```
ESTADISTICA → RESUMIR → LISTAR CASOS   En el Cuadro de diálogo

VARIABLE(S): DIS_1, PROB, DIS1_1, DIS2_1, DIS3_1, DIS4_1
NUMERO DE CASOS
ACEPTAR
```

CUADRO DE DIALOGO 12.9. Listado de los valores de las variables *DIS_1*, *PROB*, *DIS1_1*, *DIS2_1*, *DIS3_1* y *DIS4_1*.

	DIS_1	PROB	DIS1_1	DIS2_1	DIS3_1	DIS4_1
117	2	1,00	,00174	,99638	,00188	,00000
146	2	1,00	,00080	,99659	,00221	,00000
178	1	,98	,97626	,02374	,00000	,00000
181	3	,80	,00000	,00007	,79783	,20210

Number of cases read: 4 Number of cases listed: 4

FIGURA 12.9. Listado de los valores de las variables *DIS_1*, *PROB*, *DIS1_1*, *DIS2_1*, *DIS3_1* y *DIS4_1*.

FERRAN ARANAZ

13
Regresión logística

Dada una variable dependiente dicotómica y un conjunto de una o más variables independientes cuantitativas o cualitativas, la regresión logística consiste en obtener una función lineal de las variables independientes que permita clasificar a los individuos en una de las dos subpoblaciones o grupos establecidos por los dos valores de la variable dependiente.

Supongamos que se sospecha que, en los pacientes con úlcera péptica que han seguido un tratamiento, el que la sintomatología ulcerosa reaparezca o no en un plazo de tiempo inferior o igual a ocho meses desde la respuesta al tratamiento depende del 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 dos semanas) se comprueba si la sintomatología ulcerosa persiste o ha desaparecido. Al cabo de los ocho meses de la desaparición para cada paciente, se comprueba si ha reaparecido o no. 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, café o antiácidos. Teniendo en cuenta que, en la mayoría de los casos, la sintomatología ha desaparecido al cabo de las ocho semanas desde la finalización del tratamiento, se descartarán todos aquellos pacientes en los que, pasadas ocho semanas, aún persiste. Para estimar la probabilidad de que la sintomatología reaparezca antes de los ocho meses, conocidos el tiempo de respuesta al tratamiento y los distintos hábitos del paciente, se aplicará la regresión logística.

FORMULACION DEL PROBLEMA

A partir de $(x_{i1}, ..., x_{ip})$, $i = 1, ..., n$, muestra de n observaciones de las variables independientes $X_1, ..., X_p$, en los dos grupos de individuos establecidos por los dos

2

valores de la variable dependiente *Y*, se trata de obtener una combinación lineal de las variables independientes que permita estimar las probabilidades de que un individuo pertenezca a cada una de las dos subpoblaciones o grupos. La probabilidad de que un individuo pertenezca a la segunda subpoblación, *p*, vendrá dada por:

$$p = \frac{e^Z}{1+e^Z} \quad \text{o, equivalentemente,} \quad p = \frac{1}{1+e^{-Z}}$$

siendo *Z* la combinación lineal:

$$Z = \beta_1 X_1 + \ldots + \beta_p X_p + \beta_0$$

donde $\beta_0, \beta_1, \ldots, \beta_p$ son parámetros desconocidos a estimar. En particular, la probabilidad de que el *i*-ésimo individuo de la muestra pertenezca a la segunda subpoblación será:

$$p_i = \frac{1}{1+e^{-(\beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \beta_0)}}$$

Si dicha probabilidad es superior o igual a 0,5, el paciente será clasificado en la segunda subpoblación; en caso contrario, será clasificado en la primera.

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

REAPARIC	Reaparición de la sintomatología ulcerosa en un plazo de tiempo inferior o igual a ocho meses desde la respuesta al tratamiento. Valores: No y Sí, codificados como 1 y 2, respectivamente.
RESPUEST	Tiempo de respuesta al tratamiento de la sintomatología ulcerosa (en semanas). Valores: 2, 4, 6 y 8 semanas, codificados numéricamente como 1, 2, 3 y 4, respectivamente.
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 combinación lineal de las variables independientes *RESPUEST*, *TABACO*, *ALCOHOL*, *CAFE* y *ANTIACID* que permita estimar las probabilidades de pertenecer a cada uno de los dos grupos establecidos por los valores de la variable dependiente *REAPARIC*. La probabilidad de que un paciente pertenezca al segundo grupo, *p*, vendrá dada por:

$$p = \frac{e^Z}{1+e^Z} \quad \text{o, equivalentemente,} \quad p = \frac{1}{1+e^{-Z}}$$

siendo *Z* la combinación lineal:

$$Z = \beta_1 RESPUEST + \beta_2 TABACO + \beta_3 ALCOHOL + \\ + \beta_4 CAFE + \beta_5 ANTIACID + \beta_0$$

donde $\beta_0, \beta_1, \ldots, \beta_p$ son parámetros desconocidos a estimar. En particular, la probabilidad de que para el *i*-ésimo paciente la sintomatología ulcerosa reaparezca en un plazo de tiempo inferior o igual a 8 meses desde la respuesta al tratamiento será:

$$p_i = \frac{1}{1+e^{-(\beta_1 RESPUEST_i + \ldots + \beta_5 ANTIACID_i + \beta_0)}}$$

Si dicha probabilidad es superior o igual a 0,5, el paciente será clasificado en el grupo correspondiente al segundo valor de *REAPARIC*. En caso contrario, será clasificado en el primero.

El objetivo que se persigue con el análisis es estimar, para cualquier paciente que haya sido sometido al tratamiento, la probabilidad de que la sintomatología ulcerosa reaparezca en un plazo de tiempo inferior o igual a los 8 meses desde la respuesta al tratamiento. Para ello, se tratará de estimar una función *Z* tal que no sólo permita estimar la probabilidad para los pacientes observados sino que, además, garantice de alguna manera que, cuando se trate de pacientes para los que se desconoce a cuál de los dos grupos pertenecen, también la clasificación, en términos de las probabilidades estimadas, sea correcta. Por esta razón, la estimación de la función *Z* se realizará considerando una muestra aleatoria de los pacientes observados y, posteriormente, se validará la capacidad de clasificación sobre el resto de los pacientes. Si el porcentaje de pacientes correctamente clasificados es elevado, es de esperar que la función *Z* también proporcione buenos resultados a la hora de predecir el valor de *REAPARIC* para cualquier paciente.

