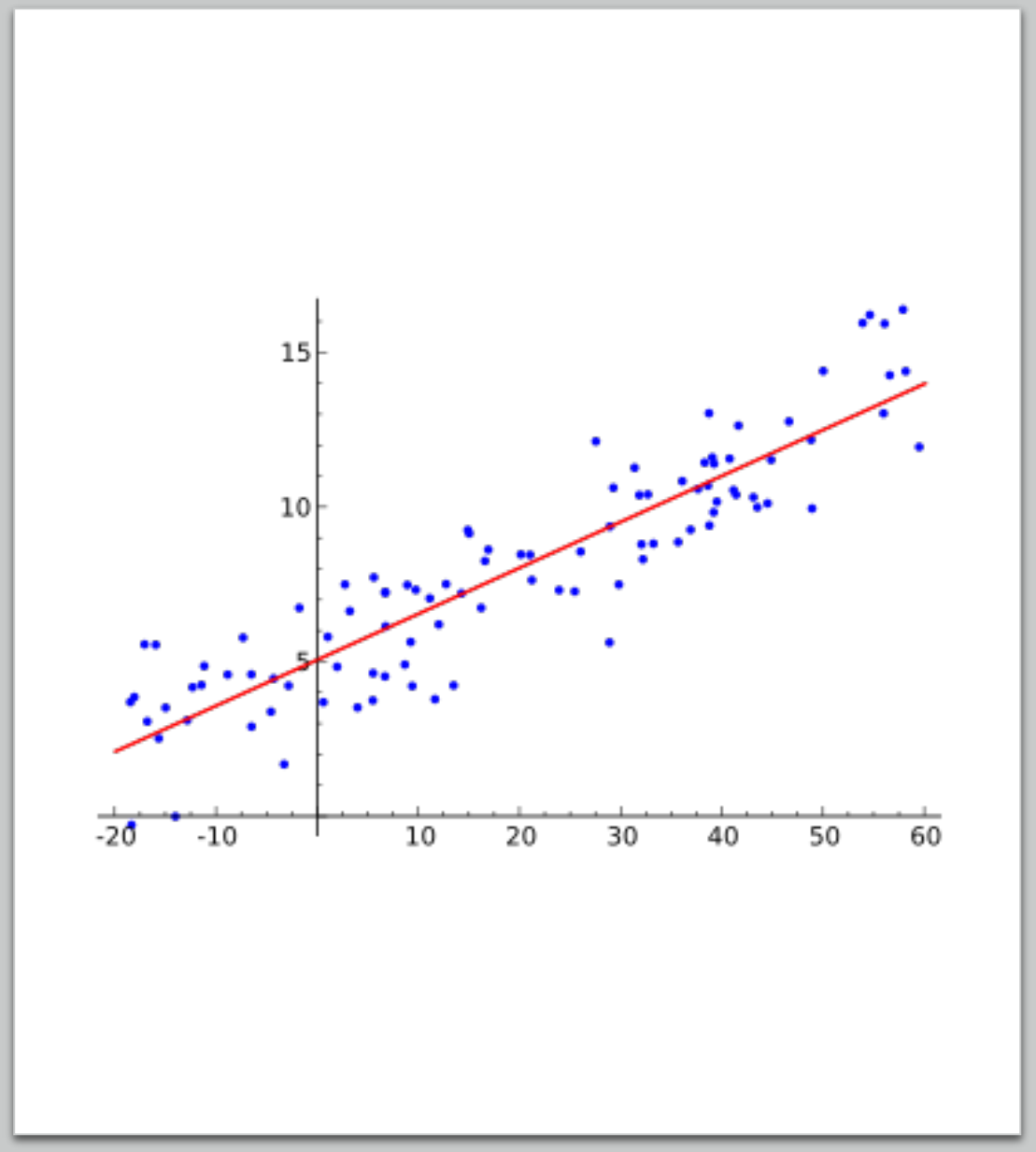
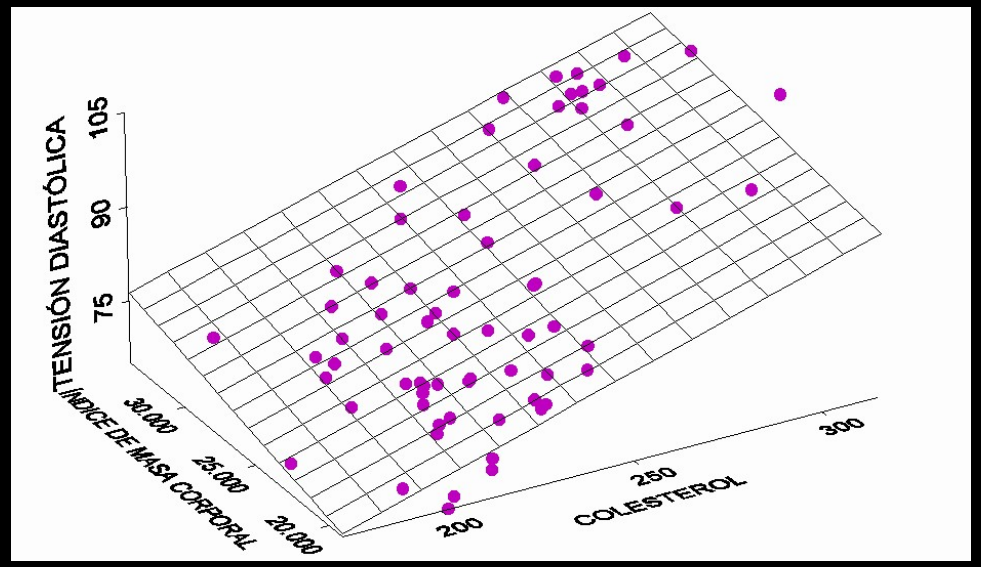
Machine Learning
Regresión Lineal

