CS498 AMO Homework 6

Team :
Minyuan Gu (minyuan3@illinois.edu, netid minyuan3)
Yanislav Shterev (shterev2@illinois.edu, netid shterev2)

(**0 points**) **Page 1**: code for regression and resulting model.

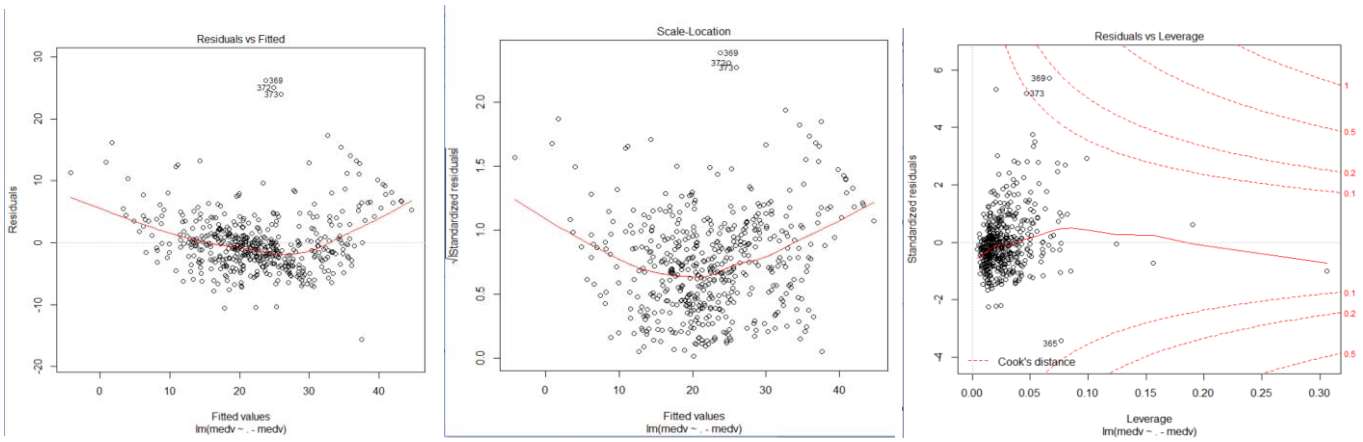
```
# Created on: 2/03/2019
library(MASS)
col_names <- c("crim","zn","indus","chas","nox","rm","age","dis","rad","tax","ptratio","b","lstat","medv")
inputdata <- read.table("./homework6/housing.data", header=F, col.names = col_names)
cooksD_lv <- c(0.1, 0.2, 0.5, 1.0)

printStats <- function(model){
    sm <- summary(model)
    print("Model: ")
    print(sm$call)
    print(paste("R squared:", sm$r.squared))
    print("Top 10 standardized residuals: ")
    print(head(sort(abs(rstandard(model)), decreasing=TRUE),10))
    print("Top 10 leverage: ")
    print(head(sort(hatvalues(model), decreasing=TRUE),10))
    print("Top 10 Cook's Distance: ")
    print(head(sort(cooks.distance(model), decreasing=TRUE),10))
}
########################################
#       Model #1 - original data       #
########################################
fit1<-lm(medv ~ . -medv, data = inputdata)
printStats(fit1)
plot(fit1, cook.levels= cooksD_lv)
```

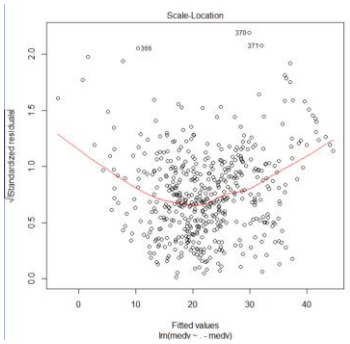
(50 points)Page 2: a screenshot of your diagnostic plot and a few sentences of your explanation.
Estimating cook distance, leverage(hat_matrix) and residuals (errors) & standardized residuals we were able to indicate 11 outlier points on row indexes: 365, 366, 368, 369, 370, 371, 372, 373, 375, 381, 413. You will see after removing them, none of the remaining points have unusually high residuals, standardized residuals or Cook's Distances. The following unfolds our analysis and decision procedure.