Mediante las opciones del Cuadro de diálogo 13.1 se seleccionará, aleatoriamente, al 90 por 100 de los pacientes de la muestra observada. Al solicitar la muestra aleatoria se generará una nueva variable, denominada *FILTER_$*, tal que sus valores serán iguales a 1 para los casos seleccionados, y a 0, para los no seleccionados. Aunque la estimación de la función *Z* se realice a partir de la información en la muestra seleccionada, para poder evaluar su utilidad en la clasificación de los pacientes no seleccionados, el procedimiento estadístico SPSS correspondiente debe ejecutarse sobre toda la muestra original. Por ello, mediante el Cuadro de diálogo 13.2, se solicita que se seleccionen todas las observaciones. En cualquier caso, mediante los valores de la variable *FILTER_$*, los pacientes de la muestra aleatoria

estarán perfectamente identificados. Por otro lado, para algunos de los pacientes se desconoce a cuál de los dos grupos pertenecen, por lo que no tiene sentido que se encuentren entre los seleccionados. Con la finalidad de evitarlo, mediante el Cuadro de diálogo 13.3, se solicita que en aquellos casos en los que el valor en *REAPARIC* no esté disponible, el valor en *FILTER_$* sea igual a 0.

```
DATOS → SELECCIONAR CASOS  En el Cuadro de diálogo

SELECCIONAR
      MUESTRA ALEATORIA DE CASOS
      MUESTRA  En el Cuadro de diálogo
         TAMAÑO DE LA MUESTRA
               APROXIMADAMENTE: 90% DE TODOS LOS CASOS
         CONTINUAR
ACEPTAR
```

CUADRO DE DIALOGO 13.1. Selección de una muestra aleatoria y generación de la variable *FILTER_$*.

```
DATOS → SELECCIONAR CASOS  En el Cuadro de diálogo

SELECCIONAR: TODOS LOS CASOS
ACEPTAR
```

CUADRO DE DIALOGO 13.2. Recuperación de todos los casos.

```
TRANSFORMAR → CALCULAR              En el Cuadro de diálogo

VARIABLE DE DESTINO: FILTER_$
EXPRESION NUMERICA: 0
SI        En el Cuadro de diálogo
          INCLUIR SI EL CASO SATISFACE LA CONDICION: SYSMIS(REAPARIC)
          CONTINUAR
ACEPTAR
```

CUADRO DE DIALOGO 13.3. Selección de los casos tales que el valor en la variable *REAPARIC* es desconocido.

Una vez preparados los datos, la regresión logística sobre las variables independientes *RESPUEST*, *ALCOHOL*, *CAFE*, *ANTIACID* y *TABACO*, en los dos grupos establecidos por los valores de la variable *REAPARIC*, se solicita en el Cuadro de diálogo 13.4. Los resultados se disponen en las Figuras 13.4*a* a 13.4*d*. Obsérvese que en el Cuadro de diálogo se indica que la función se estime considerando únicamente la información proporcionada por los pacientes seleccionados, aquellos tales que su valor en la variable *FILTER_$* es igual a 1. En concreto, como podrá comprobarse más adelante, el número total de pacientes observados es igual a 312. De ellos, 266 son los seleccionados aleatoriamente para el análisis y, de entre los 46 restantes, cuatro no disponen del valor en la variable *REAPARIC*.

VARIABLES CUALITATIVAS EN LA REGRESION LOGISTICA

Obsérvese que la variable *TABACO*, aunque sus valores hayan sido codificados como números, es cualitativa, por lo que, en principio, no debería ser introducida en la función *Z* como tal. Sin embargo, mediante una pequeña manipulación de sus valores, puede ser tratada como cualquier otra variable independiente numérica.

Si entre las independientes se encuentra alguna variable cualitativa, sus valores serán recodificados, mediante la creación de nuevas variables, a valores numéricos que correspondan en algún sentido a las categorías originales. En el caso de varia-

