Modelo de regresión lineal múltiple

1 Ajuste mínimo-cuadrático del hiperplano de regresión

En el modelo de regresión múltiple que vamos a presentar se considera que el regresando[1] es una función lineal de k regresores (de los cuales k-1 corresponden a variables explicativas o a transformaciones de las mismas y uno corresponde al término independiente) y de una perturbación. Designando por Y_t al regresando, por X_{2t}, X_{3t}, ..., X_{kt} a los regresores y por u_t a la perturbación aleatoria, el modelo teórico de regresión lineal viene dado, para la observación genérica t-ésima, por la siguiente expresión:

Modelo de regresión múltiple

$$Y_t = \beta_1 + \beta_2 X_{2t} + \cdots + \beta_k X_{kt} + u_t \qquad t = 1, 2, \cdots, T \tag{1}$$

Siendo T el tamaño de la muestra y dando valores a t desde t=1 hasta t=T, se obtiene el siguiente sistema de ecuaciones:

$$\begin{aligned} Y_1 &= \beta_1 + \beta_2 X_{21} + \beta_3 X_{31} + \cdots + \beta_k X_{k1} + u_1 \\ Y_2 &= \beta_1 + \beta_2 X_{22} + \beta_3 X_{32} + \cdots + \beta_k X_{k2} + u_2 \\ \cdots &\quad \cdots \quad \cdots \quad \cdots \\ Y_T &= \beta_1 + \beta_2 X_{2T} + \beta_3 X_{3T} + \cdots + \beta_k X_{kT} + u_T \end{aligned} \tag{2}$$

El sistema de ecuaciones anteriores se puede expresar de forma más compacta utilizando notación matricial. Así, vamos a denominar

$$\mathbf{y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_T \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & X_{21} & X_{31} & \cdots & X_{k1} \\ 1 & X_{22} & X_{32} & \cdots & X_{k2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & X_{2T} & X_{3T} & \cdots & X_{kT} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \cdots \\ \beta_k \end{bmatrix} \qquad \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \cdots \\ u_T \end{bmatrix}$$

El modelo de regresión lineal múltiple (1) expresado en notación matricial es el siguiente:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_T \end{bmatrix} = \begin{bmatrix} 1 & X_{21} & X_{31} & \cdots & X_{k1} \\ 1 & X_{22} & X_{32} & \cdots & X_{k2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & X_{2T} & X_{3T} & \cdots & X_{kT} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \cdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \cdots \\ u_T \end{bmatrix} \tag{3}$$

[1] El regresando puede ser o bien la variable endógena directamente o bien una transformación de la variable endógena (por ejemplo, el logaritmo de la variable endógena).

Si tenemos en cuenta las denominaciones dadas a vectores y matrices, el modelo de regresión lineal múltiple se puede expresar de forma compacta de la siguiente forma:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \tag{4}$$

donde, de acuerdo con la notación expuesta, **y** es un vector $T \times 1$, **X** es una matriz $T \times k$, $\boldsymbol{\beta}$ es un vector $k \times 1$ y **u** es un vector T x 1.

El correspondiente modelo ajustado será el siguiente

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \tag{5}$$

El vector de residuos es igual a la diferencia entre valores observados y ajustados, es decir,

$$\hat{\mathbf{u}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \tag{6}$$

Denominando S a la suma de los cuadrados de los residuos, se tiene que:

$$S = \hat{\mathbf{u}}'\hat{\mathbf{u}} = \begin{bmatrix} \hat{u}_1 & \hat{u}_2 & \dots & \hat{u}_T \end{bmatrix} \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \dots \\ \hat{u}_T \end{bmatrix} = \sum_{t=1}^{T} \hat{u}_t^2 \tag{7}$$

Teniendo en cuenta (6), se obtiene

$$\begin{aligned} S &= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \\ &= \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{y}'\mathbf{y} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \end{aligned} \tag{8}$$