¿Qué es la regresión lineal?

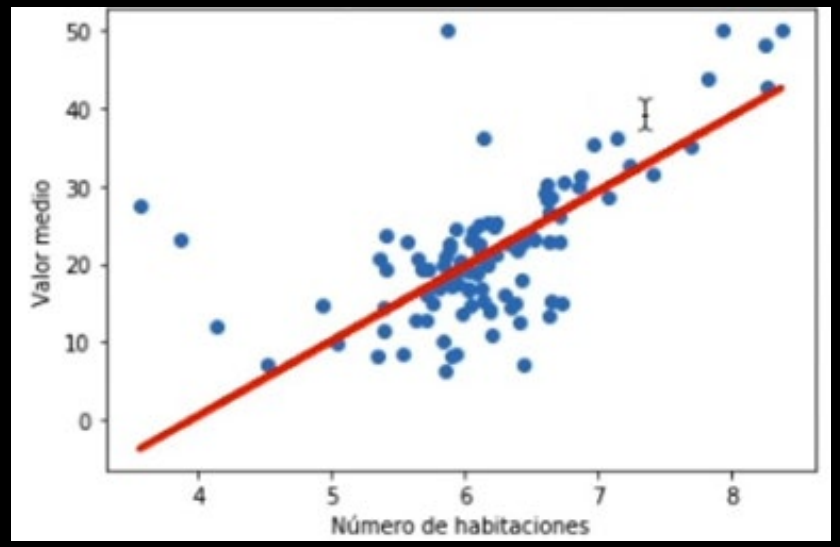
- Es un *método estadístico* que permite estudiar las relaciones entre dos variables contínuas cuantitativas.
- ¿Por qué *regresión*? Porque expresa la relación entre una variable que se llama regresando (y, dependiente) y otra que se llama regresor (x, independiente).
- ¿Por qué *lineal*? Porque el modelo que se genera es una línea, plano o hiperplano sin curvas.
- Es una *técnica paramétrica* porque hace varias suposiciones sobre el conjunto de datos.
- Uno de los métodos estadísticos de predicción más utilizados.



