Departamento de Estadística e I.O.

Máster en Estadística Aplicada

MODELOS DE RESPUESTA DISCRETA APLICACIONES BIOSANITARIAS

Tema 2
Modelos logit con variables explicativas cuantitativas

Profesores

Ana María Aguilera del Pino

Manuel Escabias Machuca

Título original: Modelos de Respuesta Discreta. Aplicaciones Biosanitarias. Tema 2: Modelos logit con variables explicativas cuantitativas

Índice general

Capítulo 2

Modelos logit con variables explicativas cuantitativas

Después de introducir en el capítulo anterior distintos modelos para explicar una variable de respuesta binaria a partir de un conjunto de variables explicativas, este capítulo estará dedicado al estudio detallado de los modelos de regresión logística en el caso de variables explicativas cuantitativas observadas sin error. Como es usual en modelización estadística, comenzaremos por la formulación del modelo e interpretación de sus parámetros, después abordaremos el problema de su estimación y contrastes de bondad de ajuste, y finalizaremos con la validación y selección del modelo más apropiado.

2.1. Modelo de regresión logística simple

Consideremos en primer lugar el caso de una única variable explicativa cuantitativa X. Recordemos que el modelo de regresión logística para una variable aleatoria binaria Y es un modelo lineal para el logaritmo de la ventaja de respuesta $Y = 1$ en cada valor observado x de la variable explicativa.

$$\ln\left[\frac{p(x)}{1-p(x)}\right] = \alpha + \beta x, \tag{2.1}$$

que equivalentemente se puede expresar de la siguiente forma en términos de la probabilidad de respuesta 1 en x:

$$p(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}},$$

curva de respuesta que es estrictamente creciente si $\beta > 0$ y estrictamente decreciente para $\beta < 0$.

2.1.1. Interpretación de parámetros

1. Si $\beta = 0$ entonces $p(x) = e^{\alpha}/(1 + e^{\alpha})$ que quiere decir que la variable Y es independiente de X puesto que $p(x)$ no depende de x.

2. α es el valor común del logaritmo de las ventajas de respuesta $Y = 1$ frente a respuesta $Y = 0$ cuando $\beta = 0$, es decir cuando la respuesta es independiente de la variable explicativa.

   Por otro lado, α se puede interpretar alternativamente como el valor del logaritmo de la ventaja de respuesta 1 para un individuo con $X = 0$.

3. La fórmula general del modelo logit simple (2.1) implica que por cada unidad de incremento en X, el logit de respuesta 1 aumenta aditivamente en β unidades. De la misma fórmula se obtiene la siguiente expresión para la ventaja de la respuesta 1 en cada x observado:

   $$\frac{p(x)}{1 - p(x)} = \exp(\alpha + \beta x) = e^{\alpha}(e^{\beta})^{x},$$

   que significa que la ventaja de respuesta 1 aumenta multiplicativamente e^{β} por cada unidad de incremento en X. De hecho,

   $$\frac{p(x+1)}{1 - p(x+1)} = \frac{p(x)}{1 - p(x)} e^{\beta}.$$

4. El cociente de ventajas de respuesta 1 para dos valores diferentes x_1 y x_2 de X es de la forma

   $$\theta(x_1, x_2) = \frac{\dfrac{p(x_1)}{1 - p(x_1)}}{\dfrac{p(x_2)}{1 - p(x_2)}} = \frac{e^{\alpha + \beta x_1}}{e^{\alpha + \beta x_2}} = e^{\beta(x_1 - x_2)}.$$

   La ventaja o preferencia de la respuesta 1 frente a la 0 toma valores en el intervalo $(0,\infty)$. Por lo tanto, el cociente de ventajas $\theta(x_1, x_2)$ tiene el mismo rango de variación y la siguiente interpretación:

   $\theta(x_1, x_2) = 1$ sii $p(x_1) = p(x_2)$

   $\theta(x_1, x_2) > 1$ sii $p(x_1) > p(x_2)$. En este caso la ventaja de respuesta 1 es $e^{\beta(x_1 - x_2)}$ veces mayor para $X = x_1$ que para $X = x_2$.

   $\theta(x_1, x_2) < 1$ sii $p(x_1) < p(x_2)$. En este caso la ventaja de respuesta 1 es $1/e^{\beta(x_1 - x_2)}$ veces mayor para $X = x_2$ que para $X = x_1$.

5. La exponencial del parámetro β es el cociente de ventajas de respuesta 1 para dos valores de X que se diferencien en una unidad. Es decir,

   $$\theta(x + 1, x) = \theta(\Delta X = 1) = e^{\beta}.$$

2.2. Modelo de regresión logística múltiple

Consideremos ahora el caso de R variables explicativas cuantitativas no aleatorias $(X_1, X_2, \ldots, X_R)$. Para cada combinación de valores observados $X_1 = x_1, X_2 = x_2, \ldots, X_R = x_R$ de las variables explicativas, la variable respuesta Y tiene distribución de Bernouilli

$$Y/(X_1 = x_1, X_2 = x_2, \ldots, X_R = x_R) \to B(1, p(x_1, x_2, \ldots, x_R)),$$

siendo

$$p(x_1, \ldots, x_R) = P[Y = 1/X_1 = x_1, \ldots, X_R = x_R] = E[Y/X_1 = x_1, \ldots, X_R = x_R].$$

El modelo de regresión logística múltiple para la variable respuesta binaria Y en términos de valores observados $X_1 = x_1, \ldots, X_R = x_R$ de las variables explicativas es de la forma

$$Y(x_1, \ldots, x_R) = p(x_1, \ldots, x_R) + \epsilon(x_1, \ldots, x_R),$$

donde $\epsilon(x_1, \ldots, x_R)$ son errores aleatorios que se consideran centrados e independientes, de modo que

$$p(x_1, \ldots, x_R) = \frac{\exp\left(\alpha + \sum_{r=1}^{R} \beta_r x_r\right)}{1 + \exp\left(\alpha + \sum_{r=1}^{R} \beta_r x_r\right)}.$$

Denotando a partir de ahora $\alpha = \beta_0$, $X = (X_0, X_1, \ldots, X_R)'$, y $x = (x_0, x_1, \ldots, x_R)'$ con $X_0 = 1$, el modelo quedará resumido como sigue:

$$p(x) = \frac{\exp\left(\sum_{r=0}^{R} \beta_r x_r\right)}{1 + \exp\left(\sum_{r=0}^{R} \beta_r x_r\right)} = \frac{\exp \beta' x}{1 + \exp \beta' x}, \tag{2.2}$$

donde β es el vector columna de parámetros $(\beta_0, \beta_1, \ldots, \beta_R)'$.

Equivalentemente el modelo de regresión logística múltiple se puede ver como un modelo de regresión lineal múltiple para la transformación logit

$$\ln\left[\frac{p(x)}{1 - p(x)}\right] = \sum_{r=0}^{R} \beta_r x_r. \tag{2.3}$$

2.2.1. Interpretación de parámetros

1. Si $\beta_r = 0 \quad \forall r = 1, \ldots, R$ entonces $p(x) = e^{\beta_0}/(1+e^{\beta_0})$ que quiere decir que la variable Y es independiente de las variables explicativas.

2. β_0 es el valor común del logaritmo de las ventajas de respuesta $Y = 1$ frente a respuesta $Y = 0$ cuando la respuesta es independiente de las variables explicativas.

   Por otro lado, β_0 es el valor del logaritmo de la ventaja de respuesta $Y = 1$ para un individuo con $X_1 = X_2 = \ldots = X_R = 0$.

3. El cociente de ventajas de respuesta $Y = 1$ para dos combinaciones diferentes de valores de las variables explicativas, $x_1 = (1, x_{11}, \ldots x_{1R})'$ y $x_2 = (1, x_{21}, \ldots x_{2R})'$, es de la forma

$$\theta(x_1, x_2) = \frac{\dfrac{p(x_1)}{1-p(x_1)}}{\dfrac{p(x_2)}{1-p(x_2)}} = \frac{\exp\left(\displaystyle\sum_{r=0}^{R} \beta_r x_{1r}\right)}{\exp\left(\displaystyle\sum_{r=0}^{R} \beta_r x_{2r}\right)} = \exp\left(\sum_{r=1}^{R} \beta_r (x_{1r} - x_{2r})\right).$$

   Para dos valores x_1 y x_2 que se diferencien en una unidad, $x_{1r} - x_{2r} = 1 \quad \forall r = 1, \ldots R$ se tendría

$$\theta(\Delta X_1 = 1, \ldots, \Delta X_R = 1) = \exp\left(\sum_{r=1}^{R} \beta_r\right) = \prod_{r=1}^{R} e^{\beta_r}.$$

4. Para dar una interpretación más intuitiva de los parámetros del modelo, vamos a calcular el cociente de ventajas de respuesta 1 cuando una de las variables explicativas se incrementa en una unidad y las otras se controlan haciendo que tomen valores fijos.

   Como ejemplo incrementamos en una unidad la variable X_l y las restantes $R-1$ variables explicativas las mantenemos fijas. Entonces, sustituyendo por $x_{1l} - x_{2l} = 1$ y $x_{1r} - x_{2r} = 0 \quad \forall r \neq l$ en la fórmula obtenida anteriormente para el cociente de ventajas se tiene

$$\theta(\Delta X_l = 1/x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_R) = e^{\beta_l} \quad \forall l = 1, \ldots, R.$$

   Esto significa que al aumentar en una unidad una de las variables y controlar las demás, la ventaja de respuesta $Y = 1$ queda multiplicada por la exponencial del coeficiente de la variable incrementada. De

este modo si la exponencial de un parámetro es mayor que uno la probabilidad de respuesta $Y = 1$ aumenta cuando aumenta la variable correspondiente y se controlan las demás, mientras que si es menor que uno la relación es inversa.

2.3. Modelos con interacción

Hasta ahora no se ha considerado la posibilidad de interacción entre las variables explicativas de un modelo de regresión logística múltiple.

2.3.1. Interacción y confusión

Observemos que en los modelos de regresión logística múltiple considerados hasta ahora, los cocientes de ventajas que miden la asociación entre la variable de respuesta y cada variable explicativa son independientes del valor fijo del resto de variables explicativas controladas. Esto significa que son modelos sin interacción porque el grado de asociación entre la variable de respuesta y cada una de las variables explicativas es el mismo en todas las combinaciones de niveles de las otras variables independientes.

