Regresión lineal

Laboratorio de Datos, IC - FCEN - UBA - 1er. Cuatrimestre 2024

¿Qué modelo uso? 🤔

¿Cómo analizarías estos datos?

0	16.99	1.01	Female	No	Sun	2
1	10.34	1.66	Male	No	Sun	3
2	21.01	3.50	Male	No	Sun	3
3	23.68	3.31	Male	No	Sun	2
4	24.59	3.61	Female	No	Sun	4
5	25.29	4.71	Male	No	Sun	4
6	8.77	2.00	Male	No	Sun	2
7	26.88	3.12	Male	No	Sun	4
8	15.04	1.96	Male	No	Sun	2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Primero, **¿qué es un modelo?**

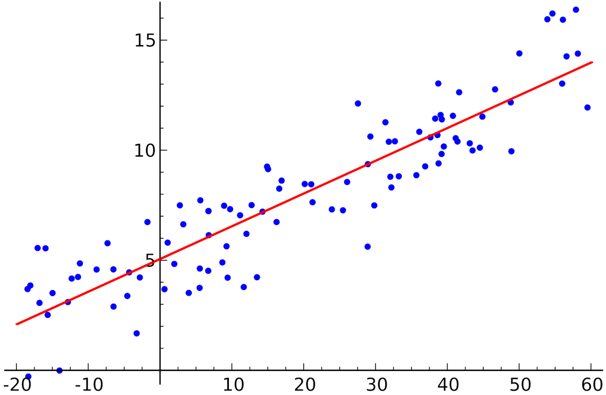
Un modelo es **una representación de fenómenos o procesos del mundo real**. En el contexto de Ciencias de Datos, los modelos son representaciones matemático-computacionales utilizadas para explicar relaciones potencialmente existentes entre las variables de los datos disponibles.

Primero, **¿qué es un modelo?**

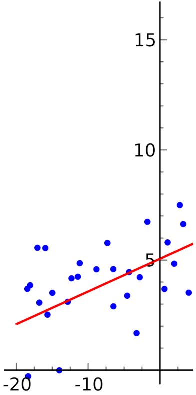
Un modelo es **una representación de fenómenos o procesos del mundo real**. En el contexto de Ciencias de Datos, los modelos son representaciones matemático-computacionales utilizadas para explicar relaciones potencialmente existentes entre las variables de los datos disponibles.

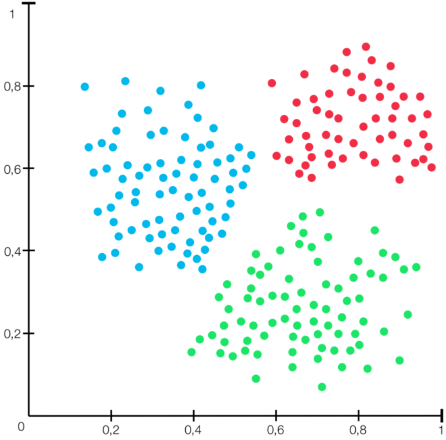
<u>**Muchos**</u> factores influyen en la elección del modelo

¿Cuál es el problema? ¿Cuál es el objetivo del análisis?

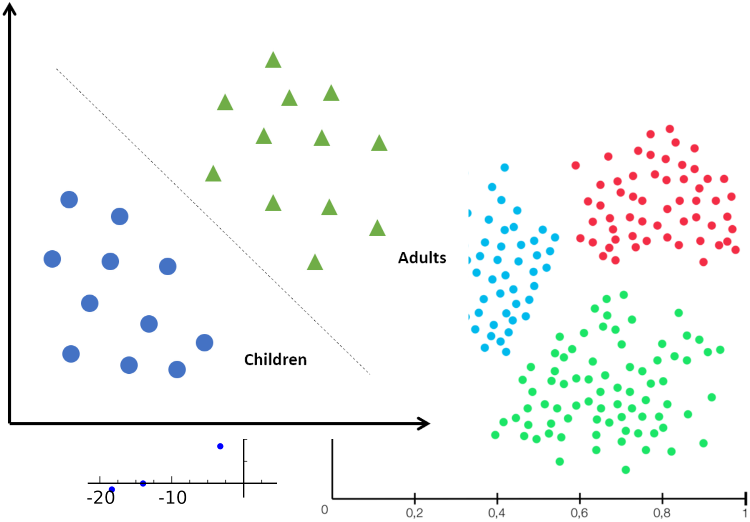


¿Cuál es el problema? ¿Cuál es el objetivo del análisis?

