Homework 2

Diana Rueda

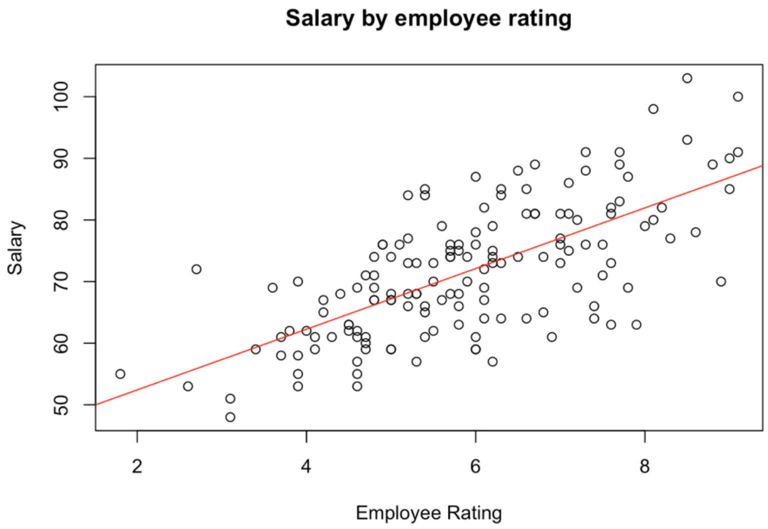
5/4/2019

Simple Linear Regression

Chapter 4

(P1)

The scatterplot representing the relationship between employee rating and salary looks like this:



The regression model for this data is:

```
##
## Call:
## lm(formula = manager$salary ~ manager$mrating)
##
## Coefficients:
##     (Intercept)  manager$mrating
##          42.575            4.925
```

The Anova table for this regression model is:

```
## Analysis of Variance Table
##
```

```
## Response: manager$salary
##                  Df Sum Sq Mean Sq F value    Pr(>F)
## manager$mrating   1 7978.5  7978.5  129.87 < 2.2e-16 ***
## Residuals       148 9092.3    61.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The t-test stat= 11.4 with p-value =P(t148>11.4) < 2.2e-16 = 0.Since p-value is smaller than α= 0.05, reject the null hypothesis and conclude that β1 is greater than 0. Likewise the F-test from the Anova table for this regression model is 129.8693212 on p-value < 2.2e-16 = 0. These F leads to the same p value as the t test, so we reject the null hypothesis and conclude that β1 is greater than 0.

(P2)

Regression of salary on ratings summary:

```
##
## Call:
## lm(formula = manager$salary ~ manager$mrating)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.4792  -4.5974  -0.1979   4.8511  18.5660
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      42.5750     2.6289    16.2   <2e-16 ***
## manager$mrating   4.9246     0.4321    11.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.838 on 148 degrees of freedom
## Multiple R-squared:  0.4674, Adjusted R-squared:  0.4638
## F-statistic: 129.9 on 1 and 148 DF,  p-value: < 2.2e-16
```

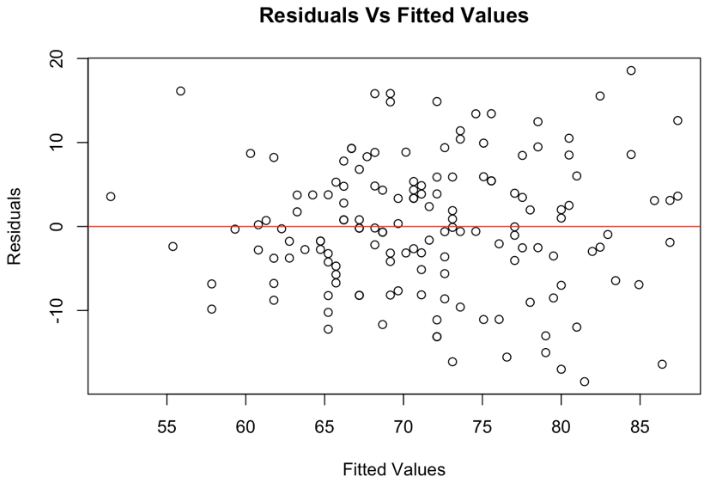
The following is the Anova table obtained by my own calculations. The values are very close to the data from the true Anova table.

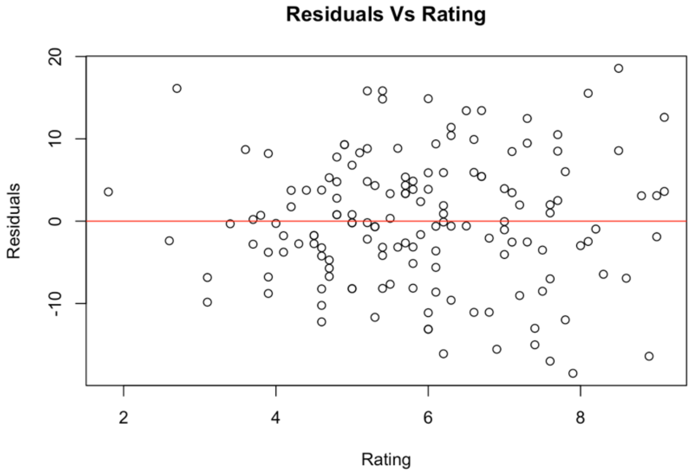
```
##                   df          SS          MS           F
## Regression     1.00000  7978.48977  7978.48977   129.86932
## Error        148.00000  9092.34356    61.43475
## Total        149.00000 17070.83333
```

(P3)

Residual plots:

For the data set manager.xlsx fit a regression of salary on ratings. Obtain the residual plot, plotting the residuals against the fitted values and also against ratings. Label both plots with a title and label the x-axis in both plots. Compare the plots and comment on similarities and differences in appearance. Make a list of model violations and aspects of model fit that can be addressed with a residual analysis. Assess each aspect with respect to both plots. Do the plots show the same or are there differences such that one plot would lead you to conclude that the model assumptions are violated but the other plot would not?





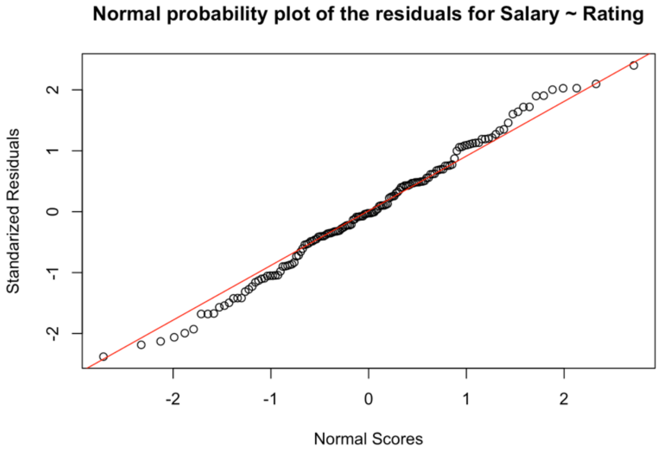
The plot against fitted values looks exactly the same as the plot against rating. The residuals seem to be evenly distributed across the zero line but their distribution along rating is not symmetrical. The data is more highly concentrated towards higher values of rating. In addition both plots show that the errors are not constant around the regression line.

a. Distribution of X values
b. Normality of errors ϵ
c. Linearity of the relationship between X and Y
d. Constant error around regression line

(P4)

For the regression model fit in problems 1-3, assess the assumption of normality of the errors. Identify an appropriate plot to use, obtain that plot and comment on the appearance.

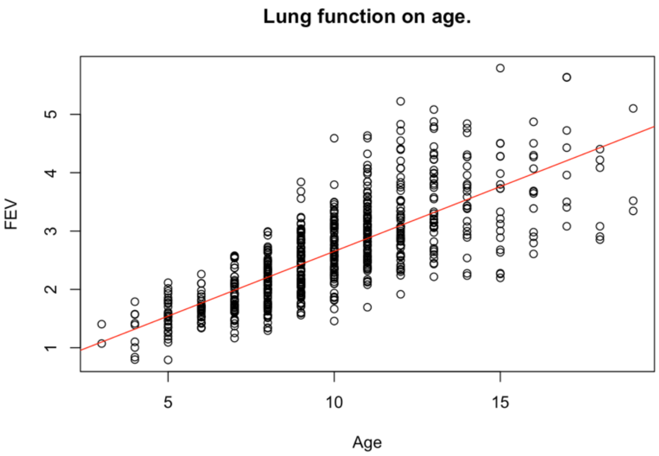
In order to assess the normality of errors we should look at the qq plot for this data.



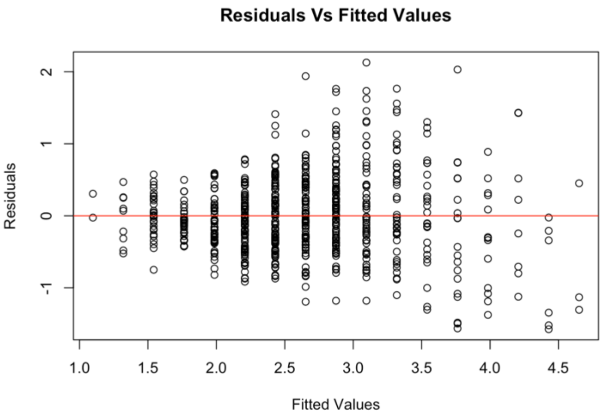
According to the distribution of residuals in this plot the unalined ends suggest a lack of fit of our current linear model. The behavior at both tails means that compared to the normal distribution there is more data located at the extremes of the distribution and less data in the center of the distribution.

(P5)

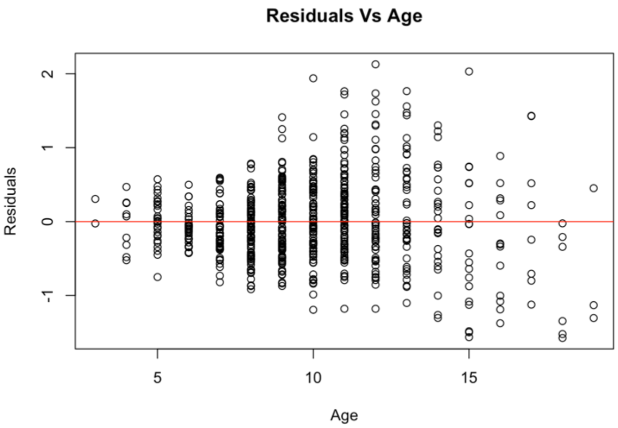
For the data set FEV.csv fit a regression model with lung function FEV (Y) as the response and age as the predictor. Carry out a residual analysis and assess which model assumptions are violated. State all conclusions clearly and in context.



Lung function on age.

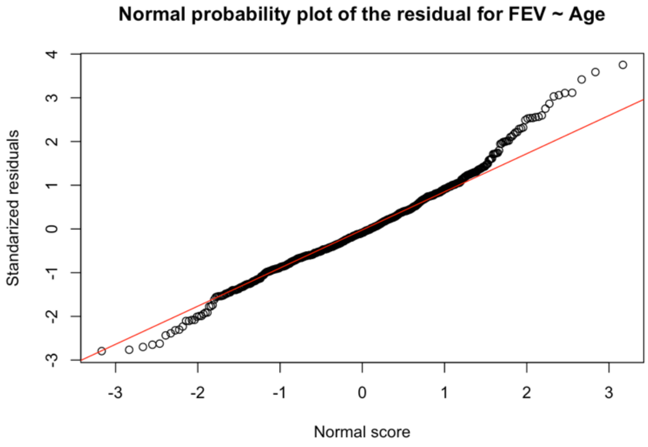


Residuals Vs Fitted Values



Residuals Vs Age

The scatterplot of lung function on age shows a positive slope but the bands covering the data above and below are not parallel which suggests a noncostant variance of errors. We also see a asymmetrical distribution of the age values. The following plots (Residuals vs age and residuals vs fitted values) look the same. However, they show the uneven distribution of errors. There is a higher concentration in the middle and again there are no two even lines that can enclose the data since the errors are closer together in the lower values of age.



Normal probability plot of the residual for FEV ~ Age

According to the distribution of residuals in this plot the unalined ends suggest a lack of fit of our current linear model. The behavior at both tails means that compared to the normal distribution there is more data located at the extremes of the distribution and less data in the center of the distribution. We have no linearity.