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

DEPENDIENTE: REAPARIC
COVARIABLES: RESPUEST, TABACO, ALCOHOL, CAFE, ANTIACID
METODO: ADELANTE: WALD
CATEGORICA  En el Cuadro de diálogo
      COVARIABLES CATEGORICAS: RESPUEST(INDICADOR), TABACO(INDICADOR)
      CONTINUAR
SELECCIONAR
      VARIABLE DE SELECCION: FILTER_$
      ESTABLECER VALOR  En el Cuadro de diálogo
            DEFINIR REGLA DE SELECCION: FILTER_$ IGUAL QUE 1
            CONTINUAR
ACEPTAR
```

CUADRO DE DIALOGO 13.4. Regresión logística de la variable *REAPARIC* sobre el conjunto de variables *RESPUEST*, *TABACO*, *ALCOHOL*, *CAFE* y *ANTIACID*.

bles con dos categorías, sus valores se recodificarán a valores 0 y 1. El valor 1 indicará la presencia de la cualidad correspondiente a una de la dos categorías, y el 0, la ausencia de dicha cualidad (en consecuencia, la presencia de la otra). Cuando una variable presente más de dos categorías, se generarán 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 efecto de la categoría referencia. Si se deseara comparar el efecto de una determinada categoría respecto al efecto promedio de las categorías, otro posible esquema de codificación sería, considerando dicha categoría como categoría referencia, mantener el esquema anterior y asignar a la categoría referencia, en cada una de las nuevas variables, el valor –1. De esta forma, el coeficiente para la categoría referencia sería el negativo de la suma de los coeficientes correspondientes a las restantes categorías.

A diferencia de la variable *TABACO*, los valores de *RESPUEST* son numéricos, por lo que podría ser considerada en el análisis como cualquier otra variable independiente numérica, sin necesidad de realizar ninguna transformación sobre ella. Sin embargo, dado que únicamente puede tomar cuatro valores diferentes, se ignorará su sentido numérico y su valores se utilizarán para distinguir cuatro tipos de pacientes, uno por cada posible tiempo de respuesta al tratamiento. En otras palabras, se considerará que *RESPUEST* es cualitativa y, por tanto, a partir de sus cuatro valores se generarán tres nuevas variables en el sentido previamente indicado.

Antes de comenzar el análisis (Figura 13.4*a*), los valores de las variables *RESPUEST* y *TABACO* son transformados, en el sentido anteriormente indicado, para poder ser considerados en la estimación de la función Z. Dado que 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 con recodificar, mediante la generación de la variable *TABACO*(1), 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. Sin embargo, para la variable *RESPUEST*, será necesario generar tres nuevas variables: *RESPUEST*(1), *RESPUEST*(2) y *RESPUEST*(3), una por cada uno de los tres primeros códigos numéricos de la variable *RESPUEST*, y tales que sus valores son iguales a 1 cuando el valor en la variable *RESPUEST* sea igual al código correspondiente, y a 0 en el resto de los casos. En consecuencia, las tres nuevas variables tomarán valor 0 cuando el valor en *RESPUEST* sea igual a 4 (categoría referencia).

SELECCION DE LAS VARIABLES

En el desarrollo del capítulo dedicado al análisis de regresión lineal múltiple, la situación planteada en el ejemplo para ilustrar la técnica era prácticamente la mis-

Dependent Variable Encoding:

	Value	Freq	Parameter Coding (1)	(2)	(3)
RESPUEST					
2 Semanas	1	99	1,000	,000	,000
4 Semanas	2	83	,000	1,000	,000
6 Semanas	3	60	,000	,000	1,000
8 Semanas	4	24	,000	,000	,000
TABACO					
Si	1	134	1,000		
No	2	132	,000		

Dependent Variable.. REAPARIC Tiempo de reaparición de los síntomas

Beginning Block Number 0. Initial Log Likelihood function

-2 Log Likelihood 357,71533

* Constant is included in the model.

Estimation terminated at iteration number 2 because
Log Likelihood decreased by less than ,01 percent.

Classification Table for REAPARIC
Selected cases FILTER_$ EQ 1

Observed		Predicted No N	Predicted Si S	Percent Correct
No	N	0	106	.00%
Si	S	0	160	100.00%
			Overall	60.15%

Classification Table for REAPARIC
Unselected cases FILTER_$ NE 1

Observed		Predicted No N	Predicted Si S	Percent Correct
No	N	0	13	.00%
Si	S	0	29	100.00%
			Overall	69.05%

Variables in the Equation

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
Constant	,4117	,1252	10,8087	1	,0010		

Beginning Block Number 1. Method: Forward Stepwise (LR)

Variables not in the Equation

Residual Chi Square 198,572 with 7 df Sig = ,0000

Variable	Score	df	Sig	R
RESPUEST	143,3554	3	,0000	,6197
RESPUEST(1)	133,2107	1	,0000	,6056
RESPUEST(2)	10,6532	1	,0011	,1555
RESPUEST(3)	47,1242	1	,0000	,3552
TABACO(1)	62,6546	1	,0000	,4118
ALCOHOL	33,6833	1	,0000	,2967
CAFE	5,8693	1	,0154	,1040
ANTIACID	8,2656	1	,0040	,1324

FIGURA 13.4*a*. Regresión logística de la variable *REAPARIC* sobre el conjunto de variables *RESPUEST, TABACO, ALCOHOL, CAFE* y *ANTIACID*.

ma que la actual. La diferencia entre los dos ejemplos estriba en que, en la regresión lineal, los valores de la variable dependiente *REAPARIC* no estaban agrupados, mientras que ahora sí. Es decir, la ecuación de regresión lineal nos permitía estimar directamente el propio valor de *REAPARIC* mientras que, en la regresión logística, a partir de la función Z, estimaremos las probabilidades de pertenecer a cada uno de los grupos y, en función de dichas probabilidades, el grupo al que pertenece cada paciente. Por lo demás, las dos técnicas son muy parecidas.

En el análisis de regresión lineal múltiple, las variables a partir de las que se construyó la ecuación podían ser seleccionadas mediante un procedimiento por pasos. El objetivo era construir la ecuación con aquel subconjunto de las variables independientes que mayor información aportase sobre los valores de la variable dependiente. Análogamente, en la regresión logística puede seleccionarse aquel subconjunto de las variables independientes que más información aporte sobre las probabilidades de pertenecer a cualquiera de los dos grupos establecidos por los valores de la variable dependiente. El método utilizado para construir la ecuación de regresión lineal múltiple, antes de intentar introducir una nueva variable en la ecuación, buscaba la posibilidad de que una variable previamente seleccionada pudiera ser eliminada. Análogamente, en la regresión logística, los dos métodos disponibles para la selección de variables paso a paso admiten también la posibilidad de eliminar variables previamente seleccionadas. Para realizar el ejemplo, el método que se utilizará para seleccionar el subconjunto de variables será el Forward y los estadísticos que utilizaremos en la selección y eliminación de variables serán la Puntuación eficiente de Rao y el estadístico de Wald, respectivamente.

En la regresión lineal múltiple, para comprobar si la información proporcionada por la variable independiente X_j era redundante, se utilizaba el p-valor asociado al estadístico T, que permitía contrastar la hipótesis nula de que el parámetro correspondiente en la ecuación de regresión era igual a cero. En este sentido, el criterio de salida consistía en eliminar aquella variable tal que el p-valor asociado, o probabilidad de salida, fuera máxima, siempre y cuando superara un mínimo valor crítico. Análogamente, si una variable era la candidata a ser seleccionada en un paso, el criterio de entrada se basaba en el p-valor asociado al estadístico T para contrastar la hipótesis nula de que el parámetro correspondiente, en el supuesto caso de que fuera seleccionada, era igual a cero. Si el p-valor, o probabilidad de entrada, era menor que un determinado valor crítico la variable era seleccionada.

Estadístico de Wald

El estadístico de Wald en la regresión logística juega el mismo papel que el estadístico T en la regresión lineal múltiple para las variables incluidas en la ecuación. Es decir, para cualquier variable independiente X_j seleccionada, si β_j es el parámetro asociado a X_j en la ecuación de regresión logística, el estadístico de Wald permite contrastar la hipótesis nula:

$$H_0 : \beta_j = 0$$

La interpretación de dicha hipótesis es que la información que se perdería al eliminar la variable X_j en el siguiente paso no es significativa. Si el p-valor asociado al estadístico de Wald es menor que α se rechazará la hipótesis nula al nivel de significación α. Bajo este punto de vista, en cada etapa del proceso de selección de variables, la candidata a ser eliminada será la que presente el máximo p-valor asociado al estadístico de Wald. Será eliminada si dicho máximo es mayor que un determinado valor crítico prefijado (si no se indica lo contrario, 0,1).

Puntuación eficiente de Rao

