Estimación por máxima verosimilitud

Curso de Series Cronológicas

Silvia Rodríguez Collazo

Facultad de Ciencias Económicas y de Administración

En las sesiones previas trabajamos con el supuesto que los parámetros poblacionales $c, \theta_1, \ldots, \theta_p, \phi_1 \ldots, \phi_p$ y σ^2 eran conocidos. Exploraremos como estimar los valores de esos parámetros en base a las observaciones de Y_t. Vamos a profundizar en cómo opera la estimación basada en maximizar la verosimilitud.
Sea $\theta = (c, \phi_1, \phi_2, ...\phi_p, \theta_1, \theta_1, ...\theta_q, \sigma^2)'$ el vector de los parámetros poblacionales. Tenemos una muestra de observaciones, de tamaño T. El método consiste en calcular la **densidad de probabilidad**:

$$f_{Y_t, Y_{t-1}, .., Y_1}(y_t, y_{t-1}, .., y_1; \theta) \tag{1}$$

que puede ser interpretada como la probabilidad de haber observado esa muestra particular. La estimación MV de θ es el valor para la cual la muestra con que contamos es la más probable de observarse, es el valor de θ que maximiza la ec. 1.

Introducción I

Este método requiere especificar una distribución particular para los Ruidos Blancos ε_t. Vamos a suponer que son Ruidos Blancos Gaussianos.

El desarrollo de la presentación recorre dos etapas:

- Calcular la función de verosimilitud
- Encontrar los valores de θ que maximicen esa función.

La función de verosimilitud de un proceso AR(1) Normal

Consideremos un proceso $\mathrm{AR}(1)$ gaussiano.

$Y_t = c + \phi Y_{t-1} + \varepsilon_t \quad$ con $\varepsilon_t \sim i.i.dN(0, \sigma^2)$.

En este caso el vector de parámetros poblacionales a ser estimado es: $\theta = (c, \phi, \sigma^2)'$.

La media del proceso AR(1) es $E(Y_t) = \frac{c}{1-\phi}$ si $|\phi| < 1$ el proceso es estacionario en sentido débil y la varianza es $V(Y_t) = \gamma_0 = \frac{\sigma^2}{(1-\phi^2)}$.

Consideremos la distribución de probabilidad de **la primera observación del proceso,** Y_1 dado que $\{\varepsilon_t\}_{t=-\infty}^{\infty}$ es un proceso gausiano,

Y_1 también es normal y su función de densidad es la siguiente:

$$f_{Y_1}(y_1, \theta) = f_{Y_1}\left(y_1, c, \phi, \sigma^2\right) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2/(1-\phi^2)}} \exp\left[-\frac{1}{2}\frac{(y_1 - [c/(1-\phi)])^2}{[\sigma^2/(1-\phi^2)]}\right] \tag{2}$$

La función de verosimilitud de un proceso AR(1) Normal II

Ahora consideremos la distribución de la segunda observación Y_2 condicional a la información observada de $Y_1 = y_1$. Entonces tenemos:

$$Y_2 = c + \phi Y_1 + \varepsilon_2 \tag{3}$$

Condicionar en $Y_1 = y_1$ significa tratar la variable aleatoria Y_1 como si fuese una constante determinista, y_1. En la ec.3, Y_2 se expresa como un constante $(c + \phi y_1)$ más $\varepsilon_2 \sim N(0, \sigma^2)$. Por lo tanto,

$$(Y_2 \mid Y_1) \sim N\left((c + \phi_1 y_1), \sigma^2\right)$$

Por lo que la función de densidad de $(Y_2 \mid Y_1)$ es:

$$f_{Y_2|Y_1}(y_2 \mid y_1, \theta) = f_{Y_2|Y_1}\left(y_2 \mid y_1, c, \phi, \sigma^2\right) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2}} \exp\left[-\frac{1}{2}\frac{(y_2 - c - \phi y_1])^2}{\sigma^2}\right] \tag{4}$$

