Modelación estadística: La regresión logística (Parte 1)

Gabriel Cavada Ch.[1,2]
[1]*División de Bioestadística, Escuela de Salud Pública, Universidad de Chile.*
[2]*Facultad de Medicina, Universidad de los Andes.*

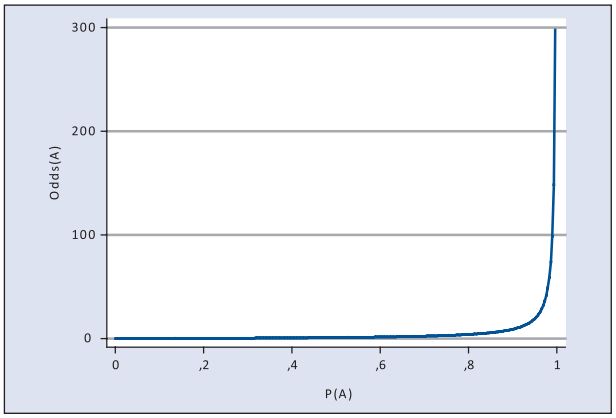
Statistical modeling: Logistic regression (Part I)

La distribución de probabilidades logística

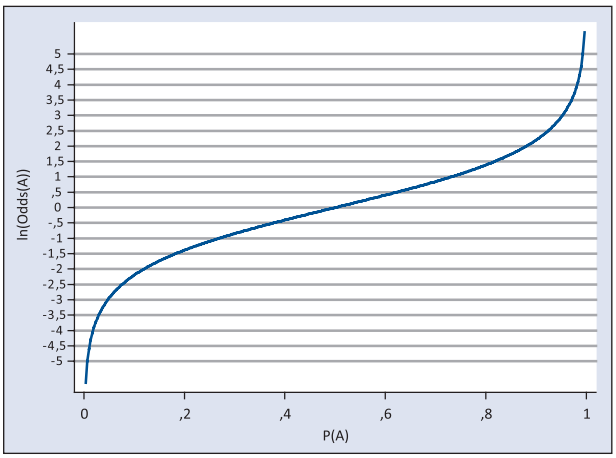
Supongamos que estamos interesados en la ocurrencia de un evento "A", cuya probabilidad de aparición es "P", es decir: $P(A) = P$ y por consiguiente la probabilidad de que "A" no ocurra es $P(A') = 1 - P$; sin embargo, sabemos que la ocurrencia de A, y por ende su probabilidad, está relacionada con el valor que tome una variable aleatoria X, esto es $P(A) = P(X \leq x)$: por ejemplo, si A: una persona muere y X es la edad de la persona, es razonable pensar que $P(morir) = P(Edad \leq edad)$. Notar que $P(A) = F(X)$, donde $F(X)$ es la función de distribución de probabilidades de X. El problema fundamental es como relacionar la probabilidad de la aparición del evento "A", con los posibles valores de la variable X.

Luego ¿Cómo hacer para que la P(A) dependa linealmente de X?; la respuesta directa a este problema sería proponer: $P(A) = \alpha + \beta \bullet X$, sin embargo, esta propuesta no es satisfactoria ya que $P(A) \in [0,1]$ y la función lineal puede tomar cualquier valor real. Si deseamos perseverar en la asociación lineal de la $P(A)$ con X, debemos pensar en una transformación de $P(A)$ que garantice que tome valores en todos los reales. Las propuestas que resuelven el problema son muchas, sin embargo, la más útil es la siguiente:

- Si consideramos el Odds del evento A, es decir $Odds(A) = \dfrac{P(A)}{1 - P(A)}$ y lo evaluamos para todos los posibles valores de P(A), obtenemos la siguiente función:



Observamos, que como es sabido que el Odds puede tomar cualquier valor real positivo, ello nos ilumina a considerar el logaritmo del Odds, ya que la función logaritmo tiene dominio en los reales positivos pero su recorrido son todos los reales, como se observa en el siguiente gráfico:



- Así entonces proponemos la relación:

$$\ln\left(\frac{P}{1-P}\right) = \alpha + \beta \bullet X$$

Que nos lleva a:

$$P(A) = F(X) = \frac{e^{\alpha+\beta\bullet X}}{1 + e^{\alpha+\beta\bullet X}}$$

