ST3131 Regression Analysis

Assignment on *FEV* Data

Name : **Chong Zhen Jie**

Student Number : **A0201613Y**

I. Introduction

This project aims to understand how the Forced Expiratory Volume (*FEV*) is affected by other variables. Data containing the *FEV* measurements of children ages 3 to 19 are used. The response variable of interest is *FEV*. In particular, I attempt to propose a final model to predict *FEV* based on regressors such as age, height, sex, and smoking status.

I first fit a preliminary model, **Model 1**: *FEV ~ Hgt + Hgt_m + Age + Sex + Smoke* .

The data given contains two regressors, height in inches and height in meters. Including both regressors in the model will not be meaningful because both regressors are clearly linearly dependent. Both indicate the same values of height, but in different units of measurement. Hence, I choose to redefine the model, eliminating the regressor *Hgt_m* since it is more realistic to interpret one unit of height in inches than in metres.

The redefined model is **Model 2**: *FEV ~ Hgt + Age + Sex + Smoke* .

From the summary tables in Appendix A, **Model 1** has R^2 = 0.7762 while **Model 2** has a similar R^2 = 0.7754. There may be slight differences in estimations due to rounding error in the unit conversions between inches and metres for height, but the impact should be minimal on the regression. It can be seen that the coefficients of both models for the intercept and all regressors except height are similar.

II. Residual Analysis of Model 2 and Model 3

It should be noted that *FEV* only takes positive values. *FEV* ranges from 0.79 to 5.79. If I were to model *FEV*, the prediction intervals could result in negative values, which would not be meaningful. Hence, a transformation on *FEV* can be considered.

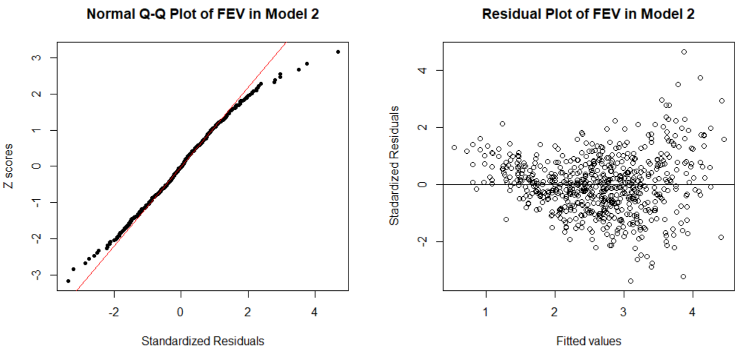

Fig. 1 Residual analysis on FEV in Model 2

This is substantiated from the Q-Q plot of *FEV* in Fig. 1, suggesting a right-skewed distribution which violates the normality assumption. This indicates the possibility of outliers. Another motivation can be seen from the residual plot of *FEV* in Fig. 1, hinting an outward-opening funnel pattern. This implies that the variance is an increasing function of *FEV*, violating the constant variance assumption.

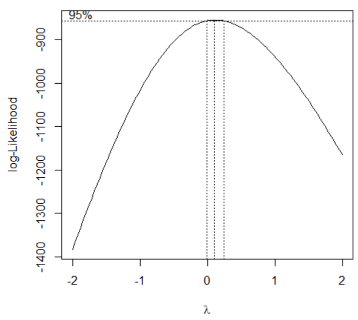

Fig. 2 Box Cox plot of FEV in Model 2

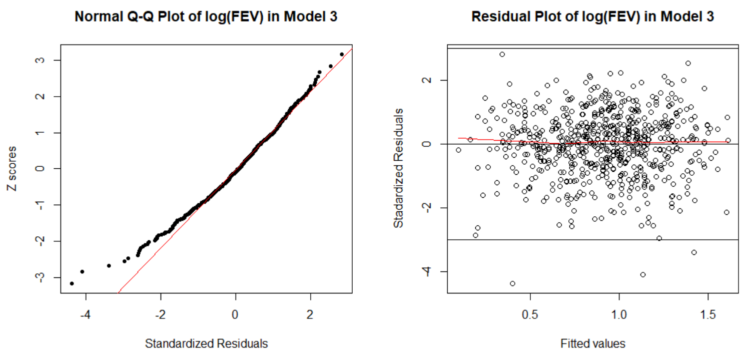

Fig. 3 Residual analysis on log(FEV) in Model 3

To correct the model inadequacy, a transformation on *FEV* can be performed using the Box-Cox Method. The Box-Cox plot is given in Fig. 2, which suggests $\lambda = 0.1$. For the ease of interpretation, I will choose $\lambda = 0$ which is in the 95% interval of λ. Hence, a transformation on the response *log(FEV)* should be used. A log transformation also ensures the prediction of *FEV* is positive, making the estimations more meaningful.

From the summary tables in Appendix A, **Model 2** has R^2 = 0.7745. The new model fitted with the response *log(FEV)*, called **Model 3**, has a better R^2 = 0.8096. From Fig. 3, normality assumption is largely satisfied because the points in the Q-Q plot mostly lie in a straight line. Independent errors and constant variance assumptions are also largely satisfied. For a large sample size of 654, the points in the residual plot are random around 0 in a horizontal band, ranging from -3 to 3. However, it should be noted that there are three large residuals, which possibly caused the Q-Q plot to slightly deviate due to the possibility of outliers. Testing for influential points should be done when proposing the final model.

