Apuntes de Econometría I

Alex Carrasco Martinez
(alex.carmar93@gmail.com)

Estudiante de primer año (maestría) - PUC Rio

Abril - Julio, 2017

Observación

El presente documento forma parte de mis notas del curso de Econometría I brindado por la PUC de Rio de Janeiro en la maestría de Economía período 2017.I. *Este documento no es oficial y no tiene ninguna relación adicional con la mencionada universidad*. Todos los errores son enteramente culpa mía.

Contents

1 Convergencia de variables aleatorias

Basado en Lukacs (1975), Newey and McFadden (1994), (White, 2001, chapter 2), y clases dictadas. Durante toda esta sección consideraré tanto variables (x_n) como vectores aleatorios $(\mathbf{x}_n)$. Dado un espacio de probabilidad $(\Omega, \mathcal{F}, \mathcal{P})$, una variable aleatoria es una función (medible) $x_n : \Omega \to \mathbb{R}$, mientras que un vector aleatorio es un vector de funciones (medibles) $\mathbf{x}_n : \Omega \to \mathbb{R}^k$.

1.1 Convergencia casi segura

Este es el concepto más cercano a la convergencia de secuencias determinísticas:

Definición 1.1. *Considere la secuencia de variables aleatorias $\{x_n(.)\}$, luego decimos que $x_n(.)$ converge de manera casi segura a $c \in \mathbb{R}$ o $x_n \xrightarrow{a.s.} c$ si*

$$\mathbb{P}[w|x_n(w) \to c] = 1 \tag{1.1}$$

- Note que en la definición se toma en cuenta la distribución conjunta de toda la secuencia, es decir, la probabilidad que toda la secuencia satisfaga cierta condición.
- La convergencia casi segura quiere decir que la probabilidad de todos los eventos que implican $x_n(w)$ converge a c, es igual a uno.
- Convergencia casi segura de secuencia de vectores aleatorios $\{\mathbf{x}_n(.)\}$ indica convergencia casi segura de cada elemento del vector aleatorio.

Teorema 1.1. ***(Teorema del Mapeo Contínuo I -TMC I)*** *Dado $g : \mathbb{R}^k \to \mathbb{R}^m$ y $\{\mathbf{x}_n(.)\}$ tal que $\mathbf{x}_n \xrightarrow{a.s.} c$. Si g es contínua en c, luego*

$$g(\mathbf{x}_n) \xrightarrow{a.s.} g(c) \tag{1.2}$$

TCM I indica conergencia de funciones de vectores aleaotorios hacia un valor (una secuencia determinística de constantes), sin embargo, a veces no es posible garantir este tipo de convergencia sino más bien que x_n comienza a seguir una secuencia determinística mientras $n \to \infty$.

Teorema 1.2. ***(Teorema del Mapeo Contínuo II - TCM II)*** *Dado $g : \mathbb{R}^k \to \mathbb{R}^m$ sobre un conjunto compacto $\mathcal{X} \subset \mathbb{R}^k$, $\{\mathbf{x}_n(.)\}$ tal que $\mathbf{x}_n \xrightarrow{a.s.} c$, y c_n una secuencia de vectores $k \times 1$ tal que $x_n - c_n \xrightarrow{a.s.} 0$. Si para algún n suficientemente grande, $c_n \in int(\mathcal{X})$, luego*

$$g(\mathbf{x}_n) - g(c_n) \xrightarrow{a.s.} 0 \tag{1.3}$$

Ahora introduciré un concepto que se deriva del de convergencia, que es muy útil para facilitar la notación.

Definición 1.2.

1. *$\{x_n\}$ es al menos de orden n^λ de manera casi segura $[x_n = O_{a.s.}(n^\lambda)]$ si $\exists\ \Delta < \infty$ and N tal que*

$$\mathbb{P}[|n^{-\lambda}x_n| < \Delta \ \ \forall n \geq N] = 1 \tag{1.4}$$

2. *$\{x_n\}$ es orden menor a n^λ de manera casi segura $[x_n = o_{a.s.}(n^\lambda)]$ si*

$$n^{-\lambda}x_n \xrightarrow{a.s.} 0 \tag{1.5}$$

1.2 Convergencia en probabilidad

Definición 1.3. *Considere la secuencia de variables aleatoria* $\{x_n(.)\}$*, luego decimos que* $x_n(.)$ *converge en probabilidad a* $c \in \mathbb{R}$ $[x_n \xrightarrow{p} c]$ *si para todo* $\epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}\Big[w\Big||x_n(w) - c| < \epsilon\Big] = 1 \tag{1.6}$$

- Note que en la definición se toma en cuenta la distribución marginal de cada elemento de la secuencia.
- La convergencia en probabilidad quiere decir que la probabilidad de todos los eventos que implican $|x_n(w) - c| < \epsilon$, converge a uno.
- Convergencia en probabilidad de secuencia de vectores aleatorios $\{\mathbf{x}_n(.)\}$ indica convergencia en probabilidad de cada elemento del vector aleatorio.
- La diferencia principal la convergencia casi segura y la convergencia en probabilidad es que la última converge de una manera más irregular, de ahí el siguiente teorema.

Teorema 1.3. *Dado una secuencia de variables aleatorias* $\{x_n(.)\}$,

1. *Si* $x_n \xrightarrow{a.s.} c$*, entonces* $x_n \xrightarrow{p} c$.
2. *Si* $x_n \xrightarrow{p} c$*, entonces existe una subsecuencia* $x_{n_j}(.)$ *tal que* $x_{n_j} \xrightarrow{a.s.} c$.

Proof. Ver Lukacs (1975). □

En Lukacs (1975) se llega a una desigualdad que puede ayudarnos a entender la diferencia entre ambos tipos de convergencia. Defina $S_{n,\epsilon} = \cap_{j=n}^{\infty}\{w : |x_j(w) - c| < \epsilon\}$, note que $S_{n,\epsilon} \subset T_{n,\epsilon} = \{w : |x_n(w) - c| < \epsilon\}$

$$\lim_{n\to\infty} \mathbb{P}(S_{n,\epsilon}) \leq \lim_{n\to\infty} \mathbb{P}(T_{n,\epsilon}) = \lim_{n\to\infty} \mathbb{P}(w : |x_n - c| < \epsilon) \leq 1$$

Note que $S_{n,\epsilon} \subset S_{n+1,\epsilon}$, entonces $\mathbb{P}(S_{n,\epsilon}) \leq \mathbb{P}(S_{n+1,\epsilon})$, así la secuencia de probabilidades del lado izquierdo es una secuencia creciente, es decir, tiene un comportamiento más regular que la del lado derecho. Lukacs (1975) muestra que convergencia casi segura es equivalente a $\lim_{n\to\infty} \mathbb{P}(S_{n,\epsilon}) = 1$, luego esta convergencia implica convergencia en probabilidad.

Teorema 1.4. ***(Teorema del Mapeo Contínuo III -TMC III)*** *Dado* $g : \mathbb{R}^k \to \mathbb{R}^m$ *y* $\{\mathbf{x}_n(.)\}$ *tal que* $\mathbf{x}_n \xrightarrow{p} c$*. Si* g *es contínua en* c*, luego*

$$g(\mathbf{x}_n) \xrightarrow{p} g(c) \tag{1.7}$$

De igual forma se puede extender este teorema para el caso de convergencia en probabilidad.

Teorema 1.5. ***(Teorema del Mapeo Contínuo IV - TCM IV)*** *Dado* $g : \mathbb{R}^k \to \mathbb{R}^m$ *sobre un conjunto compacto* $\mathcal{X} \subset \mathbb{R}^k$, $\{\mathbf{x}_n(.)\}$ *tal que* $\mathbf{x}_n \xrightarrow{a.s.} c$*, y* c_n *una secuencia de vectores* $k \times 1$ *tal que* $x_n - c_n \xrightarrow{p} 0$*. Si para algún* n *suficientemente grande,* $c_n \in int(\mathcal{X})$*, luego*

$$g(\mathbf{x}_n) - g(c_n) \xrightarrow{p} 0 \tag{1.8}$$

Asimismo, se puede definir la notación de orden.

Definición 1.4.

1. $\{x_n\}$ *es al menos de orden* n^{λ} *en probabilidad* $[x_n = O_p(n^{\lambda})]$ *si para todo* $\epsilon > 0$ *existe* $\Delta_\epsilon > 0$ *tal que*

$$\lim_{n\to\infty} \mathbb{P}\Big[w\Big||n^{-\lambda}x_n| \geq \Delta_\epsilon\Big] = 0 \tag{1.9}$$

2. $\{x_n\}$ *es orden menor a* n^{λ} *en probabilidad* $[x_n = o_p(n^{\lambda})]$ *si*

$$n^{-\lambda}x_n \xrightarrow{p} 0 \tag{1.10}$$

1.3 Convergencia en media r-ésima

Definición 1.5. *Sea $\{x_n(.)\}$ una secuencia de variables aleatorias tal que para algún $r > 0$, $\mathbb{E}|x_n|^r < \infty$. Decimos que x_n converge en media $r-$ésima $[x_n \xrightarrow{r.m.} c]$ si*

$$\mathbb{E}(|x_n - c|^r) \to 0 \tag{1.11}$$

Es fácil mostrar que si $x_n \xrightarrow{r.m.} c$ entonces $x_n \xrightarrow{s.m.} c$ para algún $s \leq r$. Es también útil presentar una desigualdad muy usada en teoría asintótica

Desigualdad de Chebyshev generalizada *Si Z es una variable aleatoria tal que $\mathbb{E}|Z|^r < \infty$ con $r > 0$, entonces para todo $\epsilon > 0$*

$$\mathbb{P}[|Z| \geq \epsilon] \leq \frac{\mathbb{E}|Z|^r}{\epsilon^r} \tag{1.12}$$

1.4 Convergencia en distribución

Definición 1.6. *Sea $x_n(.)$ una secuencia de variables aleatorias con c.d.f. $\{F_n\}$ y $Z \sim F$. Decimos que x_n converge en distribución hacia Z $[x_n \xrightarrow{d} Z]$ si*

$$F_n(a) \to F(a) \tag{1.13}$$

para todo punto a donde F es contínua.

Para el caso de vectores aleatorios el concepto es parecido pero tomando en cuenta que la convergencia es hacia la distribución conjunto de los elementos del vector. Asimismo, es intuitivo notar que si $x_n \xrightarrow{d} Z$, entonces $x_n = O_p(1)$.

Teorema 1.6. ***(Teorema del mapeo contínuo V -TCM V)*** *Dado $g : \mathbb{R}^k \to \mathbb{R}^m$ y $\{\mathbf{x}_n(.)\}$ tal que $\mathbf{x}_n \xrightarrow{d} \mathbf{Z}$. Si g es contínua en $\mathbb{R}^k$, luego*

$$g(\mathbf{x}_n) \xrightarrow{p} g(\mathbf{Z}) \tag{1.14}$$

1.5 Propiedades

1. **Relación entre convergencias**. La relación entre los distintos tipo de convergencia es resumida en la figura 1.

Figura 1: Relación entre convergencias

$$\begin{tikzcd}
\textbf{a.s.} \arrow[dr, Rightarrow] & & \\
& \mathbf{p} \arrow[r, Rightarrow] & \mathbf{d} \\
\textbf{s.m.} \arrow[ur, Rightarrow] & & \\
\textbf{r.m.} \arrow[u, Rightarrow] & &
\end{tikzcd}$$

Nota: Para $r \geq s > 0$.

2. **Regla del producto**. (i) Si $x_n = o_p(1)$ y $y_n = O_p(1)$, entonces $x_n y_n = o_p(1)$. Esta propiedad se puede extender para el caso matricial: (ii) si A_n es una matriz $n \times k$ donde cada elemento es una variable aleatoria de orden 1, $A_n = o_p(1)$, y b_n es un vector $k \times 1$ tal que $b_n = O_p(1)$, entonces $A_n b_n = o_p(1)$. Por último, (iii) si $A_n \xrightarrow{p} 0$ y $b_n \xrightarrow{d} Z$, entonces $A_n b_n \xrightarrow{p} 0$. (iv) lo mismo se cumple si se cambia p por $a.s.$

3. **Equivalencia asintótica**. Sean $\{x_n(.)\}$ y $\{y_n(.)\}$ dos secuencias de variables aleatorias tales que $x_n - y_n \xrightarrow{p} 0$ y $y_n \xrightarrow{d} Z$, entonces $x_n \xrightarrow{d} Z$.

4. **Teorema de Slutsky**. Se deriva de la regla del productos y equivalencia asintótica. Si $x_n \xrightarrow{d} x$ y $y_n \xrightarrow{p} y$, entonces: (i) $y_n x_n \xrightarrow{d} yx$, y (ii) $y_n + x_n \xrightarrow{d} y + x$.

5. Si $g : \mathbb{R}^k \to \mathbb{R}^m$ es contínua en $\mathbb{R}^k$, $x_n - b_n \xrightarrow{p} 0$, y $y_n \xrightarrow{d} Z$, entonces,

$$g(x_n) - g(y_n) \xrightarrow{p} 0 \tag{1.15}$$

$$g(x_n) \xrightarrow{d} g(Z) \tag{1.16}$$

6. **Inversa de una matriz**. No es muy complicado usar el TCM para el caso de la inversa de una matriz, ya que, los elementos de la inversa de una matriz es una función contínua de los elementos de la matriz inicial (bajo ciertas hipótesis comúnes en econometría). Sea A una matriz $k \times k$ y simétrica, entonces existe Q (ortogonal) tal que $A = Qdiag(\lambda_1, \ldots, \lambda_k)Q'$, luego $f(A) = Qf[diag(\lambda_1, \ldots, \lambda_k)]Q' = Qdiag[f(\lambda_1), \ldots, f(\lambda_k)]Q'$.

7. **Teorema del valor medio**. Si $g : \mathbb{R}^k \to \mathbb{R}^m$ es definido sobre un conjunto abierto y convexo $\Theta \subset \mathbb{R}^k$ tal que g es diferenciable en Θ. Luego, para todo j y $\theta \in \Theta$ existe $\overline{\theta}_j = \alpha_j\theta + (1 - \alpha_j)\theta_0$ para algún $\alpha_j \in [0, 1]$ tal que

$$g_j(\theta) = g_j(\theta_0) + \frac{\partial g_j(\theta)}{\theta'}\Big|_{\theta=\overline{\theta}_j}(\theta - \theta_0) \tag{1.17}$$

donde $g_j(.)$ es la función j-ésima del vector de funciones y

$$\frac{\partial g_j(\theta)}{\theta'}\Big|_{\theta=\overline{\theta}_j} = \begin{bmatrix} \frac{\partial g_j(\theta)}{\theta_1}\Big|_{\theta=\overline{\theta}_j} & \cdots & \frac{\partial g_j(\theta)}{\theta_k}\Big|_{\theta=\overline{\theta}_j} \end{bmatrix}$$

2 Ley de los grandes números

Basado en Lukacs (1975), Newey and McFadden (1994), (White, 2001, chapter 3), y clases dictadas. Estas leyes establecen que, bajo ciertos supuestos, algunas variables aleatorias poseen la propiedad de convergencia estocástica. Si la convergencia es casi segura suele llamarse "ley fuerte de los grandes números" mientras que si la convergencia es en probabilidad se les llama "ley débil de los grandes números". Los supuestos que se hacen son sobre los momentos de los elementos de la secuencia y sobre el tipo de muestra que se tiene en consideración. A lo largo de esta sección, usaré la notación $\overline{x}_n$ para refermirme a un promedio de muestra tamaño n, i.e., $\overline{x}_n = \sum_{i=1}^{n} \frac{x_i}{n}$. Antes de comenzar a exponer los distintos tipos de leyes sería útil introducir algunas proposiciones.

Proposicición 2.1. *Sea $g : \mathbb{R}^k \to \mathbb{R}^m$ una función contínua. Luego*

1. *Si $\mathbf{x}_i$ y $\mathbf{x}_j$ son idénticamente distribuidas, luego $g(\mathbf{x}_i)$ y $g(\mathbf{x}_j)$ son idénticamente distribuidas.*
2. *Si $\mathbf{x}_i$ y $\mathbf{x}_j$ son inpendientes, luego $g(\mathbf{x}_i)$ y $g(\mathbf{x}_j)$ son independientes.*

Hölder's inequality*. Sean x and y son variables alatorias y $p, q > 1$ tal que $1/p + 1/q = 1$. Si $\mathbb{E}|x|^p, \mathbb{E}|y|^q < \infty$, entonces*

$$\mathbb{E}|xy| \leq [\mathbb{E}|x|^p]^{1/p}[\mathbb{E}|x|^q]^{1/q} \tag{2.1}$$

De aquí se deriva la desigualdad de Cauchy-Shwartz.

2.1 Observaiones independiente e idénticamente distribuidas: muestra aleatoria

Ley (fuerte) de los grandes números de Kolmogórov. *Sea $\{\mathbf{x}_i(.)\}$ una secuencia de vectores aleatorios i.i.d tal que $\mathbb{E}(\mathbf{x}_i) = \mu$. Si $\mathbb{E}|\mathbf{x}_i| < \infty$, entonces*

$$\overline{\mathbf{x}}_n \xrightarrow{a.s.} \mu \tag{2.2}$$

La correspondiente ley débil (*ley de los grandes números de Khinchin*) se desprende de los mismos supuestos pero con la correspondiente convergencia en probabilidad, $\overline{\mathbf{x}}_t \xrightarrow{p} \mu$.

