1

Regresión lineal simple

1.1 Introducción

Fue introducido por Francis Galton (1908). El modelo de regresión lineal simple está formado típicamente por:

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

Donde:

y	=	variable dependiente o variable de respuesta.
x	=	variable independiente o explicativo o predictor.
β_0	=	intercepto y.
β_1	=	pendiente.
ϵ	=	error aleatorio.

Una presentación más general de un modelo de regresión sería:

$$y = E(y) + \epsilon,$$

Donde: $E(y)$ es la esperanza matemática de la variable respuesta.Cuando $E(y)$ es una combinación lineal de las variables explicativas $x_1, x_2, \ldots, x_k$ la regresión es una regresión lineal. Con $E(\epsilon_i) = 0$ y $Var(\epsilon_i) = \sigma^2$. Todos los ϵ_i son independientes.

Ahora debemos hallar buenos estimadores para β_0 y β_1.

1.2 Estimaciones por mínimos cuadrados

El principal objetivo de los mínimos cuadrados para un modelo de regresión lineal simple es hallar los estimadores b_0 y b_1 tales que la suma de la distancia al cuadrados de la respuesta real y_i y las respuesta de las pronosticadas $\hat{y}_i = \beta_0 + \beta_1 x_i$ alcanza el mínimo entre todas las opciones posibles de coeficientes de regresión β_0 y β_1. Es decir,

$$(b_0, b_1) = \arg \min_{(\beta_0, \beta_1)} \sum_{i=1}^{n} \left[\beta_0 + \beta_1 x_i\right]^2.$$

Matemáticamente, las estimaciones de mínimos cuadrados de la regresión lineal simple se obtienen resolviendo el siguiente sistema:

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_i x_i)]^2 = 0 \tag{1.1}$$

$$\frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_i x_i)]^2 = 0 \tag{1.2}$$

Supongamos que b_0 y b_1 son soluciones del sistema de arriba, podemos describir la relación entre x e y por la regresión lineal $\hat{y} = b_0 + b_1 x$, el cual es llamado la **recta de regresión ajustada**. Es más conveniente resolver para b_0 y b_1 usando el modelo lineal centralizado:

$$y_i = \beta_0 + \beta_1 \overline{x} - \beta_1 \overline{x} + \beta_1 x_i + \epsilon_i \quad \Rightarrow \quad y_i = \beta_0^* + \beta_1 (x_i - \overline{x}) + \epsilon_i,$$

donde $\beta_0 = \beta_0^* - \beta_1 \overline{x}$. Necesitamos resolver para

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

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

Realizando la derivada parcial para β_0 y β_1 tenemos

$$\sum_{i=1}^{n} [y_i - (\beta_0^* + \beta_1 (x_i - \overline{x}))] = 0$$

$$\sum_{i=1}^{n} [y_i - (\beta_0^* + \beta_1 (x_i - \overline{x}))] (x_i - \overline{x}) = 0$$

Notemos que

$$\sum_{i=1}^{n} y_i = n\beta_0^* + \sum_{i=1}^{n} \beta_1 (x_i - \overline{x}) = n\beta_0^* \tag{1.3}$$

Por lo tanto, tenemos

$$\beta_0^* = \frac{1}{n} \sum_{i=1}^{n} y_i = \overline{y}.$$

Luego, sustituyendo β_0^* por $\overline{y}$ en (2.3) obtenemos

$$\sum_{i=1}^{n} [y_i - (\overline{y} + \beta_1 (x_i - \overline{x}))] (x_i - \overline{x}) = 0$$

Después denotamos b_0 y b_1 las soluciones de los sistemas (2.1) y (2.2). Ahora, es fácil ver que

$$b_1 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \frac{\frac{1}{n} \sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})}{\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}} \tag{1.4}$$

y

$$b_0 = b_0^* - b_1 \overline{x} = \overline{y} - b_1 \overline{x} \tag{1.5}$$

El valor ajustado de la regresión lineal simple es definida como $\hat{y} = \beta_0 + \beta_1 x_i$. La diferencia entre y_i y el valor ajustado $\hat{y}_i$ es $e_i = y_i - \hat{y}_i$, que se refiere al residuo de la regresión. Los residuos de regresión se pueden calcular a partir de las respuestas observadas y_i y los valores ajustados y_i, por lo tanto, los residuos son observables. Cabe señalar que el término de error ϵ_i en el modelo de regresión no es observable. El error de regresión es la cantidad por la cual una observación difiere de su valor esperado; este último se basa en la población total de la que se eligió aleatoriamente la unidad estadística. El valor esperado, el promedio de toda la población, normalmente no es observable.

