Aprendizaje Supervisado: Regresión Logística.

Guillermo Santamaría-Bonfil

Agenda

- Regresión Logística.
  - Odds/Log odds
  - Porque no regresión lineal
  - Derivación de la función

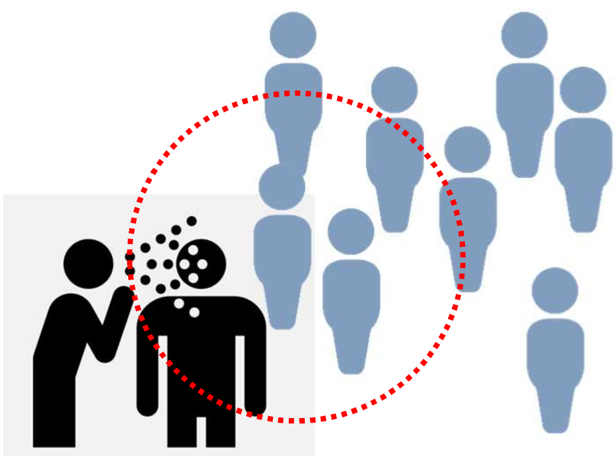
Antecedentes: Odds Ratio

Una de las principales funciones en la estadística para cuantificar:

Cuan fuerte es
la <u>*presencia*</u> o <u>*ausencia*</u> de la propiedad **<u>A</u>** asociada con
la <u>*presencia*</u> o <u>*ausencia*</u> de la propiedad **<u>B</u>**
en una población.

<u>Razón de Momios</u>
<u>Razón de oportunidades</u>
<u>Razón de probabilidades</u>

Odds ratio



E.g.
La razón de probabilidad (odds ratio)
describe cuantitativamente la asociación entre la presencia/ausencia de A (**<u>gripa</u>**) y la presencia/ausencia de B (**<u>exposición al virus</u>**)
para los individuos de la población.

Cálculo de Odds Ratio (OR)

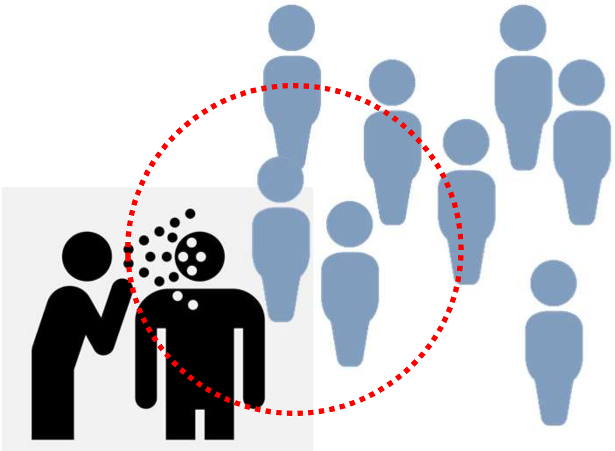
1. Para un individuo que tiene B (expuesto a la gripa) calcular los ***odds*** de que también tenga A (gripa).

2. Para un individuo que no tiene B (no estuvo expuesto) calcular los ***odds*** de que también tenga A (gripa).

3. Dividir los odds de 1 entre los odds de 2, para obtener OR.

Si $OR = 1$ → La exposición a B no afecta A
Si $OR < 1$ → A se considera asociada con B (baja la posibilidad)
Si $OR > 1$ → A se considera asociada con B (sube la posibilidad)

