10 Regresión logística

Gonzalo Sánchez Vizcaíno

1. INTRODUCCIÓN

En las disciplinas académicas no son raras las ocasiones en las que el investigador está interesado en predecir si un determinado suceso ocurrirá o no en función de una serie de variables explicativas. En el campo de las ciencias sociales, es frecuente que se presenten tales situaciones por cuanto son muchas las variables objeto de estudio que presentan serias dificultades —cuando no una clara imposibilidad— para ser representadas de forma cuantitativa. Valgan como muestra algunos ejemplos: el director de marketing de una empresa de comunicaciones se interesa por conocer hasta qué punto ciertas características socioeconómicas (estado civil, ingresos, nivel de estudios, edad o número de hijos) influyen en que un individuo contrate un nuevo servicio de televisión por cable; por su parte, un analista financiero quisiera estudiar la relación entre el comportamiento de una serie de ratios financieros y económicos con la posibilidad de que una empresa entre en situación de quiebra; por último, un responsable de política económica desearía conocer si en el hecho de que una empresa industrial sea innovadora o no influyen una serie de características como su tamaño, sector de actividad, complejidad organizacional o formalización de su estructura.

Los problemas representados en estos tres ejemplos siguen un esquema común. Se trata de construir un modelo que describa la relación entre una serie de características que conforman un conjunto de variables independientes de tipo categórico o continuo (estado civil, ingresos, edad, ratios financieros, tamaño, sector de actividad) y una variable dependiente dicotómica o binaria que sólo puede tomar dos valores que definen opciones o características opuestas o mutuamente excluyentes (contratar el servicio de televisión por cable o no; situación de quiebra o no; empresa innovadora o no innovadora). El análisis discriminante serviría para abordar situaciones como las descritas; sin embargo, la posibilidad de que coexistan variables independientes de naturaleza cuantitativa y categórica viola la asunción de normalidad multivariante.

En el presente capítulo se introduce una técnica de análisis multivariante, *la regresión logística,* que no sólo solventa las dificultades planteadas por el análisis discriminante[1], sino que también suple las limitaciones del modelo de regresión lineal respecto a la naturaleza dicotómica de la variable dependiente. Así pues, el modelo de regresión logística es un procedimiento por medio del cual se intenta analizar las relaciones de asociación entre una variable dependiente dicotómica (binaria o *dummy*) Y y una o varias variables independientes (regresores o predictores) X_n cuantitativas o categóricas, todo ello a fin de lograr los siguientes objetivos: determinar la existencia o ausencia de relación entre una o más variables independientes y la variable dependiente; medir la magnitud de dicha relación y estimar o predecir la probabilidad de que se produzca (o no) el suceso definido por la variable dependiente en función de los valores que adopten las variables independientes.

Los orígenes de esta técnica, y en general de los modelos *logit,* vienen de muy atrás en el tiempo. En efecto, como señala Cramer (1991), a mediados del siglo pasado fue diseñada la función logística como una curva de crecimiento y, ya en los años treinta, la bioestadística configuró el modelo de probabilidad bivariante, definido inicialmente como un modelo *probit* (utilizando como función de enlace la distribución normal de probabilidad). Sin embargo, es a finales de los años sesenta cuando esta técnica se convierte en un método estándar para el análisis de regresión de datos dicotómicos, principalmente en el campo de las ciencias de la salud como la bioestadística o la epidemiología[2]. Posteriormente, la utilización del análisis de regresión logística se ha extendido al resto de ciencias sociales como la sociología o las ciencias empresariales, donde ha sido objeto de numerosas aplicaciones, entre las que destacamos a modo de ejemplo el comportamiento del consumidor ante la compra de determinados productos o servicios, el análisis del éxito en la introducción de nuevos productos, el estudio del fracaso empresarial, la predicción de situaciones de quiebra o suspensión de pagos sobre la base de ratios económico-financieros o la predicción de actividades de I+D en empresas innovadoras.

2. FORMULACIÓN DEL MODELO

A continuación se exponen los aspectos más significativos del modelo a partir de las limitaciones del modelo de regresión lineal.

2.1. Limitaciones del modelo de regresión lineal

Supongamos que una entidad financiera prepara el lanzamiento de un nuevo producto. Con el fin de diseñar una adecuada política de promoción, el departamen-

[1] La regresión logística no establece ninguna restricción sobre la distribución de las variables independientes.

[2] Hosmer y Lemeshow (1989) y Cramer (1991) ofrecen una amplia bibliografía al respecto.

to de marketing estaría interesado en conocer hasta qué punto la aceptación del producto está relacionada con el nivel de ingresos de sus clientes. Con esta intención se pregunta a una muestra aleatoria de los mismos si estarían dispuestos a suscribir el nuevo producto. Los resultados de la encuesta se muestran en la tabla 10.1, en la cual la variable respuesta Y, «¿adquiriría usted el producto A?», ha sido codificada con valor 1 en caso de respuesta afirmativa y con valor cero en caso contrario. Por otra parte, la variable independiente X, que se supone influye en la anterior, representa el nivel de ingresos de cada encuestado (en 10^5 pesetas).

TABLA 10.1

*Datos ejemplo 1**

Caso	Compra	Ingresos	Caso	Compra	Ingresos	Caso	Compra	Ingresos
1	1	50,2	15	1	62	29	1	50,8
2	1	70,3	16	0	50,8	30	0	37,5
3	1	62,9	17	1	56,2	31	0	41,3
4	1	48,5	18	0	43,2	32	1	63,6
5	1	57,2	19	1	50,4	33	1	54
6	1	75	20	0	44,1	34	0	45
7	1	46,2	21	0	38,3	35	1	68
8	1	57	22	0	55	36	1	62,1
9	1	64,1	23	0	46,1	37	0	35
10	0	32	24	0	35	38	0	34,5
11	1	73,4	25	0	37,3	39	0	39,4
12	1	71,9	26	0	41,8	40	0	37
13	0	56,2	27	0	37	41	1	54,5
14	1	49,3	28	0	33,4	42	1	38,2

En estas condiciones, el modelo más sencillo que puede plantearse entre ambas variables es el lineal:

$$Y = \beta X + e$$

de tal modo que cada valor $Y = y_i$, se obtendrá:

$$y_i = \beta_1 + \beta_2 x_i + e_i$$

Obsérvese que en este caso la variable independiente Y sólo puede tomar dos valores, cero o uno, de forma que la probabilidad de que suceda una u otra opción dependerá de cada valor que tome la variable explicativa X. Así pues, la variable respuesta (aleatoria) sigue una distribución binomial $B(1, p_i)$, donde p_i es la proba-

* Véase fichero en la dirección www.ugr.es/~tluque.

bilidad de que un cliente con un nivel de ingresos x_i se muestre partidario de adquirir el nuevo producto, es decir:

$$P(Y = 1/X = x_i) = p_i$$

y, por tanto,

$$P(Y = 0/X = x_i) = 1 - p_i$$

En el modelo general de regresión lineal, la esperanza de la variable Y para un valor de $X = x_i$ es

$$E(Y/x_i) = E(\beta_1 + \beta_2 X_i + e_i) = \beta_1 + \beta_2 x_i$$

ya que $E(e_i) = 0$.

Pero en el caso concreto que nos ocupa, donde la variable dependiente sigue distribución binomial, la esperanza condicionada a los valores de X es igual a:

$$E(Y/x_i) = 1 \cdot p_i + 0 \cdot (1 - p_i) = p_i$$

en consecuencia:

$$p_i = \beta_1 + \beta_2 x_i$$

En definitiva, esto implica que las predicciones obtenidas con el modelo $(\hat{y}_i)$ pueden interpretarse en términos de probabilidad. Esto es, el modelo de regresión lineal aplicado a una variable dependiente dicotómica estima la probabilidad de que la característica estudiada esté presente en los elementos de la población definidos por $X = x_i$.

En nuestro ejemplo, el modelo de regresión lineal ajustado por MCO sería:

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i = -0{,}983 + 0{,}03005 \text{ (ingresos)} = \hat{p}_i$$

que proporciona una estimación de la probabilidad de que un individuo de la muestra esté dispuesto a adquirir el nuevo producto dado su nivel de ingresos.

A pesar de su simplicidad y sencillez, no sólo en el planteamiento sino también en su interpretación, la formulación de un modelo lineal para explicar el comportamiento de variables independientes dicotómicas en términos de probabilidad presenta serios problemas que lo hacen inapropiado[3]:

[3] Véase, por ejemplo, Hosmer y Lemeshow (1989, 5-7); Peña (1994, 501); Ruiz-Maya et al. (1995, 543-545); Jovell (1995, 2).

1. Puesto que $\hat{\beta}_1 + \hat{\beta}_2 x_i$ estima una probabilidad, es obligado que, para cualquier valor de la variable independiente X (nivel de ingresos), dicha predicción debe de estar comprendida entre 0 y 1. Sin embargo, no existe garantía de que esto ocurra siempre. En la figura 10.1 se han representado los valores estimados por el ajuste de regresión lineal para los 42 individuos encuestados. Se aprecia claramente que algunas de las probabilidades estimadas superan el valor 1 y otras se encuentran por debajo del valor 0, algo evidentemente inadmisible. Por ejemplo:

$$\hat{y}_6 = -0{,}983 + 0{,}03005 x_6 = 1{,}2707$$
$$\hat{y}_{10} = -0{,}983 + 0{,}03005 x_{10} = -0{,}0214$$

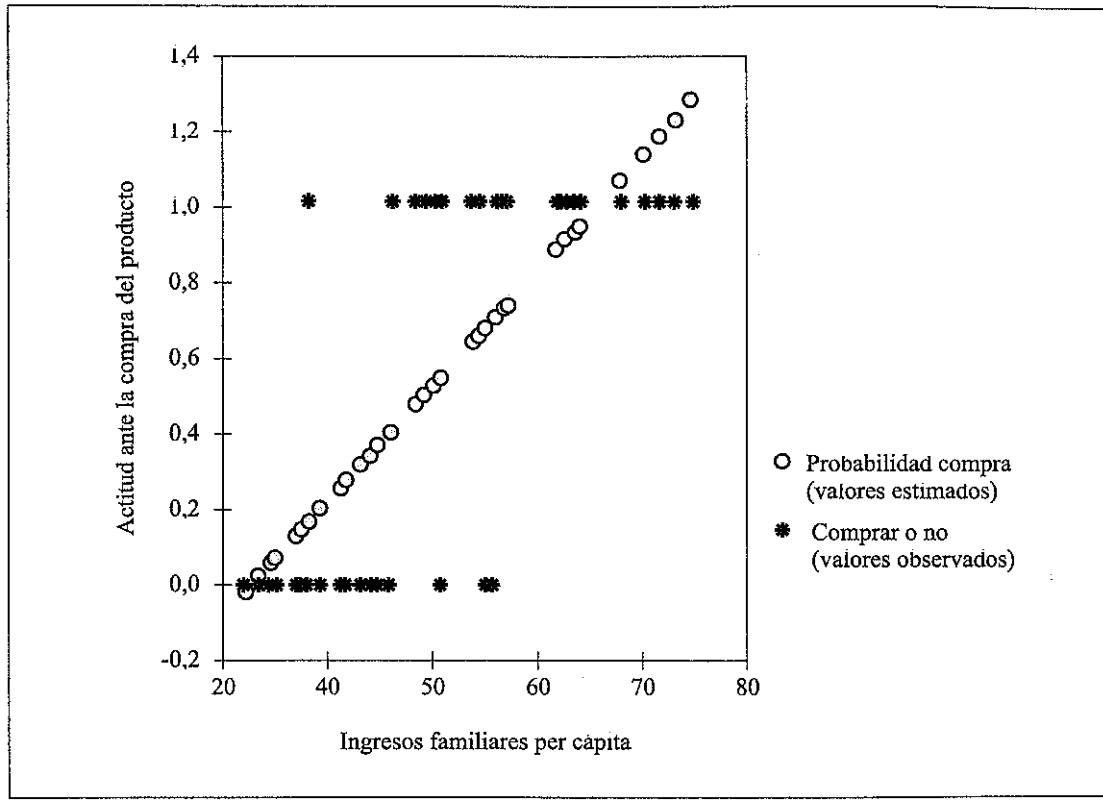

Figura 10.1. Ajuste de los datos ejemplo 1 con regresión lineal.

