Econometría

Maestría en Economía - Maestría en Econometría

Lecture 3

Agenda

1. Estimación por Máxima Verosimilitud
   - Introducción
   - Estimación
   - Propiedades
   - Máxima Verosimilitud: MCC
   - Máxima Verosimilitud: Inferencia Estadística
   - La Motivación Económica
2. Costos Económicos de Nacer con Peso Bajo
3. Modelo de Variable Dependiente Binaria e Inferencia Causal

Agenda

Definiciones

- Suponga que $x = (x_1, x_2, \ldots, x_n)'$ es una muestra aleatoria de

$$f(x_t; \theta), \quad t = 1, 2, \ldots, n$$

- Definamos:
  - Función de Verosimilitud (FV): $L(x; \theta) = f(x_1, x_2, \ldots, x_n; \theta)$
  - Logaritmo de la FV: $ln[L(x; \theta)]$
  - Función Score: $\frac{\partial ln[L(x;\theta)]}{\partial \theta}$
- El estimador de Máxima Verosimilitud (MLE) se obtiene de:

$$\hat{\theta}_{MLE} : argmax_{\theta \in \Theta} ln[L(x; \theta)]$$

Agenda

Máxima Verosimilitud: Estimación

- En el caso de que la función de verosimilitud sea diferenciable, entonces el estimador MLE se obtiene de la solución a las siguientes ecuaciones:
  a) $\frac{\partial ln[L(x;\theta)]}{\partial \theta} = 0$
  b) $H(\theta) = \frac{\partial^2 ln[L(x;\theta)]}{\partial \theta^2}|_{\theta=\hat{\theta}} < 0$
- Ejemplo: suponga que $(x_1, x_2, \ldots, x_n)$ es una muestra aleatoria de una $N(\mu, \sigma^2)$. Definamos a $\theta = (\mu, \sigma^2)'$. Entonces,

$$\begin{aligned} L(x;\theta) &= f(x_1, x_2, \ldots, x_n; \theta) = \prod_{t=1}^{n} f(x_t; \theta) \\ &= \prod_{t=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{1}{2\sigma^2}(x_t - \mu)^2] \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp[-\frac{1}{2\sigma^2}\sum_{t=1}^{n}(x_t - \mu)^2] \end{aligned}$$

Máxima Verosimilitud: Estimación

- Por lo tanto,

$$ln[L(x;\theta)] = -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{n}(x_t - \mu)^2$$

- Las condiciones de primer orden vienen dadas por,

$$\begin{aligned}
\frac{\partial ln[L(x;\theta)]}{\partial \mu} &= -\frac{1}{\sigma^2}\sum_{t=1}^{n}(x_t - \mu) = 0 \Longrightarrow \hat{\mu} = \frac{1}{n}\sum_{t=1}^{n}x_t \\
\frac{\partial ln[L(x;\theta)]}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{t=1}^{n}(x_t - \mu)^2 = 0
\end{aligned}$$

Máxima Verosimilitud: Estimación

- De la segunda ecuación tenemos,

$$\Longrightarrow \sigma^2 = \frac{1}{n}\sum_{t=1}^{n}(x_t - \mu)^2$$

$$\Longrightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{t=1}^{n}(x_t - \hat{\mu})^2$$

- Las condiciones de segundo orden son,

$$\frac{\partial^2 ln[L(x;\theta)]}{\partial \mu^2} = -\frac{n}{\sigma^2}; \quad \frac{\partial^2 ln[L(x;\theta)]}{\partial \sigma^4} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{t=1}^{n}(x_t - \mu)^2$$

y

$$\frac{\partial^2 ln[L(x;\theta)]}{\partial \mu \partial \sigma^2} = -\frac{1}{\sigma^4}\sum_{t=1}^{n}(x_t - \mu)$$

Máxima Verosimilitud: Estimación

- Evaluando las derivadas segundas en $\theta = \hat{\theta}$ tenemos:

$$\sum_{t=1}^{n}(x_t - \hat{\mu}) = 0 \Longrightarrow \frac{\partial^2 ln[L(x;\theta)]}{\partial\mu\partial\sigma^2}|_{\theta=\hat{\theta}} = 0$$

$$\sum_{t=1}^{n}(x_t - \hat{\mu})^2 = n\hat{\sigma}^2 \Longrightarrow \frac{\partial^2 ln[L(x;\theta)]}{\partial\sigma^4}|_{\theta=\hat{\theta}} = -\frac{n}{2\hat{\sigma}^4}$$

- La matriz hessiana queda entonces,

$$H(\hat{\theta}) = \begin{bmatrix} -\frac{n}{\hat{\sigma}^2} & 0 \\ 0 & -\frac{n}{2\hat{\sigma}^4} \end{bmatrix}$$

$-\frac{n}{\hat{\sigma}^2} < 0$ y determinante de $H(\hat{\theta}) = \frac{n^2}{2\hat{\sigma}^6} > 0$ y $H(\hat{\theta})$ es negativa definida.

Agenda

Máxima Verosimilitud: Propiedades

- Invariancia: Sea $g : \Theta \to \mathbb{R}^k$ y $\hat{\theta}$ el estimador MLE de θ entonces $g(\hat{\theta})$ es el MLE de $g(\theta)$.
- En general los estimadores MLE no son insesgados. Ejemplo,

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2$$

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-1)$$

- Por lo tanto,

$$E(\frac{n\hat{\sigma}^2}{\sigma^2}) = n-1 \Longrightarrow \frac{n}{\sigma^2}E(\hat{\sigma}^2) = n-1 \Longrightarrow E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2 \neq \sigma^2$$

y

$$E(\hat{\mu}) = E(\frac{1}{n}\sum_{t=1}^{n}x_t) = \mu$$

Máxima Verosimilitud: Propiedades

- Eficiencia. Teorema de Cramer-Rao: Sea $\hat{\theta}$ un estimador insesgado de $\theta = (\theta_1, \theta_2, \ldots, \theta_k)'$. Entonces bajo ciertas condiciones de regularidad,

$$Var(\hat{\theta}) - I_n(\theta)^{-1} \geq 0$$

donde

$$I_n(\theta) = E\left[\left(\frac{\partial ln[L(x;\theta)]}{\partial \theta}\right)\left(\frac{\partial ln[L(x;\theta)]}{\partial \theta}\right)'\right] = E\left[-\frac{\partial^2 ln[L(x;\theta)]}{\partial \theta \partial \theta'}\right]$$

es la matriz de información de Fisher.

- $$I_n(\theta) = -E\begin{bmatrix} \frac{\partial^2 ln[L(x;\theta)]}{\partial \theta_1^2} & \frac{\partial^2 ln[L(x;\theta)]}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 ln[L(x;\theta)]}{\partial \theta_1 \partial \theta_k} \\ \frac{\partial^2 ln[L(x;\theta)]}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 ln[L(x;\theta)]}{\partial \theta_2^2} & \cdots & \frac{\partial^2 ln[L(x;\theta)]}{\partial \theta_2 \partial \theta_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 ln[L(x;\theta)]}{\partial \theta_k \partial \theta_1} & \frac{\partial^2 ln[L(x;\theta)]}{\partial \theta_k \partial \theta_2} & \cdots & \frac{\partial^2 ln[L(x;\theta)]}{\partial \theta_k^2} \end{bmatrix}$$

Máxima Verosimilitud: Propiedades

- Por lo tanto $Var(\hat{\theta}_i) \geq I_n(\theta)^{-1}_{ii}$.
- Un estimador insesgado se dice completamente eficiente si su varianza "alcanza la cota de Cramer-Rao" $Var(\hat{\theta}_i) = I_n(\theta)^{-1}_{ii}$.
- Si el estimador es sesgado, entonces su eficiencia se calcula con el MSE

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = Var(\hat{\theta}) + Bias(\hat{\theta})^2$$

- en este caso, $MSE(\hat{\theta}) \geq CRLB(\theta)$,

$$CRLB(\theta) = \left(1 + \frac{\partial Bias(\theta)}{\partial \theta}\right)^2 / E\left[-\frac{\partial^2 ln[L(x;\theta)]}{\partial \theta^2}\right]$$

Máxima Verosimilitud: Propiedades

- Ejemplo (continuación):

$$H(\theta) = \begin{bmatrix} -\frac{n}{\sigma^2} & -\frac{1}{\sigma^4}\sum_{t=1}^{n}(x_t-\mu) \\ -\frac{1}{\sigma^4}\sum_{t=1}^{n}(x_t-\mu) & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{t=1}^{n}(x_t-\mu)^2 \end{bmatrix}$$

