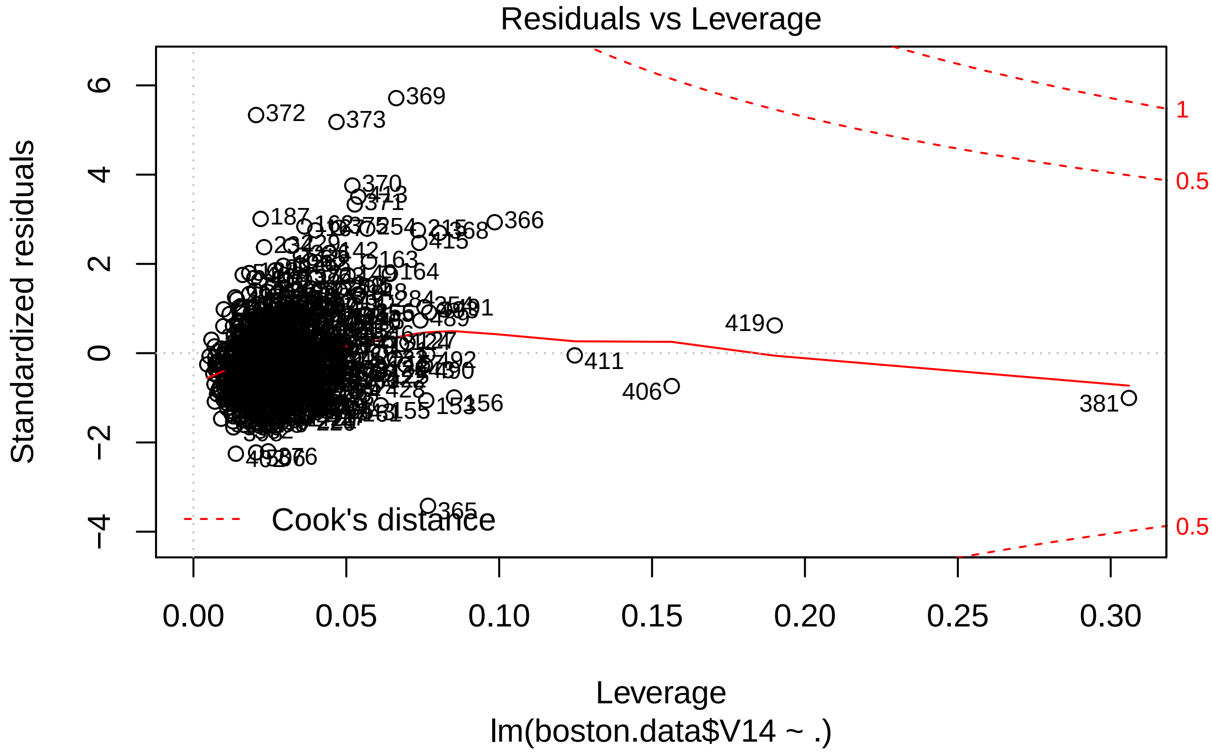
CS498 AML HW6

Pengyu Cheng

10/19/2018

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

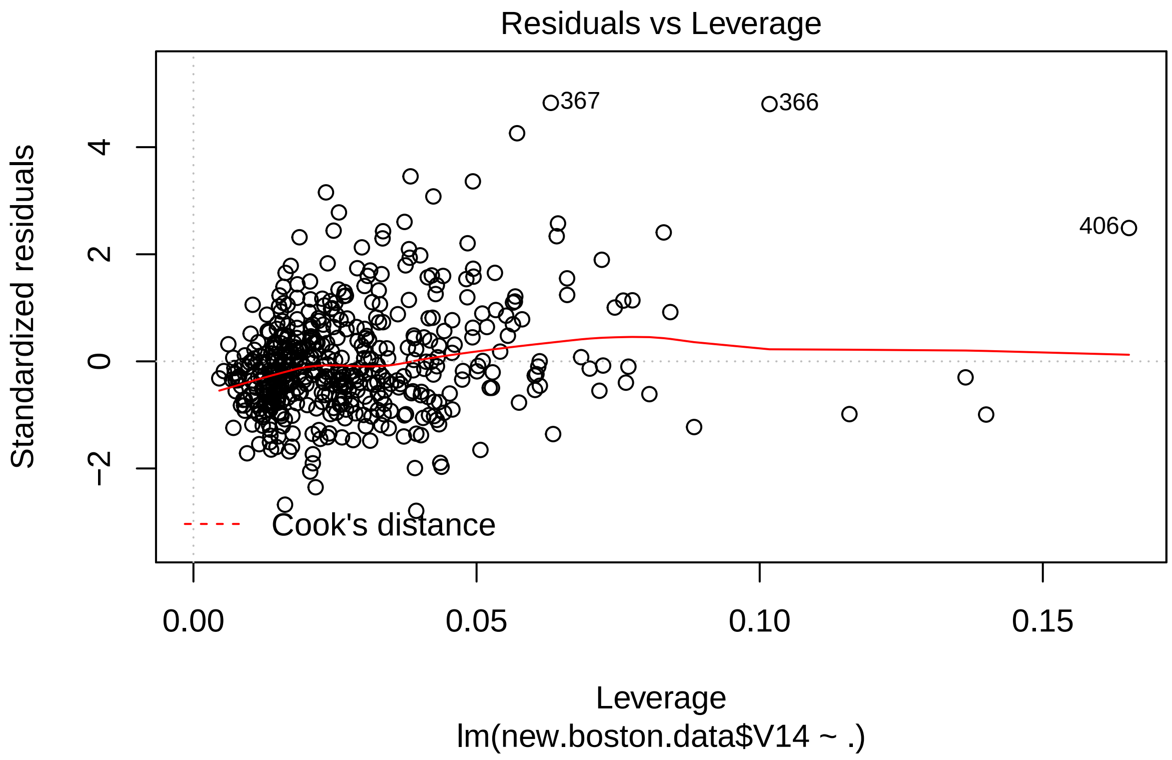
```
library(data.table)
boston.data = fread('https://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.data')
num.points = nrow(boston.data)
model = lm(boston.data$V14~., data = boston.data)
#plot all points id for identification
plot(model, which = 5, id.n=num.points)
```