Tipos de regresión lineal



Regresión lineal múltiple



Regresión lineal simple

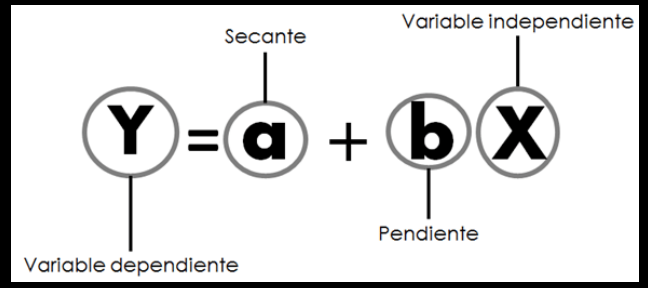
Regresión lineal múltiple

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

- β_0 es el término independiente. Es el valor esperado de Y cuando $X_1, \ldots, X_p$ son cero.
- $\beta_1, \beta_2, \ldots \beta_p$ son los coeficientes parciales de la regresión:
  - β_1 mide el cambio en Y por cada cambio unitario en X_1, manteniendo $X_2, X_3, \ldots, X_p$ constantes.
  - β_2 mide el cambio en Y por cada cambio unitario en X_2, manteniendo $X_1, X_3, \ldots, X_p$ constantes.
  - ...
  - β_p mide el cambio en Y por cada cambio unitario en X_p, manteniendo $X_1, \ldots, X_{p-1}$ constantes.
- ε es el error de observación debido a variables no controladas.

Regresión lineal simple

$$Y = \beta_0 + \beta X + \varepsilon$$

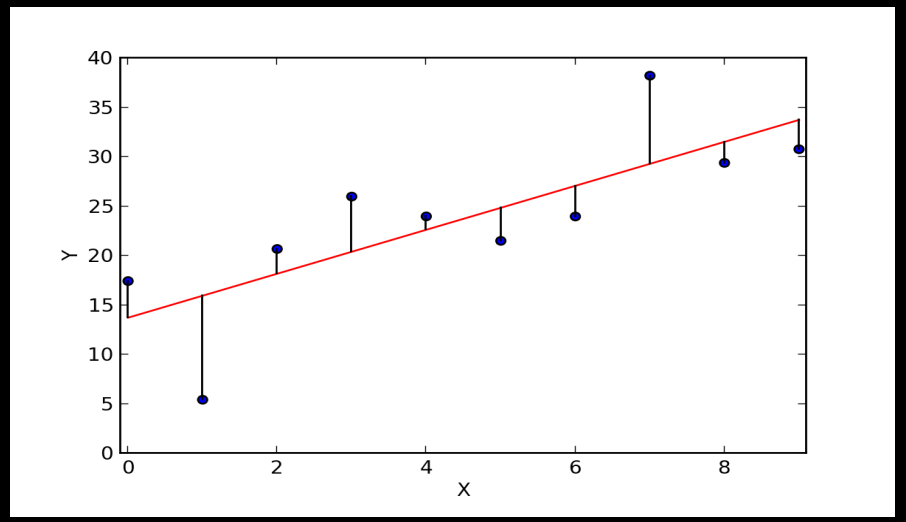


¿Cómo medimos el error cometido?

Loss function

Necesitamos una métrica que nos diga cómo de bien o mal predice el modelo: **error cuadrático medio (Mean Squared Error)**

$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y - \hat{y}\right)^2}_{\text{The square of the difference between actual and predicted}}$$

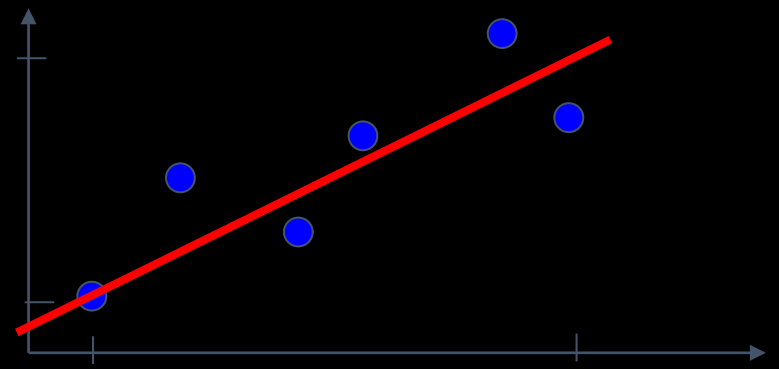


Que sería la **loss function** o función de costes de la regresión lineal

Perfecto, definida nuestra métrica de calidad del modelo, ¿ahora qué viene?
Que nuestra regresión tenga la mínima cantidad de errores, ¿cómo lo hago?
Hay que encontrar aquellos Ws que me minimicen la función de costes

¿Cómo consigo los coeficientes que me minimizan el error?