After regressing the house price on all the other 13 variables with all data points, we have the following plots with an R-squared value of 0.7406427. From the residuals plots, we noticed points 369,372 & 373 have the high residuals & standardized residuals which are more than 5 times std deviation away and suggest they may affect our model significantly (due to the nature of squaring that large values). From the residuals and leverage plot we reconfirmed 369 & 373 have highest Cook's distances (influential to our model and might well be an outlier). We further printed out the top 10 with highest standardized residuals, leverages and cook's distances, which confirmed point 372 is also among the top 10 Cook's distances. So we decided to remove them first. We also observed point 381 has the largest leverage, however it has very small residual & Cook's D, we leave it at later step below.

```
[1] "Top 10 standardized residuals: "
     369      372      373      370      413      365      371      187      366      162
5.713855 5.335291 5.180330 3.756430 3.505841 3.420118 3.332992 3.007767 2.933770 2.839032
[1] "Top 10 leverage: "
       381        419        406        411        366        156        491        368        493        365
0.30595949 0.19010096 0.15643251 0.12470699 0.09851493 0.08527666 0.08206742 0.08043019 0.07711251 0.07672256
 [1] "Top 10 Cook's Distance: "
       369        373        365        366        370        413        368        371        215        372
0.16567369 0.09409651 0.06942966 0.06718425 0.05526255 0.05004117 0.04541181 0.04419639 0.04292457 0.04255531
```



After the removal of the above 3 suspected outliers, we repeat the above procedure: replot and analyze. We could see the R-squared value increased to 0.7778 with 3 points removed. And we observed another set of points 370, 371, 366 shown in the new plots which have > 4 std deviation standardized residuals as well as among top 10 largest Cook's distances (point 366 also shows up at the 5th largest leverage). So they are good candidates for us to remove and test next.




```
[1] "Top 10 standardized residuals: "
     370      371      366      413      368      365      162      187      375      167
4.813466 4.308905 4.216479 3.900341 3.752002 3.686437 3.283152 3.191735 3.140576 3.064341
[1] "Top 10 leverage: "
       381        419        406        411        366        156        368        491        365        153
0.30637805 0.19014434 0.15648033 0.12613749 0.10678546 0.08611433 0.08581017 0.08227720 0.07781152 0.07755950
[1] "Top 10 Cook's Distance: "
       366        370        368        365        371        413        381        415        215        375
0.15181971 0.09793988 0.09438446 0.08190490 0.07909482 0.06194224 0.05073578 0.03833781 0.03764522 0.03521632
```

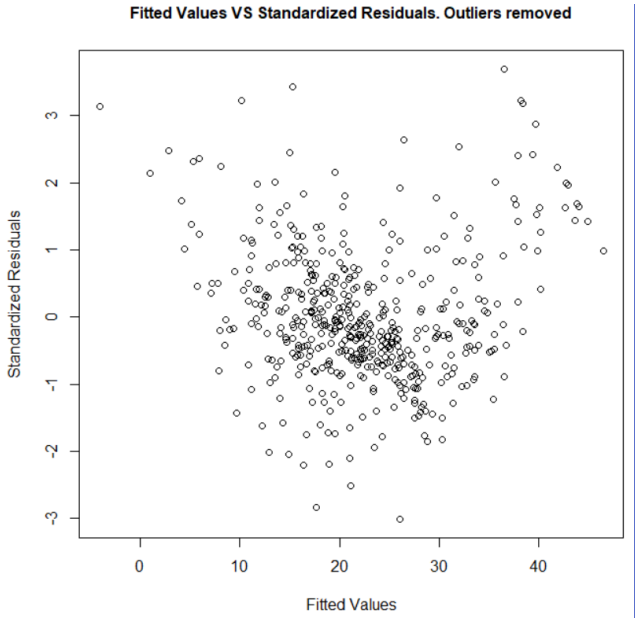
So we repeat the process to look for more outliers. Point 381 has largest leverage since beginning which means it was predicted more from itself rather than other samples. Now after above step, it starts showing up in top 10 high Cook's distances, so we decided to remove it. As the same procedure was repeated, we finally stopped at 11 points: 365, 366, 368, 369, 370, 371, 372, 373, 375, 381 & 413. We stopped since none of the points have larger than 4 standard deviations in standardized residuals(<4), high Cook's distances (<0.26) or high leverage (<0.2), with R^2=0.8283. Refer to next page for details.

