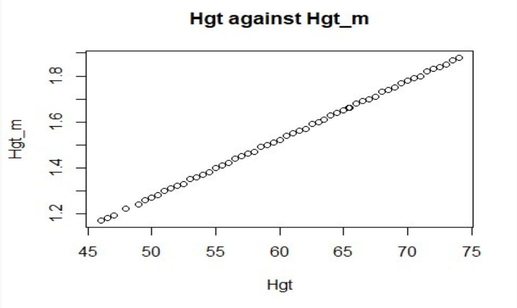
# ST3131 Assignment

```
## Call:
## lm(formula = FEV ~ Age + Sex + Smoke + Hgt + Hgt_m, 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 ***
## 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    
## Hgt          0.312051   0.142227   2.194   0.0286 *  
## Hgt_m       -8.197478   5.605713  -1.462   0.1441    
## ---
## 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
```

```
## Analysis of Variance Table
## 
## Response: FEV
##            Df  Sum Sq Mean Sq   F value    Pr(>F)    
## Age         1 280.893 280.893 1657.0034 < 2.2e-16 ***
## Sex         1  17.040  17.040  100.5212 < 2.2e-16 ***
## Smoke       1   1.144   1.144    6.7503  0.009586 ** 
## Hgt         1  81.478  81.478  480.6441 < 2.2e-16 ***
## Hgt_m       1   0.363   0.363    2.1384  0.144132    
## Residuals 648 109.848   0.170                        
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

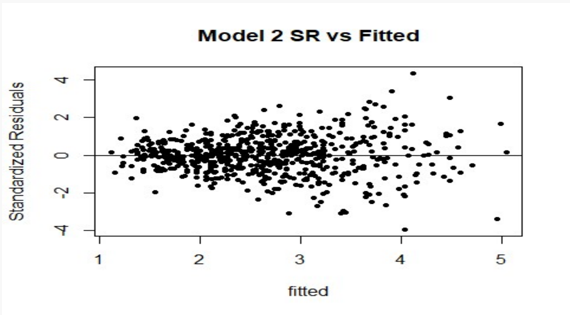


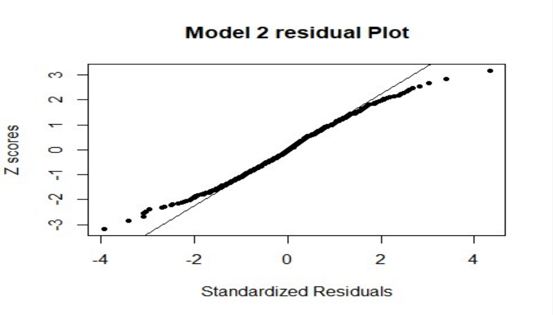
**(Referring to R-code part 1)** From the above plot, we can see that the variables Hgt is nearly linearly dependent on Hgt_M. Since this gives rise to the possibility of multicollinearity, and from anova table of model1, I know that Hgt_m has a larger p-value than Hgt, I will remove Hgt_m since it is not as significant as Hgt.

Additionally, since logically we can see that certain regressors will affect another, we will introduce certain interaction terms in the model. For now, the interaction terms that will be introduced will be (Smoke * Age), (Sex * Smoke), (Sex * Hgt), (Age * Hgt). (Smoke * Hgt) will be excluded since it makes the least sense to me. Just because someone is tall does not mean that the person is more likely to smoke.

Lastly, I will also add in quadratic terms (I(Age^2) and I(Hgt^2)) as I want to maximise the fit for now. I will remove regressors that are not essential to the model later on.

```
## Call:
## lm(formula = FEV ~ Age + Sex + Smoke + Hgt + Smoke * Age + Sex * 
##     Smoke + Age * Hgt + Sex * Hgt + I(Age^2) + I(Hgt^2), data = data)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.49590 -0.22845  0.01279  0.23456  1.67474 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)   
## (Intercept)  4.558e-01  2.314e+00   0.197  0.84390   
## Age         -3.560e-01  1.416e-01  -2.514  0.01219 * 
## Sex1        -8.601e-01  3.916e-01  -2.197  0.02841 *
## Smoke1       4.995e-01  3.328e-01   1.501  0.13387
## Hgt          1.542e-02  9.359e-02   0.165  0.86920
## I(Age^2)    -2.903e-03  2.997e-03  -0.969  0.33313
## I(Hgt^2)     2.959e-05  9.539e-04   0.031  0.97526
## Age:Smoke1  -4.914e-02  2.525e-02  -1.946  0.05210 .
## Sex1:Smoke1  6.187e-02  1.109e-01   0.558  0.57706
## Age:Hgt      7.937e-03  2.976e-03   2.667  0.00784 **
## Sex1:Hgt     1.585e-02  6.530e-03   2.427  0.01551 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3903 on 643 degrees of freedom
## Multiple R-squared:  0.8004, Adjusted R-squared:  0.7973
## F-statistic: 257.9 on 10 and 643 DF,  p-value: < 2.2e-16