- Por definición, $I_n(\theta) = -E[H(\theta)]$. Entonces,

$$E(-\frac{n}{\sigma^2}) = -\frac{n}{\sigma^2}, \quad E[-\frac{1}{\sigma^4}\sum_{t=1}^{n}(x_t-\mu)] = -\frac{1}{\sigma^4}\sum_{t=1}^{n}E(x_t-\mu) = 0$$

y

$$\begin{aligned} E[\frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{t=1}^{n}(x_t-\mu)^2] &= \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{t=1}^{n}E(x_t-\mu)^2 \\ &= \frac{n}{2\sigma^4} - \frac{n\sigma^2}{\sigma^6} = -\frac{n}{2\sigma^4} \end{aligned}$$

Máxima Verosimilitud: Propiedades

- $$I_n(\theta) = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}, \quad I_n(\theta)^{-1} = \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}$$
- Como $E(\hat{\mu}) = E(\frac{1}{n}\sum_{t=1}^{n} x_t) = \mu$ podemos chequear que,

$$Var(\hat{\mu}) = \frac{1}{n^2}\sum_{t=1}^{n} Var(x_t) = \frac{\sigma^2}{n} = I_n(\theta)_{11}^{-1}$$

  y el estimador de μ es completamente eficiente.
- Para el estimador de la varianza tenemos $E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2 = \sigma^2 - \frac{1}{n}\sigma^2$.
- Por lo tanto,

$$\frac{\partial Bias}{\partial \sigma^2} = -\frac{1}{n}$$

Máxima Verosimilitud: Propiedades

- Entonces,

$$CRLB = (1 - 1/n)^2 / \frac{n}{2\sigma^4} = \frac{(n-1)^2}{n^2} \frac{2\sigma^4}{n}$$

- Además, $Var(\hat{\sigma}^2) = \frac{2(n-1)\sigma^4}{n^2}$, por lo tanto

$$MSE = \frac{2(n-1)\sigma^4}{n^2} + (-\frac{1}{n}\sigma^2)^2 \neq CRLB$$

  y el estimador de la varianza no es completamente eficiente.
- Alcanza s^2 (el estimador de la varianza de MCC) la cota de Cramer-Rao?
- Sabemos que $s^2 = \frac{1}{n-1} \sum_{t=1}^{n} (x_t - \hat{\mu})^2$ es un estimador insesgado de σ^2.

Máxima Verosimilitud: Propiedades

- Para saber si s^2 es completamente eficiente calculamos,

$$Var(\frac{(n-1)s^2}{\sigma^2}) = 2(n-1), \implies \frac{(n-1)^2}{\sigma^4} Var(s^2) = 2(n-1)$$

$$\implies Var(s^2) = \frac{2\sigma^4}{(n-1)} \neq I_n(\theta)^{-1}_{22}$$

  y el estimador de la varianza de MCC tampoco es completamente eficiente.

- **Propiedades asintóticas:** Sea $\hat{\theta}_n$ el estimador MLE de θ entonces:
  a) $\hat{\theta}_n \xrightarrow{p} \theta$.
  b) $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N[0, V(\theta)]$.
  c) $V(\theta) = I(\theta)^{-1}$.

  donde $I(\theta) = \lim_{n\to\infty}[\frac{1}{n} I_n(\theta)]$ es la matriz de información de Fisher asintótica.

Máxima Verosimilitud: Propiedades

- Estimadores de $I(\theta)$,

  (i)

$$-\frac{1}{n}\frac{\partial^2 ln[L(x,\hat{\theta})]}{\partial\theta\partial\theta'} \xrightarrow{p} I(\theta)$$

  (ii)

$$\frac{1}{n}\sum_{t=1}^{n}\left[\left(\frac{\partial ln[L(x,\hat{\theta})]}{\partial\theta}\right)\left(\frac{\partial ln[L(x,\hat{\theta})]}{\partial\theta}\right)'\right] \xrightarrow{p} I(\theta)$$

- Estimadores de $Var(\hat{\theta}_n)$

$$\left(\frac{\partial^2 ln[L(x,\hat{\theta})]}{\partial\theta\partial\theta'}\right)^{-1}$$

ó,

$$\left[\sum_{t=1}^{n}\left(\frac{\partial ln[L(x,\hat{\theta})]}{\partial\theta}\right)\left(\frac{\partial ln[L(x,\hat{\theta})]}{\partial\theta}\right)'\right]^{-1}$$

Máxima Verosimilitud: Propiedades

- Ejemplo (continuación)

$$\hat{\theta}_n = \begin{bmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{bmatrix}$$

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

- $$I(\theta) = \lim_{n\to\infty}[\frac{1}{n} I_n(\theta)] = \lim_{n\to\infty} \frac{1}{n} \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}$$

- $$\begin{bmatrix} \sqrt{n}(\hat{\mu} - \mu) \\ \sqrt{n}(\hat{\sigma}^2 - \sigma^2) \end{bmatrix} \xrightarrow{d} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix} \right]$$

Máxima Verosimilitud: Propiedades

- ó

$$\begin{bmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{bmatrix} \sim AN\left[\begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}, \begin{pmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/n \end{pmatrix}\right]$$

- Note que el estimador de la varianza es asintóticamente eficiente.

Agenda

- **Ejemplo 2: MCC**.

$$y = x\beta + u, \quad u \sim N(0, \sigma^2 I_n) \quad \theta = (\beta', \sigma^2)'$$

- $$f(y) = f(u) = (2\pi\sigma^2)^{n/2} \exp[-\frac{1}{2\sigma^2}(y - x\beta)'(y - x\beta)] \equiv L(x; \theta)$$

- $$ln[L(x; \theta)] = -\frac{n}{2} ln(2\pi) - \frac{n}{2} ln(\sigma^2) - \frac{1}{2\sigma^2}(y'y - \beta' x' x\beta - 2\beta' x' y)$$

Máxima Verosimilitud: MCC

- Condiciones de primer orden:

$$\frac{\partial ln[L(x;\theta)]}{\partial \beta} = -\frac{1}{2\sigma^2}(2x'x\beta - 2x'y) = \frac{1}{\sigma^2}(x'y - x'x\beta) = 0 \quad (1)$$

$$\frac{\partial ln[L(x;\theta)]}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(y - x\beta)'(y - x\beta) = 0 \quad (2)$$

- $$(1) \Longrightarrow x'y - x'x\beta = 0 \Longrightarrow \hat{\beta} = (x'x)^{-1}x'y \quad (3)$$

$$(2)y(3) \Longrightarrow \hat{\sigma}^2 = \frac{1}{n}(y - x\hat{\beta})'(y - x\hat{\beta}) = \frac{\hat{u}'\hat{u}}{n} \quad (4)$$

Máxima Verosimilitud: MCC

- Condiciones de segundo orden,

$$\frac{\partial^2 ln[L(x;\theta)]}{\partial\beta\partial\beta'} = -\frac{1}{\sigma^2}x'x; \quad \frac{\partial^2 ln[L(x;\theta)]}{\partial\sigma^4} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}u'u \tag{5}$$

$$\frac{\partial^2 ln[L(x;\theta)]}{\partial\beta\partial\sigma^2} = -\frac{1}{2\sigma^4}(x'y - x'x\beta) = -\frac{1}{2\sigma^4}x'(y - x\beta) = -\frac{1}{2\sigma^4}x'u \tag{6}$$

- Esperanzas,

$$E\left(\frac{\partial^2 ln[L(x;\theta)]}{\partial\beta\partial\beta'}\right) = E(-\frac{1}{\sigma^2}x'x) = -\frac{1}{\sigma^2}x'x$$

$$E\left(\frac{\partial^2 ln[L(x;\theta)]}{\partial\sigma^4}\right) = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}E(u'u) = \frac{n}{2\sigma^4} - \frac{n\sigma^2}{\sigma^6} = -\frac{n}{2\sigma^4}$$

- $$E\left(\frac{\partial^2 ln[L(x;\theta)]}{\partial\beta\partial\sigma^2}\right) = -\frac{1}{2\sigma^4}E(x'u) = 0 \tag{7}$$
- Por lo tanto, la matriz de información de Fisher muestral es,

$$I_n(\theta) = -E\left(\frac{\partial^2 ln[L(x;\theta)]}{\partial\theta\partial\theta'}\right) = \begin{bmatrix} \frac{1}{\sigma^2}x'x & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}$$

- La inversa de la matriz de información de Fisher es,

$$I_n(\theta)^{-1} = \begin{bmatrix} \sigma^2(x'x)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}$$

- Propiedades de los Estimadores
- $$E(\hat{\beta}) = \beta, \quad Var(\hat{\beta}) = \sigma^2(x'x)^{-1} = I_n(\theta)_{11}^{-1}$$

  y el estimador de MV de β es completamente eficiente.

