Assignment

## R Markdown

Section 1: Exploring Response Variable and their relationship with the regressors In our dataset, there are 6 regressors and 1 response variable. Our starting model will hence be : FEV = β0+β1(Smoke)+β2(Age)+β3(Hgt)+β4(Sex)+β5(Hgt\_m) However, this model can be further modified by adding interaction terms and higher order terms through our model-building process. We observe that our variables Sex and Smoke takes on the values 0 and 1 only. Hence, we treat them as categorical variables and convert them into factors.

library(ggplot2)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

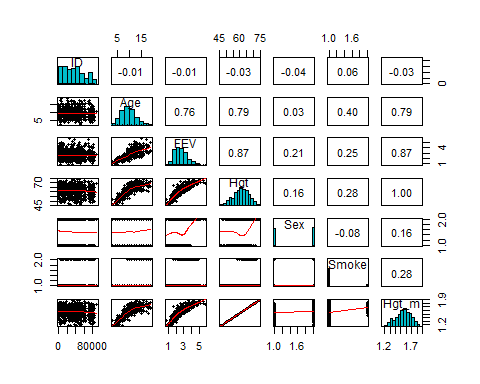
library(psych)

## Warning: package 'psych' was built under R version 4.0.5

##   
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

data<-read.csv("FEV.csv")  
data$Smoke<-as.factor(data$Smoke)  
data$Sex<-as.factor(data$Sex)  
attach(data)  
model<-lm(FEV~ Age + Hgt + Sex + Smoke + Hgt\_m,data = data)  
pairs.panels(data,   
 method = "pearson", # correlation method  
 hist.col = "#00BECC",  
 density = FALSE, # show density plots  
 ellipses = FALSE # show correlation ellipses  
 )



Hgt and Hgt\_m have a high positive correlation to FEV with correlation = 0.87. Since, Hgt\_m and Hgt are same measurements in different units, their correlation is 1. Hence, we can drop either one of them to build our model. We will be dropping Hgt\_m in this instance and our initial model will be of the form: FEV = β0+β1(Smoke)+β2(Age)+β3(Hgt)+β4(Sex)

Following that, Age also has a high correlation value with FEV where the value is 0.76.However, it is vital to note that Age has high correlation with Height (correlation = 0.79). As these two regressors are highly correlated, we will have to check for presence of multicollinearity in order to ensure adequacy of our model in our next step.

Detecting Multicollinearity

# Centering data: Subtract the mean, divide by sqrt(S\_xx)  
# sqrt(S\_xx) = (n-p) \* MS\_res  
# n = 654, p = , n-p = 31  
A <- (Age-mean(Age))/(sqrt(var(Age))\*653)  
H <- (Hgt - mean(Hgt))/(sqrt(var(Hgt))\*653)  
#Hm <- (Hgt\_m - mean(Hgt\_m))/(sqrt(var(Hgt\_m))\*653)  
#S <- (Sex - mean(Sex))/(sqrt(var(Sex))\*653)  
#Sm <- (Smoke - mean(Smoke))/(sqrt(var(Smoke))\*653)  
x <- cbind(A, H )  
x <- cor(x) # Correlation matrix of x, X'X  
x

## A H  
## A 1.0000000 0.7919436  
## H 0.7919436 1.0000000

#condition number is:  
cond <- max(eigen(x)$values)/min(eigen(x)$values)  
#condition indices:  
max(eigen(x)$values)/eigen(x)$values

## [1] 1.000000 8.612778

C<-solve(x) #this is (X'X)^(-1) where X'X is in correlation form  
VIF <- diag(C)   
VIF

## A H   
## 2.682221 2.682221

The values of VIF are 2.682221, 2.682221 Since the condition number = 8.6127 (rounded to 5sf) is lesser than 100 and values of VIF are below 10, we can conclude that there is no serious multi collinearity in our data.

Fitting our initial model: FEV = β0+β1(Smoke)+β2(Age)+β3(Hgt)+β4(Sex)

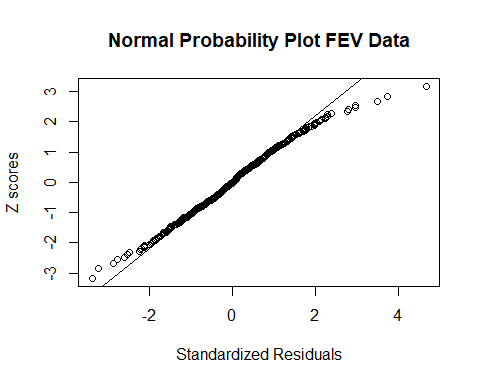
model<-lm(FEV ~ Smoke + Age + Hgt + Sex ,data=data)  
summary(model)

##   
## Call:  
## lm(formula = FEV ~ Smoke + Age + Hgt + Sex, 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 \*\*\*  
## Smoke1 -0.086846 0.059235 -1.466 0.143   
## Age 0.065510 0.009486 6.906 1.19e-11 \*\*\*  
## Hgt 0.104182 0.004756 21.904 < 2e-16 \*\*\*  
## Sex1 0.156909 0.033197 4.727 2.80e-06 \*\*\*  
## ---  
## 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

