**#QMB Assignment 4**

**#Name – Manish Sanjay Talekar**

**#U54500765**

**#Preprocessing**

#q1.

> master\_data <- read\_xlsx("6304 Module 4 Assignment Data.xlsx",

+ sheet = 'Tamil Anatomy')

> colnames(master\_data)=tolower(make.names(colnames(master\_data)))

> attach(master\_data)

#q2

> set.seed(54500765)

> primary\_data <- sample\_n(master\_data, 70)

**#Analysis**

#q1.

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

[1] 0.5548621

INTERPRETATION – Correlation between the two variables exists though it is not a very strong relation.

#q2.

> lm\_out <- lm(left.foot.length~height, data = primary\_data)

#q2.a

> summary(lm\_out)

Call:

lm(formula = left.foot.length ~ height, data = primary\_data)

Residuals:

Min 1Q Median 3Q Max

-0.81692 -0.24075 0.01019 0.28136 0.63057

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.53487 1.64977 0.324 0.747

height 0.13328 0.02423 5.500 6.21e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3695 on 68 degrees of freedom

Multiple R-squared: 0.3079, Adjusted R-squared: 0.2977

F-statistic: 30.25 on 1 and 68 DF, p-value: 6.212e-07

> confint(lm\_out)

2.5 % 97.5 %

(Intercept) - 2.75718648 3.8269329

height 0.08492451 0.1816418

REPORT –

Coefficient B0 (Intercept): 0.53487 P value - 0.747 Confidence Interval – (-2.75718648, 3.8269329)

Coefficient B1 (Height): 0.13328 P value - 6.21e-07

Confidence Interval – (0.08492451, 0.1816418)

#q2.b

INTERPRETATION –

Equation <= left.foot.length = 0.53487 + 0.13328 \* height

1. For height of 0-inch; the left foot length will be 0.53487 inches.
2. For every 10-inch increase in height; the left foot length will increase by 1.3328.

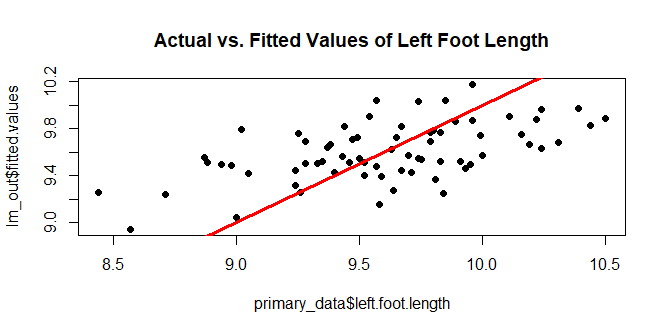
#q2.c

#Linearity

> plot(primary\_data$left.foot.length , lm\_out$fitted.values, pch=19,

+ main = "Actual vs. Fitted Values of Left Foot Length")

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

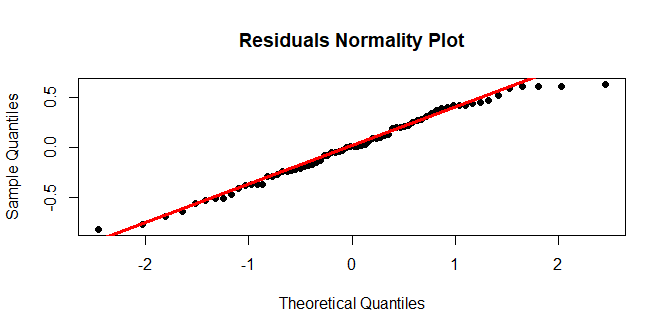


INTERPRETATION – based on the above plot I can conclude that the data is not linear as almost all the points lie outside the linear line. Therefore, the model does not conform the linearity assumption.

