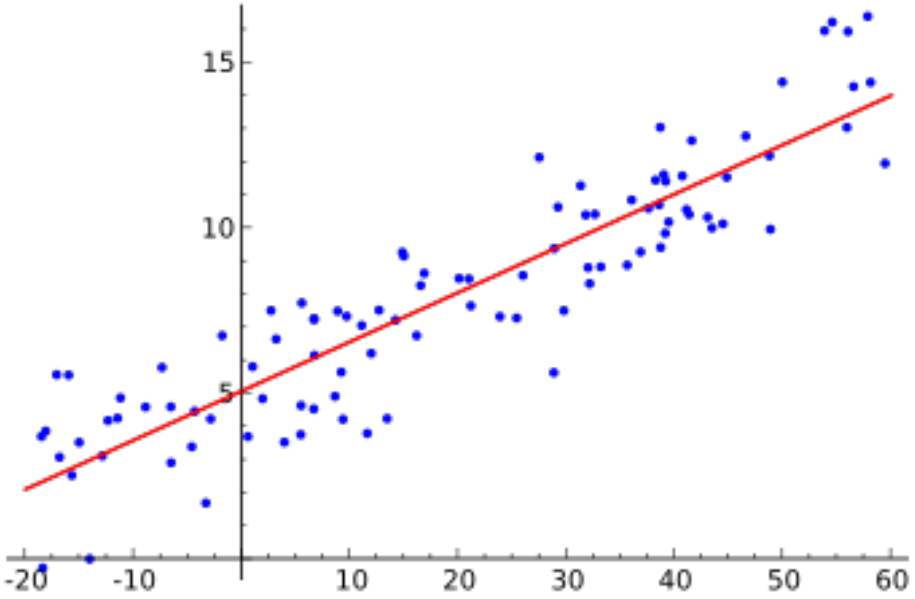
Regresión Lineal

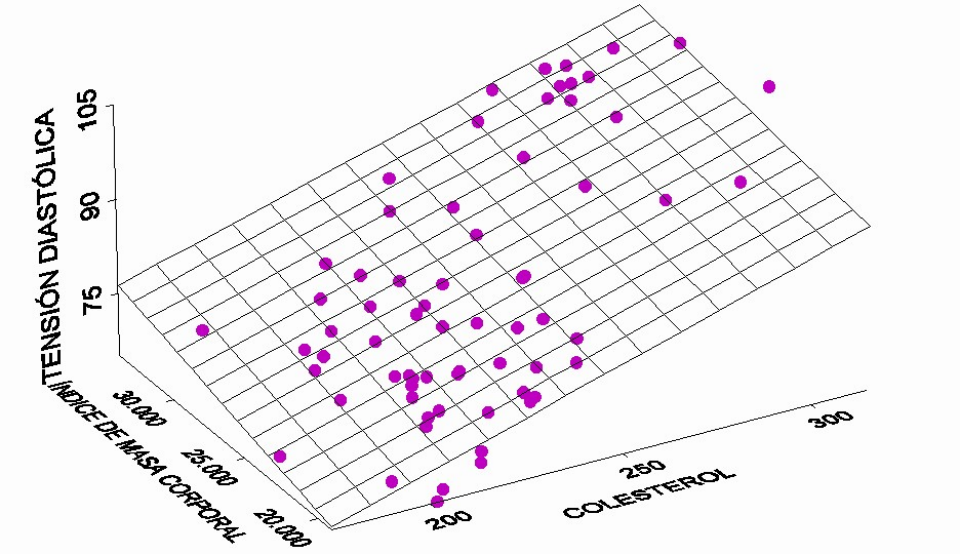
Rafael Zambrano

¿Qué es la regresión lineal?

