Parte II

REGRESIÓN LOGÍSTICA

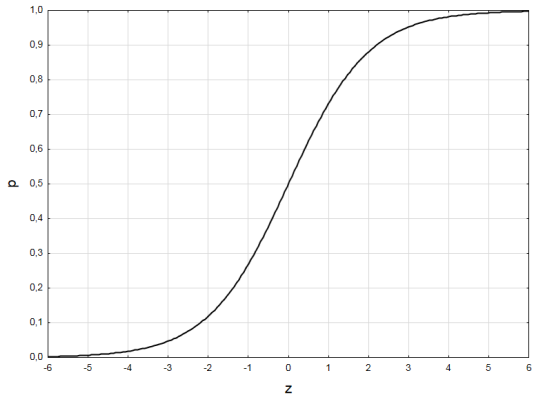
Capítulo 3

REGRESIÓN LOGÍSTICA

Consideramos una variable dependiente Y con distribución de Bernouilli (1=éxito, 0=fracaso) y un conjunto de variables independientes X_j con $j = 1, .., k$, las cuales pueden ser numéricas continuas (regresores) o variables dummy procedentes de la codificación de uno o más factores. Suponemos que, fijados unos valores x_j de las variables independientes, se verifica que

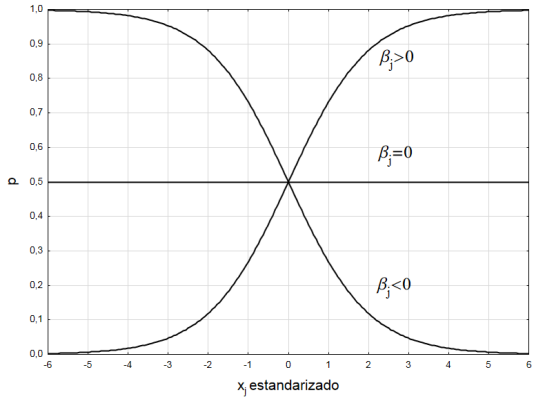
$$E(Y) = p\,(Y = 1) = p = \frac{1}{1 + e^{-z}}$$

siendo $z = \beta_0 + \sum_{j=1}^{k} \beta_j x_j$ para ciertos valores desconocidos β_j, con $j = 0, 1, ..., k$, y que pretendemos estimar. Observar que los valores de p en función de z constituyen una función estrictamente creciente, con un punto de inflexión en $z = 0$, $\lim_{z \to -\infty} f(z) = 0$ y $\lim_{z \to \infty} f(z) = 1$. Por tanto, todos los posibles valores de $p \in (0, 1)$ pueden ser obtenidos con esta función y además hay una correspondencia biunívoca entre valores de p y valores de z. Gráficamente, se tiene



Observar también que, si $\beta_j > 0$ entonces z es una función estrictamente creciente de x_j si las demás variables independientes se mantienen fijas, y por tanto, también p es una función estrictamente creciente de x_j. De modo similar, si $\beta_j < 0$ se tiene que p es una función estrictamente decreciente de x_j si las demás variables independientes se mantienen fijas. Finalmente, si $\beta_j = 0$ entonces el valor p no depende de la variable independiente X_j. Suponiendo que para el valor medio de la variable x_j es $p = 0.5$, gráficamente

se tiene:



Para poder cuantificar el efecto de cada uno de los parámetros β_j, con $j = 1, ..., k$, sobre la probabilidad de éxito p resulta conveniente definir, para cada valor p, la **odds** (ventaja) como $o = \dfrac{p}{1-p}$. De esta definición se deduce que la odds o expresa cuántas veces es mayor o menor la probabilidad de éxito p que la probabilidad de fracaso $1 - p$, es decir, en términos de probabilidad, la ventaja o desventaja del éxito frente al fracaso. Por ejemplo, si un suceso tiene probabilidad 0.8, la ventaja del éxito frente al fracaso es $0.8/0.2 = 4$, es decir el éxito es cuatro veces más probable que el fracaso o, en el lenguaje de los juegos de azar, las apuestas están 4 a 1 $(4 : 1)$ a favor del éxito. Por tanto, hablar de una probabilidad de éxito $p = 0.8$ es equivalente a hablar de una odds $o = 4$. Del mismo modo, si un suceso tiene probabilidad 0.2, la desventaja del éxito frente al fracaso es $0.2/0.8 = 1/4$, es decir, el éxito es cuatro veces menos probable que el fracaso, o las apuestas están 1 a 4 en contra del éxito. Por tanto, hablar de una probabilidad de éxito $p = 0.2$ es equivalente a hablar de una odds $o = 0.25$. Observar que, como la probabilidad p varía en el intervalo $(0, 1)$, entonces la odds o puede tomar cualquier valor del intervalo $(0, \infty)$.

Teniendo en cuenta la expresión propuesta para la probabilidad p en el modelo de regresión logística, es inmediato demostrar que, en términos de la odds o, el modelo puede formularse como

$$o = \frac{p}{1-p} = e^z$$

con $z = \beta_0 + \sum_{j=1}^{k} \beta_j x_j$. Por tanto, si $o\,(x_j)$ representa el valor actual de la odds y $o\,(x_j + \varepsilon)$ representa la nueva odds que se obtendría incrementando el valor de x_j en una cantidad ε y manteniendo constantes el resto de variables independientes, es fácil comprobar que $o\,(x_j + \varepsilon) = o\,(x_j)\ e^{\varepsilon\beta_j}$. Entonces, al cociente $\dfrac{o\,(x_j + \varepsilon)}{o\,(x_j)} = e^{\varepsilon\beta_j}$ se le denomina **odds-ratio** para una unidad de cambio igual a ε. En particular, si $\varepsilon = 1$, e^{β_j} es la odds-ratio para una unidad de cambio y representa el número por el que se multiplica la odds si x_j se incrementa en una unidad manteniendo constantes el resto de las variables independientes. En términos relativos,

$$\frac{o\,(x_j + 1) - o\,(x_j)}{o\,(x_j)} = e^{\beta_j} - 1$$

representa el cambio relativo en la odds si x_j se incrementa en una unidad manteniendo constantes el resto de las variables independientes.