2.2 Observaciones independientes y heterogéneamente disbruidas

Ley (fuerte) de los grandes números de Markov. *Sea $\{\mathbf{x}_i(.)\}$ una secuencia de vectores independientes tal que $\mathbb{E}(\mathbf{x}_i) = \mu_i$. Si para algún $\delta > 0$, $\sum_{i=1}^{\infty} i^{-(1+\delta)}\mathbb{E}|\mathbf{x}_i - \mu_i|^{1+\delta} < \infty$, entonces*

$$\overline{\mathbf{x}}_n - \overline{\mu}_n \xrightarrow{a.s.} 0 \tag{2.3}$$

Para garantizar la condición de Markov (sí, aquella sumatoria de expectativas) es sufciente que $\mathbb{E}|\mathbf{x}_i|^{1+\delta} < \Delta < \infty$ para algún $\delta > 0$ y todo i. Kolmogórov tiene una versión para este caso pero cae dentro de la ley de Markov cuando $\delta = 1$.

Ley (débil) de los grandes números de Chebyshex. *Sea $\{\mathbf{x}_i(.)\}$ una secuencia de vectores aleatorios tal que $\mathbb{E}(\mathbf{x}_i) = \mu_i < \infty$, $var(\mathbf{x}_i) = Q_i$. Si para todo i, $\lim_{n\to\infty}(n\overline{Q}_n)^{-1}Q_i = 0$, entonces*

$$\overline{\mathbf{x}}_n - \overline{\mu}_n \xrightarrow{p} 0 \tag{2.4}$$

2.3 Observaciones dependientes e idénticamente distribuidas

La independencia de las observaciones es un supuesto poco realista para series de tiempo, por tanto, muestras sin esta propiedad merecen un tratamiento distinto a los revisados. En particular, veremos que la ley de los grandes números continúa valiendo cuando podedmos garantizar que la dependencia desaperece asintóticamente.

Definición 2.1. *Si $(\Omega, \mathcal{F})$ es un espacio medible, una función $g : \Omega \to \mathbb{R}$ es $\mathcal{F}-$medible si para todo $a \in \mathbb{R}$, $\{\omega : g(\omega) \leq a\} \in \mathcal{F}$. Asimismo, una transformación uno a uno $T : \Omega \to \Omega$ es medible si $T^{-1}(\mathcal{F}) \subset \mathcal{F}$.*

Por ejemplo, considere el siguiente evento medible $\omega = (\ldots, x_{t-1}, x_t, x_{t+1}, \ldots)$, luego $T\omega = (\ldots, x_t, x_{t+1}, x_{t+2}, \ldots)$ transforma ω cambiando sus coordenadas una posición atrás. Usando este operador y una función medible X (i.e., variable aleatoria) podemos definir $X_1(\omega) = X(\omega)$, $X_2(\omega) = X(T\omega)$, $X_3(\omega) = X(T^2\omega)$, etc; o de forma compacta $X_t(\omega) = X_{t-1}(T\omega)$ para $t \geq 2$.

Definición 2.2. *Sea $(\Omega, \mathcal{F}, \mathbb{P})$ un espacio de probabilidad, la transformación $T : \Omega \to \Omega$ es preservadora de medida si es medible y si $\mathbb{P}(T^{-1}F) = \mathbb{P}(F)$ para todo $F \in \mathcal{F}$.*

Luego, si la variable aleatoria definida en el párrafo anterior es inducidad por una transformación preservadora de medida se cumple $\mathbb{P}(X_1 \leq a) = \mathbb{P}[\omega : X(\omega) \leq a] = \mathbb{P}[\omega : X(T\omega) \leq a] = \mathbb{P}(X_2 \leq a)$, i.e., las observaciones serían idénticamente distribuídas. Adicionalmente, estas variables tienen una propiedad más fuerte, *estacionariedad.*

Definición 2.3. *Sea G_τ la función distribución acumulada de la secuencia $\{\mathbf{X}_\tau, \mathbf{X}_{\tau+1}, \ldots\}$. Luego, $\{\mathbf{X}_t\}$ es estacionaria si $G_\tau = G_{\tau+1}$ para todo $\tau \geq 1$.*

Note que la estacionariedad es una propiedad más fuerte que el supuesto de idénticamente distribuídas pero más débil que el supuesto de i.i.d.

Definición 2.4. *Sea $(\Omega, \mathcal{F}, \mathbb{P})$ un espacio de probabilidad, sea $\{X_t\}$ una secuencia estacionaria y T una transformación preservadora de medida, luego $\{X_t\}$ es **ergódica** si*

$$\lim_{n\to\infty} n^{-1}\sum_{t=1}^{n} \mathbb{P}(F \cap T^t G) = \mathbb{P}(F)\mathbb{P}(G) \tag{2.5}$$

para todos los eventos $F, G \in \mathcal{F}$.

Si observamos detenidamente la anterior definición nos daremos cuenta que hace referencia a una especie de *independencia asintótica promedio.*

Teorema 2.1. Teorema ergódico. *Sea $\{X_t\}$ una secuencia ergódica con $\mathbb{E}|X_t| < \infty$, luego*

$$\overline{x}_n \xrightarrow{a.s.} \mu = \mathbb{E}(X_t) \tag{2.6}$$

2.4 Observaciones dependientes y heterogeneammente distribuidas

Ahora presentaremos un concepto más fuerte que el de ergodicidad, *condiciones de mixturas.*

Definición 2.5. *Sean $\mathcal{G}$ y $\mathcal{H}$ $\sigma-$algebras, luego podemos definir*

$$\phi(\mathcal{G},\mathcal{H}) \equiv \sup_{\{G\in\mathcal{G}, H\in\mathcal{H}: \mathbb{P}(G)>0\}} |\mathbb{P}(H|G) - \mathbb{P}(H)|$$
$$\alpha(\mathcal{G},\mathcal{H}) \equiv \sup_{\{G\in\mathcal{G}, H\in\mathcal{H}\}} |\mathbb{P}(G\cap H) - \mathbb{P}(G)\mathbb{P}(H)|$$

Intuitivamente, ϕ y α mide el grado de dependencia entre los eventos de $\mathcal{G}$ y $\mathcal{H}$. Ahora si, $\mathcal{B}^n_{-\infty}$ representa toda la información contenida en el pasado de la secuencia $\{X_t\}$ hasta el momento n y $\mathcal{B}^\infty_{n+m}$ representa toda la información contenida en el futura de la secuencia $\{X_t\}$ desde el momento $n+m$ en adelante, luego podemos definir

Definición 2.6.

$$\phi(m) \equiv \sup_n \phi(\mathcal{B}^n_{-\infty}, \mathcal{B}^\infty_{n+m})$$
$$\alpha(m) \equiv \sup_n \alpha(\mathcal{B}^n_{-\infty}, \mathcal{B}^\infty_{n+m})$$

*Si $\{X_t\}$ es tal que $\lim_{m\to\infty}\phi(m) = 0$, la secuencia es llamada **ϕ-mixta**; en cuanto, si $\{X_t\}$ es tal que $\lim_{m\to\infty}\alpha(m) = 0$, la secuencia es llamada **α-mixta***

Debemos saber que $\lim_{m\to\infty}\phi(m) = 0$ implica $\lim_{m\to\infty}\alpha(m) = 0$, luego la siguiente proposición indica que las condiciones de mixturas imponen mayor restricciones que la ergodicidad.

Proposicición 2.2. *Sea $\{X_t\}$ una secuncia estacionaria. Si $\lim_{m\to\infty}\alpha(m) = 0$, luego $\{X_t\}$ es érgodica.*

Definición 2.7. *Sea $a \in \mathbb{R}$. (i) Si $\phi(m) = O(m^{-a-\epsilon})$ para algún $\epsilon > 0$, luego ϕ es de orden $-a$. (ii) Si $\alpha(m) = O(m^{-a-\epsilon})$ para algún $\epsilon > 0$, luego α es de orden $-a$.*

Teorema 2.2. Ley de grandes números de McLeish. *Sea $\{X_t\}$ una secuencia de escalares con $\mu_t = \mathbb{E}(X_t)$. Si $\sum_{t=1}^{\infty}[\mathbb{E}(|X_t - \mu_t|^{r+\delta})/t^{r+\delta}]^{1/r} < \infty$ para algún $\delta \in (0, r]$ donde $r \geq 1$, y ϕ es de orden $-r/(2r-1)$ o α es de orden $-r/(r-1)$ con $r > 1$; luego $\overline{x}_n - \overline{\mu}_n \xrightarrow{a.s.} 0$.*

Es evidente que, aunque relevante, el teorema de McLeish es muy díficil de verificar teóricamente y quizá imposible de hacerlo empíricamente. Por tanto, necesitamos un concepto que sea manejable tanto en teoría como en la práctica.

Definición 2.8. *$\{X_t\}$ tiene elementos asintóticamente incorrelacionados si existen $\{\rho_\tau, \tau \geq 0\}$ tal que $\rho_\tau \in [0,1]$, $\sum_{\tau\geq 0} \rho_\tau < \infty$ y $cov(X_t, X_{t+\tau}) \leq \rho_\tau [var(X_t) var(X_{t+\tau})]^{1/2}$ para todo $\tau > 0$, donde $var(X_t) < \infty$ para todo t.*

Teorema 2.3. *Sea $\{X_t\}$ una secuencia asintóticamente incorrelacionada con $\mu_t = \mathbb{E}(X_t)$ y $\sigma_t^2 = var(X_t) < \infty$. Luego, $\overline{x}_n - \overline{\mu}_n \xrightarrow{a.s.} 0$.*

Tanto secuencias estacionarias en covarianza como ergódicas suelen ser asintóticamente incorrelacionadas.

2.5 Procesos martingales en diferencias

Proposicición 2.3. *Sean $\mathcal{G}$ y $\mathcal{H}$ dos σ-álgebras tal que: i) $\mathcal{G} \subset \mathcal{Y}$, y ii) $\mathbb{E}(Y|\mathcal{H})$ es un función medible en $\mathcal{G}$. Luego*

$$\mathbb{E}(Y|\mathcal{H}) = \mathbb{E}(Y|\mathcal{G}) \tag{2.7}$$

Definición 2.9. *Sea $\{Y_t\}$ una secuencia de variables aleatorias y $\{\mathcal{F}_t\}$ una filtración (secuencias de σ-álgebras crecientes). Si Y_t es medible en $\mathcal{F}_t$, luego $\{Y_t, \mathcal{F}_t\}$ es llamado un proceso estocástico adaptado.*

Definición 2.10. *Sea $\{Y_t, \mathcal{F}_t\}$ proceso estocástico adaptado. Luego $\{Y_t, \mathcal{F}_t\}$ es un martiganle en diferencias si*

$$\mathbb{E}(Y_t|\mathcal{F}_{t-1}) = 0, \qquad t \geq 1 \tag{2.8}$$

Note que de todo proceso adaptado es posible generar un proceso martingale en diferencias, $\{Y_t - \mathbb{E}(Y_t|\mathcal{F}_{t-1}), \mathcal{F}_t\}$. Además, es importante mencionar que este proceso está muy relacionado al concepto de *expectativas racionales*.

Teorema 2.4. ***(Teorema de Chow)*** *Sea $\{Y_t, \mathcal{F}_t\}$ un proceso martingal en diferencias. Si para algún $r \geq 1$, $\sum_{t=1}^{\infty} \frac{\mathbb{E}|Y_t|^{2r}}{t^{1+r}} < \infty$, luego $\overline{Y}_n \xrightarrow{a.s.} 0$*

3 Teorema de limite central

Basado en Lukacs (1975), Newey and McFadden (1994), (White, 2001, chapter 5), y clases dictadas. Primero introduciré un lema que nos permite facilitar la presentación de los teoremas.

Estrategia de Cramer-Wold. *Si $\mathbf{x}_n \xrightarrow{d} \mathbf{x}$, esto es equivalente a $\lambda' \mathbf{x}_n \xrightarrow{d} \lambda' \mathbf{x}$ para todo $\lambda \in \mathbb{R}^k$ tal que $\lambda'\lambda = 1$.*

Por tanto, si establecemos proposiciones respecto a secuencias de variables aleatorias es sencillo extrapolar éstas proposiciones a secuencias de vectores aleatorios.

3.1 Independiente e idénticamente distribuitdas: muestra aleatoria

Teorema del límite central de Lindeberg-Lévy. *Sea $\{\mathbf{x}_i(.)\}$ una secuencia de vectores i.i.d. tal que $\mathbb{E}(|\mathbf{x}_i|^2) < \infty$, entonces*

$$\sqrt{n}(\overline{\mathbf{x}}_n - \mu) \xrightarrow{d} \mathcal{N}(0, V) \tag{3.1}$$

donde, $\mu = \mathbb{E}(\mathbf{x}_i)$ y $V = \mathbb{E}[(\mathbf{x_i} - \mu)(\mathbf{x}_i - \mu)']$. También se puede escribir

$$\sqrt{n} V^{-1/2}(\overline{\mathbf{x}}_n - \mu) \xrightarrow{d} \mathcal{N}(0, I_k) \tag{3.2}$$

3.2 Observaciones independientes y heterogéneamente distribuídas

Teorema 3.1. ***(Lindeberg-Feller)*** *Sea $\{X_t\}$ una secuencia aleatoria de variables independientes con $\mu_t = \mathbb{E}(X_t)$, $\sigma_t^2 = var(X_t) < \infty$ y distribuídas según F_t. Luego*

$$\sqrt{n}(\overline{x}_n - \overline{\mu}_n)/\overline{\sigma}_n \xrightarrow{d} \mathcal{N}(0,1) \tag{3.3}$$

y $\lim_{n\to\infty} \max_{1\le t\le n} n^{-1}(\sigma_t^2/\overline{\sigma}_n^2) = 0$ si y sólo si $\forall \epsilon > 0$

$$\lim_{n\to\infty} \overline{\sigma}_n^2 n^{-1} \sum_{t=1}^{n} \int_{(x-\mu_t)^2 > \epsilon n \overline{\sigma}_n^2} (z-\mu_t)^2 dF_t(z) = 0$$

Sin embargo, en general la condición de Lindeberg es muy difícil de verificar, por tanto, sería muy útil contar con alguna condición un poco más simple.

Teorema 3.2. ***(Lindeberg-Feller)*** *Sea $\{X_t\}$ una secuencia aleatoria de variables independientes con $\mu_t = \mathbb{E}(X_t)$, $\sigma_t^2 = var(X_t)$ y $\mathbb{E}|X_t - \mu_t|^{2+\delta} < \Delta < \infty$ para algún $\delta > 0$ y todo t. Si $\overline{\sigma}_n^2 > \delta' > 0$ para todo n suficientemente grande, luego*

$$\sqrt{n}(\overline{x}_n - \overline{\mu}_n)/\overline{\sigma}_n \xrightarrow{d} \mathcal{N}(0,1) \tag{3.4}$$

Ahora sólo basta verificar $\mathbb{E}|X_t|^{2+\delta} < \infty$.

3.3 Observaciones dependientes e identicamente distribuídas

Sea $\{X_t, \mathcal{F}_t\}$ un proceso estocástico adaptado, luego podemos expresar a X_t como una suma de procesos martingales en diferencias más algo más.

$$\begin{aligned}
X_t &= X_t - \mathbb{E}(X_t|\mathcal{F}_{t-1}) + \mathbb{E}(X_t|\mathcal{F}_{t-1}) \\
&= X_t - \mathbb{E}(X_t|\mathcal{F}_{t-1}) + \mathbb{E}(X_t|\mathcal{F}_{t-1}) - \mathbb{E}(X_t|\mathcal{F}_{t-2}) + \mathbb{E}(X_t|\mathcal{F}_{t-2}) \\
&\vdots \\
X_t &= \sum_{j=0}^{m-1} R_{tj} + \mathbb{E}(X_t|\mathcal{F}_{t-m}), \quad m = 1, 2, \ldots
\end{aligned}$$

donde $R_{tj} = \mathbb{E}(X_t|\mathcal{F}_{t-j}) - \mathbb{E}(X_t|\mathcal{F}_{t-j-1})$. La gran pregunta aquí es, ¿$\mathbb{E}(X_t|\mathcal{F}_{t-m}) \to 0$?

Definición 3.1. *Sea $\{X_t, \mathcal{F}_t\}$ un proceso estocástico adaptado con $\mathbb{E}(X_t)$ finito. Luego dicho proceso es un* ***mixingale*** *si existe dos secuentas nonegativas $\{c_t\}$ y $\{\gamma_m\}$ tal que $\lim_m \gamma_m = 0$ y*

$$\{\mathbb{E}[\mathbb{E}(X_t|\mathcal{F}_{t-m})]\}^{1/2} \le c_t \gamma_m, \quad m = 1, 2, \ldots \tag{3.5}$$

Luego,

Teorema 3.3. ***Scott*** *Sea $\{X_t, \mathcal{F}_t\}$ un proceso estocástico adpatado estacionario, ergódico, y mixingale con γ_m de orden -1. Luego $\overline{\sigma}_n^2 = var(n^{-1}\sum_{t=1}^n X_t) \to \overline{\sigma}^2 < \infty$ y si $\overline{\sigma}^2 > 0$,*

$$n^{-1/2}\overline{x}_n/\overline{\sigma} \xrightarrow{d} \mathcal{N}(0,1) \tag{3.6}$$

3.4 Método Delta

Si $\sqrt{n}(\mathbf{x}_n - \mu) \xrightarrow{d} \xi$ y $g : \mathbb{R}^k \to \mathbb{R}^m$ es contínuamente diferenciable en una vecindad de μ, entonces

$$\sqrt{n}[g(\mathbf{x}_n) - g(\mu)] \xrightarrow{d} G'\xi \tag{3.7}$$

donde $G = \frac{\partial}{\partial u'} g(u)\Big|_{u=\mu}$.

