REGRESSION ANALYSIS: SOME ADDITIONAL TECHNIQUES

**CHAPTER OVERVIEW**

This chapter discusses some additional tools and concepts that are useful in regression analysis. The presentation includes expansions of the basic ideas and techniques of regression analysis that were introduced in Chapters 9 and 10.

En este capítulo se analizan algunas herramientas y conceptos adicionales que resultan útiles en el análisis de regresión. La presentación incluye ampliaciones de las ideas y técnicas básicas del análisis de regresión que se presentaron en los capítulos 9 y 10.

**TOPICS**

11.1 INTRODUCTION

11.2 QUALITATIVE INDEPENDENT VARIABLES

11.3 VARIABLE SELECTION PROCEDURES

11.4 LOGISTIC REGRESSION

11.5 SUMMARY

**LEARNING OUTCOMES**

After studying this chapter, the student will

1. understand how to include qualitative variables in a regression analysis.

1. comprender cómo incluir variables cualitativas en un análisis de regresión.

2. understand how to use automated variable selection procedures to develop regression models.

2. comprender cómo utilizar procedimientos de selección automatizada de variables para desarrollar modelos de regresión.

3. be able to perform logistic regression for dichotomous and polytomous dependent variables.

3. Ser capaz de realizar regresión logística para variables dependientes dicotómicas y politómicas.

**11.1 INTRODUCTION**

The basic concepts and methodology of regression analysis are covered in Chapters 9 and 10. In Chapter 9 we discuss the situation in which the objective is to obtain an equation that can be used to make predictions and estimates about some dependent variable from knowledge of some other single variable that we call the independent, predictor, or explanatory variable. In Chapter 10 the ideas and techniques learned in Chapter 9 are expanded to cover the situation in which it is believed that the inclusion of information on two or more independent variables will yield a better equation for use in making predictions and estimations. Regression analysis is a complex and powerful statistical tool that is widely employed in health sciences research. To do the subject justice requires more space than is available in an introductory statistics textbook. However, for the benefit of those who wish additional coverage of regression analysis, we present in this chapter some additional topics that should prove helpful to the student and practitioner of statistics.

Los conceptos básicos y la metodología del análisis de regresión se cubren en los Capítulos 9 y 10. En el Capítulo 9 analizamos la situación en la que el objetivo es obtener una ecuación que se pueda utilizar para hacer predicciones y estimaciones sobre alguna variable dependiente a partir del conocimiento de alguna otra variable única que llamamos variable independiente, predictora o explicativa. En el Capítulo 10, las ideas y técnicas aprendidas en el Capítulo 9 se amplían para cubrir la situación en la que se cree que la inclusión de información sobre dos o más variables independientes producirá una mejor ecuación para usar en la realización de predicciones y estimaciones. El análisis de regresión es una herramienta estadística compleja y poderosa que se utiliza ampliamente en la investigación de las ciencias de la salud. Hacer justicia al tema requiere más espacio del que está disponible en un libro de texto introductorio de estadística. Sin embargo, para beneficio de aquellos que desean una cobertura adicional del análisis de regresión, presentamos en este capítulo algunos temas adicionales que deberían resultar útiles para el estudiante y el profesional de la estadística.

**Regression Assumptions Revisited** As we learned in Chapters 9 and 10, there are several assumptions underlying the appropriate use of regression procedures. Often there are certain measurements that strongly influence the shape of a distribution or impact the magnitude of the variance of a measured variable. Other times, certain independent variables that are being used to develop a model are highly correlated, leading to the development of a model that may not be unique or correct.

**Supuestos de regresión revisados** ​​Como aprendimos en los capítulos 9 y 10, existen varios supuestos que sustentan el uso adecuado de los procedimientos de regresión. A menudo, hay ciertas mediciones que influyen fuertemente en la forma de una distribución o impactan la magnitud de la varianza de una variable medida. Otras veces, ciertas variables independientes que se utilizan para desarrollar un modelo están altamente correlacionadas, lo que lleva al desarrollo de un modelo que puede no ser único o correcto.

Non-Normal Data Many times the data that are used to build a regression model are not normally distributed. One may wish to explore the possibility that some of the observed data points are outliers or that they disproportionately affect the distribution of the data. Such an investigation may be accomplished informally by constructing a scatter plot and looking for observations that do not seem to fit with the others. Alternatively, many computer packages produce formal tests to evaluate potential outlying observations in either the dependent variable or the independent variables. It is always up to the researcher, however, to justify which observations are to be removed from the data set prior to analysis.

**Datos no normales** Muchas veces los datos que se utilizan para construir un modelo de regresión no se distribuyen normalmente. Uno puede desear explorar la posibilidad de que algunos de los puntos de datos observados sean valores atípicos o que afecten desproporcionadamente la distribución de los datos. Tal investigación puede llevarse a cabo de manera informal construyendo un diagrama de dispersión y buscando observaciones que no parezcan encajar con las demás. Alternativamente, muchos paquetes de computadora producen pruebas formales para evaluar posibles observaciones atípicas en la variable dependiente o en las variables independientes. Sin embargo, siempre depende del investigador justificar qué observaciones se eliminarán del conjunto de datos antes del análisis.

Often one may wish to attempt a transformation of the data. Mathematical transformations are useful because they do not affect the underlying relationships among variables. Since hypothesis tests for the regression coefficients are based on normal distribution statistics, data transformations can sometimes normalize the data to the extent necessary to perform such tests. Simple transformations, such as taking the square root of measurements or taking the logarithm of measurements, are quite common.

A menudo, se puede intentar una transformación de los datos. Las transformaciones matemáticas son útiles porque no afectan las relaciones subyacentes entre las variables. Dado que las pruebas de hipótesis para los coeficientes de regresión se basan en estadísticas de distribución normal, las transformaciones de datos a veces pueden normalizar los datos en la medida necesaria para realizar dichas pruebas. Las transformaciones simples, como tomar la raíz cuadrada de las mediciones o tomar el logaritmo de las mediciones, son bastante comunes.

**EXAMPLE 11.1.1**

Researchers were interested in blood concentrations of delta-9-tetrahydrocannabinol ($\Delta$-9-THC), the active psychotropic component in marijuana, from 25 research subjects. These data are presented in Table 11.1.1, as are these same data after using a log10 transformation.

Los investigadores se interesaron por las concentraciones sanguíneas de delta-9-tetrahidrocannabinol ()-9-THC), el componente psicotrópico activo de la marihuana, de 25 sujetos de investigación. Estos datos se presentan en la Tabla 11.1.1, al igual que estos mismos datos después de utilizar una transformación log10.

Box-and-whisker plots from SPSS software for these data are shown in Figure 11.1.1. The raw data are clearly skewed, and an outlier is identified (observation 25). A log10 transformation, which is often useful for such skewed data, removes the magnitude of the outlier and results in a distribution that is much more nearly symmetric about the median. Therefore, the transformed data could be used in lieu of the raw data for constructing the regression model. Though symmetric data do not, necessarily, imply that the data are normal, they do result in a more appropriate model. Formal tests of normality, as previously mentioned, should always be carried out prior to analysis.

En la Figura 11.1.1 se muestran los diagramas de caja y bigotes del software SPSS para estos datos. Los datos brutos están claramente sesgados y se identifica un valor atípico (observación 25). Una transformación log10, que suele ser útil para estos datos sesgados, elimina la magnitud del valor atípico y da como resultado una distribución que es mucho más simétrica respecto de la mediana. Por lo tanto, los datos transformados podrían utilizarse en lugar de los datos brutos para construir el modelo de regresión. Aunque los datos simétricos no implican necesariamente que sean normales, sí dan como resultado un modelo más apropiado. Las pruebas formales de normalidad, como se mencionó anteriormente, siempre deben realizarse antes del análisis.

**Unequal Error Variances** When the variances of the error terms are not equal, we may obtain a satisfactory equation for the model, but, because the assumption that the error variances are equal is violated, we will not be able to perform appropriate hypothesis tests on the model coefficients. Just as was the case in overcoming the non-normality problem, transformations of the regression variables may reduce the impact of unequal error variances.

**Varianzas de error desiguales** Cuando las varianzas de los términos de error no son iguales, podemos obtener una ecuación satisfactoria para el modelo, pero, debido a que se viola el supuesto de que las varianzas de error son iguales, no podremos realizar pruebas de hipótesis adecuadas sobre los coeficientes del modelo. Tal como sucedió al superar el problema de no normalidad, las transformaciones de las variables de regresión pueden reducir el impacto de las varianzas de error desiguales.

**Correlated Independent Variables** Multicollinearity is a common problem that arises when one attempts to build a model using many independent variables. Multicollinearity occurs when there is a high degree of correlation among the independent variables. For example, imagine that we want to find an equation relating height and weight to blood pressure. A common variable that is derived from height and weight is called the body mass index (BMI). If we attempt to find an equation relating height, weight, and BMI to blood pressure, we can expect to run into analytical problems because BMI, by definition, is highly correlated with both height and weight.

**Variables independientes correlacionadas** La multicolinealidad es un problema común que surge cuando se intenta construir un modelo utilizando muchas variables independientes. La multicolinealidad ocurre cuando hay un alto grado de correlación entre las variables independientes. Por ejemplo, imaginemos que queremos encontrar una ecuación que relacione la altura y el peso con la presión arterial. Una variable común que se deriva de la altura y el peso se llama índice de masa corporal (IMC). Si intentamos encontrar una ecuación que relacione la altura, el peso y el IMC con la presión arterial, podemos esperar encontrarnos con problemas analíticos porque el IMC, por definición, está altamente correlacionado tanto con la altura como con el peso.

The problem arises mathematically when the solutions for the regression coefficients are derived. Since the data are correlated, solutions may not be found that are unique to a given model. The least complex solution to multicollinearity is to calculate correlations among all of the independent variables and to retain only those variables that are not highly correlated. A conservative rule of thumb to remove redundancy in the data set is to eliminate variables that are related to others with a significant correlation coefficient above 0.7.

El problema surge matemáticamente cuando se derivan las soluciones para los coeficientes de regresión. Como los datos están correlacionados, es posible que no se encuentren soluciones que sean exclusivas de un modelo determinado. La solución menos compleja para la multicolinealidad es calcular las correlaciones entre todas las variables independientes y conservar solo aquellas variables que no están altamente correlacionadas. Una regla práctica conservadora para eliminar la redundancia en el conjunto de datos es eliminar las variables que están relacionadas con otras con un coeficiente de correlación significativo superior a 0,7.

**EXAMPLE 11.1.2**

A study of obesity and metabolic syndrome used data collected from 15 students, and included systolic blood pressure (SBP), weight, and BMI. These data are presented in Table 11.1.2.

En un estudio sobre la obesidad y el síndrome metabólico se utilizaron datos recopilados de 15 estudiantes, entre los que se incluían la presión arterial sistólica (PAS), el peso y el IMC. Estos datos se presentan en la Tabla 11.1.2.

Correlations for the three variables are shown in Figure 11.1.2. The very large and significant correlation between the variables weight and BMI suggests that including both of these variables in the model is inappropriate because of the high level of redundancy in the information provided by these variables. This makes logical sense since BMI is a function of weight. The researcher is now faced with the task of deciding which of the variables to retain for constructing the regression model.

Las correlaciones entre las tres variables se muestran en la Figura 11.1.2. La correlación muy grande y significativa entre las variables peso e IMC sugiere que la inclusión de ambas variables en el modelo es inadecuada debido al alto nivel de redundancia en la información proporcionada por estas variables. Esto tiene sentido lógico ya que el IMC es una función del peso. El investigador ahora se enfrenta a la tarea de decidir cuál de las variables conservar para construir el modelo de regresión.

**11.2 QUALITATIVE INDEPENDENT VARIABLES**

The independent variables considered in the discussion in Chapter 10 were all quantitative; that is, they yielded numerical values that were either counts or measurements in the usual sense of the word. For example, some of the independent variables used in our examples and exercises were age, education level, collagen porosity, and collagen tensile strength. Frequently, however, it is desirable to use one or more qualitative variables as independent variables in the regression model. Qualitative variables, it will be recalled, are those variables whose “values” are categories and that convey the concept of attribute rather than amount or quantity. The variable marital status, for example, is a qualitative variable whose categories are “single,” “married,” “widowed,” and “divorced.” Other examples of qualitative variables include sex (male or female), diagnosis, race, occupation, and immunity status to some disease. In certain situations an investigator may suspect that including one or more variables such as these in the regression equation would contribute significantly to the reduction of the error sum of squares and thereby provide more precise estimates of the parameters of interest.

Las variables independientes consideradas en el análisis del Capítulo 10 fueron todas cuantitativas; es decir, produjeron valores numéricos que eran recuentos o mediciones en el sentido habitual de la palabra. Por ejemplo, algunas de las variables independientes utilizadas en nuestros ejemplos y ejercicios fueron la edad, el nivel de educación, la porosidad del colágeno y la resistencia a la tensión del colágeno. Sin embargo, con frecuencia es deseable utilizar una o más variables cualitativas como variables independientes en el modelo de regresión. Las variables cualitativas, se recordará, son aquellas variables cuyos "valores" son categorías y que transmiten el concepto de atributo en lugar de cantidad. La variable estado civil, por ejemplo, es una variable cualitativa cuyas categorías son "soltero", "casado", "viudo" y "divorciado". Otros ejemplos de variables cualitativas incluyen sexo (masculino o femenino), diagnóstico, raza, ocupación y estado de inmunidad a alguna enfermedad. En ciertas situaciones, un investigador puede sospechar que incluir una o más variables como estas en la ecuación de regresión contribuiría significativamente a la reducción de la suma de cuadrados del error y, por lo tanto, proporcionaría estimaciones más precisas de los parámetros de interés.

Suppose, for example, that we are studying the relationship between the dependent variable systolic blood pressure and the independent variables weight and age. We might also want to include the qualitative variable sex as one of the independent variables. Or suppose we wish to gain insight into the nature of the relationship between lung capacity and other relevant variables. Candidates for inclusion in the model might consist of such quantitative variables as height, weight, and age, as well as qualitative variables such as sex, area of residence (urban, suburban, rural), and smoking status (current smoker, ex-smoker, never smoked).

Supongamos, por ejemplo, que estamos estudiando la relación entre la variable dependiente presión arterial sistólica y las variables independientes peso y edad. También podríamos querer incluir la variable cualitativa sexo como una de las variables independientes. O supongamos que deseamos comprender mejor la naturaleza de la relación entre la capacidad pulmonar y otras variables relevantes. Las variables candidatas para su inclusión en el modelo podrían consistir en variables cuantitativas como la altura, el peso y la edad, así como variables cualitativas como el sexo, la zona de residencia (urbana, suburbana, rural) y el estado de tabaquismo (fumador actual, exfumador, nunca ha fumado).

Dummy Variables In order to incorporate a qualitative independent variable in the multiple regression model, it must be quantified in some manner. This may be accomplished through the use of what are known as dummy variables.

**Variables ficticias** Para incorporar una variable independiente cualitativa en el modelo de regresión múltiple, es necesario cuantificarla de alguna manera. Esto se puede lograr mediante el uso de lo que se conoce como variables ficticias.

**DEFINITION**

A dummy variable is a variable that assumes only a finite number of values (such as 0 or 1) for the purpose of identifying the different categories of a qualitative variable.

Una variable ficticia es una variable que asume solo un número finito de valores (como 0 o 1) con el propósito de identificar las diferentes categorías de una variable cualitativa.

The term “dummy” is used to indicate the fact that the numerical values (such as 0 and 1) assumed by the variable have no quantitative meaning but are used merely to identify different categories of the qualitative variable under consideration. Qualitative variables are sometimes called indicator variables, and when there are only two categories, they are sometimes called dichotomous variables.

