Instructions:

Be verbose. Clearly explain your reasoning, methods, and results in your written work. Write clear code that is well documented. With 99% certainty, you cannot write too many code comments.

Written answers are worth 8 points. Code is worth 2 points. 10 points total.

Homework is due Saturday, 1/27, at 8am. We will pull answers from your Github at that time.

1. When finished, respond to the questions in Canvas as "done." We will record your grade there.
2. In your code repository, create a folder called "Week02."
3. In that folder, include:
    a. a document (PDF) with your responses.
    b. All code
    c. A README file with instructions for us to run your code.

Problem 0
Create a repository on Github for this class. Email Qiyu (qiyu.wu@duke.edu) and I (dpazzula@dumac.duke.edu) with the link to your repo. Make sure that repo is public so that we may access it. Please do this ASAP.

Problem 1 (3pts)
Given the dataset in problem1.csv:

a. calculate the first four moments values by using normalized formula in the "Week1 - Univariate Stats".

b. calculate the first four moments values again by using your chosen statistical package.

c. Is your statistical package functions biased? Prove or disprove your hypothesis respectively. Explain your conclusion.

Problem 2 (5pts)
Assume the multiple linear regression model $Y = X\beta + \epsilon,\ where\ Y \in R^n, X \in R^{n \times p},\ \beta \in R^p,\ \epsilon \in R^N$:

a. Fit the data in problem2.csv using OLS. Then, fit the data using MLE given the assumption of normality. Compare their beta and standard deviation of the OLS error to the fitted MLE σ. What's your finding? Explain any differences.

b. Fit the data in problem2.csv using MLE given the assumption of a T distribution of errors. Show the fitted parameters. Compare the fitted parameters among MLE under normality assumption and T distribution assumption. Which is the best of fit?

c. Fit the data in problem2_x.csv using MLE given $X = [X_1, X_2]$ follows the multivariate normal distribution. Assume X as a random variable, follows the fitted gaussian distribution, X_1 (problem2_x1.csv) are a part of observed value of X, What's the distribution of X_2 given each observed value? Plot the expected value along with the 95% confidence interval.

d. (Extra Credit: 1 point) Assume $\epsilon \sim N(0, \sigma^2 I_n)$, using Maximum Likelihood Estimation (MLE), derive the estimator for β and σ^2. Show your detailed proof.

Problem 3 (2pts)
Fit the data in problem3.csv using AR (1) through AR (3) and MA (1) through MA (3), respectively. Which is the best of fit?