```
##
## Call:
## lm(formula = FEV$FEV ~ FEV$Age)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.57539 -0.34567 -0.04989  0.32124  2.12786
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.431648   0.077895   5.541 4.36e-08 ***
## FEV$Age     0.222041   0.007518  29.533  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5675 on 652 degrees of freedom
## Multiple R-squared:  0.5722, Adjusted R-squared:  0.5716
## F-statistic: 872.2 on 1 and 652 DF,  p-value: < 2.2e-16
```

From the summary table above we can see a difference on the p value of the F-statistic and t-value for this data. However, a t-value of 5.542 with p < 4.36e-08 still indicates a possitive β1 since p-value is smaller than α= 0.05. Thus we reject the null hypothesis and conclude that β1 is greater than 0. The same comes from a high F-statistic with p-value < 2.2e-16.

(P6)

Using the data set FEV.csv carry out a lack of fit test. State the steps needed to take and provide the R code you used. State your conclusions in context.

```
## Analysis of Variance Table
##
```

```
## Model 1: FEV$FEV ~ FEV$Age
## Model 2: FEV$FEV ~ factor(FEV$Age)
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)    
## 1    652 210.00                                 
## 2    637 195.01 15    14.986 3.2633 2.91e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to the low F-statistic = 3.2633 and p-value < 2.91e-05 obtained from the lack of fit test, since p-value is smaller than α= 0.05, we reject the null hypothesis and beta1 is greated than zero, but we sould be better off fitting another model due to model violations.

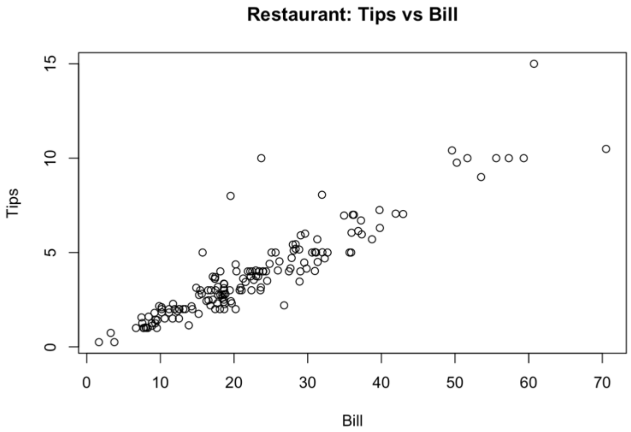
The calculations made to obtain the anova table values are:

```
#Calculations to obtain anova values for lack of fit test
SSPE <- 195.01
SSE <- 210.00
c <- 16
n <- 652
SSLF <- SSE - SSPE
MSLF <- SSLF / (c-2)
MSPE <- SSPE / (n-c)
F1 <- MSLF / MSPE
```

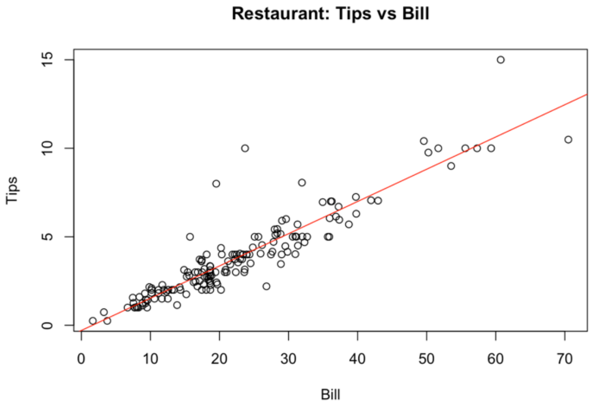
(P7)

Carry out a comprehensive data analysis of the association between tips and price of a meal for the restauranttips.xlsx data set. Clearly state all the steps you should take, fit the model, check assumptions and state conclusions in context. Answer the question: do tips increase linearly, on average, with increasing price of a meal? Justify your answer with the analysis you carried out.

