Máxima Verosimilitud

Fundamentos de Econometría

Juan Palomino[1]

[1]Magister en Economía Aplicada con Mención Estudios Regionales
juan.palominoh@pucp.pe

Departamento de Economía

Índice

Introducción

MCO

- Las estimaciones de MCO se derivan de ajustar un modelo lineal que representa el promedio de los datos.
- El modelo se obtiene al minimizar la suma de las desviaciones o errores al cuadrado.

Máxima Verosimilitud

- La función de verosimilitud indica qué tan probable es reproducir la muestra observada en función de los posibles valores de los parámetros.
- El fin es maximizar la función de verosimilitud con base en los parámetros que tienen más probabilidades de producir los datos observados.

Introducción

- Supongamos que contamos con una muestra de 7 observaciones de una variable aleatoria Y, siendo estas $Y_1, Y_2, Y_3, Y_4, Y_5, Y_6, Y_7$.
- No sabemos que distribución generó estos datos, pero supongamos que fue una variable aleatoria normalmente distribuida.
- Una distribución normal queda plenamente definida con los valores de su media μ_Y y su varianza σ_Y^2.
- **¿Qué valores de estos parámetros μ_Y y σ_Y^2 generaron a los datos observados con mayor probabilidad?**

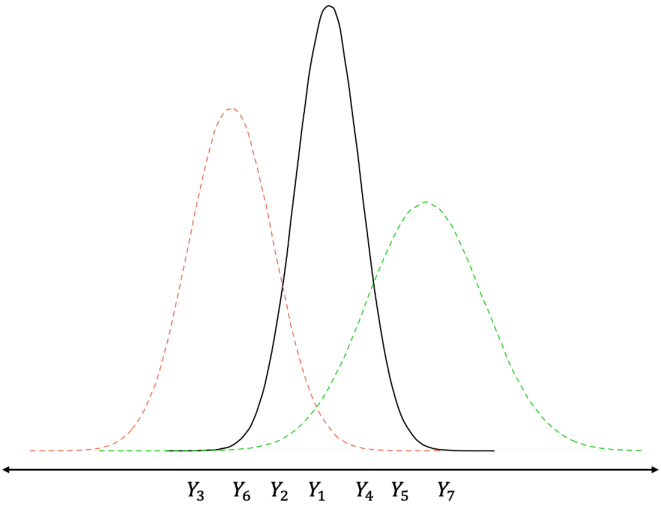

Figura: Datos observados y distribuciones alternativas

Introducción

Las ventajas de Máxima Verosimilitud sobre MCO son:

- Mejores estimaciones de los parámetros de distribución.
- Menor varianza de los parámetros (eficiencia)
- Confiabilidad en la medición de los intervalos de confianza y en las pruebas de hipótesis de los parámetros.

Máxima Verosimilitud es bastante utilizado para las regresiones no lineales y muestras grandes, por ejemplo, modelos de elección discreta (Probit, Logit), heterocedasticidad condicional como GARCH y EGARCH, modelos censurados y truncados, entre otros.

Índice

Función de Densidad Conjunta

- Sea $\{y_i\}$ una muestra aleatoria simple de $y_1, y_2, ..., y_n$, con lo cual las y_i son variables aleatorias iid con función de densidad $f(y_i \mid x_i; \theta)$, donde $\theta' = [\theta_1, \theta_2, ..., \theta_k]$ es un vector de parámetros desconocidos. Bajo el supuesto de independencia, la función de densidad conjunta es:

$$\begin{aligned} f(y_1, y_2, ..., y_n \mid x_i; \theta) &= f(y_1 \mid x_i; \theta) \cdot f(y_2 \mid x_i; \theta) \cdot ... \cdot f(y_n \mid x_i; \theta) \\ &= \prod_{i=1}^{n} f(y_i \mid x_i; \theta) \end{aligned}$$

- Esta densidad conjunta indica la probabilidad de obtener la muestra $y_1, y_2, ..., y_n$ dado el vector de parámetros θ.

Índice

Función de Verosimilitud

- La funcion de verosimilitud denotada por L es:

$$L(\theta, y \mid X) = \prod_{i=1}^{n} L(\theta; y_i \mid x_i) = \prod_{i=1}^{n} f(y_i \mid x_i; \theta)$$

- donde $y = (y_1, ..., y_n)$.
  - $L(\theta; y_i \mid x_i)$ es la contribución de verosimilitud de la observación i
  - $L(\theta, y \mid X)$ es la función de verosimilitud de toda la muestra
- La función de verosimilitud señala que para cualquier muestra dada $y \mid X$, la estimación de verosimilitud es encontrar un set de parámetros estimados, es decir $\hat{\theta}$, tal que esta verosimilitud es maximizada.

