#amulya Geereddy

Preprocessing:

library(rio)

mydata=import("6304 Module 4 Assignment Data.xlsx", sheet="Tamil Anatomy")

colnames(mydata)=tolower(make.names(colnames(mydata)))

attach(mydata)

set.seed(62067273)

primary=mydata[sample(1:nrow(mydata),70),]

Analysis:

1.

cor(primary$height,primary$left.foot.length)

> cor(primary$height,primary$left.foot.length)

[1] 0.6201473

Correlation value is slightly close to 1, So it indicates a slightly positive relation between two variables height and left foot length.

2.

a)

datalm=lm(left.foot.length~height, data=primary)

summary(datalm)

Graphical user interface, text, application, Word

Description automatically generated

From the above summary it can be seen that beta coefficients of the model are intercept

-1.7706 , 0.0168 is the slope of the regression equation and p values of the intercept and slope are 0.315 and 1.03e-08 respectively.

confint(datalm)

> confint(datalm)

2.5 % 97.5 %

(Intercept) -5.2643683 1.7231546

height 0.1157783 0.2179307

Beta coefficients for confidence intervals are also calculated as -5.26 and 1.72 as intercepts at 2.5% and 97.5% respectively and slopes are 0.115 and 0.217 respectively.

b)

From the summary we find that the P is at significant height and length foot are proportional if value of one goes up other also increases and vice versa.

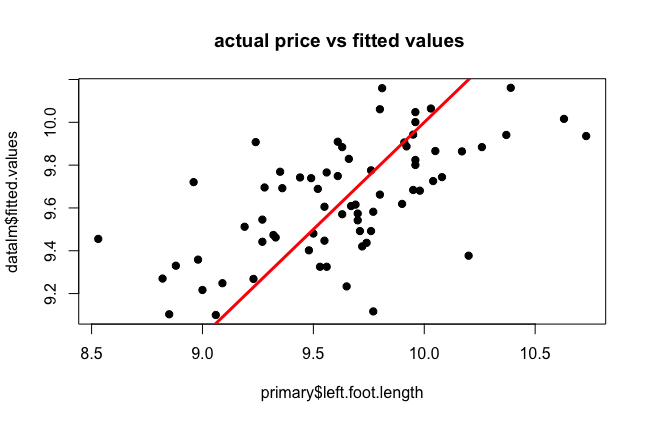
However, there are other parameters that explain length in foot as R square value is 0.38 .

c)

#Lineraity

plot(primary$left.foot.length,datalm$fitted.values,pch=19,main="actual price vs fitted values ")

abline(0,1,col="red",lwd=3)

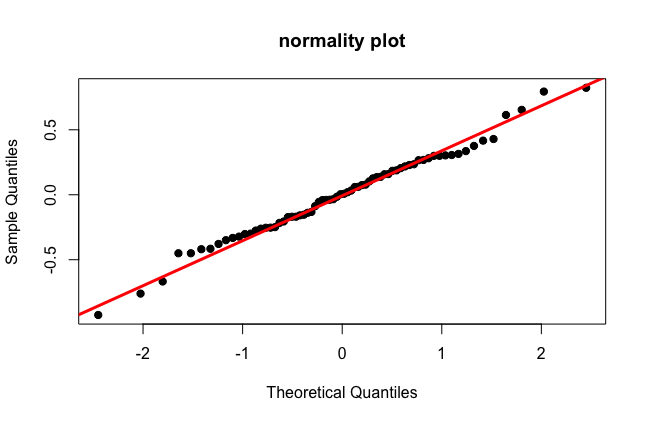


From the above data it can be seen that the data is linear and linearity assumption is satisfied.

#normality

qqnorm(datalm$residuals,pch=19,main = "normality plot")

qqline(datalm$residuals,col="red",lwd=3)

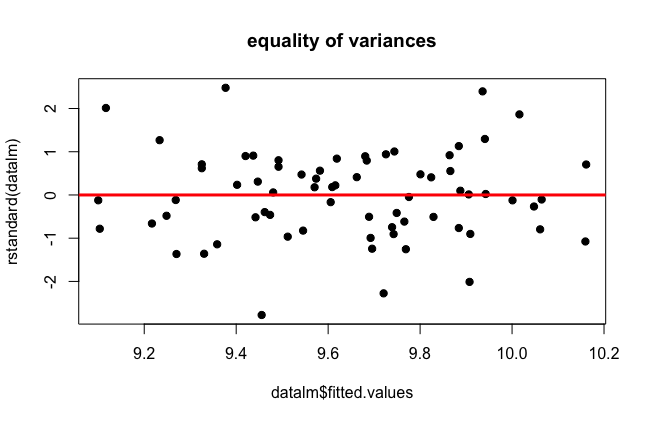


From the above graph it can be seen that though there are few outlines, the qqnorm and qqline mostly coincide which can be used to say normality assumption is satisfied.

#equality of varience

plot(datalm$fitted.values,rstandard(datalm),pch=19,main = "equality of variances")

abline(0,0,col="red",lwd=3)



From the above data fitted values vs standard , we can say that with the change in fitted values rstandard is also changing hence this is heteroscedatic

D)

min(height)

max(height)

mean(height)

> min(height)

[1] 59.84

> max(height)

[1] 74.84

> mean(height)

[1] 68.38197

pred1=data.frame(height=5.5\*12)

predict(datalm,pred1,interval = "confidence")

> pred1=data.frame(height=5.5\*12)

> predict(datalm,pred1,interval = "confidence")

fit lwr upr

1 9.24179 9.095867 9.387714

predict(datalm,pred1,interval = "predict")

> predict(datalm,pred1,interval = "predict")

fit lwr upr

1 9.24179 8.553879 9.929701

The length of his left foot would be 9.24 inches. The length is same in both the case , but while predicting we have a wider interval when compared to confidence . This is due to the reason that predict is for a single value, whereas the confidence is for average mean of the value.

3)

min(height)

max(height)

pred3=data.frame(height=48)

predict(datalm,pred3,interval = "predict")

> min(height)

[1] 59.84

>

> max(height)

[1] 74.84

>

> pred3=data.frame(height=48)

>

> predict(datalm,pred3,interval = "predict")

fit lwr upr

1 6.238409 4.996459 7.48036

A son whose height is 4ft would be equal to 48 inches, whereas our model has a range of value between 59.84 and 74.84 and mean 68.38, So the value 48 inches would be an outlier to the data due to which the model prediction would not be accurate. Hence, we cannot predict the length of the left foot accurately.