ST 453 Advanced computing for statistical reasoning
Homework problem set 4

September 16, 2024

No `R` packages are permitted for use in this assignment.

1. In previous assignments you were asked to propose a statistical model with population features that may be appropriate to estimate from the real data set you have chosen for your class project, and you were asked to consider various estimation procedures and algorithms. To evaluate an estimation procedure and algorithm, you will want to generate synthetic data from your posited statistical model. The "true" parameter values should be set as the parameter estimates from the real data set, and the point of a simulation study of synthetic data is to determine if you are able to correctly estimate the 'true" parameter values on synthetic data. If so, the simulation study will lend credibility to the fitted real data model, and if not, the simulation study will help to identify and correct any mistakes in the estimation procedure/algorithm code or any shortcomings in the model formulation. Using pseudocode, describe how you will be able to generate synthetic data from your posited statistical model.

2. (a) Write an `R` function that takes as input `( y, X)`, where `y` is an n-dimensional vector and `X` is an $n \times p$ matrix, and returns the least squares coefficient estimates by solving the normal equations

$$X'Xb = X'y$$

with Gaussian elimination and the back-substitution algorithms.

   (b) Generate synthetic regression data for various choices of n and p to test whether your least squares estimation procedure in part (a) works. Note, you should compare your least squares solution to the "true" coefficient values that you used to generate the data. Show that quantity $\|\widehat{b} - b\|_2 < 10^{-4}$ for sufficiently large values of n.

   (c) Plot the regression line using the coefficients from part (a), for synthetic simple linear regression data (i.e., $p = 2$ and b_0 is an intercept.)

3. Repeat parts (a), (b), and (c) for problem 2 using the singular value decomposition of $X = UDV'$ to simplify the task of solving the normal equations $X'Xb = X'y$. Do you still need to use the Gaussian elimination and back-substitution algorithms?