El término “variable ficticia” se utiliza para indicar que los valores numéricos (como 0 y 1) que asume la variable no tienen un significado cuantitativo, sino que se utilizan simplemente para identificar diferentes categorías de la variable cualitativa en cuestión. Las variables cualitativas a veces se denominan variables indicadoras y, cuando solo hay dos categorías, se las denomina variables dicotómicas.

The following are some examples of qualitative variables and the dummy variables used to quantify them:

A continuación se presentan algunos ejemplos de variables cualitativas y las variables ficticias utilizadas para cuantificarlas:

**Qualitative Variable**

Sex (male, female):

$$

x\_1= \left\{ \begin{array}{cl}

1 & \text{ for male } \\

0 & \text{ for female }

\end{array} \right.

$$

Place of residence (urban, rural, suburban):

$$

x\_1= \left\{ \begin{array}{cl}

1 & \text{ for urban } \\

0 & \text{ for rural and suburban }

\end{array} \right.

$$

$$

x\_2= \left\{ \begin{array}{cl}

1 & \text{ for rural } \\

0 & \text{ for urban and suburban }

\end{array} \right.

$$

Smoking status [current smoker, ex-smoker

(has not smoked for 5 years or less), ex-smoker

(has not smoked for more than 5 years), never

smoked]:

$$

x\_1= \left\{ \begin{array}{cl}

1 & \text{ for current smoker } \\

0 & \text{ for otherwise }

\end{array} \right.

$$

$$

x\_2= \left\{ \begin{array}{cl}

1 & \text{ for ex-smoker (\leq 5 years) } \\

0 & \text{ for otherwise }

\end{array} \right.

$$

$$

x\_3= \left\{ \begin{array}{cl}

1 & \text{ for ex-smoker (\> 5 years) } \\

0 & \text{ for otherwise }

\end{array} \right.

$$

Note in these examples that when the qualitative variable has k categories, k - 1 dummy variables must be defined for all the categories to be properly coded. This rule is applicable for any multiple regression containing an intercept constant. The variable sex, with two categories, can be quantified by the use of only one dummy variable, while three dummy variables are required to quantify the variable smoking status, which has four categories.

Obsérvese en estos ejemplos que, cuando la variable cualitativa tiene k categorías, se deben definir k - 1 variables ficticias para que todas las categorías se codifiquen correctamente. Esta regla es aplicable a cualquier regresión múltiple que contenga una constante de intersección. La variable sexo, con dos categorías, se puede cuantificar mediante el uso de una sola variable ficticia, mientras que se requieren tres variables ficticias para cuantificar la variable estado de tabaquismo, que tiene cuatro categorías.

The following examples illustrate some of the uses of qualitative variables in multiple regression. In the first example we assume that there is no interaction between the independent variables. Since the assumption of no interaction is not realistic in many instances, we illustrate, in the second example, the analysis that is appropriate when interaction between variables is accounted for.

Los siguientes ejemplos ilustran algunos de los usos de las variables cualitativas en la regresión múltiple. En el primer ejemplo, suponemos que no hay interacción entre las variables independientes. Dado que el supuesto de que no hay interacción no es realista en muchos casos, ilustramos, en el segundo ejemplo, el análisis que resulta adecuado cuando se tiene en cuenta la interacción entre variables.

**EXAMPLE 11.2.1**

In a study of factors thought to be associated with birth weight, a simple random sample of 100 birth records was selected from the North Carolina 2001 Birth Registry (A-1). Table 11.2.1 shows, for three variables, the data extracted from each record. There are two independent variables: length of gestation (weeks), which is quantitative, and smoking status of mother (smoke), a qualitative variable. The dependent variable is birth weight (grams).

En un estudio de factores que se cree que están asociados con el peso al nacer, se seleccionó una muestra aleatoria simple de 100 registros de nacimientos del Registro de nacimientos de Carolina del Norte de 2001 (A-1). La Tabla 11.2.1 muestra, para tres variables, los datos extraídos de cada registro. Hay dos variables independientes: la duración de la gestación (semanas), que es cuantitativa, y el tabaquismo de la madre (fuma), una variable cualitativa. La variable dependiente es el peso al nacer (gramos).

TABLE 11.2.1 Data from a Simple Random Sample of 100 Births from the North Carolina Birth Registry, Example 11.2.1

**Solution:**

For the analysis, we quantify smoking status by means of a dummy variable that is coded 1 if the mother is a smoker and 0 if she is a nonsmoker. The data in Table 11.2.1 are plotted as a scatter diagram in Figure 11.2.1. The scatter diagram suggests that, in general, longer periods of gestation are associated with larger birth weights.

Para el análisis, cuantificamos el estado de tabaquismo mediante una variable ficticia que se codifica como 1 si la madre es fumadora y como 0 si no fuma. Los datos de la Tabla 11.2.1 se representan como un diagrama de dispersión en la Figura 11.2.1. El diagrama de dispersión sugiere que, en general, los períodos más largos de gestación se asocian con pesos al nacer más elevados.

To obtain additional insight into the nature of these data, we may enter them into a computer and employ an appropriate program to perform further analyses. For example, we enter the observations y\_1 = 3147, x 11 = 40, x 21 = 0, for the first case; y2 = 2977, x 12 = 41, x 22 = 0 for the second case; and so on. Figure 11.2.2 shows the computer output obtained with the use of the MINITAB multiple regression program.

Para obtener más información sobre la naturaleza de estos datos, podemos introducirlos en una computadora y emplear un programa adecuado para realizar análisis adicionales. Por ejemplo, introducimos las observaciones y1 = 3147, x 11 = 40, x 21 = 0, para el primer caso; y2 = 2977, x 12 = 41, x 22 = 0 para el segundo caso; y así sucesivamente. La figura 11.2.2 muestra el resultado de la computadora obtenido con el uso del programa de regresión múltiple MINITAB.

We see in the printout that the multiple regression equation is

Vemos en la impresión que la ecuación de regresión múltiple es

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

yN j = -1724 + 130x 1j - 294bN 2x 2j

(11.2.1)

To observe the effect on this equation when we wish to consider only the births to smoking mothers, we let x2j = 1. The equation then becomes

Para observar el efecto en esta ecuación cuando deseamos considerar solo los nacimientos de madres fumadoras, hacemos x2j = 1. La ecuación entonces se convierte en

yN j = -1724 + 130x1j - 294112

= -2018 + 130x1j

(11.2.2)

which has a y -intercept of -2018 and a slope of 130. Note that the y-intercept for the new equation is equal to 1bN 0 + bN 12 = 3-1724 + 1-29424 = -2018.

que tiene una intersección con el eje y de -2018 y una pendiente de 130. Nótese que la intersección con el eje y para la nueva ecuación es igual a 1bN 0 + bN 12 = 3-1724 + 1-29424 = -2018.

Now let us consider only births to nonsmoking mothers. When we let x 2 = 0, our regression equation reduces to

Consideremos ahora sólo los nacimientos de madres no fumadoras. Cuando hacemos que x 2 = 0, nuestra ecuación de regresión se reduce a

yNj = -1724 + 130x1j - 294102

= -1724 + 130x1j

(11.2.3)

The slope of this equation is the same as the slope of the equation for smoking mothers, but the y-intercepts are different. The y-intercept for the equation associated with nonsmoking mothers is larger than the one for the smoking mothers. These results show that for this sample, babies born to mothers who do not smoke weighed, on the average, more than babies born to mothers who do smoke, when length of gestation is taken into account. The amount of the difference, on the average, is 294 grams. Stated another way, we can say that for this sample, babies born to mothers who smoke weighed, on the average, 294 grams less than the babies born to mothers who do not smoke, when length of gestation is taken into account. Figure 11.2.3 shows the scatter diagram of the original data along with a plot of the two regression lines (Equations 11.2.2 and 11.2.3).

La pendiente de esta ecuación es la misma que la pendiente de la ecuación para las madres fumadoras, pero los puntos de corte en y son diferentes. El punto de corte en y para la ecuación asociada con las madres no fumadoras es mayor que el de las madres fumadoras. Estos resultados muestran que para esta muestra, los bebés nacidos de madres que no fuman pesaron, en promedio, más que los bebés nacidos de madres que sí fuman, cuando se toma en cuenta la duración de la gestación. La cantidad de la diferencia, en promedio, es de 294 gramos. Dicho de otra manera, podemos decir que para esta muestra, los bebés nacidos de madres que fuman pesaron, en promedio, 294 gramos menos que los bebés nacidos de madres que no fuman, cuando se toma en cuenta la duración de la gestación. La Figura 11.2.3 muestra el diagrama de dispersión de los datos originales junto con un gráfico de las dos líneas de regresión (Ecuaciones 11.2.2 y 11.2.3).

**EXAMPLE 11.2.2**

At this point a question arises regarding what inferences we can make about the sampled population on the basis of the sample results obtained in Example 11.2.1. First of all, we wish to know if the sample difference of 294 grams is significant. In other words, does smoking have an effect on birth weight? We may answer this question through the following hypothesis testing procedure.

En este punto surge una pregunta sobre qué inferencias podemos hacer sobre la población muestreada en base a los resultados de la muestra obtenidos en el Ejemplo 11.2.1. En primer lugar, deseamos saber si la diferencia de muestra de 294 gramos es significativa. En otras palabras, ¿el tabaquismo tiene un efecto sobre el peso al nacer? Podemos responder a esta pregunta mediante el siguiente procedimiento de prueba de hipótesis.

Solution:

1. Data. The data are as given in Example 11.2.1.

2. Assumptions. We presume that the assumptions underlying multiple regression analysis are met.

2. Supuestos. Suponemos que se cumplen los supuestos que fundamentan el análisis de regresión múltiple.

3. Hypotheses. H0 : b 2 = 0; HA : b 2 Z 0. Suppose we let a = .05.

4. Test statistic. The test statistic is t = 1bN 2 - 02>sbN .2

5. Distribution of test statistic. When the assumptions are met and H0 is true the test statistic is distributed as Student’s t with 97 degrees of freedom.

5. Distribución del estadístico de prueba. Cuando se cumplen los supuestos y H0 es verdadera, el estadístico de prueba se distribuye como una t de Student con 97 grados de libertad.

6. Decision rule. We reject H0 if the computed t is either greater than or equal to 1.9848 or less than or equal to -1.9848 (obtained by interpolation).

6. Regla de decisión. Rechazamos H0 si el valor t calculado es mayor o igual a 1,9848 o menor o igual a -1,9848 (obtenido por interpolación).

7. Calculation of test statistic. The calculated value of the test statistic appears in Figure 11.2.2 as the t ratio for the coefficient associated with the variable appearing in Column 3 of Table 11.2.1. This coefficient, of course, is bN 2. We see that the computed t is -2.17.

7. Cálculo del estadístico de prueba. El valor calculado del estadístico de prueba aparece en la Figura 11.2.2 como el cociente t del coeficiente asociado con la variable que aparece en la Columna 3 de la Tabla 11.2.1. Este coeficiente, por supuesto, es bN 2. Vemos que el t calculado es -2,17.

8. Statistical decision. Since -2.17 6 -1.9848, we reject H0.

9. Conclusion. We conclude that, in the sampled population, whether the mothers smoke is associated with a reduction in the birth weights of their babies.

9. Conclusión. Concluimos que, en la población estudiada, el hecho de que las madres fumen o no está asociado con una reducción del peso de sus bebés al nacer.

10. p value. For this test we have p = .033 from Figure 11.2.2.

A Confidence Interval for B 2 Given that we are able to conclude that in the sampled population the smoking status of the mothers does have an effect on the birth weights of their babies, we may now inquire as to the magnitude of the effect. Our best point estimate of the average difference in birth weights, when length of gestation is taken into account, is 294 grams in favor of babies born to mothers who do not smoke. We may obtain an interval estimate of the mean amount of the difference by using information from the computer printout by means of the following expression:

Un intervalo de confianza para B 2 Dado que podemos concluir que en la población muestreada el estado de tabaquismo de las madres tiene un efecto sobre el peso al nacer de sus bebés, ahora podemos preguntarnos por la magnitud del efecto. Nuestra mejor estimación puntual de la diferencia promedio en el peso al nacer, cuando se toma en cuenta la duración de la gestación, es de 294 gramos a favor de los bebés nacidos de madres que no fuman. Podemos obtener una estimación de intervalo de la cantidad media de la diferencia utilizando la información de la impresión de la computadora mediante la siguiente expresión:

bN 2 ; tsbN 2

For a 95 percent confidence interval, we have

Para un intervalo de confianza del 95 por ciento, tenemos

-294.4 ; 1.98481135.82

-563.9, -24.9

Thus, we are 95 percent confident that the difference is somewhere between about 564 grams and 25 grams.

Por lo tanto, estamos 95 por ciento seguros de que la diferencia está entre 564 gramos y 25 gramos.

Advantages of Dummy Variables The reader may have correctly surmised that an alternative analysis of the data of Example 11.2.1 would consist of fitting two separate regression equations: one to the subsample of mothers who smoke and another to the subsample of those who do not. Such an approach, however, lacks some of the advantages of the dummy variable technique and is a less desirable procedure when the latter procedure is valid. If we can justify the assumption that the two separate regression lines have the same slope, we can get a better estimate of this common slope through the use of dummy variables, which entails pooling the data from the two subsamples. In Example 11.2.1 the estimate using a dummy variable is based on a total sample size of 100 observations, whereas separate estimates would be based on a sample of 85 smokers and only 15 nonsmokers. The dummy variables approach also yields more precise inferences regarding other parameters since more degrees of freedom are available for the calculation of the error mean square.

Ventajas de las variables ficticias El lector puede haber supuesto correctamente que un análisis alternativo de los datos del Ejemplo 11.2.1 consistiría en ajustar dos ecuaciones de regresión separadas: una a la submuestra de madres que fuman y otra a la submuestra de las que no fuman. Sin embargo, este enfoque carece de algunas de las ventajas de la técnica de la variable ficticia y es un procedimiento menos deseable cuando este último procedimiento es válido. Si podemos justificar el supuesto de que las dos líneas de regresión separadas tienen la misma pendiente, podemos obtener una mejor estimación de esta pendiente común mediante el uso de variables ficticias, lo que implica agrupar los datos de las dos submuestras. En el Ejemplo 11.2.1, la estimación utilizando una variable ficticia se basa en un tamaño de muestra total de 100 observaciones, mientras que las estimaciones separadas se basarían en una muestra de 85 fumadores y solo 15 no fumadores. El enfoque de las variables ficticias también produce inferencias más precisas con respecto a otros parámetros, ya que hay más grados de libertad disponibles para el cálculo del cuadrado medio del error.

Use of Dummy Variables: Interaction Present Now let us consider the situation in which interaction between the variables is assumed to be present. Suppose, for example, that we have two independent variables: one quantitative variable X1 and one qualitative variable with three response levels yielding the two dummy variables X2 and X3. The model, then, would be

Uso de variables ficticias: interacción presente Ahora consideremos la situación en la que se supone que existe interacción entre las variables. Supongamos, por ejemplo, que tenemos dos variables independientes: una variable cuantitativa X1 y una variable cualitativa con tres niveles de respuesta que dan como resultado las dos variables ficticias X2 y X3. El modelo, entonces, sería

yj = b 0 + b 1X1j + b 2X2j + b 3X3j + b 4X1j X2j + b 5X1j X3j + Pj

(11.2.4)

in which b 4X1j X2j and b 5X1j X3j are called interaction terms and represent the interaction between the quantitative and the qualitative independent variables. Note that there is no need to include in the model the term containing X2j X3j ; it will always be zero because when X2 = 1, X3 = 0, and when X3 = 1, X2 = 0. The model of Equation 11.2.4 allows for a different slope and Y-intercept for each level of the qualitative variable.

donde b 4X1j X2j y b 5X1j X3j se denominan términos de interacción y representan la interacción entre las variables independientes cuantitativas y cualitativas. Nótese que no es necesario incluir en el modelo el término que contiene X2j X3j ; siempre será cero porque cuando X2 = 1, X3 = 0, y cuando X3 = 1, X2 = 0. El modelo de la ecuación 11.2.4 permite una pendiente y una intersección con el eje Y diferentes para cada nivel de la variable cualitativa.

