A school is considering modifying a policy to allow only grade 12 students to leave campus during their lunchtime. Currently, no students are allowed to leave campus at lunchtime. The 120 Grade 11 students and the 150 Grade 12 students were surveyed about their opinions on the proposed policy. The marginal frequencies are shown in the table below.

	Favor	Oppose	Total
Grade 11	a	b	120
Grade 12	c	d	150
Total	144	126	270

Which of the following would indicate that there is *no association* between grade level and opinion about the policy?

(A) $a = b$ ~~(A)~~

(B) $a = c$ ~~(B)~~

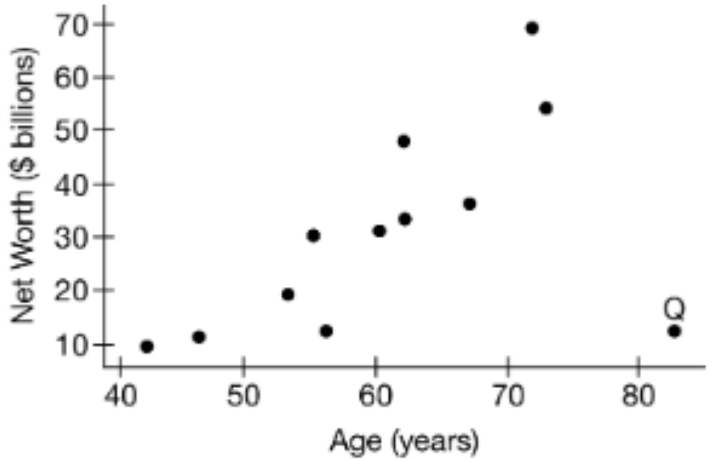
(C) $a = 0$ ~~(C)~~

(D) $a = 64$ ~~(D)~~

(E) There is not enough information to answer this question. ~~(E)~~

D

The following scatterplot shows the ages, in years, of 12 of the wealthiest people in the world along with their net worth, in billions of dollars. The data point at age 83 is labeled Q.



Suppose point Q is removed from the data set. Which of the following is likely not affected by the removal?

(A) The correlation coefficient

(B) The sign of the slope coefficient ✔

(C) The value of the slope coefficient

(D) The sum of the squared residuals

(E) The net worth intercept

B

Each accountant at a large accounting firm was classified according to accountant level (junior or senior) and method of transportation to work (walk, bus, drive alone, or carpool). The responses of the 320 accountants at the firm are summarized in the table.

	Junior	Senior	Total
Walk	25	3	28
Bus	87	12	99
Drive alone	96	25	121
Carpool	52	20	72
Total	260	60	320

What proportion of the accountants at the firm are at the senior level and carpool to work?

(A) $\frac{20}{60}$ (A)

(B) $\frac{20}{72}$ (B)

(C) $\frac{20}{320}$ ✔ (C)

(D) $\frac{112}{320}$ (D)

(E) $\frac{132}{320}$ (E)

C

Analysing Departures from Linearity

- Linearity — 线性
- Non-linearity — 非线性
- Scatter plot — 散点图
- Residual — 残差
- Residual plot — 残差图
- Trend — 趋势
- Curvature — 曲率 / 弯曲
- Correlation — 相关性
- Outlier — 异常值 / 离群点
- Best-fit line — 最佳拟合直线

- Use a **residual plot** to assess linearity.
  - **Patterns** (curves, trends) in residuals indicate nonlinearity.
- Residuals should be **randomly scattered around zero** for a good linear fit.
  - Increasing or decreasing spread suggests **non-constant variability**.
  - Strong outliers can distort the regression model and should be noted.

A researcher investigates the relationship between hours studied (x) and exam score (y) for a group of students. A scatterplot suggests a positive association.

A linear regression model is fitted, giving:

$$\hat{y} = 45 + 5.2x$$

The residual plot shows a clear curved pattern.

(a)

Interpret the slope of the regression line in context.

(b)

Explain what a **curved pattern in the residual plot** suggests about the appropriateness of the linear model.

(c)

The researcher considers fitting a quadratic model instead. Explain why this may be more appropriate.

(d)

One student is identified as an outlier with a large negative residual. Explain what this means in context.

(e)

State one limitation of using this regression model to predict exam scores for students who study 0 hours or 20 hours.

(a)

Each extra hour studied increases predicted score by **5.2 marks**.

(b)

A curved residual plot shows the **linear model is not appropriate**; relationship is **non-linear**.

(c)

Quadratic model may fit because it can **capture curvature** in the relationship.

(d)

The student scored **much lower than predicted** by the model.

(e)

Model may be unreliable outside observed data range / **extrapolation not valid**.

1. Model Choice and Justification

A dataset shows a strong association between temperature (x) and ice cream sales (y), but the scatterplot is curved upward.

A linear model is fitted first, then a logarithmic model is considered.

(a) Explain why a linear model may be inappropriate.
(b) Describe how you would decide between the linear and logarithmic model using residuals.
(c) Explain what pattern in a residual plot would indicate a good model fit.

---

2. Transformation Reasoning

A researcher models the relationship between time (x) and bacterial growth (y). The scatterplot shows exponential growth.

(a) Suggest a transformation of y that may linearise the data.
(b) Explain why this transformation works.
(c) State one limitation of interpreting the transformed regression line.

---

3. Influence and Outliers

A regression analysis shows one point with a very high leverage and large residual.

