Regresión lineal y correlación

DRA. CONSUELO VARINIA GARCÍA MENDOZA

Regresión lineal y correlación

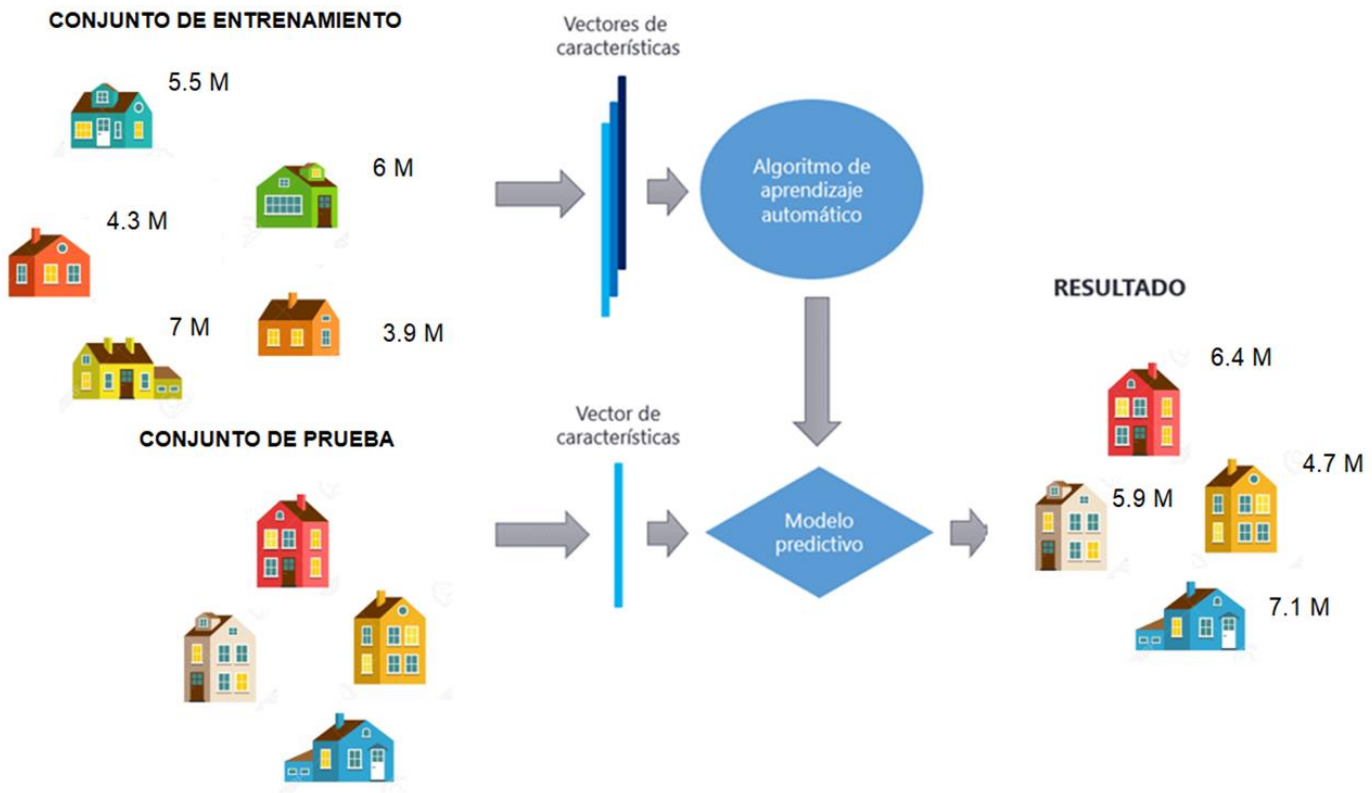
En la práctica a menudo se requiere resolver problemas que implican conjuntos de variables que están correlacionadas.