La función de densidad conjunta de las dos observaciones se obtiene mediante el producto de f_{Y_1} y $f_{Y_1|Y_2}$:

$$f_{Y_2,Y_1}(y_2, y_1, \theta) = f_{Y_2|Y_1}(y_2 \mid y_1; \theta) f_{Y_1}(y_1, \theta) \tag{5}$$

La función de verosimilitud de un proceso AR(1) Normal III

Similarmente podemos obtener la distribución de la tercer observación, Y_3 condicional a observar Y_1 y Y_2:

$$f_{Y_3|Y_2,Y_1}(y_3 \mid y_2, y_1; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{(y_3 - c - \phi y_2)^2}{\sigma^2}\right] \tag{6}$$

La densidad conjunta de $f_{Y_3,Y_2,Y_1}(y_3, y_2, y_1, \theta)$ queda expresada como:

$$f_{Y_3,Y_2,Y_1}(y_3, y_2, y_1, \theta) = f_{Y_3|Y_2,Y_1}(y_3 \mid y_2, y_1; \theta) f_{Y_2,Y_1}(y_2, y_1; \theta) \tag{7}$$

En general, los valores de $Y_1, \ldots Y_{t-1}$ importan para Y_t solamente mediante el valor de la observación anterior Y_{t-1}, y la densidad de la observación t condicional en la observaciones anteriores hasta $t-1$ esta dada por:

$$f_{Y_t|Y_{t-1},Y_{t-2},..,Y_1}(y_t \mid y_{t-1}, y_{t-2}, .., y_1; \theta) = f_{Y_t|Y_{t-1}}(y_t \mid y_{t-1}; \theta) \tag{8}$$

$$f_{Y_t|Y_{t-1}}(y_t \mid y_{t-1}; \theta) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2}} \exp\left[-\frac{1}{2}\frac{(y_t - c - \phi y_{t-1}])^2}{\sigma^2}\right] \tag{9}$$

La densidad conjunta de las primeras t observaciones es entonces:

$$f_{Y_t,Y_{t-1},..,Y_1}(y_t,y_{t-1},..,y_1;\theta) = f_{Y_t|Y_{t-1}}(y_t \mid y_{t-1};\theta) f_{Y_{t-1},Y_{t-2},..,Y_1}(y_{t-1},y_{t-2},..,y_1;\theta) \tag{10}$$

Si consideramos una muestra de tamaño T, **la función de verosimilitud** para la muestra completa se expresa como:

$$f_{Y_t,Y_{t-1},..,Y_1}(y_t,y_{t-1},..,y_1;\theta) = f_{Y_1}(y_1;\theta) \cdot \prod_{t=2}^{T} f_{Y_t|Y_{t-1}}(y_t \mid y_{t-1};\theta) \tag{11}$$

El valor de θ que maximiza la ec. 11 es el mismo valor que maximiza la log-verosimilitud:

$$\mathscr{L}(\theta) = \log f_{Y_1}(y_1;\theta) + \sum_{t=2}^{T} \log f_{Y_t|Y_{t-1}}(y_t \mid y_{t-1};\theta) \tag{12}$$

La función de verosimilitud de un proceso AR(1) Normal V

Si sustituimos en la ec. 12 los términos $f_{Y_1}(y_1;\theta)$ y $f_{Y_t|Y_{t-1}}(y_t \mid y_{t-1};\theta)$ obtenemos: el log de la verosimilitud para una muestra de tamaño T de un AR(1).

$$\mathscr{L}(\theta) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\left[\sigma^2/\left(1-\phi^2\right)\right] - \frac{\left(y_1 - [c/(1-\phi)]\right)^2}{2\sigma^2/(1-\phi^2)} - [(T-1)/2]\log(2\pi)$$
$$-[(T-1)/2]\log\left(\sigma^2\right) - \sum_{t=2}^{T}\left[\frac{\left(y_t - c - \phi y_{t-1}\right)^2}{2\sigma^2}\right] \tag{13}$$

