CS 498 AML HW6 REPORT

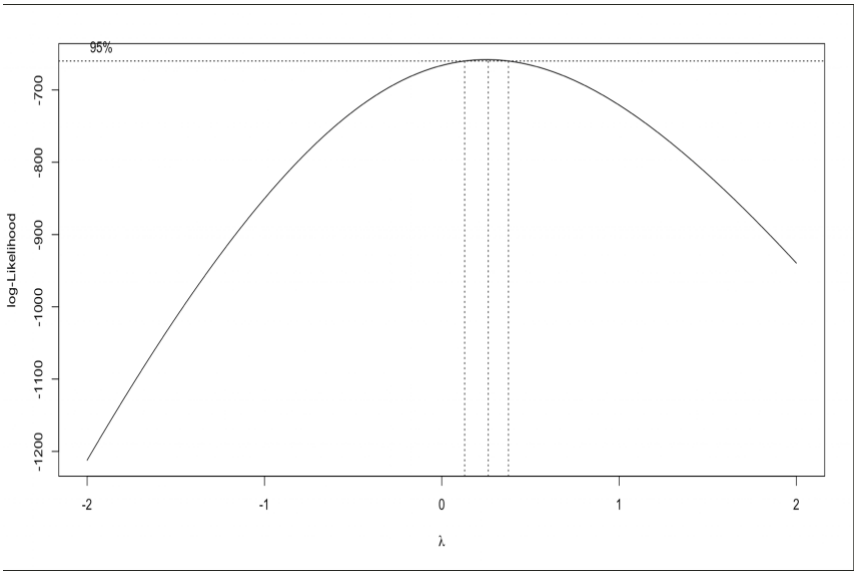
Pengyu Cheng pcheng11

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

    - The points I removed are: 369, 372, 373, 370, 366, 365, 381, 419, 406, 411

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?

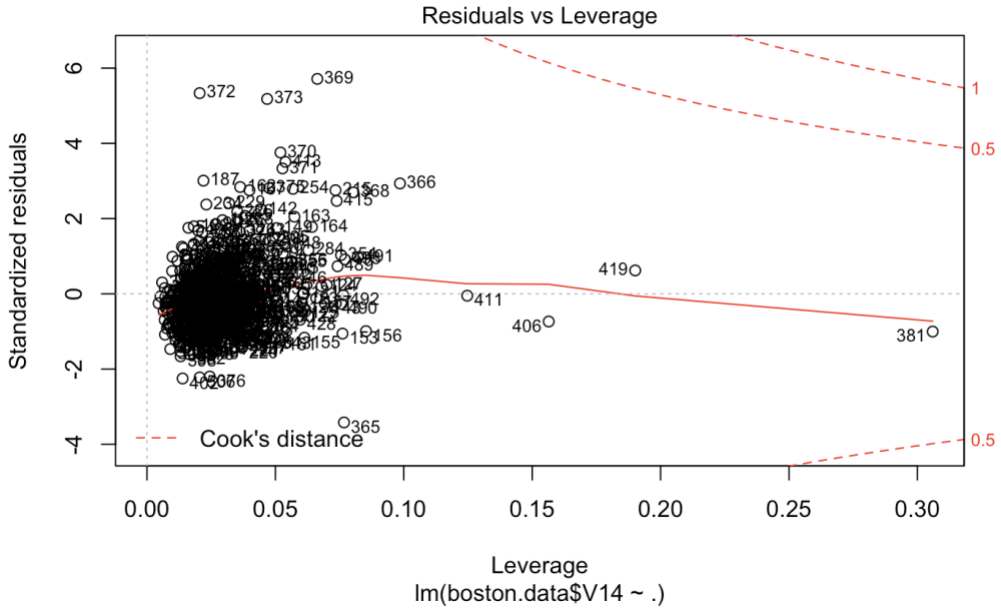
    - The best value of parameter I got is 0.2626263
    - Box-Cox transformation curve:



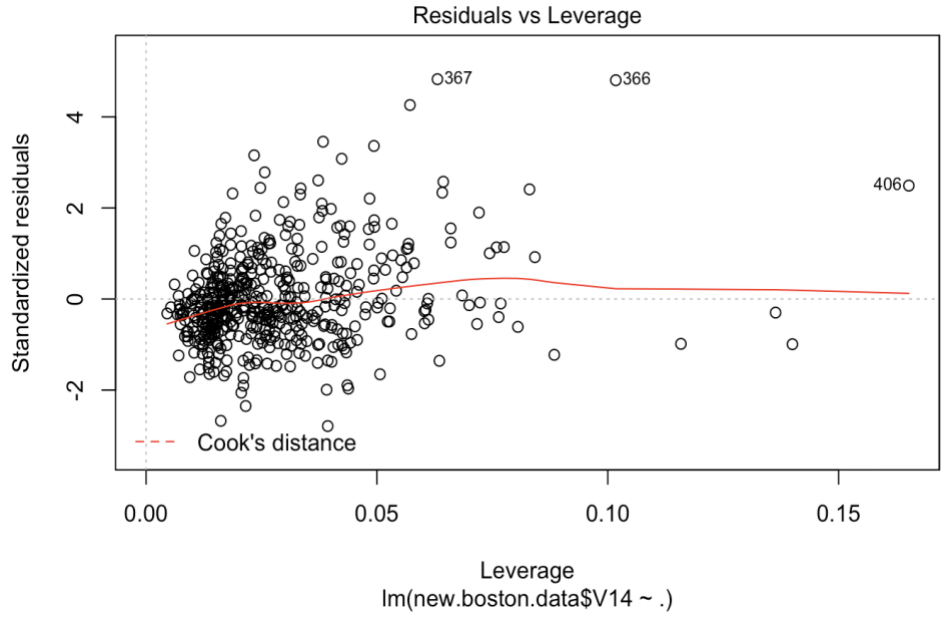
CS 498 AML HW6 REPORT
Pengyu Cheng pcheng11

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.

- Standard residuals vs Leverage vs Cook's distance plots without removing outliers
(I use "plot(model, which = 5, id.n=num.points)" to specify all the index of the points)

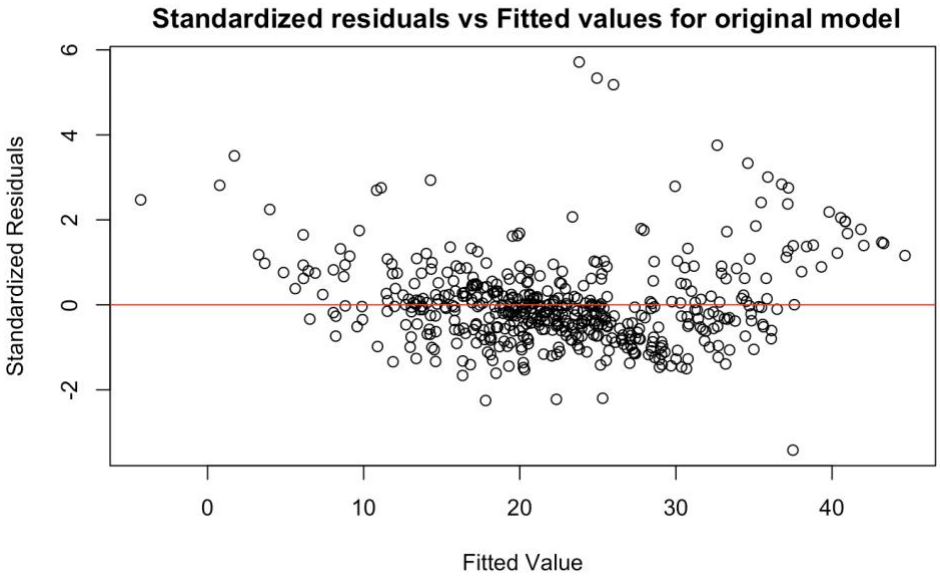


- Standard residuals vs Leverage vs Cook's distance plots after removing outliers