Se pueden considerar interacciones de distintos órdenes. Las más simples son las de orden uno (entre dos variables explicativas) que representan el grado en que la asociación entre la variable de respuesta y una variable depende de los valores de una tercera que interacciona con esta última. Las interacciones de orden dos involucran a tres variables y así sucesivamente. Por simplicidad no consideraremos interacciones de orden superior a uno que conllevarían productos entre tres o más variables.

En epidemiología es usual distinguir entre el factor de riesgo que puede ser causa de una enfermedad y otras covariables de interés que hay que controlar en el estudio estadístico para analizar la asociación entre el factor de riesgo y el padecimiento de la enfermedad. En este tipo de estudios es usual distinguir entre factores de confusión y factores modificadores del efecto del factor de riesgo sobre la enfermedad.

Una variable es de confusión cuando está asociada con el factor de riesgo de modo que la asociación marginal entre la variable de respuesta y el factor de riesgo cambia significativamente al incluirla en el análisis estadístico.

Una variable es modificadora del efecto cuando la asociación entre la variable de respuesta y el factor de riesgo cambia en función de sus valores. Es decir, se trata de una variable que interacciona con el factor de riesgo.

De lo anterior se deduce que los factores de confusión tienen que ser considerados forzosamente en el modelo, aunque puede que no interaccionen

con el factor de riesgo.

2.3.2. Formulación de modelos con interacción

La interacción entre dos variables cuantitativas se incluye en el modelo de regresión logística múltiple como producto de ambas variables. En general, el término de interacción entre dos variables cuantitativas X_r y X_s es de la forma $\beta_{rs}X_rX_s$.

Como consecuencia el modelo de regresión logística múltiple con interacciones entre cada par de covariables es de la forma

$$\ln\left[\frac{p(x)}{1-p(x)}\right] = \sum_{r=0}^{R}\beta_r x_r + \sum_{r=1}^{R}\sum_{s>r}\beta_{rs}x_r x_s.$$

En este caso el cociente de ventajas de respuesta $Y=1$ cuando se incrementa en una unidad una variable y se controlan fijas las demás depende del valor de las variables controladas

$$\theta(\Delta X_l = 1/x_1,\ldots,x_{l-1},x_{l+1},\ldots,x_R) = e^{\beta_l+\sum_{r\neq l}\beta_{lr}x_r} \quad \forall l=1,\ldots,R,$$

poniendo claramente en evidencia la interacción de cada variable con el resto.

2.4. Ajuste de modelos logit

En esta sección vamos a abordar el problema de la estimación de los parámetros de los modelos logit.

Los datos están constituidos por una muestra de tamaño N de la v.a. respuesta Y. Es decir, se dispone de N observaciones de N variables de Bernouilli independientes (sucesión de unos y ceros), a cada una de las cuales corresponde una determinada combinación de niveles, $(x_0, x_1, \ldots, x_R)$ de las R variables explicativas $X_1,\ldots,X_R$.

Si denotamos por $x_q = (x_{q0}, x_{q1}, \ldots, x_{qR})'$ $(q = 1,\ldots,Q)$ a la q-ésima combinación de valores de las R variables explicativas en la muestra, pueden ocurrir dos casos:

1. Para cada individuo muestral existe una combinación diferente de niveles de las R variables explicativas $(Q = N)$. Esto significa que hay una única observación de la v.a. de respuesta Y en cada combinación de valores de las variables explicativas, y suele ocurrir cuando las variables explicativas son todas continuas.

2. A individuos muestrales diferentes corresponden valores iguales de las variables explicativas ($Q < N$). Esto quiere decir que hay más de una observación de la v.a. de respuesta en cada combinación de valores de las variables explicativas.

Denotando por n_q al número de observaciones muestrales con $X = x_q$ y por y_q al número de respuestas $Y = 1$ de entre estas n_q observaciones, se dispone de una muestra de Q variables aleatorias independientes Y_q con distribuciones $B(n_q, p_q)$, donde $p_q = P[Y = 1/X = x_q]$. Por lo tanto, $E[Y_q] = n_q p_q$ y $\sum_{q=1}^{Q} n_q = N$.

Observemos que Y_q representa en general el número de respuestas $Y = 1$ en cada $X = x_q$. En el caso en que no hay observaciones repetidas ($Q = N$) los valores observados y_q corresponden a las respuestas binarias individuales (1 o 0).

El modelo de regresión logística muestral es de la forma

$$p_q = \frac{\exp\left(\sum_{r=0}^{R} \beta_r x_{qr}\right)}{1 + \exp\left(\sum_{r=0}^{R} \beta_r x_{qr}\right)}. \tag{2.4}$$

En forma lineal,

$$L_q = \ln\left[\frac{p_q}{1 - p_q}\right] = \sum_{r=0}^{R} \beta_r x_{qr},$$

y equivalentemente en forma matricial

$$L = X\beta,$$

donde L es el vector $Q \times 1$ de transformaciones logit

$$L = (L_1, \ldots, L_Q)',$$

β es el vector $(R + 1) \times 1$ de parámetros

$$\beta = (\beta_0, \beta_1, \ldots, \beta_R)',$$

y X es la matriz del diseño que contiene las observaciones de las variables explicativas

$$X = \begin{pmatrix} x_{10} & x_{11} & \ldots & x_{1r} & \ldots & x_{1R} \\ x_{20} & x_{21} & \ldots & x_{2r} & \ldots & x_{2R} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ x_{q0} & x_{q1} & \ldots & x_{qr} & \ldots & x_{qR} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ x_{Q0} & x_{Q1} & \ldots & x_{Qr} & \ldots & x_{QR} \end{pmatrix}$$

2.4.1. Estimación por máxima verosimilitud

Es conocido que los estimadores de máxima verosimilitud (MV) son los valores de los parámetros que dan máxima probabilidad (verosimilitud) a los datos observados. Para encontrarlos tendremos que maximizar la función de verosimilitud de los datos respecto de los parámetros del modelo logit.

La función de verosimilitud es el producto de las Q funciones masa de probabilidad de las Q binomiales independientes Y_q dada por

$$\prod_{q=1}^{Q} \left(\begin{array}{c} n_q \\ y_q \end{array} \right) p_q^{y_q} (1-p_q)^{n_q - y_q},$$

cuyo núcleo (que alcanza el máximo en el mismo punto) es

$$\prod_{q=1}^{Q} p_q^{y_q} (1-p_q)^{n_q - y_q} = \left(\prod_{q=1}^{Q} (1-p_q)^{n_q} \right) \left(\prod_{q=1}^{Q} \left(\frac{p_q}{1-p_q} \right)^{y_q} \right).$$

Maximizaremos como es usual el logaritmo del núcleo de la función de verosimilitud que bajo el modelo de regresión logística (2.4) es de la forma

$$\begin{aligned} L(\beta) &= \sum_{q=1}^{Q} n_q \ln(1-p_q) + \sum_{q=1}^{Q} y_q \ln\left(\frac{p_q}{1-p_q} \right) \\ &= -\sum_{q=1}^{Q} n_q \ln\left(1 + \exp\left(\sum_{r=0}^{R} \beta_r x_{qr} \right) \right) + \sum_{q=1}^{Q} y_q \left(\sum_{r=0}^{R} \beta_r x_{qr} \right) \\ &= -\sum_{q=1}^{Q} n_q \ln\left(1 + \exp\left(\sum_{r=0}^{R} \beta_r x_{qr} \right) \right) + \sum_{r=0}^{R} \left(\sum_{q=1}^{Q} y_q x_{qr} \right) \beta_r, \end{aligned}$$

que depende de los datos observados y_q sólo a través de los estadísticos suficientes $\{\sum_{q=1}^{Q} y_q x_{qr},\ r = 0, \ldots, R\}$.

Derivando respecto a cada uno de los parámetros β_r se obtiene

$$\frac{\Delta L(\beta)}{\Delta \beta_r} = \sum_{q=1}^{Q} y_q x_{qr} - \sum_{q=1}^{Q} n_q x_{qr} \left(\frac{\exp\left(\sum_{r=0}^{R} \beta_r x_{qr} \right)}{1 + \exp\left(\sum_{r=0}^{R} \beta_r x_{qr} \right)} \right).$$

Igualando a cero se tienen las ecuaciones de verosimilitud

$$\sum_{q=1}^{Q} y_q x_{qr} - \sum_{q=1}^{Q} n_q \hat{p}_q x_{qr} = 0 \quad r = 0, \ldots, R,$$

donde $\hat{p}_q$ es el estimador MV de p_q

$$\hat{p}_q = \frac{\exp\left(\sum_{r=0}^{R} \hat{\beta}_r x_{qr}\right)}{1 + \exp\left(\sum_{r=0}^{R} \hat{\beta}_r x_{qr}\right)}, \tag{2.5}$$

y $\hat{\beta}_r$ los estimadores MV de los parámetros.

Las ecuaciones de verosimilitud se pueden escribir equivalentemente en forma matricial como

$$X'y = X'\hat{m}$$

donde X es la matriz del diseño, y es el vector $Q \times 1$ del número de respuestas $Y = 1$ observadas y_q, y $\hat{m}$ es el vector $Q \times 1$ cuyas componentes son las frecuencias esperadas de respuesta $Y = 1$ (estimación MV de las medias de Y_q) bajo el modelo de regresión logística (2.4), dadas por $\hat{m}_q = n_q \hat{p}_q$.

En el caso de una única observación en cada combinación de valores de las variables explicativas ($Q = N$), se tiene que $\hat{m} = \hat{p}$, siendo $\hat{p}$ el vector de valores predichos para las observaciones binarias y_i, que suele denotarse por $\hat{y}$.

Observemos que las ecuaciones de verosimilitud obtenidas son similares a las del modelo de regresión lineal múltiple

$$Y = X\beta + \epsilon,$$

con $Y = (y_1, \ldots, y_N)'$ vector de observaciones independientes de una v.a. Y con distribución normal $N(\mu_q, \sigma^2)$, siendo $\mu_q = E[Y/X = x_q]$, y $x_q = (1, x_{q1}, \ldots, x_{qR})$ la q-ésima fila de la matriz del diseño X que contiene las observaciones de las variables explicativas para cada individuo muestral.