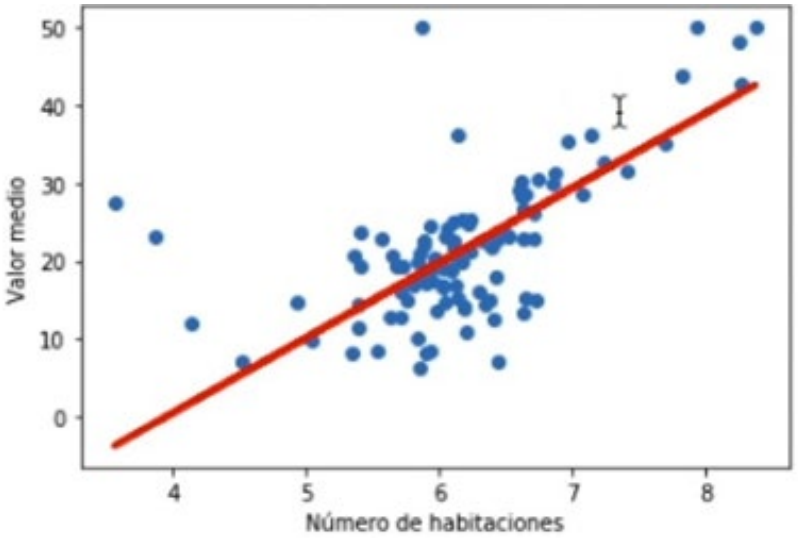
- Es un método estadístico que permite estudiar las relaciones entre variables continuas cuantitativas.
- Expresa la relación entre una variable que se llama dependiente (y) y otras independientes (X).
- ¿Por qué lineal? Porque el modelo que se genera es una línea, plano o hiperplano sin curvas.
- Uno de los métodos estadísticos de predicción más utilizados.



¿Qué es la regresión lineal?