Mínimos cuadrados (ecuación normal)



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

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

x	y
4	50
5	75
6	60
7	80
9	110
10	85

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

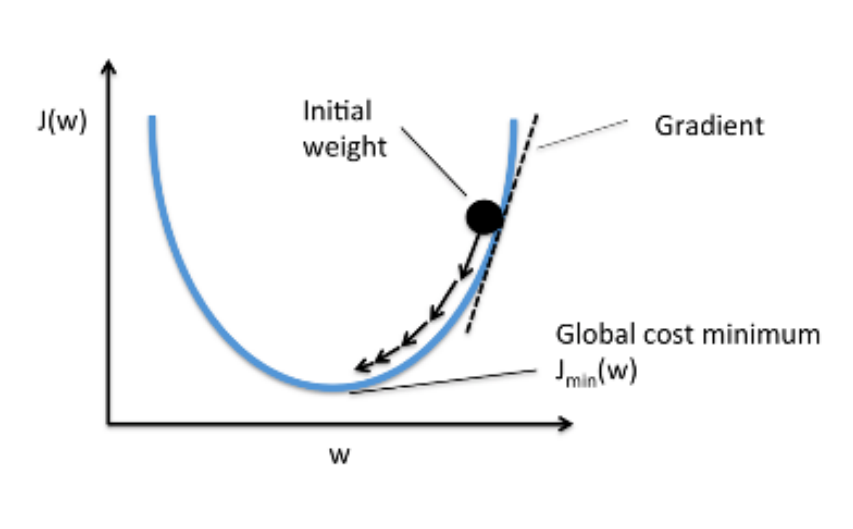
$$= \begin{bmatrix} 27.85 \\ 7.14 \end{bmatrix}$$

