

AP Statistics - Least-Squares Regression and Residuals - Lesson

Common Core AP STAT 2.6, AP STAT 2.7, AP STAT 2.8, AP STAT 2.9, HSS-ID.B.6, HSS-ID.C.7 · maths-pdf.fr/en/

The least-squares regression line minimizes the sum of squared residuals and has the form $\hat{y} = a + bx$, with slope $b = r(s_y/s_x)$. The slope is the predicted change in y per one-unit increase in x , the residual is actual minus predicted, and r squared is the proportion of variation in y explained by the line.

AP Statistics · Common Core State Standards: **AP STAT 2.6, AP STAT 2.7, AP STAT 2.8, AP STAT 2.9, HSS-ID.B.6, HSS-ID.C.7**

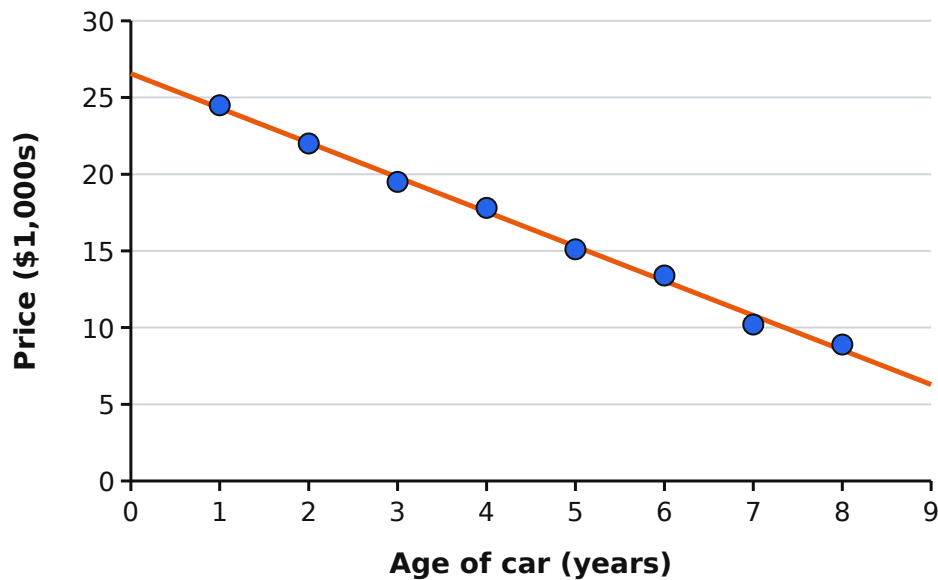
Key facts

- LSRL: $\hat{y} = a + bx$, with $b = r s_y/s_x$ and $a = \bar{y} - b\bar{x}$; it passes through $(\bar{x}, \bar{y})$.
- Interpret the slope as the predicted change in y per unit of x , in context.
- Residual = actual - predicted; a good linear model has a residual plot with no pattern.
- r squared is the percent of variation in y explained by the linear model; s is the typical prediction error.
- Avoid extrapolation and check for influential points.

What is the least-squares regression line?

A **regression line** is a straight line that models how the response variable y changes with the explanatory variable x . The **least-squares regression line** (LSRL) is the line that minimizes the sum of the squared vertical distances from the points to the line. Its equation is $\hat{y} = a + bx$.

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



Least-squares line for car price (in \$1,000s) against age: $\hat{y} = 26.56 - 2.25(\text{age})$. Dashed segments are residuals.

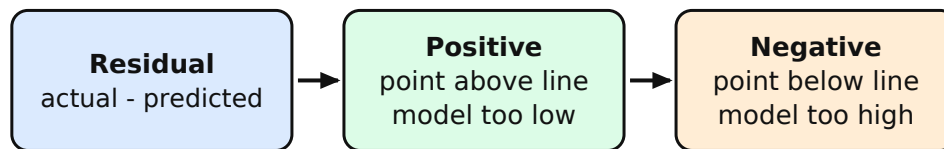
For the car data, $\hat{y} = 26.56 - 2.25x$, where x is age in years and $\hat{y}$ is predicted price in thousands of dollars. The line always passes through $(\bar{x}, \bar{y})$.

How do you interpret the slope and intercept?

Slope: for each additional year of age, the predicted price decreases by about \$2,252.