2. La distribución de los errores (e_i) es discreta con valores $-(\beta_1 + \beta_2 x)$ y $1 - (\beta_1 + \beta_2 x)$ según y_i sea igual a 0 o 1, respectivamente. Se viola la asunción de normalidad de los errores y, consecuentemente, los estimadores minimocuadráticos no serán eficientes.
3. La varianza de e_i puede calcularse fácilmente una vez conocidas las probabilidades con que la varianza aleatoria e_i toma los dos valores citados:

$$P[e_i = 1 - (\beta_1 + \beta_2 x_i)] = P[Y = 1/i] = p_i = \beta_1 + \beta_2 x_i$$
$$P[e_i = -(\beta_1 + \beta_2 x_i)] = P[Y = 0/i] = 1 - p_i = 1 - (\beta_1 + \beta_2 x_i)$$

Así,

$$\text{Var}(e_i) = p_i(1 - \beta_i - \beta_2 x_i)^2 + (1 - p_i)(-\beta_1 - \beta_2 x_i)^2 =$$
$$= p_i(1 - p_i)^2 + (1 - p_i)p_i^2 = p_i(1 - p_i) = (\beta_1 + \beta_2 x_i)(1 - \beta_1 - \beta_2 x_i)$$

por lo que el término error presenta heteroscedasticidad.

4. La hipótesis de normalidad de la variable dependiente tampoco se cumple cuando es dicotómica, como en este caso[4].

2.2. El modelo de regresión logística

Una excelente alternativa para garantizar que la respuesta prevista esté entre 0 y 1 es utilizar una función de enlace no lineal que sea monótona, creciente y acotada entre dichos valores. En estas circunstancias cabría utilizar cualquier función de distribución de variables aleatorias, de tal modo que el modelo quedaría:

$$p_i = F(\beta_0 + \beta_1 x_i)$$

es decir, que la probabilidad de que un cliente adquiera el nuevo producto ($P[Y = 1/i] = p_i$) viene expresada por una función de distribución (no lineal) de su nivel de ingresos (x_i).

El modelo de regresión logística surge cuando se utiliza la función de distribución *logística* para modelizar la relación entre la probabilidad de $Y = 1$, condicionada a un determinado valor de la variable (o variables) independiente, x_i:

$$p_i = \frac{e^{\beta_1 + \beta_2 x_i}}{1 + e^{\beta_1 + \beta_2 x_i}} \tag{1}$$

$$(1 - p_i) = \frac{1}{1 + e^{\beta_1 + \beta_2 x_i}} \tag{2}$$

Para el ajuste de este modelo y la estimación de los parámetros $\hat{\beta}_1$ y $\hat{\beta}_2$ no puede seguirse, como en el caso de la regresión lineal, el método de mínimos cuadrados puesto que, como ya hemos comentado, cuando se aplica al caso de variables dependientes dicotómicas, el modelo resultante presenta heteroscedasticidad.

Una alternativa de uso general para la estimación de los parámetros consiste en utilizar el procedimiento de *estimación por máxima verosimilitud* (EMV). En sínte-

[4] Algunos de estos inconvenientes pueden ser parcialmente superados con un procedimiento de estimación en dos etapas (*MCO* y *MCP*) y utilizando las frecuencias relativas de aceptación p_i como variable dependiente continua con una distribución normal en lugar de la variable original Y dicotómica, para lo cual será necesario contar con observaciones repetidas para cada valor de X.

sis, este método proporciona unos valores ($\hat{\beta}_1$ y $\hat{\beta}_2$) para los parámetros desconocidos (β_1 y β_2) que maximizan la probabilidad de que con ellos se obtengan los valores observados. Para aplicar la EMV se precisa construir, en primer lugar, la denominada función de verosimilitud (L) que expresa la probabilidad de los datos observados como una función de parámetros desconocidos. Los valores que maximizan la función L serán los *estimadores maximoverosímiles* de dichos parámetros[5]. Los principales paquetes estadísticos (GLIM, BMDP, SAS, SPSS, etc.) contienen un modelo de regresión logística. En concreto, utilizaremos el procedimiento *regresión logística* del programa SPSS 7.5.

Así pues, una vez ajustado el modelo y obtenidos los estimadores maximoverosímiles $\hat{\beta}_1$ y $\hat{\beta}_2$, la estimación de la probabilidad $\hat{p}_i$ es inmediata:

$$\hat{p}_i = \frac{e^{\hat{\beta}_1+\hat{\beta}_2 x_i}}{1+e^{\hat{\beta}_1+\hat{\beta}_2 x_i}} \tag{3}$$

$$(1-\hat{p}_i) = \frac{1}{1+e^{\hat{\beta}_1+\hat{\beta}_2 x_i}} \tag{4}$$

Para los datos del ejemplo 1:

$$\hat{p}_i = p\ (\text{comprar el producto}) = \frac{e^{-11,449+0,2371\ (\text{ingresos})}}{1+e^{-11,449+0,2371\ (\text{ingresos})}}$$

$$(1-\hat{p}_i) = p\ (\text{no comprar el producto}) = \frac{1}{1+e^{-11,449+0,2371\ (\text{ingresos})}}$$

De este modo, la probabilidad estimada de que un cliente de la entidad financiera se muestre dispuesto a adquirir el producto en cuestión es una función no lineal (logística) de sus ingresos, tal y como se muestra en la figura 10.2.

Para familiarizar al lector con las salidas de SPSS, ofrecemos la obtenida con los datos del ejemplo (tabla 10.2), donde B representa los estimadores $\hat{\beta}_i$. El resto de la información será tratada en apartados posteriores cuando generalicemos el modelo al caso multivariante:

TABLA 10.2

Variable	*B*	SE	Wald	Df	Sig.	*R*	Exp (*B*)
INGRESOS	0,2371	0,0698	11,5564	1	0,0007	0,4055	1,2676
CONSTANT	−11,4491	3,3797	11,4762	1	0,0007		

[5] No vamos a entrar en el procedimiento de estimación maximoverosímil para la regresión logística, limitándonos a remitir al lector a las publicaciones específicas sobre el tema. Por ejemplo, Ruiz-Maya et al. (1995), Sharma (1996), Hosmer y Lemeshow (1989) o Peña (1994).

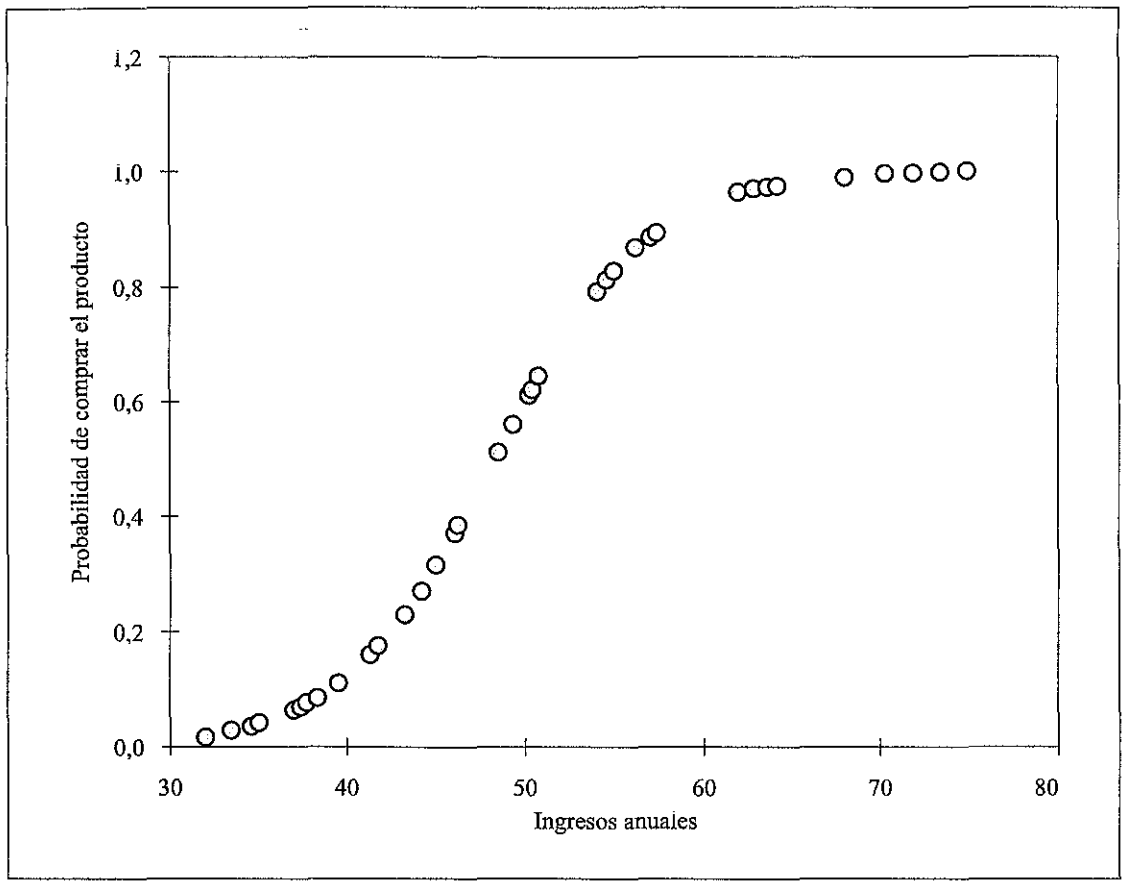

Figura 10.2. Probabilidades estimadas de comprar el producto.

Para terminar esta visión introductoria al modelo de regresión logística es necesario definir dos conceptos básicos relacionados con el mismo y que serán de suma utilidad para su más completa comprensión.

El primero es el *odds* o «ventaja»[6] de que un suceso ocurra. Se define como el cociente entre la probabilidad de que ocurra un suceso y su probabilidad complementaria, esto es, de que no ocurra: $p_i/1 - p_i$, e indica la «preferencia» de elegir la opción 1 de la variable respuesta frente a la opción 0 (Ruiz-Maya et al., 1995, 548). Por ejemplo, la ventaja de obtener «cara» en el lanzamiento de una moneda es 0,5/0,5 = 1, mientras que la ventaja de extraer una carta del palo de espadas de una baraja es 0,25/0,75 = 1/3. Estos resultados implican, en el primer caso, que la ventaja de obtener cara en el lanzamiento de una moneda es de 1 a 1, y en el segundo que la ventaja de extraer una espada de la baraja es de 1 a 3, o 0,33 a 1. En definitiva, ventaja y probabilidades proporcionan la misma información, aunque de forma diferente.

[6] Siguiendo a Ruiz-Maya et al. (1995), hemos traducido el término inglés *odds* por «ventaja». Este término es muy utilizado en el mundo de las apuestas en los países anglosajones.

Así pues, operando con las expresiones (3) y (4) podemos obtener la ventaja estimada de la opción 1 de la variable respuesta frente a la opción 0 ($\hat{\Omega}_i$):

$$\hat{\Omega}_i = \frac{\hat{p}_i}{1-\hat{p}_i} = e^{\hat{\beta}_1+\hat{\beta}_2 x_i} = e^{\hat{\beta}_1} \cdot e^{\hat{\beta}_2 x_i} \tag{5}$$

expresión que, como veremos más adelante, facilita la interpretación de los parámetros. En nuestro ejemplo, la ventaja de la opción «adquirir el producto» frente a «no adquirirlo» será:

$$\hat{\Omega}\ (\text{comprar}) = \frac{\hat{p}\ (\text{comprar})}{\hat{p}\ (\text{no comprar})} = e^{-11{,}449+0{,}2371\ (\text{ingresos})}$$

En el caso de un nivel de ingresos de 56,2, la ventaja estimada de que el cliente adquiera el producto es:

$$\hat{\Omega}_{56,2} = e^{-11{,}449} \cdot e^{0{,}2371(56{,}2)} = 6{,}527$$

es decir, que un cliente con unos ingresos de 56,2 tiene una «preferencia» por comprar el producto 6,527 veces mayor que por no comprarlo. Por otro lado, la probabilidad estimada de que dicho cliente adquiera el producto será $\hat{p}_{56,2} = 0{,}867$.

El segundo concepto al que hacíamos referencia es la denominada *transformación logística*, definida como el logaritmo de la ventaja o preferencia de la opción 1 frente a la opción 0. Aplicando esta transformación a (5) obtenemos una expresión equivalente:

$$\ln \hat{\Omega}_i = \ln\left[\frac{\hat{p}_i}{1-\hat{p}_i}\right] = \hat{\beta}_1 + \hat{\beta}_2 x_i \tag{6}$$

Hay que destacar que mientras que la probabilidad se expresa a través de un modelo no lineal (logístico), el logaritmo de las ventajas (también denominado *logit*) sí lo es, lo cual facilita la interpretación del modelo.

