Estimadores no paramétricos y semi-paramétricos

Gabriel Montes-Rojas

Estimación de la media condicional no paramétrica

Esperanzas condicionales

- Toda variable aleatoria y se puede **descomponer** en dos partes ortogonales entre sí:

$$y = E(y|\boldsymbol{x}) + u,$$

donde

  - (i) $E(u|\boldsymbol{x}) = 0$,
  - (ii) $E(h(\boldsymbol{x})u) = 0$ para cualquier función $h(.)$.

Prueba: (i) Definamos $u \equiv y - E(y|\boldsymbol{x})$. Tomando esperanzas $E(u|\boldsymbol{x}) = E(y|\boldsymbol{x}) - E(y|\boldsymbol{x}) = 0$. (ii) Usando la ley de esperanzas iteradas $E(h(\boldsymbol{x})u) = E(E(h(\boldsymbol{x})u|\boldsymbol{x})) = E(h(\boldsymbol{x})E(u|\boldsymbol{x}) = 0$.

- Un resultado importante es que $E(u|\boldsymbol{x}) = 0$ implica que $E(u) = 0$. Esto es por la propiedad de esperanzas iteradas que dice que $E_u(u) = E_{\boldsymbol{x}}[E_u(u|\boldsymbol{x})]$, donde la primera esperanza es con respecto a u y la segunda a $\boldsymbol{x}$.
- También, que $E(u|\boldsymbol{x}) = 0$ implica que $E(\boldsymbol{x}u) = 0$.
- Entonces, $cov(\boldsymbol{x}, u) \equiv E(\boldsymbol{x}u) - E(\boldsymbol{x})E(u) = 0$.
- Es decir, el supuesto $E(u|\boldsymbol{x}) = 0$ implica que los errores u de un modelo de regresión no están correlacionados con las $\boldsymbol{x}$.

Esperanzas condicionales

- La esperanza condicional es la solución al problema de minimización del valor esperado de las desviaciones al cuadrado, o sea $E(y|\boldsymbol{x}) = arg\min_{m(\boldsymbol{x})} E((y - m(\boldsymbol{x}))^2)$.
  Prueba: $(y - m(\boldsymbol{x}))^2 = ((y - E(y|\boldsymbol{x})) + (E(y|\boldsymbol{x}) - m(\boldsymbol{x})))^2 =$ $(y - E(y|\boldsymbol{x}))^2 + (E(y|\boldsymbol{x}) - m(\boldsymbol{x}))^2 + 2\ (y - E(y|\boldsymbol{x}))(E(y|\boldsymbol{x}) - m(\boldsymbol{x}))$. Notemos que el primer término no depende de $m(\boldsymbol{x})$, mientras que el tercero se puede escribir como $u(x)(E(y|\boldsymbol{x}) - m(\boldsymbol{x})) = u(x)h(x)$. Si tomamos la esperanza condicional del tercero tenemos 0 por (ii).
- Sin embargo, no sabemos la forma funcional de $E(y|\boldsymbol{x})$. Si la ésta fuese conocida se podría estimar por regresión lineal o no lineal.
- Si la esperanza condicional es lineal, entonces $E(y|\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\beta}$. Para cualquier variable aleatoria y, tenemos la proyección poblacional sobre el espacio generado por las $\boldsymbol{x}$, $r(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\beta}$ donde $\boldsymbol{\beta} = arg\min_{\boldsymbol{b}} E((y - \boldsymbol{x}\boldsymbol{b})^2)$.
- Cada vez que corremos una regresión estamos estimando $E(y|\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{b}$ asumiendo que es lineal en los parámetros. Conviene entonces decir que estamos estimando una esperanza condicional y levantar el supuesto de regresores no estocásticos.

