Regresión lineal

Maestría en Ciencia de Datos, CUCEA, Universidad de Guadalajara.

Guadalajara, Jal., agosto de 2025

Regresión lineal

Ejemplo

- Queremos modelar el precio de los departamentos en una ciudad según su superficie (en m^2).
- La **variable de entrada** (independiente) nos ayuda a predecir el valor de la variable de salida (superficie del piso).
- La **variable de salida** (dependiente) es la que queremos predecir (el costo mensual)

Superficie [metros cuadrados]	Renta [pesos]
150	4500
120	3800
170	4800
80	2700

Regresión lineal

Ecuación de una recta

- En el caso de una regresión lineal, asumimos que y (la renta) es una función lineal de x (la superficie) y entonces el modelo lineal se escribe como:

- $\hat{y} = \beta_0 + \beta_1 x$, i.e., Renta = β_0 + β_1 * superficie.

- El objetivo de una regresión lineal es encontrar de forma estadística unos valores que sean significativos para β_0 y β_1, de tal forma que minimizan las diferencias entre los valores reales de y (los que realmente tenemos) y los que prediga nuestro modelo ($\hat{y}$).

- Con estos parámetros tendremos una ecuación que será capaz de predecir los valores de y en función de la variables (o variables) de entrada x.

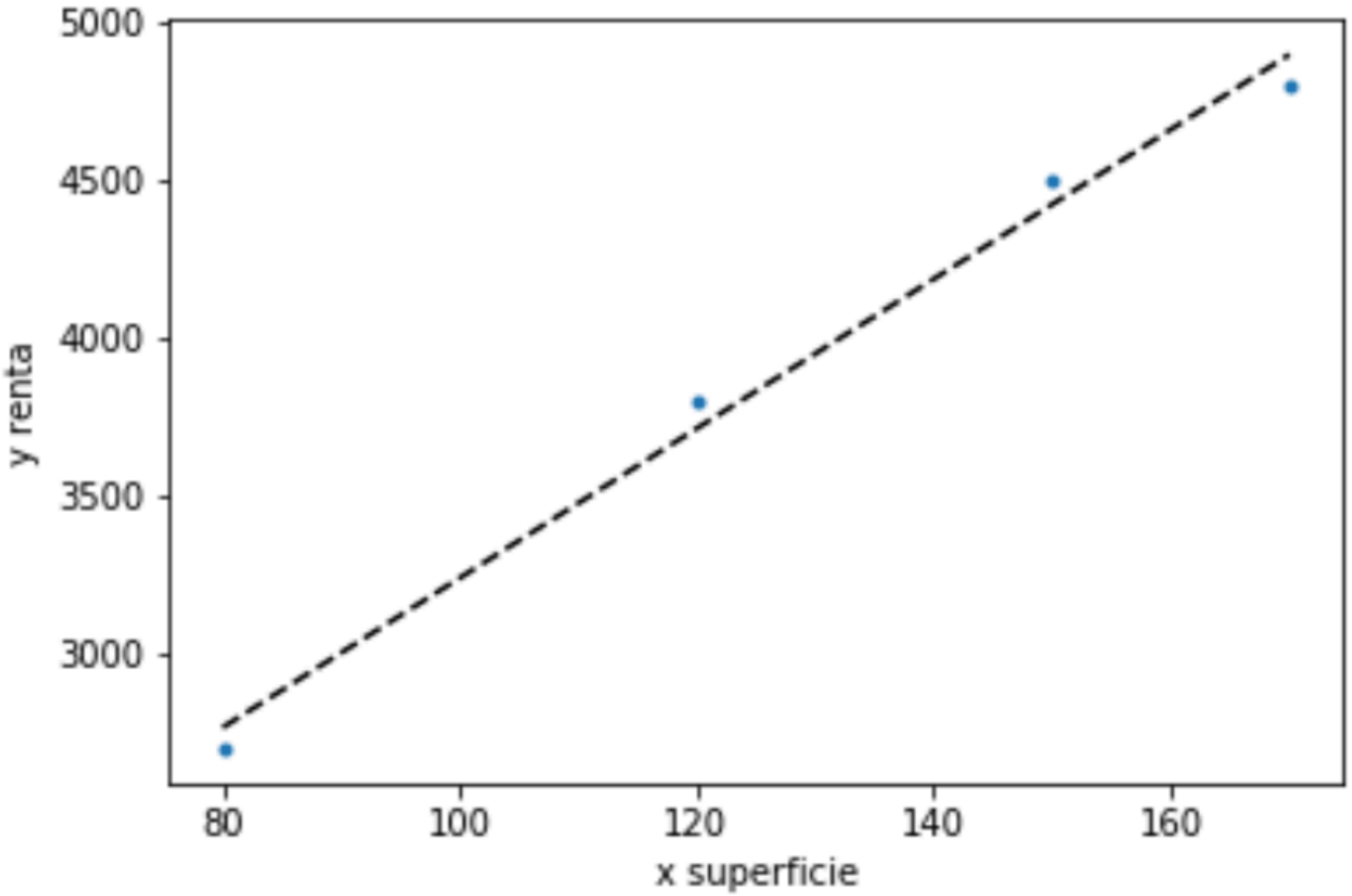
Regresión lineal

Ecuación de una recta

- Requiere de una serie de datos históricos con tantos valores de entrada como uno quiera modelar pero solo un parámetro de salida.
- Por ejemplo, con ciertos datos históricos (una base de datos anterior) podríamos crear un modelo lineal con valores $\beta_0 = 869.6$ y $\beta_1 = 23.7$.
- El modelo lineal sería $\hat{y} = 869.6 + 23.7x$.
- Con esto, por ejemplo, podríamos encontrar el alquiler de cualquier casa , por ejemplo una de 110 metros cuadrados.
- $\hat{y} = f(x = 110) = 869.6 + 23.7(110) = 3476.6$ pesos.

Regresión lineal

Ecuación de una recta



$$\hat{y} = 869.6 + 23.7x$$

$$\text{renta} = 869.6 + 23.7 * \text{superficie}$$

Interpretación de los parámetros:

- β_0 (intercept): también llamado ordenada al origen, representa el punto donde la recta de regresión cruza el eje y.

- β_1: Es el cambio esperado en y por cada unidad de aumento en x, manteniendo todo lo demás constante.

Regresión lineal

Modelo supervisado de ML

- La **regresión lineal** es un modelo de aprendizaje supervisado utilizado para predecir una **variable continua** a partir de una o más variables de entrada.

- Su objetivo es encontrar una relación lineal entre las variables independientes x y la variable dependiente y, ajustando una línea (o un hiperplano en dimensionas mayores) que minimiza la diferencia entre los valores predichos y los observados (datos).

- En el contexto de Machine Learning, se considera un **modelo base o de referencia,** debido a su simplicidad, interpretabilidad y utilidad para entender patrones en los datos, además de ser un punto de comparación frente a modelos más complejos.

