TEMA 4
Regresión logística

José R. Berrendero
Departamento de Matemáticas
Universidad Autónoma de Madrid

Análisis de Datos - Grado en Biología

Esquema del tema

- Variable respuesta dicotómica. Ejemplo.
- El modelo de regresión logística.
- Estimación e inferencia sobre los parámetros.
- Salida de SPSS.
- Aplicación al problema de clasificación.

Variable respuesta dicotómica. Ejemplo.

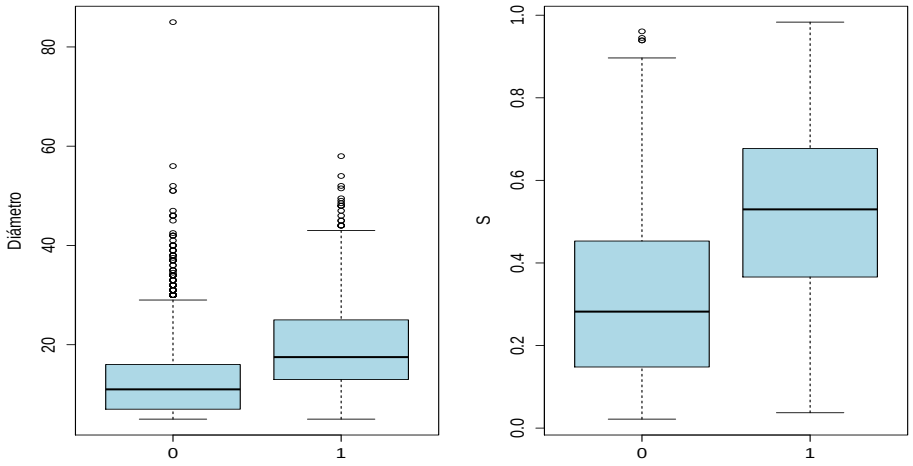
El 4 de julio de 1999 una tormenta con vientos que excedían las 90 millas por hora azotó el nordeste de Minnesota, en EE.UU., causando graves daños en los bosques de un parque natural de la zona.

Los científicos analizaron los efectos de la tormenta determinando para más de 3600 árboles del parque:

- Su diámetro en cm (variable `D`).
- Una medida de la severidad local de la tormenta relacionada con el porcentaje inerte de área basal de cuatro de las especies (variable `S`).
- Una variable que registraba si cada árbol había muerto ($y = 1$) o si había sobrevivido ($y = 0$).

El problema es determinar cómo influyen las dos primeras variables en la tercera. En este ejemplo la variable respuesta es dicotómica.

Variable respuesta dicotómica. Ejemplo.



- Los árboles que sobreviven tienden a tener un menor diámetro. La fuerza de la tormenta tiende a ser menor en las zonas correspondientes a los árboles supervivientes.
- Parece que el diámetro y la variable S pueden ser útiles para estimar la probabilidad de supervivencia de un árbol.
- El modelo de regresión logística relaciona $\mathrm{P}(Y = 1)$ con un conjunto de variables regresoras.

El modelo de regresión logística

Disponemos de n observaciones. Cada observación $(Y_i, x_{i1}, \ldots, x_{ik})$ está formada por el valor de la variable respuesta Y_i y un vector de variables regresoras $x_i = (x_{i1}, \ldots, x_{ik})$.

Las variables $Y_1, \ldots, Y_n$ son independientes y tienen distribución de Bernoulli.

Denotamos $p_i = P(Y_i = 1 \mid x_i)$. La probabilidad de "éxito" depende de las variables regresoras.

Una relación lineal $p_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$ no es adecuada (¿por qué?)