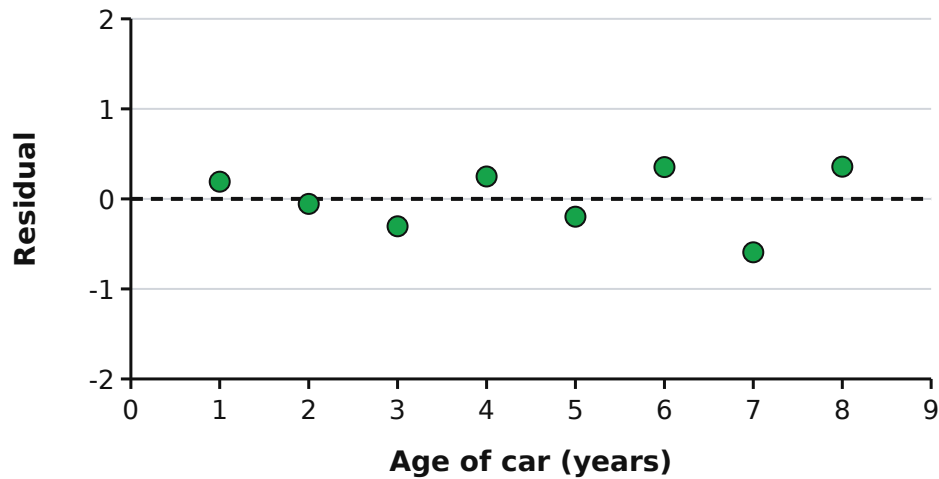
Intercept: the predicted price of a car of age 0 is about \$26,561. Always use the words "predicted" and the context. Use of the line far outside the data range is **extrapolation** and is unreliable.

What is a residual?

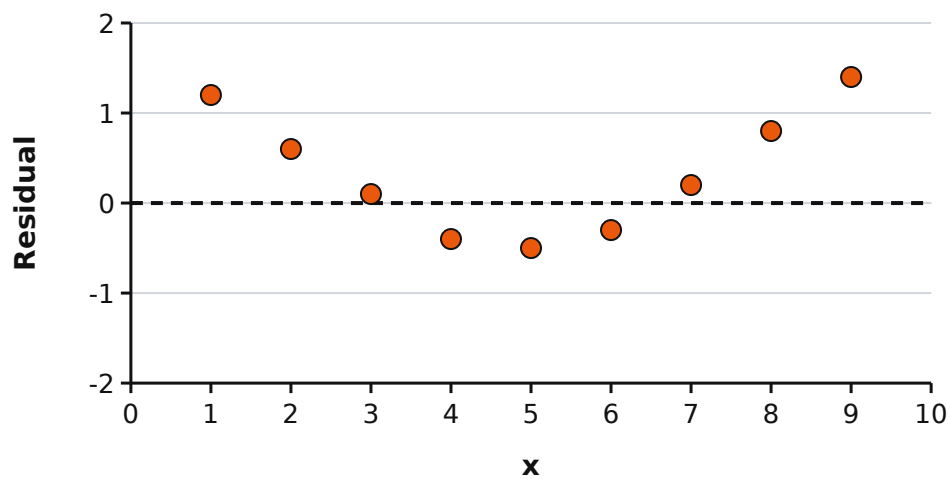


$$\text{Residual} = y - \hat{y}.$$

A **residual** is the actual value minus the predicted value: $e = y - \hat{y}$. A residual plot graphs residuals against x . Random scatter around zero (no pattern) supports a linear model; a curve means the line is not appropriate.



A residual plot with no pattern: a linear model is appropriate.



A curved residual pattern signals that a straight line is not the right model.

How do you measure how well the line fits?

The **coefficient of determination** r^2 is the proportion of variation in y explained by the linear relationship with x . Here $r^2 = 0.996$, so about 99.6% of the variation in price is explained by age. The **standard deviation of the residuals** s measures the typical prediction error:

$s = \sqrt{\frac{\sum e_i^2}{n-2}}$. Here $s = 0.37$, about \$374.

Predictor	Coef	SE Coef
Constant	26.561	0.2915
Age	-2.252	0.0577
S = 0.3741	R-Sq = 99.6%	

Computer output: intercept, slope, S, and R-Sq.

Points far from the others in the x direction can be **influential**: removing them changes the line a lot. Check for such points before you trust a regression.

Questions about least-squares regression and residuals

What does the slope of the regression line mean?

The slope is the predicted change in the response for each one-unit increase in the explanatory variable. Always state it with 'predicted' and with the units and context of the data.

How do I find a residual?

Plug the x value into the regression equation to get the predicted value, then subtract:
 $\text{residual} = \text{actual } y - \text{predicted } y$. A positive residual means the point is above the line.

What is extrapolation, and why is it risky?

Extrapolation is predicting for x values outside the range of the data. The pattern may not continue there, so the prediction can be badly wrong.

What does r squared tell me?

It gives the percent of the variation in y that is accounted for by the linear relationship with x. For example, $r^2 = 0.96$ means 96% of the variation in price is explained by age.

When is a point influential?

A point is influential if removing it substantially changes the slope, intercept, or correlation. Points with extreme x values far from the other points are the most likely to be influential.