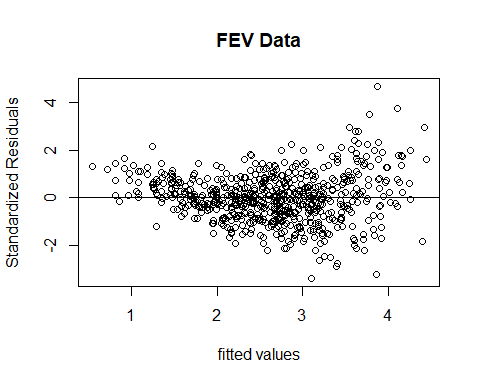
anova(model)

## Analysis of Variance Table  
##   
## Response: FEV  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Smoke 1 29.598 29.598 174.292 < 2.2e-16 \*\*\*  
## Age 1 253.425 253.425 1492.352 < 2.2e-16 \*\*\*  
## Hgt 1 93.739 93.739 552.003 < 2.2e-16 \*\*\*  
## Sex 1 3.794 3.794 22.341 2.801e-06 \*\*\*  
## Residuals 649 110.211 0.170   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

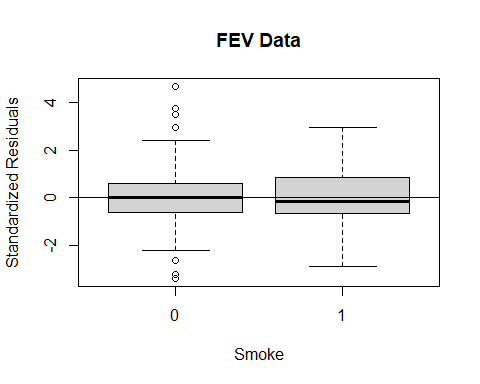
#Normal Probabiltiy plot  
qqnorm(rstandard(model),datax = TRUE, ylab = "Standardized Residuals", xlab = "Z scores", main = "Normal Probability Plot FEV Data")  
qqline(rstandard(model),datax = TRUE)



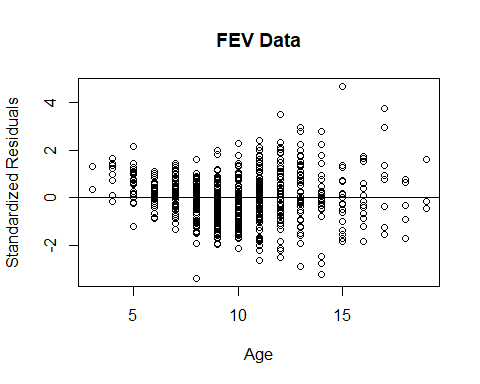
#standardized residuals vs fitted:  
plot(model$fitted.values,rstandard(model), xlab="fitted values", ylab= "Standardized Residuals", main = "FEV Data")  
abline(h=0)



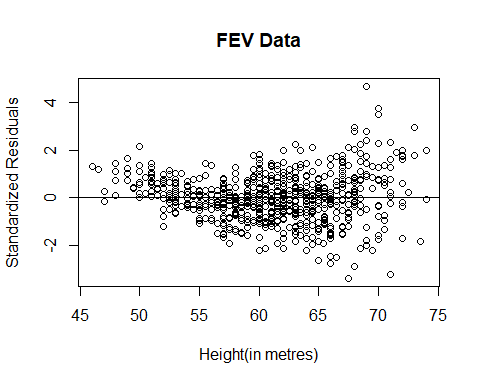
#standardized residuals vs Smoke:  
plot(Smoke,rstandard(model), xlab="Smoke", ylab= "Standardized Residuals", main = "FEV Data")  
abline(h=0)



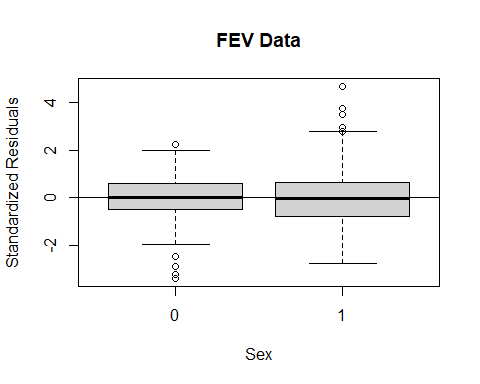
#standardized residuals vs Age:  
plot(Age,rstandard(model), xlab="Age", ylab= "Standardized Residuals", main = "FEV Data")  
abline(h=0)



#standardized residuals vs Height:  
plot(Hgt,rstandard(model), xlab="Height(in metres)", ylab= "Standardized Residuals", main = "FEV Data")  
abline(h=0)



#standardized residuals vs Sex:  
plot(Sex,rstandard(model), xlab="Sex", ylab= "Standardized Residuals", main = "FEV Data")  
abline(h=0)

 Conducting linear regression on this model, we obtain a R2 value of 0.7754 with F-Statistic of 560.2 with 4 and 649 df, and a p-value smaller than 2.2e-16. Overall, the model is significant. Hgt is observed to help in predicting FEV as seen in its statistically significant non-zero estimated coefficient, βˆ3. (P-value <2e-16 for the t-test)

However, it may not be the most ideal model and we can attempt to get a better model.