Estimación por método de máxima verosimilitud exacta para procesos AR(1)

La estimación MV de θ es el valor que maximiza la ec. 13. Para obtener el valor que maximice la función de verosimilitud es necesario derivar la función respecto de θ (espacio de parámetros) y luego igualar a cero. Finalmente despejando θ se obtendrá el máximo verosímil para la muestra dada. En la práctica es necesario trabajar con sistemas de ecuaciones no lineales en θ y $\{y_1, y_2, ...y_T\}$ y no hay una solución simple para θ en términos de $\{y_1, y_2, ...y_T\}$. Maximizar la ec.13 requiere procedimientos numéricos iterativos.

- Grid Search
- Stepest Ascent
- Newton-Raphson
- Davidon-Fletchet-Powell

Una alternativa a la maximización numérica de la función de verosimilitud exacta es tomar el valor de y_1 como determinístico y maximizar la verosimilitud condicional a esa primer observación.

Estimación MV condicional de los parámetros de un proceso AR(1)

$$f_{Y_T, Y_{T-1}, .., Y_2 | Y_1}(y_T, y_{T-1}, .., y_2 \mid y_1; \theta) = \prod_{t=2}^{T} f_{Y_t | Y_{t-1}}(y_t \mid Y_{t-1}; \theta) \tag{14}$$

el objetivo es maximizar es la log-verosimilitud:

$$\begin{aligned} &\log f_{Y_T, Y_{T-1}, .., Y_2 | Y_1}(y_T, y_{T-1}, .., y_2 \mid y_1; \theta) \\ &\qquad = -[(T-1)/2]\log(2\pi) - [(T-1)/2]\log(\sigma^2) \\ &\qquad\quad - \sum_{t=2}^{T} \left[\frac{(y_t - c - \phi y_{t-1})^2}{2\sigma^2} \right] \end{aligned} \tag{15}$$

Maximizar esta función con respecto a c y a ϕ es equivalente a minimizar:

$$\sum_{t=2}^{T} (y_t - c - \phi y_{t-1})^2 \tag{16}$$

Esto puede resolverse mediante una regresión de MCO de Y_t sobre una constante y su rezago.

Estimación MV condicional I

La estimación de máxima verosimilitud condicional de la constante c y ϕ está dada por:

$$\begin{bmatrix} \hat{c} \\ \hat{\phi} \end{bmatrix} = \begin{bmatrix} (T-1) & \sum_{t=2}^{T} Y_{t-1} \\ \sum_{t=2}^{T} Y_{t-1} & \sum_{t=2}^{T} Y_{t-1}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{t=2}^{T} y_t \\ \sum_{t=2}^{T} y_{t-1} y_t \end{bmatrix} \tag{17}$$

Para obtener la estimación de MV condicional de σ^2, se diferencia la ecuación (15) con respecto a σ^2 y se iguala el resultado a cero :

$$\frac{-(T-1)}{2\sigma^2} + \sum_{t=2}^{T} \left[\frac{(y_t - c - \phi y_{t-1})^2}{2\sigma^4} \right] = 0 \tag{18}$$

Operando obtenemos:

$$\hat{\sigma}^2 = \sum_{t=2}^{T} \left[\frac{\left(y_t - \hat{c} - \hat{\phi} y_{t-1}\right)^2}{T-1} \right] \tag{19}$$

MV condicional es la media de los residuos al cuadrado de la regresión MCO.

