IE307: Statistical Computing
Assignment 2

Student Name: Nguyen Minh Duc
Student ID: 20202026

Problem 1. Let

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

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

$$S_{XY}^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right).$$

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

(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

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

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

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

Solution.

(a)

$$\begin{aligned} S_X^2 &= \frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(X_i - \bar{X}\right) \\ &= \frac{1}{n-1}\begin{bmatrix} X_1 - \bar{X} & X_2 - \bar{X} & \cdots & X_n - \bar{X} \end{bmatrix} \begin{bmatrix} X_1 - \bar{X} \\ X_2 - \bar{X} \\ \vdots \\ X_n - \bar{X} \end{bmatrix} \end{aligned}$$

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

Note that

$$\begin{aligned}
X - \mathbf{1}\bar{X} &= X - \mathbf{1}\frac{1}{n}\mathbf{1}^T X \\
&= I_n X - \frac{1}{n}\mathbf{1}\mathbf{1}^T X \\
&= \left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) X.
\end{aligned}$$

Thus,

$$\begin{aligned}
S_X^2 &= \frac{1}{n-1}\left[\left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) X\right]^T \left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) X \\
&= \frac{1}{n-1} X^T \left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)^T \left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) X
\end{aligned}$$

(b) First, let's simplify

$$\begin{aligned}
\left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)^T &= I_n^T - \left(\frac{1}{n}\mathbf{1}\mathbf{1}^T\right)^T \\
&= I_n - \frac{1}{n}\left(\mathbf{1}\mathbf{1}^T\right)^T \\
&= I_n - \frac{1}{n}\left(\mathbf{1}^T\right)^T \mathbf{1}^T \\
&= I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T
\end{aligned}$$

Now, we have

$$\begin{aligned}
\left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)^T \left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) &= \left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) \\
&= I_n I_n - \frac{1}{n} I_n \mathbf{1}\mathbf{1}^T - \frac{1}{n}\mathbf{1}\mathbf{1}^T I_n + \frac{1}{n^2}\mathbf{1}\mathbf{1}^T\mathbf{1}\mathbf{1}^T \\
&= I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T - \frac{1}{n}\mathbf{1}\mathbf{1}^T + \frac{1}{n^2}\mathbf{1}\left(\mathbf{1}^T\mathbf{1}\right)\mathbf{1}^T \\
&= I_n - \frac{2}{n}\mathbf{1}\mathbf{1}^T + \frac{1}{n^2}\mathbf{1}n\mathbf{1}^T \\
&= I_n - \frac{2}{n}\mathbf{1}\mathbf{1}^T + \frac{1}{n}\mathbf{1}\mathbf{1}^T \\
&= I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T.
\end{aligned}$$

(c)

$$\begin{aligned}
S_{XY}^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) \\
&= \frac{1}{n-1}\begin{bmatrix} X_1 - \bar{X} & X_2 - \bar{X} & \cdots & X_n - \bar{X} \end{bmatrix} \begin{bmatrix} Y_1 - \bar{Y} \\ Y_2 - \bar{Y} \\ \vdots \\ Y_n - \bar{Y} \end{bmatrix} \\
&= \frac{1}{n-1}\left[\left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) X\right]^T \left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) Y
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n-1} X^T \left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)^T \left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) Y \\
&= \frac{1}{n-1} X^T \left[\left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)^T \left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\right] Y \\
&= \frac{1}{n-1} X^T \left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) Y.
\end{aligned}
$$

(d)

$$
\begin{aligned}
\begin{bmatrix} S_X^2 & S_{XY} \\ S_{XY} & S_Y^2 \end{bmatrix} &= \begin{bmatrix} S_X^2 & S_{YX} \\ S_{XY} & S_Y^2 \end{bmatrix} \\
&= \begin{bmatrix} \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 & \frac{1}{n-1} X^T \left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) Y \\ \frac{1}{n-1} Y^T \left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) X & \frac{1}{n-1} Y^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) Y \end{bmatrix} \\
&= \frac{1}{n-1} \begin{bmatrix} X^T \left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) X & X^T \left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) Y \\ Y^T \left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) X & Y^T \left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) Y \end{bmatrix} \\
&= \frac{1}{n-1} \begin{bmatrix} \begin{bmatrix} X_1 & X_2 & \dots & X_n \end{bmatrix} \left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) X & \begin{bmatrix} X_1 & X_2 & \dots & X_n \end{bmatrix} \left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) Y \\ \begin{bmatrix} Y_1 & Y_2 & \dots & Y_n \end{bmatrix} \left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) X & \begin{bmatrix} Y_1 & Y_2 & \dots & Y_n \end{bmatrix} \left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) Y \end{bmatrix} \\
&= \frac{1}{n-1} \begin{bmatrix} X_1 & X_2 & \dots & X_n \\ Y_1 & Y_2 & \dots & Y_n \end{bmatrix} \left[\left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) X \;\middle|\; \left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) Y\right] \\
&= \frac{1}{n-1} M^T \left[\left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \;\middle|\; \left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}\right] \\
&= \frac{1}{n-1} M^T \left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) \begin{bmatrix} X_1 & Y_1 \\ X_2 & Y_2 \\ \vdots & \vdots \\ X_n & Y_n \end{bmatrix} \\
&= \frac{1}{n-1} M^T \left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) M
\end{aligned}
$$

