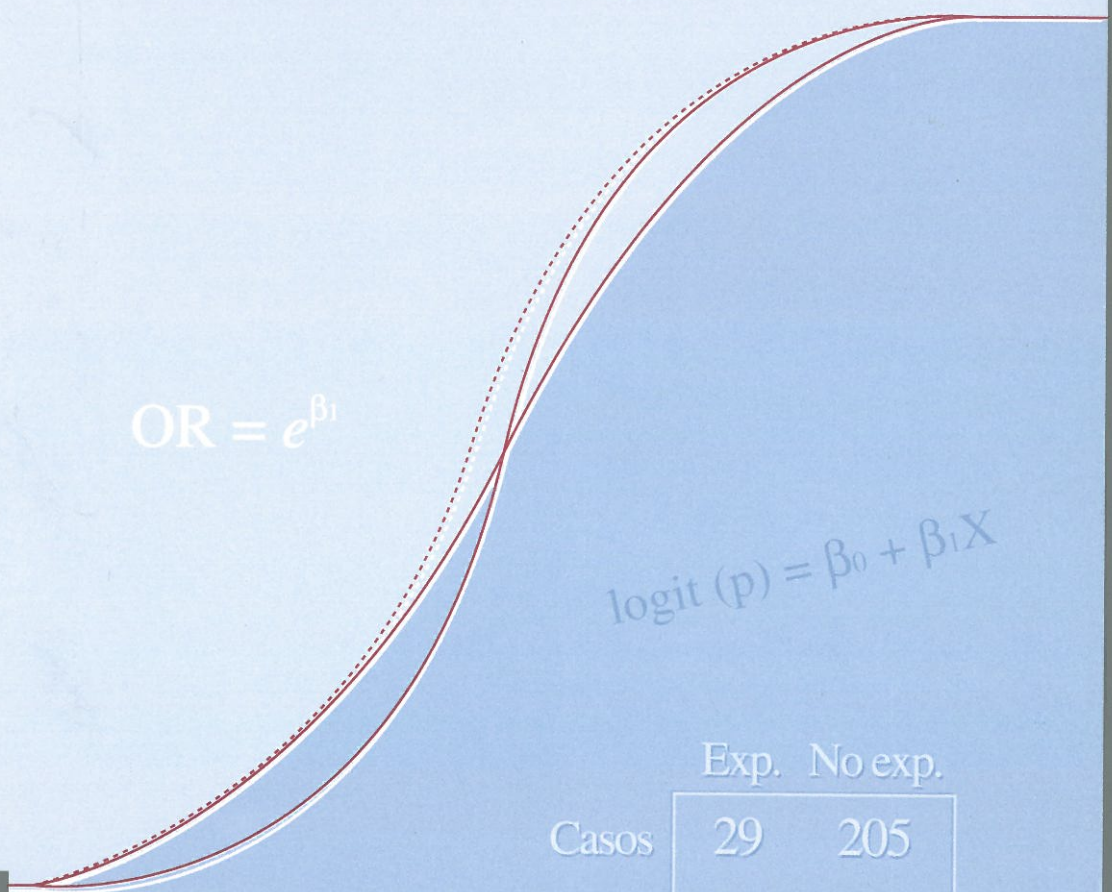
Escuela Andaluza de Salud Pública

Serie Monografías Nº 26

REGRESIÓN LOGÍSTICA EN SALUD PÚBLICA

Emilio Sánchez-Cantalejo Ramírez



JUNTA DE ANDALUCIA

REGRESIÓN LOGÍSTICA EN SALUD PÚBLICA

Emilio Sánchez-Cantalejo Ramírez
Escuela Andaluza de Salud Pública

2000

JUNTA DE ANDALUCIA

EASP
Escuela Andaluza
de Salud Pública

Catalogación por la Biblioteca de la EASP

SÁNCHEZ-CANTALEJO RAMÍREZ, Emilio
Regresión logística en salud pública/ Emilio Sánchez-Cantalejo Ramírez.-
Granada: Escuela Andaluza de Salud Pública, 2000.- (Monografías; 26)

1. Análisis de regresión. 2. Salud Pública. I. Título. II. Serie.
Library of Congress Classification QA 278

Edita:

ESCUELA ANDALUZA DE SALUD PÚBLICA
Campus Universitario de Cartuja. Apdo. de Correos 2070
18080 Granada España

ISBN: 84-87385-53-2

Depósito Legal: 1067-2000

Diseño de cubierta y maquetación: Miguel Salvatierra

Imprime: Gráficas Alhambra

A mi "Tabla 2x2", con un cero estructural

Esperanza	Julia
Ana	

Índice

Introducción 9

CAPÍTULO I
Regresión logística binaria simple 11

CAPÍTULO II
Regresión logística binaria múltiple 41

CAPÍTULO III
Diagnóstico en regresión logística 77

CAPÍTULO IV
Regresión logística policotómica y ordinal 123

Bibliografía 161

Anexo 165

Índice de materias 169

Introducción

En los últimos años, tanto la utilización de los métodos estadísticos como su sofisticación han sufrido un importante incremento en el campo de la investigación sanitaria. Esta relación entre la estadística y la investigación en salud ha sido un acicate para el desarrollo de la primera pues, ante nuevas preguntas de investigación, los estadísticos han tenido que proponer nuevos modelos para tratar de responderlas; un ejemplo de esta relación es el modelo de regresión logística para analizar los estudios de seguimiento y los de casos y controles. Actualmente, la regresión logística es el método multivariante más utilizado en el ámbito de la investigación sanitaria.

Esta monografía sobre regresión logística tiene su origen en las notas de clase de los cursos impartidos por el autor en los diez últimos años en la Escuela Andaluza de Salud Pública. Está pensada para un lector del campo de la salud, pero también puede ser de interés para investigadores y estudiantes de estadística y de ciencias sociales; por este motivo se ha tratado de hacer más énfasis en la interpretación de resultados que en la formulación de los distintos modelos. De cualquier forma, y para una compresión más rigurosa, ha sido inevitable alguna que otra fórmula que espero que no desanime demasiado al lector; los conocimientos de matemáticas necesarios para su lectura no van más allá del bachiller pero sí es necesaria la familiaridad con los conceptos básicos de estadística: descriptiva, probabilidad, intervalos de confianza y tests de hipótesis; algún conocimiento de la regresión lineal puede facilitar la lectura.

La monografía está estructurada en cuatro capítulos. En el primero se introduce el modelo de regresión logística binaria simple, tanto para el caso de una predictora categórica como para una continua. En el segundo capítulo se utiliza el modelo de regresión múltiple como mecanismo para controlar el efecto de confusión entre las predictoras así como para modelar las posibles interacciones; además se discuten las distintas estrategias de selección de variables y de comparación de modelos. En el tercer capítulo se presentan distintas técnicas para evaluar características que son desea-

bles que cumpla el modelo que se ajuste. Por último, el cuarto capítulo está dedicado a distintos modelos logísticos en donde la respuesta tiene más de dos categorías, distinguiendo entre respuestas nominales y ordinales. Los dos primeros capítulos constituyen el material para un curso básico de regresión logística; en una primera lectura los apartados 1.8 y 1.9 se pueden omitir sin pérdida de comprensión.

Evidentemente han quedado por tratar aspectos de la regresión logística muy interesantes que se han desarrollado en los últimos años desde diversas perspectivas; los estudios de cohorte con medidas repetidas han dado lugar a nuevos modelos de regresión logística que contemplan la dependencia entre los distintos valores de la respuesta en las sucesivas mediciones; la existencia ocasional de una estructura jerárquica entre las predictoras ha llevado a la construcción de modelos logísticos multinivel; también se ha propuesto el modelo de regresión logístico bajo la perspectiva de la estadística bayesiana, etc. En general, estas técnicas requieren una mayor sofisticación estadística que cae fuera del nivel de esta monografía.

Son muchas las personas que de una forma u otra tienen que ver con esta monografía y a las que quiero manifestar mi agradecimiento. Los alumnos, especialmente los Residentes de Medicina Preventiva y Salud Pública, por el interés que siempre mostraban en conocer algo que para algunos pertenecía al mundo de lo esotérico; ellos son, en gran parte, el porqué de esta historia. Emilio Perea, Miguel Angel Martínez, Jordi Sunyer, Mª Victoria Zunzunegui y Rosa Miñarro me ayudaron mucho con sus críticas y aportaciones. Quiero agradecer especialmente a Ricardo Ocaña su contribución y el apoyo que me prestó con el programa S-Plus; Carmen Martínez fue un estímulo en los momentos de distracción. Tan solo me queda por reclamar exclusivamente para mí la responsabilidad de cualquier error en lo que sigue.

CAPÍTULO 1

REGRESIÓN LOGÍSTICA BINARIA SIMPLE

En este capítulo, tras mostrar la necesidad de introducir modelos de regresión para el estudio de variables dicotómicas, definiremos el modelo de regresión logística con una sola predictora; veremos cómo interpretar y estimar sus coeficientes, la adecuación del modelo a los distintos estudios epidemiológicos y, por último, introduciremos algunos conceptos de gran utilidad para el resto de la monografía.

1.1 Introducción

Tanto en Salud Pública como en el ámbito de la Medicina Clínica, es frecuente la situación en que se dispone de una variable resultado con sólo dos posibles valores, es decir, una variable dicotómica. Tras una campaña publicitaria diseñada para cambiar el hábito de fumar en una determinada población, la variable resultado en cada individuo se puede medir en base a que haya o no abandonado tal hábito; para un clínico puede ser de interés estudiar qué factores pueden estar asociados al hecho de que un individuo presente o no rechazo al trasplantársele un riñón; ¿qué factores pueden estar asociados a la presencia o ausencia de un peso anormalmente bajo en los recién nacidos? Estos y otros muchos ejemplos ponen de manifiesto las múltiples situaciones posibles en que un investigador de la salud puede estar interesado en estudiar relaciones entre variables en donde la variable resultado es binaria, es decir, sólo puede tomar dos valores: rechazo sí o rechazo no; bajo peso sí o bajo peso no, etc.

Una aproximación al estudio de tales relaciones consiste en hacer uso de la distribución chi-cuadrado para el análisis de la tabla de contigencia generada; ya que la variable resultado es dicotómica, la tabla correspondiente será del tipo 2xc, siendo c el número de categorías de la variable predictora o independiente, la variable que queremos ver si está o no relacionada con la variable resultado. Si la predictora es propiamente categórica, se utilizarán sus categorías sin más y si fuese cuantitativa tendríamos que categorizarla, con la pérdida de información que ello conlleva. De cualquier modo, formada la tabla, mediante un test de la chi-cuadrado podemos contrastar la hipótesis de independencia de la variable resultado con la predictora.

Sin embargo, la situación más frecuente no es la de disponer de una sola variable predictora; en general, para cada individuo objeto de estudio se suelen medir más variables de las que algunas podrían afectar a la variable resultado. Siguiendo con la estrategia antes expuesta, se trataría de categorizar las predictoras y formar una tabla

de contingencia de tantas dimensiones como sea el número de predictoras más uno. Esta manera de enfocar el problema, que en el ámbito de la epidemiología se conoce como *estratificación,* presenta un grave problema: cuando el número de predictoras es superior a 4 ó 5, caso frecuente por otra parte, el número de tablas o *estratos* a formar crece tanto que la mayoría de las casillas de las subtablas generadas suelen estar vacías por lo que las estimaciones que se hagan a partir de estas subtablas serán poco precisas y los tests basados en la distribución chi-cuadrado no son aplicables.

Consideremos, por ejemplo, que se está interesado en estudiar la posible relación entre el bajo peso al nacer y el hábito de fumar de la madre; ya que puede haber otras predictoras que perturben la relación entre las dos variables sería de interés controlar su efecto. Ejemplos de posibles variables que pueden alterar tal asociación son: el consumo de alcohol, el nivel socioeconómico, la edad, etc. Aunque tan sólo se considerasen estas cuatro variables y, a pesar de los inconvenientes antes señalados, las convirtiésemos en dicotómicas, tendríamos que formar (2)(2)(2)(2)=16 estratos, es decir, 16 tablas 2x2 en donde estudiar la relación bajo peso y hábito de fumar de la madre; en términos más realistas, si consideramos 3 categorías para la paridad, 4 para el nivel de consumo de alcohol, 6 para el nivel socioeconómico y 5 para la edad, el número de estratos a formar sería (3)(4)(6)(5) = 360. A no ser que se dispusiese de un tamaño de muestra muy grande, las estimaciones que hiciésemos a partir de tales tablas serían muy imprecisas.

En resumen, si seguimos esta aproximación al problema, a la pérdida de información derivada de la categorización de las variables predictoras, habría que sumar la poca precisión de las estimaciones de los riesgos en los distintos estratos. Estos argumentos justifican la búsqueda de otras aproximaciones al problema planteado.

1.2 Modelos de regresión

Bajo este nombre se engloban una serie de modelos estadísticos que tratan de cuantificar la asociación o dependencia entre una variable resultado y una o varias variables predictoras; es costumbre representar por la letra Y a la variable resultado o variable respuesta y por X a la variable predictora o covariable; en caso de disponer de un número m de predictoras las representaremos por los símbolos X_1, X_2, ..., X_m. Pues bien, mientras que para ningún modelo de regresión existen restricciones acerca de la naturaleza de las predictoras, la de la variable resultado es la que condiciona el tipo de modelo de regresión.

En orden cronológico, el primero que fué propuesto y, por otra parte, el mejor conocido, es el llamado *modelo de regresión lineal*; este tiene como objetivo estudiar las relaciones entre una variable resultado continua, por ejemplo, la presión sistólica

(PS), y un conjunto de predictoras; para el caso univariante, el de una sola predictora, por ejemplo, el índice de masa corporal (IMC), el modelo establece que

$$PS = \beta_0 + \beta_1 IMC + e$$

donde β_0 y β_1 son los llamados *parámetros o coeficientes del modelo*; e es un término aleatorio distribuido según una normal de media cero y varianza constante. Es decir, la presión de un individuo es una constante β_0, más el producto de otra β_1 por su índice de masa corporal, más un término aleatorio que cambia de unos individuos a otros. Una manera alternativa de formular ese modelo es mediante la expresión

$$E(PS) = \beta_0 + \beta_1 IMC$$

donde $E(PS)$ representa la media de la presión.

Escrito de esta forma podemos entender que de lo que se trata realmente es de estudiar la relación entre el valor medio de la respuesta y la predictora; para ello se propuso este modelo tan sencillo, pues establece que tal relación se puede representar gráficamente mediante una recta, la curva más simple.

Como quiera que la presión depende, aparte del IMC, de otras características individules, la incorporación de éstas al modelo puede ayudarnos a entender, al menos parcialmente, la variabilidad de las presiones entre los distintos individuos. Entonces en el caso en que disponemos de varias predictoras, el modelo es

$$E(PS) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m$$

donde X_1, X_2, ..., X_m (IMC, edad, género, consumo de alcohol, colesterol, etc.) son las m predictoras que utilizamos para explicarnos los cambios del valor medio de la variable resultado.

Aunque como se dijo anteriormente, este modelo es muy bien conocido desde hace tiempo -Galton ya lo utilizó en el estudio del componente genético implicado en la altura de los individuos-, el modelo lineal presenta dificultades prácticamente insalvables si la variable respuesta es categórica; por esta razón se han propuesto otros modelos para esta nueva situación, de los que los *modelos de regresión logística* son el objeto de esta monografía.

Otro tipo de variable respuesta que se presenta con frecuencia en la investigación sanitaria es el tiempo que transcurre desde un momento determinado hasta la ocurrencia de un fenómeno de interés; por ejemplo, el tiempo que pasa desde el diagnóstico de un cáncer de mama hasta la muerte de la paciente. Al tratarse del tiempo,

podría pensarse que podríamos considerarla como una variable continua y, por tanto, utilizar el modelo de regresión lineal; sin embargo, la novedad que presentan este tipo de estudios es que para algunos individuos no va a ser posible medir exactamente la variable respuesta. Imaginemos un estudio sobre supervivencia de cáncer de mama que comenzó en 1986; si decidimos como periodo de estudio el decenio 1986-1995, afortunadamente no todas las mujeres diagnosticadas en ese periodo habrán fallecido al finalizar del estudio. Por tanto de las mujeres que no han fallecido no conocemos el tiempo de supervivencia; así, de una mujer diagnosticada a principios de 1991 y que no haya fallecido al finalizar 1995 no podemos decir que ha sobrevivido 5 años, sino que al menos ha sobrevivido 5 años; de estos tiempos de supervivencia se dice que están censurados. Esta característica pone en evidencia la necesidad de utilizar nuevos modelos que puedan tratar con este tipo de variable respuesta; son los llamados *modelos de regresión de supervivencia* de los que el modelo de Cox (1972) es el más utilizado en la literatura sanitaria.

1.3 La distribución binomial

Consideremos una población en la que los individuos pueden ser clasificados en función de la presentación o no de una determinada característica, por ejemplo, en una población de recién nacidos la característica en cuestión puede ser el tener bajo peso al nacer, definido como pesar menos de 2.500 gramos. Vamos a definir una variable Y tal que $Y=1$ en los niños con bajo peso, mientras que en los niños sin bajo peso $Y=0$; a una variable de este tipo se conoce como variable de Bernoulli. Si en la población de interés el 7% de los recién nacidos son de bajo peso, podemos establecer que la probabilidad p de que al elegir al azar un recién nacido éste sea de bajo peso es de 0,07; escrito formalmente $p=P(Y=1)=0{,}07$ y, por tanto, $1-p=P(Y=0)=0{,}93$. Estas dos igualdades describen la función de probabilidad de la variable dicotómica Y.

Supongamos ahora que tomamos dos niños al azar y definimos la variable Y cuyos valores son el número de niños, de entre los dos elegidos, que tienen bajo peso; evidentemente esta variable así definida puede tomar tres posibles valores: $Y=0$ si ninguno de los dos niños es de bajo peso, $Y=1$ si un solo niño tiene bajo peso y, por último, $Y=2$ en caso de que los dos lo tengan. Veamos como definir su función de probabilidad, es decir, la regla que nos permite calcular las probabilidades de que ocurra cada una de las tres posibles situaciones.

Si aceptamos que la presencia de la característica de interés, el bajo peso en un niño, no influye en la presencia de ella en otro niño de madre distinta, es decir, estamos hablando de sucesos independientes, la ley multiplicativa de las probabilidades nos permite decir que la probabilidad de que ninguno de los niños sea de bajo peso es igual a $(1-p)(1-p)=(1-p)^2$; por tanto $P(Y=0)=(1-p)^2$. La probabilidad de que el primer niño

sea de bajo peso y el segundo no, es $p(1-p)$, y la probabilidad de que primero no lo sea y el segundo sí es $(1-p)p$; por tanto, la probabilidad de que uno de ellos tenga bajo peso será $P(Y=1)=2p(1-p)$. Por último, la probabilidad de que los dos sean de bajo peso será $p.p=p^2$, es decir, $P(Y=1)=p^2$. Así, $0{,}93^2 = 0{,}865$, $2(0{,}07)(0{,}93)=0{,}130$ y $0{,}07^2=0{,}005$ son las probabilidades de que al tomar una muestra de dos niños ninguno, uno o los dos, respectivamente, sean de bajo peso.

Estas dos situaciones en que hemos utilizado muestras de tamaños 1 y 2, respectivamente, se pueden generalizar a cualquier tamaño de muestra. Si elegimos una muestra de n recién nacidos, podemos definir la variable Y cuyos posibles valores son el número de niños con bajo peso; evidentemente esos valores son todos los números enteros comprendidos entre cero y n. Pues bien, se puede demostrar que la probabilidad de que en una muestra de n niños haya r de ellos con bajo peso se puede calcular mediante la expresión

$$P(Y=r)=\binom{n}{r}p^r(1-p)^{(n-r)}$$

donde

$$\binom{n}{r} = \frac{n!}{r!\ (n-r)!}$$

y $n!$ es un símbolo que representa el producto de n por todos los números enteros positivos inferiores a él. La expresión anterior que permite calcular estas probabilidades es la llamada función de probabilidad binomial y de la variable Y se dice que sigue la distribución binomial. Ya que tal función nos permite calcular cualquier probabilidad con sólo conocer n y p se dice que estos son los parámetros de la distribución binomial.

Imaginemos que hemos tomado una muestra de 20 recién nacidos, ¿cuál es la probabilidad de que en esa muestra haya 2 niños con bajo peso? En este caso, $n=20$ y $r=2$ por lo que

$$P(Y=2)=\binom{20}{2}0{,}07^2(1-0{,}07)^{(20-2)}=\frac{20!}{2!\ (20-2)!}\ 0{,}07^2\ 0{,}93^{18}=$$

$$=\frac{(20)(19)\ldots(2)(1)}{(2)(1)(18)(17)\ldots(2)(1)}\ 0{,}07^2\ 0{,}93^{18}=0{,}25$$

lo que se puede interpretar diciendo que un 25% de las muestras de 20 niños que tomemos tendrán dos recién nacidos de bajo peso.

Si nos dedicaramos a tomar muchas muestras, evidentemente unas tendrían un número de niños de bajo peso y otras otros pero ¿cuál sería el número medio por

muestra y cómo de distintos serían tales números? En otras palabras, ¿cual es la media y varianza de una variable binomial de parámetros n, p? Se puede demostrar que la media y la varianza vienen dadas por las expresiones $n \cdot p$ y $n \cdot p \cdot (1-p)$, respectivamente. Por tanto, para muestras de 20 recién nacidos, el número medio de casos de bajo peso por muestra es $20(0,07)=1,4$ y $20(0,07)(0,93)=1,3$ es una medida de la variabilidad del número de casos de bajo peso entre las distintas muestras.

1.4 La transformación logit

Consideremos ahora la situación en que la variable Y es dicotómica tal que $Y=1$ si el individuo presenta la característica de interés e $Y=0$ en caso contrario. Por tanto se puede admitir que la variable respuesta sigue una distribución binomial de parámetros 1 y p, donde p representa la probabilidad de que un individuo presente la característica de interés, es decir, la probabilidad de que Y tome el valor 1; esta situación teórica aparece en el caso en que para cada individuo consideramos la variable Y. Como para una variable binomial de parámetros n y p su media es el producto $n \cdot p$, en este caso la media será $1 \cdot p = p$. De esta forma el modelo de regresión lineal se podría escribir de la forma

$$E(Y) = p = \beta_0 + \beta_1 X$$

Ya que no existen restricciones sobre los valores de los parámetros del modelo, es posible que las estimaciones de los parámetros sean tales que la suma

$$\hat{\beta}_0 + \hat{\beta}_1 X$$

sea, en alguna ocasión, o superior a la unidad o inferior a cero; en cualquiera de esas dos situaciones nos encontramos con un problema muy serio: podemos tener estimaciones de la probabilidad de presentar la característica o bien mayores que la unidad o bien negativas, lo cual carece totalmente de sentido; esta situación no causaba problemas en regresión lineal pues al ser Y normal condicionada a los valores de la variable X, cualquier valor real puede ser lícito para la media de Y. Este argumento invalida cualquier intento de utilizar la metodología de la regresión lineal para el caso de que Y sea binomial; además, la hipótesis de varianza constante tampoco se mantiene en esta nueva situación.

Una medida muy utilizada tanto en ciencias sociales como en epidemiología es la que denominaremos *ventaja* u *oportunidad*, una de las varias traducciones propuestas para la palabra *odds* de los anglosajones. Sea p la probabilidad definida anteriormente; el cociente

$$\frac{p}{1-p}$$

es decir, la probabilidad de presentar la característica dividida por la probabilidad de no presentarla, se denomina ventaja de la citada característica; que una ventaja de una determinada característica sea 5 significa que, en la población correspondiente, es 5 veces más probable presentar la característica que no presentarla. Los posibles valores de una ventaja pueden ser cualquier número positivo; en efecto, no puede ser negativa porque tanto p como $1\text{-}p$ son siempre como poco cero, es decir, valores no negativos; por otra parte, si la característica de interés es poco frecuente, es decir, p es próxima a cero, así lo será también la ventaja y si la característica es muy frecuente, es decir, p es próxima a uno, la ventaja será muy grande.

Si consideramos la transformación, mediante el logaritmo neperiano, de este parámetro

$$\log \frac{p}{1\text{-}p}$$

sus posibles valores pueden ser cualquier número real, tanto positivo como negativo, con lo que desaparece el problema antes comentado; esta transformación de p se denomina la *transformación logística* o *transformación logit* de la probabilidad p

$$\log \frac{p}{1\text{-}p} = logit(p)$$

De esta manera el modelo que nos puede permitir, en principio, resolver el problema que tenemos planteado puede representarse en la forma

$$logit(p) = \log \frac{p}{1\text{-}p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m$$

Este modelo, aunque fue propuesto a principios de los años sesenta, no alcanzó sin embargo su mayor popularidad hasta mediados de los setenta en que se generalizó el uso de los ordenadores pues, como luego veremos, las estimaciones de los parámetros conllevan una complejidad de cálculo que es prácticamente insalvable sin el auxilio de las máquinas. Este modelo se planteó en el ámbito del análisis de los estudios de cohorte también se puede aplicar a otros diseños epidemiológicos como los estudios de casos y controles y los transversales. El gran atractivo de este modelo está en que sus parámetros son interpretables como una medida de riesgo asociado a las predictoras, como se verá más adelante.

1.5 El modelo logístico binario simple

Consideremos por ahora la situación en que sólo se dispone de una variable predictora, dicho en términos epidemiológicos, un sólo posible *factor de riesgo*. El moti-

vo de considerar este caso particular se debe más que nada a su más fácil comprensión que el caso con varias predictoras. Según el modelo logístico antes propuesto, para el caso de una sola predictora, toma la forma

$$logit(p) = \log \frac{p}{1-p} = \beta_0 + \beta_1 X$$

donde p representa la probabilidad de que un individuo presente la característica de interés y X es la única predictora. La expresión anterior es equivalente a esta otra

$$\frac{p}{1-p} = e^{\beta_0 + \beta_1 X}$$

y despejando p obtenemos otra forma de escribir el modelo logístico

$$E(Y) = p = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

Si consideramos el caso en que la variable X es cuantitativa y representamos en un sistema de dos ejes los valores de p en función de los valores de X, podremos comprobar que la representación gráfica del modelo es como aparece en la Figura 1.1,

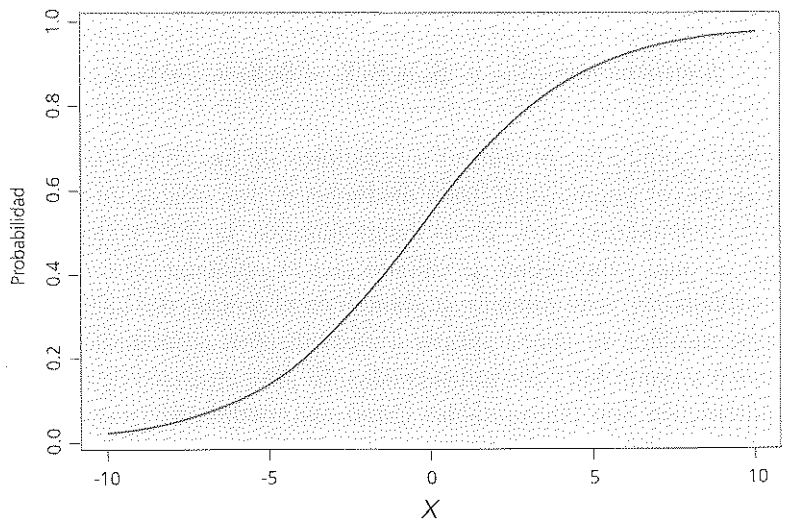


Figura 1.1. *Función logística con* $\beta_1 > 0$.

o bien en la Figura 1.2, dependiendo de que el valor del parámetro β_1 sea positivo o negativo, respectivamente. Es decir, el modelo logístico postula que el cambio de p, en función de X, es lento al principio, luego se hace más rápido creciendo prácticamente como si fuese una línea recta y, al final, vuelve a enlentecerse.

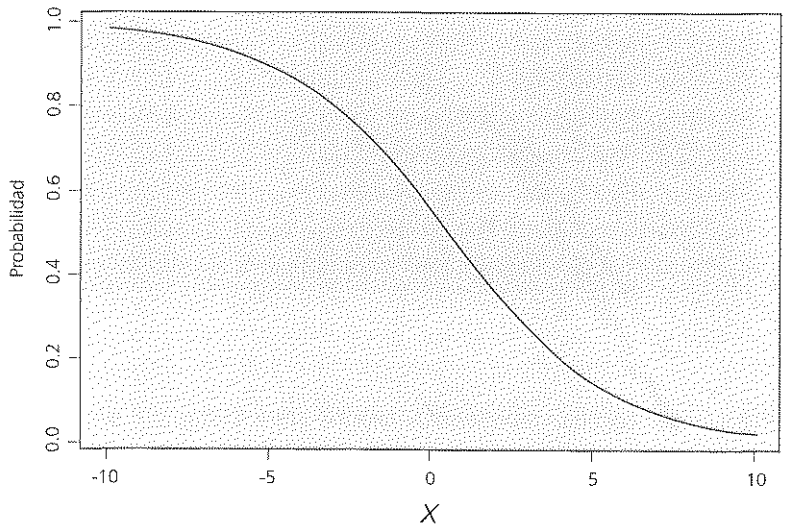


Figura 1.2. *Función logística con* $\beta_1 < 0$.

Obsérvese la diferencia con el modelo lineal que suponía una velocidad de cambio constante a lo largo del rango de los valores de X. Lo que sí cambia de forma lineal es el *logit(p)* como aparece en la Figura 1.3.

También la última formulación del modelo logístico nos permite comprobar fácilmente que el valor p, la media de la variable resultado, se mantiene dentro de los valores permitidos para una probabilidad; en efecto, ya que

$$E(Y) = p = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

estamos expresando la probabilidad como un cociente entre dos cantidades no negativas, pues sean cuales sean los valores de los parámetros, la cantidad $e^{\beta_0 + \beta_1 X}$ no puede ser negativa; como, además, el numerador es siempre menor o igual que el denominador, el cociente va a variar entre 0 y 1, el rango de valores permitido para una probabilidad.

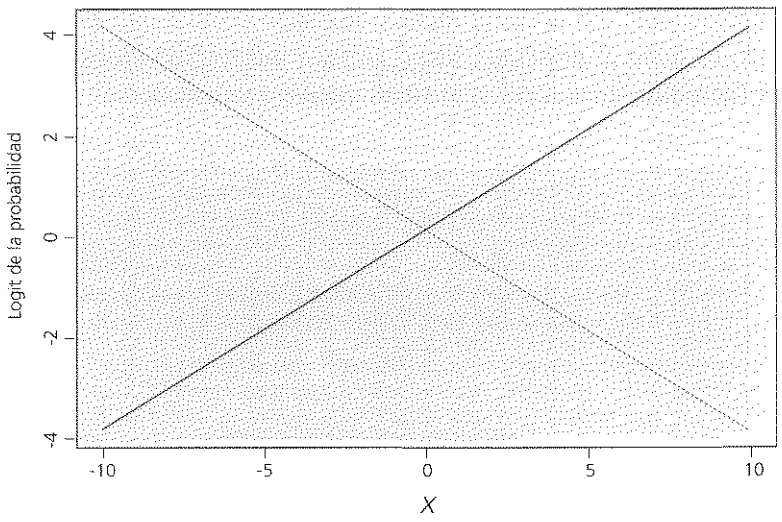


Figura 1.3. *La línea continua representa la relación entre logit(p) y X para* $\beta_1 > 0$*; la línea rayada es para el caso* $\beta_1 < 0$.

1.6 Interpretación de los parámetros

Antes se avanzó que una de las ventajas de la utilización del modelo de regresión logística era la facilidad de la interpretación de sus coeficientes. En efecto, consideremos dos sujetos con valores x_1 y x_2 de la variable X; conocidos tales valores y considerando el modelo

$$\log \frac{p}{1\text{-}p} = \beta_0 + \beta_1 X$$

tendremos que para el primer individuo

$$\log \frac{p_1}{1\text{-}p_1} = \beta_0 + \beta_1 x_1$$

y, para el segundo

$$\log \frac{p_2}{1\text{-}p_2} = \beta_0 + \beta_1 x_2$$

donde p_1 es la probabilidad de que un individuo con valor x_1 de la variable X presente la característica de interés; p_2 es lo análogo para un individuo con valor x_2 de la variable X. Restando estas dos igualdades tenemos

$$\log \frac{p_1}{1-p_1} - \log \frac{p_2}{1-p_2} = (\beta_0 + \beta_1 x_1) - (\beta_0 + \beta_1 x_2) = \beta_1 x_1 - \beta_1 x_2 = \beta_1 (x_1 - x_2)$$

y ya que la diferencia de dos logaritmos es igual al logaritmo del cociente, en definitiva tenemos la siguiente expresión,

$$\log \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}} = \beta_1 (x_1 - x_2) \tag{1.1}$$

El cociente

$$\frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}} = \frac{p_1 (1-p_2)}{p_2 (1-p_1)} = (OR)$$

es una razón entre dos ventajas por lo que se conoce con el nombre de *razón de ventajas* que representaremos mediante el símbolo *OR* (*odds ratio*).

La expresión anterior se puede escribir de esta otra forma

$$\log(OR) = \beta_1 (x_1 - x_2)$$

o lo que es igual

$$OR = e^{\beta_1 (x_1 - x_2)}$$

que podemos traducir diciendo que la razón de ventajas entre dos individuos con valores x_1 y x_2 de la predictora se puede conseguir elevando el número e al producto $\beta_1 (x_1 - x_2)$; considerando el caso particular en que $x_1 = x_2 + 1$, es decir, si los individuos se diferencian sólo en una unidad en términos de la variable X, tendremos

$$\log(OR) = \beta_1 \{ x_1 - (x_1 - 1) \} = \beta_1$$

que es igual a decir que β_1 se puede interpretar como el logaritmo de la razón de ventajas de presentar la característica entre dos individuos que se diferencian en una unidad respecto a la predictora; la última expresión también puede adoptar esta nueva forma

$$OR = e^{\beta_1}$$

por lo que β_1 es un valor que cumple la siguiente propiedad: elevando el número e a β_1, el resultado no es más que la *OR* entre dos individuos que se diferencia en una unidad en términos de la predictora.

El hecho de que β_1 sea cero supone que

$$E(Y)= p = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} = \frac{e^{\beta_0+0X}}{1+e^{\beta_0+0X}} = \frac{e^{\beta_0}}{1+e^{\beta_0}}$$

cantidad que no depende de la predictora X; es decir, que β_1 sea cero implica que p es una constante, no depende de X; de otra manera, que β_1 sea cero significa la independencia entre la variable resultado y la predictora en cuestión. Otra forma de establecer la independencia es apoyándose en el hecho de que si $\beta_1=0$, entonces

$$OR = e^{\beta_1} = e^0 = 1$$

lo que significa que las dos variables son independientes si la *OR* es la unidad. Como en todos los modelos de regresión, el valor de las estimaciones de los coeficientes correspondientes a las predictoras dependen de las unidades de medida que se utilicen para éstas, por lo que su interpretación no estará completa sin una referencia a las unidades de medida elegidas; esta cuestión se discutirá más adelante.

1.7 Interpretación de la *OR*

La situación más simple para el cáculo e interpretación de la *OR* es el caso de una sola predictora dicotómica. En la Tabla 1.1 aparecen los resultados de dos cohortes de individuos, una de 470 hombres con unos niveles de 240 mgrs. de colesterol y otra de 254 con 210 mgrs. de colesterol; tras un periodo de seguimiento, en la primera cohorte hubo 31 casos de enfermedad coronaria y 8 en la segunda.

Tabla 1.1. *Nivel de colesterol y enfermedad coronaria.*

		ENFERMEDAD CORONARIA		
		Sí	No	Total
NIVEL DE COLESTEROL	240 mgr.	31	439	470
	210 mgr.	8	246	254
	Total	39	685	724

De la Tabla 1.1 podemos estimar la ventaja de presentar infarto en cada una de las cohortes; así, para la primera

$$\frac{\hat{p}_1}{1-\hat{p}_1} = \frac{31/470}{439/470} = \frac{0{,}0660}{0{,}9340} = 0{,}0707$$

y, para la segunda,

$$\frac{\hat{p}_2}{1-\hat{p}_2} = \frac{8/254}{246/254} = \frac{0{,}0315}{0{,}9685} = 0{,}0325$$

por tanto, una estimación de la razón de ventajas de padecer infarto entre los individuos con 240 mgrs. de colesterol respecto a los que tienen 210 es

$$\frac{0{,}0707}{0{,}0325} = 2{,}17$$

En el apartado anterior se dijo que la razón de ventajas se puede escribir como el cociente

$$OR = \frac{p_1(1-p_2)}{p_2(1-p_1)}$$

por lo que, dispuestos los datos de las dos cohortes en forma de tabla 2x2, la razón de ventajas se puede estimar así

$$OR = \frac{p_1(1-p_2)}{p_2(1-p_1)} = \frac{(31/470)\ (246/254)}{(8/254)\ (439/470)} = \frac{(31)\ (246)}{(8)\ (439)} = 2{,}17$$

como resultaba antes. De esta manera, la razón de ventajas se estima mediante el cociente de los productos de los dos números que aparecen en las dos diagonales de la tabla 2x2, motivo por el que también se denomina a esta medida *razón del producto cruzado.*

Sin embargo, en los estudios de cohorte la medida de asociación más natural es el llamado *riesgo relativo* definido como el cociente de las incidencias de la enfermedad entre los individuos expuestos y los no expuestos; si a los individuos con 240 mgr los consideramos como expuestos y a los de 210 mgr. como no expuestos, una estimación del riesgo relativo para este estudio será

$$\frac{(31/470)}{(8/254)} = 2{,}09$$

valor cercano a 2,17 con el que estimamos a la razón de ventajas.

Pero, ¿qué significa que la estimación del riesgo para el estudio de la Tabla 1.1 sea 2,17?; el valor 2,17 nos da una idea acerca del número de veces al que está a más riesgo de sufrir enfermedad coronaria, durante el periodo de estudio, un individuo con 240 mgr. de colesterol que otro individuo con 210 mgr.

En el siguiente apartado se demuestra como la regresión logística es aplicable a los estudios de casos y controles manteniéndose la interpretación del parámetro de la predictora como logaritmo de la razón de ventajas; sin embargo, el término β_0 pierde su significado a no ser que se conozcan las fracciones de muestreo entre los casos y entre los controles.

1.8 El modelo logístico en distintos tipos de muestreo *

Hasta ahora se ha visto la interpretación del coeficiente β_1 que acompaña a la variable predictora, pero ¿cómo interpretar el término independiente β_0 que aparece en el modelo? Representemos por Y la presencia, Y=1, o ausencia, Y=0, de una determinada enfermedad; si X es un factor de riesgo binario, de tal forma que X=1 si el individuo estuvo expuesto a tal factor y X=0 en caso contrario, para un individuo no expuesto, el modelo logístico queda en la forma

$$logit(p_0) = \beta_0 + \beta_1 0 = \beta_0$$

donde p_0 es la probabilidad de enfermar, en el periodo de seguimiento, en los que no están expuestos; por tanto, p_0 es la incidencia de la enfermedad entre los no expuestos. En resumen, β_0 se puede interpretar como el logit de la probabilidad de enfermar entre los no expuestos.

El diseño natural para el modelo logístico que se acaba de presentar es el estudio de cohortes; en efecto, en este tipo de estudio lo que se conoce o fija de antemano es el nivel de exposición y entonces, prospectivamente tras un determinado periodo de seguimiento, se mide el estado de salud, por lo que tiene sentido preguntarse por la probabilidad de que, tras el periodo de seguimiento, un individuo expuesto o no expuesto, desarrolle la enfermedad; dicho más formalmente, en los estudios de seguimiento tiene sentido el cálculo de la probabilidad condicionada

$$p = P(Y=1/X)$$

donde $P(Y=1/X)$ indica la probabilidad, condicionada al valor de X, de que el individuo enferme. Sin embargo, no es la mayor razón de la popularidad del modelo logístico su aplicabilidad a los estudios de cohortes sino su uso en los estudios de casos y controles; en este diseño, el esquema de muestreo es distinto, pues se elige a los sujetos de estudio según los niveles de la variable resultado, los casos o enfermos y los controles o sanos, midiendo retrospectivamente la variable de exposición. Afortunadamente, el modelo logístico puede adaptarse perfectamente al estudio de casos y controles manteniendo el coeficiente β_1 el mismo significado pero con diferencias en cuanto a la interpretación del término independiente, (Breslow (1980)).

* Esta sección puede omitirse en una primera lectura

En efecto, si notamos por Z la variable que toma el valor 1 si el individuo ha sido elegido para entrar en el estudio de casos y controles y 0 en otro caso, sean

$$\pi_1 = P(Z=1 \mid Y=1) \quad \text{y} \quad \pi_0 = P(Z=1 \mid Y=0)$$

las probabilidades de ser elegidos como caso y como control, respectivamente, es decir, las fracciones de muestreo entre los casos y controles. Haciendo uso del teorema de Bayes se puede escribir que la probabilidad de que un individuo elegido con un valor determinado de la predictora, sea un caso es

$$P(Y=1 \mid Z=1, X) = \frac{P(Z=1 \mid Y=1, X)\, P(Y=1 \mid X)}{P(Z=1 \mid Y=1, X)\, P(Y=1 \mid X) + P(Z=1 \mid Y=0, X)\, P(Y=0 \mid X)}$$

donde $P(Y=1 \mid X)$ y $P(Y=0 \mid X)$ son las probabilidades de presentar y no presentar la característica, respectivamente, condicionadas al valor de la variable predictora X; es decir, si

$$P(Y=1 \mid X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

entonces,

$$P(Y=0 \mid X) = 1 - P(Y=1 \mid X) = \frac{1}{1 + e^{\beta_0+\beta_1 X}}$$

sustituyendo estos valores en la expresión anterior y después de alguna manipulación algebráica se llega a esta otra igualdad

$$P(Y=1 \mid Z=1, X) = \frac{\pi_1\, e^{\beta_0+\beta_1 X}}{\pi_1\, e^{\beta_0+\beta_1 X} + \pi_0}$$

y dividiendo numerador y denominador por π_0 se llega a que

$$P(Y=1 \mid Z=1, X) = \frac{\frac{\pi_1}{\pi_0}\, e^{\beta_0+\beta_1 X}}{\frac{\pi_1}{\pi_0}\, e^{\beta_0+\beta_1 X} + 1}$$

o lo que es igual

$$P(Y=1 \mid Z=1, X) = \frac{e^{[\log(\frac{\pi_1}{\pi_0})+\beta_0]+\beta_1 X}}{e^{[\log(\frac{\pi_1}{\pi_0})+\beta_0]+\beta_1 X} + 1}$$

que es la misma expresión conocida del modelo logístico salvo que el término independiente pasa a ser ahora

$$\beta_0^* = \log\left(\frac{\pi_1}{\pi_0}\right) + \beta_0$$

donde π_1 y π_0 son las fracciones de muestreo para los casos y los controles respectivamente. Ya que casi nunca se conocen tales fracciones de muestreo, a partir de un estudio de casos y controles no se puede conocer la incidencia de la enfermedad entre los no expuestos. Es de advertir que la probabilidad de que un caso sea elegido como caso condicionada a un valor de X, es decir, $P(Z=1 \mid Y=1, X)$, se ha sustituido por $P(Z=1 \mid Y=1)$, lo que significa que estamos asumiendo que las probabilidades de muestreo sólo dependen de la variable resultado Y y no de la variable predictora; dicho más claramente, los casos y controles se deben elegir independientemente de su situación en cuanto a la exposición.

Sin embargo, es fácil comprobar que β_1 sigue teniendo la misma interpretación que en los estudios de cohorte; en efecto, ya que

$$P(Y=1 \mid Z=1, X) = \frac{e^{\beta_0^* + \beta_1 X}}{1 + e^{\beta_0^* + \beta_1 X}}$$

que no es más que la probabilidad de que un individuo del estudio, con un determinado valor de la predictora, tenga la enfermedad, podemos decir que para un individuo expuesto, $X=1$, tal probabilidad es

$$\frac{e^{\beta_0^* + \beta_1}}{1 + e^{\beta_0^* + \beta_1}}$$

por lo que la ventaja para los expuestos es

$$\frac{\dfrac{e^{\beta_0^* + \beta_1}}{1 + e^{\beta_0^* + \beta_1}}}{\dfrac{1}{1 + e^{\beta_0^* + \beta_1}}} = e^{\beta_0^* + \beta_1}$$

De igual manera, para los no expuestos, $X=0$, la probabilidad de tener la enfermedad es

$$\frac{e^{\beta_0^*}}{1 + e^{\beta_0^*}}$$

por lo que para ellos la ventaja será

$$\frac{\dfrac{e^{\beta_0^*}}{1 + e^{\beta_0^*}}}{\dfrac{1}{1 + e^{\beta_0^*}}} = e^{\beta_0^*}$$

El cociente de las dos ventajas, es decir, la razón de ventajas es entonces

$$OR = \frac{e^{\beta_0^* + \beta_1}}{e^{\beta_0^*}} = e^{\beta_1}$$

por lo que el coeficiente β_1 de la predictora sigue teniendo la misma interpretación que en los estudios de seguimiento como logaritmo de la razón de ventajas.

1.9 Estimación de los parámetros del modelo *

En el modelo de regresión lineal, el método de elección para la estimación de los coeficientes es el de los mínimos cuadrados; este criterio de estimación se basa en calcular las estimaciones de los coeficientes de tal forma que la suma de los cuadrados de los residuales, la diferencia entre lo observado y lo predicho por el modelo, sea lo más pequeña posible. En regresión logística, como en otros muchos modelos de regresión, el método de estimación de los parámetros es el de *máxima verosimilitud (maximum likelihood)*. También este criterio de estimación tiene una base intuitiva fácil de aceptar: se trata de estimar los coeficientes del modelo bajo el supuesto de que lo que ha ocurrido, la experiencia observada, es lo más probable de lo que podía haber ocurrido.

Como ejemplificación del método consideremos la siguiente situación: se quiere estimar la probabilidad de que un recién nacido sea niño para lo cual se toma al azar una muestra aleatoria de 500 recién nacidos de los cuales 269 son niños; ¿cómo estimar la probabilidad de nacer niño? Cualquiera dirá que tal estimación es el cociente 269/500=0,538; veamos cómo estimar esta probabilidad según el método de máxima verosimilitud. Ya que el sexo de un recién nacido no condiciona el de los restantes, es decir, los sexos de los recién nacidos son independientes, la probabilidad de 269 niños y, por tanto, 231 niñas viene dada, según se vió en el Apartado 1.3, por

$$\binom{500}{269} p^{269}(1-p)^{231}$$

donde p representa la probabilidad de nacer niño; esta expresión que representaremos por $L(p)$, entendida como función del parámetro p a estimar, se denomina *función de verosimilitud*. Pues bien, la estimación de máxima verosimilitud de p es el valor, que representaremos con el símbolo $\hat{p}$, que hace que la función de verosimilitud sea lo más grande posible.

Esto se puede conseguir tanto gráfica como numéricamente; en efecto, se trataría de representar gráficamente la función de verosimilitud para distintos valores de p y el estimador de máxima verosimilitud sería el valor de p en el que tal función alcan-

* Esta sección puede omitirse en una primera lectura

ce su valor más grande; la Figura 1.4 muestra tal representación gráfica, en donde se observa que el valor máximo de $L(p)$ corresponde a p =0,538; por tanto, 0,538 es la estimación de máxima verosimilitud de la probabilidad de nacer niño.

