Explanatory Data Analysis (econometrics) on different datasets in R

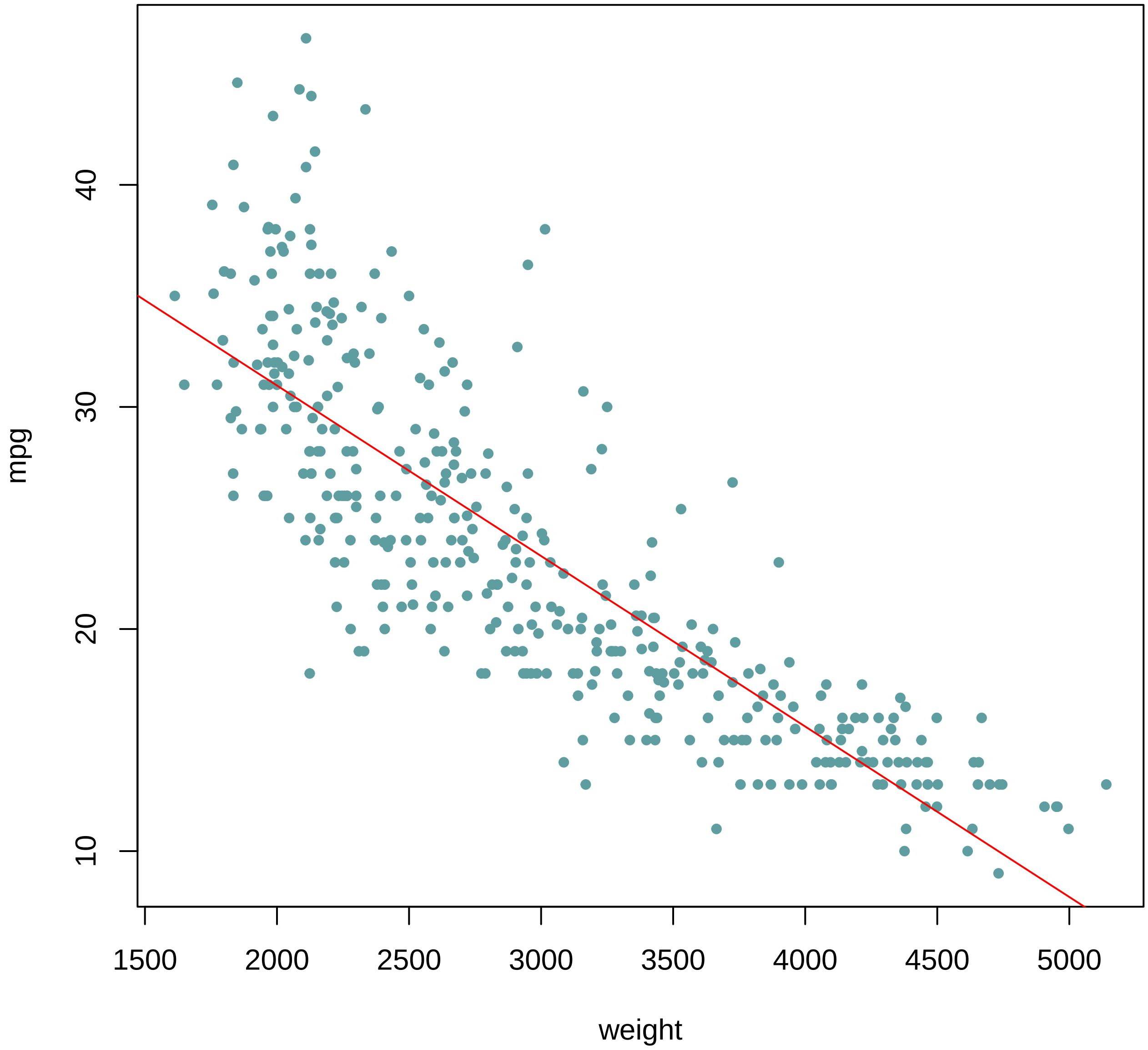
Pratyusha Bala
STAT019
Regn no.-19214220019
SEM-6, UG-3
Presidency University

7/6/2022

Dataset 1:

The dataset auto-mpg.csv, comes from the 1983 American Statistical Association Exposition. The response variable of interest is fuel consumption, measured in miles per gallon.Also attribute of cars like the weight was also recorded for each car. We will study the relationship between mpg and weight (in lbs).

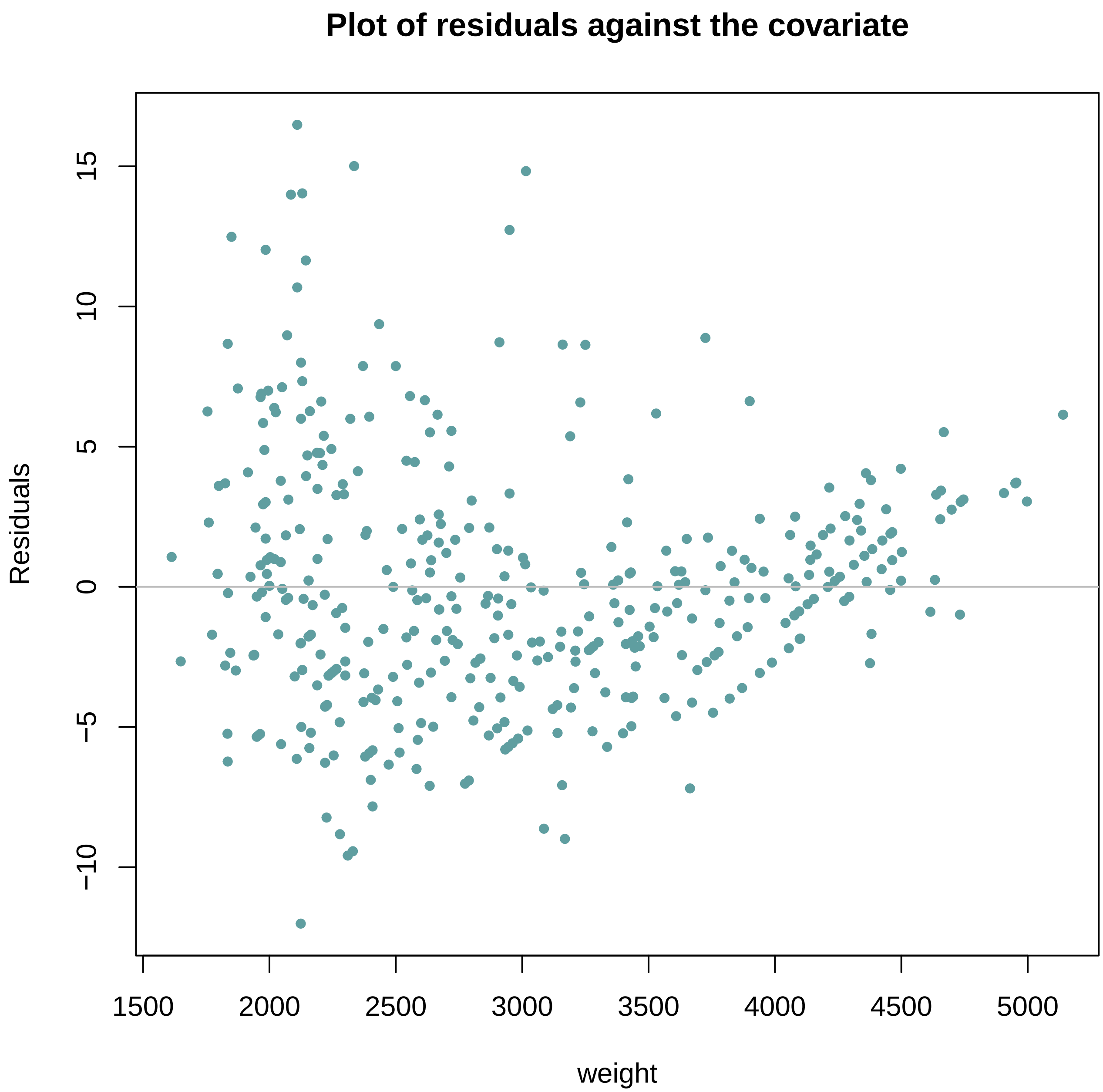
At first we fit a linear regression model to this data



