Extensiones del Modelo Lineal

Walter Sosa-Escudero

Universisad de San Andres y CONICET

Interpretacion basica

$Y = X\beta + u$. Bajo exogeneidad:

$$E(Y|X) = X\beta$$

Bajo diferenciabilidad y si las X no estan funcionalmente relacionadas:

$$\frac{\partial E(Y|X)}{\partial X_j} = \beta_j$$

Interpretacion causal? Mas adelante....

Variables binarias

$$Y_i = \beta_1 + \beta_2 D_i + \beta_3 X_i + u_i$$

D_i, binaria (0 o 1). Denota 'evento'.

$$\begin{aligned} E(Y_i|D_i = 1, X) &= \beta_1 + \beta_2 + \beta_3 X_i \\ E(Y_i|D_i = 0, X) &= \beta_1 + \beta_3 X_i \end{aligned}$$

$$\beta_2 = E(Y_i|D_i = 1, X) - E(Y_i|D_i = 0, X)$$

Graficamente....

MCO funciona (por que?)

Lineal en parametros vs. en variables

El modelo lineal es bastante menos lineal que lo que parece....

$$Y = X\beta + u$$

Lineal:

- En variables (X)
- En parametros (β)

MCO: relevante linealidad en ....

Polinomios e interacciones

Polinomios:

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + u_i$$

Interacciones:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{2i} X_{3i} + u_i$$

Ambos no lineales en variables, pero lineales en parametros: MCO funciona. Interpretaciones?

Logs

$$Y_i = b_1 X_i^{b_2} e_i^u$$

$$y_i = \beta_1 + \beta_2 x_i + u_i,$$

$$y_i \equiv \ln Y_i,\ \beta_1 \equiv \ln b_1,\ \beta_2 = B_2, x_i = \ln X_i$$

- Lineal en variables y parametros.
- Es 'linealizable'
- MCO de los logs.
- Ojo con las interpretaciones: $\beta 2$ derivada logaritmica.
- Semilog: $Y_i = \exp(\beta_1 + \beta_2 X_i + u_i)$

Anatomia de MCO

- $\hat{\beta} = (X'X)^{-1}X'Y$
- Mejor: Gauss/Markov
- Bueno: sesgo varianza

Que hace a MCO: a) impreciso, b) sesgado?

Antes, un resultado crucial...

Teorema de Frisch-Waugh-Lovell

$$Y = X\beta + u, \;\; \hat{\beta} = (X'X)^{-1}X'Y$$

Dos matrices, P y M

$$\hat{Y} \equiv X\hat{\beta} = X(X'X)^{-1}X'Y = PY$$
$$e \equiv Y - \hat{Y} = Y - PY = (I - P)Y = MY$$

- $P \equiv X(X'X)^{-1}X'$.
- $M \equiv I - X(X'X)^{-1}X'$.

Propiedades:

- M and P *simetricas:* $M = M', P = P'$
- M and P *idempotentes*: $M'M = M$, $P'P = P$
- $M + P = I, \;\; MP = 0.$
- $PX = X, \;\; MX = 0.$
- $e = MY = M(X\beta + u) = Mu$

M 'genera' (*makes*) residuos de regresar en X.

Demostrar todo. Facil

Frisch-Waugh-Lovell

Modelo lineal: $Y = X\beta + u$

Partamoslo asi: $Y = X_1\beta_1 + X_2\beta_2 + u$

X_1, X_2 matrices de k_1 and k_2 columnas. $X = [X_1 \;\; X_2]$, $\beta' = (\beta_1' \;\; \beta_2')'$ y $K = k_1 + k_2$.

$M_2 \equiv I - X_2(X_2'X_2)^{-1}X_2'$, genera residuos de regresar en X_2.

$Y^* \equiv M_2Y$, $X_1^* \equiv M_2X_1$, respectivamente, residuos de regresar Y en X_2, y todas las columnas de X_1 en X_2.

Supongamos que queremos estimar β_1. Dos metodos:

1. En un paso, regresar Y en X, $\hat{\beta} = (\hat{\beta}_1' \ \ \hat{\beta}_2')' = (X'X)^{-1}X'Y$. $\hat{\beta}_1$ es lo que buscamos.
2. En dos pasos, obtener y regresar Y^* en X_1^*, y tomar $\tilde{\beta}_1 = (X_1^{*\prime} X_1^*)^{-1} X_1^{*\prime}\, Y^*$

Sean e_1 y e_2 los residuos ambos metodos.

Teorema (Frisch y Waugh, 1933, Lovell, 1963): $\hat{\beta}_1 = \tilde{\beta}_1$, y $e_1 = e_2$.

