ECE 492-45 Homework 4

Material Covered: Linear Regression in Matrix-Vector Form, Confidence Interval, Hypothesis Test

Problem 1 (20 points) [Simulated Data in R] Complete *ISLR-3.6.5, 3.7.13.*

Problem 2 (20 points) [Interpretation of Confidence Interval]

(a) Given a regression function $f(x) = 3x + 1$ and a linear model $Y = f(X) + e$, where $e \sim N(0, 1)$ and $X \sim \text{Uniform}(-1, 1)$, generate 50 pairs of (x_i, y_i).

(b) Use the equations in the lecture slides, calculate the all estimates, namely, $\hat{\beta}_0$ and $\hat{\beta}_1$, and their standard errors. Note that when calculating standard errors, use the estimated value $\text{RSS}/(n-2)$ to replace the theoretical quantity σ^2.

(c) Calculate the confidence interval for β_1. Is 3 included in the interval?

(d) Repeat (a)–(c) 1000 times. What is the chance that 3 is included a calculated confidence interval?

(e) Now, can you explain what is a confidence interval?

Problem 3 (20 points) [Hypothesis Test] Download the `advertising` dataset from ISLR's website. Using formulas from the lecture slides, manually fit a linear model using `sales` against `TV` and with intercept. Re-calculate all entries of tables on slides 11 and 13, except the F-statistics. When calculating the p-values, using a standard normal distribution in lieu of the t distribution. Are the re-calculated values consistent with those in the tables?

Problem 4 (20 points) [Least-Squares Estimator in Matrix-Vector Form] An ECE student John is doing an electronic circuits lab during which he needs to determine the conductance of a resistor using a voltage meter, a current meter, and a DC power source. The voltage meter is connected in parallel with the resistor and the current meter is in series with the resistor. Both meters are analog devices so the readings recorded by John have errors. The power source is tunable and has a range of 1 to 5 V. Each time John will try a uniformly random input voltage level and record the readings of both voltage and current meters. Denote the

voltage reading as x_i and the current reading as y_i for the ith measurement. Assume the true conductance $G = 2$ mΩ^{-1}.

(a) Using a linear model $y_i = Gx_i + e_i$, where e_i is normally distributed with zero-mean and standard deviation $\sigma_e = 0.1$ mA, simulate a dataset of $n = 10$ measurements.

(b) Express the linear model in a matrix-vector form. Clearly indicate what are $\underset{\sim}{y}, \mathbf{X}, \beta$, and $\underset{\sim}{e}$. Directly implement the formula of the least-squares estimator, $\hat{\beta} = (X^T X)^{-1} X^T \underset{\sim}{y}$, into a computer function that takes as two input vectors $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$, and output a number $\hat{G}$. Apply your function to the simulated data. What is the value of $\hat{G}$? (Hint: $\mathbf{X}$ is a N-by-1 "matrix," and β is a 1-by-1 "vector." In R, you may generate a matrix using the `matrix()` function and carry out a matrix multiplication using operator `%*%`.)

(c) John's friend proposed a more intuitive estimator for the conductance: $\tilde{G} = \frac{1}{n} \sum_{i=1}^{N} \frac{y_i}{x_i}$. Write a computer function that takes as two input vectors $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$, and output a number $\tilde{G}$. Apply your function to the simulated data. What is the value of $\tilde{G}$?

(d) Generate 1,000 datasets. Repeatly apply function written in (b) and collect 1,000 estimates and plot the histogram using the `hist()` function with 20 cells/bins.

(e) Use the 1,000 datasets generated in (d). Repeatly apply function written in (c) and collect 1,000 estimates and plot the histogram. Which histogram has a narrower spread? Which estimator is better?

(f) (Bonus, 5 points) Find the analytic expression for $\hat{G}$. Show that both $\hat{G}$ and $\tilde{G}$ are unbiased. (Recall that if $\mathbb{E}[\hat{\theta}] - \theta = 0$ then $\hat{\theta}$ is an unbiased estimator. Consider y_i as a random variable and x_i as a fixed value.)

(g) (Bonus, 5 points) Prove that $\mathrm{Var}(\hat{G}) = \left(\sum_{i=1}^{n} x_i^2 \right)^{-1} \sigma_e^2$ and $\mathrm{Var}(\tilde{G}) = \left(n^{-2} \sum_{i=1}^{n} x_i^{-2} \right) \sigma_e^2$.

Problem 5 (20 points) [Simple Linear Regression without Intercept] Complete *ISLR-3.7.12*.