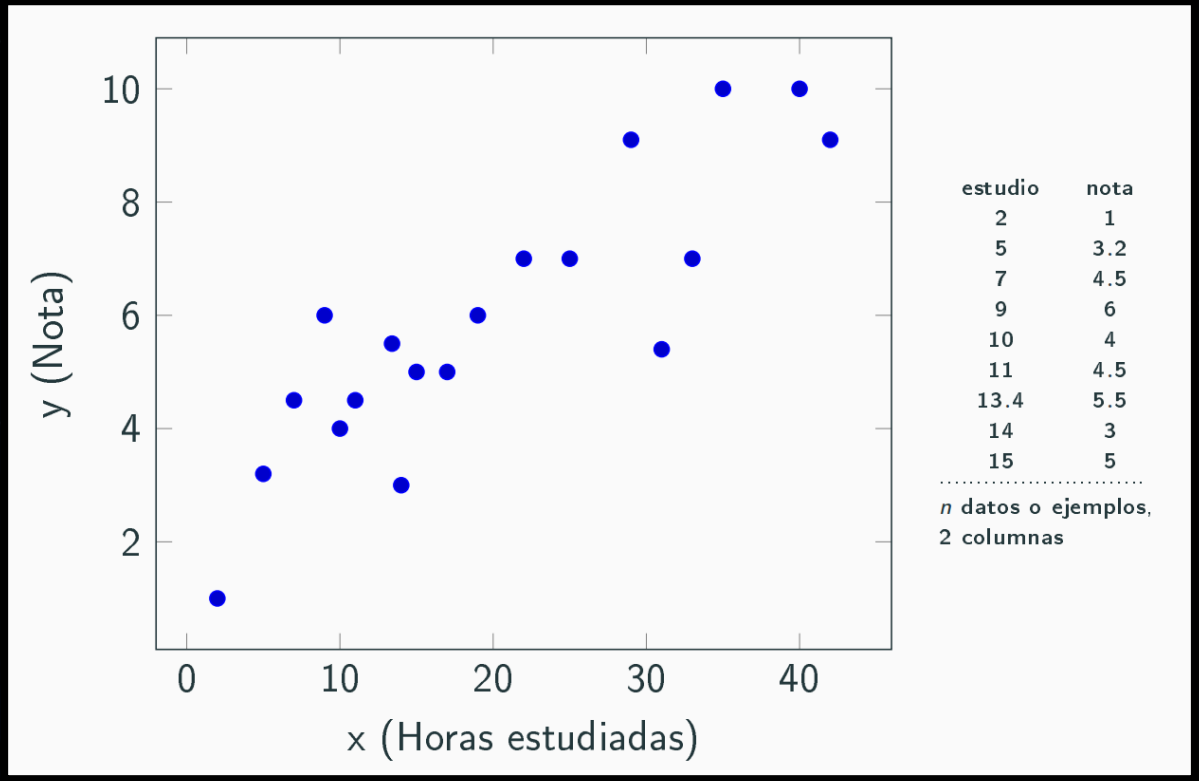
Regresión lineal

Problema de regresión



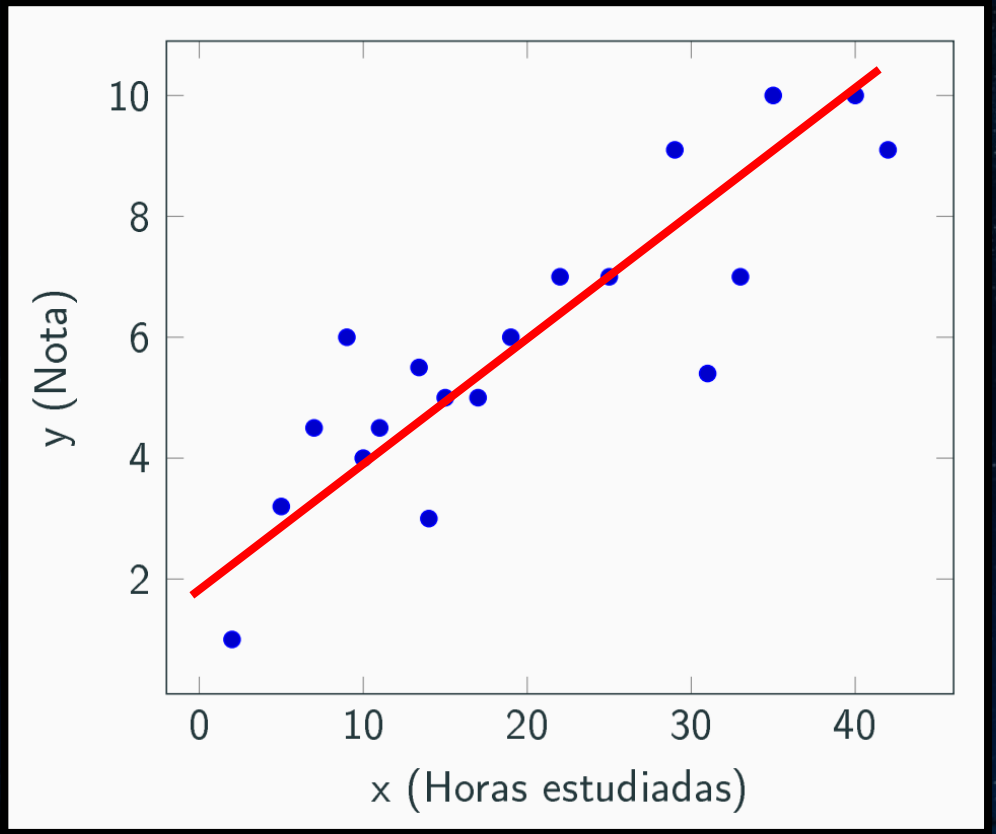
estudio	nota
2	1
5	3.2
7	4.5
9	6
10	4
11	4.5
13.4	5.5
14	3
15	5

n datos o ejemplos, 2 columnas

Si un nuevo alumno estudió x = 20hs,

¿cuál será su nota?

Problema de regresión