4 Estimadores extremos 0: Caso general

Basado en Newey and McFadden (1994) y clases dictadas. En general, un estimador es denominado extremo si

$$\hat{\theta}_n = arg \max_{\theta \in \Theta} \hat{Q}_n(\theta) \tag{4.1}$$

Observe que (i) las condiciones de primer orden pueden no identificar al verdadero parámetros θ_0 incluso cuando sólo exista un máximo global, y (ii) los resultados de consistencia aplican para el máximo global. A lo largo de esta sección θ_0 hará referencia al verdadero parámetro del proceso generador de datos.

4.1 Consistencia

Se dice que un estimador, $\hat{\theta}_n$, de θ_0 es consistente si $\hat{\theta}_n \xrightarrow{p} \theta_0$.

Teorema 4.1. *Si existe una función $Q_0(\theta)$ tal que (i) $Q_0(\theta)$ posee un único máximo en θ_0, (ii) Θ es compacto, (iii) $Q_0(\theta)$ es una función contínua, y (iv) $sup_{\theta \in \Theta} |\hat{Q}_n(\theta) - Q_0(\theta)| \xrightarrow{p} 0$ (convergencia uniforme en probabilidad); entonces*

$$\hat{\theta}_n \xrightarrow{p} \theta_0 \tag{4.2}$$

Proof. Para todo $\epsilon > 0$, con probabilidad acercándose a uno tenemos que

1. De la definición del estimador: $\hat{Q}_n(\hat{\theta}_n) \geq \hat{Q}_0(\theta_0) > \hat{Q}_0(\theta_0) - \epsilon/3$.
2. Del supuestos (iv): $\hat{Q}_n(\hat{\theta}_n) - Q_0(\hat{\theta}_n) \leq |\hat{Q}_n(\hat{\theta}_n) - Q_0(\hat{\theta}_n)| \leq sup_{\theta \in \Theta} |\hat{Q}_n(\theta) - Q_0(\theta)| < \epsilon/3$. Luego, $Q_0(\hat{\theta}_n) > \hat{Q}_n(\hat{\theta}_n) - \epsilon/3$.
3. Del supuesto (iv): $Q_0(\theta_0) - \hat{Q}_n(\theta_0) \leq |\hat{Q}_n(\theta_0) - Q_0(\theta_0)| < \epsilon/3$, luego $\hat{Q}_n(\theta_0) > Q_0(\theta_0) - \epsilon/3$.

Luego,

$$Q_0(\hat{\theta}_n) > \hat{Q}_n(\hat{\theta}_n) - \epsilon/3 > \hat{Q}_0(\theta_0) - 2\epsilon/3 > Q_0(\theta_0) - \epsilon$$

Ahora, sea $\mathcal{H}$ un conjunto abierto en Θ tal que $\theta_0 \in \mathcal{H}$, entonces $\Theta \cap \mathcal{H}^c$ es compacto, por tanto, para un $\theta^* \in \Theta \cap \mathcal{H}^c$, $sup_{\theta \in \Theta \cap \mathcal{H}^c} Q_0(\theta) = Q_0(\theta^*) < Q_0(\theta_0)$. Definiendo $\epsilon = Q_0(\theta_0) - Q_0(\theta^*)$, luego, con probabilidad acercándose a uno, $Q_0(\hat{\theta}_n) > Q_0(\theta^*) \Rightarrow \hat{\theta}_n \in \mathcal{H}$ o $\lim_{n \to} \mathbb{P}(\hat{\theta}_n \in \mathcal{H}) = 1$. Defina para cualquier $\delta > 0$, $\mathcal{H} \equiv \mathcal{H}_\delta = \{\theta \in \Theta : |\theta - \theta_0| < \delta\}$, entonces $\lim_{n \to} \mathbb{P}(|\hat{\theta}_n - \theta_0| < \delta) = 1$ para cualquier $\delta > 0$. □

Note que este teorema es muy general, incluye el caso en que θ es de dimensión infinita, e.g., el caso no paramétrico.

Definición 4.1. *Se dice que el parámetro poblacional, θ_0, está identificado si la distribución de la muestra dado θ_0 es diferente a la distribución de los datos dado $\theta \in \Theta$, $\theta \neq \theta_0$; o $f(z|\theta) \neq f(z|\theta_0)$.*

La identificación θ_0 es una hipótesis necesaria que exista un único maximo en $Q_0(\theta)$. Si añadimos otros supuestos generales se puede establecer identifcación del estimador de máxima verosimilitud y de método generalizados de momentos. Antes de pasar a estudiar la normalidad asintótica de estimadores extremos es útil introducir el siguiente lema.

Lema 4.1. Ley de los grandes números uniforme. *Si: (i) datos,$\{z_i\}_{i=1}^n$, provienen de una muestra aleatoria y Θ es compacto, (ii) $a(z_i, \theta)$ es una función contínua en Θ con probabilidad uno, y (iii) $\exists d(z)$ tal que $|a(z, \theta)| \leq d(z)$ para todo $\theta \in \Theta$ y $\mathbb{E}[d(z)] < \infty$*

$$\mathbb{E}(a(z, \theta)) \textit{ es contnua} \tag{4.3}$$

$$sup_{\theta \in \Theta} |n^{-1} \sum_{i=1}^{n} a(z_i, \theta) - \mathbb{E}[a(z, \theta)]| \xrightarrow{p} 0 \tag{4.4}$$

4.2 Normalidad asintótica

Teorema 4.2. Normalidad Asintótica I. *Si $\hat{\theta}_n \xrightarrow{p} \theta_0$, (i) $\theta_0 \in int(\Theta)$, (ii) $\hat{Q}_n(\theta) \in \mathcal{C}^2(\mathcal{H})$ donde $\mathcal{H}$ es una vencidad convexa de θ_0, (iii) $\sqrt{n}\nabla_\theta \hat{Q}_n(\theta_0) \xrightarrow{d} \mathcal{N}(0,\Sigma)$, (iv) $sup_{\theta\in\Theta}|\nabla_{\theta\theta}\hat{Q}_n(\theta) - H(\theta)| \xrightarrow{p} 0$ donde $H(\theta)$ es una función contínua en θ_0, (v) $H \equiv H(\theta_0)$ es no singular; entonces*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, H^{-1}\Sigma H^{-1}) \tag{4.5}$$

Proof. Definamos $\hat{\mathbb{1}}_n = \mathbb{I}(\hat{\theta}_n \in \mathcal{H})$, luego es evidente que $\hat{\mathbb{1}}_n \xrightarrow{p} 1$. Por la definición del estimador, se debe cumplir $\hat{\mathbb{1}}_n \nabla_\theta \hat{Q}_n(\hat{\theta}_n) = 0$, y usando el teorema de valor medio tenemos que

$$0 = \hat{\mathbb{1}}_n \nabla_\theta \hat{Q}_n(\theta_0) + \hat{\mathbb{1}}_n \overline{H}(\hat{\theta}_n - \theta_0)$$

donde $\overline{H}$ es una matriz cuya fila $j$ésima es igual a la $j$ésima fila de $\nabla_{\theta\theta}\hat{Q}_n(\theta)\Big|_{\theta=\overline{\theta}_j}$ y $\overline{\theta}_j = \alpha_j\hat{\theta}_n + (1-\alpha_j)\theta_0$ con $\alpha_j \in [0,1]$. Note que (iv) implica que $\overline{H} \xrightarrow{p} H$. Ahora, definamos $\overline{\mathbb{1}}_n = \mathbb{I}(\hat{\theta}_n \in \mathcal{H} \wedge \overline{H}\ \textit{no singular})$, por tanto,

$$\begin{aligned} 0 &= -\overline{\mathbb{1}}_n \overline{H}^{-1} \nabla_\theta \hat{Q}_n(\theta_0) - \overline{\mathbb{1}}_n(\hat{\theta}_n - \theta_0) \\ \sqrt{n}(\hat{\theta}_n - \theta_0) &= -\overline{\mathbb{1}}_n \overline{H}^{-1} \sqrt{n}\nabla_\theta \hat{Q}_n(\theta_0) + \underbrace{\sqrt{n}(1-\overline{\mathbb{1}}_n)(\hat{\theta}_n - \theta_0)}_{\xrightarrow{p} 0} \end{aligned}$$

por el teorema de Slutsky

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, H^{-1}\Sigma H^{-1})$$

□

Note que si $\hat{Q}_n(\theta) = n^{-1}\sum_{i=1}^n q(z_i,\theta)$, $q(z_i,\theta)$ fuese contínua en Θ, y $\mathbb{E}[sup_{\theta\in\Theta}|\nabla_{\theta\theta}q(z_i,\theta)|] < \infty$; por la ley de los grandes número uniforme se tiene que $H(\theta) = \mathbb{E}[\nabla_{\theta\theta}q(z_i,\theta)]$ es contínua en θ y $\sup_{\theta\in\Theta}|\hat{Q}_n(\theta) - H(\theta)|$, i.e., $H = H(\theta_0) = \mathbb{E}[\nabla_{\theta\theta}q(z_i,\theta)]$. El supuesto de doble diferenciabilidad contínua puede ser relajada en algunos casos.

Teorema 4.3. Normalidad Asintótica II. *Sea $\hat{Q}_n(\theta) = -\hat{g}_n(\theta)'\hat{W}\hat{g}_n(\theta)$, donde $\hat{W} \xrightarrow{p} W$, W es positiva semidefinida y $\hat{W}$ es simétrica, $\hat{\theta}_n \xrightarrow{p} \theta_0$, (i) $\theta_0 \in int(\Theta)$, (ii) $\hat{g}_n(\theta) \in \mathcal{C}^1(\mathcal{H})$ donde $\mathcal{H}$ es una vecindad convexa de θ_0, (iii) $\sqrt{n}\hat{g}_n(\theta_0) \xrightarrow{d} \mathcal{N}(0,\Omega)$, (iv) $sup_{\theta\in\mathcal{H}}|\frac{\partial}{\partial\theta'}\hat{g}_n(\theta) - G(\theta)| \xrightarrow{p} 0$ y $G(\theta)$ es contínua en θ_0, y (v) $G'WG$ es no singular donde $G = G(\theta_0)$; entonces*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, (G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}) \tag{4.6}$$

Proof. Definamos $\hat{\mathbb{1}}_n = \mathbb{I}(\hat{\theta}_n \in \mathcal{H})$, luego es evidente que $\hat{\mathbb{1}}_n \xrightarrow{p} 1$. Por la definición del estimador, se debe cumplir $\hat{\mathbb{1}}_n\hat{G}(\hat{\theta}_n)'\hat{W}\hat{g}_n(\hat{\theta}_n) = 0$ donde $\hat{G}(\theta) = \frac{\partial}{\partial\theta'}\hat{g}_n(\theta)$. Aplicando el teorema de valor medio a $\hat{g}_n(\hat{\theta}_n)$

$$\hat{\mathbb{1}}_n\hat{g}_n(\hat{\theta}_n) = \hat{\mathbb{1}}_n\hat{g}_n(\theta_0) + \hat{\mathbb{1}}_n\overline{G}(\hat{\theta}_n - \theta_0)$$

donde $\overline{G}$ es una matriz cuya fila $j$ésima es igual a la $j$ésima fila de $\hat{G}(\overline{\theta}_j)$ y $\overline{\theta}_j = \alpha_j\hat{\theta}_n + (1-\alpha_j)\theta_0$ con $\alpha_j \in [0,1]$. Note que (iv) implica $\overline{G} \xrightarrow{p} G$ y $\hat{G}(\hat{\theta}_n) \xrightarrow{p} G$. Ahora, definamos $\overline{\mathbb{1}}_n = \mathbb{I}(\hat{\theta}_n \in \mathcal{H} \wedge \hat{G}(\hat{\theta}_n)'\hat{W}\overline{G}\ \textit{no singular})$, por tanto,

$$\overline{\mathbb{1}}_n\hat{G}(\hat{\theta}_n)'\hat{W}\hat{g}_n(\hat{\theta}_n) = \overline{\mathbb{1}}_n\hat{G}(\hat{\theta}_n)'\hat{W}\hat{g}_n(\theta_0) + \overline{\mathbb{1}}_n\hat{G}(\hat{\theta}_n)'\hat{W}\overline{G}(\hat{\theta}_n - \theta_0)$$

$$
\begin{aligned}
0 &= \bar{1}_n \hat{G}(\hat{\theta}_n)'\hat{W}\hat{g}_n(\theta_0) + \bar{1}_n \hat{G}(\hat{\theta}_n)'\hat{W}\overline{G}(\hat{\theta}_n - \theta_0) \\
0 &= -\bar{1}_n[\hat{G}(\hat{\theta}_n)'\hat{W}\overline{G}]^{-1}\hat{G}(\hat{\theta}_n)'\hat{W}\hat{g}_n(\theta_0) - \bar{1}_n(\hat{\theta}_n - \theta_0) \\
\sqrt{n}(\hat{\theta}_n - \theta_0) &= -\bar{1}_n[\hat{G}(\hat{\theta}_n)'\hat{W}\overline{G}]^{-1}\hat{G}(\hat{\theta}_n)'\hat{W}\sqrt{n}\hat{g}_n(\theta_0) + \underbrace{\sqrt{n}(1-\bar{1}_n)(\hat{\theta}_n - \theta_0)}_{\xrightarrow{p} 0}
\end{aligned}
$$

por el teorema de Slutsky

$$
\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, (G'WG)^{-1}G'W\Omega WG(G'WG)^{-1})
$$

□

5 Estimadores extremos 1: Máxima verosimilitud

Basado en Newey and McFadden (1994), Greene (2003), y clases dictadas.

5.1 Introducción

Ahora trataremos el primero caso de extremadores extremos. Considere $\{z_i\}_{i=1}^n$ es una muestra aleatoria con densidad $f(z|\theta_0)$. La función de verosimilitad es

$$
L(\theta) \equiv L(\theta|z_1, \ldots, z_n) = \prod_{i=1}^{n} f(z_1|\theta) \tag{5.1}
$$

Note que esta función no debe ser interpretada como una función densidad de los parámetros ya que ni si quiera sabes que tiene masa 1. El logaritmo de esta función suele denominarse función log-verosímil $\ln L(\theta) = \sum_{i=1}^n \ln f(z_i|\theta)$. Luego el estimador de verosimilitud es un estimador extremo ya que es aquel que maximiza la función log-verosímil o, equivalente, $\hat{Q}_n(\theta) = n^{-1}\sum_{i=1}^n \ln f(z_i|\theta)$, es decir,

$$
\hat{\theta}_{ML} = arg \max_{\theta \in \Theta} \hat{Q}_n(\theta) \tag{5.2}
$$

5.2 Consistencia

Identificación de MV - Desigualdad Informativa. *Si θ_0 es identificable y $\mathbb{E}|\ln f(z|\theta)| < \infty$ para todo $\theta \in \Theta$, entonces $Q_0(\theta) = \mathbb{E}[\ln f(z|\theta)]$ posee un único máximo en θ_0.*

Teorema 5.1. *Suponga que $\{z_i\}_{i=1}^n$ es una muestra aleatoria con $f(z_i|\theta_0)$, (i) θ_0 es identificable, (ii) $\theta_0 \in \Theta$ el cuál es compacto, (iii) $\ln f(z_i|\theta)$ es contínua en todo $\theta \in \Theta$ con probabildiad 1, y (iv) $\mathbb{E}[sup_{\theta\in\Theta}|\ln f(z|\theta)|] < \infty$. Entonces $\hat{\theta}_{ML} \xrightarrow{p} \theta_0$.*

Proof. Verifiquemos las condiciones para la consistencia de estimadores extremos. La condición de identificación viene dada por (i), (iv) y la desigualdad informativa. La condición de compacidad por (ii). La condición de continuidad y de convergencia uniforme se deriva del supuesto de muestra aleatoria, (iii), (iv) $[d(z) = sup_{\theta\in\Theta}|\ln f(z|\theta)|]$, y la ley uniforme de los grandes números. □

5.3 Normalidad asintótica

Asumiendo diferenciabilidad (las necesarias), definamos como *score* a $s_n(\theta) \equiv \nabla_\theta \ln L(\theta) = \sum_{i=1}^n s(z_i, \theta)$ donde $s(z_i,\theta) \equiv \nabla_\theta \ln f(z_i|\theta)$, por otro lado $H_n(\theta) \equiv \nabla_{\theta\theta} \ln L(\theta) = \sum_{i=1}^n H(z_i,\theta)$ donde $H(z_i,\theta) \equiv \nabla_{\theta\theta} \ln f(z_i,\theta)$.

