9

Regresión lineal

Dada una variable dependiente y un conjunto de una o más variables independientes, todas ellas cuantitativas, la regresión lineal consiste en obtener una función lineal de las variables independientes que permita explicar o predecir el valor de la dependiente.

Supongamos que se sospecha que, en los pacientes con úlcera péptica que han seguido un tratamiento, el tiempo que tarda en reaparecer la sintomatología ulcerosa está relacionado con el tiempo que tarda el paciente en responder al tratamiento. Para comprobarlo, se somete al tratamiento a un conjunto de pacientes con úlcera péptica, siendo todos ellos fumadores, y periódicamente (cada semana) se comprueba si la sintomatología ulcerosa persiste o ha desaparecido. Una vez desaparecida, el paciente sigue sometido a revisiones mensuales para comprobar el tiempo que tardan en reaparecer los síntomas. Antes de comenzar el tratamiento, algunos de los pacientes han decidido abandonar el hábito de fumar, por lo que se sospecha que en la reaparición de los síntomas, además del tiempo de respuesta al tratamiento, puede influir el abandono del tabaco, así como otros aspectos relacionados con los hábitos del individuo, tales como el consumo de alcohol, de café o de antiácidos. Para predecir el tiempo de reaparición de los síntomas, conocidos el tiempo de respuesta al tratamiento y los distintos hábitos del paciente, se aplicará el análisis de regresión lineal.

FORMULACION DEL PROBLEMA

A partir de $(y_1, x_{11}, \ldots, x_{1p}), \ldots, (y_n, x_{n1}, \ldots, x_{np})$, muestra de n observaciones de las variables $Y, X_1, \ldots, X_p$ se trata de aproximar los valores de Y, variable dependiente, mediante una función de las variables $X_1, \ldots, X_p$ variables independientes, que exprese la asociación lineal entre Y y $X_1, \ldots, X_p$:

$$Y = \beta_1 X_1 + \ldots + \beta_p X_p + \beta_0 + e$$

donde β_0 ,.., β_p son parámetros desconocidos a estimar y e es una variable error $N(0,\sigma^2)$. En particular, para cada observación se tendrá:

$$y_i = \beta_1 x_{i2} + \ldots + \beta_p x_{ip} + \beta_0 + e_i \qquad i = 1, \ldots, n$$

donde los e_i, $i = 1, \ldots, n$, son independientes entre sí.

En nuestro ejemplo, a partir de una muestra de 312 observaciones de las variables:

REAPARIC	Tiempo de reaparición de la sintomatología ulcerosa (en meses).
RESPUEST	Tiempo de respuesta al tratamiento de la sintomatología ulcerosa (en semanas).
TABACO	El paciente ha dejado de fumar durante el tratamiento. Valores: Sí y No, codificados numéricamente como 1 y 2 respectivamente.
ALCOHOL	Consumo de alcohol (gramos diarios).
CAFE	Consumo de café. Valores: 0, ..., 9 (de nada a mucho).
ANTIACID	Consumo de antiácidos. Valores: 0, ..., 9 (de nada a mucho).

Se trata de obtener una función lineal de las variables independientes *RESPUEST*, *TABACO*,*ALCOHOL*, *CAFE* y *ANTIACID*que permita estimar el tiempo de reaparición de los síntomas, *REAPARIC*, para cualquier paciente *i*:

$$REAPARIC_i = \beta_1 RESPUEST_i + \beta_2 TABACO_i + \beta_3 ALCOHOL_i + \\ + \beta_4 CAFE_i + \beta_5 ANTIACID_i + \beta_0 + e_i$$

donde $\beta_0, \ldots, \beta_5$ son parámetros desconocidos a estimar y los e_i proceden de variables independientes, normales, de media 0 y la misma varianza.

Obsérvese que la variable *TABACO*, aunque sus valores hayan sido codificados como números, es cualitativa, por lo que, en principio, no debería ser considerada en el análisis. En los siguientes apartados ignoraremos su existencia, pero al final del capítulo veremos cómo manipular sus valores para que pueda ser introducida en la ecuación de regresión.

Antes de proceder a la estimación del modelo de regresión analicemos, mediante el coeficiente de correlación de Pearson, el grado de asociación lineal entre cada par de variables.

ANALISIS DE LA CORRELACION ENTRE PARES DE VARIABLES: EL COEFICIENTE DE CORRELACION LINEAL SIMPLE

El coeficiente de correlación lineal simple, ρ, mide el grado de asociación lineal entre dos variables medidas en escala de intervalo o de razón, tomando valores

entre –1 y 1. Valores de ρ próximos a 1 indicarán fuerte asociación lineal positiva (a medida que aumentan los valores de una de las dos variables, aumentan los de la otra); valores de ρ próximos a –1 indicarán fuerte asociación lineal negativa (a medida que aumentan los valores de una de las dos variables, disminuyen los de la otra), y valores de ρ próximos a 0 indicarán no asociación lineal (lo que no significa que no pueda existir otro tipo de asociación). El estimador muestral para ρ es el coeficiente de correlación muestral, r.

El coeficiente de correlación es una medida del grado de asociación lineal que depende del tamaño muestral: un mismo valor del coeficiente de correlación muestral calculado a partir de muestras de distinto tamaño de dos pares de variables, no corresponde a un mismo grado de asociación lineal. Para determinar si la asociación es estadísticamente significativa, se puede plantear la hipótesis nula de que el coeficiente de correlación lineal es igual a cero:

$$H_0: \rho = 0$$

El estadístico de contraste se construye a partir del coeficiente de correlación muestral, r, y del tamaño de la muestra, n. Si el p-valor asociado es menor que α, se rechazará la hipótesis nula al nivel de significación α.

La matriz de correlaciones entre las variables cuantitativas *REAPARIC*, *RESPUEST*, *ALCOHOL*, *CAFE* y *ANTIACID* se solicita en el Cuadro de diálogo 9.1. Los resultados se disponen en la Figura 9.1. La matriz de correlaciones es una matriz simétrica respecto a la diagonal principal y con unos en dicha diagonal, por lo que basta con analizar los elementos situados por encima o por debajo de ella. Si centramos nuestra atención en la relación entre la variable dependiente *REAPARIC* y cada una de las independientes, dado que en todos los casos el tamaño muestral es el mismo ($n = 312$) y, por tanto, los distintos valores de r son comparables, podemos observar que la máxima asociación lineal corresponde a la variable *RESPUEST* ($r = -0{,}769$). Para determinar si dicha asociación es estadísticamente significativa, podemos contrastar la hipótesis nula de que las variables *REAPARIC* y *RESPUEST* están incorreladas:

$$H_0: = \rho_{REAPARIC,\ RESPUEST} = 0$$

```
ESTADISTICA → CORRELACIONES → BIVARIADAS  En el Cuadro de diálogo

VARIABLES: REAPARIC, RESPUEST, ALCOHOL, CAFE, ANTIACID
COEFICIENTES DE CORRELACION: PEARSON
ACEPTAR
```

CUADRO DE DIALOGO 9.1. Matriz de correlaciones entre las variables *REAPARIC*, *RESPUEST*, *ALCOHOL*, *CAFE* y *ANTIACID*.

- - Correlation Coefficients - -

	REAPARIC	RESPUEST	ALCOHOL	CAFE	ANTIACID
REAPARIC	1,0000	-,7690	-,5465	-,2553	,3008
	(312)	(312)	(312)	(312)	(312)
	P= ,	P= ,000	P= ,000	P= ,000	P= ,000
RESPUEST	-,7690	1,0000	,0053	-,0530	,0074
	(312)	(312)	(312)	(312)	(312)
	P= ,000	P= ,	P= ,926	P= ,351	P= ,897
ALCOHOL	-,5465	,0053	1,0000	,5612	-,5855
	(312)	(312)	(312)	(312)	(312)
	P= ,000	P= ,926	P= ,	P= ,000	P= ,000
CAFE	-,2553	-,0530	,5612	1,0000	-,4202
	(312)	(312)	(312)	(312)	(312)
	P= ,000	P= ,351	P= ,000	P= ,	P= ,000
ANTIACID	,3008	,0074	-,5855	-,4202	1,0000
	(312)	(312)	(312)	(312)	(312)
	P= ,000	P= ,897	P= ,000	P= ,000	P= ,

(Coefficient / (Cases) / 2-tailed Significance)

FIGURA 9.1. Matriz de correlaciones entre las variables *REAPARIC*, *RESPUEST*, *ALCOHOL*, *CAFE* y *ANTIACID*.

El *p*-valor asociado al estadístico de contraste («*P* =0,000») es menor que 0,05, luego, al nivel de significación 0,05, se puede rechazar la hipótesis nula. Con este mismo criterio podemos observar que la asociación lineal entre la variable *REAPARIC* y cada una de las independientes es estadísticamente significativa. Respecto a la relación entre las variables independientes, mientras todos los posibles pares de asociaciones entre las variables *ALCOHOL*, *CAFE* y *ANTIACID* son estadísticamente significativas, en el caso de la variable *RESPUEST* la hipótesis de incorrelación respecto a cada una de las tres anteriores no puede ser rechazada. Teniendo en cuenta el signo de las correlaciones estadísticamente significativas, podemos concluir que:

— a mayor tiempo de respuesta al tratamiento, menor es el tiempo de reaparición de los síntomas;
— a mayor consumo de alcohol, a mayor consumo de café y a menor consumo de antiácidos, menor es el tiempo de reaparición de los síntomas;

— a mayor consumo de alcohol, mayor es el consumo de café y menor el de antiácidos, y viceversa; y
— a mayor consumo de café, menor es el consumo de antiácidos y viceversa.

Dado que el mayor grado de asociación lineal detectado con la variable tiempo de reaparición de los síntomas corresponde a la variable tiempo de respuesta al tratamiento, en el siguiente apartado trataremos de predecir los valores de la variable *REAPARIC* a partir de los de *RESPUEST*, mediante el ajuste de una ecuación de regresión lineal simple.

REGRESION LINEAL SIMPLE

En el caso de una única variable independiente, X, se habla de regresión lineal simple. La correspondiente ecuación de regresión será del tipo:

$$Y = \beta_1 X + \beta_0 + e$$

y, en particular, para cada observación:

$$y_i = \beta_1 x_i + \beta_0 + e_i \qquad i = 1, \ldots, n$$

En nuestro caso, se trata de obtener una función lineal de la variable independiente, *RESPUEST*, que permita estimar el tiempo de reaparición de los síntomas, *REAPARIC*, para cualquier paciente i:

$$REAPARIC_i = \beta_1 RESPUEST_i + \beta_0 + e_i$$

donde β_0 y β_1 son parámetros desconocidos a estimar y los e_i proceden de variables independientes, normales, de media 0 y la misma varianza.