Then the regression model is:

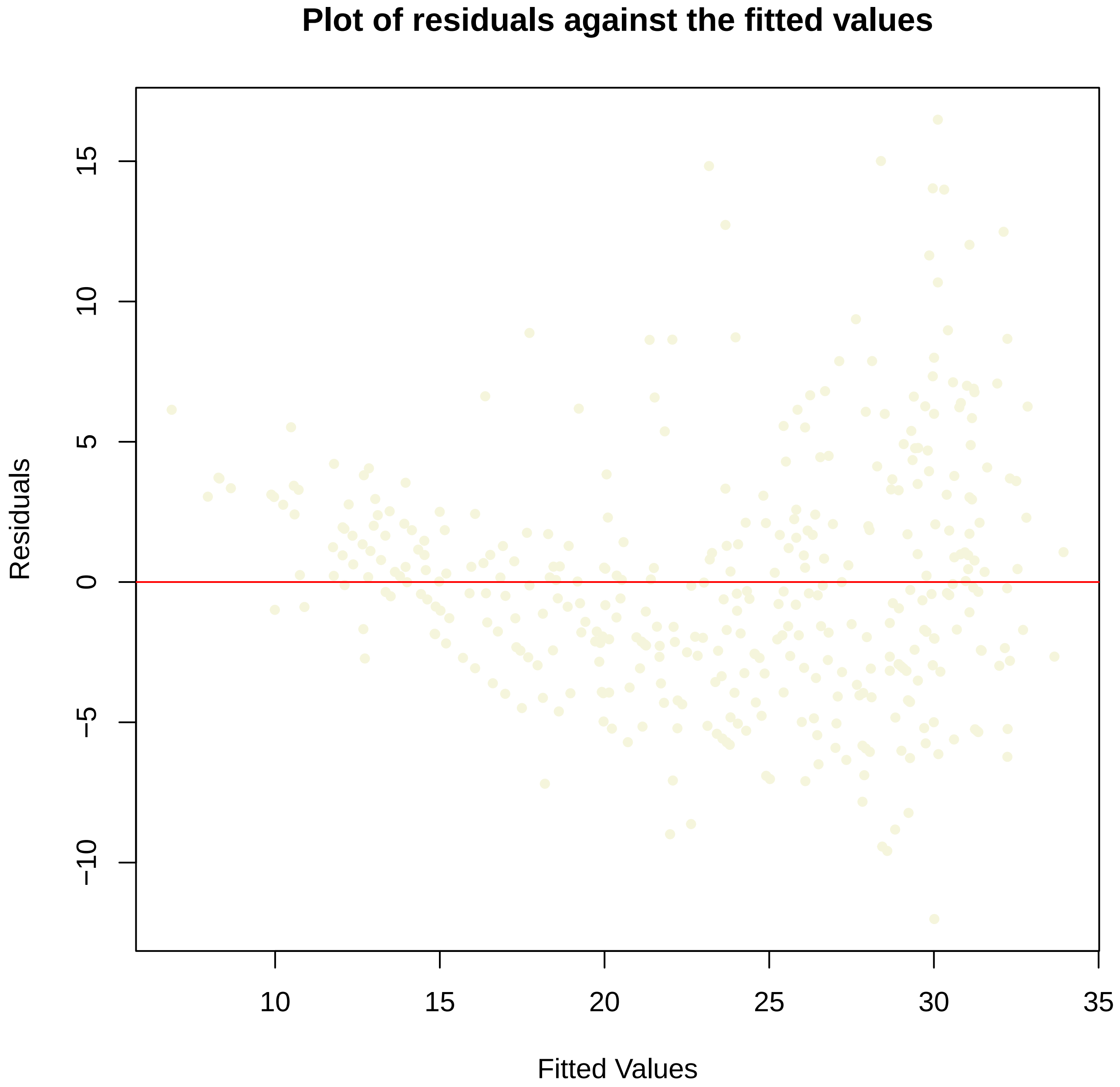
$$\hat{y}_i = 46.317364 - 0.007677 x_i \forall i$$

Now we make a plot of residuals against the covariate to check if this linear regression model would be a good fit for this data or not.



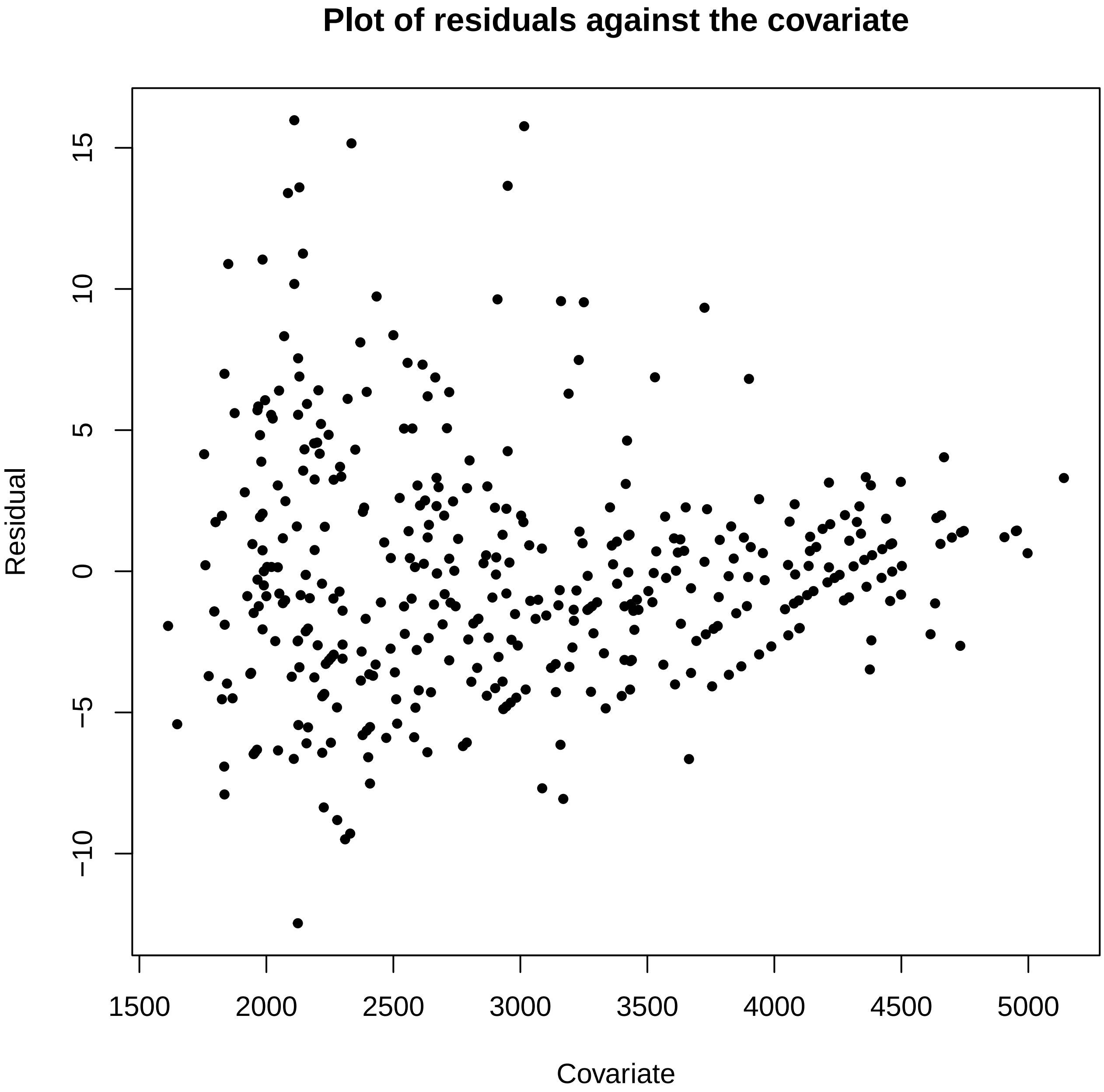
This plot suggests that the linearity assumption is wrong, since clearly there is curved or stepped patterns here, a lot of deviated from the line of 0 slope.

Now we make a plot of residuals against the fitted values.