Resumiendo, tenemos tres expresiones equivalentes del modelo de regresión logística:

1. $\hat{p}_i = e^{\hat{\beta}_1+\hat{\beta}_2 x_i}/1 + e^{\hat{\beta}_1+\hat{\beta}_2 x_i}$, que estima la probabilidad de que un sujeto elija la opción 1 dado un determinado valor de $X = x_i$. En este caso se cumple que $(0 \leqslant \hat{p}_i \leqslant 1)$.
2. $\hat{\Omega}_i = \hat{p}_i/1 - \hat{p}_i = e^{\hat{\beta}_1+\hat{\beta}_2 x_i} = e^{\hat{\beta}_1} \cdot e^{\hat{\beta}_2 x_i}$, que estima la ventaja o preferencia de un individuo por la opción 1 frente a la opción 0 de la variable dependiente para cada valor de la variable (variables) independiente; de modo que si $P(Y=1) = 0$, entonces, $\Omega_i = 0$; y si $P(Y=1) = 1$, $\Omega_i = \infty$; por tanto, $(0 \leqslant \Omega_i \leqslant \infty)$.
3. $\ln \hat{\Omega}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$, que estima el logaritmo de la ventaja Ω_i, o *logit*, cuyo campo de variación estará entre $-\infty$ y ∞: para $P(Y=1) = 0$ y $P(Y=1) = 1$, respectivamente.

3. EL MODELO MULTIVARIANTE

Al igual que en el modelo de regresión lineal, la generalización del modelo univariante de regresión logística a un contexto multivariante es inmediata. Se trata de estimar la probabilidad de que una respuesta binaria ocurra, $P(Y = 1)$, en función de los valores que tomen un conjunto de variables explicativas o predictivas, que pueden ser continuas o categóricas, x_{1i}, x_{2i}, ..., x_{ni}. Así:

$$\hat{p}_i = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_{1i}+\hat{\beta}_2 x_{2i}+\cdots+\hat{\beta}_n x_{ni}}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_{1i}+\hat{\beta}_2 x_{2i}+\cdots+\hat{\beta}_n x_{ni}}} \tag{7}$$

o sus expresiones equivalentes:

$$\hat{\Omega}_i = \frac{\hat{p}_i}{1-\hat{p}_i} = e^{\hat{\beta}_0+\hat{\beta}_1 x_{1i}+\hat{\beta}_2 x_{2i}+\cdots+\hat{\beta}_n x_{ni}} \tag{8}$$

$$\ln \hat{\Omega} = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_n x_{ni} \tag{9}$$

Al igual que en el modelo univariante, el ajuste del modelo y la estimación de los parámetros se realiza por el método de máxima verosimilitud.

Sobre este modelo multivariante, y apoyándonos en un ejemplo, iremos analizando los aspectos centrales del modelo de regresión logística. Comenzaremos por una aclaración necesaria sobre la codificación de las variables independientes de naturaleza categórica, para continuar con las pruebas de significación de los parámetros estimados, su interpretación, la evaluación de la bondad del ajuste y, por último, comentaremos los métodos más usuales de selección de variables independientes.

4. EL PROCESO DE REGRESIÓN LOGÍSTICA

4.1. Codificación de las variables independientes categóricas

Al diseñar el modelo, puede resultar conveniente la inclusión de variables independientes de naturaleza categórica, tales como sexo, estado civil, lugar de residencia, etc. En estos casos no parece correcto su inclusión en el modelo representado por las expresiones (7), (8) y (9) como si se tratara de variables continuas. La razón es que los posibles valores que pueden tomar este tipo de variables no tienen un significado numérico, sino que tan sólo indican la presencia (o ausencia) de cierto atributo (en el caso de variables dicotómicas) o la pertenencia del sujeto a una determinada categoría o nivel, del cual el valor tomado por la variable es un simple

identificador. Así pues, la incorporación de variables cualitativas al modelo no es automática, sino que precisa de un procedimiento previo de adaptación[7].

De este modo, ante la presencia de una variable cualitativa con k niveles, será necesario diseñar o definir $k-1$ variables dicotómicas («ficticias o de diseño»), cada una de las cuales representará un nivel o categoría de la variable original. Puesto que hay $k-1$ variables ficticias y k niveles, la pertenencia de un individuo al nivel no representado explícitamente por ninguna variable de diseño se entenderá que sucede si no pertenece a ninguno de los otros niveles. El siguiente paso será codificar dichas variables ficticias. Se ha optado por un criterio de codificación sencillo y que facilite una clara interpretación de los parámetros. Veamos una aplicación a partir de un ejemplo:

Ejemplo 2: El responsable de marketing de la entidad financiera a la que nos hemos referido considera que la predisposición de sus clientes a adquirir el nuevo producto depende no sólo de los ingresos, sino de otras variables, en concreto de la edad, del grado de confianza en el sistema público de pensiones y del hecho de ser propietario o no de vivienda. Las escalas de medida originales de estas variables son:

— EDAD: «menos de 40 años», «de 40 a 60 años» y «más de 60 años».

— VALPENSI: puntuar en una escala desde 1 (ninguna confianza) a 9 (plena confianza).

— VIVIENDA: 1 (propietario), 2 (no propietario).

En el ejemplo encontramos dos variables cualitativas: VIVIENDA y EDAD. Para la primera, al ser dicotómica, no habría que definir ninguna nueva variable, pues bastaría con recodificar la existente dando valor 1 al hecho de ser propietario de una vivienda y 0 en caso contrario. Para la variable EDAD se definirían dos variables ficticias $EDAD_1$ y $EDAD_2$. La primera tomaría valor 1 cuando el individuo tenga entre 40 y 60 años y 0 en caso contrario, mientras que la segunda sería igual a 1 para aquellos sujetos mayores de 60 años y 0 para cualquier otra situación. Lógicamente, cuando el encuestado tenga menos de 40 años (la categoría restante de la variable EDAD), $EDAD_1$ y $EDAD_2$ deberán ser ambas igual a 0. En la tabla 10.3 se ilustra el sistema de codificación para las variables EDAD y VIVIENDA.

Con este método de codificación el nivel de la variable original, definido por defecto («menos de 40 años», «no propietario») se denomina categoría de referencia, ya que es sobre este nivel sobre el que se realiza la interpretación de los parámetros, como se comenta más adelante. El modelo de regresión logística en SPSS ofrece la posibilidad de elegir entre varios métodos de codificación de variables ficticias que, por otro lado, son creadas automáticamente una vez que el usuario las identifica como tales en la correspondiente ventana de diálogo del programa. El que hemos comentado es definido por SPSS como «INDICADOR».

[7] Hosmer y Lemeshow (1989) y Ruiz-Maya et al. (1995).

TABLA 10.3.A)

Edad	Variables de diseño	
	$EDAD_1$	$EDAD_2$
Menos 40 años	0	0
De 40 a 60	1	0
Más de 60	0	1

TABLA 10.3.B)

Vivienda	Variable de diseño
	$VIVIENDA_1$
Propietario	1
No propietario	0

La incorporación de las variables ficticias al modelo ajustado hace que éste cambie su formulación. En el caso del ejemplo 2, obtendríamos las siguientes expresiones:

$$\hat{p}\,(\text{comprar}) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 INGRESOS+\hat{\beta}_2 VALPENSI+\hat{\beta}_3 EDAD_1+\hat{\beta}_4 EDAD_2+\hat{\beta}_5 VIVIENDA_1}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 INGRESOS+\hat{\beta}_2 VALPENSI+\hat{\beta}_3 EDAD_1+\hat{\beta}_4 EDAD_2+\hat{\beta}_5 VIVIENDA_1}}$$

y análogamente,

$$\hat{\Omega}_i = \frac{\hat{p}\,(\text{comprar})}{\hat{p}\,(\text{no comprar})} =$$

$$= e^{\hat{\beta}_0+\hat{\beta}_1 INGRESOS+\hat{\beta}_2 VALPENSI+\hat{\beta}_3 EDAD_1+\hat{\beta}_4 EDAD_2+\hat{\beta}_5 VIVIENDA_1}$$

$$\ln \hat{\Omega}\,(\text{comprar}) =$$

$$= \hat{\beta}_0 + \hat{\beta}_1 INGRESOS + \hat{\beta}_2 VALPENSI + \hat{\beta}_3 EDAD_1 + \hat{\beta}_4 EDAD_2 + \hat{\beta}_5 VIVIENDA_1$$

4.2. Contraste de hipótesis sobre la significación de los coeficientes de regresión

Una vez ajustado el modelo y estimados sus coeficientes, el investigador debe centrar su atención en comprobar si las variables independientes que lo integran están relacionadas «significativamente» con la variable respuesta o dependiente. Como en el caso del modelo de regresión lineal, esto implica plantear y contrastar hipótesis estadísticas sobre los coeficientes de regresión, bien sea de forma individual o conjunta.

En síntesis, el contraste de hipótesis sobre la significación de los coeficientes de las variables es un intento de responder a la siguiente pregunta: ¿qué modelo nos aporta una información más acertada sobre el comportamiento de la variable respuesta, el que contiene la(s) variable(s) en cuestión o el que no la(s) incluye? Si la respuesta a esta pregunta se resuelve a favor de la primera alternativa y los valores

estimados por el modelo con la(s) variable(s), entonces diremos que la(s) variable(s) en cuestión es «significativa».

Al igual que en el modelo de regresión lineal, las pruebas de significación de las variables se formulan en los siguientes términos: contrastar la hipótesis nula, H_0, de que un coeficiente de regresión o un conjunto de ellos es cero contra la hipótesis alternativa H_1, derivada del rechazo de lo establecido por H_0.

Para el modelo ajustado derivado del ejemplo 2 vamos a distinguir entre el contraste de significación de un solo coeficiente y el contraste conjunto de todos los que intervienen en el modelo.

Una de las formas más comunes de contrastar la hipótesis de que un coeficiente de regresión es cero (H_0: $\beta = 0$) se basa en el estadístico W de Wald, que para un grado de libertad es igual al cuadrado de la razón entre el estimador maximoverosímil del coeficiente de la variable independiente y un estimador de su error estándar:

$$W = \left[\frac{\hat{\beta}_i}{S\hat{E}(\hat{\beta}_i)}\right]^2$$

El estadístico resultante, bajo la hipótesis nula de que $\hat{\beta} = 0$, sigue una distribución ji-cuadrado (χ^2). Para variables categóricas, el número de grados de libertad será igual al número de categorías menos uno. Puesto que los estimadores de los errores estándar de los parámetros estimados son suministrados por los paquetes informáticos, obviaremos cualquier discusión sobre su cálculo. En concreto la salida de SPSS ofrece para cada coeficiente, además de esta información, el valor del estadístico W, sus grados de libertad y su nivel de significación. La tabla 10.4 muestra los resultados que ofrece SPSS para el ejemplo que estamos manejando.

TABLA 10.4

Variable	*B*	SE	Wald	Df	Sig.	*R*	Exp (*B*)
INGRESOS	0,1880	0,0922	4,1579	1	0,0414	0,1927	1,2068
VALPENSI	–0,0804	0,40009	0,0402	1	0,8410	0,0000	0,9227
EDAD			4,2444	2	0,1198	0,0648	
$EDAD_1$	4,8888	2,3957	4,1642	1	0,0413	0,1930	132,7999
$EDAD_2$	3,0575	1,9446	2,4721	1	0,1159	0,0901	21,2737
$VIVIENDA_1$	3,5774	1,7567	4,1469	1	0,0417	0,1922	35,7808
CONSTANT	–13,7898	5,2369	6,9338	1	0,0085		

El estadístico W para la variable INGRESOS es:

$$W = \left[\frac{\hat{\beta}_i}{S\hat{E}(\hat{\beta}_i)}\right]^2 = \left[\frac{0{,}188}{0{,}0922}\right]^2 = 4{,}1579$$

puesto que $P(\chi^2 \geq 4{,}1579) = 0{,}0414$, la hipótesis nula de que $\hat{\beta}_1 = 0$ puede ser rechazada para un nivel de significación del 5% y la variable INGRESOS es, pues, significativa. Igual ocurre para VIVIENDA ($p = 0{,}0417$) y $EDAD_1$ ($p = 0{,}0413$).

