BIOSTATISTICS\_Daniel\_C10

TER

MULTIPLE REGRESSION

AND CORRELATION

CHAPTER OVERVIEW

This chapter provides extensions of the simple linear regression and bivariate correlation models discussed in Chapter 9. The concepts and techniques discussed here are useful when the researcher wishes to consider simultaneously the relationships among more than two variables. Although the concepts, computations, and interpretations associated with analysis of multiple-variable data may seem complex, they are natural extensions of material explored in previous chapters.

Este capítulo proporciona extensiones de los modelos de regresión lineal simple y correlación bivariada discutidos en el Capítulo 9. Los conceptos y técnicas aquí discutidos son útiles cuando el investigador desea considerar simultáneamente las relaciones entre más de dos variables. Aunque los conceptos, cálculos e interpretaciones asociados con el análisis de datos de múltiples variables pueden parecer complejos, son extensiones naturales del material explorado en capítulos anteriores.

TO P I C S

10.1INTRODUCTION

10.2THE MULTIPLE LINEAR REGRESSION MODEL

10.3OBTAINING THE MULTIPLE REGRESSION EQUATION

10.4EVALUATING THE MULTIPLE REGRESSION EQUATION

10.5USING THE MULTIPLE REGRESSION EQUATION

10.6THE MULTIPLE CORRELATION MODEL

10.7SUMMARY

LEARNING OUTCOMES

After studying this chapter, the student will

Después de estudiar este capítulo, el estudiante

1. understand how to include more than one independent variable in a regression equation.

comprender cómo incluir más de una variable independiente en una ecuación de regresión.

2. be able to obtain a multiple regression model and use it to make predictions.

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

3. be able to evaluate the multiple regression coefficients and the suitability of the regression model.

Ser capaz de evaluar los coeficientes de regresión múltiple y la idoneidad del modelo de regresión.

4. understand how to calculate and interpret multiple, bivariate, and partial correlation coefficients.

comprender cómo calcular e interpretar coeficientes de correlación múltiples, bivariados y parciales.

**10.1 INTRODUCTION**

In Chapter 9 we explored the concepts and techniques for analyzing and making use of the linear relationship between two variables. We saw that this analysis may lead to an equation that can be used to predict the value of some dependent variable given the value of an associated independent variable.

En el Capítulo 9 exploramos los conceptos y técnicas para analizar y hacer uso de la relación lineal entre dos variables. Vimos que este análisis puede conducir a una ecuación que puede usarse para predecir el valor de alguna variable dependiente dado el valor de una variable independiente asociada.

Intuition tells us that, in general, we ought to be able to improve our predicting ability by including more independent variables in such an equation. For example, a researcher may find that intelligence scores of individuals may be predicted from physical factors such as birth order, birth weight, and length of gestation along with certain hereditary and external environmental factors. Length of stay in a chronic disease hospital may be related to the patient’s age, marital status, sex, and income, not to mention the obvious factor of diagnosis. The response of an experimental animal to some drug may depend on the size of the dose and the age and weight of the animal. A nursing supervisor may be interested in the strength of the relationship between a nurse’s performance on the job, score on the state board examination, scholastic record, and score on some achievement or aptitude test. Or a hospital administrator studying admissions from various communities served by the hospital may be interested in determining what factors seem to be responsible for differences

in admission rates.

La intuición nos dice que, en general, deberíamos poder mejorar nuestra capacidad de predicción incluyendo más variables independientes en dicha ecuación. Por ejemplo, un investigador puede descubrir que las puntuaciones de inteligencia de los individuos pueden predecirse a partir de factores físicos como el orden de nacimiento, el peso al nacer y la duración de la gestación, junto con ciertos factores ambientales hereditarios y externos. La duración de la estancia en un hospital de enfermedades crónicas puede estar relacionada con la edad, el estado civil, el sexo y los ingresos del paciente, sin mencionar el factor obvio del diagnóstico. La respuesta de un animal de experimentación a algún fármaco puede depender del tamaño de la dosis y de la edad y peso del animal. Un supervisor de enfermería puede estar interesado en la solidez de la relación entre el desempeño de una enfermera en el trabajo, la puntuación en el examen de la junta estatal, el historial académico y la puntuación en alguna prueba de rendimiento o aptitud. O un administrador de hospital que estudie las admisiones de varias comunidades atendidas por el hospital puede estar interesado en determinar qué factores parecen ser responsables de las diferencias. en las tarifas de admisión.

The concepts and techniques for analyzing the associations among several variables are natural extensions of those explored in the previous chapters. The computations, as one would expect, are more complex and tedious. However, as is pointed out in Chapter 9, this presents no real problem when a computer is available. It is not unusual to find researchers investigating the relationships among a dozen or more variables. For those who have access to a computer, the decision as to how many variables to include in an analysis is based not on the complexity and length of the computations but on such considerations as their meaningfulness, the cost of their inclusion, and the importance of their contribution.

Los conceptos y técnicas para analizar las asociaciones entre varias variables son extensiones naturales de los explorados en los capítulos anteriores. Los cálculos, como era de esperar, son más complejos y tediosos. Sin embargo, como se señala en el Capítulo 9, esto no presenta ningún problema real cuando hay una computadora disponible. No es raro encontrar investigadores que investiguen las relaciones entre una docena o más de variables. Para quienes tienen acceso a una computadora, la decisión sobre cuántas variables incluir en un análisis no se basa en la complejidad y extensión de los cálculos, sino en consideraciones tales como su significado, el costo de su inclusión y la importancia de su contribución.

In this chapter we follow closely the sequence of the previous chapter. The regression model is considered first, followed by a discussion of the correlation model. In considering the regression model, the following points are covered: a description of the model, methods for obtaining the regression equation, evaluation of the equation, and the uses that may be made of the equation. In both models the possible inferential procedures and their underlying assumptions are discussed.

En este capítulo seguimos de cerca la secuencia del capítulo anterior. Primero se considera el modelo de regresión, seguido de una discusión del modelo de correlación. Al considerar el modelo de regresión, se cubren los siguientes puntos: una descripción del modelo, métodos para obtener la ecuación de regresión, evaluación de la ecuación y los usos que se pueden hacer de la ecuación. En ambos modelos se discuten los posibles procedimientos inferenciales y sus supuestos subyacentes.

**10.2 THE MULTIPLE LINEAR REGRESSION MODEL**

In the multiple regression model we assume that a linear relationship exists between some variable Y, which we call the dependent variable, and k independent variables, X1, X2, . . . , Xk. The independent variables are sometimes referred to as explanatory variables, because of their use in explaining the variation in Y. They are also called predictor variables, because of their use in predicting Y.

En el modelo de regresión múltiple suponemos que existe una relación lineal entre alguna variable Y, a la que llamamos variable dependiente, y k variables independientes, X1, X2,. . . , Xk. A las variables independientes a veces se les llama variables explicativas, debido a su uso para explicar la variación en Y. También se les llama variables predictoras, debido a su uso para predecir Y.

**Assumptions**

**Suposiciones**

The assumptions underlying multiple regression analysis are as follows.

Los supuestos que subyacen al análisis de regresión múltiple son los siguientes.

1. The Xi are nonrandom (fixed) variables. This assumption distinguishes the multiple regression model from the multiple correlation model, which will be presented in Section 10.6. This condition indicates that any inferences that are drawn from sample data apply only to the set of X values observed and not to some larger collection of X’s. Under the regression model, correlation analysis is not meaningful. Under the correlation model to be presented later, the regression techniques that follow may be applied.

Las Xi son variables no aleatorias (fijas). Este supuesto distingue el modelo de regresión múltiple del modelo de correlación múltiple, que se presentará en la Sección 10.6. Esta condición indica que cualquier inferencia que se extraiga de datos de muestra se aplica sólo al conjunto de valores de X observados y no a una colección más grande de X. Según el modelo de regresión, el análisis de correlación no tiene sentido. Bajo el modelo de correlación que se presentará más adelante, se pueden aplicar las técnicas de regresión que siguen.

2. For each set of Xi values there is a subpopulation of Y values. To construct certain confidence intervals and test hypotheses, it must be known, or the researcher must be willing to assume, that these subpopulations of Y values are normally distributed. Since we will want to demonstrate these inferential procedures, the assumption of normality will be made in the examples and exercises in this chapter.

Para cada conjunto de valores Xi existe una subpoblación de valores Y. Para construir ciertos intervalos de confianza y probar hipótesis, se debe saber, o el investigador debe estar dispuesto a suponer, que estas subpoblaciones de valores de Y se distribuyen normalmente. Como queremos demostrar estos procedimientos inferenciales, en los ejemplos y ejercicios de este capítulo se asumirá la normalidad.