Teorema 5.2. *Sea $\{z_i\}_{i=1}^n$ una muestra aleatoria tal que los supuestos para la consistencia se cumplen, además (i) $\theta_0 \in int(\Theta)$, (ii) $f(z|\theta) \in \mathcal{C}^2(\mathcal{H})$ y $f(z|\theta)$ en una vencidad $\mathcal{H}$ de θ_0, (iii) $\int sup_{\theta\in\mathcal{H}}|\nabla_\theta f(z|\theta)|dz < \infty$, $\int sup_{\theta\in\mathcal{H}}|\nabla_{\theta\theta} f(z|\theta)|dz < \infty$, (iv) $J(\theta_0) = \mathbb{E}[s(z,\theta_0)s(z,\theta_0)']$ existe y es no singular, y (v) $\mathbb{E}[sup_{\theta\in\mathcal{H}}|H(z,\theta)|] < \infty$. Entonces $\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \xrightarrow{d} \mathcal{N}(0, J(\theta_0)^{-1})$.*

Note que como la muestra es aleatoria $H(\theta) = \mathbb{E}[H(z_i, \theta)]$ es el mismo para todo i, luego $\mathbb{E}[H_n(\theta)] = nH(\theta)$, asimismo, $\mathcal{I}(\theta) = var[s_n(\theta)]$. $\mathcal{I}(\theta_0)$ es denominado la *matriz de información de Fisher*. Las siguientes proposiciones serán importantes para caracterizar la eficiencia asintótica de $\hat{\theta}_{ML}$.

Proposicición 5.1. $\mathbb{E}[s(z, \theta_0)] = 0$ *y* $\mathbb{E}[s_n(\theta_0)] = 0$.

Proposicición 5.2. Matriz de información de Fisher*.* $\mathcal{I}(\theta_0) = var[s_n(\theta_0)] = \mathbb{E}[s_n(\theta_0)s_n(\theta_0)'] = nJ(\theta_0) = -\mathbb{E}[\nabla_{\theta\theta} \ln L(\theta)|_{\theta=\theta_0}] = -\mathbb{E}[H_n(\theta_0)] = -nH(\theta_0)$. *De aquí,* $J(\theta_0) = -H(\theta_0)$

Proposicición 5.3. Cota inferior de Cramer-Rao*. Si* $f(z|\theta)$ *es doble continua diferenciable, su soporte no depende de* θ*, y* $\hat{\theta}$ *es un estimador insesgado de* θ_0 *entonces*

$$var(\hat{\theta}) \geq \mathcal{I}(\theta_0)^{-1} \tag{5.3}$$

Del teorema de normalidad asintótica se puede mostrar que

$$\hat{\theta}_{ML} \overset{a}{\sim} \mathcal{N}(0, [nJ(\theta_0)]^{-1}) = \mathcal{N}(0, \mathcal{I}(\theta_0)^{-1}) \tag{5.4}$$

Algunas veces, se suelen plantear las siguientes condiciones de regularidad

Condiciones de regularidad*.*

R0 $\{z_i\}_{i=1}^n$ *es una muestra aleatoria,* $\theta_0 \in \Theta$ *el cuál es compacto y* θ_0 *es identificable.*

R1 *Las tres primera derivadas de* $\ln f(z_i|\theta)$ *con respecto a* θ *son contínuas y finitas para casi todo* z_i *y todo* $\theta \in \Theta$*.*

R2 *El soporte de* $f(z|\theta)$ *no depende de* θ *y las condiciones necesarias para la existencia de la expectativa de la primera y segunda derivada de* $\ln f(z_i|\theta)$ *se cumplen.*

R3 $\mathbb{E}[H(z_i, \theta_0)]$ *es invertible y para todo* $\theta \in \Theta$

$$\mathbb{E}\Big|\frac{\partial^3 \ln f(z|\theta)}{\partial\theta_j \partial\theta_k \partial\theta_l}\Big| < \infty$$

Estas condiciones estan muy relacionadas a los supuestos que garantizan la convergencia y normalidad asintótica (efeciencia) del estimador máxima verosimilitu. a) Para consistencia bastaría con **R0** y **R1**: por el supuestos de diferenciabilidad $\ln f(z|\theta)$ es contínua en todo $\theta \in \Theta$, compacidad más continuidad $|\ln f(z|\theta)| < B < \infty$ para todo $\theta \in \Theta$ entonces defina $d(z) = sup_{\theta\in\Theta}|\ln f(z|\theta)|$ luego $\mathbb{E}[d(z)] < \infty$, y finalmente aplicar el teorema de consistencia de estimadores extremos.

5.4 Estimadores de la matriz de información

Podemos construir estimadores consistentes de $J(\theta_0)$ y de ahí obetener una versión para la matriz de información de Fiser.

1. Si se puede obtener $J(\theta)$, entonces estimar mediante $J(\hat{\theta}_{ML})$. Luego el estimador de la matriz de información de Fiher es $\mathcal{I}(\hat{\theta}_{ML}) = nJ(\hat{\theta}_{ML})$. La consistencia de $J(\hat{\theta}_{ML})$ es a causa de la consistencia de $\hat{\theta}_{ML}$ y del TMC III.

2. Si fuese muy complicado obtener la esperanza, aún puede construirse un estimador consiste: $\hat{J}(\hat{\theta}_{ML}) = n^{-1}\sum_{i=1}^n H(z_i, \hat{\theta}_{ML})$. Luego el estimador para la matriz de Fiher es $\hat{\mathcal{I}}(\hat{\theta}_{ML}) = n\hat{J}(\hat{\theta}_{ML})$. La consistencia de $\hat{J}(\hat{\theta}_{ML})$ provienve de la consistencia de $\hat{\theta}_{ML}$ y de la ley de los grandes números.

3. Si fuese difícil obtener la matriz Hessiana y la esperanza, entonces podemos utilizar el estimador BHHH: $\hat{\hat{J}}(\hat{\theta}_{ML}) = n^{-1}\sum_{i=1}^n s(z_i, \hat{\theta}_{ML})s(z_i, \hat{\theta}_{ML})'$. El estimador BHHH de $\mathbb{I}(\theta_0)$ es $\hat{\hat{\mathcal{I}}}(\hat{\theta}_{ML}) = n\hat{\hat{J}}(\hat{\theta}_{ML})$. La consistencia de $\hat{\hat{J}}(\hat{\theta}_{ML})$ provienve de la consistencia de $\hat{\theta}_{ML}$ y de la ley de los grandes números.

6 Estimadores extremos 2: GMM

Basado en Newey and McFadden (1994).

6.1 Introducción: Método de momentos

Considere la muestra $\{z_i\}_{i=1}^n$ de la función de densidad $f(z|\theta)$. Sabemos que los k primeros momentos depende de θ

$$\mathbb{E}Z^j = \mu_j(\theta_0), \quad j = 1, \ldots, k \tag{6.1}$$

Definiendo el momento muestral como $m_j = n^{-1}\sum_{i=1}^n z_i^j$ para $j = 1, \ldots, k$. El estimador de método de momentos $\hat{\theta}_{MM} = T(m_1, \ldots, m_k)$ es aquel vector en $\mathbb{R}^k$ que resuelve

$$\begin{aligned} m_1 &= \mu_1(\hat{\theta}_{MM}) \\ &\vdots \\ m_k &= \mu_k(\hat{\theta}_{MM}) \end{aligned}$$

Note que podemos definir $g : \mathbb{R}^k \to \mathbb{R}^k$ con $g_j(z_i, \theta) = x_i^j - \mu_j(\theta)$ y usar la condición qué $\mathbb{E}[g(z_i, \theta_0)] = 0$. En este caso el estimador de momentos es aquel que, bajo algunos supuestos de que este sistema tenga solución, iguale $\hat{g}_n(\theta) \equiv n^{-1}\sum_{i=1}^n g(z_i, \theta)$ a cero, i.e., $\hat{g}_n(\hat{\theta}_{MM}) = 0$.

6.2 Consistencia

Identificación GMM. *Si W es positiva semi-definida, $g(\theta) = \mathbb{E}[g(z, \theta)]$ con $g(\theta_0) = 0$, y $Wg(\theta) \neq 0$ para todo $\theta \neq \theta_0$; entonces $Q_0(\theta) = -g(\theta)'Wg(\theta)$ posee un único máximo en θ_0.*

Proof. Sea R tal que $R'R = W$, luego $R'Rg(\theta) \neq 0$ implica $Rg(\theta_0) \neq 0$, luego $Q_0(\theta) = -(Rg(\theta)'(Rg(\theta)) < Q_0(\theta_0) = 0$ para todo $\theta \neq \theta_0$. □

Observaciones

- Una condición necesaria para la identificación es que $l \geq k$ de lo contrario $g(\theta) = 0$ no tendría solución.
- Note que si W es no singular (o positiva definida) la identifación se lograría con $g(\theta) \neq 0$ para todo $\theta \neq \theta_0$.
- Si la función de momento fuese linear, $g(z, \theta) = g(z) + G(z)\theta$, luego la condición de necesaria e suficiente para la identifcación es $l \geq k$ y $W\mathbb{E}[G(z)]$ posea rango k. Además, si W es no singular, la condición de rango sólo debe aplicarse a $\mathbb{E}[G(z)]$.

Teorema 6.1. *Sea $\{z_i\}_{i=1}^n$ una muestra aleatoria, $\hat{W} \xrightarrow{p} W$, (i) W positiva semi-definida con θ_0 identificable, (ii) $\theta \in \Theta$ el cuál es compacto, (iii) $g(z, \theta)$ es contínua en cada $\theta \in \Theta$ con probabilidad uno, y (iv) $\mathbb{E}[sup_{\theta\in\Theta}|g(z, \theta)|] < \infty$. Entonces $\hat{\theta}_{GMM} \xrightarrow{p} \theta_0$.*

Proof. Verifiquemos las condiciones para la consistencia de estimadores extremos. La condición de identificación se desprende (i) y la condición de compacidad de (ii). Defina $d(z) = sup_{\theta\in\Theta}|g(z, \theta)|$, por la ley uniforme de los grandes números $g(\theta) = \mathbb{E}[g(z, \theta)]$ es contínua y $sup_{\theta\in\Theta}\theta \in \Theta|\hat{g}_n(\theta) - g(\theta)| \xrightarrow{p} 0$, entonces $Q_0(\theta) = -g(\theta)'Wg(\theta)$ es contínua; esto es, la condición de continuidad se deriva de (iii) y (iv). Considere la siguiente desigualdad

$$\begin{aligned} |\hat{Q}_n(\theta) - Q_0(\theta)| &= |\hat{g}_n(\theta)'\hat{W}\hat{g}_n(\theta) - g(\theta)'Wg(\theta) \pm g(\theta)'\hat{W}g(\theta) \pm \hat{g}_n(\theta)'\hat{W}g(\theta)| \\ &= |\hat{g}_n(\theta)'\hat{W}[\hat{g}_n(\theta) - g(\theta)] + [\hat{g}_n(\theta) - g(\theta)]'\hat{W}g(\theta) \\ &\qquad + g(\theta)'[\hat{W} - W]g(\theta) \pm g(\theta)'\hat{W}[\hat{g}_n(\theta) - g(\theta)]| \end{aligned}$$

$$
\begin{aligned}
&= |[\hat{g}_n(\theta) - g(\theta)]'\hat{W}[\hat{g}_n(\theta) - g(\theta)] + \\
&\qquad g(\theta)'[\hat{W} + \hat{W}'][\hat{g}_n(\theta) - g(\theta)] + g(\theta)'[\hat{W} - W]g(\theta)| \\
&\leq |\hat{W}||\hat{g}_n(\theta) - g(\theta)|^2 + 2|\hat{W}||\hat{g}_n(\theta) - g(\theta)||g(\theta)| + |\hat{W} - W||g(\theta)|^2
\end{aligned}
$$

entonces,

$$
\begin{aligned}
sup_{\theta\in\Theta}|\hat{Q}_n(\theta) - Q_0(\theta)| \leq |\hat{W}|sup_{\theta\in\Theta}|\hat{g}_n(\theta) - g(\theta)|^2 + 2|\hat{W}|sup_{\theta\in\Theta}|\hat{g}_n(\theta) - g(\theta)||g(\theta)| + \\
+ |\hat{W} - W|sup_{\theta\in\Theta}|g(\theta)|^2
\end{aligned}
$$

luego $sup_{\theta\in\Theta}|\hat{Q}_n(\theta) - Q_0(\theta)| \xrightarrow{p} 0$. Por tanto, la condición de convergencia uniforme y continuidad se derivan de (iii) y (iv). □

7 Prueba de Hipótesis

Basado en Casella and Berger (2001). En este contexto, una hipótesis es una proposición sobre un parámetro poblacional. Después de observar la muestra se debe decidir si rechazar la hipótesis o no: para algunas muestras se rechará y para otras no. Sin embargo, no es necesario expresar el procedimiento en términos de toda la muestra, para esto es útil el uso de estadísticos suficientes $T(X)$ (*data reduction*): si sólo observamos $T(X)$ y no X, luego dos muestras X e Y tendrán la misma información sobre θ si $T(X) = T(Y)$.

Definición 7.1. Principio de estadísitco suficiente. *Si $T(X)$ es un estadístico suficiente para θ, luego cualquier inferencia sobre θ debe depender de X sólo por medio del valor $T(X)$.*

Teorema 7.1. Teorema de factorización. *$T(X)$ es un estadístico suficiente para θ si y solo si existe $g(t|\theta)$ y $h(x)$ tal que*

$$f(X|\theta) = g(T(X)|\theta)h(X) \tag{7.1}$$

Definición 7.2. Estadístico mínimo suficiente. *$T(X)$ es un estadístico mínimo suficiente si para todo otros estadístico suficiente $\tilde{T}(X)$, $T(X)$ es una función de $\tilde{T}(X)$.*

Definición 7.3. Monotone Likelihood Ratio (MLR). *Una familia $\{g(t|\theta) : \theta \in \Theta\}$ posee MLR si para todo $\theta_2 > \theta_1$, $\frac{g(t|\theta_2)}{g(t|\theta_1)}$ es monotóna en relación a t en $\{t : g(t|\theta_1) > 0 \vee g(t|\theta_2) > 0\}$.*

Para tomar la decisión de rechazar una hipótesis o no es común el uso de una región de rechazo R. De esta forma un test suele tener los siguientes elementos: (i) hipóteis nula, $\mathbb{H}_0 : \theta_0 \in \Theta_0$; (ii) hipótesis alternativa, $\mathbb{H}_1 : \theta_0 \in \Theta_0^c$; y (iii) Región de rechazo R, si $X \in R$ entonces rechazar $\mathbb{H}_0$. Dentro de las propiedades de un test tenemos

- *Función poder*: mide la probabilidad de rechazar la hipótesis nula en función de los posibles valores del parámetros pobacinal, $\beta(\theta_0) = \mathbb{P}_{\theta_0}(X \in R)$.
- *Tamaño de un test*: Un test tiene tamaño α si $sup_{\theta\in\Theta_0}\beta(\theta) = \alpha$.
- *Nivel de un test*: Un test es de nivel α si $sup_{\theta\in\Theta_0}\beta(\theta) \leq \alpha$.
- *Error tipo I*: Rechazar la hipótesis nula cuando $\theta_0 \in \Theta_0$, $\mathbb{P}[Error\ I] = \beta(\theta_0)$ si $\theta_0 \in \Theta_0$.
- *Error tipo I*: No rechazar la hipótesis nula cuando $\theta_0 \in \Theta_0^c$, $\mathbb{P}[Error\ II] = \beta(\theta_0)$ si $\theta_0 \in \Theta_0^c$.
- *Poder de un test*: $1 - \mathbb{P}[Error\ II]$.

Definición 7.4. *Un test con función poder $\beta(\theta_0)$ es **insesgado** si $\beta(\theta') \geq \beta(\theta'')$ para todo $\theta' \in \Theta_0^c$ y $\theta'' \in \Theta_0$.*

Definición 7.5. *Si $\mathcal{C}$ es la clase de test con $\mathbb{H}_0 : \theta_0 \in \Theta_0$ versus $\mathbb{H}_1 : \theta_0 \in \Theta_0^c$. Un test en $\mathcal{C}$, es **uniformemente más poderoso (UMP)** de la clase $\mathcal{C}$ si $\beta(\theta_0) \geq \tilde{\beta}(\theta_0)$ para todo $\theta \in \Theta_0^c$, donde $\tilde{\beta}(\theta_0)$ es la función de cualquier otro test en $\mathcal{C}$.*

Teorema 7.2. Teorema de Karlin-Rubin. *Considere la clase de test con $\mathbb{H}_0 : \theta_0 \leq \tilde{\theta}_0$ versus $\mathbb{H}_1 : \theta_0 > \tilde{\theta}_0$. Sea $T(.)$ un estadístico suficiente para θ_0 y $\{g(t|\theta) : \theta \in \Theta\}$ posee MLR. Entonces para cualquier t_0, el test con región de dechazo $R = \{X|T(X) > t_0\}$ es UMP con nivel α, donde $\alpha = \mathbb{P}_{\theta_0}[T(X) > t_0]$*

Si consideramos la clase de test con $\mathbb{H}_0 : \theta_0 \geq \tilde{\theta}_0$ versus $\mathbb{H}_1 : \theta_0 < \tilde{\theta}_0$, luego sólo debemos cambiar la región de rechazo por $R = \{X|T(X) < t_0\}$ y $\alpha = \mathbb{P}_{\theta_0}[T(X) < t_0]$.

Para muestras grandes se introducen generalmente dos nuevos conceptos respecto a pruebas de hipótesis.

