Linear Regression & Influential Points

AP Statistics Worksheet · Grade 11–12

Name: _____ Date: _____

Learning Objectives

- Interpret the slope, y-intercept, and correlation coefficient of a least-squares regression line
- Identify and explain the influence of unusual points (outliers, high-leverage points) on a regression model
- Compare two regression models to determine how removing a data point changes the LSRL

Problems

1. A least-squares regression line for predicting test score (y) from hours studied (x) is given below. Identify the slope and the y-intercept, and explain what each means in context.

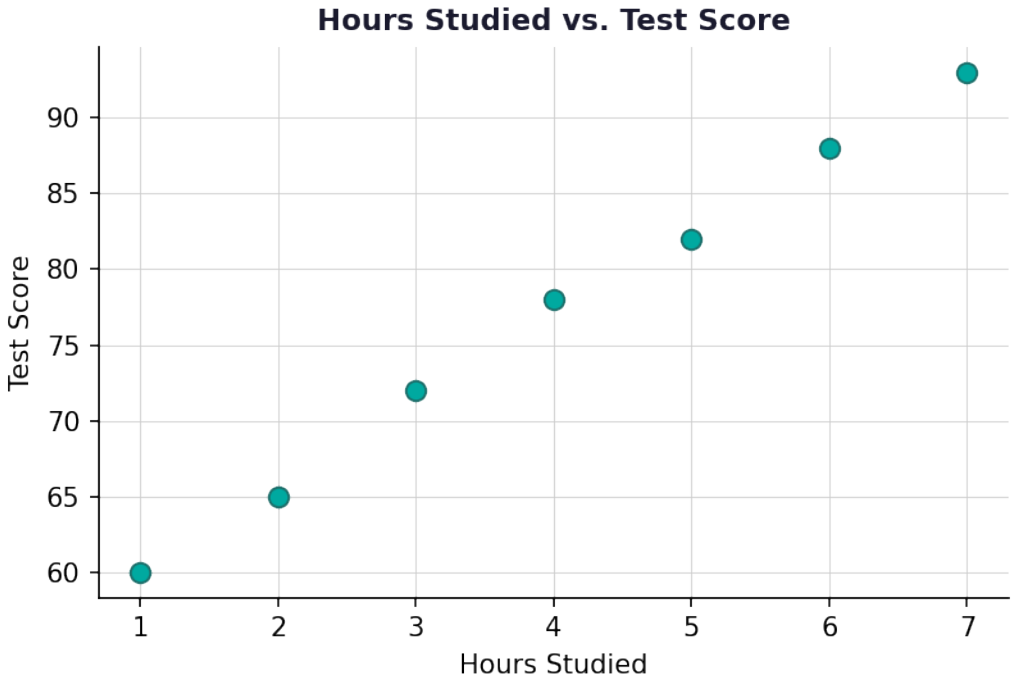
$$\hat{y} = 52.4 + 6.8x$$

2. A regression analysis reports R-squared = 0.476. Calculate the correlation coefficient r and describe the strength and direction of the linear relationship. Assume the slope is positive.

$$r = \sqrt{R^2}$$

3. Use the scatter plot of hours studied versus test score to describe the association between the two variables (form, direction, and strength).

Scan to watch



4. Two regression models are fit to the same data set, one with point P included and one with point P removed. Compare the two models using the information in the table and describe how much influence point P has.

Model	y-intercept (a)	Slope (b)	R^2
With P	8.107	0.4919	0.476
Without P	11.123	0.150	0.025

5. Using the regression outputs with and without point P (shown below), write the two least-squares regression equations in y-hat notation.

Model	Constant	Slope
With P	8.107	0.4919
Without P	11.123	0.150

6. A data set has R-squared = 0.025 after an influential point is removed. Calculate r and interpret what this value tells you about the linear relationship remaining in the data.