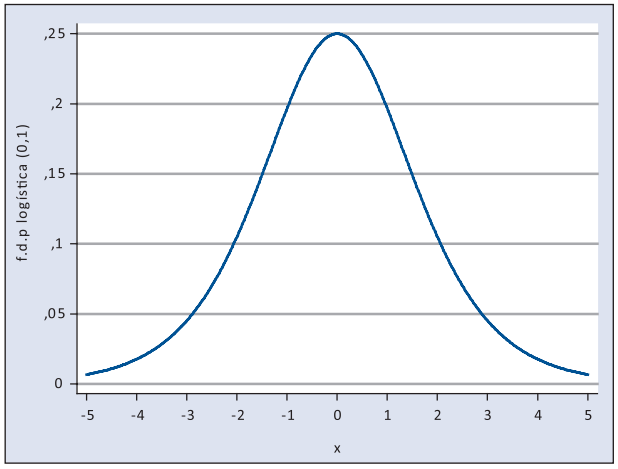
De donde deducimos que la función densidad de probabilidades es:

$$f(X) = \frac{\beta e^{\alpha+\beta\bullet X}}{\left(1 + e^{\alpha+\beta\bullet X}\right)^2}$$

Particularmente si consideramos $\alpha = 0$ y $\beta = 1$, la función densidad de probabilidades es:

$$f(X) = \frac{e^{X}}{\left(1 + e^{X}\right)^2}$$

Cuyo gráfico es el siguiente:



La esperanza y la varianza de la distribución logística estándar son respectivamente:

$$E[X] = 0$$

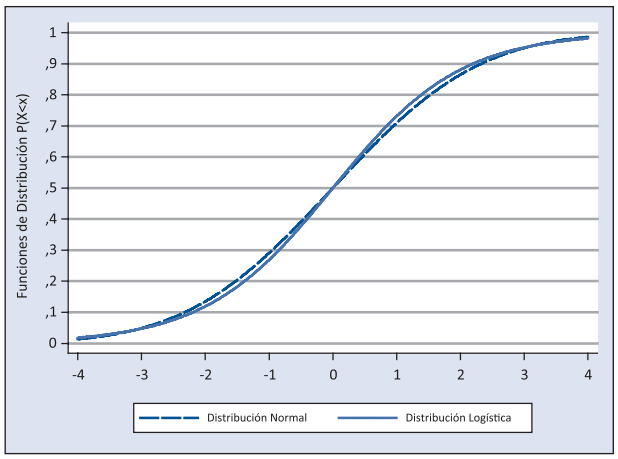
$$Var[X] = \frac{\pi^2}{3}$$

En consecuencia para la distribución logística de parámetros α y β se tiene:

$$E[X] = \alpha$$

$$Var[X] = \frac{(\beta\pi)^2}{3}$$

Usando estos resultados se encuentra un hecho sorprendente: la función de distribución de la logística estándar, difiere muy poco con la función de distribución de la $N(0,\pi \uparrow 2/3)$, como lo muestra el siguiente gráfico:



Para la distribución logística estándar se verifica:

- $1 - F(X) = \frac{1}{1 + e^X}$
- $f(X) = F(X)\,[1 - F(X)]$

La regresión logística

Nos interesa modelar la aparición de un evento, A, explicándolo por un perfil definido como una combinación lineal de variables:

$$X\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_p X_p$$

La respuesta la codificamos de la siguiente forma:

$$Y = \begin{cases} 1, \textit{ si el evento A aparece} \\ 0, \textit{ si el evento A no aparece} \end{cases}$$

Definiendo $P(Y = 1 \mid X\beta) = P(A) = \pi(X)$, es claro que la distribución de probabilidades de Y es Bernoullí con probabilidad de éxito $\pi(X)$, es decir, la función de cuantía de probabilidades es:

$$P(Y = y) = (1 - \pi(X))^{1-y}\,\pi(X)^y, \textit{ con } y = 0,1$$

Al asumir que $\pi(X) = F(X)$ donde F(X) es la función de distribución logística evaluada en el perfil Xβ, la cuantía de probabilidades de Bernoullí se puede escribir como:

$$P(Y = y \mid X) = (1/(1 + e^{\uparrow} X\beta))^{\uparrow}(1 - y)(e^{\uparrow} X\beta/(1 + e^{\uparrow} X\beta))^{\uparrow} y, \textit{ con } y = 0,1$$