Estimación MV condicional II

- La estimación por MV condicional de σ^2 resulta ser el promedio de los residuos al cuadrado de una regresión por MCO.
- Al contrario de MV exacta que necesita solución númerica para resolver un sistema de ecuaciones no lineales, MV condicional presenta una solución cerrada (MCO).
- Si la muestra es suficientemente larga, la primera observación tiene una contribución pequeña en el total de la verosimilitud.
- En el caso de $|\phi| < 1$ tienen la misma distribución muestral.
- Si $|\phi| > 1$, MV condicional es consistente mientras la maximización de la log-verosimilitud no lo es.Por eso en muchas ocasiones los parámetros de los procesos AR se estiman por MV condicional (MCO) más que por MV exacta.

Función de verosimilitud condicional

Calcular la función de verosimilitud para un proceso autorregresivo es más sencillo si condicionamos en los valores iniciales de Y. Igualmente, el cálculo de la función de verosimilitud para un proceso de media móvil es más simple si se condiciona en los valores iniciales de los ε's.

Considere el proceso $MA(1)$:

$$Y_t = \mu + \theta\varepsilon_{t-1} + \varepsilon_t$$

con $\varepsilon \sim i.i.d.N(0, \sigma^2)$.
Sean $\theta = (\mu, \theta, \sigma^2)'$ los parámetros poblacionales a estimar. Si el valor de ε_{t-1} fuese conocido con certeza, entonces:

$$Y_t \mid \varepsilon_{t-1} \sim N((\mu + \theta\varepsilon_{t-1}), \sigma^2)$$

$$f_{Y_t|\varepsilon_{t-1}}(y_t \mid \varepsilon_{t-1}; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y_t-\mu-\theta\varepsilon_{t-1})^2}{2\sigma^2}} \tag{20}$$

La función de verosimilitud de un proceso MA(1) Normal I

Supóngase que conocemos con certeza que $\varepsilon_0 = 0$, entonces:

$$(Y_1 \mid \varepsilon_0 = 0) \sim N(\mu, \sigma^2)$$

Además, dada la observación y_1, el valor de ε_1 es entonces conocido con certeza:

$$\varepsilon_1 = y_1 - \mu$$

Lo cual permite aplicar la ec 20 (verosimilitud condicional en ε_0) nuevamente:

$$f_{Y_2|Y_1,\varepsilon_0=0}(y_2 \mid y_1, \varepsilon_0 = 0; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y_2-\mu-\theta\varepsilon_1)^2}{2\sigma^2}} \tag{21}$$

Como ε_1 es conocido, ε_2 puede ser calculado como:

$$\varepsilon_2 = y_2 - \mu - \theta\varepsilon_1$$

La función de verosimilitud de un proceso MA(1) Normal II

Procediendo de esta manera, conociendo $\varepsilon_0 = 0$, la secuencia completa $\{\varepsilon_1, \varepsilon_2, ..., \varepsilon_T\}$ puede ser calculada a partir de $\{y_1, y_2, ..., y_T\}$ iterando en:

$$\varepsilon_t = y_t - \mu - \theta\varepsilon_{t-1}$$

para todo $t = 1, 2, ..., T$, empezando desde $\varepsilon_0 = 0$.
La función de densidad condicional es:

$$f_{Y_t|Y_{t-1}, Y_{t-2}..... Y_1, \varepsilon_0=0}(y_t \mid y_{t-1}, y_{t-2}, ...y_1, \varepsilon_0 = 0; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-\varepsilon_t^2}{2\sigma^2}} \qquad (22)$$

La verosimilitud muestral se puede escribir como el producto de las densidades individuales:

$$f_{Y_t|Y_{t-1}, Y_{t-2}..... Y_1, \varepsilon_0=0}(y_t \mid y_{t-1}, y_{t-2}, ...y_1, \varepsilon_0 = 0; \theta) = f_{Y_1|\varepsilon_0=0}(y_1 \mid \varepsilon_0 = 0; \theta) * \prod_{t=2}^{T} f_{Y_t|Y_{t-1}, Y_{t-2}..... Y_1, \varepsilon_0=0}(y_t \mid y_{t-1}, y_{t-2}, ...y_1, \varepsilon_0 = 0; \theta) \quad (23)$$

