ST 518 Spring 2017: Homework 3

You can submit one R markdown file with your solutions, or you can submit a .pdf file that includes the answers to the content questions and a .R file with any R code that you used to get your solutions.

Please feel free to discuss questions (without actually sharing your answers) on the discussion board.

R Question

1. (4 points) A researcher is interested in studying how (if?) GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of undergraduate institution are associated with admission into graduate school. The response variable, Admit, is a binary variable (1 == admit). This dataset can be found in `admissions.csv` available for download right underneath where you downloaded/opened this homework assignment. Treat the variables `GRE` and `GPA` as continuous, and treat `RANK`, which takes values 1 through 4, as a factor variable. A rank of 1 indicates that the student's undergraduate institution has the highest prestige, while a rank of 4 indicates that it has the lowest prestige.
   (a) Fit a logistic regression model to these data, with the variable `admit` as the response and `gpa`, `gre`, and `rank` as explanatory variables. Fit another model without `gre`. Comment on how these models are different.
   (b) Produce a scatterplot of `admit` against `GPA` and overlay 4 separate fitted lines, one for each rank, from the regression with `gpa` and `rank` as explanatory variables.
   (c) Write a short paragraph discussing your findings.

Conceptual Questions

2. (3 points) In 1846, the Donner and Reed families left Springfield, Illinois, for California by covered wagon. Along the way, more families and individuals joined the Donner Party, as it came to be known, until it reached its full size of 87 people. The group become stranded in the eastern Sierra Nevada mountains when the region was hit by heavy snows in late October. By the time the last survivor was rescued in April 1847, 40 of the 87 members had died from famine and exposure to extreme cold.
   (a) One assumption underlying the correct use of logistic regression is that observations are independent of each other. Is there some basis for thinking this assumption might be violated in the Donner Party data?
   (b) Why should one be reluctant to draw conclusions about the ratio of male and female odds of survival for Donner Party members over 50? ($Hint$: Look again the graph of the Donner Party data from lecture, where status is plotted against age.)
   (c) In this week's lecture, it was found that the estimated logisitic regression equation is:

$$\hat{\text{logit}}(p) = 1.63 - 0.078Age + 1.60Female$$

   where $Female$ is an indicator variable equal to one for females and zero for males. What is the age at which the estimated probability of survival is 50% for women? What about for men?

3. (3 points) The common brushtail possum of the Australia region is a bit cuter than its distant cousin, the American opossum. We consider 104 brushtail possums from two regions in Australia, where the possums may be considered a random sample from a larger population. The first region is Victoria, and the second region consists of New South Wales and Queensland.

   We use logistic regression to differentiate between possums in these two regions. The outcome variable `population` takes value 1 if the possum is from Victoria and 0 if it is from New South Wales and Queensland. Five predictors are considered: `sex_male`, an indicator for a possum being male, `head_length`, `skull_width`, `total_length`, and `tail_length`. A full and reduced logistic model are summarized in the following table:

	Full Model				Reduced Model			
	Estimate	SE	Z	Pr($>\|Z\|$)	Estimate	SE	Z	Pr($>\|Z\|$)
(Intercept)	39.2349	11.5368	3.40	0.0007	33.5095	9.9053	3.38	0.0007
`sex_male`	-1.2376	0.6662	-1.86	0.0632	-1.4207	0.6457	-2.20	0.0278
`head_length`	-0.1601	0.1386	-1.16	0.2480				
`skull_width`	-0.2012	0.1327	-1.52	0.1294	-0.2787	0.1226	-2.27	0.0231
`total_length`	0.6488	0.1531	4.24	0.0000	0.5687	0.1322	4.30	0.0000
`tail_length`	-1.8708	0.3741	-5.00	0.0000	-1.8057	0.3599	-5.02	0.0000

   (a) The variable `head_length` was taken out for the reduced model based on its p-value in the full model. Why did the remaining estimates change between the two models?

   (b) Suppose we see a male possum with a 65 mm wide skull, a 32 cm long tail, and a total length of 80 cm. If we know this possum was captured in the wild in Australia, what is the probability that this possum is from Victoria (using the reduced model)?