

Statistics and Econometrics

Set working Directory

```
setwd("C:\\Users\\VB\\Desktop\\PROJECTS")
```

Read Data

```
LungCapData <- read.csv("LungCapData.txt",header = T,sep = "\t")
```

```
attach(LungCapData)
names(LungCapData)
```

```
## [1] "LungCap" "Age" "Height" "Smoke" "Gender" "Caesarean"
```

Exploratory Data Analysis

UNIVARITE ANALYSIS- *to check the summary of the data to get rough idea about Data.*

```
summary(LungCapData)
```

```
##      LungCap      Age      Height      Smoke      Gender
##  Min.   : 0.507   Min.   : 3.00   Min.   :45.30   no :648   female:358
##  1st Qu.: 6.150   1st Qu.: 9.00   1st Qu.:59.90   yes: 77   male  :367
##  Median : 8.000   Median :13.00   Median :65.40
##  Mean   : 7.863   Mean    :12.33   Mean    :64.84
##  3rd Qu.: 9.800   3rd Qu.:15.00   3rd Qu.:70.30
##  Max.   :14.675   Max.    :19.00   Max.    :81.80
##  Caesarean
##  no :561
##  yes:164
##
##
##
##
```

Frequency Distribution

to examine the outliers and significant trends are the relative abundance of each particular target data within dataset.

```
breaks <- seq(from=min(Age),to=max(Age),by=3)
pop <- cut(Age,breaks = breaks,right = TRUE,include.lowest = FALSE)
title <- (cbind(table(pop)))
colnames(title) <- c("frequency");title
```

```
##          frequency
## (3,6]           51
## (6,9]          118
## (9,12]         177
## (12,15]        189
## (15,18]        140
```

```
hist(title)
```

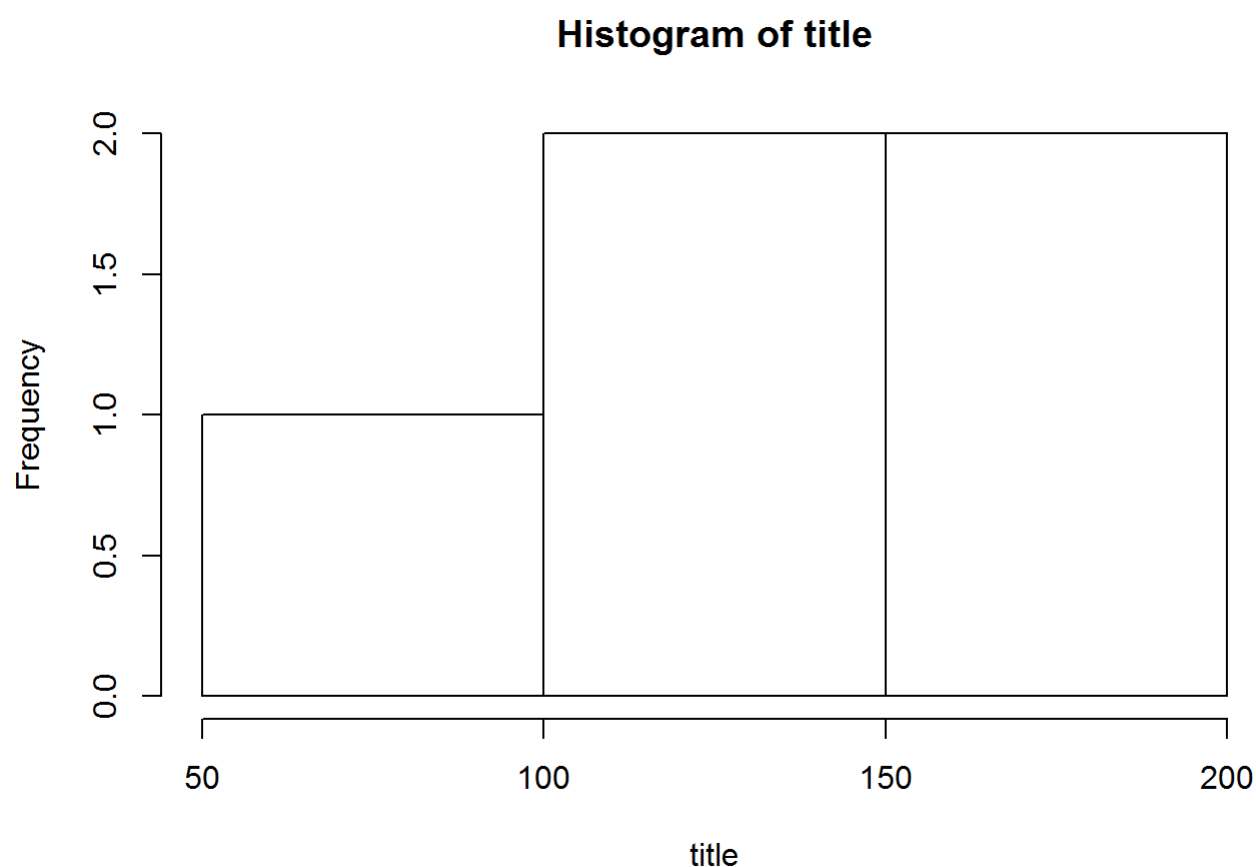


Table for proportion - to examine the number of observations for a variables.

```
table(Smoke)
```

```
## Smoke
## no yes
## 648 77
```

```
table(Smoke)/length(Smoke)
```

```
## Smoke
##      no      yes
## 0.8937931 0.1062069
```

Create two way table or contingency table

```
table(Smoke,Gender)
```

```
##      Gender
## Smoke female male
##   no      314  334
##   yes      44   33
```

Calculate Dimensions of Data - to examine the number of observations and number of variables in dataset.

```
dim(LungCapData)
```

```
## [1] 725   6
```

Measures of Central Tendency

Calculate the count of variable which is going to analyze

```
length(LungCap)
```

```
## [1] 725
```

Calculate the Sum of the LungCap

```
sum(LungCap)
```

```
## [1] 5700.782
```

Calculate the Mean of LungCap

```
mean(LungCap)
```

```
## [1] 7.863148
```

Calculate the Median of of Lungcap

```
median(LungCap)
```

```
## [1] 8
```

Calculate the Mode of Lungcap

```
mode <- function(x){  
  ux <- unique(x)  
  ux[which.max(tabulate(match(x,ux)))]  
}
```

```
mode(LungCap)
```

```
## [1] 8.35
```

Measures of Dispersions

Calculate the Range of the LungCap

```
range(LungCap)
```

```
## [1] 0.507 14.675
```

Calculate quantile of the LungCap

```
quantile(LungCap)
```

```
##      0%      25%      50%      75%     100%  
## 0.507  6.150  8.000  9.800 14.675
```

Calculate IQR of the LungCap

```
IQR(LungCap)
```

```
## [1] 3.65
```

Calculate min

```
min(LungCap)
```

```
## [1] 0.507
```

Calculate Max

```
max(LungCap)
```

```
## [1] 14.675
```

Calculate Variance

```
var(LungCap)
```

```
## [1] 7.086288
```

Calculate Standard Deviation

```
sd(LungCap)
```

```
## [1] 2.662008
```

Calculate Square root

```
sqrt(var(LungCap))
```

```
## [1] 2.662008
```

calculate probabilities

```
quantile(LungCap,probs = c(0.20,0.5,0.9,1))
```

```
##      20%      50%      90%     100%  
##  5.645   8.000  11.205  14.675
```

Measures of Shapes

Measures of Shapes - *Describe the distribution of the data within dataset.*

Calculate the Skewness

```
library(e1071)  
skewness(LungCap)
```

```
## [1] -0.2269314
```

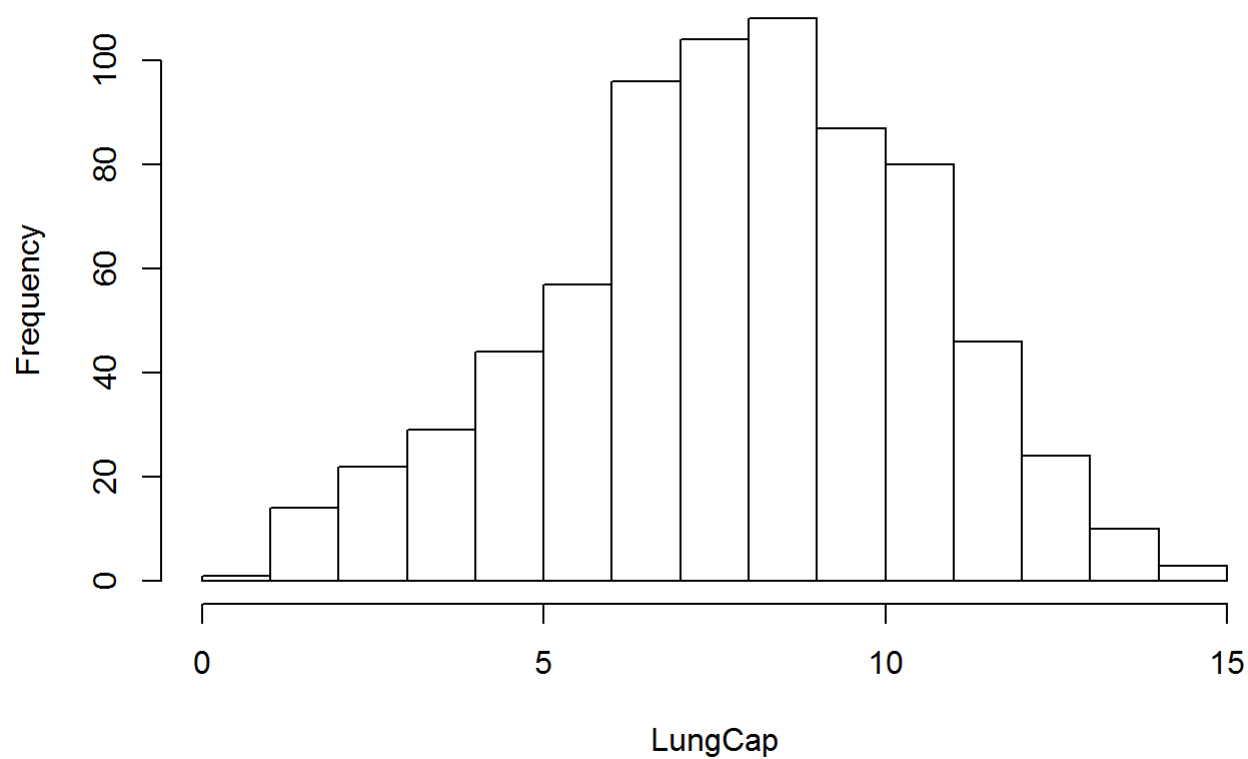
Shapiro test is used to check normality.

```
shapiro.test(LungCap)
```

```
##  
##  Shapiro-Wilk normality test  
##  
## data:  LungCap  
## W = 0.99305, p-value = 0.001886
```

```
hist(LungCap)
```

Histogram of LungCap



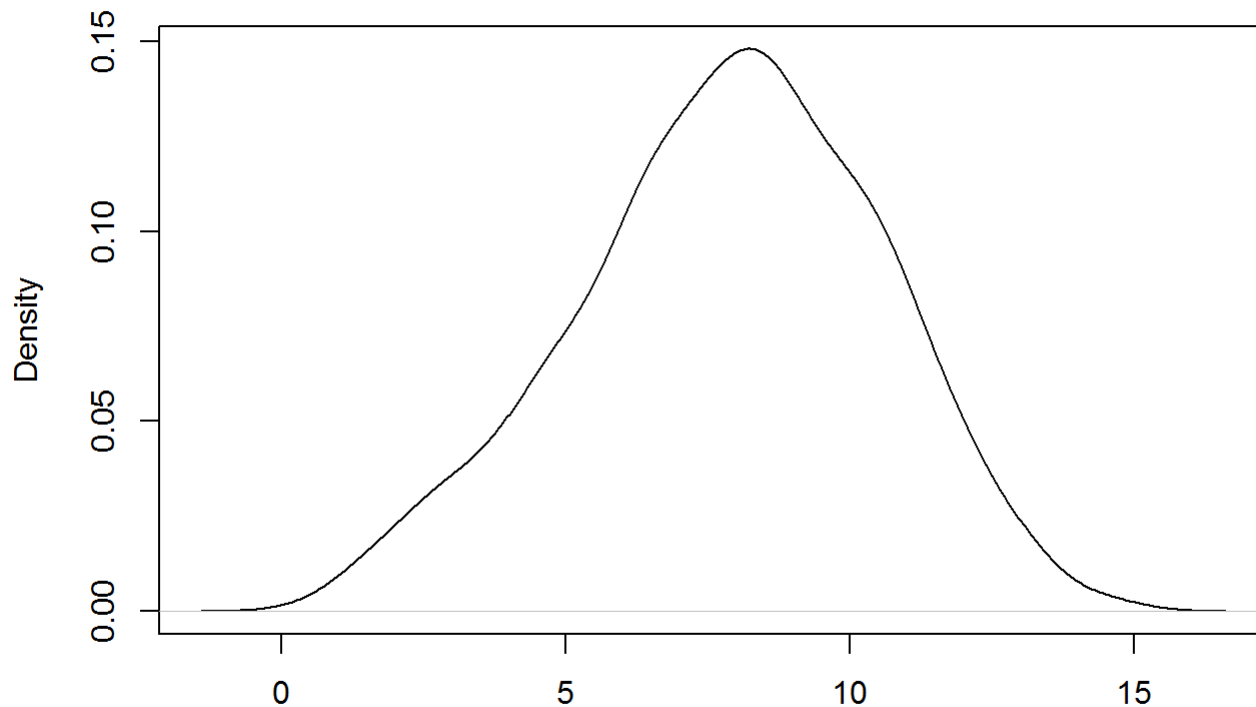
Calculate the Kurtosis

```
kurtosis(LungCap)
```

```
## [1] -0.3259122
```

```
plot(density(LungCap))
```

density.default(x = LungCap)



N = 725 Bandwidth = 0.6418

BIVARIATE ANALYSIS

Bivariate Analysis Deals with two sets of data. this paired data come from related sources or samples.