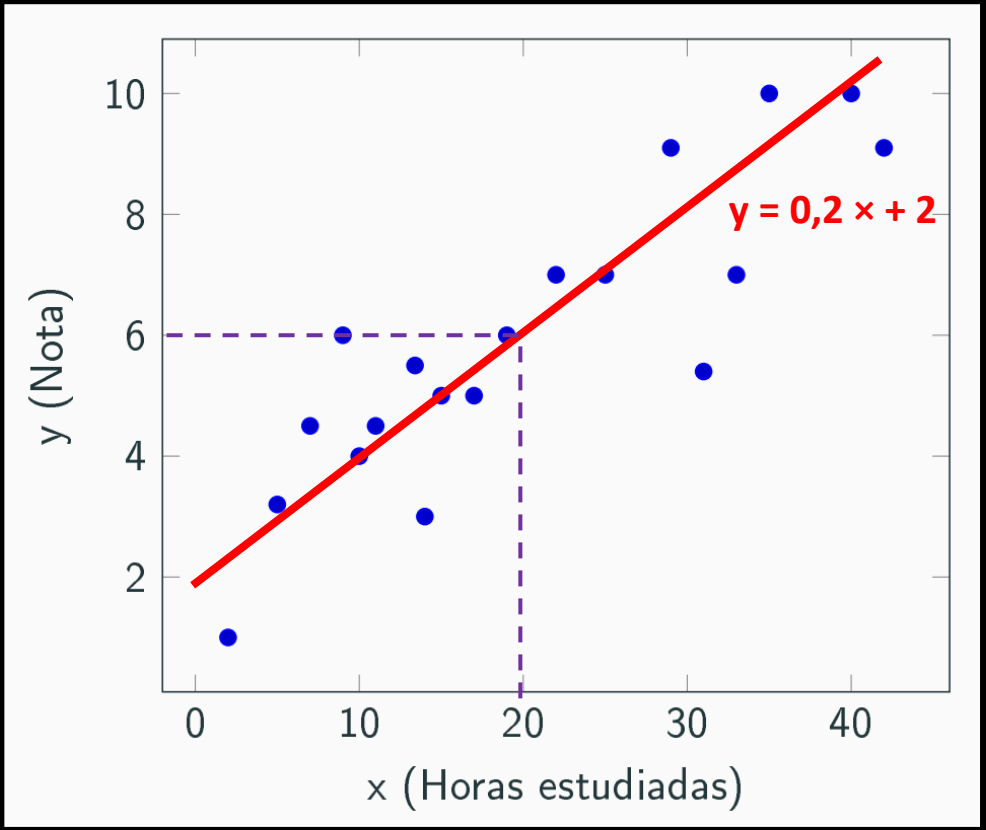


Asumimos que hay una relación lineal entre x e y:

Modelo= $y = mX + b$

Sólo necesitaríamos calcular los parámetros ***m y b***

Modelo de regresión lineal



Supongamos

m = 0.2

b = 2

f(x) = 0,2 × + 2

¿Qué nota predice el modelo si x=20?

