Exploring response variable

Relationship between Regressor and Response

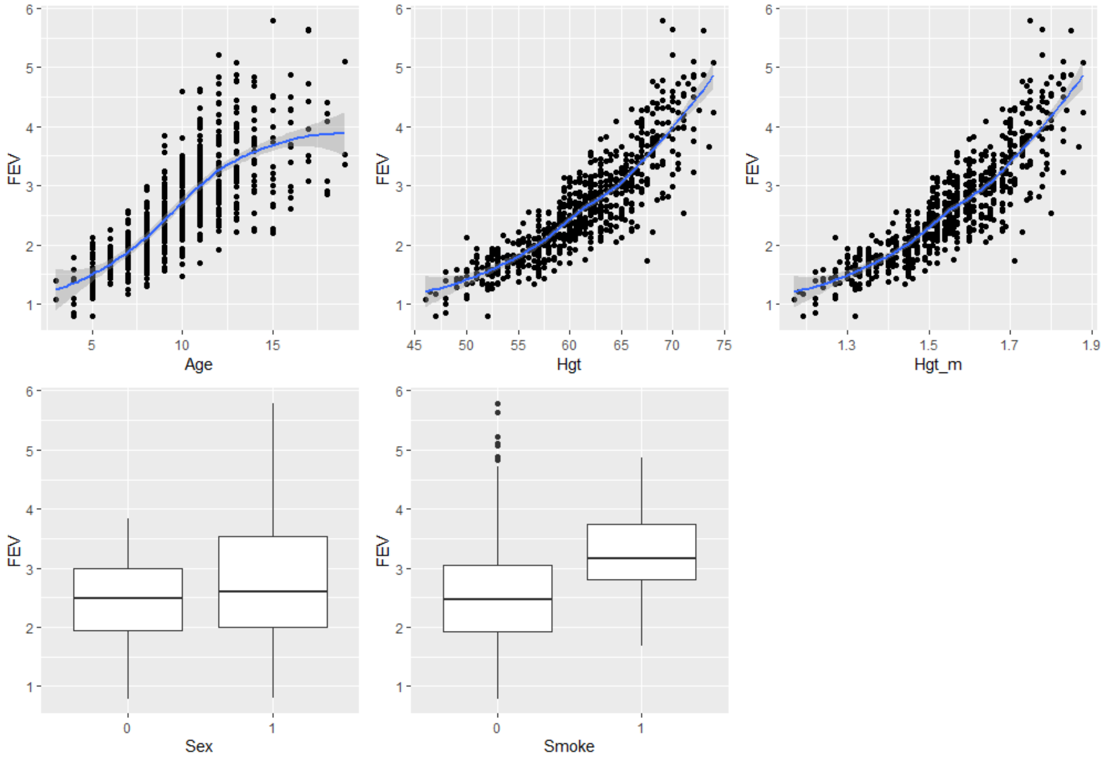


Fig 1: Plot of regressors against Response

FEV, Height and Height_m are continuous, numeric variables.
From the scatter plots in Fig 1, the relationship between Height and Height(meters) and FEV appear non-linear. The regressors may be transformed to a higher order term instead. There appears to be some points that are far from the rest, especially at minimum and maximum height values, indicating the possibility of outliers. The range of height values are (46,74)

Age is a discrete numeric variable.
The relationship between Age and FEV appears to not be linearily related, a higher order regressor term may be used to ensure a better fit. There appears to be some points that are far from the rest, especially at maximum age values, indicating the possibility of outliers. The range of Age values are (3, 19)
For Sex and Smoke, which are discrete categorical variables, and as factors in R, a boxplot was used to compare their distributions instead.
For Sex, males on average appear to have a slightly higher mean FEV than females, with heavier tails.
For Smoke, smokers on have a mean 0.6 units FEV higher than the mean FEV for non smokers, however, with lighter tails.