Regresión lineal simple

Regresión lineal simple

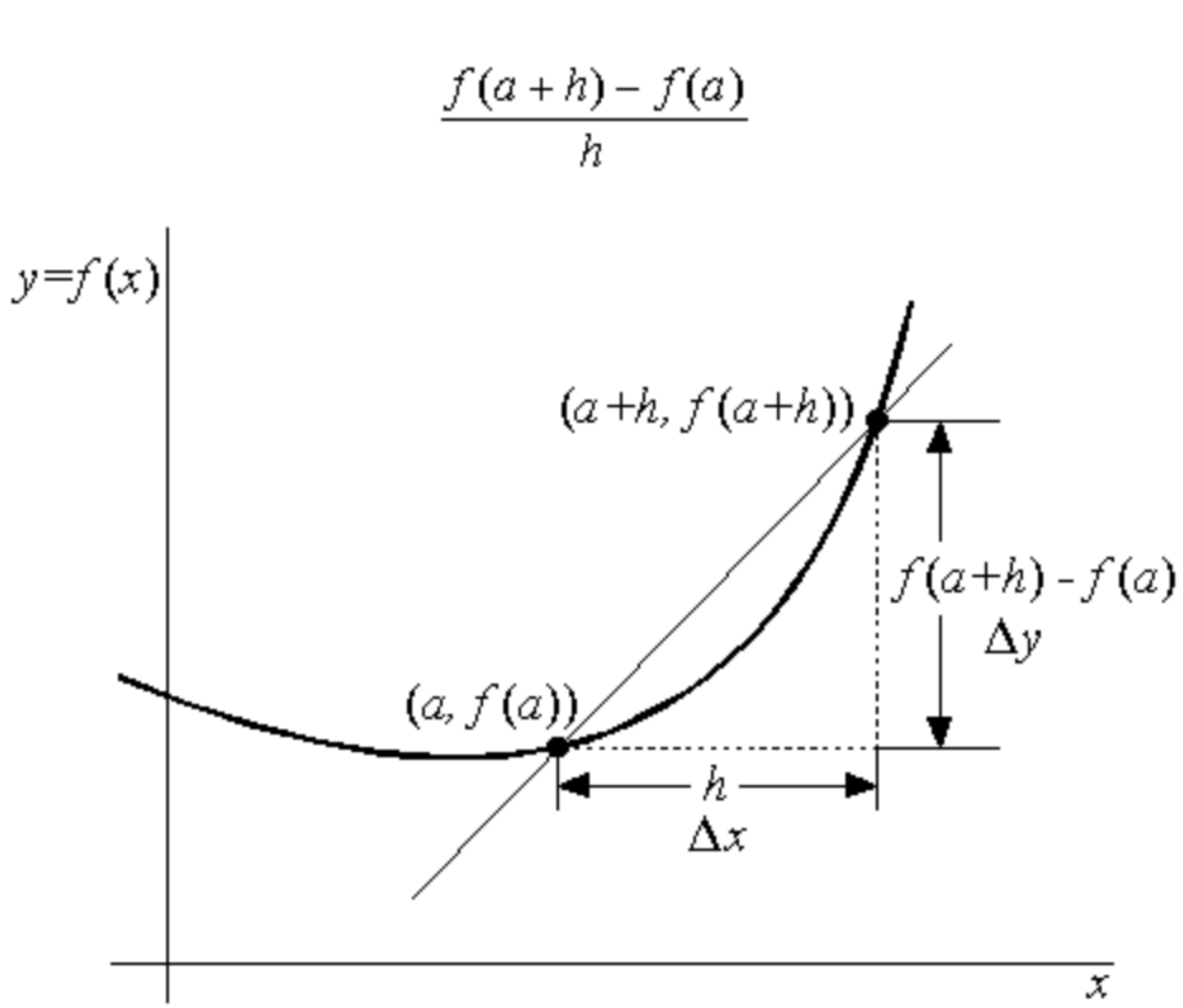
Mínimos cuadrados

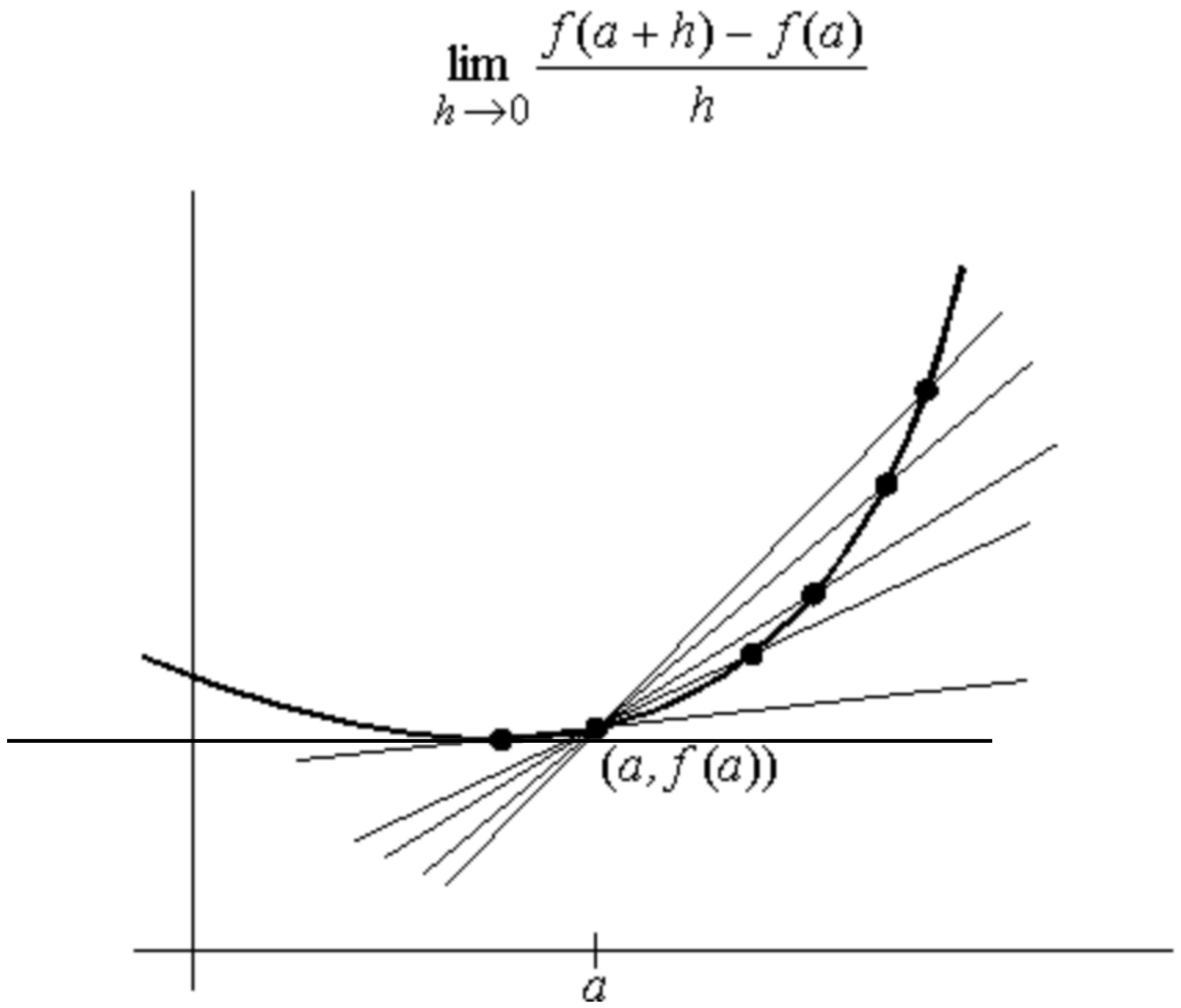
- ¿Cómo podemos encontrar los parámetros del modelo lineal?
- La diferencia entre el valor real y el estimado (residuales) se puede escribir como:
- $E = (\hat{y}_i - y_i)$, donde $\hat{y}_i$ son los valores predecidos por cada x_i, y y_i son los valores históricos (asociados a cada x_i).
- El objetivo es minimizar la suma de errores al cuadrado sobre todos los puntos del data set $X = \{(x_i, y_i)\}_{i=1}^{n}$
- $\min \sum_{i=1}^{n} E^2 = \sum_{i=1}^{n} (\hat{y}_i(x_i) - y_i)^2 = \sum_{i=1}^{n} ((\beta_0 + \beta_1 x_i) - y_i)^2 = e$