Suppose we use dummy variable coding to quantify the qualitative variable as follows:

Supongamos que utilizamos codificación de variable ficticia para cuantificar la variable cualitativa de la siguiente manera:

1 for level 1

0 otherwise

1 for level 2

X3 = e

0 otherwise

X2 = e547

The three sample regression equations for the three levels of the qualitative variable, then, are as follows:

Las tres ecuaciones de regresión de muestra para los tres niveles de la variable cualitativa son, entonces, las siguientes:

Level 1 (X2 " 1, X3 " 0)

yNj = bN 0 + bN 1x1j + bN 2112 + bN 3102 + bN 4x1j 112 + bN 5x1j 102

= bN 0 + bN 1x1j + bN 2 + bN 4 x1j

= 1bN 0 + bN 22 + 1bN 1 + bN 42x1j

(11.2.5)

Level 2 (X2 " 0, X3 " 1)

yNj = bN 0 + bN 1x1j + bN 2102 + bN 3112 + bN 4x1j 102 + bN 5x1j 112

= bN 0 + bN 1x 1j + bN 3 + bN 5 x 1j

= 1bN 0 + bN 32 + 1bN 1 + bN 52x1j

(11.2.6)

Level 3 (X2 " 0, X3 " 0)

yNj = bN 0 + bN 1x1j + bN 2102 + bN 3102 + bN 4 x1j 102 + bN 5 x1j 102

= bN 0 + bN 1x1j

(11.2.7)

Let us illustrate these results by means of an example.

**EXAMPLE 11.2.3**

A team of mental health researchers wishes to compare three methods (A, B, and C) of treating severe depression. They would also like to study the relationship between age and treatment effectiveness as well as the interaction (if any) between age and treatment.

Un equipo de investigadores de salud mental desea comparar tres métodos (A, B y C) para tratar la depresión grave. También les gustaría estudiar la relación entre la edad y la eficacia del tratamiento, así como la interacción (si la hay) entre la edad y el tratamiento.

Each member of a simple random sample of 36 patients, comparable with respect to diagnosis and severity of depression, was randomly assigned to receive treatment A, B, or C. The results are shown in Table 11.2.2. The dependent variable Y is treatment effectiveness, the quantitative independent variable X1 is patient’s age at nearest birthday, and the independent variable type of treatment is a qualitative variable that occurs at three levels. The following dummy variable coding is used to quantify the qualitative variable:

Cada miembro de una muestra aleatoria simple de 36 pacientes, comparables con respecto al diagnóstico y la gravedad de la depresión, fue asignado aleatoriamente para recibir el tratamiento A, B o C. Los resultados se muestran en la Tabla 11.2.2. La variable dependiente Y es la efectividad del tratamiento, la variable independiente cuantitativa X1 es la edad del paciente en el cumpleaños más cercano y la variable independiente tipo de tratamiento es una variable cualitativa que ocurre en tres niveles. La siguiente codificación de variable ficticia se utiliza para cuantificar la variable cualitativa:

1 for treatment A

0 otherwise

1 for treatment B

X3 = e

0 otherwise

X2 = e

The scatter diagram for these data is shown in Figure 11.2.4. Table 11.2.3 shows the data as they were entered into a computer for analysis. Figure 11.2.5 contains the printout of the analysis using the MINITAB multiple regression program.

El diagrama de dispersión de estos datos se muestra en la Figura 11.2.4. La Tabla 11.2.3 muestra los datos tal como se ingresaron en una computadora para su análisis. La Figura 11.2.5 contiene la impresión del análisis realizado con el programa de regresión múltiple MINITAB.

**Solution:**

Now let us examine the printout to see what it provides in the way of insight into the nature of the relationships among the variables. The least-squares equation is

Ahora examinemos la impresión para ver qué nos proporciona en cuanto a información sobre la naturaleza de las relaciones entre las variables. La ecuación de mínimos cuadrados es

yNj = 6.21 + 1.03x1j + 41.3x 2j + 22.7x 3j - .703x1j x 2j - .510x1j x 3j548

The three regression equations for the three treatments are as follows:

Las tres ecuaciones de regresión para los tres tratamientos son las siguientes:

Treatment A (Equation 11.2.5)

yNj = 16.21 + 41.32 + 11.03 - .7032x 1j

= 47.51 + .327x 1j1

Treatment B (Equation 11.2.6)

yN j = 16.21 + 22.72 + 11.03 - .5102x1j

= 28.91 + .520x1j

Treatment C (Equation 11.2.7)

yN j = 6.21 + 1.03x1j

Figure 11.2.6 contains the scatter diagram of the original data along with the regression equations for the three treatments. Visual inspection of Figure 11.2.6 suggests that treatments A and B do not differ greatly with respect to their slopes, but their y-intercepts are considerably different. The graph suggests that treatment A is better than treatment B for younger patients, but the difference is less dramatic with older patients. Treatment C appears to be decidedly less desirable than both treatments A and B for younger patients but is about as effective as treatment B for older patients. These subjective impressions are compatible with the contention that there is interaction between treatments and age.

La figura 11.2.6 contiene el diagrama de dispersión de los datos originales junto con las ecuaciones de regresión para los tres tratamientos. La inspección visual de la figura 11.2.6 sugiere que los tratamientos A y B no difieren mucho con respecto a sus pendientes, pero sus intersecciones con el eje y son considerablemente diferentes. El gráfico sugiere que el tratamiento A es mejor que el tratamiento B para pacientes más jóvenes, pero la diferencia es menos dramática con pacientes mayores. El tratamiento C parece ser decididamente menos deseable que los tratamientos A y B para pacientes más jóvenes, pero es casi tan efectivo como el tratamiento B para pacientes mayores. Estas impresiones subjetivas son compatibles con la afirmación de que existe interacción entre los tratamientos y la edad.

Inference Procedures

**Procedimientos de inferencia**

The relationships we see in Figure 11.2.6, however, are sample results. What can we conclude about the population from which the sample was drawn?

Sin embargo, las relaciones que vemos en la Figura 11.2.6 son resultados de muestra. ¿Qué podemos concluir acerca de la población de la que se extrajo la muestra?

For an answer let us look at the t ratios on the computer printout in Figure 11.2.5.

Para obtener la respuesta, observemos las relaciones t en la impresión de la computadora en la Figura 11.2.5.

Each of these is the test statistic

Cada uno de estos es la estadística de prueba.

bN i - 0

t =

sbN i550

for testing H0 : b i = 0. We see by Equation 11.2.5 that the y-intercept of the regression line for treatment A is equal to bN 0 + bN 2. Since the t ratio of 8.12 for testing H0 : b 2 = 0 is greater than the critical t of 2.0423 (for a = .05), we can reject H0 that b 2 = 0 and conclude that the y-intercept of the population regression line for treatment A is different from the y-intercept of the population regression line for treatment C, which has a y-intercept of b 0. Similarly, since the t ratio of 4.46 for testing H0 : b 3 = 0 is also greater than the critical t of 2.0423, we can conclude (at the .05 level of significance) that the y-intercept of the population regression line for treatment B is also different from the y-intercept of the population regression line for treatment C. (See the y-intercept of Equation 11.2.6.)

para probar H0 : b i = 0. Vemos por la Ecuación 11.2.5 que la intersección con el eje y de la línea de regresión para el tratamiento A es igual a bN 0 + bN 2. Dado que la razón t de 8,12 para probar H0 : b 2 = 0 es mayor que la t crítica de 2,0423 (para a = .05), podemos rechazar H0 que b 2 = 0 y concluir que la intersección con el eje y de la línea de regresión poblacional para el tratamiento A es diferente de la intersección con el eje y de la línea de regresión poblacional para el tratamiento C, que tiene una intersección con el eje y de b 0. De manera similar, dado que la razón t de 4,46 para probar H0 : b 3 = 0 también es mayor que la t crítica de 2,0423, podemos concluir (en el nivel de significancia .05) que la intersección con el eje y de la línea de regresión poblacional para el tratamiento B también es diferente de la intersección con el eje y de la línea de regresión poblacional para el tratamiento C. (Véase la intersección con el eje y de la ecuación 11.2.6.)

Now let us consider the slopes. We see by Equation 11.2.5 that the slope of the regression line for treatment A is equal to bN 1 (the slope of the line for treatment C) + bN 4. Since the t ratio of -6.45 for testing H0 : b 4 = 0 is less than the critical t of -2.0423, we can conclude (for a = .05) that the slopes of the population regression lines for treatments A and C are different. Similarly, since the computed t ratio for testing H0 : b 5 = 0 is also less than -2.0423, we conclude (for a = .05) that the population regression lines for treatments B and C have different slopes (see the slope of Equation 11.2.6). Thus we conclude that there is interaction between age and type of treatment. This is reflected by a lack of parallelism among the regression lines in Figure 11.2.6.

Ahora consideremos las pendientes. Vemos por la Ecuación 11.2.5 que la pendiente de la línea de regresión para el tratamiento A es igual a bN 1 (la pendiente de la línea para el tratamiento C) + bN 4. Puesto que el cociente t de -6,45 para probar H0 : b 4 = 0 es menor que el t crítico de -2,0423, podemos concluir (para a = .05) que las pendientes de las líneas de regresión poblacional para los tratamientos A y C son diferentes. De manera similar, puesto que el cociente t calculado para probar H0 : b 5 = 0 también es menor que -2,0423, concluimos (para a = .05) que las líneas de regresión poblacional para los tratamientos B y C tienen pendientes diferentes (ver la pendiente de la Ecuación 11.2.6). Por lo tanto, concluimos que hay interacción entre la edad y el tipo de tratamiento. Esto se refleja en una falta de paralelismo entre las líneas de regresión en la Figura 11.2.6.

Another question of interest is this: Is the slope of the population regression line for treatment A different from the slope of the population regression line for treatment B? To answer this question requires computational techniques beyond the scope of this text. The interested reader is referred to books devoted specifically to regression analysis.

Otra pregunta de interés es la siguiente: ¿la pendiente de la línea de regresión poblacional para el tratamiento A es diferente de la pendiente de la línea de regresión poblacional para el tratamiento B? Para responder a esta pregunta se requieren técnicas computacionales que van más allá del alcance de este texto. Se recomienda al lector interesado consultar libros dedicados específicamente al análisis de regresión.

In Section 10.4 the reader was warned that there are problems involved in making multiple inferences from the same sample data. Again, books on regression analysis are available that may be consulted for procedures to be followed when multiple inferences, such as those discussed in this section, are desired.

En la Sección 10.4 se advirtió al lector que existen problemas al realizar múltiples inferencias a partir de los mismos datos de muestra. Nuevamente, existen libros sobre análisis de regresión que pueden consultarse para conocer los procedimientos que se deben seguir cuando se desean realizar múltiples inferencias, como las que se analizan en esta sección.

We have discussed only two situations in which the use of dummy variables is appropriate. More complex models involving the use of one or more qualitative independent variables in the presence of two or more quantitative variables may be appropriate in certain circumstances. More complex models are discussed in the many books devoted to the subject of multiple regression analysis.

Hemos analizado sólo dos situaciones en las que resulta adecuado el uso de variables ficticias. En determinadas circunstancias, pueden resultar adecuados modelos más complejos que impliquen el uso de una o más variables independientes cualitativas en presencia de dos o más variables cuantitativas. En los numerosos libros dedicados al tema del análisis de regresión múltiple se analizan modelos más complejos.

**EXERCISES**

For each exercise do the following:

Para cada ejercicio haga lo siguiente:

(a) Draw a scatter diagram of the data using different symbols for the different categorical variables.

(b) Use dummy variable coding and regression to analyze the data.

(c) Perform appropriate hypothesis tests and construct appropriate confidence intervals using your choice of significance and confidence levels.

(d) Find the p value for each test that you perform.

11.2.1 For subjects undergoing stem cell transplants, dendritic cells (DCs) are antigen-presenting cells that are critical to the generation of immunologic tumor responses. Bolwell et al. (A-2) studied lymphoid DCs in 44 subjects who underwent autologous stem cell transplantation. The outcome variable is the concentration of DC2 cells as measured by flow cytometry. One of the independent variables is the age of the subject (years), and the second independent variable is the mobilization method. During chemotherapy, 11 subjects received granulocyte colony-stimulating factor (G-CSF) mobilizer (mg/kg/day) and 33 received etoposide (2 g>m2). The mobilizer is a kind of blood progenitor cell that triggers the formation of the DC cells. The results were as follows:

11.2.1 En los sujetos sometidos a trasplantes de células madre, las células dendríticas (CD) son células presentadoras de antígenos que son fundamentales para la generación de respuestas inmunológicas tumorales. Bolwell et al. (A-2) estudiaron las CD linfoides en 44 sujetos sometidos a trasplante autólogo de células madre. La variable de resultado es la concentración de células CD2 medida mediante citometría de flujo. Una de las variables independientes es la edad del sujeto (años) y la segunda variable independiente es el método de movilización. Durante la quimioterapia, 11 sujetos recibieron movilizador de factor estimulante de colonias de granulocitos (G-CSF) (mg/kg/día) y 33 recibieron etopósido (2 g>m2). El movilizador es un tipo de célula progenitora sanguínea que desencadena la formación de las células CD. Los resultados fueron los siguientes:

11.2.2 According to Pandey et al. (A-3) carcinoma of the gallbladder is not infrequent. One of the primary risk factors for gallbladder cancer is cholelithiasis, the asymptomatic presence of stones in the gallbladder. The researchers performed a case-control study of 50 subjects with gallbladder cancer and 50 subjects with cholelithiasis. Of interest was the concentration of lipid peroxidation products in gallbladder bile, a condition that may give rise to gallbladder cancer. The lipid peroxidation product melonaldehyde (MDA, mg/mg) was used to measure lipid peroxidation. One of the independent variables considered was the cytochrome P-450 concentration (CYTO, nmol/mg). Researchers used disease status (gallbladder cancer vs. cholelithiasis) and cytochrome P-450 concentration to predict MDA. The following data were collected.

11.2.2 Según Pandey et al. (A-3), el carcinoma de vesícula biliar no es infrecuente. Uno de los principales factores de riesgo del cáncer de vesícula biliar es la colelitiasis, la presencia asintomática de cálculos en la vesícula biliar. Los investigadores realizaron un estudio de casos y controles de 50 sujetos con cáncer de vesícula biliar y 50 sujetos con colelitiasis. De interés fue la concentración de productos de peroxidación lipídica en la bilis de la vesícula biliar, una condición que puede dar lugar al cáncer de vesícula biliar. El producto de peroxidación lipídica melonaldehído (MDA, mg/mg) se utilizó para medir la peroxidación lipídica. Una de las variables independientes consideradas fue la concentración de citocromo P-450 (CYTO, nmol/mg). Los investigadores utilizaron el estado de la enfermedad (cáncer de vesícula biliar frente a colelitiasis) y la concentración de citocromo P-450 para predecir el MDA. Se recopilaron los siguientes datos.