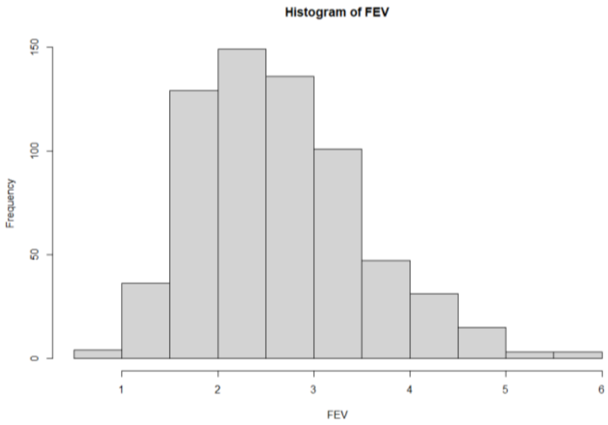


Fig 2: Histogram of Response Values

Referring to Fig2, the range of the response values (FEV) is (0.79, 5.79). The distribution is not symmetric and is right-skewed. There does not appear to be outliers.

Relationships between regressors

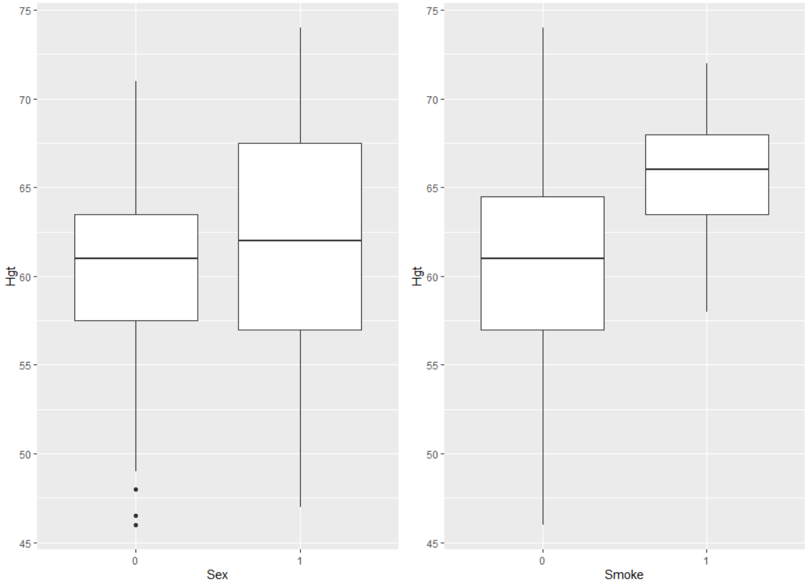
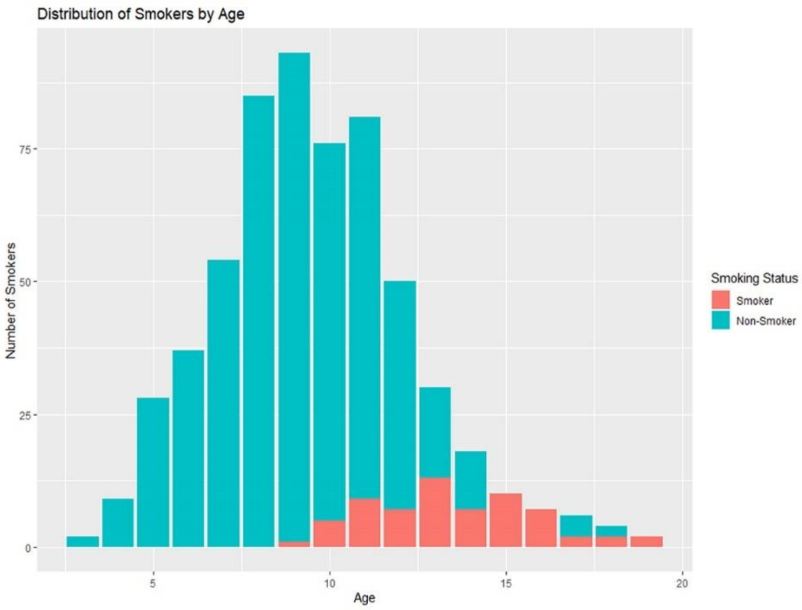
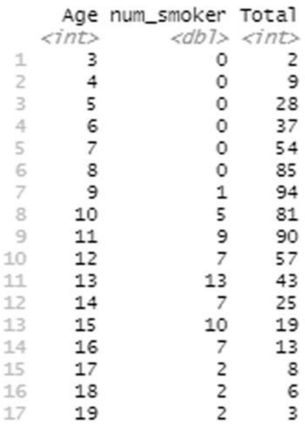