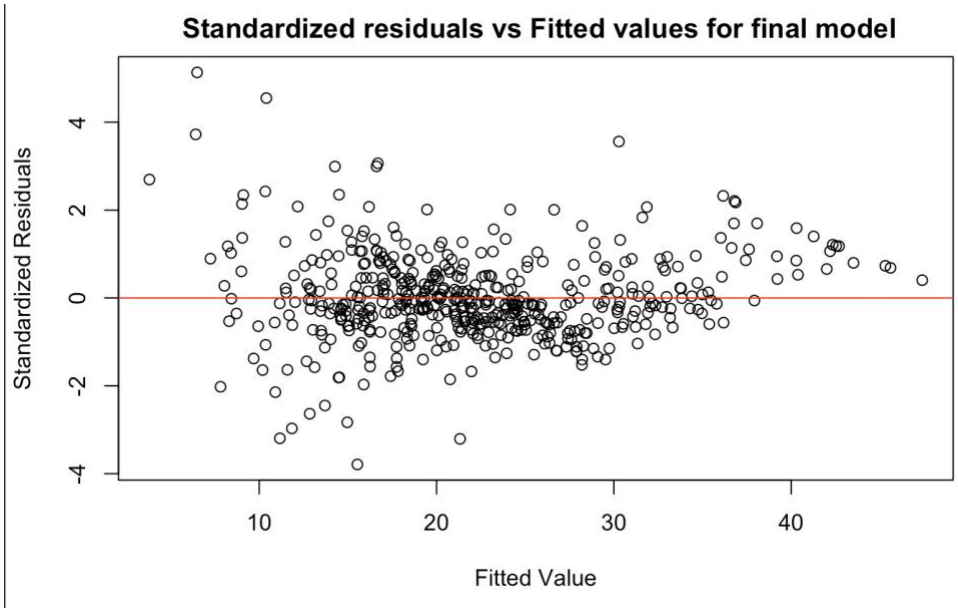


CS 498 AML HW6 REPORT
Pengyu Cheng pcheng11

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



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

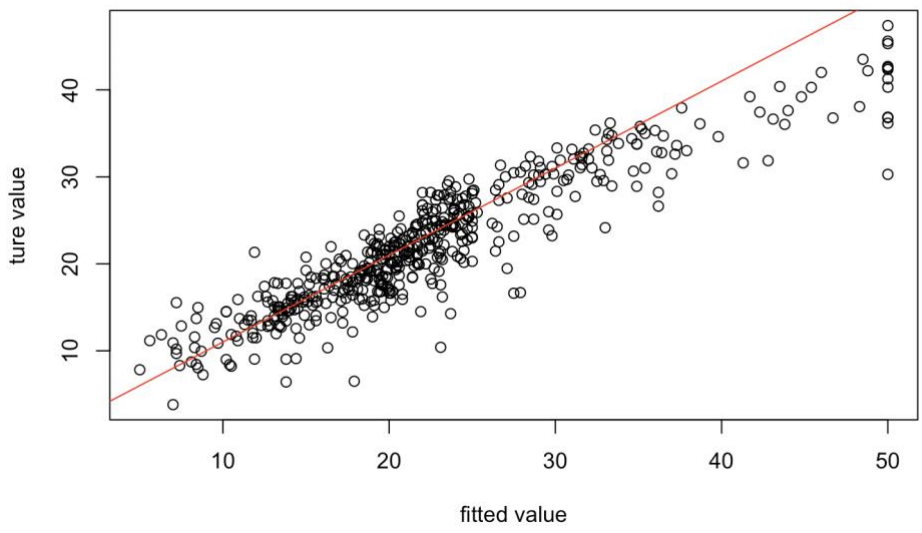


6. Compare the two plots. What do you observe?

   We observed that with the transformation and removing outliers, we get a better plot in the sense that the standardized residuals are more aligned to 0 and the residual seems to be uncorrelated to the predicted (which is a good sign) compared to the plot for the original model which has some non-linear relationship between the fitted values and the standardized residuals which suggests non-linear transformation will be helpful.

CS 498 AML HW6 REPORT
Pengyu Cheng pcheng11

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



The relationship between the true house price and the predicted house price is mostly linear and follows the line y = x well although there exist some inaccuracies.

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.

```
### Question a): Regress house price (variable 14) against all others, and use leverage, Cook's
distance, and standardized residuals to find possible outliers. Produce a diagnostic plot that
allows you to identify possible outliers
```{r}
library(data.table)
boston.data = fread('https://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.data
')
num.points = nrow(boston.data)
model = lm(boston.data$V14~., data = boston.data)
#plot all points id for identification
plot(model, which = 5, id.n=num.points)
```

Question b): Remove all the points you suspect as outliers, and compute a new regression.
Produce a diagnostic plot that allows you to identify possible outliers.
```{r}
new.boston.data = boston.data[-c(369, 372, 373, 370, 366, 365, 381, 419, 406, 411),]
new.model = lm(new.boston.data$V14~., data=new.boston.data)
plot(new.model, which = 5)
```

Question c): Apply a Box-Cox transformation to the dependent variable – what is the best value
of the parameter?
```{r}
library(MASS)
boxcox.transform = boxcox(new.model)
best.lambda = boxcox.transform$x[which(boxcox.transform$y == max(boxcox.transform$y))]
```
The best parameter of λ is `r best.lambda`

Question d): Now transform the dependent variable, build a linear regression, and check the
standardized residuals. If they look acceptable, produce a plot of fitted
house price against true house price.
```{r}
transform.model = lm(((new.boston.data$V14^best.lambda - 1)/best.lambda) ~., data=new.boston.data)
plot(transform.model, which = 5)
plot(new.boston.data$V14, (fitted.values(transform.model) * best.lambda + 1)^(1/best.lambda), xlab
= "fitted value", ylab = "ture value")
abline(1,1,col='red')
```
Plot of Standardized residuals vs Fitted values for the final linear regression model obtained
before removing all outliers and transforming the dependent variable
```{r}
std.res = rstandard(model)
plot(model$fitted.values, std.res, ylab="Standardized Residuals", xlab="Fitted Value", main="
Standardized residuals vs Fitted values for original model")
abline(0, 0, col='red')
```
Plot of Standardized residuals vs Fitted values for the final linear regression model obtained
after removing all outliers and transforming the dependent variable
```{r}
transform.std.res = rstandard(transform.model)
#scale back to original range
transform.fitted.values = (fitted.values(transform.model) * best.lambda + 1)^(1/best.lambda)
plot(transform.fitted.values, transform.std.res, ylab="Standardized Residuals", xlab="Fitted
Value", main=" Standardized residuals vs Fitted values for final model")
abline(0, 0, col='red')
```
```