Statistical Data Analysis

Mrunal Dhiwar

June 5, 2022

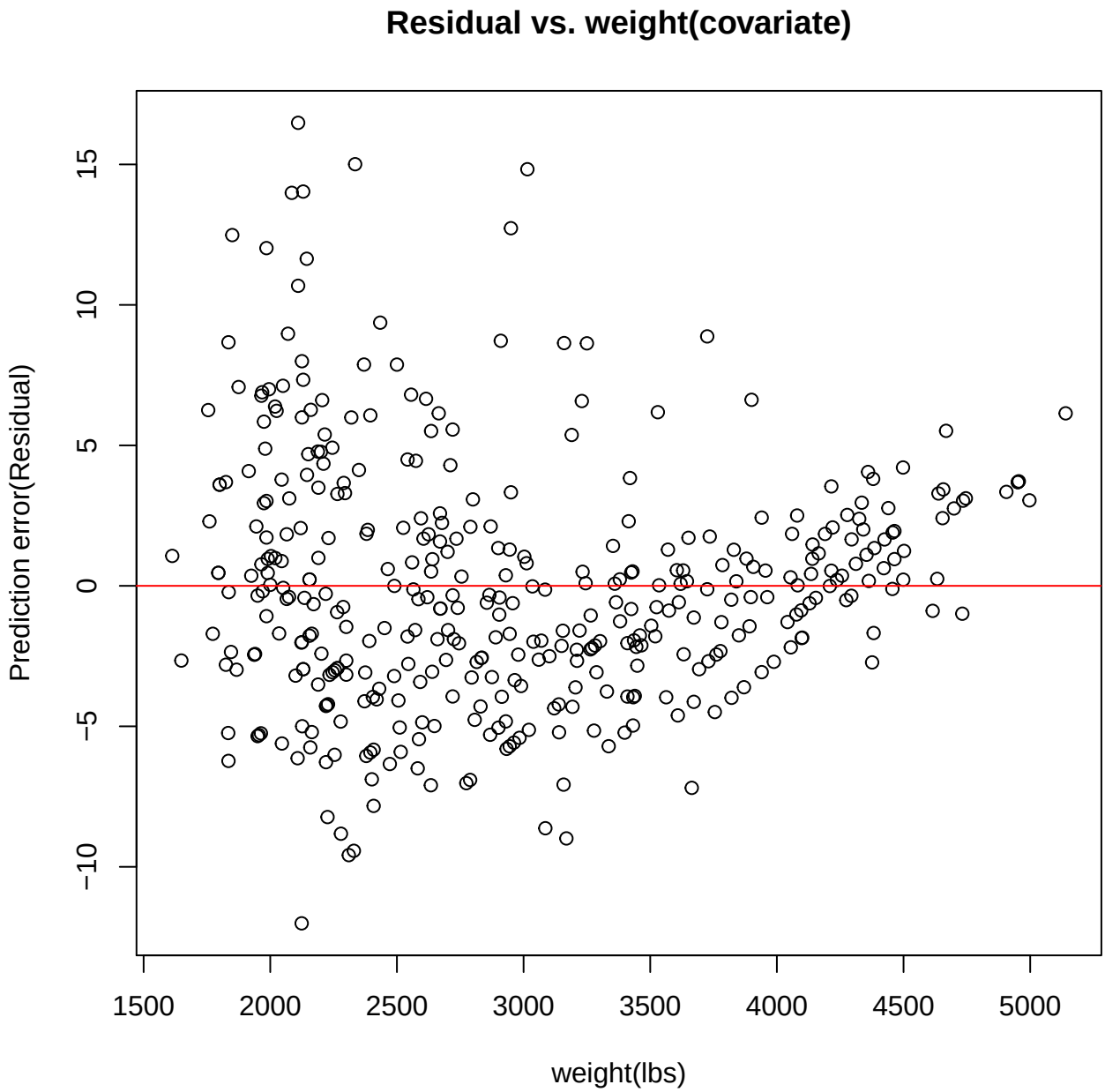
❑ Analysis of the auto-mpg dataset:

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. Other attributes of cars like the weight, horsepower, number cylinders, and acceleration time were also recorded for each car.

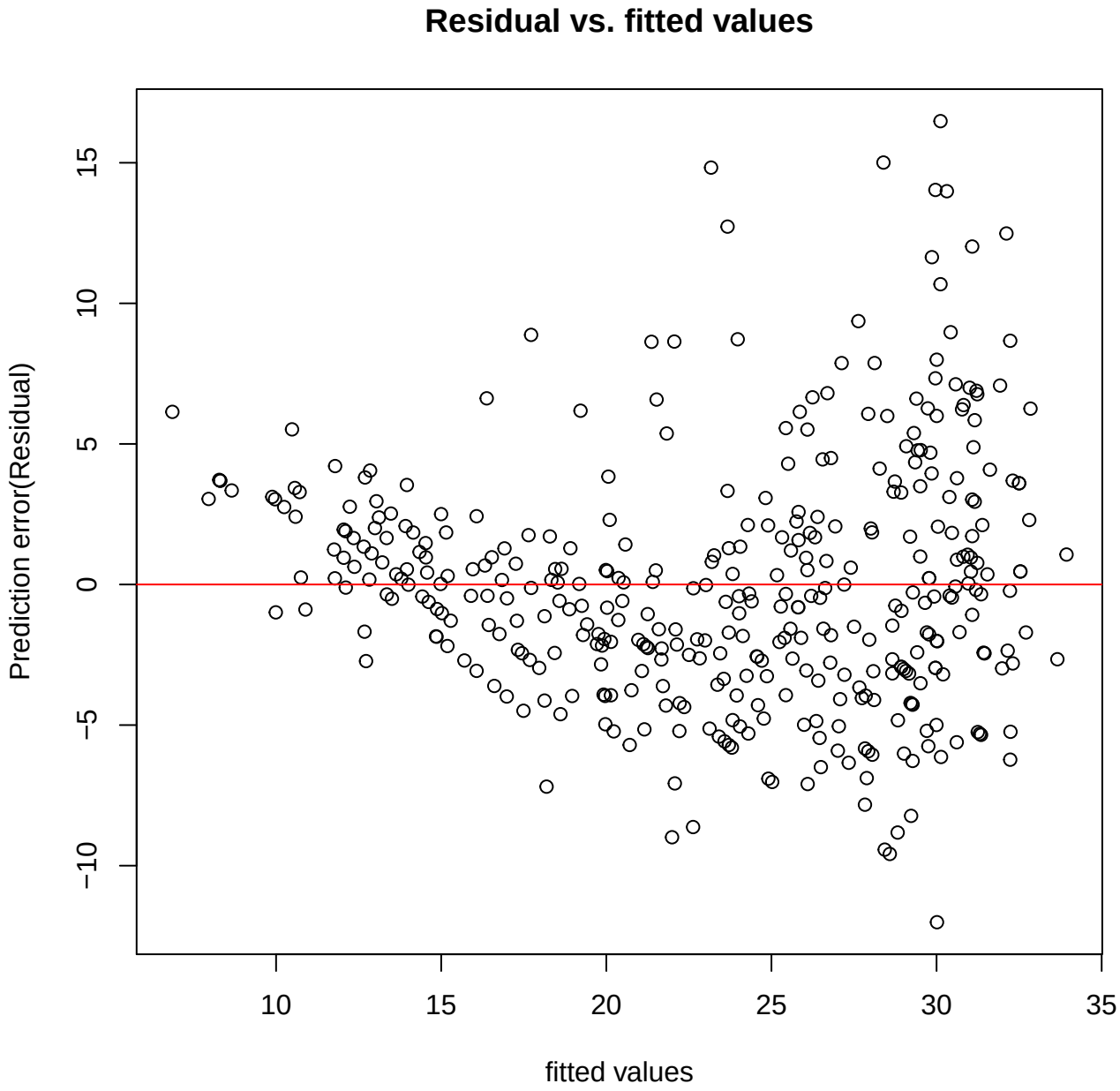
Now, we will study the relationship between mpg and weight (in lbs).

❖ Diagnostics:

We will check whether a linear regression model would be a good fit or not.



