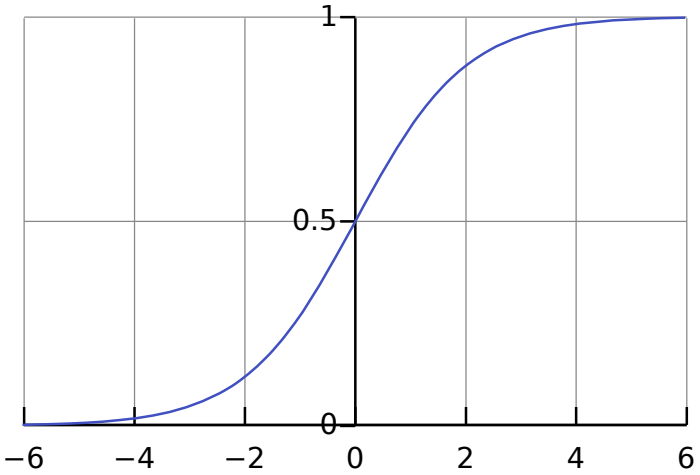
Regresión logística

¿Qué es regresión logística?

- Es un modelo lineal generalizado[1]:
  - Que generaliza la regresión lineal que estudiamos (obvio)
  - Permite una variable de salida (dependiente) con "errores" que **no** tienen una **distribución normal**
  - Relaciona la **distribución** de la variable de salida con un modelo lineal a través de una **función de enlace**
  - En este caso: la **función logística estándar**