3. The variances of the subpopulations of Y are all equal.

Las varianzas de las subpoblaciones de Y son todas iguales.

4. The Y values are independent. That is, the values of Y selected for one set of X

values do not depend on the values of Y selected at another set of X values.

Los valores de Y son independientes. Es decir, los valores de Y seleccionados para un conjunto de valores X no dependen de los valores de Y seleccionados en otro conjunto de valores X.

**The Model Equation**

**La ecuación modelo**

The assumptions for multiple regression analysis may be stated in more compact fashion as

Los supuestos para el análisis de regresión múltiple pueden expresarse de manera más compacta como

y j = b 0 + b 1x 1j + b 2 x 2j + . . . + b k x kj + P j

(10.2.1)

where yj is a typical value from one of the subpopulations of Y values; the b i are called the regression coefficients; x 1j , x 2j , Á , x kj are, respectively, particular values of the independent variables X1, X2, Á Xk; and Pj is a random variable with mean 0 and variance s2, the common variance of the subpopulations of Y values. To construct confidence intervals for and test hypotheses about the regression coefficients, we assume that the Pj are normally and independently distributed. The statements regarding Pj are a consequence of the assumptions regarding the distributions of Y values. We will refer to Equation 10.2.1 as the multiple linear regression model.

When Equation 10.2.1 consists of one dependent variable and two independent variables, that is, when the model is written

donde yj es un valor típico de una de las subpoblaciones de valores de Y; los b i se denominan coeficientes de regresión; x 1j , x 2j , Á , x kj son, respectivamente, valores particulares de las variables independientes X1, X2, Á Xk; y Pj es una variable aleatoria con media 0 y varianza s2, la varianza común de las subpoblaciones de valores de Y. Para construir intervalos de confianza y probar hipótesis sobre los coeficientes de regresión, suponemos que los Pj se distribuyen normal e independientemente. Las afirmaciones sobre Pj son consecuencia de los supuestos sobre las distribuciones de los valores de Y. Nos referiremos a la Ecuación 10.2.1 como modelo de regresión lineal múltiple.

Cuando la Ecuación 10.2.1 consta de una variable dependiente y dos variables independientes, es decir, cuando el modelo se escribe

yj = b 0 + b 1x 1j + b 2 x 2j + P j

(10.2.2)

a plane in three-dimensional space may be fitted to the data points as illustrated in Figure 10.2.1. When the model contains more than two independent variables, it is described geometrically as a hyperplane.

Se puede ajustar un plano en un espacio tridimensional a los puntos de datos como se ilustra en la Figura 10.2.1. Cuando el modelo contiene más de dos variables independientes, se describe geométricamente como un hiperplano.

In Figure 10.2.1 the observer should visualize some of the points as being located above the plane and some as being located below the plane. The deviation of a point from the plane is represented by

En la Figura 10.2.1, el observador debe visualizar algunos de los puntos ubicados sobre el plano y otros debajo del avión. La desviación de un punto respecto del plano está representada por

P j = y j - b 0 - b 1x 1j - b 2 x 2j

(10.2.3)

In Equation 10.2.2, b 0 represents the point where the plane cuts the Y-axis; that is, it represents the Y-intercept of the plane. b 1 measures the average change in Y for a unit change in X1 when X2 remains unchanged, and b 2 measures the average change in Y for a unit change in X2 when X1 remains unchanged. For this reason b 1 and b 2 are referred to as partial regression coefficients.

En la Ecuación 10.2.2, b 0 representa el punto donde el plano corta al eje Y; es decir, representa la intersección en Y del avión. b 1 mide el cambio promedio en Y para un cambio unitario en X1 cuando X2 permanece sin cambios, y b 2 mide el cambio promedio en Y para un cambio unitario en X2 cuando X1 permanece sin cambios. Por esta razón, b 1 y b 2 se denominan coeficientes de regresión parcial.

**10.3 OBTAINING THE MULTIPLE REGRESSION EQUATION**

OBTENCIÓN DE LA ECUACIÓN DE REGRESIÓN MÚLTIPLE

Unbiased estimates of the parameters b0, b1, . . . , bk of the model specified in Equation 10.2.1 are obtained by the method of least squares. This means that the sum of the squared deviations of the observed values of Y from the resulting regression surface is minimized. In the three-variable case, as illustrated in Figure 10.2.1, the sum of the squared deviations of the observations from the plane are a minimum when b0, b1, and b2 are estimated by the method of least squares. In other words, by the method of least squares, sample estimates of b0, b1, . . . , bk are selected in such a way that the quantity

Estimaciones insesgadas de los parámetros b0, b1, . . . , bk del modelo especificado en la Ecuación 10.2.1 se obtienen mediante el método de mínimos cuadrados. Esto significa que se minimiza la suma de las desviaciones al cuadrado de los valores observados de Y de la superficie de regresión resultante. En el caso de tres variables, como se ilustra en la Figura 10.2.1, la suma de las desviaciones al cuadrado de las observaciones desde el plano es mínima cuando b0, b1 y b2 se estiman mediante el método de mínimos cuadrados. En otras palabras, mediante el método de mínimos cuadrados, las estimaciones muestrales de b0, b1, . . . , bk se seleccionan de tal manera que la cantidad

gPj2 = g1y j - b 0 - b 1 x 1j - b 2 x 2j - . . . - b k x kj22

is minimized. This quantity, referred to as the sum of squares of the residuals, may also

be written as

gPj2 = g1yj - yN j22

(10.3.1)

indicating the fact that the sum of squares of deviations of the observed values of Y from the values of Y calculated from the estimated equation is minimized.

Estimates of the multiple regression parameters may be obtained by means of arithmetic calculations performed on a handheld calculator. This method of obtaining the estimates is tedious, time-consuming, subject to errors, and a waste of time when a computer is available. Those interested in examining or using the arithmetic approach may consult earlier editions of this text or those by Snedecor and Cochran (1) and Steel and Torrie (2), who give numerical examples for four variables, and Anderson and Bancroft (3), who illustrate the calculations involved when there are five variables. In the following example we use SPSS software to illustrate an interesting graphical summary of sample data collected on three variables. We then use MINITAB to illustrate the application of multiple regression analysis.

Las estimaciones de los parámetros de regresión múltiple se pueden obtener mediante cálculos aritméticos realizados en una calculadora portátil. Este método de obtener las estimaciones es tedioso, requiere mucho tiempo, está sujeto a errores y es una pérdida de tiempo cuando se dispone de una computadora. Aquellos interesados en examinar o utilizar el enfoque aritmético pueden consultar ediciones anteriores de este texto o las de Snedecor y Cochran (1) y Steel y Torrie (2), quienes dan ejemplos numéricos para cuatro variables, y Anderson y Bancroft (3), quienes Ilustre los cálculos involucrados cuando hay cinco variables. En el siguiente ejemplo utilizamos el software SPSS para ilustrar un interesante resumen gráfico de datos de muestra recopilados sobre tres variables. Luego usamos MINITAB para ilustrar la aplicación del análisis de regresión múltiple.

**EXAMPLE 10.3.1**

Researchers Jansen and Keller (A-1) used age and education level to predict the capacity to direct attention (CDA) in elderly subjects. CDA refers to neural inhibitory mechanisms that focus the mind on what is meaningful while blocking out distractions. The study collected information on 71 community-dwelling older women with normal mental status. The CDA measurement was calculated from results on standard visual and auditory measures requiring the inhibition of competing and distracting stimuli. In this study, CDA scores ranged from -7.65 to 9.61 with higher scores corresponding with better attentional functioning. The measurements on CDA, age in years, and education level (years of schooling) for 71 subjects are shown in Table 10.3.1. We wish to obtain the sample multiple regression equation.

Los investigadores Jansen y Keller (A-1) utilizaron la edad y el nivel educativo para predecir la capacidad de dirigir la atención (CDA) en sujetos de edad avanzada. CDA se refiere a mecanismos inhibidores neuronales que enfocan la mente en lo que es significativo mientras bloquean las distracciones. El estudio recopiló información sobre 71 mujeres mayores que vivían en la comunidad y tenían un estado mental normal. La medición CDA se calculó a partir de los resultados de medidas visuales y auditivas estándar que requieren la inhibición de estímulos competitivos y que distraen. En este estudio, las puntuaciones CDA oscilaron entre -7,65 y 9,61 y las puntuaciones más altas se corresponden con un mejor funcionamiento de la atención. Las mediciones de CDA, edad en años y nivel de educación (años de escolaridad) para 71 sujetos se muestran en la Tabla 10.3.1. Deseamos obtener la ecuación de regresión múltiple de muestra.

