Análisis de regresión logística binaria

La regresión logística binaria se caracteriza por disponer de una variable dependiente cualitativa con dos valores (categorías o grupos) que configuran la presencia y la ausencia de una determinada característica. Por ejemplo, los ciudadanos que se abstienen en las elecciones y los que no, los que votan a un partido y los que no, los consumidores que compran un producto y los que no, las personas que están en paro y las que no, las personas que reinciden en un delito y las que no, las personas que tienen un riesgo contraer una enfermedad y las que no, las que devolverán un préstamo y las que no, etc.

La característica definida por la variable dependiente se pretende explicar en función de una serie de variables independientes o predictoras que nos determinan en qué se diferencian los dos grupos. Si consideramos tan sólo una variable independiente podemos hablar de regresión logística simple, si consideramos dos o más variables independientes el modelo de regresión logística es múltiple. En el contexto de la regresión logística estas variables se denominan también covariables. Como resultado del análisis se obtienen unos pesos o coeficientes que nos miden la importancia de cada variable independiente para diferenciar los grupos, y en segundo término obtenemos criterios para pronosticar la clasificación de los individuos o casos.

La relación logística

En el modelo de regresión lineal la relación entre las variables se expresa de forma general como:

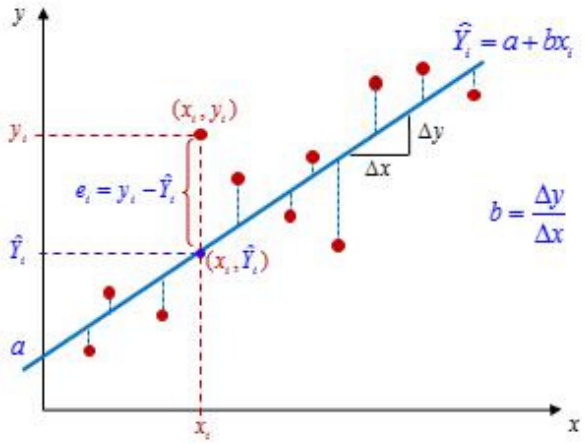
$$y_i = a + b_1 x_{1i} + b_1 x_{1i} + \cdots + b_p x_{pi} + e_i \qquad (1)$$

En que i es el dato i-ésimo y e_i es un error aleatorio. Es evidente que, si uno tuviera este conocimiento (relación entre las variables), podría plantear un modelo de estimación de la salida yi del tipo:

$$\hat{y}_i = a + b_1 x_{1i} + b_1 x_{1i} + \cdots + b_p x_{pi} \qquad (2)$$

En el caso particular del modelo de regresión simple, la ecuación $\hat{y}_i = a + b_1 x_{1i}$, se representa gráficamente mediante una recta en el plano que se ajusta por el método de mínimos cuadrados (Gráfico 1).

Gráfico 1: Modelo lineal de regresión simple



La recta de regresión se extiende de forma ilimitada entre -∞ y +∞, si bien los valores de la recta de regresión se interpretan en el rango de valores de x observados en la muestra y tienen un sentido interpretativo, descartando valores de predicción imposibles a partir de los datos estudiados. No obstante, también puede suceder que a pesar de considerar el rango de valores de la muestra los valores pronosticados sean valores imposibles. Es el caso que se puede dar cuando consideramos en la regresión lineal variables dicotómicas de la variable dependiente, codificadas con 0 y 1, donde los valores predichos pueden ser inferiores a 0 y superiores a 1, fuera del rango definido por la variable dependiente.

La regresión logística resuelve este tipo de problema usando una función no lineal como es la función logística. Con esta función se pueden efectuar predicciones comprendidas entre un mínimo y un máximo. El modelo de regresión logística es un modelo no lineal que utiliza el método de máxima verosimilitud, un procedimiento iterativo que en fases sucesivas ajusta el modelo.

