Estadística Bayesiana

Clase 20: Regresión Logística

Isabel Cristina Ramírez Guevara

Escuela de Estadística
Universidad Nacional de Colombia, Sede Medellín

Medellín, 10 de noviembre de 2020

Regresión Logística

En muchos casos las respuestas tienen solo dos categorías del tipo si/no de modo que se puede definir una variable Y_i que tome dos posibles valores 1 (éxito) y 0 (fracaso). Esta variable sigue una distribución Bernoulli tal que:

$$p(Y_i|\theta_i) = \theta_i^{y_i}(1-\theta_i)^{1-y_i} \quad y_i = 0, 1.$$

Sea n_i el número de observaciones en el grupo i, y y_i el número de éxitos en el grupo i, por lo tanto tenemos:

$$p(Y_i|\theta_i) = \binom{n_i}{y_i}\theta_i^{y_i}(1-\theta_i)^{n_i-y_i} \quad y_i = 0, \cdots, n_i.$$

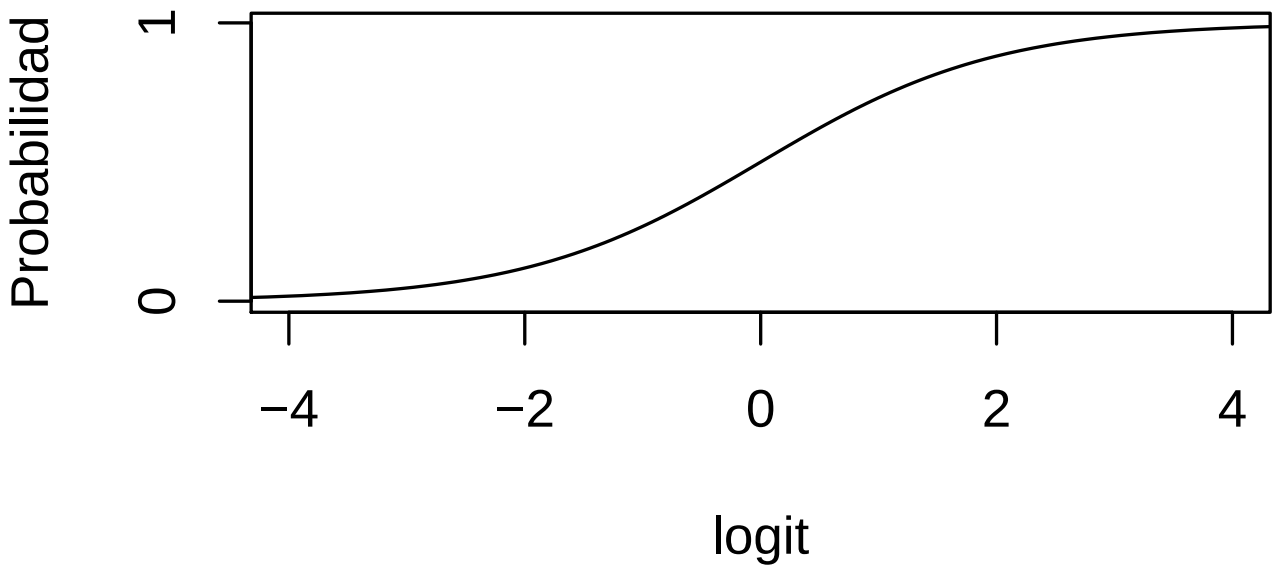
Regresión Logística

Cuando trabajamos con probabilidades, el primer problema al que nos enfrentamos es que la probabilidad sólo toma valores entre 0 y 1, y si pretendemos relacionarla directamente con los predictores puede que nos salgamos del intervalo [0,1]. Una solución sencilla es transformar la probabilidad para evitar este tipo de restricciones y relacionar esta probabilidad modificada con los predictores, esta transformación es el logit or log-odds

$$\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} = \eta_i = ln\left(\frac{\theta_i}{1-\theta_i}\right).$$

Regresión Logística

Esta transformacióon lo que consigue es que cuando la probabilidad se aproxima a 0 el logit lo hace a $-\infty$ y cuando la probabilidad se acerca a 1, el logit lo hace a $+\infty$. Si la probabilidad es 0.5, el odds es 1 y el logit 0. Logits negativos corresponden a probabilidades inferiores a $1/2$ y viceversa.



Regresión Logística

La transformación logit es uno a uno, a la inversa se le suele llamar antilogit y permite calcular la probabilidad a partir de

$$\theta_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$$

Así el modelo de regresión logística es:

$$logit(\theta_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}$$

por lo tanto:

$$\theta_i = \frac{e^{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}}}{1 + e^{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}}}$$

Interpretación de los parámetros

La interpretación de los parámetros en regresión logística está basada en la noción de los odds y la razón de odds. Los odds están definidos como la probabilidad relativa de éxito ($Y = 1$) comparada con la probabilidad de fracaso ($Y = 0$), por lo tanto

$$odds = \frac{\theta}{1 - \theta}$$

el modelo logístico se puede reescribir como

$$ln(odds_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}$$

Interpretación de los parámetros

La interpretación de las probabilidades es relativamente simple y directa. Por ejemplo, odds$=2$ implica que la probabilidad de éxito es dos veces mayor que la de fracaso, mientras que un odds $=0.6$ implica que la probabilidad de éxito es igual al 60 % de la probabilidad de fracaso. El valor de 1 es de interés central ya que implica que las probabilidades de ambos resultados son iguales a 0.5. Los valores de odds superiores a uno (>1) indican una mayor probabilidad de éxito en contraste con la probabilidad de fracaso $(\theta > 0.5)$, mientras que valores inferiores a uno (<1) indican una probabilidad de éxito inferior a la probabilidad de fracaso $(\theta <0.5)$.