El modelo de regresión logística

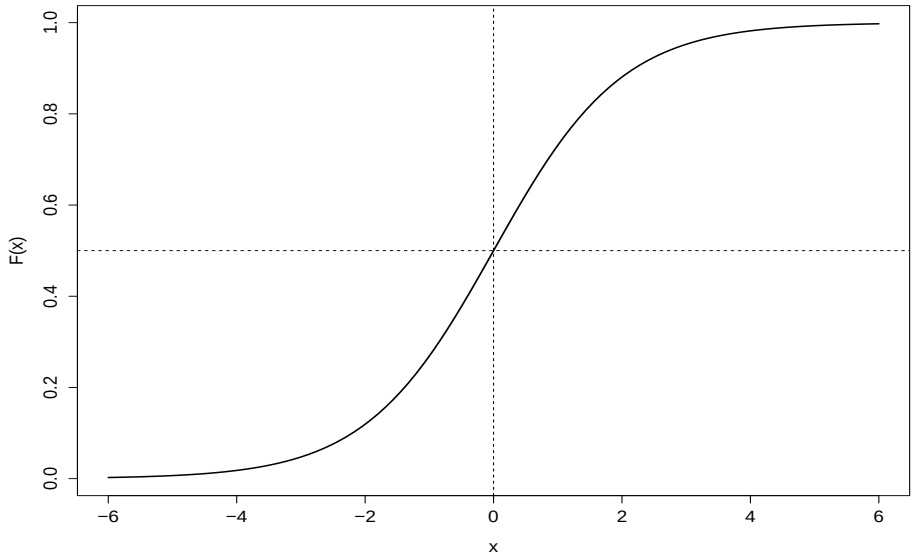
Suponemos que la relación entre p_i y x_i viene dada por

$$p_i = \frac{1}{1 + e^{-\beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik}}},$$

es decir,

$$p_i = F(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}),$$

donde $F(x) = 1/(1 + e^{-x})$ es la **función logística**.



Interpretación de los parámetros del modelo

Llamamos O_i a la **razón de probabilidades** para la observación i:

$$O_i = \frac{p_i}{1 - p_i}$$

¿Cómo se interpreta el valor de O_i? ¿Qué significa, por ejemplo, $O_i = 2$? Si se cumple el modelo de regresión logística, entonces

$$O_i = e^{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}}$$

¿Cómo varía la razón de probabilidades si la variable regresora x_{ij} se incrementa una unidad?

$$\frac{O_i'}{O_i} = \frac{e^{\beta_0 + \cdots + \beta_j (x+1) + \cdots + \beta_k x_{ik}}}{e^{\beta_0 + \cdots + \beta_j x + \cdots + \beta_k x_{ik}}} = e^{\beta_j}.$$

Por tanto e^{β_j} es la variación de la razón de probabilidades cuando la variable regresora j se incrementa en una unidad y el resto de variables permanece constante.

Estimación

Para estimar los parámetros se usa el método de máxima verosimilitud.

Por ejemplo, si observamos los datos

x_i	2	1	3
Y_i	0	1	1

entonces $\hat{\beta}_0$ y $\hat{\beta}_1$ son los valores que maximizan la función de verosimilitud

$$L(\beta_0, \beta_1) = P(Y = 0 \mid x = 2)P(Y = 1 \mid x = 1)P(Y = 1 \mid x = 3)$$

$$L(\beta_0, \beta_1) = \left(1 - \frac{1}{1 + e^{-\beta_0 - 2\beta_1}}\right)\left(\frac{1}{1 + e^{-\beta_0 - \beta_1}}\right)\left(\frac{1}{1 + e^{-\beta_0 - 3\beta_1}}\right)$$

Es necesario el ordenador para calcular estos valores mediante algún método numérico de optimización.

Ejemplo

$$p_i = P(Y_i = 1 \mid D_i, S_i) = \frac{1}{1 + e^{-\beta_0 - \beta_1 D_i - \beta_2 S_i}},$$

Resumen del modelo

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	3883,256 [a]	,274	,366

a. La estimación ha finalizado en el número de iteración 5 porque las estimaciones de los parámetros han cambiado en menos de ,001.

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1 [a]	D	,097	,005	346,022	1	,000	1,102
	S	4,424	,189	545,122	1	,000	83,412
	Constante	-3,543	,127	774,463	1	,000	,029

a. Variable(s) introducida(s) en el paso 1: D, S.