Comments on the plots obtained

For Scatter plots of Standardized Residuals (SR) vs Age, the pattern inidicates non-constant variance (heterogeneity) and for the scatter plot of SR vs Height, we see an outward funnel shape indicating non-constant variance also This suggests a transformation of the data that stabilizes the variance may be needed.

The normal probability plot indicates a heavy-tailed distribution deviating from our normality assumption. This may also indicate that there may be one or more outliers in the data. Moreover, the plot of standardised residuals vs fitted values has an outward funnel shape indicating a pattern and non-constant variance which is not favourable to us. Since, the normality assumptions are not fully satisfied, we adopt the box cox method to do a transformation on the response(FEV)

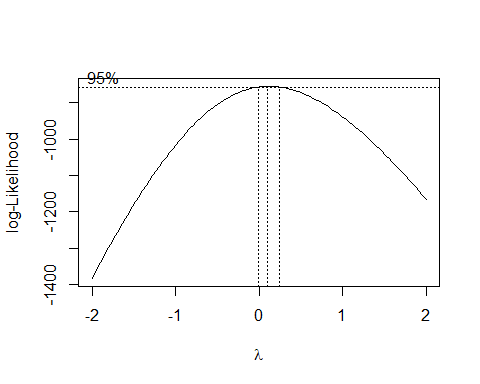
library(MASS)

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

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

## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :  
## extra argument 'optimize' will be disregarded

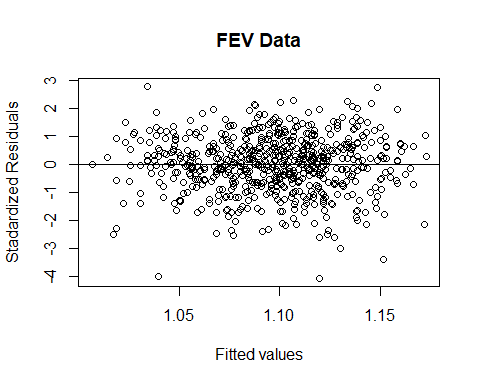


Through the box-cox method we can set our lambda to be 0.1. Hence, we can transform FEV to FEV^0.1 We then fit our model with this new transformed response variable.

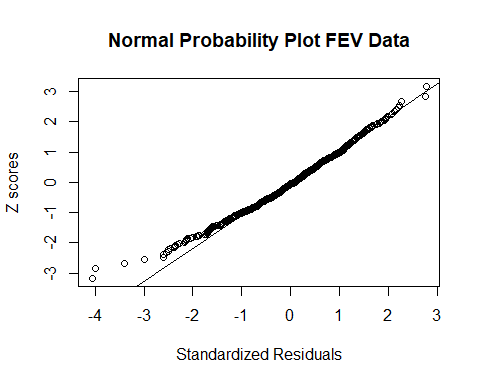
model\_tm<-lm(FEV^0.1 ~ Smoke + Age + Hgt + Sex ,data=data)  
summary(model\_tm)

##   
## Call:  
## lm(formula = FEV^0.1 ~ Smoke + Age + Hgt + Sex, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.06406 -0.00949 0.00131 0.01019 0.04397   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.7847784 0.0085932 91.326 < 2e-16 \*\*\*  
## Smoke1 -0.0048735 0.0022849 -2.133 0.03331 \*   
## Age 0.0025801 0.0003659 7.051 4.55e-12 \*\*\*  
## Hgt 0.0046574 0.0001835 25.385 < 2e-16 \*\*\*  
## Sex1 0.0035181 0.0012805 2.747 0.00617 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.0159 on 649 degrees of freedom  
## Multiple R-squared: 0.8106, Adjusted R-squared: 0.8094   
## F-statistic: 694.5 on 4 and 649 DF, p-value: < 2.2e-16

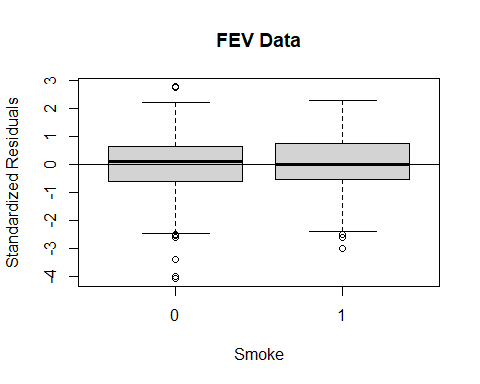
plot(model\_tm$fitted.values,rstandard(model\_tm),xlab = "Fitted values",   
ylab = "Stadardized Residuals",main = "FEV Data")  
abline(h = 0) ###



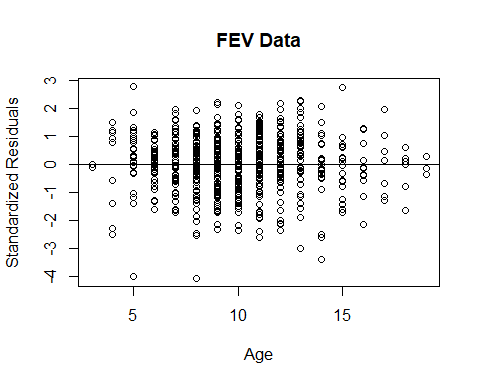
#Normal Probabiltiy plot  
qqnorm(rstandard(model\_tm),datax = TRUE, ylab = "Standardized Residuals", xlab = "Z scores", main = "Normal Probability Plot FEV Data")  
qqline(rstandard(model\_tm),datax = TRUE)