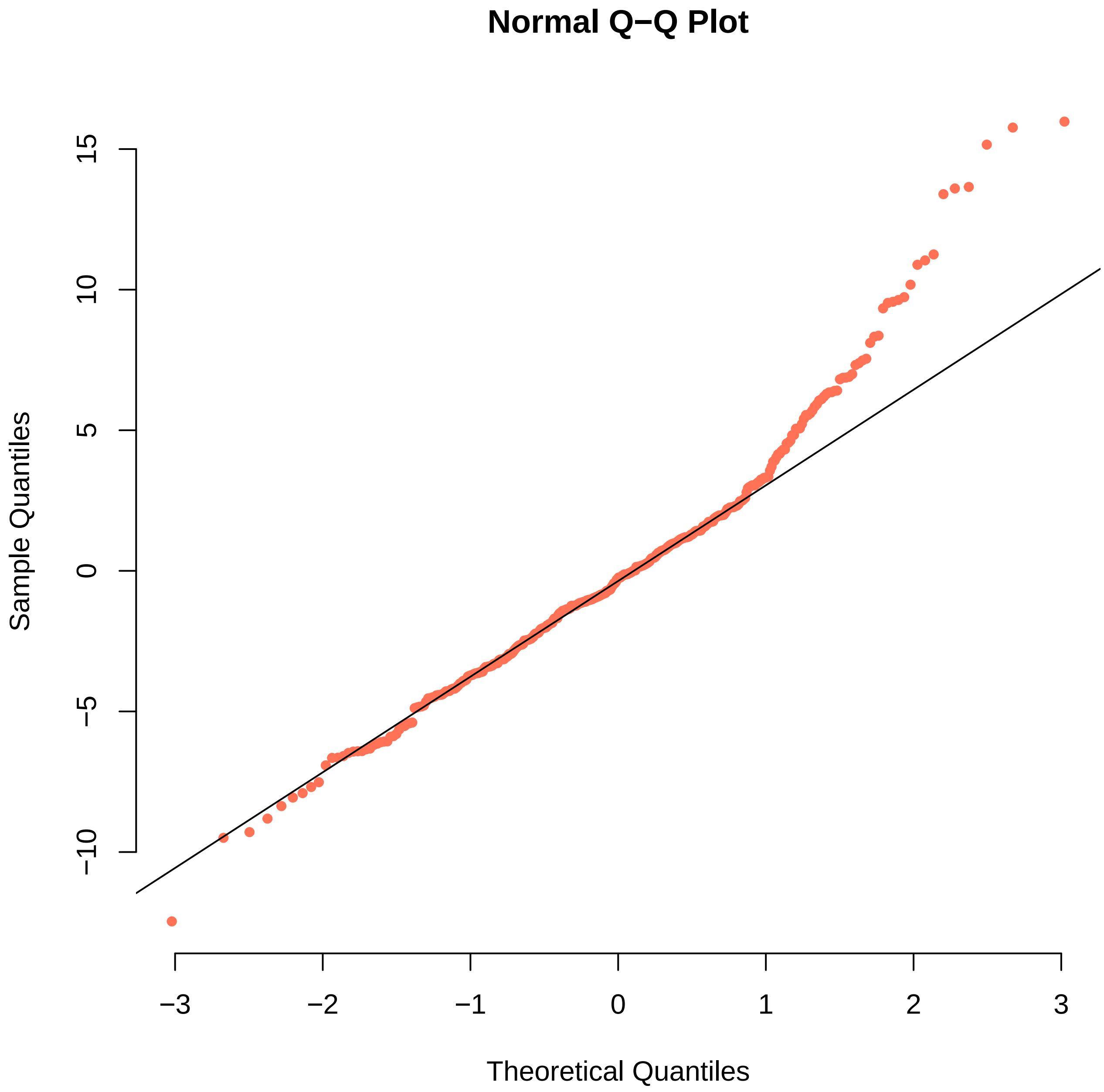
We can see that these two plots are synonymous since both have curved patterns to themselves,a lot of deviated from the line of 0 slope, indicating that the linearity assumption is wrong.

If we apply the log transformation on weight, refit the linear regression model, and produce a new residual plot:



The plot still suggests that the linearity assumption is incorrect due to the same reasons as before .

Let us check if any of our assumptions are violated.

