ECE 411 Homework 2 (Fall 2024)
Instructor: Dr. Chau-Wai Wong
Material Covered: Python and R Basics, Linear Regression, Least Squares

Problem 1 (20 points) [ML Environment and First Example] Please complete *ISLP-2.3 Lab: Introduction to Python* or *ISLR-2.3 Lab: Introduction to R*. (← "ISLP" and "ISLR" are the clickable, official links to the textbooks.) You will need to work on the task using one of the two languages: Python or R. Using Python is highly recommended.

Data files such as `Auto.data` and `Auto.csv` can be downloaded from:
https://www.statlearning.com/resources-python or
https://www.statlearning.com/resources-first-edition

Please write a report of (*i*) a few paragraphs providing output of commands and concise explanations for nontrivial commands (segments of important source code need to be included). The report should also include (*ii*) plots and their descriptions.

Coding Environment Setup for Python. Recommended setup: Use a cloud interpreter of Python on `Google Colab`, which allows you to execute Python scripts through a web browser. Other setup options: Use a locally installed Python interpreter via `VS Code`, or those suggested at the beginning of ISLP 2.3.

Coding Environment Setup for R. Install `R`, a statistical programming language, and `RStudio`, an integrated development environment (IDE) for `R`. Using `RStudio` will allow you to complete the tasks more efficiently than using the original `R` software.

For R users: Examples for nontrivial commands include: What does `attach()` do? In what cases do we need to use `as.factor()`? Examples for descriptions of plots include: What do various components of a boxplot mean? Additional hints:

- When data files are loaded, they should be placed in the same folder as displayed in the bottom-right panel of the `RStudio`.
- Try not to reuse a variable name to avoid difficult-to-debug issues. For example, `auto = na.omit(auto)` is bad. Try `auto = na.omit(auto_raw)` instead.
- To finish executing the `identify()` function, you need to click the "Finish" button at the top-right corner of the plot for which the function is called.
- Function `q()` for exiting R may not work in `RStudio`.

Problem 2 (20 points) [College Dataset] Complete *ISLP-2.4.8* and write a report.

Problem 3 (20 points) [Deriving Least-Squares Estimator]

(a) Write model $y_i = \beta_0 + \beta_1 x_i + e_i,\ i = 1, \ldots, n$ into the matrix–vector form. Clearly indicate what are $\underset{\sim}{y}, \mathbf{X}, \underset{\sim}{\beta}$, and $\underset{\sim}{e}$. Set up the cost function $J(\underset{\sim}{\beta})$ to be minimized. Take gradient with respect to $\underset{\sim}{\beta}$ and set it to zero. Express $\underset{\sim}{\hat{\beta}}$ in terms of $\mathbf{X}$ and $\underset{\sim}{y}$. What are the closed-form solution for $\hat{\beta}_0$ and $\hat{\beta}_1$. Is $\hat{\beta}_1$ consistent with that in b)?

(b) We have derived in class the least-squares estimator for p predictors. Now, show through partial differentiation that the least-squares estimator for β_1 of model $y_i = \beta_0 + \beta_1 x_i + e_i$ is

$$\hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}. \tag{1}$$

Prove that the above expression is equivalent to

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}. \tag{2}$$

Problem 4 (20 points) [Data-Driven Formula Search] It is known that the sum of the squares of n from $n = 0$ to $N - 1$ has a closed-form expression as follows:

$$\sum_{n=0}^{N-1} n^2 = a_0 + a_1 N + a_2 N^2 + a_3 N^3. \tag{3}$$

Given that a third-order polynomial is uniquely determined in terms of the values of the polynomial at four distinct points, derive a closed-form expression for this sum by setting up a set of linear equations and solving these equations for a_0, a_1, a_2, and a_3. Verify your solution against the formula $\sum_{k=1}^{n} k^2 = n(n+1)(2n+1)/6$.