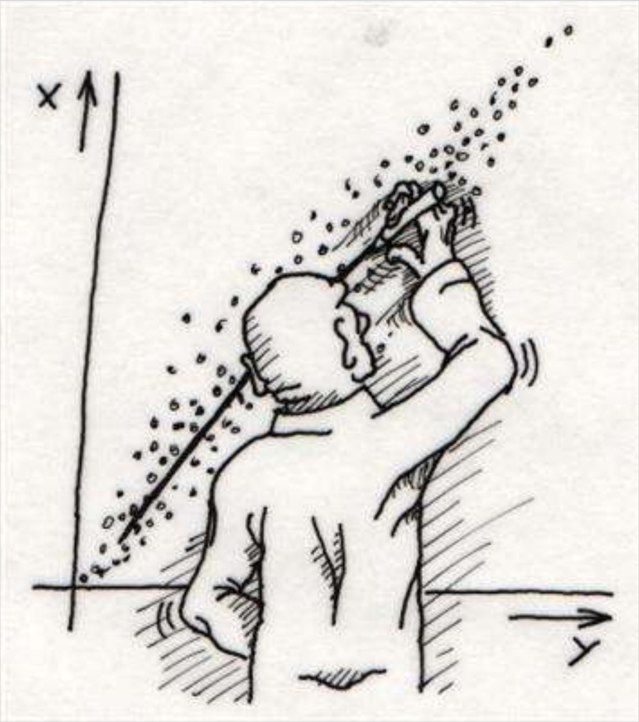


DigitalHouse >
Coding School

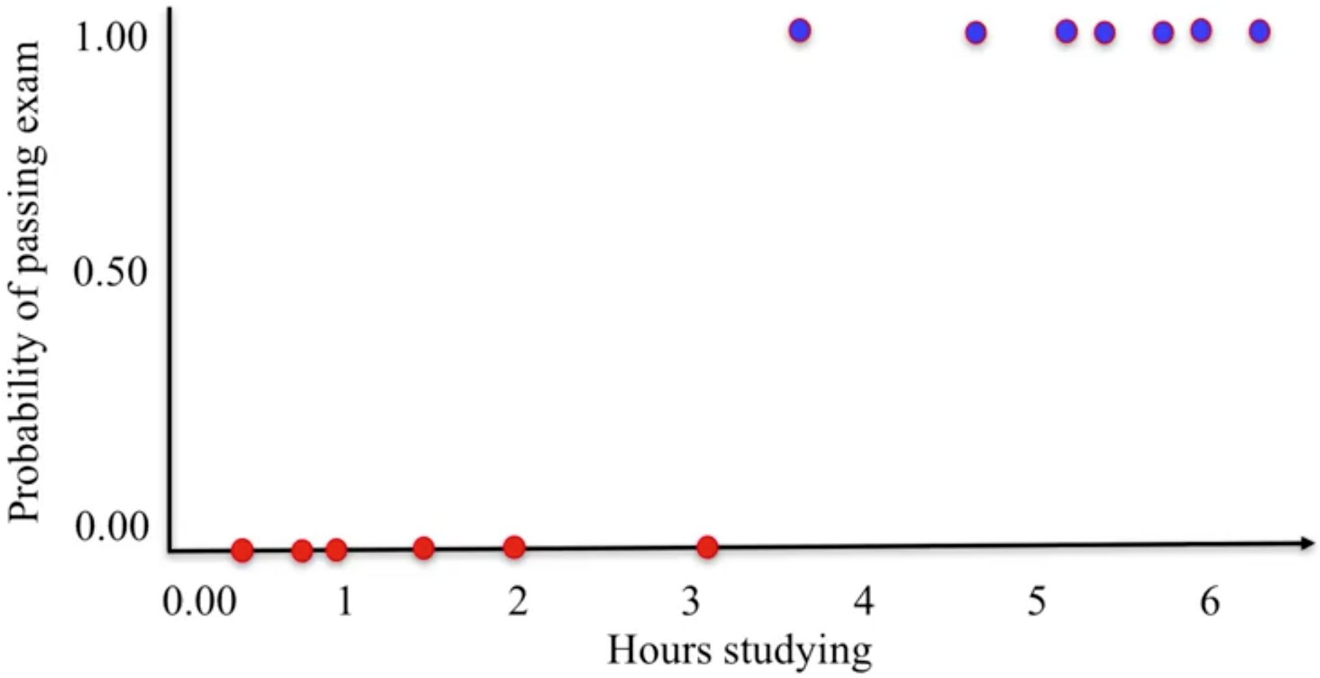
DATA SCIENCE

MÓDULO 4

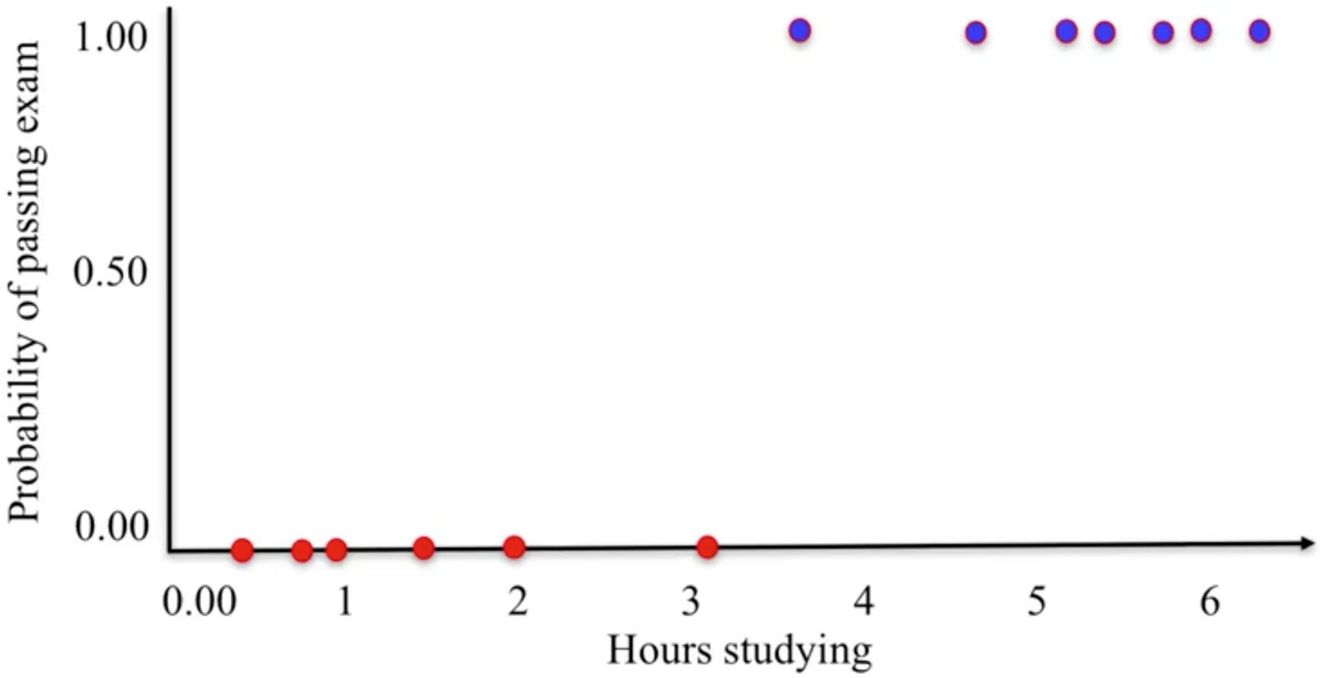
Introducción a la regresión logística

1. Describir el modelo de regresión logística
