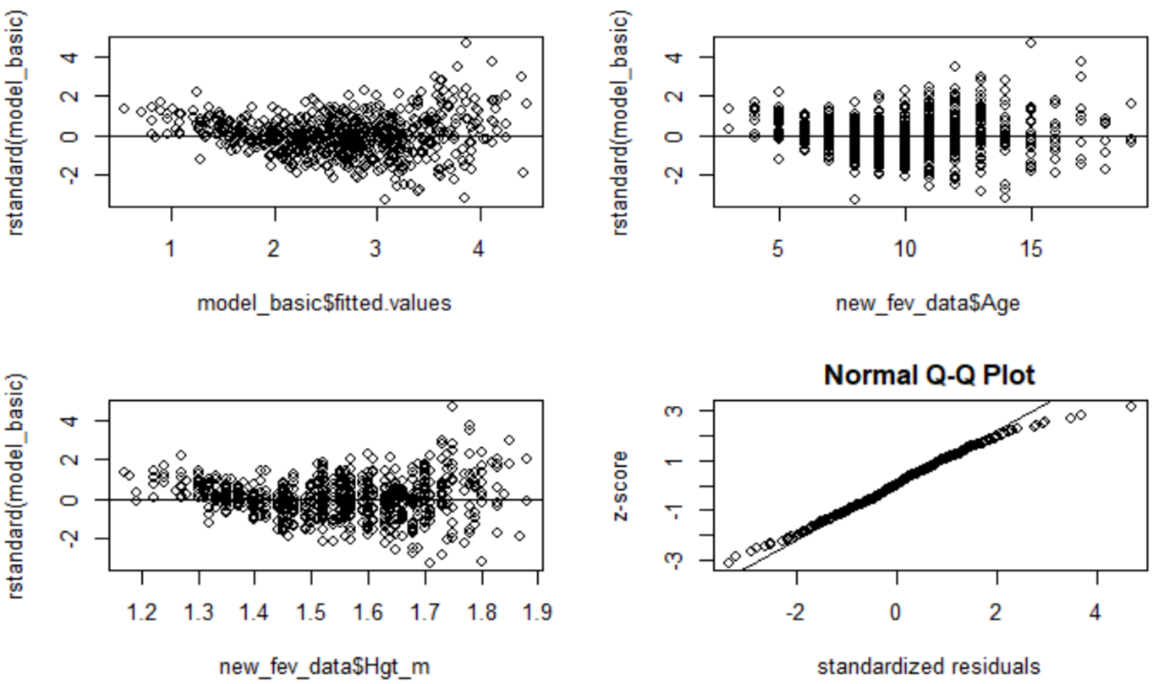

Figure 9: Residual plots (Basic model)

2.3. Model with interaction term

Since age is naturally correlated with height, we can include an interaction term by multiplying height with age. The rest of the predictors stays the same as the basic model.

<u>Model</u>

$$FEV = 0.158 * Age + 0.657 * Height + 0.0991 * Sex(type = 1) - 0.149 * Smoke\ (type = 1) + 0.119 * Age * Height + 2.51$$

<u>Model Interpretation: As every other predictor remains fixed,</u>

5. As age increases by 1 unit, the mean FEV increases by 0.158 + 0.119*Height
6. As height increases by 1 unit, the mean FEV increases by 0.657 + 0.119*Age
7. Males have a higher mean FEV as compared to females by 0.0991
8. Smokers have a lower mean FEV as compared to non-smokers by 0.149