- El rendimiento del combustible de un motor está relacionado con su volumen
- El precio de una casa está relacionado con su superficie de construcción (tamaño)

Es de interés un método de pronóstico

- El rendimiento de cualquier motor dado su volumen
- El precio de cualquier casa dado tu tamaño

Aprendizaje supervisado



Relación no determinista

Ejemplos:

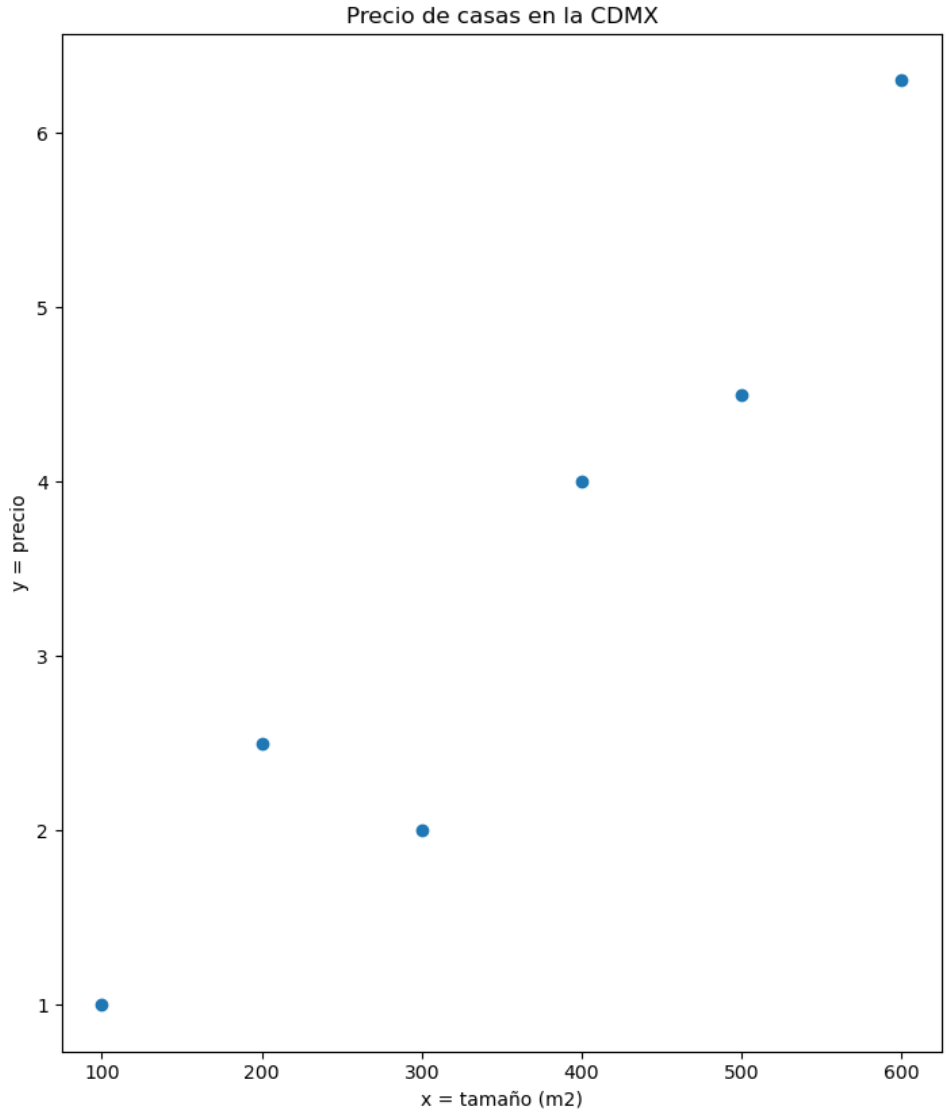
- Autos con motores del mismo volumen pueden tener distinto rendimiento de combustible
- Casas con la misma superficie de construcción distintos precios

Componente aleatorio relacionado no otras variables independientes o incluso elementos desconocidos.