Fig 3: Boxplots of Sex against Hgt and Smoking against Hgt

From fig 3, females are on average shorter than males, and this coincides with our common knowledge. Sex and Height are positively related.
It appears that smokers too are on average taller than non-smokers.





```
     Age num_smoker Total
   <int>      <dbl> <int>
 1     3          0     2
 2     4          0     9
 3     5          0    28
 4     6          0    37
 5     7          0    54
 6     8          0    85
 7     9          1    94
 8    10          5    81
 9    11          9    90
10    12          7    57
11    13         13    43
12    14          7    25
13    15         10    19
14    16          7    13
15    17          2     8
16    18          2     6
17    19          2     3
```

Fig 4: Distribution of Smoker & Non-Smokers , and table of counts of Smokers by Age

Investigating further using Fig 4. we realise that that is because the youngest person to start smoking is a 9 year old. Since people generally only start smoking after 9, and height increases with age, smokers are thus on average taller. Hence we deduce that height and smoking are positively related.

Fitting the full model

The full model was fit with all the regressors chosen, Age Height, Sex, Smoke, Height_meters.

Residual analysis

Plot of residuals against fitted values of full model

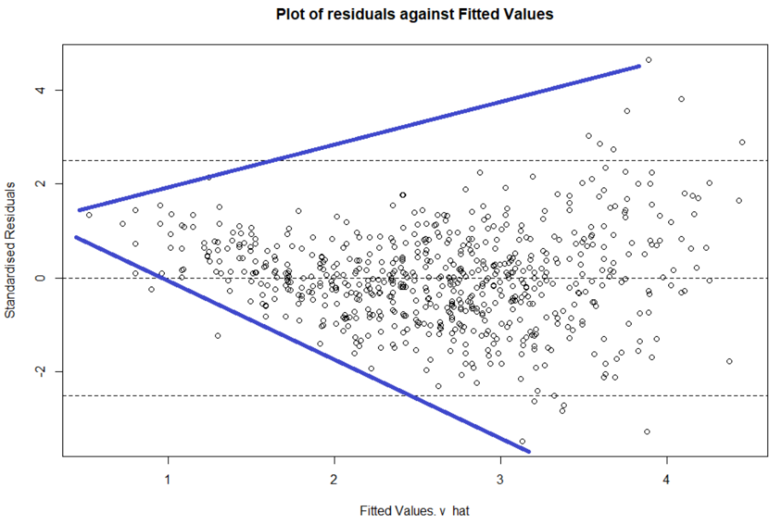
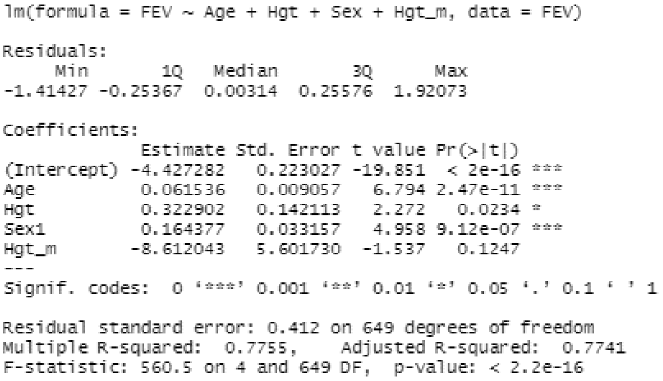


Fig 5: Scatter plot of residuals against fitted values

From the Fig 5, most of the residuals stay within the range of (-2.5, 2.5), however there exists points beyond the range, suggesting the prescence of potential outliers.
The residuals(error terms) have a outward funnel shape, violating the constant error assumption.
As such, the model is deemed to be inadequate.