2. Presentar los fundamentos matemáticos de la regresión
3. Entrenar un modelo de regresión logística utilizando las librerías Statsmodels y Scikit-Learn de Python

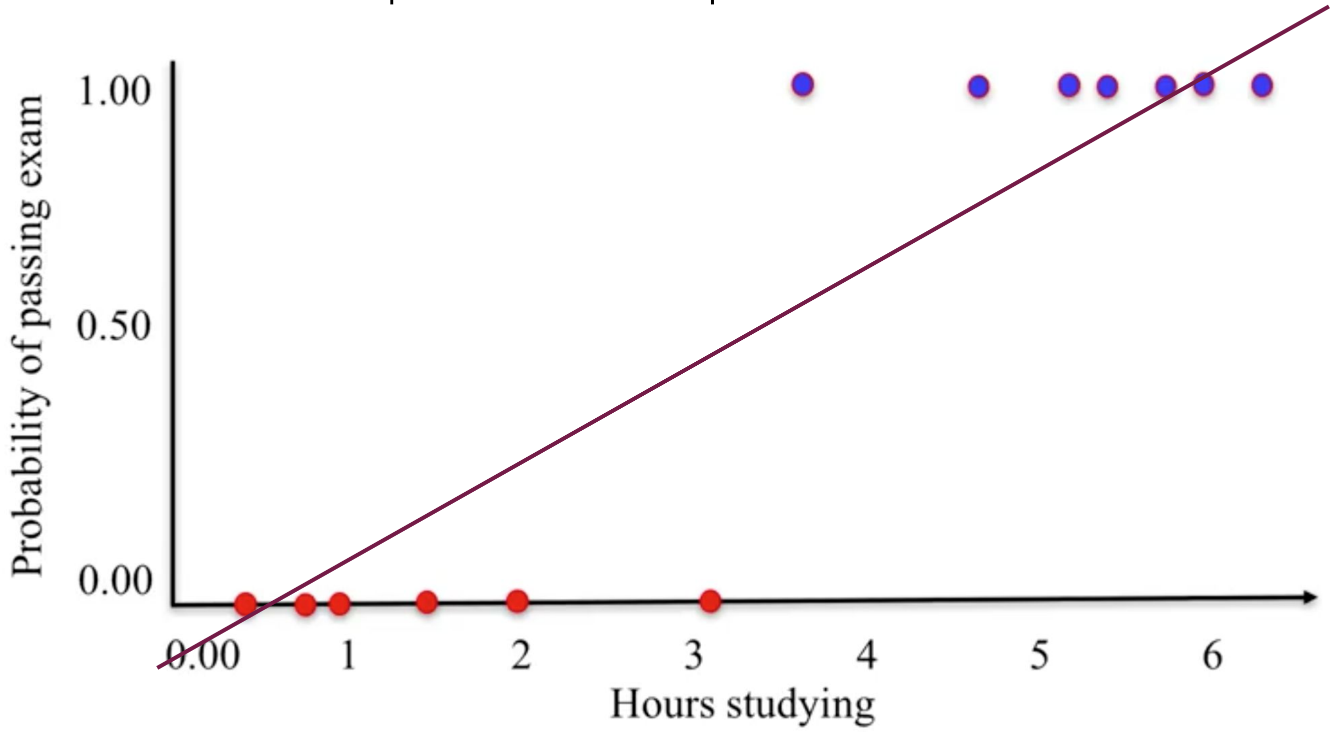
La regresión logística es un abordaje lineal para resolver **problemas de clasificación**.