Prueba. Empecemos con:

$$Y = (P + M)Y = PY + MY = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + MY.$$

Multiplicamos ambos lados por $X_1' M_2$:

$$X_1' M_2 Y = X_1' M_2 X_1 \hat{\beta}_1 + X_1' M_2 X_2 \hat{\beta}_2 + X_1' M_2 MY$$

- $M_2 X_2 = 0$ por la propiedad de M_2.
- $X_1' M_2 MY = X_1^{*\prime} e$, con $X_1^{*\prime} \equiv X_1' M_2$ y $e \equiv MY$. e son errores de regresar Y en X_1 y X_2. Notar que:

  $$X'e = X'Mu = 0,$$

  dado que $X'M = 0$. Entonces, todas las columnas de X tienen correlacion nula con e. $X_1^{*\prime}$ son residuos de regresar X_1 en X_2, esto es, la parte de X_1 no explicada linealmente por X_2. Entonces, por construccion $X_1^{*\prime}$ esta correlacionada con X_1 pero no con X_2. Entonces $X_1^{*\prime} e = 0$ ya que e no esta correlacionada con $X_1^{*\prime}$: el tercer termino es nulo.

Esto deja: $X_1' M_2 Y = X_1' M_2 X_1 \hat{\beta}_1$. Entonces:

$$\hat{\beta}_1 = (X_1' M_1 X_1)^{-1} \; X_1' M_1 Y = \tilde{\beta}_1$$

.

Intuicion:

- Dos formas de estimar β_1. Directa: regresar Y en X_1 and X_2. Indirecta: primero eliminar el efecto de X_2.
- Da contenido a la idea de 'controlar por X_2'.
- El modelo de K-variables se reduce a uno con dos variables!
- Muchas aplicaciones!. Ver Davidson and MacKinnon (1993).

Bondad del ajuste: R^2

$$Y_i = \hat{Y}_i + e_i$$

Resultado:

$$\underbrace{\sum (Y_i - \bar{Y}_i)^2}_{SCT} = \underbrace{\sum (\hat{Y}_i - \bar{Y}_i)^2}_{SCE} + \underbrace{\sum e_i^2}_{SCE}$$

$$R^2 \equiv \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$$

Interpretacion?

Prueba: $Y_i = \hat{Y}_i + e_i$. Restemos $\bar{Y}$, entonces $y_i = \hat{y}_i + e_i$ $y_i \equiv Y_i - \bar{Y}$, $\hat{y}_i = \hat{Y}_i - \bar{Y}$. Entonces $\sum y_i^2 = \sum \hat{y}_i^2 + \sum e_i^2 + 2\sum \hat{y}_i e_i$. Necesitamos probar $\sum \hat{y}_i e_i = 0$. $\sum \hat{y}_i e_i = \sum (\hat{Y}_i - \bar{Y}) e_i = \sum \hat{Y}_i e_i - \bar{Y} \sum e_i$. Notar que $\sum \hat{Y}_i e_i = \hat{Y}'e = Y'PMu = 0$, ya que $PM = 0$. Por las condiciones de primer orden del problema de MCO, $X'e = 0$. Si el modelo tiene un intercepto, la primer columna de X es un vector de unos, entonces el primer elemento de $X'e$ es $1'e = \sum e_i = 0$.

Fuentes de imprecision

$$V(\hat{\beta}_j) = \frac{\sigma^2}{n\,(1 - R_j^2)V(X_j)},$$

Resultado crucial. Cuatro factores contribuyen a la varianza:

1. σ^2: ignorancia.
2. $V(X_j)$: variabilidad de X_j.
3. R_j^2: multicolinealidad.
4. n: 'micronumerosidad'.

Tatuense esto en un brazo. Dejen lugar para una mas...

Prueba: Por FWL,

$$\hat{\beta}_j = \frac{\sum_{i=1}^n X_{ji}^* Y_i}{\sum_{i=1}^n X_{ji}^{*2}}$$

y

$$V(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^n X_{ji}^{*2}} = \frac{\sigma^2}{\frac{\sum_{i=1}^n X_{ji}^{*2}}{S_{jj}} S_{jj}}$$

en donde $X_j^* \equiv M_{-j} X_j$ y M_{-j} genera residuos de regresar X_j en el resto de los regresores. El resultado se sigue notando que:

$$R_j^2 = 1 - \frac{\sum_{i=1}^n X_{ji}^{*2}}{S_{jj}} = 1 - \frac{\sum_{i=1}^n X_{ji}^{*2}}{\sum_{i=1}^n (X_{ji} - \bar{X}_j)^2} \quad \square$$

