STA 207 HW-6
Due Date: 11/7 by 5PM

Problem-1 [25 points]: We will work with the data on median housing prices in neighborhoods in the suburbs of Boston (same as hw-5). Regress medv (Y) on lstat (X) and report the fitted model. For this model

- [5 points] Identify high leverage points, if any.
- [5 points] Identify outliers, if any.
- [15 points] Identify influential points, if any. Use Cook's distance and influence plots.

Solution: For the SLR model regressing median value of owner-occupied homes in $1000s on lower status of the population (percent), we identify

a) 34 observations which have high leverage and these are

9 33 49 124 127 142 143 144 145 146 148 149 215 374 375

385 386 387 388 389 393 399 400 401 405 409 413 415 416 417

418 438 439 491

b) 32 observations are classified as outliers and these are

99 142 162 163 164 167 181 187 196 204 205 215 225 226 229

234 257 258 262 263 268 281 283 284 369 370 371 372 373 375

413 506

c) 41 observations are identified as influential and these are

9 49 99 142 148 149 162 163 164 167 187 196 203 204 205

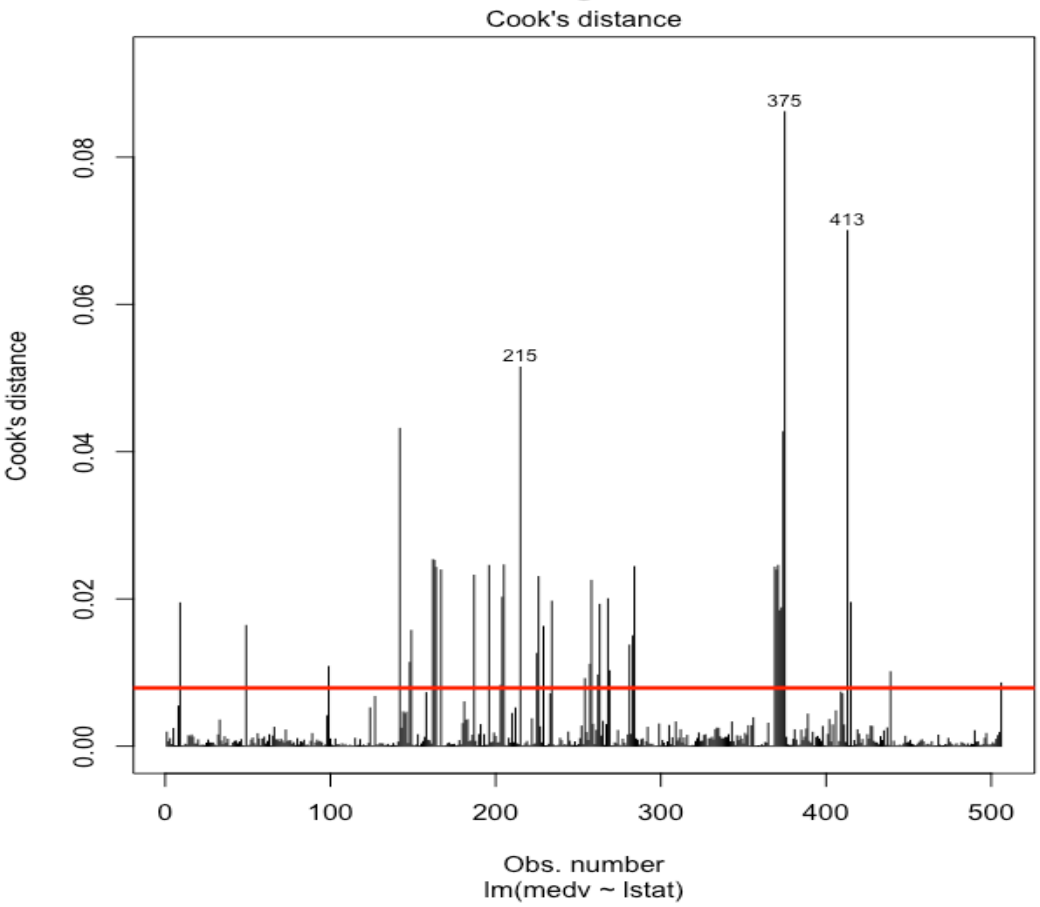
215 225 226 229 234 254 257 258 262 263 268 269 281 283 284