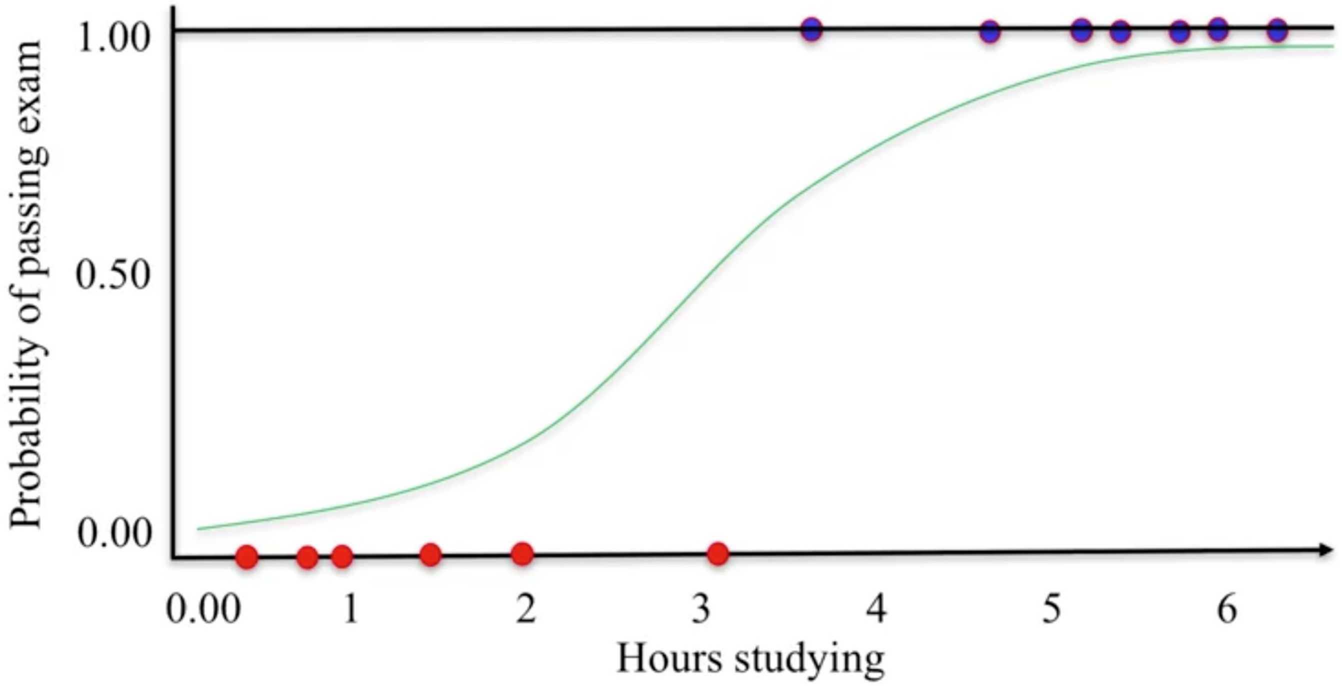
¿Podríamos usar una regresión lineal?



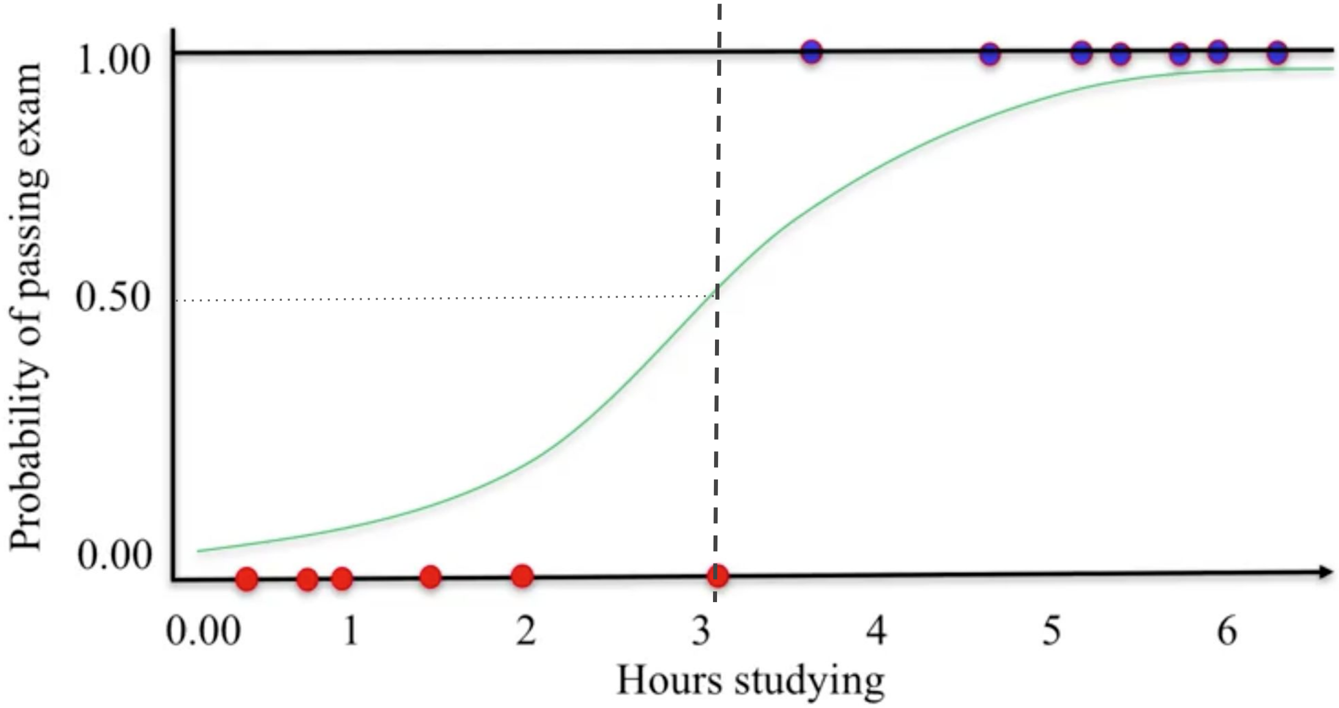
¿Podríamos usar una regresión lineal? Podríamos pero vamos a obtener valores fuera del rango [0,1], dificultando la interpretación como probabilidad.