Correlation - is a parametric measure of the linear association between 2 numeric variables.

```
cor(LungCap,Height)
```

```
## [1] 0.9121873
```

```
cor(LungCap,Age)
```

```
## [1] 0.8196749
```

Covariance - measure of how much two random variables vary together.

```
cov(LungCap,Age)
```

```
## [1] 8.738289
```

Calculating Correlation matrix

```
cor(LungCapData[,1:3])
```

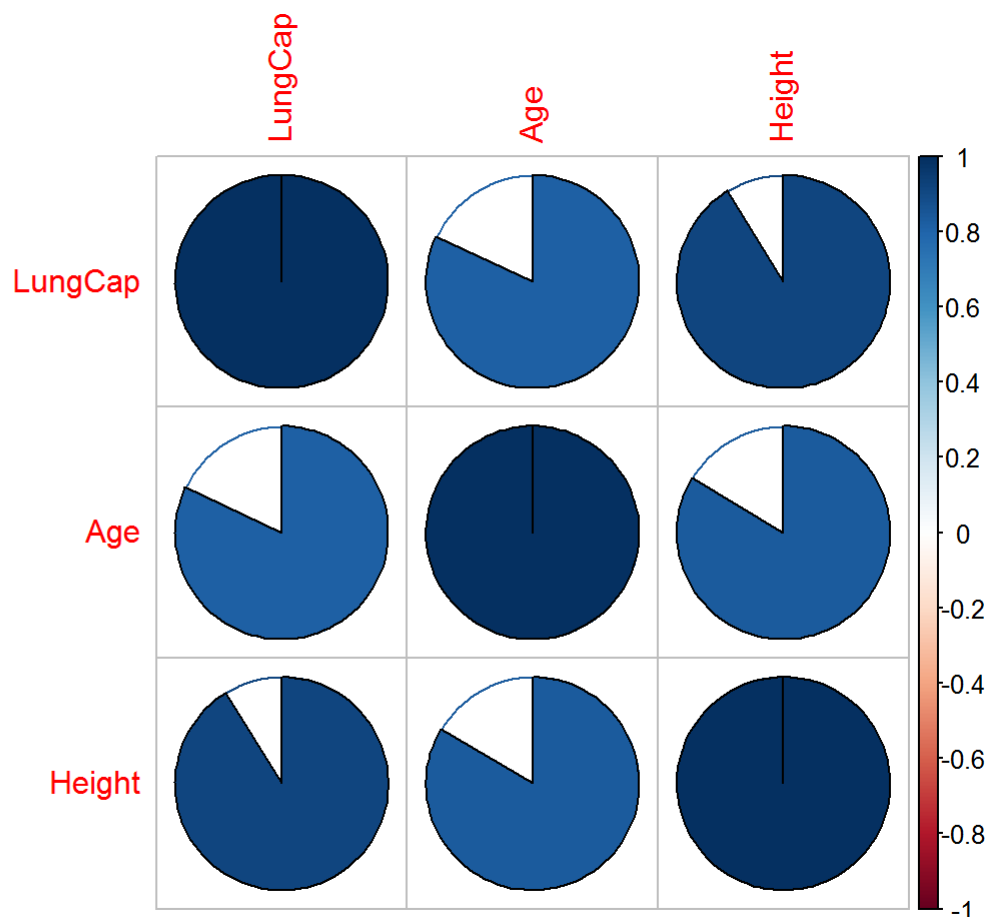
```
##           LungCap      Age      Height
## LungCap 1.0000000 0.8196749 0.9121873
## Age      0.8196749 1.0000000 0.8357368
## Height   0.9121873 0.8357368 1.0000000
```

correlation matrix plot

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
corrplot(cor(LungCapData[,1:3]),method = 'pie')
```



pair plot

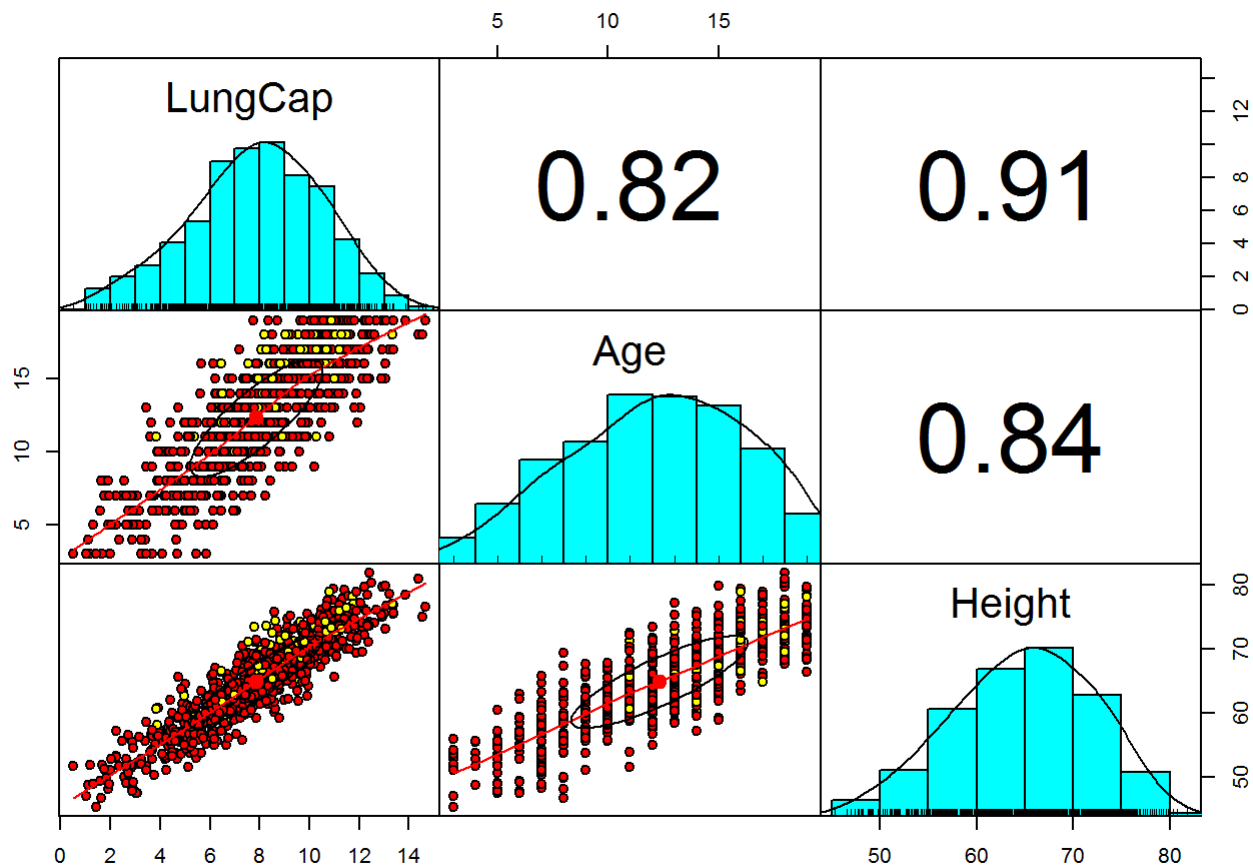
```
library(psych)
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'
```



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

```
pairs.panels(LungCapData[c(1:3)],gap=0,bg=c("red","yellow","blue")[LungCapData$Smoke],pch=21)
```



Multivariate Analysis

You try to understand a sense of relationship of all variables with one another.

```
aggregate(data.frame("LungCap"=LungCapData$LungCap,"Age"=LungCapData$Age,"Height(cm)"=LungCapData$Height), by=list(Smoke=LungCapData$Smoke), FUN=mean)
```

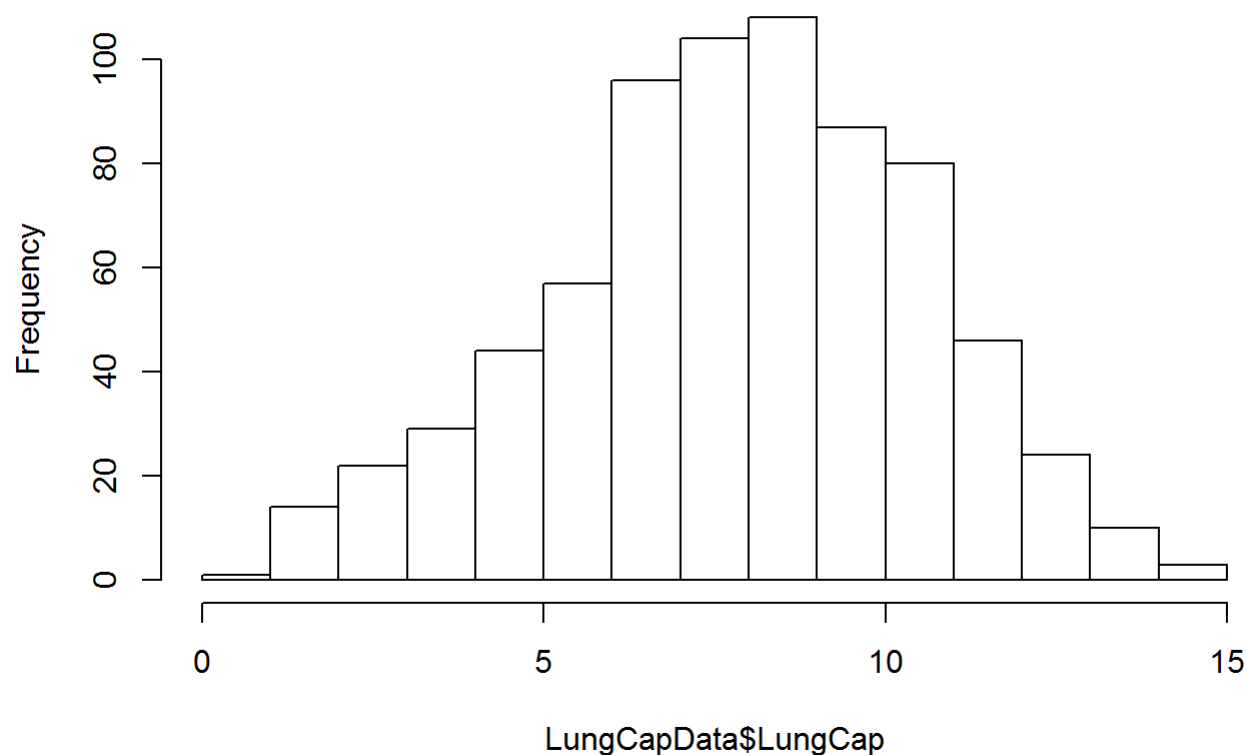
```
##   Smoke  LungCap    Age Height.cm.
## 1    no  7.770188 12.03549  64.39830
## 2    yes  8.645455 14.77922  68.52208
```

VISUALIZATION

Histogram -is a quick way to get information about a sample distribution without detailed statistical Analysis.

```
hist(LungCapData$LungCap)
```

Histogram of LungCapData\$LungCap



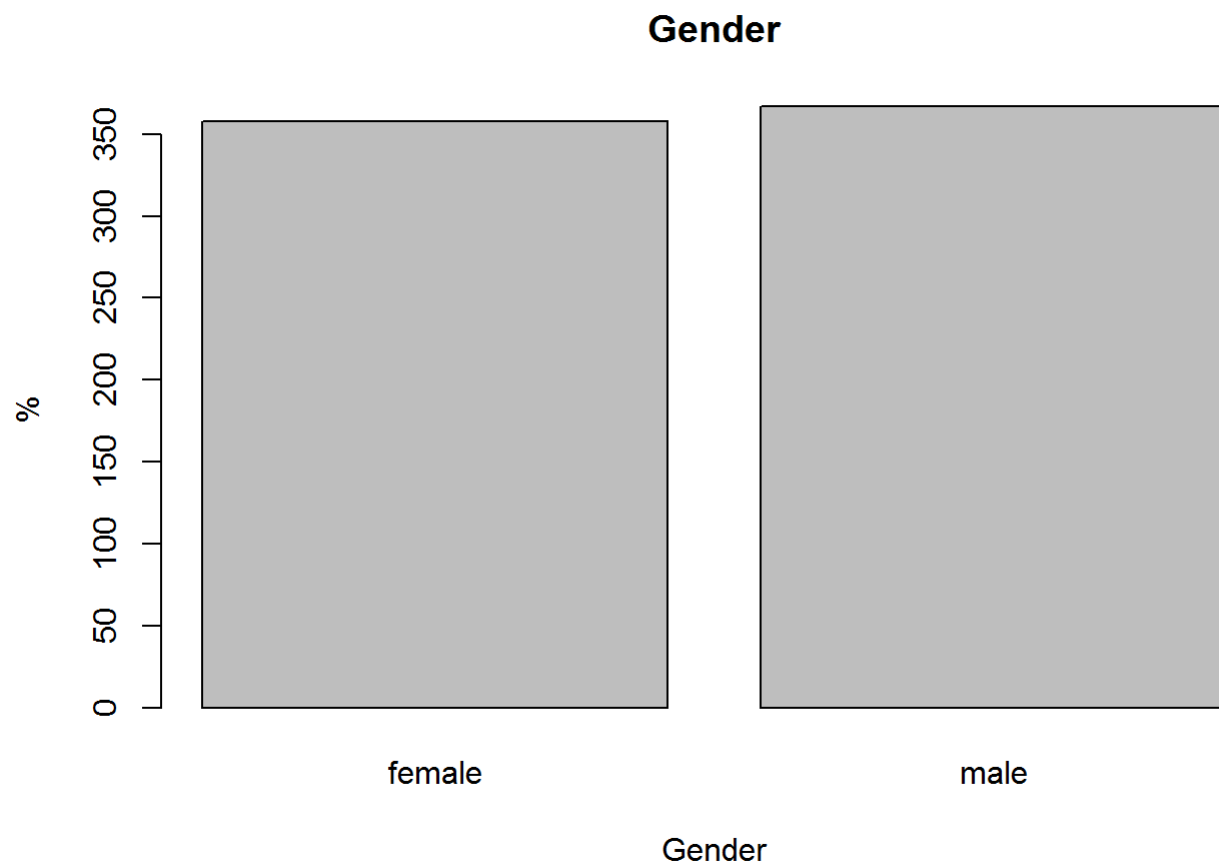
Barplot

Barplot is appropriate for summarizing the distribution of a categorical variables.

```
count <- table(Gender); count
```

```
## Gender
## female  male
##    358    367
```

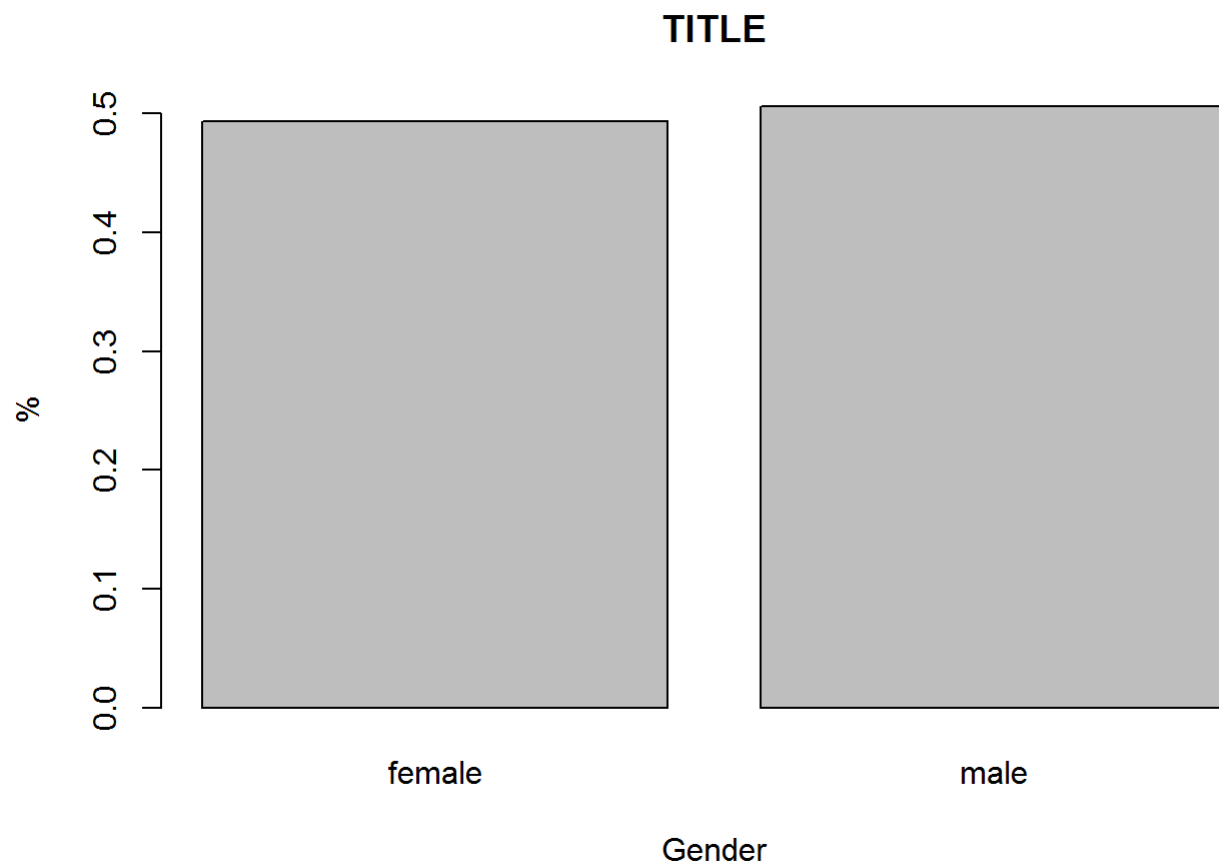
```
barplot(count,main = "Gender",xlab = "Gender",ylab = "%")
```



```
percentage <- table(Gender)/length(Gender)
```

Adding Titles to the plot

```
barplot(percentage,main = "TITLE",xlab = "Gender",ylab = "%")
```

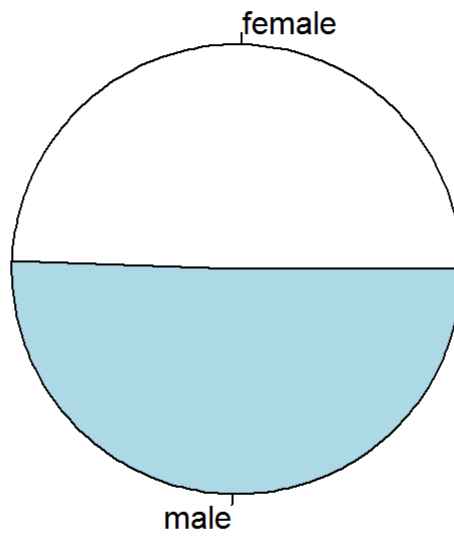


Pie Charts

Pie chart is appropriate for summarizing the distribution of a categorical variables.