11.2.3 The purpose of a study by Krantz et al. (A-4) was to investigate dose-related effects of methadone in subjects with torsades de pointes, a polymorphic ventricular tachycardia. In the study of 17 subjects, 10 were men (sex = 0) and seven were women (sex = 1). The outcome variable, is the QTc interval, a measure of arrhythmia risk. The other independent variable, in addition to sex, was methadone dose (mg/day). Measurements on these variables for the 17 subjects were as follows.

11.2.3 El objetivo de un estudio realizado por Krantz et al. (A-4) fue investigar los efectos relacionados con la dosis de metadona en sujetos con torsades de pointes, una taquicardia ventricular polimórfica. En el estudio de 17 sujetos, 10 eran hombres (sexo = 0) y siete eran mujeres (sexo = 1). La variable de resultado es el intervalo QTc, una medida del riesgo de arritmia. La otra variable independiente, además del sexo, fue la dosis de metadona (mg/día). Las mediciones de estas variables para los 17 sujetos fueron las siguientes.

11.2.4 Refer to Exercise 9.7.2, which describes research by Reiss et al. (A-5), who collected samples from 90 patients and measured partial thromboplastin time (aPTT) using two different methods: the CoaguChek point-of-care assay and standard laboratory hospital assay. The subjects were also classified by their medication status: 30 receiving heparin alone, 30 receiving heparin with warfarin, and 30 receiving warfarin and enoxaparin. The data are as follows.

11.2.4 Consulte el Ejercicio 9.7.2, que describe la investigación de Reiss et al. (A-5), quienes recogieron muestras de 90 pacientes y midieron el tiempo de tromboplastina parcial (aPTT) utilizando dos métodos diferentes: el ensayo en el punto de atención CoaguChek y el ensayo estándar de laboratorio en el hospital. Los sujetos también fueron clasificados según su estado de medicación: 30 recibieron heparina sola, 30 recibieron heparina con warfarina y 30 recibieron warfarina y enoxaparina. Los datos son los siguientes.

Use the multiple regression to predict the hospital aPTT from the CoaguCheck aPTT level as well as the medication received. Is knowledge of medication useful in the prediction? Let a = .05 for all tests.

Utilice la regresión múltiple para predecir el aPTT hospitalario a partir del nivel de aPTT de CoaguCheck, así como de la medicación recibida. ¿Es útil el conocimiento de la medicación para la predicción? Sea a = .05 para todas las pruebas.556CAPÍTULO 11 ANÁLISIS DE REGRESIÓN: ALGUNAS TÉCNICAS ADICIONALES

**11.3 VARIABLE SELECTION PROCEDURES**

11.3 PROCEDIMIENTOS DE SELECCIÓN DE VARIABLES

Health sciences researchers contemplating the use of multiple regression analysis to solve problems usually find that they have a large number of variables from which to select the independent variables to be employed as predictors of the dependent variable. Such investigators will want to include in their model as many variables as possible in order to maximize the model’s predictive ability. The investigator must realize, however, that adding another independent variable to a set of independent variables always increases the coefficient of determination R 2. Therefore, independent variables should not be added to the model indiscriminately, but only for good reason. In most situations, for example, some potential predictor variables are more expensive than others in terms of data-collection costs. The cost-conscious investigator, therefore, will not want to include an expensive variable in a model unless there is evidence that it makes a worthwhile contribution to the predictive ability of the model.

Los investigadores de las ciencias de la salud que contemplan el uso del análisis de regresión múltiple para resolver problemas suelen descubrir que tienen una gran cantidad de variables de las cuales seleccionar las variables independientes que se emplearán como predictores de la variable dependiente. Dichos investigadores querrán incluir en su modelo tantas variables como sea posible para maximizar la capacidad predictiva del modelo. Sin embargo, el investigador debe darse cuenta de que agregar otra variable independiente a un conjunto de variables independientes siempre aumenta el coeficiente de determinación R2. Por lo tanto, las variables independientes no deben agregarse al modelo indiscriminadamente, sino solo por una buena razón. En la mayoría de las situaciones, por ejemplo, algunas variables predictoras potenciales son más caras que otras en términos de costos de recopilación de datos. Por lo tanto, el investigador consciente de los costos no querrá incluir una variable costosa en un modelo a menos que haya evidencia de que hace una contribución valiosa a la capacidad predictiva del modelo.

The investigator who wishes to use multiple regression analysis most effectively must be able to employ some strategy for making intelligent selections from among those potential predictor variables that are available. Many such strategies are in current use, and each has its proponents. The strategies vary in terms of complexity and the tedium involved in their employment. Unfortunately, the strategies do not always lead to the same solution when applied to the same problem.

El investigador que desee utilizar el análisis de regresión múltiple de la manera más eficaz debe ser capaz de emplear alguna estrategia para hacer selecciones inteligentes entre las posibles variables predictoras disponibles. Actualmente se utilizan muchas de esas estrategias y cada una tiene sus defensores. Las estrategias varían en términos de complejidad y del tedio que implica su uso. Desafortunadamente, las estrategias no siempre conducen a la misma solución cuando se aplican al mismo problema.

Stepwise Regression Perhaps the most widely used strategy for selecting independent variables for a multiple regression model is the stepwise procedure. The procedure consists of a series of steps. At each step of the procedure each variable then in the model is evaluated to see if, according to specified criteria, it should remain in the model.

**Regresión por pasos** Quizás la estrategia más utilizada para seleccionar variables independientes para un modelo de regresión múltiple es el procedimiento por pasos. El procedimiento consta de una serie de pasos. En cada paso del procedimiento se evalúa cada variable del modelo para ver si, según criterios específicos, debería permanecer en el modelo.

Suppose, for example, that we wish to perform stepwise regression for a model containing k predictor variables. The criterion measure is computed for each variable. Of all the variables that do not satisfy the criterion for inclusion in the model, the one that least satisfies the criterion is removed from the model. If a variable is removed in this step, the regression equation for the smaller model is calculated and the criterion measure is computed for each variable now in the model. If any of these variables fail to satisfy the criterion for inclusion in the model, the one that least satisfies the criterion is removed. If a variable is removed at this step, the variable that was removed in the first step is reentered into the model, and the evaluation procedure is continued. This process continues until no more variables can be entered or removed.

Supongamos, por ejemplo, que deseamos realizar una regresión por pasos para un modelo que contiene k variables predictoras. Se calcula la medida de criterio para cada variable. De todas las variables que no satisfacen el criterio de inclusión en el modelo, se elimina del modelo la que menos lo satisface. Si se elimina una variable en este paso, se calcula la ecuación de regresión para el modelo más pequeño y se calcula la medida de criterio para cada variable que ahora está en el modelo. Si alguna de estas variables no satisface el criterio de inclusión en el modelo, se elimina la que menos lo satisface. Si se elimina una variable en este paso, la variable que se eliminó en el primer paso se vuelve a ingresar en el modelo y continúa el procedimiento de evaluación. Este proceso continúa hasta que no se puedan ingresar o eliminar más variables.

The nature of the stepwise procedure is such that, although a variable may be deleted from the model in one step, it is evaluated for possible reentry into the model in subsequent steps.

La naturaleza del procedimiento paso a paso es tal que, aunque una variable puede eliminarse del modelo en un paso, se evalúa su posible reingreso al modelo en pasos posteriores.

MINITAB’s STEPWISE procedure, for example, uses the associated F statistic as the evaluative criterion for deciding whether a variable should be deleted or added to the model. Unless otherwise specified, the cutoff value is F = 4. The printout of the STEPWISE results contains t statistics (the square root of F ) rather than F statistics. At each step MINITAB calculates an F statistic for each variable then in the model. If the F statistic for any of these variables is less than the specified cutoff value (4 if some other value is not specified), the variable with the smallest F is removed from the model. The regression equation is refitted for the reduced model, the results are printed, and the procedure goes to the next step. If no variable can be removed, the procedure tries to add a variable. An F statistic is calculated for each variable not then in the model. Of these variables, the one with the largest associated F statistic is added, provided its F statistic is larger than the specified cutoff value (4 if some other value is not specified). The regression equation is refitted for the new model, the results are printed, and the procedure goes on to the next step. The procedure stops when no variable can be added or deleted.

Por ejemplo, el procedimiento STEPWISE de MINITAB utiliza la estadística F asociada como criterio de evaluación para decidir si se debe eliminar o agregar una variable al modelo. A menos que se especifique lo contrario, el valor de corte es F = 4. La impresión de los resultados de STEPWISE contiene estadísticas t (la raíz cuadrada de F ) en lugar de estadísticas F. En cada paso, MINITAB calcula una estadística F para cada variable que exista en el modelo. Si la estadística F de cualquiera de estas variables es menor que el valor de corte especificado (4 si no se especifica ningún otro valor), la variable con la F más pequeña se elimina del modelo. La ecuación de regresión se reajusta para el modelo reducido, se imprimen los resultados y el procedimiento pasa al siguiente paso. Si no se puede eliminar ninguna variable, el procedimiento intenta agregar una variable. Se calcula una estadística F para cada variable que no exista en el modelo. De estas variables, se agrega la que tenga la estadística F asociada más grande, siempre que su estadística F sea mayor que el valor de corte especificado (4 si no se especifica ningún otro valor). La ecuación de regresión se reajusta para el nuevo modelo, se imprimen los resultados y el procedimiento continúa con el siguiente paso. El procedimiento se detiene cuando no se puede agregar ni eliminar ninguna variable.

The following example illustrates the use of the stepwise procedure for selecting variables for a multiple regression model.

**El siguiente ejemplo ilustra el uso del procedimiento paso a paso para seleccionar variables para un modelo de regresión múltiple.**

**EXAMPLE 11.3.1**

A nursing director would like to use nurses’ personal characteristics to develop a regression model for predicting the job performance ( JOBPER). The following variables are available from which to choose the independent variables to include in the model:

Un director de enfermería desea utilizar las características personales de las enfermeras para desarrollar un modelo de regresión para predecir el desempeño laboral (JOBPER). Las siguientes variables están disponibles para elegir las variables independientes que se incluirán en el modelo:

X1 = assertiveness 1ASRV2

asertividad

X2 = enthusiasm 1ENTH2

entusiasmo

X3 = ambition 1AMB2

X4 = communication skills 1COMM2

X5 = problem-solving skills 1PROB2

X6 = initiative 1INIT2

iniciativa

We wish to use the stepwise procedure for selecting independent variables from those available in the table to construct a multiple regression model for predicting job performance.

Deseamos utilizar el procedimiento paso a paso para seleccionar variables independientes de aquellas disponibles en la tabla para construir un modelo de regresión múltiple para predecir el desempeño laboral.

**Solution:**

Table 11.3.1 shows the measurements taken on the dependent variable, JOBPER, and each of the six independent variables for a sample of 30 nurses.

TABLE 11.3.1 Measurements on Seven Variables for Examples 11.3.1

We use MINITAB to obtain a useful model by the stepwise procedure. Observations on the dependent variable job performance (JOBPER) and the six candidate independent variables are stored in MINITAB Columns 1 through 7, respectively. Figure 11.3.1 shows the appropriate MINITAB procedure and the printout of the results.

Utilizamos MINITAB para obtener un modelo útil mediante el procedimiento paso a paso. Las observaciones sobre la variable dependiente desempeño laboral (JOBPER) y las seis variables independientes candidatas se almacenan en las columnas 1 a 7 de MINITAB, respectivamente. La figura 11.3.1 muestra el procedimiento MINITAB apropiado y la impresión de los resultados.

To obtain the results in Figure 11.3.1, the values of F to enter and F to remove both were set automatically at 4. In step 1 there are no variables to be considered for deletion from the model. The variable AMB (Column 4) has the largest associated F statistic, which is F = 19.7422 = 94.8676. Since 94.8676 is greater than 4, AMB is added to the model. In step 2 the variable INIT (Column 7) qualifies for addition to the model since its associated F of 1-2.222 = 4.84 is greater than 4 and it is the variable with the largest associated F statistic. It is added to the model. After step 2 no other variable could be added or deleted, and the procedure stopped. We see, then, that the model chosen by the stepwise procedure is a two-independent-variable model with AMB and INIT as the independent variables. The estimated regression equation is

Para obtener los resultados de la Figura 11.3.1, los valores de F para ingresar y F para eliminar se establecieron automáticamente en 4. En el paso 1 no hay variables que se deban considerar para la eliminación del modelo. La variable AMB (columna 4) tiene la estadística F asociada más grande, que es F = 19.7422 = 94.8676. Como 94.8676 es mayor que 4, se agrega AMB al modelo. En el paso 2, la variable INIT (columna 7) califica para su adición al modelo ya que su F asociada de 1-2.222 = 4.84 es mayor que 4 y es la variable con la estadística F asociada más grande. Se agrega al modelo. Después del paso 2, no se pudo agregar ni eliminar ninguna otra variable y el procedimiento se detuvo. Vemos, entonces, que el modelo elegido por el procedimiento paso a paso es un modelo de dos variables independientes con AMB e INIT como variables independientes. La ecuación de regresión estimada es

yN = 31.96 + .787x 3 - .45x 6

To change the criterion for allowing a variable to enter the model from 4 to some other value K, click on Options, then type the desired value of K in the Enter box. The new criterion F statistic, then, is K rather than 4. To change the criterion for deleting a variable from the model from 4 to some other value K, click on Options, then type the desired value of K in the Remove box. We must choose K to enter to be greater than or equal to K to remove.

Para cambiar el criterio para permitir que una variable ingrese al modelo de 4 a otro valor K, haga clic en Opciones y luego escriba el valor deseado de K en el cuadro Ingresar. El nuevo estadístico del criterio F, entonces, es K en lugar de 4. Para cambiar el criterio para eliminar una variable del modelo de 4 a otro valor K, haga clic en Opciones y luego escriba el valor deseado de K en el cuadro Eliminar. Debemos elegir que K para ingresar sea mayor o igual que K para eliminar.

Though the stepwise selection procedure is a common technique employed by researchers, other methods are available. Following is a brief discussion of two such tools. The final model obtained by each of these procedures is the same model that was found by using the stepwise procedure in Example 11.3.1.

Aunque el procedimiento de selección por pasos es una técnica común empleada por los investigadores, existen otros métodos disponibles. A continuación se presenta una breve descripción de dos de estas herramientas. El modelo final obtenido mediante cada uno de estos procedimientos es el mismo modelo que se encontró utilizando el procedimiento por pasos en el Ejemplo 11.3.1.

Forward Selection This strategy is closely related to the stepwise regression procedure. This method builds a model using correlations. Variables are retained that meet the criteria for inclusion, as in stepwise selection. The first variable entered into the model is the one with the highest correlation with the dependent variable. If this variable meets the inclusion criterion, it is retained. The next variable to be considered for inclusion is the one with the highest partial correlation with the dependent variable. If it meets the inclusion criteria, it is retained. This procedure continues until all of the independent variables have been considered. The final model contains all of the independent variables that meet the inclusion criteria.

**Selección hacia adelante** Esta estrategia está estrechamente relacionada con el procedimiento de regresión por pasos. Este método construye un modelo utilizando correlaciones. Se conservan las variables que cumplen los criterios de inclusión, como en la selección por pasos. La primera variable que se introduce en el modelo es la que tiene la correlación más alta con la variable dependiente. Si esta variable cumple con el criterio de inclusión, se conserva. La siguiente variable que se considera para su inclusión es la que tiene la correlación parcial más alta con la variable dependiente. Si cumple con los criterios de inclusión, se conserva. Este procedimiento continúa hasta que se hayan considerado todas las variables independientes. El modelo final contiene todas las variables independientes que cumplen con los criterios de inclusión.

