Unit 10 · Correlation and Regression

HARDER PRACTICE QUESTIONS · ANSWER KEY

Full marks 27

Give the marks for the points in the marking scheme; an answer in other words that says the same is right.

1 Turning the axes round

HA · 4

ANSWER

Tests: a change of scale by a **negative** number flips the sign of r; the regression coefficients carry units and change by the scale factors; $b_{yx}\, b_{xy} = r^2$ still holds.

a. $d = 2.4 - 0.1x$: a change of origin and a change of scale by -0.1, a negative number. $F = 1.8y + 32$: a scale of $+1.8$. A positive scale leaves r as it is; a negative one flips its sign (going deeper means going lower). So $r_{dF} = \mathbf{+0.996}$. Nabin is wrong: the rule holds only for positive scale factors.

b. F changes 1.8 times as much as y, and one kilometer of depth is -10 hundreds of meters of height:

$b_{Fd} = 1.8 \times (-0.5131) \times (-10) = \mathbf{9.236}$ **°F per km** of depth.

d changes -0.1 times as much as x, and one °F is $\dfrac{1}{1.8}$ °C:

$b_{dF} = (-1.9345) \times (-0.1) \div 1.8 = \mathbf{0.1075}$ **km per °F**.

c. $9.236 \times 0.1075 = 0.9926 \approx (0.996)^2 = 0.992$ ✓, and both coefficients are positive, like r. (The tempting answer: $r = -0.996$ and $b_{Fd} = -0.5131$, as if nothing had changed.)

MARKING SCHEME

✓ d is a change of scale by a negative number, so r changes sign: $+0.996$; Nabin is wrong.

✓ $b_{Fd} = 1.8 \times (-0.5131) \times (-10) = 9.236$ °F per km.

✓ $b_{dF} = 0.1075$ km per °F.

✓ The product $\approx 0.992 = r^2$, both positive.

Science Playground, Katunje, Bhaktapur
Book a class visit: https://science-playground.com/booking

2 Two wrong entries

ANSWER A · 6

Tests: correcting every sum a wrong pair enters (Σx, Σy, Σx^2, Σy^2 and Σxy); the sums form of r; how far one point off the line pulls r.

a. $n\Sigma xy - \Sigma x\Sigma y = 40\,120 - 38\,520 = 1600$; $n\Sigma x^2 - (\Sigma x)^2 = 9400 - 8100 = 1300$;
$n\Sigma y^2 - (\Sigma y)^2 = 188\,360 - 183\,184 = 5176$.

$r = \dfrac{1600}{\sqrt{1300 \times 5176}} = \dfrac{1600}{2594} = \mathbf{0.617}$.

b. Take each wrong pair out and put the right one in:

$\Sigma x = 90 - 16 + 11 = 85$; $\Sigma y = 428 - 53 + 33 = 408$;

$\Sigma x^2 = 940 - 256 + 121 = 805$; $\Sigma y^2 = 18\,836 - 2809 + 1089 = 17\,116$;

$\Sigma xy = 4012 - 6 \times 53 + 6 \times 33 - 16 \times 46 + 11 \times 46 = 4012 - 120 - 230 = 3662$.

$n\Sigma xy - \Sigma x\Sigma y = 36\,620 - 34\,680 = 1940$; $n\Sigma x^2 - (\Sigma x)^2 = 8050 - 7225 = 825$;

$n\Sigma y^2 - (\Sigma y)^2 = 171\,160 - 166\,464 = 4696$.

$r = \dfrac{1940}{\sqrt{825 \times 4696}} = \dfrac{1940}{1968.3} = \mathbf{0.986}$.

c. $b_{yx} = \dfrac{1940}{825} = 2.352$, $\bar{x} = 8.5$, $\bar{y} = 40.8$: $y - 40.8 = 2.352(x - 8.5)$, that is $\mathbf{y = 20.81 + 2.352}x$.

d. The other eight points lie close to a rising line. A wrong entry puts a point far off that line (53 marks at 6 hours, 16 hours with only 46 marks), and r works with **squares** of the deviations, so one far-off point adds a large amount to $\Sigma(y - \bar{y})^2$ or $\Sigma(x - \bar{x})^2$ while adding little to the product sum: r fell from 0.986 to 0.617. (The tempting slip in b: correcting Σx and Σy but not Σx^2, Σy^2 and Σxy.)