Problem 2. Suppose we have random samples $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ sampled from a population with an unknown joint distribution $p(x, y)$. That is, the pairs (X_1, Y_1)'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,

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

Also, let

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

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

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

(a) Prove that

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

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

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

Solution.

(a) Let $\mathbf{X} = [X_1, X_2, \ldots, X_n]^T$, $\mathbf{Y} = [Y_1, Y_2, \ldots, Y_n]^T$. The expectation can be expanded as follows

$$\begin{aligned}
E\left[S_{XY}\right] &= E\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)\right] \\
&= E\left[\frac{1}{n-1}\mathbf{X}^T\left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\mathbf{Y}\right] \\
&= \frac{1}{n-1}E\left[\mathbf{X}^T\left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\mathbf{Y}\right] \\
&= \frac{1}{n-1}E\left[Tr\left(\mathbf{X}^T\left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\mathbf{Y}\right)\right] \\
&= \frac{1}{n-1}E\left[Tr\left(\left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\mathbf{Y}\mathbf{X}^T\right)\right] \\
&= \frac{1}{n-1}Tr\left(E\left[\left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\mathbf{Y}\mathbf{X}^T\right]\right) \\
&= \frac{1}{n-1}Tr\left(\left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)E\left[\mathbf{Y}\mathbf{X}^T\right]\right)
\end{aligned}$$

Note that the covariance matrix is

$$\begin{aligned}
Cov(\mathbf{Y}, \mathbf{X}) &= E\left[\left(\mathbf{Y} - E(\mathbf{Y})\right)\left(\mathbf{X} - E(\mathbf{X})\right)^T\right] \\
&= E\left[\mathbf{Y}\mathbf{X}^T - \mathbf{Y}E(\mathbf{X})^T - E(\mathbf{Y})\mathbf{X}^T + E(\mathbf{Y})E(\mathbf{X})^T\right] \\
&= E(\mathbf{Y}\mathbf{X}^T) - E(\mathbf{Y})E(\mathbf{X})^T - E(\mathbf{Y})E(\mathbf{X})^T + E(\mathbf{Y})E(\mathbf{X})^T \\
&= E(\mathbf{Y}\mathbf{X}^T) - E(\mathbf{Y})E(\mathbf{X})^T
\end{aligned}$$

Thus, the expectation is the same as

$$E\left[S_{XY}\right] = \frac{1}{n-1}Tr\left(\left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\left(Cov(\mathbf{Y}, \mathbf{X}) + E(\mathbf{Y})E(\mathbf{X})^T\right)\right)$$

