Tema 3: Técnicas de clasificación
Regresión logística

Inmaculada Barranco Chamorro
(chamorro@us.es)

Departamento de Estadística e Investigación Operativa

Universidad de Sevilla

Indice

Objetivo:

Predecir el valor de una variable respuesta a partir de una serie de variables dadas.

• Si la variable es **numérica** $\Rightarrow$ **Regresión múltiple**

• Si la variable es **dicotómica** $\Rightarrow$ **Regresión logística**

- Ejemplo: Ver si se desarrolla o no una enfermedad a partir de variables como la edad, fumar o no, tener o no antecedentes familiares, etc.

Modelo de regresión logística

es útil para abordar este tipo de cuestiones bajo la condiciones de que

- Hayamos tenido en cuenta todas las **variables importantes** para explicar la variable respuesta
- **Muestra** se haya tomado **adecuadamente**

Objetivo:

Predecir el valor de una variable respuesta a partir de una serie de variables dadas.

• Si la variable es **numérica** $\Rightarrow$ **Regresión múltiple**

• Si la variable es **dicotómica** $\Rightarrow$ **Regresión logística**

- Ejemplo: Ver si se desarrolla o no una enfermedad a partir de variables como la edad, fumar o no, tener o no antecedentes familiares, etc.

Modelo de regresión logística

es útil para abordar este tipo de cuestiones bajo la condiciones de que

- Hayamos tenido en cuenta todas las **variables importantes** para explicar la variable respuesta
- **Muestra** se haya tomado **adecuadamente**

Conceptos básicos: Comparaciones de riesgos

En 1 de cada 200 nacimientos ocurre un parto gemelar

- **Probabilidad o riesgo**

$$R_1 = \frac{1}{200}$$

$$R_1 = \frac{\text{casos favorables}}{\text{casos posibles}}$$

- **Odds**

$$O_1 = \frac{1}{199}$$

$$O_1 = \frac{\text{núm. casos en que ocurre}}{\text{núm. casos en que no ocurre}}$$

- **Observamos** que

$$O = \frac{p}{1-p}$$

Introducimos un factor de riesgo

En mujeres que han tomado ácido fólico durante el embarazo se observa que 3 de cada 200 partos eran gemelares.

$$R_2 = \frac{3}{200}, \qquad O_2 = \frac{3}{197}$$

¿Cómo expresar numéricamente el aumento de riesgo de embarazo gemelar? Hay dos formas

- **Riesgo relativo (RR)**

$$RR = \frac{R_2}{R_1} = \frac{3/200}{1/200} = 3$$

$$RR = \frac{\text{probabilidad en expuestos}}{\text{probabilidad en no expuestos}}$$

- **Odds ratio (OR)**

$$OR = \frac{O_2}{O_1} = \frac{3/197}{1/199} = 3.03$$

$$OR = \frac{\text{oportunidad en expuestos}}{\text{oportunidad en no expuestos}}$$

Interpretación de la OR

- $OR = 1$.
  No existe tal factor de riesgo, la oportunidad en los no expuestos es la misma que en los expuestos.
- $OR > 1$.
  Se ha **localizado un factor de riesgo**, un posible efecto dañino, porque la *oportunidad* en los expuestos es mayor que en los no expuestos.
  Puede interesar p.e. en epidemiología para localizar factores de riesgo.
- $OR < 1$.
  La oportunidad de que ocurra el suceso es menor en los individuos expuestos al tratamiento que en los no expuestos.
  Interesa cuando estudiemos **tratamientos para reducir** la frecuencia de un suceso (p.e. para reducir la mortalidad)

Interpretación de la OR

- $OR = 1$.
  No existe tal factor de riesgo, la oportunidad en los no expuestos es la misma que en los expuestos.
- $OR > 1$.
  Se ha **localizado un factor de riesgo**, un posible efecto dañino, porque la *oportunidad* en los expuestos es mayor que en los no expuestos. Puede interesar p.e. en epidemiología para localizar factores de riesgo.
- $OR < 1$.
  La oportunidad de que ocurra el suceso es menor en los individuos expuestos al tratamiento que en los no expuestos.
  Interesa cuando estudiemos **tratamientos para reducir** la frecuencia de un suceso (p.e. para reducir la mortalidad)

Interpretación de la OR

- $OR = 1$.
  No existe tal factor de riesgo, la oportunidad en los no expuestos es la misma que en los expuestos.

