**CS498 AMO Homework 6**

Team :

Minyuan Gu ([minyuan3@illinois.edu](mailto:minyuan3@illinois.edu), netid minyuan3)

Yanislav Shterev ([shterev2@illinois.edu](mailto:shterev2@illinois.edu), netid shterev2)

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

Residuals:

Min 1Q Median 3Q Max

-15.595 -2.730 -0.518 1.777 26.199

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.646e+01 5.103e+00 7.144 3.28e-12 \*\*\*

crim -1.080e-01 3.286e-02 -3.287 0.001087 \*\*

zn 4.642e-02 1.373e-02 3.382 0.000778 \*\*\*

indus 2.056e-02 6.150e-02 0.334 0.738288

chas 2.687e+00 8.616e-01 3.118 0.001925 \*\*

nox -1.777e+01 3.820e+00 -4.651 4.25e-06 \*\*\*

rm 3.810e+00 4.179e-01 9.116 < 2e-16 \*\*\*

age 6.922e-04 1.321e-02 0.052 0.958229

dis -1.476e+00 1.995e-01 -7.398 6.01e-13 \*\*\*

rad 3.060e-01 6.635e-02 4.613 5.07e-06 \*\*\*

tax -1.233e-02 3.760e-03 -3.280 0.001112 \*\*

ptratio -9.527e-01 1.308e-01 -7.283 1.31e-12 \*\*\*

black 9.312e-03 2.686e-03 3.467 0.000573 \*\*\*

lstat -5.248e-01 5.072e-02 -10.347 < 2e-16 \*\*\*

---

Residual standard error: 4.745 on 492 degrees of freedom

Multiple R-squared: 0.7406, Adjusted R-squared: 0.7338

F-statistic: 108.1 on 13 and 492 DF, p-value: < 2.2e-16

1. (**50 points**) **Page 2**: a screenshot of your diagnostic plot and a few sentences of your explanation.

After regressing the house price on all the other 13 variables the R-squared is as follows:

[1] 0.7406427

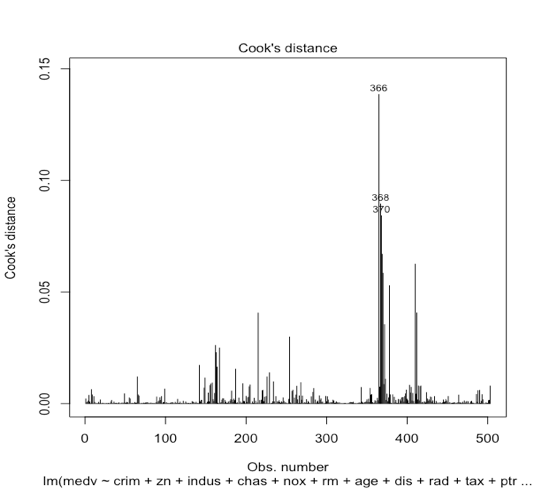
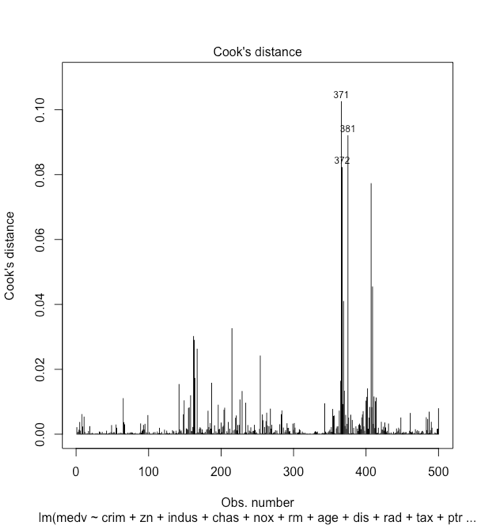
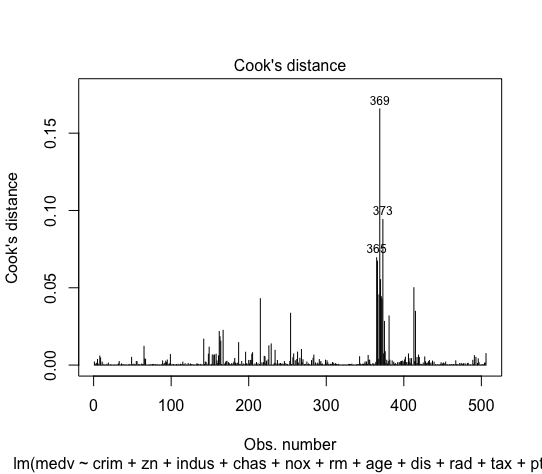
Estimating cook distance, leverage(hat\_matrix) and residuals(errors) we were able to indicate 10 outlier points on row indexes: 366, 368, 370,365, 369, 373,372,371,381,413

