Homework Assignment 7: Red Brain, Blue Brain

36-402, Advanced Data Analysis, Spring 2015

Due at 11:59 pm on Monday, 23 March 2015

The data set n90_pol.csv contains information on 90 university students who participated in a psychological experiment designed to look for relationships between the size of different regions of the brain and political views. The variables `amygdala` and `acc` indicate the volume of two particular brain regions known to be involved in emotions and decision-making, the amygdala and the anterior cingulate cortex; more exactly, these are residuals from the predicted volume, after adjusting for height, sex, and similar anatomical variables. The variable `orientation` gives the subjects' locations on a five-point scale from 1 (very conservative) to 5 (very liberal). `orientation` is an ordinal but not a metric variable, so scores of 1 and 2 are not necessarily as far apart as scores of 2 and 3.

1. (10) *Joint density of brain regions*

   (a) (5) Using `npudens`, estimate a joint probability density for the volumes of the amygdala and the ACC. What are the bandwidths?

   (b) (5) Plot the joint density. Does it suggest the two volumes are statistically independent? Should they be? You you may use three dimensions, color, contours, etc., for your plot, but you will be graded, in part, on how easy to read it is.

2. (20) *Predicting brain sizes from political views*

   (a) (1) Ignoring the fact that `orientation` is an ordinal variable, what is the correlation between it and the volume of the amygdala? Between `orientation` and the volume of the ACC?

   (b) (4) Using case resampling, give 95% bootstrap confidence intervals for these correlations.

   (c) (2) The function `rank`, applied to a data vector, returns the vector of ranks, where 1 indicates the smallest value, 2 the next-smallest, etc. What are the correlations between the ranks of `orientation` and the ranks of `amygdala`? Between `orientation` and `acc`? *Hint:* What does `cor(x,y,method="spearman")` do?

   (d) (3) Using case resampling, give 95% bootstrap confidence intervals for the rank correlations.

(e) (10) Using `npcdens`, plot the condition density of the volume of the amygdala as a function of political orientation. Do the same for the volume of the ACC. Make sure that in both cases you are treating `orientation` as an ordinal variable. You will be graded on how easy your plots are to read.

3. (5) *Creating a binary response variable*

   (a) (1) Create a vector, `conservative`, which is 1 when the subject has `orientation` ≤ 2, and 0 otherwise.

   (b) (2) Explain why the cut-off was put at an `orientation` score of 2 (as opposed to some other cut-off).

   (c) (1) Check that your `conservative` vector has the proper values, *without* manually examining all 90 entries.

   (d) (1) Add `conservative` to your data frame. (Creating a new data frame with a new name will only get you partial credit.)

4. (10) *Logistic regression*

   (a) (5) Fit a logistic regression of `conservative` (not `orientation`) on `amygdala` and `acc`. Report the coefficients to no more than three significant digits. Explain what the coefficients mean.

   (b) (5) Using case resampling, give bootstrap standard errors and 95% confidence intervals for the coefficients. Was the restriction to three significant digits reasonable?

5. (10) *Generalized additive model.* Fit a generalized additive model for `conservative` on `amygdala` and `acc`. (Be sure to smooth both the input variables.) Make sure you are using a logistic link function. Report the intercept with reasonable precision. Plot the partial response functions, and explain what they mean (be careful!).

6. (15) *Kernel conditional probability estimation*

   (a) (5) Using `npcdens`, find the conditional probability of `conservative` given `amygdala` and `acc`. Make sure `npcdens` treats `conservative` as a categorical variable and not a continuous one. Report the bandwidths.

   (b) (5) Plot the estimated conditional probability that `conservative` is 1, with `acc` set to its median value and `amygdala` running over the range $[-0.07, 0.09]$. (The plotting range for `amygdala` exceeds the range of values found in the data.) *Hint:* your code will need to provide values for `acc`, for `amygdala` *and* for `conservative` (why?).

   (c) (5) Plot the estimated conditional probability that `conservative` is 1, with `amygdala` set to its median value and `acc` running over the range $[-0.04, 0.06]$. (This plotting range also requires extrapolating outside the data.)

7. *Classification* (15) The models from problems 4–6 predict probabilities for `conservative`. If we have to make a point prediction of whether someone is conservative or not, we should predict 1 if the probability is ≥ 0.5 and 0 otherwise.

   (a) (7) Find such predictions for each subject, under each of the three models. What fraction of subjects are mis-classified? What fraction would be mis-classified by "predicting" that none of them are conservative?

   (b) (8) Re-calculate the classification error rates using leave-one-out cross-validation for each model.

RUBRIC (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Figures and tables are easy to read, with informative captions, axis labels and legends, and are placed near the text of the corresponding problems. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision. Code is either properly integrated with a tool like R Markdown or knitr, or included as a separate `.R` file. In the latter case, the code is clearly divided into sections referring to particular problems. In either case, the code is indented, commented, and uses meaningful names. All code is relevant to the text; there are no dangling or useless commands. All parts of all problems are answered with actual coherent sentences, and never with raw computer code or its output. For full credit, all code runs, and the Markdown file knits (if applicable).