Estimación Máxima Verosimilitud

Microeconomía Cuantitativa

R. Mora

Departmento of Economía
Universidad Carlos III de Madrid

Outline

Estrategias generales de estimación

Hay criterios de estimación que producen buenos estimadores

Mínimos cuadrados (MCO o MCG)

Método de momentos (MCO, MCG, VI):

$$\theta = g\left(E(Y)\right) \Rightarrow \hat{\theta} = g\left(E_N[y_i]\right)$$

Máxima Verosimilitud (MV)

Selecciona el vector $\hat{\theta}$ que maximiza la estimación de la probabilidad de la muestra.

Planteamiento básico

- Sea $\{y_1, y_2, \ldots, y_N\}$ una muestra iid de la población con densidad $f(Y; \theta_0)$. Queremos estimar θ_0
- Por el supuesto iid, la conjunta de $\{y_1, y_2, \ldots, y_N\}$ es el producto de las densidades:

$$f(y_1, y_2, \ldots, y_N; \theta_0) = f(y_1; \theta_0) f(y_2; \theta_0) \ldots f(y_N; \theta_0)$$

- La Función de Verosimilitud para una muestra dada se obtiene sustituyendo el verdadero θ_0 por cualquier θ

$$L(\theta) = f(y_1; \theta) f(y_2; \theta) \ldots f(y_N; \theta)$$

- $L(\theta)$ es una variable aleatoria porque depende de la muestra

Definición

El estimador máxima verosimilitud de θ_0, $\hat{\theta}^{ML}$, es el valor de θ que maximiza la función de verosimilitud $L(\theta)$

- Es conveniente trabajar con el logaritmo de la verosimilitud

$$l(\theta) = \sum_{i=1}^{N} log\left(f(y_i;\theta)\right)$$

- Puesto que la transformación logaritmica es monótona, $\hat{\theta}^{ML}$ también maximiza $l(\theta)$

Ejemplo: Bernoulli (1/3)

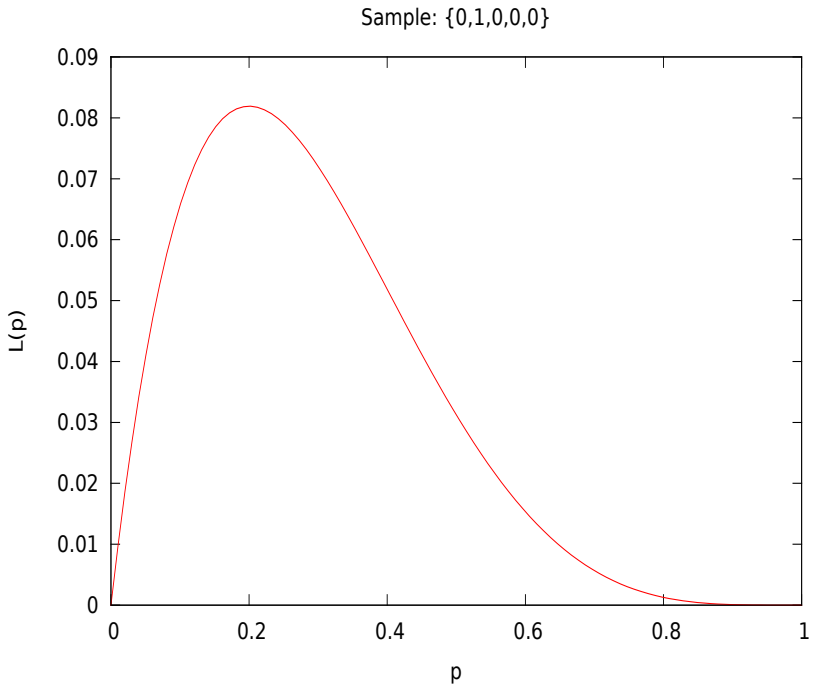
- Supón que Y es Bernoulli: $\begin{cases} 1 & \text{con probabilidad } p_0 \\ 0 & \text{con probabilidad } 1-p_0 \end{cases}$
- Verosimilitud para la observación i : $\begin{cases} p_0 & \text{si } y_i = 1 \\ 1-p_0 & \text{si } y_i = 0 \end{cases}$
- Sea n_1 el número de observaciones con valor 1. Entonces bajo muestreo iid,