- $OR > 1$.
  Se ha **localizado un factor de riesgo**, un posible efecto dañino, porque la *oportunidad* en los expuestos es mayor que en los no expuestos.
  Puede interesar p.e. en epidemiología para localizar factores de riesgo.

- $OR < 1$.
  La oportunidad de que ocurra el suceso es menor en los individuos expuestos al tratamiento que en los no expuestos.
  Interesa cuando estudiemos **tratamientos para reducir** la frecuencia de un suceso (p.e. para reducir la mortalidad)

Propiedades matemáticas de la OR

- $OR \in (0, \infty)$.
  Al tomar logaritmo, tomará valores en todo R.
  Facilita su **tratamiento matemático** (*regresión logística*)
- Utilizaremos el **modelo de regresión logística** para **obtener intervalos de confianza para la OR**:
  - Si los intervalos contienen al valor $OR = 1$, podemos suponer que no hay tal factor de riesgo (la oportunidad para los expuestos puede considerarse la misma que para los no expuestos)
  - Diremos que aumenta la oportunidad del suceso si todos los valores del intervalo de confianza son mayores que 1.
  - Diremos que disminuye la oportunidad del suceso si todos los valores del intervalo de confianza son menores que 1.
- Cuando se evalúa la eficacia de una prueba diagnóstica utilizando un estudio caso-control, siempre podremos estimar la OR.

• **Nota:** Si el suceso de interés es raro, la OR puede considerarse como una aproximación del RR, que tiene una interpretación más natural.

Regresión logística binaria

- Variable dependiente binaria (enfermar o no, vivir o no, que se presente una mutación genética o no)
- Queremos estudiar el efecto que tienen sobre ella otras variables independientes (fumar, edad).

Modelo de regresión logística binaria permite

- Dados los valores de las variables independientes, **estimar la probabilidad** de que se presente el suceso de interés (p.e. enfermar)
- **Evaluar el efecto** que cada variable independiente tiene sobre la variable respuesta en forma de OR.

(OR mayor que indica un aumento en la probabilidad del suceso que nos interesa, y OR menor que 1, implica disminución)

Regresión logística binaria

- Variable dependiente binaria (enfermar o no, vivir o no, que se presente una mutación genética o no)
- Queremos estudiar el efecto que tienen sobre ella otras variables independientes (fumar, edad).

Modelo de regresión logística binaria permite

- Dados los valores de las variables independientes, **estimar la probabilidad** de que se presente el suceso de interés (p.e. enfermar)
- **Evaluar el efecto** que cada variable independiente tiene sobre la variable respuesta en forma de OR.

(OR mayor que indica un aumento en la probabilidad del suceso que nos interesa, y OR menor que 1, implica disminución)

¿Cómo construir el modelo de regresión logística?

Necesitamos

- Conjunto de variables independientes.
- Variable dependiente o **respuesta dicotómica**.
  (Principal diferencia con el modelo de regresión múltiple)

Destacamos que

la regresión logística es una **solución óptima** para controlar múltiples variables de confusión (categóricas y continuas)

Regresión logística binaria simple

Y variable binaria

$$Y = \begin{cases} 1, & \text{si un hecho ocurre} \\ 0, & \text{si el hecho } \textbf{no} \text{ ocurre} \end{cases}$$

- $\pi = P[Y = 1]$
- **Modelo**

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta X \tag{2.1}$$

Recordemos que

$$odds = \frac{\pi(x)}{1-\pi(x)}$$

La regresión logística modela **linealmente** el logaritmo de la odds de un suceso.

Modelo

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta X \tag{2.2}$$

Observación: ¿Cómo predecir π?

