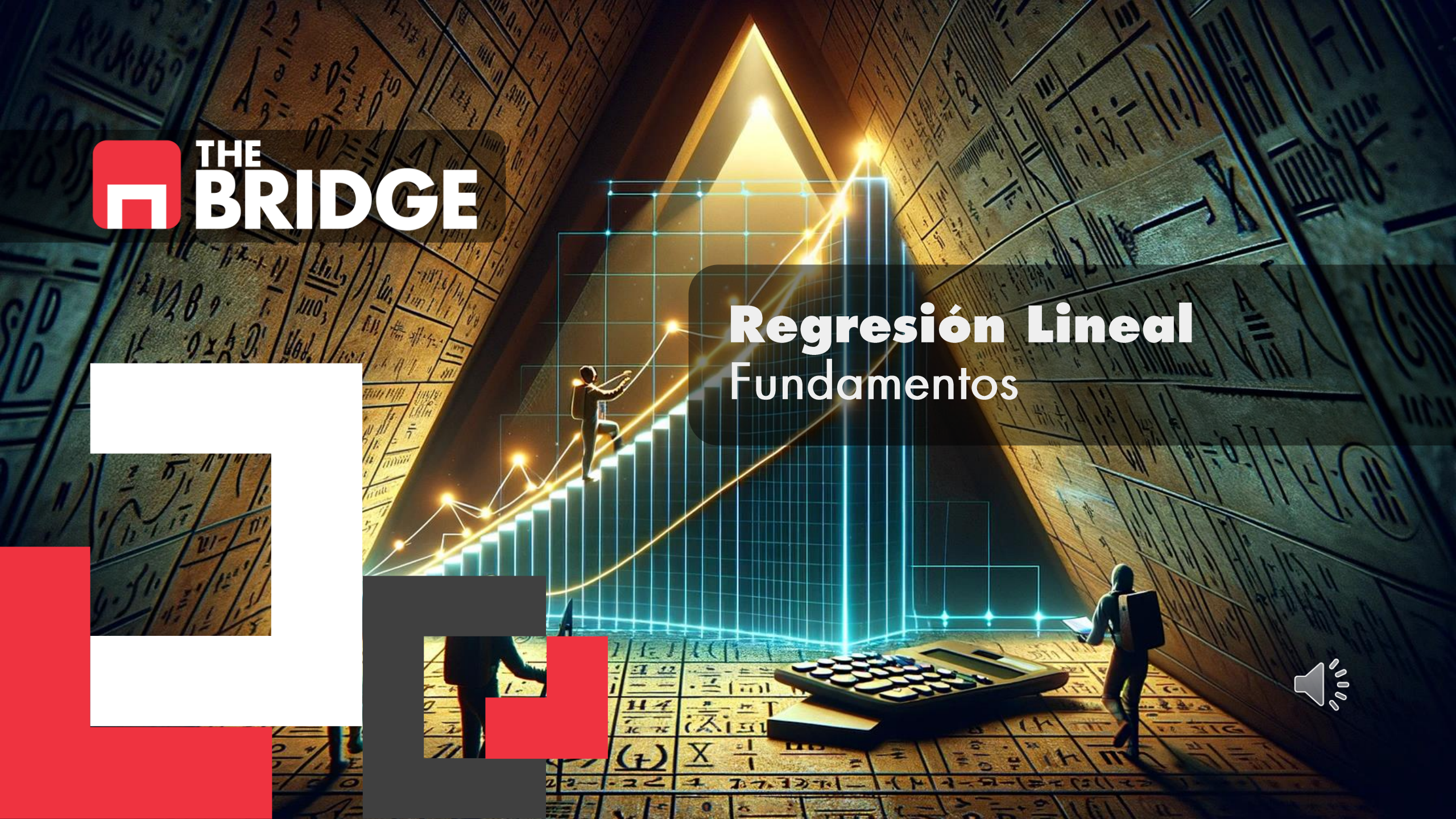
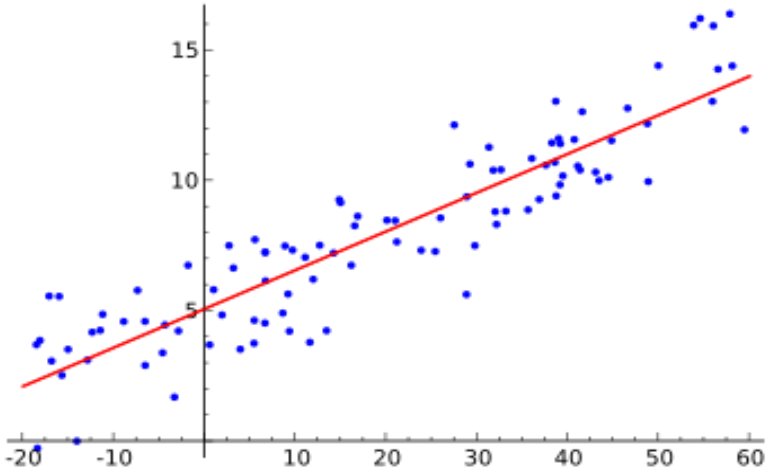
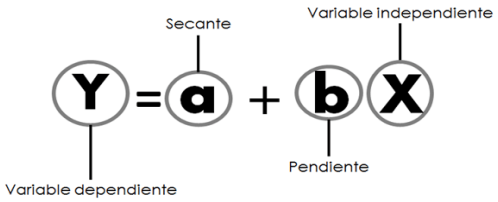