y = 27.85 + 7.14 x

Demo

Gradient Descent

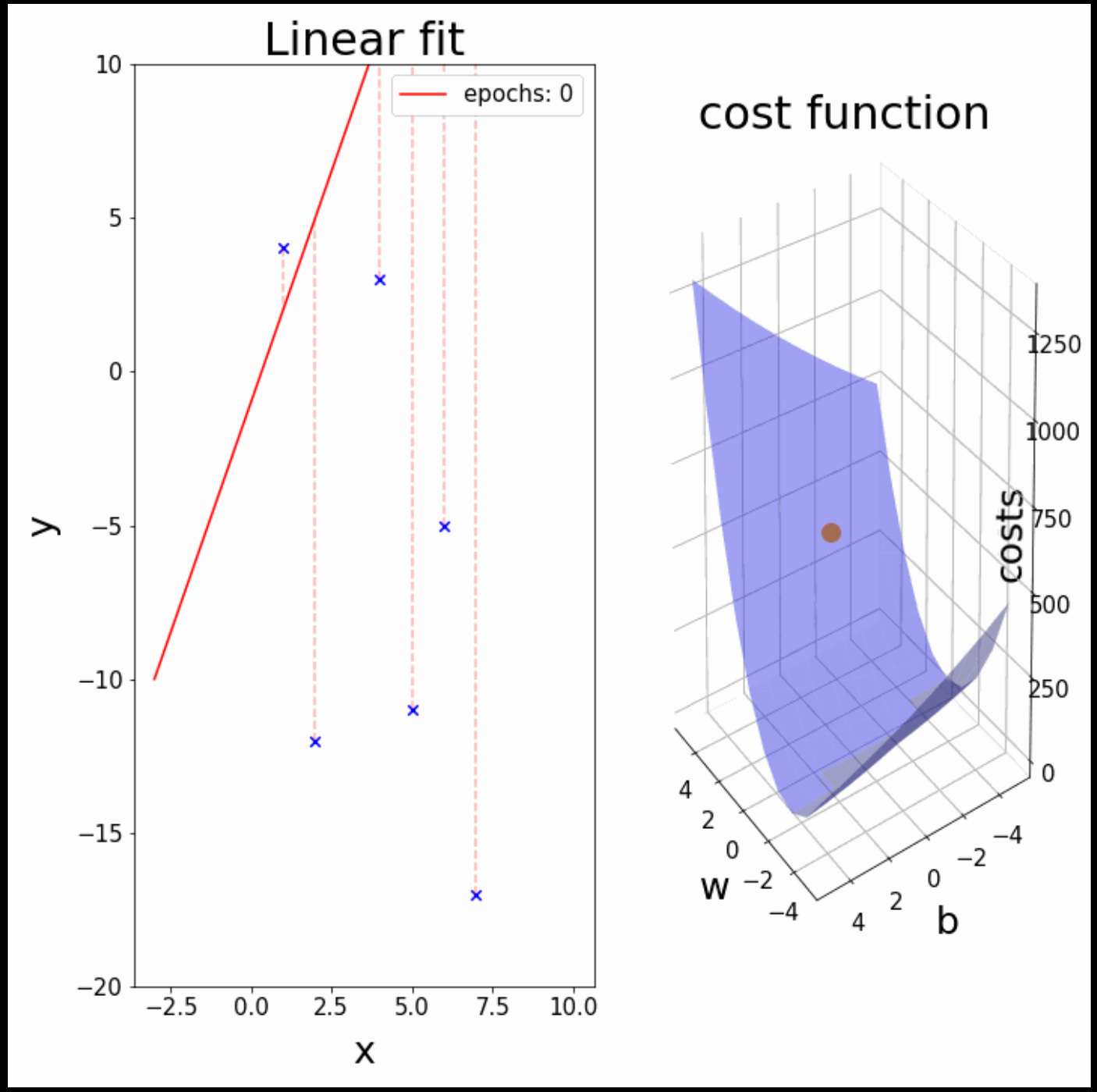
Problema de optimización matemática. El Gradient Descent es uno de los métodos más utilizados en algoritmos de aprendizaje supervisado.
¿Cuáles son los pesos W, que dan mejores resultados? Los que minimizan la función de coste



Hypothesis:	$h_\theta(x) = \theta_0 + \theta_1 x$
Parameters:	θ_0, θ_1
Cost Function:	$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$
Goal:	$\underset{\theta_0, \theta_1}{\text{minimize}}\ J(\theta_0, \theta_1)$

[Demo]

Gradient Descent



¿Cómo interpreto los pesos (w)?

Coeficientes

Si queremos predecir el precio de casas de un DF, podríamos obtener los siguientes coeficientes:

	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

E interpretaríamos la regresión lineal como:

y = w1*x1 + w2*x2 + w3*x3 + w4*x4 + w5*x5

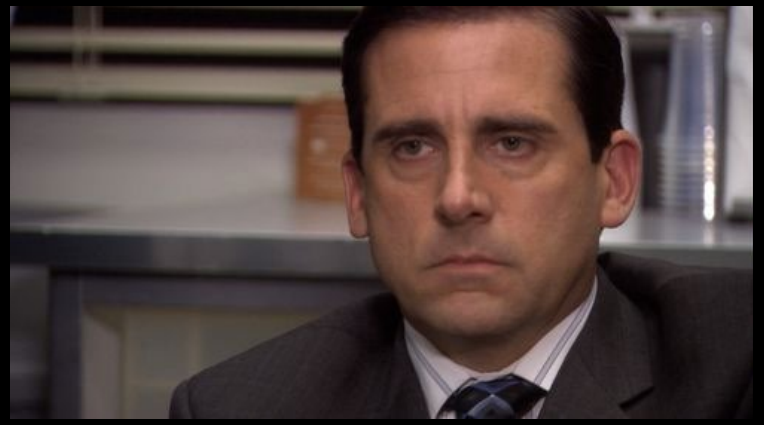
Precio casas = 20.9 * (Avg. Area Income) + 158094.41 * (Avg. Area House Age) + …

