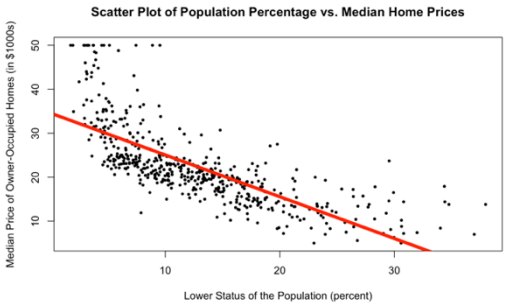
STA 207 HW-6
Due Date: 11/8 by 10:20 AM

Derin Gezgin | Camel ID: 00468038

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.



The fitted equation: $\widehat{Y} = -0.95X + 34.554$

Where:

x = Lower Status of population in percentage

Y = Median value of Owner-Occupied Homes

For this model

- **[5 points]** Identify high leverage points, if any.

  When we run the code to get the hat-values and filter the hat-values to only have the that are larger than two times the mean of the hat-values we will have the following points, which are the high leverage points. There are 34 high leverage points:

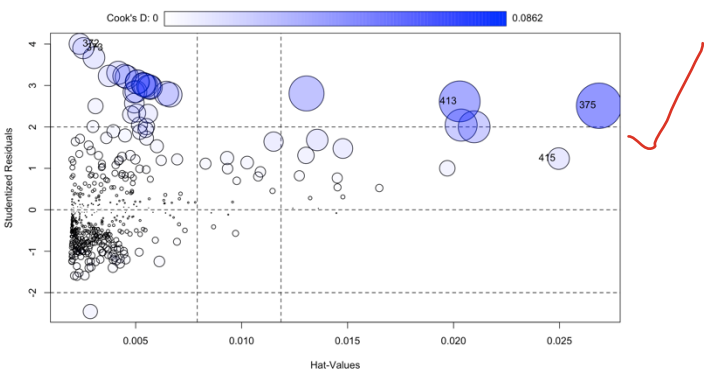
  9 33 49 124 127 142 143 144 145 146 148 149 215 374 375 385 386 387 388 389 393 399 400 401 405 409 413 415 416 417 418 438 439 491

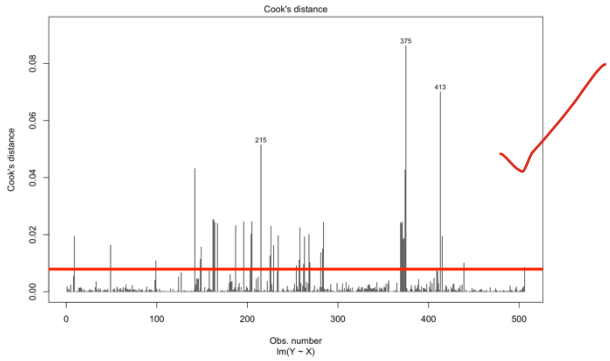
- **[5 points]** Identify outliers, if any.

  When we run to code to get the residual values larger than 2 and less than -2 we will have the following points which are the outliers. There are ~~33~~ outliers: 32

  99 142 162 163 164 167 181 187 196 204 205 215 225 226 229 234 257 258 262 263 268 281 283 284 369 370 371 372 373 374 375 413 506

- **[15 points]** Identify influential points, if any. Use Cook's distance and influence plots.





When we check the Cook's distance for the influential points, we can see that there are 41 influential points:

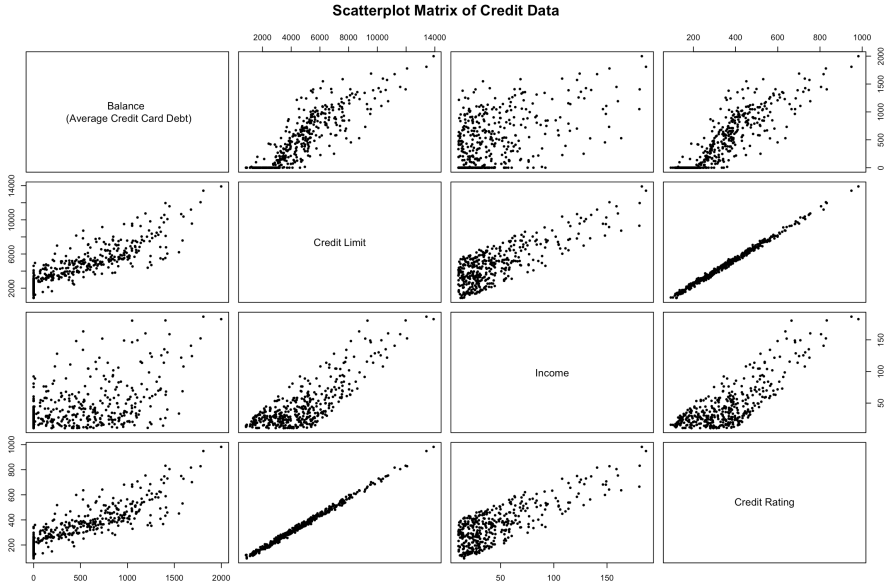
9 49 99 142 148 149 162 163 164 167 187 196 203 204 205 215 225 226 229 234 254 257 258 262 263 268 269 281 283 284 369 370 371 372 373 374 375 413 415 439 506

On the left, you can see the *Influence Plot*. In the influence plot we can see some points from my R output like 413, 375, and 372.

