Homework 8: Red Brain, Blue Brain?

36-402, Spring 2025

Due at 6:00 pm on Thursday, 20 March 2025

> AGENDA: More practice with classification models, including yet more logistic regression; density estimation; conditional densities.
>
> TIMING: Problems 1–3 and 6 involve fitting models to data, plotting, and interpretation, but no coding. Problem 5 requires doing all that and some bootstrapping, for which you will need to write a little code (along lines you have done before). Problem 7 requires fitting a model and making some plots from it, and you will (probably) need to write a little code, along the lines of examples in the book, to do so. Problem 8 requires comparing models, and you will need to either write some new code, or tweak some example code, to do 8b. The solutions to all problems take about 5 minutes to knit without a cache (and about two seconds with a cache — cache everything!).

The data file at `http://www.stat.cmu.edu/~cshalizi/uADA/25/hw/08/n90_pol.csv` contains information on 90 British university students who participated in a study looking 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 variables are residuals from the predicted volume, after adjusting for height, sex, and similar anatomical variables[1]. The variable `orientation` gives the subjects' self-reported locations on a five-point scale from 1 (very conservative) to 5 (very liberal). `orientation` is an ordinal but not a metric variable, so the difference between a score of 1 and score of 2 might or might not be as big as the difference between scores of 3 and 4.

1. *Marginal density of brain region volumes*
   (a) (4) Using `npudens`, estimate the probability density for the volume of the amygdala. Plot it and report the bandwidth.
   (b) (4) Repeat this for the volume of the ACC.
2. *Joint density of brain regions*

[1] The units are *probably* cubic centimeters.

(a) (5) Using `npudens`, estimate a joint probability density for the volumes of the amygdala and the ACC. What are the bandwidths? Are they the same as the bandwidths you got in problem 1? Should they be?

(b) (5) Plot the joint density. Does it suggest the two volumes are statistically independent? Should they be? 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.

Hint: Remember that the random variables X and Y are statistically independent when their joint pdf is the product of their marginal pdfs, $p(x, y) = p(x)p(y)$. Think about what the product of your estimated pdfs from problem 1 would look like.

3. *Predicting brain sizes from political views*

   (a) (5) Using `npcdens`, find the conditional density of the volume of the amygdala as a function of political orientation. (Make sure that you are treating `orientation` as an ordinal variable.) Report the bandwidths. Is the bandwidth for the amygdala the same as either of the previous two bandwidths you have found for it? Should it be? Plot the distribution, and comment on whether it suggests any relationship between the size of this brain region and political orientation.

   (b) (5) Repeat this for the conditional density of the ACC as a function of orientation.

4. *Creating a binary response variable* Add a column, `conservative`, to your data frame, which is 1 when the subject has `orientation` $\leq$ 2, and 0 otherwise.

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

5. *Logistic regression*

   (a) (3) 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) (3) Using case resampling, give bootstrap standard errors and 95% confidence intervals for the coefficients. Was the restriction to three significant digits reasonable?

6. *Generalized additive model.* (4) 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!).

7. *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.)

8. *Classification* The models from problems 5–7 predict probabilities for `conservative`. If we have to make a definite prediction of whether someone is conservative or not, we should predict 1 if the probability is ≥ 0.5 and 0 otherwise.

   (a) (3) Classify 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) (6) Re-calculate the mis-classification rates using leave-one-out cross-validation for each model.

9. (5) *Interpretation* Write at least a paragraph, but no more than one page, on what we can conclude about the relationship between brain anatomy and political views — at least based on this data, in this population. Support your conclusions by referring to specific findings of your various analyses.

10. (1) *Timing* How long, roughly, did you spend on this assignment? How much of that time was spent on math, on coding/debugging, and on writing?

Presentation Rubric (15): 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. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. 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. All parts of all problems are answered with actual coherent sentences, and raw computer code or output are only shown when explicitly asked for. Text from the homework assignment, including this rubric, is included only when relevant, not blindly copied.

(In Gradescope, assign *all* pages to this rubric.)

CODE RUBRIC (15): The code is logically organized and easy to read. No redundant code; no needlessly repetitive code; no unused code. Variables and functions have descriptive and appropriate names. (Loop or array indices, arguments, etc., can have short, conventional names such as `i` `x`, `df`, etc.) All functions have comments defining their purpose, their inputs, their outputs, and any dependencies on other code you wrote. Vectorization is used wherever appropriate. Allowed packages: `knitr`, `tidyverse`, `dplyr`, `ggplot2`, and those explicitly mentioned in the textbook or the assignment for implementing particular methods. (Any other packages require prior permission from the professor, which must be renewed for each assignment; record the date on which you got permission in your comments.) Code from the textbook and class examples is used wherever possible and appropriate. In particular, it should be used for tasks like bootstrapping, calibration plots, and cross-validation (*unless* the package implementing a model includes its own cross-validation functions). All plots and tables are generated by code included in the R Markdown file. Numerical results (etc.) appearing in text are neither hand-copied nor spat out with `cat()`, `print()`, `sprintf()` etc., but instead properly formatted through in-line code.

(Do not assign any pages to this rubric; instead, submit your Rmd file to the "HW 8 R Markdown File" assignment on Gradescope.)