Definición 7.6. *Un test asintótico con región de rechazo R_n, función poder $\beta_n(\theta_0)$, posee un nivel asintótico igual a α si*

$$\sup_{\theta \in \Theta_0} \lim_{n \to \infty} \beta_n(\theta) = \alpha \tag{7.2}$$

Definición 7.7. *Un test asintótico con región de rechazo R_n, función poder $\beta_n(\theta_0)$ es **consistente** si*

$$\lim_{n \to \infty} \beta_n(\theta) = 1 \tag{7.3}$$

para todo $\theta \in \Theta_0^c$

8 Test basados en MLE

Basado en Gourieroux and Monfort (1995), Greene (2003), y Newey and McFadden (1994).

Supuesto 8.1. *$\mathbb{H}_0 : \theta_0 \in \Theta_0$ donde $\Theta_0 = \{\theta \in \Theta | r(\theta) = 0\}$ y $r : \Theta \subset \mathbb{R}^k \to \mathbb{R}^q$ es contínuamente diferenciable con $\Gamma(\theta) = \frac{\partial}{\partial \theta'} r(\theta)$ posee rango q para todo $\theta \in \Theta$.*

8.1 Test de Wald

Observación: Aunque se estudiará el estadístico W_n en el contexto de esteimadore ML, este test tiene aplicación fuera de estos.

Teorema 8.1. *Si las condiciones de regularidad y el supesto 8.1 se cumplen, luego el test basado en $R_n = \{W_n \geq \chi^2_{q,1-\alpha}\}$ donde*

$$W_n = nr(\hat{\theta}_{ML})' \Big[\Gamma(\hat{\theta}_{ML}) \tilde{J}(\hat{\theta}_{ML})^{-1} \Gamma(\hat{\theta}_{ML})'\Big]^{-1} r(\hat{\theta}_{ML}) \xrightarrow{d} \chi^2_q \tag{8.1}$$

es consistente con nivel asintótico igual a α. $\tilde{J}(\hat{\theta}_{ML})$ es un estimador consistente $J(\theta_0)$.

Proof. Usando la normalidad asintótica de $\hat{\theta}_{ML}$ y el método delta, bajo la hipótesis nula

$$\sqrt{n} r(\hat{\theta}_{ML}) \xrightarrow{d} \mathcal{N}(0, \Gamma(\theta_0) J(\theta_0)^{-1} \Gamma(\theta_0))$$

$$\Rightarrow \sqrt{n} \Big[\Gamma(\theta_0) J(\theta_0)^{-1} \Gamma(\theta_0)\Big]^{-1/2} r(\hat{\theta}_{ML}) \xrightarrow{d} \mathcal{N}(0, I_q)$$

$$\Rightarrow nr(\hat{\theta}_{ML})' \Big[\Gamma(\theta_0) J(\theta_0)^{-1} \Gamma(\theta_0)'\Big]^{-1} r(\hat{\theta}_{ML}) \xrightarrow{d} \chi^2_q$$

Como $\Gamma(\theta)$ es contínua y $\tilde{J}(\theta_0)$ es un estimador consistente de $J(\theta_0)$ luego

$$W_n \xrightarrow{d} \chi^2_q$$

Note que si $\theta_0 \notin \Theta_0$ luego $g(\hat{\theta}_{ML}) \xrightarrow{p} g(\theta_0) \neq 0$ entonces W_n converge a infinito y $\lim_{n\to\infty} \mathbb{P}_\theta(R_n) = 1$. □

Note que el estadístico W_n es asintóticamente *pivotal*, es decir, su distrubción asintótica no depende de parámetros desconocidos.

8.2 Test de multiplicadores de Lagrange o test *score*

El estimador de máxima verosimilitud restringido es definido por

$$\hat{\theta}^0_{ML} = arg \max_{\theta\in\Theta_0} \hat{Q}_n(\theta) \tag{8.2}$$

el puede ser representa por las condiciones de primer orden de maximizar el lagrangeano $\mathcal{L} = \hat{Q}_n(\theta) + \lambda' g(\theta)$,

$$\frac{s_n(\hat{\theta}^0_{ML})}{n} + \Gamma(\hat{\theta}^0_{ML})'\hat{\lambda} = 0$$
$$g(\hat{\theta}^0_{ML}) = 0$$

Teorema 8.2. *Si las condiciones de regularidad y el supuesto 8.1 se cumple, el test basado en $R_n = \{LM_n \geq \chi^2_{q,1-\alpha}\}$ donde*

$$LM_n = \Big(\frac{s_n(\hat{\theta}^0_{ML})}{\sqrt{n}}\Big)' \tilde{J}(\hat{\theta}^0_{ML})^{-1} \Big(\frac{s_n(\hat{\theta}^0_{ML})}{\sqrt{n}}\Big) \xrightarrow{d} \chi^2_q \tag{8.3}$$

posee nivel asintótico igual α y sobre la hipótesos nula $W_n - LM_n \xrightarrow{p} 0$. $\tilde{J}(\theta)$ es un estimador asintótico de $J()\theta$.

En general el test de multiplicadores de Lagrange es también consistente. Una propiedad de este test es que si usamos el estimador BHHH, entonces $LM_n = nR^2$ donde R^2 es clásico indicador de ajuste de una regresión.

8.3 Test de ratio de verosimilitud

Teorema 8.3. *Si las condiciones de regularidad y el supuesto 8.1 se cumplen, el test basado en $R_n = \{LR_n \geq \chi^2_{q,1-\alpha}\}$ donde*

$$LR_n = -2[\ln L(\hat{\theta}^0_{ML}) - \ln L(\hat{\theta}_{ML})] \xrightarrow{d} \chi^2_q \tag{8.4}$$

es consistente con nivel asintótico α. Bajo la hipótesis nula $LR_n - LM_n \xrightarrow{p} 0$ y $LR_n - W_n \xrightarrow{p} 0$.

9 Sistema de Ecuaciones I

Basado en Hansen (2017), Wooldridge (2010), y clases dictadas. Consideremos una muestra aleatoria $\{(\mathbf{X}_i, \mathbf{y}_i)|i = 1, \ldots, N\}$ donde $\mathbf{X}_i$ es una matriz aleatoria de orden $G \times K$ y $\mathbf{y}_i$ es un vecto aleatorio de orden $G \times 1$. El interés cae sobre β $(K \times 1)$ en

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{u}_i \tag{9.1}$$

donde $\mathbf{u}_i$ es de orden $G \times 1$.

9.1 Estimador OLS

Supuestos:

ols1 $\mathbb{E}[\mathbf{X}_i'\mathbf{u}_i] = 0$.

ols2 $A = \mathbb{E}[\mathbf{X}_i'\mathbf{X}_i]$ es no singular.

ols3 $\mathbb{E}|\mathbf{X}_i'\mathbf{u}_i\mathbf{u}_i'\mathbf{X}_i| < \infty$.

El estimador de mínimos cuadrados va a tener buenas propiedades en este contexto:

$$\hat{\beta}_{ols} = (\sum_i \mathbf{X}_i'\mathbf{X}_i')^{-1} \sum_i \mathbf{X}_i'\mathbf{y}_i \tag{9.2}$$

Teorema 9.1. *Si se cumple ols1 y ols2:* $\hat{\beta}_{ols} \xrightarrow{p} \beta$.

Teorema 9.2. *Si se cumple ols1, ols2 y ols3:* $\sqrt{N}(\hat{\beta}_{ols} - \beta) \xrightarrow{d} \mathcal{N}(0, A^{-1}BA^{-1})$, *donde* $B = \mathbb{E}[\mathbf{X}_i'\mathbf{u}_i\mathbf{u}_i'\mathbf{X}_i]$

Para realizar inferencia con este estimador necesitamos un estimador de la varianza asintótica de $\hat{\beta}_{ols}$:

$$a\hat{var}(\hat{\beta}_{ols}) = (\sum_i \mathbf{X}_i'\mathbf{X}_i)^{-1}[\sum_i \mathbf{X}_i'\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i'\mathbf{X}_i](\sum_i \mathbf{X}_i'\mathbf{X}_i)^{-1} \tag{9.3}$$

donde $\hat{\mathbf{u}}_i = \mathbf{y}_i - \mathbf{X}_i\hat{\beta}_{ols}$. Note que las propiedades de este estimador no depende de la forma de $\Omega = \mathbb{E}(\mathbf{u}_i\mathbf{u}_i')$ ni de $\mathbb{V}(\mathbf{u}_i|\mathbf{X}_i)$.

9.2 Estimador GLS

Supuestos:

gls1 $\mathbb{E}[\mathbf{X}_i \otimes \mathbf{u}_i] = 0$.

gls2 Ω es una matriz positiva definida (p.d. en adelante) y $A_* = \mathbb{E}[\mathbf{X}_i'\Omega^{-1}\mathbf{X}_i]$ es no singular.

gls3 $\mathbb{E}|\mathbf{X}_i'\Omega^{-1}\mathbf{u}_i\mathbf{u}_i'\Omega^{-1}\mathbf{X}_i| < \infty$.

gls4 $\mathbb{E}[\mathbf{X}_i'\Omega^{-1}\mathbf{u}_i\mathbf{u}_i'\Omega^{-1}\mathbf{X}_i] = \mathbb{E}[\mathbf{X}_i'\Omega^{-1}\mathbf{X}_i]$, i.e., $B_* = A_*$.

Observación: a) Aquí el estimador GLS no está corrigiendo un problema de heteroscedasticidad condicionada sino utiliza la información de la matriz de varianzas indondicionada. b) El supuesto gls4 se cumple si se asume homocedasticidad, $\mathbb{E}[\mathbf{u}_i\mathbf{u}_i'|\mathbf{X}_i] = \Omega$.

$$\hat{\beta}_{gls} = (\sum_i \mathbf{X}_i'\Omega^{-1}\mathbf{X}_i')^{-1} \sum_i \mathbf{X}_i'\Omega^{-1}\mathbf{y}_i \tag{9.4}$$

Teorema 9.3. *Si se cumple gls1 y gls2:* $\hat{\beta}_{gls} \xrightarrow{p} \beta$.

Teorema 9.4. *Si se cumple gls1, gls2 y gls3:* $\sqrt{N}(\hat{\beta}_{gls} - \beta) \xrightarrow{d} \mathcal{N}(0, A_*^{-1}B_*A_*^{-1})$, *donde* $B_* = \mathbb{E}[\mathbf{X}_i'\Omega^{-1}\mathbf{u}_i\mathbf{u}_i'\Omega^{-1}\mathbf{X}_i]$

Teorema 9.5. *Si se cumple gls1-4,* $\hat{\beta}_{gls}$ *es el estimador asintóticamente eficiente entre los estimadores consistentes en* $\mathbb{E}[\mathbf{X}_i \otimes \mathbf{u}_i]$. *Además,* $avar(\hat{\beta}_{gls}) = A_*^{-1}/N$.

Es decir, bajo el supuesto de sistema homocedástico GLS es más eficiente que OLS. El problema con este estimador es que no es factible: no conocemos Ω!.

9.3 Estimador GLS factible

- **Paso 1**: Obtener $\hat{\beta}_{ols}$ (si gls1 y gls2 se cumple, este estimador es consistente). Obtener $\hat{\mathbf{u}}_i = \mathbf{y}_i - \mathbf{X}_i\hat{\beta}_{ols}$ y calcular

$$\hat{\Omega} = N^{-1}\sum_i \hat{\mathbf{u}}_i\hat{\mathbf{u}}_i' = N^{-1}\sum_i \mathbf{u}_i\mathbf{u}_i' + o_p(1) \tag{9.5}$$

- **Paso 2:** Calcula el estimador FGLS:

$$\hat{\beta}_{fgls} = (\sum_i \mathbf{X}_i'\hat{\Omega}^{-1}\mathbf{X}_i')^{-1}\sum_i \mathbf{X}_i'\hat{\Omega}^{-1}\mathbf{y}_i \tag{9.6}$$

Teorema 9.6. *Si se cumple gls1-3: $\hat{\beta}_{gls}$ y $\hat{\beta}_{fgls}$ son $\sqrt{N}-$equivalentes: $\sqrt{N}(\hat{\beta}_{fgls} - \hat{\beta}_{gls}) = o_p(1)$; y*

$$\sqrt{N}(\hat{\beta}_{fgls} - \beta) \xrightarrow{d} \mathcal{N}(0, A_*^{-1}B_*A_*^{-1}) \tag{9.7}$$

Teorema 9.7. *Si se cumple gls1-4, $\hat{\beta}_{fgls}$ es un estimador asintóticamente eficiente entre los estimadores consistentes en $\mathbb{E}[\mathbf{X}_i \otimes \mathbf{u}_i]$. Además, $avar(\hat{\beta}_{fgls}) = A_*^{-1}/N$.*

Para hacer pruebas de hipótesis debemos esetimar la varianza asintótica de $\hat{\beta}_{fgls}$:

$$a\hat{v}ar(\hat{\beta}_{fgls}) = (\sum_i \mathbf{X}_i'\hat{\Omega}^{-1}\mathbf{X}_i)^{-1}[\sum_i \mathbf{X}_i'\hat{\Omega}^{-1}\tilde{\mathbf{u}}_i\tilde{\mathbf{u}}_i'\hat{\Omega}^{-1}\mathbf{X}_i](\sum_i \mathbf{X}_i'\hat{\Omega}^{-1}\mathbf{X}_i)^{-1} \tag{9.8}$$

donde $\tilde{\mathbf{u}}_i = \mathbf{y}_i - \mathbf{X}_i\hat{\beta}_{fgls}$.

9.4 Modelo SUR

Cada observación i es descrita mediante un sistema de ecuaciones,

$$\begin{aligned}
y_{i1} &= \mathbf{x}_{i1}'\beta_1 + u_{i1}\\
y_{i2} &= \mathbf{x}_{i2}'\beta_2 + u_{i2}\\
&\vdots\\
y_{iG} &= \mathbf{x}_{iG}'\beta_1 + u_{iG}
\end{aligned}$$

donde $\mathbf{x}_{ig}$ es un vector aleatorio de orden $K_g \times 1$. Este modelo permite una representación como la que se ha analizado en esta sección,

$$\mathbf{y}_i = \begin{bmatrix} y_{i1}\\ \vdots \\ y_{iG}\end{bmatrix}, \ \mathbf{u}_i = \begin{bmatrix} u_{i1}\\ \vdots \\ u_{iG}\end{bmatrix}, \quad \mathbf{X}_i = \begin{bmatrix} \mathbf{x}_{i1}' & 0 & \dots & 0\\ 0 & \mathbf{x}_{i2}' & \dots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \dots & \mathbf{x}_{iG}'\end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1\\ \vdots \\ \beta_G\end{bmatrix}$$

donde $K = K1 + \cdots + K_G$. De tratarse de un modelo estructural (sin variables omitidas, errores de media o simultaneidad) es posible asumir

$$\mathbb{E}[u_{ig}|\mathbf{x}_{i1}, \dots, \mathbf{x}_{iG}] = 0, \quad g = 1, \dots, G \tag{9.9}$$

Esto implica ols1 y gls1. De no poder defender la exogeneidad estricta aún se poddría la exogeneidad débil, $\mathbb{E}[\mathbf{x}_{ig}u_{ig}] = 0$ para todo g, y usar el estimador OLS. Algunas propiedades interesantes de los estimador estudiados en el contexto de modelos SUR son

1. El estimador OLS, $\hat{\beta}_{ols}$, puede obtenerse haciendo estimación OLS ecuación por ecuación. Por tanto, si sólo estamos interesados en β_g, basta que $\mathbb{E}[\mathbf{x}_{ig}u_{ig}]$ para garantizar consistencia de $\hat{\beta}_{ols,g}$.

2. Si $\hat{\Omega}$ es una matriz diagonal, $\hat{\beta}_{ols} = \hat{\beta}_{fgls}$. Si Ω es diagonal, $\hat{\beta}_{ols}$ y $\hat{\beta}_{fgls}$ son equivalente asintóticamente.
3. Si $\mathbf{x}_{i1} = \mathbf{x}_{i2} = \cdots = \mathbf{x}_{iG}$ para todo i, entonces $\hat{\beta}_{ols} = \hat{\beta}_{fgls}$.

Resumen: (i) Estimar β usando un sistema de ecuaciones puede ser útil si desea hacer hipótesis sobre la relación de parámetro en distintas ecuaciones. (ii) OLS es más robusto, mientras que FGLS es asintóticamente más eficiente (si se cumple gls4).

9.5 Modelos de datos de Panel

Para cada unidad de la muestra,

$$y_{it} = \mathbf{x}'_{it}\beta + u_{it}, \quad t = 1, \ldots, T$$

A diferencia del contexto de modelos SUR, la variable dependiente y regresores son los mismos pero observados para puntos distintos de tiempo, y el parámetro de interés es el mismo para cada ecuación.

$$\mathbf{y}_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{iT} \end{bmatrix}, \quad \mathbf{u}_i = \begin{bmatrix} u_{i1} \\ \vdots \\ u_{iT} \end{bmatrix}, \quad \mathbf{X}_i = \begin{bmatrix} \mathbf{x}'_{i1} \\ \vdots \\ \mathbf{x}'_{iT} \end{bmatrix}$$

Es suficiente que $\mathbb{E}[\mathbf{x}_{it}u_{it}] = 0$ $(t = 1, \ldots, T)$ para que se cumple ols1. Si es posible asumir a) $\mathbb{E}[u_t^2\mathbf{x}_t\mathbf{x}'_t] = \sigma^2\mathbb{E}[\mathbf{x}_t\mathbf{x}'_t]$ para $t = 1, \ldots, T$ donde $\sigma^2 = \mathbb{E}[u_t^2]$, y b) $\mathbb{E}[u_tu_s\mathbf{x}_t\mathbf{x}'_t] = 0$ para $t \neq s$; entonces OLS es equivalente a FGLS, y por tanto, eficiente.

