report

XIAO YUE

12/6/2019

Introduction

- Regression, more generally,linear regression model, is widely used techiniques in statistics. It attempts to model the relationship between two variables by fitting a linear equation to observed data. In this report, we will use the dataset abalone from the [@University_of_California_,_Irvine_Machine_learning_Repository][^r_version], and learn how to construct the linear model of abalone age from height(in mm.) in R.

```
library(readr)
abalone <- read_csv("abalone.csv")
```

```
## Parsed with column specification:
## cols(
##   Height = col_double(),
##   Rings = col_double()
## )
```

```
dim(abalone)
```

```
## [1] 4177    2
```

Explore data and Initial model

- In this dataset, in total there is 4177 abalones observed. And we firstly examine two variable seperately by creating summary table and histogram to find the distribution of each vairable. To be clear, in each summary table, it includes the minimum,maximum,median,mean and variance of each variable.From Figure 1, we can clearly see that the largest proportion of observed abalones fall in the height range [0,0.25mm],and the summary table for Height(mm) shows that there is also one maximum height which is not in the range measuring 1.13mm.Secondly for Rings, summary tables represents that the rings count from 1 to 9,and the whole distribution of Rings are shown in figure 2, the largest proportion of the rings are in around 10.

- For height data summary

```
library(knitr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
summary_height <- data.frame(abalone) %>% select(Height) %>% summarise_all(funs(mi
n, max,median,mean,var))
```

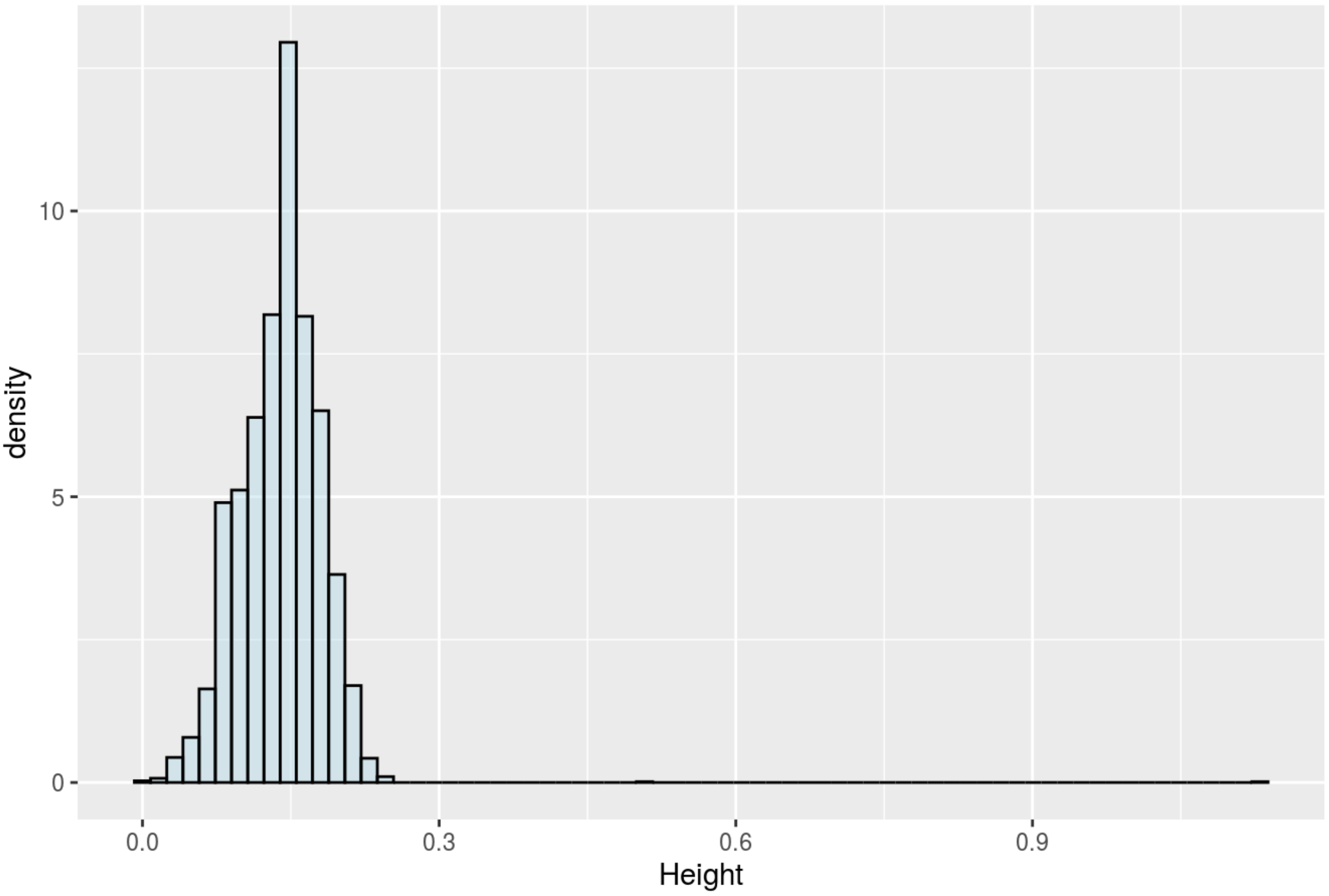
```
## Warning: funs() is soft deprecated as of dplyr 0.8.0
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once per session.
```

```
row.names(summary_height) <- "Height(mm)"
summary_height
```

```
##            min  max median      mean         var
## Height(mm)   0 1.13   0.14 0.1395164 0.001749503
```

```
plot_height <- ggplot(abalone,aes(x=Height)) +
    geom_histogram(bins = 70,aes(y=..density..),col="black",fill= "lightblue",alph
a=0.4)+
    #geom_density(size = 0.5,adjust = 0.4,col="Red")+
     ggtitle("Figure 1: Histogram for Abalone height")
plot_height
```

