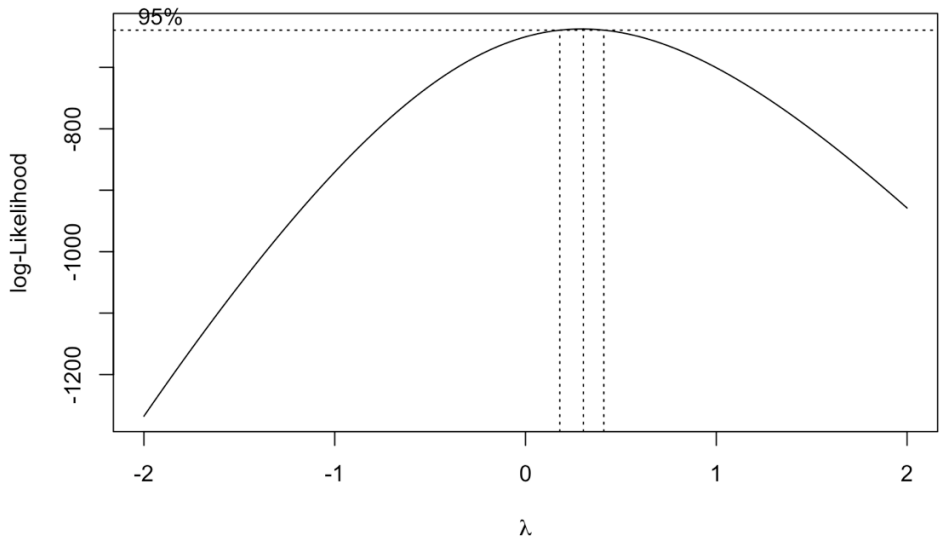
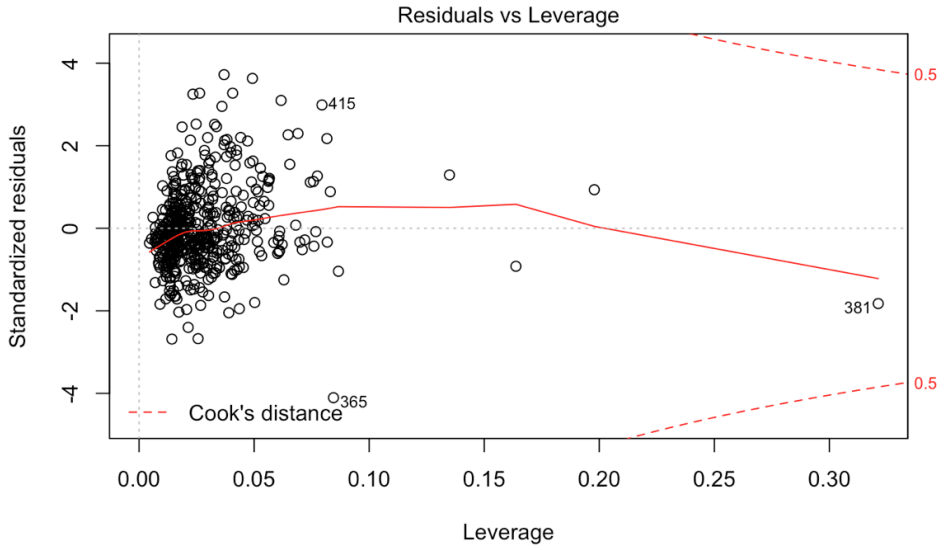
CS498 HW6

