Tema 2: El Modelo Lineal (I)

Econometría 2021-2022 — Grado Ingenierías & ADE

Introducción

Modelo Lineal Simple:

$$y_i \sim a + bx_i$$

Modelo Lineal Múltiple:
Permite estudiar el comportamiento de una variable endógena Y, a partir de k variables independientes $(X_2, X_3, \ldots, X_k)$

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + u_i$$

El modelo lineal múltiple

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + u_i$$

- el subíndice $i = 1, \ldots, n$, indica la i-ésima de las n observaciones de la muestra.
- Y_i es la i-ésima observación de la variable dependiente.
- $X_{2i}, \ldots, X_{ki}$ son las i-ésimas observaciones de las k variables independientes.
- u_i es la perturbación de la i-ésima observación, encargada de recoger aquella parte de la variable Y_i que no es explicada por $X_{1i}, X_{2i}, \ldots, X_{ki}$.
- $\beta_1, \ldots, \beta_k$ son los parámetros del modelo. Además, β_1 es el denominado término constante.

El modelo lineal múltiple

La decripción completa es:

$$\begin{aligned} Y_1 &= \beta_1 + \beta_2 X_{21} + \ldots + \beta_k X_{k1} + u_1 \\ Y_2 &= \beta_1 + \beta_2 X_{22} + \ldots + \beta_k X_{k2} + u_2 \\ &\vdots \\ Y_n &= \beta_1 + \beta_2 X_{2n} + \ldots + \beta_k X_{kn} + u_n \end{aligned}$$

Equivalentemente:

$$\begin{aligned} Y_1 &= \beta_1 X_{11} + \beta_2 X_{21} + \ldots + \beta_k X_{k1} + u_1 \\ Y_2 &= \beta_1 X_{12} + \beta_2 X_{22} + \ldots + \beta_k X_{k2} + u_2 \\ &\vdots \\ Y_n &= \beta_1 X_{1n} + \beta_2 X_{2n} + \ldots + \beta_k X_{kn} + u_n \end{aligned}$$

donde $X_1 = (1, \ldots, 1)$.

El modelo lineal múltiple

$$\vec{y} = X\vec{\beta} + \vec{u}$$

$$\vec{y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & X_{21} & \dots & X_{k1} \\ 1 & X_{22} & \dots & X_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{2n} & \dots & X_{kn} \end{pmatrix} \quad \vec{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} \quad \vec{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

Supuestos

Para el análisis estadístico del modelo, comenzaremos considerando las siguiente hipótesis básicas sobre el modelo:

- El supuesto de linealidad.
- El supuesto rango completo por columnas.
- El supuesto de exogeneidad.
- El supuesto de causalidad. El mecanismo de generación de las observaciones.
- Supuestos sobre el término perturbación.
- El supuesto de normalidad del término de perturbación.

Supuesto de Linealidad

Sobre el vector $\overrightarrow{y}$ existe una relación entre las variables a nivel poblacional X y el resto de factores omitidos que son relevantes en la explicación de la variable dependiente, $\overrightarrow{u}$. Supondremos que esta relación de causalidad es lineal, permitiéndonos así utilizar procedimientos de álgebra de matrices y vectores, lo que simplifica enormemente la operatoria. La relación muestral podemos escribirla como:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + u_i$$

con $i = 1, \ldots, n$ o equivalentemente $Y_i = \overrightarrow{X}_i^t \overrightarrow{\beta} + u_i$.

Supuesto de Rango Completo por Columnas

La matriz de observaciones muestrales $X_{n \times k}$ tiene rango completo por columnas. Al verificarse que $n > k$ se tiene que $\rho(X) = k$. Por tanto, a nivel muestral, las k columnas de la matriz X son linealmente independientes, reflejando la no existencia de una relación lineal entre las variables explicativas.
Este supuesto a efectos del procedimiento de estimación del modelo, nos permite efectuar el cálculo de la inversa de la matriz de productos cruzados de los regresores del modelo, lo que denotaremos en el próximo tema como la matriz $\left(X^t X\right)^{-1}$. En el caso de que no se verifique el supuesto mencionado, implicaría que alguna de las variables explicativas sería combinación lineal de otras de las variables, produciéndose un problema de multicolinealidad.