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

Feature importance

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

R-Squared

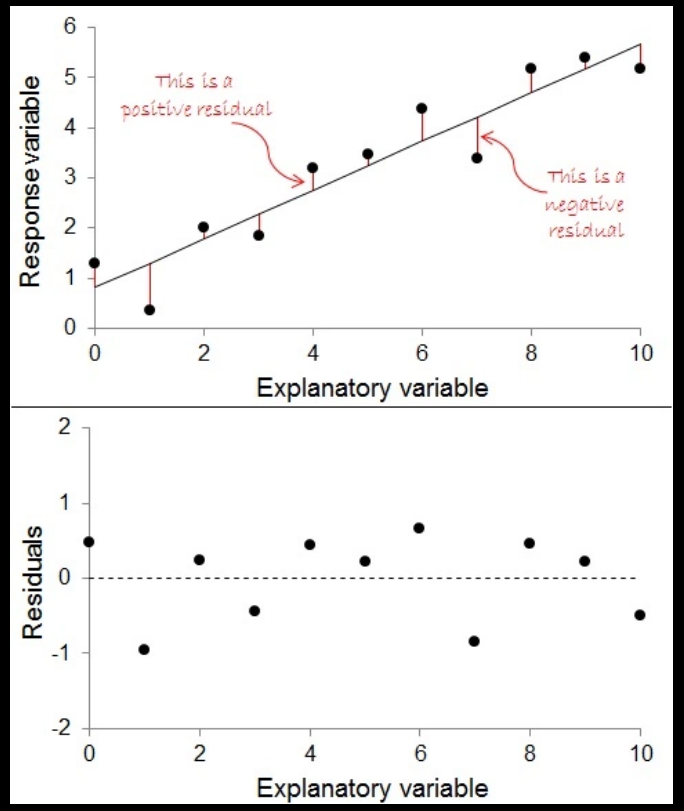
Coeficiente de determinación. Mide cuánto de bien una regresión se ajusta a los datos. También se define como la porción de variación de la variable dependiente (y) predecible mediante la independiente (x). Va de [0,1]. Cuanto mejor se ajuste, más se acercará a 1. Cuanto más cercano a 0, menos fiable será.

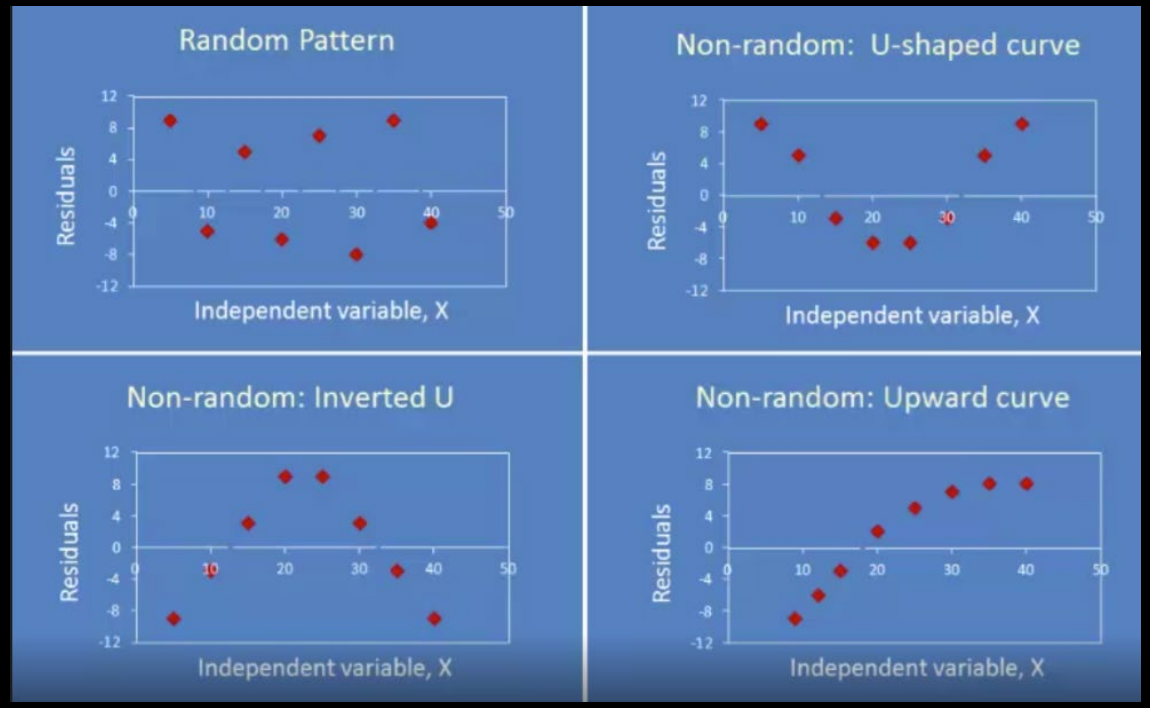
How to know if the model is best fit for your data?