Estimación de los parámetros

El criterio para obtener los coeficientes de regresión, B_0 y B_1, estimaciones de los parámetros desconocidos β_0 y β_1 respectivamente, es el de mínimos cuadrados, que consiste en minimizar la suma de los cuadrados de los residuos. Si $\hat{y}_i$ es la estimación de y_i mediante el modelo de regresión lineal:

$$\hat{y}_i = B_1 x_i + B_0$$

el residuo correspondiente, E_i, será la desviación de cada observación al valor estimado:

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

La Figura 9.2 (proporcionada por el Cuadro de diálogo 9.2) muestra la representación gráfica de los valores de *REAPARIC* frente a los valores de *RESPUEST*. Se observa que, confirmando la asociación lineal negativa detectada mediante el coeficiente de correlación lineal simple, a medida que aumentan los valores de la variable *RESPUEST* disminuyen los de *REAPARIC*. Si B_0 y B_1 son las estimaciones de los parámetros del modelo, la recta de pendiente B_1 y término independiente B_0:

$$\hat{REAPARIC} = B_1\ RESPUEST + B_0$$

atravesará la nube de puntos. Para un paciente i con valores observados iguales a $RESPUEST_i$ y $REAPARIC_i$, la estimación del valor de *REAPARIC* vendrá dada por

```
GRAFICAS → DISPERSION  En el Cuadro de diálogo

SIMPLE→ DEFINIR        En el Cuadro de diálogo

EJE Y: REAPARIC
EJE X: RESPUEST
ACEPTAR
```

CUADRO DE DIALOGO 9.2. Gráfico de los valores de *REAPARIC* frente a los valores de *RESPUEST*.

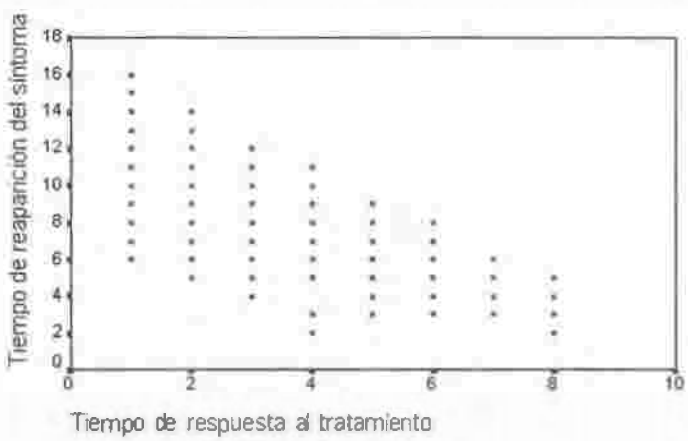


FIGURA 9.2. Gráfico de los valores de *REAPARIC* frente a los valores de *RESPUEST*.

el valor en ordenadas para el punto correspondiente en la recta al valor en abscisas *RESPUEST*$_i$. El residuo será igual a la diferencia entre los valores observado y estimado para *REAPARIC*. La recta de regresión es tal que, para cualquier otra recta que atraviese la nube de puntos, la suma de los cuadrados de las distancias entre los valores observado y estimado de *REAPARIC* es mayor.

El análisis de regresión lineal simple de la variable *REAPARIC* sobre la variable *RESPUEST* se solicita en el Cuadro de diálogo 9.3. Los resultados se disponen en la Figura 9.3. Las estimaciones de los parámetros del modelo (columna «B» en el bloque «Variables in the Equation») son:

$$B_1 = -1{,}207042 \qquad \text{y} \qquad B_0 = 12{,}186087$$

En particular, para cada valor del tiempo de respuesta al tratamiento, la predicción o estimación del tiempo de reaparición vendrá dada por:

$$\widehat{REAPARIC}_i = B_1 RESPUEST_i + B_0 = -1{,}21\ RESPUST_i + 12{,}19 \quad i = 1, \ldots, 312$$

y el residuo será igual a:

$$E_i = REAPARIC_i - \widehat{REAPARIC}_i$$

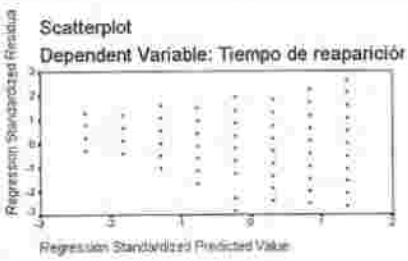
Por ejemplo, si el tiempo de respuesta al tratamiento para un paciente hubiera sido de 5 semanas, el tiempo estimado de reaparición de los síntomas (en meses) mediante la función de regresión sería:

$$-1{,}21 \times 5 + 12{,}19 = 6{,}14$$

```
ESTADISTICA → REGRESION → LINEAL  En el Cuadro de diálogo

DEPENDIENTE: REAPARIC
INDEPENDIENTE(S): RESPUEST
METODO: INTRODUCIR
ESTADISTICOS    En el Cuadro de diálogo
        COEFICIENTES DE REGRESION: ESTIMACIONES
        DURBIN-WATSON
        CONTINUAR
GRAFICAS: En el Cuadro de diálogo
        Y: *ZRESID
        X: *ZPRED
        GRAFICOS DE RESIDUOS TIPIFICADOS: GRAFICO DE PROBABILIDAD NORMAL
CONTINUAR
ACEPTAR
```

CUADRO DE DIALOGO 9.3. Regresión lineal simple de la variable *REAPARIC* sobre la variable *RESPUEST*.

```
* * * *  M U L T I P L E   R E G R E S S I O N  * * * *

Equation Number 1    Dependent Variable..   REAPARIC   Tiempo de
reaparición de los síntomas

Block Number  1.  Method:  Enter       RESPUEST

Variable(s) Entered on Step Number
   1..     RESPUEST  Tiempo de respuesta al tratamiento

Multiple R            ,76897
R Square              ,59132
Adjusted R Square     ,59000
Standard Error       1,89211

Analysis of Variance
                 DF        Sum of Squares        Mean Square
Regression        1          1605,78389          1605,78389
Residual        310          1109,82829             3,58009

F =     448,53155        Signif F =  ,0000

---------------------- variables in the Equation -----------------

Variable                B         SE B        Beta           T  Sig T

RESPUEST        -1,207042     ,056994    -,768970    -21,179   ,0000
(Constant)      12,186087     ,230292                 52,916   ,0000

Durbin-Watson Test =   1,79500
```



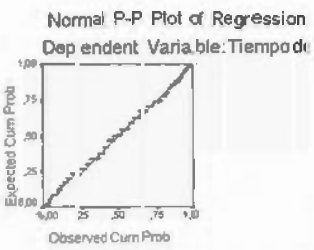


FIGURA 9.3. Regresión lineal simple de la variable *REAPARIC* sobre la variable *RESPUEST*.

Si el tiempo de reaparición observado en ese mismo paciente hubiera sido de 8 meses, el residuo sería igual a:

$$8-6,14=1,86$$

mientras que si hubiera sido de 4 meses, sería igual a:

$$4-6,14=-2,14$$

Análisis de los residuos

En la regresión lineal se supone que los verdaderos errores, e_i, son independientes con distribución $N(0, \sigma^2)$. Los residuos, E_i, son las estimaciones de los verdaderos errores, y la estimación de σ^2 es la media de los cuadrados de los residuos, s^2, donde s es el error típico de la estimación («Standard Error = 1,89211»). Si el modelo ajustado es apropiado, los residuos deben presentar características similares.

El hecho de que la media de los residuos sea igual a cero es consecuencia del método de estimación de los parámetros de la función de regresión.

Respecto a la normalidad, la distribución de la variable formada por los residuos debe ser Normal: los residuos observados y los esperados bajo hipótesis de distribución Normal deben ser parecidos. Para comprobarlo, una alternativa es el gráfico de probabilidad normal, que permite comparar, gráficamente, la función de distribución observada en la muestra con la función de distribución de una Normal(0, 1) (por lo que la variable objeto de análisis debe tener media 0 y desviación típica 1). En el gráfico de probabilidad normal para los residuos tipificados (en el ángulo inferior derecho de la Figura 9.3) se representa la función de distribución esperada bajo la hipótesis de distribución Normal(0, 1), en el eje vertical, frente a la función de distribución acumulativa de los valores observados, en el horizontal. Si la distribución de los residuos fuera Normal, dichos valores deberían ser aproximadamente iguales y, en consecuencia, los puntos del gráfico estarían situados, como ocurre en este caso, sobre la recta que pasa por el origen con pendiente igual a 1. Luego podemos aceptar que los residuos proceden de una distribución Normal.

Respecto a la independencia, el valor observado en una variable para un individuo no debe estar influenciado en ningún sentido por los valores de esta variable observados en otros individuos: los residuos no deben presentar ningún patrón sistemático respecto a la secuencia de observación. El estadístico de Durbin-Watson, D, mide el grado de autocorrelación entre el residuo correspondiente a cada observación y la anterior. Si su valor es próximo a 2, los residuos estarán incorrelados; si se aproxima a 4, estarán negativamente autocorrelados, y si se aproxima a 0, estarán positivamente autocorrelados. En nuestro caso, el hecho de que los residuos sean independientes es consecuencia directa de que las observaciones lo son (se supone que los pacientes observados son independientes entre sí). En cualquier caso, el valor del estadístico de Durbin-Watson («Durbin-Watson Test= 1.795») es próximo a 2, lo que confirma la incorrelación de los residuos.

Las varianzas de las distribuciones de la variable dependiente ligadas a los distintos valores de las variables independientes deben ser iguales: los residuos no deben presentar ningún patrón sistemático respecto de las predicciones o respecto de cada una de las variables independientes. Para analizar la homogeneidad de varianzas utilizaremos el gráfico de los residuos tipificados frente a las estimaciones tipificadas (en el ángulo inferior izquierdo de la Figura 9.3). Si la varianza de los residuos fuera constante, la nube de puntos estaría concentrada en una banda, centrada en el cero y paralela al eje de abscisas. Obsérvese que, en este caso, si nos desplazamos de izquierda a derecha en el gráfico, la dispersión de la nube de puntos va en aumento: a mayor valor en la estimación del tiempo de reaparición de los síntomas, mayor es la dispersión de los residuos. Confirmemos esta observación en el gráfico de los valores de *REAPARIC* frente a los de *RESPUEST*(Figura 9.2). A medida que aumenta el tiempo de respuesta al tratamiento (en abscisas) no sólo disminuye el tiempo de reaparición de los síntomas (en ordenadas) sino que, además, también disminuye la dispersión de las observaciones. Si los residuos presentaran varianza constante, la nube de puntos estaría concentrada en una banda, centrada en la recta de regresión y delimitada por dos rectas paralelas a la misma.

En aquellas situaciones en las que la dispersión aumenta o disminuye con la tendencia central, existe una familia de transformaciones que, en general, permite estabilizar la varianza. Antes de proceder a la búsqueda de la transformación adecuada, confirmemos, mediante la prueba de Levene, que la varianza de la variable *REAPARIC* para cada valor de la variable *RESPUEST* no es constante.

Prueba de Levene y transformaciones para estabilizar la varianza

La prueba de Levene permite contrastar la hipótesis de que la varianza de una variable Y en K subpoblaciones o grupos es la misma. Si denominamos σ_j^2 a la varianza de Y en la j-ésima subpoblación, $j = 1, \ldots, K$, la hipótesis nula que se desea contrastar es:

$$H_0: \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_K^2$$

Si el p-valor asociado al estadístico de contraste es menor que α, se rechazará la hipótesis nula de homogeneidad de varianzas al nivel de significación α.