Prior to analyzing the data using multiple regression techniques, it is useful to construct plots of the relationships among the variables. This is accomplished by making separate plots of each pair of variables, (X1, X2), (X1, Y ), and (X2, Y ). A software package such as SPSS displays each combination simultaneously in a matrix format as shown in Figure 10.3.1. From this figure it is apparent that we should expect a negative relationship between CDA and Age and a positive relationship between CDA and Ed-Level. We shall see that this is indeed the case when we use MINITAB to analyze the data.

Antes de analizar los datos utilizando técnicas de regresión múltiple, es útil construir gráficos de las relaciones entre las variables. Esto se logra haciendo gráficos separados de cada par de variables, (X1, X2), (X1, Y) y (X2, Y). Un paquete de software como SPSS muestra cada combinación simultáneamente en un formato matricial como se muestra en la Figura 10.3.1. De esta figura se desprende que deberíamos esperar una relación negativa entre CDA y Edad y una relación positiva entre CDA y Ed-Level. Veremos que este es efectivamente el caso cuando usemos MINITAB para analizar los datos.

**Solution:**

We enter the observations on age, education level, and CDA in c1 through c3 and name them X1, X2, and Y, respectively. The MINITAB dialog box and session command, as well as the output, are shown in Figure 10.3.2. We see from the output that the sample multiple regression equation, in the notation of Section 10.2, is

Ingresamos las observaciones sobre edad, nivel educativo y CDA en c1 a c3 y las denominamos X1, X2 e Y, respectivamente. El cuadro de diálogo MINITAB y el comando de sesión, así como el resultado, se muestran en la Figura 10.3.2. Vemos en el resultado que la ecuación de regresión múltiple de muestra, en la notación de la Sección 10.2, es

yNj = 5.49 - .184x 1j + .611x 2j

Other output entries will be discussed in the sections that follow. The SAS output for Example 10.3.1 is shown in Figure 10.3.3.

■

After the multiple regression equation has been obtained, the next step involves its evaluation and interpretation. We cover this facet of the analysis in the next section.

Una vez obtenida la ecuación de regresión múltiple, el siguiente paso consiste en su evaluación e interpretación. Cubrimos esta faceta del análisis en la siguiente sección.

**10.4 EVALUATING THE MULTIPLE REGRESSION EQUATION**

Before one uses a multiple regression equation to predict and estimate, it is desirable to determine first whether it is, in fact, worth using. In our study of simple linear regression we have learned that the usefulness of a regression equation may be evaluated by a consideration of the sample coefficient of determination and estimated slope. In evaluating a multiple regression equation we focus our attention on the coefficient of multiple deter- mination and the partial regression coefficients.

Antes de utilizar una ecuación de regresión múltiple para predecir y estimar, es deseable determinar primero si, de hecho, vale la pena usarla. En nuestro estudio de regresión lineal simple hemos aprendido que la utilidad de una ecuación de regresión puede evaluarse considerando el coeficiente de determinación de la muestra y la pendiente estimada. Al evaluar una ecuación de regresión múltiple centramos nuestra atención en el coeficiente de determinación múltiple y los coeficientes de regresión parcial.

**The Coefficient of Multiple Determination**

In Chapter 9 the coefficient of determination is discussed in considerable detail. The concept extends logically to the multiple regression case. The total variation present in the Y values may be partitioned into two components—the explained variation, which measures the amount of the total variation that is explained by the fitted regression surface, and the unexplained variation, which is that part of the total variation not explained by fitting the regression surface.

En el Capítulo 9 se analiza con considerable detalle el coeficiente de determinación. El concepto se extiende lógicamente al caso de regresión múltiple. La variación total presente en los valores de Y se puede dividir en dos componentes: la variación explicada, que mide la cantidad de variación total que se explica por la superficie de regresión ajustada, y la variación no explicada, que es la parte de la variación total que no se explica. explicado ajustando la superficie de regresión.

The measure of variation in each case is a sum of squared deviations. The total variation is the sum of squared deviations of each observation of Y from the mean of the observations and is designated by g1yj - y22 or SST. The explained variation, designated g1yN j - y22, is the sum of squared deviations of the calculated values from the mean of the observed Y values. This sum of squared deviations is called the sum of squares due to regression (SSR). The unexplained variation, written as g1yj - yN j22, is the sum of squared deviations of the original observations from the calculated values. This quantity is referred to as the sum of squares about regression or the error sum of squares (SSE). We may summarize the relationship among the three sums of squares with the following equation:

La medida de variación en cada caso es una suma de desviaciones al cuadrado. La variación total es la suma de las desviaciones al cuadrado de cada observación de Y de la media de las observaciones y se designa como g1yj - y22 o SST. La variación explicada, denominada g1yN j - y22, es la suma de las desviaciones al cuadrado de los valores calculados de la media de los valores Y observados. Esta suma de desviaciones al cuadrado se llama suma de cuadrados debida a la regresión (SSR). La variación inexplicable, escrita como g1yj - yN j22, es la suma de las desviaciones al cuadrado de las observaciones originales de los valores calculados. Esta cantidad se conoce como suma de cuadrados de regresión o suma de cuadrados de error (SSE). Podemos resumir la relación entre las tres sumas de cuadrados con la siguiente ecuación:

g1yj - y22 = g1yN j - y22 + g1 yj - yN j22

SST = SSR + SSE

total sum of squares = explained 1regression2 sum of squares

+ unexplained 1error2 sum of squares

(10.4.1)

2

The coefficient of multiple determination, R y.12 . . . k is obtained by dividing the explained sum of squares by the total sum of squares. That is,

R 2y.12 . . . k =

g1yN j - y22

g1yj - y2

2

=

SSR

SST

(10.4.2)

The subscript y.12 . . . k indicates that in the analysis Y is treated as the dependent variable and the X variables from X1 through Xk are treated as the independent variables. The 2 value of R y.12 . . . k indicates what proportion of the total variation in the observed Y values is explained by the regression of Y on X1, X2, . . . , Xk. In other words, we may say that 2 R y.12

. . . k is a measure of the goodness of fit of the regression surface. This quantity is analogous to r 2, which was computed in Chapter 9.

El subíndice y.12. . . k indica que en el análisis Y se trata como la variable dependiente y las variables X desde X1 hasta Xk se tratan como variables independientes. El valor 2 de R y.12 . . . k indica qué proporción de la variación total en los valores de Y observados se explica por la regresión de Y en X1, X2, . . . , Xk. En otras palabras, podemos decir que 2 R y.12 . . . k es una medida de la bondad de ajuste de la superficie de regresión. Esta cantidad es análoga a r 2, que se calculó en el capítulo 9.

**EXAMPLE 10.4.1**

Refer to Example 10.3.1. Compute R 2y .12.

**Solution**:

For our illustrative example we have in Figure 10.3.1

SST = 1061.36

SSR = 393.39

SSE = 667.97

393.39

= .3706 L .371

R 2y.12 =

1061.36499

10.4 EVALUATING THE MULTIPLE REGRESSION EQUATION

We say that about 37.1 percent of the total variation in the Y values is explained by the fitted regression plane, that is, by the linear relationship with age and education level.

■

**Testing the Regression Hypothesis**

To determine whether the overall regression is significant (that is, to determine whether R 2y.12 is significant), we may perform a hypothesis test as follows.

Para determinar si la regresión general es significativa (es decir, para determinar si R 2y.12 es significativa), podemos realizar una prueba de hipótesis de la siguiente manera.

1. **Data**. The research situation and the data generated by the research are examined to determine if multiple regression is an appropriate technique for analysis.

2. **Assumptions**. We assume that the multiple regression model and its underlying assumptions as presented in Section 10.2 are applicable.

3. **Hypotheses**. In general, the null hypothesis is H0: b 1 = b 2 = b 3 = . . . = b k = 0 and the alternative is HA: not all b i = 0. In words, the null hypothesis states that all the independent variables are of no value in explaining the variation in the Y values.

4. **Test** **statistic**. The appropriate test statistic is V.R., which is computed as part of an analysis of variance. The general ANOVA table is shown as Table 10.4.1. In Table 10.4.1, MSR stands for mean square due to regression and MSE stands for mean square about regression or, as it is sometimes called, the error mean square.