Nota: ESTO **NO IMPLICA CAUSALIDAD: B CAUSA A**



	ENFERMO (GRIPA)	SANO (NO GRIPA)
EXPUESTO	G_E=4	S_E=1
NO-EXPUESTO	G_{NE}=1	S_{NE}=2

1. Tener gripa dado que estuvo expuesto: $\frac{G_E}{S_E} = 4$

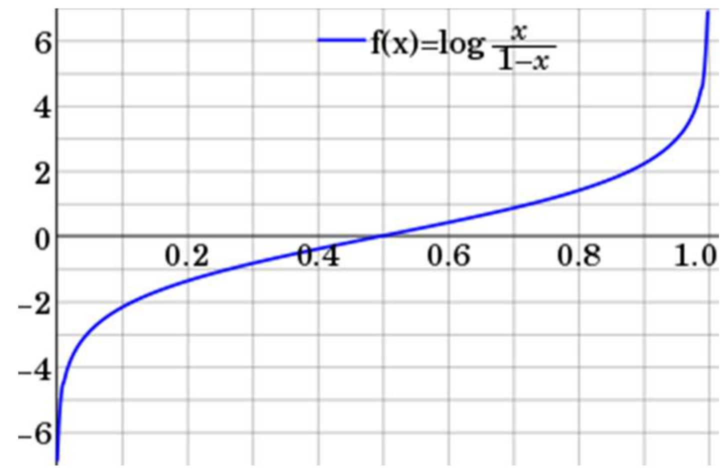
2. Tener gripa dado que no estuvo expuesto: $\frac{G_{NE}}{S_{NE}} = \frac{1}{2}$

$$3. OR = \frac{\frac{G_E}{S_E}}{\frac{G_{NE}}{S_{NE}}} = \frac{G_E S_{NE}}{G_{NE} S_E} = \frac{\frac{G_E}{G_{NE}}}{\frac{S_E}{S_{NE}}} = \frac{4}{\frac{1}{2}} = 8$$

¿Qué implica este resultado?

Antecedentes: Log Odds

- El logaritmo de la razón de momios…
- Alternativa para expresar probabilidades simplificando el proceso de actualización con nueva evidencia
- Logit
- Función inversa de la función sigmoidal logística.



$$logit(P(A)) = \log\left(\frac{P(A)}{P(\neg A)}\right)$$

$$logit(P(A)) \in (-\infty, \infty)$$

	ENFERMO (GRIPA)	SANO (NO GRIPA)
EXPUESTO	G_E=4	S_E=1
NO-EXPUESTO	G_{NE}=1	S_{NE}=2

$$logit(P(G)) = \log\left(\frac{P(G)}{P(S)}\right)$$

Antecedentes:
Log Odds

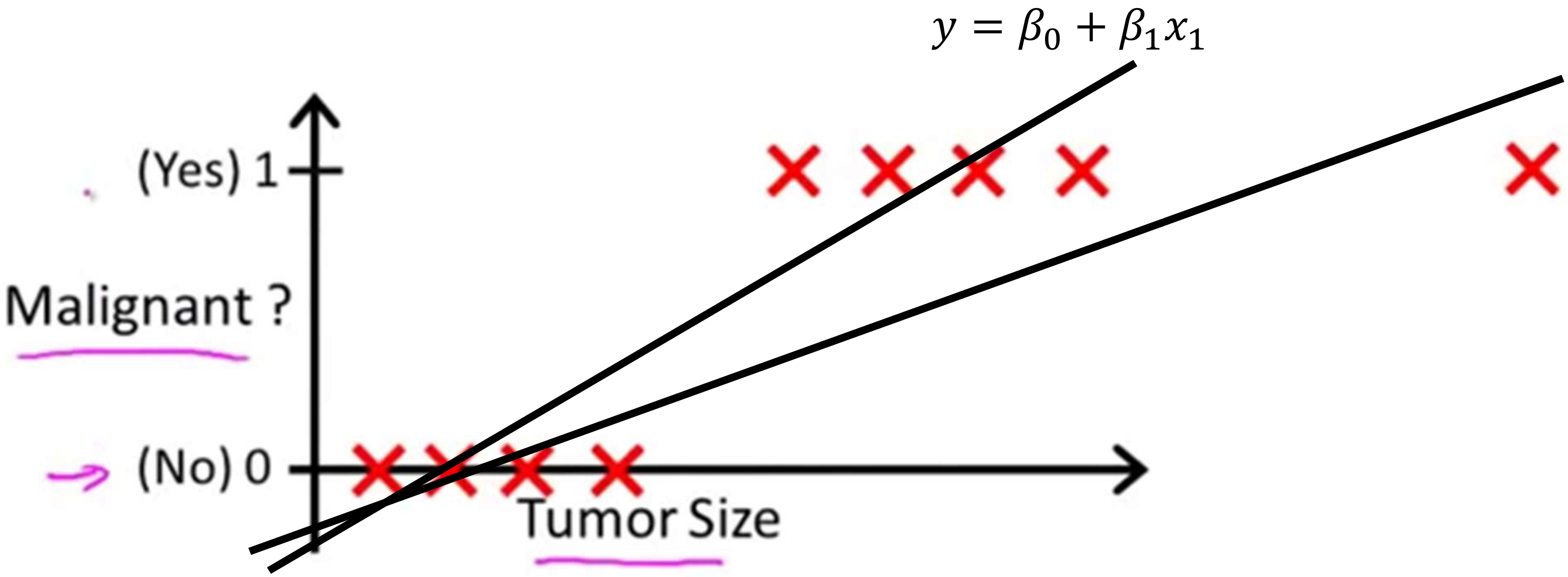
La función logística de cualquier número α esta dado por el logit inverso:

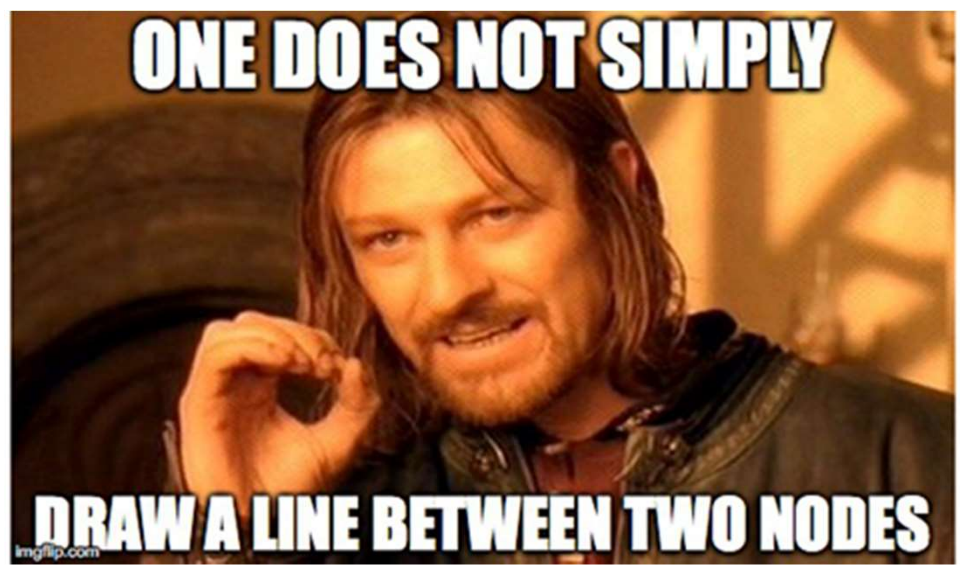
$$logit(\alpha) = \log\left(\frac{\alpha}{\neg\alpha}\right) = \log(\alpha) - \log(\neg\alpha) \rightarrow$$

…

…

$$logit^{-1}(\alpha) = logistic(\alpha) = \frac{1}{1 + \exp(-\alpha)} = \frac{\exp(\alpha)}{\exp(\alpha) + 1}$$





Pregunta:

a) ¿Qué pasa con el modelo y su error?

b) ¿Cómo son los valores devueltos por el modelo?

Regresión Logística

- Familia de Métodos
  Modelos Lineales Generalizados
  (GLM)
- 1972 Nelder and Wedderburn
  - Extensión de la MLR a
    problemas ≠ Regresión
  - La Regresión Logística es un caso
    especial de MLR

GLM

$$g(E(y)) = a + \beta x_1 + \gamma x_2$$

$g(.)$ Función liga

$E(.)$ Función de Valor Esperado

El valor esperado de una variable aleatoria es su MEDIA, a continuación se muestra el cálculo a) discreta y b) Contina:

a) $E[h(x)] = \sum_x h(x) f(x) \leftarrow h(x)$ en una variable aleatoria

b) $E(x) = \int_{-\infty}^{\infty} x f(x) dx$

Ejemplo:

