REGRESIÓN LOGISTICA

Índice

1. Objetivos

- Estudiar los modelos de regresión logística considerando sus fundamentos teóricos y sus aplicaciones.
- Se pretende que al finalizar el curso los alumnos hayan adquirido los conocimientos relativos a las formulaciones, supuestos y condiciones de aplicación del modelo de regresión logística.
- El objetivo final es que los alumnos puedan plantear y diseñar una investigación empírica en la que se requiera aplicar dichos modelos.

2. Introducción al modelo de regresión logística

Los modelos de regresión logística son modelos estadísticos en los que se desea conocer la relación entre:

Una variable dependiente cualitativa, dicotómica (regresión logística binaria o binomial) o con más de dos valores (regresión logística multinomial).

Una o más variables explicativas independientes, o covariables, ya sean cualitativas o cuantitativas, siendo la ecuación inicial del modelo de tipo exponencial, si bien su transformación logarítmica (*logit*) permite su uso como una función lineal.

Como se ve, las covariables pueden ser cuantitativas o cualitativas. Las covariables cualitativas deben ser dicotómicas, tomando valores 0 para su ausencia y 1 para su presencia (esta codificación es importante, ya que cualquier otra codificación provocaría modificaciones en la interpretación del modelo). Pero si la covariable cualitativa tuviera más de dos categorías, para su inclusión en el modelo debería realizarse una transformación de la misma en varias covariables cualitativas dicotómicas ficticias o de diseño (las llamadas variables *dummy*), de forma que una de las categorías se tomaría como categoría de referencia. Con ello cada categoría entraría en el modelo de forma individual. En general, si la covariable cualitativa posee n categorías, habrá que realizar $n-1$ covariables ficticias.

Por sus características, los modelos de regresión logística permiten dos finalidades:

- Cuantificar la importancia de la relación existente entre cada una de las covariables y la variable dependiente, lo que lleva implícito también clarificar la existencia de interacción y confusión entre covariables respecto a la variable dependiente (es decir, conocer la *odds ratio* para cada covariable).
- Clasificar individuos dentro de las categorías (presente/ausente) de la variable dependiente, según la probabilidad que tenga de pertenecer a una de ellas dada la presencia de determinadas covariables.

No cabe duda que la regresión logística es una de las herramientas estadísticas con mejor capacidad para el análisis de datos en investigación clínica y epidemiología, de ahí su amplia utilización.

El objetivo primordial que resuelve esta técnica es el de modelar cómo influye en la probabilidad de aparición de un suceso, habitualmente dicotómico, la presencia o no de diversos factores y el valor o nivel de los mismos. También puede ser usada para estimar la probabilidad de aparición de cada una de las posibilidades de un suceso con más de dos categorías (politómico).

3. Introducción a la selección de variables

Pero, del conjunto de variables que pueda tener un estudio, ¿qué variables deben introducirse en el modelo? El modelo debe ser aquél más reducido que explique los datos (principio de parsimonia), y que además sea clínicamente congruente e interpretable. Hay que tener en cuenta que un mayor número de variables en el modelo implicará mayores errores estándar.

Deben incluirse todas aquellas variables que se consideren clínicamente importantes para el modelo, con independencia de si un análisis univariado previo se demostró o no su significación estadística. Por otro lado, no debería dejarse de incluir toda variable que en un análisis univariado previo demostrara una relación "suficiente" con la variable dependiente. Como se ve, no se habla de significación estadística $(p<0,05)$, que sería un criterio excesivamente restrictivo, sino de un cierto grado de relación (por ejemplo $p<0,25$). La laxitud de esta recomendación se debe a que un criterio tan restrictivo como una $p<0,05$ puede conducir a dejar de incluir en el modelo covariables con una débil asociación a la variable dependiente en solitario pero que podrían demostrar ser fuertes predictores de la misma al tomarlas en conjunto con el resto de covariables.

Una cuestión importante a tener en cuenta es el correcto manejo de las variables cualitativas transformadas en varias variables ficticias.

Siempre que se decida incluir (o excluir) una de estas variables, todas sus correspondientes variables ficticias deben ser incluidas (o excluidas) en bloque. No hacerlo así implicaría que se habría recodificado la variable, y por tanto la interpretación de la misma no sería igual.

Otro aspecto de interés es la significación que pudiera tener cada variable ficticia. No siempre todas las variables ficticias de una covariable son significativas, o todas no significativas. En estos casos es recomendable contrastar el modelo completo frente al modelo sin la covariable mediante la prueba de razón de verosimilitud (es decir, se sacarían del modelo en bloque todas las variables ficticias de la covariable de interés). La decisión se tomaría dependiendo del resultado de la prueba y del interés clínico de la covariable: Si se obtiene significación en este contraste, la variable permanece en el modelo; si no se obtiene significación y la covariable es de interés clínico a criterio del investigador, habría que valorar la magnitud en la que se distancia de la significación para decidir si la covariable debe permanecer o no en el modelo.

Una vez se dispone de un modelo inicial debe procederse a su reducción hasta obtener el modelo más reducido que siga explicando los datos. Para ello se puede recurrir a métodos de selección paso a paso, bien mediante inclusión "hacia adelante" o por eliminación "hacia atrás", o a la selección de variables por mejores subconjuntos de covariables. Estos métodos se encuentran implementados en numerosos paquetes estadísticos, por lo que son muy populares. Dado que para la comprensión de los métodos de selección paso a paso se requiere un conocimiento previo acerca del ajuste del modelo, éste es un aspecto que debe ser tratado en otro momento; se sugiere al lector que se introduzca en este aspecto una vez tenga conocimientos sobre el análisis del ajuste del modelo. No obstante hay que advertir que su uso nunca puede sustituir a la valoración juiciosa de los modelos que van surgiendo de forma seriada en cada paso y del modelo final. No hacerlo así puede llevar a dar por bueno un modelo surgido de forma automática (por criterios preestablecidos por el paquete estadístico muchas veces mal conocidos por el usuario del software), con escaso valor clínico.

Cada vez que se encuentre ante un modelo de regresión logística (el inicial, cualquiera de los intermedios o el final), se tendrá que contrastar su significación global, mediante las pruebas de ajuste global del modelo.

Una vez se dispone un modelo preliminar, se podrían incluir factores de interacción, es decir, estudiar cómo la asociación de dos o más covariables puede influir en la variable dependiente. Existen estrategias de desarrollo de modelos de regresión por las que se recomienda la inclusión en el modelo inicial de todas las covariables necesarias más las interacciones de las mismas, o por lo menos, las interacciones de primer orden (tomadas las covariables dos a dos), a los que se les llama modelos saturados. Interacciones de mayor orden suelen ser de difícil interpretación. En cualquier caso siempre hay que tener presente las limitaciones de tamaño muestral (que se verán luego), y de interpretación desde el punto de vista clínico (no se deberían incluir interacciones de significado incierto).

Otra estrategia en el desarrollo del modelo final es el diseño y ajuste de un modelo final preliminar sin interacciones, en el que luego se ensayarían la inclusión, uno por uno, de términos de interacción que pudieran tener traducción clínica (Hosmer y Lemeshow), y valorar su significación respecto del modelo previo sin interacciones.

Una vez se haya decidido la inclusión de un factor de interacción, se tendrá en cuenta que siempre deberán estar incluidas también de forma aislada en el modelo las covariables que componen la interacción (principio jerárquico): si la interacción es "*HTA-diabetes*", en el modelo se encontrarán como covariables *HTA* y diabetes (*DM*):

$$logit = \beta_0 + \beta_1 HTA + \beta_2 DM + \beta_3 HTA\,DM + \ldots \qquad [1a]$$

Por otra parte, y en relación con la inclusión de interacciones, hay que tener en cuenta que la inclusión de las mismas puede generar multicolinealidad, tanto más probable cuanto mayor sea el número de interacciones.

Siempre debe considerarse la suficiencia del tamaño muestral para el número de covariables que se desea incluir en el modelo: modelos excesivamente grandes para muestras con tamaños muestrales relativamente pequeños implicarán errores estándar grandes o coeficientes estimados falsamente muy elevados (sobreajuste). En general se recomienda que por cada covariable se cuente con un mínimo de 10 individuos por cada evento de la variable dependiente con menor representación (Peduzzi). Por ejemplo, si la variable dependiente Y es “muerte” y en los datos hay 120 sujetos vivos y 36 sujetos muertos, el evento de Y menos representado es “muerte”, con 36 sujetos; de esta forma el modelo no debería contener más de $36/10 \sim 3$ covariables.