En el caso particular de que la hipótesis de homogeneidad de varianzas sea rechazada debido a que la varianza cambia con la media, existe una familia de transformaciones que proporciona, en general, homogeneidad en varianzas. En la práctica, se utilizan transformaciones del tipo:

$$T(Y) = \begin{cases} Y^p & p \neq 0 \\ Ln\, Y & p = 0 \end{cases}$$

donde p es el múltiplo de un medio más próximo al poder de transformación proporcionado por el gráfico de nivel y dispersión («Spread-level»). El gráfico de nivel y dispersión representa, para cada uno de los K grupos, el logaritmo neperiano del rango intercuartílico (en ordenadas) frente al logaritmo neperiano de la mediana (en abscisas) de la variable Y. El poder de transformación es igual a 1 menos la pendiente de la recta de regresión mínimo-cuadrática ajustada a los K puntos. Esta familia de transformaciones no sólo permite estabilizar la varianza sino que, incluso, puede proporcionar normalidad. Por otro lado, sólo está definida para datos positivos, por lo que, en ocasiones, antes de realizar la transformación, será necesario sumar una misma constante a todos los valores de Y.

Particularizando al caso de la variable *REAPARIC*, y teniendo en cuenta que la variable *RESPUEST* toma valores enteros entre 1 y 8, ambos inclusive, la hipótesis nula que se desea contrastar es que la varianza de *REAPARIC* en cada uno de los ocho grupos establecidos por los valores de *RESPUEST* es la misma:

$$H_0: \sigma^2_{respuest=1} = \ldots = \sigma^2_{respuest=8}$$

Si dicha hipótesis, como es de esperar, fuera rechazada, trataríamos de encontrar una transformación que homogeneizara las varianzas.

La prueba de Levene, el gráfico de nivel y dispersión y la estimación del poder de transformación para estabilizar la varianza de la variable *REAPARIC* se solicita en la parte superior del Cuadro de diálogo 9.4. Los resultados se disponen en la parte superior de la Figura 9.4. El p-valor asociado al estadístico de Levene («Significance = 0,0004») es menor que 0,05. Luego, efectivamente, al nivel de significación 0,05, la hipótesis de homogeneidad de varianzas de *REAPARIC* en los ocho grupos establecidos por los valores de *RESPUEST* puede ser rechazada. El gráfico de nivel y dispersión representa el logaritmo neperiano del rango intercuartílico de la variable *REAPARIC* (como medida de la dispersión de las observaciones) frente al logaritmo neperiano de la mediana (como medida de la tendencia central) en cada uno de los 8 grupos. El poder de transformación, estimado a partir de la pendiente de la recta de regresión ajustada a la nube de puntos («Slope = 0,716»), es igual a 0,284. Por tanto, al redondear el poder de transformación al múltiplo de un medio más próximo, las transformaciones sugeridas son la raíz cuadrada ($p = 0,5$) y el logaritmo neperiano ($p = 0$).

Al solicitar (en la parte central del Cuadro de diálogo 9.4) la prueba de Levene sobre los valores de la variable *REAPARIC* con el poder de transformación igual a un medio o, lo que es lo mismo, sobre los valores transformados por la función raíz cuadrada, el p-valor asociado («Significance = 0,41», en la parte central de la Figura 9.4) es mayor que 0,05. Luego, al nivel de significación 0,05, no podría rechazarse la hipótesis nula de homogeneidad de varianzas. Sin embargo, al solicitar (en la parte inferior del Cuadro de diálogo 9.4) la prueba con el poder de transformación igual a cero o, lo que es lo mismo, sobre los valores transformados por la función logaritmo neperiano, el p-valor asociado («Significance = 0,4908», en la parte inferior de la Figura 9.4) es mayor que el obtenido en el caso anterior. Por

```
ESTADISTICA → RESUMIR → EXPLORAR  En el Cuadro de diálogo

LISTA DEPENDIENTE: REAPARIC
LISTA DE FACTORES: RESPUEST
MOSTRAR: GRAFICOS
GRAFICAS        En el Cuadro de diálogo
        DISPERSION POR NIVEL CON PRUEBA DE LEVENE: ESTIMACION DE POTENCIA
        CONTINUAR
ACEPTAR
```

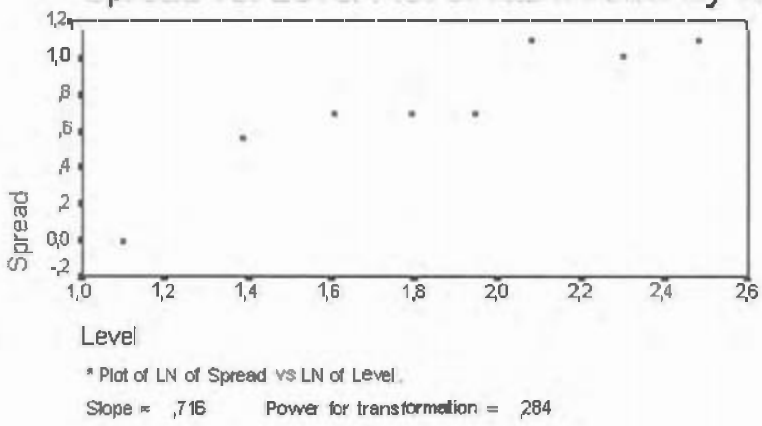
```
ESTADISTICA → RESUMIR → EXPLORAR  En el Cuadro de diálogo

LISTA DEPENDIENTE: REAPARIC
LISTA DE FACTORES: RESPUEST
MOSTRAR: GRAFICOS
GRAFICAS        En el Cuadro de diálogo
        DISPERSION POR NIVEL CON PRUEBA DE LEVENE: TRANSFORMADO: RAIZ
          CUADRADA
        CONTINUAR
ACEPTAR
```

```
ESTADISTICA → RESUMIR → EXPLORAR  En el Cuadro de diálogo

LISTA DEPENDIENTE: REAPARIC
LISTA DE FACTORES: RESPUEST
MOSTRAR: GRAFICOS
GRAFICAS        En el Cuadro de diálogo
        DISPERSION POR NIVEL CON PRUEBA DE LEVENE: TRANSFORMADO: LOG
          NATURAL
        CONTINUAR
ACEPTAR
```

CUADRO DE DIALOGO 9.4. En la parte superior: Prueba de Levene, gráfico de nivel y dispersión y estimación del poder de transformación para estabilizar la varianza de la variable *REAPARIC*; en el centro, prueba de Levene con la transformación raíz cuadrada, y en la parte inferior, prueba de Levene con la transformación logaritmo neperiano.

tanto, la transformación logaritmo neperiano proporciona una mayor estabilidad en varianzas que la raíz cuadrada. En consecuencia, el análisis proseguirá sobre la variable *LNREAPAR*, generada mediante el Cuadro de diálogo 9.5, cuyos valores son iguales al logaritmo neperiano de los valores de la variable *REAPARIC*.

La Figura 9.6 (proporcionada por el Cuadro de diálogo 9.6) muestra la representación gráfica de los valores de *LNREAPAR* frente a los valores de *RESPUEST*. En este caso, la nube de puntos sí está concentrada en una banda, centrada en la




```
* Plot of LN of Spread vs LN of Level.

Slope =     ,716          Power for transformation =     ,284

Test of homogeneity of variance          df1      df2      Significance
Levene Statistic              3,9687       7      304          ,0004
```

```
* Data transformed using P =  ,500

Slope=      ,074

Test of homogeneity of variance          df1      df2      Significance
Levene Statistic              1,0300       7      304          ,4100
```

```
* Data transformed using P =  0

Slope =    -,056

Test of homogeneity of variance          df1      df2      Significance
Levene Statistic               ,9204       7      304          ,4908
```

FIGURA 9.4. En la parte superior: prueba de Levene, gráfico de nivel y dispersión y estimación del poder de transformación para estabilizar la varianza de la variable *REAPARIC*; en el centro, prueba de Levene con la transformación raíz cuadrada, y en la parte inferior, prueba de Levene con la transformación logaritmo neperiano.

```
TRANSFORMAR → CALCULAR  En el Cuadro de diálogo

VARIABLE DE DESTINO: LNREAPAR
EXPRESION NUMERICA: LN(REAPARIC)
ACEPTAR
```

CUADRO DE DIALOGO 9.5. Generación de la variable *LNREAPAR*.

```
GRAFICAS → DISPERSION    En al Cuadro de diálogo

SIMPLE → DEFINIR         En el Cuadro de diálogo

EJE Y: LNREAPAR
EJE X: RESPUEST
ACEPTAR
```

CUADRO DE DIALOGO 9.6. Gráfico de los valores de *LNREAPAR* frente a los de *RESPUEST*.

FIGURA 9.6. Gráfico de los valores de *LNREAPAR* frente a los de *RESPUEST*.

recta de regresión y delimitada por dos rectas paralelas a la misma, sin que se observe tendencia creciente ni decreciente en la dispersión.

El análisis de regresión lineal simple de la variable *LNREAPAR* sobre la variable *RESPUEST* se solicita en el Cuadro de diálogo 9.7. Los resultados se disponen en la Figura 9.7. La ecuación de regresión obtenida («Variables in the Equation») es:

$$\widehat{LNREAPAR} = B_1\ RESPUEST + B_0 = -0{,}17\ RESPUEST + 2{,}58$$

Luego la estimación del tiempo de reaparición vendrá dada por:

$$\widehat{REAPARIC} = e^{B_1 RESPUEST + B_0} = e^{-0{,}17\ RESPUEST + 2{,}58}$$

Obsérvese que, al considerar la transformación sobre la variable dependiente, en el gráfico de los residuos tipificados frente a las estimaciones tipificadas (en la parte inferior de la Figura 9.7) la tendencia creciente de la dispersión de la nube de puntos ha sido eliminada (compárese con la representación gráfica en el ángulo inferior izquierdo de la Figura 9.3). Además, la correlación en valor absoluto entre las variables *LNREAPAR* y *RESPUEST* («Multiple $R = 0{,}77588$») es mayor que la obtenida entre *REAPARIC* y *RESPUEST* («Multiple $R = 0{,}76897$», en la Figura 9.3). En el caso de las variables *REAPARIC* y *RESPUEST* obtuvimos, mediante el *p*-valor correspondiente, que la correlación era estadísticamente significativa (Figura 9.1). Veamos a continuación cómo determinaríamos si, en este caso, el grado de asociación lineal es estadísticamente significativo. Dado que el número de observaciones sigue siendo el mismo, $n = 312$, y que la correlación en valor absoluto es mayor al considerar los valores de la variable *LNREAPAR*, con mayor motivo se rechazará la

```
ESTADISTICA → REGRESION → LINEAL     En el Cuadro de diálogo

DEPENDIENTE: LNREAPAR
INDEPENDIENTE(S): RESPUEST
METODO: INTRODUCIR
ESTADISTICA      En el Cuadro de diálogo
        COEFICIENTES DE REGRESION: ESTIMACIONES
        CONTINUAR
GRAFICAS         En el Cuadro de diálogo
        Y: *ZRESID
        X: *ZPRED
        CONTINUAR
ACEPTAR
```

