

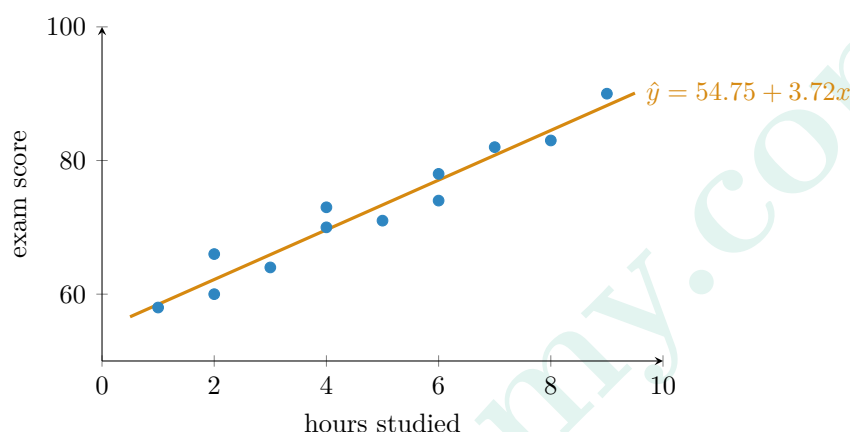
Correlation & Least-Squares Regression

AP Statistics | scatterplots, r , and the regression line

Unit 2 in the 2019 CED; Unit 5 (topics 5.1–5.5) from fall 2026.

1. Reading a scatterplot

A scatterplot shows two quantitative variables measured on the same individuals. The **explanatory variable** goes on the x -axis, the **response variable** on the y -axis. Below: hours studied and exam score for 12 students (hypothetical data).



Describe it with DUFS — always in context

Direction (positive / negative), **U**nusual features (outliers, clusters, gaps), **F**orm (linear / curved), **S**trength (strong / moderate / weak).

“There is a strong, positive, linear relationship between hours studied and exam score, with no apparent outliers.”

2. The correlation coefficient r

$$r = \frac{1}{n-1} \sum z_x z_y, \quad z_x = \frac{x - \bar{x}}{s_x}, \quad z_y = \frac{y - \bar{y}}{s_y}.$$

Each point contributes the product of its two z -scores. Points above-right or below-left of $(\bar{x}, \bar{y})$ contribute **positive** products; points in the other two corners contribute **negative** ones. For our students almost every product is positive, and $r = 0.97$.

- $-1 \leq r \leq 1$. The sign gives the direction, the size gives the strength.
- r has **no units**, and does not change if you change units (hours to minutes).
- r measures **linear** association only.
- r is **not resistant**: add one student with 9 hours and a score of 55, and r falls from 0.97 to 0.57.

$r = 0$ does not mean “no relationship”

Scores that rise and then fall (performance vs. stress, an upside-down U) can have $r = 0$: the best *straight* line is flat. $r = 0$ means no *linear* relationship. **Always plot the data first.**

3. The least-squares regression line

A line predicts $\hat{y}$ (“ y -hat”) for each x . The **residual** of a point is

$$\text{residual} = y - \hat{y} = \text{actual} - \text{predicted}.$$

Points above the line have positive residuals, points below negative ones. The **least-squares regression line** is the line that makes the **sum of the squared residuals** as small as possible. Its equation is $\hat{y} = a + bx$ with

$$b = r \frac{s_y}{s_x}, \quad a = \bar{y} - b\bar{x},$$

so the line always passes through $(\bar{x}, \bar{y})$. For our students $r = 0.97$, $s_x = 2.53$, $s_y = 9.67$, $\bar{x} = 4.75$, $\bar{y} = 72.42$, giving $\hat{y} = 54.75 + 3.72x$.

Interpreting slope and intercept

Templates AP readers look for

Slope: “For each additional hour studied, the *predicted* exam score increases by about 3.72 points.”

Intercept: “A student who studies 0 hours has a *predicted* exam score of about 54.8 points.” (If $x = 0$ makes no sense in context, say so.)

Cautions

- **Extrapolation:** our data cover 1–9 hours. At 20 hours the line predicts 129 on a test out of 100. Do not trust predictions far outside the data.
- **Correlation is not causation:** ice cream sales and sunburns are strongly correlated because both rise on hot days. Only a well-designed experiment shows cause and effect.

Where students lose points

1. Describing a scatterplot without context: name the variables.
2. Interpreting the slope without the word *predicted*, or without context.
3. Mixing up r and the slope: r has no units and lies in $[-1, 1]$; the slope has units of y per unit of x and can be any number.
4. Saying $r = 0$ means no relationship. It means no *linear* relationship.
5. Claiming causation from a correlation, or trusting an extrapolation.

Remember

- Describe: direction, unusual features, form, strength — in context.
- $-1 \leq r \leq 1$ measures the strength of a *linear* relationship only.
- $\hat{y} = a + bx$ minimizes the sum of squared residuals and passes through $(\bar{x}, \bar{y})$.
- Slope = change in *predicted* y for each 1-unit increase in x .

Now practise it

Theory sticks once you use it. On **sad-academy.com** you'll find practice questions for this exact subtopic in the AP format, and an AI that grades your free-response answers against the scoring guidelines, the way an AP reader would.

sad-academy.com

sad-academy.com

Data in this guide are hypothetical. AP[®] is a registered trademark of College Board, which is not affiliated with and does not endorse this material.