Maxima Verosimilitud

Walter Sosa Escudero

wsosa@udesa.edu.ar

Universidad de San Andrés

Introduccion

- Supongamos que $Y_i \sim N(\mu, \sigma^2 = 1)$, $i = 1, \dots, n$ y que Y_i toma los siguientes valores: 24, 26, 28, 31, 22, 26.
- Como justificariamos que 43 es una mala estimacion de μ ?
- Como justificariamos que 26 no es tan mala?

El metodo de *maxima verosimilitud* intenta buscar un estimador para μ utilizando la siguiente logica:

1. Los datos fueron generados con $N(\mu, \sigma^2 = 1)$.
2. Para los datos que dispongo, cual es el valor de μ que es *mas compatible* con el hecho de que 1) sea cierto.

Estamos intentando recorrer el camino inverso al cual los datos fueron generados. Para generar los datos primero fijamos μ y luego los generamos. Para estimar μ , primero vemos los datos y luego buscamos cual fue el μ que los pudo haber generado.

Maxima verosimilitud = maxima compatibilidad entre el modelo y los datos.

Verosimilitud

X , una variable aleatoria continua $\sim f(x; \theta_0)$. $\theta_0 \in \mathfrak{R}^K$.

$\tilde{X} = (X_1, X_2, \dots, X_n)'$ una muestra aleatoria iid $f(x; \theta_0)$.

- Que mide la *funcion de densidad*, $f(x; \theta_0) : \mathfrak{R} \rightarrow \mathfrak{R}$?
- La *funcion de verosimilitud* de X se define como:

$$L(\theta; x) : \mathfrak{R}^K \rightarrow \mathfrak{R} = f(x; \theta)$$

Análogamente, la función de verosimilitud muestral es:

$$\begin{aligned} L(\theta, \tilde{X}) &= f(X_1, \dots, X_n, \theta) \\ &= \prod_{i=1}^n f(X_i, \theta) \\ &= \prod_{i=1}^n L(\theta, X_i) \end{aligned}$$

Definamos $l(\theta, X) \equiv \ln L(\theta, X)$. Es trivial mostrar que:

$$l(\theta, \tilde{X}) = \sum_{i=1}^n l(\theta, X_i)$$

Ahora necesitamos definir algunos objetos relacionados con estas funciones

Score:

$$s(\theta, X) \equiv \frac{dl(\theta, X)}{d\theta}$$

Notar que se trata de un vector columna de K derivadas.

En forma analoga, el score de la muestra aleatoria es:

$$s(\theta, \tilde{X}) = \frac{dl(\theta; \tilde{X})}{d\theta} = \frac{d \sum_{i=1}^n l(\theta, X_i)}{d\theta} = \sum_{i=1}^n \frac{dl(\theta, X_i)}{d\theta} = \sum_{i=1}^n s(\theta, X_i)$$

que tambien es un vector columna de K elementos.

Matriz de informacion:

$$I(\theta, X) \equiv E[s(\theta, X)s(\theta, X)'],$$

la esperanza es tomada con respecto a la distribución de X .

$I(\theta)$ es una matriz $K \times K$.

La matriz de informacion muestral sera:

$$\begin{aligned} I(\theta, \tilde{X}) &= E[s(\theta, \tilde{X})s(\theta, \tilde{X})'] \\ &= E \left[\left(\sum_{i=1}^n s(\theta, X_i) \right) \left(\sum_{i=1}^n s(\theta, X_i) \right)' \right] \\ &= \sum_{i=1}^n E [s(\theta, X_i)s(\theta, X_i)'] && \text{porque?} \\ &= n I(\theta) && \text{porque?} \end{aligned}$$

Matriz Hessiana:

$$H(\theta; X) \equiv \frac{d^2 l(\theta, X)}{d\theta d\theta} = \frac{dS(\theta, X)}{d\theta}$$

Notar que se trata de una matrix $K \times K$.

Para la muestra, usando el resultado anterior, esta matriz sera:

$$H(\theta, \tilde{X}) = \frac{d^2 l(\theta, \tilde{X})}{d\theta d\theta} = \frac{d[\sum_{i=1}^n S(\theta, X_i)]}{d\theta} = \sum_{i=1}^n H(\theta, X_i)$$

Maxima Verosimilitud

Dada una muestra aleatoria $X_i \sim f(x; \theta_0); i = 1, \dots, n$, el *estimador máximo-verosímil (MV)* de θ_0 , denotado como $\hat{\theta}$ es aquel que maximiza la verosimilitud muestral:

$$\hat{\theta} \equiv \operatorname{argmax}_{\theta} L(\theta, \tilde{X})$$

Intuitivamente, $\hat{\theta}$ es el valor de los parámetros para los cuales la probabilidad de haber generado la muestra observada es máxima.