Lo anterior es válido siempre que se trate de covariables cuantitativas o cualitativas con distribuciones bien equilibradas. La situación se complica si una o más de las covariables cualitativas no tiene una distribución

equilibrada (uno de sus dos valores tiene una mínima representación); en ese caso se recomienda que en su tabla de contingencia respecto a la variable dependiente, en cada celda haya un mínimo de 10 observaciones. En el siguiente ejemplo se debería disponer de suficiente tamaño muestral como para que en cada celda haya 10 ó más sujetos (es decir, que tanto a, b, c como d sean mayores de 10).

	$x = 0$	$x = 1$
$y = 0$	a	b
$y = 1$	c	d

4. Concepto de regresión logística

La regresión logística es un instrumento estadístico de análisis bivariado o multivariado, de uso tanto explicativo como predictivo. Resulta útil su empleo cuando se tiene una variable dependiente dicotómica (un atributo cuya ausencia o presencia se ha puntuado con los valores cero y uno, respectivamente) y un conjunto de m variables predictoras o independientes, que pueden ser cuantitativas (que se denominan covariables o covariadas) o categóricas. En este último caso, se requiere que sean transformadas en variables ficticias o simuladas ("*dummy*").

El propósito del análisis es:

- Predecir la probabilidad de que a alguien le ocurra cierto evento: por ejemplo, "estar desempleado" =1 o "no estarlo" = 0; "ser pobre" = 1 o "no ser pobre" = 0; "graduarse como sociólogo" =1 o "no graduarse" = 0;
- Determinar qué variables pesan más para aumentar o disminuir la probabilidad de que a alguien le suceda el evento en cuestión.

Esta asignación de probabilidad de ocurrencia del evento a un cierto sujeto, así como la determinación del peso que cada una de las variables dependientes en esta probabilidad, se basan en las características que presentan los sujetos a los que, efectivamente, les ocurren o no estos sucesos. Por ejemplo, la regresión logística tomará en cuenta los valores que asumen en una serie de variables (edad, sexo, nivel educativo, posición en el hogar, origen migratorio, etc.) los sujetos que están efectivamente desocupados (= 1) y los que no lo están (= 0). En base a ello, predecirá a cada uno de los sujetos – independientemente de su estado real y actual – una determinada probabilidad de ser desocupado (es decir, de tener valor 1 en la variable dependiente). Es decir, si alguien es un joven no amo de casa, con baja educación y de sexo masculino y origen emigrante (aunque esté ocupado) el modelo le predecirá una alta probabilidad de estar desocupado (puesto que la tasa de desempleo de el grupo así definido es alta), generando una variable con esas probabilidades estimadas. Y procederá a clasificarlo como desocupado en una nueva variable, que será el resultado de la predicción. Además, analizará cuál es el peso de cada uno de estas variables independientes en el aumento o la disminución de esa probabilidad. Por ejemplo, cuando aumenta la educación disminuirá en algo la probabilidad de ser desocupado. En cambio, cuando el sexo pase de 0 = "mujer" a 1 = "varón", aumentará en algo la probabilidad de desempleo porque la tasa de desempleo de los jóvenes de sexo masculino es mayor que la de las mujeres jóvenes. El modelo, obviamente, estima los coeficientes de tales cambios.

Cuanto más coincidan los estados pronosticados con los estados reales de los sujetos, mejor ajustará el modelo. Uno de los primeros indicadores de importancia para apreciar el ajuste del modelo logístico es el doble logaritmo del estadístico de verosimilitud (*likelihood)*. Se trata de un estadístico que sigue una distribución similar a χ^2 y compara los valores de la predicción con los valores observados en dos momentos: (a) en el modelo sin variables independientes, sólo con la constante y (b) una vez introducidas las variables predictoras. Por lo tanto, el valor de la verosimilitud debiera disminuir sensiblemente entre ambas instancias e, idealmente, tender a cero cuando el modelo predice bien.

4.1.Regresión logística binaria

Los modelos de regresión logística binaria resultan los de mayor interés ya que la mayor parte de las circunstancias analizadas en medicina responden a este modelo (presencia o no de enfermedad, éxito o fracaso, etc.). Como se ha visto, la variable dependiente será una variable dicotómica que se codificará como 0 ó 1 (respectivamente, "ausencia" y "presencia"). Este aspecto de la codificación de las variables no es banal (influye en la forma en que se realizan los cálculos matemáticos), y habrá que tenerlo muy en

cuenta si se emplean paquetes estadísticos que no recodifican automáticamente las variables cuando éstas se encuentran codificadas de forma diferente (por ejemplo, el uso frecuente de 1 para la presencia y –1 ó 2 para la ausencia).

La ecuación de partida en los modelos de regresión logística es:

$$\Pr(y=1 \mid x) = \frac{\exp\left(b_0 + \sum_{i=1}^{n} b_i x_i\right)}{1 + \exp\left(b_0 + \sum_{i=1}^{n} b_i x_i\right)} \qquad [1b]$$

donde: $\Pr(y=1 \mid X)$ es la probabilidad de que y tome el valor 1 (presencia de la característica estudiada), en presencia de las covariables X;

X es un conjunto de n covariables $\{x_1, x_1, \ldots, x_n\}$ que forman parte del modelo;

b_0 es la constante del modelo o término independiente;

b_i los coeficientes de las covariables.

Es lo que se denomina distribución logística. En la siguiente imagen vemos un ejemplo de esta distribución: la probabilidad de padecer enfermedad coronaria en función de la edad. Como puede verse, la relación entre la variable dependiente (cualitativa dicotómica), y la covariable (edad, cuantitativa continua en este caso), no es definida por una recta (lo que correspondería un modelo lineal), sino que describe una forma sigmoidea (distribución logística).

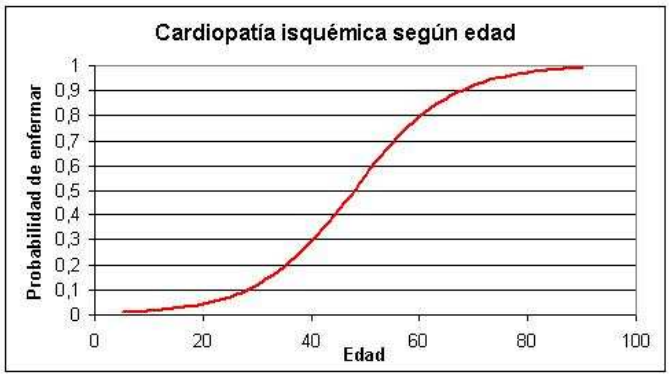

Figura 1.

Si se divide la expresión [1] por su complementario, es decir, si se construye su *odds* (en el ejemplo de presencia o no de enfermedad, la probabilidad de estar enfermo entre la probabilidad de estar sano), se obtiene una expresión de manejo matemático más fácil:

$$\frac{\Pr(y=1 \mid x)}{1 - \Pr(y=1 \mid x)} = \exp\left(b_0 + \sum_{i=1}^{n} b_i x_i\right) \qquad [2]$$

Pero esta expresión aún es difícil de interpretar. Su representación gráfica es (figura 2):

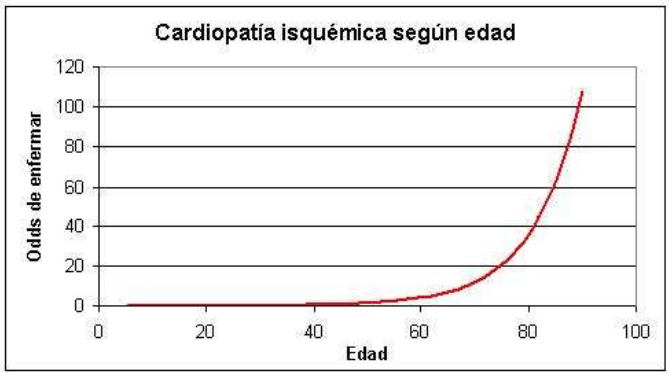

Figura 2

Si ahora se realiza su transformación logaritmo natural, se obtiene una ecuación lineal que lógicamente es de manejo matemático aún más fácil y de mayor comprensión:

$$\log\left(\frac{\Pr(y=1 \mid x)}{1-\Pr(y=1 \mid x)}\right) = b_0 + \sum_{i=1}^{n} b_i x_i \qquad [3]$$

o simplificando:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_i x_i \qquad [3a]$$