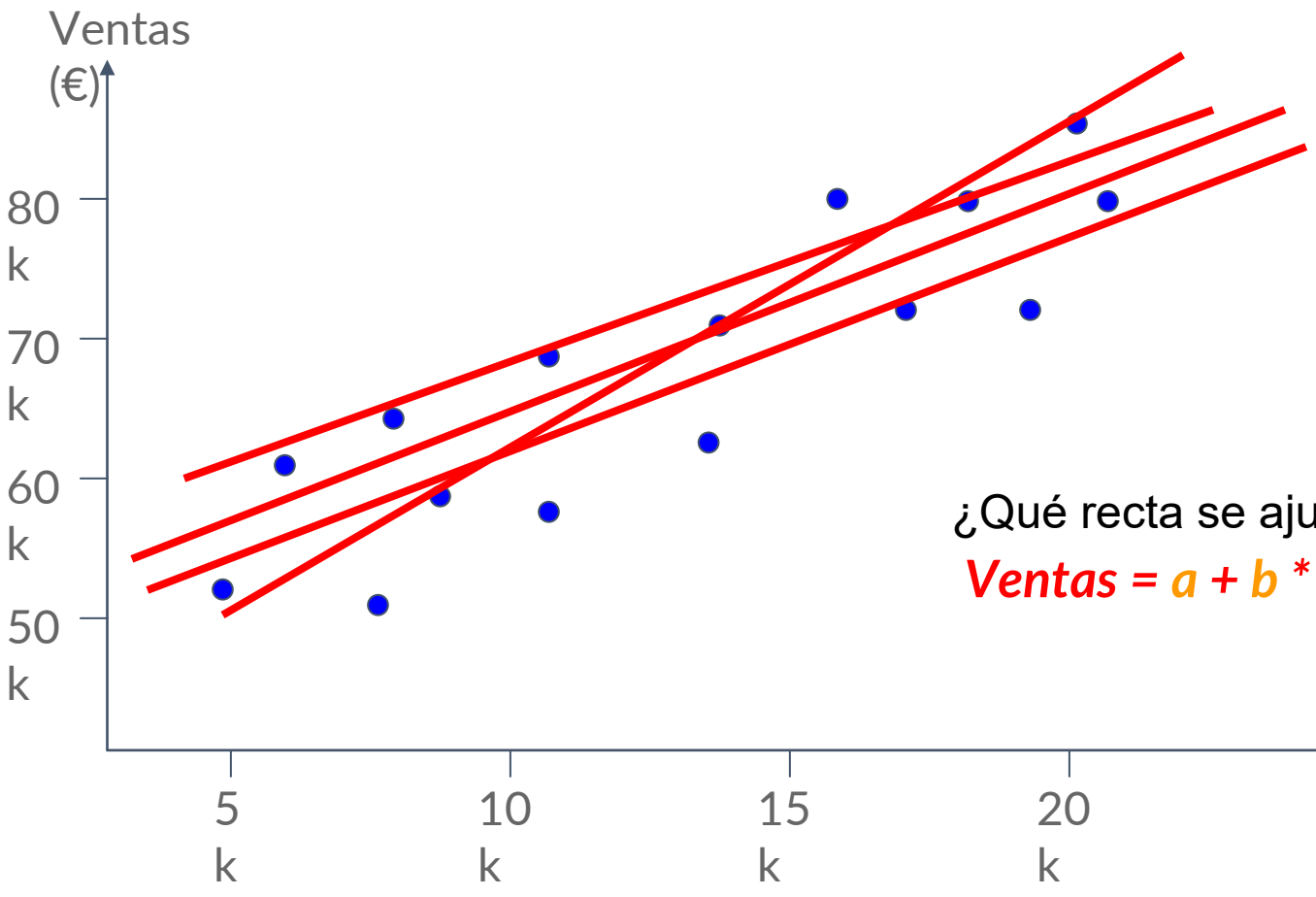


Regresión lineal múltiple



Regresión lineal simple

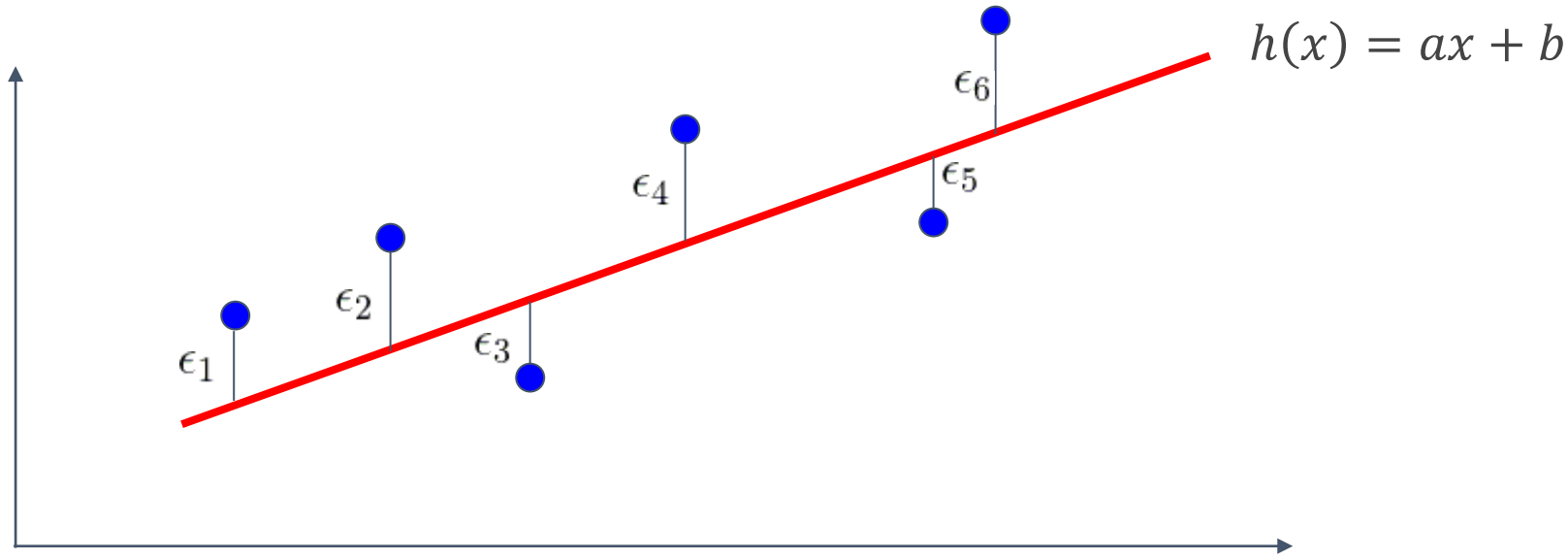
Regresión Lineal



¿Qué recta se ajusta mejor a los puntos?