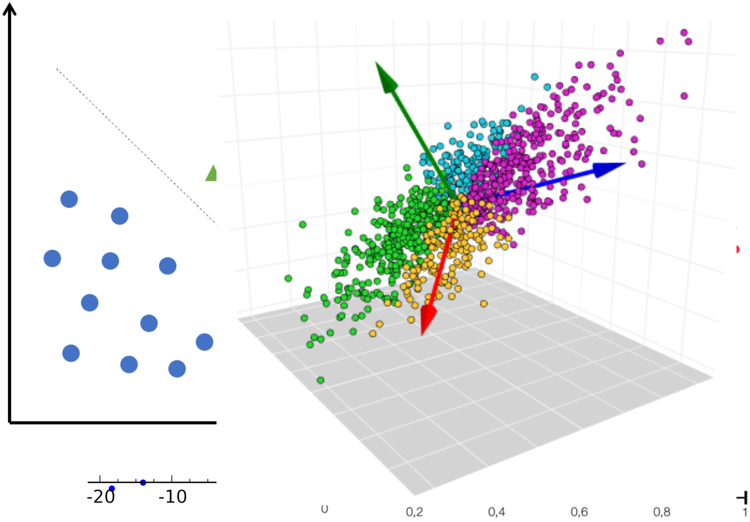




¿Cuál es el problema? ¿Cuál es el objetivo del análisis?



¿Cuál es el problema? ¿Cuál es el objetivo del análisis?



- ¿Cuál es el problema? ¿Cuál es el objetivo del análisis?

- ¿Cuál es el problema? ¿Cuál es el objetivo del análisis?
- ¿Qué tipos de variables tengo?

- ¿Cuál es el problema? ¿Cuál es el objetivo del análisis?
- ¿Qué tipos de variables tengo?
- ¿Cuántos datos tengo?

- ¿Cuál es el problema? ¿Cuál es el objetivo del análisis?
- ¿Qué tipos de variables tengo?
- ¿Cuántos datos tengo?
- ¿Tengo muchos outliers? ¿Qué tan robusto debe ser el modelo?

- ¿Cuál es el problema? ¿Cuál es el objetivo del análisis?
- ¿Qué tipos de variables tengo?
- ¿Cuántos datos tengo?
- ¿Tengo muchos outliers? ¿Qué tan robusto debe ser el modelo?
- ¿Con cuántos recursos computacionales cuento?

- ¿Cuál es el problema? ¿Cuál es el objetivo del análisis?
- ¿Qué tipos de variables tengo?
- ¿Cuántos datos tengo?
- ¿Tengo muchos outliers? ¿Qué tan robusto debe ser el modelo?
- ¿Con cuántos recursos computacionales cuento?
- ¿Es importante poder entender cómo el modelo toma decisiones?

Regresión

Regresión

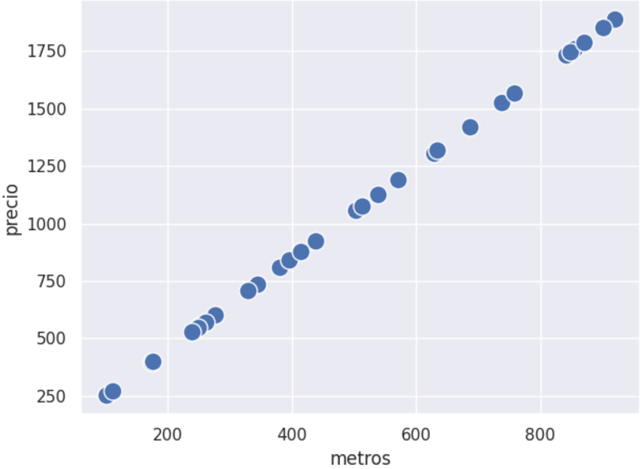
Queremos utilizar los datos que tenemos para poder **estimar datos que no conocemos** o **predecir observaciones futuras**. Los valores a predecir son valores **numéricos**, más precisamente, continuos.