```
pie(count, main = "Gender")
```

Gender

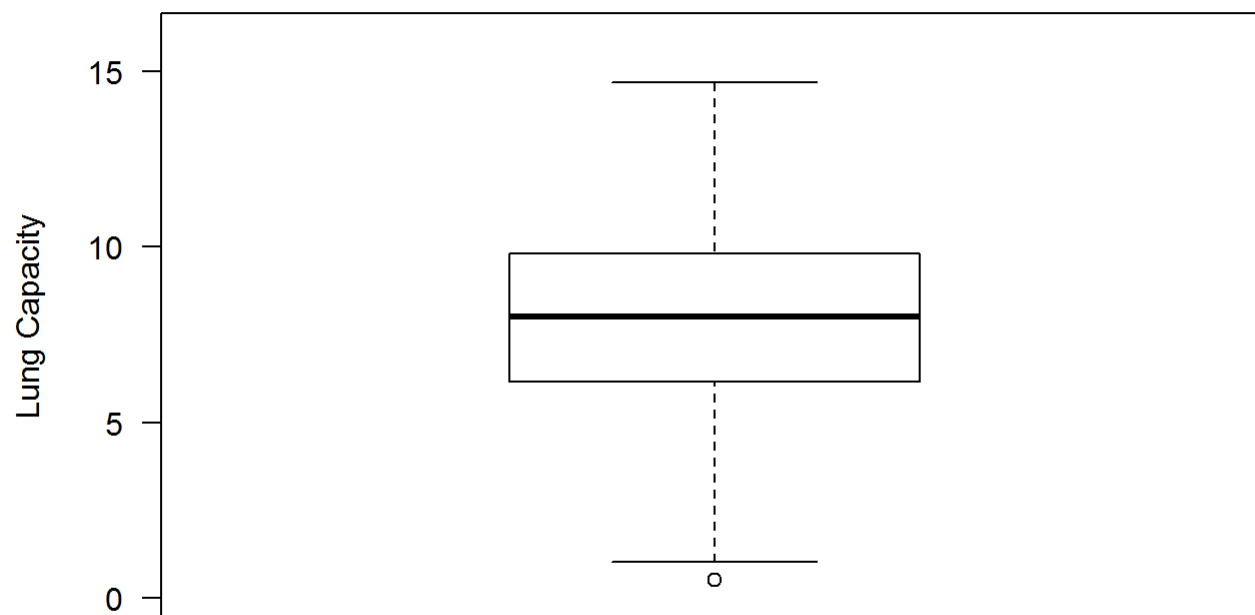


Boxplot

Boxplot is appropriate for summarizing the distribution of a numerical variables.

```
boxplot(LungCap,main='Boxplot',ylab='Lung Capacity',ylim=c(0,16),las=1)
```

Boxplot

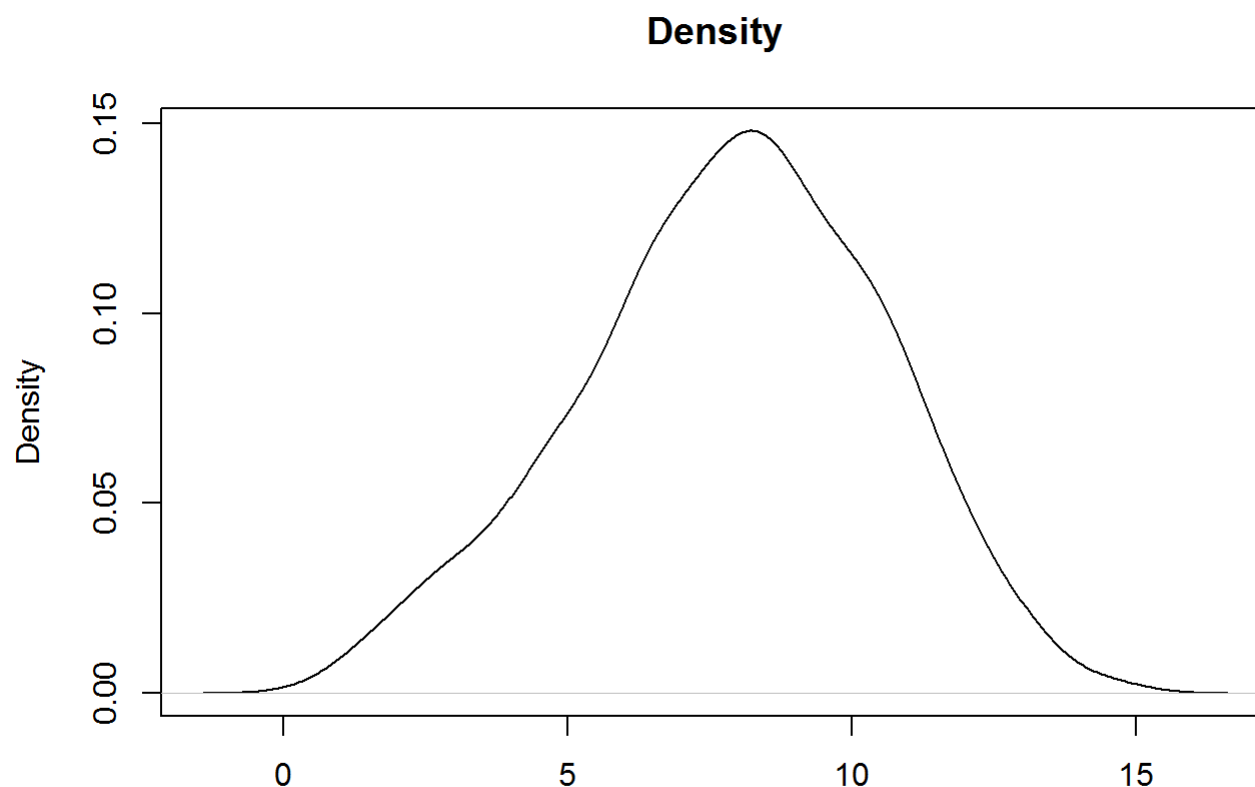


Density plot

Density plot is appropriate for summarizing the distribution of a numerical variables.

```
d <- density(LungCap)
```

```
plot(d, main = 'Density')
```



N = 725 Bandwidth = 0.6418

Stratified Boxplot

is useful for examining the relationship between a categorical variable and numerical variable with strata or groups.

```
AgeGroups <- cut(Age,breaks = c(0,13,15,17,25),labels = c("<13","14/15","15/17","18"))
```

```
Age[1:5]
```

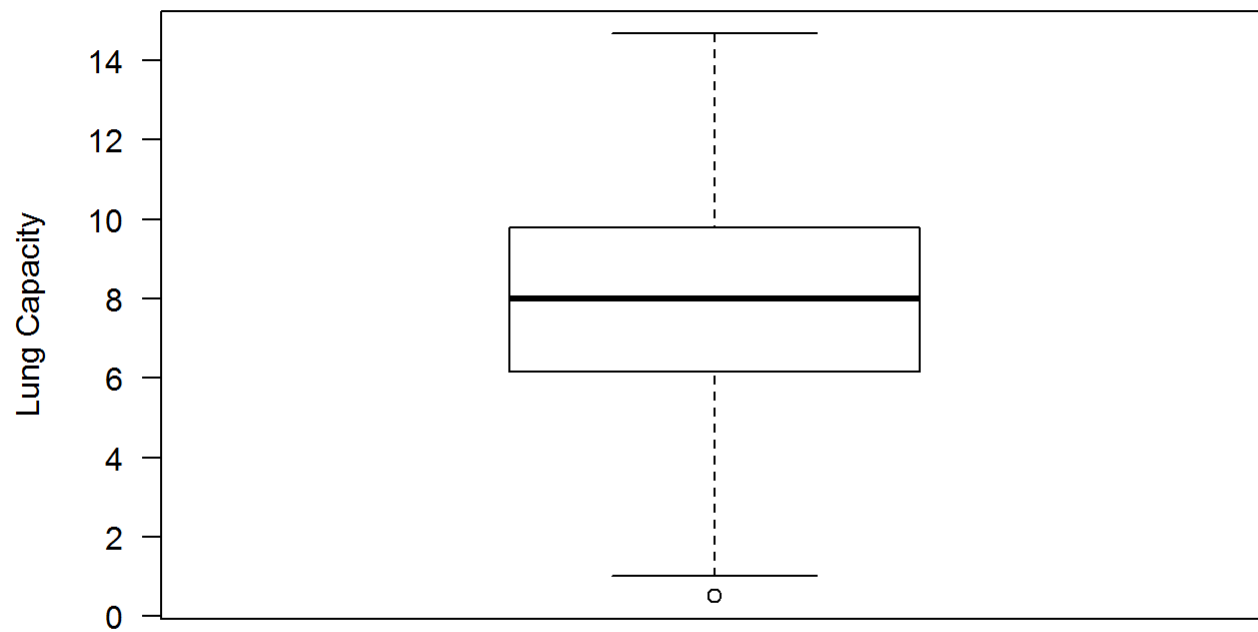
```
## [1] 6 18 16 14 5
```