CUADRO DE DIALOGO 9.7. Regresión lineal simple de la variable *LNREAPAR* sobre la variable *RESPUEST*.

```
* * * *  M U L T I P L E   R E G R E S S I O N  * * * *

Equation Number 1    Dependent Variable..   LNREAPAR   Logaritmo del
tiempo de reaparición

Block Number  1.  Method:  Enter      RESPUEST

Variable(s) Entered on Step Number
   1..     RESPUEST  Tiempo de respuesta al tratamiento

Multiple R            ,77588
R Square              ,60198
Adjusted R Square     ,60070
Standard Error        ,25484

Analysis of Variance
                 DF        Sum of Squares        Mean Square
Regression        1             30,44881           30,44881
Residual        310             20,13201             ,06494

F =     468,86194        Signif F =  ,0000

    ------------------ Variables in the Equation ------------------

Variable                B         SE B        Beta          T  Sig T

RESPUEST         -,166213     ,007676    -,775876    -21,653  ,0000
(Constant)      2,581571     ,031017                 83,232  ,0000
```

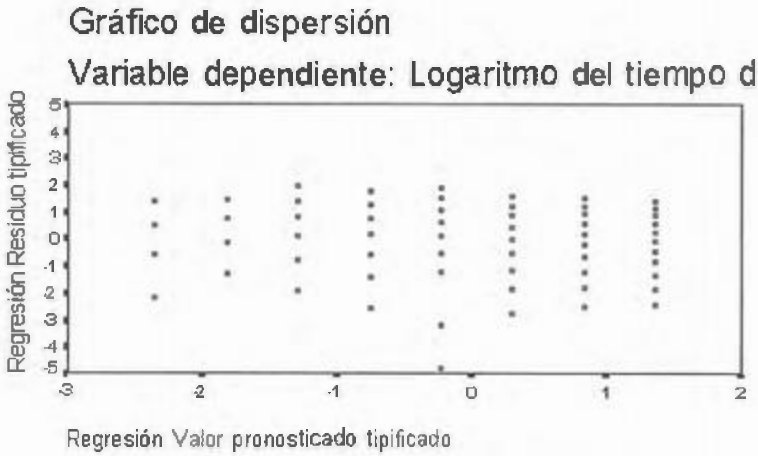


FIGURA 9.7. Regresión lineal simple de la variable *LNREAPAR* sobre la variable *RESPUEST*.

hipótesis nula de que las variables *LNREAPAR* y *RESPUEST* están incorreladas. En cualquier caso, veamos cómo se obtendría dicho resultado.

Análisis de la varianza y el coeficiente de determinación en el modelo simple

La tabla de análisis de la varianza permite comprobar hasta qué punto es adecuado el modelo de regresión lineal para estimar los valores de la variable dependiente. En el caso de no disponer de ninguna otra información más que la relativa a la muestra de observaciones de la propia variable, la estimación coincidiría con la media de las observaciones. Si las desviaciones de las estimaciones mediante el modelo lineal a la media de *LNREAPAR* fueran muy grandes respecto a las desviaciones de los valores observados a la estimación correspondiente (la pendiente de la recta de regresión muy distinta de cero y los puntos muy concentrados en torno a la recta), la variabilidad total sería debida a la dispersión de los valores de *RESPUEST*. Sin embargo, dado que se supone que la varianza de *LNREAPAR*, para los distintos valores de *RESPUEST*, es la misma, si las desviaciones de las estimaciones mediante el modelo lineal a la media de *LNREAPAR* fueran muy parecidas respecto a las desviaciones de los valores observados a la estimación correspondiente, la variabilidad total sería debida a la de los valores de *LNREAPAR*. El análisis de la varianza se realizará a partir de esta consideración, y obsérvese que su veracidad depende de la homogeneidad de las varianzas de los residuos. Luego la transformación realizada sobre la variable dependiente hace que el resultado proporcionado por el análisis de la varianza sea fiable.

El análisis de la varianza se basa en que la variabilidad total de la muestra puede descomponerse en la variabilidad explicada por la regresión y la variabilidad residual:

$$SC_{total} = SC_{reg} + SC_{res}$$

donde:

- SC_{total} mide las desviaciones de las observaciones, y_i, a la media muestral de Y.
- SC_{reg} mide las desviaciones de las estimaciones mediante el modelo de regresión lineal, $\hat{y}_i$, a la media muestral de Y.
- SC_{res} mide las desviaciones de las observaciones, y_i, a las estimaciones mediante el modelo de regresión lineal, $\hat{y}_i$.

La tabla de análisis de la varianza (Tabla 9.1) se construye a partir de esta descomposición y proporciona el estadístico F que permite contrastar la hipótosis nula de que la pendiente de la recta de regresión es igual a cero:

$$H_0: \beta_1 = 0$$

TABLA 9.1. Análisis de la varianza para el modelo de regresión lineal.

Fuente de variación	Suma de cuadrados	Grados de libertad	Medias de cuadrados	Estadístico F
Regresión	SC_{reg}	p	$MC_{reg} = \frac{SC_{reg}}{p}$	$F = \frac{MC_{reg}}{MC_{res}}$
Residual	SC_{res}	$n-p-1$	$MC_{res} = \frac{SC_{res}}{n-p-1}$	
Total	SC_{total}	$n-1$		

Además, se verifica que:

$$r^2 = \frac{SC_{reg}}{SC_{tot}}$$

donde r^2 es el cuadrado del coeficiente de correlación muestral, estimación del cuadrado del coeficiente de correlación, ρ^2, al que se denomina coeficiente de determinación. En consecuencia, el coeficiente de determinación puede interpretarse como la proporción de variabilidad total de la variable dependiente explicada mediante la recta de regresión. La hipótesis anterior será entonces equivalente a la hipótesis:

$$H_0: \rho^2 = 0$$

o, lo que es lo mismo, a la hipótesis de que Y y X están incorreladas.

Obsérvese que, según la expresión del estadístico F, cuanto mayor sea su valor mejor será la predicción mediante el modelo lineal respecto a la predicción mediante la media muestral. Si el p-valor asociado a F es menor que α, se rechazarán las dos hipótesis nulas planteadas al nivel de significación α.

Los resultados del análisis de la varianza se proporcionan en el bloque encabezado por «Analysis of Variance» (en la Figura 9.7). El p-valor asociado al estadístico F («Signif F = 0,000») es menor que 0,05, luego, al nivel de significación 0,05, se rechaza la hipótesis nula de que la pendiente de la recta de regresión, cuyo valor estimado es $B_1 = -0,166213$, es igual a cero o, equivalentemente, la hipótesis nula de que las variables *LNREAPAR* y *RESPUEST* están incorreladas. Por tanto, el modelo de regresión lineal es adecuado para mejorar nuestra estimación de los valores de la variable *LNREAPAR* y, además, la proporción de variabilidad de dicha variable explicada mediante el mismo es $r^2 = 0,60198$ («R Square»).

4. REGRESION LINEAL MULTIPLE

El hecho de que el modelo de regresión lineal simple sea adecuado no significa que no pueda ser mejorado a través de la información proporcionada por otras variables. En particular, recordando que la segunda variable más correlada con la variable *REAPARIC* era *ALCOHOL* (Figura 9.1), puede suceder que, al incorporar esta segunda variable al modelo, la proporción de variabilidad explicada aumente significativamente. Para comprobarlo, estimaremos los coeficientes del modelo de regresión lineal múltiple de la forma:

$$\widehat{LNREAPAR} = \beta_1\ RESPUEST + \beta_2\ ALCOHOL + \beta_0$$

El análisis de regresión lineal múltiple de la variable dependiente *LNREAPAR* sobre las variables independientes *RESPUEST* y *ALCOHOL* se solicita en el Cuadro de diálogo 9.8. Los resultados se disponen en la Figura 9.8. La ecuación de regresión obtenida («Variables in the Equation») es:

$$\begin{aligned} LNREAPAR &= B_1\ RESPUEST + B_2 ALCOHOL + B_0 \\ &= -0{,}17\ RESPUEST - 0{,}01\ ALCOHOL + 3{,}23 \end{aligned}$$

Luego la estimación del tiempo de reaparición vendrá dada por:

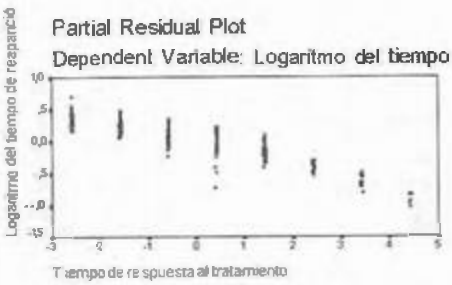
$$\begin{aligned} \widehat{REAPARIC} &= e^{B_1\ RESPUEST + B_2 ALCOHOL + B_0} \\ &= e^{-0{,}17\ RESPUEST - 0{,}01\ ALCOHOL + 3{,}23} \end{aligned}$$

Hasta ahora conocemos el grado de asociación lineal entre las variables *LNREAPAR* y *RESPUEST* y la proporción de variabilidad de la primera explicada

```
ESTADISTICA → REGRESION → LINEAL      En el Cuadro de diálogo

DEPENDIENTE: LNREAPAR
INDEPENDIENTE(S): RESPUEST, ALCOHOL
METODO: INTRODUCIR
ESTADISTICOS    En el Cuadro de diálogo
        COEFICIENTES DE REGRESION: ESTIMACIONES
        CONTINUAR
GRAFICAS    En el Cuadro de diálogo
        GENERAR TODOS LOS GRAFICOS PARCIALES
        CONTINUAR
ACEPTAR
```

CUADRO DE DIALOGO 9.8. Regresión lineal múltiple de la variable *LNREAPAR* sobre las variables *RESPUEST* y *ALCOHOL*.

```
* * * *  M U L T I P L E   R E G R E S S I O N  * * * *

Equation Number 1     Dependent Variable..    LNREAPAR    Logaritmo del
tiempo de reaparición

Block Number  1.  Method:  Enter       RESPUEST ALCOHOL

Variable(s) Entered on Step Number
   1..     ALCOHOL    Consumo de Alcohol
   2..     RESPUEST   Tiempo de respuesta al tratamiento

Multiple R              ,94199
R Square                ,88735
Adjusted R Square       ,88663
Standard Error          ,13579

Analysis of Variance
                 DF        Sum of Squares        Mean Square
Regression        2              44,88311           22,44156
Residual        309               5,69771            ,01844

F =    1217,05770        Signif F =  ,0000

---------------- Variables in the Equation ------------------

Variable            B         SE B        Beta          T  Sig T

RESPUEST       -,165606      ,004090    -,773044   -40,487  ,0000
ALCOHOL        -,011955  4,2728E-04    -,534209   -27,979  ,0000
(Constant)     3,225487      ,028334               113,838  ,0000
```



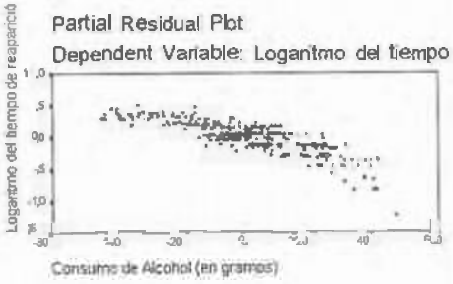


FIGURA 9.8. Regresión lineal múltiple de la variable *LNREAPAR* sobre las variables *RESPUEST* y *ALCOHOL*.

por el modelo construido a partir de la información aportada por la segunda. Análogamente, en el caso de la regresión lineal múltiple, podemos obtener una medida del grado de asociación lineal entre la variable *LNREAPAR* y el conjunto de las dos variables *RESPUEST* y *ALCOHOL*, así como la proporción de variabilidad de la primera explicada por el modelo construido a partir de la información aportada por dicho conjunto de variables.

