Teoría Econométrica I - EAE- 250-A

Mínimos cuadrado no lineales y multicolinealidad

Tatiana Rosá

Instituto de Economía - Pontificia Universidad Católica de Chile

Septiembre/Octubre 2021

Introducción

- En la sección pasada relajamos el supuesto de matriz de varianzas esférica, centrándonos en heteroscedasticidad y autocorrelación
- En esta sección haremos los siguientes supuestos:

  1. Relajamos que $\mathbb{E}(y|x,\beta)$ es lineal en parámetros.
  2. La forma funcional es conocida.

  En general, $\mathbb{E}(y|x,\theta)$=$m(x,\theta)$. Donde, $m(x,\theta)$ es conocida y diferenciable. Ejemplos:

$$m(x,\theta) = \theta_1 + \frac{\theta_2 x}{1+\theta_3 x}$$

$$m(x,\theta) = \theta_1 + \theta_2 exp\{\theta_3 x\}$$

Estimación

- Vamos a definir $\hat{\theta}_{MCNL}$
$$\hat{\theta}_{MCNL} = arg\ min\ S_N(\theta)$$
donde:
$$S_N(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - m(x,\theta))^2$$
- El problema en general es que no existe solución analítica, ya que la primera derivada puede ser no lineal
- Esto implica que para estimar los parámetros debemos usar métodos numéricos, obteniendo sólo un valor y no una forma funcional del estimador

Gauss-Newton

- Dada la continuidad y diferenciabilidad de $m(x, \theta)$, podemos aplicar el teorema de Taylor de primer orden
- Si lo hacemos en torno a θ_0, tenemos que:

$$m(x, \theta) \cong m(x_i, \theta_0) + \frac{\partial m(x_i, \theta)}{\partial \theta}|_{\theta_0} (\theta - \theta_0)$$

- Reemplazando en $S_n(\theta)$

$$S_n(\theta) \cong \frac{1}{n} \sum_{i=1}^{n} \left[y_i - m(x_i, \theta_0) - \frac{\partial m(x_i, \theta)'}{\partial \theta}|_{\theta_0} (\theta - \theta_0) \right]^2$$

- Repasemos las dimensiones de cada elemento...

Gauss-Newton

- Como es cuadrático en θ, las condiciones de primer orden con respecto a θ son lineales.

$$\underline{CPO}: -\frac{2}{n}\sum_{i=1}^{n}\frac{\partial m(x_i,\theta)}{\partial\theta}|_{\theta_0}\left[y_i - m(x_i,\theta_0) - \frac{\partial m(x_i,\theta)'}{\partial\theta}|_{\theta_0}\,(\theta-\theta_0)\right] = 0$$

- Tenemos k incógnitas, luego son k ecuaciones:

$$\sum_{i=1}^{n}\left[\frac{\partial m(x_i,\theta)}{\partial\theta}|_{\theta_0}[y_i - m(x_i,\theta_0)] - (\hat{\theta}-\theta_0)\sum_{i=1}^{n}\frac{\partial m(x_i,\theta)}{\partial\theta}'|_{\theta_0}\frac{\partial m(x_i,\theta)}{\partial\theta}|_{\theta_0}\right] = 0$$

- De esta manera obtenemos que:

$$\hat{\theta} = \theta_0 + \left[\sum_{i=1}^{n}\frac{\partial m(x_i,\theta)}{\partial\theta}|_{\theta_0}\frac{\partial m(x_i,\theta)'}{\partial\theta}|_{\theta_0}\right]^{-1}\left[\sum_{i=1}^{n}\frac{\partial m(x_i,\theta)}{\partial\theta}'|_{\theta_0}(y_i - m(x_i,\theta_0)\right]$$

Ejemplo

- Tomemos este ejemplo:

$$m(x, \theta) = \frac{\theta_1^2}{2} + exp\{\theta_2 x\}$$

- $S(\theta)$; $\frac{\partial m(x_i, \theta)}{\partial \theta}'$; Aproximación por taylor; $\hat{\theta}$

Gauss-Newton

- Así vemos que Gauss-Newton es un algortimo iterativo, asi

$$\theta_0 = \theta_{j-1}$$

$$\hat{\theta}_{MCNL} = \hat{\theta}_j$$

$$\Rightarrow \hat{\theta}_j = \hat{\theta}_{0_{j-1}} + \underbrace{\left[\sum_{i=1}^{n} \frac{\partial m(x_i,\theta)}{\partial \theta}' |_{\theta_{j-1}} \frac{\partial m(x_i,\theta)}{\partial \theta} |_{\theta_{j-1}}\right]^{-1}}_{size} \underbrace{\sum_{i=1}^{n} \left[\frac{\partial m(x_i,\theta)}{\partial \theta}' |_{\theta_{j-1}} (y_i - m(}_{step}$$