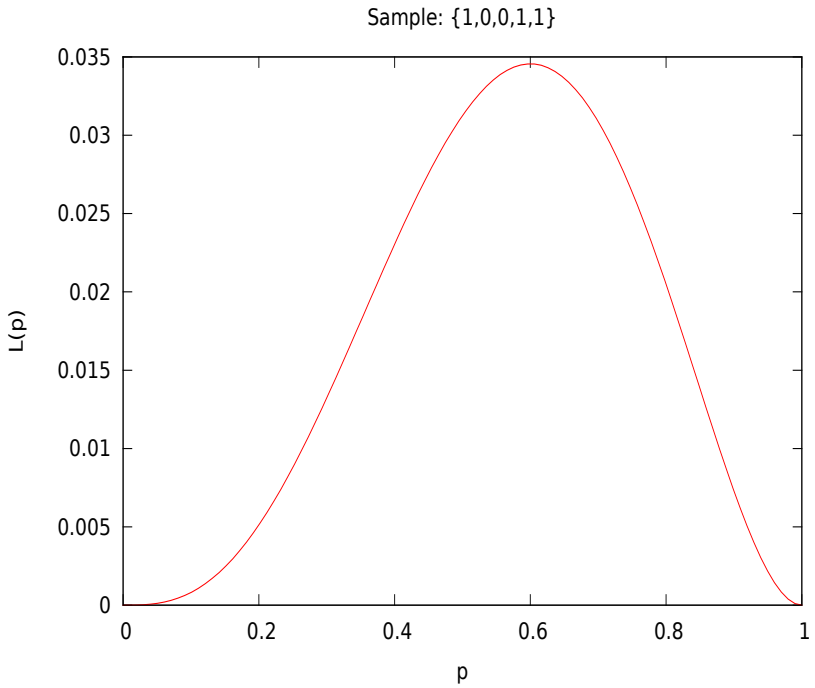
$$L(p) = p^{n_1}(1-p)^{n-n_1}$$

Para cada muestra tenemos una verosimilitud

- Con $\{0,1,0,0,0\} \Rightarrow L(p) = p(1-p)^4$
- Con $\{1,0,0,1,1\} \Rightarrow L(p) = p^3(1-p)^2$

Ejemplo: Bernoulli (2/3)





- Con $\{0,1,0,0,0\} \Rightarrow \hat{p} = 0.2$
- Con $\{1,0,0,1,1\} \Rightarrow \hat{p}^{ML} = 0.6$

Ejemplo: Bernoulli (3/3)

- El estimador máxima verosimilitud es el valor que maximiza

$$L(p) = p^{n_1}(1-p)^{n-n_1}$$

- El mismo $\hat{p}^{ML}$ maximiza el logaritmo de la verosimilitud

$$l(p) = n_1 log(p) + (n-n_1) log(1-p)$$

$$\frac{\partial l(p)}{\partial p} = 0 \Leftrightarrow \frac{n_1}{\hat{p}^{ML}} = \frac{n-n_1}{1-\hat{p}^{ML}} \Rightarrow \hat{p}^{ML} = \frac{n_1}{n}$$

- With $\{0,1,0,0,0\} \Rightarrow \hat{p}^{ML} = \frac{1}{5} = 0.2$
- With $\{1,0,0,1,1\} \Rightarrow\Rightarrow \hat{p}^{ML} = \frac{3}{5} = 0.6$

Planteamiento

- Sea $\{y_1, y_2, \ldots, y_N\}$ una muestra iid de $y|\mathbf{x} \sim N\left(\beta_0 x, \sigma_0^2\right)$.
- Queremos estimar $\theta_0 = \left(\beta_0, \sigma_0^2\right)$
- Debido al supuesto iid, la conjunta de $\{y_1, y_2, \ldots, y_N\}$ es el producto de las densidades:

$$f\left(y_1, y_2, \ldots, y_N | x_1, ..., x_N; \theta_0\right) = f(y_1|x_1;\theta_0) f(y_2|x_2;\theta_0) ... f(y_N|x_N;\theta_0)$$