## Analysis of Variance Table
##
## Response: FEV
##            Df  Sum Sq Mean Sq   F value    Pr(>F)
## Age         1 280.893 280.893 1843.9343 < 2.2e-16 ***
## Sex         1  17.040  17.040  111.8613 < 2.2e-16 ***
## Smoke       1   1.144   1.144    7.5119  0.006300 **
## Hgt         1  81.478  81.478  534.8669 < 2.2e-16 ***
## I(Age^2)    1   4.129   4.129   27.1043 2.597e-07 ***
## I(Hgt^2)    1   5.144   5.144   33.7680 9.774e-09 ***
## Age:Smoke   1   0.520   0.520    3.4148  0.065075 .
## Sex:Smoke   1   0.408   0.408    2.6798  0.102117
## Age:Hgt     1   1.162   1.162    7.6268  0.005915 **
## Sex:Hgt     1   0.897   0.897    5.8886  0.015513 *
## Residuals 643  97.950   0.152
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

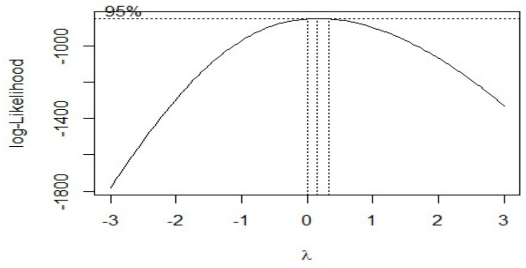




**(Referring to R-code part 2)** From the summary, even though Sex * Smoke still has quite a high p-value, I will just leave it in the model for now since the R^2 is somewhat strong, at 0.8. Next up, I will proceed to test the adequacy of the model, using the residual plot and the residual vs fitted plot.

From the fitted values vs residual plot, we can see that the constant variance assumption is violated and the points are ranging from close to -4 to 4, which is not very ideal.

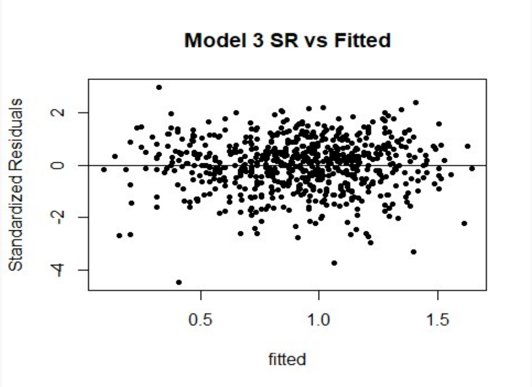
From the residual plots, we can see that the normality assumption is also violated since we can see a trend whereby the right and left tail deviates from the straight line. Hence, I will attempt to do some transformation to make the variance more constant. For the transformation, I will do a boxcox transformaton.