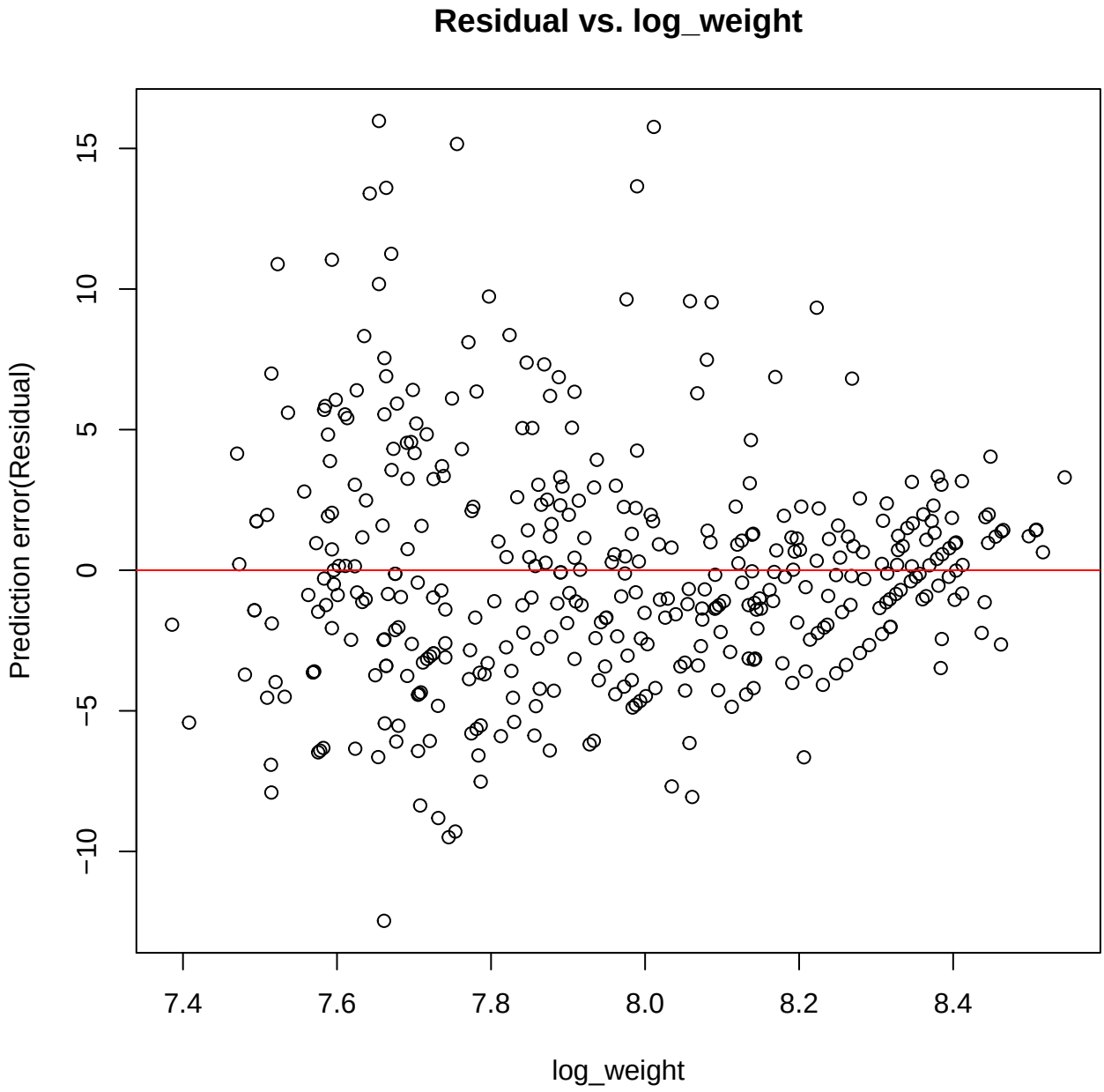
Clearly, the residual vs covariate plot doesn't look like a constant-width blur of points around a straight, flat line at height zero, rather the width of the cluster of points is decreasing gradually. That means the linearity assumption of the regression model is wrong.



We, can use this plot to answer the previous question. The interpretation of a "residual vs. covariate plot" is identical to that for a "residuals vs. fitted values plot". The reason is that, for the

❖ Transforming the predictor:

We apply the log transformation on weight and refit the regression model.

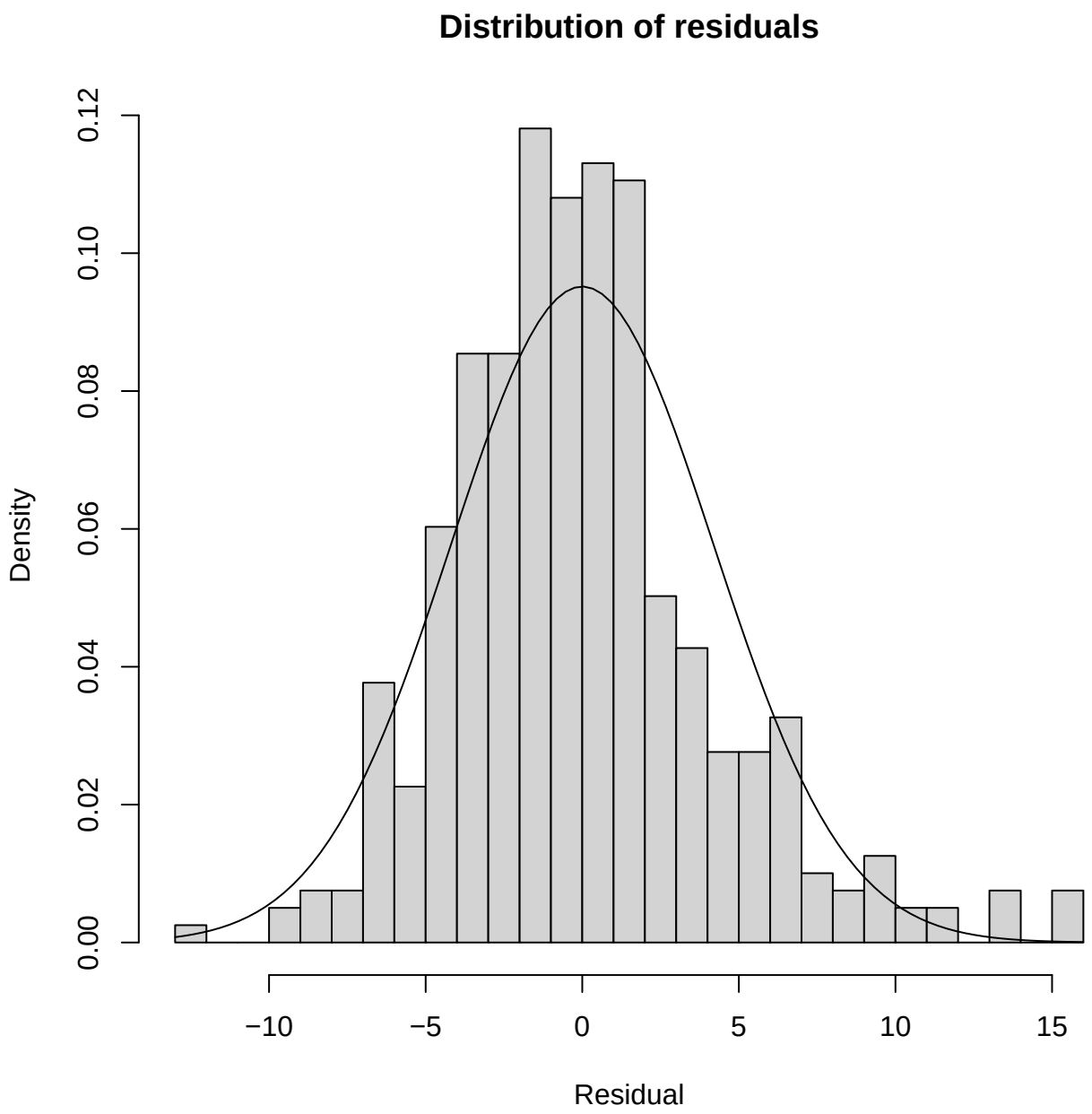


- Here also the residual vs covariate plot doesn't look like a constant-width blur of points around a straight, flat line at height zero, rather the width of the cluster of points is decreasing gradually. That means the linearity assumption of the regression model is wrong.
- Also, the decreasing width of the scatter of points indicates non-constant noise variance (technically called "heteroscedasticity").

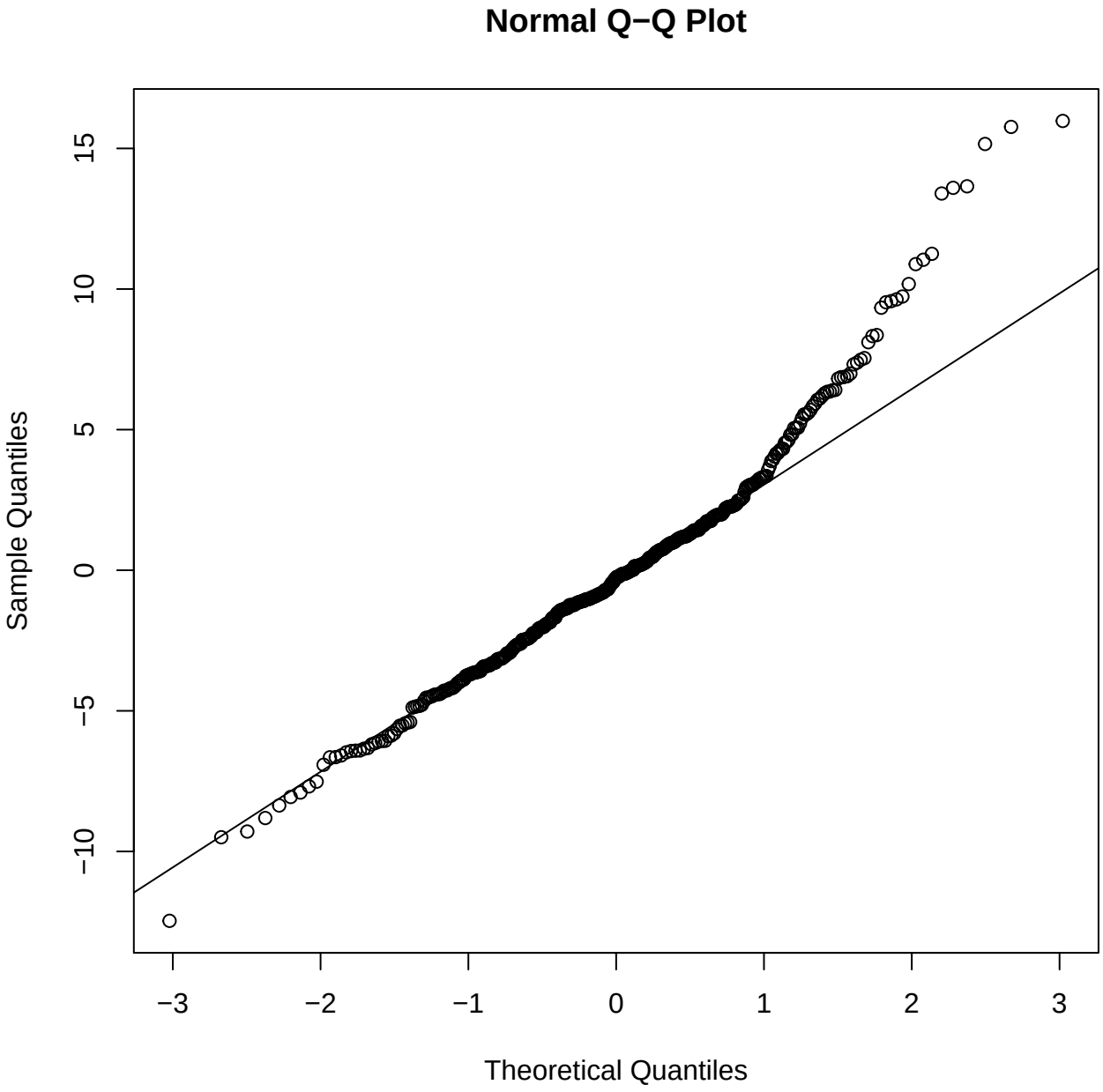
❖ Distribution of residuals:

To assess the normality of the residuals (for the model with transformed predictor) we should make plots of the distribution of the residuals, and compare that to a Gaussian.

We plot a histogram of the residuals over-laid with a Gaussian probability density of mean 0 (because we know the residuals average to 0), and the same standard deviation as the residuals (because that's the MLE of the standard deviation in a Gaussian model).



Also, we prepare a Q-Q plot of the residuals as a normal probability plot. The sample quantiles are plotted against the theoritical quantiles. The points should form an approximate straight line to satisfy the normality assumption of the residuals.



Clearly, both the plots suggest the violation of the normality assumption of the residuals.

❖ Box-cox Transformation of the response:

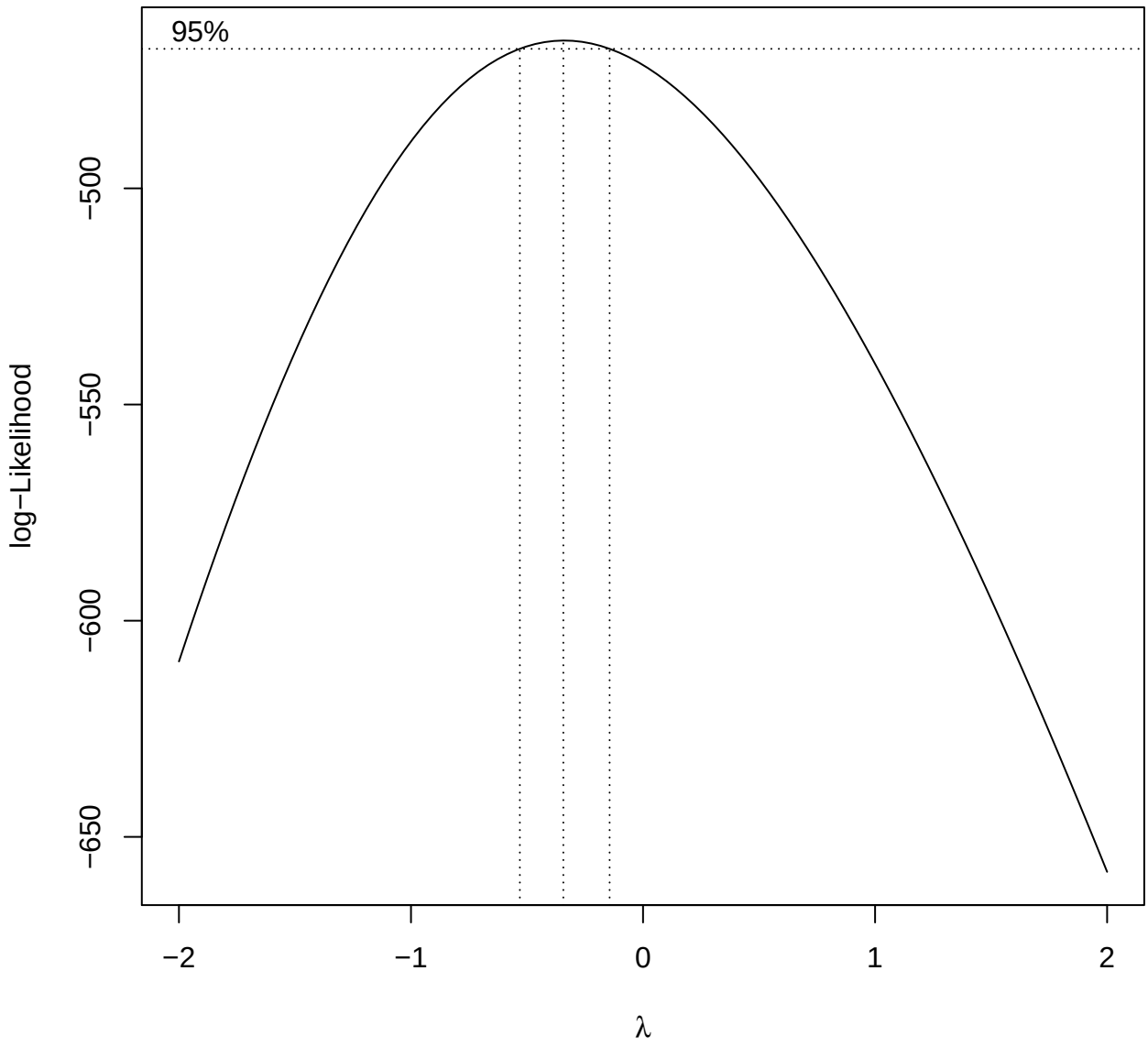
Box-cox tansformation:

$$b_\lambda(y) = \frac{y^\lambda - 1}{\lambda}$$

this becomes log y, as $\lambda \to \infty$
Here, our model is,

$$b_\lambda(Y) = \beta_0 + \beta_1 x + \epsilon \text{ ,where } \epsilon \sim N(0, \sigma^2) \text{ and } \epsilon \text{ is independent of x.}$$

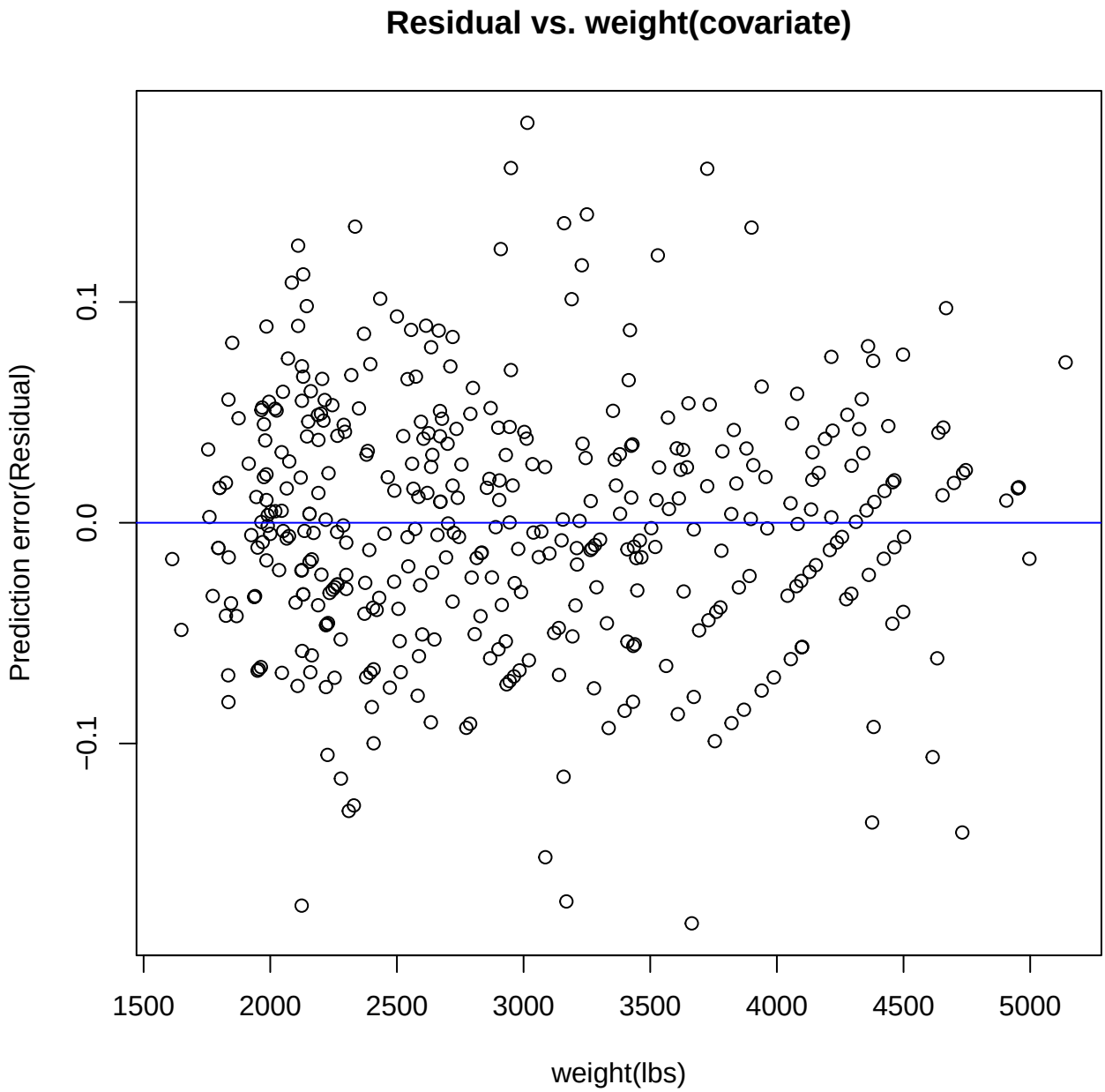
we can estimate λ by maximizing the likelihood.
Estimation of λ:




```
## [1] -0.3434343
```