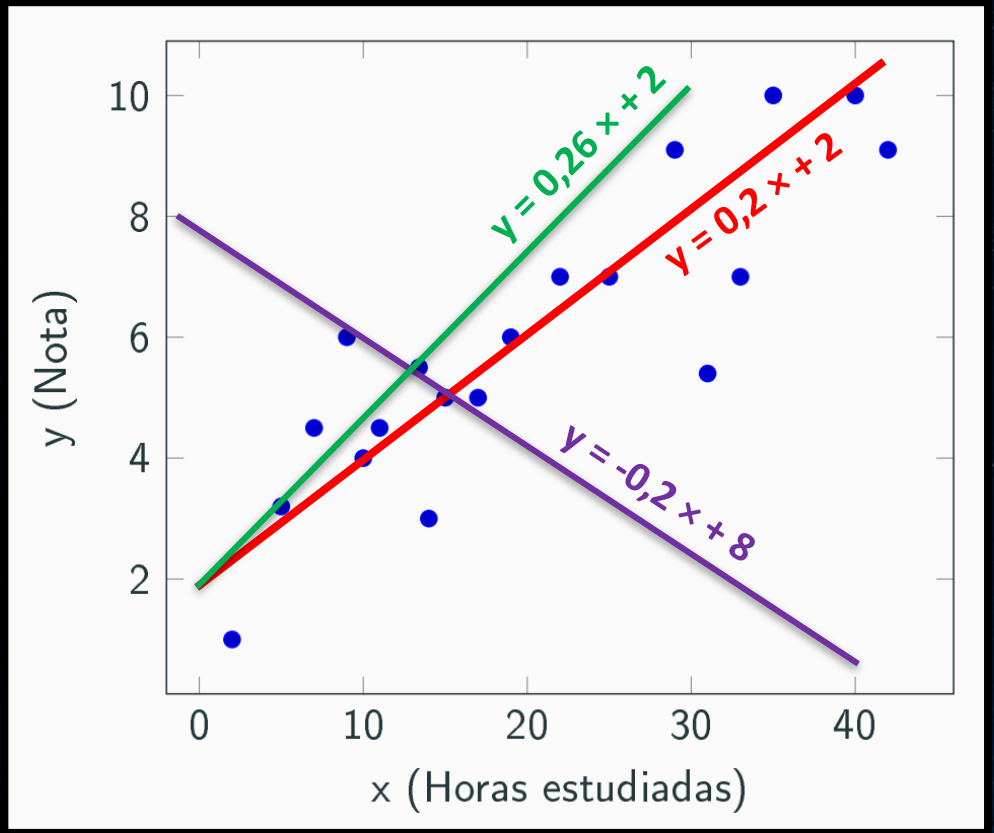
f(20) = m × 20 + b

= 0,2 × 20 + 2

= 6

Predice y=6 (nota =6)

Modelo de regresión lineal



Valores de m y b definen la recta:

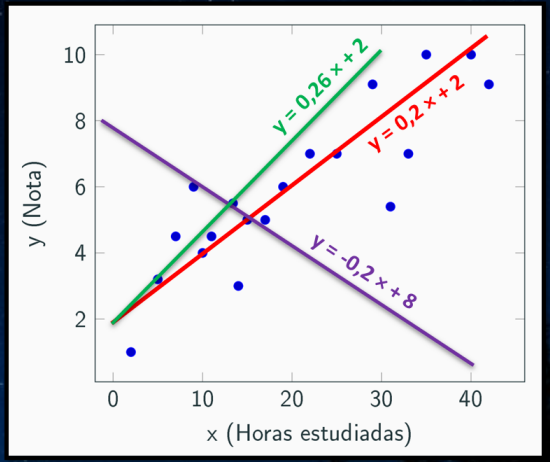
- **m** = 0.20, **b** = 2
- **m** = 0.26, **b** = 2
- **m** = -0.20, **b** = 8