III. Introduction of Interaction Terms

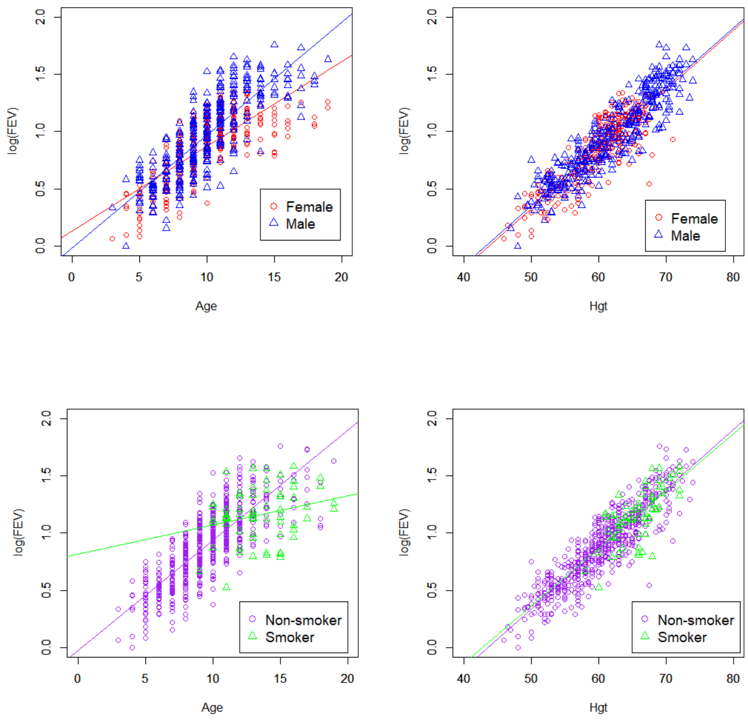

Fig. 4 Scatter plots based on the two categorical variables, Sex and Smoke

From Fig. 4, the effect of *Age* on *log(FEV)* changes with *Sex* and *Smoke* because the slopes of the regression lines for males and non-smokers are larger. However, the slopes of the regression lines are approximately the same for the *Hgt* on *log(FEV)*. This is meaningful because it is assumed that at every age, males tend to have greater lung volume due to biological reasons, and that non-smokers are generally healthier and thus better lung capacities, both leading to higher *log(FEV).* Hence, the following two interaction terms can be considered, one between *Age* and *Sex* and another between *Age* and *Smoke*.

IV. Residual Analysis of Model 4

The refitted model is **Model 4**: *log(FEV) ~ Hgt + Age + Sex + Smoke + Age*Sex + Age*Smoke* .

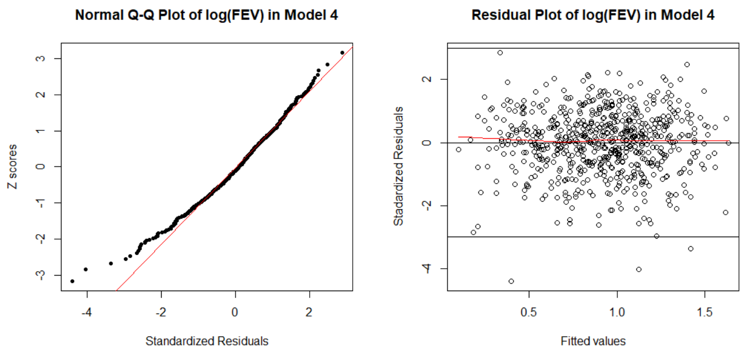

Fig. 5 Residual analysis on log(FEV) in Model 4

From Fig. 5, the Q-Q plot and residual plot of **Model 4** are satisfactory, being similar to **Model 3**'s. The three large residuals in the residual plot are still present as before.

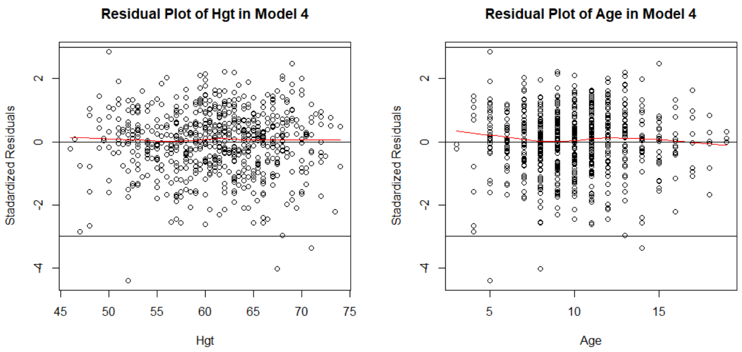

Fig. 6 Residual analysis on Hgt and Age in Model 4

From Fig. 6, linearity assumption is satisfied because the residual plots for the regressors *Age* and *Hgt* appear acceptable, being random around 0 in a horizontal band, ranging from -3 to 3. The residual plots do not indicate any nonlinear pattern. Hence, I will not consider any higher order terms of the individual regressors. The three large residuals are observed as well.

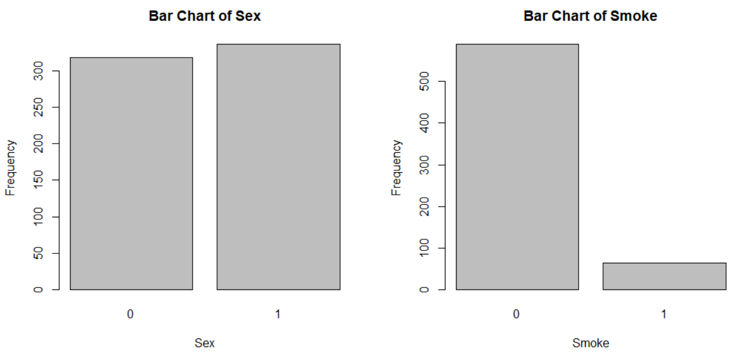

Fig. 7 Bar charts of categorical variables Sex and Smoke

Residual analysis is not performed on the categorical variables as it makes no sense to fit a linear model for them. From Fig. 7, the number of males and females is fairly symmetric. However, the number of smokers is very small relative to the number of non-smokers. This is likely a constraint because the data focuses on children ages 3 to 19, and it is often assumed that children are less likely to smoke, probably due to reasons like legal age and thus less exposure to smoking. This problem can be solved only if there is data over a wider range of age above 19.