MARKING SCHEME

- ✓ The three bracketed quantities from the entered sums (1600, 1300, 5176).
- ✓ $r = 0.617$ as entered.
- ✓ Corrected $\Sigma x = 85$, $\Sigma y = 408$, $\Sigma x^2 = 805$, $\Sigma y^2 = 17\,116$.
- ✓ Corrected $\Sigma xy = 3662$.
- ✓ Correct $r = 0.986$, and $y = 20.81 + 2.352x$.
- ✓ A point far from the line adds a large square to the deviations' sums, so it pulls r down.

3 What "best" fit means

ANSWER HA · 5

Tests: what the least squares line makes least (the sum of the squared misses), not the sum of the misses; reading the normal equation.

a. $\bar{x} = 3$, $\bar{y} = 7$. $X = -2, -1, 0, 1, 2$; $Y = -4, -3, 1, 2, 4$. $\Sigma XY = 8 + 3 + 0 + 2 + 8 = 21$, $\Sigma X^2 = 10$.
$b_{yx} = 2.1$: $y - 7 = 2.1(x - 3)$, $\mathbf{y = 0.7 + 2.1}x$.

b. $\Sigma(y - a - bx) = \Sigma y - na - b\Sigma x$, and the first normal equation says this is 0. ✓

c. Her misses: $3 - 3 = 0$, $4 - 5 = -1$, $8 - 7 = 1$, $9 - 9 = 0$, $11 - 11 = 0$: they add to 0, so the claim about them is true. (Any line through $(\bar{x}, \bar{y}) = (3, 7)$ has misses adding to zero, and hers passes through it.)

But the least squares line is the one that makes $\Sigma(\text{miss})^2$ least. Hers: $0 + 1 + 1 + 0 + 0 = 2$. The regression line predicts 2.8, 4.9, 7.0, 9.1, 11.2, so its misses are 0.2, −0.9, 1.0, −0.1, −0.2, squares adding to $0.04 + 0.81 + 1 + 0.01 + 0.04 = \mathbf{1.9}$. $1.9 < 2$, so her line is **not** as good; no line does better than 1.9. (The tempting idea: misses adding to zero make a line best.)

MARKING SCHEME

- ✓ $y = 0.7 + 2.1x$.
- ✓ $\Sigma(y - a - bx) = \Sigma y - na - b\Sigma x = 0$ by the normal equation.
- ✓ Her misses add to 0 (her line passes through the means).
- ✓ Her squared misses add to 2; the regression line's to 1.9.
- ✓ The best line makes the squares least, so hers is not as good.

Science Playground, Katunje, Bhaktapur
Book a class visit: https://science-playground.com/booking

4 The widest disagreement

ANSWER HA · 5

Tests: why ρ can't go below -1 (the largest Σd^2, with Grade 11's sums of series); working back from ρ to Σd^2.

a. $d = k - (n+1-k) = 2k - (n+1)$, so $d^2 = 4k^2 - 4(n+1)k + (n+1)^2$.

$$\Sigma d^2 = 4 \cdot \frac{n(n+1)(2n+1)}{6} - 4(n+1) \cdot \frac{n(n+1)}{2} + n(n+1)^2$$

$$= n(n+1)\left[\frac{2(2n+1)}{3} - 2(n+1) + (n+1)\right]$$

$$= n(n+1) \cdot \frac{n-1}{3} = \frac{n(n^2-1)}{3}.$$

Then $\rho = 1 - \frac{6}{n(n^2-1)} \cdot \frac{n(n^2-1)}{3} = 1 - 2 = -1$. ✓

b. $0.6 = 1 - \frac{6\Sigma d^2}{10 \times 99}$, so $\Sigma d^2 = \frac{0.4 \times 990}{6} = 66$. Corrected: $66 - 9 + 49 = 106$.

$$\rho = 1 - \frac{6 \times 106}{990} = 1 - 0.642 = \mathbf{0.358}.$$

c. No. Reversed orders give the largest possible Σd^2, $\frac{10 \times 99}{3} = 330$, and so the least ρ, -1. $\rho = -1.2$ would need $\Sigma d^2 = 363$, more than any two orders of 10 can give: a value below -1 shows an arithmetic mistake.

