14.1 Sample and Population Regression Models

This is Section 14.1 of Mind On Statistics, 3rd edition by Utts and Heckard, copyright Brroks and Cole

A **regression model** describes the relationship between a quantitative response variable (the y variable) and one or more explanatory variables (x variables). The y variable is sometimes called the **dependent variable,** and because regression models may be used to make predictions, the x variables may be called the **predictor variables.** The labels *response variable* and *explanatory variable* may be used for the variables on the y axis and x axis, respectively, even if there is not an obvious way to assign these labels in the usual sense.

Any regression model has two important components. The most obvious component is the equation that describes how the mean value of the y variable is connected to specific values of the x variable. The equation for the connection between handspan and height, Average handspan $= -3 + 0.35$ (Height), is an example. In this chapter, we focus on *linear relationships,* so a straight-line equation will be used, but it is important to note that some relationships are *curvilinear.*

The second component of a regression model describes how individuals vary from the regression line. Figure 14.1, which is identical to Figure 5.6, displays the raw data for the sample of $n = 167$ handspans and heights along with the regression line that estimates how the mean handspan is connected to specific heights. Notice that most individuals vary from the line. When we examine sample data, we will find it useful to estimate the general size of the deviations from the line. When we consider a model for the relationship within the population represented by a sample, we will state assumptions about the distribution of deviations from the line.

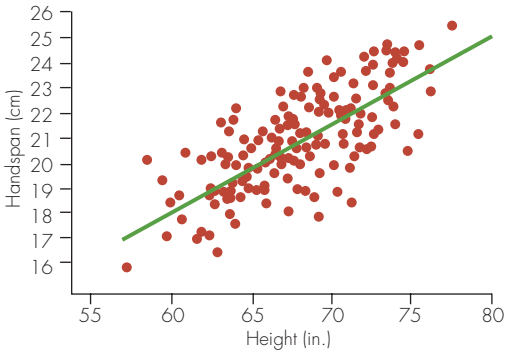

Figure 14.1 ▮ Regression line linking handspan and height for a sample of college students

14-W3527 11/3/05 11:14 AM Page 601

If the sample represents a larger population, we need to distinguish between the **regression line for the sample** and the **regression line for the population.** The observed data can be used to determine the regression line for the sample, but the regression line for the population can only be imagined. Because we do not observe the whole population, we will not know numerical values for the intercept and slope of the regression line in the population. As in nearly every statistical problem, the statistics from a sample are used to estimate the unknown population parameters, which in this case are the slope and intercept of the regression line.

The Regression Line for the Sample

In Chapter 5, we introduced this notation for the regression line that describes sample data:

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

In any given situation, the sample is used to determine values for b_0 and b_1.

- $\hat{y}$ is pronounced "y-hat" and is also referred to either as *predicted y* or *estimated y*.
- b_0 is the **intercept** of the straight line. The *intercept* is the value of $\hat{y}$ when $x = 0$.
- b_1 is the **slope** of the straight line. The *slope* tells us how much of an increase (or decrease) there is for $\hat{y}$ when the x variable increases by one unit. The sign of the slope tells us whether $\hat{y}$ increases or decreases when x increases. If the slope is 0, there is no linear relationship between x and y because $\hat{y}$ is the same for all values of x.

The equation describing the relationship between handspan and height for the sample of college students can be written as

$$\hat{y} = -3 + 0.35x$$

In this equation,

- $\hat{y}$ estimates the average handspan for any specific height x. If height = 70 in., for instance, $\hat{y} = -3 + 0.35(70) = 21.5$ cm.
- The *intercept* is $b_0 = -3$. While necessary for the line, this value does not have a useful statistical interpretation in this example. It estimates the average handspan for individuals who have height = 0 in., an impossible height far from the range of the observed heights. It also is an impossible handspan.
- The *slope* is $b_1 = 0.35$. This value tells us that the *average increase* in handspan is 0.35 cm for every 1-inch increase in height.

Reminder: The Least-Squares Criterion

In Chapter 5, we described the least-squares criterion. This mathematical criterion is used to determine numerical values of the intercept and slope of a sample regression line. The **least-squares line** is the line, among all possible lines, that has the smallest sum of squared differences between the sample values of y and the corresponding values of $\hat{y}$.