From the "Residual vs Leverage", we identify point 369, 372, 373 and 365 as possible outliers (standardized residual too large). Other possible outliers are point 381, 419, 406 and 411 (large leverage).

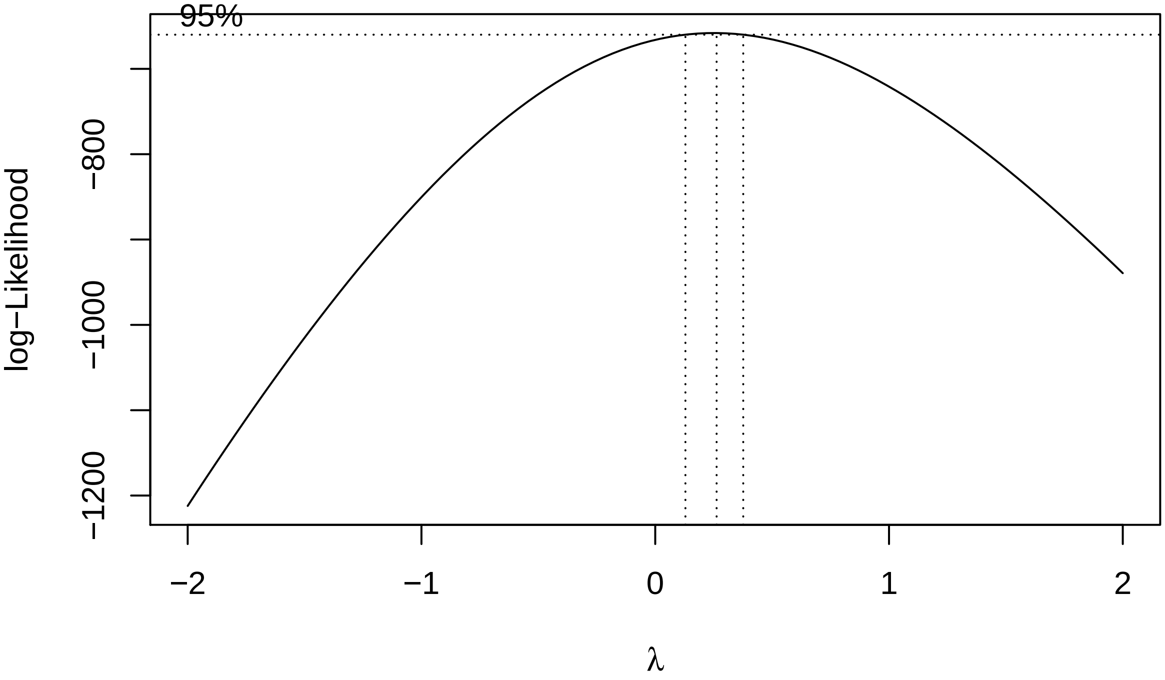
Question b): Remove all the points you suspect as outliers, and compute a new regression. Produce a diagnostic plot that allows you to identify possible outliers.

```
#remove data points 373, 369 365 and 372
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?

