

AP Statistics · Regression Analysis · Lesson

Common Core AP Statistics Unit 5 (2026-27) · mathovore.fr

This ap statistics two variable lesson explains regression analysis in one printable page, with definitions, a method and a worked example.

A scatterplot shows how two quantitative variables are related. The correlation r measures the strength and direction of a linear association, and the least-squares regression line predicts y from x . Interpret the slope, intercept and r squared in context, check a residual plot, and avoid extrapolation.

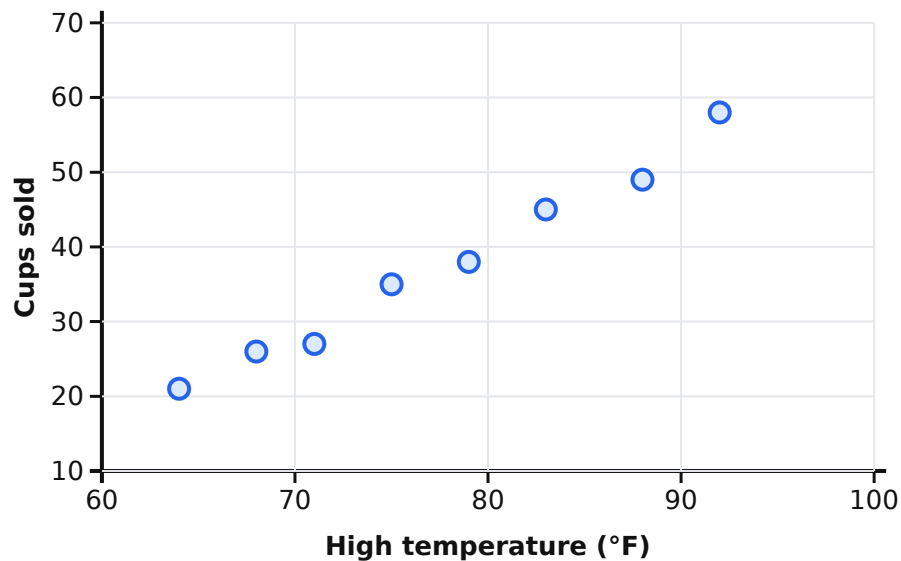
AP StatisticsCommon Core: AP Statistics Unit 5 (2026-27)

AP Statistics Two Variable Lesson: key facts

- Describe a scatterplot by direction, form, strength and unusual points, in context.
- The correlation r is between -1 and 1, has no units, and measures only linear association.
- The least-squares line has slope $b = r (s_y / s_x)$ and passes through $(\bar{x}, \bar{y})$.
- A residual is $y - \hat{y}$; a residual plot with no pattern supports a linear model.
- r squared is the percent of the variation in y explained by the linear relationship with x .

How do you describe the association in a scatterplot?

A scatterplot shows two quantitative variables for the same individuals. Put the **explanatory variable** x on the horizontal axis and the **response variable** y on the vertical axis. Describe the **direction** (positive or negative), **form** (linear or curved), **strength** and any **unusual points**, in context. Ethan's lemonade stand shows a strong positive linear association.



Lemonade sales rise with temperature: positive, linear and strong.

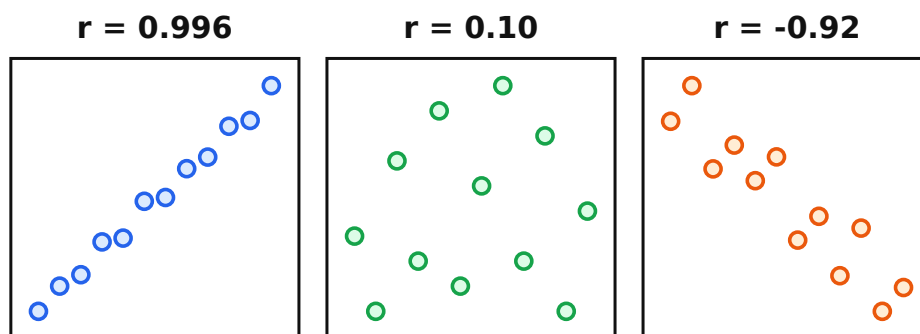
What does the correlation r measure?