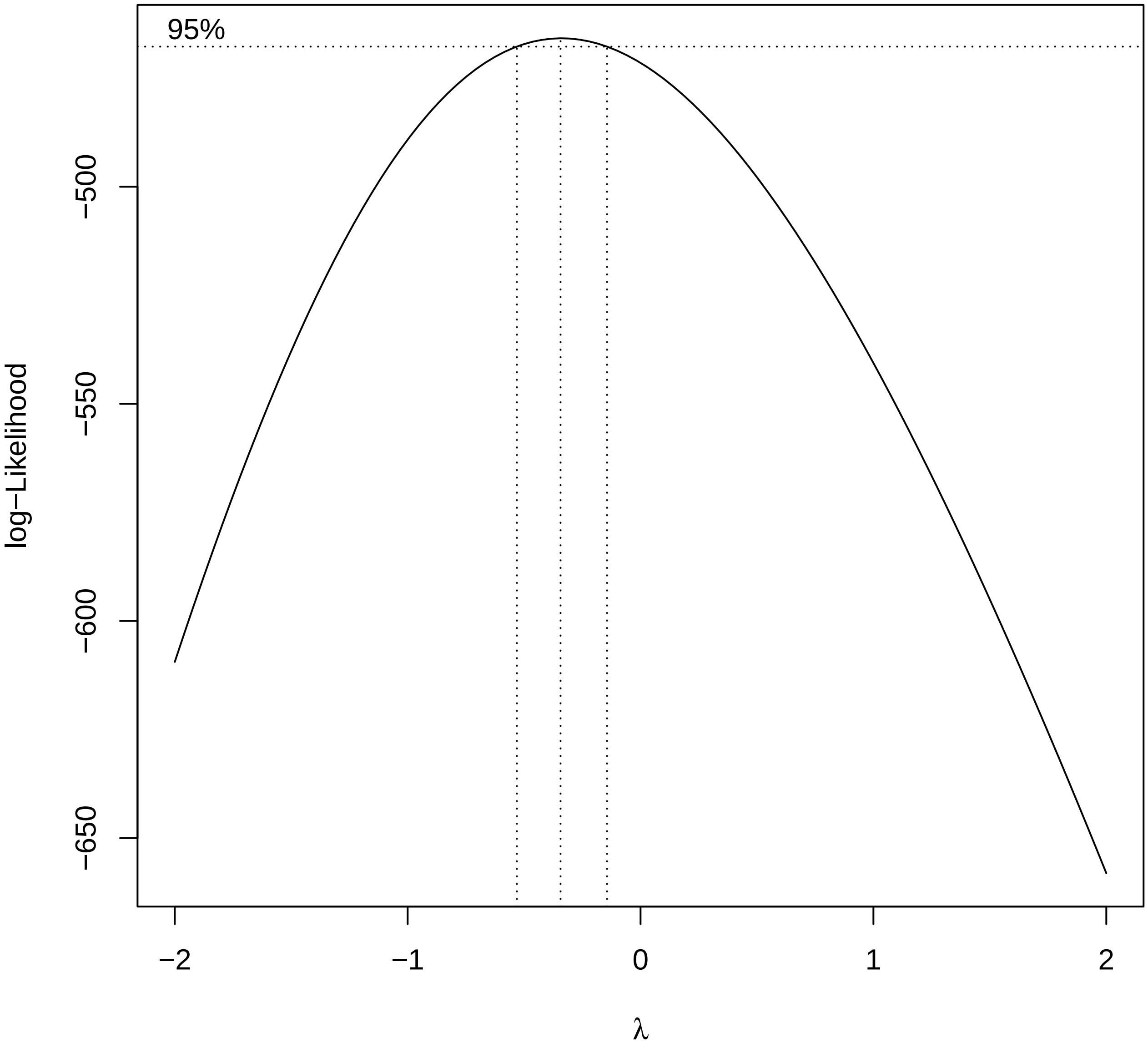


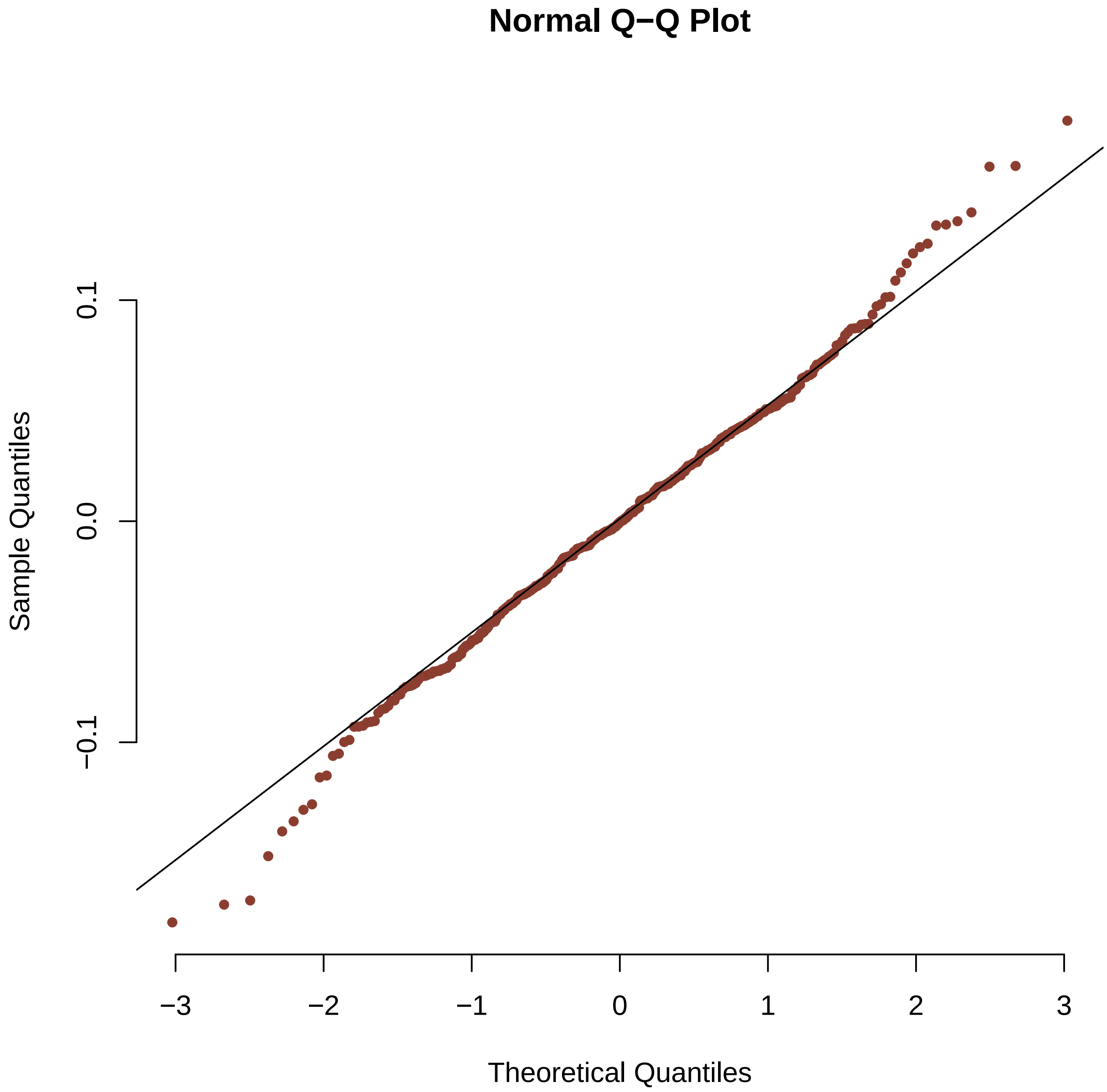
Thus the residual of the linear model does not satisfy the normality assumption.

At last, let us use the *Box-Cox procedure* to estimate the best power transformation for the response variable. Box-Cox transformation:

$b_y(y) = (y^\lambda - 1)/\lambda$

in the limit $\lambda \to 0$, this becomes $\log y$


95%
log−Likelihood
−500
−550
−600
−650
−2
−1
0
1
2
λ



After evaluating the new residual plot and normal probability plot again, we can say that the normality assumptions are better satisfied now. But the linearity assumption of the regression model is incorrect as there is still a visible stepped pattern in the plot of residuals vs. covariate.

The final model is the one we got from box-cox procedure which is

$$y = 2.2693267 - 0.0001239x + \epsilon$$

i.e The interpretation of the intercept would be when x=0 i.e at 0 weight, the fuel consumption (y) of a car is on an average 2.2693267 mgp, which is absurd. Thus, the interpretation of the intercept is misplaced here. Whereas, with unit change in weight of a car, its fuel consumption decreases by 0.0001239 mgp on an average.

Now we are interested to test whether or not there is a linear association between mpg and weight or not.

$H_0 : \hat{\beta} = 0$ **vs** $H_1 : \hat{\beta} \neq 0$ in the above model

We accept the null hypothesis if the p-value is greater that the level of significance,(which is let $\alpha = 0.05$). Here p-value $< 2^{-16}$

Conclusion: As the p-value is much less than 0.05, we reject the null hypothesis, i.e. there is a signicant relation between the variables in the linear regression model of the dataset.at 5% level of significance.

Lastly, we find a 90% confidence interval for $\hat{\beta}$

```
                        5 %          95 %
(Intercept)  2.2524782725  2.2861751007
weight      -0.0001293669 -0.0001184565
```

Dataset 2:

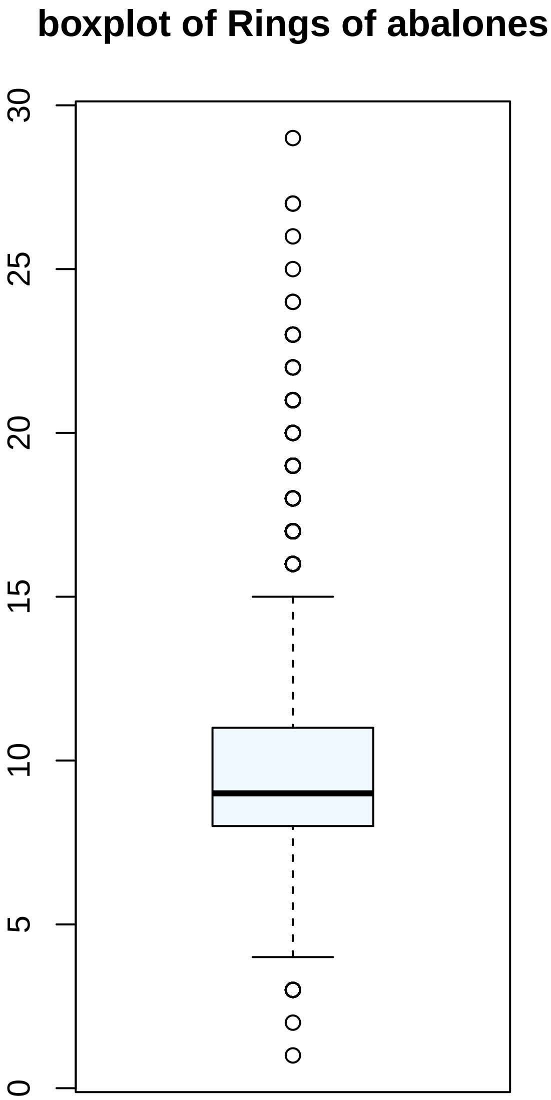
Research Problem:

