PTER

9

SIMPLE LINEAR REGRESSION AND CORRELATION

CHAPTER OVERVIEW

This chapter provides an introduction and overview of two common techniques for exploring the strength of the relationship between two variables. The first technique, linear regression, will help us find an objective way to predict or estimate the value of one variable given a value of another variable. The second technique, correlation, will help us find an objective measure of the strength of the relationship between two variables.

Este capítulo proporciona una introducción y una descripción general de dos técnicas comunes para **explorar la fuerza de la relación entre dos variables**. La primera técnica, la **regresión lineal**, nos ayudará a encontrar una forma objetiva de **predecir o estimar el valor de una variable dado el valor de otra variable**. La segunda técnica, la **correlación**, nos ayudará a encontrar una medida objetiva de la **fuerza de la relación entre dos variables.**

TO P I C S

9.1INTRODUCTION

9.2THE REGRESSION MODEL

9.3THE SAMPLE REGRESSION EQUATION

9.4EVALUATING THE REGRESSION EQUATION

9.5USING THE REGRESSION EQUATION

9.6THE CORRELATION MODEL

9.7THE CORRELATION COEFFICIENT

9.8SOME PRECAUTIONS

9.9SUMMARY

LEARNING OUTCOMES

**After studying this chapter, the student will**

1. be able to obtain a simple linear regression model and use it to make predictions.

Ser capaz de obtener un modelo de regresión lineal simple y utilizarlo para hacer predicciones.

2. be able to calculate the coefficient of determination and to interpret tests of regression coefficients.

Ser capaz de calcular el **coeficiente de determinación** e interpretar p**ruebas de coeficientes de regresión.**

3. be able to calculate correlations among variables.

4. understand how regression and correlation differ and when the use of each is appropriate.

comprender en qué se diferencian la regresión y la correlación y cuándo es apropiado el uso de cada una.

9.1 INTRODUCTION

In analyzing data for the health sciences disciplines, we find that it is frequently desirable to learn something about the relationship between two numeric variables. We may, for example, be interested in studying the relationship between blood pressure and age, height and weight, the concentration of an injected drug and heart rate, the consumption level of some nutrient and weight gain, the intensity of a stimulus and reaction time, or total family income and medical care expenditures. The nature and strength of the relationships between variables such as these may be examined by regression and correlation analysis, two statistical techniques that, although related, serve different purposes

.Al analizar datos para las disciplinas de las ciencias de la salud, encontramos que frecuentemente es deseable aprender algo sobre la **relación entre dos variables numéricas**. Nos puede interesar, por ejemplo, estudiar la relación entre la presión arterial y la edad, la altura y el peso, la concentración de un fármaco inyectado y la frecuencia cardíaca, el nivel de consumo de algún nutriente y el aumento de peso, la intensidad de un estímulo y el tiempo de reacción. , o ingresos familiares totales y gastos de atención médica. La naturaleza y la fuerza de las relaciones entre variables como estas pueden examinarse mediante análisis de regresión y correlación, dos técnicas estadísticas que, aunque relacionadas, **tienen propósitos diferentes**.

**Regression**

Regression analysis is helpful in assessing specific forms of the relationship between variables, and the ultimate objective when this method of analysis is employed usually is to predict or estimate the value of one variable corresponding to a given value of another variable. The ideas of regression were first elucidated by the English scientist Sir Francis Galton (1822–1911) in reports of his research on heredity—first in sweet peas and later in human stature. He described a tendency of adult offspring, having either short or tall parents, to revert back toward the average height of the general population. He first used the word reversion, and later regression, to refer to this phenomenon.

El **análisis** **de regresión** es útil para evaluar formas específicas de relación entre variables, y el objetivo final cuando se emplea este método de análisis suele ser predecir o estimar el valor de una variable correspondiente a un valor dado de otra variable. Las ideas de regresión fueron aclaradas por primera vez por el **científico inglés Sir Francis Galton (1822-1911)** en informes de sus investigaciones sobre la herencia, primero en los guisantes de olor y luego en la estatura humana. Describió una tendencia de la descendencia adulta, con padres altos o bajos, a volver a la altura promedio de la población general. Primero utilizó la palabra **reversión**, y luego regresión, para referirse a este fenómeno.

**Correlation**

Correlation analysis, on the other hand, is concerned with measuring the strength of the relationship between variables. When we compute measures of correlation from a set of data, we are interested in the degree of the correlation between variables. Again, the concepts and terminology of correlation analysis originated with Galton, who first used the word correlation in 1888.

El **análisis de correlación**, por otro lado, se ocupa de medir la fuerza de la relación entre variables. Cuando calculamos medidas de correlación a partir de un conjunto de datos, nos interesa el **grado de correlación entre variables**. Una vez más, los conceptos y la terminología del análisis de correlación se originaron con Galton, quien utilizó por primera vez la palabra correlación en 1888.

In this chapter our discussion is limited to the exploration of the linear relationship between two variables. The concepts and methods of regression are covered first, beginning in the next section. In Section 9.6 the ideas and techniques of correlation are introduced. In the next chapter we consider the case where there is an interest in the relationships among three or more variables.

En este capítulo nuestra discusión se limita a la **exploración de la relación lineal** entre dos variables. Los conceptos y métodos de regresión se tratan primero, a partir de la siguiente sección. En la Sección 9.6 se introducen las ideas y técnicas de correlación. En el próximo capítulo consideramos el caso en el que hay interés en las relaciones entre tres o más variables.

Regression and correlation analysis are areas in which the speed and accuracy of a computer are most appreciated. The data for the exercises of this chapter, therefore, are presented in a way that makes them suitable for computer processing. As is always the case, the input requirements and output features of the particular programs and software packages to be used should be studied carefully.

El análisis de regresión y correlación son áreas en las que se aprecia más la velocidad y precisión de una computadora. Por lo tanto, los datos de los ejercicios de este capítulo se presentan de manera que sean adecuados para el procesamiento informático. Como siempre es el caso, se deben estudiar cuidadosamente los requisitos de entrada y las características de salida de los programas y paquetes de software particulares que se utilizarán.

**9.2 THE REGRESSION MODEL**

In the typical regression problem, as in most problems in applied statistics, researchers have available for analysis a sample of observations from some real or hypothetical population. Based on the results of their analysis of the sample data, they are interested in reaching decisions about the population from which the sample is presumed to have been drawn. It is important, therefore, that the researchers understand the nature of the population in which they are interested. They should know enough about the population to be able either to construct a mathematical model for its representation or to determine if it reasonably fits some established model.

En el problema típico de regresión, como en la mayoría de los problemas de estadística aplicada, los investigadores tienen disponible para su análisis una **muestra de observaciones de alguna población real o hipotética.** Con base en los resultados de su análisis de los datos de la muestra, están interesados en **tomar decisiones sobre la población de la cual se supone que se extrajo la muestra.** Por lo tanto, es importante que los investigadores **comprendan la naturaleza de la población** que les interesa. Deben saber lo suficiente sobre la población para poder construir un modelo matemático para su representación o determinar si se ajusta razonablemente a algún modelo establecido.

A researcher about to analyze a set of data by the methods of simple linear regression, for example, should be secure in the knowledge that the simple linear regression model is, at least, an approximate representation of the population. It is unlikely that the model will be a perfect portrait of the real situation, since this characteristic is seldom found in models of practical value. A model constructed so that it corresponds precisely with the details of the situation is usually too complicated to yield any information of value. On the other hand, the results obtained from the analysis of data that have been forced into a model that does not fit are also worthless. Fortunately, however, a perfectly fitting model is not a requirement for obtaining useful results. Researchers, then, should be able to distinguish between the occasion when their chosen models and the data are sufficiently compatible for them to proceed and the case where their chosen model must be abandoned.

Un investigador que esté a punto de analizar un conjunto de datos mediante los métodos de **regresión lineal simple**, por ejemplo, debe estar seguro de que el modelo de regresión lineal simple es, al menos, una **representación aproximada de la población**. Es poco probable que el modelo sea un retrato perfecto de la situación real, ya que esta característica rara vez se encuentra en modelos de valor práctico. Un modelo construido de manera que se corresponda exactamente con los detalles de la situación suele ser demasiado complicado para proporcionar información de valor. Por otro lado, los resultados obtenidos del análisis de datos que han sido forzados a formar parte de un modelo que no se ajusta tampoco tienen valor. Afortunadamente, sin embargo, un modelo que se ajuste perfectamente no es un requisito para obtener resultados útiles. Los investigadores, entonces, deberían poder distinguir entre la ocasión en que los modelos elegidos y los datos son suficientemente compatibles para continuar y el caso en el que el modelo elegido debe ser abandonado.

**Assumptions Underlying Simple Linear Regression**

In the simple linear regression model two variables, usually labeled X and Y, are of interest. The letter X is usually used to designate a variable referred to as the independent variable, since frequently it is controlled by the investigator; that is, values of X may be selected by the investigator and, corresponding to each preselected value of X, one or more values of another variable, labeled Y, are obtained. The variable, Y, accordingly, is called the dependent variable, and we speak of the regression of Y on X. The following are the assumptions underlying the simple linear regression model.

En el modelo de regresión lineal simple, son de interés dos variables, normalmente denominadas X e Y. La letra X suele utilizarse para designar una variable denominada variable independiente, ya que f**recuentemente está controlada por el investigado**r; es decir, el investigador puede seleccionar valores de X y, correspondientes a cada valor preseleccionado de X, se obtienen uno o más valores de otra variable, denominada Y. En consecuencia, la variable Y se denomina variable dependiente y **hablamos de la regresión de Y sobre X**. Los siguientes son los supuestos subyacentes al modelo de regresión lineal simple.

1. Values of the independent variable X are said to be “fixed.” This means that the values of X are preselected by the investigator so that in the collection of the data they are not allowed to vary from these preselected values. In this model, X is referred to by some writers as a nonrandom variable and by others as a mathematical variable. It should be pointed out at this time that the statement of this assumption classifies our model as the classical regression model. Regression analysis also can be carried out on data in which X is a random variable.

1. Se dice que **los valores de la variable independiente X son "fijos"**. Esto significa que los valores de X son preseleccionados por el investigador de modo que en la recopilación de datos no se les permite variar de estos valores preseleccionados. En este modelo, algunos autores se refieren a X como una variable no aleatoria y otros como una variable matemática. Cabe señalar en este momento que la formulación de este supuesto clasifica nuestro modelo como el modelo de regresión clásico. El análisis de regresión también se puede realizar con datos en los que X es una variable aleatoria.

2. The variable X is measured without error. Since no measuring procedure is perfect,this means that the magnitude of the measurement error in X is negligible.

**La variable X se mide sin error.** Dado que ningún procedimiento de medición es perfecto, esto significa que la magnitud del error de medición en X es insignificante.

3. For each value of X there is a subpopulation of Y values. For the usual inferential procedures of estimation and hypothesis testing to be valid, these subpopulations must be normally distributed. In order that these procedures may be presented it will be assumed that the Y values are normally distributed in the examples and exercises that follow.

**Para cada valor de X existe una subpoblación de valores de Y.** Para que los procedimientos inferenciales habituales de estimación y prueba de hipótesis sean válidos, estas subpoblaciones deben estar distribuidas normalmente. Para poder presentar estos procedimientos, se asumirá que los valores de Y se distribuyen normalmente en los ejemplos y ejercicios siguientes.

4. The variances of the subpopulations of Y are all equal and denoted by **s2**.

5. **The means of the subpopulations of Y all lie on the same straight line.** This is known as the assumption of linearity. This assumption may be expressed symbolically as

my|x = b 0 + b 1x

(9.2.1)

where my|x is the mean of the subpopulation of Y values for a particular value of X, and b 0 and b 1 are called the population regression coefficients. Geometrically, b 0 and b 1 represent the y-intercept and slope, respectively, of the line on which all of the means are assumed to lie.

6. The Y values are statistically independent. In other words, in drawing the sample, it is assumed that the values of Y chosen at one value of X in no way depend on the values of Y chosen at another value of X.

**Los valores de Y son estadísticamente independientes.** En otras palabras, al extraer la muestra, se supone que los valores de Y elegidos para un valor de X no dependen de ninguna manera de los valores de Y elegidos para otro valor de X.

These assumptions may be summarized by means of the following equation, which is called the regression model:

**y = b 0 + b 1x + P**

(9.2.2)

where y is a typical value from one of the subpopulations of Y, b 0 and b 1 are as defined for Equation 9.2.1, and P is called the error term. If we solve 9.2.2 for P, we have

P = y - 1b 0 + b 1x2

(9.2.3)

= y - mu ƒx

and we see that P shows the amount by which y deviates from the mean of the subpopulation of Y values from which it is drawn. As a consequence of the assumption that the subpopulations of Y values are normally distributed with equal variances, the P’s for each subpopulation are normally distributed with a variance equal to the common variance of the subpopulations of Y values.