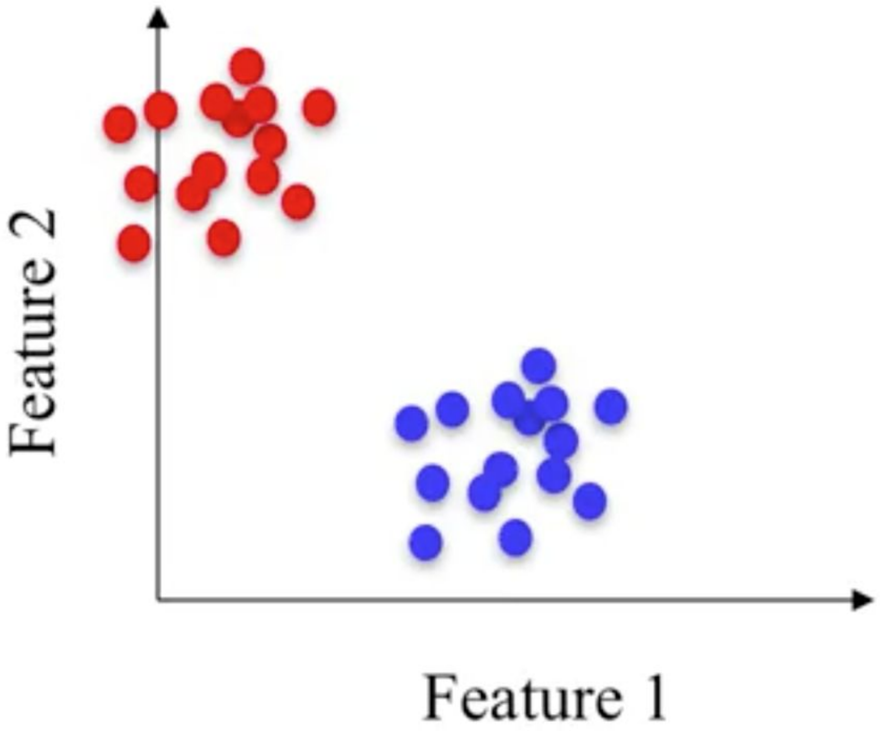
La regresión logística nos permite modelar la probabilidad de que la variable objetivo y pertenezca a una determinada categoría, dados los valores de las variables X.



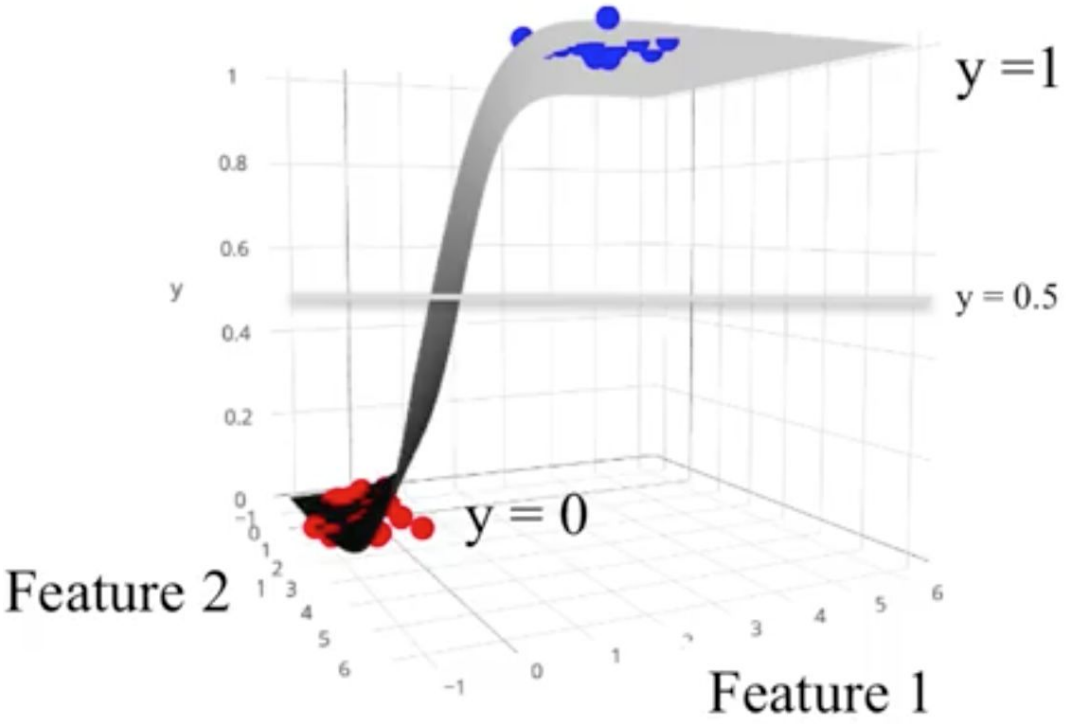
Podemos establecer una **frontera de decisión lineal**...