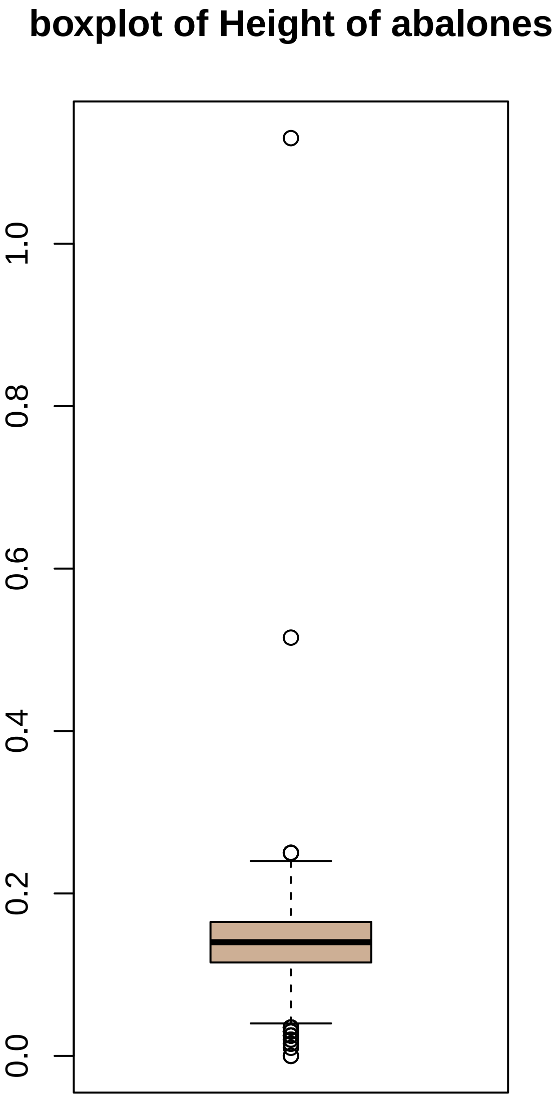
The topic of interest here is, abalones,one type of reef-dwelling marine snails, a common name for any of a group of small to very large marine gastropod molluscs in the family Haliotidae which is reknowned for being a delicacy around the world .The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope which is a tedious and time-consuming task. There are other easily attainable measurements which can be used to predict the age, in this study, the height of the abalone is of our interest.We will to build a predictive model of abalone's age from height (in mm.).

Research hypothesis:

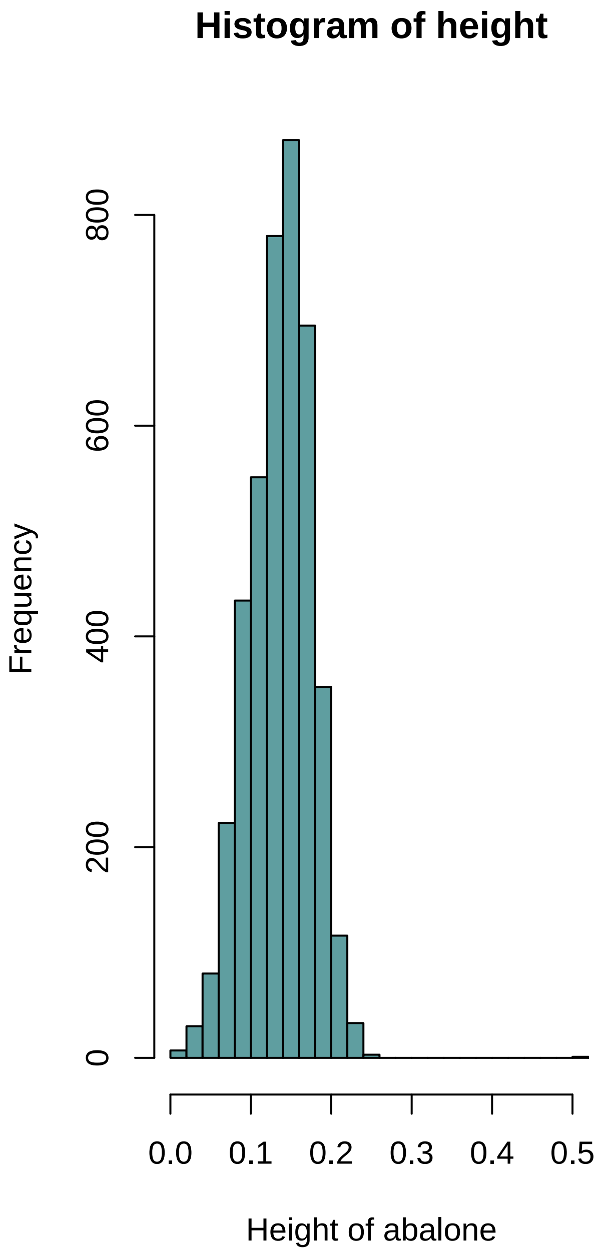
Here the research hypothesis would be that a simple linear regression model with normal error assumption is appropriate to describe the relationship between the height of abalones and their ages. Sources: (https://en.wikipedia.org/wiki/Abalone),(https://archive.ics.uci.edu/ml/datasets/Abalone.)

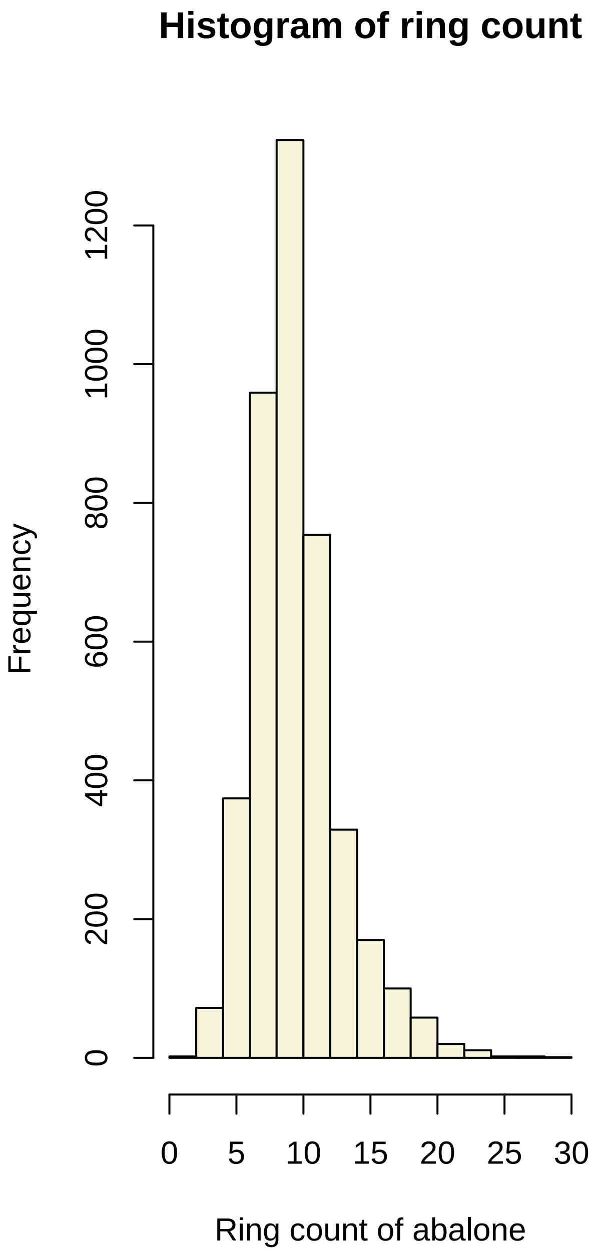
We perform EDA on the two variables individually (univariate) for starters,


