Formulación matricial del modelo lineal general

Estimadores MCO usando matrices

Mariana Marchionni
marchionni.mariana@gmail.com

Econometría I - FCE - UNLP
www.econometria1unlp.com

Temario de la clase

El modelo lineal con K variables

- Vimos el modelo

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_K X_{Ki} + u_i, \;\; i = 1, \ldots n$$

- Como vale para $i = 1, \ldots, n$, entonces podemos escribir:

$$\begin{array}{rcl} Y_1 & = & \beta_1 + \beta_2 X_{21} + \beta_3 X_{31} + \cdots + \beta_K X_{K1} + u_1 \\ Y_2 & = & \beta_1 + \beta_2 X_{22} + \beta_3 X_{32} + \cdots + \beta_K X_{K2} + u_2 \\ & \vdots & \\ Y_n & = & \beta_1 + \beta_2 X_{2n} + \beta_3 X_{3n} + \cdots + \beta_K X_{Kn} + u_n \end{array}$$

- Sistema de n ecuaciones lineales. ¿Lineales en qué?
- Todo sistema lineal puede ser expresado en matrices

Formulación matricial del modelo

- Notar que el modelo puede escribirse como:

$$
\underset{Y}{\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}} = \underset{X}{\begin{bmatrix} 1 & X_{21} & X_{31} & \cdots & X_{K1} \\ 1 & X_{22} & X_{32} & \cdots & X_{K2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{2n} & X_{3n} & \cdots & X_{Kn} \end{bmatrix}} \cdot \underset{\beta}{\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix}} + \underset{u}{\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}}
$$

- ¿Qué dimensiones tiene cada matriz/vector?

$$X = \begin{bmatrix} 1 & X_{21} & X_{31} & \cdots & X_{K1} \\ 1 & X_{22} & X_{32} & \cdots & X_{K2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{2n} & X_{3n} & \cdots & X_{Kn} \end{bmatrix}$$

X es $(n \times K)$, n filas (observaciones) y K columnas (variables explicativas)

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix} \qquad u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

Y y u son $(n \times 1)$, y β es $(K \times 1)$.

- Entonces, el modelo

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_K X_{Ki} + u_i, \ \ i = 1, \ldots n$$

- Puede escribirse como

$$Y = X\beta + u$$

- **Este es el modelo lineal general (K variables) expresado en notación matricial**
  - Ventaja: nos permitirá trabajar expresiones sin el uso de sumatorias
  - Desventaja: trabajar con álgebra matricial puede ser un poco molesto al principio

Digresión: definiciones y resultados de álgebra matricial

- **Rango de una matriz.** Def.: es el número máximo de filas y/o columnas linealmente independientes (li). Lo denotamos como $\rho(X)$
- Resultado 1: el **máximo nro. de columnas li = máximo nro. de filas li**. En otras palabras: el rango columna coincide con el rango fila.
- **Matriz no singular.** Def.: una matriz cuadrada $A_{(K\times K)}$ es no singular, si y solo si su determinante es distinto de cero, es decir, $|A| \neq 0$.
- Resultado 2: si la matriz $A_{(K\times K)}$ es no singular $\Rightarrow$ existe una única matriz, también no singular, a la que llamamos **inversa de** A y denotamos como A^{-1}, tal que $AA^{-1} = A^{-1}A = I_K$
- Resultado 3. Sea una matriz $A_{(K\times K)}$:
$$\begin{cases} \rho(A) = K & \Longrightarrow \quad |A| \neq 0 \\ \rho(A) < K & \Longrightarrow \quad |A| = 0 \end{cases}$$
- Resultado 4. Sea una matriz $X_{(n\times K)}$ con $\rho(X) = K$ (rango columna completo). Se cumple que $\rho(X) = \rho(X'X) = K$

¿Qué es $X'X$ en nuestro modelo?

Supongamos $n = 3$ y $K = 2$:

$$X = \begin{bmatrix} 1 & X_{21} \\ 1 & X_{22} \\ 1 & X_{23} \end{bmatrix} \Rightarrow X'X = \begin{bmatrix} 3 & X_{21} + X_{22} + X_{23} \\ X_{21} + X_{22} + X_{23} & X_{21}^2 + X_{22}^2 + X_{23}^2 \end{bmatrix}$$

Generalizando para cualquier n y K:

$$X'X = \begin{bmatrix} n & \sum X_{2i} & \sum X_{3i} & \cdots & \sum X_{Ki} \\ \sum X_{2i} & \sum X_{2i}^2 & \sum X_{2i}X_{3i} & \cdots & \sum X_{2i}X_{Ki} \\ \sum X_{3i} & \sum X_{2i}X_{3i} & \sum X_{3i}^2 & \cdots & \sum X_{3i}X_{Ki} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum X_{Ki} & \sum X_{2i}X_{Ki} & \sum X_{3i}X_{Ki} & \cdots & \sum X_{Ki}^2 \end{bmatrix}$$

- notar que $X'X$ es de dimensión $(K \times K)$

¿Qué nos dice el resultado 4?