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-k) \Longrightarrow E(\frac{n\hat{\sigma}^2}{\sigma^2}) = (n-k) \Longrightarrow E(\hat{\sigma}^2) = \frac{n-k}{n}\sigma^2 \neq \sigma^2$$

$$E(\hat{\sigma}^2) = \frac{n-k}{n}\sigma^2 = \sigma^2 - \frac{k}{n}\sigma^2 \Longrightarrow Bias(\hat{\sigma}^2) = -\frac{k}{n}\sigma^2$$

- Entonces,

$$Var(\frac{n\hat{\sigma}^2}{\sigma^2}) = 2(n-k) \Longrightarrow Var(\hat{\sigma}^2) = \frac{2(n-k)}{n^2}\sigma^4$$

- y el MSE es,

$$MSE(\hat{\sigma}^2) = \frac{2(n-k)}{n^2}\sigma^4 + (-\frac{k}{n}\sigma^2)^2 = \frac{2(n-k)+nk^2}{n^2}\sigma^4$$

- La cota de Cramer-Rao es,

$$\begin{aligned} CRLB &= \left(1 + \frac{\partial Bias(\theta)}{\partial \theta}\right)^2 / E\left[-\frac{\partial^2 ln[L(x;\theta)]}{\partial \theta^2}\right] \\ &= (1-\frac{k}{n})^2/(n/2\sigma^4) = \frac{2(n-k)^2\sigma^4}{n^3} \end{aligned}$$

El estimador de la varianza es sesgado y su MSE no alcanza la cota de Cramer-Rao.

Máxima Verosimilitud: MCC

- Es el estimador de la varianza de MCC es completamente eficiente?
- $$\frac{(n-k)s^2}{\sigma^2} \sim \chi^2(n-k); \quad s^2 = \frac{\hat{u}'\hat{u}}{(n-k)}$$
- Esperanza,
$$E(\frac{(n-k)s^2}{\sigma^2}) = n-k \Longrightarrow E(s^2) = \sigma^2$$
- Varianza,
$$Var(\frac{(n-k)s^2}{\sigma^2}) = 2(n-k) \Longrightarrow Var(s^2) = \frac{2\sigma^4}{n-k} \neq I_n(\theta)^{-1}_{22} = \frac{2\sigma^4}{n}$$

El estimador es insesgado pero tampoco es completamente eficiente.

- Propiedades Asintóticas:

$$\hat{\beta} \xrightarrow{p} \beta, \quad s^2 \xrightarrow{p} \sigma^2$$

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N[0,\, I(\theta)^{-1}], \;\; \hat{\theta} = (\hat{\beta}',\, s^2)'$$

$$I(\theta) = \lim_{n\to\infty}[\frac{1}{n} I_n(\theta)] = \begin{bmatrix} \frac{1}{\sigma^2}Q & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}, \quad Q = \lim_{n\to\infty} \frac{1}{n} x'x$$

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N[0,\, \sigma^2 Q^{-1}], \;\; \sqrt{n}(s^2 - \sigma^2) \xrightarrow{d} N[0,\, 2\sigma^4]$$

Ambos estimadores son asintóticamente insesgados y eficientes.

Máxima Verosimilitud: MCC

- MCC con variables explicativas aleatorias.

$$y = x\beta + u, \quad u \sim N(0, \sigma^2 I_n) \quad \theta = (\beta', \sigma^2)'$$

- Supuestos:
  1. x aleatorias con densidad $f(x)$.
  2. $f(x)$ no depende de θ.
  3. x y u se distribuyen en forma independiente.
-

$$L(y, x; \theta) = f(y|x)f(x) = f(u|x)f(x) = f(u)f(x)$$

- El logaritmo de la función de verosimilitud queda,

$$\begin{aligned} Ln[L(y, x; \theta)] &= Ln[f(u)] + Ln[f(x)] \\ &= -\frac{n}{2} ln(2\pi) - \frac{n}{2} ln(\sigma^2) - \frac{1}{2\sigma^2}(y - x\beta)'(y - x\beta) \\ &+ Ln[f(x)] \end{aligned}$$

Máxima Verosimilitud: MCC

- Por lo tanto,

$$\hat{\beta} = (x'x)^{-1}x'y, \quad \hat{\sigma}^2 = \frac{\hat{u}'\hat{u}}{n}, \quad I_n(\theta) = \begin{bmatrix} \frac{1}{\sigma^2}E(x'x) & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}$$

$$I(\theta) = \lim_{n\to\infty}[\frac{1}{n}I_n(\theta)] = \begin{bmatrix} \frac{1}{\sigma^2}Q & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}, \quad Q = \lim_{n\to\infty}\frac{1}{n}E(x'x)$$

- Propiedades asintóticas,

$$\hat{\beta} \xrightarrow{p} \beta, \quad \hat{\sigma}^2 \xrightarrow{p} \sigma^2$$

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N[0, I(\theta)^{-1}], \quad \hat{\theta} = (\hat{\beta}', \hat{\sigma}^2)'$$

$$I(\theta) = \lim_{n\to\infty}[\frac{1}{n} I_n(\theta)] = \begin{bmatrix} \frac{1}{\sigma^2}Q & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}, \quad Q = \lim_{n\to\infty} \frac{1}{n} E(x'x)$$

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N[0, \sigma^2 Q^{-1}], \quad \sqrt{n}(\hat{\sigma}^2 - \sigma^2) \xrightarrow{d} N[0, 2\sigma^4]$$

- Todos estos resultados asumieron normalidad del término de error. Sin este supuesto no hay garantías de que los estimadores de MV sean los de MCC o que el estimador de MCC alcance la cota de Cramer-Rao.
- Si el término de error no es normal, todo lo anterior implica que los estimadores de MV maximizan una función de verosimilitud que está mal especificada.

Cuasi (Pseudo) Máxima Verosimilitud: MCC

- En este caso, el método se denomina de Cuasi-Máxima Verosimilitud y tiene las siguientes propiedades.
- Todos los resultados de muestra finita que vimos para MCC estimados por MV pueden re-interpretarse como resultados de muestra finita del cuasi-MLE cuando el error es incorrectamente especificado como teniendo distribución normal.

Agenda

Máxima Verosimilitud: Inferencia Estadística

- Sea $L(x, \theta)$ la función de densidad conjunta del vector de variables aleatorias $x = (x_1, \ldots, x_n)'$, caracterizadas por el vector de parámetros θ de dimensión $(k \times 1)$.
- Queremos contrastar las siguientes hipótesis

$$H_0 : \phi(\theta) = 0 \quad vs. \quad H_1 : \phi(\theta) \neq 0$$

donde $\phi(\cdot)$ es un vector de dimensión $q \times 1$ de funciones diferenciables con $q < k$.

- Definamos a $\hat{\theta}$ como el estimador MLE sin restricciones, i.e. resuelve $argmax_\theta \; Ln[L(x, \theta)]$.
- Definamos a $\tilde{\theta}$ como el estimador MLE restringido, i.e. resuelve $argmax_\theta \; Ln[L(x, \theta)], \;\; s.t. \, \phi(\theta) = 0$.

Test de Wald

- Haciendo una expansión de Taylor de primer orden de $\phi(\hat{\theta})$ alrededor del verdadero vector de parámetros θ, tenemos

$$\begin{aligned} \phi(\hat{\theta}) &= \phi(\theta) + F(\theta)'(\hat{\theta} - \theta) + o_p(1) \Longrightarrow \\ \sqrt{n}\left[\phi(\hat{\theta}) - \phi(\theta)\right] &= F(\theta)'\sqrt{n}(\hat{\theta} - \theta) + o_p(1) \end{aligned} \tag{8}$$

  donde $F(\theta) = \frac{\partial \phi(\theta)'}{\partial \theta}$ con $rango\,[F(\theta)] = q$.
- De las propiedades asintóticas de los estimadores de máxima verosimilitud sabemos que:

$$\sqrt{n}(\hat{\theta} - \theta) \sim N(0,\, I(\theta)^{-1}) \tag{9}$$

- De (8) y (9) tenemos:

$$\sqrt{n}\left[\phi(\hat{\theta}) - \phi(\theta)\right] \sim N[0,\, F(\theta)' I(\theta)^{-1} F(\theta)]$$

Test de Wald

- Bajo la hipótesis nula $\phi(\theta) = 0$ por lo tanto,

$$\sqrt{n}\phi(\hat{\theta}) \sim N[0,\ F(\theta)'I(\theta)^{-1}F(\theta)] \tag{10}$$

- Usando la forma cuadrática de variables aleatorias normales tenemos

$$n\phi(\hat{\theta})'[F(\theta)'I(\theta)^{-1}F(\theta)]^{-1}\phi(\hat{\theta}) \sim \chi^2(q) \tag{11}$$

- Podemos aproximar consistentemente los términos del corchete del estadístico anterior evaluando en el estimador de máxima verosimilitud $\hat{\theta}$,

$$W = n\phi(\hat{\theta})'[F(\hat{\theta})'I(\hat{\theta})^{-1}F(\hat{\theta})]^{-1}\phi(\hat{\theta}) \sim \chi^2(q) \tag{12}$$

Test del Multiplicador de Lagrange (LM)

- Se resuelve el siguiente problema:

$$max_\theta \, Ln[L(x, \theta)] \;\; s.t. \; \phi(\theta) = 0$$

$$\mathbb{L}(\theta, \lambda) = Ln[L(x, \theta)] + \lambda' \phi(\theta)$$

- La solución satisface el siguiente conjunto de ecuaciones:

$$\begin{aligned} \frac{\partial Ln[L(x, \theta)]}{\partial \theta}|_{\tilde{\theta}} + F(\tilde{\theta})\tilde{\lambda} &= 0 \\ \phi(\tilde{\theta}) &= 0 \end{aligned} \tag{13}$$

  donde $\tilde{\theta}$ es la solución del problema de maximización condicionada.