Regresión

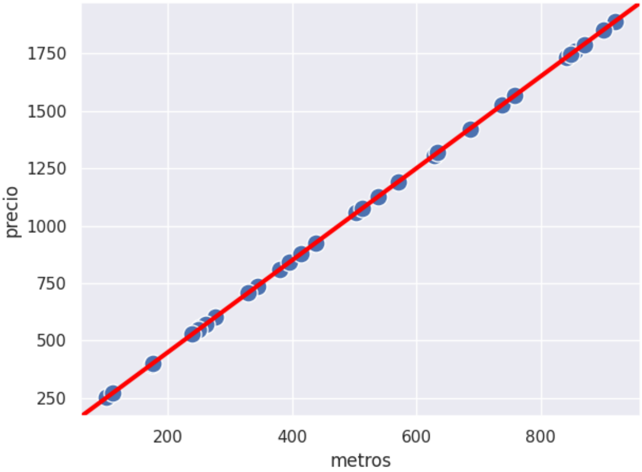
Queremos utilizar los datos que tenemos para poder **estimar datos que no conocemos** o **predecir observaciones futuras**. Los valores a predecir son valores **numéricos**, más precisamente, continuos.

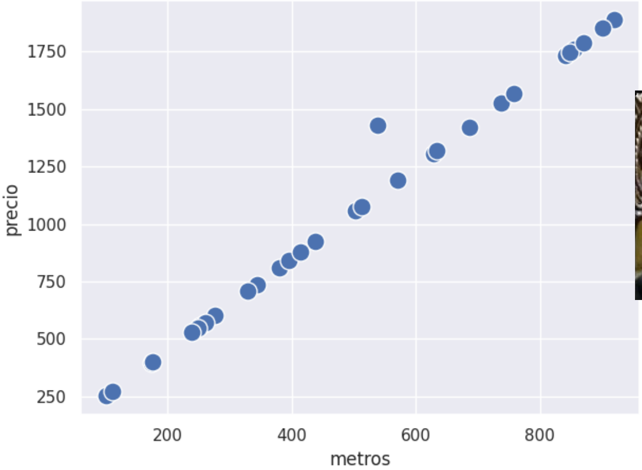
Modelo de regresión simple: $Y = f(X)$

Ejemplo: predecir el valor de un inmueble a partir de su tamaño (m^2)