- $y|\mathbf{x} \sim N\left(\beta_0 x, \sigma_0^2\right) \Rightarrow y - \beta_0 x \equiv \varepsilon \sim N\left(0, \sigma_0^2\right)$. Esto implica que

$$f_{y|x}\left(y_i|x_i;\theta_0\right) = f_\varepsilon\left(y_i - \beta x_i;\theta_0\right)$$

Densidad del término de error

- Si $\varepsilon \sim N\left(0, \sigma_0^2\right)$, ¿cuál es su densidad $f_\varepsilon\left(z; \theta_0\right)$?

1. $\varepsilon \sim N\left(0, \sigma_0^2\right) \rightarrow \frac{\varepsilon}{\sigma_0} \sim N\left(0, 1\right)$
2. $CDF_\varepsilon\left(z\right) \equiv Pr\left(\varepsilon \leq z\right) = Pr\left(\frac{\varepsilon}{\sigma_0} \leq \frac{z}{\sigma_0}\right)$
3. Por tanto, $CDF_\varepsilon\left(z\right) = \Phi\left(\frac{z}{\sigma_0}\right)$
4. La densidad de una variable continua es la primera derivada de su FDA:

$$f_\varepsilon\left(z; \theta_0\right) = \left(\frac{1}{\sigma_0}\right) \phi\left(\frac{z}{\sigma_0}\right)$$

Densidad de la muestra

- Puesto que

$$f_\varepsilon(z;\theta_0) = \left(\frac{1}{\sigma_0}\right)\phi\left(\frac{z}{\sigma_0}\right)$$

- y que

$$f_{y|x}(y_i|x_i;\theta_0) = f_\varepsilon(y_i - \beta x_i;\theta_0)$$

- y que

$$f(y_1, y_2, \ldots, y_N|x_1, \ldots, x_N;\theta_0) = f(y_1|x_1;\theta_0)f(y_2|x_2;\theta_0)\ldots f(y_N|x_N;\theta_0)$$

- entonces

$$f(y_1, y_2, \ldots, y_N|x_1, \ldots, x_N;\theta_0) = \prod_{i=1}^{N}\left\{\left(\frac{1}{\sigma_0}\right)\phi\left(\frac{y_i - \beta_0 x_i}{\sigma_0}\right)\right\}$$

La log-verosimilitud

- La verosimilitud reemplaza los valores reales de los parámetros por variables:

$$L(\beta,\sigma)=\prod_{i=1}^{N}\left\{\left(\frac{1}{\sigma}\right)\phi\left(\frac{y_i-\beta x_i}{\sigma}\right)\right\}$$

- tomando logaritmos la expresión es muy sencilla

$$log\left(L(\beta,\sigma)\right)=\sum_{i=1}^{N}\left\{log\left(\frac{1}{\sigma}\right)+log\left[\phi\left(\frac{y_i-\beta x_i}{\sigma}\right)\right]\right\}$$

- y como $\phi\left(\frac{y_i-\beta x_i}{\sigma}\right)=(2\pi)^{-\frac{1}{2}}\exp\left[-\left(\frac{y_i-\beta x_i}{\sigma}\right)^2\right]$ entonces

$$log\left(L(\beta,\sigma)\right)=Nlog\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}}-\sum_{i=1}^{N}\left(\frac{y_i-\beta x_i}{\sigma}\right)^2$$

El estimador MV: CPO

- Con respecto a β:

$$\frac{2}{\hat{\sigma}^2}\sum_{i=1}^{N} x_i\left(y_i - \hat{\beta}x_i\right) = 0$$

- es decir

$$\sum_{i=1}^{N} x_i\left(y_i - \hat{\beta}x_i\right) = 0$$

- Con respecto a σ, la CPO implica

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{\beta}x_i\right)^2$$

- MV $\hat{\beta}$ es el mismo estimador que MCO;$\hat{\sigma}^2 = \frac{N-1}{N}s^2$ es sesgado, pero su sesgo desaparece conforme aumenta N

