Problem 1  (25 points)
Consider the following data:

<table>
<thead>
<tr>
<th>y</th>
<th>x</th>
</tr>
</thead>
<tbody>
<tr>
<td>y_{11}</td>
<td>1</td>
</tr>
<tr>
<td>y_{21}</td>
<td>1</td>
</tr>
<tr>
<td>y_{31}</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>y_{n1}</td>
<td>1</td>
</tr>
<tr>
<td>y_{12}</td>
<td>0</td>
</tr>
<tr>
<td>y_{22}</td>
<td>0</td>
</tr>
<tr>
<td>y_{32}</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>y_{n2}</td>
<td>0</td>
</tr>
</tbody>
</table>

These data concern the regression of y on x, but x here indicates group membership. If x = 1 then the corresponding y value belongs to group 1, and if x = 0 the corresponding y value belongs to group 2. There are n_1 observations in group 1 and n_2 observations in group 2, a total of n = n_1 + n_2 observations. The formulas that we discussed in class on simple regression apply here as well. But there is an interesting result about \( \hat{\beta}_1 \) and \( \hat{\beta}_0 \). It is easy to see that \( \bar{x} = \frac{n_1}{n_1 + n_2} \), \( \sum_{i=1}^{n} x_i^2 = n_1 \), and \( (\sum_{i=1}^{n} x_i)^2 = n_1^2 \). Answer the following questions:

a. Show that \( \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{n_1 n_2}{n_1 + n_2} \).

\[
\sum (x_i - \bar{x})^2 = \sum x_i^2 - \bar{x}^2 = n_1 \left( \frac{n_1 + n_2}{n_1} \right) \frac{n_1^2}{n_1 + n_2} = \frac{n_1 n_2}{n_1 + n_2}.
\]

b. Show that \( \hat{\beta}_1 = \bar{y}_1 - \bar{y}_2 \), and \( \hat{\beta}_0 = \bar{y}_2 \), where \( \bar{y}_1 \) is the sample mean of the y values in group 1 and \( \bar{y}_2 \) is the sample mean of the y values in group 2.

\[
\hat{\beta}_1 = \frac{\sum x_i y_i - \frac{1}{n} (\sum x_i) (\sum y_i)}{\sum (x_i - \bar{x})^2} = \frac{n_1 \bar{y}_1 - \frac{1}{n_1 + n_2} n_1 (n_1 \bar{y}_1 + n_2 \bar{y}_2)}{n_1} = \frac{n_1 \bar{y}_1 + n_2 \bar{y}_2}{n_1 + n_2}.
\]

\[
\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x} = \bar{y} + \frac{n_1 \bar{y}_1 + n_2 \bar{y}_2}{n_1 + n_2} = \bar{y}_1.
\]

c. For a particular data set it is given that \( n_1 = 15, n_2 = 10, s^2 = 0.8, \bar{y}_1 = -0.46, \) and \( \bar{y}_2 = 0.28 \). Compute \( R^2 \).

\[
R^2 = \frac{(\bar{y} - \bar{y}_1)^2}{\sum (y_i - \bar{y})^2} = \frac{(\bar{y} - \bar{y}_1)^2}{(n-1) s^2} = \left( \frac{0.46 - 0.28}{2.8} \right)^2 = \frac{150}{2.8^2} = 17.1 \%
\]
Problem 2 (225 points)
Answer the following questions:

a. Consider the simple regression of $y$ on $x$. Suppose we transform the $x$ values using $x' = \frac{x - \bar{x}}{s_x}$ and the $y$ values using $y' = \frac{y - \bar{y}}{s_y}$. Is it true that the estimated slope is $\hat{\beta}_1 = r$ (correlation coefficient)? Please explain your answer mathematically (use the formulas to show if this is true).

\[ \hat{\beta}_1 = \frac{\sum (y_i' - \bar{y}')(x_i' - \bar{x}')} {\sum (x_i' - \bar{x})^2} = \frac{\frac{1}{n-1} \sum (y_i' - \bar{y}')(x_i' - \bar{x}')} {\frac{1}{n-1} \sum (x_i' - \bar{x})^2} = \frac{\frac{1}{n-1} \sum (y_i' - \bar{y})(x_i' - \bar{x})} {\frac{1}{n-1} \sum (x_i - \bar{x})^2} = \frac{r \sigma_{y'} \sigma_{x'}} {\sigma_x \sigma_y} = r. \]