La formulación matemática de la curva logística en el caso de la regresión logística binaria simple es:

$$y = \Pr\left(y = \frac{1}{x}\right) = \frac{e^{a+bx}}{1+e^{a+bx}} \qquad (2)$$

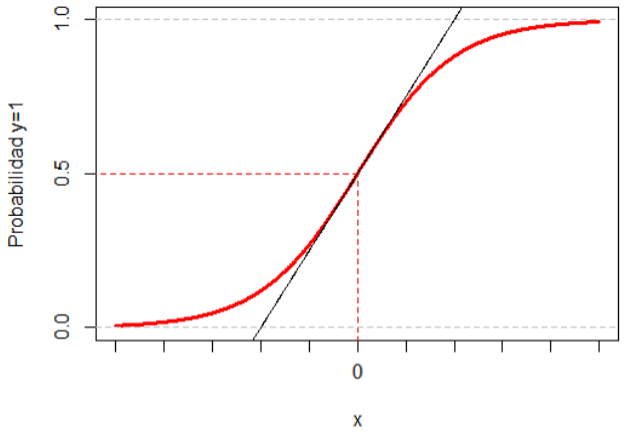
o bien, de forma equivalente:

$$y = \Pr\left(y = \frac{1}{x}\right) = \frac{1}{1+e^{-(a+bx)}} \qquad (3)$$

Es decir, la probabilidad de que la variable dependiente y tome el valor 1 (presencia de la característica estudiada) en función de la variable independiente x.

La representación gráfica de la función logística, de expresión general $y = f(x) = \frac{1}{1+e^{-x}}$, es una curva con forma sigmoidea (Gráfico 2):

Gráfico 2. La curva logística



que verifica las propiedades siguientes:

- Sus valores oscilan entre 0 y 1, $0 < f(x) < 1$, lo que permite interpretarla en términos de probabilidad.

- Su límite inferior es el valor 0 : $\lim_{x\to-\infty} \frac{1}{1+e^{-x}} = 0$
- Su límite superior es el valor 1 : $\lim_{x\to\infty} \frac{1}{1+e^{-x}} = 1$
- Cuando la x vale 0 la función vale ½ : $f(0) = \frac{1}{1+e^{-0}} = \frac{1}{2}$

La ecuación de la función logística permite asignar valores a la variable independiente para generar valores de la dependiente de la misma forma que en la regresión lineal, y su interpretación es similar. Pero en este caso los valores de predicción de la variable independiente y se situarán siempre en el intervalo (0,1), lo que facilita interpretar los resultados y los parámetros de la ecuación en términos de probabilidad para pronosticar un comportamiento.

1.2. El modelo de la regresión logística

El modelo de regresión logística binaria considera dos sucesos de un fenómeno o variable Y, excluyentes y exhaustivos, que se codifican con valores 0 y 1. Si la probabilidad de que suceda uno de ellos es P, la probabilidad de que suceda la otro es igual a 1 menos la probabilidad P:

$$Pr(y=1) = P$$
$$Pr(y=0) = 1-P$$

La cuestión es considerar la información de una (o más variables en la versión múltiple) para definir un modelo que permita pronosticar la probabilidad de la variable dependiente *y*, es decir, se trata de encontrar una o más variables que discriminen bien entre los dos posibles valores de la variable *y*.

En un modelo de regresión logística binaria simple, la ecuación logística se expresa como:

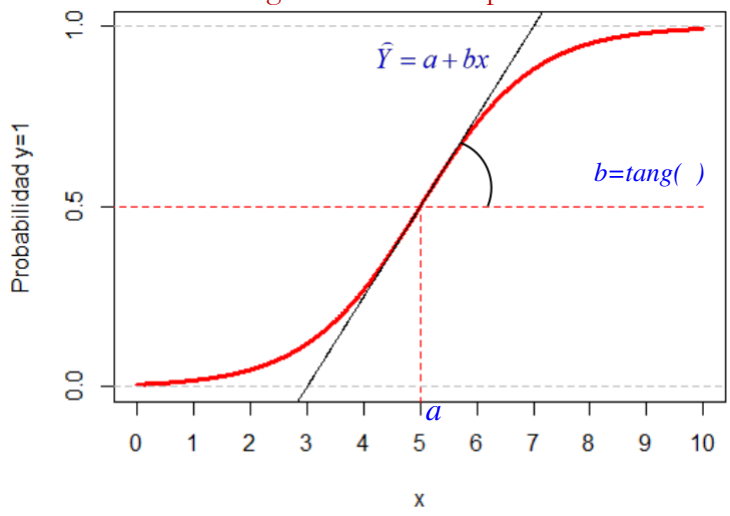
$$\Pr(y=1) = \frac{1}{1+e^{-(a+bx)}} = P$$

y, por tanto,

$$\Pr(y=0) = 1-\left(\frac{1}{1+e^{-(a+bx)}}\right) = 1-P$$

Con la representación gráfica adjunta (Gráfico 3).

