Fundamentos de Econometría

Ignacio Lobato

ITAM

Introducción

- En el capítulo anterior estudiamos la estimación de parámetros en contextos en los cuales no se conoce la población.
- En esta sección, estudiaremos el caso en el cual la forma de la distribución poblacional es conocida, excepto por algunos parámetros que nos interesa estimar.
- Supongamos que la variable aleatoria Y tiene como densidad poblacional a $f(Y;\theta)$, donde la función es conocida excepto por el parámetro θ.
- En las siguientes secciones definiremos algunas variables para poder estimar nuestro parámetro de interés θ.

La Función de Verosimilitud y el Score

- Dos variables importantes en esta sección serán la función de verosimilitud y la variable de score.

  Log-Likelihood: Es el logaritmo de $f(y;\theta)$:

  $$L = log(f(Y;\theta)) = L(Y,\theta);$$

  Score: Se define como la derivada del log-likelihood:

  $$Z = \partial log(f(Y;\theta))/\partial\theta = \partial L/\partial\theta = Z(Y;\theta)$$

- Definiendo estas variables tendremos los siguientes resultados: 1) el valor esperado del score es cero (ZES rule) y 2) la desigualdad de Cramér-Rao (CRI).

La regla ZES

- Este resultado señala que el valor esperado de la variable de score es cero.

  Demostración: Asumiremos que la variable aleatoria Y es continua.

  $$E(Z) = \int z(y;\theta)f(y;\theta)dy = \int (zf)dy$$

  como $f(y;\theta)$ es una función de densidad, debe cumplirse que:

  $$\int f(y,\theta)dy = 1 \quad \text{para todo } \theta$$

  Diferenciando en ambos lados respecto a theta:

  $$\int (\partial f(y,\theta)/\partial\theta)dy = 0$$

  Y note que:

  $$\partial f/\partial\theta = (\partial logf/\partial\theta)f = z(y;\theta)f = zf,$$

  Por lo tanto: $\int zfdy = 0$

Desigualdad de Cramér-Rao (CRI)

- **Desigualdad de Cramér-Rao (CRI):** En una muestra aleatoria de tamaño n, de una población caracterizada por $f(y;\theta)$, si $T = h(Y_1, ..., Y_n)$ y $E(T) = \theta \ \forall \ \theta$, entonces:

$$V(T) \geq 1/[nV(Z)]$$

Demostración: Primero considere el caso $n = 1$. Aquí $T = h(Y)$ con $E(T) = \theta \ \forall\theta$. Esto es:

$$\int h(y)f(y;\theta)dy = \theta \quad \forall\theta$$

Diferenciando ambos lados:

$$\int h(y)zfdy = 1, \quad \text{que significa} \quad E(TZ) = 1$$

Como sabemos que $E(Z) = 0$, entonces C(T,Z)=1.

Desigualdad de Cramér-Rao (CRI)

De la desigualdad de Cauchy-Schwarz, sabemos que:

$$V(T)V(Z) \geq C^2(T, Z) = 1$$

Por lo tanto:

$$V(T) \geq 1/V(Z)$$

Lo cual es la CRI para n=1.Para el caso $n > 1$, sea $g = g_n(y_1, ..., y_n)$ la distribución conjunta de la muestra aleatoria.
Aquí $T = h(Y_1, ..., Y_n)$ con $E(T) = \theta \ \forall \theta$. Esto es:

$$\int ... \int h(y_1, ..., y_n) g_n(y_1, ..., y_n; \theta) dy_n ... dy_1 = \theta$$

Diferenciando ambos lados:

$$\int ... \int h(\partial g / \partial \theta) dy_n ... dy_1 = 1$$

Desigualdad de Cramér-Rao (CRI)

Además:

$$\partial g/\partial\theta = (\partial log(g)/\partial\theta)g, \quad \text{y} \quad g = \prod_{i=1}^{n} f(y_i;\theta),$$

Entonces,

$$log(g) = \sum_i logf(y_i;\theta) = \sum_i logf_i$$

Así:

$$\partial log(g)/\partial\theta = \sum_i (\partial logf_i/\partial\theta) = \sum_i z(y_i,\theta) = \sum_i z_i = n\overline{Z}$$