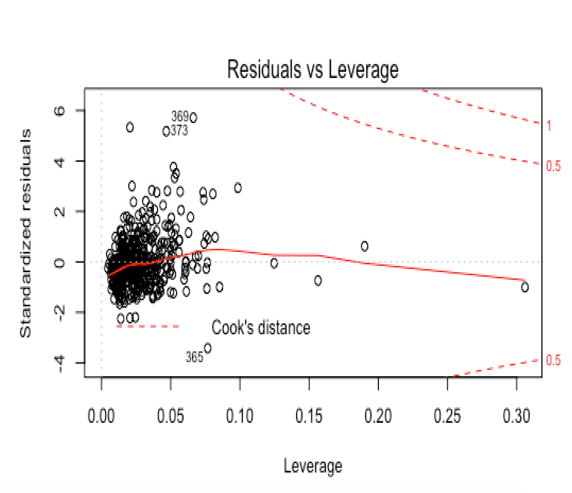
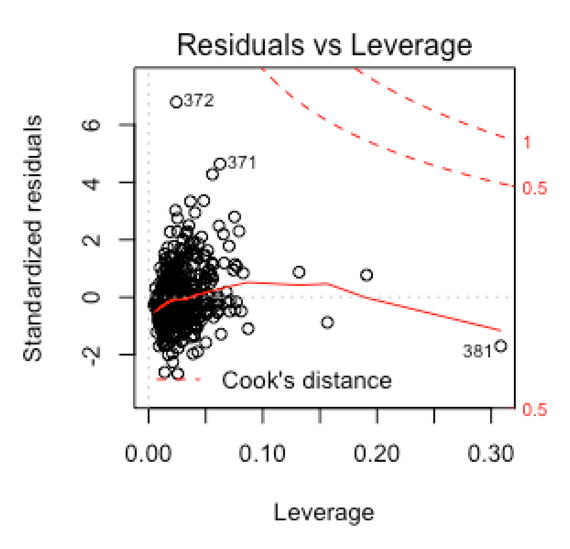
Cutoff Threshold for detecting outliers: 0.008163265

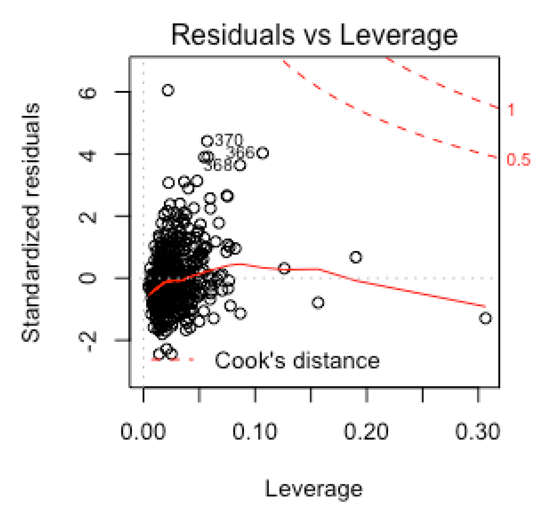
All of the listed outlier points have very high cook distance which makes them influential points. Any large difference from the 0 residuals with combination of high leverage makes them influential.

