

Simple Linear Regression

A guided model of fitted values, slopes, and residual variation

Ratiomera Statistics

Table of contents

Course-Page Summary	2
The Fitted Line	2
Inference and Diagnostics	2
Scope of the Conclusion	3
Expanded Reference	4
Purpose and foundations	4
Core ideas	4
Formula guide	5
Reading the explanatory figure	7
Interpretation checklist	8
How this topic connects	8

Course-Page Summary

Simple linear regression models the mean of one quantitative outcome from one predictor. It does not predict every case exactly. The fitted line describes the conditional mean, while each residual records one case's observed-minus-fitted difference.

The Fitted Line

The population model is

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

and the sample fitted line is

$$\hat{Y}_i = b_0 + b_1 X_i.$$

Table 1: Every regression quantity answers a different part of the fitted-model question.

Quantity	Meaning
b_0	Fitted mean outcome when $X = 0$; interpret only when zero is meaningful and supported by the data
b_1	Fitted change in mean outcome for a one-unit increase in X
$\hat{Y}_i$	Fitted mean outcome at case i 's predictor value
$e_i = Y_i - \hat{Y}_i$	Vertical residual; positive means the observed outcome lies above the line
R^2	Proportion of sample outcome variation represented by the fitted line
Residual standard error	Typical remaining outcome variation around the fitted line, in outcome units

The slope connects directly to Topic 4:

$$b_1 = \frac{s_{XY}}{s_X^2} = r_{XY} \frac{s_Y}{s_X}.$$

Covariance supplies joint movement, division by predictor variance turns it into outcome units per predictor unit, and the correlation form shows how the unit-free association is returned to the original scales. In simple regression with an intercept, $R^2 = r^2$.

Inference and Diagnostics

A slope confidence interval and test use

$$t = \frac{b_1 - 0}{SE(b_1)}, \quad df = n - 2.$$

Before trusting that inference, check whether a straight line is an adequate summary, whether residual spread is reasonably stable, whether the error distribution is compatible with the model, and whether cases are independent under the design. A residual-versus-fitted plot checks form and spread. A normal Q-Q plot checks the broad error-shape assumption. Leverage and Cook's distance identify cases for investigation, not automatic deletion.

Scope of the Conclusion

- Interpolation stays within the supported predictor range; extrapolation goes beyond it and requires stronger assumptions.
- A prediction for an individual needs more uncertainty than an interval for the mean outcome at the same predictor value.
- A small p-value does not prove causality, correct measurement, or good prediction in a new population.
- A large R^2 does not prove the model is correctly specified.

Topic 6 uses the residual idea twice: it adjusts both focal variables for a third variable and correlates what remains. That makes partial correlation the natural next step after the one-predictor line.

Expanded Reference

Purpose and foundations

Simple linear regression describes how the conditional mean of one numerical outcome changes across values of one predictor. The **outcome**, written Y , is the variable the model aims to explain or predict. The **predictor**, written X , supplies the explanatory information. “Simple” means that the model contains one predictor, and “linear” means that the modeled mean of Y changes along a straight line as X changes.

Each observed case has a predictor value x_i and an outcome value y_i . The model provides a fitted value $\hat{y}_i$, the outcome level predicted by the fitted line at that case’s predictor value. The difference between the observed and fitted outcome is the **residual** e_i . A positive residual places the point above the line; a negative residual places it below. Residuals keep the outcome’s unit and show the part of each observed value that the fitted line did not reproduce.

Regression is closely related to covariance and correlation, but its roles are different. Correlation treats the two variables symmetrically and summarizes standardized linear association. Regression assigns distinct predictor and outcome roles, retains the outcome unit, and provides an equation for fitted values. Reversing X and Y therefore produces a different regression problem even though their correlation is unchanged.

Component	Meaning	Unit
Intercept b_0	Fitted mean outcome when $X = 0$	Outcome units
Slope b_1	Fitted outcome change for a one-unit increase in X	Outcome units per predictor unit
Fitted value $\hat{y}_i$	Point on the fitted line at x_i	Outcome units
Residual e_i	Observed outcome minus fitted outcome	Outcome units

Core ideas

The intercept is mathematically required to position the line, but its substantive interpretation depends on whether zero is meaningful and represented by the data. If the observed predictor values are far from zero, the intercept is an extrapolation. In that situation, report it as a model coefficient without attaching a real-world baseline interpretation that the observed predictor range cannot support.

The slope is the central coefficient. A positive slope means that the fitted mean outcome increases as the predictor increases. A negative slope means it decreases. Its magnitude must be read with both units. A slope of two means two outcome units per one predictor unit, not a correlation of two. The slope describes an average conditional pattern, not a guaranteed change for every case.

