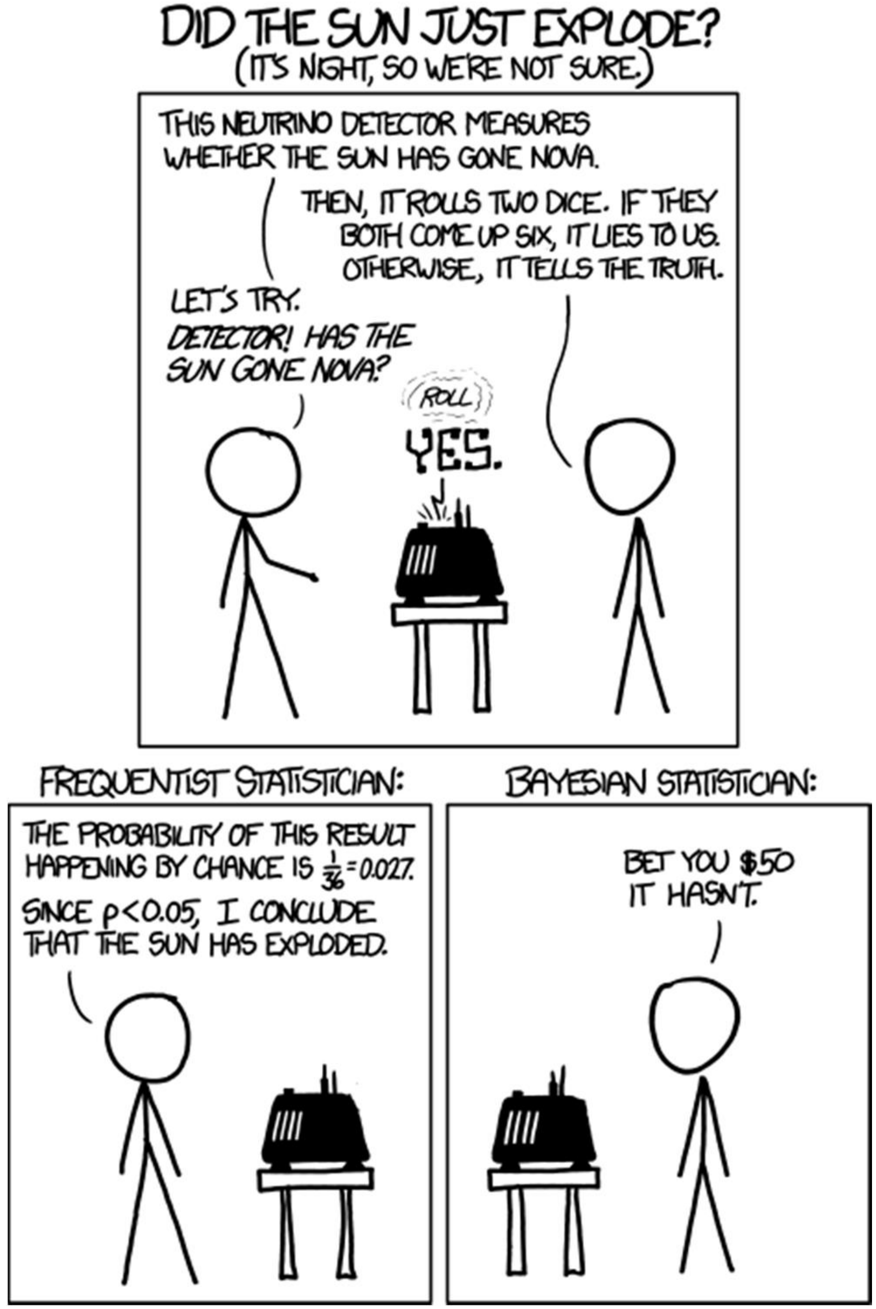
Aprendizaje Supervisado: Regresión Logística.

Guillermo Santamaría-Bonfil

Agenda

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

H_0: M miente
H_A: M No miente



$P(\text{S explotó}|\text{M si o no}) \approx 0$

https://www.explainxkcd.com/wiki/index.php/1132:_Frequentists_vs._Bayesians

Antecedentes:
Odds (Momios)

En algunos contextos Odds = Probability = Likelihood, en probabilidad **NO**.
En probabilidad nos referimos a “POSIBILIDADES”.