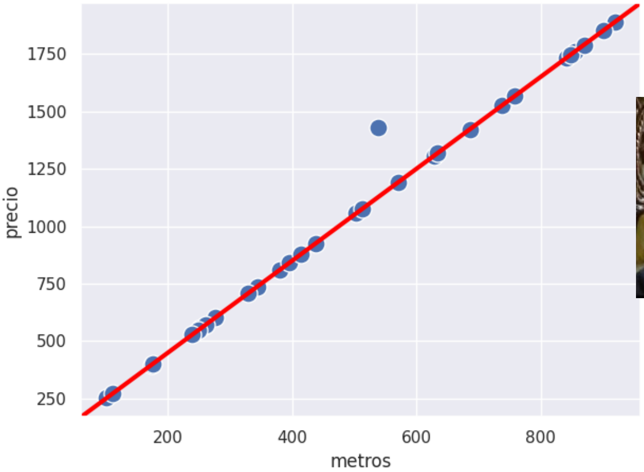

1750
1500
1250
1000
750
500
250
precio
200
400
600
800
metros

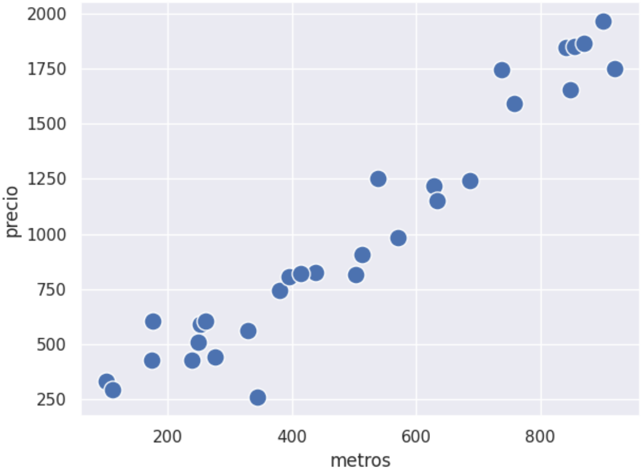
$$y = 2x + 50$$




1750
1500
1250
1000
750
500
250
precio
200
400
600
800
metros

$$y = 2x + 50$$





¿Qué hacemos ahora?

Regresión Lineal

$$\underbrace{\textbf{Regresión}}_{\substack{\text{Queremos predecir} \\ \text{un valor continuo}}} \quad \underbrace{\textbf{Lineal}}_{\text{con una recta}}$$

Modelo matemático: $Y = \beta_0 + \beta_1 X$

- β_0 es la ordenada al origen
- β_1 es la pendiente
- X es la variable predictora
- Y es la variable dependiente

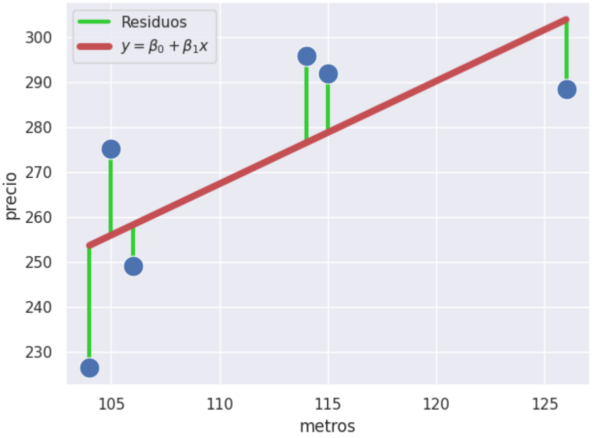
Modelo de regresión lineal (simple): $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

- (x_i, y_i) son los datos observados
- ε_i es un error aleatorio (variación de Y no explicada por X)
- β_0 y β_1 son **los parámetros del modelo**

Notación de Wilkinson: $Y \sim X$

Residuo: dados β_0 y β_1, definimos al residuo como la diferencia entre el valor observado (y_i) y el valor predicho ($\hat{y}_i$):

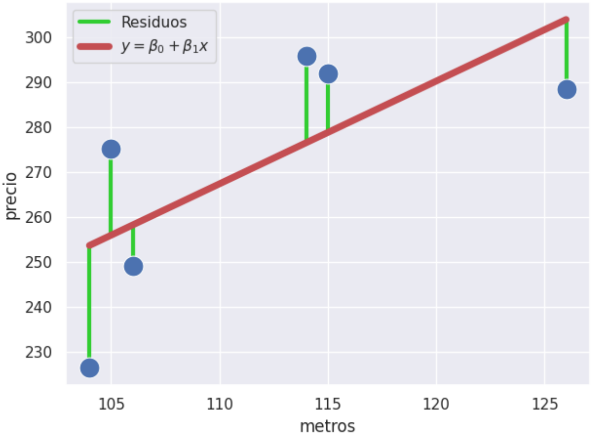
$$y_i - \underbrace{(\beta_0 + \beta_1 x_i)}_{\hat{y}_i}$$



Residuo: dados β_0 y β_1, definimos al residuo como la diferencia entre el valor observado (y_i) y el valor predicho ($\hat{y}_i$):

$$y_i - \underbrace{(\beta_0 + \beta_1 x_i)}_{\hat{y}_i}$$

$\rightarrow$ Queremos encontrar valores para β_0 y β_1 que minimicen los residuos