Figure 1: Histogram for Abalone height



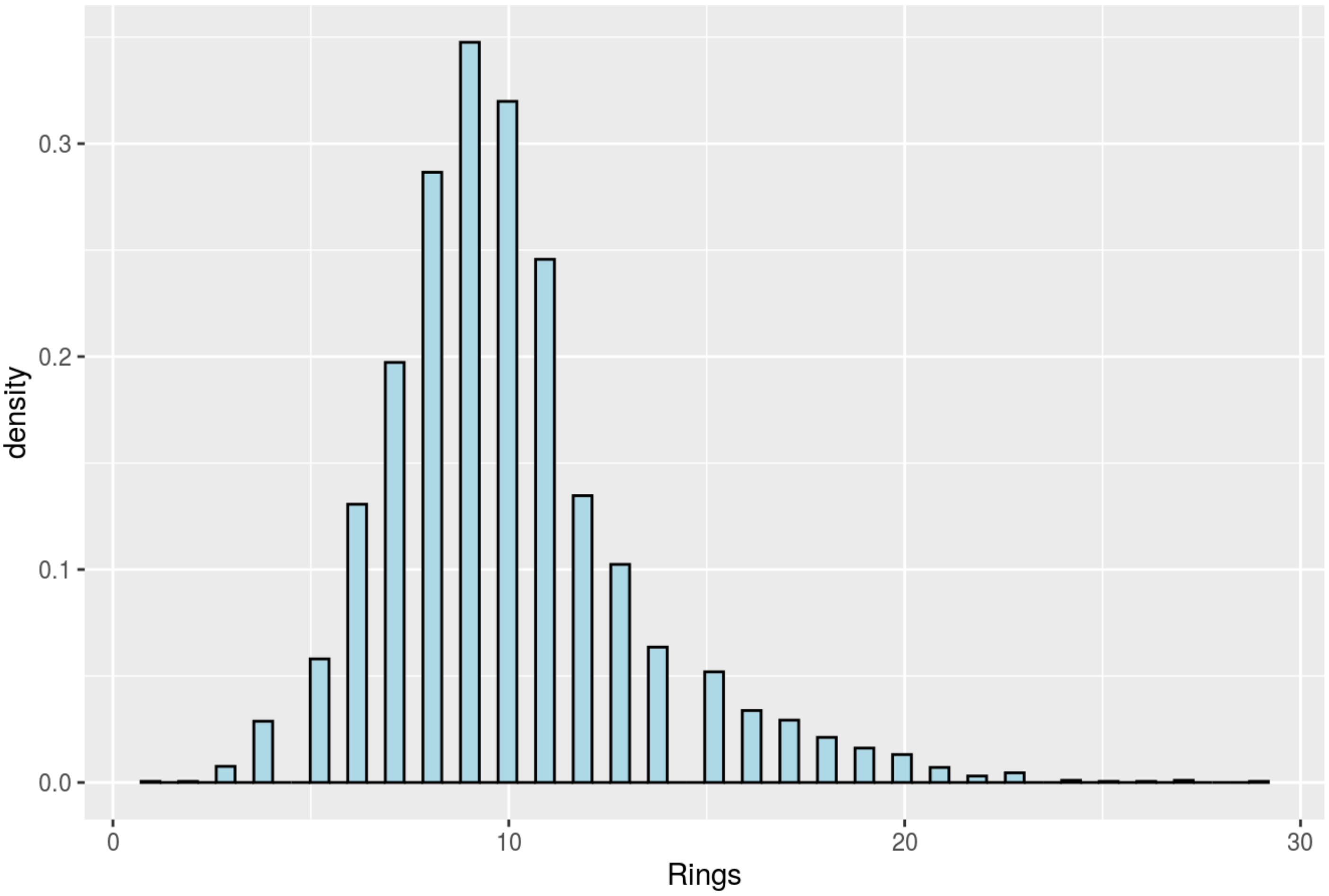
- For Rings summary

```
summary_rings <- data.frame(abalone) %>% select(Rings) %>% summarise_all(funs(min,
max,median,mean,var))
row.names(summary_rings) <- "Rings"
summary_rings
```

```
##       min max median     mean      var
## Rings   1  29      9 9.933684 10.39527
```