5. Distribution of test statistic. When H0 is true and the assumptions are met, V.R. is distributed as F with k and n - k - 1 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. See Table 10.4.1.

8. Statistical decision. Reject or fail to reject H0 in accordance with the decision rule.

9. Conclusion. If we reject H0, we conclude that, in the population from which the sample was drawn, the dependent variable is linearly related to the independent variables as a group. If we fail to reject H0, we conclude that, in the population from which our sample was drawn, there may be no linear relationship between the dependent variable and the independent variables as a group.

10. p value. We obtain the p value from the table of the F distribution.

We illustrate the hypothesis testing procedure by means of the following example.

TABLE 10.4.1 ANOVA Table for Multiple Regression

SourceSSd.f.MSV.R.

Due to regressionSSR

SSEk

n - k - 1MSR = SSR>k

MSE = SSE>1n - k - 12MSR>MSE

About regressionTotalSSTn - 1500

CHAPTER 10 MULTIPLE REGRESSION AND CORRELATION

**EXAMPLE 10.4.2**

We wish to test the null hypothesis of no linear relationship among the three variables discussed in Example 10.3.1: CDA score, age, and education level.

Deseamos probar la hipótesis nula de que no existe una relación lineal entre las tres variables analizadas en el ejemplo 10.3.1: puntuación CDA, edad y nivel educativo.

Solution:

1. Data. See the description of the data given in Example 10.3.1.

2. Assumptions. We assume that the assumptions discussed in Section 10.2 are met.

3. Hypotheses.

H0: = b 1 = b 2 = 0

HA: = not all b i = 0

4. Test statistic. The test statistic is V.R.

5. Distribution of test statistic. If H0 is true and the assumptions are met, the test statistic is distributed as F with 2 numerator and 68 denominator degrees of freedom.

6. Decision rule. Let us use a significance level of a = .01. The decision rule, then, is reject H0 if the computed value of V.R. is equal to or greater than 4.95 (obtained by interpolation).

7. Calculation of test statistic. The ANOVA for the example is shown in Figure 10.3.1, where we see that the computed value of V.R. is 20.02.

8. Statistical decision. Since 20.02 is greater than 4.95, we reject H0.

9. Conclusion. We conclude that, in the population from which the sample came, there is a linear relationship among the three variables.

10. p value. Since 20.02 is greater than 5.76, the p value for the test is less than .005.

■

**Inferences Regarding Individual B’s**

Frequently, we wish to evaluate the strength of the linear relationship between Y and the independent variables individually. That is, we may want to test the null hypothesis that b i = 0 against the alternative b i Z 0 1i = 1, 2, . . . , k2. The validity of this procedure rests on the assumptions stated earlier: that for each combination of X i values there is a normally distributed subpopulation of Y values with variance s2.

Con frecuencia deseamos evaluar la fuerza de la relación lineal entre Y y las variables independientes individualmente. Es decir, es posible que deseemos probar la hipótesis nula de que b i = 0 contra la alternativa b i Z 0 1i = 1, 2, . . . , k2. La validez de este procedimiento se basa en los supuestos establecidos anteriormente: que para cada combinación de valores X i existe una subpoblación normalmente distribuida de valores Y con varianza s2.

**Hypothesis Tests for the Bi**

To test the null hypothesis that bi is equal to some particular value, say, bi0, the following t statistic may be computed:

Para probar la hipótesis nula de que bi es igual a algún valor particular, digamos bi0, se puede calcular el siguiente estadístico t:

t =

bN i - b i0

sbN i

(10.4.3)

where the degrees of freedom are equal to n - k - 1, and sbN i is the standard deviation

of the bN i.

The standard deviations of the bN i are given as part of the output from most computer software packages that do regression analysis.

Las desviaciones estándar de bN i se dan como parte del resultado de la mayoría de los paquetes de software que realizan análisis de regresión.

**EXAMPLE 10.4.3**

Let us refer to Example 10.3.1 and test the null hypothesis that age (years) is irrelevant in predicting the capacity to direct attention (CDA).

Consultemos el ejemplo 10.3.1 y probemos la hipótesis nula de que la edad (años) es irrelevante para predecir la capacidad de dirigir la atención (CDA).

**Solution**:

1. Data. See Example 10.3.1.

2. Assumptions. See Section 10.2.

3. Hypotheses.

H 0: b 1 = 0

HA: b 1 Z 0

Let a = .05

4. Test statistic. See Equation 10.4.3.

5. Distribution of test statistic. When H0 is true and the assumptions are met, the test statistic is distributed as Student’s t with 68 degrees of freedom.

6. Decision rule. Reject H0 if the computed t is either greater than or equal to 1.9957 (obtained by interpolation) or less than or equal to -1.9957.

7. Calculation of test statistic. By Equation 10.4.3 and data from Figure 10.3.1 we compute

t =

bN 1 - 0

-.18412

=

= -3.80

sbN 1

.04851

8. Statistical decision. The null hypothesis is rejected since the computed value of t, -3.80, is less than -1.9957.

9. Conclusion. We conclude, then, that there is a linear relationship between age and CDA in the presence of education level.

10. p value. For this test, p 6 21.0052 = .01 because -3.80 6 -2.6505 (obtained by interpolation).

■

Now, let us perform a similar test for the second partial regression coefficient, b 2:

H0: b 2 = 0

HA: b 2 Z 0

a = .05

N

b2 - 0

.6108

t =

=

= 4.50

sbN 2

.1357502

CHAPTER 10 MULTIPLE REGRESSION AND CORRELATION

In this case also the null hypothesis is rejected, since 4.50 is greater than 1.9957. We conclude that t ere is a linear relationship between education level and CDA in the presence age, and that education level, used in this manner, is a useful variable for predicting CDA. [For this test, p 6 21.0052 = .01.]

En este caso también se rechaza la hipótesis nula, ya que 4,50 es mayor que 1,9957. Concluimos que existe una relación lineal entre el nivel educativo y el CDA en la edad de presencia, y que el nivel educativo, utilizado de esta manera, es una variable útil para predecir el CDA. [Para esta prueba, p 6 21.0052 = .01.]

**Confidence Intervals for the Bi**

When the researcher has been led to conclude that a partial regression coefficient is not 0, he or she may be interested in obtaining a confidence interval for this bi. Confidence intervals for the bi may be constructed in the usual way by using a value from the t distribution for the reliability factor and standard errors given above.

Cuando el investigador ha llegado a la conclusión de que un coeficiente de regresión parcial no es 0, puede estar interesado en obtener un intervalo de confianza para este bi. Los intervalos de confianza para el bi se pueden construir de la forma habitual utilizando un valor de la distribución t para el factor de confiabilidad y los errores estándar indicados anteriormente.

A 10011 - a2 percent confidence interval for b i is given by

bN i ; t 1-1a> 22,n-k-1sbN i

For our illustrative example we may compute the following 95 percent confidence intervals for b 1 and b 2.

The 95 percent confidence interval for b 1 is

-.18412 ; 1.99571.048512

-.18412 ; .0968

-.28092, -.08732

The 95 percent confidence interval for b 2 is

.6108 ; 11.995721.13572

.6108 ; .2708

.3400, .8816

We may give these intervals the usual probabilistic and practical interpretations. We are 95 percent confident, for example, that b 2 is contained in the interval from .3400 to .8816 since, in repeated sampling, 95 percent of the intervals that may be constructed in this manner will include the true parameter.

Podemos dar a estos intervalos las interpretaciones prácticas y probabilísticas habituales. Por ejemplo, tenemos una confianza del 95 por ciento en que b 2 está contenido en el intervalo de 0,3400 a 0,8816 ya que, en el muestreo repetido, el 95 por ciento de los intervalos que pueden construirse de esta manera incluirán el parámetro verdadero.

**Some Precautions**

One should be aware of the problems involved in carrying out multiple hypothesis tests and constructing multiple confidence intervals from the same sample data. The effect on a of performing multiple hypothesis tests from the same data is discussed in Section 8.2. A similar problem arises when one wishes to construct confidence intervals for two or more partial regression coefficients. The intervals will not be independent, so that the tabulated confidence coefficient does not, in general, apply. In other words, all such intervals would not be 10011 - a2 percent confidence intervals.

Se deben ser conscientes de los problemas que implica realizar múltiples pruebas de hipótesis y construir múltiples intervalos de confianza a partir de los mismos datos muestrales. El efecto sobre a de realizar múltiples pruebas de hipótesis a partir de los mismos datos se analiza en la Sección 8.2. Un problema similar surge cuando se desea construir intervalos de confianza para dos o más coeficientes de regresión parcial. Los intervalos no serán independientes, por lo que, en general, no se aplica el coeficiente de confianza tabulado. En otras palabras, todos esos intervalos no serían 10011, es decir, intervalos de confianza del 2 por ciento.