Supuesto de Exogeneidad

Las variables explicativas del modelo $X_{1i}, X_{2i}, \ldots, X_{ki}$, con $i = 1, \ldots, n$, no incorporan información alguna que nos permita predecir el término perturbación, u_i. Equivalentemente, se tiene que para cualquier valor que tomen las variables explicativas, el valor esperado de la perturbación es cero. Es decir,

$$E\left[u_i | \left(X_{1i}, X_{2i}, \ldots, X_{ki}\right)\right] = 0$$

para $i = 1, \ldots, n$. Por tanto, teniendo en cuenta que $Y_i = \overrightarrow{X}_i^t \overrightarrow{\beta} + u_i$, se obtiene de forma inmediata que

$$E\left[Y_i | \left(X_{1i}, X_{2i}, \ldots, X_{ki}\right)\right] = \overrightarrow{X}_i^t \overrightarrow{\beta}\,.$$

Supuesto de Causalidad

La relación de causalidad será siempre unidireccional, en el sentido de que serán las variables que aparecen en el lado derecho de la ecuación las que expliquen el comportamiento de la variable dependiente pero no a la inversa. Teniendo en cuenta que la variable explicada, Y_i, depende de la perturbación, se verifica que la variable dependiente adquirirá también un carácter aleatorio.
La matriz de las variables explicativas, X, se supone no estocástica. Por tanto, el proceso generador de estas observaciones será ajeno al proceso generador del modelo y de la perturbación como consecuencia del supuesto de exogeneidad.

Supuesto sobre la perturbación

La perturbación aleatoria, u_i, está centrada, es homocedástica e incorrelada. Es decir:

1. u_i está centrada: $E[u_i] = 0 \ \forall i = 1, \ldots, n$.
2. La perturbación es homocedástica: la varianza de la misma se mantiene constante para todas las observaciones, es decir, $\text{var}[u_i] = E[u_i^2] = \sigma^2 \ \forall i = 1, \ldots, n$. En el caso de que no se verifique lo indicado, estaríamos ante un problema de heterocedasticidad.
3. Si la perturbación es incorrelada provoca que $\text{cov}[u_i u_j] = E[u_i u_j] = 0 \ \forall i \neq j$ con $i = 1, \ldots, n$ y $i = j, \ldots, n$. Así que la perturbación en el momento i no está correlacionada con el su valor en otro momento determinado j. Cuando $\text{cov}[u_i u_j] \neq 0$ entonces estamos ante un problema de autocorrelación.

Como consecuencia se tiene que la matriz de varianzas-covarianzas del término perturbación es:

$$\text{var}[\vec{u}] = E\left[(\vec{u} - E[\vec{u}])(\vec{u} - E[\vec{u}])^t\right] = E[\vec{u}\,\vec{u}^t]$$

El supuesto de normalidad del término de perturbación

Teniendo en cuenta las hipótesis relacionadas con el término perturbación, y con el fin de realizar inferencia sobre los parámetros que intervienen en el modelo, se considera que la perturbación aleatoria u_i se distribuye según una normal con media cero y varianza σ^2, es decir

$$u_i \sim N(0, \sigma^2) \qquad \forall i = 1, \ldots, n$$

Por el momento supondremos que el vector $\vec{u}$ se distribuye según una distribución normal multivariante con vector de medias $\vec{0}$ y matriz de varianzas-covarianzas $\sigma^2 \text{I}_n$, es decir

$$\vec{u} \sim N\left(\vec{0}, \sigma^2 \text{I}_n\right).$$

La función de densidad asociada a dicha distribución viene dada por la siguiente expresión:

$$f(\vec{u}\,|\sigma^2) = \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^n} \exp\left[-\frac{1}{2\sigma^2}\vec{u}^t\vec{u}\right]$$

Mínimos Cuadrados Ordinarios

Consideremos el modelo lineal múltiple clásico en su notación matricial

$$\vec{y} = X\vec{\beta} + \vec{u}$$

Para estimar los parámetros $\vec{\beta}$ utilizaremos el Método de los Mínimos Cuadrados Ordinarios (MCO).
Denotando por $\vec{\widehat{y}}$ el valor ajustado de $\vec{y}$ por el modelo:

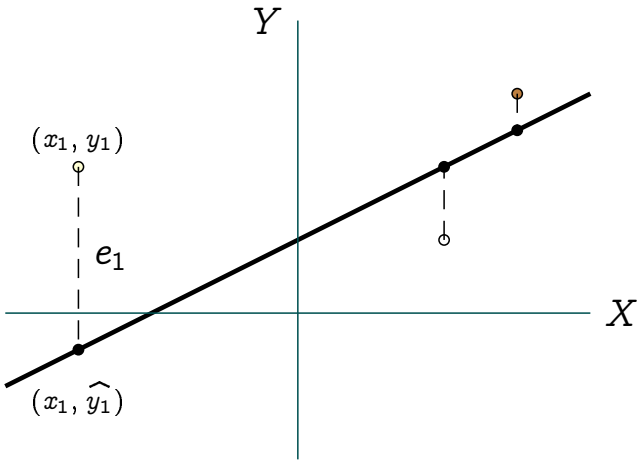
$$\vec{\widehat{y}} = X\vec{\widehat{\beta}}$$

