Page1:

1. List all the points (row numbers) you removed (indexed on the original dataset) as outlier points

Answer:

*We removed these points referring to original datasets are 369,373,365,367,370,372,368,374,377

*We removed these points in the 3 iterations. (3 points in every iteration)

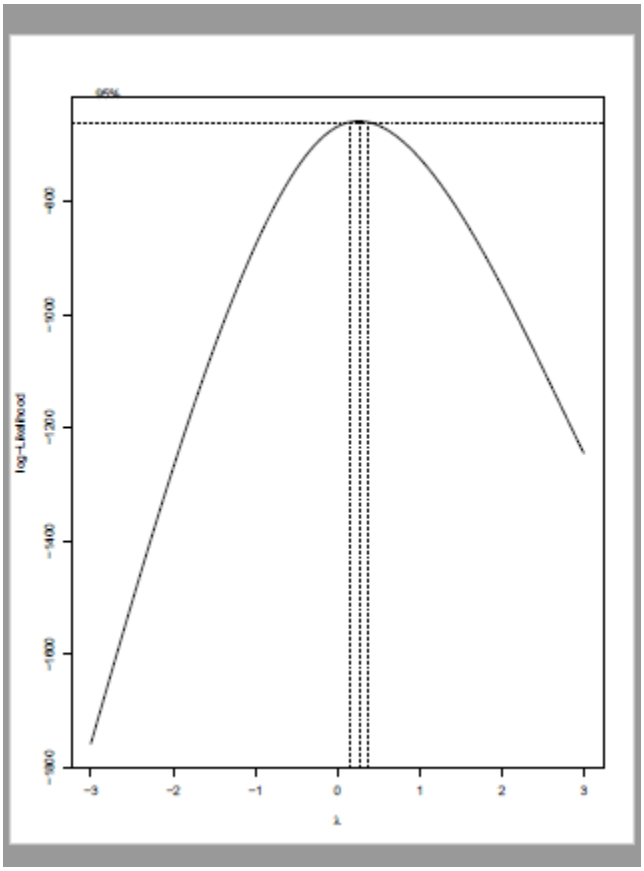
2. Box-Cox Transformation - Plot the Box-Cox transformation curve (Log-likelihood vs Parameter value). What is the best value of the parameter you got?

Answer:

We select the value of lambda where log likelihood is maximum. It is visible from the graph.

The value was calculated after removing all sets of outliers from the data (9 OUTLIERS)

Lambda=0.2727



Page 2:
3. Diagnostic plots used for identification of outliers. Please only include the Standard residuals vs Leverage vs Cook's distance plots (do not put other 3 plots you obtain for R). The final diagnostic plot obtained after removing all outliers should also be included.