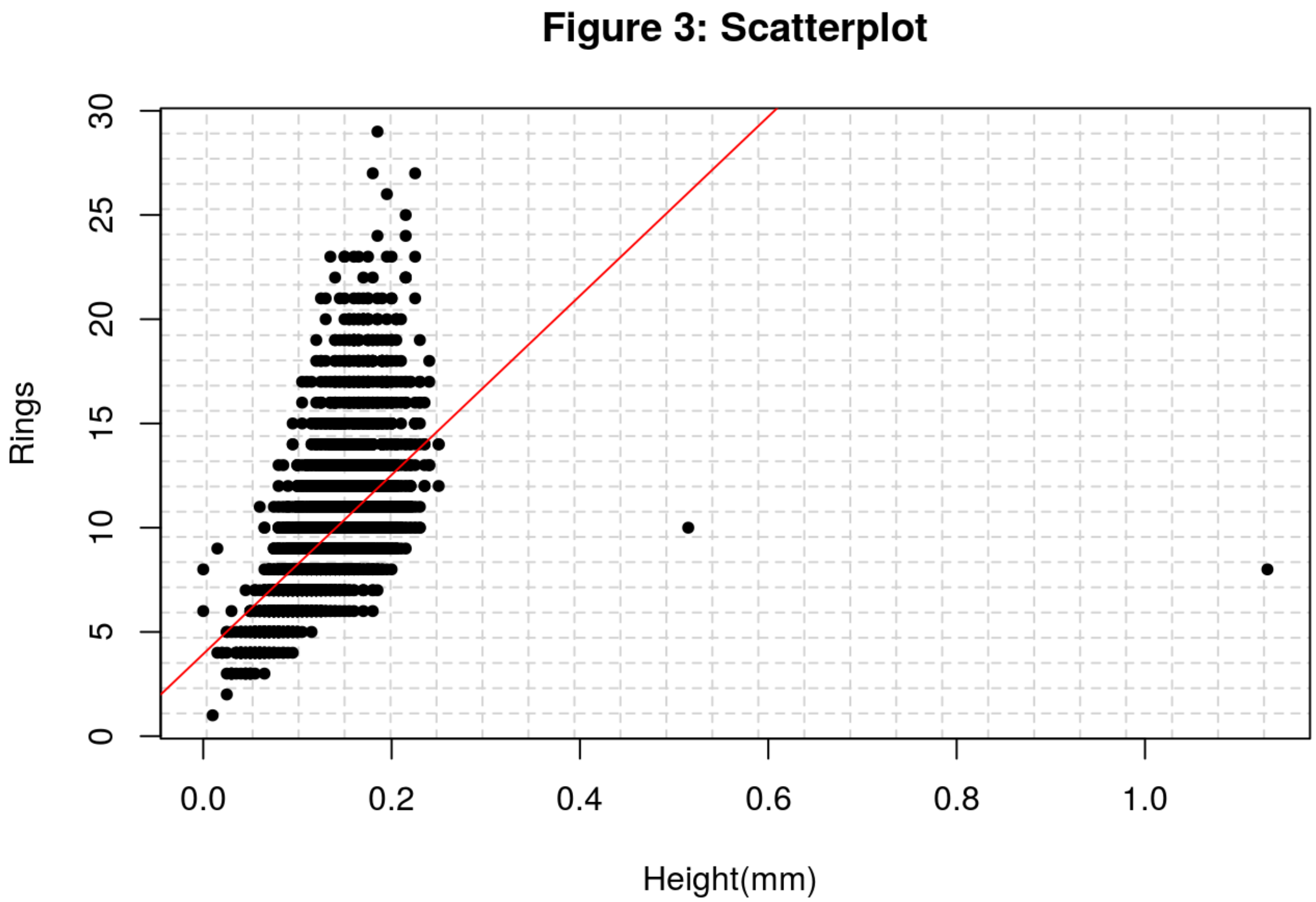
```
plot_rings <- ggplot(abalone,aes(x=Rings)) +
    geom_histogram(bins = 60,aes(y=..density..),col="black",fill= "lightblue")+
    ggtitle("Figure 2: Histogram for Abalone Rings")
plot_rings
```