Regresión lineal simple

Mínimos cuadrados

- La suma de residuales al cuadrado se tiene que minimizar, y para eso derivamos para encontrar los parámetros óptimos β_0 y β_1





Optimización

Parámetro β_0

- Queremos ver dónde se encuentra el mínimo de este error. En otras palabras, dónde su derivada (pendiente) es igual a cero. Para eso tenemos que utilizar la primera derivada.

- Primero derivamos con respecto al parámetro β_0

- $\frac{\partial}{\partial \beta_0} e = \frac{\partial}{\partial \beta_0} \left[\sum_{i=1}^{n} ((\beta_0 + \beta_1 x_i) - y_i)^2 \right] = 0$

Optimización

Parámetro β_0

- La idea es ver esta expresión de una manera más sencilla. Todo lo que está dentro del paréntesis se puede ver como **una función de β_0**, y podríamos llamarla si queremos $u(\beta_0) = \beta_0 + \beta_1 x_i - y_i$.

- **Recordar que en esta derivada parcial, β_0 es la única variable. Qué quiere decir? Que consideramos β_1 como constante, un número.**

- $\dfrac{\partial}{\partial \beta_0} e = \dfrac{\partial}{\partial \beta_0}\left[\sum_{i=1}^{n} (u(\beta_0))^2\right] = \dfrac{\partial}{\partial \beta_0}\left[\sum_{i=1}^{n} u^2(\beta_0)\right]$ Ya no se ve tan mal.!!!

Optimización

Parámetro β_0

- $\dfrac{\partial}{\partial \beta_0} e = \dfrac{\partial}{\partial \beta_0} \left[\sum_{i=1}^{n} u^2(\beta_0) \right]$. Entonces, tenemos una función anidada en otra!!!
- Tenemos una función $e(u) = \sum_{i=1}^{n} u^2$, $u(\beta_0) = \beta_0 + \beta_1 x_i - y_i$. Por regla de la cadena:
- $\dfrac{\partial e}{\partial \beta_0} = \dfrac{\partial e}{\partial u} \dfrac{\partial u}{\partial \beta_0}$

Optimización

Parámetro β_0

- $$\frac{\partial e}{\partial \beta_0} = \frac{\partial e}{\partial u}\frac{\partial u}{\partial \beta_0}$$
- Hacemos cada una por separado. Cada una de las derivadas únicamente requiere derivadas **de una ley de potencias. !!! Recordar sustituir de vuelta la variable u del comienzo.**
- $$\frac{\partial e}{\partial u} = \sum_{i=1}^{n} 2u = 2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)$$
- $$\frac{\partial u}{\partial \beta_0} = 1$$