Es fácil notar que:

$$\hat{\theta} = \operatorname{argmax}_{\theta} l(\hat{\theta}, \tilde{X}) \quad \text{Porque?}$$

- Si $l(\theta, \tilde{X})$ es diferenciable y tiene un máximo local en un punto interior el estimador máximo verosímil $\hat{\theta}$ satisface:

$$s(\hat{\theta}, \tilde{X}) = 0$$

que es un sistema de K ecuaciones no-lineales cuya solución define *implicitamente* al estimador máximo verosímil.

- Si la función de verosimilitud es globalmente cóncava, encontrar el estimador MV equivale a encontrar la solución de esta ecuación.

Tres ejemplos

Distribucion de Poisson

$Y \sim \text{Poisson}(\mu)$ si toma valores enteros y positivos (incluyendo el cero) y:

$$f(y) = \text{Pr}(Y = y) = \frac{e^{-\lambda_0} \lambda_0^y}{y!}$$

Para una muestra aleatoria iid:

$$L(\lambda, \tilde{Y}) = \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{y_i}}{y_i!}$$

cuyo logaritmo sera:

$$l(\lambda, \tilde{Y}) = \sum_{i=1}^n [-\lambda + y_i \ln \lambda - \ln y_i!]$$

$$= -n\lambda + \ln \lambda \sum_{i=1}^n y_i - \sum_{i=1}^n \ln y_i!$$

El score de la muestra sera:

$$s(\lambda, \tilde{Y}) = -n + \frac{1}{\lambda} \sum_{i=1}^n y_i$$

de modo que el estimador maximo-verosimil de λ es:

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^n y_i = \bar{y}$$

Elección binaria:

$Y \sim \text{Bernoulli}(p)$, $p \equiv \text{Pr}(Y = 1|x) = F(x'\beta)$. Cuando $F(\cdot)$ es la función acumulada de una variable aleatoria normal estandar, corresponde al caso *probit*, y cuando $F(\cdot)$ es la función de distribución logística estandar, corresponde al caso *logit*.

La función de verosimilitud muestral será:

$$L(\beta, \tilde{Y}) = \prod_{i/y_i=1} p_i \prod_{i/y_i=0} (1 - p_i) = \prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1-y_i}$$

de modo que:

$$l(\beta, \tilde{Y}) = \sum_{i=1}^n [y_i \ln F(x'_i \beta) + (1 - y_i) \ln (1 - F(x'_i \beta))]]$$

Las condiciones de primer orden para un maximo local son:

$$\sum_{i=1}^n \frac{(y_i - F_i) f_i x_i}{F_i(1 - F_i)} = 0$$

, $F_i \equiv F(x'_i \hat{\beta})$, $f_i \equiv f(x'_i \hat{\beta})$. Notar que, tanto para el caso probit como el logit, se trata de un sistema de K ecuaciones no-lineales con K incognitas. Mas aun, no es posible despejar $\hat{\beta}$ y obtener una solucion explicita.

Modelo de regresion normal:

$y = x'\beta_0 + u$ con $u \sim N(0, \sigma^2)$ O, alternativamente:

$$y \sim N(x'\beta_0, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left[-\frac{1}{2} \left(\frac{y - x'\beta}{\sigma} \right)^2 \right]$$

$$l(\beta, \sigma^2; \tilde{Y}) = -n \ln \sigma - n \ln \sqrt{2\pi} - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - x'_i \beta)^2$$

Cuanto da $\hat{\beta}$?

Propiedades asintoticas: aspectos preliminares

Desigualdad de la verosimilitud:

Sea X una variable aleatoria continua con densidad $f(x; \theta_0)$.

Supongamos que: i) Si $\theta \neq \theta_0$ entonces $f(x; \theta) \neq f(x; \theta_0)$

(supuesto de identificacion), ii) $E[|l(X; \theta)|] < \infty$. Entonces,

$$E[l(\theta; X)] < E[l(\theta_0; X)]$$

La esperanza del log de la funcion de verosimilitud tiene un maximo unico en θ_0 .