**y vemos que P muestra la cantidad en la que y se desvía de la media de la subpoblación de valores de Y de la que se extrae.** Como consecuencia del supuesto de que las subpoblaciones de valores de Y se distribuyen normalmente con varianzas iguales, las P para cada subpoblación se distribuyen normalmente con una varianza igual a la varianza común de las subpoblaciones de valores de Y.

The following acronym will help the reader remember most of the assumptions necessary for inference in linear regression analysis:

El siguiente acrónimo ayudará al lector a recordar la mayoría de los supuestos necesarios para la inferencia en el análisis de regresión lineal:

LINE [Linear (assumption 5), Independent (assumption 6), Normal (assumption 3), Equal variances (assumption 4)]

A graphical representation of the regression model is given in Figure 9.2.1.

**9.3 THE SAMPLE REGRESSION EQUATION**

In simple linear regression the object of the researcher’s interest is the population regression equation—the equation that describes the true relationship between the dependent variable Y and the independent variable X. The variable designated by Y is sometimes called the response variable and X is sometimes called the predictor variable.

En la regresión lineal simple, el objeto de interés del investigador es la **ecuación de regresión poblacional,** la ecuación que describe la verdadera relación entre la variable dependiente Y y la variable independiente X. La variable designada por Y a veces se llama **variable de respuesta** y a X se le llama a veces variable de respuesta. la **variable predictora.**

In an effort to reach a decision regarding the likely form of this relationship, the researcher draws a sample from the population of interest and using the resulting data, computes a sample regression equation that forms the basis for reaching conclusions regarding the unknown population regression equation.

En un esfuerzo por llegar a una decisión sobre la forma probable de esta relación, el investigador extrae una muestra de la población de interés y, utilizando los datos resultantes, calcula una ecuación de regresión muestral que forma la base para llegar a conclusiones sobre la ecuación de regresión de la población desconocida.

**Steps in Regression Analysis**

In the absence of extensive information regarding the nature of the variables of interest, a frequently employed strategy is to assume initially that they are linearly related. Subsequent analysis, then, involves the following steps.

En ausencia de información extensa sobre la naturaleza de las variables de interés, **una estrategia frecuentemente empleada es suponer inicialmente que están relacionadas linealmente.** El análisis posterior, entonces, implica los siguientes pasos.

1. Determine whether or not the assumptions underlying a linear relationship are met in the data available for analysis.

Determine si los supuestos subyacentes a una relación lineal se cumplen o no en los datos disponibles para el análisis.

2. Obtain the equation for the line that best fits the sample data.

3. Evaluate the equation to obtain some idea of the strength of the relationship and the usefulness of the equation for predicting and estimating.

Evalúe la ecuación para obtener una idea de la fuerza de la relación y la utilidad de la ecuación para predecir y estimar.

4. If the data appear to conform satisfactorily to the linear model, use the equation obtained from the sample data to predict and to estimate.

Si los datos parecen ajustarse satisfactoriamente al modelo lineal, utilice la ecuación obtenida de los datos de la muestra para predecir y estimar.

When we use the regression equation to predict, we will be predicting the value Y is likely to have when X has a given value. When we use the equation to estimate, we will be estimating the mean of the subpopulation of Y values assumed to exist at a given value of X. Note that the sample data used to obtain the regression equation consist of known values of both X and Y. When the equation is used to predict and to estimate Y, only the corresponding values of X will be known. We illustrate the steps involved in simpler linear regression analysis by means of the following example.

Cuando usamos la ecuación de regresión para predecir, estaremos prediciendo el valor que probablemente tendrá Y cuando X tenga un valor dado. Cuando usamos la ecuación para estimar, estimaremos la media de la subpoblación de valores de Y que se supone existen para un valor dado de X. Tenga en cuenta que los datos de muestra utilizados para obtener la ecuación de regresión consisten en valores conocidos de X e Y. Cuando se utiliza la ecuación para predecir y estimar Y, sólo se conocerán los valores correspondientes de X. Ilustramos los pasos involucrados en un análisis de regresión lineal más simple mediante el siguiente ejemplo.

**EXAMPLE 9.3.1**

Després et al. (A-1) point out that the topography of adipose tissue (AT) is associated with metabolic complications considered as risk factors for cardiovascular disease. It is important, they state, to measure the amount of intraabdominal AT as part of the evaluation of the cardiovascular-disease risk of an individual. Computed tomography (CT), the only available technique that precisely and reliably measures the amount of deep abdominal AT, however, is costly and requires irradiation of the subject. In addition, the technique is not available to many physicians. Després and his colleagues conducted a study to develop equations to predict the amount of deep abdominal AT from simple anthropometric measurements. Their subjects were men between the ages of 18 and 42 years who were free from metabolic disease that would require treatment.

Després et al. (A-1) señalan que la **topografía del tejido adiposo (TA)** se asocia con **complicaciones metabólicas** consideradas como factores de riesgo de enfermedad cardiovascular. Es importante, afirman, medir la cantidad de AT intraabdominal como parte de la evaluación del riesgo de enfermedad cardiovascular de un individuo. Sin embargo, la tomografía computarizada (TC), la única técnica disponible que mide de manera precisa y confiable la cantidad de AT abdominal profunda, es costosa y requiere irradiación del sujeto. Además, la técnica no está disponible para muchos médicos. Després y sus colegas realizaron un estudio para desarrollar ecuaciones para predecir la cantidad de AT abdominal profundo a partir de mediciones **antropométricas** simples. Sus sujetos eran hombres de entre 18 y 42 años que no padecían enfermedades metabólicas que requirieran tratamiento.

Among the measurements taken on each subject were deep abdominal AT obtained by CT and waist circumference as shown in Table 9.3.1. A question of interest is how well one can predict and estimate deep abdominal AT from knowledge of the waist circumference. This question is typical of those that can be answered by means of regression analysis. Since deep abdominal AT is the variable about which we wish to make predictions and estimations, it is the dependent variable. The variable waist measurement, knowledge of which will be used to make the predictions and estimations, is the independent variable.

Entre las mediciones tomadas en cada sujeto se encontraban la AT abdominal profunda obtenida por TC y la circunferencia de la cintura como se muestra en la Tabla 9.3.1. Una cuestión de interés es qué tan bien se puede predecir y estimar la AT abdominal profunda a partir del conocimiento de la circunferencia de la cintura. Esta pregunta es típica de las que pueden responderse mediante análisis de regresión. Dado que la AT abdominal profunda es la variable sobre la cual deseamos hacer predicciones y estimaciones, es la variable dependiente. La variable medida de la cintura, cuyo conocimiento se utilizará para realizar las predicciones y estimaciones, es la variable independiente.

**The Scatter Diagram**

A first step that is usually useful in studying the relationship between two variables is to prepare a scatter diagram of the data such as is shown in Figure 9.3.1. The points are plotted by assigning values of the independent variable X to the horizontal axis and values of the dependent variable Y to the vertical axis.

Un primer paso que suele ser útil para estudiar la relación entre dos variables es preparar un diagrama de dispersión de los datos como se muestra en la Figura 9.3.1. Los puntos se trazan asignando valores de la variable independiente X al eje horizontal y valores de la variable dependiente Y al eje vertical.

The pattern made by the points plotted on the scatter diagram usually suggests the basic nature and strength of the relationship between two variables. As we look at Figure 9.3.1, for example, the points seem to be scattered around an invisible straight line. The scatter diagram also shows that, in general, subjects with large waist circumferences also have larger amounts of deep abdominal AT. These impressions suggest that the relationship between the two variables may be described by a straight line crossing the Y-axis below the origin and making approximately a 45-degree angle with the X-axis. It looks as if it would be simple to draw, freehand, through the data points the line that describes the relationship between X and Y.

El patrón formado por los puntos trazados en el diagrama de dispersión suele sugerir la naturaleza básica y la fuerza de la relación entre dos variables. Si observamos la Figura 9.3.1, por ejemplo, los puntos parecen estar dispersos alrededor de una línea recta invisible. El diagrama de dispersión también muestra que, en general, los sujetos con circunferencias de cintura grandes también tienen mayores cantidades de AT abdominal profunda. Estas impresiones sugieren que la relación entre las dos variables puede describirse mediante una línea recta que cruza el eje Y por debajo del origen y forma aproximadamente un ángulo de 45 grados con el eje X. Parece que sería sencillo dibujar, a mano alzada, a través de los puntos de datos la línea que describe la relación entre X e Y.

It is highly unlikely, however, that the lines drawn by any two people would be exactly the same. In other words, for every person drawing such a line by eye, or freehand, we would expect a slightly different line. The question then arises as to which line best describes the relationship between the two variables. We cannot obtain an answer to this question by inspecting the lines. In fact, it is not likely that any freehand line drawn through the data will be the line that best describes the relationship between X and Y, since freehand lines will reflect any defects of vision or judgment of the person drawing the line. Similarly, when judging which of two lines best describes the relationship, subjective evaluation is liable to the same deficiencies.

Sin embargo, es muy poco probable que las líneas trazadas por dos personas cualesquiera sean exactamente las mismas. En otras palabras, por cada persona que dibuje una línea de este tipo a ojo o a mano alzada, esperaríamos una línea ligeramente diferente. Entonces surge la pregunta de qué línea describe mejor la relación entre las dos variables. No podemos obtener una respuesta a esta pregunta inspeccionando las líneas. De hecho, no es probable que cualquier línea dibujada a mano alzada a través de los datos sea la línea que mejor describa la relación entre X e Y, ya que las líneas a mano alzada reflejarán cualquier defecto de visión o juicio de la persona que dibuja la línea. De manera similar, al juzgar cuál de dos líneas describe mejor la relación, la evaluación subjetiva está sujeta a las mismas deficiencias.

What is needed for obtaining the desired line is some method that is not fraught with these difficulties.

Lo que se necesita para obtener la línea deseada es algún método que no presente estas dificultades.

**The Least-Squares Line**

The method usually employed for obtaining the desired line is known as the method of least squares, and the resulting line is called the least-squares line. The reason for calling the method by this name will be explained in the discussion that follows.

El método empleado habitualmente para obtener la recta deseada se conoce como **método de mínimos cuadrados,** y la recta resultante se denomina recta de mínimos cuadrados. La razón para llamar al método con este nombre se explicará en la discusión siguiente.

We recall from algebra that the general equation for a straight line may be written as

y = a + bx

(9.3.1)

where y is a value on the vertical axis, x is a value on the horizontal axis, a is the point where the line crosses the vertical axis, and b shows the amount by which y changes for each unit change in x. We refer to a as the y-intercept and b as the slope of the line. To draw a line based on Equation 9.3.1, we need the numerical values of the constants a and b. Given these constants, we may substitute various values of x into the equation to obtain corresponding values of y. The resulting points may be plotted. Since any two such coordinates determine a straight line, we may select any two, locate them on a graph, and connect them to obtain the line corresponding to the equation.

donde y es un valor en el eje vertical, x es un valor en el eje horizontal, a es el punto donde la línea cruza el eje vertical y **b muestra la cantidad en la que y cambia por cada cambio de unidad en x.** Nos referimos a a como la intersección con el eje y y b como la pendiente de la recta. Para trazar una línea basada en la Ecuación 9.3.1, necesitamos los valores numéricos de las constantes a y b. Dadas estas constantes, podemos sustituir varios valores de x en la ecuación para obtener los valores correspondientes de y. Los puntos resultantes se pueden trazar. Dado que dos de estas coordenadas determinan una línea recta, podemos seleccionar dos, ubicarlas en una gráfica y conectarlas para obtener la línea correspondiente a la ecuación.

**Obtaining the Least-Square Line**

The least-squares regression line equation may be obtained from sample data by simple arithmetic calculations that may be carried out by hand using the following equations

La ecuación de la línea de regresión de mínimos cuadrados se puede obtener a partir de datos de muestra mediante cálculos aritméticos simples que se pueden realizar a mano usando las siguientes ecuaciones

a 1x i - x21y1 - y2

i=1

N

b1 =

n

2

a 1x i - x2

n

(9.3.2)

i=1

bN 0 = y - bN 1x

(9.3.3)

where xi and yi are the corresponding values of each data point (X, Y ), x and y are the means of the X and Y sample data values, respectively, and bN 0 and bN 1 are the estimates of the intercept b 0 and slope b 1, respectively, of the population regression line. Since the necessary hand calculations are time consuming, tedious, and subject to error, the regression line equation is best obtained through the use of a computer software package. Although the typical researcher need not be concerned with the arithmetic involved, the interested reader will find them discussed in references listed at the end of this chapter.

donde xi y yi son los valores correspondientes de cada punto de datos (X, Y), x e y son las medias de los valores de datos muestrales de X e Y, respectivamente, y bN 0 y bN 1 son las estimaciones de la intersección b 0 y pendiente b 1, respectivamente, de la recta de regresión poblacional. Dado que los cálculos manuales necesarios requieren mucho tiempo, son tediosos y están sujetos a errores, la ecuación de la línea de regresión se obtiene mejor mediante el uso de un paquete de software de computadora. Aunque el investigador típico no necesita preocuparse por la aritmética involucrada, el lector interesado la encontrará analizada en las referencias que se enumeran al final de este capítulo.