Pasos del algoritmo

1. Valores iniciales:
   1.1 Teoría
   1.2 Truco: Alterar la función con solución analítica.
   1.3 Valores de MCO, pero nos podemos equivocar.
   1.4 Graficar la función.
2. Iteración
3. Stopping Rule
   3.1 Absolutas: $||\theta_0 - \theta_{j-1}|| <$ *tolerancia*, generalmente esta tolerancia es 10^{-6} para los computadores.
   3.2 relativas: $\frac{||\theta_0 - \theta_{j-1}||}{||\theta_{j-1}||} <$ *tolerancia*

Distribución Asintótica

- No es difícil llegar a esta expresión:

$$\sqrt{n}(\hat{\theta}-\theta) = \left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial m(x_i,\theta)}{\partial\theta}'|_{\theta}\frac{\partial m(x_i,\theta)}{\partial\theta}|_{\theta}\right]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{\partial m(x_i,\theta)}{\partial\theta}'|_{\theta_0}\underbrace{[y_i - m(x_i}_{\varepsilon_i}\right.$$

$$\sqrt{n}(\hat{\theta}-\theta) \xrightarrow{d} N(0, V_{\theta})$$

- Por simplicidad asumamos que:

$$m_{\theta i} = \frac{\partial m(x_i,\theta)}{\partial\theta}|_{\hat{\theta}}$$

- Luego,

$$\mathbb{E}(m'_{\theta i}m_{\theta i}) = \mathbb{E}\left[\frac{\partial m(x_i,\theta)}{\partial\theta}'\frac{\partial m(x_i,\theta)}{\partial\theta}\right]$$

$$\mathbb{E}(m'_{\theta i}m_{\theta i}\varepsilon_i^2) = E\left[\frac{\partial m(x_i,\theta)}{\partial\theta}'\frac{\partial m(x_i,\theta)}{\partial\theta}\varepsilon_i^2\right]$$

Distribución Asintótica

- Luego tenemos que la varianza es

$$V_\theta = (\mathbb{E}(m_{\theta i} m'_{\theta i}))^{-1} (\mathbb{E}(m_{\theta i} m'_{\theta i} \varepsilon_i^2))(\mathbb{E}(m_{\theta i} m'_{\theta i}))^{-1}$$

- Ahora, definiendo,

$$\left[\frac{1}{n}\sum_{i=1}^{n} \hat{m}'_{\theta i}\hat{m}_{\theta i}\right] = \left[\frac{1}{n}\sum_{i=1}^{n} \frac{\partial m(x_i,\theta)}{\partial\theta}'|_{\hat{\theta}} \frac{\partial m(x_i,\theta)}{\partial\theta}|_{\hat{\theta}}\right]$$

$$\left[\frac{1}{n}\sum_{i=1}^{n} \hat{m}'_{\theta i}\hat{m}_{\theta i}\varepsilon_i^2\right] = \frac{1}{n}\sum_{i=1}^{n} \left[\frac{\partial m(x_i,\theta)}{\partial\theta}'|_{\hat{\theta}} \frac{\partial m(x_i,\theta)}{\partial\theta}|_{\hat{\theta}}\ \varepsilon_i^2\right]$$

- la varianza asintótica estimada nos quedará,

$$\hat{V}_\theta = \left[\frac{1}{n}\sum_{i=1}^{n} \hat{m}'_{\theta i}\hat{m}_{\theta i}\right]^{-1} \left[\frac{1}{n}\sum_{i=1}^{n} \hat{m}'_{\theta i}\hat{m}_{\theta i}\varepsilon_i^2\right] \left[\frac{1}{n}\sum_{i=1}^{n} \hat{m}'_{\theta i}\hat{m}_{\theta i}\right]^{-1}$$

Multicolinealidad

- La multicolinealidad aparece cuando las variables explicativas en el modelo econométrico están correlacionadas entre sí.
- Si la multicolinealidad es exacta, entonces $rango(X) < K$ y el estimador MCO no existe.
- Estudiaremos la multicolinealidad "no exacta": $rango(X) = K$ pero el $det(X'X)$ es muy bajo y daña la inferencia.
- La multicolinealidad es un problema de naturaleza muestral y no tiene una manera única de ser detectada. Lo que sí tiene son algunas reglas prácticas. **Casos sospechosos de multicolinealidad:**
    - El R^2 es alto, pero los parámetros no resultan ser individualmente significativos.
    - Pequeños cambios en los datos producen importantes variaciones en las estimaciones mínimo cuadráticas.
    - Los coeficientes pueden tener signos opuestos a los esperados o una magnitud poco creíble.

Metodos para detectar multicolinealidad - 1

(a) Método basado en la correlación entre variables explicativas:

- Una de las consecuencias de la multicolinealidad es que la varianza de los estimadores es bastante alta (porque $(X'X)^{-1}$ es muy pequeño)
- Si descomponemos la matriz X de la siguiente forma:

$$X = [x_j; X_j]$$

donde x_j es un vector columna correspondiente a la j-ésima variable explicativa y X_j es una matriz de $n \times (k-1)$ con las observaciones de las variables restantes. Entonces $X'X$ puede escribirse como

$$X'X = \begin{pmatrix} x_j' x_j & x_j' X_j \\ X_j' x_j & X_j' x_j \end{pmatrix}$$

