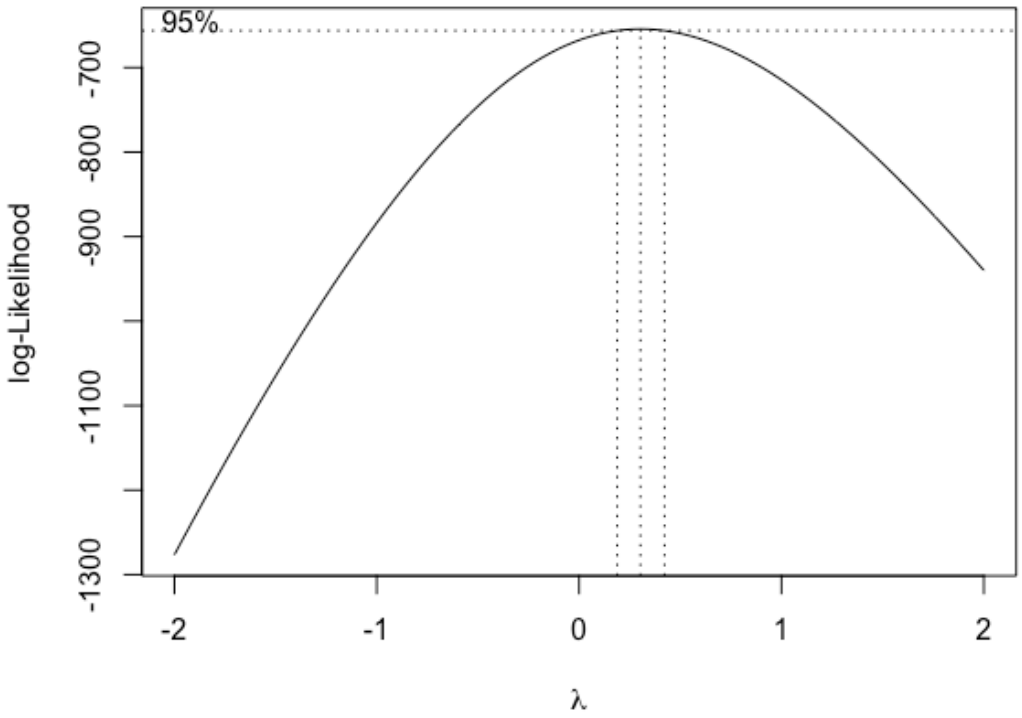
CS498AML_HW6_panz2

Outlier points:

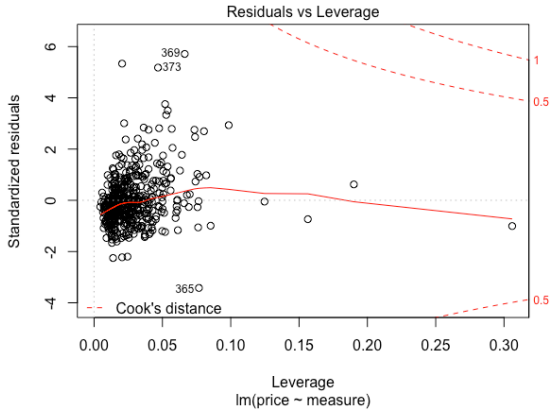
I removed 8 outliers: 366, 368, 369, 370, 371, 372, 373, 413

Box-Cox transformation curve:

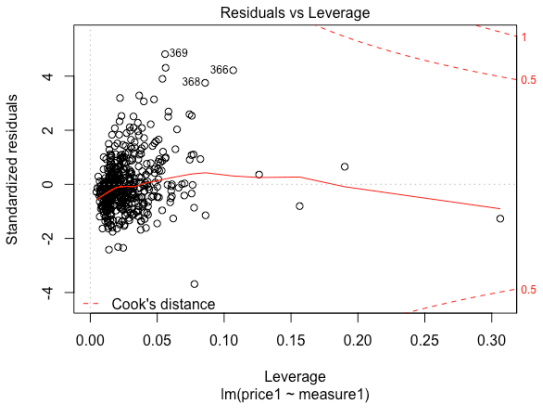


Best value of the parameter: 0.3030303

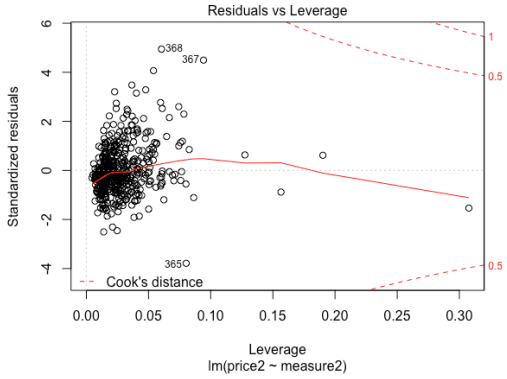
Original data:



After removing 369, 372, 273:

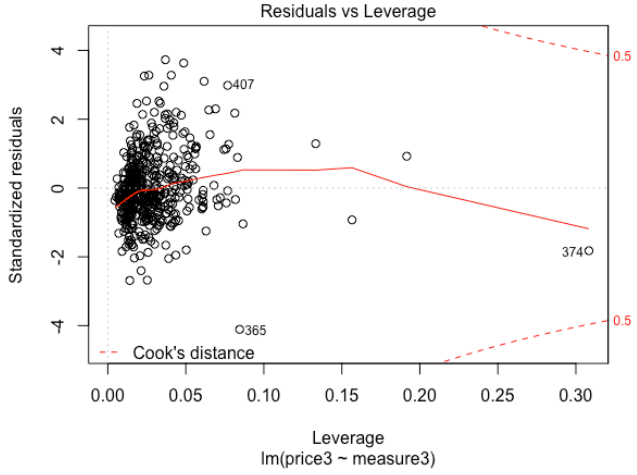


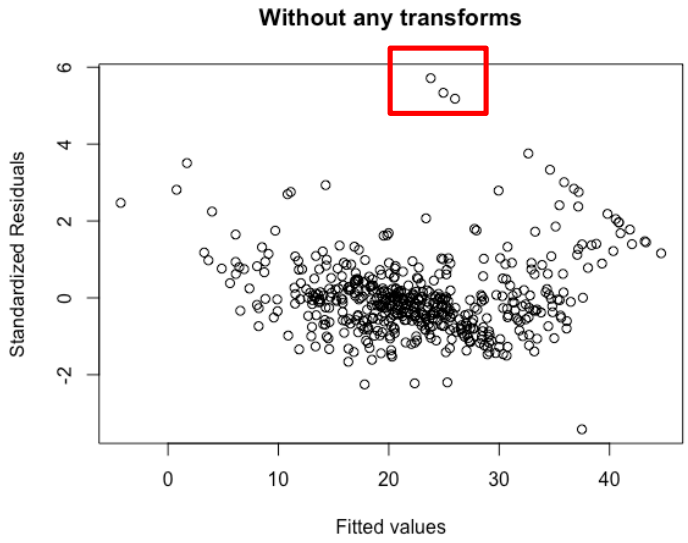
After remove 366, 369 (original 366, 370).



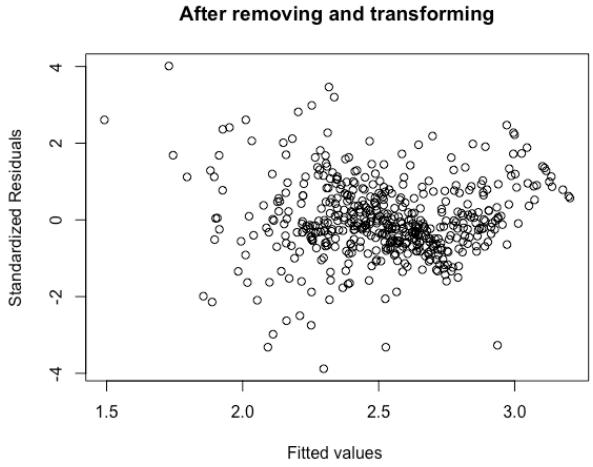
Then, remove 367, 368, 408 (original 368, 371, 413).

After removing all outliers:

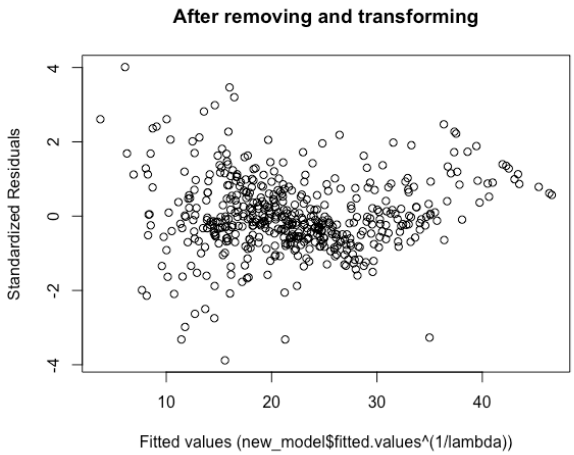




Fitted values of price^lambda:



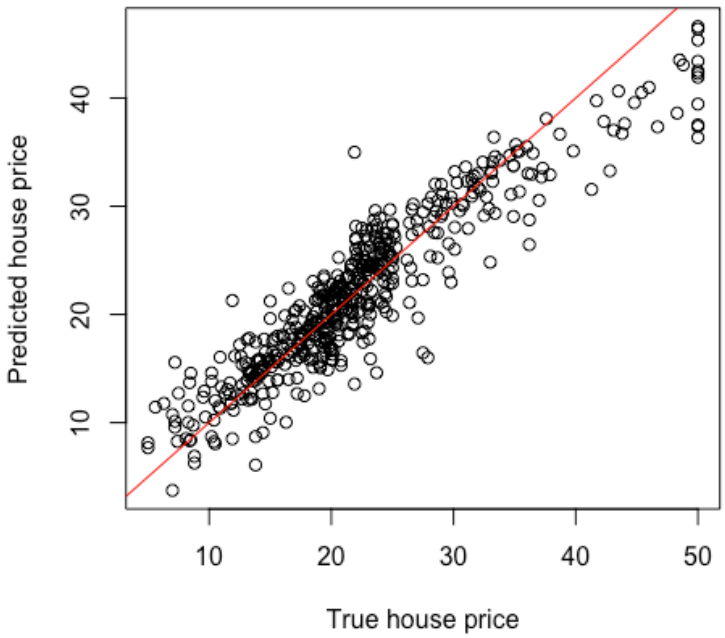
Fitted values of price:



Observe:

1. In the plot of standardized residuals against fitted values **without any transform,** there are 3 outlier points in the red box (high residuals). Despite the outlier points, all of the standardized residuals for small and for large predicted values are positive and large, which indicate a non-linear transformation of something might be helpful. And most of standardized residuals for medium predicted values are negative.
2. In the plot of standardized residuals against fitted values **after removing outliers and transforming the dependent variable,** the outlier points has gone. The weird phenomenon, that all of the standardized residuals for small predicted values are positive and large whereas most residuals for medium predicted values are negative, has gone. Although all of the standardized residuals for large predicted values are still positive, they are much more close to 0 than the plot without transformation. In summary, the

standardized residuals after removing outliers and dependent variable transformation look much more randomly and better than in the original one.



The red line is y=x

Observe:

Most of the dots are very close to the red line y=x, suggesting that the predicted house price nearly equal to the true house price. Despite the high true house price (> 40), the predicted house values almost evenly distributed on the upper and lower sides of the line, which means our model works perfectly for house price < 40. However, the predicted house price is always a little bit smaller when true house price > 40 and that is consist with that all of the standardized residuals for large predicted values are positive. In general, the final linear regression model works well.

```
library(MASS)
setwd('/Users/panzhang/Desktop/CS498AML/HW6')
data=read.table("housing.data",header = FALSE)
measure=data.matrix(data[,1:13])
price=data[,14]
model=lm(price ~ measure)
#plot(model,which=c(1:6))
stdres = rstandard(model)
plot(model$fitted.values, stdres,
        ylab="Standardized Residuals",
        xlab="Fitted values",
        main="Without any transforms")
plot(price,model$fitted.values,
     ylab="Predicted house price",
     xlab="True house price")
abline(0,1,col="red")

#Remove outlier points
data1=data[-c(369,373,372),]
measure1=data.matrix(data1[,1:13])
price1=data1[,14]
model1=lm(price1 ~ measure1)
#plot(model1,which=c(1:6))
data2=data1[-c(366,369),]
measure2=data.matrix(data2[,1:13])
price2=data2[,14]
model2=lm(price2 ~ measure2)
#plot(model2,which=c(1:6))
data3=data2[-c(367,368,408),]
measure3=data.matrix(data3[,1:13])
price3=data.matrix(data3[,14])
model3=lm(price3 ~ measure3)
#plot(model3,which=c(1:6))
#Box-Cox transformation
bc = boxcox(price3 ~ measure3)
I=which(bc$y==max(bc$y))
lambda=bc$x[I]
lambda
new_model=lm(price3^lambda~measure3)
#plot(new_model)
res3=new_model$residuals
stdres3 = rstandard(new_model)
plot(new_model$fitted.values, stdres3,
     ylab="Standardized Residuals",
     xlab="Fitted values",
     main="After removing and transforming")
plot(new_model$fitted.values^(1/lambda), stdres3,
     ylab="Standardized Residuals",
     xlab="Fitted values (new_model$fitted.values^(1/lambda))",
     main="After removing and transforming")
plot(price3,new_model$fitted.values^(1/lambda),
     ylab="Predicted house price",
     xlab="True house price")
abline(0,1,col="red")
```