1)Scatterplot of the data



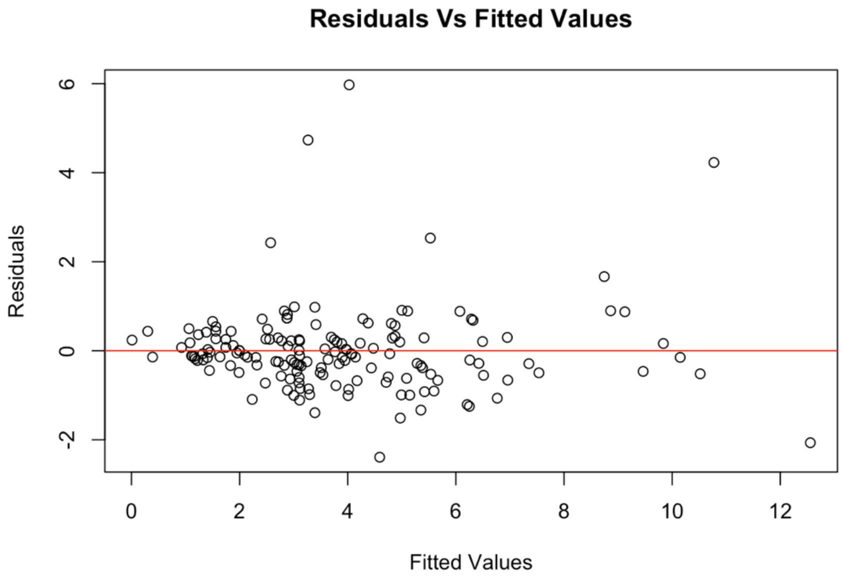
2)Fit the model

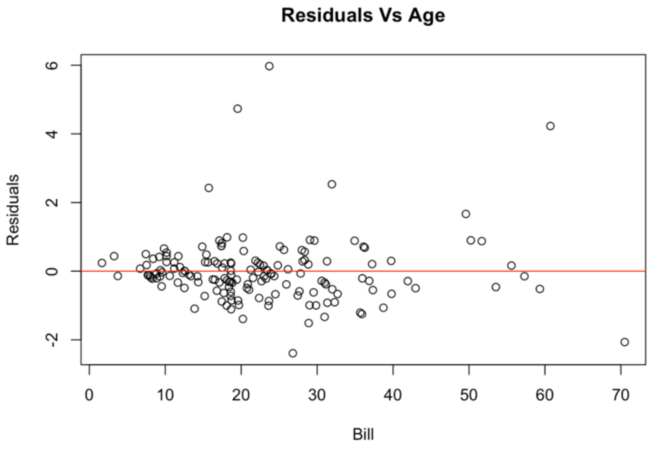


```
## Analysis of Variance Table
##
## Response: restauranttips$Tip
##                       Df Sum Sq Mean Sq F value    Pr(>F)
## restauranttips$Bill    1 765.53  765.53  797.87 < 2.2e-16 ***
## Residuals            155 148.72    0.96
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The model shows a positive slope, but the data is not evenly distributed and we have outliers. No constant error.

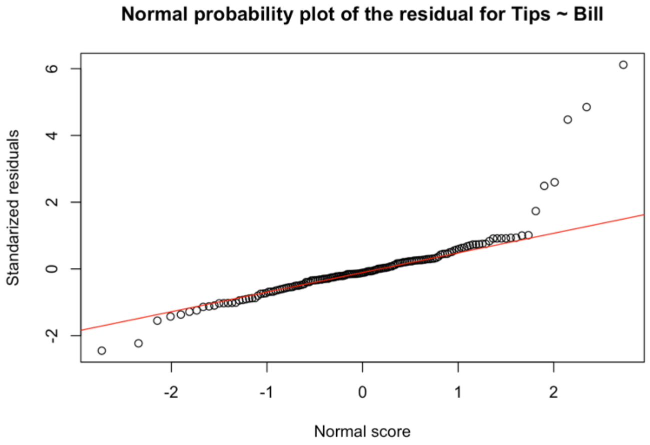
3)Residual plots





The residual plots against fitted values and bill amount are consistent. However we have nonconstant error varience and an uneven distribution of the errors across the zero line. The data seems to be concentrated in the fisrt half of the x values, and more loosely distributed in the second half.

4)Normal Q-Q Plots to assess normality of errors.



The Q-Q plot clearly showa an abnormality in the tail behavior of our data. The assumptions of normality and linearity are violated.

5)Lack of fit test

```
## Analysis of Variance Table
##
## Model 1: restauranttips$Tip ~ restauranttips$Bill
## Model 2: restauranttips$Tip ~ factor(restauranttips$Bill)
##   Res.Df     RSS  Df Sum of Sq      F  Pr(>F)
## 1    155 148.717
## 2     18   8.086 137    140.63 2.2851 0.02287 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-statistic of 2.2851 resulting of the lack of fit test with a p-value = 0.02287 which is smaller than α= 0.05 we reject the null hypothesis and conclude that β1 is greater than 0. The tips increase with bill amount.

However due to model violations it is better to fit another model to our data.

(P8)

Show b1 is normally distributed in the simple linear regression, using model assumptions made that are needed to justify the normal distribution and identify the terms involved. Using the data set manager.xlsx calculate the ki explicitly for the first 20 observations.

In yi = B0 + B1 xi + Ei, if the errors are normally distributed and B0 + B1 Xi is a non random constant, yi depends on the normality of Ei, so the sum of Ki yi is also normal. In addition remember that bi = Sum(ki yi), which indicates it is also normally distributed.

```
###(P8)
#calculaate Ki
xi <- manager$mrating[1:20]

Ki <- function(x) {
  ki <- (x - mean(x)) / sum((x - mean(x))^2)
  ki
}

K20 <- Ki(xi)
```

K1 = 0.0374232 — K11 = -0.0609273— K2 = -0.035923 — K12 = 0.0390902— K3 = -0.0075847 — K13 = 0.0257545— K4 = -0.0242542 — K14 = -0.0042507— K5 = 0.0340893 — K15 = -0.0242542— K6 = 0.0124188 — K16 = -0.0225873— K7 = 0.0174197 — K17 = -0.0075847— K8 = 0.0240875 — K18 = -0.0275882— K9 = 0.047425 — K19 = -0.0275882— K10 = 0.0140858 — K20 = -0.0092516

(P9)

Show the residuals are normally distributed and show ∑ni=1 residuals = 0.

In yi = B0 + B1 Xi +Ei, B0 + B1 Xi is a non random constant and Ei is the difference between the observed value and the estimated expected value: ei = yi - (B0+B1Xi), difference from mean or if we visualize it it is the distance from a given data point to the regression line. If the data is normally distributed we can expect the errors form this data to do the same.

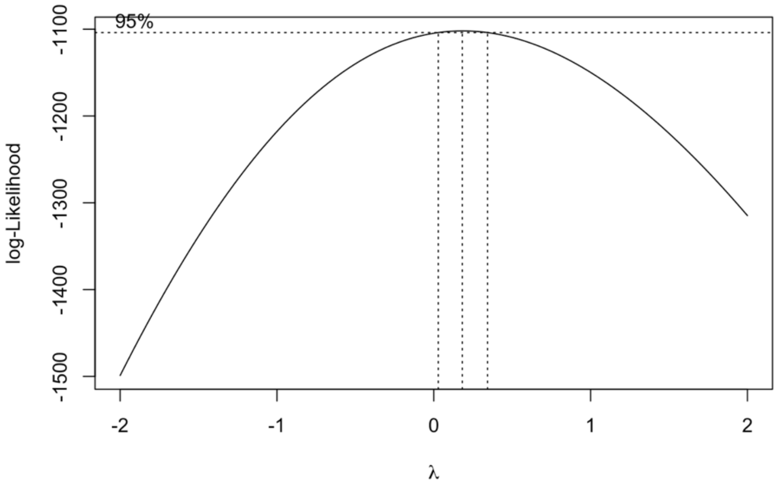
The residuals have mean zero: E[ei] = E[Yi−ˆYi] = E[Yi]−E[ˆYi] = β0+β1Xi − E[b0+b1Xi] = 0

The sum of the residuals is zero: ei = ∑(Yi−ˆYi) = ∑Yi − ∑ˆYi = ∑Yi − ∑(b0+b1Xi) = n ̄Y−nb0−nb1 =0̄

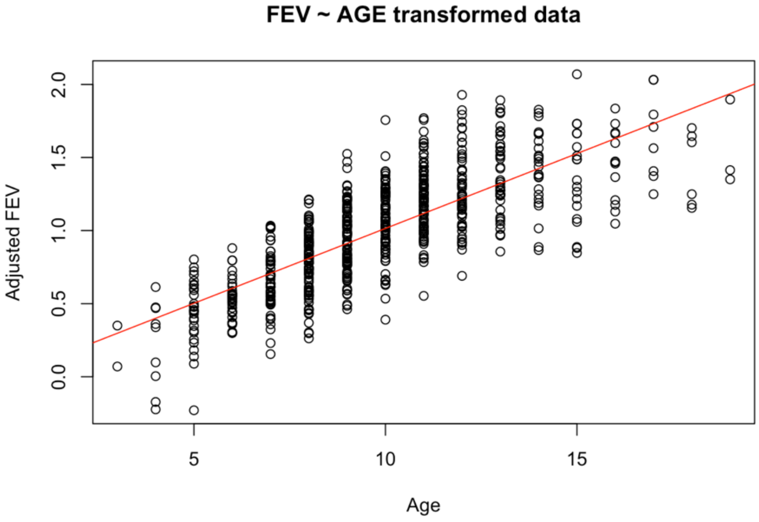
Chapter 5

(P1)

