Regresión logística y multinomial

January 2024

Modelo de regresión logística

El contexto

- Deseamos predecir una variable respuesta binaria Y que solo toma los valores 0 y 1.
  - → Además, estos valores numéricos solo indicarán la pertenencia o no a un determinado grupo.
- `Ejemplos`:
  - → Determinar si un paciente tiene o no una determinada enfermedad en función de diferentes variables (edad, presión arterial, nivel de colesterol, etc.).
    - ⇾ En este caso el valor 1 suele indicar que sí la tiene y 0 que no.
  - → Predecir si un estudiante aprueba o no un examen en función de las horas de estudio.
  - → Predecir si un mensaje de correo electrónico es spam o no en función de las palabras clave.
  - → Predecir si un cliente comprará o no un determinado producto en función de la edad y el salario.
  - → Predecir si un paciente tiene diabetes o no en función de variables como el nivel de glucosa, la presión arterial y el índice de masa corporal (IMC).

Objetivo

- `Objetivo`: predecir la variable respuesta Y a partir de k variables numéricas $X_1, \ldots, X_k$ utilizando una única función

$$h_{\boldsymbol{\theta}}(\mathbf{x}) = g(\boldsymbol{\theta}'\mathbf{x}) = g(\theta_0 + \theta_1 x_1 + \cdots + \theta_k x_k)\,,$$

  donde $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_k)' \in \mathbb{R}^{k+1}$ contiene los parámetros del modelo y $\mathbf{X} = (X_0, \ldots, X_k)'$ las variables que podemos medir en nuestros individuos para predecir Y.
- Para mejorar la notación hemos incluido una variable artificial X_0 que siempre vale 1.

¿Cómo elegir la función g?

- La función g debe transformar esos valores numéricos (lineales) en números entre 0 y 1 que nos indicarán la `probabilidad` de que el individuo pertenezca al grupo ($Y = 1$):

$$g : \mathbb{R} \to [0, 1]$$

y $h_{\boldsymbol{\theta}}(\mathbf{x}) \approx Pr(Y = 1|\mathbf{X} = \mathbf{x})$, donde $\mathbf{X} = (X_0, \ldots, X_k)'$.

- `Regla de decisión`:

$$h_{\boldsymbol{\theta}}(\mathbf{x}) \geq 0.5 \to \hat{y} = 1$$
$$h_{\boldsymbol{\theta}}(\mathbf{x}) < 0.5 \to \hat{y} = 0$$

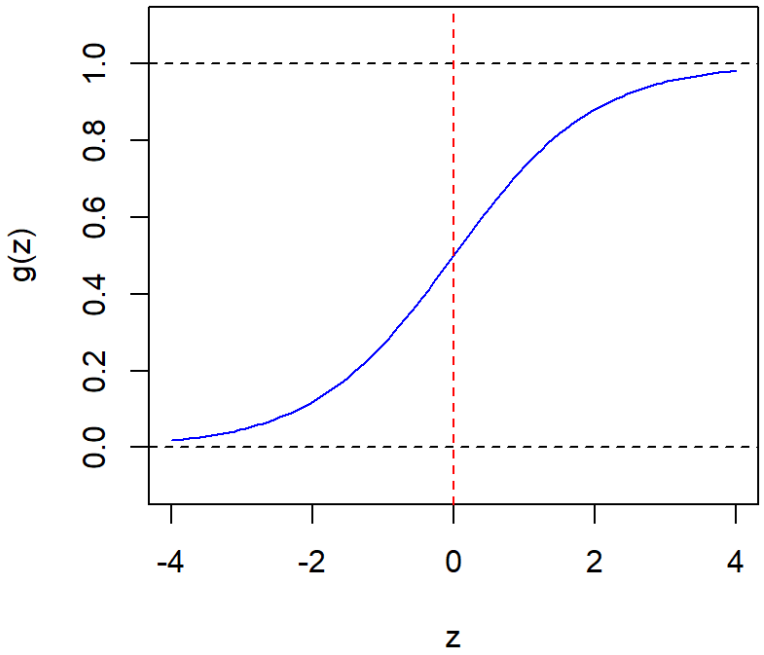
donde $\hat{y}$ representa el valor que predecimos para Y cuando $\mathbf{X} = \mathbf{x}$.

- Existen diversas opciones para determinar g, la más popular es la `función logística` (o `sigmoide`)

$$g(z) = \frac{1}{1 + \exp(-z)} = \frac{\exp(z)}{1 + \exp(z)}.$$

Función logística

- Code



Propiedades

- Es continua
- Estrictamente creciente
- Recorrido de 0 a 1,
- Transformará el valor $\boldsymbol{\theta}'\mathbf{x} \in \mathbb{R}$ en un valor $h_{\boldsymbol{\theta}}(\mathbf{x}) \in [0, 1]$.
- $g(0) = 0.5$
- `Regla de decisión`:

$$\boldsymbol{\theta}'\mathbf{x} \geq 0 \rightarrow \hat{y} = 1$$
$$\boldsymbol{\theta}'\mathbf{x} < 0 \rightarrow \hat{y} = 0$$

- La función $I_{\boldsymbol{\theta}}(\mathbf{x}) = \boldsymbol{\theta}'\mathbf{x}$ define un `índice de separación` entre las categorías de Y.

¿Cómo determinar una función costo que penalice las decisiones erróneas?