Las ecuaciones de verosimilitud del modelo de regresión lineal múltiple así formulado son de la forma $X'Y = X'\hat{Y}$, siendo $\hat{Y} = X\hat{\beta}$ y $\hat{\beta} = (X'X)^{-1}X'Y$. Esta similitud entre las ecuaciones de verosimilitud es debida a que en todos los GLM con ligadura canónica, como el caso de la regresión lineal múltiple y de la regresión logística, las ecuaciones de verosimilitud igualan los estadísticos suficientes a sus valores esperados.

Como la log-verosimilitud es una función cóncava para los modelos logit (y también para modelos probit), los estimadores MV de sus parámetros existen y son únicos excepto en ciertos casos. Supongamos que ordenamos las observaciones según una única variable explicativa, y que a todos los valores para los que la variable respuesta es cero corresponden valores de la variable explicativa menores que aquellos para los que la respuesta es uno. Este caso se conoce como separación completa y lleva a la no existencia de los estimadores MV. Está demostrado que para que existan los estimadores

MV tiene que darse cierto solapamiento en los datos. Cuando las variables explicativas son continuas es difícil encontrar separación completa en las observaciones. Un estudio más profundo sobre la separación completa y la existencia de los estimadores MV se puede ver en Ryan (1997).

Observemos finalmente que las ecuaciones de verosimilitud no son lineales en los parámetros por lo que requieren el uso de métodos de solución iterativa como el de Newton-Raphson que se presenta a continuación.

2.4.2. Estimación MV iterativa con Newton-Raphson

Recordemos que el algoritmo de Newton-Raphson es un método iterativo para calcular de forma aproximada el máximo de una función. El método requiere una aproximación inicial del valor en el que la función alcanza el máximo. En el primer ciclo la función se aproxima en un entorno de este valor inicial mediante el polinomio de segundo grado obtenido al truncar el desarrollo de Taylor de la función en ese punto y se toma como localización del máximo de la función la de este polinomio de segundo grado. Así sucesivamente, la localización del máximo de la función en el ciclo (t) corresponde a la del polinomio de segundo grado obtenido truncando el desarrollo de Taylor de la función en un entorno de la aproximación obtenida en el ciclo $(t-1)$. La sucesión de estimaciones así obtenidas convergen a la localización del máximo de la función siempre que verifique ciertas condiciones de regularidad y el valor inicial sea adecuado.

A continuación aplicaremos este método iterativo para obtener el estimador MV de β en el modelo de regresión logística múltiple. Es decir, el vector β donde alcanza el máximo la log-verosimilitud $L(\beta)$.

El desarrollo de Taylor de $L(\beta)$ en un entorno del valor inicial $\beta^{(0)}$ truncado en el tercer término es de la forma

$$Q^{(0)}(\beta) = L(\beta^{(0)}) + (D^{(0)})'(\beta - \beta^{(0)}) + \frac{1}{2}(\beta - \beta^{(0)})'H^{(0)}(\beta - \beta^{(0)}),$$

donde $D^{(0)}$ es el vector columna $(R+1) \times 1$ de derivadas primeras de $L(\beta)$ en $\beta = \beta^{(0)}$, y $H^{(0)}$ es la matriz $(R+1) \times (R+1)$ de derivadas segundas de $L(\beta)$ en $\beta = \beta^{(0)}$.

Para obtener el máximo de $Q^{(0)}$ derivamos respecto de β e igualamos a cero obteniendo la ecuación

$$D^{(0)} + H^{(0)}\beta - H^{(0)}\beta^{(0)} = 0,$$

de donde se obtiene que la localización del máximo en el primer ciclo es

$$\beta^{(1)} = \beta^{(0)} - (H^{(0)})^{-1}D^{(0)}.$$

Así sucesivamente, la localización del máximo de la log-verosimiltud en el ciclo (t) se obtiene como sigue a partir de la localización del máximo en el ciclo anterior $(t-1)$:

$$\beta^{(t)} = \beta^{(t-1)} - (H^{(t-1)})^{-1} D^{(t-1)}. \tag{2.6}$$

Según hemos obtenido en el apartado anterior, las derivadas primeras de $L(\beta)$ evaluadas en $\beta^{(t-1)}$ son:

$$D_s^{(t-1)} = \left.\frac{\Delta L(\beta)}{\Delta \beta_s}\right|_{\beta^{(t-1)}} = \sum_{q=1}^{Q} \left(y_q - n_q p_q^{(t-1)}\right) x_{qs}$$

donde $p_q^{(t-1)}$ es obtenido a partir de $\beta^{(t-1)}$ en la forma

$$p_q^{(t-1)} = \frac{\exp\left(\sum_{r=0}^{R} \beta_r^{(t-1)} x_{qr}\right)}{1 + \exp\left(\sum_{r=0}^{R} \beta_r^{(t-1)} x_{qr}\right)} \tag{2.7}$$

Calculemos ahora la matriz de derivadas segundas de $L(\beta)$

$$\begin{aligned}
\frac{\Delta^2 L(\beta)}{\Delta \beta_r \Delta \beta_s} &= \frac{\Delta}{\beta_r}\left[\sum_{q=1}^{Q} y_q x_{qs} - \sum_{q=1}^{Q} n_q x_{qs} \frac{\exp\left(\sum_{r=0}^{R} \beta_r x_{qr}\right)}{1 + \exp\left(\sum_{r=0}^{R} \beta_r x_{qr}\right)}\right] \\
&= -\sum_{q=1}^{Q} n_q x_{qr} x_{qs} \frac{\exp\left(\sum_{r=0}^{R} \beta_r x_{qr}\right)}{\left[1 + \exp\left(\sum_{r=0}^{R} \beta_r x_{qr}\right)\right]^2} \\
&= -\sum_{q=1}^{Q} x_{qr} x_{qs} n_q p_q (1 - p_q).
\end{aligned}$$

Por lo tanto, la matriz de derivadas segundas de la log-verosimilitud evaluadas en la aproximación del máximo en el paso (t-1) son de la forma

$$H_{rs}^{(t-1)} = \left.\frac{\Delta^2 L(\beta)}{\Delta \beta_r \Delta \beta_s}\right|_{\beta^{(t-1)}} = -\sum_{q=1}^{Q} x_{qr} x_{qs} n_q p_q^{(t-1)} (1 - p_q^{(t-1)}).$$

Sustituyendo en la fórmula (2.6) se obtiene la localización aproximada del máximo de $L(\beta)$ en el paso (t)

$$\beta^{(t)} = \beta^{(t-1)} + [X'Diag[n_q p_q^{(t-1)}(1 - p_q^{(t-1)})]X]^{-1}X'(y - m^{(t-1)}) \qquad (2.8)$$

siendo $m_q^{(t-1)} = n_q p_q^{(t-1)}$.

El procedimiento comienza con unos valores iniciales $\beta^{(0)}$ a partir de los cuales calcula $p^{(0)}$. A partir de $t > 0$ las aproximaciones se calculan con la fórmula iterativa anterior.

Las aproximaciones $p^{(t)}$ y $\beta^{(t)}$ convergen a los estimadores MV $\hat{p}$ y $\hat{\beta}$, respectivamente. La convergencia del método de Newton-Raphson suele ser rápida. De hecho esta convergencia es de segundo orden porque se tiene que para valores de t grandes, existe $c > 0$ verificando:

$$|\beta_r^{(t)} - \hat{\beta}_r| \leq c|\beta_r^{(t-1)} - \hat{\beta}_r|^2, \quad \forall r.$$

En el caso de modelos de regresión logística este método suele tomar pocas iteraciones para alcanzar un grado de convergencia satisfactorio. No hay acuerdo sobre el criterio de convergencia o de parada del algoritmo de Newton-Raphson que suele ser uno de los tres siguientes:

1. Cambio de menos de 10^{-9} en los parámetros β estimados en un ciclo y el siguiente. Es decir,

$$|\beta_r^{(t)} - \beta_r^{(t-1)}| \leq 10^{(-9)} \quad r = 0, \ldots, R.$$

2. Cambio de menos de 10^{-3} en la log-verosimilitud. Es decir,

$$|L(\beta^{(t)}) - L(\beta^{(t-1)})| \leq 10^{-3}.$$

3. Cambios muy pequeños en las probabilidades predichas p_q de respuesta $Y = 1$. Por ejemplo,

$$|p_q^{(t)} - p_q^{(t-1)}| \leq 10^{-3} \quad \forall q = 1, \ldots, Q.$$

Finalmente la elección de valores iniciales $\beta^{(0)}$ adecuados se puede hacer por distintos procedimientos. En Ruiz-Maya *et al.* (1995) se propone tomar $\beta^{(0)} = 0$ o bien la estimación de β obtenida mediante mínimos cuadrados ordinarios. Mediante análisis discriminante sobre las variables explicativas (adecuado en el caso de normalidad), se obtienen valores iniciales apropiados

(Hosmer y Lemeshow, 1989). En el caso del modelo de regresión logística simple estos valores iniciales para los parámetros $\beta_0^{(0)}$ y $\beta_1^{(0)}$ son de la forma

$$\begin{aligned} \beta_0^{(0)} &= \frac{\ln(\pi_1/\pi_0) - 0{,}5(\bar{x}_1^2 - \bar{x}_0^2)}{S^2} \\ \beta_1^{(0)} &= \frac{(\bar{x}_1 - \bar{x}_0)}{S^2}, \end{aligned}$$

donde π_1 es la proporción de respuestas $Y = 1$ en la muestra, definida por

$$\pi_1 = \frac{\sum_{q=1}^{Q} y_q}{N}$$

y π_0 es la proporción de respuestas $Y = 0$ en la muestra dada por $\pi_0 = 1 - \pi_1$. Por otro lado se asume que

$$(X/Y = 1) \to N(\mu_1, \sigma^2) \quad \text{y} \quad (X/Y = 0) \to N(\mu_0, \sigma^2),$$

de modo que $\bar{x}_1$ es la media muestral de X para las observaciones con $Y = 1$, y $\bar{x}_0$ es la media muestral de X para las observaciones $Y = 0$, definidas ambas como sigue

$$\begin{aligned} \bar{x}_1 &= \frac{\sum_{q=1}^{Q} x_q y_q}{a_1} \\ \bar{x}_0 &= \frac{\sum_{q=1}^{Q} x_q (n_q - y_q)}{a_0}, \end{aligned}$$

definiendo $a_1 = \sum_{q=1}^{Q} y_q$ como el número de observaciones con $Y = 1$ en la muestra, y $a_0 = N - a_1$ como el número de observaciones con $Y = 0$.