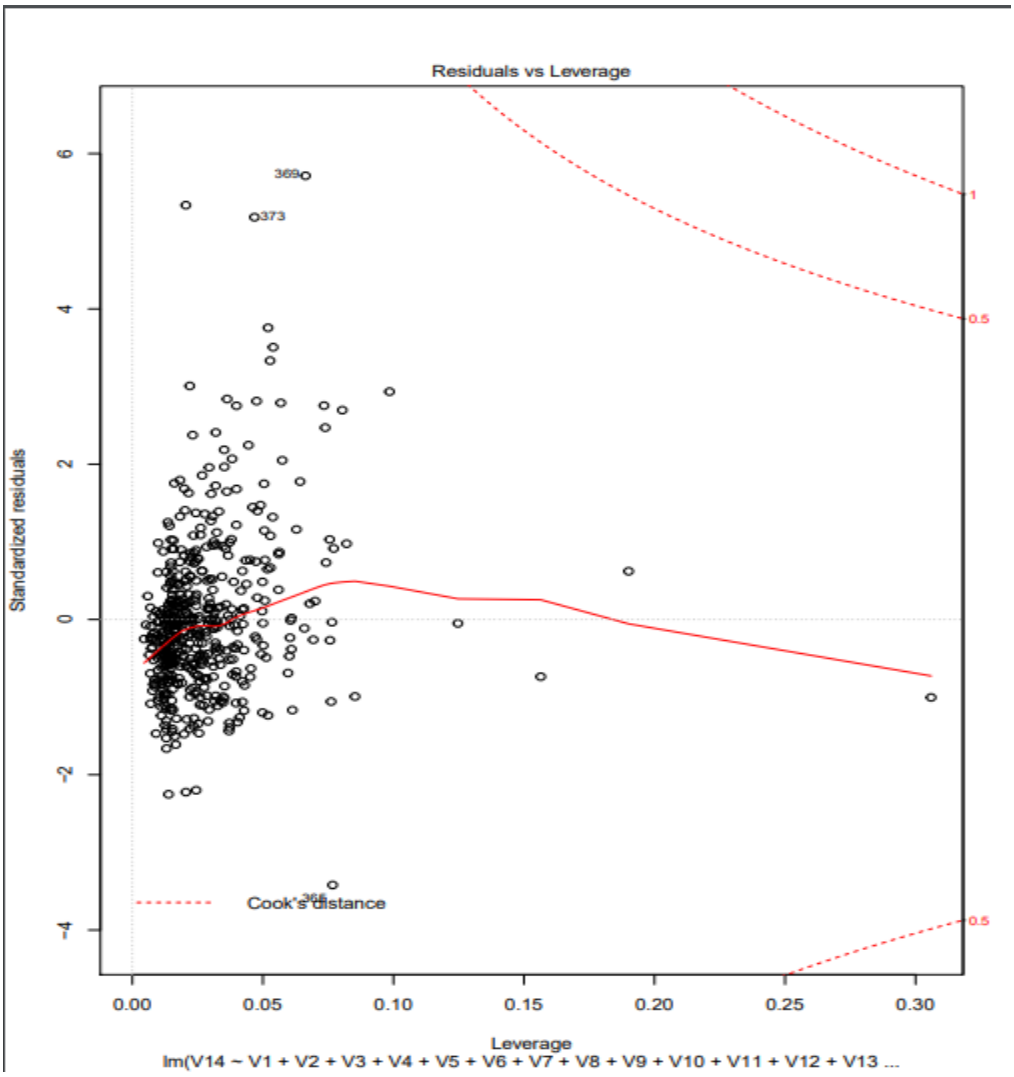
Answer:

We removed outliers in three steps.

I have shown graph for all the outliers.

About 99% of the sampled values of a standard normal random variable are in the range [−3, 3].

I removed some of the points which are outside these ranges.