Ventas = a + b * Publicidad

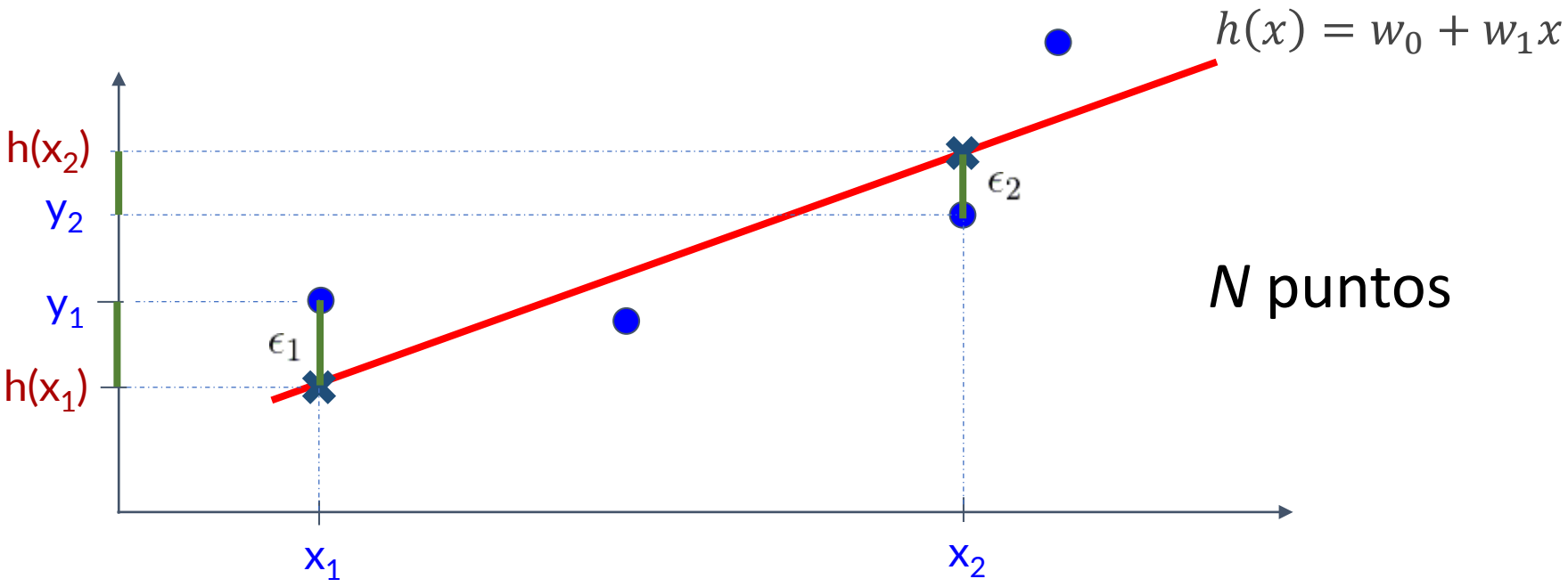
Regresión Lineal



Se busca la recta que minimice $\sum_n \epsilon_n^2 = \epsilon_1^2 + \epsilon_2^2 + \epsilon_3^2 + \epsilon_4^2 + \epsilon_5^2 + \epsilon_6^2 + \cdots + \epsilon_N^2$

Es decir, se minimiza la distancia entre la recta y todos los puntos, para obtener los coeficientes *a* y *b*

