Name: SOLUTIONS UCLA ID: ________ Section: _____

Problem 1 **(25 points)**
Consider the following data:

y	x
y_{11}	1
y_{21}	1
y_{31}	1
$\vdots$	$\vdots$
$\vdots$	$\vdots$
y_{n11}	1
y_{12}	0
y_{22}	0
y_{32}	0
$\vdots$	$\vdots$
$\vdots$	$\vdots$
y_{n22}	0

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

$$\Sigma (x_i - \bar{x})^2 = \Sigma x_i^2 - n\bar{x}^2 = n_1 - (n_1+n_2)\frac{n_1^2}{(n_1+n_2)^2}$$

$$= n_1 - \frac{n_1^2}{n_1+n_2} = \frac{n_1^2 + n_1 n_2 - n_1^2}{n_1+n_2} = \frac{n_1 n_2}{n_1+n_2}$$

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.

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

$$= \frac{n_1^2\bar{y}_1 + n_1 n_2 \bar{y}_1 - n_1^2\bar{y}_1 - n_1 n_2 \bar{y}_2}{n_1 n_2} = \bar{y}_1 - \bar{y}_2$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1+n_2} - (\bar{y}_1 - \bar{y}_2)\frac{n_1}{n_1+n_2}$$

$$= \frac{n_1\bar{y}_1 + n_2\bar{y}_2 - n_1\bar{y}_1 + n_1\bar{y}_2}{n_1+n_2} = \bar{y}_2.$$

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

$$R^2 = \frac{\hat{\beta}_1^2 \Sigma (x_i - \bar{x})^2}{SST} = \frac{(\bar{y}_1 - \bar{y}_2)^2 \frac{n_1 n_2}{n_1+n_2}}{(n-1)s_y^2} = \frac{(-0.46 - 0.28)^2 \frac{150}{25}}{24\,(0.8)}$$

$$R^2 = 17.1\%$$

Problem 2 (225 points)
Answer the following questions:

a. Consider the simple regression of y on x. Suppose we transform the x values using $x_i = \frac{x_i - \bar{x}}{s_x}$ and the y values using $y_i = \frac{y_i - \bar{y}}{s_y}$. 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).

$$\hat{\beta}_1 = \frac{\sum \frac{(y_i - \bar{y})}{s_y}\frac{(x_i - \bar{x})}{s_x}}{\frac{\sum (x_i - \bar{x})^2}{s_x^2}} = \frac{\frac{1}{n-1}\sum (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n-1}\sum (x_i - \bar{x})^2} \frac{s_x^2}{s_x s_y} = \frac{\text{cov}(x,y)}{s_x s_y} = r.$$

YES.

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

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

$\bar{y}$ OF THE TRANSFORMED y IS ZERO

$\bar{x}$ OF THE TRANSFORMED x 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.

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

$$\hat{\beta}_0^* = \bar{y}^* - \hat{\beta}_1^* \bar{x}^* = 5\bar{y} - \frac{5}{4}\hat{\beta}_1 4\bar{x} = 5\hat{\beta}_0.$$

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.

$$R^2 = \frac{\hat{\beta}_1^2 \sum (x_i - \bar{x})^2}{SST} = \frac{\frac{5}{4}\hat{\beta}_1^2 \, 4 \sum (x_i - \bar{x})^2}{5 \sum (y_i - \bar{y})^2} = R^2$$

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.

WHEN $\bar{x} = 0$

$$\boxed{\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2} \quad \hat{\beta}_0 = \bar{y}}$$

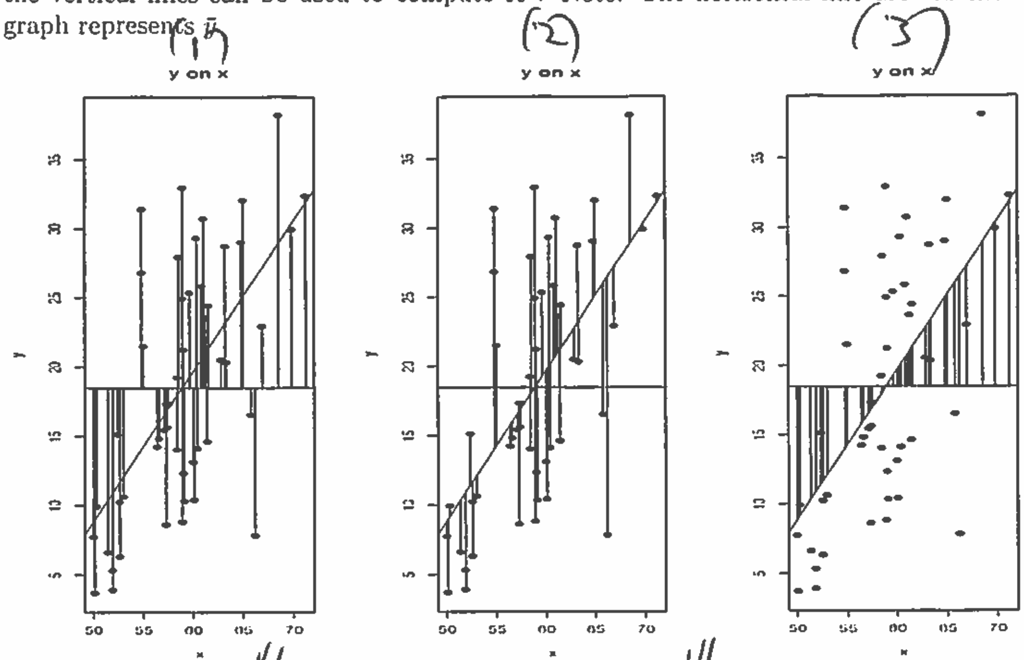
IF WE MULTIPLY x BY 4:

$$\hat{\beta}_1 = \frac{4 \sum x_i y_i}{4^2 \sum x_i^2} = \frac{\hat{\beta}_1}{4}.$$

$$\hat{\beta}_0 = \bar{y} - \frac{\hat{\beta}_1}{4}\underbrace{\bar{x}}_{0} = \hat{\beta}_0$$ 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 $\bar{y}$.



$$SST = \sum (y_i - \bar{y})^2 \quad // \quad SSE = \sum (y_i - \hat{y}_i)^2 \quad /// \quad SSR = \sum (\hat{y}_i - \bar{y})^2$$

$$R^2 = SSR/SST$$

SUM UP THE SQUARE LINES IN (3) AND DIVIDE BY SUM OF SQUARED LINES IN (1).