- El tests LM está basado en la idea de que $\tilde{\lambda}$ apropiadamente ponderado tiene distribución asintótica normal.

Test del Multiplicador de Lagrange (LM)

- Tomemos una expansión de Taylor de primer orden de $\phi(\hat{\theta})$ y $\phi(\tilde{\theta})$ alrededor del verdadero vector de parámetros θ. Ignorando los términos $o_p(1)$ tenemos,

$$\sqrt{n}\phi(\hat{\theta}) = \sqrt{n}\phi(\theta) + F(\theta)'\sqrt{n}(\hat{\theta} - \theta) \quad (14)$$
$$\sqrt{n}\phi(\tilde{\theta}) = \sqrt{n}\phi(\theta) + F(\theta)'\sqrt{n}(\tilde{\theta} - \theta) \quad (15)$$

- De las condiciones de primer orden sabemos que $\phi(\tilde{\theta}) = 0$ de forma tal que restando (15) de (14) obtenemos

$$\sqrt{n}\phi(\hat{\theta}) = F(\theta)'\sqrt{n}(\hat{\theta} - \tilde{\theta}) \quad (16)$$

- Por otro lado, tomando una expansión de Taylor de primer orden de $\frac{\partial Ln[L(x,\theta)]}{\partial \theta}|_{\theta=\hat{\theta}}$ y $\frac{\partial Ln[L(x,\theta)]}{\partial \theta}|_{\theta=\tilde{\theta}}$ alrededor del verdadero vector de parámetros θ tenemos (ignorando los términos $o_p(1)$)

$$\frac{\partial Ln[L(x,\theta)]}{\partial \theta}|_{\theta=\hat{\theta}} = \frac{\partial Ln[L(x,\theta)]}{\partial \theta} + \frac{\partial^2 Ln[L(x,\theta)]}{\partial \theta \partial \theta'}(\hat{\theta}-\theta) \quad (17)$$

$$\frac{\partial Ln[L(x,\theta)]}{\partial \theta}|_{\theta=\tilde{\theta}} = \frac{\partial Ln[L(x,\theta)]}{\partial \theta} + \frac{\partial^2 Ln[L(x,\theta)]}{\partial \theta \partial \theta'}(\tilde{\theta}-\theta) \quad (18)$$

- Note que,

$$\begin{aligned}
\frac{\partial Ln[L(x,\theta)]}{\partial \theta}|_{\theta=\hat{\theta}} &= \frac{\partial Ln[L(x,\theta)]}{\partial \theta} + \frac{\partial^2 Ln[L(x,\theta)]}{\partial \theta \partial \theta'}(\hat{\theta}-\theta) \Longrightarrow \\
\frac{1}{\sqrt{n}}\frac{\partial Ln[L(x,\theta)]}{\partial \theta}|_{\theta=\hat{\theta}} &= \frac{1}{\sqrt{n}}\frac{\partial Ln[L(x,\theta)]}{\partial \theta} + \frac{1}{n}\frac{\partial^2 Ln[L(x,\theta)]}{\partial \theta \partial \theta'}\sqrt{n}(\hat{\theta}-\theta) \Longrightarrow \\
\frac{1}{\sqrt{n}}\frac{\partial Ln[L(x,\theta)]}{\partial \theta}|_{\theta=\hat{\theta}} &= \frac{1}{\sqrt{n}}\frac{\partial Ln[L(x,\theta)]}{\partial \theta} - I(\theta)\sqrt{n}(\hat{\theta}-\theta)
\end{aligned} \quad (19)$$

Test del Multiplicador de Lagrange (LM)

- De la misma manera

$$\frac{1}{\sqrt{n}}\frac{\partial Ln[L(x,\,\theta)]}{\partial \theta}|_{\theta=\tilde{\theta}} = \frac{1}{\sqrt{n}}\frac{\partial Ln[L(x,\,\theta)]}{\partial \theta} - I(\theta)\sqrt{n}(\tilde{\theta}-\theta) \tag{20}$$

- De las condiciones de primer orden de la maximización no restringida sabemos que $\frac{\partial Ln[L(x,\theta)]}{\partial \theta}|_{\theta=\hat{\theta}} = 0$. Restando (19) de (20) tenemos,

$$\frac{1}{\sqrt{n}}\frac{\partial Ln[L(x,\,\theta)]}{\partial \theta}|_{\theta=\tilde{\theta}} = -I(\theta)\sqrt{n}(\tilde{\theta}-\hat{\theta}) = I(\theta)\sqrt{n}(\hat{\theta}-\tilde{\theta}) \tag{21}$$

- Por lo tanto

$$\sqrt{n}(\hat{\theta}-\tilde{\theta}) = I(\theta)^{-1}\frac{1}{\sqrt{n}}\frac{\partial Ln[L(x,\,\theta)]}{\partial \theta}|_{\theta=\tilde{\theta}} \tag{22}$$

Test del Multiplicador de Lagrange (LM)

- De (16) y (22) tenemos

$$\sqrt{n}\phi(\hat{\theta}) = F(\theta)' I(\theta)^{-1} \frac{1}{\sqrt{n}} \frac{\partial Ln[L(x,\, \theta)]}{\partial \theta}|_{\theta=\tilde{\theta}} \tag{23}$$

- Usando (13) obtenemos

$$\begin{aligned} \sqrt{n}\phi(\hat{\theta}) &= -F(\theta)' I(\theta)^{-1} F(\tilde{\theta}) \frac{\tilde{\lambda}}{\sqrt{n}} \\ &\Longrightarrow -F(\theta)' I(\theta)^{-1} F(\theta) \frac{\tilde{\lambda}}{\sqrt{n}} \end{aligned} \tag{24}$$

- Entonces,

$$\frac{\tilde{\lambda}}{\sqrt{n}} = -\left[F(\theta)' I(\theta)^{-1} F(\theta)\right]^{-1} \sqrt{n}\phi(\hat{\theta}) \tag{25}$$

Test del Multiplicador de Lagrange (LM)

- De (10), bajo la hipótesis nula $\sqrt{n}\phi(\hat{\theta}) \sim N[0,\ F(\theta)'I(\theta)^{-1}F(\theta)]$, por lo tanto

$$\frac{\tilde{\lambda}}{\sqrt{n}} \sim N\left[0,\ \left(F(\theta)'I(\theta)^{-1}F(\theta)\right)^{-1}\right] \tag{26}$$

- Nuevamente, usando la forma cuadrática de variables normales se obtiene

$$\frac{1}{n}\tilde{\lambda}'[F(\theta)'I(\theta)^{-1}F(\theta)]\tilde{\lambda} \xrightarrow[H_0]{d}\ \chi^2(q) \tag{27}$$

- Usando (13) otra forma del estadístico LM es

$$\frac{1}{n}\frac{\partial Ln[L(x,\ \theta)]}{\partial\theta'}|_{\theta=\tilde{\theta}}I(\theta)^{-1}\frac{\partial Ln[L(x,\ \theta)]}{\partial\theta}|_{\theta=\tilde{\theta}} \xrightarrow[H_0]{d}\ \chi^2(q) \tag{28}$$

- Para que el test LM sea operativo en la práctica debemos evaluar la matriz de información de Fisher en la estimación restringida consistente, $\tilde{\theta}$.