V. Test for Multicollinearity

```
> X<-cbind(Age, Hgt)
> X<-cor(X)
> C<-solve(X)
> VIF <- diag(C)
> VIF
     Age      Hgt
2.682221 2.682221
```

Fig. 8 Variance Inflation Factors of regressors in Model 4

I used Variance Inflation Factors (VIF) to test on the individual regressors. VIF values are not calculated for regressors involving categorical variables because it does not make sense to fit a linear model for them. From Fig. 8, the VIF values of all the other regressors, *Age* and *Hgt*, are small and well below the threshold value of 10. Hence, the problem of multicollinearity will not be considered in **Model 4**.

VI. Variable Selection for Final Model

After model adequacy checking, I arrived at the full model for evaluation, **Model 4**. I now perform all possible subset regressions of **Model 4** using R package 'olsrr' and function 'ols_step_all_possible'. I will use the R-squared, R^2, and adjusted R-squared, R_{adj}^2, as the criteria.

```
> head(model_data)
  Index k Regressors in Model  R-squared Adjusted R-squared
1     1 1                 Hgt 0.79560711         0.79529362
2     2 1                 Age 0.59584474         0.59522487
3     3 1               Smoke 0.05977974         0.05833768
4     4 1                 Sex 0.02874056         0.02725090
5     5 2             Hgt Age 0.80712202         0.80652946
6     6 2             Hgt Sex 0.79639326         0.79576774
```

Fig. 9 First 6 rows of summary statistics of all 25 possible regressions

Model 4 has 6 regressors so there are $2^6 = 64$ total regressions to be examined. However, if the interaction terms *Age*Sex*, *Age*Smoke* are present, their respective first order terms must be present. As such, data cleaning is performed to reduce to 25 possible regressions. From Fig. 9, considering only the one-regressor models, the effect of *Hgt* is the strongest at $R^2 = 0.79560711$ compared to the other individual regressors. (refer to Appendix B for the full summary statistics of all 25 possible regressions)

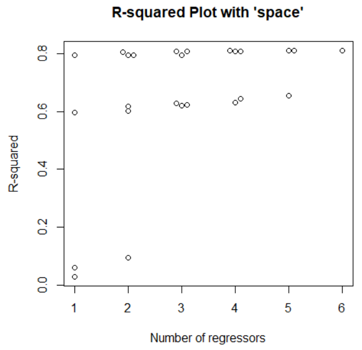

Fig. 10 R-squared plot with 'space'

From Fig. 10 and Appendix B, it should be noted that all the points at around $R^2 = 0.8$ have *Hgt* as one of the regressors. Because of the significant effect of *Hgt* as explained above, when *Hgt* is already in the model, adding further regressors only improves R^2 slightly, and will still be approximately at $R^2 = 0.8$.

```
> bestsubsets
   Index k                    Regressors in Model R-squared Adjusted R-squared
1      1 1                                    Hgt 0.7956071          0.7952936
5      5 2                                Hgt Age 0.8071220          0.8065295
11    11 3                            Hgt Age Sex 0.8092256          0.8083451
17    17 4                      Hgt Age Sex Smoke 0.8106392          0.8094722
22    22 5            Hgt Age Sex Smoke Age:Smoke 0.8111859          0.8097290
25    25 6 Hgt Age Sex Smoke Age:Sex Age:Smoke 0.8112581          0.8095078
```

Fig. 11 Best subset regressions based on adjusted R-squared for every value of k

After further data cleaning, Fig. 11 shows the best subset regressions based on adjusted R-squared for every value of k, where k is the number of regressors in the model. It can be seen that the regression models at k = 4,5,6 have similar R^2 values. At k = 4, adding further regressors will be of little use as the increase in R^2 is relatively less significant compared to values of k < 4. Hence, I propose the final model to be the model with k = 4 in Fig. 11, which is the same as **Model 3**: *log(FEV) ~ Hgt + Age + Sex + Smoke* .

Alternatively, I will perform stepwise regression using the Akaike Information Criterion (AIC) to do variable selection.

```
Start:  AIC=-1436.59
log(FEV) ~ 1

         Df Sum of Sq    RSS     AIC
+ Hgt     1    57.672 14.816 -2472.9
+ Age     1    43.192 29.297 -2027.1
+ Smoke   1     4.333 68.155 -1474.9
+ Sex     1     2.083 70.405 -1453.7
<none>                72.488 -1436.6
```

```
Step:  AIC=-2516.91
log(FEV) ~ Hgt + Age + Sex + Smoke

            Df Sum of Sq    RSS     AIC
<none>                   13.726 -2516.9
+ Age:Smoke  1    0.0396 13.687 -2516.8
+ Age:Sex    1    0.0031 13.723 -2515.1
- Smoke      1    0.1025 13.829 -2514.1
- Sex        1    0.1317 13.858 -2512.7
- Age        1    1.0323 14.759 -2471.5
- Hgt        1   13.7401 27.467 -2065.3
```