Prueba:

$$\begin{aligned} E[l(\theta_o)] - E[l(\theta)] &= E[l(\theta_o) - l(\theta)] \\ &= E \left[-\ln \frac{f(X, \theta)}{f(X, \theta_o)} \right] \\ &> -\ln E \left[\frac{f(X; \theta)}{f(X; \theta_o)} \right] && \text{Desigualdad de Jensen} \\ &> -\ln \int_{\mathfrak{R}} \frac{f(x; \theta)}{f(x; \theta_o)} f(x; \theta_o) dx \\ &> -\ln 1 \\ &> 0 \end{aligned}$$

Lema 1: Sea X una variable aleatoria con densidad $f(x; \theta_0)$.
Entonces $E(s(\theta_0, X)) = 0$.

Prueba: El resultado a demostrar es:

$$E(s(\theta_0, X)) = \int_{\mathfrak{R}} s(\theta_0; x) f(x, \theta_0) dx = 0 \quad (1)$$

Para cualquier θ es cierto que :

$$\int_{\mathfrak{R}} f(x, \theta) dx = 1$$

Derivando con respecto a θ :

$$\frac{d[\int_{\mathfrak{R}} f(x, \theta) dx]}{d\theta} = 0$$

Si es posible intercambiar las operaciones de integración y diferenciación:

$$\int_{\mathfrak{R}} \frac{df(x, \theta)}{d\theta} dx = 0$$

De acuerdo a la definición del score, para una realización x de X :

$$s(\theta, X) = \frac{d \ln f(x, \theta)}{d\theta} = \frac{df(x, \theta)/d\theta}{f(x, \theta)}$$

por lo que

$$df(x, \theta)/d\theta = s(\theta; x)f(x, \theta)$$

reemplazando este resultado:

$$\int s(\theta; x)f(x; \theta) dx = 0$$

Porque la prueba NO termina aqui?

$$\int s(\theta; x) f(x; \theta) dx = 0$$

Cuando $\theta = \theta_o$, la expresion del lado izquierdo es $E(S(\theta_o, X))$, que es lo que se queria demostrar.

Este resultado implica que para θ_0 , la información $I(\theta_0)$ es igual la varianza del score:

$$V[s(\theta_0, X)] = E[s(\theta_0, X)s(\theta_0, X)'] = I(\theta_0, X)$$

Lema 2: (*Information equality*) Si $X \sim f(x; \theta_0)$, entonces:
 $I(\theta_0, X) = -E(H(\theta_0, X))$

Prueba: del resultado anterior:

$$\int_{\mathfrak{R}} s(\theta, x) f(x, \theta) dx = 0$$

Derivando ambos miembros con respecto a θ , utilizando la regla del producto y omitiendo los argumentos de las funciones:

$$\begin{aligned} \int_{\mathfrak{R}} (s f' + s' f) dx &= 0 \\ \int_{\mathfrak{R}} s f' dx + \int_{\mathfrak{R}^n} s' f dx &= 0 \end{aligned}$$

De la prueba del Lema 1, $f' = sf$, reemplazando f' arriba:

$$\int_{\mathfrak{R}} s(\theta, X)s(\theta, X)' f(x, \theta)dx + \int_{\mathfrak{R}} s(\theta, X)' f(x, \theta)dx = 0$$

cuando $\theta = \theta_0$

$$\begin{aligned} E(s(\theta_0, X)s(\theta_0, X)') + \int_{\mathfrak{R}^n} H(\theta_0; x)dx &= 0 \\ I(\theta_0, X) + E(H(\theta_0; X)) &= 0 \end{aligned}$$

lo que implica el resultado deseado.

Desigualdad de Cramer-Rao asintotica: Si $X \sim f(x, \theta_0)$, sea $\tilde{\theta}$ un estimador consistente de θ_0 y sea $V(\tilde{\theta})$ su varianza asintotica, entonces $V(\tilde{\theta}) \geq I(\theta_0, \tilde{X})^{-1} = [nI(\theta_0, X)]^{-1}$

- Esto establece un limite inferior a la varianza asintotica de un estimador consistente, el cual esta dado por la inversa de la informacion muestral.
- Si probamos que un estimador consistente tiene varianza igual a $I(\theta_0, \tilde{X})^{-1}$, ¿que es lo que hemos probado?

Propiedades asintoticas: consistencia

Queremos probar $\hat{\theta}_n \xrightarrow{p} \theta_0$.

