Problem Set 4

Statistics 506, Fall 2017

Due: Tuesday December 5 at 9am

Instructions

- Use Rmarkdown or another proram to create PDF files (one per question) containing your answers to each of the questions posed below. Embed your R code either by setting "echo = TRUE" or by using fenced code blocks "````". Embed SAS code and PBS scripts within code blocks.
- Question 2 is optional. If choose to complete it, points earned will be added to your homework score. You can use this to make up for any questions you've struggle with in the past.
- Your code should be clearly written and it should be possible to assess it by reading it. Use appropriate variable names and comments. Your style will be graded using the style rubric (./StyleRubric.html).
- All work for problems 1 and 3 should be done using SAS.
- Some of these exercises will require you to use commands or techniques that were not covered in class or in the course notes. You can use the web as needed to identify appropriate approaches. Questions are welcome at all times, but please make an attempt to locate relevant information yourself first.

Questions

1. In this question you will use SAS to fit mixed models to the audiometry data from the 2009 NHANES survey previously used for problem set 1, question 3. As before you may treat this as a simple random sample. You may wish to refer to the previous solution when specifying the mixed models.

a. Determine how to load the audiometry and demographic data into SAS and then merge on the common identifier `seqn` . Drop all cases without audiometry data.

b. Produce a reduced data set in long format containing columns for:
   - the unique identifier `seqn`
   - [demographics] age and gender
   - [hearing threshold tests] ear, frequency, and threshold or result
   - An age group indicator for whether the person is older than 25 years of age.

   For each person use just the first test at each frequency for each ear.

c. Filter your data to contain only the 1000 Hz test for the right ear so that each unique id appears just once. Use `proc reg` to fit regression models for answering the following questions:

   i. At this frequency, is there a significant interaction between age group and gender in determining how well an individual hears form their right ear?

   ii. After controlling for age group and gender, is age still important as a continuous variable?

iii. Is the effect of age, as a continuous variable, significantly different among the older and/or younger age groups?

d. Answer the questions from part "c" using data at all frequencies and both ears using `proc mixed`.

---

2. [Optional] Repeat question 1 using R and the `lmer` function from the `lme4` package for part "d". Use `lm` for part "c".

---

3. This question repeats question one from problem set 4 (http://dept.stat.lsa.umich.edu/~kshedden/Courses/Stat506/ps4/), assigned by Professor Shedden last year. You will need to read the first part of the question for data details. Solutions to parts b and c using the `data.table` package in R are available there. You may wish to refer to his solutions in constructing your own.

a. Read the data into SAS using the script provided with the data download.

b. Use a "data" statement to create a new variable called totpay that contains the amount of money paid from Medicare to the provider for the services represented in each line of data. You will need to look at the variable definitions to figure out how to do this.

c. Use `proc sql` to answer the following questions:

i. Using the data you constructed in part (b), obtain the average cost per service for each type of medical service. Identify the service with the highest average cost, and the service with the highest average cost among all services provided 100,000 or more times.

ii. Restrict the data to individual providers (see the data documentation for details). Determine the total amount of money paid to each provider. Then do the following two analyses with this data:

- Restrict to all providers who charged more than 1 million dollars in total, create a frequency distribution of the provider types within this set of providers, and report the 10 most frequent provider types;
- Average the total amount paid to the providers within each provider type, and report the provider types with the two highest and two lowest averages.

---

4. In this question you will use parallel computing on Flux to estimate out-of-sample prediction error using cross-validation in R. To get started download a simulated data frame containing two columns 'x' and 'y' from here (./ps4_q4.RData).

a. Install the package `mgcv` and read the documentation for the `gam` function. Then use logistic regression to predict 'y' as a smooth function of 'x'.

b. Write a function `xvalidate` with an argument `folds` to estimate the out-of-sample prediction error for the model you fit in "a" using k-fold cross validation. Use a sequential `for` loop to compute each sub model and the `predict.gam` method for computing predictions on the out-of-sample fold. What is the approximate out-of-sample prediction error for this model using 10-fold cross validation?

c. Add an option "cores" to the function you wrote in part "b". Modify your function so that when "`cores > 1`" the prediction error on each of the k-folds is computed in parallel.

d. Choose a setting for `cores` between 2 and 8. Write a PBS script to estimate the out-of-sample prediction error using that many cores on Flux. Submit and run your job.

e. Modify your scripts from part d to work with a PBS job array, where the array id is used to indicate the number of folds to be passed to your function. Run the array for k = 10, 100, 1000, and 10000 (i.e. leave-one-out cross validation). Does the number of folds have a meaningful impact on the estimated prediction error? How does the number of folds impact the run time?