```
AgeGroups[1:5]
```

```
## [1] <13 18 15/17 14/15 <13  
## Levels: <13 14/15 15/17 18
```

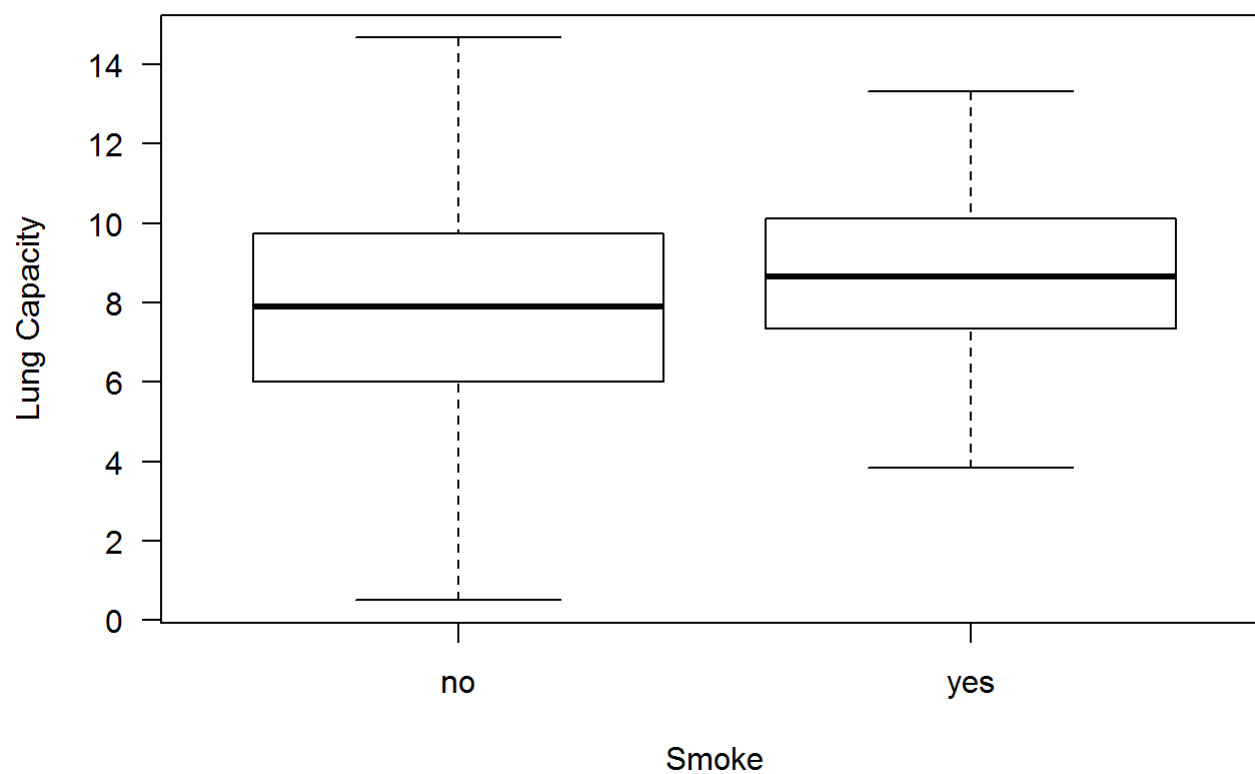
```
boxplot(LungCap,ylab='Lung Capacity',main="Boxplot of LungCap",las = 1)
```

Boxplot of LungCap



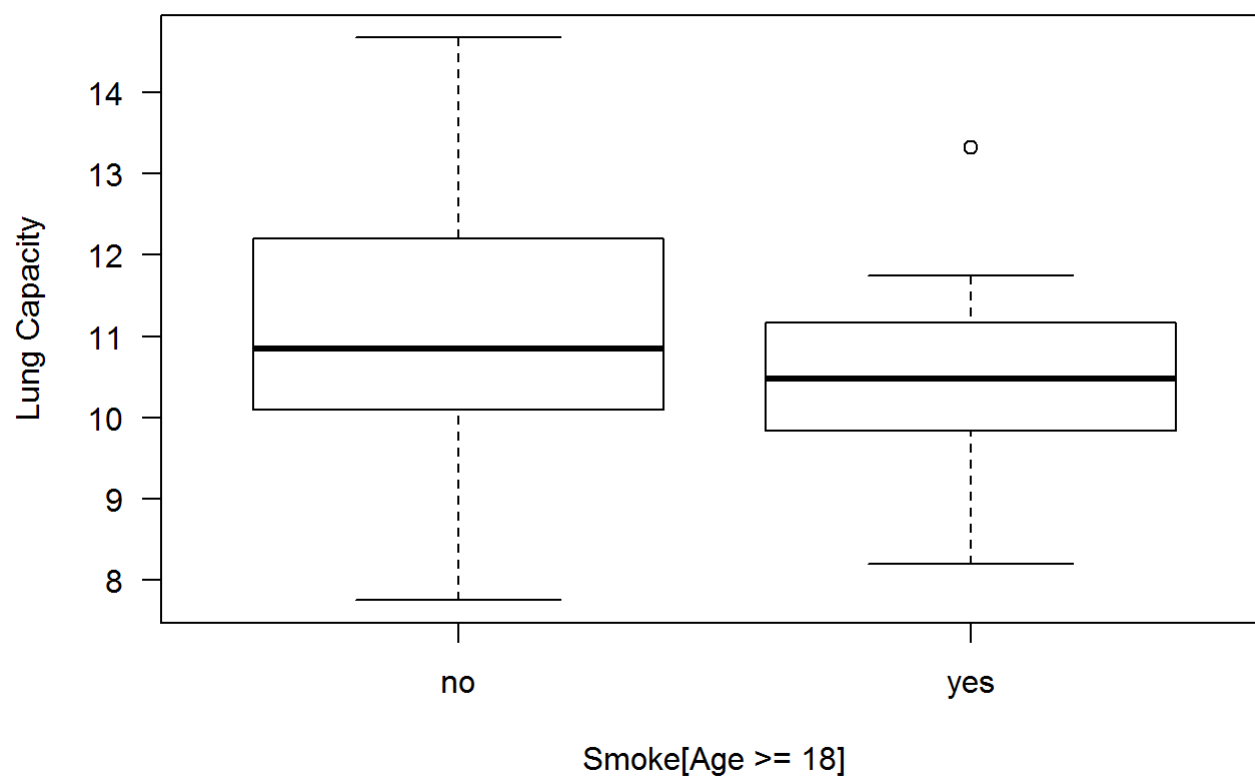
```
boxplot(LungCap ~ Smoke,ylab = "Lung Capacity",main="LungCap vs Smoke",las = 1)
```


LungCap vs Smoke



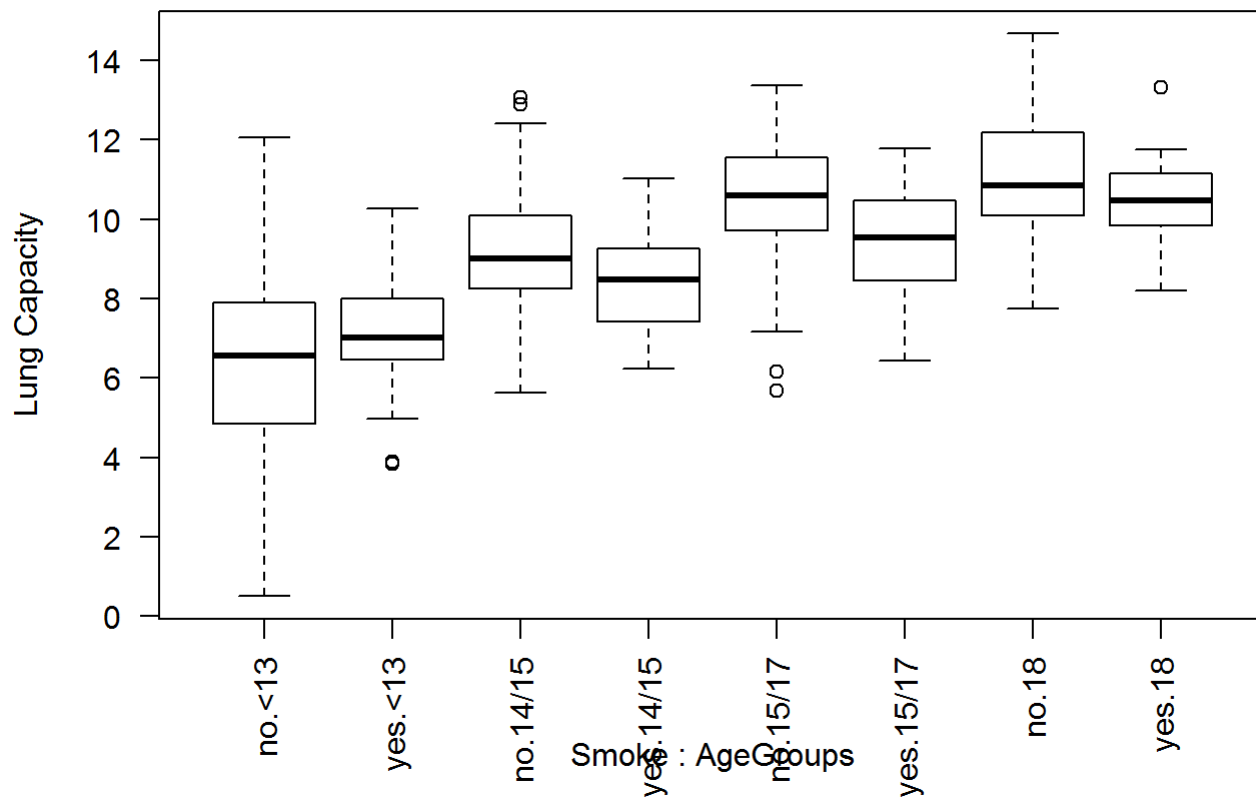
```
boxplot(LungCap[Age >= 18] ~ Smoke[Age >= 18],ylab = "Lung Capacity",main = "LungCap vs smoke, f  
or 18+",las=1)
```

LungCap vs smoke, for 18+



```
boxplot(LungCap ~ Smoke * AgeGroups,ylab = "Lung Capacity", main = "LungCap vs Smoke, by AgeGroup",las = 2)
```

LungCap vs Smoke, by AgeGroup



Steam and Leaf Plot

is appropriate for summarizing the distribution of a numeric variables and are most appropriate for smaller datasets.

```
femaleLungCap <- LungCap[Gender == "female"]
```

```
stem(femaleLungCap,scale = 2)
```

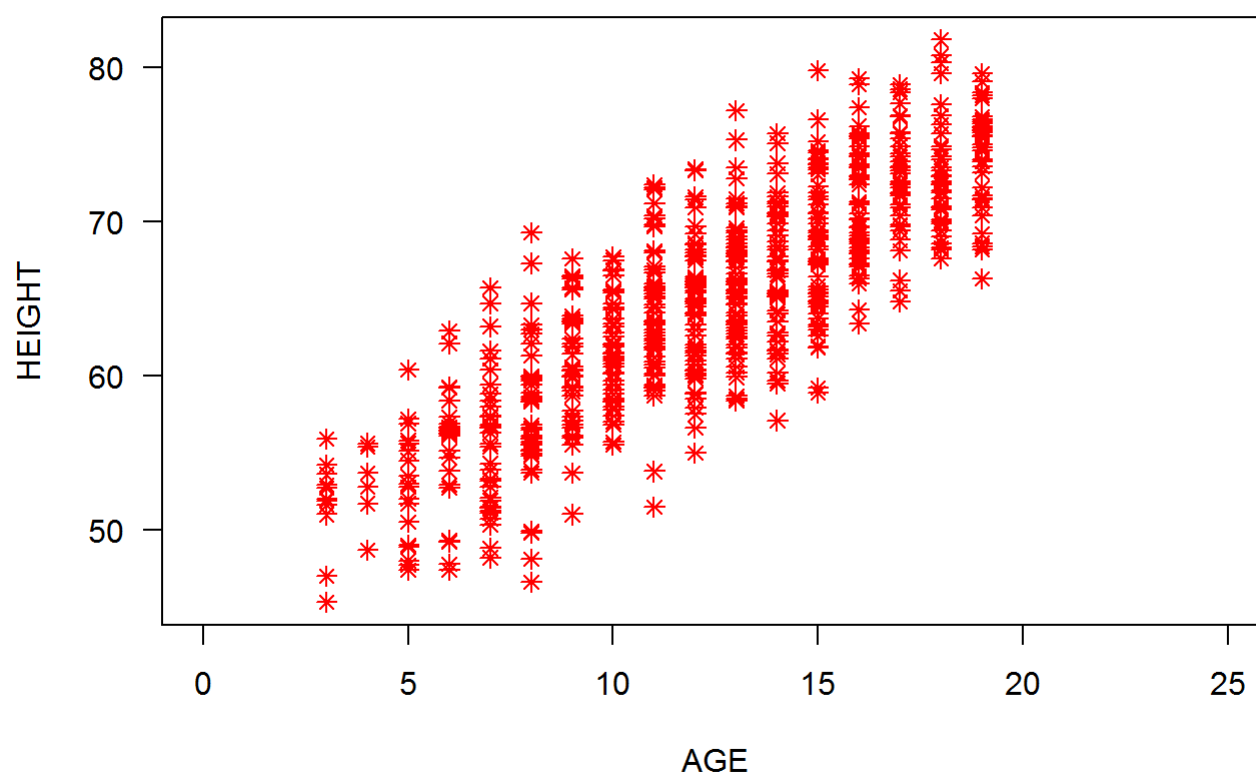
```
##
## The decimal point is at the |
##
## 0 | 5
## 1 | 013
## 1 | 5689
## 2 | 00334
## 2 | 56777789999
## 3 | 01224
## 3 | 57788999999
## 4 | 012333344
## 4 | 555556666677777899
## 5 | 00001222223344
## 5 | 66666777778999
## 6 | 000111111222222223334
## 6 | 55555666666777777788888999999
## 7 | 000123334444444444
## 7 | 5555666667778888888999999
## 8 | 00000000111112222233333444444
## 8 | 55555666666666667777788888888899
## 9 | 00000000111222233333444
## 9 | 55556666777788888999999
## 10 | 00001111122233444
## 10 | 5555666777778899
## 11 | 00111223
## 11 | 556678888
## 12 | 12224
## 12 | 79
## 13 | 1
```

Scatterplot

is appropriate for examining the relationship between 2 numerical variable

```
plot(Age,Height,main = 'Scatterplot',xlab='AGE',ylab = 'HEIGHT',las=1,xlim = c(0,25),pch=8,col=2
)
```

Scatterplot



INFERENCE STATISTICS

Hypothesis Testing *Parametric Test*

Z-test one-sided

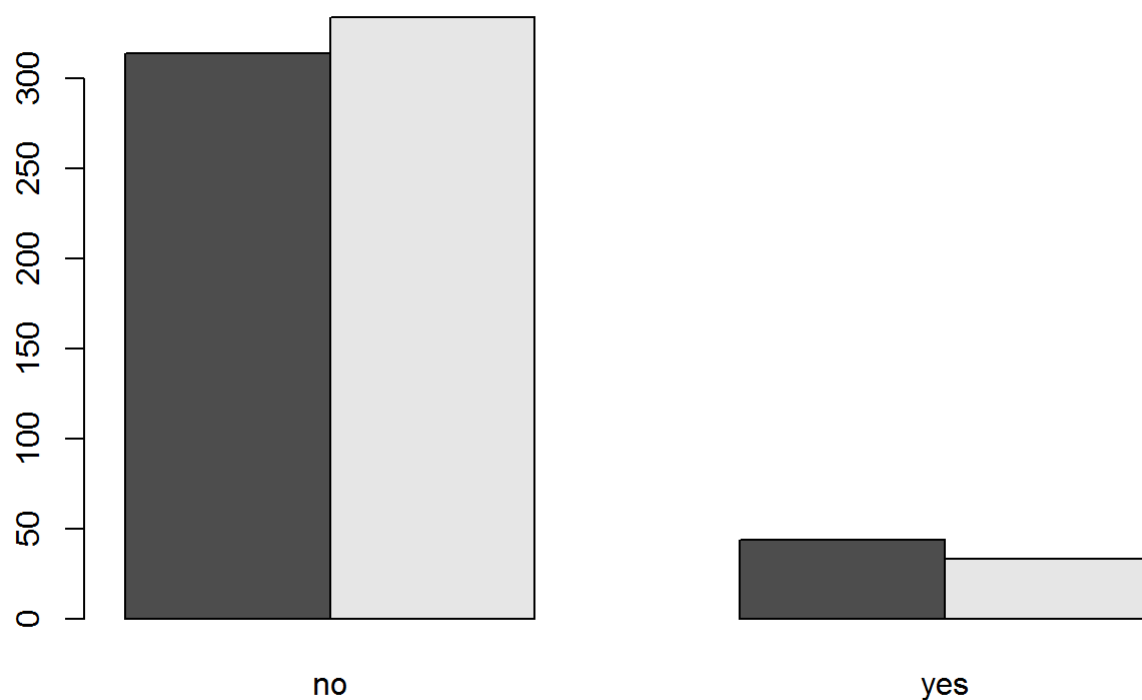
```
# H0 : Male smoker is greater than female smoker.
# CI : 95 %
```

```
t <- table(Gender,Smoke);t
```

```
##          Smoke
## Gender    no  yes
##  female 314  44
##   male  334  33
```

Make a Barplot to examine the distribution of data

```
barplot(t,beside = TRUE)
```



```
prop.test(t,correct = FALSE,alternative = 'greater')
```

```
##
## 2-sample test for equality of proportions without continuity
## correction
##
## data: t
## X-squared = 2.0773, df = 1, p-value = 0.9252
## alternative hypothesis: greater
## 95 percent confidence interval:
## -0.07064256 1.00000000
## sample estimates:
## prop 1 prop 2
## 0.8770950 0.9100817
```

Z-test two-sided

```
# H0 : female smoker is equal to male smoker
# CI : 95%

prop.test(t,correct = FALSE)
```

```
##
## 2-sample test for equality of proportions without continuity
## correction
##
## data:  t
## X-squared = 2.0773, df = 1, p-value = 0.1495
## alternative hypothesis: two.sided
## 95 percent confidence interval:
## -0.07785641  0.01188287
## sample estimates:
##   prop 1    prop 2
## 0.8770950 0.9100817
```

T-test one-side

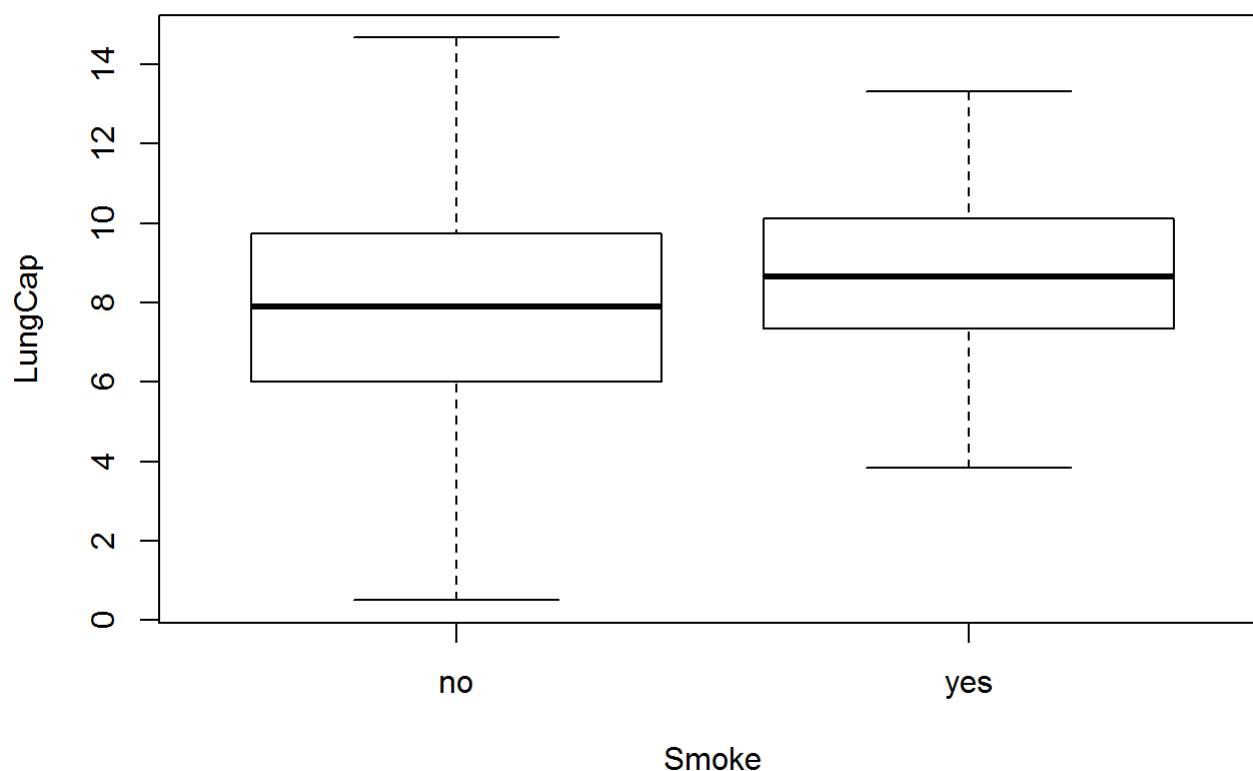
```
# H0 : mu < 8
# one sided 95% cI for mu

t.test(LungCap,mu=8,alternative = "less",conf.level = 0.95)
```

```
##
## One Sample t-test
##
## data:  LungCap
## t = -1.3842, df = 724, p-value = 0.08336
## alternative hypothesis: true mean is less than 8
## 95 percent confidence interval:
##      -Inf 8.025974
## sample estimates:
## mean of x
##  7.863148
```

t-test two-side - is parametric methods appropriate for examining the difference in means for 2 population.