Incorrecta especificación

- ¿Qué pasa si $E(y|x) = m(x)$ no es lineal pero corremos una regresión lineal? Supongamos que queremos estimar el modelo $y_i = \beta x_i + u_i$. Entonces el estimador MCO, $\hat{\beta}$ lo podemos representar como
  $\hat{\beta} = \frac{Cov(y,x)}{Var(x)} + o_p(1) = \frac{Cov(m(x)+e,x)}{Var(x)} + o_p(1) = \frac{Cov(m(x),x)}{Var(x)} + o_p(1)$.
- Tomemos la expansión de Taylor de $m(x)$, evaluada a x^*:

$$m(x) = m(x^*) + m'(x^*)(x - x^*) + \frac{m''(x^*)}{2}(x - x^*)^2 + ...$$

  - Si $m(x) = a + bx$ entonces $\beta = b$.
  - Si $m(x) = a + bx + cx^2$ entonces $\beta = b + c\frac{Cov(x^2,x)}{Var(x)}$ (como si x^2 fuera una variable omitida).
  - Para cualquier $m(x)$ tenemos

$$\beta(x^*) = m'(x^*) + \frac{m''(x^*)}{2Var(x)}\left(Cov(x^2, x) - 2x^* Var(x)\right) + ...$$

Estimación de la media condicional local

Este estimador plantea estimar "localmente" promediando los valores de y "cerca" de x.

- Consideremos la estimación de $m(x) = E[y|x]$ en un valor x dado.

$$E[y|x] = \int yf(y|x)dy = \int y\frac{f(y,x)}{f(x)}dy = \frac{g(x)}{f(x)}$$

where

- $f(y|x)$ es la densidad condicional de y, condicional en x donde por definición $f(y|x) \equiv \frac{f(y,x)}{f(x)}$;
- $f(y,x)$ es la densidad conjunta de (y,x);
- $g(x) \equiv \int yf(y,x)dy$.

Estimador Nadaraya-Watson

El estimador de Nadaraya-Watson es un promedio ponderado de los ys que se corresponden a los xs cercanos a x.

- Consideremos la estimación de la media condicional por kernels.
- Definamos $\psi_i(x) = \frac{x_i - x}{h}$ donde h es el ancho de banda que pondera las distancias de cada x_i al valor de refrencia x.
- Definamos un kernel $K(.)$.
- Entonces,

$$\hat{m}_h(x) = \frac{\sum_{i=1}^n y_i K(\psi_i)}{\sum_{i=1}^n K(\psi_i)} = \sum_{i=1}^n y_i w_i,$$

donde $w_i = \frac{K(\psi_i)}{\sum_{i=1}^n K(\psi_i)}$ son los pesos de cada observación i.

Estimador Nadaraya-Watson

- Notar que cuando $h \to \infty$, $\hat{m}_h(x) \to \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, la media incondicional de y.
  - ¿Qué significa? Para un h grande $lim_{h\to\infty}\psi_i(x) = lim_{h\to\infty}\frac{x_i - x}{h} = 0$ y entonces $lim_{h\to\infty}K(\psi_i(x)) = K(0) = max = constante$. En este caso no hay suavizamiento basado en x.
  - Resultado: Demasiado suavizamiento.
- Notar que cuando $h \to 0$, entonces $\hat{m}_h(x)$ se vuelve igual a los puntos (y_i, x_i) o el estimador del vecino más cercano.
  - ¿Qué significa? Para un h chico $lim_{h\to 0}\psi_i(x) = lim_{h\to 0}\frac{x_i - x}{h} = \infty$ excepto cuando $x_i = x$. Notar que $K(\infty) = 0$ y entonces solo los x_is iguales a x son considerados. Esto es lo mismo que $\hat{m}_0(x) = \frac{\sum_i 1[x_i = x]y_i}{\sum_i 1[x_i = x]}$, esto es, para cada x toma el valor y si hay un par $(y_i, x_i = x)$; toma el promedio de las y tal que $x_i = x$ si hay más observaciones; o toma el valor de los más cercanos $x_j = x$.
  - Resultado: Nada de suavizamiento.

Efecto marginal

- Consideremos ahora el efecto marginal de x en y.
- Definamos

$$\beta(x) \equiv \frac{dm(x)}{dx} = m'(x) = \frac{f(x)g'(x) - g(x)f'(x)}{f^2(x)}$$

$$= \frac{g'(x)}{f(x)} - g(x)\frac{f'(x)}{f^2(x)}$$

- Entonces el estimador local de kernel es

$$\hat{\beta}(x) = \frac{\hat{g}'(x)}{\hat{f}(x)} - \hat{g}(x)\frac{\hat{f}'(x)}{\hat{f}^2(x)}$$

con

$$\hat{g}(x) = \frac{1}{nh}\sum_{i=1}^{n} y_i K(\psi_i), \hat{g}'(x) = \frac{1}{nh^2}\sum_{i=1}^{n} y_i K'(\psi_i)$$