Backward Elimination This model-building procedure begins with all of the variables in the model. This strategy also builds a model using correlations and a predetermined inclusion criterion based on the F statistic. The first variable considered for removal from the model is the one with the smallest partial correlation coefficient. If this variable does not meet the criterion for inclusion, it is eliminated from the model. The next variable to be considered for elimination is the one with the next lowest partial correlation. It will be eliminated if it fails to meet the criterion for inclusion. This procedure continues until all variables have been considered for elimination. The final model contains all of the independent variables that meet the inclusion criteria.

**Eliminación hacia atrás** Este procedimiento de construcción de modelos comienza con todas las variables del modelo. Esta estrategia también construye un modelo utilizando correlaciones y un criterio de inclusión predeterminado basado en la estadística F. La primera variable que se considera para eliminar del modelo es la que tiene el coeficiente de correlación parcial más pequeño. Si esta variable no cumple con el criterio de inclusión, se elimina del modelo. La siguiente variable que se considera para eliminar es la que tiene la siguiente correlación parcial más baja. Se eliminará si no cumple con el criterio de inclusión. Este procedimiento continúa hasta que se hayan considerado todas las variables para su eliminación. El modelo final contiene todas las variables independientes que cumplen con los criterios de inclusión.

**EXERCISES**

11.3.1 Refer to the data of Exercise 10.3.2 reported by Son et al. (A-6), who studied family caregiving in Korea of older adults with dementia. The outcome variable, caregiver burden (BURDEN), was measured by the Korean Burden Inventory (KBI) where scores ranged from 28 to 140 with higher scores indicating higher burden. Perform a stepwise regression analysis on the following independent variables reported by the researchers:

11.3.1 Consulte los datos del Ejercicio 10.3.2 informados por Son et al. (A-6), quienes estudiaron el cuidado familiar en Corea de adultos mayores con demencia. La variable de resultado, la carga del cuidador (BURDEN), se midió mediante el Inventario de Carga Coreano (KBI), donde las puntuaciones oscilaron entre 28 y 140, y las puntuaciones más altas indican una carga mayor. Realice un análisis de regresión por pasos sobre las siguientes variables independientes informadas por los investigadores:

CGAGE: caregiver age (years)

edad del cuidador

CGINCOME: caregiver income (Won-Korean currency)

CGDUR: caregiver-duration of caregiving (month)

ADL: total activities of daily living where low scores indicate the elderly perform activities

independently.

MEM: memory and behavioral problems with higher scores indicating more problems.

COG: cognitive impairment with lower scores indicating a greater degree of cognitive impair-

ment.

SOCIALSU: total score of perceived social support (25–175, higher values indicating more

support). The reported data are as follows.

11.3.2 Machiel Naeije (A-7) identifies variables useful in predicting maximum mouth opening (MMO, millimeters) for 35 healthy volunteers. The variables examined were:

11.3.2 Machiel Naeije (A-7) identifica variables útiles para predecir la apertura bucal máxima (AMO, milímetros) en 35 voluntarios sanos. Las variables examinadas fueron:

AGE:

DOWN\_CON:

FORW\_CON:

Gender:

MAN\_LENG:

MAN\_WIDT:

years

downward condylar translation, mm

forward condylar translation, mm

0 " Female, 1 " Male

mandibular length, mm

mandibular width, mm

Use the following reported measurements to perform a stepwise regression.

11.3.3 One purpose of a study by Connor et al. (A-8) was to examine reactive aggression among children and adolescents referred to a residential treatment center. The researchers used the Proactive/Reactive Rating Scale, obtained by presenting three statements to clinicians who examined the subjects. The respondents answered, using a scale from 1 to 5, with 5 indicating that the statement almost always applied to the child. An example of a reactive aggression statement is, “When this child has been teased or threatened, he or she gets angry easily and strikes back.” The reactive score was the average response to three statements of this type. With this variable as the outcome variable, researchers also examined the following: AGE (years), VERBALIQ (verbal IQ), STIM (stimulant use), AGEABUSE (age when first abused), CTQ (a measure of hyperactivity in which higher scores indicate higher hyperactivity), TOTALHOS (total hostility as measured by an evaluator, with higher numbers indicating higher hostility). Perform stepwise regression to find the variables most useful in predicting reactive aggression in the following sample of 68 subjects.

11.3.3 Uno de los objetivos de un estudio realizado por Connor et al. (A-8) fue examinar la agresión reactiva entre niños y adolescentes derivados a un centro de tratamiento residencial. Los investigadores utilizaron la Escala de Calificación Proactiva/Reactiva, obtenida presentando tres afirmaciones a los médicos que examinaron a los sujetos. Los encuestados respondieron utilizando una escala del 1 al 5, donde 5 indica que la afirmación casi siempre se aplicaba al niño. Un ejemplo de una afirmación de agresión reactiva es: “Cuando este niño ha sido objeto de burlas o amenazas, se enoja fácilmente y contraataca”. La puntuación reactiva fue la respuesta media a tres afirmaciones de este tipo. Con esta variable como variable de resultado, los investigadores también examinaron lo siguiente: AGE (años), VERBALIQ (CI verbal), STIM (uso de estimulantes), AGEABUSE (edad en la que se abusó por primera vez), CTQ (una medida de hiperactividad en la que las puntuaciones más altas indican una hiperactividad más alta), TOTALHOS (hostilidad total medida por un evaluador, donde los números más altos indican una hostilidad más alta). Realice una regresión por pasos para encontrar las variables más útiles para predecir la agresión reactiva en la siguiente muestra de 68 sujetos.

**1 .4 LOGISTIC REGRESSION**

Up to now our discussion of regression analysis has been limited to those situations in which the dependent variable is a continuous variable such as weight, blood pressure, or plasma levels of some hormone. Much research in the health sciences field is motivated by a desire to describe, understand, and make use of the relationship between independent variables and a dependent (or outcome) variable that is discrete. Particularly plentiful are circumstances in which the outcome variable is dichotomous. A dichotomous variable, we recall, is a variable that can assume only one of two mutually exclusive values. These values are usually coded Y = 1 for a success and Y = 0 for a nonsuccess, or failure.

Hasta ahora, nuestra discusión del análisis de regresión se ha limitado a aquellas situaciones en las que la variable dependiente es una variable continua, como el peso, la presión arterial o los niveles plasmáticos de alguna hormona. Gran parte de la investigación en el campo de las ciencias de la salud está motivada por el deseo de describir, comprender y hacer uso de la relación entre las variables independientes y una variable dependiente (o de resultado) que es discreta. Son particularmente abundantes las circunstancias en las que la variable de resultado es dicotómica. Recordemos que una variable dicotómica es una variable que puede asumir solo uno de dos valores mutuamente excluyentes. Estos valores suelen codificarse como Y = 1 para un éxito e Y = 0 para un fracaso.

Dichotomous variables include those whose two possible values are such categories as died–did not die; cured–not cured; disease occurred–disease did not occur; and smoker–nonsmoker. The health sciences professional who either engages in research or needs to understand the results of research conducted by others will find it advantageous to have, at least, a basic understanding of logistic regression, the type of regression analysis that is usually employed when the dependent variable is dichotomous. The purpose of the present discussion is to provide the reader with this level of understanding. We shall limit our presentation to the case in which there is only one independent variable that may be either continuous or dichotomous.

Las variables dicotómicas incluyen aquellas cuyos dos valores posibles son categorías como muerto-no muerto; curado-no curado; enfermedad ocurrida-enfermedad no ocurrida; y fumador-no fumador. El profesional de las ciencias de la salud que se dedica a la investigación o necesita entender los resultados de la investigación realizada por otros encontrará ventajoso tener, al menos, un conocimiento básico de la regresión logística, el tipo de análisis de regresión que se emplea habitualmente cuando la variable dependiente es dicotómica. El propósito de la presente discusión es proporcionar al lector este nivel de comprensión. Limitaremos nuestra presentación al caso en el que sólo hay una variable independiente que puede ser continua o dicotómica.

**The Logistic Regression Model** Recall that in Chapter 9 we referred to regression analysis involving only two variables as simple linear regression analysis. The simple linear regression model was expressed by the equation

El modelo de regresión logística Recuerde que en el Capítulo 9 nos referimos al análisis de regresión que involucra sólo dos variables como análisis de regresión lineal simple. El modelo de regresión lineal simple se expresó mediante la ecuación

y = b 0 + b 1x + P

(11.4.1)

in which y is an arbitrary observed value of the continuous dependent variable. When the observed value of Y is my ƒx , the mean of a subpopulation of Y values for a given value of X, the quantity P, the difference between the observed Y and the regression line (see Figure 9.2.1) is zero, and we may write Equation 11.4.1 as

donde y es un valor observado arbitrario de la variable dependiente continua. Cuando el valor observado de Y es mi ƒx , la media de una subpoblación de valores de Y para un valor dado de X, la cantidad P, la diferencia entre el Y observado y la línea de regresión (ver Figura 9.2.1) es cero, y podemos escribir la Ecuación 11.4.1 como

my ƒ x = b 0 + b 1x(11.4.2)

E 1y ƒ x2 = b 0 + b 1x(11.4.3)

which may also be written as

Generally the right-hand side of Equations 11.4.1 through 11.4.3 may assume any value between minus infinity and plus infinity.

Generalmente, el lado derecho de las ecuaciones 11.4.1 a 11.4.3 puede asumir cualquier valor entre menos infinito y más infinito.

Even though only two variables are involved, the simple linear regression model is not appropriate when Y is a dichotomous variable because the expected value (or mean) of Y is the probability that Y = 1 and, therefore, is limited to the range 0 through 1, inclusive. Equations 11.4.1 through 11.4.3, then, are incompatible with the reality of the situation.

Aunque sólo intervienen dos variables, el modelo de regresión lineal simple no es adecuado cuando Y es una variable dicotómica porque el valor esperado (o media) de Y es la probabilidad de que Y = 1 y, por lo tanto, está limitado al rango de 0 a 1, inclusive. Las ecuaciones 11.4.1 a 11.4.3, entonces, son incompatibles con la realidad de la situación.

If we let p = P1Y = 12, then the ratio p>11 - p2 can take on values between 0 and plus infinity. Furthermore, the natural logarithm (ln) of p>11 - p2 can take on values between minus infinity and plus infinity just as can the right-hand side of Equations 11.4.1 through 11.4.3. Therefore, we may write

Si hacemos que p = P1Y = 12, entonces la razón p>11 - p2 puede tomar valores entre 0 y más infinito. Además, el logaritmo natural (ln) de p>11 - p2 puede tomar valores entre menos infinito y más infinito, al igual que el lado derecho de las ecuaciones 11.4.1 a 11.4.3. Por lo tanto, podemos escribir

ln c

p

d = b 0 + b 1x

1 - p

(11.4.4)

Equation 11.4.4 is called the logistic regression model because the transformation of my ƒx (that is, p) to ln3p>11 - p24 is called the logit transformation. Equation 11.4.4 may also be written as

La ecuación 11.4.4 se denomina modelo de regresión logística porque la transformación de mi ƒx (es decir, p) a ln3p>11 - p24 se denomina transformación logit. La ecuación 11.4.4 también se puede escribir como

p =

exp1b 0 + b 1x2

1 + exp1b 0 + b 1x2

(11.4.5)

in which exp is the inverse of the natural logarithm.

The logistic regression model is widely used in health sciences research. For example, the model is frequently used by epidemiologists as a model for the probability (interpreted as the risk) that an individual will acquire a disease during some specified time period during which he or she is exposed to a condition (called a risk factor) known to be or suspected of being associated with the disease.

El modelo de regresión logística se utiliza ampliamente en la investigación en ciencias de la salud. Por ejemplo, los epidemiólogos lo utilizan con frecuencia como modelo para la probabilidad (interpretada como el riesgo) de que un individuo contraiga una enfermedad durante un período de tiempo específico durante el cual está expuesto a una condición (llamada factor de riesgo) que se sabe o se sospecha que está asociada con la enfermedad.

Logistic Regression: Dichotomous Independent Variable The simplest situation in which logistic regression is applicable is one in which both the dependent and the independent variables are dichotomous. The values of the dependent (or outcome) variable usually indicate whether or not a subject acquired a disease or whether or not the subject died. The values of the independent variable indicate the status of the subject relative to the presence or absence of some risk factor. In the discussion that follows we assume that the dichotomies of the two variables are coded 0 and 1. When this is the case the variables may be cross-classified in a table, such as Table 11.4.1, that contains two rows and two columns. The cells of the table contain the frequencies of occurrence of all possible pairs of values of the two variables: (1, 1), (1, 0),(0, 1), and (0, 0).

**Regresión logística: variable independiente dicotómica** La situación más simple en la que se puede aplicar la regresión logística es aquella en la que tanto la variable dependiente como la independiente son dicotómicas. Los valores de la variable dependiente (o de resultado) suelen indicar si un sujeto adquirió o no una enfermedad o si el sujeto murió o no. Los valores de la variable independiente indican el estado del sujeto en relación con la presencia o ausencia de algún factor de riesgo. En la discusión que sigue, suponemos que las dicotomías de las dos variables están codificadas como 0 y 1. Cuando este es el caso, las variables pueden clasificarse de forma cruzada en una tabla, como la Tabla 11.4.1, que contiene dos filas y dos columnas. Las celdas de la tabla contienen las frecuencias de aparición de todos los pares posibles de valores de las dos variables: (1, 1), (1, 0), (0, 1) y (0, 0).

An objective of the analysis of data that meet these criteria is a statistic known as the odds ratio. To understand the concept of the odds ratio, we must understand the term odds, which is frequently used by those who place bets on the outcomes of sporting events or participate in other types of gambling activities. Using probability terminology, we may define odds as follows.

Un objetivo del análisis de los datos que cumplen estos criterios es una estadística conocida como razón de probabilidades. Para entender el concepto de razón de probabilidades, debemos entender el término probabilidades, que es utilizado con frecuencia por quienes realizan apuestas sobre los resultados de eventos deportivos o participan en otros tipos de actividades de juego. Utilizando la terminología de probabilidad, podemos definir las probabilidades de la siguiente manera.

**DEFINITION**

The odds for success are the ratio of the probability of success to the probability of failure.

Las probabilidades de éxito son la relación entre la probabilidad de éxito y la probabilidad de fracaso.

The odds ratio is a measure of how much greater (or less) the odds are for subjects possessing the risk factor to experience a particular outcome. This conclusion assumes that the outcome is a rare event. For example, when the outcome is the contracting of a disease, the interpretation of the odds ratio assumes that the disease is rare.

La razón de probabilidades es una medida de cuánto mayores (o menores) son las probabilidades de que los sujetos que poseen el factor de riesgo experimenten un resultado particular. Esta conclusión supone que el resultado es un evento poco común. Por ejemplo, cuando el resultado es contraer una enfermedad, la interpretación de la razón de probabilidades supone que la enfermedad es poco común.

Suppose, for example, that the outcome variable is the acquisition or nonacquisition of skin cancer and the independent variable (or risk factor) is high levels of exposure to the sun. Analysis of such data collected on a sample of subjects might yield an odds ratio of 2, indicating that the odds of skin cancer are two times higher among subjects with high levels of exposure to the sun than among subjects without high levels of exposure.

Supongamos, por ejemplo, que la variable de resultado es la aparición o no de cáncer de piel y que la variable independiente (o factor de riesgo) son los altos niveles de exposición al sol. El análisis de esos datos recopilados sobre una muestra de sujetos podría arrojar un cociente de probabilidades de 2, lo que indica que las probabilidades de padecer cáncer de piel son dos veces mayores entre los sujetos con altos niveles de exposición al sol que entre los sujetos sin altos niveles de exposición.

Computer software packages that perform logistic regression frequently provide as part of their output estimates of b0 and b1 and the numerical value of the odds ratio. As it turns out the odds ratio is equal to exp1b12.