The interpretation of the variables is similar to that of the basic model.

```
Call:
lm(formula = new_fev_data$FEV ~ Age_standardized + Hgt_m_standardized + 
    sex_factor + smoke_factor + Age_standardized * Hgt_m_standardized)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.58596 -0.21994 -0.00202  0.22055  1.74034 

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)                          2.50707    0.02468 101.592  < 2e-16 ***
Age_standardized                     0.15777    0.02725   5.790 1.10e-08 ***
Hgt_m_standardized                   0.65719    0.02726  24.112  < 2e-16 ***
sex_factor1                          0.09914    0.03262   3.039  0.00247 ** 
smoke_factor1                       -0.14867    0.05724  -2.597  0.00961 ** 
Age_standardized:Hgt_m_standardized  0.11901    0.01512   7.873 1.46e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3948 on 648 degrees of freedom
Multiple R-squared:  0.7942,	Adjusted R-squared:  0.7926 
F-statistic: 500.1 on 5 and 648 DF,  p-value: < 2.2e-16

Analysis of Variance Table

Response: new_fev_data$FEV
                                     Df  Sum Sq Mean Sq   F value    Pr(>F)    
Age_standardized                      1 280.893 280.893 1802.1247 < 2.2e-16 ***
Hgt_m_standardized                    1  94.865  94.865  608.6233 < 2.2e-16 ***
sex_factor                            1   3.958   3.958   25.3920 6.076e-07 ***
smoke_factor                          1   0.387   0.387    2.4808    0.1157    
Age_standardized:Hgt_m_standardized   1   9.662   9.662   61.9876 1.458e-14 ***
Residuals                           648 101.002   0.156                        
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 10: Summary and ANOVA (Interaction term model)

Model Adequacy

From figure 10, the p-value of t-test of smoke is 0.1157. This suggests that the predictor smoke is still not significant even after inclusion of an interaction term. Adjusted R-squared value improved slightly from 0.7731 to 0.7926. We should continue to improve this.

From figure 11, we observe that there is a funnel shape in the graph. The variance of the residuals is increasing. This is not ideal as well as one of the basic assumptions of a linear mode is that the variance of the errors should be constant. We can also observe a funnel shape for residuals against the numerical predictors. The normal probability plot also tells us that the residuals have a deviation from normality. Both the tails are heavier than normal. As such, we should continue to improve the model.

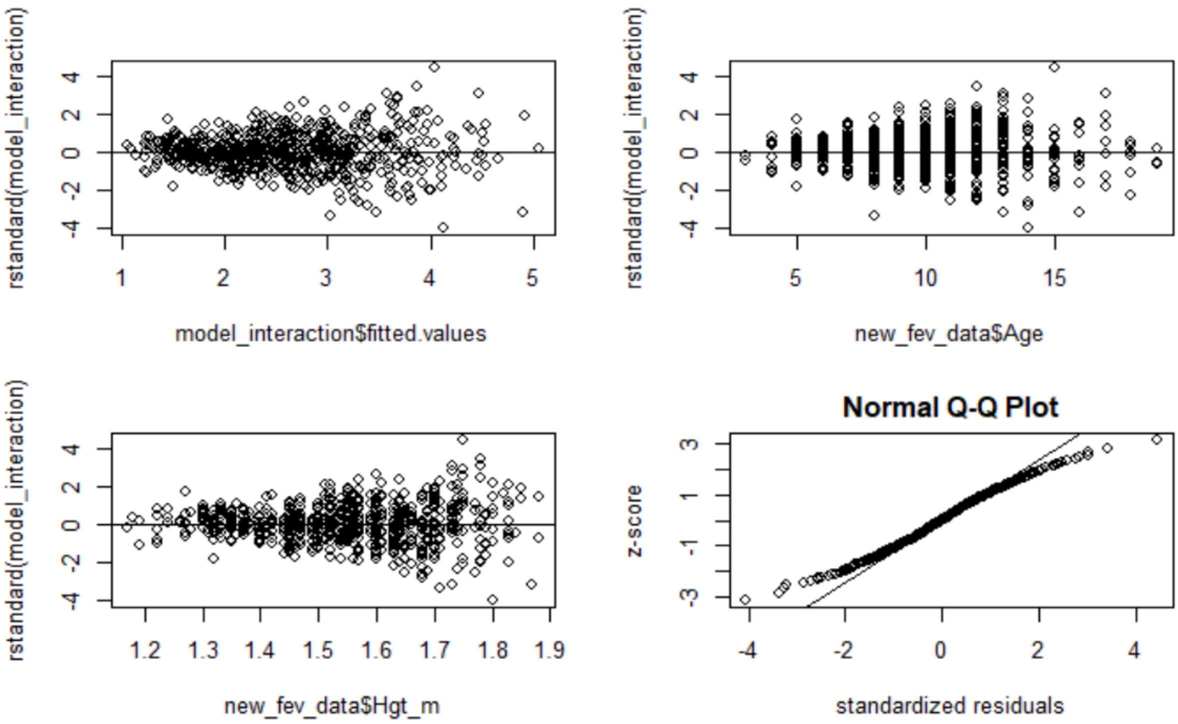

Figure 11: Residuals plots (Interaction term model)

2.4. Transformation on response

Since there is a non-linear relationship on the response which violates the linearity assumption of linear regression, we aim to transform the response using the box-cox method. According to figure 12, we should use log transformation since lambda = 0.

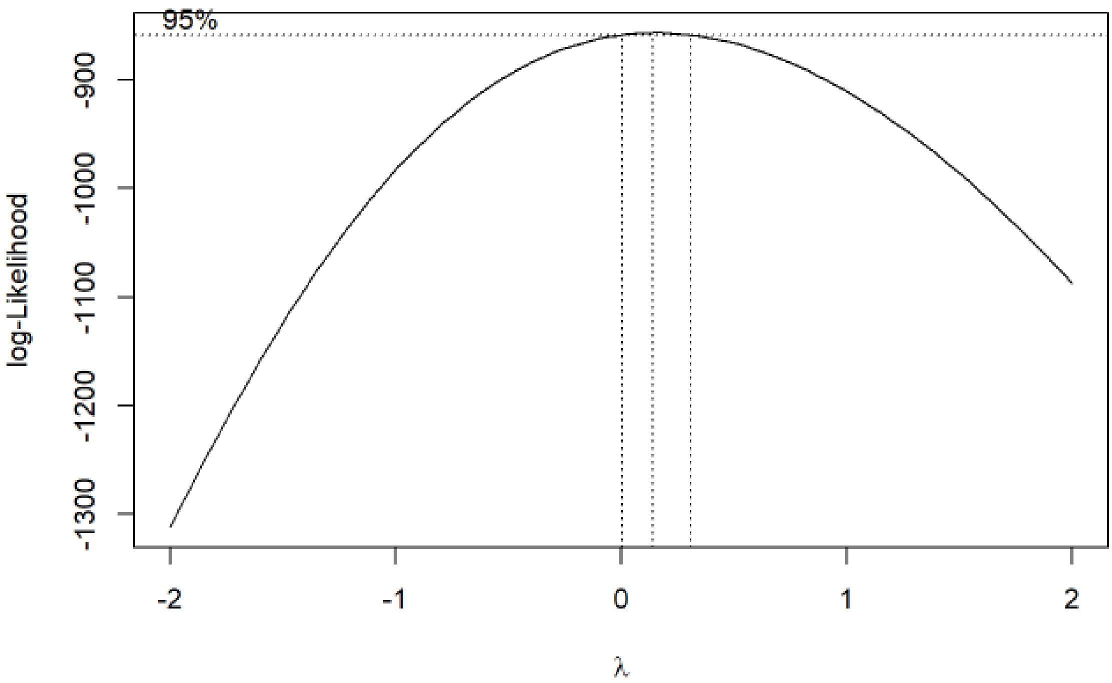

Figure 12: Box-Cox for transformation of response

Model

$$\log(FEV) = 0.0704 * Age + 0.242 * Height + 0.0297 * Sex(type = 1) - 0.0461$$
$$* Smoke\ (type = 1) - 0.00196 * Age * Height + 0.907$$

Model Interpretation: As every other predictor remains fixed,

1. As age increases by 1 unit, the mean of log(FEV) increases by 0.0704 - 0.00196*Height
2. As height increases by 1 unit, the mean of log(FEV) increases by 0.242 - 0.00196*Age
3. Males have a higher mean log(FEV) as compared to females by 0.0297
4. Smokers have a lower mean log(FEV) as compared to non-smokers by 0.0461

The interpretation of median of FEV will only be done for the final model. This model is still inadequate.