Figure 2: Histogram for Abalone Rings



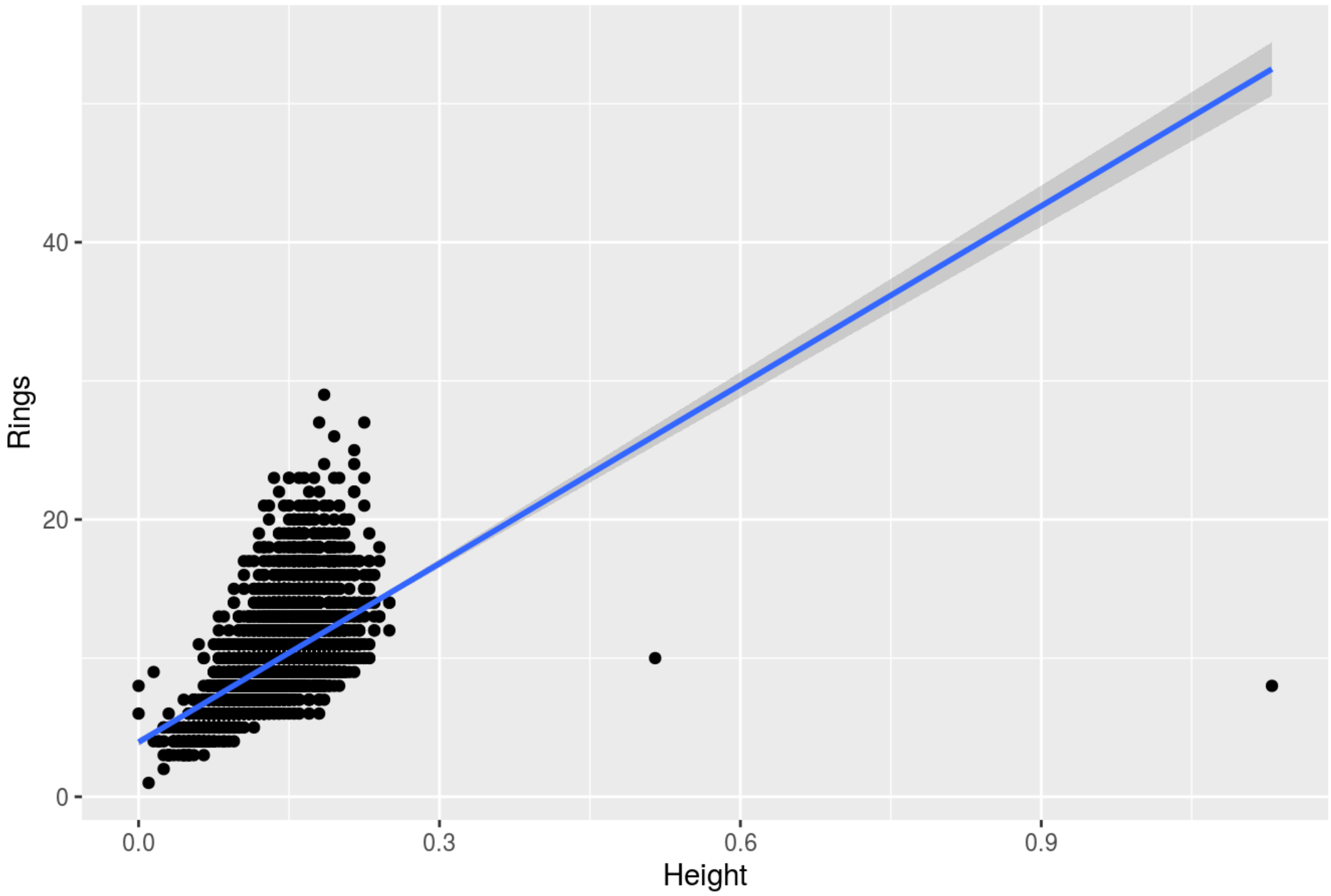
- Then we simply construct linear regression model to predict the number of rings using height of the abalones, which is shwon in Figure 2. In order to have a clearer trend of the line,figure 4 is represented by adding a geom_smooth layer to display a regression line with confidence intervals (95% CI by default).For the third one,we simply just plot the linear regresion model,and notice that the two extremely large heights seems to significantly influence the fit. And it appears that a linear model does not fit on this scale.

```
height <- abalone$Height
rings <- abalone$Rings
model <- lm(rings~height)
lm_plot <- plot(height,rings, main="Figure 3: Scatterplot",pch=20,
   xlab="Height(mm) ", ylab="Rings",panel.first = grid(25, lty = 2, lwd = 1))
abline(model, col="red",lwd=1)
```

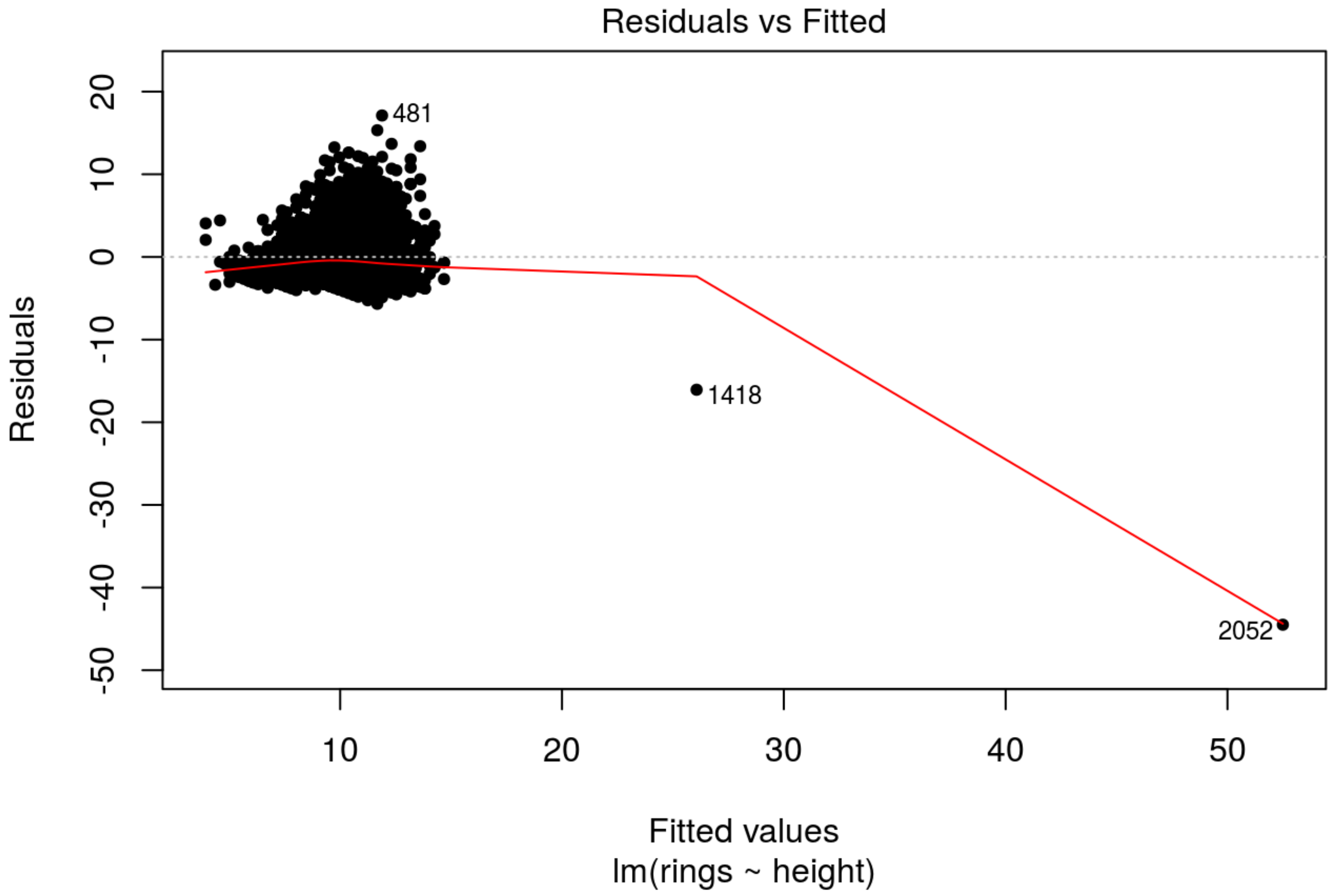


```
ggplot(abalone, aes(Height, Rings)) +
   geom_point()+
   geom_smooth(method = "lm",span=0.6)+
   ggtitle("Figure 4: Plot with smooth")
```