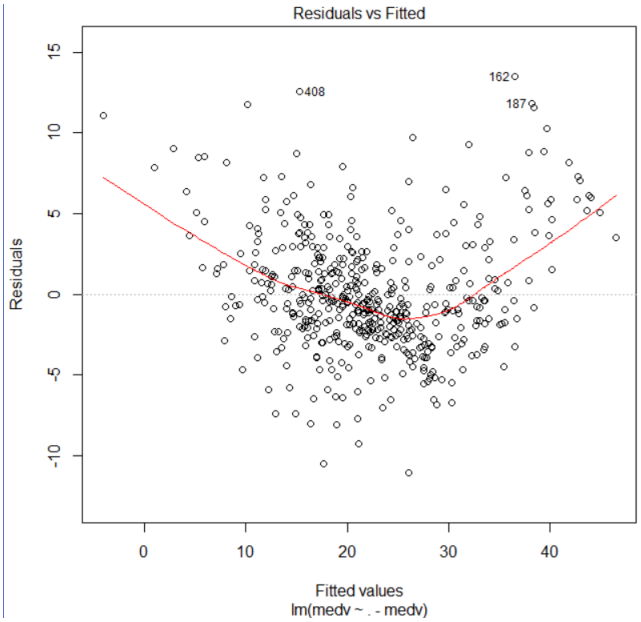
```
[1] "R squared: 0.828325078406639"
[1] "Top 10 standardized residuals: "
     162      408      187      367      167      415      376      163      402      182
3.698090 3.433237 3.221316 3.218697 3.185159 3.131236 3.013168 2.865499 2.840508 2.645116
[1] "Top 10 leverage: "
       419        406        411        415        156        491        215        153        493        492
0.25616985 0.22367134 0.15572621 0.09140755 0.08794674 0.08354652 0.08263255 0.08210704 0.07780236 0.07696812
[1] "Top 10 Cook's Distance: "
       415        406        163        215        162        167        367        408        164        407
0.07045575 0.04212467 0.04086557 0.03838112 0.03835376 0.03181224 0.02905559 0.02282503 0.02209394 0.02136026
```

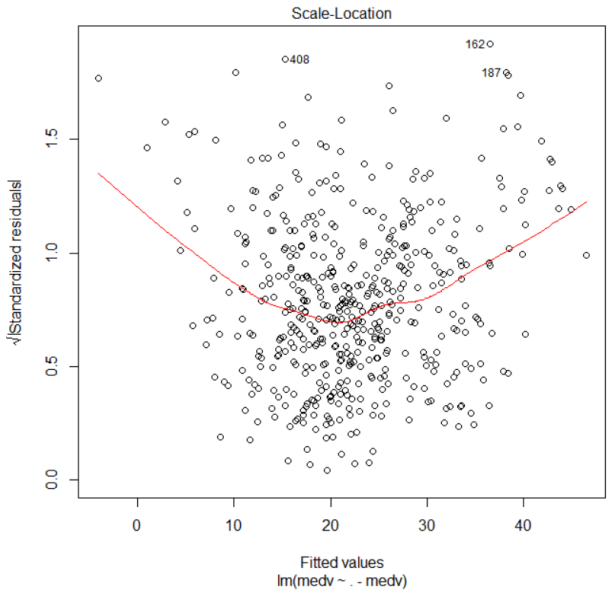
(**20 points**) **Page 3**: a screenshot of your new diagnostic plot.

