

Regression Analysis: Lines, Residuals and r

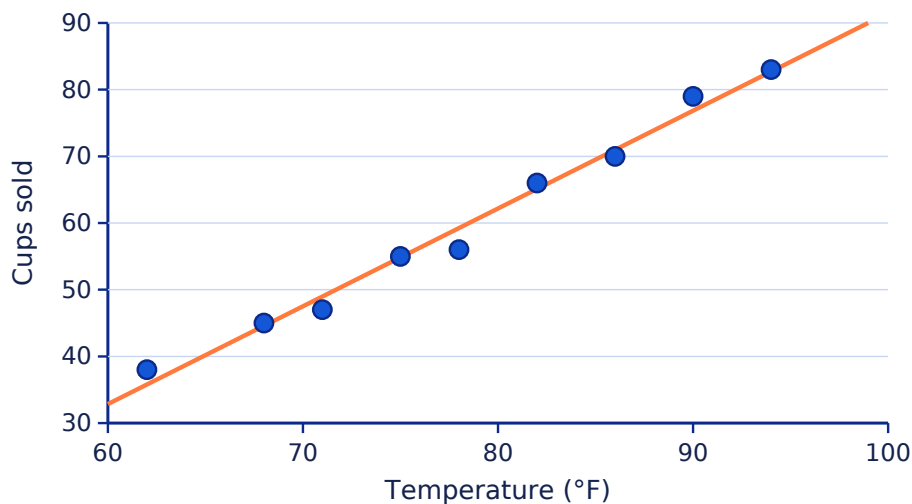
Grade 12 – AP Statistics math · Lesson · Standards: AP Statistics Unit 5 (2026-27)

Quick answer: Regression describes the linear relationship between an explanatory and a response variable. Correlation r measures the strength and direction of a linear pattern, the least-squares line $\hat{y} = a + bx$ predicts the response, r^2 gives the percent of variation explained, and residual plots check whether a line is appropriate.

Standards: AP Statistics Unit 5 (2026-27) · **Level:** Grade 12 AP Statistics

How do you describe a scatterplot?

A scatterplot shows two quantitative variables measured on the same individuals. The **explanatory** variable goes on the x-axis and the **response** variable on the y-axis. Describe the **direction** (positive or negative), **form** (linear or curved), **strength**, and unusual features such as outliers.



Smoothie stand: temperature and cups sold.

The smoothie stand data show a strong, positive, linear association between temperature and cups sold, with no outliers.

What does the correlation r measure?

Correlation

$$r = \frac{1}{(n-1)} \sum \frac{(x_i - \bar{x})}{s_x} \cdot \frac{(y_i - \bar{y})}{s_y}$$

$-1 \leq r \leq 1$. The sign gives the direction and the size gives the strength of a *linear* pattern.

Changing units does not change r , and r is not resistant to outliers.

Here $r = 0.9933$. Correlation does not show causation: a lurking variable can drive both variables.

How do you find and use the least-squares line?

Least-squares regression line

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

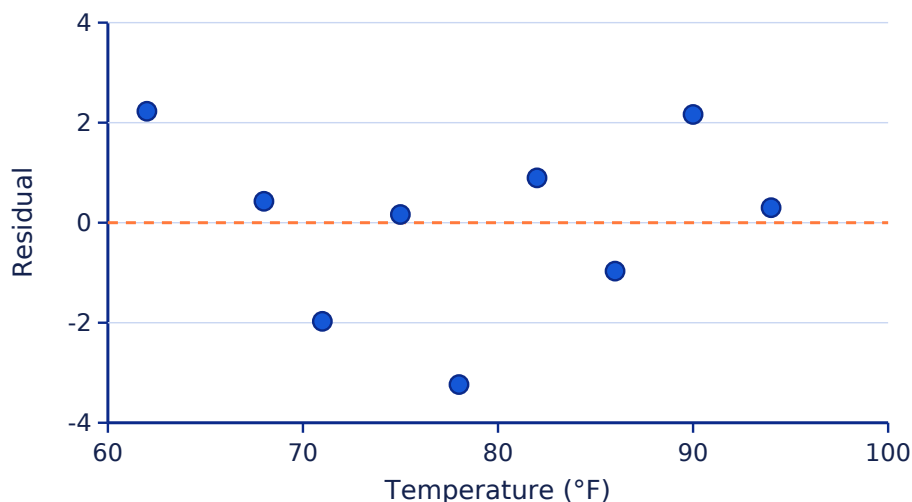
The line always passes through $(\bar{x}, \bar{y})$. Residual = $y - \hat{y}$ (actual minus predicted).