Fig. 12 First and last step in stepwise regression

From Fig. 12, the proposed final model under stepwise regression is the same as **Model 3** as well. Before I confirm my selection of the final model, I will test for the significance of the interaction terms in the model.

```
> anova(model_4)
Analysis of Variance Table

Response: log(FEV)
           Df Sum Sq Mean Sq   F value    Pr(>F)    
Hgt         1 57.672  57.672 2727.3102 < 2.2e-16 ***
Age         1  0.835   0.835   39.4727  6.11e-10 ***
Sex         1  0.152   0.152    7.2108  0.007432 ** 
Smoke       1  0.102   0.102    4.8461  0.028061 *  
Age:Sex     1  0.003   0.003    0.1476  0.700979    
Age:Smoke   1  0.042   0.042    1.9737  0.160535    
Residuals 647 13.682   0.021                        
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fig. 13 ANOVA table of Model 4

From the ANOVA table of **Model 4** in Fig. 13, under H_0: $\beta_5 = \beta_6 = 0$ vs. H_1: $\beta_j \neq 0$ for at least one $j = 5, 6$, the test statistic is $F = \frac{SS_R(\beta_5\beta_6|\beta_1\beta_2\beta_3\beta_4\beta_0)/2}{MS_{Res}} = \frac{(0.042+0.003)\,/\,2}{0.021} = 1.071429 \sim F_{2,647}$. Using R, the p-value = 0.3431256 > 0.15, so there is not enough evidence against H_0. The interaction terms are statistically insignificant. Without further information to determine whether the interaction terms are meaningful, I choose not to include the interaction terms and stick to **Model 3** as my final model.

VII. Adequacy Checking of Final Model – Model 3

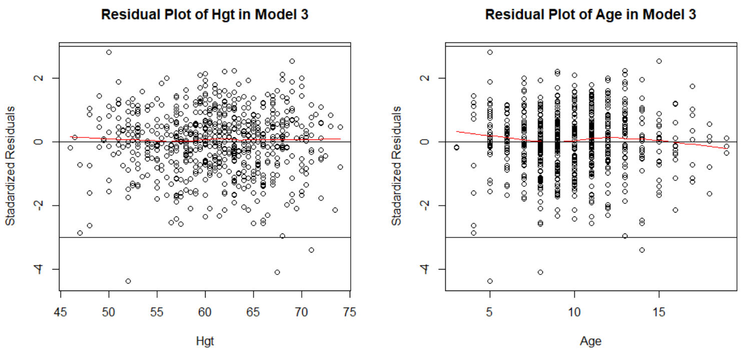

Fig. 14 Residual analysis on Hgt and Age in Model 3

In section II, I have already performed residual analysis on *log(FEV)* in **Model 3**, where the Q-Q plot and residual plot is satisfactory. For regressors, from Fig. 14, the residual plots of *Age* and *Hgt* do not indicate any nonlinearity, as the points are random about 0 and in a horizontal band, ranging from -3 to 3.

For the test of multicollinearity, since the VIF values cannot be computed from regressors involving the categorical variables, the results are the same from earlier in Fig. 8. The VIF values are well below the threshold value of 10, so the problem of multicollinearity is not considered in **Model 3** as well.

However, it is noted that there are three large residuals in the residual plots previously, which are data points 2, 140 and 473. Hence, there is a need to test for possible influential points.

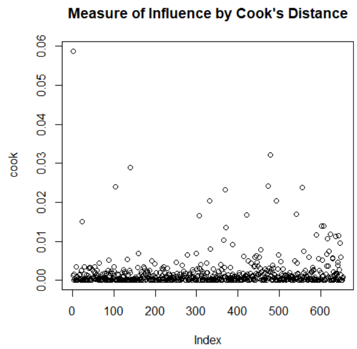

Fig. 15 Plot of Cook's Distance values

To measure the influence of the possible outliers, I use the Cook's Distance and the threshold value of $D_i > 1$ to determine that a data point is an influential point. However, from the plot in Fig. 15, none of the data points satisfy $D_i > 1$. I conclude that the three data points are non-influential points.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.942930   0.078618 -24.713  < 2e-16 ***
Hgt          0.042783   0.001679  25.488  < 2e-16 ***
Age          0.023387   0.003348   6.986 7.01e-12 ***
Sex1         0.029236   0.011716   2.496   0.0128 *
Smoke1      -0.046015   0.020905  -2.201   0.0281 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1454 on 649 degrees of freedom
Multiple R-squared:  0.8106,	Adjusted R-squared:  0.8095
F-statistic: 694.6 on 4 and 649 DF,  p-value: < 2.2e-16
```