La prueba se basa en los siguientes cuatro resultados:

1. $\hat{\theta}_n = \operatorname{argmax}_{\theta} l(\theta, \tilde{X})/n$
2. $l(\theta, \tilde{X})/n = (1/n) \sum_{i=1}^n l(\theta, X_i) \xrightarrow{p} E[l(\theta, X)]$.
3. $\operatorname{argmax}_{\theta} E[l(\theta, X)] = \theta_0$
- 4.

$$\frac{l(\theta)}{n} \xrightarrow{p} E[l(\theta, X)] \implies \operatorname{argmax}_{\theta} \frac{l(\theta)}{n} \xrightarrow{p} \operatorname{argmax}_{\theta} E[l(\theta, X)]$$

1. Simplemente la definicion del estimador MV, obtenido como maximizador de la funcion $l(\theta)/n$ (notar que maximizar $l(\theta)$ es lo mismo).
2. La funcion a maximizar tiene un limite $E[l(\theta, X)]$ 'bien definido'. La prueba pasa por usar 'bien' la ley de grandes numeros.
3. Esta funcion limite es maximizada por θ_0 . Ya lo probamos...
Cuando?
4. Si una sucesion de funciones converge a una funcion limite, entonces la sucesion de maximizadores de estas funciones converge al maximizador de la funcion limite. Este ultimo es el que es complejo probar. Ver Newey y McFadden (1994).

Propiedades asintóticas: normalidad asintótica

$$\sqrt{n} (\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I(\theta_0, X)^{-1})$$

Prueba: También daremos alguna intuición acerca de la prueba de este resultado. El estimador máximo verosímil satisface:

$$s(\hat{\theta}, \tilde{X}) = 0$$

Consideremos una expansión de Taylor de $s(\hat{\theta})$ en θ_0 :

$$s(\hat{\theta}, \tilde{X}) \approx s(\theta_0, \tilde{X}) + H(\theta_0, \tilde{X})(\hat{\theta} - \theta_0) = 0$$

Esta aproximación anterior tiende a ser exacta cuando $\hat{\theta} \rightarrow \theta_0$, lo cual, de acuerdo al resultado anterior, ocurre cuando $n \rightarrow \infty$.

Despejando y multiplicando por $\sqrt{n}$:

$$\sqrt{n} (\hat{\theta} - \theta_0) = \left(-\frac{H(\theta_0, \tilde{X})}{n} \right)^{-1} \left(\frac{s(\theta_0, \tilde{X})}{\sqrt{n}} \right)$$

Comentario: a partir de aqui, la prueba es identica a la que hicimos con el estimador MCO: mostraremos que el primer factor converge a una constante, y que el segundo tiene distribucion normal asintotica.

$$\sqrt{n} (\hat{\theta} - \theta_0) = \left(-\frac{H(\theta_0, \tilde{X})}{n} \right)^{-1} \left(\frac{s(\theta_0, \tilde{X})}{\sqrt{n}} \right)$$

Con respecto al primer factor, recordar que:

$$\frac{H(\theta_0, \tilde{X})}{n} = \frac{\sum_{i=1}^n H(\theta_0, X_i)}{n} \xrightarrow{p} E[H(\theta_0, X)]$$

ya que la muestra es iid, usando la ley de grandes numeros elemento por elemento. De acuerdo al Lema 2:

$$-E[H(\theta_0, X)] = I(\theta_0, X)$$

Entonces, por continuidad de la inversion de matrices:

$$\left(-\frac{H(\theta_0, \tilde{X})}{n} \right)^{-1} \xrightarrow{p} I(\theta_0, X)^{-1}$$

Con respecto al segundo factor, notar que:

$$\frac{s(\theta_0, \tilde{X})}{\sqrt{n}} = \sqrt{n} \frac{s(\theta_0, \tilde{X})}{n}$$

Por Cramer-Wold es lo mismo establecer la distribución de:

$$\sqrt{n} c' \left(\frac{s(\theta_0, \tilde{X})}{n} \right)$$

para cualquier $c \in \mathfrak{R}^K$. Entonces

$$\sqrt{n} c' \left(\frac{s(\theta_0, \tilde{X})}{n} \right) = \sqrt{n} c' \frac{\sum s(\theta_0, X_i)}{n} = \sqrt{n} \frac{\sum c' s(\theta_0, X_i)}{n} \equiv \sqrt{n} \frac{\sum z_i}{n}$$

con $z_i \equiv c' s(\theta_0, X_i)$.

$$\sqrt{n} \frac{\sum z_i}{n}, \quad z_i \equiv c' s(\theta_0, X_i)$$

Es facil chequear que:

- $E(z_i) = 0$ por el Lema 1 e iid.
- $V(z_i) = c' I(\theta_0) c < \infty$, por el Lema 2 e iid.

