

Unit 5: Regression Analysis

Unit 5 turns two-variable descriptions into a working prediction model. It covers scatterplots and describing association, the correlation coefficient, the least-squares regression line, residuals and residual plots, and the coefficient of determination.

AP Statistics

Regression Analysis

About 12 minutes to read

How to use this guide

Read it in order the first time because the topics build on each other. Scatterplots let you see the relationship, correlation measures its linear strength, the regression line models it, and residuals tell you whether the model was a good idea. Exam questions often give you a scatterplot or a regression equation and ask you to describe, predict, or judge the model.

After the first read, use the trap boxes and the comparison table to review the distinctions that exam questions test most often. Finish with the practice questions, then complete the recall check on the last page out loud and note any items you cannot explain yet.

Why this unit matters. Regression is where description becomes prediction, and the free-response section regularly asks you to interpret a slope, an intercept, or an r^2 in context. Those interpretation sentences follow fixed templates, and this guide gives you each one. Learn them word for word; free-response rubrics reward the exact phrasing.

5.1 Scatterplots and Describing Association

When you measure two quantitative variables on the same individuals, you get **bivariate quantitative data**: ordered pairs like (hours studied, test score) for each student. The **scatterplot** graphs each pair as a point, with the **explanatory variable** on the x-axis and the **response variable** on the y-axis. The explanatory variable is the one you use to explain or predict; the response is the one being explained. Which is which is a judgment call about the story, not a fact about the numbers. If you are predicting test scores from study hours, study hours go on the x-axis.

Every scatterplot description needs four pieces: **form**, **direction**, **strength**, and **unusual features**. Form is the shape of the pattern, described as linear or nonlinear. Direction is positive when response values tend to increase as explanatory values increase, and negative when they tend to decrease. Strength is how closely the points follow the pattern, described as strong, moderate, or weak. Unusual features are clusters of points or individual points that do not fit the general pattern. A complete description sounds like this: "There is a strong, positive, linear association between study hours and test score, with one student who scored much lower than expected for their hours."

Trap. Direction and form are separate judgments. A scatterplot can show a strong nonlinear association, and "negative direction" only means the points trend downward, not that the relationship is weak. Describe all four pieces every time, and name unusual features instead of ignoring them.

5.2 Correlation

The **correlation coefficient** r summarizes the strength and direction of the *linear* association between two quantitative variables. It is unit-free, so changing hours to minutes or points to percentages does not change r , and it always falls between -1 and 1 inclusive. The sign gives the direction and the distance from zero gives the strength. The closer r is to -1 or 1 , the stronger the linear association; $r = 0$ means no linear association, and $r = \pm 1$ means a perfect linear association with every point exactly on a line.

Two warnings come with r . First, it measures only *linear* association. A scatterplot with a perfect U-shaped curve can have r near 0 even though the variables are strongly related, so a value near ± 1 does not guarantee a linear model is appropriate, and a value near 0 does not mean there is no relationship at all. Always look at the scatterplot before trusting r . Second, **correlation does not imply causation**. A strong correlation between ice cream sales and drowning deaths does not mean ice cream causes drowning; both rise in summer. A perceived or real relationship between two variables never by itself proves that changes in one cause changes in the other.

Trap. The most expensive mistake in this unit is reading causation into a correlation. On the exam, any conclusion that one variable *caused* the change in the other is wrong unless the data come from a randomized experiment. Observational data can show association, never causation.

Trap. " $r = 0$ " means no *linear* association, not no association. If the scatterplot shows a clear curve and r is 0.08 , the correct reading is that a straight-line summary misses the real pattern, not that the variables are unrelated.

5.3 The Regression Line

When the scatterplot looks linear, a **linear regression model** approximates the relationship with the equation $\hat{y} = a + bx$, where x is the explanatory variable and $\hat{y}$ (y-hat) is the **predicted response value** for a given x . The **slope** b is the predicted change in the response for a one-unit increase in x . The **y-intercept** a is the predicted response when $x = 0$. For example, if the model for test scores is $\hat{y} = 52 + 4.5x$ with x as study hours, the slope says each additional hour of study predicts a 4.5-point increase in test score, and the intercept predicts a score of 52 for a student who studies 0 hours.

Predictions come in two kinds. **Interpolation** predicts using an x -value inside the range of the data that built the line, and it is the safe kind. **Extrapolation** predicts using an x -value beyond that range, and it gets less reliable the further out you go, because you have no evidence the linear pattern continues. Predicting a score for 3 hours when the data run from 1 to 6 hours is interpolation. Predicting for 20 hours is extrapolation.

Trap. Extrapolation is the prediction trap the exam loves. A question that asks for a prediction at an x -value far outside the data is testing whether you notice the danger, not whether you can plug into the equation. Compute the value if asked, but flag it as unreliable.