Fuente de sesgos

$$Y = X_1\beta_1 + X_2\beta_2 + u$$

- Interes en β_1. Regresar Y en X_1 y X_2 (o usar TFWL).
- Supongamos que regresamos Y en solo X_1:

$$\hat{\beta}_1^* = (X_1'X_1)^{-1}X_1'Y$$

Resultado (Sesgo por variable omitida):

$$E(\hat{\beta}_1^*|X) = \beta_1 + \underbrace{\delta_{21}\ \beta_2}_{\text{sesgo}},$$

$\delta_{21} = (X_1'X_1)^{-1}X_1'X_2$

- Si $\beta_2 \neq 0$, omitir X_2 no necesariamente induce sesgo.
- Sesgo: relevancia de X_2 y correlacionada con X_1.
- Puedo omitir: variables irrelevantes $(\beta_2 = 0)$ o relevantes no correlacionadas con la variable de interes.

Tatuense esto un brazo.

Prueba:

$$\begin{aligned}\hat{\beta}_1^* &= (X_1'X_1)^{-1}X_1'Y \\ &= (X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + u) \\ E(\hat{\beta}_1^*|X) &= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\ \beta_2\end{aligned}$$

Idea clave: lo que dejemos fuera del modelo pertenece al termino de error.

Modelos grandes vs chicos

Interes en β_1

$$Y = X_1\beta_1 + X_2\beta_2 + u$$

Que hacer con X_2?

- $\beta_2 = 0$. Omitir. Por que?
- $\beta_2 \neq 0$
  - $\delta_{21} = 0$, se puede omitir X_2.
  - $\delta_{21} \neq 0$, incluir X_2 para evitar sesgo.

En la practica, incluir X_2 o no?

Sesgo por variables omitidas: un ejemplo

Datos ficticios, compatibles con Appleton, French and Vanderpump ("Ignoring a Covariate: an Example of Simpon's Paradox", The American Statistician, 50, 4, 1996)

- Y = riesgo de muerte.
- $SMOKE$ = consumo de cigarrillos.

```
. reg y smoke

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =  194.34
       Model |  7613.25147     1  7613.25147           Prob > F      =  0.0000
    Residual |  3839.18734    98  39.1753811           R-squared     =  0.6648
-------------+------------------------------           Adj R-squared =  0.6614
       Total |  11452.4388    99    115.6812           Root MSE      =   6.259

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |  -1.819348   .1305081   -13.94   0.000    -2.078337   -1.560359
       _cons |   158.5975   4.774249    33.22   0.000     149.1231    168.0718
------------------------------------------------------------------------------
```

```
. reg y smoke age

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  2,    97) = 5424.58
       Model |  11350.9524     2  5675.47622           Prob > F      =  0.0000
    Residual |  101.486373    97  1.04625126           R-squared     =  0.9911
-------------+------------------------------           Adj R-squared =  0.9910
       Total |  11452.4388    99    115.6812           Root MSE      =  1.0229

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |   .9431267    .050902    18.53   0.000     .8421004    1.044153
         age |   .9804631   .0164039    59.77   0.000     .9479059     1.01302
       _cons |   12.84084   2.560392     5.02   0.000     7.759169    17.92251
------------------------------------------------------------------------------

. cor y smoke age
(obs=100)

             |        y    smoke      age
-------------+---------------------------
           y |   1.0000
       smoke |  -0.8153   1.0000
         age |   0.9797  -0.9080   1.0000
```

Toda la econometria basica en 3 formulas

1. MCO: $\hat{\beta} = (X'X)^{-1}X'Y$
2. Varianza: $V(\hat{\beta}_j) = \frac{\sigma^2}{n\ (1-R_j^2)V(X_j)}$
3. Sesgo: $E(\hat{\beta}_1^*|X) = \beta_1 + \delta_{21}\ \beta_2$

Propiedades asintoticas

Por que una teoria asintotica?

- $\hat{\beta} = (X'X)^{-1}X'Y$, $E(\hat{\beta}) = \beta$, facil, linealidad (dos veces).
- $\hat{\beta}^* = g(X, Y)$?
- Tuvimos que *suponer* normalidad para hacer inferencia.
- Muestra grande como aproximacion util

Convergencia

$\{z_n\}$ converge en probabilidad a una constante no aleatoria a si para todo $\epsilon > 0$:

$$\lim_{n \to \infty} Pr\left[|z_n - a| > \epsilon\right] = 0$$

$\{z_n\}$ con FDA $F_n(z)$ converge en distribucion a una VA escalar z con FDA F si

$$\lim_{n \to \infty} F_n(z) = F(z)$$

para todos los puntos de continuidad de $F(z)$.