```
boxplot(LungCap ~ Smoke)
```



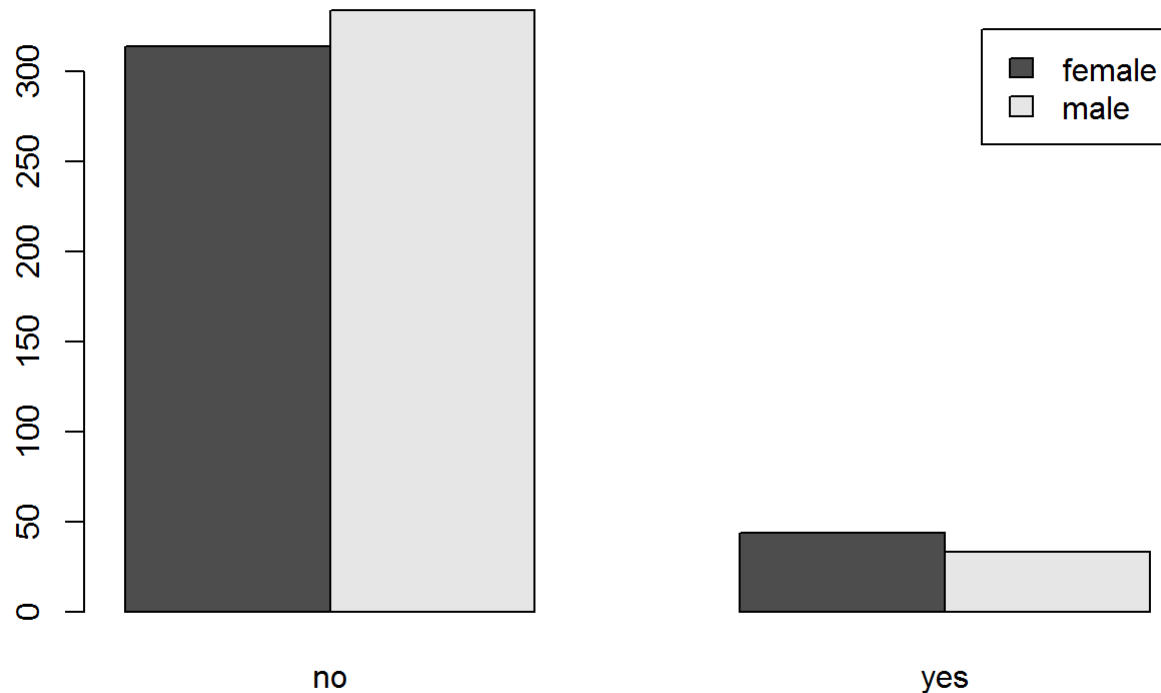
```
# Ho : mean Lung cap of smokers = of non-smokers
# assume non-equal variances

t.test(LungCap ~ Smoke,mu=0,alt='two.side',conf=0.95,var.eq=F,paired =F)
```

```
##
## Welch Two Sample t-test
##
## data: LungCap by Smoke
## t = -3.6498, df = 117.72, p-value = 0.0003927
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -1.3501778 -0.4003548
## sample estimates:
## mean in group no mean in group yes
## 7.770188 8.645455
```

Chi-square - Appropriate for testing independence between two categorical variables.


```
# For chi-square test produce a contingency table  
  
TAB <- table(Gender,Smoke)  
  
# produce a barplot to check the distribution.  
  
barplot(TAB,beside = T,legend=T)
```



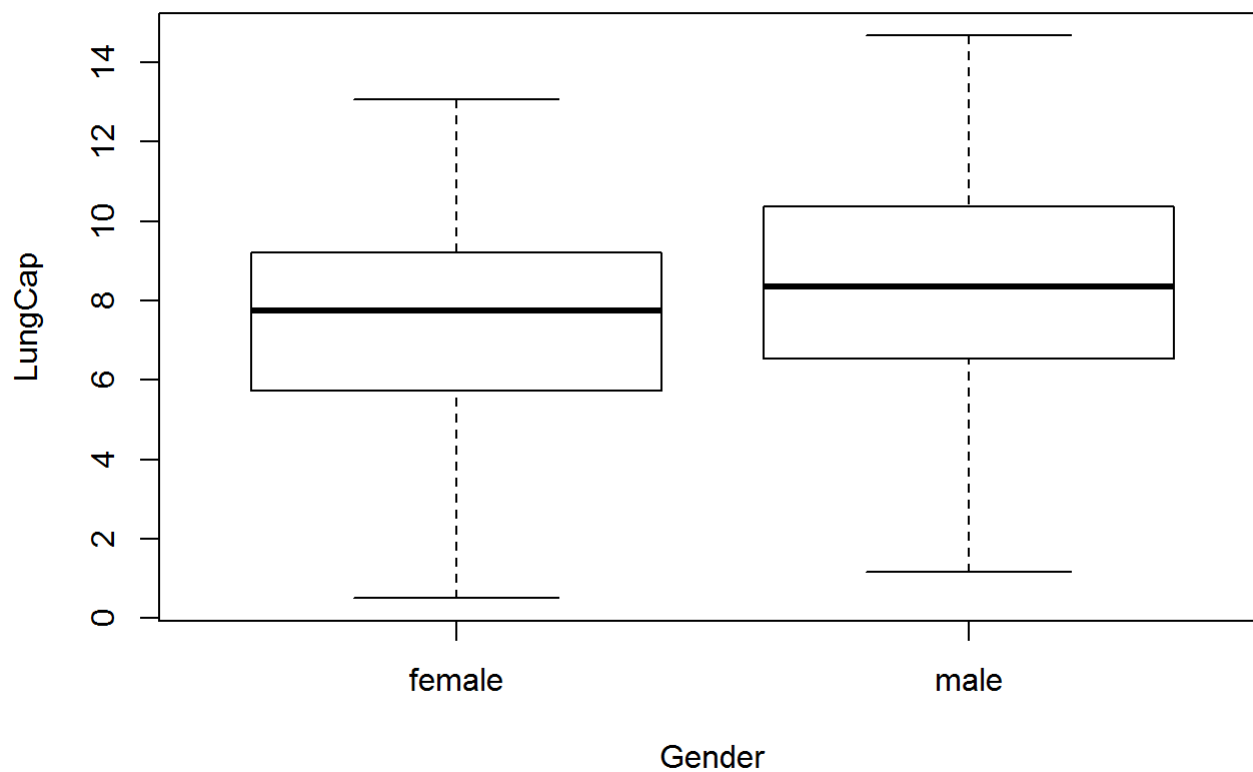
```
CHI <- chisq.test(TAB,correct = T)  
CHI
```

```
##  
## Pearson's Chi-squared test with Yates' continuity correction  
##  
## data: TAB  
## X-squared = 1.7443, df = 1, p-value = 0.1866
```

F-Test

ANOVA - appropriate for comparing the means for 2 or more independent populations.

```
# produce a box plot to check the distribution of Lung cap variable and gender variable.  
  
boxplot(LungCap ~ Gender)
```



```
# H0 : Mean Lungcap is the same for all Genders
```

```
ANOVA1 <- aov(LungCap ~ Gender)
summary(ANOVA1)
```

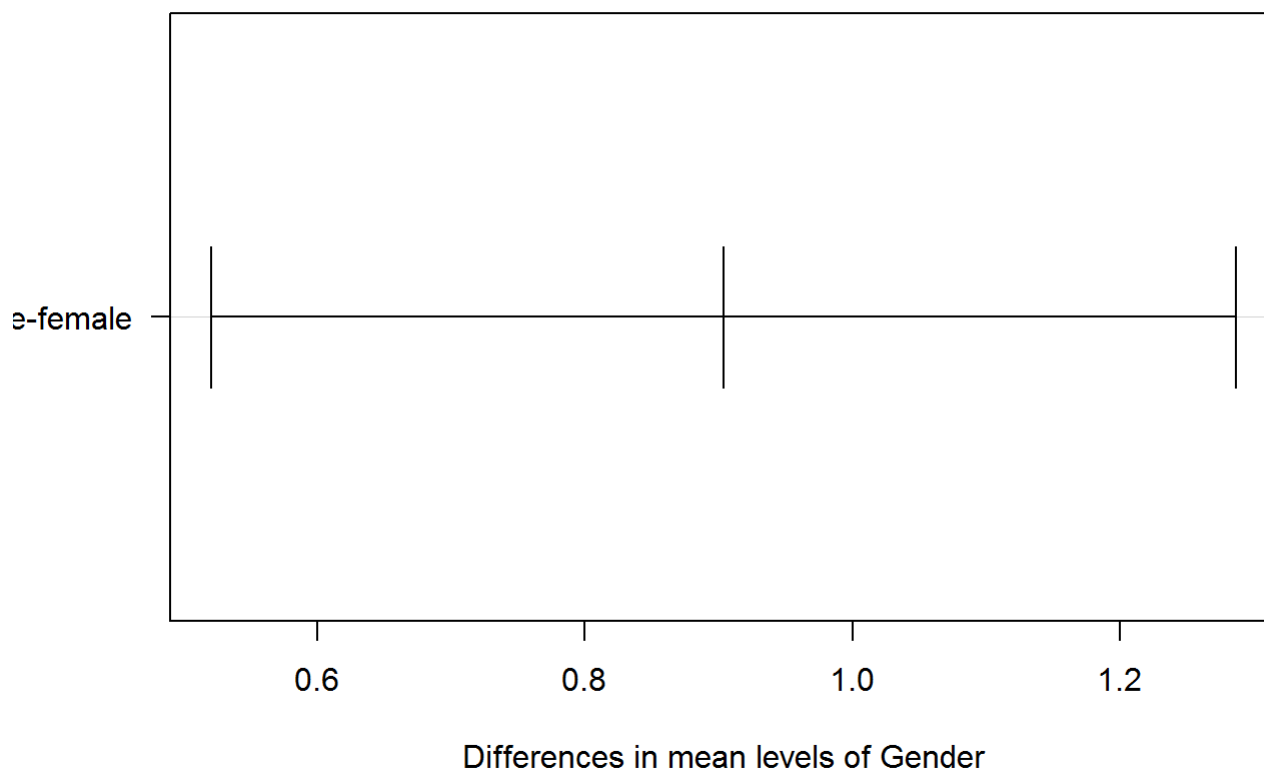
```
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Gender      1    148   147.96    21.47 4.26e-06 ***
## Residuals 723   4983     6.89
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(ANOVA1)
```

```
## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = LungCap ~ Gender)
##
## $Gender
##           diff      lwr      upr    p adj
## male-female 0.9035866 0.5207397 1.286434 4.3e-06
```

```
plot(TukeyHSD(ANOVA1),las=1)
```

95% family-wise confidence level



Non-Parametric Test

Wilcoxon Signed Rank Test - *Appropriate for examining the median Difference in observations for 2 populations.*

Mann-Whitney U Test A.K.A Wilcoxon Rank sum test - *appropriate for examining the difference in Medians for 2 independent populations*

```
# Ho : Median Lung Capacity of Smokers = that of non smokers
# two sided test
```

```
wilcox.test(LungCap ~ Smoke,mu=0,alt='two.sided',conf.int=T,conf.level = 0.95,paired=F,exact=F,correct =T)
```

```
##
## Wilcoxon rank sum test with continuity correction
##
## data: LungCap by Smoke
## W = 20128, p-value = 0.005538
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
## -1.3999731 -0.2499419
## sample estimates:
## difference in location
## -0.8000564
```

Kruskal Wallis Test - equivalent to one-way Analysis of Variance

```
kruskal.test(LungCap ~ Gender)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  LungCap by Gender
## Kruskal-Wallis chi-squared = 18.325, df = 1, p-value = 1.862e-05
```

Fisher's Exact test* - *alternative to the chi-square test, it is used when the assumptions of chi-square test not met. we may consider using Fisher's Exact Test*

```
fisher.test(TAB, conf.int = T, conf.level = 0.99)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  TAB
## p-value = 0.1845
## alternative hypothesis: true odds ratio is not equal to 1
## 99 percent confidence interval:
##  0.3625381 1.3521266
## sample estimates:
## odds ratio
##  0.7054345
```

PROBABILITY DISTRIBUTION

Binomial Distribution - *X is Binomially Distributed with $n = 20$ trials and $p = 1/6$ prob of success.*

```
# In R dbinom command is used to find values for the probability density function of x, f(x)
# suppose 80% adults are smoker and out of 10 are caesarean what
# is the probability that they are male exactly seven
# observations is n = 10
# success or events of male is x = 7
# p=0.8

dbinom(x=7, size = 10, prob = 0.8)
```

```
## [1] 0.2013266
```

```
# Probability of having exactly 7 males is 20.13%
```

Possion Distribution - *is the probability distribution of independent occurances in an interval.*

```
# support there are 12 adults smoking per minute on an average,  
# find the probability of having seventeen or more adults smoking in a  
# particular minutes  
  
# probability of haveing sixteen or less adults smoking in a particular  
# minute is given by the fuction ppois.  
  
ppois(16,lambda = 12) # Lower tail
```

```
## [1] 0.898709
```

```
# Here, the probability of having seventeen or more adults smoking in a # minute is in the upper  
tail of the probability density function  
  
ppois(16,lambda = 12,lower = FALSE) # UPPER TAIL
```

```
## [1] 0.101291
```

```
# if there are twelve adults smoking per minute on an average, the probability of having sevente  
en ore more adults smoking in a particuler minute is 10.1%
```

Normal Distribution