Entonces, por el TCL aplicado a $\sqrt{n}\bar{z}$:

$$\sqrt{n}(\bar{z} - 0) = c' \left(\frac{s(\theta_0, \tilde{X})}{\sqrt{n}} \right) \xrightarrow{d} N(0, c' I(\theta_0, X) c)$$

Por Cramer-Wald:

$$\left(\frac{s(\theta_0, \tilde{X})}{\sqrt{n}} \right) \xrightarrow{d} N(0, I(\theta_0, X))$$

Reemplazando:

$$\sqrt{n} \left(\hat{\theta}_n - \theta_0 \right) = \left(- \frac{H(\theta_0, \tilde{X})}{n} \right)^{-1} \left(\frac{s(\theta_0, \tilde{X})}{\sqrt{n}} \right)$$
$$\xrightarrow{p} I(\theta_0, X)^{-1} \xrightarrow{d} N(0, I(\theta_0, X))$$

Por el Teorema de Slutsky y propiedad de linealidad la distribución normal:

$$\sqrt{n} \left(\hat{\theta}_n - \theta_0 \right) \xrightarrow{p} N \left(0, I(\theta_0, X)^{-1} I(\theta_0, X) I(\theta_0, X)^{-1} \right) = N(0, I(\theta_0, X)^{-1})$$

Eficiencia asintótica e invariancia

- *Eficiencia asintótica*: $\hat{\theta}$ es asintóticamente eficiente dentro de la clase de estimadores consistentes y uniformemente asintóticamente normales. Prueba: la matriz de varianza asintótica coincide con la cota inferior de Cramer-Rao.
- *Invariancia*: Si $\hat{\theta}$ es el estimador MV de θ_0 , entonces el estimador MV de $h(\theta_0)$ es $h(\hat{\theta})$. La prueba de este resultado, si bien sencilla, también será omitida.

Estimacion de la varianza asintotica

Anteriormente probamos el resultado de normalidad asintotica:

$$\sqrt{n} (\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I(\theta_0)^{-1})$$

el cual es potencialmente util para realizar tests de hipotesis. El problema es que, en general, $I(\theta_0)^{-1}$ dependera de θ_0 , que es desconocido.

Entonces, la 'solucion' (usando el Teorema de Slutsky y el de continuidad) pasa por encontrar un estimador consistente para $I(\theta_0)$. Dos metodos consistentes.

1. *Matriz Hessiana Promedio:*

$$\hat{I}_H = -\frac{\sum_{i=1}^n H(\hat{\theta}_n, X_i)}{n}$$

es facil probar (algo muy similar hicimos antes), que $\hat{I}_H \xrightarrow{p} I(\theta_0)$, en base a la ley de grandes numeros.

2. *Producto Externo del Gradiente (OPG):*

$$\hat{I}_{OPG} = \frac{\sum_{i=1}^n s(\hat{\theta}_n, X_i)s(\hat{\theta}_n, X_i)'}{n}$$

que tambien es muy facil probar que $\hat{I}_{OPG} \xrightarrow{p} I(\theta_0)$ por la ley de grandes numeros.

Apendice: condiciones de regularidad

1. $x_i, i = 1, \dots, n, iid \sim f(x_i; \theta_0)$
2. Si $\theta \neq \theta_0 \Rightarrow f(x'_i; \theta) \neq f(x_i; \theta_0)$.
3. $\theta \in \Theta, \Theta$ es compacto.
4. $\ln f(x_i; \theta)$ continua para cada $\theta \in \Theta$ con probabilidad uno.
5. $E[\sup_{\theta} |\ln f(x; \theta)|] < \infty$.

6. θ_0 es un punto interior de Θ .
 7. $f(x; \theta)$ es dos veces diferenciable y estrictamente positiva en un entorno de θ_0 .
 8. $\int \sup_{\theta} \nabla_{\theta} f(x; \theta) dx < \infty$ y $\int \sup_{\theta} \nabla_{\theta\theta} f(x; \theta) dx < \infty$
 9. $I(\theta_0, x) \equiv E(s(\theta_0, x)s(\theta_0, x)')$ existe y es no-singular.
 10. $E[\sup_{\theta} ||H(x; \theta)||] < \infty$.
1. a 5. son necesarios para consistencia, el resto hace tambien falta para normalidad asintotica. Ver Newey y McFadden (1994) para mas detalles.