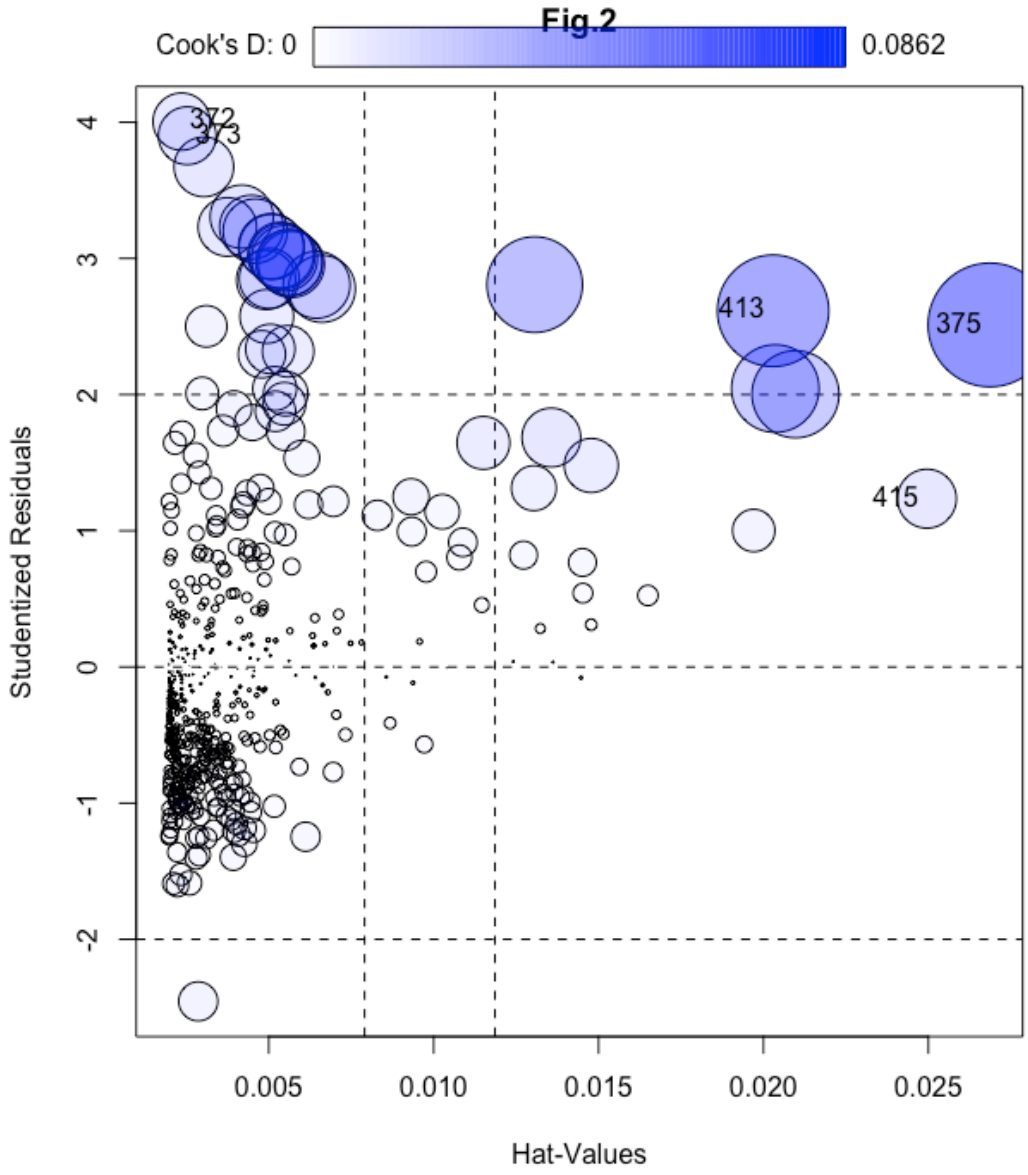
369 370 371 372 373 374 375 413 415 439 506

Fig.1 below shows the observations whose Cook's distance values are above the threshold of 4/n (shown with the red line).