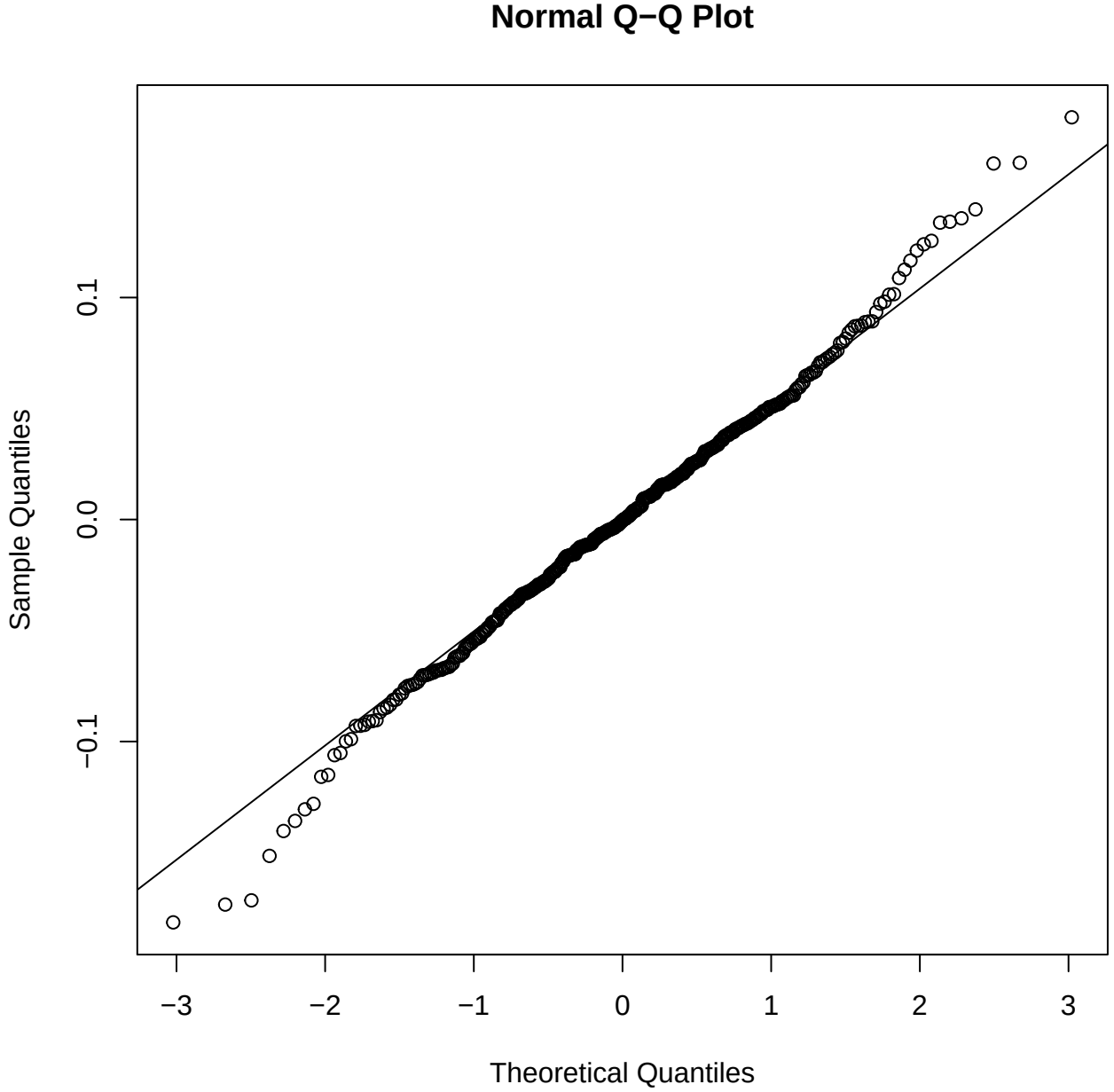
Therefore, the best power transformation for the response variable is for $\hat{\lambda} = -0.3434343$.

- **New Residual plot**:



Clearly, the residual vs covariate plot seems better to look like a constant-width blur of points around a straight, flat line at height zero. That means the linearity assumption of this box-cox transformed model is better satisfied now.

- **New Normal Probability plot**:



Clearly, the normal probability plot suggests better assumption of the normality of the residuals.
Parameter interpretation:

```
##   (Intercept)        weight
##  2.2693266866 -0.0001239117
```

- **Interpretation of the estimated intercept ($\hat{\beta}_0$)**: $\hat{\beta}_0$(=2.269) is the expected value of $b_\lambda(Y)$ when the weight of the car is of 0 unit. But here this interpretation is not reasonable as '0' weight implies no car.
- **Interpretation of the estimated slope ($\hat{\beta}_1$)**: If we select two sets of cases from the un-manipulated distribution where the weight(x) differs by 1, we expect $b_\lambda(Y)$ to differ by $\hat{\beta}_1$(=-0.0001239) accordingly.

Test of linearity:

We have to test whether there is a linear association between the 'transformed mpg' and 'weight', at 5% level of significance.

Now, we know correlation measures the linear association between two variables. So, here we use the cor.test() function for testing.

Alternative Hypothesis (H_1): There is a negative correlation between the 'transformed mpg' and 'weight'

Decision Rule: $p \leq \alpha \implies$ Null hypothesis is rejected and go in favour of H_1

and $p > \alpha \implies H_0$ can't be rejected at level of significance α

```
##
##  Pearson's product-moment correlation
##
## data:  cars.data$weight and cars.data$bc.mpg
## t = -37.449, df = 396, p-value < 2.2e-16
## alternative hypothesis: true correlation is less than 0
## 95 percent confidence interval:
##  -1.0000000 -0.8634569
## sample estimates:
##        cor
## -0.8830687
```

Here, the p-value of the test is 2.2×10^{-16} which is less than $\alpha = 0.05$

Hence, we can conclude that there is a negative correlation between the 'transformed mpg' and 'weight' at 5% level of significance.

90% confidence interval of the slope($\hat{\beta}_1$):

```
##                   5 %          95 %
## weight -0.0001293669 -0.0001184565
```

Interpretation: The interval [-0.0001293669, -0.0001184565] will include the true value of $\hat{\beta}_1$, whatever it is, with 90% confidence.

❑ Analysis of the Abalone dataset:

Abalones, also called ear-shells or sea ears, are one type of reef-dwelling marine snails. The flesh of abalones is widely considered to be a desirable food, and is consumed raw or cooked in a variety of cultures. It is difficult to tell the ages of abalones because their shell sizes not only depend on how old they are, but also depend on the availability of food. The study of age is usually by obtaining a stained sample of the shell and looking at the number of rings through a microscope. A research group are interested in using some of abalones' physical measurements, especially the height measurement to predict their ages. The research group believe that a simple linear regression model with normal error assumption is appropriate to describe the relationship between the height of abalones and their ages, and particularly, that a larger height is associated with an older age.

❖ Research Problem:

Predicting the age of abalone from physical measurements, especially the height measurement.

The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope -- a boring and time-consuming task. So, other measurements, which are easier to obtain, are used to predict the age (e.g. height measurement).

(Source: Data Set Information from https://archive.ics.uci.edu/ml/datasets/Abalone)

Research Hypothesis:

It is hypothesized that the relationship between the height of abalones and their ages is linear, and particularly, that a larger height is associated with an older age.