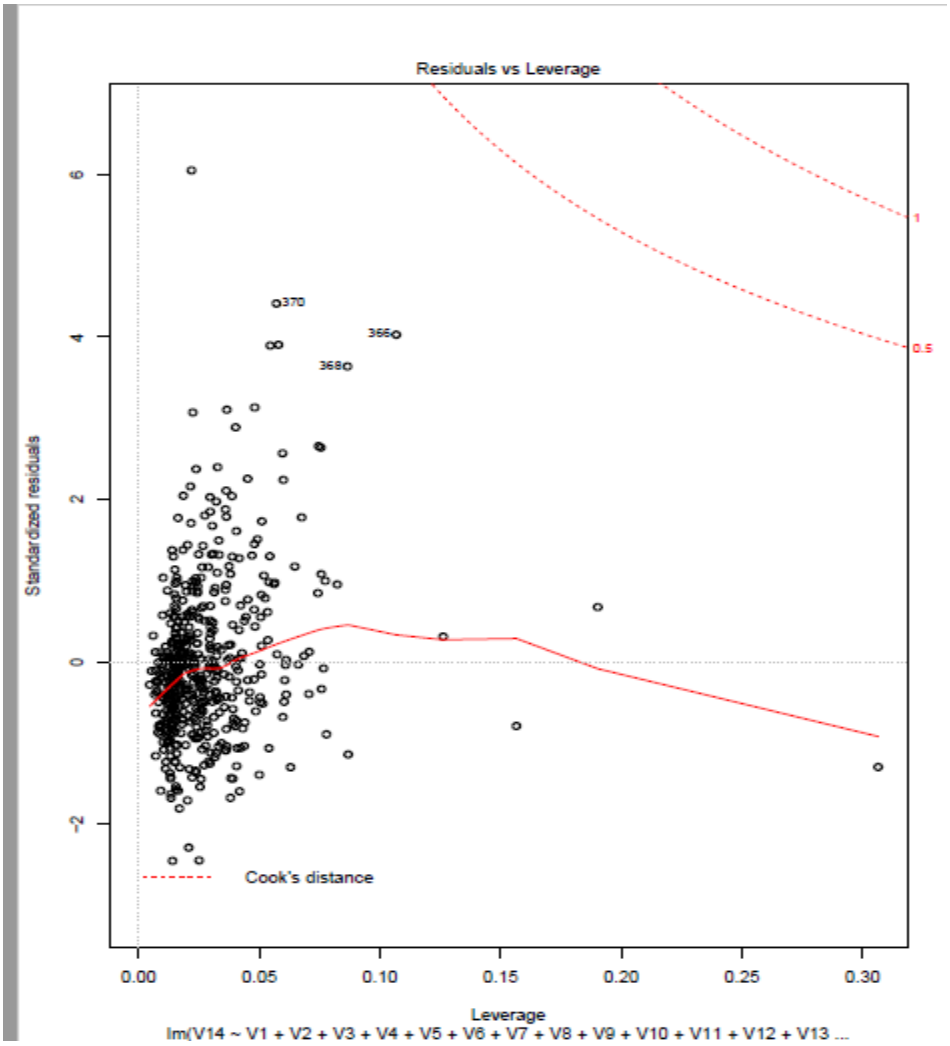
Page 2:
3. Diagnostic plots used for identification of outliers. Please only include the Standard residuals vs Leverage vs Cook's distance plots (do not put other 3 plots you obtain for R). The final diagnostic plot obtained after removing all outliers should also be included.

Answer:

This graph was also used to remove outliers.

About 99% of the sampled values of a standard normal random variable are in the range [−3, 3].

I removed some of the points which are outside these ranges.



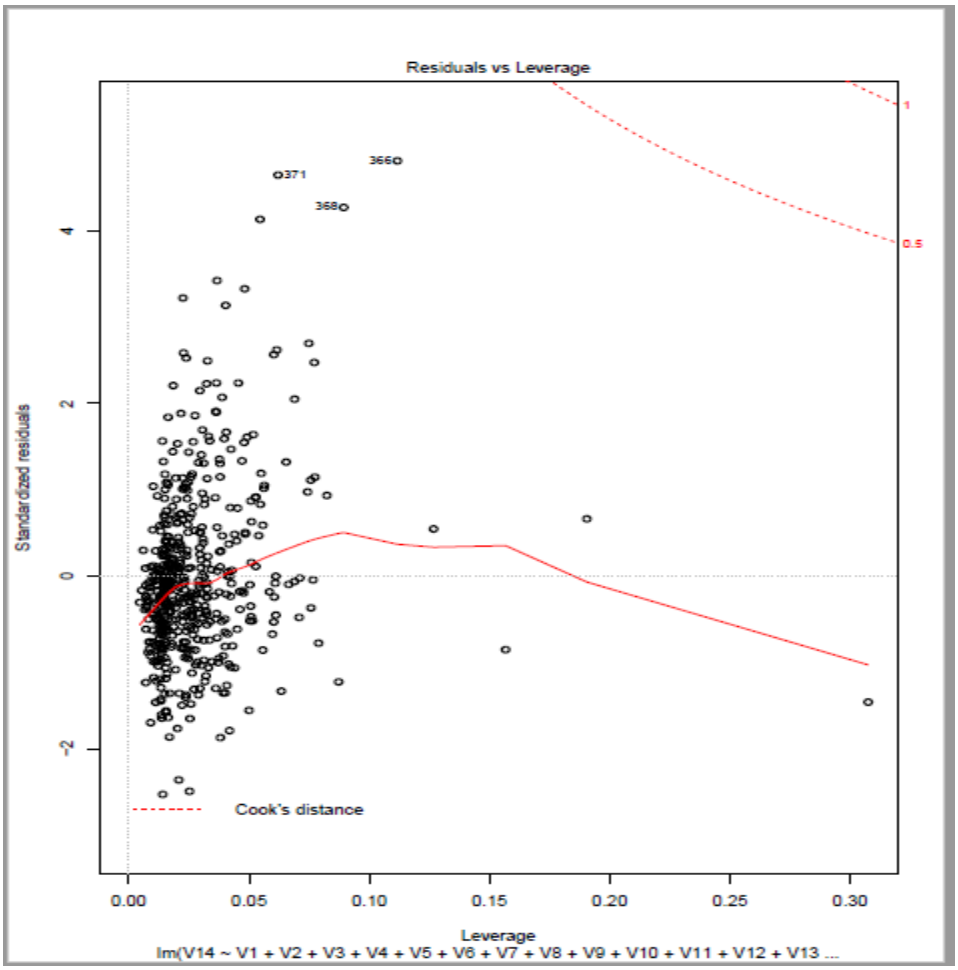
Page 2:
3. Diagnostic plots used for identification of outliers. Please only include the Standard residuals vs Leverage vs Cook's distance plots (do not put other 3 plots you obtain for R). The final diagnostic plot obtained after removing all outliers should also be included.