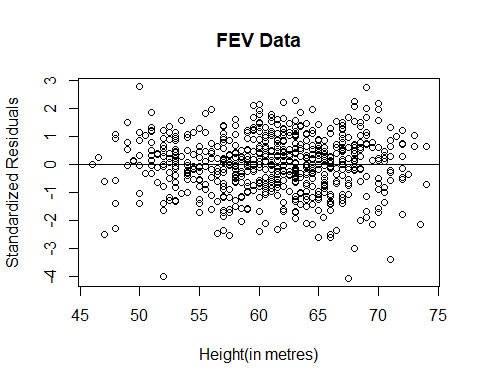
#standardized residuals vs Smoke:  
plot(Smoke,rstandard(model\_tm), xlab="Smoke", ylab= "Standardized Residuals", main = "FEV Data")  
abline(h=0)



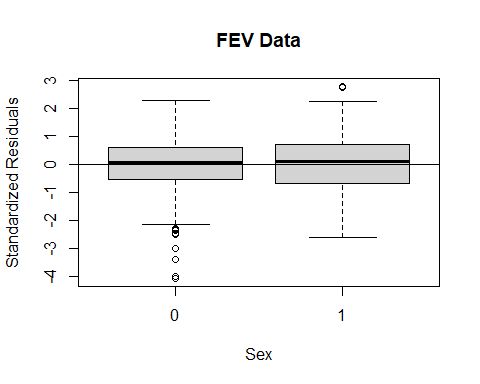
#standardized residuals vs Age:  
plot(Age,rstandard(model\_tm), xlab="Age", ylab= "Standardized Residuals", main = "FEV Data")  
abline(h=0)



#standardized residuals vs Height:  
plot(Hgt,rstandard(model\_tm), xlab="Height(in metres)", ylab= "Standardized Residuals", main = "FEV Data")  
abline(h=0)



#standardized residuals vs Sex:  
plot(Sex,rstandard(model\_tm), xlab="Sex", ylab= "Standardized Residuals", main = "FEV Data")  
abline(h=0)

 After transforming our response variable (FEV), we see that the points in the plot of standardised residuals vs fitted values are now random and no longer have a funnel shape as seen before. The points fall within the band of -4 to 3. The plot of SR vs Height is also more random with our new fitted model. By looking at the summary(model\_tm), we observe an increase in adjusted R square value compared to initial model where the adjusted R square value of initial model is 0.774 while in our model\_tm it is 0.8094. This suggests a better fit for the model. However, as observed from our QQ plot many points lie closer to the qqline but yet some still seem to deviate from the qqline which is not yet very ideal.

We then go on to add interaction terms to further investigate our model. We will call this model model\_int where it is:

FEV^(0.1) = β0+β1(Age)+β2(Hgt)+β3(Sex)+β4(Smoke)+β5(Hgt\_m)+β6(Smoke*Age)+β7(Smoke*Hgt)+β8(Smoke*Sex)+β9(Age*Hgt)+β10(Age*Sex)+β11(Hgt*Sex)

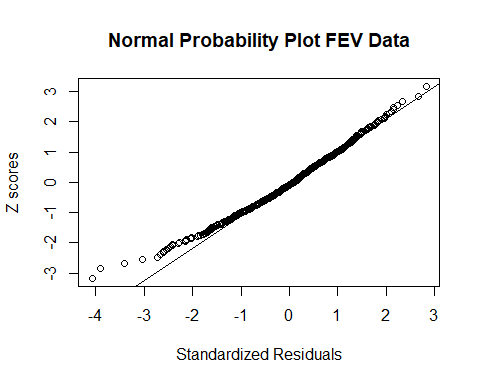
model\_int<-lm(FEV^0.1 ~ Smoke + Age + Hgt + Sex + Smoke\*Age+Smoke\*Hgt+Smoke\*Sex+Age\*Hgt+Age\*Sex+Hgt\*Sex,data=data)  
summary(model\_int)

##   
## Call:  
## lm(formula = FEV^0.1 ~ Smoke + Age + Hgt + Sex + Smoke \* Age +   
## Smoke \* Hgt + Smoke \* Sex + Age \* Hgt + Age \* Sex + Hgt \*   
## Sex, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.064416 -0.009319 0.001334 0.010602 0.044476   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.076e-01 2.341e-02 34.503 <2e-16 \*\*\*  
## Smoke1 -1.205e-02 5.086e-02 -0.237 0.813   
## Age 2.037e-03 2.606e-03 0.782 0.435   
## Hgt 4.221e-03 4.266e-04 9.895 <2e-16 \*\*\*  
## Sex1 -1.240e-02 1.876e-02 -0.661 0.509   
## Smoke1:Age -1.574e-03 9.696e-04 -1.624 0.105   
## Smoke1:Hgt 4.188e-04 7.992e-04 0.524 0.600   
## Smoke1:Sex1 8.986e-04 5.369e-03 0.167 0.867   
## Age:Hgt 1.480e-05 4.127e-05 0.359 0.720   
## Age:Sex1 -1.321e-04 7.542e-04 -0.175 0.861   
## Hgt:Sex1 2.818e-04 3.876e-04 0.727 0.467   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.01592 on 643 degrees of freedom  
## Multiple R-squared: 0.8118, Adjusted R-squared: 0.8089   
## F-statistic: 277.3 on 10 and 643 DF, p-value: < 2.2e-16

anova(model\_int)