381, 413 have very high leverage

365- has very negative residual (more than 3 standard deviations below the mean)

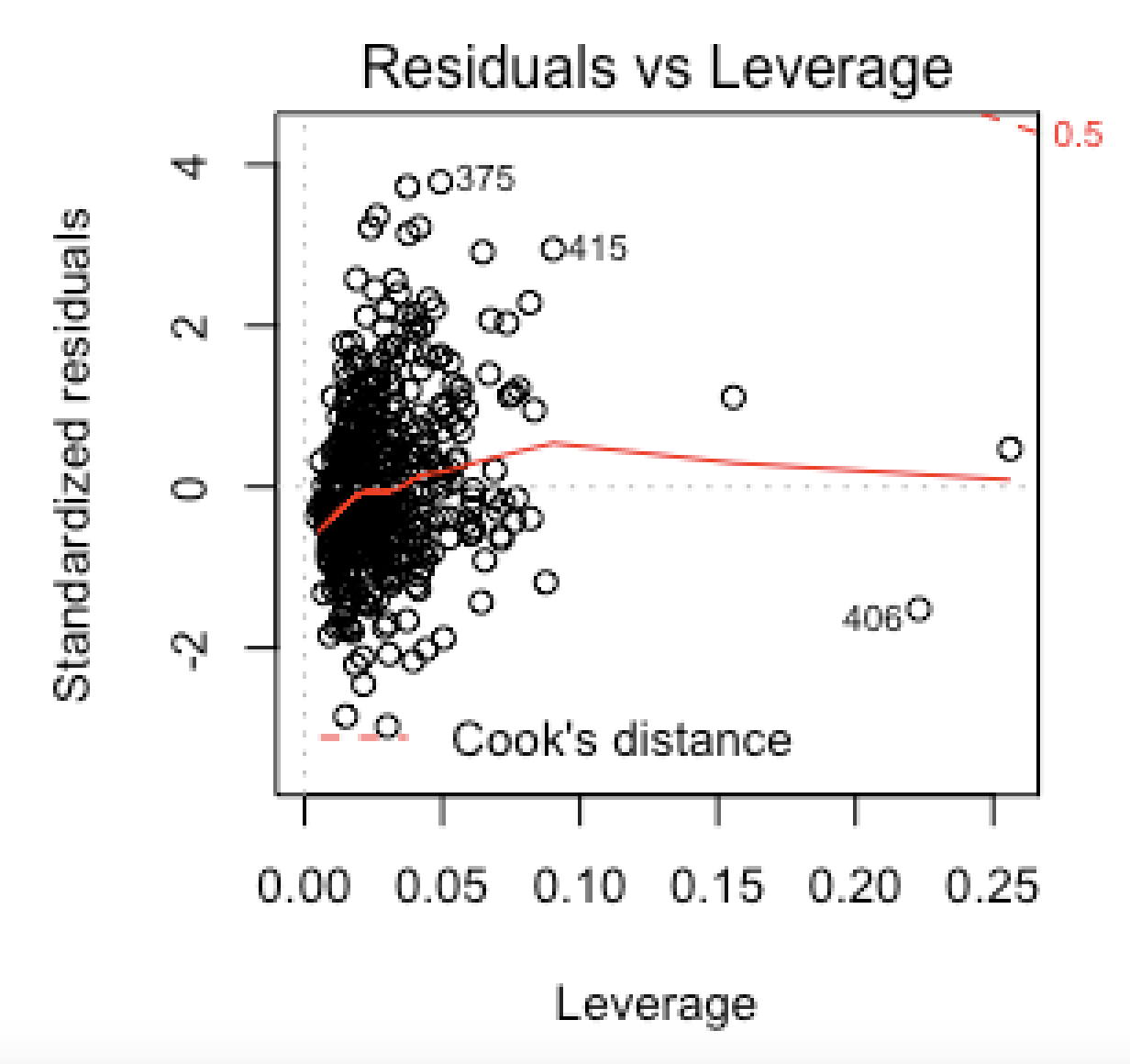
366, 368, 370, 369, 373, 372, 371 have very high positive error values(residuals) more than 4 standard deviations above the mean.