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\psi_i), \hat{f}'(x) = \frac{1}{nh^2}\sum_{i=1}^{n} K'(\psi_i)$$

Regresión polinómica local

- El estimador Nadaraya-Watson se obtiene como $\hat{m}_h(x) = argmin_a \sum_{i=1}^n K(\psi_i)(y_i - a)^2$. Es decir, podemos pensarlo como una regresión ponderada, donde las ponderaciones son los kernels, $K(\psi_i)$.
- Ahora consideremos una extensión,

$$(\hat{a}_h(x), \hat{b}_h(x)) = argmin_{(a,b)} \sum_i (y_i - a - b(x_i - x))^2 K(\psi_i).$$

$$\begin{pmatrix} \widehat{a}_h(x) \\ \widehat{b}_h(x) \end{pmatrix} = \left[\sum_{i=1}^n K(\psi_i) \begin{pmatrix} 1 & (x_i - x) \\ (x_i - x) & (x_i - x)^2 \end{pmatrix} \right]^{-1} \sum_{i=1}^n K(\psi_i) \begin{pmatrix} y_i \\ y_i(x_i - x) \end{pmatrix}$$

- Se puede mostrar que $\widehat{a}_h(x)$ es un estimador de $m(x)$ y $\widehat{b}_h(x)$ es un estimador de $m'(x)$.

Regresión polinómica local

- Este estimador puede usarse en un polinomio de orden mayor, o sea $y_i - a - b(x_i - x) - c(x_i - x)^2$. De hecho para la estimación de $\beta(x) = \frac{dm(x)}{dx}$ es mejor usar un polinomio cuadrático para reducir el sesgo.
- Notar que $h \to \infty$, el estimador de regresión local se vuelve $\lim_{h\to\infty} \hat{a}_h(x) + \hat{b}_h(x) \cdot x = \hat{x}$, el estimador de media no condicional.
- ¿Qué pasa cuando $h \to 0$?

Propiedades asintóticas

Supuestos

Consideremos los siguientes supuestos

1. m y f son dos veces diferenciables en un vecindario de x. f es acotada en un vecindario de x. $x \in int(\mathcal{X})$ [No puede haber "saltos".]
2. El kernel K es una función simétrica con (i) $\int K(\psi)d\psi = 1$; (ii) $\int \psi K(\psi)d\psi = 0$; (ii) $\int \psi^2 K(\psi)d\psi = \mu_2 < \infty$.
3. $h = h_n \to 0$, $nh_n \to \infty$ as $n \to \infty$.
4. Las xs son iid independentes del error, $y = m(x) + error$.

Propiedades asintóticas
Nadaraya-Watson

Teorema (Pagan&Ullah p. 101) *Con los supuestos de arriba*

$$Sesgo(\hat{m}_h(x)) = \frac{h^2}{2f}\mu_2(m''f + 2f'm') + O(n^{-1}h^{-1}) + o(h^2)$$

$$Var(\hat{m}_h(x)) = \frac{\sigma^2}{nhf}\int K^2(\psi)d\psi + o(n^{-1}h^{-1})$$

Entonces el ancho de banda óptimo es $h_n \propto n^{-1/5}$.

Propiedades asintóticas
Regresión lineal local

Teorema (Pagan&Ullah p. 105) *Con los supuestos de arriba*

$$Sesgo(\hat{m}_h(x)) = \frac{m'' h^2}{2}\mu_2 + O(n^{-1}h^{-1}) + o(h^2)$$

$$Var(\hat{m}_h(x)) = \frac{\sigma^2}{nhf}\int K^2(\psi)d\psi + o(n^{-1}h^{-1})$$

La maldición de la dimensionalidad

- Estos resultados se pueden usar para más variables de control, digamos q.
- Sin embargo, cuando q crece, la tasa de convergencia empeora, en particular $O_p(n^{-2/(q+4)})$. Comparemos esto con la teoría asintótica de MCO $O_p(n^{-1/2})$ para cualquier q.
- En particular,

$$h_{opt} \propto n^{-1/(q+4)}$$

STATA

Nadaraya-Watson y regresión local se puede implementar con el comando `lpoly`:
`http://www.stata.com/manuals13/rlpoly.pdf`
Ver también `http://www.stata.com/manuals13/rlowess.pdf` y `http://www.stata.com/manuals13/rsmooth.pdf`