On the right, you can see the Cook's Distance plot and similar to the influence plot, we can see that 375 and 413 are points with large influence.

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.



Using this scatterplot, we can explain the relationship between our outcome variable balance (average credit card debt) and predictor variables (Credit limit, Income, and Credit rating).

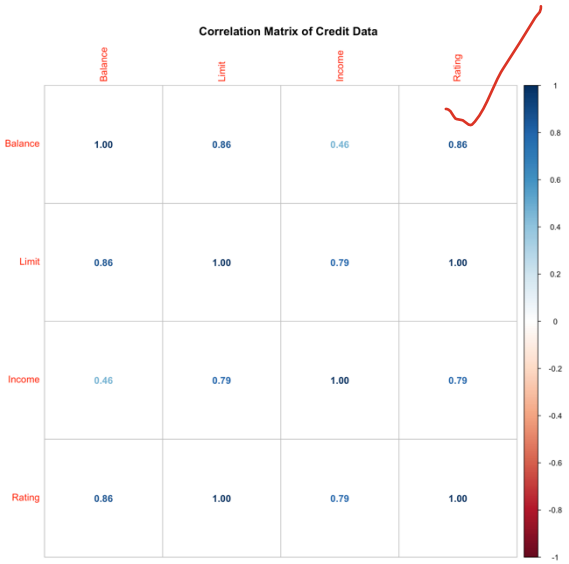
- ***Relationship between average credit card debt and credit limit***
  From the intersection of $\hat{Y}$=Average credit card debt and X=credit limit, we can see that there is a relationship between these variables in a medium-strong strength. There is definitely a relationship between them as we can see the upward trend but it is far from being very strong.

- ***Relationship between average credit card debt and income***
  From the intersection of $\hat{Y}$ = Average credit card debt and X=income, we can see that there is not a significant linear relationship between the variables.

- ***Relationship between average credit card debt and credit rating***
  From the intersection of $\hat{Y}$ = Average credit card debt and X = credit rating, we can see that the relationship is really similar to the relationship between the average credit card debt and the credit limit which is in a medium-strong strength. There is definitely a relationship considering the upward trend but it is not very strong at all.

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.

Using this correlation plot, we can reconsider my relationship estimations between the variables.



- ***Relationship between average credit card debt and credit limit***
  We can see that the correlation coefficient value between the balance (average credit card debt) and credit limit is ***0.86***, which is actually a strong positive relationship as it is close to 1. It can even be considered as very strong.

- ***Relationship between average credit card debt and income***
  We can see that the correlation coefficient value between the balance (average credit card debt) and credit limit is ***0.46***, which supports my theory of there is no significant linear relationship between the variables.

- ***Relationship between average credit card debt and credit rating***
  Similar to the previous question, we can see that the correlation coefficient value between the balance (average credit card debt) and credit rating is ***0.86***, which is actually a strong positive relationship as it is close to 1. It can even be considered as very strong.

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

$$\hat{Y} = -489.727 - 7.719x_1 + 2.699x_2 + 0.085x_3$$

Where;

Y = Average credit card debt | x_1= Income of the individual

x_2 = Credit rating of the individual | x_3 = Credit card limit of the individual

x_1 Income of the individual
The average change in the balance (average credit card debt) is -7.719 for a 1$ increase in the income of an individual with a certain credit rating and credit card limit.

x_2 Credit rating of the individual
The average change in the balance (average credit card debt) is 2.699 for a 1-point increase in the credit rating of an individual with a certain income and credit card limit.

x_3 Credit card limit of an individual
The average change in the balance (average credit card debt) is 0.085 for a 1$ increase in the credit card limit of an individual with a certain income and credit rating.

Intercept

The mean balance (average credit card debt) is -489.727 for an individual with 0$ income, 0 credit rating and 0 credit card limit.