Función Log-Verosimilitud

- La función de log-verosimilitud es:

$$lnL(\theta, y \mid X) = lnL(\theta) = ln\underbrace{\left(\prod_{i=1}^{n} f(y_i \mid x_i; \theta)\right)}_{f(y|X;\theta)} = \sum_{i=1}^{n} lnf(y_i \mid x_i; \theta)$$

- La función log-verosimilitud es una función creciente monotónica de $L(\theta, y \mid X)$:
    - Cualquier valor de maximización $\hat{\theta}$ de $lnL(\theta, y \mid X)$ también debe maximizar $L(\theta, y \mid X)$.
- Tomando logaritmos que convierte productos en suma.
    - Permite simplicar en la determinación numérica del EMV.
    - Valores de verosimilitud son extremadamente pequeños.
    - Problemas de optimización numérica

Índice

Función Score

- Derivando la función log-verosimilitud respecto a θ e igualando a cero, se tiene:
$$\frac{\partial lnL(\theta; y_1, ..., y_n \mid x_i)}{\partial \theta} = \underset{k\times 1}{0}$$
- Los estimadores de MV son aquellos $\hat{\theta}_{MV}$ que resuelven estas k ecuaciones.

Índice

Propiedades asintóticas

Consistencia

Todos los estimadores máximo-verosímiles son consistentes.

$$plim \hat{\theta}_{MV} = \theta$$

Propiedades asintóticas

Normalidad Asintótica

Cuando la muestra es muy grande, la distribución de $\hat{\theta}_{MV}$ es asintóticamente normal, para cualquier función de densidad $f(y_i \mid x_i; \theta)$:

$$\sqrt{n}(\hat{\theta}_{MV} - \theta) \xrightarrow{d} N(0, nI^{-1}(\theta))$$

Alcanzando la cota mínima de Cramér-Rao $nI^{-1}(\theta)$, donde $I^{-1}(\theta)$ es la inversa de la "matriz de información" definida como:

$$I(\theta) = -E[\frac{\partial^2 lnL}{\partial\theta\partial\theta'}]$$

Para n grande, se cumple que la distribución asintótica de $\hat{\theta}_{MV}$ es:

$$\hat{\theta}_{MV} \overset{a}{\sim} N(\theta, I^{-1}(\theta))$$

Propiedades asintóticas

Eficiencia Asintótica

Los estimadores de máxima verosimilitud son los estimadores más eficiente, dentro de la clase de estimadores consistentes, asintóticamente insesgados y asintóticamente normales.

Invarianza

Si $\hat{\theta}_{MV}$ es un estimador de máxima verosimilitud y $w = g(\theta)$ es una función continua de θ; entonces, $\hat{w} = g(\hat{\theta}_{MV})$ es el estimador máximo verosímil de w.

Índice

Estimador Restringido

- Se pueden imponer restricciones lineales o no lineales en la estimación por máxima verosimilitud.
- Si queremos estimar un parámetro θ por este método, pero sujeto a que se cumpla la restricción $C(\theta) = r$, el problema de optimización a resolver es:

$$max\ lnL(\theta)\ s.a.\ C(\theta) = r$$

- El Lagrangiano de este problema es:

$$\mathscr{L} = lnL(\theta) + \lambda(r - C(\theta))$$

Estimador Restringido

- Las condiciones de primer orden son:

$$\frac{\partial \mathscr{L}}{\partial \theta} = \frac{\partial lnL(\theta)}{\partial \theta} - \lambda \frac{\partial C(\theta)}{\partial \theta} = 0$$

$$\frac{\partial \mathscr{L}}{\partial \lambda} = r - C(\theta) = 0$$

- La solución a estas ecuaciones es el estimador de máxima verosimilitud restringido $\tilde{\theta}_{MV}$.

Índice

Estimación del Modelo Bivariado

- Tenemos n observaciones de dos variables aleatorias (X_i e Y_i), que se relacionan entre sí mediante el modelo $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$.
- Dado que:

$$Y_i | X_i \sim N(\beta_1 + \beta_2 X_i, \sigma^2)$$

- La función de densidad de probabilidad condicional (pdf) es:

$$f(Y_i \mid X_i) = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{\left[-\frac{(Y_i - \beta_1 - \beta_2 X_i)^2}{2\sigma_0^2}\right]}$$

Estimación del Modelo Bivariado

- La función de verosimilitud es:

$$L(\beta_1, \beta_2, \sigma^2) = \prod_{i=1}^{n} f(Y_i \mid X_i) = (\frac{1}{2\pi\sigma_0^2})^{n/2} \prod_{i=1}^{n} e^{[-\frac{(Y_i-\beta_1-\beta_2 X_i)^2}{2\sigma_0^2}]}$$

$$= (\frac{1}{2\pi\sigma_0^2})^{n/2} e^{[-\frac{\sum(Y_i-\beta_1-\beta_2 X_i)^2}{2\sigma_0^2}]}$$

