Estimadores M

Gabriel V. Montes-Rojas

Marco general

- Sea $m(\boldsymbol{x}, \boldsymbol{\theta})$ un modelo paramétrico para $E(y|\boldsymbol{x})$ (o cualquier momento de interés) donde
  - $m(.)$ es una función conocida de $\boldsymbol{x}$ y $\boldsymbol{\theta}$.
  - $\boldsymbol{x}$ es un vector de K variables explicativas.
  - $\boldsymbol{\theta}$ es un vector $P \times 1$ de parámetros. $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^P$.
- Un modelo está **correctamente especificado** para la media condicional $E(y|\boldsymbol{x})$ si para algún $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$,

$$E(y|\boldsymbol{x}) = m(\boldsymbol{x}, \boldsymbol{\theta}_0)$$

Marco general

- Tomemos $q(\boldsymbol{w}, \boldsymbol{\theta})$ como una función del vector aleatorio $\boldsymbol{w}$ y el vector de parámetros $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.
  - En general tenemos $\boldsymbol{w} = (y, \boldsymbol{x})$ con elementos típicos $\boldsymbol{w}_i$ de una muestra $\{\boldsymbol{w}_i : i = 1, 2, ..., N\}$.
  - $q(.)$ es una función conocida de $\boldsymbol{w}$ y $\boldsymbol{\theta}$.
  - $\boldsymbol{\theta}$ es un vector $P \times 1$. $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^P$.
- Un **estimador M** (M-estimator en inglés) de $\boldsymbol{\theta}_0$ resuelve el problema

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} N^{-1} \sum_{i=1}^{N} q(\boldsymbol{w}_i, \boldsymbol{\theta}),$$

$$\hat{\boldsymbol{\theta}} = arg \min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} N^{-1} \sum_{i=1}^{N} q(\boldsymbol{w}_i, \boldsymbol{\theta})$$

- Definamos la función $\boldsymbol{\theta} \mapsto M_N(\boldsymbol{\theta})$ donde $M_N(\boldsymbol{\theta}) \equiv N^{-1} \sum_{i=1}^{N} q(\boldsymbol{w}_i, \boldsymbol{\theta})$ y $\boldsymbol{\theta} \mapsto M_N(\boldsymbol{\theta})$ donde $M(\theta) \equiv E[q(\boldsymbol{w}, \boldsymbol{\theta})]$.
- Notemos que $\hat{\boldsymbol{\theta}}$ es una función de la muestra aleatoria $\{\boldsymbol{w}_i : i = 1, 2, ..., N\}$. Algunas veces usaremos la notación $\hat{\boldsymbol{\theta}}_N$ para enfatizar que depende de la muestra.

Marco general

- En general asumimos que

$$\theta_0 = arg \min_{\theta \in \Theta} E[q(\boldsymbol{w}, \theta)].$$

- $E[q(\boldsymbol{w}, \theta)]$ se puede ver en términos estadísticos como una **función de pérdida** (loss function en inglés).

  Ejemplos:
  - cuadrática: $(.)^2$, que es la base de los estimadores MCO
  - valor absoluto: $|.|$, que es la base de la regresión en la mediana.
  - máxima verosimilitud: $q = -\ell$ el log de la función de verosimilitud (log-likelihood).

Identificación

La identificación requiere que la solución a la minimización sea única.

$$E[q(\boldsymbol{w}, \boldsymbol{\theta}_0)] < E[q(\boldsymbol{w}, \boldsymbol{\theta})], \text{ para todo } \boldsymbol{\theta} \in \boldsymbol{\Theta}, \boldsymbol{\theta} \neq \boldsymbol{\theta}_0.$$

En otros contextos se usan otras definiciones de identificación. Por ejemplo, en modelos lineales de regresión se requiere que los parámetros se puedan obtener como funciones de esperanzas poblacionales.

Consistencia

Dado que $\hat{\theta}$ depende de la función $\theta \mapsto M_N(\theta)$ necesitamos "convergencia funcional".

Convergencia uniforme en probabilidad (uniform convergence in probability):

$$\max_{\theta \in \Theta} \left| N^{-1} \sum_{i=1}^{N} q(\boldsymbol{w}_i, \theta) - E[q(\boldsymbol{w}, \theta)] \right| \xrightarrow{p} 0.$$

Ley uniforme y débil de los grandes números

Ley uniforme y débil de los grandes números: *Supongamos que* $\boldsymbol{w}$ *es un vector aleatorio que toma valores en* $\mathcal{W} \subset \mathbb{R}^M$, $\boldsymbol{\Theta} \subset \mathbb{R}^P$, *y* $q : \mathcal{W} \times \boldsymbol{\Theta} \to \mathbb{R}$ *es una función real. Asumamos que*
(a) $\boldsymbol{\Theta}$ *es compacto;*
(b) para cada $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $q(., \boldsymbol{\theta})$ *es medible en términos de Borel (Borel measurable) sobre* $\mathcal{W}$*;*
(c) para cada $\boldsymbol{w} \in \mathcal{W}$, $q(\boldsymbol{w}, .)$ *es contínua en* $\boldsymbol{\Theta}$*; y*
(d) $|q(\boldsymbol{w}, \boldsymbol{\theta})| < b(\boldsymbol{w})$ *para todo* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, *donde b es una función no negativa de* $\mathcal{W}$ *tal que* $E[b(\boldsymbol{w})] < \infty$.
Entonces

$$\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left| N^{-1} \sum_{i=1}^{N} q(\boldsymbol{w}_i, \boldsymbol{\theta}) - E[q(\boldsymbol{w}, \boldsymbol{\theta})] \right| \xrightarrow{p} 0.$$