- A pesar de que no se pude hacer un pronóstico exacto si es posible hacer un pronóstico estimado o ajustado

Diagrama de dispersión

tamaño (m^2)	precio
100	1
200	2.5
300	2
400	4
500	4.5
600	6.3



Pronóstico estimado o ajustado $\hat{y}$

Relación lineal

$$\hat{y} = b_0 + b_1 x$$

donde

b_0 : intersección

b_1 : pendiente

x : variable independiente o regresor (tamaño de la casa, volumen del motor)

$\hat{y}$: variable dependiente o respuesta estimada (precio, rendimiento del combustible)

Diagrama de dispersión con rectas de regresión

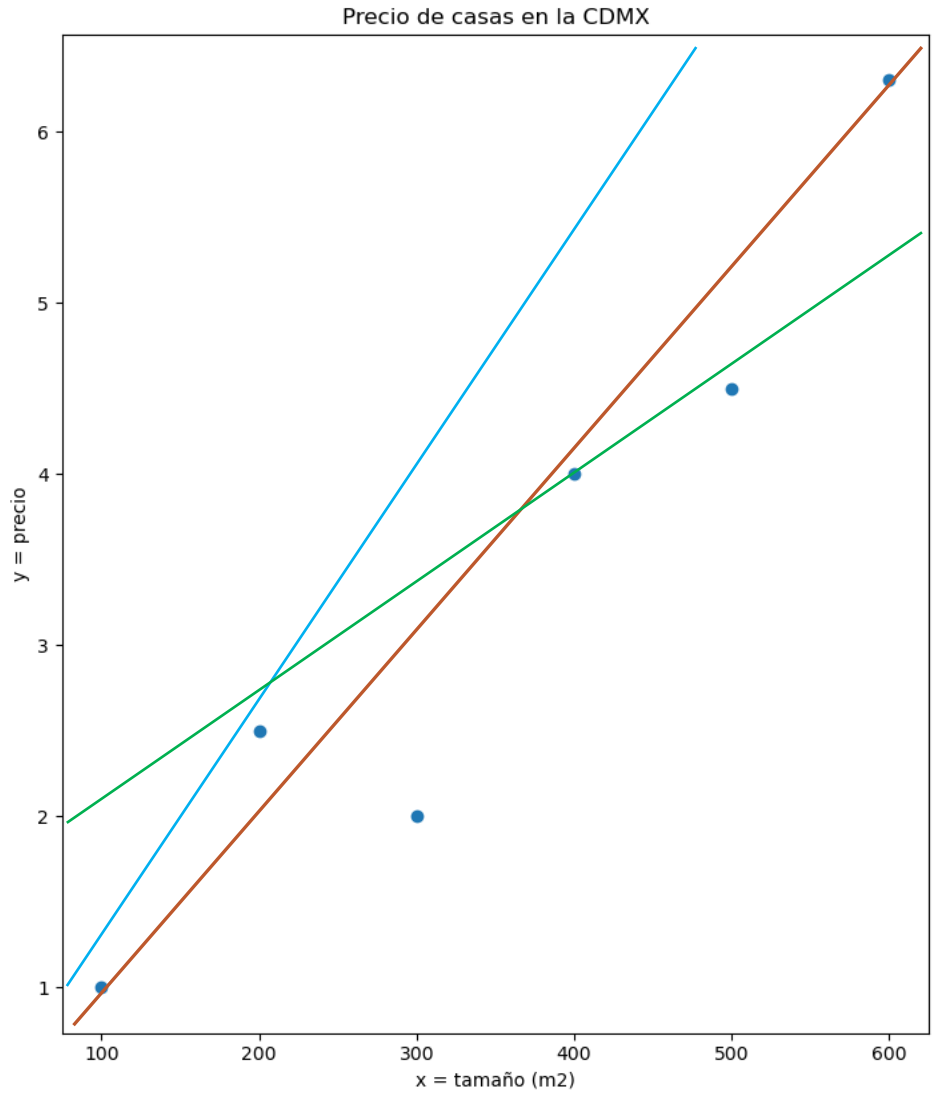
$$\widehat{precio} = b_0 + b_1 tamaño$$

donde

b_0 : intersección

b_1 : pendiente

b_0 y b_1 son los coeficientes de valores reales que el modelo debe aprender



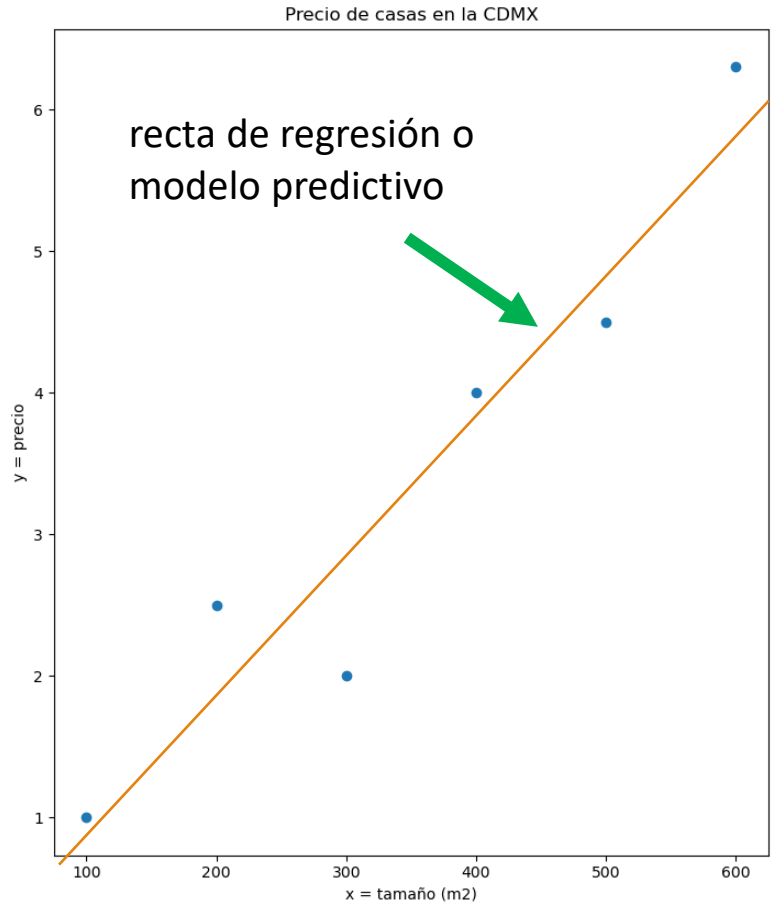
Enfoque matemático y de aprendizaje supervisado

tamaño (m^2) regresor/característica	precio respuesta/etiqueta
100	1
200	2.5
300	2
400	4
500	4.5
600	6.3
120	$\hat{y}$ (120)
180	$\hat{y}$ (180)
240	$\hat{y}$ (240)
375	$\hat{y}$ (375)