Interpretación de los parámetros

La cantidad (odds - 1) x 100 da información sobre el porcentaje de incremento o disminución (dependiendo del signo) de la probabilidad de éxito en comparación con la probabilidad de fracaso. Por ejemplo, un valor de odds=1.6 indica que la probabilidad de éxito es 60 % mayor que la probabilidad de fracaso. De la misma manera, un odds=0.6 indica que la probabilidad de éxito es 40 % menor que la probablidad de fracaso.

Interpretación de odds

$$odds = \frac{\theta}{1-\theta} = a$$

- Si a=1 $\rightarrow \theta = 1 - \theta$ =0.5.
- Si $a < 1 \rightarrow \theta < 0.5 < 1 - \theta$.
- Si $a > 1 \rightarrow \theta > 0.5 > 1 - \theta$.
- La probabilidad de éxito (Y=1) es a veces mayor que la probabilidad de fracaso (Y=0).
- Si $a > 1$, entonces la probabilidad de éxito (Y=1) es (a-1)x100 % veces mayor que la probabilidad de fracaso (Y=0).
- Si $a < 1$, entonces la probabilidad de éxito (Y=1) es (1-a)x100 % veces menor que la probabilidad de fracaso (Y=0).

Interpretación de *OR*

La razón de dos odds es llamado razón de odds (OR) y proporcionar el cambio relativo de las probabilidades en dos condiciones diferentes, denotado por $X = 1, 2$ y subíndices 1 y 2:

$$OR_{12} = \frac{odds(X = 1)}{odds(X = 2)}.$$

Cuando OR_{12}= 1, los *odds* condicionales bajo comparación son iguales, lo que indica que no hay diferencia en probabilidad relativa de Y bajo $X = 1$ y $X = 2$. Usando un enfoque similar, (OR_{12}-1)x100 proporciona el cambio porcentual de los *odds* para X= 1 en comparación con los odds correspondientes cuando X= 2.

Interpretación de *OR*

$$OR_{12} = \frac{odds(X = 1)}{odds(X = 2)} = b$$

- Si b=1 $\rightarrow$ $odds(X = 1) = odds(X = 2)$.
- Si $b < 1 \rightarrow odds(X = 1) < odds(X = 2)$.
- Si $b > 1 \rightarrow odds(X = 1) > odds(X = 2)$.

Interpretación de los parámetros

Suponga que tenemos una única covariable, por lo tanto,

$$\begin{aligned} ln\left(\frac{\theta}{1-\theta}\right) &= \beta_0 + \beta_1 x \\ ln(odds) &= \beta_0 + \beta_1 x \\ odds &= e^{\beta_0} e^{\beta_1 x} \\ &= B_0 B_1^x \end{aligned}$$

donde $B_j = e^{\beta_j}$ para $j = 0, 1$. Ahora calculamos $OR_{x+1,x}$,

$$OR_{x+1,x} = \frac{odds(x+1)}{odds(x)} = \frac{B_0 B_1^{x+1}}{B_0 B_1^x} = B_1$$

Por lo tanto $B_1 = e^{\beta_1}$ denota la magnitud relativa de los *odds* cuando X incrementa una unidad. Notemos que para $X = -\beta_0/\beta_1$, calculamos el valor de X para el cual ambas probabilidades es igual a 0.5.

Ejemplo

A 54 personas mayores se les realizó la prueba que mide la escala de inteligencia de adultos de Wechsler (WAIS) que da como resultado una puntuación discreta con un rango de 0 a 20. El objetivo de este estudio fue identificar personas con síntomas de senilidad (variable binaria) mediante la puntuación WAIS. Por lo tanto, se está interesado en calcular el valor umbral de X para el cual $\theta > 0.5$ lo que permite identificar el estado del paciente directamente utilizando el valor de la covariable X, el puntaje WAIS.

En este ejemplo la variable de respuesta, síntoma de senilidad, es binaria y se tiene una única covariable x el puntaje de WAIS. Se obtienen los siguientes resultados:

Ejemplo

	2.5 %	50 %	97.5 %
β_0	0.3173625	2.6030000	5.4010250
β_1	-0.6228	-0.3451	-0.1317
e^{β_0}	1.373925	13.510000	221.702500
$OR = e^{\beta_1}$	0.5365	0.7081	0.8766
WAIS ($\theta = 0.5$)	2.280000	7.491500	9.521025

Se puede concluir que hay una asociación negativa entre el puntaje WAIS y la presencia de síntomas de senilidad. Los odds de síntomas de senilidad para un individuo cuyo puntaje WAIS es cero tiene un valor esperado posterior igual a 13.51. Adicionalmente, por cada punto que aumenta el puntaje WAIS, la probabilidad de tener la enfermedad disminuye en 29 % en promedio. Con respecto al valor umbral de enfermedad, el modelo muestra que si el puntaje de WAIS ≤ 7 se considera que el paciente tiene la enfermedad, si el puntaje de WAIS ≥ 8 se considera que la persona esta sana.