Sobre el estadístico W hay que hacer notar que puede presentar un comportamiento anómalo en presencia de coeficientes de regresión demasiado altos, como consecuencia de errores típicos elevados[8]. En estas circunstancias parece recomendable acudir a otro procedimiento alternativo, para evaluar la significación de un coeficiente de regresión como el basado en el test de la razón de verosimilitud, que precisaremos a continuación.

La tabla 10.4 también ofrece los valores del estadístico R_A de Atkinson, que mide la correlación parcial entre cada variable independiente y la variable dependiente, pudiendo variar desde –1 a +1[9]. Los valores positivos indican que cuando la variable incrementa su valor, también lo hace la verosimilitud de que el suceso representado por la variable dependiente ocurra. R_A se puede considerar como una medida de la contribución parcial de cada variable al modelo, y su formulación es:

$$R_A = \sqrt{\frac{W - 2p}{-2 \ln L_0}} \tag{11}$$

donde W es el estadístico de Wald, p el número de parámetros estimados (será mayor que 1 en el caso de variables categóricas como EDAD) y $-2 \ln L_0$ representa –2 veces el logaritmo de la verosimilitud del modelo que contiene sólo el término independiente.

4.2.1. Significación global de los coeficientes de regresión

En este caso, la hipótesis nula que se desea contrastar es que todos los coeficientes de las variables independientes son iguales a cero, es decir: $H_0 = \hat{\beta}_1 = \hat{\beta}_2 = \cdots = \hat{\beta}_n = 0$. En un modelo con término independiente $\hat{\beta}_0$, no rechazar H_0 sería tanto como admitir que el modelo que sólo incluye el término constante predice mejor los valores observados de Y que el modelo ajustado en cuestión con n variables predictoras (X_n). Por el contrario, si H_0 fuese rechazada, esto indicaría que al menos uno de los coeficientes (y probablemente todos) es distinto de cero.

En la regresión logística este contraste se realiza por medio del test G o prueba de la razón de verosimilitud que se define:

$$G = -2 \ln \left[\frac{\text{Verosimilitud del modelo sólo con la constante } (L_0)}{\text{Verosimilitud del modelo seleccionado } (L_p)} \right] \tag{12}$$

[8] Véase Hosmer y Lemeshow (1989) y Ato y López (1996).

[9] El signo del parámetro en cuestión es el aplicado al estadístico R_A.

que se distribuye como una χ^2 con $p-1$ grados de libertad, donde p representa el número de parámetros en el modelo sometido a estudio. Este estadístico se basa en la función de verosimilitud de cada modelo y, en definitiva, compara la probabilidad de que los datos estimados por cada uno de los modelos representen a los valores realmente observados de la variable respuesta[10].

En nuestro ejemplo, el valor de $G = 43{,}240$ con 5 grados de libertad permite rechazar la H_0 antes formulada, ya que la probabilidad de error ante este rechazo (significación) es prácticamente nula. SPSS nos ofrece este contraste bajo la denominación *model chi-square,* donde se puede observar el valor del estadístico G, sus grados de libertad y la probabilidad $P(\chi^2 > G)$ sobre la que se realiza el contraste de H_0, de tal modo que si dicha probabilidad es menor que el nivel de significación α fijado por el investigador (por ejemplo, $\alpha = 0{,}05$), entonces se podrá rechazar H_0 de que el valor de todos los coeficientes estimados es igual a 0, porque al menos uno de ellos es distinto de 0.

Como veremos en posteriores epígrafes, la prueba de razón de verosimilitud también se emplea para evaluar la bondad del ajuste del modelo seleccionado y para medir la mejora en el ajuste del modelo a los datos cuando se incluyen o se excluyen una o varias variables independientes. Por último, el lector habrá deducido que, en el caso univariante, el estadístico G sirve para determinar la significación de la única variable independiente ofreciendo una información alternativa al estadístico W de Wald.

4.3. Medidas de la bondad de ajuste

Las pruebas de contraste de hipótesis sobre la significación de las variables descritas en el apartado anterior no pueden ser consideradas como un medio de evaluar la bondad del ajuste del modelo, puesto que se limitan a valorar la mayor o menor adecuación de los datos estimados por dos modelos a los valores realmente observados de la variable respuesta (Jovell, 1995). Por medidas de la bondad del ajuste hemos de entender aquellas pruebas o procedimientos que evalúen el grado de efectividad absoluta del modelo considerado en cuanto a la descripción de la variable dependiente, es decir, cuán cerca están los valores estimados $\hat{y}_i$ de los realmente observados y_i. Analizaremos tres grupos de medidas de bondad del ajuste: las basadas en pruebas estadísticas de contraste de hipótesis, las derivadas de la comparación directa entre los valores estimados y observados de la variable respuesta y, por último, las que son análogas al coeficiente de determinación múltiple (R^2) de la regresión lineal.

[10] La expresión operativa de este estadístico puede verse, por ejemplo, en Hosmer y Lemeshow (1989, 15) o en Ruiz-Maya et al. (1995, 682).

4.3.1. Bondad del ajuste: contraste de hipótesis

Este tipo de medidas de bondad del ajuste se basa en contrastar la hipótesis nula H_0 de que el modelo seleccionado ajusta bien los datos por medio de un estadístico con una distribución conocida.

a) Desvianza (deviance)

El estadístico desvianza (D) se define como una función del logaritmo neperiano del cociente de la función de verosimilitud del modelo seleccionado y la del modelo saturado. Un modelo saturado es aquel que contiene tantos parámetros como datos y que predice perfectamente los valores observados. La desvianza tiene la siguiente expresión:

$$D = -2 \ln \left[\frac{\text{Verosimilitud del modelo seleccionado}}{\text{Verosimilitud del modelo saturado}} \right] \tag{13}$$

La cantidad entre corchetes se denomina razón de verosimilitud y el propio estadístico D es también llamado test o prueba de razón de verosimilitud[11]. Este estadístico se distribuye aproximadamente como una χ^2 con $N-p$ grados de libertad, donde N es el número de observaciones y p el número de parámetros contenidos en el modelo. En algunos tests y paquetes informáticos se suele utilizar $-2 \ln L(\hat{\beta})$ o -2 ln *likelihood* para referirse a la desvianza de un determinado modelo. SPSS utiliza esta última notación y ofrece tanto la desvianza del modelo sólo con la constante, como la del modelo seleccionado.

El lector puede comprobar cómo D es el origen de la prueba G de contraste de hipótesis para la significación conjunta de todas las variables incluidas en el modelo y que, en definitiva, este último estadístico no hace sino recoger el cambio en D debido a incluir las variables independientes en el modelo que sólo contiene el término constante. Así:

$G = D$ (para el modelo sin las variables, sólo con la constante) $- D$ (para el modelo con las variables)

Puesto que la verosimilitud del modelo saturado es la misma en ambos valores de D, la diferencia puede expresarse:

$$G = -2 \ln \left[\frac{\text{Verosimilitud del modelo sin la variable (sólo con la constante)}}{\text{Verosimilitud del modelo con las variables}} \right]$$

[11] La expresión operativa de la desvianza puede verse, por ejemplo, en Ruiz-Maya et al. (1995, 682) o en Hosmer y Lemeshow (1989, 14).

En nuestro ejemplo, SPSS ofrece en primer lugar el valor de la desvianza del modelo sólo con el término independiente -2 ln *likelihood* = 58,129, y una vez realizada la estimación maximoverosímil, la desvianza del modelo con todas las variables incluidas: -2 ln *likelihood* = 14,889. Aunque el programa no facilita los grados de libertad ni el valor p asociado a cada valor de D, es fácil realizar el contraste de la hipótesis nula de que el modelo en cuestión ajusta bien los datos para $\alpha = 0{,}05$ conociendo que D se distribuye como una χ^2 con 41 grados de libertad $(42 - 1)$ en el primer caso, y con 36 grados de libertad $(42 - 6)$ en el segundo. De este modo se rechazará H_0 para el modelo con el término constante, mientras que para el modelo seleccionado, la hipótesis no podría ser rechazada para un $\alpha = 0{,}05$, es decir, no podríamos rechazar que dicho modelo ajuste bien los datos (su verosimilitud no difiere estadísticamente de 1 para un nivel de significación de 5%)[12].

b) Prueba de la ji-cuadrado

Tanto esta prueba como la siguiente son medidas de bondad del ajuste que se basan en comparar los valores observados y los estimados por el modelo que se desea evaluar (valores esperados), todo ello, una vez más, bajo la H_0 de que dicho modelo ajusta bien los datos observados.

Esta prueba se basa en la obtención de un estadístico χ^2 que mide el nivel de discordancia que puede existir al comparar, para cada uno de los diferentes patrones de predictores existentes, el número de respuestas (afirmativas) observadas con la probabilidad estimada por el modelo[13]. Por patrón de predictores se entiende cada una de las diferentes combinaciones de valores que pueden adoptar las variables independientes incluidas en el modelo. Por ejemplo, las variables SEXO (1 = hombre; 0 = mujer) y ESTADO CIVIL (1 = soltero; 0 = casado) determinarían cuatro patrones de covariables, puesto que cada uno de los individuos que componen la muestra pueden clasificarse en uno de los siguientes grupos (patrones): hombre-soltero; hombre-casado; mujer-soltera; y mujer-casada. El estadístico χ^2, cuando el número de patrones de predictores $M < N$, es:

$$\chi^2 = \sum_{i=1}^{M} \frac{m_i(y_i - \hat{p}_i)^2}{\hat{p}_i(1 - \hat{p}_i)} \tag{14}$$

donde m_i es el número de casos incluidos en cada patrón de predictores, y_i la opción de la variable respuesta y $\hat{p}_i$ la probabilidad estimada por el modelo para el patrón de covariables i. Para grandes muestras el estadístico se distribuye, obviamente, como una ji-cuadrado con $M - p$ grados de libertad.

[12] El valor teórico de χ^2 con 41 grados de libertad para $\alpha = 0{,}05$ es $56{,}942 < 58{,}129$, mientras que el valor teórico de χ^2 con 36 grados de libertad para $\alpha = 0{,}05$ es $50{,}998 > 14{,}889$.

[13] Jovell (1995, 78) y Ruiz-Maya et al. (1995, 682).

En presencia de variables continuas, el número de patrones de predictores es muy probable que sea igual al número de observaciones muestrales $M \approx N$. En estos casos la prueba χ^2 tomaría la expresión:

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - \hat{p}_i)^2}{\hat{p}_i(1 - \hat{p}_i)} \tag{15}$$

puesto que n_i sería igual a 1. Hosmer y Lemeshow (1989) advierten de la obtención de valores p incorrectos cuando $M \approx N$; sin embargo, estos autores sugieren que, en los casos en que el modelo ajustado es el correcto, se puede utilizar la prueba χ^2 con $N - p$ grados de libertad con unos resultados razonables.

En nuestro ejemplo, que contiene variables continuas, el valor del estadístico χ^2 con 36 grados de libertad (42 – 6) es de 29,815, mientras que el valor teórico de la distribución para $\alpha = 0{,}05$ es de 50,998, lo cual indica que no se debe rechazar la hipótesis nula de que el modelo seleccionado ajusta bien los datos. El programa SPSS identifica esta prueba como *goodness of fit*.

c) *Prueba de Hosmer-Lemeshow*

Esta prueba es especialmente adecuada para evaluar la bondad del ajuste de aquellos modelos que incluyan una o varias variables independientes de tipo continuo y que cuenten con un número de patrones de predictores prácticamente igual al número de casos observados ($M \approx N$). Estos autores proponen ordenar de menor a mayor las N probabilidades estimadas (una para cada caso observado) y a continuación agruparlas en diez grupos de tal modo que en el primero de ellos se encuentren los $n_1 = N/10$ sujetos que tengan las probabilidades estimadas más bajas y en el último los $n_{10} = N/10$ sujetos con las probabilidades estimadas más elevadas. Estos grupos son conocidos como «deciles de riesgo». El *estadístico de bondad del ajuste de Hosmer-Lemeshow*, $\hat{C}$, se obtiene calculando el estadístico ji-cuadrado de Pearson de una tabla de 2×10 referida a las frecuencias observadas y estimadas para cada uno de los diez grupos. Aunque los principales paquetes estadísticos que desarrollan la regresión logística ofrecen una salida con el resultado de esta prueba, reproducimos a continuación la fórmula de cálculo de $\hat{C}$:

$$\hat{C} = \sum_{k=1}^{10} \frac{(o_k - n_k \bar{p}_k)^2}{n_k \bar{p}_k (1 - \bar{p}_k)} \tag{16}$$

donde n_k es el número de patrones de predictores del grupo k-ésimo,

$$o_k = \sum_{i=1}^{n_k} y_i$$

es decir, el número de respuestas afirmativas registradas para la variable respuesta ($Y = 1$) para los n_k patrones de predictores, y

$$\bar{p}_k = \sum_{i=1}^{n_k} \frac{m_i \hat{p}_i}{n_k}$$

la media de la probabilidad estimada.