- Recordemos el resultado 4: Si $\rho(X) = K$ (rango columna completo), se cumple que $\rho(X) = \rho(X'X) = K$
- ¿Qué se requiere para que $\rho(X) = K$?

$$X = \begin{bmatrix} 1 & X_{21} & X_{31} & \cdots & X_{K1} \\ 1 & X_{22} & X_{32} & \cdots & X_{K2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{2n} & X_{3n} & \cdots & X_{Kn} \end{bmatrix}$$

- Por el resultado 4 sabemos que si $\rho(X) = K$ entonces $\rho(X'X) = K$. En otras palabras, que $\rho(X) = K$ garantiza que $X'X$ tiene rango completo y por lo tanto es una matriz no singular (resultado 3), lo que implica que $\exists (X'X)^{-1}$ (resultado 2).
  - Que $X'X$ tenga inversa es muy importante, ya veremos...

Digresión: derivación con matrices

- **Regla 1:** sean a y b dos vectores $(K \times 1)$, entonces:
$$\frac{\partial(b'a)}{\partial b} = a$$

- **Regla 2:** sea A una matriz simétrica $(K \times K)$ y b un vector $(K \times 1)$, entonces:
$$\frac{\partial(b'Ab)}{\partial b} = 2Ab$$

Chequeamos la regla 1: $\frac{\partial(b'a)}{\partial b} = a$

Supongamos $K = 2$: $a = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$ y $b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$

- Notar que $b'a = b_1 a_1 + b_2 a_2$ es un escalar (dimensión 1x1)

$\Rightarrow \frac{\partial(b'a)}{\partial b}$ es un escalar derivado por un vector. ¿Y eso?

- Derivar por un vector $(K \times 1)$ es derivar por cada uno de los K elementos del vector.
  - Derivamos $b'a = b_1 a_1 + b_2 a_2$ primero respecto de b_1 y obtenemos a_1; después derivamos respecto de b_2 y obtenemos a_2. Luego "apilamos" los resultados en un nuevo vector (vector de derivadas):

$$\frac{\partial(b'a)}{\partial b} = \begin{bmatrix} \frac{\partial(b'a)}{\partial b_1} \\ \frac{\partial(b'a)}{\partial b_2} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = a$$

Chequeamos la regla 2: $\frac{\partial(b'Ab)}{\partial b} = 2Ab$

Supongamos otra vez $K = 2$: $b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$ y $A = \begin{bmatrix} A_{11} & A_{12} \\ A_{12} & A_{22} \end{bmatrix}$

- Notar que $b'Ab = b_1^2 A_{11} + b_2^2 A_{22} + 2b_1 b_2 A_{12}$ es una función cuadrática en b (y es un escalar)
- Entonces, ¿qué es $\frac{\partial(b'Ab)}{\partial b}$?
- Derivamos:

$$\frac{\partial(b'Ab)}{\partial b} = \begin{bmatrix} \frac{\partial(b'Ab)}{\partial b_1} \\ \frac{\partial(b'Ab)}{\partial b_2} \end{bmatrix} = \begin{bmatrix} 2b_1 A_{11} + 2b_2 A_{12} \\ 2b_2 A_{22} + 2b_1 A_{12} \end{bmatrix} = 2Ab$$

MCO en matrices

Recordemos la formulación matricial del modelo lineal general:
$Y = X\beta + u$

- Sea $\hat{\beta}$ el vector que apila los estimadores del vector de parámetros

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_K \end{bmatrix}$$

Definiciones

- Vector de estimaciones de Y $(n \times 1)$:
  $\hat{Y} \equiv X\hat{\beta}$
- Vector de residuos o errores de estimación $(n \times 1)$:
  $e \equiv Y - \hat{Y} = Y - X\hat{\beta}$

Recordemos la función de pérdida del método de MCO y veamos cómo escribirla con la nueva notación matricial:

$$SRC \equiv \sum_{i=1}^{n} e_i^2 = \begin{bmatrix} e_1 & e_2 & \cdots & e_n \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