Testing for variables(T-test) and selection



```
lm(formula = FEV ~ Age + Hgt + Sex + Hgt_m, data = FEV)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.41427 -0.25367  0.00314  0.25576  1.92073 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -4.427282   0.223027 -19.851  < 2e-16 ***
Age          0.061536   0.009057   6.794 2.47e-11 ***
Hgt          0.322902   0.142113   2.272   0.0234 *  
Sex1         0.164377   0.033157   4.958 9.12e-07 ***
Hgt_m       -8.612043   5.601730  -1.537   0.1247    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.412 on 649 degrees of freedom
Multiple R-squared:  0.7755,	Adjusted R-squared:  0.7741 
F-statistic: 560.5 on 4 and 649 DF,  p-value: < 2.2e-16
```

Fig 6 : Summary table of full model

Referring to Fig 6, there is weak evidence against the null for Hgt_m, as its p-value is very low. Since it is merely a different unit of measurement from Hgt and thus will not contribute to the response prediction given that Hgt is already included, it will be removed. We are unable to conclude on the rejection of the Hypothesis test as the significance level is not given.

Fitting Multiple Models

To correct Model Inadequecies(transformation of response/regressor)

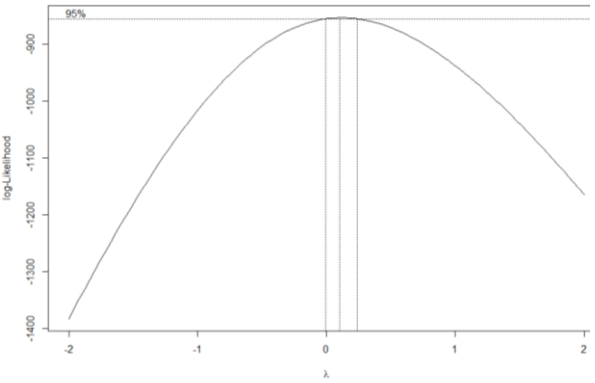
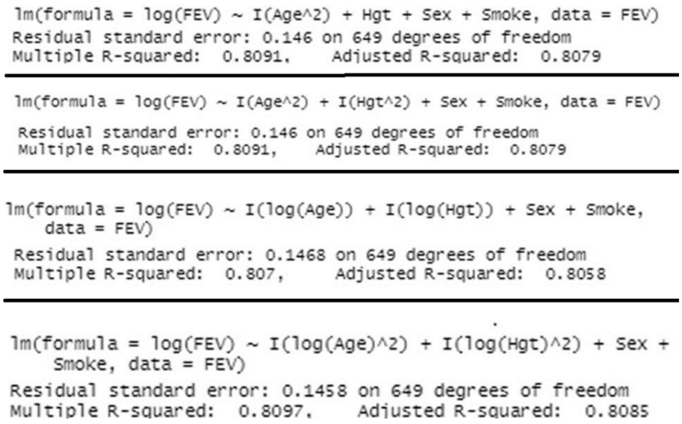


Fig 7: Box-Cox graph

Using box-cox(Fig 7) to decide a transformation for the response variable, the optimal value of lamba was found to be 0. Given this and the previous scatter plot of regressors against the response, a ln transformation of the response is chosen.

