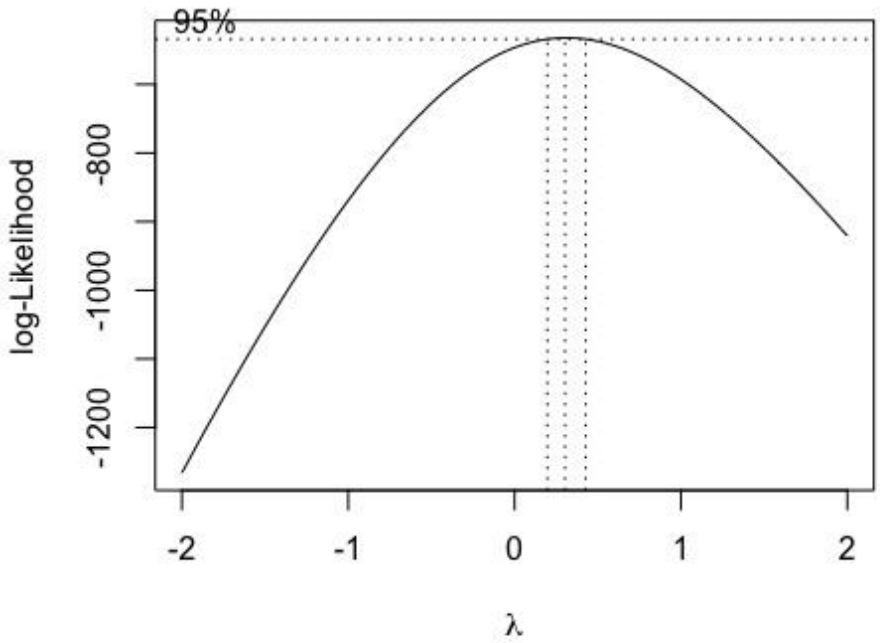
CS 498 AML HW6

Yidi Yang(yyang160)
Huiyun Wu(hwu63)

1. Removed Row Numbers and BoxCox Curve.

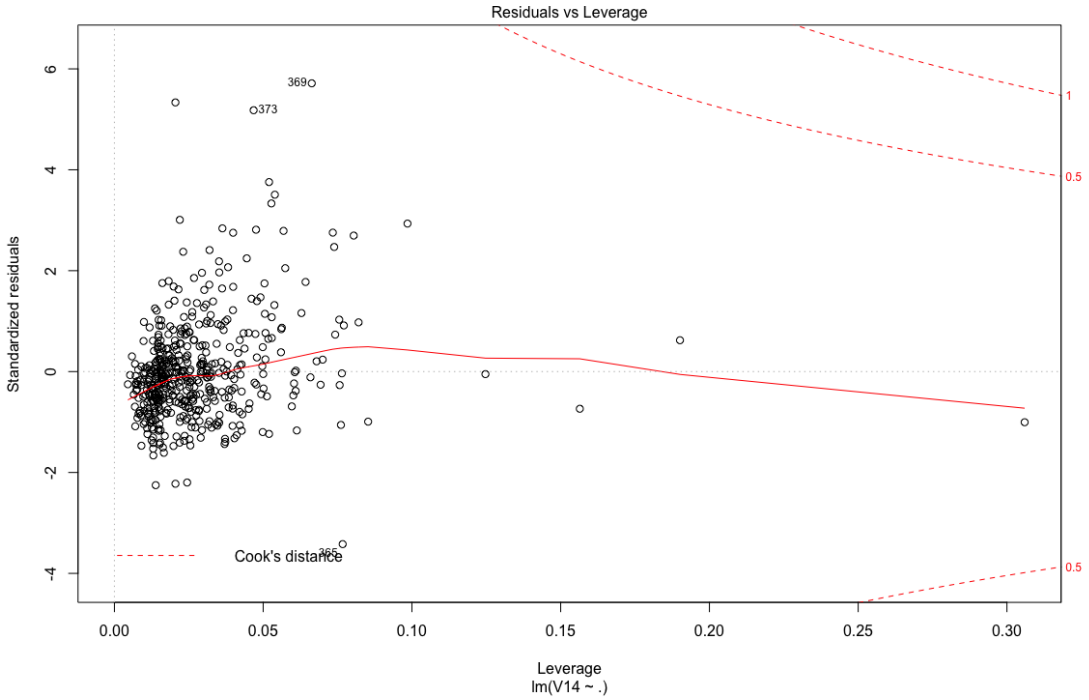
    a. row numbers removed : 365, 366, 369, 372, 373, 368, 370, 371, 413