Función costo

- Para $z = h_{\boldsymbol{\theta}}(\mathbf{x})$, definimos la función costo

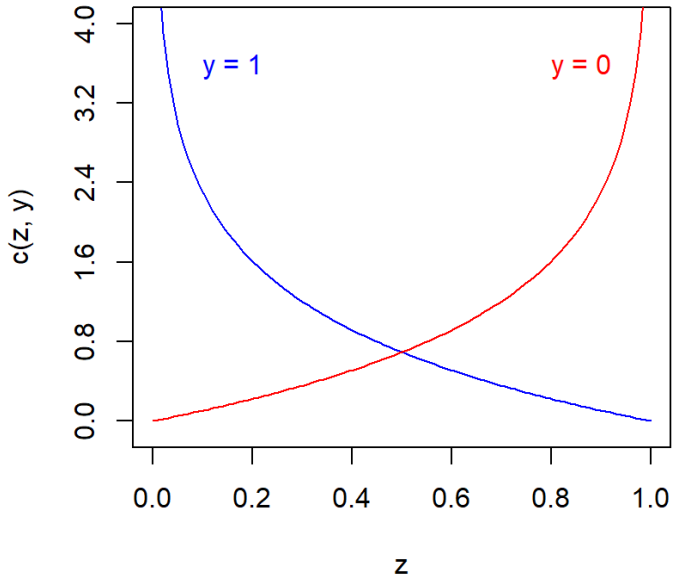
$$c(z, y) = \begin{cases} -\log(z) & \text{si } y = 1 \\ -\log(1-z) & \text{si } y = 0 \end{cases}$$

- O, equivalentemente,

$$c(z, y) = -y \log(z) - (1-y) \log(1-z)\,,$$

donde $y \in \{0, 1\}$ y los logaritmos son neperianos.

▶ Code



Criterio

- Minimizar el valor esperado de la función costo

$$\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = E[c(h_{\boldsymbol{\theta}}(\mathbf{X}), Y)]$$

- Para determinar los valores óptimos de los parámetros:
  - → Dispondremos de una muestra (`training sample`) de individuos en los que se conozcan tanto los valores de $\mathbf{x}$ como los valores de y (`aprendizaje supervisado`).
  - → Calcularemos los costos en los valores muestrales.
  - → Determinaremos los valores de los parámetros $\boldsymbol{\theta}$ que minimizan estos costos.

Otra formulación del problema

- Si $p = Pr(Y = 1)$ este modelo es equivalente a suponer que existe una relación lineal entre las variables $X_1, \ldots, X_k$ y la función `log-odd de p`

$$\log \frac{p}{1-p} = \boldsymbol{\theta}'\mathbf{X}.$$

- Puesto que esto es equivalente a suponer que

$$p = Pr(Y = 1) = \frac{\exp(\boldsymbol{\theta}'\mathbf{X})}{1 + \exp(\boldsymbol{\theta}'\mathbf{X})} = g(\boldsymbol{\theta}'\mathbf{X})$$

Inferencia y predicción

Función de costo empírica

- Datos muestrales: $(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(n)}, y^{(n)})$.
- Función de costo:

$$J(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^{n} c(h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}), y^{(i)}) .$$

- Desarrollando la función $\boldsymbol{c}$ obtenemos

$$\begin{aligned} J(\boldsymbol{\theta}) &= -\frac{1}{n} \sum_{i=1}^{n} [y^{(i)} \log(h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)})) + (1 - y^{(i)}) \log(1 - h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}))] \\ &= -\frac{1}{n} \sum_{i=1}^{n} [y^{(i)} \log(g(\boldsymbol{\theta}' \mathbf{x}^{(i)})) + (1 - y^{(i)}) \log(1 - g(\boldsymbol{\theta}' \mathbf{x}^{(i)}))] \end{aligned}$$

Función de costo empírica en forma matricial

- Denotando
  - → $M = (x_j^{(i)})$ a la matriz de datos,
  - → $\mathbf{y} = (y^{(i)})$ al vector columna con los valores de Y y
  - → $h := g(M\boldsymbol{\theta})$ al vector columna con los ajustes en cada individuo, entonces

$$J(\boldsymbol{\theta}) := -\frac{1}{n}[\mathbf{y}' \log(h) + (\mathbf{1}_n - \mathbf{y})' \log(\mathbf{1}_n - h)]$$

donde $\mathbf{1}_n$ representa un vector columna de dimensión n.

Objetivo

- Ajustar el parámetro $\boldsymbol{\theta}$ para que J tome el menor valor posible.
  - → `Solución` → Algoritmos iterativos de búsqueda como, por ejemplo, el algoritmo del gradiente descendente.
    - ⇾ Práctica complementaria de regresión logística.
  - → Existen varias librerías de `R` que permiten obtener estimaciones de los parámetros del modelo logístico.
    - ⇾ Práctica de regresión logística.

Un ejemplo sencillo

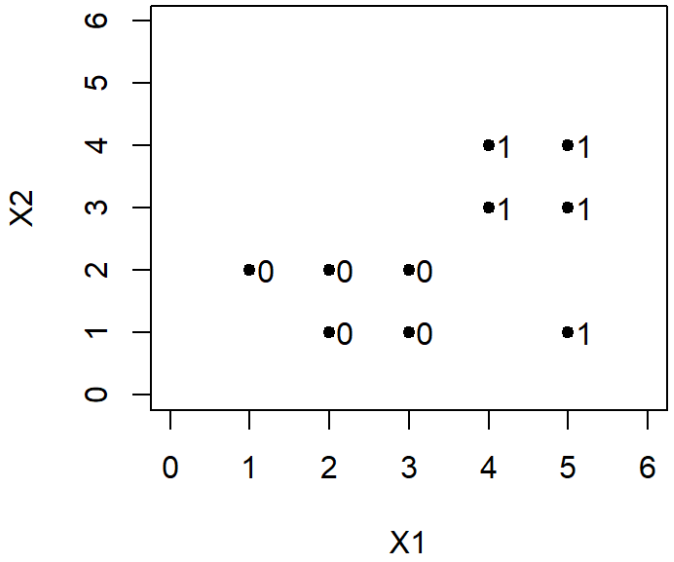
Datos muestrales

- Como en técnicas anteriores usaremos un ejemplo sencillo para comprobar cómo funciona nuestro modelo.
- Supongamos que tenemos dos variables predictoras X_1 y X_2 (k = 2) y los datos siguientes:

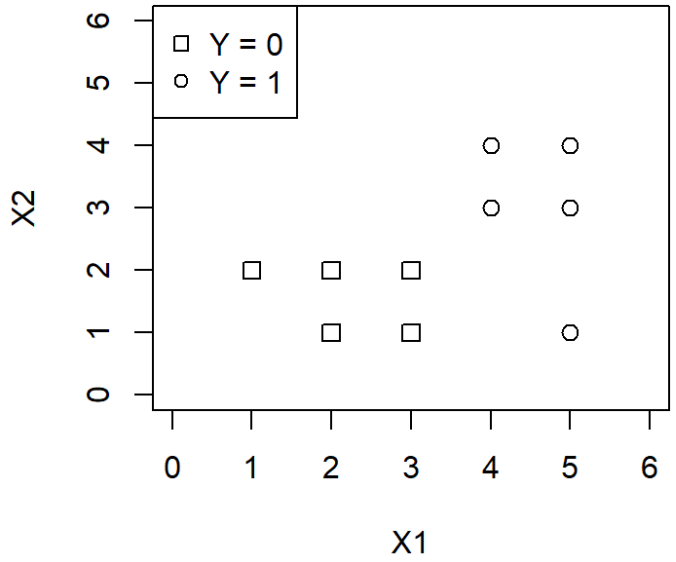
Individuo	X_1	X_2	Y
1	1	2	0
2	2	1	0
3	3	1	0
4	2	2	0
5	5	1	1
6	5	3	1
7	3	2	0
8	4	3	1
9	4	4	1
10	5	4	1

- Lo primero que tenemos que hacer (si es posible) es dibujar estos puntos añadiendo una etiqueta para distinguir los de cada grupo.

▶ Code



▶ Code



- En ambas gráficas podemos observar que los dos grupos se pueden separar muy bien con rectas.
- Por lo tanto nuestro modelo será

$$h_{\boldsymbol{\theta}}(\mathbf{x}) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)\,.$$

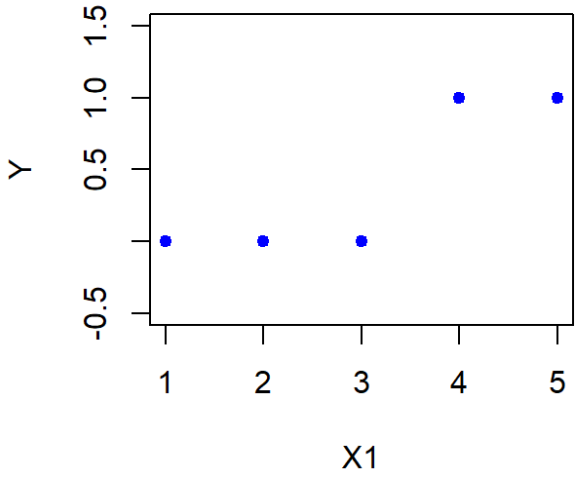
- Otra forma de analizar los grupos es calcular medidas descriptivas en cada uno de ellos.
  → Por ejemplo podemos calcular las medias en cada grupo:

▶ Code

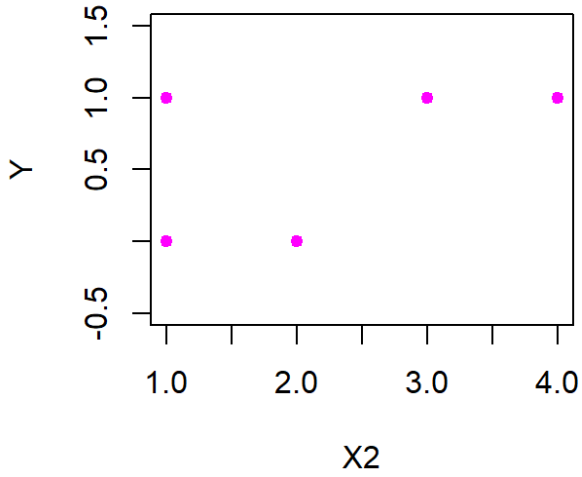
```
            Y = 0 Y = 1
Media de X1   2.2   4.6
Media de X2   1.6   3.0
```

- Estas diferencias también se pueden ver representado $\boldsymbol{x}^{(i)}$ frente a $\boldsymbol{Y}$.

▶ Code



▶ Code



- Podemos observar cómo la primera variable separa mejor a los grupos que la segunda (en los valores muestrales).
- De forma similar se pueden representar histogramas o diagramas caja-bigote para comparar las variables en cada grupo.