Finalmente, S^2 se define en la forma

$$S^2 = \frac{(a_0 - 1)S_0^2 + (a_1 - 1)S_1^2}{a_0 + a_1 - 1},$$

donde S_0^2 y S_1^2 son las varianzas muestrales de X para las observaciones con $Y = 0$ y con $Y = 1$, respectivamente, que se definen como es usual en la

forma

$$S_1^2 = \frac{\sum_{q=1}^{Q} x_q^2 y_q}{a_1} - \bar{x}_1^2$$

$$S_0^2 = \frac{\sum_{q=1}^{Q} x_q^2 (n_q - y_q)}{a_0} - \bar{x}_0^2.$$

2.4.3. Estimación por mínimos cuadrados ponderados

Como método alternativo para la estimación de los parámetros del modelo de regresión logística múltiple, se podría usar el método de mínimos cuadrados. En este caso las varianzas de las observaciones no son las mismas para todos los valores de las variables explicativas (heterocedasticidad), de modo que en lugar de mínimos cuadrados ordinarios se utiliza el método de mínimos cuadrados ponderados para la estimación de los parámetros de un determinado modelo. En este tipo de ajuste cada observación se pondera por el valor inverso de su varianza.

Consideremos de nuevo el modelo logit muestral

$$L = X\beta.$$

Denotando por $f_q = y_q/n_q$ a la proporción observada de respuestas $Y = 1$ en la q-ésima combinación de valores de las variables explicativas, el ajuste por mínimos cuadrados ponderados de este modelo logit corresponde al ajuste del modelo lineal

$$Z = X\beta + \epsilon,$$

siendo Z el vector columna de elementos $z_q = \ln[f_q/(1 - f_q)]$, y ϵ el vector de errores que se consideran centrados e independientes. Los valores z_q se suelen llamar transformaciones logit muestrales ya que son los estimadores MV de los logit poblacionales L_q, cuya distribución es asintóticamente normal (Ruiz-Maya et al., 1995).

El estimador de mínimos cuadrados ponderados (WLS) de β en el modelo lineal general $Z = X\beta + \epsilon$ es de la forma $\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}Z$, donde V es la matriz de covarianzas de los residuos ϵ.

En nuestro caso $V = Cov(\epsilon) = [Diag(n_q f_q(1 - f_q))]^{-1}$, de modo que la estimación por mínimos cuadrados ponderados de sus parámetros viene dada por:

$$\hat{\beta} = (X'Diag(n_q f_q(1 - f_q))X)^{-1}X'Diag(n_q f_q(1 - f_q))Z.$$

La justificación para el uso de mínimos cuadrados ponderados es que los estimadores obtenidos son asintóticamente óptimos si los n_q son grandes. Por ello no tiene sentido usar este método de estimación para muestras pequeñas.

Una dificultad obvia añadida es cuando $n_q = 1$, f_q es 1 o 0 y por lo tanto z_q no está definido. En este caso como el peso que le corresponde a dicha observación, obtenido a partir de la inversa de la matriz de covarianzas, es también nulo, se podría argumentar que el método de mínimos cuadrados ponderados simplemente ignora estos casos. En el caso de variables continuas este tipo de estimación podría llevar a ignorar casi toda la información e incluso todos los datos. Una corrección ad hoc simple para este problema suele ser sumar una cantidad pequeña a aquellas proporciones estimadas f_q que sean cero o uno. Lo usual es sustituir por 0.5 las frecuencias de respuesta con y_q nulas.

Observemos que cuando los valores de partida para el método de Newton-Raphson son $p_q^{(0)} = f_q$, la estimación de mínimos cuadrados ponderados coincide con la aproximación de los estimadores MV proporcionada por Newton-Raphson en el primer paso. Además, el método de Newton-Raphson también requiere ajustar aquellos valores f_q iguales a 0 o 1 antes de comenzar.

Observemos además que el valor aproximado en el paso (t) mediante Newton-Raphson se puede expresar equivalentemente en la forma

$$\beta^{(t)} = [X'Diag[n_q p_q^{(t-1)}(1 - p_q^{(t-1)})]X]^{-1} X'Diag[n_q p_q^{(t-1)}(1 - p_q^{(t-1)})]Z^{(t-1)} \tag{2.9}$$

donde $Z^{(t-1)}$ es la forma muestral linealizada de la transformación logit, que tiene como elementos:

$$z_q^{(t-1)} = \ln\left[\frac{p_q^{(t-1)}}{1 - p_q^{(t-1)}}\right] + \frac{y_q - n_q p_q^{(t-1)}}{n_q p_q^{(t-1)}(1 - p_q^{(t-1)})},$$

que están muy cercanos a las transformaciones logit muestrales z_q cuando $p_q^{(t-1)}$ está próximo a f_q, que es el caso del primer ciclo del método de Newton-Raphson con valores iniciales $p_q^{(0)} = f_q$.

De esto se deduce que $\beta^{(t+1)}$ es el estimador de mínimos cuadrados ponderados para el modelo lineal general $Z^{(t)} = X\beta^{(t)} + \epsilon^{(t)}$ con errores $\epsilon_q^{(t)}$ incorrelados de varianza $1/(n_q p_q^{(t)}(1 - p_q^{(t)})$.

El estimador MV de β es el límite de una sucesión de estimadores de mínimos cuadrados ponderados, de modo que la matriz de las ponderaciones cambia en cada ciclo. Por ello al método iterativo de Newton-Raphson para el cálculo aproximado de los estimadores MV se le llama método de *mínimos cuadrados iterativamente reponderados*.

Sin embargo, la diferencia fundamental está en que el método de Newton-Raphson converge a los estimadores MV independientemente de los valores

de partida (siempre que cumplan las condiciones requeridas). El método de mínimos cuadrados ponderados si depende crucialmente de los valores iniciales.

2.4.4. Propiedades de los estimadores MV

Wald en 1943 proporcionó resultados asintóticos generales para los estimadores de máxima verosimilitud demostrando que, bajo ciertas condiciones de regularidad, los estimadores de máxima verosimilitud tienen distribución asíntótica normal con media el valor poblacional del parámetro estimado y matriz de covarianzas dada por la inversa de la matriz de información de Fisher. Esto significa que los estimadores de máxima verosimilitud son asintóticamente insesgados y además se puede hacer inferencia sobre ellos basándose en dicha distribución normal cuando el tamaño muestral es suficientemente grande.

En nuestro caso, el estimador MV $\hat{\beta}$ del vector de parámetros β del modelo de regresión logística múltiple converge en distribución a la siguiente distribución normal

$$\hat{\beta} \underset{N \to \infty}{\overset{d}{\longrightarrow}} N(\beta, Cov(\hat{\beta})) \ ,$$

donde la matriz de covarianzas es

$$Cov(\hat{\beta}) = \Sigma_{\hat{\beta}} = \left(-E \left[\frac{\Delta^2 L}{\Delta\beta_r \Delta\beta_s} \right] \right)^{-1} = (X' Diag[n_q p_q (1 - p_q)] X)^{-1}. \quad (2.10)$$

Por lo tanto, las raíces cuadradas de los elementos de la diagonal de esta matriz son los errores estándar (ASE) de los estimadores de los parámetros del modelo.

La estimación MV de esta matriz de covarianzas se obtiene sustituyendo en la expresión anterior p_q por su estimación MV $\hat{p}_q$.

El método de Newton-Raphson garantiza que las matrices de derivadas segundas $H^{(t)}$ convergen a la matriz de derivadas segundas de la log-verosimilitud evaluada en el estimador MV $\hat{\beta}$ que denotaremos por $\hat{H}$. Entonces, las matrices $-(H^{(t)})^{-1}$ convergen a la estimación MV de la matriz de covarianzas asintótica de los parámetros $-(\hat{H})^{-1}$, que se obtiene, por lo tanto, como un subproducto del método de Newton-Raphson.

2.5. Inferencia en regresión logística

Una vez estimados los parámetros del modelo de regresión logística múltiple nos proponemos hacer inferencia para extrapolar los resultados muestrales a la población.

2.5.1. Contrastes de bondad de ajuste

Recordemos que y_q representa el número respuestas $Y = 1$ (éxitos) en las n_q observaciones correspondientes a la q-ésima combinación de valores de las variables explicativas.

Una vez estimados los parámetros, a partir de ellos se estiman los logit, y a partir de estos las probabilidades $\hat{p}_q$. Por lo tanto, las frecuencias esperadas de respuesta $Y = 1$, estimadas bajo el modelo de regresión logística, son en este caso de la forma $\hat{m}_q = n_q\hat{p}_q$.

Para contrastar la bondad del ajuste global del modelo cuando el número de observaciones n_q en cada combinación de valores de las variables explicativas es grande se dispone del estadístico chi-cuadrado de Pearson y del estadístico de Wilks de razón de verosimilitudes. Cuando n_q no es suficientemente grande se usará el estadístico de Hosmer y Lemeshow que es una versión modificada del estadístico chi-cuadrado de Pearson.

El test global de bondad de ajuste del modelo de regresión logística múltiple contrasta la hipótesis nula

$$H_0 : p_q = \frac{\exp\left(\sum_{r=0}^{R} \beta_r x_{qr}\right)}{1 + \exp\left(\sum_{r=0}^{R} \beta_r x_{qr}\right)} \quad \forall q = 1, \ldots, Q,$$

frente a la hipótesis alternativa

$$H_1 : p_q \neq \frac{\exp\left(\sum_{r=0}^{R} \beta_r x_{qr}\right)}{1 + \exp\left(\sum_{r=0}^{R} \beta_r x_{qr}\right)} \quad \text{para algún q.}$$

Test chi-cuadrado de Pearson

El estadístico chi-cuadrado de Pearson de bondad de ajuste a un modelo M de regresión logística de la forma (2.5) es

$$X^2(M) = \sum_{q=1}^{Q} \frac{(y_q - n_q\hat{p}_q)^2}{n_q\hat{p}_q(1-\hat{p}_q)} = \sum_{q=1}^{Q} \frac{n_q(y_q - \hat{m}_q)^2}{\hat{m}_q(n_q - \hat{m}_q)},$$

donde $\hat{p}_q$ es la estimación MV de p_q bajo el modelo M y $\hat{m}_q = n_q\hat{p}_q$ es la estimación de MV de los valores esperados $m_q = n_q p_q$.

