Unit 10 · Correlation and Regression

HARDER PRACTICE QUESTIONS

Full marks 27

Write your answers on separate paper. Show your working. Give r to three decimal places.

1 Turning the axes round [4]

For five of Nepal's weather stations, the height x in hundreds of meters and the average year's temperature y in °C gave $r = -0.996$, $b_{yx} = -0.5131$ (°C per 100 m) and $b_{xy} = -1.9345$ (hundreds of meters per °C).

A climber prefers to measure the depth d below a ridge 2,400 m high, in kilometers, so $d = \dfrac{2400 - 100x}{1000}$, and her thermometer reads in °F, $F = 1.8y + 32$.

a. Find the correlation coefficient between d and F. Nabin says it must still be -0.996, because "a change of origin and scale never changes r". Is he right?

b. Find the regression coefficient of F on d, with its units, and the regression coefficient of d on F.

c. Check your two answers in b against your answer in a.

2 Two wrong entries [6]

Ten students' hours of study a week x and marks y were entered in a spreadsheet, which gave $n = 10$, $\Sigma x = 90$, $\Sigma y = 428$, $\Sigma x^2 = 940$, $\Sigma y^2 = 18\,836$, $\Sigma xy = 4012$.

Later two entries were found to be wrong: the pair (6, 33) had been entered as (6, 53), and the pair (11, 46) as (16, 46).

a. Find r from the sums as they were entered.

b. Find the correct sums and the correct r.

c. Find the correct regression line of y on x.

d. Two wrong entries out of ten changed r a lot. Explain why such an entry can do so much harm.

3 What "best" fit means [5]

Five pairs of data: (1, 3), (2, 4), (3, 8), (4, 9), (5, 11).

a. Find the regression line of y on x.

b. From the first normal equation, $\Sigma y = na + b\Sigma x$, show that the misses $y - (a + bx)$ of the least squares line always add up to zero.

c. Susan draws the line through the first and the last points, $y = 1 + 2x$. She says: "My misses add up to zero too, so my line is just as good." Check her claim about the misses, and decide whether her line is as good, by the rule the least squares line is chosen by.

4 The widest disagreement [5]

a. Two judges rank n singers, and Judge 2's order is Judge 1's exactly reversed, so singer k gets ranks k and $n + 1 - k$. Using $\Sigma k = \dfrac{n(n+1)}{2}$ and $\Sigma k^2 = \dfrac{n(n+1)(2n+1)}{6}$, show that $\Sigma d^2 = \dfrac{n(n^2-1)}{3}$, and hence that $\rho = -1$.

b. For 10 students ranked in two subjects, Nirjala found $\rho = 0.6$. She then noticed that one difference had been written as 3 when it was really 7. Find the correct ρ.

c. Could a slip in copying ever give $\rho = -1.2$ for 10 students with no ties? Give a reason.

5 From two lines to the spread [4]

The two regression lines of some data are $2x + 5y = 32$ and $5x + 8y = 62$, and the standard deviation of x is 3.

a. Find the standard deviation of y.

b. Pratima wants the value of x when $y = 6$. She puts $y = 6$ into $2x + 5y = 32$ and gets $x = 1$. Find the right estimate, and say why hers is not it.

6 A curve that hides from r [3]

The five points $(-2, 4)$, $(-1, 1)$, $(0, 0)$, $(1, 1)$, $(2, 4)$ all lie on $y = x^2$, and their r is 0.

a. Add the point $(3, 9)$, also on the curve, and find the new r.

b. A sixth point (t, t^2) on the curve is added instead. Find every value of t for which r stays 0.