Chapter 3

Coefficient of determination r^2 The fraction of the variation in the values of *y* that is accounted for by the least-squares regression line of *y* on *x*. We can calculate r^2 using the following formula: $r^2 = 1 - \frac{\text{SSE}}{\text{SST}}$ where SSE = Σ residual^2 = $\Sigma\ (y_i - \hat{y})^2$ and SST = $\Sigma\ (y_i - \overline{y})^2$.

Correlation Measures the direction and strength of the linear relationship between two quantitative variables. Correlation is usually written as *r*.

Equation of the least-squares regression line $\hat{y} = a + bx$ with slope $b = r\frac{s_y}{s_x}$ and *y* intercept $a = \overline{y} - b\overline{x}$.

Explanatory variable A variable that may help explain or influences changes in a response variable.

Extrapolation The use of a regression line for prediction far outside the interval of values of the explanatory variable *x* used to obtain the line. Such predictions are often not accurate.

Influential An observation is influential for a statistical calculation if removing it would markedly change the result of the calculation. Points that are outliers in the *x* direction of a scatterplot are often influential for the least-squares regression line.

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

Negative association When above-average values of one variable tend to accompany below-average values of the other, and vice versa.

Overall pattern In any graph of data, look for the overall pattern and for striking departures from that pattern. You can describe the overall pattern of a scatterplot by the *direction, form, and strength* of the relationship.

Outlier An observation that lies outside the overall pattern of the other observations. Points that are outliers in the *y* direction but not the *x* direction of a scatterplot have large residuals. Other outliers may not have large residuals.

Positive association When above-average values of one variable tend to accompany above-average values of the other, and below-average values also tend to occur together.

Predicted value $\hat{y}$ *(read "y hat")* is the **predicted value** of the response variable y for a given value of the explanatory variable x.

Regression line A 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*.

Residual The difference between an observed value of the response variable and the value predicted by the regression line. That is,

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

Residual plot A scatterplot of the regression residuals against the explanatory variable (or equivalently, against the predicted *y*-values). Residual plots help us assess how well a regression line fits the data.

Response variable A variable that measures an outcome of a study.

Scatterplot Shows the relationship between two quantitative variables measured on the same individuals. The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis. Each individual in the data appears as a point in the graph.

Slope Suppose that *y* is a response variable (plotted on the vertical axis) and *x* is an explanatory variable (plotted on the horizontal axis). A regression line relating *y* to *x* has an equation of the form $\hat{y} = a + bx$. In this equation, *b* is the slope, the amount by which *y* is predicted to change when *x* increases by one unit.

Standard deviation of the residuals (*s*) If we use a least-squares line to predict the values of a response variable *y* from an explanatory variable *x*, the standard deviation of the residuals (*s*) is given by

$$s = \sqrt{\frac{\sum \text{residuals}^2}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y})^2}{n-2}}$$

This value gives the approximate size of a "typical" or "average" prediction error (residual).

***y* intercept** Suppose that *y* is a response variable (plotted on the vertical axis) and *x* is an explanatory variable (plotted on the horizontal axis). A regression line relating *y* to *x* has an equation of the form $\hat{y} = a + bx$. In this equation, the number *a* is the *y* intercept, the predicted value of *y* when $x = 0$.