- Podemos calcular la función $\boldsymbol{J(\theta)}$ en R con los datos anteriores.

```r
X1 <- c(1, 2, 3, 2, 5, 5, 3, 4, 4, 5)
X2 <- c(2, 1, 1, 2, 1, 3, 2, 3, 4, 4)
Y <- c(0, 0, 0, 0, 1, 1, 0, 1, 1, 1)
n <- length(Y)
k <- 2
X0 = rep(1, n)
M <- matrix(c(X0, X1, X2), nrow = n, ncol = k + 1, byrow = FALSE)
g <- function(z){
  g = exp(z)/(1 + exp(z))
  return(g)
}
J <- function(theta){
  J = - sum(Y * log(g(M %*% theta)) + (1 - Y) * log(1 - g(M %*% theta)))/n
  return(J)
}
```

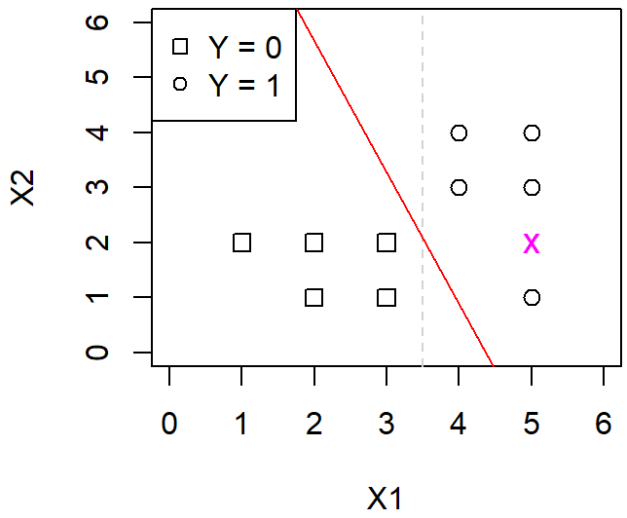
- Podemos aplicar un `método iterativo` para la obtención del óptimo.
  - → Por ejemplo, el `método del gradiente descendiente` (se detallará su aplicación en la práctica complementaria de regresión logística).

▶ Code

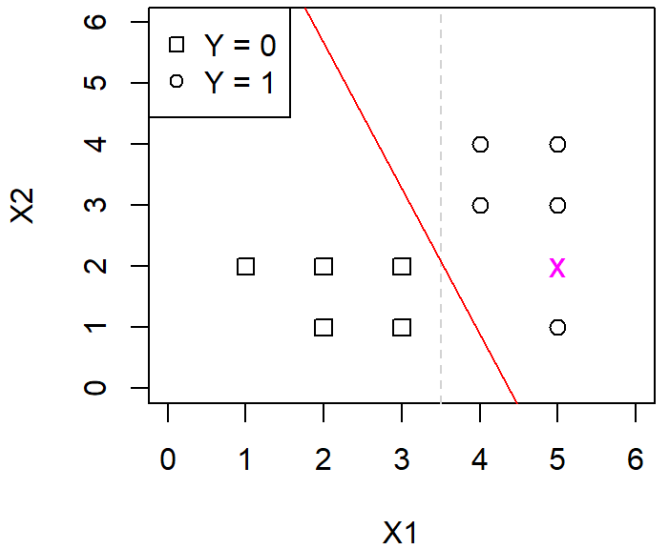
- Partiendo del punto inicial $\boldsymbol{\theta}^{(0)} = (-3.5, 1, 0)$ (recta vertical de separación $x_1 = 3.5$) y después de `1000 iteraciones` obtendremos $\hat{\theta}_0$ = -10.7505, $\hat{\theta}_1$ = 2.4594 y $\hat{\theta}_3$ = 1.0287, con valor de $J(\hat{\boldsymbol{\theta}})$ = 0.0597.
- En este caso,
  - → $I_{\hat{\boldsymbol{\theta}}}(x_1, x_2)$ = -10.7505 + 2.4594 x_1 + 1.0287 x_2.
- La recta que marca la frontera de esta solución será
  - → -10.7505 + 2.4594 x_1 + 1.0287 x_2 = 0.
- Esto es, x_2 = 10.4506 - 2.3908 x_1.

- Si queremos `predecir` el grupo para un nuevo individuo con valores $x_1 = 5$ y $x_2 = 2$,
  - → evaluamos la función $I_{\hat{\boldsymbol{\theta}}}(x_1, x_2)$ = -10.7505 + 2.4594 x_1 + 1.0287 x_2,
  - → $I_{\hat{\boldsymbol{\theta}}}(5, 2)$ = 3.6037 > 0, por lo que el individuo se clasifica en el grupo $y = 1$.
- Representamos los valores muestrales incluyendo la recta que marca la frontera y el punto $(5, 2)$.

▶ Code



▶ Code



- Para medir cómo de fiable es esta clasificación podemos
  - → observar cómo se distribuyen los puntos en esta gráfica (cuando sea posible) o
  - → calcular las `probabilidades a posteriori`.
- $Pr(Y = 1|X_1 = 5, X_2 = 2) \approx g(I_{\hat{\boldsymbol{\theta}}}(5, 2))$ = 0.9735.
- $Pr(Y = 0|X_1 = 5, X_2 = 2) \approx 1 - g(I_{\hat{\boldsymbol{\theta}}}(5, 2))$ = 0.0265.

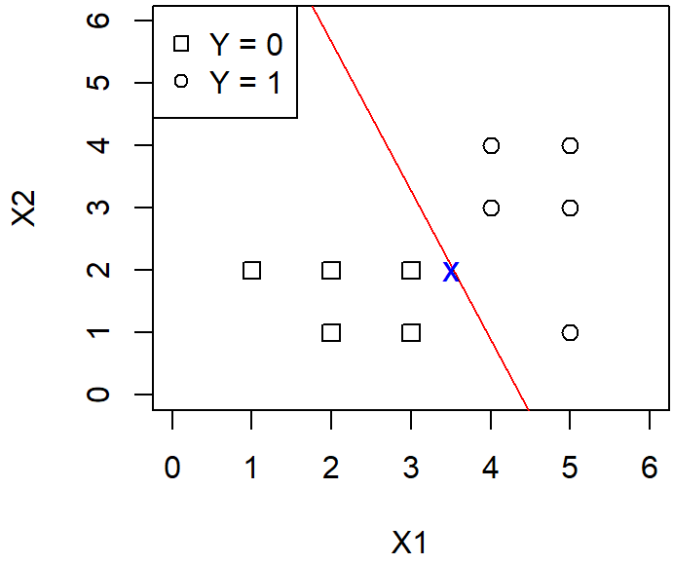
- Observando la gráfica con los valores muestrales parece razonable que el nuevo individuo con valores $\mathbf{x} = (5, 2)$ sea clasificado en el grupo $y = 1$.