Observar que, si $\beta_j > 0$ entonces $e^{\beta_j} > 1$, mientras que $e^{\beta_j} < 1$ si $\beta_j < 0$. Por ejemplo, si $\beta_j = 0.1$ entonces $e^{\beta_j} = 1.11$ y la odds se multiplica por 1.11 si x_j se incrementa en una unidad, es decir, que la

odds crece, en términos relativos, $e^{\beta_j} - 1 = 0.11$, lo que supone un 11 %. De modo similar, si $\beta_j = -0.1$ entonces $e^{\beta_j} = 0.90$ y la odds se multiplica por 0.90 si x_j se incrementa en una unidad, es decir, que la odds decrece, en términos relativos, $e^{\beta_j} - 1 = -0.10$, lo que supone un 10 %. Por tanto, las cantidades e^{β_j} y $e^{\beta_j} - 1$ nos permiten cuantificar el efecto del parámetro β_j sobre la odds en términos absolutos y en términos relativos respectivamente.

Finalmente, si para un p dado definimos el **logit** de p como $\ln\left(\frac{p}{1-p}\right)$, es evidente que el modelo de regresión logística puede formularse como

$$\ln\left(\frac{p}{1-p}\right) = z = \beta_0 + \sum_{j=1}^{k} \beta_j x_j$$

y por tanto el parámetro β_j representa el cambio lineal que se produce en el logit por cada unidad que aumentemos el valor de x_j manteniendo constantes el resto de los valores de las variables independientes.

Supongamos ahora que disponemos de una serie de observaciones $(y_i, x_{i1}, ..., x_{ik})$ con $i = 1, ..., n$ y queremos estimar los valores de los parámetros β_j, con $j = 0, 1, ..., k$. Para ciertos valores β_j de dichos parámetros podemos considerar los valores $z_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$, $p_i = \dfrac{1}{1 + e^{-z_i}}$ y observar que, si $y_i = 1$, entonces

$$p\,(Y = y_i) = p_i = \frac{1}{1 + e^{-z_i}} = \frac{e^{z_i}}{1 + e^{z_i}}$$

mientras que, si $y_i = 0$, entonces

$$p\,(Y = y_i) = 1 - p_i = 1 - \frac{e^{z_i}}{1 + e^{z_i}} = \frac{1}{1 + e^{z_i}}$$

Por tanto, para cualquier y_i, podemos escribir

$$p\,(Y = y_i) = \frac{e^{y_i z_i}}{1 + e^{z_i}}$$

Si consideramos el vector respuesta $\mathbf{y} = (y_1, ..., y_n)' \in \mathbb{R}^n$ con los valores observados de 0 y 1, la matriz $\mathbf{X} = (\mathbf{x_0}, \mathbf{x_1}, ..., \mathbf{x_k}) \in M_{n\times(k+1)}$ con $\mathbf{x_0} = \mathbf{1_n} = (1, ..., 1)' \in \mathbb{R}^n$ y $\mathbf{x}_j = (x_{1j}, ..., x_{nj})' \in \mathbb{R}^n$, el vector $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)' \in \mathbb{R}^{k+1}$, el vector $\mathbf{z} = (z_1, ..., z_n)' = \mathbf{X}\boldsymbol{\beta} \in \mathbb{R}^n$, el vector $\mathbf{p} = (p_1, ..., p_n)' \in \mathbb{R}^n$ con $p_i = \dfrac{1}{1 + e^{-z_i}}$ y el vector $\mathbf{0} = (0, ..., 0)' \in \mathbb{R}^n$, podemos calcular la verosimilitud de la muestra observada como

$$L\,(\boldsymbol{\beta}) = \prod_{i=1}^{n} p\,(Y = y_i) = \prod_{i=1}^{n} \frac{e^{y_i z_i}}{1 + e^{z_i}}$$

y parece lógico estimar los parámetros β_j de modo que la verosimilitud de la muestra sea máxima. Es decir, podemos plantear el problema matemático

$$\max_{\boldsymbol{\beta}} L\,(\boldsymbol{\beta}) = \max_{\boldsymbol{\beta}} \prod_{i=1}^{n} \frac{e^{y_i z_i}}{1 + e^{z_i}}$$

(**Nota**: Al igual que en regresión lineal múltiple, suponemos siempre que $rg(\mathbf{X}) = k + 1$ y ninguno de los vectores $\mathbf{x}_j$ es combinación lineal del resto. En regresión logística esto puede no ser cierto en muchas ocasiones, pero para explicar toda la teoría nos vamos a limitar a este caso).

Por tanto, para estimar los parámetros β_j vamos a utilizar el método de máxima verosimilitud. Ahora bien, es evidente que los valores β_j que maximizan $L(\boldsymbol{\beta})$ serán los mismos que minimizan la función auxiliar $\Lambda(\boldsymbol{\beta}) = -2\ln[L(\boldsymbol{\beta})]$, la cual siempre toma valores positivos por ser $0 < L(\boldsymbol{\beta}) < 1$. El problema puede formularse entonces como:

$$\min_{\boldsymbol{\beta}} \Lambda(\boldsymbol{\beta}) = \min_{\boldsymbol{\beta}} \{-2\ln[L(\boldsymbol{\beta})]\} = \min_{\boldsymbol{\beta}} \left\{-2\sum_{i=1}^{n}[y_i z_i - \ln(1+e^{z_i})]\right\}$$

Teniendo en cuenta que $\frac{\partial z_i}{\partial \beta_j} = x_{ij}$ para $j = 0, 1, ..., k$, podemos derivar la función objetivo respecto a cada uno de los β_j y obtenemos:

$$\frac{\partial \Lambda(\boldsymbol{\beta})}{\partial \beta_j} = \frac{\partial \Lambda(\boldsymbol{\beta})}{\partial z_i}\frac{\partial z_i}{\partial \beta_j} = -2\sum_{i=1}^{n} x_{ij}\left(y_i - \frac{e^{z_i}}{1+e^{z_i}}\right) = -2\sum_{i=1}^{n} x_{ij}\left(y_i - \frac{1}{1+e^{-z_i}}\right)$$