Answer :

This graph was also used to remove outliers.

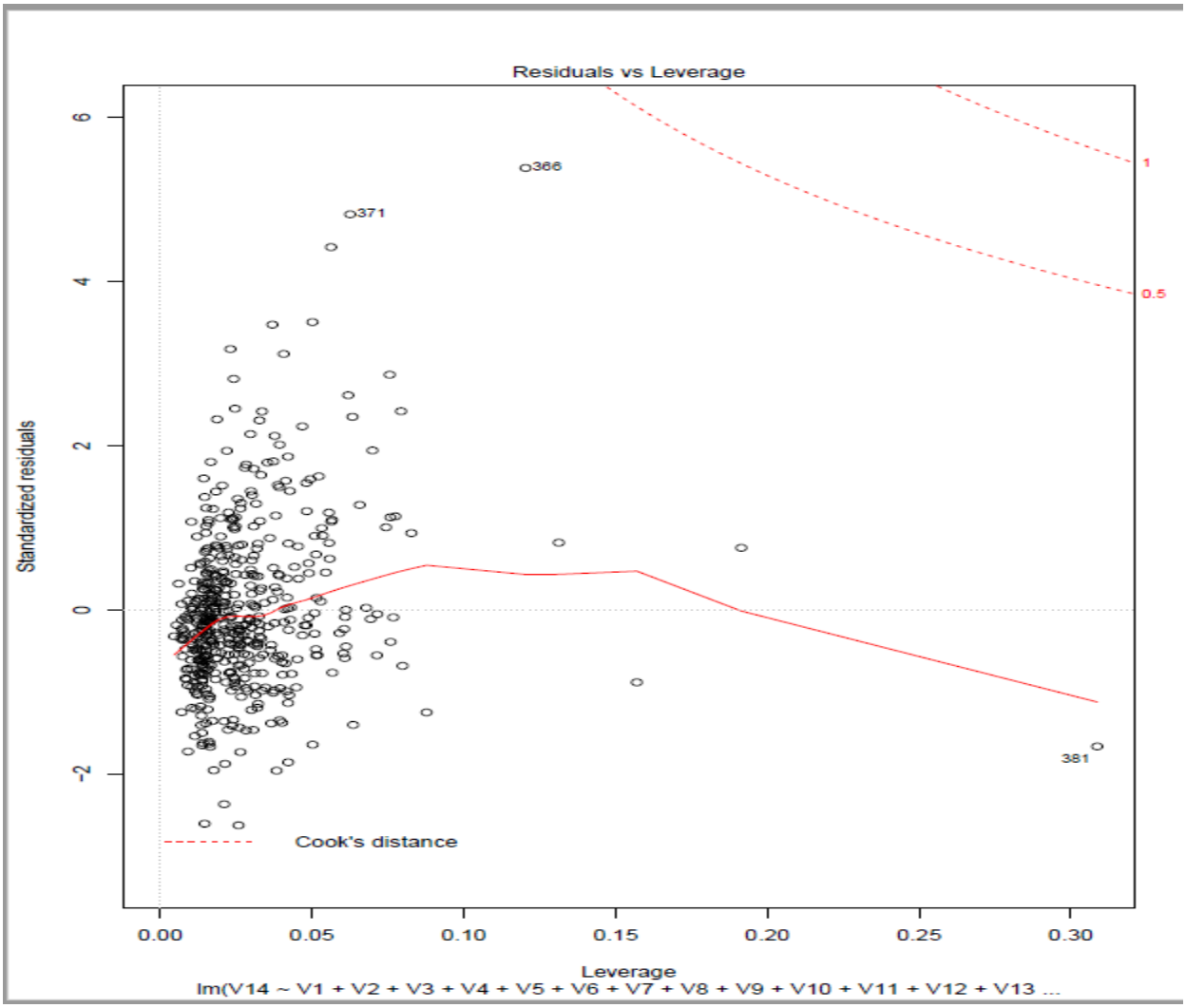
About 99% of the sampled values of a standard normal random variable are in the range [−3, 3].