1. Outlier points removed: 373,369,405,413,373,368,372,370,366,371
2. Box-Cox Transformation:

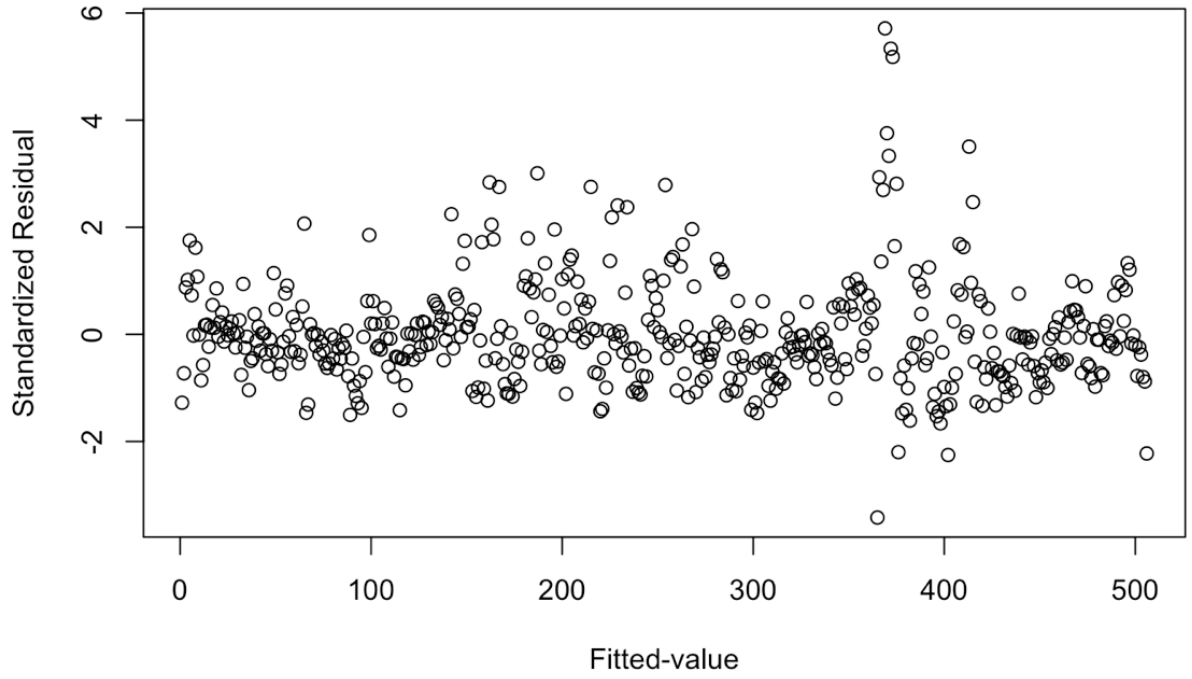


The best value of parameter is 0.3030303

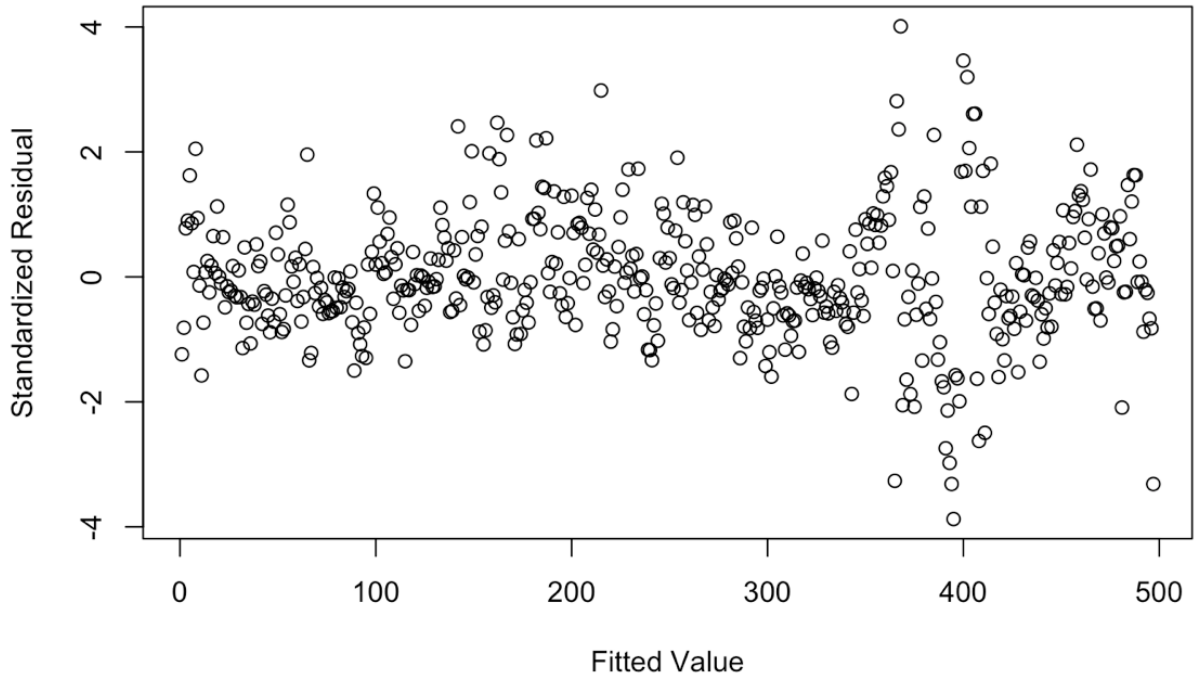
3. Diagnostic plots used for identification of outliers:



4. Standardized residuals vs Fitted values for the linear regression model obtained without any transforms:

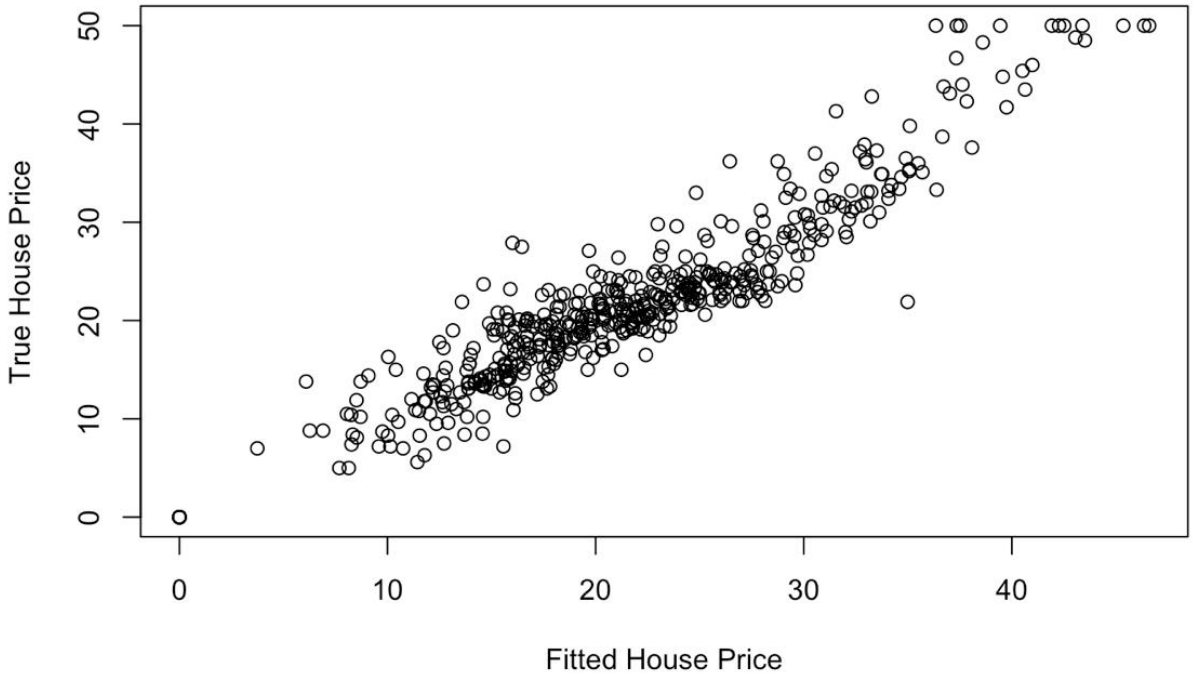


5. Standardized residuals vs Fitted values for the final linear regression model obtained after removing all outliers and transforming the dependent variable:



6. By comparing the two plot, I observed that after removing the outliers with very high standardized residuals, the standardized residuals (as shown in plot for question 5) become more evenly distributed around 0.

7. Final plot of Fitted house price vs True house price:



Observation:

The Fitted House Price is predicted quite well, as it is shown in a linear relationship with True House Price with a coefficient of approximately 1.

8. Code screenshot:
    a. Linear regression:

```
#plots for question 1
```{r}
oriline.lm = lm( V14 ~ V1 + V2 + V3 + V4 + V5 + V6 + V7 + V8 + V9 + V10 + V11 + V12 + V13 ,data = data_ori)
plot(oriline.lm)
#abline(oriline.lm)
a = rstandard(oriline.lm)
plot(a)
a = rstandard(oriline.lm)
plot(a, xlab = "Fitted-value", ylab = "Standardized Residual")
```
```

    b. Box-cox transformation:

```
#question 3 the boxcox
```{r}
#install.packages("MASS")
library(MASS)
bc = boxcox(oriline.lm)
best.lam = bc$x[which(bc$y == max(bc$y))]
best.lam
fullmodel.inv = lm((V14)^(best.lam)~ V1 + V2 + V3 + V4 + V5 + V6 + V7 + V8 + V9 + V10 + V11 + V12 + V13 ,data = data_new)
plot(fullmodel.inv)
a = rstandard(fullmodel.inv)
plot(a, xlab = "Fitted Value", ylab = "Standardized Residual")
#fullmodel.inv$fitted.values^(1/best.lam)
```
```

    c. Transformation of dependent variable:

```
#question 3 the boxcox
```{r}
#install.packages("MASS")
library(MASS)
bc = boxcox(oriline.lm)
best.lam = bc$x[which(bc$y == max(bc$y))]
best.lam
fullmodel.inv = lm((V14)^(best.lam)~ V1 + V2 + V3 + V4 + V5 + V6 + V7 + V8 + V9 + V10 + V11 + V12 + V13 ,data = data_new)
plot(fullmodel.inv)
a = rstandard(fullmodel.inv)
plot(a, xlab = "Fitted Value", ylab = "Standardized Residual")
#fullmodel.inv$fitted.values^(1/best.lam)
```
```

9. More Code:

```
#last part, plot
```{r}
pre_list = numeric(506)
act_list = numeric(506)
#fullmodel.inv$fitted.values
for(i in 1:496){
  pre = fullmodel.inv$fitted.values[i]^(1/best.lam)
  actually = data_new[i,14]
  pre_list[i] = pre
  act_list[i] = actually
}
as.data.frame(pre_list)
as.data.frame(act_list)
data_ori$V15 = pre_list
data_ori$V16 = act_list
plot(data_ori$V15,data_ori$V16,xlab= "Fitted House Price", ylab = "True House Price")
```
```