For the data in Table 9.3.1 we obtain the least-squares regression equation by means of MINITAB. After entering the X values in Column 1 and the Y values in Column 2 we proceed as shown in Figure 9.3.2.

For now, the only information from the output in Figure 9.3.2 that we are interested in is the regression equation. Other information in the output will be discussed later.

From Figure 9.3.2 we see that the linear equation for the least-squares line that describes the relationship between waist circumference and deep abdominal AT may be written, then, as

En la Figura 9.3.2 vemos que la ecuación lineal para la línea de mínimos cuadrados que describe la relación entre la circunferencia de la cintura y la AT abdominal profunda se puede escribir, entonces, como

yN = -216 + 3.46x

This equation tells us that since bN 0 is negative, the line crosses the Y-axis below the origin, and that since bN 1 the slope, is positive, the line extends from the lower left-hand corner of the graph to the upper right-hand corner. We see further that for each unit increase in x, y increases by an amount equal to 3.46. The symbol y denotes a value of y computed from the equation, rather than an observed value of Y.

Esta ecuación nos dice que como bN 0 es negativo, la línea cruza el eje Y debajo del origen, y que como bN 1 la pendiente es positiva, la línea se extiende desde la esquina inferior izquierda del gráfico hasta la esquina superior derecha. -esquina de la mano. Vemos además que por cada aumento unitario en x, y aumenta en una cantidad igual a 3,46. El símbolo y denota un valor de y calculado a partir de la ecuación, en lugar de un valor observado de Y.

By substituting two convenient values of X into Equation 9.3.2, we may obtain the necessary coordinates for drawing the line. Suppose, first, we let X = 70 and obtain

Sustituyendo dos valores convenientes de X en la ecuación 9.3.2, podemos obtener las coordenadas necesarias para trazar la línea. Supongamos, primero, que hacemos X = 70 y obtenemos

yN = -216 + 3.461702 = 26.2

If we let X = 110 we obtain

yN = -216 + 3.4611102 = 164

The line, along with the original data, is shown in Figure 9.3.3.

**The Least-Squares Criterion**

Now that we have obtained what we call the “best fit” line for describing the relationship between our two variables, we need to determine by what criterion it is considered best. Before the criterion is stated, let us examine Figure 9.3.3. We note that generally the least-squares line does not pass through the observed points that are plotted on the scatter diagram. In other words, most of the observed points deviate from the line by varying amounts.

Ahora que hemos obtenido lo que llamamos la línea de "mejor ajuste" **para describir la relación entre nuestras dos variables,** necesitamos determinar con qué criterio se considera mejor. Antes de establecer el criterio, examinemos la Figura 9.3.3. Observamos que generalmente la línea de mínimos cuadrados no pasa por los puntos observados que se trazan en el diagrama de dispersión. En otras palabras, la mayoría de los puntos observados se desvían de la línea en cantidades variables.

The line that we have drawn through the points is best in this sense:

**La línea que hemos trazado a través de los puntos es mejor en este sentido:**

The sum of the squared vertical deviations of the observed data points (yi) from the least-squares line is smaller than the sum of the squared vertical deviations of the data points from any other line.

La suma de las desviaciones verticales al cuadrado de los puntos de datos observados (yi) desde la línea de mínimos cuadrados es menor que la suma de las desviaciones verticales al cuadrado de los puntos de datos desde cualquier otra línea.

In other words, if we square the vertical distance from each observed point ( yi) to the least-squares line and add these squared values for all points, the resulting total will be smaller than the similarly computed total for any other line that can be drawn through the points. For this reason the line we have drawn is called the least-squares line.

En otras palabras, si elevamos al cuadrado la distancia vertical desde cada punto observado (yi) a la línea de mínimos cuadrados y sumamos estos valores al cuadrado para todos los puntos, el total resultante será menor que el total calculado de manera similar para cualquier otra línea que pueda ser dibujado a través de los puntos. Por esta razón la línea que hemos dibujado se llama línea de mínimos cuadrados.

**9.4 EVALUATING THE REGRESSION EQUATION**

Once the regression equation has been obtained it must be evaluated to determine whether it adequately describes the relationship between the two variables and whether it can be used effectively for prediction and estimation purposes.

**Una vez obtenida la ecuación de regresión, se debe evaluar para determinar si describe adecuadamente la relación entre las dos variables y si se puede utilizar de manera efectiva con fines de predicción y estimación.**

**When H0 : B1 =0 Is Not Rejected**

If in the population the relationship between X and Y is linear, b 1, the slope of the line that describes this relationship, will be either positive, negative, or zero. If b 1 is zero, sample data drawn from the population will, in the long run, yield regression equations that are of little or no value for prediction and estimation purposes. Furthermore, even though we assume that the relationship between X and Y is linear, it may be that the relationship could be described better by some nonlinear model. When this is the case, sample data when fitted to a linear model will tend to yield results compatible with a population slope of zero.

Si en la población la relación entre X e Y es lineal, b 1, la pendiente de la recta que describe esta relación, será positiva, negativa o cero. Si b 1 es cero, los datos muestrales extraídos de la población producirán, a largo plazo, ecuaciones de regresión que tienen poco o ningún valor para fines de predicción y estimación. Además, aunque suponemos que la relación entre X e Y es lineal, es posible que la relación pueda describirse mejor mediante algún modelo no lineal. Cuando este es el caso, los datos muestrales, cuando se ajustan a un modelo lineal, tenderán a producir resultados compatibles con una pendiente poblacional de cero.

Thus, following a test in which the null hypothesis that b 1 equals zero is not rejected, we may conclude (assuming that we have not made a type II error by accepting a false null hypothesis) either (1) that although the relationship between X and Y may be linear it is not strong enough for X to be of much value in predicting and estimating Y, or (2) that the relationship between X and Y is not linear; that is, some curvilinear model provides a better fit to the data. Figure 9.4.1 shows the kinds of relationships between X and Y in a population that may prevent rejection of the null hypothesis that b1 = 0.

Por lo tanto, después de una prueba en la que no se rechaza la hipótesis nula de que b 1 es igual a cero, podemos concluir (asumiendo que no hemos cometido un error de tipo II al aceptar una hipótesis nula falsa) ya sea (1) que aunque la relación entre X e Y puede ser lineal, no es lo suficientemente fuerte como para que X sea de mucho valor para predecir y estimar Y, o (2) que la relación entre X e Y no es lineal; es decir, algún modelo curvilíneo proporciona un mejor ajuste a los datos. La figura 9.4.1 muestra los tipos de relaciones entre X e Y en una población que pueden evitar el rechazo de la hipótesis nula de que b1 = 0.

**When H0 : B1 = 0 Is Rejected**

Now let us consider the situations in a population that may lead to rejection of the null hypothesis that b 1 = 0. Assuming that we do not commit a type I error, rejection of the null hypothesis that b 1 = 0 may be attributed to one of the following conditions in the population: (1) the relationship is linear and of sufficient strength to justify the use of sample regression equations to predict and estimate Y for given values of X; and (2) there is a good fit of the data to a linear model, but some curvilinear model might provide an even better fit. Figure 9.4.2 illustrates the two population conditions that may lead to rejection of H0 : b 1 = 0.

Consideremos ahora las situaciones en una población que pueden llevar al rechazo de la hipótesis nula de que b 1 = 0. Suponiendo que no cometemos un error tipo I, el rechazo de la hipótesis nula de que b 1 = 0 puede atribuirse a una de las siguientes condiciones en la población: (1) la relación es lineal y de fuerza suficiente para justificar el uso de ecuaciones de regresión muestral para predecir y estimar Y para valores dados de X; y (2) hay un buen ajuste de los datos a un modelo lineal, pero algún modelo curvilíneo podría proporcionar un ajuste aún mejor. La figura 9.4.2 ilustra las dos condiciones poblacionales que pueden llevar al rechazo de H0 : b 1 = 0.

Thus, we see that before using a sample regression equation to predict and estimate, it is desirable to test H0 : b 1 = 0. We may do this either by using analysis of variance and the F statistic or by using the t statistic. We will illustrate both methods. Before we do this, however, let us see how we may investigate the strength of the relationship between X and Y.

Por lo tanto, vemos que antes de usar una ecuación de regresión muestral para predecir y estimar, es deseable probar H0 : b 1 = 0. Esto se puede hacer usando el análisis de varianza y el estadístico F o usando el estadístico t. Ilustraremos ambos métodos. Sin embargo, antes de hacer esto, veamos cómo podemos investigar la **fuerza de la relación entre X e Y.**

**The Coefficient of Determination**

One way to evaluate the strength of the regression equation is to compare the scatter of the points about the regression line with the scatter about y, the mean of the sample values of Y. If we take the scatter diagram for Example 9.3.1 and draw through the points a line that intersects the Y-axis at y and is parallel to the X-axis, we may obtain a visual impression of the relative magnitudes of the scatter of the points about this line and the regression line. This has been done in Figure 9.4.3.424

**Una forma de evaluar la fuerza de la ecuación de regresión es comparar la dispersión de los puntos alrededor de la línea de regresión con la dispersión alrededor de y, la media de los valores muestrales de Y**. Si tomamos el diagrama de dispersión del ejemplo 9.3.1 y dibujamos a través de los puntos una línea que corta el eje Y en y y es paralela al eje X, podemos obtener una impresión visual de las magnitudes relativas de la dispersión de los puntos alrededor de esta línea y la línea de regresión. Esto se ha hecho en la Figura 9.4.

FIGURE 9.4.1 Conditions in a population that may prevent rejection of the null hypothesis that b 1 = 0. (a) The relationship between X and Y is linear, but b 1 is so close to zero that sample data are not likely to yield equations that are useful for predicting Y when X is given. (b) The relationship between X and Y is not linear; a curvilinear model provides a better fit to the data; sample data are not likely to yield equations that are useful for predicting Y when X is given.

Condiciones en una población que pueden impedir el rechazo de la hipótesis nula de que b 1 = 0. (a) La relación entre X e Y es lineal, pero b 1 es tan cercano a cero que no es probable que los datos muestrales produzcan ecuaciones que sean útiles para predecir Y cuando se da X. (b) La relación entre X e Y no es lineal; un modelo curvilíneo proporciona un mejor ajuste a los datos; No es probable que los datos muestrales produzcan ecuaciones que sean útiles para predecir Y cuando se da X.

It appears rather obvious from Figure 9.4.3 that the scatter of the points about the regression line is much less than the scatter about the y line. We would not wish, however, to decide on this basis alone that the equation is a useful one. The situation may not be always this clear-cut, so that an objective measure of some sort would be much more desirable. Such an objective measure, called the coefficient of determination, is available.

Parece bastante obvio en la Figura 9.4.3 que la dispersión de los puntos alrededor de la línea de regresión es mucho menor que la dispersión alrededor de la línea y. Sin embargo, no deseamos decidir únicamente sobre esta base si la ecuación es útil. La situación puede no ser siempre tan clara, por lo que sería mucho más deseable una medida objetiva de algún tipo. **Existe una medida objetiva de este tipo, denominada coeficiente de determinación.**

**The Total Deviation**

Before defining the coefficient of determination, let us justify its use by examining the logic behind its computation. We begin by considering the point corresponding to any observed value, yi, and by measuring its vertical distance from the y line. We call this the total deviation and designate it 1yi - y2.

Antes de definir el coeficiente de determinación, justifiquemos su uso examinando la lógica detrás de su cálculo. Comenzamos considerando el punto correspondiente a cualquier valor observado, yi, y midiendo su distancia vertical desde la recta y. A esto lo llamamos desviación total y lo designamos 1yi - y2.

**The Explained Deviation**

If we measure the vertical distance from the regression line to the y line, we obtain 1yN i - y2, which is called the explained deviation, since it shows by how much the total deviation is reduced when the regression line is fitted to the points.

Si medimos la distancia vertical desde la recta de regresión hasta la recta y, obtenemos 1yN i - y2, que se llama desviación explicada, ya que muestra en cuánto se reduce la desviación total cuando se ajusta la recta de regresión a los puntos.

FIGURE 9.4.2 Population conditions relative to X and Y that may cause rejection of the null hypothesis that b 1 = 0. (a) The relationship between X and Y is linear and of sufficient strength to justify the use of a sample regression equation to predict and estimate Y for given values of X. (b) A linear model provides a good fit to the data, but some curvilinear model would provide an even better fit.