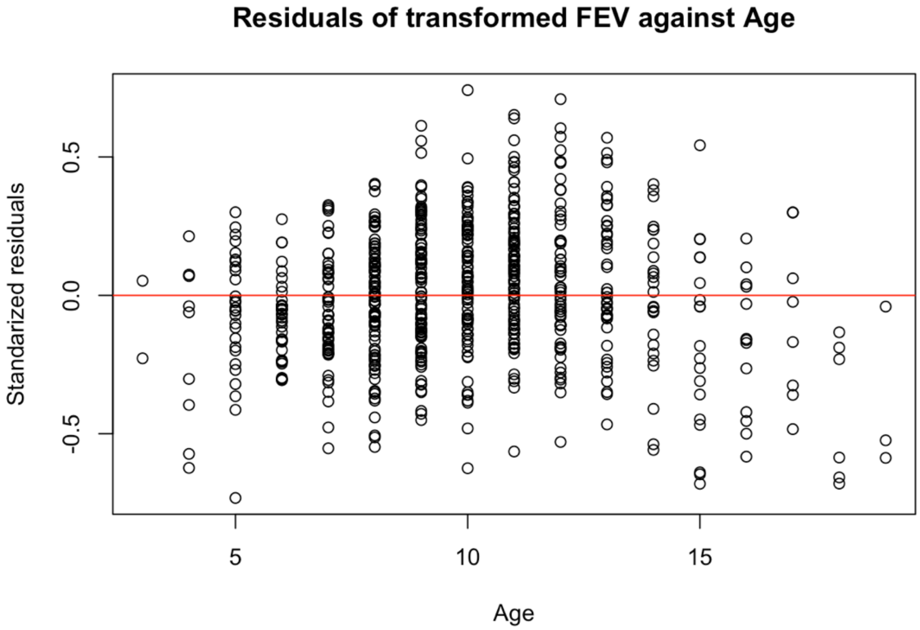
Using the data set FEV.csv use a Box-Cox transformation to identify the best transformation of FEV for the regression of FEV on age. Carry out the regression analysis with the transformed data set.



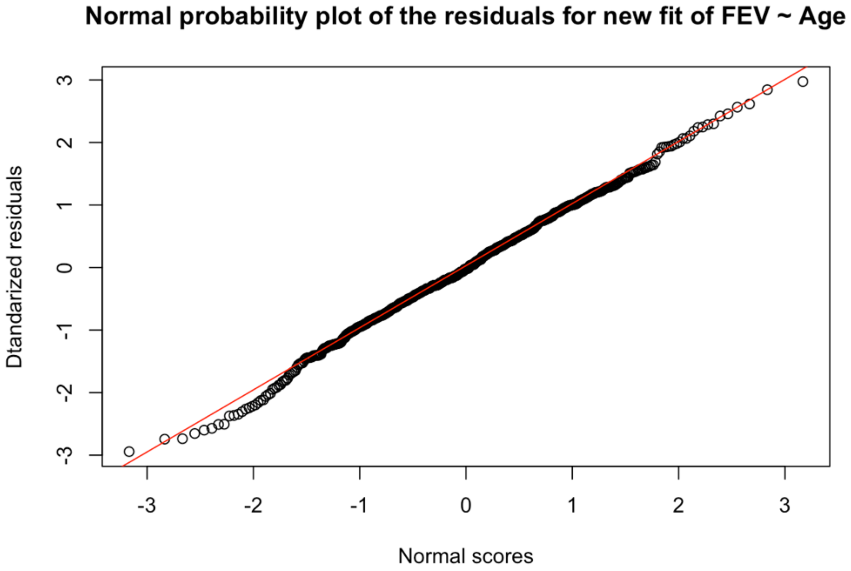
Using Box-cox we found that lambda = 0.1818182.



```
## Analysis of Variance Table
##
## Response: newY
##            Df Sum Sq Mean Sq F value    Pr(>F)
## FEV$Age     1 59.864  59.864  962.24 < 2.2e-16 ***
## Residuals 652 40.563   0.062
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Residuals of transformed FEV against Age



Normal probability plot of the residuals for new fit of FEV ~ Age

The first plot “FEV~ Age transformed data” show the adjusted model for lung function on age. The data is more evenly distributed with what looks like constant error variance. It can also be enclosed in two stright lines parallel to the regression line which the original model could not do.

Next we have the anova table whih shows a higher F-statistic of 962.24 and a p-value = 2.2e-16 which is smaller than alpha = 0.05. Now we can reject the null hypothesis and conclude that beta1 is greater than zero. Note the F-statistic is also higher than in the original model.

The second graph is the plot of the residuals against age. The transformed model shows a more even distribution of residuals across the zero line and along the age values which means a more constant error than before the transformation.

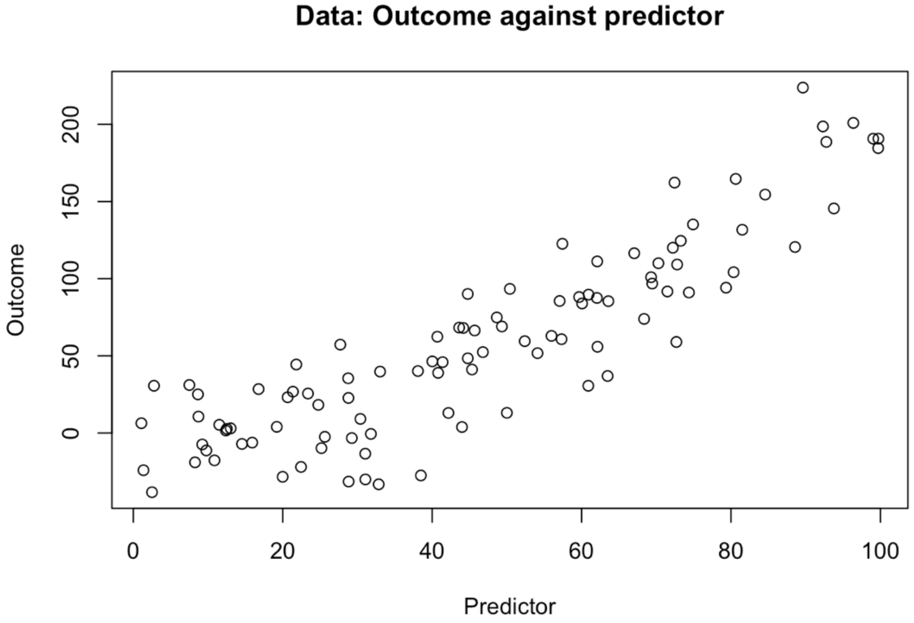
The last graph is the Normal Q-Q plot. We can se a change in the behavior of the tails of the data after the transformation of y. All the data is closer to the line which suggest normality of errors and linearity.

In conclussion the new model for the regression FEV~Age based on lambda = 0.1818182 is a better fit than the original linear model.

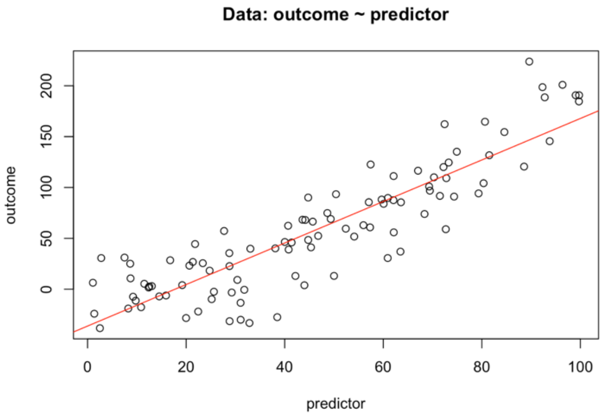
(P2)

Using the data set hw2chapter5p2.csv carry out a regression analysis with model fit assessment. If necessary, determine a transformation of either outcome (Y) or predictor (X) variable. Justify your choice with model assumptions and residual analysis or any other tests as appropriate.

1)Scatterplot of the data



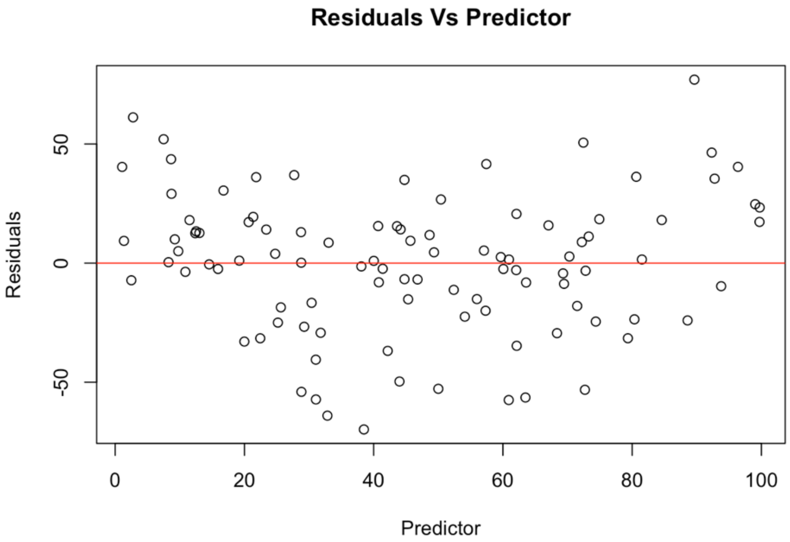
2)Fit the model