1000 clientes

Necesitamos predecir la probabilidad.

clientes compran / no-compran

$$y \in \left\{ \begin{matrix} 0 & - \text{no-compra} \\ 1 & - \text{compra} \end{matrix} \right\}$$

a) Empezamos por plantear una regresión lineal

con $y \to g(y)$

$$g(y) = \beta_0 + \beta_1 (\text{age}) \quad (a)$$

En Reg. Logística se busca:

- Estimar $P(\cdot)$ del resultado de y

Para ello $g(\cdot)$ requiere 2 cosas:

i) probabilidad de Exito (COMPRA)

ii) ✓ de Falla (NO-COMPRA)

Estas probabilidades deben satisfacer:

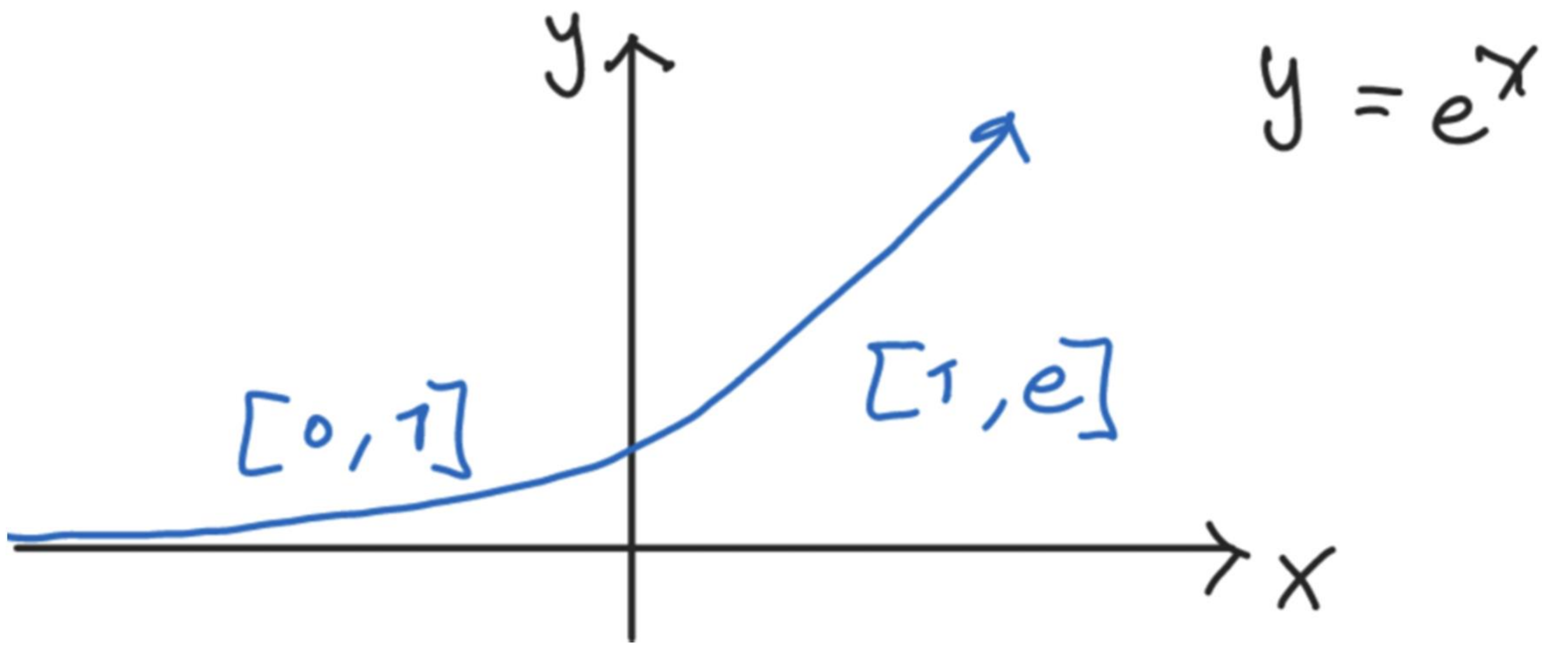
1. $P(\cdot) \geq 0$

2. $P(a) + P(b) \leq 1$

Debido a estas condiciones (Prob ≥ 0) expresaremos la regresión lineal como una función exponencial

¿Porqué?