Parámetros del modelo

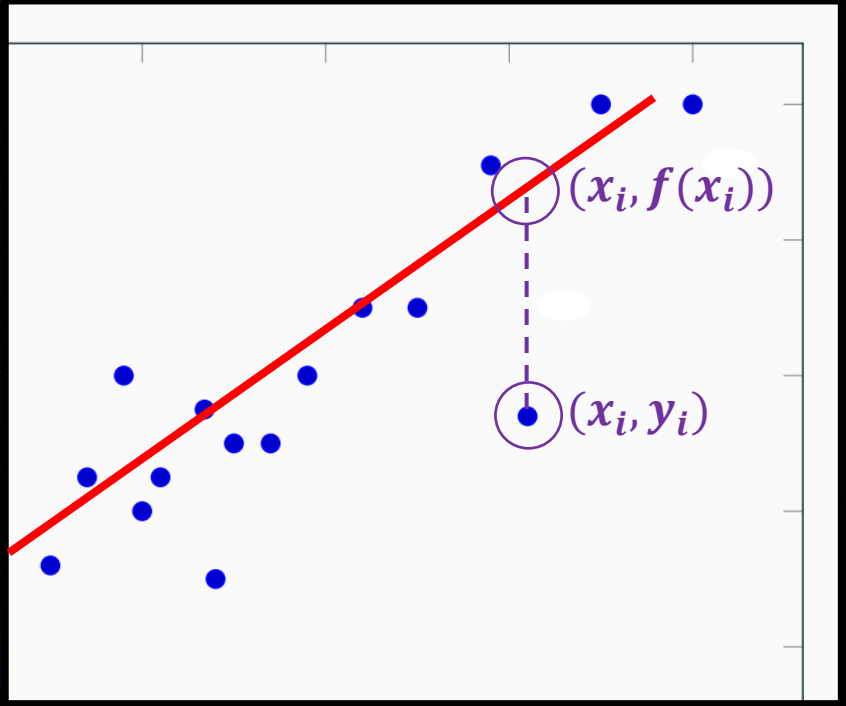
- m indica la pendiente
- b la ordenada a la origen

Modelo lineal

- Es el modelo más simple (polinomio de grado 1)
- Una recta nunca va a poder pasar por todos los puntos. Cada punto (x, y) se aproxima con cierto error.
- ¿Cómo elegimos un modelo? Necesitamos una medida de error.



Error del modelo

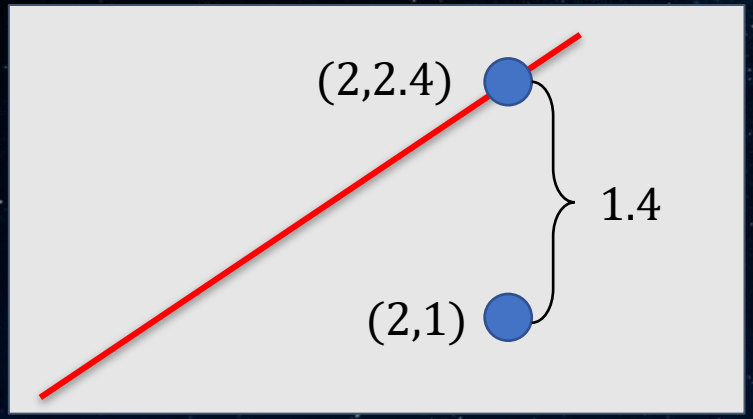


Error de un dato:

- $E_i(m, b) =$ Error del dato i para m y b
- Distancia cuadrática entre el valor esperado (y_i) y el predicho por el modelo ($f(x_i)$)
- $E_i(m, b) = (y_i - f(x_i))^2$