- Ahora queremos `predecir` el grupo para otro nuevo individuo con valores $x_1 = 3.5$ y $x_2 = 2$. En este caso,

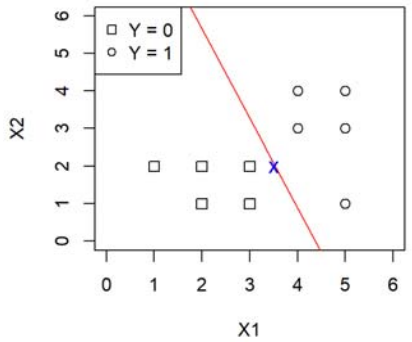
  → $I_{\hat{\boldsymbol{\theta}}}(x_1, x_2)$ = -0.0853 < 0, por lo que el individuo se clasifica en el grupo $y = 0$.

- Representamos gráficamente.

▶ Code



▶ Code



- Observamos dónde se encuentra el nuevo individuo en la gráfica.
- Calculamos las `probabilidades a posteriori`:
  - → $Pr(Y = 1|X_1 = 3.5, X_2 = 2) \approx g(I_{\hat{\boldsymbol{\theta}}}(3.5, 2))$ = 0.4787.
  - → $Pr(Y = 0|X_1 = 3.5, X_2 = 2) \approx 1 - g(I_{\hat{\boldsymbol{\theta}}}(3.5, 2))$ = 0.5213.

- Recordemos que en realidad no estamos seguros de que esos valores sean realmente esas probabilidades.
- De esta forma intuimos que esta clasificación no es muy fiable ya que ese punto está `muy cerca de la frontera`.

Regresión logística multinomial

El contexto

- Generalización del modelo de regresión logística binaria.
- La variable dependiente tiene `más de dos categorías`, sin/con un orden implícito.
  - → Primer caso: considera variables de `respuesta nominal`,
    - ⇾ Por ejemplo, el país de procedencia, el color de un automóvil, etc.
  - → Segundo caso: trata variables de `respuesta ordinal`,
    - ⇾ Por ejemplo, el nivel educativo, la fase de una enfermedad, etc.

Objetivo

- Estimar la probabilidad de que un individuo presente cada una de estas categorías en función de los valores que se observen de las variables explicativas.

Modelo teórico

Formulación

- La variable respuesta Y puede presentar g `categorías`.

  → Y toma los valores $1, 2, \ldots, g$, que indican la pertenencia a cada grupo definido por cada categoría, con probabilidades $p_1, p_2, \cdots, p_g$, respectivamente, tales que

$$\sum_{j=1}^{g} p_j = 1\,.$$

- Consideramos como referencia una de las categorías, por ejemplo, la última, g.

- Establecemos un modelo `logit` para cada categoría con respecto a esta:

$$\log \frac{p_j}{p_g} = \log \frac{Pr[Y=j]}{Pr[Y=g]} = \boldsymbol{\theta}_j' \mathbf{X}, \quad j = 1, \ldots, g-1\,,$$

  donde $\boldsymbol{\theta}_j = (\theta_{j0}, \ldots, \theta_{jk})' \in \mathbb{R}^{k+1}$ contiene los parámetros del modelo para cada categoría j y $\mathbf{X} = (X_0, \ldots, X_k)'$ las variables que podemos medir en nuestros individuos para predecir Y.
- Al cociente $\dfrac{p_j}{p_g}$ se le denomina `odds` de la categoría j respecto de la categoría g.
- Se ha considerado un término constante en el modelo incluyendo la variable artificial X_0 que siempre vale 1.
- Cada uno de los coeficientes se interpreta como el efecto de cada variable explicativa sobre el logaritmo de los `odds` de la categoría j respecto de la categoría de referencia g.
- Cuando $g = 2$, el modelo se reduce a una única ecuación equivalente a la propuesta en la regresión logística.

Observaciones

- Si comparamos las probabilidades para dos categorías diferentes, i y j, utilizando el modelo anterior obtenemos que:

$$\log \frac{p_i}{p_j} = \log \frac{\frac{p_i}{p_g}}{\frac{p_j}{p_g}} = \log \frac{p_i}{p_g} - \log \frac{p_j}{p_g} = \boldsymbol{\theta}_i' \mathbf{X} - \boldsymbol{\theta}_j' \mathbf{X}$$
$$= (\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)' \mathbf{X} = (\theta_{i0} - \theta_{j0}) + (\theta_{i1} - \theta_{j1}) X_1 + \cdots (\theta_{ik} - \theta_{jk}) X_k \, .$$

→ De esta forma, se obtiene una ecuación `logit` de la categoría i con respecto a la categoría j, donde $\theta_0 = \theta_{i0} - \theta_{j0}$, $\theta_1 = \theta_{i1} - \theta_{j1}$,..., $\theta_k = \theta_{ik} - \theta_{jk}$.