Definición (Residuo)

Diferencia existente entre el valor observado de la variable explicada y su estimación:

$$\vec{e} = \vec{y} - \vec{\widehat{y}} = \vec{y} - X\vec{\widehat{\beta}}$$

Introducción



Minimizaremos la suma de los cuadrados de los residuos del ajuste:

$$\min_{\vec{\widehat{\beta}}} \sum_{i=1}^{n} e_i^2$$

donde $e_i = y_i - \widehat{y}_i = y_i - \widehat{\beta}_1 - \widehat{\beta}_2 x_{2i} - \cdots - \widehat{\beta}_k x_{ki}$ (componentes de $\vec{e}$)
Denotaremos a la función a minimizar por: $f(\vec{\widehat{\beta}}) = \vec{e}^{\,t}\vec{e}$.

Regresión por Mínimos Cuadrados

Paso 1 Definimos la expresión a minimizar. Se define la suma de los cuadrados de los residuos como:

$$\begin{aligned} f\left(\overrightarrow{\beta}\right) &= \left(\overrightarrow{y} - X\overrightarrow{\beta}\right)^t \left(\overrightarrow{y} - X\overrightarrow{\beta}\right) \\ &= \overrightarrow{y}^t\overrightarrow{y} - \overrightarrow{y}^t X\overrightarrow{\beta} - \overrightarrow{\beta}^t X^t\overrightarrow{y} + \overrightarrow{\beta}^t X^t X\overrightarrow{\beta} \\ &= \overrightarrow{y}^t\overrightarrow{y} - 2\overrightarrow{\beta}^t X^t\overrightarrow{y} + \overrightarrow{\beta}^t X^t X\overrightarrow{\beta} \end{aligned}$$

Paso 2 Derivamos con respecto a $\overrightarrow{\beta}$ e igualamos a cero.

$$\nabla f\left(\overrightarrow{\beta}\right) = -2X^t\overrightarrow{y} + 2X^t X\overrightarrow{\beta} = \overrightarrow{0}$$

Regresión por Mínimos Cuadrados

Paso 3 Despejamos el vector de parámetros.

$$-2X^t\overrightarrow{y} + 2X^t X\overrightarrow{\beta} = \overrightarrow{0} \Leftrightarrow X^t X\overrightarrow{\beta} = X^t\overrightarrow{y}$$

Considerando el supuesto de rango completo por columnas, sabemos que se verifica que $\rho\left(X^t X\right) = k$, condición necesaria para la existencia de la matriz $\left(X^t X\right)^{-1}$, luego:

$$\overrightarrow{\beta} = \left(\mathrm{X^t X}\right)^{-1}\left(\mathrm{X^t}\overrightarrow{y}\right)$$

Paso 4 Para que la solución sea mínimo se ha de cumplir:

$$Hess(f) = 2X^t X \succ 0 \ \ (A \succ 0 \text{ sii } z^t Az > 0)$$

$$\overrightarrow{\beta} = \left(\mathrm{X^t X}\right)^{-1}\left(\mathrm{X^t}\overrightarrow{y}\right)$$

es el Estimador Mínimo Cuadrático Ordinario (EMCO) del vector de parámetros $\overrightarrow{\beta}$ del modelo de regresión lineal múltiple $\overrightarrow{y} = X\overrightarrow{\beta} + \overrightarrow{u}$.

Regresión por Mínimos Cuadrados

Si el modelo tiene término independiente entonces se tiene

$$\overrightarrow{\hat{\beta}} = \begin{pmatrix} n & \sum_{i=1}^{n} X_{2i} & \dots & \sum_{i=1}^{n} X_{ki} \\ \sum_{i=1}^{n} X_{2i} & \sum_{i=1}^{n} X_{2i}^2 & \dots & \sum_{i=1}^{n} X_{2i}X_{ki} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} X_{ki} & \sum_{i=1}^{n} X_{ki}X_{2i} & \dots & \sum_{i=1}^{n} X_{ki}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_{2i}Y_i \\ \ddots \\ \sum_{i=1}^{n} X_{ki}Y_i \end{pmatrix}$$