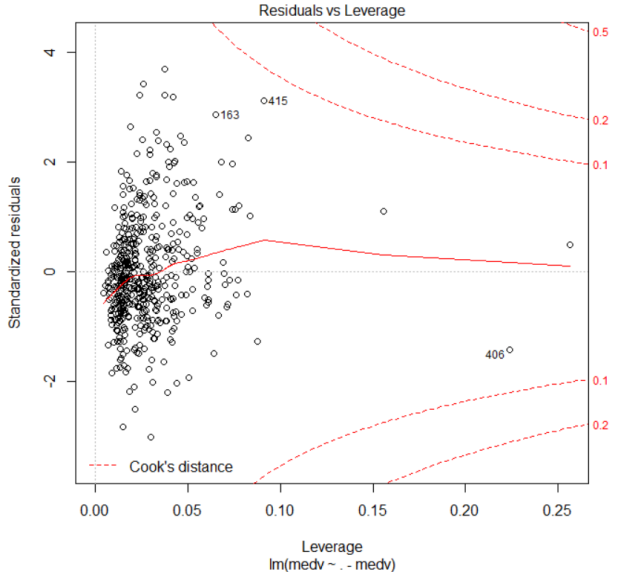
After removing the 11 outliers, R-squared metric became 0.828325, increased from initial 0.7406427. We also observed the max leverage reduced from 0.30 to 0.25 and all the points are within 4 standard deviations from the mean. Given there are 506 observations in the data set, 11 outliers means around 2% of the data. As said, we stopped since none of the points have large standardized residuals, high Cook's distances or high leverage, it will be prudent to stop at 2% data removed as outlier and not go beyond that.

We also noticed data point 415 has negative predicted value (-4.084973) so far, and it will be addressed by the Box-cox transformation later, please refer to next page for discussion.









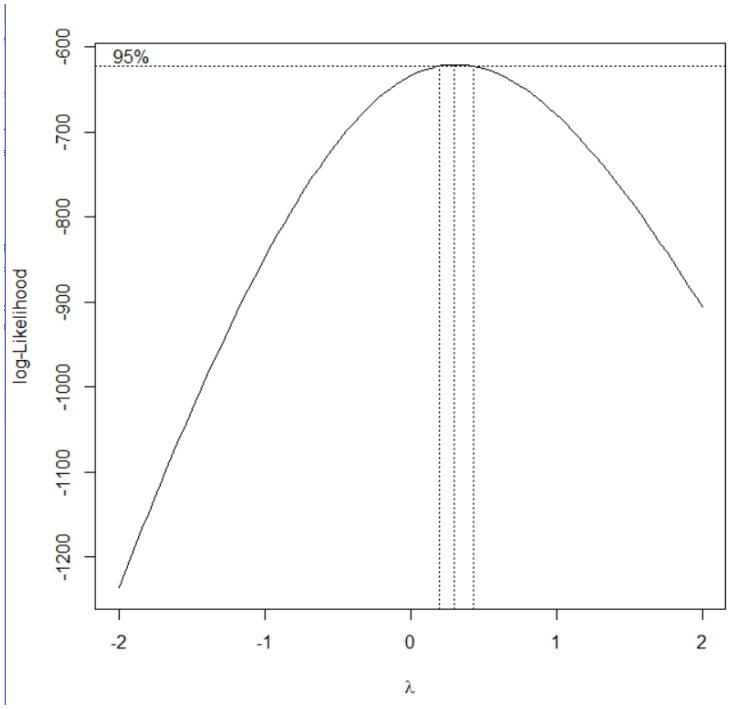
(**10 points**) **Page 4**: a screenshot of your code for subproblem 2.

```
########################################
#   Model #2 - outliers removed data    #
########################################
# now remove the outliers of point 369, 372 & 373 - due to high standardized residuals and high cook's distance
removed <- inputdata[-c(369, 372, 373),]
fit2<-lm(medv ~ . - medv, data = removed)
printStats(fit2)
plot(fit2, cook.levels= cooksD lv)
# also removed  366, 368, 370, 371 which have large standardized residual & cook's distance
# and we repeated the above procedure.
# we skipped some test cycles here and jump to final removal.
# we remove total 365, 366, 368, 369, 370, 371, 372, 373, 375, 381, 413
removed <- inputdata[-c(365, 366, 368, 369, 370, 371, 372, 373, 375, 381, 413),]
fit2<-lm(medv ~ . - medv, data = removed)
printStats(fit2)
plot(fit2, cook.levels= cooksD lv)
plot(fitted(fit2), rstandard(fit2), main="Fitted Values VS Standardized Residuals. Outliers removed", xlab="Fitted
Values",ylab="Standardized Residuals", cex.main=1)
```

(**10 points**) **Page 5**: a screenshot of Box-Cox transformation plot and the best value you chose.