Correlation

$$r = \frac{1}{n-1} \sum \left(\frac{x - \bar{x}}{s_x} \right) \left(\frac{y - \bar{y}}{s_y} \right)$$

r is between -1 and 1. It measures the strength and direction of a *linear* association only, has no units, and is not resistant to outliers.

For the lemonade data $r \approx 0.993$. Even a high r does not prove that x causes y .



Correlation measures how closely points follow a line, and its sign gives the direction.

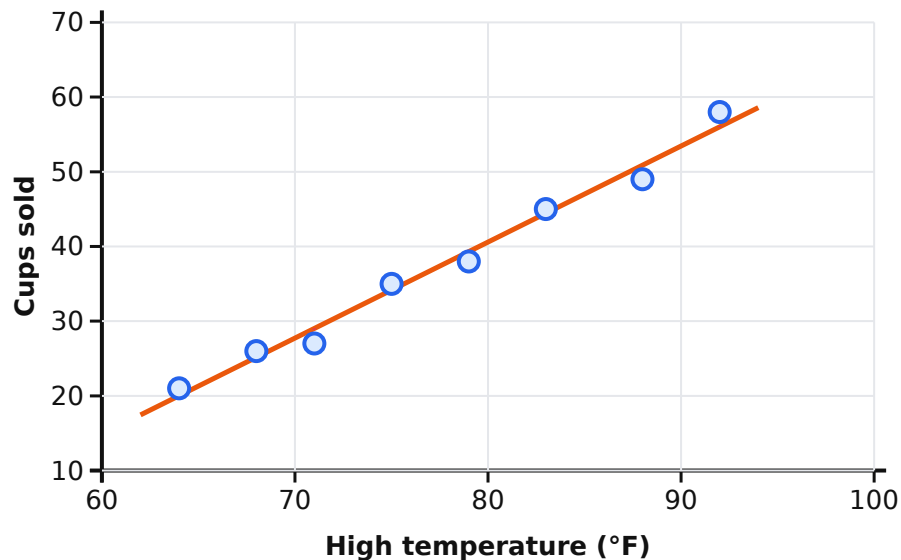
How do you find the least-squares regression line?

The **least-squares regression line** $\hat{y} = a + bx$ makes the sum of the squared residuals as small as possible. A **residual** is $y - \hat{y}$. Compute the slope and intercept from summary

statistics, since the line always passes through $(\bar{x}, \bar{y})$.

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

Example. Here $\bar{x} = 77.5$, $s_x \approx 9.81$ and $\bar{y} = 37.375$ and $s_y \approx 12.70$, with $r \approx 0.993$. Then $b = 0.993 \cdot \frac{12.70}{9.81} \approx 1.286$ and $a = 37.375 - 1.286(77.5) \approx -62.26$, so $\hat{y} = -62.26 + 1.286x$.



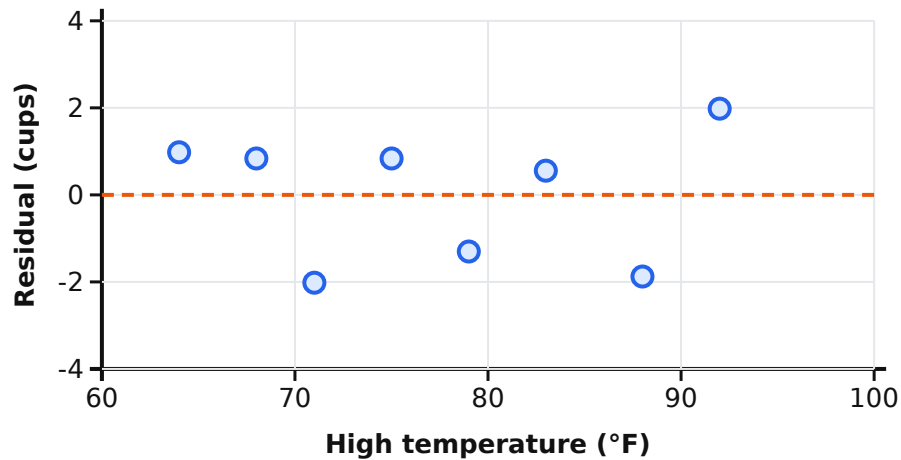
The least-squares line passes through $(77.5, 37.375)$; dashed segments are residuals.

