Quiz 1

Date: 28-Aug-2025

Full Marks: 15 Time: 45 minutes

1. (5 points) Suppose we have k independent random samples from populations with a common variance σ^2:

$$X_{i1}, X_{i2}, \ldots, X_{in_i}, \quad i = 1, 2, \ldots, k.$$

For each sample, let

$$S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2.$$

Suggest an appropriate formula for the *pooled sample variance* using the S_i's and show that your suggestion is an unbiased estimator of the common variance σ^2.

2. (5 points) Consider the multiple linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n),$$

where $\mathbf{X}$ is appropriately defined. Let $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ be the least-square fitted values. Define

$$\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2, \quad \text{MSS} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2, \quad \text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Derive the decomposition TSS = MSS + RSS.

3. (5 points) Recall the following programming problem from Assignment-1:

==========

Consider the simple linear regression model $y = 50 + 10x + \varepsilon$ where ε is NID$(0, 16)$. Suppose that $n = 20$ pairs of observations are used to fit this model. Generate 500 samples of 20 observations, drawing one observation for each level of $x = 1, 1.5, 2, \ldots, 10$ for each sample.

1. For each sample, compute the least-squares estimates of the slope and intercept. Construct histograms of the sample values of $\hat{\beta}_0$ and $\hat{\beta}_1$. Discuss the shape of these histograms.

==========

The correct solution is shown in Figure 1. However you end up writing the following code for solving the Assignment Problem above. Assuming the same sequence of random numbers as in the correct code, plot the output histograms for your code and explain your results. Give an approximate (visual) estimate of the variances of $\hat{\beta}_0$ and $\hat{\beta}_1$ from your histograms and comment about their correctness with respect to the original problem.

500 samples.

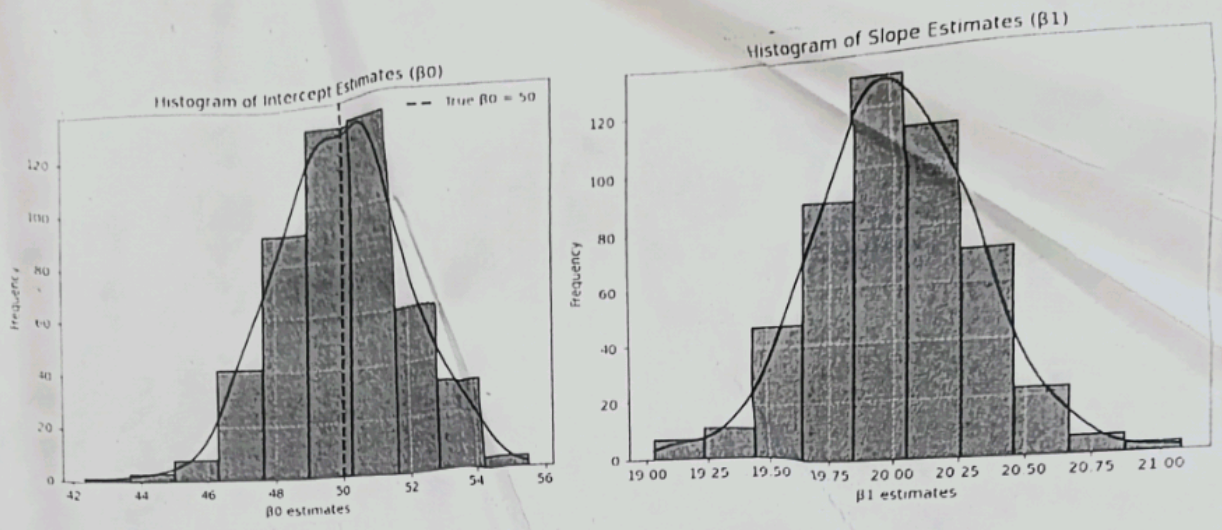

Figure 1: Correct Solution to Assignment Question

```
import numpy as np; import matplotlib.pyplot as plt; import seaborn as sns
np.random.seed(42)
beta0, beta1, sigma, n_samples, n_obs, temp = 50 , 10, 4, 500, 20, 1
x_values = []
for i in range( n_obs ):
    x_values.append( temp )
    temp += 0.5
x_values = np.array( x_values )
all_samples_X = []
all_samples_Y = []
for _ in range(n_samples):
    eps = np.random.normal(loc=0, scale=sigma , size=n_obs)
    y_values = beta0 + 2*beta1 * x_values + eps
    all_samples_X.append(x_values)
    all_samples_Y.append(y_values)
all_samples_X = np.array(all_samples_X)  # shape (500, 20)
all_samples_Y = np.array(all_samples_Y)  # shape (500, 20)
beta0_estimates = []
beta1_estimates = []
for i in range(n_samples) :
    x_mean = np.mean(all_samples_X[i])
    y_mean = np.mean(all_samples_Y[i])
    beta1_hat = np.sum((all_samples_X[i] - x_mean) * (all_samples_Y[i] - y_mean))
    / np.sum((all_samples_X[i] - x_mean)**2)
    beta0_hat = y_mean - beta1_hat * x_mean
    beta0_estimates.append(beta0_hat)
    beta1_estimates.append(beta1_hat)
beta0_estimates = np.array(beta0_estimates)
beta1_estimates = np.array(beta1_estimates)

# Assume correct Code for plotting histograms for \beta_0 and \beta_1
# using the variables beta0_estimates and beta1_estimates
```