En la expresión [3] se ve a la izquierda de la igualdad el llamado *logit*, es decir, el logaritmo natural de la *odds* de la variable dependiente (esto es, el logaritmo de la razón de proporciones de enfermar, de fallecer, de éxito, etc.). El término a la derecha de la igualdad es la expresión de una recta, idéntica a la del modelo general de regresión lineal:

$$y = a + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n \qquad [4]$$

Siguiendo el ejemplo de las figuras 1 y 2, se puede representar el *logit* frente a la edad como se observa en la figura 3:

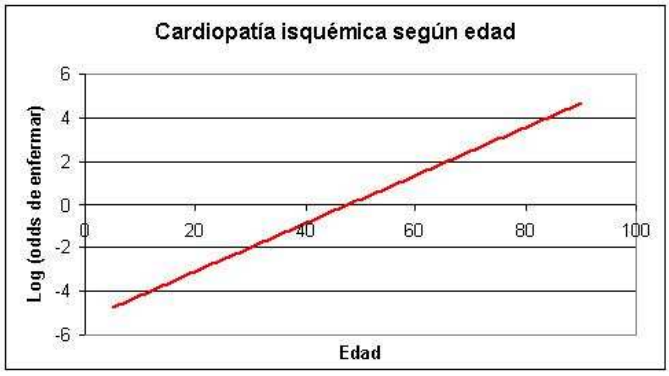

Figura 3.

Pero la regresión lineal presenta una diferencia fundamental respecto al modelo de regresión logística. En el modelo de regresión lineal se asume que los errores estándar de cada coeficiente siguen una distribución normal de media 0 y varianza constante (homoscedasticidad). En el caso del modelo de regresión logística no pueden realizarse estas asunciones pues la variable dependiente no es continua (sólo puede tomar dos valores, 0 ó 1, pero ningún valor intermedio). Llamando ε al posible error de predicción para cada covariable x_i se tendrá que el error cometido dependerá del valor que llegue a tomar la variable dependiente, tal como se ve en [5].

$$y = \Pr(x) + \varepsilon \qquad \begin{cases} y=1 \Rightarrow \varepsilon = 1-\Pr(x) \\ y=0 \Rightarrow \varepsilon = -\Pr(x) \end{cases} \qquad [5]$$

Esto implica que ε sigue una distribución binomial, con media y varianza proporcionales al tamaño muestral y a $\Pr\left(y=1 \mid x_i\right)$ (la probabilidad de que $y=1$ dada la presencia de x_i).

5. Requisitos y etapas de la regresión logística

- Recodificar las variables independientes categóricas u ordinales en variables ficticias o simuladas y de la variable dependientes en 0 y 1;
- Evaluar efectos de confusión y de interacción del modelo explicativo;
- Evaluar la bondad de ajuste de los modelos;
- Analizar la fuerza, sentido y significación de los coeficientes, sus exponenciales y estadísticos de prueba (Wald).

6. Estimación de los coeficientes del modelo y de sus errores estándar

Para la estimación de los coeficientes del modelo y de sus errores estándar se recurre al cálculo de estimaciones de máxima verosimilitud, es decir, estimaciones que hagan máxima la probabilidad de obtener los valores de la variable dependiente Y proporcionados por los datos de nuestra muestra. Estas estimaciones no son de cálculo directo, como ocurre en el caso de las estimaciones de los coeficientes de regresión de la regresión lineal múltiple por el método de los mínimos cuadrados. Para el cálculo de estimaciones máximo–verosímiles se recurre a métodos iterativos, como el método de Newton–Raphson. Dado que el cálculo es complejo, normalmente hay que recurrir al uso de rutinas de programación o a paquetes estadísticos. De estos métodos surgen no sólo las estimaciones de los coeficientes de regresión, sino también de sus errores estándar y de las covarianzas entre las covariables del modelo.

El siguiente paso será comprobar la significación estadística de cada uno de los coeficientes de regresión en el modelo. Para ello se pueden emplear básicamente tres métodos: el estadístico de Wald, el estadístico G de razón de verosimilitud y la prueba Score.

6.1. El estadístico de Wald

Contrasta la hipótesis de que un coeficiente aislado es distinto de 0, y sigue una distribución normal de media 0 y varianza 1. Su valor para un coeficiente concreto viene dado por el cociente entre el valor del coeficiente y su correspondiente error estándar. La obtención de significación indica que dicho coeficiente es diferente de 0 y merece la pena su conservación en el modelo. En modelos con errores estándar grandes, el estadístico de Wald puede proporcional falsas ausencias de significación (es decir, se incrementa el error tipo II). Tampoco es recomendable su uso si se están empleando variables de diseño.

6.2. El estadístico G de razón de verosimilitud

Se trata de ir contrastando cada modelo que surge de eliminar de forma aislada cada una de las covariables frente al modelo completo. En este caso cada estadístico G sigue una χ^2 con un grado de libertad (no se asume normalidad). La ausencia de significación implica que el modelo sin la covariable no empeora respecto al modelo completo (es decir, da igual su presencia o su ausencia), por lo que según la estrategia de obtención del modelo más reducido (principio de parsimonia), dicha covariable debe ser eliminada del modelo ya que no aporta nada al mismo. Esta prueba no asume ninguna distribución concreta, por lo que es la más recomendada para estudiar la significación de los coeficientes.

6.3. La prueba Score

Su cálculo para el caso de una única variable viene dado por:

$$S = \frac{\sum_{i=1}^{n} x_i \left(y_i - \overline{y}\right)}{\sqrt{\overline{y}\left(1-\overline{y}\right)\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}} \qquad [6]$$

En el caso de múltiples covariables hay que utilizar cálculo matricial, si bien no requiere un cálculo iterativo (precisamente su rapidez de cálculo sería su aspecto más favorable). En contra del mismo dos aspectos:

(a) Se sabe que este estadístico se incrementa conforme aumenta el número de covariables (es decir tiende a dar significación con mayor frecuencia).

(b) Este estadístico también asume una distribución normal con media 0 y varianza 1.

Al igual que en los casos anteriores, si alcanza significación indica que la covariable debería permanecer en el modelo. Su uso en algunos paquetes estadísticos ha quedado relegado a la selección de variables en métodos paso a paso (por la mayor rapidez de cálculo).

Cuando la covariable es cualitativa con n categorías (siendo $n>2$), en el modelo se analizará la significación de cada una de sus $n-1$ variables ficticias, así como la significación global de la covariable comparando la presencia en bloque frente a la ausencia en bloque de sus $n-1$ covariables ficticias.

En el siguiente ejemplo, tomado de Hosmer y realizado con SPSS®, se analiza la variable edad (AGE) y la variable IVHX (usuario de drogas por vía parenteral); esta segunda era una variable con tres categorías ("nunca", "previa" y "reciente"), por lo que se crearon dos variables ficticias: IVHX(1) e IVHX(2); el resultado es una estimación de los β con sus errores estándar, la significación para IVHX(1) e IVHX(2), y la significación de IVHX considerada como la entrada frente a la salida en bloque del modelo de IVHX(1) e IVHX(2).

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1[a]	AGE	,045	,017	7,216	1	,007	1,046
	IVHX			18,618	2	,000	
	IVHX(1)	-,681	,279	5,952	1	,015	,506
	IVHX(2)	-1,002	,236	18,081	1	,000	,367
	Constante	-2,033	,527	14,909	1	,000	,131

a. Variable(s) introducida(s) en el paso 1: AGE, IVHX.

Figura 4.

