Estimación de máxima verosimilitud

La estimación de máxima verosimilitud (MLE) es una técnica utilizada para estimar los parámetros de una distribución determinada, utilizando algunos datos observados. Por ejemplo, si se sabe que una población sigue una distribución normal pero se desconocen la media y la varianza, se puede utilizar MLE para estimarlas utilizando una muestra limitada de la población, encontrando valores particulares de la media y la varianza para que la observación sea el resultado más probable que haya ocurrido.

Sean $x_1, x_2,\ldots, x_n$ observaciones de n variables aleatorias independientes e idénticamente distribuidas extraídas de una distribución de probabilidad p0 donde se sabe que p0 pertenece a una familia de distribuciones p que dependen de algunos parámetros θ. Por ejemplo, se podría saber que p0 pertenece a la familia de distribuciones normales p, que dependen de los parámetros σ (desviación estándar) y μ (media), y $x_1, x_2,\ldots, x_n$ serían observaciones de p0. El objetivo de MLE es maximizar la función de verosimilitud (**likelihood function**):

$$\mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta}) \equiv p(\boldsymbol{x}|\boldsymbol{\theta}) = p(x_1|\boldsymbol{\theta}) \times p(x_2|\boldsymbol{\theta}) \times \ldots \times p(x_n|\boldsymbol{\theta}) = \prod_{\boldsymbol{i=1}}^{\boldsymbol{n}} p(x_i|\boldsymbol{\theta})$$

A menudo, es más fácil trabajar con la función de probabilidad logarítmica (**log-likelihood**):

$$\ell(\boldsymbol{x}|\boldsymbol{\theta}) = \log\big(\mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta})\big) = \sum_{i=1}^{n} \log p(x_i|\boldsymbol{\theta})$$

Hay varias maneras en que MLE podría terminar funcionando: podría descubrir parámetros θ en términos de las observaciones dadas, podría descubrir múltiples parámetros que maximicen la función de verosimilitud, podría descubrir que no existe un máximo. No existe una forma cerrada al máximo y es necesario un análisis numérico para encontrar un MLE.

Aunque los MLE no son necesariamente óptimos (en el sentido de que existen otros algoritmos de estimación que pueden lograr mejores resultados), tienen varias propiedades atractivas, la más importante de las cuales es la coherencia: una secuencia de MLE (en un número creciente de observaciones) convergen al valor real de los parámetros.

El modelo de regresión lineal

El objetivo es estimar los parámetros del modelo de regresión lineal: $y_i = \theta_0 + \theta_1 x_i + e_i$

Suponiendo que la hipótesis es $\hat{y}_i \equiv h(\boldsymbol{x}|\boldsymbol{\theta}) = \theta_0 + \theta_1 x_i$, entonces el modelo de regresión lineal es:

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

Cuando y_i es la variable dependiente, x_i es un vector de 1 x K de regresores, $\boldsymbol{\theta}$ es el vector K x 1 de los coeficientes de regresión a estimar y e_i es un término de error no observable.

La muestra está compuesta por observaciones N IID (Independientes e Idénticamente Distribuidas) (x_i, y_i). El modelo de regresión lineal en notación matricial es:

Y el total de respueta

$$\begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_i \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_i \\ \vdots \\ e_N \end{bmatrix}$$

donde

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_N \end{bmatrix} \quad ; \quad \boldsymbol{A} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_i \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix} \quad ; \quad \boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} \quad ; \quad \boldsymbol{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_i \\ \vdots \\ e_N \end{bmatrix}$$

Las ecuaciones de regresión se pueden escribir en forma de matriz como:

$$\boldsymbol{y} = A\boldsymbol{\theta} + \boldsymbol{e}$$

$$\boldsymbol{y} = \hat{\boldsymbol{y}} + \boldsymbol{e}$$

$$\boldsymbol{e} = \boldsymbol{y} - A\boldsymbol{\theta}$$

donde $\hat{\boldsymbol{y}} = A\boldsymbol{\theta}$ es la hipótesis del modelo lineal, el vector N x 1 de las observaciones de la variable dependiente se denota por $\boldsymbol{y}$, la matriz N x K de regresores (matriz de diseño) se denota por $\boldsymbol{A}$ y el vector N x 1 de los términos de error se denotan por $\boldsymbol{e}$.