Figure 4: Plot with smooth



```
plot(model,which = 1,pch=20)
```

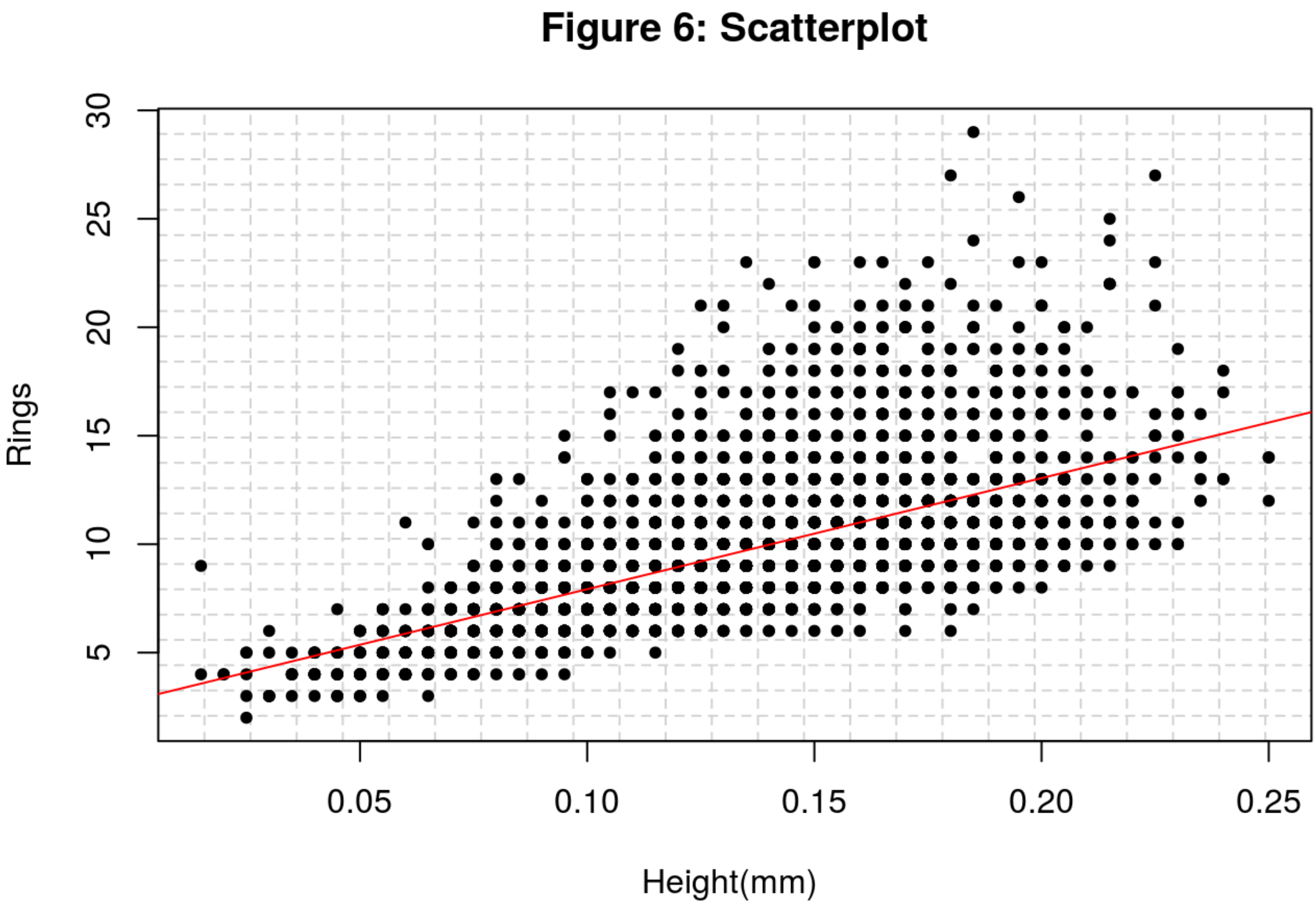


Modeling and diagnostics

- By noticing that there are two observations of abalones with height greater than 0.4 mm, which affect our model, we remove these observations and focus on a smaller set of abalone. We also discard those observations with height 0 and those height smaller than the minimum height since there counts on rings are really large. By doing this, we avoid some factors that will have bad influence on our results. In the end, we can conclude that our final form based on our chosen model will be $ = + *Height $ Height ∈ [0.025,0.25]