El coeficiente de correlación múltiple y análisis de la varianza en el modelo múltiple

El coeficiente de correlación múltiple, $\rho_{Y.1...p}$, es una medida del grado de asociación lineal entre la variable dependiente, Y, y el conjunto de variables independientes, $X_1,..., X_p$. Es la máxima correlación entre Y y cualquier Y' que sea combinación lineal de $X_1, ..., X_p$. Su valor está comprendido entre 0 y 1. Si es próximo a 1, el ajuste del plano de regresión será casi perfecto, y si es próximo a 0, el plano de regresión no mejorará la predicción de Y sobre la predicción obtenida con la media muestral de Y. El estimador muestral de $\rho_{Y.1...p}$ es el coeficiente de correlación múltiple muestral, R. En el caso particular de una única variable independiente, el coeficiente de correlación múltiple coincide con el valor absoluto del coeficiente de correlación simple.

Como en el modelo de regresión simple, también en el modelo múltiple la variabilidad total puede descomponerse en la variabilidad explicada por la regresión y la variabilidad residual:

$$SC_{total} = SC_{reg} + SC_{res}$$

donde SC_{total}, SC_{reg} y SC_{res} se definen exactamente igual que en el modelo simple. En este caso, el estadístico F proporcionado por la tabla de análisis de la varianza (Tabla 9.1) permite contrastar la hipótesis nula de que la pendiente del plano de regresión es igual a cero, es decir:

$$H_0: \beta_1 = ... = \beta_p = 0$$

Además, se verifica que:

$$R^2 = \frac{SC_{reg}}{SC_{tot}}$$

donde R^2 es el cuadrado del coeficiente de correlación múltiple muestral, estimación del cuadrado del coeficiente de correlación múltiple, $\rho^2_{Y.1...p}$, al que se denomina coeficiente de determinación. Es decir, el coeficiente de determinación se define a partir de la correlación múltiple y, en el caso particular de una única variable independiente, coincidirá con el cuadrado del coeficiente de correlación simple. La

generalización de la interpretación del coeficiente de determinación al caso múltiple es automática: proporción de variabilidad total de la variable dependiente explicada mediante el plano de regresión. La hipótesis nula anterior es entonces equivalente a:

$$H_0: \rho^2_{Y.1...p} = 0$$

o, lo que es lo mismo, a la hipótesis de que Y está incorrelada con cualquier combinación lineal del conjunto de variables $X_1, .., X_p$.

Los resultados del análisis de la varianza se proporcionan en el bloque encabezado por «Analysis of Variance» (en la Figura 9.8). El p-valor asociado al estadístico F («Signif F = 0,000») es menor que 0,05, luego, al nivel de significación 0,05, se rechazará la hipótesis nula de que la pendiente del plano de regresión es igual a cero:

$$H_0: \beta_1 = \beta_2 = 0$$

siendo las estimaciones de β_1 y β_2:

$$B_1 = -0{,}165606 \text{ y } B_2 = -0{,}011955$$

Equivalentemente, se rechazará la hipótesis nula de que la variable *LNREAPAR* está incorrelada con cualquier combinación lineal de las variables *RESPUEST* y *ALCOHOL*. Concretamente, el grado de asociación lineal entre *LNREAPAR* y las variables *RESPUEST* y *ALCOHOL* es 0,94199 («Multiple R») y la proporción de variabilidad explicada mediante el plano de regresión es $R^2 = 0{,}88735$ («R Square»). Teniendo en cuenta que, cuando *RESPUEST* era la única variable independiente, el coeficiente de determinación era igual a 0,60198 (Figura 9.7), al considerar la información proporcionada por la variable *ALCOHOL*, la proporción de variabilidad explicada ha aumentado en 0,28 aproximadamente.

El coeficiente de determinación presenta el inconveniente de que, a mayor número de variables en el modelo, mayor es su valor, por lo que, en general, se considera que R^2 tiende a sobrestimar el verdadero valor de $\rho^2_{Y.1...p}$. El coeficiente de determinación ajustado por el número de observaciones y el número de variables independientes incluidas en la ecuación de regresión:

$$R^2_a = \frac{(n-1)R^2 - p}{n-1-p}$$

corrige la sobrestimación de R^2.

Comparando el coeficiente de determinación ajustado para el modelo con dos variables («Adjusted R Square = 0,88663») con el del modelo con una única varia-

ble («Adjusted R Square = 0,60070», en la Figura 9.7), podemos concluir que, al introducir la información de la variables *ALCOHOL*, la variabilidad explicada ha aumentado en un porcentaje superior al 28%.

En este caso, la mejora parece bastante considerable. Sin embargo, podría suceder que, a pesar de que el coeficiente de determinación ajustado fuera algo mayor, el incremento no fuera lo suficientemente grande como para considerar la información de la segunda variable. En este sentido, deberíamos analizar la información proporcionada por cada variable en particular.

Estadísticos para las variables independientes

Una alternativa para comparar la contribución de las distintas variables al modelo sería comparar los coeficientes correspondientes en la ecuación de regresión: a mayor coeficiente, mayor influencia de los valores de la variable correspondiente en la estimación del valor de la variable dependiente. Sin embargo, los coeficientes de la ecuación de regresión presentan el inconveniente de que su magnitud es relativa. Teniendo en cuenta que el coeficiente asociado a una variable independiente es igual al incremento (positivo o negativo) que se produciría en la variable dependiente al variar en una unidad el valor de la independiente, un mismo coeficiente para dos variables en distintas unidades de medida podría indicar una importancia relativa muy distinta. Por ejemplo, mientras que los valores de la variable *RESPUEST* están comprendidos entre 1 y 8 (semanas), los valores de la variable *ALCOHOL* estarán, en general, por encima de 10 (gramos diarios). Luego, si las dos variables presentaran el mismo coeficiente, los valores de la variable *ALCOHOL* tendrían mayor influencia en la estimación de los valores de la variable *LNREAPAR*.

En el sentido anterior, para eliminar el efecto de las distintas unidades de medida de las variables independientes, sería más adecuado considerar los coeficientes de regresión tipificados. Los coeficientes de regresión tipificados son los coeficientes de las variables cuando la ecuación de regresión se expresa como función de las variables tipificadas. Dado que al tipificar las variables se homogeniza la unidad de medida, el coeficiente de regresión tipificado se puede interpretar como una medida de la contribución relativa de la variable correspondiente al modelo. El plano de regresión construido a partir de la tipificación de las variables pasará por el origen y, por tanto, el coeficiente correspondiente al término independiente será igual a cero.

Siguiendo con nuestro ejemplo, dado que el coeficiente de regresión tipificado para la variable *RESPUEST* («Beta = –0,773044») es mayor, en valor absoluto, que el de la variable *ALCOHOL* («Beta= –0,534209»), la contribución de la primera variable al modelo será mayor.

El hecho de que una variable contribuya más que otra no significa necesariamente que su contribución sea grande. Podría suceder que la contribución de las dos fuera pequeña. Mediante el estadístico *F* asociado a la descomposición de la

varianza contrastamos la hipótesis nula de que la pendiente del plano de regresión era igual a cero:

$$H_0: \beta_1 = \ldots = \beta_p = 0$$

El rechazar esta hipótesis no implica que, aunque el conjunto de las variables independientes mejore la estimación de los valores de la variable dependiente respecto a la media, todas ellas contribuyan a la mejora. Para comprobarlo, contrastaremos la hipótesis anterior sobre cada parámetro en particular:

$$H_0: \beta_i = 0 \; \forall_i = 1, \ldots, p$$

En este caso, la hipótesis nula significa que la variable X_i no mejora la predicción de Y sobre la regresión obtenida con las $p-1$ variables restantes. El estadístico de contraste («T») es igual a:

$$\frac{B_i}{S_{B_i}}$$

donde S_{B_i} («SE B») es el error típico del coeficiente B_i. Si el p-valor asociado al estadístico de contraste T es menor que α, se rechazará la hipótesis nula al nivel de significación α.

La hipótesis nula anterior también podría plantearse para el término independiente del modelo:

$$H_0: \beta_0 = 0$$

En este caso, se interpretaría como que el plano de regresión pasa por el origen.

Tanto para cada una de las variables *RESPUEST* y *ALCOHOL* como para el término independiente, el p-valor asociado al estadístico T («Sig T = 0,0000», en las líneas *RESPUEST*, *ALCOHOL* y (Constant), respectivamente) es menor que 0,05. Luego, al nivel de significación 0,05, la hipótesis nula correspondiente puede ser rechazada en los tres casos. En consecuencia, la contribución de cualquiera de las dos variables es significativamente distinta de cero y, además, el plano de regresión no pasa por el origen.

Analicemos gráficamente el hecho de que introducir una segunda variable en el modelo aporte nueva información. En la parte inferior de la Figura 9.8 se muestra la representación de los gráficos de residuos parciales. En el ángulo inferior izquierdo se representan los residuos de la regresión de la variable *LNREAPAR* sobre la variable *RESPUEST* frente a los residuos de la regresión de la variable *ALCOHOL* sobre la variable *RESPUEST*. En el ángulo inferior derecho se realiza el mismo tipo de representación, pero intercambiando los papeles de las variables *RESPUEST* y *ALCOHOL*. Es decir, los gráficos de residuos parciales permiten analizar la asocia-

ción entre la variable dependiente y cada una de las independientes después de eliminar el efecto de las restantes independientes. Por ejemplo, en el caso del primer gráfico, el que la nube de puntos se concentre en una recta con pendiente distinta de cero significa que la información de la variable dependiente no aportada por la variable *RESPUEST* está muy relacionada con la información que no presentan en común las variables *ALCOHOL* y *RESPUEST*. En consecuencia, la variable *ALCOHOL* añade información respecto a la aportada por la variable *RESPUEST*. Análogamente, mediante el gráfico de la derecha, también llegaríamos a la conclusión de que la variable *RESPUEST* hubiera añadido información a la aportada por la variable *ALCOHOL*, en el supuesto caso de que ésta hubiera sido elegida en primer lugar. Para medir cuál de las dos variables añade más información a la proporcionada por la otra, bastaría con calcular la correlación entre los residuos del gráfico correspondiente, denominada correlación parcial entre la variable dependiente y la variable independiente en ordenadas, después de eliminar el efecto de la otra variable independiente.

El coeficiente de correlación parcial

Dada una variable dependiente, Y, y un conjunto de variables independientes, $X_1, \ldots, X_p$, todas ellas medidas en escala de intervalo o de razón, el coeficiente de correlación parcial, $r_{YXj \cdot C}$, entre la variable dependiente y una de las independientes, X_j, mide el grado de asociación lineal entre ellas dos, después de eliminar el efecto lineal del conjunto de las $p - 1$ restantes variables independientes, al que denominaremos C. Los valores del coeficiente de correlación parcial se interpretan igual que los del coeficiente de correlación simple. El estimador muestral para $\rho_{YXj \cdot C}$ es el coeficiente de correlación parcial muestral, $r_{YXj \cdot C}$.