La función de probabilidad

Dado que las observaciones de la muestra son independientes, la probabilidad de la muestra es igual al producto de las probabilidades de las observaciones individuales:

$$\mathcal{L}(\boldsymbol{\theta}, \sigma^2; \boldsymbol{y}, \boldsymbol{A}) = \prod_{i=1}^{N} f_Y(x_i|\boldsymbol{\theta})$$

$$\mathcal{L}(\boldsymbol{\theta}, \sigma^2; \boldsymbol{y}, \boldsymbol{A}) = \prod_{i=1}^{N} (2\pi\sigma^2)^{-1/2} exp\left[-\frac{1}{2}\left(\frac{y_i - \hat{y}_i}{\sigma}\right)^2\right]$$

donde el modelo de regresión lineal (hipótesis) es $\hat{y}_i = \theta_0 + \theta_1 x_i$, es decir $y_i = \hat{y}_i + e_i$, entonces el termino de error $e_i = y_i - \hat{y}_i$. Entonces la función de probabilidad (likelihood function) se define como:

$$\mathcal{L}(\boldsymbol{\theta}, \sigma^2; \boldsymbol{y}, \boldsymbol{A}) = (2\pi\sigma^2)^{-N/2} \, exp\left[-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{y_i - \hat{y}_i}{\sigma}\right)^2\right]$$

$$\mathcal{L}(\boldsymbol{\theta}, \sigma^2; \boldsymbol{y}, \boldsymbol{A}) = (2\pi\sigma^2)^{-N/2} \, exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2\right]$$

→ suma de los errores al cuadrado

o bien sustituyendo $e_i = y_i - \hat{y}_i$ en la ecuación anterior, obtenemos:

$$\mathcal{L}(\boldsymbol{\theta}, \sigma^2; \boldsymbol{y}, \boldsymbol{A}) = (2\pi\sigma^2)^{-N/2} \, exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N} e_i^2\right]$$

Si $SSE \equiv \sum_{i=1}^{N} e_i^2 = \|\boldsymbol{e}\|_F^2$ es la suma de los errores al cuadrado y sustituyendo en la ecuación anterior, obtenemos:

$$\mathcal{L}(\boldsymbol{\theta}, \sigma^2; \boldsymbol{y}, \boldsymbol{A}) = (2\pi\sigma^2)^{-N/2} \, exp\left[-\frac{SSE}{2\sigma^2}\right]$$

O bien

$$\mathcal{L}(\boldsymbol{\theta}, \sigma^2; \boldsymbol{y}, \boldsymbol{A}) = (2\pi\sigma^2)^{-N/2} \, exp\left[-\frac{\|\boldsymbol{e}\|_F^2}{2\sigma^2}\right]$$

La función de probabilidad logarítmica (log-likelihood function)

Se obtiene tomando el logaritmo natural de la función de probabilidad:

$$\ell(\boldsymbol{\theta}, \sigma^2; \boldsymbol{y}, \boldsymbol{A}) = ln\left[\mathcal{L}(\boldsymbol{\theta}, \sigma^2; \boldsymbol{y}, \boldsymbol{A})\right]$$

$$= ln\left((2\pi\sigma^2)^{-N/2} \, exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2\right]\right)$$

$$= ln\left((2\pi\sigma^2)^{-N/2}\right) + ln\left(exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2\right]\right)$$

$$= -\frac{N}{2}\, ln(2\pi) - \frac{N}{2}\, ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

Entonces la función de probabilidad logarítmica (log-likelihood function) se define como:

$$\ell(\boldsymbol{\theta}, \sigma^2; \boldsymbol{y}, \boldsymbol{A}) = -\frac{N}{2}\, ln(2\pi) - \frac{N}{2}\, ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

o bien sustituyendo $e_i = y_i - \hat{y}_i$ en la ecuación anterior, obtenemos:

$$= -\frac{N}{2}\, ln(2\pi) - \frac{N}{2}\, ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} e_i^2$$