```
library(MASS)
boxcox.transform = boxcox(new.model)
```



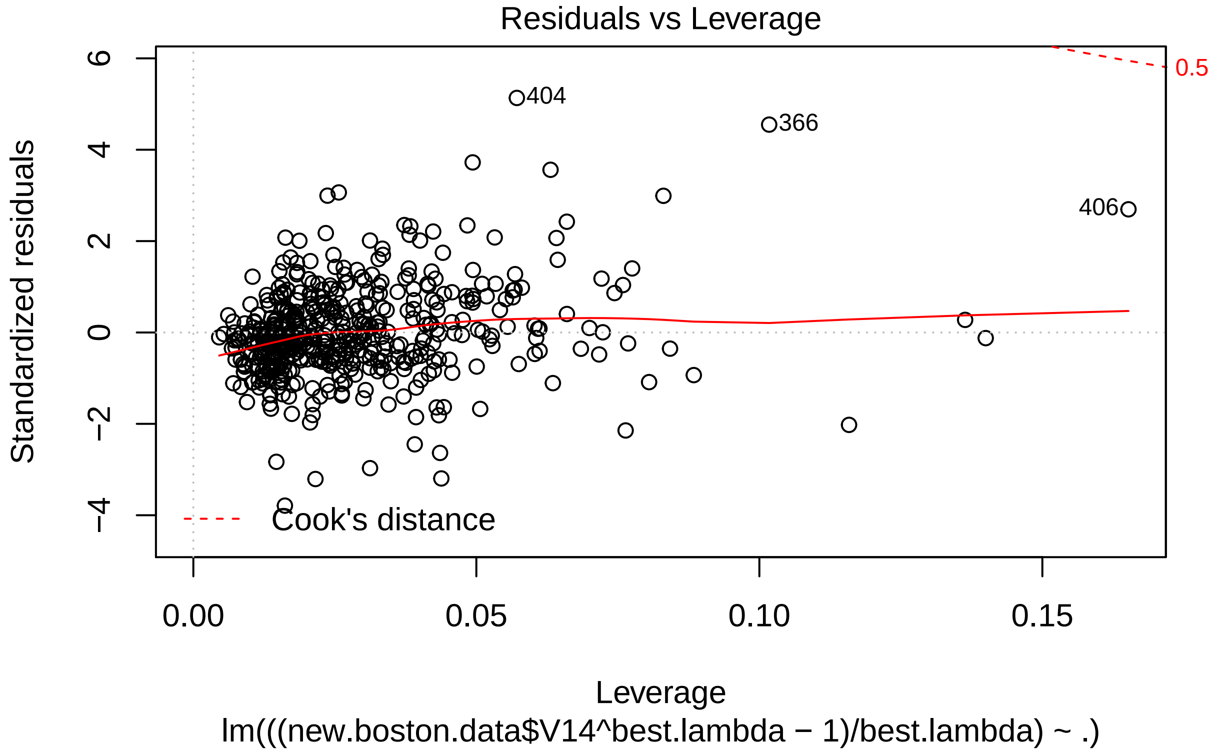
```
best.lambda = boxcox.transform$x[which(boxcox.transform$y == max(boxcox.transform$y))]
```

The best parameter of λ is 0.2626263

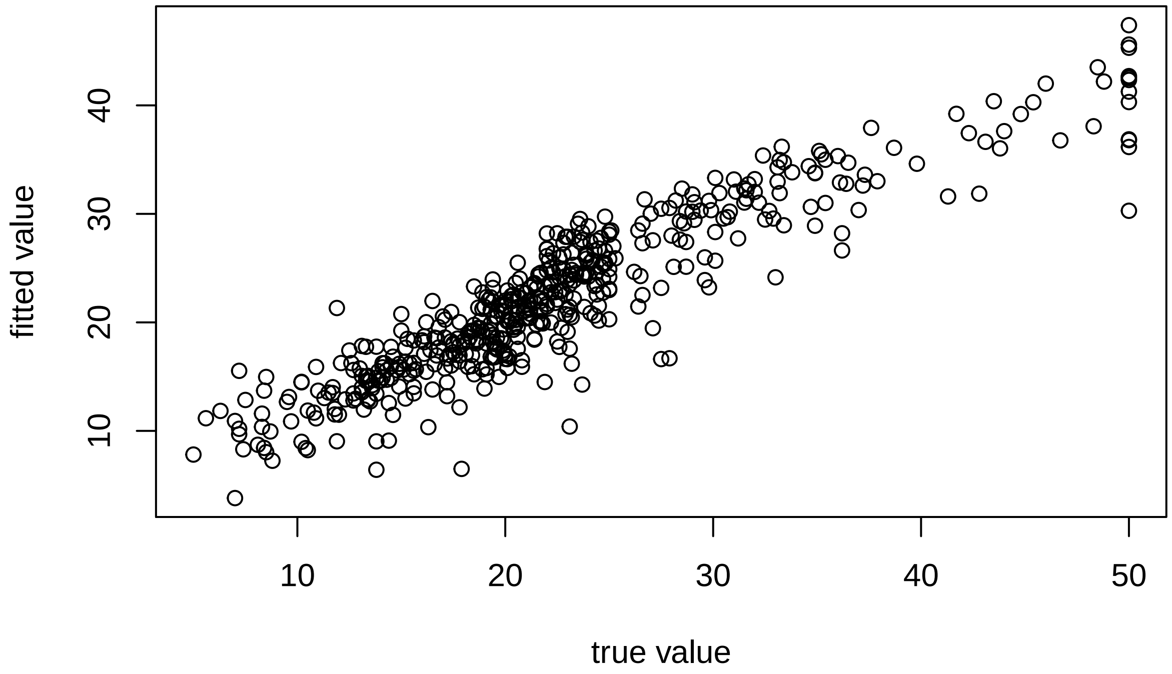
Question d): Now transform the dependent variable, build a linear regression, and check the standardized residuals. If they look acceptable, produce a plot of fitted

house price against true house price.