Por tanto, para encontrar la solución es necesario resolver el siguiente sistema de ecuaciones no lineales:

$$\left.\begin{array}{c} \sum_{i=1}^{n}\left(y_i - \frac{1}{1+e^{-z_i}}\right) = 0 \\ \sum_{i=1}^{n} x_{i1}\left(y_i - \frac{1}{1+e^{-z_i}}\right) = 0 \\ \vdots \\ \sum_{i=1}^{n} x_{ik}\left(y_i - \frac{1}{1+e^{-z_i}}\right) = 0 \end{array}\right\}$$

Matricialmente, utilizando los vectores definidos anteriormente, este sistema puede representarse como

$$\nabla[\Lambda(\boldsymbol{\beta})] = -2\mathbf{X}'(\mathbf{y} - \mathbf{p}) = \mathbf{0}$$

Además, teniendo en cuenta que

$$\frac{\partial p_i}{\partial \beta_j} = \frac{\partial p_i}{\partial z_i}\frac{\partial z_i}{\partial \beta_j} = \frac{e^{-z_i}}{(1+e^{-z_i})^2}x_{ij} = \left(\frac{1}{1+e^{-z_i}}\right)\left(1 - \frac{1}{1+e^{-z_i}}\right)x_{ij} = p_i(1-p_i)x_{ij}$$

podemos calcular también las derivadas parciales segundas de la función auxiliar $\Lambda(\boldsymbol{\beta})$ como

$$\frac{\partial^2 \Lambda(\boldsymbol{\beta})}{\partial \beta_{j_1}\partial \beta_{j_2}} = \sum_{i=1}^{n}\frac{\partial\left(\frac{\partial \Lambda(\boldsymbol{\beta})}{\partial \beta_{j_1}}\right)}{\partial p_i}\frac{\partial p_i}{\partial \beta_{j_2}} = 2\sum_{i=1}^{n} x_{ij_1}p_i(1-p_i)x_{ij_2}$$

y por tanto la matriz hessiana de la función auxiliar $\Lambda(\boldsymbol{\beta})$ es

$$\mathbf{H}[\Lambda(\boldsymbol{\beta})] = \frac{d^2[\Lambda(\boldsymbol{\beta})]}{d\,\boldsymbol{\beta}^2} = 2\mathbf{X}'\mathbf{W}\mathbf{X} \in M_{(k+1)\times(k+1)}$$

donde $\mathbf{W} \in M_{n\times n}$ es la matriz diagonal con elementos $p_i(1-p_i) > 0$ en la diagonal principal. Observar que esta matriz siempre tiene inversa por ser $rg(\mathbf{X}'\mathbf{W}\mathbf{X}) = rg(\mathbf{X}'\mathbf{X}) = rg(\mathbf{X}) = k+1$. Además, es siempre simétrica y definida positiva.

Por tanto, para resolver el sistema de ecuaciones $\nabla[\Lambda(\boldsymbol{\beta})] = -2\mathbf{X}'(\mathbf{y} - \mathbf{p}) = \mathbf{0}$, podemos utilizar el método iterativo de Newton-Raphson a partir de la función de iteración

$$\boldsymbol{\beta}_m = \boldsymbol{\beta}_{m-1} - \mathbf{H}\left[\Lambda\left(\boldsymbol{\beta}_{m-1}\right)\right]^{-1}\nabla\left[\Lambda\left(\boldsymbol{\beta}_{m-1}\right)\right] = \boldsymbol{\beta}_{m-1} + \left(\mathbf{X}'\mathbf{W}_{m-1}\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{y} - \mathbf{p}_{m-1}\right)$$

con un cierto vector inicial de partida $\boldsymbol{\beta}_0$, $\mathbf{p}_0$ obtenido a partir de $\boldsymbol{\beta}_0$ con las expresiones $p_i = \dfrac{1}{1+e^{-z_i}}$ y $z_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$, y $\mathbf{W}_0$ la matriz diagonal con elementos $p_i(1-p_i)$ en la diagonal principal.

Una vez alcanzada la convergencia, podemos calcular el vector de parámetros estimados del modelo $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k\right)' \in \mathbb{R}^{k+1}$, el vector $\hat{\mathbf{z}} = (\hat{z}_1, ..., \hat{z}_n)' \in \mathbb{R}^n$ con los valores estimados $\hat{z}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{ij}$, el vector de probabilidades predichas $\hat{\mathbf{p}} = (\hat{p}_1, ..., \hat{p}_n)' \in \mathbb{R}^n$ con $\hat{p}_i = \dfrac{1}{1+e^{-\hat{z}_i}}$, la matriz diagonal final $\hat{\mathbf{W}} \in M_{n\times n}$ con elementos $\hat{p}_i(1-\hat{p}_i)$ en la diagonal principal, el valor mínimo $\Lambda_f = \Lambda\left(\hat{\boldsymbol{\beta}}\right)$ de la función auxiliar $\Lambda(\boldsymbol{\beta})$ (denominado **deviance**, o desviación en castellano) y la máxima verosimilitud $L_f = L\left(\hat{\boldsymbol{\beta}}\right) = e^{-0.5\Lambda_f}$.

La matriz de varianzas-covarianzas del vector $\hat{\boldsymbol{\beta}}$ puede estimarse como

$$\mathbf{B} = Cov\left(\hat{\boldsymbol{\beta}}\right) = \left(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X}\right)^{-1} \in M_{(k+1)\times(k+1)}$$

y por tanto el error estándar de $\hat{\beta}_j$ puede calcularse como la raíz cuadrada del elemento diagonal j-ésimo de dicha matriz, es decir,

$$s.e.\left(\hat{\beta}_j\right) = \sqrt{b_{jj}} \text{ para } j = 0, 1, ..., k$$

Además, teniendo en cuenta que $\hat{\mathbf{z}} = (\hat{z}_1, ..., \hat{z}_n)' = \mathbf{X}\hat{\boldsymbol{\beta}}$, podemos calcular también la matriz de varianzas-covarianzas del vector $\hat{\mathbf{z}}$ como

$$\mathbf{C} = Cov\left(\hat{\mathbf{z}}\right) = \mathbf{X}\left(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}' \in M_{n\times n}$$

y por tanto el error estándar de $\hat{z}_i$ es

$$s.e.\left(\hat{z}_i\right) = \sqrt{c_{ii}} \text{ para } i = 1, ..., n$$

siendo c_{ii} el elemento diagonal i-ésimo de la matriz $\mathbf{C}$.