Si el estadístico de Wald en la regresión logística juega el mismo papel que el estadístico T en la regresión lineal múltiple para las variables incluidas en la ecuación, la Puntuación eficiente de Rao juega el de la T para las variables no incluidas. Supongamos que β_j es el parámetro asociado a la variable X_j, supuesto que entrara en la ecuación de regresión en el siguiente paso. El estadístico Puntuación eficiente de Rao permite contrastar la hipótesis nula:

$$H_0 : \beta_j = 0$$

La interpretación de dicha hipótesis es que, si la variable X_j fuera seleccionada en el siguiente paso, la información que aportaría no sería significativa. Si el p-valor asociado al estadístico Puntuación eficiente de Rao es menor que α se rechazará la hipótesis nula al nivel de significación α. Bajo este punto de vista, en cada etapa del proceso de selección de variables, la candidata a ser seleccionada será la que presente el mínimo p-valor asociado al estadístico Puntuación eficiente de Rao. Será seleccionada si dicho mínimo es menor que un determinado valor crítico prefijado (si no se indica lo contrario, 0,05).

Método Forward para la selección de variables

Si el proceso comienza con el modelo ajustado considerando únicamente el término independiente, entonces:

1. En el primer paso se introduce la variable que presente el mínimo p-valor asociado al estadístico Puntuación eficiente de Rao, siempre y cuando verifique el criterio de selección. En caso contrario, el proceso finalizará sin que ninguna variable sea seleccionada y, en consecuencia, no será posible construir la función Z a partir de la información de las variables independientes.
2. En el segundo paso se introduce la variable que presente el mínimo p-valor asociado al estadístico Puntuación eficiente de Rao, siempre que verifique el criterio de selección. En caso contrario, el proceso finalizará, y la función Z se construirá a partir de la información de la variable independiente introducida en el primer paso.

3. En el siguiente paso se introduce la variable que presente el mínimo *p*-valor asociado al estadístico Puntuación eficiente de Rao, siempre que verifique el criterio de selección. Si, al introducir una variable, el máximo *p*-valor asociado al estadístico de Wald para las previamente incluidas verifica el criterio de eliminación, antes de proceder a la selección de una nueva variable, se eliminará la variable correspondiente.
4. Cuando ninguna variable verifique el criterio de eliminación, se vuelve a la etapa 3. La etapa 3 se repite hasta que ninguna variable no seleccionada satisfaga el criterio de selección y ninguna de las seleccionadas satisfaga el de eliminación.

Si el proceso comienza con una o más variables seleccionadas, en el primer paso se analizará la posibilidad de seleccionar a las que no lo están.

Siguiendo con nuestro ejemplo (Figura 13.4*a*), la variable candidata a ser seleccionada en el primer paso, la que presenta el mínimo *p*-valor asociado al estadístico Puntuación eficiente de Rao o, lo que es equivalente, el máximo valor en dicho estadístico, es *RESPUEST* («Score = 143,3554»). El *p*-valor correspondiente («Sig = 0,0000») es menor que 0,05, por lo que no sólo es la candidata sino que, además, será seleccionada (Figura 13.4*b*, «Variable(s) Entered on Step Number: 1»). Recordemos que la variable *RESPUEST* era cualitativa y que, a partir de sus valores, se habían generado las variables *RESPUEST*(1), *RESPUEST*(2) y *RESPUEST*(3). Obsérvese que, al ser seleccionada *RESPUEST*, las tres variables correspondientes entran en bloque («Variables in the Equation»), es decir, son tratadas como un único grupo de información. Esta es una de las ventajas que presenta el método por pasos en la regresión logística frente a los métodos por pasos en la regresión lineal o en el análisis discriminante. En este sentido se considera que la regresión logística es una técnica que admite variables independientes cuantitativas y cualitativas.

Continuando con el proceso de selección de variables, veamos qué sucede en el siguiente paso. De entre las restantes variables independientes («Variables not in the Equation»), la candidata a ser seleccionada es la que presenta el máximo valor en la Puntuación eficiente de Rao («Score = 68,9473»), *TABACO*(1). El *p*-valor asociado («Sig = 0,0000») es menor que 0,05, por lo que será seleccionada en el segundo paso (Figura 13.4*c*, «Variable(s) Entered on Step Number: 2»). Una vez seleccionada una variable el siguiente paso sería, en general, tratar de eliminar variables pero, dado que nos encontramos en el segundo paso y que, por tanto, únicamente hay dos variables seleccionadas, no tiene sentido tratar de eliminar a una de las dos (la candidata a ser eliminada siempre será la segunda, y última, introducida en el análisis). En cualquier caso, para ilustrar el proceso de selección, procedamos como si hubiera más de dos variables.

De entre las variables incluidas en el análisis después del segundo paso («Variables in the Equation»), la candidata a ser eliminada sería aquella que presentara el máximo *p*-valor asociado al estadístico de Wald. En este caso, *RESPUEST*(3). Dicho *p*-valor («Sig = 0,8573») es mayor que 0,1, en consecuencia, debería ser eliminada. Sin embargo, recordemos que las variables *RESPUEST*(1), *RESPUEST*(2)

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

Estimation terminated at iteration number 8 because
Log Likelihood decreased by less than ,01 percent.

-2 Log Likelihood	188,285		
Goodness of Fit	242,002		
	Chi-Square	df	Significance
Model Chi-Square	169,431	3	,0000
Improvement	169,431	3	,0000

Classification Table for REAPARIC
Selected cases FILTER_$ EQ 1

Observed		Predicted No	Predicted Si	Percent Correct
No	N	84	22	79,25%
Si	S	15	145	90,63%
			Overall	86,09%

Classification Table for REAPARIC
Unselected cases FILTER_$ NE 1

Observed		Predicted No	Predicted Si	Percent Correct
No	N	7	6	53,85%
Si	S	6	23	79,31%
			Overall	71,43%

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

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
RESPUEST			72,6849	3	,0000	,4318	
RESPUEST(1)	-10,9255	20,3389	,2886	1	,5911	,0000	,0000
RESPUEST(2)	-8,1201	20,3386	,1594	1	,6897	,0000	,0003
RESPUEST(3)	-5,1252	20,3620	,0634	1	,8013	,0000	,0059
Constant	9,2027	20,3370	,2048	1	,6509		

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

Residual Chi Square 90,100 with 4 df Sig = ,0000

Variable	Score	df	Sig	R
TABACO(1)	68,9473	1	,0000	,4326
ALCOHOL	55,1624	1	,0000	,3855
CAFE	18,1615	1	,0000	,2126
ANTIACID	15,8872	1	,0001	,1970

FIGURA 13.4*b*. Regresión logística de la variable *REAPARIC* sobre el conjunto de variables *RESPUEST, TABACO, ALCOHOL, CAFE* y *ANTIACID*.

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

Estimation terminated at iteration number 9 because
Log Likelihood decreased by less than .01 percent.

 -2 Log Likelihood       111,658
 Goodness of Fit        3889,463

                     Chi-Square   df Significance

 Model Chi-Square       246,058    4        ,0000
 Improvement             76,627    1        ,0000

Classification Table for REAPARIC      Classification Table for REAPARIC
Selected cases FILTER_$ EQ 1           Unselected cases FILTER_$ NE 1

              Predicted                              Predicted
            No       Si    Percent                 No       Si    Percent
             N   |    S    Correct                  N   |    S  | Correct
Observed   |-----|-----|               Observed   |-----|-----|
  No    N  | 105 |   1 |   99.06%        No    N  |  12 |   1 |   92.31%
           |-----|-----|                          |-----|-----|
  Si    S  |  27 | 133 |   83.13%        Si    S  |   7 |  22 |   75.86%
           |-----|-----|                          |-----|-----|
                 Overall 89.47%                          Overall 80.95%

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

Variable             B      S.E.      Wald    df     Sig       R    Exp(B)

RESPUEST                             35,3572   3    ,0000   ,2865
 RESPUEST(1)   -13,6692   28,3138     ,2331    1    ,6293   ,0000    ,0000
 RESPUEST(2)    -8,9782   28,2973     ,1007    1    ,7510   ,0000    ,0001
 RESPUEST(3)    -5,0911   28,3134     ,0323    1    ,8573   ,0000    ,0062
TABACO(1)       -4,8389    1,0373   21,7608    1    ,0000  -,2350    ,0079
Constant        13,3391   28,3132     ,2220    1    ,6376

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

Residual Chi Square      44,449 with      3 df     Sig =  ,0000

Variable           Score    df      Sig       R

ALCOHOL          44,2614     1    ,0000   ,3437
CAFE             11,4903     1    ,0007   ,1629
ANTIACID         12,7949     1    ,0003   ,1737
```

