Modelo de Regresión Logística

Modelo de regresión que explica el comportamiento de una variable dependiente discreta, Y, dicotómica en función de una o más variables independientes cualitativas o cuantitativas.

Los valores que toma la variable dependiente son:
Y=1 e Y=0

El objetivo es construir un modelo capaz de describir el efecto de los cambios de las variables explicativas sobre la probabilidad de que **Y** valga 1 (probabilidad del suceso de interés que denominamos éxito).

Sea p=P(Y=1) la probabilidad de que ocurra el suceso de interés. Por ejemplo, probabilidad de que un trabajador sufra un accidente; 1-p=P(Y=0) es la probabilidad del suceso contrario, que denominamos fracaso (no sufra accidente)

Modelo de Regresión Logística

Dada una variable independiente X, el modelo de regresión logística simple es

$$\log it = \ln \frac{p}{1-p} = \beta_0 + \beta_1 x$$

donde los logits son funciones lineales de las variables explicativas (pero no las probabilidades)

Se denomina **Odds** (o ventaja) a la razón de una probabilidad a su valor complementario

$$Odds = \frac{p(Y=1)}{p(Y=0)} = \frac{p(Y=1)}{1-p(Y=1)}$$

El modelo expresado de forma equivalente en términos de Odds es:

$$\frac{p}{1-p} = e^{\beta_0 + \beta_1 x}$$

Modelo de Regresión Logística

Otro modo equivalente de expresar el modelo es en términos de la probabilidad

$$\frac{p}{1-p} = e^{\beta_0+\beta_1 x} \Rightarrow p = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

Dados dos valores de la variable independiente X se puede determinar la razón de odds a partir del coeficiente del modelo, correspondiente a dicha variable.

$$RO(x2/x1) = \frac{Odd(x2)}{Odd(x1)} = e^{\beta 1(x2-x1)}$$

Modelo de Regresión Logística

Por ejemplo, la razón de odds

$$RO(x+1/x) = \frac{Odd(x+1)}{Odd(x)} = e^{\beta 1}$$

permite comparar por cociente las odds de la variable respuesta en dos situaciones caracterizadas por los valores adoptados por la variable independiente x

Este estadístico da una idea de cúanto es más (o menos) probable el suceso al pasar de x a x+1, es decir, al aumentar una unidad la variable independiente.

Nota: Observe que

$$\ln(RO) = \log it(x+1) - \log it(x) = \beta_1$$

Modelo de Regresión Logística

El modelo de **regresión logística múltiple** con k variables explicativas es

$$\log it = \beta_0 + \sum_{j=1}^{k} \beta_j x_j$$

Que también podemos expresar mediante:

$$P(Y = 1/ X) = \frac{e^{\beta_0 + \sum_k \beta_k x_k}}{1 + e^{\beta_0 + \sum_k \beta_k x_k}}$$

Para decidir qué factores o variables independientes son importantes para describir la probabilidad de ocurrencia del suceso de interés se utilizan los contrastes de hipótesis de nulidad de los coeficientes del modelo.

Modelo de Regresión Logística

Contraste de hipótesis de nulidad de los parámetros

Uno de los más usados es el **test de Wald** que se efectúa para cada una de las variables que intervienen en el modelo. Para un coeficiente cualquiera, **βj**, se verifica (para muestras suficientemente grandes) que bajo la hipótesis nula **H0: βj = β0**, el estadístico **w** definido por:

$$w = \frac{(b_j - \beta_0)^2}{Var(b_j)} \rightarrow \chi_1$$

sigue un modelo Chi-cuadrado con 1 g.l.

En R el contraste viene expresado aproximando a una z con su correspondiente p-valor mediante la función summary aplicada al objeto (modelos glm ajustado)

Permite establecer qué variables son importantes para explicar la probabilidad del suceso (Y=1), mediante el contraste de hipótesis
H0: βj = 0 frente a la alternativa
H1: βj ≠0

Decisión:
Si la sig. del estadístico o p-valor es inferior a 0,05, se rechaza H0 al nivel alfa del 5%. Caso contrario, se acepta.

Modelo de Regresión Logística

Comparación de modelos anidados

Un test muy corriente en glm es el que permite establecer la significatividad de un solo o de varios términos de un modelo, en base al cambio registrado en la deviance.

Este test es muy útil para comparar modelos anidados. Los términos del modelo menos complejo forman parte del modelo más complejo.

En R el contraste se realiza mediante la función anova

Permite establecer qué variable o conjunto de variables son importantes para explicar la probabilidad, mediante un contraste Chi-cuadrado.

Decisión:
Si la sig. del estadístico o p-valor es inferior a 0,05, el término o al menos alguno del conjunto de términos contrastados sirven para explicar el modelo al nivel alfa del 5%.
Caso contrario, se acepta la nulidad de los términos.