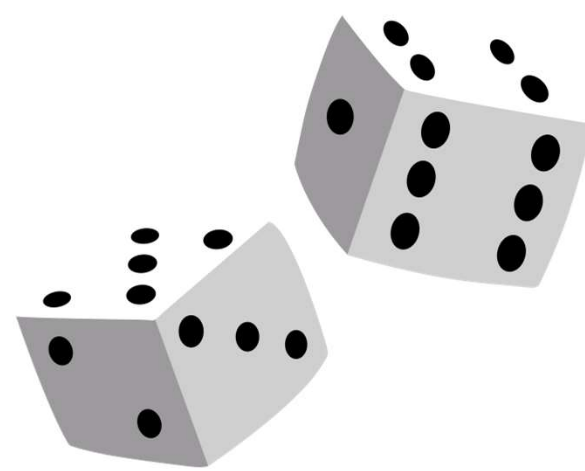
- La probabilidad de un evento como una proporción $\frac{\#Eventos}{Total}$

- La palabra “odds” se usa para representar la probabilidad de (POSIBILIDAD) un evento como probabilidades relativas . Es decir, la proporción de probabilidad de que un evento ocurra respecto a la probabilidad de que no ocurra.

$$\text{Otros Eventos} : \#\, Eventos$$

- Ej: La probabilidad de que en un dado justo caiga 6:

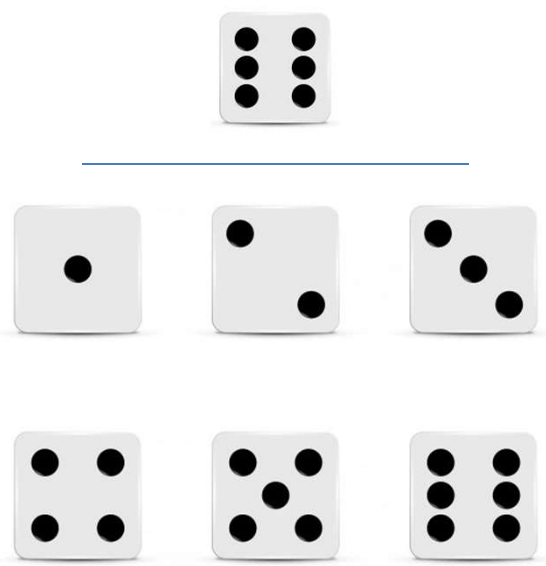
a) 1/6
b) 5:1
b.1) (5/6) : (1/6)