$$
\begin{aligned}
&= \frac{1}{n-1} Tr\left(Cov(\mathbf{Y},\mathbf{X}) + E(\mathbf{Y})E(\mathbf{X})^T - \frac{1}{n}\mathbf{1}\mathbf{1}^T Cov(\mathbf{Y},\mathbf{X}) - \frac{1}{n}\mathbf{1}\mathbf{1}^T E(\mathbf{Y})E(\mathbf{X})^T \right) \\
&= \frac{1}{n-1} Tr\left(Cov(\mathbf{Y},\mathbf{X}) + \mathbf{1}\bar{Y}\mathbf{1}^T\bar{X} - \frac{1}{n}\mathbf{1}\mathbf{1}^T Cov(\mathbf{Y},\mathbf{X}) - \frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{1}\bar{Y}\mathbf{1}^T\bar{X} \right) \\
&= \frac{1}{n-1} Tr\left(Cov(\mathbf{Y},\mathbf{X}) + \bar{Y}\bar{X}\mathbf{1}\mathbf{1}^T - \frac{1}{n}\mathbf{1}\mathbf{1}^T Cov(\mathbf{Y},\mathbf{X}) - \bar{Y}\bar{X}\frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{1}\mathbf{1}^T \right) \\
&= \frac{1}{n-1} Tr\left(Cov(\mathbf{Y},\mathbf{X}) + \bar{Y}\bar{X}\mathbf{1}\mathbf{1}^T - \frac{1}{n}\mathbf{1}\mathbf{1}^T Cov(\mathbf{Y},\mathbf{X}) - \bar{Y}\bar{X}\frac{1}{n}\mathbf{1}n\mathbf{1}^T \right) \\
&= \frac{1}{n-1} Tr\left(Cov(\mathbf{Y},\mathbf{X}) + \bar{Y}\bar{X}\mathbf{1}\mathbf{1}^T - \frac{1}{n}\mathbf{1}\mathbf{1}^T Cov(\mathbf{Y},\mathbf{X}) - \bar{Y}\bar{X}\mathbf{1}\mathbf{1}^T \right) \\
&= \frac{1}{n-1} Tr\left(Cov(\mathbf{Y},\mathbf{X}) - \frac{1}{n}\mathbf{1}\mathbf{1}^T Cov(\mathbf{Y},\mathbf{X}) \right) \\
&= \frac{1}{n-1} \left(Tr\left(Cov(\mathbf{Y},\mathbf{X})\right) - Tr\left(\frac{1}{n}\mathbf{1}\mathbf{1}^T Cov(\mathbf{Y},\mathbf{X}) \right) \right) \\
&= \frac{1}{n-1} \left(Tr\left(Cov(\mathbf{X},\mathbf{Y})^T\right) - Tr\left(\frac{1}{n}\mathbf{1}^T Cov(\mathbf{X},\mathbf{Y})^T\mathbf{1} \right) \right) \\
&= \frac{1}{n-1} \left(Tr\left(Cov(\mathbf{X},\mathbf{Y})\right) - \frac{1}{n}\mathbf{1}^T Cov(\mathbf{X},\mathbf{Y})^T\mathbf{1} \right) \\
&= \frac{1}{n-1} \left(n\sigma_{XY} - \frac{1}{n}\mathbf{1}^T Cov(\mathbf{X},\mathbf{Y})^T\mathbf{1} \right)
\end{aligned}
$$

Note that X_i and Y_j are independent for every $i \neq j$, hence, the covariance matrix:

$$
Cov(\mathbf{X},\mathbf{Y}) = \begin{bmatrix} \sigma_{XY} & 0 & \cdots & 0 \\ 0 & \sigma_{XY} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{XY} \end{bmatrix} = Cov(\mathbf{X},\mathbf{Y})^T.
$$

This results in

$$
\begin{aligned}
E\left[S_{XY}\right] &= \frac{1}{n-1}\left(n\sigma_{XY} - \frac{1}{n}n\sigma_{XY} \right) \\
&= \frac{1}{n-1}\left(n\sigma_{XY} - \sigma_{XY} \right) \\
&= \frac{1}{n-1}\left(n-1 \right)\sigma_{XY} \\
&= \sigma_{XY}.
\end{aligned}
$$

(b) Since $Y = aX + b$, $\sigma_Y^2 = a^2\sigma_X^2 \implies \sigma_Y = |a|\sigma_X$.
The population covariance is:

$$
\begin{aligned}
\sigma_{XY} &= Cov(X,Y) \\
&= Cov(X, aX+b) \\
&= Cov(X, aX) + Cov(X, b) \\
&= aCov(X,X) + 0 \\
&= aVar(X) \\
&= a\sigma_X^2.
\end{aligned}
$$

Thus, the population correlation is

$$
\begin{aligned}
\rho_{XY} &= \frac{\sigma_{XY}}{\sigma_X\sigma_Y} \\
&= \frac{a\sigma_X^2}{\sigma_X|a|\sigma_X} \\
&= \frac{a}{|a|} = sign(a).
\end{aligned}
$$

This is true because $\dfrac{a}{|a|} = \begin{cases} \frac{a}{a} = 1 & \text{if } a > 0 \\ \frac{a}{-a} = -1 & \text{if } a < 0 \end{cases} = sign(a).$

Similarly, $S_Y^2 = a^2 S_X^2 \implies S_Y = |a| S_X$, and $S_{XY} = a S_X^2$. Therefore, the sample correlation is

$$\begin{aligned} r_{XY} &= \frac{S_{XY}}{S_X S_Y} \\ &= \frac{a S_X^2}{S_X |a| S_X} \\ &= \frac{a}{|a|} = sign(a). \end{aligned}$$

Problem 3. Suppose that we have two datasets

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

Suppose that in both D_1 and D_2,

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

where $\epsilon_{11}, \ldots, \epsilon_{25}$ 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

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

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

Suppose that you can look at 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.

