ASRM 563/ STAT 558 **Risk Modeling and Analysis**

Coding Assignment 1

This assignment is designed to provide hands-on experience of data preprocessing and analysis, and to highlight risk-factor analysis in business/research questions.

Description for the dataset

The dataset containing information of users from a medical insurance company, including the response variable *charges* (the medical fees the user will have to pay) and other explanatory variables such as *age* and *sex*. The goal of the analysis is to predict for the response variable and to reveal significant risk factors for medical expenditure.

Questions

1. Perform data cleaning/preprocessing. Document your findings and how you deal with them. Hint: are there unreasonable values, missing data, outliers, unbalanced groups, etc.?

2. Explore the data. Show distribution and summary statistics for individual variables, correlation between variables, etc. Explain your findings.

3. Perform Ordinary Least Squares (OLS) regression with individual features. Interpret the result. Optional: Are there more questions you have from the data/results/residuals, or improvement for the OLS model? Choose your own model for additional analysis.)

4. What are the risk factors that have significant influence on the medical expenditure? How do they affect the response value?

5. Please report the R^2 and Mean Squared Error (MSE) for your final model on the training dataset. Provide your prediction of the testing set (Prediction performance will be measured by MSE).

Submission

Please submit on Canvas

1. An RMarkdown style file (such as Python notebook or Matlab live script) with the code, the output, and your explanation.

2. A file that reports your prediction of the test data (preferably in csv format).