https://en.wikipedia.org/wiki/Odds#Mathematical_relations

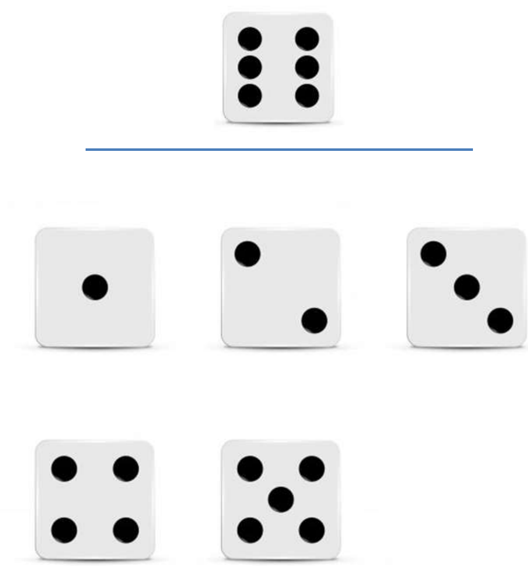
Ejemplo: caer un 6

Probabilidad



$= 1/6$

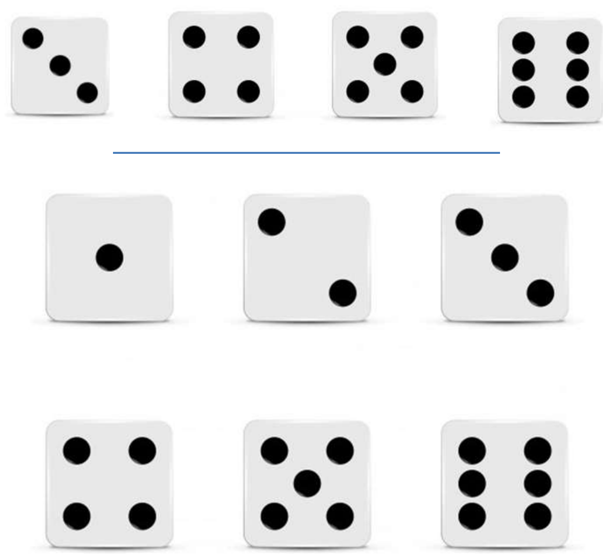
Odds



$= 1/5$

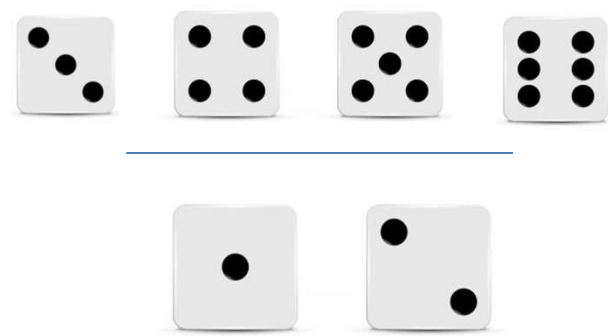
Ejemplo: caer un 3 o 4 o 5o 6

Probabilidad



$= 4/6$

Odds



$= 4/2$

De forma general, la diferencia entre probabilidad y mómios es

Probabilidad

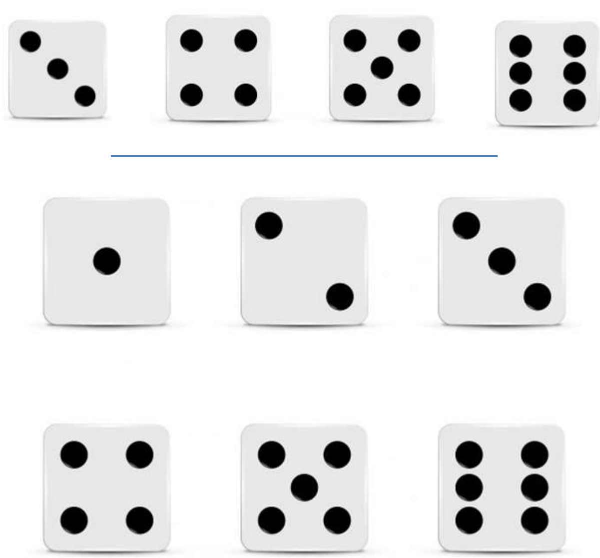
$$\frac{\text{La proporción de veces que SI SUCEDE un evento}}{\text{TOTAL de eventos}}$$

Odds

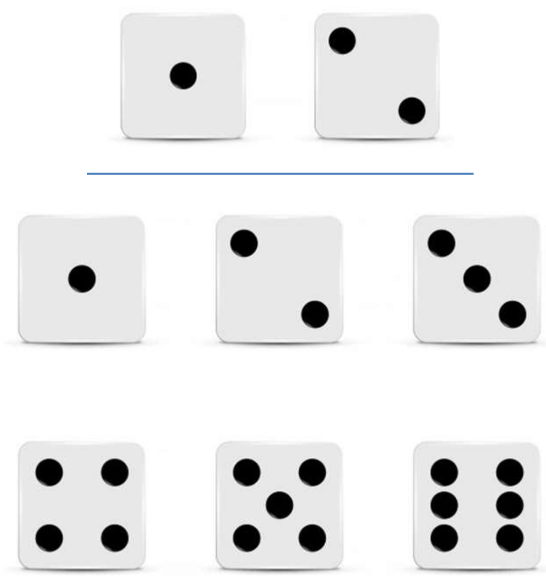
$$\frac{\text{La proporción de veces que SI SUCEDE un evento}}{\text{La proporción de que NO SUCEDE un evento}}$$