The most common metrics to look at while selecting the model are:

STATISTIC	CRITERION
R-Squared	Higher the better *(> 0.70)*
Adj R-Squared	Higher the better
F-Statistic	Higher the better
Std. Error	Closer to zero the better
t-statistic	Should be greater 1.96 for p-value to be less than 0.05
AIC	Lower the better
BIC	Lower the better
Mallows cp	Should be close to the number of predictors in model
MAPE (Mean absolute percentage error)	Lower the better
MSE (Mean squared error)	Lower the better
Min_Max Accuracy => mean(min(actual, predicted)/max(actual, predicted))	Higher the better

Residuos





Si el modelo predice bien, los residuos siguen una distribución normal

Evaluación del modelo

Regresión lineal: errores

- Mean Absolute Error (MAE)

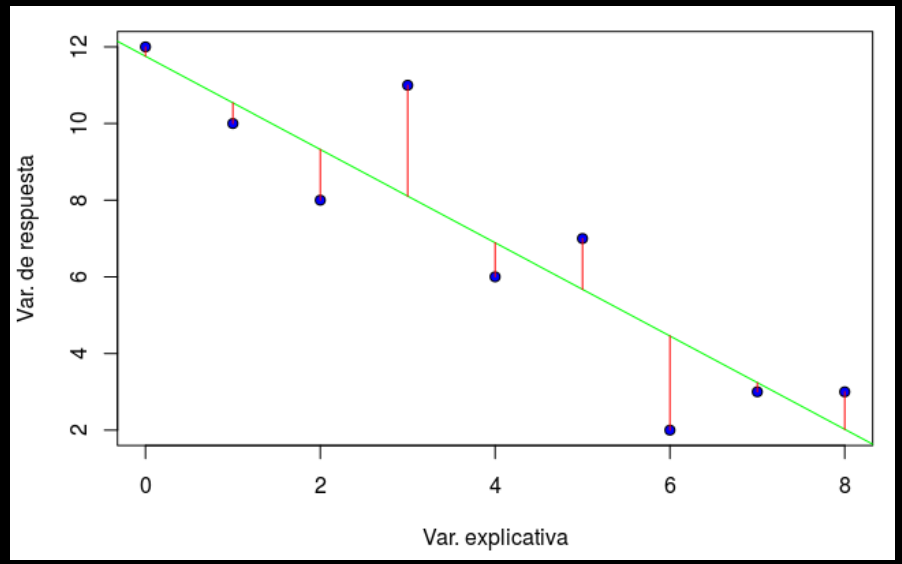
$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

- Mean Squared Error (MSE)