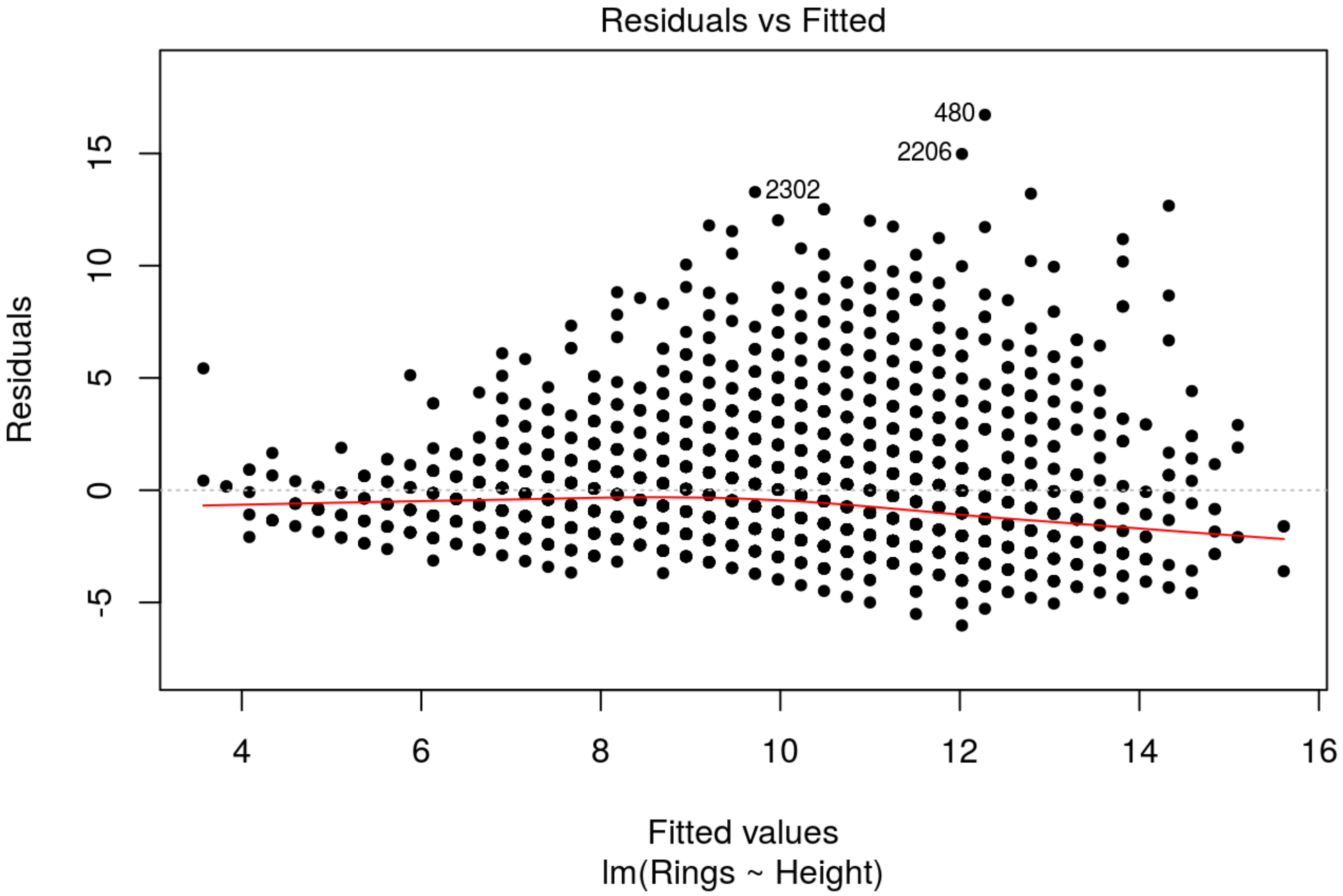
```
abalone_re <- abalone %>% filter(Height <0.4,Height>0.01395)
model_re <- lm(Rings~Height,data = abalone_re)
```

```
heights_1 <- abalone_re$Height
rings_1 <- abalone_re$Rings
plot(heights_1,rings_1, main="Figure 6: Scatterplot",pch=20,
   xlab="Height(mm) ", ylab="Rings",panel.first = grid(25, lty = 2, lwd = 1))
abline(model_re, col="red",lwd=1)
```



- We can see here that by modelling smaller set of abalone observations,we get a more linearer model than plots before represented.