Calcular los Odds a partir de probabilidades

$$\frac{\text{La } \underline{\textbf{\textit{probabilidad}}} \text{ de que SI SUCEDE un evento}}{\text{La } \underline{\textbf{\textit{probabilidad}}} \text{ de que NO SUCEDE un evento}} = \frac{\frac{4}{6}}{\frac{2}{6}} = 2 \qquad = \frac{p}{1-p}$$



$= 4/6$



$= 2/6$

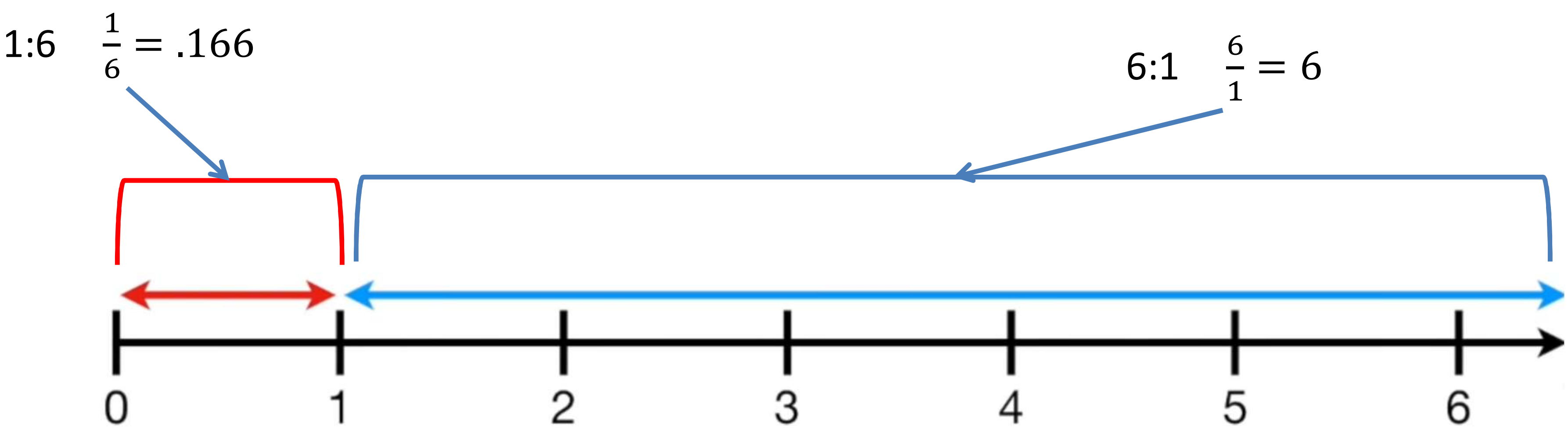
¿Cuál es el rango de valores de los Odds?

La ***probabilidad*** de que NO SUCEDE un evento

La ***probabilidad*** de que SI SUCEDE un evento

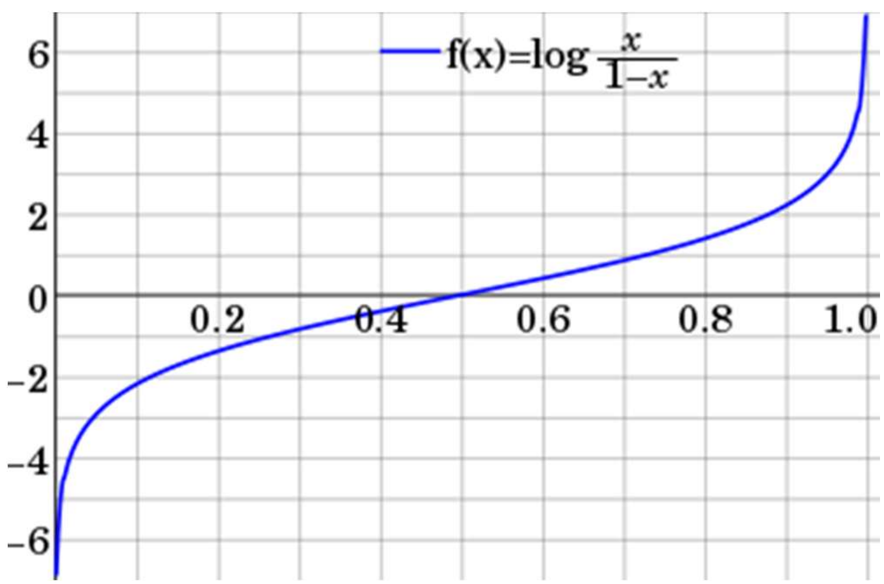
¿Cómo comparamos valores?