    b. Box-Cox transformation curve

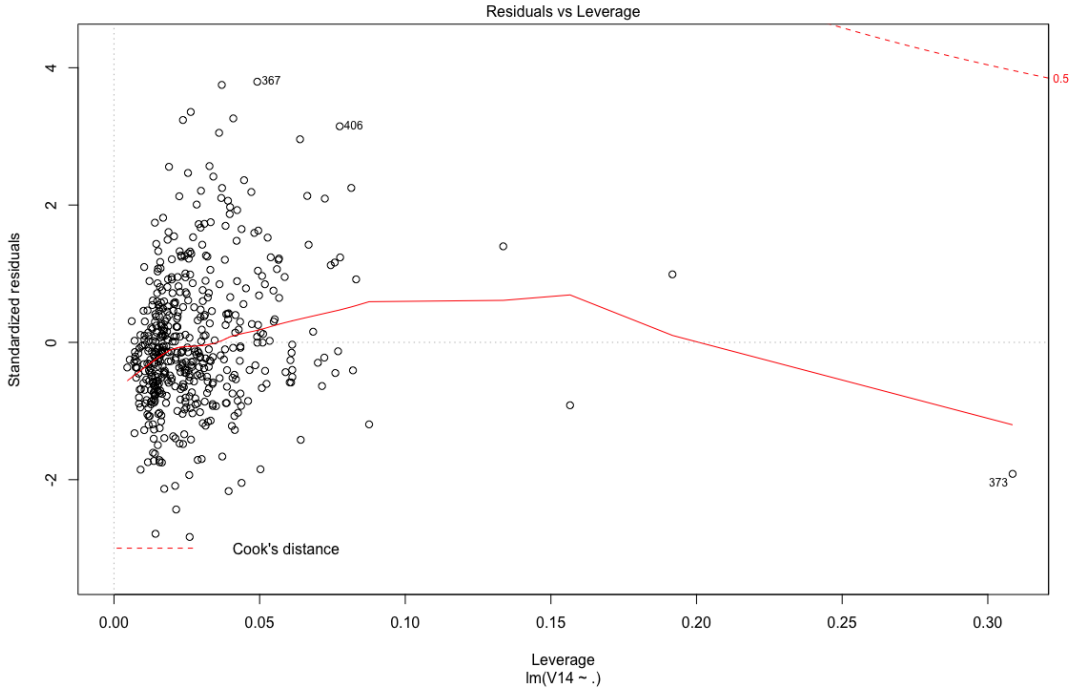


Best value of the parameter: 0.303030303

2. Plot for Identification of Outliers (3 metrics in 1 plot).

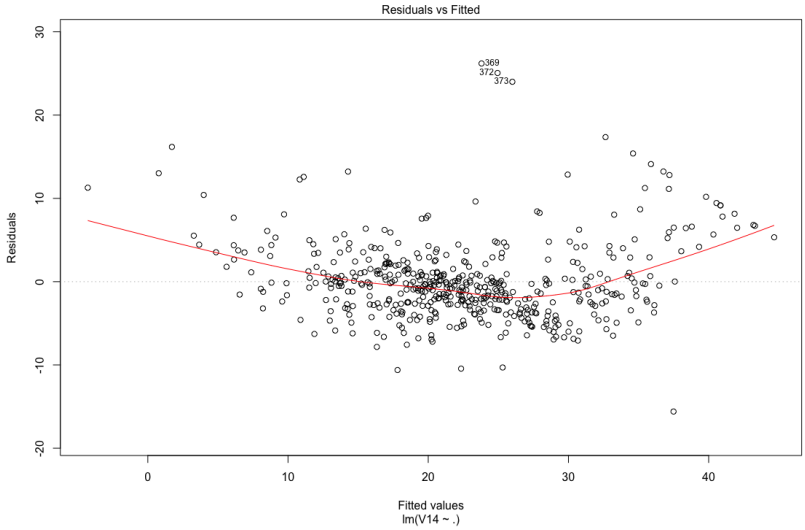


Final diagnostic plot after removing all outliers (but before tranforming y).