Metodos para detectar multicolinealidad - 1

- De esta forma, el elemento (1,1) de $(X'X)^{-1}$, de acuerdo a la formula de la inversa por bloques, es:

$$[(x_j'x_j) - x_j'X_j(X_j'X_j)^{-1}(X_j'x_j)]^{-1} = (x_j'M_jx_j)^{-1}$$

donde $M_j = I_n - X_j(X_j'X_j)^{-1}X_j'$ y donde $x_j'M_jx_j$ corresponde a la suma de los residuos al cuadrado de una regresión de x_j sobre X_j, de esta forma se siente que:

$$Var(\hat{\beta}_j) = \frac{\sigma_u^2}{x_j'M_jx_j}$$

- Lo que tiene la siguiente expresión:

$$Var(\hat{\beta}_j) = \frac{\sigma_u^2}{ST_j(1 - R_j^2)}$$

donde ST_j es la suma de total al cuadrado ($ST_j = \sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2$) y R_j^2 es el coeficiente R^2 de esta misma regresión (con constante).

Metodos para detectar multicolinealidad - 1

- La varianza de $\hat{\beta}_j$ depende de tres cosas:
  - (i) La varianza del término de error, que es independiente del grado de correlación entre las x's.
  - (ii) La suma total propia de la variable x_j, la que depende sólo de esta variable.
  - (iii) El coeficiente de determinación R_j^2, el que si depende del grado de correlación entre la variable x_j y las restantes, es decir, depende del grado de multicolinealidad.
- La cota inferior para la varianza de $\hat{\beta}_j$, cuando $R_j^2 = 0$, es:

$$Var(\hat{\beta}_j^0) = \frac{\sigma_u^2}{ST_j}$$

Por lo que la relación entre las varianzas de la estimación de β_j en un caso de correlación entre variables explicativas y el caso de independencia es el **Variance Inflation Factor**

$$VIF = \frac{Var(\hat{\beta}_j)}{Var(\hat{\beta}_j^0)} = \frac{1}{1 - R_j^2}$$

Ejemplo usando el VIF

- **EJEMPLO 1**: si $VIF = 9$ para j, el error estándar es 3 veces mayor que el caso con $VIF = 1$. El test-t queda dividido por 3.
  De acuerdo con este análisis, los coeficientes de determinación obtenidos en las regresiones de cada variable explicativa con el resto son un buen indicador de una posible situación de multicolinealidad.

Metodos para detectar multicolinealidad - 2

(b) Método basado en el tamaño de la matriz X'X:

- Cuando tenemos multicolinealidad la matriz X'X es casi singular, de esta manera una medida de tamaño de esta matriz nos permite detectar la presencia de multicolinealidad.
- El determinante no es una medida buena, ya que tiene problemas de sensibilidad a los cambios de unidades.
- Pero sabemos que el determinante de una matriz simétrica es igual al producto de sus valores propios, y por lo tanto el examen de estos valores nos da una idea del tamaño de la matriz.
- Belsey propone la siguiente medida para ver el grado de multicolinealidad:

$$\gamma = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$$

- Esta medida se denomina ***número de condición*** de la matriz X, y números de este indicador mayores a 25 o 30 suelen considerarse problemáticos.

Metodos para detectar multicolinealidad - 2

- Los λ's corresponden a los valores propios de la matriz $B = S(X'X)S$, donde S es la siguiente matriz diagonal:

$$H = \begin{bmatrix} \frac{1}{\sqrt{x_2'x_2}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{x_3'x_3}} & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{\sqrt{x_k'x_k}} \end{bmatrix}$$

- Esta matriz nos permite librarnos del problema de unidad en el tamaño de los valores propios, ya que normaliza cada una de las variables al dividir todas las observaciones por la raíz de la norma euclidiana.
- El número de condición de la matriz $X(\gamma)$ implica que mientras mayor es este valor, el valor de λ_{min} es realmente pequeño al compararlo con λ_{max}, indicando el potencial problema de multicolinealidad

¿Cómo podemos solucionarlo?

- La solución más sencilla es eliminar de la regresión las variables que se sospeche son la causa de la multicolinealidad. Obviamente de este método surge problemas de especificación, como la omisión de variables relevantes.
- Una alternativa es usar el **estimador de "ridge"**:

$$\hat{\beta}_r = (X'X + \lambda I)^{-1}X'Y$$

  Donde $\lambda > 0$ es un parámetro que ayuda a corregir el problema. Este estimador es sesgado pero posee menor varianza que el estimador MCO.
- Note que

$$\mathbb{E}[\hat{\beta}_r|X] = (X'X + \lambda I)^{-1}X'X\beta$$

  y su varianza condicional está dada por:

$$Var(\hat{\beta}_r|X) = \sigma^2(X'X + \lambda I)^{-1}X'X(X'X + \lambda I)^{-1}$$

- Note que

$$\mathbb{E}[\hat{\beta}_r|X] = (X'X + \lambda I)^{-1}X'X\beta$$

  y su varianza condicional está dada por:

$$Var(\hat{\beta}_r|X) = \sigma^2(X'X + \lambda I)^{-1}X'X(X'X + \lambda I)^{-1}$$