Entonces,

$$\int \ldots \int hn\overline{Z}gdy_n \ldots dy_1 = 1,$$

Desigualdad de Cramér-Rao (CRI)

Por lo tanto:

$$E(Tn\overline{Z}) = 1$$

y como $E(\overline{Z}) = E(Z) = 0$, entonces:

$$C(\overline{Z}, T) = 1/n$$

Luego usando la Desigualdad de Cauchy-Schwarz:

$$V(T)V(\overline{Z}) \geq \frac{1}{n^2}$$

y utilizando que $V(\overline{Z}) = V(Z)/n$, tenemos que:

$$V(T) \geq \frac{1}{nV(Z)}$$

- La CRI no nos provee de un estimador directamente, pero establece un estándar a partir del cual los estimadores insesgados deben ser evaluados.
- Por ejemplo, si encontráramos un estimador insesgado T con $V(T) = 1/nV(Z)$, podemos detener la búsqueda de un mejor estimador insesgado.

Information Rule

- Existe otra forma de plantear la CRI, definiendo lo que se conoce como la **Variable de Información**, entendida como:

$$W = -\partial Z / \partial \theta = -\partial^2 logf / \partial \theta^2$$

A partir de esta definición, se establece el siguiente resultado:

La Regla de Información: La esperanza de la variable de información es igual a la varianza del score.

Demostración: Sabemos que $E(Z) = 0 = \int zfdy$. Diferenciando respecto a θ:

$$\int (\partial zf / \partial \theta) dy = 0$$

Además,

$$\partial (zf) / \partial \theta = z(\partial f / \partial \theta) + (\partial z / \partial \theta) f = z(\partial logf / \partial \theta) f - wf$$
$$= z^2 f - wf = (z^2 - w) f$$

Information Rule

Entonces tenemos que:

$$\int (z^2 - w) f dy = 0 = E(Z^2 - W)$$

Por lo tanto:

$$E(Z^2) = V(Z) = E(W)$$

Estimación por Regla ZES

- Otra función que tiene la ZES es proveer un estimador para θ.
- Supongamos que tenemos una distribución conocida $f(y;\theta)$ excepto por el parámetro θ. Sea la variable de score $Z = z(y;\theta)$. Como $E(Z) = 0$, podemos caracterizar θ como el valor c que satisface $E[z(Y,c)] = 0$.
- Por el principio de analogía, podemos estimar θ resolviendo:

$$(1/n)\sum_i z(y_i;c) = 0$$

Entonces el estimador ZES se obtiene resolviendo $\sum_i \hat{Z}_i = 0$.

Estimación por Máxima Verosimilitud

- Existe otro enfoque (más conocido), que produce el mismo estimador dado por la regla ZES. Este es el de Máxima Verosimilitud (ML).
- Considere una población en la cual la variable aleatoria Y tiene como función de distribución a $f(y;\theta)$, con la función f conocida pero con θ desconocido.
- Bajo una muestra aleatoria de tamaño n se tendría como densidad conjunta:

  $$\mathcal{L} = g_n(y_1, ..., y_n; \theta) = \prod_i f(y_i; \theta)$$

  A esta función se le llamará función de verosimilitud para θ.
- El estimador de Máxima Verosimilitud (ML) para θ es el valor para θ que maximiza $\mathcal{L}$ en la muestra. Además note que si $\mathcal{L}$ es maximizado, su logaritmo también lo está:

  $$log\mathcal{L} = \sum_i log(f(y_i; \theta)) = \sum_i L_i$$

Diferenciando respecto a θ obtenemos:

$$\partial log \mathcal{L}/\partial\theta = \sum_i \partial L_i/\partial\theta = \sum_i Z_i = \sum_i z(Y_i;\theta).$$

E igualando a cero, se tiene que el estimador de ML resuelve:

$$\sum_i z(y_i; T) = 0$$

A partir de esto último, observe que el estimador por ML es equivalente al obtenido por la regla ZES.

