10-3 Regression

Key Concept Suppose that we have a collection of paired data and we use the methods of Section 10-2 to conclude that there is a linear correlation between two variables. This section presents methods for finding the equation of the straight line that best fits a scatterplot of the sample data. That best-fitting straight line is called the *regression line*, and its equation is called the *regression equation*. We can use the regression equation to make predictions for the value of one of the variables given some specific value of the other variable. In Part 2 of this section we discuss marginal change, influential points, and residual plots as tools for analyzing correlation and regression results.

Part 1: Basic Concepts of Regression

In some cases, two variables are related in a *deterministic* way, meaning that given a value for one variable, the value of the other variable is exactly determined without any error, as in the equation $y = 2.54x$ for converting a distance x from inches to centimeters. Such equations are considered in algebra courses, but statistics courses focus on *probabilistic* models, which are equations with a variable that is not determined completely by the other variable. For example, the height of a child cannot be determined completely by the height of the father and/or mother. Sir Francis Galton (1822–1911) studied the phenomenon of heredity and showed that when tall or short couples have children, the heights of those children tend to *regress,* or revert to the more typical mean height for people of the same gender. We continue to use Galton's "regression" terminology, even though our data do not involve the same height phenomena studied by Galton.

DEFINITIONS

Given a collection of paired sample data, the **regression line** (or *line of best fit,* or *least-squares line*) is the straight line that "best" fits the scatterplot of the data. (The specific criterion for the "best-fitting" straight line is the "least-squares" property described later.)

The **regression equation**

$$\hat{y} = b_0 + b_1 x$$

algebraically describes the regression line.

The regression equation expresses a relationship between x (called the **explanatory variable,** or **predictor variable,** or **independent variable**) and $\hat{y}$ (called the **response variable,** or **dependent variable**). The preceding definition shows that in statistics, the typical equation of a straight line $y = mx + b$ is expressed in the form $\hat{y} = b_0 + b_1 x$, where b_0 is the y-intercept and b_1 is the slope.

The values of the slope b_1 and y-intercept b_0 can be easily found by using any one of the many computer programs and calculators designed to provide those values. (See "Using Technology" at the end of this section.)

Prediction Worth $1 Million

Netflix is a large movie rental company that recently sponsored a contest with a $1 million prize for developing a system for predicting whether someone will like a movie based on how much they liked other movies. Netflix had developed its own system it called Cinematch, but the contest required a new system with a substantial improvement over Cinematch. The $1 million prize was won by a team called BellKor's Pragmatic Chaos, which included members of the Statistics Research Department at AT&T and others.

Contestants used data from 100 million movie ratings. The objective was to use the past movie ratings to predict which movies people would prefer, and the predictions were compared to movies that the people later viewed and rated. The winning prediction system is quite complex and is beyond the scope of this text.