```
Call:
lm(formula = log(new_fev_data$FEV) ~ Age_standardized + Hgt_m_standardized +
    sex_factor + smoke_factor + Age_standardized * Hgt_m_standardized)

Residuals:
     Min       1Q   Median       3Q      Max
-0.62944 -0.08608  0.01087  0.09205  0.41287

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                          0.906557   0.009122  99.377  < 2e-16 ***
Age_standardized                     0.070412   0.010073   6.990 6.85e-12 ***
Hgt_m_standardized                   0.242266   0.010075  24.046  < 2e-16 ***
sex_factor1                          0.029667   0.012059   2.460   0.0141 *
smoke_factor1                       -0.046079   0.021160  -2.178   0.0298 *
Age_standardized:Hgt_m_standardized -0.001961   0.005588  -0.351   0.7257
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1459 on 648 degrees of freedom
Multiple R-squared:  0.8096,    Adjusted R-squared:  0.8081
F-statistic: 551.1 on 5 and 648 DF,  p-value: < 2.2e-16

Analysis of Variance Table

Response: log(new_fev_data$FEV)
                                     Df Sum Sq Mean Sq   F value    Pr(>F)
Age_standardized                      1 43.192  43.192 2027.8505 < 2.2e-16 ***
Hgt_m_standardized                    1 15.237  15.237  715.3642 < 2.2e-16 ***
sex_factor                            1  0.148   0.148    6.9501  0.008582 **
smoke_factor                          1  0.107   0.107    5.0323  0.025216 *
Age_standardized:Hgt_m_standardized   1  0.003   0.003    0.1232  0.725701
Residuals                           648 13.802   0.021
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 13: Summary and ANOVA (Transformation on response)

Model Adequacy

From figure 13, the p-value for t-test of smoke is 0.0298. Upon transforming the response, our smoke predictor is now significant to the model. However, our interaction became insignificant with a p-value for t-test of 0.7257. Adjusted R-squared value further improved from 0.7926 to 0.8081. However, this model is still not finalised as we have to remove the interaction term.