```
# the mean of Lung capacity is 7, and standard deviation is 2.66.  
# What is the percentage of Lung capacity of female which has 9 or more lung cap.  
  
pnorm(9,mean = 7,sd=2.66, lower.tail = FALSE)
```

```
## [1] 0.2260617
```

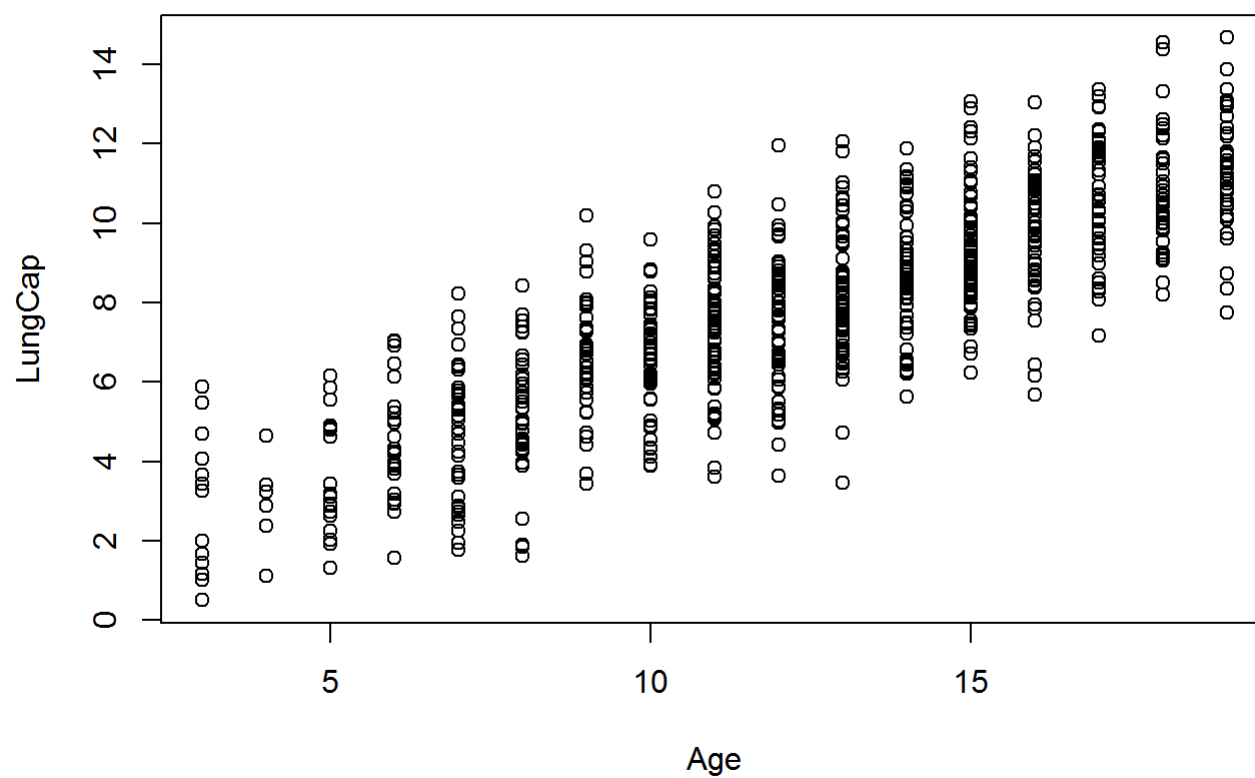
```
# The percentage of female having Lung cap 9 or more is 22.6%
```

MODELLING

Simple Linear Regression - useful for examining or modelling the relationship between 2 numeric variables.

```
# Model the relationship between Age and Lung Capacity  
  
plot(Age,LungCap,main = "Scatterplot")
```

Scatterplot



```
# Calculate the correlation
```

```
cor(Age,LungCap)
```

```
## [1] 0.8196749
```

```
# Fit a Linear Model
```

```
mod <- lm(LungCap ~ Age)
```

```
# Model Evaluation
# Function that returns Root Mean Squared Error
```

```
rmse <- function(error)
{
  sqrt(mean(error^2))
}
```

```
# Function that returns Mean Absolute Error
```

```
mae <- function(error)
{
  mean(abs(error))
}
```

```
rmse(mod$residuals)/100
```

```
## [1] 0.01523824
```

```
mae(mod$residuals)/100
```

```
## [1] 0.01218942
```

```
summary(mod)
```

```
##
## Call:
## lm(formula = LungCap ~ Age)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.7799 -1.0203 -0.0005  0.9789  4.2650
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.14686    0.18353   6.249 7.06e-10 ***
## Age          0.54485    0.01416  38.476 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.526 on 723 degrees of freedom
## Multiple R-squared:  0.6719, Adjusted R-squared:  0.6714
## F-statistic: 1480 on 1 and 723 DF,  p-value: < 2.2e-16
```

```
# Predict the Model
```

```
predict(mod, data.frame(Age = 16))
```

```
##          1
## 9.864432
```

Multiple Linear Regression -

useful for modelling the relationship between more than 2 numeric variables.

```
# Fit model
```

```
model1 <- lm(LungCap ~ Age + Height)
```

```
# Summary
```

```
summary(model1)
```

```
##
## Call:
## lm(formula = LungCap ~ Age + Height)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.4080 -0.7097 -0.0078  0.7167  3.1679
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11.747065   0.476899  -24.632  < 2e-16 ***
## Age          0.126368   0.017851   7.079 3.45e-12 ***
## Height       0.278432   0.009926  28.051  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.056 on 722 degrees of freedom
## Multiple R-squared:  0.843, Adjusted R-squared:  0.8425
## F-statistic: 1938 on 2 and 722 DF, p-value: < 2.2e-16
```

```
# Calculate Pearson's correlation between Age and Height
```

```
cor(Age,Height,method = "pearson")
```

```
## [1] 0.8357368
```

```
# ask for confidence intervals for the model coefficients
```

```
confint(model1,conf.level=0.95)
```



```
##                2.5 %      97.5 %  
## (Intercept) -12.68333877 -10.8107918  
## Age         0.09132215   0.1614142  
## Height      0.25894454   0.2979192
```

MODEL DIAGNOSTIC

Check Outliers

```
library(car)
```

```
## Loading required package: carData
```

```
##  
## Attaching package: 'car'
```

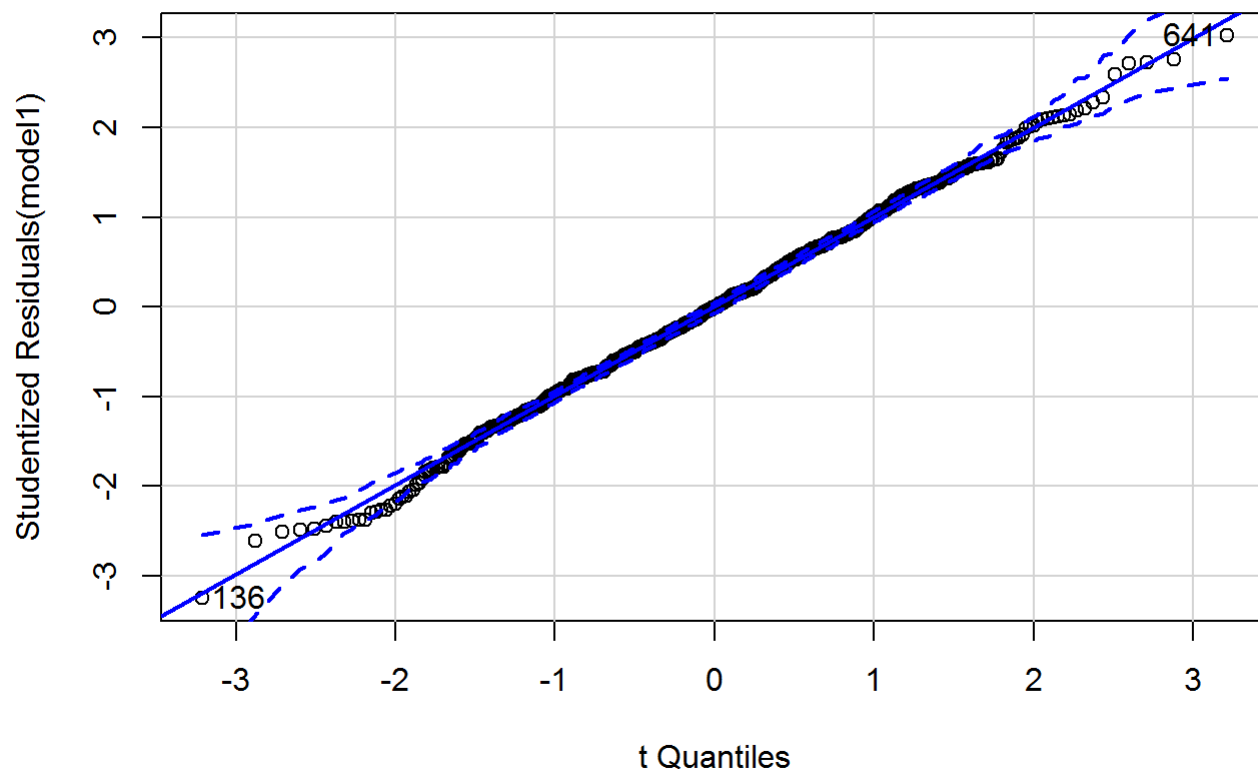
```
## The following object is masked from 'package:psych':  
##  
##      logit
```

```
outlierTest(model1)
```

```
## No Studentized residuals with Bonferroni p < 0.05  
## Largest |rstudent|:  
##      rstudent unadjusted p-value Bonferroni p  
## 136 -3.250022      0.0012075      0.87547
```

```
qqPlot(model1,main="QQPLOT")
```

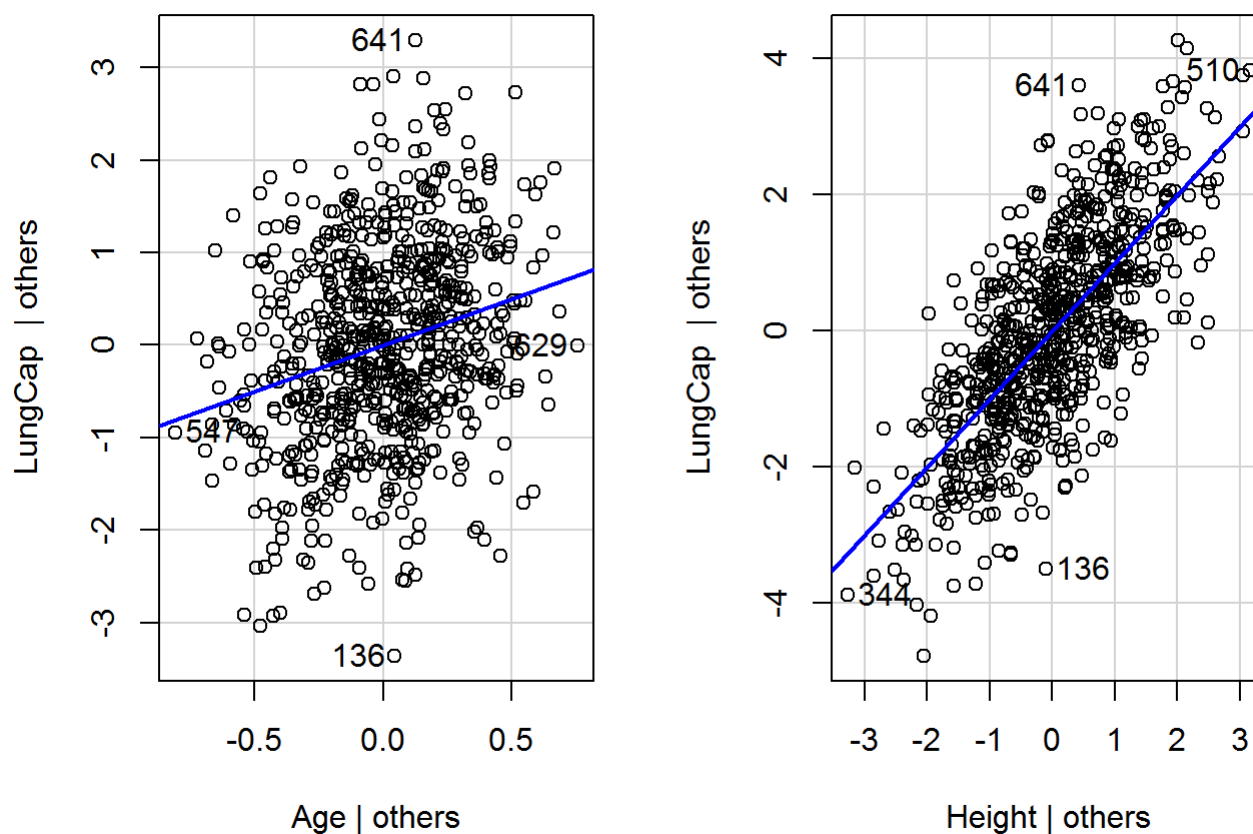
QQPLOT



```
## [1] 136 641
```

```
leveragePlots(model1)
```

Leverage Plots



Examine the multicollinearity problem in the model

```
vif(model1) # Variable value should be less than 10 then we can conclude there is no Multicollinearity issue.
```

```
##      Age      Height
## 3.316266 3.316266
```

Checking the Heteroscedasticity problem in the model

```
ncvTest(model1) # P-value should be less than 0.05 then we can conclude there is no heteroscedasticity
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 1.347551, Df = 1, p = 0.24571
```

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##  
## Attaching package: 'zoo'
```