De (2.2), se tiene que

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \tag{2.3}$$

Ejemplo

Datos de un estudio de la relación entre la edad y la mortalidad por cardiopatía isquémica (CI) en diabéticos

		Causa	Muerte		
		CI	Otra	n_i	f_i
Edad	20-30	1	9	10	0.10
	30-35	2	13	15	0.13
	35-40	3	9	12	0.25
	40-45	5	10	15	0.33
	45-50	6	7	13	0.46
	50-55	5	3	8	0.63
	55-60	13	4	17	0.76
	60-70	8	2	10	0.80
	Total	43	53	100	0.43

Predecir la probabilidad de muerte por CI a partir de la edad

- **No** podemos (debemos) utilizar **regresión lineal** porque la variable dependiente es binaria
- Utilizaremos regresión logística
- $$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = a + bX$$

  $$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = -5{,}091 + 0{,}105X$$

Interpretación del coeficiente *b*

¿Qué se puede hace igual que en regresión lineal?

- Denotamos $s^* = -5{,}091 + 0{,}105X$
- Algunos aspectos serán similares a la regresión lineal

signo de b

- **Signo de b positivo**: al aumentar X aumenta π

  (Al aumentar la edad, aumenta la probabilidad de enfermar)

- Como en regresión lineal, se puede **determinar si una variable es estadísticamente significativa**, para un nivel de significación α dado.

Método de Wald

- **Contraste**

$$\begin{aligned} H_0 &: \beta = 0 \\ H_1 &: \beta \neq 0 \end{aligned}$$

Considerar el intervalo de confianza

$$\left(b - z_{1-\alpha/2} s.e.(b),\ b + z_{1-\alpha/2} s.e.(b)\right),$$

con $s.e.(b)$: estimador del error estándar de b.

a) Si ese intervalo **contiene al cero**, no rechazo $H_0 : \beta = 0$. La variable X no es relevante en el modelo, puedo omitirla

b) Si ese intervalo **no contiene al cero**, rechazo H_0. La variable dependiente X es relevante (es estadísticamente significativa).

- El contraste anterior se puede hacer **también con el p-valor**. (Salida de programas estadísticos).

Interpretación del coeficiente b

¿Qué es diferente?
La **interpretación de los coeficientes**. En concreto de sus **magnitudes**

- ¿Qué quiere decir 0.105 en la siguiente ecuación?

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = -5{,}091 + 0{,}105 \ \text{Edad}$$

- Por cada año adicional, $\ln\left(\frac{\pi(x)}{1-\pi(x)}\right)$ aumenta 0.105 unidades.

- **¿Qué es $\ln\left(\frac{\pi(x)}{1-\pi(x)}\right)$?**

- **¿Cómo realizar una predicción?**
  Denotemos por $s^* = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = -5{,}091 + 0{,}105x$

  - P.e. para $x = 40$

$$\begin{aligned} s^* &= -5{,}091 + 0{,}105 \times 40 = -0{,}891 \\ odds &= \pi/(1-\pi) = e^{s*} = e^{-0{,}891} = 0{,}410 \\ \pi &= \frac{odds}{1+odds} = 0{,}291 \end{aligned}$$

Impacto de un año más en la ecuación que hemos estimado

$$x = 40, \quad s^* = -0{,}891, \quad odds = e^{-0{,}891} = 0{,}410, \quad \pi = 0{,}291$$
$$x = 41, \quad s^* = -0{,}786, \quad odds = e^{-0{,}786} = 0{,}456, \quad pi = 0{,}313$$

Cociente de odds

$$\frac{odds(41)}{odds(40)} = 1{,}112$$

- Es la OR del aumento del valor de x en una unidad.
- Veremos que es igual a e^b

$$\frac{odds(41)}{odds(40)} = e^b = e^{0{,}105} = 1{,}112$$

Propiedad

En general dados x_1, x_2

$$\frac{odds(x_2)}{odds(x_1)} = e^{b(x_2 - x_1)}$$

Para el **caso particular** en que $x_2 = x_1 + 1$

$$\frac{odds(x+1)}{odds(x)} = e^{b}$$

Conclusión

e^b es la OR del aumento del valor de x en una unidad.