From figure 14, we observe that there the residuals do not seem to have any funnel or quadratic shape anymore. It is now randomly scattered around 0. However, the standardized residuals range from -4 to 3. This is not ideal as this would mean that the residuals have a large variance. We also observe no clear relationship in the plot of residuals against the numerical predictors. The normal probability plot tells us that the residuals have a slight deviation from normality. The left tail is heavier than normal.

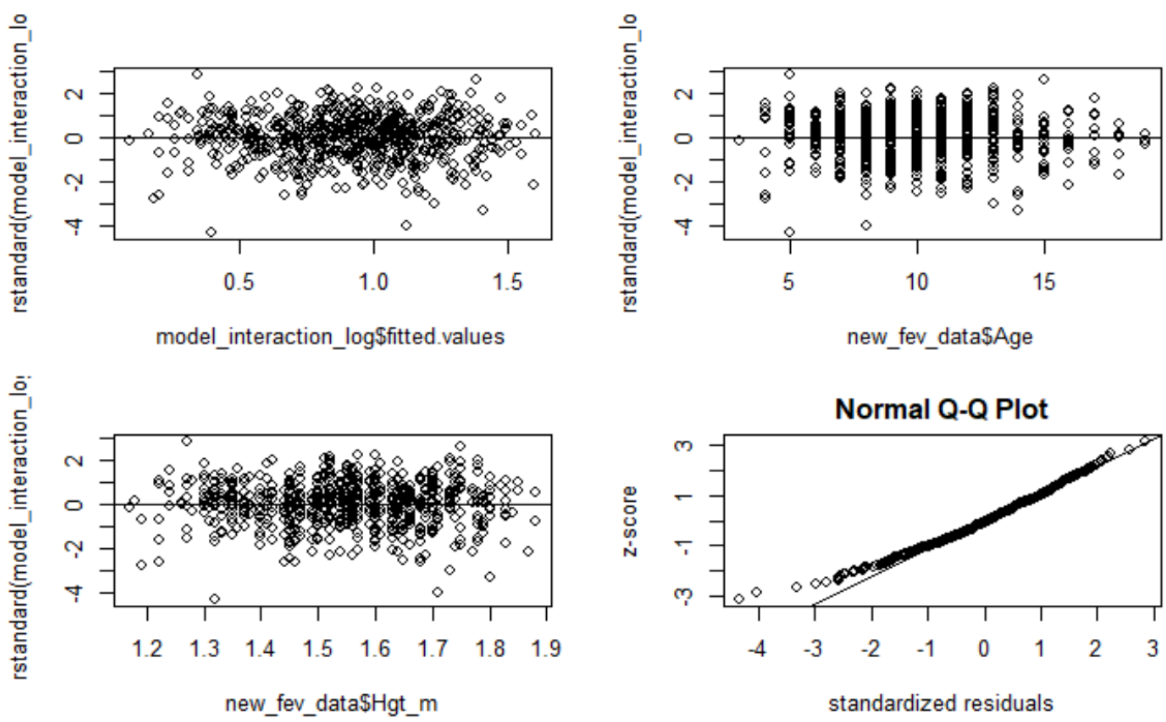

Figure 14: Residuals plot (Transformation on response)

2.5. Removal of interaction term

As mentioned in 2.3. Model with interaction term, the interaction term is insignificant to the model. Hence, we will be removing it.

<u>Model</u>

$\log(FEV) = \ 0.0698 * Age + 0.243 * Height + 0.0287 * Sex(type = 1) - 0.0471 *$

