Unit 10 · Correlation and Regression

STUDY NOTES

WHAT NEB'S GRADE 12 EXAM ASKS

- **CDC's grid (2078):** Statistics and Probability (Units 10–11 together): 6 of 75 marks; 1 MCQ and 1 short question; no long question
- **The paper:** 11 MCQs × 1 · 8 short × 5 · 3 long × 8; 3 hours
- **CDC's model paper:** from a five-row table, a regression estimate (3 marks) and the correlation coefficient (2 marks)
- **MCQ (1 mark)**
  - where r lies
  - Spearman's formula
  - what $r = 0$ means
  - r from b_{yx}, b_{xy}
  - the means from two lines
  - $r = -1$: the lines
  - change of scale
  - ρ from Σd^2
  - r from the three sums
  - impossible pair of coefficients
  - r from two lines
- **Short (5 marks, in parts)**
  - Pearson's r, interpreted (not cause)
  - both regression lines; an estimate
  - rank correlation of two subjects
  - $r^2 = b_{yx}b_{xy}$; means and r from two lines
  - scatter diagram, r, both lines, which to use

Key terms

Bivariate data: pairs (x, y), two numbers measured on each thing.

Scatter diagram: each pair a point, x across, y up.

Correlation: how closely two quantities go together.

Positive / negative / none: points rise / fall to the right / no tilt.

Perfect: all on one line; **linear**: round a line; **curvilinear**: round a curve (r doesn't measure it).

Regression line: the line that estimates one variable from the other.

Least squares: the line making the sum of the squared misses least.

Regression coefficient b_{yx}: the slope; the average change in y per unit of x.

Coefficient of determination r^2: the share of the spread in y the line accounts for.

Formulas

$$r = \frac{\Sigma (x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma (x-\bar{x})^2 \Sigma (y-\bar{y})^2}}$$ whole-number means **Use:** Pearson's r

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$ from the data's sums **Use:** Pearson's r

$$\rho = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$ d: difference of ranks **Use:** Spearman

$$y - \bar{y} = b_{yx}(x - \bar{x}) \qquad b_{yx} = \frac{\Sigma (x-\bar{x})(y-\bar{y})}{\Sigma (x-\bar{x})^2} = r\frac{\sigma_y}{\sigma_x}$$ **Use:** to estimate y

$$x - \bar{x} = b_{xy}(y - \bar{y}) \qquad b_{xy} = \frac{\Sigma (x-\bar{x})(y-\bar{y})}{\Sigma (y-\bar{y})^2} = r\frac{\sigma_x}{\sigma_y}$$ **Use:** to estimate x

$$b_{yx} = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$ from the data's sums **Use:** slope

$$b_{yx}b_{xy} = r^2$$ $r = \pm\sqrt{b_{yx}b_{xy}}$, their sign **Use:** both lines

- $\sigma = \sqrt{\Sigma (x-\bar{x})^2/n}$ (Grade 11): divide by n

Correlation and its nature

- draw the scatter diagram first: rising positive; falling negative; no tilt none
  - strength: how tightly the points hug a line

- examples
  - positive: a child's age and height; fertilizer and yield
  - negative: height and temperature; price and sales
- **correlation is not cause**: a third thing can drive both (ice cream and drowning: hot months)

Pearson's r

- **why it works**: cross at $(\bar{x}, \bar{y})$; top-right and bottom-left points give $(x-\bar{x})(y-\bar{y}) > 0$ ("together"), the other corners negative
- dividing by $\sigma_x \sigma_y$ removes the units: $r = \text{cov}(x, y)/(\sigma_x \sigma_y)$
- **short form**: $\Sigma (x-\bar{x})(y-\bar{y}) = \Sigma xy - \Sigma x \Sigma y/n$; $\Sigma (x-\bar{x})^2 = \Sigma x^2 - (\Sigma x)^2/n$
- whole-number means: deviations $X = x - \bar{x}$, $Y = y - \bar{y}$; a table of X, Y, XY, X^2, Y^2

Properties (stated without proof):

1. $-1 \le r \le 1$; ± 1 perfect; 0 no **linear** correlation
2. sign the direction, size the strength: $|r| \ge 0.75$ high, 0.5–0.75 moderate, below 0.5 low (cut-offs vary)
3. no units; $r_{xy} = r_{yx}$
4. **unchanged by a change of origin and scale**: $u = (x-a)/h$, $v = (y-b)/k$, $h, k > 0 \rightarrow r_{uv} = r_{xy}$ (opposite signs: only the sign flips)
5. r^2: share of the spread in y explained

- five stations, height (100 m) and temperature: $\Sigma XY = -109.8$, $\Sigma X^2 = 214$, $\Sigma Y^2 = 56.76 \rightarrow$ $r = -0.996$, $r^2 = 0.99$
- probable error $0.6745(1-r^2)/\sqrt{n}$: older books; not in CDC's 2078 curriculum or model paper

Rank correlation

- only an order (two judges, two subjects), or one far-out value would swamp r
- rank both variables **the same way** (1 for the smallest, or the largest, in both)
- ρ is Pearson's r on the ranks: $-1 \le \rho \le 1$
  - $\rho = 1$: orders agree (all $d = 0$); -1: one reversed
- tied values share the mean of their ranks
- check: $\Sigma d = 0$
- 15 stations, height and rain: $\Sigma d^2 = 732$, $\rho = 1 - 4392/3360 = -0.31$

Regression lines

- least squares, $y = a + bx$: normal equations $\Sigma y = na + b\Sigma x$, $\Sigma xy = a\Sigma x + b\Sigma x^2$
  - the first $\div n$: **the line passes through the means**
- **two lines**: y on x makes the up-and-down misses least; x on y the across misses
- properties
  1. both through $(\bar{x}, \bar{y})$: solve the two equations together for the means
  2. $b_{yx} b_{xy} = r^2$; same sign as r; can't both be more than 1 in size
  3. $r = \pm 1$: one line; $r = 0$: $y = \bar{y}$ and $x = \bar{x}$, at right angles
- two lines given: take one as y on x, the other as x on y; if $b_{yx} b_{xy} > 1$, swap them ($r^2 \le 1$)
- stations: $b_{yx} = -0.5131$, $y = 25.70 - 0.5131x$; 1,000 m $\rightarrow$ 20.57 °C
  - $b_{xy} = -1.9345$; 20 °C $\rightarrow$ about 1,113 m
- estimate only inside the range of the data (no station above 2,400 m: nothing about Everest)
- 15 stations: temperature $r = -0.995$ (height explains 99%); rain $r = -0.22$ (about 5%)

COMMON MISTAKES

✗ x from a given y by solving the y-on-x line
✓ use the x-on-y line (they differ unless $r = \pm 1$).

✗ $r = 0$: no relation
✓ no **straight-line** relation; there may be a curve.

✗ a strong r proves cause
✓ a third thing may drive both.

✗ r changes when heights go from m to km
✓ unchanged by origin and scale.

✗ $r = +\sqrt{b_{yx} b_{xy}}$ always
✓ the sign of the two coefficients.

✗ ρ with one variable ranked from the largest, the other from the smallest
✓ the same way for both.

✗ σ dividing by $n - 1$
✓ by n, as in Grade 11.

✗ estimates far outside the data
✓ only within its range.