Modelo de regresión logística binaria múltiple

$$\log\left(\frac{\pi(\underline{x})}{1-\pi(\underline{x})}\right) = b_0 + b_1 x_1 + \ldots + b_k x_k \qquad (2.4)$$

donde π es la probabilidad de que ocurra el suceso de interés, x_i son las variables independientes, y b_i son los coeficientes asociados.

Predicciones de las probabilidades

• Dado el valor de las variables independientes, podemos calcular la estimación de la probabilidad o riesgo de que ocurra el suceso que nos interesa como:

$$\pi(\underline{x}) = \frac{e^s}{1+e^s}, \qquad s = b_0 + b_1 x_1 + \ldots + b_k x_k$$

Ejemplo

Estudiamos la aparición o no de una enfermedad coronaria (EC) en varones durante un cierto periodo de tiempo.

Consideramos $k = 3$ variables que se miden al comienzo del estudio, y que presumiblemente influyen en el proceso.
X_1 : edad del individuo (EDAD)
X_2 : hábito de fumar (HF) (1:fuma, 0:no)
X_3 : tensión arterial sistólica (TAS) en mm

De los datos recogidos en el estudio, se estima el modelo

$$logit(\pi(\underline{x})) = \ln\left(\frac{\pi(\underline{x})}{1-\pi(\underline{x})}\right) = -6{,}614 + 0{,}075X_1 + 0{,}312X_2 + 0{,}018X_3$$

Interpretación del signo de b_2

X_2 es una variable **dicotómica**

Predicciones

• Para un varón de 58 años, fumador ($x_2 = 1$), y con TAS de 150 mm, predecir la probabilidad de que aparezca EC.

Denotamos predicción

$$s^* = ln\left(\frac{\pi(\underline{x})}{1-\pi(\underline{x})}\right) = -6{,}614 + 0{,}075 \times 58 + 0{,}312 \times 1 + 0{,}018 \times 150 = 0{,}750$$

$$\begin{aligned} s^* &= 0{,}750 \\ odds &= \pi(\underline{x})/(1-\pi(\underline{x})) = e^{s^*} = e^{0{,}750} = 2{,}12 \\ \pi(\underline{x}) &= \frac{odds}{1+odds} = 0{,}679 \end{aligned}$$

Comparación de 2 perfiles concretos

$X^1 = (X_1^1, \ldots, X_k^1)$
$X^0 = (X_1^0, \ldots, X_k^0)$

$$\frac{odds(X^1)}{odds(X^0)} = \exp\left\{\sum_{i=1}^{k} bi(x_i^1 - x_i^0)\right\}$$

Medida relativa que permite valorar el riesgo de un perfil respecto de otro

Ejemplo

¿Cuánto más peligro tiene de desarrollar una EC un individuo de 65 años, fumador, y con TAS de 175mm, que uno de 58, no fumador, y con TAS de 150?
$X^1 = (65, 1, 175)$
$X^0 = (58, 0, 150)$

$$\begin{aligned}\frac{odds(X^1)}{odds(X^0)} &= \exp\left\{\sum_{i=1}^{k} bi(x_i^1 - x_i^0)\right\} \\ &= \exp\{b_1(65-58) + b_2(1-0) + b_3(175-150)\} = 3{,}62\end{aligned}$$

La primera situación es 3.6 veces más peligrosa que la segunda

Si los perfiles son iguales salvo en una de las variables X_i

$$\frac{odds(X^1)}{odds(X^0)} = \exp\left\{bi(x_i^1 - x_i^0)\right\}$$

En particular, si $x_i^1 = x_i^0 + 1$ entonces

$$\frac{odds(X^1)}{odds(X^0)} = \exp\{bi\}$$

Ejemplo

Dos individuos que sólo difieren en que uno fuma y otro no

$$\frac{odds(X^1)}{odds(X^0)} = \exp\{0{,}312\} = 1{,}37$$

Es 1.37 veces más peligroso, teniendo iguales valores de edad y TAS.

