A **regression line** is a straight line that describes how a response variable y changes as an explanatory variable x changes. We often use a regression line to predict the value of y for a given value of x.

Suppose that y is a response variable (plotted on the vertical axis) and x is an explanatory variable (plotted on the horizontal axis). A straight line relating y to x has an equation of the form

$$y = a + bx$$

In this equation, b is the **slope,** the amount by which y changes when x increases by one unit. The number a is the **intercept,** the value of y when $x = 0$.

The **least-squares regression line** of y on x is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.

A **residual** is the difference between an observed value of the response variable and the value predicted by the regression line. That is, a residual is the prediction error that remains after we have chosen the regression line:

$$\text{residual} = \text{observed } y - \text{predicted } y$$

A **residual plot** is a scatterplot of the regression residuals against the explanatory variable. Residual plots help us assess how well a regressi

An observation is **influential** for a statistical calculation if removing it would markedly change the result of the calculation.

Extrapolation is the use of a regression line for prediction far outside the range of values of the explanatory variable x that you used to obtain the line. Such predictions are often not accurate.

A **lurking variable** is a variable that is not among the explanatory or response variables in a study and yet may influence the interpretation of relationships among those variables.