Fig. 16 Summary statistics of Model 3

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.980826   0.076407 -25.925  < 2e-16 ***
Hgt          0.043812   0.001633  26.835  < 2e-16 ***
Age          0.021524   0.003249   6.625  7.3e-11 ***
Sex1         0.022113   0.011351   1.948   0.0518 .
Smoke1      -0.047946   0.020186  -2.375   0.0178 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1403 on 646 degrees of freedom
Multiple R-squared:  0.8209,	Adjusted R-squared:  0.8198
F-statistic: 740.3 on 4 and 646 DF,  p-value: < 2.2e-16
```

Fig. 17 Summary statistics of Model 5

I attempt to refit the model discarding the three non-influential points, called **Model 5**. Comparing the summary statistics in Fig. 16 and Fig. 17 (refer to Appendix A for full summary tables), R^2 improved slightly from $R^2 = 0.8106$ to $R^2 = 0.8209$. However, deleting the three data points has almost no effect on the estimations. Without further information, the three data points should not be discarded.

Hence, I keep to **Model 3** instead. From the above adequacy checking, **Model 3** satisfies the assumptions of linearity, constant variance, normality and independent errors, and there is no problem of multicollinearity as well.

VIII. Interpretation of Final Model

From the summary table of **Model 3** in Appendix A, **Model 3** provides a good fit with a strong $R^2 = 0.8096$. Before I interpret the model, I test for both the significance of **Model 3** and the individual regression coefficients.

Under H_0: $\beta_1 = \beta_2 = \ldots = \beta_7$ vs. H_1: $\beta_j \neq 0$ for at least one j, F-statistic: $F = 694.6 \sim F_{4, 649}$, has very small p-value < 2.2e-16 so data provides strong evidence against H_0 and **Model 3** is statistically significant.

Under H_0: $\beta_j = 0$ vs. H_1: $\beta_j \neq 0$, from the summary table of **Model 3** in Appendix A,

Regressor	**Regression coefficient**	***T*-statistic**	**p-value**
Hgt	β_1	$25.488 \sim t_{649}$	< 2e-16
Age	β_2	$6.986 \sim t_{649}$	7.01e-12
Sex	β_3	$2.496 \sim t_{649}$	0.0128
Smoke	β_4	$-2.201 \sim t_{649}$	0.0281

Fig. 18 T-statistics and p-values for test on individual regression coefficients

From Fig. 18, the tests on β_j for all regressors have p-values < 0.05, so data provides strong evidence against H_0. All the regressors are statistically significant.

I now attempt to interpret the model coefficients. Let $y = \log(FEV)$. From the summary table of **Model 3** in Appendix A, **Model 3** is given by:

$$\hat{y} = \beta_0 + \beta_1 * Hgt + \beta_2 * Age + \beta_3 * I(Sex = 1) + \beta_4 * I(Smoke = 1)$$

$$\hat{y} = -1.942930 + 0.042783 * Hgt + 0.023387 * Age + 0.029236 * I(Sex = 1) - 0.046015 * I(Smoke = 1)$$

I first interpret the categorical variables. The estimated mean of *log(FEV)* for males is larger than females by 0.028735. Using R, the 95% confidence interval (CI) for *Sex* is (0.005652577, 0.05181742). This means than I am 95% confident that the mean of *log(FEV)* for males is higher than females by a value between 0.005652577 and 0.05181742.

On the other hand, the estimated mean of *log(FEV)* for smokers is smaller than non-smokers by 0.047056. Using R, the 95% CI of *Smoke* is (-0.08821753, -0.005894472). I am also 95% confident that the mean of *log(FEV)* for smokers is lower than non-smokers by a value between 0.08821753 and 0.005894472, which makes sense because of the negative CI.

For the other regressors, as *Age* increases by 1 unit, the estimated *log(FEV)* increases by 0.023628. As *Hgt* increases by 1 unit, the estimated *log(FEV)* increases by 1.681433.

Another interpretation would be to exponentiate the model equation to get:

$$\widehat{FEV} = exp(\beta_0) * exp(\beta_1 * Hgt) * exp(\beta_2 * Age) * exp(\beta_3 * I(Sex = 1)) * exp(\beta_4 * I(Smoke = 1))$$

This implies that *FEV* has a multiplicative relationship with the regressors instead of the usual additive relationship, which allow for interaction effects of all regressors when estimating the mean of *FEV*. This seems to suggest that a non-smoking male that has relatively larger age and taller height will have the estimated mean of FEV to be proportionally higher than the others, judging from the signs of the coefficients.

Keeping other regressors constant, I can differentiate $\hat{y}$ with respect to *Age*:

$$\left(\frac{\partial}{\partial\, Age}\right)\hat{y} = \left(\frac{\partial\, \widehat{FEV}}{\partial\, Age}\right) * \frac{1}{\widehat{FEV}} = \beta_2 \rightarrow \frac{\partial\, \widehat{FEV}}{\widehat{FEV}} = \partial\, Age * \beta_2$$

We see that $\frac{\partial\, \widehat{FEV}}{\widehat{FEV}}$ is the rate of change of $\widehat{FEV}$ and $\partial\, Age$ is the marginal unit of *Age*. To visualize better, I multiply each side by 100 to represent the values in percentages:

$$\%\, change\ in\ \widehat{FEV} = 100 * \frac{\partial\, \widehat{FEV}}{\widehat{FEV}} = 100 * \partial\, Age * \beta_2$$

This means that for every unit of *Age* increased, the estimated mean of *FEV* increases by $100 * \beta_2 = 2.3387\%$.

Similarly, the final equation for *Hgt* can be derived the same way (refer to Appendix C for the full steps):

$$\%\, change\ in\ \widehat{FEV} = 100 * \frac{\partial\, \widehat{FEV}}{\widehat{FEV}} = 100 * \partial\, Hgt * \beta_1$$

This means that for every unit of *Hgt* increased, the estimated mean of *FEV* increases by $100 * \beta_1 = 4.2783\%$. This is much easier to interpret using height in inches compared to using *Hgt_m* because every unit would be in metres, which would be less reasonable for a person's height to increase in units of metres.

In conclusion, **Model 3** seems to be suggesting a multiplicative interpretation on the *FEV* data, with a strong $R^2 = 0.8096$. Improvements can be made by including data over a wider range of ages to make the proportion of smokers and non-smokers more symmetric. Adding more data can also help to determine whether the three large residuals seen in the residual plot of **Model 3** are bad outliers or actually part of the distribution. The model can then be improved.

Appendix

The R code is appended at the end of the Appendix.

Appendix A: Summary statistics of all models

```
> summary(model_1)