Los resultados de aplicar la prueba de Hosmer-Lemeshow al ejemplo 2 se muestran en la tabla 10.5.

TABLA 10.5

Prueba de Hosmer-Lemeshow

Grupo	Compra = No		Compra = Sí		Total
	Observad	Esperad	Observad	Esperad	
1	4,000	3,997	0,000	0,003	4,000
2	4,000	3,985	0,000	0,015	4,000
3	4,000	3,937	0,000	0,063	4,000
4	3,000	3,772	1,000	0,228	4,000
5	4,000	2,597	0,000	1,403	4,000
6	1,000	1,421	3,000	2,579	4,000
7	0,000	0,167	4,000	3,833	4,000
8	0,000	0,083	4,000	3,917	4,000
9	0,000	0,030	4,000	3,970	4,000
10	0,000	0,010	6,000	5,990	6,000

	Ji-cuadrado	*gl*	Significación
Test de bondad del ajuste	5,5128	8	0,7016

Por ejemplo, la frecuencia observada de los clientes que comprarían el producto ($Y = 1$) para el sexto decil de riesgo es 3. Este valor se obtiene de sumar los valores estimados de la variables respuesta para los Y individuos de este grupo y se representaría como $o_6 = 3$. De forma similar, la correspondiente frecuencia estimada esperada para este decil es 2,579, que es la suma de las cuatro probabilidades estimadas por el modelo para esos cuatro clientes. Si asumimos que $n_6 = 4$ (un patrón de predictores para cada sujeto del grupo), entonces $n_6\bar{p}_6 = 2,579$ puesto que:

$$\bar{p}_6 = \sum_{i=1}^{4} \frac{m_i \hat{p}_i}{4} = \frac{2,579}{4}$$

ya que $m_1 = m_2 = m_3 = m_4 = 1$ (un individuo por cada uno de los patrones de predictores del decil). La frecuencia observada de los clientes que no compran el producto ($Y = 0$) para este decil es $4 - 3 = 1$ y la frecuencia estimada esperada es $4 - 2,579 = 1,421$.

El valor del estadístico $\hat{C}_0$ ofrecido por SPSS es de 5,5128, cuya probabilidad calculada para una distribución ji-cuadrado con ocho grados de libertad (10 – 2) es $P(\chi^2_8 > 5{,}5128) = 0{,}7016$, lo cual implica que el modelo ajusta bien los datos (no se puede rechazar H_0).

d) Bondad del ajuste: eficacia predictiva[14]

Otro modo de evaluar la bondad del ajuste del modelo seleccionado consiste en comparar las predicciones del mismo con los datos muestrales observados, siendo la tabla de clasificación —y una serie de medidas derivadas de la misma—, el procedimiento más utilizado para este fin. Junto a la tabla de clasificación también se comentará una salida adicional ofrecida por SPSS llamada histograma de probabilidades estimadas.

La *tabla de clasificación* es una tabla de doble entrada donde se clasifican los casos que componen la muestra según los valores observados de la variable respuesta (1, 0; sí, no; ausencia, presencia) y los valores pronosticados por el modelo estimado, de tal modo que, dado un valor de corte (generalmente 0,5), todos los casos cuya probabilidad estimada sea igual o mayor que este valor serán clasificados en el grupo que denota la presencia de la característica representada por la variable dependiente, mientras que aquellas observaciones que obtengan una probabilidad menor que 0,5 lo serán en el grupo que implica la ausencia de dicha característica. Una vez construida la tabla es conveniente arbitrar algunas medidas que actúan como índices de la eficacia predictiva del modelo. De acuerdo con Ato y López (1996, 196), la tabla de clasificación obtenida adoptará la siguiente forma:

TABLA 10.6

Observados	Pronosticados		Totales
	Negativo	Positivo	
Negativo	A	B	$(A + B)$
Positivo	C	D	$(C + D)$
Totales	$(A + C)$	$(B + D)$	N

Donde A y D son los casos clasificados correctamente por el modelo y B y D los incorrectamente clasificados. De este modo se puede definir los siguientes índices:

[14] Este grupo de medidas es considerado por algunos autores como una forma de evaluar la eficacia predictiva del modelo más que la bondad del ajuste (Ato y López, 1996) o incluso como una mera forma de clasificación de los resultados obtenidos (Sharma, 1996). No obstante, siguiendo a Jovell (1995), los hemos introducido en este apartado por cuanto sirven para evaluar la idoneidad del modelo considerado.

— Tasa de aciertos: $(A + D)/N$.

— Tasa de errores: $(B + C)/N$.

— Especificidad: proporción entre la frecuencia de negativos correctos y el total de resultados negativos observados $(A/(A + B))$.

— Sensibilidad: razón entre los positivos correctos y el total de positivos observados $(D/(C + D))$.

— Tasa de falsos negativos: $C/(A + C)$.

— Tasa de falsos positivos: $B/(B + A)$.

La tabla de clasificación correspondiente a nuestro ejemplo aparece detallada en la tabla 10.7, tal y como es ofrecida por la salida de SPSS:

TABLA 10.7

Tabla de clasificación de resultados

Observados	Pronosticados		
	No	Sí	
No	20	0	100%
Sí	2	20	90,91%
Global			95,24%

Es fácil comprobar que el programa ofrece la especificidad, $20/(20 + 0) = 100\%$; la sensibilidad, $20/(2 + 20) = 90{,}91\%$, y la tasa de aciertos $(20 + 20)/42 = 95{,}24\%$. De la misma forma se obtienen el resto de medidas:

— Tasa de errores: $(2 + 0)/42 = 4{,}76\%$.

— Tasa de falsos negativos: $2/(20 + 2) = 9{,}1\%$.

— Tasa de falsos positivos: $0/(0 + 20) = 0$.

Aunque la interpretación de estos resultados, y en especial de la tasa de aciertos, puede conducir a afirmar que el modelo goza de una alta eficacia predictiva, cabe preguntarse hasta qué punto son «buenas» estas tasas de clasificación. Para ello se puede recurrir a la comprobación de la significación estadística de la tasa global de aciertos siguiendo procedimientos similares a los discutidos en el análisis discriminante. Por ejemplo, a partir del test de Huberty, el número esperado de casos correctamente clasificados debidos al azar es:

$$e = \frac{1}{42}(20^2 + 22^2) = 21{,}05$$

obteniéndose el valor del estadístico Z^*, que se distribuye, aproximadamente, como una normal:

$$Z^* = \frac{(40 - 21{,}05)\sqrt{42}}{\sqrt{21{,}05(42 - 21{,}05)}} = 5{,}848$$

Para un nivel de significación $\alpha = 0{,}05$, el valor del estadístico Z^* (5,848 > 1,96) conduce a rechazar la hipótesis nula de que el número de casos correctamente clasificados por el modelo no difiere de la clasificación esperada sólo por efecto del azar, es decir, que la tasa de aciertos del modelo es significativamente mayor que la que se obtendría debido al azar.

El modelo de regresión logística de SPSS permite obtener una imagen adicional sobre la eficacia predictiva del modelo estimado por medio del denominado *histograma de probabilidades estimadas*. La figura 10.3 representa sobre un eje de coordenadas los casos agrupados en función de su probabilidad estimada (eje de abscisas), figurando con una notación que representa su pertenencia real (observada) de cada uno de los dos grupos definidos por la variable respuesta.

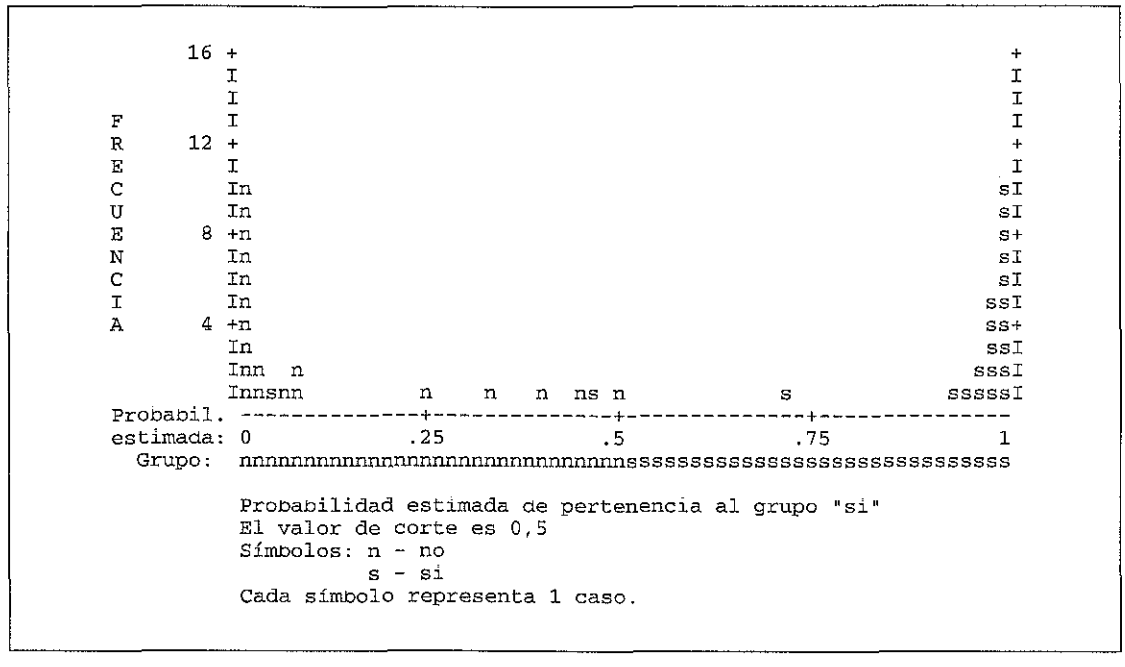

Figura 10.3. Probabilidades estimadas y grupos observados.

Si el modelo estimado distingue acertadamente los dos grupos, los casos para los que se ha observado que ocurre el fenómeno a estudiar ($Y = 1$) deberían estar situados a la derecha del punto de corte elegido (0,5), mientras que aquellos casos para los que se ha observado la ausencia del evento ($Y = 0$) se situarían a la izquierda de 0,5. Cuanto más agrupados estén ambos grupos en sus respectivos extremos mayor será la eficacia predictiva del modelo. En la figura 10.3 puede observarse que de los individuos que han declarado que comprarían el producto, sólo para

dos de ellos se ha estimado una baja probabilidad de que esto suceda: son los dos únicos casos mal clasificados por el modelo.

4.3.2. Bondad del ajuste: medidas similares a R^2

Para terminar, recogemos una medida de la bondad del ajuste en cierto modo análoga al coeficiente de determinación múltiple (R^2) de la regresión lineal. Aunque a priori un estadístico de este tipo podría parecer la solución idónea para evaluar la bondad del ajuste, las fuentes bibliográficas consultadas desaconsejan su uso para este fin, advirtiendo de sus deficiencias en cuanto a su capacidad de discriminación y de explicación de la variabilidad en relación con el coeficiente de determinación propiamente dicho utilizado en la regresión lineal[15]. Por esta razón, nos limitaremos a presentar esta medida sin entrar en más consideraciones.

$$R_L^2 = 1 - \frac{-2 \ln L(\hat{\beta})}{-2 \ln L(\hat{\beta}_0)} \tag{17}$$

donde $L(\hat{\beta}_0)$ es la función de verosimilitud del modelo que sólo contiene el término independiente y $L(\hat{\beta}_0)$ la verosimilitud del modelo seleccionado. En nuestro ejemplo R_L^2 toma el valor:

$$R_L^2 = 1 - \frac{14{,}889}{58{,}129} = 0{,}7438$$

La prueba R_L^2 ofrece valores entre 0 y 1, siendo los más cercanos a este último los indicativos de un buen ajuste.

4.4. Interpretación de los resultados

Una vez determinadas las variables significativas en el modelo seleccionado, el siguiente paso será determinar cómo influyen en la variable respuesta objeto de estudio, para lo cual se ha de proceder a interpretar sus respectivos coeficientes, $\hat{\beta}_i$. Igual que en el análisis de la regresión lineal, los coeficientes estimados de las variables independientes indican la magnitud de la influencia de cada una de estas variables sobre la variable dependiente. En el caso de la regresión lineal, la interpretación de los coeficientes es directa, ya que expresan el cambio producido en la escala de medida a la variable dependiente ante cambios unitarios de las variables independientes.