$$f(z) = \frac{1}{1+e^{-z}}$$

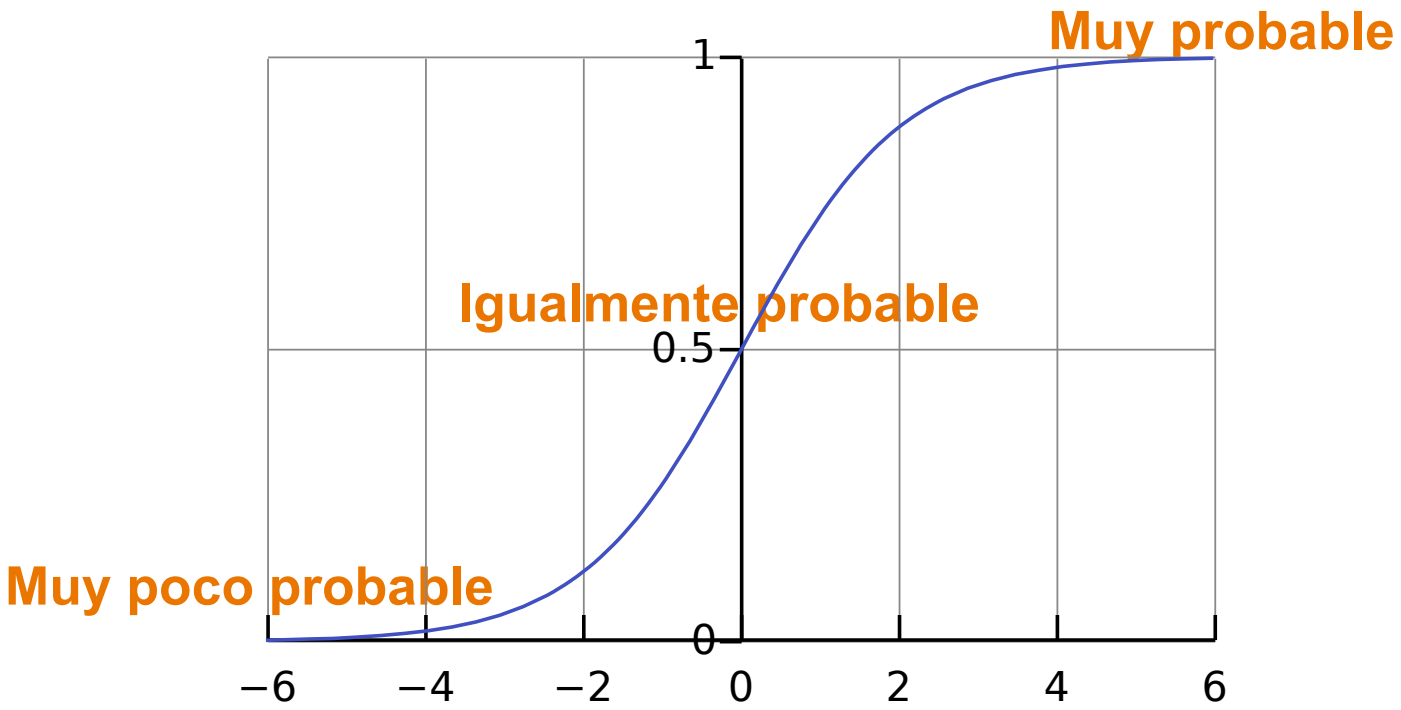


¿Y para qué sirve?

- ¿Qué gracia tiene esta función?
  - Si nos fijamos, describe una **transición de cero a uno**
  - ¿Qué tal si esa transición es de **probabilidad** de que ocurra algún evento?

¿Y para qué sirve?

- ¿Qué gracia tiene esta función?
  - Si nos fijamos, describe una **transición de cero a uno**
  - ¿Qué tal si esa transición es de **probabilidad** de que ocurra algún evento?



Ejemplo

- Veamos un ejemplo
  - Bien trillado, tanto, ¡que es el aparece en Wikipedia![2]

Ejemplo

- Veamos un ejemplo
  - Bien trillado, tanto, ¡que es el aparece en Wikipedia![2] :
    - probabilidad de pasar un examen *versus* horas de estudio

Ejemplo

- Veamos un ejemplo
  - Bien trillado, tanto, ¡que es el aparece en Wikipedia![2] :
    - probabilidad de pasar un examen *versus* horas de estudio
    - supongamos las siguientes respuestas (en orden creciente):

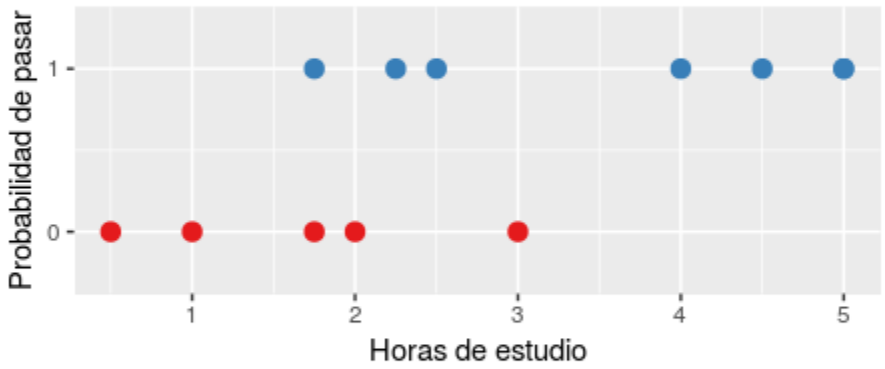
Horas	0.50	1.00	1.75	1.75	2.00	2.25	2.50	3.00	4.00	4.50	5.00	5.00
Pasa	0	0	0	1	0	1	1	0	1	1	1	1

Ejemplo

- Veamos un ejemplo
  - Bien trillado, tanto, ¡que es el aparece en Wikipedia![2] :
    - probabilidad de pasar un examen *versus* horas de estudio
    - supongamos las siguientes respuestas (en orden creciente):

Horas	0.50	1.00	1.75	1.75	2.00	2.25	2.50	3.00	4.00	4.50	5.00	5.00
Pasa	0	0	0	1	0	1	1	0	1	1	1	1

    - en un gráfico...