Supongamos, que estoy apostando en un juego de azar, y los odds de ganar son:



Antecedentes:
Log Odds

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



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

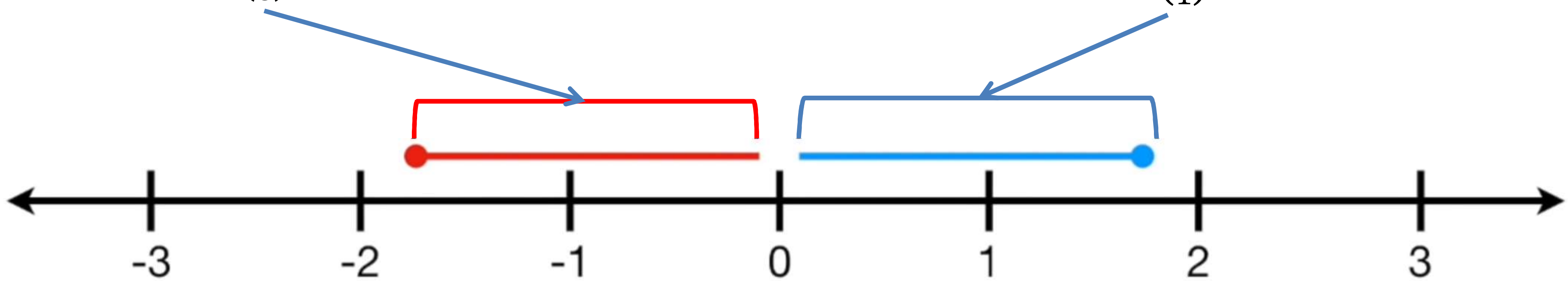
$$logit\big(P(A)\big) \in (-\infty, \infty)$$

Comparar las “posibilidades” es más fácil usando Log Odds

Supongamos, que estoy apostando en un juego de azar, y los odds de ganar son:

1:6 $\log\left(\frac{1}{6}\right) = -1.79$

6:1 $\log\left(\frac{6}{1}\right) = 1.79$



Antecedentes:
Log Odds

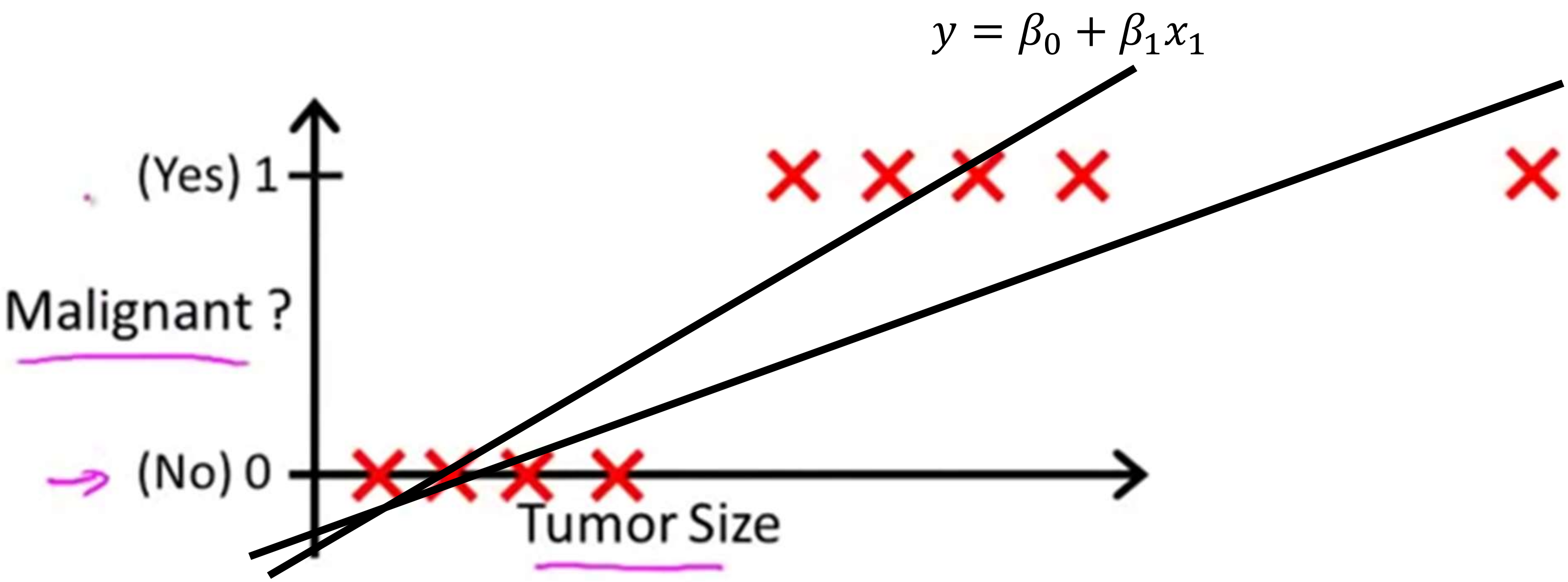
La función logística de cualquier número α esta dado por el logit inverso:

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

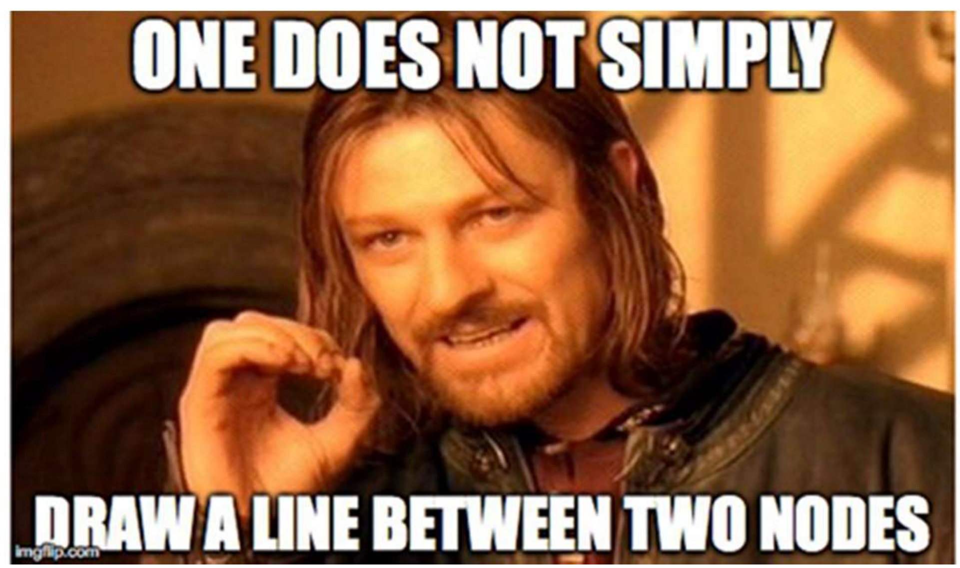
…

…

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



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) continua:

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

$$y \in \begin{Bmatrix} 0 & - & \text{no-compra} \\ 1 & - & \text{compra} \end{Bmatrix}$$

a) Empezamos por plantear una regresión lineal

con $y \rightarrow g(y)$

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

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:

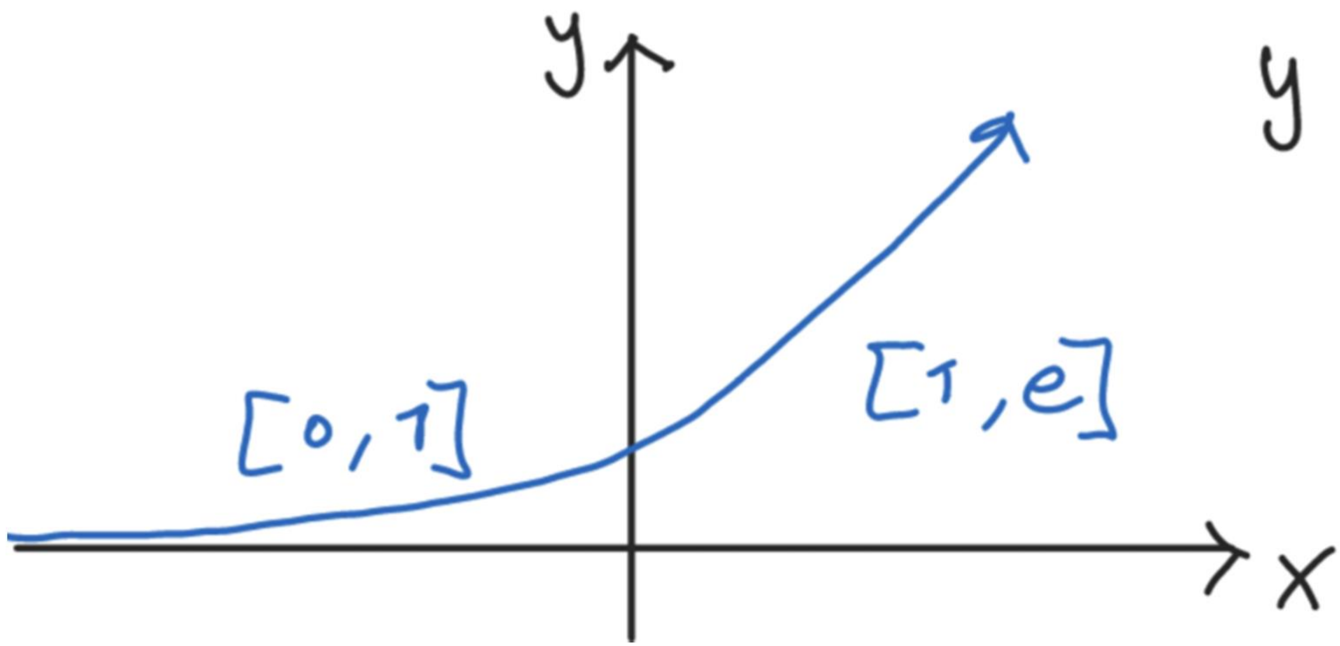
1. $P(\cdot) \geq 0$

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

Debido a estas condiciones ($Prob \geq 0$) 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 $p = g(\cdot)$

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

Para garantizar que $P \leq 1$

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

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!!

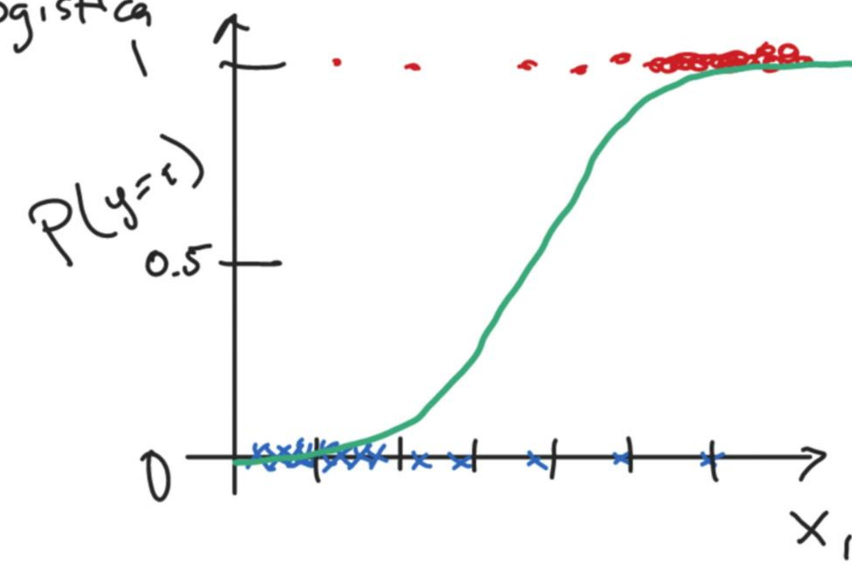
Pregunta: $\log\left(\frac{P}{1-P}\right)$ es lineal? <u>NO</u>

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%.