SML 310 **Mini-Project #1 -- Linear Regression** **Spring 2022**

The following mini-project is focused on sharpening your Data Science skills in the context of Linear Regression! Your writeup should be submitted as a well-documented Python / Jupyter Notebook (as .ipynb and PDF) by e-mail to jonhanke@princeton.edu and sulinl@princeton.edu by 7pm (EST) on Monday February 21th, 2021.

Collaboration: All class projects are required to be solely your work, with no collaboration on modelling or design choices, and you will be required to certify this when you submit your assignment. If you want to ask others about or share general references, please feel free to post them on Slack where all comments are clearly visible.

References: Please feel free to use online resources (with clear inline citations in your code) to better understand the process of using your basic data science tools (Python / Jupyter / Pandas / Scikit-Learn) to perform a specific task (i.e. make a scatter plot, fill in missing values, etc.), but ***you may not consult*** online references specific to your given dataset, or solicit/receive comments outside of course discussions about the choices needed to perform EDA or modelling specific to your dataset. When in doubt about a reference, please feel free to ask on our Slack channel, or during Class / Precept / Office Hours.

Mini-Project Questions:

1. Probability and Linear Regression (20%)

    a. Please explain what it means to perform a linear regression on a collection of pairs $\{(x^{(i)}, y^{(i)})\}$with $1 \leq i \leq n$, drawn from two random variables X and Y? What does this produce? How do we understand it? And what does this tell us about the underlying relationship between X and Y?

    b. Please generate a set of 100 data points $\{(x^{(i)}, y^{(i)})\}$ where X is distributed according to the Gaussian distribution N(0,1), and $Y = X^2$, and another set where X is distributed according to the uniform distribution on [0,1], and$Y = X^2$. Find the linear regression lines for each, and explain why it's reasonable that they are the same/different.

    c. Please explain how it is possible to have a random variable Y that is positively correlated with X_1 but whose coefficient $b_1 < 0$ in a multivariable linear regression model$Y = b_0 + b_1 X_1 + b_2 X_2$. What does this mean about how we interpret a mult-variable model? Can this happen in a single-variable linear

regression model as well? Why/Why not?

d. Please create a dataset that demonstrates your explanation in 1c.

2. Anscombe's Quartet (20%)

   a. Please load each of the four datasets making Anscombe's quartet from the "`Anscombe_Quartet.xlsx`" data file into a Pandas `DataFrame` with columns labelled by "x" and "y".

   b. Verify that each of these datasets has the same mean and variance for each of the variables, and also the same linear regression line. To what extent do the means and variances of the columns (x and y) alone determine the regression line?

   c. Please use your skills as a data scientist to give what you feel is the best model for explaining the variable y in terms of x for each of these four datasets, being sure to carefully explain your reasoning!

3. Diabetes Dataset (60%)

   a. Please load the scikit-learn diabetes dataset (e.g. with the Python `sklearn.datasets.load_diabetes()` command) into a Pandas `DataFrame` that includes all features as labeled columns (e.g. series), as well as the (labeled) disease progression measure in the last column.

   b. Please determine what you feel is the best single variable linear model (i.e. using only one feature) to explain disease progression in terms of the data features, being sure to carefully explain your reasoning!

   c. Please determine what you feel is the best multivariable variable linear model (i.e. using more than one feature) to explain disease progression in terms of the data features, being sure to carefully explain your reasoning!