## Analysis of Variance Table  
##   
## Response: FEV^0.1  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Smoke 1 0.05215 0.05215 205.7571 < 2.2e-16 \*\*\*  
## Age 1 0.46880 0.46880 1849.5199 < 2.2e-16 \*\*\*  
## Hgt 1 0.17906 0.17906 706.4249 < 2.2e-16 \*\*\*  
## Sex 1 0.00191 0.00191 7.5246 0.006256 \*\*   
## Smoke:Age 1 0.00044 0.00044 1.7383 0.187820   
## Smoke:Hgt 1 0.00020 0.00020 0.7833 0.376453   
## Smoke:Sex 1 0.00004 0.00004 0.1689 0.681191   
## Age:Hgt 1 0.00012 0.00012 0.4857 0.486084   
## Age:Sex 1 0.00007 0.00007 0.2754 0.599931   
## Hgt:Sex 1 0.00013 0.00013 0.5286 0.467447   
## Residuals 643 0.16298 0.00025   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

qqnorm(rstandard(model\_int),datax = TRUE, ylab = "Standardized Residuals", xlab = "Z scores", main = "Normal Probability Plot FEV Data")  
qqline(rstandard(model\_int),datax = TRUE)

 model\_int has a slightly lower R squared value of 0.8089 as compared to model\_tm which had a adjusted R square value of 0.8094.The F-statistic with p-value less than 2.2e-16 also implies that the model is statistically significant. We then perform step-wise regression to decide which variables to keep.

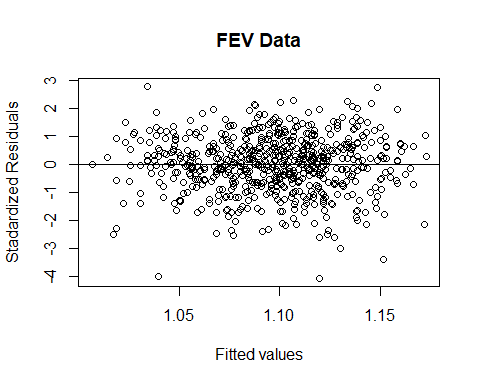
model\_new <- step(model\_int, direction = c("both"))