Consistencia

Consistencia de los estimatores M: *Bajo los supuestos de la convergencia uniforme en probabilidad, y asumiendo identificación, entonces el estimador M* $\hat{\theta}$ *(o sea,* $\hat{\theta} = arg\,\min_{\theta \in \Theta} N^{-1} \sum_{i=1}^{N} q(\boldsymbol{w}_i, \theta)$*) satisface* $\hat{\theta} \xrightarrow{p} \theta$.

Supongamos que $\hat{\theta} \xrightarrow{p} \theta$*, y asumamos que* $r(\boldsymbol{w}, \theta)$ *satisface los mismos supuestos de* $q(\boldsymbol{w}, \theta)$*. Entonces,* $N^{-1} \sum_{i=1}^{N} r(\boldsymbol{w}, \hat{\theta}) \xrightarrow{p} E[r(\boldsymbol{w}, \theta_0)]$.

Ejemplo MCO

- En los modelos MCO $m(\boldsymbol{x}, \boldsymbol{\beta}) = E(y|\boldsymbol{x}) = \boldsymbol{x\beta}$ y $q(y, \boldsymbol{x}, \boldsymbol{\beta}) = [y - m(\boldsymbol{x}, \boldsymbol{\beta})]^2$.
- Para la identificación,

$$\begin{aligned} E[q(y, \boldsymbol{x}, \boldsymbol{\beta}_0)] &= E[y^2 + m(\boldsymbol{x}, \boldsymbol{\beta}_0)^2 - 2ym(\boldsymbol{x}, \boldsymbol{\beta}_0)^2] \\ &= E[(\boldsymbol{x\beta}_0)^2 + u^2 + (\boldsymbol{x\beta}_0)^2 - 2(\boldsymbol{x\beta}_0)^2] = \sigma^2 \end{aligned}$$

$$\begin{aligned} E[q(y, \boldsymbol{x}, \boldsymbol{\beta})]_{\boldsymbol{\beta} \neq \boldsymbol{\beta}_0} &= E[y^2 + m(\boldsymbol{x}, \boldsymbol{\beta})^2 - 2ym(\boldsymbol{x}, \boldsymbol{\beta})^2] \\ &= E[(\boldsymbol{x\beta}_0)^2 + u^2 + (\boldsymbol{x\beta})^2 - 2(\boldsymbol{x\beta}_0)(\boldsymbol{x\beta})] \\ &= \sigma^2 + E(\boldsymbol{x\beta}_0 - \boldsymbol{x\beta})^2 \end{aligned}$$

Entonces,

$$E[q(\boldsymbol{x}, \boldsymbol{\beta}_0)] < E[q(\boldsymbol{x}, \boldsymbol{\beta})] \text{ para todo } \boldsymbol{\beta} \in \boldsymbol{B}, \boldsymbol{\beta} \neq \boldsymbol{\beta}_0.$$

Score de la función objetivo

Si $q(\boldsymbol{w}, .)$ es contínuamente diferenciable en el interior de $\boldsymbol{\Theta}$, entonces con probabilidad tendiendo a uno, $\hat{\boldsymbol{\theta}}$ resuelve la condición de primer orden

$$\sum_{i=1}^{N} \boldsymbol{s}(\boldsymbol{w}_i, \hat{\boldsymbol{\theta}}) = \mathbf{0},$$

donde

$$\boldsymbol{s}(\boldsymbol{w}, \boldsymbol{\theta})' = \nabla_{\theta} q(\boldsymbol{w}, \boldsymbol{\theta}) = \left(\frac{\partial q(\boldsymbol{w}, \boldsymbol{\theta})}{\partial \theta_1}, \frac{\partial q(\boldsymbol{w}, \boldsymbol{\theta})}{\partial \theta_2}, ..., \frac{\partial q(\boldsymbol{w}, \boldsymbol{\theta})}{\partial \theta_P} \right)'$$

es el **score de la función objetivo**, un vector $P \times 1$.
Esta condición se satisface en general con igualdad (pero no siempre, ver regresión por cuantiles). Para MCO siempre es con igualdad.