In order to maintain approximate 10011 - a2 confidence intervals for partial regression coefficients, adjustments must be made to the calculation of errors in the previous equations. These adjustments are sometimes called family-wise error rates, and can be found in many computer software packages. The topic is discussed in detail by Kutner, et al. (4).

Para mantener intervalos de confianza aproximados de 10011 - a2 para los coeficientes de regresión parcial, se deben realizar ajustes en el cálculo de los errores en las ecuaciones anteriores. Estos ajustes a veces se denominan tasas de error familiares y se pueden encontrar en muchos paquetes de software informático. El tema es discutido en detalle por Kutner, et al. (4).

Another problem sometimes encountered in the application of multiple regression is an apparent incompatibility in the results of the various tests of significance that one may perform. In a given problem for a given level of significance, one or the other of the following situations may be observed.

Otro problema que a veces se encuentra en la aplicación de la regresión múltiple es una aparente incompatibilidad en los resultados de las diversas pruebas de significancia que se pueden realizar. En un problema dado para un nivel de significancia dado, se puede observar una u otra de las siguientes situaciones.

1. R 2 and all bN i significant

2. R 2 and some but not all bN i significant

3. R 2 significant but none of the bN i significant

4. All bN i significant but not R 2

5. Some bN i significant, but not all nor R 2

6. Neither R 2 nor any bN i significant

Notice that situation 1 exists in our illustrative example, where we have a significant R 2 and two significant regression coefficients. This situation does not occur in all cases. In fact, situation 2 is very common, especially when a large number of independent variables have been included in the regression equation.

Observe que la situación 1 existe en nuestro ejemplo ilustrativo, donde tenemos un R 2 significativo y dos coeficientes de regresión significativos. Esta situación no se da en todos los casos. De hecho, la situación 2 es muy común, especialmente cuando se ha incluido una gran cantidad de variables independientes en la ecuación de regresión.

**10.5 USING THE MULTIPLE REGRESSION EQUATION**

As we learned in the previous chapter, a regression equation may be used to obtain a computed value of Y, yN when a particular value of X is given. Similarly, we may use our multiple regression equation to obtain a yN value when we are given particular values of the two or more X variables present in the equation.

Como aprendimos en el capítulo anterior, se puede utilizar una ecuación de regresión para obtener un valor calculado de Y, yN cuando se da un valor particular de X. De manera similar, podemos usar nuestra ecuación de regresión múltiple para obtener un valor de yN cuando se nos dan valores particulares de dos o más variables X presentes en la ecuación.

Just as was the case in simple linear regression, we may, in multiple regression, interpret a yN value in one of two ways. First we may interpret yN as an estimate of the mean of the subpopulation of Y values assumed to exist for particular combinations of Xi values. Under this interpretation yN is called an estimate, and when it is used for this purpose, the equation is thought of as an estimating equation. The second interpretation of yN is that it is the value Y is most likely to assume for given values of the Xi. In this case yN is called the predicted value of Y, and the equation is called a prediction equation. In both cases, intervals may be constructed about the yN value when the normality assumption of Section 10.2 holds true. When yN is interpreted as an estimate of a population mean, the interval is called a confidence interval, and when yN is interpreted as a predicted value of Y, the interval is called a prediction interval. Now let us see how each of these intervals is constructed.

Tal como fue el caso en la regresión lineal simple, en la regresión múltiple podemos interpretar un valor de yN de dos maneras. Primero, podemos interpretar yN como una estimación de la media de la subpoblación de valores de Y que se supone existen para combinaciones particulares de valores de Xi. Según esta interpretación, yN se denomina estimación y, cuando se utiliza para este propósito, la ecuación se considera una ecuación de estimación. La segunda interpretación de yN es que es el valor que Y es más probable que asuma para valores dados de Xi. En este caso, yN se denomina valor predicho de Y y la ecuación se denomina ecuación de predicción. En ambos casos, se pueden construir intervalos alrededor del valor de yN cuando se cumple el supuesto de normalidad de la sección 10.2. Cuando yN se interpreta como una estimación de una media poblacional, el intervalo se denomina intervalo de confianza, y cuando yN se interpreta como un valor predicho de Y, el intervalo se denomina intervalo de predicción. Veamos ahora cómo se construye cada uno de estos intervalos.

**The Confidence Interval for the Mean of a Subpopulation of Y Values Given Particular Values of the X i**

We have seen that a 10011 - a2 percent confidence interval for a parameter may be constructed by the general procedure of adding to and subtracting from the estimator a quantity equal to the reliability factor corresponding to 1 - a multiplied by the standard error of the estimator.

Hemos visto que se puede construir un intervalo de confianza de 10011 - a2 por ciento para un parámetro mediante el procedimiento general de sumar y restar al estimador una cantidad igual al factor de confiabilidad correspondiente a 1 - a multiplicado por el error estándar del estimador.

We have also seen that in multiple regression the estimator is

yN j = bN 0 + bN 1x 1j + bN 2 x 2j + . . . + bN k x k j

(10.5.1)

If we designate the standard error of this estimator by syN , the 10011 - a2 percent confidence interval for the mean of Y, given specified X i is as follows:

Si designamos el error estándar de este estimador como syN, el intervalo de confianza de 10011 - a2 por ciento para la media de Y, dado X i especificado, es el siguiente:

yN j ; t 11-a>22,n-k -1syNj

(10.5.2)

**The Prediction Interval for a Particular Value of Y Given Particular Values of the X i**

When we interpret yN as the value Y is most likely to assume when particular values of the Xi are observed, we may construct a prediction interval in the same way in which the confidence interval was constructed. The only difference in the two is the standard error. The standard error of the prediction is slightly larger than the standard error of the estimate, which causes the prediction interval to be wider than the confidence interval.

Cuando interpretamos yN como el valor que Y es más probable que adopte cuando se observan valores particulares de Xi, podemos construir un intervalo de predicción de la misma manera en que se construyó el intervalo de confianza. La única diferencia entre los dos es el error estándar. El error estándar de la predicción es ligeramente mayor que el error estándar de la estimación, lo que hace que el intervalo de predicción sea más amplio que el intervalo de confianza.

If we designate the standard error of the prediction by s y¿N , the 10011 - a2 percent prediction interval is

Si designamos el error estándar de la predicción como s y¿N, el intervalo de predicción de 10011 - a2 por ciento es

yN j ; t 11-a>22, n-k-1s¿yN j

(10.5.3)

The calculations of syN j and s y¿N j in the multiple regression case are complicated and will not be covered in this text. The reader who wishes to see how these statistics are calculated may consult the book by Anderson and Bancroft (3), other references listed at the end of this chapter and Chapter 9, and previous editions of this text. The following example illustrates how MINITAB may be used to obtain confidence intervals for the mean of Y and prediction intervals for a particular value of Y.

Los cálculos de syN j y s y¿N j en el caso de regresión múltiple son complicados y no se tratarán en este texto. El lector que desee ver cómo se calculan estas estadísticas puede consultar el libro de Anderson y Bancroft (3), otras referencias enumeradas al final de este capítulo y en el capítulo 9, y ediciones anteriores de este texto. El siguiente ejemplo ilustra cómo se puede utilizar MINITAB para obtener intervalos de confianza para la media de Y e intervalos de predicción para un valor particular de Y.

**EXAMPLE 10.5.1**

We refer to Example 10.3.1. First, we wish to construct a 95 percent confidence interval for the mean CDA score (Y) in a population of 68-year-old subjects 1X12 who completed 12 years of education 1X22. Second, suppose we have a subject who is 68 years of age and has an education level of 12 years. What do we predict to be this subject’s CDA score?

Nos referimos al Ejemplo 10.3.1. Primero, deseamos construir un intervalo de confianza del 95 por ciento para la puntuación CDA media (Y) en una población de sujetos de 68 años 1X12 que completaron 12 años de educación 1X22. En segundo lugar, supongamos que tenemos un sujeto que tiene 68 años y un nivel educativo de 12 años. ¿Cuál predecimos que será la puntuación CDA de este sujeto?