## Start: AIC=-5404.38  
## FEV^0.1 ~ Smoke + Age + Hgt + Sex + Smoke \* Age + Smoke \* Hgt +   
## Smoke \* Sex + Age \* Hgt + Age \* Sex + Hgt \* Sex  
##   
## Df Sum of Sq RSS AIC  
## - Smoke:Sex 1 0.00000710 0.16299 -5406.4  
## - Age:Sex 1 0.00000778 0.16299 -5406.4  
## - Age:Hgt 1 0.00003261 0.16301 -5406.3  
## - Smoke:Hgt 1 0.00006960 0.16305 -5406.1  
## - Hgt:Sex 1 0.00013399 0.16312 -5405.8  
## <none> 0.16298 -5404.4  
## - Smoke:Age 1 0.00066815 0.16365 -5403.7  
##   
## Step: AIC=-5406.36  
## FEV^0.1 ~ Smoke + Age + Hgt + Sex + Smoke:Age + Smoke:Hgt + Age:Hgt +   
## Age:Sex + Hgt:Sex  
##   
## Df Sum of Sq RSS AIC  
## - Age:Sex 1 0.00000495 0.16299 -5408.3  
## - Age:Hgt 1 0.00002918 0.16302 -5408.2  
## - Smoke:Hgt 1 0.00012161 0.16311 -5407.9  
## - Hgt:Sex 1 0.00013544 0.16312 -5407.8  
## <none> 0.16299 -5406.4  
## - Smoke:Age 1 0.00066434 0.16365 -5405.7  
## + Smoke:Sex 1 0.00000710 0.16298 -5404.4  
##   
## Step: AIC=-5408.34  
## FEV^0.1 ~ Smoke + Age + Hgt + Sex + Smoke:Age + Smoke:Hgt + Age:Hgt +   
## Hgt:Sex  
##   
## Df Sum of Sq RSS AIC  
## - Age:Hgt 1 0.00002644 0.16302 -5410.2  
## - Smoke:Hgt 1 0.00011672 0.16311 -5409.9  
## - Hgt:Sex 1 0.00023359 0.16323 -5409.4  
## <none> 0.16299 -5408.3  
## - Smoke:Age 1 0.00065946 0.16365 -5407.7  
## + Age:Sex 1 0.00000495 0.16299 -5406.4  
## + Smoke:Sex 1 0.00000427 0.16299 -5406.4  
##   
## Step: AIC=-5410.23  
## FEV^0.1 ~ Smoke + Age + Hgt + Sex + Smoke:Age + Smoke:Hgt + Hgt:Sex  
##   
## Df Sum of Sq RSS AIC  
## - Smoke:Hgt 1 0.00014910 0.16317 -5411.6  
## - Hgt:Sex 1 0.00033124 0.16335 -5410.9  
## <none> 0.16302 -5410.2  
## - Smoke:Age 1 0.00063979 0.16366 -5409.7  
## + Age:Hgt 1 0.00002644 0.16299 -5408.3  
## + Smoke:Sex 1 0.00000238 0.16302 -5408.2  
## + Age:Sex 1 0.00000221 0.16302 -5408.2  
##   
## Step: AIC=-5411.63  
## FEV^0.1 ~ Smoke + Age + Hgt + Sex + Smoke:Age + Hgt:Sex  
##   
## Df Sum of Sq RSS AIC  
## - Hgt:Sex 1 0.00038070 0.16355 -5412.1  
## <none> 0.16317 -5411.6  
## - Smoke:Age 1 0.00052995 0.16370 -5411.5  
## + Smoke:Hgt 1 0.00014910 0.16302 -5410.2  
## + Age:Hgt 1 0.00005882 0.16311 -5409.9  
## + Smoke:Sex 1 0.00005589 0.16311 -5409.9  
## + Age:Sex 1 0.00000129 0.16317 -5409.6  
##   
## Step: AIC=-5412.11  
## FEV^0.1 ~ Smoke + Age + Hgt + Sex + Smoke:Age  
##   
## Df Sum of Sq RSS AIC  
## - Smoke:Age 1 0.000441 0.16399 -5412.3  
## <none> 0.16355 -5412.1  
## + Hgt:Sex 1 0.000381 0.16317 -5411.6  
## + Age:Sex 1 0.000259 0.16329 -5411.1  
## + Age:Hgt 1 0.000201 0.16335 -5410.9  
## + Smoke:Hgt 1 0.000199 0.16335 -5410.9  
## + Smoke:Sex 1 0.000164 0.16339 -5410.8  
## - Sex 1 0.002069 0.16562 -5405.9  
## - Hgt 1 0.143280 0.30683 -5002.6  
##   
## Step: AIC=-5412.35  
## FEV^0.1 ~ Smoke + Age + Hgt + Sex  
##   
## Df Sum of Sq RSS AIC  
## <none> 0.16399 -5412.3  
## + Smoke:Age 1 0.000441 0.16355 -5412.1  
## + Hgt:Sex 1 0.000291 0.16370 -5411.5  
## + Age:Sex 1 0.000205 0.16379 -5411.2  
## + Age:Hgt 1 0.000102 0.16389 -5410.8  
## + Smoke:Sex 1 0.000088 0.16390 -5410.7  
## + Smoke:Hgt 1 0.000069 0.16392 -5410.6  
## - Smoke 1 0.001149 0.16514 -5409.8  
## - Sex 1 0.001907 0.16590 -5406.8  
## - Age 1 0.012564 0.17655 -5366.1  
## - Hgt 1 0.162833 0.32682 -4963.3

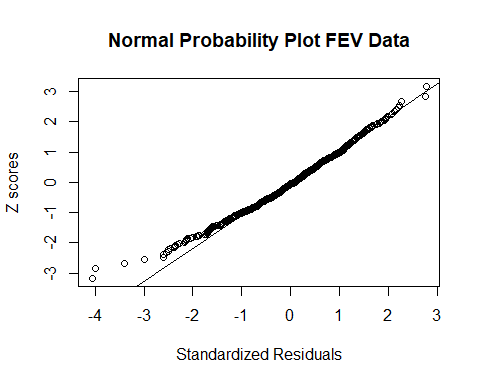
summary(model\_new)

##   
## Call:  
## lm(formula = FEV^0.1 ~ Smoke + Age + Hgt + Sex, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.06406 -0.00949 0.00131 0.01019 0.04397   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.7847784 0.0085932 91.326 < 2e-16 \*\*\*  
## Smoke1 -0.0048735 0.0022849 -2.133 0.03331 \*   
## Age 0.0025801 0.0003659 7.051 4.55e-12 \*\*\*  
## Hgt 0.0046574 0.0001835 25.385 < 2e-16 \*\*\*  
## Sex1 0.0035181 0.0012805 2.747 0.00617 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.0159 on 649 degrees of freedom  
## Multiple R-squared: 0.8106, Adjusted R-squared: 0.8094   
## F-statistic: 694.5 on 4 and 649 DF, p-value: < 2.2e-16

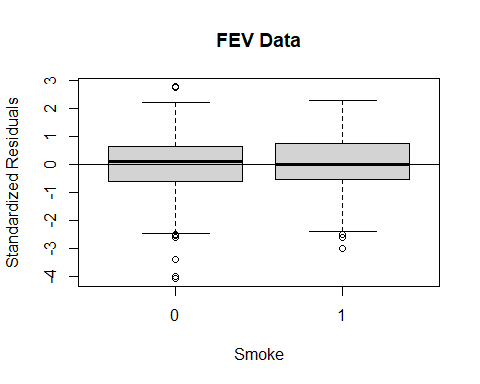
plot(model\_new$fitted.values,rstandard(model\_new),xlab = "Fitted values",   
ylab = "Stadardized Residuals",main = "FEV Data")  
abline(h = 0) ###