Ejemplo

- Supongamos que $Y \sim Bernoulli(\theta)$. Entonces:

$$Z = (Y - \theta)/[\theta(1-\theta)]$$

Entonces las CPO nos dan:

$$\sum_i \hat{Z}_i = \sum_i (Y_i - T)/[T(1-T)] = 0,$$

Por lo tanto:

$$\sum_i Y_i = nT \quad \text{entonces} \quad T = \overline{Y}$$

Propiedades del Estimador por ML

- Existe otra analogía que conduce directamente a la maximización del logaritmo de la función de verosimilitud.
- θ puede ser caracterizado en la población como el valor c que maximiza la esperanza de la log-verosimilitud. Ahora considere:

$$D(c) = logf(y; c) - log(f; \theta) = log(f(y; c)/f(y; \theta))$$

Como el logaritmo es una función cóncava, por la desigualdad de Jensen tendremos que:

$$E[D(c)] \leq logE[f(y; c)/f(y; \theta)]$$

Con igualdad si $c = \theta$ pero

$$E[f(y; c)/f(y; \theta)] = \int [f(y; c)/f(y; \theta)]f(y; \theta)dy = \int f(y; c)dy = 1,$$

Propiedades del Estimador por ML

- Entonces tenemos que:

$$E[D(c)] \leq log(1) = 0$$

  Con igualdad si $c = \theta$. Esto nos dice que θ es el valor para c que maximiza la media de la log-verosimilitud.

- La ventaja de este enfoque adicional es que este resuelve el caso donde las CPO tienen múltiples soluciones, los cuales ocurren, por ejemplo, en los casos no lineales.

- Este problema de elección es resuelto tomando la solución que da un máximo global.

Propiedades del Estimador por ML

- Si consideramos una muestra aleatoria de una población cuya variable de Score es $Z = \partial logf(y;\theta)/\partial\theta$ y cuya variable de información es $W = -\partial Z/\partial\theta$, entonces si T es el estimador por ML de θ se cumple que:

$$T \xrightarrow{p} \theta,$$

$$\sqrt{n}(T-\theta) \xrightarrow{d} N(0,\phi^2) \quad \text{donde} \quad \phi^2 = 1/V(Z) = 1/E(W),$$

$$T \xrightarrow{d} N(\theta, \phi^2/n),$$

$$T \quad \text{es un estimador BAN de } \theta$$

- Otra propiedad interesante del estimador por ML es la **invarianza**: Si $\alpha = h(\theta)$ es una función monótona de θ y T es el estimador por ML de θ, entonces A=h(T) es el estimador por ML de α.

Estimando una Relación Poblacional

- Con lo visto anteriormente, tenemos herramientas para poder estimar la relación entre 2 variables Y y X en una población bivariada.
- En secciones anteriores vimos que cuando se tiene una población univariada, la media muestral $\overline{Y}$ es un estimador insesgado y consistente de la media poblacional μ_y y que es asintóticamente normal.
- Recuerde que este resultado puede ser obtenido al menos desde dos analogías: 1) la media poblacional es el mejor predictor constante de Y en la población y 2) μ_y es el valor constante para c que minimiza $E(U^2)$, donde $U = Y - c$.

Estimando una Relación Poblacional

- Asimismo, vimos que para estimar la proyección lineal entre dos variables de una población bivariada, era suficiente emplear los análogos muestrales de los parámetros asociados a dicha proyección, obteniéndose consistencia y normalidad asintótica.

- En este caso, la proyección lineal muestral proviene de la analogía en la muestra de minimizar $\sum_i u_i^2/n$ donde ahora $u_i = y_i - (a + bx_i)$. Esto se puede entender como la analogía de los mínimos cuadrados.

- Por otro lado, el BLP es la línea que satisface $E(U) = 0$ y $E(XU) = 0$. La proyección lineal en la muestra tiene las mismas propiedades pues hace $\sum_i u_i/n = 0$ y $\sum_i x_i u_i/n = 0$. Esto es conocido como la analogía de **variable instrumental o de condición de ortogonalidad**.

Estimando una Relación Poblacional

- Con este background, ahora exploramos la estimación de $E(Y/X) = h(X;\theta)$, donde la función $h(X;\theta)$ es conocida excepto por θ.
- En particular, exploraremos los casos en los cuales $h(X;\theta)$ es lineal y no lineal.