Para llegar a la última igualdad de la expresión anterior se ha tenido en cuenta la igualdad entre los dos escalares siguientes

$$\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} = \mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}}$$

ya que un escalar es igual a su traspuesto, es decir,

$$(\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y})' = \mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}}$$

Aplicar el criterio mínimo-cuadrático expuesto en el tema de regresión lineal simple es equivalente a minimizar el escalar *S.* Para ello se calcula la primera derivada de *S* con respecto al vector de coeficientes mínimo-cuadráticos, $\hat{\boldsymbol{\beta}}$, en la expresión (8) y se iguala a 0[2]:

$$\frac{\partial S}{\partial \hat{\boldsymbol{\beta}}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{0} \tag{9}$$

[2] Para la derivación de escalares, expresados mediante productos matriciales, respecto a un vector, véase el anexo 1 de *Econometría aplicada*.

Por tanto,

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} \qquad (10)$$

Al sistema anterior se le denomina genéricamente sistema de ecuaciones normales del hiperplano. Cuando k=2 se obtiene el sistema de ecuaciones normales de la recta; cuando k=3 se obtiene el sistema de ecuaciones normales del plano; finalmente, cuando k>3 se obtiene específicamente el sistema de ecuaciones normales del hiperplano, el cuál no es susceptible de ser representado físicamente.

En notación matricial expandida, el sistema de ecuaciones normales es el siguiente:

$$\begin{bmatrix} T & \sum_{t=1}^{T} X_{2t} & \cdots & \sum_{t=1}^{T} X_{kt} \\ \sum_{t=1}^{T} X_{2t} & \sum_{t=1}^{T} X_{2t}^2 & \cdots & \sum_{t=1}^{T} X_{2t}X_{kt} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{t=1}^{T} X_{kt} & \sum_{t=1}^{T} X_{kt}X_{2t} & \cdots & \sum_{t=1}^{T} X_{kt}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \sum_{t=1}^{T} Y_t \\ \sum_{t=1}^{T} X_{2t}Y_t \\ \vdots \\ \sum_{t=1}^{T} X_{kt}Y_t \end{bmatrix} \qquad (11)$$

Obsérvese que:

a) $\mathbf{X}'\mathbf{X}/T$ es la matriz de momentos muestrales de segundo orden, con respecto al origen, de los regresores.

b). $\mathbf{X}'\mathbf{y}/T$ es el vector de momentos muestrales de segundo orden, con respecto al origen, entre el regresando y los regresores.

Para poder resolver el sistema (10) respecto a $\hat{\boldsymbol{\beta}}$ unívocamente, se debe cumplir que el rango de la matriz $\mathbf{X}'\mathbf{X}$ sea igual a k. Si se cumple esta condici6n, se pueden premultiplicar ambos miembros de (10) por $\left[\mathbf{X}'\mathbf{X}\right]^{-1}$

$$\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{y}$$

con lo cual se obtiene la expresión del vector de estimadores mínimo-cuadráticos:

$$\hat{\boldsymbol{\beta}} = \left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{y} \qquad (12)$$

S presenta un mínimo en $\hat{\boldsymbol{\beta}}$, ya que la matriz de segundas derivadas, $2\mathbf{X}'\mathbf{X}$, es definida positiva. Para comprobarlo, consideremos el vector **a**, de orden $k \times 1$, distinto de cero. Entonces,

$$\mathbf{a}'\mathbf{X}'\mathbf{X}\mathbf{a} = \left[\mathbf{X}\mathbf{a}\right]'\left[\mathbf{X}\mathbf{a}\right]$$

será el producto escalar del vector $\left[\mathbf{X}\mathbf{a}\right]'$ por su transpuesto, producto que no será negativo, por ser una suma de cuadrados. Si el rango de $\mathbf{X}'\mathbf{X}$ es k, entonces queda garantizado que dicho producto es positivo. Por otra parte, al ser $\mathbf{X}'\mathbf{X}$ definida positiva, se sigue, evidentemente, que $2\mathbf{X}'\mathbf{X}$ también es definida positiva.