$Smoke\ (type = 1) + 0.906$

<u>Model Interpretation: As every other predictor remains fixed,</u>

1. As age increases by 1 unit, the mean of log(FEV) increases by 0.0698
2. As height increases by 1 unit, the mean of log(FEV) increases by 0.243
3. Males have a higher mean log(FEV) as compared to females by 0.0287
4. Smokers have a lower mean log(FEV) as compared to non-smokers by 0.0471

```
Call:
lm(formula = log(new_fev_data$FEV) ~ Age_standardized + Hgt_m_standardized +
    sex_factor + smoke_factor)

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)         0.905582   0.008684 104.286  < 2e-16 ***
Age_standardized    0.069795   0.009912   7.041 4.86e-12 ***
Hgt_m_standardized  0.243331   0.009601  25.344  < 2e-16 ***
sex_factor1         0.028735   0.011755   2.445   0.0148 *
smoke_factor1      -0.047056   0.020962  -2.245   0.0251 *
---
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

Analysis of Variance Table

Response: log(new_fev_data$FEV)
                    Df Sum Sq Mean Sq  F value    Pr(>F)
Age_standardized     1 43.192  43.192 2030.5938 < 2.2e-16 ***
Hgt_m_standardized   1 15.237  15.237  716.3319 < 2.2e-16 ***
sex_factor           1  0.148   0.148    6.9595  0.008537 **
smoke_factor         1  0.107   0.107    5.0391  0.025117 *
Residuals          649 13.805   0.021
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 15: Summary and ANOVA (Transformed model without interaction term)

Model Adequacy

From figure 15, the p-value for t-test of all predictors is less than 0.05. All the predictors are now significant. This is also in line with the ANOVA table which provide p-value for f-test. Adjusted R-squared value further improved slightly from 0.8081 to 0.8084.

From figure 16, we observe that there the residuals were not affected upon removal of interaction term. It is still randomly scattered around 0. However, the standardized residuals still range from -4 to 3. This is not ideal as this would mean that the residuals have a large variance. We also observe no clear relationship in the plot of residuals against the numerical predictors. The normal probability plot tells us that the residuals have a slight deviation from normality. The left tail is heavier than normal.

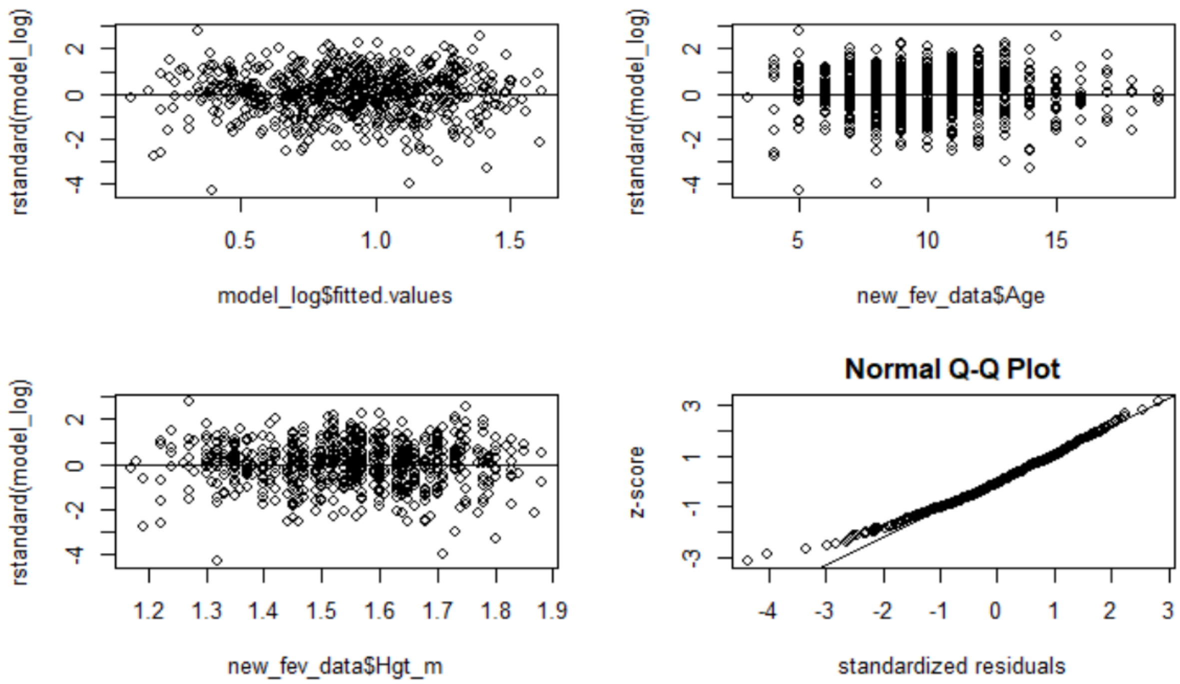

Figure 16: Residual plots (Transformed model without interaction term)

2.6. Transformation on regressors

One final improvement to the model is to apply transformation on the predictors. For this purpose, we will be using the Box and Tidwell method. The Box and Tidwell method requires all regressor to be strictly positive. I also removed the standardization procedure as it provides lambda which are not feasible (lambda of height (standardized) was 4). From figure 17, we can observe that the proposed lambda of Age is 1 and the proposed of lambda for Height is 0.5. Hence, we will square root Height and not transform Age.