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

Residuals:
     Min       1Q   Median       3Q      Max 
-1.41306 -0.25696  0.00108  0.26249  1.89828 

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

Residual standard error: 0.4117 on 648 degrees of freedom
Multiple R-squared:  0.7762,	Adjusted R-squared:  0.7744 
F-statistic: 449.4 on 5 and 648 DF,  p-value: < 2.2e-16
```

```
> summary(model_2)

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

Residuals:
     Min       1Q   Median       3Q      Max 
-1.38104 -0.24963  0.00817  0.25462  1.91721 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -4.455340   0.222770 -20.000  < 2e-16 ***
Hgt          0.104182   0.004756  21.904  < 2e-16 ***
Age          0.065510   0.009486   6.906 1.19e-11 ***
Sex1         0.156909   0.033197   4.727 2.80e-06 ***
Smoke1      -0.086846   0.059235  -1.466    0.143    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4121 on 649 degrees of freedom
Multiple R-squared:  0.7754,	Adjusted R-squared:  0.774 
F-statistic: 560.2 on 4 and 649 DF,  p-value: < 2.2e-16
```

```
> summary(model_3)

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

Residuals:
     Min       1Q   Median       3Q      Max 
-0.63443 -0.08644  0.01167  0.09492  0.40904 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.942930   0.078618 -24.713  < 2e-16 ***
Hgt          0.042783   0.001679  25.488  < 2e-16 ***
Age          0.023387   0.003348   6.986 7.01e-12 ***
Sex1         0.029236   0.011716   2.496   0.0128 *  
Smoke1      -0.046015   0.020905  -2.201   0.0281 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1454 on 649 degrees of freedom
Multiple R-squared:  0.8106,	Adjusted R-squared:  0.8095 
F-statistic: 694.6 on 4 and 649 DF,  p-value: < 2.2e-16
```

```
> summary(model_4)

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

Residuals:
     Min       1Q   Median       3Q      Max 
-0.63606 -0.08771  0.01224  0.09554  0.41337 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.898652   0.090011 -21.094  < 2e-16 ***
Hgt          0.041841   0.001813  23.078  < 2e-16 ***
Age          0.024647   0.003911   6.302 5.43e-10 ***
Sex1         0.011485   0.040541   0.283    0.777    
Smoke1       0.112222   0.114177   0.983    0.326    
Age:Sex1     0.001983   0.003987   0.497    0.619    
Age:Smoke1  -0.011936   0.008496  -1.405    0.161    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1454 on 647 degrees of freedom
Multiple R-squared:  0.8113,	Adjusted R-squared:  0.8095 
F-statistic: 463.5 on 6 and 647 DF,  p-value: < 2.2e-16
```

```
> summary(model_5)

Call:
lm(formula = log(FEV) ~ Hgt + Age + Sex + Smoke, data = data_2)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.43274 -0.08562  0.01154  0.08902  0.41192 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.980826   0.076407 -25.925  < 2e-16 ***
Hgt          0.043812   0.001633  26.835  < 2e-16 ***
Age          0.021524   0.003249   6.625  7.3e-11 ***
Sex1         0.022113   0.011351   1.948   0.0518 .  
Smoke1      -0.047946   0.020186  -2.375   0.0178 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1403 on 646 degrees of freedom
Multiple R-squared:  0.8209,	Adjusted R-squared:  0.8198 
F-statistic: 740.3 on 4 and 646 DF,  p-value: < 2.2e-16
```

Appendix B: Summary statistics of all 25 possible regressions

```
> model_data # 25 possible regressions
   Index k                   Regressors in Model  R-squared Adjusted R-squared
1      1 1                                   Hgt 0.79560711         0.79529362
2      2 1                                   Age 0.59584474         0.59522487
3      3 1                                 Smoke 0.05977974         0.05833768
4      4 1                                   Sex 0.02874056         0.02725090
5      5 2                               Hgt Age 0.80712202         0.80652946
6      6 2                               Hgt Sex 0.79639326         0.79576774
7      7 2                             Hgt Smoke 0.79564111         0.79501328
8      8 2                               Age Sex 0.61748203         0.61630686
9      9 2                             Age Smoke 0.60129869         0.60007380
10    10 2                             Sex Smoke 0.09533354         0.09255422
11    11 3                           Hgt Age Sex 0.80922555         0.80834506
12    12 3                         Hgt Age Smoke 0.80882219         0.80793983
13    13 3                         Hgt Sex Smoke 0.79639852         0.79545882
14    14 3                       Age Sex Age:Sex 0.62950379         0.62779381
15    15 3                   Age Smoke Age:Smoke 0.62362406         0.62188694
16    16 3                         Age Sex Smoke 0.62108985         0.61934103
17    17 4                     Hgt Age Sex Smoke 0.81063925         0.80947215
18    18 4                   Hgt Age Sex Age:Sex 0.80930912         0.80813383
19    19 4               Hgt Age Smoke Age:Smoke 0.80919058         0.80801456
20    20 4               Age Sex Smoke Age:Smoke 0.64457833         0.64238775
21    21 4                 Age Sex Smoke Age:Sex 0.63216347         0.62989637
22    22 5           Hgt Age Sex Smoke Age:Smoke 0.81118592         0.80972902
23    23 5             Hgt Age Sex Smoke Age:Sex 0.81068230         0.80922152
24    24 5       Age Sex Smoke Age:Sex Age:Smoke 0.65589104         0.65323588
25    25 6 Hgt Age Sex Smoke Age:Sex Age:Smoke 0.81125807         0.80950776
```

Appendix C: Interpretation of Regression Coefficients

For *Age:*

$$\left(\frac{\partial}{\partial\, Age}\right)\hat{y} = \left(\frac{\partial\, \widehat{FEV}}{\partial\, Age}\right) * \frac{1}{\widehat{FEV}} = \beta_2 \rightarrow \frac{\partial\, \widehat{FEV}}{\widehat{FEV}} = \partial\, Age * \beta_2$$

$$\%\ change\ in\ \widehat{FEV} = 100 * \frac{\partial\, \widehat{FEV}}{\widehat{FEV}} = 100 * \partial\, Age * \beta_2$$

For *Hgt:*

$$\left(\frac{\partial}{\partial\, Hgt}\right)\hat{y} = \left(\frac{\partial\, \widehat{FEV}}{\partial\, Hgt}\right) * \frac{1}{\widehat{FEV}} = \beta_1 \rightarrow \frac{\partial\, \widehat{FEV}}{\widehat{FEV}} = \partial\, Hgt * \beta_1$$

$$\%\ change\ in\ \widehat{FEV} = 100 * \frac{\partial\, \widehat{FEV}}{\widehat{FEV}} = 100 * \partial\, Hgt * \beta_1$$