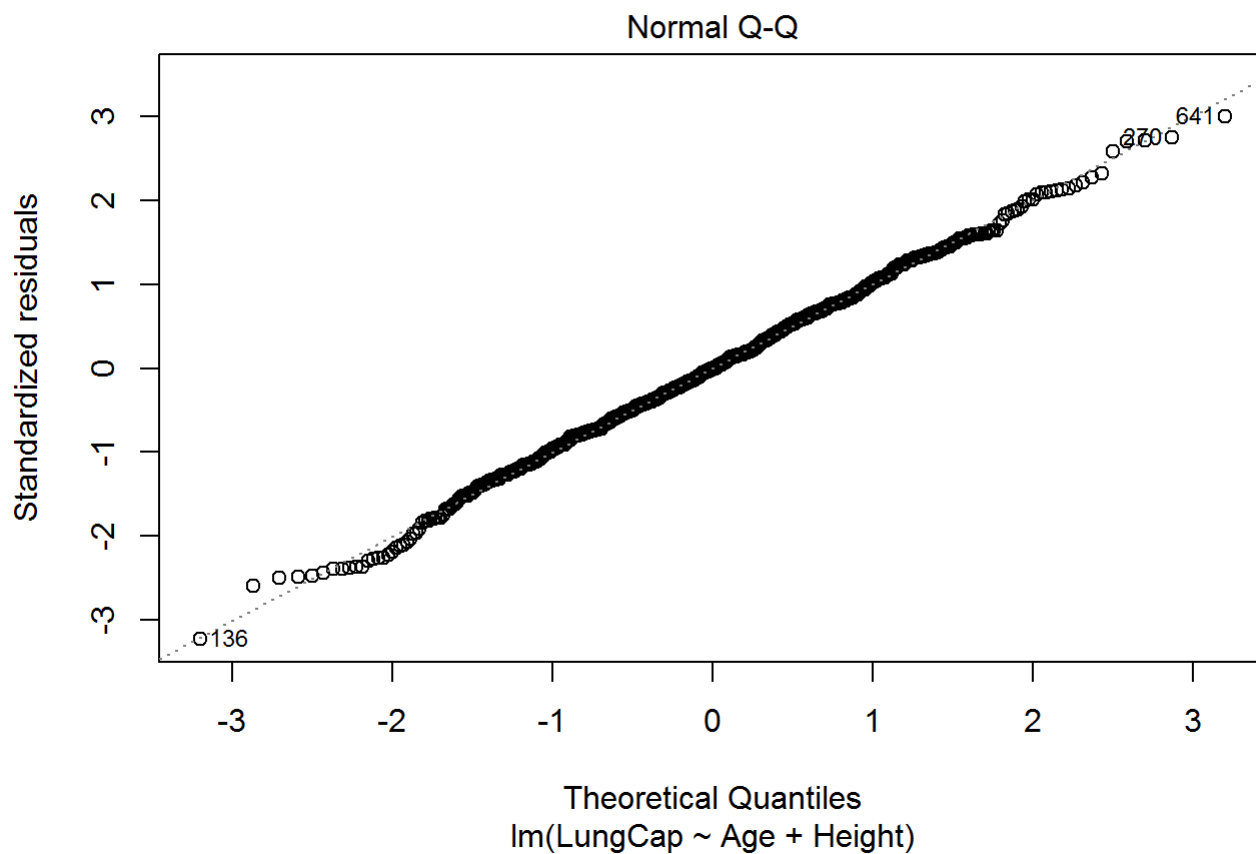
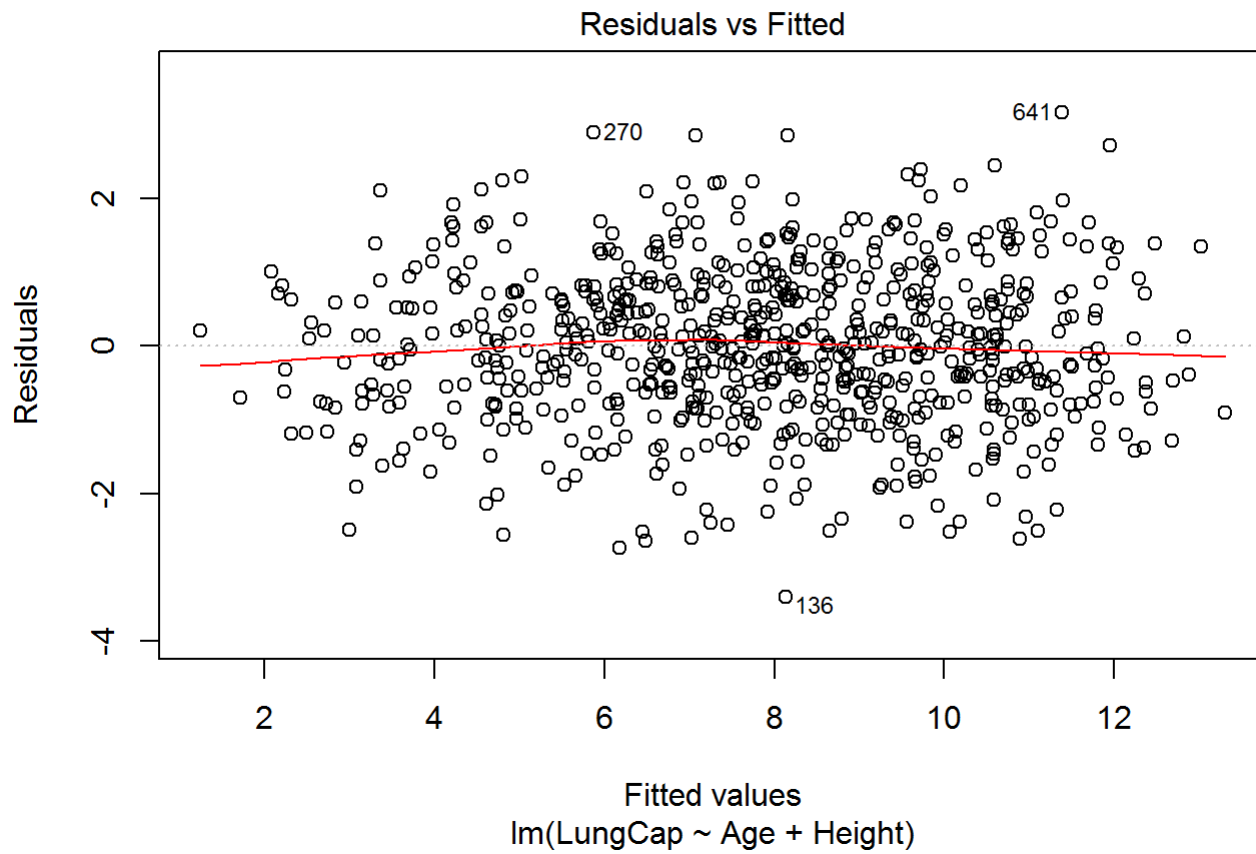
```
## The following objects are masked from 'package:base':  
##  
##    as.Date, as.Date.numeric
```

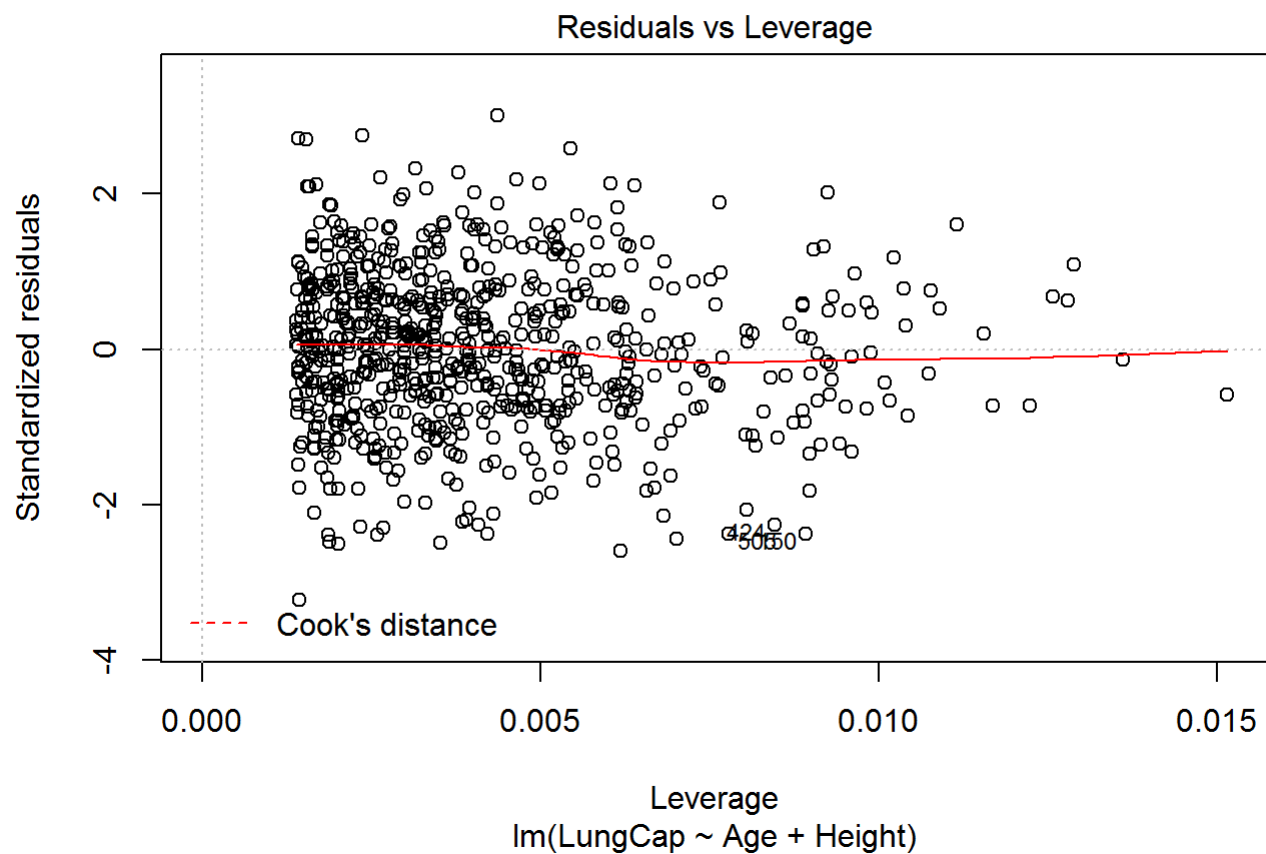
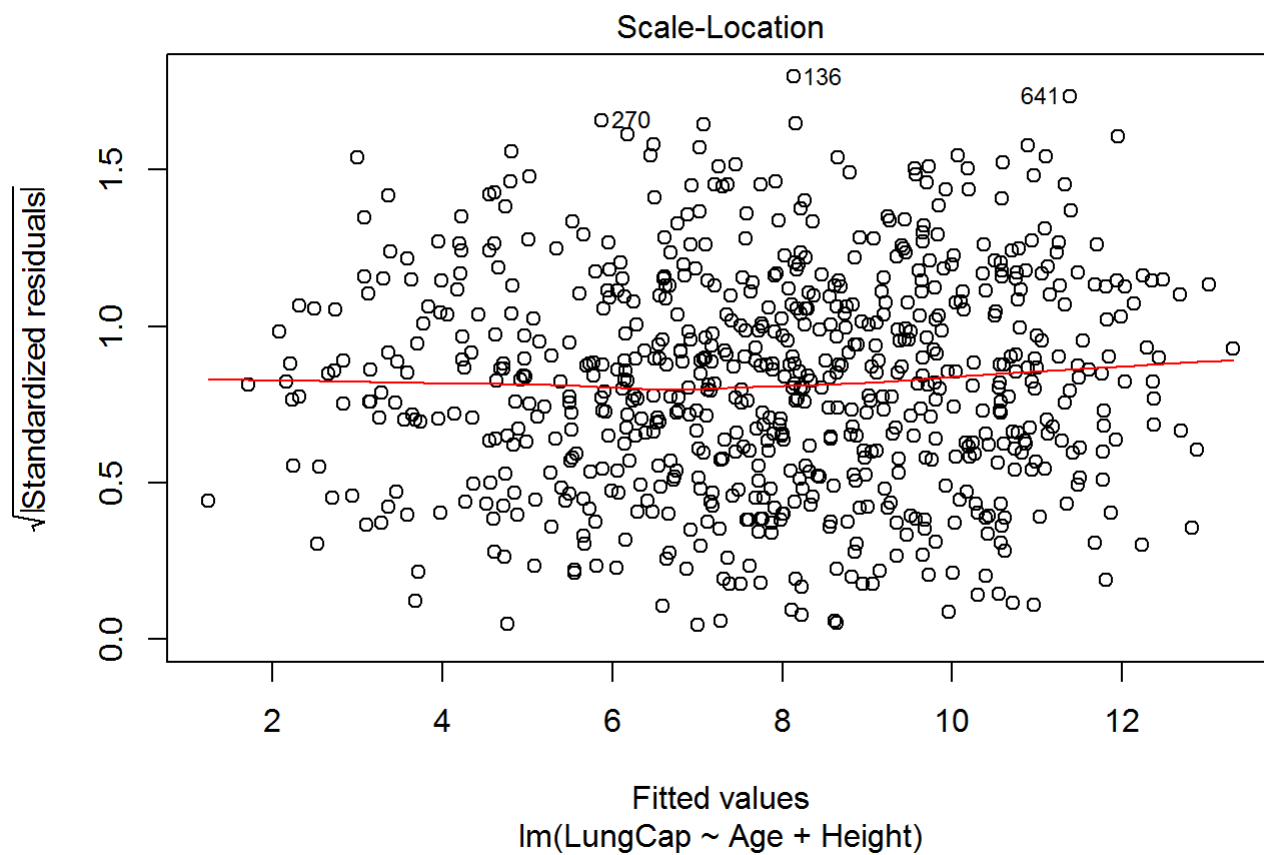
```
bptest(model1)
```

```
##  
## studentized Breusch-Pagan test  
##  
## data:  model1  
## BP = 1.7152, df = 2, p-value = 0.4242
```

Checking Non-Linearity problem in the model

```
library(corrplot)  
plot(model1)
```





Checking AUtocorrelation problem in the model

```
dwtest(model1)
```

```
##  
## Durbin-Watson test  
##  
## data: model1  
## DW = 1.8348, p-value = 0.01296  
## alternative hypothesis: true autocorrelation is greater than 0
```

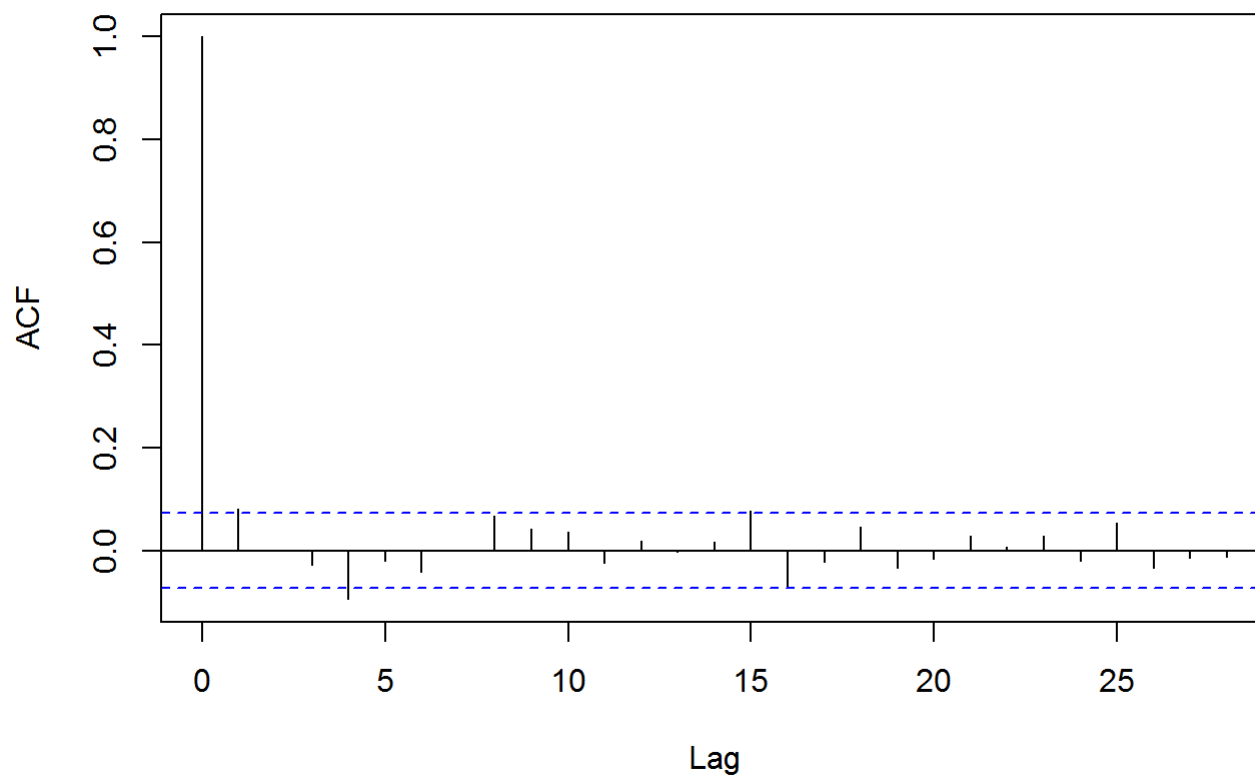
p-value is < 2.0 hence we accept H_a , can conclude there is no autocorrelation in this model

```
durbinWatsonTest(model1)
```

```
## lag Autocorrelation D-W Statistic p-value  
## 1 0.0815809 1.83481 0.02  
## Alternative hypothesis: rho != 0
```

```
acf(model1$residuals)
```