Condiciones de la población relativas a X e Y que pueden causar el rechazo de la hipótesis nula de que b 1 = 0. (a) La relación entre X e Y es lineal y de fuerza suficiente para justificar el uso de una ecuación de regresión muestral para predecir y estimar Y. para valores dados de X. (b) Un modelo lineal proporciona un buen ajuste a los datos, pero algún modelo curvilíneo proporcionaría un ajuste aún mejor.

**Unexplained Deviation**

Finally, we measure the vertical distance of the observed point from the regression line to obtain 1yi - yN i2, which is called the unexplained deviation, since it represents the portion of the total deviation not “explained” or accounted for by the introduction of the regression line. These three quantities are shown for a typical value of Y in Figure 9.4.4. The difference between the observed value of Y and the predicted value of Y, 1yi - yN i2, is also referred to as a residual. The set of residuals can be used to test the underlying linearity and equal-variances assumptions of the regression model described in Section 9.2. This procedure is illustrated at the end of this section.

Finalmente, medimos la distancia vertical del punto observado desde la línea de regresión para obtener 1yi - yN i2, que se llama desviación inexplicable, ya que representa la porción de la desviación total no “explicada” o contabilizada por la introducción del línea de regresión. Estas tres cantidades se muestran para un valor típico de Y en la Figura 9.4.4. La diferencia entre el valor observado de Y y el valor predicho de Y, 1yi - yN i2, también se denomina residual. El conjunto de residuos se puede utilizar para probar los supuestos subyacentes de linealidad y varianzas iguales del modelo de regresión descrito en la sección 9.2. Este procedimiento se ilustra al final de esta sección.

It is seen, then, that the total deviation for a particular yi is equal to the sum of the explained and unexplained deviations. We may write this symbolically as

Se ve, entonces, que la desviación total para un yi particular es igual a la suma de las desviaciones explicadas y no explicadas. Podemos escribir esto simbólicamente como

1yi - y2 = 1yN i - y2 + 1yi - yN i2

total

deviation

explained unexplained

deviation deviation

FIGURE 9.4.3 Scatter diagram, sample regression line, and y line for Example 9.3.1.

Diagrama de dispersión, línea de regresión muestral y línea y para el ejemplo 9.3.1.

If we measure these deviations for each value of yi and yNi , square each deviation, and add up the squared deviations, we have

Si medimos estas desviaciones para cada valor de yi e yNi, elevamos al cuadrado cada desviación y sumamos las desviaciones al cuadrado, tenemos

g 1yi - y22 = g 1yN i - y22 + g 1yi - yN i22

total

sum

of squares

explained

sum

of squares

(9.4.2)

unexplained

sum

of squares

These quantities may be considered measures of dispersion or variability.

**Total Sum of Squares**

The total sum of squares (SST), for example, is a measure of the dispersion of the observed values of Y about their mean y; that is, this term is a measure of the total variation in the observed values of Y. The reader will recognize this term as the numerator of the familiar formula for the sample variance.

La suma total de cuadrados (SST), por ejemplo, es una medida de la dispersión de los valores observados de Y respecto de su media y; es decir, este término es una medida de la variación total en los valores observados de Y. El lector reconocerá este término como el numerador de la fórmula familiar para la varianza muestral.

**Explained Sum of Squares**

The explained sum of squares measures the amount of the total variability in the observed values of Y that is accounted for by the linear relationship between the observed values of X and Y. This quantity is referred to also as the sum of squares due to linear regression (SSR).

La suma de cuadrados explicada mide la cantidad de variabilidad total en los valores observados de Y que se explica por la relación lineal entre los valores observados de X e Y. Esta cantidad también se conoce como suma de cuadrados debido a la regresión lineal. (RSS).

FIGURE 9.4.4 Scatter diagram showing the total, explained, and unexplained deviations for a selected value of Y, Example 9.3.1.

Diagrama de dispersión que muestra las desviaciones totales, explicadas y no explicadas para un valor seleccionado de Y, Ejemplo 9.3.1.

**Unexplained Sum of Squares**

The unexplained sum of squares is a measure of the dispersion of the observed Y values about the regression line and is sometimes called the error sum of squares, or the residual sum of squares (SSE). It is this quantity that is minimized when the least-squares line is obtained.

La suma de cuadrados inexplicada es una medida de la dispersión de los valores de Y observados alrededor de la línea de regresión y, a veces, se denomina suma de cuadrados de error o suma de cuadrados residual (SSE). Es esta cantidad la que se minimiza cuando se obtiene la línea de mínimos cuadrados.

We may express the relationship among the three sums of squares values as

Podemos expresar la relación entre las tres sumas de valores de cuadrados como

SST = SSR + SSE

The numerical values of these sums of squares for our illustrative example appear in the analysis of variance table in Figure 9.3.2. Thus, we see that SST = 354531, SSR = 237549, SSE = 116982, and

354531 = 237549 + 116982

354531 = 354531

**Calculating r2**

It is intuitively appealing to speculate that if a regression equation does a good job of describing the relationship between two variables, the explained or regression sum of squares should constitute a large proportion of the total sum of squares. It would be of interest, then, to determine the magnitude of this proportion by computing the ratio of the explained sum of squares to the total sum of squares. This is exactly what is done in evaluating a regression equation based on sample data, and the result is called the sample coefficient of determination, r 2. That is,

Es intuitivamente atractivo especular que si una ecuación de regresión describe bien la relación entre dos variables, la suma de cuadrados explicada o de regresión debería constituir una gran proporción de la suma total de cuadrados. Sería interesante, entonces, determinar la magnitud de esta proporción calculando la relación entre la suma de cuadrados explicada y la suma total de cuadrados. Esto es exactamente lo que se hace al evaluar una ecuación de regresión basada en datos de muestra, y el resultado se llama **coeficiente de determinación de muestra**, r 2. Es decir,

r2 =

g1yN i - y22

g1yi - y 2

2

=

SSR

SST

In our present example we have, using the sums of squares values from Figure 9.3.2,

En nuestro ejemplo actual tenemos, usando las sumas de valores de cuadrados de la Figura 9.3.2,

r2 =

237549

= .67

354531

The sample coefficient of determination measures the closeness of fit of the sample regression equation to the observed values of Y. When the quantities 1yi - yN i2, the vertical distances of the observed values of Y from the equations, are small, the unexplained sum of squares is small. This leads to a large explained sum of squares that leads, in turn, to a large value of r 2. This is illustrated in Figure 9.4.5.

El coeficiente de determinación muestral mide la cercanía del ajuste de la ecuación de regresión muestral a los valores observados de Y. Cuando las cantidades 1yi - yN i2, las distancias verticales de los valores observados de Y de las ecuaciones, son pequeñas, la suma inexplicable de los cuadrados son pequeños. Esto conduce a una suma grande de cuadrados explicada que conduce, a su vez, a un valor grande de r 2. Esto se ilustra en la Figura 9.4.5.

In Figure 9.4.5(a) we see that the observations all lie close to the regression line, and we would expect r 2 to be large. In fact, the computed r 2 for these data is .986, indicating that about 99 percent of the total variation in the yi is explained by the regression.

En la figura 9.4.5(a) vemos que todas las observaciones se encuentran cerca de la línea de regresión y esperaríamos que r 2 fuera grande. De hecho, el r 2 calculado para estos datos es 0,986, lo que indica que alrededor del 99 por ciento de la variación total en el yi se explica por la regresión.

In Figure 9.4.5(b) we illustrate a case in which the yi are widely scattered about the regression line, and there we suspect that r 2 is small. The computed r 2 for the data is .403; that is, less than 50 percent of the total variation in the yi is explained by the regression.

En la figura 9.4.5(b) ilustramos un caso en el que los yi están muy dispersos alrededor de la línea de regresión y allí sospechamos que r 2 es pequeño. El r 2 calculado para los datos es 0,403; es decir, menos del 50 por ciento de la variación total del yi se explica por la regresión.

The largest value that r 2 can assume is 1, a result that occurs when all the variation in the yi is explained by the regression. When r 2 = 1 all the observations fall on the regression line. This situation is shown in Figure 9.4.5(c).

El valor más grande que r 2 puede asumir es 1, resultado que ocurre cuando toda la variación en yi se explica por la regresión. Cuando r 2 = 1 todas las observaciones caen en la recta de regresión. Esta situación se muestra en la Figura 9.4.5(c).

The lower limit of r 2 is 0. This result is obtained when the regression line and the line drawn through y coincide. In this situation none of the variation in the yi is explained by the regression. Figure 9.4.5(d) illustrates a situation in which r 2 is close to zero.

El límite inferior de r 2 es 0. Este resultado se obtiene cuando la recta de regresión y la recta trazada por y coinciden. En esta situación, la regresión no explica ninguna variación del yi. La figura 9.4.5(d) ilustra una situación en la que r 2 es cercano a cero.

When r 2 is large, then, the regression has accounted for a large proportion of the total variability in the observed values of Y, and we look with favor on the regression equation. On the other hand, a small r 2 which indicates a failure of the regression to account for a large proportion of the total variation in the observed values of Y, tends to cast doubt on the usefulness of the regression equation for predicting and estimating purposes. We do not, however, pass final judgment on the equation until it has been subjected to an objective statistical test.

Cuando r 2 es grande, entonces, la regresión ha explicado una gran proporción de la variabilidad total en los valores observados de Y, y miramos con buenos ojos la ecuación de regresión. Por otro lado, un r 2 pequeño que indica que la regresión no explica una gran proporción de la variación total en los valores observados de Y, tiende a poner en duda la utilidad de la ecuación de regresión para propósitos de predicción y estimación. Sin embargo, no emitiremos un juicio final sobre la ecuación hasta que haya sido sometida a una prueba estadística objetiva.

**Testing H0 : B1 = 0 with the F Statistic**

The following example illustrates one method for reaching a conclusion regarding the relationship between X and Y.

El siguiente ejemplo ilustra un método para llegar a una conclusión sobre la relación entre X e Y.

**EXAMPLE 9.4.1**

Refer to Example 9.3.1. We wish to know if we can conclude that, in the population from which our sample was drawn, X and Y are linearly related.

Consulte el Ejemplo 9.3.1. Deseamos saber si podemos concluir que, en la población de la que se extrajo nuestra muestra, X e Y están relacionados linealmente.

FIGURE 9.4.5 r 2 as a measure of closeness-of-fit of the sample regression line to the sample observations.

**Solution:**

The steps in the hypothesis testing procedure are as follows:

1. Data. The data were described in the opening statement of Example 9.3.1.

2. Assumptions. We presume that the simple linear regression model and its underlying assumptions as given in Section 9.2 are applicable.

3. Hypotheses.

H0 : b 1 = 0

HA : b 1 Z 0

a = .05430

CHAPTER 9 SIMPLE LINEAR REGRESSION AND CORRELATION

TABLE 9.4.1 ANOVA Table for Simple Linear Regression

Source of

Variation

SSd.f.MSV.R.

SSR

SSE1

n&2MSR " SSR/ 1

MSE " SSE/ (n & 2)MSR/MSE

ResidualTotalSSTn&1

Linear regression

4. **Test** **statistic**. The test statistic is V.R. as explained in the discussion that follows.

From the three sums-of-squares terms and their associated degrees of freedom the analysis of variance table of Table 9.4.1 may be constructed.

In general, the degrees of freedom associated with the sum of squares due to regression is equal to the number of constants in the regression equation minus 1. In the simple linear case we have two estimates, b0 and b1; hence the degrees of freedom for regression are 2 - 1 = 1.

5. **Distribution** **of** **test** **statistic**. It can be shown that when the hypothesis of no linear relationship between X and Y is true, and when the assumptions underlying regression are met, the ratio obtained by dividing the regression mean square by the residual mean square is distributed as F with 1 and n - 2 degrees of freedom.

6. **Decision** **rule**. Reject H0 if the computed value of V.R. is equal to or greater than the critical value of F.

7. **Calculation** **of** **test** **statistic**. As shown in Figure 9.3.2, the computed value of F is 217.28.

8. **Statistical** **decision**. Since 217.28 is greater than 3.94, the critical value of F (obtained by interpolation) for 1 and 107 degrees of freedom, the null hypothesis is rejected.

9. **Conclusion**. We conclude that the linear model provides a good fit to the data.

10. **p** **value**. For this test, since 217.28 7 8.25, we have p 6 .005.