Test del Multiplicador de Lagrange (LM)

- Podemos aproximar $I(\theta)$ con:

$$-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 Ln[L(x_i,\,\theta)]}{\partial\theta\partial\theta'}\Big|_{\theta=\tilde{\theta}} \tag{29}$$

- O con

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial Ln[L(x_i,\,\theta)]}{\partial\theta}\Big|_{\theta=\tilde{\theta}}\frac{\partial Ln[L(x_i,\,\theta)]}{\partial\theta'}\Big|_{\theta=\tilde{\theta}} \tag{30}$$

- Si elegimos la segunda aproximación el test LM queda,

$$LM=\frac{1}{n}\frac{\partial Ln[L(x,\,\tilde{\theta})]}{\partial\theta'}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial Ln[L(x_i,\,\tilde{\theta})]}{\partial\theta}\frac{\partial Ln[L(x_i,\,\tilde{\theta})]}{\partial\theta'}\right]^{-1}\frac{\partial Ln[L(x,\,\tilde{\theta})]}{\partial\theta} \tag{31}$$

Test del Multiplicador de Lagrange (LM)

- Note que,

$$\frac{\partial Ln[L(x,\tilde{\theta})]}{\partial \theta'} = \sum_{i=1}^{n} \frac{\partial Ln[L(x_i,\tilde{\theta})]}{\partial \theta'}$$

por lo que el test LM queda,

$$LM = \sum_{i=1}^{n} \frac{\partial Ln[L(x_i,\tilde{\theta})]}{\partial \theta'} \left[\sum_{i=1}^{n} \frac{\partial Ln[L(x_i,\tilde{\theta})]}{\partial \theta} \frac{\partial Ln[L(x_i,\tilde{\theta})]}{\partial \theta'}\right]^{-1} \sum_{i=1}^{n} \frac{\partial Ln[L(x_i,\tilde{\theta})]}{\partial \theta}$$

- Recuerde la definición del R^2 no centrado para el modelo $y = X\beta + u$:

$$R^2 = \frac{\hat{y}'\hat{y}}{y'y} = \frac{(X\hat{\beta})'X\hat{\beta}}{y'y} = \frac{y'X(X'X)^{-1}X'X(X'X)^{-1}X'y}{y'y} = \frac{y'X(X'X)^{-1}X'y}{y'y} \qquad (32)$$

Test del Multiplicador de Lagrange (LM)

- Defina $y = \mathbf{1} = (1\,1\,\ldots\,1)'$ al vector de unos de dimensión $n \times 1$ y $X = \left[\frac{\partial Ln[L(x_1,\tilde{\theta})]}{\partial\theta'}\ \frac{\partial Ln[L(x_2,\tilde{\theta})]}{\partial\theta'}\ \ldots\ \frac{\partial Ln[L(x_n,\tilde{\theta})]}{\partial\theta'}\right]'$ a la matriz $n \times k$ de valores de las variables explicativas.
- En la ecuación del R^2 no centrado tenemos:

$$y'X = \mathbf{1}'X = \sum_{i=1}^{n} \frac{\partial Ln[L(x_i,\,\tilde{\theta})]}{\partial\theta'}$$

y

$$X'X = \sum_{i=1}^{n} \frac{\partial Ln[L(x_i,\,\tilde{\theta})]}{\partial\theta} \frac{\partial Ln[L(x_i,\,\tilde{\theta})]}{\partial\theta'}$$

Test del Multiplicador de Lagrange (LM)

- Entonces,

$$R^2 = \frac{\mathbf{1}'X(X'X)^{-1}X'\mathbf{1}}{\mathbf{1}'\mathbf{1}} =$$

$$\frac{\sum_{i=1}^{n} \frac{\partial Ln[L(x_i,\tilde{\theta})]}{\partial \theta'} \left[\sum_{i=1}^{n} \frac{\partial Ln[L(x_i,\tilde{\theta})]}{\partial \theta} \frac{\partial Ln[L(x_i,\tilde{\theta})]}{\partial \theta'}\right]^{-1} \sum_{i=1}^{n} \frac{\partial Ln[L(x_i,\tilde{\theta})]}{\partial \theta}}{n}$$

$$\Longrightarrow LM = nR^2 \sim \chi^2(q)$$

- Este resultado sugiere que el test LM no es más que el R^2 no centrado de una regresión auxiliar de un vector de unos sobre los *scores*, evaluados en la estimación de máxima verosimilitud restringida, multiplicado por el número de observaciones.

- Estadístico del cociente de verosimilitud (LR test)

$$\mu = max_{\phi(\theta)=0} L(x, \theta) / max_{\theta} L(x, \theta) = \frac{L(x, \tilde{\theta})}{L(x, \hat{\theta})}$$

- Entonces el estadístico de contraste es,

$$LR = -2Ln(\mu) = 2\{Ln[L(x, \hat{\theta})] - Ln[L(x, \tilde{\theta})]\} \xrightarrow{d}_{H_0} \chi^2(q)$$

- Note que, haciendo una expansión de Taylor de segundo orden

$$\begin{aligned} Ln[L(x, \tilde{\theta})] &\simeq Ln[L(x, \hat{\theta})] + (\hat{\theta} - \tilde{\theta}) \frac{\partial Ln[L(x, \hat{\theta})]}{\partial \hat{\theta}} \\ &+ \frac{1}{2} (\hat{\theta} - \tilde{\theta})' \frac{\partial^2 Ln[L(x, \hat{\theta})]}{\partial \hat{\theta} \partial \hat{\theta}'} (\hat{\theta} - \tilde{\theta}) \end{aligned}$$

- La ecuación anterior implica que,

$$\begin{aligned} -2\{Ln[L(x,\hat{\theta})] - Ln[L(x,\tilde{\theta})]\} &= (\hat{\theta}-\tilde{\theta})' \frac{\partial^2 Ln[L(x,\hat{\theta})]}{\partial\hat{\theta}\partial\hat{\theta}'}(\hat{\theta}-\tilde{\theta}) \Longrightarrow \\ 2\{Ln[L(x,\hat{\theta})] - Ln[L(x,\tilde{\theta})]\} &= n(\hat{\theta}-\tilde{\theta})' I(\theta)(\hat{\theta}-\tilde{\theta}) \end{aligned}$$

- Donde hicimos uso de,

$$\frac{\partial Ln[L(x,\hat{\theta})]}{\partial\hat{\theta}} = 0$$

$$I(\theta) = -\frac{1}{n}\frac{\partial^2 Ln[L(x,\hat{\theta})]}{\partial\hat{\theta}\partial\hat{\theta}'}$$

Máxima Verosimilitud: Inferencia Estadística

- Por lo tanto, como $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N[0,\ I(\theta)^{-1}]$, se desprende que,

$$n(\hat{\theta} - \tilde{\theta})' I(\theta)(\hat{\theta} - \tilde{\theta}) \xrightarrow{d}_{H_0} \chi^2(q)$$

Agenda

Sesgo de Selección por Truncamiento Incidental

- En economía un caso emblemático de aplicación de máxima verosimilitud es el modelo de oferta de trabajo de las mujeres (Gronau, 1974; Heckman, 1976). Este modelo consiste de dos ecuaciones, una ecuación de salarios que representa la diferencia entre el salario de mercado de una persona y su salario de reserva en función de características tales como la edad, educación, número de hijos etc.
- La segunda ecuación es una ecuación de horas deseadas de trabajo que depende del salario, de la presencia de hijos pequeños en el hogar, del estado civil, etc.
- El problema del truncamiento es que en la segunda ecuación observamos las horas reales solo si la persona está trabajando. Esto es, solo si el salario de mercado excede al salario de reserva. En este caso se dice que la variable horas en la segunda ecuación está incidentalmente truncada.

Sesgo de Selección por Truncamiento Incidental

- Definiciones: Suponga que y y z tienen una distribución bivariada con correlación ρ. Nosotros estamos interesados en la distribución de y dado que z excede un determinado valor. Esto es, la función de densidad conjunta de y y z es:

$$f(y, z|z > a) = \frac{f(y, z)}{Pr(z > a)}$$

- Teorema 20.4 (Greene, 1997, Cap. 20, página 975): Si y y z tienen una distribución normal bivariada con medias μ_y y μ_z, desvíos estándar σ_y y σ_z y correlación ρ, entonces:

$$E(y|z > a) = \mu_y + \rho\sigma_y\lambda(\alpha_z)$$

$$Var(y|z > a) = \sigma_y^2[1 - \rho^2\delta(\alpha_z)],$$

donde, $\alpha_z = (a - \mu_z)/\sigma_z$, $\lambda(\alpha_z) = \phi(\alpha_z)/[1 - \Phi(\alpha_z)]$ y $\delta(\alpha_z) = \lambda(\alpha_z)[\lambda(\alpha_z) - \alpha_z]$.

Sesgo de Selección por Truncamiento Incidental

- Para poner el ejemplo de la oferta de trabajo de las mujeres en un marco general de análisis, digamos que la ecuación que determina la selección muestral es:

$$z_i^* = \gamma' w_i + u_i,$$

donde z_i^* es la diferencia entre el salario de mercado y el salario de reserva de la persona i.

- La ecuación de interés es,

$$y_i = \beta' x_i + \epsilon_i,$$

donde y_i es la oferta de trabajo (en horas) de la persona i.

- La regla es que y_i es observada solo cuando z_i^* es mayor a cero.

Sesgo de Selección por Truncamiento Incidental

- Asumamos que u_i y ϵ_i tienen distribución normal bivariada con medias cero y correlación ρ. Aplicando el teorema 20.4 tenemos,

$$\begin{aligned} E(y_i|y_i\text{es observada}) &= E(y_i|z_i^* > 0) \\ &= E(y_i|u_i > -\gamma' w_i) \\ &= \beta' x_i + E(\epsilon_i|u_i > -\gamma' w_i) \\ &= \beta' x_i + \rho\sigma_\epsilon \lambda_i(\alpha_u) \\ &= \beta' x_i + \beta_\lambda \lambda_i(\alpha_u) \end{aligned}$$

  donde $\alpha_u = -\gamma' w_i/\sigma_u$ y $\lambda_i(\alpha_u) = \phi(\gamma' w_i/\sigma_u)/\Phi(\gamma' w_i/\sigma_u)$.
- Entonces, la ecuación de interés puede escribirse como,

$$y_i|z_i^* > 0 = \beta' x_i + \beta_\lambda \lambda_i(\alpha_u) + v_i$$

  donde v_i es un término de error con media cero.