  $= (y_i - mx_i + b)^2$

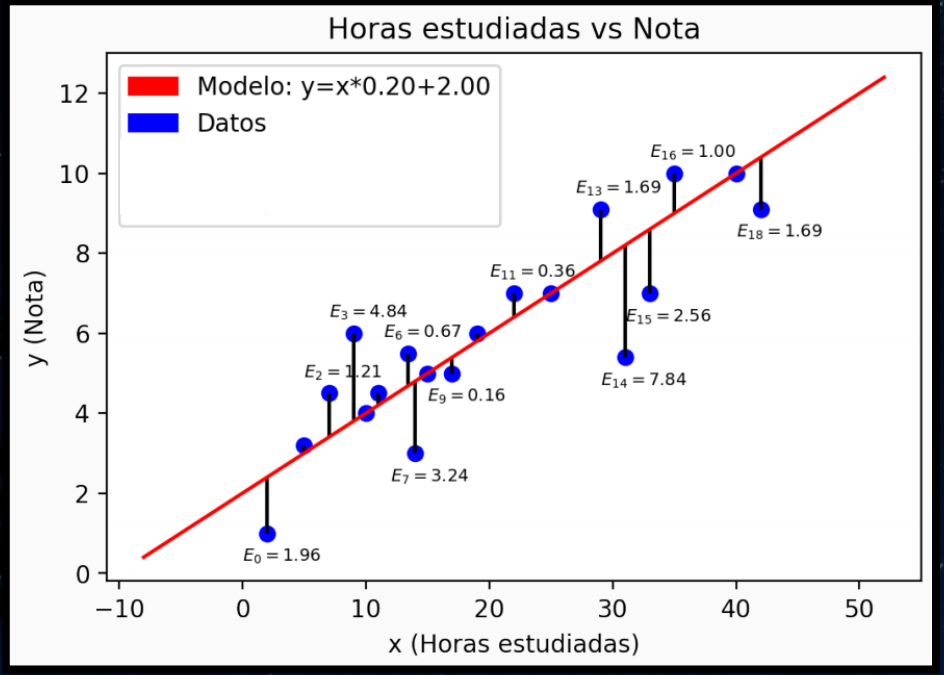
Error del modelo



$y_i-f(x_i)$	$= -1.4$
$\lvert y_i-f(x_i)\rvert$	$= 1.4$
$(y_i-f(x_i))^2$	$= 1.96$
$(y_i-f(x_i))^4$	$= 3.84$

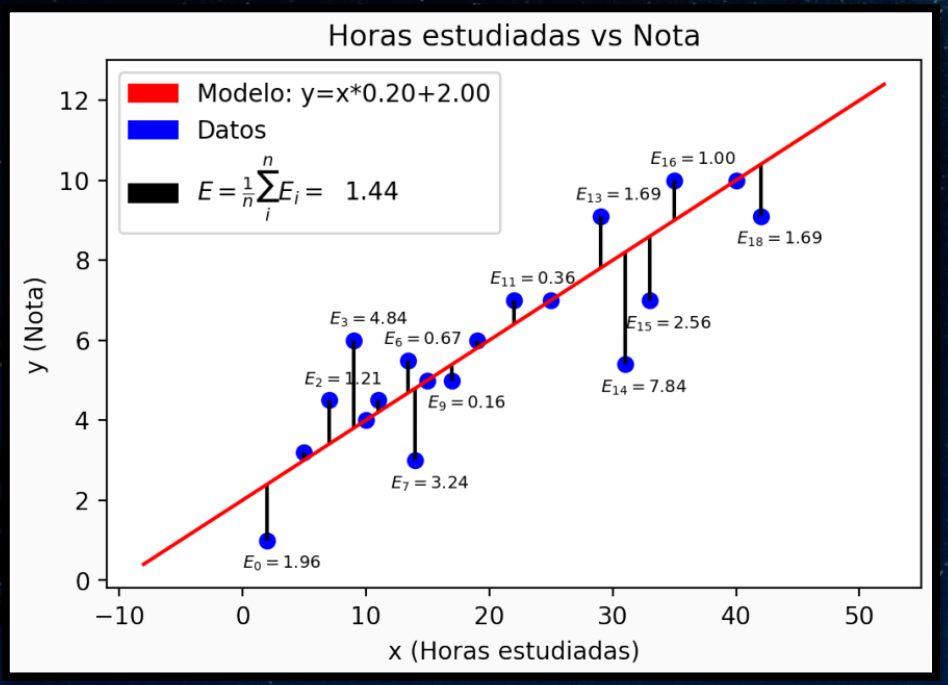
- $\mathbf{E_i = (y_i-f(x_i))^2}$
  ¿Por qué esta función de error?
- ¿Por qué no usar $\mathbf{y_i-f(x_i)}$?
  Valores negativos
- ¿Por qué no usar $\mathbf{|y_i-f(x_i)|}$?
  No es una función derivable
  Difícil de optimizar
- ¿Qué efecto tiene el ^{2}?
  Penaliza más errores grandes
  $0.5^2=0.25$, $1^2=1$, $5^2=25$
- ¿Por qué no usar $\mathbf{(y_i-f(x_i))^4}$?
  Posible, pero penalizaría demasiado

Error del modelo



Necesitamos ahora evaluar todos los errores generados por el modelo.