Transforming different regressors



```
lm(formula = log(FEV) ~ I(Age^2) + Hgt + Sex + Smoke, data = FEV)
Residual standard error: 0.146 on 649 degrees of freedom
Multiple R-squared:  0.8091,    Adjusted R-squared:  0.8079
```

```
lm(formula = log(FEV) ~ I(Age^2) + I(Hgt^2) + Sex + Smoke, data = FEV)

 Residual standard error: 0.146 on 649 degrees of freedom
 Multiple R-squared:  0.8091,    Adjusted R-squared:  0.8079
```

```
lm(formula = log(FEV) ~ I(log(Age)) + I(log(Hgt)) + Sex + Smoke,
    data = FEV)
 Residual standard error: 0.1468 on 649 degrees of freedom
 Multiple R-squared:  0.807,     Adjusted R-squared:  0.8058
```

```
lm(formula = log(FEV) ~ I(log(Age)^2) + I(log(Hgt)^2) + Sex +
    Smoke, data = FEV)
Residual standard error: 0.1458 on 649 degrees of freedom
Multiple R-squared:  0.8097,    Adjusted R-squared:  0.8085
```

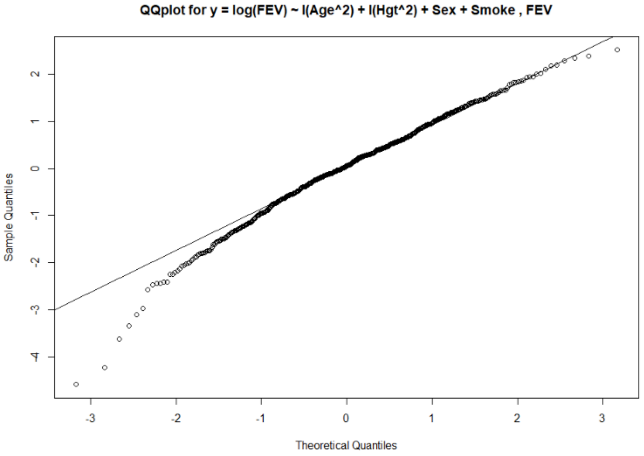


Fig 8 : Summary tables of multiple regressors and

Refering to Fig 8, attempts to transform variables by log, log(x^2), x^2 all resulted in lower adjusted R-squared than model 2 of 0.8095 and were thus rejected. That meant that the goodness of fit of those models was even worse than before transformation. Additionally, plotting the residual plots and qqplots of these models did not prove them to be better.

Plot of residuals against fitted values of model2

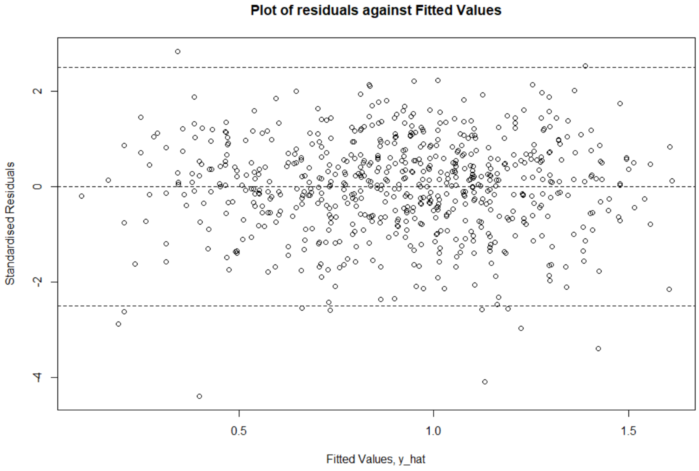


Fig 9 : Plot of residuals against fitted values of model2

From Fig 9, most of the residuals stay within the range of (-2.5, 2.5), however there exists points beyond the range, suggesting the prescence of potential outliers.

The residuals(error terms) are scatter randomly within the channel , and thus satisf the constant variance assumption.