Es interesante considerar el caso especial de que no existan variables independientes en el modelo, es decir, el caso especial $\beta_j = 0$ para $j = 1, ..., k$ y por tanto $z = \beta_0$. A este modelo se le denomina el **modelo nulo**. En esta situación se tendría $z_i = \beta_0$ para $i = 1, ..., n$ y la única ecuación del sistema anterior sería

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

Si denotamos por $\hat{p}_0$ a la proporción de éxitos en la muestra, esta ecuación puede escribirse como

$$\frac{1}{1+e^{-\beta_0}} = \bar{y} = \hat{p}_0$$

y por tanto la solución sería

$$\hat{\beta}_0 = \ln\left(\frac{\hat{p}_0}{1-\hat{p}_0}\right)$$

(este valor $\hat{\beta}_0$, junto con $\beta_j = 0$ para $j = 1, ..., k$, suele utilizarse como punto inicial en el proceso iterativo de Newton-Raphson para la búsqueda de la solución, es decir, $\boldsymbol{\beta}_0 = (\hat{\beta}_0, 0, ..., 0)' \in \mathbb{R}^{k+1}$).

Para este modelo nulo la máxima verosimilitud sería

$$L_0 = L\left(\hat{\beta}_0\right) = (\hat{p}_0)^{\sum_{i=1}^{n} y_i} (1-\hat{p}_0)^{n-\sum_{i=1}^{n} y_i} = (\hat{p}_0)^{n(\hat{p}_0)} (1-\hat{p}_0)^{n(1-\hat{p}_0)}$$

y el mínimo valor de la función auxiliar Λ sería

$$\Lambda_0 = -2n\left[\hat{p}_0 \ln(\hat{p}_0) + (1-\hat{p}_0)\ln(1-\hat{p}_0)\right]$$

Se puede demostrar entonces que, bajo la hipótesis nula $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$, se tiene

$$\Lambda_0 - \Lambda_f \rightsquigarrow \chi^2_k$$

y podemos construir un test para contrastar dicha hipótesis nula, denominado test global del modelo de regresión logística. Este test también recibe el nombre de **test de la razón de verosimilitud**, por el hecho de que

$$\Lambda_0 - \Lambda_f = -2\ln\left(\frac{L_0}{L_f}\right)$$

y $\dfrac{L_0}{L_f}$ es el cociente de las verosimilitudes. En concreto, si λ_f y λ_0 son los valores observados para el modelo inicial y el modelo nulo respectivamente, el p-valor para el test global del modelo será

$$\text{p-valor} = p\left(\chi^2_k > \lambda_0 - \lambda_f\right)$$

Con un gran número de variables en el modelo (k muy grande) podemos hacer que λ_f se acerque a cero artificialmente y por ello conviene introducir una nueva medida que tenga en cuenta el número de variables en el modelo y nos permita comparar modelos con diferente número de variables. Así, se definen el **criterio de información de Akaike AIC** como

$$AIC = \Lambda_f + 2(k+1)$$

donde $k+1$ es el número de parámetros en el modelo, o el **criterio de información de Schwarz SC** como

$$SC = \Lambda_f + (k+1)\ln n$$

donde n es el número de puntos en el modelo. En general, se prefieren modelos con valores más bajos de AIC o SC.

Un segundo test global del modelo para contrastar la hipótesis nula $H_0 : \beta_j = 0$ para todo $j = 1, ..., k$ puede obtenerse con el estadístico **Score** (o puntuación en castellano) definido como

$$\text{Score} = (\mathbf{y}-\mathbf{p}_0)'\,\mathbf{X}\,(\mathbf{X}'\mathbf{W}_0\mathbf{X})^{-1}\,\mathbf{X}'\,(\mathbf{y}-\mathbf{p}_0)$$

siendo $\mathbf{p}_0$ la solución del modelo nulo, es decir, $\mathbf{p}_0 = (\hat{p}_0, ..., \hat{p}_0)' = \hat{p}_0\mathbf{1_n} \in \mathbb{R}^n$ con $\hat{p}_0$ igual a la proporción de éxitos en la muestra, y $\mathbf{W}_0 \in M_{n\times n}$ la matriz diagonal con elementos $\hat{p}_0(1-\hat{p}_0)$ en la diagonal principal, es decir, $\mathbf{W}_0 = \hat{p}_0(1-\hat{p}_0)\mathbf{I_{n\times n}}$. Bajo la citada hipótesis nula, este estadístico tiene una distribución de probabilidad χ^2_k, es decir,

$$\text{Score} = \frac{(\mathbf{y}-\mathbf{p}_0)'\,\mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\,(\mathbf{y}-\mathbf{p}_0)}{\hat{p}_0(1-\hat{p}_0)} \rightsquigarrow \chi^2_k$$

y por tanto podemos construir un nuevo test global del modelo, denominado **test Score**, que es una alternativa al test de la razón de verosimilitud.