- Es decir, la suma de residuos cuadráticos se puede reescribir como:

$$SRC \equiv e'e$$

- Recordando que $e \equiv Y - X\hat{\beta}$ es fácil ver que SRC es una función de $\hat{\beta}$:

$$SRC(\hat{\beta}) \equiv (Y - X\hat{\beta})'(Y - X\hat{\beta})$$

El problema de MCO:

$$\underset{\hat{\beta}}{Min}\ SRC(\hat{\beta}) \equiv e'e = (Y - X\hat{\beta})'(Y - X\hat{\beta})$$

- Notar que la función a minimizar es escalar (1×1)
- Hay que derivar esa expresión (1×1) respecto del vector $\hat{\beta}$
- Las CPO igualan el vector de derivadas al vector cero:

$$\frac{\partial SRC(\hat{\beta})}{\partial \hat{\beta}} = 0$$

- Sistema de K ecuaciones con K incógnitas

Obtención de los estimadores MCO

$$e'e = \left(Y-\hat{Y}\right)'\left(Y-\hat{Y}\right) = \left(Y-X\hat{\beta}\right)'\left(Y-X\hat{\beta}\right)$$

$$= Y'Y - Y'X\hat{\beta} - \hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$$

$$= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$$

- En el último paso, notar que los términos 2do y 3ro son iguales porque (1) ambos son escalares, (2) si trasponemos $(Y'X\hat{\beta})$ obtenemos $(\hat{\beta}'X'Y)$ y viceversa, y (3) un escalar traspuesto siempre es igual al mismo escalar.
- Ahora hay que derivar respecto del vector $\hat{\beta}$
  - El primer término no depende de $\hat{\beta}$
  - Notar que el 2do término es de la forma $b'a$, con $b=\hat{\beta}$ y $a=X'Y$ (recordar regla 1)
  - Notar que el 3er término es de la forma $b'Ab$, con $b=\hat{\beta}$ y $A=X'X$ (recordar regla 2)

- La función a minimizar:

$$e'e = Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$$

- Las CPO:

$$\frac{\partial e'e}{\partial \hat{\beta}} = 0 - 2X'Y + 2X'X\hat{\beta} = 0 \Rightarrow X'X\hat{\beta} = X'Y$$

- Si existe $(X'X)^{-1}$, entonces:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

- $\hat{\beta}$ es el vector de estimadores MCO

¿De qué depende que exista $(X'X)^{-1}$?

- Vimos antes (resultado 4) que si $\rho(X) = K$ (rango columna completo), entonces $\exists (X'X)^{-1}$
- El supuesto de no multicolinealidad perfecta justamente dice que $\rho\,(X) = K$
- **La ausencia de multicolinealidad perfecta es necesaria para que exista el vector de estimadores MCO**

Propiedades algebraicas en notación matricial

Propiedad 1: los estimadores de MCO son lineales, es decir tienen la forma

$$\hat{\beta} = AY$$

donde A es una matriz $(K \times n)$ con elementos *no estocásticos* (no aleatorios).

Prueba:

- Los estimadores MCO son $\hat{\beta} = (X'X)^{-1}X'Y$
- Si llamamos A a la matriz $(X'X)^{-1}X'$ (de dimensión $K \times n$), $\hat{\beta}$ queda escrito en la forma lineal

Propiedad 2: $X'e = 0$

- Puede obtenerse a partir de la CPO: $X'X\hat{\beta} = X'Y$
- Mostrar que esta propiedad implica 2 resultados que ya vimos antes:
  1. $\sum_{i=1}^{n} e_i = 0$
  2. $\sum_{i=1}^{n} X_{ki} e_i = 0 \;\; k = 2, ...K$
- ¿Qué nos dicen estos resultados intuitivamente?

Propiedad 3: similar a la propiedad anterior:

$$\hat{Y}'e = 0$$

Propiedad 4: el punto $(\bar{X}, \bar{Y})$ pertenece a la función de regresión estimada por MCO:

$$\bar{Y} = \bar{X}\hat{\beta}$$

Bondad del ajuste usando notación matricial

La descomposición de la suma de cuadrados es

$$\begin{aligned}
\sum_{i=1}^{N}(Y_i-\bar{Y})^2 &= \sum_{i=1}^{N}(\hat{Y}_i-\bar{Y})^2 + \sum_{i=1}^{N}e_i^2 \\
\underbrace{Y'Y-n\bar{Y}^2} &= \underbrace{\hat{Y}'\hat{Y}-n\bar{Y}^2} + \underbrace{e'e} \\
STC &= SEC + SRC
\end{aligned}$$