FIGURA 13.4c. Regresión logística de la variable *REAPARIC* sobre el conjunto de variables *RESPUEST, TABACO, CAFE, ALCOHOL* y *ANTIACID.*

y *RESPUEST*(3) son tratadas en bloque, por lo que *RESPUEST*(3) no puede ser eliminada independientemente. Es decir, los únicos *p*-valores sujetos al criterio de eliminación son los correspondientes a *RESPUEST* y *TABACO*(1), y cualquiera de los dos son inferiores a 0,1. En consecuencia, ninguna de las dos puede ser eliminada.

Una vez comprobado que ninguna variable puede ser eliminada, el siguiente paso será comprobar si el *p*-valor asociado a la Puntuación eficiente de Rao correspondiente a la variable candidata a ser seleccionada es menor que 0,05. La candidata es *ALCOHOL* y, además, el *p*-valor correspondiente («Sig = 0,0000») es menor que 0,05, luego será seleccionada en el tercer paso («Variable(s) Entered on Step Number: 3», en la Figura 13.4*d*). De entre las variables incluidas en el análisis después del tercer paso, la candidata a ser eliminada es *TABACO*(1), pero, dado que el *p*-valor asociado al estadístico de Wald («Sig = 0,0024») es menor que 0,1, no será eliminada. Por otro lado, la candidata a ser seleccionada en el cuarto paso es *ANTIACID*, pero el *p*-valor asociado al estadístico Puntuación eficiente de Rao («Sig = 0,7672») es mayor que 0,05. En consecuencia, dado que ninguna variable más puede ser eliminada o seleccionada, la estimación de los parámetros de la función *Z* se realizará a partir de los valores de las variables *RESPUEST, ALCOHOL* y *TABACO*.

ESTIMACION DE LOS PARAMETROS

Recordemos que, a partir del modelo de regresión logística, la probabilidad de que un individuo pertenezca a la segunda subpoblación vendrá dada por:

$$p = \frac{1}{1 + e^{-Z}}$$

siendo *Z* la combinación lineal:

$$Z = \beta_1 X_1 + \ldots + \beta_p X_p + \beta_0$$

donde $\beta_0, \beta_1, \ldots, \beta_p$ son parámetros desconocidos a estimar.

El criterio para obtener los coeficientes $B_0, B_1, \ldots, B_p$, estimaciones de los parámetros desconocidos $\beta_0, \beta_1, \ldots, \beta_p$ es el de máxima verosimilitud. A partir de $B_0, B_1, \ldots, B_p$, las probabilidades estimadas de que un individuo pertenezca a las subpoblaciones segunda y primera son, respectivamente:

$$\hat{p} = \frac{1}{1+e^{-\hat{Z}}} \quad \text{y} \quad \hat{q} = 1 - \hat{p}$$

donde:

$$\hat{Z} = B_1 X_1 + \ldots + B_p X_p + B_0$$

Teniendo en cuenta que:

$$lg\left(\frac{\hat{p}}{\hat{q}}\right) = e^{\hat{Z}}$$

```
Variable(s) Entered on Step Number
3..       ALCOHOL   Consumo de Alcohol (gramos diarios)

Estimation terminated at iteration number 9 because
Log Likelihood decreased by less than ,01 percent.

 -2 Log Likelihood          53,228
 Goodness of Fit            59,150

                        Chi-Square    df Significance

 Model Chi-Square          304,488     5        ,0000
 Improvement                58,430     1        ,0000

Classification Table for REAPARIC      Classification Table for REAPARIC
Selected cases FILTER_$ EQ 1           Unselected cases FILTER_$ NE 1

             Predicted                              Predicted
            No       Si     Percent                No       Si     Percent
             N   |    S     Correct                 N   |    S   | Correct
Observed    |--------|--------|          Observed  |--------|--------|
  No    N   | 100    |     6  |  94.34%    No    N  |  11    |     2  |  84.62%
            |--------|--------|                    |--------|--------|
  Si    S   |   5    | 155    |  96.88%    Si    S  |   5    |    24  |  82.76%
            |--------|--------|                    |--------|--------|
                      Overall 95.86%                          Overall 83.33%

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

Variable             B        S.E.      Wald    df       Sig        R    Exp(B)

RESPUEST                             24,8201     3     ,0000    ,2294
 RESPUEST(1)  -18,2091     26,1670     ,4843     1     ,4865    ,0000     ,0000
 RESPUEST(2)   -8,4171     26,1483     ,1036     1     ,7475    ,0000     ,0002
 RESPUEST(3)   -4,8939     26,0831     ,0352     1     ,8512    ,0000     ,0075
TABACO(1)      -7,3969      2,4337    9,2382     1     ,0024   -,1422     ,0006
ALCOHOL          ,2495       ,0552   20,4120     1     ,0000    ,2269    1,2833
Constant        2,0766     26,0844     ,0063     1     ,9365

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

Residual Chi Square        ,163 with      2 df     Sig =  ,9218

Variable          Score    df       Sig        R

CAFE              ,0779     1     ,7801    ,0000
ANTIACID          ,0876     1     ,7672    ,0000

No more variables can be deleted or added.
```

FIGURA 13.4*d*. Regresión logística de la variable *REAPARIC* sobre el conjunto de variables *RESPUEST, TABACO, CAFE, ALCOHOL* y *ANTIACID*.

una expresión alternativa para el modelo de regresión logística es:

$$\frac{\hat{p}}{\hat{q}} = e^{B_0}(e^{B_1})^{X_1} \ldots (e^{B_p})^{X_p}$$

Luego para valores fijos de los restantes términos, cuanto mayor sea el coeficiente B_i mayor será el cociente entre las probabilidades y, en consecuencia, mayor será la probabilidad de pertenecer al segundo grupo.