QQplot

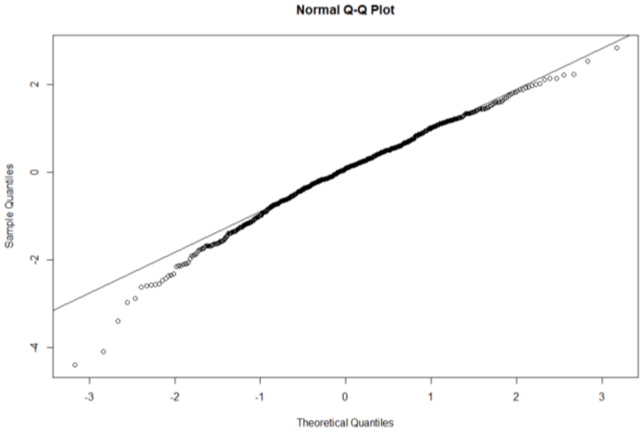


Fig 10 : QQplot

Referring to Fig 10, the qqplot appears to be adequate, however it appears to be right-skewed.

Thus the normality assumption of the residuals are satisfied.

Hence given both the normality assumption and the constant variance assumption is satisfied, the model is adequate.

Outlier detection

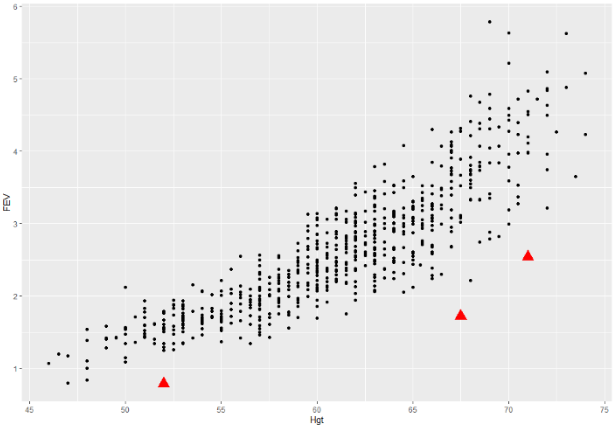
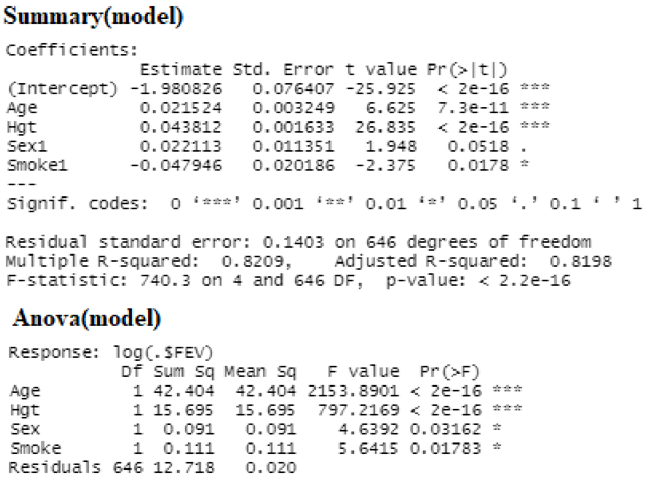


Fig 11: Plot of Height against FEV, with potential outliers highlighted as Red Triangles



Summary(model)

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.980826   0.076407 -25.925  < 2e-16 ***
Age          0.021524   0.003249   6.625  7.3e-11 ***
Hgt          0.043812   0.001633  26.835  < 2e-16 ***
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
```

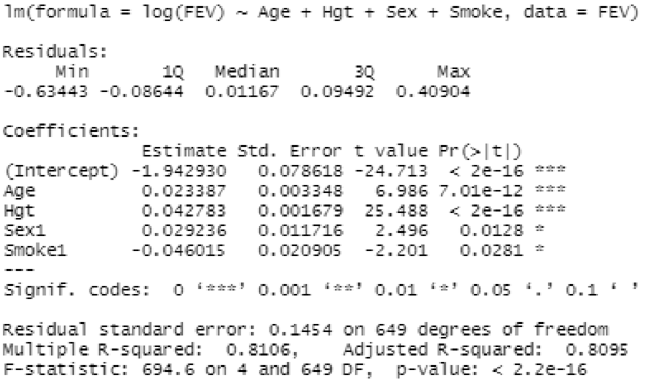
Anova(model)