- Notacion: $z_n \stackrel{p}{\rightarrow} a$, o *plim* $z_n = a$. Se extiende a vectores or matrices, elemento a elemento.
- Notacion: $z_n \stackrel{d}{\rightarrow} z$.
- F es la *distribucion limite* de $\{z_n\}$.

LGN y TCL

$$\{\bar{z}_n\} \equiv \frac{\sum_{i=1}^{n} z_i}{n}$$

Ley de Grandes Numeros (Kolmogorov): $\{z_n\}$ iid, $E(z_i) = \mu$ finita, entonces

$$\bar{z}_n \xrightarrow{p} \mu$$

.

Teorema Central del Limite (Lindeberg-Levy): $\{z_n\}$ iid con $E(z_i) = \mu$ y $V(z_i) = \sigma^2$ finitas, entonces

$$\sqrt{n}\,\frac{\bar{z}_n - \mu}{\sigma} \xrightarrow{d} N(0,1)$$

Resultados de continuidad

Si $g(.)$ continua:

- $z_n \xrightarrow{p} c \Rightarrow g(z_n) \xrightarrow{p} g(c)$
- $z_n \xrightarrow{d} z \Rightarrow g(z_n) \xrightarrow{d} g(z)$

Slutzky. $z_n \xrightarrow{d} z$ y $c_n \xrightarrow{p} c$:

- $z_n + c_n \xrightarrow{d} z + c$
- $z_n c_n \xrightarrow{d} zc$
- $z_n / c_n \xrightarrow{d} z/c, \quad c \neq 0$

Cramer Wold Device: sean z_n, z, vectores de k VA's. $z_n \xrightarrow{d} z$ sii $\lambda' z_n \xrightarrow{d} \lambda' z, \forall \lambda \in \Re^k$

Estimadores como sucesiones

Un estimador $\hat{\theta}_n$ es una funcion de la muestra. $\{\hat{\theta}_n\}_n$ refiere a una sucesion de estimadores, aumentando la muestra progresivamente.

- $\hat{\theta}_n$ es consistente para θ_0 si $\hat{\theta}_n \xrightarrow{p} \theta_0$.
- $\hat{\theta}_n$ es asintoticamente normal si $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \Sigma)$.
- Σ es la varianza asintotica. Cuidado!
- A estos estimadores se los llama $\sqrt{n}$-consistentes.

Notacion

Reescribamos el modelo lineal

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + u_i, \qquad i = 1, \ldots, n$$

como: $y_i = x_i'\beta + u_i$, con x_i un vector $K \times 1$, $(x_{1i}, x_{2i}, \ldots, x_{Ki})'$.

$x_i x_i'$ es una $K \times K$ *matriz*. Verificar

- $X'X = \sum_{i=1}^n x_i x_i'$, $X'Y = \sum_{i=1}^n x_i y_i$
- $\hat{\beta}_n = (X'X)^{-1} X'Y = \left(\sum_{i=1}^n x_i x_i'\right)^{-1} \left(\sum_{i=1}^n x_i y_i\right)$

Supuestos para analisis asintotico

1. Linealidad: $y_i = x_i'\beta_0 + u_i \qquad i = 1, \ldots, n.$
2. Muestra aleatoria: $\{x_i, u_i\}$ iid.
3. Predeterminacion: $E(x_{ik}u_i) = 0$ para todo i y $k = 1, \ldots, K$.
4. Rango: $\Sigma_x \equiv E(x_i x_i')$ positiva.
5. $V(x_i u_i) = S$ positiva.

Resultados

- Consistencia: $\hat{\beta}_n \xrightarrow{p} \beta_0$.
- Normalidad asintotica: $\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) \xrightarrow{p} N(0, \Sigma_x^{-1} S \Sigma_x^{-1})$

Caso importante: si $E(u_i^2|x_i) = \sigma_0^2$, entonces usando la LEI:

$$S = V(x_i u_i) = E(x_i u_i u_i x_i') = \sigma_0^2 E(x_i x_i') = \sigma_0^2 \; \Sigma_x$$

Entonces

$$\Sigma_x^{-1} S \Sigma_x^{-1} = \sigma_0^2 \Sigma_x^{-1} \Sigma_x \Sigma_x^{-1} = \sigma_0^2 \Sigma_x^{-1} = \sigma_0^2 E(X'X)^{-1},$$

nuestro resultado anterior. En sintesis: no hace falta homocedasticidad para normalidad asintotica.