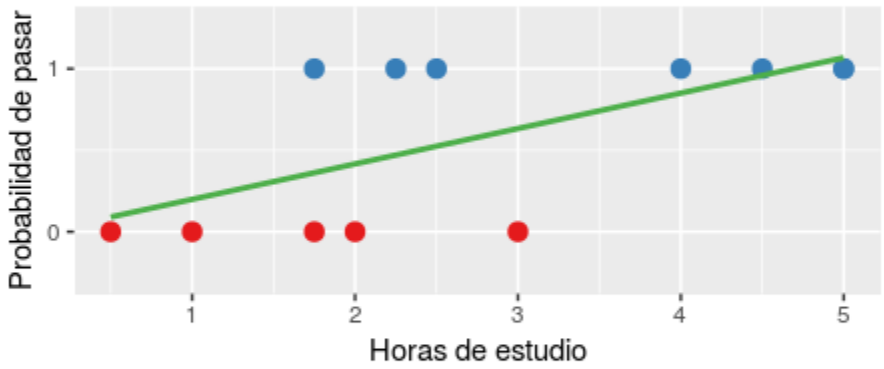


Ejemplo

- Veamos un ejemplo
  - Bien trillado, tanto, ¡que es el aparece en Wikipedia![2] :
    - probabilidad de pasar un examen *versus* horas de estudio
    - supongamos las siguientes respuestas (en orden creciente):

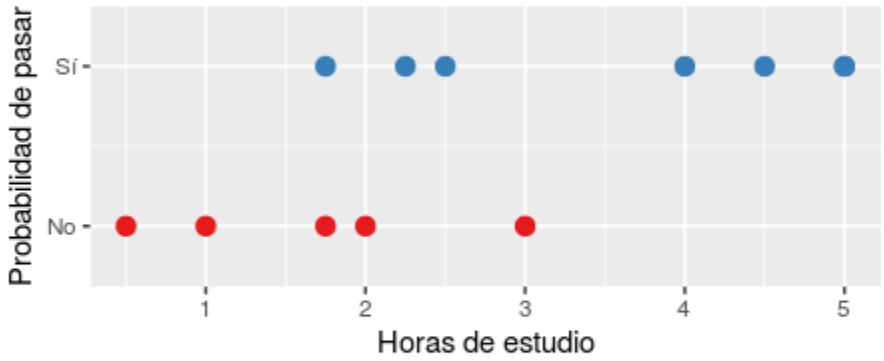
Horas	0.50	1.00	1.75	1.75	2.00	2.25	2.50	3.00	4.00	4.50	5.00	5.00
Pasa	0	0	0	1	0	1	1	0	1	1	1	1

- en un gráfico...
- claramente una recta **no es un modelo adecuado**
- ¿por qué?



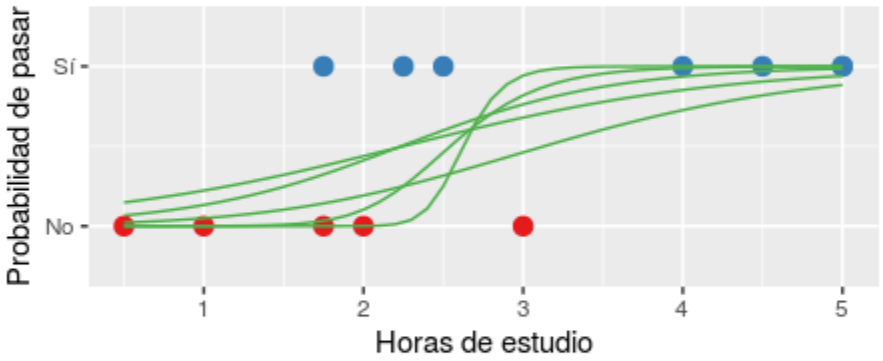
Ejemplo

- Porque la variable de salida no es numérica
  - Es nominal: Sí pasa o no pasa el examen
    - no podemos asociarla a una distribución normal
    - sino que a una **distribución binomial** (por ejemplo)