Los paquetes de software informáticos que realizan regresión logística suelen proporcionar como parte de sus resultados estimaciones de b0 y b1 y el valor numérico de la razón de probabilidades. Resulta que la razón de probabilidades es igual a exp1b12.

**EXAMPLE 11.4.1**

LaMont et al. (A-9) tested for obstructive coronary artery disease (OCAD) among 113 men and 35 women who complained of chest pain or possible equivalent to their primary care physician. Table 11.4.2 shows the cross-classification of OCAD with gender. We wish to use logistic regression analysis to determine how much greater the odds are of finding OCAD among men than among women.

LaMont et al. (A-9) realizaron pruebas para detectar la enfermedad coronaria obstructiva (EOC) en 113 hombres y 35 mujeres que se quejaron de dolor en el pecho o un posible equivalente a su médico de atención primaria. La Tabla 11.4.2 muestra la clasificación cruzada de la EOC con el género. Deseamos utilizar un análisis de regresión logística para determinar cuánto mayores son las probabilidades de encontrar EOC entre los hombres que entre las mujeres.

**Solution:**

We may use the SAS® software package to analyze these data. The independent variable is gender and the dependent variable is status with respect to having obstructive coronary artery disease (OCAD). Use of the SAS® command PROC LOGIST yields, as part of the resulting output, the statistics shown in Figure 11.4.1.

Podemos utilizar el paquete de software SAS® para analizar estos datos. La variable independiente es el género y la variable dependiente es el estado con respecto a la presencia de enfermedad coronaria obstructiva (EOC). El uso del comando SAS® PROC LOGIST genera, como parte de la salida resultante, las estadísticas que se muestran en la Figura 11.4.1.

We see that the estimate of a is -1.4773 and the estimate of b 1 is 1.7649. The estimated odds ratio, then, is OR = exp11.76492 = 5.84. Thus we estimate that the odds of finding a case of obstructive coronary artery disease to be almost six times higher among men than women.

Vemos que la estimación de a es -1,4773 y la estimación de b 1 es 1,7649. La razón de probabilidades estimada es, entonces, OR = exp11,76492 = 5,84. Por lo tanto, estimamos que las probabilidades de encontrar un caso de enfermedad coronaria obstructiva son casi seis veces mayores entre los hombres que entre las mujeres.

Logistic Regression: Continuous Independent Variable Now let us consider the situation in which we have a dichotomous dependent variable and a continuous independent variable. We shall assume that a computer is available to perform the calculations. Our discussion, consequently, will focus on an evaluation of the adequacy of the model as a representation of the data at hand, interpretation of key elements of the computer printout, and the use of the results to answer relevant questions about the relationship between the two variables.

Regresión logística: variable independiente continua Consideremos ahora la situación en la que tenemos una variable dependiente dicotómica y una variable independiente continua. Supondremos que se dispone de una computadora para realizar los cálculos. En consecuencia, nuestro análisis se centrará en la evaluación de la idoneidad del modelo como representación de los datos disponibles, la interpretación de los elementos clave de la impresión de la computadora y el uso de los resultados para responder a preguntas relevantes sobre la relación entre las dos variables.

**EXAMPLE 11.4.2**

According to Gallagher et al. (A-10), cardiac rehabilitation programs offer “information, support, and monitoring for return to activities, symptom management, and risk factor modification.” The researchers conducted a study to identify among women factors that are associated with participation in such programs. The data in Table 11.4.3 are the ages of 185 women discharged from a hospital in Australia who met eligibility criteria involving discharge for myocardial infarction, artery bypass surgery, angioplasty, or stent. We wish to use these data to obtain information regarding the relationship between age (years) and participation in a cardiac rehabilitation program (ATT = 1, if participated, and ATT = 0, if not). We wish also to know if we may use the results of our analysis to predict the likelihood of participation by a woman if we know her age.

Según Gallagher et al. (A-10), los programas de rehabilitación cardíaca ofrecen “información, apoyo y seguimiento para el retorno a las actividades, el manejo de los síntomas y la modificación de los factores de riesgo”. Los investigadores llevaron a cabo un estudio para identificar entre las mujeres los factores que se asocian con la participación en dichos programas. Los datos de la Tabla 11.4.3 corresponden a las edades de 185 mujeres dadas de alta de un hospital de Australia que cumplían los criterios de elegibilidad que incluían el alta por infarto de miocardio, cirugía de bypass arterial, angioplastia o colocación de stent. Deseamos utilizar estos datos para obtener información sobre la relación entre la edad (años) y la participación en un programa de rehabilitación cardíaca (ATT = 1, si participó, y ATT = 0, si no). También deseamos saber si podemos utilizar los resultados de nuestro análisis para predecir la probabilidad de participación de una mujer si conocemos su edad.

**Solution**:

The independent variable is the continuous variable age (AGE), and the dependent or response variable is status with respect to attendance in a cardiac rehabilitation program. The dependent variable is a dichotomous variable that can assume one of two values: 0 = did not attend, and 1 = did attend. We use the SAS® software package to analyze the data. The SAS® command is PROC LOGISTIC, but if we wish to predict attendance in the cardiac program, we need to use the “descending” option with PROC LOGISTIC. (When you wish to predict the outcome labeled “1” of the dependent variable, use the “descending option” in SAS®. Consult SAS® documentation for further details.) A partial printout of the analysis is shown in Figure 11.4.2.

La variable independiente es la variable continua edad (AGE), y la variable dependiente o de respuesta es el estado con respecto a la asistencia a un programa de rehabilitación cardíaca. La variable dependiente es una variable dicotómica que puede asumir uno de dos valores: 0 = no asistió y 1 = asistió. Utilizamos el paquete de software SAS® para analizar los datos. El comando SAS® es PROC LOGISTIC, pero si deseamos predecir la asistencia al programa cardíaco, debemos utilizar la opción “descendente” con PROC LOGISTIC. (Cuando desee predecir el resultado etiquetado como “1” de la variable dependiente, utilice la “opción descendente” en SAS®. Consulte la documentación de SAS® para obtener más detalles). En la Figura 11.4.2 se muestra una impresión parcial del análisis.

The slope of our regression is -.0379, and the intercept is 1.8744. The regression equation, then, is

yN i = 1.8744 - .0379x i

where yNi = ln3pN i>11 - pN i24 and pN i is the predicted probability of attending cardiac rehabilitation for a woman aged x i.

**Test of H0 that B 1 " 0**

We reach a conclusion about the adequacy of the logistic model by testing the null hypothesis that the slope of the regression line is zero. The test statistic is z = bN 1>sbN 1 where z is the standard normal statistic, bN 1 is the sample slope 1-.03792, and sbN 1 is its standard error (.0146) as shown in Figure 11.4.2. From these numbers we compute z = -.0379>.0146 = -2.5959, which has an associated two-sided p value of .0094. We conclude, therefore, that the logistic model is adequate. The square of z is chi-square with 1 degree of freedom, a statistic that is shown in Figure 11.4.2.

Llegamos a una conclusión sobre la adecuación del modelo logístico probando la hipótesis nula de que la pendiente de la línea de regresión es cero. El estadístico de prueba es z = bN 1>sbN 1 donde z es el estadístico normal estándar, bN 1 es la pendiente de la muestra 1-.03792, y sbN 1 es su error estándar (.0146) como se muestra en la Figura 11.4.2. A partir de estos números calculamos z = -.0379>.0146 = -2.5959, que tiene un valor p bilateral asociado de .0094. Concluimos, por lo tanto, que el modelo logístico es adecuado. El cuadrado de z es chi-cuadrado con 1 grado de libertad, un estadístico que se muestra en la Figura 11.4.2.

**Using the Logistic Regression to Estimate p**

We may use Equation 11.4.5 and the results of our analysis to estimate p, the probability that a woman of a given age (within the range of ages represented by the data) will attend a cardiac rehabilitation program. Suppose, for example, that we wish to estimate the probability that a woman who is 50 years of age will participate in a rehabilitation program. Substituting 50 and the results shown in Figure 11.4.2 into Equation 11.4.5 gives

Podemos utilizar la ecuación 11.4.5 y los resultados de nuestro análisis para estimar p, la probabilidad de que una mujer de una edad determinada (dentro del rango de edades representado por los datos) asista a un programa de rehabilitación cardíaca. Supongamos, por ejemplo, que deseamos estimar la probabilidad de que una mujer de 50 años participe en un programa de rehabilitación. Sustituyendo 50 y los resultados que se muestran en la Figura 11.4.2 en la ecuación 11.4.5 obtenemos

pN =

exp31.8744 - 1.0379215024

1 + exp31.8744 - 1.0379215024

= .49485

SAS® calculates the estimated probabilities for the given values of X. We can see the estimated probabilities of attending cardiac rehabilitation programs for the age range of the subjects enrolled in the study in Figure 11.4.3. Since the slope was negative, we see a decreasing probability of attending a cardiac rehabilitation program for older women.

SAS® calcula las probabilidades estimadas para los valores dados de X. Podemos ver las probabilidades estimadas de asistir a programas de rehabilitación cardíaca para el rango de edad de los sujetos inscritos en el estudio en la Figura 11.4.3. Como la pendiente era negativa, vemos una probabilidad decreciente de asistir a un programa de rehabilitación cardíaca para mujeres mayores.

Multiple Logistic Regression Practitioners often are interested in the relationships of several independent variables to a response variable. These independent variables may be either continuous or discrete or a combination of the two.

Los profesionales de la regresión logística múltiple suelen estar interesados ​​en las relaciones de varias variables independientes con una variable de respuesta. Estas variables independientes pueden ser continuas o discretas o una combinación de ambas.

Multiple logistic models are constructed by expanding Equations 11.4.1 to 11.4.4. If we begin with Equation 11.4.4, multiple logistic regression can be represented as

Los modelos logísticos múltiples se construyen ampliando las ecuaciones 11.4.1 a 11.4.4. Si comenzamos con la ecuación 11.4.4, la regresión logística múltiple se puede representar como

lnc

p

d = b 0 + b 1x1j + b 2x 2j + Á + b k x kj

1 - p

(11.4.6)

Using the logit transformation, we now have

p =

exp1b 0 + b 1x 1j + b 2x 2j + Á + b k x kj2

1 + exp1b 0 + b 1x 1j + b 2x 2j + Á + b k x kj2

(11.4.7)

**EXAMPLE 11.4.3**

Consider the data presented in Review Exercise 24. In this study by Fils-Aime et al. (A-21), data were gathered and classified with regard to alcohol use. Subjects were classified as having either early (% 25 years) or late ($ 25 years) onset of excessive alcohol use.

Considere los datos presentados en el Ejercicio de Revisión 24. En este estudio de Fils-Aime et al. (A-21), se recopilaron y clasificaron datos con respecto al consumo de alcohol. Los sujetos se clasificaron como de inicio temprano (% 25 años) o tardío ($ 25 años) de consumo excesivo de alcohol.

Levels of cerebrospinal fluid (CSF) tryptophan (TRYPT) and 5-hydroxyindoleacetic acid (5-HIAA) concentrations were also obtained.

También se obtuvieron los niveles de triptófano (TRYPT) y de ácido 5-hidroxiindolacético (5-HIAA) en el líquido cefalorraquídeo (LCR).

**Solution:**

The independent variables are the concentrations of TRYPT and 5-HIAA, and the dependent variable is the dichotomous response for onset of excessive alcohol use. We use SPSS software to analyze the data. The output is presented in Figure 11.4.3.

Las variables independientes son las concentraciones de TRYPT y 5-HIAA, y la variable dependiente es la respuesta dicotómica para el inicio del consumo excesivo de alcohol. Utilizamos el software SPSS para analizar los datos. El resultado se presenta en la Figura 11.4.3.

The equation can be written as

yN i = 2.076 - .013x 1j + 0x 2j

Note that the coefficient for TRYPT is 0, and therefore it is not playing a role in the model.

Test of H0 that B 1 " 0

Tests for significance of the regression coefficients can be obtained directly from Figure 11.4.3. Note that both the constant (intercept) and the 5-HIAA variables are significant in the model (both have p values, noted as “Sig.” in the table, % .05); however, TRYPT is not significant and therefore need not be in the model, suggesting that it is not useful for identifying those study participants with early or late alcoholism onset.

Las pruebas de significancia de los coeficientes de regresión se pueden obtener directamente de la Figura 11.4.3. Nótese que tanto la constante (intersección) como las variables 5-HIAA son significativas en el modelo (ambas tienen valores p, indicados como “Sig.” en la tabla, % .05); sin embargo, TRYPT no es significativo y por lo tanto no necesita estar en el modelo, lo que sugiere que no es útil para identificar a aquellos participantes del estudio con inicio temprano o tardío del alcoholismo.

As above, probabilities can be easily obtained by using equation 11.4.7 and substituting the values obtained from the analysis.

Como se indicó anteriormente, las probabilidades se pueden obtener fácilmente utilizando la ecuación 11.4.7 y sustituyendo los valores obtenidos del análisis.

■

Polytomous Logistic Regression Thus far we have limited our discussion to situations in which there is a dichotomous response variable (e.g., successful or unsuccessful). Often we have a situation in which multiple categories make up the response. We may, for example, have subjects that are classified as positive, negative, and undetermined for a given disease (a standard polytomous response). There may also be times when we have a response variable that is ordered. We may, for example, classify our subjects by BMI as underweight, ideal weight, overweight, or obese (an ordinal polytomous response). The modeling process is slightly more complex and requires the use of a computer program. For those interested in exploring these valuable methods further, we recommend the book by Hosmer and Lemeshow (1).

Regresión logística politómica Hasta ahora hemos limitado nuestra discusión a situaciones en las que hay una variable de respuesta dicotómica (p. ej., exitosa o no exitosa). A menudo tenemos una situación en la que múltiples categorías conforman la respuesta. Podemos, por ejemplo, tener sujetos que se clasifican como positivos, negativos e indeterminados para una enfermedad dada (una respuesta politómica estándar). También puede haber momentos en los que tenemos una variable de respuesta que está ordenada. Podemos, por ejemplo, clasificar a nuestros sujetos por IMC como bajo peso, peso ideal, sobrepeso u obesidad (una respuesta politómica ordinal). El proceso de modelado es ligeramente más complejo y requiere el uso de un programa de computadora. Para aquellos interesados ​​en explorar estos valiosos métodos más a fondo, recomendamos el libro de Hosmer y Lemeshow (1).

Further Reading We have discussed only the basic concepts and applications of logistic regression. The technique has much wider application. Stepwise regression analysis may be used with logistic regression. There are also techniques available for constructing confidence intervals for odds ratios. The reader who wishes to learn more about logistic regression may consult the books by Hosmer and Lemeshow (1) and Kleinbaum (2).

Lectura adicional Hemos analizado únicamente los conceptos y aplicaciones básicos de la regresión logística. La técnica tiene una aplicación mucho más amplia. El análisis de regresión por pasos se puede utilizar con la regresión logística. También existen técnicas disponibles para construir intervalos de confianza para los odds ratios. El lector que desee aprender más sobre la regresión logística puede consultar los libros de Hosmer y Lemeshow (1) y Kleinbaum (2).

**EXERCISES**

11.4.1 In a study of violent victimization of women and men, Porcerelli et al. (A-11) collected information from 679 women and 345 men ages 18 to 64 years at several family-practice centers in the metropolitan Detroit area. Patients filled out a health history questionnaire that included a question about victimization. The following table shows the sample subjects cross-classified by gender and whether the subject self-identified as being “hit, kicked, punched, or otherwise hurt by someone within the past year.” Subjects answering yes to that question are classified “violently victimized.” Use logistic regression analysis to find the regression coefficients and the estimate of the odds ratio. Write an interpretation of your results.