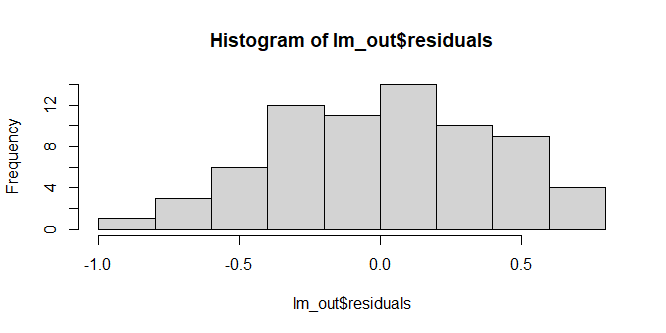
#Normality

> qqnorm(lm\_out$residuals,pch=19,main="Residuals Normality Plot")

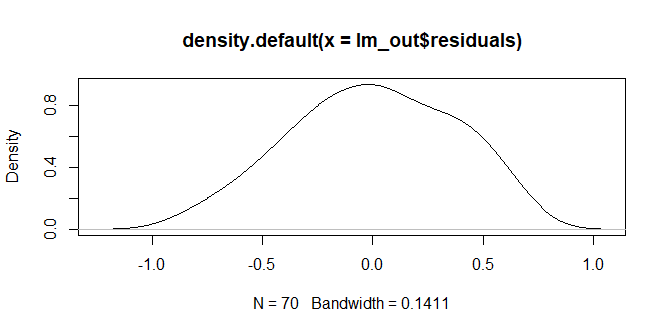
> qqline(lm\_out$residuals,col="red",lwd=3)



> hist(lm\_out$residuals)



> plot(density(lm\_out$residuals))



> skewness(lm\_out$residuals)

[1] -0.1666593

> kurtosis(lm\_out$residuals)

[1] 2.270872

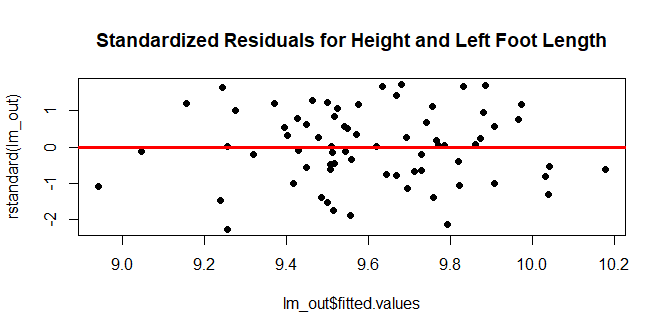
INTERPRETATION – Based on above analysis I can conclude that the residuals are not perfectly normally distributed as it is little skewed to left. Therefore, I can conclude that the assumption of normality is failed.

#Equality of Variances

> plot(lm\_out$fitted.values,rstandard(lm\_out),pch=19,

+ main="Standardized Residuals for Height and Left Foot Length")

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



INTERPRETATION – as seen on the above graph, there does not seem to be any pattern between the residuals and therefore we can conform the assumption of equality of variances.

#q2.d

> newdata=data.frame(height=66) #5.5 feet = 66 Inches

> predict(lm\_out,newdata,interval="predict")

fit lwr upr

1 9.33156 8.582477 10.08064

INTERPRETATION – Gives 95% likelihood interval (8.582477, 10.08064) for the individual occurrence of the height variable specified(5.5 feet). Therefore, the interval is very wide.

> predict(lm\_out,newdata,interval="confidence")

fit lwr upr

1 9.33156 9.198867 9.464253

INTERPRETATION - Gives 95% likelihood confidence interval (CI) for the sample, whose mean will lie in this CI of (9.198867, 9.464253). Therefore, the interval is tight. For example, when many samples of height 66 inches are expected in the sample data the mean of the sample will lie in this CI.

#q3.

> boy=data.frame(height=48) #4 feet = 48 Inches

> predict(lm\_out,boy,interval="predict")

fit lwr upr

1 6.932464 5.711179 8.153748

> predict(lm\_out,boy,interval="confidence")

fit lwr upr

1 6.932464 5.958802 7.906126

INTERPRETATION – We have got the prediction value for the left foot length of 6.932464 however, the confidence interval is very wide and therefore we cannot trust the prediction of the model. Additionally, the model is trained with the data of Tamil Adults and we are trying to predict the left foot length for the Tamil Child which is out of the scope/bound for this model.