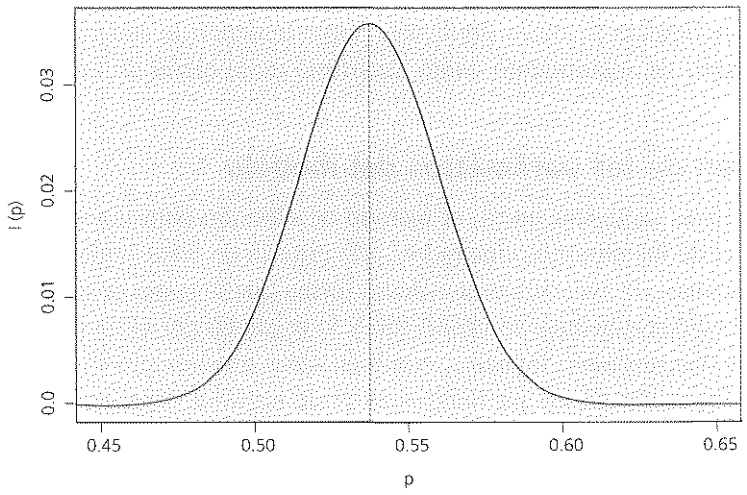


Figura 1.4. *Representación de la función de verosimilitud; la vertical corresponde a p=0,538.*

Desde el punto de vista analítico, ya que el máximo de una función coincide con el máximo de su logaritmo y este es más fácil de tratar matemáticamente, buscaremos el valor de p que haga lo más grande posible el logaritmo de la expresión anterior; por tanto, la función, salvo una constante, a maximizar es

$$l = \log(L) = 269 \log(p)+231 \log(1-p)$$

Los extremos de una función se consiguen calculando su primera derivada, igualándola a cero y resolviendo la ecuación resultante; en el caso que nos ocupa la derivada de la función anterior es

$$269 \frac{1}{p} + 231 \frac{-1}{1-p}$$

que igualada a cero y despejando el lector puede comprobar que da lugar a una estimación $\hat{p}$=269/500=0,538; es decir, 0,538 es la estimación de máxima verosimilitud de la probabilidad de nacer niño, en base a la muestra de recién nacidos elegida, lo que está de acuerdo con la Figura 1.4.

El método de máxima verosimilitud es un procedimiento muy utilizado en estadística a la hora de estimar los parámetros o coeficientes de un modelo. Veamos a continuación un caso un poco más complicado de estimación; el de un estudio de

cohorte muy simple pues solo vamos a considerar una predictora dicotómica por lo que tendremos dos cohortes de tamaños n_1, n_2. Notemos por Y_1 la variable número de individuos que presentan,tras el seguimiento, la característica de interés al tomar muestras de tamaño n_1 de la población de la que proviene la primera cohorte, e Y_2, n_2 lo análogo en la segunda cohorte; sean p_1 y p_2 las probabilidades de presentar la característica en esas dos poblaciones y representemos por y_1 e y_2 los valores de Y_1 e Y_2 en las dos cohortes concretas elegidas; por último, sea X la variable predictora dicotómica con valores $X=1$ para los expuestos y $X=0$ para los no expuestos. Si la elección de los individuos que constituyen las cohortes se hace de forma independiente, podemos aceptar que Y_1 e Y_2 son dos variables binomiales independientes de parámetros (n_1 , p_1) y (n_2 , p_2).

En estas condiciones, la probabilidad de que en la primera cohorte, la de los expuestos, se hayan encontrado tras el seguimiento y_1 individuos con la característica viene dada por la expresión

$$P(Y_1=y_1) = \binom{n_1}{y_1} p_1^{y_1},(1-p_1)^{n_1-y_1}$$

y, de igual forma, la probabilidad de que en la segunda cohorte y_2 individuos presenten la característica es

$$P(Y_2=y_2) = \binom{n_2}{y_2} p_2^{y_2},(1-p_2)^{n_2-y_2}$$

Por tanto, apoyándose en la independencia de las dos variables, la probabilidad de que en la primera cohorte haya habido y_1 individuos con la característica e y_2 en la segunda será

$$P(Y_1=y_1)\cdot P(Y_2=y_2) = \prod_{i=1}^{2} P(Y_i=y_i) = \prod_{i=1}^{2} \binom{n_1}{y_1} p_i^{y_i}(1-p_i)^{n_i-y_i}$$

donde el símbolo $\prod_{i=1}^{2}$ representa el producto de los dos probabilidades.

Ya que p_i depende de X a través de los parámetros β_i , el producto anterior es una función de β_0 y β_1 . La expresión

$$L(\beta_0 , \beta_1) = \prod_{i=1}^{2} P(Y_i=y_i)$$

contemplada como función de los parámetros β_0 y β_1, es, salvo una constante, la función de verosimilitud. La cuestión se reduce por tanto al cálculo de los valores de los parámetros que hagan máxima a la función

$$l(\beta_0 , \beta_1) = \log \{L(\beta_0 , \beta_1)\}$$

que se puede escribir así

$$l(\beta_0, \beta_1) = \log \prod_{i=1}^{2} P(Y_i=y_i) = \sum_{i=1}^{2} \log \{P(Y_i=y_i)\}$$

que, salvo la constante antes aludida, es igual a

$$\sum_{i=1}^{2} \{y_i \log \left(\frac{p_i}{1-p_i}\right) + n_i \log(1-p_i)\}$$

o bien,

$$\sum_{i=1}^{2} \{y_i \log (p_i) + (n_i - y_i) \log(1-p_i)\}$$

Sustituyendo p_i por su valor

$$p_i = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

y manipulando algebraicamente la expresión se llega a esta otra

$$L(\beta_0, \beta_1) = \sum_{i=1}^{2} y_i \ (\beta_0+\beta_1 x_i) - \sum_{i=1}^{2} n_i \log \{1 + e^{\beta_0+\beta_1 X}\}$$

El máximo de esta función se consigue derivando respecto de los parámetros β_0 y β_1 e igualando a cero ambas derivadas; en definitiva, obtenemos un sistema de dos ecuaciones con dos incógnitas β_0 y β_1, cuyas soluciones son las estimaciones de máxima verosimilitud de los dos parámetros.

A efectos de demostración de la gran complejidad de cálculo que implica este proceso, y para lectores con conocimientos de calculo diferencial, vamos a desarrollar este caso, el más simple por otra parte. La derivada del logaritmo de la función de verosimilitud respecto al parámetro β_0 es

$$\frac{\partial l\ (\beta_0, \beta_1)}{\partial \beta_0} = y_1 + y_2 - \frac{n_1\, e^{\beta_0+\beta_1}}{1 + e^{\beta_0+\beta_1}} - \frac{n_2\ e^{\beta_0}}{1 + e^{\beta_0}}$$

y la derivada respecto a β_1 es

$$\frac{\partial l\ (\beta_0, \beta_1)}{\partial \beta_1} = y_1 - \frac{n_1\ e^{\beta_0+\beta_1}}{1 + e^{\beta_0+\beta_1}}$$

Igualando estas dos derivadas a cero y tras alguna manipulación se llega al siguiente sistema de dos ecuaciones con dos incógnitas, las estimaciones de máxima verosimilitud,

$$\frac{e^{\beta_0+\beta_1}}{1 + e^{\beta_0+\beta_1}} = \frac{y_1}{n_1} \qquad \text{y} \qquad \frac{e^{\beta_0}}{1 + e^{\beta_0}} = \frac{y_2}{n_2}$$

cuyas soluciones son

$$\hat{\beta}_0 = \log \frac{y_2}{n_2 - y_2} \qquad \hat{\beta}_1 = \log \frac{\frac{y_1}{n_1 - y_1}}{\frac{y_2}{n_2 - y_2}}$$

es decir, la estimación del coeficiente β_1 es el logaritmo de la estimación del la razón de ventajas.

Para el ejemplo anterior del colesterol y la enfermedad coronaria, Tabla 1.1, tenemos dos cohortes de tamaños n_1=470 y n_2=254, en las que tras el seguimiento se han detectado y_1=31, e y_2=8 individuos con la característica de interés, la enfermedad coronaria; por tanto,

$$\hat{\beta}_0 = \log \frac{8}{254-8} = -3{,}426 \qquad \hat{\beta}_1 = \log \frac{\frac{31}{470-31}}{\frac{8}{254-8}} = \log(2{,}17) = 0{,}775$$

donde el valor 0,775 de $\hat{\beta}_1$ no es más que el logaritmo del valor 2,17 de la *OR* deducida de la Tabla 1.1

Como se dijo antes, este caso desarrollado es el más sencillo posible porque disponemos tan sólo de una predictora que además es dicotómica. Sin embargo, un análisis más fino de la relación entre enfermedad coronaria y nivel de colesterol debería tener en cuenta los valores individuales de colesterol sin tener que pagar el precio de la categorización, pues ésta conlleva a una pérdida de información; si, como en este ejemplo, la predictora es continua por naturaleza y se han seguido a n individuos, es posible que la mayoría de los valores x_i de X sean distintos y como p_i depende del valor x_i esas probabilidades serán distintas por lo que, en este caso, cada individuo representaría una cohorte distinta de tamaño n_i=1, es decir, tendremos que considerar n variables binomiales Y_i de parámetros (1 , p_i) independientes. Siguiendo un procedimiento similar al anterior, la función de verosimilitud es ahora

$$\sum_{i=1}^{2} \{y_i \log(p_i) + (1-y_i) \log(1-p_i)\}$$

y las ecuaciones que dan lugar a las estimaciones de máxima verosimilitud son

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}} = 0$$

y

$$\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \frac{x_i\, e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}} = 0$$

donde y_i=1 si el individuo i presenta la enfermedad e y_i=0 en caso contrario. Este sistema de ecuaciones es más complicado de resolver que el anterior y se necesitan métodos de cálculo numérico para encontrar sus soluciones; tales métodos conllevan procesos iterativos de gran complejidad en su ejecución. Esta es la razón de que, aunque estos modelos se propusieron al principio de la década de los sesenta, su uso no se haya extendido hasta que se popularizó la utilización de los ordenadores para los que esos procesos de cálculo no son, en general, un inconveniente serio.

Una cuestión importante, y de la que haremos uso a lo largo de esta monografía, es que los estimadores de máxima verosimilitud se distribuyen, para el caso de grandes muestras, según una distribución normal de media precisamente el parámetro a estimar; este hecho nos posibilitará la realización de contrastes de hipótesis y el cálculo de intervalos de confianza para los parámetros del modelo mediante la expresión

$$\hat{\beta}_1 \pm t_\alpha \, [ee \, (\hat{\beta}_1)]$$

donde $ee \, (\hat{\beta}_1)$ es el error estandar del estimador del parámetro y t_α es la cantidad a buscar en la distribución normal estandarizada al error α considerado.

1.10 Ejemplo

En el Apartado 1.9 se ha introducido el método de estimación de los parámetros en regresión logística, el método de máxima verosimilitud; este método porporciona, además de la estimación de los parámetros, los errores estandar de tales estimaciones, lo que nos va a permitir calcular intervalos de confianza para tales parámetros.

Consideremos los datos del estudio que aparecen en la Tabla 1.2, diseñado para evaluar la asociación entre el rendimiento escolar y el estado nutricional de los niños.

Tabla 1.2. *Distribución del rendimiento escolar según el estado nutricional.*

		ESTADO NUTRICIONAL	
		Malo	Bueno
RENDIMIENTO ESCOLAR	Malo	105	15
	Aceptable	80	300

Aquí, la variable respuesta es el rendimiento escolar que vamos a codificar como Y=1 si el niño tuvo un mal rendimiento escolar e Y=0 en caso de un rendimiento aceptable. De forma análoga, la predictora *ESTN*, el estado nutricional, la codificaremos

como *ESTN*=1 si éste es malo y *ESTN*=0 si el estado nutricional es bueno. Utilizando cualquier programa informático de regresión logística, los resultados de este estudio son los que aparecen en la Tabla 1.3; en ésta aparece una primera columna con la variable y el término independiente, también llamado constante; otra columna en donde aparecen las estimaciones de los dos parámetros del modelo y, una tercera en donde están sus correspondientes errores estandar.

Tabla 1.3. *Estimación y error estandar de los parámetros del modelo que contiene a la variable ESTN.*

VARIABLE	ESTIMACIÓN	E.E.
Constante	-2.996	0.265
ESTN	3.268	0.303

Por tanto el modelo estimado es

$$logit(\hat{p}) = \log \frac{\hat{p}}{1-\hat{p}} = -2{,}996 + 3{,}268\ ESTN$$

Según estos resultados, 3.268 es una estimación de β_1, y como antes se dijo que la *OR* estimada era e^{β_1}, podemos decir que, $OR = e^{3{,}268} = 26{,}26$, lo que se puede interpretar diciendo que la estimación del riesgo de tener mal rendimiento escolar es, aproximadamente, 26 veces más alto entre los niños mal nutridos que entre los bien nutridos. Esta interpretación es así y no de otra manera por la asignación de valores que hemos hecho a las dos variables implicadas; en efecto, ya que *Y*=1 representa el mal rendimiento escolar, 26,26 es una estimación del riesgo de tener esa característica de un niño mal nutrido, *ESTN*=1, respecto a uno bien nutrido, *ESTN*=0.

Imaginemos que hubiésemos codificado a la variable predictora al revés, es decir, *ESTN*=1 para los bien nutridos y *ESTN*=0 para los mal nutridos; en tal caso la estimación del coeficiente de la predictora hubiese sido el mismo pero con distinto signo, es decir, -3,268; por tanto, el riesgo sería $e^{-3{,}268} = 0{,}038$, que ahora se interpretaría como una medida del riesgo de mal rendimiento escolar de un bien nutrido respecto de un mal nutrido; como era de esperar 1/26,26 es, salvo errores de redondeo, precisamente 0,038.

La Tabla 1.3 también muestra el error estandar para el estimador de β_1; debido a que, como se señaló anteriormente, los estimadores de máxima verosimilitud se distribuyen aproximadamente según una normal, un intervalo de confianza para β_1 será

$$\hat{\beta}_1 \pm t_\alpha\ [ee\ (\hat{\beta}_1)]$$

por lo que un intervalo *aproximado*, al 95% de confianza, para β_1, el logaritmo de la *OR* es 3,268 ± 1,96 (0,303) = (2,674 , 3,862). Ya que $OR = e^{\beta_1}$, un intervalo para la *OR* será

$$e^{\hat{\beta}_1 \pm t_\alpha [ee(\hat{\beta}_1)]}$$

que para nuestro ejemplo es

$$(e^{2,674} , e^{3,862}) = (14,50 , 47,56)$$

Este intervalo significa que tenemos una confianza del 95% de que la verdadera *OR*, el parámetro a estimar, debe valer como poco 14,50 y como mucho 47,56; dicho de otra forma, con una confianza del 95% podemos decir que un mal nutrido está como poco a 14,50 veces y como mucho a 47,56 veces más riesgo de tener mal rendimiento escolar que un niño bien nutrido.

Analicemos ahora la tabla por el método clásico del test basado en la distribución chi-cuadrado; para los datos de la Tabla 1.2 la *OR* estimada es

$$OR = \frac{(105)\ (300)}{(15)\ (80)} = 26,25$$

Un intervalo de confianza para la *OR* en tablas 2x2 propuesto por Miettinen es

$$OR^{1 \pm t_\alpha \sqrt{\chi^2}}$$

donde χ^2 es el valor de la chi-cuadrado de la tabla; en este caso, el lector puede comprobar que tal valor es 172,66 por lo que el intervalo, al 95% de confianza, será

$$26,25^{\ 1 \pm 1,96\ /\ \sqrt{172,66}} = (16,12\ \ ,\ \ 42,74)$$

Otra alternativa a la construcción de un intervalo de confianza para una razón de ventajas es el llamado método de Woolf; en este caso, una estimación aproximada del error estandar para el logaritmo de la *OR* viene dada por la raiz cuadrada de la suma de los inversos de los valores de las casillas, por lo que un intervalo de confianza, al 95%, para log(*OR*) viene dado por

$$\log(OR) \pm 1,96\sqrt{(1/a)+(1/b)+(1/c)+(1/d)}$$

donde *a*, *b*, *c*, *d* son los cuatro valores de las casillas de la tabla 2x2. Para nuestro ejemplo,

$$\log(26,25) \pm 1,96\sqrt{(1/105)+(1/15)+(1/80)+(1/300)} = (2,673\ ,\ 3,862)$$

por lo que el intervalo para la *OR* es $(e^{2,673}, e^{3,862}) = (14,48, 47,56)$, intervalo también muy parecido al calculado a partir del modelo logístico.

1.11 Evaluación de la bondad del ajuste

Ante los datos de un estudio, proponemos un modelo para explicar la relacción entre la variable resultado y la predictora y estimamos sus parámetros, es decir, ajustamos el modelo a los datos. El resultado del ajuste de un modelo a un conjunto de datos puede contemplarse como la sustitución de los valores observados de la variable resultado por las estimaciones de sus valores medios, estimaciones derivadas de la función de regresión. En el modelo lineal clásico, una medida del grado de desajuste entre lo observado y lo predicho por el modelo viene dada por la suma de cuadrados de los residuales.

En regresión logística lo que haremos será comparar la verosimilitud del modelo ajustado con la del modelo "perfecto", el llamado *modelo saturado*; éste es el modelo que reproduce exactamente las observaciones realizadas, pero siempre es un modelo demasiado complicado pues tiene tantos parámetros como sujetos de estudio; sin embargo, lo podemos utilizar como referencia con el que comparar nuestro modelo ajustado. Si acaso el modelo ajustado fuese parecido, desde el punto de vista estadístico, al modelo saturado, podríamos pensar que el modelo ajustado explica nuestros datos suficientemente bien, con la gran ventaja de tener muchos menos parámetros y, por tanto, mucho más sencillo; si, por el contrario, nuestro modelo ajustado fuese muy distinto del saturado, tendríamos que concluir que nuestro modelo no da una explicación suficiente a los datos de nuestro estudio.

La medida que utilizaremos en regresión logística para comparar nuestro modelo ajustado con el modelo saturado se denomina *lejanía (deviance)* que en general de define mediante la expresión

$$Lejanía = -2 \log \frac{verosimilitud\ del\ modelo\ ajustado}{verosimilitud\ del\ modelo\ saturado}$$

donde la razón de tomar -2 veces el logaritmo de la *razón de verosimilitudes (likelihood ratio)* se debe al hecho de que esta razón de verosimilitudes no tiene una distribución conocida y, sin embargo, si se conoce que la lejanía se distribuye aproximadamente según una chi-cuadrado, aunque ahora veremos que no siempre es así.

La lejanía es una medida que como poco vale 0, cuando nuestro modelo explique los datos perfectamente, y que cuanto mayor sea su valor peor explicará nuestro modelos los datos observados. Para el caso de la regresión logística la lejanía viene dada por

$$lejanía = 2 \sum_{i=1}^{2} \{ y_i \log \frac{y_i}{\hat{y}_i} + (n_i - y_i) \log \frac{n_i - y_i}{n_i - \hat{y}_i} \}$$

donde k es el número de grupos de sujetos de estudio en relación a la predictora. En el ejemplo de la Tabla 1.2, $k=2$, los bien nutridos y los malnutridos; n_i es el número de individuos en cada grupo, 185 y 315; y_i representa el número de sujetos con la característica de interés, el mal rendimiento escolar en cada grupo, 105 y 80; por último, $\hat{y}_i$ son los valores correspondientes a éstos que predice el modelo ajustado.

Veamos cuales son estos valores predichos para nuestro ejemplo; ya que el modelo estimado fue

$$logit(\hat{p}) = \log \frac{\hat{p}}{1-\hat{p}} = -2{,}996 + 3{,}268\ ESTN$$

podemos afirmar que el logit de la probabilidad de tener un mal rendimiento escolar para un niño malnutrido, $ESTN=1$, es -2,996+3,268(1)=0.272, por lo que la probabilidad de tener mal rendimiento escolar entre los niños malnutridos es

$$\frac{e^{0,272}}{1 + e^{0,272}} = 0{,}568$$

es decir, nuestro modelo predice que el 56.8% de los malnutridos tiene dificultades en el colegio; por tanto, de los 185 niños malnutridos de la Tabla 1.2 es esperable que haya 185.(0.568), es decir, 105 salvo errores de redondeo. Para los niños bien nutridos, $ESTN=0$, el logit de la probabilidad de tener un mal rendimiento escolar será -2,996+3,268(0)=-2,996, por lo que la probabilidad de tener mal rendimiento escolar entre los niños bien nutridos es

$$\frac{e^{-2,996}}{1 + e^{-2,996}} = 0{,}0476$$

por lo que entre los 315 bien nutridos es esperable que haya 315·(0,476), es decir, 15.

Si los fracasos escolares observados y predichos por el modelo son idénticos lo mismo ocurrirá para los que tienen un buen rendimiento escolar, luego el modelo ajustado reproduce las observaciones perfectamente; dicho de otra forma, el modelo ajustado y el saturado se comportan igual por lo que la lejanía para nuestro modelo debería ser 0. En efecto, sustituyendo en la expresión de la lejanía tendremos

$$2 \left(105 \log \frac{105}{105} + (185\text{-}105) \log \frac{185\text{-}105}{185\text{-}85} + 15 \log \frac{15}{15} + (315\text{-}15) \log \frac{315\text{-}15}{315\text{-}15}\right) = 0$$

Antes se avanzó que en ciertas ocasiones la lejanía no sigue una distribución chi-cuadrado y esto es más la regla que la excepción. Para el caso de *datos agrupados*, es decir, cuando la o las predictoras sean todas categóricas y siempre que en cada casilla haya un númro suficiente de observaciones, lejanía sigue una distribución χ^2 con k-ν grados de libertad, siendo k el número de patrones distintos de la o las predictoras y ν el número de parámetros que se estiman en el modelo; para nuestro ejemplo k=2, porque tenemos solo dos grupos distintos de niños en relación al estado nutricional, y también hemos estimado dos parámetros luego ν=2; por tanto, el modelo ajustado tiene 0 grados de libertad. Este hecho nos va a permitir considerar, en el caso de datos agrupados, a la lejanía como una medida de la *bondad del ajuste* realizado comparando el valor de ella con el de la distribución chi-cuadrado con k-ν grados de libertad, al error α que se considere. Así, siempre que la lejanía sea menor que el valor de la chi-cuadrado correspondiente no hay evidencias de mal ajuste y diremos que el modelo ajustado reproduce nuestros resultados suficientemente bien. Sin embargo, en el caso de *datos no agrupados* n_i=1, es decir, en las casillas haya pocos individuos o cuando la predictora sea una variable continua, casi cada individuo tendrá un patrón de predictoras distinto; pues bien, para este caso no se conoce la distribución de la lejanía por lo que no podemos saber si un valor concreto de ella implica necesariamente evidencia de un mal ajuste.

Otra medida de la bondad del ajuste de un modelo, con las mismas limitaciones que la lejanía, es el estadístico χ^2, definido según la siguiente expresión

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - n_i\hat{p}_i)^2}{n_i\hat{p}_i\,(1 - \hat{p}_i)}$$

dejando al lector la comprobación de que para nuestro modelo ajustado tiene el mismo valor que la lejanía, es decir, 0. La razón del mayor uso de la lejanía está justificada por el hecho de que, para *modelos anidados*, modelos que unos son casos particulares de otros, la lejanía es una medida aditiva, lo que permite medir la contribución relativa de un conjunto de variables a la bondad del ajuste; sin embargo, el estadístico de Pearson no goza de esa propiedad de aquí que casi se haga, de ahora en adelante, uso exclusivo de la lejanía. De estas propiedades haremos uso en los capítulos posteriores.

1.12 Comparación de modelos

Una pregunta interesante a responder es ¿está asociado el rendimiento escolar al estado nutricional de los niños? o lo que es igual ¿hay diferencias entre los bien nutridos y los malnutridos en cuanto a su rendimiento escolar?. Esta cuestión la podemos contestar afirmativamnte pues vimos en el Apartado 1.10 que el intervalo de confianza

para la razón de ventajas es (14,50 , 47,56), por lo que el valor 1 que es el que establece la independencia no es un valor pausible a la vista de los datos.

Sin embargo, vamos a ver otro método para conseguir lo mismo pero que tiene ventajas importantes sobre éste. Para ello supongamos que no existen tales diferencias en el rendimiento escolar en los dos grupos de niños; bajo esta hipótesis, el modelo es el siguiente

$$logit(p) = \log \frac{p}{1-p} = \beta_0$$

pues partimos de la hipótesis de que no vamos a tener en cuenta el estado nutricional de los niños por lo que esta variable no aparece; los resultados de este nuevo modelo ajustado están en la Tabla 1.4.

Tabla 1.4. *Estimación del parámetro para el modelo que establece la independencia.*

VARIABLE	ESTIMACIÓN	E.E.
Constante	-1.153	0.1047

Según estas estimaciones

$$\frac{e^{-1,153}}{1 + e^{-1,153}} = 0,240$$

es una estimación de la probabilidad de mal rendimiento escolar, sea cual sea el estado nutricional del niño; obsérvese que este 0.24 no es más que el porcentaje de malnutridos, 105+15, entre el total de niños, 185+315. Si esto es así, de entre los 185 niños malnutridos este nuevo modelo predice 185(0,24)=44,4 en contra de los 105 observados; entre los 315 bien nutridos son esperables 315(0,240)=75,6 cuando realmente hubo 15. El lector puede comprobar que la lejanía de este nuevo modelo es de 177,39; los grados de libertad son ahora 2-1=1, pues tenemos dos patrones distintos de la predictora pero se ha ajustado un modelo con un sólo parámetro a estimar, el β_0. Como 177.39 es un valor muy grande para la distribución chi-cuadrado con 1 grado el modelo ajustado es estadísticamente distinto del modelo saturado, es decir, no reproduce fehacientemente los datos observados. En resumen, el rendimiento escolar está asociado al estado nutricional en el sentido de que los malnutridos están a más riesgo del fracaso escolar. El lector puede comprobar que para el modelo que no contiene a la predictora, el valor del estadístico χ^2 de Pearson es 172.66, valor parecido al de la lejanía.

El mayor interés de la lejanía está en su uso, más que como medidad de la bondad del ajuste, como herramienta para comparar distintos modelos ajustados; esto se

debe al hecho de que la diferencia entre los lejanías de dos modelos siempre, para datos agrupados o no agrupados, sigue una distribución chi-cuadrado, lo que posibilita la comparación de los dos modelos. Siguiendo con nuestro ejemplo, para el modelo que contiene a la predictora *ESTN*, la lejanía es 0 mientras que para el modelo en que no figura tal variable, es decir, el que establece la independencia entre el estado nutricional y el rendimiento escolar la lejanía es 177,39; parece claro que la diferencia entre esas dos lejanías, 177,39-0=177,39, es el precio a pagar, en desajuste, por no considerar el estado nutricional. Pues bien, se demuestra que, bajo la hipótesis de que los dos modelos explican los datos de la misma manera, la diferencia entre las lejanías correspondientes a los dos modelos sigue una distribución χ^2 cuyos grados de libertad son la diferencia de los correspondientes a las dos lejanías, en nuestro caso 1-0=1; por tanto, ya que el valor 177,39 es un valor extremadamente alto para la distribución chi-cuadrado con un grado de libertad, no podemos achacar, salvo el error α que en este caso es muy pequeño, al azar la diferencia entre las dos lejanías; por tanto, podemos rechazar la hipótesis anterior de equivalencia de los dos modelos y concluir que el rendimiento escolar está asociado al estado nutricional.

En general, y esto vale también para el caso multivariante; supongamos dos modelos M_1, M_2 ajustados a unos datos y tales que M_1 sea un caso particular de M_2, es decir, el primero tiene sólo parte de las predictoras del segundo; se conoce que si D_1, D_2 son sus correspondientes lejanías, siempre se cumple que $D_1 \geq D_2$. Pues bien, bajo la hipótesis de que los dos modelos explican las observaciones de forma equivalente, la diferencia $D_1 - D_2$ de las dos lejanías sigue una distribución χ^2 cuyos grados de libertad son la diferencia de los grados de libertad de D_1 y de D_2. Por tanto, para decidir con qué modelo quedarnos, calcularemos la diferencia $D_1 - D_2$ y la compararemos con la distribución χ^2 al error α elegido. Si la diferencia en lejanías es un valor mayor que el de la distribución teórica, significa que debemos rechazar la hipótesis de equivalencia de los dos modelos y por tanto nos quedaremos con el modelo M_2; en caso contrario, si la diferencia entre las lejanías es un valor común para la distribución teórica, no hay motivos para pensar que los dos modelos no son equivalentes, por lo que nos decidiremos por el más sencillo, el M_1.

Un hecho importante a recordar es que, aunque dispongamos de datos no agrupados y por tanto las lejanías no sigan distribuciones chi-cuadrado, la diferencia entre lejanías si sigue tal distribución, por lo que el método de comparación de modelos que se acaba de exponer es válido en cualquier ocasión.

Una forma alternativa para contrastar que $\beta_1 = 0$, podría ser, apoyándose en la normalidad de los estimadores de máxima verosimilitud, dividir $\hat{\beta}_1$ por su error estandar y comparar tal cociente 3,268/0,303 = 10,79 con la distribución normal, lo que nos permite también rechazar la hipótesis de independencia de las dos variables. Este nuevo

método para evaluar la asociación entre la variable resultado y la predictora es llamado *método de Wald* y es el que muestra cualquier paquete informático. Ya que este método no es válido en ciertas ocasiones, como norma general se debe utilizar el método basado en la comparación de las lejanías pues es un método más fiable.

CAPÍTULO II

REGRESIÓN LOGÍSTICA BINARIA MÚLTIPLE

Una vez definido el modelo logístico simple, en este capitulo introducimos la versión multivariante; discutimos cómo controlar el efecto de las posibles confusoras y cómo modelar la interacción entre las predictoras. El tratamiento de éstas cuando son categóricas así como la selección del modelo más parsimonioso, ocupan el resto del capítulo.

2.1 Introducción

Como ya se dijo en el capítulo I, no suele bastar con una sola predictora a la hora de estudiar cómo cambia una variable resultado; los fenómenos de la naturaleza suelen ser más complejos y, por tanto, será necesario el conocimiento de varias predictoras para explicar, al menos en una parte sustancial, la variabilidad de la variable resultado. Como ya se avanzó, una posible estrategia de análisis podría ser la estratificación mediante la categorización de las predictoras, pero ya se comentó la necesidad de muestras tremendamente grandes para poder realizar estimaciones de los parámetros de interés que sean fiables. El modelo logístico multiple o multivariante es un método mucho más eficiente pues permite estimar los efectos de varias variables simultáneamente sin tener que pagar el precio de muestras tan grandes y tiene la ventaja añadida de no tener que categorizar las variables cuantitativas.

2.2 El modelo logístico binario múltiple

Consideremos ahora la variable dicotómica respuesta Y y un conjunto de predictoras $X_1, X_2, \ldots, X_m$ medidas en n individuos. El modelo logístico multivariante establece que

$$logit(p) = \log \frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m$$

donde los β_i son los parámetros desconocidos del modelo. De manera análoga al caso univariante, otra forma de expresar este mismo modelo es

$$p = \frac{e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m)}}$$

o bien,

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m)}}$$

En definitiva, este modelo es muy parecido en su definición al modelo logístico simple, la única diferencia es que ahora entran en juego mas variables con la esperanza de que nos ayuden a entender mejor porqué varía la respuesta de unos individuos a otros. Según la definición del modelo multivariante, para un individuo concreto, cuanto mayor sea el valor de

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m$$

llamado *componente sistemático* del modelo, mayor será la probabilidad de que presente la característica de interés.

Como para cualquier modelo de regresión, no se establece ninguna restricción acerca de la naturaleza de las variables predictoras X_i ; pueden ser categóricas, discretas o continuas.

2.3 Interpretación de los coeficientes

En el modelo de regresión logística simple, la interpretación del coeficiente de la variable predictora era clara: si X es una variable dicotómica que toma el valor 1 para el estado de "exposición", estar malnutrido para el ejemplo de la Tabla 1.2, y el valor 0 para la "no exposición", estar bien nutrido, el coeficiente β_1 de X se podía interpretar como el logaritmo natural de la razón de ventajas de estar "enfermo", tener mal rendimiento escolar, respecto a "no estar enfermo", tener buen rendimiento escolar. Dicho de otra forma, elevando el número e a β_1 obtenemos el riesgo de un individuo expuesto respecto a otro no expuesto, es decir, $OR=e^{\beta_1}$.

Veamos ahora la interpretación de los coeficientes en el modelo logístico multivariante y, para fijar ideas, nos vamos a centrar en el coeficiente β_1 de la variable X_1 siendo ésta una variable dicotómica con valor 1 para los individuos "expuestos" y 0 para los "no expuestos". Consideremos ahora dos individuos: el primero, el A, expuesto y el segundo, el individuo B, no expuesto en relación a X_1, pero idénticos en cuanto a los valores del resto de las variables predictoras. Por tanto, el logaritmo de la ventaja para el primero es

$$\log \frac{p_A}{1-p_A} = \beta_0 + \beta_1(1) + \beta_2 X_2 + \ldots + \beta_m X_m$$

y para el individuo B

$$\log \frac{p_B}{1-p_B} = \beta_0 + \beta_1(0) + \beta_2 X_2 + \ldots + \beta_m X_m$$

donde p_A y p_B son las probabilidades de presentar la característica los individuos A y B respectivamente. Restando estas dos igualdades

$$\log \frac{p_A}{1-p_A} - \log \frac{p_B}{1-p_B} = (\beta_0 + \beta_1(1) + \beta_2 X_2 + \ldots + \beta_m X_m) - (\beta_0 + \beta_1(1) + \beta_2 X_2 + \ldots + \beta_m X_m)$$

$$= \beta_1(1 - 0) = \beta_1$$

pues antes dijimos que los dos eran iguales en todas la predictoras, a excepción de X_1; por tanto,

$$\log \frac{p_A \ / \ (1 - p_A)}{p_B \ / \ (1 - p_B)} = \log (OR) = \beta_1$$

por lo que la interpretación de β_1 sigue siendo la misma que en el caso univariante, hecha la importante salvedad de la coincidencia de los dos individuos en los valores de las restantes variables predictoras. Dicho de otra forma: β_1 es el logaritmo de la razón de ventajas respecto a la variable X_1 pero *controlando* por las otras variables presentes en el modelo; es decir, si A y B son dos individuos que son *iguales* en cuanto a todas las predictoras excepto para la X_1 en que A toma el valor 1 y B el valor 0, e^{β_1} es la *OR* del individuo A respecto del B. Si la variable X_1 es numérica, discreta o continua, con valores x_{A1} y x_{B1} en los individuos A y B respectivamente, la *OR* del individuo A respecto del B es entonces

$$OR = e^{\beta_1 (x_{A1} - x_{B1})}$$

Por tanto ya sabemos cómo comparar los riesgos entre dos individuos que son distintos solo en una variable predictora. Por ejemplo, si en el modelo figuran las predictoras edad, género y tabaco, ya sabemos como calcular la *OR* entre dos individuos de distinta edad, pero de igual género y hábito de fumar, o bien, entre un hombre y una mujer pero con la condición de que tengan la misma edad y fumen igual cantidad de tabaco, etc. Pero aparte de este tipo de comparaciones, podemos estar interesados en evaluar el número de veces que está a más riesgo un hombre cualquiera respecto de una mujer cualquiera, por ejemplo, un hombre de 50 años fumador de 25 cigarrillos respecto a una mujer de 40 años fumadora de 10 cigarrillos. Luego la pregunta a responder es: ¿cómo calcular la razón de ventajas entre dos individuos cualesquiera?

Consideremos que el individuo A tiene como valores de las predictoras $x_{A1}, x_{A2}, \ldots, x_{Am}$ y el individuo B, $x_{B1}, x_{B2}, \ldots, x_{Bm}$; por tanto, para el individuo A, el modelo será

$$\log \frac{p_A}{1-p_A} = \beta_0 + \beta_1 x_{A1} + \beta_2 x_{A2} + \ldots + \beta_m x_{Am}$$

y para el individuo B

$$\log \frac{p_B}{1-p_B} = \beta_0 + \beta_1 x_{B1} + \beta_2 x_{B2} + \ldots + \beta_m x_{Bm}$$

Restando estas dos expresiones tenemos que

$$\log \frac{p_A / (1-p_A)}{p_B /(1-p_B)} = \beta_1(x_{A1}- x_{B1}) + \beta_2(x_{A2}- x_{B2}) +\ldots+ \beta_m(x_{Am}- x_{Bm})$$

por lo que

$$OR = e^{\beta_1(x_{A1}- x_{B1}) + \beta_2(x_{A2}- x_{B2}) +\ldots+ \beta_m(x_{Am}- x_{Bm})} = e^{\beta_1(x_{A1}- x_{B1})} \cdot e^{\beta_2(x_{A2}- x_{B2})} \cdot \ldots \cdot e^{\beta_m(x_{Am}- x_{Bm})}$$

Recuérdese que en el modelo univariante, si x_A y x_B eran los valores de la predictora X en dos individuos, la razón de ventajas venía dada por la expresión $OR= e^{\beta_1(x_A- x_B)}$; lo que se acaba de demostrar es que, para el caso del modelo logístico multivariante, la razón de ventajas del individuo A respecto del B se consigue así

$$OR = e^{\beta_1(x_{A1}- x_{B1})} \cdot e^{\beta_2(x_{A2}- x_{B2})} \cdot \ldots \cdot e^{\beta_m(x_{Am}- x_{Bm})}$$

Obsérvese que esto es un producto donde cada factor es la razón de ventajas del individuo A respecto del B en relación a cada predictora en la que difieren, controlando por las restantes en cada caso. En definitiva, la *OR* entre dos individuos con distintos valores para varias predictoras se puede conseguir multiplicando las *OR* correspondientes a cada una de ellas. Siguiendo con el ejemplo anterior, el riesgo del hombre de 50 años y fumador de 25 cigarrillos, respecto a la mujer de 40 años y fumadora de 10 cigarrillos se consigue multiplicando el riesgo asociado al género, controlado por edad y tabaco, por el riesgo asociado al distinto número de cigarrillos, controlado por la edad y el género, y multiplicado por el riesgo asociado a la diferencia de 15 años de edad, controlado por género y tabaco. Ya que el riesgo global lo conseguimos como producto de riesgos asociados a las distintas predictoras, tiene sentido hablar del caracter multiplicativo del modelo de regresión logística.

2.4 Estimación de los coeficientes

En cuanto a la estimación de los coeficientes, el método sigue siendo el de máxima verosimilitud. Para el caso univariante, el logaritmo de la función de verosimilitud era

$$l(\beta_0 , \beta_1) = \sum_{i=1}^{n} y_i(\beta_0+\beta_1 x_i) - \sum_{i=1}^{n} n_i \log (1 + e^{\beta_0+\beta_1 x_i})$$

y ahora, de forma similar, es

$$l(\beta_0 , \beta_1, \ldots, \beta_m) = \sum_{i=1}^{n} y_i(\beta_0+\beta_1 x_{i1}+\ldots+\beta_m x_{im}) - \sum_{i=1}^{n} n_i \log (1 + e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_m x_{im}})$$

que derivando respecto a los $m+1$ parámetros e igualando a cero cada una de esas $m+1$ derivadas obtenemos un sistema de ecuaciones, evidentemente mucho más com-

plejo que en el caso de una sola predictora, que se tiene que resolver también por procedimientos iterativos.

En el ejemplo que aparece en la Tabla 1.2, también se disponía del género de los individuos de tal manera que la tabla 2x2 anterior se puede dividir en dos, una para niños y otra para niñas, es decir, se puede *estratificar* por género como aparece en la Tabla 2.1.

Tabla 2.1. *Rendimiento escolar y estado nutricional para niños y niñas.*

		ESTADO NUTRICIONAL			
		Niños		Niñas	
		Malo	Bueno	Malo	Bueno
RENDIMIENTO	Pobre	50	7	55	8
ESCOLAR	Aceptable	38	140	42	160

Cuando se pretende estudiar la relación entre dos variables, es muchas veces necesario tener en cuenta otras variables que pueden afectar a tal relación. Consideremos que estamos interesados en la relación entre Y y X; si se piensa que otra variable Z puede afectar a la asociación de interés y las tres variables son dicotómicas, podemos formar tres tablas : para cada uno de los dos valores de la variable Z podemos formar una tabla para medir la asociación entre Y y X. A estas dos tablas se les denomina *tablas parciales* y en cada una de ellas el valor de la variable Z es constante. Combinando las dos tablas parciales se puede formar otra tabla, la *tabla marginal*, en donde el valor de Z deja de ser constante, no es el mismo para todos los individuos observados. Pues bien, como se verá más adelante, las tablas parciales pueden mostrar un patrón de asociación entre Y y X no necesariamente igual al mostrado por la tabla marginal. Para el ejemplo que nos ocupa la variable Y es el rendimiento escolar, X es el estado nutricional y Z representa el género de los sujetos de estudio; las dos tablas parciales son las correspondientes a niños y a niñas y la tabla marginal es la que engloba a las 500 sujetos. Lo que se acaba de exponer se puede extender al caso en que X y Z tengan cualquier número de categorías.

Consideremos, en primer lugar, el modelo que contiene como única predictora al estado nutricional de los individuos, es decir, el análisis de la tabla marginal; el modelo ajustado es el que aparece en la Tabla 2.2,

Tabla 2.2. *Estimación del modelo con la variable ESTN.*

VARIABLE	ESTIMACIÓN	ERROR ESTÁNDAR
Constante	-2,996	0,265
ESTN	3,268	0,303

el mismo que se obtuvo en la Tabla 1.3 pues no se está teniendo en cuenta el género, con una lejanía de 0,00025 y 2 grados de libertad ya que ahora disponemos de 4 binomiales, es decir, cuatro grupos de individuos en relación a las predictoras: dos de niños, los bien y los mal nutridos, y otros de para las niñas; como se han estimado dos parámetros, grados de libertad serán 4-2=2, como se dijo en el apartado 1.11.

Cuando se añade al modelo la variable *SEXO*, codificada como *SEXO*=0 para los niños y 1 para las niñas, el análisis correspondiente a las tablas parciales, da lugar a una lejanía = 0,00006 y 1 grado de libertad, pues ahora hemos estimado un parámetro más que en el modelo anterior; las estimaciones correspondientes al nuevo modelo son las que aparecen en la Tabla 2.3.

Tabla 2.3. *Estimación del modelo con las variables ESTN y SEXO.*

VARIABLE	ESTIMACIÓN	ERROR ESTÁNDAR
Constante	-2,994	0,298
ESTN	3,268	0,303
SEXO	-0,0036	0,259

Obsérvese como la introducción del género en el modelo, es decir, cuando se controla por género, la relación entre el rendimiento escolar y el estado nutricional, medida mediante la *OR* sigue siendo $e^{3,268}=26,26$, es decir, el género no afecta a la relación de interés, la del estado nutricional con el rendimiento escolar; esto concuerda con el hecho de que las razones de ventajas para las tablas parciales coinciden con la de la marginal; en efecto, para niños

$$OR = \frac{(50)\ (140)}{(7)\ (38)} = 26,32$$

y para niñas

$$OR = \frac{(55)\ (160)}{(8)\ (42)} = 26,19$$

muy parecidos a 26,26, la razón de ventajas de la tabla marginal.

Ajustados estos dos modelos distintos una pregunta inmediata es, ¿cual de los dos modelos ajustados elegir?; evidentemente, cuantas más variables predictoras consideremos, menor será la lejanía y, por tanto, mayor acuerdo habrá entre los valores observados y los predichos por el modelo. Sin embargo, el *principio de parsimonia* recomienda explicar lo observado de la forma más sencilla posible, es decir, con el mínimo número de predictoras, siempre que no se pierda información sustantiva. En estas circuns-

tancias la pregunta es: ¿aporta el género información importante sobre la que ya aporta el estado nutricional a la hora de predecir el rendimiento escolar? La diferencia en lejanía entre los dos modelos es de 0,00019 y la diferencia en grados de libertad es de 1 pues el segundo modelo contiene un parámetro más, el correspondiente al género, que el primer modelo. Como se vió en el apartado 1.12, bajo la hipótesis de equivalencia de los dos modelos, la diferencia en lejanía sigue una chi-cuadrado cuyos grados de libertad son la diferencia entre el número de parámetros de los dos modelos; por tanto, para constrastar esa hipótesis comparamos el valor 0,00019 con el valor de la chi-cuadrado con 1 grado de libertad. Ya que 0,00019 no es, en absoluto, un valor grande para tal distribución, no hay razones para rechazar la equivalencia entre los dos modelos y apoyándonos en el principio de parsimonia elegiremos el más sencillo, el que no contempla el género de los individuos, es decir,

$$logit(p)=-2{,}996+3{,}268ESTN$$

Obsérvese que en el modelo que contiene al género como predictora, el coeficiente estimado para esta variable es -0,0036 y como los valores asignados a esta variable han sido el 0 a los niños y el 1 a las niñas, la cantidad $e^{-0,0036}=0,996$ es la estimación de la razón de ventajas de tener mal rendimiento escolar de las niñas respecto de los niños a igualdad del estado nutricional. Ya que el error estandar correspondiente al género es 0,259, el estadístico de Wald para evaluar la significación estadística de esta variable es 0,0036/0,259=0,01; este resultado nos indica que, controlando por el estado nutricional, el rendimiento escolar no está asociado al género, es decir, para individuos con igual estado nutricional, no hay diferencias de rendimiento escolar entre los géneros, por lo que se entiende que no sea necesario incorporar la variable género al modelo que contiene al estado nutricional.

2.5 El concepto de confusión y su control

En el ejemplo que acabamos de ver, el introducir la variable *SEXO* en el modelo logístico no alteraba para nada la relación entre el estado nutricional y el rendimiento escolar; dicho con otras palabras, la variable *SEXO no confunde* esa relación; ello significa que no es necesario desagregar por género a la hora de estudiar la relación entre el rendimiento escolar y el estado nutricional. Sin embargo hay muchas situaciones en que este no es el caso y, no solamente esto, a veces la relación entre dos variables puede depender del nivel que se considere de otra u otras variables. En este y en el siguiente apartado se trata de modelar situaciones como las referidas.

Los datos que aparecen en la Tabla 2.4 corresponden a la experiencia de mortalidad de un conjunto de pacientes en función del tratamiento recibido y según el hospital donde recibieron tal tratamiento.

Tabla 2.4. *Experiencia de supervivencia según tipo de tratamiento y hospital donde se recibió.*

	HOSPITAL			
	H1		H2	
	Vivo	Muerto	Vivo	Muerto
TRAT. A	150	50	240	560
TRAT. B	400	400	25	175

En este ejemplo, la variable respuesta es la supervivencia del paciente, que vamos a codificar con valor 1 si el individuo sobrevive y como 0 en caso de muerte. Al tratamiento recibido le asignaremos un 1 si fué el A y un 0 en caso de que el tratamiento fuese el B; en esas condiciones, ajustando un modelo con el tratamiento como única predictora, se obtienen unas estimaciones que aparecen en la Tabla 2.5 con una lejanía de 239,14 y 2 grados de libertad.

Tabla 2.5. *Estimación del modelo con la variable TRAT.*

VARIABLE	ESTIMACIÓN	ERROR ESTÁNDAR
Constante	-0.302	0.064
TRAT	-0.145	0.091

Como ya se sabe, el coeficiente -0,145 de la variable tratamiento significa que una estimación de la ventaja de vivir tras el tratamiento A es $e^{-0,145}$=0,865 veces la de B, pues al tratamiento A se le asignó el valor 1 y al tratamiento B el 0; como el lector puede comprobar, esta *OR* de 0,865 es la razón de ventajas que resulta de la Tabla 2.4 colapsando por la variable hospital. Al ser esta razón de ventajas inferior a la unidad significa que es más probable sobrevivir si se recibió en tratamiento B que el A, en concreto, 1/0,865=1,16 veces más probable.

Consideremos ahora el modelo que permite evaluar la misma cuestión pero controlando por el posible efecto que pueda tener el hospital donde recibió el tratamiento, codificado como 0 para el hospital H1 y 1 para el H2; el modelo ajustado correspondiente para esta situación da lugar a una lejanía = 0,0000 con 1 grado de libertad y las estimaciones que aparecen en la Tabla 2.6

Tabla 2.6. *Estimación del modelo con las variables TRAT y HOSP.*

VARIABLE	ESTIMACIÓN	ERROR ESTÁNDAR
Constante	0,000	0,069
TRAT	1,099	0,140
HOSP	-1,946	0,141

Lo primero que salta a la vista es el cambio tan sustancial ocurrido en la estimación del coeficiente correspondiente a la variable tratamiento; ha pasado de valer -0,145 a ser 1,099, por el hecho de tener en cuenta el hospital donde se recibió el tratamiento. De esta forma, ajustando por el hospital, la razón de ventajas del tratamiento A respecto del B pasa a ser $e^{1.099}=3$, es decir, cuando no se controla por la variable hospital, el tratamiento B es superior, en términos de supervivencia, al tratamiento A, resultado diferente a cuando no se tiene el cuenta el hospital donde se recibió el tratamiento. Resumiendo, la relación entre el tipo de tratamiento y la supervivencia de los pacientes depende de que se tenga o no en cuenta una tercera variable, el hospital donde se recibió el tratamiento; en esta circunstancia se dice que la variable hospital confunde o que es una *variable confundente (confounding variable)* para la asociación entre supervivencia y tratamiento. Se puede comprobar que 3 es precisamente la razón de ventajas en cada uno de los dos hospitales; en efecto, según la Tabla 2.4, para el hospital H1, la estimación de la *OR* del tratamiento A respecto al B es

$$\frac{(150)\ (400)}{(50)\ (400)} = 3$$

y para el hospital H2

$$\frac{(175)\ (240)}{(25)\ (560)} = 3$$

Las condiciones para que una variable, en este caso el hospital donde se recibe el tratamiento, sea confusora para la relación entre la supervivencia y el tratamiento recibido, es que aquella esté asociada con cada una de estas y, además, no se encuentre en el camino causal de estas dos. En nuestro ejemplo, el hospital está relacionado con el tipo de tratamiento que reciben los pacientes pues, si no distinguimos entre vivos y muertos, obtenemos la Tabla 2.7 en donde se observa que el tratamiento A predomina en el hospital H2 y el tratamiento B en el hospital H1.