Computando el estimador MV

- Los estimadores MV son frecuentemente fáciles de obtener (como en los ejemplos anteriores)
- A veces, sin embargo, no hay solución algebráica para el problema de maximización
- Entonces se hace necesario utilizar un procedimiento numérico de maximización

Procedimientos de maximización

El método de Newton

- Empieza con un valor inicial $\hat{\theta}^0$
- En cada iteración, $\hat{\theta}^{j+1} = \hat{\theta}^j - H^{-1}g$
  - g es la primera derivada de la verosimilitud (el gradiente)
  - H es la segunda derivada (el Hessiano)
- Comprueba si hay convergencia

- ¿Qué $\Delta\hat{\theta}$ incrementa más la aproximación cuadrática de Taylor de la función $L\left(\hat{\theta} + \Delta\hat{\theta}\right)$,

$$L\left(\hat{\theta} + \Delta\hat{\theta}\right) \simeq L\left(\hat{\theta}\right) + g\left(\hat{\theta}\right)\Delta\hat{\theta} + \tfrac{1}{2}H\left(\hat{\theta}\right)\Delta\hat{\theta}^2?$$

Métodos cuasi-Newton

- El método de Newton falla cuando el hessiano no es definido negativo
- En tales casos frecuentemente se reemplaza el hessiano por una matriz que, por construcción, sea siempre definida negativa
- Esta estrategia incluye a todos los procedimientos conocidos como "métodos cuasi-Newton"
- `gretl` utiliza uno de ellos: el algoritmo BFGS (Broyden, Fletcher, Goldfarb and Shanno)

Consistencia

Supuestos

1. Identificación en muestra pequeña: $l(\theta)$ toma valores diferentes para diferentes θ
2. Muestreo: $\frac{1}{n}\sum_i l_i(\hat{\theta})$ satisface una ley de grandes números
3. Identificación asintótica: max $l(\theta)$ proporciona una forma única para determinar el parámetro en el límite conforme aumenta n

- Bajo estas condiciones, el estimador MV es consistente

$$plim\left(\hat{\theta}^{ML}\right) = \theta_0$$

Identificación

- Estos son los supuestos cruciales que nos ayudan a explotar con éxito la propiedad de que la esperanza de la verosimilitud alcanza su máximo en el verdadero valor θ_0
- Si estas condiciones no se satisfacen, entonces habría un valor θ_1 tal que θ_0 y θ_1 generarían una distribución idéntica de los datos observados
- Entonces no podríamos distinguir entre los dos parámetros incluso con una muestra infinita...
- Y diríamos que los dos parámetros son observacionalmente equivalentes y que el modelo no está identificado

Normalidad asintótica

Supuestos

1. Consistencia
2. $l(\theta)$ es diferenciable y alcanza el máximo en un punto interior
3. Un TCL se puede aplicar al gradiente

- Bajo estas condiciones, el estimador MV es asintóticamente normal

$$n^{1/2}\left(\hat{\theta} - \theta_0\right) \to N(0, \Sigma)) \quad \text{conforme } n \to \infty$$

donde $\Sigma = -\left(plim \frac{1}{n} \sum H_i\right)^{-1}$

Eficiencia asintótica y estimación de la varianza

Si $l(\theta)$ es diferenciable y alcanza un máximo interior

- el estimador MV debe ser al menos tan eficiente como cualquier otro estimador consistente que sea insesgado asintóticamente

Estimadores consistentes de la matriz de Varianzas-Covarianzas

- Hessiano empírico: $var_H(\hat{\theta}) = -\left[\frac{1}{n}\sum H_i^{-1}(\hat{\theta})\right]^{-1}$
- BHHH, $var_{BHHH}(\hat{\theta}) = \left[\left(\frac{1}{n}\sum g_i(\hat{\theta})\right)^T \left(\frac{1}{n}\sum g_i(\hat{\theta})\right)\right]^{-1}$
- El estimador "sandwich": válido incluso si el modelo está mal especificado

Resumen

- Los estimadores MV son los valores que maximizan la verosimilitud
- Bajo supuestos generales, MV es consistente, asintóticamente normal, y asintóticamente eficiente