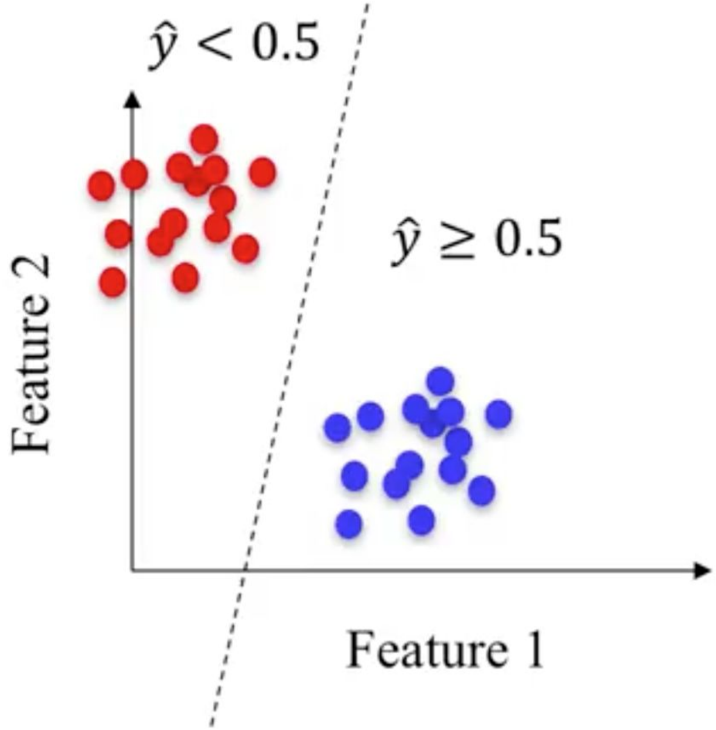
Si tuviéramos dos features:



Si tuviéramos dos features:



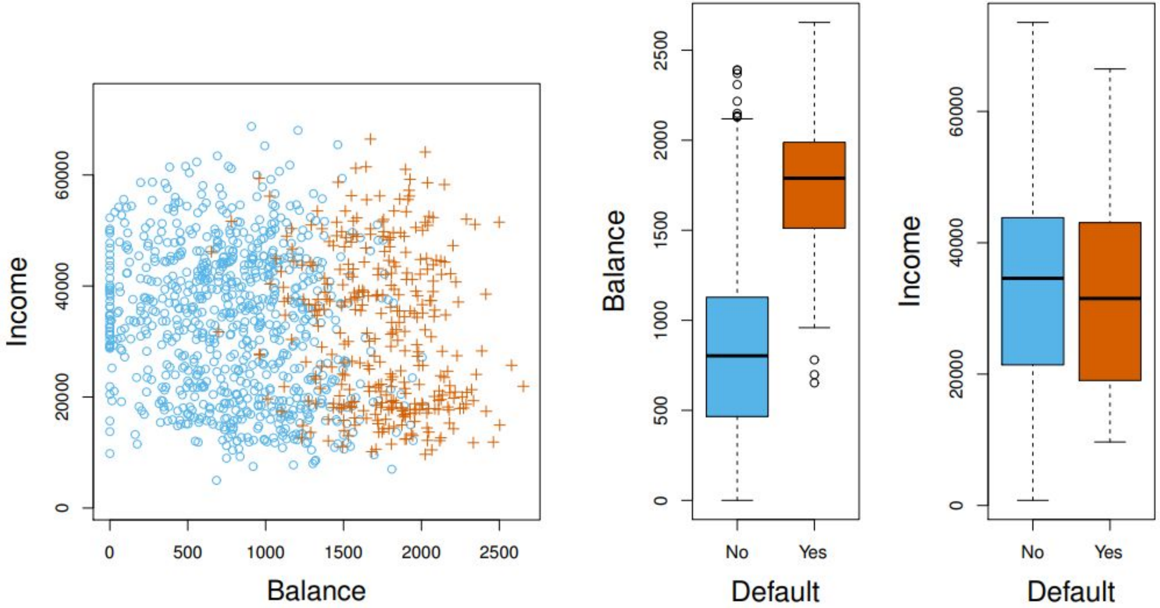
También tenemos una **frontera de decisión lineal**:



- Queremos predecir la **probabilidad** de que un cliente no pague su tarjeta de crédito (entre en *Default*):

$$Y = \begin{cases} 0 & \text{if } \texttt{No} \\ 1 & \text{if } \texttt{Yes}. \end{cases}$$

- Los **features** son
    - ingresos (*income*)
    - Deuda en su tarjeta de crédito (*balance*)



- Observando los datos, vamos a usar la variable *balance*. Es decir, queremos predecir p(y = 1 | balance)
- Si la p(y=1|balance) > 0.5 => default = Yes (podríamos elegir otro umbral)