Además, considerando el vector $\tilde{\boldsymbol{\beta}} = \left(\hat{\beta}_1, ..., \hat{\beta}_k\right) \in \mathbb{R}^k$ y la matriz $Cov\left(\tilde{\boldsymbol{\beta}}\right) \in M_{k\times k}$ que se obtiene eliminado la primera fila y la primera columna de la matriz $Cov\left(\hat{\boldsymbol{\beta}}\right) = \left(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X}\right)^{-1} \in M_{(k+1)\times(k+1)}$, un tercer test global del modelo que contrasta la hipótesis nula $H_0 : \beta_j = 0$ para todo $j = 1, ..., k$, es el denominado **test de Wald**, que utiliza el estadístico

$$\text{Wald} = \tilde{\boldsymbol{\beta}}' \left[Cov\left(\tilde{\boldsymbol{\beta}}\right)\right]^{-1} \tilde{\boldsymbol{\beta}}$$

cuya distribución de probabilidad bajo la hipótesis nula H_0 es también una χ^2_k, es decir,

$$\text{Wald} = \tilde{\boldsymbol{\beta}}' \left[Cov\left(\tilde{\boldsymbol{\beta}}\right)\right]^{-1} \tilde{\boldsymbol{\beta}} \rightsquigarrow \chi^2_k$$

Una vez que hemos obtenido la solución del modelo y hemos testado su significación global, es necesario evaluar la calidad del ajuste. Para ello conviene observar que $0 < \Lambda_f < \Lambda_0$, $L_0 = e^{-0.5\Lambda_0}$, $L_f = e^{-0.5\Lambda_f}$ y por tanto $L_0 < L_f < 1$. La calidad de ajuste del modelo será mayor cuanto más se acerque el valor de L_f a 1, es decir, cuando la verosimilitud del modelo esté próxima a 1. Entonces, imitando a lo que se hace en el modelo lineal general, Cox-Snell (1989, pp. 208-209) han propuesto el siguiente coeficiente de determinación generalizado o **pseudo-R^2 de Cox-Snell**

$$R^2 = 1 - \left(\frac{L_0}{L_f}\right)^{\frac{2}{n}}$$

de modo que si L_f está próximo a L_0 entonces $R^2 \approx 0$ y cuando L_f se acerca a 1 se obtienen valores más grandes con $0 < R^2 < 1 - (L_0)^{\frac{2}{n}} < 1$. Es decir, cuanto mayor sea el coeficiente pseudo-R^2 de Cox-Snell mayor será la calidad de ajuste del modelo. Ahora bien, teniendo en cuenta que nunca se puede alcanzar el valor 1, conviene ajustar dicho coeficiente en relación a su máximo valor $R^2_{\text{máx}} = 1 - (L_0)^{\frac{2}{n}}$. Por ello Nagelkerke (1991) propuso el siguiente coeficiente **pseudo-R^2 de Nagelkerke** ajustado

$$\tilde{R}^2 = \frac{R^2}{R^2_{\text{máx}}}$$

de modo que ahora $0 < \tilde{R}^2 < 1$ y este coeficiente puede interpretarse como el habitual coeficiente de determinación en el modelo lineal general.

Resulta también interesante diseñar test de hipótesis e intervalos de confianza para cada uno de los coeficientes β_j del modelo de regresión logística . Para ello, si eliminásemos del modelo la variable X_j (es decir, si $\beta_j = 0$), podríamos volver a resolver el modelo, denotando por Λ_{-j} al mínimo valor de la función auxiliar $\Lambda(\boldsymbol{\beta})$, con $0 < \Lambda_f < \Lambda_{-j}$, y por $L_{-j} = e^{-0.5\Lambda_{-j}}$ a la correspondiente verosimilitud del modelo. Entonces, de modo similar que para $\Lambda_0 - \Lambda_f$, se verifica que, bajo la hipótesis nula $H_0 : \beta_j = 0$, se tiene

$$\Lambda_{-j} - \Lambda_f = -2\ln\left(\frac{L_{-j}}{L_f}\right) \rightsquigarrow \chi^2_1$$

y, si denotamos por λ_{-j} al valor observado para Λ_{-j}, podemos calcular el p-valor del test $H_0 : \beta_j = 0$ como

$$\text{p-valor} = p\left(\chi^2_1 > \lambda_{-j} - \lambda_f\right)$$

Otra forma de llevar a cabo el test $H_0 : \beta_j = 0$ anterior es el denominado **test chi-cuadrado de Wald**, que consiste en utilizar el estadístico

$$\left(\frac{\hat{\beta}_j}{s.e.\left(\hat{\beta}_j\right)}\right)^2 = \frac{\left(\hat{\beta}_j\right)^2}{b_{jj}} \rightsquigarrow \chi^2_1$$

Habitualmente los paquetes de software suelen utilizar este último test en vez del anterior basado en el cociente de verosimilitudes $\frac{L_{-j}}{L_f}$, porque el cálculo de Λ_{-j} requiere el ajuste de un nuevo modelo que no contenga a la variable X_j.

Para calcular intervalos de confianza de los parámetros β_j del modelo suele recurrirse a la normalidad asintótica de los estimadores $\hat{\beta}_j$, es decir,

$$\frac{\hat{\beta}_j - \beta_j}{s.e.\left(\hat{\beta}_j\right)} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{b_{jj}}} \tilde{\rightsquigarrow} N(0,1)$$

y por tanto se obtienen los siguientes intervalos, denominados **intervalos de confianza de Wald** para los parámetros

$$\hat{\beta}_j \pm z_{\alpha/2}\ \sqrt{b_{jj}}$$

Pueden calcularse también otros intervalos de confianza alternativos basados en la razón de verosimilitudes, más complicados de evaluar porque se requieren procedimientos iterativos, y que nosotros no estudiaremos.