Una vez se ha estimado los coeficientes de regresión y sus correspondientes errores estándar se calcularán los correspondientes intervalos de confianza para las estimaciones, bajo la hipótesis de que dichos coeficientes se distribuyen según respectivas distribuciones normales. Para un determinado coeficiente, su intervalo de confianza al 95 % vendrá dado por:

$$IC\,95\%(\beta) = \left[(\beta - 1{,}96\,ee)\,,\,(\beta + 1{,}96\,ee)\right]$$
$$IC\,95\%(OR) = \left[\exp(\beta - 1{,}96\,ee)\,,\,\exp(\beta + 1{,}96\,ee)\right] \qquad [7]$$

Junto a la significación del estadístico empleado para contrastar la significación de los coeficientes de regresión, la inclusión de la unidad en el intervalo de confianza es, lógicamente, indicativa de la ausencia de significación.

En ocasiones existirán modelos que llaman la atención por la falta de sentido de sus estimaciones. Esta sorpresa suele venir dada por la presencia de estimaciones de grandes errores estándar, con frecuencia asociadas a estimaciones de coeficientes de regresión también anormalmente elevados. Las posibles causas de este hecho pueden ser:

(a) Presencia de una frecuencia cero en una tabla de contingencia Y x X. Cuando esto ocurre provoca en el cálculo de la correspondiente *odds* la presencia de un 0 en el denominador (y por tanto no es calculable). Si esta covariable se intenta introducir en el modelo de regresión que se está diseñando, el software puede comportarse de forma incorrecta: desde excluirla por entender que predice perfectamente la variable dependiente, a incluirla y comunicar un error (porque la rutina de iteración para el cálculo de estimaciones de máxima verosimilitud o bien no llega a converger o bien llega al máximo de iteraciones prefijadas). Esta circunstancia puede y debe ser detectada durante el análisis univariado. En el caso de tratarse de una variable cualitativa con más de dos categorías, una solución es colapsar dos de esas categorías.

(b) También puede ocurrir que se incluyan interacciones que impliquen una excesiva estratificación para la muestra disponible. El resultado puede ser una estimación elevada del correspondiente coeficiente de

regresión y de su error estándar. En el análisis univariado, al realizar efectivamente las dos tablas de contingencia de la estratificación, se observará que alguna de las ocho celdas contiene el cero. Si no puede recurrir al colapso de categorías, puede decidirse diseñar una nueva variable que sea la combinación de las dos covariables con sus correspondientes categorías, e incluirla como tal en el modelo.

Presencia de una o más covariables que discriminan perfectamente las dos categorías de la variable dependiente. Algunos ejemplos servirán para explicar esta circunstancia: Si siempre que se administran antimicrobianos los sujetos con una determinada enfermedad infecciosa viven y siempre que no se administran mueren, la covariable "antimicrobianos" discrimina perfectamente a la variable "muerte"; o si siempre que se tienen más de 65 años se padece de cardiopatía isquémica y por debajo no, la covariable "edad" discrimina perfectamente a la variable "cardiopatía isquémica". En la práctica esta circunstancia impide que se puedan realizar estimaciones de coeficientes por máxima verosimilitud, lo que no quiere decir que el paquete estadístico necesariamente no de falsas estimaciones, como en el punto anterior.

Este problema está en estrecha relación con el tamaño muestral y el número de covariables que se desean introducir en el modelo: la probabilidad de discriminación completa es elevada en los modelos con muestras con tamaños muestrales pequeños, sobre todo cuando una de las categorías de la variable dependiente está poco representada, y tanto más cuanto mayor es el número de covariables introducidas en el modelo.

(c) Multicolinealidad. Si bien existen pruebas que permiten comprobar la existencia de colinealidad entre covariables (que se verá más adelante), cabe reseñar aquí que al igual que en los casos anteriores, los modelos con multicolinealidad entre las covariables introducidas llamarán la atención por la presencia de grandes errores estándar, y frecuentemente, estimaciones de coeficientes anormalmente elevadas. Sin embargo la multicolinealidad no afecta al sentido de las estimaciones (la multicolinealidad no hará que aparezca significación donde no la hay, y viceversa).

7. Multicolinealidad

Se dice que existe multicolinealidad cuando dos o más de las covariables del modelo mantienen una relación lineal. Cuando la colinealidad es perfecta, es decir, cuando una covariable puede determinarse según una ecuación lineal de una o más de las restantes covariables, es posible estimar un único coeficiente de todas las covariables implicadas. En estos casos debe eliminarse la covariable que actúa como dependiente.

Normalmente lo que se hallará será una multicolinealidad moderada, es decir, una mínima correlación entre covariables. Si esta correlación fuera de mayor importancia, su efecto sería, como ya se vio anteriormente, el incremento exagerado de los errores estándar, y en ocasiones, del valor estimado para los coeficientes de regresión, lo que hace las estimaciones poco creíbles.

Un primer paso para analizar este aspecto puede ser examinar la matriz de coeficientes de correlación entre las covariables. Coeficientes de correlación muy elevados llevarán a investigar con mayor profundidad. Sin embargo, este método, bueno para detectar colinealidad entre dos covariables, puede conducir a no poder detectar multicolinealidad entre más de dos de ellas.

Existen otros procedimientos analíticos para detectar multicolinealidad. Puede desentenderse por el momento de la variable dependiente y realizar sendos modelos en los que una de las covariables actuará como variable dependiente y las restantes covariables como variables independientes de aquella. A cada uno de estos modelos se le puede calcular la R^2 (o dispersión total, medida de ajuste que se verá más adelante). Se denomina tolerancia al complementario de R^2, $\left(1-R^2\right)$, y factor de inflación de la varianza (*FIV*) al inverso de la tolerancia, $1/\left(1-R^2\right)$. Cuando existe estrecha relación entre covariables la tolerancia tiende a ser 0, y por tanto *FIV* tiende al infinito. Como regla general debería preocupar tolerancias menores de 0,1 y *FIV* mayores de 10. SPSS ofrece la matriz de correlaciones, pero no aporta índices de multicolinealidad para la regresión logística.

La solución a la multicolinealidad no es fácil:

- Puede intentarse eliminar la variable menos necesaria implicada en la colinealidad, a riesgo de obtener un modelo menos válido;
- Puede intentar cambiarse la escala de medida de la variable en conflicto (es decir, transformarla), para evitar sacarla del modelo, si bien no siempre se encontrará una transformación de forma directa. Algunas transformaciones frecuentes son el centrado respecto de la media, la estandarización o la creación de variables sintéticas mediante un análisis previo de componentes principales (que es otro

tipo de análisis multivariado). Estas transformaciones por el contrario hacen al modelo muy dependiente de los datos actuales, invalidando su capacidad predictiva;

- También se puede recurrir a aumentar la muestra para así aumentar la información en el modelo, lo que no siempre será posible.

8. Las variables simuladas (*dummy)*

A veces se necesita incorporar al modelo de regresión logística variables independientes que no son numéricas sino categóricas. Supóngase, por ejemplo, que se quiere predecir la probabilidad de una persona de "ser pobre".

Tal vez resulte importante incorporar variables que no son cuantitativas: por ejemplo, la categoría ocupacional ("empleador", "trabajador por cuenta propia", "asalariado" "trabajador sin remuneración"). En este caso, esta variable podría ser incorporada a la ecuación si se la transforma en una variable simulada. Ello consiste en generar $n-1$ variables dicotómicas con valores cero y uno, siendo n el número de categorías de la variable original. Para el caso de la variable categoría ocupacional, la transformación sería la siguiente:

categoría ocupacional	variable fictícia		
	empleador	cuenta propia	asalariado
empleador	1	0	0
cuenta propia	0	1	0
asalariado	0	0	1
trabajador sin remuneración	0	0	0

Se crearían tres variables dicotómicas: la primera de ellas sería "empleador". Quien lo sea tendrá valor 1 en esa variable y valor cero en las variables "cuenta propia" y "asalariado". Los "por cuenta propia" tendrán valor 1 en la segunda variable y cero en las otras, etc. No se necesita crear, en cambio, una variable llamada "trabajador sin remuneración": lo será quien tenga valores cero en las tres anteriores. Esta última es la categoría "base" de las variables simuladas.

Una vez realizada esta transformación, estas variables pueden ser incorporadas en una ecuación de regresión: sus valores sólo pueden variar entre cero y uno y sus coeficientes b indicarán, en cada caso, cuanto aumentan o disminuyen los "*odds*" de probabilidad del evento que se procura predecir cuando una de estas variables pasa de cero a uno (por ejemplo, cuando alguien es un empleador, seguramente la probabilidad de que sea pobre disminuirá, lo que se expresará en un coeficiente b negativo en la ecuación logística).