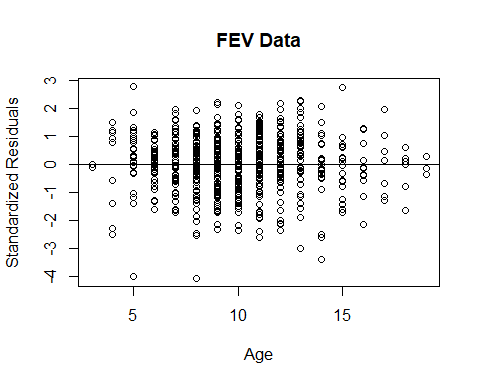
#Normal Probabiltiy plot  
qqnorm(rstandard(model\_new),datax = TRUE, ylab = "Standardized Residuals", xlab = "Z scores", main = "Normal Probability Plot FEV Data")  
qqline(rstandard(model\_new),datax = TRUE)



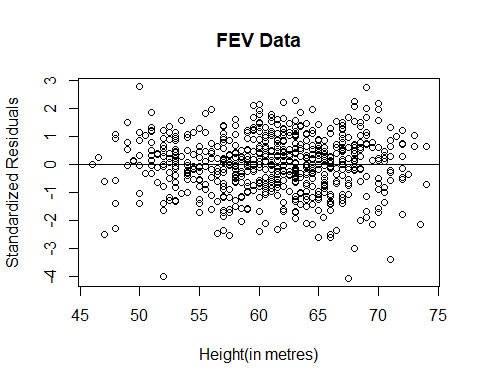
#standardized residuals vs Smoke:  
plot(Smoke,rstandard(model\_new), xlab="Smoke", ylab= "Standardized Residuals", main = "FEV Data")  
abline(h=0)



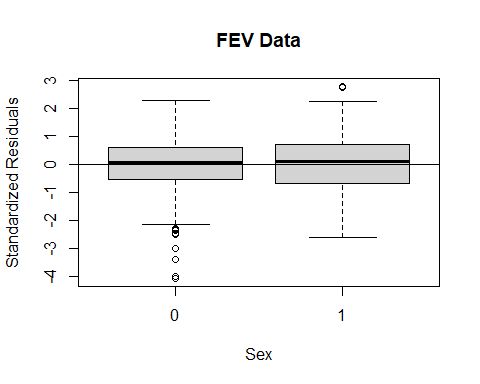
#standardized residuals vs Age:  
plot(Age,rstandard(model\_new), xlab="Age", ylab= "Standardized Residuals", main = "FEV Data")  
abline(h=0)



#standardized residuals vs Height:  
plot(Hgt,rstandard(model\_new), xlab="Height(in metres)", ylab= "Standardized Residuals", main = "FEV Data")  
abline(h=0)



#standardized residuals vs Sex:  
plot(Sex,rstandard(model\_new), xlab="Sex", ylab= "Standardized Residuals", main = "FEV Data")  
abline(h=0)

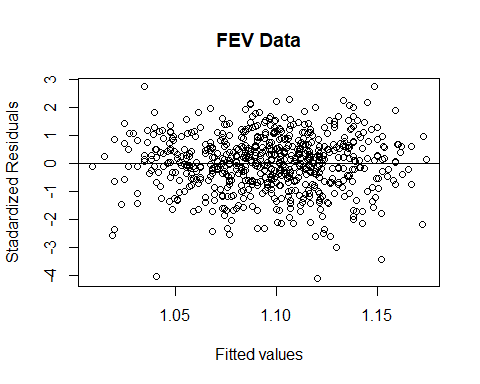
 Smaller AIC Value is preferred and hence our resulting model is given by the output in R as: FEV^(0.1) ~ Smoke + Age + Hgt + Sex

With reference to the scatterplots in the introduction, we notice a non-linear relationship between FEV and Height. From Topic 5 (slide 6), we transform height variable into log(Hgt) to perform a suitable transformation.

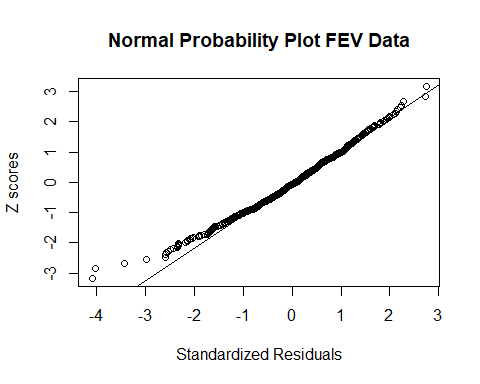
# For the Age variable  
model3 <- lm(FEV^0.1 ~ Smoke + I(Age \* log(Age)) + Hgt+ Sex, data = data)  
gamma1 <- model3$coefficients["I(Age \* log(Age))"]  
beta1 <- model\_tm$coefficients["Age"]  
power1 <- as.numeric((gamma1/beta1) + 1)  
#power1  
model3 <- lm(FEV\*\*0.1~ Smoke + I(Age\*\*1.290987) + Hgt+ Sex, data = data)  
summary(model3)