Optimización

Parámetro β_0

- Entonces, sólo nos queda sustituir:
- $$\frac{\partial e}{\partial \beta_0} = \frac{\partial e}{\partial u}\frac{\partial u}{\partial \beta_0} = \left[2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)\right][1] = 0$$
- Que podemos reordenar, despejar:
- $$\Rightarrow \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 = \sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i$$
- $$\Rightarrow \sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i$$

Optimización

Parámetro β_1

- Derivando con respecto al parámetro β_1 es algo muy similar, pero hay un factor x_i adicional por ahí.
- $\dfrac{\partial}{\partial \beta_1}(e) = \left[2 \sum_{i=1}^{n} \left[-x_i \right] (y_i - \beta_0 - \beta_1 x_i) \right] = 0$
- $\Rightarrow \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0 = \sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2$
- $\Rightarrow \sum_{i=1}^{n} x_i y_i = \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2$

Optimización

Parámetro óptimos : sistema de ecuaciones

- Entonces tenemos un sistema de dos ecuaciones:

$$\sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i \qquad (1)$$

$$\sum_{i=1}^{n} x_i y_i = \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 \qquad (2)$$

- Podemos resolver este sistema de ecuaciones por el método de sustitución.

Optimización

Parámetro β_0

- De la ecuación 1, podemos dividir entre n:
- $\frac{1}{n}\sum_{i=1}^{n} y_i = \beta_0 + \beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i$
- $\Rightarrow \bar{y} = \beta_0 + \beta_1 \bar{x}$
- $\Rightarrow \beta_0 = \bar{y} - \beta_1 \bar{x}$

Optimización

Parámetro β_1

- Entonces podemos sustituir esta expresión para β_0 en la ecuación (2).
- $$\sum_{i=1}^{n} x_i y_i = \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = (\bar{y} - \beta_1 \bar{x}) \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2$$
- $$= \bar{y} \sum_{i=1}^{n} x_i - \beta_1 \bar{x} \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2$$
- Dividimos entre n,
- $$\frac{1}{n} \sum_{i=1}^{n} x_i y_i = \bar{y} \frac{1}{n} \sum_{i=1}^{n} x_i - \beta_1 \bar{x} \frac{1}{n} \sum_{i=1}^{n} x_i + \beta_1 \frac{1}{n} \sum_{i=1}^{n} x_i^2$$
- $$\Rightarrow \frac{1}{n} \sum_{i=1}^{n} x_i y_i = \bar{y}\bar{x} - \beta_1 \bar{x}\bar{x} + \beta_1 \frac{1}{n} \sum_{i=1}^{n} x_i^2$$

Optimización

Parámetro β_1

- Despejando tenemos que:
- $$\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} = \beta_1 \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 \right)$$
- $$\Rightarrow Cov(x, y) = \beta_1 \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 \right)$$

Optimización

Parámetro β_1

- Por otro lado, tenemos que la varianza puede expresarse de otra forma:
- $$Var(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\left(\sum_{i=1}^{n}x_i^2 - 2\sum_{i=1}^{n}x_i\bar{x} + \sum_{i=1}^{n}\bar{x}^2\right)$$
- $$= \frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2\bar{x}\frac{1}{n}\sum_{i=1}^{n}x_i + \frac{1}{n}n\bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2$$
- $$Var(x) = \sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2$$

Optimización

Parámetro β_1

- $Cov(x,y) = \sigma_{xy} = \beta_1 \left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2 \right) = \beta_1 \sigma_x^2$
- $\beta_1 = \frac{\sigma_{xy}}{\sigma_x^2}$