boxplot of Rings of abalones

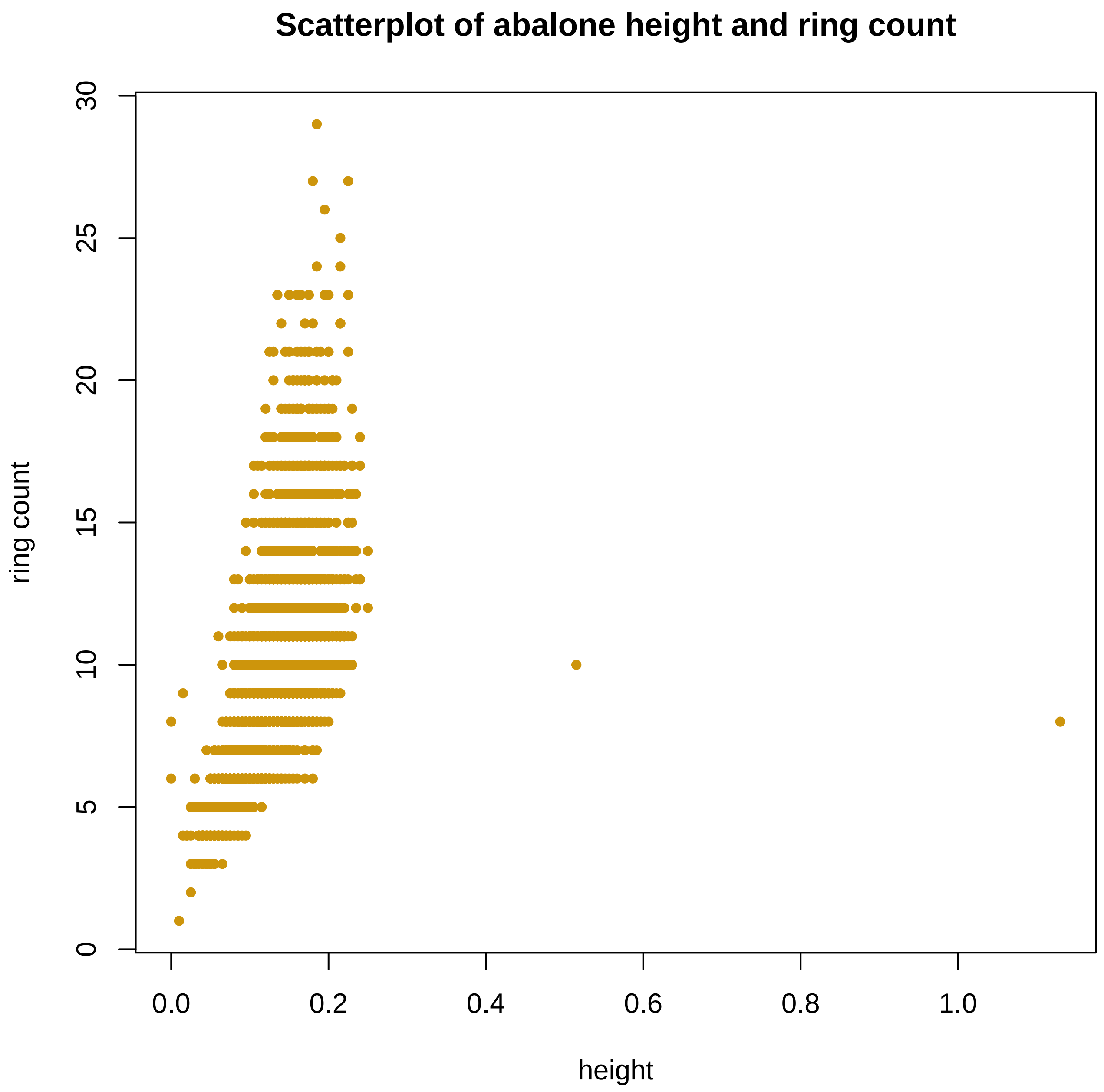

boxplot of Height of abalones

There are two potential outliers in the covariate: height 0.515 mm and 1.13 mm.This regressor has mean 0.1395 mm, with variance 0.001749084. The number of rings ranges from 1 to 29, with only one abalone having 29 rings. The mean of the ring count is 9.934, and the median 9, with variance 10.39278.



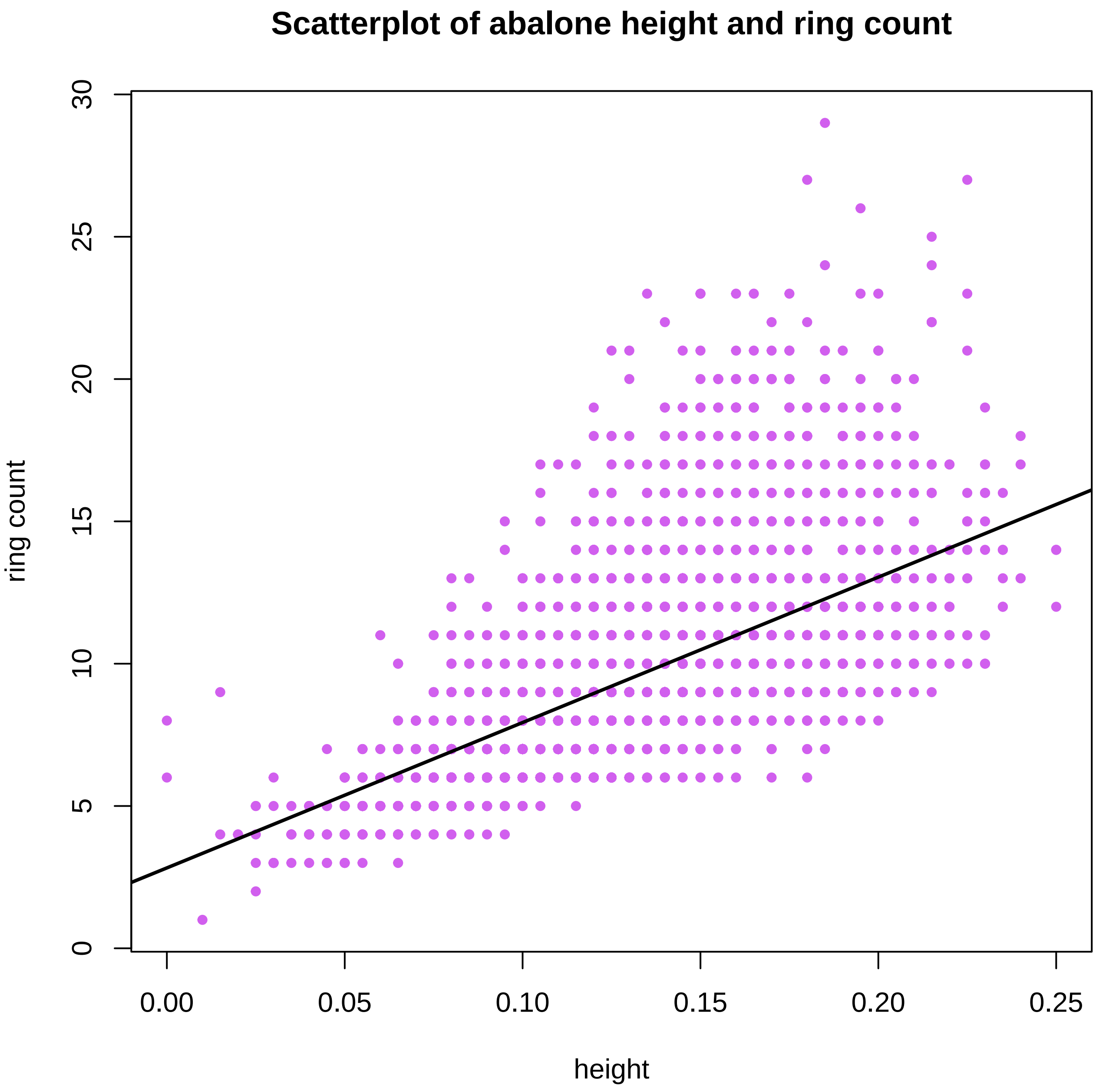


The distributions of height and ring count are positively skewed and leptokurtic.