```
      MLE of lambda Score Statistic (z) Pr(>|z|)
Age         1.03532              0.0777   0.9381
Hgt_m       0.58256             -0.7945   0.4269

iterations =  3
```

Figure 17: Box-Tidwell method

Model

$$\log(FEV) = 0.0234 * Age + 4.16 * \sqrt{Height} + 0.0322 * Sex(type = 1) - 0.0443 * Smoke\ (type = 1) - 4.50$$

Model Interpretation: As every other predictor remains fixed,

1. As age increases by 1 unit, the median of FEV increases by $e^{0.0234}$
2. At a fixed height H, as height increases by 1 unit, the median of FEV increases by $e^{4.16*(\sqrt{H+1}-\sqrt{H})}$

> Explanation:
>
> $$\log(FEV) = \cdots + 4.16 * \sqrt{Height}\ , where \ldots is\ the\ rest\ of\ the\ equation$$
>
> $$FEV = e^{\ldots + 4.16*\sqrt{H}} \rightarrow Equation\ 1$$
>
> $$FEV = e^{\ldots + 4.16*\sqrt{H+1}} \rightarrow Equation\ 2$$
>
> $$\frac{e^{\ldots + 4.16*\sqrt{H+1}}}{e^{\ldots + 4.16*\sqrt{H}}} \rightarrow Taking\ equation\ 2\ over\ 1, which\ evaluates\ to:$$
>
> $$e^{4.16*(\sqrt{H+1}-\sqrt{H})}$$

3. The median of FEV of males is expected to change by $e^{0.0322}$ times the median of FEV of females.
4. The median of FEV of smokers is expected to change by $e^{-0.0443}$ times the median of FEV of non-smokers.

Since FEV does not follow a normal distribution, log transformation only preserves the median and not the mean. Hence, we are unable to get conclusion on the mean of FEV without other tools.

```
Call:
lm(formula = log(FEV) ~ Age + I(sqrt(Hgt_m)) + sex_factor + smoke_factor,
    data = new_fev_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.62503 -0.08593  0.01108  0.09401  0.42124

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -4.503788   0.178180 -25.277  < 2e-16 ***
Age             0.023438   0.003358   6.980 7.31e-12 ***
I(sqrt(Hgt_m))  4.156890   0.163809  25.376  < 2e-16 ***
sex_factor1     0.032197   0.011718   2.748  0.00617 **
smoke_factor1  -0.044293   0.020954  -2.114  0.03492 *
---
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:  0.8098,	Adjusted R-squared:  0.8086
F-statistic: 690.8 on 4 and 649 DF,  p-value: < 2.2e-16

Analysis of Variance Table

Response: log(FEV)
                Df Sum Sq Mean Sq   F value  Pr(>F)
Age              1 43.192  43.192 2033.2110 < 2e-16 ***
I(sqrt(Hgt_m))   1 15.232  15.232  717.0482 < 2e-16 ***
sex_factor       1  0.182   0.182    8.5896 0.00350 **
smoke_factor     1  0.095   0.095    4.4680 0.03492 *
Residuals      649 13.787   0.021
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 18: Summary and ANOVA (Transformation on predictors)

Model Adequacy

From figure 18, the p-value for t-test of all predictors is less than 0.05. All the predictors are significant. This is also in line with the ANOVA table which provide p-value for f-test. Adjusted R-squared value further improved slightly again from 0.8084 to 0.8086. The model can explain 80.86% of the variation in the data.

From figure 19, we observe that all the residuals against fitted values and predictors is randomly scattered around 0. The standardized residuals still range from -4 to 3. However, we observe that a huge proportion of the residuals actually lie between -2 to 2. Those minority of the points that lie outside this range could be outliers, but we are unable to confirm this without understanding the origin of the data. The normal probability plot tells us that the residuals have a slight deviation from normality. The left tail is slightly heavier than normal.