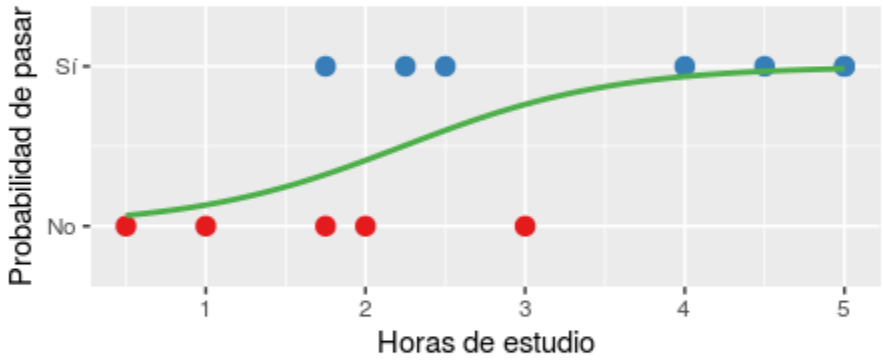
Ejemplo

- Porque la variable de salida no es numérica
  - Es nominal: Sí pasa o no pasa el examen
    - no podemos asociarla a una distribución normal
    - sino que a una **distribución binomial** (por ejemplo)
    - una **curva sigmoide** se ajusta mejor a este gráfico
    - ¿pero cuál elegir?