THE BRIDGE
Regresión Lineal
Fundamentos

Concepto

Método estadístico que modela la relación entre una variable continua y una o más variables independientes



1. Estimación de las ventas a partir del gasto en marketing
2. Estimación del consumo de gasolina en función de la distancia.
3. Predicción de precios de casas en función de los metros cuadrados (entre otras variables)



$Y = 5 + 6X$

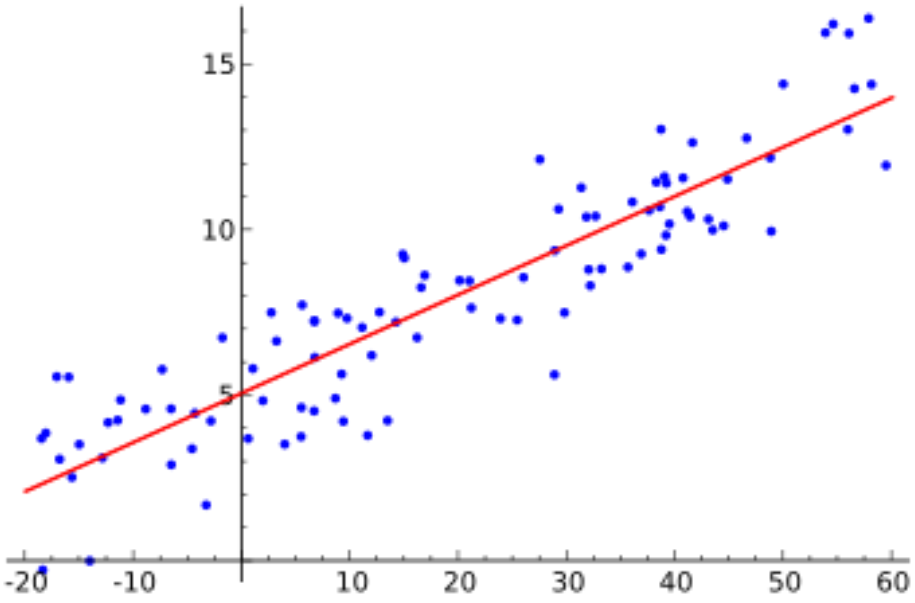
Concepto (cont.)

¿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*? Los parámetros de la ecuación se incorporan de forma lineal

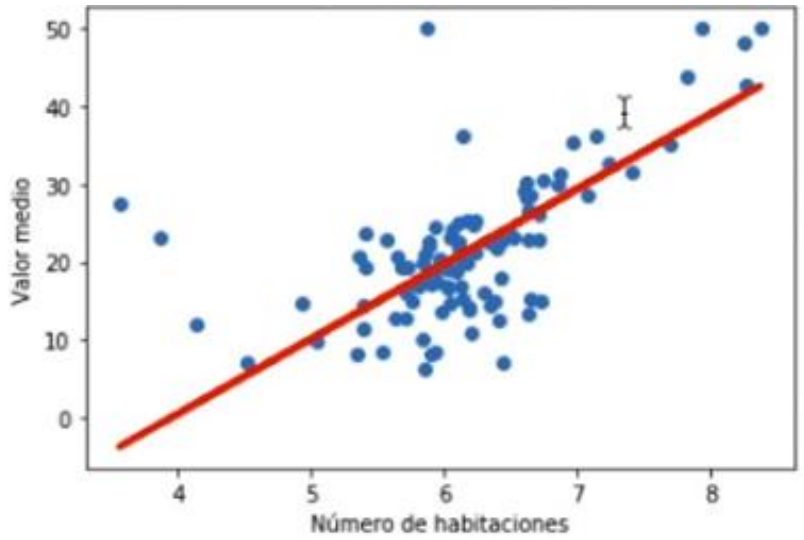
Es una *técnica paramétrica* porque hace varias suposiciones sobre el conjunto de datos.

