STAT 315 Assignment 2

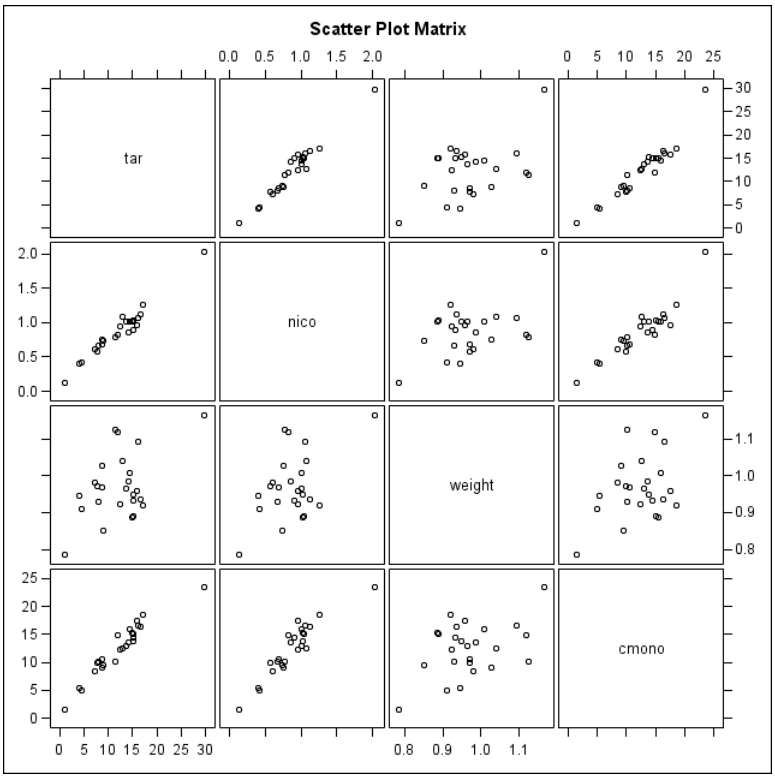
May 18, 2014

Lab: 31 March. Due Date: 11 April, 4pm. Submit via Learn. Maximum possible marks: 12.
Download the files `cigarettes.txt`, `Cigarettes.sas`, `galapagos.txt` and `Tortoise.sas`. Save them in the folder `P:\stat315`.

1. The file `cigarettes.txt` contains data on tar content(mg), nicotine content(mg), weight(g) and carbon monoxide content(mg) for 25 brands of cigarettes. The data were collected by the US Federal Trade Commission.

   (a) Run the SAS script to load in the data, and look at the output of PROC CORR. Based on the correlation matrix and/or pairs plot, which variable do you think is likely to be the best predictor of carbon monoxide content (`cmono`)? Explain your answer. Would you classify any of the observations as outliers? (2)

   Here is the pairs plot:



   Both tar and nicotine are highly-correlated with cmono. They have correlations with cmono of 0.96 and 0.93 respectively. So either of these variables is probably the best predictor. [1 mark, explanation required]

One observation, Bull Durham, is an outlier with unusually high values for all variables (note that it is possible for an observation to be a multivariate outlier even if its individual values look unremarkable.) [1 mark]

Here several people said that it wasn't really an outlier because it "fits into the pattern". This doesn't really have to do with being an outlier or not, which is a vague statement about how "far" it is from the other data points.