Para determinar si el grado de asociación lineal entre las variables Y y X_j, después de eliminar el efecto de las restantes variables independientes, es estadísticamente significativo, se puede plantear la hipótesis nula de que el coeficiente de correlación parcial es igual a cero:

$$H_0: \rho_{YXj \cdot C} = 0$$

El estadístico de contraste se construye a partir del coeficiente de correlación parcial muestral, $r_{YXj \cdot C}$, del tamaño de la muestra, n, y del número de variables independientes en el conjunto C, $k = p$-1. Si el p-valor asociado es menor que α, se rechazará la hipótesis nula al nivel de significación α.

Recordemos que, al calcular la matriz de correlaciones entre las variables *REAPARIC*, *RESPUEST*, *ALCOHOL*, *CAFE* y *ANTIACID* (Figura 9.1), la variable más correlada con *REAPARIC* era *RESPUEST*, seguida de *ALCOHOL*, razón por la que el modelo de regresión múltiple se construyó con dichas variables independientes. Sin embargo, podría suceder que la variable *ALCOHOL*, a pesar de estar más correlada con la variable *REAPARIC* que las variables *CAFE* y *ANTIACID*, añadiera menos información a la previamente aportada por *RESPUEST* que cual-

quiera de estas dos. En otras palabras, podría suceder que la correlación parcial entre las variables *REAPARIC* y *ALCOHOL*, después de eliminar el efecto de la variable *RESPUEST*, fuera menor, en valor absoluto, que la correlación parcial considerando cualquiera de las otras dos variables. Teniendo en cuenta esta observación y que, por otro lado, el análisis se está realizando sobre el logaritmo neperiano de la variable *LNREAPAR*, comprobemos, calculando las correlaciones parciales entre *LNREAPAR* y *ALCOHOL*, *CAFE* y *ANTIACID*, eliminando el efecto de *RESPUEST*, que la variable independiente más adecuada para construir el modelo de regresión múltiple con dos variables es *ALCOHOL*.

La matriz de correlaciones parciales entre las variables *LNREAPAR*, *ALCOHOL*, *CAFE* y *ANTIACID*, después de eliminar el efecto de *RESPUEST*, se solicita en el Cuadro de diálogo 9.9. Los resultados se disponen en la Figura 9.9. La matriz de correlaciones parciales es una matriz simétrica respecto a la diagonal principal, por lo que basta con analizar los elementos situados por encima o por debajo de ella. Si centramos nuestra atención en la relación entre la variable dependiente *LNREAPAR* y cada una de las independientes, podemos observar que, dado que en todos los casos el tamaño muestral es el mismo y que, por tanto, los distintos valores son comparables, la máxima correlación parcial muestral, en valor absoluto, corresponde a *ALCOHOL*. Para determinar si el valor 0,8467 indica que la asociación lineal es estadísticamente significativa, contrastaremos la hipótesis nula de que no existe asociación:

$$H_0: \rho_{LNREPARALCOHOL\cdot RESPUEST} = 0$$

El p-valor asociado al estadístico de contraste ($P = 0{,}000$) es menor que 0,05, luego, al nivel de significación 0,05, rechazaremos la hipótesis nula.

La mayor asociación detectada con *LNREAPAR*, después de eliminar el efecto de *RESPUEST*, corresponde a *ALCOHOL*. Luego, efectivamente, la variable independiente más adecuada, junto con la variable *RESPUEST*, para construir el mode-

```
ESTADISTICA → CORRELACIONES → PARCIALES        En el Cuadro de diálogo

VARIABLES: LNREAPAR, ALCOHOL, CAFE, ANTIACID
CONTROLAR PARA: RESPUEST
ACEPTAR
```

CUADRO DE DIALOGO 9.9. Matriz de correlaciones parciales entre las variables *LNREAPAR*, *ALCOHOL*, *CAFE* y *ANTIACID*, eliminando el efecto de la variable *RESPUEST*.

```
- PARTIAL CORRELATION COEFFICIENTS -

Controlling for..  RESPUEST

            LNREAPAR        ALCOHOL         CAFE          ANTIACID

LNREAPAR     1,0000          -,8467         -,4483          ,5055
             (    0)         ( 309)         ( 309)         ( 309)
             P= .            P= ,000        P= ,000        P= ,000

ALCOHOL      -,8467          1,0000          ,5622         -,5855
             ( 309)          (    0)         ( 309)         ( 309)
             P= ,000         P= .            P= ,000        P= ,000

CAFE         -,4483           ,5622         1,0000         -,4204
             ( 309)          ( 309)         (    0)        ( 309)
             P= ,000         P= ,000        P= .           P= ,000

ANTIACID      ,5055          -,5855         -,4204         1,0000
             ( 309)          ( 309)         ( 309)         (    0)
             P= ,000         P= ,000        P= ,000        P= .

(Coefficient / (D.F.) / 2-tailed Significance)
```

FIGURA 9.9. Matriz de correlaciones parciales entre las variables *LNREAPAR*, *ALCOHOL*, *CAFE* y *ANTIACID*, eliminando el efecto de la variable *RESPUEST*.

lo de regresión múltiple con dos variables es la que habíamos considerado, *ALCOHOL*. Si se deseara mejorar el modelo introduciendo una tercera variable, se procedería de la misma forma: calculando la correlación parcial entre la variable *LNREAPAR* y cada una de las dos variables restantes, *CAFE* y *ANTIACID*, eliminando el efecto de las variables *RESPUEST* y *ALCOHOL*. Se seleccionaría aquella que presentara máxima correlación en valor absoluto. En otras palabras, el modelo de regresión múltiple se puede construir paso a paso, seleccionando en cada uno de ellos la variable que más información añada a la aportada por las variables previamente seleccionadas.

El Método Stepwise

El método Stepwise es un método de construcción de la ecuación de regresión lineal múltiple que selecciona las variables paso a paso. Frente a otros métodos, presenta la ventaja de admitir que una variable seleccionada en un paso puede ser

eliminada en otro posterior. Por ejemplo, supongamos que el tiempo de reaparición de los síntomas dependiera del consumo de alcohol, de café y de antiácidos, pero que cada uno de ellos en particular proporcionara menos información que el tiempo de respuesta al tratamiento (dado que éste podría depender a su vez de los tres consumos simultáneamente). Podría suceder que, aunque en el primer paso del proceso de selección de las variables, la seleccionada fuera *RESPUEST*, sí, en los tres pasos siguientes, las variables *ALCOHOL*, *CAFE* y *ANTIACID* también fueran seleccionadas, la información de *RESPUEST* fuera redundante y, en consecuencia, en el cuarto paso sería eliminada. Obsérvese en este ejemplo que, si no se establece un criterio de parada, la variable *RESPUEST*, al ser la única no incluida en la ecuación, volvería a ser seleccionada en el siguiente paso y, en consecuencia, eliminada en el siguiente, y así sucesivamente. Es decir, tanto para establecer si la información que aportará una nueva variable al ser seleccionada es significativa como para establecer si la de una variable previamente seleccionada es redundante, habrá que fijar algún criterio.

En el apartado «Estadísticos para las variables independientes» analizamos, mediante el *p*-valor asociado al estadístico *T*, si la información proporcionada por cada una de las variables podía ser redundante. En este sentido, un posible criterio de salida sería eliminar aquella variable tal que el *p*-valor asociado, o probabilidad de salida, fuera máximo, siempre y cuando superara un mínimo valor. Análogamente, si la variable X_j es la candidata a ser seleccionada en ese paso, un posible criterio de entrada se basa en el *p*-valor asociado al estadístico *T* para contrastar la hipótesis:

$$H_0: \beta_j = 0$$

siendo β_j el parámetro asociado a X_j en el supuesto caso de que fuera seleccionada. En dicho caso, la ecuación se construiría con X_j y con todas las previamente seleccionadas. Si el *p*-valor, o probabilidad de entrada, es menor que un determinado valor crítico la variable será seleccionada. Con la finalidad de que una variable no pueda entrar y salir de la ecuación en dos pasos consecutivos, el valor crítico de la probabilidad de salida debe ser mayor que el de la probabilidad de entrada (si no se indica lo contrario, el valor crítico de la probabilidad de entrada será igual a 0,05, mientras que el de la probabilidad de salida será igual a 0,1). En cualquier caso, para evitar que el proceso de selección se convierta en un proceso cíclico, es conveniente establecer un límite para el número de pasos (si no se indica lo contrario, el doble del número de variables independientes).

Otra ventaja del método Stepwise es que el proceso de selección puede comenzar a partir de la ecuación construida con un subconjunto de las variables independientes e, incluso, con todas ellas. En dicho caso, el proceso comenzaría eliminando variables. El hecho de introducir simultáneamente dos o más variables presenta el riesgo de que alguna de ellas puede ser una combinación lineal de las restantes. En dicho caso, las estimaciones de los parámetros del modelo no serían fiables. Para evitar esta situación, se utilizará el criterio de la tolerancia.

La tolerancia de una variable X_j con las variables $X_1, .., X_{j-1}, X_{j+1}, .., X_p$ se define como:

$$Tol_j = 1 - R_j^2$$

donde R_j^2 es el cuadrado del coeficiente de correlación múltiple entre X_j y $X_1, .., X_{j-1}, X_{j+1}, ..., X_p$. Si el valor de la tolerancia es próximo a 0, la variable X_j será casi una combinación lineal de las restantes variables, y si es próximo a 1, la variable X_j puede reducir la parte de variabilidad de Y no explicada por las restantes. Teniendo en cuenta esta definición, antes de construir la ecuación con todas las variables del subconjunto, se analizará la tolerancia de cada variable con las restantes. Si la tolerancia para una variable es muy pequeña, será excluida de la ecuación.

El criterio de la tolerancia puede ser utilizado, como un criterio adicional, a la probabilidad de entrada. En la construcción de la ecuación, para que la variable candidata a ser seleccionada en un paso pueda serlo, la tolerancia con las variables incluidas en la ecuación deberá superar un cierto valor mínimo. Por otro lado, al entrar la variable, la tolerancia de cualquier variable en la ecuación con las restantes también deberá superar ese mínimo valor.

Concretando, si el proceso comienza sin ninguna variable en la ecuación, el método Stepwise consiste en:

1. En el primer paso se introduce la variable más correlada con la dependiente, siempre que verifique el criterio de entrada. En caso contrario, el proceso finalizará sin que ninguna variable sea seleccionada y, en consecuencia, no tendrá sentido construir el modelo de regresión lineal a partir de la información de las variables independientes.
2. En el segundo paso se introduce la variable con mayor coeficiente de correlación parcial con la dependiente respecto de la independiente introducida en el primer paso, siempre que verifique el criterio de entrada. En caso contrario, el proceso finalizará y el modelo de regresión lineal será un modelo simple construido a partir de la información de la variable independiente introducida en el primer paso.
3. En el siguiente paso se introduce la variable con mayor correlación parcial con la dependiente respecto de las independientes que se encuentran en la ecuación, siempre que verifique el criterio de entrada. Si al introducir una variable, alguna de las previamente incluidas verifica el criterio de salida, antes de proceder a la selección de una nueva se eliminarán, paso a paso, las variables que verifiquen el criterio de salida.
4. Cuando ninguna variable en la ecuación verifique el criterio de salida se vuelve a la etapa 3. La etapa 3 se repite hasta que ninguna variable fuera de la ecuación satisfaga el criterio de entrada y ninguna de las variables en la ecuación satisfaga el de salida, o se alcance el máximo número de pasos.

