Semana 12. Máxima Verosimilitud

Equipo Econometría Avanzada

Universidad de los Andes

3 de noviembre de 2022

Contenido

Máxima Verosimilitud (MV)

Situación: Sean $\{Y_i\}_{i=1}^N$ una familia variables aleatorias i.i.d. tales que $Y_i \sim f(y;\theta)$, donde $f(\cdot)$ es la función de densidad (o la función de masa de probabilidad) asociada Y_i y θ es un vector de parámetros desconocido.

- Cuando suponemos que las Y_i dependen de otras variables X, entonces consideramos la distribución condicional. Esto es, suponemos que $\{Y_i\}_{i=1}^N$ es una familia variables aleatorias i.i.d. tales que $Y_i \sim f(y|X,\theta)$.
- A veces no determinamos directamente la distribución de la familia $\{Y_i\}_{i=1}^N$, sino que lo hacemos indirectamente a través de una relación funcional.

  Ejemplo:
  $$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i; \qquad \epsilon_i \sim \mathcal{N}(0,\sigma^2)$$

  Entonces
  $$Y_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2) \rightarrow f(y_i|x_i,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\beta_0-\beta_1 x_i)^2}{2\sigma^2}}$$

Dada una muestra aleatoria fija $\{(Y_i = y_i, X_i = x_i)\}_{i=1}^{N}$, la "probabilidad" conjunta de observar la muestra dado un parámetro θ se conoce como la función de verosimilitud condicional:

$$L(\theta|X, Y) = f(y_1, y_2, ..., y_N|X = x, \theta)$$

y dado que las y_i's son i.i.d. condicional en X, entonces,

$$L(\theta|X, Y) = \prod_{i=1}^{N} f(y_i|X = x, \theta)$$

Por simplicidad, en lugar de trabajar con $L(\cdot)$, solemos usar la función de log-verosimilitud $\ell(\cdot)$:

$$\ell(\theta|X, Y) = \ln(L(\theta|X, Y)) = \sum_{i=1}^{N} \ln(f(y_i|X = x, \theta))$$

Estimador de MV

Objetivo: Hallar un argumento que maximice la función de verosimilitud/log-verosimilitud. A este argumento se le conoce como un estimador de máxima verosimilitud. Si suponemos que $Y_i \sim f(y|x,\theta)$, entonces $\hat{\theta}_{MV} = \hat{\theta}(y,x)$.

Entonces, el problema de maximización es

$$\max_{\theta \in \Omega} \sum_{i=1}^{N} \ln(f(y_i|X=x,\theta))$$

Note que si $\ell(\cdot)$ es una función cóncava y alcanza un máximo al interior de Ω, entonces las condiciones de primer orden son necesarias y suficientes para encontrar el máximo. Entonces $\hat{\theta}_{MV}$ satisface

$$\sum_{i=1}^{N} \frac{\partial \ln(f(y_i|X=x,\theta))}{\partial \theta} = 0$$

Dicha ecuación puede resolverse analíticamente o usando métodos numéricos.

Propiedades de los estimadores de MV

Los estimadores de MV son ampliamente usados pues tienen algunas propiedades muy deseables. En particular, si $f(\cdot)$ y Ω son lo suficientemente regulares (Ej. cuando f pertenece a la familia exponencial de distribuciones) entonces:

- Los estimadores de máxima verosimilitud son consistentes, aunque no necesariamente insesgados.
- Asintóticamente normales:

$$\hat{\theta}_{MV} \xrightarrow{D} \mathcal{N}(\theta, [NI(\theta)]^{-1})$$

donde $I(\theta)$ es la matriz de información de Fisher. Sea $\mathbf{l}(\theta|y,x) = \ln(f(y|x,\theta))$, luego dicha matriz está dada por

$$I(\theta) = \mathbb{E}[(\mathbf{l}'(\theta|y,x))^2] \qquad (= -\mathbb{E}[\mathcal{H}_\theta(\mathbf{l}(\theta|y,x))] = \text{var}(\mathbf{l}'(\theta|y,x))$$

- Los estimadores son asintóticamente eficientes, pues su varianza asintótica es la cota inferior de Crámer-Rao.

Ejercicio - Máxima Verosimilitud

Suponga que usted tiene una muestra aleatoria $\{(Y_i, X_i)\}_{i=1}^{N}$ donde $Y_i \sim Poisson(\lambda)$ y donde λ es desconocido. Entonces, la función de masa de probabilidad del problema es:

$$P(Y_i = y|\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}$$

Ejercicio: Plantee la función de log-verosimilitud, obtenga el estimador de MV para λ y encuentre su varianza asintótica.