Trap. The intercept often has no sensible interpretation. If $x = 0$ is outside the data or gives an impossible value, like a negative height, say so plainly instead of inventing a story about it. The definition allows a predicted value at $x = 0$; context decides whether that value means anything.

5.4 Residuals and Checking the Model

A **residual** is the model's miss on one data point: observed y minus predicted $\hat{y}$. A positive residual means the model underpredicted (the point sits above the line); a negative residual means it overpredicted (the point sits below). If a student scored 78 but the model predicted 74, the residual is $78 - 74 = +4$.

The **residual plot** graphs each residual against its x -value (or against its predicted $\hat{y}$), and it is how you judge whether a straight line was the right model. When the **residual plot looks like random scatter** with no pattern, the linear form is confirmed and the model is appropriate. When the plot shows **curvature**, the linear model is not the best choice: the data have a bend the line cannot follow. A fan shape, where the spread of residuals grows as x grows, is another sign the straight-line model is struggling.

If the residual plot shows a curve, one standard fix is to **transform the data** before fitting the line. Taking the logarithm of the response variable, for example, can straighten a relationship that curves upward, because the log of exponential growth is linear. You then fit the regression to the transformed values and check the new residual plot. The transformation does not change the data; it changes the scale so that a linear model becomes appropriate.

Trap. A pattern in the residual plot is a verdict on the *model*, not on the data. Curvature means your straight line is the wrong tool, not that the points are wrong. The correct response is to reconsider the model, possibly with a transformation, rather than to delete points until the plot looks random.

5.5 Least Squares, r^2 , and Interpreting in Context

The **least-squares regression line** (LSRL) is the specific line that minimizes the sum of the squared residuals, which is why it is called the line of best fit. Squaring keeps positive and negative misses from canceling each other out and punishes large misses more than small ones. It is calculated with technology, and it always passes through the point $(\bar{x}, \bar{y})$, the mean of x and the mean of y .

The **coefficient of determination r^2** is the square of the correlation, and it tells you how much of the story the line explains. It is the proportion of the variation in the response variable explained by the linear relationship with the explanatory variable. If $r^2 = 0.72$, then 72 percent of the variation in test scores is explained by the linear relationship with study hours, and the remaining 28 percent is left to other factors and randomness.

Free-response questions demand interpretations in context, and each follows a template. For the **slope**: state the predicted increase or decrease in the response variable, with units, for a one-unit increase in the explanatory variable. "For each additional hour studied, the predicted test score increases by 4.5 points." For the **y-intercept**: state the predicted response when $x = 0$, with units, in context, and note when it has no reasonable interpretation. "The model predicts a test score of 52 points for a student who studies 0 hours." Keep the words "predicted" in both sentences. The line gives predictions, not guarantees.

One more distinction worth having: an **outlier** is a point with a large residual, far from the line in the y -direction, while an **influential point** is a point whose removal would noticeably change the regression line, usually because its x -value is far from the rest. A point can be influential without having a big residual if it sits far out on the x -axis and pulls the line toward itself. When you see an unusual point, ask both questions: how far is it from the line, and how far is it from the pack?

Trap. r^2 is a proportion of variation explained, not a percent-correct score. Saying "the model is 72 percent accurate" is wrong. The correct sentence always has the shape "72 percent of the variation in [response] is explained by the linear relationship with [explanatory]."

Confusions That Cost Points

PAIR	HOW TO KEEP THEM STRAIGHT
Explanatory vs response variable	The explanatory variable goes on the x-axis and does the predicting. The response goes on the y-axis and gets predicted. If the story is "predict score from hours," hours are explanatory.
Correlation vs causation	Correlation measures linear association in observational data. Causation needs a randomized experiment. A large r never proves cause on its own.
r vs r^2	r measures the strength and direction of linear association and ranges from -1 to 1 . r^2 is the proportion of variation in the response explained by the line and ranges from 0 to 1 .
Interpolation vs extrapolation	Interpolation predicts inside the data range and is trustworthy. Extrapolation predicts outside it and grows less reliable the further out you go.
Residual sign	Residual = observed – predicted. Positive means the point is above the line and the model underpredicted. Negative means below the line and overpredicted.
Residual plot patterns	Random scatter means the linear model fits. Curvature means it does not, and a transformation may help. A pattern indicts the model, not the data.
Outlier vs influential point	An outlier has a large residual. An influential point changes the line if removed, usually from an extreme x-value. A point can be one, both, or neither.
Slope vs intercept interpretation	Slope: predicted change in y per one-unit increase in x . Intercept: predicted y when $x = 0$. Both need units, context, and the word "predicted."