Score de la función objetivo

- Si $E[q(\boldsymbol{w}, \theta)]$ es contínuamente diferenciable en el interior de $\boldsymbol{\Theta}$, entonces una condición para la miniminzación es

$$\nabla_\theta E[q(\boldsymbol{w}, \theta)]_{\theta=\theta_0} = \mathbf{0}.$$

- Si las derivadas y las esperanzas se pueden intercambiar (Teorema de Convergencia Dominada), entonces

$$E[\nabla_\theta q(\boldsymbol{w}, \theta_0)] = E[\boldsymbol{s}(\boldsymbol{w}, \theta_0)] = \mathbf{0}.$$

- Además si $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$, entonces bajo las condiciones estándar de regularidad

$$N^{-1} \sum_{i=1}^{N} \boldsymbol{s}(\boldsymbol{w}_i, \hat{\boldsymbol{\theta}}) \xrightarrow{p} E[\boldsymbol{s}(\boldsymbol{w}, \theta_0)].$$

Score de la función objetivo

Un estimador que se define por

$$\hat{\theta} = arg \min_{\theta \in \Theta} \| \sum_{i=1}^{N} \boldsymbol{s}(\boldsymbol{w}_i, \theta) \|_M$$

se llama estimador Z (del inglés Z-estimator, por "zero"). Aquí $\| \|_M$ *es una norma pre-especificada (en general la normal euclidiana).*
Para este caso deberíamos usar una condición de identificación diferente:

$$E[\boldsymbol{s}(\boldsymbol{w}, \theta)] = \mathbf{0} \iff \theta = \theta_0.$$

Hessiano de la función objetivo

- Si $q(\boldsymbol{w}, .)$ es diferenciable al menos dos veces en el interior de $\boldsymbol{\Theta}$, entonces definamos el hessiano de la función objetivo como,

$$\boldsymbol{H}(\boldsymbol{w}, \boldsymbol{\theta}) = \nabla^2_{\theta} q(\boldsymbol{w}, \boldsymbol{\theta}) = \frac{\partial^2 q(\boldsymbol{w}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \frac{\partial \boldsymbol{s}(\boldsymbol{w}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}$$

- Por el teorema del valor medio aplicado a $q(\boldsymbol{w}, \boldsymbol{\theta})$ en el valor $\boldsymbol{\theta}_0$,

$$\sum_{i=1}^{N} \boldsymbol{s}(\boldsymbol{w}_i, \hat{\boldsymbol{\theta}}) = \sum_{i=1}^{N} \boldsymbol{s}(\boldsymbol{w}_i, \boldsymbol{\theta}_0) + \left(\sum_{i=1}^{N} \ddot{\boldsymbol{H}}_i \right) (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0),$$

  donde $\ddot{\boldsymbol{H}}_i \equiv \boldsymbol{H}(\boldsymbol{w}_i, \tilde{\boldsymbol{\theta}})$ y $\tilde{\boldsymbol{\theta}}$ es un vector donde cada elemento está en el segmento entre $\hat{\boldsymbol{\theta}}$ y $\boldsymbol{\theta}_0$.
- Notar que la solución implica $\sum_{i=1}^{N} \boldsymbol{s}(\boldsymbol{w}_i, \hat{\boldsymbol{\theta}}) = \boldsymbol{0}$.
- Por otro lado notar que $N^{-1} \sum_{i=1}^{N} \boldsymbol{s}(\boldsymbol{w}_i, \hat{\boldsymbol{\theta}})$ es un promedio, que converje en probabilidad a $E[\boldsymbol{s}(\boldsymbol{w}, \boldsymbol{\theta}_0)]$ dado que $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$ cuando $N \to \infty$, usando LGN.

Expansiones asintóticas

- Ahora con $N^{-1}\sum_{i=1}^{N} \boldsymbol{s}(\boldsymbol{w}_i, \theta_0) \xrightarrow{p} E[\boldsymbol{s}(\boldsymbol{w}, \theta_0)] = \boldsymbol{0}$ también podemos aplicar la LGN.
- También vamos a usar el TCL, $N^{-1/2}\sum_{i=1}^{N} \boldsymbol{s}(\boldsymbol{w}_i, \theta_0) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{B}_0)$, donde $\boldsymbol{B}_0$ se define más abajo, la varianza de $\boldsymbol{s}(\boldsymbol{w}, \theta_0)$.
- Usando argumentos similares para el score $\hat{\theta} \xrightarrow{p} \theta_0$, entonces por teoría asintótica (LGN y teorema de Slutsky)

$$N^{-1}\sum_{i=1}^{N} \boldsymbol{H}(\boldsymbol{w}_i, \tilde{\theta}) \xrightarrow{p} E[\boldsymbol{H}(\boldsymbol{w}, \theta_0)].$$

- Para estimadores M, si θ_0 está identificado entonces $E[\boldsymbol{H}(\boldsymbol{w}, \theta_0)]$ es definida positiva.

Expansiones asintóticas

- La expansión en el valor real de los parámetros θ_0 juega un rol fundamental en el análisis asintótico.
- Tomemos la igualdad anterior y multipliquemos ambos lados de la igualdad por $1/\sqrt{N}$,

$$0 = 1/\sqrt{N} \sum_{i=1}^{N} \boldsymbol{s}(\boldsymbol{w}_i, \theta_0) + 1/N \left(\sum_{i=1}^{N} \ddot{\boldsymbol{H}}_i \right) \sqrt{N}(\hat{\theta} - \theta_0).$$

- Reordenando los términos obtenemos la expansión asintótica de primer orden de $\hat{\theta}$:

$$\sqrt{N}(\hat{\theta} - \theta_0) = \left(N^{-1} \sum_{i=1}^{N} \ddot{\boldsymbol{H}}_i \right)^{-1} \left[-N^{-1/2} \sum_{i=1}^{N} \boldsymbol{s}_i(\theta_0) \right]$$

donde $\boldsymbol{s}_i(\theta_0) = \boldsymbol{s}(\boldsymbol{w}_i, \theta_0)$.