Consistencia

$$\begin{aligned}\hat{\beta}_n &= \beta_0 + (X'X)^{-1}X'u \\ &= \beta_0 + \left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'u}{n}\right)\end{aligned}$$

Idea: mostrar a) $n^{-1}X'u \xrightarrow{p} 0$ y $\left(\frac{X'X}{n}\right)^{-1}$ finita.

El elemento $h-$esimo de $\left(\frac{X'u}{n}\right)$ es

$$\frac{1}{n}\sum_{i=1}^{n} x_{hi}u_i = \frac{1}{n}\sum_{i=1}^{n} z_i, \qquad z_i \equiv x_{hi}u_i$$

con $E(z_i) = 0$. Por la LGN de Kolmogorov

$$\frac{1}{n}\sum_{i=1}^{n} z_i = \frac{\sum_{i=1}^{n} x_{hi}u_i}{n} \quad \xrightarrow{p} \quad E(x_{hi}u_i) = 0$$

entonces

$$\left(\frac{X'u}{n}\right) \xrightarrow{p} 0$$

El elemento (h, j) de $\frac{X'X}{n}$ es $\frac{\sum_{i=1}^{n} x_{hi} x_{ji}}{n}$. Como $E(x_{hi} x_{ji}) = \Sigma_{x,hj}$ y es finita, por LGN (elemento a elemnto)

$$\frac{X'X}{n} \xrightarrow{p} \Sigma_x$$

Dado que Σ_x es invertible y por continuidad:

$$\left(\frac{X'X}{n}\right)^{-1} \xrightarrow{p} \Sigma_x^{-1}$$

En sintesis:

$$\hat{\beta}_n = \beta_0 + \underset{\xrightarrow{p} \Sigma_x^{-1}}{\left(\frac{X'X}{n}\right)^{-1}} \quad \underset{\xrightarrow{p} 0}{\left(\frac{X'u}{n}\right)}$$

Por la regla del producto: $\hat{\beta}_n \xrightarrow{p} \beta_0$

Normalidad asintotica
Partimos de

$$\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) = \left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'u}{\sqrt{n}}\right)$$

Ya mostramos:

$$\left(\frac{X'X}{n}\right)^{-1} \xrightarrow{p} \Sigma_x^{-1}$$

Encontraremos la distribucion asintotica de $\left(\frac{X'u}{\sqrt{n}}\right)$ y luego usamos continuidad.

$$\left(\frac{X'u}{\sqrt{n}}\right) = \sqrt{n}\,\frac{X'u}{n}$$

es un vector de K VA's. Por el 'artilugio' de Cramer-Wold, buscamos la distribucion asintotica de:

$$\sqrt{n}\ c'\left(\frac{X'u}{n}\right)$$

para cualquier vector $c \in \Re^K$. Notar que:

$$\sqrt{n}\ c'\left(\frac{X'u}{n}\right) = \sqrt{n}\ c'\,\frac{\sum x_i u_i}{n} = \sqrt{n}\,\frac{\sum c' x_i u_i}{n} \equiv \sqrt{n}\ \frac{\sum z_i}{n}$$

con $z_i \equiv c' x_i u_i$, una VA escalar.

Es facil chequear que:

- $E(z_i) = 0$
- $V(z_i) = c'Sc < \infty$.

Entonces, por el TCL aplicado a $\sqrt{n}\bar{z}$:

$$\sqrt{n}\;(\bar{z} - 0) = c' \left(\frac{X'u}{\sqrt{n}} \right) \xrightarrow{d} N(0, c'Sc)$$

y de vuelta por Cramer-Wald:

$$\left(\frac{X'u}{\sqrt{n}} \right) \xrightarrow{d} N(0, S)$$

Entonces:

$$\begin{array}{ccccc} \sqrt{n}\left(\hat{\beta}_n - \beta_0\right) & = & \left(\frac{X'X}{n}\right)^{-1} & & \left(\frac{X'u}{\sqrt{n}}\right) \\ & & \xrightarrow{p} \Sigma_x^{-1} & & \xrightarrow{d} N(0, S) \end{array}$$

Por el teorema de Slutzky's y linealidad de la normal:

$$\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) \xrightarrow{p} N(0, \Sigma_x^{-1} S \Sigma_x^{-1})$$

En sintesis

Bajo supuestos muy generales (clasicos)

1. $\hat{\beta} \xrightarrow{p} \beta$.
2. $\hat{\beta}|X \sim N(\beta, \sigma^2(X'X)^{-1})$

La normal aparece como resultado asintotico.

Nos va a permitir pensar mucho mas alla del modelo lineal y MCO. Relevancia de los resultados de continuidad.