Sesgo de Selección por Truncamiento Incidental

- Como queda claro de este desarrollo, estimar por MCC la ecuación de interés usando solo los datos observados, produce estimadores inconsistentes de β por el argumento estándar de variables omitidas (i.e. estamos omitiendo $\lambda_i(\cdot)$).
- Cómo podemos obtener estimaciones consistentes de la ecuación de horas de trabajo utilizando solo los datos observados.
- En principio tenemos un problema similar al de la variable habilidad omitida en la ecuación del salario. En este caso, la variable $\lambda_i(\cdot)$ no es observada.
- Note que aún cuando observáramos $\lambda_i(\cdot)$, MCC no nos daría estimadores eficientes porque los errores de la ecuación de interés, v_i, son heterocedásticos (i.e. $Var(v_i) = \sigma_\epsilon^2(1 - \rho^2\delta_i)$ de acuerdo al teorema 20.4 de Greene).

Sesgo de Selección por Truncamiento Incidental

- Una posible solución es estimar la ecuación de selección para obtener los $\hat{\gamma}$ y construir la variable omitida como $\hat{\lambda}_i = \phi(\hat{\gamma}' w_i)/\Phi(\hat{\gamma}' w_i)$. Luego en un segundo paso estimar la ecuación de interés por MCC en una regresión de y sobre x y $\hat{\lambda}$.
- El único problema de esta solución es que la variable dependiente de la ecuación de selección, z_i^*, no es observada. Lo que podemos observar es $z_i = 1$ si $z_i^* > 0$, es decir si la persona está trabajando; o $z_i = 0$ si $z_i^* < 0$ si la persona no está trabajando.
- Es decir que el modelo que podemos estimar es:

$$z_i = \gamma' w_i + u_i, \tag{33}$$

donde z_i es una variable binaria.

Modelo de Probabilidad Lineal

- Una posibilidad es estimar la ecuación de selección por MCC. Para ver porque esta no es la mejor solución considere la esperanza condicional de la variable dependiente.

$$E(z_i|w_i) = \gamma' w_i$$

Además desde la definición de esperanza matemática $E(z_i|w_i) = 0 * (1 - P_i) + 1 * P_i = P_i$ donde P_i es la probabilidad de que $z_i = 1$. Entonces,

$$P_i = Pr[z_i = 1|w_i] = E(z_i|w_i) = \gamma' w_i$$

de aquí el nombre de Modelo de Probabilidad Lineal.

- El problema es que la estimación por MCC de z, $\hat{z}_i = \hat{\gamma}' w_i$, es la estimación de la probabilidad de que la persona i se encuentre trabajando. En la práctica, como MCC estima una recta, los valores de $\hat{z}_i$ pueden caer fuera del intervalo (0, 1).

Modelo de Probabilidad Lineal

- Como z_i solo adopta dos valores, $z_i = 1$ ó $z_i = 0$, los errores del modelo pueden ser $u_i = 1 - \gamma' w_i$ ó $u_i = -\gamma' w_i$ con probabilidades P_i y $1 - P_i$ respectivamente. Entonces la distribución de los errores es,

u_i	$f(u_i)$
$1 - \gamma' w_i$	$\gamma' w_i$
$-\gamma' w_i$	$1 - \gamma' w_i$

- Una consecuencia de la distribución anterior es que

$$\begin{aligned} Var(u_i) &= \gamma' w_i (1 - \gamma' w_i)^2 + (1 - \gamma' w_i)(\gamma' w_i)^2 \\ &= \gamma' w_i (1 - \gamma' w_i) \\ &= E(z_i|w_i)[1 - E(z_i|w_i)]. \end{aligned}$$

y los errores son heterocedásticos.

Modelo de Probabilidad Lineal

- Golberger (1964) sugirió el siguiente procedimiento de estimación: Primero, estime (33) por MCC. Segundo, calcule $\hat{z}_i(1-\hat{z}_i)$ y use MCGE.
- Los problemas del MPL son los siguientes
  1. En la práctica $\hat{z}_i(1-\hat{z}_i)$ puede ser cero o negativo y el método de MCGE no puede aplicarse.
  2. Como la esperanza condicional de la variable dependiente $E(z_i|w_i)$ se interpreta como una probabilidad, en la práctica su estimación por MCGE puede arrojar valores fuera del intervalo $(0, 1)$.
- Una forma alternativa para formular el modelo que resuelve los problemas anteriores es considerar que $z_i = 1$ cuando $z_i^* > 0$ de forma tal que

$$\begin{aligned} P_i = Pr[z_i = 1] &= Pr[z_i^* > 0] = Pr[\gamma' w_i + u_i > 0] \\ &= Pr[u_i > -\gamma' w_i] = Pr[u_i < \gamma' w_i] \\ &= \Phi(\gamma' w_i) \end{aligned}$$

Modelo Probit

- A diferencia de lo que ocurría con el MPL en este caso existe una relación no lineal con los parámetros del modelo y por lo tanto no puede usarse el método de mínimos cuadrados para la estimación.
- Para realizar la estimación del modelo debemos recurrir al método de máxima verosimilitud.
- Como sabemos la función de probabilidad de los errores y tenemos una muestra aleatoria (es decir, compuesta por variables aleatorias independientes) la función de verosimilitud es simplemente la multiplicación de las funciones de probabilidad para todas las observaciones que hay en la muestra.
-

$$L(\gamma; w) = \prod_{i=1}^{n} \Phi(\gamma' w_i)^{z_i} [1 - \Phi(\gamma' w_i)]^{(1-z_i)}$$

Modelo Probit: Estimación

- El logaritmo natural de la función de verosimilitud es,

$$l(\hat{\gamma}; w_i) = \sum_{i=1}^{n} \left[z_i ln\{\Phi(\hat{\gamma}' w_i)\} + (1 - z_i) ln\{1 - \Phi(\hat{\gamma}' w_i)\} \right]$$

- Las condiciones de primer orden para la maximización de esta función son,

$$S(\hat{\gamma}) = \frac{\partial l(\cdot)}{\partial \hat{\gamma}} = \sum_{i=1}^{n} \left[\frac{z_i - \Phi(\hat{\gamma}' w_i)}{\Phi(\hat{\gamma}' w_i)[1 - \Phi(\hat{\gamma}' w_i)]} \phi(\hat{\gamma}' w_i) \right] w_i = 0$$

- Como se puede observar en las condiciones de primer orden las incógnitas, $\hat{\gamma}$, entran en forma no lineal y por lo tanto no pueden resolverse usando métodos lineales.

Sesgo de Selección por Truncamiento Incidental

- Para resolver las condiciones de primer orden hay que recurrir a algoritmos numéricos.
- Amemiya (1985) demostró que la función de verosimilitud del modelo Probit es globalmente cóncava por lo que las condiciones de segundo orden para un máximo se cumplen y no es necesario chequearlas.
- Sin embargo, necesitamos las condiciones de segundo orden para calcular la matriz de información,

$$\begin{aligned} I(\hat{\gamma}) &= E\left(-\frac{\partial^2 l(\cdot)}{\partial\hat{\gamma}\partial\hat{\gamma}'}\right) \\ &= \sum_{i=1}^{n} \frac{[\phi(\hat{\gamma}'w_i)]^2}{\Phi(\hat{\gamma}'w_i)[1-\Phi(\hat{\gamma}'w_i)]} w_i w_i' \end{aligned}$$

- Entonces la matriz de varianzas y covarianzas asintótica viene dada por $[I(\hat{\gamma})]^{-1}$.

Sesgo de Selección por Truncamiento Incidental

- Adicionalmente, usando el algoritmo de Newton-Raphson, comenzando con un valor inicial $\hat{\gamma}_0$ el nuevo valor de $\hat{\gamma}_1$ se obtiene,

$$\hat{\gamma}_1 = \hat{\gamma}_0 + [I(\hat{\gamma}_0)]^{-1} S(\hat{\gamma}_0)$$

este procedimiento iterativo se repite hasta converger.

- Volviendo al ejemplo de la estimación de la oferta de trabajo de las mujeres, Heckman (1979) sugirió utilizar el siguiente procedimiento en dos etapas.
  1. Estimar la ecuación de selección usando un Probit para obtener estimaciones de γ. Luego para cada observación de la muestra se construye,

$$\hat{\lambda}_i = \frac{\phi(\hat{\gamma}' w_i)}{\Phi(\hat{\gamma}' w_i)}$$

     También vamos a necesitar $\hat{\delta}_i = \hat{\lambda}_i(\hat{\lambda}_i + \hat{\gamma}' w_i)$.
  2. Estime β y β_λ por MCC en una regresión de y sobre x y $\hat{\lambda}$.