La función de verosimilitud de un proceso MA(1) Normal III

La función de log-verosimilitud es:

$$\mathscr{L}(\theta) = log\ f_{Y_t|Y_{t-1},Y_{t-2}.....Y_1,\varepsilon_0=0}(y_t \mid y_{t-1}, y_{t-2}, ...y_1, \varepsilon_0 = 0; \theta) =$$

$$= -\frac{T}{2} log(2\pi) - \frac{T}{2} log(\sigma^2) - \sum_{t=1}^{T} \frac{\varepsilon_t^2}{2\sigma^2} \qquad (24)$$

La log verosimilitud condicional,$\mathscr{L}(\theta)$ es una función de los ε al cuadrado.

Esa log verosimilitud es una función no lineal de μ y θ, por lo que la expresión analítica de máximo de la función de log verosimilitud respecto de μ y θ no es sencilla de calcular.Por lo que la estimación MV condicional delos parámetros de un proceso MA(1) requieren el uso de optimización numérica.

Si $|\theta| < 1$ el efecto de imponer que $\varepsilon_0 = 0$ se pierde rápidamente y la verosimilitud condicional será una buena aproximación de la verosilimitud incondicional, para una muestra razonablemente grande.

Pero si $|\theta| > 1$ las consecuencias de imponer que $\varepsilon_0 = 0$ se acumulan en el tiempo y la aproximación condicional no es buena.Si de la aproximación numérica resulta un $|\theta| > 1$ debe ser descartado, se debiera comenzar nuevamente y se debe usar el inverso de $\hat{\theta}$ como valor de inicio en el procedimiento.

Estimar $\hat{\theta}$ por MV, implica maximizar el logaritmo de la verosimilitud $\mathscr{L}(\theta)$. Las fórmulas presentadas se apoyan en ciertos supuestos : Se supone que los datos provienen de un proceso estacionario y que ni los parámetros estimados $\hat{}$, ni el verdadero valor están en los límites del espacio de parámetros.

Error estándar asintótico para estimadores de Máxima Verosimilitud

Si el tamaño de la muestra, T, es suficientemente grande, se puede decir que la distribución del estimador de máxima verosimilitud $\hat{\theta}$ puede ser aproximado por la siguiente distribución:

$$\hat{\theta} \approx N(\theta_0, T^{-1}\mathscr{I}^{-1}) \tag{25}$$

Donde θ_0 es el verdadero valor del vector de parámetros, $\mathscr{I}$ es la llamada Matriz de Información, esta puede ser estimada al menos de dos formas:
A través del **estimador de segunda derivada de la matriz de información**

$$\hat{\mathscr{I}}_{2D} = -T^{-1}\frac{\partial^2 \mathscr{L}(\theta)}{\partial\theta\partial\theta'}|_{\theta=\hat{\theta}} \tag{26}$$

$\mathscr{L}(\theta)$ es el logaritmo de la verosimilitud:

$$\mathscr{L}(\theta) = \sum_{t=1}^{T} log f_{Y_t|\mathscr{Y}_{\sqcup-\infty}}(y_t|\mathscr{Y}_{t-1};\theta) \tag{27}$$

Donde $\mathscr{Y}_t$ simboliza la historia de las observaciones de y a través del tiempo t.

La matriz de segunda derivada del logaritmo de la verosimilitud generalmente se calcula de forma numérica. Sustituyendo la fórmula $\hat{\mathscr{I}}_{2D}$ en $\hat{\theta} \approx N(\theta_0, T^{-1}\mathscr{I}^{-1})$.