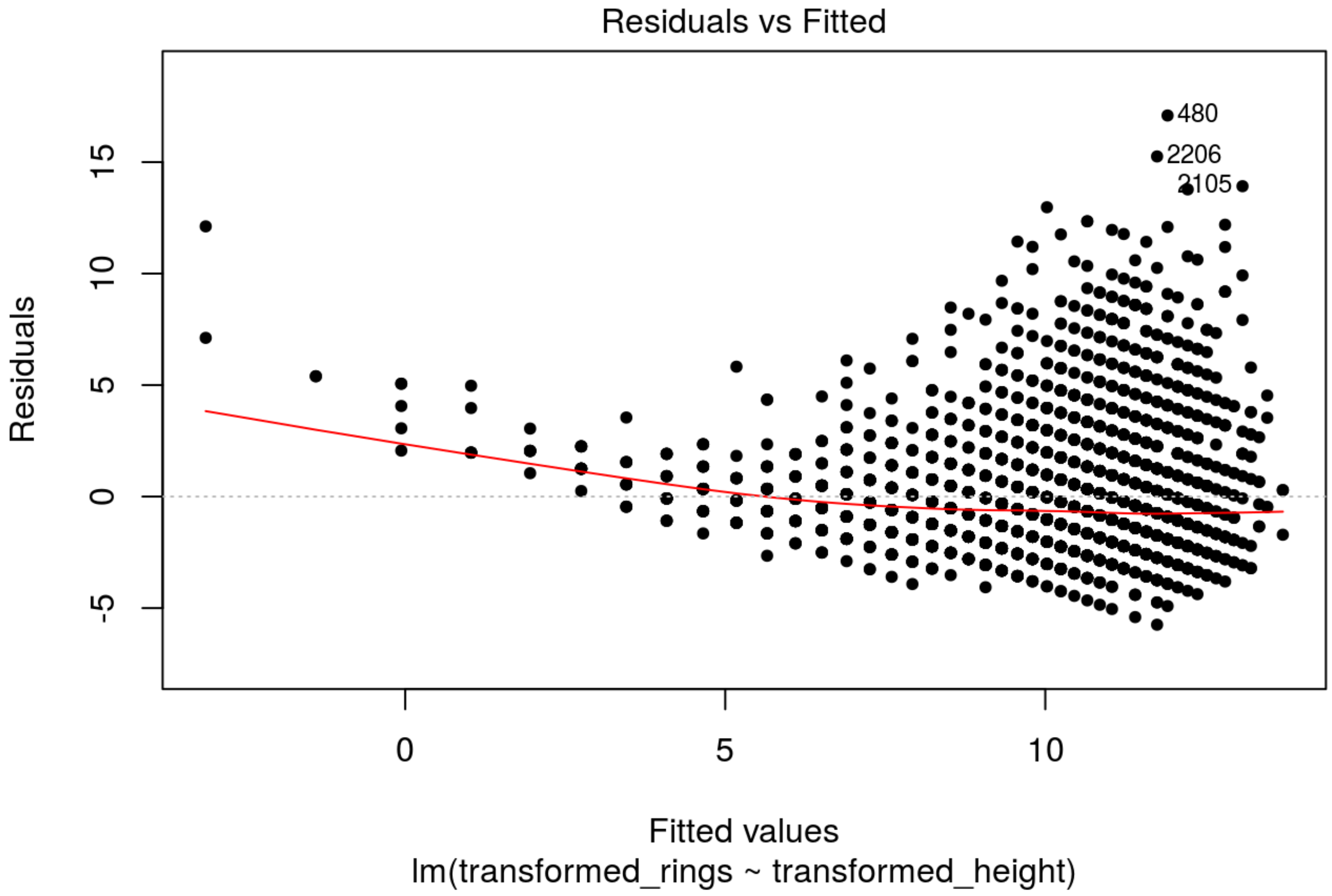
```
plot(model_re,which = 1,pch=20)
```



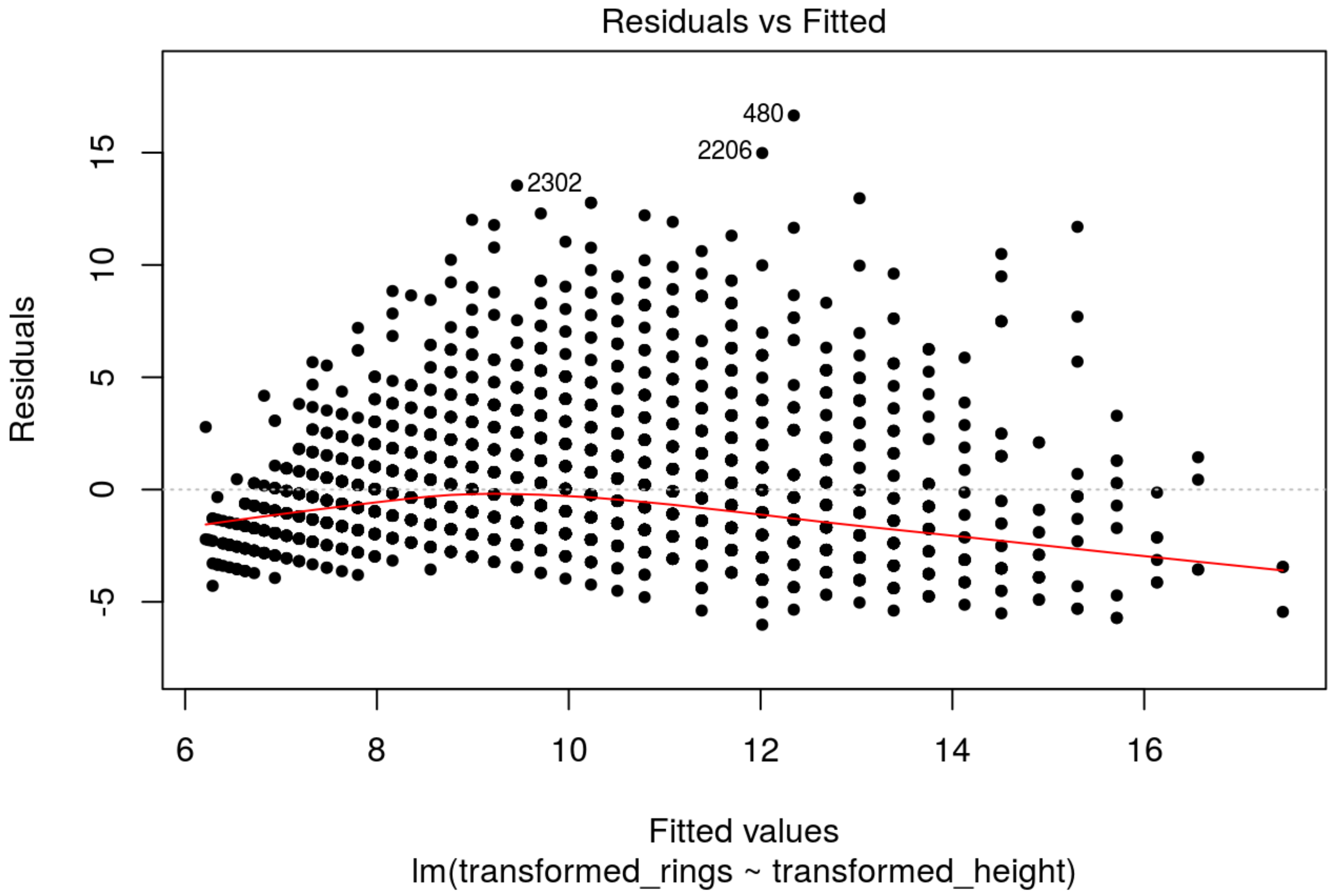
Transformed model attempts

- Then we tried two different kinds of tranformed height, which is respectively square of the height and the log of the height. These two graph, represented below, do not show healthy residuals.So we stick to the better model previously constructed , which is model_re.

```
transformed_height<-log(abalone_re$Height)
transformed_rings <- abalone_re$Rings
#transformed_rings <- as.numeric(unlist(transformed_rings))
model_transformed <- lm(transformed_rings~transformed_height)
plot(model_transformed,which=1,pch=20)
```



```
transformed_height<-abalone_re$Height^2
transformed_rings <- abalone_re$Rings
#transformed_rings <- as.numeric(unlist(transformed_rings))
model_transformed_sq <- lm(transformed_rings~transformed_height)
plot(model_transformed_sq,which=1,pch=20)
```



### Results

Since our model works,we then try to find the paramter estimates by using summary function. The results show that $\hat{\beta 1}$ is 2.8036 and $\hat{\beta 0}$ is 51.2185,so our linear formula is $ = 2.8036 + 51.2185*Height $ Height ∈ [0.025,0.25].What's more, we can see that the p-value is less than 2e-16,which indicates that there is a relationship bewtween height and its number of rings.

```
summary(model_re)
```

```
##
## Call:
## lm(formula = Rings ~ Height, data = abalone_re)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0229 -1.6694 -0.5351  0.8184 16.7210
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.8036     0.1491   18.80   <2e-16 ***