Ordinary least squares chooses the intercept and slope that minimize the sum of squared residuals. Squaring prevents positive and negative residuals from canceling and gives greater weight to larger discrepancies. The fitted line passes through $(\bar{x}, \bar{y})$ when an intercept is included. The residuals then sum to approximately zero, apart from numerical rounding.

Diagnostic feature	Desired pattern	Concern suggested by a visible pattern
Residuals versus fitted values	Unstructured band around zero	Curvature, changing spread, or omitted structure

Diagnostic feature	Desired pattern	Concern suggested by a visible pattern
Residual spread	Roughly similar across fitted values	Nonconstant conditional variance
Normal quantile comparison	Approximate straight pattern when normal-error inference is used	Strong tail departures or unusual residuals
Leverage and influence	No single case dominates the fitted line	A case with unusual predictor position may strongly affect coefficients

R^2 compares the residual variation after fitting the line with the total variation around the outcome mean. It ranges from zero to one for an ordinary intercept model. A larger value means the fitted line accounts for more sample variation in Y , but it does not establish causality, guarantee accurate predictions for individuals, or prove that the model form is appropriate. A high R^2 can coexist with a systematic residual pattern.

Inference for the slope asks whether a population linear association is compatible with zero under the model. A confidence interval shows which slope values are compatible with the estimate and its standard error. The assumptions concern the conditional relationship: linear mean structure, independent observations, suitable residual variance, and a residual distribution adequate for the intended inference. The predictor itself does not have to be normally distributed.

Point prediction requires care. A fitted value estimates the conditional mean outcome at a chosen predictor value; it does not guarantee the outcome for one individual. Prediction within the observed range is interpolation. Prediction beyond that range is extrapolation and relies on an untested continuation of the fitted line, so the unsupported range must be identified explicitly.

Formula guide

The population model separates a systematic line from an individual error term:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

β_0 and β_1 are population coefficients. The error ε_i represents the difference between case i and the population conditional mean. After fitting sample data, the estimated line is:

$$\hat{y}_i = b_0 + b_1 x_i$$

The residual compares the observed value with this fitted value:

$$e_i = y_i - \hat{y}_i$$

The least-squares slope can be written using the cross-products introduced in covariance:

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

The intercept then positions the line through the two sample means:

$$b_0 = \bar{y} - b_1 \bar{x}$$

The coefficient of determination is the fraction of total outcome variation represented by the fitted model:

$$R^2 = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

With one predictor and an intercept, $R^2 = r_{xy}^2$. The slope can also be expressed as $b_1 = r_{xy}(s_y/s_x)$, which shows how standardized association is translated back into the variables' original units.

The fitted line separates each observed deviation into a model component and a residual component:

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i).$$

For an OLS model with an intercept, squaring and summing across all cases produces

$$SS_{\text{total}} = SS_{\text{model}} + SS_{\text{error}},$$

with

$$SS_{\text{total}} = \sum_i (y_i - \bar{y})^2, \quad SS_{\text{model}} = \sum_i (\hat{y}_i - \bar{y})^2, \quad SS_{\text{error}} = \sum_i (y_i - \hat{y}_i)^2.$$

The exact squared partition is a full-sample OLS result. Do not square the three signed distances for one case and expect the same identity to hold case by case.

The residual standard error estimates the typical error spread in outcome units:

$$s_e = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n-2}}.$$

The denominator uses $n-2$ because the line estimates an intercept and a slope. This quantity is different from the slope's standard error, which measures uncertainty in b_1 across hypothetical repeated samples:

$$SE(b_1) = \frac{s_e}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}.$$

For $H_0 : \beta_1 = 0$, the coefficient test and matching two-sided interval are

$$t = \frac{b_1}{SE(b_1)}, \quad df = n - 2,$$

$$b_1 \pm t_{1-\alpha/2, n-2} SE(b_1).$$

If the matching interval excludes zero, the matching two-sided test rejects the zero-slope null. Inclusion of zero means failure to reject, not proof that the population slope is exactly zero.

The model table divides the model and error sums of squares by their degrees of freedom:

$$F = \frac{MS_{\text{model}}}{MS_{\text{error}}}.$$

With one predictor, the global model test and the two-sided slope test ask the same null question, so $F = t^2$ and their p-values agree under the same model.

Prediction question	Meaning	Required caution
Fitted value at x_0	Estimated conditional mean $\hat{y} = b_0 + b_1 x_0$	It is not a guaranteed outcome for one case

Prediction question	Meaning	Required caution
Interpolation	x_0 lies within the observed predictor range	The model form and diagnostics still matter
Extrapolation	x_0 lies outside the observed predictor range	The relationship may differ where no data were observed

Measurement quality belongs in the interpretation. Under the classical predictor-measurement-error model developed in the supplied material, noise in X generally pulls the simple-regression slope toward zero. This attenuation can make an underlying relationship look weaker. It does not imply that every possible measurement-error process creates the same bias.

