Statistical Methods for Data Science (Spring 2018)

Mini Project 4

Instructions:

- Due date: March 29, 2018.
- Total points = 20.
- Submit a typed report.
- You can work on the project either individually or in a group of no more than two students. In case of the latter, submit only one report for the group, and include a description of the contribution of each member.
- It is OK to discuss the project with other students in the class (even those who are not in your group), but each group must write its own code and answers. If the submitted report (including code and answer) is similar (either partially or fully) to someone else's, this will be considered evidence of academic dishonesty, and you will referred to appropriate university authorities.
- Do a good job.
- You must use the following template for your report:

  Mini Project #
  Name
  Names of group members (if applicable)
  Contribution of each group member
  Section 1. Answers to the specific questions asked
  Section 2: R code. Your code must be annotated. No points may be given if a brief look at the code does not tell us what it is doing.

---

1. In the class, we talked about bootstrap in the context of one-sample problems. But the idea of nonparametric bootstrap is easily generalized to more general situations. For example, suppose there are two *dependent* variables X_1 and X_2 and we have i.i.d. data on (X_1, X_2) from n independent subjects. In particular, the data consist of (X_{i1}, X_{i2}), $i = 1, \ldots, n$, where the observations X_{i1} and X_{i2} come from the ith subject. Let θ be a parameter of interest — it's a feature of the distribution of (X_1, X_2). We have an estimator $\hat{\theta}$ of θ that we know how to compute from the data. To obtain a draw from the bootstrap distribution of $\hat{\theta}$, all we need to do is the following: randomly select n subject IDs with replacement from the original subject IDs, extract the observations for the selected IDs (yielding a resample of the original sample), and compute the estimate from the resampled data. This process can be repeated in the usual manner to get the bootstrap distribution of $\hat{\theta}$ and obtain the desired inference.

   Now, consider the `advertising` data stored in the `Advertising.csv` file available on eLearning. Make scatterplots of `sales` against `TV` and `radio`, and comment on the strength of linear relationship between `sales` and `TV` and `sales` and `radio`. Let ρ_1 and ρ_2 respectively denote the population correlation between `sales` and `TV` and between `sales` and `radio`. For each of the two correlations, provide a point estimate, bootstrap estimates of bias and standard error of the point estimate, and 95% confidence interval computed using percentile bootstrap. Interpret the results. (To review population and sample correlations, look at Sections 3.3.5 and 11.1.4 of the textbook. The sample correlation provides an estimate of the population correlation and can be computed using `cor` function in `R`.)

2. Consider the dataset stored in the file `singer.txt` on eLearning. This dataset contains heights in inches of the singers from a choral society. The data are grouped according to voice part. There are four voice parts, namely, Bass, Tenor, Alto, and Soprano. The vocal range for each voice part increases in pitch from Bass to Soprano.

   (a) Perform an exploratory analysis of the data by examining the distributions of the heights of the singers in the four groups. Comment on what you see. Do the four distributions seem similar? Justify your answer.

   (b) Is there any difference in the mean heights of Alto and Soprano singers? If yes, how much is the difference? Answer these questions by constructing an appropriate confidence interval. Clearly state the assumptions, if any, and be sure to verify the assumptions.

   (c) How does your conclusion in (b) compare with what you expected from the exploratory analysis in (a)?