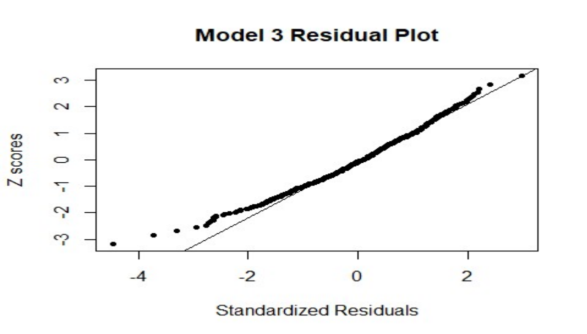


**(Referring to R-code part 3)** From the boxcox diagram, I will take 0 to transform my response since it is easier to interpret as compared to 0.2. As such, my response will be transformed to log(FEV).

```
##
## Call:
## lm(formula = log(FEV) ~ Age + Sex + Smoke + Hgt + Smoke * Age +
##     Sex * Smoke + Age * Hgt + Sex * Hgt + I(Age^2) + I(Hgt^2),
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.64357 -0.08702  0.01360  0.09503  0.42723
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.5833218  0.8615887  -4.159 3.63e-05 ***
## Age         -0.0643388  0.0527379  -1.220   0.2229
## Sex1        -0.1003664  0.1458068  -0.688   0.4915
## Smoke1       0.1251306  0.1239266   1.010   0.3130
## Hgt          0.1126094  0.0348499   3.231   0.0013 **
## I(Age^2)    -0.0010640  0.0011161  -0.953   0.3408
## I(Hgt^2)    -0.0007350  0.0003552  -2.069   0.0389 *
## Age:Smoke1  -0.0129784  0.0094032  -1.380   0.1680
## Sex1:Smoke1  0.0095665  0.0412876   0.232   0.8168
## Age:Hgt      0.0018119  0.0011081   1.635   0.1025
## Sex1:Hgt     0.0022023  0.0024315   0.906   0.3654
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1453 on 643 degrees of freedom
## Multiple R-squared:  0.8127, Adjusted R-squared:  0.8097
## F-statistic: 278.9 on 10 and 643 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 2045.0700 < 2.2e-16 ***
## Sex         1  1.568   1.568   74.2639 < 2.2e-16 ***
## Smoke       1  0.262   0.262   12.3828 0.0004639 ***
## Hgt         1 13.740  13.740  650.5752 < 2.2e-16 ***
## I(Age^2)    1  0.007   0.007    0.3501 0.5542517
## I(Hgt^2)    1  0.010   0.010    0.4812 0.4881399
## Age:Smoke   1  0.041   0.041    1.9592 0.1620779
## Sex:Smoke   1  0.011   0.011    0.5250 0.4689785
## Age:Hgt     1  0.059   0.059    2.7918 0.0952344 .
## Sex:Hgt     1  0.017   0.017    0.8204 0.3654153
## Residuals 643 13.580   0.021
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Model 3 SR vs Fitted


Model 3 Residual Plot

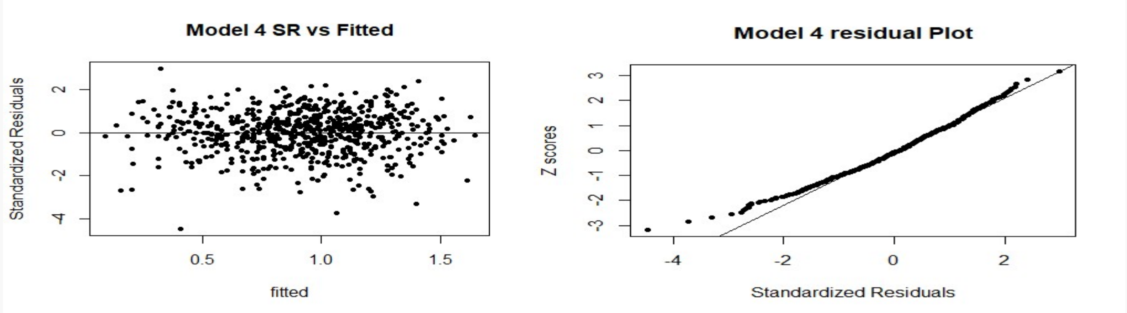
```
##         Age         Sex       Smoke         Hgt    I(Age^2)    I(Hgt^2)
##  750.355261  164.456604   42.568344 1221.537411  155.003704 1865.808259
##   Age:Smoke   Sex:Smoke     Age:Hgt     Sex:Hgt
##   46.285015    2.015118 1961.201395  179.709062
```

**(Referring to R-code part 4)** From Model3 SR vs fitted plot, it appears that the variance is acceptable as there is no obvious non-constant variance and now there is only one point that is less than -4. However, the VIF values are not acceptable since they are in the hundreds and even thousands. As such, I will conduct unit length scaling to reduce VIF. Note that age_new refers to Age that has been scaled and hgt_new refers to Hgt that has been scaled.