```
### R Code for ST3131 Assignment

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

# Linear independence of the two height variables

model_1<-lm(FEV~Hgt+Hgt_m+Age+Sex+Smoke,data=data)
summary(model_1)

model_2<-lm(FEV~Hgt+Age+Sex+Smoke,data=data)
summary(model_2)

# Residual analysis on Model 2

qqnorm(rstandard(model_2),datax = TRUE, ylab = "Standardized Residuals", xlab = "Z scores",
main = "Normal Q-Q Plot of FEV in Model 2",pch=20)
qqline(rstandard(model_2),datax = TRUE, col='red')

plot(model_2$fitted.values,rstandard(model_2), xlab = "Fitted values",
ylab = "Stadardized Residuals",main = "Residual Plot of FEV in Model 2")
abline(h = 0)

# We perform Box-Cox method on Model 2.

library(MASS)

boxcox(model_2, lambda=seq(-2, 2, by=0.5),optimize=TRUE,plotit = TRUE)

# Refit model with transformation on response log(FEV)

model_3<-lm(log(FEV)~Hgt+Age+Sex+Smoke,data=data)
summary(model_3)
```

```
qqnorm(rstandard(model_3),datax = TRUE, ylab = "Standardized Residuals", xlab = "Z scores",
main = "Normal Q-Q Plot of log(FEV) in Model 3",pch=20)
qqline(rstandard(model_3),datax = TRUE, col='red')

plot(model_3$fitted.values,rstandard(model_3), xlab = "Fitted values",
ylab = "Stadardized Residuals",main = "Residual Plot of log(FEV) in Model 3")
with(data, lines(loess.smooth(model_3$fitted.values, rstandard(model_3)), col = "red"))
abline(h = 0)
abline(h = 3)
abline(h = -3)

# Check if there is a need for interaction terms

plot(log(FEV)[data$Sex=="0"]~Age[data$Sex=="0"],
pch = 1,col="red",xlim=c(0,20),ylim=c(0,2), xlab="Age",ylab="log(FEV)")
abline(lm(log(FEV)[data$Sex=="0"]~Age[data$Sex=="0"],data=data),col="red")
par(new=T)
plot(log(FEV)[data$Sex=="1"]~Age[data$Sex=="1"],
pch = 2,col="blue", xlim=c(0,20),ylim=c(0,2), xlab="",ylab="")
abline(lm(log(FEV)[data$Sex=="1"]~Age[data$Sex=="1"],data=data),col="blue")
par(new=F)
legend(14,.5,legend=c("Female", "Male"),
col=c("red", "blue"), pch=1:2, cex=1.2)

plot(log(FEV)[data$Sex=="0"]~Hgt[data$Sex=="0"],
pch = 1,col="red",xlim=c(40,80),ylim=c(0,2), xlab="Hgt",ylab="log(FEV)")
abline(lm(log(FEV)[data$Sex=="0"]~Hgt[data$Sex=="0"],data=data),col="red")
par(new=T)
plot(log(FEV)[data$Sex=="1"]~Hgt[data$Sex=="1"],
pch = 2,col="blue", xlim=c(40,80),ylim=c(0,2), xlab="",ylab="")
abline(lm(log(FEV)[data$Sex=="1"]~Hgt[data$Sex=="1"],data=data),col="blue")
par(new=F)
legend(67,0.4,legend=c("Female", "Male"),
col=c("red", "blue"), pch=1:2, cex=1.2)

plot(log(FEV)[data$Smoke=="0"]~Age[data$Smoke=="0"],
pch = 1,col="purple",xlim=c(0,20),ylim=c(0,2), xlab="Age",ylab="log(FEV)")
abline(lm(log(FEV)[data$Smoke=="0"]~Age[data$Smoke=="0"],data=data),col="purple")
```

```
par(new=T)
plot(log(FEV)[data$Smoke=="1"]~Age[data$Smoke=="1"],
pch = 2,col="green", xlim=c(0,20),ylim=c(0,2), xlab="",ylab="")
abline(lm(log(FEV)[data$Smoke=="1"]~Age[data$Smoke=="1"],data=data),col="green")
par(new=F)
legend(12,0.4,legend=c("Non-smoker", "Smoker"),
col=c("purple", "green"), pch=1:2, cex=1.2)

plot(log(FEV)[data$Smoke=="0"]~Hgt[data$Smoke=="0"],
pch = 1,col="purple",xlim=c(40,80),ylim=c(0,2), xlab="Hgt",ylab="log(FEV)")
abline(lm(log(FEV)[data$Smoke=="0"]~Hgt[data$Smoke=="0"],data=data),col="purple")
par(new=T)
plot(log(FEV)[data$Smoke=="1"]~Hgt[data$Smoke=="1"],
pch = 2,col="green", xlim=c(40,80),ylim=c(0,2), xlab="",ylab="")
abline(lm(log(FEV)[data$Smoke=="1"]~Hgt[data$Smoke=="1"],data=data),col="green")
par(new=F)
legend(65,0.4,legend=c("Non-smoker", "Smoker"),
col=c("purple", "green"), pch=1:2, cex=1.2)