Fig.2 below shows the influence plot with high Cook's distance observation represented using darker blue color.

Fig.1


Cook's distance
375
413
215
0.08
0.06
0.04
0.02
0.00
Cook's distance
0
100
200
300
400
500
Obs. number
lm(medv ~ lstat)

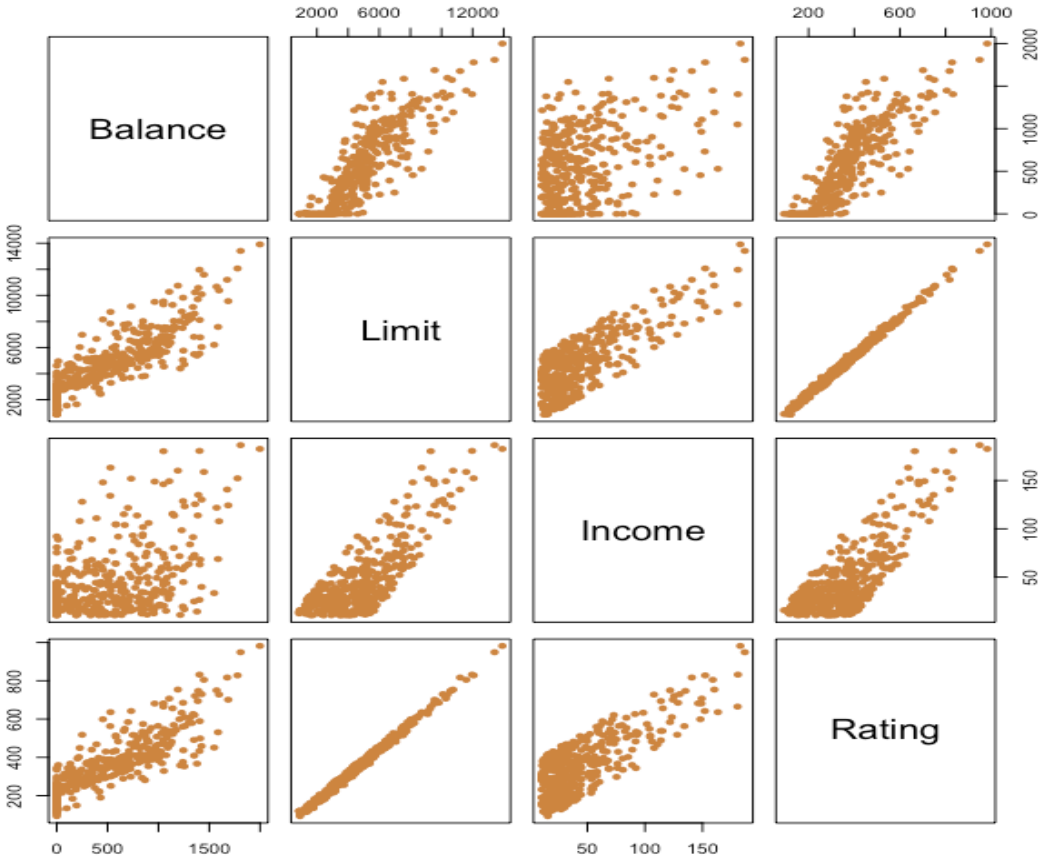


Problem-2 [75 points]: We will need R package ISLR for this problem so install it first. In R package ISLR, we will use dataset called Credit. This is a simulated data set containing information on ten thousand customers. The aim here is to predict which customers will default on their credit card debt. The outcome variable of interest is the credit card debt of 400 individuals. Other variables like income, credit limit, credit rating, and age are included as well. Note that the Credit data is not based on real individuals' financial information, but rather is a simulated dataset used for educational purposes. Using outcome variable average credit card debt (called Balance) and three predictor variables cardholder's credit limit (Limit), Income and Credit rating (Rating), do the following:

a.) [7.5 points] Make a scatterplot matrix and explain relationship of outcome variable with all three predictor variables.