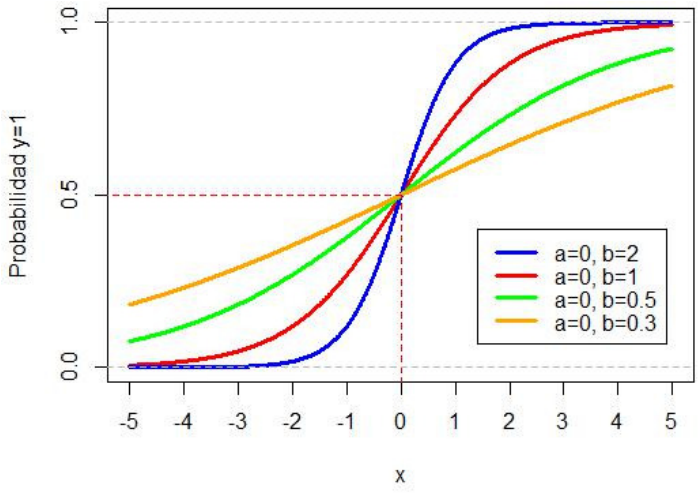
Gráfico 3. Representación del modelo de regresión logística binaria simple



El coeficiente *a* representa la posición de la curva sobre el eje horizontal o de abscisas, y sitúa la curva más hacia la derecha o hacia la izquierda. El coeficiente *b* representa la pendiente de la curva en su punto de inflexión, en función de su valor más alto o más bajo tendremos una pendiente de la curva más inclinada o menos.

Por tanto, nos podemos encontrar con una familia de curvas que varían en función de los valores de *a* y de *b* (Gráfico 4).

Gráfico 4. Familia de curvas logísticas según valores distintos de la pendiente b



La variación de la pendiente implicará una distinta capacidad discriminatoria de los valores de *y*. Una buena variable independiente predictora es la que genera una curva con una elevada pendiente, cuando el valor absoluto de *b* es alto; si *b* se acerca al valor 0 su capacidad predictora se reduce. Por tanto, el objetivo del análisis de regresión logística consiste en encontrar las variables con el mayor coeficiente asociado.