MARKING SCHEME

- ✓ $d = 2k - (n+1)$ and Σd^2 written with the two series.
- ✓ Simplified to $\frac{n(n^2-1)}{3}$, so $\rho = -1$.
- ✓ $\Sigma d^2 = 66$ from $\rho = 0.6$.
- ✓ Corrected $\Sigma d^2 = 106$, $\rho = 0.358$.
- ✓ No: Σd^2 can be at most 330, so $\rho \geq -1$.

5 From two lines to the spread

ANSWER A · 4

Tests: $b_{yx} = r\frac{\sigma_y}{\sigma_x}$ used with $r^2 = b_{yx}b_{xy}$ to reach a spread; the right line for the right estimate.

The means: $2x + 5y = 32$ and $5x + 8y = 62$ give $\bar{x} = 6$, $\bar{y} = 4$.

Taking $2x + 5y = 32$ as y on x: $b_{yx} = -0.4$; $5x + 8y = 62$ as x on y: $b_{xy} = -1.6$. Product $0.64 \leq 1$, allowed. (The other way gives $(-0.625)(-2.5) = 1.5625 > 1$, impossible.) So $r = -0.8$.

a. $\sigma_y = \frac{b_{yx}\sigma_x}{r} = \frac{-0.4 \times 3}{-0.8} = \mathbf{1.5}$.

b. To estimate x use x on y: $x = 6 - 1.6(6 - 4) = \mathbf{2.8}$. Pratima used the line of y on x, which makes the misses in y least and is the wrong line for estimating x; the two agree only when $r = \pm 1$.

MARKING SCHEME

- ✓ The means (6, 4) and which line is which, with the product test.
- ✓ $r = -0.8$.
- ✓ $\sigma_y = 1.5$.
- ✓ $x = 2.8$ from the line of x on y, with the reason.

6 A curve that hides from r

ANSWER

Tests: r measures only a straight-line relation; working with $\Sigma(x - \bar{x})(y - \bar{y}) = \Sigma xy - n\bar{x}\bar{y}$.

a. Six points: $\bar{x} = \frac{3}{6} = 0.5$, $\bar{y} = \frac{19}{6}$. $\Sigma xy = -8 - 1 + 0 + 1 + 8 + 27 = 27$, so

$\Sigma(x - \bar{x})(y - \bar{y}) = 27 - 6 \times 0.5 \times \frac{19}{6} = 17.5$. $\Sigma(x - \bar{x})^2 = 19 - 1.5 = 17.5$,

$\Sigma(y - \bar{y})^2 = 115 - \frac{361}{6} = 54.83$.

$$r = \frac{17.5}{\sqrt{17.5 \times 54.83}} = \mathbf{0.565}.$$

b. $r = 0$ exactly when $\Sigma xy - n\bar{x}\bar{y} = 0$. With the sixth point, $\Sigma xy = t^3$, $\bar{x} = \frac{t}{6}$, $\bar{y} = \frac{10 + t^2}{6}$:

Science Playground, Katunje, Bhaktapur
Book a class visit: https://science-playground.com/booking

$t^3 - 6 \cdot \frac{t}{6} \cdot \frac{10 + t^2}{6} = \frac{5t^3 - 10t}{6} = \frac{5t(t^2 - 2)}{6} = 0$, so $t = \mathbf{0}$ **or** $t = \pm\sqrt{\mathbf{2}}$. With $(\sqrt{2}, 2)$ every point is still on $y = x^2$, and r is still 0.

MARKING SCHEME

✓ $r = 0.565$ with $(3, 9)$.

✓ The covariance sum in terms of t: $\frac{5t(t^2 - 2)}{6}$.

✓ $t = 0, \pm\sqrt{2}$.

Marks by level	Knowledge	Understanding	Application	Higher ability	Total
Marks	0	0	10	17	**27**

Science Playground, Katunje, Bhaktapur
Book a class visit: https://science-playground.com/booking