La estimación de la función Z a partir de los valores de las variables seleccionadas, *RESPUEST*(1), *RESPUEST*(2), *RESPUEST*(3), *ALCOHOL* y *TABACO*(1) (Figura 13.4*d*, columna «B» en el bloque «Variables in the Equation») será:

$$\hat{Z} = -18{,}21\ RESPUEST(1) - 8{,}42\ RESPUEST(2) - 4{,}89\ RESPUEST(3) - 7{,}40\ TABACO(1) + 0{,}25\ ALCOHOL + 2{,}08$$

Por otro lado, la expresión alternativa para el modelo de regresión logística es:

$$\frac{\hat{p}}{\hat{q}} = e^{-18{,}21\ RESPUEST(1)}\, e^{-8{,}42\ RESPUEST(2)}\, e^{-4{,}89\ RESPUEST(3)}\, e^{-7{,}40\ TABACO(1)}\, e^{0{,}25\ ALCOHOL}\, e^{2{,}08}$$

o, lo que es equivalente (ver elementos en la columna «Exp(B)»):

$$\frac{\hat{p}}{\hat{q}} = (0{,}0000)^{RESPUEST(1)}(0{,}0002)^{RESPUEST(2)}(0{,}0075)^{RESPUEST(3)}(0{,}0006)^{TABACO(1)}(1{,}2833)^{ALCOHOL}e^{2{,}08}$$

En particular, para el paciente *i*-ésimo, el cociente entre las probabilidades estimadas de pertenecer a la segunda y a la primera subpoblación es:

$$\frac{\hat{p}_i}{\hat{q}_i} = (0{,}0000)^{RESPUEST(1)_i}(0{,}0002)^{RESPUEST(2)_i}(0{,}0075)^{RESPUEST(3)_i}(0{,}0006)^{TABACO(1)_i}(1{,}2833)^{ALCOHOL_i}e^{2{,}08}$$

Supongamos que el tiempo de respuesta al tratamiento del paciente *i*-ésimo ha sido de 8 semanas (*RESPUEST* = 4 o, lo que es equivalente, *RESPUEST*(1) = *RESPUEST*(2) = *RESPUEST*(3) = 0), y que el paciente no ha dejado de fumar (*TABACO*(1) = 0), entonces:

$$\frac{\hat{p}_i}{\hat{q}_i} = (1{,}2833)^{ALCOHOL_i}e^{2{,}08}$$

Si, además, su consumo de alcohol fuera de 20 gramos diarios, la probabilidad de que, en un plazo de tiempo inferior o igual a 8 meses desde la respuesta al tratamiento, los síntomas reaparezcan sería 1.175 veces superior a la probabilidad de que no reaparezca:

$$\frac{\hat{p}_i}{\hat{q}_i} = (1{,}2833)^{20} e^{2{,}08} \approx 1.175$$

mientras que, si redujera a 10 gramos el consumo diario de alcohol, sería 97 veces superior:

$$\frac{\hat{p}_i}{\hat{q}_i} = (1{,}2833)^{10} e^{2{,}08} \approx 97$$

Teniendo en cuenta que la suma de las probabilidades de pertenecer a cada uno de los dos grupos es igual a 1, en el primer caso:

$$\hat{p}_i \approx \frac{1.175}{1.176} = 0{,}999$$

mientras que en el segundo:

$$\hat{p}_i \approx \frac{97}{98} = 0{,}990$$

En otras palabras, si el tiempo de respuesta al tratamiento del paciente *i*-ésimo ha sido de 8 semanas (el máximo observado) y el paciente no ha dejado de fumar, la probabilidad de que la sintomatología ulcerosa reaparezca en un plazo de tiempo inferior o igual a 8 meses desde la respuesta al tratamiento es muy grande y, además, cuanto mayor sea el consumo diario de alcohol, mayor es dicha probabilidad.

BONDAD DEL AJUSTE

Hemos determinado cómo obtener las probabilidades estimadas de que el *i*-ésimo individuo pertenezca a cada una de las subpoblaciones. En concreto, hemos estimado la probabilidad de pertenecer a la segunda bajo dos situaciones diferentes. La diferencia entre la probabilidad observada de pertenecer a la segunda subpoblación y la estimada mediante el modelo de regresión logística es el residuo:

$$E_i = p_i - \hat{p}_i$$

donde p_i es 1 o 0, dependiendo de si el individuo pertenece o no, respectivamente, a la segunda subpoblación.

Comprobar la bondad del ajuste es analizar cuán probables son los resultados muestrales a partir del modelo ajustado. La probabilidad de los resultados observados se denomina verosimilitud, y se basa en comparar el número de individuos observado en la segunda subpoblación con el número esperado si el modelo fuera válido. El número esperado será igual al total de individuos en la muestra multiplicado por la probabilidad estimada para la segunda subpoblación. Para comprobar que la verosimilitud no difiere de 1 (que el modelo se ajusta perfectamente a los datos) se utiliza el estadístico:

$$-2LL = -2 \times \textit{Logaritmo de la verosimilitud}$$

Alternativamente, mediante el estadístico Bondad de ajuste, se podrían comparar las probabilidades observadas y las estimadas a partir del modelo:

$$Z^2 = \frac{\sum_{i=1}^{n} E_i^2}{\hat{p}_i (1 - \hat{p}_i)}$$

Ambos estadísticos, bajo la hipótesis nula de que el modelo se ajusta a los datos observados, siguen una distribución *ji*-cuadrado con *n*-2 grados de libertad.

Los valores de ambos estadísticos para el modelo con las variables independientes *RESPUEST*, *ALCOHOL* y *TABACO* («-2 Log Likelihood = 53,228» y «Goodness of Fit = 59,150», en la parte superior de la Figura 13.4*d*) son tales que, salvo que recurriéramos a las tablas de la *ji*-cuadrado con $n - 2$ ($n = 266$) grados de libertad, no proporcionan demasiada información. Sin embargo, sabemos que cuanto menor sea el valor del estadístico *ji*-cuadrado, mayor será el *p*-valor asociado y, en consecuencia, menos motivos tendremos para rechazar la hipótesis nula de que el modelo es adecuado. Obsérvese cómo, a lo largo de los tres pasos del procedimiento de selección de variables (Figuras 13.4*a* a 13.4*d*), los valores del primer estadístico («-2 Log Likelihood») han ido disminuyendo paulatinamente. Sin embargo, para el segundo («Goodness of Fit»), aunque al final del proceso su valor es inferior al inicial, en el segundo paso su valor aumenta respecto al anterior. Para clarificar estos resultados, analicemos desde otro punto de vista si el modelo es adecuado y cuál ha sido la mejora obtenida en cada paso del proceso de selección de variables.

Para todas las variables introducidas en la función *Z* tenemos garantías de que, por el criterio de eliminación en el proceso de selección, el *p*-valor asociado al estadístico de Wald es menor que 0,1 o, lo que es equivalente, la hipótesis de que el parámetro correspondiente es nulo puede ser rechazada al nivel de significación 0,1. En este sentido, para comprobar que el modelo es adecuado, una alternativa sería contrastar, en una única hipótesis nula, que todos los parámetros correspondientes al conjunto de variables incluidas en el modelo son iguales a cero. Recordemos que, en el análisis de regresión lineal múltiple, dicho contraste era

10

equivalente a contrastar que el coeficiente de determinación era igual a cero. El estadístico de contraste era el estadístico *F* de la tabla de análisis de la varianza. Además, mediante el coeficiente de determinación ajustado podíamos evaluar la mejora obtenida en cada paso del proceso de selección de variables. En la regresión logística, para contrastar la hipótesis nula de que, en cada etapa, para todas las variables incluidas en el modelo los parámetros asociados son nulos utilizaremos el estadístico *Ji*-cuadrado del modelo. Para evaluar la mejora obtenida en cada paso o, lo que es equivalente, el cambio producido en el estadístico *Ji*-cuadrado respecto al paso anterior, utilizaremos el estadístico de Mejora.