```
transform.model = lm(((new.boston.data$V14^best.lambda - 1)/best.lambda) ~., data=new.boston.data)
plot(transform.model, which = 5)
```

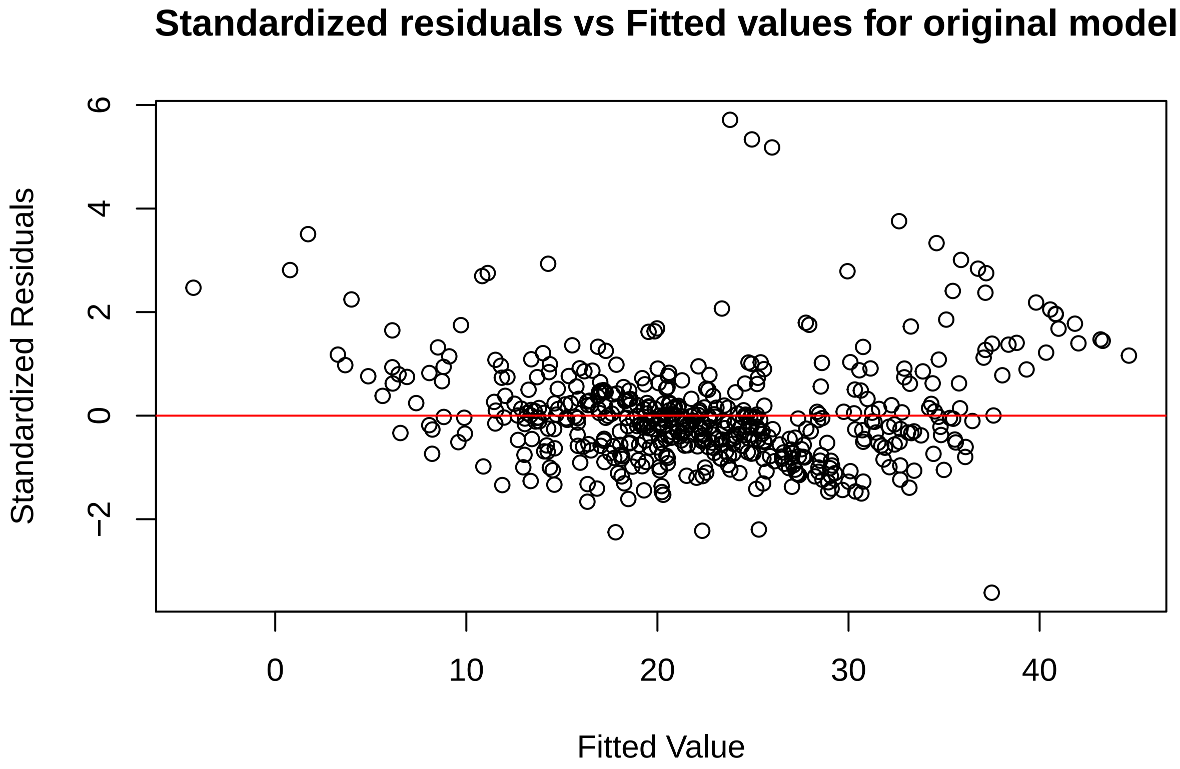


