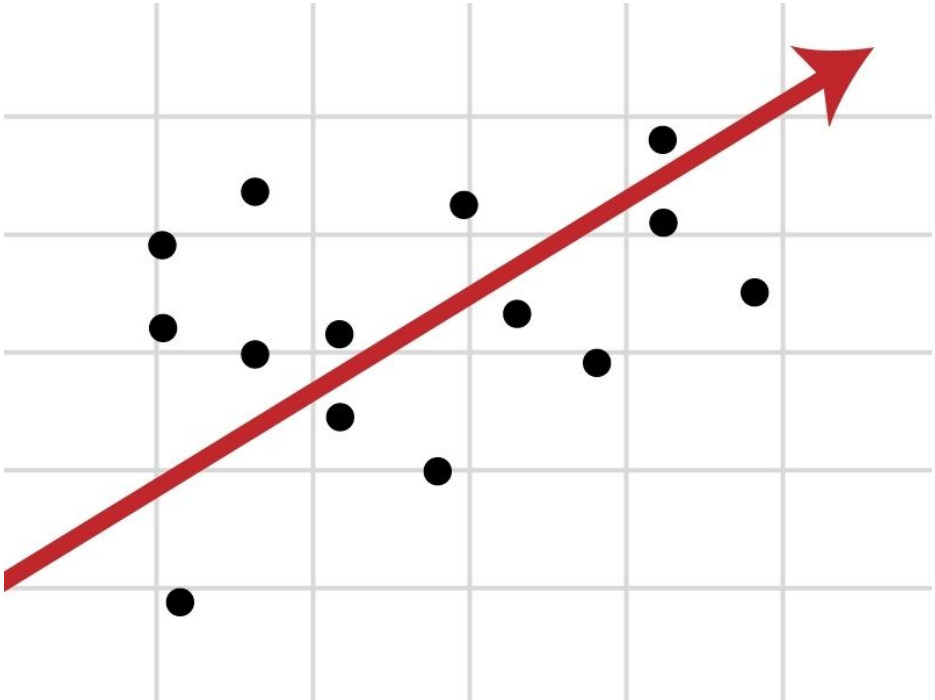
Aprendizaje de Máquinas

Primer Semestre 2025

Regresión Lineal Simple

Regresión Lineal Simple

Análisis de regresión. Suena como parte de la psicología freudiana. En realidad, una regresión es una herramienta estadística aparentemente ubicua que aparece en legiones de artículos científicos, y el análisis de regresión es un método para medir el vínculo entre dos o más fenómenos.



Extracto de:
https://news.mit.edu/2010/explained-reg-analysis-0316

Regresión Lineal Simple

Imagine que desea conocer la conexión entre los pies cuadrados de las casas y sus precios de venta. Una regresión traza tal vínculo, al hacerlo señalando “un efecto causal promedio,” como el economista del MIT Josh Angrist y su coautor Jorn-Steffen Pischke de la London School of Economics lo pusieron en su libro de 2009, “Mostly Harmless Econometrics.”

Regresión Lineal Simple

“Para comprender el concepto básico, tome la forma más simple de una regresión: una regresión lineal bivariada, que describe una relación inmutable entre dos (y no más) fenómenos. Ahora suponga que se está preguntando si hay una conexión entre el tiempo que los estudiantes de secundaria pasan haciendo la tarea de francés y las calificaciones que reciben. Estos tipos de datos se pueden trazar como puntos en un gráfico, donde el eje x es el número promedio de horas por semana que un estudiante estudia, y el eje y representa puntajes de examen de 100. Juntos, los puntos de datos normalmente se dispersan un poco en el gráfico. El análisis de regresión crea la línea única que mejor resume la distribución de puntos.”

Regresión Lineal Simple

Matemáticamente, la línea que representa una regresión lineal simple se expresa a través de una ecuación básica:

$Y = a_0 + a_1 X.$

Aquí X pasan horas estudiando por semana, la “variable independiente.”

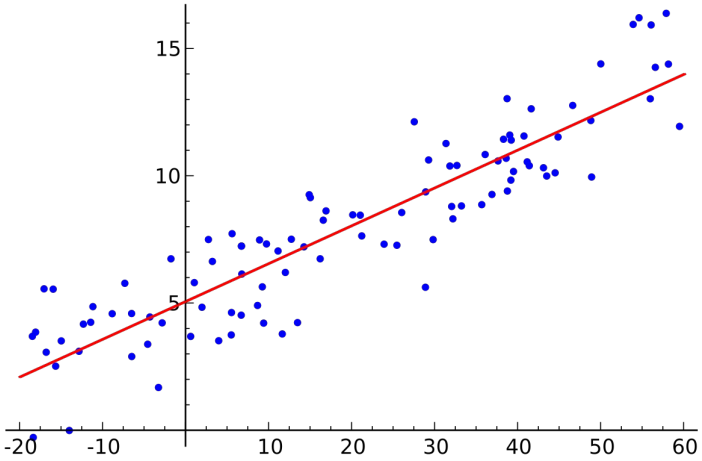
Y son los puntajes de los exámenes, la variable dependiente “,” ya que — creemos que — esos puntajes dependen del tiempo dedicado a estudiar.