I removed some of the points which are outside these ranges.



Page 2:
3. Diagnostic plots used for identification of outliers. Please only include the Standard residuals vs Leverage vs Cook's distance plots (do not put other 3 plots you obtain for R**. The final diagnostic plot obtained after removing all outliers should also be included.**

Answer: The final plot after removing all the residuals.



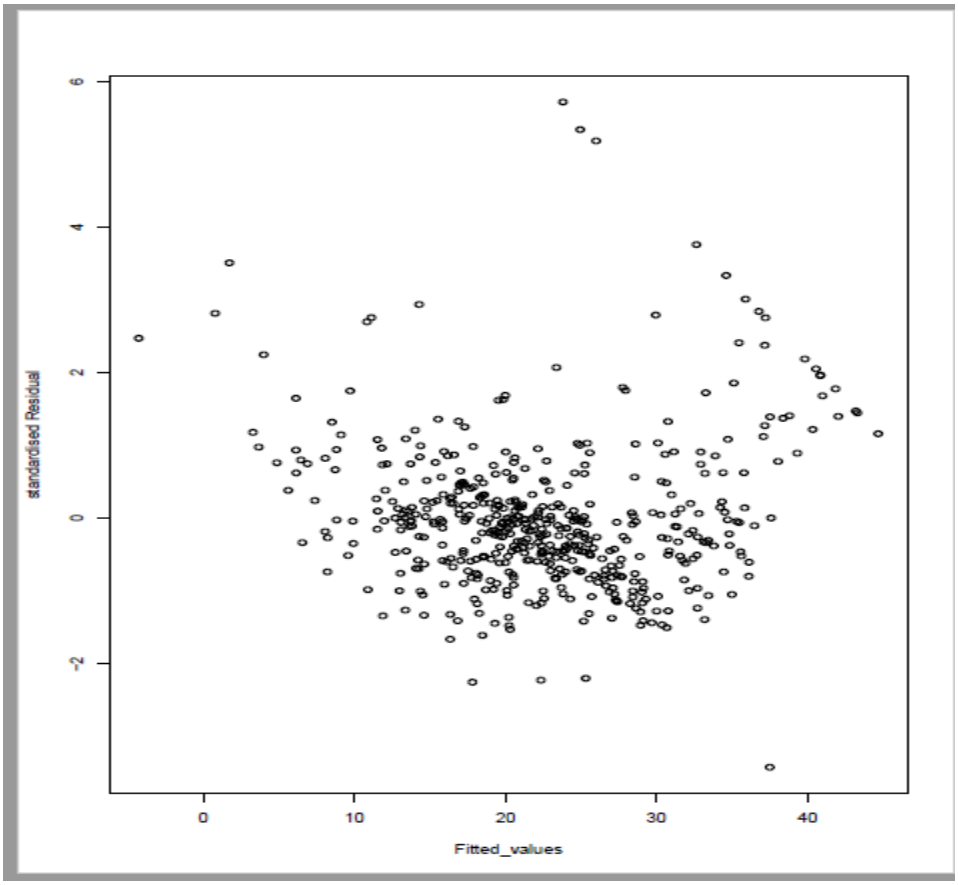
Page 3:

4. Plot of Standardized residuals vs Fitted values for the linear regression model **obtained without any transforms (like removing outliers or transforming dependent variables)**

Answer:

This is plot when the outliers are not removed and there is no transformation in the dependent variables.

We see that standardized residuals have some pattern with fitted values. There is some pattern in the below graph.

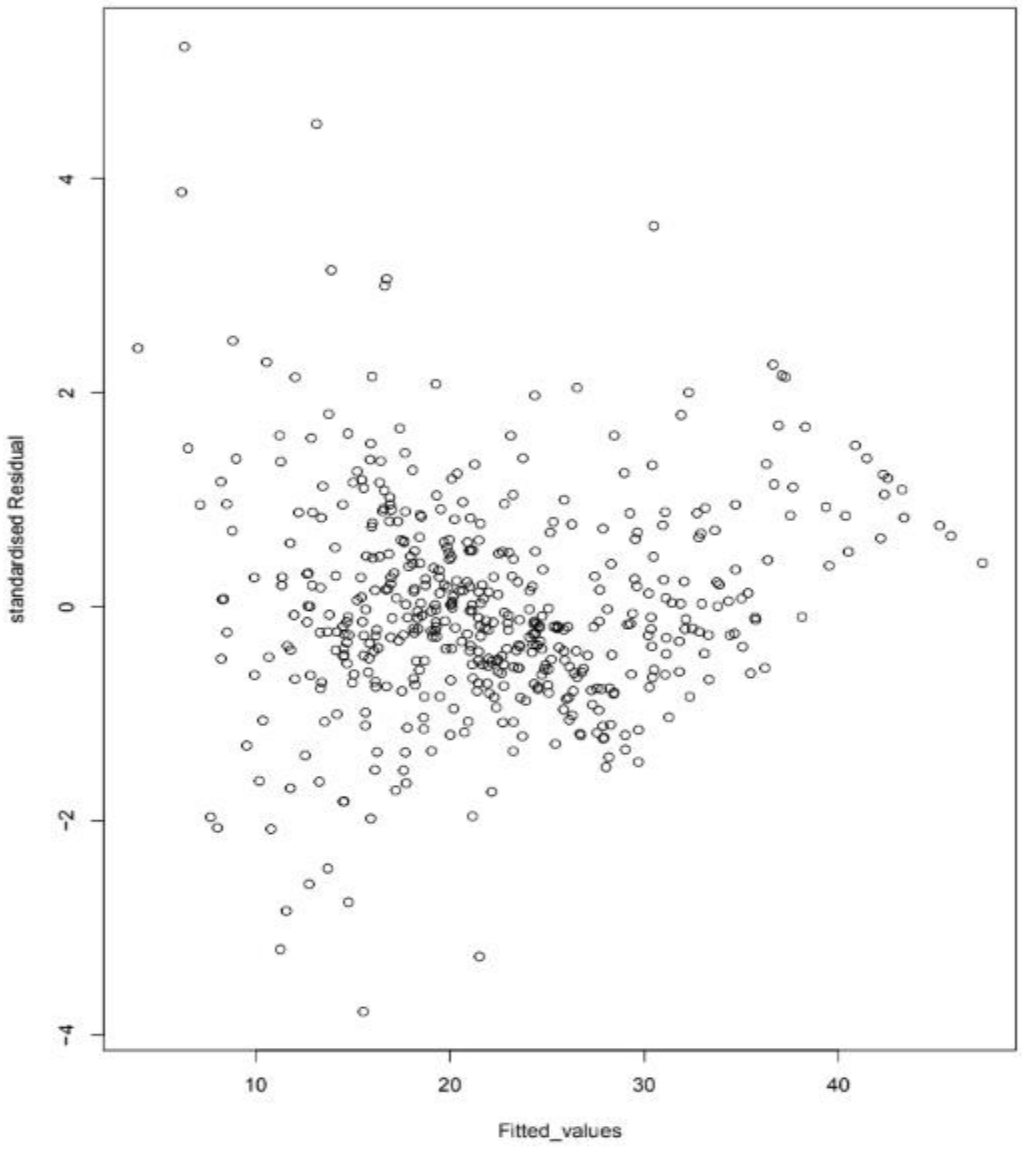


Page3:

5. Plot of Standardized residuals vs Fitted values for **the final linear regression model obtained after removing all outliers and transforming the dependent variable**

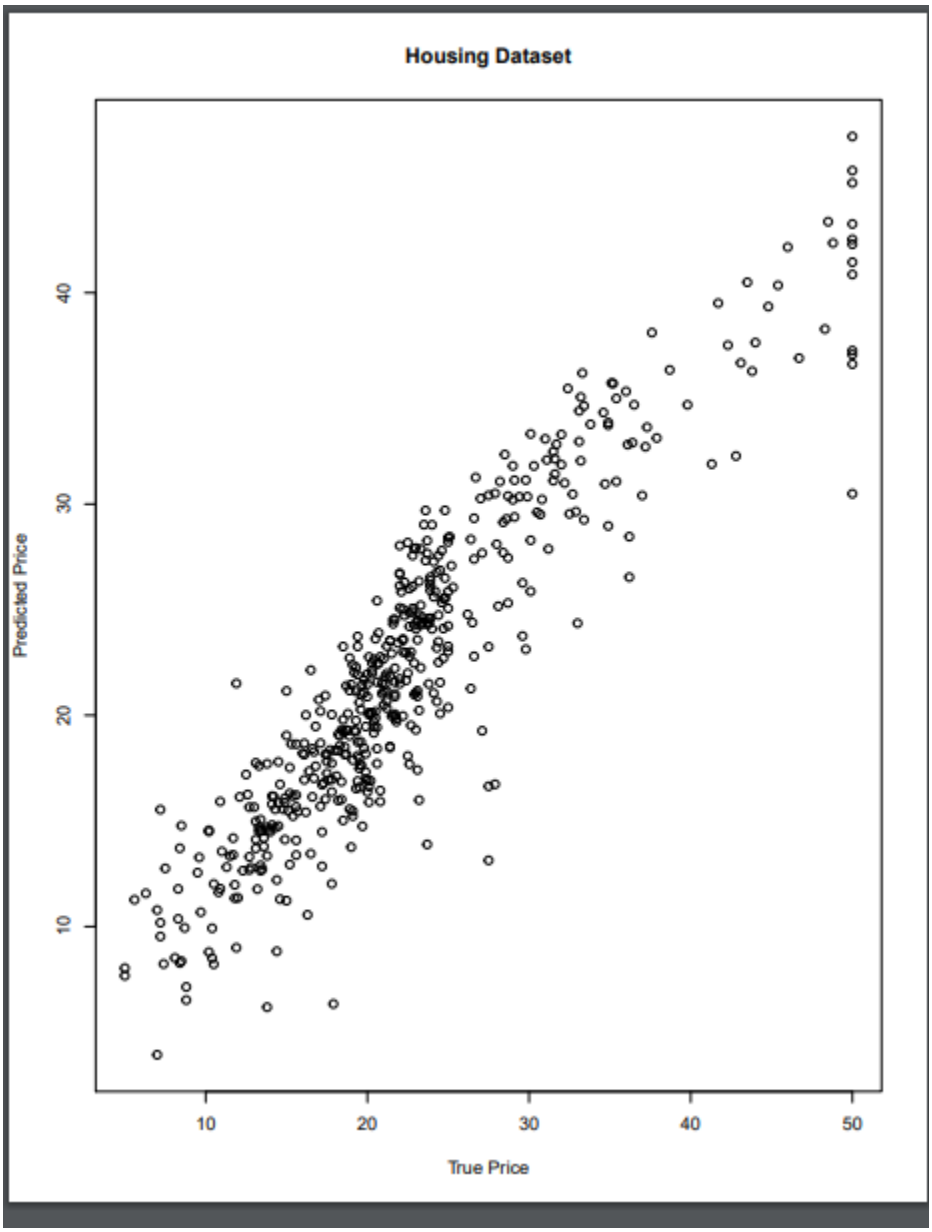
Compare the two plots. What do you observe ?