11.4.1 En un estudio sobre la victimización violenta de mujeres y hombres, Porcerelli et al. (A-11) recopilaron información de 679 mujeres y 345 hombres de 18 a 64 años en varios centros de medicina familiar en el área metropolitana de Detroit. Los pacientes completaron un cuestionario de antecedentes médicos que incluía una pregunta sobre la victimización. La siguiente tabla muestra los sujetos de la muestra clasificados de forma cruzada por género y si el sujeto se autoidentificó como “golpeado, pateado, golpeado o lastimado de otra manera por alguien en el último año”. Los sujetos que respondieron sí a esa pregunta se clasificaron como “victimizados violentamente”. Utilice el análisis de regresión logística para encontrar los coeficientes de regresión y la estimación de la razón de probabilidades. Escriba una interpretación de sus resultados.

11.4.2 Refer to the research of Gallagher et al. (A-10) discussed in Example 11.4.2. Another covariate of interest was a score using the Hospital Anxiety and Depression Index. A higher value for this score indicates a higher level of anxiety and depression. Use the following data to predict whether a woman in the study participated in a cardiac rehabilitation program.

11.4.2 Consulte la investigación de Gallagher et al. (A-10) analizada en el Ejemplo 11.4.2. Otra covariable de interés fue una puntuación obtenida con el Índice de ansiedad y depresión hospitalaria. Un valor más alto de esta puntuación indica un nivel más alto de ansiedad y depresión. Utilice los siguientes datos para predecir si una mujer del estudio participó en un programa de rehabilitación cardíaca.

**11.5 SUMMARY**

This chapter is included for the benefit of those who wish to extend their understanding of regression analysis and their ability to apply techniques to models that are more complex than those covered in Chapters 9 and 10. In this chapter we present some additional topics from regression analysis. We discuss the analysis that is appropriate when one or more of the independent variables is dichotomous. In this discussion the concept of dummy variable coding is presented. A second topic that we discuss is how to select the most useful independent variables when we have a long list of potential candidates. The technique we illustrate for the purpose is stepwise regression analysis. Finally, we present the basic concepts and procedures that are involved in logistic regression analysis. We cover two situations: the case in which the independent variable is dichotomous, and the case in which the independent variable is continuous.

Este capítulo se incluye para el beneficio de aquellos que desean ampliar su comprensión del análisis de regresión y su capacidad para aplicar técnicas a modelos que son más complejos que los cubiertos en los Capítulos 9 y 10. En este capítulo presentamos algunos temas adicionales del análisis de regresión. Analizamos el análisis que es apropiado cuando una o más de las variables independientes son dicotómicas. En esta discusión se presenta el concepto de codificación de variables ficticias. Un segundo tema que discutimos es cómo seleccionar las variables independientes más útiles cuando tenemos una larga lista de candidatos potenciales. La técnica que ilustramos para este propósito es el análisis de regresión por pasos. Finalmente, presentamos los conceptos y procedimientos básicos que están involucrados en el análisis de regresión logística. Cubrimos dos situaciones: el caso en el que la variable independiente es dicotómica y el caso en el que la variable independiente es continua.

Since the calculations involved in obtaining useful results from data that are appropriate for analysis by means of the techniques presented in this chapter are complicated and time-consuming when attempted by hand, it is recommended that a computer be used to work the exercises.

Dado que los cálculos necesarios para obtener resultados útiles a partir de datos apropiados para el análisis mediante las técnicas presentadas en este capítulo son complicados y consumen mucho tiempo cuando se intentan realizar a mano, se recomienda utilizar una computadora para realizar los ejercicios.

**REVIEW QUESTIONS AND EXERCISES**

1.What is a qualitative variable?

2.What is a dummy variable?

2.¿Qué es una variable ficticia?

3.Explain and illustrate the technique of dummy variable coding.

3. Explique e ilustre la técnica de codificación de variables ficticias.

4.Why is a knowledge of variable selection techniques important to the health sciences researcher?

4.¿Por qué es importante el conocimiento de las técnicas de selección de variables para el investigador en ciencias de la salud?

5.What is stepwise regression?

5.¿Qué es la regresión por pasos?

6.Explain the basic concept involved in stepwise regression.

6. Explique el concepto básico involucrado en la regresión por pasos.

7.When is logistic regression used?

8.Write out and explain the components of the logistic regression model.

9.Define the word odds.

10.What is an odds ratio?

11.Give an example in your field in which logistic regression analysis would be appropriate when the independent variable is dichotomous.

12.Give an example in your field in which logistic regression analysis would be appropriate when the independent variable is continuous.

13.Find a published article in the health sciences field in which each of the following techniques is employed:

(a) Dummy variable coding

(b) Stepwise regression

(c) Logistic regression

Write a report on the article in which you identify the variables involved, the reason for the choice of the technique, and the conclusions that the authors reach on the basis of their analysis.

Redacte un informe sobre el artículo en el que identifique las variables involucradas, el motivo de la elección de la técnica y las conclusiones a las que llegan los autores con base en su análisis.

14. In Example 10.3.1, we saw that the purpose of a study by Jansen and Keller (A-12) was to predict the capacity to direct attention (CDA) in elderly subjects. The study collected information on 71 community-dwelling older women with normal mental status. Higher CDA scores indicate better attentional functioning. In addition to the variables age and education level, the researchers performed stepwise regression with two additional variables: IADL, a measure of activities of daily living (higher values indicate greater number of daily activities), and ADS, a measure of attentional demands (higher values indicate more attentional demands). Perform stepwise regression with the data in the following table and report your final model, p values, and conclusions.

14. En el Ejemplo 10.3.1, vimos que el propósito de un estudio de Jansen y Keller (A-12) era predecir la capacidad de dirigir la atención (CDA) en sujetos de edad avanzada. El estudio recopiló información sobre 71 mujeres mayores que vivían en la comunidad y tenían un estado mental normal. Las puntuaciones más altas en CDA indican un mejor funcionamiento atencional. Además de las variables edad y nivel educativo, los investigadores realizaron una regresión por pasos con dos variables adicionales: IADL, una medida de las actividades de la vida diaria (los valores más altos indican un mayor número de actividades diarias) y ADS, una medida de las demandas atencionales (los valores más altos indican más demandas atencionales). Realice una regresión por pasos con los datos de la siguiente tabla e informe su modelo final, los valores p y las conclusiones.

15. In the following table are the cardiac output (L/min) and oxygen consumption (VO2) values for a sample of adults (A) and children (C), who participated in a study designed to investigate the relationship among these variables. Measurements were taken both at rest and during exercise. Treat cardiac output as the dependent variable and use dummy variable coding and analyze the data by regression techniques. Explain the results. Plot the original data and the fitted regression equations.

15. En la siguiente tabla se muestran los valores de gasto cardíaco (L/min) y consumo de oxígeno (VO2) de una muestra de adultos (A) y niños (C) que participaron en un estudio diseñado para investigar la relación entre estas variables. Las mediciones se tomaron tanto en reposo como durante el ejercicio. Considere el gasto cardíaco como la variable dependiente y utilice la codificación de variables ficticias y analice los datos mediante técnicas de regresión. Explique los resultados. Dibuje los datos originales y las ecuaciones de regresión ajustadas.

16. A simple random sample of normal subjects between the ages of 6 and 18 yielded the data on total body potassium (mEq) and total body water (liters) shown in the following table. Let total potassium be the dependent variable and use dummy variable coding to quantify the qualitative variable. Analyze the data using regression techniques. Explain the results. Plot the original data and the fitted regression equations.

16. Una muestra aleatoria simple de sujetos normales de entre 6 y 18 años de edad arrojó los datos sobre potasio corporal total (mEq) y agua corporal total (litros) que se muestran en la siguiente tabla. Sea el potasio total la variable dependiente y utilice la codificación de variable ficticia para cuantificar la variable cualitativa. Analice los datos utilizando técnicas de regresión. Explique los resultados. Dibuje los datos originales y las ecuaciones de regresión ajustadas.

17. The data shown in the following table were collected as part of a study in which the subjects were preterm infants with low birth weights born in three different hospitals. Use dummy variable coding and multiple regression techniques to analyze these data. May we conclude that the three sample hospital populations differ with respect to mean birth weight when gestational age is taken into account? May we conclude that there is interaction between hospital of birth and gestational age? Plot the original data and the fitted regression equations.

17. Los datos que se muestran en la siguiente tabla se recopilaron como parte de un estudio en el que los sujetos fueron bebés prematuros con bajo peso al nacer nacidos en tres hospitales diferentes. Utilice la codificación de variables ficticias y técnicas de regresión múltiple para analizar estos datos. ¿Podemos concluir que las tres poblaciones hospitalarias de muestra difieren con respecto al peso medio al nacer cuando se tiene en cuenta la edad gestacional? ¿Podemos concluir que existe interacción entre el hospital de nacimiento y la edad gestacional? Dibuje los datos originales y las ecuaciones de regresión ajustadas.

18. Refer to Chapter 9, Review Exercise 18. In the study cited in that exercise, Maria Mathias (A-13) investigated the relationship between ages (AGE) of boys and improvement in measures of hyperactivity, attitude, and social behavior. In the study, subjects were randomly assigned to two different treatments. The control group (TREAT = 0) received standard therapy for hyperactivity, and the treatment group (TREAT = 1) received standard therapy plus pet therapy. The results are shown in the following table. Create a scatter plot with age as the independent variable and ATT (change in attitude with positive numbers indicating positive change in attitude) as the dependent variable. Use different symbols for the two different treatment groups. Use multiple regression techniques to determine whether age, treatment, or the interaction are useful in predicting ATT. Report your results.

18. Consulte el Capítulo 9, Ejercicio de revisión 18. En el estudio citado en ese ejercicio, Maria Mathias (A-13) investigó la relación entre las edades (EDAD) de los niños y la mejora en las medidas de hiperactividad, actitud y comportamiento social. En el estudio, los sujetos fueron asignados aleatoriamente a dos tratamientos diferentes. El grupo de control (TREAT = 0) recibió terapia estándar para la hiperactividad, y el grupo de tratamiento (TREAT = 1) recibió terapia estándar más terapia con mascotas. Los resultados se muestran en la siguiente tabla. Cree un diagrama de dispersión con la edad como variable independiente y ATT (cambio de actitud con números positivos que indican un cambio positivo en la actitud) como variable dependiente. Utilice símbolos diferentes para los dos grupos de tratamiento diferentes. Utilice técnicas de regresión múltiple para determinar si la edad, el tratamiento o la interacción son útiles para predecir ATT. Informe sus resultados.

For each study described in Exercises 19 through 21, answer as many of the following questions as possible:

(a) Which is the dependent variable?

(b) What are the independent variables?

(c) What are the appropriate null and alternative hypotheses?

(d) Which null hypotheses do you think were rejected? Why?

(e) Which is the more relevant objective, prediction or estimation, or are the two equally relevant? Explain your answer.

(f) What is the sampled population?

(g) What is the target population?

(h) Which variables are related to which other variables? Are the relationships direct or inverse?

(i) Write out the regression equation using appropriate numbers for parameter estimates.

( j) Give numerical values for any other statistics that you can.

(k) Identify each variable as to whether it is quantitative or qualitative.

(l) Explain the meaning of any statistics for which numerical values are given.

19. Golfinopoulos and Arhonditsis (A-14) used a multiple regression model in a study of trihalomethanes (THMs) in drinking water in Athens, Greece. THMs are of concern since they have been related to cancer and reproductive outcomes. The researchers found the following regression model useful in predicting THM:

19. Golfinopoulos y Arhonditsis (A-14) utilizaron un modelo de regresión múltiple en un estudio de trihalometanos (THM) en el agua potable en Atenas, Grecia. Los THM son motivo de preocupación, ya que se han relacionado con el cáncer y los problemas reproductivos. Los investigadores encontraron que el siguiente modelo de regresión era útil para predecir los THM:

THM = -.26chla + 1.57 pH + 28.74Br - 66.72Br 2

-43.63S + 1.13Sp + 2.62T \* S - .72T \* CL

The variables were as follows: chla = chlorophyll concentration, pH = acid/base scale, Br = bromide concentration, S = dummy variable for summer, Sp = dummy variable for spring, T = temperature, and CL = chlorine concentration. The researchers reported R = .52, p 6 .001.

Las variables fueron las siguientes: chla = concentración de clorofila, pH = escala ácido/base, Br = concentración de bromuro, S = variable ficticia para verano, Sp = variable ficticia para primavera, T = temperatura y CL = concentración de cloro. Los investigadores informaron un R = .52, p 6 .001.

20. In a study by Takata et al. (A-15), investigators evaluated the relationship between chewing ability and teeth number and measures of physical fitness in a sample of subjects ages 80 or higher in Japan. One of the outcome variables that measured physical fitness was leg extensor strength. To measure the ability to chew foods, subjects were asked about their ability to chew 15 foods (peanuts, vinegared octopus, and French bread, among others). Consideration of such variables as height, body weight, gender, systolic blood pressure, serum albumin, fasting glucose concentration, back pain, smoking, alcohol consumption, marital status, regular medical treatment, and regular exercise revealed that the number of chewable foods was significant in predicting leg extensor strength 1bN 1 = .075, p = .03662. However, in the presence of the other variables, number of teeth was not a significant predictor 1bN 1 = .003, p = .93732.

20. En un estudio de Takata et al. (A-15), los investigadores evaluaron la relación entre la capacidad de masticación y el número de dientes y las medidas de aptitud física en una muestra de sujetos de 80 años o más en Japón. Una de las variables de resultado que medían la aptitud física era la fuerza extensora de las piernas. Para medir la capacidad de masticar alimentos, se preguntó a los sujetos sobre su capacidad para masticar 15 alimentos (cacahuetes, pulpo avinagrado y pan francés, entre otros). La consideración de variables como la altura, el peso corporal, el sexo, la presión arterial sistólica, la albúmina sérica, la concentración de glucosa en ayunas, el dolor de espalda, el tabaquismo, el consumo de alcohol, el estado civil, el tratamiento médico regular y el ejercicio regular reveló que el número de alimentos masticables era significativo para predecir la fuerza extensora de las piernas 1bN 1 = .075, p = .03662. Sin embargo, en presencia de las otras variables, el número de dientes no fue un predictor significativo 1bN 1 = .003, p = .93732.

21. Varela et al. (A-16) examined 515 patients who underwent lung resection for bronchogenic carcinoma. The outcome variable was the occurrence of cardiorespiratory morbidity after surgery. Any of the following postoperative events indicated morbidity: pulmonary atelectasis or pneumonia, respiratory or ventilatory insufficiency at discharge, need for mechanical ventilation at any time after extubation in the operating room, pulmonary thromboembolism, arrhythmia, myocardial ischemia or infarct, and clinical cardiac insufficiency. Performing a stepwise logistic regression, the researchers found that age 1p 6 .0012 and postoperative forced expiratory volume 1p = .0032 were statistically significant in predicting the occurrence of cardiorespiratory morbidity.

21. Varela et al. (A-16) examinaron a 515 pacientes que se sometieron a resección pulmonar por carcinoma broncogénico. La variable de resultado fue la aparición de morbilidad cardiorrespiratoria después de la cirugía. Cualquiera de los siguientes eventos postoperatorios indicó morbilidad: atelectasia pulmonar o neumonía, insuficiencia respiratoria o ventilatoria al alta, necesidad de ventilación mecánica en cualquier momento después de la extubación en el quirófano, tromboembolia pulmonar, arritmia, isquemia o infarto de miocardio e insuficiencia cardíaca clínica. Al realizar una regresión logística por pasos, los investigadores encontraron que la edad 1p 6 .0012 y el volumen espiratorio forzado postoperatorio 1p = .0032 fueron estadísticamente significativos para predecir la aparición de morbilidad cardiorrespiratoria.