Solución - Ejercicio Máxima Verosimilitud

Para encontrar el EMV necesitamos resolver

$$\underset{\{\lambda \in (0,\infty)\}}{\text{máx}} \quad \sum_{i=1}^{N} \ln(P(Y_i = y_i|\lambda))$$

Notemos que

$$\mathbf{l}(\lambda|y_i) = \ln(P(Y_i = y_i|\lambda)) = -\lambda + y_i \ln(\lambda) - \ln(y_i!)$$

Por lo que la condición de primer orden del problema es (con respecto a λ):

$$-n + \frac{1}{\lambda} \sum_{i=1}^{n} y_i = 0$$

De manera que

$$\hat{\lambda}_{MV} = \bar{Y}$$

Solución - Unicidad

Observe que la segunda derivada de la log-verosimilitud es

$$\ell(\lambda_{MV}|Y)'' = -\frac{1}{\lambda^2}\sum_{i=1}^{n} y_i$$

En particular, en el punto crítico

$$\ell(\hat{\lambda}_{MV}|Y)'' = -\frac{n^2}{\sum_{i=1}^{n} y_i}$$

Así las cosas, $\ell(\lambda|Y)$ es cóncava, luego el estimador máximiza la verosimilitud y es único.

Solución - Varianza asintótica

Para calcular la varianza asintótica, basta hallar la información de Fisher. Note que

$$I(\lambda) = \text{var}(\mathbf{l}'(\lambda|y)) = \text{var}\left(-1 + \frac{y}{\lambda}\right) = \frac{1}{\lambda^2}\text{var}(Y) = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda}$$

y sabemos que $\frac{1}{N} I^{-1}(\lambda)$ es la varianza asintótica.

Algunas claridades

Dependiendo del libro, puede cambiar la definición de la matriz de información de Fisher. **Hay que leer bien cuál se usa en los textos**. Las dos más populares son:

- **Definición 1:** Tipo Score:
  La matriz de información de Fisher $I(\theta)$ está dada por

$$I(\theta) = -\mathbb{E}[\mathcal{H}_\theta(\mathbf{l}(\theta|y,x))]; \qquad \mathbf{l}(\theta|y,x) = \ln(f(y|x,\theta))$$

  donde $\mathcal{H}$ es la matriz Hessiana. La varianza asintótica del estadístico de MV es $[NI(\theta)]^{-1}$.

- **Definición 2:** Usando la log-verosimilitud:
  La matriz de información de Fisher $I(\theta)$ está dada por

$$I(\theta) = -\mathbb{E}[\mathcal{H}_\theta(\ell(\theta|y,x))]; \qquad \ell(\theta|y,x)) = \sum_{i=1}^{N} \ln(f(y_i|x_i,\theta))$$

  donde $\mathcal{H}$ es la matriz Hessiana. La varianza asintótica del estadístico de MV es $[I(\theta)]^{-1}$.

Una aplicación de MV

Attanasio et al. (2018) utilizan datos de niños y niñas de comunidades vulnerables de la intervención que mejoró la calidad de los hogares comunitarios FAMI.

- Intervención en **hogares comunitarios** que mejoró la calidad de los servicios ofrecidos.
- La intervención:
  - Se enfocó en la crianza.
  - Tuvo un componente nutricional.
  - Se capacitó, supervisó y orientó a las madres comunitarias.
- **Aleatoriamente** se asignaron 87 municipios al grupo de tratamiento.

Preguntas de investigación

$$Attrition_i = \begin{cases} 0 & \text{si } i \text{ fue encuestado en el seguimiento} \\ 1 & \text{si } i \textbf{ no} \text{ fue encuestado en el seguimiento} \end{cases}$$

1. ¿Cuál es el efecto de la asignación al tratamiento (T_i) sobre la atrición?

$$Pr\left(Attrition_i = 1\right) = f\left(\alpha, T_i\beta, X_i\gamma\right)$$

- ¿Puede darse una respuesta causal a (1)?

Elección Binaria

Los modelos de elección binaria se utilizan para caracterizar la respuesta de una variable Y_i que es binaria frente a cambios de otras variables X. Los modelos más utilizados son:

- **MPL**: $\mathbb{P}(Y_i = 1|X) = D_i\beta + X_i\gamma$
- **Logit**: $\mathbb{P}(Y_i = 1|X) = \Lambda(D_i\beta + X_i\gamma)$
- **Probit**: $\mathbb{P}(Y_i = 1|X) = \Phi(D_i\beta + X_i\gamma)$

MPL es estimado mediante MCO. Los modelos Probit y Logit son estimados mediante Máxima Verosimilitud.

Dada la forma funcional de $\mathbb{P}(Y_i = 1|X)$, el efecto marginal:

- **MPL**: Es constante (β_k).
- **Logit**: Varía entre $i's$ ($\Lambda'(D_i\beta + X_i\gamma)\beta_k$).
- **Probit**: Varía entre $i's$ ($\Phi'(D_i\beta + X_i\gamma)\beta_k$).