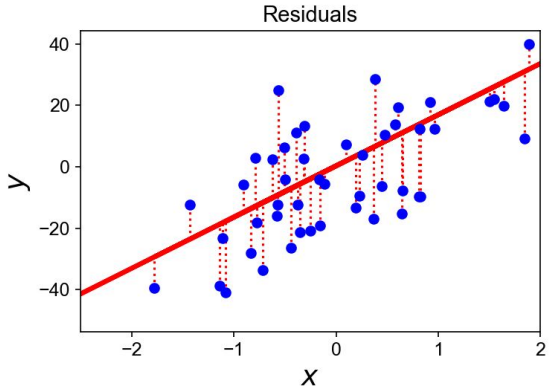
Además, a_0 es el y-intercept (el valor de Y cuando X es cero) y a_1 es la pendiente de la línea, que caracteriza la relación entre las dos variables.

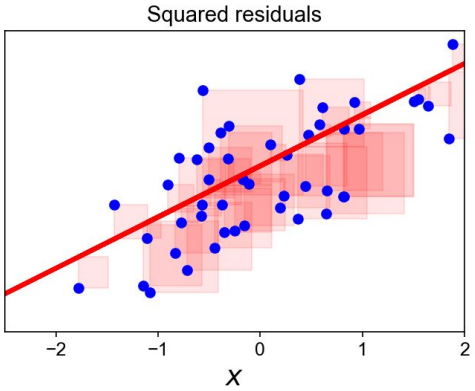
Regresión Lineal Simple- Mínimos Cuadrados Ordinarios

Usando dos ecuaciones ligeramente más complejas, la “ecuaciones normales” para la línea de regresión lineal básica, podemos conectar todos los números para X e Y, resolver para un a_0 y a_1 igualmente en realidad dibujar la línea.

Esa línea a menudo representa el agregado más bajo de los cuadrados de las distancias entre todos los puntos y él mismo, el método “Ordinary Least Squares” (OLS), mencionado en montañas de trabajos académicos.





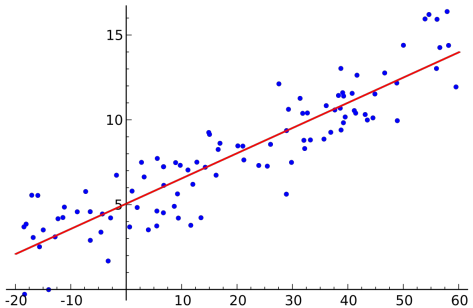


Regresión Lineal Simple- Mínimos Cuadrados Ordinarios

Para ver por qué OLS es lógico, imagine una línea de regresión que ejecuta 6 unidades por debajo de un punto de datos y 6 unidades por encima de otro punto; está a 6 unidades de los dos puntos, en promedio.

Ahora supongamos que una segunda línea corre 10 unidades por debajo de un punto de datos y 2 unidades por encima de otro punto; también está a 6 unidades de los dos puntos, en promedio. Pero si cuadramos las distancias involucradas, obtenemos diferentes resultados: $6^2 + 6^2 = 72$ en el primer caso, y $10^2 + 2^2 = 104$ en el segundo caso.

Así que la primera línea produce la cifra más baja — la “minimos cuadrados” — y es una reducción más consistente de la distancia desde los puntos de datos. (Los métodos adicionales, además de OLS, pueden encontrar la mejor línea para formas más complejas de análisis de regresión.)



Regresión Lineal Simple

Mínimos Cuadrados Ordinarios (MCO)

Es un **método de estimación** en regresión lineal.

Se utiliza para encontrar los coeficientes de una función lineal que mejor se ajusta a los datos minimizando la **suma de los errores al cuadrado**.