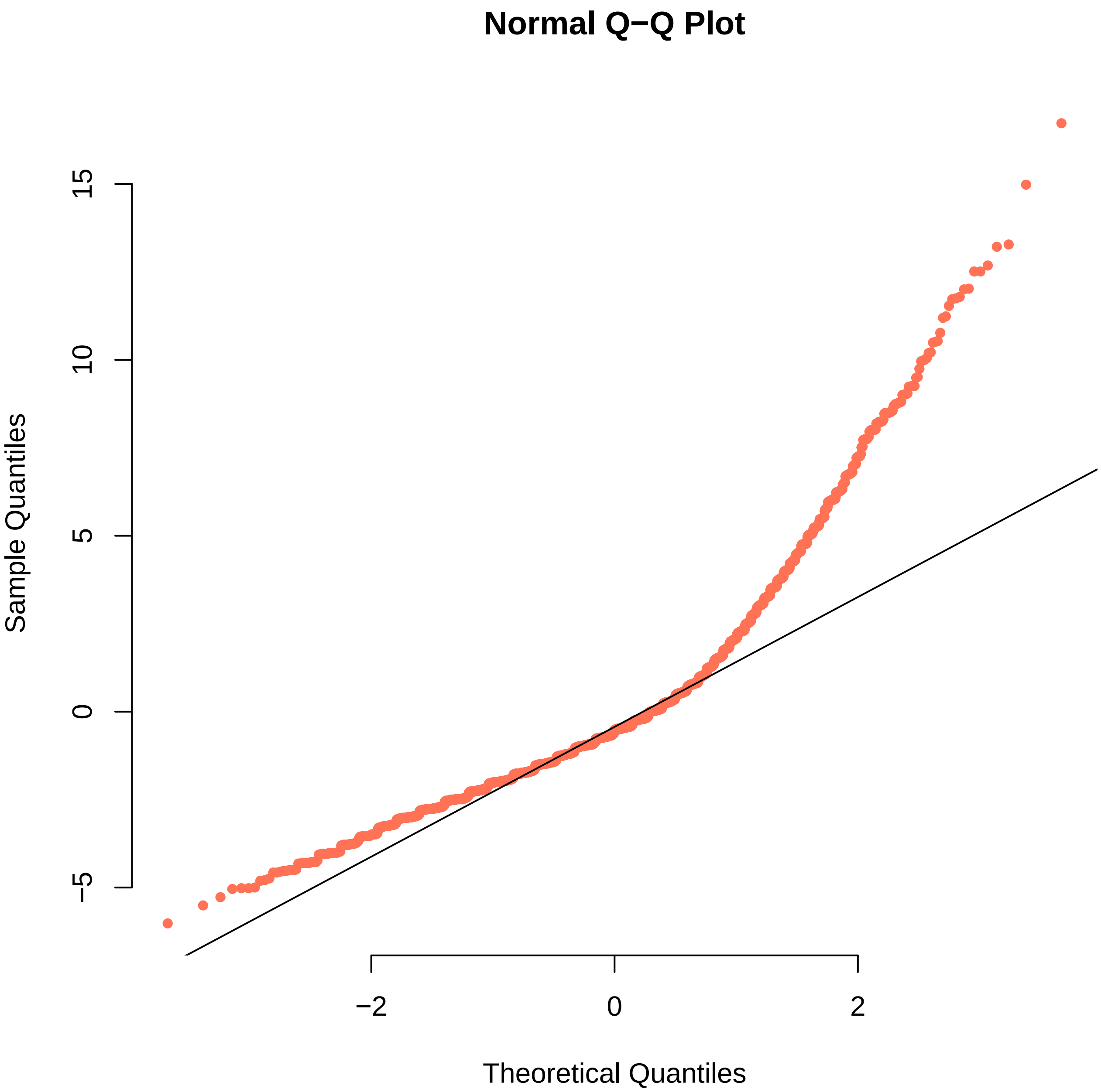


It is actually quite apparent that if we exclude the two outliers having extreme height, the relationship can be assumed to be positively correlated, since with increase in height we can expect increase in ring count which in turn means increase in age. Let us actually exclude the two outliers and fit a simple linear regression. If we denote Ring count by Y and Height by X, then the estimated model is

$$\hat{y} = 2.825 + 51.078x$$



The data does not show any signs of linearity, rather it seems that it would fit other non-linear regression. Let us check whether the model assumptions are met or not.

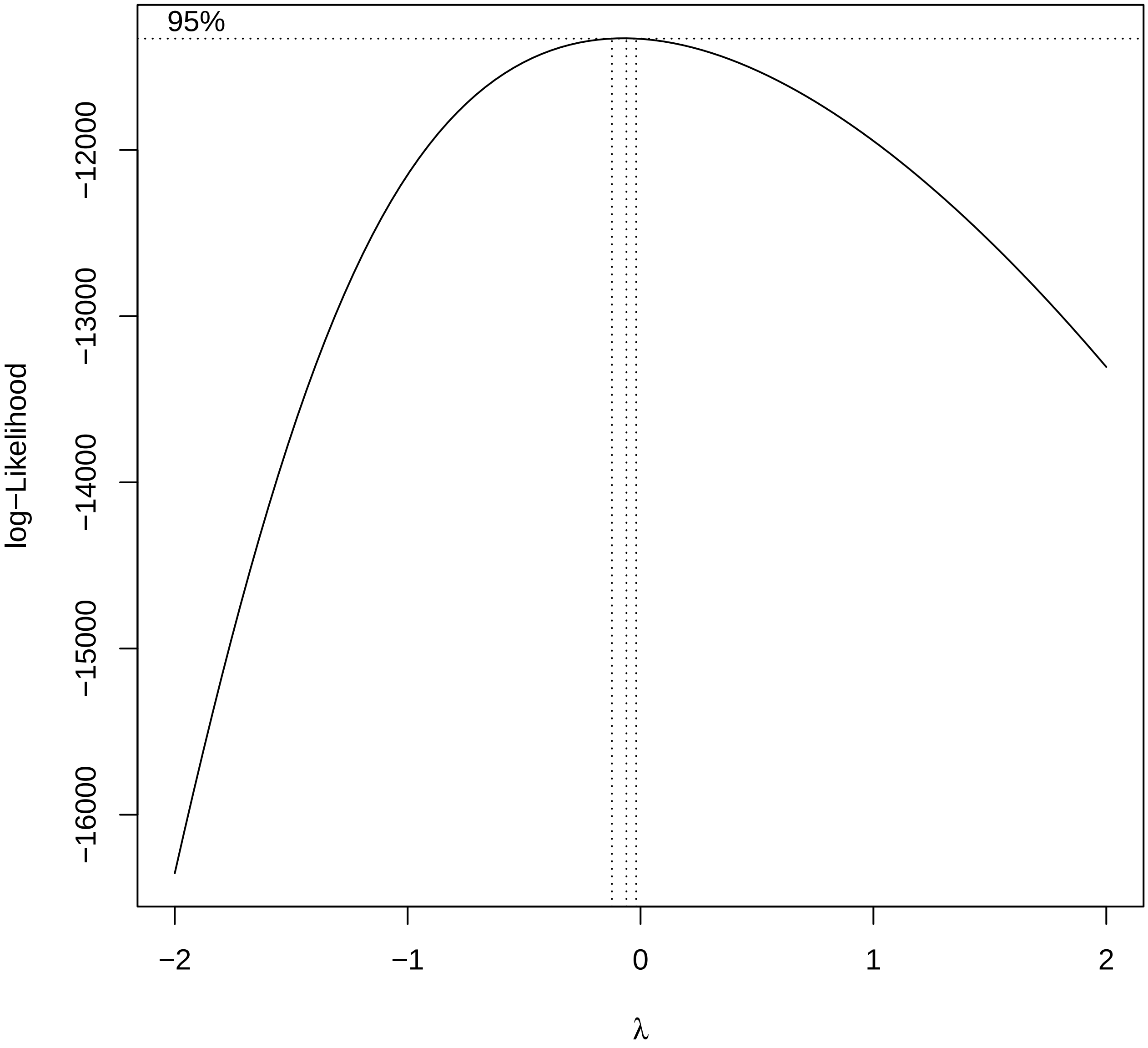


Thus the residual of the linear model does not satisfy the normality assumption and our linearity assumptions are wrong.