- Tomando el logaritmo natural:

$$lnL(\beta_1, \beta_2, \sigma^2) = -\frac{n}{2} ln(2\pi) - \frac{n}{2} ln(\sigma_0^2) - \frac{1}{2\sigma_0^2} \sum (Y_i - \beta_1 - \beta_2 X_i)^2$$

- Derivando respecto a los parámetros:

$$\frac{\partial lnL(\beta_1, \beta_2, \sigma^2)}{\partial \beta_1} = \frac{1}{\sigma^2} \sum (Y_i - \beta_1 - \beta_2 X_i) = 0$$

$$\frac{\partial lnL(\beta_1, \beta_2, \sigma^2)}{\partial \beta_2} = \frac{1}{\sigma^2} \sum (Y_i - \beta_1 - \beta_2 X_i) X_i = 0$$

$$\frac{\partial lnL(\beta_1, \beta_2, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum (Y_i - \beta_1 - \beta_2 X_i) = 0$$

- La solución para β_1y β_2 son las mismas que las de mínimas cuadrados ordinarios:

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2\bar{X}$$

$$\hat{\beta}_2 = \frac{\sum X_iY_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n\bar{X}^2} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

- Luego reemplazamos $\hat{\beta}_1$ y $\hat{\beta}_2$ en $\frac{\partial lnL(\beta_1,\beta_2,\sigma^2)}{\partial\sigma^2}$ y obtenemos:

$$\hat{\sigma}^2 = \frac{\sum(Y_i - \hat{\beta}_1 - \hat{\beta}_2X_i)^2}{n} = \frac{\sum(\hat{\varepsilon}^2)}{n} = \frac{SCR}{n}$$

- El estimador de $\hat{\sigma}^2$ hallado por MV no varía a pesar del cambiante número de variables independientes que tenga el modelo, haciendolo sesgado.

Índice

Estimación con K variables

- Consideremos que $\{y_i, x_i\}$ es iid, entonces el modelo de regresión lineal:

$$y_i = x_i^{'}\beta + \varepsilon_i \quad \varepsilon_i \mid x_i \sim N(0, \sigma_0^2)$$

- Por lo tanto la función de densidad de probabilidad condicional (pdf) es:

$$f(y_i \mid x_i; \theta_0) = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{[-\frac{(y_i - x_i^{'}\beta_0)^2}{2\sigma_0^2}]}$$

- La pdf conjunta de la muestra es:

$$\prod_{i=1}^{n} f(y_i \mid x_i; \theta_0) = (\frac{1}{2\pi\sigma_0^2})^{n/2} \prod_{i=1}^{n} e^{[-\frac{(y_i - x_i^{'}\beta_0)^2}{2\sigma_0^2}]}$$

Estimación con K variables

- La función log-verosimilitud condicional de $y_i \mid x_i \sim N(x_i^{'}\beta_0, \sigma_0^2)$ es:

$$\begin{aligned}
log(\prod_{i=1}^{n} f(y_i \mid x_i; \theta_0)) &= \frac{n}{2} log(\frac{1}{2\pi\sigma_0^2}) + \sum_{i=1}^{n} \frac{-(y_i - x_i^{'}\beta_0)^2}{2\sigma_0^2} \\
&= \frac{n}{2} log(1) - \frac{n}{2} log(2\pi\sigma_0^2) - \sum_{i=1}^{n} \frac{(y_i - x_i^{'}\beta_0)^2}{2\sigma_0^2} \\
&= -\frac{n}{2} log(2\pi) - \frac{n}{2} log(\sigma_0^2) - \sum_{i=1}^{n} \frac{(y_i - x_i^{'}\beta_0)^2}{2\sigma_0^2}
\end{aligned}$$

Estimación con K variables

- El objetivo es buscar los valores de β y σ^2 que maximicen esta función. Entonces, las condiciones de primer orden son

$$
\begin{aligned}
\frac{\partial lnL}{\partial \beta} &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(y_i - x_i^{'}\beta_0)(-x_i^{'}) = 0 \\
&= \sum_{i=1}^{n}(y_i x_i^{'} - x_i^{'}\beta_0 x_i^{'}) = 0 \\
&= \sum_{i=1}^{n} y_i x_i^{'} - \beta_0 \sum_{i=1}^{n} x_i x_i^{'} = 0 \\
\hat{\beta}_{MV} &= (\sum_{i=1}^{n} x_i x_i^{'})^{-1} \sum_{i=1}^{n} x_i y_i = \hat{\beta}_{MCO}
\end{aligned}
$$

