HW1: Simple Linear Regression Overview

MATH/STAT 571A

DUE: 09/08/2023 11:59pm

Homework Guidelines

Please submit your answers on Gradescope as a PDF with pages matched to question answers.

One way to prepare your solutions to this homework is with R Markdown, which provides a way to include mathematical notation, text, code, and figures in a single document. A template `.Rmd` file is available through D2L.

Make sure all solutions are clearly labeled, and please utilize the question pairing tool on Gradescope. You are encouraged to work together, but your solutions, code, plots, and wording should always be your own. Come and see me during office hours or schedule an appointment when you get stuck and can't get unstuck.

I. Mathematical Derivations [20 pts]

(1) [4 pts] Let $Y_1, \ldots, Y_n$ be an independent and identically distributed sample with $\mathrm{E}[Y_i] = \mu$ and $\mathrm{Var}(Y_i) = \sigma^2$. Define $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and $e_i = Y_i - \bar{Y}$. Show that $\mathrm{Cov}(e_i, e_j) = -\sigma^2/n$ (hint: recall that for independent RVs Y_i and Y_j, $\mathrm{E}[Y_i Y_j] = \mathrm{E}[Y_i]\mathrm{E}[Y_j]$).

(2) [3 pts] ALRM Exercise 1.5

(3) [3 pts] ALRM Exercise 1.7

(4) [3 pts] ALRM Exercise 1.12 (a)–(c)

(5) [4 pts] ALRM Exercise 1.32

(6) [3 pts] ALRM Exercise 1.36 (hint: you can use the normal equations 1.9a, 1.9b)

II. Tree Cover in Tucson [20 pts]

Download the Tree Equity Scores data set from Module 1 on D2L (`Tree_Equity_Scores_Tucson_noNA.csv`). Use the `read.csv()` function to import the comma spaced values as a dataframe in R. There are several variables recorded in this data set, but we will focus on two (read more here: https://gisdata.tucsonaz.gov/datasets/cotgis::tree-equity-scores-tucson-1/about). The first is `PCTTreeCover`, the percent of each corresponding census tract that is covered by tree canopy. The second is `PCTPoverty`, the percentage of households in each census tract with an income below the poverty line.

(7) [3 pts] Make a scatter plot of percent tree cover on the vertical axis and percent below the poverty line on the horizontal axis. Do you see any statistical relationship in the pattern of the points? If yes, what do you see? If no, why not?

Treat percent tree cover as the response variable and percent below poverty line as a predictor. Assume the simple linear regression model (eq. 1.1) is appropriate.

(8) [3 pts] Obtain the least-squares estimates of the intercept and slope, and state the estimated regression function.

(9) [4 pts] Add a line representing the linear regression function to the scatterplot you made above (you can turn in a single plot for the HW). Does the estimated function appear to fit the data well?

(10) [3 pts] Give a point estimate of the mean percent tree cover for census tracts with 50% of households below the poverty line.

(11) [3 pts] How much is percent tree cover expected to change when the percent of households below the poverty line increased by 20%?

(12) [4 pts] Use R to compute the MLE for σ^2. How does it compare to the MSE? In what units is σ expressed?