Con esto, **la matriz de covarianzas** de $\hat{\theta}$ se puede aproximar de la siguiente forma:

$$E(\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)' \cong [-\frac{\partial^2 \mathscr{L}(\theta)}{\partial\theta\partial\theta'}|_{\theta=\hat{\theta}}]^{-1} \tag{28}$$

Como segunda alternativa se puede **estimar el producto externo**

$$\hat{\mathscr{I}}_{OP} = T^{-1} \sum_{t=1}^{T} [h(\hat{\theta}, \mathscr{Y}_t][h(\hat{\theta}, \mathscr{Y}_t]' \tag{29}$$

Donde

$$h(\hat{\theta}, \mathscr{Y}_t) = \frac{\partial log f(y_t|y_{t-1}, y_{t-2}, ...; \theta)}{\partial\theta}|_{\theta=\hat{\theta}} \tag{30}$$

es el vector de derivadas del log de la densidad condicional de la r-ésima observación con respecto al elemento a del vector de parámetros θ con su derivada evaluada en el máximo, $\hat{\theta}$.

Error estándar asintótico para estimadores de Máxima Verosimilitud I

En este caso, la matrix de varianzas y covarianzas de $\hat{\theta}$ es aproximada por:

$$E(\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)' \cong [\sum_{t=1}^{T} [h(\hat{\theta}, \mathscr{Y}_t)][h(\hat{\theta}, \mathscr{Y}_t)]^{-1} \tag{31}$$

Se puede suponer que la log verosimilitud toma la forma de:

$$\mathscr{L}(\theta) = -1,5\theta_1^2 - 2\theta_2^2 \tag{32}$$

Entonces:

$$\frac{\partial^2 \mathscr{L}(\theta)}{\partial \theta \partial \theta'} = \begin{bmatrix} -3 & 0 \\ 0 & -4 \end{bmatrix}$$

Junto con la ecuación $E(\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)'$ **la varianza** del estimador máximo verosímil $\hat{\theta}_2$ puede ser aproximado por $1/4$. El **estimador** máximo verosímil para este ejemplo es $\hat{\theta}_2 = 0$
El **intervalo de confianza** al 95 % para θ_2 sería:

$$0 \pm 2\sqrt{\frac{1}{4}} = \pm 1$$

En general se necesitan calcular los elementos de la matriz $\hat{\mathscr{I}}$ e invertirla para obtener errores estándar para cualquier parámetro dado, a no ser que los elementos de la diagonal de $\hat{\mathscr{I}}$ sean cero.

Prueba de razón de Verosimilitud

La prueba de razón de verosimilitud es otra manera de contrastar hipótesis sobre los parámetros que son estimados por máxima verosimilitud.

Se supone una hipótesis nula que implica m restricciones distintas, en el valor de un vector de parámetros $(a \times 1)$ de θ.

Primeramente se maximiza la función de verosimilitud ignorando estas restricciones, para obtener la estimación máxima verosímil $\hat{\theta}$ no restringida. Luego, se encuentra un estimador $\tilde{\theta}$ que maximice la verosimilitud, satisfaciendo todas las restricciones. Para ello, se define un nuevo vector de dimensión λ de $[(a-m) \times 1]$, en donde todos los elementos de Θ satisfacen todas las restricciones.

Por ejemplo, si la restricción es que los últimos m elementos de Θ sean cero, entonces λ serán los primeros m elementos de θ.

Sea $\mathscr{L}(\hat{\theta})$ el valor de la función log verosimilitud en la estimación no restringida, y $\mathscr{L}(\tilde{\theta})$ que describe el valor de la log verosimilitud en la estimación restringida.

$\mathscr{L}(\hat{\theta}) > \mathscr{L}(\tilde{\theta})$, luego tenemos que:

$$2[\mathscr{L}(\hat{\theta}) - \mathscr{L}(\tilde{\theta})] \approx \chi^2(m)$$

Observación final

Observación:
Estas presentaciones no constituyen una nota completa sobre el tema, sólo son una guía para comprender el tema de estimación por Máxima Verosimiliud, se recomienda la lectura del capítulo 5 de Hamilton, J. (1994)

Referencia bibliográfica

- Hamilton, J. (1994) *"Time Series Analysis"*. Princeton University Press