How do you interpret the slope, the intercept and r squared?

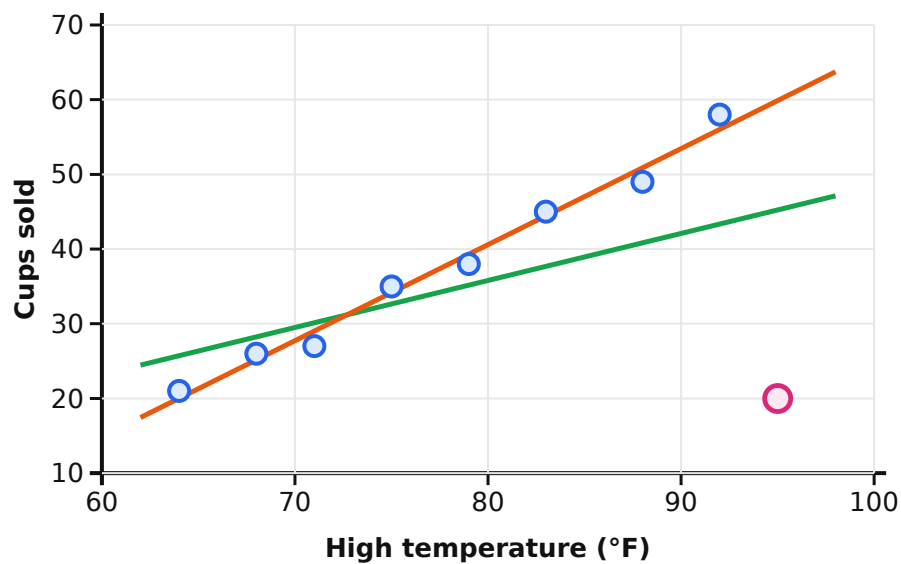
- **Slope:** for each additional 1 °F, the predicted number of cups sold increases by about 1.29.
- **Intercept:** at 0 °F the model predicts -62.3 cups, which is not meaningful because 0 °F is far outside the data.
- **r squared:** $r^2 \approx 0.986$ means about 98.6% of the variation in cups sold is explained by the linear relationship with temperature.

How do residual plots, influential points and extrapolation affect the model?

If the residual plot shows no pattern, a line is appropriate; a curve in the plot means the form is not linear. A point that is extreme in the x direction can be **influential**: removing it changes the slope or r a lot.



No pattern in the residuals, so a line is an appropriate model.



The added point (95, 20), circled in pink, drags the line from slope 1.29 (orange) to 0.63 (green).

Extrapolation, predicting outside the observed x values, is unreliable. The line gives ≈ 66.3 cups at 100 °F, but no data support that.

Regression Analysis: frequently asked questions

What is the difference between r and r squared?

The correlation r gives direction and strength of the linear association, from -1 to 1. The coefficient of determination r squared is the fraction of variation in y explained by the line, always between 0 and 1. If r is -0.9, then r squared is 0.81.

How do I interpret the slope of a regression line?

Say that for each additional 1 unit of x , the predicted y changes by the slope, naming both

variables and units. For a slope of 1.29, each extra degree is associated with about 1.29 more cups sold.

When is the intercept meaningful?

Only when $x = 0$ is plausible and within the range of the data. Otherwise just say it is the predicted y when x is 0 and note that it has no practical meaning.

What does a pattern in the residual plot mean?

A curve or a fan shape means a straight line is not a good model, even if r is high. Random scatter around zero with no pattern means the linear model is appropriate.

What is an influential point?

It is a point whose removal changes the slope, intercept or correlation substantially. Points far from the other x values are the most likely to be influential, even when their residual is small.

How to use this ap statistics two variable lesson

Use this ap statistics two variable lesson before the worksheet, then check what you learned with the quiz.