En el primer paso del proceso de selección el estadístico *Ji*-cuadrado para el modelo con la variable *RESPUEST* como única independiente es igual a 169,431 («Model Chi-Square» en el primer bloque de resultados de la Figura 13.4*b*). Dado que se trata del primer paso, coincide con el valor del estadístico de mejora («Improvement»). Concentrémonos, por tanto, en el primer estadístico. El *p*-valor asociado al estadístico *Ji*-cuadrado para el modelo («Significance») es menor que 0,05, luego, al nivel de significación 0,05, se rechaza la hipótesis nula de que los parámetros asociados a las tres variables generadas a partir de los valores de *RESPUEST* son nulos (aunque dicha hipótesis, enunciada en particular para cada variable no era rechazada). En el segundo paso (Figura 13.4*c*), al introducir la variable *TABACO* en el modelo, el valor del estadístico *Ji*-cuadrado para el modelo con las variables *RESPUEST* y *TABACO* aumenta hasta 246,058; la Mejora que se produce es de 76,627. El *p*-valor asociado al estadístico de Mejora («Significance») es menor que 0,05, por lo que se rechaza la hipótesis nula de que la mejora no es significativa. Finalmente, en el tercer paso (Figura 13.4*d*), al introducir la variable *ALCOHOL* en el modelo, el valor del estadístico *Ji*-cuadrado para el modelo con las variables *RESPUEST*, *TABACO* y *ALCOHOL* aumenta hasta 304,488 (la Mejora es de 58,43). También en este caso el *p*-valor asociado al estadístico de Mejora es menor que 0,05, por lo que puede concluirse que la mejora es significativa. Obsérvese que, para ambos estadísticos, los grados de libertad («df») varían de un paso a otro. La magnitud de su valor depende del número de variables incluidas en el modelo en cada paso. En consecuencia, una misma mejora en dos pasos distintos no será igual de significativa.

Veamos cómo se traduce la mejora obtenida en cada paso en términos de la clasificación de los individuos en cada uno de los dos grupos. Es decir, veamos en qué medida aumenta el porcentaje de casos correctamente clasificados al ir introduciendo la información de las distintas variables.

CLASIFICACION DE LOS INDIVIDUOS

Validación de los resultados

La clasificación de los individuos en uno u otro grupo se realizará a partir de la probabilidad estimada de pertenecer al segundo grupo. Si, para un determinado

individuo, la probabilidad estimada de pertenecer a la segunda subpoblación es mayor o igual que 0,5, será clasificado en dicha subpoblación. En caso contrario, será clasificado en la primera. El porcentaje de casos correctamente clasificados será un índice de la efectividad del modelo. Si el modelo es efectivo sobre la muestra observada, es de esperar que también lo sea cuando se trate de clasificar a un individuo para el que se desconoce a cuál de los dos grupos pertenece.

La Figura 13.4*d* muestra el resumen de los resultados de la clasificación, tanto para los casos que forman parte de la muestra seleccionada aleatoriamente como para los que no. Respecto al primer conjunto («Classification Table for *REAPARIC*: Selected cases *FILTER_$ EQ* 1»), el porcentaje de casos correctamente clasificado en el primer grupo («Percent Correct: Observed = No») es igual a 94,34 por 100, mientras que en el segundo («Percent Correct: Observed = Sí») es igual a 96,88 por 100. En el segundo conjunto de casos («Classification Table for *REAPARIC*: Selected cases *FILTER_$ NE* 1»), dichos porcentajes son iguales a 84,62 por 100 y 82,76 por 100. En términos generales, de un total de 266 pacientes en el primer conjunto, 255 han sido correctamente clasificados o, lo que es equivalente, el 95,86 por 100. De un total de 42 pacientes en el segundo conjunto, 35 han sido correctamente clasificados o, lo que es equivalente, el 83,33 por 100.

Analicemos cuál ha sido la mejora obtenida a lo largo de las tres etapas del proceso de selección de variables sobre la muestra objeto de análisis. Al inicio del proceso, antes de considerar la información de cualquiera de las variables independientes o, lo que es equivalente, considerando únicamente la relativa al número de casos en cada grupo, el porcentaje de casos correctamente clasificados (Figura 13.4*a*) era del 60,15 por 100. Al introducir la información de la variable *RESPUEST* (Figura 13.4*b*), el porcentaje aumentó al 86,09 por 100; al considerar simultáneamente la información de *RESPUEST* y *TABACO* (Figura 13.4*c*), aumentó al 89,47 por 100 y, finalmente, al incorporar la de *ALCOHOL* a las dos anteriores (Figura 13.4*d*), el porcentaje de casos correctamente clasificado es igual a 95,86 por 100, lo que supone un incremento total, desde el principio hasta el final del proceso, del 35,71 por 100. Sin embargo, sobre la otra muestra, la considerada para validar el análisis, el incremento en cada etapa ha sido menor, siendo el incremento total del 14,28 por 100. Obsérvese que, en esta segunda muestra, el porcentaje de pacientes en el segundo grupo (Figura 13.4*a*) es del 69,05 por 100, en consecuencia, desde el inicio del proceso el porcentaje de casos correctamente clasificado era elevado, razón por la que la mejora no resulta tan significativa como en la primera muestra.

En conclusión, sin considerar la información proporcionada por las variables *RESPUEST*, *TABACO* y *ALCOHOL*, el valor estimado de *REAPARIC* sería, en todos los casos, igual a 2 y, supuesto que la muestra de pacientes considerada en el análisis es representativa de la población, el porcentaje de casos que se esperaría clasificar correctamente sería del 60,15 por 100. Mediante el modelo de regresión logística, con la información aportada por las variables *RESPUEST*, *TABACO* y *ALCOHOL*, el porcentaje de casos correctamente clasificado ha sido del 95,86 por 100, lo que supone una mejora del 35,71 por 100. Podemos concluir entonces que la información aportada por las variables *RESPUEST*, *ALCOHOL* y *TABACO* es

significativa. Además, sobre la muestra de pacientes no considerada en el análisis, el porcentaje de casos correctamente clasificado es del 83,33 por 100. En consecuencia, cuando se trate de predecir, para aquellos pacientes que acaban de responder al tratamiento, si la sintomatología ulcerosa reaparecerá o no en un plazo de tiempo inferior o igual a los 8 meses, también el porcentaje de predicciones correctas será elevado.

Predicción

Una vez comprobado que, mediante la función estimada a partir de los valores de las variables *RESPUEST*, *TABACO* y *ALCOHOL*, el porcentaje de casos correctamente clasificados es elevado, es de esperar que la función también proporcione buenos resultados a la hora de predecir el valor de *REAPARIC* para cualquier paciente.