Estimando una Media Condicional - El Caso Lineal

- Supongamos que la función de media condicional (CEF) es lineal pero se desconocen el intercepto y la pendiente, entonces la esperanza condicional coincide con el Mejor Predictor Lineal (BLP). Esto es:

  $$E(Y/X) = \alpha + \beta X$$

  con

  $$\beta = \sigma_{xy}/\sigma_x^2 \quad \text{y} \quad \alpha = \mu_y - \beta\mu_x$$

  Las dos analogías aplican de nuevo y el estimador natural es otra vez la proyección lineal muestral $\hat{Y} = A + BX$.
- En efecto, tendremos que cuando $E(Y/X)$ es lineal, A y B son estimadores **insesgados** de α y β como se demuestra a continuación.

Estimando una Media Condicional - El Caso Lineal

- El siguiente teorema enuncia un resultado importante.

  Teorema: En una muestra aleatoria en la cual se cumple que en la población $E(Y/X) = \alpha + \beta X$, el intercepto muestral A y la pendiente muestral B son estimadores insesgados de α y β.

  Demostración: Primero re-expresamos la varianza y la covarianza muestral de una forma más conveniente:

$$S_x^2 = (1/n)\sum_i (X_i - \overline{X})^2 = \sum_i X_i^2/n - \overline{X}^2 = (1/n)\sum_i (X_i - \overline{X})X_i$$

$$S_{xy} = \sum_i X_i Y_i/n - \overline{X}\,\overline{Y} = (1/n)\sum_i (X_i - \overline{X})Y_i$$

Estimando una Media Condicional - El Caso Lineal

Además:

$$B = S_{xy}/S_x^2 = \sum_i (X_i - \overline{X})Y_i / \sum_i (X_i - \overline{X})^2$$

$$= \sum_i \left[(X_i - \overline{X}) / \sum_{h=1}^{n} (X_h - \overline{X})^2 \right] Y_i = \sum_i W_i Y_i$$

donde

$$W_i = (X_i - \overline{X}) / \sum_{h=1}^{n} (X_h - \overline{X})^2 \quad \text{(i=1,...,n)}$$

Asimismo, note que W_i satisface las siguientes propiedades:

$$\sum_i W_i = 0, \qquad \sum_i W_i X_i = 1 \qquad \sum_i W_i^2 = 1/\left[\sum_i (X_i - \overline{X})^2 \right]$$

Ahora note que W_i es una función de $X = x = (x_1, ..., x_n)'$. Entonces tenemos:

Estimando una Media Condicional - El Caso Lineal

$$E(B/x) = E(\sum_i W_i Y_i/x) = \sum_i E(W_i Y_i/x) = \sum_i w_i E(Y_i/x)$$

$$= \sum_i w_i(\alpha + \beta x_i) = \alpha \sum_i w_i + \beta \sum_i w_i x_i = \beta$$

Por lo tanto, B es independiente en media de X y $E(B) = \beta$.
Similarmente para A:

$$E(A/x) = E(\overline{Y}/x) - E(B/x)\overline{x} = (\alpha + \beta\overline{x}) - \beta\overline{x} = \alpha$$

Por lo tanto, $E(A) = \alpha$. Esto concluye la prueba.

Estimando una Media Condicional - El Caso Lineal

- Note que durante toda la demostración anterior se utilizó que $E(Y_i/x) = E(Y_i/x_i)$. Esto tiene justificación mientras trabajemos en una muestra aleatoria.
- Así, es suficiente probar que $E(Y_1/x_1, x_2) = E(Y_1/x_1)$. Note que:

$$g(y_1/x_1, x_2) = f(y_1, x_1, x_2)/f_2(x_1, x_2)$$

y

$$f(y_1, x_1, x_2) = g_1(y_1, x_1)f_1(x_2) \quad \text{y} \quad f_2(x_1, x_2) = f_1(x_1)f_1(x_2)$$

Por lo tanto.

$$g(y_1/x_1, x_2) = g_1(y_1, x_1)/f_1(x_1) = g_2(y_1/x_1)$$

Como ambas distribuciones son iguales, entonces sus momentos son iguales. Luego tenemos que: $E(Y_1/x_1, x_2) = E(Y_1/x_1)$

Estimando una Media Condicional - El Caso Lineal

- En el caso de la varianza de nuestro estimador, bajo una muestra aleatoria y con CEF lineal tenemos que:

$$V(B/x) = V\left(\sum_i W_i Y_i/x\right) = \sum_i w_i^2 V(Y_i/x) = \sum_i w_i^2 V(Y_i/x_i)$$

  Además, sabemos que:

$$V(B) = E_x[V(B/x)] + V[E(B/x)] = E_x[V(B/x)]$$

- Un caso particular de este resultado es aquel en el cual $V(Y/X)$ es constante. Este caso es conocido como *homocedasticidad*.