(a) Explain the difference between **leverage** and **outlier**.
(b) Describe the effect of this point on the regression line.
(c) State whether this point should always be removed. Justify.

4. Extrapolation and Model Validity

A linear model is used to predict house prices based on distance from city centre. Data is collected only between 1 km and 10 km.

(a) Explain why predicting at 25 km is unreliable.
(b) Describe one assumption of linear regression that may be violated outside the data range.
(c) Suggest how the model could be improved.

---

5. Comparing Models

Two models are fitted:

- Model A: linear regression
- Model B: quadratic regression

Model B has a slightly higher R^2.

(a) Explain why a higher R^2 does not automatically mean Model B is better.
(b) State one method (other than R^2) to compare models.
(c) Describe what residual plots you would expect for the better model.

1. Model Choice

(a) Relationship is curved → linear model underestimates/overestimates.

(b) Compare residual plots: smaller, random scatter = better model.

(c) Random scatter around 0, no pattern.

---

2. Transformation

(a) $\ln(y)$ or log transformation.

(b) Linearises exponential growth.

(c) Interpretation applies to transformed scale, not original units.

---

3. Influence & Outliers

(a) Leverage = extreme x; outlier = extreme y (large residual).

(b) Can strongly pull regression line.

(c) Not always; depends if it's valid data or error.

4. Extrapolation

(a) Outside data range → model may not hold.
(b) Linearity assumption may fail.
(c) Collect more data / use non-linear model.

---

5. Comparing Models

(a) Higher R^2 may overfit.
(b) Residual plots / adjusted R^2.
(c) Better model: random residual scatter, no pattern.

Inference for linear regression (SLOPE)

- Slope (β_1) — 斜率（β_1）
- Regression line — 回归直线
- Null hypothesis (H_0) — 原假设
- Alternative hypothesis (H_1) — 备择假设
- t-test — t 检验
- t-statistic — t 统计量
- p-value — p 值
- Standard error (SE) — 标准误
- Confidence interval — 置信区间
- Degrees of freedom (n−2) — 自由度（n−2）

Inference for Linear Regression (Slope Test) — Key Points

- We test whether there is a **linear relationship between two quantitative variables**.
- The focus is the **population slope (β_1)**, not the sample slope (b_1).
- **Null hypothesis:** $H_0 : \beta_1 = 0$ (no linear relationship).
- **Alternative hypothesis:** $H_a : \beta_1 \neq 0$ (or > 0, < 0).
- Use a **t-test for slope** with:

$$t = \frac{b_1 - 0}{SE(b_1)}$$

- Degrees of freedom: $n - 2$.
- A **small p-value** means strong evidence of a **linear relationship**.
- Conditions: **linear relationship, independent observations, normal residuals, constant variance.**
- A **confidence interval for β_1** gives a plausible range for the true slope.
- If 0 is in the CI → **not statistically significant**.

A researcher studies the relationship between hours of sleep per night (x) and reaction time in milliseconds (y) for a group of students.

A linear regression output is given:

$$\hat{y} = 520 - 18.5x$$

Standard error of slope: $SE = 4.2$
Sample size: $n = 28$

(a)

Interpret the slope in context.

(b)

State appropriate hypotheses for testing whether sleep is related to reaction time.

(c)

Calculate the test statistic for the slope.

(d)

The p-value is 0.0003. Interpret this in context.

(e)

At $\alpha = 0.05$, state and justify the conclusion.

(f)

State one condition that must be met for this inference to be valid.

(a)

Each extra hour of sleep decreases predicted reaction time by **18.5 ms**.

(b)

$H_0 : \beta_1 = 0$ (no relationship)
$H_a : \beta_1 \neq 0$ (relationship exists)

(c)

$$t = \frac{-18.5}{4.2} \approx -4.40$$

(d)

If the true slope is 0, the probability of getting a result this extreme is **0.0003**.

(e)

Reject H_0; there is **strong evidence of a linear relationship** between sleep and reaction time.

(f)

One valid condition: **linear relationship between variables** (or normal residuals / independent observations / constant variance).

Fill in the Gaps: Degrees of Freedom in Linear Regression

In simple linear regression, the degrees of freedom are _____ because we estimate _____ parameters from the data.

These parameters are the _____ and the _____ of the regression line.

Each estimated parameter uses up _____ degree of freedom.

Therefore:

$$\mathrm{df} = n - \_\_\_\_$$

After fitting the regression line, the remaining degrees of freedom are used to measure the _____ variation.

To perform inference on the slope, we use a _____ distribution with _____ degrees of freedom.

We subtract 2 because a line is defined by at least _____ parameters.

In simple linear regression, the degrees of freedom are $n - 2$ because we estimate **2** parameters from the data.

These parameters are the **intercept** and the **slope** of the regression line.

Each estimated parameter uses up **one** degree of freedom.

Therefore:

$$\mathrm{df} = n - 2$$

After fitting the regression line, the remaining degrees of freedom are used to measure the **residual** variation.

To perform inference on the slope, we use a **t** distribution with $n - 2$ degrees of freedom.

We subtract 2 because a line is defined by at least **two** parameters.

TRUE or FALSE

- A curved pattern in a residual plot suggests a linear model is appropriate.

TRUE

TRUE or FALSE

- If residuals are randomly scattered around zero, a linear model is reasonable.

TRUE

TRUE or FALSE

- Transformations (e.g., log or square root) can help correct non-linearity.

TRUE