```
## Analysis of Variance Table
##
## Response: data$outcome
##                  Df Sum Sq Mean Sq F value    Pr(>F)
## data$predictor  1 301103  301103  351.78 < 2.2e-16 ***
## Residuals       98  83882     856
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

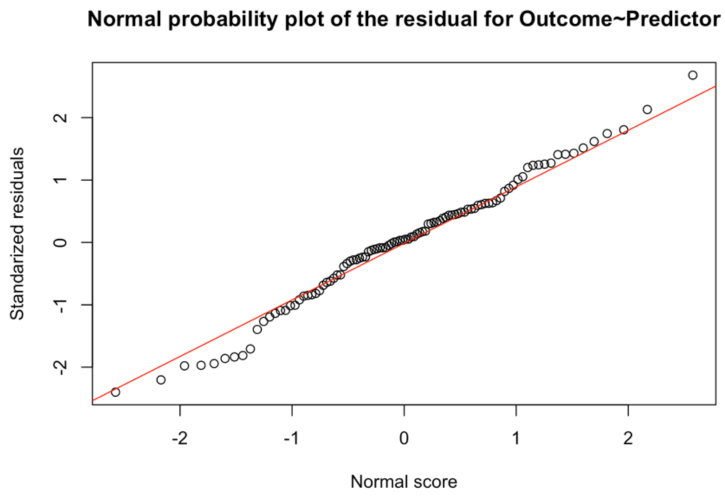
The model shows a positive slope, but the data can not be enclosed in two parallel lines. It would seem x^2 would fit the data better. Error variance seems constant. The F-statistic of 351.78 with a p-value = 2.2e-16 which is smaller than α= 0.05 indicates we must reject the null hypothesis and conclude that β1 is greater than 0.

3)Residual plots



The residual plot shows an uneven distribution of residuals across the zero line but symmetrical along the predictor values. This indicates another model migth be a better fit, but the error variance is constant.

4)Normal Q-Q Plots to assess normality of errors.



The Q-Q plot clearly showa an abnormality in the tail behavior of our data. The assumptions of normality and linearity are violated.

5)Lack of fit test

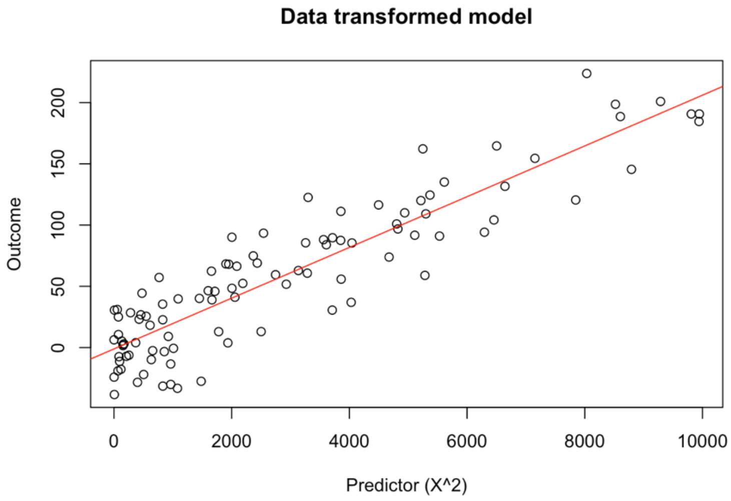
```
## Warning in anova.lm(data_model2): ANOVA F-tests on an essentially perfect
## fit are unreliable
## Analysis of Variance Table
##
## Response: data$outcome
##                        Df Sum Sq Mean Sq F value Pr(>F)
## factor(data$predictor) 99 384985  3888.7
## Residuals               0      0
```

From the information of the anova table for the lack of fit test we have a nearly perfect data set, so we reject the null hypothesis and conclude that β1 is greater than 0.

