1 Mark for Review

Jacques, an artisan cheese maker, collects data on every step of the cheese-making process for each batch he makes. Jacques noticed that the daily high temperature in his shop on the day he made a batch of cheese was related to the pH of the cheese the next morning. He computed the correlation between the daily high temperature and the pH of the cheese to be -0.64. What information does the correlation provide about the relationship between the daily high temperature and the pH of the cheese?

(A) The relationship is linear because the correlation is negative. (A)

(B) The relationship is not linear because the correlation is negative. (B)

(C) The morning pH of the cheese tends to be higher when the daily high temperature in the shop is warmer, compared to when the daily high temperature is cooler. (C)

(D) The morning pH of the cheese tends to be higher when the daily high temperature in the shop is cooler, compared to when the daily high temperature is warmer. (D)

(E) There is no relationship between the daily high temperature and the pH of the cheese. (E)

D

Correlation

Correlation	相关
Positive correlation	正相关
Negative correlation	负相关
Strength	强度
Direction	方向

- **Correlation (r)** measures the **strength and direction of a linear relationship** between two quantitative variables.
- The value of **r** ranges from **–1 to +1**.
- Correlation is **unitless** and does not change with linear transformations.
- **Outliers** can strongly affect the value of r.
- **Correlation does not imply causation** and only describes linear relationships.

A study collects data on **hours of sleep per night** and **students' reaction times** in a cognitive test.

(a) Describe the **direction** and **strength** of the relationship if the correlation coefficient is $r = -0.75$.

(b) Explain what a correlation of $r = 0$ would indicate about the relationship.

(c) Identify one limitation of using correlation to describe the relationship between two variables.

(a)

- **Direction:** Negative correlation (as hours of sleep increase, reaction time decreases)
- **Strength:** Strong (because $|r| = 0.75$ is close to 1)

(b) A correlation of $r = 0$ indicates **no linear relationship** between hours of sleep and reaction time.

(c) Correlation **does not imply causation**; it only measures the **linear relationship** between variables and can be affected by **outliers**.

Stretch and Challenge

6. Multiple Variables

Two variables x and y have $r = 0.5$. Another variable z is added, which is strongly related to x but not y.

Question:

Discuss how the presence of z could affect interpretation of the original correlation between x and y.

---

7. Residual Analysis

A least-squares regression line is fitted, but residuals show a clear curve.

Question:

What does this indicate about the correlation and linear model? What should you do next?

---

8. High Leverage Points

A single x-value far outside the rest of the data has y close to the regression line.

Question:

Explain how this point affects correlation and slope. Is it influential?

---

6. Multiple Variables

- z can **confound** x-y correlation.
- Must consider **partial correlation** or **context**.

---

7. Residual Analysis

- Curved residuals → **linear model inappropriate**
- Correlation underestimates true association → consider **nonlinear regression**.

---

8. High Leverage Points

- High-leverage x near line → little effect on slope/correlation
- Not necessarily **influential**, but must check if extreme y is present.

1. Regression Basics

- **Regression Line (回归线)** – line that models the relationship between x and y
- **Least-Squares Regression Line (最小二乘回归线)** – line minimizing the sum of squared residuals
- **Slope (斜率)** – change in y for a one-unit increase in x
- **Intercept (截距)** – predicted y when x = 0
- **Predicted Value ($\hat{y}$) (预测值)** – y-value estimated from the regression line

---

2. Correlation & Fit

- **Correlation Coefficient (r) (相关系数)** – measures direction and strength of linear relationship
- **Coefficient of Determination (r^2) (决定系数)** – proportion of variation in y explained by x

---

3. Residuals & Errors

- **Residual (残差)** – observed y minus predicted $\hat{y}$
- **Residual Plot (残差图)** – used to check linearity, constant variance, and outliers
- **Outlier (离群值)** – point far from the trend
- **Influential Point (有影响点)** – point that significantly affects slope or intercept

1. Purpose

- Models the **linear relationship** between an explanatory variable x and a response variable y.
- Used for **prediction, interpretation, and understanding associations**.

---

2. Equation

$$\hat{y} = a + bx$$

- b = **slope** → change in y per unit change in x
- a = **intercept** → predicted y when x = 0

---

3. Least-Squares Criterion

- Regression line **minimizes the sum of squared residuals**:

$$\sum (y_i - \hat{y}_i)^2$$

---

4. Residuals

- **Residual** = observed y – predicted y
- Residuals help check:
  - **Linearity** (should be random)
  - **Constant variability**
  - **Outliers or influential points**

A company wants to predict monthly sales (y, in \$1000s) based on advertising expenditure (x, in \$100s). A sample of 10 months produced the following regression output:

$$\hat{y} = 5 + 2x$$

- $r = 0.85$
- $r^2 = 0.7225$

Tasks:

1. Interpret the **slope** and **intercept** in context.
2. Predict sales if the company spends \$500 on advertising.
3. Explain what $r^2 = 0.7225$ means in context.
4. If a residual for one month is −3, explain what this indicates.
5. Discuss any caution you would have in predicting sales for \$2000 in advertising.

1. Interpret Slope and Intercept

- **Slope** $b = 2$ → For each additional \$100 spent on advertising, predicted monthly sales increase by \$2000.
- **Intercept** $a = 5$ → Predicted sales are \$5000 when advertising expenditure is \$0.

---

2. Prediction

$$x = 500 \text{ dollars } = 5 \text{ hundreds}$$

$$\hat{y} = 5 + 2(5) = 15$$

Predicted sales = \$15,000.

---

3. Interpret r^2

$r^2 = 0.7225$ → **72.25% of the variation in monthly sales** is explained by advertising expenditure.

---

4. Residual Interpretation

Residual = observed − predicted = −3 → **sales were \$3000 lower than predicted** for that month.

---

5. Caution for Extrapolation

- Predicting for \$2000 (x = 20 hundreds) is **outside the observed data range**.
- Relationship may **not remain linear**, so predicti may be unreliable.

TRUE or FALSE

- Extrapolation beyond the observed data is generally less reliable than interpolation.

True

TRUE or FALSE

- If all residuals are positive, the regression line perfectly fits the data.

FALSE

TRUE or FALSE

- The regression line minimizes the sum of the absolute values of residuals.

FALSE – minimizes the sum of the residuals

Regression line equation

$$\hat{y} = a + bx$$

Where:

- $b = \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$ → slope
- $a = \bar{y} - b\bar{x}$ → intercept

1. **Objective:** Minimize the sum of squared residuals

$$S = \sum (y_i - (a + bx_i))^2$$

2. Take partial derivatives with respect to a and b:

$$\frac{\partial S}{\partial a} = 0, \quad \frac{\partial S}{\partial b} = 0$$

3. Solve these equations → yields:

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}, \quad a = \bar{y} - b\bar{x}$$