STAT 315 Assignment 2

March 28, 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)

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

   (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)

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)

   (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)

   (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)

   (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)

   (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)

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