Expansiones asintóticas

- Además, usando $\boldsymbol{A}_0 \equiv E[\boldsymbol{H}(\boldsymbol{w}, \theta_0)]$,

$$\begin{aligned}
\sqrt{N}(\hat{\theta} - \theta_0) &= \boldsymbol{A}_0^{-1} \left[-N^{-1/2} \sum_{i=1}^{N} \boldsymbol{s}_i(\theta_0) \right] \\
&\quad + \left(\left(N^{-1} \sum_{i=1}^{N} \ddot{\boldsymbol{H}}_i \right)^{-1} - \boldsymbol{A}_0^{-1} \right) \left[-N^{-1/2} \sum_{i=1}^{N} \boldsymbol{s}_i(\theta_0) \right] \\
&= \boldsymbol{A}_0^{-1} \left[-N^{-1/2} \sum_{i=1}^{N} \boldsymbol{s}_i(\theta_0) \right] + o_p(1) \cdot O_p(1) \\
&= \boldsymbol{A}_0^{-1} \left[-N^{-1/2} \sum_{i=1}^{N} \boldsymbol{s}_i(\theta_0) \right] + o_p(1)
\end{aligned}$$

- Definamos la función de influencia de $\hat{\theta}$ como $\boldsymbol{e}(\boldsymbol{w}_i, \theta_0) \equiv \boldsymbol{e}_i(\theta_0) \equiv \boldsymbol{A}_0^{-1} \boldsymbol{s}_i(\theta_0)$. Esta mide la "influencia" de cada observación particular i en el estimador.

Normalidad asintótica

Normalidad asintótica: *Además de los supuestos necesarios para consistencia, asumamos que*
(a) $\theta_0 \in int\Theta$;
(b) $\boldsymbol{s}(\boldsymbol{w}, .)$ *es continuamente diferenciable en el interior de* Θ *para todo* $\boldsymbol{w} \in \mathcal{W}$;
(c) Cada elemento en $\boldsymbol{H}(\boldsymbol{w}, \theta)$ *esta acotado en valor absoluto por una función* $b(\boldsymbol{w})$, *donde* $E[b(\boldsymbol{w})] < \infty$;
(d) $\boldsymbol{A}_0 \equiv E[\boldsymbol{H}(\boldsymbol{w}, \theta_0)]$ *es definida positiva;*
(e) $E[\boldsymbol{s}(\boldsymbol{w}, \theta_0)] = \boldsymbol{0}$; *y*
(f) Cada elemento en $\boldsymbol{s}(\boldsymbol{w}, \theta_0)$ *tiene momento segundo finito (varianza finita).*

Entonces,

$$\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{A}_0^{-1}\boldsymbol{B}_0\boldsymbol{A}_0^{-1}),$$

donde $\boldsymbol{B}_0 \equiv E[\boldsymbol{s}(\boldsymbol{w}, \theta_0)\boldsymbol{s}(\boldsymbol{w}, \theta_0)'] = Var[\boldsymbol{s}(\boldsymbol{w}, \theta_0)]$. *Así,*

$$Avar(\hat{\theta}) = \boldsymbol{A}_0^{-1}\boldsymbol{B}_0\boldsymbol{A}_0^{-1}/N.$$

Notemos la **fórmula sandwich**: $\boldsymbol{A}_0^{-1}\boldsymbol{B}_0\boldsymbol{A}_0^{-1}$. *Esta aparecerá muchas veces en el futuro.*

Ejemplo MCO

- Resolver este caso para MCO.
- Notar que

$$\boldsymbol{s}(\boldsymbol{w}, \boldsymbol{\beta}) = \nabla_\theta q(\boldsymbol{w}, \boldsymbol{\beta}) = -2\boldsymbol{x}'(y - \boldsymbol{x}\boldsymbol{\beta}),$$

  que nos da la clásica condición $E[\boldsymbol{x}'u] = \boldsymbol{0}$.
- El hessiano es

$$\boldsymbol{A}_0 = E[\boldsymbol{H}(\boldsymbol{w}, \boldsymbol{\beta}_0)] = 2E[\boldsymbol{x}'\boldsymbol{x}].$$

- Si asumimos homoscedasticidad ($E[\boldsymbol{u}\boldsymbol{u}'|\boldsymbol{x}] = \sigma^2\boldsymbol{I}$),

$$\boldsymbol{B}_0 = E[\boldsymbol{s}(\boldsymbol{w}, \boldsymbol{\beta}_0)\boldsymbol{s}(\boldsymbol{w}, \boldsymbol{\beta}_0)'] = 4\sigma^2 E[\boldsymbol{x}'\boldsymbol{x}]$$

