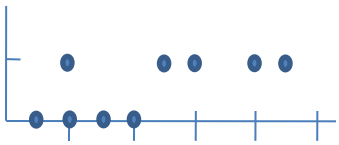
Regresión Logística

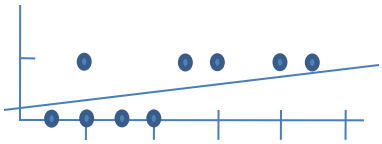
Dr. Jesús Alberto Mellado Bosque

La regresión logística se utiliza cuando se tiene una variable independiente continua y una variable dependiente con dos valores (dicotómica). Por ejemplo, la variable independiente es la cantidad en gramos de sal que consume una persona al día en promedio y la variable dependiente es si la persona tiene hipertensión o no (valores 0 y 1). Los datos pueden ser como la siguiente tabla:



Grs de sal	0.5	0.5	1	1	1.5	2	2.5	3	4	4.5
Hipertensión	0	0	0	1	0	0	1	1	1	1

Un primer intento puede ser calcular la línea de regresión lineal, pero el resultado de la variable dependiente no es de dos valores, así que no es aplicable.



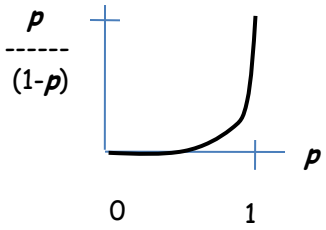
La solución propuesta es cambiar la variable dependiente, en lugar de que sea la hipertensión, será la proporción (odds) de la probabilidad de que tenga hipertensión. Si ***p*** es la probabilidad de que tenga hipertensión, entonces la proporción será:

$$\frac{p}{(1-p)}$$

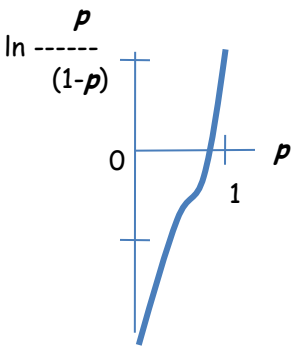
Entonces la línea de regresión se puede definir como:

$$\frac{p}{(1-p)} = ax + b$$

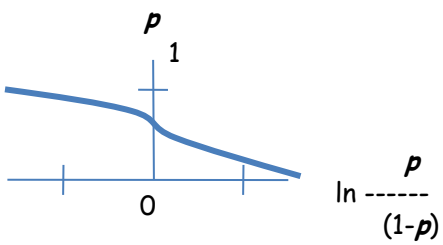
Desafortunadamente, la proporción $p/(1-p)$ no se distribuye lineal, sino exponencial, como se muestra en la siguiente gráfica, donde el eje horizontal es la probabilidad ***p*** y el eje vertical es $p/(1-p)$.



Para convertir la línea exponencial en una línea lo mas recta posible, se le aplica un logaritmo natural (la función inversa a una exponencial), de tal manera que la gráfica queda como se muestra:



Pero al despejar ***p***, la gráfica tiene como variable independiente una línea y como variable dependiente un resultado entre 0 y 1.



Entonces si se asocia la línea de regresión con la nueva variable $\ln(p/(1-p))$, se construye la siguiente expresión.

$$\ln \frac{p}{(1-p)} = ax+b$$

Al hacer las operaciones algebraicas necesarias se tiene que:

$$p = \frac{1}{1 + e^{(-a-bx)}}$$

Con esta ecuación se puede saber la probabilidad de que ocurra un evento (e.g. hipertensión) de acuerdo a una variable independiente.
Además de conocer la probabilidad de que ocurra el evento para cierto valor de la variable independiente, es importante conocer cómo se comporta esa probabilidad, es decir, la pendiente de la línea de probabilidad.

El "odds" (proporción p/(1-p)) indica que tantas veces es mayor la probabilidad de que ocurra un evento respecto a que no ocurra, por ejemplo, si la proporción es 3, significa que es tres veces mas probable a que no ocurra (p=0.75; 1-p=0.25).

Si se toma un x_1 muy bajo, y se toma un x_2 muy alto en la variable independiente, y se calcula su "odds" para cada uno de los valores, entonces se pueden comparar los "odds" de la siguiente manera:

$$\frac{Odds_1}{Odds_2} = \frac{p_1(1-p_1)}{p_2(1-p_2)} = \text{Odds ratio (índice de riesgo)}$$

Como existe la dificultad de generalizar para un x_1 y un x_2, entonces se ha seleccionado un estimador. El "odds ratio" o el índice de riesgo, se puede estimar como:

$$\text{Odds ratio (índice de riesgo)} = e^b$$

El valor ***b*** es la pendiente del resultado de la regresión lineal simple.

Ahora falta validar el modelo, para lo cual se aplica una prueba de bondad de ajuste utilizando la distribución chi-cuadrada. Algunos autores proponen el método general, algunos otros el de Wald o el de Hosmer-Lemeshow.

Se ha propuesto reportar los resultados de una prueba re regresión logística mediante la siguiente tabla.

Variable	Indice de riesgo OR (odds ratio)	Err Std	Chi cuad	Nivel significancia
Hipertensión	2.3	2.4	4.2	p<0.05

En la tercera columna se muestra el error estándar de la diferencia del modelo propuesto a los datos originales. En la cuarta columna el resultado de la prueba de bondad de ajuste y finalmente el nivel de significancia de la bondad de ajuste.