La muestra contiene cuatro pacientes para los que se desconoce el valor de *REAPARIC*. Para predecir el grupo al que pertenecen cada uno de ellos, y con la finalidad de considerar toda la información disponible, estimaremos de nuevo los coeficientes de la función *Z* sobre toda la muestra, la formada por los 308 pacientes (Cuadro de diálogo 13.5). En este caso, la estimación de la función *Z* (Figura 13.5) es:

$$\hat{Z} = -17{,}79\ RESPUEST(1) - 9{,}45\ RESPUEST(2) - 6{,}36\ RESPUEST(3) - 6{,}85\ TABACO(1) + 0{,}23\ ALCOHOL + 3{,}31$$

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

DEPENDIENTE: REAPARIC
COVARIABLES: RESPUEST, TABACO, ALCOHOL
METODO: INTRODUCIR
CATEGORICA       En el Cuadro de diálogo
      COVARIABLES CATEGORICAS: RESPUEST(INDICADOR), TABACO(INDICADOR)
      CONTINUAR
GUARDAR  En el Cuadro de diálogo
      VALORES PRONOSTICADOS: PROBABILIDADES, GRUPO DE PERTENENCIA
      CONTINUAR
ACEPTAR
```

CUADRO DE DIALOGO 13.5. Regresión logística de la variable *REAPARIC* sobre las variables *RESPUEST*, *TABACO* y *ALCOHOL*.


```
Dependent Variable..   REAPARIC   Tiempo de reaparición de los síntomas

Beginning Block Number  0.  Initial Log Likelihood Function

-2 Log Likelihood    410,9297

* Constant is included in the model.

Beginning Block Number  1.  Method: Enter

Variable(s) Entered on Step Number
1..        RESPUEST   Tiempo de respuesta al tratamiento
           TABACO     Paciente ha dejado de fumar
           ALCOHOL    Consumo de Alcohol (gramos diarios)

Estimation terminated at iteration number 9 because
Log Likelihood decreased by less than ,01 percent.

 -2 Log Likelihood          74,407
 Goodness of Fit            73,458

                    Chi-Square      df Significance

 Model Chi-Square      336,523        5          ,0000
 Improvement           336,523        5          ,0000

Classification Table for REAPARIC
                  Predicted
                    No        Si        Percent Correct
                     N         S
Observe d      |----------|----------|
   No      N   |   110    |     9    |   92.44%
               |----------|----------|
   Si      S   |     9    |   180    |   95.24%
               |----------|----------|
                          Overall 94,16%

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

Variable              B        S.E.      Wald     df       Sig        R    Exp(B)

RESPUEST                                32,0301     3     ,0000    ,2517
 RESPUEST(1)    -17,7938    24,9767      ,5075     1     ,4762    ,0000    ,0000
 RESPUEST(2)     -9,4484    24,9570      ,1433     1     ,7050    ,0000    ,0001
 RESPUEST(3)     -6,3596    24,9156      ,0651     1     ,7985    ,0000    ,0017
TABACO(1)        -6,8528     1,8617   13,5497     1     ,0002   -,1676    ,0011
ALCOHOL           ,2338      ,0452    26,7943     1     ,0000    ,2456   1,2633
Constant         3,3186    24,9158      ,0177     1     ,8940

2 new variables have been created.
  Name          Contents

  PRE_1         Predicted Value
  PGR_1         Predicted Group
```

FIGURA 13.5. Regresión logística de la variable *REAPARIC* sobre las variables *RESPUEST*, *TABACO* y *ALCOHOL*.

Obsérvese que, respecto a la función estimada a partir de la muestra aleatoria con 266 pacientes (Figura 13.4*d*), el cambio experimentado en los coeficientes es poco significativo. Luego parece que los coeficientes de la función son independientes de cuál sea la muestra considerada en su estimación. Es de esperar entonces que los resultados de la clasificación sobre la población objeto de análisis sean parecidos a los obtenidos sobre la muestra. Considerando la muestra de 308 pacientes, el porcentaje de casos correctamente clasificados (Figura 13.5) es del 94,16 por 100. Por tanto, a la hora de predecir el valor de *REAPARIC*, el porcentaje de aciertos será elevado.

En la parte inferior de la Figura 13.5 se indica la generación de las variables *PRE_*1 y *PGR_*1 («2 new variables have been created»), cuyos valores son: la probabilidad de pertenecer al segundo grupo y el grupo en el que ha sido clasificado el caso, respectivamente. Comprobemos, a partir de la información proporcionada por estas variables para los cuatro pacientes previamente mencionados, cuál es el grupo en el que han sido clasificados y con qué probabilidad.

La Figura 13.7 contiene un listado de los cuatro pacientes (casos 117, 146, 178 y 181) para los que se desea estimar, conocidos el tiempo de respuesta al tratamiento, el consumo de alcohol y el abandono o no del hábito de fumar, la probabilidad de que la sintomatología ulcerosa reaparezca en un plazo de tiempo inferior o igual a los 8 meses desde la respuesta al tratamiento. El listado ha sido solicitado mediante el Cuadro de diálogo 13.7, pero, previamente (Cuadro de diálogo 13.6), ha sido necesario indicar que se limite a los cuatro casos con valor desconocido en la variable *REAPARIC*. La primera columna, correspondiente a la variable *PRE_*1, contiene la probabilidad estimada de pertenecer al segundo grupo, y la segunda, correspondiente a la variable *PGR_*1, contiene la identificación del grupo en el que ha sido clasificado cada uno de ellos. Analizando los resultados, podría concluirse que, para los pacientes 117, 146 y 178, clasificados en el segundo grupo con probabilidades próximas a 1, es muy probable que la sintomatología reaparezca en un plazo de tiempo inferior o igual a los 8 meses, mientras que para el paciente 181, clasificado en el primer grupo y con probabilidad próxima a cero de pertenecer al segundo, parece poco probable.

```
DATOS → SELECCIONAR CASOS   En el Cuadro de diálogo

SI SE SATISFACE LA CONDICION
SI   En el Cuadro de diálogo
       SYSMIS(REAPARIC)
       CONTINUAR
ACEPTAR
```

CUADRO DE DIALOGO 13.6. Selección de los casos tales que el valor de la variable *REAPARIC* es desconocido.


```
ESTADISTICA → RESUMIR → LISTAR CASOS   En el Cuadro de diálogo

VARIABLE(S): PRE_1, PGR_1
NUMERO DE CASOS
ACEPTAR
```

CUADRO DE DIALOGO 13.7. Listado de los valores de las variables *PRE_*1 y *PGR_*1.

```
          PRE_1 PGR_1

  117     .99993     2
  146     .99818     2
  178     .99981     2
  181     .00160     1
```

FIGURA 13.7. Listado de los valores de las variables *PRE_*1 y *PGR_*1.