- Entonces,

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\boldsymbol{0}, \sigma^2 E[\boldsymbol{x}'\boldsymbol{x}]^{-1}).$$

Ejemplo MCO

- Si tenemos heteroscedasticidad:

$$\boldsymbol{B}_0 = E[\boldsymbol{s}(\boldsymbol{w}, \beta_0)\boldsymbol{s}(\boldsymbol{w}, \beta_0)'] = 4E[\boldsymbol{x}'u'u\boldsymbol{x}]$$

  y entonces $\boldsymbol{B}_0 \neq \boldsymbol{A}_0$.
- En este caso,

$$\sqrt{N}(\hat{\beta} - \beta_0) \xrightarrow{d} N(\boldsymbol{0}, E[\boldsymbol{x}'\boldsymbol{x}]^{-1}E[\boldsymbol{x}'u'u\boldsymbol{x}]E[\boldsymbol{x}'\boldsymbol{x}]^{-1}).$$

Método delta

- La expansión asintótica se puede usar para funciones de parámetros. Consideremos la función $\theta \mapsto g(\theta)$. Nos gustaría saber la distribución de $\sqrt{N}(g(\hat{\theta}) - g(\theta_0))$.
- Asumamos que
  - $\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{V}_{\hat{\theta}})$,
  - $\theta \mapsto g(\theta)$ es diferenciable en θ_0 (o sea, la primera derivada existe y es contínua en θ_0).

  Entonces,

$$\sqrt{N}(g(\hat{\theta}) - g(\theta_0)) \xrightarrow{d} N\left(\mathbf{0}, \nabla_\theta g(\theta_0) \boldsymbol{V}_{\hat{\theta}} \nabla_\theta g(\theta_0)'\right)$$

- Por el teorema de Slutsky:

$$\nabla_\theta g(\hat{\theta}) \hat{\boldsymbol{V}}_{\hat{\theta}} \nabla_\theta g(\hat{\theta})' \xrightarrow{p} \nabla_\theta g(\theta_0) \boldsymbol{V}_{\hat{\theta}} \nabla_\theta g(\theta_0)',$$

donde $\hat{\boldsymbol{V}}_{\hat{\theta}}$ es un estimador consistente de $\boldsymbol{V}_{\hat{\theta}}$.

Ejemplo: Elasticidades de largo plazo

- Consideremos el modelo $y_t = \beta_0 + x_t\beta_1 + y_{t-1}\alpha_1 + u_t$ donde (y, x) estan en logaritmos.
- La elasticidad de corto plazo viene dado por del efecto de x en y viene dado por β_1.
- La elasticidad de largo plazo viene dado por del efecto de x en y viene dado por $\frac{\beta_1}{1-\alpha_1}$.
- Obtener la distribución de $\frac{\hat{\beta}_1}{1-\hat{\alpha}_1}$.

Método de los momentos generalizados

- Supongamos el modelo de regresión lineal $y = \boldsymbol{x}\boldsymbol{\beta} + u$. En este modelo podemos tener que $\boldsymbol{x}$ es endógena (o algunas xs), por lo que β no puede ser estimado con MCO. Supongamos que $\boldsymbol{x}$ es un vector $1 \times K$ y β es $K \times 1$. $\boldsymbol{X}$ es la matriz de datos $N \times K$. $\boldsymbol{y}$ y $\boldsymbol{u}$ son vectores $N \times 1$.
- Supongamos un conjunto de J variables exógenas, $\boldsymbol{z}$ que dan lugar a las condiciones de momento, $E(u|\boldsymbol{z}) = \mathbf{0}_J$.
- Si $J = K$ el modelo está exactamente identificado. Ejemplo:
  - $E(u|\boldsymbol{z}) = \mathbf{0}_J$ da lugar a las condiciones de momento $E(\boldsymbol{z}'u) = \mathbf{0}_J$ (probar).
  - Entonces podemos plantear las siguientes condiciones empíricas $\boldsymbol{X}'\boldsymbol{u}(\hat{\boldsymbol{\beta}}) = \boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \mathbf{0}_J$. La solución es MCO.
- Si $J > K$ el modelo está sobre-identificado (*over-identified*). Notemos que en este caso la solución no es única. Para cada combinación de K condiciones de momento tenemos un estimador diferente...