**Estimating the Population Coefficient of Determination**

The sample coefficient of determination provides a point estimate of r2 the population coefficient of determination. The population coefficient of determination, r2 has the same function relative to the population as r 2 has to the sample. It shows what proportion of the total population variation in Y is explained by the regression of Y on X. When the number of degrees of freedom is small, r 2 is positively biased. That is, r 2 tends to be large. An unbiased estimator of r2 is provided by

El coeficiente de determinación muestral proporciona una estimación puntual de r2, el coeficiente de determinación poblacional. El coeficiente de determinación poblacional, r2, tiene la misma función con respecto a la población que r2 con respecto a la muestra. Muestra qué proporción de la variación poblacional total en Y se explica por la regresión de Y sobre X. Cuando el número de grados de libertad es pequeño, r 2 tiene un sesgo positivo. Es decir, r 2 tiende a ser grande. Un es

~2

r = 1 -

g1 yi - yN i22>1n - 22

g1 yi - y22>1n - 12

(9.4.3)9.4 EVALUATING THE REGRESSION EQUATION

431

Observe that the numerator of the fraction in Equation 9.4.3 is the unexplained mean square and the denominator is the total mean square. These quantities appear in the analysis of variance table. For our illustrative example we have, using the data from Figure 9.3.2,

Observe que el numerador de la fracción en la Ecuación 9.4.3 es el cuadrado medio inexplicable y el denominador es el cuadrado medio total. Estas cantidades aparecen en la tabla de análisis de varianza. Para nuestro ejemplo ilustrativo tenemos, utilizando los datos de la Figura 9.3.2,

r~ 2 = 1 -

116982>107

354531>108

= .66695

This quantity is labeled R-sq(adj) in Figure 9.3.2 and is reported as 66.7 percent. We see that this value is less than

r2 = 1 -

116982

= .67004

354531

We see that the difference in r 2 and r~2 is due to the factor 1n - 12>1n - 22. When n is large, this factor will approach 1 and the difference between r 2 and r~2 will approach zero.

Vemos que la diferencia en r 2 y r~2 se debe al factor 1n - 12>1n - 22. Cuando n es grande, este factor se acercará a 1 y la diferencia entre r 2 y r~2 se acercará a cero.

**Testing H0 : B1 = 0 with the t Statistic**

When the assumptions stated in Section 9.2 are met, bN 0 and bN 1 are unbiased point estimators of the corresponding parameters b 0 and b 1. Since, under these assumptions, the subpopulations of Y values are normally distributed, we may construct confidence intervals for and test hypotheses about b 0 and b 1. When the assumptions of Section 9.2 hold true, the sampling distributions of bN 0 and bN 1 are each normally distributed with means and variances as follows:

Cuando se cumplen los supuestos establecidos en la sección 9.2, bN 0 y bN 1 son estimadores puntuales insesgados de los parámetros correspondientes b 0 y b 1. Dado que, bajo estos supuestos, las subpoblaciones de valores de Y se distribuyen normalmente, podemos construir intervalos de confianza para y probar hipótesis sobre b 0 y b 1. Cuando los supuestos de la sección 9.2 son ciertos, las distribuciones muestrales de bN 0 y bN 1 se distribuyen normalmente con medias y varianzas de la siguiente manera:

mbN 0 = b 0(9.4.4)

s2bN 0 =(9.4.5)

s2y>x g x i2

ng1x i - x22

mbN 1 = b 1

(9.4.6)

and

sb2N 1

=

s2y>x

g1x i - x22

(9.4.7)

In Equations 9.4.5 and 9.4.7 s2y>x is the unexplained variance of the subpopulations of Y values.

En las ecuaciones 9.4.5 y 9.4.7, s2y>x es la varianza no explicada de las subpoblaciones de valores de Y.

With knowledge of the sampling distributions of bN 0 and bN 1 we may construct confidence intervals and test hypotheses relative to b 0 and b 1 in the usual manner. Inferences regarding a are usually not of interest. On the other hand, as we have seen, a great deal of interest centers on inferential procedures with respect to bN 1. The reason for this is the fact that b 1 tells us so much about the form of the relationship between X and Y. When X and Y are linearly related a positive bN 1 indicates that, in general, Y increases as X increases, and we say that there is a direct linear relationship between X and Y. A negative bN 1 indicates that values of Y tend to decrease as values of X increase, and we say that there is an inverse linear relationship between X and Y.

Con conocimiento de las distribuciones muestrales de bN 0 y bN 1 podemos construir intervalos de confianza y probar hipótesis relativas a b 0 y b 1 de la manera habitual. Las inferencias sobre a no suelen ser de interés. Por otro lado, como hemos visto, gran parte del interés se centra en los procedimientos inferenciales con respecto a bN 1. La razón de esto es el hecho de que b 1 nos dice mucho sobre la forma de la relación entre X e Y. Cuando X e Y están relacionados linealmente, un bN 1 positivo indica que, en general, Y aumenta a medida que aumenta X, y decimos que existe una relación lineal directa entre X e Y. Un bN 1 negativo indica que los valores de Y tienden a disminuir. a medida que los valores de X aumentan, decimos que existe una relación lineal inversa entre X e Y.

FIGURE 9.4.6 Scatter diagrams showing (a) direct linear relationship, (b) inverse linear relationship, and (c) no linear relationship between X and Y.

Diagramas de dispersión que muestran (a) una relación lineal directa, (b) una relación lineal inversa y (c) ninguna relación lineal entre X e Y.

When there is no linear relationship between X and Y, bN 1 is equal to zero. These three situations are illustrated in Figure 9.4.6.

**The Test Statistic**

For testing hypotheses about b 1 the test statistic when s2y>x is known is

z =

bN 1 - 1b 120

sbN 1

(9.4.8)

where 1b 120 is the hypothesized value of b 1. The hypothesized value of b 1 does not have to be zero, but in practice, more often than not, the null hypothesis of interest is that b 1 = 0.

As a rule s2y|x is unknown. When this is the case, the test statistic is

t =

bN 1 - 1b 120

sbN1

(9.4.9)

where sbN 1 is an estimate of sbN 1 and t is distributed as Student’s t with n - 2 degrees of freedom.

If the probability of observing a value as extreme as the value of the test statistic computed by Equation 9.4.9 when the null hypothesis is true is less than a>2 (since we have a two-sided test), the null hypothesis is rejected.

Si la probabilidad de observar un valor tan extremo como el valor del estadístico de prueba calculado por la Ecuación 9.4.9 cuando la hipótesis nula es verdadera es menor que a>2 (dado que tenemos una prueba bilateral), la hipótesis nula se rechaza .

**EXAMPLE 9.4.2**

Refer to Example 9.3.1. We wish to know if we can conclude that the slope of the population regression line describing the relationship between X and Y is zero.

Consulte el Ejemplo 9.3.1. Deseamos saber si podemos concluir que la pendiente de la recta de regresión poblacional que describe la relación entre X e Y es cero.

**Solution**:

1. **Data**. See Example 9.3.1.

2. **Assumptions**. We presume that the simple linear regression model and its underlying assumptions are applicable.

3. **Hypotheses**.

H0 : b 1 = 0

HA : b 1 Z 0

a = .05

4. **Test** **statistic**. The test statistic is given by Equation 9.4.9.

5. **Distribution** **of** **test** **statistic**. When the assumptions are met and H0 is true, the test statistic is distributed as Student’s t with n - 2 degrees of freedom.

6. **Decision** **rule**. Reject H0 if the computed value of t is either greater than or equal to 1.9826 or less than or equal to -1.9826.

7. **Calculation** **of** **statistic**. The output in Figure 9.3.2 shows that

bN 1 = 3.4589, sbN1 = .2347, and

t =

3.4589 - 0

= 14.74

.2347

8. **Statistical** **decision**. Reject H0 because 14.74 7 1.9826.

9. **Conclusion**. We conclude that the slope of the true regression line is not zero.

10. **p** **value**. The p value for this test is less than .01, since, when H0 is true, the probability of getting a value of t as large as or larger than 2.6230 (obtained by interpolation) is .005, and the probability of getting a value of t as small as or smaller than -2.6230 is also .005. Since 14.74 is greater than 2.6230, the probability of observing a value of t as large as or larger than 14.74 (when the null hypothesis is true) is less than .005. We double this value to obtain 21.0052 = .01.

Either the F statistic or the t statistic may be used for testing H0: b 1 = 0. The value of the variance ratio is equal to the square of the value of the t statistic 1i.e., t 2 = F 2 and, therefore, both statistics lead to the same conclusion. For the current example, we see that 114.7422 = 217.27, the value obtained by using the F statistic in Example 9.4.1.

The practical implication of our results is that we can expect to get better predictions and estimates of Y if we use the sample regression equation than we would get if we ignore the relationship between X and Y. The fact that b is positive leads us to believe that b 1 is positive and that the relationship between X and Y is a direct linear relationship. ■

As has already been pointed out, Equation 9.4.9 may be used to test the null hypothesis that b 1 is equal to some value other than 0. The hypothesized value for b 1, 1b 120 is substituted into Equation 9.4.9. All other quantities, as well as the computations, are the same as in the illustrative example. The degrees of freedom and the method of determining significance are also the same.

Como ya se ha señalado, la ecuación 9.4.9 se puede utilizar para probar la hipótesis nula de que b 1 es igual a algún valor distinto de 0. El valor hipotético de b 1, 1b 120 se sustituye en la ecuación 9.4.9. Todas las demás cantidades, así como los cálculos, son los mismos que en el ejemplo ilustrativo. Los grados de libertad y el método para determinar la importancia también son los mismos.

**A Confidence Interval for B 1**

Once we determine that it is unlikely, in light of sample evidence, that b 1 is zero, we may be interested in obtaining an interval estimate of b 1. The general formula for a confidence interval,

Una vez que determinamos que es poco probable, a la luz de la evidencia de la muestra, que b 1 sea cero, es posible que nos interese obtener una estimación de intervalo de b 1. La fórmula general para un intervalo de confianza,

estimator ; 1reliability factor21standard error of the estimate2

may be used. When obtaining a confidence interval for b 1, the estimator is bN 1, the reliability factor is some value of z or t (depending on whether or not s2y ƒx is known), and the standard error of the estimator is

puede ser usado. Al obtener un intervalo de confianza para b 1, el estimador es bN 1, el factor de confiabilidad es algún valor de z o t (dependiendo de si se conoce o no s2y ƒx), y el error estándar del estimador es

sbN 1 =

s2y ƒ x

C g1x i - x22

When s2y ƒx is unknown, sb is estimated by

sbN 1 =

s y2 ƒ x

C g1x i - x22

where s y2 ƒ x = MSE

In most practical situations our 10011 - a2 percent confidence interval for b is

bN 1 ; t 11-a> 22sbN 1

(9.4.10)

For our illustrative example we construct the following 95 percent confidence interval for b :

3.4589 ; 1.98261.23472

2.99, 3.92

We interpret this interval in the usual manner. From the probabilistic point of view we say that in repeated sampling 95 percent of the intervals constructed in this way will include b 1. The practical interpretation is that we are 95 percent confident that the single interval constructed includes b1.

Interpretamos este intervalo de la manera habitual. Desde el punto de vista probabilístico decimos que en el muestreo repetido el 95 por ciento de los intervalos construidos de esta manera incluirán b1. La interpretación práctica es que tenemos una confianza del 95 por ciento de que el intervalo único construido incluye b1.

**Using the Confidence Interval to Test H0: B1 = 0**

It is instructive to note that the confidence interval we constructed does not include zero, so that zero is not a candidate for the parameter being estimated. We feel, then, that it is unlikely that b1 = 0. This is compatible with the results of our hypothesis test in which we rejected the null hypothesis that b 1 = 0. Actually, we can always test H0: b 1 = 0 at the a significance level by constructing the 10011 - a2 percent confidence interval for b 1, and we can reject or fail to reject the hypothesis on the basis of whether or not the interval includes zero. If the interval contains zero, the null hypothesis is not rejected; and if zero is not contained in the interval, we reject the null hypothesis.

Es instructivo observar que el intervalo de confianza que construimos no incluye cero, por lo que cero no es candidato para el parámetro que se estima. Creemos, entonces, que es poco probable que b1 = 0. Esto es compatible con los resultados de nuestra prueba de hipótesis en la que rechazamos la hipótesis nula de que b 1 = 0. En realidad, siempre podemos probar H0: b 1 = 0 en Para obtener el nivel de significancia a se construye el intervalo de confianza por ciento 10011 - a2 para b 1, y podemos rechazar o no rechazar la hipótesis en función de si el intervalo incluye o no el cero. Si el intervalo contiene cero, la hipótesis nula no se rechaza; y si el cero no está contenido en el intervalo, rechazamos la hipótesis nula.

**Interpreting the Results**

It must be emphasized that failure to reject the null hypothesis that b 1 = 0 does not mean that X and Y are not related. Not only is it possible that a type II error may have been committed but it may be true that X and Y are related in some nonlinear manner. On the other hand, when we reject the null hypothesis that b 1 = 0, we cannot conclude that the true relationship between X and Y is linear. Again, it may be that although the data fit the linear regression model fairly well (as evidenced by the fact that the null hypothesis that b 1 = 0 is rejected), some nonlinear model would provide an even better fit. Consequently, when we reject H0 that b1 = 0, the best we can say is that more useful results (discussed below) may be obtained by taking into account the regression of Y on X than in ignoring it.