❖ Detailed analysis of the two variables:

- **Summary Measures of the height variable**:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.1150  0.1400  0.1395  0.1650  1.1300
## [1] 0.001749503
```

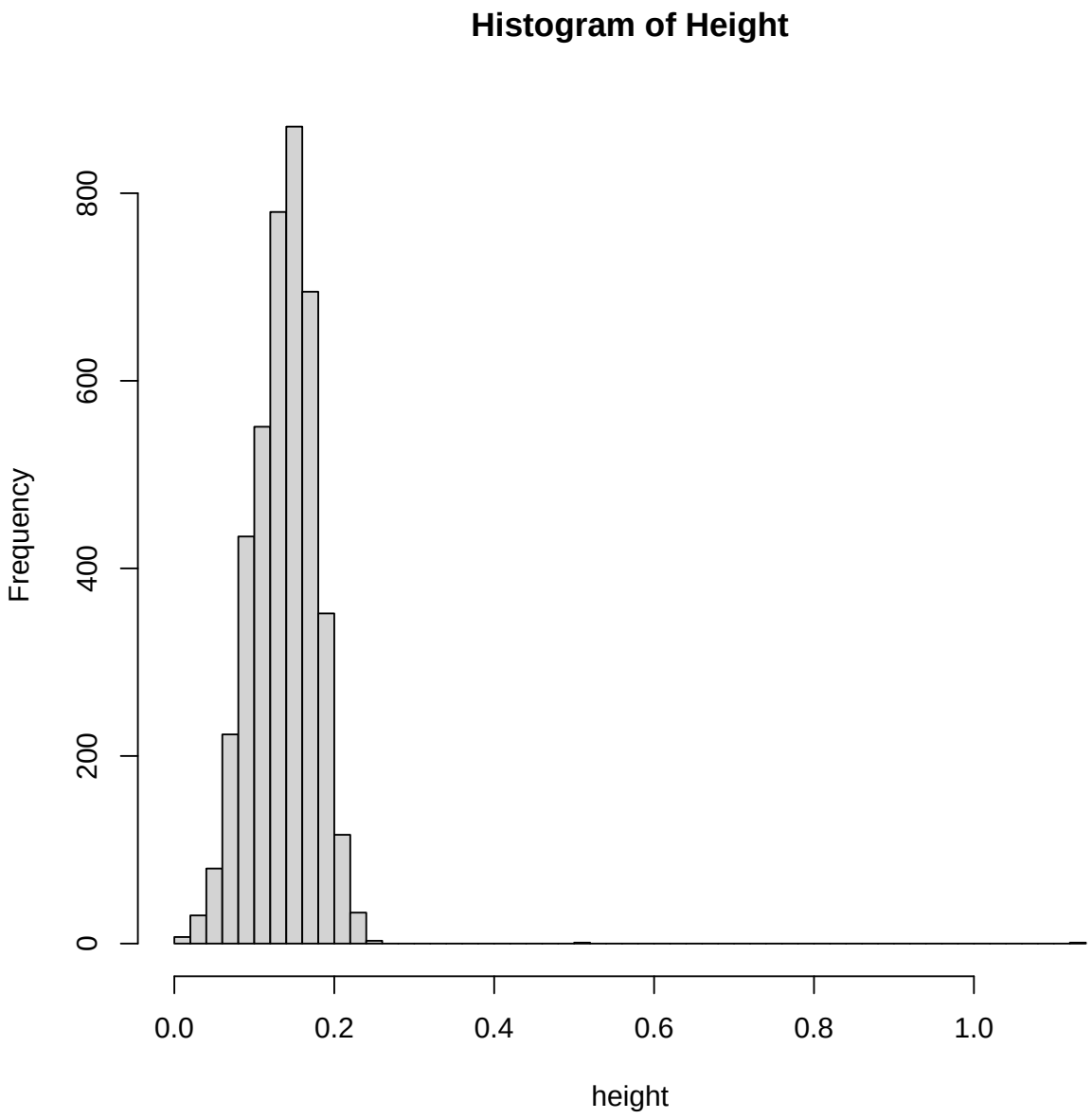
The unit of height is millimeter(mm). (Source: Attribute Information from https://archive.ics.uci.edu/ml/datasets/Abalone

- **Summary Measures of the rings variable**:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   8.000   9.000   9.934  11.000  29.000
## [1] 10.39527
```

Clearly, the ring count is a discrete integer valued variable.

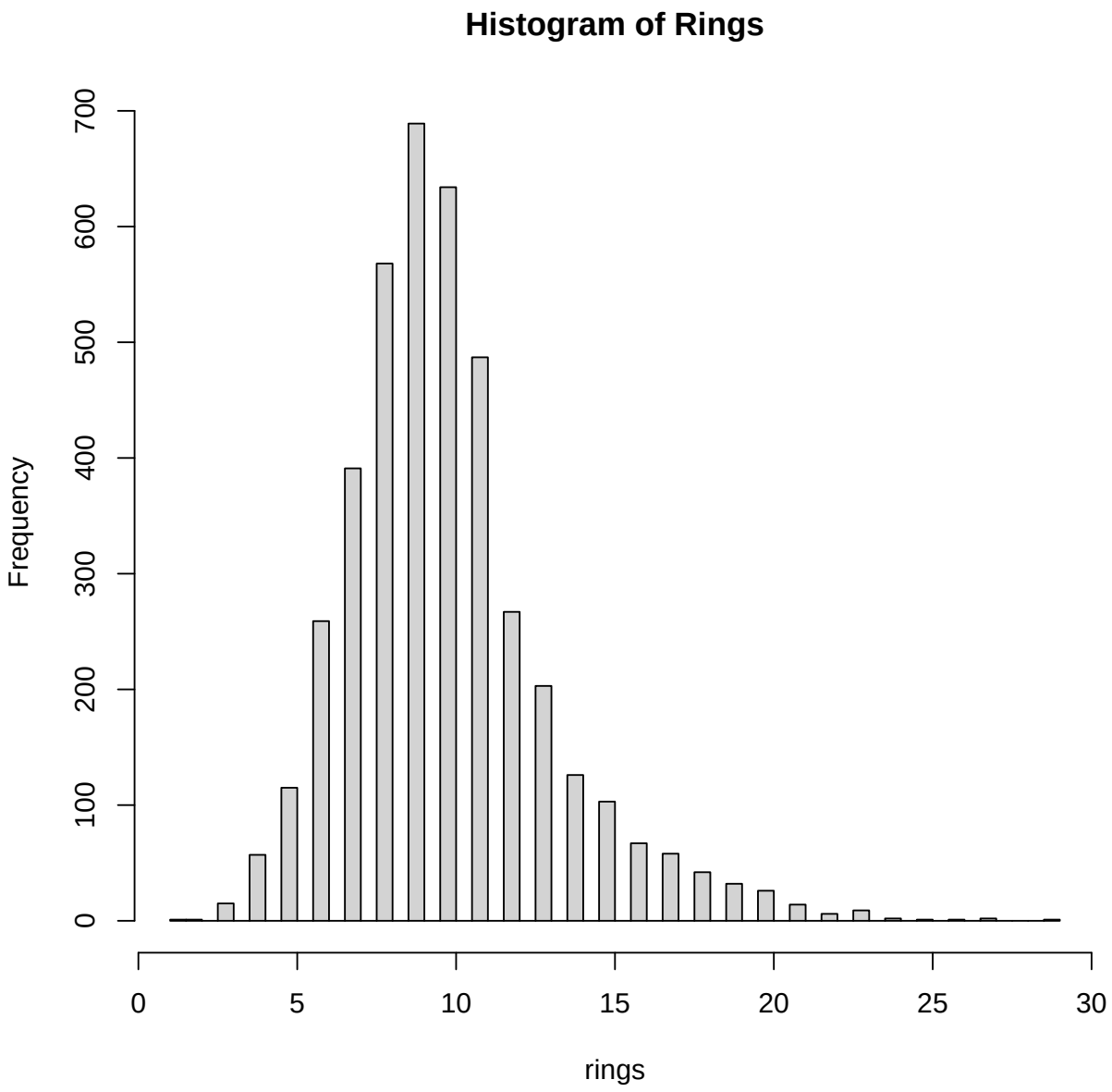
- **Graphical representation of the height variable**:




```
## [1] 0.515 1.130
```

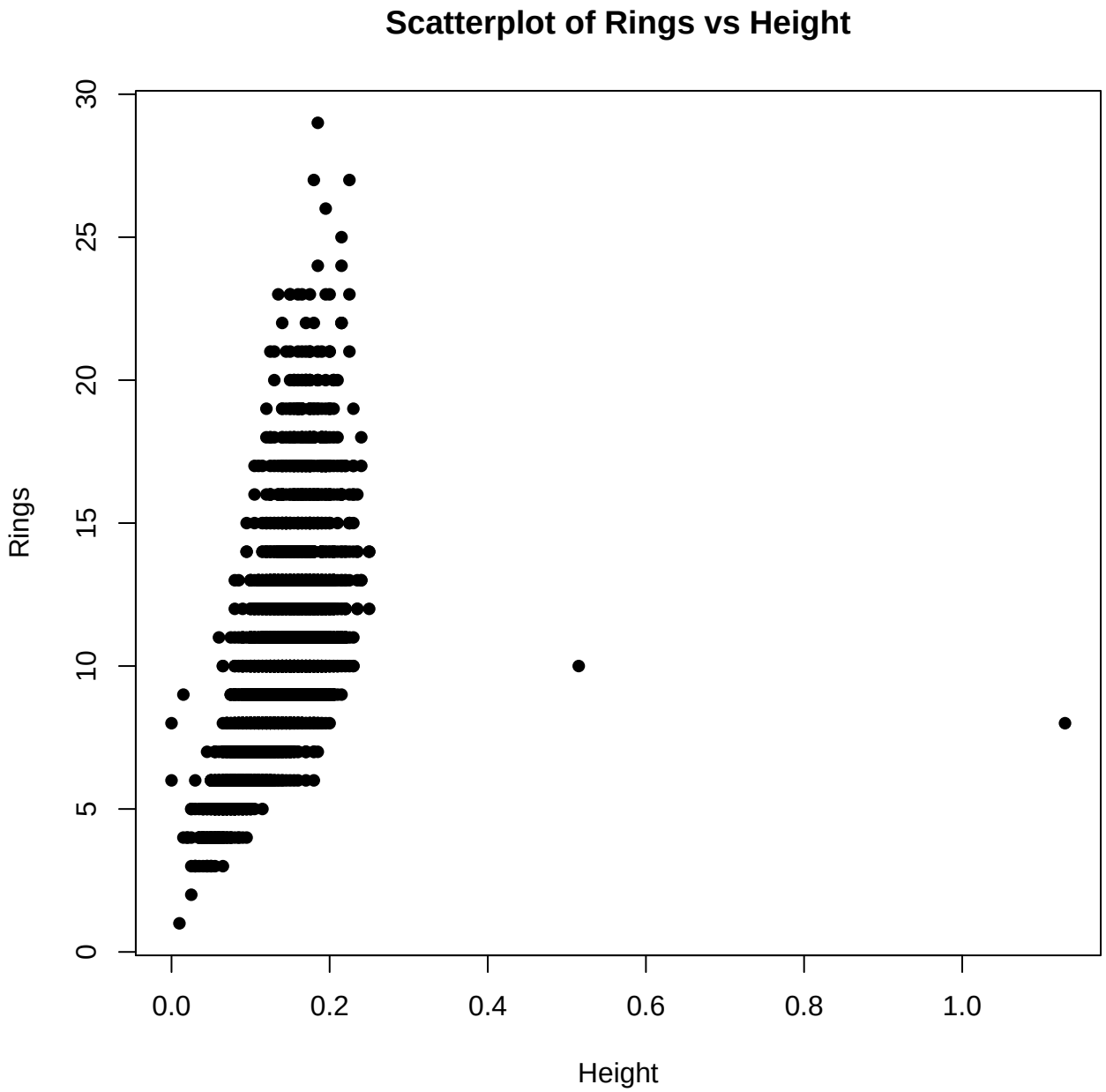
Clearly, for majority of the observed abalones the height range is [0,0.25mm.], with the mode (highest frequency of Height) nearly 0.17 mm, except two outliers whose heights are 0.515mm. and 1.13 mm. respectively.