Tabla 2.7. *Tipo de tratamiento según el hospital.*

	H1	H2
Trat. A	200	800
Trat. B	800	200

De igual forma, el tipo de hospital está relacionado con la supervivencia de los enfermos; en efecto, si no distinguimos el tipo de tratamiento, de la Tabla 2.4 podemos generar la Tabla 2.8.

Tabla 2.8. *Supervivencia según el hospital.*

	H1	H2
Vivo	550	265
Muerto	450	735

En esta se observa que en el hospital H2 existe un mayor riesgo de muerte que el hospital H1. En definitiva, la variable hospital donde se recibe el tratamiento está asociada tanto con el tratamiento como con la supervivencia. El hecho de que la relación para la tabla marginal no solo cambie sino que además tenga sentido contrario a la que se observa en la tablas parciales se conoce con el nombre de *paradoja de Simpson.*

Mantel y Haenszel propusieron un método para combinar diferentes estimaciones de una misma *OR* estimada en varios estratos con el objeto de controlar el posible efecto confusor de la variable por la que se estratifica; el estadístico propuesto por estos autores es una suma ponderada de los estimadores de la *OR* en cada estrato, siendo los pesos asignados inversamente porporcionales a la varianza del logaritmo de cada estimador. El estimador de Mantel-Haenszel toma la forma

$$OR_{MH} = \frac{\sum \frac{a_i d_i}{n_i}}{\sum \frac{b_i c_i}{n_i}}$$

donde la suma se extiende a todos los estratos y a_i, b_i, c_i y d_i son los valores de las casillas en cada estrato y n_i es el número de individuos en ese estrato. Para nuestro ejemplo, para el primer estrato, el hospital H1, a_1=150, b_1=50, c_1=400, d_1=400 y para el segundo estrato, el hospital H2, a_2=240, b_2=560, c_2=25 y d_2=175, con $n_1 = n_2$ =1000, por lo que

$$OR_{MH} = \frac{\frac{(150)\ (400)}{1000} + \frac{(240)\ (175)}{1000}}{\frac{(50)\ (400)}{1000} + \frac{(560)\ (25)}{1000}} = 3$$

que, como se observa, coincide con la estimación dada por el modelo logístico cuando se controla por el tipo de hospital. Para este ejemplo concreto, ya que las estimaciones en los dos estratos coindicen exactamente, ese valor común es, evidentemente, el valor del estadístico de Mantel-Haenszel.

2.6 Interacción entre las predictoras

En la situación que se acaba de describir la relación entre la variable resultado, supervivencia del paciente, y la predictora de interés, el tratamiento recibido, es la misma en cada una de las categorías de la otra variable considerada, el hospital donde recibió el tratamiento. Sin embargo, en algunas ocasiones esto no es así y la relación entre dos variables puede cambiar dependiendo del nivel de una tercera. Consideremos los datos hipotéticos de la Tabla 2.9, de un estudio de casos y controles sobre el riesgo de cáncer de pulmón, según el nivel de exposición al asbesto; el tabaco se consideró como covariable, es decir, variable por la que controlar la relación de interés.

Tabla 2.9. *Cancer de pulmón según consumo de tabaco y exposición al asbesto.*

	FUMA			
	Sí		No	
	ASBESTO			
	Sí	No	Sí	No
Casos	600	200	100	100
Controles	50	100	100	200

Para construir el fichero de datos notemos por 1 a los casos y 0 a los controles, 1 para los fumadores y 0 para los que no fumadores y, por último, 1 a los expuestos al asbesto y 0 en caso contrario. El modelo para el asbesto da lugar a una lejanía = 230,35 con 2 grados de libertad y unas estimaciones que aparecen en la Tabla 2.10.

Tabla 2.10. *Estimación del modelo que solo contiene a ASBE.*

VARIABLE	ESTIMACIÓN	ERROR ESTÁNDAR
Constante	0,000	0,082
ASBE	1,540	0,121

Si a este modelo se añade la variable *FUMA*, se obtiene una lejanía = 17,038 con 1 grado de libertad y las estimaciones de la Tabla 2.11

Tabla 2.11. *Estimación de los coeficientes de ASBE y FUMA.*

VARIABLE	ESTIMACIÓN	ERROR ESTÁNDAR
Constante	-0,937	0,112
ASBE	1,251	0,132
FUMA	1,874	0,132

Este modelo tiene una lejanía muy grande para 1 grado de libertad lo que se puede interpretar, al disponer de datos agrupados, como que el modelo que contiene a esas dos variables no explica suficientemente bien los datos observados.

En el modelo modelo ajustado de la Tabla 2.12 está implícita la suposición de que el efecto del asbesto es el mismo para los fumadores que para los no fumadores, pues antes se dijo que el coeficiente 1,251 del asbesto indica una estimación del logaritmo de la *OR* en dos individuos, uno expuesto al asbesto y el otro no, pero que son iguales respecto a la otra variable presente en el modelo, en este caso, la variable *FUMA*; es decir, ese riesgo es el mismo tanto si los dos individuos son fumadores como si los dos no lo son. Sin embargo, es posible que la relación entre el cáncer del pulmón y el asbesto sea diferente en los fumadores y en los no fumadores. Si acaso ocurriese esta situación se dice que existe *interacción* entre el asbesto y el tabaco indicando con ello que la relación entre asbesto y cancer de pulmón *depende* de la categoría de la variable *FUMA*. La pregunta que ahora se plantea es, ¿cómo modelar esta nueva situación?

Para ello definimos una nueva variable mediante el producto de *ASBE* y *FUMA* y establecemos el nuevo modelo

$$logit(p) = \beta_0 + \beta_1 ASBE + \beta_2 FUMA + \beta_3 (ASBE)(FUMA)$$

donde aparecen tanto los denominados *efectos principales*, los términos correspondientes a *ASBE* y a *FUMA*, como el término de interacción, el correspondiente al producto (*ASBE*)(*FUMA*).

Examinemos detenidamente este nuevo modelo; para los individuos fumadores, *FUMA*=1, el modelo que contempla la interacción toma la forma

$$logit(p) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) ASBE$$

y para los no fumadores, *FUMA*=0,

$$logit(p) = \beta_0 + \beta_1 ASBE$$

por lo que el modelo propuesto contempla la posibilidad de una distinta asociación entre asbesto y cáncer en cada una de las categorías de la variable *FUMA*, pues

$$OR_{Fumadores} = e^{\beta_1 + \beta_3} \qquad OR_{No\ fumadores} = e^{\beta_1}$$

ya que $\beta_1+\beta_3$ es el coeficiente del asbesto para el modelo de los fumadores, siendo β_1 el correspondiente para el modelo de los no fumadores.

Si β_3 es distinto de cero, el logaritmo de la *OR* para los fumadores viene dado por la cantidad $\beta_1+\beta_3$ que es distinto a β_1, el logaritmo de la *OR* correspondiente a los

no fumadores; si β_3 es mayor que cero el riesgo del asbesto será mayor en los fumadores que en los no fumadores y en el caso en que β_3 sea menor que cero, el riesgo será menor. Si β_3 es cero, el modelo se reduce al que solamente tiene en cuenta los efectos principales, por lo que la asociación de interés es la misma en los dos estratos de la variable *FUMA*. De aquí se deduce que para evaluar la posible interacción entre asbesto y tabaco no hay más que contrastar la hipótesis de que el coeficiente β_3 correspondiente al término de interacción sea cero.

Ajustando el modelo que contiene tanto a los efectos principales como a la interacción obtenemos una lejanía igual a 0,00 con 0 grados de libertad y las estimaciones de la Tabla 2.12.

Tabla 2.12. *Estimación del modelo con la interacción de las variables ASBE y FUMA.*

VARIABLE	ESTIMACIÓN	ERROR ESTÁNDAR
Constante	-0,693	0,122
ASBE	0,693	0,187
FUMA	1,386	0,173
ASBE.FUMA	1,099	0,268

Ya que la diferencia en lejanía entre los modelos con y sin interacción es de 17,038-0 =17,038 para un grado de libertad, el modelo que contiene a la interacción explica mejor los datos que el modelo que no la contiene, es decir, podemos rechazar la hipótesis de que β_3 es cero; dicho de otra manera, el número 1,099, estimación de β_3, es suficientemente distinto, desde el punto de vista estadístico, de cero. Ya que el signo de esta estimación es positivo, podemos afirmar que el riesgo del asbesto es mayor entre los fumadores que entre los no fumadores, es decir, el tabaco potencia el efecto del asbesto.

Si consideramos la relación cancer de pulmón y asbesto sin controlar por el hecho de que el individuo fume o no fume, es decir, estimamos la *OR* según el primer modelo ajustado, tal estimación viene dada por la cantidad $e^{1.54}$=4,665, pero considerando el tercer modelo

$$logit(p) = -0{,}6931 + 0{,}6931ASBE + 1{,}386FUMA + 1{,}099(ASBE)(FUMA)$$

que es el que se ajusta bien a los datos observados, el efecto del asbesto depende de que individuo fume o no fume; así, para los individuos no fumadores, una estimación de la *OR* es $e^{0,6931}$=2,00, mientras que para los fumadores la *OR* viene estimada por la cantidad $e^{0,6931+1,099}$=6. El tabaco triplica el riesgo asociado al asbesto.

Obsérvese como la relación entre cáncer de pulmón y asbesto no es constante en las dos categorías de la variable *FUMA*; es decir, la relación entre esas dos variables depende del nivel de la variable *FUMA*; por tanto el coeficiente de la variable *ASBE* deja de tener la interpretación que hasta ahora se venía manteniendo, pues ahora esa relación depende de la categoría de la variable *FUMA*; a una variable con las características de la variable *FUMA* se le denomina *modificadora del efecto*, en este caso del asbesto. Facilmente se puede comprobar que las tres estimaciones anteriores son las que se derivan de la Tabla 2.9 antes presentada; el valor 4.665, salvo errores de redondeo, resulta de colapsar dicha tabla por la variable *FUMA*, como aparece en la Tabla 2.13.

Tabla 2.13. *Resultado de colapsar la Tabla 2.9 por la variable FUMA.*

	Asbesto sí	Asbesto no
Casos	700	300
Controles	150	300

$$\frac{(700)\ (300)}{(150)\ (300)} = 4{,}667$$

Las otras dos estimaciones correspondientes a los riesgos asociados al asbesto en los no fumadores y en los fumadores aparecen en la Tabla 2.14 y la Tabla 2.15, respectivamente.

Tabla 2.14. *Cáncer de pulmón y exposición al asbesto en no fumadores.*

	Asbesto +	Asbesto -
Casos	100	100
Controles	100	200

$$\frac{(100)\ (200)}{(100)\ (100)} = 2$$

Tabla 2.15. *Cáncer de pulmón y exposición al asbesto en fumadores.*

	Asbesto +	Asbesto -
Casos	600	200
Controles	50	100

$$\frac{(600)\ (100)}{(200)\ (50)} = 6$$

A partir del modelo de la Tabla 2.12 también se puede estimar la *OR* para un fumador expuesto al asbesto respecto a un no fumador no expuesto al asbesto; en efecto, como se vió en el Apartado 2.3, el componente sistemático para un fumador, $FUMA=1$, expuesto al asbesto, $ASBE=1$, será

$$-0{,}693 + 0{,}693(1) + 1{,}386(1) + 1{,}099(1)(1)$$

mientras que para uno no fumador, $FUMA=0$, y no expuesto al asbesto $ASBE=0$,

$$-0{,}693 + 0{,}693(0) + 1{,}386(0) + 1{,}099(0)(0)$$

por lo que su diferencia, 0.693 + 1.386 + 1.099, no es más que el logaritmo de la razón de ventajas del primero respecto del segundo. Por tanto, el riesgo, razón de ventajas, del primero respecto al segundo individuo es $e^{0.6931+1.386+1.099}=24$, que también se puede estimar a partir de la Tabla 2.16.

Tabla 2.16. *Distribución de dos categorías de exposición en casos y en controles.*

	Asbesto sí Fumadores	Asbesto no No fumadores
Casos	600	100
Controles	50	200

$$OR = \frac{(600)\ (200)}{(100)\ (50)} = 24$$

El modelo con el componente de interacción que se acaba de ajustar es un nuevo ejemplo de modelo saturado pues ahora se disponía de cuatro binomiales, cuatro grados de libertad, y dicho modelo tiene cuatro parámetros a estimar; es por eso que la lejanía vale cero, así como sus grados de libertad asociados.

2.7 Ejemplo. Contraceptivos orales e infarto de miocardio

A continuación se presentan en la Tabla 2.17 los datos de un estudio caso-control diseñado con al objeto de evaluar la posible relación que pudiera existir entre el uso de contraceptivos orales (AO) y el infarto de miocardio, Shapiro (1979).

Tabla 2.17. *Casos de infarto según uso de anticonceptivos, edad y tabaco.*

			GEDA				
FUMA	AO		1	2	3	4	5
0	0	Caso	1	0	3	10	20
		Control	106	175	153	165	155
	1	Caso	0	0	0	1	3
		Control	25	13	8	4	2
1	0	Caso	0	5	11	21	42
		Control	79	142	119	130	96
	1	Caso	1	1	1	0	0
		Control	25	10	11	4	1
2	0	Caso	1	7	19	34	31
		Control	39	73	58	67	50
	1	Caso	3	8	3	5	3
		Control	12	10	7	1	2

Como los autores del estudio pensaban que esta relación podía venir confundida por no controlar algunas variables, también registraron para cada una de las mujeres variables tales como la edad, el número de cigarrillos que fumaban diariamente, etc.; nosotros tan sólo vamos a tratar de controlar por estas dos variables citadas. La variable *FUMA* correspondiente al tabaco está codificada así: 0 para las no fumadoras, 1 para fumadoras entre 1 y 24 cigarrillos diarios y 2 para las que fuman 25 o más cigarrillos diarios. La segunda variable corresponde al uso de anticonceptivos orales indicando el 1 su utilización y 0 lo contrario; por último, la variable *GEDA* es la edad codificada como sigue: 1 corresponde al grupo de edad entre 25 y 29 años, 2 a las mujeres entre 30 y 34 años, 3 a las que tienen entre 35 y 39, 4 entre 40 y 44 y 5 entre 45 y 49 años. De todas formas, tomaremos como *EDAD* de las mujeres un representante del grupo de edad al que pertenecen: así, supondremos que todas las mujeres del primer grupo tienen 27 años, 32 las del segundo, etc.

El interés del estudio está en la evaluación de los anticonceptivos orales como factor de riesgo para el infarto de miocardio; una primera aproximación al problema sería medir la asociación cruda, es decir, sin tener en cuenta ningún otro factor, entre el infarto y los anticonceptivos. Para ello ajustamos un modelo logístico que contiene sólo a los anticonceptivos

Tabla 2.18. *Estimación del modelo que solo contiene a los anticonceptivos.*

VARIABLE	ESTIMACIÓN	ERROR ESTÁNDAR
Constante	-2.059	0.074
AO	0.521	0.218

que da lugar a una lejanía de 302,34 y 28 grados de libertad y unas estimaciones que están en la Tabla 2.18, de donde se puede, en principio, deducir que el riesgo de padecer infarto de miocardio entre las mujeres que toman anticonceptivos es $e^{0.521} = 1{,}684$ veces superior al riesgo que tienen las mujeres que no los toman. El lector puede comprobar como a partir de la Tabla 2.17, agregando por edad y tabaco, se puede construir la Tabla 2.19, la tabla marginal.

Tabla 2.19. *Los mismos datos que en la Tabla 2.17 colapsada por FUMA y GEDA.*

		INFARTO	
		Sí	No
AO	Sí	29	135
	No	205	1607

que efectivamente da lugar a una razón de ventajas de 1,684. Evaluar la significación de la *OR* de la que 1,684 es una estimación, se puede conseguir comparando las lejanía de este modelo y la del que no contiene ninguna predictora; esta última lejanía es 307,58 con 29 grados de libertad, por lo que la diferencia 307,58- 302,34=5,24 es significativa para una χ^2 con 1 grado de libertad; por lo que, en principio y con muchas reservas, podríamos hablar de la asociación entre el infarto de miocardio y la utilización de los anticonceptivos orales; luego volveremos sobre esto.

2.8 Variables indicadoras

Hasta este momento solo hemos utilizado como predictoras a variables dicotómicas a las que hemos asignado valores 0 y 1 para sus dos categorías para poder tratar informáticamente el fichero de datos; pero, ¿ porqué esta asignación de valores y no otra distinta? En realidad, asignando dos valores enteros positivos consecutivos, por ejemplo, 1 y 2 ó 34 y 35, la interpretación de β_1 sigue siendo la misma pues

$$e^{\beta_1(35-34)} = e^{\beta_1(2-1)} = e^{\beta_1(1-0)} = e^{\beta_1}$$

Aunque los valores 0 y 1 son los que se utilizan mas frecuentemente, también son posibles otras asignaciones, por ejemplo, -1 y 1; lo que pasa es que ahora el riesgo del segundo individuo respecto al primero es

$$e^{\beta_1(1-(-1))} = e^{2\beta_1}$$

por lo que, con esta asignación de valores, sería erróneo decir que el riesgo es e^{β_1}.

La variable *FUMA*, aunque en principio se podía haber tratado también como numérica, expresada como número de cigarrillos, los autores del trabajo la categorizaron como se explicó anteriormente; alguien podría pensar que de lo que se trata es de asignar los valores, por ejemplo, 0, 1 y 2 a las tres categorías y tratarla como numérica. Sin embargo, esa estrategia no resiste una mínima crítica; en efecto, los valores asignados a esta variable lo han sido con la única intención de identificar a los tres grupos de mujeres en su relación con el tabaco, pero nada más; así, una mujer con valor 2 no significa que fume el doble que una mujer con valor 1; por tanto, estas etiquetas 0, 1 y 2 no tienen sentido estrictamente numérico; ¿cómo tratar entonces a una predictora categórica?

En situaciones como ésta en que se dispone de una variable categórica ordinal en la que la asignación propiamente numérica no es clara en absoluto, o en el caso de una variable no ordinal, el camino a seguir para tratarla consiste en construir, a partir de ella, un conjunto de variables indicadoras (*dummy variables*). Antes se vió que en el caso de una variable dicotómica, una variable con dos posibles valores podía identificar perfectamente a todos los individuos; en el caso de la variable *FUMA* que tiene tres categorías podemos identificarlos mediante dos variables indicadoras *FUMA(1)* y *FUMA(2)* definidas como sigue: una mujer no fumadora vendría identificada por la pareja de valores (0,0), las fumadoras entre 1 y 24 cigarrillos corresponderían al par (1,0) y la pareja (0,1) correspondería a las fumadoras de más de 24 cigarrillos, como aparece en la Tabla 2.20.

Tabla 2.20. *Construcción de dos variables indicadoras para la variable FUMA.*

		FUMA(1)	*FUMA(2)*
FUMA	0	0	0
	1	1	0
	2	0	1

En general ante una variable categórica con k niveles habría que construir k-1 variables indicadoras; como puede adivinarse esto puede ser molesto a la hora de construir la base de datos. Por tanto, en lugar de construir el fichero de datos con los valo-

res 0, 1, 2 de la variable *FUMA*, habría que introducir los valores correspondientes de las variables *FUMA(1)* y *FUMA(2)*; más incómodo sería el caso de una predictora categórica con 5 categorías pues tendríamos que definir 4 variables indicadoras. Afortunadamente, la mayoría de los programas de regresión logística disponen de órdenes que permiten la construcción de variables indicadoras sin tener que construirlas físicamente en el fichero de datos.

Aunque no estaba en los objetivos del estudio analicemos la asociación entre infarto y la variable FUMA; para ello se ajusta el modelo

$$logit(p) = \beta_0 + \beta_1 FUMA(1) + \beta_2 FUMA(2)$$

que da lugar a una lejanía de 184,51 con 27 grados de libertad pues ahora se han estimado tres coeficientes, que aparecen en la Tabla 2.21. Como el modelo que no contiene predictoras tenía una lejanía de 307,58 con 29 grados de libertad, la diferencia 307,58-184,51=123,07 es una valor muy grande para la distribución chi-cuadrado con 29 - 27=2 grados de libertad; es decir, existe asociación significativa entre el tabaco y el infarto. Pero ¿cómo interpretar los coeficientes de las variables indicadoras utilizadas?

Tabla 2.21. *Estimación del modelo para la variable FUMA.*

VARIABLE	ESTIMACIÓN	ERROR ESTÁNDAR
Constante	-3,054	0,166
FUMA(1)	1,036	0,203
FUMA(2)	2,026	0,198

Ya que el modelo estimado es

$$logit(p) = -3{,}054 + 1{,}036 FUMA(1) + 2{,}026 FUMA(2)$$

para el caso de una no fumadora, es decir, una mujer para la que los dos variables indicadoras valen cero, $FUMA(1)=FUMA(2)=0$, tendremos

$$logit(p) = -3{,}054$$

y para una fumadora entre 1 y 24 cigarrillos, $FUMA(1)=1$ y $FUMA(2)=0$, la estimación del modelo será

$$logit(p) = -3{,}054 + 1{,}036$$

con lo que la diferencia entre este logit y el anterior, es decir, el logaritmo del riesgo de la fumadora entre 1 y 24 cigarrillos respecto a la no fumadora es (-3,054+1,036)

- (-3,054) =1,036. De esta manera el coeficiente 1,036 de la variable *FUMA(1)* corresponde al logaritmo de la razón de ventajas de sufrir infarto entre las fumadoras de entre 1 y 24 cigarrillos respecto a las no fumadoras; es decir, este tipo de fumadoras está a $e^{1,036}=2,82$ veces más riesgo de sufrir infarto que las no fumadoras. De la misma forma, el coeficiente 2,026 de la variable *FUMA(2)* indica que las fumadoras de más de 24 cigarrillos están a $e^{2.026}=7,58$ veces más riesgo de sufir infarto que las no fumadoras. Para comparar una fumadora de más de 24 cigarrillos con una que fume entre 1 y 24 no hay más que calcular sus logits que son -3,054+2,026 y -3,054+1,036 y restarlos; esa diferencia es 2,026-1,036, por lo que el riesgo entre esos dos tipos de mujeres será $e^{2.026-1,036}=2,69$.

La Tabla 2.22 muestra la distribución de los casos y controles respecto a la variabla FUMA, así como los logits para cada categoría; la diferencia entre los correspondientes a las categorías 1 y 0 es -2,018-(-3,054)=1,036, por tanto, ya que la diferencia entre los logits es el logaritmo de la *OR*, 1,036 será el logaritmo de la *OR* entre una fumadora de entre 1 y 24 cigarrillos y una no fumadora: esta es la misma solución que se desprende de la Tabla 2.22.

Tabla 2.22. *Ventajas y logit de sufrir infarto para las distintas categorías de fumadoras.*

		INFARTO			
		Sí	No	Ventaja	logit(p)
FUMA	0	38	806	0,04715	-3,054
	1	82	617	0,13290	-2,018
	2	114	319	0,35737	-1,029
	Total	234	1742		

Con la asignación de valores realizada, los coeficientes de las variables indicadoras no son más que el logaritmo del riesgo de cada categoría respecto a las no fumadoras, denominada *categoría de referencia*, que obsérvese que es la categoría que tiene valor 0 para todas las variables indicadoras. Por tanto, ante una predictora categórica debemos decidir cual de sus categorías queremos tomar como referencia y esta será la que tenga valor cero para todas las variables indicadoras; es aconsejable tomar como categoría de referencia la categoría más numerosa por razones de fiabilidad estadística, ya que si tomásemos una categoría con pocas observaciones, todas las comparaciones con las categorías restantes van a ser poco fiables.

Según la Tabla 2.21, ya que las no fumadoras son la categoría de referencia, evidentemente está a riesgo uno respecto de sí misma, o lo que es igual, tiene una *OR* de 1, por lo que su logaritmo es 0; para las fumadoras entre 1 y 24 cigarrillos, una estimación del logaritmo de la *OR* es 1,036 y 2,026 para las fumadoras de más de 24

cigarrillos; es decir, existe aproximadamente la misma diferencia entre las no fumadoras y las fumadoras entre 1 y 24 cigarrillos (1,036-0), que entre éstas y las fumadoras de más de 24, (2,026-1,036); esto hace pensar que los valores asignados 0, 1, 2, a la variable *FUMA* se podrían, en principio, tratar como valores numéricos. Lo que se acaba de sugerir puede comprobarse gráficamente, como aparece en la Figura 2.1, representando los logits para las tres categorías de *FUMA*.

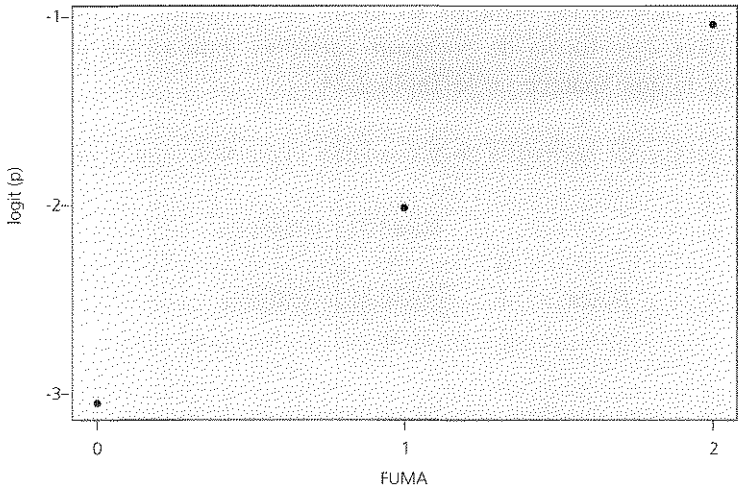


Figura 2.1. *Relación entre logit(p) y FUMA.*

Considerando a la variable *FUMA* como numérica, con valores 0,1 y 2, y ajustando el modelo correspondiente sólo aparecerá un coeficiente estimado para esta variable, pues ahora se considera como numérica; las estimaciones de este nuevo ajuste se muestran en la Tabla 2.23 con una lejanía de 184,54 y 28 grados de libertad; el coeficiente 1,009 es el aumento en logit por unidad de aumento de la variable *FUMA*.

Tabla 2.23. *Estimación del modelo considerada como numérica la variable FUMA.*

VARIABLE	ESTIMACIÓN	ERROR ESTÁNDAR
Constante	-3,041	0,1393
FUMA	1,009	0,0956

Así, $e^{(1-0)1,009}$=2,74 es la *OR* para las fumadoras de entre 1 y 24 cigarrillos respecto a las no fumadoras y $e^{(2-0)1,009}$=7,52 es la *OR* para las fumadoras de más de 24 también respecto a las no fumadoras. Las lejanías de los dos modelos, el que contiene a *FUMA* como categórica, mediante las dos variables indicadoras, y el que la contempla como numérica, tan solo se diferencian, en términos de lejanía, en 184,54-184,51 =0,03,

pero, sin embargo, se gana un grado de libertad al considerar *FUMA* como variable numérica ya que se estima un parámetro menos. Dicho de otra manera, como los dos modelos dan los mismos riesgos y el mismo grado de ajuste y éste último contiene menos parámetros a estimar, es el que debemos de elegir en virtud del principio de parsimonia.

La codificación utilizada anteriormente para las variables indicadoras es, con diferencia, la más utilizada en los estudios en salud; sin embargo, no es la única. Una alternativa, es decir, otra forma de construir variables indicadoras para la variable *FUMA* podría ser la que aparece en la Tabla 2.24

Tabla 2.24. *Construcción alternativa de dos variables indicadoras para la variable FUMA.*

		FUMA(1)	*FUMA(2)*
	0	-1	-1
FUMA	1	1	0
	2	0	1

que, como se verá a continuación, supone un cambio en la interpretación de los coeficientes del modelo. En efecto, dado el modelo

$$logit(p) = \beta_0 + \beta_1 FUMA(1) + \beta_2 FUMA(2)$$

con *FUMA(1)* y *FUMA(2)* definidas como aparece en la Tabla 2.24, el logit para una no fumadora, *FUMA(1)=FUMA(2)*=-1, será

$$\beta_0 + \beta_1(-1) + \beta_2(-1)$$

y para una fumadora entre 1 y 24 cigarrillos su logit será

$$\beta_0 + \beta_1(1) + \beta_2(0)$$

por lo que restando de este logit el anterior tendremos el logaritmo de la razón de ventajas

$$2\beta_1 + \beta_2$$

por lo que el riesgo de una fumadora de 1 a 24 cigarrillos respecto a una no fumadora será

$$OR = e^{2\beta_1 + \beta_2}$$

De forma similar, el riesgo de una fumadora de más de 24 cigarrillos respecto a una no fumadora será

$$OR = e^{\beta_1 + 2\beta_2}$$

El modelo estimado para esta nueva definición de las variables indicadoras resulta ser

$$logit(p) = -2{,}034 + 0.016FUMA(1) + 1{,}005FUMA(2)$$

por lo que, según se acabar de ver, los riesgos de una fumadora entre 1 y 24 y una de más de 24 cigarrillos, respecto a la no fumadora serán, respectivamente

$$e^{2(0{,}016)\,+\,1.005} = 2{,}82 \qquad e^{0{,}016\,+\,2(1{,}005)} = 7{,}58$$

que son los mismos que se calcularon anteriormente cuando las variables indicadoras se definieron de forma distinta. Igual que con la primera codificación el coeficiente de una variable indicadora era el logit de la categoría correspondiente respecto a la categoría a la que se asignaron valores 0 para todas las indicadoras, la categoría de referencia, en el segundo esquema se puede demostrar fácilmente que el coeficiente de una variable indicadora, por ejemplo 0.016, es el logit de las fumadoras de 1 a 24 cigarrillos respecto a la *categoría promedio* de las tres categorías.

Lo que sí es algo más complicado con esta definición de indicadoras es el cálculo de los intervalos de confianza; en efecto, ahora el logaritmo de la razón de ventajas para las fumadoras entre 1 y 24 cigarrillos es ahora $2\beta_1 + \beta_2$ por lo que su error estandar no es tan evidente como en el primer caso en que el logaritmo del riesgo era β_1. En el apartado 2.10 se muestra el método de cálculo del error estandar para situaciones como esa.

Existen otras formas de definir variables indicadoras que no vamos a comentar pero lo que sí es muy importante es ser conscientes del esquema adoptado, pues según se acaba de demostrar, la interpretación de los coeficientes depende de éste y los programas informáticos ofrecen distintas alternativas. De todas formas, ya que en la mayoría de los estudios epidemiológicos el interés se centra en estudiar el riesgo de los distintos grados de exposición respecto a la no exposición, el tipo de asignación de la Tabla 2.20 es la más utilizada en la literatura.

2.9 Estrategias de selección de variables

Habitualmente son varias las variables que se miden en cualquier estudio con la finalidad de evaluar su posible carácter confusor de la relación entre el factor de riesgo y la enfermedad. Ya que casi siempre existe un número bastante limitado de individuos a observar, el número de variables a entrar en el modelo está obligado a ser pequeño, de lo contrario las estimaciones de los coeficientes del modelo son poco precisas lo que se traduce en errores estándar muy grandes; aparte de esta cuestión está el hecho de que de lo que se trata no es tanto de elegir el modelo de mejor ajuste sino

más bien de un modelo que explique sufucientemente bien los datos y además sea suficientemente simple de interpretar; de poco vale un ajuste perfecto con una imposible interpretabilidad. Subyacente a estos hechos está la cuestión de qué variables introducir como confundentes y como términos de interacción en el modelo. Como el objetivo del estudio estaba centrado en la evaluación del uso de anticonceptivos como factor de riesgo para el infarto de miocardio, tanto la edad como el tabaco se midieron con el único objeto de controlar su posible efecto confusor. La cuestión que ahora se plantea es, ¿ qué criterio seguir para incluir estas variables en el modelo multivariante con el objeto de controlar la posible confusión?

En epidemiología el problema de la selección de las variables posibles confundentes es un debate que dista mucho de estar cerrado; mientras algunos autores como Fleiss (1986) son defensores de la utilización de los tests de hipótesis para seleccionar las variables confundentes, hay otros, como por ejemplo Miettinen (1976), que proponen como criterio de selección la repercusión en la estimación del efecto de la exposición.

En primer lugar es necesario distinguir dos tipos de variables: por una parte el posible factor de riesgo y las variables que por estudios anteriores se saben que son factores confundentes, y por otra, el resto de las variables. El primer grupo de variables deben entrar automáticamente en el modelo; los factores confundentes conocidos deben entrar en el modelo, independientemente de la significación estadística de su asociación, si tal inclusión cambia la estimación del efecto del factor de riesgo, Breslow (1980). Por otra parte, las variables del segundo grupo son las que se someten al proceso de selección.

El criterio de selección basado en los tests de hipótesis propone seleccionar las variables en función de la significación estadística de su asociación con la enfermedad, lo que puede llevarse a cabo siguiendo diferentes estrategias. El método de *selección hacia adelante (forward selection)* consiste en, una vez construido el modelo con las variables que obligatoriamente deben entrar, de entre las restantes variables se selecciona aquella que, introducida en el modelo, de lugar a una estimación más significativa; si tal estimación es estadísticamente significativa para un error, por ejemplo, de 0.15, la variable se introduce en el modelo. Este proceso se continua hasta que no haya variable que cumpla con la condición anterior. En el método de *selección hacia atrás (backward selection)*, se incluyen todas las variables en el modelo inicial, eligiendo aquella que tenga un coeficiente menos significativo; si su error correspondiente es mayor que el previamente elegido, se saca la variable del modelo; este proceso se repite hasta que no haya variables que cumplan el requisito anterior. Una combinación de estos dos algoritmos se conoce como método de *selección paso a paso (stepwise selection)*.

El otro criterio de seleccionar variables a incluir en el modelo está basado en el cambio producido en la estimación del efecto del factor de riesgo por el hecho de introducir una nueva variable en el modelo; en este caso, a partir de un modelo inicial, la estrategia hacia adelante selecciona como primera variable aquella que al introducirla en el modelo cause un mayor cambio en la estimación del efecto del factor de riesgo; si este cambio es mayor que uno preestablecido, por ejemplo, un 10% de cambio sobre la *OR*, se introduce la variable; este proceso se continúa hasta que no haya variables que produzcan al menos ese cambio. En el método hacia atrás, a partir del modelo que contiene a todas las variables se selecciona la variable que al quitarla produzca un cambio más pequeño; si este cambio no alcanza al valor prefijado, la variable se omite del modelo. Este procedimiento se sigue hasta que no se pueda sacar ninguna variable.

Schall y Zucchini (1990) piensan que, en principio, toda variable a seleccionar debería figurar como modificador del efecto en el modelo que ellos denominan *modelo operativo*. Aunque este modelo es el que mejor describe las observaciones, es el que contempla más parámetros a estimar y por tanto las estimaciones correspondientes tienen menos precisión, mayor error estándar; cualquier modelo, caso particular del modelo operativo, tendrá menos parámetros a estimar y sus estimaciones tendrán errores estandars más pequeños. El método que estos autores proponen consiste en elegir el modelo que explique suficientemente bien las observaciones y que dé unas estimaciones de las *OR* lo más precisas posible; para ello proponen una medida de discrepancia entre modelos. Como se ha dicho antes, actualmente hay un gran debate acerca de estos dos criterios de selección de variables con el objeto de medir el efecto del factor de riesgo; son mayoría los partidarios del criterio del cambio en el efecto y parece ser que la inferioridad de los métodos basados en la significación estadística se debe a la baja potencia de los tests debido al casi siempre pequeño número de individuos observados. Mickey (1989), en uno de los pocos estudios comparativos de los dos criterios, encuentran que el criterio basado en el cambio del efecto tiende a ser mejor aunque señalan que los métodos basados en los tests pueden comportarse aceptablemente tomando niveles de significación del orden del 20% o superior.

Estos métodos de seleccionar el modelo final basados en criterios clásicos como la lejanía, el criterio de información de Akaike, u otro estadístico cualquiera no están exentos de algunas dificultades, por lo que en algunas ocasiones han sido puestos en tela de juicio. Argumentos que avalan estas dudas son los estudios experimentales que demuestran que en ocasiones, y especialmente con muestras grandes, se encuentran errores pequeños para el rechazo de una hipótesis que es verdadera, lo que conlleva declarar significativas asociaciones entre variables cuando no existe tal asociación. Por otra parte, téngase en cuenta que si son el número de predictoras, son 2^m el número de posibles modelos a elegir; en concreto, con veinte variables pasan del millón,

1.048.576, el número de modelos distintos, eso sin considerar ningún término de interacción; en situaciones como esta no es improbable encontrar varios modelos que expliquen los datos suficientemente bien y que entren en contradicción a la hora de explicar el fenómeno estudiado. Además, el modelo seleccionado puede variar según el criterio de selección de las variables.

De todo lo anterior se desprende un ambiente de incertidumbre a la hora de seleccionar un modelo. Una de las aproximaciones más interesantes a este problema viene desde lo que se conoce como Estadística Bayesiana, enfoque cada día más relavante en sus aplicaciones a la investigación en salud, Berry (1996). El criterio de selección de variables desde un enfoque bayesiano incorpora tal incertidumbre al análisis mediante la utilización del criterio de información bayesiano BIC (Bayesian Information Criterion); el artículo de Raftery (1995) es una excelente introducción al tema.

2.10 Ejemplo del infarto y los anticonceptivos: continuación

Volviendo al ejemplo del infarto y los anticonceptivos, si elegimos el método de selección hacia adelante, consideremos como modelo inicial al que solo contiene el posible factor de riesgo y añadamos la variable *EDAD*, en años, para evaluar su efecto confusor; para ello ajustamos el modelo

$$logit(p) = \beta_0 + \beta_1 AO + \beta_2 EDAD$$

modelo con una lejanía de 158,39, por lo que el cambio producido en la lejanía respecto al modelo inicial es de 143,95 con 1 grado de libertad que evidentemente es significativo y con unas estimaciones como las que aparecen en la Tabla 2.25.

Tabla 2.25. *Estimación del efecto de los anticonceptivos controlando por la edad.*

VARIABLE	ESTIMACIÓN	ERROR ESTÁNDAR
Constante	-7,764	0,501
AO	1,336	0,247
EDAD	0,142	0,013

Obsérvese el cambio producido en la estimación del efecto de los anticonceptivos; mientras en el modelo inicial, Tabla 2.18, la estimación del efecto, medido mediante la *OR*, de la toma de anticonceptivos era $e^{0,521} = 1,684$, ahora, ajustando por la edad, pasa a ser $e^{1,336} = 3,804$; es decir, entre dos mujeres con la misma edad, sea cual sea, la usuaria de anticonceptivos está a un riesgo de infarto 3,804 veces superior al riesgo de la no usuaria; por tanto, ha habido un aumento en la estimación del ries-

go del 125% por el hecho de controlar por la edad. En resumen, bajo los dos criterios de selección de variables, la *EDAD* tiene caracter confusor.

Si se hace el análisis correspondiente al tabaco, el modelo resultante da lugar a una lejanía de 182,54, por lo que la diferencia en lejanía respecto al modelo inicial, el que solo incluye a los anticonceptivos, es de 119,8, también con 1 grado de libertad; las estimaciones aparecen en la Tabla 2.26.

Tabla 2.26. *Estimación del efecto de los anticonceptivos contrlando por tabaco.*

VARIABLE	ESTIMACIÓN	ERROR ESTÁNDAR
Constante	-3,063	0,140
AO	0,329	0,227
FUMA	0,999	0,096

Aunque desde el punto de vista del criterio basado en los tests de hipótesis las dos variables se comportan de manera equivalente, ambas están significativamente asociadas al padecimiento del infarto, es evidente que la variable *EDAD* produce un cambio superior en la *OR* asociada a los anticonceptivos orales. Siguiendo los algoritmos antes descritos, la variable *EDAD* es la primera elegida.

El modelo que contiene a los tres efectos principales

$$logit(p) = \beta_0 + \beta_1 AO + \beta_2 EDAD + \beta_2 FUMA$$

da lugar a una lejanía de 34,877 con 26 grados de libertad y unas estimaciones que son las que aparecen en la Tabla 2.27.

Tabla 2.27. *Estimación del efecto de los anticonceptivos controlando por edad y tabaco.*

VARIABLE	ESTIMACIÓN	ERROR ESTÁNDAR
Constante	-9,252	0,626
AO	1,184	0,261
EDAD	0,152	0,014
FUMA	1,061	0,101

Una vez incluida la *EDAD*, la variable *FUMA* aporta información estadísticamente significativa pues hay una diferencia de lejanía entre este modelo y el anterior de 158,39-34,877 = 123,51 con 1 grado de libertad; si se estiman como importantes des-

censos superiores al 10%, el incluir *FUMA* produce un descenso en la estimación de la *OR* del 14% aproximadamente. Por tanto, también bajo los dos criterios hay que incluir, además de la *EDAD,* la variable *FUMA*.

Como ya se dijo, el modelo que solo contiene los efectos principales, implica que el riesgo de infarto con relación a los anticonceptivos es el mismo en las distintas categorías de las variables *EDAD* y *FUMA*; según los resultados de la Tabla 2.27, una usuaria de anticonceptivos orales está a un riesgo $e^{1.184}=3,27$ veces superior a una no usuaria, con la condición de que las dos sean iguales en cuanto al hábito de fumar, sea este el que sea, y de que tengan la misma edad, sea cual sea.

Sin embargo, podría ocurrir que el efecto de los anticonceptivos cambie dependiendo de las categorías de esas variables; por ejemplo, nos podríamos plantear preguntas del tipo: ¿entraña el mismo riesgo el uso de anticonceptivos en las mujeres de 25 a 29 que en las que tienen de 45 a 49 años?, o, ¿potencia el tabaco el riesgo de los anticonceptivos? Para estudiar estos problemas es necesario introducir las interacciones del factor de riesgo, los anticonceptivos, con las restantes variables; es costumbre considerar tan solo interacciones de primer orden, las construidas en base a productos de dos términos, pues las interacciones de orden superior son de difícil interpretación.

Veamos en primer lugar la interacción de los anticonceptivos con la edad, con el objetivo de estudiar si el riesgo de los anticonceptivos depende de la edad de la mujer; para ello añadimos al modelo anterior la nueva variable de interacción *(AO)·(EDAD)*, por lo que el nuevo modelo a ajustar sería

$$logit(p) = \beta_0 + \beta_1 AO + \beta_2 EDAD + \beta_2 FUMA + \beta_2 (AO)(EDAD)$$

que da lugar a una lejanía de 34,640 por lo que el cambio producido respecto al que aparece en la Tabla 2.27 es 34,877-34,640=0,237, que para un grado de libertad 1 es no significativo; es decir, no hay evidencias de que el riesgo de sufrir infarto por la toma de anticonceptivos cambie con la edad. En cuanto a la otra interacción, los anticonceptivos con el tabaco, definimos la variable *(AO).(FUMA)* y ajustado el modelo correspondiente da lugar a una lejanía 33,519 con un cambio de 1,358 con 1 grado de libertad que tampoco evidencia cambio del riesgo del uso de los anticonceptivos con el aumento del consumo del tabaco, P = 0.24. Por tanto, el modelo de elección es el que solo contiene el factor de riesgo y las dos variables confundentes.

Un intervalo de confianza aproximado, al 95% de confianza, para la *OR* de las usuarias respecto a las no usuarias de anticonceptivos es

$$e^{1.184 \pm 1,96(0,261)}=(1,96\ ,\ 5,45)$$

por lo que podemos afirmar que el riesgo de que una usuaria de anticonceptivos sufra infarto de miocardio es entre dos y cinco veces y media mayor que el riesgo de las no usuarias, controlando por edad y tabaco. De forma similar se pueden calcular los intervalos para las *OR* correspondientes a las otras dos variables aunque no tiene mucho sentido pues la variable de riesgo considerada es tan solo la toma de anticonceptivos figurando las otras dos variables en el modelo con el único fin de evitar el sesgo de confusión que supondría el no considerarlas.

De todas formas, y para ilustrar una situación que hasta ahora no hemos contemplado, calculemos un intervalo de confianza para el riesgo, según el modelo de la Tabla 2.27, de las fumadoras de más de 24 cigarrillos respecto a las no fumadoras; una estimación del logaritmo del riesgo vendrá dada por (2-0) $\hat{\beta}_3 = 2\hat{\beta}_3$, que tiene un error estandar igual a $2[e.e.(\hat{\beta}_3)]$, es decir, $2(0{,}101)=0{,}202$; de aquí que un intervalo para el logaritmo del riesgo sea

$$2(1{,}061) \pm 1{,}96\ (0{,}202) = (1{,}726\ ,\ 2{,}518)$$

por lo que el intervalo de confianza para el riesgo será

$$e^{(1{,}726\ ,\ 2{,}518)} = (5{,}62\ ,\ 12{,}40)$$

En general, dados dos individuos A y B que tienen valores x_{Ai}, x_{Bi} para la variable predictora X_i, el logaritmo del riesgo viene estimado por

$$(x_{Ai} - x_{Bi})\ \hat{\beta}_i$$

cuyo error estándar es

$$(x_{Ai} - x_{Bi})\ [e.e.(\hat{\beta}_i)]$$

por lo que un intervalo de confianza para el riesgo del individuo A respecto del B será

$$e^{(x_{Ai} - x_{Bi})\ \hat{\beta}_i \pm t_\alpha (x_{Ai} - x_{Bi})\ .\ [e.e.(\hat{\beta}_i)]}$$

que para el ejemplo anterior sería

$$e^{(3-1)\ 1{,}061\ \pm\ 1{,}96(3-1)\ \cdot\ (0{,}1010)} = (5{,}62\ ,\ 12{,}40)$$

Cuando, como en nuestro caso, el modelo no contiene términos de interacción, el cálculo de los intervalos de confianza para las *OR* no entraña dificultad alguna, como se acaba de ver. Imaginemos, sin embargo, que la interacción entre *AO* y *FUMA* hubiese dado significativa y que nuestro modelo de elección estimado hubiese sido el que aparece en la Tabla 2.28.

Tabla 2.28. *Estimación del modelo que incluye tanto los efectos principales como la interacción de la edad y el tabaco.*

VARIABLE	ESTIMACIÓN	ERROR ESTÁNDAR
Constante	-9,231	0,630
AO	0,631	0,575
EDAD	0,153	0,014
FUMA	1,022	0,106
AO.FUMA	0,399	0,354

¿Cómo calcular en esta situación los intervalos de confianza para las *OR* de los anticonceptivos en las distintas categorías de la variable *FUMA*? Para responder a esta pregunta hay que hacer uso de una propiedad acerca de la varianza de la suma de dos variables aleatorias; para nuestro ejemplo, según este modelo de interacción,

$$logit(p) = \beta_0 + \beta_1 AO + \beta_2 EDAD + \beta_3 FUMA + \beta_4 (AO)(FUMA)$$

o lo que es igual,

$$logit(p) = \beta_0 + (\beta_1 + \beta_4 FUMA)AO + \beta_2 EDAD + \beta_3 (FUMA)$$

por lo que la estimación del logaritmo de la *OR* para los anticonceptivos es $(\beta_1 + \beta_4 FUMA)$, expresión que, para un valor constante de la variable *FUMA*, depende de la suma de dos estimadores, es decir, variables aleatorias. Por otra parte, la varianza de una suma (diferencia) de dos variables es la suma de las varianzas de cada una de ellas más (menos) el doble de la covarianza, medida ésta que indica el cambio de una estimación cuando cambia la otra; es decir, si k es una constante,

$$var(\hat{\beta}_1 + k\hat{\beta}_4) = var(\hat{\beta}_1) + k^2\, var(\hat{\beta}_4) + 2k\, cov(\hat{\beta}_1\ ,\ \hat{\beta}_4)$$

por lo que, según los resultados de la Tabla 2.28,

$$var(\log(OR)) = 0{,}575^2 + (FUMA)^2\, 0{,}354^2 + 2(FUMA)\, cov(\hat{\beta}_1\ ,\ \hat{\beta}_4)$$

donde por $cov(\hat{\beta}_1\ ,\ \hat{\beta}_4)$ se representa la covarianza de las estimaciones de los parámetros de *AO* y *(AO)(FUMA)*; por tanto, para calcular la varianza de log(*OR*) necesitamos conocer la llamada *matriz de varianzas-covarianzas*, información que suelen dar los programas de regresión logística y que aparece en la Tabla 2.29.