Solution.

In D_1, the x_{ji} values are distinct and evenly spaced from 1 to 5. This indicates that there is a good degree of variability in the x_{ji} values. While in D_2, the x_{ji} values are not distinct, with repeating values of 2 and 4.

Intuitively, the dataset D_1 seems to be a better choice to make an inference on β_0 and β_1 as each data point would have equal weight and the model wouldn't bias towards any groups of clusters, not like the dataset D_2.

More formally, let's look at the variance of β_0 and β_1 in each dataset. Let $n = 5$, then the variance formula of the estimates are

$$\begin{aligned} \text{Var}(\beta_1) &= \frac{\sigma^2}{S_{XX}}, \\ \text{Var}(\beta_0) &= \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{S_{XX}}. \end{aligned}$$

The mean of each x_i's in each dataset is

$$\begin{aligned} \overline{x_1} &= \frac{1+2+3+4+5}{5} = 3, \\ \overline{x_2} &= \frac{2+2+3+4+4}{5} = 3. \end{aligned}$$

In D_1, the estimates' variances are

$$\begin{aligned} \text{Var}\left(\beta_{11}\right) &= \frac{1}{\sum_{i=1}^{n} (x_{1i} - \overline{x_1})^2} = \frac{1}{10}, \\ \text{Var}\left(\beta_{10}\right) &= \frac{\sigma^2}{n} + \bar{x_1}^2 \frac{\sigma^2}{S_{X_1 X_1}} = \frac{1}{5} + 3^2 \times \frac{1}{10} = \frac{11}{10}. \end{aligned}$$

In D_2, the estimates' variances are

$$\begin{aligned}
\operatorname{Var}(\beta_{21}) &= \frac{1}{\sum_{i=1}^n (x_{2i} - \overline{x_2})^2} = \frac{1}{4}, \\
\operatorname{Var}(\beta_{20}) &= \frac{\sigma^2}{n} + \bar{x_2}^2 \frac{\sigma^2}{S_{X_2 X_2}} = \frac{1}{5} + 3^2 \times \frac{1}{4} = \frac{49}{20}.
\end{aligned}$$

Note that $\operatorname{Var}(\beta_{11}) < \operatorname{Var}(\beta_{21})$, and $\operatorname{Var}(\beta_{10}) < \operatorname{Var}(\beta_{20})$. This means the estimates' variance in D_1 is smaller than in D_2, making D_1 a better choice for inferencing β_0 and β_1.

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

$$R^2 = r_{Y\hat{Y}}^2,$$

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

Solution.

The sample correlation between Y and $\hat{Y}$ is

$$\begin{aligned}
r_{Y\hat{Y}} &= \frac{\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}} \\
&= \frac{\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{S_{YY} S_{\hat{Y}\hat{Y}}}}.
\end{aligned}$$

Recall that $\beta_0 = \bar{y} - \beta_1 \bar{x} \implies \bar{y} = \beta_0 + \beta_1 \bar{x}$. Also $\bar{\hat{y}} = \frac{1}{n} \sum_{i=1}^n \beta_0 + \beta_1 x_i \implies \bar{\hat{y}} = \beta_0 + \beta_1 \bar{x} = \bar{y}$. Thus,

$$r_{Y\hat{Y}} = \frac{\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{S_{YY} S_{\hat{Y}\hat{Y}}}}.$$

The numerator is equivalent to

$$\begin{aligned}
\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{y}) &= \sum_{i=1}^n (y_i - \bar{y}) \left(\bar{y} + \beta_1 (x_i - \bar{x}) - \bar{y}\right) \\
&= \sum_{i=1}^n (y_i - \bar{y}) \left(\beta_1 (x_i - \bar{x})\right) \\
&= \beta_1 \sum_{i=1}^n (y_i - \bar{y}) (x_i - \bar{x}) \\
&= \frac{S_{XY}}{S_{XX}} S_{XY} \\
&= \frac{S_{XY}^2}{S_{XX}}
\end{aligned}$$