**Solution**:

The point estimate of the mean CDA score is

yN = 5.494 - .184121682 + .61081122 = .3034

The point prediction, which is the same as the point estimate obtained pre-

viously, also is

yN = 5.494 - .184121682 + .61081122 = .3034

To obtain the confidence interval and the prediction interval for the parameters for which we have just computed a point estimate and a point prediction, we use MINITAB as follows. After entering the information for a regression analysis of our data as shown in Figure 10.3.1, we click on Options in the dialog box. In the box labeled “Prediction intervals for new observations,” we type 68 and 12 and click OK twice. In addition to the regression analysis, we obtain the following output:

Para obtener el intervalo de confianza y el intervalo de predicción para los parámetros para los cuales acabamos de calcular una estimación puntual y una predicción puntual, utilizamos MINITAB de la siguiente manera. Después de ingresar la información para un análisis de regresión de nuestros datos como se muestra en la Figura 10.3.1, hacemos clic en Opciones en el cuadro de diálogo. En el cuadro denominado "Intervalos de predicción para nuevas observaciones", escribimos 68 y 12 y hacemos clic en Aceptar dos veces. Además del análisis de regresión, obtenemos el siguiente resultado:

New Obs

1

Fit

0.303

SE Fit

0.672

95.0% CI

(&1.038, 1.644)

95.0% PI

(&6.093, 6.699)

We interpret these intervals in the usual ways. We look first at the confidence interval. We are 95 percent confident that the interval from -1.038 to 1.644 includes the mean of the subpopulation of Y values for the specified combination of X i values, since this parameter would be included in about 95 percent of the intervals that can be constructed in the manner

shown.

Interpretamos estos intervalos de la forma habitual. Primero nos fijamos en el intervalo de confianza. Tenemos un 95 por ciento de confianza en que el intervalo de -1,038 a 1,644 incluye la media de la subpoblación de valores Y para la combinación especificada de valores X i, ya que este parámetro se incluiría en aproximadamente el 95 por ciento de los intervalos que se pueden construir en el manera mostrado.

Now consider the subject who is 68 years old and has 12 years of education. We are 95 percent confident that this subject would have a CDA score somewhere between -6.093 and 6.699. The fact that the P.I. is wider than the C.I. should not be surprising. After all, it is easier to estimate the mean response than it is estimate an individual observation.

Consideremos ahora al sujeto que tiene 68 años y 12 años de educación. Estamos 95 por ciento seguros de que este tema tendría una puntuación CDA entre -6,093 y 6,699. El hecho de que el P.I. es más ancho que el C.I. No debería sorprendernos. Después de todo, es más fácil estimar la respuesta media que estimar una observación individual.

■

**10.6 THE MULTIPLE CORRELATION MODEL**

We pointed out in the preceding chapter that while regression analysis is concerned with the form of the relationship between variables, the objective of correlation analysis is to gain insight into the strength of the relationship. This is also true in the multivariable case, and in this section we investigate methods for measuring the strength of the relationship among several variables. First, however, let us define the model and assumptions on which our analysis rests.

Señalamos en el capítulo anterior que, si bien el análisis de regresión se ocupa de la forma de la relación entre variables, el objetivo del análisis de correlación es comprender la fuerza de la relación. Esto también es cierto en el caso multivariable, y en esta sección investigamos métodos para medir la fuerza de la relación entre varias variables. Sin embargo, primero definamos el modelo y los supuestos en los que se basa nuestro análisis.

**The Model Equation**

We may write the correlation model as

yj = b 0 + b 1x 1j + b 2 x 2j + Á + b k x kj + Pj

(10.6.1)

where yj is a typical value from the population of values of the variable Y, the b’s are the regression coefficients defined in Section 10.2, and the x ij are particular (known) values of the random variables Xi. This model is similar to the multiple regression model, but there is one important distinction. In the multiple regression model, given in Equation 10.2.1, the Xi are nonrandom variables, but in the multiple correlation model the Xi are random variables. In other words, in the correlation model there is a joint distribution of Y and the Xi that we call a multivariate distribution. Under this model, the variables are no longer thought of as being dependent or independent, since logically they are interchangeable and either of the Xi may play the role of Y.

donde yj es un valor típico de la población de valores de la variable Y, los b son los coeficientes de regresión definidos en la sección 10.2 y los x ij son valores particulares (conocidos) de las variables aleatorias Xi. Este modelo es similar al modelo de regresión múltiple, pero hay una distinción importante. En el modelo de regresión múltiple, dado en la ecuación 10.2.1, las Xi son variables no aleatorias, pero en el modelo de correlación múltiple las Xi son variables aleatorias. En otras palabras, en el modelo de correlación existe una distribución conjunta de Y y Xi que llamamos distribución multivariada. Bajo este modelo ya no se piensa que las variables sean dependientes o independientes, ya que lógicamente son intercambiables y cualquiera de las Xi puede desempeñar el papel de Y.

Typically, random samples of units of association are drawn from a population of interest, and measurements of Y and the Xi are made.

Normalmente, se extraen muestras aleatorias de unidades de asociación de una población de interés y se realizan mediciones de Y y Xi.

A least-squares plane or hyperplane is fitted to the sample data by methods described in Section 10.3, and the same uses may be made of the resulting equation. Inferences may be made about the population from which the sample was drawn if it can be assumed that the underlying distribution is normal, that is, if it can be assumed that the joint distribution of Y and Xi is a multivariate normal distribution. In addition, sample measures of the degree of the relationship among the variables may be computed and, under the assumption that sampling is from a multivariate normal distribution, the corresponding parameters may be estimated by means of confidence intervals, and hypothesis tests may be carried out.

Se ajusta un plano o hiperplano de mínimos cuadrados a los datos de la muestra mediante los métodos descritos en la Sección 10.3, y se pueden hacer los mismos usos de la ecuación resultante. Se pueden hacer inferencias sobre la población de la que se extrajo la muestra si se puede suponer que la distribución subyacente es normal, es decir, si se puede suponer que la distribución conjunta de Y y Xi es una distribución normal multivariada. Además, se pueden calcular medidas muestrales del grado de relación entre las variables y, bajo el supuesto de que el muestreo se realiza a partir de una distribución normal multivariada, se pueden estimar los parámetros correspondientes mediante intervalos de confianza y se pueden realizar pruebas de hipótesis. .

Specifically, we may compute an estimate of the multiple correlation coefficient that measures the dependence between Y and the Xi. This is a straightforward extension of the concept of correlation between two variables that we discuss in Chapter 9. We may also compute partial correlation coefficients that measure the intensity of the relationship between any two variables when the influence of all other variables has been removed.

Específicamente, podemos calcular una estimación del coeficiente de correlación múltiple que mide la dependencia entre Y y Xi. Ésta es una extensión directa del concepto de correlación entre dos variables que analizamos en el capítulo 9. También podemos calcular coeficientes de correlación parcial que miden la intensidad de la relación entre dos variables cualesquiera cuando se ha eliminado la influencia de todas las demás variables.

**The Multiple Correlation Coefficient**

As a first step in analyzing the relationships among the variables, we look at the multiple correlation coefficient.

Como primer paso para analizar las relaciones entre las variables, observamos el coeficiente de correlación múltiple.

The multiple correlation coefficient is the square root of the coefficient of multiple determination and, consequently, the sample value may be computed by taking the

square root of Equation 10.4.2. That is,

El coeficiente de correlación múltiple es la raíz cuadrada del coeficiente de determinación múltiple y, en consecuencia, el valor de la muestra puede calcularse tomando la raíz cuadrada de la Ecuación 10.4.2. Eso es,

R y.12 . . . k =

2R 2y.12 . . . k =

g1yN j - y22

C g1yj - y22

=

SSR

A SST

(10.6.2)

To illustrate the concepts and techniques of multiple correlation analysis, let us consider an example.

**EXAMPLE 10.6.1**

Wang et al. (A-4), using cadaveric human femurs from subjects ages 16 to 19 years, investigated toughness properties of the bone and measures of the collagen network within the bone. Two variables measuring the collagen network are porosity (P, expressed as a percent) and a measure of collagen network tensile strength 1S2. The measure of toughness (W, Newtons), is the force required for bone fracture. The 29 cadaveric femurs used in the study were free from bone-related pathologies. We wish to analyze the nature and strength of the relationship among the three variables. The measurements are shown in the following table.

Wang y cols. (A-4), utilizando fémures humanos cadavéricos de sujetos de entre 16 y 19 años, investigaron las propiedades de dureza del hueso y las medidas de la red de colágeno dentro del hueso. Dos variables que miden la red de colágeno son la porosidad (P, expresada como porcentaje) y una medida de la resistencia a la tracción de la red de colágeno 1S2. La medida de tenacidad (W, Newtons), es la fuerza necesaria para la fractura del hueso. Los 29 fémures cadavéricos utilizados en el estudio estaban libres de patologías relacionadas con los huesos. Deseamos analizar la naturaleza y fuerza de la relación entre las tres variables. Las medidas se muestran en la siguiente tabla.