En estos últimos dos casos, en vez de reportar un efecto marginal para cada individuo, pueden ser reportados **(1)** el efecto marginal promedio, **(2)** el efecto marginal de una unidad promedio/representativa o **(3)** el efecto marginal de una unidad específica.

Tipos de efectos marginales

Los EMa pueden variar a través de i. Sin pérdida de generalidad, suponga que se busca el efecto marginal de una x_k continua.

- **Efecto marginal en el promedio**

$$\left.\frac{\partial \widehat{Pr}\left(Attrition_i = 1\right)}{\partial x_k}\right|_{X_i = \bar{X}}$$

- **Efecto marginal promedio**

$$\frac{1}{N}\sum_i \frac{\partial \widehat{Pr}\left(Attrition_i = 1\right)}{\partial x_k}$$

- **Efecto marginal en un valor específico**

$$\left.\frac{\partial \widehat{Pr}\left(Attrition_i = 1\right)}{\partial x_k}\right|_{X_i = X_0}$$

Elección Binaria

1. **Modelo de probabilidad lineal (MPL):**

   ¿Se puede estimar un MCO cuando la variable dependiente es binaria? ¡Sí! (ver vídeo)

$$Attrition_i = \alpha + T_i\beta + X_i\gamma + \varepsilon_i$$

Efectos marginales

- EMa de T:

$$E\left[Attrition_i | T_i = 1, X\right] - E\left[Attrition_i | T_i = 0, X\right] = \beta$$

- EMa de x_k continuo:

$$\frac{\partial E\left[Attrition_i | X_{-x_k}, T\right]}{\partial x_k} = \gamma_k$$

Posibles problemas:

- Modelo es heterocedastico por construcción.
- Probabilidades predichas no están acotadas entre 0 y 1.

No obstante, estos **NO** son problemas asociados a la identificación causal (ver vídeo o leer tweet)

3. **Logit:** [*Supuesto*] Los términos de error de las utilidades aleatorias siguen una distribución valor extremo tipo I o, equivalentemente, el término de error de la utilidad normalizada sigue una distribución logística.

   Así, estimamos los parámetros de interés mediante modelos de máxima verosimilitud, tal que:

$$Pr\left(Attrition_i = 1\right) = \Lambda\left(\alpha + T_i\beta + X_i\gamma\right) = \frac{e^{(\alpha + T_i\beta + X_i\gamma)}}{1 + e^{(\alpha + T_i\beta + X_i\gamma)}}$$

   donde $\Lambda(\cdot)$ es la función de distribución acumulada asociada a una distribución logística.

   Efectos marginales

   - EMa de T:

$$\Lambda\left(\alpha + \beta + X_i\gamma\right) - \Lambda\left(\alpha + X_i\gamma\right)$$

   - EMa de x_k continuo:

$$\frac{\partial Pr\left(Attrition_i = 1\right)}{\partial x_k}$$

   Por definición, $\widehat{Pr}\left(Attrition_i = 1\right) \in [0, 1]$.

2. **Probit:** [*Supuesto*] Los términos de error de las utilidades aleatorias normalizadas siguen una distribución normal estándar. Estimamos las probabilidades predichas mediante modelos de máxima verosimilitud, tal que:

$$Pr\left(Attrition_i = 1\right) = \Phi\left(\alpha + T_i\beta + X_i\gamma\right)$$

donde $\Phi(\cdot)$ es la función de distribución acumulada asociada a una distribución normal estándar.

Efectos marginales

- EMa de T:

$$\Phi\left(\alpha + \beta + X_i\gamma\right) - \Phi\left(\alpha + X_i\gamma\right)$$

- EMa de x_k continuo:

$$\frac{\partial Pr\left(Attrition_i = 1\right)}{\partial x_k}$$

Por definición, $\widehat{Pr}\left(Attrition_i = 1\right) \in [0, 1]$.

Elección Binaria: MPL vs. Probit y Logit

Se ha argumentado que en caso de que la variable de resultado es binaria, se debería usar Probit y Logit versus MPL. Los argumentos para esto son:

- Las predicciones del modelo pueden no estar en $[0, 1]$.
- El modelo es heterocedástico por construcción.

Dado que la heterocedasticidad la podemos corregir con errores robustos, el principal problema es el de predicción. Se ha mostrado que si el objetivo es calcular los efectos marginales, entonces los modelos dan resultados parecidos, pero MPL tiene la ventaja de que es más robusto (no necesita suponer distribuciones del error). Así las cosas, hay relativo consenso de que MPL es superior para **buscar efectos marginales**, mientras que Probit y Logit para **predicción**.

Gracias!