Codificación de las variables

Para interpretar adecuadamente los resultados de un modelo de regresión logística se deben seguir las siguientes recomendaciones:

• **Variable dependiente**: se codifica como 1 que ocurra el suceso que nos interesa y como 0 que no ocurra.

• **Variables independientes**: pueden ser varias y de distintos tipos.

Distinguimos

- **Caso dicotómico**
- **Caso categórico**
- **Caso de variable numérica**

Caso dicotómico

Se codifica **como 1** la situación que creemos que favorece la ocurrencia del suceso, y como 0 el caso contrario.

- Ejemplo de partos gemelares, codificaríamos como

1: tomaron ácido fólico $\equiv$ casos expuestos,

0: no tomaron ácido fólico $\equiv$ no expuestos al posible factor de riesgo, casos de referencia o grupo control

Caso categórico

- Nos referimos a una variable categórica con **más de 2 categorías**, tenemos así una variable independiente que puede tomar más de 2 valores.
- Se codifican las categorías utilizando variables indicadoras (*dummy*), de forma similar a como se hace en el modelo de regresión lineal múltiple. Algunos programas ayudan a hacerlo.
- Es necesario destacar una modalidad que represente el caso de referencia, al que le deben corresponder la codificación con todas las variables indicadoras puestas a 0.

Ejemplo

Resultados sobre la presencia de anticuerpos inherentes a ciertos virus según zonas de una región: Norte, Sur, Este, Oeste

Zonas	Z_1	Z_2	Z_3
Norte	0	0	0
Sur	1	0	0
Este	0	1	0
Oeste	0	0	1

Caso de variable numérica

Pueden darse 2 situaciones:

- Si creemos que por cada unidad que aumente la variable, la odds aumenta en un factor multiplicativo constante, entonces podemos **usar la variable tal cual** en el modelo. Si tenemos dudas, pasar a la opción siguiente.

- Creemos que la variable numérica puede afectar a la respuesta, pero no tenemos muy claro cómo, podemos **categorizar** la variable. P.e. **estratificando** la variable en valores pequeños, medianos y grandes. Los puntos de corte los podemos elegir nosotros manualmente, o usar cortes automáticos basados en que cada categoría tenga el mismo número de observaciones (usando p.e. percentiles). Algunos programas ayudan a hacerlo.

Recapitulación

• **Odds para los individuos de referencia o control** (aquellos para los que x_i valen 0, si seguimos las referencias de codificación) sería:

$$e^{b_0}$$

• Para **cualquier otro coeficiente** del modelo, se tiene que

$$e^{b_i}$$

coincide con la OR del aumento del valor de x_i en una unidad con respecto a aquellos individuos que presentan los valores de todas las demás variables iguales

- Si se ha seguido el criterio de codificación recomendado, y la variable que estamos considerando es dicotómica, tenemos la OR del factor de riesgo x_i.
- Si la variable es numérica, como p.e. el número de bypass coronarios, se está estimando la OR del factor de riesgo tener un bypass más.

Requisitos y limitaciones

Destacamos los siguientes factores a tener en cuenta para que el modelo sea válido.

- Los parámetros del modelo se estiman por **máxima verosimilitud**.
- Para que este método sea válido debemos tener un número suficientemente alto de observaciones para cada combinación de variables independientes.
- Si las estimaciones de los parámetros son anormalmente grandes, es posible que se viole esta condición.
- Se puede intentar solucionar el problema agrupando categorías (donde tenga sentido).

Requisitos

• **No** se deben introducir variables **innecesarias**.

• **No** se debe **excluir** ninguna variable **relevante** en el proceso. Si identificamos variables confusoras, deben tenerse en cuenta, introduciéndolas en el modelo, o estratificando el estudio en submuestras.

• Al igual que ocurría en regresión lineal múltiple, **puede haber problemas de colinealidad**. Si los errores típicos en la estimación de los coeficientes, o los intervalos de confianza son anormalmente grandes, es posible que se esté dando este problema.