Series model1\$residuals



we see all the vertical lines are within significance bounce except 0 and more or less line 4 is crossing the significance bounce but not so high. we can conclude there is no autocorrelation in the model.

```
bgtest(model1)
```

```
##  
## Breusch-Godfrey test for serial correlation of order up to 1  
##  
## data: model1  
## LM test = 4.8678, df = 1, p-value = 0.02736
```

p value is < 2.0 hence we accept H_a , can conclude there is no autocorrelation in this model

```
bgtest(model1, order = 2)
```

```
##  
## Breusch-Godfrey test for serial correlation of order up to 2  
##  
## data: model1  
## LM test = 4.8876, df = 2, p-value = 0.08683
```

p value is < 2.0 hence we accept H_a , can conclude there is no autocorrelation in this model

REGRESSION ASSUMPTIONS

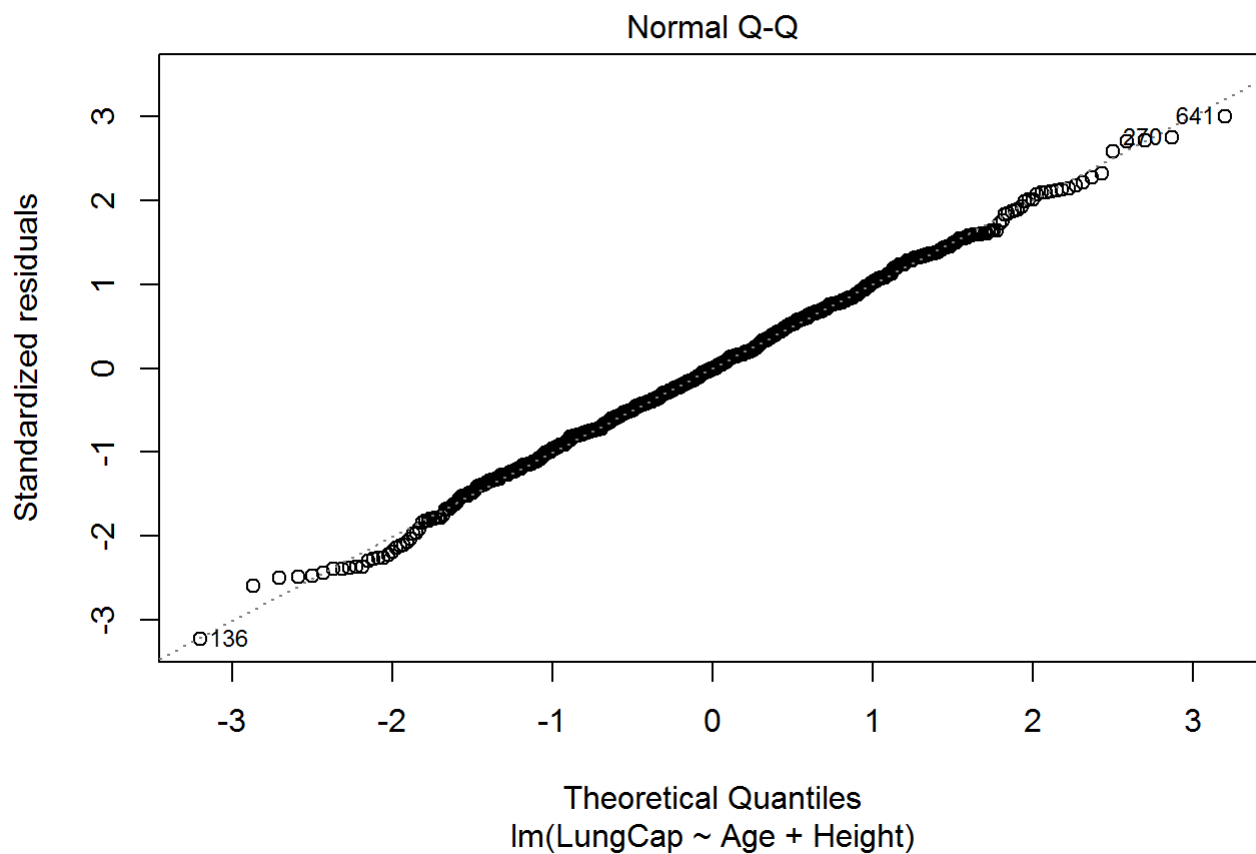
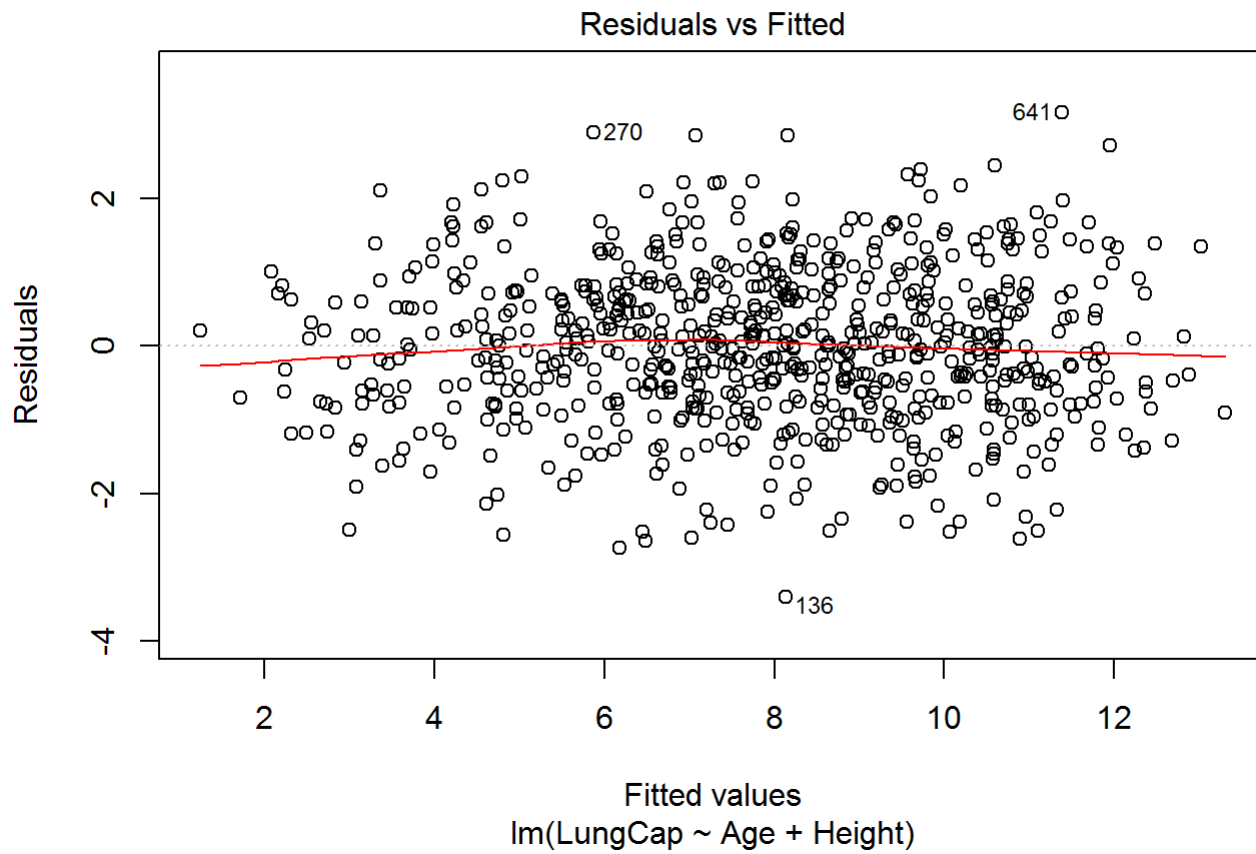
1 Assumption : The Y-values (or the errors, "e") are independent!

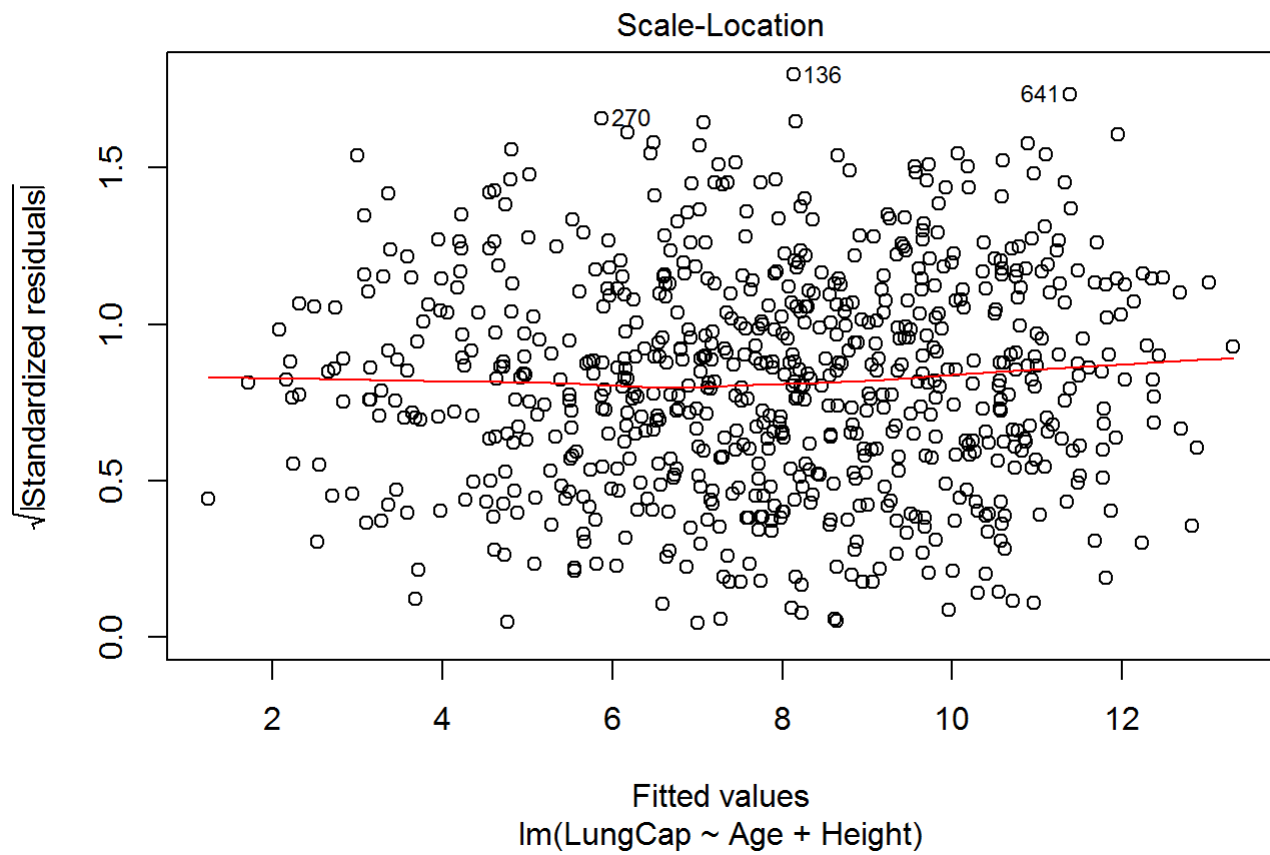
2 Assumption : The Y-values can be expressed as a linear function of the X variables

3 Assumption : Variation of observations around the regression line (the residual SE is constant (homoscedasticity))

4 Assumption : for given values of X, Y values (or the error) are Normally distributed

```
plot(model1)
```



LOGISTIC REGRESSION - is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

```
# Read Data
```

```
data <- read.csv("LungCapData.txt",header = T,sep = "\t")
str(data)
```

```
## 'data.frame': 725 obs. of 6 variables:
## $ LungCap : num 6.47 10.12 9.55 11.12 4.8 ...
## $ Age : int 6 18 16 14 5 11 8 11 15 11 ...
## $ Height : num 62.1 74.7 69.7 71 56.9 58.7 63.3 70.4 70.5 59.2 ...
## $ Smoke : Factor w/ 2 levels "no","yes": 1 2 1 1 1 1 1 1 1 1 ...
## $ Gender : Factor w/ 2 levels "female","male": 2 1 1 2 2 1 2 2 2 2 ...
## $ Caesarean: Factor w/ 2 levels "no","yes": 1 1 2 1 1 1 2 1 1 1 ...
```

```
data <- data[c(1:4)]
```

```
# Normalize Data
```

```
data$LungCap <- scale(data$LungCap)
data$Age <- scale(data$Age)
data$Height <- scale(data$Height)
```

```
# Convert data to binary
```

```
levels(data$Smoke)=0:1
head(data)
```

```
##      LungCap      Age      Height Smoke
## 1 -0.5214663 -1.5798483 -0.3799252     0
## 2  0.8496790  1.4165938  1.3695539     1
## 3  0.6336766  0.9171868  0.6753161     0
## 4  1.2253352  0.4177798  0.8558180     0
## 5 -1.1506905 -1.8295518 -1.1019324     0
## 6 -0.6153804 -0.3313307 -0.8520068     0
```

```
# Check the class imbalance
```

```
table(data$Smoke)
```

```
##
## 0 1
## 648 77
```

Data Partition

```
set.seed(1234)
ind <- sample(2,nrow(data),replace = T,prob = c(0.8,0.2))
train <- data[ind == 1,]
test <- data[ind == 2,]
```

Handling Class Imbalance

```
library(ROSE)
```

```
## Loaded ROSE 0.0-3
```

```
both <- ovun.sample(Smoke ~., data = train,method = 'both',p=0.5,seed=222,N=573)$data
table(both$Smoke)
```

```
##
##    0    1
## 287 286
```

```
over <- ovun.sample(Smoke ~.,data = train,method = 'over',N=1296)$data
table(over$Smoke)
```

```
##
##    0    1
## 511 785
```

```
under <- ovun.sample(Smoke ~., data = train,method = 'under',N=154)$data
table(under$Smoke)
```

```
##
##    0    1
##  92  62
```

Build a Model

```
mymodel <- glm(Smoke ~ LungCap + Age + Height,data = both, family = 'binomial')
```

Predict Model on Training Dataset

```
p <- predict(mymodel,both,type = 'response')
pred <- ifelse(p>0.5,1,0)
tab <- table(Predicted=pred,Actual=both$Smoke)
1-sum(diag(tab))/sum(tab)
```

```
## [1] 0.2879581
```

```
# Predict model on test data
```

```
p1 <- predict(mymodel,test,type = 'response')  
pred1 <- ifelse(p1 >0.5,1,0)  
tab1 <- table(Predicted= pred1, Actual=test$Smoke)  
1-sum(diag(tab1))/sum(tab1)
```

```
## [1] 0.2960526
```