YES.

b. Refer to question (a). Find $\hat{\beta}_0$.

\[ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \]

\[ \bar{y} \text{ of the transformed } y \text{ is zero} \]

\[ \bar{x} \text{ of the transformed } x \text{ is zero} \]

c. Let $\hat{\beta}_1$ and $\hat{\beta}_0$ be the estimated slope and intercept of the simple regression of $y$ on $x$. If we multiply the $y$ variable by 5 and the $x$ variable by 4 and we regress $5y$ on $4x$ give the new estimates of the slope and intercept in terms of $\hat{\beta}_1$ and $\hat{\beta}_0$. Show all your work.

\[ \hat{\beta}_1 = \frac{\sum (5y_i - 5\bar{y})(4x_i - 4\bar{x})} {\sum (4x_i - 4\bar{x})^2} = \frac{4 \times 5 \sum (x_i - \bar{x})(y_i - \bar{y})} {4^2 \sum (x_i - \bar{x})^2} = \frac{5} {4} \hat{\beta}_1. \]

\[ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 5\bar{y} - \frac{5} {4} \hat{\beta}_1 4\bar{x} = 5 \hat{\beta}_0. \]

d. Refer to question (c). Will $R^2$ of the regression of $5y$ on $4x$ be the same with the $R^2$ of the regression of $y$ on $x$? Please explain your answer mathematically.

\[ R^2 = \frac{\hat{\beta}^2 \sum (x_i - \bar{x})^2} {\sum (y_i - \bar{y})^2} = \frac{\frac{5} {4} \hat{\beta}_1^2 \sum (x_i - \bar{x})^2} {\sum (y_i - \bar{y})^2} = \frac{5} {4} R^2. \]

SAME.

e. Suppose in a simple regression of $y$ on $x$ it happens that $\bar{x} = 0$. Find $\hat{\beta}_1$ and $\hat{\beta}_0$. For the same data set, find the new estimates of $\hat{\beta}_1$ and $\hat{\beta}_0$ in terms of the old ones if we multiply the $x$ variable by 4.

IF we multiply $x$ by 4:

\[ \hat{\beta}_1 = \frac{\sum x_i y_i} {\sum x_i^2} = \frac{4 \sum x_i y_i} {4^2 \sum x_i^2} = \frac{1} {4} \hat{\beta}_1. \]

\[ \hat{\beta}_0 = \bar{y} - \frac{\hat{\beta}_1}{\bar{x}} \bar{x} = \bar{y} - \frac{\hat{\beta}_1}{\bar{x}} = \hat{\beta}_0 \text{ same as before}. \]
Problem 3 (25 points)
Answer the following questions:

a. Consider the regression of $y$ on $x$. Please describe what you see in the following graphs. Explain how the vertical lines can be used to compute $R^2$. Note: The horizontal line around the middle of the graph represents $\bar{y}$.

\[
\begin{align*}
\text{SST} & = \sum (y_i - \bar{y})^2 \\
\text{SSE} & = \sum (y_i - \hat{y}_i)^2 \\
\text{SSR} & = \sum (\hat{y}_i - \bar{y})^2
\end{align*}
\]

\[
R^2 = \frac{\text{SSR}}{\text{SST}} \quad \text{SQUARE LINES IN (3) SUM THE UP} \quad \text{AND DIVIDE BY SUM OF SQUARES (1) LINES IN (1)}.
\]

b. Explain how the permutation test can be used to test the hypothesis $H_0: \beta_1 = 5$ against the alternative $H_0: \beta_1 \neq 5$. Please give all the details.

\[
y_i = \hat{y}_0 + \hat{\beta}_1 x_i + \varepsilon_i.
\]

PERMUTE $y_i - \hat{y}_i$ TO GENERATE MANY DATA SPS. COUNT HOW MANY OF THE SIMULATED DATA ACCEPT THE ACTUAL $\hat{\beta}_1$ AND DIVIDE BY 1000 TO GET THE P-VALUE. IF P-VALUE $< 0.01$ REJECT $H_0$.