- Si estimáramos p(y=1 | X) con una regresión lineal, nuestro modelo asumiría la siguiente forma:

$$p(X) = \beta_0 + \beta_1 X_1$$

- donde para abreviar, definimos $p(X) = p(Y = 1|X)$

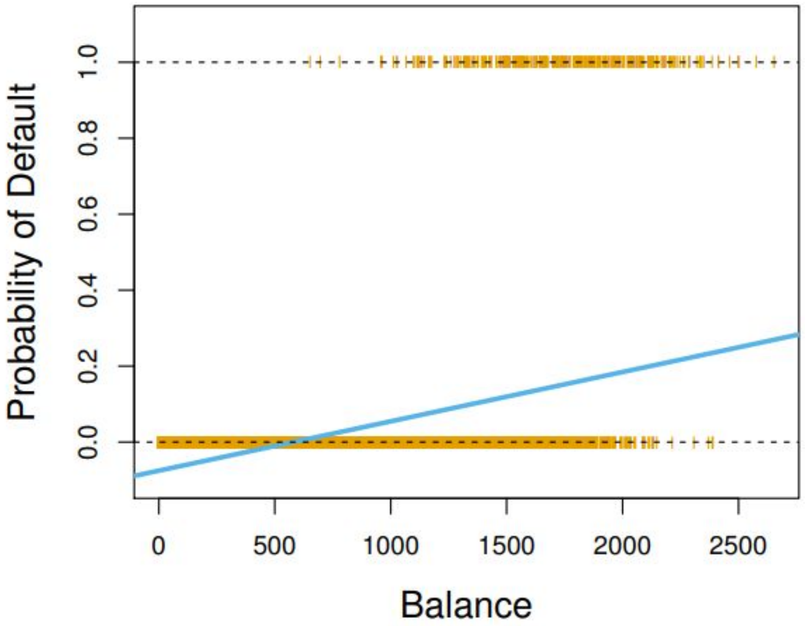
- Como vimos, esto arrojaría valores fuera del rango válido para una probabilidad [0,1]. Adicionalmente, cuando el problema de clasificación es multi-clase, el modelo tendería a interpretar a las diferentes clases como valores numéricos.

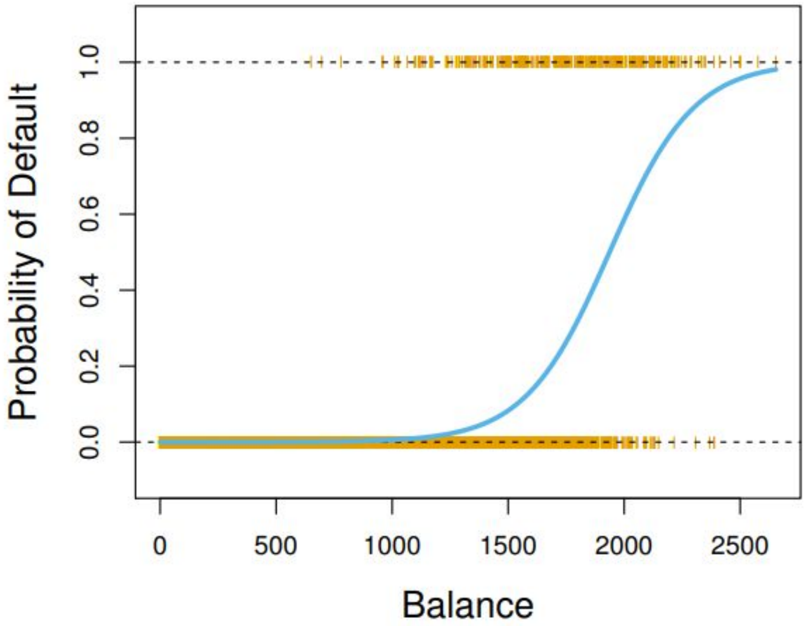
- Tenemos que buscar una función que nos garantice que las estimaciones que hagamos estarán dentro del rango válido de una probabilidad.
    - Podemos usar la función logística:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

- Vemos ahora que, sin que importe qué valores tome **X** siempre vamos a predecir valores dentro del rango [0,1].

La Regresión Logística





- Si manipulamos un poco la función logística que vimos hace algunas slides podemos llegar a la siguiente expresión:

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$$

- La cantidad p(X)/1-p(X) se denomina "odds-ratio" y lo que expresa es la relación entre la probabilidad de que y=1 (**p(X)**) y la probabilidad de que y=0 (**1-p(X)**).
- El odds-ratio toma valores entre 0 e infinito.

DigitalHouse>
Coding School

- Si tomamos logaritmos de la expresión anterior (odds-ratio)

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1$$

- vemos que el logaritmo del odds-ratio tiene una relación lineal con **X**.
- En el modelo lineal, los $\boldsymbol{\beta}_p$ eran el cambio promedio en Y ante un cambio unitario en **X**.
- En la regresión logística, incrementar una unidad en **X**, cambia el logaritmo del odds-ratio en $\boldsymbol{\beta}_p$. O, lo que es lo mismo, multiplica el odds por $e^{\beta p}$.
- La relación entre p(X) y X no es una línea recta => **cuánto cambia p(X) ante un cambio unitario en X** depende de los valores de **X**. Aún así, el signo de $\boldsymbol{\beta}_p$ expresa la dirección de cambio en p(X) (independientemente del valor de X).


Probability
1.0
0.8
0.6
0.4
0.2
0.0
−10 −5 0 5 10
Variable independiente
Odds
25000
20000
15000
10000
5000
0
−10 −5 0 5 10
Variable independiente
Log odds
10
5
0
−5
−10
−10 −5 0 5 10
Variable independiente

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.6513	0.3612	−29.5	<0.0001
balance	0.0055	0.0002	24.9	<0.0001

- Se observa que un incremento de $1 en el *balance* incrementa 0.0055 unidades el "log odds ratio".
- Hay un estadístico z que es análogo al estadístico t en regresión lineal.
  - Ho β_1 = 0 (o en otras palabras que la probabilidad de default no depende del balance
  - Ha β_1 <> 0

- Una vez que hemos estimados los coeficientes del modelo podemos hacer predicciones y computar la probabilidad de default para algún valor dado de balance.
- Por ejemplo, para un *balance* $1000 tenemos que está por debajo del 1%

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1,000}}{1+e^{-10.6513+0.0055\times 1,000}} = 0.00576,$$

- En cambio, para un balance de $2000 es mucho más grande y está cerca del 58%

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−3.5041	0.0707	−49.55	<0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004

- La regresión logística admite también variables cualitativas. Entrenamos un modelo utilizando la condición de ser estudiante sobre la probabilidad de default.
- En este caso, el coeficiente es positivo, lo cual indica que ser estudiante tiene una relación positiva con ser potencial moroso.
- ¿Qué pueden decir de la significación del coeficiente?