Debe enfatizarse que no rechazar la hipótesis nula de que b 1 = 0 no significa que X e Y no estén relacionados. No sólo es posible que se haya cometido un error de tipo II, sino que también puede ser cierto que X e Y estén relacionados de alguna manera no lineal. Por otro lado, cuando rechazamos la hipótesis nula de que b 1 = 0, no podemos concluir que la verdadera relación entre X e Y sea lineal. Nuevamente, puede ser que aunque los datos se ajusten bastante bien al modelo de regresión lineal (como lo demuestra el hecho de que se rechaza la hipótesis nula de que b 1 = 0), algún modelo no lineal proporcionaría un ajuste aún mejor. En consecuencia, cuando rechazamos H0 y b1 = 0, lo mejor que podemos decir es que se pueden obtener resultados más útiles (que se analizan más adelante) teniendo en cuenta la regresión de Y sobre X que ignorándola.

**Testing the Regression Assumptions**

The values of the set of residuals, 1yi - yN i2, for a data set are often used to test the linearity and equal-variances assumptions (assumptions 4 and 5 of Section 9.2) underlying the regression model. This is done by plotting the values of the residuals on the y-axis and the predicted values of y on the x-axis. If these plots show a relatively random scatter of points above and below a horizontal line at 1yi - yN i2 = 0, these assumptions are assumed to have been met for a given set of data. A non-random pattern of points can indicate violation of the linearity assumption, and a funnel-shaped pattern of the points can indicate violation of the equal-variances assumption. Examples of these patterns are shown in Figure 9.4.7.

Los valores del conjunto de residuos, 1yi - yN i2, para un conjunto de datos se utilizan a menudo para probar los supuestos de linealidad y varianzas iguales (supuestos 4 y 5 de la sección 9.2) subyacentes al modelo de regresión. Esto se hace trazando los valores de los residuos en el eje y y los valores predichos de y en el eje x. Si estos gráficos muestran una dispersión relativamente aleatoria de puntos por encima y por debajo de una línea horizontal en 1yi - yN i2 = 0, se supone que estos supuestos se han cumplido para un conjunto de datos dado. Un patrón de puntos no aleatorio puede indicar una violación del supuesto de linealidad, y un patrón de puntos en forma de embudo puede indicar una violación del supuesto de varianzas iguales. En la Figura 9.4.7 se muestran ejemplos de estos patrones.

Many computer packages will provide residual plots automatically. These plots often use standardized values 1i.e., ei> 1MSE2 of the residuals and predicted values, but are interpreted in the same way as are plots of unstandardized values.

Muchos paquetes informáticos proporcionarán gráficos residuales automáticamente. Estos gráficos suelen utilizar valores estandarizados 1es decir, ei> 1MSE2 de los residuos y valores predichos, pero se interpretan de la misma manera que los gráficos de valores no estandarizados.

**EXAMPLE 9.4.3**

Refer to Example 9.3.1. We wish to use residual plots to test the assumptions of linearity and equal variances in the data.

Consulte el Ejemplo 9.3.1. Deseamos utilizar gráficos de residuos para probar los supuestos de linealidad y varianzas iguales en los datos.

Solution:

A residual plot is shown in Figure 9.4.8.

Since there is a relatively equal and random scatter of points above and below the residual 1yi - yN i2 = 0 line, the linearity assumption is presumed to be valid. However, the funneling tendency of the plot suggests that as the predicted value of deep abdominal AT area increases, so does the amount of error. This indicates that the assumption of equal variances may not be valid for these data.

■

**9.5 USING THE REGRESSION EQUATION**

If the results of the evaluation of the sample regression equation indicate that there is a relationship between the two variables of interest, we can put the regression equation to practical use. There are two ways in which the equation can be used. It can be used to predict what value Y is likely to assume given a particular value of X. When the normality assumption of Section 9.2 is met, a prediction interval for this predicted value of Y may be constructed.

Si los resultados de la evaluación de la ecuación de regresión muestral indican que existe una relación entre las dos variables de interés, podemos darle un uso práctico a la ecuación de regresión. Hay dos formas en las que se puede utilizar la ecuación. Puede usarse para predecir qué valor es probable que asuma Y dado un valor particular de X. Cuando se cumple el supuesto de normalidad de la Sección 9.2, se puede construir un intervalo de predicción para este valor predicho de Y.

We may also use the regression equation to estimate the mean of the subpopulation of Y values assumed to exist at any particular value of X. Again, if the assumption of normally distributed populations holds, a confidence interval for this parameter may be constructed. The predicted value of Y and the point estimate of the mean of the subpopulation of Y will be numerically equivalent for any particular value of X but, as we will see, the prediction interval will be wider than the confidence interval.

También podemos usar la ecuación de regresión para estimar la media de la subpoblación de valores de Y que se supone existen en cualquier valor particular de X. Nuevamente, si se cumple el supuesto de poblaciones normalmente distribuidas, se puede construir un intervalo de confianza para este parámetro. El valor predicho de Y y la estimación puntual de la media de la subpoblación de Y serán numéricamente equivalentes para cualquier valor particular de X pero, como veremos, el intervalo de predicción será más amplio que el intervalo de confianza.

**Predicting Y for a Given X**

If it is known, or if we are willing to assume that the assumptions of Section 9.2 are met, and when s2y ƒ x is unknown, then the 10011 - a2 percent prediction interval for Y is given by

Si se sabe, o si estamos dispuestos a suponer que se cumplen los supuestos de la sección 9.2, y cuando s2y ƒ x es desconocido, entonces el intervalo de predicción de 10011 - a2 por ciento para Y está dado por

yN ; t 11-a> 22sy ƒ x

C

1 +

1x p - x22

1

+

n

g1x i - x22

(9.5.1)

where x p is the particular value of x at which we wish to obtain a prediction interval for Y and the degrees of freedom used in selecting t are n - 2.

donde x p es el valor particular de x en el que deseamos obtener un intervalo de predicción para Y y los grados de libertad utilizados para seleccionar t son n - 2.

**Estimating the Mean of Y for a Given X**

The 10011 - a2 percent confidence interval for my ƒ x , when s2y ƒ x is unknown, is given by

1x p - x22

1

yN ; t 11-a> 22sy ƒ x

+

Cn

g1x i - x22

(9.5.2)

We use MINITAB to illustrate, for a specified value of X, the calculation of a 95 percent confidence interval for the mean of Y and a 95 percent prediction interval for an individual Y measurement.

Usamos MINITAB para ilustrar, para un valor específico de X, el cálculo de un intervalo de confianza del 95 por ciento para la media de Y y un intervalo de predicción del 95 por ciento para una medición individual de Y.

Suppose, for our present example, we wish to make predictions and estimates about AT for a waist circumference of 100 cm. In the regression dialog box click on “Options.” Enter 100 in the “Prediction interval for new observations” box. Click on “Confidence limits,” and click on “Prediction limits.”

Supongamos, para nuestro ejemplo actual, que deseamos hacer predicciones y estimaciones sobre AT para una circunferencia de cintura de 100 cm. En el cuadro de diálogo de regresión, haga clic en "Opciones". Ingrese 100 en el cuadro "Intervalo de predicción para nuevas observaciones". Haga clic en "Límites de confianza" y haga clic en "Límites de predicción".

We obtain the following output:

Fit

129.90

Stdev.Fit

3.69

95.0% C.I.

(122.58, 137.23)

95.0% P.I.

(63.93, 195.87)438

CHAPTER 9 SIMPLE LINEAR REGRESSION AND CORRELATION

We interpret the 95 percent confidence interval (C.I.) as follows.

If we repeatedly drew samples from our population of men, performed a regression analysis, and estimated my ƒx =100 with a similarly constructed confidence interval, about 95 percent of such intervals would include the mean amount of deep abdominal AT for the population. For this reason we are 95 percent confident that the single interval constructed contains the population mean and that it is somewhere between 122.58 and 137.23.

Si extrajemos repetidamente muestras de nuestra población de hombres, realizamos un análisis de regresión y estimamos mi ƒx = 100 con un intervalo de confianza construido de manera similar, alrededor del 95 por ciento de dichos intervalos incluirían la cantidad media de AT abdominal profunda para la población. Por esta razón, tenemos una confianza del 95 por ciento en que el intervalo único construido contiene la media poblacional y que está entre 122,58 y 137,23.

Our interpretation of a prediction interval (P.I.) is similar to the interpretation of a confidence interval. If we repeatedly draw samples, do a regression analysis, and construct prediction intervals for men who have a waist circumference of 100 cm, about 95 percent of them will include the man’s deep abdominal AT value. This is the probabilistic interpretation. The practical interpretation is that we are 95 percent confident that a man who has a waist circumference of 100 cm will have a deep abdominal AT area of somewhere between 63.93 and 195.87 square centimeters.

Nuestra interpretación de un intervalo de predicción (P.I.) es similar a la interpretación de un intervalo de confianza. Si extraemos muestras repetidamente, hacemos un análisis de regresión y construimos intervalos de predicción para hombres que tienen una circunferencia de cintura de 100 cm, aproximadamente el 95 por ciento de ellos incluirá el valor de AT abdominal profundo del hombre. Ésta es la interpretación probabilística. La interpretación práctica es que tenemos un 95 por ciento de confianza en que un hombre con una circunferencia de cintura de 100 cm tendrá un área AT abdominal profunda de entre 63,93 y 195,87 centímetros cuadrados.

Simultaneous confidence intervals and prediction intervals can be calculated for all possible points along a fitted regression line. Plotting lines through these points will then provide a graphical representation of these intervals. Since the mean data point 1X, Y2 is always included in the regression equation, as illustrated by equations 9.3.2 and 9.3.3, plots of the simultaneous intervals will always provide the best estimates at the middle of the line and the error will increase toward the ends of the line. This illustrates the fact that estimation within the bounds of the data set, called interpolation, is acceptable, but that estimation outside of the bounds of the data set, called extrapolation, is not advisable since the pridiction error can be quite large. See Figure 9.5.1. Figure 9.5.2 contains a partial printout of the SAS® simple linear regression analysis of the data of Example 9.3.1.

Se pueden calcular intervalos de confianza e intervalos de predicción simultáneos para todos los puntos posibles a lo largo de una línea de regresión ajustada. Trazar líneas a través de estos puntos proporcionará una representación gráfica de estos intervalos. Dado que el punto de datos medio 1X, Y2 siempre se incluye en la ecuación de regresión, como lo ilustran las ecuaciones 9.3.2 y 9.3.3, las gráficas de los intervalos simultáneos siempre proporcionarán las mejores estimaciones en el medio de la línea y el error aumentará. hacia los extremos de la línea. Esto ilustra el hecho de que la estimación dentro de los límites del conjunto de datos, llamada interpolación, es aceptable, pero que la estimación fuera de los límites del conjunto de datos, llamada extrapolación, no es aconsejable ya que el error de predicción puede ser bastante grande. Ver Figura 9.5.1. La Figura 9.5.2 contiene una copia impresa parcial del análisis de regresión lineal simple de SAS® de los datos del Ejemplo 9.3.1.

**Resistant Line**

Frequently, data sets available for analysis by linear regression techniques contain one or more “unusual” observations; that is, values of x or y, or both, may be either considerably larger or considerably smaller than most of the other measurements. In the output of Figure 9.3.2, we see that the computer detected seven unusual observations in the waist circumference and deep abdominal AT data shown in Table 9.3.1.

Con frecuencia, los conjuntos de datos disponibles para el análisis mediante técnicas de regresión lineal contienen una o más observaciones “inusuales”; es decir, los valores de x o y, o de ambos, pueden ser considerablemente mayores o considerablemente menores que la mayoría de las otras mediciones. En el resultado de la Figura 9.3.2, vemos que la computadora detectó siete observaciones inusuales en la circunferencia de la cintura y los datos de AT abdominal profundo que se muestran en la Tabla 9.3.1.

The least-squares method of fitting a straight line to data is sensitive to unusual observations, and the location of the fitted line can be affected substantially by them. Because of this characteristic of the least-squares method, the resulting least-squares line is said to lack resistance to the influence of unusual observations. Several methods have been devised for dealing with this problem, including one developed by John W. Tukey. The resulting line is variously referred to as Tukey’s line and the resistant line.

El método de mínimos cuadrados para ajustar una línea recta a los datos es sensible a observaciones inusuales, y la ubicación de la línea ajustada puede verse afectada sustancialmente por ellas. Debido a esta característica del método de mínimos cuadrados, se dice que la línea de mínimos cuadrados resultante carece de resistencia a la influencia de observaciones inusuales. Se han ideado varios métodos para abordar este problema, incluido uno desarrollado por John W. Tukey. La línea resultante se conoce como línea de Tukey y línea resistente.