Tabla 2.29. *Matriz de varianzas-covarianzas de las estimaciones del modelo con la interacción de AO y FUMA.*

	β_0	β_1	β_2	β_3	β_4
β_0	0,3972				
β_1	-0,0637	0,3306			
β_2	-0,0087	0,00097	0,0002		
β_3	-0,0211	0,0139	0,0002	0,0112	
β4	0,0038	-0,1801	0,0002	-0,0109	0,1256

Esta matriz da, en la diagonal principal, las varianzas, cuadrado de los errores estandars, de las estimaciones de los distintos parámetros, y en las casillas que no están en esa diagonal aparecen las covarianzas entre las distintas estimaciones; así, 0,3306 es la varianza del estimador del coeficiente de *AO*, que en efecto, coincide con el cuadrado del valor 0,575, error estandar dado en la Tabla 2.28; por otra parte, -0.1801 es la covarianza entre la estimación del coeficiente de *AO* y del coeficiente de la interacción *(AO)(FUMA)*; sustituyendo estos valores en la expresión anterior queda

$$var(\log(OR)) = 0{,}3306 + (FUMA)^2\ 0{,}1256 + 2(FUMA)\ (-0{,}1801)$$

Con la estimación del logaritmo de la *OR* para el término de interacción y su varianza podemos construir el intervalo de confianza; así, para las fumadoras de 1 a 24 cigarrillos, que son las mujeres cuyo valor de *FUMA* es 1, el riesgo asociado a los anticonceptivos orales, a igualdad de edad, viene dado por

$$\log(OR) = 0{,}631 + 0{,}399\ (1) = 1{,}03$$

y su varianza es

$$var(\log(OR)) = 0{,}3306 + (1)^2\ 0{,}1256 + 2(1)(-0{,}1801) = 0{,}096$$

por lo que el error estándar es la raiz cuadrada de 0,096, es decir, 0,3098; de esta manera $e^{1{,}03 \pm 1{,}96\ (0{,}3098)} = (1{,}53\ ,\ 5{,}14)$ es el intervalo buscado para la *OR* de las usuarias de anticonceptivos respecto a las no usuarias, pero sólo para las mujeres fumadoras entre 1 y 24 cigarrillos, sea cual sea la edad de las mujeres.

Merece la pena hacer la siguiente consideración acerca de las unidades de medida de las predictoras que hemos utilizado en el modelo; la variable *EDAD* ha sido tratada como numérica con valores 27, 32, 27, 42 y 47 para los 5 grupos de edad; sin embargo, en la construcción del fichero se había introducido la variable *GEDA* con valores 1, 2, 3, 4 y 5. Veamos que ocurre si utilizamos como predictora la variable *GEDA* en lugar de la *EDAD*; recuérdese que estaban ligadas por la relación *EDAD*=22+5(*GEDA*) por lo que *GEDA*=(*EDAD*-22)/5=*EDAD*/5-4.4; obsérvese como

la variable *GEDA* es igual a la quinta parte de la variable *EDAD* menos una constante. La lejanía y las estimaciones de los coeficientes del modelo que contiene *AO*, *GEDA* y *FUMA* son las que aparecen en la Tabla 2.30.

El coeficiente 0,152 de la *EDAD* en la Tabla 2.27 se podía interpretar como el cambio que experimenta el logit de la probabilidad de sufrir infarto por cada unidad de aumento de la variable *EDAD*, es decir, por cada año que cumpla una mujer; ahora el coeficiente 0.760 de *GEDA* es el cambio que experimenta el mismo logit por cada unidad de aumento en la variable *GEDA*, es decir, cada cinco años y obsérvese como, en efecto, 0.152 es precisamente el cociente 0.760/5. Esto viene a confirmar lo que ya se avanzó sobre la dependencia de los valores de las estimaciones de los coeficientes respecto de las unidades de medida que se utilicen para las variables.

Tabla 2.30. *El mismo modelo que en la Tabla 2.25 pero con la edad expresada en quinquenios.*

VARIABLE	ESTIMACIÓN	ERROR ESTÁNDAR
Constante	-5,907	0,332
AO	1,184	0,261
GEDA	0,760	0,070
FUMA	1,061	0,101

Este hecho debe prevenir al lector sobre el peligro que existe a la hora de comparar los coeficientes del modelo entre sí como forma de hacer comparaciones entre los riesgos asociados a las distintas variables que figuran en el modelo, ya que, como se acaba de comprobar, el valor del coeficiente depende de las unidades de medida en que se exprese la variable correspondiente. Una solución a este problema es utilizar los valores de las predictoras estandarizados.

2.11 Construcción de un índice de riesgo

Volviendo de nuevo al modelo más parsimonioso que explica nuestras observaciones, el modelo que no contempla ninguna interacción, se pueden calcular las *OR* estimadas a partir del modelo para cualquier combinación de las predictoras. Para ello tomemos como categoría de referencia las no usuarias de anticonceptivos, no fumadoras y con edad en el grupo mas joven, el representado por el valor 27. Una vez ajustado un modelo, los programas suelen generar una variable cuyos elementos son las estimaciones del componente sistemático del modelo para cada combinación de las predictoras; en la Tabla 2.31 aparecen los valores de las predictoras así como los del componente sistemático (CS) para cada tipología de mujer. Ya que estos últimos valores no son más que el logit de cada perfil mujer, podemos calcular todas las estimaciones de las distintas *OR* correspondientes a los riesgos asociados a cualquier mujer en relación a la categoría de referencia; así, en las dos primeras filas de la tabla

aparecen los valores -5,14627 y -4,38590 correspondientes a la categoría de referencia y a la categoría de las no usuarias, del grupo de 32 años y no fumadoras; por tanto, la diferencia -4,38590-(-5,14627)=0,76037 es la estimación del logaritmo del riesgo para las mujeres del grupo de 32 años respecto del grupo de 27, controlando por anticonceptivos y tabaco; por tanto, una estimación de esta *OR* es $e^{0,76037} = 2,14$.

De esta manera, cuanto mayor sea el valor del componente sistemático de una mujer, mayor será la diferencia con el correspondiente a la mujer de referencia; ya que esa diferencia no es más que el logaritmo del riesgo, cuanto mayor sea el valor del componente sistemático la mujer correspondiente estará a mayor riesgo de sufrir la enfermedad. De aquí que el componente sistemático se pueda considerar como un *índice de riesgo*.

Tabla 2.31. *Valores de las tres predictoras y del componente sistemático (CS).*

AO	EDAD	FUMA	CS
0,000	27,00	1,000	-5,14627
0,000	32,00	1,000	-4,38590
0,000	37,00	1,000	-3,62553
0,000	42,00	1,000	-2,86515
0,000	47,00	1,000	-2,10478
1,000	27,00	1,000	-3,96206
1,000	32,00	1,000	-3,20169
1,000	37,00	1,000	-2,44132
1,000	42,00	1,000	-1,68095
1,000	47,00	1,000	-0,92058
0,000	27,00	2,000	-4,08548
0,000	32,00	2,000	-3,32511
0,000	37,00	2,000	-2,56474
0,000	42,00	2,000	-1,80437
0,000	47,00	2,000	-1,04400
1,000	27,00	2,000	-2,90128
1,000	32,00	2,000	-2,14091
1,000	37,00	2,000	-1,38053
1,000	42,00	2,000	-0,62016
1,000	47,00	2,000	0,14021
0,000	27,00	3,000	-3,02470
0,000	32,00	3,000	-2,26433
0,000	37,00	3,000	-1,50395
0,000	42,00	3,000	-0,74358
0,000	47,00	3,000	0,01679
1,000	27,00	3,000	-1,84049
1,000	32,00	3,000	-1,08012
1,000	37,00	3,000	-0,31975
1,000	42,00	3,000	0,44062
1,000	47,00	3,000	1,20099

Comparando de esta manera cada combinación de las predictoras con la categoría de referencia se puede obtener la Tabla 2.32 que muestra las 30 estimaciones de las correspondientes *OR*.

Tabla 2.32. *Razones de ventajas de las distintas categorías de mujeres respecto a las no usuarias de anticonceptivos, no fumadoras y del grupo de 25 a 29 años.*

		E D A D				
FUMA	AO	27	32	37	42	47
No	No	1,00	2,14	4,58	9,79	20,95
	Sí	3,27	7,00	14,77	32,01	68,51
1-24	No	2,89	6,18	13,23	28,29	60,55
	Sí	9,45	20,21	43,26	92,51	198,00
≥ 25	No	8,35	17,87	38,24	81,75	174,93
	Sí	27,30	58,43	125,04	267,32	572,02

Como se indicó en el Apartado 1.8, en los estudios de casos y controles el término independiente no se puede interpretar a no ser que se conozcan las fracciones de muestreo en los casos y en los controles. Es por eso que nos hemos tenido que conformar hasta ahora con el conocimiento de las *OR*; es decir, en los estudios de casos y controles nos tenemos que contentar con saber cuántas veces es mayor el riesgo de una mujer respecto a otra, es decir, con una medida de riesgo relativa. Sin embargo no podemos responder a preguntas del tipo: ¿cuál es el riesgo de que una mujer determinada sufra un infarto?

La Tabla 2.33 muestra los resultados parciales, debidos a Halperin (1971), del estudio de cohorte de Framimgham diseñado para evaluar factores de riesgo asociados con la enfermedad coronaria; estos resultados corresponden a una subcohorte de 742 hombres con edades comprendidas entre 40 y 49 años seguidos durante 12 años y libres de enfermedad coronaria al comienzo del estudio;

Tabla 2.33. *Modelo estimado derivado del seguimiento de 742 hombres durante 12 años, para evaluar factores de riesgo asociados a enfermedad coronaria.*

VARIABLE	ESTIMACIÓN	ERROR ESTÁNDAR
Constante	-13,257	
Edad	0,122	0,044
Colesterol	0,007	0,002
Presión sistólica	0,007	0,006
Peso relativo	0,026	0,009
Hemoglobina	-0,001	0,01
Tabaco	0,422	0,103
ECG	0,721	0,401

las variables se codificaron así: la edad en años, el colesterol en mg/100 ml, la presión en mm de Hg., el peso relativo como el cociente entre su peso y la mediana del peso para los hombres de su altura, expresado en tanto por cien, la hemoglobina en gr/100 ml, el tabaco como 0 si nunca fumó, 1 si fuma menos de un paquete, 2 si fuma un paquete y 3 en otro caso; por último, 0 si el electrocardiograma (ECG) fué normal y 1 en caso contrario.

Al tratarse de un estudio de seguimiento, no sólo se pueden calcular razones de ventajas, sino que podemos evaluar riesgos individuales de padecer la enfermedad, es decir, probabilidades de que un individuo con determinadas características desarrolle la enfermedad en los 12 años de estudio; así, un hombre con 43 años, 230 mg/100 ml de colesterol, 150 mm. de presión, un peso relativo del 90%, una hemoglobina de 110 gr/100 ml, fumador de 20 cigarrillos y electro normal, tendrá un componente sistemático según el modelo estimado de -13,257 + 0,122(43) + 0,007(230) + 0,007(150) + 0,026(90) -0,001(110) + 0,422(2) + 0,721(0) = -2,267, por lo que la probabilidad de que un individuo como este desarrolle la enfermedad a lo largo de los 12 años de estudio es

$$\frac{e^{-2,267}}{1 + e^{-2,267}} = 0,09$$

lo que se puede interpretar diciendo que, de cada cien hombres con las características del anterior, 9 de ellos desarrollarán la enfermedad en un periodo de 12 años.

Aunque en el apartado 1.8 se dijo que, en los estudios de cohorte,

$$\frac{e^{\hat{\beta}_0}}{1 + e^{\hat{\beta}_0}}$$

se podía interpretar como una estimación de la incidencia en los no expuestos, es decir, en los individuos donde el valor de las predictoras es 0, en este caso no podemos decir nada acerca del valor -13,2573 ya que no tiene sentido hablar de un hombre sin edad, sin colesterol, etc. Para evitar esta dificultad lo que se suele hacer es posibilitar el que cualquier variable predictora pueda tomar valor 0, lo que se puede conseguir con el *centrado (centering)* de las variables cuantitativas; todo consiste en restar a los valores de las tales variables el valor de su media (mediana) por lo que esta nueva variable tomará valor cero cuando la variable original valga el valor de su media (mediana). Shapiro (1982) publicó un estudio de cohorte sobre factores de riesgo asociados a infección tras histerectomía para lo que midieron la existencia de profilaxis, codificada como 0 en caso de que no se realizase profilaxis y 1 en otro caso, el tipo de cirugía empleada, 0 si fué vaginal y 1 en caso de cirugía abdominal, tipo de servicio sanitario donde se realizó la intervención, 0 para los privados y 1 para los públicos,

edad de la paciente y duración de la operación; estas dos últimas variables cuantitativas las centraron en la mediana de sus distribuciones que fueron 40 años y 120 minutos, respectivamente; los resultados del modelo de regresión logística son los que aparecen en la Tabla 2.34.

Tabla 2.34. *Factores de riesgo asociados a infección tras histerectomía.*

VARIABLE	ESTIMACIÓN
Constante	-2,090
Profilaxis	-0,878
Tipo de operación	0,642
Tipo de servicio	0,602
Edad - 40 años	-0,018
Duración de la operación - 120 minutos	0,003
Profilaxis . (duración - 120 minutos)	0,011

Con estas transformaciones de las variables cuantitativas sí tiene sentido hablar de una mujer con valor 0 para todas las predictoras; ella sería una mujer a la que no se le administró profilaxis, se le hizo abordaje vaginal en un servicio privado, con 40 años de edad y para la que la operación duró 120 minutos; entonces,

$$\frac{e^{-2,09}}{1 + e^{-2,09}} = 0,11$$

es la probabilidad de que una mujer de esas características llegue a tener infección, o lo que es igual, de cada 100 mujeres de esas características, 11 de ellas resultarán infectadas.

CAPÍTULO III

DIAGNÓSTICO EN REGRESIÓN LOGÍSTICA

Todo modelo estadístico, en particular el modelo de regresión logística, está basado en unas hipótesis previas de las que es necesario evaluar su cumplimiento. Estas cuestiones junto a la presencia de observaciones extrañas o muy influyentes en el modelo ajustado son el objetivo del presente capítulo.

3.1. Introducción

Aunque se haya ajustado un modelo estadístico como mecanismo generador plausible del patrón subyacente de unos datos observados, este no debe ser el final de ningún análisis estadístico. Hay toda una fase posterior de análisis que tiene que ver con la adecuación de las observaciones al modelo propuesto; aunque un ajuste determinado explique suficientemente bien los datos, hay una serie de cuestiones que hay que investigar. Podría ocurrir que: a) otra función distinta de la logística pudiese describir mejor nuestras observaciones; b) por otra parte, el modelo logístico, tal como se ha presentado, es restrictivo en el sentido de que exige que la relación entre el logit de la probabilidad del suceso de interés y las predictoras sea lineal, exigencia que no tiene porqué ocurrir en todas las ocasiones, en cuyo caso puede ser necesario realizar algún tipo de transformación de las predictoras para poder cumplir tal condición de linealidad; c) también es fundamental en cualquier análisis estadístico el detectar algunas observaciones "raras" para el conjunto de datos que estamos manejando, o bien, observaciones que tengan una influencia exagerada en las estimaciones de los parámetros del modelo. El objetivo de este capítulo es el tratamiento de estas cuestiones que se ejemplificarán con los datos que utiliza Pregibon (1981) en su artículo sobre diagnóstico en regresión logística.

3.2. Ejemplo. Vasoconstricción en la piel de los dedos e inspiración profunda

La experiencia se debe a Finney (1947) y consiste en un ensayo para estudiar un reflejo fisiológico de vasoconstricción en la piel de los dedos después de una inspiración profunda; la respuesta es la presencia o no de vasoconstricción, *RESP*, y la hipótesis del estudio es la dependencia de esta respuesta dependiendo del volumen de aire inspirado, *VOL*, y de la velocidad o tasa a la que se inspira, *TASA*. La Tabla 3.1 presenta los datos de los 39 individuos del estudio, junto a la variable de identificación *ID* para cada uno de ellos.

La variable respuesta *RESP* se codificó así: *RESP*=1 si el sujeto presentaba vasoconstricción y *RESP*=0 en caso contrario. Ya que en este caso tenemos datos no agrupados debemos considerar 39 binomiales de parámetro n=1.

Tabla 3.1. *Datos del estudio de Finney.*

ID	VOL	TASA	RESP	ID	VOL	TASA	RESP	ID	VOL	TASA	RESP
1	3,7	0,825	1	14	1,4	2,33	1	27	1,8	1,5	1
2	3,5	1,09	1	15	0,75	3,75	1	28	0,95	1,9	0
3	1,25	2,5	1	16	2,3	1,64	1	29	1,9	0,95	1
4	0,75	1,5	1	17	3,2	1,6	1	30	1,6	0,4	0
5	0,8	3,2	1	18	0,85	1,415	1	31	2,7	0,75	1
6	0,7	3,5	1	19	1,7	1,06	0	32	2,35	0,03	0
7	0,6	0,75	0	20	1,8	1,8	1	33	1,1	1,83	0
8	1,1	1,7	0	21	0,4	2,0	0	34	1,1	2,2	1
9	0,9	0,75	0	22	0,95	1,36	0	35	1,2	2,0	1
10	0,9	0,45	0	23	1,35	1,35	0	36	0,8	3,33	1
11	0,8	0,57	0	24	1,5	1,36	0	37	0,95	1,9	0
12	0,55	2,75	0	25	1,6	1,78	1	38	0,75	1,9	0
13	0,6	3,0	0	26	0,6	1,5	0	39	1,3	1,625	1

Ajustando distintos modelos, se obtienen los siguientes resultados: para el modelo nulo, la lejanía es 54,040 con 38 grados de libertad; cuando se añade la variable *TASA*, el lejanía pasa a ser 49,655 con un cambio de 4,39 que para 1 grado de libertad es significativo, es decir, la variable *TASA* está asociada a la variable respuesta. Si a este modelo se le añade el volumen de aire inspirado, la lejanía desciende al valor 29,772 con 36 grados de libertad, por lo que el cambio producido es de 19,883 para 1 grado de libertad. En definitiva, las dos predictoras aportan información interesante; las estimaciones de los parámetros de este último modelo aparecen en la Tabla 3.2.

Tabla 3.2. *Estimaciones y errores estándar de los coeficientes de las variables TASA y VOL.*

VARIABLE	ESTIMACION	E.E.
Constante	-9,530	3,224
TASA	2,649	0,912
VOL	3,882	1,425

Clásicamente, estos datos no se han ajustado como se acaba de hacer sino que tanto Finney como Pregibon utilizan las variables *LTAS* y *LVOL*, transformaciones logarítmicas de las predictoras, que dan lugar a las tres lejanías siguientes: 54.040,

48,857 y 29,227, respectivamente. Como se ve, las dos últimas son muy parecidas a las de los modelos con las variables no transformadas, por lo que no hay razones, al menos estadísticas, para elegir entre un modelo y otro; nosotros elegiremos, por respeto a los citados autores, este último modelo cuyos parámetros estimados aparecen en la Tabla 3.3.

Tabla 3.3. *Estimaciones y errores estándar de los coeficientes de las variables LTAS y LVOL.*

VARIABLE	ESTIMACION	E.E.
Constante	-2,875	1,319
LTAS	4,562	1,837
LVOL	5,179	1,864

En definitiva, el modelo estimado es

$$logit(\hat{p}) = -2{,}875 + 4{,}562\ LTAS + 5{,}179\ LVOL$$

y como los coeficientes de las dos predictoras son ambos positivos, ello implica que al aumentar cualquiera de éstas, aumenta la probabilidad de vasoconstricción; por ejemplo, para la misma tasa de aire inspirado, la ventaja de tener vasoconstricción es $e^{0{,}5179}=1{,}68$ veces mayor por cada décima de unidad de aumento en el logaritmo del volumen de aire inspirado; téngase en cuenta que si un individuo tiene una décima más de que otro en el logaritmo del volumen de aire inspirado, ello significa que el volumen del primero es $e^{0{,}1}=1{,}105$ veces el volumen del segundo.

3.3. Alternativas al modelo de regresión logística

Como ya se ha dicho en repetidas ocasiones, el modelo de regresión logística establece que

$$p(Y=1/X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

Según esta expresión, la probabilidad de que Y sea igual a 1 viene dada por una función que que varía entre 0 y 1, los valores permitidos para una probabilidad, y que, en el caso en que $\beta_1 > 0$, es una función no decreciente, es decir, cuando aumenta el valor de la predictora la probabilidad de presentar la característica de interés aumenta o permanece constante. Por tanto, en principio, cualquier función

$$F(\beta_0 + \beta_1 X_1)$$

que varíe entre 0 y 1 y sea no decreciente puede ser candidata para modelar la relación entre Y y X.

En farmacología es frecuente el estudio de la tolerancia a un determinado fármaco; sea X la dosis administrada del medicamento e Y una variable dicotómica tal que Y=1 si el sujeto experimental presenta intoxicación por el fármaco e Y=0 en caso contrario; sea D la tolerancia de un sujeto al fármaco, de tal manera que si la dosis X administrada es mayor que D, el individuo presenta intoxicación, es decir,

$$P(Y = 1) = P(D \leq X)$$

Ya que en muchas situaciones, la tolerancia de un individuo suele distribuirse según una normal de una combinación lineal de la dosis administrada, podemos escribir que

$$p = P(Y=1 \ /X) = P(D \leq X) = \phi(\beta_0 + \beta_1 X)$$

donde ϕ es la función de distribución de la normal estandarizada; por tanto

$$\phi^{-1}(p) = \beta_0 + \beta_1 X$$

Este modelo, en el que la función de nexo es la inversa de la normal estandarizada, es el llamado *modelo probit* y es una alternativa para estudiar la relación entre una variable dicotómica Y y una predictora X.

Tanto el modelo logístico como el probit son algo restrictivos en cuanto a la relación entre p y X, en el sentido de que implícitamente establecen que p se aproxima a la misma velocidad a 0 que a 1, es decir, son simétricas respecto a la horizontal p=0.5.

Una forma de salvar esta restricción es mediante el *modelo log-log*

$$\log(-\log(p)) = \beta_0 + \beta_1 X$$

que se puede escribir tambien de la siguiente forma

$$p = \exp(-e^{\beta_0 + \beta_1 X})$$

modelo que utiliza como nexo la función log-log y que se aproxima más rápidamente a 0 que a 1.

Por último, el modelo

$$\log(-\log(1 - p)) = \beta_0 + \beta_1 X$$

que utiliza la función de *log-log complementaria*, es similar al anterior, salvo la diferencia de que se aproxima a 1 más rápidamente que 0. Evidentemente, es de esperar que la interpretación de los coeficientes de estos modelos alternativos no sea la misma que el modelo logístico; en efecto, sean dos individuos con valores x_1 y x_2 de la variable X; bajo éste último modelo, para el primer individuo tendremos

$$\log(-\log(1-p_1)) = \beta_0 + \beta_1 x_1$$

y para el segundo

$$\log(-\log(1-p_2)) = \beta_0 + \beta_1 x_2$$

con p_1 y p_2 las probabilidades de presentar la característica el primer y segundo individuos, respectivamente. Restando de esta última expresión la anterior se obtiene

$$\log(-\log(1-p_2)) - \log(-\log(1-p_1)) = \beta_1(x_2 - x_1)$$

es decir,

$$\frac{\log(1-p_2)}{\log(1-p_1)} = e^{\beta_1(x_2 - x_1)}$$

luego

$$1 - p_2 = (1 - p_1)e^{\beta_1(x_2 - x_1)}$$

Resumiendo, bajo el modelo que utiliza la función log-log complementaria, la probabilidad de que el segundo individuo no presente la característica es igual a la probabilidad correspondiente al primero elevada a la potencia

$$e^{\beta_1(x_2 - x_1)}$$

Para el modelo con la función log-log, no hay más que cambiar la no presentación por la presentación de la característica de interés.

Los programas estadísticos disponen de varias funciones de nexo, entre ellas la función logit, la probit y la log-log complementaria, por lo que los modelos que acabamos de describir se pueden ajustar. En la Figura 3.1 aparecen las representaciones gráficas de las funciones logit, probit y log-log complementaria.

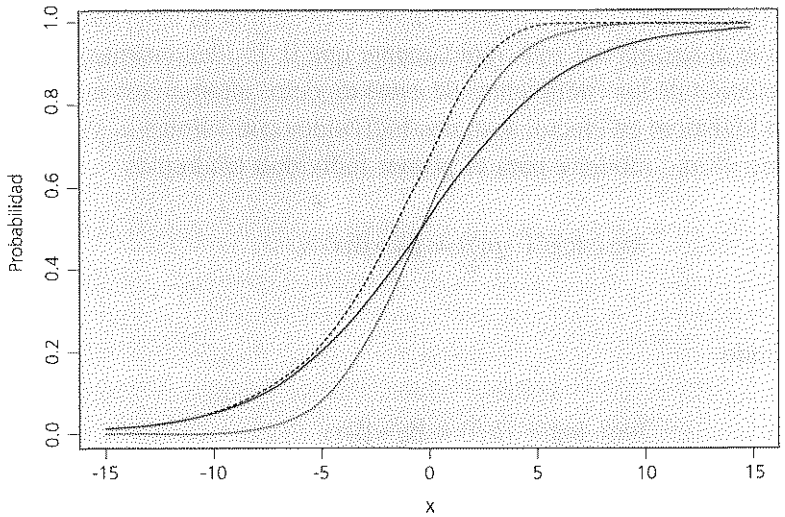


Figura 3.1. *Representación gráfica de las funciones logit (línea continua), probit (punteada) y log-log complementaria (rayada).*

Aranda-Ordaz (1981) propuso una familia de funciones de nexo a partir de la cual se pueden derivar tests para evaluar la bondad de la función logística como nexo. Por último, Guerrero (1982) propusieron una familia de transformaciones de la razón de ventajas de la forma

$$\left(\frac{p_i}{1 - p_i}\right)^{(\lambda)} = \beta_0 + \beta_1 X_1 + \ldots + \beta_m X_m$$

donde

$$\left(\frac{p_i}{1 - p_i}\right)^{(\lambda)} = \log \frac{p_i}{1 - p_i} \qquad \lambda = 0$$

$$= \frac{\left(\frac{p_i}{1 - p_i}\right)^{\lambda} - 1}{\lambda} \qquad \lambda \neq 0$$

que incluye, para el caso particular λ=0, al modelo logístico. Como los β_i, λ es un parámetro a estimar mediante máxima verosimilitud. Dependiendo de la función de nexo elegida el ajuste puede ser más o menos bueno y, en términos generales, las variables a incluir en el modelo final van a depender del nexo elegido.

De todas formas, la transformación logística es muy popular en los estudios epidemiológicos debido a la interpretabilidad de sus coeficientes como logaritmos de razones de ventajas, lo que hace que sea aplicable a los estudios de casos y controles.

3.4. Residuales en regresión logística

Como en los modelos de regresión lineal, parece natural que, a la hora de evaluar la adecuación de un modelo ajustado, juegue un papel fundamental la medida que capte la diferencia entre las respuestas observadas y las predichas por tal modelo; esa medida recibe el nombre de *residual*. Sin embargo, al contrario que en los modelos lineales, para el modelo logístico se pueden definir los residuales de varias maneras; así, en una primera aproximación, la diferencia

$$y_i - n_i \hat{p}_i$$

son los *residuales crudos*, que no son más que la respuesta observada menos la predicha por el modelo; en caso de datos agrupados, n_i representa el número de individuos con un patrón i de predictoras, y_i es el número de estos que presentan la característica de interés y $\hat{p}_i$ la probabilidad predicha para un individuo con tal patrón de predictoras; estos residuales crudos divididos por su error estandar

$$r_i = \frac{(y_i - n_i \hat{p}_i)}{\sqrt{n_i \hat{p}_i(1 - \hat{p}_i)}}$$

reciben el nombre de *residuales de Pearson* o *residuales crudos estandarizados*. Para el caso de datos no agrupados, es decir, n_i=1, los residuales anteriores toman la forma

$$y_i - \hat{p}_i$$

y

$$r_i = \frac{(y_i - \hat{p}_i)}{\sqrt{\hat{p}_i(1 - \hat{p}_i)}}$$

respectivamente.

Los residuales de Pearson no son más que la raiz cuadrada de la contribución de cada observación al valor del estadístico χ^2 correspondiente al modelo ajustado; es decir, sumando para cada respuesta los cuadrados de estos residuales se obtiene el estadístico χ^2.

Otro residual de interés es d_i, la raiz cuadrada de la contribución de cada observación a la lejanía; según se vió en el apartado 1.11, la lejanía del modelo logístico en el caso de datos agrupados toma la forma

$$2\sum_{i=1}^{k}\left(y_i \log\frac{y_i}{n_i\hat{p}_i} + (n_i - y_i)\log\frac{n_i - y_i}{n_i - n_i\hat{p}_i}\right)$$

con k el número de binomiales; por tanto, el llamado *residual de la lejanía* toma la forma

$$d_i = \pm\sqrt{2\left(y_i \log\frac{y_i}{n_i\hat{p}_i} + (n_i - y_i)\log\frac{n_i - y_i}{n_i - n_i\hat{p}_i}\right)}$$

donde el signo es el mismo que el de la diferencia

$$y_i - n_i\hat{p}_i$$

Para datos no agrupados, es decir, cuando todas las binomiales tengan como parámetro $n_i=1$, se pueden distinguir dos situaciones; en el caso en que la respuesta observada sea $y_i=0$ estos residuales se pueden expresar abreviadamente como sigue

$$d_i = -\sqrt{-2\log(1-\hat{p}_i)}$$

y en el caso en que $y_i=1$

$$d_i = -\sqrt{-2\log(\hat{p}_i)}$$

Se puede comprobar que entre estos dos residuales existe la siguiente relación

$$d_i^2 = 2\log(1 + r_i^2)$$

Aunque se pueden definir más tipos de residuales para los modelos lineales generalizados, y por tanto para los modelos de regresión logística, estos dos definidos y algunas transformaciones de ellos, que trataremos más adelante, son los más utilizados.

Una vez ajustado un modelo logístico, los programas informáticos de regresión logística proporcionan para cada individuo, entre otras variables de interés, las probabilidades estimadas de presentar la característica, en este caso la vasoconstricción, y los residuales tanto de Pearson como los de la lejanía; para nuestro ejemplo, la Tabla 3.4 presenta el valor de la respuesta y los dos tipos de residuales para cada uno de los 39 individuos del estudio.

Tabla 3.4. *Valores de la respuesta (RESP), probabilidades predichas ($\hat{p}_i$), residuales de Pearson (r_i) y de la lejanía (d_i).*

ID	RESP	$\hat{p}_i$	r_i	d_i
1	1	0,954	0,220	0,308
2	1	0,982	0,135	0,190
3	1	0,921	0,292	0,405
4	1	0,075	3,518	2,278
5	1	0,782	0,529	0,702
6	1	0,729	0,609	0,794
7	0	0,001	-0,033	-0,046
8	0	0,510	-1,020	-1,194
9	0	0,009	-0,094	-0,132
10	0	0,001	-0,029	-0,041
11	0	0,001	-0,037	-0,052
12	0	0,205	-0,507	-0,677
13	0	0,375	-0,775	-0,970
14	1	0,939	0,256	0,356
15	1	0,841	0,435	0,589
16	1	0,976	0,157	0,222
17	1	0,995	0,071	0,100
18	1	0,106	2,904	2,119
19	0	0,535	-1,072	-1,237
20	1	0,945	0,240	0,335
21	0	0,011	-0,107	-0,152
22	0	0,150	-0,419	-0,569
23	0	0,512	-1,024	-1,198
24	0	0,652	-1,368	-1,453
25	1	0,899	0,335	0,461
26	0	0,025	-0,159	-0,224
27	1	0,883	0,364	0,499
28	0	0,447	-0,899	-1,088
29	1	0,554	0,898	1,088
30	0	0,010	-0,099	-0,140
31	1	0,722	0,620	0,806
32	0	0,000	-0,001	-0,001
33	0	0,593	-1,206	-1,340
34	1	0,771	0,545	0,721
35	1	0,774	0,540	0,716
36	1	0,811	0,483	0,647
37	0	0,447	-0,899	-1,088
38	0	0,192	-0,487	-0,653
39	1	0,668	0,705	0,899

Representando gráficamente los valores de estos dos tipos de residuales, como muestra la Fig. 3.2, se puede comprobar su parecido comportamiento; obsérvese como los individuos 4 y 18 tienen ambos residuales grandes; más adelante haremos uso de estos residuales como herramienta fundamental para todo el proceso diagnóstico del modelo.

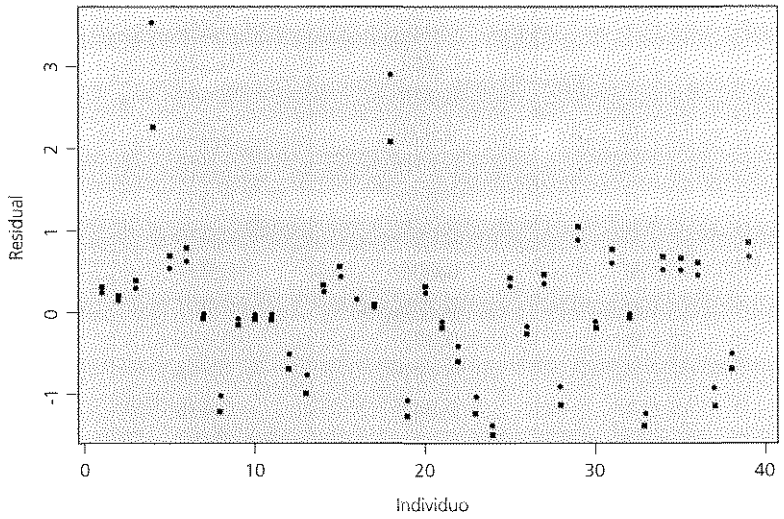


Figura 3.2. *Representación de los residuales de Pearson (•) y de la lejanía (■).*

3.5. Evaluación de la bondad de ajuste

El uso de la lejanía como medida de la bondad del ajuste de un modelo a unos datos se basa en la propiedad de que este estadístico se distribuye asintóticamente como una χ^2. Sin embargo, como se dijo en el apartado 1.11, esto no siempre es así, pues para ello se han de cumplir dos importantes condiciones: la primera exige la independencia de las observaciones mientras que la segunda establece que los parámetros n_i de las binomiales deben ser "grandes", lo que en la práctica significa que no más del 20% de los valores predichos por el modelo pueden ser menores de 5. Si no se cumplen estas dos condiciones, la distribución de la lejanía deja de ser la χ^2; dicho de otra forma, en tales condiciones un valor grande de la lejanía no implica necesariamente evidencia de un mal ajuste. En el caso de datos no agrupados, donde cada individuo sigue una binomial, con n_i =1, McCullagh (1989) demuestra que la distribución de la lejanía no es la χ^2, por lo que ésta no se puede tomar como medida de la bondad del ajuste. Para el caso de datos agrupados, tanto la lejanía como el estadístico de Pearson sí pueden tomarse como medidas de la bondad del ajuste. Sin

embargo, a la hora de comparar dos modelos, sean cuales sean los parámetros de las binomiales, es decir, dispongamos o no de datos agrupados, la diferencia entre dos lejanías sigue aproximadamente una χ^2 cuyos grados de libertad son la diferencia entre el número de parámetros a estimar en los dos modelos.

Un método para la evaluación de la bondad del ajuste de un determinado modelo en caso de que los datos no sean agrupados o que, aunque agrupados, haya más de un 20% de valores esperados menores de 5, fué propuesto por Hosmer (1980), que puede describirse como sigue: una vez ajustado un determinado modelo, se pueden calcular las probabilidades estimadas para cada individuo. A partir de estos resultados se pueden formar los llamados *grupos de riesgo* que se consiguen dividiendo el conjunto ordenado de esas probabilidades estimadas en k grupos. Estos autores recomiendan, en base a los 9 deciles, tomar $k=10$ grupos, que demominan deciles de riesgo y que representamos por D_j, con $j=1, 2, ...,10$, por lo que en el primer decil de riesgo D_1 estarán los individuos cuyas probabilidades $\hat{p}_i$ asociadas estén comprendidas entre 0 y el primer decil; el segundo decil de riesgo lo formarán los individuos con probabilidades comprendidas entre el primer y segundo deciles, etc. Sean n_{1j} el número de casos, (Y=1), en el grupo D_j y por n_{0j} el número de controles, (Y=0); con estos valores podemos representar lo siguiente

	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}
Y=0	n_{01}	n_{02}	n_{03}	n_{04}	n_{05}	n_{06}	n_{07}	n_{08}	n_{09}	n_{010}
	$\hat{n}_{01}$	$\hat{n}_{02}$	$\hat{n}_{03}$	$\hat{n}_{04}$	$\hat{n}_{05}$	$\hat{n}_{06}$	$\hat{n}_{07}$	$\hat{n}_{08}$	$\hat{n}_{09}$	$\hat{n}_{010}$
Y=1	n_{11}	n_{12}	n_{13}	n_{14}	n_{15}	n_{16}	n_{17}	n_{18}	n_{19}	n_{110}
	$\hat{n}_{11}$	$\hat{n}_{12}$	$\hat{n}_{13}$	$\hat{n}_{14}$	$\hat{n}_{15}$	$\hat{n}_{16}$	$\hat{n}_{17}$	$\hat{n}_{18}$	$\hat{n}_{19}$	$\hat{n}_{110}$

donde

$$\hat{n}_{1j} = \sum_{i \in D_j} \hat{p}_i$$

es la suma de las probabilidades predichas para los casos pertenecientes al decil j, es decir, el número de casos esperado en ese decil de riesgo; por tanto,

$$\hat{n}_{0j} = (n_{0j} + n_{1j}) - \hat{n}_{1j}$$

que es la diferencia entre el número de individuos en el decil j menos los casos esperados, será el numero de controles esperados para ese decil.

Hosmer y Lemeshow propusieron el estadístico

$$C = \sum_{s=0}^{1} \sum_{j=1}^{10} \frac{(n_{sj} - \hat{n}_{sj})^2}{\hat{n}_{sj}}$$

como medida de la bondad del ajuste del modelo propuesto y comprobaron que si éste es válido, su estadístico se distribuye aproximadamente como una χ^2 con 10-2 grados de libertad. Hosmer y Lemeshow recomiendan utilizar k=10 grupos de riesgo y, según su experiencia, si se utilizan menos de seis grupos la potencia del test es muy baja. Otro estadístico alternativo propuesto por los mismos autores se construye de forma similar a éste pero en vez de tomar los deciles de riesgo toma unos determinados puntos de corte, por ejemplo, 0.1, 0.2, ..., 0.9 para formar los grupos de riesgo.

En regresion lineal, el coeficiente de determinación cuantifica R^2 la parte de la variabilidad de la respuesta que es explicada por las predictoras o, dicho de otra manera, este coeficiente describe la fuerza de la asociación entre la respuesta y el componente sistemático del modelo; así, si R^2=1 implica que podemos predecir perfectamente la respuesta a partir de las predictoras. Para el caso de la regresión logística se han hecho varios intentos para definir una medida análoga a este coeficiente aunque ninguna de las propuestas goza del conceso general, Agresti (1990); un reciente estudio comparativo entre ellas, Mittlböck (1996), recomienda, entre otras, el cuadrado del coeficiente de correlación de Pearson entre la respuesta y las probabilidades predichas, es decir,

$$R^2 = \frac{\left(\sum_{i=1}^{n} (y_i - \bar{p})(\hat{p}_i - \bar{p})\right)^2}{\sum_{i=1}^{n} (y_i - \bar{p})^2 \sum_{i=1}^{n} (\hat{p}_i - \bar{p})^2}$$

donde $\bar{p}$ es el porcentaje de casos en la muestra, que coincide con la probabilidad predicha media para todos los individuos. Para el ejemplo que venimos estudiando, con los resultados de la Tabla 3.4, se puede comprobar que $\bar{p}$=20/39=0,51 y R^2=0,73; en definitiva, las dos predictoras del modelo ajustado explican un 73% de la variabilidad de la respuesta.

3.6. Elección de la forma funcional de las predictoras. Métodos basados en el alisamiento

Un tipo de cuestiones, muy importante por otra parte, tiene que ver con la elección de la mejor escala para las variables predictoras; es decir, la falta de bondad de ajuste puede provenir de una especificación incorrecta de la forma funcional de las predictoras; en esta cuestión, los métodos gráficos son las principales herramientas

disponibles actualmente para establecer una aproximación a la forma funcional correcta. Retomemos el ejemplo del Capítulo II sobre los anticonceptivos y el infarto y estudiemos la relación entre la edad de las mujeres y el riesgo de infarto; en este caso estamos ante una variable con caracter numérico que, aunque continua, aquí se ha categorizado en el sentido de que se ha asignado un mismo valor de la edad a cada grupo; de cualquier forma, los valores 27, 32, 37, 42 y 47 tienen naturaleza numérica; el modelo logístico establece que existe una relación lineal entre el *logit(p)* y EDAD.

Un método de evaluar la adecuación de nuestros datos al modelo es representar en un sistema de coordenadas los logits estimados para cada edad y ver si existe tal tendencia lineal; para ello, las distribuciones de los casos y de los controles en los diferentes grupos de edad son las que aparecen en la Tabla 3.5.

		INFARTO			
		Sí	No	Ventaja	logit(p)
EDAD	27	6	286	0,02098	-3,864
	32	21	423	0,04965	-3,003
	37	37	356	0,10393	-2,264
	42	71	371	0,19137	-1,654
	47	99	306	0,32353	-1,128
	Total	234	1742		

Tabla 3.5. *Ventajas y logit(p) de sufrir infarto para las distintas edades.*

Si se representa en el eje vertical los logits estimados y en el eje de abcisas la edad, la nube constituida por estos cinco puntos es razonablemente lineal, como aparece en la Figura 3.3. Si acaso no lo fuese habría que someter a la variable predictora de la que se trate, en este caso la EDAD, a algún tipo de transformación; la transformación logarítmica o la inclusión de términos elevados al cuadrado, al cubo, etc., muchas veces resuelven estos problemas de no linealidad; más adelante se discuten otras técnicas para detectar la no linealidad de las predictoras.

Cuando se tiene una variable continua con los datos no agrupados, las cosas se complican. Si acaso se dispusiesen de varios valores de la variable resultado para cada valor de la predictora, se podría proceder como en el caso de datos agrupados; sin embargo, en los estudios no experimentales, este hecho casi nunca ocurre, por lo que

no podremos calcular los logits para cualquier valor de las predictoras. Ante esta situación caben dos alternativas: categorizar la predictora según el problema del que se trate, lo que casi nunca es buena política, o bien, utilizar métodos más sofisticados, algunos de los cuales se exponen en los siguientes apartados.

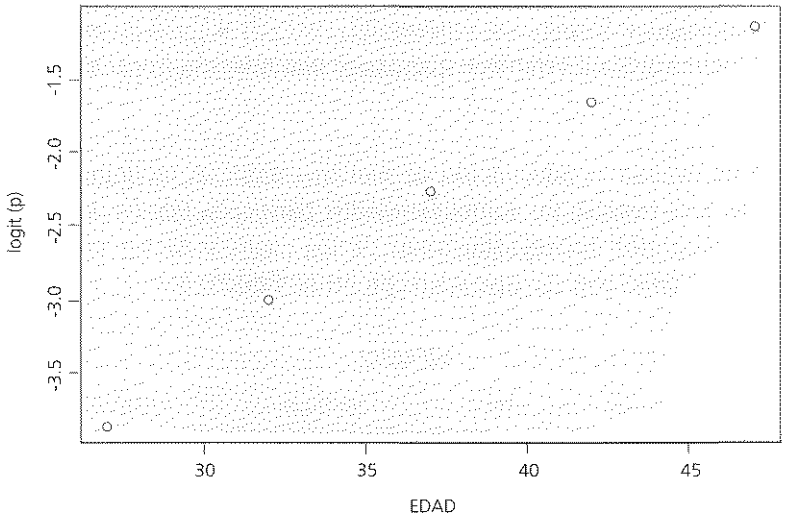


Figura 3.3. *Logit de infarto para distintas edades.*

3.6.1. Alisado para una respuesta continua

En estadística, los modelos de regresión clásicos que establecen la dependencia entre una variable resultado y las predictoras a través de un conjunto de parámetros, se denominan *modelos paramétricos*. Estos presentan la ventaja de proporcionar una descripción más o menos sencilla de tal dependencia y, por otra parte, permiten cuantificar el cambio que experimenta la variable resultado por unidad de aumento de la(s) predictora(s). Como inconveniente está el hecho de que en ocasiones las relaciones entre variables son más complicadas y no se pueden, desgraciadamente, describir mediante esos modelos. Para soslayar estas dificultades se han propuesto nuevos modelos, tanto paramétricos como no paramétricos, basados todos en el concepto de *alisamiento* (*smoothing*).

Para facilitar la compresión de lo que sigue consideremos una variable respuesta Y y una predictora X ambas continuas. El principio que subyace en el alisamiento es que la relación entre la variable dependiente y las predictoras no presenta discontinuidades, es decir, para dos valores "próximos" en el espacio de las predictoras, los valores correspondientes de la respuesta también son "parecidos"; en particular, en el caso de una sola predictora, para valores de X próximos, sus correspondientes valo-

res de Y también son próximos; por tanto, a la hora de estimar el valor medio de Y correspondiente a un valor x_i de X, algo deben tener que decir los valores de Y correspondientes a valores de X cercanos a x_i.

Sean $(x_i, y_i)1$ con $i=1,2,...,n$ un conjunto de n observaciones de las variables X e Y que, sin pérdida de generalidad vamos a suponer $x_1 \leq x_2 \leq ... x_n$; la representación gráfica de estas n parejas de valores da lugar a una nube de puntos sobre el plano. El alisamiento de esta nube de puntos consiste en el cálculo de otra nube definida por los puntos (x_i, y_i^s) con las mismas abcisas que la anterior pero cuyas nuevas ordenadas y_i^s son los valores alisados de los y_i, estimación del valor medio de Y en $X=x_i$; para valores distintos de los x_i, pero en el rango de x_1 a x_n, las estimaciones se consiguen interpolando.

Consideremos el caso más simple de una sola predictora y para ello supongamos que queremos estudiar el cambio del índice de masa corporal (IMC) con la edad en los individuos adultos; para ello, pensemos en una muestra de adultos de edades entre 18 y 60 años de los que conocemos tanto su edad como su IMC. En una primera aproximación, podríamos considerar una serie de categorías de la edad, por ejemplo, los menores de 30 años, los que tiene entre 30 y 40 años, los comprendidos entre 40 y 50 y los mayores de 50 y menores de 60 años. En estas condiciones, ¿cómo estimar el cambio del IMC según el grupo de edad? Parece natural estimar el IMC para cada categoría de edad como el valor medio de los IMC de los individuos pertenecientes a tal grupo etáreo; la Figura 3.4 muestra un gráfico en escalera, representación de la relación entre el IMC y la edad; en ese gráfico se puede apreciar el aumento del IMC con el aumento de la edad de los adultos.