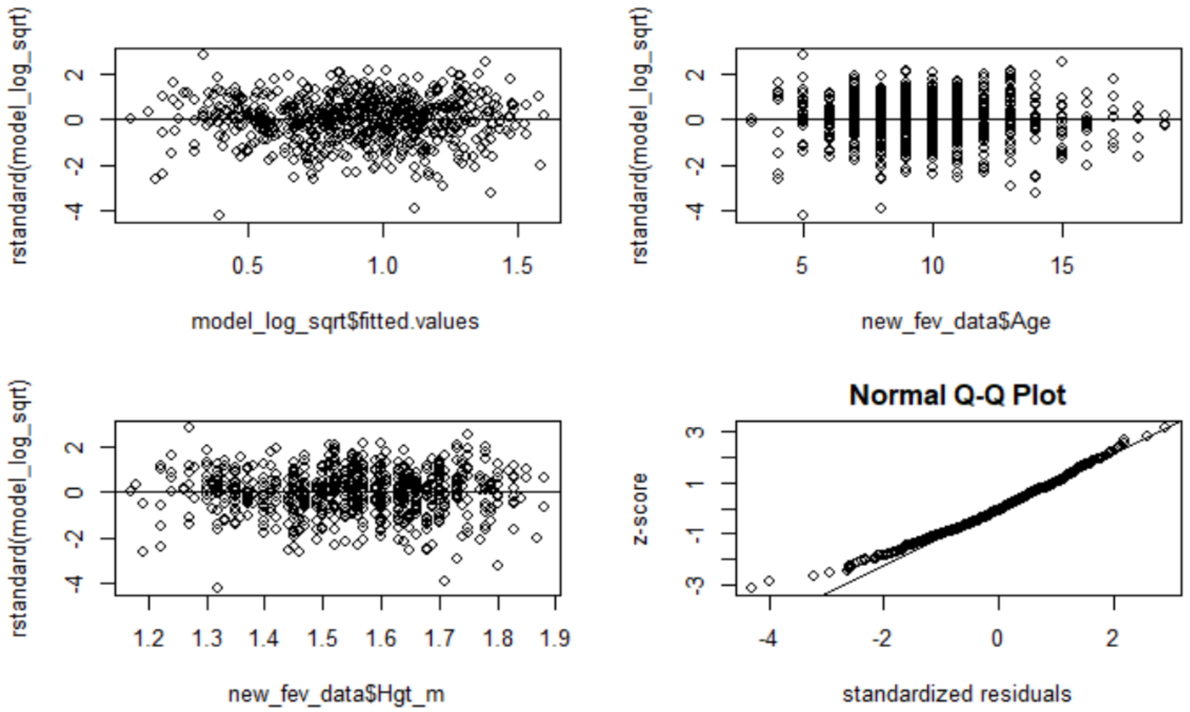

Figure 19: Residual plots for model (Transformation on predictors)

Through figure 20, we can observe that before the transformation, there seem to be a curvature in the graph of FEV against Age in the graph on the left. However, after doing a log transform on FEV as suggested by the Box-Cox method, the new plot on the right seems to be linear. Similarly, through figure 21, we can observe that before the transformation, there is also a curvature in the graph of FEV against Height on the graph on the left. However after doing a log transform on FEV and square root of Height as suggested by the Box-Tidwell method, the new plot on the right seems to be linear.