```
Response: log(.$FEV)
           Df Sum Sq Mean Sq   F value  Pr(>F)    
Age         1 42.404  42.404 2153.8901 < 2e-16 ***
Hgt         1 15.695  15.695  797.2169 < 2e-16 ***
Sex         1  0.091   0.091    4.6392 0.03162 *  
Smoke       1  0.111   0.111    5.6415 0.01783 *  
Residuals 646 12.718   0.020                      
```

Fig 12 : Summary & Anova table of model fitted without the potential outlier points

The points were removed and fitted with the same regressors. The anova and summary tables were taken.
Refering to Fig 12, since the removal of the potential outlier points did not change the coefficient estimates, the R-adjusted or MS_{res} significantly, we consider that they are not overly influential.

F-test for model2



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

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 ***
Age          0.023387   0.003348   6.986 7.01e-12 ***
Hgt          0.042783   0.001679  25.488  < 2e-16 ***
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 13: Summary table of model2

$H_0 = B_1 = B_2 = \ldots = B_5 = 0$ vs H1 : $B_j \neq 0$ for at least one j

Since the p-value for the F-statistic is less than $2.2 \text{ x } 10^{-16}$, data provide strong evidence against the null. The fully fitted model is thus signficant.

The R^2_{adj} is decent at 0.8095, suggesting a decent goodness of fit.

T-test for model2

$H_0 = B_1 = 0$ vs $H_0 = B_1 \neq 0$

$H_0 = B_2 = 0$ vs $H_0 = B_2 \neq 0$

$H_0 = B_3 = 0$ vs $H_0 = B_3 \neq 0$

$H_0 = B_4 = 0$ vs $H_0 = B_4 \neq 0$

Refering to the P-values for the regressors in Fig13, there is weak evidence against the null for all the regressors. Thus they are all significant.We are unable to conclude on the rejection of the Hypothesis test as the significance level is not given.

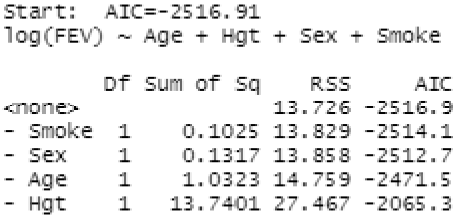
Multicollinearity

VIF = 2.0294506 1.0651047 0.7164495 0.1889952

Since non of the VIF(Variance Inflation Factor) values exceed 5, we deduce that there is no multicollinearity between variables in the model.

The condition number is 10.73811 and agrees with this deduction.

Stepwise Regression for variable selection



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

        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    1    1.0323 14.759 -2471.5
- Hgt    1   13.7401 27.467 -2065.3
```

Fig 14: Stepwise Regression

Referring to Fig 14, the final model selected is the model above of log(FEV) ~ Age + Hgt + Sex + Smoke. This agrees with the leaps function, with the method argument set as adjusted R-squared.

Final Model

Model Description

log(FEV) = -1.94293 + 0.02339 * Age + 0.04278 * Hgt + 0.029248 * I(Sex = 1) - 0.04602 * I(Smoke = 1)

Interpretation

FEV = exp(-1.94293 + 0.02339 * Age + 0.04278 * Hgt + 0.029248 * I(Sex = 1) - 0.04602 * I(Smoke = 1))

Keeping all other regressors constant, males have a mean FEV 1.029672 or exp(0.029248) times higher than females. Keeping all other regressors constant Smokers have a mean FEV 0.9550229 or exp(- 0.04602) of non-smokers.

Keeping all other regressors constant, for every additional year in age, the mean FEV is exp(0.02339) times higher, and respectively for every every additional inch of height, the mean FEV is exp(0.04278) times higher.

