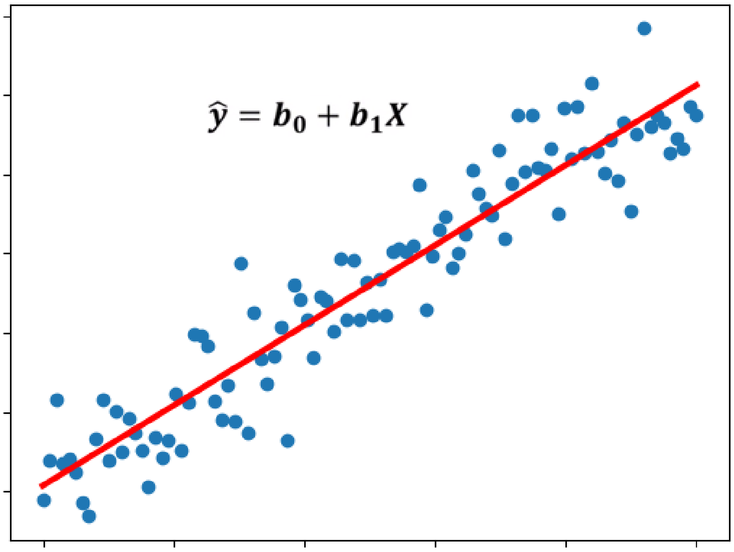
APRENDE CON ELI

ESTADÍSTICA

Fórmulas en regresión

DEFINICIÓN DEL MODELO DE REGRESIÓN

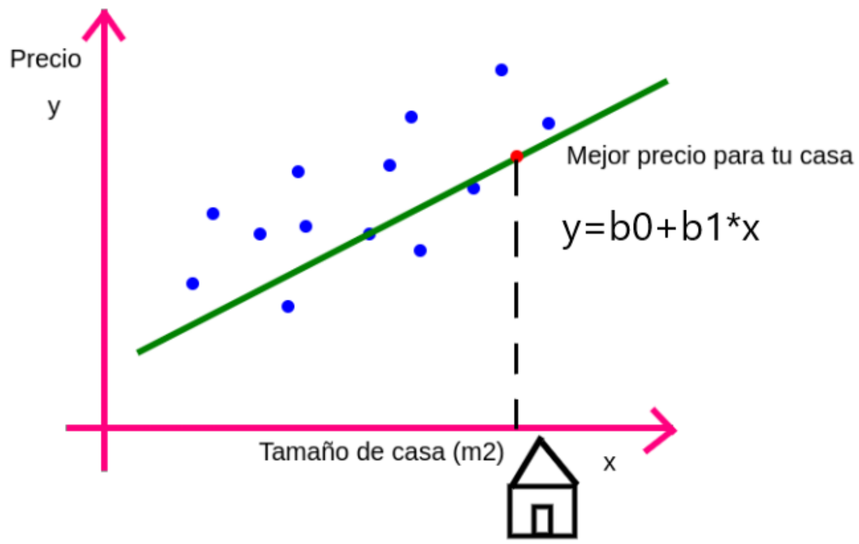
La idea del modelo de regresión simple es ajustar una línea recta que mejor represente el patrón de relación lineal positiva o negativa que siguen los datos que tenemos.



En la imagen anterior, los puntos azules son nuestros datos muestrales.

EJEMPLO

Por ejemplo, el eje X (horizontal) podría representar el tamaño en metros cuadrados de las viviendas y el eje Y (vertical) podría representar el precio de las viviendas. Si lo pensamos, es lógico que haya una relación directa y positiva entre el precio y el tamaño de la vivienda, mientras más grande más cuesta. Por eso, una vez que sabemos que hay una relación lineal (positiva o negativa) entre la variable respuesta que en nuestro ejemplo son los Precios y la variable explicativa que en nuestro ejemplo es el Tamaño, podemos estimar un modelo que represente esa relación, ese patrón.



VENTAJAS DEL MODELO DE REGRESIÓN

Y algunos dirán, si tenemos en nuestros datos todos los tamaños y precios de esas viviendas, ¿para qué queremos un modelo que los represente? Pues para datos futuros, viviendas nuevas de las que sólo podemos medir el tamaño, y nos gustaría poder estimar un precio para ellas correctamente.

Entonces las ventajas del modelo de regresión son muchas, pero principalmente se usa para hacer predicciones de la variable respuesta respecto a datos futuros, desconocidos, de la variable explicativa.

¿CÓMO ESTIMAMOS EL MODELO DE REGRESIÓN?

Una vez que tenemos claro la utilidad del modelo de regresión, vamos a ver cómo podemos estimar el modelo de una forma correcta, con el menor error posible y que ese modelo o esa recta represente lo mejor posible a nuestros datos muestrales.

Para estimar la recta de regresión, siendo Y la variable respuesta que queremos poner en función de X, se puede utilizar esta fórmula:

$$y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x})$$

Donde tendríamos que calcular y sustituir los valores de:

- $\bar{y}$: la media muestral de la variable respuesta Y
- $\bar{x}$: la media muestral de la variable explicativa X
- s_{xy}: la covarianza muestral entre X e Y
- s_x^2: la varianza muestral de X

Después de sustituir todo se despejaría la variable Y en la ecuación y quedaría en función de X.

Lo que se obtiene es justamente la ecuación de la recta de regresión que siempre será de esta forma:

$$y = \beta_0 + \beta_1 x$$

Donde β_0 es una constante (un número que no acompaña a la variable) y β_1 es el coeficiente que acompaña a la variable X.

INTERPRETACIÓN DE LOS COEFICIENTES

El β_0 se puede interpretar como el intercepto de la recta de regresión con el eje vertical. Y el β_1 se puede interpretar como la pendiente de la recta. El signo del β_1 nos dirá si la relación lineal es positiva (β_1 tiene signo +) o negativa (β_1 tiene signo -).

OTRA FORMA DE HALLAR LOS COEFICIENTES

Una forma alternativa (pero equivalente) de hallar los coeficientes es:

$$\beta_1 = r_{xy}\frac{s_y}{s_x}$$

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

Es decir, para esto necesitamos las dos desviaciones típicas de ambas variables: s_x y s_y. También el coeficiente de correlación entre ambas: r_{xy}. Y simplemente, hallamos primero el β_1 sustituyendo la correlación y las desviaciones típicas, y luego hallamos el β_0 sustituyendo el β_1 que acabamos de calcular, y las dos medias muestrales $\bar{x}$ y $\bar{y}$. Por último, cuando se tengan β_0 y β_1, la ecuación de regresión será:

$$y = \beta_0 + \beta_1 x$$

Con esta ecuación ya calculada, en nuestro ejemplo, si quisiéramos saber el precio de nuestra vivienda, sólo tendríamos que pasarle el valor de X, el tamaño de la vivienda, y con la ecuación podríamos calcular el precio estimado.