```
##
## Call:
## lm(formula = log(FEV) ~ age_new + Sex + Smoke + hgt_new + Smoke *
##     age_new + Sex * Smoke + age_new * hgt_new + Sex * hgt_new +
##     I(age_new^2) + I(hgt_new^2), data = data)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.64357 -0.08702  0.01360  0.09503  0.42723
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.910684   0.010160  89.635  < 2e-16 ***
## age_new          0.890967   0.146186   6.095 1.89e-09 ***
## Sex1             0.034291   0.013000   2.638  0.00855 **
## Smoke1          -0.003760   0.038808  -0.097  0.92284
## hgt_new          2.768092   0.180379  15.346  < 2e-16 ***
## I(age_new^2)    -1.318409   1.382910  -0.953  0.34077
## I(hgt_new^2)    -3.394948   1.640610  -2.069  0.03891 *
## age_new:Smoke1  -0.456841   0.330995  -1.380  0.16800
## Sex1:Smoke1      0.009566   0.041288   0.232  0.81684
## age_new:hgt_new  4.334717   2.650975   1.635  0.10251
## Sex1:hgt_new     0.149681   0.165259   0.906  0.36542
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1453 on 643 degrees of freedom
## Multiple R-squared:  0.8127, Adjusted R-squared:  0.8097
## F-statistic: 278.9 on 10 and 643 DF,  p-value: < 2.2e-16
```

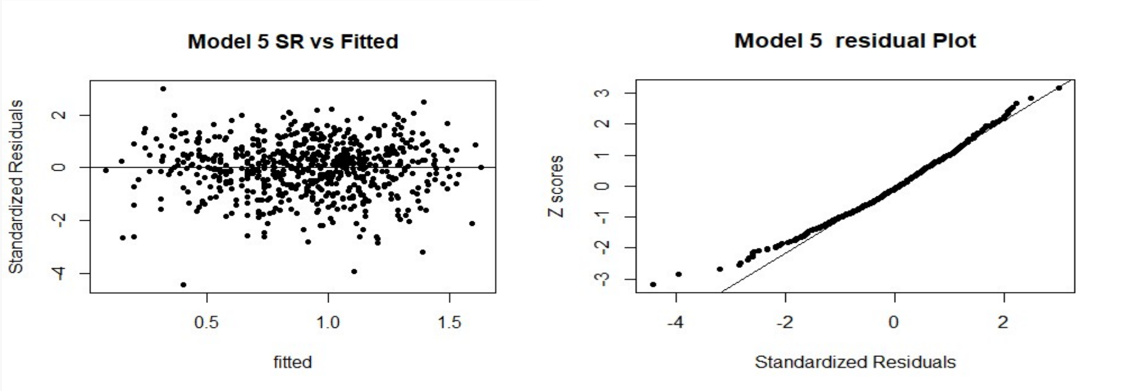
```
## Analysis of Variance Table
## Response: log(FEV)
##                  Df Sum Sq Mean Sq   F value    Pr(>F)
## age_new           1 43.192  43.192 2045.0700 < 2.2e-16 ***
## Sex               1  1.568   1.568   74.2639 < 2.2e-16 ***
## Smoke             1  0.262   0.262   12.3828 0.0004639 ***
## hgt_new           1 13.740  13.740  650.5752 < 2.2e-16 ***
## I(age_new^2)      1  0.007   0.007    0.3501 0.5542517
## I(hgt_new^2)      1  0.010   0.010    0.4812 0.4881399
## age_new:Smoke     1  0.041   0.041    1.9592 0.1620779
## Sex:Smoke         1  0.011   0.011    0.5250 0.4689785
## age_new:hgt_new   1  0.059   0.059    2.7918 0.0952344 .
## Sex:hgt_new       1  0.017   0.017    0.8204 0.3654153
## Residuals       643 13.580   0.021
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



```
##         age_new             Sex           Smoke         hgt_new    I(age_new^2)
##        4.653080        1.307371        4.174384        7.084445        6.160395
##    I(hgt_new^2)   age_new:Smoke       Sex:Smoke age_new:hgt_new     Sex:hgt_new
##        6.207006        4.628144        2.015118       12.849294        3.794035
```