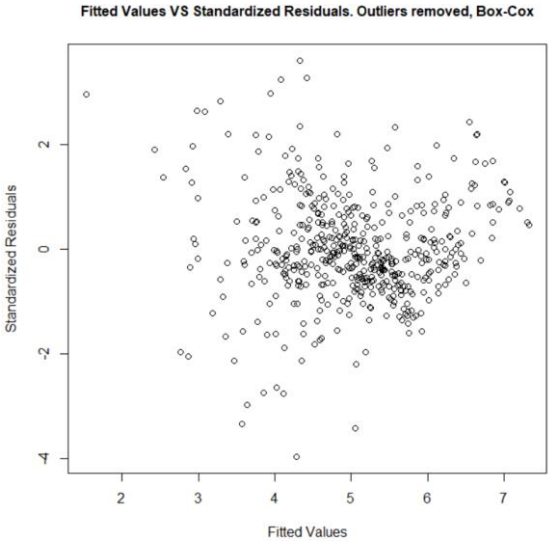
The best value of lambda is 0.3030303, resulting from below graph and codes:

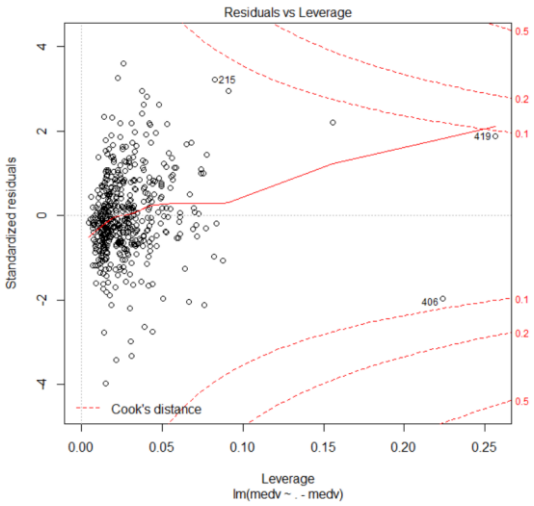
```
results <- boxcox(fit2)
lambda <- results$x[which.max(results$y)]
print("Suggested lambda from boxcox: ")
print(lambda)
> [1] "Suggested lambda from boxcox: "
> [1] 0.3030303
```



(**10 points**) **Page 6**: result of the standardized residuals of the regression after Box-Cox transformation and a plot of fitted house price against true house price.

After fitting linear regression with Box-Cox transformation the new R-squared value became: 0.84306. Also point 415, that used to have negative predicted value (-4.084973), now has a value of 3.512946 (post reversal of Box-Cox transformation). Below are the standardized residuals plots (Noted: fitted values in below plots are the transformed values, but not the final prediction since they needs to be reversed from Box-Cox transformation.)





Below is the plot of fitted house prices against true house prices.

(**0 points**) **Page 7**: code for subproblems 3 and 4.

```
############################################################
#   Model #3 - apply boxcox with outliers removed data     #
############################################################
results <- boxcox(fit2)
lambda <- results$x[which.max(results$y)]
print("Suggested lambda from boxcox: ")
print(lambda)
# now calculate using lambda transformed Y
transformed <- removed
transformed$medv<- ((transformed$medv)^lambda - 1) / lambda
fit3<- lm(medv ~ . - medv, data = transformed)
printStats(fit3)
plot(fit3, cook.levels= cooksD_lv)
plot(fitted(fit3), rstandard(fit3), main="Fitted Values VS Standardized Residuals. Outliers removed, Box-Cox", xlab="Fitted
Values",ylab="Standardized Residuals", cex.main=1)

# now print the predicted data against true data
predicted = (predict(fit3, removed[,-14])*lambda+1)^(1/lambda)
plot(removed[,14], predicted, xlab="True House Price",ylab="Fitted House Price",main='True House Price vs Fitted Price -
outlier removed')
```

Libraries used & Reference:

David Forsyth's book - Applied Machine Learning
Trevor Walker's lecture and sample code – CS-498 Lecture videos
Accelerometer dataset - https://archive.ics.uci.edu/ml/datasets/Dataset+for+ADL+Recognition+with+Wrist-worn+Accelerometer
R MASS library - https://cran.r-project.org/web/packages/MASS/index.html
R - https://cran.r-project.org/