Name: _____ Date: _____

UNIT 4

SL 4.4 — Correlation & Regression

Scatter plots • Pearson's correlation (r) • Spearman's rank (r_s) • Regression line • Interpolation vs Extrapolation

A. KEY IDEAS

Write short, clear answers. Use context words (units, variables, what increases/decreases).

1. **What does Pearson's correlation coefficient r measure? (Mention the type of relationship.)**

   _____

2. **State the possible range of r and what the values r = 1, r = -1, and r = 0 mean.**

   _____

   _____

3. **Explain why a strong correlation does *not* automatically imply causation. (One sentence.)**

   _____

B. PEARSON'S PRODUCT-MOMENT CORRELATION (PPMC)

Use technology to calculate **r**. Then interpret the strength and direction in context.

Math (x)	**60**	**45**	**70**	**36**	**91**	**75**	**68**
Reading (y)	57	78	76	42	61	81	29

1. **Using technology, calculate Pearson's correlation coefficient r for the data above.**

   Formula (for reference): $r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$

   r = _____

2. **Interpret your value of r in context (direction + strength).**

   _____

3. **Sketch a scatter plot (rough is fine) and describe whether a straight-line model is reasonable.**

   _____

C. SPEARMAN'S RANK CORRELATION (SRCC)

SRCC measures how well a **monotonic** relationship fits the data. It does not need to be linear.

Rank of Math (x)	**5**	**6**	**3**	**7**	**1**	**2**	**4**
Rank of Reading (y)	5	2	3	6	4	1	7

1. **For each student, compute d = x-rank - y-rank and then compute ∑d².**

   SRCC formula: $r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$

   ∑d² = _____

2. **Calculate Spearman's rank correlation coefficient r_s.**

   r_s = _____

3. **Compare your r (from Part B) and r_s. What does the difference suggest about the pattern in the data?**

   _____

D. LINEAR REGRESSION

Use technology to find the regression line of **y** on **x** and use it for prediction.

Hours studying (x)	**1**	**3**	**5**	**11**	**15**	**20**	**27**
Exam score (y)	15	28	39	47	59	64	83

1. Find the regression line of y on x in the form $y = ax + b$ **.**

a = _____ **b =** _____

2. Interpret a (gradient) and b (intercept) in context.

_____

3. Predict the exam score for someone who studies 10 hours using $\hat{y} = ax + b$ **.**

_____

4. Is this interpolation or extrapolation? Explain briefly.

_____

E. EXTENSION (EXAM-STYLE)

Write full sentences where asked. Justify your reasoning.

1. **A dataset has r = 0.92, but one point is a clear outlier far from the rest of the data. Explain two reasons why you should be cautious about using the regression line to predict values.**

   _____

   _____

2. **Two students claim:**

   Student A: "Because r is close to 1, the relationship must be causal."
   Student B: "Because r is close to 0, there is no relationship at all."

   For each claim, state whether it is correct and give a one-sentence explanation.

   A: _____

   B: _____

3. **The "hours studying vs exam score" model is used to predict a score for someone who studies 40 hours. (a) Is this interpolation or extrapolation? (b) Why is it riskier?**

   (a) _____

   (b) _____

4. **Residuals are defined as actual - predicted. If a residual plot shows a curved pattern, what does that suggest about using a linear regression model?**

   _____

ANSWER KEY

Name: _____ Date: _____

UNIT 4

SL 4.4 — Correlation & Regression — Answers

Scatter plots • Pearson's correlation (r) • Spearman's rank (r_s) • Regression line • Interpolation vs Extrapolation

A. KEY IDEAS (ANSWERS)

1. Pearson's **r** measures the **strength** and **direction** of a **linear** relationship between two quantitative variables.
2. Range: $-1 \le r \le 1$. $r = 1$ means perfect positive linear correlation, $r = -1$ means perfect negative linear correlation, and $r = 0$ means no linear correlation.
3. Correlation describes association, but confounders, coincidence, or reverse causation can also explain the pattern.

B. PEARSON'S PRODUCT-MOMENT CORRELATION (ANSWERS)

1. **$r \approx 0.186$**
2. This is a **weak positive** correlation: higher math scores tend to go with slightly higher reading scores, but the relationship is not strong.
3. The scatter would look fairly spread out, so a straight-line model is only mildly reasonable and predictions would have large uncertainty.

C. SPEARMAN'S RANK CORRELATION (ANSWERS)

1. **$\sum d^2 = 24$**
2. **$r_s \approx 0.357$** (with n = 7)
3. **r_s** is larger than **r**, suggesting the relationship may be more monotonic than strictly linear.

D. LINEAR REGRESSION (ANSWERS)

1. Regression line: $y \approx 2.353x + 20.295$ so **a ≈ 2.353**, **b ≈ 20.295**.
2. **a**: Each extra hour of studying increases the predicted exam score by about **2.35 points** on average. **b**: At 0 hours, the model predicts about **20.3 points**.
3. For x = 10: $\hat{y} = 2.353(10) + 20.295 \approx 43.824$ so the prediction is **≈ 43.8**.
4. This is **interpolation** because 10 is within the observed x-range (1 to 27).

E. EXTENSION — ANSWERS

1. Outliers can strongly change **r** and the regression line, and predictions may be driven by the unusual point rather than the overall trend.
2. Student A: **incorrect** (correlation ≠ causation). Student B: **incorrect** (r ≈ 0 means no *linear* relationship; a non-linear relationship may still exist).
3. (a) **Extrapolation**. (b) Riskier because 40 is outside the data range, so the linear trend may not continue and uncertainty increases.
4. A curved residual pattern suggests the relationship is not linear, so a different model/transform may be more appropriate.