9. Función de verosimilitud

Se sabe que cualquier variable dependiente de otra u otras variables, toma valores según los valores de las variables de las que depende. Por otra parte, esa variable dependiente irá tomando valores siguiendo o describiendo una determinada distribución de frecuencias; es decir, tomen los valores que tomen las variables independientes, si el experimento se repite múltiples veces, la variable dependiente tomará para esos valores de las independientes un determinado valor, y la probabilidad de ocurrencia de dicho valor vendrá dado por una distribución de frecuencias concreta: una distribución normal, una distribución binomial, una distribución hipergeométrica, etc. En el caso de una variable dependiente dicotómica (como el caso que nos ocupa), la distribución de frecuencias que seguirá será la binomial, que depende de la tasa de éxitos (X sujetos de un total de N , que sería el elemento variable), para un determinado tamaño muestral N y probabilidad $\Pr(\bullet)$ de ocurrencia del evento valorado por la variable dependiente (parámetros constantes). La función de densidad de esta distribución de frecuencias vendrá dada por la siguiente expresión:

$$\Pr(y) \approx f(x) = \binom{N}{x} p^x (1-p)^{N-x} \qquad [8]$$

Si en la expresión anterior introducimos los datos concretos de nuestra muestra de N sujetos (es decir, se convierte el elemento variable X en parámetro), y se hace depender el resultado de la función de densidad del parámetro "probabilidad de ocurrencia" (p , que de esta forma se convierte en variable), se está generando su función de verosimilitud, $f(p \mid x)$ (función dependiente de p dado el valor muestral de x) o $L(p)$ (L de *likelihood*), que ofrece como resultados las probabilidades de la función de densidad ajustada a los datos:

$$f(p \mid x) = \binom{N}{x} p^{x} (1-p)^{N-x} \qquad [9]$$

Se deduce que, para una muestra concreta, esa probabilidad será diferente según qué valores tome el parámetro "probabilidad de ocurrencia":

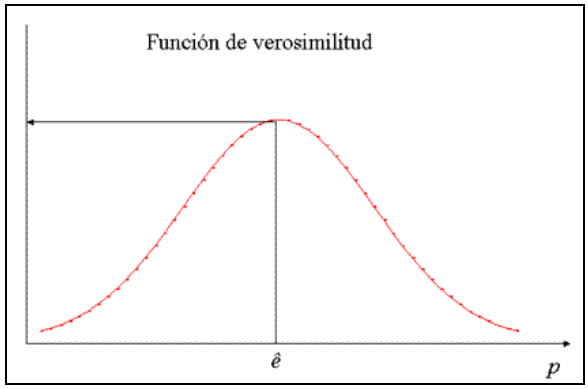

Figura 5.

Se demuestra que la mejor estimación del parámetro $\hat{e}$ es aquel valor que maximiza esta función de verosimilitud, ya que son estimadores consistentes (conforme crece el tamaño muestral, la estimación se aproxima al parámetro desconocido), suficientes (aprovechan la información de toda la muestra), asintóticamente normales y asintóticamente eficientes (con mínima varianza), si bien no siempre son insesgados (no siempre la media de las estimaciones para diferentes muestras tenderá hacia el parámetro desconocido).

9.1. Método de Newton-Raphson

Se trata de un método iterativo, empleado en diversos problemas matemáticos, como en la determinación de las raíces de ecuaciones, y en el presente caso, en la estimación de los coeficientes de regresión β por el procedimiento de máxima verosimilitud.

Por facilidad de cálculo toda la formulación se expresará en forma de matrices. Las particularidades del cálculo matricial escapan del ámbito de este documento. Téngase presente la base de datos (una tabla con filas y columnas). Se dispondrá de:

- Una variable Y , que es la variable dependiente. Expresada como matriz será una matriz de N filas y una columna, cuyo contenido será de ceros y unos (ya que se trata de una variable dicotómica).

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{pmatrix} \qquad [10]$$

- Un conjunto de M covariables, que pueden expresarse como una matriz de N filas y M columnas. Sin embargo, dado que el modelo contiene una constante, ésta se puede expresar como una columna

adicional en la que todos sus elementos son "1". Por tanto la matriz X queda como una matriz con N filas y $(M+1)$ columnas, de la forma:

$$X = \begin{pmatrix} 1 & x_{1,2} & \dots & x_{1,m+1} \\ 1 & x_{2,2} & \dots & x_{2,m+1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,2} & \dots & x_{n,m+1} \end{pmatrix} \qquad [11]$$

- Y por último un conjunto de coeficientes de regresión β, uno para cada covariable, incluida la covariable creada para la constante, con una fila y $(M+1)$ columnas

$$\beta = (\beta_1, \beta_2, \dots, \beta_{m+1}) \qquad [12]$$

El proceso se inicia construyendo la función de verosimilitud (*likelihood function*) de la ecuación de regresión logística:

$$L(\beta) = p_i^{\sum y_i} (1-p_i)^{\left(N-\sum y_i\right)} \qquad [13]$$

o mejor, su transformación logarítmica (*log likelihood*):

$$LL(\beta) = \sum y_i \ln(p_i) + \left(N - \sum y_i\right) \ln(1-p_i) \qquad [14]$$

donde p_i es la probabilidad de ocurrencia de $y=1$ con los valores muestrales de las covariables $X = \{x_1, x_2, \dots, x_{m+1}\}$, para el sujeto $i = \{1, 2, \dots, N\}$. El valor $2LL(\beta)$ se llama devianza y mide en qué grado el modelo se ajusta a los datos: cuanto menor sea su valor, mejor es el ajuste.

Se trata de conocer aquellos valores de β que hacen máxima la función de verosimilitud (o su logaritmo). Se sabe que si se iguala a cero la derivada parcial de una función respecto a un parámetro, el resultado es unos valores de dicho parámetro que hacen llevar a la función a un valor máximo o un valor mínimo (un punto de inflexión de la curva). Para confirmar que se trata de un máximo y no de un mínimo, la segunda derivada de la función respecto a dicho parámetro debe ser menor de cero.

La primera derivada de $LL(\beta)$ respecto de β (llamada función *score*) en su forma matricial es:

$$U(\beta) = \frac{\partial LL(\beta)}{\partial \beta} = X'(Y-p) \qquad [15]$$

donde: p es una matriz de N filas y una columna que contiene las probabilidades de cada individuo de que tengan su correspondiente evento y_i

La segunda derivada, llamada matriz informativa o hessiana, es:

$$H(\beta) = \frac{\partial^2 LL(\beta)}{\partial \beta \, \partial \beta} = -\mathbf{X}'\mathbf{W}\mathbf{X} \qquad [16]$$

donde: $\mathbf{W}$ es una matriz diagonal[1] de n filas y n columnas, en la que los elementos de su diagonal vienen dados por los respectivos productos $p_i(1-p_i)$, de manera que $\mathbf{W}$ queda de la forma siguiente:

[1] Matriz cuadrada en la que todos sus elementos son 0 excepto su diagonal

$$\mathbf{W} = \begin{pmatrix} p_1(1-p_1) & 0 & \cdots & 0 \\ 0 & p_2(1-p_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p_n(1-p_n) \end{pmatrix} \qquad [17]$$

y para cada fila su p_i es:

$$p_i = \frac{1}{1+\exp\left(-\sum_{j=1}^{m+1}\beta_j\, x_{ij}\right)} \qquad [18]$$

Una vez se dispone todos los elementos necesarios, se procede a explicar como tal el método iterativo para la determinación de los coeficientes de regresión.

Se le asigna un valor inicial empírico a los coeficientes de regresión, en general cero a todos ellos

En cada iteración t la matriz de nuevos coeficientes de regresión experimentales resulta de sumar matricialmente un gradiente a la matriz de coeficientes experimentales del paso anterior. Este gradiente es el resultado del cociente entre la primera derivada y la segunda derivada de la función de verosimilitud de la ecuación de regresión.

$$\hat{\beta}_t = \hat{\beta}_{t-1} + \left(\mathbf{X}'\mathbf{W}_{\mathbf{t\text{-}1}}\mathbf{X}\right)' \mathbf{X}'\left(\mathbf{Y} - p_{t-1}\right) \qquad [19]$$

El segundo paso se repite tantas veces como sea necesario hasta que la diferencia entre la matriz de coeficientes de regresión en dicha iteración y la matriz de la iteración previa, sea 0 o prácticamente 0 (por ejemplo $<10^{-6}$). Los paquetes estadísticos suelen tener un límite de iteraciones que pueden modificarse si no se obtiene convergencia inicialmente. SPSS® tiene además otras condiciones de parada:

- $LL(\beta)$ muy cercana a cero;
- Diferencia entre $LL(\beta)$ de dos iteraciones consecutivas muy cercana a cero.