O bien

$$= -\frac{N}{2}\, ln(2\pi) - \frac{N}{2}\, ln(\sigma^2) - \frac{SSE}{2\sigma^2}$$

Entonces la función de probabilidad logarítmica (log-likelihood function) se expresa como:

$$\ell(\boldsymbol{\theta}, \sigma^2; \boldsymbol{y}, \boldsymbol{A}) = -\frac{N}{2}\, ln(2\pi) - \frac{N}{2}\, ln(\sigma^2) - \frac{\|\boldsymbol{e}\|_F^2}{2\sigma^2}$$

o bien la función de probabilidad logarítmica negativa (negative log-likelihood function) se define como:

$$n\ell\ell(\boldsymbol{\theta}, \sigma^2; \boldsymbol{y}, \boldsymbol{A}) \equiv -\ell(\boldsymbol{\theta}, \sigma^2; \boldsymbol{y}, \boldsymbol{A}) = \frac{N}{2}\, ln(2\pi) + \frac{N}{2}\, ln(\sigma^2) + \frac{\|\boldsymbol{e}\|_F^2}{2\sigma^2}$$

o bien la función de probabilidad logarítmica negativa promedio (average negative log-likelihood function) se define como:

$$an\ell\ell(\boldsymbol{\theta}, \sigma^2; \boldsymbol{y}, \boldsymbol{A}) \equiv -\ell(\boldsymbol{\theta}, \sigma^2; \boldsymbol{y}, \boldsymbol{A})/N = \frac{1}{2}\, ln(2\pi) + \frac{1}{2}\, ln(\sigma^2) + \frac{1}{2\sigma^2}\frac{\|\boldsymbol{e}\|_F^2}{N}$$

donde $MSE \equiv \frac{SSE}{N} = \frac{\|\boldsymbol{e}\|_F^2}{N}$, sustituyendo en la ecuación anterior se obtiene:

$$an\ell\ell(\boldsymbol{\theta}, \sigma^2; \boldsymbol{y}, \boldsymbol{A}) = \frac{1}{2}\, ln(2\pi) + \frac{1}{2}\, ln(\sigma^2) + \frac{1}{2\sigma^2}\, MSE$$

→ Utilizada para optimizar

Los estimadores de máxima verosimilitud

Los estimadores resuelven el siguiente problema de maximización

Los estimadores resuelven el siguiente problema de maximización

$$\widehat{\boldsymbol{\theta}}, \hat{\sigma}^2 = arg \max_{\boldsymbol{\theta}, \sigma^2} \ell(\boldsymbol{\theta}, \sigma^2; \boldsymbol{y}, \boldsymbol{A})$$

Las condiciones de primer orden para un máximo son

$$\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}, \sigma^2; \boldsymbol{y}, \boldsymbol{A}) = 0$$

$$\nabla_{\sigma^2} \ell(\boldsymbol{\theta}, \sigma^2; \boldsymbol{y}, \boldsymbol{A}) = 0$$

Los estimadores resuelven el siguiente problema de minimización

$$\widehat{\boldsymbol{\theta}}, \hat{\sigma}^2 = arg \min_{\boldsymbol{\theta}, \sigma^2} n\ell\ell(\boldsymbol{\theta}, \sigma^2; \boldsymbol{y}, \boldsymbol{A})$$

Las condiciones de primer orden para un mínimo son

$$\nabla_{\boldsymbol{\theta}} n\ell\ell(\boldsymbol{\theta}, \sigma^2; \boldsymbol{y}, \boldsymbol{A}) = 0$$

$$\nabla_{\sigma^2} n\ell\ell(\boldsymbol{\theta}, \sigma^2; \boldsymbol{y}, \boldsymbol{A}) = 0$$

Los estimadores de máxima verosimilitud de los coeficientes de regresión y de la varianza de los términos de error son:

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{y}$$

$$\widehat{\boldsymbol{y}} = \boldsymbol{A}\widehat{\boldsymbol{\theta}}$$

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