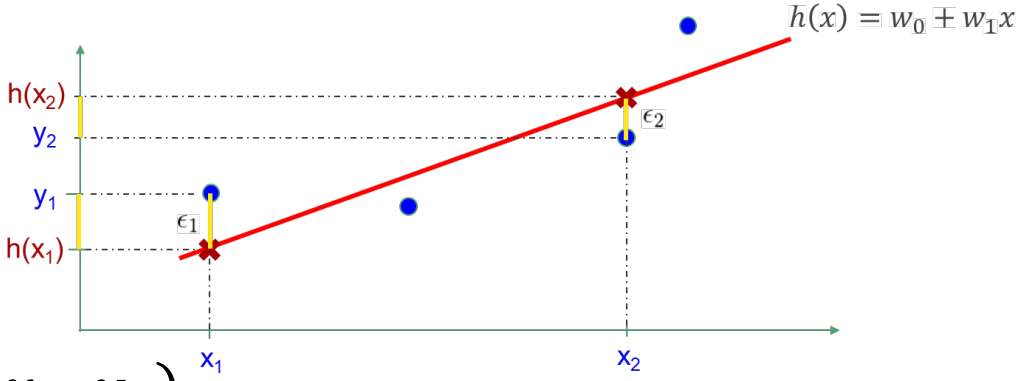
Regresión Lineal



$$\epsilon_1^2 = (h(x_1) - y_1)^2$$
$$\epsilon_2^2 = (h(x_2) - y_2)^2$$
$$\Rightarrow E = \frac{1}{N}\sum_{n=1}^{N}(h(x_n) - y_n))^2$$

Función de error
(error cuadrático medio)

Regresión Lineal



- **Datos:** $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

$$\begin{aligned} h(x_1) &= w_0 + w_1 x_1 \\ h(x_2) &= w_0 + w_1 x_2 \\ &\vdots \\ h(x_n) &= w_0 + w_1 x_n \end{aligned} \quad \Rightarrow \quad \underset{X}{\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}} \underset{w}{\begin{bmatrix} w_0 \\ w_1 \end{bmatrix}}$$

- **Errores cuadráticos:**

$$\begin{aligned} \epsilon_1^2 &= (w_0 + w_1 x_1 - y_1)^2 \\ \epsilon_2^2 &= (w_0 + w_1 x_2 - y_2)^2 \\ &\vdots \\ \epsilon_n^2 &= (w_0 + w_1 x_n - y_n)^2 \end{aligned} \quad \Rightarrow \quad |X \quad w \quad - y|^2$$

$$\left(\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} - \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \right)^2$$

Regresión Lineal

Minimizar $E_{in}(\boldsymbol{w}) = \frac{1}{N}|X\boldsymbol{w} - \boldsymbol{y}|^2$

$$\nabla E_{in}(\boldsymbol{w}) = \frac{2}{N} X^T(X\boldsymbol{w} - \boldsymbol{y}) = \boldsymbol{0}$$

$$X^T X\boldsymbol{w} - X^T\boldsymbol{y} = 0$$

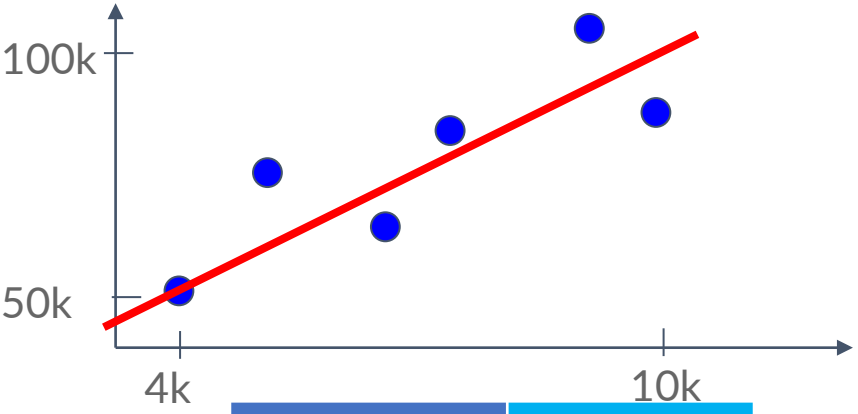
$$X^T X\boldsymbol{w} = X^T\boldsymbol{y}$$

$$\left(X^T X\right)^{-1} X^T X\boldsymbol{w} = \left(X^T X\right)^{-1} X^T\boldsymbol{y}$$

$$I \cdot \boldsymbol{w} = \left(X^T X\right)^{-1} X^T\boldsymbol{y}$$

$$\Rightarrow \boxed{\boldsymbol{w} = \left(X^T X\right)^{-1} X^T\boldsymbol{y}}$$