En el caso de que el modelo carezca del término independiente entonces:

$$\overrightarrow{\hat{\beta}} = \begin{pmatrix} \sum_{i=1}^{n} X_{1i}^2 & \sum_{i=1}^{n} X_{1i}X_{2i} & \dots & \sum_{i=1}^{n} X_{1i}X_{ki} \\ \sum_{i=1}^{n} X_{2i}X_{1i} & \sum_{i=1}^{n} X_{2i}^2 & \dots & \sum_{i=1}^{n} X_{2i}X_{ki} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} X_{ki}X_{1i} & \sum_{i=1}^{n} X_{ki}X_{2i} & \dots & \sum_{i=1}^{n} X_{ki}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{n} X_{1i}Y_i \\ \sum_{i=1}^{n} X_{2i}Y_i \\ \ddots \\ \sum_{i=1}^{n} X_{ki}Y_i \end{pmatrix}$$

Propiedades algebraicas de MCO (I)

Considerando que SÍ hay término independiente en el modelo ($X_{1i} = 1$), se tiene:

- Las variables exógenas son ortogonales al vector de los residuos: $X^t \overrightarrow{e} = \overrightarrow{0}$.
  La condición necesaria de EMCO era:

$$-2X^t \overrightarrow{y} + 2X^t X \overrightarrow{\hat{\beta}} = \overrightarrow{0}$$

  Equivalentemente:

$$X^t \left(\overrightarrow{y} - X \overrightarrow{\hat{\beta}} \right) = \overrightarrow{0}$$

  Como $\overrightarrow{\hat{y}} = X \overrightarrow{\hat{\beta}}$, tenemos que: $X^t \left(\overrightarrow{y} - \overrightarrow{\hat{y}} \right) = \overrightarrow{0}$, ó equivalentemente

$$X^t \overrightarrow{e} = \overrightarrow{0}.$$

Propiedades algebraicas de MCO (II)

- La suma de los residuos mínimo cuadráticos es cero: $\sum_{i=1}^{n} e_i = 0$.
- La suma de los valores observados coincide con la suma de los valores estimados: $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \widehat{Y}_i$.
- Definiendo:

$$SCT = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \overrightarrow{y}^t\overrightarrow{y} - n \cdot \overline{Y}^2$$

$$SCR = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \overrightarrow{y}^t\overrightarrow{y} - \overrightarrow{\widehat{\beta}}^t X^t \overrightarrow{y}$$

$$SCE = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 = \overrightarrow{\widehat{\beta}}^t X^t \overrightarrow{y} - n \cdot \overline{Y}^2$$

  Luego: $SCT = SCR + SCE$, donde SCT es la suma de los cuadrados totales, SCR es la suma de los cuadrados de los residuos y SCE es la suma de los cuadrados explicada.
- Se cumple que $\overrightarrow{\widehat{y}}^t \overrightarrow{e} = 0$.

Las Ecuaciones Normales

$$-2X^t\overrightarrow{y} + 2X^tX\overrightarrow{\widehat{\beta}} = \overrightarrow{0} \Rightarrow X^t(y - \widehat{y}) = 0$$

Las siguientes expresiones son equivalentes:

- $X^t\overrightarrow{e} = 0$.
- $$\begin{pmatrix} 1 & X_{21} & \dots & X_{k1} \\ 1 & X_{22} & \dots & X_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{2n} & \dots & X_{kn} \end{pmatrix}^t \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$
- Ecuaciones normales:
$$\begin{cases} e_1 + \cdots + e_n = 0, \\ X_{21}e_1 + \cdots + X_{2n}e_n = 0, \\ \vdots = \vdots \\ X_{k1}e_1 + \cdots + X_{kn}e_n = 0, \end{cases}$$

Bondad del Ajuste: R^2 y $\overline{R}^2$

Definición (Coeficiente de determinación R^2)

Porcentaje de variabilidad explicada por el modelo. Por tanto, éste se obtendrá como el cociente entre la varianza explicada por la estimación y la total:

$$R^2 = \frac{\frac{1}{n}\cdot\sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{Y}\right)^2}{\frac{1}{n}\cdot\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} = \frac{\sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} = \frac{SCE}{SCT}.$$

Si el modelo tiene término independiente: $R^2 = 1 - \frac{SCR}{SCT}$ y los valores del coeficiente estarán comprendidos entre 0 y 1.