Este tipo de modelos sirve para estudiar relaciones dinámicas y permite controlar por heterogeneidad no observable. Será más estudiado más adelante. Asimismo, al incrementar el tamaño de la muestra obtenemos estimadores más precisos y permite realizar pruebas de hipótesis más elaboradas.

10 Sistema de Ecuaciones II: Problema de endogeneidad

Basado en Hansen (2017), Wooldridge (2010), Hansen and Singleton (1982), y clases dictadas. Ahora trabajaremos en un modelo donde $\mathbb{E}[\mathbf{X}'_i\mathbf{u}_i] \neq 0$. El modelo es descrito por

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{u}_i \tag{10.1}$$

Los supuestos que haremos,

siv1 $\mathbb{E}[\mathbf{Z}'_i\mathbf{u}_i] = 0$, donde $\mathbf{Z}_i$ es una matriz de variables instrumentales de orden $G \times L$.

siv2 $rank\{C\} = K$, donde $C = \mathbb{E}[\mathbf{Z}'_i\mathbf{X}_i]$ (condición de rango).

siv3 $\hat{W} \xrightarrow{p} W$, donde W es simétrica de orden $L \times L$y p.d.

siv4 $W = \Lambda^{-1}$ donde $\Lambda = \mathbb{E}[\mathbf{Z}'_i\mathbf{u}_i\mathbf{u}'_i\mathbf{Z}_i]$.

siv5 $\mathbb{E}[\mathbf{Z}'_i\mathbf{u}_i\mathbf{u}'_i\mathbf{Z}_i] = \mathbf{E}[\mathbf{Z}_i\Omega\mathbf{Z}_i]$ donde $\Omega = \mathbb{E}[\mathbf{u}_i\mathbf{u}'_i]$.

siv6 $\mathbb{E}[\mathbf{Z}_i \otimes \mathbf{u}_i] = 0$ para todo i.

Note que, siv2 implica que $L \geq K$ (condición de orden). Un resultado directo es que *la identificación de un modelo SUR es equivalente a la identificación de ecuación por ecuación*, i.e., siv3 $\Leftrightarrow rank\{\mathbb{E}[\mathbf{z}_{ig}\mathbf{x}'_{ig}]\} = K_g$, donde $\mathbf{z}_{ig}$ es de orden $L_g \times 1$.

La función de momentos es $\mathbb{E}[\mathbf{Z}'_i(\mathbf{y}_i - \mathbf{X}_i\beta)] = 0$, y el estimador GMM es

$$\hat{\beta}_{gmm} = arg\min_{\beta}[N^{-1}\sum_i \mathbf{Z}'_i(\mathbf{y}_i - \mathbf{X}_i\beta)]'\hat{W}[N^{-1}\sum_i \mathbf{Z}'_i(\mathbf{y}_i - \mathbf{X}_i\beta)]$$

lo que da

$$\hat{\beta}_{gmm} = [(\sum_i \mathbf{X}_i'\mathbf{Z}_i)\hat{W}(\sum_i \mathbf{Z}_i'\mathbf{X}_i)]^{-1}[(\sum_i \mathbf{X}_i'\mathbf{Z}_i)\hat{W}(\sum_i \mathbf{Z}_i'\mathbf{y}_i)] \tag{10.2}$$

Teorema 10.1. *Si se cumple siv1-3, entonces* $\hat{\beta}_{gmm} \xrightarrow{p} \beta$. *Además,* $\sqrt{N}(\hat{\beta}_{gmm} - \beta) \xrightarrow{d} \mathcal{N}(0, (C'WC)^{-1}C'W\Lambda WC(C'WC)^{-1})$.

Teorema 10.2. *Si se cumple siv1-4, entonces* $\hat{\beta}_{gmm}$ *se convierte en el estimador asintóticamente más eficiente entre todos los estimadores de tipo GMM. Además,* $avar(\hat{\beta}_{gmm}) = (C'\Lambda C)^{-1}/N$.

10.1 Estimador IV: sistema identificado

Si $L = K$, decimos que el sistema está identificado (siv1-3). Luego

$$\hat{\beta}_{iv} = (\sum_i \mathbf{Z}_i'\mathbf{X}_i)^{-1}\sum_i \mathbf{Z}_i'\mathbf{y}_i \tag{10.3}$$

Note que este estimador no depende de la elección de $\hat{W}$, por tanto, el conjunto de estimadores GMM es singleton y no tiene mucho sentido hablar del más eficiente en este contexto.

10.2 Estimador 2SLS: sistemas sobreidentificados

Este estimador usa $\hat{W} = (N^{-1}\sum_i \mathbf{Z}_i'\mathbf{Z}_i)^{-1} \xrightarrow{p} (\mathbb{E}[\mathbf{Z}_i'\mathbf{Z}_i])^{-1}$. Luego

$$\hat{\beta}_{2sls} = [(\sum_i \mathbf{X}_i'\mathbf{Z}_i)(\sum_i \mathbf{Z}_i'\mathbf{Z}_i)^{-1}(\sum_i \mathbf{Z}_i'\mathbf{X}_i)]^{-1}[(\sum_i \mathbf{X}_i'\mathbf{Z}_i)(\sum_i \mathbf{Z}_i'\mathbf{Z}_i)^{-1}(\sum_i \mathbf{Z}_i'\mathbf{y}_i)] \tag{10.4}$$

Si se aplica este estimador a un sistema de ecuaciones tipo SUR, entonces es equivalente a usar 2SLS ecuación por ecuación.

10.3 Estimador "Two-step GMM" o mínimo chi-cuadrado

Es el estimador GMM eficiente factible.

- **Paso 1**: Dado un $\tilde{W}$, estimar $\tilde{\beta}_{gmm}$, obtener $\tilde{\mathbf{u}}_i = \mathbf{y}_i - \mathbf{X}_i\tilde{\beta}_{gmm}$ para todo i y calcular

$$\hat{\Lambda} = N^{-1}\sum_i \mathbf{Z}_i'\tilde{\mathbf{u}}_i\tilde{\mathbf{u}'}_i\mathbf{Z}_i \tag{10.5}$$

- **Paso 2:** Calcular $\hat{W} = \hat{\Lambda}^{-1}$ y estimar $\hat{\beta}_{gmm}$. Para hacer inferencia,

$$a\hat{var}(\hat{\beta}_{gmm}) = [(\sum_i \mathbf{X}_i'\mathbf{Z}_i)(\sum_i \mathbf{Z}_i'\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i'\mathbf{Z}_i)^{-1}(\sum_i \mathbf{Z}_i\mathbf{X}_i)]^{-1} \tag{10.6}$$

donde $\hat{\mathbf{u}}_i = \mathbf{y}_i - \mathbf{X}_i\hat{\beta}_{gmm}$.

Note que si $\mathbf{Z}_i = \mathbf{X}_i$, entonces $a\hat{var}(\hat{\beta}_{gmm})$ es igual al estimador robusto de la varianza asintótica de OLS. Si $\mathbf{Z}_i = \hat{\Omega}^{-1}\mathbf{X}_i$, entonces $a\hat{var}(\hat{\beta}_{gmm})$ es igual al estimador robusto de la varianza asintótica de FGLS.

10.4 Estimador 3SLS

- **Paso 1**: Dado un $\tilde{W}$, estimar $\tilde{\beta}_{gmm}$, obtener $\tilde{\mathbf{u}}_i = \mathbf{y}_i - \mathbf{X}_i\tilde{\beta}_{gmm}$ para todo i y calcular

$$\hat{\Omega} = N^{-1}\sum_i \tilde{\mathbf{u}}_i \tilde{\mathbf{u}'}_i \tag{10.7}$$

- **Paso 2:** Calcular $\hat{W} = (N^{-1}\sum_i \mathbf{Z}_i'\hat{\Omega}\mathbf{Z}_i)^{-1}$ y estimar $\hat{\beta}_{3sls} \equiv \hat{\beta}_{gmm}$.

Teorema 10.3. *Si se cumple siv1-3 y siv5, $\hat{\beta}_{3sls}$ es el estimador GMM óptimo.*

Entonces, para hacer inferencia,

$$a\hat{var}(\hat{\beta}_{3sls}) = [(\sum_i \mathbf{X}_i'\mathbf{Z}_i)(\sum_i \mathbf{Z}_i'\hat{\Omega}\mathbf{Z}_i)^{-1}(\sum_i \mathbf{Z}_i\mathbf{X}_i)]^{-1} \tag{10.8}$$

Note que este estimador de varianza asintótica no es robusto. Es evidente, que “two-step GMM” es al menos tan bueno como 3SLS e incluso es más robusto a la validez de siv5. Entonces, ¿como justificar el uso de 3sls?

1. Las bondades de 3SLS han sido más estudiados a lo largo de la historia, mientras que GMM óptimo es más reciente.
2. Si se cumple siv5, 3SLS podría tener mejores propiedades en muestras finitas.

Además, si $\hat{\Omega}$ es diagonal entonces $\hat{\beta}_{2sls} = \hat{\beta}_{3sls}$, y si en todo caso, Ω fuera diagonal, ambos estimadores son asintóticamente equivalentes.

10.5 Estimador GIV

Podemos usar una transformación a la ecuación antes de estimar similar a la usada en GLS. Entonces, el estimador generalizado de variables instrumentales, $\tilde{\mathbf{Z}}_i = \Omega^{-1/2}\mathbf{Z}_i$ es

$$\hat{\beta}_{giv} = [(\sum_i \mathbf{X}_i'\Omega^{-1}\mathbf{Z}_i)(\sum_i \mathbf{Z}_i'\Omega^{-1}\mathbf{Z}_i)(\sum_i \mathbf{Z}_i'\Omega^{-1}\mathbf{Z}_i)]^{-1}(\sum_i \mathbf{X}_i'\Omega^{-1}\mathbf{Z}_i)(\sum_i \mathbf{Z}_i'\Omega^{-1}\mathbf{Z}_i)(\sum_i \mathbf{Z}_i'\Omega^{-1}\mathbf{y}_i) \tag{10.9}$$

Para que $\hat{\beta}_{giv}$ sea consistente precisamos del supuesto siv6. Para obtener la contraparte factible de este estimador, remplazamos Ω por un estimador consistente de este, $\hat{\Omega}$.

10.6 Pruebas de hipótesis

Aquí asumiremos que $\hat{\beta}$ es el estimador mínimo chi-cuadrado. Para hacer pruebas de hipótesis podemos usar pruebas de tipo Wald, pero en el contexto de estimador GMM aperece un estadístico que sigue la misma idea del estadístico de ratios de verosimilitud. Si la hipótesis nula es $H_0 : h(\beta) = 0$ donde $h : \mathbb{R}^k \to \mathbb{R}^Q$, y $\tilde{\beta}$ es el estimador GMM restringido a $h(\beta) = 0$ con la misma matriz de pesos $\hat{W}$, entonces bajo $H0$,

$$(\sum_i \mathbf{Z}_i'\tilde{\mathbf{u}}_i)'\hat{W}(\sum_i \mathbf{Z}_i'\tilde{\mathbf{u}}_i) - (\sum_i \mathbf{Z}_i'\hat{\mathbf{u}}_i)'\hat{W}(\sum_i \mathbf{Z}_i'\hat{\mathbf{u}}_i)/N \xrightarrow{d} \chi^2_Q \tag{10.10}$$

Si queremos testear la hipótesis de sobreidentificación: $H0 : \mathbb{E}[\mathbf{Z}_i'\mathbf{u}_i] = 0$, usamos el estadístico J

$$(\sum_i \mathbf{Z}_i'\hat{\mathbf{u}}_i)'\hat{W}(\sum_i \mathbf{Z}_i'\hat{\mathbf{u}}_i)/N \xrightarrow{d} \chi^2_{L-K} \tag{10.11}$$

Algunas observaciones: (i) Si usamos más instrumentos, los grados de libertad aumentan y estos reduce la probabilidad de rechazar $H0$, (ii) si G es pequeño en relación a N el test será sobrerechazado, y (iii) debido a estos problemas podemos usar el test annterior si el modelo puede ser anidado a un modelo general alternativo.

10.7 Ejemplo: Hansen and Singleton (1982)

Considere

$$\max \mathbb{E}_t\Big\{\sum_{s=0}^{\infty}\beta^s u(c_{t+s})\Big\}$$

$$s.a.:$$

$$\sum_{j=1}^{N} P_{jt}Q_{jt} + c_t = \sum_{j=1}^{N} X_{jt}Q_{j,t-j} + w_t$$

donde P_{jt} es el precio del activo con madurez j que promete un único pago de $X_{j,t+j}$. Las CPO del problem es el sistema

$$1 = \beta^j \mathbb{E}_t\Big[R_{j,t+j}\frac{u'(c_{t+j})}{u'(c_t)}\Big], \quad j = 1, \ldots, N$$

donde $R_{j,t+j} = X_{j,t+j}/P_{jt}$. Si usamos una función de utilidad CRRA, $u(c) = (1-\rho)^{-1}c^{1-\rho}$, y definimos $\theta = [\beta, \rho]$. El sistema que nos interesa sería

$$1 = \beta^j \mathbb{E}_t\Big[R_{j,t+j}\Big(\frac{c_{t+j}}{c_t}\Big)^{-\rho}\Big], \quad j = 1, \ldots, N$$

y la muestra es, aunque en este contexto no es aleatoria, $\{y_t\}_{i=1}^{T+N}$ donde $y_t = \Big(\frac{c_{t+1}}{c_t}, R_{1,t+1}, \ldots, \frac{c_{t+N}}{c_t}, R_{N,t+N}\Big)'$. Si definimos

$$u_t = \begin{bmatrix} \beta R_{j,t+1}\Big(\frac{c_{t+1}}{c_t}\Big)^{-\rho} - 1 \\ \beta^2 R_{j,t+2}\Big(\frac{c_{t+2}}{c_t}\Big)^{-\rho} - 1 \\ \vdots \\ \beta^N R_{j,t+N}\Big(\frac{c_{t+N}}{c_t}\Big)^{-\rho} - 1 \end{bmatrix} \tag{10.12}$$

La condición de primer orden y el supuesetos de expectativas racionales nos dan la condición de exogeneidad $\mathbb{E}[u_t|z_t] = 0$, donde z_t es un vector de información de L variables en el momento t. Podemos usar la función de momentos $\mathbb{E}[u_t \otimes z_t] = 0$, luego

$$\hat{\theta}_{gmm} = arg\min_{\theta}(T^{-1}\sum_{t=1}^{T} u_t \otimes z_t)'W_T(T^{-1}\sum_{t=1}^{T} u_t \otimes z_t)$$

A este tipo de estimadores se le conoce como estimador GMM mediante Ecuación de Euler.

11 Modelos de datos de Panel: Básico

Basado en Greene (2003), Wooldridge (2010), y clases dictadas. Solo estudiaremos modelos donde la heterogeneidad no observable es contante en el tiempo, caracterizados por T fijo y $N \to \infty$ (*short panel*). De nuevo la muestra es aleatoria: $\{(\mathbf{y}_i, \mathbf{X}_i)|i = 1, \ldots, T\}$. El modelo es descrito por

$$y_{it} = \mathbf{x}_{it}'\beta + v_{it} \tag{11.1}$$

$$v_{it} = c_i + u_{it}, \quad t = 1, \ldots, T \tag{11.2}$$

o en su contraparte compacta

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{v}_i \tag{11.3}$$

$$\mathbf{v}_i = c_i\mathbf{1}_T + \mathbf{u}_i \tag{11.4}$$

donde c_i es una variable aleatoria no observada. Los supuestos de estimación establecen relaciones $c_i, \mathbf{x}_{it}$, y u_{it}.

11.1 Estimador "Pooled OLS"

Supuestos

p1 $\mathbb{E}[\mathbf{x}_{it}u_{it}] = 0$ y $\mathbb{E}[\mathbf{x}_{it}c_i] = 0$ para todo t.

p2 $rank\{\mathbb{E}[\sum_{t=1}^{T}\mathbf{x}_{it}\mathbf{x}_{it}']\} = K$.