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

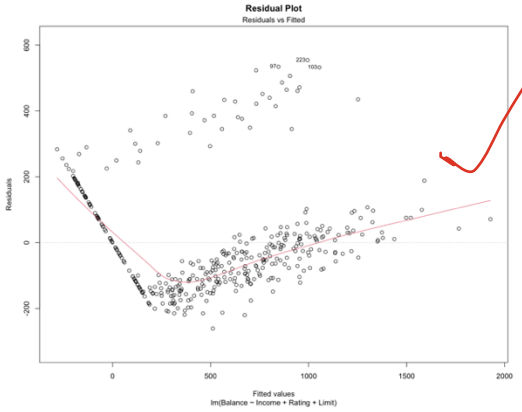
Coefficient of Determination

The multiple R-Squared value is 0.8762. Which means that 87.6% of the variability on the balance (average credit card debt) of an individual can be explained by income, credit rating and credit card limit of an individual. This indicates a strong fit.

Variability in Errors

The residual standard error is 162.4 which is the average distance observed balance (average credit card debt) values fall from the regression line. Considering the scope of the balance, $162.4 variance is a very big value which would not occur in a strong fit.

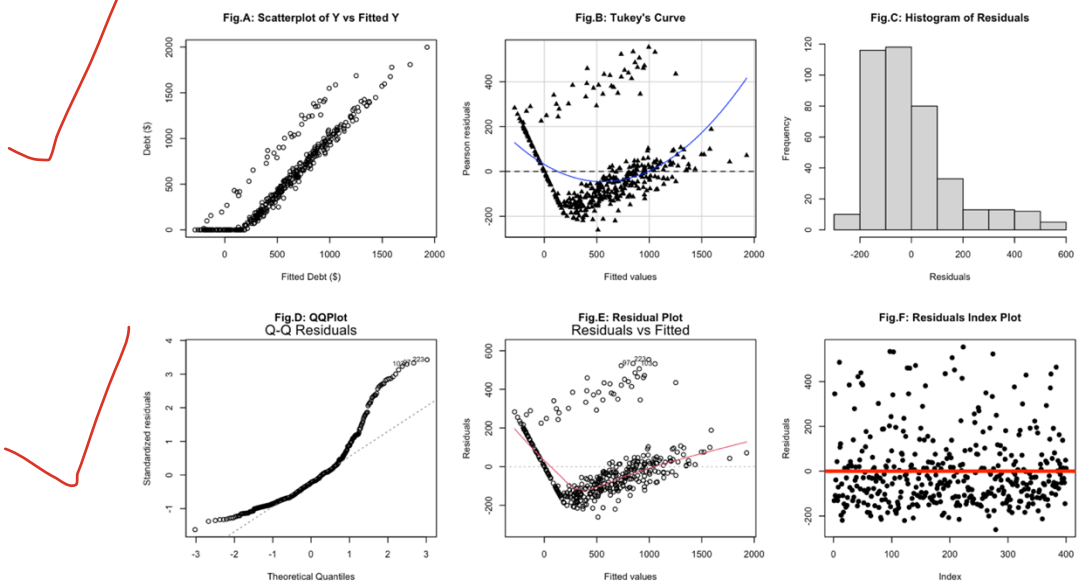
Residual Plot



From the residual plot, we can see that there is a very clear pattern throughout the whole plot. This is a sign of a not a very good fit as we can see a significant pattern.

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

L: There is a linear relationship between the X and Y



To test the linearity, we can check Figures A and B.

From figure A, we can see that there is a *weird* linear relationship. We can see that on the right side, there is a very strong linear relationship while there is also another line on the left of that line. Also, there is a straight line in the beginning of the graph, which does not support the linearity. We can say that this scatterplot does not support the linearity assumption.

From figure B (Tukey's Curve), we can see that the blue fit line has significant deviation from the flat line, which does not support the linearity assumption. We can do the Tukey's curve test to confirm the non-linearity:

Null Hypothesis (H_0): Linearity assumption holds
Alternative Hypothesis (H_A): Linearity assumption fails.
Significance level α → 0.05 (default value)
Test Statistic → 7.783
P-Value → 0.0000000000000007
P-Value < α → Reject H_0
Conclusion: At 5% level of significance linearity assumption **fails.**

I: Independent Errors

For this assumption we check the index plot of the residuals which is Figure F. While we can see that the distribution is not random, we cannot say that the distribution of residuals has a cyclic pattern. I would say weak-accept to the independence of errors assumption.

N: Normally distributed errors

For this assumption, there are two different figure we can check: Figure C (Histogram of residuals) and Figure D (Q-Q Plot).

From Figure C, we can clearly see that the distribution of residuals has a right skew and this violates the normality assumption.

Similarly, in the Q-Q plot, we can see a clear curve towards the end of the Q-Q plot and this also violates the normality assumption.

In conclusion, I can definitely say that the normality assumption **fails**.

