Regresión y Correlación Lineal Simple

El modelo de regresión simple tiene por objetivo explicar el comportamiento de una variable endógena, explicada o dependiente, que se designa como Y, utilizando la información proporcionada por los valores tomados por una variable exógena, explicativa o independiente, denominada X.

Regresión	Correlación	Lineal	Simple
Encontrar la ecuación del modelo que mejor ajuste a los datos observados	Obtener el grado de asociación entre las variables	La ecuación se puede explicar por medio de la ecuación de la recta	Hay una sola variable independiente en el modelo

Si X es la variable explicativa e Y es la variable explicada entonces:

Modelo Poblacional

$$Y = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Modelo Muestral

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i + e_i$$

Supuestos del modelo lineal simple

1. La variable aleatoria ε_i es estadísticamente independiente de los valores x_i.
2. La variable aleatoria ε_i tiene distribución normal
3. La esperanza de ε_i vale 0. $E(\varepsilon_i) = 0$.
4. Para todo par se tiene que son estadísticamente independientes por lo tanto la $Cov(\varepsilon_i, \varepsilon_j) = 0$.
5. La $V(\varepsilon_i) = \sigma^2$ es finita y constante para todo i.

Recta de regresión muestral

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$$

$$e_i = y_i - \hat{y}_i$$

Dada una serie de valores muestrales de x e y se utiliza el método de mínimos cuadrados ordinarios MCO para encontrar la recta que mejor ajuste (la mejor ecuación) la recta va a ser la que reduce al mínimo las desviaciones cuadradas entre los valores estimados y real de la variable dependiente para los datos muestrales.

Una vez formulada la ecuación de regresión, puede servir para estimar el valor de la variable dependiente dado el valor de la variable independiente originalmente muestreada, ya que no existe base estadística para suponer que la línea de regresión es adecuada fuera de estos límites.

Método de mínimos cuadrados

$Minimizar\ \sum_{i=1}^{n} e_i^2$

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

$Minimizar\ \sum_{i=1}^{n} \left(y_i - (\hat{b}_0 + \hat{b}_1 x_i)\right)^2$

$Minimizar\ \sum_{i=1}^{n} \left(y_i - \ \hat{b}_0 - \hat{b}_1 x_i\right)^2$

Función que se quiere minimizar en el modelo de regresión lineal simple.

$$\varphi(\hat{b}_0, \hat{b}_1) = \sum_{i=1}^{n} (y_i - \ \hat{b}_0 - \hat{b}_1 x_i)^2$$

La condición de mínimo es que las derivadas parciales respecto de $\hat{b}_0\ y\ \hat{b}_1$sean cero.

$$\frac{\delta\varphi}{\delta\hat{b}_0} = 0$$

$$\frac{\delta\varphi}{\delta\hat{b}_1} = 0$$

$$\frac{\delta\varphi}{\delta\hat{b}_0} = 2\sum_{i=1}^{n} (y_i - \ \hat{b}_0 - \hat{b}_1 x_i)(-1) = 0$$

$$\frac{\delta\varphi}{\delta\hat{b}_1} = 2\sum_{i=1}^{n} (y_i - \ \hat{b}_0 - \hat{b}_1 x_i)(-x_i) = 0$$

Dividiendo por 2 a ambas ecuaciones, aplicando distributiva respecto de la sumatoria y de la derivada:

$$\frac{\delta\varphi}{\delta\hat{b}_0} = -\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} \ \hat{b}_0 + \sum_{i=1}^{n} \hat{b}_1 x_i = 0$$

$$\frac{\delta\varphi}{\delta\hat{b}_1} = -\sum_{i=1}^{n} y_i x_i + \sum_{i=1}^{n} \ \hat{b}_0 x_i + \sum_{i=1}^{n} \hat{b}_1 x_i^2 = 0$$

$$\sum_{i=1}^{n} \ \hat{b}_0 + \sum_{i=1}^{n} \hat{b}_1 x_i = \sum_{i=1}^{n} y_i$$

$$\sum_{i=1}^{n} \ \hat{b}_0 x_i + \sum_{i=1}^{n} \hat{b}_1 x_i^2 = \sum_{i=1}^{n} y_i\, x_i$$

$$\sum_{i=1}^{n} \ \hat{b}_0 \ = \ \hat{b}_0 + \ \hat{b}_0 + \cdots + \ \hat{b}_0 = n\ \hat{b}_0$$

$$\sum_{i=1}^{n} \hat{b}_1 x_i = \hat{b}_1 x_1 + \hat{b}_1 x_2 + \cdots + \hat{b}_1 x_n = \hat{b}_1 \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} \hat{b}_0 x_i = \hat{b}_0 x_1 + \hat{b}_0 x_2 + \cdots + \hat{b}_0 x_n = \hat{b}_0 \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} \hat{b}_1 x_i^2 = \hat{b}_1 x_1^2 + \hat{b}_1 x_2^2 + \cdots + \hat{b}_1 x_n^2 = \hat{b}_1 \sum_{i=1}^{n} x_i^2$$

Sistema de Ecuaciones Normales de Gauss

$$n \hat{b}_0 + \hat{b}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\hat{b}_0 \sum_{i=1}^{n} x_i + \hat{b}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

¿Cómo se obtienen los estimadores $\hat{b}_0$ y $\hat{b}_1$?

Utilizando determinantes:

$$\hat{b}_1 = \frac{\begin{vmatrix} n & \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} y_i x_i \end{vmatrix}}{\begin{vmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{vmatrix}} = \frac{n \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$\hat{b}_0 = \frac{\begin{vmatrix} \sum_{i=1}^{n} y_i & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} y_i x_i & \sum_{i=1}^{n} x_i^2 \end{vmatrix}}{\begin{vmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{vmatrix}} = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} y_i x_i \sum_{i=1}^{n} x_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

Se puede obtener en primer lugar $\hat{b}_1$ y luego $\hat{b}_0$ en función de $\hat{b}_1$.

$$\frac{n \hat{b}_0}{n} + \hat{b}_1 \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{n} y_i}{n}$$

$$\frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

$$\frac{\sum_{i=1}^{n} y_i}{n} = \bar{y}$$

$$\hat{b}_0 + \hat{b}_1 \bar{x} = \bar{y}$$

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$$

Covarianza

$$\hat{b}_1 = \frac{\mathrm{Cov(x,y)}}{\mathrm{V(x)}} = \frac{\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}}{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$\widehat{b}_0 = \bar{y} - \hat{b}_1\bar{x}$$

Ecuación de la recta de regresión

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$$