

Vocabulary

Term	Definition
Association	Direction: A positive direction or association means that, in general, as one variable increases, so does the other. When increases in one variable generally correspond to decreases in the other, the association is negative.
	Form: The form we care about most is straight, but you should certainly describe other patterns you see in scatterplots.
	Strength: A scatterplot is said to show a strong association if there is little scatter around the underlying relationship.
Correlation	Correlation is a numerical measure of the direction and strength of a linear association.
Explanatory variable	The x -variable in a relationship that explains, predicts, or is otherwise responsible for the y -variable. Sometimes called the <i>independent variable</i> .
Extrapolation	Although linear models provide an easy way to predict values for y for a given value of x , it is unsafe to predict for values of x far from the ones used to find the linear model equation. Such extrapolation may pretend to see into the future, but the predictions should not be trusted.
Influential point	If omitting a point from the data results in a very different regression model, then that point is called an influential point.
Intercept	The intercept, b_0 , gives a starting value in y -units. It's the y -value when x is 0.
Ladder of Powers	The Ladder of Powers places in order the effects that many re-expressions have on the data
Least squares	The least squares criterion specifies the unique line that minimizes the variance of the residuals, or, equivalently, the sum of the squared residuals.
Leverage	Data points whose x -values are far from the mean of x are said to exert leverage on a linear model. High-leverage points pull the line close to them, and so they can have a large effect on the line, sometimes completely determining the slope and intercept. With high enough leverage, their residuals can appear to be deceptively small.
Linear model	A linear model is an equation of the form $\hat{y} = b_0 + b_1x$. To interpret a linear model we need to know the variables (along with their W's) and their units.
Lurking variable	A variable other than x and y that simultaneously affect both variables, accounting for the correlation between the two.
Lurking variable	A variable that is not explicitly part of a model but affects the way the variables in the model appear to be related is called a lurking variable. Because we can never be certain that observational data are not hiding a lurking variable that influences both x and y , it is never safe to conclude that a linear model demonstrates a causal relationship, no matter how strong the linear association.

Vocabulary

Term	Definition
Model	An equation or formula that simplifies and represents equality.
Outlier	A point that does not fit the overall pattern seen in the scatterplot.
Outlier	Any data point that stands away from the others can be called an outlier. In regression, outliers can be extraordinary in two ways, by having a large residual or by having high leverage.
Predicted value	The value of y found for each x -value in the data. A predicted value is found by substituting the x -value in the regression equation. The predicted values are the values on the fitted line; the points (x, \hat{y}) all lie exactly on the fitted line.
R^2	R^2 is the square of the correlation between x and y . R^2 gives the fraction of the variability of y accounted for by the least squares linear regression on x . R^2 is an overall measure of how successful the regression is in linearly related y to x .
Re-express data	We re-express data by taking the logarithm, the square root, the reciprocal, or some other mathematical operation on all values in the data set.
Regression line or Line of best fit	The particular equation $\hat{y} = b_0 + b_1x$ that satisfies the least squares criterion is called the least squares regression line. Casually, we often just call it the regression line, or the line of best fit.
Regression to the mean	Because the correlation is always less than 1.0 in magnitude, each predicted y tends to be fewer standard deviations from its mean than its corresponding x was from its mean. This is called regression to the mean.
Residuals	Residuals are the differences between data values and the corresponding values predicted by the regression model - or, more generally, values predicted by any model. Residual = observed value - predicted value = $y - \hat{y}$
Response variable	The y -variable that you hope to predict or explain. Sometimes called the <i>dependent variable</i> .
Scatterplots	A scatterplot shows the relationship between two quantitative variables measured on the same cases.
Slope	The slope gives a value in "y-units per x-unit." Changes of one unit in x are associated with changes of b_1 units in predicted values of \hat{y} .
Subset	One unstated condition for finding a linear model is that the data be homogeneous. If, instead, the data consist of two or more groups that have been thrown together, it is usually best to fit different linear models to each group than to try to fit a single model to all of the data. Displays of the residuals can often help you find subsets in the data.