The sample correlation now becomes

$$\begin{aligned}
r_{Y\hat{Y}} &= \frac{\frac{S_{XY}^2}{S_{XX}}}{\sqrt{S_{YY} S_{\hat{Y}\hat{Y}}}} \\
&= \frac{S_{XY}^2}{S_{XX} \sqrt{S_{YY} S_{\hat{Y}\hat{Y}}}}
\end{aligned}$$

The regression sum of squares, i.e., $S_{\hat{Y}\hat{Y}}$ can be rewritten as

$$S_{\hat{Y}\hat{Y}} = \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^{n} \left(\bar{y} + \beta_1 (x_i - \bar{x}) - \bar{y}\right)^2 \\
&= \beta_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \\
&= \frac{S_{XY}^2}{S_{XX}^2} S_{XX} \\
&= \frac{S_{XY}^2}{S_{XX}}.
\end{aligned}
$$

The sample correlation can be further simplified as follows

$$
\begin{aligned}
r_{Y\hat{Y}} &= \frac{S_{XY}^2}{S_{XX}\sqrt{S_{YY}S_{\hat{Y}\hat{Y}}}} \\
&= \frac{S_{XY}^2}{S_{XX}\sqrt{S_{YY}\frac{S_{XY}^2}{S_{XX}}}} \\
&= \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}.
\end{aligned}
$$

Squaring the sample correlation now gives the R^2 coefficient,

$$
r_{Y\hat{Y}}^2 = \frac{S_{XY}^2}{S_{XX}S_{YY}} = R^2.
$$

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

$$
y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, 3, \ldots, 10,
$$

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 $\hat{\beta}_0$. (Set $z_{0.025} = 2$ and assume that we know $\sigma^2 = 1$.)

Solution.

(a) Let $n = 10$. First, let's calculate the sample mean of x and y:

$$
\begin{aligned}
\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10} \times 10 = 1, \\
\bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{10} \times 20 = 2.
\end{aligned}
$$

Recall that

$$
\beta_1 = \frac{S_{XY}}{S_{XX}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.
$$

Expanding the numerator, we obtain

$$\begin{aligned}
\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) &= \sum_{i=1}^{n}\left(x_iy_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y}\right) \\
&= \sum_{i=1}^{n} x_iy_i - \sum_{i=1}^{n} x_i\bar{y} - \sum_{i=1}^{n} \bar{x}y_i + \sum_{i=1}^{n} \bar{x}\bar{y} \\
&= 30 - \bar{y}\sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} y_i + n\bar{x}\bar{y} \\
&= 30 - 2\times 10 - 1\times 20 + 10\times 1\times 2 \\
&= 10.
\end{aligned}$$

Expanding the denominator, we obtain

$$\begin{aligned}
\sum_{i=1}^{n}(x_i-\bar{x})^2 &= \sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right) \\
&= \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2x_i\bar{x} + \sum_{i=1}^{n} \bar{x}^2 \\
&= 100 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2 \\
&= 100 - 2\times 1\times 10 + 10\times 1^2 \\
&= 90.
\end{aligned}$$

The least-squares estimate $\hat{\beta}_1$ is $\dfrac{10}{90} = \dfrac{1}{9}$.

The least-squares estimate $\hat{\beta}_0$ is $\bar{y} - \beta_1\bar{x} = 2 - \dfrac{1}{9}\times 1 = \dfrac{17}{9}$.

(b) Since we know that $\sigma = 1$, the variance of $\hat{\beta}_0$ is

$$\operatorname{Var}(\hat{\beta}_0) = \frac{\sigma^2}{n} + \bar{x}^2\frac{\sigma^2}{S_{XX}} = \frac{1^2}{10} + 1^2\times\frac{1^2}{90} = \frac{1}{9}.$$

The standard deviation of $\hat{\beta}_0$ is

$$\sigma_{\beta_0} = \sqrt{\operatorname{Var}(\hat{\beta}_0)} = \sqrt{\frac{1}{9}} = \frac{1}{3}.$$

The formula for the 95% confidence interval of $\hat{\beta}_0$ is

$$\frac{17}{9} \pm z_{0.025}\sigma_{\beta_0}.$$

Substitute the values gives us the interval

$$\frac{17}{9} \pm \frac{2}{3}.$$

Simplifying, we have

$$\frac{17\pm 6}{9}.$$

Therefore, the 95% confidence interval of $\hat{\beta}_0$ is $\left[\dfrac{11}{9}, \dfrac{23}{9}\right]$.