Fig. 3 shows pairwise scatterplots for all variables. We see that balance is positively related to Limit, Income, and Rating. This relation between balance is linear with limit and rating, whereas it is not as strong between balance and income. We can also see a very strong positive relation between limit and rating, which might lead to multi-collinearity. Likewise, there is positive and linear relationship between limit and income.

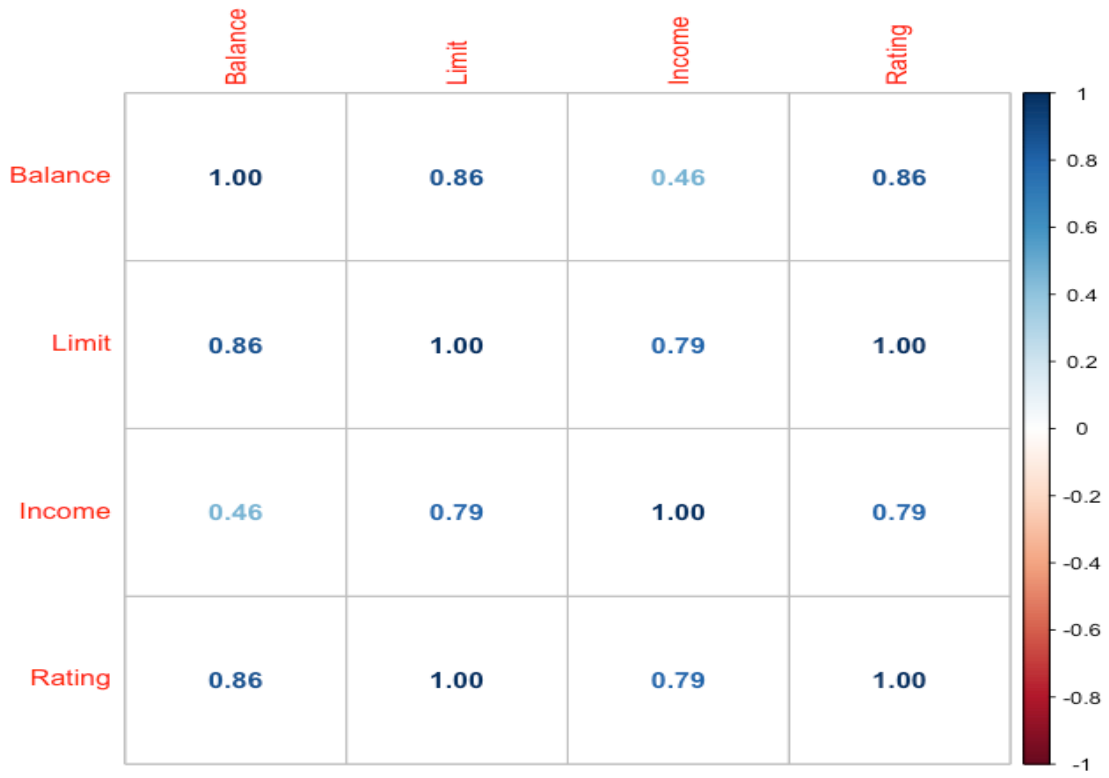
Fig.3: Matrix Scatterplot



b.) [7.5 points] Make a corrplot and see if your explanation in (a) match with the strength and direction of the correlation coefficient for each relationship.

Fig. 4 shows pairwise correlations for all variables. We see that balance is positively related to Limit, Income, and Rating. This relation between balance is strong with limit and rating, whereas it is mild between balance and income. This matched with what the observations from the scatterplot pairs in Fig.6. We can also see a perfect (very strong) positive relation between limit and rating, which will definitely lead to multi-collinearity. Likewise, there is positive and linear relationship between limit and income.

Fig.4: Correlation Plot



c.) [10 points] Regression outcome variable on the three predictors and report the fitted model. Interpret the regression coefficients.