[15] Hosmer y Lemeshow (1989), Cramer (1991), Novales (1993) y Jovell (1995).

En los modelos de regresión logística, la interpretación de los coeficientes no es tan directa, debido fundamentalmente a la naturaleza no lineal del modelo original. Las transformaciones del mismo recogidas en las expresiones (5) y, sobre todo, (6) facilitan algo la tarea interpretativa; pero incluso en esta última el modelo lineal resultante lo es respecto al *logit*, y, por tanto, la interpretación de los coeficientes dependerá de que se sea capaz de comprender el significado de los cambios en esta escala de medida. Además, la distinta naturaleza de las variables independientes (dicotómicas, categóricas y continuas) conlleva singularidades en cuanto a la interpretación de sus parámetros. Así pues, y siguiendo el esquema propuesto por Hosmer y Lemeshow (1989), pasaremos a considerar la interpretación de los coeficientes en un modelo de regresión logística para cada una de las posibles escalas de medida de la variable independiente.

A) Variables independientes dicotómicas

Sea un modelo de regresión logística en el que una de las variables independientes es de naturaleza *dicotómica*. Según se ha definido anteriormente, la ventaja de la opción 1 de la variable respuesta ($Y = 1$) frente a la opción 0 para un determinado valor de la variable independiente X_{ni} es $\Omega_i = p_i/1 - p_i$. Puesto que en el caso que nos ocupa X_n sólo puede tomar los valores 1 y 0, la ventaja de la opción 1 para $x_n = 1$ será $\Omega_1 = p(Y = 1)/1 - p(Y = 1/x = 1) = p_1/1 - p_1$, y para $x_n = 0$, $\Omega_0 = p_0/1 - p_0$. Llegado a este punto, se define el *cociente de ventajas*, Ψ, como la razón de la ventaja de la opción 1 para $x_n = x_{ni}$ respecto a la ventaja para $x_n = x_{nj}$. En el caso de que x_n sólo tome dos valores (1, 0) el cociente de ventajas se referirá a $x_n = 1$ respecto a $x_n = 0$; por tanto, el cociente de ventajas $\Psi_{1,0}$ tomará la expresión:

$$\Psi_{1,0} = \frac{\Omega_1}{\Omega_0} = \frac{\dfrac{e^{(\beta_0+\beta_1)} / (1 + e^{(\beta_0+\beta_1)})}{1 / (1 + e^{(\beta_0+\beta_1)})}}{\dfrac{e^{\beta_0} / (1 + e^{\beta_0})}{1 / (1 + e^{\beta_0})}} = \frac{e^{(\beta_0+\beta_1)}}{e^{\beta_0}} = e^{\beta_1} \qquad (18)$$

Así pues, el cociente de ventajas referido a una determinada variable independiente dicotómica es una medida de asociación que indica cuánto más probable (o improbable) es que se presente el suceso que se está investigando ($Y = 1$) entre aquellos sujetos para los que $x_n = 1$ que entre aquellos otros con $x_n = 0$. En la tabla 10.4 que muestran los resultados obtenidos para nuestro ejemplo; la última columna denominada exp (B) ofrece los valores e^{β_n} para cada una de las variables independientes. En el caso concreto de la variable dicotómica VIVIENDA, el valor estimado de 35,78 significa que, manteniendo constantes el resto de las variables, el mostrarse favorable a la adquisición del nuevo producto financiero es 35,78 veces más probable que ocurra entre los sujetos que poseen vivienda propia (VIVIENDA = 1) que entre los que no la poseen (VIVIENDA = 0). En otras palabras, e^{β_n} indica el factor de cambio en la ventaja de que ocurra el suceso represen-

tado por la variable dependiente cuando la variable independiente dicotómica toma el valor 1 frente a la opción 0. Dicho de otro modo, la ventaja de la opción 1 de la variable respuesta «sí estaría dispuesto a adquirir el nuevo producto financiero» cuando el sujeto posee vivienda propia es superior en 35,78 veces a la correspondiente cuando el sujeto no posee vivienda propia (VIVIENDA = 0).

La interpretación de los coeficientes de las variables categóricas con más de dos opciones es similar, como ya se señaló anteriormente. Estas variables independientes han de ser recodificadas en tantas variables dicotómicas o variables de diseño como número de categorías existentes menos uno. Utilizando el método de codificación de la «categoría de referencia», la interpretación de cada uno de los coeficientes de las variables dicotómicas recodificadas se realizará de la forma antes expuesta. Claro que ahora e^{β_n} representa el factor de cambio en la ventaja de que ocurra el suceso representado por la variable dependiente cuando el sujeto pertenece a la categoría representada por la correspondiente variable de diseño frente al hecho de que el sujeto pertenezca a la categoría de referencia (representada por los valores 0 de las variables de diseño). Veamos la aplicación práctica con el ejemplo 2.

Recordamos que el modelo estimado contiene dos variables de diseño dicotómicas, $EDAD_1$ y $EDAD_2$, quedando la categoría de «menos de 40 años» como categoría de referencia, esto es, sobre la que se realiza la comparación (tabla 10.3). La tabla 10.4 muestra el valor de los coeficientes estimados para $EDAD_1$ y $EDAD_2$, $\hat{\beta}_3 = 4{,}88$ y $\hat{\beta}_4 = 3{,}057$. Manteniendo constantes el resto de las variables, se puede afirmar que:

— El estar dispuesto a adquirir el nuevo producto financiero es 132,79 veces ($e^{4,88}$) más probable que ocurra cuando el sujeto tiene entre 40 y 60 años ($EDAD_1 = 1$) que cuando tiene menos de 40 años (categoría de referencia).

— La ventaja de la opción 1 de la variable respuesta «estaría dispuesto a adquirir el nuevo producto financiero» cuando el individuo tiene más de 60 años ($EDAD_2 = 1$) es superior en 21,27 veces ($e^{3,057}$) a la correspondiente cuando el sujeto tiene menos de 40 años (categoría de referencia)[16].

B) Variables independientes continuas

En el caso de variables independientes continuas, la interpretación de sus coeficientes también se realiza sobre la base de la razón de ventajas. Así, ante un cambio unitario en la escala de medida de la variable independiente, la ventaja de la opción 1 de la variable dependiente se incrementará (o disminuirá) en un factor igual a $e^{\hat{\beta}_n}$. En el ejemplo 1, la variable independiente INGRESOS es de naturaleza continua, y una vez ajustado el modelo, el coeficiente $\hat{\beta}_1$ toma un valor de 0,2371, mientras que $e^{\hat{\beta}_1} = 1{,}267$ (tabla 10.2). Esto significa que un incremento de 100.000 pesetas en los ingresos (la unidad de medida de INGRESOS es 10^5 pesetas) provo-

[16] Se comprueba fácilmente a partir de la expresión (18) que:

$$\hat{\Psi}_{(EDAD(2),\ menos\ de\ 40)} = \hat{\Omega}_{(EDAD(2))}/\hat{\Omega}_{(menos\ 40)} = e^{\hat{\beta}_4} = e^{3,057}$$

cará un incremento multiplicativo por un factor 1,267 de la ventaja de la opción 1 de la variable dependiente. No obstante, el modelo puede interpretarse no sólo respecto al cociente de ventajas, sino también respecto a las propias ventajas o a la escala original de la variable respuesta, esto es, la escala de probabilidades. Veamos un ejemplo. Sean dos clientes *A* y *B* de la entidad financiera con unos ingresos de $50,5 \times 10^5$ y $38,5 \times 10^5$ pesetas, respectivamente. Aplicando el modelo de regresión logística estimado donde, $\hat{\beta}_0 = -11,4491$ y $\hat{\beta}_1 = 0,2371$, podría llegarse a la siguiente conclusión:

$$\hat{\Psi}_{(50,5;38,5)} = \frac{e^{-11,4491}e^{0,2371(50,5)}}{e^{-11,4491}e^{0,2371(38,5)}} = e^{0,237(50,5-38,5)} = 17,205$$

El valor 17,205 indica que la ventaja de la opción «adquirir el nuevo producto» para el cliente *A* es superior en más de 17 veces a la ventaja de «adquirir el nuevo producto» para el cliente *B*. Asimismo, dichas ventajas para los niveles de ingresos de ambos clientes tomarán dos valores:

$$A) \quad \hat{\Omega}_{50,5} = e^{-11,4491}e^{0,2371(50,5)} = 1,6895 = \frac{\hat{p}_{50,5}}{1-\hat{p}_{50,5}} = \frac{0,62818}{0,3718}, \text{ y}$$

$$B) \quad \hat{\Omega}_{38,5} = e^{-11,4491}e^{0,2371(38,5)} = 0,09819 = \frac{\hat{p}_{38,5}}{1-\hat{p}_{38,5}} = \frac{0,08941}{0,91058}$$

En este caso, los valores obtenidos reflejan la situación relativa en que se muestran las dos opciones de la variable dependiente cuando la variable explicativa INGRESOS toma, respectivamente, el valor 50,5 y 38,5. El cliente *A* es más «probable» que se incline por adquirir el nuevo producto (opción 1) que por «no adquirirlo» en una relación 1,689 a 1. Por su parte, el cliente *B* preferirá «adquirir el nuevo producto» en una proporción de 0,09819 a 1, o lo que es igual, preferirá «no adquirir el nuevo producto» en una relación 10,18 a 1. Se observa que el incremento en la ventaja de 0,09819 a 1,6895 es $17,205 = e^{\hat{\beta}_1(50,5-38,5)}$.

Por último, también la probabilidad estimada de que los clientes *A* y *B* declaren que estarían dispuestos a adquirir el nuevo producto es, respectivamente:

$$\hat{p}_{50,5} = \frac{e^{-11,4491+0,2371(50,5)}}{1+e^{-11,4491+0,2371(50,5)}} = 0,62218$$

$$\hat{p}_{38,5} = \frac{e^{-11,4491+0,2371(38,5)}}{1+e^{-11,4491+0,2371(38,5)}} = 0,08941$$

4.5. Valores extremos y colinealidad

A la hora de examinar la idoneidad del modelo de regresión logística seleccionado es importante valorar la posible presencia de *valores extremos (outliers)* que

puedan alterar el ajuste de los datos. Para ello existe un conjunto de *métodos de diagnóstico* basados en estadísticos o indicadores que examinan la relación existente entre los valores observados y los estimados por el modelo para cada caso o sujeto. En general, estas medidas pueden sintetizarse en dos grupos: valores residuales y medidas de influencia. Las primeras se apoyan en diferentes análisis de los residuos (diferencia entre los valores observados de la variable respuesta y los valores derivados del modelo) para cada observación a fin de detectar aquellos casos para los cuales el modelo no ajusta bien los datos. Entre éstos podemos citar:

— *El residuo* propiamente dicho.
— *El residuo estandarizado o tipificado:* es el residuo dividido por una estimación de su desviación típica. Para muestras grandes se distribuye como una normal (0,1).
— *La desvianza:* compara la probabilidad estimada de que el caso en cuestión pertenezca al grupo correcto respecto a la predicción perfecta (igual a 1). Como el anterior, este estadístico sigue aproximadamente una distribución normal y grandes valores del mismo indican que el modelo no ajusta bien el caso estudiado.
— *El residuo ajustado por el método de Student (studentized residual):* para cada caso mide el cambio en la desvianza del modelo si el caso fuera excluido.

Las medidas de influencia identifican aquellos sujetos que ejercen una notable influencia en las estimaciones derivadas del modelo. Entre otras son:

— El *valor de influencia (leverage)* se utiliza para detectar aquellos casos que tienen gran impacto sobre el ajuste del modelo.
— La *distancia de Cook* es una medida que cuantifica el cambio en los residuos de todos los casos cuando una determinada observación es excluida del cálculo de los coeficientes de regresión.
— El *cambio en los coeficientes del modelo* cuando se excluye un caso concreto. Este estadístico mide el cambio para cada coeficiente incluido el término independiente. Obviamente, la presencia de valores de cambio altos identifica observaciones que deberían ser examinadas con más detalle.

