# Machine Learning - Problem Set 3

PPHA 30546 - Professor Clapp
Winter 2024

This assignment must be handed in via Gradescope on Canvas by **11:45pm Central Time on Wednesday, February 14th. Happy Valentines Day!** You are welcome (and encouraged!) to form study groups (of no more than 2 students) to work on the problem sets and mini-projects together. But you must write your own code and your own solutions. Please be sure to include the names of those in your group on your submission.

You should submit your answers in one of two ways:

1. As a single PDF containing BOTH a write-up of your solutions that directly integrates any relevant supporting output from your code (e.g., estimates, tables, figures) AND your code appended to the end of your write up. You may type your answers or write them out by hand and scan them (as long as they are legible). Your original code may be a Python (*.py) or Jupyter Notebook (*.ipynb) file converted to PDF format. OR

2. As a single PDF of a Jupyter Notebook (*.ipynb) file with your your solutions and explanations written in Markdown.[1]

Regardless of how you submit your answers, be sure to make it clear what question you are answering by labeling the sections of your write up well and assigning your answers to the appropriate question in Gradescope. Also, be sure that it is immediately obvious what supporting output from your code (e.g., estimates, tables, figures) you are referring to in your answers. In addition, your answers should be direct and concise. Points will be taken off for including extraneous information, even if part of your answer is correct. You may use bullet points if they are beneficial. Finally, for your code, please also be sure to practice the good coding practices you learned in PPHA 30537/8 and comment your code, cite any sources you consult, etc.

You are allowed to consult the textbook authors' websites, Python documentation, and websites like StackOverflow for general coding questions. You are not allowed to consult material from other classes (e.g., old problem sets, exams, answer keys) or websites that post solutions to the textbook questions.

1. Do the following questions from Chapter 5 of the *Introduction to Statistical Learning* textbook:

   (a) Question 6

   - In parts (a) and (d), the problem references the `sm.glm()` function. You can use the `sm.Logit()` function instead if you'd prefer. The `sm.glm()` function is more general way to estimate different linear models (hence the GLM name), but it will estimate a logistic regression with the appropriate arguments.

[1] Converting a Jupyter Notebook to PDF is not always straightforward (e.g., some methods don't wrap text properly). Please ensure that your PDF is legible! We will deduct points if we cannot read your PDF (even if you have the correct answers in your Notebook).

- Part (a) also references the `summarize()` function. This is a typo. The authors mean the `.summary()` method as a way to view the results of a `sm.glm()` fit.
- In part (c), please draw 1,000 bootstrap samples when bootstrapping your standard errors.

(b) Question 8

- In part (a), note that the code you're given sets a random seed equal to 1.
- In part (c), please keep the same random seed as in part (a).
- In part (d), please set a random seed equal to 2.

2. Do the following questions from Chapter 6 of the *Introduction to Statistical Learning* textbook:

(a) Question 11

- In part (a), use forward stepwise & backwards stepwise selection (FSS & BSS) instead of the methods the book lists. Do so based on using a mathematical adjustment approach (*AIC*) and 5-Fold Cross-Validation (5FCV) to estimate the test error. Use the entire dataset for 5FCV, shuffle the data randomly for splitting, and set `random_state=23`). This means you will select a model four different ways: FSS-*AIC*, FSS-5FCV, BSS-*AIC*, BSS-5FCV.
- As part of your answer for part (b), be sure to explain why the the different methods you use may select different models.