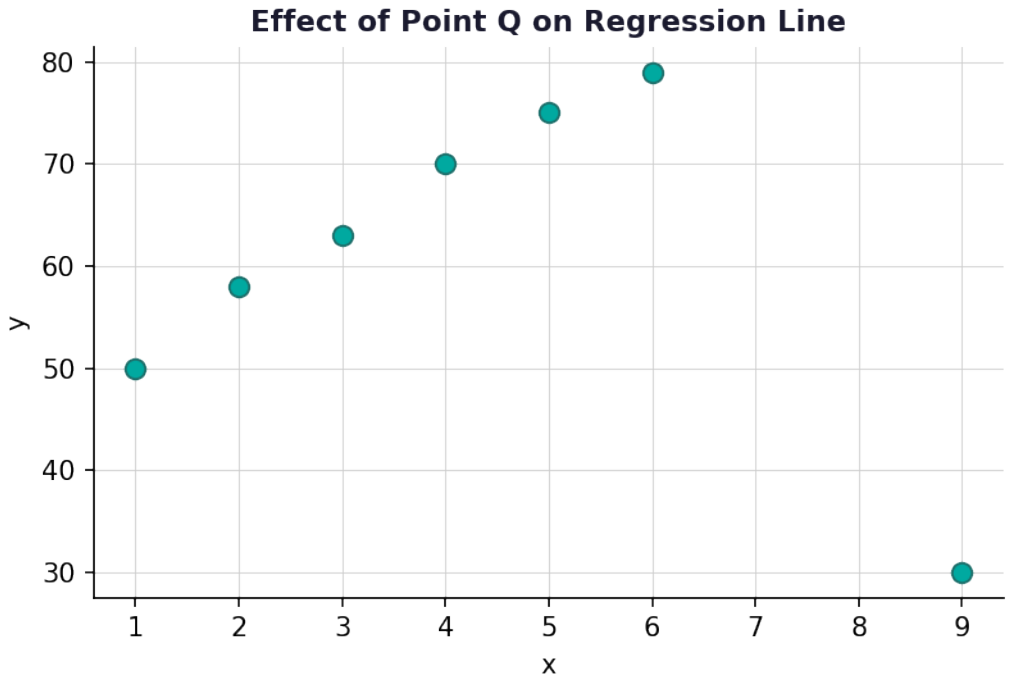
Scan to watch

$$r = \sqrt{0.025}$$

7. A student recorded the following data for hours of exercise (x) and resting heart rate (y). Plot the data on the scatter plot, then sketch the approximate least-squares regression line and describe its direction.

Hours of Exercise (x)	Resting Heart Rate (y)
1	82
2	78
3	74
4	70
5	66
6	63

8. A regression line is computed with and without a suspected influential point Q. The original regression (with Q) has a positive slope, but after Q is moved to the lower-right region of the scatter plot, the regression line changes direction. Explain why this happens and what it tells you about point Q's influence.



Scan to watch

9. Two regression models are compared. With point P: y-intercept = 8.107, slope = 0.4919, and $r \approx 0.689$. Without point P: y-intercept = 11.123, slope = 0.150, and $r \approx 0.158$. Write a complete justification (3–4 sentences) explaining whether point P is an influential point, referencing slope, y-intercept, and correlation coefficient.

10. A regression study on advertising spending (x, in thousands of dollars) and weekly sales (y, in thousands of dollars) produces the output below. (a) Write the LSRL equation. (b) Predict weekly sales when advertising spending is 8 thousand dollars. (c) Compute r from R-squared. (d) The data point (12, 14) is suspected to be influential. If this point were removed and the slope decreased from 3.2 to 1.1 while R-squared dropped from 0.81 to 0.09, would you classify it as influential? Justify fully.