$\hat{Y} = -489.73 + 0.085X_1 - 7.72X_2 + 2.70X_3$, where

$\boldsymbol{Y}$: average credit card debt in $(called Balance)

$\boldsymbol{X_1}$: cardholder's credit limit (Limit),

$\boldsymbol{X_2}$: Income and

$\boldsymbol{X_3}$: Credit rating (Rating)

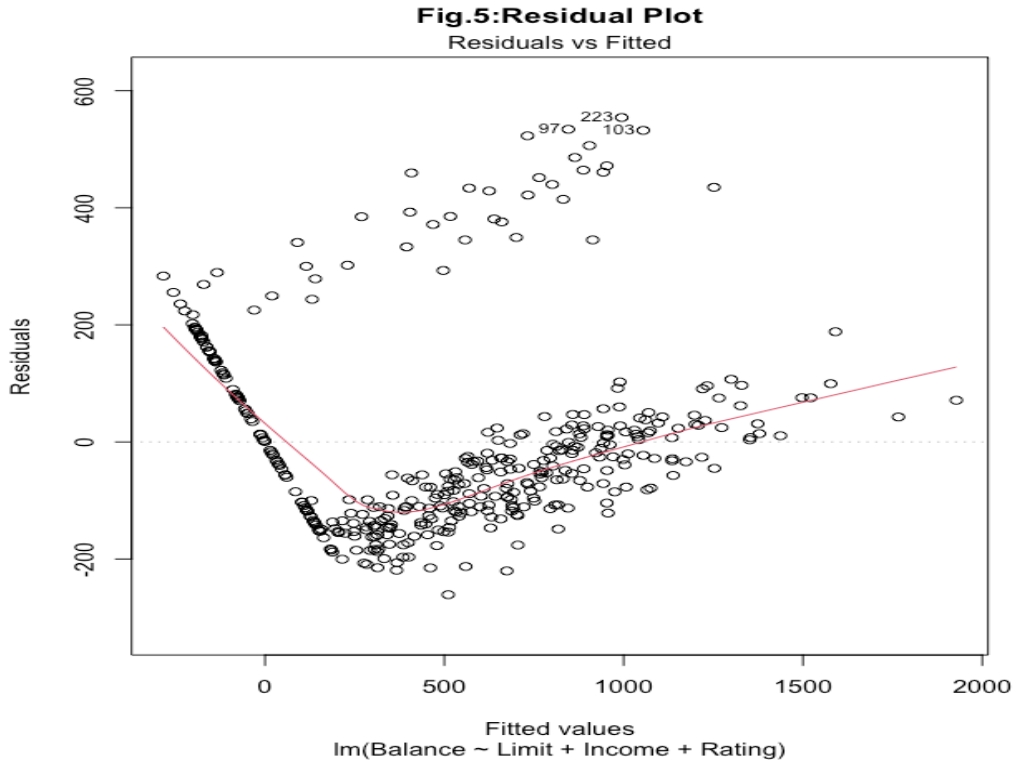
d.) [10 points] Report goodness of fit for the above model.

Goodness of fit:

i) $R^2 = 0.8762$ showing that approximately 87.6% change in average credit card debt is explained together by cardholder's credit limit, income and their credit rating.

ii) $\hat{\sigma} = 162.4$ meaning that on an average the fitted values of average credit card debt vary 162$ from the observed debt.

iii) Residual Plot: In Fig.5 there is a clear curve pattern in the residual plot, meaning that the model is not extracting all information present in the average credit card debt.



e.) [25 points] Report if the LINE conditions are met or not.