Por lo tanto, si se tiene una muestra aleatoria de "n" perfiles asociados a sus respectivas respuestas "y", la función de verosimilitud que estima los parámetros β del modelo es:

$$L = \prod_{i=1}^{n} \left(\frac{1}{1 + e^{X_i\beta}}\right)^{1-y_i} \left(\frac{e^{X_i\beta}}{1 + e^{X_i\beta}}\right), \textit{ con } y_i = 0,1$$

Esta función de verosimilitud corresponde al modelo logístico de respuesta binaria. Los parámetros hay que estimarlos mediante el método iterativo de Newton-Raphson, como se revisó en el capítulo I.

Como se estableció anteriormente:

$$ln\left((P(Y = 1 \mid X\beta))/(1 - P(Y = 1 \mid X\beta))\right) = ln\left(Odds\,(Y = 1 \mid X\beta)\right) = X\beta$$

Rincón de la Bioestadística

Esta relación permite comparar dos perfiles: **X** y **X'** pues al evaluar la expresión anterior en cada uno de estos perfiles y luego restar estas ecuaciones se obtiene:

$$ln\left(Odds\,(Y=1|X\beta)\right)=X\beta$$

$$ln\left(Odds\,(Y=1|X'\beta)\right)=X'\beta$$

$$\downarrow$$

$$ln\big(Odds\,(Y=1|X\beta)\big)-ln\big(Odds\,(Y=1|X'\beta)\big)=X\beta-X'\beta=(X-X')\beta$$

O equivalentemente:

$$ln=\left(\frac{Odds\,(Y=1|X\beta)}{Odds\,(Y=1|X'\beta)}\right)=\ln(OR)=X\beta-X'\beta=(X-X')\beta$$

Por lo tanto, β, es el cambio del *ln(OR)* por cambio de perfil, de donde se deduce que:

$$OR=e^{(X-X')\beta}$$

Si X es una variable dicotómica, por ejemplo X = 1 y X = 0 denoten exposición y no exposición respectivamente, la expresión del OR es:

$$OR=e^{(X-X')\beta}=e^{(1-0)\beta}=e^{\beta}$$

Cuya interpretación ya es conocida.

La novedad es que si X es una variable continua y comparamos el perfil X con el perfil X+1, la expresión que define el OR entre perfiles es:

$$OR=e^{(X-X')\beta}=e^{(X+1-X')\beta}=e^{\beta}$$

Que representa el cambio de riesgo cuando la variable X se incrementa en "una unidad".

Los programas estadísticos dan la opción de reportar los resultados en términos de coeficientes o si se desea en Odds Ratios.

Ejemplo 1: Estimar la fuerza de la asociación en la siguiente tabla:

	Cáncer de vesícula	Control	
Consumo ají rojo	30	45	75
No consumo de ají rojo	10	55	65
	40	100	140

Ca	Odds Ratio	Error estándar	p-value	Intervalo de confianza 95%	
Ají	3,67	1,53	0,002	1,62	8,30

Es decir, el riesgo de estar expuesto al consumo de ají es 367% mayor en los sujetos con Cáncer de vesícula, si el consumo del ají en los controles se produjera por azar.

Ejemplo 2: Estimar la fuerza de la asociación de la glicemia con la mortalidad intrahospitalaria por IAM ajustada por género.

Mortalidad intrahospitalaria	Odds Ratio	Error estándar	p-value	Intervalo de confianza (95%)	
Glicemia	1,01	0,00	0,0000	1,00	1,01
Sexo femenino	2,59	0,94	0,0080	1,28	5,27

La interpretación de estos resultados es: por cada punto de aumento en la glicemia de ingreso el riesgo de muerte crece en 1% si en el nivel anterior la muerte se produjera por azar, ajustando por género. O el riesgo de morir por ser mujer es 259% mayor que si en los hombres la muerte se produjera por azar, ajustando por glicemia.