Worked example

For the smoothie data, $\bar{x} = 78.44$, $s_x = 10.561$, $\bar{y} = 59.89$, $s_y = 15.592$. Then $b = 0.9933 \times 15.592/10.561 = 1.467$ and $a = 59.89 - 1.467(78.44) = -55.15$. The line is $\hat{y} = -55.15 + 1.47x$. **Slope:** each additional degree is associated with about 1.47 more cups sold, on average. **Intercept:** at 0°F the model predicts -55.1 cups, which is outside the data and not meaningful. For 85°F , $\hat{y} = 69.5$ cups.

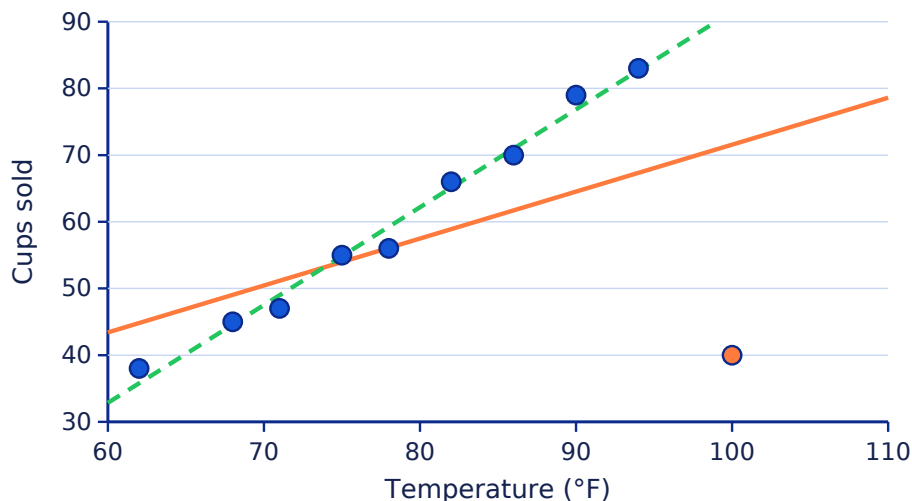
How do you judge whether the line fits?

The coefficient of determination $r^2 = 0.9866$ says that about 98.7% of the variation in cups sold is explained by the linear relationship with temperature. A **residual plot** should show random scatter around 0; a curved pattern means a line is not appropriate.



Residual plot: no pattern.

An **influential point** is one whose removal changes the line noticeably, typically a point far from the others in the x direction. **Extrapolation**, predicting outside the range of the data, is unreliable.



One high-leverage point changes the line (solid: with it; dashed: without).

Mathbot's tip

Interpret slope with the words on average, and name both variables with units. Never say that x causes y unless the data come from a randomized experiment.

Key points to remember

- Describe a scatterplot by direction, form, strength and unusual features.
- Slope $b = r \cdot s_y / s_x$; the line passes through $(\bar{x}, \bar{y})$.
- Residual = actual y – predicted $\hat{y}$.
- r^2 is the proportion of variation in y explained by the line.
- Extrapolation and influential points can make a model misleading.

Regression Analysis: Lines, Residuals and r : frequently asked questions

What does the slope of a regression line mean?

The slope is the predicted change in the response for each one-unit increase in the explanatory variable. Use the words predicted or on average, and include units, for example: each extra degree is associated with about 1.4 more cups sold.

What does r^2 tell you?

It is the proportion of the variation in the response that is accounted for by the linear relationship with the explanatory variable. An r^2 of 0.96 means about 96% of the variation in y is explained by the line; the rest is unexplained.

What is a residual and what does a residual plot show?

A residual is the actual y minus the predicted y . A residual plot graphs residuals against x or $\hat{y}$.

Random scatter around 0 means a line is appropriate; a curve or a fan shape shows the linear model is not a good fit.

What is the difference between an outlier and an influential point?

An outlier has an unusual y value given its x, giving a large residual. An influential point is one whose removal would substantially change the slope or intercept. Points far away in the x direction are often influential even with a small residual.

Why is extrapolation risky?

The linear pattern is only known for the range of x-values in the data. Beyond that range the relationship may change, so predictions can be far from the truth. Interpolating within the range is much safer.