Regresión lineal simple

Parámetros óptimos

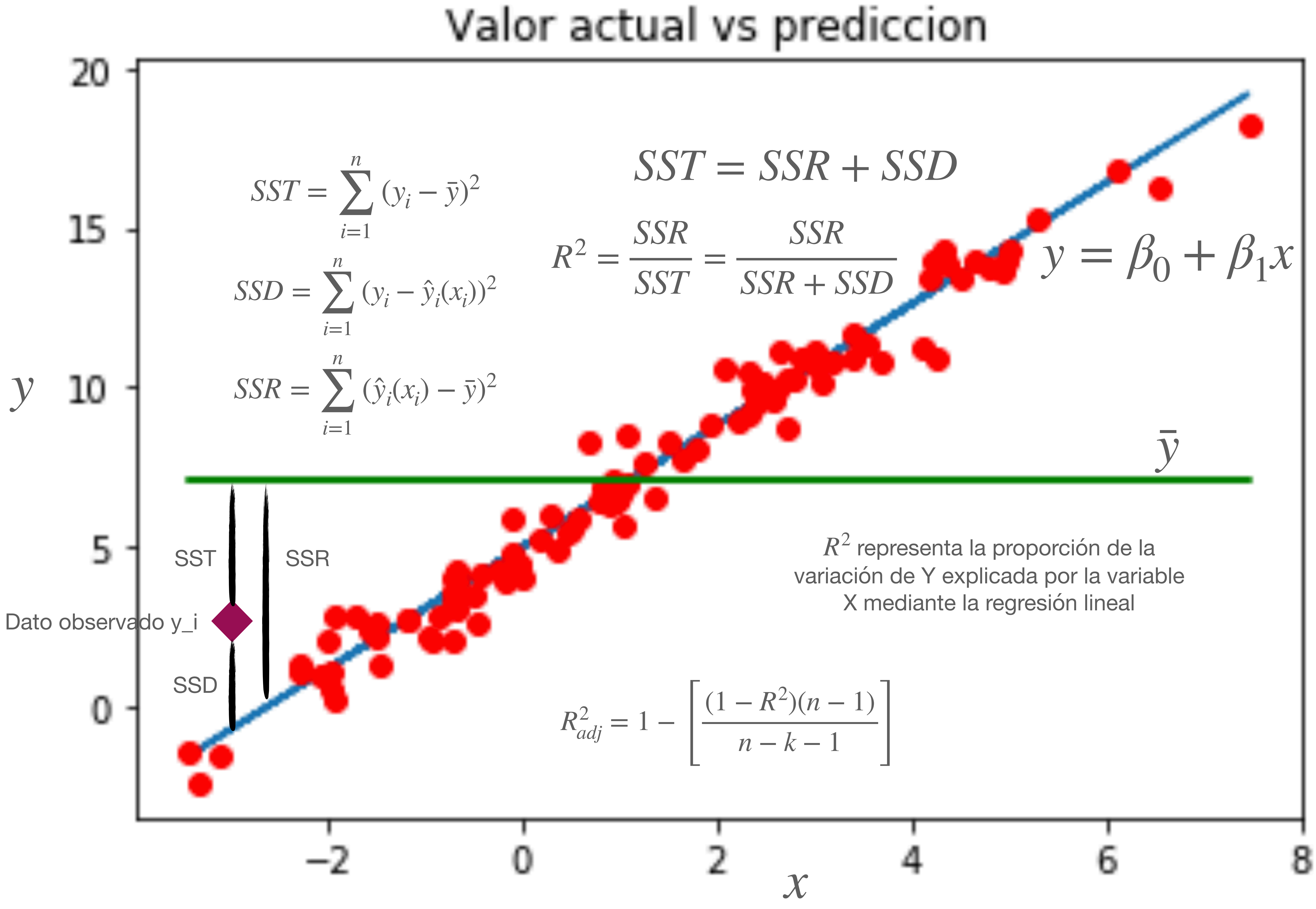
- Los parámetros que minimizan la suma de los residuos al cuadrado son:

$$\beta_1 = \frac{\sigma_{xy}}{\sigma_x^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$A = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$B = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$



Valor actual vs prediccion
$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$
$SSD = \sum_{i=1}^{n} (y_i - \hat{y}_i(x_i))^2$
$SSR = \sum_{i=1}^{n} (\hat{y}_i(x_i) - \bar{y})^2$
$SST = SSR + SSD$
$R^2 = \frac{SSR}{SST} = \frac{SSR}{SSR + SSD}$
$y = \beta_0 + \beta_1 x$
$\bar{y}$
SST
SSR
SSD
Dato observado y_i
R^2 representa la proporción de la variación de Y explicada por la variable X mediante la regresión lineal
$R^2_{adj} = 1 - \left[\frac{(1 - R^2)(n-1)}{n-k-1}\right]$
y
x
20
15
10
5
0
−2
0
2
4
6
8