Cuadrados Mínimos

Cuadrados Mínimos

Minimizar la suma de los residuos al cuadrado:

$$
\begin{aligned}
RSS(\beta_0, \beta_1) = & \quad (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots + (y_n - \hat{y}_n)^2 = \\
= & \quad (y_1 - (\beta_0 + \beta_1 x_1))^2 + \cdots + (y_n - (\beta_0 + \beta_1 x_n))^2 = \\
= & \quad \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2
\end{aligned}
$$

Calculamos $\hat{\beta}_0$ y $\hat{\beta}_1$ tales que $\nabla RSS(\hat{\beta}_0, \hat{\beta}_1) = (0,0)$:

$$\begin{aligned} \hat{\beta}_1 = & \frac{\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \hat{\beta}_0 = & \bar{y} - \hat{\beta}_1 \bar{x} \end{aligned}$$

donde:

$$\begin{aligned} \bar{x} = & \frac{1}{n}\sum_{i=1}^{n} x_i \\ \bar{y} = & \frac{1}{n}\sum_{i=1}^{n} y_i \end{aligned}$$

¿Qué tan bien ajusta el modelo?

Vamos a ver dos formas de responder a esta pregunta:

¿Qué tan bien ajusta el modelo?

Vamos a ver dos formas de responder a esta pregunta:

- Error Cuadrático Medio

¿Qué tan bien ajusta el modelo?

Vamos a ver dos formas de responder a esta pregunta:

- Error Cuadrático Medio
- El coeficiente de determinación R^2

¿Qué tan bien ajusta el modelo? - ECM

Error cuadrático medio (ECM): cuantifica qué tan cerca está un valor predicho del valor real:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \underbrace{(\beta_0 + \beta_1 x_i)}_{\hat{y}_i})^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

¿Qué tan bien ajusta el modelo? - ECM

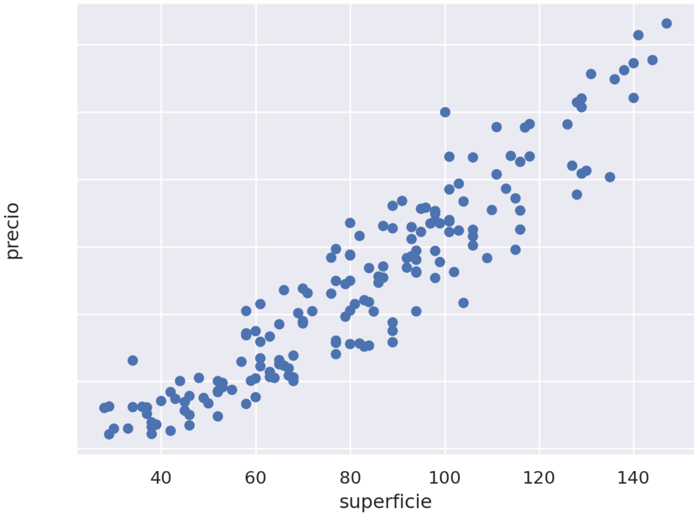
Error cuadrático medio (ECM): cuantifica qué tan cerca está un valor predicho del valor real:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \underbrace{(\beta_0 + \beta_1 x_i)}_{\hat{y}_i})^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

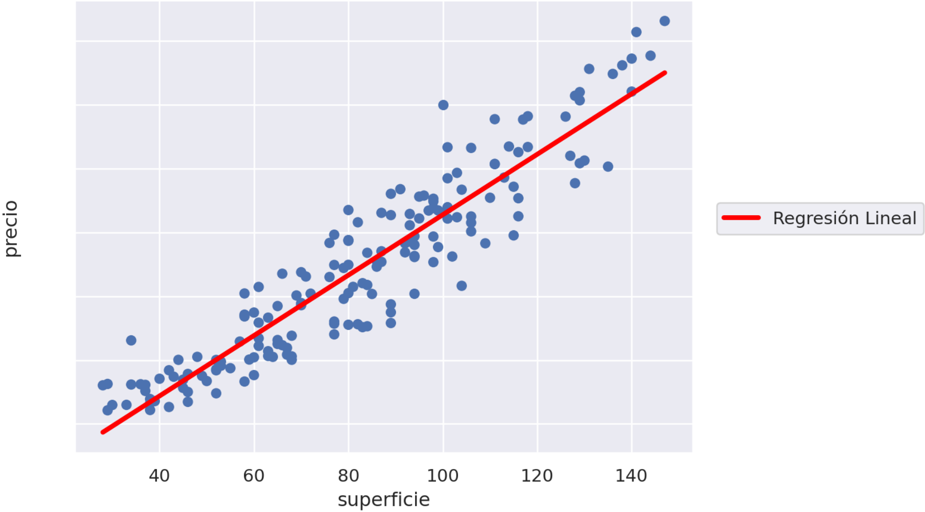
Para facilitar la interpretación, usamos la **Raíz del Error Cuadrático Medio (RECM)**:

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

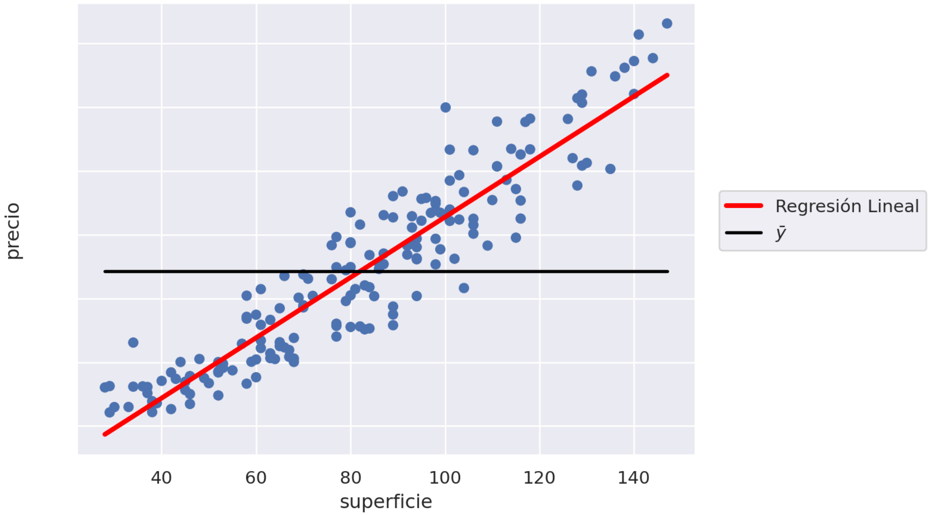
¿Qué tan bien ajusta el modelo? - Coeficiente de determinación



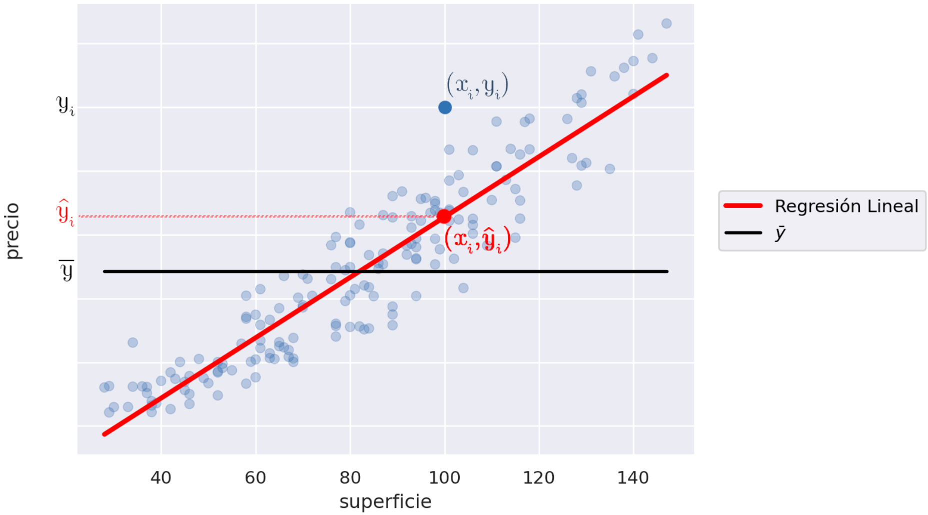
¿Qué tan bien ajusta el modelo? - Coeficiente de determinación