Un residual, por otro lado, es una estimación observable de un error no observable. El caso más simple implica una muestra aleatoria de n hombres cuyas alturas se miden. El promedio de la muestra se utiliza como una estimación del promedio de la población. Entonces, la diferencia entre la altura de cada hombre de la muestra y el promedio de la población no observable es un error, y la diferencia entre la altura de cada hombre de la muestra y el promedio de la muestra observable es un residuo. Dado que los residuales son observables, podemos usar los residuales para estimar el error del modelo no observable. La discusión detallada se proporcionará más adelante.

1.3 Propiedades estadísticas de la estimación por mínimos cuadrados

Primero discutiremos las propiedades estadísticas sin el supuesto de distribución del termino de error. Pero asumiremos que $\mathrm{E}(\epsilon_i) = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2$ y ϵ_i para $i = 1, 2, \ldots, n$ son independientes.

Teorema 1.1 El estimador de mínimos cuadrados b_0 es un estimador insesgado de β_0.

Demostración.-

$$\begin{aligned}
E(b_0) &= E(\overline{y} - b_1\overline{x}) \\
&= E\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) - E\left(b_1\overline{x}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E(y_i) - \overline{x}E(b_1) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i) - \beta_1\overline{x} \\
&= \frac{1}{n}\sum_{i=1}^{n}\beta_0 + \beta_1\frac{1}{n}\sum_{i=1}^{n} x_i - \beta_1\overline{x} \\
&= \frac{1}{n}\sum_{i=1}^{n}\beta_0 + \beta_1\frac{1}{n}\sum_{i=1}^{n} x_i - \beta_1\frac{1}{n}\sum_{i=1}^{n} x_i \\
&= \frac{n\beta_0}{n} \\
&= \beta_0.
\end{aligned}$$

■

Teorema 1.2 El estimador de mínimos cuadrados b_1 es un estimador insesgado de β_1.

Demostración.-

$$\begin{aligned}
\mathrm{E}(b_1) &= \mathrm{E}\left(\frac{S_{xy}}{S_{xx}}\right) \\
&= \frac{1}{S_{xx}}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})(x_i-\overline{x})\right] \\
&= \frac{1}{s_{xx}}\frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n}(x_i-\overline{x})y_i-\sum_{i=1}^{n}(x_i-\overline{x})\overline{y}\right] \\
&= \frac{1}{s_{xx}}\frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n}(x_i-\overline{x})y_i-\overline{y}\sum_{i=1}^{n}(x_i-\overline{x})\right] \quad \text{ya que } \overline{y} \text{ es constante.}
\end{aligned}$$

Sabemos que $\sum_{i=1}^{n}(x_i-\overline{x}) = \sum_{i=1}^{n}x_i - n\overline{x} = \sum_{i=1}^{n}x_i - n\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right) = 0$, por lo que

$$\begin{aligned}
&= \frac{1}{S_{xx}}\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})\,\mathrm{E}(y_i) \\
&= \frac{1}{S_{xx}}\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(\beta_0+\beta_1 x_i) \\
&= \frac{1}{S_{xx}}\frac{1}{n}\left[\beta_0\sum_{i=1}^{n}(x_i-\overline{x})+\sum_{i=1}^{n}(x_i-\overline{x})\beta_1 x_i\right] \\
&= \frac{1}{S_{xx}}\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})\beta_1 x_i \\
&= \frac{1}{S_{xx}}\frac{1}{n}\left[\sum_{i=1}^{n}(x_i-\overline{x})\,\beta_1 x_i-\sum_{i=1}^{n}(x_i-\overline{x})\,\beta_1\overline{x}\right] \\
&= \frac{1}{S_{xx}}\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})\beta_1(x_i-\overline{x}) \\
&= \frac{1}{S_{xx}}\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\beta_1 \\
&= \frac{S_{xx}}{S_{xx}}\beta_1 \\
&= \beta_1.
\end{aligned}$$

■

Teorema 1.3 $\text{Var}(b_1) = \dfrac{\sigma^2}{nS_{xx}}$.