Estimando una Media Condicional - El Caso Lineal

- Note que en el caso de homocedasticidad, si $V(Y/X) = \sigma^2$, entonces:

  $$V(B/x) = \sum_i w_i^2 \sigma^2 = \sigma^2 \sum_i w_i^2 = \sigma^2 / \sum_i (x_i - \overline{x})^2 = (\sigma^2/n)(1/S_x^2)$$

  Entonces tenemos que:

  $$V(B) = E_x[V(B/X)] = (\sigma^2/n)E(1/S_x^2)$$

  Par n grande y fijo, esta varianza exacta es aproximadamente $(\sigma^2/n)(1/\sigma_x^2)$, la cual es en efecto la varianza asintótica de B.

- En el caso homocedástico con σ^2 y $E(1/S_x^2)$ desconocidos, podemos obtener un error estándar aproximado para B tomando la raíz cuadrada de

  $$\hat{V}(B) = S^2/(nS_x^2)$$

  con $S^2 = \sum_i e_i^2/n$ y $e_i = Y_i - A - BX_i$, pues S^2 y $1/S_x^2$ son estimadores consistentes para σ^2 y $1/\sigma_x^2$

Estimando una Media Condicional - El Caso No-Lineal

- La teoría precedente es aplicable también para casos en los cuales la CEF es no lineal.
- Por ejemplo, supongamos que $E(Y/X) = exp(\theta_1 + \theta_2 X)$ con θ_1 y θ_2 desconocidos. ¿Cómo deberíamos estimar θ_1 y θ_2?.
- Nuevamente recurriremos a las analogías de 1) mínimos cuadrados y de 2) variables instrumentales.
- **1) Mínimos Cuadrados No-Lineales:** Sabemos que la CEF es el mejor predictor. En particular, es el mejor predictor de la forma $h(X; c_1, c_2)$. En la población, θ_1 y θ_2 son los valores para c_1 y c_2 que minimizan $E(U^2)$, donde:

$$U = Y - exp(c_1 + c_2 X) = Y - h(X; c_1, c_2)$$

En la muestra tenemos entonces:

$$u_i = y_i - exp(c_1 + c_2 x_i)$$

y escogemos c_1, c_2 para minimizar $\phi = \phi(c_1, c_2) = \sum_i u_i^2$

Estimando una Media Condicional - El Caso No-Lineal

Sacando la CPO obtenemos:

$$\partial\phi/\partial c_1 = 2\sum_i u_i(\partial u_i/\partial c_1) = -2\sum_i u_i h_i$$

$$\partial\phi/\partial c_2 = 2\sum_i u_i(\partial u_i/\partial c_2) = -2\sum_i u_i h_i x_i$$

donde $h_i = h(x_i; c_1, c_2)$. Entonces tenemos:

$$\sum_i h_i u_i = 0 \quad \text{y} \quad \sum_i h_i x_i u_i = 0$$

Al estimador obtenido a partir de esto se le conoce como **Estimador por Mínimos No-Lineales (NLS)**.

Estimando una Media Condicional - El Caso No-Lineal

- **2. Variables Instrumentales:** Las desviaciones sobre la CEF tienen media cero y covarianza cero con X. Esto es, sea $U = Y - E(Y/X)$, luego $E(U) = 0$ y $E(XU) = 0$. Entonces estimamos θ_1 y θ_2 a través de los valores c_1 y c_2 que resuelven $\sum_i u_i/n = 0$ y $\sum_i x_i u_i/n = 0$.
- Estas son dos ecuaciones lineales que al ser resueltas nos dan estimadores por **Variables Instrumentales** para θ_1 y θ_2.
- Note que en este caso particular, utilizamos como instrumento a X. No obstante, podemos utilizar como instrumento a cualquier función de X por la ausencia de correlación entre U y X.