Based on medians, which, as we have seen, are descriptive measures that are themselves resistant to extreme values, the resistant line methodology is an exploratory data analysis tool that enables the researcher to quickly fit a straight line to a set of data consisting of paired x, y measurements. The technique involves partitioning, on the basis of the independent variable, the sample measurements into three groups of as near equal size as possible: the smallest measurements, the largest measurements, and those in between. The resistant line is the line fitted in such a way that there are an equal number of values above and below it in both the smaller group and the larger group. The resulting slope and y -intercept estimates are resistant to the effects of either extreme y values, extreme x values, or both. To illustrate the fitting of a resistant line, we use the data of Table 9.3.1 and MINITAB. The procedure and output are shown in Figure 9.5.3.

Basada en medianas, que, como hemos visto, son medidas descriptivas que en sí mismas son resistentes a valores extremos, la metodología de la línea resistente es una herramienta exploratoria de análisis de datos que permite al investigador ajustar rápidamente una línea recta a un conjunto de datos que consta de datos pareados. medidas x, y. La técnica implica dividir, sobre la base de la variable independiente, las mediciones de la muestra en tres grupos de tamaño lo más parecido posible: las mediciones más pequeñas, las mediciones más grandes y las intermedias. La línea resistente es la línea ajustada de tal manera que hay un número igual de valores encima y debajo de ella tanto en el grupo más pequeño como en el grupo más grande. Las estimaciones resultantes de la pendiente y del intercepto en el eje y son resistentes a los efectos de los valores extremos de y, los valores extremos de x o ambos. Para ilustrar el ajuste de una línea resistente, utilizamos los datos de la Tabla 9.3.1 y MINITAB. El procedimiento y el resultado se muestran en la Figura 9.5.3.

We see from the output in Figure 9.5.3 that the resistant line has a slope of 3.2869 and a y-intercept of -203.7868. The half-slope ratio, shown in the output as equal to .690, is an indicator of the degree of linearity between x and y. A slope, called a half-slope, is computed for each half of the sample data. The ratio of the right half-slope, b R, and the left half-slope, b L, is equal to b R > b L. If the relationship between x and y is straight, the half-slopes will be equal, and their ratio will be 1. A half-slope ratio that is not close to 1 indicates a lack of linearity between x and y.

Vemos en el resultado de la Figura 9.5.3 que la línea de resistencia tiene una pendiente de 3,2869 y una intersección con el eje y de -203,7868. La relación de media pendiente, que se muestra en el resultado igual a 0,690, es un indicador del grado de linealidad entre xey. Se calcula una pendiente, llamada media pendiente, para cada mitad de los datos de la muestra. La relación entre la media pendiente derecha, b R, y la media pendiente izquierda, b L, es igual a b R > b L. Si la relación entre x e y es recta, las medias pendientes serán iguales, y su relación será 1. Una relación de media pendiente que no sea cercana a 1 indica una falta de linealidad entre x e y.

The resistant line methodology is discussed in more detail by Hartwig and Dearing (1), Johnstone and Velleman (2), McNeil (3), and Velleman and Hoaglin (4).

**9.6 THE CORRELATION MODEL**

In the classic regression model, which has been the underlying model in our discussion up to this point, only Y, which has been called the dependent variable, is required to be random. The variable X is defined as a fixed (nonrandom or mathematical) variable and is referred to as the independent variable. Recall, also, that under this model observations are frequently obtained by preselecting values of X and determining corresponding values of Y.

En el modelo de regresión clásico, que ha sido el modelo subyacente en nuestra discusión hasta este punto, sólo se requiere que Y, que ha sido denominada variable dependiente, sea aleatoria. La variable X se define como una variable fija (no aleatoria o matemática) y se la denomina variable independiente. Recuerde también que, según este modelo, las observaciones se obtienen frecuentemente preseleccionando valores de X y determinando los valores correspondientes de Y.

When both Y and X are random variables, we have what is called the correlation model. Typically, under the correlation model, sample observations are obtained by selecting a random sample of the units of association (which may be persons, places, animals, points in time, or any other element on which the two measurements are taken) and taking on each a measurement of X and a measurement of Y. In this procedure, values of X are not preselected but occur at random, depending on the unit of association selected in the sample.

Cuando tanto Y como X son variables aleatorias, tenemos lo que se llama modelo de correlación. Normalmente, bajo el modelo de correlación, las observaciones muestrales se obtienen seleccionando una muestra aleatoria de las unidades de asociación (que pueden ser personas, lugares, animales, momentos en el tiempo o cualquier otro elemento sobre el cual se toman las dos mediciones) y tomando cada una es una medida de X y una medida de Y. En este procedimiento, los valores de X no se preseleccionan sino que ocurren al azar, dependiendo de la unidad de asociación seleccionada en la muestra.

Although correlation analysis cannot be carried out meaningfully under the classic regression model, regression analysis can be carried out under the correlation model. Correlation involving two variables implies a co-relationship between variables that puts them on an equal footing and does not distinguish between them by referring to one as the dependent and the other as the independent variable. In fact, in the basic computational procedures, which are the same as for the regression model, we may fit a straight line to the data either by minimizing g1yi - yNi22 or by minimizing g1x i - xN i22 . In other words, we may do a regression of X on Y as well as a regression of Y on X. The fitted line in the two cases in general will be different, and a logical question arises as to which line to fit.

Aunque el análisis de correlación no se puede realizar de manera significativa con el modelo de regresión clásico, el análisis de regresión se puede realizar con el modelo de correlación. La correlación que involucra dos variables implica una correlación entre variables que las coloca en pie de igualdad y no distingue entre ellas al referirse a una como variable dependiente y a la otra como variable independiente. De hecho, en los procedimientos computacionales básicos, que son los mismos que para el modelo de regresión, podemos ajustar una línea recta a los datos minimizando g1yi - yNi22 o minimizando g1x i - xN i22. En otras palabras, podemos hacer una regresión de X sobre Y así como una regresión de Y sobre X. La línea ajustada en los dos casos en general será diferente, y surge una pregunta lógica sobre qué línea ajustar.

If the objective is solely to obtain a measure of the strength of the relationship between the two variables, it does not matter which line is fitted, since the measure usually computed will be the same in either case. If, however, it is desired to use the equation describing the relationship between the two variables for the purposes discussed in the preceding sections, it does matter which line is fitted. The variable for which we wish to estimate means or to make predictions should be treated as the dependent variable; that is, this variable should be regressed on the other variable.

Si el objetivo es únicamente obtener una medida de la fuerza de la relación entre las dos variables, no importa qué línea se ajuste, ya que la medida normalmente calculada será la misma en ambos casos. Sin embargo, si se desea utilizar la ecuación que describe la relación entre las dos variables para los propósitos analizados en las secciones anteriores, sí importa qué línea se ajuste. La variable para la cual deseamos estimar medias o hacer predicciones debe tratarse como variable dependiente; es decir, esta variable debe retroceder en la otra variable.

**The Bivariate Normal Distribution**

Under the correlation model, X and Y are assumed to vary together in what is called a joint distribution. If this joint distribution is a normal distribution, it is referred to as a bivariate normal distribution. Inferences regarding this population may be made based on the results of samples properly drawn from it. If, on the other hand, the form of the joint distribution is known to be nonnormal, or if the form is unknown and there is no justification for assuming normality, inferential procedures are invalid, although descriptive measures may be computed.

Según el modelo de correlación, se supone que X e Y varían juntos en lo que se denomina distribución conjunta. Si esta distribución conjunta es una distribución normal, se la denomina distribución normal bivariada. Se pueden hacer inferencias sobre esta población a partir de los resultados de muestras adecuadamente extraídas de ella. Si, por otra parte, se sabe que la forma de la distribución conjunta no es normal, o si se desconoce la forma y no hay justificación para suponer normalidad, los procedimientos inferenciales no son válidos, aunque se pueden calcular medidas descriptivas.

**Correlation Assumptions**

The following assumptions must hold for inferences about the population to be valid when sampling is from a bivariate distribution.

Los siguientes supuestos deben cumplirse para que las inferencias sobre la población sean válidas cuando el muestreo se realiza a partir de una distribución bivariada.

1. For each value of X there is a normally distributed subpopulation of Y values.

2. For each value of Y there is a normally distributed subpopulation of X values.

3. The joint distribution of X and Y is a normal distribution called the bivariate normal distribution.

4. The subpopulations of Y values all have the same variance.

5. The subpopulations of X values all have the same variance.

The bivariate normal distribution is represented graphically in Figure 9.6.1. In this illustration we see that if we slice the mound parallel to Y at some value of X, the cutaway reveals the corresponding normal distribution of Y. Similarly, a slice through the mound parallel to X at some value of Y reveals the corresponding normally distributed subpopulation of X.

La distribución normal bivariada se representa gráficamente en la Figura 9.6.1. En esta ilustración vemos que si cortamos el montículo paralelo a Y en algún valor de X, el corte revela la distribución normal correspondiente de Y. De manera similar, un corte a través del montículo paralelo a X en algún valor de Y revela la correspondiente distribución normal subpoblación de X.

**9.7 THE CORRELATION COEFFICIENT**

The bivariate normal distribution discussed in Section 9.6 has five parameters, sx , sy, mx , my, and r. The first four are, respectively, the standard deviations and means associated with the individual distributions. The other parameter, r, is called the population correlation coefficient and measures the strength of the linear relationship between X and Y.

La distribución normal bivariada analizada en la sección 9.6 tiene cinco parámetros, sx, sy, mx, my y r. Los primeros cuatro son, respectivamente, las desviaciones estándar y las medias asociadas con las distribuciones individuales. El otro parámetro, r, se llama coeficiente de correlación poblacional y mide la fuerza de la relación lineal entre X e Y.

The population correlation coefficient is the positive or negative square root of r2, the population coefficient of determination previously discussed, and since the coefficient of determination takes on values between 0 and 1 inclusive, r may assume any value between -1 and +1. If r = 1 there is a perfect direct linear correlation between the two variables, while r = -1 indicates perfect inverse linear correlation. If r = 0 the two variables are not linearly correlated. The sign of r will always be the same as the sign of b 1, the slope of the population regression line for X and Y.

El coeficiente de correlación poblacional es la raíz cuadrada positiva o negativa de r2, el coeficiente de determinación poblacional discutido anteriormente, y dado que el coeficiente de determinación toma valores entre 0 y 1 inclusive, r puede asumir cualquier valor entre -1 y +1. Si r = 1 existe una correlación lineal directa perfecta entre las dos variables, mientras que r = -1 indica una correlación lineal inversa perfecta. Si r = 0 las dos variables no están correlacionadas linealmente. El signo de r siempre será el mismo que el signo de b 1, la pendiente de la recta de regresión poblacional para X e Y.

The sample correlation coefficient, r, describes the linear relationship between the sample observations on two variables in the same way that r describes the relationship in a population. The sample correlation coefficient is the square root of the sample coefficient of determination that was defined earlier.

El coeficiente de correlación muestral, r, describe la relación lineal entre las observaciones muestrales de dos variables de la misma manera que r describe la relación en una población. El coeficiente de correlación muestral es la raíz cuadrada del coeficiente de determinación muestral que se definió anteriormente.

Figures 9.4.5(d) and 9.4.5(c), respectively, show typical scatter diagrams where r : 0 1r 2 : 02 and r = +1 1r 2 = 12. Figure 9.7.1 shows a typical scatter diagram where r = -1.

We are usually interested in knowing if we may conclude that r Z 0, that is, that X and Y are linearly correlated. Since r is usually unknown, we draw a random sample from the population of interest, compute r, the estimate of r, and test H0 : r = 0 against the alternative r Z 0. The procedure will be illustrated in the following example.

Generalmente nos interesa saber si podemos concluir que r Z 0, es decir, que X e Y están correlacionados linealmente. Como r suele ser desconocido, se toma una muestra aleatoria de la población de interés, se calcula r, la estimación de r y se prueba H0 : r = 0 contra la alternativa r Z 0. El procedimiento se ilustrará en el siguiente ejemplo.

**EXAMPLE 9.7.1**

The purpose of a study by Kwast-Rabben et al. (A-7) was to analyze somatosensory evoked potentials (SEPs) and their interrelations following stimulation of digits I, III, and V in the hand. The researchers wanted to establish reference criteria in a control population. Thus, healthy volunteers were recruited for the study. In the future this information could be quite valuable as SEPs may provide a method to demonstrate functional disturbances in patients with suspected cervical root lesion who have pain and sensory symptoms. In the study, stimulation below-pain-level intensity was applied to the fingers.