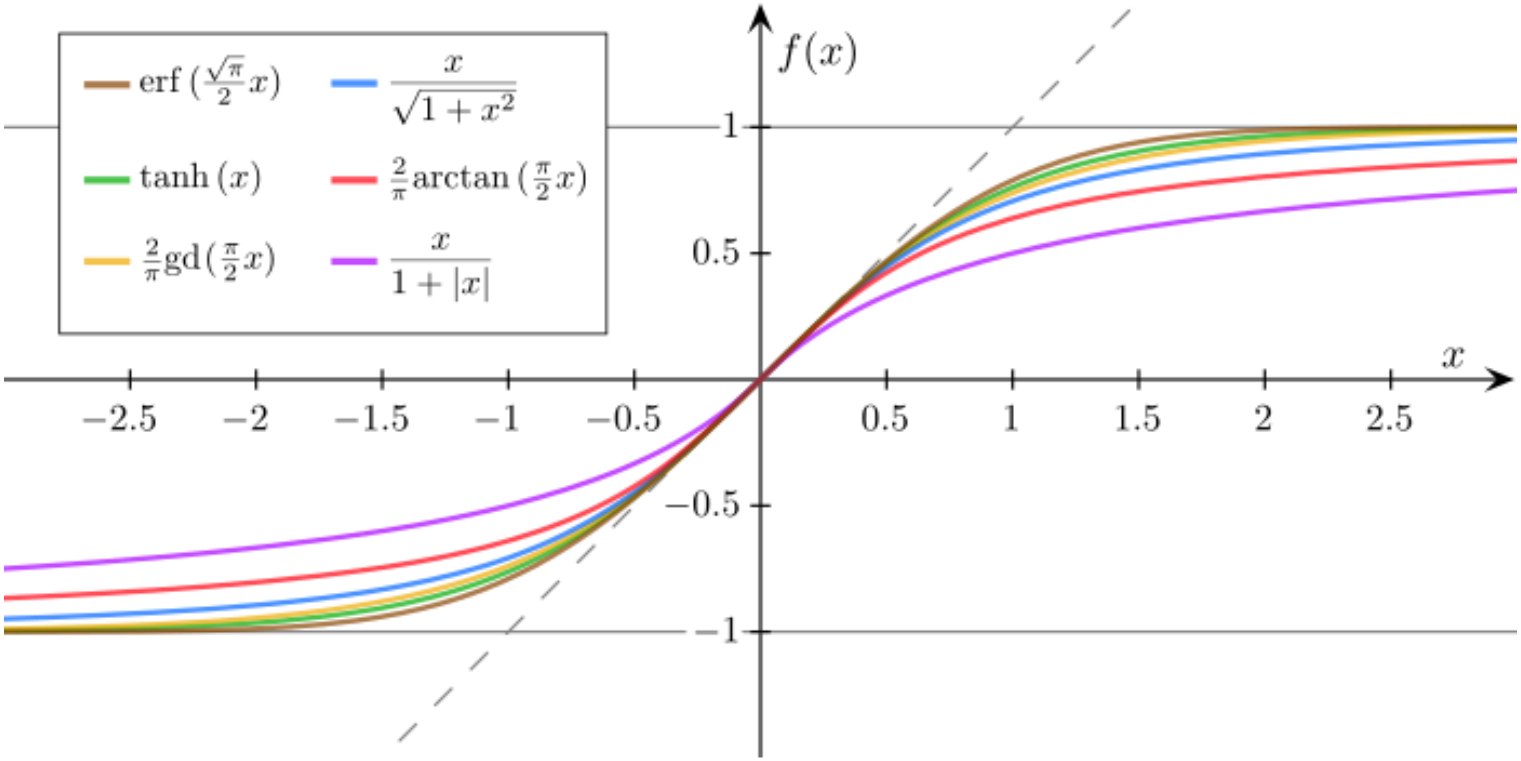
Ejemplo

- Porque la variable de salida no es numérica
  - Es nominal: Sí pasa o no pasa el examen
    - no podemos asociarla a una distribución normal
    - sino que a una **distribución binomial** (por ejemplo)
    - una **curva sigmoide** se ajusta mejor a este gráfico
    - ¿pero cuál elegir?
    - una opción es el **modelo logístico**



¿Por qué la función logística?

- Obviamente hay otras
  - Como la arcotangente, la tangente hiperbólica, la función error, la función logística generalizada, etc.



Fuente de la figura: https://commons.wikimedia.org/wiki/File:Gjl-t(x).svg

¿Por qué la función logística?

- Pero esta función es ***natural***[3]
  - En teoría (e.g. en estadística Bayesiana), y práctica
    - en varias áreas (notoriamente, en **medicina**) y en algunos países, resulta normal hablar de ***odds***[4]
    - ¡que no tiene traducción directa al castellano!
    - "*against all odds*" en Linguee: A pesar de todo, haciendo frente a adversidades de todo tipo, contra todo pronóstico, contra viento y marea, en desafío a todas **las probabilidades**, ...
    - esto es un problema, puesto que *odds* **no es lo mismo** que *probability* en inglés

¿Por qué la función logística?

- Veamos un ejemplo cotidiano[4]
  - Registros históricos dicen que en junio llueven 12 días
    - es decir, la probabilidad de día lluvioso es

$$p = 12 / 30 = 0{,}4$$

    - pero, los *odds* de día lluvioso *versus* día no lluvioso:

$$odds = 12 / 18 = 0{,}67$$

    - representan **la misma información**, de forma alternativa

¿Por qué la función logística?

- Veamos un ejemplo cotidiano[4]
  - Registros históricos dicen que en junio llueven 12 días
    - es decir, la probabilidad de día lluvioso es

$$p = 12 / 30 = 0{,}4$$

    - pero, los *odds* de día lluvioso *versus* día no lluvioso:

$$odds = 12 / 18 = 0{,}67$$

    - representan **la misma información**, de forma alternativa
    - sí, es una **medida** de la **verosimilitud** de un evento
    - como sabemos, podemos **relacionar** ambas medidas (suponiendo que el **logaritmo de los *odds*** sigue una distribución normal):

$$z = \log\left(\frac{p}{1-p}\right) \qquad p = \frac{1}{1+e^{-z}}$$

Usando regresión logística

- Pero esta condición ya la hemos visto
  - Por ejemplo, en la regresión lineal
    - luego podemos relacionar z con otra variable

$$z = \beta_0 + \beta_1 x$$

    - y así la función logística puede **modelar la relación** entre (el aumento de) **una variable** (e.g. horas de estudio) con la **probabilidad** de que un evento ocurra (e.g. pasar el examen)