To test linearity condition, we use

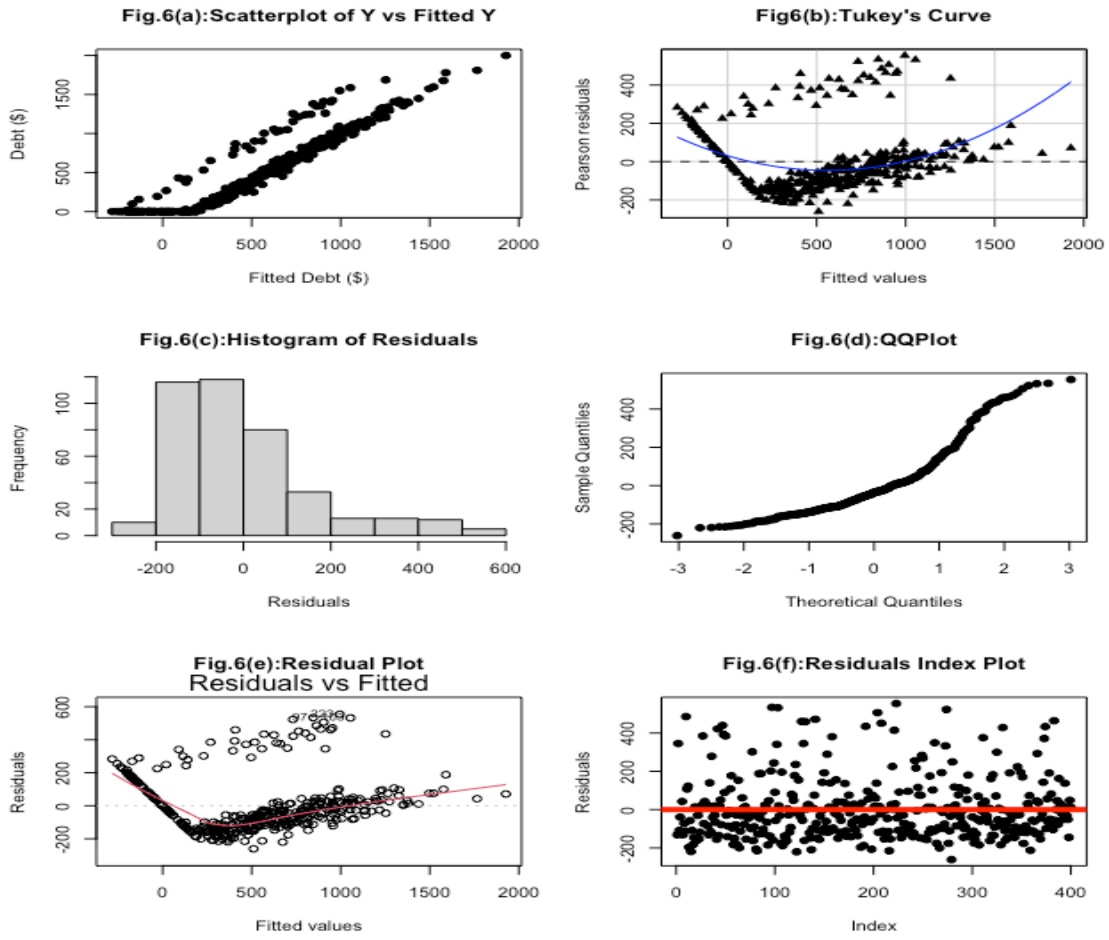
- scatterplot of debt versus estimated debt: Fig 6(a) shows a linear relation between te two with two parallel set of dots.
- Residual plot: In Fig. 6(e) shows a clear pattern.
- Tukey's curve test: In Fig. 6(b) we see a clear curve with P-value~0 thus rejecting the linearity.

The linearity assumption fails.

To test normality of errors we use histogram in Fig.6(c) which appears right skewed and QQplot in Fig. 6(d) which shows a curve, thus normality assumption is also failing.

To test if errors are homoscedastic, we use residual plot in Fig. 6(e) which shows some patterns of different variability. Then, we use the BP test which results in P-value=0.5974. Since P-value>0.05, we have evidence of homoscedasticity (do not reject homoscedasticity hypothesis).

The independence of errors assumption is tested using Fig.6(f) which shows random spread without any obvious patterns; therefore the ind



f.) [15 points] Check if there is multi-collinearity in the three predictor variables.

The initial scatterplot shows that variables Limit and Rating have a strong, positive correlation and correlation matrix confirms that there is a perfect linear correlation between Limit and Rating. The variance inflation factors for these variables are 161.19 and 160.71, respectively. This is a clear sign of multi-collinearity problem due to high correlation between Limit and Rating.