Estimación con K variables

- Con respecto a σ^2:

$$\frac{\partial lnL}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - x_i^{'}\beta_0)^2 = 0$$

$$\hat{\sigma}^2_{MV} = \frac{\sum_{i=1}^{n}(y_i - x_i^{'}\beta_0)^2}{n} = \frac{\sum_{i=1}^{n}\varepsilon_i{}^2}{n}$$

Estimación con K variables

- El estimador $\hat{\sigma}^2_{MV}$ no es igual al de MCO, de hecho es un estimador sesgado, pero asintóticamente insesgado. Veamos:

$$E[\hat{\sigma}^2_{MV}] = E[\frac{\hat{\varepsilon}'\hat{\varepsilon}}{n}] = \frac{1}{n}E[\hat{\varepsilon}'\hat{\varepsilon}] = \frac{(n-k)\sigma^2}{n} \neq \sigma^2_{MCO}$$

- Podemos decir que es asintóticamente insesgado, pues:

$$\lim_{n\to\infty} E[\hat{\sigma}^2_{MV}] = \lim_{n\to\infty} (\frac{n-k}{n})\sigma^2 = \sigma^2$$

Estimación con K variables

- Estos dos estimadores cumplen las propiedades de los estimadores máximo verosímiles.
- En concreto, nos interesa resaltar que se cumplirá que son consistentes y también son asintóticamente normales:

$$plim(\hat{\beta}_{MV}) = \beta \quad plim(\hat{\sigma}^2_{MV}) = \sigma^2$$

$$\begin{bmatrix} \hat{\beta}_{MV} \\ \hat{\sigma}^2_{MV} \end{bmatrix} \overset{a}{\sim} N\left(\begin{bmatrix} \beta \\ \sigma^2 \end{bmatrix}, \quad I^{-1}(\beta \;, \sigma^2) \right)$$

- Para calcular la matriz de varianzas y covarianzas asintótica de $\hat{\beta}_{MV}$ y $\hat{\sigma}^2_{MV}$, debemos hacer el cálculo de la matriz de información a través de las segundas derivadas:

$$\frac{\partial^2 lnL}{\partial\beta\partial\beta'} = -\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i x_i^{'}$$

$$\frac{\partial^2 lnL}{\partial\beta\partial\sigma^2} = \frac{1}{\sigma^4}(\sum_{i=1}^{n} x_i^{'}\varepsilon_i)$$

$$\frac{\partial^2 lnL}{\partial(\sigma^2)} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n} {\varepsilon_i}^2$$

- Multiplicando por (-1) y tomando el valor esperado, se obtiene:

$$-E[\frac{\partial^2 lnL}{\partial\beta\partial\beta'}] = \frac{X'X}{\sigma^2}$$
$$-E[\frac{\partial^2 lnL}{\partial\beta\partial\sigma^2}] = 0$$

$$\begin{aligned}-E[\frac{\partial^2 lnL}{\partial(\sigma^2)}] &= -\frac{n}{2\sigma^4} + \frac{1}{\sigma^6}E[\sum_{i=1}^{n}\varepsilon_i{}^2] \\ &= -\frac{n}{2\sigma^4} + \frac{1}{\sigma^6}\sigma^2 \\ &= -\frac{n}{2\sigma^4} + \frac{n}{\sigma^4} = \frac{n}{2\sigma^4}\end{aligned}$$

- La matriz de información es:

$$I(\beta,\sigma^2)=\begin{bmatrix} -E[\frac{\partial^2 lnL}{\partial\beta\partial\beta'}] & -E[\frac{\partial^2 lnL}{\partial\beta\partial\sigma^2}] \\ -E[\frac{\partial^2 lnL}{\partial\sigma^2\partial\beta}] & -E[\frac{\partial^2 lnL}{\partial(\sigma^2)^2}] \end{bmatrix}$$

$$=\begin{bmatrix} \frac{1}{\sigma^2}X'X & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}$$

Matriz de Información

- La matriz de varianzas y covarianzas asintótica:

$$I^{-1}(\beta,\sigma^2) = \begin{bmatrix} \sigma^2(X'X)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}$$

Referencias

- **Capítulo 4.4; Apéndice 4A y 7A.4** - Gujarati, D., & Porter, D. (2010). Econometría (Quinta edición ed.). & P. Carril Villareal, Trad.) México: Mc Graw Hill educación.
- **Chapter 2.6 Maximum Likelihood Estimation** - Scott Long, J. (1997). Regression models for categorical and limited dependent variables. Advanced quantitative techniques in the social sciences, 7.
- **Capítulo 14.1-14.10 Maximum Likelihood Estimation** - Greene, W. H. (2018). Econometric analysis. Pearson Education India.