Método de los momentos generalizados

- Tomemos $\boldsymbol{g}(\boldsymbol{w}, \boldsymbol{\theta})$ un vector $J \times 1$, $\boldsymbol{g}(\boldsymbol{w}, \boldsymbol{\theta}) = (g_1(\boldsymbol{w}, \boldsymbol{\theta}), \ldots, g_J(\boldsymbol{w}, \boldsymbol{\theta}))'$ que mapea $\mathcal{W} \times \boldsymbol{\Theta} \mapsto \boldsymbol{g}(\boldsymbol{w}, \boldsymbol{\theta})$, y también $\boldsymbol{g}_i(\boldsymbol{\theta}) \equiv \boldsymbol{g}(\boldsymbol{w}_i, \boldsymbol{\theta})$.
- Asumir que $E[\boldsymbol{g}(\boldsymbol{w}_i, \boldsymbol{\theta}_0)] = \boldsymbol{0}$.
- Una condición necesaria mínima para la identificación es $\boldsymbol{\theta}_0$ is $J \geq P$ (que haya al menos tantos momentos como parámetros a estimar).
- Cuando $J = P$, entonces $\boldsymbol{\theta}_0$ se estima por la contraparte muestral $N^{-1} \sum_{i=1}^{N} \boldsymbol{g}(\boldsymbol{w}, \boldsymbol{\theta}_0) = 0$. Este es el caso de MCO donde $\boldsymbol{g}(\boldsymbol{w}, \boldsymbol{\theta}) = \boldsymbol{x}'(y - \boldsymbol{x}\boldsymbol{\beta})$, el modelo lineal que teníamos antes.
- Cuando $J > P$, entonces el Método de los momentos generalizados (GMM en inglés) usa una métrica cuadrada para minimizar $\sum_{i=1}^{N} \boldsymbol{g}(\boldsymbol{w}, \boldsymbol{\theta})$:

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left[\sum_{i=1}^{N} \boldsymbol{g}(\boldsymbol{w}_i, \boldsymbol{\theta}) \right]' \hat{\boldsymbol{\Xi}} \left[\sum_{i=1}^{N} \boldsymbol{g}(\boldsymbol{w}_i, \boldsymbol{\theta}) \right]$$

  donde $\hat{\boldsymbol{\Xi}}$ es una matriz de pesos $J \times J$ simétrica y semidefinida positiva.
- $\hat{\boldsymbol{\Xi}}$ tiene un rol central en el modelo GMM. En particular se encarga de "pesar" cada condición de momento g_j. Ejemplo: Supongamos que g_1 tiene más varianza que g_2. Entonces deberíamos usar más información de g_2 que de g_1.

Método de los momentos generalizados

- Definamos $Q_N(\theta) = \left[\sum_{i=1}^N \boldsymbol{g}(\boldsymbol{w}_i, \theta)\right]' \hat{\Xi} \left[\sum_{i=1}^N \boldsymbol{g}(\boldsymbol{w}_i, \theta)\right]$. Bajo condiciones de regularidad estándares $Q_N(\theta)$ converge uniformemente a $\{E[\boldsymbol{g}_i(\theta)]\}' \Xi_0 \{E[\boldsymbol{g}_i(\theta)]\}$ donde $\hat{\Xi} \xrightarrow{p} \Xi_0$.
- Entonces el estimador GMM es

$$\hat{\theta}_{GMM} = argmin_{\theta \in \Theta} \left[\sum_{i=1}^N \boldsymbol{g}(\boldsymbol{w}_i, \theta)\right]' \hat{\Xi} \left[\sum_{i=1}^N \boldsymbol{g}(\boldsymbol{w}_i, \theta)\right].$$

Método de los momentos generalizados

- Bajo el supuesto que $\boldsymbol{g}(\boldsymbol{w}, .)$ es contínuamente diferenciable en $int(\boldsymbol{\Theta})$, $\theta_0 \in \boldsymbol{\Theta}$, entonces la condición de primer orden para $\hat{\theta}$ es

$$\left[\sum_{i=1}^{N} \nabla_\theta \boldsymbol{g}_i(\hat{\theta})\right]' \hat{\boldsymbol{\Xi}} \left[\sum_{i=1}^{N} \boldsymbol{g}_i(\hat{\theta})\right] \equiv \mathbf{0}.$$

- Definamos
  - matriz $J \times P$ de rango P, $\boldsymbol{G}_0 \equiv E[\nabla_\theta \boldsymbol{g}_i(\hat{\theta})]$;
  - matriz $P \times P$ de rango P, $\boldsymbol{A}_0 \equiv \boldsymbol{G}_0' \boldsymbol{\Xi}_0 \boldsymbol{G}_0$;
  - matriz $P \times P$ de rango P, $\boldsymbol{B}_0 \equiv \boldsymbol{G}_0' \boldsymbol{\Xi}_0 \boldsymbol{\Lambda}_0 \boldsymbol{\Xi}_0 \boldsymbol{G}_0$;
  - matriz $J \times J$ de rango J, $\boldsymbol{\Lambda}_0 \equiv E[\boldsymbol{g}_i(\theta_0)\boldsymbol{g}_i(\theta_0)'] = Var[\boldsymbol{g}_i(\theta_0)]$.
- Con manipulaciones algebraicas llegamos a

$$\mathbf{0} = \boldsymbol{G}_0' \boldsymbol{\Xi}_0 N^{-1/2} \sum_{i=1}^{N} \boldsymbol{g}_i(\theta_0) + \boldsymbol{A}_0 \sqrt{N}(\hat{\theta} - \theta_0) + o_p(1)$$

- Entonces,

$$\sqrt{N}(\hat{\theta} - \theta_0) = -\boldsymbol{A}_0^{-1} \boldsymbol{G}_0' \boldsymbol{\Xi}_0 N^{-1/2} \sum_{i=1}^{N} \boldsymbol{g}_i(\theta_0) + o_p(1) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{A}_0^{-1} \boldsymbol{B}_0 \boldsymbol{A}_0^{-1})$$

Método de los momentos generalizados

- El estimador GMM más eficiente requiere $\Xi_0 = \Lambda_0^{-1}$ porque

$$(G_0'\Xi_0 G_0)^{-1}(G_0'\Xi_0\Lambda_0\Xi_0 G_0)(G_0'\Xi_0 G_0)^{-1} - (G_0'\Lambda_0^{-1}G_0)^{-1},$$

  o sea, la diferencia entre dos estimadores GMM es una matriz semidefinida positiva.
- Notemos que no podemos tener Λ_0 antes de estimar $\hat{\theta}$. Sin embargo, solo necesitamos un estimador consistente, $\tilde{\theta}$, para estimar primero Λ_0, y luego se construye el estimador GMM.