Este estadístico se obtiene como la suma de los estadísticos chi-cuadrado de bondad de ajuste a cada una de las distribuciones $B(n_q, p_q)$ que generan a los datos muestrales bajo la hipótesis nula de que las probabilidades p_q verifiquen el modelo M. Es decir,

$$\begin{aligned} X^2(M) &= \sum_{q=1}^{Q} \left[\frac{(y_q - n_q\hat{p}_q)^2}{n_q\hat{p}_q} + \frac{((n_q - y_q) - (n_q - n_q\hat{p}_q))^2}{(n_q - n_q\hat{p}_q)} \right] \\ &= \sum_{q=1}^{Q} \frac{n_q(y_q - n_q\hat{p}_q)^2}{n_q\hat{p}_q(n_q - n_q\hat{p}_q)} = \sum_{q=1}^{Q} \frac{(y_q - n_q\hat{p}_q)^2}{n_q\hat{p}_q(1-\hat{p}_q)}. \end{aligned}$$

Este estadístico tiene distribución asintótica chi-cuadrado con grados de libertad obtenidos como la diferencia entre el número de parámetros p_q (transformaciones logit muestrales) y el número de parámetros independientes en el modelo. Es decir,

$$X^2(M) \underset{n_q \to \infty}{\overset{d}{\longrightarrow}} \chi^2_{Q-(R+1)} \ .$$

Por lo tanto, se rechazará la hipótesis nula H_0 al nivel de significación α cuando se verifica que

$$X^2(M)_{Obs} \geq \chi^2_{Q-(R+1);\alpha},$$

siendo $\chi^2_{Q-(R+1);\alpha}$ el cuantil de orden $(1-\alpha)$ que es el valor de la distribución $\chi^2_{Q-(R+1)}$ que acumula a su derecha probabilidad α. Equivalentemente, si se define el p-valor del contraste como la probabilidad acumulada a la derecha del valor observado

$$p - valor = P[X^2(M) \geq X^2(M)_{Obs}],$$

se rechazará la hipótesis nula con nivel de significación α cuando se verifique que $p - valor \leq \alpha$.

Test chi-cuadrado de razón de verosimilitudes

El estadístico de Wilks de razón de verosimilitudes para el contraste de bondad de ajuste de un modelo de regresión logística múltiple M es de la forma

$$G^2(M) = 2\left[\sum_{q=1}^{Q}(n_q - y_q)\ln\left(\frac{n_q - y_q}{n_q - \hat{m}_q}\right) + \sum_{q=1}^{Q} y_q \ln\left(\frac{y_q}{\hat{m}_q}\right)\right].$$

Recordemos que el estadístico de Wilks de razón de verosimilitudes se obtiene como menos dos veces el logaritmo del cociente entre el supremo de la verosimilitud bajo la hipótesis nula y el supremo de la verosimilitud en la población. Es decir, si denotamos por V a la función de verosimilitud de los datos

$$G^2(M) = -2\ln\left(\frac{\underset{\{p \text{ bajo } H_0\}}{Sup} V(p)}{\underset{\{p\}}{Sup} V(p)}\right) = 2\left[\underset{\{p\}}{Sup} L(p) - \underset{\{p \text{ bajo } H_0\}}{Sup} L(p)\right],$$

donde L es la log-verosimilitud. Ya se ha demostrado que

$$\underset{\{p \text{ bajo } H_0\}}{Sup} L(p) = L(\hat{p})$$

donde $\hat{p}$ es el vector cuyas componentes son las estimaciones MV, $\hat{p}_q$, de las probabilidades p_q bajo el modelo M. Por otro lado, derivando la log-verosimilitud respecto de cada p_q e igualando a cero, se demuestra fácilmente que

$$\underset{\{p\}}{Sup} L(p) = L(f)$$

donde f es el vector $Q \times 1$ cuyas componentes son las proporciones observadas de respuesta $Y = 1$ en cada combinación de valores de las variables explicativas, definidas por $f_q = y_q/n_q$. Observemos que $L(f)$ corresponde al máximo de la log-verosimilitud bajo el *modelo saturado* que es aquel que se ajusta perfectamente a los datos (las frecuencias esperadas de respuesta $Y = 1$ bajo este modelo coinciden con las observadas, $\hat{m}_q = n_q f_q = y_q$) y tiene tantos parámetros libres como observaciones diferentes de las variables explicativas (transformaciones logit diferentes en nuestro caso). Un ejemplo simple de modelo saturado es el de regresión lineal simple con sólo dos datos.

Para simplificar denotaremos al máximo de la log-verosimilitud bajo un modelo por L_M. En particular, L_S representará el máximo de la log-verosimilitud bajo el modelo saturado.

Recordemos que la log-verosimilitud de los datos es

$$L(p) = \sum_{q=1}^{Q} \ln \left(\begin{array}{c} n_q \\ y_q \end{array} \right) + \sum_{q=1}^{Q} y_q \ln (p_q) + \sum_{q=1}^{Q} (n_q - y_q) \ln (1 - p_q).$$

Por lo tanto, operando adecuadamente se tiene la expresión del estadístico $G^2(M)$ en términos de las frecuencias esperadas bajo el modelo $\hat{m}_q = n_q \hat{p}_q$

$$\begin{aligned} G^2(M) &= 2(L_S - L_M) = 2(L(f) - L(\hat{p})) \\ &= 2 \left[\sum_{q=1}^{Q} (n_q - y_q) \ln \left(\frac{n_q - y_q}{n_q - \hat{m}_q} \right) + \sum_{q=1}^{Q} y_q \ln \left(\frac{y_q}{\hat{m}_q} \right) \right]. \end{aligned}$$

El estadístico de Wilks de razón de verosimilitudes tiene distribución asintótica chi-cuadrado con grados de libertad obtenidos como la diferencia entre la dimensión del espacio paramétrico y la dimensión de este espacio bajo la hipótesis nula. En el caso del modelo de regresión logística múltiple el número de grados de libertad es la diferencia entre el número de parámetros p_q y el número de parámetros β_r bajo el modelo. En total $Q - (R + 1)$ grados de libertad

$$G^2(M) \underset{n_q \to \infty}{\overset{d}{\longrightarrow}} \chi^2_{Q-(R+1)} \ .$$

Por lo tanto, se rechazará la hipótesis nula H_0 al nivel de significación α cuando se verifica que

$$G^2(M)_{Obs} \geq \chi^2_{Q-(R+1);\alpha},$$

o bien cuando el p-valor del contraste sea menor o igual que el nivel de significación

$$p - valor = P[G^2(M) \geq G^2(M)_{Obs}] \leq \alpha.$$

El estadístico $G^2(M)$ suele ser llamado *devianza* y juega un papel similar al de la suma de los cuadrados de los residuos en regresión lineal.

Observemos que los estadísticos de bondad de ajuste de Pearson y de razón de verosimilitudes tienen la misma distribución asintótica chi-cuadrado con los mismos grados de libertad. Para que $X^2(M)$ y $G^2(M)$ tengan distribuciones chi-cuadrado aproximadas debe ocurrir que el número de observaciones n_q en cada nivel observado de las covariables (variables explicativas) sea grande. Por ello, en el caso de covariables continuas estos contrastes de bondad de ajuste no tienen distribución aproximada chi-cuadrado debido a que lo usual es obtener observaciones distintas de las variables explicativas

para cada individuo de la muestra. La teoría asintótica se aplica de forma más natural a modelos logísticos con variables explicativas categóricas. En el caso de variables continuas los valores observados se agrupan en intervalos que definan una partición del rango de valores de las variables explicativas dando lugar al test de Hosmer y Lemeshow que se presenta a continuación.

Test de Hosmer y Lemeshow

Cuando no hay un número suficiente de observaciones n_q en cada combinación de valores x_q de las variables explicativas, no se puede asumir la distribución chi-cuadrado de los estadísticos de Pearson y de razón de verosimilitudes como buena. La norma para poder usar estos contrastes es que el 80 % de las frecuencias estimadas bajo el modelo, $\hat{m}_q = n_q\hat{p}_q$, sean mayores que cinco y todas mayores que uno.

El estadístico de Hosmer y Lemeshow es el estadístico chi-cuadrado de Pearson de bondad de ajuste al modelo después de agrupar adecuadamente los datos en intervalos, de modo que su valor depende fuertemente del número de clases resultantes de la agrupación.

Supongamos que agrupamos las variables explicativas obteniendo como resultado G grupos o clases, denotando por n'_g al número total de observaciones en el g-ésimo grupo, por u_g al número de respuestas $Y = 1$ en el g-ésimo grupo, y por $\bar{p}_g$ a la probabilidad estimada bajo el modelo de respuesta $Y = 1$ para el g-ésimo grupo obtenida como la media de las probabilidades $\hat{p}_q$ de los valores x_q de dicho grupo. Entonces, el estadístico de Hosmer y Lemeshow es de la forma

$$C = \sum_{g=1}^{G} \frac{(u_g - n'_g\bar{p}_g)^2}{n'_g\bar{p}_g(1-\bar{p}_g)},$$

y tiene también distribución asintótica chi-cuadrado con $G - 2$ grados de libertad.

El problema en la aplicación práctica de este estadístico es la selección de las clases. Se sabe que si el número de clases G es menor que seis, este contraste lleva casi siempre a aceptar el modelo como adecuado (potencia muy baja). Sin embargo usar muchos grupos puede llevar a tamaños muestrales pequeños en cada grupo. Hosmer y Lemeshow (1985) aconsejan $G = 10$ construyendo las clases en base a los deciles de las probabilidades estimadas bajo el modelo.

Medidas globales de bondad de ajuste

Para cuantificar la bondad del ajuste global del modelo se dispone de medidas como la *Tasa de clasificaciones correctas* (CCR) y medidas *R-cuadrado*

alternativas al coeficiente R^2 de la regresión lineal.