Podemos estimar las probabilidades de entrar en default siendo estudiante y no siéndolo (la variable dummy cuyos coeficientes habíamos estimado previamente).

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=Yes}) = \frac{e^{-3.5041+0.4049\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.0431,$$

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=No}) = \frac{e^{-3.5041+0.4049\times 0}}{1+e^{-3.5041+0.4049\times 0}} = 0.0292.$$

- Ahora, de forma análoga al caso de regresión lineal, pensemos en el problema de predecir una variable cualitativa binaria con una serie de p features.
- Nuestro modelo de regresión logística múltiple quedaría definido:

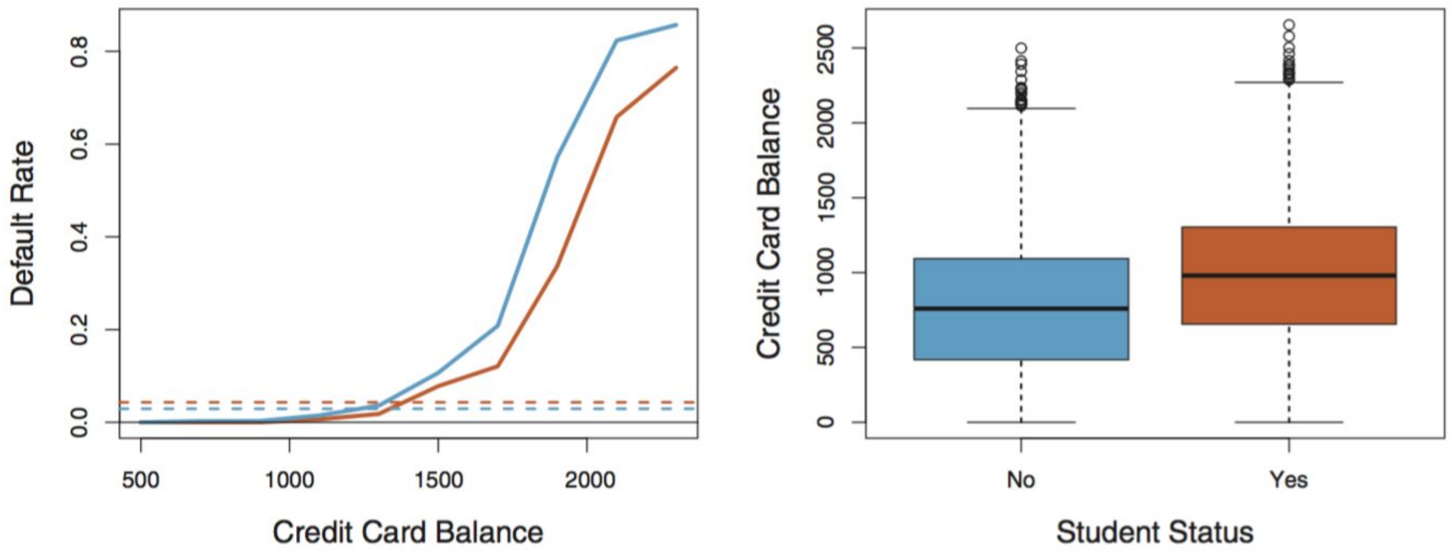
$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_p X_p}}$$

- Pudiendo ser reescrito en términos de logs odds:

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

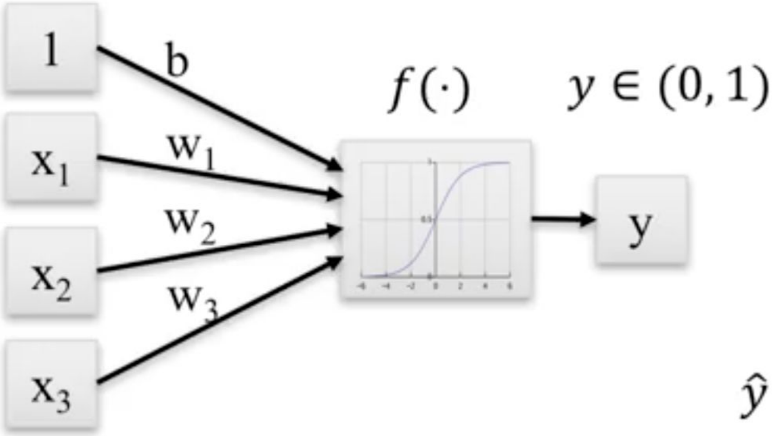
	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.8690	0.4923	−22.08	<0.0001
balance	0.0057	0.0002	24.74	<0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	−0.6468	0.2362	−2.74	0.0062

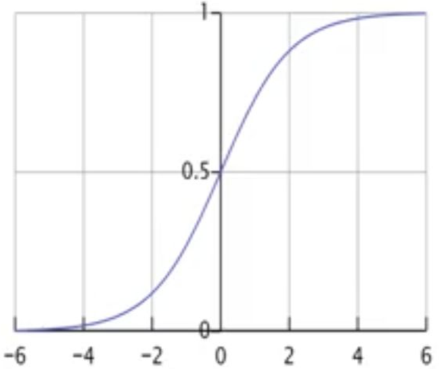
- Veamos los resultados de aplicar este modelo para predecir la probabilidad de default según el ingreso, el balance y la condición de estudiante.
- ¿Qué pueden decir de los resultados?
- ¿Ven algo raro?