Método de los momentos generalizados

- Siguiendo con el modelo lineal sobre-identificado, podemos plantear el problema como encontrar un cero a $\boldsymbol{Z}'\boldsymbol{u}(\boldsymbol{\beta}) = \boldsymbol{0}_J$ donde $\boldsymbol{u}(\boldsymbol{\beta}) = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$ es un vector $N \times 1$. Como $J > P = K$ tenemos más ecuaciones que parámetros, entonces no hay solución única.
- Para transformar esto en un problema de solución única, lo transformamos en una forma cuadrática. Para cualquier matriz Ξ positiva semi-definida $J \times J$,

$$\hat{\boldsymbol{\beta}}^{\Xi}_{GMM} = arg\min_{\boldsymbol{\beta}} (\boldsymbol{Z}'u(\boldsymbol{\beta}))'\Xi(\boldsymbol{Z}'u(\boldsymbol{\beta})) = arg\min_{\boldsymbol{\beta}} \left(\boldsymbol{Z}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right)' \Xi \left(\boldsymbol{Z}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right)$$

$$= arg\min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{N} \boldsymbol{z}_i' u_i(\boldsymbol{\beta})\right]' \Xi \left[\sum_{i=1}^{N} \boldsymbol{z}_i' u_i(\boldsymbol{\beta})\right]$$

De esta manera J condiciones de momento se transformaron en una función cuadrática a ser minimizada, que tienen solución única.

- También lo podemos escribir en función de lo anterior como

$$\hat{\boldsymbol{\beta}}_{GMM} = arg\min_{\boldsymbol{\beta}} Q_N(\boldsymbol{\beta})$$

Método de los momentos generalizados

- Tomemos ahora la primera derivada de esta función y resolvamos para que evaluada en $\hat{\beta}$ sea 0:

$$\frac{\partial Q_N(\hat{\beta})}{\partial \beta} = \mathbf{0}_J = \frac{\partial d}{d\hat{\boldsymbol{u}}} \left(\hat{\boldsymbol{u}}'(\boldsymbol{Z}\Xi\boldsymbol{Z}')\hat{\boldsymbol{u}}\right) \frac{\partial d\left(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}\right)}{d\hat{\beta}} = \frac{2}{N}\hat{\boldsymbol{u}}'\boldsymbol{Z}\Xi\boldsymbol{Z}'(-\boldsymbol{X}),$$

donde $\hat{\boldsymbol{u}} = \boldsymbol{y} - \boldsymbol{X}\hat{\beta}$ y usamos $\partial \boldsymbol{A}\boldsymbol{b}/\partial \boldsymbol{b} = \boldsymbol{A}$ y $\partial(\boldsymbol{b}'\boldsymbol{A}\boldsymbol{b})/\partial \boldsymbol{b} = 2\boldsymbol{b}'\boldsymbol{A}$, donde $\boldsymbol{b}$ es un vector columna y $\boldsymbol{A}$ es una matriz simétrica. Entonces tenemos,

$$\hat{\beta}^{\Xi}_{GMM} = (\boldsymbol{X}'\boldsymbol{Z}\Xi\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}\Xi\boldsymbol{Z}'\boldsymbol{y},$$

donde enfatizamos que depende de Ξ.

Método de los momentos generalizados

- Si las condiciones de momento son válidas, el estimador es consistente:

$$\hat{\boldsymbol{\beta}}^{\Xi}_{GMM} = (\boldsymbol{X}'\boldsymbol{Z}\Xi\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}\Xi\boldsymbol{Z}'\boldsymbol{y} = (\boldsymbol{X}'\boldsymbol{Z}\Xi\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}\Xi\boldsymbol{Z}'(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{u})$$

$$= \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{Z}\Xi\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}\Xi\boldsymbol{Z}'\boldsymbol{u} \xrightarrow{p} \boldsymbol{\beta}, N \to \infty$$

  usando $\frac{1}{N}\boldsymbol{Z}'\boldsymbol{u} \xrightarrow{p} \boldsymbol{0}_J$ y que $\frac{1}{N}\boldsymbol{Z}'\boldsymbol{X}$ converge a una matríz no nula cuando $N \to \infty$.
- Este estimador no es factible porque no tenemos Ξ. Para ello necesitamos calcular la varianza asintótica...