- **Graphical representation of the rings variable**:



The ring count of the majority of the observed abalones ranges between 1 to 29, with the mode (highest frequency of ring count) of nine.

❖ **Scatterplot of the data:**



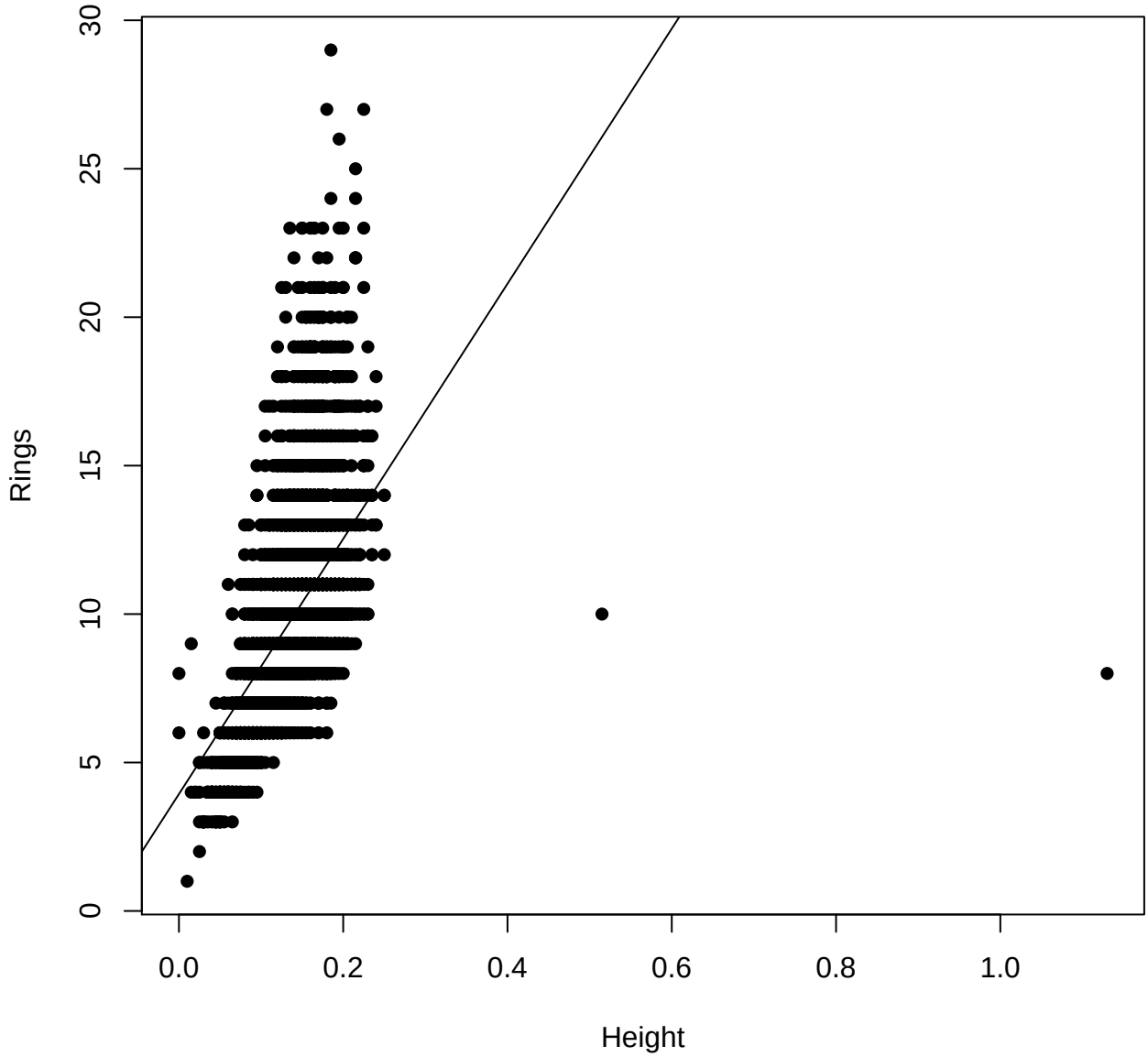
Clearly, the two exceptionally large heights are outliers. Except the outliers, as Height increases the number of Rings also inceases in a rapid manner (exponentially). So, even we restrict our attention to the non-extreme cases, a linear model won't be suitable in this case.

❖ Fitting simple linear regression to the data:

Fitting a simple linear regression to the data predicting number of rings (Y) using height of the abalones (x).

```
##
## Call:
## lm(formula = Rings ~ Height, data = abalone)
```

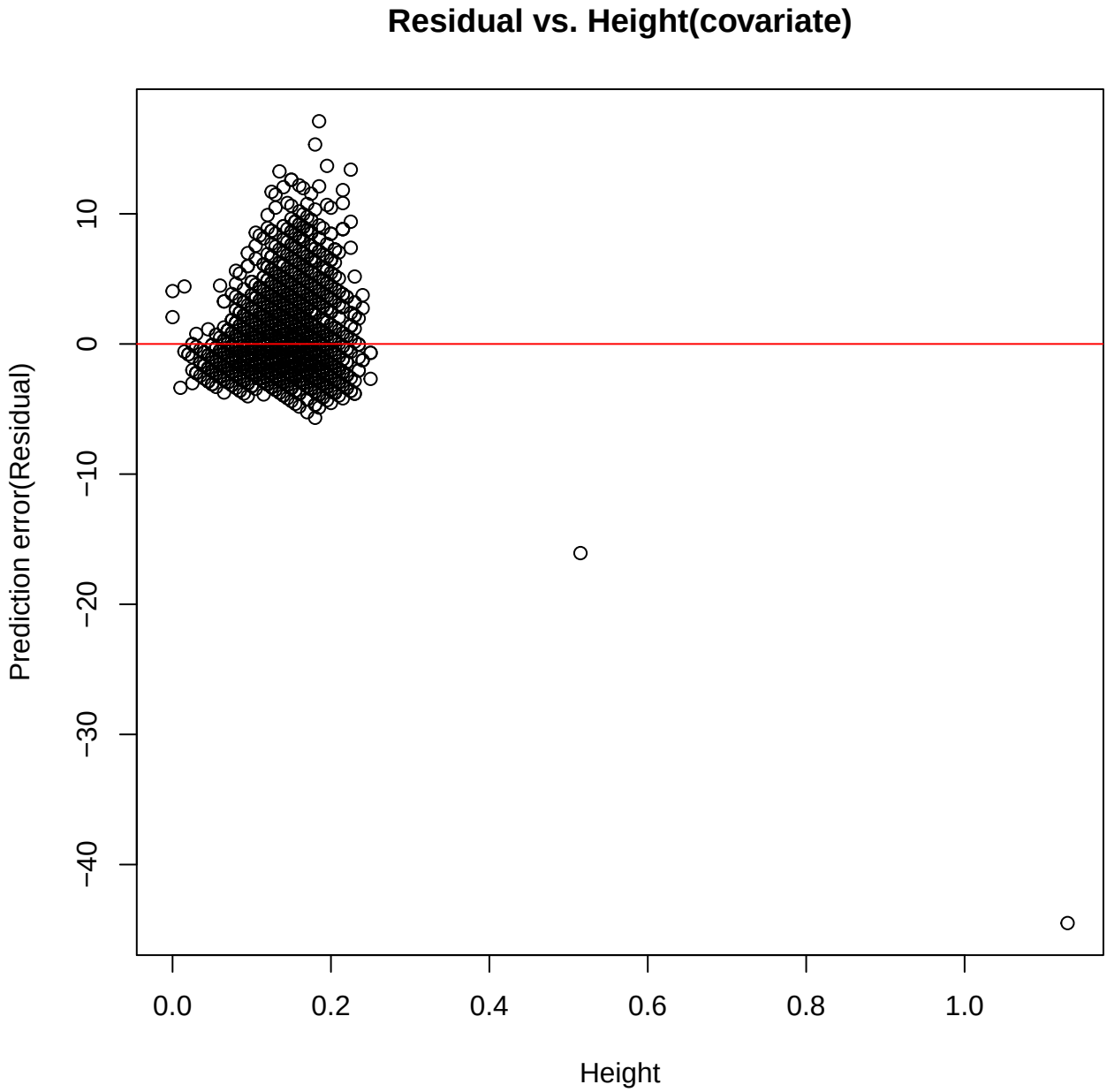
```
##
## Coefficients:
## (Intercept)       Height
##       3.938       42.971
```



It seems that presence of the outliers significantly affected the fit. Furthermore, even we restrict our attention to the non-extreme cases, it seems that a linear model won't be suitable in this case for fitting.