Bondad del Ajuste: R^2 y $\overline{R}^2$

Definición (Coeficiente de determinación corregido $\overline{R}^2$)

Porcentaje de variación de la variable explicada considerando el número de variables incluidas en el modelo, es decir, considerando el valor de k. Se define como:

$$\overline{R}^2 = 1 - \frac{SCR/(n-k)}{SCT/(n-1)} = 1 - (1-R^2)\cdot\frac{n-1}{n-k}.$$

Bondad del Ajuste: R^2 y $\overline{R}^2$

Adviértase que obtener un R^2 o $\overline{R}^2$ cercano a 1 no indica que los resultados sean fiables, ya que, por ejemplo, puede ser que no se cumpla alguna de las hipótesis básicas y los resultados no ser válidos.
Es necesario señalar que ambos coeficientes no son capaces de detectar si alguna de las variables incluidas en el modelo son o no estadísticamente significativas.
Por tanto, estos indicadores han de ser considerados como una herramienta más a tener en cuenta dentro del análisis.

Criterios de Akaike y Schwarz

Para comparar distintos modelos podríamos utilizar el coeficiente de determinación. Sin embargo, para poder llevar cabo dicha comparación será nes necesario que tengan la misma variable explicada.
Con el fin de buscar una solución al problema utilizaremos los criterios de selección de modelos de Akaike (AIC) y el bayesiano de Schwarz (BIC). Estos criterios se obtienen a partir de la suma de cuadrados de los residuos y de un factor que penaliza la inclusión de parámetros.
El criterio de información de Akaike:

$$AIC = \ln\left(\frac{SCR}{n}\right) + \frac{2k}{n},$$

El criterio de información de Schwarz:

$$BIC = \ln\left(\frac{SCR}{n}\right) + \frac{k}{n} \cdot \ln(n).$$

Utilizando estos criterios se escogería aquel modelo con un menor valor de AIC o BIC.

Ejemplo

Usando los siguientes datos, consumo nacional (C_t) y renta nacional (R_t) en España para el periodo 1995-2005 a precios corrientes (10^9 euros), obtenga las estimaciones por MCO, así como las sumas de cuadrados total, explicada y residual, y el coeficiente de determinación, para el modelo de regresión $C_t = \beta_1 + \beta_2 R_t + u_t$.

Año	C_t	R_t
1995	349	388
1996	368	408
1997	388	433
1998	414	465
1999	444	498
2000	484	538
2001	518	574
2002	550	614
2003	586	656
2004	635	699
2005	686	748

Ejemplo

Para el modelo $Y_t = \beta_1 + \beta_2 v_t + \beta_3 w_t + u_t$ se tienen los siguientes datos:

$$n = 12, \qquad SCT = 104'9167,$$

$$(X^t X)^{-1} = \begin{pmatrix} 0'6477 & -0'041 & -0'0639 \\ -0'041 & 0'0071 & -0'0011 \\ -0'0639 & -0'0011 & 0'0152 \end{pmatrix}, \qquad X^t Y = \begin{pmatrix} 91 \\ 699 \\ 448 \end{pmatrix}.$$

Se pide:

a) Ajustar el modelo por el método de MCO y calcular el coeficiente de determinación.

Ejemplo

En un estudio de los determinantes de la inversión se usaron 20 datos anuales, correspondientes a las siguientes variables: inversión anual en billones de pesetas (Y), tipo de interés en porcentaje (X_1) y variación anual de PIB en billones de pesetas (X_2). Se dispone de la siguiente información:

$$\begin{array}{lll} \sum X_{1t} = 100 & \sum X_{2t} = 24 & \sum Y_t = 5 \\ \sum X_{1t}Y_t = -255 & \sum X_{2t}Y_t = 146 & \sum X_{1t}X_{2t} = 100 \\ \sum X_{1t}^2 = 680 & \sum X_{2t}^2 = 48'8 & \sum \left(Y_t - \overline{Y}\right)^2 = 1200 \end{array}$$

Se pide:

a) Obtenga las estimaciones por MCO del modelo $Y_t = \alpha + \beta X_{1t} + \delta X_{2t} + u_2$.

Ejemplo

Para estimar el modelo $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t$ se ha obtenido una muestra de la cual ha resultado:

$$X^tX = \begin{pmatrix} 14 & 7 & 14 \\ 7 & 4'5 & 7 \\ 14 & 7 & 15 \end{pmatrix}, \qquad X^tY = \begin{pmatrix} 10 \\ 6 \\ 12 \end{pmatrix}, \qquad Y^tY = 14.$$

Se pide:

a) Estimar los coeficientes del modelo por MCO.