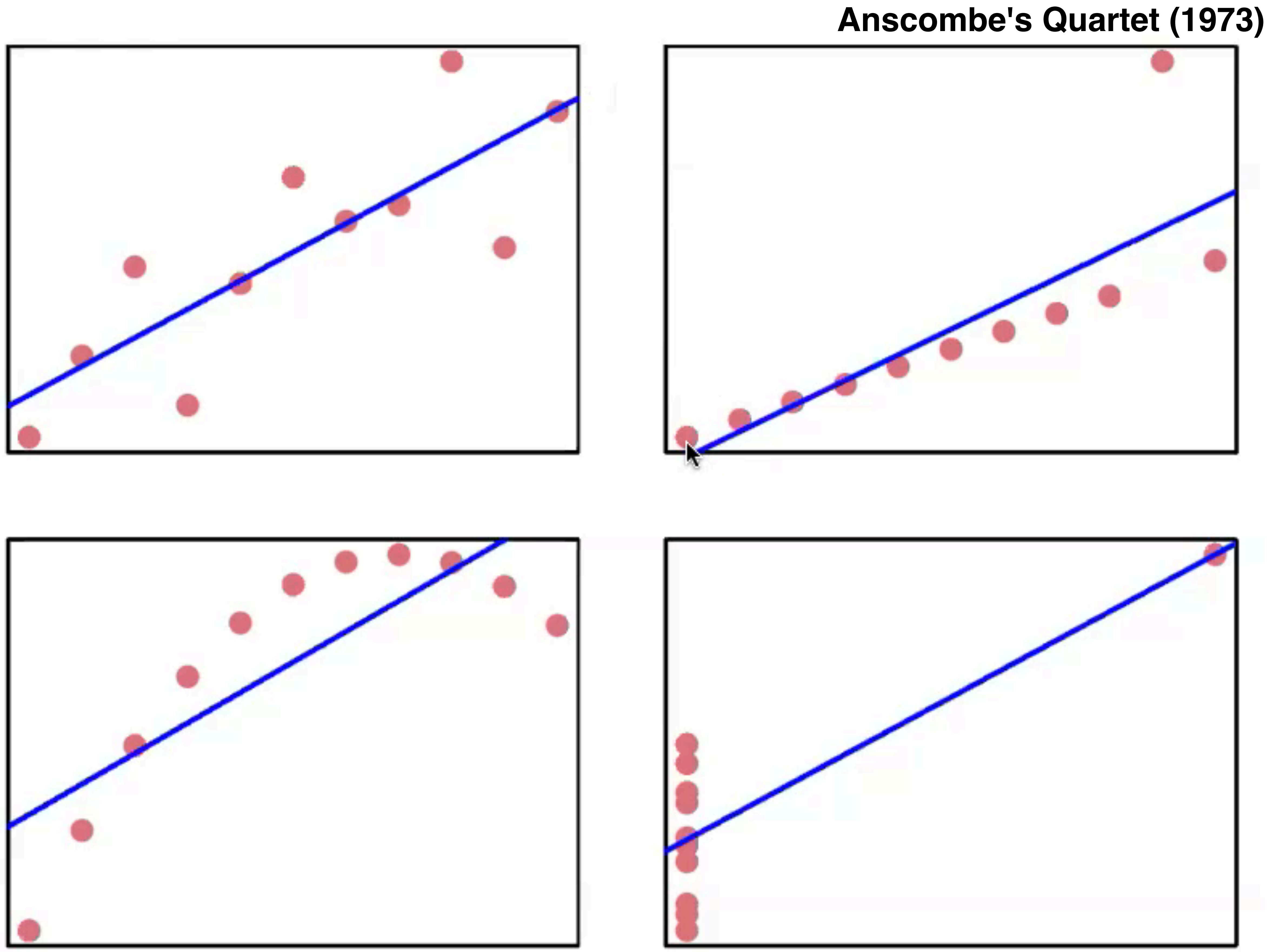
Regresión lineal

La componente de error

- En un mundo ideal, el modelo sería perfectamente lineal $\hat{y} = \beta_0 + \beta_1 x$.
- Pero en realidad, siempre tendremos una componente de error o residuo ϵ.
- $y = \beta_0 + \beta_1 x + \epsilon$
- El residuo ϵ será una variable aleatoria con **distribución normal** $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.
- ϵ_i es usualmente llamado **error irreducible**, siempre está presente. La idea de ajustar (o entrenar) un modelo no es eliminar esté error verdaderamente aleatorio, pues esto llevaría a un **sobreentrenamiento.**
- Siempre se supone, o se hacen pruebas, para ver que $E(\epsilon_i) = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2$. A esto se le llama **homocedasticidad.**
- Para un ejemplo práctico sobre esto, revisa el notebook de Jupyter de esta sección.

Regresión lineal

Ejemplos de algunos modelos ajustados



En todos los 4 casos, los valores de $\beta_0, \beta_1, r_{xy}, R^2$, son iguales.