Usando regresión logística

- Pero esta condición ya la hemos visto
  - Por ejemplo, en la regresión lineal
    - luego podemos relacionar z con otra variable

$$z = \beta_0 + \beta_1 x$$

    - y así la función logística puede **modelar la relación** entre (el aumento de) **una variable** (e.g. horas de estudio) con la **probabilidad** de que un evento ocurra (e.g. pasar el examen)
    - lo mejor, es que ya tenemos procedimientos para obtener la regresión anterior
    - solo nos falta darle utilidad: ahora podemos **predecir un evento** y **clasificar** un objeto en una de dos categorías

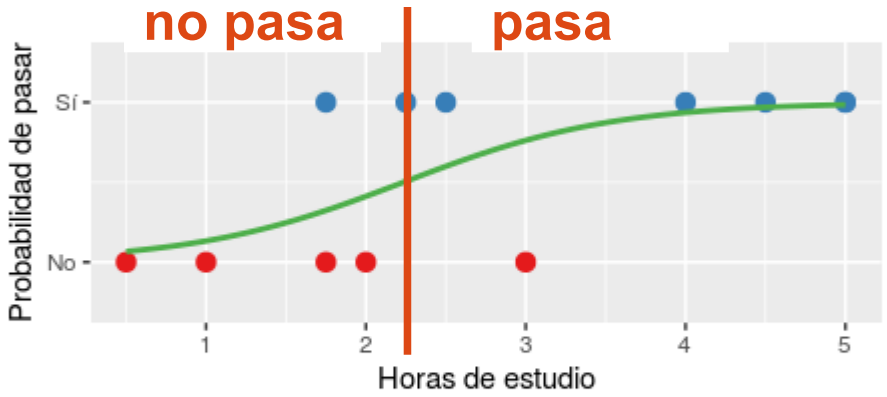
Usando regresión logística

- Equiprobable
  - Cuando el evento tiene igual “chance” de ocurrir o no ocurrir
    - ¿podemos considerar “chance” como traducción de *odds*?
    - p = 0,5; $odds$ = 1; log($odds$) = logit(p) = z = 0

    este punto genera una división, **un umbral**:
    - **a la izquierda** de este punto, se predice “no ocurre”;
      a la derecha se predice “sí ocurre”
    - a la izquierda de este punto, se predice clase A;
      a la derecha se predice clase B

Usando regresión logística

- En nuestro ejemplo
  - Podemos clasificar a los estudiantes
    - dos clases: pasan el examen, no pasan el examen
    - podemos ver que se cometen **errores en esta clasificación**
    - en general, la **calidad de un modelo** de clasificación se mide con esta tasa de errores
    - aunque también se puede mejorar con **umbrales distintos**



Referencias

[1] PennState (2018). STAT 504: Analysis of Discrete Data. The Pennsylvania State University, Eberly College of science. Obtenido en línea en agosto 2018 desde https://onlinecourses.science.psu.edu/stat504

[2] Wikipedia contributors (2018). Logistic regression. In Wikipedia, The Free Encyclopedia. Obtenido en línea en agosto 2018 desde https://en.wikipedia.org/w/index.php?title=Logistic_regression&oldid=854182949Obt

[3] Christopher M. Bishop (2006). Pattern Recognition and Machine Learning; Information Science and Statistics, Springer.

[4] Jaime Cerda, Claudio Vera, & Gabriel Rada (2013). Odds ratio: aspectos teóricos y prácticos. Revista médica de Chile, 141(10), 1329-1335. https://dx.doi.org/10.4067/S0034-98872013001000014