14-W3527 11/3/05 11:14 AM Page 602

Deviations from the Regression Line in the Sample

The terms *random error, residual variation,* and *residual error* all are used as synonyms for the term **deviation.** Most commonly, the word **residual** is used to describe the deviation of an observed y value from the sample regression line. A *residual* is easy to compute. It simply is the difference between the observed y value for an individual and the value of $\hat{y}$ determined from the x value for that individual.

Example 14.1

Residuals in the Handspan and Height Regression Consider a person who is 70 inches tall whose handspan is 23 centimeters. The sample regression line is $\hat{y} = -3 + 0.35x$, so $\hat{y} = -3 + 0.35(70) = 21.5$ cm for this person. The *residual* = observed y − predicted $y = y - \hat{y} = 23 - 21.5 = 1.5$ cm. Figure 14.2 illustrates this residual.

Statistics Now™

Watch a video example at **http://1pass.thomson.com** or on your CD.

For software help, download your Minitab, Excel, TI-83, SPSS, R, and JMP manuals from **http://1pass.thomson.com**, or find them on your CD.

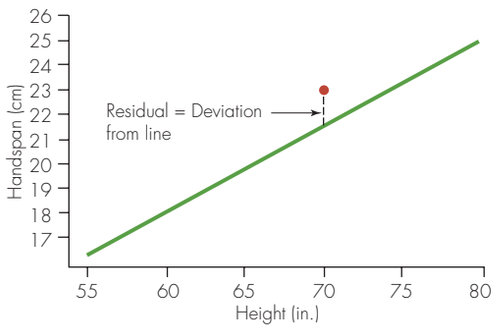

Figure 14.2 ▮ Residual for a person 70 inches tall with a handspan = 23 cm. The residual is the difference between observed $y = 23$ and $\hat{y} = 21.5$, the predicted value for a person 70 inches tall. ■

definition

For an observation y_i in the sample, the **residual** is

$e_i = y_i - \hat{y}_i$

y_i = the value of the response variable for the observation.

$\hat{y}_i = b_0 + b_1 x_i$, where x_i is the value of the explanatory variable for the observation.

technical note

The sum of the residuals is 0 for any least-squares regression line. The least-squares formulas for determining the equation always result in $\Sigma y_i = \Sigma \hat{y}_i$, so $\Sigma e_i = 0$.

The Regression Line for the Population

The regression equation for a simple linear relationship in a population can be written as

$$E(Y) = \beta_0 + \beta_1 x$$

14-W3527 11/3/05 11:14 AM Page 603

- $E(Y)$ represents the mean or expected value of y for individuals in the population who all have the same particular value of x. Note that $\hat{y}$ is an estimate of $E(Y)$.
- β_0 is the **intercept** of the straight line in the **population.**
- β_1 is the **slope** of the straight line in the **population.** Note that if the slope $\beta_1 = 0$, there is no linear relationship in the population.

Unless we measure the entire population, we cannot know the numerical values of β_0 and β_1. These are population parameters that we estimate using the corresponding sample statistics. In the handspan and height example, $b_1 = 0.35$ is a sample statistic that estimates the population parameter, β_1, and $b_0 = -3$ is a sample statistic that estimates the population parameter β_0.

technical note

Multiple Regression

In **multiple regression,** the mean of the response variable is a function of two or more explanatory variables. Put another way, in multiple regression we use the values of more than one explanatory (predictor) variable to predict the value of a response variable. For example, a college admissions committee might predict college GPA for an applicant based on using Verbal SAT, Math SAT, high school GPA, and class rank. The general structure of an equation for doing this might be

$$\text{College GPA} = \beta_0 + \beta_1 \text{ Verbal SAT} + \beta_2 \text{ Math SAT} + \beta_3 \text{ HS GPA} + \beta_4 \text{ Class Rank}$$

As in simple regression, numerical estimates of the parameters β_0, β_1, β_2, β_3, and β_4 would be determined from a sample. On the CD for this book, multiple regression is covered in Supplemental Topic 3.

Assumptions About Deviations from the Regression Line in the Population

To make statistical inferences about the population, two assumptions about how the y values vary from the population regression line are necessary. First, we assume that there is a **constant variance,** meaning that the general size of the deviation of y values from the line is the same for all values of the explanatory variable (x). This assumption may or may not be correct in any particular situation, and a scatterplot should be examined to see whether it is reasonable or not. In Figure 14.1 (p. 600), the constant variance assumption looks reasonable because the magnitude of the deviation from the line appears to be about the same across the range of observed heights.