Cuestiones y observaciones

- ¿Cuál es el valor de $\hat{\beta}_1$? ¿Cuál es el error típico de este estimador?
- Interpreta el valor del estimador anterior.
- ¿Qué puede significar la nota bajo la tabla que comienza *La estimación ha finalizado en el número de iteración 5...*?
- El valor 3883,256 (-2 log de la verosimilitud) corresponde a:

$$-2 \log L(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = 3883{,}256$$

- Los coeficientes de Cox-Snell y de Nagelkerke son medidas de la bondad del ajuste del modelo a los datos. Se interpretan de forma similar al coeficiente de determinación.

Intervalos y contrastes sobre los parámetros

Los valores

$$z_j = \frac{\hat{\beta}_j - \beta_j}{\text{error típico de } \beta_j}$$

tienen aproximadamente distribución normal estándar.

Intervalo de confianza para β_i:

$$\text{IC}_{1-\alpha}(\beta_i) = [\hat{\beta}_i \mp z_{\alpha/2} \times \text{error típico de } \hat{\beta}_i].$$

Intervalo de confianza para e^{β_i}:

$$\text{IC}_{1-\alpha}(e^{\beta_i}) = [\exp(\hat{\beta}_i - z_{\alpha/2} \times \text{e.t. de } \hat{\beta}_i), \exp(\hat{\beta}_i + z_{\alpha/2} \times \text{e.t. de } \hat{\beta}_i)].$$

Intervalos y contrastes sobre los parámetros

Bajo $H_0 : \beta_i = 0$, el **estadístico de Wald**

$$z_i = \frac{\hat{\beta}_i}{\text{error típico de } \hat{\beta}_i}$$

tiene aproximadamente una distribución normal estándar.

Este valor (elevado al cuadrado) aparece en la columna `Wald` del cuadro de SPSS.

Región de rechazo (con nivel de significación aproximado α) para $H_0 : \beta_i = 0$:

$$R = \left\{ \left(\frac{\hat{\beta}_i}{\text{error típico de } \hat{\beta}_i} \right)^2 > \chi^2_{1;\alpha} \right\}.$$

El p-valor correspondiente aparece en la columna `Sig` de la tabla de SPSS.

Ejemplos

- Calcula un IC de nivel 0,95 para β_1.
- ¿Podemos afirmar a nivel 0,01 que el diámetro de los árboles está relacionado con la probabilidad de su supervivencia?
- Calcula un IC de nivel 0,90 para e^{β_2}

Aplicación al problema de clasificación

Problema de clasificación: dado un conjunto de observaciones que se sabe que pertenecen a uno de dos posibles grupos, clasificar una nueva observación cuyo grupo es desconocido.

Predecir el valor de Y_0 dada una nueva observación $x_0 = (x_{01}, \ldots, x_{0k})$ equivale a clasificar x_0 en uno de los dos grupos.

Si $\hat{p}_0$ es la probabilidad estimada de $Y_0 = 1$ dado x_0,

$$\hat{p}_0 = \frac{1}{1 + e^{-\hat{\beta}_0 - \hat{\beta}_1 x_{01} - \cdots - \hat{\beta}_k x_{0k}}},$$

¿cómo se puede usar el valor $\hat{p}_0$ para clasificar x_0?

Regla de clasificación:

$$\hat{Y}_0 = 1 \iff \hat{p}_0 > 1/2 \iff \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \cdots + \hat{\beta}_k x_{0k} > 0.$$

Ejemplo

En el ejemplo, clasificamos un árbol como **no superviviente** ($\hat{Y}_0 = 1$) si las correspondientes variables D_0 y S_0 verifican

$$-3{,}543 + 0{,}097 \cdot D_0 + 4{,}424 \cdot S_0 > 0$$

¿En qué grupo clasificarías un árbol cuyo diámetro es 20 cm y está situado en una zona en la que la tormenta ha tenido una fuerza S=0.3?

En la figura siguiente se representan los pares (D_i, S_i) en color rojo (no supervivientes) o color verde (supervivientes) junto con la recta que divide las zonas en las que un nuevo árbol se clasifica como superviviente o no superviviente.

Ejemplo