Statistical Computing / Basic Math for AI: HW 2

Due April 10 11:59 pm, 2024

Write the answers on a paper, scan it, and submit it as a pdf file in Blackboard.

1. Let

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \quad S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2,$$

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}).$$

Also, let $X = [X_1, X_2, \cdots, X_n]^T, Y = [Y_1, Y_2, \cdots, Y_n]^T, \mathbf{1} = [1, 1, \cdots, 1]^T \in \mathbb{R}^n$. Answer the following questions.

(a) Prove that

$$S_X^2 = \frac{1}{n-1} X^T \left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)^T \left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) X.$$

(b) Prove that

$$\left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)^T \left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T.$$

(c) Prove that

$$S_{XY} = \frac{1}{n-1} X^T \left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) Y.$$

(d) Let

$$M = \begin{bmatrix} X_1 & X_2 & \cdots & X_n \\ Y_1 & Y_2 & \cdots & Y_n \end{bmatrix}^T.$$

Express the following matrix using **matrix operation** that involves M.

$$\begin{bmatrix} S_X^2 & S_{XY} \\ S_{XY} & S_Y^2 \end{bmatrix}.$$

2. Suppose we have random samples $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$ sampled from a population with unknown joint distribution $p(x, y)$. That is, the pairs (X_i, Y_i)'s are i.i.d. distributed. (When $i \neq j$, X_i and X_j are independent, Y_i and Y_j are independent, and X_i and Y_j are independent.)

Let σ_{XY} and S_{XY} be the population covariance and sample covariance respectively,

$$\sigma_{XY} = E[(X - E(X))(Y - E(Y))], \quad S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}).$$

Also, let

$$\sigma_X^2 = E[(X - E(X))^2], \quad \sigma_Y^2 = E[(Y - E(Y))^2], \quad \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \quad S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2, \quad r_{XY} = \frac{S_{XY}}{S_X S_Y}.$$

(ρ_{XY} and r_{XY} are population correlation and sample correlation respectively.)

(a) Prove that

$$E[S_{XY}] = \sigma_{XY}.$$

You can use any style of proof you want, either using matrix operation or not.

(b) Suppose $Y = aX + b$ $(a \neq 0)$. Prove that $\rho_{XY} = sign(a)$ and $r_{XY} = sign(a)$, where

$$sign(a) = \begin{cases} 1 & \text{if } a > 0 \\ -1 & \text{if } a < 0 \end{cases}$$

3. Suppose that we have two datasets:

$$D_1 = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{15} \\ y_{11} & y_{12} & \cdots & y_{15} \end{bmatrix}, \quad D_2 = \begin{bmatrix} x_{21} & x_{22} & \cdots & x_{25} \\ y_{21} & y_{22} & \cdots & y_{25} \end{bmatrix}.$$

Suppose that in both D_1 and D_2,

$$y_{ji} = \beta_0 + \beta_1 x_{ji} + \varepsilon_{ji}, \quad j = 1, 2 \ \& \ i = 1, 2, \cdots, 5$$

where $\varepsilon_{11}, \cdots, \varepsilon_{25}$ are iid with mean 0 and variance 1. (Hence, the values of β_0, β_1 are the same in D_1 and D_2.)

Suppose that

$$x_{11} = 1, \ x_{12} = 2, \ x_{13} = 3, \ x_{14} = 4, \ x_{15} = 5,$$

$$x_{21} = 2, \ x_{22} = 2, \ x_{23} = 3, \ x_{24} = 4, \ x_{25} = 4.$$

Suppose that you can look only one dataset. As a statistician, which one would you choose to make a *better* inference of β_0 and β_1? Explain your response.

4. Prove that in simple linear regression with least-squares estimation,

$$R^2 = r^2_{Y\hat{Y}},$$

where $r_{Y\hat{Y}}$ is the sample correlation of $Y = [y_1, y_2, \cdots, y_n]^T$ and $\hat{Y} = [\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n]^T$.

5. Suppose we conduct linear regression on outcome variable (y) and explanatory variable (x). We posit the following model,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \cdots, 10,$$

where ϵ_i's are iid with distribution $\mathcal{N}(0, 1)$. Suppose

$$\sum_{i=1}^{10} x_i = 10, \quad \sum_{i=1}^{10} x_i^2 = 100, \quad \sum_{i=1}^{10} y_i = 20, \quad \sum_{i=1}^{10} x_i y_i = 30.$$

(a) What are the least-squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$?

(b) Construct a 95% confidence interval of β_0. (Set $z_{0.025} = 2$ and assume that we know that $\sigma^2 = 1$.)

Homework Collaboration Policy Collaboration on homework is allowed, however, you should follow the following rules.

- After discussion with collaborators, write the answers and codes in *your own words.*
- Make sure you acknowledge the person you got help from, for each exercise.