Estimando una Media Condicional - El Caso No-Lineal

- Note que tanto para el caso de NLS como para el de IV, la computación de los estimadores obtenidos en el caso No-Lineal puede ser complicada y muchas veces se deben recurrir a métodos numéricos.
- Asimismo, las soluciones obtenidas por los estimadores IV y por NLS no siempre coinciden a pesar de que ambos son consistentes (pero no insesgados).
- Para obtener la distribución asintótica de ambos estimadores, se sigue un procedimiento similar al realizado para el estimador de Máxima Verosimilitud.
- Asimismo, note que la estimación por NLS puede ser visto como un caso particular de estimación IV, pues como el producto de las desviaciones de la CEF con cualquier función de X tiene esperanza cero, entonces NLS utiliza unas funciones en particular (instrumentos específicos).

Un Ejemplo - Modelo Probit

- Para ilustrar los métodos de estimación estudiados antes, estudiaremos un modelo de elección binaria. Aquí Y es una variable binaria, y solo toma valores 0 y 1. En particular:

$$E(Y/X) = \Phi(\theta_1 + \theta_2 X),$$

  donde $\Phi(.)$ es la cdf de la normal estándar. Además $\phi(.)$ denotará la densidad de la normal estándar.
- Vamos a considerar 3 estimadores: 1) NLS, 2) IV y 3) MV.
- En el caso de NLS, se escogen c_1 y c_2 que minimizan $\sum_i u_i^2$, donde:

$$u_i = y_i - \Phi_i, \qquad \Phi_i = \Phi(c_1 + c_2 x_i)$$

Un Ejemplo - Modelo Probit

- Luego tendremos como CPO para NLS:

$$2\sum_i u_i(\partial u_i/\partial c_i) = -2\sum_i u_i\phi_i,$$

$$2\sum_i u_i(\partial u_i/\partial c_2) = -2\sum_i u_i\phi_i x_i$$

Entonces tenemos que resolver:

$$\sum_i \phi_i u_i = 0 \quad \text{y} \quad \sum_i \phi_i x_i u_i = 0$$

Un Ejemplo - Modelo Probit

- En el caso de estimación IV, buscamos valores c_1, c_2 que hacen $\sum_i u_i/n = 0$ y $\sum_i x_i u_i/n = 0$. Entonces resolvemos:

$$\sum_i u_i = 0 \quad \text{y} \quad \sum_i x_i u_i = 0$$

  Note que este sistema de ecuaciones es diferente al del NLS.

- Por último, podemos emplear también el método de MV o la regla ZES, pues conocemos la distribución condicional de Y respecto a X.
- Como Y es una variable binaria con $E(Y/X) = \Phi(\theta_1 + \theta_2 X)$, es claro que condicional en X, Y tiene una distribución Bernoulli con parámetro $F(\theta_1 + \theta_2 X)$.
- Asimismo, como los parámetros θ_1 y θ_2 no aparecen en la distribución marginal de X, solo hay que maximizar la verosimilitud condicional. De este modo, tendremos que la función objetivo es:

Un Ejemplo - Modelo Probit

$$L_i = y_i log(\Phi_i) + (1 - y_1) log(1 - \Phi_i),$$

Como tenemos 2 parámetros, tendremos 2 variables de score:

$$Z_{1i} = \partial L_i / \partial \theta_1 = (y_i / \Phi_i)\phi_i - [(1 - y_i)/(1 - \Phi_i)]\, \phi_i$$

$$= w_i (y_i - \Phi_i) = w_i u_i$$

$$Z_{2i} = \partial L_i / \partial \theta_2 = w_i x_i (y_i - F_i) = w_i x_i u_i$$

donde $w_i = \phi_i / [\Phi_i (1 - \Phi_i]$. Como ambas variables score tienen esperanza cero escogemos c_1 y c_2 que resuelven:

$$\sum_i w_i u_i = 0 \quad \text{y} \quad \sum_i w_i x_i u_i = 0$$

Este sistema no lineal nos dará los estimadores para θ_1 y θ_2 por ML.

- Los tres estimadores serán diferentes en cada muestra pero todos serán consistentes (aunque no insesgados) para θ_1 y θ_2.
- Para elegir uno de estos estimadores, podemos basarnos en la propiedad de BAN de la estimación por ML.