Una vez finalizadas las iteraciones, la inversa de la matriz informativa de la última iteración ofrece los valores de varianzas y covarianzas de las estimaciones de los coeficientes de regresión estimados. En concreto, el error estándar de cada coeficiente de regresión coincide con la raíz cuadrada del elemento respectivo de la diagonal principal (es decir el elemento (1,1) sería el cuadrado del error estándar del coeficiente β_1, el elemento (2,2) el cuadrado del error estándar del coeficiente β_2, y así sucesivamente). Por debajo de esta diagonal quedan las covarianzas de cada pareja de covariables (es decir, el elemento (2,1) es la covarianza de β_1 y β_2, el elemento (3,2) es la covarianza de β_2 y β_3, etc.). Hay programas estadísticos que ofrecen esta matriz de varianzas y covarianzas; SPSS® no lo hace, sino que ofrece la matriz de correlaciones. En ese caso se podrá calcular la matriz de varianzas y covarianzas sabiendo que la covarianza de dos variables es igual al producto del coeficiente de correlación de ambas (r) y los dos respectivos errores estándar:

$$\mathrm{cov}(\beta_1, \beta_2) = r(\beta_1, \beta_2)\, ee(\beta_1)\, ee(\beta_2) \qquad [20]$$

Entender esta formulación y el algoritmo de las iteraciones puede ser de gran utilidad, pues con conocimientos básicos de programación facilita el desarrollo de rutinas propias, por ejemplo en *VisualBasic*® dentro de una base de datos de *Access*®, que pueden liberar de la dependencia de costosos paquetes estadísticos.

Odds ratio: Es un cociente de proporciones de enfermos por cada sano entre el grupo con un factor de riesgo y el grupo sin dicho factor de riesgo. Supóngase el siguiente ejemplo:

		factor de riesgo		
		si	no	
enfermedad	si	20	30	50
	no	80	270	359
		100	300	400

En este caso, entre los que tienen el factor de riesgo hay 20 enfermos por cada 80 sanos (0,25), y entre los que no tienen el factor de riesgo hay 30 enfermos por cada 270 sanos (0,11), por lo que las personas con el factor de riesgo tienen un riesgo de enfermar 2,25 veces superior (0,25/0,11) que las personas sin el factor de riesgo.

Principio jerárquico: siempre que se incluya en el modelo un término de interacción, el modelo debe incluir también todos los términos de orden inferior, y si el término de interacción resultase significativo y permaneciese en el modelo, también deberían permanecer los términos de orden inferior, aunque no se lograra demostrar significación para ellos.

Modelo con interacción de primer orden:

$$y = a + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2$$

Modelo con interacción de segundo orden:

$$y = a + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_1 x_2 + b_5 x_1 x_3 + b_6 x_2 x_3 + b_7 x_1 x_2 x_3$$

Principio de parsimonia: En igualdad de condiciones la solución más sencilla que explique completamente un problema es probablemente la correcta (Guillermo de Ockham). Según este principio, cuando más de un modelo se ajuste a las observaciones, se retendrá el modelo más simple que explique dichas observaciones con un grado adecuado de precisión.

10. Tipos de modelos de regresión logística

Modelo logístico univariante simple

$$y = \frac{\exp(a + b x)}{1 + \exp(a + b x)}$$

Modelo logístico univariante múltiple

$$y = \frac{\exp(\beta_{i,0} + \beta_1 x_1 + \beta_{i,2} x_2 + \ldots + \beta_{i,k} x_k)}{1 + \exp(\beta_{i,0} + \beta_1 x_1 + \beta_{i,2} x_2 + \ldots + \beta_{i,k} x_k)}$$

Modelo logístico multivariante simple

$$y_i = \frac{\exp(a + b x)}{1 + \exp(a + b x)}$$

Modelo logístico multivariante múltiple

$$y_i = \frac{\exp(\beta_{i,0} + \beta_1 x_1 + \beta_{i,2} x_2 + \ldots + \beta_{i,k} x_k)}{1 + \exp(\beta_{i,0} + \beta_1 x_1 + \beta_{i,2} x_2 + \ldots + \beta_{i,k} x_k)}$$

11. Interpretación del modelo logístico

Los parámetros del modelo son: β_0, la ordenada en el origen, y $\boldsymbol{\beta}_\mathbf{i} = \{\beta_1, \beta_2, \ldots, \beta_k\}$. A veces, se utilizan también como parámetros $\exp(\beta_0)$ y $\exp(\boldsymbol{\beta}_\mathbf{i})$, que se denominan *odds ratios* o razón de probabilidades. Estos valores indican cuánto se modifican las probabilidades por unidad de cambio en las variables x. De [3a] se deduce que:

$$O_i = \frac{p_i}{1-p_i} = \exp(\beta_0)\prod_{j=1}^{k}\exp(\beta_j)^{x_j}$$

Supóngase que dos elementos tienen valores iguales en todas las variables menos en una.

Sean $(x_{i,1}, x_{i,2}, \ldots, x_{i,h}, \ldots, x_{i,k})$ los valores de las variables para el primer elemento y $(x_{j,1}, x_{j,2}, \ldots, x_{j,h}, \ldots, x_{j,k})$ para el segundo, y todas las variables son las mismas en ambos elementos menos en la variable h donde $x_{i,h} = x_{j,h} + 1$. Entonces, el *odds ratio* para estas dos observaciones es:

$$\frac{O_i}{O_j} = \exp(\beta_h)$$

e indica cuánto se modifica el *ratio* de probabilidades cuando la variable x_j aumenta en una unidad.

Si se considera $p_i = 0,5$ en el modelo *logit*, entonces

$$\log\frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_k x_{k,i} = 0$$

es decir,

$$x_{1,i} = -\frac{\beta_0}{\beta_1} - \sum_{j=2}^{k}\frac{\beta_j x_{i,j}}{\beta_1}$$

donde $x_{1,i}$ representa el valor de x_1 que hace igualmente probable que un elemento cuyas restantes variables son $(x_{2,i}, \ldots x_{k,i})$ pertenezca a la primera o la segunda población.

12. Medidas de confiabilidad del modelo

12.1. Devianza

Es similar a la suma de cuadrados del error de la regresión lineal y se define como:

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

Si D es mayor que una χ^2 con $n-p$ grados de libertad para un nivel de significación dado entonces el modelo logístico es confiable.

12.2. Criterio AIC de Akaike

Se define como

$$AIC = D + 2(p+1)$$

donde p es el número de variables predictoras.

12.3. Prueba de bondad de ajuste de Hosmer-Lemeshov

Se define como

$$c = \sum_{i=1}^{g}\frac{(O_i - n'_i\,\overline{p}_i)^2}{n'_i\,\overline{p}_i\,(1-\overline{p}_i)}$$

donde g es el número de grupos;

n_i' es el numero de observaciones en el *i*–ésimo grupo;

O_i es la suma de las Y en el *i*–ésimo grupo; y

$\overline{p}_i$ es el promedio de las p_i en el *i*–ésimo grupo.

13. Estadísticos influenciales para regresión logística

Existen varios tipos de residuales que permiten cotejar si una observación es influencial o no.

13.1. Residuales de Pearson

Definidos como:

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

donde y_i representa el número de veces que $y=1$ entre las m_i repeticiones de X_i si los valores de la variable de respuesta están agrupadas,

El residual de Pearson es similar al residual estudentizado usado en regresión lineal. Así, un residual de Pearson mayor que 2 indica un dato anormal.

Si el modelo es correcto, los residuales de Pearson serán variables de media cero y varianza unidad que pueden servir para hacer el diagnóstico de dicho modelo. El estadístico $\chi_0^2 = \sum_{i=1}^{k} e_i^2$ permite realizar un contraste global de la bondad del ajuste. Se distribuye asintóticamente como una χ^2 con $\left(n-k-1\right)$ grados de libertad, donde $k+1$ es el número de parámetros en el modelo.