Un ejemplo ficticio

- Supongamos que deseamos estudiar cómo influye el sexo del neonato en la aparición de determinados problemas durante el parto.
- Se contemplan únicamente tres opciones posibles para los partos (Y):
  - $\rightarrow$ $Y = 1$: parto con el problema A
  - $\rightarrow$ $Y = 2$: parto con el problema B
  - $\rightarrow$ $Y = 3$: parto sin problemas
- La tercera opción se toma como la opción de referencia.
- Se introduce una variable binaria X para representar el sexo del neonato:
  - $\rightarrow$ $X = 0$: si es niño
  - $\rightarrow$ $X = 1$: si es niña

- Planteamos un modelo de regresión logística para predecir la variable categórica Y en función de X:

$$\log \frac{p_j}{p_3} = \log \frac{Pr[Y = j]}{Pr[Y = 3]} = \theta_{j0} + \theta_{j1} x \, , \quad j = 1, 2.$$

Interpretación de los parámetros

- Considerando la primera de las ecuaciones:

$$\log \frac{p_1}{p_3} = \log \frac{Pr[Y=1]}{Pr[Y=3]} = \theta_{10} + \theta_{11} x \,,$$

  → Si $X = 0$, $\dfrac{p_1}{p_3} = \exp(\theta_{10})$.

  → Si $X = 1$, $\dfrac{p_1}{p_3} = \exp(\theta_{10} + \theta_{11})$.

- Por lo tanto, respecto a la probabilidad de un parto normal,

  → la probabilidad de la presencia del problema A se multiplica por $\exp(\theta_{10})$ en el caso de niños y por $\exp(\theta_{10} + \theta_{11})$ en el caso de niñas.

- Si, por ejemplo, resultara $\theta_{11} = 0$,

  → el sexo no influiría sobre la probabilidad de que aparezca el problema A.

- Razonando de forma análoga, si resultara $\theta_{21} > 0$ se concluiría que la aparición del problema B es más probable en niñas que en niños.

Recordemos...

Estimador de máxima verosimilitud

- Sea X una variable aleatoria, con función de densidad o función puntual de probabilidad $x \mapsto f_X(x;\boldsymbol{\theta})$, donde $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^k$.
- Consideramos una m.a.s $(X_1, \ldots, X_n)$.
- Para un valor concreto de $(X_1, \ldots, X_n)$, que denotamos por $(x_1, \ldots, x_n)$, la `función de verosimilitud` L_n es una función de $\boldsymbol{\theta}$, $L_n : \Theta \subset \mathbb{R}^k \to \mathbb{R}^+$, definida como

$$L_n(\boldsymbol{\theta}) = f_{X_1,\ldots,X_n}(x_1, \ldots, x_n; \boldsymbol{\theta}) = \prod_{i=1}^{n} f_{\mathbf{X}}(x_i; \boldsymbol{\theta}) \,.$$

- El `estimador de máxima verosimilitud` $\hat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$ es cualquier valor de $\boldsymbol{\theta}$ admisible que maximiza la función $L_n(\boldsymbol{\theta})$,

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Theta} L_n(\boldsymbol{\theta}) \,.$$

Criterio de máxima de verosimilitud para nuestro modelo

Criterio

- Para estimar los parámetros del modelo utilizaremos el criterio de máxima verosimilitud:
  - → Calculamos la función de verosimilitud
  - → Maximizamos esta función para obtener los estimadores de máxima verosimilitud (MLE, `maximum likelihood estimator`).
- Redefiniremos la variable Y en g variables indicadoras $(Y_1, \ldots, Y_g)$:
  - → Y_j toma el valor 1 si la respuesta pertenece al grupo j y 0 en otro caso.
  - → Tendremos que $\sum_{j=1}^{g} Y_j = 1$.

Función de verosimilitud

- Supongamos que disponemos de n observaciones independientes de la variable Y y de las variables explicativas.Para cada individuo i tendremos:

  → Las observaciones $(y_1^{(i)}, \ldots, y_g^{(i)})$, donde

  $$\sum_{j=1}^{g} y_j^{(i)} = 1\,.$$

  → Los valores de las variables explicativas observados $\mathbf{x}^{(i)} = (x_0^{(i)}, \ldots, x_k^{(i)})'$.

- Entonces, la `función de verosimilitud` adopta la expresión

$$L(\boldsymbol{\theta}_1, \ldots \boldsymbol{\theta}_{g-1}) \propto \prod_{i=1}^{n} \prod_{j=1}^{g} p_j(\mathbf{x}^{(i)})^{y_j^{(i)}}\,.$$

donde el símbolo $\propto$ indica `proporcional a`.

Función de log-verosimilitud

- En lugar de maximizar directamente la función de verosimilitud consideraremos su logaritmo neperiano:
  - → función más manejable que simplifica los cálculos,
  - → permite utilizar métodos numéricos de optimización más eficientes y estables al transformar productos en sumas.
- La log-verosimilitud adopta la expresión

$$\log L(\boldsymbol{\theta}_1, \dots \boldsymbol{\theta}_{g-1}) \propto \log \prod_{i=1}^{n} \prod_{j=1}^{g} p_j(\mathbf{x}^{(i)})^{y_j^{(i)}} = \sum_{i=1}^{n} \sum_{j=1}^{g} y_j^{(i)} \log p_j(\mathbf{x}^{(i)}) \, .$$

- En términos de una función costo, el criterio de máxima verosimilitud equivale a minimizar la función

$$J(\boldsymbol{\theta}_1, \dots \boldsymbol{\theta}_{g-1}) = -\log L(\boldsymbol{\theta}_1, \dots \boldsymbol{\theta}_{g-1}) \, .$$

- En ocasiones en lugar de la función de log-verosimilitud se utiliza la función auxiliar $\Lambda = -2 \log L$, denominada la `deviance` del modelo.