Tasa de clasificaciones correctas

La tasa de clasificaciones correctas es la proporción de individuos clasificados correctamente por el modelo obtenida como el cociente entre el número de observaciones clasificadas correctamente y el tamaño muestral N.

Un individuo es clasificado correctamente por el modelo logit cuanto su valor observado de respuesta (1 o 0) coincide con su valor estimado por el modelo. Para asignar respuesta $Y = 1$ o $Y = 0$ bajo el modelo a los datos se elige un punto de corte (cut-off), $p \in (0, 1)$, de modo que a una observación con valor $X = x_q$ se le estima respuesta $Y = 1$ si $\hat{p}_q \geq p$ y se le estima respuesta $Y = 0$ cuando $\hat{p}_q < p$.

Como punto de corte para clasificar en unos y ceros a las observaciones se suele elegir 0.5 aunque es más apropiado elegir la proporción de unos en la muestra. También se puede calcular la tasa de clasificaciones correctas para una partición muy fina de puntos de corte en (0,1) de modo que el estadístico pueda elegir como más apropiado el punto de corte que proporciona la máxima tasa de clasificaciones correctas.

Medidas tipo R^2

Es conocido que en regresión R^2 da la reducción proporcional en varianza entre la varianza condicional de la respuesta y la varianza marginal, describiendo la magnitud de la asociación entre el predictor lineal y la respuesta. Para el caso de datos de respuesta categórica se han definido medidas análogas a R^2 pero ninguna de ellas es tan útil como este coeficiente. Siguiendo esta filosofía el coeficiente R^2 de bondad de ajuste global del modelo de regresión logística se podría definir como la siguiente medida de reducción proporcional en el error:

$$1 - \frac{\sum_{q=1}^{Q}(y_q - n_q\hat{p}_q)^2}{\sum_{q=1}^{Q}(y_q - \bar{y})^2}$$

que en el caso de ajustar el modelo de probabilidad lineal por mínimos cuadrados, coincide con el cuadrado del coeficiente de correlación lineal entre $\{y_q\}$ y $\{\hat{m}_q\}$.

Una desventaja importante de esta última medida es que no tiene en cuenta en la estructura del error la dependencia de la varianza de Y_q respecto de p_q. Además, dado un modelo y un conjunto de datos, los valores estimados de los parámetros que maximizan esta medida no son los estimadores MV ni tampoco estimadores eficientes. Además, al calcularla con estimadores MV,

esta medida puede decrecer al añadir una variable explicativa al modelo. También se ha demostrado que este coeficiente puede llegar a tomar valores pequeños cuando el ajuste es casi perfecto. Estos problemas se podrían corregir definiendo una medida que pondere las desviaciones cuadráticas por el inverso de las varianzas estimadas pero la medida resultante es sin duda más complicada de cara a la interpretación.

Teniendo en cuenta que en el caso del modelo de regresión lineal el coeficiente de determinación se podía definir de forma equivalente a partir del cociente entre el máximo de la verosimilitud bajo el modelo nulo dado sólo por un término constante (V_0) y el máximo de la verosimilitud bajo el modelo ajustado con todos los parámetros (V_M), en regresión logística se calcula R^2 de la siguiente forma (ver Ryan (1997)):

$$R^2_{CN} = 1 - \left(\frac{V_0}{V_M}\right)^{\frac{2}{N}},$$

que recibe el nombre de R^2 de Cox y Snell.

Al aumentar el número de parámetros de un modelo, aumenta el máximo de la verosimilitud, y por otro lado, como las probabilidades están entre cero y uno, la medida R^2 toma valores entre 0 y 1.

Aunque esté acotada entre cero y uno, la medida R^2 así calculada no toma necesariamente el valor uno como máximo sino

$$\text{máx}R^2_{CN} = 1 - (V_0)^{\frac{2}{N}}.$$

Además este valor puede ser próximo a cero cuando hay pocos datos por lo que se propone como medida de bondad de ajuste el siguiente coeficiente de determinación ajustado:

$$R^2_N = \frac{R^2_{CN}}{\text{máx}R^2_{CN}},$$

que recibe el nombre de R^2 de Nagelkerke.

2.5.2. Contrastes sobre los parámetros del modelo

En este apartado pretendemos contrastar si un subconjunto de los parámetros β_r del modelo de regresión logística, que denotaremos por $\gamma = (\gamma_1, \ldots, \gamma_l)'$, es nulo. Por lo tanto la hipótesis nula del contraste es

$$H_0 : \gamma = 0.$$

Contrastes de Wald

Están basados en la normalidad asintótica de los estimadores de máxima verosimilitud.

Consideremos el contraste de hipótesis

$$\begin{aligned} H_0 : &\quad \gamma = 0 \\ H_1 : &\quad \gamma \neq 0. \end{aligned}$$

El estimador MV de γ, denotado por $\hat{\gamma}$, tiene distribución normal asintótica de media γ y matriz de covarianzas estimada $\widehat{Cov}(\hat{\gamma})$ obtenida de forma inmediata a partir de la matriz de covarianzas estimada para todos los parámetros $\widehat{Cov}(\hat{\beta})$ correspondiente a la expresión 2.10.

Como consecuencia el estadístico de Wald de este contraste es la siguiente forma cuadrática

$$\hat{\gamma}'[\widehat{Cov}(\hat{\gamma})]^{-1}\hat{\gamma},$$

que bajo la hipótesis nula tiene distribución chi-cuadrado asintótica con l grados de libertad (número de parámetros nulos bajo la hipótesis nula).

Por lo tanto, se rechazará la hipótesis nula al nivel de significación α cuando el valor observado de este estadístico sea mayor o igual que el cuantil de orden $(1-\alpha)$ de la distribución χ^2_l, denotado por $\chi^2_{l;\alpha}$.

Si en un determinado modelo M se quiere contrastar la igualdad a cero de un solo parámetro

$$\begin{aligned} H_0 : &\quad \beta_r = 0 \\ H_1 &\quad \beta_r \neq 0, \end{aligned}$$

el estadístico del contraste será

$$W = \frac{\hat{\beta}_r^2}{\hat{\sigma}^2(\hat{\beta}_r)},$$

que bajo la hipótesis nula tiene distribución chi-cuadrado asintótica con un grado de libertad, por ser el cuadrado de una normal estándar.

Por lo tanto se rechazará la hipótesis nula al nivel de significación α cuando se verifique que

$$W_{Obs} \geq \chi_{1;\alpha}.$$

Contrastes condicionales de razón de verosimilitudes

Supongamos que un modelo de regresión logística M_G se ajusta bien y queremos contrastar si un subconjunto de sus parámetros $\gamma = (\gamma_1, \ldots, \gamma_l)$

son nulos. Denotemos por M_P al modelo más simple que resulta al hacer cero estos parámetros en M_G, de modo que el modelo particular M_P está anidado en el modelo general M_G. Las hipótesis de este contraste se pueden expresar como

$$\begin{aligned} H_0: &\quad \gamma = 0 \ (M_P \ \text{ se verifica}) \\ H_1: &\quad \gamma \neq 0 \ \text{ asumiendo cierto } \ M_G. \end{aligned}$$

Asumiendo que M_G se verifica, el estadístico del test de razón de verosimilitudes para contrastar si M_P se verifica es de la forma:

$$\begin{aligned} G^2(M_P|M_G) &= -2(L_P - L_G) \\ &= -2(L_P - L_S) - (-2(L_G - L_S)) \\ &= G^2(M_P) - G^2(M_G), \end{aligned}$$

donde L_S, L_P y L_G son los máximos de la log-verosimilitud bajo la suposición de que se verifican los modelos saturado, M_P y M_G, respectivamente. Esto quiere decir que el test de razón de verosimilitudes para contrastar dos modelos anidados es la diferencia de los contrastes de razón de verosimilitudes de bondad de ajuste para cada modelo.

Además, el estadístico $G^2(M_P|M_G)$ tiene, bajo el modelo M_P (hipótesis nula), distribución chi-cuadrado con grados de libertad igual a la diferencia entre los grados de libertad de las distribuciones chi-cuadrado asintóticas de $G^2(M_P)$ y $G^2(M_G)$, que coincide con el número de parámetros que se anulan bajo H_0. En este caso,

$$G^2(M_P/M_G) \longrightarrow \chi^2_l.$$

Este estadístico tiene el mismo papel en regresión logística que el numerador del test F parcial en regresión lineal.

Igual que con el estadístico de Wald, se rechazará la hipótesis nula al nivel de significación α cuando el valor observado $G^2_{Obs}(M_P/M_G)$ sea mayor o igual que el cuantil de orden $(1-\alpha)$ de la distribución χ^2_l, denotado por $\chi^2_{l;\alpha}$.

Para el contraste sobre un parámetro de hipótesis nula $H_0 : \beta_r = 0$, se construye con la misma filosofía el estadístico $G^2(M_P/M_G)$ que en este caso tiene distribución asintótica χ^2_1.

Existen casos en los que el test de Wald no es tan potente como el test de razón de verosimilitudes, proporcionando a veces resultados no deseables. Por ello se aconseja usar el test de razón de verosimilitudes en los procedimientos de selección de variables.

Test score

El test score está basado en la distribución de las derivadas parciales de la log-verosimilitud y su principal ventaja es que reduce los cálculos con respecto al test de Wald y al test de razón de verosimilitudes. El uso de este test está muy limitado porque no es proporcionado por muchos paquetes estadísticos. Este contraste se puede utilizar para introducir variables en el procedimiento stepwise de selección forward.

Consideremos de nuevo el contraste de hipótesis

$$\begin{aligned} H_0 : &\quad \gamma = 0 \\ H_1 : &\quad \gamma \neq 0. \end{aligned}$$

Sea $\widetilde{\beta}$ el estimador de máxima verosimilitud MV de los parámetros del modelo (M_P) obtenido haciendo $\gamma = 0$ en el modelo completo (M_G), y $D(\widetilde{\beta})$ el vector de derivadas parciales de la log-verosimilitud evaluado en $\widetilde{\beta}$

$$D(\widetilde{\beta}) = X'(y - \widetilde{m}),$$

siendo $\widetilde{m}$ el vector de componentes $\widetilde{m}_q = n_q\widetilde{p}_q$, con $\widetilde{p}_q$ los estimadores MV de p_q obtenidos a partir de los $\widetilde{\beta}_q$.