**Solution**

We use MINITAB to perform the analysis of our data. Readers interested in the derivation of the underlying formulas and the arithmetic procedures involved may consult the texts listed at the end of this chapter and Chapter 9, as well as previous editions of this text. If a least-squares prediction equation and multiple correlation coefficient are desired as part of the analysis, we may obtain them by using the previously described MINITAB multiple regression procedure. When we do this with the sample values of Y, X1, and X2, stored in Columns 1 through 3, respectively, we obtain the output shown in Figure 10.6.1.

Utilizamos MINITAB para realizar el análisis de nuestros datos. Los lectores interesados en la derivación de las fórmulas subyacentes y los procedimientos aritméticos involucrados pueden consultar los textos enumerados al final de este capítulo y el Capítulo 9, así como ediciones anteriores de este texto. Si se desea una ecuación de predicción de mínimos cuadrados y un coeficiente de correlación múltiple como parte del análisis, podemos obtenerlos utilizando el procedimiento de regresión múltiple MINITAB descrito anteriormente. Cuando hacemos esto con los valores de muestra de Y, X1 y X2, almacenados en las columnas 1 a 3, respectivamente, obtenemos el resultado que se muestra en la Figura 10.6.1.

The least-squares equation, then, is

yN j = 35.61 + 1.451x 1j + 2.3960x 2j

This equation may be used for estimation and prediction purposes and may be evaluated by the methods discussed in Section 10.4.

As we see in Figure 10.6.1, the multiple regression output also gives us the coefficient of multiple determination, which, in our present example, is

R 2y.12 = .294

The multiple correlation coefficient, therefore, is

R y.12 = 1.294 = .542

**Interpretation of R y.12**

We interpret R y.12 as a measure of the correlation among the variables force required to fracture, porosity, and collagen network strength in the sample of 29 femur bones from subjects ages 16 to 19. If our data constitute a random sample from the population of such persons, we may use R y.12 as an estimate of ry.12, the true population multiple correlation coefficient. We may also interpret R y.12 as the simple correlation coefficient between yj and yN , the observed and calculated values, respectively, of the “dependent” variable. Perfect correspondence between the observed and calculated values of Y will result in a correlation coefficient of 1, while a complete lack of a linear relationship between observed and calculated values yields a correlation coefficient of 0. The multiple correlation coefficient is always given a positive sign.

Interpretamos R y.12 como una medida de la correlación entre las variables fuerza requerida para fracturar, porosidad y resistencia de la red de colágeno en la muestra de 29 huesos de fémur de sujetos de 16 a 19 años. Si nuestros datos constituyen una muestra aleatoria de la población de dichas personas, podemos utilizar R y.12 como estimación de ry.12, el verdadero coeficiente de correlación múltiple de la población. También podemos interpretar R y.12 como el coeficiente de correlación simple entre yj e yN, los valores observados y calculados, respectivamente, de la variable “dependiente”. Una correspondencia perfecta entre los valores observados y calculados de Y dará como resultado un coeficiente de correlación de 1, mientras que una falta total de una relación lineal entre los valores observados y calculados produce un coeficiente de correlación de 0. Al coeficiente de correlación múltiple siempre se le da un signo positivo.

We may test the null hypothesis that ry.12 . . . k = 0 by computing

F =

R 2y.12 . . . k

1 - R 2y.12 . . . k

#n - k - 1

k

(10.6.3)

The numerical value obtained from Equation 10.6.3 is compared with the tabulated value of F with k and n - k - 1 degrees of freedom. The reader will recall that this is identical to the test of H0: b 1 = b 2 = . . . = b k = 0 described in Section 10.4.

El valor numérico obtenido de la Ecuación 10.6.3 se compara con el valor tabulado de F con k y n - k - 1 grados de libertad. El lector recordará que ésta es idéntica a la prueba de H0: b 1 = b 2 = . . . = b k = 0 descrito en la Sección 10.4.

For our present example let us test the null hypothesis that ry.12 = 0 against the alternative that ry.12 Z 0. We compute

F =

.294 # 29 - 2 - 1

= 5.41

1 - .294

2

Since 5.41 is greater than 4.27, p 6 .025, so that we may reject the null hypothesis at the .025 level of significance and conclude that the force required for fracture is correlated with porosity and the measure of collagen network strength in the sampled population.

The computed value of F for testing H0 that the population multiple correlation coefficient is equal to zero is given in the analysis of variance table in Figure 10.6.1 and is 5.42. The two computed values of F differ as a result of differences in rounding in the intermediate calculations.

■

**Partial Correlation**

The researcher may wish to have a measure of the strength of the linear relationship between two variables when the effect of the remaining variables has been removed. Such a measure is provided by the partial correlation coefficient. For example, the partial sample correlation coefficient ry.12 is a measure of the correlation between Y and X1 after controlling for the effect of X2.

Es posible que el investigador desee tener una medida de la fuerza de la relación lineal entre dos variables cuando se ha eliminado el efecto de las variables restantes. Esta medida la proporciona el coeficiente de correlación parcial. Por ejemplo, el coeficiente de correlación de muestra parcial ry.12 es una medida de la correlación entre Y y X1 después de controlar el efecto de X2.

The partial correlation coefficients may be computed from the simple correlation coefficients. The simple correlation coefficients measure the correlation between two variables when no effort has been made to control other variables. In other words, they are the coefficients for any pair of variables that would be obtained by the methods of simple correlation discussed in Chapter 9.

Los coeficientes de correlación parciales pueden calcularse a partir de los coeficientes de correlación simples. Los coeficientes de correlación simples miden la correlación entre dos variables cuando no se ha hecho ningún esfuerzo por controlar otras variables. En otras palabras, son los coeficientes de cualquier par de variables que se obtendrían mediante los métodos de correlación simple analizados en el capítulo 9.

Suppose we have three variables, Y, X1, and X2. The sample partial correlation coefficient measuring the correlation between Y and X1 after controlling for X2, for example, is written ry1.2. In the subscript, the symbol to the right of the decimal point indicates the variable whose effect is being controlled, while the two symbols to the left of the decimal point indicate which variables are being correlated. For the three-variable case, there are two other sample partial correlation coefficients that we may compute. They are ry2.1 and r12.y.

Supongamos que tenemos tres variables, Y, X1 y X2. El coeficiente de correlación parcial de la muestra que mide la correlación entre Y y X1 después de controlar X2, por ejemplo, se escribe ry1.2. En el subíndice, el símbolo a la derecha del punto decimal indica la variable cuyo efecto se está controlando, mientras que los dos símbolos a la izquierda del punto decimal indican qué variables se están correlacionando. Para el caso de tres variables, hay otros dos coeficientes de correlación parcial muestrales que podemos calcular. Son ry2.1 y r12.y.

**The Coefficient of Partial Determination**

The square of the partial correlation coefficient is called the coefficient of partial determination. It provides useful information about the interrelationships among variables. Consider ry1.2, for example. Its square, r y1.2 tells us what proportion of the remaining variability in Y is explained by X1 after X2 has explained as much of the total variability in Y as it can.

El cuadrado del coeficiente de correlación parcial se llama coeficiente de determinación parcial. Proporciona información útil sobre las interrelaciones entre variables. Considere ry1.2, por ejemplo. Su cuadrado, r y1.2, nos dice qué proporción de la variabilidad restante en Y es explicada por X1 después de que X2 haya explicado la mayor parte posible de la variabilidad total en Y.

**Calculating the Partial Correlation Coefficients**

For three variables the following simple correlation coefficients may be calculated:

ry1, the simple correlation between Y and X1

ry2, the simple correlation between Y and X2

r12, the simple correlation between X1 and X2

The MINITAB correlation procedure may be used to compute these simple correlation coefficients as shown in Figure 10.6.2. As noted earlier, the sample observations are stored in Columns 1 through 3. From the output in Figure 10.6.2 we see that r12 = -.08, ry1 = .043, and ry2 = .535.