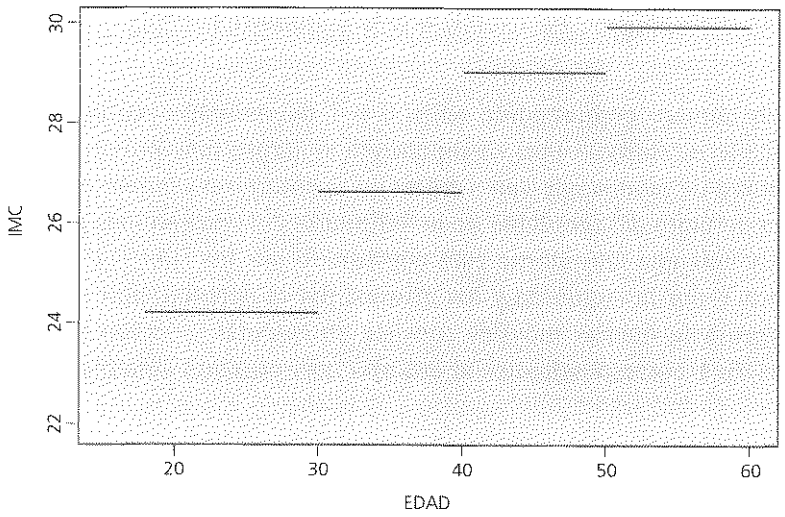


Figura 3.4. *Relación entre IMC y la EDAD en adultos.*

Sin embargo, cuando la predictora es continua, como en este caso, es una mala solución categorizarla y, por otra parte, el alisamiento deja que desear pues la gráfica resultante es una función en escalera, con discontinuidades en los puntos que definen las distintas categorías. En tal caso, si se dispusiera de varios valores de IMC por cada valor de edad, las estimaciones correspondientes a cada una de estas edades se podría conseguir mediante la media de los IMC de los individuos con tal edad. Como antes se dijo, en los estudios observacionales no disponemos de observaciones repetidas para cada valor de la predictora; una forma de obtener una estimación del valor medio de Y para $X=x_i$ es mediante la combinación de los valores de Y correspondientes a valores de X comprendidos en un entorno de x_i, es decir, valores próximos a x_i; por ejemplo, para estimar el IMC para los individuos de 43 años consideraremos los IMC de ellos y de los que tengan una edad parecida. Dependiendo de como combinemos los IMC y de lo que entendamos por "edad parecida", así tendremos distintos tipos de alisamiento.

Consideremos a continuación un método de alisamiento, utilizado tradicionalmente en el análisis de series temporales, conocido con el nombre de *medias móviles* (*running-mean smoother*) que evita los saltos de la gráfica anterior. Consideremos que los individuos entre 37 y 48 años tienen una edad parecida a 43; pues bien, estimaremos el IMC para los 43 años mediante la media de los IMC de los sujetos que tienen entre 37 y 48 años. Un tipo de entorno es el denominado *entorno más próximo simétrico* definido como sigue: supuestos ordenados los n valores de X en forma creciente, una definición formal del entorno N_i más próximo simétrico correspondiente al valor x_i es el conjunto de valores de X correspondientes a los órdenes definidos así

$$N_i = \{ \max\left(i - \frac{[kn] - 1}{2}, 1\right), \ldots, i\text{-}1, i, i\text{+}1, \ldots, \min\left(i + \frac{[kn] - 1}{2}, n\right) \}$$

donde k es un número comprendido entre 0 y 1 denominado *parámetro de alisamiento*. Mediante el símbolo $[kn]$ se representa el impar más próximo al producto kn y es el número de puntos que van a entrar en el entorno o *ventana de alisamiento*. Esa expresión un tanto desagradable de la definición del entorno más próximo simétrico no indica más que tal entorno está constituido por los $([kn]-1)/2$ valores de X más cercanos a x_i por la izquierda y los mismos por la derecha, más el propio x_i. El alisamiento mediante medias móviles consiste en calcular y_i^s como el valor medio de los valores de Y correspondientes al entorno más próximo simétrico; escrito formalmente

$$y_i^s = \frac{\sum_{j \in N_i} y_j}{[kn]}$$

Veamos como funciona este método con un ejemplo muy simple; sean las n=21 parejas de valores que aparecen en la Tabla 3.6 y consideremos que cada entorno va a contener el 40% de los puntos, lo que es igual que escribir k=0,4; por tanto, la ventana de alisamiento contendrá en este caso [21(0,4)] = [8,4]=9 puntos. ¿Cómo calcular el valor alisado correspondiente al valor x_{10}=10,107? Ya que la ventana tiene 9 puntos, tendremos que elegir los cuatro puntos más próximos a 10,107 por la izquierda y los cuatro más próximos por la derecha; en definitiva, debemos considerar los valores

$$y_6 , y_7 , y_8 , y_9 , y_{10} , y_{11} , y_{12} , y_{13} , y_{14}$$

que son 14,10, 17,37, 18,89, 16,95, 21,86, 19,03, 22,68, 23,43, 19,94.

Tabla 3.6. *Veintiún parejas de puntos (x,y).*

Observación	X	Y
1	1,155	10,98
2	2,795	11,83
3	3,056	11,34
4	3,843	13,96
5	5,214	16,27
6	5,597	14,10
7	5,828	17,37
8	8,697	18,89
9	9,428	16,95
10	10,107	21,86
11	11,778	19,03
12	11,959	22,68
13	12,083	23,43
14	12,751	19,94
15	13,739	20,92
16	16,475	24,56
17	16,833	24,23
18	17,625	29,38
19	18,070	26,80
20	18,935	28,38
21	19,758	28,76

Por tanto, el valor alisado correspondiente a $x_{10} = 10{,}107$ será la media aritmética de estos nueve valores, es decir, $y_{10}^{s} = 19{,}36$. En la Figura 3.5 aparece una representación gráfica de la ventana elegida para el alisamiento, así como el valor alisado, representado mediante el símbolo *, correspondiente a x_{10}.

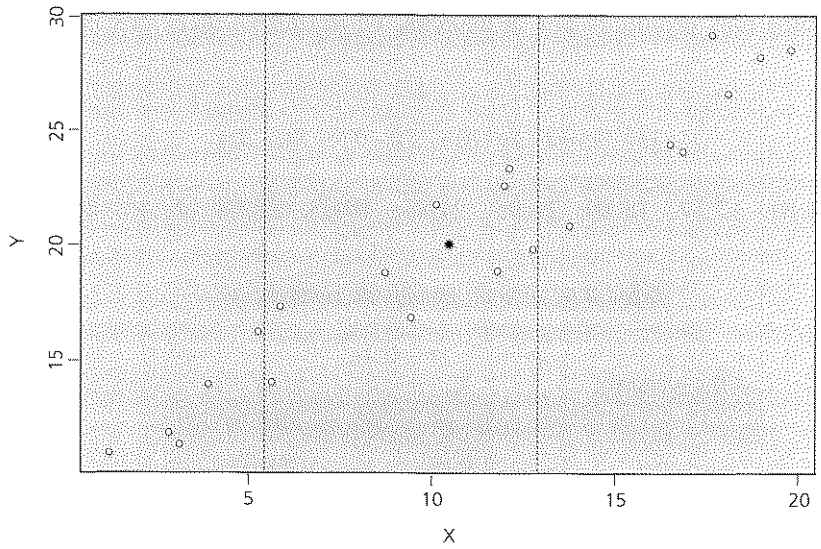


Figura 3.5. *Representación del valor alisado correspondiente a x_{10} y la ventana de alisamiento.*

Procediendo de igual forma para los restantes puntos e interpolando entre ellos, obtendríamos el alisamiento de la nube de puntos mediante el método del entorno simétrico más próximo. La única dificultad que nos encontramos para llevar a cabo esto es cuando queremos calcular el valor alisado correspondiente a un valor de X que esté en un extremo de su distribución; consideremos, por ejemplo, el cálculo del valor estimado correspondiente al valor $x_2 = 2{,}795$. La novedad que se presenta es que aquí no podemos elegir un intervalo simétrico pues a la izquierda de este valor tan solo está el $x_1 = 1{,}155$. En este caso el entorno de x_2 estará constituido por los cuatro puntos más próximos de su derecha, él mismo y los que se puedan tomar por la izquierda, el x_1; es decir, los valores de x_1 a x_6; en el extremo superior de la distribución se opera de forma similar. En la Figura 3.6 aparece la nube de puntos y su alisamiento representado por la línea continua.

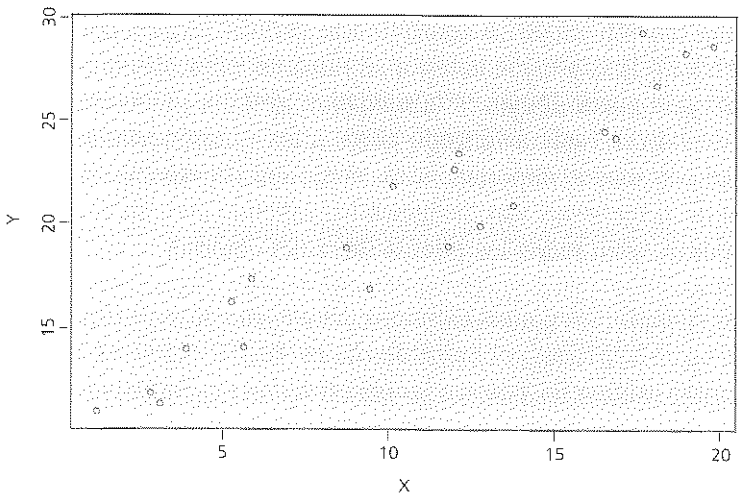


Figura 3.6. *La línea continua representa el alisamiento mediante medias móviles.*

El lector atento habrá observado que una manera de mejorar este método consiste en evitar, como se ha establecido implícitamente, el dar la misma importancia a todos los valores de un entorno dado. Antes definimos el entorno correspondiente a los 43 años como los individuos entre 37 y 48 años; no es dificil admitir que para hacer la estimación del IMC a los 43 años se debe dar más *peso* o importancia a los IMC de los que tienen 42 años que a los que tienen 37. Si se representa por w_j los pesos asignados a los valores y_j de un entorno dado, calcularemos el valor alisado mediante la siguiente expresión

$$y_i^s = \frac{\sum_{j \in N_i} w_j \, y_j}{\sum_{j \in N_i} w_j}$$

que no es más que la media ponderada de los y_j del entorno considerado.

Lo que queda por definir es la manera de asignar tales pesos; los llamados *alisamientos mediante núcleos (kernel smoothers)* definen los pesos mediante una función denominada *núcleo* que asigna la máxima importancia al punto x_i en el que se hace la estimación y la importancia de los puntos restantes del entorno va decreciendo dependiendo de su distancia a x_i. Quizás una de las formas más populares de asignar los pesos es mediante la llamada *función tricubo*, función definida como sigue

$$w_j = \left(1 - \left\{ \frac{|x_i - x_j|}{d_i} \right\}^3 \right)^3$$

donde d_i es la máxima distancia desde x_i a cualquier punto de su entorno N_i. De la expresión de la función tricubo podemos adivinar alguna de sus propiedades: en pri-

mer lugar, es una función simétrica respecto a la vertical trazada por x_i pues la diferencia que aparece en la expresión anterior se toma en valor absoluto; por otra parte, el máximo valor que puede tomar es la unidad y lo toma precisamente en el valor $x_j=x_i$, es decir, cuando se están asignando pesos a los y_j correspondientes a los x_j del entorno, el valor con mayor importancia es, como era de esperar, el y_i. Cuando nos vamos alejando de x_i los pesos van decreciendo, valiendo cero en el punto del entorno más alejado de x_i, y en todos los puntos fuera del entorno.

Para nuestro ejemplo, los puntos extremos del entorno de $x_{10}=10{,}107$ son los valores $x_6=5{,}597$ y $x_{14}=12{,}751$; así, el valor de d_{10} es 10,107-5,597=4.51. Por tanto, la función tricubo para asignar los pesos a las observaciones y_j correspondientes al entorno de x_{10} será

$$w_j = \left(1 - \left\{\frac{|x_i - x_j|}{4.51}\right\}^3\right)^3$$

Veamos un ejemplo; ¿qué peso asigna esta función al valor $y_{12}=22{,}68$?; según la ecuación anterior,

$$w_{12} = \left(1 - \left\{\frac{|10{,}107 - 11{,}959|}{4{,}51}\right\}^3\right)^3 = 0{,}806$$

La Figura 3.7 representa gráficamente los pesos asignados correspondientes a los puntos del entorno de x_{10}; como se puede apreciar, el máximo peso, la unidad, se le asigna al valor $y=21{,}86$ correspondiente a x_{10}, mientras los pesos decrecen conforme nos alejamos de x_{10}; gráficamente se puede comprobar como al valor $y=11{,}959$ le corresponde un peso de 0,80, aproximadamente.

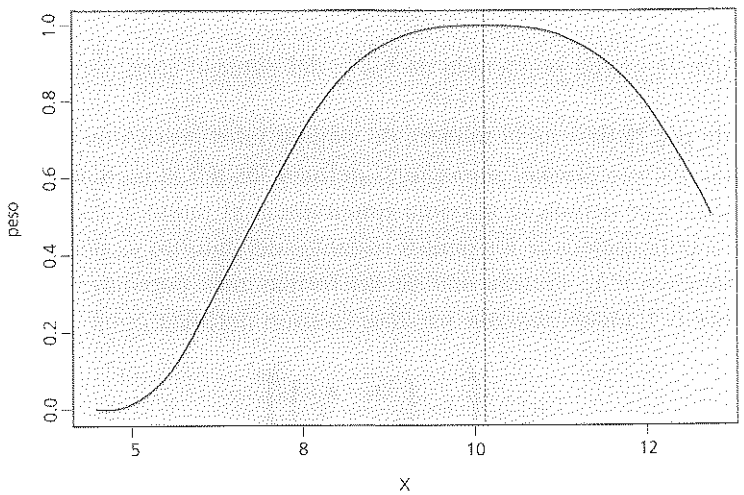


Figura 3.7. *La función tricubo en el entorno del valor 10,107.*

Obteniendo los pesos para los 9 puntos del entorno, no tenemos más que calcular la media ponderada según esos pesos; el valor estimado y_{10}^{s} es 20,30. Repitiendo el proceso para todos los puntos obtenemos el alisado de la nube de puntos; en la Figura 3.8 aparece la nube de puntos original, el alisado correspondiente a la técnica no ponderada, la curva de trazado continuo y, por último, la curva rayada correspondiente al criterio de ponderación mediante la función tricubo.

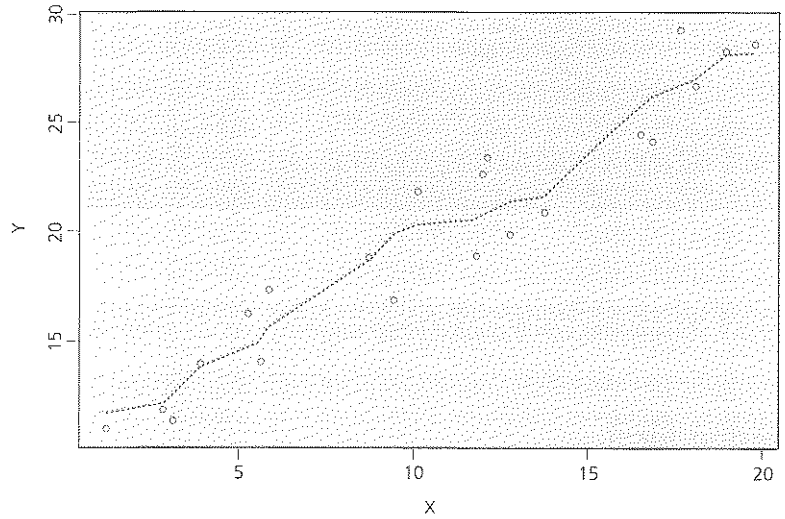


Figura 3.8. *Dos tipos de alisamiento: no ponderado (línea continua) y ponderado (línea rayada).*

Una cuestión importante no discutida hasta ahora es la proporción k de puntos a elegir para que entren en los entornos, pues eso condicionará el valor $[kn]$. No hay criterio claro para elegir el valor de k pues eso depende del número total de puntos de la nube; en general, a mayor tamaño de muestra, menor k; así, valores de k entre 0,1 y 0,7 son frecuentes de encontrar en la literatura. Pero, ¿qué importancia tiene el tomar una mayor o menor proporción de puntos?. Para responder a esta cuestión vamos a cambiar el valor de k y tomemos ahora k=0,6, lo que implica que el número de puntos por entorno será [(0,6)21=12,6] = 13 y repitamos el alisado con el criterio de ponderación mediante la función tricubo. En la Figura 3.9 aparece el nuevo alisado de la nube; en esta situación se dice que hay más alisamiento, la nueva curva que se puede construir con las nuevas estimaciones es más "regular" que en el caso anterior que era más "quebrada". Llevando las cosas a un extremo, si cada entorno contuviese un solo punto, evidentemente, y_i^s coincide con y_i, se ha reproducido la misma nube de puntos, el alisamiento producido es nulo; en el otro extremo, si en cada entorno entran todos los puntos de la nube, el alisamiento es máximo, esto es, *the trade-off between bias and variance*, Segal (1988).

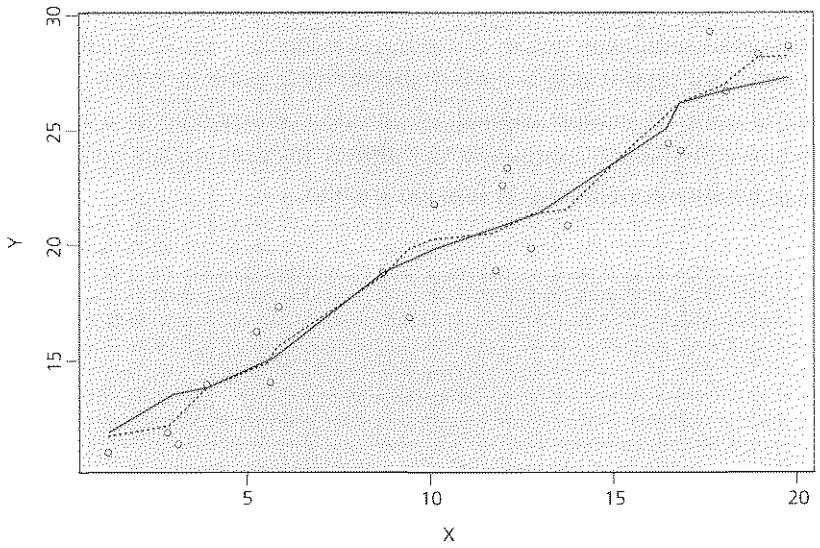


Figura 3.9. *Alisamiento con 9 puntos (línea rayada) y con 13 (línea continua).*

3.6.2. Regresión local

La metodología descrita hasta ahora se puede generalizar en varios sentidos; en primer lugar, la forma del entorno no tiene porqué ser simétrica; es decir, al hacer la estimación en x_{10} se podrían elegir los y_j correspondientes a los puntos más próximos, sin considerar cuantos quedan a la izquierda o a la derecha de x_{10}, dicho de otro modo, sin considerar que el entorno sea simétrico. En segundo lugar, los valores de y_j se pueden combinar de manera distinta a como se ha hecho hasta ahora, es decir, calculando su media ponderada o no; en este sentido es frecuente el uso de la llamada *regresión mínimo-cuadrática local* que, a pesar de su nombre, está basada en una idea bastante simple: se trata de calcular la recta de regresión a partir de las parejas de valores del entorno de x_i y calcular el valor y_i^s correspondiente a x_i mediante la ecuación

$$y_i^s = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

donde $\hat{\beta}_0$ y $\hat{\beta}_1$ son los coeficientes de la recta estimada para el entorno mediante mínimos cuadrados. Consideremos la estimación correspondiente al valor x_{10}; para ello calculamos la recta de regresión, correspondiente a los 9 puntos de la nube de su entorno, cuyos parámetros estimados son, como puede comprobar el lector, $\hat{\beta}_0$=10,87 y $\hat{\beta}_1$=0,8663, por lo que el valor alisado en x_{10}=10,107 será 10,87+0,8663(10,107)=19,63 La Figura 3.10 ilustra lo que se acaba de exponer

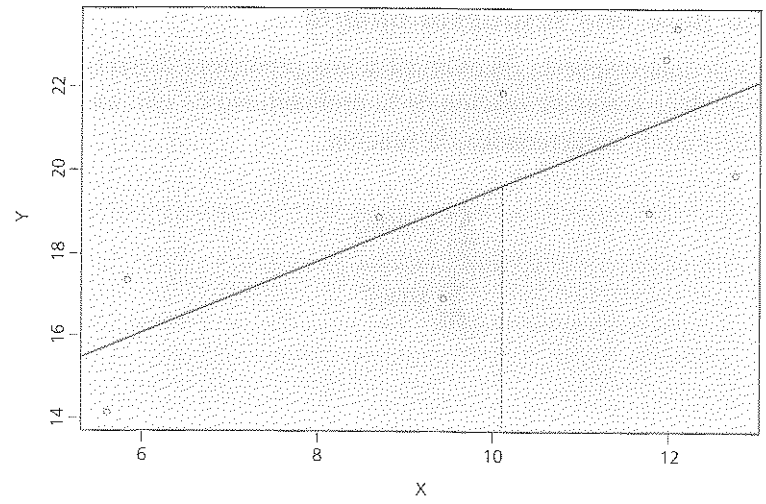


Figura 3.10. *Valor alisado correspondiente a x_{10} según la regresión local.*

Aunque a primera vista este método parezca muy distinto al anterior, en realidad es bastante similar en su filosofía pues, al fin y al cabo, en regresión lo que se estima es una media y en regresión lineal a cada observación se le asigna la misma importancia. Cleveland (1979) con su procedimiento LOWESS (*LOcally WEighted Scatterplot Smoothing*) generaliza esta situación considerando regresión lineal ponderada, asignando los pesos mediante la función tricubo antes mencionada y describe también un proceso iterativo para dar mayor robustez al alisado. La Figura 3.11 es el alisado de la misma nube mediante el método de Cleveland, tomando el 40% de los puntos en cada ventana y utilizando la función tricubo y sin ninguna interacción.

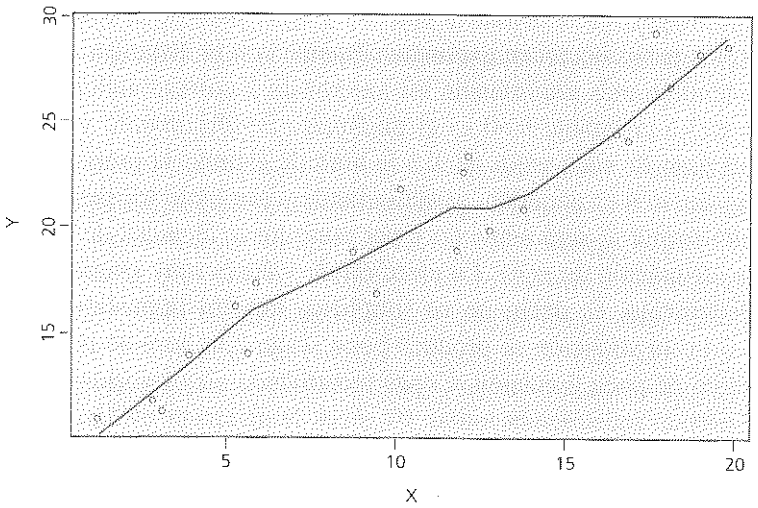


Figura 3.11. *Alisamiento según el método LOWESS.*

3.6.3. Alisado para una respuesta binaria

Para el caso de una respuesta dicotómica y una predictora discreta o continua X, Copas (1983) establece como estimación de la probabilidad de que un individuo, con valor x_i de X, presente la característica, la función

$$y_i^s = \frac{\sum_{j=1}^{n} y_j \psi(w_j)}{\sum_{j=1}^{n} \psi(w_j)}$$

donde

$$\psi(w_j) = e^{-w_j^2/2}$$

y

$$w_j = \frac{x_i - x_j}{h}$$

En definitiva, obsérvese que de lo que se trata es de calcular la estimación de la probabilidad en el valor x_i de X como una media ponderada de los valores y_i, que ahora son 0 o 1, correspondientes a todos los valores de X; ya que a la hora de calcular la probabilidad estimada en x_i, el peso de la observación viene dado por la expresión

$$\psi(w_j) = e^{-\frac{1}{2}\left(\frac{x_i - x_j}{h}\right)^2}$$

por lo que cuanto mayor sea la distancia entre x_i y x_j, el peso de la observación j será menor; es decir, los pesos decrecen exponencialmente con la distancia al valor x_i considerado, luego tendrán más importancia los valores de Y correspondientes a los valores próximos a x_i. La constante $h>0$ es el parámetro de alisamiento y de él también va a depender el peso de cada valor de X a la hora de estimar la probabilidad correspondientes en x_i; así, si h es pequeño, w_j será grande y por tanto los pesos $\psi(w_j)$ serán pequeños; de otra forma w_j será pequeño y por tanto los x_j alejados de x_i tendrán más influencia en la estimación de $p(x_i)$. La elección de valores muy pequeños de h lleva a muy poco alisamiento, es decir, no se simplifica nuestra impresión visual de las observaciones. En esta situación se produce poco sesgo a la hora de la estimación de $p(x_i)$ pero al precio de poca precisión. Copas recomienda elegir varios valores de h y como valor de partida recomienda tomar 10 veces la distancia media entre los valores de X. Representando mediante una nube de puntos el logit de la probabilidad estimada contra los valores de X se puede obtener una aproximación a la forma funcional en que X debe entrar en el modelo.

Royston (1992) propone otro método para evaluar la asociación entre la probabilidad y una variable continua, basado en sumas acumulativas. Hastie y Tibshirani (1986) publicaron una artículo sobre los llamados modelos aditivos generalizados como extensión de los lineales aunque manteniendo el caracter aditivo; el componente sistemático de tales modelos es de la forma

$$\alpha + f_1(x_1) + f_2(x_2) + ... + f_m(x_m)$$

donde las m funciones $f_i(x_i)$ son desconocidas y estimadas mediante alisamiento. Herman (1990) utiliza esta metodología para evaluar factores pronósticos de muerte neonatal.

3.6.4. Regresión por splines

Las distintas variantes de alisamiento vistas hasta ahora son ténicas de regresión no paramétrica pues la estimación de la variable respuesta no viene a través de la estimación de parámetros; el resultado de la estimación es una representación gráfica. A continuación vamos a ver otra forma de alisamiento mediante regresión paramétrica, la denominada regresión por *splines*. Antes se dijo que en ocasiones, la falta de no linealidad de la predictora se puede resolver añadiendo al modelo alguna alguna transformación como una potencia, logaritmo, etc.; por ejemplo, añadiendo un término al cuadrado, el modelo logístico quedaría de la forma

$$logit(p) = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2$$

pero el éxito de esta solución es limitado, aunque se consideren potencias de grado superior; dicho de otra forma, un polinomio de grado superior es más flexible que una recta, pero no lo suficiente, y se conocen relaciones entre variables que no pueden ser descritas mediante polinomios; por ejemplo, la relación entre dos variables tal que a partir de un cierto valor de la predictora la respuesta permanece prácticamente constante, se alcanza un *plateau*, no se puede describir mediante un polinomio. Sin embargo, y aunque parezca contradictorio con lo que se acaba de decir, para cualquier forma de relacion entre variables hay solución en base a un tipo especial de polinomios, los denominados splines.

El problema de los polinomios es que, aunque en ciertos rangos de valores de la predictora puedan representar bien a la relación entre las variables, en otros rangos de valores las cosas no son así; una solución por tanto podría venir del ajuste de distintos polinomios en distintos rangos de valores de la predictora. El spline más simple es el llamado lineal o de grado 1, cuya representación gráfica es un polinomio de

ese grado, es decir, un a recta, para cada rango de valores; un spline lineal se representa de la forma

$$\beta_0+\beta_1X+\beta_2(X-a)_+ +\beta_3(X-b)_+ +\beta_4(X-c)_+$$

donde los valores $a < b < c$, llamados *puntos de unión,* son los que permiten definir los, en este caso, cuatro rangos de valores de la predictora; el simbolo $(X-a)_+$ vale cero para todos los valores de X menores o iguales que a, y para valores mayores vale $(X-a)$; concretamente este es un spline lineal con tres puntos de unión a, b y c.

Este spline permite modelar la relación de forma distinta en cada uno de los cuatro intervalos que definen los tres puntos de unión a, b y c. En efecto, según este spline, para el rango de valores de la predictora que sean inferiores al punto de unión a, los componentes $(X-a)_+$, $(X-b)_+$ y $(X-c)_+$ son cero, por lo que, para este intervalo, el spline se reduce a $\beta_0+\beta_1X$, es decir, una recta de pendiente β_1. Para el rango de valores entre a y b, los componentes $(X-b)_+$, $(X-c)_+$ se anulan y $(X-a)_+=(X-a)$, por lo que el spline es de la forma

$$\beta_0+\beta_1X+\beta_2(X-a)_+ = (\beta_0-\beta_2a) + (\beta_1+\beta_2)X$$

es decir, para este segundo intervalo también es una recta, pero distinta a la anterior pues esta tiene pendiente, $\beta_1+\beta_2$, potencialmente diferente; el lector puede comprobar que para los otros dos intervalos las pendientes de las rectas asociadas son $\beta_1+\beta_2+\beta_3$ y $\beta_1+\beta_2+\beta_3+\beta_4$. En definitiva, para cada intervalo definido hay una recta potencialmente distinta, en funcion de que se anule alguno o algunos de los parámetros β_2, β_3 o β_4. En caso de un tamaño de muestra grande, es decir, si disponemos de muchos valores de la predictora, podríamos hacer un análisis más fino considerando más intervalos, por ejemplo 10, mediante la definición de 9 puntos de unión. Por tanto, un spline lineal no es más que un conjunto de rectas, polinomios de grado 1; por esta razón tambien se les conoce con el nombre de *polinomios a trozos (piecewise polynomials).*

Esta situación se puede generalizar permitiendo flexibilizar las relaciones dentro de cada intervalo mediante polinomios de grado superior a 1. Los más utilizados son los splines de grado 3 que son de la forma

$$\beta_0+\beta_1X+\beta_2X^2+\beta_3X^3+\beta_{13}(X-t_1)_+^3+\beta_{23}(X-t_2)_+^3+\ \ldots+\beta_{k3}(X-t_k)_+^3$$

siendo $t_1 < t_2 < \ldots < t_k$ los puntos de unión elegidos.

Para evitar algunos problemas de estimación que se suelen presentar en las colas de la distribución, es decir, cuando $X < t_1$ o $X > t_k$, Stone (1985) propuso que estas

funciones sean lineales en las colas. Eso conlleva una simplificación en la expresión del spline, que para el caso cúbico toma la forma

$$\beta_0+\beta_1 X+\sum_{i=1}^{k-2}\beta_{i+1}Z_i$$

donde las k-2 funciones Z_i son definidas como sigue

$$Z_i = (X - t_i)_+^3 - (X - t_{k-1})_+^3 \, (t_k - t_i) / (t_k - t_{k-1}) + (X - t_k)_+^3 \, (t_{k-1} - t_i) / (t_k - t_{k-1})$$

Ya que el spline cúbico es lineal en los parámetros, se pueden utilizar los paquetes estandars para su estimación mediante el método de los mínimos cuadrados, sin más que definir esas k-2 funciones Z_i. Para evaluar la no linealidad de una determinada predictora en el modelo logístico, no tendremos más que comparar la lejanía del modelo

$$logit(p) = \beta_0+\beta_1 X+\sum_{i=1}^{k-2}\beta_{i3}Z_i$$

con la de este otro

$$logit(p) = \beta_0+\beta_1 X$$

pues la diferencia entre esas dos lejanías se distribuye según una χ^2 con k-2 grados de libertad, la diferencia del número de parámetros entre los dos modelos. La representación gráfica del spline estimado contra X nos puede dar una pista, en caso de no linealidad, sobre la transformación a la que someter a la predictora.

Como ventajas del alisamiento mediante splines está el hecho de ser lo suficientemente flexible y, por otra parte, pueden utilizarse con los programas estandar; por otra parte, es posible evaluar la linealidad mediante contrastes de hipótesis.

Un inconveniente de esta metodología es la necesidad de establecer a priori los puntos de unión, por lo que hay que responder a la pregunta: ¿cuántos y cuáles elegir? No hay respuestas claras a estas dos preguntas y aunque la mayoría de los autores recomiendan tomar entre 3 y 5 puntos de unión, por los estudios de simulación publicados parece que no tiene demasiada transcendencia la elección. En caso de tomar 5 puntos de unión, se recomiendan tomar los percentiles 5, 25, 50, 75 y 95 de la distribución de X. Harrell (1988) presenta ejemplos de esta metodología aplicada a regresión logística y al análisis de supervivencia mediante el modelo de Cox.

Consideremos el problema de la linealidad de *LVOL* desde lo expuesto sobre los splines. Como el número de datos es pequeño, vamos a tomar solo 3 puntos de unión

que van a ser los percentiles 10, 50 y 90. Esos percentiles son los valores $t_1 = -0{,}5108$ $t_2 = 0{,}09531$ y $t_3 = 0{,}9933$, respectivamente. Recuérdese que para el spline cúbico, se debían definir tantas funciones Z_i como puntos de unión menos 2; como hemos elegido tres puntos de unión, tan solo tenemos que definir una variable, la Z_1. Según la expresión que antes se dió, para el caso que nos ocupa el modelo a ajustar sería

$$logit(p) = \beta_0 + \beta_1 X + \beta_2 Z_1$$

donde

$$Z_1 = (LVOL - t_1)_+^3 - (LVOL - t_2)_+^3 (t_3 - t_1) / (t_3 - t_2) + (LVOL - t_3)_+^3 (t_2 - t_1) / (t_3 - t_2)$$

que para los puntos de unión elegidos será

$$Z_1 = (LVOL - (-0{,}5108))_+^3 - (LVOL - 0{,}09531)_+^3 (0{,}9933 - (-0{,}5108)) / (0{,}9933 - 0{,}09531) + (LVOL - 0{,}9933)_+^3 (0{,}09531 - (-0{,}5108)) / (0{,}9933 - 0{,}09531)$$

Conocidas las estimaciones de β_0, β_1 y β_2, el spline cúbico resultante se puede expresar así

$$\beta_0 + \beta_1 LVOL + \beta_2 (LVOL - (-0{,}5108))_+^3 + \beta_3 (LVOL - 0{,}09531)_+^3 + \beta_4 (LVOL - 0{,}9933)_+^3$$

donde

$$\beta_3 = -\beta_2 (0{,}9933 - (-0{,}5108) / (0{,}9933 - 0{,}09531)$$

$$\beta_3 = -\beta_2 (0{,}09531 - (-0{,}5108) / (0{,}9933 - 0{,}09531)$$

Por tanto, para evaluar la linealidad de *LVOL* compararíamos la lejania del modelo

$$logit(p) = \beta_0 + \beta_1 X + \beta_2 Z_1$$

con la de este otro

$$logit(p) = \beta_0 + \beta_1 X$$

Si la diferencia en lejanías es mayor que el valor teórico para una χ^2 con 1 grado de libertad, la diferencia de parámetros estimados en cada modelo, nos informaría de la no linealidad de la predictora; en caso de similares lejanías los dos modelos serían indistinguibles, por lo que elegiríamos el más simple, que establece la linealidad de la variable. La Fig. 3.12 muestra el ajuste del modelo lineal y el del spline cúbico con los puntos de unión anteriores

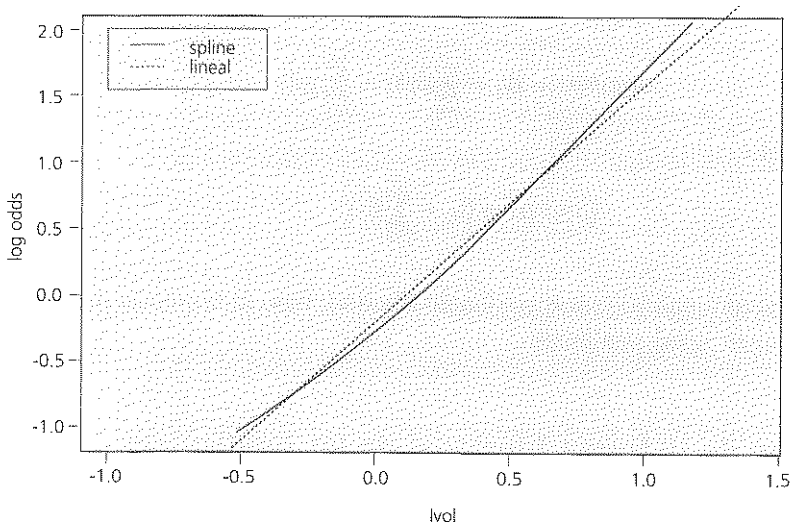


Fig. 3.12. *Ajuste lineal y mediante un spline cúbico.*

La diferencia en lejanía entre el modelo con *LVOL* y el que contiene a *LVOL* y a Z_1 es de 0,04 que, evidentemente, es no significativa para una χ^2 con k-2 = 3-2 = 1 grados de libertad, lo que está de acuerdo con la Fig. 3.12.

Minkin (1989 b) propone un método similar a la regresión por splines; se trata de una regresión por segmentos que también permite identificar el no cumplimiento de la hipótesis de linealidad. El método puede describirse así: sea X_m la variable de la que queremos estudiar la forma funcional; como en la regresión por splines, definimos k puntos de unión $v_1, v_2, ..., v_k$ y, a partir de ellos, consideremos las k+1 variables siguientes

$$\begin{aligned} Z_j &= 0 && \text{si} \quad X_m < v_{j-1} \\ &= X_m - v_{j-1} && \text{si} \quad v_{j-1} \leq X_m < v_j \\ &= v_j - v_{j-1} && \text{si} \quad X_m > v_j \end{aligned}$$

para j=2, 3, ..., k. Para j=1

$$\begin{aligned} Z_1 &= X_m && \text{si} \quad X_m < v_1 \\ &= v_1 && \text{si} \quad X_m \geq v_1 \end{aligned}$$

y, por último, para j=k+1

$$\begin{aligned} Z_{k+1} &= X_m - v_k && \text{si} \quad X_m > v_k \\ &= 0 && \text{si} \quad X_m \leq v_k \end{aligned}$$

Con estas variables definidas, la diferencia entre la lejanía del modelo

$$logit(p) = \beta_0 + \sum_{i \neq m}^{p} \beta_i X_i + \sum_{j=1}^{k+1} \gamma_j Z_j$$

y la del modelo

$$logit(p) = \beta_0 + \sum_{j=1}^{p} \beta_i X_i$$

se distribuye según una χ^2 con k grados de libertad, bajo la hipótesis de que todos los segmentos tengan igual pendiente.

3.7. Elección de la forma funcional de las predictoras. Métodos basados en los residuales

3.7.1. Nube de puntos para una variable añadida

Como muchos métodos diagnósticos en regresión logística, la *nube de puntos de una variable añadida, (the added variable plot)*, es una adaptación del existente para el caso de la regresión lineal. El objetivo del método es la evaluación de la necesidad de, dado un modelo logístico ajustado con predictoras X_1, X_2, ... X_m, introducir nuevas variables y, como caso particular, transformaciones de las variables ya incluidas en el modelo ajustado.

Supongamos que Z representa una nueva variable o una transformación: potencia, logaritmo, etc., de una variable presente en el modelo ajustado y z_i los valores de tal variable en los distintos individuos. Los residuales

$$(z_i - \hat{z}_i)\sqrt{w_i}$$

se denominan *z-residuales* o *residuales de la variable añadida*, siendo $\hat{z}_i$ los valores predichos mediante la regresión lineal, por mínimos cuadrados ponderados, entre Z y las variables X_1, X_2, ... X_m incluidas en el modelo ajustado, $w_i = n_i\hat{p}_i(1-\hat{p}_i)$ son los pesos con los que se pondera la regresión mínimo-cuadrática y $\hat{p}_i$ son las probabilidades, según el modelo logístico que no contiene a Z, asociadas a los distintos individuos. La nube de puntos entre los residuales de Pearson definidos en el apartado 3.4 y los z-residuales es un método gráfico para evaluar la pertinencia de la inclusión de la variable Z en el modelo ya que esta representación gráfica es un reflejo, como demostró Wang(1985), de la relación entre el logit(p) y la variable Z. Así, si no hay ninguna tendencia en esta nube de puntos es una indicación de que la variable Z no aporta nada nuevo al ajuste, mientras que una tendencia lineal en tal nube de puntos habla de la necesidad de incluir la variable Z en el modelo ajustado previamente.

Supongamos ajustado el modelo logístico que contiene como predictora a la variable *LTAS* y estudiemos la pertinencia de incluir en tal modelo a la variable *LVOL*; una vez ajustado el modelo que contiene solo a *LTAS* como predictora, cualquier programa proporciona las probabilidades ajustadas, por lo que podremos calcular los pesos w_i.

En este caso particular, la variable *LVOL* juega el papel de la variable Z por lo que deberemos estimar los valores predichos para *LVOL* mediante la regresión por mínimos cuadrados ponderados tomando como predictora a *LTAS*, la variable incluida en el modelo logístico, lo que se puede conseguir con cualquier programa estandar. Por tanto, podemos tener toda la información necesaria para calcular los *LVOL*-residuales; enfrentando éstos con los de Pearson del modelo ajustado sólo con *LTAS* obtenemos la nube de puntos de la Fig. 3.13, donde la línea continua representa el alisamiento de tal nube. De esta representación gráfica se deducen claros indicios de linealidad, lo que hace pensar en la necesidad de introducir *LVOL* en el modelo que incluye a *LTAS*.

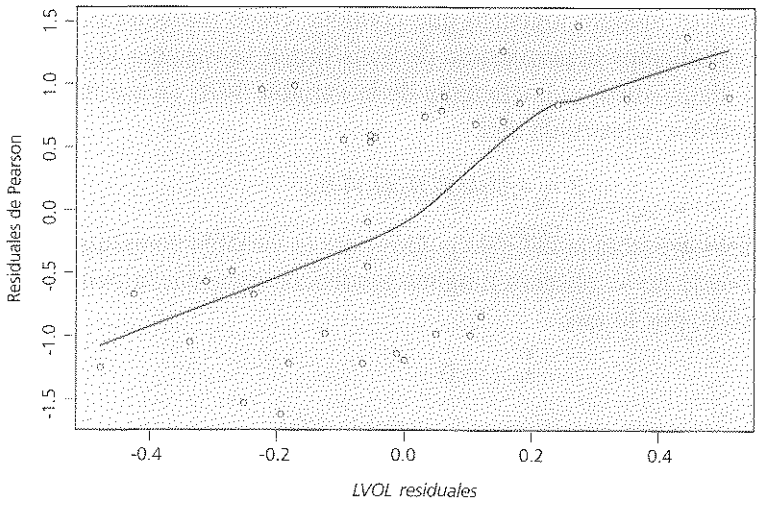


Figura 3.13. *Alisado de la nube de puntos para la variable añadida LVOL.*

3.7.2. Nube de puntos para una variable construida

Box (1962) estableció un método general como ayuda a la selección de la transformación de las variables predictoras para el modelo lineal, que fue adaptado por Wang (1987) para los modelos lineales generalizados.

Sea X una predictora; se trata se evaluar si la transformación

$$X^{(\lambda)} = \begin{cases} \dfrac{X^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ \log X, & \lambda = 0 \end{cases}$$

produce un mejor ajuste del modelo, donde para el caso $\lambda=0$ la transformación es la logarítmica. Para ver si este parámetro λ es distinto de la unidad, se define la que denominaremos *variable construida* Z de la siguiente forma $Z = \hat{\beta} X \log X$ donde $\hat{\beta}$ es el coeficiente estimado para la variable X del modelo logístico ajustado. El método diagnóstico, denominado *nube de puntos para la variable construida (constructed variable plot)*, no es más que una nube de puntos para la variable añadida Z; de manera similar al procedimiento anterior, las cantidades

$$(z_i - \hat{z}_i)\sqrt{w_i}$$

se conocen con el nombre de *residuales de la variable construida*. La nube de puntos que resulta de enfrentar los residuales de Pearson con los z-residuales de la variable construida nos permite evaluar la transformación a elegir, pues el citado autor demuestra que la pendiente de la nube de puntos es una estimación de λ-1; así, si la pendiente de la nube de puntos es cero, la variable X no necesitará ser transformada.

Consideremos el modelo logístico que contiene como predictoras a las variables *LTAS* y *VOL* y estudiemos la pertinencia de transformar la variable *VOL* haciendo uso de este método diagnóstico; el coeficiente de la variable *VOL* en ese modelo es 4,504 por lo que la variable construida será

$$Z = 4{,}504(VOL)(LVOL)$$

La Figura 3.14 presenta la nube de puntos para la variable construida que se acaba de definir, en donde no se observa ninguna tendencia, es decir, la pendiente es aproximadamente cero, por lo que

$$\hat{\lambda} - 1 \approx 0$$

por tanto no es necesario transformar la variable *VOL*.

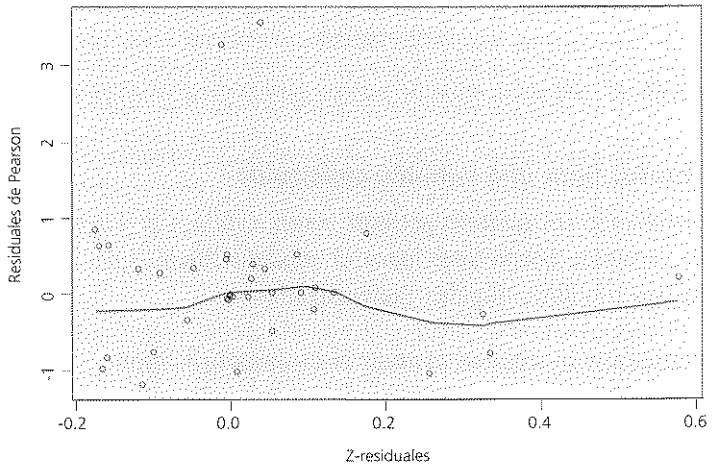


Figura 3.14. *Nube de puntos para la variable construida.*

3.7.3. Nube de puntos de los residuales parciales

Landwehr (1984) propone el método denominado *nube de puntos de los residuales parciales (partial residuals plots)*, que permite evaluar la forma funcional óptima, en el sentido de mayor bondad de ajuste, en que una predictora debe entrar en el componente sistemático del modelo logístico. El método de Landwehr puede describirse brevemente así: sea Z la variable de la que se quiere estudiar la forma funcional; estos autores establecen el modelo

$$logit(p) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \delta f(Z)$$

donde $f(Z)$ es precisamente la función desconocida de la variable Z que se trata de conocer. Para ello proponen proceder de la siguiente manera; en primer lugar, se ajusta el modelo

$$logit(p) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \delta Z$$

que da lugar a unas estimaciones de δ, β_i y p_i, donde p_i es la probabilidad correspondiente al individuo i; a partir de estas estimaciones se pueden calcular los denominados *residuales parciales*

$$r_i^{par} = \frac{y_i - \hat{p}_i}{\hat{p}_i (1 - \hat{p}_i)} + \hat{\delta} z_i$$

Estos autores demuestran que mediante un procedimiento de alisado de la nube de puntos resultante de representar en el eje de ordenadas los valores r_i^{par} y z_i en el de abcisas, se obtiene una estimación de la forma funcional de $f(Z)$ desconocida; si esta relación es lineal es indicativa de la no necesidad de transformación alguna. Parece pertinente, una vez elegida $f(Z)$, realizar el correspondiente test para determinar si es interesante o no el aumento de la complejidad del modelo.

Con el modelo ajustado se ha establecido la dependencia lineal entre *logit(p)* y las predictoras *LTAS* y *LVOL*; esta suposición no tiene porqué cumplirse por lo que es necesario comprobar las posible no linealidad; para ello utilizaremos el gráfico de los residuales parciales con su alisamiento correspondiente. Si el interés es la linealidad de *LVOL*, ajustando el modelo con las dos variables obtenemos un valor $\hat{\delta} = 5{,}179$, no tenemos más que sustituir en la expresión para obtener los residuales parciales.

La Figura 3.15 muestra los residuales parciales y su alisamiento. Como se ve, la relación entre los residuales parciales y *LVOL* es lineal; así, parece que no es necesario introducir en el modelo ninguna transformación de la variable *LVOL*. La recta de mínimos cuadrados de esa nube de puntos, la recta rayada, tiene como pendiente precisamente el coeficiente, 5,179, de la variable *LVOL*.