**(Referring to R-code part 5)** After the scaling, it is evident that the VIF values have decreased tremendously. However, there is still one VIF value belonging to age_new * hgt_new that is too large (above 10), hinting at possibility of multicollinearity. Hence, I will remove the regressor from the model.

```
##
## Call:
## lm(formula = log(FEV) ~ age_new + Sex + Smoke + hgt_new + Smoke *
##     age_new + Sex * Smoke + Sex * hgt_new + I(age_new^2) + I(hgt_new^2),
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.63865 -0.08616  0.01331  0.09575  0.43123
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.906677   0.009873  91.836  < 2e-16 ***
## age_new         0.868455   0.145725   5.960 4.17e-09 ***
## Sex1            0.034563   0.013016   2.655  0.00812 **
## Smoke1         -0.002629   0.038852  -0.068  0.94607    
## hgt_new         2.747201   0.180160  15.249  < 2e-16 ***
## I(age_new^2)    0.436501   0.873245   0.500  0.61734    
## I(hgt_new^2)   -1.164072   0.912335  -1.276  0.20244    
## age_new:Smoke1 -0.509199   0.329870  -1.544  0.12317    
## Sex1:Smoke1     0.017854   0.041028   0.435  0.66359    
## Sex1:hgt_new    0.159978   0.165354   0.967  0.33366    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.1455 on 644 degrees of freedom
## Multiple R-squared:  0.8119, Adjusted R-squared:  0.8092 
## F-statistic: 308.8 on 9 and 644 DF,  p-value: < 2.2e-16

## Analysis of Variance Table
## Response: log(FEV)
##                 Df Sum Sq Mean Sq   F value    Pr(>F)    
## age_new          1 43.192  43.192 2039.7689 < 2.2e-16 ***
## Sex              1  1.568   1.568   74.0714 < 2.2e-16 ***
## Smoke            1  0.262   0.262   12.3507 0.0004717 ***
## hgt_new          1 13.740  13.740  648.8888 < 2.2e-16 ***
## I(age_new^2)     1  0.007   0.007    0.3492 0.5547651    
## I(hgt_new^2)     1  0.010   0.010    0.4799 0.4887036    
## age_new:Smoke    1  0.041   0.041    1.9542 0.1626214    
## Sex:Smoke        1  0.011   0.011    0.5236 0.4695545    
## Sex:hgt_new      1  0.020   0.020    0.9360 0.3336622    
## Residuals      644 13.637   0.021                        
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



```
##       age_new           Sex         Smoke       hgt_new  I(age_new^2) 
##      4.611807      1.307158      4.173058      7.048902      2.450000 
##  I(hgt_new^2) age_new:Smoke     Sex:Smoke   Sex:hgt_new 
##      1.914491      4.584832      1.984749      3.788526
```

**(Referring to R-code part6)** After removing age_new * hgt_new from the model, we can see that the qqplot is still approximately a straight line and the variance of the SR vs fitted model is still acceptable. And all the VIF values are now below 10, which is a good sign. Since the model is now adequate, i.e. the SR vs fitted diagram has constant variance, residual plots more or less follows a straight line, and all VIF values are acceptable, I will conduct variable selection now using backward selection.