Uno de los métodos estadísticos de predicción más utilizados.



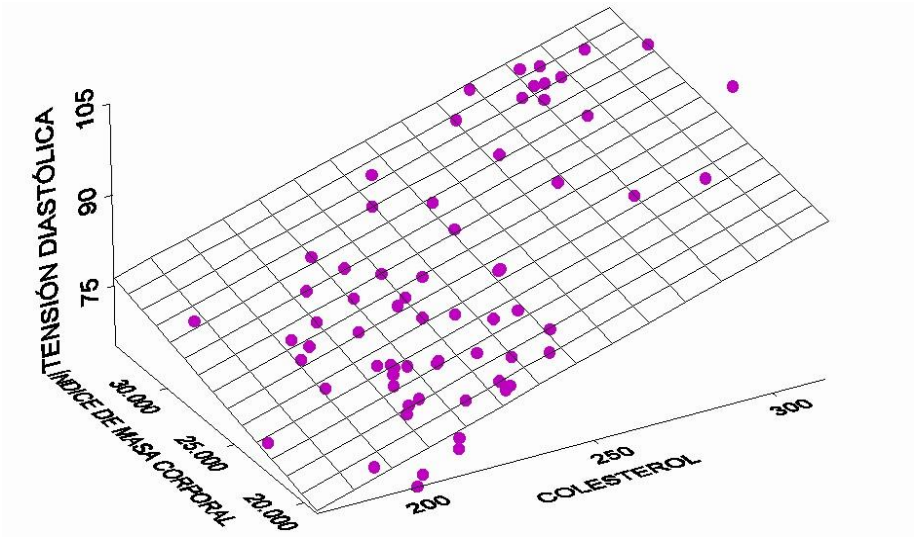
Tipos de regresión lineal

TV	radio	newspaper	sales
230.1	37.8	69.2	22100.0
44.5	39.3	45.1	10400.0
17.2	45.9	69.3	9300.0
151.5	41.3	58.5	18500.0



Regresión lineal simple

$$Y = \beta_0 + \beta_1 X_1$$



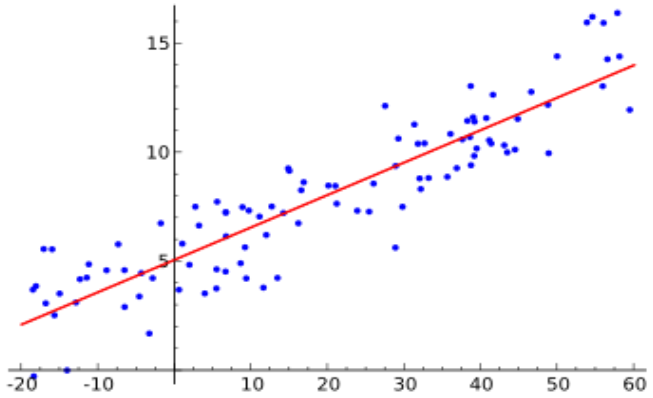
Regresión lineal múltiple

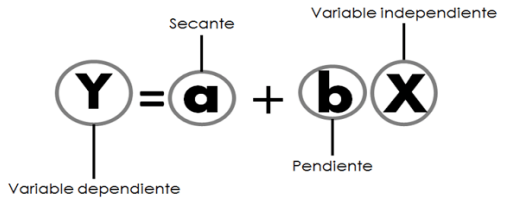
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots. + \beta_j X_j$$

Parámetros: ¿Qué son los coeficientes?

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots. + \beta_j X_j$$

1. β_0: valor de variable respuesta para cuando todos los predictores son 0
2. β_j: cuánto aumenta la variable respuesta cuando el predictor j incrementa en una unidad





Y = 5 + 6X

Parámetros: ¿Qué son los coeficientes?

- Los coeficientes determinan el peso de una feature en la estimación del target
- Por eso muchas veces utilizaremos el nombre de "peso" para refererirnos a los coeficientes y a algunos parámetros de los modelos

TV	radio	newspaper	sales
230.1	37.8	69.2	22100.0
44.5	39.3	45.1	10400.0
17.2	45.9	69.3	9300.0
151.5	41.3	58.5	18500.0

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

$$Sales = \beta_0 + \beta_1 * (TV) + \beta_2 * (radio) + \beta_3 * (newspaper)$$

$$Sales = 500 + 50* (TV) + 20* (radio) + 15* (newspaper)$$

Parámetros: ¿Cómo se interpretan?