Utilizando estos intervalos pueden calcularse también estimaciones puntuales e intervalos de confianza para las odds-ratio con una magnitud del cambio igual ε. En concreto, la estimación puntual será $e^{\varepsilon\hat{\beta}_j}$ y el intervalo de confianza será

$$\left(e^{\varepsilon\left(\hat{\beta}_j - z_{\alpha/2}\ \sqrt{b_{jj}}\right)}, e^{\varepsilon\left(\hat{\beta}_j + z_{\alpha/2}\ \sqrt{b_{jj}}\right)}\right)$$

De modo similar, teniendo en cuenta la normalidad asintótica de las estimaciones $\hat{z}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{ij}$, pueden calcularse intervalos de confianza para los logit, z_i, de las observaciones muestrales. Específicamente,

$$\frac{\hat{z}_i - z_i}{s.e.\left(\hat{z}_i\right)} = \frac{\hat{z}_i - z_i}{\sqrt{c_{ii}}} \tilde{\rightsquigarrow} N(0,1)$$

y el intervalo de confianza será

$$\hat{z}_i \pm z_{\alpha/2}\ \sqrt{c_{ii}}$$

Por tanto, la predicción puntual para las probabilidades p_i de las observaciones muestrales serán $\hat{p}_i = \dfrac{1}{1+e^{-\hat{z}_i}}$ y los correspondientes intervalos de confianza serán

$$\left(\left[1+e^{-\hat{z}_i + z_{\alpha/2}\ \sqrt{c_{ii}}}\right]^{-1}, \left[1+e^{-\hat{z}_i - z_{\alpha/2}\ \sqrt{c_{ii}}}\right]^{-1}\right)$$

Por último, para una nueva observación $\mathbf{x}_h = (1, x_{h1}, ..., x_{hk})' \in \mathbb{R}^{k+1}$, las predicciones puntuales serán $\hat{z} = (\mathbf{x}_h)'\,\hat{\boldsymbol{\beta}}$ y $\hat{p} = \dfrac{1}{1+e^{-\hat{z}}}$, y los correspondientes intervalos de confianza serán

$$\hat{z} \pm z_{\alpha/2}\ \sqrt{(\mathbf{x}_h)'\left(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{x}_h}$$

y

$$\left(\left[1+e^{-\hat{z}+z_{\alpha/2}\ \sqrt{(\mathbf{x}_h)'\left(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{x}_h}}\right]^{-1},\left[1+e^{-\hat{z}-z_{\alpha/2}\ \sqrt{(\mathbf{x}_h)'\left(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{x}_h}}\right]^{-1}\right)$$

Comparando los valores observados con los predichos podemos evaluar también la calidad de ajuste del modelo. Para ello consideramos todos los pares de observaciones con un éxito y un fracaso, y denotamos por t al número total de pares, es decir

$$t=\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n}(1-y_i)\right)$$

Entonces diremos que uno de estos pares de observaciones es concordante si la observación con el valor 0 tiene una probabilidad predicha estrictamente menor; diremos que es discordante si es estrictamente mayor y diremos que hay empate si ambas probabilidades predichas son iguales. Podemos denotar entonces por n_c y n_d, respectivamente, al número de pares concordantes y discordantes, de modo que el número de empates será $t-n_c-n_d$. Por tanto, el **porcentaje de concordancia** del modelo es $\frac{n_c}{t}100$, el **porcentaje de discordancia** es $\frac{n_d}{t}100$ y el **porcentaje ligado** es $\frac{t-n_c-n_d}{t}100$. Valores altos del porcentaje de concordancia y valores bajos del porcentaje de discordancia indican una buena calidad de ajuste del modelo. Se define entonces un **índice c de calidad del ajuste** como

$$c=\frac{n_c+0.5\,(t-n_c-n_d)}{t}\in[0,1]$$

de modo que cuanto mayor sea c mejor es la calidad del ajuste. En la práctica es deseable que $c>0.7$.

Otros índices para evaluar la capacidad predictiva del modelo son:

$$\textbf{D de Sommers (o coeficiente de Gini)}=D=\frac{n_c-n_d}{t}$$

$$\textbf{Gamma de Goodman-Kruskal}=\gamma=\frac{n_c-n_d}{n_c+n_d}$$

$$\textbf{Tau-a de Kendall}=\tau_a=\frac{2\,(n_c-n_d)}{n(n-1)}$$

de modo que valores altos de estos índices implican una mayor capacidad predictiva. Observar que los tres índices están siempre en el intervalo $[-1,1]$ y, si no hay empates, entonces $D=\gamma=2c-1$.

Con el mismo propósito de evaluar la calidad del ajuste y la capacidad predictiva del modelo, para cada $p\in[0,1]$ y para cada y_i observado, podemos definir el valor predicho $\hat{y}_i(p)$ como

$$\hat{y}_i(p)=\begin{cases}0 & \text{si } \hat{p}_i<p\\ 1 & \text{si } \hat{p}_i\geq p\end{cases}$$

y diremos que el modelo predice positivo $(+)$ cuando $\hat{y}_i(p)=1$ y predice negativo $(-)$ cuando $\hat{y}_i(p)=0$.

A partir de estos valores predichos, podemos definir la **sensibilidad del modelo** como

$$S(p)=\frac{\sum_{i=1}^{n}\hat{y}_i(p)y_i}{\sum_{i=1}^{n}y_i}$$

y la **especificidad del modelo** como

$$E(p) = \frac{\sum_{i=1}^{n} [1 - \hat{y}_i(p)]\,(1 - y_i)}{\sum_{i=1}^{n} (1 - y_i)}$$

de modo que $S(p)$ representa la proporción de éxitos (1) correctamente clasificados como éxitos (+) para el nivel de probabilidad p, y $E(p)$ representa la proporción de fracasos (0) correctamente clasificados como fracasos (−) para el nivel de probabilidad p.

Por tanto, para cada nivel de probabilidad p, podríamos construir una tabla de contingencia 2×2 con los valores observados y los predichos del siguiente modo:

		Predichos		
		−	+	
Observados	0	$n_{0-} = \sum_{i=1}^{n} [1 - \hat{y}_i(p)]\,(1 - y_i)$	$n_{0+} = \sum_{i=1}^{n} \hat{y}_i(p)(1 - y_i)$	$n_0 = \sum_{i=1}^{n} (1 - y_i)$
	1	$n_{1-} = \sum_{i=1}^{n} [1 - \hat{y}_i(p)]\,y_i$	$n_{1+} = \sum_{i=1}^{n} \hat{y}_i(p) y_i$	$n_1 = \sum_{i=1}^{n} y_i$
		$n_- = \sum_{i=1}^{n} [1 - \hat{y}_i(p)]$	$n_+ = \sum_{i=1}^{n} \hat{y}_i(p)$	n

