5.6 Regression and correlation

Name: _____ Class: _____ Date: _____ **Total: 14 marks**

KEY WORDS correlation • least squares • residual

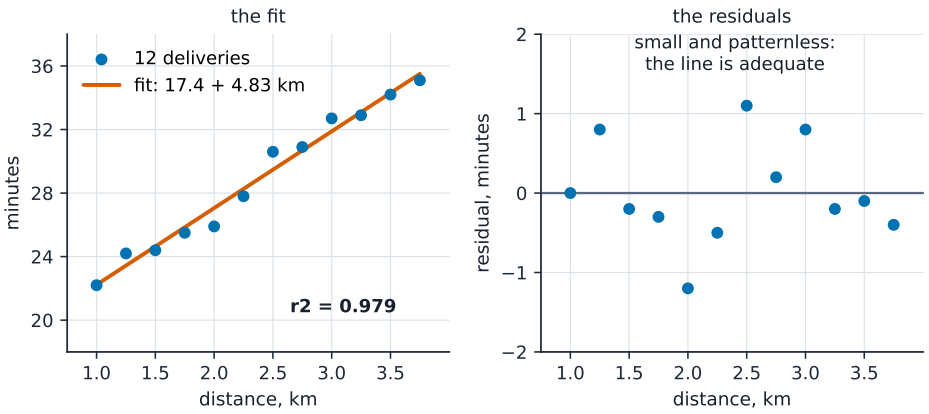
Objective

Fit and interpret a simple regression; read residual plots; keep correlational data from making causal claims.

1 Key knowledge and worked examples

■ **The line, the residuals, the police**

Least squares gives $\hat{y} = a + bx$; r^2 is the shared variance. The **residual** plot grades the fit – small and patternless, or the line is wrong somewhere.
Correlation is association; only design earns causation.



The fit is half the report - the residuals are the other half

■ **Worked example. The course fit**

$\hat{y} = 17.4 + 4.83\,\text{km}$, $r = 0.989$, $r^2 = 0.979$; residuals -1.2 to $+1.1$, patternless ✓. Each kilometre travels with 4.83 more minutes IN THIS SAMPLE – the causal rate needs routes held fixed.

2 Practice

A bakery tracks 10 batches: oven temperature setting (1–5) and crust score. Fitted: score $= 2.1 + 0.62 \times$ setting; $r = 0.71$; residuals show a gentle U-shape.

2.1 Interpret the slope and predict the score at setting 4. [3]

2.2 Compute r^2 and say what it does – and does not – mean. [3]

2.3 The U-shaped residuals: what do they tell you about the straight line, and what would you try next? [2]

2.4 "Each setting step CAUSES +0.62 score." Rewrite the sentence honestly, and name the study that could earn the causal version. [3]

2.5 True or false, one line each: (a) $r = 0.71$ means the fit is 71% accurate; (b) the intercept 2.1 is the score at setting zero, a setting that does not exist; (c) predicting at setting 9 is extrapolation. [3]

exercises.lol exercises.com.cn

Solutions

2.1

- Each one-step setting increase travels with +0.62 score on average in this sample.
- At 4: $2.1 + 0.62 \times 4 = 2.1 + 2.48$.
- $=$ **4.6** (4.58).

2.2

- $r^2 = 0.71^2 =$ **0.50**.
- It means half the score variance moves with the setting in this sample.
- It does NOT mean 50%-accurate predictions or any causal share – variance shared, nothing more.

2.3

- The U-shape says the true relation curves; a straight line systematically misses at the ends.
- Try a quadratic term (or bin the settings) – and say so before quoting any slope.

2.4

- Honest: "setting and score move together ($r^2 = 0.50$); higher settings travel with higher scores in these 10 batches."
- A causal rate needs temperature randomly assigned across batches (holding recipe and time fixed).
- Then the slope becomes an effect of the intervention, not an association.

2.5

- (a) False – r is strength and direction of association; accuracy claims need prediction errors.
- (b) True – and that is exactly why the intercept is not interpreted here.
- (c) True – no data beyond setting 5 vouches for the line.

exercises.lol exercises.com.cn