Parameter	Estimate
Constant (a)	5.6
Slope (b)	3.2
R^2	0.81

Scan to watch

Linear Regression & Influential Points — Answer Key

AP Statistics Worksheet · Grade 11–12

Answer Key

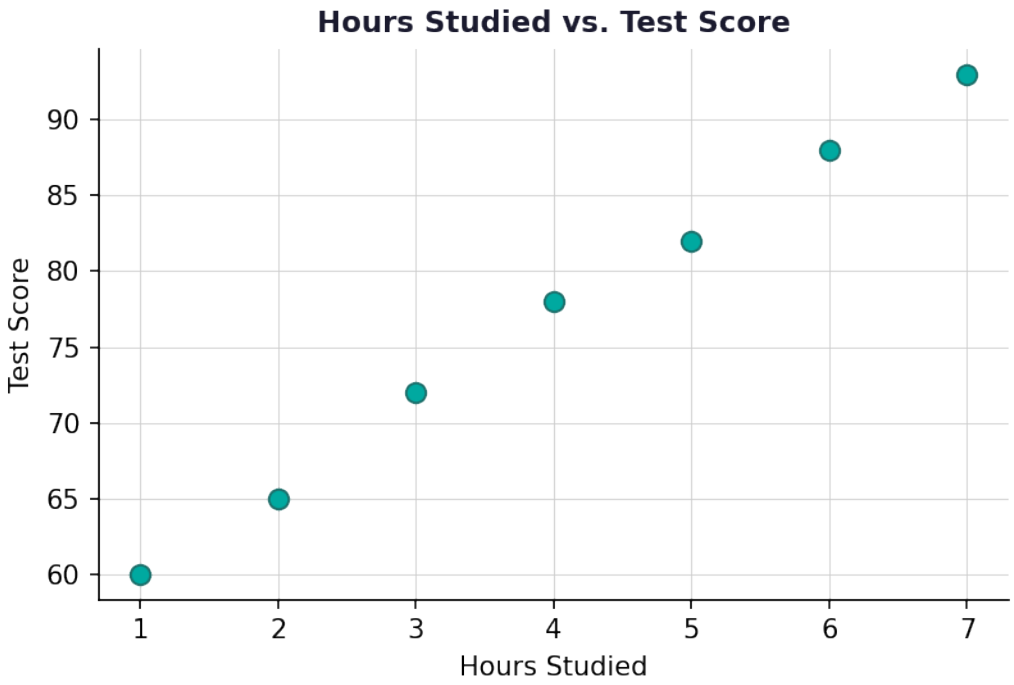
1. Answer: Slope = 6.8 (score increases ~6.8 pts per extra hour); y-intercept = 52.4 (predicted score with 0 hours studied)

- The y-intercept is 52.4 — the predicted test score when x = 0 hours are studied.
- The slope is 6.8 — for each additional hour studied, the predicted test score increases by 6.8 points.

2. Answer: r ≈ 0.689; moderate positive linear relationship

- r = √(0.476) ≈ 0.6899, which rounds to 0.689 (or about 69%).
- Since the slope is positive and r ≈ 0.689, there is a moderate positive linear relationship between the two variables.

3. Answer: Linear, positive, moderately strong association



- Form: The points follow a roughly linear pattern.
- Direction: As hours studied increase, test scores increase — positive association.
- Strength: The points are fairly close to a line, indicating a moderately strong relationship.

4. Answer: Point P has a large influence — removing it changes the slope from 0.49 to 0.15 and R^2 from 0.476 to 0.025

Scan to watch

- With P: ■ = 8.107 + 0.4919x, R^2 = 0.476, r ≈ 0.689.
- Without P: ■ = 11.123 + 0.150x, R^2 = 0.025, r ≈ 0.158.
- The slope decreased from 0.49 to 0.15 and R^2 dropped dramatically, indicating point P had a large influence on the regression line.

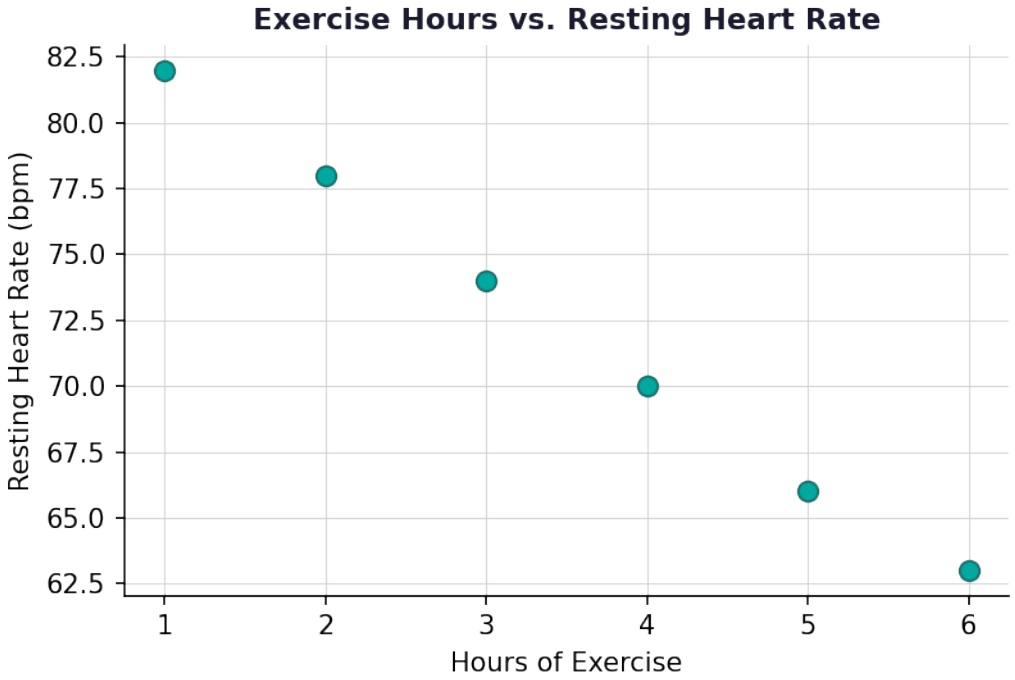
5. Answer: With P: ■ = 8.107 + 0.4919x; Without P: ■ = 11.123 + 0.150x