Regresión Lineal



Publicidad x_1	Ventas y
4	50
5	75
6	60
7	80
9	110
10	85

$$\boldsymbol{w} = \left(X^T X\right)^{-1} X^T \boldsymbol{y}$$

$$\mathrm{X} = \begin{bmatrix} 1 & 4 \\ 1 & 5 \\ 1 & 6 \\ 1 & 7 \\ 1 & 9 \\ 1 & 10 \end{bmatrix}, \mathrm{y} = \begin{bmatrix} 50 \\ 75 \\ 60 \\ 80 \\ 110 \\ 85 \end{bmatrix}$$

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 4 & 5 & 6 & 7 & 9 & 10 \end{bmatrix}$$

$$\Rightarrow \mathbf{w} = \left(X^T X\right)^{-1} X^T \boldsymbol{y} = \begin{bmatrix} 27.85 \\ 7.14 \end{bmatrix}$$

Ventas = 27.85 + 7.14 * Publicidad

Evaluación de modelos de regresión

- Error absoluto medio MAE (*Mean Absolute Error*)

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

Real	Predicción
32	35
23	28
85	95

$$MAE = \frac{1}{3}(|32-35| + |23-28| + |85-97| = \frac{1}{3}(3+5+10) = 6$$

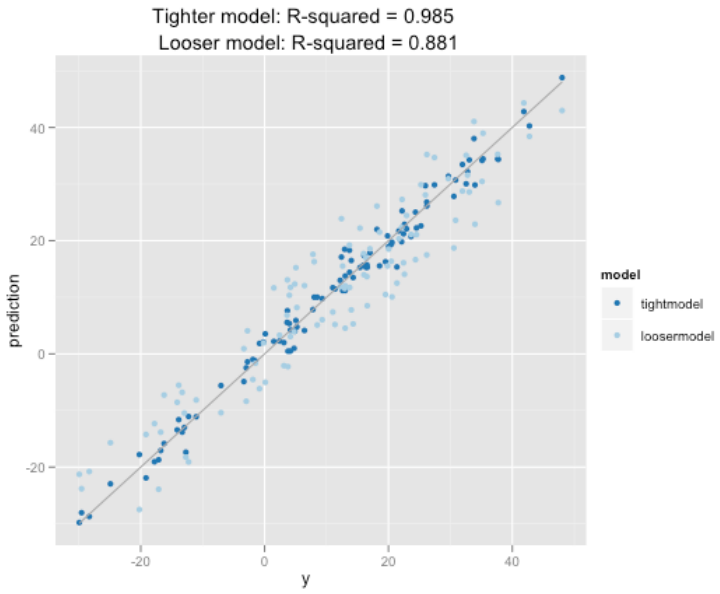
- Error cuadrático medio RMSE (*Root Mean Square Error*): Tiene la ventaja de penalizar errores grandes

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

$$RMSE = \sqrt{\frac{1}{3}(3^2 + 5^2 + 10^2)} = 6.68$$

Evaluación de modelos de regresión

- R^2 (coeficiente de determinación): explica cómo de bien el modelo explica la variación en los datos (valor entre 0 y 1)



https://youtu.be/w2FKXOa0HGA

Interpretación del modelo

Si, por ejemplo, tenemos los siguientes coeficientes para predecir el precio de unas casas

	Coefficient
Avg. Area Income	20.920136
Avg. Area House Age	158094.410454
Avg. Area Number of Rooms	123512.191322
Avg. Area Number of Bedrooms	-3031.996047
Area Population	15.971469