2. Propiedades descriptivas en la regresión lineal múltiple

Las propiedades que se exponen a continuación son propiedades derivadas exclusivamente de la aplicación del método de estimación por mínimos cuadrados al modelo de regresión (1) en el que se incluye como primer regresor el término independiente.

1. *La suma de los residuos mínimo-cuadráticos es igual a cero*:

$$\sum_{t=1}^{T} \hat{u}_t = 0 \qquad (13)$$

Demostración.

Por definición de residuo

$$\hat{u}_t = Y_t - \hat{Y}_t = Y_t - \hat{\beta}_1 - \hat{\beta}_2 X_{2t} - \cdots - \hat{\beta}_k X_{kt} \qquad t = 1, 2, \cdots, T \qquad (14)$$

Si sumamos para las T observaciones se obtiene:

$$\sum_{t=1}^{T} \hat{u}_t = \sum_{t=1}^{T} Y_t - T\hat{\beta}_1 - \hat{\beta}_2 \sum_{t=1}^{T} X_{2t} - \cdots - \hat{\beta}_k \sum_{t=1}^{T} X_{kt} \qquad (15)$$

Por otra parte, la primera ecuación del sistema de ecuaciones normales (11) es igual a

$$T\hat{\beta}_1 + \hat{\beta}_2 \sum_{t=1}^{T} X_{2t} + \cdots + \hat{\beta}_k \sum_{t=1}^{T} X_{kt} = \sum_{t=1}^{T} Y_t \qquad (16)$$

Al comparar (15) y (16), se concluye que necesariamente debe cumplirse (13).

Obsérvese que, al cumplirse (13), se cumplirá también que

$$\sum_{t=1}^{T} Y_t = \sum_{t=1}^{T} \hat{Y}_t$$

y, al dividir por T, se obtiene

$$\overline{Y} = \overline{\hat{Y}} \qquad (17)$$

2. El hiperplano de regresión pasa necesariamente por el punto $(\overline{Y}, \overline{X}_2, \cdots, \overline{X}_k)$.

Demostración.

En efecto, dividiendo la ecuación (16) por T se obtiene:

$$\overline{Y} = \hat{\beta}_1 + \hat{\beta}_2 \overline{X}_2 + \cdots + \hat{\beta}_k \overline{X}_k \qquad (18)$$

3. Los momentos de segundo orden entre cada regresor y los residuos son iguales a 0.

Para el conjunto de los regresores se puede expresar así:

$$\mathbf{X'\hat{u} = 0} \qquad (19)$$

En efecto,

$$\mathbf{X}'\hat{\mathbf{u}} = \mathbf{X}'\left[\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right] = \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{y} = \mathbf{0}$$

Para llegar a la última igualdad se ha tenido en cuenta (6) y (10).

4. Los momentos de segundo orden entre $\hat{\mathbf{y}}$ *y los residuos son 0*, es decir,

$$\hat{\mathbf{y}}'\hat{\mathbf{u}} = \mathbf{0} \qquad (20)$$

Demostración.

En efecto, si se tiene en cuenta (5) y (19) resulta que

$$\hat{\mathbf{y}}'\hat{\mathbf{u}} = \left[\mathbf{X}\hat{\boldsymbol{\beta}}\right]' \hat{\mathbf{u}} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\hat{\mathbf{u}} = \hat{\boldsymbol{\beta}}'\mathbf{0} = \mathbf{0}$$

3 Hipótesis estadísticas básicas del modelo

I Hipótesis sobre la forma funcional

La relación entre el regresando, los regresores y la perturbación aleatoria es lineal:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \cdots + \beta_k X_{kt} + u_t \qquad t = 1, 2, \cdots, T$$

o, en forma matricial,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \qquad (21)$$

II Hipótesis sobre el vector de perturbaciones aleatorias

La perturbación aleatoria u_t *es una variable aleatoria no observable.*

a) La esperanza matemática del vector de perturbaciones aleatorias es cero.