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

	Coefficient
Avg. Area Income	21.625799
Avg. Area House Age	165590.392746
Avg. Area Number of Rooms	119827.783390
Avg. Area Number of Bedrooms	2361.095262
Area Population	15.216581

E interpretaríamos la regresión lineal como:

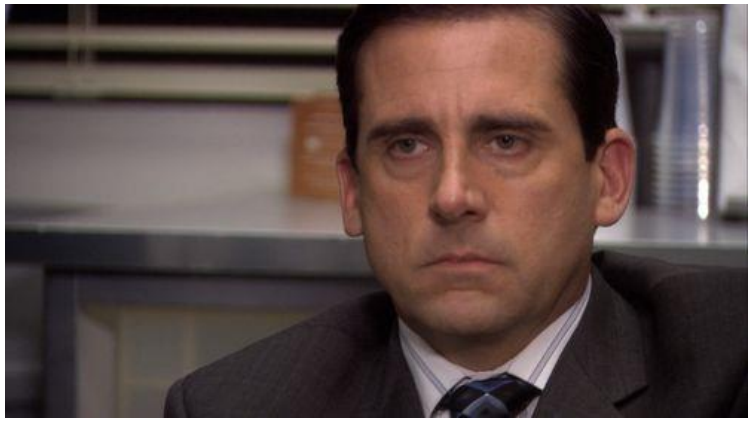
$y = w1*x1 + w2*x2 + w3*x3 + w4*x4 + w5*x5$

Precio casas = 21.6 * (Avg. Area Income) + 165590.4 * (Avg. Area House Age).....

¿Cómo se interpreta esto? Por cada unidad de *Avg. Area Income*, aumenta 21.6 el precio

Feature importance

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



	Coefficient
Avg. Area Income	21.625799
Avg. Area House Age	165590.392746
Avg. Area Number of Rooms	119827.783390
Avg. Area Number of Bedrooms	2361.095262
Area Population	15.216581

	coefficient
Avg. Area House Age	165590.392746
Avg. Area Number of Rooms	119827.783390
Avg. Area Number of Bedrooms	2361.095262
Avg. Area Income	21.625799
Area Population	15.216581

NO! Estamos comparando unidades diferentes. ¿La edad de la casa es más importante que la media de ingresos de la zona?

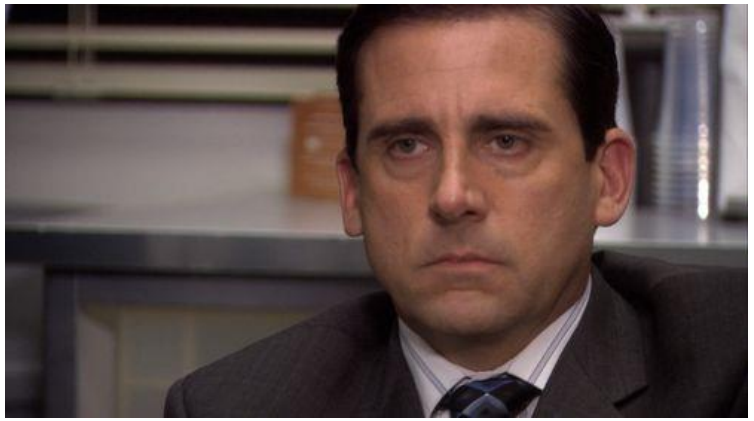
Precio ($) = Edad(años) + Habitaciones(n° habitaciones)...

¿Solución? Estandarizar los datos

$$z = \frac{x - \mu}{\sigma}$$

Feature importance

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



	Coefficient
Avg. Area Income	21.625799
Avg. Area House Age	165590.392746
Avg. Area Number of Rooms	119827.783390
Avg. Area Number of Bedrooms	2361.095262
Area Population	15.216581

	coefficient
Avg. Area House Age	165590.392746
Avg. Area Number of Rooms	119827.783390
Avg. Area Number of Bedrooms	2361.095262
Avg. Area Income	21.625799
Area Population	15.216581

NO! Estamos comparando unidades diferentes. ¿La edad de la casa es más importante que la media de ingresos de la zona?

Precio ($) = Edad(años) + Habitaciones(n° habitaciones)...

¿Solución? Estandarizar los datos

$$z = \frac{x - \mu}{\sigma}$$


THE
BRIDGE