Sesgo de Selección por Truncamiento Incidental

- Para poder hacer inferencia estadística en la regresión del segundo paso hay que tener en cuenta dos problemas: heterocedasticidad en v_i y el hecho de que una de las variables explicativas de la regresión está construída a partir de una estimación anterior.
- Heckman (1979) deriva la verdadera matriz de varianzas y covarianzas de los estimadores del segundo paso.
- Recuerde que $\widehat{Var(v_i)} = \hat{\sigma}_\epsilon^2(1 - \hat{\rho}^2\hat{\delta}_i)$ usando el teorema (20.4) de Greene. Donde $\hat{\sigma}_\epsilon^2 = \frac{e'e}{n} + \bar{\delta}\hat{\beta}_\lambda^2$, y $\bar{\delta} = \frac{1}{n}\sum_i \hat{\delta}_i$.
- Entonces la estimación correcta de la matriz de varianzas y covarianzas del segundo paso es,

$$Var[\hat{\beta}, \hat{\beta}_\lambda] = \hat{\sigma}_\epsilon^2(\sum_i x_i^{*'} x_i^*)^{-1} \left[\sum_i (1 - \hat{\rho}^2\hat{\delta}_i) x_i^* x_i^{*'} + Q\right] (\sum_i x_i^{*'} x_i^*)^{-1}$$

Sesgo de Selección por Truncamiento Incidental

- Donde,

$$Q = \hat{\rho}^2(X^{*'}\Delta W)[Avar(\hat{\gamma})](W'\Delta X^*)$$

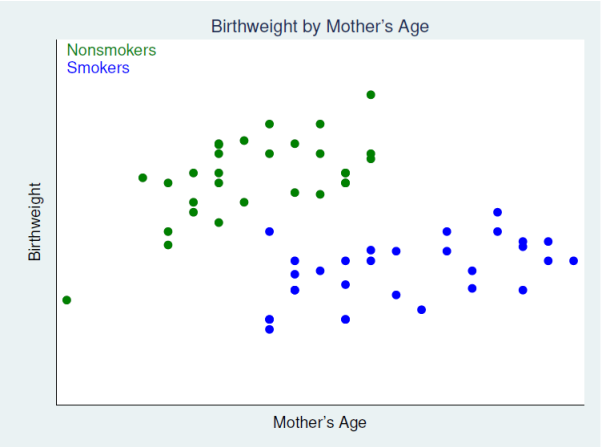
y Δ es una matriz diagonal con $\hat{\delta}_i$ en la diagonal principal.

Costos Económicos de Nacer con Peso Bajo

- Considere un caso hipotético como el de Almond et al. (2005).
- La pregunta que se quiere responder es si fumar durante el embarazo afecta el peso de un recién nacido.
- Las unidades en este caso son mujeres embarazadas, algunas de las cuales fumaron durante el embarazo.
- La variable de resultado es el peso del bebé al nacer.

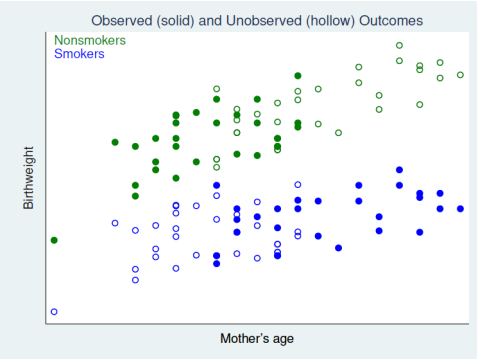
Costos Económicos de Nacer con Peso Bajo

- La figura muestra el peso del bebé al nacer para madres fumadoras y no fumadoras como función de la edad de la madre.



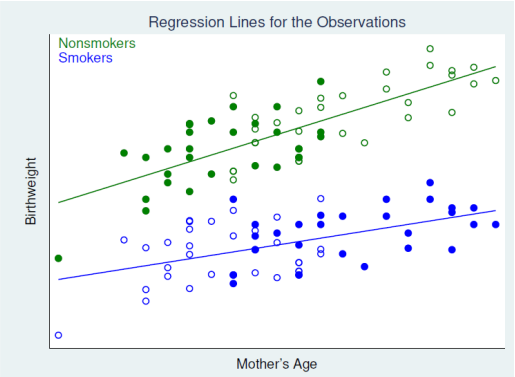
Costos Económicos de Nacer con Peso Bajo

- La figura sugiere que las mujeres fumadoras tienden a tener mayor edad que las no fumadoras.
- Para las mujeres fumadoras de mayor edad y las no fumadoras más jóvenes no parece haber un soporte común.
- Supongamos que observamos los resultados potenciales de cada mujer embarazada:



Costos Económicos de Nacer con Peso Bajo

- Lo que hace el método de regresión es estimar una regresión (con los datos observados, círculos sólidos) del peso del bebé sobre la edad de la madre para el grupo de madres fumadoras y para el grupo de madres no fumadoras.
- Luego se usa la linea de regresión muestral de las madres fumadoras como resultado contrafáctico de las madres no fumadoras y viceversa.



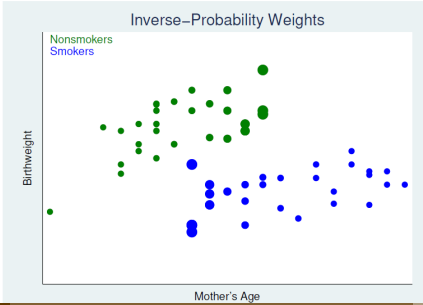
Estimación

- La estimación del ATE condicional es directa a través del método de mínimos cuadrados clásicos.
- En la práctica estimamos $E[y \mid \mathbf{w}, s_T = 1]$ con el plano de regresión muestral, $\widehat{m}_1(\mathbf{w}, \hat{\delta}_1)$, con las observaciones del tratamiento y estimamos $E[y \mid \mathbf{w}, s_T = 0]$ con el plano de regresión muestral, $\widehat{m}_0(\mathbf{w}, \hat{\delta}_0)$, con las observaciones del control.

$$\widehat{\tau}_w = \frac{1}{N} \sum_{i=1}^{N} \left[\widehat{m}_1(\mathbf{w_i}, \hat{\delta}_1) - \widehat{m}_0(\mathbf{w_i}, \hat{\delta}_0) \right]$$

Inversa del propensity score como ponderador

- Una alternativa a esto es ponderar las esperanzas matemáticas por la inversa de la probabilidad condicional de recibir el tratamiento (**propensity score**).
- El método del propensity score ve a los resultados potenciales no observados (círculos sin relleno) como observaciones faltantes y usa ponderadores para corregir las estimaciones de las esperanzas para las unidades tratadas y no tratadas.
- El método aplica mayor ponderación a los círculos verdes sólidos correspondientes a madres de mayor edad y menor ponderación a los correspondientes a madres más jóvenes.



Inversa del propensity score como ponderador

- En términos formales, una forma de estimar el ATE es utilizando la inversa del propensity score ($\pi(x)$) como ponderador.
- Recordando que $sy = sy_1$ tenemos

$$\begin{aligned} E\left[\frac{sy}{\pi(x)}\middle| x\right] &= E\left[\frac{sy_1}{\pi(x)}\middle| x\right] = E\left\{E\left[\frac{sy_1}{\pi(x)}\middle| x, s\right]\middle| x\right\} \\ &= E\left\{\frac{sE(y_1|x,\, s)}{\pi(x)}\middle| x\right\} = E\left\{\frac{sE(y_1|x)}{\pi(x)}\middle| x\right\} \\ &= E\left\{\frac{s}{\pi(x)}\middle| x\right\} E(y_1|x) = E(y_1|x) \end{aligned}$$

- Haciendo uso de $E(s|x) = \pi(x)$.
- Usando un argumento similar se puede mostrar que:

$$E\left[\frac{(1-s)y}{1-\pi(x)}\middle| x\right] = E(y_0|x)$$

Inversa del propensity score como ponderador

- Usando álgebra se puede mostrar que,

$$E(y_1 - y_0|x) = E\left[\frac{[s - \pi(x)]y}{\pi(x)[1 - \pi(x)]}\middle| x\right]$$

- Y usando expectativas iteradas,

$$\tau_{ATE} = E(y_1 - y_0) = E\left[\frac{[s - \pi(x)]y}{\pi(x)[1 - \pi(x)]}\right] \tag{34}$$

Inversa del propensity score como ponderador

- Note que $\pi(x) = Pr[s = 1|x]$ es la probabilidad de recibir el tratamiento y ex-ante no es observable.
- Lo único que observamos en la práctica es si la persona recibe el tratamiento ($s = 1$) o no lo recibe ($s = 0$).
- Supongamos que una persona se autoselecciona en el tratamiento dependiendo de la utilidad que le brinda.
- Denotemos por

$$U_i^T = x_i\gamma_T + u_i^T \tag{35}$$

a la utilidad que le brinda al individuo i autoseleccionarse en el tratamiento y de la misma manera

$$U_i^C = x_i\gamma_C + u_i^C \tag{36}$$

a la utilidad de no recibir el tratamiento.