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

$H_0 : \beta_1 - 5 = 0$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$
$$y_i - 5x_i = \beta_0 + (\beta_1 - 5)x_i + \varepsilon_i$$
$$y_i - 5x_i = \beta_0 + \beta_1^* x_i + \varepsilon_i$$

PERMUTE $y_i - 5x_i$ TO GENERATE (1000) MANY DATA SETS. COUNT HOW MANY OF THE SIMULATED BETAS EXCEED THE ACTUAL $\hat{\beta}_1^*$ AND DIVIDE BY 1000 TO GET THE P-VALUE. IF P-VALUE $< \alpha = 0.01$ REJECT H_0.

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?

$$0.0525 = 0.5^2(0.16) + 0.5^2(0.25) + 2(0.5)(0.5)\,\mathrm{cov}(A,B) \Rightarrow \mathrm{cov}(A,B) = -0.1$$

$$\mathrm{CORR}(A,B) = \mathrm{COV}(A,B) / \mathrm{SD}(A)\,\mathrm{SD}(B) = \frac{-0.1}{\sqrt{0.18}\sqrt{0.25}} = -0.5$$

d. Consider a portfolio consisting of 5 stocks. How many variances and how many covariances are there. Show them on the variance covariance matrix.

5 VARIANCES

10 COVARIANCES

$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} \\ 2 & \sigma_{21} & \sigma_2^2 & \sigma_{23} & \sigma_{24} & \sigma_{21} \\ 3 & \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{34} & \sigma_{35} \\ 4 & \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 & \sigma_{45} \\ 5 & \sigma_{51} & \sigma_{52} & \sigma_{53} & \sigma_{54} & \sigma_5^2 \end{array}$$

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):

```
summary(a)
       Cd                Co
 Min.   :0.1350   Min.   : 1.552
 1st Qu.:0.6525   1st Qu.: 6.660
 Median :1.1000   Median : 9.840
 Mean   :1.2882   Mean   : 9.439
 3rd Qu.:1.6800   3rd Qu.:12.100
 Max.   :5.1290   Max.   :20.600

var(a$Cd)
0.7380493

var(a$Co)
12.73241

#First 3 observations of the data set:
head(a)
     Cd    Co
1 1.570  8.28
2 2.045 10.80
3 1.203 12.00
.   .     .
.   .     .
```

$$h_{11} = \frac{1}{n} + \frac{(x_1-\bar{x})^2}{\Sigma(x_i-\bar{x})^2}$$

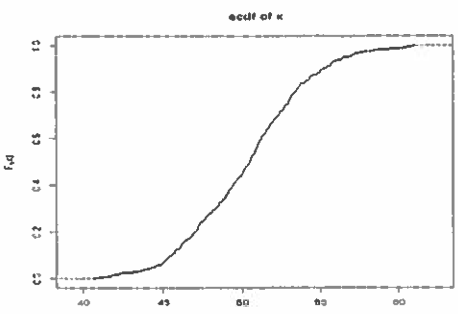
$$= \frac{1}{359} + \frac{(8.28 - \cancel{9.439})^2}{(359-1)\,12.73241}$$

$$h_{11} = 0.0031 < \underbrace{2 \times \frac{2}{359}}_{0.0111}$$

No!

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.



SYMMETRICAL!

b. Data on 6 pairs of x and y gave the following information:
$\sum_{i=1}^{6} x_i = 9.545$, $\sum_{i=1}^{6} y_i = 61.668$, $\sum_{i=1}^{6} x_i^2 = 15.78468$, $\sum_{i=1}^{6} y_i^2 = 719.9573$, $\sum_{i=1}^{6} x_i y_i = 96.43722$.
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.

New $\Sigma x_i = 9.545 - 1.565 = 7.98$ New $\Sigma y_i = 61.668 - 3.508 = 58.16$

New $\Sigma x_i^2 = 15.78468 - 1.565^2 = 13.33545$ New $\Sigma y_i^2 = 719.9573 - 3.508^2 = 707.6512$

New $\Sigma x_i y_i = 96.43722 - (1.565)(3.508) = 90.9472$ $\hat{\beta}_1 = \dfrac{90.9472 - \frac{1}{5}(7.98)(58.16)}{13.33545 - \frac{7.98^2}{5}}$

c. Refer to question (b). Find R^2 using the data set after the last pair was deleted form the data set.

$\Rightarrow \hat{\beta}_1 = -3.13$

$$R^2 = \frac{\hat{\beta}_1^2\,\Sigma(x_i-\bar{x})^2}{\Sigma(y_i-\bar{y})^2} = \frac{(-3.13)^2\left(13.33545 - 7.98^2/5\right)}{707.6512 - \frac{58.16^2}{5}}$$

$$\hat{\beta}_0 = \frac{58.16}{5} - (-3.13)\frac{7.98}{5}$$

$\hat{\beta}_0 = 16.63$

$R^2 = 18.86\,\%$