Appendix:

```
library(ggplot2)
library(gridExtra)
library(leaps)
library(dplyr)
FEV <- read.csv("../data/FEV.csv")
# account for categorical values
FEV$Sex <- as.factor(FEV$Sex)
FEV$Smoke <- as.factor(FEV$Smoke)
# fit full model
# scatterplot
# plot(FEV$Hgt_m, FEV$FEV)
feeder <- ggplot(FEV)
p1 <- feeder + geom_point(aes(Age,FEV)) + geom_smooth(aes(x= Age, y = FEV), method = "loess", formula =
"y~x")
p2 <- feeder + geom_point(aes(Hgt,FEV)) + geom_smooth(aes(x = Hgt, y = FEV), method = "loess", formula =
"y~x")
p3 <- feeder + geom_point(aes(Hgt_m,FEV)) + geom_smooth(aes(x = Hgt_m, y = FEV), method = "loess",
formula = "y~x")
p4 <- feeder + geom_boxplot(aes(Sex,FEV))
p5 <- feeder + geom_boxplot(aes(Smoke,FEV))
grid.arrange(p1, p2, p3, p4, p5,nrow = 2, ncol = 3)

b1 <- feeder + geom_boxplot(aes(Sex, Hgt))
b2 <- feeder + geom_boxplot(aes(Smoke, Hgt))
grid.arrange(b1,b2, ncol = 2)

FEV %>% group_by(Age) %>% summarise(num_smoker = sum(if_else(Smoke==1, 1, 0)), non_smoker =
n_distinct(ID) - num_smoker) %>% ggplot() +
  geom_col(aes(Age, non_smoker, fill = "green")) + geom_col(aes(Age, num_smoker, fill = "brown")) +
labs(title = "Distribution of Smokers by Age", y = "Number of Smokers") +
  scale_fill_discrete(name = "Smoking Status", labels = c("Smoker","Non-Smoker"))

# equal number of male/female smokers
# FEV %>% group_by(Sex) %>% count(Smoke)
feeder + geom_col(aes(Sex, Smoke))

#========================================
" plot against all regressors + response or only variable against response "
#========================================


#full model
model <- lm(FEV ~ Age + Hgt + Sex + Smoke + Hgt_m, FEV)
plot(model$fitted.values, rstandard(model), xlab = "Fitted Values, y_hat", ylab = "Standardised
Residuals", main = "Plot of residuals against Fitted Values")
abline(h = c(-2.5, 0, 2.5), lty = 2)

#box-cox for regressor transformation
library(MASS)
boxcox(model, lambda=seq(-2, 2, by=0.5),optimize=TRUE,plotit = TRUE)


model2 <- lm(log(FEV) ~ Age + Hgt + Sex + Smoke , FEV)
plot(model2$fitted.values, rstandard(model2), xlab = "Fitted Values, y_hat", ylab = "Standardised
Residuals", main = "Plot of residuals against Fitted Values")
abline(h = c(-2.5, 0, 2.5), lty = 2)
qqnorm(rstandard(model2))
qqline(rstandard(model2))

tmodel <- lm(log(FEV) ~ I(Age^2) + I(Hgt^2) + Sex + Smoke , FEV)
qqnorm(rstandard(tmodel))
qqline(rstandard(tmodel))


########## outlier detection
```

```
plot(model2, which =2)
feeder + geom_point(aes(Hgt,FEV)) + geom_point(data = data.frame(a = Hgt[c(2,140,473)], b
=FEV$FEV[c(2,140,473)]), aes(a,b), color = "Red", size = 5, shape = 17)

# Removal of outliers
drop <- c(2,140,473)
FEV[!seq_len(nrow(FEV)) %in% drop,] %>% NROW()
FEV %>% filter(!row_number() %in% drop) %>% lm(log(.$FEV) ~ Age + Hgt + Sex + Smoke , .) %>% summary()#
anova() ##plot(which = 2)
model2 %>% anova()

####
# multicollinearity
####Hgt, FEV$Sex, FEV$Smoke)
r <- cor(x)
eigen(r)$values

# stepwise regression
x <- FEV %>% select(-ID, -FEV)
y <- FEV$FEV
leaps(x, y , method = "adjr2")


####Stepwise Regression
sw <- step(model2, direction = "both")
summary(sw)
```