Inversa del propensity score como ponderador

- Entonces

$$\begin{aligned}\pi(x) &= Pr[s=1|x] = Pr[U_i^T > U_i^C] = Pr[u_i^C - u_i^T \leq x_i(\gamma_T - \gamma_C)] \\ &= Pr[u_i \leq x_i\gamma] = \Phi(x_i\gamma)\end{aligned}$$

- Donde asumimos que el error u_i tiene distribución normal estándar (modelo Probit).
- En términos de nuestro ejemplo,

$$\Phi(x_i\gamma) = \int_{-\infty}^{x_i\gamma} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt$$

existe una relación nolineal entre la probabilidad de recibir el tratamiento y los parámetros del modelo.

Modelo Probit: estimación

- Para realizar la estimación del modelo debemos recurrir al método de máxima verosimilitud.
- La función de verosimilitud es la multiplicación de la probabilidad de recibir el tratamiento para todas las observaciones que hay en la muestra.
-

$$L(\gamma; x) = \prod_{i=1}^{n} \Phi(x_i\gamma)^{s_i}[1 - \Phi(x_i\gamma)]^{(1-s_i)}$$

Modelo Probit: Estimación

- El logaritmo natural de la función de verosimilitud es,

$$l(\hat{\gamma}; x_i) = \sum_{i=1}^{n} [s_i ln\{\Phi(x_i\hat{\gamma})\} + (1 - s_i) ln\{1 - \Phi(x_i\hat{\gamma})\}]$$

- Las condiciones de primer orden para la maximización de esta función son,

$$S(\hat{\gamma}) = \frac{\partial l(\cdot)}{\partial \hat{\gamma}} = \sum_{i=1}^{n} \left[\frac{s_i - \Phi(x_i\hat{\gamma})}{\Phi(x_i\hat{\gamma})[1 - \Phi(x_i\hat{\gamma})]} \phi(x_i\hat{\gamma}) \right] x_i = 0$$

- Como se puede observar en las condiciones de primer orden las incógnitas, $\hat{\gamma}$, entran en forma no lineal y por lo tanto no pueden resolverse usando métodos lineales.

Modelo Probit: Estimación

- Para resolver las condiciones de primer orden hay que recurrir a algoritmos numéricos.
- Amemiya (1985) demostró que la función de verosimilitud del modelo Probit es globalmente cóncava por lo que las condiciones de segundo orden para un máximo se cumplen y no es necesario chequearlas.
- Sin embargo, necesitamos las condiciones de segundo orden para calcular la matriz de información,

$$\begin{aligned} I(\hat{\gamma}) &= E\left(-\frac{\partial^2 l(\cdot)}{\partial\hat{\gamma}\partial\hat{\gamma}'}\right) \\ &= \sum_{i=1}^{n} \frac{[\phi(x_i\hat{\gamma})]^2}{\Phi(x_i\hat{\gamma})[1-\Phi(x_i\hat{\gamma})]} x_i x_i' \end{aligned}$$

- Entonces la matriz de varianzas y covarianzas asintótica viene dada por $[I(\hat{\gamma})]^{-1}$.

- Adicionalmente, usando el algoritmo de Newton-Raphson, comenzando con un valor inicial $\hat{\gamma}_0$ el nuevo valor de $\hat{\gamma}_1$ se obtiene,

$$\hat{\gamma}_1 = \hat{\gamma}_0 + [I(\hat{\gamma}_0)]^{-1} S(\hat{\gamma}_0)$$

  este procedimiento iterativo se repite hasta converger.
- Una vez estimado el propensity score usando un Probit estimamos el ATE,

$$\hat{\tau}_{ATE} = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{[s - \hat{\pi}(x)]y}{\hat{\pi}(x)[1 - \hat{\pi}(x)]} \right] \tag{37}$$

Variable Dependiente Binaria y Efectos Marginales

- En los modelos lineales los coeficientes que acompañan a las variables explicativas tienen la interpretación de ser directamente el efecto marginal de un cambio en la variable explicativa sobre la variable dependiente.
- En el caso de modelos no lineales como el Probit los coeficientes que acompañan a las variables explicativas no tienen esta interpretación.
- Para ver esto, considere el siguiente modelo:

$$\Phi(\alpha + \beta x_i + z\gamma) = \int_{-\infty}^{\alpha+\beta x_i+z\gamma} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt$$

donde z es un vector de variables explicativas adicionales a x y γ un vector de parámetros poblacionales.

- En este caso el efecto marginal viene dado por la derivada parcial de la probabilidad ex-ante con respecto a alguna variable explicativa.

$$\frac{\partial Pr[y_i = 1|\cdot]}{\partial x_i} = \phi(\alpha + \beta x_i + z\gamma) \times \beta$$

Variable Dependiente Binaria y Efectos Marginales

- La expresión anterior corresponde a una variable explicativa x_i continua.
- Si x_i es una variable discreta (binaria) el efecto marginal viene dado por:

$$Pr[y_i = 1|x_i = 1] - Pr[y_i = 1|x_i = 0] = \{\Phi(\alpha + \beta + z\gamma) - \Phi(\alpha + z\gamma)\} \quad (38)$$

- Note que a diferencia de los modelos lineales, el efecto marginal es una función y no una constante.
- Para obtener un resultado numérico puntual, es estándar en la práctica calcular el denominado efecto marginal promedio.
- En términos matemáticos,

$$E\{Pr[y_i = 1|x_i = 1] - Pr[y_i = 1|x_i = 0]\} = E\{\Phi(\alpha + \beta + z\gamma) - \Phi(\alpha + z\gamma)\} \quad (39)$$

Variable Dependiente Binaria y Efectos Marginales

- Supongamos ahora que queremos analizar la eficacia de una vacuna contra el COVID-19 en un experimento aleatorizado.
- La "política" es la vacunación.
- La variable de resultado es si el individuo se contagia o no de COVID.
- Aquí el resultado potencial es una variable binaria y en un modelo lineal el ATE viene dado por:

$$\begin{aligned} ATE &= E(Y_s(u)|s_T = 1) - E(Y_s(u)|s_T = 0) = E(Y_T(u)) - E(Y_C(u)) \\ &= Pr(Y_s = 1|s_T = 1) - Pr(Y_s = 1|s_T = 0) \end{aligned} \tag{40}$$

  donde u denota al individuo, Y_s es una variable binaria que adopta el valor unitario si el individuo se contagia de COVID-19 y vale cero si no se contagia y $s_T = 1$ si el individuo está vacunado.
- La segunda igualdad viene dada por el hecho de que la esperanza matemática condicional de un modelo de variable dependiente binaria es igual a la probabilidad de ocurrencia del evento analizado.

Variable Dependiente Binaria y Efectos Marginales

- Supongamos que estimamos las probabilidades utilizando un modelo Probit

$$E\left[\mathrm{Y}_i \mid \mathrm{X}_i, \mathrm{S}_i\right] = \Phi\left[\mathrm{X}_i'\beta_0 + \beta_1 \mathrm{S}_i\right] \tag{41}$$

- Ahora la "switching equation" es

$$E\left[\mathrm{Y}_i \mid \mathrm{X}_i, \mathrm{S}_i\right] = \Phi\left[\mathrm{X}_i'\beta_0\right] + \left\{\Phi\left[\mathrm{X}_i'\beta_0 + \beta_1\right] - \Phi\left[\mathrm{X}_i'\beta_0\right]\right\} \times \mathrm{S}_i \tag{42}$$

- y el ATE es

$$\begin{aligned} ATE &= E\left\{Pr(Y_i = 1|\mathrm{X}_i, \mathrm{S}_i = 1) - Pr(Y_i = 1|\mathrm{X}_i, \mathrm{S}_i = 0)\right\} \\ &= E\left\{\Phi\left[\mathrm{X}_i'\beta_0 + \beta_1\right] - \Phi\left[\mathrm{X}_i'\beta_0\right]\right\} \end{aligned} \tag{43}$$

Variable Dependiente Binaria y Efectos Marginales

- El principal resultado de esta discusión es que estimar el efecto tratamiento promedio usando un modelo Probit involucra calcular el efecto marginal promedio.
- En la práctica no hay grandes diferencias entre el efecto marginal promedio y el coeficiente sobre la variable de asignación del tratamiento en un modelo de regresión lineal.
- Esta regularidad empírica más la complejidad adicional para hacer inferencia estadística (i.e. se necesitan errores estándar para los efectos marginales promedio) hace que muchos investigadores cuando trabajan con evaluaciones de impacto prefieran aproximar la estimación de las esperanzas matemáticas condicionales usando un modelo de probabilidad lineal.