##   
## Call:  
## lm(formula = FEV^0.1 ~ Smoke + I(Age^1.290987) + Hgt + Sex, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.064569 -0.009708 0.001205 0.010201 0.043431   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.7874427 0.0088280 89.198 < 2e-16 \*\*\*  
## Smoke1 -0.0050712 0.0022941 -2.210 0.02742 \*   
## I(Age^1.290987) 0.0009804 0.0001393 7.038 4.98e-12 \*\*\*  
## Hgt 0.0047184 0.0001770 26.650 < 2e-16 \*\*\*  
## Sex1 0.0034136 0.0012789 2.669 0.00779 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.0159 on 649 degrees of freedom  
## Multiple R-squared: 0.8106, Adjusted R-squared: 0.8094   
## F-statistic: 694.2 on 4 and 649 DF, p-value: < 2.2e-16

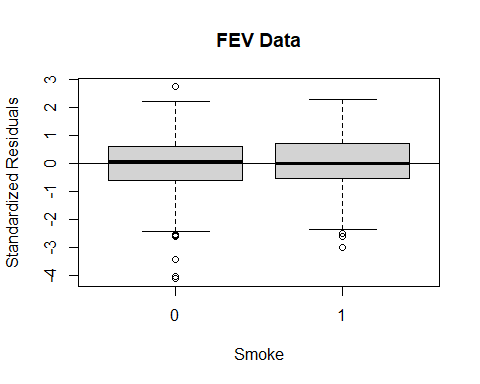
plot(model3$fitted.values,rstandard(model3),xlab = "Fitted values",   
ylab = "Stadardized Residuals",main = "FEV Data")  
abline(h = 0) ###



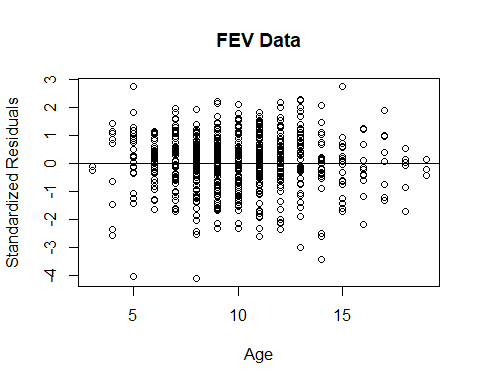
#Normal Probabiltiy plot  
qqnorm(rstandard(model3),datax = TRUE, ylab = "Standardized Residuals", xlab = "Z scores", main = "Normal Probability Plot FEV Data")  
qqline(rstandard(model3),datax = TRUE)



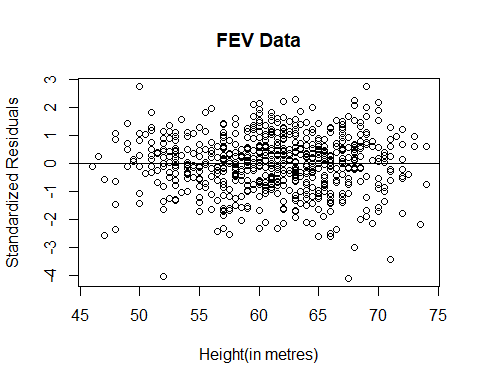
#standardized residuals vs Smoke:  
plot(Smoke,rstandard(model3), xlab="Smoke", ylab= "Standardized Residuals", main = "FEV Data")  
abline(h=0)



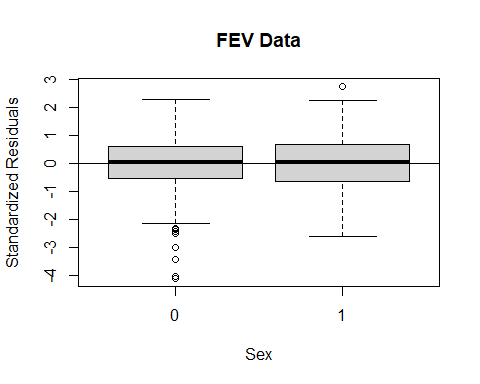
#standardized residuals vs Age:  
plot(Age,rstandard(model3), xlab="Age", ylab= "Standardized Residuals", main = "FEV Data")  
abline(h=0)



#standardized residuals vs Height:  
plot(Hgt,rstandard(model3), xlab="Height(in metres)", ylab= "Standardized Residuals", main = "FEV Data")  
abline(h=0)



#standardized residuals vs Sex:  
plot(Sex,rstandard(model3), xlab="Sex", ylab= "Standardized Residuals", main = "FEV Data")  
abline(h=0)



table(data$Age)

##   
## 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19   
## 2 9 28 37 54 85 94 81 90 57 43 25 19 13 8 6 3

Age^1.290987 Examining FEV and Age model1<-lm(FEV ~ Age ,data=data) summary(model1) Examining FEV with regressor Smoke(The R square is not very good. However, the p value is small) model2<-lm(FEV ~ Smoke ,data=data) summary(model2) modelz<-lm(FEV ~ Hgt ,data=data) summary(modelz) plot(Hgt,FEV) anova(model3) model4<-lm(FEV~Hgt\_m,data=data) summary(model4) model<-lm(FEV ~ Smoke + Age + Hgt + Hgt\_m +Sex ,data=data) sw<-step(model, direction = “backward”) summary(sw) sw$anova As observed by the scatterplot in the diagram above, we also notice a non-linear(indicating curvature) relationship between FEV and Hgt. From Topic 5 (slide 6), we transform height variable into log(Hgt) to perform a suitable transformation.