Por otro lado, si el proceso comienza con alguna variable en la ecuación, antes de intentar introducir alguna más, se tratará de eliminar a las que están.

El mitodo Stepwise para la selección de las variables se solicita en el Cuadro de diálogo 9.10. Los resultados se disponen en las Figuras 9.10*a* y 9.10*b*. La primera variable seleccionada es *RESPUEST*, la más correlada con *LNREAPAR*. En consecuencia, la ecuación de regresión lineal simple y los resultados asociados coinciden con los de la Figura 9.7. Veamos qué sucede en el siguiente paso. De entre las restantes variables independientes («Variables not in tbe equation»), la variable candidata a entrar es la que presenta mayor correlación parcial en valor absoluto («Partial = –0,46748»), *ALCOHOL*. Además, la probabilidad de entrada, o *p*-valor asociado al estadístico *T* («Sig *T*=0,0000»), es menor que 0,05. En consecuencia, no sólo es la candidata sino que entrará en el segundo paso (Figura 9.10*b*).

Observemos que, como era de esperar, el valor del estadístico *T* correspondiente a la variable *ALCOHOL* es el mismo antes y después de que la variable haya sido introducida en la ecuación. En el segundo paso, las variables incluidas en la ecuación («Variables in the Equation») son *RESPUEST* y *ALCOHOL*. En consecuencia, la ecuación de regresión lineal múltiple y los resultados asociados coinciden con los de la Figura 9.8.

Una vez seleccionada una variable, el siguiente paso sería, en general, tratar de eliminar variables. Dado que nos encontramos en el segundo paso y que, por tanto, únicamente hay dos variables incluidas en la ecuación, no tiene sentido tratar de eliminar a una de las dos (la candidata a ser eliminada siempre será la segunda, y última, introducida). En cualquier caso, para ilustrar el método de selección Stepwise, procedamos como si hubiera más de dos variables. La variable candidata a ser eliminada sería aquella que presentara la máxima probabilidad de salida, o máximo *p*-valor asociado al estadístico *T* (o, lo que es equivalente, el mínimo valor absoluto

```
ESTADISTICA → REGRESION → LINEAL      En el Cuadro de diálogo

DEPENDIENTE: LNREAPAR
INDEPENDIENTE(S): RESPUEST, ALCOHOL, CAFE, ANTIACID
METODO: PASOS SUCESIVOS
ESTADISTICOS     En el Cuadro de diálogo
        COEFICIENTES DE REGRESION: ESTIMACIONES
        CONTINUAR
GUARDAR   En el Cuadro de diálogo
        VALORES PRONOSTICADOS: NO TIPIFICADOS
        RESIDUOS: NO TIPIFICADOS
        CONTINUAR
ACEPTAR
```

CUADRO DE DIALOGO 9.10. Regresión lineal múltiple de la variable *LNREAPAR* sobre las variables *RESPUEST*, *ALCOHOL*, *CAFE* y *ANTIACID*.

```
* * * *  M U L T I P L E   R E G R E S S I O N  * * * *

Equation Number 1 Dependent Variable.. LNREAPAR Logaritmo del tiempo de r

Block Number  1. Method: Stepwise    Criteria   PIN  ,0500   POUT  ,1000
   RESPUEST  ALCOHOL  CAFE  ANTIACID

Variable(s) Entered on Step Number
   1..     RESPUEST  Tiempo de respuesta al tratamiento

Multiple R            ,77588
R Square              ,60198
Adjusted R Square     ,60070
Standard Error        ,25484

Analysis of Variance
                    DF        Sum of Squares        Mean Square
Regression           1             30,44881           30,44881
Residual           310             20,13201            ,06494

F =      468,86194        Signif F =  ,0000

---------------- Variables in the Equation -----------------

Variable                B         SE B        Beta           T  Sig T

RESPUEST          -,166213     ,007676    -,775876    -21,653  ,0000
(Constant)       2,581571     ,031017                 83,232  ,0000

---------------- Variables not in the Equation -----------------

Variable      Beta In     Partial   Min Toler         T    Sig T

ALCOHOL      -,534209    -,846748     ,999972   -27,979    ,0000
CAFE         -,283192    -,448250     ,997194    -8,815    ,0000
ANTIACID      ,318944     ,505536     ,999945    10,300    ,0000
```

FIGURA 9.10*a*. Regresión lineal múltiple de la variable *LNREAPAR* sobre las variables *RESPUEST*, *ALCOHOL*, *CAFE* y *ANTIACID*.

de *T*); en este caso, *ALCOHOL*. Pero, dado que la probabilidad de salida es menor que 0,1, no será eliminada.

Comprobado que ninguna variable puede ser eliminada, analizaríamos si la variable candidata a ser seleccionada, la que presenta mayor correlación parcial en valor absoluto, puede serlo. La candidata sería *CAFE*, pero, al ser su probabilidad de entrada («Sig $T = 0{,}2669$») mayor que 0,05, no será seleccionada. En consecuencia, dado que ninguna variable más puede ser eliminada o seleccionada, el proceso finaliza con las variables *RESPUEST* y *ALCOHOL* incluidas en la ecuación.

```
    Variable(s) Entered on Step Number
    2..     ALCOHOL    Consumo de Alcohol

Multiple R              ,94199
R Square                ,88735
Adjusted R Square       ,88663
Standard Error          ,13579

Analysis of Variance
                    DF        Sum of Squares       Mean Square
Regression           2             44,88311          22,44156
Residual           309              5,69771           ,01844

F =    1217,05770        Signif F =  ,0000

------------------ Variables in the Equation ------------------

Variable                B         SE B        Beta           T  Sig T

RESPUEST         -,165606     ,004090    -,773044    -40,487  ,0000
ALCOHOL          -,011955  4,2728E-04    -,534209    -27,979  ,0000
(Constant)       3,225487     ,028334                113,838  ,0000

---------------- Variables not in the Equation ----------------

Variable     Beta In     Partial     Min Toler         T     Sig T

CAFE         ,025703     ,063243      ,681970     1,112     ,2669
ANTIACID     ,009342     ,022564      ,657107      ,396     ,6923

End Block Number   1   PIN =     ,050 Limits reached.

From Equation   1:   2 new variables have been created.

  Name          Contents
  ----          --------
  PRE_1         Predicted Value
  RES_1         Residual
```

FIGURA 9.10*b*. Regresión lineal múltiple de la variable *LNREAPAR* sobre las variables *RESPUEST*, *ALCOHOL*, *CAFE* y *ANTIACID*.

Parece entonces que el modelo construido no es mejorable con la información disponible. Sin embargo, recordemos la existencia de una quinta variable independiente, *TABACO*. Podría suceder que dicha variable incorporara más información, pero, como ya se mencionó, presenta el inconveniente de que sus valores corresponden a categorías. En cualquier caso, analicemos su posible relación con la variable *LNREAPAR* después de eliminar el efecto de las variables *RESPUEST* y *ALCOHOL*.

Obsérvese que, al final del bloque de resultados de la Figura 9.10b, se indica la generación de dos variables («From Equation 1: 2 new variables have been created»), denominadas *PRE_1* y *RES_1*, cuyos valores corresponden a la predicción de *LNREAPAR* mediante el modelo y a los residuos correspondientes respectivamente. En la representación gráfica de los residuos frente a las predicciones (en la parte izquierda de la Figura 9.11, solicitada en la parte superior del Cuadro de diálogo 9.11) se distinguen dos grupos según los dos valores de la variable *TABACO*. En general, los residuos correspondientes a casos tales que *TABACO* es igual a «Sí» son mayores que 0, mientras que los residuos correspondientes a casos tales que *TABACO* es igual a «No» son menores. Luego, al ajustar la ecuación de regresión, parece que las predicciones para los pacientes que han dejado de fumar están en general por debajo del verdadero valor, mientras que para los que no han dejado de fumar están por encima. Esta observación puede confirmarse en la representación gráfica de los valores de *LNREAPAR* frente a las predicciones (en la parte derecha de la Figura 9.11, solicitada en la parte inferior del Cuadro de diálogo 9.11). En dicha representación, para un mismo valor de la predicción, el valor observado es mayor en general en los casos tales que *TABACO* es igual a «Sí». Luego parece que, en el tiempo de reaparición de los síntomas, no sólo influyen el tiempo de respuesta al tratamiento y el consumo de alcohol, sino también el hecho de haber dejado o no de fumar. En lo que sigue veremos cómo introducir la información de la variable *TABACO* en el modelo de regresión.

```
GRAFICAS → DISPERSION   En el Cuadro de diálogo

SIMPLE → DEFINIR        En el Cuadro de diálogo

EJE Y: RES_1
EJE X: PRE_1
ETIQUETAR CASOS POR: TABACO
ACEPTAR
```

```
GRAFICAS → DISPERSION   En el Cuadro de diálogo

SIMPLE → DEFINIR        En el Cuadro de diálogo

EJE Y: LNREAPAR
EJE X: PRE_1
ETIQUETAR CASOS POR: TABACO
ACEPTAR
```

CUADRO DE DIALOGO 9.11. Gráficos de los valores de *RES_1*, en la parte superior, y de los de *LNREAPAR*, en la parte inferior, frente a los de *PRE_1* (cada punto aparece identificado por el valor del caso correspondiente en la variable *TABACO*).

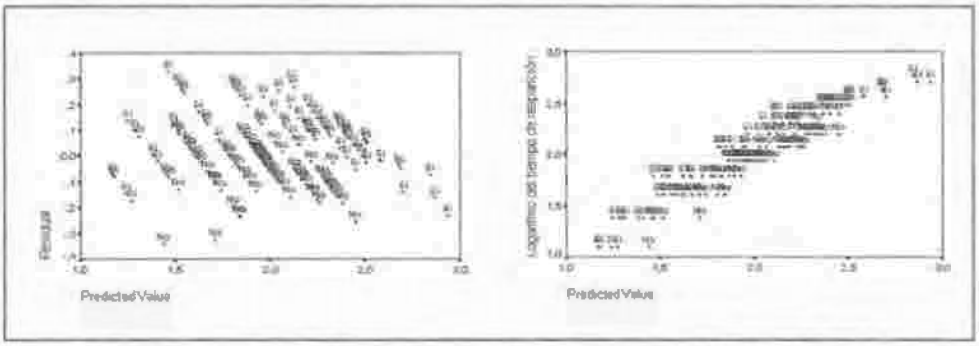


FIGURA 9.11. Gráficos de los valores de *RES_1*, en la parte izquierda, y de los de *LNREAPAR*, en la parte derecha, frente a los de *PRE_1* (cada punto aparece identificado por el valor del caso correspondiente en la variable *TABACO*).