(b) Perform the multiple regression and examine the plots produced by SAS. Is there evidence of collinearity? (1)

Yes, both nico and tar have a vif in excess of 21. A vif of 5 or more is usually considered to be suspicious. There is probably some collinearity among the predictors. (In fact, the correlation between nico and tar is very high, as we've already seen that it's 0.977.) [1 mark]

(c) Perform stepwise model selection using the supplied code. What is the smallest value of Mallows' C_p found by SAS? Which variables are in the final model? (2)

The smallest value of C_p is the top row of the table outputted by SAS. It is 0.4672 [1 mark]. The same table tells us that the only variable in the final model is tar. [1 mark]

2. The program `Tortoise.sas` reads in the file `galapagos.txt`. This file contains the counts of number of species of tortoise on 30 Galapagos islands, based on the following variables:

- *Endemics*; whether or not over 50 endemic species are present.
- *Area*; Area of the island in hectares.
- *Elevation*; Elevation in metres.
- *Nearest*; Distance in km to closest island.
- *Scruz*; Distance to Santa Cruz in km.
- *Adjacent*; Area of adjacent island in hectares.

(a) What is R^2 for the regression of *Species* on the other variables? (1)

This is given in the SAS output. It's 0.9035. [1 mark]

(b) What is the mean squared error for leave-one-out cross-validation on this data set (remember to divide PRESS by the number of data points to get the mean squared error)? Based on this value, do you think that the regression gives a good prediction for *Species*? (2)

PRESS is 551818. This is the predicted sum of squared errors calculated from LOOCV. You need to divide by 30 to get the predicted mean squared error on a new observation. This is $551818/30 = 18393.93$. [1 mark] This is the *squared* error, so you would expect the error for an unseen observation to be about $\pm\sqrt{18393.93} = \pm 135.6$. That is a very wide margin of error, which suggests the model does not predict well. [1 mark] *Remark: the reason why PRESS is not widely used is that it doesn't come with a standard error, or any indication of how accurate it is. Cross-validation is better for this reason.*

(c) The rest of the code splits off ten islands to use as a validation set and fits the model to the rest. What is the mean squared error on the validation set? (1)

It's given as mean squared residual in the output: 1264.15. [1 mark]

(d) What is the mean squared error if `seed=100` in PROC SURVEYSELECT is changed to `seed=101`? The effect of this is to choose a different random sample. (1)

Similar to the last part, 89595.77 [1 mark]

(e) Try setting the seed to a few other values. Why do you think the mean squared error on the test set varies so wildly from sample to sample? If the model is so bad at prediction, how come it was able to achieve such a high R^2 (Hint: look at the plot of Species vs. Predicted Value produced by PROC REG)? (2)

The mean squared error varies wildly from sample to sample because there are some very big islands. You will notice that when these islands are in the validation set, the mean squared error on the validation set tends to be bigger. For example, if Isabela is in the validation set then it usually has a huge residual. [1 mark] The model was able to achieve such a high R^2 because R^2 is the ratio of the variance of the predicted y–values divided by the variance of the actual y–values. In this case, the predicted y–values for the big islands are also big and those for the small islands are small, hence there is a "good" R^2. In a sense, the model is good at predicting whether an island will have a lot of species or just a few, but it can't predict the exact number of species very accurately at all. [1 mark]

Some people said that R^2 is high because there are so many predictors compared to the number of data points. I decided that this is also a reasonable answer.

(f) (Bonus point) As a data analyst, what would be your first step towards improving this model? (1)

There are many possible answers. Some people suggest breaking off the group of big islands and making a separate model for them. The problem with this approach is that when you remove outliers, there always seem to be new outliers in the remainder of the data. Also, removing data is a bad idea here as we have so little.

The most obvious thing to do is to take logs (of everything except endemics), as almost all the variables in this data set are positive by definition. (Distance to Santa Cruz is an exception as it contains a zero value.) If you know GLMs, you might also consider a Poisson regression, since you want to predict a count. But the *first* step is definitely to take logs of everything and look at the logged data.

Since this is a bonus question, no marks will be awarded for other answers, although they might well be sensible. [1 mark]

Some people suggested throwing out some of the data. This is a bad idea, especially since we had so little to begin with.

Some people suggested deleting predictor variables. This is a good idea, but I would not want to do it without transforming the data first.

In this question, one student actually looked up the Galapagos Islands in Wikipedia and used that information in an answer, which is an excellent idea.