The second assumption about the population is that for any specific value of x, the distribution of y values is a normal distribution. Equivalently, this assumption is that deviations from the population regression line have a normal curve distribution. Figure 14.3 (see next page) illustrates this assumption along with the other elements of the population regression model for a linear relationship. The line $E(Y) = \beta_0 + \beta_1 x$ describes the mean of y, and the normal curves describe deviations from the mean.

14-W3527 11/3/05 11:14 AM Page 604

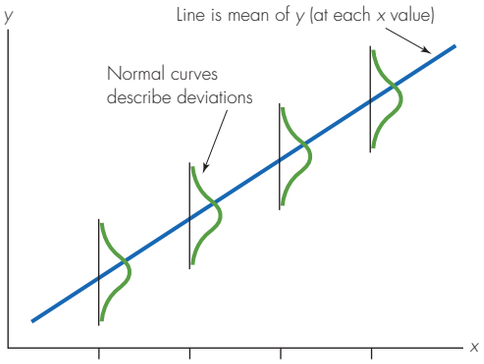

Figure 14.3 ■ Regression model for population

The Simple Regression Model for a Population

A useful format for expressing the components of the population regression model is

$$y = \text{Mean} + \text{Deviation}$$

This conceptual equation states that for any individual, the value of the response variable (y) can be constructed by combining two components:

1. The *mean,* which in the population is the line $E(Y) = \beta_0 + \beta_1 x$ if the relationship is linear. There are other possible relationships, such as curvilinear, a special case of which is a quadratic relationship, $E(Y) = \beta_0 + \beta_1 x + \beta_2 x^2$. Relationships that are not linear will not be discussed further in this book.
2. The individual's *deviation* = y − *mean,* which is what is left unexplained after accounting for the mean y value at that individual's x value.

This format also applies to the sample, although technically, we should use the term *estimated mean* when referring to the sample regression line.

Example 14.2

Mean and Deviation for Height and Handspan Regression Recall from Example 14.1 that the sample regression line for handspans (y) and heights (x) is $\hat{y} = -3 + 0.35x$. Although it is not likely to be true, let's assume for convenience that this equation also holds in the population. If your height is $x = 70$ in. and your handspan is $y = 23$ cm, then

$$\text{Mean} = -3 + 0.35(70) = 21.5$$
$$\text{Deviation} = y - \text{Mean} = 23 - 21.5 = 1.5$$
$$y = 23 = \text{Mean} + \text{Deviation} = 21.5 + 1.5$$

In other words, the mean handspan for people with your height is 21.5 cm, and your handspan is 1.5 cm above that mean. ■

In the theoretical development of procedures for making statistical inferences for a regression model, the collection of all *deviations* in the population is assumed to have a normal distribution with mean 0 and standard deviation σ

14-W3527 11/3/05 11:14 AM Page 605

(so the variance is σ^2). The value of the standard deviation σ is an unknown population parameter that is estimated by using the sample. This standard deviation can be interpreted in the usual way in which we interpret a standard deviation. It is, roughly, the average distance between individual values of y and the mean of y as described by the regression line. In other words, it is roughly the size of the average deviation across all individuals in the range of x values.

Keeping the regression notation straight for populations and samples can be confusing. Although we have not yet introduced all relevant notation, a summary at this stage will help you to keep it straight.

14.1 Exercises are on page 626.

in summary

Simple Linear Regression Model: Population and Sample Versions

For (x_1, y_1), (x_2, y_2), . . . , (x_n, y_n), a sample of n observations of the explanatory variable x and the response variable y from a large population, the **simple linear regression model** describing the relationship is as follows.

Population Version

Mean: $E(Y) = \beta_0 + \beta_1 x$

Individual: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = E(Y_i) + \varepsilon_i$

The deviations ε_i are assumed to follow a normal distribution with mean 0 and standard deviation σ.

Sample Version

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

Individual: $y_i = b_0 + b_1 x_i + e_i = \hat{y}_i + e_i$

where e_i is the *residual* for individual i. The sample statistics b_0 and b_1 *estimate* the population parameters β_0 and β_1. The mean of the residuals is 0, and the residuals can be used to estimate the population standard deviation σ.