❖ Diagnostics:

We will check whether a simple linear regression model would be a good fit or not.



Clearly, if we don't consider the two influential outliers though, the residual vs covariate plot doesn't look like a constant-width blur of points around a straight, flat line at height zero. That means the linearity assumption of the regression model is wrong.

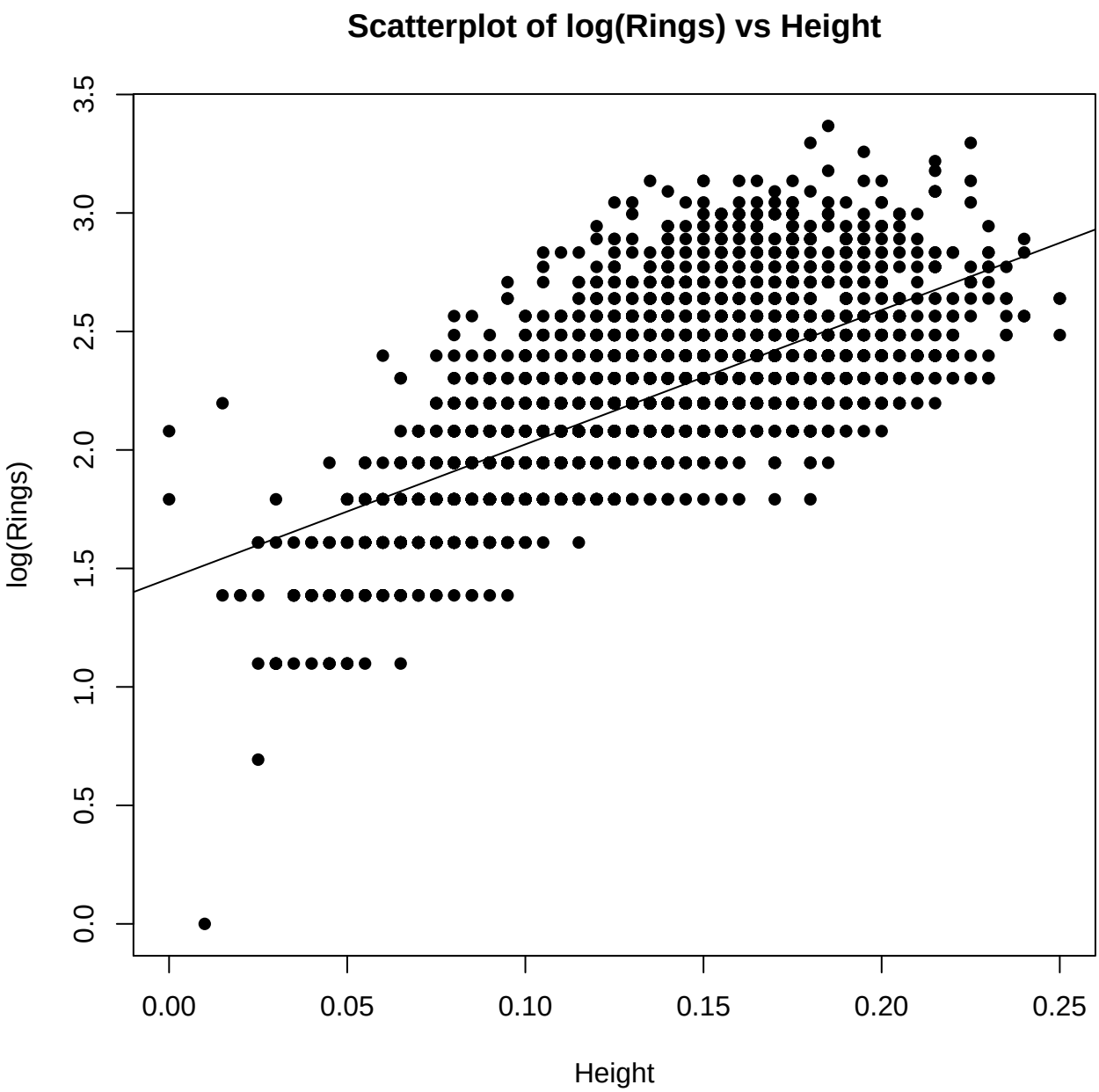
❖ Transforming the response:

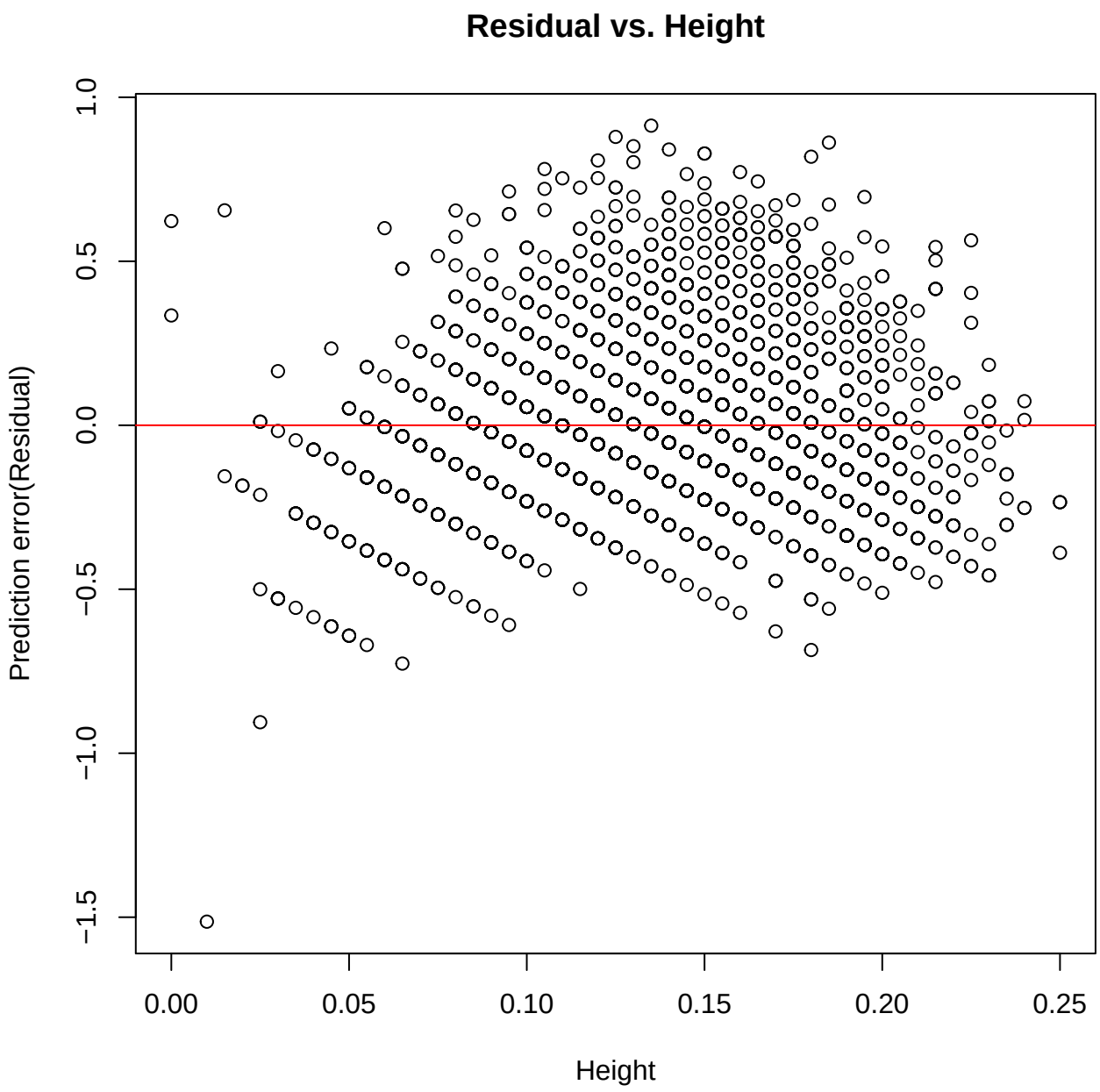
At first we remove the two influential outliers from the data.

We apply log transformation on the number of rings (response variable) and refit the regression model as,

$$logY = \alpha + \beta x + \epsilon$$

where, $\epsilon \sim N\left(0, \sigma^2\right)$, independent of x.


