

Problem Set 2.

1. Show that if r is the correlation coefficient between n pairs of variables (X_i, Y_i) , then the squared correlation between the n pairs $(aX_i + b, cY_i + d)$, where a, b, c and d are constants, is also r^2 .

Answer:

Define

$$\begin{aligned} X'_i &= aX_i + b, \\ Y'_i &= cY_i + d. \end{aligned} \tag{1}$$

Averaging over all the observations gives

$$\begin{aligned} \bar{X}' &= a\bar{X} + b, \\ \bar{Y}' &= c\bar{Y} + d. \end{aligned}$$

Subtracting the above expressions from equation (1) gives

$$\begin{aligned} x'_i &= ax_i, \\ y'_i &= cy_i. \end{aligned}$$

Next, the correlation coefficient between X'_i and Y'_i is given by

$$\begin{aligned} r' &= \frac{\sum x'_i y'_i}{\sqrt{\sum (x'_i)^2} \sqrt{\sum (y'_i)^2}} = \\ &= \frac{\sum a x_i c y_i}{\sqrt{\sum a^2 x_i^2} \sqrt{\sum c^2 y_i^2}} = \\ &= \frac{ac \sum x_i y_i}{ac \sqrt{\sum x_i^2} \sqrt{\sum y_i^2}} = \\ &= \frac{\sum x_i y_i}{\sqrt{\sum x_i^2} \sqrt{\sum y_i^2}} = r. \end{aligned}$$

2. From a sample of 200 observations the following quantities were calculated:

$$\begin{aligned}\sum X &= 11.34, & \sum Y &= 20.72, & \sum X^2 &= 12.16 \\ \sum Y^2 &= 84.96, & \sum XY &= 22.13.\end{aligned}$$

Estimate both regression equations: the regression of Y on X , and the regressions of X on Y .

In order to estimate the slope coefficients:

$$\beta_{YX} = \frac{\sum yx}{\sum x^2}, \quad \beta_{XY} = \frac{\sum yx}{\sum y^2}.$$

we use the following expressions

$$\begin{aligned} \sum yx &= \sum YX - N\bar{Y}\bar{X} = \\ \sum YX - N\frac{\sum Y}{N}\frac{\sum X}{N} &= \sum YX - \frac{\sum Y \sum X}{N}, \\ \sum x^2 &= \sum X^2 - N(\bar{X})^2 = \\ \sum X^2 - N\frac{(\sum X)^2}{N^2} &= \sum X^2 - \frac{(\sum X)^2}{N}, \\ \sum y^2 &= \sum Y^2 - N(\bar{Y})^2 = \\ \sum Y^2 - N\frac{(\sum Y)^2}{N^2} &= \sum Y^2 - \frac{(\sum Y)^2}{N}, \end{aligned}$$

We obtain

$$\begin{aligned} \sum yx &= 22.13 - \frac{20.72 \times 11.34}{200} = 20.95, \\ \sum x^2 &= 12.16 - \frac{(11.34)^2}{200} = 11.52, \\ \sum y^2 &= 84.96 - \frac{(20.72)^2}{200} = 82.81. \end{aligned}$$

It follows that

$$\beta_{YX} = \frac{20.95}{11.52} = 1.82,$$
$$\beta_{XY} = \frac{20.95}{82.81} = 0.25,$$

and

$$\alpha_{YX} = \bar{Y} - \beta_{YX}\bar{X} = 0.10 - 1.82 \times 0.06 = -0.01,$$
$$\alpha_{XY} = \bar{X} - \beta_{XY}\bar{Y} = 0.06 - 0.25 \times 0.10 = 0.035.$$

3. Prove that r^2 (the correlation coefficient squared) in the regression of X on Y can be expressed as

$$r^2 = \beta_{yx}\beta_{xy},$$

where the β 's are the LS slopes in the respective regressions.

