Regresión Logística

Introducción

- En un problema de regresión, no siempre la respuesta (condicionada) es normal
- A veces, la respuesta ni siquiera es cuantitativa

Supongamos que nos interesa modelizar las siguientes variables:

- Si una persona vota o no por un determinado candidato
- Si un estudiante aprueba o no un examen
- Si mañana lloverá o no

Es decir, nos interesa modelizar una variable Y, una **variable respuesta categórica binaria**:

$$Y = \begin{cases} 1 \text{ si mañana llueve} \\ 0 \text{ en caso contrario} \end{cases}$$

en función de ciertas variables explicativas potenciales...

En otros contextos se habla de un **problema de clasificación**

Según los valores que puede tomar Y, ¿qué modelo de probabilidad podemos asumir?

$$Y_i \mid \pi_i \sim \text{Bern}(\pi_i)$$

$$\mathbb{E}(Y_i) = \pi_i$$

En la regresión normal que conocíamos, teníamos

$$Y_i \mid \mu_i \sim \mathcal{N}(\mu_i, \sigma^2)$$

$$\mathbb{E}(Y_i) = \mu_i$$

Por analogía, ¿podemos hacer $\pi_i = \beta_0 + \beta_1 x_{1_i} + \beta_2 x_{2_i} + ...$?

¿Qué problemas identificamos?

Tendremos que hacer

$$g(\pi_i) = \beta_0 + \beta_1 x_{1_i} + \beta_2 x_{2_i} + ...$$

$g(\cdot)$ se conoce como función de enlace (*link function*). ¿Cuál es una g apropiada en este caso?

Si π_i es la probabilidad del evento de interés, $\frac{\pi_i}{1-\pi_i}$ es la chance (*odds*) del evento de interés.

Mientras que $\pi_i \in [0, 1]$, $\frac{\pi_i}{1-\pi_i} \in [0, +\infty)$

Establecemos un modelo lineal para el *log-odds* del evento de interés

$$\log(\text{odds}_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{1_i} + \beta_2 x_{2_i} + ...$$

La función $g(x) = \log\left(\frac{x}{1-x}\right)$ se conoce como función **logit**. Es una función no lineal.

Analicemos lo que vimos hasta ahora:

- ¿Cuál es el dominio de la función logit?
- ¿Para qué necesitamos la función $g(\cdot)$?
- ¿Cuál es la relación entre el predictor lineal y la variable respuesta?
- ¿Cuál es la distribución de la variable respuesta?

Consideremos el caso con una sola variable explicativa

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{1_i}$$

Se cumple:

$$\frac{\pi_i}{1-\pi_i} = e^{\beta_0+\beta_1 x_{1_i}} \qquad \pi_i = \frac{e^{\beta_0+\beta_1 x_{1_i}}}{1+e^{\beta_0+\beta_1 x_{1_i}}}$$

¡La esperanza de la variable respuesta se relaciona de manera no lineal con las variables explicativas!

