Regression - practice questions

EXERCISE 1
Data have been collected for 19 observations of two variables, y and x, in order to run a regression of y on x. You are given that \( s_y = 10 \), \( \sum_{i=1}^{19} (y_i - \hat{y}_i)^2 = 180 \).

a. Compute the proportion of the variation in y that can be explained by x. [Ans. 0.90]

b. Compute the standard error of the estimate (s_e). [Ans. 3.25]

EXERCISE 2
Data on y and x were collected to run a regression of y on x. The intercept is included. You are given the following: \( \bar{x} = 76 \), \( \bar{y} = 880 \), \( \sum_{i=1}^{n} (x_i - \bar{x})^2 = 6800 \), \( \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = 14200 \), \( r_{xy} = 0.72 \), \( s_e = 20.13 \).

a. What is the value of \( \hat{\beta}_1 \)? [Ans. 2.088]

b. What is the value of \( \hat{\beta}_0 \)? [Ans. 721.312]

c. What is the value of \( \sum_{i=1}^{n} (y_i - \bar{y})^2 \)? [Ans. 57188]

d. What is the sample size n? [Ans. 70]

EXERCISE 3
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>\vdots</td>
<td>\vdots</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>\vdots</td>
<td>\vdots</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 \( n_1 \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{n_1 n_2}{n_1 + n_2} \).

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.

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 \( R^2 \).
**EXERCISE 4**

Answer the following questions:

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

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

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.

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.

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.

**EXERCISE 5**

Answer the following questions:

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

```r
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
```

```r
var(a$Cd)
0.7380493

var(a$Co)
12.73241
```

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

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

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

$\sum_{i=1}^{6} x_i = 95.45$, $\sum_{i=1}^{6} y_i = 61.668$, $\sum_{i=1}^{6} x_i^2 = 157.8468$, $\sum_{i=1}^{6} y_i^2 = 719.9573$, $\sum_{i=1}^{6} x_i y_i = 96.43722$.

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.

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