Estimación de media condicional semi-paramétrica

Modelos parcialmente lineales: $y = x'\beta + g(z) + u$

Consideremos el modelo

$$y_i = x_i'\beta + g(z_i) + u_i, \quad i = 1, 2, ..., n,$$

- x_i es un vector $p \times 1$ de controles;
- z_i es un vector $q \times 1$ de controles;
- $g(.)$ es una función no conocida;
- $u_i \equiv y_i - E(y_i|x_i, z_i)$ entonces $E(u_i|x_i, z_i) = 0$ y $E(u_i^2|x_i, z_i) = \sigma^2(x_i, z_i)$ (puede haber heterocedasticidad).
- Ejemplo: Supongamos que podemos asumir un modelo lineal para algunas variables (x) pero no para otras (z).
- Este modelo evita la maldición de la dimensionalidad si tenemos unos "pocos" o un z.
- Al separar entre variables lineales y no-lineales se incrementa la precisión de los estimadores.
- Vamos a tener que $\hat{\beta}$ es $\sqrt{n}$-consistente (como MCO).

Modelos parcialmente lineales: $y = x'\beta + g(z) + u$

Robinson (1988)

- Consideremos $E(y_i|z_i) = E(x_i|z_i)'\beta + g(z_i) + E(u_i|z_i) = E(x_i|z_i)'\beta + g(z_i)$ $(porque\ E(u|z) = 0)$.
- Entonces, restando del modelo original $y_i - E(y_i|z_i) = (x_i - E(x_i|z_i))'\beta + u_i$. Llamemos $\tilde{y}_i = y_i - E(y_i|z_i)$ y $\tilde{x}_i = x_i - E(x_i|z_i)$, entonces podemos obtener

$$\tilde{\beta} = \left[\sum_i^n \tilde{x}_i \tilde{x}_i'\right]^{-1} \sum_i^n \tilde{x}_i \tilde{y}_i$$

  Notar que este estimador no es factible porque no tenemos $E(y_i|z_i)$ ni $E(x_i|z_i)$.
- $E(y|z)$ y $E(x|z)$ se pueden obtener **no paramétricamente** como $\widehat{E(y|z)}$ y $\widehat{E(x|z)}$.
- Entonces el estimador de β es $\widehat{\tilde{\beta}} = \left[\sum_i^n \widehat{\tilde{x}}_i \widehat{\tilde{x}}_i'\right]^{-1} \sum_i^n \widehat{\tilde{x}}_i \widehat{\tilde{x}}_i$, donde $\widehat{\tilde{x}}_i = x_i - \widehat{E(x_i|z_i)}$ y $\widehat{\tilde{y}}_i = x_i - \widehat{E(y_i|z_i)}$.
- $g(z_i)$ se puede obtener de $\hat{g}(z_i) = \widehat{\tilde{y}}_i - \widehat{\tilde{x}}_i'\tilde{\beta}$.

Partially linear models: $y = x'\beta + g(z) + u$

Yatchew's (1998) estimator

- Ordenemos los datos de acuerdo a z, o sea, $z_{(1)}, z_{(2)}, ..., z_{(n)}$.
- Consideremos el estimador de primeras diferencias

$$\Delta y = \Delta x'\beta + \Delta g(z) + \Delta u$$

- Si $g(.)$ es suave con derivada acotada en un espacio compacto, entonces $\Delta g(z) \approx 0$ cuando $n \to \infty$.
- Entonces β se puede obtener de la regresión de Δy en Δx, como $\hat{\beta}$.
- $g(z)$ se puede obtener como un estimador no paramétrico de $y - x\hat{\beta}$ en z.

Modelos de índice: $y = g(x'\beta) + u$

Un modelo semi-paramétrico de índice es

$$y_i = g(x_i'\beta) + u_i, \quad i = 1, 2, ..., n,$$

- x_i es un vector $q \times 1$;
- $g(.)$ es una función no conocida;
- Nota: esto es diferente a $y_i = g(x_i) + u_i$;
- $u_i \equiv y_i - E(y_i|x_i)$ entonces $E(u_i|x_i) = 0$ y $E(u_i^2|x_i) = \sigma^2(x_i)$ (i.e. puede haber heterocedasticidad).