```
plot(new.boston.data$V14, (fitted.values(transform.model) * best.lambda + 1)^(1/best.lambda), xlab = "t
```



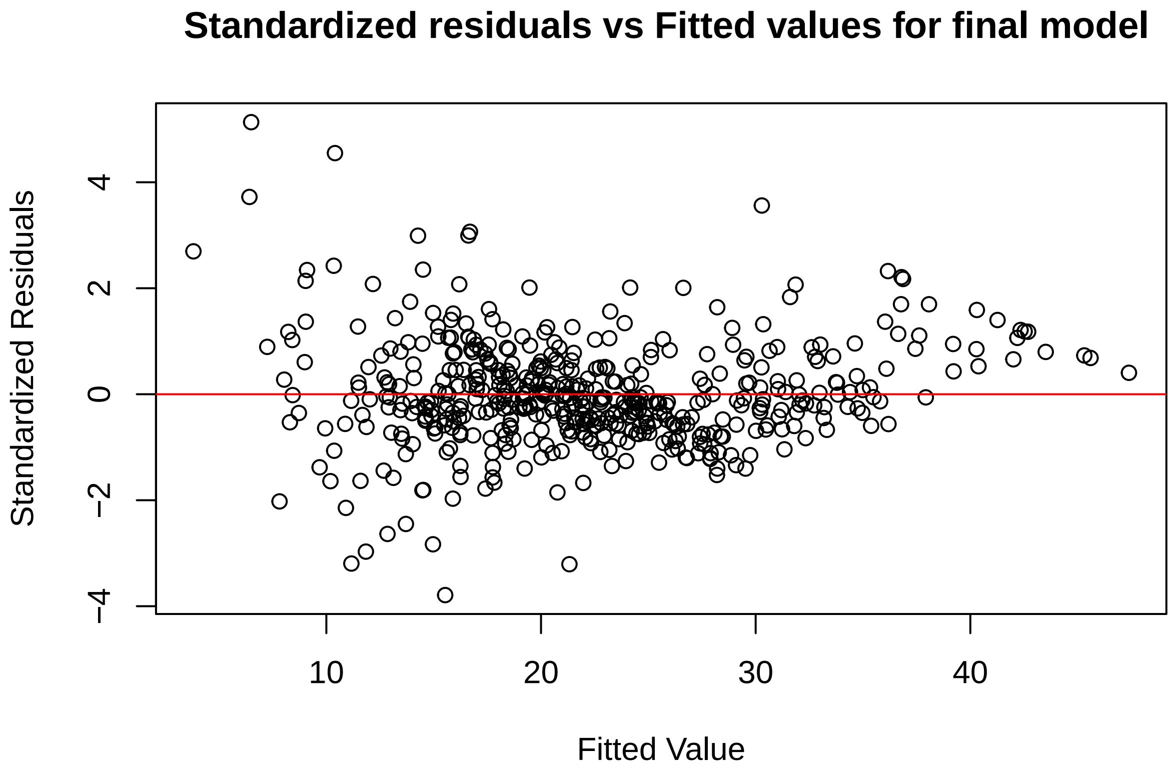
4. Plot of Standardized residuals vs Fitted values for the linear regression model obtained without any transforms (like removing outliers or transforming dependent variables)

```
std.res = rstandard(model)
plot(model$fitted.values, std.res, ylab="Standardized Residuals", xlab="Fitted Value", main=" Standardi
abline(0, 0, col='red')
```



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

```
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", ma
abline(0, 0, col='red')
```



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