Para cualquier valor de la pendiente m y peso de la var. independiente el valor será POSITIVO ✓



Para facilitar la notación $p = g(\cdot)$

$$p = e^{(\beta_0 + \beta_1 (age))} \quad (b)$$

Para garantizar que $P \leq 1$

$$p = \frac{e^{(\beta_0 + \beta_1 (age))}}{e^{(\beta_0 + \beta_1 (age))} + 1} \quad (c)$$

Recordemos que la función liga requiere la probabilidad:

a) De **ÉXITO**
b) De **FALLA**

De acuerdo a la ecuación ©,
la probabilidad de ÉXITO se expresa:

$$p = \frac{e^y}{e^y + 1}$$

en este caso p es la probabilidad de ÉXITO

Función LOGIT

Entonces para $p(\text{FALLA}) = q$

$$q = 1 - p = 1 - \left(\frac{e^y}{e^y+1}\right)$$

Si dividimos p/q

$$\frac{p}{q} = \frac{\frac{e^y}{e^y+1}}{1 - \frac{e^y}{e^y+1}} = \frac{\frac{e^y}{e^y+1}}{\frac{\cancel{e^y}+1-\cancel{e^y}}{e^y+1}} = \frac{\frac{e^y}{e^y+1}}{\frac{1}{e^y+1}}$$

$$= \frac{e^{2y}+e^y}{e^y+1} = \frac{e^y\cancel{(e^y+1)}}{\cancel{e^y+1}} = \boxed{e^y}$$

Entonces si P/Q

$$P/Q = \boxed{\frac{p}{1-p} = e^{y}}$$

$$\log\left(\frac{p}{1-p}\right) = y$$

Función Liga!!

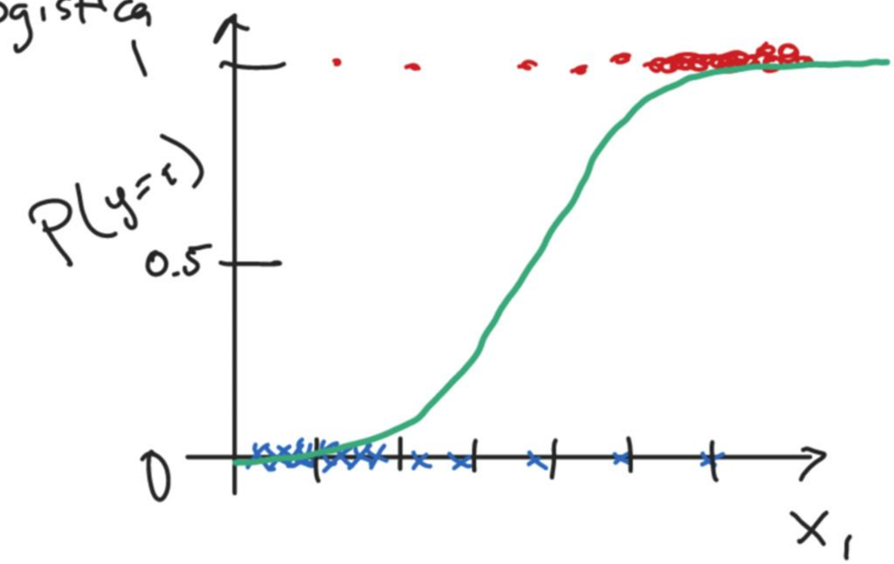
Pregunta: $\log\left(\frac{p}{1-p}\right)$ es lineal? NO

Una asociación NO-LINEAL con un modelo lineal

El modelo entonces queda:

$$\underbrace{\log\left(\frac{p}{1-p}\right)}_{\text{log-odds!!}} = \beta_0 + \beta_1 X_1$$

- Una gráfica típica de la regresión logística



- Comentarios del Modelo:

I. Cuando $\log\left(\frac{p}{1-p}\right) > 0$

↳ la probabilidad de ÉXITO es >50%.