Función de costo del modelo



Error cuadrático medio

$$E = \frac{1}{n}\sum_{i}^{n} E_i$$

$$E = \frac{1}{n}\sum_{i}^{n} (y'_i - y_i)^2$$

- y'_i = el valor predicho de mi modelo.
- y_i = el valor esperado (real) para x.

Error cuadrático medio - Ejemplo

$m = 0.2$

$b = 2$

estudio	nota
2	1
10	4
14	5
30	9
40	10

$f(x_i)$	$f(x_i)-y_i$	E_i
$0.2 * 2 + 2 = 2.4$	$2.4 - 1 = 1.4$	$(1.4)^2 = 1.96$
$0.2 * 10 + 2 = 4$	$4 - 4 = 0$	$(0)^2 = 0$
$0.2 * 14 + 2 = 4.8$	$4.8 - 5 = -0.2$	$(-0.2)^2 = 0.04$
$0.2 * 30 + 2 = 8$	$8 - 9 = -1$	$(-1)^2 = 1$
$0.2 * 40 + 2 = 10$	$10 - 10 = 0$	$(0)^2 = 0$

Error cuadrático medio

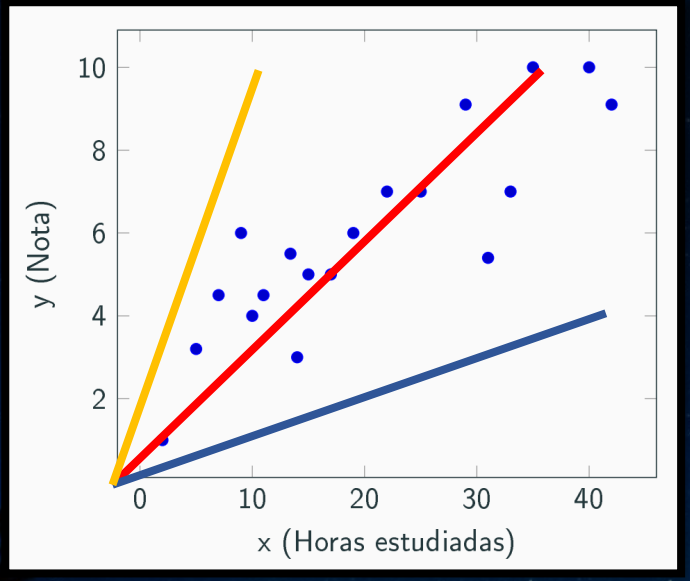
$$E = \frac{1}{n}\sum_i^n E_i = \frac{1.96+0+0.04+1+0}{5} = \frac{3}{5} = 0.6$$