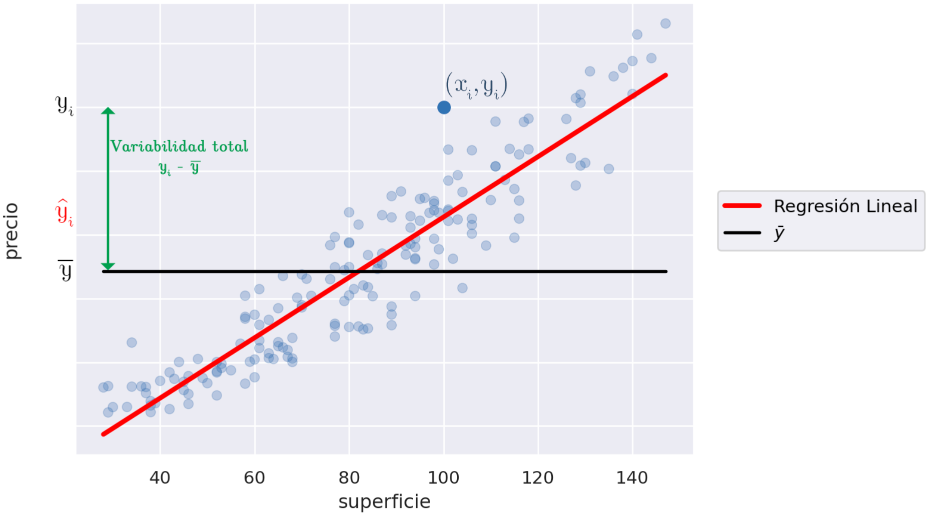
¿Qué tan bien ajusta el modelo? - Coeficiente de determinación



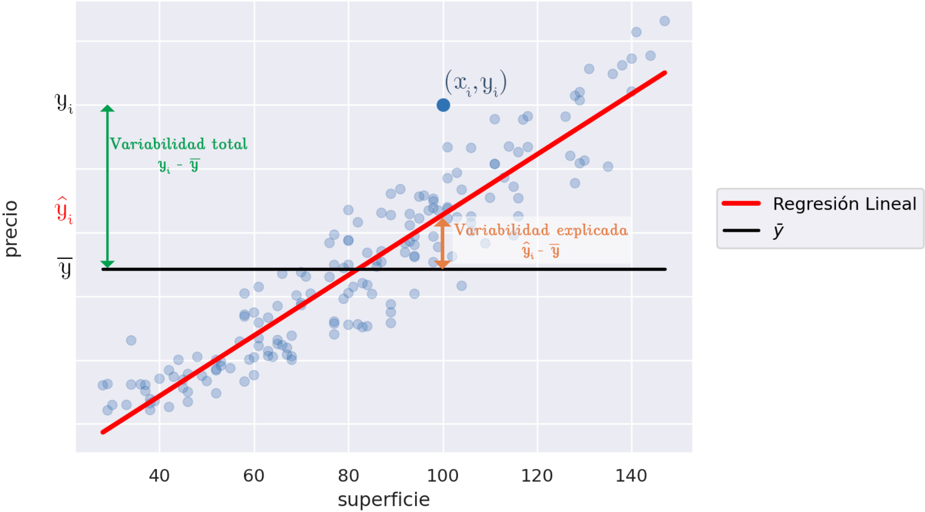
¿Qué tan bien ajusta el modelo? - Coeficiente de determinación