- En el cuadro de la izquierda, en **naranja** se representan a los estudiantes y en **azul** a los no estudiantes. La línea punteada representa el % total de defaults, mientras que la línea sólida representa el % de defaults en función del balance de la tarjeta de crédito

Input features





$$\hat{y} = \text{logistic}(\hat{b} + \hat{w}_1 \cdot x_1 + \cdots \hat{w}_n \cdot x_n)$$

$$= \frac{1}{1 + \exp\left[-\left(\hat{b} + \hat{w}_1 \cdot x_1 + \cdots \hat{w}_n \cdot x_n\right)\right]}$$

Loss (error) function:

The loss function measures the discrepancy between the prediction ($\hat{y}^{(i)}$) and the desired output ($y^{(i)}$). In other words, the loss function computes the error for a single training example.

$$L(\hat{y}^{(i)}, y^{(i)}) = -(y^{(i)} \log(\hat{y}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)})$$

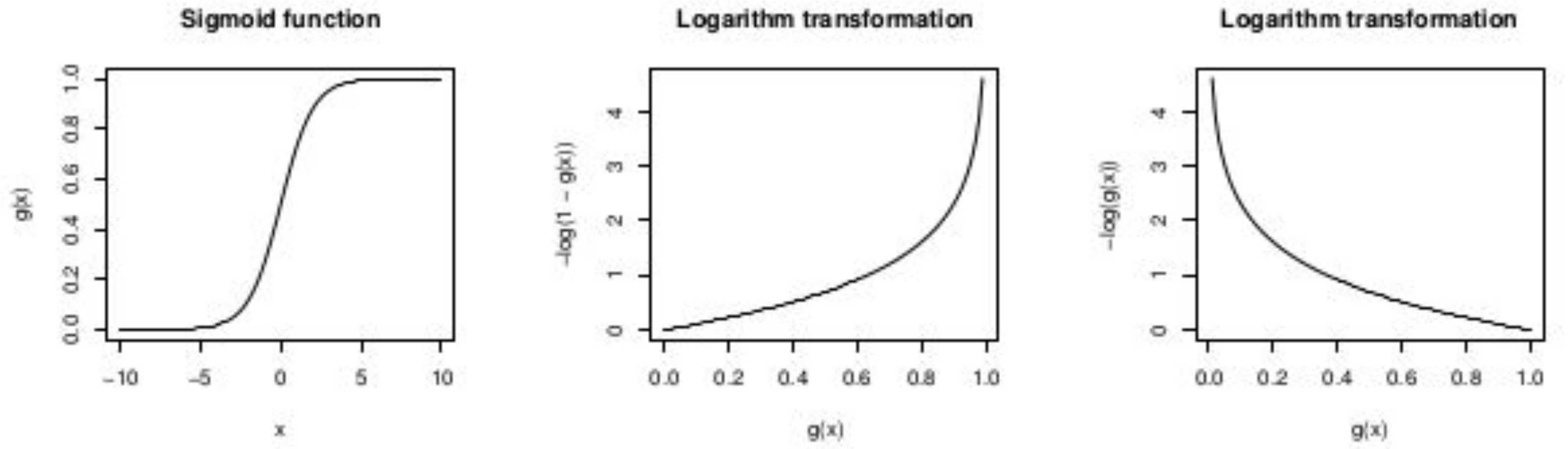
- If $y^{(i)} = 1$: $L(\hat{y}^{(i)}, y^{(i)}) = -\log(\hat{y}^{(i)})$ where $\log(\hat{y}^{(i)})$ and $\hat{y}^{(i)}$ should be close to 1
- If $y^{(i)} = 0$: $L(\hat{y}^{(i)}, y^{(i)}) = -\log(1 - \hat{y}^{(i)})$ where $\log(1 - \hat{y}^{(i)})$ and $\hat{y}^{(i)}$ should be close to 0

Cost function

The cost function is the average of the loss function of the entire training set. We are going to find the parameters w *and* b that minimize the overall cost function.

$$J(w, b) = \frac{1}{m} \sum_{i=1}^{m} L(\hat{y}^{(i)}, y^{(i)}) = -\frac{1}{m} \sum_{i=1}^{m} [(y^{(i)} \log(\hat{y}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)})]$$

La Regresión Logística: función de costo



(a) *Sigmoid function.* (b) *Cost for $y = 0$.* (c) *Cost for $y = 1$.*

Es posible aplicar **regularización** a la regresión logística, por ejemplo aplicando regularización con norma L2:

$$J(w,b) = -\frac{1}{m}\sum_{i=1}^{m}[(y^{(i)}\log(\hat{y}^{(i)}) + (1-y^{(i)})\log(1-\hat{y}^{(i)})] + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

Hay que tener en cuenta que en Scikit-Learn, la regresión logística aplica regularización por default. El parámetro *lambda* se reemplaza por su inversa, el parámetro *C*. A mayor valor de C, menos penalización. El valor por default de Scikit-Learn es C=1.

- Modelo para abordar problemas de clasificación con una variable target cualitativa o categórica, generalmente binaria.

- La relación entre la variable dependiente y los predictores es lineal al realizar la transformación logística de los datos.

- Pueden interpretarse los valores predichos por el modelo como "probabilidades" de cada uno de las categorías de la variable.

- Podemos realizar la interpretación de la influencia de las variables predictoras en términos de odd-ratio