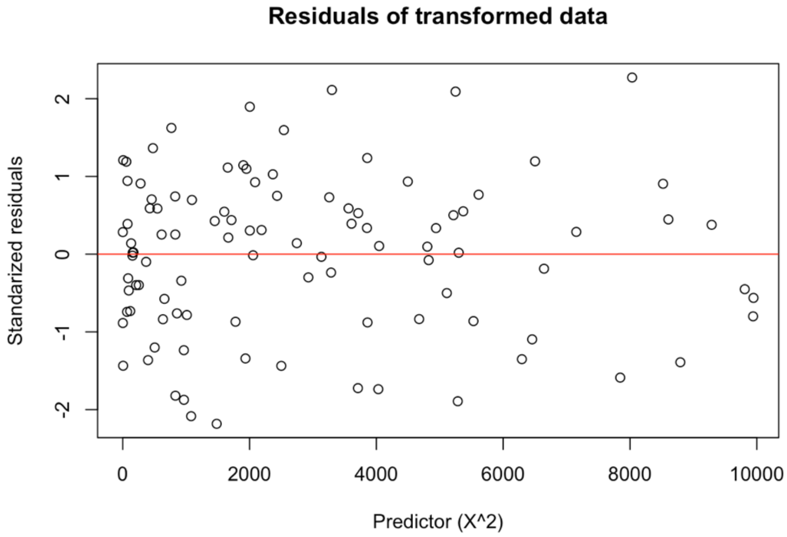
However due to model violations it is better to fit another model to our data.

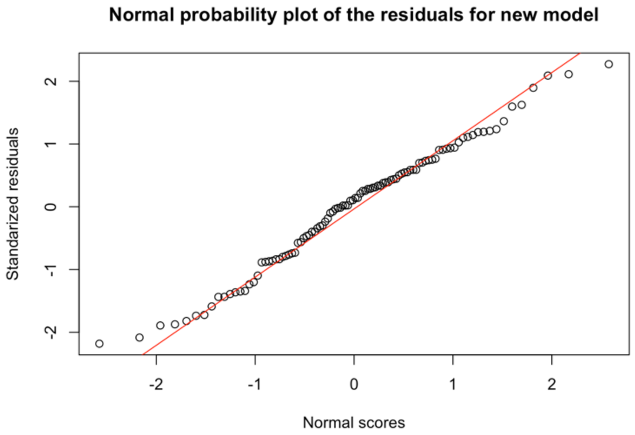
6)It would seem an appropiate transformation, due to constant error, would be on X.

```
## Analysis of Variance Table
##
## Response: data$outcome
##              Df Sum Sq Mean Sq F value    Pr(>F)
## data_x_trans  1 317012  317012  457.05 < 2.2e-16 ***
## Residuals    98  67973     694
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



The data in the scatterplot shows a more constant distribution of data. The data set can be enclosed in two stright lines parallel to the regression line. The only violation that seems to appear is a non constant error variance.





The residual plot confirms the distribution of our data, but also the non constant variance of error. The normal probability plot continues to show abnormal behavior of our errors.

Code Appendix

```
### (P1)
#load data and plot
library(readxl)
manager <- read_excel("~/Downloads/manager.xlsx")

plot(manager$mrating, manager$salary, main = "Salary by employee rating",
xlab = "Employee Rating", ylab = "Salary")

salary_on_rating <- lm(manager$salary ~ manager$mrating)
abline(salary_on_rating, col = "red")

#Regression model
salary_on_rating
#anova table
anova(salary_on_rating)

#F-test
f_value <- anova(salary_on_rating)["manager$mrating", "F value"]

###(P2)

summary(salary_on_rating)

#Create Anova table from sumnary table data
intercept <- coef(lm(salary_on_rating))["(Intercept)"]
slope <- coef(lm(salary_on_rating))["manager$mrating"]

Yhat <- function(x){
  estimate <- intercept + slope * x
  estimate
}
```

```
SSR <- sum((Yhat(manager$mrating) - mean(manager$salary))^2)
SSE <- sum((manager$salary - Yhat(manager$mrating))^2)
TotalSS <- sum((manager$salary - mean(manager$salary))^2)

n <- 150

MSR <- SSR / 1
MSE <- SSE / (n - 2)

f <- MSR / MSE

#make table
anova_data <- matrix(c(1, SSR, MSR, f, n-2, SSE, MSE, NA, n-1, TotalSS, NA,
NA), ncol = 4, byrow = TRUE)
colnames(anova_data) <- c("df", "SS", "MS", "F")
rownames(anova_data) <- c("Regression", "Error", "Total")
anova_table <- as.table(anova_data)

anova_table
###(P3)
#plot residuals against fitted values and against rating

plot(salary_on_rating$fitted.values, salary_on_rating$residuals, main =
"Residuals Vs Fitted Values", xlab = "Fitted Values", ylab = "Residuals")
abline(0, 0, col = "red")

plot(manager$mrating, resid(salary_on_rating), main = "Residuals Vs Rating",
xlab = "Rating", ylab = "Residuals")
abline(0, 0, col = "red")

###(P4)
#qqplot to assess normality of errors.
salary_rating_stdres <- rstandard(salary_on_rating)
qqnorm(salary_rating_stdres, main = "Normal probability plot of the residuals
for Salary ~ Rating", xlab = "Normal Scores" , ylab = "Standarized
Residuals")
qqline(salary_rating_stdres, col = "red")

###(P5)
#FEV regression model

FEV <- read_excel("~/Downloads/FEV.xls")
lung_on_age <- lm(FEV$FEV ~ FEV$Age)

plot(FEV$Age, FEV$FEV, main = "Lung function on age.", xlab = "Age", ylab =
"FEV")
abline(lung_on_age, col = "red")