E: Errors are homoscedastic, $Var(\epsilon) = \sigma^2$

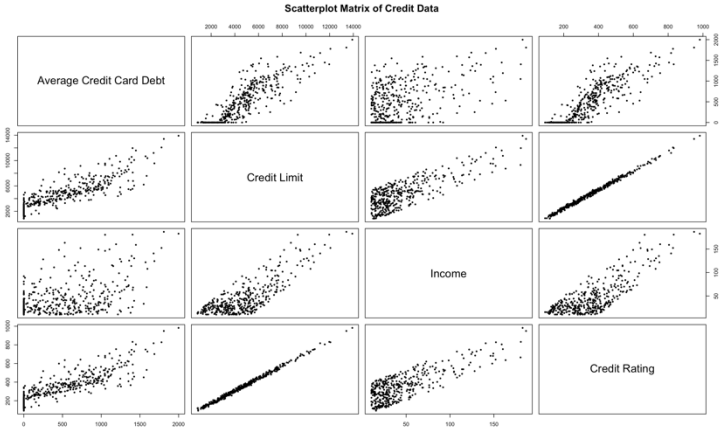
For the homoscedasticity assumption, we should check Plot E which is our residual plot. We can see that the residuals have constant variance throughout the graph. This supports the homoscedasticity assumption.

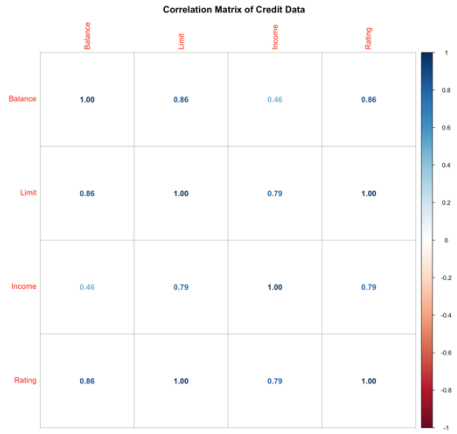
Similarly, we can do the BP-Test like the following:

Null Hypothesis (H_0): Homoscedasticity of errors
Alternative Hypothesis (H_A): Heteroscedasticity of errors
Significance level α → 0.05 (default value)
Test Statistic → 1.881
P-Value → 0.597
P-Value < α → Do not reject the null hypothesis
Conclusion: At 5% level of confidence, we can conclude that the errors are homoscedastic.

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

- ***Interpreting the scatterplot and correlation matrices***





From the scatterplot and correlation matrices, and the Variance Inflation Factor (VIF) values we can check the predictor variables individually;

Income of the Individual

We can see that income and credit card limit has a correlation coefficient of 0.79. This shows a semi-strong positive relationship between these two predictor variables. Similarly, we can see the scatter plot of Credit Limit v. Income which shows us the semi-strong linear relationship.

Similarly, income and the credit card rating have a correlation coefficient of 0.79. This shows a semi-strong positive relationship between these two predictor variables. Similarly, we can see the scatter plot of Credit rating v. Income which shows us the semi-strong linear relationship.

The *VIF* value for the Income is 2.687 which shows moderate correlation between Income and other predictor variables (Rating and Credit Limit). **From the matrices and VIF value, we can conclude a moderate multicollinearity problem.**

Credit-Rating of the Individual

We can see that credit-rating and credit card limit has a correlation coefficient of 1.00. This shows an extremely strong positive relationship between these two predictor variables. Similarly, we can see the scatter plot of Credit-Rating v. Credit Card limit which shows us an extreme linear relationship, as well.

On the other hand, credit rating and income has a correlation coefficient of 0.79. This shows a semi-strong positive relationship between these two predictor variables. Similarly, we can see the scatter plot of Credit-Rating v. Income which shows us the semi-strong linear relationship.

The *VIF* value for the Rating is 160.708 which shows a critical level of multicollinearity between Rating and other predictor variables (Income and Credit Limit). **From the matrices and VIF value, we can conclude a strong multicollinearity problem.**

Credit-Limit of the Individual

We can see that credit-limit and credit-rating has a correlation coefficient of 1.00. This shows an extremely strong positive relationship between these two predictor variables. Similarly, we can see the scatterplot of Credit-Limit v. Credit Rating which shows us an extreme linear relationship, as well.

On the other hand, credit limit and income have a correlation coefficient of 0.79. This shows a semi-strong positive relationship between these two predictor variables. Similarly, we can see the scatter plot of Credit Limit v. Income which shows us the semi-strong linear relationship.

The *VIF* value for the Credit-Limit is 161.193 which shows a critical level of multicollinearity between credit-limit and other predictor variables (Income and Rating). **From the matrices and VIF value, we can conclude a strong multicollinearity problem.**