En lugar de los residuos de Pearson se pueden utilizar, también, las desviaciones o pseudoresiduos definidos por:

$$d_i = -2\left(y_i \log \hat{p}_i + \left(1-y_i\right)\log\left(1-\hat{p}_i\right)\right)$$

13.2. Residuales de devianza

Definidos como:

$$D_i = -2\,sign\left(y_i - m_i\,p_i\right)\sqrt{y_i \log\frac{m_i\,p_i}{y_i} + \left(m_i - y_i\right)\log\frac{m_i\left(1-\hat{p}_i\right)}{m_i - y_i}}$$

Si la devianza es mayor que 4 entonces la observación correspondiente es anormal.

13.3. Uso de la regresión logística en clasificación

Para efectos de clasificación la manera más fácil de discriminar es considerar que si $p>0,5$ entonces la observación pertenece a la clase que interesa. Pero algunas veces esto puede resultar injusto sobre todo si se conoce si una de las clases es menos frecuente que la otra.

Métodos alternativos son:

(a) Representar gráficamente el porcentaje de observaciones que poseen el evento (o sean que pertenecen al grupo (1) y que han sido correctamente clasificadas (sensibilidad) frente a distintos niveles de probabilidad y el porcentaje de observaciones de la otra clase que han sido correctamente clasificadas (especificidad) frente a los mismos niveles de probabilidad anteriormente usados, en la misma gráfica. La probabilidad que se usará para clasificar las observaciones se obtienen cortando las dos curvas.

(b) Usar la curva ROC (*receiver operating characteristic*). En este caso se representa gráficamente la sensibilidad frente a (1-especifidad)100 %, y se escoge como el p ideal aquel que está más cerca a la esquina superior izquierda, o sea al punto (100 , 0).

14. Diagnóstico en regresión logística

Verificar que el modelo es adecuado, (bondad de ajuste):

- Con datos agrupados: deviancia residual;
- Con datos individuales hace falta una referencia, que puede obtenerse a partir del modelo saturado, siempre que se trabaje con pocas variables y éste sea estimable;
- Otros estadísticos:
  - $\frac{\sum(O-E)^2}{E}$ sobre cada observación;
  - Hosmer y Lemeshow: $\frac{\sum(O-E)^2}{E}$ sobre 10 categorías de p.

15. Modelos predictivos

El objetivo del modelo puede ser:

- Generar una ecuación con capacidad predictiva, como una clasificación (análisis discriminante);
- Buscar qué factores tienen capacidad predictiva

Si la respuesta es la aparición de un evento, pueden llamarse modelos pronósticos. En este tipo de estudios es típico contar con un gran número de variables a explorar.

16. Métodos de selección automática

16.1. Hacia adelante

1. Se inicia con un modelo vacío (sólo α);
2. Se ajusta un modelo y se calcula el p valor de incluir cada variable por separado;
3. Se selecciona el modelo con la más significativa;
4. Se ajusta un modelo con la(s) variable(s) seleccionada(s) y se calcula el p valor de añadir cada variable no seleccionada por separado;
5. Se selecciona el modelo con la más significativa;
6. Se repite 4 — 5 hasta que no queden variables significativas para incluir.

16.2. Hacia atrás

1. Se inicia con un modelo con TODAS las variables candidatas;
2. Se eliminan, una a una, cada variable y se calcula la pérdida de ajuste al eliminar;
3. Se selecciona para eliminar la menos significativa;
4. Se repite 2 – 3 hasta que todas las variables incluidas sean significativas y no pueda eliminarse ninguna sin que se pierda ajuste.

16.3. *Stepwise*

- Se combinan los métodos adelante y atrás;
- Puede empezarse por el modelo vacío o por el completo, pero en cada paso se exploran las variables incluidas, por si deben salir y las no seleccionadas, por si deben entrar;
- No todos los métodos llegan a la misma solución necesariamente;

16.4. Consideraciones

- Criterio exclusivamente estadístico: no se tienen en cuenta otros "conocimientos" sobre las variables más interesantes a incluir (aunque se puede forzar a que algunas variables siempre estén en el modelo);
- Si hay un conjunto de variables muy correlacionadas, sólo una será seleccionada;
- No es fácil tener en cuenta interacciones entre variables (los modelos deben ser jerárquicos).

17. Valoración de la capacidad predictiva del modelo

- El modelo permite calcular una predicción del resultado en escala de probabilidad;
- Puede decidirse clasificar un individuo en el grupo de sucesos si su probabilidad supera un valor π:

$$clasificación = \begin{cases} \Pr > \pi \Rightarrow y_e = 1 \\ \Pr \leq \pi \Rightarrow y_e = 0 \end{cases}$$

17.1.Clasificación

		realidad y_0	
		1	**0**
modelo y_e	**1**	VP	FP
	0	FN	VN

- Sensibilidad = VP/(VP+FN)
- Especificidad = VN/(VN+FP)
- Area bajo la curva ROC construida para todos los posibles puntos de corte de π para clasificar los individuos:

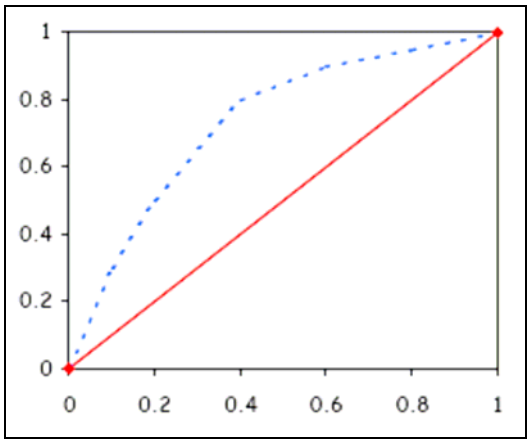

Figura 8.

17.2. Cálculo del área bajo la curva ROC

- Guardar los valores que predice el modelo (esperados)
- Calcular la U de Mann-Whitney respecto a los esperados:

$$AUC = 1 - \frac{U}{n_1 \, n_0}$$

donde n_1 y n_0 son respectivamente el número esperado de "1" y "0".

Test Statistics[a]

	Predicted Value
Mann-Whitney U	26273.500
Wilcoxon W	69933.500
Z	-7.866
Asymp. Sig. (2-tailed)	.000

a. Grouping Variable: GROUP

GROUP

		Frequency	Percent
Valid	control	295	50.8
	caso	286	49.2
	Total	581	100.0

Figura 9.

$$AUC = 1 - \frac{U}{n_1\, n_0} = 1 - \frac{26273}{295 \times 286} = 0,69$$

Un $AUC = 0,5$ corresponde a una capacidad predictiva nula. El máximo es 1.

17.3. Elección del punto de corte óptimo

- Debe optimizarse la sensibilidad y la especificidad, y elegir un punto según la naturaleza del modelo predictivo
- El cambio en el punto de corte corresponde a emplear diferentes constantes en el modelo logístico
- Con frecuencia la constante estimada, α, consigue una sensibilidad y especificidad máxima, pero puede no ser el caso.

18. Validación del modelo

- El cálculo de la capacidad predictiva (CP) del modelo sobre la misma muestra que lo generó siempre es optimista, y debe validarse;
- Diferentes estrategias:
  - Probar el modelo en otra muestra diferente;
  - Elaborar el modelo con un 75 % de la muestra y calcular la CP en el 25 % restante;
  - Usar la misma muestra, pero calcular los indicadores de CP mediante técnicas de bootstrap o validación cruzada, que corrigen el “optimismo”.

19. Regresión logística condicional

- Estudios con datos apareados;
- Generalización (modelo) del test de McNemar;
- Las observaciones no son independientes. Si se ignora, la correlación (positiva) entre las observaciones genera un sesgo hacia la hipótesis nula;
- Se pierde poder estadístico.