El propósito de un estudio de Kwast-Rabben et al. (A-7) fue analizar los potenciales evocados somatosensoriales (SEP) y sus interrelaciones después de la estimulación de los dedos I, III y V de la mano. Los investigadores querían establecer criterios de referencia en una población de control. Por tanto, se reclutaron voluntarios sanos para el estudio. En el futuro, esta información podría ser muy valiosa ya que los SEP pueden proporcionar un método para demostrar alteraciones funcionales en pacientes con sospecha de lesión de la raíz cervical que presentan dolor y síntomas sensoriales. En el estudio, se aplicó en los dedos una estimulación por debajo del nivel del dolor.

Recordings of spinal responses were made with electrodes fixed by adhesive electrode cream to the subject’s skin. One of the relationships of interest was the correlation between a subject’s height (cm) and the peak spinal latency (Cv) of the SEP. The data for 155 measurements are shown in Table 9.7.1.

Los registros de las respuestas de la columna se realizaron con electrodos fijados mediante una crema adhesiva para electrodos a la piel del sujeto. Una de las relaciones de interés fue la correlación entre la altura de un sujeto (cm) y la latencia espinal máxima (Cv) del SEP. Los datos de 155 mediciones se muestran en la Tabla 9.7.1.

Solution:

The scatter diagram and least-squares regression line are shown in Figure 9.7.2.

Let us assume that the investigator wishes to obtain a regression equation to use for estimating and predicting purposes. In that case the sample correlation coefficient will be obtained by the methods discussed under the regression model.

**The Regression Equation**

Let us assume that we wish to predict Cv levels from knowledge of heights. In that case we treat height as the independent variable and Cv level as the dependent variable and obtain the regression equation and correlation coefficient with MINITAB as shown in Figure 9.7.3. For this example r = 1.719 = .848. We know that r is positive because the slope of the regression line is positive. We may also use the MINITAB correlation procedure to obtain r as shown in Figure 9.7.4.

Supongamos que deseamos predecir los niveles de Cv a partir del conocimiento de las alturas. En ese caso tratamos la altura como la variable independiente y el nivel de Cv como la variable dependiente y obtenemos la ecuación de regresión y el coeficiente de correlación con MINITAB como se muestra en la Figura 9.7.3. Para este ejemplo r = 1,719 = 0,848. Sabemos que r es positivo porque la pendiente de la recta de regresión es positiva. También podemos usar el procedimiento de correlación MINITAB para obtener r como se muestra en la Figura 9.7.4.

The printout from the SAS® correlation procedure is shown in Figure 9.7.5. Note that the SAS® procedure gives descriptive measures for each variable as well as the p value for the correlation coefficient.

La impresión del procedimiento de correlación SAS® se muestra en la Figura 9.7.5. Tenga en cuenta que el procedimiento SAS® proporciona medidas descriptivas para cada variable, así como el valor p para el coeficiente de correlación.

When a computer is not available for performing the calculations, r may be obtained by means of the following formulas:

An alternative formula for computing r is given by

r =

n g x i yi - 1g x i21g yi2

2n g x i2 - 1g x i22 2ng y i2 - 1g yi22

(9.7.2)

An advantage of this formula is that r may be computed without first computing b. This is the desirable procedure when it is not anticipated that the regression equation will be used.

Remember that the sample correlation coefficient, r, will always have the same sign as the sample slope, b.

■

**EXAMPLE 9.7.2**

Refer to Example 9.7.1. We wish to see if the sample value of r = .848 is of sufficient magnitude to indicate that, in the population, height and Cv SEP levels are correlated.

Solution:

We conduct a hypothesis test as follows.

1. **Data**. See the initial discussion of Example 9.7.1.

2. **Assumptions**. We presume that the assumptions given in Section 9.6 are applicable.

3. **Hypotheses**.

H0 : r = 0

HA : r Z 0

4. **Test** **statistic**. When r = 0, it can be shown that the appropriate test statistic is

t = r

n - 2

A1 - r2

(9.7.3)

5. **Distribution** **of** **test** **statistic**. When H0 is true and the assumptions are met, the test statistic is distributed as Student’s t distribution with n - 2 degrees of freedom.

6. **Decision** **rule**. If we let a = .05, the critical values of t in the present example are ; 1.9754 (by interpolation). If, from our data, we compute a value of t that is either greater than or equal to +1.9754 or less than or equal to -1.9754, we will reject the null hypothesis.

7. **Calculation** **of** **test** **statistic**. Our calculated value of t is

t = .848

153

= 19.787

A 1 - .719

8. **Statistical** **decision**. Since the computed value of the test statistic does exceed the critical value of t, we reject the null hypothesis.

9. **Conclusion**. We conclude that, in the population, height and SEP levels in the spine are linearly correlated.

10. p value. Since t = 19.787 7 2.6085 (interpolated value of t for 153, .995), we have for this test, p 6 .005.

■

**A Test for Use When the Hypothesized R Is a Nonzero Value**

The use of the t statistic computed in the above test is appropriate only for testing H0 : r = 0. If it is desired to test H0 : r = r0, where r0 is some value other than zero, we must use another approach. Fisher (5) suggests that r be transformed to z r as follows:

El uso del estadístico t calculado en la prueba anterior es apropiado sólo para probar H0 : r = 0. Si se desea probar H0 : r = r0, donde r0 es algún valor distinto de cero, debemos usar otro enfoque. Fisher (5) sugiere que r se transforme en z r de la siguiente manera:

zr =

1 1 + r

ln

2 1 - r

(9.7.4)

where ln is a natural logarithm. It can be shown that z r is approximately normally distributed with a mean of z r = 12 ln511 + r2>11 - r26 and estimated standard deviation of

szp =

1

1n - 3

(9.7.5)

To test the null hypothesis that r is equal to some value other than zero, the test statistic is

Z =

zr - zr

1> 1n - 3

(9.7.6)

which follows approximately the standard normal distribution.

To determine z r for an observed r and z r for a hypothesized r, we consult Table I, thereby avoiding the direct use of natural logarithms.

Suppose in our present example we wish to test

H0 : r = .80

against the alternative

HA : r Z .80

at the .05 level of significance. By consulting Table I (and interpolating), we find that for

r = .848,z r = 1.24726

r = .80,z r = 1.09861

and for

Our test statistic, then, is

Z =

1.24726 - 1.09861

= 1.83

1> 1155 - 3450

CHAPTER 9 SIMPLE LINEAR REGRESSION AND CORRELATION

Since 1.83 is less than the critical value of z = 1.96, we are unable to reject H0. We conclude that the population correlation coefficient may be .80.

For sample sizes less than 25, Fisher’s Z transformation should be used with caution, if at all. An alternative procedure from Hotelling (6) may be used for sample sizes equal to or greater than 10. In this procedure the following transformation of r is employed:

z\* = z r -

3z r + r

4n

(9.7.7)

The standard deviation of z\* is

sz\* =

The test statistic is

Z\* =

where

z\* - z\*

1> 1n - 1

1

1n - 1

(9.7.8)

= 1z\* - z\*2 1n - 1

z\* 1pronounced zeta2 = z r -

(9.7.9)

13z r + r2

4n

Critical values for comparison purposes are obtained from the standard normal distribution.

In our present example, to test H0 : r = .80 against HA : r Z .80 using the Hotelling transformation and a = .05, we have

z\* = 1.24726 -

z\* = 1.09861 -

311.247262 + .848

411552

311.098612 + .8

411552

= 1.2339

= 1.0920

Z \* = 11.2339 - 1.09202 1155 - 1 = 1.7609

Since 1.7609 is less than 1.96, the null hypothesis is not rejected, and the same conclusion is reached as when the Fisher transformation is used.

**Alternatives**

In some situations the data available for analysis do not meet the assumptions necessary for the valid use of the procedures discussed here for testing hypotheses about a population correlation coefficient. In such cases it may be more appropriate to use the Spearman rank correlation technique discussed in Chapter 13.

En algunas situaciones, los datos disponibles para el análisis no cumplen con los supuestos necesarios para el uso válido de los procedimientos discutidos aquí para probar hipótesis sobre un coeficiente de correlación poblacional. En tales casos, puede ser más apropiado utilizar la técnica de correlación de rangos de Spearman que se analiza en el Capítulo 13.

**Confidence Interval for R**

Fisher’s transformation may be used to construct 10011 - a2 percent confidence intervals for r. The general formula for a confidence interval

estimator ; (reliability factor)(standard error)EXERCISES

451

is employed. We first convert our estimator, r, to zr , construct a confidence interval about z r, and then reconvert the limits to obtain a 10011 - a2 percent confidence interval about r. The general formula then becomes

z r ; z 11> 1n - 32

(9.7.10)

For our present example the 95 percent confidence interval for z r is given by

1.24726 ; 1.9611> 1155 - 32

1.08828, 1.40624

Converting these limits (by interpolation in Appendix Table I), which are values

of z r , into values of r gives

zr

1.08828

1.40624

r

.7962

.8866

We are 95 percent confident, then, that r is contained in the interval .7962 to .88866. Because of the limited entries in the table, these limits must be considered as only approximate.

**9.8 SOME PRECAUTIONS**

Regression and correlation analysis are powerful statistical tools when properly employed. Their inappropriate use, however, can lead only to meaningless results. To aid in the proper use of these techniques, we make the following suggestions:

Los análisis de regresión y correlación son herramientas estadísticas poderosas cuando se emplean adecuadamente. Sin embargo, su uso inadecuado sólo puede conducir a resultados sin sentido. Para ayudar en el uso adecuado de estas técnicas, hacemos las siguientes sugerencias:

1. The assumptions underlying regression and correlation analysis should be reviewed carefully before the data are collected. Although it is rare to find that assumptions are met to perfection, practitioners should have some idea about the magnitude of the gap that exists between the data to be analyzed and the assumptions of the proposed model, so that they may decide whether they should choose another model; proceed with the analysis, but use caution in the interpretation of the results; or use the chosen model with confidence.

Los supuestos subyacentes al análisis de regresión y correlación deben revisarse cuidadosamente antes de recopilar los datos. Aunque es raro encontrar que los supuestos se cumplan a la perfección, los profesionales deben tener alguna idea sobre la magnitud de la brecha que existe entre los datos a analizar y los supuestos del modelo propuesto, de modo que puedan decidir si deben elegir otro. modelo; proceder con el análisis, pero tener precaución en la interpretación de los resultados; o utilice el modelo elegido con confianza.

2. In simple linear regression and correlation analysis, the two variables of interest are measured on the same entity, called the unit of association. If we are interested in the relationship between height and weight, for example, these two measurements are taken on the same individual. It usually does not make sense to speak of the correlation, say, between the heights of one group of individuals and the weights of another group.

En el análisis de correlación y regresión lineal simple, las dos variables de interés se miden en la misma entidad, llamada unidad de asociación. Si nos interesa la relación entre altura y peso, por ejemplo, estas dos medidas se toman en el mismo individuo. Generalmente no tiene sentido hablar de correlación, digamos, entre las alturas de un grupo de individuos y el peso de otro grupo.

3. No matter how strong is the indication of a relationship between two variables, it should not be interpreted as one of cause and effect. If, for example, a significant sample correlation coefficient between two variables X and Y is observed, it can mean one of several things:

No importa cuán fuerte sea el indicio de una relación entre dos variables, no debe interpretarse como una relación de causa y efecto. Si, por ejemplo, se observa un coeficiente de correlación muestral significativo entre dos variables X e Y, puede significar una de varias cosas:

a. X causes Y.

b. Y causes X.

c. Some third factor, either directly or indirectly, causes both X and Y.

d. An unlikely event has occurred and a large sample correlation coefficient has been generated by chance from a population in which X and Y are, in fact, not correlated.

e. The correlation is purely nonsensical, a situation that may arise when measurements of X and Y are not taken on a common unit of association.

4. The sample regression equation should not be used to predict or estimate outside the range of values of the independent variable represented in the sample. As illustrated in Section 9.5, this practice, called extrapolation, is risky. The true relationship between two variables, although linear over an interval of the independent variable, sometimes may be described at best as a curve outside this interval. If our sample by chance is drawn only from the interval where the relationship is linear, we have only a limited representation of the population, and to project the sample results beyond the interval represented by the sample may lead to false conclusions. Figure 9.8.1 illustrates the possible pitfalls of extrapolation.

La ecuación de regresión muestral no debe usarse para predecir o estimar fuera del rango de valores de la variable independiente representada en la muestra. Como se ilustra en la Sección 9.5, esta práctica, llamada extrapolación, es arriesgada. La verdadera relación entre dos variables, aunque lineal a lo largo de un intervalo de la variable independiente, a veces puede describirse, en el mejor de los casos, como una curva fuera de este intervalo. Si nuestra muestra por casualidad se extrae sólo del intervalo donde la relación es lineal, tenemos sólo una representación limitada de la población, y proyectar los resultados de la muestra más allá del intervalo representado por la muestra puede llevar a conclusiones falsas. La Figura 9.8.1 ilustra los posibles errores de la extrapolación.