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.$$

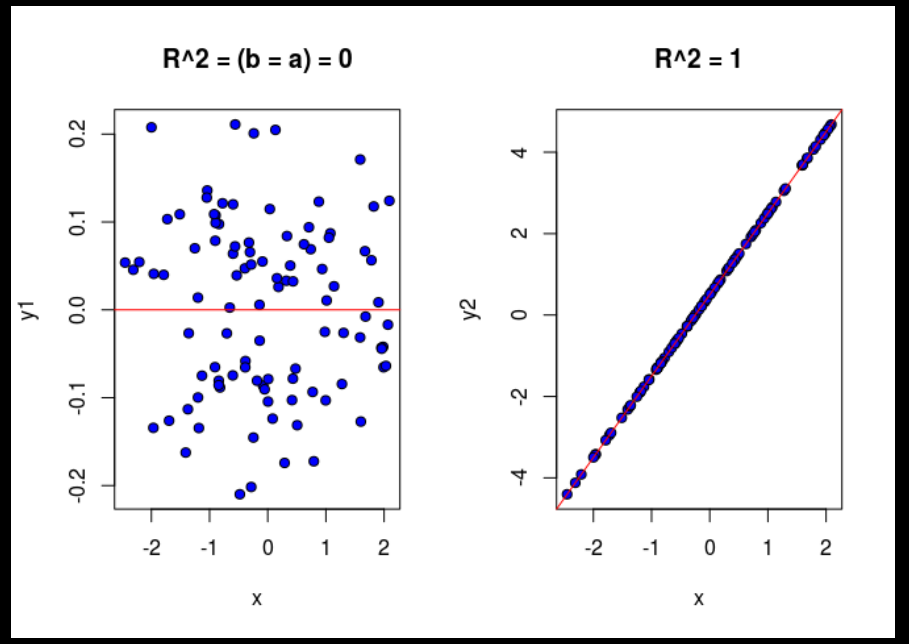
- Root Mean Squared Error (RMSE)

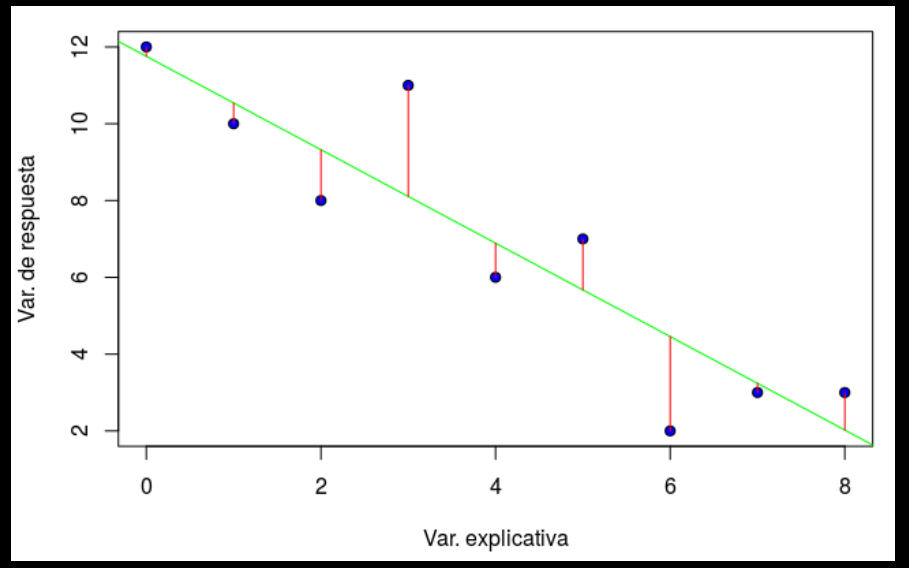
$$\mathrm{RMSD} = \sqrt{\frac{\sum_{t=1}^{T} (\hat{y}_t - y_t)^2}{T}}.$$



Regresión lineal: meta

- *Objetivo*: encontrar la relación lineal entre todas las variables del problema. Encontrar ‘a’ y ‘b’.
- El valor añadido es poder predecir valores inexistentes.
- Tiene ciertas limitaciones. Un ejemplo, datos no lineales.
- Se genera un error global que es la distancia entre todos los datos y nuestro modelo (línea, plano, hiperplano).





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"*

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

- Si ambas variables tienen una relación lineal, por ejemplo $\text{Job Level} = 0.2 * \text{Working Years} + 1$ estas tres ecuaciones generarían el mismo resultado:

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

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

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

Job Level = 0.2 * Working Years + 1

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)

Preguntas