- With P: ■ = 8.107 + 0.4919x (use constant as a and slope as b in ■ = a + bx).
- Without P: ■ = 11.123 + 0.150x.
- Both equations follow the form ■ = a + bx, where a is the y-intercept and b is the slope.

6. Answer: r ≈ 0.158; very weak positive linear relationship

- r = √(0.025) ≈ 0.1581.
- Since r ≈ 0.158, this is a very weak positive linear relationship.
- Removing the influential point caused most of the linear association to disappear.

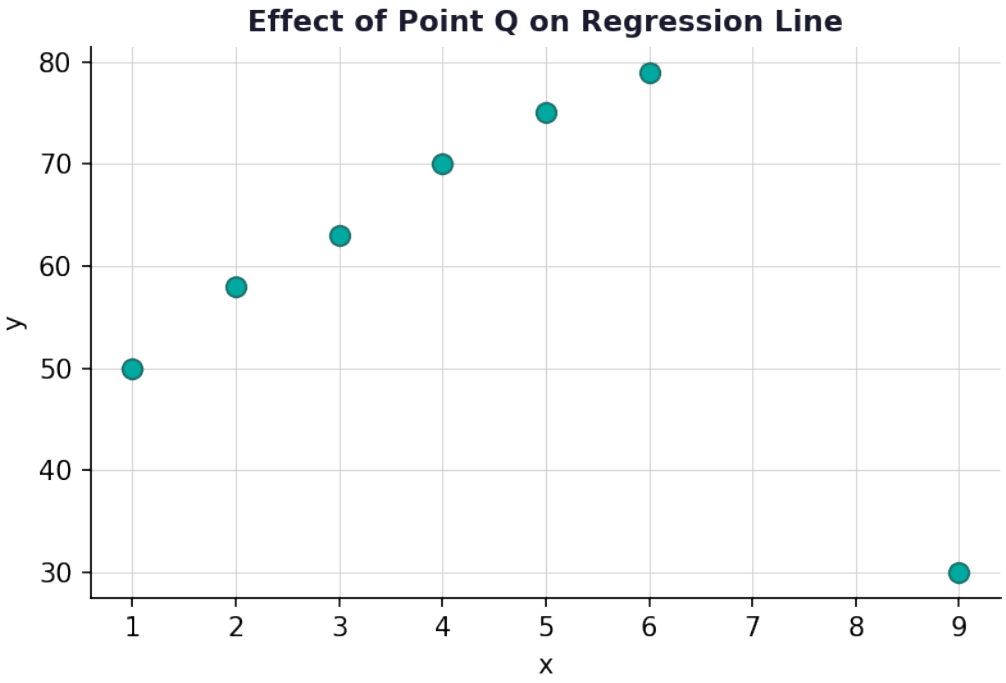
7. Answer: Negative linear association; as exercise increases, resting heart rate decreases



- Plot each (x, y) pair on the scatter plot.
- The points trend downward from left to right, indicating a negative association.
- The LSRL would have a negative slope, meaning more exercise is associated with a lower resting heart rate.