Demostración.- Usando la propiedad $Var(X) = a^2 Var(X)$ con respecto a y, se tiene

$$\begin{aligned}
\text{Var}(b_1) &= \text{Var}\left(\frac{S_{xy}}{S_{xx}}\right) \\
&= \left(\frac{1}{S_{xx}}\right)^2 \text{Var}\left[\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})(x_i-\overline{x})\right] \\
&= \frac{1}{S_{xx}^2}\,\text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} y_i(x_i-\overline{x})\right] \\
&= \frac{1}{S_{xx}^2}\frac{1}{n^2}\sum_{i=1}^{n}(x_i-\overline{x})^2\,\text{Var}(y_i) \\
&= \frac{1}{S_{xx}^2}\frac{1}{n^2}\sum_{i=1}^{n}(x_i-\overline{x})^2\sigma^2 \\
&= \frac{\sigma^2}{nS_{xx}}.
\end{aligned}$$

■

Teorema 1.4 **El estimador de mínimos cuadrados b_1 e $\overline{y}$ no están correlacionados.** Bajo el supuesto de normalidad de y_i para $i = 1, 2, \ldots, n$, b_1 e $\overline{y}$ se distribuyen normalmente y son independientes.

Demostración.-

$$\begin{aligned}
\text{Cov}(b_1, \overline{y}) &= \text{Cov}\left(\frac{S_{xy}}{S_{xx}}, \overline{y}\right) \\
&= \frac{1}{S_{xx}}\,\text{Cov}(S_{xy}, \overline{y}) \\
&= \frac{1}{nS_{xx}}\,\text{Cov}\left[\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}), \overline{y}\right] \\
&= \frac{1}{n^2S_{xx}}\,\text{Cov}\left[\sum_{i=1}^{n}(x_i-\overline{x})y_i, \sum_{i=1}^{n} y_i\right] \\
&= \frac{1}{n^2S_{xx}}\sum_{i=1}^{n}(x_i-\overline{x})\,\text{Cov}(y_i, y_j)
\end{aligned}$$

Notemos que $E(\epsilon_i) = 0$ y ϵ_i son independientes. De donde podemos escribir

$$\text{Cov}(y_i, y_j) = E\left\{[y_i - E(y_j)][y_j - E(y_j)]\right\} = E(\epsilon_i, \epsilon_j) = \begin{cases} \sigma^2 & si \quad i = j \\ 0 & si \quad i \neq j. \end{cases}$$

Concluimos que

$$\text{Cov}(b_1, \overline{y}) = \frac{1}{n^2} S_{xx} \sum_{i=1}^{n} (x_i - \overline{x})\sigma^2 = 0.$$

Recuerde que la correlación cero es equivalente a la independencia entre dos variables normales. Por lo tanto, concluimos que b_1 e $\overline{y}$ son independientes. ■

Teorema 1.5 $\text{Var}(b_0) = \left(\frac{1}{n} + \frac{\overline{x}^2}{nS_{xx}} \right) \sigma^2.$

Demostración.-

$$\begin{aligned} \text{Var}(b_0) &= \text{Var}(\overline{y} - b_1 \overline{x}) \\ &= \text{Var}(\overline{y}) + (\overline{x})^2 \, \text{Var}(b_1) \\ &= \frac{\sigma^2}{n} + \overline{x}^2 \frac{\sigma^2}{nS_{xx}} \\ &= \left(\frac{1}{n} + \frac{\overline{x}^2}{nS_{xx}} \right) \sigma^2 \end{aligned}$$

■