Problem Set 1.

Questions.

1. Consider the numerical example in Johnston and Dinardo in section 1.4.5.

a) Use two alternative ways to calculate the correlation coefficient between X and Y (r_{xy}).

b) Show that using the following form you will get the same answer as in part a:

$$\frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}}$$

2. Assume that Y is a linear combination of X. Prove that $r_{xy} = 1$.

3. A sample of 20 observations corresponding to the model

$$Y = a + bX + e,$$

where the e's are normally and independently distributed with zero mean and constant variance, gave the following data:

$$\begin{aligned} \sum Y &= 21.9, \ \sum(Y - \overline{Y})^2 = 86.9, \\ & \sum(X - \overline{X})(Y - \overline{Y}) \boxed{=} 106.4, \\ \sum X &= 186.2, \ \sum(X - \overline{X})^2 = 215.4. \end{aligned}$$

Estimate a and b.

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

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

Compute r_{xy}.

5. Solve Exercise 1.15a in Johnston and Dinardo.

Answers

1. a)

$$r = \frac{\sum yx}{\sqrt{\sum x^2}\sqrt{\sum y^2}} = \frac{70}{\sqrt{40}\sqrt{124}} = \frac{70}{70.34} = 0.99.$$

b)

$$r = \frac{\sum yx}{\sqrt{\sum x^2}\sqrt{\sum y^2}}.$$

We will use the following expressions

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

Substituting the above expressions into the equation for r gives

$$r = \frac{\sum YX - \frac{\sum Y \sum X}{N}}{\sqrt{\sum X^2 - \frac{(\sum X)^2}{N}}\sqrt{\sum Y^2 - \frac{(\sum Y)^2}{N}}}.$$

Multiply and divide by N gives

$$r = \frac{N \sum YX - \sum Y \sum X}{\sqrt{N \sum X^2 - (\sum X)^2}\sqrt{N \sum Y^2 - (\sum Y)^2}}.$$

2. Let

$$Y = a + bX.$$

From the above expression it follows:

$$\begin{aligned} \mathsf{Var}(Y) &= b^2\mathsf{Var}(X), \\ \mathsf{Cov}(Y,X) &= b\mathsf{Cov}(X,X) = b\mathsf{Var}(X). \end{aligned}$$

The theoretical correlation between the two variables is given by

$$\rho = \frac{\mathsf{Cov}(Y,X)}{\sqrt{\mathsf{Var}(X)}\sqrt{\mathsf{Var}(Y)}}.$$

Using the above expressions, we have

$$\rho = \frac{b\mathsf{Var}(X)}{\sqrt{\mathsf{Var}(X)}\sqrt{b^2\mathsf{Var}(X)}} = \frac{b\mathsf{Var}(X)}{b\mathsf{Var}(X)} = 1.$$

3. The estimated value of the slope coefficient is given by

$$\beta = \frac{\sum(X - \overline{X})(Y - \overline{Y})}{\sum(X - \overline{X})^2} = \frac{106.4}{215.4} = 0.49.$$

The estimated value for the constant is given by

$$\begin{aligned} \alpha &= \overline{Y} - \beta\overline{X} = \frac{21.9}{20} - 0.49\frac{186.2}{20} \\ &= 1.09 - 0.49 \times 9.31 = -3.47. \end{aligned}$$

4. The sample correlation coefficient is given by

$$
\begin{aligned}
r &= \frac{N\sum YX - \sum Y \sum X}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}} = \\
&= \frac{200 \times 22.13 - 20.72 \times 11.34}{\sqrt{200 \times 12.16 - (11.34)^2}\sqrt{200 \times 84.96 - (20.72)^2}} \\
&= \ldots
\end{aligned}
$$

5. You have to use the following form

$$
r = \frac{\sum yx}{\sqrt{\sum x^2}\sqrt{\sum y^2}}.
$$