20. Estudios de casos y controles apareados

- Para cada caso se elige uno o varios controles que son idénticos (o muy parecidos) al caso en variables que se quiere controlar forzosamente: sexo, edad, residencia, fecha de diagnóstico;
- Estas variables quedarán igualadas entre casos y controles: no se podrá estimar un efecto;
- Debe controlarse la correlación que generan: modelo condicional;
- Cada caso se compara exclusivamente con sus controles;
- Las parejas (caso–control) que sean iguales en el factor de interés no son informativas;
- Es un método menos eficiente

$$I(\beta, x) = \sum_{sets} \frac{\exp(\alpha + \beta x_0)}{\exp(\alpha + \beta x_0) + \exp(\alpha + \beta x_1) + \ldots + \exp(\alpha + \beta x_r)}$$

- La constante del modelo se anula (es igual para casos y para controles).

Ajuste del modelo

- Software:
  - Se puede usar un programa para analizar modelos de Cox de supervivencia, pues la función de verosimilitud es la misma.
- Es un modelo lineal generalizado, por lo que pueden emplearse los mismos métodos para valorar efectos;
- Los coeficientes $\beta_.$ son $\log(OR)$: $OR = e^{\beta}$.

21. Regresión multinomial

- La variable dependiente es categórica con más de dos grupos;
- Puede analizarse con regresión logística politómica (modelo multinomial);
- Se elige una categoría como referencia y se modelan varios *logits* simultáneamente, uno para cada una de las restantes categorías respecto a la de referencia.

Ejemplo: hábito tabáquico

- La variable resultado tiene tres categorías:
  - Fumador;
  - Exfumador;
  - No fumador (referencia).
- Se modelan dos *logits* simultáneamente:
  - $logit(fumador / no\ fumador \mid z) = \alpha_1 + \beta_1 z$;
  - $logit(ex-fumador / no\ fumador \mid z) = \alpha_2 + \beta_2 z$
- Las covariables z son comunes pero se estiman coeficientes diferentes para cada *logit* (incluso diferente constante).

22. Regresión ordinal

- La variable respuesta tiene más de dos categorías ordenadas;
- Se modela un único *logit* que recoge la relación (de tendencia) entre la respuesta y las covariables;
- Hay varios modelos posibles según interese modelar la tendencia:
  - *odds* proporcionales (acumulado);
  - categorías adyacentes (parejas).

22.1 *Odds*-proporcionales

- Se compara un promedio de los posibles logia acumulados (respecto a la primera categoría):

logit	respuesta			
	muy bajo	bajo	alto	muy alto
1				
2				
3				

- Cada *logit* tiene una constante diferente pero comparten el coeficiente de las covariables

- Modelo de *odds* proporcionales:

$$logit_k\left(y > y_k \mid z\right) = \alpha_k + \beta z$$

donde $y = 1,2,\ldots,C$;

$k = 2,3,\ldots,C$.

- Supone que el cambio entre diferentes puntos de corte de la respuesta es constante (β), pero parte de diferentes niveles (α_k).

22.2 Categorías adyacentes

- Compara cada categoría con la siguiente:

logit	respuesta			
	muy bajo	bajo	alto	muy alto
1				
2				
3				

- Cada *logit* tiene una constante diferente pero comparten el coeficiente de las covariables.

Modelo de categorías adyacentes:

$$logit_k\left(y_k > y_{k-1} \mid z\right) = \alpha_k + \beta z$$

donde $y = 1,2,\ldots,C$;

$k = 2,3,\ldots,C$.

- Supone que el cambio entre categorías adyacentes de la respuesta es constante (β), pero parte de diferentes niveles (α_k).

23. Elementos de interés en regresión logística

- Parámetros: (α, β)
- Matriz de varianza–covarianza:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \mathrm{var}(\alpha) & \mathrm{cov}(\alpha,\beta_1) & \mathrm{cov}(\alpha,\beta_1) \\ \mathrm{cov}(\alpha,\beta_1) & \mathrm{var}(\beta_1) & \mathrm{cov}(\beta_1,\beta_2) \\ \mathrm{cov}(\alpha,\beta_1) & \mathrm{cov}(\beta_2,\beta_1) & \mathrm{var}(\beta_2) \end{pmatrix}$$

- Valor de $\log L$ cuando es máximo:
  - "*Likelihood value*"
  - Tiene asociados unos "grados de libertad": *g.l. = #observaciones – #parámetros – 1*

24. Elementos derivados

$OR = e^{\beta}$

- permite: interpretar los coeficientes como riesgos.

Errores estándar de β: $ee_{\beta} = \sqrt{\text{var}(\beta)}$

- permite: calcular intervalos de confianza y realizar tests de hipótesis.

Deviance: $D = -2\log L$

- permite: valorar el ajuste del modelo (datos agrupados);
- realizar test de hipótesis (comparando modelos).

25. Bibliografía

1 Fagerland MW, Hosmer DW, Bofin AM. Multinomial goodness-of-fit tests for logistic regression models. Statist Med. 2008; 27:4238–53.

2. Wendy A. Bergerud. Introduction to logistic regression models with worked forestry examples biometrics Information. Handbook 7. British Columbia Ministry of Forests Research Program.

3. Muller R, Möckel M. Logistic regression and CART in the analysis of multimarker studies Clin Chim Acta 2008; 394: 1–6.

4. Hsiehl FY, Bloch DA, Larsen MD. A simple method of sample size calculation for linear and logistic regression Statist. Med. 1998;17:1623-34.

5. Gil JF, Pérez A, Castañeda JA. Empirical power from existing sample size formulae for logistic regression. Proceedings of the Annual Meeting of the American Statistical Association, 2001.

6. Ortega Calvo M, Cayuela Domínguez A. Regression logística no condicionada y tamaño de muestra: una revisión bibliográfica. Rev Esp Salud Pública 2002; 76: 85-93.

7. Alderete AM. Fundamentos del análisis de regresión logística en la investigación psicológica. Evaluar, 2006; 6: 52–67.

8. Silva C, Salinas M. Modelos de regresión y correlación III. Regresión logística. Ciencia y trabajo 2007; 24: 81-4.

9. Hidalgo Montesinos MD, Gómez Benito J. Comparación de la eficacia de regresión logística politómica y análisis discriminante logístico en la detección del DIF no uniforme. Psicothema 2000;12(Supl. 2), 298-300.

10. Bender Grouven U. Logistic regression models used in medical research are poorly presented. BMJ 1996; 313: 628.

11. Silva Ayçaguer LC. Excursión a la regresión logística en ciencias de la salud. Madrid: Díaz de Santos, 1995

12. Hosmer DW, Lemeshow S. Applied logistic regression. New York: Wiley, 1989.

13. Abraira Santos V, Pérez de Vargas Luque A. Métodos multivariantes en bioestadística Madrid: Centro de estudios Ramón Areces, 1996.

14. Silva Ayçaguer LC, Barroso Ultra IM. Regresión logística. Madrid: La muralla, 2004.

15. Pérez López C. Métodos estadísticos avanzados con SPSS. Thomson, 2005

16. Faulín Fajardo FJ. Bioestadística amigable. Madrid:Díaz de Santos, 2006

17. Sánchez-Cantalejo Ramírez E. Regresión logística en salud pública. EASP 2000

18. Hosmer DW, Lemeshow S. Applied logistic regression. Chichester: Wiley. 2000

19. Christensen, R. Log-linear models and logistic regression. Berlin: Springer. 1997

20. www.seh-lelha.org/rlogis1.htm

21. www.seh-lelha.org/pdf/rlogis2.pdf

22. www.hrc.es/bioest/Reglog_1.html

23. www.hrc.es/bioest/Reglog_2.html

24. halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/DM/tema2dm.pdf

25. ciberconta.unizar.es/LECCION/logis/inicio.html

26. saei.org/hemero/epidemiol/nota6.html

27. http://www.5campus.com/leccion/logis

28. patoral.umayor.cl/anestbas/rl.html

29. www.uoc.edu/in3/emath/docs/T10_Reg_Logistica.pdf

30. math.uprm.edu/~edgar/clasifall6.pdf

31. halweb.uc3m.es/esp/Personal/personas/amalonso/esp/bstat-tema9.pdf
32. www.salvador.edu.ar/csoc/idicso/docs/aephc1.pdf
33. es.geocities.com/r_vaquerizo/VL_regresion_logistica.htm
34. www.u-. cursos.cl/medicina/2008/2/TM1EPISP3/1/material_alumnos/objeto/20781
35. ciberconta.unizar.es/leccion/logis/060.HTM