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**

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 X i 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 X i 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.

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

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 + b1 x 1j + b2 x 2j + ... + bk x kj + Pj

(10.2.1)

where y j 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 P j 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 P j 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.

donde y j 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 P j es una variable aleatoria con media 0 y varianza s2, la varianza común de las subpoblaciones de valores 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 P j 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.

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

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

y j = b 0 + b1 x 1j + b2 x 2j + Pj

(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 el 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 - b1 x 1j - b2 x2j

(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 X 1 when X 2 remains unchanged, and b 2 measures the average change in Y for a unit change in X 2 when X 1 remains unchanged. For this reason b 1 and b 2 are referred to as partial regression coefﬁcients.