```
## Start:  AIC=-2511.21
## log(FEV) ~ age_new + Sex + Smoke + hgt_new + Smoke * age_new + 
##     Sex * Smoke + Sex * hgt_new + I(age_new^2) + I(hgt_new^2)
## 
##                 Df Sum of Sq    RSS     AIC
## - Sex:Smoke      1  0.004010 13.641 -2513.0
## - I(age_new^2)   1  0.005291 13.642 -2513.0
## - Sex:hgt_new    1  0.019820 13.656 -2512.3
## - I(hgt_new^2)   1  0.034472 13.671 -2511.6
## <none>                       13.637 -2511.2
## - age_new:Smoke  1  0.050456 13.687 -2510.8
## 
## Step:  AIC=-2513.02
## log(FEV) ~ age_new + Sex + Smoke + hgt_new + I(age_new^2) + I(hgt_new^2) + 
##     age_new:Smoke + Sex:hgt_new
## 
##                 Df Sum of Sq    RSS     AIC
## - I(age_new^2)   1  0.005138 13.646 -2514.8
## - Sex:hgt_new    1  0.026899 13.668 -2513.7
## - I(hgt_new^2)   1  0.034446 13.675 -2513.4
## <none>                       13.641 -2513.0
## - age_new:Smoke  1  0.047715 13.688 -2512.7
##
## Step:  AIC=-2514.77
## log(FEV) ~ age_new + Sex + Smoke + hgt_new + I(hgt_new^2) + age_new:Smoke +
##     Sex:hgt_new
##
##                 Df Sum of Sq    RSS     AIC
## - Sex:hgt_new    1  0.026412 13.672 -2515.5
## - I(hgt_new^2)   1  0.030764 13.677 -2515.3
## <none>                       13.646 -2514.8
## - age_new:Smoke  1  0.042861 13.689 -2514.7
##
## Step:  AIC=-2515.51
## log(FEV) ~ age_new + Sex + Smoke + hgt_new + I(hgt_new^2) + age_new:Smoke
##
##                 Df Sum of Sq    RSS     AIC
## - I(hgt_new^2)   1    0.0147 13.687 -2516.8
## - age_new:Smoke  1    0.0369 13.709 -2515.7
## <none>                       13.672 -2515.5
## - Sex            1    0.1587 13.831 -2510.0
## - hgt_new        1   11.9183 25.590 -2107.5
##
## Step:  AIC=-2516.8
## log(FEV) ~ age_new + Sex + Smoke + hgt_new + age_new:Smoke
##
##                 Df Sum of Sq    RSS     AIC
## - age_new:Smoke  1    0.0396 13.726 -2516.9
## <none>                       13.687 -2516.8
## - Sex            1    0.1446 13.832 -2511.9
## - hgt_new        1   12.0771 25.764 -2105.1
##
## Step:  AIC=-2516.91
## log(FEV) ~ age_new + Sex + Smoke + hgt_new
##
##           Df Sum of Sq    RSS     AIC
## <none>                 13.726 -2516.9
## - Smoke    1    0.1025 13.829 -2514.1
## - Sex      1    0.1317 13.858 -2512.7
## - age_new  1    1.0323 14.759 -2471.5
## - hgt_new  1   13.7401 27.467 -2065.3
```

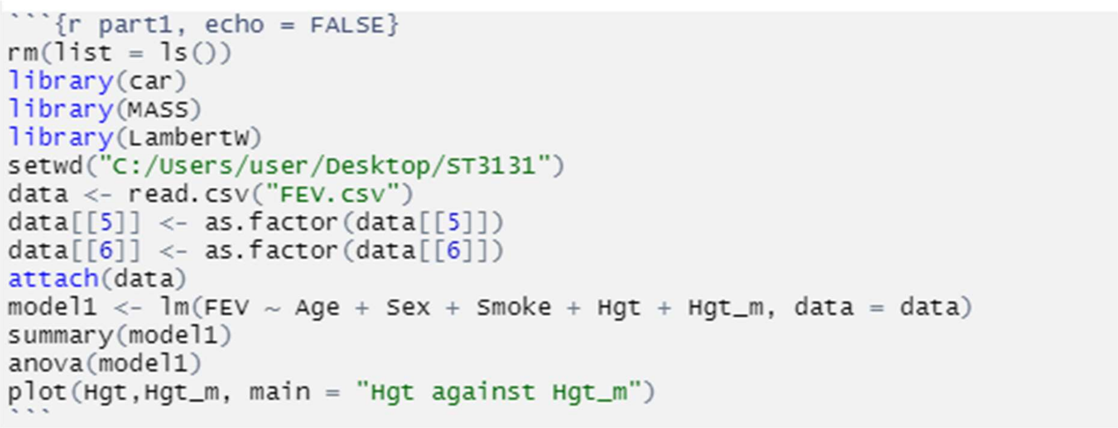
```
##
## Call:
## lm(formula = log(FEV) ~ age_new + Sex + Smoke + hgt_new, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.63443 -0.08644  0.01167  0.09492  0.40904
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.905221   0.008657 104.571  < 2e-16 ***
## age_new      0.823228   0.117836   6.986 7.01e-12 ***
## Sex1         0.029236   0.011716   2.496   0.0128 *
## Smoke1      -0.046015   0.020905  -2.201   0.0281 *
## hgt_new      2.907738   0.114082  25.488  < 2e-16 ***
## ---
## 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
```

```
## Analysis of Variance Table
## Response: log(FEV)
##            Df Sum Sq Mean Sq  F value    Pr(>F)
## age_new     1 43.192  43.192 2042.151 < 2.2e-16 ***
## Sex         1  1.568   1.568   74.158 < 2.2e-16 ***
## Smoke       1  0.262   0.262   12.365 0.0004679 ***
## hgt_new     1 13.740  13.740  649.647 < 2.2e-16 ***
## Residuals 649 13.726   0.021
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

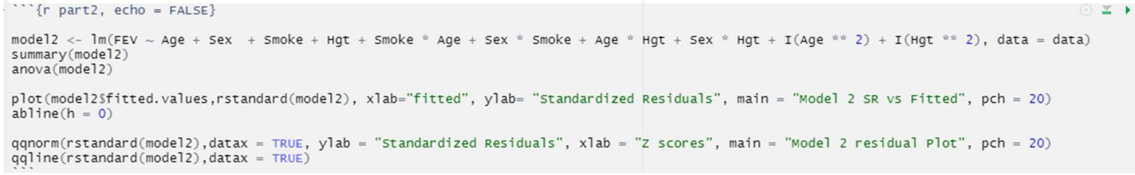
**(Referring to R-code part7)** From the summary, the final model that I will pick is log(FEV) = 0.905221 + 0.823228 * age_new + 0.029236 * Sex - 0.046015 * Smoke + 2.907738 * hgt_new since it is adequate and all the regressors are significant as seen from anova table.