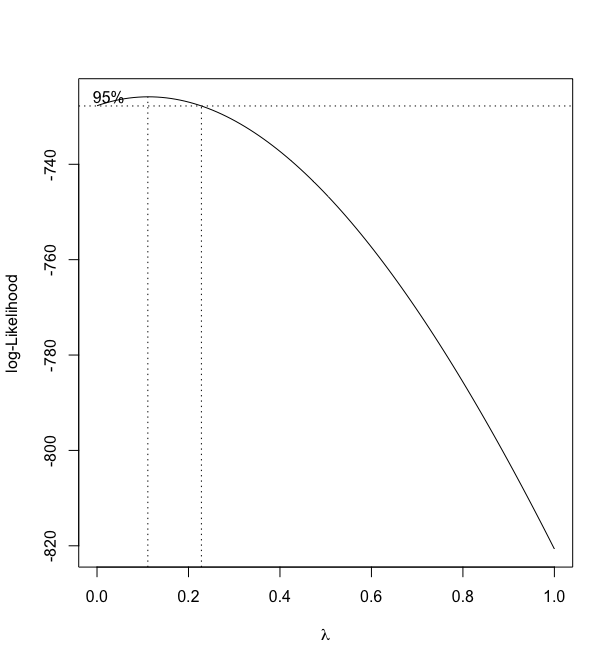
1. (**20 points**) **Page 3**: a screenshot of your new diagnostic plot.

After removing the outliers R-squared metric became 0.8233966. We can see how the max leverage shrunk from 0.30 to 0.25 and all the points are within 4 standard deviations far from the mean. This is still far comparing with the general advice to try to keep them between 3 standard deviations but also, we have ~500 observations and we already took out 10 which is already 2% of the data and we can see there are wide number of points being more than 3 standard deviations from the positive side of the mean. Which shows that the data is a little skewed.



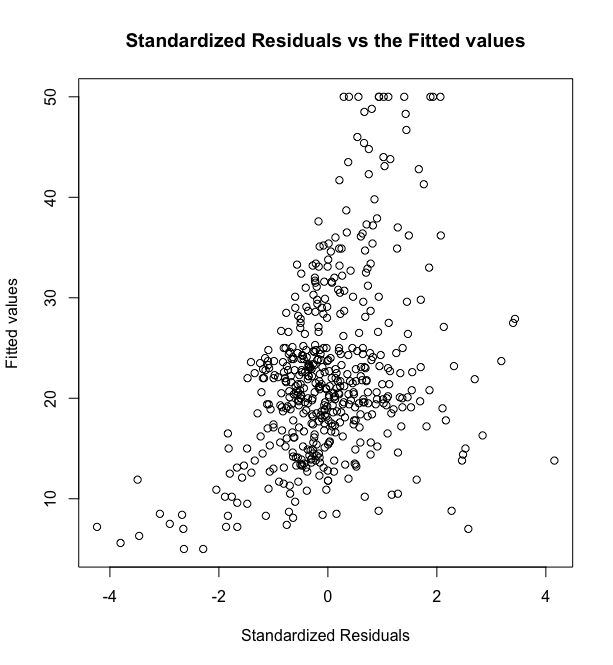
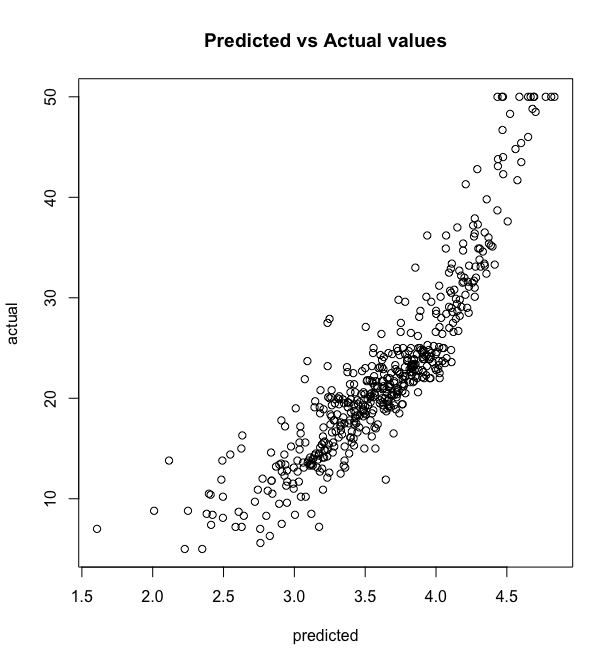
1. (**10 points**) **Page 4**: a screenshot of your code for subproblem 2.
2. (**10 points**) **Page 5**: a screenshot of Box-Cox transformation plot and the best value you chose.

The best value of lambda is 0.1111111



1. (**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: 0.8348808



1. (**0 points**) **Page 7**: code for subproblems 3 and 4.

**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>