Answer: The standard residuals does not have any pattern with fitted values. They are almost and complete random.



Page 4:
7. Final plot of Fitted house price vs True house price. What do you observe?

Answer: The regression looks somewhat good. The line appears to be approximated linear .The predicted price is somewhat near to true price.



Page 5:
8. 1 page Code screenshot. It should include code for Linear regression, Box-Cox transformation and how you used the parameter value to transform the dependent variable.

Applying Linear regression

```
library(MASS)
housing <- read.table("~/Desktop/housing.data.txt", quote="\"", comment.char="")
View(housing)
length(housing)
dim(housing)
attach(housing)
summary(housing)
table(is.na(housing))

# Applyting linear Regression

linear_model<-lm(V14~V1+V2+V3+V4+V5+V6+V7+V8+V9+V10+V11+V12+V13,data=housing)
plot(linear_model)

# Plot standardresidual vs Fitted values
st=rstandard(linear_model)
plot(fitted(linear_model),st,xlab = "Fitted_values",ylab = "standardised Residual")


outlier_1<-housing[c(369,373,365), ]

```

```
# Removed first set of outliers


myData<-housing[-c(369,373,365), ]
linear_model_1<-lm(V14~V1+V2+V3+V4+V5+V6+V7+V8+V9+V10+V11+V12+V13,data=myData)
plot(linear_model_1)

outlier_2<-myData[c(366,370,368), ]

# Removed 2nd set of outliers

myData_1<-myData[-c(366,370,368), ]
linear_model_2<-lm(V14~V1+V2+V3+V4+V5+V6+V7+V8+V9+V10+V11+V12+V13,data=myData_1)
plot(linear_model_2)


outlier_3<-myData_1[c(371,368,366), ]

# Removed 3rd set of outliers

myData_2<-myData_1[-c(371,368,366), ]
linear_model_3<-lm(V14~V1+V2+V3+V4+V5+V6+V7+V8+V9+V10+V11+V12+V13,data=myData_2)
plot(linear_model_3)

```

Box-Cox transformation and how you used the parameter value to transform the dependent variable.

```
46
47  # combining all the outliers
48
49  outlier<-rbind(outlier_1,outlier_2,outlier_3)
50  rownames(outlier) <- NULL
51
52  # Box Cox Transformation after removing outliers
53
54  bc=boxcox(linear_model_3,lambda=seq(-3,3))
55
56  # maximum value of lambda where log likelihood is maximum
57
58  b_lam=bc$x[bc$y==max(bc$y)]
59
60
61
62  #After box transformation apply linear regression
63
64  bx<-lm((myData_2$V14)^b_lam ~ myData_2$V1+myData_2$V2+myData_2$V3+myData_2$V4+myData_2$V5+myData_
65
66  #standard residual values vs fitted values after transformation
67
68  st=rstandard(bx)
69  plot(fitted(bx)^(1/(b_lam)),st,xlab = "Fitted_values",ylab = "standardised Residual")
84:8  (Top Level) ÷                                                                    R Script
```

```
72  # Plot True price vs Predicted price
73
74  plot(myData_2$V14,(fitted(bx))^(1/(b_lam)),xlab="True Price",ylab = "Predicted Price",main="Housing Dataset")
75
76
77
78  #Points (outliers)refering to the original Dataset
79  k<-c()
80  for (i in 1:dim(outlier)[1]) {
81    for (j in 1:dim(housing)[1]) {
82      if (sum(outlier[i,]==housing[j,])==14) {
83            k<-c(k,j)
84        }
85      }
86  }
87
88
```