Si se cumple p1 y p2, entonces $\hat{\beta}_{pols}$ es consistente.

$$\hat{\beta}_{pols} = (\sum_i \mathbf{X}_i'\mathbf{X}_i)^{-1}(\sum_i \mathbf{X}_i'\mathbf{y}_i) = (\sum_i\sum_t \mathbf{x}_{it}\mathbf{x}_{it}')^{-1}(\sum_i\sum_t \mathbf{x}_{it}y_{it}) \tag{11.5}$$

En general, este estimador tiene las propiedades de OLS para un sistema de ecuaciones (ver arriba). Note que bajo los supuestos re1-3 (ver método de efectos aleatorios),

$$\mathbb{V}[\hat{\beta}_{pols}|\mathbf{X}] = (\sum_i \mathbf{X}_i'\mathbf{X}_i)^{-1}(\sum_i \mathbf{X}_i'\Omega\mathbf{X}_i)(\sum_i \mathbf{X}_i'\mathbf{X}_i)^{-1} \tag{11.6}$$

y la varianza asintótica es estimada robustamente por

$$a\hat{v}ar(\hat{\beta}_{pols}) = (\sum_i \mathbf{X}_i'\mathbf{X}_i)^{-1}(\sum_i \mathbf{X}_i'\hat{\mathbf{v}}_i\hat{\mathbf{v}}_i'\mathbf{X}_i)(\sum_i \mathbf{X}_i'\mathbf{X}_i)^{-1} \tag{11.7}$$

11.2 Método de efectos aleatorios

Los siguientes supuestos son más fuertes de lo necesario pero más fáciles de interpretar.

re1 a) $\mathbb{E}[u_{it}|\mathbf{X}_i, c_i] = 0$ para todo t, y b) $\mathbb{E}[c_i|\mathbf{X}_i] = \mathbb{E}[c_i] = 0$.

re2 $rank\{\mathbb{E}[\mathbf{X}_i'\Omega^{-1}\mathbf{X}_i']\} = K$, donde $\Omega = \mathbb{E}[\mathbf{v}_i\mathbf{v}_i']$.

re3 a) $\mathbb{E}[\mathbf{u}_i\mathbf{u}_i'|\mathbf{X}_i, c_i] = \sigma_u^2 I_T$, y b) $\mathbb{E}[c_i^2|\mathbf{X}_i] = \sigma_c^2$.

De esta forma la distribución de $c_i - \overbrace{\mathbb{E}[c_i]}^{=0} = c_i$ no depende de las características del individuo i. Además, si se cumplen re1 y re3:

$$\begin{aligned}
\mathbb{E}[\mathbf{v}_i\mathbf{v}_i'|\mathbf{X}_i] &= \mathbb{E}[(c_i\mathbf{1}_T + \mathbf{u}_i)(c_i\mathbf{1}_T + \mathbf{u}_i)'|\mathbf{X}_i] \\
&= \mathbb{E}[c_i^2\mathbf{1}_T\mathbf{1}_T' + \mathbf{u}_i\mathbf{u}_i'|\mathbf{X}_i] \\
&= \sigma_u^2 I_t + \sigma_c^2\mathbf{1}_T\mathbf{1}_T'
\end{aligned}$$

cuando se cumplen dichos supuestos, se dice que Ω posee la **estructura de efectos aleatorios**. Note también que se cumple el supuesto de sistema homocedástico (gls4): $\mathbb{E}[\mathbf{v}_i\mathbf{v}_i'|\mathbf{X}_i] = \mathbb{E}[\mathbf{v}_i\mathbf{v}_i']$. Así, los supuestos re1-3 implican gls1-4, y por tanto, la eficiencia del estimador GLS e FGLS. En general, note que el estimador no sólo es consistente sino también es insesgado si r1-2 es válido.

- **Paso 1**: Obtener $\tilde{\beta}_{pols}$ (que bajo re1-2, es consistente), y estimar[1]

$$\hat{\sigma}_v^2 = (NT - K)^{-1}\sum_i\sum_t \tilde{v}_{it}^2$$

$$\hat{\sigma}_c^2 = [NT(T-1)/2 - K]^{-1}\sum_i\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\tilde{v}_{it}\tilde{v}_{is}$$

$$\hat{\sigma}_u^2 = \hat{\sigma}_v^2 - \hat{\sigma}_c^2$$

[1]Dada la estructura es fácil mostrar que

$$\mathbb{E}[\sum_{t=1}^{T-1}\sum_{s=t+1}^{T} v_{it}v_{is}] = \sigma_c^2\frac{T}{2}(T-1)$$

- **Paso 2**: Calcular $\hat{\Omega} = \hat{\sigma}_u^2 I_T + \hat{\sigma}_c^2 \mathbf{1}_t \mathbf{1}_T'$ y estimar

$$\hat{\beta}_{re} = (\sum_i \mathbf{X}_i \hat{\Omega}^{-1} \mathbf{X}_i)^{-1} (\sum_i \mathbf{X}_i' \hat{\Omega}^{-1} \mathbf{y}_i) \tag{11.8}$$

El hecho que $\hat{\sigma}_c^2 < 0$ es un indicio que re3 no se cumple. Para hacer inferencia es prudente usar el estimador robusto de varianza asintótica de los estimadores:

$$a\hat{v}ar(\hat{\beta}_{re}) = (\sum_i \mathbf{X}_i' \hat{\Omega}^{-1} \mathbf{X}_i^{-1})(\sum_i \mathbf{X}_i' \hat{\Omega}^{-1} \hat{\mathbf{v}}_i \hat{\mathbf{v}}_i' \hat{\Omega}^{-1} \mathbf{X}_i)(\sum_i \mathbf{X}_i' \hat{\Omega}^{-1} \mathbf{X}_i)^{-1}) \tag{11.9}$$

Si se satisfacen re1-3, el resultado que la correlación serial de $\{v_{it}\}$ viene dado por la varianza de c_i permite testear la existenccia de heterogeneidad no observable, $H_0 0 : \sigma_c^2 = 0$. Sabemos que bajo esos supuestos y H_0

$$\begin{aligned}
&N^{-1/2} \sum_i \sum_{t=1}^{T-1} \sum_{s=t+1}^{T} \hat{v}_{it} \hat{v}_{is} \xrightarrow{d} \mathcal{N}(0, \mathbb{E}[(\sum_{t=1}^{T-1} \sum_{s=t+1}^{T} v_{it} v_{is})^2]) \\
&\Rightarrow \frac{\sum_i \sum_{t=1}^{T-1} \sum_{s=t+1}^{T} \hat{v}_{it} \hat{v}_{is}}{\left[\sum_i \left(\sum_{t=1}^{T-1} \sum_{s=t+1}^{T} \hat{v}_{it} \hat{v}_{is} \right)^2 \right]^{1/2}} \xrightarrow{d} \mathcal{N}(0,1)
\end{aligned}$$

Note que si se rechaza, es un indicador de existencia de correlación serial mas no garantiza que re3 sea un supuesto razonable. De no existir efectos no observables, el estimador $\hat{\beta}_{pols}$ es eficiente y toda inferencia se puede realizar con estadísticos basados en dicho estimador.

Podemos relajar el supuesto re3 a simplemente asumir $\mathbb{E}[\mathbf{v}_i \mathbf{v}_i'] = \Omega$ y usar $\hat{\Omega} = N^{-1} \sum_i \tilde{\mathbf{v}}_i \tilde{\mathbf{v}}_i'$ en el paso 1 y 2 (*análisis FGLS general*).

11.3 Método de efectos fijos

A diferencia del método con efectos aleatorios aquí sí se permite correlación arbitraria entre $\mathbf{x}_{it}$ y c_i, es decir se relaja el supuesto re1b. Por tanto, no debe sorprendernos que para obtener estimadores consistentes se debe transformar el sistema con c_i a una ecuación de estimación donde no aparezca c_i y poder utilizar el estimador POLS. Supuestos

fe1 $\mathbb{E}[u_{it} | \mathbf{X}_i, c_i] = 0$ para todo $t = 1, \ldots, T$.

fe2 $rank\{\mathbb{E}[\tilde{\mathbf{X}}_i' \tilde{\mathbf{X}}_i]\} = rank\{\sum_{t=1}^T \mathbb{E}[\tilde{\mathbf{x}}_{it} \tilde{\mathbf{x}}_{it}']\} = K$, donde $\tilde{\mathbf{x}}_{it} = \mathbf{x}_{it} - \overline{\mathbf{x}}_i$, $\overline{\mathbf{x}}_i = T^{-1} \sum_{t=1}^T \mathbf{x}_{it}$, $\tilde{\mathbf{X}}_i = Q_T \mathbf{X}_i$.

fe3 $\mathbb{E}[\mathbf{u}_i \mathbf{u}_i' | \mathbf{X}_i, c_i] = \sigma_u^2 I_t$.

fe4 $\mathbb{E}[\mathbf{u}_i \mathbf{u}_i' | \mathbf{X}_i, c_i] = \Lambda$ es una matriz p.d.

Si expresamos la relación arbitraria entre c_i y $\mathbf{X}_i$ mediante $c_i = h(\mathbf{X}_i)$, entonces el modelo puede reescribirse

$$y_{it} = \mathbf{x}_{it}' \beta + c_i + \overbrace{e_{it} + \tilde{c}_i - h(\mathbf{X}_i)}^{u_{it}} \tag{11.10}$$

donde $\mathbb{E}[\tilde{c}_i | \mathbf{X}_i] = h(\mathbf{X}_i) = c_i$, y el supuesto fe1 sigue siendo válido. Por tanto, los supuestos de estimación de efectos fijos permiten que cualquier variación de la heterogeneidad inobservable causada por características distintas a $\mathbf{x}_{it}$, $\tilde{c}_i - c_i$, sea absorbida por la perturbación. Nótese que, *la heterogeneidad es tratada como una variable fija para cada grupo, $\mathbf{X}_i$, mas no quiere decir que no sea aleatoria.*

Transformación "between"

$$\overline{y}_i = \overline{\mathbf{x}}_i'\beta + c_i + \overline{u}_i \tag{11.11}$$

donde $\overline{y}_i = T^{-1}\sum_t y_{it}$ y $\overline{u}_i = T^{-1}\sum_t u_{it}$. Está transformación no elimina la heterogeneidad no observable de la ecuación, por tanto, para que el estimador

$$\hat{\beta}_{between} = (\sum_i \overline{\mathbf{x}}_i\overline{\mathbf{x}}_i')^{-1}\sum_i \overline{\mathbf{x}}_i\overline{y}_i \tag{11.12}$$

sea consistente precisamos del supuesto re1. Sin embargo, si asumimos re1, es mejor usar $\hat{\beta}_{re}$. En particular note que (bajo re1-2),

$$\mathbb{V}(\hat{\beta}_{between}|\mathbf{X}) = (\sum_i \overline{\mathbf{x}}_i\overline{\mathbf{x}}_i')^{-1}(\sum_i \sigma^2_{\overline{v},i}\overline{\mathbf{x}}_i\overline{\mathbf{x}}_i')(\sum_i \overline{\mathbf{x}}_i\overline{\mathbf{x}}_i')^{-1} \tag{11.13}$$

donde $\sigma^2_{\overline{v},i} = T^{-2}\mathbb{E}[\mathbf{1}_T'Q(\mathbf{X}_i)\mathbf{1}_T]$ y $Q(\mathbf{X}_i) = \mathbb{E}[\mathbf{v}_i\mathbf{v}_i'|\mathbf{X}_i]$. Si añadimos re3, $Q(\mathbf{X}_i) = Q \Rightarrow \mathbf{1}_T'Q\mathbf{1}_T = T\sigma_u^2 + T^2\sigma_c^2$, luego

$$\mathbb{V}(\hat{\beta}_{between}|\mathbf{X}) = [\sigma_u^2/T + \sigma_c^2](\sum_i \overline{\mathbf{x}}_i\overline{\mathbf{x}}_i')^{-1} \tag{11.14}$$

Transformación "within"

$$\tilde{y}_{it} = \tilde{\mathbf{x}}_{it}'\beta + \tilde{u}_{it}, \quad t = 1,\dots,T \tag{11.15}$$

Aquí sí se elimino la heterogeneidad no observable. Note que esta ecuación es sólo para la estimación la interpretación de los parámetros sigue dependiendo del modelo de regresión inicial. Además, *todos los regresores deben variar con el tiempo*, i.e., no podemos incluir *directamente* regresores como género, raza, industria, etc. El modelo puede ser reexpresado usando la matriz $Q_T = I_T - \mathbf{1}_T(\mathbf{1}_T'\mathbf{1}_T)^{-1}\mathbf{1}_T'$,

$$\tilde{\mathbf{y}}_i = \tilde{\mathbf{X}}_i\beta + \tilde{\mathbf{u}}_i \tag{11.16}$$

donde $\tilde{\mathbf{y}}_i = Q_T\mathbf{y}_i$ y $\tilde{\mathbf{u}}_i = Q_T\mathbf{u}_i$. De esta forma el estimador de efectos fijos,

$$\hat{\beta}_{fe} = (\sum_i \tilde{\mathbf{X}}_i'\tilde{\mathbf{X}}_i)^{-1}\sum_i \tilde{\mathbf{X}}_i'\tilde{\mathbf{y}}_i = (\sum_i\sum_t \tilde{\mathbf{x}}_{it}\tilde{\mathbf{x}}_{it}')^{-1}\sum_i\sum_t \tilde{\mathbf{x}}_{it}\tilde{y}_{it} \tag{11.17}$$

Si se cumple fe1-3, entonces: a) el estimador FE es consistente e insesgado, y b) tenemos qué $\mathbb{E}[\mathbf{u}_i\mathbf{u}_i'|\tilde{\mathbf{X}}_i] = \sigma_u^2 I_T$, luego el estimador puede ser reescrito

$$\begin{aligned}
\sqrt{N}(\hat{\beta}_{fe} - \beta) &= (\sum_i \tilde{\mathbf{X}}_i'\tilde{\mathbf{X}}_i)^{-1}\sum_i \tilde{\mathbf{X}}_i'\tilde{\mathbf{u}}_i \\
&= (\sum_i \tilde{\mathbf{X}}_i'\tilde{\mathbf{X}}_i)^{-1}\sum_i \mathbf{X}_i'\underbrace{Q_T'Q_T}_{Q_T}\mathbf{u}_i \\
&= (\sum_i \tilde{\mathbf{X}}_i'\tilde{\mathbf{X}}_i)^{-1}\sum_i \tilde{\mathbf{X}}_i'\mathbf{u}_i
\end{aligned}$$

por tanto,

$$\sqrt{N}(\hat{\beta}_{fe} - \beta) \xrightarrow{d} \mathcal{N}(0, \sigma_u^2\Big[\mathbb{E}(\tilde{\mathbf{X}}_i'\tilde{\mathbf{X}}_i)\Big]^{-1}) \tag{11.18}$$

Note que aunque $\{\tilde{u}_{it}\}_{t=1}^T$ tiene correlación serial

$$\begin{aligned}
\mathbb{E}[\tilde{u}_{it}^2] &= \sigma_u^2(1 - 1/T) \\
\mathbb{E}[\tilde{u}_{it}\tilde{u}_{is}] &= -\sigma_u^2/T \\
corr(\tilde{u}_{it}\tilde{u}_{is}) &= -\frac{1}{T-1}
\end{aligned}$$

El estimador $\hat{\beta}_{fe}$ tiene varianza asintótica igual al de un estimador eficiente. Para estimar la varianza asintótica debemos tener cuidado:

$$\sum_{t=1}^{T} \mathbb{E}[\tilde{u}_{it}^2] = (T-1)\sigma_u^2$$

Un estimador ideal sería $[N(T-1)]^{-1}\sum_i \sum_t \tilde{u}_{it}^2$, pero los residuos verdaderos no se conocen así que el estimador que normalmente se utiliza es

$$\hat{\sigma}_u^2 = \frac{\sum_i \sum_t \hat{u}_{it}^2}{N(T-1)-K} \tag{11.19}$$

donde $\hat{u}_{it}^2 = \tilde{y}_{it} - \tilde{\mathbf{x}}_{it}'\beta$. Asimismo, podemos estimar σ_v^2 con $\hat{\sigma}_v^2 = (NT-K)^{-1}\sum_i\sum_t (y_{it} - \mathbf{x}_{it}'\hat{\beta}_{fe})^2$ y $\hat{\sigma}_c^2 = \hat{\sigma}_v^2 - \hat{\sigma}_u^2$.