Scatterplot of log(Rings) vs Height
log(Rings)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0.00
0.05
0.10
0.15
0.20
0.25
Height



Clearly, here the residual vs covariate plot seems better to look like a constant-width blur of points around a straight, flat line at height zero. That means the linearity assumption of this log transformed model is better satisfied now.

Interpretation of the Parameters:

The estimated value of the parameters are:

```
## (Intercept)      Height
##    1.456909    5.667076
```

- **Interpretation of the intercept**(α):

The expected value of $log\,(Y)$ is 1.456909, when the height of the abalone is 0 mm.

i.e. the expected number of rings on a abalone is almost $4(\approx e^{1.456909})$ when the height of the abalone is 0 mm. This interpretation is meaningless. So, in this case we should fit a model without the intercept term.

- **Interpretation of the slope**(β):

If we select two sets of cases from the un-manipulated distribution where Height differs by 1 mm., we expect the value of $log\,(Y)$ to differ by 5.667076.

95% Confidence Intervals for α, β:

```
##                2.5 %   97.5 %
## (Intercept) 1.430312 1.483506
## Height      5.482899 5.851252
```

- **Interpretation**:

The interval [1.430312,1.483506] will include the true value of the parameter α with 95% confidence.

However, in this context this interpretation is meaningless as, adding the intercept term in this model makes no sense.

The interval [5.482899, 5.851252] will include the true value of the parameter β with probability 0.95 (i.e. very high probability).

Test of the relationship between the height and the number of rings of abalones:

We have to test whether there is a statistically significant relationship between the 'transformed number of rings of abalones' and 'height', at 5% level of significance.

Now, we know correlation measures the linear association between two variables. So, here we use the cor.test() function for testing.

Alternative Hypothesis (H_1): There is a non-zero correlation between the two variables i.e. $\rho \neq 0$

Decision Rule: $p \leq \alpha \implies$ Null hypothesis is rejected and go in favour of H_1

and $p > \alpha \implies H_0$ can't be rejected at level of significance α where p is the p-value of the test.

```
##
##  Pearson's product-moment correlation
##
## data:  trans.abalone.new$Height and trans.abalone.new$ln_ring
## t = 60.325, df = 4173, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6659711 0.6983929
## sample estimates:
##       cor
## 0.6825176
```

Here, the p-value of the test is 2.2×10^{-16} which is less than $\alpha = 0.05$

Hence, we can conclude that there is a non-zero correlation between the 'transformed number of rings of abalones (and hence, the age)' and 'height' at 5% level of significance.

Therefore, there is a statistically significant linear relationship between the two variables.

Note:

- Hence, the point estimate of the average number of rings for abalones with height at 0.128 mm. is 9 (~8.86663).

- Therefore, a 95% confidence interval for the average of the logarithm of number of rings for abalones with height at 0.128 mm. is (2.174943, 2.189647)

Interpretation: The interval (2.174943, 2.189647) contains the true value of the average of the logarithm of number of rings (i.e. population average, which is impossible to get in real life) for abalones of height at 0.128 mm. with probability 0.95 (i.e. very high probability).

- The predicted number of rings for an abalone with height at 0.132 mm. is 9 (~9.069917).

- The 99% prediction interval for the logarithm of number of rings of an abalone with height at 0.132 mm. is (1.602896, 2.80703) i.e. the 99% prediction interval for the number of rings of an abalones with height at 0.132 mm. is $(e^{1.602896}, e^{2.80703})$ ≈(4.967397, 16.560660).

Key finding:

The linear relationship between the height of the abalones and the logarithm of number of rings of abalones (and hence, the age) is statistically significant. Potential reason for this is that there is a biological process which affects or relate height with age.

Suggestions and recommendations for the researchers:

1. Study for potential variables which may be affecting both height and number of rings.
2. Remove the outliers before analysis.
3. Take log-normal error assumption rather than gaussian error assumption.
4. Do not include the intercept in the linear model. This is because the intercept is the expected age of abalones with non-existant height, which is practically impossible.

❑ Appendix:

(R codes)

1. Analysis of the auto-mpg dataset:

```
cars.data=read.table("D:\\Econometrics_AKG Sir\\Assignments\\auto-mpg.csv",header=T,sep=",")
attach(cars.data)
```

   (a) (Diagnostics and Transformations)

      i. linear regression model:

         A.
```
model<-lm(mpg~weight,data=cars.data)
plot(weight,residuals(model),xlab="weight(lbs)",ylab="Prediction error(Residual)",main="Res
abline(h=0,col="red")
```

         B.
```
plot(fitted(model),residuals(model),xlab="fitted values",ylab="Prediction error(Residual)",
abline(h=0,col="red")
```

ii.
```
x=log(weight)
ncars.data=data.frame(cars.data,ln_wgt=x)
cars.lm<-lm(mpg~ln_wgt,data=ncars.data)
plot(ncars.data$ln_wgt,residuals(cars.lm),xlab="log_weight",ylab="Prediction error(Residual)",
abline(h=0,col="red")
```

iii.
```
hist(residuals(cars.lm),breaks=40,prob=T,xlab="Residual",main="Distribution of residuals")
y=min(residuals(cars.lm))
z=max(residuals(cars.lm))
x=seq(from=y,to=z,by=0.1)
curve(dnorm(x,mean=0,sd=sd(residuals(cars.lm))),add=T)
qqnorm(residuals(cars.lm))
qqline(residuals(cars.lm))
```

iv. Box-cox procedure:

```
library(MASS)
boxcox(model)
bc.cars<-boxcox(mpg~weight,data=cars.data,plotit=T)
lambda.hat<-bc.cars$x[which.max(bc.cars$y)]
lambda.hat
```

```
cars.data$bc.mpg<-((mpg^lambda.hat)-1)/lambda.hat
bc.lm<-lm(bc.mpg~weight,data=cars.data)
coefficients(bc.lm)
plot(weight,residuals(bc.lm),xlab="weight(lbs)",ylab="Prediction error(Residual)",main="Residu
abline(h=0,col="blue")

qqnorm(residuals(bc.lm))
qqline(residuals(bc.lm))
```

(b) Based on the previous model:

i. interpretation.

ii.
```
cor.test(cars.data$weight,cars.data$bc.mpg,alternative="less",conf.level=0.95)
```

iii.
```
confint(bc.lm,"weight",level=0.90)
```

2. Analysis of the Abalone dataset:

```
abalone<-read.table("D:\\Econometrics_AKG Sir\\Assignments\\abalone.csv",header=T,sep=",")
attach(abalone)
```

(a) Research problem and hypothesis.

(b)
```
summary(Height)
var(Height)
```

```
summary(Rings)
var(Rings)
```

```
hist(Height,xlab="height",breaks=50)
Height[Height>0.25]
```

```
hist(Rings,xlab="rings",breaks=50)
```

(c)
```
plot(Height,Rings,main="Scatterplot of Rings vs Height",pch=16)
```

(d)
```
ab.lm=lm(Rings~Height,data=abalone)
ab.lm
```

(e)
```
plot(Height,Rings,pch=16)
abline(ab.lm)
```

(f)
```
plot(Height,residuals(ab.lm),xlab="Height",ylab="Prediction error(Residual)",main="Residual vs. He
abline(h=0,col="red")
```

```
abalone.new=abalone[-c(which(abalone$Height==0.515),which(abalone$Height==1.130)),]
x=log(abalone.new$Rings)
trans.abalone.new=data.frame(abalone.new,ln_ring=x)
nab.lm<-lm(ln_ring~Height,data=trans.abalone.new)
plot(trans.abalone.new$Height,trans.abalone.new$ln_ring,xlab="Height",ylab="log(Rings)",main="Scat
abline(nab.lm)
plot(trans.abalone.new$Height,residuals(nab.lm),xlab="Height",ylab="Prediction error(Residual)",ma
abline(h=0,col="red")
```

(g)
```
coefficients(nab.lm)
```

```
confint(nab.lm)
```

(h)
```
cor.test(trans.abalone.new$Height,trans.abalone.new$ln_ring,alternative="two.sided",conf.level=0.9
```

(i)
```
#...conditional mean of rings for abalones with height at 0.128
y=coefficients(nab.lm)[1]+coefficients(nab.lm)[2]*0.128
condmean.est=exp(y)
```

```
n=dim(abalone.new)[1]
RSS=sum(residuals(nab.lm)^2)
S=sqrt(RSS/(n-2))
x.var=((n-1)/n)*var(Height)
SE.condmean=(S/sqrt(n))*sqrt(1+((0.128-mean(Height))^2)/x.var)
condmean.lb=y-(qt(0.025,df=n-2,lower.tail=F))*SE.condmean
condmean.ub=y+(qt(0.025,df=n-2,lower.tail=F))*SE.condmean
condmean.CI=data.frame(CI.lower=condmean.lb,CI.upper=condmean.ub)
row.names(condmean.CI)<-c("conditional_mean")
```

(j)
```
#...predicted number of rings for an abalone with height at 0.132
a=coefficients(nab.lm)[1]+coefficients(nab.lm)[2]*0.132
pred=exp(a)
pred
```

```
SE.pred=S*sqrt(1+(1/n)+((0.132-mean(Height))^2)/(n*x.var))
pred.lb=a-(qt(0.005,df=n-2,lower.tail=F))*SE.pred
pred.ub=a+(qt(0.005,df=n-2,lower.tail=F))*SE.pred
pred.CI=data.frame(CI.lower=pred.lb,CI.upper=pred.ub)
row.names(pred.CI)<-c("pred_interval")
pred.CI
```