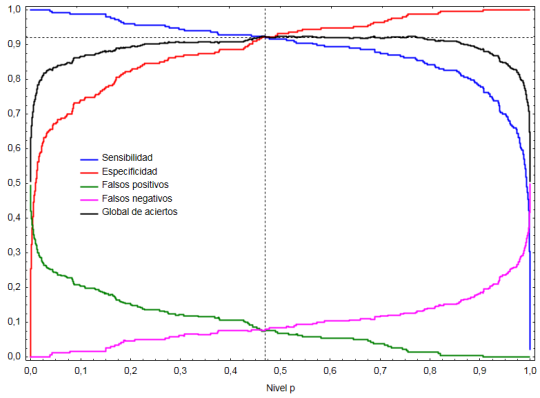
Observar que $S(p) = \dfrac{n_{1+}}{n_1}$ y por tanto la sensibilidad estima la probabilidad condicionada $p(+/1)$, es decir la probabilidad de acertar cuando el verdadero valor es un éxito. Del mismo modo, $E(p) = \dfrac{n_{0-}}{n_0}$ y la especificidad estima la probabilidad condicionada $p(-/0)$, es decir la probabilidad de acertar cuando el verdadero valor es un fracaso. Obviamente, en un modelo con buena calidad de ajuste tendremos valores altos (próximos a 1) para la sensibilidad y la especificidad. También la **proporción global de aciertos**, definida como el cociente $\dfrac{n_{0-} + n_{1+}}{n}$, debería ser próxima a 1 para tener una buena calidad de ajuste.

Además, podemos estimar la probabilidad condicionada $p(1/-)$ como $\dfrac{n_{1-}}{n_-}$ y este cociente recibe el nombre de **proporción de falsos negativos**. Del mismo modo, la probabilidad condicionada $p(0/+)$ puede estimarse como $\dfrac{n_{0+}}{n_+}$ y este cociente recibe el nombre de **proporción de falsos positivos**. Para que la calidad predictiva del modelo sea buena será necesario también valores bajos (próximos a 0) para las proporciones de falsos negativos y falsos positivos.

Es importante observar que estos cinco índices (proporción de aciertos, sensibilidad, especificidad, falsos positivos y falsos negativos) dependen del nivel de probabilidad p fijado para predecir éxito o fracaso, y por tanto pueden ser mejores o peores en función de dicho nivel. Por ello resulta interesante plantear el problema de la elección del nivel de probabilidad de corte p para que el modelo proporcione resultados más satisfactorios. A priori, puede parecer más objetivo elegir $p = 0.5$ como punto de corte, pero no siempre es ésta la mejor opción. Por ejemplo, si la muestra inicial utilizada para ajustar el modelo no ha sido dirigida, es decir, si el cociente $\dfrac{n_1}{n}$ estima la probabilidad global de éxito en la población, puede parecer también adecuado elegir como punto de corte $p = \dfrac{n_1}{n}$. O también, teniendo en cuenta que la función $S(p)$ es decreciente mientras que $E(p)$ es creciente, y queremos que ambos valores sean altos,

puede resultar adecuado elegir como punto de corte el valor p para el cual se verifica que $S(p) = E(p)$, es decir, el punto de corte entre la sensibilidad y la especificidad.

En el siguiente gráfico se muestran estos cinco índices, en función del nivel de probabilidad p, para una regresión logística con $n = 500$:

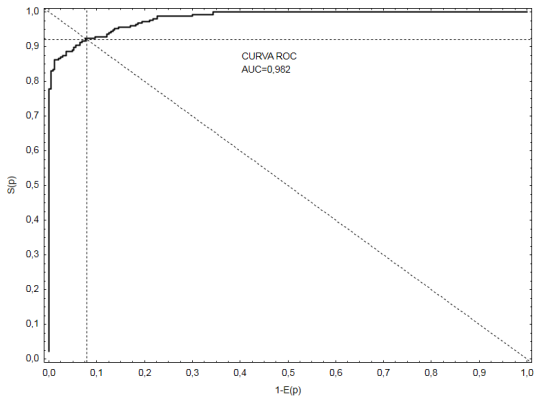


Para valorar globalmente la calidad de ajuste del modelo, sin depender de un nivel p de probabilidad de corte, se define la **curva ROC** (Receiver Operation Characteristic) del modelo de regresión logística como la función escalonada definida por los puntos $(1 - E(\hat{p}_i), S(\hat{p}_i))$, es decir, representando los valores de la sensibilidad en el eje de ordenadas y de uno menos la especificidad en el eje de abscisas. Esta curva siempre contiene al punto $(1, 1)$ (obtenido para el mínimo valor de $\hat{p}_i$), y suele completarse con el punto $(0, 0)$ que se obtendría para el punto de corte $p = 1$ (observar que en este caso todos los puntos serían clasificados como fracasos y por tanto la especificidad sería 1 y la sensibilidad sería 0). El modelo sería perfecto cuando, para cualquier nivel de probabilidad de corte p, la sensibilidad y la especificidad fuesen siempre igual a 1, es decir cuando la curva ROC fuese constantemente igual a 1 y por tanto el área encerrada bajo ella fuese también igual a 1. Es obvio que esta curva ideal no podrá alcanzarse en ningún caso, pero cuanto más se acerque a 1 el área encerrada bajo la curva ROC más cerca estará el modelo de la situación perfecta. Por tanto, un buen indicador global de la calidad de ajuste del modelo puede ser el área encerrada bajo la curva ROC, que suele representarse por **AUC**. Teniendo en cuenta que $1 - E(\hat{p}_i) + S(\hat{p}_i) > 0$ es claro que todos los puntos de la curva ROC están siempre por encima de la bisectriz del primer cuadrante y por tanto el área bajo la curva ROC estará siempre en el intervalo $(0.5, 1)$ y la calidad de ajuste del modelo será mejor cuanto más cerca esté del valor 1. Además, puede demostrarse que un buen estimador del valor AUC es siempre el índice de calidad del ajuste c definido con anterioridad, es decir, $AUC \approx c$.