Un último aspecto a tener en cuenta en el análisis de idoneidad de un modelo de regresión logística es la presencia de *multicolinealidad,* es decir, de una elevada correlación entre las variables independientes. La presencia de una multicolinealidad elevada puede producir unos coeficientes estimados sesgados y unos errores estándar elevados, lo que alteraría tanto los valores estimados de las probabilidades como el resultado de las pruebas de Wald. A fin de no resultar reiterativos, remitimos al lector al capítulo de la regresión lineal para los aspectos relacionados con este fenómeno (diagnosis y posibles soluciones).

4.6. Selección de las variables independientes

La selección de las variables independientes a incluir en el modelo como predictores del fenómeno que se desea estudiar ha de realizarse siguiendo dos criterios que no tienen por qué conducir a resultados coincidentes: modelización estadística y modelización sustantiva (Jovell, 1995). Según el criterio estadístico tan sólo se incluirán en el modelo aquellas variables que tienen una capacidad de predicción estadísticamente significativa, es decir, que contribuyan a la mejora de la bondad del ajuste del modelo. En el criterio sustantivo, el investigador decide qué variables independientes debe incluir en función de la base teórica en la que se apoya la hipótesis de investigación que se pretende verificar.

Muchos paquetes estadísticos incorporan métodos de selección de variables independientes en el módulo de regresión logística. Antes de describir brevemente los contenidos en SPSS, hay que hacer dos advertencias. En primer lugar, parece haber un consenso generalizado al considerar que no existe un procedimiento de selección que garantice el «mejor» modelo en términos estadísticos, por lo que el empleo de uno u otro método puede conducir a diferentes resultados que habrán de ser interpretados en base a criterios de interpretabilidad y parsimonia. En segundo lugar, el modelo es seleccionado para ajustar una determinada muestra, por lo que la modelización estadística está condicionada por variaciones en los datos empíricos.

El programa SPSS presenta dos métodos de selección de variables que complementan al procedimiento de introducción conjunta de todas las variables en el modelo (INTRODUCIR): el método de introducción por pasos de las variables (ADELANTE) y el método de eliminación por pasos (ATRÁS).

— Selección hacia delante: se parte del modelo tan sólo con la constante. En cada paso se introduce la variable independiente que presente el nivel de significación más bajo en el estadístico *score*[17] siempre y cuando se encuentre por debajo del punto de corte (por defecto fijado en 0,05), y se examina el modelo resultante a fin de eliminar del mismo, de entre todas las variables que superen el punto de corte de salida (por defecto 0,1), aquella con el mayor nivel de significación en el estadístico de Wald (opción *WALD*) o en el estadístico *G* de cambio en la razón de verosimilitud (opción *RV*). El proceso concluye cuando ninguna variable pueda ser introducida ni retirada o cuando se llega a un modelo anteriormente considerado.

— Selección hacia atrás: se parte del modelo que contiene todas las variables independientes consideradas y se procede con el procedimiento anterior pero comenzando por la eliminación de las variables (opción *WALD* o *RV*). Para la introducción de las variables se utiliza también el estadístico *score*.

[17] El estadístico *score* es utilizado como alternativa al estadístico de Wald para la prueba de significación estadística de los coeficientes de regresión por su mayor facilidad de cálculo (Hosmer y Lemeshow, 1989, 17).

En la aplicación práctica de este capítulo se comentará con más detalle el proceso de selección de variables; tan sólo apuntar que todas las variables utilizadas para representar a la misma variable categórica entran o salen del modelo juntas.

5. CONSIDERACIONES FINALES

Para terminar el análisis del modelo de regresión logística creemos conveniente realizar unos breves comentarios sobre la metodología del diseño del estudio y sobre la relación entre la regresión logística y el modelo *logit*.

El propósito de la técnica del análisis de regresión logística es la obtención de una ecuación en virtud de la cual se pueda predecir, en términos de probabilidad, la ocurrencia de un determinado suceso objeto de la investigación en función de una serie de variables o predictores de carácter categórico o continuo. Esto implica que el diseño de la metodología de la investigación ha de cumplir las premisas básicas que, a nivel teórico, informan de la presencia de causalidad entre una variable dependiente (efecto) y una o varias variables independientes (causa) (Luque, 1997). En este sentido, los estudios de carácter longitudinal (esencialmente prospectivos) son los que a priori pueden garantizar el cumplimiento de tales asunciones, mientras que en un estudio transversal, la interpretación de la relación entre las variables debería realizarse en términos de asociación o correlación, a no ser que las características o naturaleza de las variables implicadas permita asumir en un sentido amplio las condiciones previas de causalidad.

No obstante, no es extraordinario que el investigador se enfrente con el problema de que la variable objeto de estudio presente a nivel poblacional una probabilidad de ocurrencia (o de no ocurrencia) muy baja. Así, ante estudios que intenten, por ejemplo, predecir el carácter innovador o no innovador de una empresa, o estimar la probabilidad de que un consumidor utilice la venta por catálogo o por Internet, el diseño de la recogida de datos mediante muestreo aleatorio entraña el riesgo de no registrar el volumen necesario de observaciones que presenten una de las dos opciones de la variable dependiente. En estos casos es corriente el recurso a un diseño retrospectivo en el cual se realiza un muestreo aleatorio entre los individuos o casos que presentan una de las opciones de la variable respuesta, y otro para aquellas observaciones en las que se produzca la segunda alternativa. En este tipo de estudios, muy utilizados en investigaciones médicas, parece más apropiado hablar de asociación entre la variable dependiente y las independientes que de predicción de un determinado valor de la primera dados unos valores de las segundas. Como afirma Christensen (1990, 235), «el análisis de datos como éstos sirve para describir las características de los dos grupos en términos de los factores explicativos». Hecha esta salvedad, el investigador puede ajustar un modelo de regresión logística con datos procedentes de estudios retrospectivos utilizando cualquier paquete estadístico que contenga un módulo específico para la regresión logística.

Por último, unas consideraciones sobre los modelos de respuesta discreta que utilizan la función logística como función de enlace. Siguiendo a Ato y López (1996, 190-191), existen dos enfoques alternativos:

— El *análisis de regresión logística* o *modelo logit con datos no agrupados* se utiliza con una o varias variables independientes continuas cuya presencia hace difícil encontrar casos individuales con los mismos valores en todas las variables.

— El *análisis logit* o *modelo logit con datos agrupados* pertenece al tipo del análisis de la varianza (o covarianza) del modelo lineal y se utiliza cuando es posible agrupar observaciones con valores iguales en las variables independientes, lo cual suele ocurrir cuando éstas son todas categóricas.

El análisis de regresión logística es más general y puede abordar cualquier situación representativa del análisis *logit*. Sin embargo, algunos paquetes estadísticos, como el SPSS, utilizan programas diferentes para cada uno de los enfoques. En la bibliografía se referencian algunos trabajos de naturaleza empírica que han utilizado uno u otro enfoque.

6. APLICACIÓN PRÁCTICA

Una empresa especializada en estudios de mercado recibe el encargo de un cliente de investigar los principales factores que pueden influir en aumentar la probabilidad de que un nuevo producto sea introducido con éxito en el mercado. Con este fin la empresa ha realizado una encuesta a 240 empresas industriales de las cuales 156 declararon haber intentado introducir en el mercado un nuevo producto. La codificación de la información recogida en los cuestionarios de estas 156 empresas está reflejada en la tabla 10.7.

Así pues, la empresa de investigación se enfrenta a una situación en la que se desea explicar el comportamiento en términos de probabilidad de una variable dependiente dicotómica (éxito o fracaso en el lanzamiento de un nuevo producto), en función de un conjunto de variables predictoras de naturaleza continua, categórica o dicotómica. En tales circunstancias parece adecuado proceder al ajuste de un modelo de regresión logística.

La resolución del problema se va a realizar mediante el módulo REGRESIÓN LOGÍSTICA del programa SPSS, destacando las fases que se relacionan:

1. Creación de la base de datos con la información disponible.
2. Seleccionar el módulo REGRESIÓN LOGÍSTICA en el menú ESTADÍSTICOS.
3. Elegir la variable dependiente (dicotómica) de la lista de variables desplegada e introducirla en la ventana dispuesta a tal efecto en el programa. En nuestro ejemplo dicha variable es ÉXITO.

TABLA 10.7

*Cuadro de codificación (aplicación práctica)**

Variable	Descripción	Valores
ÉXITO (dependiente)	¿Cuál ha sido el resultado de su intento por introducir un nuevo producto en el mercado?	0. Fracaso 1. Éxito
PUBLICID	Gastos en promoción y publicidad	En 10^4 euros
GRADNOVE	Grado de novedad del nuevo producto	1. Mejoras sustanciales en productos existentes 2. Productos totalmente nuevos
TIPO	Tipo de producto	1. Consumo final 2. Industrial
IMASD	¿Posee un departamento formal de investigación y desarrollo?	1. Sí 2. No
SECTECNG	Intensidad tecnológica del sector de actividad al que pertenece la empresa	1. Baja intensidad tecnológica 2. Intensidad media 3. Alta intensidad tecnológica
PERSONAL	Personal asalariado que participa en las tareas de investigación, desarrollo y lanzamiento de los nuevos productos	Número de empleados

4. Seleccionar las variables independientes e introducirlas en la ventana de COVARIABLES. En este caso las variables serían: PUBLICID, GRADNOVE, TIPO, IMASD, SECTECNG y PERSONAL.
5. Entre las independientes, marcar las que son categóricas (GRADNOVE, TIPO, IMASD y SECTECNG).
6. Recodificación de las variables categóricas.

* Véase fichero en la dirección www.ugr.es/~tluque.

7. Especificar las OPCIONES (estadísticos, método, etc.).
8. Ejecutar el programa.
9. Análisis de los resultados.

La recodificación de las variables categóricas originales es efectuada por el programa creando para cada una de ellas tantas *variables de diseño* como número de categorías menos una. De entre los criterios de recodificación disponibles, se ha optado por el criterio INDICADOR, tomando como categoría de referencia (no representada por ninguna variable de diseño) la última de cada variable original. La elección de la categoría de referencia es una opción del investigador que deberá calibrar, desde un punto de vista teórico, cuál de ellas es más conveniente a efectos de base de comparación. Las variables de diseño obtenidas se muestran en la siguiente tabla.

TABLA 10.8

Variables de diseño

Variable original		Variables de diseño	
		$SECTECNG_1$	$SECTECNG_2$
SECTECNG			
Baja intensidad	(1)	1	0
Intensidad media	(2)	0	1
Alta intensidad	(3)	0	0
GRADNOVE		$GRADNOVE_1$	
Productos mejorados	(1)	1	
Productos nuevos	(2)	0	
TIPO		$TIPO_1$	
Producto de consumo	(1)	1	
Producto industrial	(2)	0	
IMASD		$IMASD_1$	
Sí	(1)	1	
No	(2)	0	

Con esta información el programa realiza el *ajuste del modelo de regresión logística* introduciendo todas las variables (opción *INTRODUCIR*). En la siguiente tabla aparecen los coeficientes estimados (*B*), el error estándar (*EE*), el valor del estadístico de Wald y su significación, el valor del estadístico *R* y, por último, el valor de *e* elevado al coeficiente (Exp (*B*)):

TABLA 10.9

Variables en el modelo

Variable	*B*	*EE*	Wald	*gl*	Sig.	*R*	Exp (*B*)
PUBLICID	2,0224	0,5002	16,3443	1	0,0001	0,2602	7,5562
$GRADNOVE_1$	1,2240	0,5340	5,2537	1	0,0219	0,1239	3,4007
$TIPO_1$	–0,1427	0,7794	0,0335	1	0,8547	0,0000	0,8670
$I+D_1$	3,5733	1,3678	6,8249	1	0,0090	0,1509	35,6346
SECTECNG			9,2739	2	0,0097	0,1578	
$SECTECNG_1$	–2,1377	0,8252	6,7112	1	0,0096	–0,1491	0,1179
$SECTECNG_2$	–0,6543	0,8574	0,5823	1	0,4454	0,0000	0,5198
PERSONAL	0,4209	0,1492	7,9610	1	0,0048	0,1677	1,5234
Constante	–10,3346	2,4026	18,5015	1	0,0000		

Las pruebas de contraste de hipótesis sobre la *significación de los coeficientes* de regresión (*B*) arrojan los siguientes resultados:

— Sólo hay dos variables cuyos coeficientes no son significativos ($p > 0,05$) según el estadístico de Wald: $TIPO_1$ y $SECTECNG_2$.