- Puesto que $p_g = 1 - (p_1 + \ldots p_{g-1})$, la contribución del individuo i en la función de la log-verosimilitud sería

$$\begin{aligned}\sum_{j=1}^{g} y_j^{(i)} \log p_j(\mathbf{x}^{(i)}) &= \sum_{j=1}^{g-1} y_j^{(i)} \log p_j(\mathbf{x}^{(i)}) + \\ &\left(1 - \sum_{j=1}^{g-1} y_j^{(i)}\right) \log \left(1 - \sum_{j=1}^{g-1} p_j(\mathbf{x}^{(i)})\right) \\ &= \sum_{j=1}^{g-1} y_j^{(i)} \log \frac{p_j(\mathbf{x}^{(i)})}{1 - \sum_{h=1}^{g-1} p_h(\mathbf{x}^{(i)})} + \log \left(1 - \sum_{j=1}^{g-1} p_j(\mathbf{x}^{(i)})\right).\end{aligned}$$

- Según el modelo logístico multinomial,

$$\log \frac{p_j(\mathbf{x}^{(i)})}{p_g(\mathbf{x}^{(i)})} = \boldsymbol{\theta}_j' \mathbf{x}^{(i)} \, .$$

- Por otra parte, $p_g(\mathbf{x}^{(i)})$ puede escribirse como

$$p_g(\mathbf{x}^{(i)}) = 1 - \sum_{j=1}^{g-1} p_j(\mathbf{x}^{(i)}) = 1 - p_g(\mathbf{x}^{(i)}) \sum_{j=1}^{g-1} \exp(\boldsymbol{\theta}_j' \mathbf{x}^{(i)}) \, ,$$

- Y despejando ahora $p_g(\mathbf{x}^{(i)})$ en la expresión anterior se obtiene que

$$p_g(\mathbf{x}^{(i)}) = \frac{1}{1 + \sum_{h=1}^{g-1} \exp\left(\boldsymbol{\theta}_j' \mathbf{x}^{(i)}\right)} \, .$$

- Y por tanto,

$$p_j(\mathbf{x}^{(i)}) = \frac{\exp(\boldsymbol{\theta}_j'\mathbf{x}^{(i)})}{1 + \sum_{h=1}^{g-1} \exp(\boldsymbol{\theta}_h'\mathbf{x}^{(i)})} \, .$$

- Sustituyendo ahora en la función de log-verosimilitud se tendrá que

$$\begin{aligned}
\log L(\boldsymbol{\theta}_1, \ldots \boldsymbol{\theta}_{g-1}) &\propto \sum_{i=1}^{n} \sum_{j=1}^{g} y_j^{(i)} \log p_j(\mathbf{x}^{(i)}) \\
&= \sum_{i=1}^{n} \left[\sum_{j=1}^{g-1} y_j^{(i)} (\boldsymbol{\theta}_j'\mathbf{x}^{(i)}) - \log \left(1 + \sum_{j=1}^{g-1} \exp(\boldsymbol{\theta}_j'\mathbf{x}^{(i)}) \right) \right] \\
&= \sum_{i=1}^{n} \sum_{j=1}^{g-1} y_j^{(i)} (\boldsymbol{\theta}_j'\mathbf{x}^{(i)}) - \sum_{i=1}^{n} \log \left(1 + \sum_{j=1}^{g-1} \exp(\boldsymbol{\theta}_j'\mathbf{x}^{(i)}) \right)
\end{aligned}$$

¿Cómo obtenemos en la práctica las estimaciones de $\boldsymbol{\theta}_1 \ldots, \boldsymbol{\theta}_{g-1}$?

- Para obtener valores de los parámetros $\boldsymbol{\theta}_1 \ldots, \boldsymbol{\theta}_{g-1}$ que maximicen la log-verosimilitud (o equivalentemente, minimicen la función costo $\boldsymbol{J}$), podremos
  - → Aplicar algoritmos iterativos de búsqueda como el algoritmo del gradiente descendente.
  - → Hacer uso de funciones implementadas en librerías de `R` que permiten obtener estimaciones de los parámetros del modelo logístico multinomial, como la función `multinom()` de la librería `nnet`.
    - ⇾ Práctica de regresión logística multinomial.