$$E(\mathbf{u}) = \mathbf{0} \qquad (22)$$

b) Todas las perturbaciones aleatorias tienen la misma varianza, es decir, las perturbaciones son homoscedásticas

$$E(u_t)^2 = \sigma^2 \qquad t = 1, 2, \ldots, T \qquad (23)$$

c) Las perturbaciones aleatorias con distintos subíndices no están correlacionadas entre sí, es decir, las perturbaciones no están autocorrelacionadas:

$$E(u_t u_s) = 0 \qquad t \neq s \qquad (24)$$

La formulación de las hipótesis b) y c) permiten especificar la matriz de covarianzas del vector de perturbaciones. (La matriz de covarianzas de un vector que contiene T variables aleatorias es una matriz cuadrada y simétrica de orden $T \times T$, en cuya diagonal principal aparecen las varianzas de cada uno de los elementos del vector y fuera de la diagonal principal aparecen las covarianzas entre cada par de elementos.)

En concreto, la matriz de covarianzas del vector de perturbaciones es la siguiente:

$$E\left[\left[\mathbf{u}-E(\mathbf{u})\right]\left[\mathbf{u}-E(\mathbf{u})\right]'\right]=E\left[\left[\mathbf{u}-\mathbf{0}\right]\left[\mathbf{u}-\mathbf{0}\right]'\right]=E\left[\left[\mathbf{u}\right]\left[\mathbf{u}\right]'\right]$$

$$=E\left[\begin{bmatrix}u_1\\u_2\\\vdots\\u_T\end{bmatrix}\begin{bmatrix}u_1 & u_2 & \cdots & u_T\end{bmatrix}\right]=E\begin{bmatrix}u_1^2 & u_1u_2 & \cdots & u_1u_T\\u_2u_1 & u_2^2 & \cdots & u_2u_T\\\vdots & \vdots & \ddots & \vdots\\u_Tu_1 & u_Tu_2 & \cdots & u_T^2\end{bmatrix}$$

$$=\begin{bmatrix}E(u_1^2) & E(u_1u_2) & \cdots & E(u_1u_T)\\E(u_2u_1) & E(u_2^2) & \cdots & E(u_2u_T)\\\vdots & \vdots & \ddots & \vdots\\E(u_Tu_1) & E(u_Tu_2) & \cdots & E(u_T^2)\end{bmatrix}=\begin{bmatrix}\sigma^2 & 0 & \cdots & 0\\0 & \sigma^2 & \cdots & 0\\\vdots & \vdots & \ddots & \vdots\\0 & 0 & \cdots & \sigma^2\end{bmatrix}$$

Para llegar a la última igualdad se ha tenido en cuenta que las varianzas de cada uno de los elementos del vector es constante e igual a σ^2 de acuerdo con (23) y que las covarianzas entre cada par de elementos es 0 de acuerdo con (24).

El resultado anterior se puede expresar de forma sintética:

$$E(\mathbf{u}\mathbf{u}')=\sigma^2\mathbf{I} \qquad (25)$$

A la anterior matriz se le denomina *matriz escalar*, ya que es igual a un escalar (σ^2, en este caso) multiplicada por la matriz identidad.

d) La perturbación aleatoria tiene una distribución normal multivariante

Todas las hipótesis sobre el vector de perturbaciones se pueden formular de la siguiente forma:

$$\mathbf{u}\sim N(\mathbf{0},\sigma^2\mathbf{I}) \qquad (26)$$

III Hipótesis sobre el regresor X

a) La matriz de regresores, **X***, es una matriz fija.*

De acuerdo con esta hipótesis, los distintos regresores del modelo toman los mismos valores para diversas muestras del regresando. Éste es un supuesto fuerte en el caso de las ciencias sociales, en el que es poco viable experimentar. Los datos se obtienen por observación, y no por experimentación. Para que dicho supuesto se cumpliera, los regresores deberían ser susceptibles de ser controlados por parte del investigador. Es importante señalar que los resultados que se obtienen utilizando este supuesto se mantendrían prácticamente idénticos si supusiéramos que los regresores son estocásticos, siempre que introdujéramos el supuesto adicional de independencia entre los regresores y la perturbación aleatoria. Este supuesto alternativo se puede formular así:

a) La matriz de regresores,* ***X****, se distribuye independientemente del vector de perturbaciones aleatorias*

$$E(\mathbf{X}'\mathbf{u}) = \mathbf{0} \qquad (27)$$

b) La matriz de regresores, **X**, *tiene rango k.*

$$\rho(\mathbf{X}) = k \qquad (28)$$

Recordemos que la matriz de regresores contiene *k* columnas, correspondientes a los *k* regresores del modelo, y *T* filas, correspondientes al número de observaciones. La hipótesis b) tiene dos implicaciones:

1. El número de observaciones, *T,* debe ser igual o mayor que el numero de regresores, *k.*

2. Todas las columnas de la matriz de regresores deben ser linealmente independientes, lo cual implica que no puede existir una relación lineal exacta entre ningún subconjunto de regresores. En caso contrario, el rango de la matriz **X** sería menor que *k,* y, por tanto, la matriz $\mathbf{X}'\mathbf{X}$ sería singular (carecería de inversa), con lo cual no se podría determinar los valores del vector de estimadores de los parámetros del modelo. Si se diera este caso se dice que existe *multicolinealidad perfecta*. Si existe una relación lineal aproximada – es decir, no exacta –, entonces se pueden obtener estimaciones de los parámetros, si bien la fiabilidad de los mismos quedaría afectada. En este último caso se dice que existe *multicolinealidad no perfecta*.

c) La matriz de regresores, **X**, *no contiene errores de observación o de medida*

Ésta es una hipótesis que raramente se cumple en la práctica, ya que los instrumentos de medición en economía son escasamente fiables (piénsese en la multitud de errores que es posible cometer en una recogida de información, mediante encuesta, sobre los presupuestos familiares). Aunque es difícil encontrar instrumentos para contrastar esta hipótesis, la naturaleza del problema y, sobre todo, la procedencia de los datos utilizados pueden ofrecer evidencia favorable o desfavorable a la hipótesis enunciada.

IV Hipótesis sobre el vector de parámetros β

El vector de parámetros β es constante.

Si no se adopta esta hipótesis el modelo de regresión sería muy complicado de manejar. En todo caso, puede ser aceptable que los parámetros del modelo se mantienen estables en el tiempo (si no se trata de períodos muy extendidos) o en el espacio (si está relativamente acotado).

4 Propiedades probabilísticas del modelo

Distribución del regresando

El regresando es función lineal del vector de perturbaciones aleatorias, que, por la hipótesis II d), sigue una distribución normal. Por lo tanto, el regresando, **y**, seguirá también una distribución normal.

La esperanza matemática de **y**, teniendo en cuenta la hipótesis II a)[3], viene dada por

$$E(\mathbf{y}) = E\left[\mathbf{X}\boldsymbol{\beta} + \mathbf{u}\right] = \mathbf{X}\boldsymbol{\beta} + E(\mathbf{u}) = \mathbf{X}\boldsymbol{\beta} \qquad (29)$$

La matriz de varianzas covarianzas, teniendo en cuenta las hipótesis II a) a II c), serán

$$Var(\mathbf{y}) = E\left[\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)'\right] = E(\mathbf{u}\mathbf{u}') = \sigma^2\mathbf{I} \qquad (30)$$

En consecuencia, e1 regresando, **y**, tiene una distribución normal multivariante con vector de medias $\boldsymbol{\beta}$ y con una matriz de varianzas-covarianzas, $\sigma^2\mathbf{I}$, escalar.

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}) \qquad (31)$$

UU

[3] Las hipótesis de los bloques III y IV se tendrán implícitamente en cuenta, aunque no se mencionen.