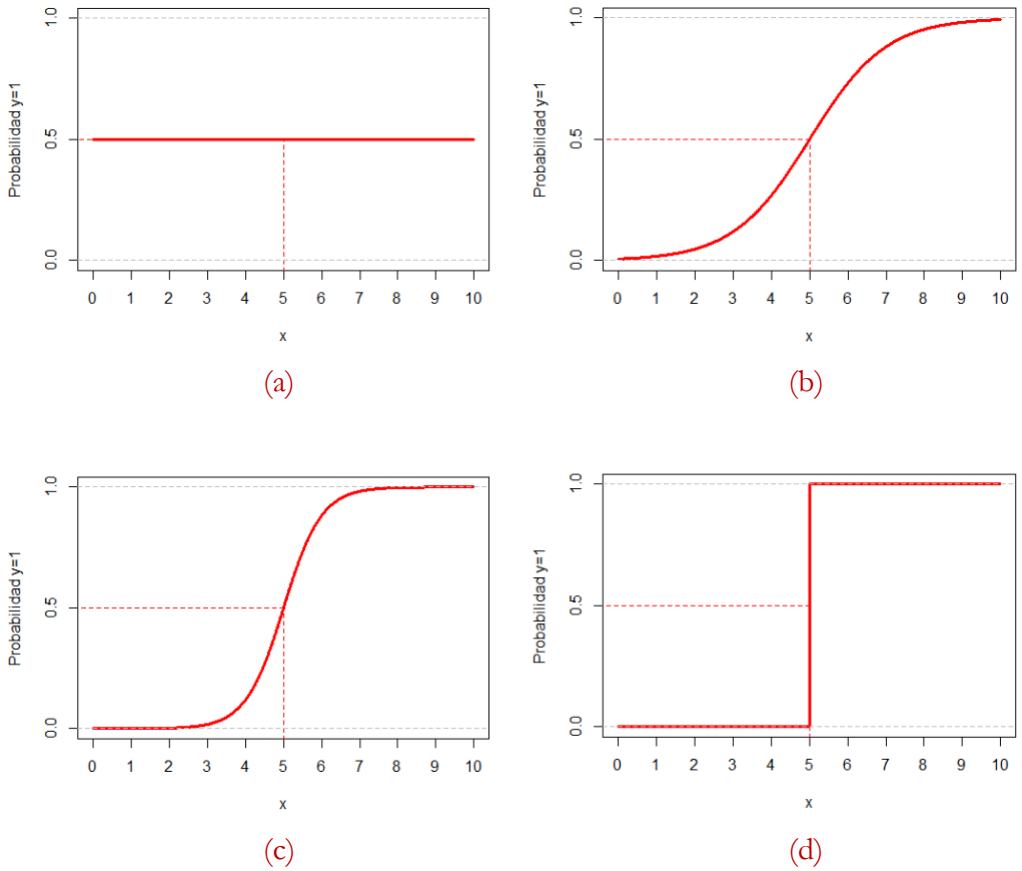
En las cuatro representaciones presentadas en el Gráfico 5 podemos observar cuatro casos de curvas logísticas ordenadas desde la mínima capacidad discriminatoria (figura a) hasta la máxima capacidad de discriminar los valores de la variable dependiente (figura d).

La interpretación de los coeficientes de la regresión logística difiere del caso de la regresión lineal. Aquí el coeficiente no es la medida de cuánto variará *y* ante una variación en una unidad de *x*, sino el cambio producido por una variación de una unidad de *x* en el logaritmo neperiano (*ln*) del cociente de probabilidades de los dos sucesos, la denominada transformación logit.

La transformación logit surge de considerar la relación o el cociente de probabilidad entre dos sucesos, llamada ventaja o razón (como traducción de la expresión inglesa *odds*). La razón de un suceso es el cociente entre la probabilidad de que éste suceda y la probabilidad de que no suceda:

$$Odds = \frac{P}{1-P} = \frac{\text{Probabilidad de que ocurra un suceso}}{\text{Probabilidad de que no ocurra un suceso}} \qquad (4)$$

Gráfico 5. Representación de curvas logísticas con diferente capacidad explicativa



Así, por ejemplo, si el 75% de la población vota en unas elecciones (la probabilidad es del 0,75), el 25% se abstiene y el *odds* será 3: $\frac{P}{1-P} = \frac{0,75}{0,25} = 3$. De la misma forma que pasamos de las probabilidades a las razones, podemos pasar de las razones a las probabilidades:

$$P = \frac{odds}{odds + 1}$$

Si el odds es 3 la probabilidad es: $\frac{3}{3+1} = 0,75$ .En ambos casos se cuantifica qué tan probable es un suceso, su "riesgo". El riesgo relativo es el cociente de probabilidades de un suceso en dos condiciones distintas. El *odds ratio* (o razón de razones de probabilidad) es el cociente de dos *odds*. Si en el municipio A vota el 80% y en el Municipio B el 50%, el odds ratio será 4:

A partir de las expresiones anteriores de Pr(y=1) y Pr(y=0), obtenemos:

$$\frac{\Pr(y=1)}{\Pr(y=0)} = \frac{\dfrac{1}{1+e^{-(a+bx)}}}{1-\left(\dfrac{1}{1+e^{-(a+bx)}}\right)} = \frac{P}{1-P}$$

La expresión se puede simplificar para obtener:

$$\frac{\Pr(y=1)}{\Pr(y=0)} = \frac{P}{1-P} = e^{a+bx}$$