instancias del conjunto de entrenamiento

conjunto de prueba

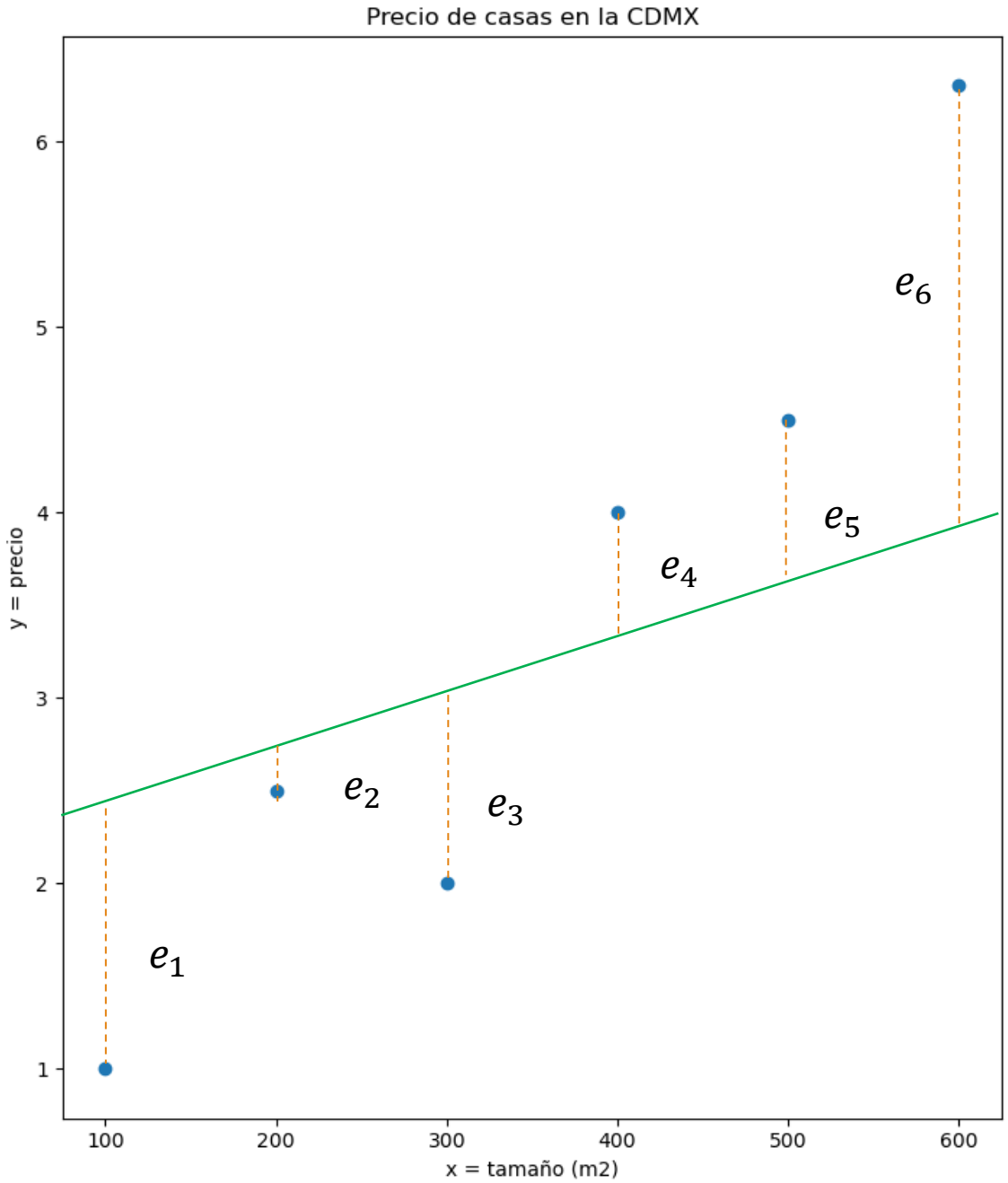
respuesta estimada o hipótesis



Error de estimación

Cada recta de regresión que elijamos para modelar los datos tendrá asociada una suma del error respecto a la recta de regresión estimada