## Height       51.2185     1.0323   49.62   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.555 on 4170 degrees of freedom
## Multiple R-squared:  0.3712, Adjusted R-squared:  0.371
## F-statistic:  2462 on 1 and 4170 DF,  p-value: < 2.2e-16
```

- Here we compute the confidence interval for height = 0.128, which is in the range(9.278721,9.437726)

```
new_data_1 <- data.frame(Height = 0.128)
predict(model_re,newdata = new_data_1,interval = 'confidence')
```

```
##        fit      lwr      upr
## 1 9.359581 9.278721 9.440441
```

- Here we compute the confidence interval for height = 0.132, which is in the range(9.485499,9.6434411)

```
new_data_2 <- data.frame(Height = 0.132)
predict(model_re,newdata = new_data_2,interval = 'confidence')
```

```
##        fit      lwr      upr
## 1 9.564455 9.485499 9.643411
```

Conclusion and Discussion:

- Based on the plots and table we generated, there is a strong linear relationship between height and rings, so that we can use the height to predict its number of rings,and thus its age. In order to get better performance of the model, removing unnecessary data is extremely important in our case,which makes our model less error. Also, by evaluating p-value, our hypothesis which is there is linear relationship between the height of abalone and its number of rings is true. Furthermore, this research could be done better by other researchers by working on more complex linear model of height, or even nonparametric model to get more accurate data.