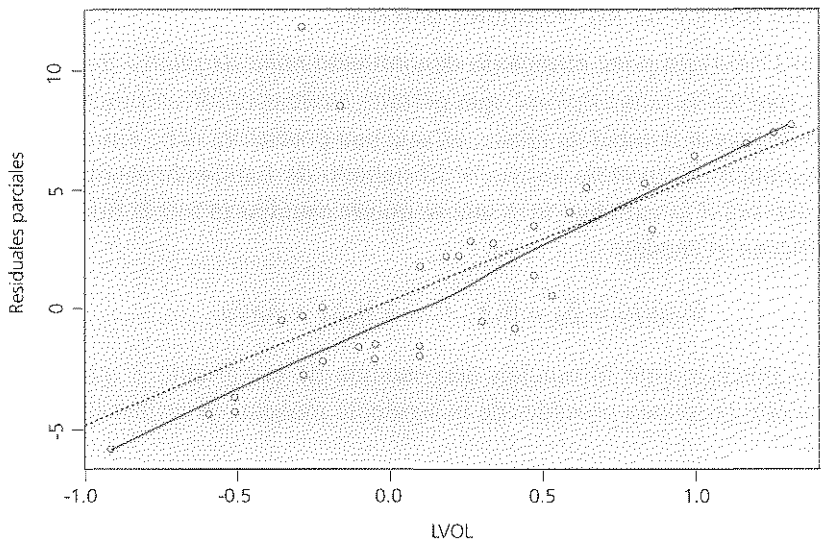


Figura 3.15. *Alisado de los residuales parciales para la variable LVOL.*

Fowlkes (1987) da una version "alisada" de estos residuales parciales, definidos mediante la siguiente expresión

$$r_i^{par} = \frac{\hat{p}_i^s - \hat{p}_i}{\hat{p}_i(1 - \hat{p}_i)} + \hat{\delta}_{Z_i}$$

donde $\hat{p}_i^s$ es el valor alisado multivariante de las observaciones de la variable resultado y z_i son los valores de la variable Z, respectivamente, en el individuo i.

3.8. Importancia de las distintas observaciones

Hasta ahora se han expuesto algunos de los estadísticos que dan idea de la bondad del ajuste del modelo, aunque estos presentan el inconveniente de dar una idea global del ajuste; sin embargo, es deseable disponer de técnicas que puedan detectar "observaciones especiales" que no sigan el patrón general de los datos. En regresión lineal existen una gran variedad de procedimientos diagnósticos definidos con la finalidad de poner en evidencia diferentes aspectos relativos al modelo ajustado; las monografías de Cook and Weisberg (1982) y de Atkinson (1985) son dos referencias excelentes en este tema. Aunque el estudio de estas cuestiones viene de lejos en el caso de la regresión lineal, para el modelo de regresión logística es Pregibon (1981) quien establece las bases de tales procedimientos diagnósticos.

3.8.1. Cambios en la lejanía y el estadístico χ^2

Un problema importante en la selección del modelo es el hecho de que unas pocas observaciones pueden tener una influencia exagerada en el ajuste del modelo; la consideración o no de un individuo en el análisis puede afectar notablemente a la estimación de alguno o algunos coeficientes del modelo, o bien, es posible que un individuo o categoría de individuos determine si una predictora está o no asociada a la variable resultado. Cuando se dispone de una sola predictora, las representaciones gráficas son herramientas muy útiles para tratar con este problema pero cuando, como en la mayoría de las ocasiones, se dispone de varias predictoras, las cosas se complican.

Un primer aspecto interesante es el examen de la repercusión que puedan tener pequeñas perturbaciones en los datos observados sobre los estadísticos que miden la bondad del ajuste al modelo; una forma común de efectuar tales perturbaciones consiste en la omisión de una observación; es decir, de lo que se trata es de medir la alteración que produce la ausencia de cada observación aislada sobre los diversos estadísticos del modelo ajustado. Para ello notemos por D la lejanía del modelo ajustado con todas las observaciones y por D_i la lejanía calculada para el mismo modelo pero sin considerar la observación i; por tanto, la repercusión de la omisión de esta observación sobre la lejanía será

$$\Delta D_i = D\text{-}D_{(i)}$$

El cálculo de este estadístico puede ser tedioso cuando el número de individuos sea grande, pues tendríamos que ajustar n+1 modelos, por lo que Pregibon (1981) propuso la siguiente aproximación

$$\Delta D_i \doteq d_i^2 + \frac{r_i^2\, h_i}{(1 - h_i)}$$

donde r_i y d_i son los residuales de Pearson y de la lejanía correspondientes al modelo ajustado con todas las observaciones, definidos en el apartado 3.4; obsérvese cómo se puede conseguir el cambio aproximado en lejanía con sólo ajustar el modelo que contiene a todas las observaciones. Sustituyendo d_i por r_i se obtiene esta otra expresión

$$\Delta D_i = \frac{d_i^2}{(1 - h_i)}$$

que sigue, aproximadamente, una distribución χ^2 con un grado de libertad, por lo que su raiz cuadrada, denominada *residual de la lejanía estandarizado*,

$$d_i^a \doteq \frac{d_i}{\sqrt{1 - h_i}}$$

sigue una normal estandarizada. Los valores h_i que aparecen en estas expresiones son los denominados *leverage*, los componentes de la diagonal principal de la *matriz sombrero (hat)*; este nombre proviene del hecho de que tal matriz H, en regresión lineal, tiene la propiedad de que

$$\hat{y} = H\, y$$

es decir, multiplicando tal matriz por los valores de la predictora se consiguen las estimaciones de la respuesta, según el modelo ajustado.

En regresión lineal los valores h_i son números comprendidos entre 0 y 1 y gozan de la propiedad de que valores grandes de h_i indican que los individuos correspondientes son extremos en el espacio de diseño, lo que significa que los valores de las predictoras no son valores "estandar"; dicho de otro modo, un leverage grande indica que el individuo correspondiente se aleja del "valor medio" en términos de las predictoras. Valores de h_i superiores a la cantidad $2(m+1)/n$ se consideran como indicadores de que la observación correspondiente es potencialmente influyente.

En regresión logística no necesariamente todo valor alto de h_i implica valor extremo en el espacio de las predictoras, ni la matriz sombrero cumple exactamente la propiedad antes comentada; Hosmer and Lemeshow (1989) discuten el significado del leverage en regresión logística.

Un estudio similar al hecho con la lejanía se puede realizar con el otro estadístico de bondad de ajuste, la χ^2; sin embargo, mientras la eliminación de una observación hace siempre disminuir el valor de la lejanía, es posible que el valor de la χ^2 aumente, aunque realmente solo en ocasiones excepcionales. De manera análoga a lo antes visto para la lejanía, el cambio producido en la χ^2 por el hecho de no considerar la observación i es

$$\Delta\chi_i^2 = \chi^2 - \chi_{(i)}^2$$

o su aproximación

$$\Delta\chi_i^2 = \frac{r_i^2}{1 - h_i}$$

La raiz de este estadístico es el residual de Pearson estandarizado que toma la forma

$$r_i^a = \frac{r_i}{\sqrt{1 - h_i}}$$

también conocido como *residual ajustado*.

En la Fig. 3.16 aparecen los valores del leverage para todos los individuos, siendo 0.163, 0.168 y 0.246 los correspondientes a las observaciones 13, 29 y 31, respectivamente.

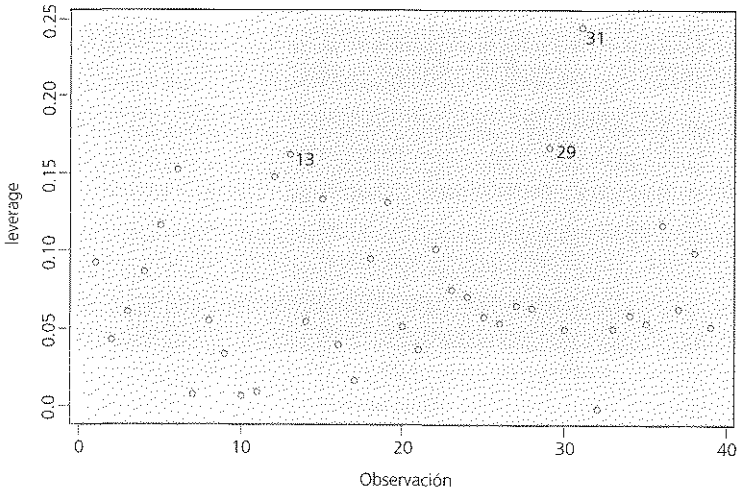


Figura 3.16. *Valores h_i para los distintos individuos.*

La observación 31, la que tiene mayor leverage, no tiene como valores de las predictoras que estén fuera del patrón general; antes se apuntó esta posibilidad en regresión logística.

Las Figuras 3.17 y 3.18 muestran los cambios en lejanía y chi-cuadrado, junto con sus aproximaciones.

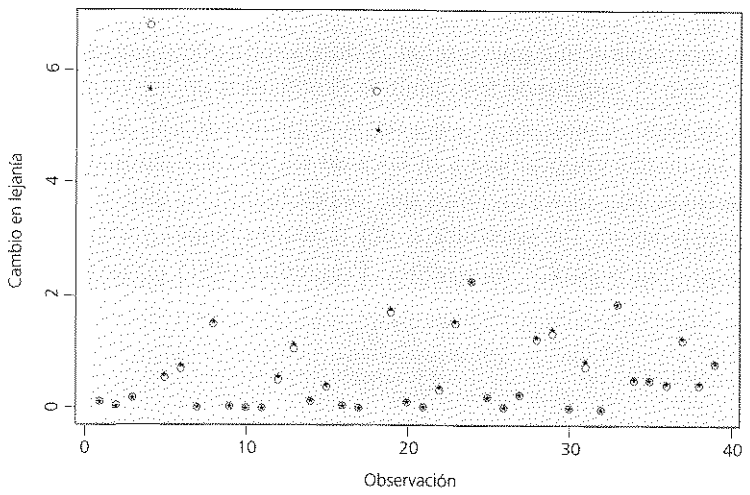


Figura 3.17. *Cambio exacto y aproximado (*) en la lejanía al suprimir cada observación.*

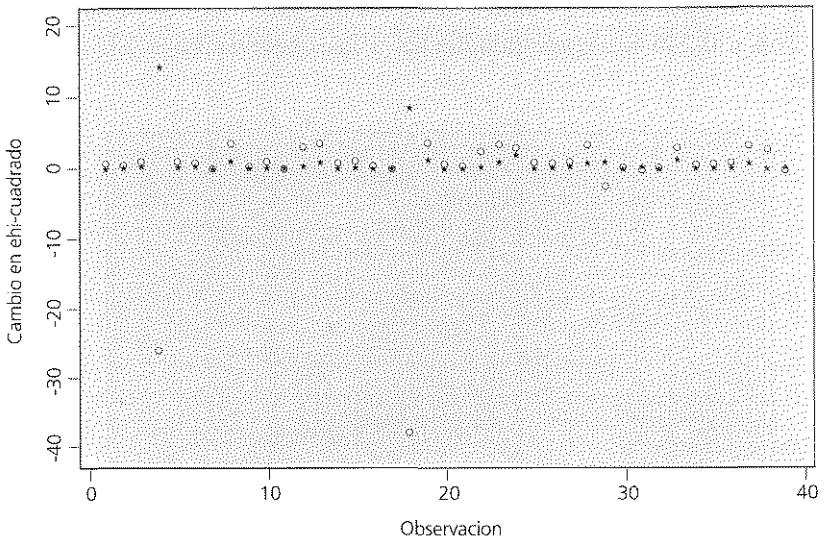


Figura 3.18. *Cambio exacto y aproximado (*) en la chi-cuadrado al suprimir cada observación.*

De la Figura 3.17 se puede deducir como, a excepción de los puntos 4 y 18, hay bastante acuerdo entre el cambio en la lejanía y la aproximación a él. Sin embargo, como muestra la Figura 3.18, la aproximación al cambio en la χ^2 no es tan buena y además, el omitir algunos puntos a la hora de ajustar el modelo puede producir un aumento en la χ^2, de ahí la presencia de algunos valores negativos la representación gráfica.

3.8.2. Cambios en la estimaciones de los parámetros

Aparte de la repercusión que pueda tener la omisión de uno o más individuos sobre la lejanía y la χ^2, está el estudio de los cambios producidos en las estimaciones de los coeficientes del modelo. Como en el caso de la regresión lineal, no todas las observaciones tienen la misma importancia a la hora de estimar los parámetros del modelo de regresión logística; hay observaciones que dependiendo de que se tengan o no en cuenta, las estimaciones de los parámetros del modelo pueden variar sensiblemente; esos son los denominados *puntos influyentes*. Pregibon (1981) propuso el estadístico

$$c_i^a = \frac{r_i^2 \, h_i}{(1 - h_i)^2}$$

como medida del cambio que experimenta el conjunto de las estimaciones de los coeficientes del modelo por la omisión de la observación i; esta medida es la análoga a la *distancia de Cook* utilizada en regresión lineal; así, valores grandes de este esta-

dístico corresponden a los posibles puntos influyentes; el valor del estadístico c_i^a es directamente proporcional tanto a los residuales de Pearson como a los leverage; así, una observación con un valor alto de leverage será necesariamente un punto influyente, a menos que su correpondiente residual de Pearson sea muy pequeño.

Por tanto, conocidos los leverage y los residuales de Pearson, se puede conseguir la distancia de Cook cuyos valores aparecen representados en la Figura 3.19.

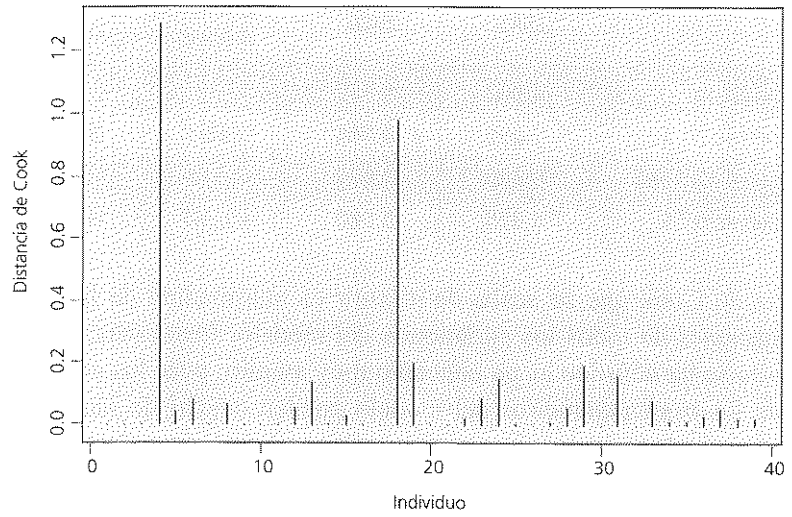


Figura 3.19. *Distancia de Cook para cada observación.*

Como se puede observar, los puntos 4 y 18 son los que tienen mayor distancia de Cook por lo que esos pueden ser puntos influyentes. Es importante resaltar el hecho de que la distancia de Cook depende tanto del leverage como del residual de Pearson; la Tabla 3.7 pone de manifiesto como aunque las observaciones 13, 29 y 31 eran las que tenían mayor leverage, sus influencias son, aproximadamente, seis veces menores que las de las observaciones 4 y 18.

Tabla 3.7. *Influencia del leverage y el residual de Pearson en la distancia de Cook.*

Observación	r_i^2	h_i	c_i^a
4	12,357	0,087	1,286
18	8,435	0,095	0,983
13	0,601	0,163	0,139
29	0,807	0,168	0,196
31	0,384	0,246	0,166

Para ver la repercusión real de estas dos observaciones en la estimaciones de los parámetros del modelo, también lo hemos ajustado sin ellas. En la Tabla 3.8 aparecen los resultados de tales ajustes junto con los del modelo sin la observación 5 cuya distancia de Cook es pequeña; es evidente la influencia de las observaciones 4 y 18 en el modelo ajustado.

Tabla 3.8. *Coeficientes del modelo según la presencia o ausencia de las observaciones 4, 18 y 5.*

	LTAS	LVOL	Diferencias con el modelo completo	
Modelo completo	4,562	5,179		
Sin la observación 4	7,455	8,468	-2,893	-3,289
Sin la observación 18	6,880	7,671	-2,318	-2,492
Sin la observación 5	4,315	5,080	+0,247	+0,099

La distancia de Cook que se acaba de discutir es una medida de la repercusión sobre las estimaciones de todos los parámetros del modelo. Realmente, la influencia como hasta ahora se ha definido es un concepto muy amplio, pues los cambios en las estimaciones de los $m+1$ parámetros del modelo pueden ocurrir de muchas formas; por ejemplo, una observación puede ser muy influyente para la estimación de unos pocos parámetros y nada influyentes para el resto, o bien, afectar moderadamente a todas las estimaciones. Además, hay ocasiones en que el interés está en el cambio producido en la estimación de un solo parámetro; como ya se dijo en el Capítulo II, en muchos estudios epidemiológicos una de las variables, el posible factor de riesgo, juega un papel diferente al resto de las variables debido a que éstas están presentes en el modelo con el ánimo exclusivo de controlar la posible confusión; el objetivo está en medir, lo más exactamente posible, la estimación del parámetro correspondiente al factor de riesgo; por tanto, en situaciones como ésta interesará medir la influencia de cada observación en la estimación de este parámetro y no en el conjunto de todas las estimaciones. Para evitar, como en casos anteriores, el hecho de tener que ajustar $n+1$ modelos, se ha propuesto una aproximación al cambio estandarizado dada por la expresión

$$\Delta i\,\hat{\beta}_j = \frac{V_j\, x_i\, (y_i - y_i)}{e.e.\ (\hat{\beta}_j)\ (1 - h_i)}$$

donde V_j es la j-ésima fila de la matriz de varianzas-covarianzas del modelo que contiene a todas las observaciones y x_i la matriz columna de los valores las predictoras correspondientes a la observación i; el indice j varía ahora entre 1 y $m+1$, correspondiendo el valor 1 al término independiente, el valor 2 a la variable que aparece en primer lugar en el modelo, etc.

Consideremos ahora que el parámetro de interés es solo el correspondiente a la variable *LTAS* y que la presencia de *LVOL* es con el ánimo exclusivo de controlar el posible efecto confusor entre *LTAS* y la respuesta *RESP*; veamos como calcular una aproximación al cambio de la estimación del coeficiente de *LTAS* debido a la ausencia de la observación 12. Para ello una vez ajustado el modelo en el que aparece *LTAS* como primera variable y que contiene a todas las observaciones, podemos obtener la matriz de varianzas-covarianzas de las estimaciones de los parámetros de ese modelo, cuya segunda fila, la correspondiente a *LTAS*, tiene de componentes (-2,261, 3,374, 2,753); los valores de las predictoras para la observación 12 son 1,0116 y -0,59784, la respuesta es 0 y el valor predicho para ella por el modelo es 0.20469; por último, el error estándar del estimador del coeficiente de *LTAS* es 1,837, la raiz de su varianza 3,374, y su leverage es 0,1481. Por tanto, una aproximación al cambio por la ausencia de la observación 12 será

$$\frac{(-2{,}261 \quad 3{,}374 \quad 2{,}753)\begin{pmatrix} 1 \\ 1{,}0116 \\ -0{,}59784 \end{pmatrix}(0 - 0{,}20469)}{1{,}837\,(1 - 0{,}1481)} = 0{,}065$$

La estimación del parámetro de *LTAS* tras eliminar la observación 12 es 4,458 por lo que

$$\frac{(4{,}562 - 4{,}458)}{1{,}837} = 0{,}057$$

es el cambio estandarizado exacto, parecido al aproximado antes calculado; en la Figura 3.20 se puede observar que los puntos más influyentes sobre este coeficiente vuelven a ser el 4 y el 18.

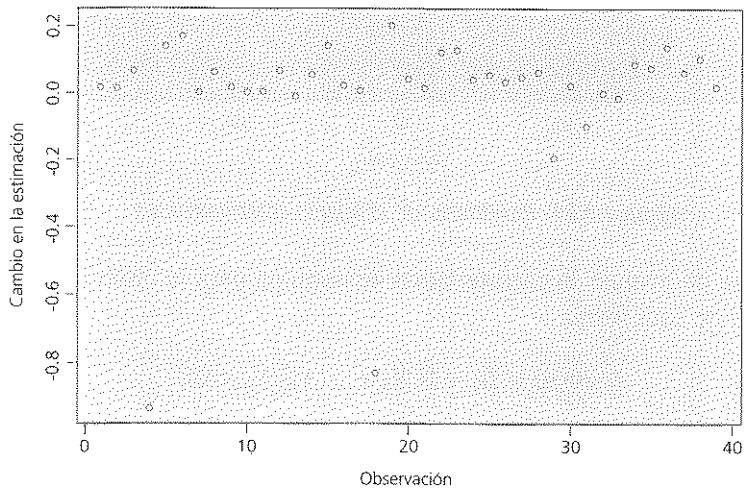


Figura 3.20. *Cambio estandarizado del coeficiente de LTAS eliminando cada observación.*

Hasta ahora la influencia de una observación se ha contemplado como su repercusión sobre la estimación de los parámetros del modelo; sin embargo, esta no es la única faceta de interés de un modelo ajustado. Así, Johnson (1985) propone medidas para detectar la influencia de las observaciones en relación a las probabilidades estimadas y a la clasificación de futuros individuos.

3.9. Observaciones extremas

Aparte de la posible influencia que pueden ejercer algunas observaciones en la estimación del modelo, hay observaciones que no se ajustan bien al patrón general de los datos; son las llamadas *observaciones extremas (outliers)*; ya que los residuales dan idea del ajuste de cada observación al modelo propuesto, aquellas que tengan residuales "grandes" serán observaciones extremas. Para tal fin se suelen utilizar las nubes de puntos de los residuales ajustados, bien de Pearson bien de la lejanía, que se definieron en el apartado 3.8.1. La Figura 3.21 muestra los residuales de la lejanía estandarizados, donde se puede observar la magnitud de los correspondientes a las observaciones 4 y 18.

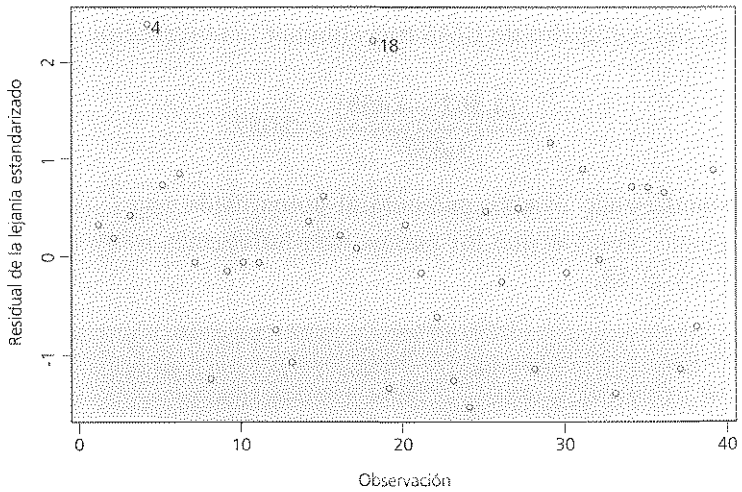


Figura 3.21. *Residuales de la lejanía estandarizados.*

Otra manera de evaluar la característica de extrema para la observación i es construir el modelo de esta forma

$$logit(p) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \delta Z$$

donde Z es una variable indicador que toma el valor cero para todas las observaciones salvo para el individuo i que toma el valor 1; de esta manera contrastando la hipótesis $\delta = 0$, se puede comprobar si la observación i sigue el patrón general o no.

Consideremos la observación $i=4$ por lo que $Z=1$ para ese individuo y cero para los restantes; si ajustamos el modelo que contiene como predictoras *LTAS*, *LVOL* y *Z* da lugar a una lejanía de 22,426 con un cambio, respecto al modelo que contiene a *LTAS* y *LVOL*, de 6,80 para 1 grado de libertad. De la misma forma, para la observación 18 el cambio producido en la lejanía es 5,64, también con un grado de libertad.

Antes de decidir declararlas o no como observaciones raras conviene discutir algunas cuestiones previas. En primer lugar, no todo residual grande corresponde necesariamente a una observación de este tipo; téngase en cuenta que estamos tratando con una distribución de probabilidad y valores raros, aunque infrecuentes, se pueden dar. En segundo lugar, hay que distinguir si se ha hecho o no alguna hipótesis previa sobre la posible rareza de una determinada observación. Así, si antes del análisis, por el conocimiento del problema tratado, se establece tal hipótesis, el test para decidir si declararla como rara está, aproximadamente, basado en la distribución normal. Por otra parte, si como es lo más frecuente, sospechamos de alguna observación una vez vistos los residuales, hay que tomar alguna precaución; el procedimiento de Bonferroni establece un error α/n.

¿Qué hacer una vez detectada una observación como extrema? Si su ausencia no afecta a cuestiones de interés, no habrá problema en dejarla en el conjunto de datos. Si acaso la observación es influyente lo más sensato puede ser la presentación de los resultados derivados de los modelos con y sin tal o tales observaciones. De todas formas, la discusión de estos extremos en un problema concretova más allá de la estadística y se hace necesaria la colaboración del investigador del problema tratado.

3.10 Colinealidad

Un problema que se puede presentar al utilizar la regresión logística es, como en otros modelos de regresión, que dos o más predictoras estén tan interrrelacionadas entre sí que sea dificil separar sus efectos sobre la variable resultado. Este fenómeno, conocido como *colinealidad* se define como la existencia de una dependencia lineal entre algunas de las predictoras del modelo y tiene como consecuencia la imprecisión en la estimación de los coeficientes del modelo. Aunque casi siempre existen algún grado de colinealidad entre las predictoras, es cuando ésta es grande cuando se puede presentar el problema.

A modo de ejemplo, consideremos el modelo ajustado de la Tabla 3.3 y supongamos que tenemos otra predictora *X* muy correlacionada con *LVOL* y *LTAS*; tal variable *X* se puede conseguir añadiendo a la suma *LVOL+LTAS* una pequeña perturbación mediante una variable que siga una distribución normal de media cero y desviación

típica 0,05; pues bien, ajustando el modelo que contiene a *LVOL*, *LTAS* y a *X*, se obtienen los resultados de la Tabla 3.9.

Tabla 3.9. *Estimaciones y errores estándar de los coeficientes de las variables LTAS y LVOL en presencia de colinealidad con X.*

VARIABLE	ESTIMACIÓN	E.E.
Constante	-2,710	1,356
LTAS	-2,735	10,564
LVOL	-1,830	10,136
X	7,042	10,170

Ahí se puede apreciar el gran cambio producido en las estimaciones y errores estandar, en relación a la Tabla 3.3, por el hecho de introducir en el modelo una variable que da lugar a un alto grado de colinealidad. Definida y vistas sus consecuencias, una primera cuestión a resolver es la detección de la posible colinealidad. Una estrategia empleada frecuentemente, pero no por ello menos errónea, es calcular el coeficiente de correlación, de Pearson o de Spearman, entre todas las parejas de predictoras; evidentemente, así no se puede detectar correlaciones entre más de dos variables. El uso del coeficiente de determinación, R^2, resuelve parcialmente este problema; dada la predictora X_j, su coeficiente de terminación correspondiente R_j^2 mide el parte de variabilidad de X que puede ser explicada por el resto de las predictoras; como consecuencia de esta definición, $0 \leq R_j^2 \leq 1$, así que valores del coeficiente de determinación próximos a la unidad indicarán existencia de colinealidad; sin embargo, Belsley (1980) demuestran que este coeficiente también puede fallar a la hora de detectar algunas colinealidades.

3.11. Sobredispersión

Con este término se nombra la situación en que la varianza de la variable respuesta Y es mayor que la varianza nominal $np(1\text{-}p)$; el no tener en cuenta este hecho puede llevar a una subestimación de la variabilidad de los estimadores del modelo, lo que llevaría a declarar como significativas asociaciones que realmente no lo son.

La presencia de sobredispersión puede deberse tanto a efectos del diseño empleado en el estudio como, en general, a la ausencia de predictores importantes en el modelo ajustado; una situación en la que cabe esperar la presencia de este fenómeno es en los estudios en donde se utiliza el muestreo por conglomerados (*cluster*). En efecto, supongamos que se quiere estimar la prevalencia de una enfermedad en una población determinada y para ello tomamos una muestra de n individuos; si defini-

mos para cada individuo de la muestra una variable binaria Y_i tal que $y_i=1$ si el individuo i presenta la enfermedad e $y_i=0$ en caso contrario, sabemos que la prevalencia p de la enfermedad en la población se puede estimar mediante el cociente entre el número de enfermos y el número de individuos en la muestra, es decir,

$$\hat{p} = \frac{\sum_{i=1}^{n} y_i}{n}$$

y que la varianza de este estimador viene dada por

$$\sigma^2(\hat{p}) = \frac{p\ (1-p)}{n}$$

Consideremos ahora que la población consiste en S conglomerados y, por razones de sencillez en la exposición, el tamaño de muestra n' elegido en cada uno es el mismo, es decir, $n'=n/S$. Supongamos que la prevalencia de enfermedad p_j en el conglomerado j es la misma para todos, es decir, $p_1=p_2=\ldots=p_s=p$; si notamos por y_{ij} el valor correspondiente al individuo i del conglomerado j, la prevalencia de enfermedad en el conglomerado j viene estimada por el cociente

$$\hat{p}_j = \frac{\sum_{i=1}^{n} y_{ij}}{n'}$$

con una varianza

$$\sigma^2(\hat{p}_j) = \frac{p\ (1-p)}{n'}$$

en caso de independencia entre las observaciones dentro de los conglomerados.

Sin embargo esta condición de independencia es más la excepción que la regla en la vida real; piénsese, por ejemplo, en un estudio para estimar prevalencia de una enfermedad infecciosa y que se realiza un muestreo por familias, los conglomerados. Si las observaciones dentro de cada estrato no son independientes, y notamos por ρ_j la correlación entre las respuestas en el conglomerado j, se puede demostrar que la varianza dentro de cada estrato es ahora

$$\sigma^2(\hat{p}_j) = \frac{p\ (1-p)}{n'}\ [1 + (n'-1)\rho_j\ ]$$

y que la varianza para toda la muestra es

$$\sigma^2(\hat{p}) = \frac{p\ (1-p)}{n'}\ [1 + (n'-1)\overline{\rho}_j\ ]$$

donde

$$\hat{p} = \frac{\sum_{i=1}^{n} y_i}{n}$$

En definitiva, cuando existe dependencia entre los individuos de los conglomerados, la varianza está afectada por la cantidad

$$[1+(n'-1)\overline{\rho}]$$

Si la correlacion entre los valores de la respuesta es positiva, tal cantidad es mayor que la unidad y por tanto la varianza real es mayor que la nominal. Obsérvese como el efecto de la sobredispersión es tanto mayor cuanto mayor sea el tamaño muestral n' de cada estrato; evidentemente, sólo es posible la presencia de sobredispersión cuando $n>1$. Un tratamiento exhaustivo de este problema en regresión logística puede verse en Collett (1991).

Conceptualmente, una situación similar es la de los estudios de seguimiento, en donde se mide la variable respuesta varias veces a lo largo del tiempo; en este caso, los sujetos seguidos juegan el papel de los conglomerados y las distintas mediciones serían los individuos. Esto ha posibilitado que la investigación epidemiológica haya sido un estímulo para el desarrollo de nuevos modelos que puedan incorporar la correlación intra-conglomerado, Pendergast (1996) y Ashby (1992) son excelentes referencias.

Como ejemplo de aplicación de uno de estos métodos, Gange (1996) utiliza la distribución beta-binomial para estudiar la inadecuación de la estancia hospitalaria. Para cada paciente (conglomerado) de la muestra, midieron la inadecuación de cada día de su estancia hospitalaria. El objetivo principal del estudio era estudiar la frecuencia de inadecuación entre el año 1988 y 1990 fechas entre las que hubo una intervención en la gestión del hospital; pero ya que los autores disponían de varias observaciones para un mismo individuo, también evalúan la correlación dentro de cada paciente, que interpretan como propagación de la inadecuación en los distintos dias de estancia, propagación que, haciéndola depender de una variable indicadora, antes o después de la intervención, permite evaluar el efecto de la intervención sobre la propagación de la inadecuación.

CAPÍTULO IV

REGRESIÓN LOGÍSTICA POLICOTÓMICA Y ORDINAL

En este capítulo vamos a extender la definición del modelo logístico binario al caso en que la variable resultado tenga más de dos categorías; así trataremos con el modelo logístico policotómico. Como quiera que en ocasiones hay un orden entre las categorías de la respuesta, también estudiaremos modelos que contemplan esta circunstancia.

4.1. Introducción

El modelo de regresión logística que se ha expuesto hasta ahora es, con mucho, el método más utilizado para analizar relaciones entre una variable resultado binaria y un conjunto de predictoras. Sin embargo, con frecuencia se presenta la situación de tener una variable resultado con más de dos categorías. Dubin y Pasternack (1986) presentan un estudio de casos y controles sobre cáncer de mama en el que los casos son de dos tipos: los que denominan cáncer mínimo (cáncer in situ o invasivo de menos de 1 cm.) y los que llaman cánceres clínicos (los restantes); en esta situación la variable respuesta tiene tres categorías: cáncer clínico, cáncer mínimo y control. Frank y Kamlet (1989) presentan un ejemplo del campo de la salud mental; en base a una encuesta de 2800 entrevistas registraron, como variable respuesta, el tipo de atención psiquiátrica demandada por los individuos y categorizada así: 1) no busca ayuda, 2) la busca de un no profesional de la medicina (familia, amigos, religiosos, etc.), 3) de un médico general y 4) de un especialista en salud mental. Como predictoras consideran características del que presta el servicio así de como de la salud general y mental del individuo posible demandante de atención psiquiátrica; el interés está en estudiar qué variables predictoras están asociadas a la elección de la ayuda solicitada y cómo es esa asociación.

Sin embargo, no todas las variables categóricas son de las misma naturaleza; se pueden diferenciar dos grupos dependiendo de que entre sus categorías haya o no algún tipo de orden o jerarquía, distinguiendo así las respuestas *ordinales* de las *nominales*. Esta distinción no es baladí desde el punto de vista del análisis estadístico, como se verá más adelante.

Para las variables nominales, la estrategia de análisis puede ser parecida a la de la regresión logística binaria; en ésta, se toma una de las categorías de la respuesta como la de referencia y se enfrenta a la otra. Para una variable nominal con c categorías elegiremos una como referencia y la compararemos con las c-1 categorías restantes.

Es más frecuente el caso del clínico o del epidemiólogo que se encuentra con variables respuesta medidas en una escala ordinal. A veces estas variables pueden proceder de la categorización, por criterios prácticos, de una variable cuantitativa; por ejemplo, el índice de masa corporal, aunque es una variable que se mide en escala continua, kg./m^2, en la práctica se suele categorizar y se habla de individuos delgados, normales, obesos y muy obesos. En otras ocasiones, las variables ordinales proceden de una variable cuantitativa subyacente no cuantificable; características de tanto interés como la salud percibida, la satisfacción con la atención recibida, el dolor o la capacidad funcional son de este tipo. Sea cual sea el origen de la variable ordinal, lo sustantivo es la gradación entre las categorías.

En principio, ante respuestas ordinales se pueden adoptar distintas estrategias de análisis, ninguna de ellas exenta de dificultades. Una primera alternativa puede ser no considerar el caracter ordinal y proceder como en el caso de las variables nominales. Una segunda solución puede ser asignar valores numéricos a las categorías y utilizar el modelo de regresión lineal; podríamos asignar 0 a las categoría de los delgados, 1 a los normales, 2 a los obesos y 3 a los muy obesos; sin embargo, es dificil defender esta asignación numérica y no otra distinta, con el grave inconveniente de que distintas asignaciones pueden dar lugar a resultados diferentes. Como tercera alternativa podríamos dicotomizar la respuesta considerando un grupo constituido por los delgados y los normales y el otro grupo formado por los obesos y muy obesos y utilizar el modelo de regresión logística binaria; esta estrategia no es necesariamente incorrecta pero se recomienda utilizarla con mucha precaución, Strömberg (1996).

Estas dificultades han llevado en los últimos años al desarrollo de distintos modelos que incorporan en su definición el caracter ordinal de la respuesta; entre otros, dos propuestos por McCullagh (1989) que son los más utilizados, quizás por la mayor disponibilidad de programas informáticos más o menos estándar para su estimación, y otro más general, Anderson (1984).

4.2. La distribución multinomial

En el Apartado 1.3 se vió que la distribución de probabilidad asociada a la regresión logística binaria es la binomial; ello significa que en cada individuo medimos la variable resultado Y que es dicotómica con valores Y=1 si éste presenta la característica de interés e Y=0 en caso contrario. Si p es la probabilidad de que un individuo cualquiera presente la característica, es decir, $P(Y=1)=p$, todo individuo puede pertenecer a una de los dos categorías de la respuesta con probabilidades p y 1-p. Según vimos, elegidos n individuos, la probabilidad de que entre ellos haya exactamente r que tengan la característica de interés viene dada por la expresión

$$P(Y=r) = \binom{n}{r} p^r (1-p)^{n-r}$$

Vamos a generalizar esta situación considerando una respuesta con más de dos categorías. Sea ahora la variable respuesta Y con c categorías distintas 1, 2,..., c y con probabilidades asociadas $\pi_1, \pi_2, \ldots, \pi_c$, respectivamente; si tomamos una muestra aleatoria de n individuos, la probabilidad de que haya r_1 de ellos en la categoría 1, r_2 en la categoría 2 y así, hasta r_c en la categoría c, viene dada por la función

$$\frac{n!}{r_1!\, r_2! \ldots r_c!} \pi_1^{r_1} \pi_2^{r_2} \ldots \pi_c^{r_c} = \frac{n!}{\prod_{i=1}^{c} r_i!} \prod_{i=1}^{c} \pi_i^{r_i}$$

que es la llamada *función de probabilidad multinomial*. Ya que las categorías de la respuesta deben ser exhaustivas y excluyentes, se debe cumplir que la suma de los individuos en cada categoría debe ser el número total de individuos, es decir, $r_1 + r_2 + \ldots + r_c = n$ y, por otra parte, $\pi_1 + \pi_2 + \ldots + \pi_c = 1$; veamos a continuación un ejemplo concreto.

La salud percibida es un indicador de mucho interés para los investigadores sanitarios; consideremos una determinada población de mayores de 65 años en la que el 25% de ellos tienen sensación de mala salud, el 50% piensa que su salud es regular y el 20% y 5% creen que su salud es buena y muy buena, respectivamente; es decir, π_1=0,25, π_2=0,5, π_3=0,2 y π_4=0,05. Si tomásemos una muestra aleatoria de 100 viejos de esa población, ¿cuál será la probabilidad de que entre los 100 haya 30 viejos con mala percepción de su salud, 47 con salud regular, 17 buena y 6 con muy buena salud percibida? En este caso n=100, r_1=30, r_2=47, r_3=17 y r_4=6, por lo que la probabilidad pedida será, según la expresión anterior,

$$\frac{100!}{30!\ 47!\ 17!\ 6!}\ 0{,}25^{30}\ 0{,}5^{47}\ 0{,}2^{17}\ 0{,}05^{6} = 0{,}0007$$

Lo mismo se puede conseguir para cada posible muestra elegida.

En caso de que el número de categorías de la respuesta sea $c=2$, tendremos dos probabilidades π_1, π_2 tales $\pi_1+\pi_2 = 1$; es decir, si notamos por p a la probabilidad de pertenecer a la primera categoría, es decir, $p=\pi_1$, entonces $\pi_2 = 1-p$; por otra parte, como $r_2 = n-r_1$, la función de probabilidad anterior se puede escribir así

$$\frac{n!}{r_1!\,(n-r_1)!}\, p^{r_1}\, (1-p)^{n-r_1}$$

que no es otra cosa que la función de probabilidad de la distribución binomial; así, una binomial es un caso particular de la multinomial para el caso de solo dos categorías.

4.3. Modelos de regresión logística para variables nominales

El objetivo de este apartado es extender el modelo de regresión logística binaria al caso de una respuesta con más de 2 categorías. En el modelo binario, si Y=1 representa el poseer la característica de interés (enfermedad) e Y=0 indica la categoría de referencia (control), la probabilidad de presentar la característica venía dada por

$$P(Y=1) = \frac{e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)}}{1 + e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)}}$$

que también escribíamos así,

$$logit[P(Y=1)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

En regresión logística binaria, se tomó una categoría, la Y=0, como referencia, y se definió la ventaja, $P(Y=1)/P(Y=0)$, de la otra categoría de la respuesta, Y=1, respecto de ella; en el caso policotómico, como hay c categorías, tomamdo como referencia a una de ellas, por ejemplo a la categoría 1, podemos definir c-1 ventajas del tipo

$$\frac{P(Y=s)}{P(Y=1)} \qquad\qquad s = 2, 3, \ldots, c$$

y, por tanto, los c-1 logits correspondientes

Generalizando la situación anterior, sea Y una variable respuesta nominal con c categorías con valores Y=1, Y=2,..., Y=c, y sean X_1, X_2, ... ,X_m un conjunto de m predictoras. El *modelo de regresión logística policotómica* establece que

$$logit[P(Y=s)] = \log \frac{P(Y=s)}{P(Y=1)} = \beta_{0s} + \beta_{1s} X_1 + \beta_{2s} X_2 + \ldots + \beta_{ms} X_m$$

Obsérvese que esta definición quiere decir que para cada categoría s distinta de la de referencia tenemos un conjunto de parámetros β_{0s}, β_{1s},..., β_{ms}. Por tanto, el modelo policotómico realmente contempla un conjunto de c-1 ecuaciones de regresión y como para cada una de ellas tenemos m+1 parámetros, en el modelo logístico están implicados (c-1)(m+1) parámetros. La definición del modelo policotómico conlleva por tanto que para cada predictora X_i tengamos que estimar c-1 efectos (coeficientes).

Otra manera de formularlo es mediante la expresión

$$\pi_s = P(Y=s) \frac{e^{(\beta_{0s} + \beta_{1s} X_1 + \beta_{2s} X_2 + \ldots + \beta_{ms} X_m)}}{1 + \sum_{s=2}^{c} e^{(\beta_{0s} + \beta_{1s} X_1 + \beta_{2s} X_2 + \ldots + \beta_{ms} X_m)}}$$

para cualquier categoría s distinta de la de referencia, y

$$\pi_1 = P(Y=1) \frac{1}{1 + \sum_{s=2}^{c} e^{(\beta_{0s} + \beta_{1s}X_1 + \beta_{2s}X_2 + \ldots + \beta_{ms}X_m)}}$$

para ésta.

4.3.1 Interpretación de los coeficientes

La interpretación de los coeficientes del modelo de regresión logística policotómica también guarda paralelismo con el caso binario; consideremos, por ejemplo, la variable predictora dicotómica X_i tal que $X_i = 1$ para los expuestos a un cierto factor de riesgo y $X_i = 0$ para los no expuestos. Si A es un individuo expuesto y B uno no expuesto en relación a la variable X_i pero iguales valores en relación a las restantes predictoras, el logit de la probabilidad de que A pertenezca a la categoría s es

$$\log \frac{P_A(Y=s)}{P_A(Y=1)} = \beta_{0s} + \beta_{1s}X_1 + \ldots + \beta_{is} \cdot 1 + \ldots + \beta_{ps}X_p$$

y, para el B

$$\log \frac{P_B(Y=s)}{P_B(Y=1)} = \beta_{0s} + \beta_{1s}X_1 + \ldots + \beta_{is} \cdot 0 + \ldots + \beta_{ps}X_p$$

Restando esas dos igualdades se llega a esta expresión

$$\log \frac{P_A(Y=s)/P_A(Y=1)}{P_B(Y=s)/P_B(Y=1)} = \beta_{is}(1-0) = \beta_{is}$$

en donde tenemos que el logaritmo de la razón de ventajas de estar en la categoría s respecto a la categoría de referencia, del individuo A expuesto al factor representado por X_i respecto a uno B no expuesto, viene dado precisamente por el coeficiente de tal variable; es decir,

$$\psi_{is} = e^{\beta_{is}}$$

donde ψ_{is} es la razón de ventajas de estar en la categoría s respecto a la de referencia, entre esos dos individuos que se diferencian en una unidad de la variable X_i. Es evidente que la notación se complica algo pues ahora para una misma variable tenemos c-1 de estas *OR*, y en el modelo binario teníamos solo una, pues entonces c=2.

Por un argumento similar, si X_i es una variable cuantitativa con valores x_{iA} , x_{iB} en esos dos individuos, siendo iguales respecto a las restantes variables, la razón de ventajas de pertenecer a la categoría s respecto a la de referencia del individuo A respecto del B vendrá dada por la expresión

$$\psi_{is} = e^{\beta_{is}(x_{iA} - x_{iB})}$$

Por último, si A y B difieren en más de una predictora, la razón de ventajas vendrá dada, de manera análoga al modelo binario, por

$$e^{\sum_i \beta_{is}(x_{iA}-x_{iB})} = \prod_i e^{\beta_{is}(x_{iA}-x_{iB})}$$

donde la suma y el producto se extienden a todos los subíndices correspondientes a las variables en que difieren A y B; esta expresión pone también en evidencia el caracter multiplicativo de este modelo pues el riesgo respecto a varias variables se consigue multiplicando los riesgos individuales para cada variable, evidentemente, en el caso en que no exista interacción entre las predictoras.

En regresión logística binaria un coeficiente $\beta_i > 0$ indicaba un aumento en el logit con el aumento de la predictora x_i, lo que traía consigo un aumento de la probabilidad de que la respuesta tomase el valor 1, es decir, el individuo presente la característica de interés. Sin embargo, en regresión policotómica esto no es así necesariamente; en efecto, si $\beta_{is}>0$, un aumento en x_i implica un aumento en el logit

$$\log \frac{P(Y=s)}{P(Y=1)}$$

pero ahora esto no conlleva necesariamente el aumento de $P(Y=s)$, pues podría ocurrir que tal probabilidad fuese parecida y lo que ocurriera es que disminuyese $P(Y=1)$; esto no puede ocurrir en el modelo binario pues, al ser solo dos categorías, si la probabilidad de una disminuye, la de la otra debe necesariamente aumentar.

El modelo policotómico, además de proporcionar los riesgos de cada categoría respecto a la de referencia, también permite comparar dos categorías cualesquiera; en efecto, de manera análoga que para la categoría s, para la categoría q

$$\log \frac{P(Y=q)}{P(Y=1)} = \beta_{0q} + \beta_{1q}X_1 + \beta_{2q}X_2 + \ldots + \beta_{mq}X_m$$

Restando esas dos igualdades se obtiene

$$\log \frac{P(Y=s)}{P(Y=q)} = (\beta_{0s}-\beta_{0q}) + (\beta_{1s}-\beta_{1q})X_1 + \ldots + (\beta_{ms}-\beta_{mq})X_m$$

por lo que las diferencias $\beta_{is}-\beta_{iq}$ se pueden interpretar como el logaritmo de la razón de ventajas de estar en la categoría s respecto de la q entre un expuesto al factor X_i y un no expuesto, es decir,

$$e^{\beta_{is}-\beta_{iq}}$$

es la *OR* correspondiente.

Una ventaja de este modelo sobre los que más adelante se discutirán es que es aplicable tanto a estudios transversales, de cohortes o de casos y controles.