# Residual analysis on Model 4

model_4<-lm(log(FEV)~Hgt+Age+Sex+Smoke+Age*Sex+Age*Smoke,data=data)

qqnorm(rstandard(model_4),datax = TRUE, ylab = "Standardized Residuals", xlab = "Z scores",
main = "Normal Q-Q Plot of log(FEV) in Model 4",pch=20)
qqline(rstandard(model_4),datax = TRUE, col='red')

plot(model_4$fitted.values,rstandard(model_4), xlab = "Fitted values",
ylab = "Stadardized Residuals",main = "Residual Plot of log(FEV) in Model 4")
with(data, lines(loess.smooth(model_4$fitted.values, rstandard(model_4)), col = "red"))
abline(h = 0)
abline(h = 3)
abline(h = -3)

plot(Hgt,rstandard(model_4), xlab = "Hgt",
ylab = "Stadardized Residuals",main = "Residual Plot of Hgt in Model 4")
with(data, lines(loess.smooth(Hgt, rstandard(model_4)), col = "red"))
abline(h = 0)
```

```
abline(h = 3)
abline(h = -3)

plot(Age,rstandard(model_4), xlab = "Age",
ylab = "Stadardized Residuals",main = "Residual Plot of Age in Model 4")
with(data, lines(loess.smooth(Age, rstandard(model_4)), col = "red"))
abline(h = 0)
abline(h = 3)
abline(h = -3)

plot(Sex, xlab="Sex",ylab="Frequency",main="Bar Chart of Sex")
plot(Smoke, xlab="Smoke",ylab="Frequency",main="Bar Chart of Smoke")

# Test for multicollinearity using VIF

X<-cbind(Age, Hgt)
X<-cor(X)
C<-solve(X)
VIF <- diag(C)
VIF

# Evaluating all possible regressions

library(olsrr)
model_full<-lm(log(FEV)~Hgt+Age+Sex+Smoke+Age*Sex+Age*Smoke,data=data)
models<-ols_step_all_possible(model_full)
models_cleaned<-models[-c(2:3,
7:8,12:17,19,
22:26,28:29,32:36,38:39,
42:48,52:53,55,57:58,60),] #delete rows where the interaction terms that do not have the first order terms present

model_data<-data.frame(models_cleaned)[,c(1:5)]

names(model_data)[1] = 'Index'
names(model_data)[2] = 'k'
names(model_data)[3] = 'Regressors in Model'
names(model_data)[4] = 'R-squared'
names(model_data)[5] = 'Adjusted R-squared'
```

```
model_data["Index"] <- 1:nrow(model_data)
row.names(model_data) <- 1:nrow(model_data)

model_data # 25 possible regressions
head(model_data)

library(gplots)
plot(space(model_data$k,model_data$"R-squared"),
xlab="Number of regressors",ylab="R-squared",
main="R-squared Plot with 'space'")

bestsubsets<-model_data[c(1,5,11,17,22,25),]
bestsubsets

# Variable selection using stepwise regression

model_s<-lm(log(FEV)~ 1, data = data)

model_e<-stepAIC(model_s, direction = "both",
scope = log(FEV)~Hgt+Age+Sex+Smoke+Age*Sex+Age*Smoke,data=data)

summary(model_e)

summary(model_4)
anova(model_4)
# F-statistic is 1.071429.
1-pf(1.071429,2,647) # p-value = 0.3431256

anova(model_3,model_4)

# Residual analysis on regressor in Model 3

plot(Hgt,rstandard(model_3), xlab = "Hgt",
ylab = "Stadardized Residuals",main = "Residual Plot of Hgt in Model 3")
with(data, lines(loess.smooth(Hgt, rstandard(model_3)), col = "red"))
abline(h = 0)
abline(h = 3)
abline(h = -3)
```

```
plot(Age,rstandard(model_3), xlab = "Age",
ylab = "Stadardized Residuals",main = "Residual Plot of Age in Model 3")
with(data, lines(loess.smooth(Age, rstandard(model_3)), col = "red"))
abline(h = 0)
abline(h = 3)
abline(h = -3)

# Test for influential points

cook<-cooks.distance(model_3)
plot(cook, main="Measure of Influence by Cook's Distance")
abline(h = 1)
order(rstandard(model_3)) #the three possible outliers are data points 2,140,473

# Refit model, discarding the three possible outliers

data_2<-data[-c(2,140,473),]
attach(data_2) #reattach data
model_5<-lm(log(FEV)~Hgt+Age+Sex+Smoke,data=data_2)
summary(model_5)

# Confidence intervals for categorical terms

qt(0.975,649)
CI_Sex<-cbind(CIlower = 0.028735 - qt(0.975,649) * 0.011755,
CIupper = 0.028735 + qt(0.975,649) * 0.011755)
CI_Sex
CI_Smoke<-cbind(CIlower = -0.047056 - qt(0.975,649) * 0.020962,
CIupper = -0.047056 + qt(0.975,649) * 0.020962)
CI_Smoke
```