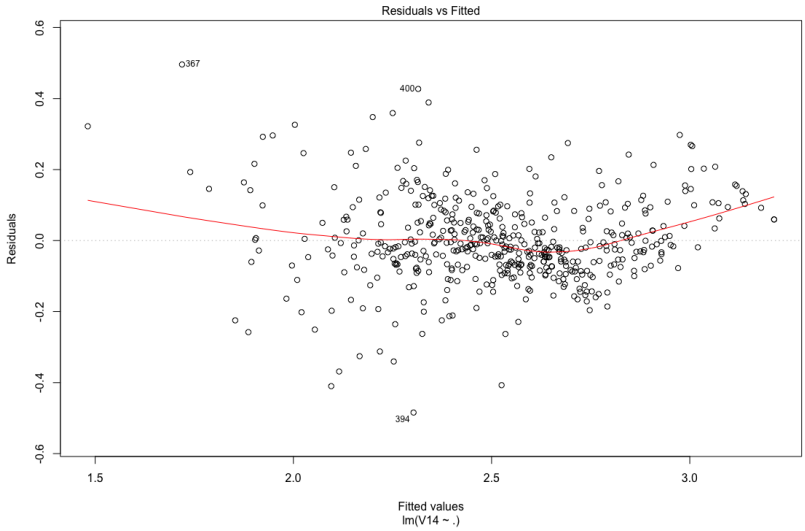


3. Plots of Standardized Residuals VS Fitted Values.

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



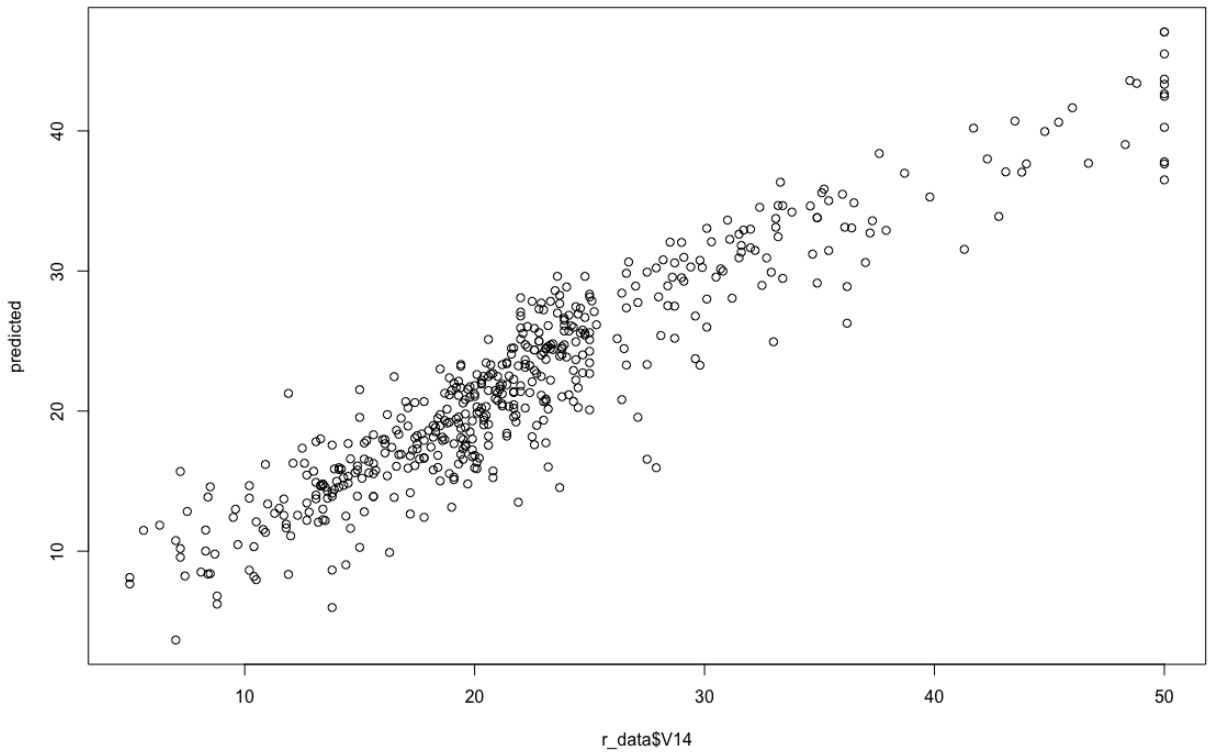
b. Plot of Standardized residuals vs Fitted values for the final linear regression model obtained after removing all outliers and transforming the dependent variable.



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

We observed that the second plot looks more noisy. The residual seems to be uncorrelated to the predicted value, and the mean of the residual seems to be zero, which are both good signs as mentioned on the textbook.

4. Final plot of Fitted house price vs True house price.



The true price values and the predicted ones seem to be linearly correlated, indicating that removing the outliers helps predict the house price more accurately.

5. Code screenshot.

```
library(car)
library(MASS)
library(pracma)
Data <- read.table("~/Documents/AML_hw6/Boston_Housing_data", quote="\"", comment.char="")
linearMod <- lm(V14 ~ ., data=Data)
summary(linearMod)
plot(linearMod)

r_data <- Data[-c(365,366,369,372,373, 368, 370, 371, 413),]
row.names(r_data) <- 1:nrow(r_data)
fit <- lm(V14 ~ ., data= r_data)
summary(fit)
plot(fit)
cd <- cooks.distance(fit)
cd <- data.frame(1:nrow(r_data), cd)
colnames(cd) <- c("idx", "value")
plot(cd$idx ~ cd$value)
text(cd$idx ~ cd$value, labels=idx, data=cd, cex=0.9, font=2)

bc <- boxcox(V14 ~ .,  data=r_data)
lambda <- bc$x[which.max(bc$y)]
print(lambda)

new_data <- r_data
new_data$V14 <- new_data$V14 ** lambda
new_fit <- lm(V14 ~ ., data=new_data)
summary(new_fit)
plot(new_fit)
plot(rstandard(new_fit))
predicted <- fitted(new_fit) ** (1/lambda)

plot(predicted ~ r_data$V14)

```