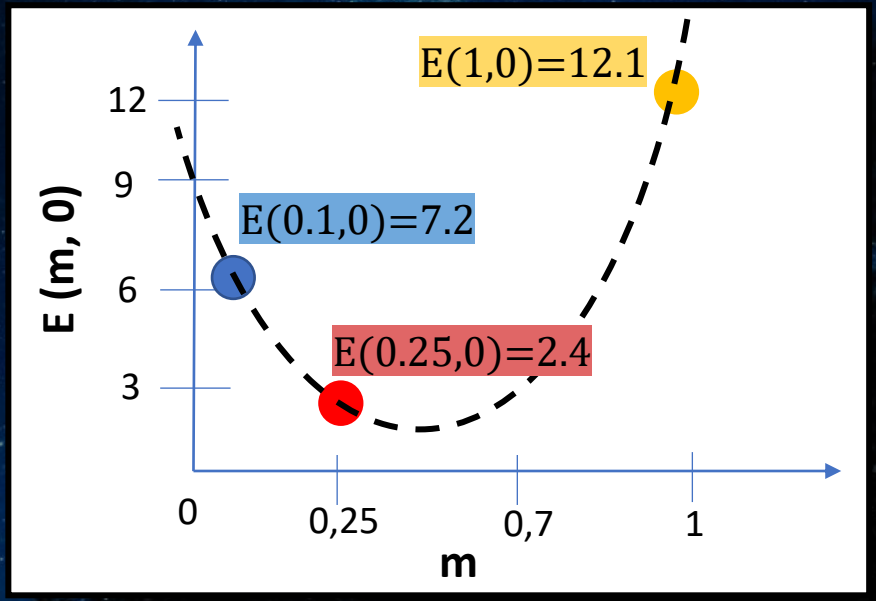
Error cuadrático medio

Asumiendo b=0

Probamos con m= 0.1 o 0.25 o 1

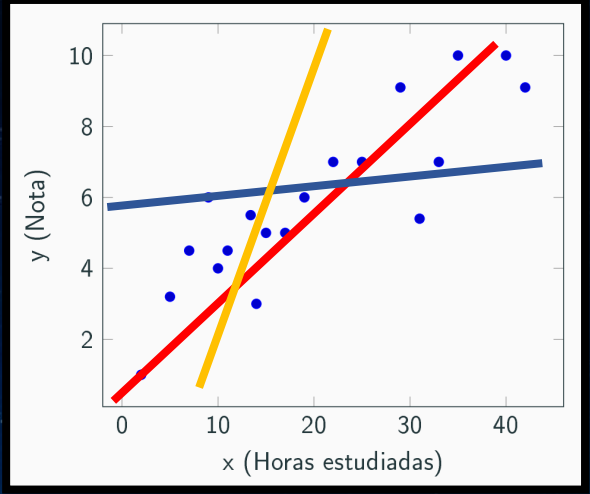
- Error en 1D.
- Parábola.

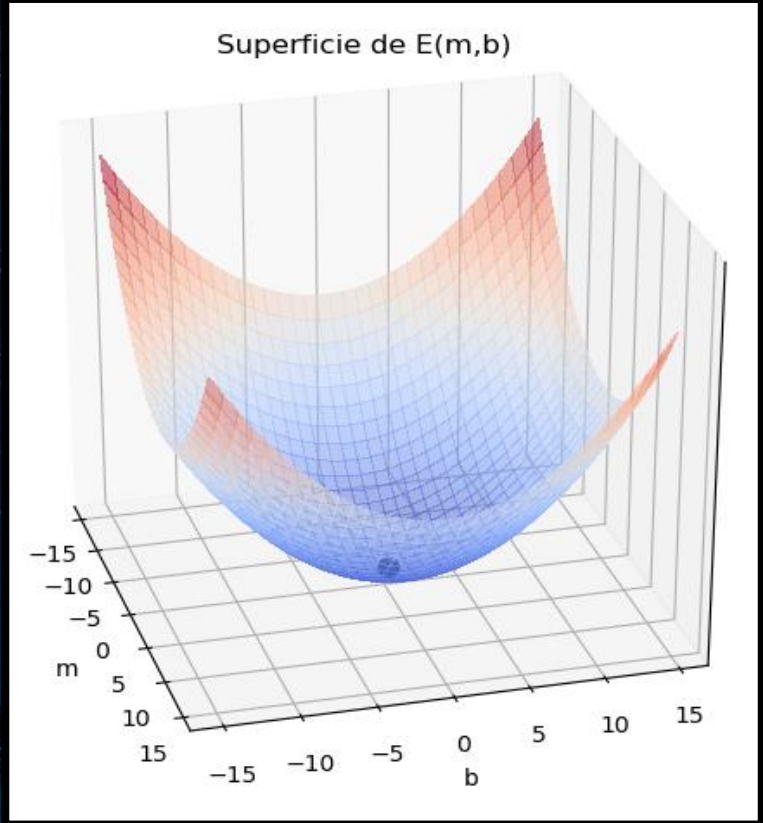




Error para m y b

- E(m,b) es un paraboloide.
- Siempre es convexa.
- Posee un solo mínimo local, que es el mínimo global.





Resumen

Regresión Lineal

- f(x)=m x + b
- Modelo más simple
- Asume relación lineal entre x e y (es aproximada)

Parámetros: m y b

Función de Error:

- Error cuadrático medio
- Promedio del Error de cada elemento.