Homework 2 (20 Points)

Due: 2022/10/09

Submission Guidlines

1. You need to write your homework in *Rmd/* ipynb/Rnotebook.
2. Please submit two files: one script file and one generated *pdf file.

In this assignment, you will explore multiple linear regression on a real dataset. Note: when you are performing linear regression, you are **not** allowed to use built-in linear regression modules. Please implement the code by yourself.

1. Median sale prices data for Los Angeles County Housing in Aug 2013 from the Los Angeles Times were compiled into the file LAhousingpricesaug2013.txt. Let Y = sales of single family homes in August, X1 = median price of a single family residence (SFR) in thousands of dollars, X2 = median price of a condo in thousands of dollars, and X3 = median home price per square foot, in dollars. Each of these 4 vectors initially has length 269. If any row has an "n/a" in it for any of these 4 variables, then remove this entire row. Now each vector will have length 217. Please report your code of reading and cleaning the data.
2. Perform regression (with intercept) of Y on X = {X1, X2, X3} to compute a vector of parameter estimates, $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$, where β_0 is the estimated intercept and for i = 1,2,3, β_i is the slope corresponding to explanatory variable Xi. Please report $\hat{\beta}$. Note: you need to implement two methods from the lecture for linear regression: Vectorized Gauss-Jordan Elimination, Sweep Operator, QR with Gram-Schmidt, QR with Householder)
3. Let $i = 1$ Perform regression with intercept of Y on X with row i removed from the dataset. Let $(-i)$ denote your resulting vector of parameter estimates, so that $\beta^{(-i)}$ is your estimate of the slope with i^{th} row dropped. Please record $\beta^{(-i)}$
4. Repeat step 3 for $i = 2, 3, \cdots, 217$ and record $\beta^{(-i)}$.
5. Plot the influences of $\beta_1^{(-i)}$, versus i. That is, the x-axis will span from i = 1 to 217, and the y-axis will be $\beta_1^{(-1)}, \cdots \beta_1^{(-217)}$ which indicates the influence of observation i on the estimated slope. Please briefly describe your observation.