Practice Questions

Original questions written for this guide in the style of the AP exam. Answers and explanations are on the next page, so complete the questions before checking them.

1. A scatterplot of study hours (x) versus test score (y) shows points rising from left to right in a fairly tight straight-line pattern, with one student who scored far below the rest at a similar number of hours. Which description is correct?

- A. Weak, negative, nonlinear association with no unusual features
- B. Strong, positive, linear association with one unusual point
- C. Moderate, positive, linear association; the unusual point proves causation
- D. No association, because one point does not fit the pattern

2. For the same data, the correlation coefficient is $r = -0.87$. Which statement is correct?

- A. There is a strong negative linear association between the variables
- B. 87% of the variation in the response is explained by the model
- C. The variables have a weak association because r is negative
- D. Decreasing x by one unit causes y to decrease by 0.87 units

3. The least-squares regression line for predicting test score (y) from study hours (x) is $\hat{y} = 52 + 4.5x$. Which is the correct interpretation of the slope?

- A. For each additional hour studied, the predicted test score increases by 4.5 points
- B. A student who studies 0 hours will score exactly 52 points
- C. Studying causes test scores to rise by 4.5 points per hour
- D. 4.5% of the variation in test scores is explained by study hours

4. A residual plot of the regression in question 3 shows the residuals forming a clear U-shaped curve. What does this suggest?

- A. The linear model is appropriate because the residuals are all small
- B. The linear model is not appropriate; the relationship appears nonlinear
- C. The data contain an error and the curved points should be removed
- D. The correlation coefficient must be close to 1

Answer Key

- 1. B.** The points rise left to right (positive direction), follow a straight pattern (linear form), cluster tightly (strong), and one point breaks the pattern (unusual feature). A gets direction, strength, and form all wrong. C correctly starts but then claims causation, which a scatterplot can never show. D lets one point veto the clear pattern in the rest.
- 2. A.** The sign gives direction (negative) and $|r| = 0.87$ is close to 1, so the linear association is strong. B confuses r with r^2 ; the proportion of variation explained is $r^2 = 0.76$, not 0.87. C mistakes the negative sign for weakness; sign is direction only. D reads causation and a slope meaning into r , which measures neither.
- 3. A.** The slope template is "predicted change in y per one-unit increase in x , with units and context." B interprets the intercept instead of the slope, and drops the word "predicted." C claims causation from observational regression data. D confuses the slope with r^2 .
- 4. B.** Curvature in a residual plot means a straight line is the wrong model for the relationship. A mistakes small residuals for a good fit; the pattern is what matters, not the size. C blames the data instead of the model; the correct response is to reconsider the model, possibly with a transformation. D is backwards: a curved relationship is exactly when r understates what is going on.

One-Page Recall Check

- Explain what makes data bivariate quantitative and how a scatterplot is set up.
- Describe a scatterplot using form, direction, strength, and unusual features.
- State what r measures, its range, and why it is unit-free.
- Explain why r near 0 does not mean no relationship.
- State the correlation-does-not-imply-causation rule and when causation is allowed.
- Write the regression equation $\hat{y} = a + bx$ and identify each part.
- Interpret a slope and an intercept in context, with units and the word "predicted."
- Distinguish interpolation from extrapolation and explain the risk.
- Define a residual and explain what its sign tells you.
- Explain what a residual plot shows and what random scatter versus curvature means.
- Describe how a transformation can fix a curved relationship.
- State what the least-squares line minimizes and which point it always passes through.
- Interpret r^2 as a proportion of variation explained, in context.
- Distinguish an outlier from an influential point.

Where to go next. Turn every missed item above into flashcards and drill them spaced out over several days rather than in one sitting. In Rycal, open the Regression Analysis deck under AP Statistics at <https://rycal.web.app/apstats>. The deck covers the terms in this guide, and its practice questions target the same traps named here. If you have a test date, add it in the Test Planner. You can also start your next review with a Brain Dump, then check what you missed against this guide.

Key terms for this unit

Bivariate quantitative data, Scatterplot, Explanatory variable, Response variable, Form of association, Direction of association, Strength of association, Unusual features in a scatterplot, Correlation coefficient (r), Interpreting the correlation coefficient, Correlation does not imply causation, Linear regression model, Predicted response value ($\hat{y}$), Slope of the regression line, y -intercept of the regression line, Extrapolation, Interpolation, Residual, Residual plot, Assessing linearity with residual plots, Least-squares regression line (LSRL), Coefficient of determination (r^2), Interpreting the slope in context, Interpreting the y -intercept in context.

About this guide. Written for Rycal and aligned to the AP Statistics course framework, Unit 5: Regression Analysis. All questions and explanations are original Rycal writing. Rycal is independent and is not affiliated with or endorsed by the College Board.