\[
y_i - \hat{y}_i = \hat{y}_0 + (\hat{\beta}_1 - 5) x_i + \varepsilon_i,
\]

\[
y_i - \hat{y}_i = \hat{y}_0 + \hat{\beta}_1 x_i + \varepsilon_i.
\]

c. Assume that the variance of stock $A$ is 0.16 and the variance of stock $B$ is 0.25. The variance of a portfolio consisting of 50% $A$ and 50% $B$ is 0.0525. What is the correlation between stocks $A$ and $B$?

\[
\sigma_{A,B} = \sigma^2 (0.16) + \sigma^2 (0.25) + 2 (0.5) (0.5) \text{ COV (A,B)} \Rightarrow \text{COV (A,B)} = -0.1
\]

\[
\text{CORR (A,B)} = \frac{\text{COV (A,B)}}{\text{SD (A)} \text{ SD (B)}} = \frac{-0.1}{\sqrt{0.16} \sqrt{0.25}} = -0.5
\]

d. Consider a portfolio consisting of 5 stocks. How many variances and how many covariances are there?

Show them on the variance covariance matrix.

\[
\begin{array}{c}
\text{5 VARIANCES} \\
\text{10 COVARIANCES}
\end{array}
\]

\[
\begin{pmatrix}
\sigma_1 & \sigma_2 & \sigma_3 & \sigma_4 & \sigma_5 \\
\sigma_2 & \sigma_6 & \sigma_7 & \sigma_8 & \sigma_9 \\
\sigma_3 & \sigma_7 & \sigma_{10} & \sigma_{11} & \sigma_{12} \\
\sigma_4 & \sigma_8 & \sigma_{11} & \sigma_{13} & \sigma_{14} \\
\sigma_5 & \sigma_9 & \sigma_{12} & \sigma_{14} & \sigma_{15}
\end{pmatrix}
\]
Problem 4  (25 points)
Answer the following questions:

a. You are given the following information on two variables (ppm of Cadmium and Cobalt at 359 spatial locations):

\[
\begin{align*}
    h_{11} &= \frac{1}{n} + \frac{\sum (x_i - \bar{x})^2}{n \sum (x_i - \bar{x})^2} \\
    &= \frac{1}{359} + \frac{\left(2.28 - \frac{4.39}{359-1} \frac{12.7326}{12.7326}ight)^2}{359-1} \\
    &= 0.0031 < \frac{2}{359} \times 0.0111
\end{align*}
\]

1. Consider the regression of Cd on Co. Compute the leverage value of the first data point. Is it a high leverage point? Explain.

2. Describe the shape of the histogram that corresponds to the ecdf shown below.

b. Data on 6 pairs of x and y gave the following information:

\[
\begin{align*}
    \sum x_i &= 9.545, \sum y_i = 61.668, \sum x_i^2 = 15.78468, \sum y_i^2 = 719.9573, \sum x_i y_i = 96.43722.
\end{align*}
\]

It was discovered that the last pair of x, y values (1.565, 3.508) was not part of the data set and it was deleted. Find \( \hat{\beta}_1 \) and \( \hat{\beta}_0 \) after the last pair was deleted form the data set.

\[
\begin{align*}
    \sum x_i &= 9.545 - 1.565 = 7.98 \\
    \sum y_i &= 61.668 - 3.508 = 58.16 \\
    \sum x_i^2 &= 15.78468 - 1.565^2 = 13.33545 \\
    \sum y_i^2 &= 719.9573 - 3.508^2 = 707.6512 \\
    \sum x_i y_i &= 96.43722 - (1.565)(3.508) = 90.9472
\end{align*}
\]

\[
\begin{align*}
    \hat{\beta}_1 &= \frac{90.9472}{13.33545 - \frac{7.98^2}{5}} = \frac{7.98}{5} \\
    \hat{\beta}_0 &= \frac{58.16 - \hat{\beta}_1 \frac{7.98}{5}}{5}
\end{align*}
\]

\( R^2 \) using the data set after the last pair was deleted form the data set.

\[
R^2 = \frac{\hat{\beta}_1 \left( \frac{\sum (x_i - \bar{x})^2}{\sum (x_i^2 - \bar{x}^2)} \right)}{\sum (y_i - \bar{y})^2} = \frac{(7.98)^2}{707.6512} \approx 0.1886
\]

\[\hat{\beta}_1 = -3.13 \]

\[\hat{\beta}_0 = 16.63\]