For each of the data sets given in Exercises 22 through 29, do as many of the following as you think appropriate:

Para cada uno de los conjuntos de datos dados en los ejercicios 22 a 29, realice tantas de las siguientes operaciones como considere apropiadas:

(a) Apply one or more of the techniques discussed in this chapter.

(b) Apply one or more of the techniques discussed in previous chapters.

(c) Construct graphs.

(d) Formulate relevant hypotheses, perform the appropriate tests, and find p values.

(e) State the statistical decisions and clinical conclusions that the results of your hypothesis tests justify.

(f) Describe the population(s) to which you think your inferences are applicable.

22. A study by Davies et al. (A-17) was motivated by the fact that, in previous studies of contractile responses to b -adrenoceptor agonists in single myocytes from failing and nonfailing human hearts, they had observed an age-related decline in maximum response to isoproterenol, at frequencies where the maximum response to high Ca2 + in the same cell was unchanged. For the present study, the investigators computed the isoproterenol/Ca2 + ratio (ISO/CA) from measurements taken on myocytes from patients ranging in age from 7 to 70 years. Subjects were classified as older ( 750 years) and younger. The following are the (ISO/CA) values, age, and myocyte source of subjects in the study. Myocyte sources were reported as donor and biopsy.

22. Un estudio de Davies et al. (A-17) fue motivado por el hecho de que, en estudios previos de respuestas contráctiles a agonistas de los receptores b-adrenérgicos en miocitos individuales de corazones humanos con insuficiencia y sin insuficiencia, habían observado una disminución relacionada con la edad en la respuesta máxima al isoproterenol, en frecuencias en las que la respuesta máxima al Ca2+ alto en la misma célula no había cambiado. Para el presente estudio, los investigadores calcularon la relación isoproterenol/Ca2+ (ISO/CA) a partir de mediciones tomadas en miocitos de pacientes con edades comprendidas entre 7 y 70 años. Los sujetos se clasificaron como mayores (750 años) y menores. A continuación se presentan los valores (ISO/CA), la edad y la fuente de miocitos de los sujetos del estudio. Las fuentes de miocitos se informaron como donante y biopsia.

23. Hayton et al. (A-18) investigated the pharmacokinetics and bioavailability of cefetamet and cefetamet pivoxil in infants between the ages of 3.5 and 17.3 months who had received the antibiotic during and after urological surgery. Among the pharmacokinetic data collected were the following measurements of the steady-state apparent volume of distribution (V). Also shown are previously collected data on children ages 3 to 12 years (A-19) and adults (A-20). Weights (W) of subjects are also shown.

23. Hayton et al. (A-18) investigaron la farmacocinética y la biodisponibilidad de cefetamet y cefetamet pivoxil en niños de entre 3,5 y 17,3 meses de edad que habían recibido el antibiótico durante y después de una cirugía urológica. Entre los datos farmacocinéticos recopilados se encontraban las siguientes mediciones del volumen de distribución aparente en estado estacionario (V). También se muestran datos recopilados previamente sobre niños de entre 3 y 12 años de edad (A-19) y adultos (A-20). También se muestran los pesos (W) de los sujetos.

24. According to Fils-Aime et al. (A-21), epidemiologic surveys have found that alcoholism is the most common mental or substance abuse disorder among men in the United States. Fils-Aime and associates investigated the interrelationships of age at onset of excessive alcohol consumption, family history of alcoholism, psychiatric comorbidity, and cerebrospinal fluid (CSF) monoamine metabolite concentrations in abstinent, treatment-seeking alcoholics. Subjects were mostly white males classified as experiencing early (25 years or younger) or late (older than 25 years) onset of excessive alcohol consumption. Among the data collected were the following measurements on CSF tryptophan (TRYPT) and 5-hydroxyindoleacetic acid (5-HIAA) concentrations (pmol/ml).

24. Según Fils-Aime et al. (A-21), las encuestas epidemiológicas han descubierto que el alcoholismo es el trastorno mental o de abuso de sustancias más común entre los hombres en los Estados Unidos. Fils-Aime y colaboradores investigaron las interrelaciones entre la edad de inicio del consumo excesivo de alcohol, los antecedentes familiares de alcoholismo, la comorbilidad psiquiátrica y las concentraciones de metabolitos de monoamina en el líquido cefalorraquídeo (LCR) en alcohólicos abstinentes que buscaban tratamiento. Los sujetos eran en su mayoría varones blancos clasificados como que experimentaban un inicio temprano (25 años o menos) o tardío (más de 25 años) de consumo excesivo de alcohol. Entre los datos recopilados se encontraban las siguientes mediciones de las concentraciones de triptófano (TRYPT) y ácido 5-hidroxiindolacético (5-HIAA) en el LCR (pmol/ml).

25. The objective of a study by Abrahamsson et al. (A-22) was to investigate the anti-thrombotic effects of an inhibitor of the plasminogen activator inhibitor-1 (PAI-1) in rats given endotoxin. Experimental subjects were male Sprague–Dawley rats weighing between 300 and 400 grams. Among the data collected were the following measurements on PAI-1 activity and the lung 125I-concentration in anesthetized rats given three drugs:

25. El objetivo de un estudio realizado por Abrahamsson et al. (A-22) fue investigar los efectos antitrombóticos de un inhibidor del inhibidor del activador del plasminógeno-1 (PAI-1) en ratas a las que se les administró endotoxina. Los sujetos experimentales fueron ratas Sprague-Dawley macho que pesaban entre 300 y 400 gramos. Entre los datos recopilados se encontraban las siguientes mediciones de la actividad del PAI-1 y la concentración pulmonar de 125I en ratas anestesiadas a las que se les administraron tres fármacos:

26. Pearse and Sylvester (A-23) conducted a study to determine the separate contributions of ischemia and extracorporeal perfusion to vascular injury occurring in isolated sheep lungs and to determine the oxygen dependence of this injury. Lungs were subjected to ischemia alone, extracorporeal perfusion alone, and both ischemia and extracorporeal perfusion. Among the data collected were the following observations on change in pulmonary arterial pressure (mm Hg) and pulmonary vascular permeability assessed by estimation of the reflection coefficient for albumin in perfused lungs with and without preceding ischemia:

26. Pearse y Sylvester (A-23) realizaron un estudio para determinar las contribuciones independientes de la isquemia y la perfusión extracorpórea a la lesión vascular que se produce en pulmones aislados de ovejas y para determinar la dependencia del oxígeno de esta lesión. Los pulmones se sometieron a isquemia únicamente, a perfusión extracorpórea únicamente y a ambas, isquemia y perfusión extracorpórea. Entre los datos recopilados se encontraban las siguientes observaciones sobre el cambio en la presión arterial pulmonar (mmHg) y la permeabilidad vascular pulmonar evaluada mediante la estimación del coeficiente de reflexión de la albúmina en pulmones perfundidos con y sin isquemia previa:

27. The purpose of a study by Balzamo et al. (A-24) was to investigate, in anesthetized rabbits, the effects of mechanical ventilation on the concentration of substance P (SP) measured by radioimmunoassay in nerves and muscles associated with ventilation and participating in the sensory innervation of the respiratory apparatus and heart. SP is a neurotransmitter located in primary sensory neurons in the central and autonomic nervous systems. Among the data collected were the following measures of SP concentration in cervical vagus nerves (X) and corresponding nodose ganglia (NG), right and left sides:

27. El objetivo de un estudio realizado por Balzamo et al. (A-24) fue investigar, en conejos anestesiados, los efectos de la ventilación mecánica sobre la concentración de sustancia P (SP) medida por radioinmunoensayo en nervios y músculos asociados con la ventilación y que participan en la inervación sensorial del aparato respiratorio y el corazón. La SP es un neurotransmisor ubicado en neuronas sensoriales primarias en los sistemas nerviosos central y autónomo. Entre los datos recopilados se encuentran las siguientes mediciones de concentración de SP en los nervios vagos cervicales (X) y los ganglios nodosos correspondientes (NG), lados derecho e izquierdo:

28. Scheeringa and Zeanah (A-25) examined the presence of posttraumatic stress disorder (PTSD), the severity of posttraumatic symptomatology, and the pattern of expression of symptom clusters in relation to six independent variables that may be salient to the development of a posttraumatic disorder in children under 48 months of age. The following data were collected during the course of the study.

28. Scheeringa y Zeanah (A-25) examinaron la presencia de trastorno de estrés postraumático (TEPT), la gravedad de la sintomatología postraumática y el patrón de expresión de los grupos de síntomas en relación con seis variables independientes que pueden ser relevantes para el desarrollo de un trastorno postraumático en niños menores de 48 meses de edad. Los siguientes datos se recopilaron durante el curso del estudio.

Wit./exper.

0 = subject witnessed but did not directly experience trauma

1 = subject directly experienced the trauma

Threat to caregiver

0 = caregiver was not threatened in the trauma

1 = caregiver was threatened in the trauma

Reexp " Reexperiencing cluster symptom count

Numb " Numbing of responsiveness/avoidance cluster symptom count

Arous " Hyperarousal cluster symptom count

FrAgg " New fears/aggression cluster symptom count

29. One of the objectives of a study by Mulloy and McNicholas (A-26) was to compare ventilation and gas exchange during sleep and exercise in chronic obstructive pulmonary disease (COPD). The investigators wished also to determine whether exercise studies could aid in the prediction of nocturnal desaturation in COPD. Subjects (13 male, 6 female) were ambulatory patients attending an outpatient respiratory clinic. The mean age of the patients, all of whom had severe, stable COPD, was 64.8 years with a standard deviation of 5.2. Among the data collected were measurements on the following variables:

29. Uno de los objetivos de un estudio realizado por Mulloy y McNicholas (A-26) fue comparar la ventilación y el intercambio de gases durante el sueño y el ejercicio en la enfermedad pulmonar obstructiva crónica (EPOC). Los investigadores también deseaban determinar si los estudios de ejercicio podían ayudar a predecir la desaturación nocturna en la EPOC. Los sujetos (13 hombres, 6 mujeres) eran pacientes ambulatorios que asistían a una clínica respiratoria ambulatoria. La edad media de los pacientes, todos ellos con EPOC grave y estable, era de 64,8 años con una desviación estándar de 5,2. Entre los datos recopilados se encontraban mediciones de las siguientes variables:

Exercises for Use with the Large Data Sets Available on the Following Website:

www.wiley.com/college/daniel

1. The goal of a study by Gyurcsik et al. (A-27) was to examine the usefulness of aquatic exercise-related goals, task self-efficacy, and scheduling self-efficacy for predicting aquatic exercise attendance by individuals with arthritis. The researchers collected data on 142 subjects participating in Arthritis Foundation Aquatics Programs. The outcome variable was the percentage of sessions attended over an 8-week period (ATTEND). The following predictor variables are all centered values. Thus, for each participant, the mean for all participants is subtracted from the individual score. The variables are:

1. El objetivo de un estudio realizado por Gyurcsik et al. (A-27) fue examinar la utilidad de las metas relacionadas con el ejercicio acuático, la autoeficacia en la tarea y la autoeficacia en la programación para predecir la asistencia al ejercicio acuático por parte de personas con artritis. Los investigadores recopilaron datos sobre 142 sujetos que participaban en los programas acuáticos de la Arthritis Foundation. La variable de resultado fue el porcentaje de sesiones a las que se asistió durante un período de 8 semanas (ATTEND). Las siguientes variables predictoras son todas valores centrados. Por lo tanto, para cada participante, la media de todos los participantes se resta de la puntuación individual. Las variables son:

GOALDIFF—higher values indicate setting goals of higher participation.

GOALSPEC—higher values indicate higher specificity of goals related to aquatic exercise.

INTER—interaction of GOALDIFF and GOALSPEC.

TSE—higher values indicate participants’ confidence in their abilities to attend aquatic

classes.

SSE—higher values indicate participants’ confidence in their abilities to perform eight tasks

related to scheduling exercise into their daily routine for 8 weeks.

MONTHS—months of participation in aquatic exercise prior to start of study.

With the data set AQUATICS, perform a multiple regression to predict ATTEND with each of the above variables. What is the multiple correlation coefficient? What variables are significant in predicting ATTEND? What are your conclusions?

Con el conjunto de datos AQUATICS, realice una regresión múltiple para predecir ATTEND con cada una de las variables anteriores. ¿Cuál es el coeficiente de correlación múltiple? ¿Qué variables son significativas para predecir ATTEND? ¿Cuáles son sus conclusiones?

2. Rodehorst (A-28) conducted a prospective study of 212 rural elementary school teachers. The main outcome variable was the teachers’ intent to manage children demonstrating symptoms of asthma in their classrooms. This variable was measured with a single-item question that used a seven-point Likert scale (INTENT, with possible responses of 1 = extremely probable to 7 = extremely improbable). Rodehorst used the following variables as independent variables to predict INTENT:

2. Rodehorst (A-28) realizó un estudio prospectivo de 212 maestros de escuelas primarias rurales. La variable de resultado principal fue la intención de los maestros de manejar a los niños que presentaban síntomas de asma en sus aulas. Esta variable se midió con una pregunta de un solo ítem que utilizó una escala Likert de siete puntos (INTENT, con posibles respuestas de 1 = extremadamente probable a 7 = extremadamente improbable). Rodehorst utilizó las siguientes variables como variables independientes para predecir la INTENCIÓN:

SS = Social Support. Scores range from 7 to 49, with higher scores indicating higher perceived social support for managing children with asthma in a school setting.

ATT = Attitude. Scores range from 15 to 90, with higher scores indicating more favorable attitudes toward asthma.

KNOW = Knowledge. Scores range from 0 to 24, with higher scores indicating higher general knowledge about asthma.

CHILD = Number of children with asthma the teacher has had in his or her class during his or her entire teaching career.

SE = Self-efficacy. Scores range from 12 to 60, with higher scores indicating higher self-efficacy for managing children with asthma in the school setting.

YRS = Years of teaching experience.

With the data TEACHERS, use stepwise regression analysis to select the most useful variables to include in a model for predicting INTENT.

3. Refer to the weight loss data on 588 cancer patients and 600 healthy controls (WGTLOSS). Weight loss among cancer patients is a well-known phenomenon. Of interest to clinicians is the role played in the process by metabolic abnormalities. One investigation into the relationships among these variables yielded data on whole-body protein turnover (Y ) and percentage of ideal body weight for height (X ). Subjects were lung cancer patients and healthy controls of the same age. Select a simple random sample of size 15 from each group and do the following:

3. Consulte los datos de pérdida de peso de 588 pacientes con cáncer y 600 controles sanos (WGTLOSS). La pérdida de peso entre los pacientes con cáncer es un fenómeno bien conocido. A los médicos les interesa el papel que desempeñan las anomalías metabólicas en el proceso. Una investigación sobre las relaciones entre estas variables arrojó datos sobre el recambio proteico de todo el cuerpo (Y ) y el porcentaje del peso corporal ideal para la altura (X ). Los sujetos eran pacientes con cáncer de pulmón y controles sanos de la misma edad. Seleccione una muestra aleatoria simple de tamaño 15 de cada grupo y haga lo siguiente:

(a) Draw a scatter diagram of the sample data using different symbols for each of the two groups.

(b) Use dummy variable coding to analyze these data.

(c) Plot the two regression lines on the scatter diagram. May one conclude that the two sampled

populations differ with respect to mean protein turnover when percentage of ideal weight is taken

into account?

May one conclude that there is interaction between health status and percentage of ideal body weight? Prepare a verbal interpretation of the results of your analysis and compare your results with those of your classmates.