REGRESION LOGISTICA MULTIPLE

Jorge Galbiati R.

La regresión logística permitediscriminar entre dos poblaciones, en términos de un conjunto de variables numéricas, en el papel de predictores.

También para representar probabilidades de ocurrencia de un evento, como función de una serie de variables predictoras.

Por último, sirve para representar una variabole asociada a un fenómeno, que dependa de un conjunto de variables predictoras, cuyo comportamiento sea aproximadamente lineal, dentro de un cierto rango de los predictores, y tiendan a mantenerse constantes fuera de él.

De este modo el modelo de regresión logística resulta más realista que un modelo de regresión lineal. Es irreal que la relación entre una variable respuesta y un grupo de predictores sea de tipo lineal, en todo el rango de estas últimas.

Los predictores pueden ser también variables dicotómicas, en tal caso se utilizan variables dummy para representarlas.

El modelo de regresión logística es un modelo lineal generalizado $\gamma = \eta(x) + e$

en que la función de enlace η es del tipo:

$$\eta(x) = \frac{\exp\{\beta_0+\beta_1 x_1+........+\beta_k x_k\}}{1+\exp\{\beta_0+\beta_1 x_1+........+\beta_n x_k\}}$$

Las variables $x_1, x_2, ..., x_k$ son los predictores , la variable γ es la respuesta, que toma los valores 1 o 0, los $\beta_0, \beta_1, ..., \beta_k$ son los parámetros del modelo.

e es un error aleatorio, tal que $0 < \eta(x) + e < 1$

Si el modelo se utiliza para clasificar, los valores 0 y 1 de la respuesta y identifican a ambas poblaciones en que se puede clasificar una observación.

Normalmente 1 corresponde a la población con presencia de algún atributo, o con la ausencia de él. (por ej: alguna enfermedad, un estado de desarrollo, etc.).

<u>Caso particular:</u> Un predictor.

El modelo se reduce a: $\gamma = \frac{\exp\{\beta_0+\beta_1 x_1\}}{1+\exp\{\beta_0+\beta_1 x_1\}} + e$ (1)

Si β_1 es positivo, entonces: $Lim_{x\to-\infty}\gamma = 0$ y $Lim_{x\to\infty}\gamma = 1$

Si el coeficiente β_1 es positivo, γ es creciente con x, en forma sigmoidal, en caso contrario es negativo. El valor $x_{\frac{1}{2}}$ para el cual γ alcanza 1/2 es $x_{\frac{1}{2}} = -\beta_0/\beta_1$. En ese punto se alcanza la máxima pendiente, que es igual a $\beta_1/4$.

La Figura 1 representa la función, en el caso de β_1 positivo.

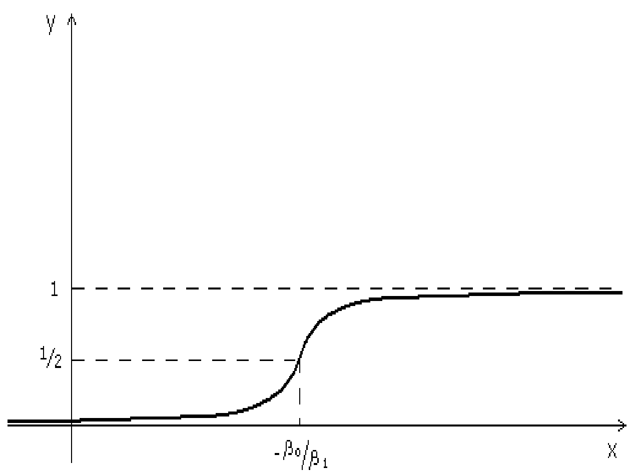

Figura 1

Figura 1: Regiones de discriminacón.

Una forma de estimar los parámetros es recurriendo al método de mínimos cuadrados. Si previamente se linealiza el modelo entonces se puede encontrar una solución analítica.

En efecto, despejando, se puede expresar (1) como

$$\exp\{\beta_0 + \beta_1 x_1 + ... + \beta_k x_k\} = \frac{\eta(x)}{1-\eta(x)}$$

De donde

$$log\left(\frac{\eta(x)}{1-\eta(x)}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

Si se hace $y = \eta(x)$, es decir, se prescinde del término de error e, entonces queda

$$log\left(\frac{y}{1-y}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

Se transforman los valores y a $z = log\left(\frac{y}{1-y}\right)$, y se estiman los parámetros como en una regresión lineal de z en términos de los predictores $x_1, x_2, ..., x_k$.por el método de mínimos cuadrados.

Obsérvese que los valores de y no pueden valer 0 ó 1. En tal caso se deben estimar directamente del modelo (1), como un modelo no lineal, por mínimos cuadrados, recurriendo a una minimización por métodos numéricos.

Tipos de modelos de regresión logística

Modelo logístico univariante simple.

$$y = \frac{\exp\{a+bx\}}{1+\exp\{a+bx\}}$$

Modelo logístico univariante multiple.

$$y = \frac{\exp\{\beta_{j,0}+\beta_1 x_1+\beta_{j,2}x_2+...+\beta_{j,k}x_k\}}{1+\exp\{\beta_{j,0}+\beta_{j,1}x_1+\beta_{j,2}x_2+...+\beta_{j,k}x_k\}}$$

Modelo logístico multivariante simple.

$$y_j = \frac{\exp\{a+bx\}}{1+\exp\{a+bx\}}$$

Modelo logístico multivariante multiple.

$$y_j = \frac{\exp\{\beta_{j,0}+\beta_1 x_1+\beta_{j,2}x_2+...+\beta_{j,k}x_k\}}{1+\exp\{\beta_{j,0}+\beta_{j,1}x_1+\beta_{j,2}x_2+...+\beta_{j,k}x_k\}}$$