Si $\{u_{it}\}_t$ posee correlación serial, entonces el supuesto fe3 no se cumple; por tanto, podemos hacer la prueba de hipótesis de ausencia de correlación serial. Si $T=2$, para cualquier tipo de correlación serial de u_{it}, siempre tenemos que $\tilde{u}_{i1} = u_{i1} - (u_{i1}+u_{i2})/2 = (u_{i1}-u_{i2})/2 = -u_{i2} + (u_{i1}+u_{i2})/2 = -\tilde{u}_{i2}$, por tanto es inútil intentar hacer inferencia en cualquier escenario. Si $T \geq 3$, se puede hacer una regresión de $\hat{u}_{it}$ con $\hat{u}_{i,t-1}$ y hacer la prueba de hipótesis de $H0 : \delta = -\frac{1}{T-1}$. De encontrarse correlación serial, es mejor hacer inferencia con el estimador robusto de varianza asintótica de $\hat{\beta}_{fe}$,

$$a\hat{var}(\hat{\beta}_{fe}) = (\sum_i \tilde{\mathbf{X}}_i'\tilde{\mathbf{X}}_i)^{-1}[\sum_i \tilde{\mathbf{X}}_i'\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i'\tilde{\mathbf{X}}_i](\sum_i \tilde{\mathbf{X}}_i'\tilde{\mathbf{X}}_i)^{-1} \tag{11.20}$$

Regresión con variables dummy. Considere $dj_i = \mathbb{I}(j=i)$, y estime un POLS de y_{it} sobre $d1_i, \ldots, dN_i, \mathbf{x}_{it}$, luego la heterogeneidad no observable c_j es estimador por el coeficiente que acompaña dj_i, $\hat{c}_j$. El estimador de los coeficientes de $\mathbf{x}_{it}$ es igual a $\hat{\beta}_{fe}$, además la estimación de σ_u^2 considera, correctamente, $NT - N - K = N(T-1) - K$ grados de libertad. Sin embargo, la estimación de c_j no es consistente cuando $N \to \infty$. Podemos estimar c_i mediante un criterio de mínimos cuadrados, $\min_{c_i} \sum_i (\overline{y}_i - \overline{\mathbf{x}}_i'\beta - c_i)$,

$$\hat{c}_i = \overline{y}_i - \overline{\mathbf{x}}_i'\hat{\beta}_{fe}, \qquad i = 1, \ldots, N \tag{11.21}$$

De esta forma podemos obtener una distribución de la heterogeneidad no observable. Ahora hagamos un poco de algebra, supongamos que quisieramos esetimar $\{c_i\}_{i=1}^N$,

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{1}_T c_i + v_{it}$$

Debe notarse que toda variable invariante en el tiempo es tratada como heterogeneidad en c_i. Defina la matriz *stack*: $\mathbf{y} = (\mathbf{y}_1', \ldots, \mathbf{y}_N')'$, $\mathbf{X} = (\mathbf{X}_1', \ldots, \mathbf{X}_N')'$, $D = I_N \otimes \mathbf{1}_T$, $c = (c_1, \ldots, c_N)'$, y $\mathbf{u} = (\mathbf{u}_1', \ldots, \mathbf{u}_N')'$; luego

$$\mathbf{y} = \mathbf{X}\beta + Dc + \mathbf{u} \tag{11.22}$$

Definamos $M_D = I_{NT} - D(D'D)^{-1}D' = I_{NT} - (I_N \otimes \mathbf{1}_T)[(I_N \otimes \mathbf{1}_T)'(I_N \otimes \mathbf{1}_T)]^{-1}(I_N \otimes \mathbf{1}_T)' = I_{NT} - (I_N \otimes \mathbf{1}_T)[I_N \otimes T^{-1}](I_N \otimes \mathbf{1}_T') = I_{NT} - (I_N \otimes T^{-1}\mathbf{1}_T\mathbf{1}_T')$, luego

$$M_D = \begin{bmatrix} Q_T & 0 & \ldots & 0 \\ 0 & Q_T & \ldots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \ldots & Q_T \end{bmatrix} \tag{11.23}$$

Por el teorema FWL,

$$\hat{\beta}_{fe} = (\mathbf{X}'M_D\mathbf{X})^{-1}(\mathbf{X}'M_D\mathbf{y}) = (\sum_i \mathbf{X}_i'Q_T\mathbf{X}_i)^{-1}\sum_i \mathbf{X}_i'Q_t\mathbf{y}_i \tag{11.24}$$

Despues

$$\hat{c} = (D'D)^{-1}D'(\mathbf{y} - \mathbf{X}\hat{\beta}_{fe}) = \overline{\mathbf{y}} - \overline{\mathbf{X}}\hat{\beta}_{fe} \tag{11.25}$$

GLS con efectos fijos

Si no se cumple fe3 podemos todavía asumir fe4, donde no se especficica una forma de la varianza de $\mathbf{u}_i$, y así usar el estimador GLS dada la ecuación de estimación de efectos fijos. Problema: $\mathbb{E}[\tilde{\mathbf{u}}_i\tilde{\mathbf{u}}_i'|\tilde{\mathbf{X}}_i] = Q_T\Lambda Q_T$ es no invertible, rango $T-1$. Solución: Botar una ecuación del sistema (e.g., la ecuación correspondiente a T).

- **Paso 1**: Estimar $\hat{\beta}_{fe}$, y después tirar la observación temporal T, calcular

$$\hat{\Omega} = N^{-1}\sum_i \hat{\mathbf{u}}_i\hat{\mathbf{u}}_i' \tag{11.26}$$

  donde $\hat{\mathbf{u}}_i = \tilde{\mathbf{y}}_i - \tilde{\mathbf{x}}_i'\hat{\beta}_{fe}$.

- **Paso 2**: Calcular el estimador FEGLS

$$\hat{\beta}_{fegls} = (\sum_i \tilde{\mathbf{X}}_i'\hat{\Omega}^{-1}\tilde{\mathbf{X}}_i)^{-1}\sum_i \tilde{\mathbf{X}}_i'\hat{\Omega}^{-1}\tilde{\mathbf{y}}_i \tag{11.27}$$

  siempre y cuando $rank\{\mathbb{E}[\sum_i \tilde{\mathbf{X}}_i'\Omega^{-1}\tilde{\mathbf{X}}_i]\} = K$, $\hat{\beta}_{fegls}$ es consistente; además, si $\Lambda \neq \sigma_u^2 I_T$, generalmente el estimador FEGLS es más eficiente que FE. Para hacer inferencia usar

$$a\hat{var}(\hat{\beta}_{fegls}) = (\sum_i \tilde{\mathbf{X}}_i'\hat{\Omega}^{-1}\tilde{\mathbf{X}}_i)^{-1} \tag{11.28}$$

  Observación: note que $\Omega = \mathbb{E}[\tilde{\mathbf{u}}_{i,T-1}\tilde{\mathbf{u}}_{i,T-1}']$ que es la matriz $(T-1)\times(T-1)$ que resta despues de quitar la T-ésima columna y fila de $Q_T\Lambda Q_T$.

Efecto temporal no observable

Considere

$$y_{it} = \mathbf{x}_{it}'\beta + \mu + c_i + \delta_t + u_{it} \tag{11.29}$$

con la restricción $\sum_i c_i = \sum_t \delta_t = 0$. Podemos obtener un estimador para β mediante regresión particionada (Teorema de FWL), $\hat{\beta}$, luego

$$\hat{\mu} = \overline{\overline{y}} - \overline{\overline{\mathbf{x}}}'\hat{\beta}$$
$$\hat{c}_i = (\overline{y}_i - \overline{\overline{y}}) - (\overline{\mathbf{x}}_i - \overline{\overline{\mathbf{x}}})'\hat{\beta}$$
$$\hat{\delta}_t = (\overline{y}_t - \overline{\overline{y}}) - (\overline{\mathbf{x}}_t - \overline{\overline{\mathbf{x}}})'\hat{\beta}$$

Note que si $D_c = I_N \otimes \mathbf{1}_T$, $D_\delta = \mathbf{1}_N \otimes I_T$, y $\delta = (\delta_1, \delta_2, \ldots, \delta_T)'$ entonces

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{1}_{NT}\mu + D_c c + D_\delta\delta + \mathbf{u}$$

Aquí la estimación es restringida como ya mencionamos. El caso asímetrico (donde tomamos $t=1$ como base del efecto temporal), donde $D_\delta = \mathbf{1}_N \otimes [\mathbf{0}_{T-1\times 1}, I_{T-1}]'$ y $\delta = (\delta_2, \ldots, \delta_T)'$

$$\mathbf{y} = \mathbf{X}\beta + D_c c + D_\delta\delta + \mathbf{u}$$

Si $D = [D_c, D_\delta]$ y definidmos $M_D = I_{NT} - D(D'D)^{-1}D'$, entonces

$$\hat{\beta} = (\mathbf{X}'M_D\mathbf{X})^{-1}\mathbf{X}'M_D\mathbf{y} \tag{11.30}$$

No es muy complicado mostrar que $D_c'D_\delta = \mathbf{1}_N \otimes \mathbf{1}_{T-1}'$, $D_\delta'D_\delta = (N \otimes I_{T-1})$, la algebra matricial que sigue es más complicada. De este modo (usando el supuesto fe4),

$$\mathbb{V}[\hat{\beta}|\mathbf{X}] = \sigma_u^2(\mathbf{X}'M_D\mathbf{X})^{-1} \tag{11.31}$$

y

$$\hat{\sigma}_u^2 = \frac{\sum_i\sum_t \hat{u}_{it}^2}{(N-1)(T-1)-K} \tag{11.32}$$

11.4 Método de primeras diferencias

$$\Delta y_{it} = \Delta \mathbf{x}'_{it}\beta + \Delta u_{it}, \quad t = 2, \dots, T \tag{11.33}$$

Supuestos,

fd1 fe1.

fd2 $rank\{\sum_t \mathbb{E}[\Delta \mathbf{x}_{it}\Delta \mathbf{x}'_{it}]\} = K$.

fd3 $\mathbb{E}[\mathbf{e}_i\mathbf{e}'_i|\mathbf{x}_{it}, \dots, \mathbf{x}_{iT}, c_i] = \sigma_e^2 I_{T-1}$, donde $e_{it} = \Delta u_{it}$ para $t = 2, \dots, T$.

Sobre fd1, tenemos que $\mathbb{E}[\Delta u_{it}|\Delta \mathbf{x}_{i2}, \dots, \Delta \mathbf{x}_{iT}] = 0$. El estimador en primera diferencias es

$$\hat{\beta}_{fd} = (\sum_i \Delta \mathbf{X}'_i \Delta \mathbf{X}_i)^{-1} \sum_i \Delta \mathbf{X}'_i \Delta \mathbf{y}_i = (\sum_i \sum_{t=2}^{T} \Delta \mathbf{x}_{it}\Delta \mathbf{x}'_{it})^{-1} \sum_i \sum_{t=2}^{T} \Delta \mathbf{x}_{it} \Delta y_{it} \tag{11.34}$$

Teorema 11.1. *Si se cumple fd1-3, $\hat{\beta}_{fd}$ es el estimador más eficiente entre los que usan el supuesto fe1. Y el estimador de varianza asintótica es*

$$a\hat{var}(\hat{\beta}_{fd}) = \hat{\sigma}_e^2(\sum_i \Delta \mathbf{X}'_i \Delta \mathbf{X}_i)^{-1} \tag{11.35}$$

donde $\hat{\sigma}_e^2 = [N(T-1) - K]^{-1} \sum_i \sum_{t=2}^{T} \hat{e}_{it}^2$ *y* $\hat{e}_{it} = \Delta y_{it} - \Delta \mathbf{x}'_{it}\hat{\beta}_{fd}$.

Si el supuesto fd3 no se satisface es mejor usar un estimador robusto de la varianza asintótica,

$$a\hat{var}(\hat{\beta}_{fd}) = (\sum_i \Delta \mathbf{X}'_i \Delta \mathbf{X}_i)^{-1}(\sum_i \Delta \mathbf{X}'_i \hat{\mathbf{e}}_i \hat{\mathbf{e}}_i \Delta \mathbf{X}_i)(\sum_i \Delta \mathbf{X}'_i \Delta \mathbf{X}_i)^{-1} \tag{11.36}$$

Para garantizar la eficiencia de este estimador es preciso tener evidencia de ausencia de correlación serial en $\{e_{it}\}$. Podemos hacer la regresión: $\hat{e}_{it} = \hat{p}_1 \hat{e}_{i,t-1} + error_{it}$, con $t = 3, \dots, T$; $i = 1, \dots, N$, y luego testear $H_0 : \rho_1 =$ y $H_0 : \rho_1 = -.5$ [u_{it} no correlacionado].

11.5 Comparando estimadores

FE vs FD

Si $T = 2$, entonces $FE = FD$. Cuando $T \geq 3$, si existe evidencia de ausencia de correlación serial de u_{it} entonces estimar $\hat{\beta}_{fe}$; por otro lado, si la evidencia es a favor de que u_{it} es un random walk, usar $\hat{\beta}_{fd}$.

De ser las estimaciones muy distintas debemos desconfiar del supuesto fe1 y testear la exogeneidad estricta de los regresores (Test de Haussman). De no existir problemas, también podemos usar el estimador FEGLS o FDGLS, que son asintóticamente equivalentes.

FE vs RE

Si $\mathbf{x}_{it}$ no tiene mucha variabilidad es posible violemos los supuestos fe2 y fd2, entonces sería bueno estimar $\hat{\beta}_{re}$ incluso si desconfiamos de la ortoganilidad de c_i y $\mathbf{x}_{it}$. Si efectivamente estas variables son ortoganales, el estimador RE puede ser más preciso que FE. Es posible mostrar que, dada la estructura de efecto aleatorios, la ecuación de estimación de RE es

$$y_{it} - \lambda \overline{y}_i = (\mathbf{x}_{it} - \lambda \overline{\mathbf{x}}_i)'\beta + u_{it} - \lambda \overline{u}_i \tag{11.37}$$

donde $\lambda = 1 - \left[\frac{\sigma_u^2}{\sigma_u^2 + T\sigma_c^2}\right]^2 = 1 - \left[\frac{1}{1 + T\sigma_c^2/\sigma_u^2}\right]^2$. Si T es muy alto, entonces no debe sorprendernos que FE y RE son muy proximos (ya que, la correlación serial de $\tilde{u}_{it}$ es muy pequeña). El estimador RE esta basado en $\hat{\lambda}$, el cual depende de $\hat{\sigma}_u^2$ y $\hat{\sigma}_c^2$, por tanto, incluso cuando T es bajo $FE \approx RE$ si la varianza de c_i es muy alta relativa a la varianza de u_{it} (la precisión de RE disminuye).

Test de Haussman

Se usa como supuesto auxiliar re3. De esta forma, $H0 : \mathbb{E}[c_i|\mathbf{X}_i] = \mathbf{E}[c_i]$, bajo esta hipótesis FE y RE son consistentes y RE es más eficiente; y bajo $H1 : \mathbb{E}[c_i|\mathbf{X}_i] \neq \mathbf{E}[c_i]$, FE es consistente y RE es inconsiste. Estas aspectos caracterizan un test de Haussman: $\hat{\delta}_{re}$, estimador de los parámetros asociados a las M variables cambiantes en tiempo, y $\hat{\beta}_{fe}$ los estimadores de efectos fijos, luego

$$H = (\hat{\delta}_{fe} - \hat{\delta}_{re})'[a\hat{var}(\hat{\delta}_{fe}) - a\hat{var}(\hat{\delta}_{re})]^{-1}(\hat{\delta}_{fe} - \hat{\delta}_{re}) \xrightarrow{d} \chi^2_M \tag{11.38}$$

Digresión: Estimador "between" y "within"

Cualquiera sean los supuestos de estimación podemos expresar el modelo en tres formas

$$\begin{aligned} y_{it} &= \beta_0 + \mathbf{x}'_{it}\beta + v_{it} \\ \overline{y}_i &= \beta_0 + \overline{\mathbf{x}}'_i\beta + \overline{v}_{it} \\ y_{it} - \overline{y}_i &= (\mathbf{x}_{it} - \overline{\mathbf{x}}_i)'\beta + v_{it} - \overline{v}_{it} \end{aligned}$$

y estimar por OLS. Para eso definamos,

$$\begin{aligned} S^t_{xx} &= \sum_i \sum_t (\mathbf{x}_{it} - \overline{\overline{\mathbf{x}}})(\mathbf{x}_{it} - \overline{\overline{\mathbf{x}}})', \quad S^t_{xy} = \sum_i \sum_t (\mathbf{x}_{it} - \overline{\overline{\mathbf{x}}})(y_{it} - \overline{\overline{y}}) \\ S^b_{xx} &= \sum_i T(\overline{\mathbf{x}}_i - \overline{\overline{\mathbf{x}}})(\overline{\mathbf{x}}_i - \overline{\overline{\mathbf{x}}})', \quad S^t_{xy} = \sum_i T(\overline{\mathbf{x}}_i - \overline{\overline{\mathbf{x}}})(\overline{y}_i - \overline{\overline{y}}) \\ S^t_{xx} &= \sum_i \sum_t (\mathbf{x}_{it} - \overline{\mathbf{x}}_i)(\mathbf{x}_{it} - \overline{\mathbf{x}}_i)', \quad S^t_{xy} = \sum_i \sum_t (\mathbf{x}_{it} - \overline{\mathbf{x}}_i)(y_{it} - \overline{y}_i) \end{aligned}$$

No es díficil mostrar que

$$\begin{aligned} S^t_{xx} &= S^b_{xx} + S^w_{xx} \\ S^t_{xy} &= S^b_{xy} + S^w_{xy} \end{aligned}$$

y luego que

$$\hat{\beta}^t = \mathbf{F}^w\hat{\beta}^w + \mathbf{F}^b\hat{\beta}^b \tag{11.39}$$

donde $\mathbf{F}^w = [S^w_{xx} + S^b_{xx}]^{-1}S^w_{xx} = I - \mathbf{F}^b$.

References

Casella, G. and Berger, R. (2001). *Statistical Inference.* Duxbury Resource Center.

Gourieroux, C. and Monfort, A. (1995). *Statistics and Econometric Models.* Number v. 2 in Statistics and Econometric Models 2 volume set. Cambridge University Press.

Greene, W. H. (2003). *Econometric Analysis.* Prentice Hall, Upper Saddle River, NJ, 5. edition.

Hansen, B. (2017). *Econometrics.* manuscript.

Hansen, L. and Singleton, K. (1982). Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica*, 50(5):1269–86.

Lukacs, E. (1975). *Stochastic Convergence.* Probability and Mathematical Statistics - Academic Press. Academic Press.

Newey, W. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In Engle, R. F. and McFadden, D., editors, *Handbook of Econometrics*, volume 4, chapter 36, pages 2111–2245. Elsevier, 1 edition.

White, H. (2001). *Asymptotic Theory for Econometricians.* Economic theory, econometrics, and mathematical economics. Academic Press.

Wooldridge, J. (2010). *Econometric Analysis of Cross Section and Panel Data.* Econometric Analysis of Cross Section and Panel Data. MIT Press.