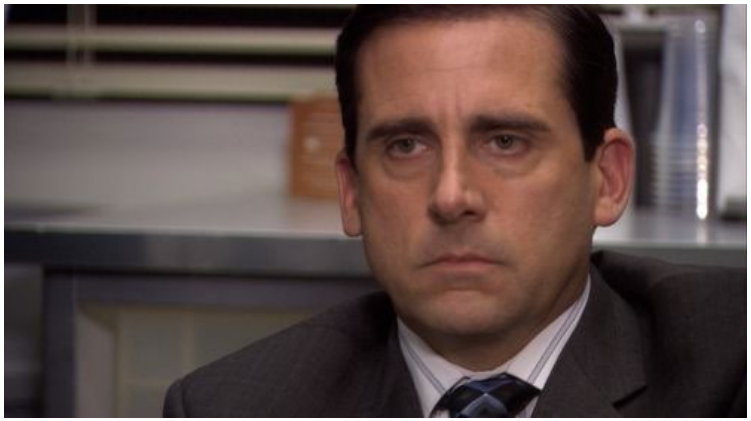
Precio casas = 20.9 * (Avg. Area Income) + 158094.41 * (Avg. Area House Age) + …

¿Cómo se interpreta esto? Por cada unidad de *Avg. Area Income*, aumenta 20.9 el precio

Interpretación del modelo

Vale, entonces cuanto más alto es el coeficiente, mayor es la importancia de la variable...

	Coefficient
Avg. Area Income	20.920136
Avg. Area House Age	158094.410454
Avg. Area Number of Rooms	123512.191322
Avg. Area Number of Bedrooms	-3031.996047
Area Population	15.971469



NO! Estamos comparando unidades diferentes. ¿El numero de habitaciones es menos importante que la edad de la casa?

¿Solución? Estandarizar los datos

Multicolinearidad

Hay que **evitar** que las variables en un modelo de regresión lineal múltiple estén **altamente correlacionadas**

Ejemplo: queremos predecir el salario de empleados usando dos variables: *"Job Level"* y *"Working Years"*

$$\text{Salary} = \mathbf{a} * \text{Job Level} + \mathbf{b} * \text{Working Years} + \mathbf{c}$$

- Si ambas variables tienen una relación lineal, por ejemplo

$$\text{Job Level} = \mathbf{0.2} * \text{Working Years} + \mathbf{1}$$

estas tres ecuaciones generarían el mismo resultado:

$$\text{Salary} = \mathbf{1000} * \text{Job Level} + \mathbf{0} * \text{Working Years} + \mathbf{1000}$$

$$\text{Salary} = \mathbf{500} * \text{Job Level} + \mathbf{100} * \text{Working Years} + \mathbf{1500}$$

$$\text{Salary} = \mathbf{0} * \text{Job Level} + \mathbf{200} * \text{Working Years} + \mathbf{2000}$$

Multicolinearidad

- Por ejemplo, si *Working Years = 5 → Job Level = 0.2 * 5 + 1 = 2*

Salary = **1000** * Job Level + **0** * Working Years + **1000** *= 1000 * 2 + 0 * 5 + 1000 = 3000*

Salary = **500** * Job Level + **100** * Working Years + **1500** *= 500 * 2 + 100 * 5 + 1500 = 3000*

Salary = **0** * Job Level + **200** * Working Years + **2000** *= 0 * 2 + 200 * 5 +2000 = 3000*

- ¿Un incremento de 1 año en Working Years incrementa tu salario en 200€? ¿100€? ¿0€?
- Esto supone un problema, ya que los coeficientes de las variables no son confiables
- Es recomendable eliminar variables que tengan alta correlación con otra (por ejemplo superior a 0.9)

Resumen

- El objetivo de la regresión lineal es encontrar la relación lineal entre todas las variables del problema
- Se genera un error global que es la distancia entre todos los datos y nuestro modelo (línea, plano, hiperplano).
- El valor añadido es poder predecir valores inexistentes.
- Tiene ciertas limitaciones. Por ejemplo, con datos no lineales.