Entonces, la bondad del ajuste puede escribirse como:

$$R^2 = \frac{\hat{Y}'\hat{Y}-n\bar{Y}^2}{Y'Y-n\bar{Y}^2} = 1-\frac{e'e}{Y'Y-n\bar{Y}^2}$$

R^2 no decrece con K

- El R^2 nunca decrece con el agregado de variables explicativas
- En general, cuantas más variables mayor será el R^2
- El R^2 tiende a aumentar aunque las variables que agregamos no sean estadísticamente significativas
- Entonces, trivialmente, a mayor número de variables mejor ajusta el modelo. ¿Por qué?

R^2 no decrece con K

- Suponer el modelo $Y = X\beta + Z\gamma + u$
  X matriz $n \times K_1$ (incluye constante), Z matriz $n \times K_2$, con $K = K_1 + K_2$
- **Problema 1:** regresión de Y en X y Z.
  $\underset{\hat{\beta},\hat{\gamma}}{Min}\ SRC(\hat{\beta},\hat{\gamma}) = (Y - X\hat{\beta} - Z\hat{\gamma})'(Y - X\hat{\beta} - Z\hat{\gamma}) \rightarrow SRC_1^*$
- **Problema 2**: regresión de Y en X solamente
  igual al problema 1 pero le agregamos una restricción: s.a
  $\hat{\gamma} = 0 \rightarrow SRC_2^*$
- Siempre $SRC_1^* \leq SRC_2^* \Rightarrow R_1^2 \geq R_2^2$
  con desigualdad estricta si $\hat{\gamma} \neq 0$

R^2 no decrece con K

- Pensemos en que tenemos un modelo y agregamos una nueva variable. El R^2 aumenta siempre que el coeficiente de regresión estimado de la nueva variable tenga un valor distinto de cero.
- Numéricamente distinto de cero $\neq$ estadísticamente distinto de cero.
- Implicancia: no tiene sentido basarse en el R^2 para comparar el ajuste de modelos con distinta cantidad de variables.
- ¿Entonces?

R^2 ajustado

$$R^2 ajustado = 1 - \frac{SRC}{STC}\frac{(n-1)}{(n-K)}$$

¿Qué pasa con el $R^2 ajustado$ si agregamos una nueva variable al modelo?

- STC no cambia. ¿Por qué?
- $(n-K)$ cae en una unidad
- SRC también cae. ¿Cuánto? Depende...
  - Cuanto más relevante sea la nueva variable, más cae SRC. ¿Por qué?
  - Entonces, cuanto más relevante sea la nueva variable, más aumenta $R^2 ajustado$.
  - Si la variable es poco relevante, $R^2 ajustado$ puede caer

R^2 ajustado

$$R^2 ajustado = 1 - \frac{SRC}{STC} \frac{(n-1)}{(n-K)}$$

- **Intuición.** El R^2*ajustado* incorpora el trade-off que se genera cuando agregamos más variables al modelo: por un lado baja *SRC* y por el otro perdemos grados de libertad, lo que implica estimadores más ineficientes.
- **En la práctica**:
  - R^2*ajustado* aumenta siempre que el estadístico $|t| > 1$. Entonces puede darse que el R^2*ajustado* aumente aunque la nueva variable NO sea estadísticamente significativa.
  - R^2*ajustado* baja siempre que el estadístico $|t| < 1$.

Supuestos clásicos en notación matricial

- Modelo lineal:

$$Y = X\beta + u$$

- Supuestos clásicos:
  1. $E(u) = 0$
  2. $VarCov(u) = \sigma^2 I_n$
  3. X es una matriz $(n \times K)$ no estocástica (no aleatoria) con $\rho(X) = K$ (rango columna completo)

Esperanza de u

- Recordemos que u es un vector aleatorio
- $E(u)$ es el vector de las esperanzas

$$u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \Longrightarrow E(u) = \begin{bmatrix} E(u_1) \\ E(u_2) \\ \vdots \\ E(u_n) \end{bmatrix}$$