La curva ROC definida anteriormente puede utilizarse también para elegir el nivel de probabilidad de corte p, teniendo en cuenta que el valor ideal para la curva ROC es el punto $(0, 1)$ que proporcionaría una sensibilidad y una especificidad iguales a 1. Por tanto, podríamos elegir como nivel de corte aquel valor $\hat{p}_i$ para el cual el punto de la curva ROC $(1 - E(\hat{p}_i), S(\hat{p}_i))$ esté lo más cerca posible del punto $(0, 1)$. Teniendo en cuenta que la diagonal del cuadrado unidad dentro del cual se encuentra la curva ROC y

que pasa por el punto $(0,1)$ es la recta $1 - E(p) + S(p) = 1$, este punto se obtendrá aproximadamente cuando $E(p) = S(p)$ y por tanto el criterio del punto de corte entre la sensibilidad y la especificidad del que hablamos antes queda reafirmado.

El siguiente gráfico muestra la curva ROC de la regresión logística con $n = 500$ citada anteriormente:



Terminaremos analizando la adecuación del modelo mediante la evaluación de la falta de ajuste. Para ello vamos a denotar por J al número de filas distintas que hay en la matriz X (es decir, el número de perfiles distintos de las variables independientes) y por m_j, con $j = 1, ..., J$, al número de puntos correspondientes a cada perfil. Supondremos que, como será habitual en la práctica, $J > k + 1$. Es evidente que para todos los puntos que tengan un mismo perfil la probabilidad predicha por el modelo es siempre la misma y vamos a denotarla por $\hat{\pi}_j$. Además, vamos a denotar por o_j al número de éxitos observados en cada perfil j, con $j = 1, ..., J$. Definimos entonces los **residuales de Pearson** del modelo como

$$r_j = \frac{o_j - m_j\hat{\pi}_j}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)}}$$

y a partir de ellos construimos el **estadistíco chi-cuadrado de Pearson** como

$$\chi^2 = \sum_{j=1}^{J} r_j^2 = \sum_{j=1}^{J} \frac{(o_j - m_j\hat{\pi}_j)^2}{m_j\hat{\pi}_j(1-\hat{\pi}_j)}$$

Si $J << n$ y el modelo ajustado es correcto, el anterior estadístico tiene una distribución χ^2 con $J - (k+1)$ grados de libertad y por tanto podemos construir un test para la falta de ajuste del modelo, denominado **test chi-cuadrado de Pearson**, utilizando el p-valor $p\left(\chi^2_{J-k-1} \geq \chi^2\right)$. Si el modelo es adecuado este test debería proporcionar un p-valor mayor que 0.05.

Alternativamente, podemos definir el **residual deviance** del modelo como

$$d_j = \pm\left\{2\left[o_j \ln\left(\frac{o_j}{m_j\hat{\pi}_j}\right) + (m_j - o_j)\ln\left(\frac{m_j - o_j}{m_j(1-\hat{\pi}_j)}\right)\right]\right\}^{1/2}$$

donde el signo de d_j es el mismo que el de $o_j - m_j\hat{\pi}_j$. Para los perfiles con $o_j = 0$ suponemos que

$$d_j = -\sqrt{2m_j\left|\ln\left(1-\hat{\pi}_j\right)\right|}$$

y para los perfiles con $o_j = m_j$ suponemos que

$$d_j = \sqrt{2m_j \left|\ln \hat{\pi}_j\right|}$$

A partir de estos residuales podemos construir el **estadístico deviance** D definido como

$$D = \sum_{j=1}^{J} d_j^2$$

cuya distribución de probabilidad es también una distribución χ^2 con $J - (k+1)$ grados de libertad si $J << n$ y el modelo ajustado es correcto. Por tanto podemos construir un test para la falta de ajuste del modelo, denominado **test deviance** del modelo, utilizando el p-valor $p\left(\chi^2_{J-k-1} \geq D\right)$. Si el modelo es adecuado este test debería proporcionar también un p-valor mayor que 0.05. Es fácil comprobar que, si $J = n$ y por tanto $m_j = 1$ y $\hat{\pi}_j = \hat{p}_i$, entonces el valor del estadístico deviance D coincide con la deviance del modelo $\Lambda_f = \Lambda\left(\hat{\boldsymbol{\beta}}\right)$ definida con anterioridad.

El problema de los dos tests anteriores es que requieren $J << n$ y en los casos prácticos puede haber situaciones en que J esté próximo a n o incluso $J = n$. En estas situaciones los tests anteriores no son correctos y es necesario buscar un test alternativo. Para ello Hosmer y Lemeshow (2000), suponiendo que $J = n$, propusieron construir g grupos (habitualmente $g = 10$) calculando $g - 1$ cuantiles (habitualmente los 9 deciles) de las probabilidades $\hat{p}_i$ predichas por el modelo e incluyendo en cada grupo a todos los puntos cuyas probabilidades predichas se encuentran entre dos cuantiles (habitualmente deciles) consecutivos. Para cada una de las clases denotamos por n_j al número de puntos que componen la clase j (habitualmente la parte entera de $0.1n$ para las 9 primeras clases y el resto para la última clase), por o_j al número de éxitos observados entre los puntos que componen la clase j y por $\hat{\pi}_j$ a la media de las probabilidades predichas $\hat{p}_i$ para los puntos de la clase j. Se calcula entonces el estadístico de Hosmer-Lemeshow como

$$\chi^2_{HL} = \sum_{j=1}^{g} \frac{\left(o_j - n_j \hat{\pi}_j\right)^2}{n_j \hat{\pi}_j (1 - \hat{\pi}_j)}$$

cuya distribución de probabilidad, suponiendo que $J = n$ y que el modelo ajustado es correcto, es una χ^2 con $g - 2$ grados de libertad (habitualmente 8). Por tanto podemos construir el **test de Hosmer-Lemeshow** para la falta de ajuste del modelo calculando el p-valor $p\left(\chi^2_{g-2} \geq \chi^2_{HL}\right)$. Si el modelo es adecuado este test debería proporcionar un p-valor mayor que 0.05. Aunque no fue analizado por los autores, el test se considera igualmente válido cuando $J \approx n$.