8STT108 – Big Data Statistics Analytical Tools
Lab #4
Logistic regression

Fall 2024 **Professor: Sara Séguin**

Python is required for the lab assignment. Use any relevant libraries such as: pandas, numpy, statistics, matplotlib.

The goal of the lab is for the students to experiment with the libraries, therefore, you should read any required documentation to understand the tools.

- The spreadsheet **crabs is used for this assignment.**
  https://users.stat.ufl.edu/~aa/cda/data.html
- The following details are given in the file:
  - the color of her shell
  - the condition of her spine
  - the width of her carapace shell (in centimeters)
  - the number of male satellites
  - the weight of the female (in grams)

We want to predict the probability of a female having one or more satellites, based on the width of her shell. A satellite is a male that stays around the female.

Recall the logistic regression equation:

$$log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_i.$$

1. Load the dataset in Python.

Part 1

2. Fit a logistic regression. First, using the "logit" and second the "glm". Report the regression equations and all relevant code. Analyze the results, p-values, and so forth.
3. What is the difference between these 2 models?

4. What are the log odds of a 25 cm femal having stallites? Calculate this value.
5. Transform this log odds into a probability, using calculations.

Part 2

6. Fit a logistic regression. We want to investigate the relationship between the weight and the log odds of having satellites.
7. What is the regression equation?
8. Consider the weight of the female is 2000 grams. What is the probability that she has one or more satellites?