— La prueba de la razón de verosimilitud o test *G* para contrastar la significación global de todos los coeficientes es ofrecido por la salida de SPSS con la denominación *model chi-square*. Este estadístico toma un valor de 107,499 con 7 grados de libertad que permite rechazar la hipótesis nula de que todos los coeficientes son iguales a cero. Se concluye que al menos uno (y probablemente todos) es distinto de cero.

La efectividad absoluta del modelo ajustado se evalúa por medio de tres medidas de *bondad del ajuste* basadas en el contraste de hipótesis mediante estadísticos. En las siguientes tablas (10.10 y 10.11) se resumen los resultados obtenidos con estas pruebas (entre paréntesis, la denominación que aparece en SPSS).

Junto con estas medidas, la prueba de bondad del ajuste de Hosmer y Lemeshow también permite afirmar que el modelo seleccionado ajusta bien los datos observados (no se puede rechazar la hipótesis nula de que el modelo ajusta bien los datos) (tabla 10.11).

La *interpretación* de los coeficientes de regresión obtenidos en el modelo ajustado es la siguiente:

— Tanto la variable $GRADNOVE_1$ como $IMASD_1$ son estadísticamente significativas en el modelo y presentan signo positivo. De este modo, y manteniendo constantes el resto de las variables, el valor estimado Exp (*B*) para $GRADNOVE_1$ indica que el «éxito en la introducción de un nuevo producto» es 3,4007 veces más probable que ocurra para aquellos «productos mejorados de otros ya existentes» que en el caso de que se trate de «productos total-

TABLA 10.10

*Bondad del ajuste**

Estadístico	Valor	*gl*	Sig.	H_0	Resultado
Desvianza inicial *(–2 log likelihood)*	211,908	155	0,0004	El modelo sólo con el término independiente ajusta bien los datos	Rechazar
Desvianza modelo *(–2 log likelihood)*	104,409	148	0,9974	El modelo con todas las variables incluidas ajusta bien los datos	No se puede rechazar
Test ji-cuadrado *(goodness of fit)*	123,971	148	0,9252	El modelo considerado ajusta bien los datos	No se puede rechazar

* Tanto los grados de libertad como el nivel de significación no son proporcionados por el programa en la versión utilizada.

TABLA 10.11

Prueba de bondad del ajuste de Hosmer-Lemeshow

Grupo	Éxito = 0		Éxito = 1		Total
	Observad	Esperad	Observad	Esperad	
1	15,000	15,746	1,000	0,254	16,000
2	15,000	14,588	1,000	1,412	16,000
3	12,000	12,868	4,000	3,132	16,000
4	12,000	9,518	4,000	6,482	16,000
5	5,000	6,049	11,000	9,951	16,000
6	3,000	3,674	14,000	13,326	17,000
7	3,000	1,876	13,000	14,124	16,000
8	0,000	0,620	16,000	15,380	16,000
9	0,000	0,590	16,000	15,941	16,000
10	0,000	0,000	11,000	11,000	11,000
			Ji-cuadrado	*gl*	Significación
Test C_0 de Hosmer-Lemeshow			6,1683	8	0,6284

mente nuevos». Es decir, que la ventaja de que la introducción de un producto sea un éxito cuando dicho producto es una mejora de uno ya existente es superior en 3,4007 veces a la correspondiente cuando el producto es totalmente nuevo. El impacto de la presencia de un departamento de I+D en la mejora de la probabilidad de que el nuevo producto sea un éxito es aún ma-

yor: en concreto es 35,6346 veces más probable que esto ocurra cuando existe un departamento de investigación y desarrollo que cuando no existe.

— La pertenencia de la empresa a sectores de baja intensidad tecnológica, $SECTECNG_1$, es también una variable significativa en el modelo, aunque en este caso su influencia sobre la variable dependiente es de signo negativo. Esto significa que el «éxito en la introducción de un nuevo producto» es 0,1179 veces *menos* probable que ocurra cuando la empresa pertenece a un sector de baja intensidad tecnológica en comparación con aquellas que pertenecen a sectores industriales tecnológicamente intensivos (categoría de referencia que no aparece explicitada en el modelo). Es decir, que la ventaja de la opción 1 de la variable dependiente («éxito en la introducción de un nuevo producto») cuando la empresa pertenece a industrias de baja intensidad tecnológica es inferior en 0,1179 veces a la correspondiente cuando se trata de una empresa inscrita en un sector de alta intensidad tecnológica.

— Las dos variables continuas, PUBLICID y PERSONAL, presentan un coeficiente estimado de signo positivo. Esto implica que, si permanecen constantes el resto de las variables, un incremento unitario de inversión en publicidad (10.000 euros) provocará un incremento multiplicativo por un factor 7,5562 de la ventaja de la opción «éxito en la introducción del nuevo producto». Asimismo, también se podría calcular el impacto que sobre la variable dependiente tendría un incremento de, por ejemplo $1{,}65 \times 10^4$ euros (desde 3 a $4{,}65 \times 10^4$ euros): Exp $(2{,}0224 \times 1{,}65) = 28{,}1335$, es decir, que la ventaja de la opción «éxito en la introducción de nuevos productos» es 28,1335 veces mayor cuando se gastan $4{,}65 \times 10^4$ euros que cuando se gastan 3×10^4 euros. Consideraciones similares se realizarían para la variable PERSONAL teniendo en cuenta que en este caso la unidad de medida es «una persona».

En relación con la *capacidad predictiva* del modelo, presenta una eficacia predictiva óptima, tal y como se desprende de los resultados que muestra la tabla de clasificación: un 85,26% de casos bien clasificados (tasa de aciertos); de los 65 casos en los que el nuevo producto «ha fracasado», el modelo predice correctamente el 80% (especificidad); de los 91 casos en los que el nuevo producto ha sido un «éxito», el modelo predice correctamente el 89,01% (sensibilidad):

TABLA 10.12

Clasificación para ÉXITO (el valor de corte es 0,50)

Valores observados	Valores pronosticados		Porcentaje de aciertos
	Fracaso (0)	Éxito (1)	
Fracaso (0)	52	13	80,00%
Éxito (1)	10	81	89,01%
Total			85,26%

El resto de medidas de la eficacia predictiva son fáciles de calcular:

— Tasa de errores: (10 + 13)/156 = 14,74%.

— Tasa de falsos negativos: 10/(52 + 10) = 16,13%.

— Tasa de falsos positivos: 13/(13 + 81) = 13,83%.

Asimismo, para comprobar que la tasa de aciertos del modelo es significativamente mayor que el número esperado de casos correctamente clasificados debido al azar (e), utilizamos el estadístico Z^*:

$$e = (1/156) \cdot (65^2 + 91^2) = 80{,}16$$

$$Z^* = \frac{(133 - 80{,}16)\sqrt{156}}{\sqrt{80{,}16(156 - 80{,}16)}} = 8{,}4644$$

cuyo resultado conduce a rechazar la hipótesis nula (para un nivel de significación del 5%) de que los casos correctamente clasificados por el modelo no difieren de los esperados sólo por efecto del azar.

El programa SPSS también proporciona un *histograma de probabilidades estimadas* en el que se puede apreciar cómo los casos estimados por el modelo (éxito y fracaso) se concentran en los extremos del gráfico, lo que viene a confirmar su eficacia predictiva:

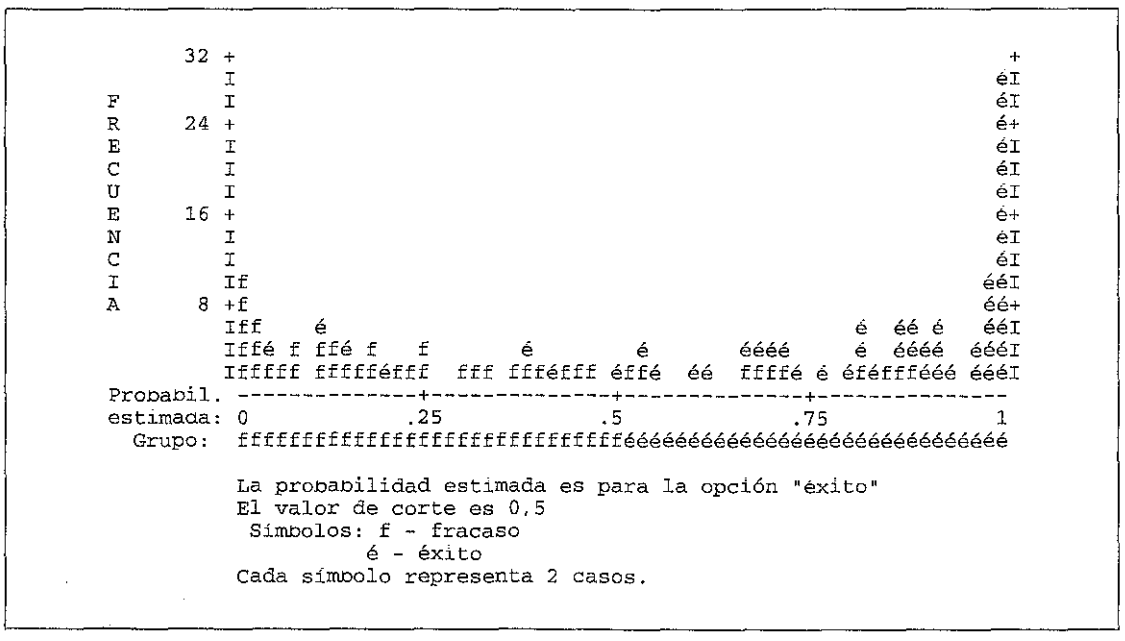

Figura 10.4. Histograma de probabilidades estimadas y grupos observados.

INVENTARIO DE TÉRMINOS Y CONCEPTOS

- Estimación por máxima verosimilitud.
- *Odds* o ventaja.
- Transformación logística o *logit.*
- Variable de diseño.
- Patrón de predictores.
- Estadístico *W* de Wald.
- R_A de Atkinson.
- Test *G* o prueba de la razón de verosimilitud.
- Desvianza.
- Prueba de Hosmer-Lemeshow.
- Tabla de clasificación.
- Histograma de probabilidades estimadas.
- Cociente de ventajas.
- Métodos de diagnóstico: valores residuales y medidas de influencia.
- Modelización estadística y modelización sustantiva.

BIBLIOGRAFÍA

Ato García, M., y López García, J. J. (1996): *Análisis estadístico para datos categóricos,* Síntesis Psicología, Madrid.

Casey, C., y Bartczak, N. (1985): «Using operating cash flow data to predict financial distress: some extensions», *Journal of Accounting Research,* vol. 23, núm. 1, pp. 384-385.

Christensen, R. (1990): *Log-lineal models,* Springer-Verlag, Nueva York.

Cramer, J. S. (1991): *The logit model: an introduction for economists,* Edward Arnold, Londres.

Hall, G. (1994): «Factors distinguishing survivors from failures amongst small firms in the UK construction sector», *Journal of Management Studies,* vol. 31, núm. 5, pp. 737-760.

Hosmer, D. W., y Lemeshow, S. (1989): *Applied logistic regresion,* John Wiley & Sons, Nueva York.

Jovell, A. J. (1995): *Análisis de regresión logística,* Centro de Investigaciones Sociológicas, colección «Cuadernos Metodológicos», núm. 15, Madrid.

López, J., Gandía, J. L., y Molina, R. (1998): «La suspensión de pagos en las Pymes: una aproximación empírica», *Revista Española de Financiación y Contabilidad,* vol. XXVII, núm. 94, pp. 71-97.

Mora, A. (1994): «Los modelos de predicción del fracaso empresarial: una aplicación empírica del logit», *Revista Española de Financiación y Contabilidad,* vol. XXIII, núm. 85, pp. 203-233.

Norusis, M. J. (1986): *Advanced statistics. SPSS/PC+,* SPSS Inc., Chicago.

Novales, A. (1993): *Econometría,* McGraw-Hill, Madrid.

Peña, D. (1994): *Estadística. Modelos y métodos,* vol. 2, Alianza, Madrid.

Rodríguez, M. A., y Frías, D. M. (1997): «Segmentación del mercado de las tarjetas de crédito», *Actualidad Financiera,* núm. 8, págs. 47-62.

Ruiz-Maya, L., et al. (1995): *Análisis estadístico de encuestas: datos cualitativos,* AC, Madrid.

Sánchez, G. (1997): *La innovación tecnológica y la pequeña y mediana empresa en Andalucía: un estudio empírico,* tesis doctoral, Departamento de Administración de Empresas y Marketing, Universidad de Granada.

Sharma, S. (1996): *Applied multivariate techniques,* John Wiley & Sons, Nueva York.