En términos matemáticos, si tenemos una variable dependiente Y y una variable independiente X, el modelo de regresión lineal es:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Los coeficientes β_0 y β_1 se eligen para minimizar la **suma de los residuos al cuadrado**:

$$\sum (Y_i - \hat{Y}_i)^2$$

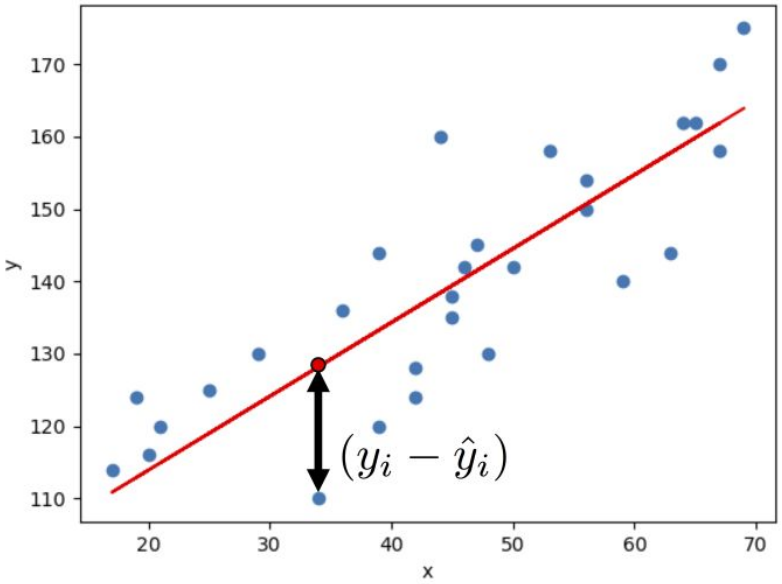
Uso principal: Se usa en modelos de regresión para encontrar la línea que mejor se ajusta a los datos.

Regresión Lineal Simple

A su vez, la distancia típica entre la línea y todos los puntos (a veces llamada “error estándar”) indica si el análisis de regresión ha capturado una relación que es fuerte o débil. Cuanto más cerca esté una línea de los puntos de datos, en general, más fuerte será la relación.

La historia se completa al considerar el porcentaje de error medio que permite evaluar la precisión del modelo de regresión lineal

La parte β0+β1X es el modelo de regresión lineal, siendo β0 y β1 los coeficientes de la regresión lineal y ϵ el error cometido por el modelo.



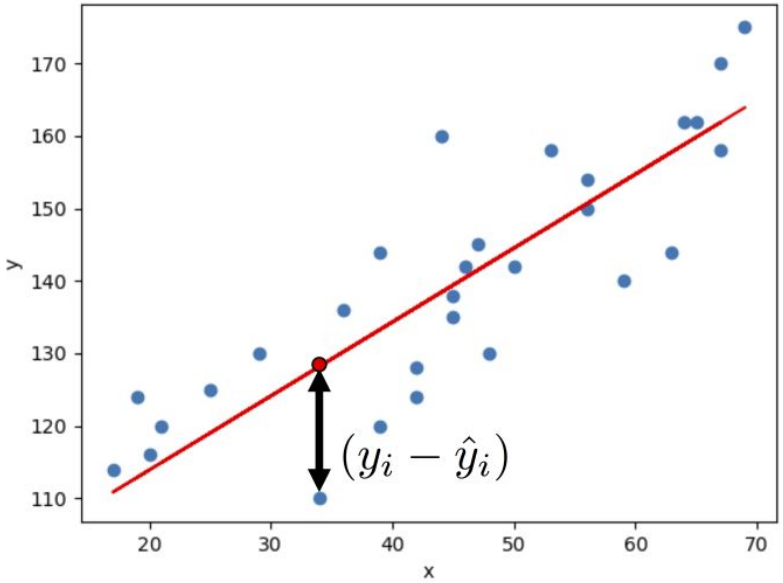
Regresión Lineal Simple

La pérdida permite medir la diferencia existente entre los datos reales (y) y los datos obtenidos tras realizar la Regresión Lineal (que en adelante llamaremos (ŷ).

Existen diferentes formas de definir matemáticamente esta pérdida, pero la más usada es a través del error cuadrático medio (ECM), definido de la siguiente manera:

N es el número total de datos que se tienen originalmente e Yi e ŷi son cada uno de los datos originales y los obtenidos a partir de la regresión, y la resta

se eleva al cuadrado para que todas las diferencias que están en la sumatoria sean positivas y no se cancelen entre sí.



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

Ejemplo Python

Demo:
https://colab.research.google.com/drive/13VwzY9pp13C4Ie8VkZEtgvn9VoiLqEEW#scrollTo=-tV4gGkqHSyc

Realizar actividad de regresión Lineal:

https://www.aprendemachinelearning.com/regresion-lineal-en-espanol-con-python/