Regresión lineal simple
Suposiciones

- **Libre de error.** Las variables predictoras son libre de error, i. e., que no son variables aleatorias.
- **Linealidad.** Se asume que la relación entre las variables independientes y la variable dependiente es lineal.
- **Varianza constante.** El error en la variable de respuesta es constante sobre los valores de entrada.
- **No hay multicolinealidad.** Ninguno de los parámetros de entrada son redundantes uno de otro.

Regresión lineal múltiple

Regresión lineal múltiple

Overview

- La regresión lineal que hemos visto tiene la forma
- $y_{model} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$
- Se podría pensar que todas las variables toman valores numéricos.
- Se pueden usar variables dummy para variables categóricas.

Regresión lineal múltiple

Definición

- En una regresión lineal múltiple se trabajan con más de 1 variable independiente.

- $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_n x_{ni} + \epsilon_i$

- O de forma equivalente, $y_i = \beta_0 + \left(\sum_{j=1}^{k} \beta_j x_{ij} \right) + \epsilon_i$ para k variables independientes.

Regresión lineal múltiple

Forma matricial

- O puesto de otra forma,
- $y_1 = \beta_0 + \sum_{j=1}^{k} \beta_j x_{1j} + \epsilon_1$
- $y_2 = \beta_0 + \sum_{j=1}^{k} \beta_j x_{2j} + \epsilon_2$
- $y_3 = \beta_0 + \sum_{j=1}^{k} \beta_j x_{3j} + \epsilon_3$
- $\vdots$
- $y_n = \beta_0 + \sum_{j=1}^{k} \beta_j x_{nj} + \epsilon_n$

$$\boxed{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}}$$

$$\downarrow$$

$$\boxed{\mathbf{Y} = \mathbf{X}\beta + \epsilon}$$

Regresión lineal múltiple

Mínimos cuadrados

- Queremos minimizar $\sum_{i=1}^{n} \hat{\epsilon}_i^{\,2} = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \ldots - \hat{\beta}_k x_{ki})^2$
- O puesto en forma matricial
- $\epsilon^T \epsilon = (Y^T - B^T X^T)(Y - BX)$
- Que se puede desarrollar y llegar a que:

$B^T X^T Y = Y^T X B$

- $\epsilon^T \epsilon = Y^T Y - 2B^T X^T Y + B^T X^T X B$

Regresión lineal múltiple

Mínimos cuadrados y parámetros óptimos

- Entonces, tomando la primera derivada,
- $\dfrac{\partial(\epsilon^T\epsilon)}{\partial B} = -2X^TY + 2X^TXB$
- Al igualar a cero. Ya estamos considerando solamente los estimadores de la minimización $\hat{\beta}$,
- $-2X^TY + 2X^TX\beta = 0 \Rightarrow X^TX\beta = X^TY$
- Multiplicando por la inversa de la matriz X^TX
- $\beta = (X^TX)^{-1}X^TY$

Regresión lineal múltiple

Error estándar residual

- RSE es la desviación estándar del término de error (desviación de la parte de los datos que el modelo no es capaz de explicar ya sea por falta de información o por parámetros adicionales.)
- En una regresión lineal simple: $RSE = \sqrt{\frac{\sum (y_i - \hat{y}_i(x_i))^2}{n-2}} = \sqrt{\frac{SSD}{n-2}}$ siendo n el número de datos.
- En una regresión lineal múltiple (k número de variables de entrada):
- $RSE = \sqrt{\frac{SSD}{n-k-1}}$

Regresión lineal múltiple

El problema de la multicolinealidad

La **multicolinealidad** es una relación de dependencia (lineal) fuerte entra más de dos variables de entrada en una regresión lineal múltiple. Esto puede llevar a lo siguiente:

- Las varianzas de los estimadores son muy grandes.
- Podría pasar que la hipótesis nula de que un estimador es cero sea aceptadas, aún cuando la correspondiente variable sea relevante.
- Los coeficientes estimados serán muy sensibles ante pequeños cambios en los datos.
- Este problema puede reducirse con la eliminación de los represores de las variables con alta relación lineal entre ellas.

Regresión lineal múltiple

Factor de inflación de la varianza

- Un criterio utilizado para verificar multicolinealidad es el uso del factor de inflación de la varianza (VIF), dado por:
- $VIF = \dfrac{1}{1 - R^2}$
- Cuando este valor supera a 5, se considera que se tiene un problema alto de multicolinealidad, que se tendrá que resolver de alguna manera.