- El supuesto 1 establece que el vector de esperanzas es igual al vector nulo, es decir:

$$E(u_i) = 0 \text{ para } i = 1, \ldots n$$

Matriz de varianzas y covarianzas de u

- La matriz de varianzas y covarianzas de un vector aleatorio u se define como:

$$VarCov(u) = E\left[(u - E(u))(u - E(u))'\right]$$

$$= \begin{bmatrix} V[u_1] & Cov[u_1, u_2] & \cdots & Cov[u_1, u_n] \\ Cov[u_2, u_1] & V[u_2] & \cdots & Cov[u_2, u_n] \\ \vdots & \vdots & \ddots & \vdots \\ Cov[u_n, u_1] & Cov[u_n, u_2] & \cdots & V[u_n] \end{bmatrix}$$

- Notar que $VarCov[u]$ es una matriz $n \times n$ y simétrica
- ¿Qué significa entonces suponer que $VarCov(u) = \sigma^2 I_n$?
- Notar: $VarCov(u) = E(uu')$

Propiedades estadísticas de los estimadores MCO

- $\hat{\beta}$ es un vector que contiene K variables aleatorias
- ¿Qué propiedades conocían en el caso de MCO con 2 variables? ¿De qué dependían?
- Vamos a demostrar esas mismas propiedades usando notación matricial

Insesgadez de $\hat{\beta}$

$$\begin{aligned}
\hat{\beta} &= (X'X)^{-1}X'Y \\
&= (X'X)^{-1}X'(X\beta + u) \\
&= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'u \\
&= \beta + (X'X)^{-1}X'u \qquad (*)
\end{aligned}$$

$$\begin{aligned}
E[\hat{\beta}] &= \beta + (X'X)^{-1}X'E[u] \\
&= \beta
\end{aligned}$$

Matriz de varianzas y covarianzas de $\hat{\beta}$

Mostraremos que $VarCov(\hat{\beta}) = \sigma^2 (X'X)^{-1}$

$$\begin{aligned}
VarCov[\hat{\beta}] &= E[(\hat{\beta} - E(\hat{\beta}))(\hat{\beta} - E(\hat{\beta}))'] \\
&= E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] \\
&= E[(X'X)^{-1}X'u((X'X)^{-1}X'u)'] \\
&= E[(X'X)^{-1}X'uu'X(X'X)^{-1}] \\
&= (X'X)^{-1}X'E[uu']X(X'X)^{-1} \\
&= (X'X)^{-1}X'\sigma^2 I_n X(X'X)^{-1} \\
&= \sigma^2 (X'X)^{-1}X'X(X'X)^{-1} \\
&= \sigma^2 (X'X)^{-1}
\end{aligned}$$

- Es importante saber qué supuestos fueron utilizados para obtener esta expresión.

Matriz de varianzas y covarianzas de $\hat{\beta}$

- ¿Qué dimensión tiene $VarCov(\hat{\beta}) = \sigma^2(X'X)^{-1}$?

$$VarCov[\hat{\beta}] = \begin{bmatrix} V[\hat{\beta}_1] & Cov[\hat{\beta}_1,\hat{\beta}_2] & \cdots & Cov[\hat{\beta}_1,\hat{\beta}_K] \\ Cov[\hat{\beta}_1,\hat{\beta}_2] & V[\hat{\beta}_2] & \cdots & Cov[\hat{\beta}_2,\hat{\beta}_K] \\ \vdots & \vdots & \ddots & \vdots \\ Cov[\hat{\beta}_1,\hat{\beta}_K] & Cov[\hat{\beta}_2,\hat{\beta}_K] & \cdots & V[\hat{\beta}_K] \end{bmatrix}$$

- Notar:
  - cada elemento de la diagonal es $V[\hat{\beta}_k] = \sigma^2 A_{kk}$, $k = 1, \ldots K$, donde A_{kk} es el elemento en la fila k y columna k de la matriz $(X'X)^{-1}$
  - cada elemento fuera de la diagonal es $Cov[\hat{\beta}_j, \hat{\beta}_k] = \sigma^2 A_{jk}$, $j \neq k$, donde A_{jk} es el elemento en la fila j y columna k de la matriz $(X'X)^{-1}$

$VarCov(\hat{\beta})$ en la práctica

- $VarCov(\hat{\beta})$ depende de σ^2, un valor desconocido.
- En su lugar usaremos un estimador insesgado:

$$S^2 = \frac{1}{n-K}\sum_{i=1}^{n} e_i^2 = \frac{e'e}{n-K}$$

- Luego, el estimador de la matriz de varianzas y covarianzas es:

$$Var\hat{C}ov(\hat{\beta}) = S^2(X'X)^{-1}$$

Bibliografía para esta clase

- Notas de clase, capítulo 2 secciones 2.1 a 2.9.
- Wooldridge, capítulo 6.3 (R cuadrado y R cuadrado ajustado)