#residual plots
plot(lung_on_age$fitted.values, lung_on_age$residuals, main = "Residuals Vs
Fitted Values", xlab = "Fitted Values", ylab = "Residuals")
abline(0, 0, col = "red")

plot(FEV$Age, resid(lung_on_age), main = "Residuals Vs Age", xlab = "Age",
ylab = "Residuals")
abline(0, 0, col = "red")
```

```
#qq plot
lung_age_stdres <- rstandard(lung_on_age)
qqnorm(lung_age_stdres, main = "Normal probability plot of the residual for
FEV ~ Age", xlab = "Normal score", ylab = "Standarized residuals")
qqline(lung_age_stdres, col = "red")

#tables
summary(lung_on_age)

#goodness of fit test
lung_on_age2 <- lm(FEV$FEV ~ factor(FEV$Age))
#anova(lung_on_age)
#anova(lung_on_age2)
anova(lung_on_age, lung_on_age2)
#Calculations to obtain anova values for lack of fit test
SSPE <- 195.01
SSE <- 210.00
c <- 16
n <- 652
SSLF <- SSE - SSPE
MSLF <- SSLF / (c-2)
MSPE <- SSPE / (n-c)
F1 <- MSLF / MSPE

###(P7)
library(readr)
restauranttips <- read_csv("~/Downloads/restauranttips.csv")

plot(restauranttips$Bill, restauranttips$Tip, main = "Restaurant: Tips vs
Bill", xlab = "Bill", ylab = "Tips")
#fit the model
tips_on_bill <- lm(restauranttips$Tip ~ restauranttips$Bill)
plot(restauranttips$Bill, restauranttips$Tip, main = "Restaurant: Tips vs
Bill", xlab = "Bill", ylab = "Tips")
abline(tips_on_bill, col = "red")

anova(tips_on_bill)

#Residual plots
plot(tips_on_bill$fitted.values, tips_on_bill$residuals, main = "Residuals Vs
Fitted Values", xlab = "Fitted Values", ylab = "Residuals")
abline(0, 0, col = "red")

plot(restauranttips$Bill, resid(tips_on_bill), main = "Residuals Vs Age",
xlab = "Bill", ylab = "Residuals")
abline(0, 0, col = "red")

#q-q plot
tips_bill_stdres <- rstandard(tips_on_bill)
qqnorm(tips_bill_stdres, main = "Normal probability plot of the residual for
Tips ~ Bill", xlab = "Normal score", ylab = "Standarized residuals")
qqline(tips_bill_stdres, col = "red")

#lack of fit test
tips_on_bill2 <- lm(restauranttips$Tip ~ factor(restauranttips$Bill))
anova(tips_on_bill, tips_on_bill2)
```

```
###(P8)
#calculaate Ki
xi <- manager$mrating[1:20]

Ki <- function(x) {
  ki <- (x - mean(x)) / sum((x - mean(x))^2)
  ki
}

K20 <- Ki(xi)

###Cahpter 5
###(P1)
#Box - cox
library(MASS)
x <- FEV$FEV
y <- FEV$Age
bc <- boxcox(x ~ y, plotit = T)

lambda <- bc$x[which.max(bc$y)]
#transformation
powerTransform <- function(y, lambda1, lambda2 = NULL, method = "boxcox") {
  boxcoxTrans <- function(x, lam1, lam2 = NULL) {
    lam2 <- ifelse(is.null(lam2), 0, lam2)
    if (lam1 == 0L) {
      log(y + lam2)
    } else {
      (((y + lam2)^lam1) - 1) / lam1
    }
  }
  switch(method
         , boxcox = boxcoxTrans(y, lambda1, lambda2)
         , tukey = y^lambda1
  )
}

newY <- powerTransform(FEV$FEV, lambda)
FEV_new <- lm(newY ~ FEV$Age)

#regression analisys
plot(FEV$Age, newY, main = "FEV ~ AGE transformed data", xlab = "Age", ylab =
"Adjusted FEV")
abline(FEV_new, col = "red")
anova(FEV_new)

#residual plots
FEV_new_stdres <- rstandard(FEV_new)
plot(FEV$Age, FEV_new$residuals, main = "Residuals of transformed FEV against
Age", xlab = "Age", ylab = "Standarized residuals")
abline(0, 0, col = "red")

#qq plot for new model
qqnorm(FEV_new_stdres, main = "Normal probability plot of the residuals for
new fit of FEV ~ Age", xlab = "Normal scores", ylab = "Dtandarized
residuals")
qqline(FEV_new_stdres, col = "red")
```

```
###(P2)
#visualize data
library(readxl)
data <- read_excel("~/Downloads/hw2ch4problemlast.xlsx")
plot(data$predictor, data$outcome , main = "Data: Outcome against predictor",
xlab = "Predictor", ylab = "Outcome")
#fit the model
data_model <- lm(data$outcome ~ data$predictor)
plot(data$predictor, data$outcome, main = "Data: outcome ~ predictor", xlab =
"predictor", ylab = "outcome")
abline(data_model, col = "red")

anova(data_model)

#Residual plots
plot(data$predictor, resid(data_model), main = "Residuals Vs Predictor", xlab
= "Predictor", ylab = "Residuals")
abline(0, 0, col = "red")

#q-q plot
data_stdres <- rstandard(data_model)
qqnorm(data_stdres, main = "Normal probability plot of the residual for
Outcome~Predictor", xlab = "Normal score", ylab = "Standarized residuals")
qqline(data_stdres, col = "red")

#lack of fit test
data_model2 <- lm(data$outcome ~ factor(data$predictor))
anova(data_model2)

#transformation on X
data_x_trans <- data$predictor^(2)

model_trans <- lm(data$outcome ~ data_x_trans)
anova(model_trans)
plot(data_x_trans, data$outcome, main = "Data transformed model", xlab =
"Predictor (X^2)", ylab = "Outcome")
abline(model_trans, col = "red")

#residual plots
new_stdres <- rstandard(model_trans)
plot(data_x_trans, new_stdres, main = "Residuals of transformed data", xlab =
"Predictor (X^2)", ylab = "Standarized residuals")
abline(0, 0, col = "red")

#qq plot for new model
qqnorm(new_stdres, main = "Normal probability plot of the residuals for new
model", xlab = "Normal scores", ylab = "Standarized residuals")
qqline(new_stdres, col = "red")
```