4.3.2 Estimación de los coeficientes

El método de estimación de los coeficientes sigue siendo el método de máxima verosimilitud. Como ya ha quedado dicho, en esta ocasión hay que estimar $(c\text{-}1)(m+1)$ parámetros β_{is} con $s=2, 3, ..., c$, e $i=0, 1, ..., m$. Sean n_i el número de individuos con un patrón de valores de las predictoras

$$x_{1i}\ ,\ x_{2i}\ ,\ldots,\ x_{mi}$$

y sean

$$n_{i1}\ ,\ n_{i2}\ ,\ldots,\ n_{ic}$$

el número de ellos que pertenecen a las categorías 1, 2, ..., c, respectivamente. Si la probabilidades asociadas a esas categorías las notamos por π_{is}, la probabilidad de lo ocurrido es, según se dijo antes,

$$\frac{n!}{\prod_{s=1}^{c} n_{is}!}\prod_{s=1}^{c}\pi_{is}^{n_{is}}$$

Considerando los k patrones distintos de las predictoras, la probabilidad de lo ocurrido será

$$\prod_{s=1}^{k}\frac{n!}{\prod_{s=1}^{c} n_{is}!}\prod_{s=1}^{c}\pi_{is}^{n_{is}}$$

que considerada como función de los parámetros β_{is} del modelo es, salvo una constante, la función de verosimilitud; en concreto, eliminando la parte de esta función que no depende de los parámetros a estimar y tomamdo logaritmos obtenemos la función

$$\sum_{i=1}^{k}\sum_{s=1}^{c} n_{is}\log(\pi_{is})$$

que, considerada como función de los parámetros del modelo, es la función de verosimilitud. Pues bien, como para el caso de la regresión binaria, las estimaciones de los parámetros del modelo son los valores de éstos que hacen máxima tal función; para este caso, el proceso de estimación da lugar a un sistema de ecuaciones no lineales con $(c\text{-}1)(m+1)$ incógnitas, el número de parámetros a estimar, cuya resolución se consigue mediante procedimientos iterativos. El concepto de lejanía es análogo al caso binario y la estrategia de comparación de modelos descrita en el apartado 1.12 es también vigente ahora.

4.3.3 Ejemplo con una sola predictora

Con el objetivo de fijar ideas vamos a tratar con un caso muy sencillo; los datos que a continuación se muestran que provienen de una encuesta realizada en 1986 por la Escuela Andaluza de Salud Pública a la población de viejos, personas de 65 o más años de edad, del distrito sanitario de Guadix (Granada); entre otras muchas preguntas, los entrevistados debían responder acerca de la precepción de su propia salud, variable que se registró como mala, regular, buena o muy buena; la distribución de la salud percibida según la edad de los viejos aparece en la Tabla 4.1.

Tabla 4.1. *Distribución de la salud percibida, según edad.*

		SP			
		Mala	Regular	Buena	Muy buena
EDAD	65-74	82	123	91	21
	> 74	51	103	47	6

La variable respuesta, que por ahora vamos a considerar nominal, va a ser la salud percibida *(SP)* que tiene $c = 4$ categorías que las vamos a codificar así: 1 = mala, 2 = regular, 3 = buena y 4 = muy buena; como variable predictora consideramos la edad *(EDAD)* categorizada como 0 en el grupo de 65-74 años y como 1 en el de más de 74 años. El modelo policotómico antes propuesto para el caso de una sola predictora es

$$\log \frac{P(Y=s)}{P(Y=1)} = \beta_{0s} + \beta_{1s} X_1$$

que para este ejemplo concreto será

$$\log \frac{P(Y=s)}{P(Y=1)} = \beta_{0s} + \beta_{1s} EDAD$$

con $s=2, 3, 4$, donde se ha elegido como categoría de referencia la salud percibida "mala", $SP=1$. ¿Cómo interpretar los tres coeficientes β_{1s} del modelo? Si consideramos la categoría salud percibida "muy buena" ($SP=4$), el logit de esta categoría respecto de la "mala" para los individuos con una edad superior a los 74 años, $EDAD=1$, es

$$\log \frac{P\ (muy\ buena)}{P\ (mala)} = \beta_{04} + \beta_{14} \cdot 1$$

y para los individuos de entre 65 a 74 años, $EDAD=0$, ese mismo logit es $\beta_{04}+\beta_{14}\cdot 0$. Por tanto, la diferencia de estos dos logits, o lo que es igual, el logaritmo de la razón de ventajas del grupo de los mayores respecto de los más jóvenes será

$$(\beta_{04} + \beta_{14} \cdot 1) - (\beta_{04} + \beta_{14} \cdot 0) = \beta_{14}$$

es decir, $e^{\beta_{14}}$ es la razón de ventajas de los de más de 74 años respecto al grupo de edad de 65-74, de percibir su salud como "muy buena" en relación a percibirla como "mala". Igual interpretación tienen los coeficientes $e^{\beta_{13}}$ y $e^{\beta_{12}}$ para las categorías "buena" y "regular", respectivamente.

El modelo ajustado con la edad como predictora es el que aparece en la Tabla 4.2 que tiene una lejanía asociada de 0 con 0 grados de libertad

Tabla 4.2. *Estimaciones y errores estándar de los coeficientes de la edad.*

	Regular/Mala	Buena/Mala	Muy buena/Mala
Constante	0,406 (0,143)	0,104 (0,152)	-1,363 (0,245)
EDAD	0,298 (0,223)	-0,186 (0,253)	-0,778 (0,496)

Para evaluar la dependencia entre la salud percibida y la edad en los viejos, lo haremos ajustando el modelo que no contiene a ninguna predictora, que tiene una lejanía de 8,9 con 3 grados de libertad, y el que se acaba de ajustar; por tanto tenemos que comparar la diferencia 8,9-0=8,9 con una chi-cuadrado con 3-0=3 grados de libertad; ya que valores superiores a 8,9 ocurren tan solo en un, aproximadamente, 3% de la ocasiones en una chi-cuadrado con 3 grados de libertad, tenemos argumentos para rechazar la hipótesis de que los dos modelos comparados son equivalentes. En definitiva, la edad está asociada al grado de percepción de salud de los viejos.

Según la Tabla 4.2, las estimaciones de β_{12}, β_{13}, β_{14} son 0,298, -0,186 y -0,778 respectivamente. A partir de ella se pueden obtener los tres modelos logísticos correspondientes a la comparación de cada categoría con la de referencia; estos modelos estimados son

$$\log \frac{P\ (regular)}{P\ (mala)} = 0{,}406 + 0{,}298 EDAD$$

$$\log \frac{P\ (buena)}{P\ (mala)} = 0{,}104 - 0{,}186 EDAD$$

$$\log \frac{P(muy\ buena)}{P\ (mala)} = -1{,}363 - 0{,}778 EDAD$$

En el Apartado 4.3.1 acaba de verse que elevando el número e a estas estimaciones podíamos conseguir las razones de ventajas del grupo de $EDAD=1$, los mayores de 74 años, respecto del grupo más joven, $EDAD=0$. En efecto, según la Tabla 4.1, la razón de ventajas de percibir la salud como "muy buena" ($SP=4$) en contra de "muy mala" para los mayores de 74 años respecto a los que tienen entre 65 y 74 años es

$$(6)(82) / (21)(51) = 0,459$$

valor que coincide con $e^{-0,778}$. Para la clase "buena" la razón de ventajas es

$$(47)(82) / (91)(51) = 0,830$$

que no es más que $e^{-0,186}$ y, por último, para la clase "regular" la razón de ventajas es

$$(103)(82) / (123)(51) = 1,346$$

que vuelve a coincidir, como era de esperar, con $e^{0,298}$.

En una primera aproximación, como interpretación del papel de la edad en la percepción del estado de salud de los viejos podemos decir que los más viejos tienden a percibir su salud peor que los viejos más jóvenes, pues salvo para la categoría regular, las razones de ventajas que comparan los mayores con los más jóvenes, son menores que la unidad.

Antes se comentó que, aparte de estimar las razones de ventajas de cada clase respecto a la de referencia, también se pueden comparar cualquier pareja de clases de la variable respuesta. Consideremos las categorías "buena" y "regular" cuyos coeficientes asociados son -0,186 y 0,298, respectivamente. Según se dijo, el logaritmo de la razón de ventajas de la clase buena respecto a la regular se puede obtener mediante la diferencia -0,186-0,298=-0,484, por lo que $e^{-0,484}$=0,616 es la razón de ventajas entre esas dos clases; en efecto, si se considera la clase "regular" como referencia respecto de la "buena" da origen a la Tabla 4.3

Tabla 4.3. *Datos para las categorías buena y regular.*

		SP	
		Regular	Buena
EDAD	65-74	123	91
	> 74	103	47

y, como era de esperar, el producto cruzado (123)(47)/(91)(103)=0,617 coincide con la anterior estimación, salvo redondeo.

La Tabla 4.4 muestra la matriz de varianzas-covarianzas de los parámetros estimados, donde la diagonal de esta matriz son las varianzas, cuadrado de los errores estandar, de las estimaciones del modelo; 1, 2 y 3 se refieren a las tres constantes y 4, 5 y 6 a los tres coeficientes de la edad en el modelo estimado de la Tabla 4.2.

Tabla 4.4. *Varianzas y covarianzas de las estimaciones de los parámetros del modelo que contiene a la edad.*

1	0,0203					
2	0,0122	0,0232				
3	0,0122	0,0122	0,0598			
4	-0,0122	-0,0122	-0,0318	0,0496		
5	-0,0232	-0,0122	-0,0318	0,0318	0,0641	
6	-0,0122	-0,0598	-0,0318	0,0318	0,0318	0,2461
	1	2	3	4	5	6

Esta matriz nos permite hacer contrastes y calcular intervalos de confianza para cualquier pareja de categorías de la respuesta; en efecto, según se dijo en el apartado 4.3.1, β_{is}-β_{iq} era el logaritmo de la razón de ventajas de estar en la categoría s respecto de la q, de un individuo que supere a otro en una unidad en relación a la predictora X_i. Por tanto, entre la categoría "buena" y "regular" tal logaritmo de la *OR* será (-0,186)-0,298=-0,484, por lo que la razón de ventajas es 0,616. Para calcular un intervalo de confianza para esa *OR* necesitamos conocer el error estándar de la diferencia de las estimaciones; según la Tabla 4.4, las varianzas son 0,0641 y 0,0496 y su covarianza 0,0318 por lo que la varianza de la diferencia será 0,0641+0,0496-2(0,0318)=0,0501 y la raíz cuadrada será su error estandar, 0,224. Así, un intervalo al 95% para el logaritmo de la *OR* será

$$-0{,}186-0{,}298\pm 1{,}96(0{,}224)=(-0{,}923\ ,\ -0{,}045)$$

por lo que el intervalo buscado será (0,40 , 0,96). El lector puede comprobar que este mismo intervalo es el que se deriva para la *OR* de la Tabla 4.3 según el método de Woolf visto en el Apartado 1.10.

Con el modelo logístico multinomial se obtienen las estimaciones de los coeficientes maximizando la verosimilitud y considerando simultáneamente c-1 ecuaciones, siendo c el número de categorías de la variable respuesta. Otra posible alternativa sería ajustar c-1 modelos logísticos binarios de cada categoría respecto a la de referencia. Begg (1984) estudió esta solución alternativa y aunque comprobó que los estimadores eran menos eficientes, tal ineficiencia disminuye con la prevalencia de la categoría de referencia, por lo que si no nos vemos forzados a la elección de tal categoría, es deseable tomar como referencia la que sea más prevalente. Así, ajustando un modelo binario para la categoría "muy buena" respecto de la "mala" da lugar a un coeficiente asociado a la *EDAD* de -0,778 con un error estándar de 0,496, los mismos resultados que en modelo policotómico.

Aparte del inconveniente de la ineficiencia, esta forma de atacar el problema no permite comparar dos categorías cualesquiera ya que no podemos conocer la cova-

rianza entre los estimadores de distintas categorías por el hecho de ajustarse de forma independiente.

4.3.4 Ejemplo con varias predictoras

Los datos que aparecen en la Tabla 4.5 son de la misma encuesta pero donde, además de la edad, también se muestran otras variables, a saber: *SEXO* codificada como 0=Hombre, 1=Mujer y otra variable *ACOM* que tiene que ver con el sentimiento de soledad, codificada como *ACOM*=0 si la persona se siente sola muy frecuentemente, 1 si eso le ocurre algunas veces y 2 si no se siente solo. Como quiera que esta última variable es categórica con 3 categorías, serán necesarias dos variables indicadoras para poner utilizarla como predictora de la salud percibida; las dos variables indicadoras *ACOM(1)* y *ACOM(2)* las definimos así:

ACOM	*ACOM(1)*	*ACOM(2)*
0	0	0
1	1	0
2	0	1

lo que equivale a tomar a los que se siente solos casi siempre como categoría de referencia; el coeficiente de *ACOM(1)* medirá el efecto de sentirse solo algunas veces respecto de sentirse casi siempre.

Acabamos de estudiar la relación entre la salud percibida y la edad de los viejos y existe sospechas de que ambas variables están asociadas; ¿no podría ocurrir que esta asociación estuviese mediatizada por el género o por el grado de compañía? Se trata ahora de estudiar la relación entre la salud percibida y esas tres predictoras.

Tabla 4.5. *Distribución de la salud percibida según grado de compañía, edad y género.*

			Salud percibida			
ACOM	*EDAD*	*SEXO*	1(mala)	2(regular)	3(buena)	4(muy buena)
0	0	0	7	7	4	1
		1	22	19	9	0
	1	0	8	7	1	1
		1	14	14	3	1
1	0	0	4	9	7	2
		1	9	13	9	2
	1	0	3	8	6	0
		1	10	11	4	0
2	0	0	16	43	34	7
		1	24	32	28	9
	1	0	4	27	17	2
		1	12	36	16	2

Con los datos de la Tabla 4.5 se han ajustado distintos modelos cuyas predictoras, lejanías y grados de libertad asociados se muestran en la Tabla 4.6.

Tabla 4.6. *Distintos modelos ajustados junto con sus lejanías y grados de libertad.*

Predictoras	Lejanía	g.l.
Nulo	65,05	33
SEXO	54,46	30
EDAD	56,15	30
ACOM	32,25	27
SEXO, EDAD	45,65	27
SEXO, ACOM	25,95	24
EDAD, ACOM	23,58	24
SEXO, EDAD, ACOM	17,41	21

Esta tabla nos permite comparar distintos modelos y, por tanto, evaluar los efectos de cada predictora; por ejemplo, el modelo ajustado con las tres predictoras tiene una lejanía asociada de 17,41 con 21 grados de libertad y para el modelo que solo tiene en cuenta la edad y el grado de compañía la lejanía correspondiente es 23,58 con 24 grados de libertad; esto quiere decir que 23,58-17,41=6,17, la diferencia en lejanías, es una medida del ajuste perdido por no considerar el género. Ya que la diferencia en grados de libertad es 24-21=3, con un error aproximado de 0,1, podemos decir que la variable *SEXO* aporta información después de la que aportan las otras dos variables; el lector puede comprobar como el modelo que contiene a las tres variables explica las cosas mejor que cualquier otro y como ninguna de las interacciones entre las predictoras es significativa, este será el modelo elegido.

Que el modelo con las tres predictoras sea la propuesta para explicar los datos no significa necesariamente que su ajuste sea satisfactorio. Para ello lo comparamos con el modelo saturado que sabemos que tiene tanto la lejanía como los grados de libertad igual a 0; por tanto, entre nuestro modelo elegido y el saturado hay una diferencia de lejanía de 17,41-0=17,41 y de 21-0=21 en grados de libertad; ya que 17,41 no es en absoluto un valor extraño para la χ^2 con 21 grados de libertad ello significa que entre nuestro modelo y el saturado no hay grandes diferencias, por lo que el modelo con las tres predictoras explica las cosas suficientemente bien.

Tabla 4.7. *Estimaciones de los parámetros junto a sus errores estándar. Debajo aparece la estimación mediante el método de Begg.*

Constante		SEXO	EDAD	ACOM(1)	ACOM(2)
log P(regular)	0,104 (0,279)	-0,464 (0,235)	0,332 (0,228)	0,495 (0,325)	0,927 (0,260)
/P(mala)	0,091 (0,279)	-0,471 (0,236)	0,360 (0,230)	0,540 (0,325)	0,933 (0,261)
log P(buena)	-0,624 (0,345)	-0,611 (0,260)	-0,134 (0,262)	1,054 (0,397)	1,530 (0,330)
/P(mala)	-0,617 (0,345)	-0,644 (0,265)	-0,126 (0,270)	1,070 (0,399)	1,542 (0,331)
log P(muy buena)	-2,261 (0,675)	-0,490 (0,437)	-0,721 (0,501)	0,936 (0,804)	1,707 (0,654)
/P(mala)	-2,239 (0,670)	-0,616 (0,445)	-0,522 (0,517)	0,963 (0,806)	1,691 (0,656)

De la Tabla 4.7 se puede deducir que, a igualdad de edad y nivel de compañía, las mujeres tienen una peor percepción de su salud que los hombres pues los riesgos de percibir la salud como regular, buena o muy buena respecto a percibirla como mala, de las mujeres respecto de los hombres, son

$$e^{-0,464}=0,63 \quad , \quad e^{-0,611}=0,54 \quad , \quad e^{-0,490}=0,61$$

es decir, los riesgos de percibir la salud como regular, buena o muy buena, respecto de percibirla como mala son 1,59, 1,85 y 1,64 veces superior, respectivamente, en los hombres que en las mujeres. En cuanto a la edad, los mayores de 74 años perciben su salud peor que los de edades comprendidas entre 65 y 74 años, pues tienen como riesgos de 1,39, 0,87 y 0,49 correspondientes a las categorías regular, buena o muy buena de salud percibida. El hecho de no haber encontrado interacciones significativas implica que, por ejemplo, la diferencia de percepción entre los hombres y las mujeres no depende de la edad que tengan ni de lo más o menos acompañados que se sientan.

La indicadora *ACOM(1)* tiene coeficientes estimados todos positivos y la variable *ACOM(2)* también y, además, mayores que los anteriores; ello nos indica una mejor percepción de la salud cuanto mayor es el grado de acompañamiento; por ejemplo, a igualdad de sexo y edad, las *OR* de un individuo que se siente algunas veces solo respecto a otro que se siente solo casi siempre son

$$e^{0,495}=1,64 \quad , \quad e^{1,054}=2,87 \quad , \quad e^{0,936}=2,55$$

y para el que no se siente solo

$$e^{0,927}=2,53 \quad , \quad e^{1,53}=4,62 \quad , \quad e^{1,707}=5,51$$

El modelo de la Tabla 4.7 se podría simplificar si consideramos que, a igualdad de sexo y edad, los logits de los que se sienten solos a veces respecto a los que se

sienten solo casi siempre, 0,495, 1,054 y 0,936 son, aproximadamente, los mismos que los logits de los que no se sienten solos respecto a los que lo hacen a veces 0,927-0,495=0,432, 1,530-1,054=0,476 y 1,707-0,936=0,771. Este hecho hace pensar en la posibilidad de tratar la variable *ACOM* como numérica con sus valores 0, 1 y 2; ajustando de nuevo con la variable *ACOM* obtenemos unas estimaciones como las de la Tabla 4.8, con una lejanía de 18.37 con 24 grados de libertad, lo que no hace sospechar de un desajuste global. Este modelo respecto al anterior ha aumentado la lejanía en 0.96 pero ganamos 3 grados de libertad, por lo que este sería, por el momento, el modelo de elección.

Tabla 4.8. *El mismo modelo de la Tabla 4,7 con ACOM como numérica.*

Constante		SEXO	EDAD	ACOM
log P(regular)/P(mala)	0,102 (0,271)	-0,464 (0,234)	0,333 (0,228)	0,469 (0,130)
log P(buena)/P(mala)	-0,510 (0,318)	-0,612 (0,260)	-0,127 (0,261)	0,724 (0,156)
log P(muy buena)/P(mala)	-2,244 (0,613)	-0,490 (0,437)	-0,719 (0,501)	0,853 (0,305)

En resumen, los hombres perciben su salud como mejor que las mujeres, a igualdad de edad y grado de acompañamiento; la tendencia de los coeficientes de la edad para los tres logits habla que, en general, los viejos de entre 65 y 74 años tienen mejor percepción de su salud que los que tienen 75 o más, a igualdad de sexo y sensación de soledad; por último, los valores positivos de los coeficientes del grado de acompañamiento significan que los que se sienten solos perciben su salud como peor que los que se sienten acompañados, a igualdad de edad y sexo.

Ajustado un modelo, podemos estimar las predicciones para cada tipología de viejos y cada categoría de salud percibida. Por ejemplo, consideremos los hombres de 65 a 74 años que casi siempre se sienten solos; estos son 7+7+4+1=19 hombres, ¿cuantos dice el modelo que deberían percibir su salud como mala? En el Apartado 4,3 se vió que para la categoría de referencia, la salud percibida mala, la probabilidad venía dada por la expresión

$$\pi_1 = P(Y=1)\,\frac{1}{1+\sum_{s=2}^{c} e^{(\beta_{0s}+\beta_{1s}X_1+\beta_{2s}X_2+\ldots+\beta_{ms}X_m)}}$$

y ya que estamos considerando hombres de 65 a 74 años y casi siempre solos significa, por la codificación de la variables, que $SEXO=EDAD=ACOM=0$, por lo que la expresión anterior queda de la forma

$$\pi_s = P(Y=1)\,\frac{1}{1+\sum_{s=2}^{4} e^{\beta_{0s}}} = \frac{1}{1+e^{\beta_{02}}+e^{\beta_{03}}+e^{\beta_{04}}}$$

por lo que una estimación de la probabilidad de percibir la salud como mala es

$$\frac{1}{1 + e^{0,102} + e^{-0,510} + e^{-2,244}} = 0,36$$

y como hay 19 viejos de ese tipo, el número esperado para esa casilla es 19(0,36)=6,8. Este proceso se puede repetir para los distintos patrones de predictoras y podemos estimar los sujetos en cada casilla, como aparece en la Tabla 4.9.

Tabla 4.9. *Valores observados y predichos por el modelo de la Tabla 4.8.*

			SALUD PERCIBIDA			
ACOM	*EDAD*	*SEXO*	1(mala)	2(regular)	3(buena)	4(muy buena)
0	0	0	7 (6,8)	7 (7,5)	4 (4,1)	1 (0,7)
		1	22 (24,0)	19 (16,7)	9 (7,8)	0 (1,6)
	1	0	8 (5,4)	7 (8,4)	1 (2,9)	1 (0,3)
		1	14 (14,0)	14 (13,6)	3 (4,0)	1 (0,4)
1	0	0	4 (5,2)	9 (9,1)	7 (6,4)	2 (1,3)
		1	9 (11,2)	13 (12,5)	9 (7,6)	2 (1,7)
	1	0	3 (3,6)	8 (9,0)	6 (4,0)	0 (0,4)
		1	10 (7,8)	11 (12,1)	4 (4,6)	0 (0,6)
2	0	0	16 (14,3)	43 (40,6)	34 (36,7)	7 (8,4)
		1	24 (20,6)	32 (36,6)	28 (28,5)	9 (7,4)
	1	0	4 (6,7)	27 (26,4)	17 (15,0)	2 (1,9)
		1	12 (13,5)	36 (33,6)	16 (16,5)	2 (2,4)

4.3.5 Diagnóstico en regresión policotómica

Como en regresión logística binaria, en el caso policotómico existen dos estadísticos de interés derivados del modelo ajustado: la χ^2 y la lejanía. La expresión del estadístico χ^2 es similar a la del caso binario, solo que con una notación algo más complicada; sea n_{is} el número de individuos observados con un patrón de predictoras

$$x_{1i}, x_{2i}, \ldots, x_{mi}$$

en la categoría s de la variable respuesta y representemos por n_i la suma de los n_{is} para todas las categorías de la respuesta, es decir,

$$n_i = \sum_{s=1}^{c} n_{is}$$

el número de individuos con tal patrón de predictoras; para el caso de datos no agrupados, n_i=1.

Para el ejemplo de la Tabla 4.8, el primer patrón de predictoras serían los valores 0 para *ACOM*, 0 para *EDAD* y 0 para *SEXO*, por lo que n_{11}=7, n_{12}=7, n_{13}=4, n_{14}=1; por tanto n_1=19, y así para todas las filas de la tabla.

La expresión del estadístico χ^2 es ahora

$$\chi^2 = \sum_{i=1}^{k} \sum_{s=1}^{c} \frac{(n_{is} - n_i\hat{p}_{is})^2}{n_i\hat{p}_{is}}$$

donde k es el número de patrones distintos de las predictoras, en nuestro caso k=8, y los productos $n_i\hat{p}_{is}$ son los valores predichos por el modelo, para el patrón anterior, para la categoría s de la respuesta, que nos puede ofrecer cualquier programa estandar. La raiz cuadrada de las contribuciones, para cada categoría de la respuesta y para cada patrón de las predictoras, a la χ^2, es decir,

$$r_{is} = \frac{n_{is} - n_i\hat{p}_{is}}{\sqrt{n_i\hat{p}_{is}}}$$

son ahora los residuales de Pearson para el modelo logístico policotómico. El lector puede comprobar facilmente que para el caso de dos categorías de la respuesta estos residuales coinciden con los definidos para el caso binario.

De igual forma, la lejanía es ahora

$$\chi^2 = \sum_{i=1}^{k} \sum_{s=1}^{c} n_{is} \log \frac{n_{is}}{n_i\hat{p}_{is}}$$

por lo que la contribución del patrón de predictoras anterior para la categoría s, es decir, el residual de la lejanía para la casilla correspondiente de la tabla, es ahora

$$d_{is} = \sqrt{2n_{is} \log \frac{n_{is}}{n_i\hat{p}_{is}}}$$

que, aunque muchos programas no lo proporcionan, puede calcularse facilmente; el signo a elegir es el de la diferencia $n_{is} - n_i\hat{p}_{is}$, es decir, el número observado de individuos con el patrón de predictoras i que pertenecen a la categoría s de la respuesta, menos el número predicho por el modelo.

El lector debe tener presente la discusión en el capítulo anterior acerca de estas medidas como criterio de bondad de ajuste; aunque los datos de la Tabla 4.5 son agrupados, hay 16 casillas con valores esperados menores de 5, sobre un total de 48, lo que representa un 33%; por tanto, la aproximación de la distribución de la lejanía a la chi-cuadrado no es muy buena en este caso; de todas formas, ya que la lejanía para

el modelo ajustado es 17,41 para 21 grados de libertad, nos puede dejar suficientemente satisfechos del ajuste del modelo.

El estudio de los residuales adolece de los inconvenientes anteriores; de todas formas, el residual de Pearson más grande es 1,35, que corresponde a la casilla de los hombres (*SEXO*=0), del mayor grupo de edad (*EDAD*=1) y que se sienten siempre solos (*ACOM*=0), valor del cociente

$$\frac{1 - 0{,}28}{\sqrt{0{,}28}}$$

salvo efectos de redondeo; el residual de la lejanía más alto vale 2,98. Por tanto, de aquí tampoco tenemos evidencias de un mal ajuste general de los datos al modelo propuesto.

Para el modelo policotómico se puede extender el test de bondad de ajuste propuesto por Hosmer y Lemeshow para el caso binario y que se presentó en el apartado 3.5. Lesaffre (1986) recomienda utilizar la versión extendida por él a partir del test de Tsiatis (1980) para el caso binario; en este test en lugar de formar grupos en función de las probabilidades predichas por el modelo, lo que se hace es dividir el espacio de valores de las predictoras en distintas regiones. Sin embargo, las distribuciones de los estadísticos que de estos tests se derivan no son muy bien conocidas y/o la potencia es baja.

La definición de observación rara es análoga al caso binario; para el modelo policotómico, la decisión sobre si observación i perteneciente a la categoría s es rara se consigue comparando el modelo

$$\log \frac{P(Y=s)}{P(Y=1)} = \beta_{0s} + \beta_{1s}X_1 + \beta_{2s}X_2 + \ldots + \beta_{ms}X_m$$

con este otro

$$\log \frac{P(Y=s)}{P(Y=1)} = \beta_{0s} + \beta_{1s}X_1 + \beta_{2s}X_2 + \ldots + \beta_{ms}X_m + \delta_s Z$$

donde Z es una variable que vale 1 para la observación i y 0 en otro caso; las distintas elecciones de los parámetros δ_s se traducen en distintos tipos de observaciones extrañas. Lesaffre (1986) discute estas cuestiones y propone estadísticos similares a los descritos en el Capítulo III para el caso binario. Ya que existe poca disponibilidad de programas que realicen el diagnóstico para regresión policotómica, muchos autores recomiendan utilizar los métodos del Capítulo III a cada uno de los c-1 modelos logísticos binarios.

4.4 Modelos logísticos para variables ordinales

Vamos a considerar ahora una nueva situación en el sentido de disponer de una variable respuesta con más de dos categorías pero de naturaleza ordinal. En ciencias de la salud es bastante común el medir variables respuesta que son ordinales; piénsese en la variable "evolución tras un tratamiento", "salud percibida por los individuos", "satisfacción de los usuarios con los servicios sanitarios", etc.. Estas y otras muchas variables de interés para los salubristas tienen como denominador común su caracter ordinal; de ellas podríamos construir, por ejemplo, cinco categorías que guardan un cierto orden: muy mala, mala, regular, buena y muy buena. Hasta ahora se ha utilizado el modelo policotómico para analizar respuestas con más de dos categorías pero sin utilizar la información del caracter ordinal que algunas respuestas tienen. El objetivo de este apartado es la introducción de modelos logísticos que tienen en cuenta esta circunstancia; los *modelos de regresión logística ordinal*, también llamados *modelos de regresión ordinal*, se vienen desarrollando desde principios de la década de los ochenta a raiz de los artículos de Fienberg (1979), McCullagh (1980) y Anderson (1984). Greenland (1994) y Armstrong (1989) son dos buenas referencias sobre estos modelos.

El modelo policotómico no ha utilizado el carácter ordinal que tiene la salud percibida del ejemplo anterior; con variables ordinales, aparte de la probabilidad de que un individuo pertenezca a una determinada categoría, podemos definir otras probabilidades que pueden ser de interés; por ejemplo, la probabilidad de percibir la salud, a lo sumo, como buena; es decir, las suma de la probabilidad de percibirla como mala, regular o buena. A partir de estas nuevas probabilides que implican a varias categorías, podemos definir los logits correspondientes. Dependiendo de la definición que se adapte para estos logits, tendremos distintos modelos.

La utilización de estos modelos es cada día más frecuente en la literatura; por ejemplo, Brazer (1991) publicó un trabajo en donde trataban de estudiar las relaciones entre el cancer colorectal y un grupo de predictoras como la edad, el sexo, el valor hematocrito, dolor abdominal, etc; en base a una muestra de sujetos sometidos a una colonoscopia, estos fueron clasificados como 0 si no tenían neoplasia, 1 si presentaban un pólipo adenomatoso con tamaño inferior a 5 mm., 2 si era mayor de ese tamaño y 3 en caso de cancer; estos autores utilizaron un modelo de regresión logística ordinal, pues es evidente la jerarquía entre las categorías de la respuesta en cuanto a la salud.

4.4.1 El modelo de ventajas proporcionales

Con una variable ordinal como respuesta tiene sentido definir las siguientes nuevas probabilidades; consideremos la categoría s de la variable respuesta ordinal y notemos por $Y \leq s$ la nueva categoría constituida por todas las categorías inferiores o igual a s; por tanto,

$$P(Y \leq s) = P(Y=1) + P(Y=2) + \ldots + P(Y=s) = \pi_1 + \pi_2 + \ldots + \pi_s$$

es la probabilidad de que la variable respuesta tome un valor menor o igual que s; esta probabilidad es la *probabilidad acumulativa* correspondiente a la categoría s. El cociente

$$\frac{P(Y \leq s)}{1 - P(Y \leq s)} = \frac{P(Y \leq s)}{P(Y > s)}$$

se puede interpretar como la ventaja, *ventaja acumulativa*, de tener una respuesta menor o igual que s, respecto a tenerla mayor que s.

Consideremos a continuación una predictora X dicotómica con valores X=1 para los expuestos y X=0 para los no expuestos; para cada una de estas dos categorías podemos definir la ventaja acumulativa anterior, por lo que el cociente

$$\frac{P(Y \leq s \,/\, X=1) \,/\, P(Y > s \,/\, X=1)}{P(Y \leq s \,/\, X=0) \,/\, P(Y > s \,/\, X=0)}$$

es la razón de la ventajas de que la respuesta tome un valor menor o igual que s en la categoría de X=1, los expuestos al factor representado por X, respecto de la categoría X=0, los no expuestos. A esta nueva medida se le conoce también con el nombre de *razón de ventajas acumulativas* y aunque, por comodidad, hablemos de ventajas, para este modelo se debe entender que son acumulativas.

Siguiendo con los datos de la Tabla 4.1, puede ser de interés plantearse la siguiente pregunta: ¿cuál es la ventaja acumulativa de percibir la salud como mala entre los individuos de 65 a 74 años? Según se acaba de definir esta ventaja, la podemos estimar mediante el cociente (82/317) /((123+91+21)/317)=0,349; la misma ventaja acumulativa para los viejos de más de 74 años es 0,327, por lo que la razón de ventajas acumulativas de la clase mala, para los viejos de de 65 a 74 años respecto de los más mayores es 0,327/0,349=0,94; este resultado se puede interpretar diciendo que el riesgo de percibir la salud como mala respecto a percibirla como regular, buena o muy buena es 0,94 veces mayor en los viejos de más de 74 que entre los que tienen de 65 a 74 años. Dicho de otra forma, estos dos grupos de viejos son parecidos en cuanto a percibir su salud como mala.

De la misma manera, ¿cuál es la razón de ventajas de percibir la salud como mala o regular en el segundo grupo de edad, respecto del primero?; por un argumento análogo tal razón de ventajas acumulativas es (112)(154)/(205)(53)=1,59; por último, la razón de ventajas de percibir la salud como mala, regular o buena respecto de muy buena, en el segundo grupo de edad respecto del primero es (201)(21)/(296)(6)=2,38.

Como acabamos de ver, a partir de la Tabla 4.1 que es una tabla de 2 filas y cuatro columnas, es decir, una tabla 2x4, hemos calculado 3 razones de ventajas acumulativas que, de alguna manera, nos describen la relación de la salud percibida con la edad; pero ¿cómo interpretar estos resultados?. Como para la razón de ventajas clásica, estas tres razones de ventajas acumulativas nos indican, en principio, una mayor tendencia a percibir la salud como mejor en el grupo de edad 65-74 que en el grupo de mayores de 75. Si se quisiera obtener una medida sumario de esa relación, alguien podría estar tendado de utilizar el método de Mantel-Haenzsel sin reparar en que esa metodología está pensada para calcular una razón de ventajas común a partir de estratos independientes, condición que no tienen las tres tablas que hemos debido construir para calcular las 3 razones de ventajas acumulativas anteriores.

En general, en una tabla *r*x*c* hay *r*-1 parejas de filas que se pueden comparar de esta forma y, por cada una de ellas, se pueden formar *c*-1 ventajas acumulativas; en definitva, se pueden formar (*r*-1)(*c*-1) razones de ventajas acumulativas. En el ejemplo de la Tabla 4.1 donde *r*=2 y *c*=4, se han formado (2-1)(4-1)=3 razones ventajas acumulativas que aparecen en la Tabla 4.10. Si *Y* y *X* son variables independientes, las *r*-1 ventajas acumulativas correspondientes a la categoría *s*, deben ser iguales, y esto para las c-1 categorías; esto implica que todas las (*r*-1)(*c*-1) razones de ventajas sean iguales a la unidad.

Tabla 4.10. *Ventajas acumulativas y logits correspondientes a la Tabla 4.1.*

	Mala / Regular, Buena o Muy buena	Mala o Regular / Buena o Muy buena	Mala, Regular o Buena/ Muy buena
Grupo 65-74 Ventaja	0,349	1,83	14,095
Grupo >74 Ventaja	0,327	2,906	33,5
Razón de ventajas acumulativa	0,94	1,59	2,38
Logit acumulativo	-0,065	0,462	0,865

Veamos a continuación como construir un modelo de regresión para unos nuevos logits, los logaritmos de las ventajas acumulativas, que denominaremos *logits acumulativos*; de manera análoga al caso binario, estas ventajas acumulativas tratan a la variable respuesta como binaria, en el sentido que combinan las primeras *s* categorías y las enfrentan a las *c*-*s* categorías restantes. A partir de esas combinaciones, se pueden formar los logits acumulativos

$$\text{logit}\,[P(Y\leq s)] = \log\frac{P(Y\leq s)}{P(Y>s)}$$

$$s = 1, 2, \ldots, c\text{-}1$$

El modelo más simple para describir esos c-1 logits es

$$\text{logit}[P(Y \leq s)] = \alpha_s$$

Los c-1 parámetros α_s son cantidades que crecen con el aumento de s ya que los logits acumulativos aumentan, o al menos no disminuyen, cuando se añaden probabilidades al numerador y se sustraen del denominador, de la ventaja correspondiente; es decir, entre esos parámetros existe la relación

$$\alpha_1 \leq \alpha_2 \leq .... \leq \alpha_{c-1}$$

Consideremos ahora una predictora X; McCullagh (1980) propuso el modelo

$$\text{logit}\ [P(Y \leq s)] = \alpha_s + \beta X$$

para describir los c-1 logits acumulativos, en donde β es el parámetro de interés pues informa del cambio de los logits acumulativos con el cambio de la predictora X.

Para el modelo como se ha definido, el hecho de que β sea mayor que cero, implica que las ventajas acumulativas tienden a ser mayores con el aumento de X, lo que conlleva que los valores pequeños de Y están asociados a valores grandes de X. Ya que en la mayoría de los modelos de regresión suele ser común interpretar un coeficiente positivo como indicativo del aumento de la respuesta con el aumento de la predictora, es por lo que el modelo de ventajas proporcionales se suele expresar en esta otra forma

$$\text{logit}[P(Y \leq s)] = \alpha_s - \beta X$$

pues haciéndolo así, el hecho de que β sea positivo significa que a valores grandes de X le corresponden valores grandes de Y. Una expresión análoga a la anterior es la siguiente

$$P(Y \leq s) = \frac{e^{\alpha_s - \beta X}}{1 + e^{\alpha_s - \beta X}}$$

Para dos individuos con valores x_1 y x_2 de la predictora, notemos por

$$\text{logit}(P(Y \leq s)/x_1) \qquad \text{y} \qquad \text{logit}(P(Y \leq s)/x_2)$$

los logits acumulativos correspondientes para la categoría s en cada uno de esos dos individuos; entonces, la diferencia entre estos dos logits es

$$(\alpha_s - \alpha_s) - \beta(x_1 - x_2) = -\beta(x_1 - x_2)$$

o lo que es igual

$$\log\left(\frac{P(Y \leq s / x_1) / P(Y > s / x_1)}{P(Y \leq s / x_2) / P(Y > s / x_2)}\right) = -\beta(x_1 - x_2)$$

relación que establece que el logaritmo de la razón de ventajas acumulativas es proporcional a la diferencia entre los valores de la variable predictora, y no depende de la categoría s de la respuesta. Esta propiedad es el argumento para el nombre de *modelo de ventajas proporcionales (proportional odds model)*. Por tanto, la razón de ventajas acumulativas de tener un valor de respuesta menor o igual a s es

$$e^{-\beta(x_1 - x_2)}$$

veces superior en $X=x_1$ que en $X=x_2$.

La versión multivariante del modelo de ventajas proporcionales es

$$\text{logit}[P(Y \leq s)] = \alpha_s - (\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m)$$

Si A y B son dos individuos con valores

$$x_{1A}, x_{2A}, \ldots, x_{mA}$$
$$x_{1B}, x_{2B}, \ldots, x_{mB}$$

el logaritmo de la razón de ventajas acumulativas viene dado por

$$-\beta_1(x_{1A} - x_{1B}) - \ldots - \beta_m(x_{mA} - x_{mB}) = -\sum_{i=1}^{m} \beta_i(x_{iA} - x_{iB})$$

por lo que la razón de ventajas será

$$e^{-\sum_{i=1}^{m} \beta_i(x_{iA} - x_{iB})} = \prod_{i=1}^{m} e^{-\beta_i(x_{iA} - x_{iB})}$$

Si los individuos A y B son iguales respecto a todas las predictoras excepto a una de ellas, la X_i, la razón de ventajas acumulativa anterior queda reducida a

$$e^{-\beta_i(x_{iA} - x_{iB})}$$

por lo que la interpretación de los β_i es similar al modelo binario salvo que ahora de trata de razones de ventajas acumulativas, es decir, $-\beta_i$ es el logaritmo de la razón de ventajas entre dos individuos que se diferencien en una unidad en términos de la variable X_i, pero iguales respecto a las restantes variables presentes en el modelo. Obsérvese como este β_i no depende del punto de corte s elegido para formar el logit acumulativo.

Ajustemos, en primer lugar, el modelo que establece la independencia de la edad y la salud percibida

$$\text{logit}[P(Y \leq s)] = \alpha_s$$

a los datos de la Tabla 4.1; las estimaciones de los tres parámetros α_1, α_2, α_3 son –1,078, 0,777 y 2,913 respectivamente; estas estimaciones no son más que los logaritmos de las ventajas acumulativas de la distribución marginal de la salud percibida, es decir, de su distribución sin tener en cuenta la edad. En efecto, como quiera que en total hay 524 viejos de los cuales 133 perciben su salud como mala, independientemente de la edad, la ventaja acumulativa para la clase mala será 133/(524-133) y su logaritmo es precisamente el valor -1,078. Este modelo ajustado tiene una lejanía de 8,90 que, para 3 grados de libertad, es un valor muy grande, por tanto, este modelo no reproduce suficientemente los datos de la Tabla 4.1.

Ajustando un nuevo modelo de ventajas proporcionales con la edad como única preditora, los resultados son los que aparecen en la Tabla 4.11; el signo positivo de la edad nos debe hacer pensar en la posibilidad de que a mayor edad de los viejos, tengan peor percepción de su salud; de todas formas, y sin controlar por ninguna otra variable, la relación entre la edad y la salud percibida de los viejos no está clara, P=0,147 y, por otra parte, este modelo no reproduce suficientemente las observaciones pues tiene una lejanía de 6,8 para 2 grados de libertad, lo que supone una P=0,033, poniendo de manifiesto un desajuste significativo.

Tabla 4.11. *Modelo de ventajas proporcionales con EDAD como única predictora.*

VARIABLE	ESTIMACIÓN	ERROR ESTÁNDAR
α_1	-1,180	0,123
α_2	0,681	0,115
α_3	2,824	0,207
EDAD	0,237	0,164

Ya que la variable respuesta, la salud percibida ha sido codificada de 1 a 4 para las categorías mala, regular, buena y muy buena, y para los viejos de 65 a 74 años se les ha asignado el valor 0 y a los de más de 74 el 1, el signo positivo del coeficiente de la edad indica que a más edad peor salud percibida; en concreto, la ventaja estimada de que los más mayores tengan una percepción negativa de su salud es

$$e^{0,237}=1,27$$

veces mayor que la ventaja estimada para los que tienen entre 65 y 74 años; esta estimación del efecto de la edad es un resumen de las *OR* acumulativas 0,94, 1,59 y 2,37 que se pueden construir de la Tabla 4.1. A partir de la Tabla 4.11, podemos estimar las probabilidades de percibir la salud desde mala a muy buena en los dos grupos de edad; en efecto, el modelo para la clase codificada como de salud percibida mala es

$$P(Y\leq 1) = P(mala) = \frac{e^{-1,180+0,237EDAD}}{1 + e^{-1,180+0,237EDAD}}$$

por lo que, para los más jóvenes, $EDAD$=0, la probabilidad predicha de percibir la salud como mala es

$$\frac{e^{-1,180}}{1 + e^{-1,180}} = 0,235$$

mientras que para los de edad superior a 74 años, $EDAD$=1, es

$$\frac{e^{-1,180+0,237(1)}}{1 + e^{-1,180+0,237(1)}} = 0,280$$

De forma similar, la probabilidad de percibir la salud como mala o regular viene dada por

$$P(Y\leq 2) = P(mala\ o\ regular) = \frac{e^{0,681+0,237EDAD}}{1 + e^{0,681+0,237EDAD}}$$

por lo que para los más jóvenes será

$$P(Y\leq 2) = P(mala\ o\ regular) = \frac{e^{0,681}}{1 + e^{0,681}} = 0,664$$

y como la categoría de mala salud tenía una probabilidad de 0,235, la probabilidad de percibir la salud como regular será 0,664-0,235=0,429. Procediendo de esta manera se pueden calcular todas las probabilidades.

Con las probabilidades calculadas y teniendo en cuenta que hay 317 menores de 74 años, el número de viejos que es esperable que perciban su salud como mala serán 317(0,235)=74,5; en la Tabla 4.12 aparecen, para cada casilla, las frecuencias observadas y las predichas por el modelo; como se puede observar no existe gran acuerdo entre lo observado y lo predicho por el modelo que contiene como única predictora a la edad de los viejos, lo que concuerda con el desajuste antes mencionado.

Tabla 4.12. *Frecuencias observadas y predichas por el modelo de ventajas proporcionales.*

		SALUD PERCIBIDA			
		Mala	Regular	Buena	Muy buena
EDAD	65-74	82	123	91	21
		74,5	136,0	88,8	17,8
	> 74	51	103	47	6
		58,0	90,0	49,7	9,2

Utilizando las tres predictoras junto con sus interacciones, ninguna de estas fué significativa y el modelo con los tres efectos principales tiene una lejanía asociada de 26,76 con 29 grados de libertad; además, controlando por compañía y sexo, la edad no muestra asociación con la salud percibida, P=0,16. Ajustado el modelo solo con las variables *ACOM* y *SEXO,* los coeficientes de las dos variables indicadoras generadas a partir de *ACOM* son 0,74 y 1,01, con errores estándar de 0,26 y 0,21, lo que da pistas para pensar que las dos categorías que representan esas dos variables indicadoras, los que alguna vez se sienten solos y los que nunca perciben el sentimiento de soledad, se comportan de forma similar en cuanto a la percepción de su salud. Por tanto, recategorizamos esta variable y definimos una nueva variable *ACOMD* distinguiendo los que se sienten casi siempre solos, la categoría que tomamos como referencia, del resto; ajustado este nuevo modelo tiene una diferencia de lejanía con el modelo anterior de 2,6 que, como hemos ganado un grado de libertad al tener que estimar un parámetro menos para *ACOMD,* no es significativo para una χ^2 con un grado de libertad. En definitiva, el modelo final elegido es el que aparece en la Tabla 4.13.

Tabla 4.13. *Estimación del modelo de ventajas proporcionales final.*

VARIABLE	ESTIMACIÓN	ERROR ESTÁNDAR
α_1	-0,564	0,213
α_2	1,393	0,222
α_3	3,567	0,284
SEXO	0,376	0,166
ACOMD	-1,001	0,203

El signo postivo del sexo indica que los hombres, la categoría de referencia, tienen mejor percepción de su salud que las mujeres; el signo negativo de la variable que mide el grado de acompañamiento indica una mejor percepción de la salud entre los que no se sienten solos en relación a los que dicen sentirse solos muy frecuentemente. Un intervalo de confianza para el efecto del género será

$$e^{0,376 \pm 1,96(0,166)} = (1,05\ ,\ 2,02)$$

lo que se puede interpretar diciendo que las mujeres tienden a percibir su salud peor que los hombres, entre 1,05 y 2,02 veces, en iguales condiciones de soledad percibida. Los que se sienten acompañados tienen una razón de ventajas entre

$$e^{-1,001 \pm 1,96(0,203)} = (0,25\ ,\ 0,55)$$

de percibir su salud como mala en relación a los que se sienten solos.

Como para todo modelo ajustado, dado un patrón de predictoras, es posible calcular los valores predichos por el modelo; en la Tabla 4.14 se muestran los valores observados y los predichos por este modelo, calculados como antes se ha mostrado.

Tabla 4.14. *Frecuencias observadas y predichas por el modelo de la Tabla 4.13.*

		SALUD PERCIBIDA			
ACOMD	*SEXO*	1(mala)	2(regular)	3(buena)	4(muy buena)
0	0	15 (13,1)	14 (15,8)	5 (6,1)	2 (1,0)
	1	36 (37,2)	33 (32,8)	12 (10,6)	1 (1,4)
1	0	27 (32,7)	87 (80,1)	64 (62,7)	11 (13,5)
	1	55 (50,7)	92 (97,6)	57 (57,9)	13 (10,9)

Una característica interesante del modelo de ventajas proporcionales es que los efectos de las predictoras se mantienen si se colapsan categorías adyacentes de la respuesta; por tanto, dos estudios que utilicen distintas categorizaciones de la variable respuesta tendrán resultados similares. Por otra parte, y es algo poco conocido, al contrario que en regresión logística binaria, el modelo de ventajas porporcionales no es aplicable en los estudios en que el muestreo se hace según el valor de la variable resultado, por ejemplo en casos y controles; Greenland (1994) demuestra que con este diseño, las estimaciones de los parámetros del modelo son dependientes de la fracciones de muestro, por lo que, salvo en contadas ocasiones, sus estimaciones serán sesgadas.