Answer:

We have:

$$\begin{aligned} r &= \frac{\sum yx}{\sqrt{\sum x^2}\sqrt{\sum y^2}} \Rightarrow r^2 = \frac{\sum yx \sum yx}{\sum x^2 \sum y^2} \\ &= \frac{\sum yx}{\sum x^2} \times \frac{\sum yx}{\sum y^2} = \beta_{YX}\beta_{XY}. \end{aligned}$$

4. Show that if r is the correlation coefficient between n pairs of variables (X_i, Y_i) , then the squared correlation coefficient may be interpreted as the proportion of Y variation attributable to the linear regression on X .

Answer:

We have to prove that

$$\sum y^2 = \sum \hat{e}^2 + r^2 \sum y^2.$$

We start from the bivariate regression:

$$\begin{aligned} Y &= \alpha + \beta X + \hat{e} \Rightarrow \\ \bar{Y} &= \alpha + \beta \bar{X}. \end{aligned}$$

Subtracting the second equation from the first gives

$$\begin{aligned} y &= \beta x + \hat{e} \Rightarrow \hat{e} = y - \beta x \\ \Rightarrow \hat{e}^2 &= y^2 + \beta^2 x^2 - 2\beta yx \Rightarrow \\ \sum \hat{e}^2 &= \sum y^2 + \beta^2 \sum x^2 - 2\beta \sum xy. \end{aligned} \quad (2)$$

Next, recall that

$$\begin{aligned}\beta &= \frac{\sum xy}{\sum x^2} \Rightarrow \beta \sum x^2 = \sum xy \Rightarrow \\ &\Rightarrow \beta^2 \sum x^2 = \beta \sum xy.\end{aligned}$$

Thus substituting the above expression into equation (2), yields

$$\begin{aligned}\sum \hat{e}^2 &= \sum y^2 + \beta \sum xy - 2\beta \sum xy = \\ &= \sum y^2 - \beta \sum xy \Rightarrow \\ \sum y^2 &= \beta \sum xy + \sum \hat{e}^2.\end{aligned}$$

Using

$$\begin{aligned}\beta &= \frac{\sum xy}{\sum x^2} \Rightarrow \beta \sum xy = \frac{\sum xy}{\sum x^2} \times \sum xy = \frac{(\sum xy)^2}{\sum x^2} = \\ &= \frac{(\sum xy)^2}{\sum x^2} \times \frac{\sum y^2}{\sum y^2} = \frac{(\sum xy)^2}{\sum x^2 \sum y^2} \times \sum y^2 = r^2 \sum y^2,\end{aligned}$$

we obtain

$$\sum y^2 = \sum \hat{e}^2 + r^2 \sum y^2,$$

or

$$\text{TSS} = \text{RSS} + \text{ESS}.$$

Rearranging the above expression gives

$$r^2 = 1 - \frac{\sum \hat{e}^2}{\sum y^2},$$

or

$$r^2 = 1 - \frac{\text{RSS}}{\text{TSS}}.$$

5. Consider the numerical example 1.4.5 in Johnston and Dinardo:

a) Obtain the regression coefficients α and β .

b) Calculate the explained and residual sum of squares as well as the correlation coefficient between Y and X (r).

c) Obtain the estimated standard errors of the regression coefficients.

Answer:

5. a)

$$\beta = \frac{\sum xy}{\sum x^2} = \frac{70}{40} = 1.75,$$
$$\alpha = \bar{Y} - \beta\bar{X} = 8 - 1.75 \times 4 = 1.$$

b)

$$r^2 = \frac{(\sum yx)^2}{\sum x^2 \sum y^2} = \frac{70^2}{40 \times 124} = \frac{4900}{4960} = 0.99.$$

Moreover,

$$\text{ESS} = r^2 \sum y^2 = 123,$$

$$\text{RSS} = \sum y^2 - r^2 \sum y^2 = 124 - 123 = 1.$$

c)

$$s^2 = \frac{\sum \hat{e}^2}{N - 2} = \frac{1}{3} = 0.33.$$