El estadístico puntuación del contraste score es la siguiente forma cuadrática:

$$S = (D(\widetilde{\beta}))'(X'Diag[n_q\widetilde{p}_q(1 - \widetilde{p}_q)]X)^{-1}D(\widetilde{\beta}),$$

que bajo la hipótesis nula tiene distribución chi-cuadrado con l grados de libertad.

En el caso en que se quiere contrastar la igualdad a cero de un único parámetro ($\beta_r = 0$) el estadístico puntuación se calcula de igual forma siendo $\widetilde{\beta}$ el estimador MV de los parámetros del modelo resultante de hacer $\beta_r = 0$ en el modelo completo M_G. En este caso el estadístico del contraste score tiene bajo la hipótesis nula ($\beta_r = 0$) distribución chi-cuadrado con un grado de libertad.

2.5.3. Intervalos de confianza

En esta sección se construirán intervalos de confianza aproximados basados en la distribución normal asintótica de los estimadores MV.

Intervalos de confianza para los parámetros

Vamos a construir un intervalo de confianza (I.C.) aproximado con nivel de confianza $1-\alpha$ para cada uno de los parámetros β_r del modelo de regresión

logística múltiple. Para ello recordemos que la distribución asintótica de $\hat{\beta}_r$ es $N(\beta_r, \hat{\sigma}^2(\hat{\beta}_r))$.

Entonces se tiene que

$$P\left[-z_{\frac{\alpha}{2}} \leq \frac{\hat{\beta}_r - \beta_r}{\hat{\sigma}(\hat{\beta}_r)} \leq z_{\frac{\alpha}{2}}\right] = 1 - \alpha,$$

de donde se obtiene el siguiente intervalo de confianza aproximado para β_r al nivel $(1 - \alpha)$:

$$\hat{\beta}_r \pm z_{\frac{\alpha}{2}} \hat{\sigma}(\hat{\beta}_r).$$

Intervalos de confianza para las transformaciones logit

Sea la transformación logit

$$L(x) = \ln\left[\frac{p(x)}{1 - p(x)}\right] = x'\beta,$$

donde $p(x) = P[Y = 1/X = x]$ con $x = (1, x_1, \ldots, x_R)'$ y β el vector de parámetros del modelo de regresión logística múltiple.

Para construir un I.C. aproximado para $L(x)$ nos basaremos en la distribución normal asintótica de su estimador MV definido por

$$\hat{L}(x) = x'\hat{\beta}.$$

A partir de la distribución normal de $\hat{\beta}$ se obtiene que $\hat{L}(x)$ tiene distribución asintótica normal de media $L(x)$ y varianza estimada $\hat{\sigma}^2(\hat{L}(x)) = x'\widehat{Cov}(\hat{\beta})x$.

Como consecuencia, un I.C. aproximado para el logit poblacional $L(x)$ al nivel $1 - \alpha$ es de la forma

$$\hat{L}(x) \pm z_{\frac{\alpha}{2}} \hat{\sigma}(\hat{L}(x)).$$

Intervalos de confianza para las probabilidades de respuesta Y=1

Haciendo uso de la transformación: $\hat{p}(x) = \exp(\hat{L}(x))/[1 + \exp(\hat{L}(x))]$ se obtiene el siguiente I.C. aproximado al nivel de confianza $1 - \alpha$ para la probabilidad $p(x)$ de respuesta $Y = 1$ dado el valor $X = x$:

$$\frac{\exp\left(\hat{L}(x) \pm z_{\frac{\alpha}{2}} \hat{\sigma}(\hat{L}(x))\right)}{1 + \exp\left(\hat{L}(x) \pm z_{\frac{\alpha}{2}} \hat{\sigma}(\hat{L}(x))\right)}.$$

Intervalos de confianza para los cocientes de ventajas

En el caso del modelo de regresión logística múltiple se ha demostrado que los cocientes de ventajas de respuesta $Y = 1$ cuando se incrementa en una unidad cierta variable y se controlan las demás es de la forma

$$\theta(\Delta X_l = 1/X_r = x_r(r \neq l)) = e^{\beta_l} \quad \forall l = 1, \ldots, R.$$

En este caso tomando exponenciales en el intervalo obtenido para cada uno de los parámetros se obtienen los siguientes intervalos de confianza aproximados al nivel de confianza $1 - \alpha$ para dichos cocientes de ventajas

$$\exp\left(\hat{\beta}_r \pm z_{\frac{\alpha}{2}} \hat{\sigma}(\hat{\beta}_r)\right),$$

o equivalentemente

$$\hat{\theta}(\Delta X_l = 1/X_r = x_r(r \neq l)) \exp\left(\pm z_{\frac{\alpha}{2}} \hat{\sigma}(\hat{\beta}_r)\right).$$

2.6. Validación y diagnosis de modelos logit

Los estadísticos G^2 y X^2 son medidas de la calidad global del ajuste. Una vez contrastado que un modelo se ajusta globalmente bien, se procede a estudiar mediante medidas alternativas la bondad del ajuste observación a observación, así como la naturaleza de la falta de ajuste.

Los métodos gráficos suelen ser de mucha ayuda. Otra forma habitual de validar un modelo es el estudio de los residuos que comparan el número observado de éxitos, en cada combinación de valores de las variables explicativas, con su valor ajustado por el modelo. En el caso de modelos GLM el estudio de los residuos puede poner de manifiesto si la falta de ajuste se debe a una elección inapropiada de la ligadura o a la falta de linealidad en los efectos de las variables explicativas.

También se pueden detectar observaciones influyentes calculando residuos y aproximando el efecto que produce sobre los parámetros borrar observaciones simples.

2.6.1. Residuos

En base a los estadísticos X^2 y G^2 se definen dos tipos de residuos en cada combinación de valores x_q de las variables explicativas

1. *Residuos de Pearson o residuos estandarizados*

$$r_q = \frac{y_q - n_q \hat{p}_q}{[n_q \hat{p}_q (1 - \hat{p}_q)]^{1/2}}.$$

Observemos que el estadístico chi-cuadrado de Pearson se descompone como

$$X^2 = \sum_{q=1}^{Q} r_q^2.$$

Una vez estimados los residuos se contrasta su significación estadística mediante el test

$$\begin{aligned} H_0 : &\quad r_q = 0 \\ H_1 : &\quad r_q \neq 0. \end{aligned}$$

Bajo la hipótesis nula r_q tiene distribución asintótica normal con media cero y varianza estimada $\hat{\sigma}^2(r_q) < 1$. Esto significa que los residuos r_q tienen menor variabilidad que una v.a. normal estándar. A pesar de esto los residuos de Pearson r_q suelen ser tratados como normales estándar, considerándose significativos cuando sus valores absolutos son mayores que dos (falta de ajuste).

Para evitar este problema se definen los *residuos de Pearson ajustados* que tienen distribuciones asintóticas normales estándar

$$r_q^s = \frac{r_q}{(1 - h_{qq})^{1/2}},$$

donde h_{qq} es el elemento diagonal de la matriz

$$H = W^{\frac{1}{2}} X (X'WX)^{-1} X' W^{\frac{1}{2}},$$

con $W = Diag[n_q \hat{p}_q (1 - \hat{p}_q)]$. Observemos que la matriz H es la equivalente a la matriz *hat* del modelo de regresión lineal múltiple siendo h_{qq} la influencia (*leverage*) de la observación x_q.

Entonces se toma como estadístico del contraste

$$\begin{aligned} H_0 : &\quad r_q = 0 \\ H_1 : &\quad r_q \neq 0, \end{aligned}$$

r_q^s que bajo la hipótesis nula tiene distribución asintótica normal estándar, o bien su cuadrado que tiene distribución chi-cuadrado con un grado de libertad.

Finalmente se rechazará la hipótesis nula (residuo significativamente distinto de cero) al nivel de significación α cuando se verifique

$$|r_q^s| \geq z_{\alpha/2}.$$

2. *Residuos de la devianza o residuos estudentizados*

$$d_q = \text{signo}(y_q - \hat{m}_q)\left(2\left[y_q \ln\left(\frac{y_q}{\hat{m}_q}\right) + (n_q - y_q)\ln\left(\frac{n_q - y_q}{n_q - \hat{m}_q}\right)\right]\right)^{\frac{1}{2}}$$

Observemos que el estadístico chi-cuadrado de razón de verosimilitudes se descompone como

$$G^2 = \sum_{q=1}^{Q} d_q^2.$$

De nuevo, bajo la hipótesis nula $H_0 : d_q = 0$, el residuo d_q tiene distribución asintótica normal con media cero y varianza estimada $\hat{\sigma}^2(d_q) < 1$. Para evitar este problema se definen los *residuos de la devianza ajustados* o *estandarizados*

$$d_q^s = \frac{d_q}{(1 - h_{qq})^{1/2}},$$

que bajo la hipótesis nula del contraste

$$\begin{aligned} H_0 : &\quad d_q = 0 \\ H_1 : &\quad d_q \neq 0, \end{aligned}$$

tienen distribución asintótica normal estándar.

Por lo tanto, se rechazará la hipótesis nula (residuo significativamente distinto de cero) al nivel de significación α cuando se verifique

$$|d_q^s| \geq z_{\alpha/2}.$$

La diferencia entre ambos tipos de residuos es que los de la devianza convergen más rápidamente a la distribución normal que los de Pearson.

Como alternativa se pueden usar también los *residuos de la devianza modificados*

$$d_q^* = d_q + \frac{1 - 2\hat{p}_q}{(n_q \hat{p}_q(1 - \hat{p}_q))^{1/2}},$$

que bajo la hipótesis nula $H_0 : d_q = 0$ tienen distribución asintótica normal incluso cuando los tamaños muestrales n_q son pequeños.

2.6.2. Medidas de influencia

Igual que en el caso de la regresión lineal, pueden existir observaciones que se sitúen lejos del resto influyendo en las estimaciones de los parámetros del modelo. Las medidas de influencia permiten detectar dichos puntos influyentes estimando el cambio que se produce en los residuos cuando se eliminan observaciones influyentes.