8. Answer: Point Q is highly influential — it pulls the regression line toward a negative slope, reversing the direction of the association

Scan to watch



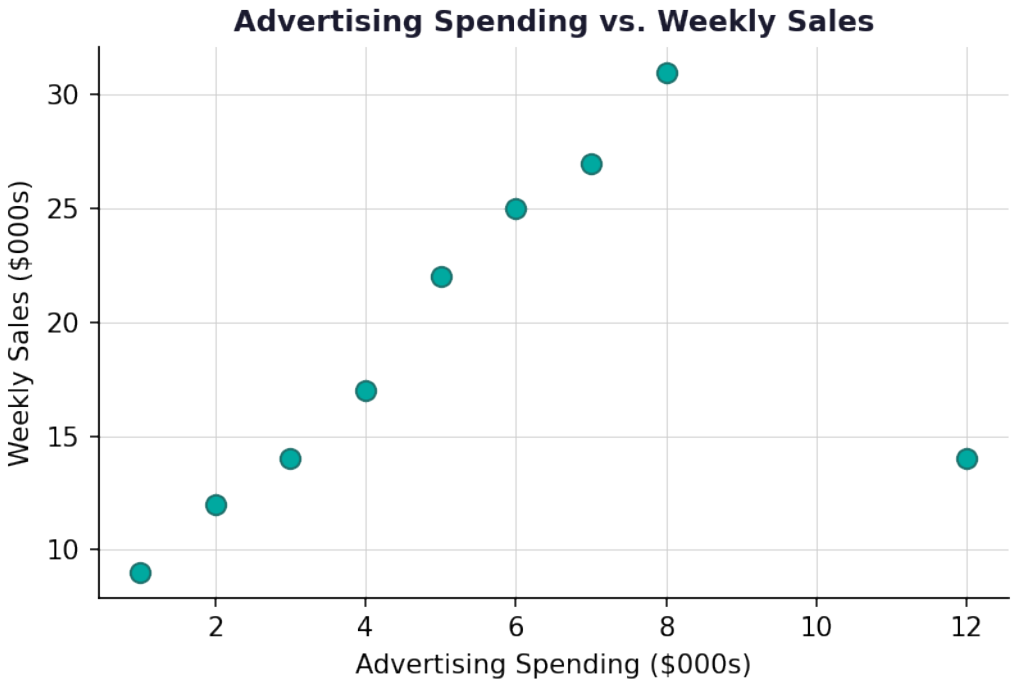
• Without Q, the remaining points show a positive linear trend (slope > 0).
• When Q is placed in the lower-right corner (high x, low y), it exerts a strong downward pull on the right side of the line.
• This can reverse the slope from positive to negative, demonstrating that Q is a high-leverage, influential point.
• An influential point is one whose removal causes a substantial change in the regression line.

9. Answer: Yes — point P is influential because removing it substantially changes the slope (0.49 to 0.15), raises the y-intercept (8.1 to 11.1), and weakens the correlation dramatically (r: 0.689 to 0.158)

• Step 1 — Compare slopes: The slope dropped from 0.4919 to 0.150, a large change indicating P affects the steepness of the line.
• Step 2 — Compare y-intercepts: The y-intercept increased from 8.107 to 11.123 when P was removed.
• Step 3 — Compare correlation: r went from 0.689 (moderate linear association) to 0.158 (very weak), meaning the linear relationship nearly disappears without P.
• Step 4 — Conclusion: Because removing P changes the slope, y-intercept, AND correlation coefficient substantially, point P is classified as an influential point.

10. Answer: (a) ■ = 5.6 + 3.2x; (b) ■ = 31.2 thousand dollars; (c) r = 0.90; (d) Yes, highly influential

Scan to watch



• (a) LSRL: ■ = 5.6 + 3.2x, where x = advertising spending and ■ = predicted weekly sales (both in $000s).
• (b) Predicted sales at x = 8: ■ = 5.6 + 3.2(8) = 5.6 + 25.6 = 31.2 thousand dollars.
• (c) r = √(0.81) = 0.90 (positive because slope > 0); strong positive linear relationship.
• (d) Removing (12, 14) changes the slope from 3.2 to 1.1 (a decrease of 2.1) and drops R^2 from 0.81 to 0.09, meaning r falls from 0.90 to 0.30.
• These are large changes in both the slope and the strength of the association, so (12, 14) is classified as a highly influential point — it substantially alters the regression model.

Scan to watch