STAT 315 Assignment 3

April 10, 2014

Lab: 12 May. Due Date: 16 May, 4pm. Submit via Learn. Maximum possible marks: 15.
Save the files `Xenon.txt` and `Nordic.txt` and the SAS file `a3code.sas` in the folder `P:\stat315`.

1. This question builds on your knowledge of SAS from Assignments 1 and 2. It is important to know how to interpret SAS output in a way that would make sense to a non-statistician.

   The file `Xenon.txt` contains measurements of pressure (kPa), temperature (millions of degrees celsius) and volume (cubic metres) for samples of 150g of Xenon from a star.

   (a) Run the SAS script to load in the data, and perform the regression of T on P and V. The R^2 is quite high. Do you have any reservations about the assumptions of linear regression here? Which plot or plots look(s) suspicious? (2)

   (b) Run the `proc gam` code and look at the output. The model fitted by SAS is

   $$T = \beta_0 + \beta_1 P + \beta_2 V + \beta_3 s(P) + \beta_4 s(V)$$

   where the $s(P)$ and $s(V)$ are smoothing terms. Only one of these terms is significantly different from zero at the $\alpha = 0.05$ level. Which one? (1)

   (c) Now create a new data set `Xenon2` with a new variable `PV`, defined as the product $P \times V$. (Look at the SAS code for Assignment 1 to see how to do this.) Run another regression with `PV` added to the linear model from part (a). Give the SAS code for this part in your solution. (2)

   (d) Using only the SAS output, which of the three models do you think is the best? How would you explain your choice to a non-statistician scientist? Which plot or plots would you show them? (3)

2. The file `Nordic.txt` contains the result of the Sochi 2014 Nordic Combined 10k/Normal Hill event. The competition is decided by who performs the best in a combination of ski jumping and cross-country skiing. The variable `SkiJump` is the ski jump score and `CrossCountry` is the cross-country time in seconds. Source: `http://www.sochi2014.com/en/nordic-combined-ind-gund-nh-10-km-cross-c-free-race`

   (a) One way of combining the scores is to use the first principal component. Why might this be a good idea? (1)

   (b) If the competitors were ranked based on the first principal component, who would have won the bronze medal? (1)

   (c) What do you think the *second* principal component represents? Are the data adequately summarized by one principal component? (2)

   (d) The IOC wants to introduce a new snowmobile half-pipe event and is considering dropping the Nordic combined on the grounds that ability in cross-country skiing and ski jumping are more or less equivalent. Do you think this is reasonable? Explain your answer referring to the SAS output. (1)

   (e) Would it be better to run a PCA on the covariance matrix instead of the correlation matrix in this example? Who would be the gold medallist in that case? (2)