ST 495 Advanced computing for statistical methods
Homework problem set 5

February 20, 2024

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

1. In the previous assignment you were asked to provide pseudocode for how to generate synthetic data from the statistical model posited for your midterm project. The motivation was described as synthetic data will be helpful for evaluating an estimation procedure and algorithm. Describe how you will design a simulation study based on your synthetic data to evaluate the estimation procedure and algorithm that you proposed a few weeks ago.

2. Recall that in the case that $X \in \mathbb{R}^{n \times p}$ does *not* have full column rank, $(X'X)^{-1}$ does not exist, and so $P_X = X(X'X)^{-1}X'$ does not exist. However, using the SVD of X we can still construct an orthogonal projection matrix onto $\text{col}(X)$. Write an `R` function that takes as input `(X)`, where `X` is an $n \times p$ matrix, and returns the orthogonal projection matrix onto col(`X`), regardless of the column rank of `X`.

3. Write an `R` function that takes as input `(X)`, where `X` is an $n \times p$ matrix with full column rank, and returns, via the Gram-Schmidt orthonormalization algorithm, an $n \times p$ orthonormal matrix `Q` such that col(`Q`) = col(`X`), along with a $p \times p$ upper-triangular matrix `R` such that `X` = `Q R`.

4. (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

$$X'Xb = X'y$$

   using the QR decomposition of X.

   (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.)