VARIABLES CUALITATIVAS EN EL MODELO DE REGRESION LINEAL

Si entre las independientes se encuentra alguna variable cualitativa, sus valores deben ser recodificados, mediante la creación de nuevas variables, a valores numéricos que correspondan en algún sentido a las categorías originales. En el caso de variables con dos categorías, sus valores se pueden recodificar a valores 0 y 1. El valor 1 indicará la presencia de la cualidad correspondiente a una de la dos categorías, y el valor 0, la ausencia de dicha cualidad (en consecuencia, la presencia de la otra). Cuando una variable presente más de dos categorías deberán generarse tantas variables como el total de categorías menos uno. Cada nueva variable tomará valor 1 para una determinada categoría y 0 en el resto, de tal forma que los individuos en una misma categoría tomarán valor 1 en una misma variable y 0 en el resto. La categoría no considerada, o categoría referencia, estará representada por el valor 0 en todas las nuevas variables. Mediante este esquema de codificación, los coeficientes de las nuevas variables reflejarán el efecto de las categorías representadas respecto al de la categoría referencia.

Los posibles valores que puede tomar la variable *TABACO* son 1 y 2, en los pacientes que han dejado de fumar y en los que no, respectivamente. Bastará entonces con recodificar, mediante el Cuadro de diálogo 9.12, el valor 2 al 0. En definitiva, sólo se trata de una nueva codificación de los valores de *TABACO* que podría haber sido asignada desde el principio. A partir de esta recodificación y teniendo en cuenta la información proporcionada por los gráficos anteriores, es de esperar que, si *TABACO* es introducida en el modelo junto con *RESPUEST* y *ALCOHOL*, la

```
TRANSFORMAR → RECODIFICAR → EN LAS MISMAS VARIABLES
En el Cuadro de diálogo

VARIABLES: TABACO
VALORES ANTIGUOS Y NUEVOS   En el Cuadro de diálogo
        ANTIGUO → NUEVO: 2 → 0
        CONTINUAR
ACEPTAR
```

CUADRO DE DIALOGO 9.12. Recodificación de los valores de la variable *TABACO*.

ecuación sea tal que, para valores fijos de éstas, la estimación de *LNREAPAR* tome mayor valor cuando *TABACO* sea igual a 1. En otras palabras, es de esperar que el coeficiente asociado, supuesto que la variable sea seleccionada, sea positivo.

El método Stepwise para la selección de las variables, con la variable *TABACO* incluida entre las independientes, se solicita en el Cuadro de diálogo 9.13. Además, se indica que el proceso de selección parta de la situación final en la regresión anterior; es decir, con las variables *RESPUEST* y *ALCOHOL* incluidas en la ecuación. Los resultados se disponen en las Figuras 9.13*a* y 9.13*b*. Obsérvese que los resultados de la Figura 9.13*a* coinciden con los de la Figura 9.10*b*, con la única excepción de que, entre las variables que no están en la ecuación, se encuentra *TABACO*. En consecuencia, dado que sabemos que, respecto a las variables *RESPUEST*, *ALCOHOL*, *CAFE* y *ANTIACID*, el proceso había finalizado, lo único que podría suceder es que en el siguiente paso entrara *TABACO*.

```
ESTADISTICA → REGRESION → LINEAL      En el Cuadro de diálogo

DEPENDIENTE: LNREAPAR
BLOQUE 1 DE 1:  INDEPENDIENTE(S): RESPUEST, ALCOHOL
                        METODO: INTRODUCIR
BLOQUE 2 DE 2:  INDEPENDIENTE(S): CAFE, ANTIACID, TABACO
                        METODO: PASOS SUC.
ESTADISTICOS    En el Cuadro de diálogo
     COEFICIENTES DE REGRESION: ESTIMACIONES
     CONTINUAR
ACEPTAR
```

CUADRO DE DIALOGO 9.13. Regresión lineal múltiple de la variable *LNREAPAR* sobre las variables *RESPUEST*, *ALCOHOL*, *CAFE*, *ANTIACID* y *TABACO*.

```
* * * *  M U L T I P L E   R E G R E S S I O N  * * * *

Equation Number 1 Dependent Variable.. LNREAPAR Logaritmo del tiempo de r

Block Number  1.  Method:  Enter       RESPUEST ALCOHOL

Variable(s) Entered on Step Number
   1 .     ALCOHOL    Consumo de Alcohol
   2..     RESPUEST   Tiempo de respuesta al tratamiento

MulTipIe R            ,94199
R Square              ,88735
Adjusted R Square     ,88663
Standard Error        ,13579

Analysis of Variance
                  DF        Sum of Squares       Mean Square
Regression         2             44,88311          22,44156
Residual         309              5,69771           ,01844

F =    1217,05770        Signif F =  ,0000

------------------ Variables in the Equation ------------------

Variable                B         SE B        Beta           T  Sig T

RESPUEST          -,165606     ,004090    -,773044    -40,487  ,0000
ALCOHOL           -,011955  4,2728E-04    -,534209    -27,979  ,0000
(Constant)       3,225487     ,028334                113,838  ,0000

-------------- Variables not in the Equation ----------------

Variable     Beta In     Partial   Min Toler         T     Sig T

CAFE         ,025703     ,063243    ,681970     1,112     ,2669
ANTIACID     ,009342     ,022564    ,657107      ,396     ,6923
TABACO       ,201839     ,589769    ,961762    12,817     ,0000

End Block Number   1   All requested variables entered.
```

FIGURA 9.13*a*. Regresión lineal múltiple de la variable *LNREAPAR* sobre las variables *RESPUEST*, *ALCOHOL*, *CAFE*, *ANTIACID* y *TABACO*.

La probabilidad de entrada asociada a *TABACO* («Sig T = 0,0000») no sólo coincide con la mínima de entre las correspondientes a las tres variables no incluidas en la ecuación sino que, además, es menor que 0,05. Luego, efectivamente, la información que aporta *TABACO* puede ser considerada significativa y es introducida en la ecuación (Figura 9.13*b*).

```
* * * *  M U L T I P L E   R E G R E S S I O N  * * * *

Equation Number 1     Dependent Variable..   LNREAPAR   Logaritmo del
tiempo de reaparición

Block Number 2.  Method: Stepwise   Criteria   PIN  ,0500   POUT  ,1000

   CAFE      ANTIACID TABACO

Variable(s) Entered on Step Number
   3..    TABACO     Paciente ha dejado de fumar

Multiple R            ,96257
R Square              ,92654
Adjusted R Square     ,92582
Standard Error        ,10984

Analysis of Variance
                 DF        Sum of Squares       Mean Square
Regression        3             46,86493          15,62164
Residual        308              3,71589           ,01206

F =    1294,83547        Signif F =  ,0000

---------------- Variables in the Equation -----------------

Variable              B         SE B        Beta          T  Sig T

RESPUEST        -,157157     ,003374    -,733604   -46,584  ,0000
ALCOHOL         -,011993  3,4563E-04    -,535938   -34,700  ,0000
TABACO           ,162567     ,012684     ,201839    12,817  ,0000
(Constant)      3,117636     ,024415                127,694  ,0000

--------------- Variables not in the Equation --------------

Variable     Beta In     Partial   Min Toler         T     Sig T

CAFE         ,019709     ,060030    ,681542     1,054     ,2928
ANTIACID     ,023484     ,070118    ,654927     1,232     ,2190

End Block Number   2   PIN =     ,050 Limits reached.
```

FIGURA 9.13*b*. Regresión lineal múltiple de la variable *LNREAPAR* sobre las variables *RESPUEST*, *ALCOHOL*, *CAFE*, *ANTIACID* y *TABACO*.

Seleccionada *TABACO*, el siguiente paso será tratar de eliminar alguna variable. La candidata en el paso siguiente será aquella que presente máxima probabilidad de salida, o máximo p-valor asociado al estadístico T (o, lo que es equivalente, el mínimo valor absoluto de T); en este caso, *TABACO*. Pero, dado que la probabi-

lidad de salida es menor que 0,1, no será eliminada. A pesar de que en el ejemplo anterior ni *CAFE* ni *ANTIACID* fueron incluidas en la ecuación, es posible que ahora, dado que entre las incluidas también se encuentra *TABACO*, pudiera cambiar la situación. La candidata a ser seleccionada, la que presenta máxima correlación en valor absoluto, es *ANTIACID* (obsérvese que en el ejemplo anterior era *CAFE*, luego al menos en este aspecto sí ha cambiado la situación), pero su probabilidad de entrada («Sig $T = 0{,}2190$») es mayor que 0,05. En consecuencia, dado que ninguna variable más puede ser eliminada o seleccionada, el proceso finaliza con las variables *RESPUEST*, *ALCOHOL* y *TABACO* incluidas en la ecuación.

El grado de asociación lineal entre la variable dependiente y el conjunto formado por las variables *RESPUEST*, *ALCOHOL* y *TABACO* es próximo a 1 («Multiple $R = 0{,}96257$»). Por otro lado, la proporción de variabilidad de la variable dependiente explicada, mediante el modelo lineal, por el conjunto de las tres variables es muy alta («Adjusted R Square = 0,92582»). Comparándola con la explicada mediante el modelo con dos variables («Adjusted R Square = 0,88663», en la Figura 9.10*b*), la mejora obtenida al introducir la información de la variable *TABACO* es aproximadamente igual a 0,04. Veamos en qué se traduce esta mejora.

La ecuación de regresión obtenida («Variables in the Equation») es:

$$\begin{aligned} \hat{LNREAPAR} &= B_1\ RESPUEST + B_2 ALCOHOL + B_3\ TABACO + B_0 \\ &= -0{,}16\ RESPUEST - 0{,}01\ ALCOHOL + 0{,}16\ TABACO + 3{,}12 \end{aligned}$$

Luego la estimación del tiempo de reaparición de los síntomas vendrá dada por:

$$\begin{aligned} \hat{REAPARIC} &= e^{B_1\ RESPUEST + B_2 ALCOHOL + B_3\ TABACO + B_0} \\ &= e^{-0{,}16\ RESPUEST - 0{,}01\ ALCOHOL + 0{,}16\ TABACO + 3{,}12} \end{aligned}$$

En particular, si un paciente siguiera fumando durante el tratamiento:

$$\hat{REAPARIC}_{TABACO=0} = e^{-0{,}16 RESPUEST - 0{,}01\ ALCOHOL + 3{,}12}$$

mientras que, si dejara de fumar:

$$\begin{aligned} \hat{REAPARIC}_{TABACO=1} &= e^{-0{,}16 RESPUEST - 0{,}01\ ALCOHOL + 0{,}16 + 3{,}12} = \\ &= e^{0{,}16}\ \hat{REAPARIC}_{TABACO=0} \end{aligned}$$

Es decir, el tiempo esperado de reaparición de los síntomas se incrementaría en un factor de $e^{0{,}16} = 1{,}17$ si el paciente dejara de fumar. Por ejemplo, si el tiempo de respuesta para un determinado paciente fuera de 4 semanas, el consumo de alcohol de 40 gramos diarios y el paciente no dejara de fumar, su tiempo esperado de reaparición de los síntomas sería de 8 meses:

$$\hat{REAPARIC} = e^{-0{,}64 - 0{,}4 + 3{,}12} = e^{2{,}08} = 8$$

mientras que si dejara de fumar, sería de 9,32 meses:

$$\hat{REAPARIC} = e^{-0,16}\, e^{2,08} = e^{2,24} = 9,32$$

Si, además, redujera su consumo de alcohol a 10 gramos diarios, sería de 12,68 meses:

$$\hat{REAPARIC} = e^{-0,64 - 0,1 + 0,16 + 3,12} = e^{2,54} = 12,68$$