Assignment 4 – Regression & Association Rules

2181MATH4990 Mateen Shaikh

Due: March 19 by 3:30pm

Instructions:

Complete the following assignment, answering all questions and providing plots when as appropriate. The total length of the assignment should not be more than 5 pages. Make sure that you don't just copy/paste output in your assignment.

Marks will be deducted for embedded code/output. Submit code in a separate document.

autompg

Redownload the autompg data and

- remove the column indicating the type of car
- change the variable corresponding to orig with one that is a factor variable

Do not remove the missing observatiosn (R will perform point-wise deletion as appropriate).

1. Estimate the average effects of the variables with mileage. Choose one predictor variable and comment on the corresponding slope, interpretation of slope, and its significance. Ensure that when you comment on the significance, you explain what this means to someone in the context of vehicles who doesn't understand significance. Also comment on the overall fit of the model. Do not copy/paste any R output, you'll receive 0 for the question.
2. Because the response is strictly positive, we can consider the log-linear model by modelling the logarithm of mileage (still use all variables). Compare the results with the previous model including the overall fit.
3. Repeat steps 1 and 2 with liters per 100km rather than with miles per gallon (see assignment 1 for the conversion). You may do wish to use a scatterplot matrix but I am particularly interested with
   - the results of lp100km with mpg
   - the results of the logarithm of lp100km with mpg

- the results of the logarithm of lp100km with the logarithm of mpg

4. One of the above relationships is perfectly linear. If you understand the mathematics, explain why this is so. If you are uncomfortable with the mathematics, just report the slope and intercept of this line (you can use the lm function for this too!)

Missing Items & association rules

5. How does maintaining missing values change support of some items? To answer this, I recommend you consider a simpler case:

A	B
0	0
0	1
1	0
1	1
1	?

Comment on the support, confidence, and lift of the rule $A \Rightarrow B$, when

a. the entire transaction is deleted (all calculations performed assuming #transactions=n=4)

b. the missing datum is point-wise deleted (n=5 for $P(A)$ and $n = 4$ for $P(B)$ and $P(AB)$)

c. the final transaction is included, (all calculations performed assuming n=5). Note that $P(B) = 0.4$ in this case and equivalent to treating the missingness as another category.

Congressional Voting

Download the congressional voting data from the UCI machine learning database:

http://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records

The data set is in the 'Data Folder' shown at the top and has a .Data extension. However, this is just a plain text csv file. Read this data in noting that there are no headers in the file (in R use `header=F` as an option).

Note on missing data: ***to simplify my marking I require you to treat missing data as a third category*** for each voting outcome. Again, this is to making marking easier; context determines which course of action is most appropriate when dealing with missing data.

6. This data set has the character '?' for missing values. Read the information about what the missing values represent from the UCI webpage and appropriately cite some source (even if it's just wikipedia) on what abstention is and why it's used. For these data, justify your opinion on the most appropriate course of treating the missing data.

7. Mine association rules with at most 3 items, and minimum support of 0.2 and the default confidence of 0.8. Summarize the 4 rules with highest values of lift and comment on the items most often in the rules. Change the confidence threshold to 0.75 and 0.85 and note how the top rules change accordingly.