ECE 3502 / CS 3501: Foundations of Data Analysis **Due Th 4/7**

Homework 3: Linear Regression

Instructions: Submit a single Jupyter notebook (.ipynb) of your work to Collab by 11:59pm on the due date. All code should be written in Python. **Be sure to show all the work involved in deriving your answers! If you just give a final answer without explanation, you may not receive credit for that question.**

You may discuss the concepts with your classmates, but write up the answers entirely on your own. Do not look at another student's answers, do not use answers from the internet or other sources, and do not show your answers to anyone. **Cite any sources you used outside of the class material (webpages, etc.), and list any fellow students with whom you discussed the homework concepts.**

1. Implement a function to solve the multilinear regression problem for a given vector y of dependent values and a matrix X of independent values. Your function should return the least-squares solution for the parameter vector, $\hat{\beta}$. **Hint:** Be sure to add a column of all 1's to your X matrix for the intercept term. **Hint 2:** See the SVD example code for matrix operations using `numpy`. In addition to those, you will need to perform a matrix inverse, which you can do with `numpy.linalg.solve`.

2. In this problem we will be analyzing data from the World Health Organization (WHO) on life expectancy in various countries. Download the CSV of the data from the class schedule page. [1]

   (a) Use `Life expectancy` as your y variable, and the following features for your X predictor variables: `Status` (a binary variable for developing or developed country), `Year`, and `BMI` (average body mass index). Run your regression model and report the $\hat{\beta}$ vector of parameters.

   (b) For each of your parameter estimates (each entry of $\hat{\beta}$) write a brief sentence or two about what it tells you about the data.

   (c) Re-run your regression model, this time with only `Status` and `Year` as X features. Plot a scatterplot of life expectancy vs. year. Color the points two different colors for developed vs. developing countries. Now use your $\hat{\beta}$ values to plot two regression lines of life expectancy vs. year, one for developing and one for developed countries.

   (d) Using the formula for the R^2 statistic from class, what is the proportion of variance explained by your two regression models above? Which model has a better (higher) R^2, and can you explain why this is the case?

[1] Data is from here: `https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who`