We can interpret the model as follows: If age_new increases by 1 unit, with all variables being constant, FEV increases by (0.029236 * 100) percent. If hgt_new increases by 1 unit, with all variables being constant, FEV increases by (2.907738* 100) percent. With this fitted model, the estimated mean FEV for Sex = 1 (ie male) is (0.029236 * 100) percent more than for Sex = 0 (ie female) With this fitted model, the estimated mean FEV for Smoke = 1 (ie current smoker) is (0.046015 * 100) percent more than for Smoke = 0 (ie current non-smoker).

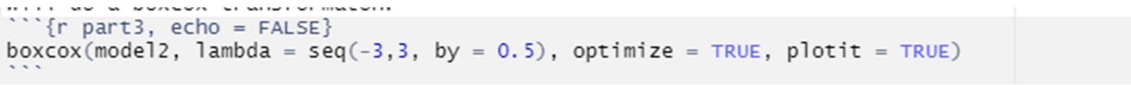
R codes:



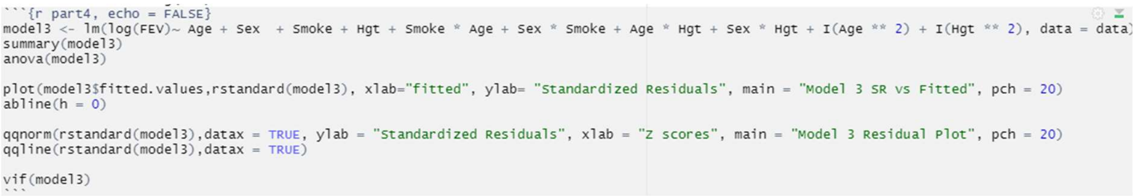
~~~
```{r part1, echo = FALSE}
rm(list = ls())
library(car)
library(MASS)
library(LambertW)
setwd("C:/Users/user/Desktop/ST3131")
data <- read.csv("FEV.csv")
data[[5]] <- as.factor(data[[5]])
data[[6]] <- as.factor(data[[6]])
attach(data)
model1 <- lm(FEV ~ Age + Sex + Smoke + Hgt + Hgt_m, data = data)
summary(model1)
anova(model1)
plot(Hgt,Hgt_m, main = "Hgt against Hgt_m")
```
~~~



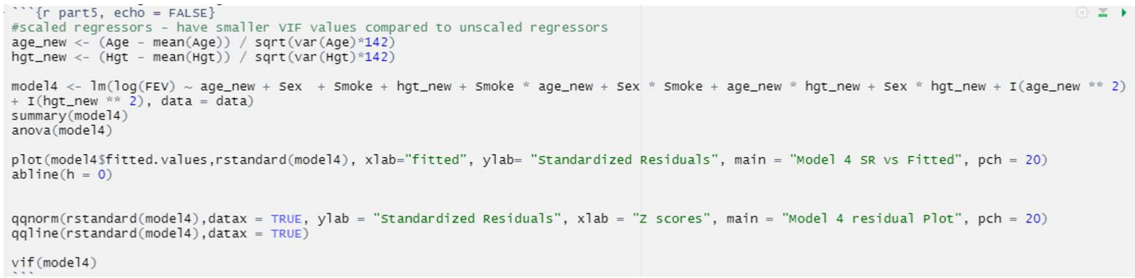
~~~
```{r part2, echo = FALSE}

model2 <- lm(FEV ~ Age + Sex  + Smoke + Hgt + Smoke * Age + Sex * Smoke + Age * Hgt + Sex * Hgt + I(Age ** 2) + I(Hgt ** 2), data = data)
summary(model2)
anova(model2)

plot(model2$fitted.values,rstandard(model2), xlab="fitted", ylab= "Standardized Residuals", main = "Model 2 SR vs Fitted", pch = 20)
abline(h = 0)

qqnorm(rstandard(model2),datax = TRUE, ylab = "Standardized Residuals", xlab = "Z scores", main = "Model 2 residual Plot", pch = 20)
qqline(rstandard(model2),datax = TRUE)
```
~~~