En caso de que algún residuo resulte significativo, se estudiará su influencia sobre el ajuste del modelo mediante las *distancias de Cook*

$$D_q = \frac{1}{R+1}(r_q^s)^2 \left(\frac{h_{qq}}{1-h_{qq}}\right),$$

y las *distancias de Cook modificadas*

$$D_q^* = \left[\frac{N-(R+1)}{R+1}\frac{h_{qq}}{1-h_{qq}}\right]^{1/2} |r_q^{s*}|,$$

definiendo

$$r_q^{s*} = \frac{y_q - n_q\hat{p}_{(q)}}{(n_q\hat{p}_{(q)}(1-\hat{p}_{(q)})(1-h_{qq}))^{1/2}},$$

siendo $\hat{p}_{(q)}$ el estimador de p_q obtenido eliminando las observaciones para las que $X = x_q$.

Como medidas de influencia alternativas se puede calcula también el cambio en los coeficientes estimados $\hat{\beta}$ cuando se elimina por ejemplo el caso x_q. Estas medidas reciben el nombre de DFbetas y se definen en la forma

$$\Delta\beta_q = \frac{(X'WX)^{-1}x_q'(y_q - n\hat{p}_q)}{(1-h_{qq})}.$$

2.6.3. Métodos gráficos

Un procedimiento estándar para identificar la falta de ajuste en análisis de regresión es representar gráficamente los residuos frente a los valores predichos por el modelo ajustado, en nuestro caso las probabilidades estimadas. Esta representación debería dar lugar a una banda horizontal alrededor del cero.

Un gráfico equivalente consiste en representar gráficamente las observaciones frente a los valores predichos, en este caso las proporciones observadas de respuesta Y=1 frente a las ajustadas por el modelo. Este gráfico debería tener forma de banda alrededor de la recta de pendiente uno y constante cero. En el caso de la regresión logística con observaciones no repetidas, este

gráfico debería mostrar que valores predichos cerca de cero tienen la mayoría de sus observaciones iguales a cero; valores predichos cerca de uno tienen la mayoría de sus observaciones iguales a uno; y valores predichos próximos a 0.5 deberían de tener igual número de valores observados ceros y unos.

El gráfico anterior es muy difícil de interpretar de modo que lo que se suele hacer es dividir las probabilidades predichas en diez intervalos de amplitud 0.1 y estimar en cada intervalo el número esperado de respuestas Y=1 multiplicando el número de observaciones con probabilidades predichas en dicho intervalo por la marca de clase del intervalo. De este modo se podrían calcular los residuos de Pearson asociados a cada intervalo y sumando sus cuadrados se obtendría una medida resumen de la bondad del ajuste del modelo.

Curva Roc

Es una gráfica que permite evaluar la capacidad del modelo para discriminar. En el caso en que $Y = 1$ signifique padecer cierta enfermedad, el área bajo la curva ROC (*Receiver operating characteristic*) representa la probabilidad de que un individuo enfermo elegido al azar tenga mayor probabilidad estimada de padecer la enfermedad que un individuo no enfermo elegido también al azar. En la práctica este área es el porcentaje de pares de individuos enfermos y no enfermos en los que el enfermo tiene mayor probabilidad estimada de padecer la enfermedad que el no enfermo.

Para construir la curva ROC se dispone de un test de diagnóstico, en nuestro caso el modelo de regresión logística, que se usa para detectar si los individuos de una población tienen cierta característica (Y=1). Los test de diagnóstico son muy usados en medicina para detectar si un individuo padece cierta enfermedad. Un resultado positivo del test (T=+) predice que el individuo tiene la enfermedad y un resultado negativo (T=-) que no la tiene. Cuando el test de diagnóstico es el modelo de regresión logística ajustado, resultado positivo quiere decir que se le predice valor $\hat{Y} = 1$ porque su probabilidad predicha $\hat{p}$ es mayor o igual que el cut-point elegido para discriminar.

Para que un test de diagnóstico sea efectivo debe tener sensibilidad y especificidad altas. La sensibilidad del test es la probabilidad de que el test dé resultado positivo dado que el sujeto tiene la enfermedad $P[T = +/Y = 1]$. La especificidad es la probabilidad de que el test sea negativo dado que el sujeto no tiene la enfermedad $P[T = -/Y = 0]$. Los resultados del test se suelen representar en la siguiente *matriz de clasificación:*

	T=+	T=-
Y=1	A	B
Y=0	C	D

A partir de esta matriz de clasificación de los datos muestrales se definen la *tasa de verdaderos positivos* (sensibilidad muestral) como el cociente (A/(A+B)), y la *tasa de falsos positivos* (uno menos la especificidad muestral como el cociente (C/(C+D)). Entonces, la curva ROC se obtiene representando gráficamente la tasa de verdaderos positivos frente a la tasa de falsos positivos para distintos test de diagnóstico que corresponden a diferentes formas de definir un positivo, en nuestro caso distintos cut-point. También se puede elegir una partición muy fina de puntos de corte en el intervalo [0,1] y calcular el porcentaje de clasificaciones correctas y las tasas de verdaderos y falsos positivos en cada caso, de modo que podamos seleccionar el punto de corte más adecuado en función de estos parámetros.

Finalmente cuando el área bajo la curva ROC es al menos 0.7 el modelo logit ajustado se considera preciso con capacidad de discriminación alta.

2.7. Selección de modelos logit

Una vez conocido el procedimiento de ajuste de modelos logit, el siguiente paso es el desarrollo de estrategias para seleccionar las variables que mejor explican a la variable de respuesta binaria. Para ello se adoptará el principio de parsimonia que consiste en seleccionar el modelo que con menor número de parámetros se ajuste bien a los datos y lleve a una interpretación sencilla en términos de cocientes de ventajas.

El procedimiento que se utilizará en este curso para la inclusión o eliminación de variables es el de selección paso a paso (*stepwise*) tanto hacia adelante (*forward*) como hacia atrás (*backward*), que es el implementado en la mayoría de los paquetes estadísticos estándar.

A continuación describiremos brevemente el procedimiento de selección *stepwise forward-backward* basado en contrastes condicionales de razón de verosimilitudes aunque también se podrían usar test de Wald y test score. Este método consiste en partir de un modelo inicial de partida, de modo que en cada paso se ajustarán todos aquellos modelos logit que resultan de la inclusión en el modelo seleccionado en el paso anterior de cada una de las variables explicativas que no están en dicho modelo. Entonces se llevan a cabo contrastes condicionales de razón de verosimilitudes que tienen en la hipótesis nula el modelo seleccionado en el paso anterior y en la hipótesis alternativa el modelo resultante de la inclusión de cada variable. De este modo se seleccionarán

aquellas variables para las que este contraste es significativo(p-valor menor o igual que el nivel de significación α_1 fijado para la inclusión de variables), y se incluirá en el modelo aquella variable asociada al mínimo p-valor de entre todos los menores o iguales que α_1. La introducción de variables mediante este método continua hasta que ninguno de estos contrastes condicionales resulte significativo.

Por otro lado se considerará en cada paso la posibilidad de eliminar alguno de los parámetros del modelo seleccionado en el paso anterior(selección *backward*). Para evitar eliminar en un paso la variable que acaba de entrar en el anterior, se fijará para la eliminación de variables un nivel de significación α_2 mayor que el nivel de significación α_1 fijado para la inclusión de variables. Para la eliminación de variables se realizarán contrastes condicionales de razón de verosimilitudes que tienen en la hipótesis nula el modelo que resulta de la eliminación de cada una de las variables y en la hipótesis alternativa el modelo seleccionado en el paso anterior. De este modo se considerarán para la eliminación aquellas variables cuyo p-valor es mayor que α_2 y se eliminará la asociada al máximo de estos p-valores. La eliminación de variables continua hasta que todos estos contrastes condicionales resulten significativos.

Partiendo del modelo inicial que por defecto suele ser el modelo con la constante $L_q = \beta_0 \quad q = 1, \ldots, Q$, en el primer paso se ajustan todos los modelos logit simples que resultan de la inclusión en el modelo inicial de cada una de las variables explicativas ($L_q = \beta_0 + \beta_r x_{qr} \quad r = 1, \ldots, R.$) Una vez ajustados se realizan contrastes condicionales de razón de verosimilitudes de la forma

$$\begin{aligned} H_0 : &\qquad L_q = \beta_0 \\ H_1 : &\quad L_q = \beta_0 + \beta_r x_{qr}, \end{aligned}$$

para decidir que variables son candidatas a entrar en el modelo logit. Podrían entrar todas aquellas para las que este contraste condicional sea significativo (p-valor menor o igual que el nivel de significación α_1). Finalmente entrará aquella variable asociada al mínimo p-valor de entre todos los menores o iguales que α_1. Si la variable introducida en el paso 1 ha sido X_l, el modelo resultante es $L_q = \beta_0 + \beta_l x_{ql}$.

Por otro lado, en este primer paso sólo podría ser eliminada la constante del modelo pero por defecto suelen considerarse siempre modelos con término constante.

En el segundo paso, se parte del modelo seleccionado en el primer paso $L_q = \beta_0 + \beta_l X_{ql}$. Para decidir que variable se introduce en el modelo se

realizarán los contrastes condicionales de la forma

$$\begin{aligned} H_0 : &\qquad L_q = \beta_0 + \beta_l x_{ql} \\ H_1 : &\quad L_q = \beta_0 + \beta_l x_{ql} + \beta_r x_{qr} \quad r \neq l, \end{aligned}$$

introduciendo aquella variable X_m asociada al mínimo p-valor de entre los p-valores menores o iguales que el nivel de significación α_1.

Por otro lado la variable X_l que entró en el primer paso no se considera para salir en el segundo, tal y como se indicó al fijar los niveles de significación de entrada y salida en la forma $\alpha_2 > \alpha_1$. El único parámetro que podría salir en caso de considerarlo adecuado es el término constante en base al contraste condicional

$$\begin{aligned} H_0 : &\qquad L_q = \beta_l x_{ql} \\ H_1 : &\quad L_q = \beta_0 + \beta_l x_{ql}. \end{aligned}$$

Si el p-valor de este contraste es mayor que α_2 la constante saldría mientras que en caso contrario se quedaría en el modelo.

Así sucesivamente, el procedimiento *stepwise* continuará hasta llegar a un paso en el que ninguno de los contrastes condicionales de introducción de variables sean significativos y todos los de eliminación de variables sean significativos.