Optional Assignment

Mathematics and Computation

This goal of this assignment is to give you experience in mathematics and computation fundamental to understanding more advanced statistical methods. Use R Markdown to produce an HTML (or PDF) file that includes your responses to each of the questions on this assignment. Please adhere to the following guidelines for further formatting your assignment:

- All graphics should be resized so that they do not take up more room than necessary and should have an appropriate caption. Learn how to do this in a code chunk using knitr syntax.
- Any typed mathematics (equations, matrices, vectors, etc.) should be appropriately typeset within the document using Markdown's display equations. See here for some examples of how mathematics can be typeset in R Markdown.
- All syntax should be hidden (i.e., not displayed) unless specifically asked for. Any messages or warnings produced from loading packages should also be hidden.

For each question, specify the question number using a section header. Submit your HTML (or PDF) file and your RMD file via email that you send to both Jonathan and Andy.

Part I: Matrix Algebra and Regression

Use the following data to answer the questions in this section.

```
##    wage age sex
##   12.00  32   M
##    8.00  33   F
##   16.26  32   M
##   13.65  33   M
##    8.50  26   M
```

Consider the regression model that includes an intercept, and the main-effects of `age` and `sex` to predict `wage`. The `lm()` formula for this model would be `wage ~ 1 + age + sex`. For consistency, use females as the reference group.

1. Write out the design matrix (i.e., the X-matrix) for the model.
2. What are the dimensions of the design matrix?
3. Using matrix algebra, compute and report the **b** vector (i.e., the vector of the regression coefficients). Show any relevant work.
4. Using matrix algebra, compute and report the standard errors for each of the regression coefficients in the model. Show any relevant work.
5. Using the values from Questions 3 and 4, compute and report the t-statistic for each of the regression coefficients. Show any relevant work. You may need to refresh your memory about what how a t-value is computed. One place to start may be your introductory statistics textbook, or any of a number of websites online.
6. Use the `pt()` function to compute the p-value (two-sided) for each of the regression coefficients. Show any relevant work. Again, you may need to refresh your memory about what a p-value is, and how they are computed.

Now consider the regression model that includes an intercept, the main-effects of `age` and `sex`, and the interaction effect between `age` and `sex` to predict `wage`. The `lm()` formula for this model would be `wage ~ 1 + age + sex + age:sex`. For consistency, use females as the reference group.

7. Write out the design matrix for the model.
8. Try to compute the **b** vector. You get an error message saying: `Error in solve.default(t(X) %*% X) : system is computationally singular`. Explain, using the language of matrix algebra, what this error message means.

Part II: Computation and Bootstrapping: Mean

Read The Bootsrap Method for Standard Errors and Confidence Intervals form *Biostatistics for Dummies*. You might want to actually do do the experiment they outline (at least a few times) to better understand the procedure.

9. You are going to write an R function that bootstraps the mean. This function should take as inputs (1) a vector of data (e.g., the IQ scores mentioned in the article), and (2) the number of bootstrap replications you want to carry out. The function should give as output (1) the estimated bootstrap standard error and (2) the 2.5th and 97.5th centiles (this is called a 95% percentile interval). It should also store the bootstrapped means so someone can access them (but not print them as direct output). Provide the function's syntax in your Markdown document.

To make sure you are on the right track with this, using the original data, compute the standard error for the mean by computing the standard deviation of the IQ scores divided by the square root of the sample size. Also compute the 95% confidence interval for the mean using the `t.test()` function. The bootstrapped SE and percentile interval should be similar to these values.

10. Run your function using the IQ data and 500 bootstrap replications. Plot the distribution of the bootstrapped mean values. Annotate your plot to indicate the orignal data's mean (i.e., the sample mean), and the lower and upper limits of the percentile interval.
11. The percentile interval for the mean is called a *symmetric interval.* Explain, by referring to the plot and annotation in Question 10, why this is called a *symmetric interval.*
12. Compute the confidence interval by adding and subtracting two standard errors (use the SE you computed from your function) to the original data's mean. How do the limits on the CI compare to the limits of the percentile interval.

Part II: Computation and Bootstrapping: R^2

13. Use the *homework-education-gpa.csv* data set to fit a regression model that uses time spent on homework (`homework`) and parent education level (`parentEd`) to explain variation in students' GPAs (`gpa`). Indicate the value of the estimated R^2 statistic.

You are going to bootstrap R^2. The algorithm is simlar to that which you used for the mean, with some slight modifications. Akin to that described in the *Biostatistics for Dummies* article, the algorithm would be:

- Write the measurements (time spent on homework, parent education level, and GPA) for a student on a slip of paper. Use a separate slip of paper for each student. (Each slip of paper should have three measurement, and there should be as many slips of paper as there are students.)
- Put all the slips of paper into a bag. Reach in and draw out one slip, write the three measurement down, and put the slip back into the bag. Repeat this as many times as needed to match the number ofstudents there were in the original sample.

- Fit the same regression model to this bootstrap sample that you fitted to the original data. Calculate R^2 and record it.
- Repeat this process many, many times.
- Compute the standard deviation of the R^2 values. This is called the *standard error* of R^2.
- Compute the 95% percentile interval for R^2 in the same manner you did for the mean.

14. Write a function that bootstraps $R2 from a simple regression model. This function should take as inputs (1) a data frame, and (2) the number of bootstrap replications you want to carry out. (If you really want to get crazy, it could also take the `lm()` formula. This would make it more generalizable to other data sets.) The function should give as output (1) the estimated bootstrap standard error and (2) the 2.5th and 97.5th centiles (this is called a 95% percentile interval). It should also store the bootstrapped R^2 values so someone can access them (but not print them as direct output). Provide the function's syntax in your Markdown document.

15. Use your function to carry out 500 bootstrap replications on the data. Plot the distribution of the bootstrapped R^2 values. Annotate your plot to indicate the orignal data's R^2 value, and the lower and upper limits of the percentile interval.

16. Explain, by referring to the plot and annotation in Question 12, why the percentile interval for R^2 is called an *asymmetric interval.*

17. Explain why computing the percentile interval for R^2 is more appropriate for obtaining the interval estimate than computing the confidence interval; adding and subtracting two standard errors to the original data's R^2.