The sample partial correlation coefficients that may be computed from the simple correlation coefficients in the three-variable case are:

1. The partial correlation between Y and X1 after controlling for the effect of X2:

ry1.2 = 1ry1 - ry2r122> 211 - r 2y2211 - r 12

2(10.6.4)

2

2

ry2.1 = 1ry2 - ry1r122> 211 - r y1

211 - r 12

2(10.6.5)

2

2

r12.y = 1r12 - ry1ry22> 211 - r y1

211 - r y2

2(10.6.6)

2. The partial correlation between Y and X2 after controlling for the effect of X1:

3. The partial correlation between X1 and X2 after controlling for the effect of Y:10.6 THE MULTIPLE CORRELATION MODEL

511

Dialog box:Session Command:

Stat ➤ Basic Statistics ➤ CorrelationMTB> CORRELATION C1-C3

Type C1-C3 in Variables. Click OK.

Output:

X1

X2

Y

0.043

0.823X1

0.535

0.003-0.080

0.679

Cell Contents: Pearson correlation

P-Value

FIGURE 10.6.2 MINITAB procedure for calculating the simple correlation coefficients for

the data in Table 10.6.1.

**EXAMPLE 10.6.2**

To illustrate the calculation of sample partial correlation coefficients, let us refer to Example 10.6.1, and calculate the partial correlation coefficients among the variables force to fracture 1Y 2, porosity 1X12, and collagen network strength 1X22.

Para ilustrar el cálculo de los coeficientes de correlación parcial de la muestra, consultemos el Ejemplo 10.6.1 y calculemos los coeficientes de correlación parcial entre las variables fuerza de fractura 1Y 2, porosidad 1X12 y resistencia de la red de colágeno 1X22.

**Solution**:

Instead of computing the partial correlation coefficients from the simple correlation coefficients by Equations 10.6.4 through 10.6.6, we use MINITAB to obtain them.

En lugar de calcular los coeficientes de correlación parcial a partir de los coeficientes de correlación simples de las Ecuaciones 10.6.4 a 10.6.6, usamos MINITAB para obtenerlos.

The MINITAB procedure for computing partial correlation coefficients is based on the fact that a given partial correlation coefficient is itself the simple correlation between two sets of residuals. A set of residuals is obtained as follows. Suppose we have measurements on two variables, X (independent) and Y (dependent). We obtain the least-squares prediction equation, yN = bN 0 + bN x. For each value of X we compute a residual, which is equal to 1yi - yN i2, the difference between the observed value of Y and the predicted value of Y associated with the X.

El procedimiento MINITAB para calcular coeficientes de correlación parcial se basa en el hecho de que un coeficiente de correlación parcial dado es en sí mismo la correlación simple entre dos conjuntos de residuos. Un conjunto de residuos se obtiene de la siguiente manera. Supongamos que tenemos medidas de dos variables, X (independiente) e Y (dependiente). Obtenemos la ecuación de predicción de mínimos cuadrados, yN = bN 0 + bN x. Para cada valor de X calculamos un residual, que es igual a 1yi - yN i2, la diferencia entre el valor observado de Y y el valor predicho de Y asociado con X.

Now, suppose we have three variables, X1, X2, and Y. We want to compute the partial correlation coefficient between X1 and Y while holding X2 constant. We regress X1 on X2 and compute the residuals, which we may call residual set A. We regress Y on X2 and compute the residuals, which we may call residual set B. The simple correlation coefficient measuring the strength of the relationship between residual set A and residual set B is the partial correlation coefficient between X1 and Y after controlling for the effect of X2.

Ahora, supongamos que tenemos tres variables, X1, X2 e Y. Queremos calcular el coeficiente de correlación parcial entre X1 e Y manteniendo constante X2. Hacemos una regresión X1 en X2 y calculamos los residuos, que podemos llamar conjunto residual A. Hacemos una regresión Y en X2 y calculamos los residuos, que podemos llamar conjunto residual B. El coeficiente de correlación simple que mide la fuerza de la relación entre el conjunto residual A y el conjunto residual B es el coeficiente de correlación parcial entre X1 e Y después de controlar el efecto de X2.

When using MINITAB we store each set of residuals in a different column for future use in calculating the simple correlation coefficients between them.

Cuando usamos MINITAB, almacenamos cada conjunto de residuos en una columna diferente para usarlos en el futuro al calcular los coeficientes de correlación simples entre ellos.

We use session commands rather than a dialog box to calculate the partial correlation coefficients when we use MINITAB. With the observations on X1, X2, and Y stored in Columns 1 through 3, respectively, the procedure for the data of Table 10.6.1 is shown in Figure 10.6.3. The output shows that ry1.2 = .102, r12.y = -.122, and ry2.1 = .541.

Usamos comandos de sesión en lugar de un cuadro de diálogo para calcular los coeficientes de correlación parcial cuando usamos MINITAB. Con las observaciones de X1, X2 e Y almacenadas en las columnas 1 a 3, respectivamente, el procedimiento para los datos de la Tabla 10.6.1 se muestra en la Figura 10.6.3. El resultado muestra que ry1.2 = .102, r12.y = -.122 y ry2.1 = .541.

Partial correlations can be calculated directly using SPSS software as seen in Figure 10.6.5. This software displays, in a succinct table, both the partial correlation coefficient and the p value associated with each partial correlation.

Las correlaciones parciales se pueden calcular directamente utilizando el software SPSS como se ve en la Figura 10.6.5. Este software muestra, en una tabla sucinta, tanto el coeficiente de correlación parcial como el valor p asociado con cada correlación parcial.

■

**Testing Hypotheses About Partial Correlation Coefficients**

We may test the null hypothesis that any one of the population partial correlation coefficients is 0 by means of the t test. For example, to test H0: ry1.2 . . . k = 0, we compute

t = ry1.2 . . . k

n - k - 1

2

A 1 - r y1.2

...k

(10.6.7)

which is distributed as Student’s t with n - k - 1 degrees of freedom.

Let us illustrate the procedure for our current example by testing H0: ry1.2 = 0

against the alternative, HA: ry1.2 Z 0. The computed t is

t = .102

29 - 2 - 1

= .523

A 1 - 1.10222

Since the computed t of .523 is smaller than the tabulated t of 2.0555 for 26 degrees of freedom and a = .05 (two-sided test), we fail to reject H0 at the .05 level of significance and conclude that there may be no correlation between force required for fracture and porosity after controlling for the effect of collagen network strength. Significance tests for the other two partial correlation coefficients will be left as an exercise for the reader. Note that p values for these tests are calculated by MINITAB as shown in Figure 10.6.3.

Dado que la t calculada de .523 es menor que la t tabulada de 2.0555 para 26 grados de libertad y a = .05 (prueba bilateral), no podemos rechazar H0 al nivel de significancia de .05 y concluimos que puede haber no hay correlación entre la fuerza requerida para la fractura y la porosidad después de controlar el efecto de la resistencia de la red de colágeno. Las pruebas de significancia para los otros dos coeficientes de correlación parcial se dejarán como ejercicio para el lector. Tenga en cuenta que los valores de p para estas pruebas los calcula MINITAB como se muestra en la Figura 10.6.3.

The SPSS statistical software package for the PC provides a convenient procedure for obtaining partial correlation coefficients. To use this feature choose “Analyze” from the menu bar, then “Correlate,” and, finally, “Partial.” Following this sequence of choices the Partial Correlations dialog box appears on the screen. In the box labeled “Variables:,” enter the names of the variables for which partial correlations are desired. In the box labeled “Controlling for:” enter the names of the variable(s) for which you wish to control. Select either a two-tailed or one-tailed level of significance. Unless the option is deselected, actual significance levels will be displayed. For Example 10.6.2, Figure 10.6.4 shows the SPSS computed partial correlation coefficients between the other two variables when controlling, successively, for X1 (porosity), X2 (collagen network strength), and Y (force required for fracture).

El paquete de software estadístico SPSS para PC proporciona un procedimiento conveniente para obtener coeficientes de correlación parcial. Para utilizar esta función, elija "Analizar" en la barra de menú, luego "Correlacionar" y, finalmente, "Parcial". Siguiendo esta secuencia de opciones, aparece en la pantalla el cuadro de diálogo Correlaciones parciales. En el cuadro denominado “Variables:”, ingrese los nombres de las variables para las cuales se desean correlaciones parciales. En el cuadro denominado “Controlando por:” ingrese los nombres de las variables que desea controlar. Seleccione un nivel de significancia de dos colas o de una cola. A menos que se deseleccione la opción, se mostrarán los niveles de significancia reales. Para el ejemplo 10.6.2, la Figura 10.6.4 muestra los coeficientes de correlación parcial calculados por SPSS entre las otras dos variables al controlar, sucesivamente, X1 (porosidad), X2 (resistencia de la red de colágeno) e Y (fuerza requerida para la fractura).