Un caso sencillo

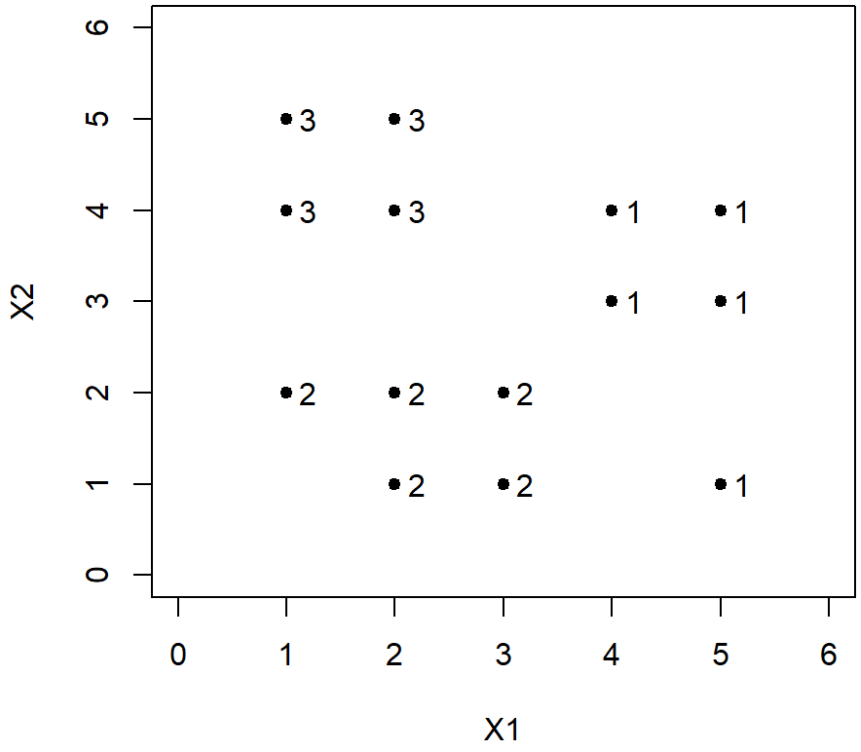
Cálculo de la verosimilitud

- Supongamos que tenemos una variable respuesta (Y) con tres categorías posibles y dos variables explicativas X_1 y X_2, cuyas observaciones están recogidas en la tabla adjunta.

Individuo	X_1	X_2	Y
1	1	2	2
2	2	1	2
3	1	5	3
4	2	4	3
5	3	1	2
6	2	2	2
7	2	5	3
8	5	1	1
9	5	3	1
10	3	2	2
11	4	3	1
12	4	4	1
13	5	4	1
14	1	4	3

- Para la implementación en R de estos cálculos introduciremos los datos en vectores y representaremos los datos gráficamente.

▶ Code



- Incluimos los valores de la variable artificial $X_0 = 1$.
- Asociada a la variable Y definimos tres variables indicadoras, Y_j, $j = 1, 2, 3$, de manera que $Y_j = 1$ si $Y = j$ y 0 en otro caso.

▶ Code

	X1	X2	Y	X0	Y1	Y2	Y3
1	1	2	2	1	0	1	0
2	2	1	2	1	0	1	0
3	1	5	3	1	0	0	1
4	2	4	3	1	0	0	1
5	3	1	2	1	0	1	0
6	2	2	2	1	0	1	0
7	2	5	3	1	0	0	1
8	5	1	1	1	1	0	0
9	5	3	1	1	1	0	0
10	3	2	2	1	0	1	0
11	4	3	1	1	1	0	0
12	4	4	1	1	1	0	0
13	5	4	1	1	1	0	0
14	1	4	3	1	0	0	1

- Empezamos implementando la función `J` utilizando código en `R`.

```r
J <- function(theta){
  C = exp(t(theta)%*%t(X))
  D = colSums(C)
  E = matrix(rep(1 + D, g - 1), nrow = g - 1, ncol = n, byrow = TRUE)
  P = C/E
  Pg = 1/(1+D)
  PT =  rbind(P, Pg)
  J = -sum(YY*log(t(PT)))
  return(J)
}
```

- Introducimos los valores muestrales de las variables.

▶ Code

- Evaluamos en

$$\boldsymbol{\theta} = \begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 2 & 3 \end{pmatrix}$$

▶ Code

```
[1] 111.0598
```

- Evaluamos en

$$\boldsymbol{\theta} = \begin{pmatrix} -2.3 & 1.5 \\ 0.5 & 2 \\ 2.1 & 3.5 \end{pmatrix}$$

▶ Code

```
[1] 155.5009
```

Estimación de los parámetros

- Utilizaremos la función `multinom()` de la librería `nnet` de R.

  → Las opciones de esta función se verán con más detalle en la práctica de regresión logística multinomial.

▶ Code

```
# weights:  12 (6
variable)
initial  value 15.380572
iter  10 value 0.194988
iter  20 value 0.010826
iter  30 value 0.006357
iter  40 value 0.005358
iter  50 value 0.004594
iter  60 value 0.003156
iter  70 value 0.002732
iter  80 value 0.002396
iter  90 value 0.002072
iter 100 value 0.001924
final  value 0.001924
stopped after 100
iterations
```

▶ Code

```
  (Intercept)        X1          X2
1   -19.79766 12.677892  -5.560149
2    27.10762  2.521385 -10.282168
```

Modelo estimado

- Ecuaciones del `modelo estimado`:

$$\log \frac{p_1}{p_3} = -19.798 + 12.678x_1 - 5.56x_2$$

$$\log \frac{p_2}{p_3} = 27.108 + 2.521x_1 - 10.282x_2$$

- Recordemos que los coeficientes del modelo miden la variación del logaritmo de los `odds` por unidad de cambio en el correspondiente predictor.
- Tomando exponenciales sobre los coeficientes, medimos las variaciones producidas sobre los `odds` directamente.
- El algoritmo ha parado después de 100 iteraciones.
- Valor de la $-\log L$: 0.0019242
- Valor de la `deviance` del modelo: 0.0038484

Modelo logístico multinomial con categorías ordinales

Modelo teórico

- Este tipo de modelo se utiliza cuando las categorías de la variable dependiente representan un orden lógico o jerarquía.
- Expresión general del `modelo`:

$$\log \frac{Pr[Y \leq j]}{1 - Pr[Y \leq j]} = \boldsymbol{\theta}_j' \mathbf{X}, \quad j = 1, \ldots, g-1\,,$$

donde $\boldsymbol{\theta}_j = (\theta_{j0}, \ldots, \theta_{jk})' \in \mathbb{R}^{k+1}$ contiene los parámetros del modelo para cada categoría j y $\mathbf{X} = (X_0, \ldots, X_k)'$ las variables que podemos medir en nuestros individuos para predecir Y.

- `Interpretación de los coeficientes`: entender cómo un cambio en una variable predictora afecta a la razón de probabilidades de que la variable dependiente sea menor o igual a una categoría específica en comparación con las categorías superiores.