Bootstrap and Resampling Methods, Spring 2023

Homework exercise 1

Due date: 3 April 2023 (second day of Passover vacation)

Note: You can use R, Python or any other tool. The simple way I know of getting the data used is through the bootstrap library in R. If it is not directly available in Python you can save it from R and then model it in Python.

Submission format: Please include your code in your submission as an appendix.

1. **Sources of variance in bootstrap estimation.**
   In this problem we will examine and compare the variance of the bootstrap estimator due to the original sampling to the variance resulting from using B bootstrap iterations instead of ∞ (the variance in the "Bootstrap world").
   This problem uses the (LSAT,GRE) dataset from the book (recall it is considered a population, not a sample), which is available as the object law82 in the bootstrap package. As in the book we will consider bootstrap estimation of $\theta = \text{se}(\hat{\text{cor}}(\text{LSAT,GRE}))$. The book shows one example of a bootstrap estimate derived using $B = 3200$ bootstrap samples (0.132), which comes out very close to the standard error estimate based on resampling from the "true" population (0.131).

   (a) Repeat the following exercise 20 times:
      i. Draw a random sample of size 15 from the (LSAT,GRE) population
      ii. Generate 50 independent bootstrap estimates of θ based on this sample with each of the following bootstrap sample number: $B \in \{100, 1000, 3000\}$

   (b) Using the 50 repetitions, empirically estimate the variance resulting from the bootstrap estimation within each of the 20 samples and for each bootstrap samples number. Estimate an overall variance "in the bootstrap world" at each samples number B by averaging the 20 estimates at this B

   (c) For the largest samples number $B = 3000$, estimate the variance of the bootstrap estimates "in the real world" by calculating the average of the 50 bootstrap estimates and empirically calculating the variance of the 20 bootstrap averages you obtained

   (d) Compare your estimates from the previous two items and conclude about the sources of variance when performing bootstrap estimation: is most of the variance due to using finite B or due to the randomness of the original sample?

   (e) Assuming that the true population parameter is 0.131, how representative is the bootstrap estimate 0.132 obtained in the book? In other words, with different random samples, is the typical bootstrap estimate this close to the truth?

2. **Problem 7.6 from the book: Comparing PCA with and without scaling**
   This problem uses the test score data used for principal component analysis (object scor),

   (a) First, perform the PCA + bootstrap as we did in class, to get:
      i. PCA estimates of leading eigenvalues and eigenvectors

ii. Use non-parametric bootstrap to calculate $B = 200$ bootstrap estimates of the top two PC's, and estimate the standard error of the percentage of variance explained by the top two eigenvalues, and the standard error of each of the $2 \times 5 = 10$ loadings of these two vectors

(b) Now, we want to compare these results to the analysis on the correlation matrix instead of the covariance matrix (for example, using the option scale=TRUE in the functionprcomp in R)

(c) Compare the results of the two analyses and comment on them

3. **Problem 9.7 from the book: Least squares estimation in semi-parametric vs non-parametric bootstrap**
In class we have shown that when we use the semi-parametric bootstrap for inference on the parameters, the least squares estimate in each bootstrap sample is:

$$\hat{\beta}^* = (C^T C)^{-1} C^T y^*,$$

where y^* is the bootstrap sample generated as we noted in class. If we instead employ the non-parametric approach of resampling $(\mathbf{c}, y)$ pairs in the bootstrap world, how would this estimate change? Write a diagram of the sampling in the bootstrap world in both cases and explain how it affects the calculation above.