Método de los momentos generalizados

- Calculemos la varianza asintótica,

$$Var(\hat{\boldsymbol{\beta}}^{\Xi}_{GMM}) = plim_{N\to\infty}(\boldsymbol{X}'\boldsymbol{Z}\Xi\boldsymbol{Z}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Z}\Xi\boldsymbol{Z}'\boldsymbol{u}\boldsymbol{u}'\boldsymbol{Z}\Xi\boldsymbol{Z}'\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{Z}\Xi\boldsymbol{Z}'\boldsymbol{X})^{-1}.$$

- Si hacemos $\hat{\Xi} = \frac{1}{N}\boldsymbol{Z}'\boldsymbol{u} \xrightarrow{p} E[\boldsymbol{z}_i' u_i u_i' \boldsymbol{u}_i] = (\Xi^*)^{-1}$, entonces tenemos la varianza óptima:

$$Var(\hat{\boldsymbol{\beta}}^{*}_{GMM}) = plim_{N\to\infty}(\boldsymbol{X}'\boldsymbol{Z}\Xi^*\boldsymbol{Z}'\boldsymbol{X})^{-1}.$$

Prueba: Para ver que este es el estimador más eficiente calculemos $Var(\hat{\boldsymbol{\beta}}^{*}_{GMM}) - Var(\hat{\boldsymbol{\beta}}^{\Xi}_{GMM}) = plim_{N\to\infty}(\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{Z}'\boldsymbol{u}\boldsymbol{u}'\boldsymbol{Z}\boldsymbol{Z}'\boldsymbol{X})^{-1} - (\boldsymbol{X}'\boldsymbol{Z}\Xi\boldsymbol{Z}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Z}\Xi\boldsymbol{Z}'\boldsymbol{u}\boldsymbol{u}'\boldsymbol{Z}\Xi\boldsymbol{Z}'\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{Z}\Xi\boldsymbol{Z}'\boldsymbol{X})^{-1} = (\Sigma_{XZ}\Xi^*\Sigma_{XZ}')^{-1} - (\Sigma_{XZ}\Xi\Sigma_{XZ}')^{-1}(\Sigma_{XZ}\Xi\Xi^*\Xi\Sigma_{XZ}')(\Sigma_{XZ}\Xi\Sigma_{XZ}')^{-1}$ donde $\Sigma_{XZ} = \frac{1}{N}plim_{N\to\infty}\boldsymbol{X}'\boldsymbol{Z}$. Esto puede escribirse como una forma cuadrática $= -\boldsymbol{D}[\boldsymbol{I} - \boldsymbol{H}(\boldsymbol{H}'\boldsymbol{H})^{-1}\boldsymbol{H}']\boldsymbol{D}$ donde $\boldsymbol{D} = (\Sigma_{XZ}\Xi\Sigma_{XZ}')^{-1}(\Sigma_{XZ}\Xi(\Xi^*)^{1/2})$ y $\boldsymbol{H} = (\Xi^*)^{-1/2}\Sigma_{XZ}$.

- El estimador GMM se tiene que armar con ponderaciones que son inversamente proporcionales a la varianza de las condiciones de momento.
- Estos modelos reciben el nombre de **mínimos cuadrados generalizados** (MCG).

Método de los momentos generalizados

- Supongamos que $u_i \sim i.i.d.(0, \sigma^2)$ y que $\boldsymbol{Z} = \boldsymbol{X}$, es decir son exógenas. Entonces tenemos los supuestos de Gauss-Markov. En este caso, $\Xi^{-1} = Var(\boldsymbol{Z}'\boldsymbol{u}) = E(\boldsymbol{Z}'\boldsymbol{u}\boldsymbol{u}'Z) = E[\boldsymbol{Z}E(\boldsymbol{u}\boldsymbol{u}'|\boldsymbol{Z})\boldsymbol{Z}'] = \sigma^2 E[\boldsymbol{Z}\boldsymbol{Z}'] = \sigma^2 E[\boldsymbol{X}\boldsymbol{X}']$, entonces $\hat{\boldsymbol{\beta}}^*_{GMM} = \hat{\boldsymbol{\beta}}_{MCO}$.
- Supongamos que $u_i|x_i \sim i.i.d.(0, h(x_i))$ y que $\boldsymbol{Z} = \boldsymbol{X}$. Entonces $\Xi^{-1} = Var(\boldsymbol{Z}'\boldsymbol{u})$, que no se puede simplificar.
- Supongamos que $E(\boldsymbol{u}\boldsymbol{u}'|\boldsymbol{X}) = \Omega$ y que $\boldsymbol{Z} = \boldsymbol{X}$. Modelo de efectos aleatorios, clusters, etc. (ver datos en panel)
- Supongamos que $\boldsymbol{Z} \neq \boldsymbol{X}$. Variables instrumentales.

Referencias

Estas notas se basan en

- *Capítulos 12, 13 y 14 de Wooldridge.*
- *Newey, W.K., y McFadden, D. (1994), "Large Sample Estimation and Hypothesis Testing," en Handbook of Econometrics, Volumen 4, ed. R.F. Engle y D. McFadden. Amsterdam: North Holland, 2111–2245.*
- *Van der Vaart, A.W. (1998), Asymptotic Statistics. Cambridge University Press.*