Reading the explanatory figure

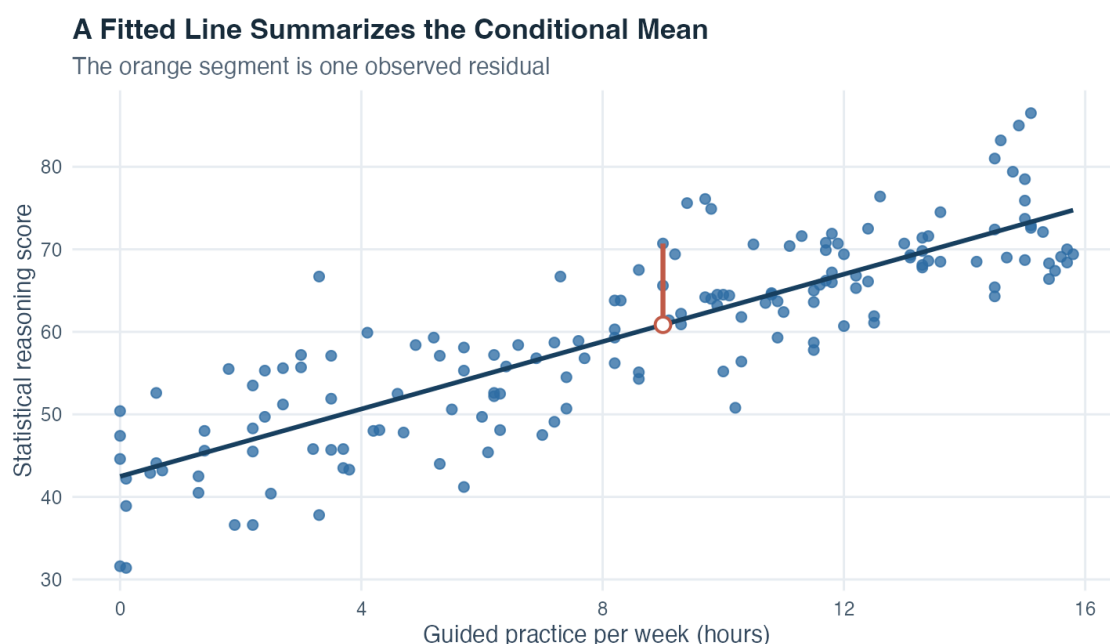


Figure 1: Scatterplot of guided-practice hours and statistical-reasoning scores with a rising fitted line and one orange vertical residual segment.

The horizontal axis gives weekly guided-practice hours, and the vertical axis gives statistical-reasoning scores. Each blue point is one simulated case. The dark line rises, so the fitted mean score is higher at larger practice values in this dataset. The line describes the average modeled pattern. Individual points remain scattered above and below it, reminding you that predictor information does not determine every person's outcome.

At about nine practice hours, the hollow circle marks the fitted score on the line. The observed point for that case is higher. The orange vertical segment is the residual: observed score minus fitted score. Its direction is positive and its length is measured in score points. Least squares performs this comparison for every point and chooses the line with the smallest total squared residual length.

The graph supports a statement about linear association in the simulated data. It does not show that additional practice caused higher scores. Prior preparation, selection, tutorial format, or other variables may be connected with both. It also does not show whether the residual assumptions hold; separate diagnostic plots are needed for curvature, changing variance, unusual residuals, and influence.

Interpretation checklist

Name the outcome and predictor and state their units. Inspect their distributions and scatterplot. Confirm that a straight-line summary is appropriate across the observed range. Report the fitted equation and translate the slope into a full sentence containing both units. Interpret the intercept only if its reference value is meaningful. Report R^2 as sample variation represented by the model, not as a causal percentage.

Inspect residual and influence diagnostics before relying on inference. Report the slope estimate, standard error, confidence interval, test statistic, degrees of freedom, and p-value when relevant. Distinguish an estimated conditional mean from a guaranteed individual outcome, and identify extrapolation. Describe simulated results as simulated and keep association language separate from causal language.

How this topic connects

This model makes the connection with covariance and correlation concrete. Covariance supplies the slope's numerator, predictor variance supplies its denominator, and correlation standardizes the same linear pairing. Regression adds direction and units: it asks how the fitted mean of a chosen outcome changes across the predictor.

The next steps address a central limitation of the one-predictor line. A third variable may account for part of the observed association. Partial correlation removes linear components associated with that third variable from both focal variables and correlates what remains. Multiple regression places several predictors in one model, so each coefficient describes a conditional association while the others are held fixed. The residual idea introduced here becomes the common language for both extensions.

Document ID: topic-05-simple-linear-regression-summary-en

Notes 1 of 3

Simple Linear Regression | Topic 5

Notes 2 of 3

Simple Linear Regression | Topic 5

Notes 3 of 3

Simple Linear Regression | Topic 5