Antes hemos visto que el formulación del modelo de ventajas proporcionales implica la llamada hipótesis de proporcionalidad, lo que significa imponer que cada predictora debe tener un efecto constante sobre la respuesta, independientemente del punto de corte elegido; el cumplimiento de esta hipótesis facilita evidentemente la interpretación de los resultados; parece entonces importante disponer de algún mecanismo para contrastar tal constancia de los efectos de las predictoras. Una aproximación exploratoria a esta cuestión es la representación gráfica de los efectos, *OR* y sus intervalos de confianza, de las variables predictoras para los distintos puntos de corte; evidentemente, esto vale solo para predictoras dicotómicas. La Figura 4.1 ilustra este proceder y parece que el efecto de la edad no es el mismo para todos los puntos de corte; mientras que los dos grupos de edad son parecidos para el primer punto de corte, la mala salud, para los otros dos son cada vez más evidentes las diferencias entre los viejos de más de 74 años respecto a los de 65 a 74. En consecuencia, hay fundadas sospechas del no cumplimiento de la hipótesis de proporcionalidad en relación a la edad.

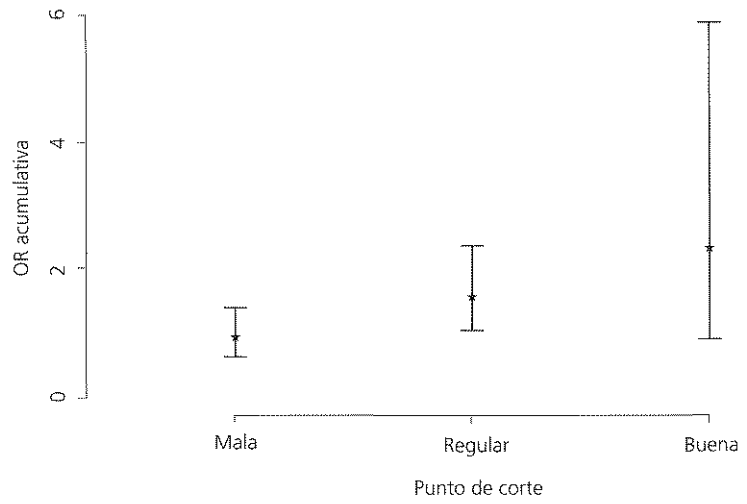


Figura 4.1. *Razones de ventajas acumulativas según el punto de corte.*

Lipsitz et al. (1996) ha propuesto un test para la bondad del ajuste del modelo de ventajas proporcionales, test que es una generalización del de Hosmer y Lemeshow visto en el Apartado 3.5; este test tiene todo su interés para el caso de predictoras continuas, es decir, para datos no agrupados donde la lejanía del modelo no es una buena medida de la bondad del ajuste realizado.

4.4.2 El modelo de ventajas proporcionales parciales

El gran atractivo del modelo de ventajas proporcionales está en la facilidad de su interpretación al suponer constante el efecto de las variables sea cual sea el corte elegido en la variable respuesta; tan solo necesitamos estimar un coeficiente por cada variable predictora para conocer su efecto. Desgraciadamente esta constancia del efecto de las predictoras no es la norma y es común encontrar situaciones en que tal efecto depende del punto de corte considerado; es decir, no se cumple la hipótesis de proporcionalidad. Peterson (1990) propuso una extensión del modelo de ventajas proporcionales, el modelo de ventajas proporcionales parciales (*partial proportional odds model*), que permite que algunas predictoras no tengan porqué cumplir la proporcionalidad; imaginemos sin pérdida de generalidad, que las $k \leq m$ primeras variables son las que supuestamente no cumplen esta hipótesis; el modelo de ventajas proporcionales parciales establece que

$$logit[P(Y \leq s)] = -\alpha_s - (\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m) - (\gamma_{1s} X_1 + \gamma_{2s} X_2 + \ldots + \gamma_{ks} X_k)$$

donde los γ_{is} son coeficientes adicionales para las variables que quizás no cumplan la proporcionalidad pues estos coeficientes sí dependen de la categoría o punto de corte s que se considere.

Como en el modelo de ventajas proporcionales, para la categoría de la respuesta tomada como referencia estos coeficientes se anulan, es decir, $\gamma_{i1}=0$, para cualquier variable. Veamos su interpretación para las restantes categorías; por motivos de sencillez supongamos que X_1 es una de las predictoras que se sospecha que no tiene un efecto constante y que es dicotómica; consideremos dos individuos con valores 1 y 0 de esta predictora, pero iguales iguales en el resto de ellas; si nos fijamos en la categoría s, el modelo anterior para el primer individuo será

$$logit[P(Y\leq s)] = -\alpha_s - (\beta_1(1) + \beta_2 X_2 + \ldots + \beta_m X_m) - (\gamma_{1s}(1) + \gamma_{2s}X_2 + \ldots + \gamma_{ks}X_k)$$

y

$$logit[P(Y\leq s)] = -\alpha_s - (\beta_1(0) + \beta_2 X_2 + \ldots + \beta_m X_m) - (\gamma_{1s}(0) + \gamma_{2s}X_2 + \ldots + \gamma_{ks}X_k)$$

para el segundo, por lo que la diferencia entre los logits correspondientes a la categoría s es ahora $(\beta_1+\gamma_{1s})$; por tanto, la *OR* entre estos dos individuos será

$$e^{-(\beta_1+\gamma_{1s})}$$

En definitiva, mientras en el modelo de ventajas proporcionales la estimación del efecto de X_1 era constante, ahora depende de la categoría que se considere; así, los c-1 riesgos son

$$e^{-\beta_1}\ ,\ e^{-(\beta_1+\gamma_{12})},\ldots,\ e^{-(\beta_1+\gamma_{1c-1})}$$

Desgraciadamente, la falta de programas informáticos para el ajuste de este nuevo modelo no nos permite compararlo con el de ventajas porporcionales.

4.4.3 El modelo de razones de continuidad

Otro modelo, propuesto por Fienberg (1979), para estudiar la asociación entre una respuesta ordinal y un conjunto de predictoras es el llamado *modelo de razón de continuidad (continuation ratio model)*. Este modelo está basado en la probabilidad condicional de que la respuesta tome el valor s supuesto que toma valores superiores o iguales a s, es decir, la probabilidad

$$\delta_s = \frac{\pi_s}{\pi_s + \pi_{s+1} + \ldots + \pi_c} = \frac{P(Y=s)}{P(Y\geq s)}$$

A partir de esta probabilidad condicional podemos formar las ventajas

$$\frac{\delta_s}{1 - \delta_s} = \frac{P(Y=s)}{P(Y>s)}$$

El modelo, para una sola predictora, se expresa como sigue

$$logit(\delta_s) = \log \frac{P(Y=s)}{P(Y>s)} = \theta_s - \beta X$$

apareciendo el signo negativo delante de los parámetros de las predictoras del modelo por los mismos argumentos dados para el modelo de ventajas proporcionales. La definición que se acaba de hacer de este modelo también implica que el coeficiente β de la predictora no depende de la categoría s que se considere.

Dados dos individuos con valores x_1 y x_2 de la predictora, notemos por

$$logit(\delta_s/x_1) \qquad \text{y} \qquad logit(\delta_s/x_2)$$

los logits correspondientes para la categoría s en cada uno de esos dos individuos; entonces,

$$logit(\delta_s/x_1) - logit(\delta_s/x_2) = (\theta_s - \theta_s) - \beta(x_1 - x_2) = -\beta(x_1 - x_2)$$

o lo que es igual

$$\log \left(\frac{P(Y=s/x_1) \,/\, P(Y>s/x_1)}{P(Y=s/x_2) \,/\, P(Y>s/x_2)} \right) = -\beta(x_1 - x_2)$$

donde la primera igualdad no es más que el logaritmo de la razón de las nuevas ventajas definidas; por tanto

$$e^{-\beta(x_1 - x_2)}$$

será la nueva razón de ventajas que no depende de la categoría s considerada.

De manera análoga a como se formaron las tablas para el modelo de ventajas proporcionales a partir de la Tabla 4.1 de salud percibida y edad, para el modelo de Fienberg las debemos formar considerando cada categoría de la respuesta contra todas las categorías posteriores. De esta manera, la primera a considerar es la Tabla 4.16

Tabla 4.16. *Salud percibida y edad.*

		SALUD PERCIBIDA	
		Mala	Regular, Buena o Muy buena
EDAD	65-74	82	235
	>74	51	156

que ya vimos que tiene una razón de ventajas igual a (51/156)/(82/235)=0,94; para la siguiente razón de ventajas, la tabla a construir es la Tabla 4.17 que aparece a continuación

Tabla 4.17. *Salud percibida y edad.*

		SALUD PERCIBIDA	
		Regular	Buena o Muy buena
EDAD	65-74	123	112
	>74	103	53

que da lugar a una razón de ventajas de 1,77; esta se puede interpretar como que un viejo con más de 74 años está a un riesgo 1,77 veces superior de percibir su salud como regular, respecto a buena o muy buena, en relación a uno con edad entre 65 y 74 años. Por último, la otra razón de ventajas que se puede construir es la que enfrenta la categoría "buena" con la "muy buena" a partir de la Tabla 4.18

Tabla 4.18. *Salud percibida y edad.*

		SALUD PERCIBIDA	
		Buena	Muy buena
EDAD	65-74	91	21
	>74	47	6

con un valor de 1,81. En resumen, las tres razones de ventajas estimadas 0,94, 1,77 y 1,81 cuantifican el riesgo de estar en las categorías de salud percibida mala, regular y buena, respectivamente, en relación a categorías superiores, de los más viejos respecto a los mas jóvenes; es decir, los que tienen más de 74 años tienden a percibir su salud peor que los que tienen entre 65 y 74 años.

Un primer modelo a ajustar sería el que no tiene ninguna predictora

$$\log \frac{P(Y=s)}{P(Y>s)} = \theta_s$$

que también se puede escribir así

$$\frac{P(Y=s)}{P(Y>s)} = e^{\theta_s}$$

que nos va a permitir interpretar los términos independientes θ_s; coeficientes de estos tenemos uno para cada una de las categorías de salud percibida mala, regular y buena;

para la primera categoría, mala salud, e^{θ_1} se puede interpretar como cuantas veces es más probable tener mala salud que tenerla mejor; e^{θ_2} cuantifica cuanto más probable es percibir la salud como regular en contra de sentirla como buena o muy buena; por último, e^{θ_3} es el número de veces que es más probable percibir la salud como buena que como muy buena. Ajustar este modelo equivale a estudiar la salud percibida en los viejos de la tabla 4.1 sin tener en cuenta la edad, por lo que en definitiva tenemos 133 viejos con mala salud percibida, 226 con una percepción regular de su salud, 138 con buena salud y 27 con muy buena; así, el logartimo del cociente entre la probabilidad de percibir la salud como mala en contra de las restantes categorías es 133/(226+138+27)=0,340, por lo que su logaritmo es -1,078. La probabilidad de percibir la salud como regular entre la correspondiente a buena o muy buena es 226/(138+27)=1,370 y su logaritmo 0,315. Por último, una estimación de la probabilidad de percibir la salud como buena respecto a muy buena es 138/27=5,11, cuyo logaritmo es 1,632.

Sin embargo, el programa utilizado para el ajuste, *S-Plus*, da como estimaciones -1,078, 1,393 y 2,710; el primer valor es el logaritmo de la ventaja de tener mala salud contra percibirla mejor; sin embargo, el valor 1,393 no es el logaritmo de la segunda ventaja sino que 1,393=0,315-(-1,078), es decir, el incremento del segundo logit respecto del primero; ya que esta diferencia es una estimación de esta otra

$$\log \frac{P(regular)}{P(buena,\ muy\ buena)} - \log \frac{P(mala)}{P(buena,\ muy\ buena)} =$$

$$\log \frac{\dfrac{P(regular)}{P(buena,\ muy\ buena)}}{\dfrac{P(mala)}{P(regular,\ buena,\ muy\ buena)}}$$

podemos asegurar que $e^{1,393}=4,027$ es una estimación del número de veces que es más probable el cociente del numerador que el del denominador en la expresión anterior; en efecto, 1,370/0,340 es, salvo errores de redondeo, precisamente 4,027. Si lo que nos interesa es la estimación de la probabilidad de la categoría regular contra la de las siguientes, no tenemos más que calcular la diferencia entre las dos estimaciones 1,393-1,078=0,315 y elevar *e* a ella. Por último, la diferencia 2,710-1,078=1,632 será la estimación del logartimo de pertenecer a la categoría de salud percibida buena en relación a muy buena; en efecto, la estimación de esa probabilidad es 138/27=5,111, por lo que el logaritmo es aproximadamente 1,632. En definitiva las estimaciones obtenidas son las correspondientes a θ_1, $\theta_2-\theta_1$ y $\theta_3-\theta_1$

El ajuste del modelo que contiene como predictora a la edad da lugar a una lejanía de 5,117 con 2 grados de libertad y unas estimaciones como las que aparecen en la Tabla 4.19

Tabla 4.19. *Estimación de modelo de razones de continuidad con la EDAD como predictora.*

VARIABLE	ESTIMACIÓN	ERROR ESTÁNDAR
θ_1	-1,190	0,117
θ_2-θ_1	1,398	0,144
θ_3-θ_1	2,740	0,234
EDAD	0,271	0,140

El coeficiente positivo de la *EDAD* indica que a valores mayores de edad le corresponden valores menores de la respuesta, es decir, peor salud percibida; en efecto, ya que los viejos más jóvenes están codificados como 0 y los más mayores como 1, dicho valor positivo es interpretable como que cuanta más edad tienen los viejos peor perciben su salud. Ya que antes vimos que β no depende de la categoría considerada,

$$e^{0,271}=1,31$$

es la estimación común de las tres razones calculadas 0,94, 1,77 y 1,81 a partir de las tablas 4.16, 4.17 y 4.18.

La versión multivariante del modelo de Fienberg es

$$\log \frac{P(Y=s)}{P(Y>s)} = \theta_s - (\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m)$$

donde los coeficientes de las predictoras tienen la misma interprteación que en modelo univariante, salvo que ahora se controla por las restantes predictoras; evidentemente, aquí se sigue manteniendo la constancia del coeficiente de cada variable, sea cual sea la categoría s de la respuesta. Más adelante veremos como contrastar el cumplimiento de esta restricción.

La Tabla 4.20 muestra las estimaciones de los coeficientes del modelo que contiene las predictoras edad, sexo y grado de acompañamiento; la lejanía resultante es 29,034 con 29 grados de libertad. Los signos positivos de la edad y el sexo indican que a más edad peor salud percibida y que las mujeres, $SEXO$=1, perciben su salud peor que los hombres, $SEXO$=0; por el contrario, los signos negativos de la variable que indica el grado de compañía expresan que cuanto más acompañados se sienten los viejos mejor percepción tienen de su salud.

Tabla 4.20. *Modelo de razones de continuidad con varias predictoras.*

VARIABLE	ESTIMACIÓN	ERROR ESTÁNDAR
θ_1	-0,819	0,200
θ_2-θ_1	1,513	0,149
θ_3-θ_1	2,927	0,241
SEXO	0,265	0,141
EDAD	0,257	0,142
ACOM(1)	-0,635	0,225
ACOM(2)	-0,915	0,184

Así, la probabilidad de poseer una mala percepción de la salud respecto a percibirla al menos como regular es

$$e^{0.257} = 1{,}29$$

veces mayor en los viejos de más de 74 años respecto a los de edades comprendidas entre 65 y 74 años, a igualdad de sexo y compañía. De igual forma, y por la restricción que impone el modelo acerca de la independencia del coeficiente y la categoría s considerada, 1,29 es la ventaja de los más viejos de percibir su salud como regular en contra de buena o muy buena y, por último, 1,29 es también la razón de ventajas de percibir la salud como buena en contra de muy buena. Para el género, a igualdad de las otras variables,

$$e^{0.265} = 1{,}30$$

estima la razón de ventajas de una peor percepción de la salud por parte de las mujeres. Por último, son la razones de los que se sienten frecuentemente acompañados, $ACOM$=1, y en los que siempre se sienten acompañados, $ACOM$=2, respectivamente, respecto a los que se sienten casi siempre solos, $ACOM$=0, a igualdad de edad y género.

$$e^{-0.6346} = 0{,}53 \quad \text{y} \quad e^{-0.9146} = 0{,}40$$

Como las dos indicadoras creadas para el grado de compañía tienen coeficientes parecidos, vistos sus errores estandar, dicotomizamos el grado de compañía como se hizo antes; las estimaciones resultantes aparecen en la Tabla 4.21.

Tabla 4.21. *Modelo final.*

VARIABLE	ESTIMACIÓN	ERROR ESTÁNDAR
θ_1	-0,724	0,199
θ_2-θ_1	1,505	0,149
θ_3-θ_1	2,913	0,240
SEXO	0,275	0,141
EDAD	0,262	0,142
ACOMD	-0,846	0,178

La diferencia en lejanías entre estos dos modelos ajustados es de 2,28 que, para un grado de libertad, resulta ser P=0,13. En definitiva, nos quedaríamos con este último modelo por ser más parsimonioso; a partir de él, una estimación del efecto del grado de compañía es

$$e^{-0.846} = 0{,}43$$

con un intervalo de confianza de

$$e^{-0.846 \pm 1{,}96(0.178)} = (0{,}30\ ,\ 0{,}61)$$

4.4.4 El modelo de razones de continuidad extendido

El modelo de Fienberg también impone la restricción de la independencia de los coeficientes respecto a la categoría considerada; esta limitación se puede obviar permitiendo que los parámetros varien con la categoría de la respuesta como se hizo para el modelo de ventajas proporcionales. De esta manera el modelo univariante de Fienberg para la edad queda en la forma

$$\log \frac{P(Y=s)}{P(Y>s)} = \theta_s - \beta_s EDAD$$

donde ahora tenemos 3 parámetros, $\beta_1, \beta_2, \beta_3$ para las categorías mala, regular y buena. Aplicado a los datos de la Tabla 4.1, da lugar a una lejanía de 0,00 con 0 grados de libertad; como se vió en el apartado anterior, el modelo que establecía la independencia del coeficiente de la edad en relación a la categoría de la respuesta tiene una lejanía de 5,117 con 2 grados de libertad, por lo que el cambio producido en lejanía es de 5,117 y de 2 en los grados de libertad, que corresponde a un valor P=0,077. En definitiva, hay sospechas fundadas de que la suposición de que los parámetros β_s son independientes de la categoría de la respuesta no es sostenible. Las estimaciones de los parámetros del modelo, junto a sus errores estandar, aparecen en la Tabla 4.22.

Tabla 4.22. *Modelo que permite diferentes efectos para la edad.*

Constante	EDAD	
log P(Y=1) / P(Y>1)	-1,053 (0,128)	-0,065 (0,206)
log P(Y=2) / P(Y>2)	1,146 (0,183)	0,571 (0,297)
log P(Y=3) / P(Y>3)	2,519 (0,274)	0,592 (0,538)

Según este modelo, las estimaciones de los coeficientes para la edad, según la categoría de la respuesta son -0,065, 0,571 y 0,592 y, como el lector puede comprobar, estas estimaciones no son más que los logartimos de las razones de ventajas 0,94, 1,77

y 1,81, calculadas de las tablas 4.16, 4.17 y 4.18; el hecho de que este modelo reproduzca exactamente nuestras observaciones es la razón de que tenga una lejanía nula.

La versión multivariante del modelo de Fienberg también se puede generalizar haciendo depender los parámetros del modelo del punto de corte elegido

$$\log \frac{P(Y=s)}{P(Y>s)} = \theta_s - (\beta_{1s}X_1 + \ldots + \beta_{ms}X_m)$$

Ajustando este modelo, el único parámetro que parece cambiar en función de la respuesta es el correspondiente a la edad, con una diferencia en lejanía de 5,608 para 2 grados de libertad, P=0,06, respecto al modelo con efecto constante para todas las categorías; como en el modelo de ventajas proporcionales, tampoco se aprecian diferencias entre las categorías 1 y 2 del grado de compañía por lo que se pueden colapsar. En definitva, los coeficientes del sexo y del grado de compañía sí son los mismos para las distintas categorías pero no el de la edad; los nuevos coeficientes estimados, junto con sus errores estandars, aparecen en la Tabla 4.23.

	SEXO	EDAD	ACOMD
log P(Y=1) / P(Y>1)	0,276 (0,141)	-0,094 (0,210)	-0,853 (0,178)
log P(Y=2) / P(Y>2)	0,276 (0,141)	0,572 (0,301)	-0,853 (0,178)
log P(Y=3) / P(Y>3)	0,276 (0,141)	0,608 (0,611)	-0,853 (0,178)

Tabla 4.23. *Coeficientes estimados con sus errores estándar según sexo y edad.*

4.4.5 Comparación de los modelos de McCullagh y Fienberg

Aunque los coeficientes en el modelo de Fienberg y en el de ventajas proporcionales juegan un papel parecido, no son directamente comparables pues mientras este predice probabilidades acumulativas, el primero trata con probabilidades condicionales; realmente, el modelo de Fienberg es el modelo de riesgos proporcionales propuesto por Cox (1972) para análisis de supervivencia con tiempos discretos.

Antes de plantearse qué modelo elegir para un caso concreto, es conveniente conocer las fortalezas y debilidades de cada uno de ellos. Una característica interesante del modelo de ventajas proporcionales es que los efectos de las predictoras se mantienen si se colapsan categorías adyacentes de la respuesta; por tanto, dos estudios que utilicen distintas categorizaciones de la variable respuesta tendrán resultados similares.

Por otra parte, la magnitud de las estimaciones también se mantiene si se invierte el orden de la variable respuesta (creciente o decreciente). Ninguna de estas dos deseables propiedades las cumple el modelo de Fienberg; con este modelo los resultados de una u otra codificación no son equivalentes pues dependerán de si se estudia la mejoría o el empeoramiento de la salud percibida.

Por otra parte, y es algo poco conocido, al contrario que en regresión logística binaria, ninguno de los modelos ordinales vistos es aplicable en los estudios en que el muestreo se hace según el valor de la variable resultado, por ejemplo, en los estudios de casos y controles; Greenland (1994) demuestra que con este diseño, las estimaciones de los parámetros del modelo son dependientes de la fracciones de muestro, por lo que, salvo en contadas ocasiones, sus estimaciones serán sesgadas; esta propiedad es una limitación para su aplicación en el análisis de los estudios epidemiológicos ya que una parte sustantiva de ellos son estudios de casos y controles.

Mientras que para cualquier punto de corte el modelo de ventajas proporcionales, por tratar probabilidades acumulativas, hace entrar en juego a todos los sujetos de estudio, el modelo de razones de continuidad solo los utiliza para el primer punto; en los siguientes cortes utiliza cada vez menos sujetos al tratar con probabilidades condicionales. Estos hechos hacen que sea esperable una mayor variación aleatoria para las estimaciones del modelo de razones de continuidad.

La elección de un modelo debe estar en relación a la hipótesis del estudio; el modelo de ventajas proporcionales es más facil de interpretar pues representa el riesgo entre cualquiera de las particiones que se pueden de la respuesta manteniendo su orden; a favor del modelo de Fienberg juega el hecho de poder ser ajustado con programas de regresión logística binaria reescribiendo el fichero de datos, (Armstrong, 1989).

En caso que ninguno de los modelos sean buenas representaciones para nuestros datos, queda la solución propuesta por Anderson (1984) (*the stereotype model*), que es similar al modelo policotómico pero que tiene en cuenta el orden de la variable respuesta. El modelo policotómico se definió en el Apartado 4.3 mediante la expresión

$$P(Y=s)\ \frac{e^{(\beta_{0s}+\beta_{1s}X_1+\beta_{2s}X_2+\ldots+\beta_{ms}X_m)}}{1+\sum_{s=2}^{c} e^{\beta_{0s}+(\beta_1X_1+\beta_2X_2+\ldots+\beta_mX_m)}}$$

lo que significa que el efecto de cada variable sobre la respuesta varía para cada categoría distinta a la de referencia y, además, sin ninguna restricción. Por tanto, el modelo policotómico es muy flexible, en el sentido que no impone restricciones, pero al precio de ser un poco parsimonioso y por tanto es complicado explicar sus resultados, especialmente cuando la respuesta tiene muchas categorías.

Una forma de simplificar el modelo anterior es escribiendo los coeficientes β_{is} en la forma $\beta_i t_s$; entonces el modelo anterior se puede escribir de esta otra manera

$$P(Y=s)\ \frac{e^{(\beta_{0s} + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m)\ t_s}}{1 + \sum_{s=2}^{c} e^{\beta_{0s} + (\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m)\ t_s}}$$

por lo que la *OR* de pertenecer a la categoría s respecto a la de referencia es ahora

$$e^{\beta_i t_s}$$

por unidad de aumento en la predictora X_i; por tanto, este efecto depende por una parte de la variable considerada y, por otra, del parámetro t_s que tiene que ver exclusivamente con la categoría s de la respuesta y, por tanto, es igual para todas las predictoras.

Esta nueva formulación simplifica mucho el modelo; en efecto, en el modelo policotómico, por cada predictora había que estimar c-1 parámetros, por lo que para las m predictoras son $m(c\text{-}1)$ parámetros; en el modelo de Anderson hay que estimar uno por cada predictora más uno por cada categoría distinta a la de referencia, en total $(c\text{-}1)+m$ que siempre es un número mucho menor que $m(c\text{-}1)$; además, este modelo es aplicable en cualquier diseño epidemiológico, no cambian los efectos de las predictoras si se invierte el sentido de la variable resultado; estas propiedades hacen que sea un modelo atractivo para modelar respuestas ordinales pero, desgraciadamente, los programas para estimar sus parámetros son de difícil acceso.

4.5 ¿Regresión ordinal o regresión binaria?

En este capítulo se han descrito distintos modelos de regresión aplicables en caso de una respuesta con caracter ordinal. Una pregunta no respondida hasta ahora es qué ganamos con utilizar estos métodos en relación a dicotomizar la respuesta y utilizar el modelo de regresión logística binaria.

Una primera crítica a la dicotomización de la respuesta es que el punto de corte que se elija es uno entre varios, por lo que su elección puede tener un componente importante de arbitrariedad; los resultados para el punto de corte elegido no tienen por qué ser iguales a si se cambia de criterio de dicotomización.

Por otra parte, y desde el punto de vista de la eficiencia estadística, la regresión ordinal es casi siempre más interesante que la regresión binaria. Armstrong (1989) estudió esta cuestión y comprobó que, salvo que se elija el punto de corte óptimo, la pérdida de eficiencia puede ser relevante. En definitiva, los modelos de regresión para variables ordinales pueden ser una herramienta valiosa y deben ser incorporados como métodos de análisis en los estudios tanto clínicos como epidemiológicos.

BIBLIOGRAFIA

Agresti, A. (1984): *Analysis of ordinal categorical data.* John Wiley and Sons. New York.

Agresti, A. (1990): *Categorical Data Analysis.* John Wiley and Sons. New York.

Anderson, J.A. (1984): Regression and ordered categorical variables. *J Royal Statist Soc (B)* 46, 1-30.

Aitkin, M., Anderson, D., Francis, B. and Hinde, J. (1989): *Statistical modelling in GLIM.* Oxford University Press. Oxford.

Aranda-Ordaz, F.J. (1981): On two families of transformations to additivity for binary response data. *Biometrika* 68, 357-363.

Armstrong, B.G., Sloan, M. (1989): Ordinal regression models for epidemiologic data. *Am J Epidemiol* 129, 191-204.

Ashby, M., Neuhaus J.M., Hauck W.W. et al.(1992): An annotated bibliography of methos for analysing correlated categorical data. *Statist Med*, 11, 67-99.

Atkinson, A.A. (1985) *Plots. transformations and regression.* Oxford University Press. London.

Begg, C.B., Gray, R. (1984): Calculation of polytomous logistic regression parameters using individualized regressions. *Biometrika* 71, 11-18.

Belsley, D.A., Kuh E, Welsch R.E. (1980): *Regression Diagnostics.* John Wiley and Sons. New York.

Berry, D.A., Stangl D.K. (1996). *Bayesian Biostatistics.* Marcel Dekker. New York.

Box, G.E.P., Tidwell, P.W. (1962): Transformations of the independent variables. *Technometrics* 4, 531-550.

Brazer, S.R., Pancotto, F.S., Long III, T.T. et al. (1991): Using ordinal logistic regression to estimate the likelihood of colorectal neoplasia. *J Clin Epidemiol* 44, 1263-1270.

Breslow, N.E., Day, N.E. (1980): *Statistical methods in cáncer research, Vol 1: The analysis of case-control studies.* IARC Scientific Publications, Lyon.

Cleveland, W.S. (1979): Robust locally weighted regression and smoothing scatterplots. *J Am Statist Assoc* 74, 828-836.

Collett D (1991). *Modelling binary data.* Chapman and Hall, London.

Cook, R.D. and Weisberg, S. (1982): *Residuals and Influence in regression.* Chapman and Hall.

Copas, J.B. (1983): Plotting p against x. *Appl Statist* 32, 25-31.

Cox, D.R. (1972): Regression models and life tables (with discussion). *J Royal Statist Soc (B)* 74, 187-220.

Dubin, N. and Pasternak, B.S. (1986): Risk assessment for case-control sugbroups by polychotomous logistic regression. *Am J Epidemiol* 123, 1101-1117.

Fienberg, S.E. (1980): *The Analysis of Cross-Classified Categorical Data*, 2nd Ed., MIT Press, Cambridge, MA.

Fienberg, S.E. (1979): Identification and estimation of age, period and cohort models in the analysis of discrete archival data. *Sociological Methodology*, 1-67.

Finney, D.J. (1947): The estimation from original records of the relationship between dose and quantal response. *Biometrika* 34, 320-334.

Fleiss, J.L. (1986): Significance tests have a role in epidemiologic research: reactions to A.M. Walker. *Am J Public Health* 76, 559-560.

Fowlkes, E.B. (1987): Some diagnostics for binary logistic regression via smoothing. *Biometrika* 74, 503-515.

Frank, G., Kamlet, M. (1989): Determining provider choice for the treatment of mental disorder: the role of health and mental health status. *Health Serv Res* 24, 83-103.

Gange, S.J., Muñoz, A., Saez, M., et al. (1996): Use of the beta-binomial distribution to model the effect of policy changes on appropiateness of hospital stays. *Appl Statist* 45, 371-382.

Greenland, S. (1994): Alternative models for ordinal logistic regression. *Statist Med* 13, 1665-1677.

Guerrero, V.M., Johnson R.A. (1982): Use of Box-Cox transformation with binary response models. *Biometrika* 69, 309-314.

Halperin, M., Blackwelder, W.C., Verter, J.I. (1971): Estimation of the multivariate logistic risk function: a comparasion of the discriminant function and maximum likelihood approaches. *J Chronic Dis* 24, 125-158.

Harrell F.E., Lee, K.L. and Pollock, B.G. (1988): Regression models in clinical studies: Determining relationships between predictors and response. *J Nat Cancer Inst* 80, 1198-1202.

Hart, J.S., George, S.L., Frei III E. et al. (1977): Prognostic significance of pretreatment proliferative activity in adult acute leukemia. *Cancer* 39, 1603-1617.

Hastie, T., Tibshirani, R. (1986): Generalized additive models. *Statist Sci* 1, 297-318.

Hastie, T., Tibshirani, R. (1987): Non-parametric logistic and proportional odds regression. *Appl Statist* 36, 260-276.

Herman, AA, Hastie, T. (1990): An analysis of gestational age, neonatal size and neonatal death using nonparametric logistic regression. *J Clin Epidemiol* 43, 1179-1190.

Hosmer, D.W., Lemeshow, S. (1980): Goodness of fit test for the multiple logistic regression model. *Comm Statist* A9, 1043-1069.

Hosmer, D.W., Lemeshow, S. (1989): *Applied logistic regression*. Wiley, New York.

Johnson, W. (1985): Influence measures for logistic regression: Another point of view. *Biometrika* 72, 59-65.

Kay, R., Little, S. (1987): Transformations of the explanatory variables in the logistic regression model for binary data. *Biometrika* 74, 495-501.

Landwehr, J.M., Pregibon, D., Shoemaker, A.C. (1984): Graphical methos for assesing logistic regression models. *J Amer Statist Assoc* 79, 61-71.

Lesaffre E. (1986): *Logistic Discriminant Analysis with Applications in Electrocardiography.* Doctoral Degree in Science Dissertation, Univ of Leuven, Belgium.

Lipsitz, S.R., Fitzmaurice, G.M., Molenberghs G. (1996): Goodness-of-fit tets for ordinal response regression models. *Appl Statist* 45, 175-190.

Mc Cullagh, P. (1980): Regression models for ordinal data (with discussion). *J Royal Statist Soc*, Series B 42, 109-127.

McCullagh, P., Nelder, J.A. (1989): *Generalized linear models*, 2nd Ed. Chapman and Hall, London.

Mickey, R.M., Greenland, S. (1989): The impact of confounder selection criteria on effect estimation. *Am J Epidemiol* 129, 125-137.

Miettinen, O.S. (1976): Stratification by a multivariate confounder score. *Am J Epidemiol* 104, 609-620.

Minkin, S. (1989b): Fit assessment and identification of functional form in logistic regression. *Appl Statist* 38, 343-350.

Mittlböck, M., Schemper, M. (1996): Explained variation for logistic regression. *Statist Med* 15, 1987-1997.

Pendergast, J.F., Gange, S.J., Newton, M.A. et al. (1996): A survey of methods for analyzing clustered binary response data. *Int Statist Rev* 64, 89-118.

Peterson, B., Harrell, F.E. (1990): Partial proportinal odds models for ordinal response variables. *Appl Statist* 39, 205-217.

Pregibon, D. (1981): Logistic regression diagnostics. *Ann Statist* 9, 705-724.

Raftery, A.E. (1995): Bayesian model selection in social research (with discussion). In *Sociological Methodology* 1995, ed. Marsden P.V., Blackwell, Oxford.

Royston, P. (1992): The use of cusums and other techniques in modelling continuous covariates in logistic regression. *Statist Med* 11, 1115-1129.

SAS Technical Report P-200 (1990): SAS/STAT software: CALIS and LOGISTIC Procedures, Release 6.04, SAS Institute Inc.

Schall, R., Zucchini, W. (1990): Model selection and the estimation of odds ratio in the presence of extraneous factors. *Statist Med* 9, 1131-1141.

Segal, M.R., Weiss, S.T., Speizer, F.E. et al. (1988): Smoothing methods for epidemiologic analysis. *Statist Med* 7, 601-611.

Shapiro, M., Muñoz, A., Tager, I.B. et al. (1982): Risk factors for infection at the operative site after abdominal or vaginal hysterectomy. *N Engl J Med* 307, 1661-1666.

Shapiro, S., Slone, D., Rosenberg, L. et al. (1979): Oral-contraceptive use in relation to myocardial infarction. *Lancet* 1, 743-747.

Stone, C.J., Koo, C.Y. (1985): Additive splines in statistics. *Proc Statist Comput Sect ASA*, 45-48.

Tsiatis, A.A. (1980): A note on the goodness of fit for the logistic regression model. *Biometrika* 67, 250-251.

Wang, P.C. (1985): Adding a variable in generalized linear models. *Technometrics* 27, 273-276.

Wang, P.C. (1987): Residual plots for detecting nonlinearity in generalized linear models. *Technometrics* 29, 435-438.

Williams, D.A. (1982): Extra-binomial variation in logistic linear models. *Appl Statist* 31, 144-148.

ANEXO

PROGRAMAS INFORMÁTICOS PARA ANALIZAR DATOS MEDIANTE REGRESIÓN LOGÍSTICA

Actualmente son varios los programas informáticos que incorporan procedimientos para analizar datos mediante regresión logística, especialmente para el caso binario. No es fácil la elección entre ellos debido a que son varias las características a considerar en la relación coste-beneficio. Partiendo de este hecho, solo comentaremos algunas características de dos de ellos: SPSS y S-Plus.

1. Programa SPSS

La elección del programa SPSS viene condicionada por ser uno de los paquetes estadísticos más difundidos en nuestro país, especialmente en el ámbito sanitario. La versión 9.0 para Windows permite capturar distintos tipos de bases de datos, entre otras, Excell, Access, dBase y también ficheros de texto.

Una vez leídos los datos, en la ventana **Analizar** se encuentran la mayoría de los procedimientos estadísticos que se pueden realizar con este programa; al activarla aparece otra que se denomina **Regresión,** dentro de la cual está logística binaria. Al pulsarla aparece una pantalla encabezada con el nombre **Regresión logística** que es la pantalla principal

En la parte izquierda hay una ventana donde están todas las variables que aparecen en el fichero de datos

En la parte central de la pantalla aparecen cuatro ventanas; en la superior, denominada **Dependiente**, es donde se debe incluir la variable dependiente que, en general, se codifica con los valores 0 y 1. A continuación se incluyen en la segunda ventana, llamada **Covariables**, la o las variables independientes y sus posibles interacciones; una tercera ventana, **Método**, permite elegir alguno de los procedimientos automáticos de selección de variables hacia adelante o hacia atrás, o bien, incluir a todas en el modelo, que es el criterio por defecto y que se denomina INTRODUCIR. Por último, la cuarta ventana permite seleccionar una parte de los individuos del fichero con relación a alguna condición respecto a alguna de las variables.

Para el caso de variables predictoras categóricas con más de dos categorías, SPSS permite construir las variables *dummy* correspondientes; para ello, pulsando el botón **Categórica** aparecen dos nuevas ventanas. En la de la izquierda están todas las predictoras y se trata de pasar a la ventana de la derecha las predictoras categóricas; esta pantalla también permite elegir distintos métodos de formación de variables *dummy*; en esta monografía se ha utilizado el que SPSS denomina *Indicador*.

Todavía queda por especificar qué categoría de la predictora se tomará como referencia para lo que se dispone de dos opciones: a) la categoría cuyo valor asignado sea el más grande, la opción por defecto; b) el más pequeño. Una vez especificada esta información y tras pulsar el botón **Continuar** se vuelve a la pantalla anterior.

De nuevo en la pantalla principal hay otro botón denominado **Guardar,** que lleva a una nueva pantalla que nos permite obtener los estadísticos necesarios para el diagnóstico del modelo; marcando las distintas opciones podemos obtener variables cuyos valores son las probabilidades predichas, el grupo predicho según el modelo definido, la distancia de Cook, la influencia, los delta-beta y distintos tipos de residuales para cada uno de los individuos del fichero; pulsando el botón **Continuar** volvemos de nuevo a la pantalla principal.

Por último, pulsando el botón **Opciones** permite, entre otras facilidades, el cálculo del estadístico de Hosmer-Lemeshow, la elección de los criterios de entrada y salida de las variables en los criterios de selección automática, la elección del nivel de confianza para los intervalos de la razón de ventajas y la elección del punto de corte para la clasificación predicha en uno de los dos grupos de la variable respuesta; pulsando el botón **Continuar** de nuevo se vuelve a la pantalla principal.

Una vez realizado todo el proceso anterior, pulsando el botón **Aceptar** de la pantalla principal y después de una serie de iteraciones, SPSS nos lleva a la pantalla de resultados.

Con la versión 9.0 de SPSS para Windows también se puede llevar a cabo la regresión policotómica o **Logística multinomial**; también está dentro de la ventana **Analizar** y **Regresión** y tiene un formato parecido al de la regresión logística binaria; por defecto, la categoría de referencia de la variable resultado es la que tiene un valor mayor, de tal forma que si se quiere utilizar otra categoría como referencia es necesario recodificarla de tal manera que a ésta se le asigne el valor superior. Para las predictoras dispone de dos ventanas para distinguir las variables categóricas, de las que forma las correspondientes dummy, de las numéricas. Para la versión 10 se anuncia la incorporación al paquete de algún modelo de regresión ordinal.

2. Programa S-Plus

S-Plus es un paquete estadístico de gran versatilidad, con muchas posibilidades gráficas; aunque ciertamente es más complejo de utilizar que SPSS, la última versión S-Plus 2000 está construida bajo entorno Windows, lo que facilita bastante su manejo. Incluye el modelo de regresión logística en la clase de modelos lineales generalizados, por lo que su ajuste se realiza mediante la función **glm** *(generalized linear models)*. La especificación del modelo se realiza mediante una fórmula y el argumento **"family=binomial"** de la función **glm**. La sintaxis para ajustar el modelo es la siguiente

$$> modelo <- glm\ (Y \sim X_1 + X_2 + \ldots + X_P,\ family=binomial)$$

donde Y es el nombre de la variable dependiente, que en este caso deberá ser dicotómica, X_1, X_2, ..., X_P son las variables predictoras y *modelo* es el objeto donde se almacenarán los resultados del ajuste. Una vez creado el objeto *modelo* es posible obtener tanto las estimaciones de los parámetros como los estadísticos necesarios para el diagnóstico.

Con S-Plus se puede evaluar la forma funcional de las predictoras pues este programa posibilita la utilización de muchas técnicas de alisamiento, en particular, mediante los modelos aditivos generalizados ***(gam)***.

Con relación a los contenidos de esta monografía, la nueva versión S-Plus 2000 trae incorporadas una serie de *librerías*, entre ellas *Design*, que contiene la función **lrm** que permite estimar los modelos para la regresión logística ordinal. Un gran atractivo de este programa es la existencia de un grupo coordinado por el Departamento de Estadística de la *Carnegie Mellon University*, http://www.stat.cmu.edu/S/, en donde aparecen muchas rutinas a ser utilizadas.

Los cálculos y gráficos de esta monografía se han realizado con S-Plus.

Índice de materias

Ajuste lineal, 105
Alisado para una respuesta binaria, 100
Alisado, 94, 110
Alisamiento, 89, 90, 91
Alisamiento mediante núcleos, 95
Alisamiento no ponderado, 97
Alisamiento ponderado, 97
Alternativas al modelo de regresión logística, 79
Anticonceptivos, 66

Bajo peso al nacer, 12
Bondad del ajuste, 37, 86, 87

Cálculo diferencial, 30
Cáncer de pulmón, 51
Categoría de referencia, 60
Categoría promedio, 63
Centrado de las variables cuantitativas, 75, 76
Cohorte de Framimgham, 74
Colinealidad, 119
Componente semántico del modelo, 42
Componente sistemático, 72, 73
Conceptos básicos de estadística, 9
Confusión, 47
Contraceptivos orales, 55

Diagnóstico en regresión logística, 77
Diagnóstico en regresión policotómica, 138
Dicotomización, 160
Distancia de Cook, 114, 115, 116
Distribución binomial, 14
Distribución chi-cuadrado, 11, 12, 34
Distribución multinomial, 124

Escuela Andaluza de Salud Pública, 9
Estadística bayesiana, 10
Estadística de Wald, 47
Estadístico χ^2 de Pearson, 38, 86
Estadístico χ^2, 37
Estimación de los coeficientes, 44, 129
Estratificación, 12
Estudios de cohorte, 10

Factor de riesgo, 17
Forma funcional de la predictoria, 106
Función de probabilidad, 14
Función de probabilidad multinomial, 125
Función de verosimilitud, 27, 29, 30, 31
Función Logit, 81, 82
Función log-log complementaria, 81, 82
Función probit, 81, 82
Función tricubo, 95, 96

Grados de libertad, 47
Grupos de riesgo, 87

Hábito de fumar de la madre, 12
Histerectomía, 75

Índice de riesgo, 71, 72
Infarto de miocardio, 55, 66
Inspiración profunda, 77
Interacción, 52
Interpretación de los coeficientes, 127
Investigación sanitaria, 9

Lejanía, 35, 38, 47, 87
Ley multiplicativa de las probabilidades, 14
Logaritmo neperiano, 17
Logit (p), 18, 19, 24, 36, 41, 47, 52, 53, 66, 67, 70, 79, 88, 89, 103
Logit acumulativo

Matriz de varianza-covarianza, 72, 132
Matriz sombrero, 112
Máxima verosimilitud, 27, 28, 30, 32, 33
Medias móviles, 92
Método ajustado, 35
Método de Wald, 40
Modelo anidado, 37
Modelo de Cox, 14
Modelo de estereotipo, 159
Modelo de McCullagh, 158
Modelo de razón de continuidad, 151, 155, 156, 159
Modelo de razones de continuidad extendido, 157
Modelo de regresión, 12
Modelo de regresión de supervivencia, 14
Modelo de regresión lineal, 12
Modelo de regresión logística, 13
Modelo de regresión logística binaria simple, 9, 17
Modelo de regresión logística oridinal, 141

Modelo de regresión logística policotómica, 126
Modelo de regresión múltiple, 9
Modelo de regresión ordinal, 141
Modelo de ventajeas proporcionales, 141, 145, 146, 147, 148, 149
Modelo de ventajas proporcionales parciales, 150, 159
Modelo logístico, 10, 19, 24
Modelo logístico binario múltiple, 41
Modelo logístico multinivel, 10
Modelo logístico múltiple (o multivariante), 41
Modelo logístico multinominal, 133
Modelo log-log, 80
Modelo operativo, 65
Modelo paramétrico, 90
Modelo policotómico, 128, 141, 159, 160
Modelo probit, 80
Modelo saturado, 35
Modelo univariante de Fienberg, 152, 157, 158, 159
Muestreo por conglomerado, 120

Nube de puntos de los residuales parciales, 109
Nube de puntos para una variable añadida 106
Nube de puntos para una variable construida, 107
Nutrición infantil, 36

Observaciones extremas, 118
Odds ratio (OR) 21, 23 ,27, 34, 43, 46, 52, 53, 55, 60, 72
Outlier (ver Observación extrema) 118

Paquete informático, 40
Paradoja de Simpson, 50
Parámetro de alisamiento, 92
Parámetros o coeficientes del modelo, 13
Polinomio, 101
Polinomios a trozos, 102
Principio de parsimonia, 46
Probabilidad acumulativa, 142
Procedimiento LOWESS,
Programas informáticos de regresión logística, 84
Programa S-Plus, 154, 167
Programa SPSS, 165,166
Proportional odds model 143, 150
Puntos influyentes, 114
Puntos de unión, 102

Razón de producto cruzado, 23
Razón de ventaja, 21, 23, 27, 41, 74
Razón de verosimilitud, 35

Regresión local, 98
Regresión logística, 9
Regresión logística binaria, 128, 160
Regresión logística ordinaria, 123, 160
Regresión logística policotómica, 123
Regresión mínimo-cuadrática, 98
Regresión por *splines*, 101
Rendimiento escolar, 36
Residual, 83, 84, 86, 106
Residual ajustado, 112
Residuales crudos, 83
Residuales crudos estandarizados (r. de Pearson), 83
Residual de la lejanía 139, 140
Residual de la lejanía estandarizado, 111
Residuales de la variable añadida (z-residuales), 106
Residuales de la variable construida, 108
Residuales de Pearson, 11, 112, 115
Residuales parciales, 109, 110
Respuesta nominal, 10
Respuesta ordinal, 10
Riesgo asociado a infección, 74
Riesgo relativo, 23

Selección de variables, 63
Selección hacia adelante, 64
Selección hacia atrás, 64
Selección paso a paso, 64
Smoothing (ver Alisamiento) 90, 98
Sobredispersión, 120, 122
Splines, 101, 103
Spline cúbico, 103, 104
Stereotype model 159
Supervivencia, 51

Tabla 2x2, 5, 23
Tabla de contingencia (tabla 2 x 2), 11, 23
Tabla marginal, 45
Tablas parciales, 45
Teorema de Bayes, 25
Test de chi-cuadrado, 11
Test de Hosmer-Lemeshow 166
Tipos de muestreo, 24
Transformación de muestreo, 24
Transformación logística, 17
Transformación logit, 16, 17

Unidades de medida de las predictoras, 71, 72

Variable categórica, 123
Variable confundente, 49, 64
Variable construida, 108
Variable cuantitativa, 41
Variable de bernoulli, 14
Variable dicotómica, 11
Variable indicadora, 57, 58, 59
Variable nominal, 123, 126
Variable ordinal, 123, 141
Variable predictora, 11, 51, 88
Variable resultado, 11, 41 ,51
Vasoconstricción en la piel, 77
Ventaja u oportunidad, 16
Ventaja acumulativa, 142
Ventana de alisamiento, 92, 94
Verosimilitud del modelo ajustado, 35
Verosimilitud del modelo saturado, 35