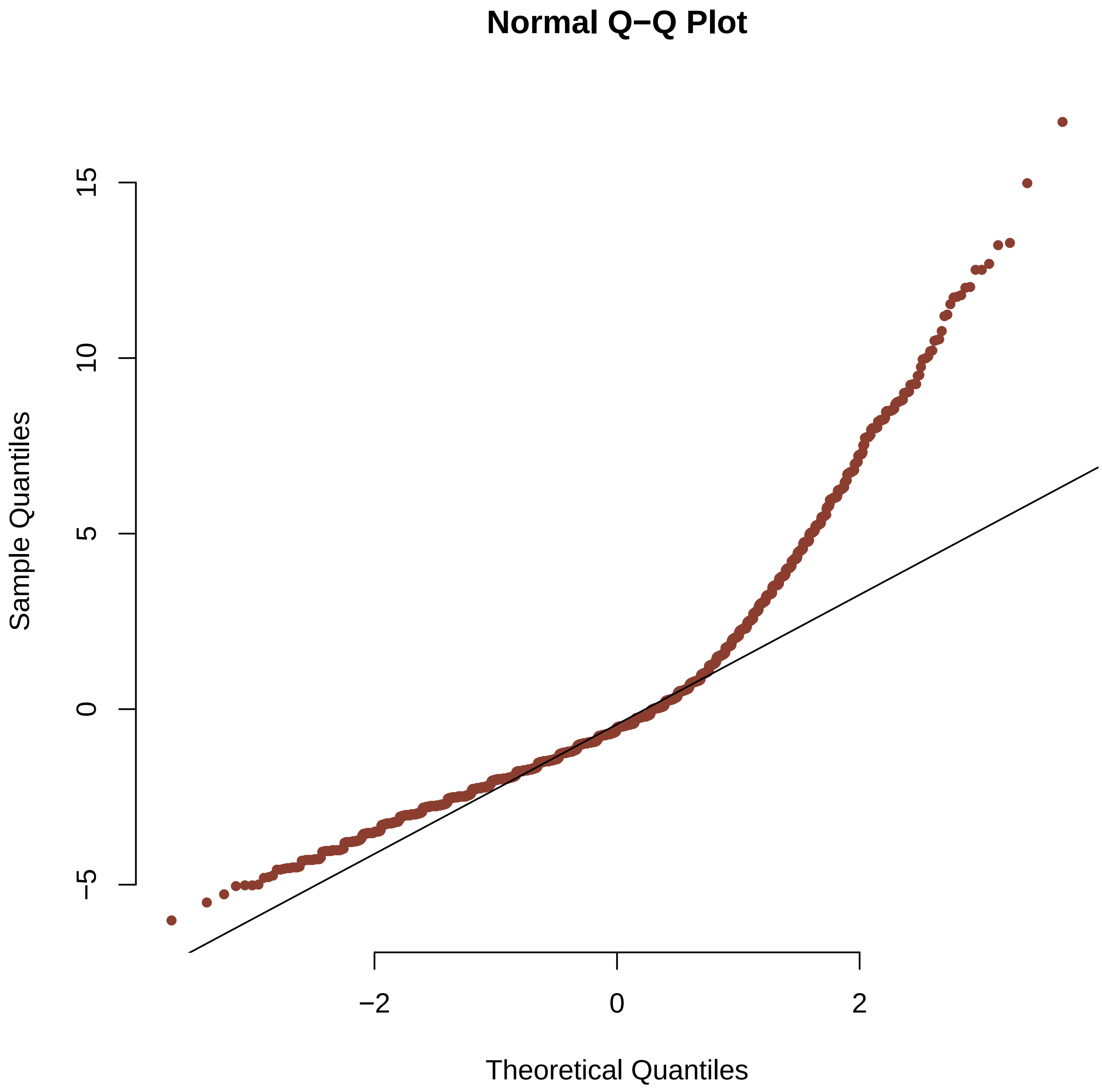
let us use the *Box-Cox procedure* to estimate the best power transformation for the response variable.

$b_y(y) = (y^\lambda - 1)/\lambda$

in the limit $\lambda \to 0$, this becomes $\log y$



```
## Error in eval(expr, envir, enclos):  object 'Height' not found
```



After evaluating the new residual plot and normal probability plot again, the normality assumptions did not improve at all. Hence the regression model is incorrect as there is still a visible stepped pattern in the qqplot of residuals.

Interpretation of final parameters: α : denotes the ring count when Height = 0, but it has no scientific interpretation. So in this case we can interpret the value $\hat{\alpha}$= 2.825 as the baseline number of rings for an abalone, but then again there are sample points for which ring count is less than 2.825. So, there is this dilemma.

β : denotes the change in expected number of rings with unit change in the value of height. Now, $\hat{\beta}$=51.078 means that with 1mm increase in height means 51 increase in ring count. The 95% confidence interval associated with height=0.3 is (17.81473,18.48131).

A t-test for a significant association between ring count and height yields a p-value less than 2.2^{-16} (with *** meaning significant at <0.001 level), so we can conclude there is a significant relationship between the height and number of rings of abalones,at 5% level of significance, just not a simple linear relationship .While we can use the height measurement to predict their ages using our linear regression line, the error is too large to ignore. Thus this relationship is not statistically significant.

A 95% confidence interval for the average number of rings for abalones with height at 0.128 mm. is (9.275917, 9.437726). The point estimate for the average number of rings for abalones with height at 0.128mm is 9.362609. The predicted value and a 99% prediction interval for the number of rings for an abalone with height at 0.132mm is 9.566921 and (4.554,14.5797) respectively.

Conclusions:-

We conclude that the height of an abalone is not that useful in predicting the number of rings, and thus the abalone's age. Even though there exists quite a positive association between them, simple linear regression does not seem to fit the data and thus it is not statistically significant enough for us to predict the number of rings using some of abalones' height measurement.

R codes:

Dataset 1

```
a=read.csv("C:\\Users\\pratyusha\\Downloads\\R datasets\\auto-mpg.csv",header=T,sep=",")
summary(a)
model=lm(mpg~weight,data=a)
plot(mpg~weight,data=a,pch=20,col="cadetblue")
abline(model,col="red")

plot(a$weight, residuals(model), pch=20, ylab="Residuals", xlab="weight", col="cadetblue", main="Plot of resi
abline(h = 0, col = "grey")#...a.i.A

plot(predict(model),residuals(model), pch=20, xlab="Fitted Values", ylab="Residuals", main="Plot of residuals
abline(h = 0, col = "red")#...a.i.B

model2=lm(mpg~log(weight),data=a)
plot(a$weight,residuals(model2),pch=20,ylab="Residual", xlab="Covariate", col="black", main="Plot of residual
qqnorm(residuals(model2), pch=20 , col="coral1", bty="n")
qqline(residuals(model2))#...a.iii

library(MASS)
boxcox(mpg~weight,data=a)
```

```
model3 = boxcox(mpg~weight, data=a
lambda = model3$x[which.max(model3$y)]
lambda
a$mpg=(a$mpg^lambda - 1)/lambda
model4 = lm(mpg~weight, data=a)
model4
plot(a$weight,residuals(model4),pch=20, ylab="Residual", xlab="Covariate", main="Plot of residuals against th
qqnorm(residuals(model4), pch=20 , col="coral4", bty="n")
qqline(residuals(model4))#...a.iv

summary(model4)
confint(model4, level=0.9)#...b
```

dataset 2

```
aba=read.table("C:\\Users\\pratyusha\\Downloads\\R datasets\\abalone.csv", header=T, sep=",")
summary(aba)
aba[aba$Height>"0.25",]
var(aba$Height)*4176/4177
var(aba$Rings)*4176/4177
par(mfrow=c(1,2))
boxplot(aba$Rings,col="aliceblue",main="boxplot of Rings of abalones")
boxplot(aba$Height,col="peachpuff3",main="boxplot of Height of abalones") #...a

par(mfrow=c(1,2))
hist(aba$Height, breaks =80, main = "Histogram of height", xlab="Height of abalone", col = "cadetblue", xlim=
hist(aba$Rings, main = "Histogram of ring count", xlab="Ring count of abalone", col = "beige")#...b

par(mfrow=c(1,1))
plot(aba$Height,aba$Rings,pch=20,col="darkgoldenrod3",xlab="height",ylab="ring count",main="Scatterplot of ab

a=aba[-which(aba$Height>"0.25"),]
plot(a$Height,a$Rings,pch=20,col="mediumorchid2",xlab="height",ylab="ring count",main="Scatterplot of abalone
model=lm(Rings~Height,data=a)
abline(model,lwd = 2)#...d

qqnorm(residuals(model), pch=20 , col="coral1", bty="n")
qqline(residuals(model))#...e

library(MASS)
boxcox(Rings~Height,data=a)
model2 = boxcox(Rings~Height,data=a)
lambda = model2$x[which.max(model2$y)]
```

```
a$Rings=(a$Rings^Height - 1)/lambda
model3 = lm(Rings~Height, data=a)
model3
plot(a$Height,residuals(model3),pch=20, ylab="Residual", xlab="Covariate", main="Plot of residuals against th
qqnorm(residuals(model3), pch=20 , col="coral4", bty="n")
qqline(residuals(model3))#...f

predict(model, newdata = data.frame(Height=c(0.2,0.3,0.4)), interval = "confidence")#...g

anova(model)#...h

predict(model, newdata = data.frame(Height=c(0.128)), interval = "confidence")
predict(model, newdata = data.frame(Height=c(0.128)))#...i

predict(model, newdata = data.frame(Height=c(0.132)))
predict(model, newdata = data.frame(Height=c(0.132)), interval = "prediction")#...j
```