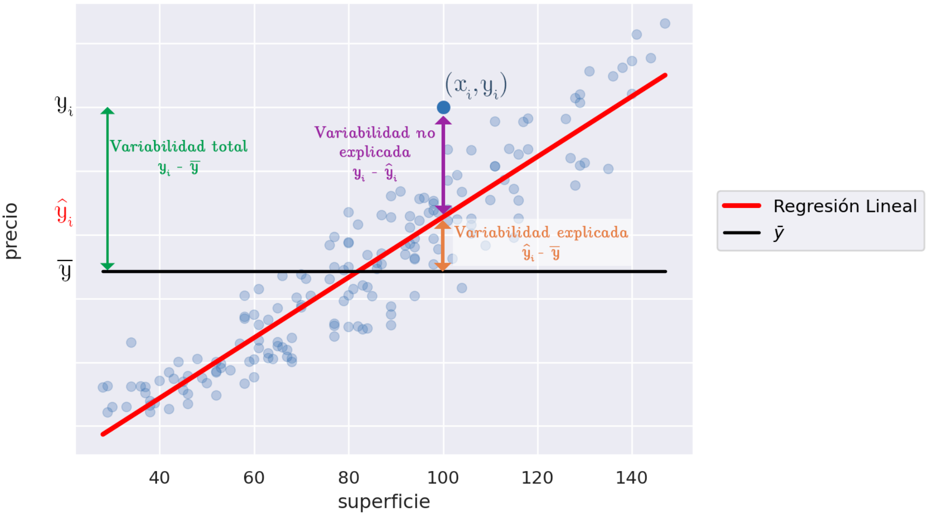
¿Qué tan bien ajusta el modelo? - Coeficiente de determinación



¿Qué tan bien ajusta el modelo? - Coeficiente de determinación



¿Qué tan bien ajusta el modelo? - Coeficiente de determinación



¿Qué tan bien ajusta el modelo? - Coeficiente de determinación

Variabilidad del modelo:

$$\begin{aligned}
\text{Variabilidad total:} \quad & \textstyle\sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad (\approx \text{Varianza muestral}) \\
\text{Variabilidad no explicada:} \quad & \textstyle\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (\text{RSS}) \\
\text{Variabilidad explicada:} \quad & \textstyle\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2
\end{aligned}$$

¿Qué tan bien ajusta el modelo?

La proporción de la variabilidad de Y explicada por X se puede explicar como:

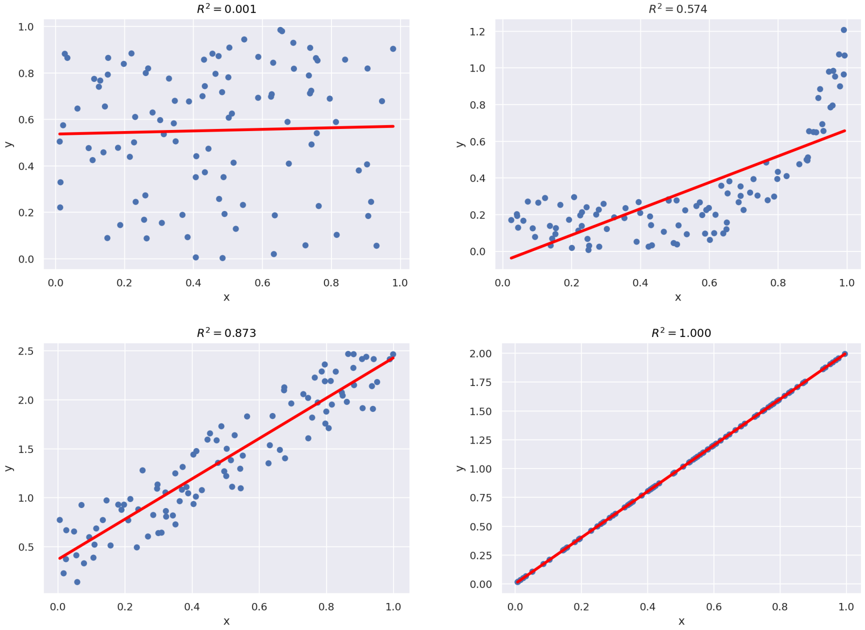
$$R^2 = \frac{\text{Variabilidad explicada}}{\text{Variabilidad total}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

¿Qué tan bien ajusta el modelo?

La proporción de la variabilidad de Y explicada por X se puede explicar como:

$$R^2 = \frac{\text{Variabilidad explicada}}{\text{Variabilidad total}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Obs: $0 \leq R^2 \leq 1$
A mayor R^2 más cercanos están los puntos a la recta y, por lo tanto, tiene más poder de predicción.



$R^2 = 0.001$
$R^2 = 0.574$
$R^2 = 0.873$
$R^2 = 1.000$
x
y