$$\sum_{I=1}^{n} |y_i - \hat{y}_i|$$



Mínimos cuadrados

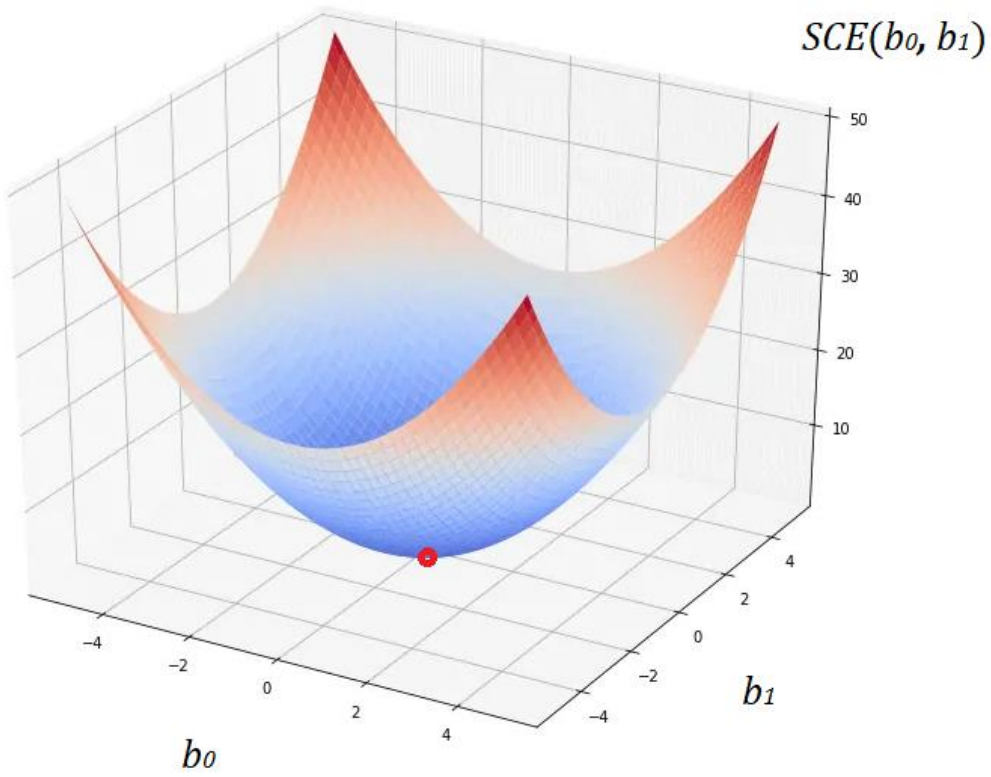
Suma de los cuadrados del error respecto a la recta de regresión estimada

$$\mathrm{SCE}(b_{0,}\, b_1) = \sum_{I=1}^{n} {e_i}^2 = \sum_{I=1}^{n} (y_i - \hat{y_i})^2$$

Problema de optimización

$$\frac{\partial \mathrm{SCE}(b_{0,}\, b_1)}{\partial b_0} = 0$$

$$\frac{\partial \mathrm{SCE}(b_{0,}\, b_1)}{\partial b_1} = 0$$



Método Analítico

$$\frac{\partial \sum_{I=1}^{n}(y_i - b_0 - b_1 x)^2}{\partial b_0} = 0$$

$$\frac{\partial \sum_{I=1}^{n}(y_i - b_0 - b_1 x)^2}{\partial b_1} = 0$$

$$b_0 = \frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i}{n}$$

$$b_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n \sum_{i=1}^{n} {x_i}^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Actividad

Medidas de la reducción de los sólidos y de la demanda de oxígeno químico

Reducción de sólidos, x (%)	Reducción de la demanda de oxígeno, y (%)	Reducción de sólidos, x (%)	Reducción de la demanda de oxígeno, y (%)
3	5	36	34
7	11	37	36
11	21	38	38
15	16	39	37
18	16	39	36
27	28	39	45
29	27	40	39
30	25	41	41
30	35	42	40
31	30	42	44
31	40	43	37
32	32	44	44
33	34	45	46
33	32	46	46
34	34	47	49
36	37	50	51
36	38		

Uno de los problemas más desafiantes que enfrenta el campo del control de la contaminación del agua lo representa la industria de la peletería, ya que sus desechos son químicamente complejos; se caracterizan por valores elevados de la demanda de oxígeno químico, sólidos volátiles y otras medidas de contaminación. La tabla contiene los datos experimentales que se obtuvieron de 33 muestras de desechos tratados químicamente en un estudio. Se registraron los valores de x, la reducción porcentual de los sólidos totales, y de y, el porcentaje de disminución de la demanda de oxígeno químico.

a) Encuentra los valores de b_0 y b_1
b) ¿Cuál es la recta de regresión estimada $\hat{y}_i$