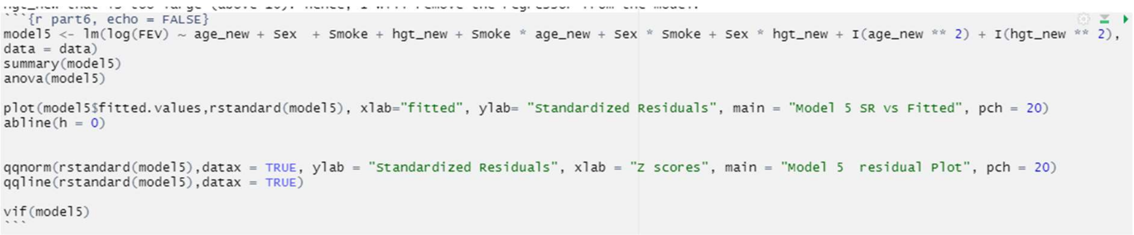
~~~
#### do a boxcox transformation.
```{r part3, echo = FALSE}
boxcox(model2, lambda = seq(-3,3, by = 0.5), optimize = TRUE, plotit = TRUE)
```
~~~



~~~
```{r part4, echo = FALSE}
model3 <- lm(log(FEV)~ Age + Sex  + Smoke + Hgt + Smoke * Age + Sex * Smoke + Age * Hgt + Sex * Hgt + I(Age ** 2) + I(Hgt ** 2), data = data)
summary(model3)
anova(model3)

plot(model3$fitted.values,rstandard(model3), xlab="fitted", ylab= "Standardized Residuals", main = "Model 3 SR vs Fitted", pch = 20)
abline(h = 0)

qqnorm(rstandard(model3),datax = TRUE, ylab = "Standardized Residuals", xlab = "Z scores", main = "Model 3 Residual Plot", pch = 20)
qqline(rstandard(model3),datax = TRUE)

vif(model3)
```
~~~

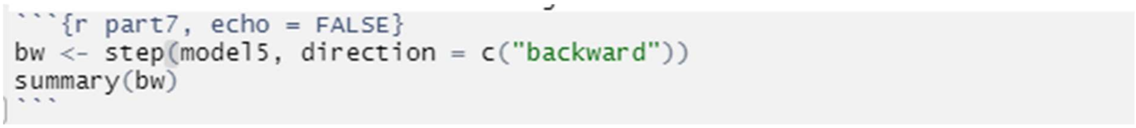


~~~
```{r part5, echo = FALSE}
#scaled regressors - have smaller VIF values compared to unscaled regressors
age_new <- (Age - mean(Age)) / sqrt(var(Age)*142)
hgt_new <- (Hgt - mean(Hgt)) / sqrt(var(Hgt)*142)

model4 <- lm(log(FEV) ~ age_new + Sex  + Smoke + hgt_new + Smoke * age_new + Sex * Smoke + age_new * hgt_new + Sex * hgt_new + I(age_new ** 2)
+ I(hgt_new ** 2), data = data)
summary(model4)
anova(model4)

plot(model4$fitted.values,rstandard(model4), xlab="fitted", ylab= "Standardized Residuals", main = "Model 4 SR vs Fitted", pch = 20)
abline(h = 0)


qqnorm(rstandard(model4),datax = TRUE, ylab = "Standardized Residuals", xlab = "Z scores", main = "Model 4 residual Plot", pch = 20)
qqline(rstandard(model4),datax = TRUE)

vif(model4)
```
~~~

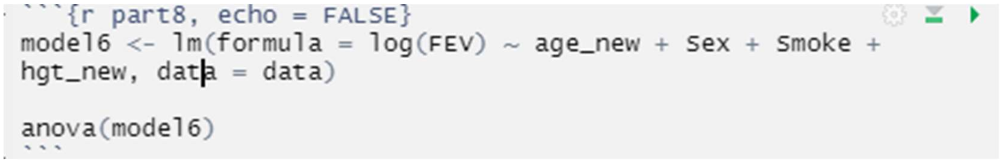


hgt_new that is too large (above 10). Hence, I will remove the regressor from the model.

~~~
```{r part6, echo = FALSE}
model5 <- lm(log(FEV) ~ age_new + Sex  + Smoke + hgt_new + Smoke * age_new + Sex * Smoke + Sex * hgt_new + I(age_new ** 2) + I(hgt_new ** 2),
data = data)
summary(model5)
anova(model5)

plot(model5$fitted.values,rstandard(model5), xlab="fitted", ylab= "Standardized Residuals", main = "Model 5 SR vs Fitted", pch = 20)
abline(h = 0)


qqnorm(rstandard(model5),datax = TRUE, ylab = "Standardized Residuals", xlab = "Z scores", main = "Model 5  residual Plot", pch = 20)
qqline(rstandard(model5),datax = TRUE)

vif(model5)
```
~~~



~~~
```{r part7, echo = FALSE}
bw <- step(model5, direction = c("backward"))
summary(bw)
```
~~~



~~~
```{r part8, echo = FALSE}
model6 <- lm(formula = log(FEV) ~ age_new + Sex + Smoke +
hgt_new, data = data)

anova(model6)
```
~~~