Modelos de índice: $y = g(x'\beta) + u$

El estimador de Ichimura (1993) consiste en asumir que hay un parámetro, β_0, donde

$$y_i = g(x_i'\beta_0) + u_i, \quad i = 1, 2, ..., n.$$

Sin embargo definimos $E(y|x'\beta)$ para cualquier β. Notar que $E(y|x'\beta) \neq g(x'\beta)$ a menos que $\beta = \beta_0$.

- Consideremos una grilla $\beta \in \mathcal{B}$, $\mathcal{B} = [\beta^{(1)}, \beta^{(2)}, ..., \beta^{(M)}]$;
- Para cada $j = 1, 2, ..., M$ y para cada observación $i = 1, 2, ..., n$, estimar $\hat{g}_{-i}(x_i'\beta^{(j)})$ no paramétricamente usando el estimador leave-one-out con kernel $g_i(x_i'\beta)$;
- Elegir β como

$$\hat{\beta} = argmin_{\beta \in \mathcal{B}} \sum_i^n (y_i - \hat{g}_{-i}(x_i\beta))^2$$

Modelos de índice de elección binaria

Supongmos que y es binaria, o sea $y \in \{0, 1\}$. Esto es una alternativa a probit y logit, Klein y Spady (1993).
Sea $g(x_i\beta) = Pr[y = 1|x_i]$. Entonces aplicar Ichimura a la cuasi-verosimilitud

$$\hat{\beta} = argmax_{\beta \in \mathcal{B}} \sum_{i}^{n} (1 - y_i) ln(1 - \hat{g}_{-i}(x_i'\beta)) + y_i ln(\hat{g}_{-i}(x_i'\beta))$$

Nota: Comparar con probit y logit,

$$\hat{\beta} = argmax_{\beta \in \mathcal{B}} \sum_{i}^{n} (1 - y_i) ln(1 - F(x_i'\beta)) + y_i ln(F(x_i'\beta))$$

donde F es la función de probabilidad acumulada normal o logística.

STATA

- El estimador de Robinson (1988) se puede obtener con `semipar`.
- El estimador de Yatchew (1988) se puede obtener con `plreg`.
- `sml` estima el estimador semi-paramétrico de Klein y Spady (1993).
  `http://www.stata-journal.com/sjpdf.html?articlenum=st0144`

Modelo de filtros

- El filtro de series de tiempo más conocido es el de Hodrick-Prescott, "HP filter".

$$min_{\tau_t} \sum_{t=1}^{T} \left[(x_t - \tau_t)^2 + \lambda((\tau_{t+1} - \tau_t) - (\tau_t - \tau_{t-1}))^2 \right]$$

  - τ_t: tendencia
  - $z_t = x_t - \tau_t$: business cycle
  - λ: parámetro de penalización de fluctuaciones
- En STATA ver el comando `tsfilter hp`.
- `http://repec.org/nasug2006/TSFiltering_beamer.pdf`

Otros modelos alternativos

- Tranformación de Box-Cox:
  `http://www.stata.com/manuals13/rboxcox.pdf`
- Splines:
  `http://www.stata.com/manuals13/rmkspline.pdf`

Referencias

- *Gutierrez, R.G., Linhart, J.M. and Pintblado, J.S. (2003), "From the help desk: Local polynomial regression and Stata plugins", Stata Journal, 3(4), 412–419.*
- *Ichimura, H. (1993), "Semiparametric least squares (SLS) and weighted SLS estimation of single-index models", Journal of Econometrics, 58, 71–120.*
- *Klein, R.W. and Spady, R.H. (1993), "An efficient semiparametric estimator for binary response models", Econometrica, 61, 387–421.*
- *Pagan, A. and Ullah, A. (1999), Nonparametric Econometrics. Cambridge: Cambridge University Press.*
- *Racine, J. (2008), "Nonparametric Econometrics: A Primer", Foundations and Trends in Econometrics, 3(1), 1-88.*
- *Robinson, P. (1988), "Root-n-consistent semi-parametric regression", Econometrica, 56, 931–954.*
- *Yatchew, A. (1998), "Nonparametric regression techniques in economics", Journal of Economic Literature, 36, 669–721.*