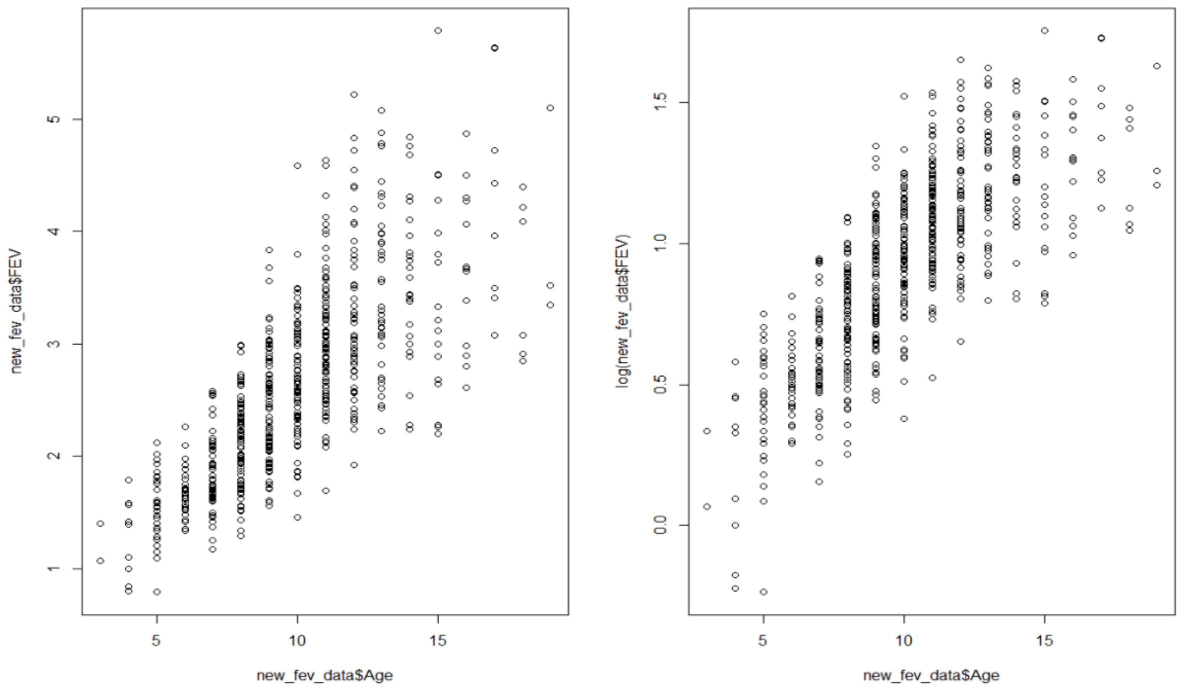

Figure 20: Before and after transformation on FEV

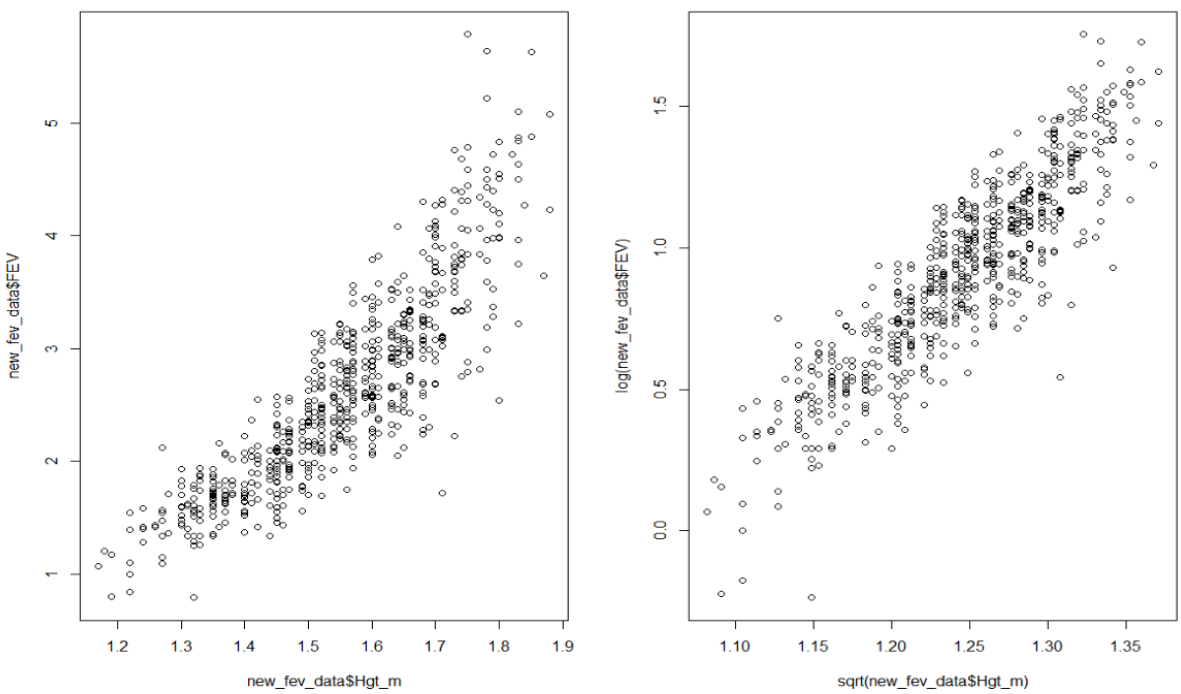

Figure 21: Before and after transformation on FEV and Height

2.7. Multicollinearity

To check for the presence of multicollinearity in the finalised model, we deploy 3 methods

1. Correlation Matrix

The correlation matrix is only produced for numerical regressors. The correlation between age and square root of height is 0.79, which is rather high.

	Age	Height_m_sqrt
Age	1.0000000	0.7925206
Height_m_sqrt	0.7925206	1.0000000

Figure 22: Correlation matrix

2. Variance inflation factor

Despite a high correlation number, the variance inflation factor shows a low value. 2.689 is less than 5, which shows that multicollinearity should not be an issue for the final model created.

Age	Height_m_sqrt
2.688815	2.688815

Figure 23: Variance inflation factor

3. Condition number

The condition number is found to be 8.636, which is much lower than 100. This also supports the view that multicollinearity should not be a serious problem in our model.