Ejemplo

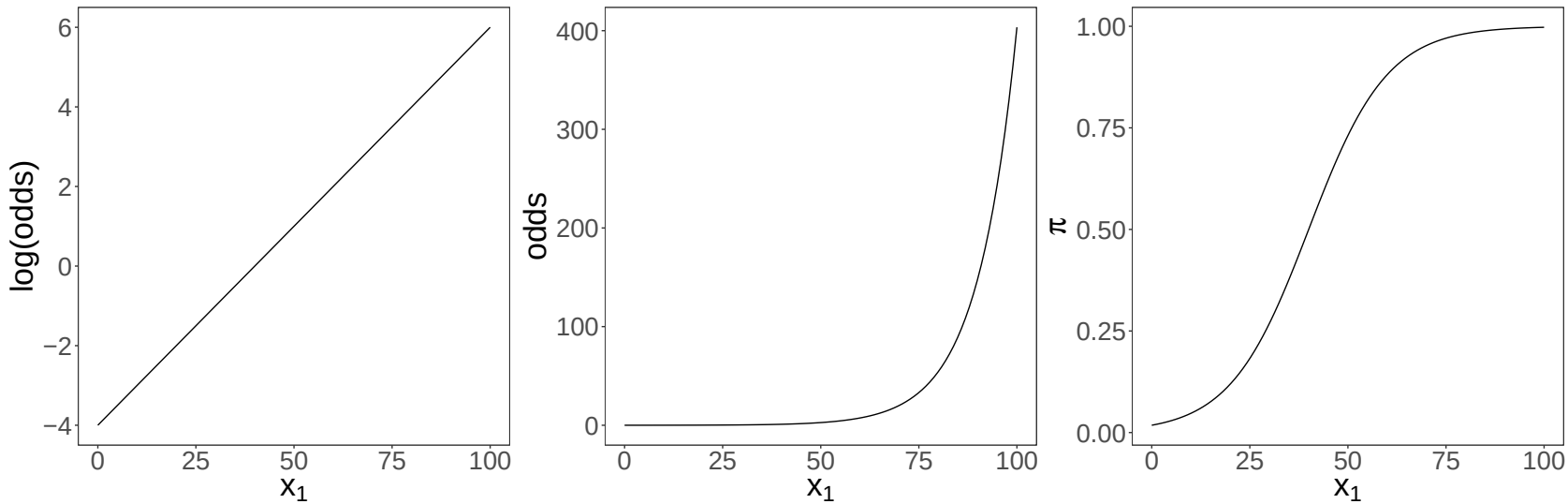
Consideremos la siguiente relación

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = -4 + 0.1\ x_{1_i}$$

e imaginemos que π_i es la probabilidad de que llueva el día i y x_{1_i} la humedad a las 9 de la mañana del día anterior al día i.

```r
x1 <- seq(0, 100, length.out = 100)
beta0 <- -4
beta1 <- 0.1

data <- tibble(x1 = x1,
               log_odds = beta0 + beta1*x1,
               odds = exp(log_odds),
               pi = odds/(1+odds))
```


log(odds)
6
4
2
0
−2
−4
0
25
50
75
100
x1
odds
400
300
200
100
0
0
25
50
75
100
x1
π
1.00
0.75
0.50
0.25
0.00
0
25
50
75
100
x1

Interpretación de los coeficientes

β_0

β_0 es la log-chance (*log-odds*) del evento de interés cuando todas las variables explicativas valen 0. e^{β_0} es la chance. En términos del problema: e^{β_0} *es la chance de que llueva mañana si la humedad de hoy a las 9 de la mañana es 0*.

β_1

β_1 no es el incremento en la probabilidad del evento de interés cuando x_1 aumenta en una unidad...

- odds_x es la chance del evento de interés cuando $x_1 = x$
- $\text{odds}_{x+\Delta x}$ es la chance del evento de interés cuando $x_1 = x + \Delta x$

$$\log(\text{odds}_x) = \log\left(\frac{\pi_x}{1-\pi_x}\right) = \beta_0 + \beta_1 x$$

$$\log(\text{odds}_{x+\Delta x}) = \log\left(\frac{\pi_{x+\Delta x}}{1-\pi_{x+\Delta x}}\right) = \beta_0 + \beta_1 (x + \Delta x)$$

Entonces

$$\log(\text{odds}_{x+\Delta x}) - \log(\text{odds}_x) = \beta_1 \Delta x$$

$$e^{\beta_1 \Delta x} = \frac{\text{odds}_{x+\Delta x}}{\text{odds}_x}$$

La chance del evento de interés aumenta $e^{\beta_1 \Delta x}$ veces cuando x_1 aumenta en Δx (y el resto de las variables se mantienen constantes). En términos del problema: *La chance de que llueva mañana aumenta e^{β_1} veces si la humedad aumenta en 1*.

Más en términos del problema:

$$e^{\beta_1} = \frac{\text{odds}_{x+1}}{\text{odds}_x} \Rightarrow \text{odds}_{x+1} = e^{\beta_1}\text{odds}_x$$

$$e^{\beta_1} = 1.11$$

- La chance de que llueva mañana aumenta 1.11 veces cuando la humedad a las 9 de la mañana de hoy aumenta en una unidad
- La chance de que llueva mañana aumenta en un 11% cuando la humedad a las 9 de la mañana de hoy aumenta en una unidad

Analicemos juntos el siguiente caso:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = 1.1 - 0.2\ \text{despierto}_i$$

e imaginemos que π_i es la probabilidad de que un estudiante i apruebe el parcial de Análisis de Datos de Duración i y despierto_i la cantidad de horas que el estudiante i estuvo despierto la noche anterior al parcial.

¿Y Bayes?

La especificación del modelo se completa con la elección de distribuciones *a priori* para $\beta_0,\ \beta_1,\ .....$

Cada cantidad que dependa de los $\beta_0,\ \beta_1,\ ...$ tendrá una distribución de probabilidad.

Las predicciones también son probabilísticas.

¿Cómo se estiman $\beta_0,\ \beta_1,\ \ldots$? Como siempre. Aplicando la Regla de Bayes. Solo que la verosimilitud ahora es Bernoulli.