Please answer the following questions. Do not spend more than 12 hours on this test (it should require no more than 4). Please email back a zipped file of your html, PDF or notebook responses to the theory and data parts. Separately, please return your coding solution.

**Theory** (for this part clearly legible handwritten answers are acceptable)

1. Suppose you run two OLS regression models on disjoint datasets, one of Y1 (Mx1) onto X1 (Mx1) and the other of Y2 (Nx1) onto X2 (Nx1). Both models come without intercept. Let the resulting coefficients be β1 and β2 respectively. Suppose now you co-fit another OLS model on the union of two datasets also without intercept. You therefore have a linear regression of Y ((M+N)x1) onto X ((M+N)x1), where Y is a concatenation of Y1 and Y2 and X is a concatenation of X1 and X2.
   1. Given β1 and β2, what is the possible range of the coefficient β if you run the OLS regression of Y onto X?
   2. What if instead all three regressions of Y1 on X1, of Y2 on X2 and of Y on X are run with intercept? Let β1, β2 and β be the corresponding parameters associated with X1, X2 and X. Given β1 and β2, what is the possible range of β? You may assume the answer is an interval without gaps.
   3. Again, assume no intercepts are fitted in any of the three regressions as in part a. Further assume that all (x, y) pairs in (X, Y) are drawn i.i.d. from a zero-mean 2D multivariate Gaussian distribution. Given β1 and β2, what’s your best guess at β? You may take reasonable approximations in your solution.
2. Suppose that you have two matrices of regressors X1 and X2 where X1 is Nxk and X2 is N\*(f-k). You fit a multiple regression model for a target variable Y onto X1 obtaining residuals e and a QR decomposition of the matrix X1.
   1. You want to find which one of the additional variables contained in X2 will reduce the residual sum of squares the most when included with the ones from X1. Detail an efficient procedure to obtain this.
   2. Derive an efficient procedure to compute the new coefficients of the regression using all variables in X1 plus the variable from X2 you want to add based on (a).

**Data Exercise**

You are given three data sets, X.csv, Y.csv and Z.csv. You can read these data sets in an environment where you would be comfortable analyzing, plotting and describing them (most candidates have chosen R or Python). You can use a Jupyter notebook to share your results or present a file with your main functions and a separate document with the results and comments on your analysis. Your task is to build a model to forecast the vector Y using the variables in X and Z. You can assume that the conditional expectation of Y given X is linear in X.

1. Examine and present the main characteristics of the data.
2. Propose a forecasting model for Y only using the variables in X without Z and explain its properties.
3. Further improve the modeling from (2) with both X and Z.
4. Evaluate the quality of your models and of their parameter estimates. Which one produces the best forecast? Interpret why.