

Avances Recientes en Quantile Regression

Walter Sosa Escudero
(wsosa@udesa.edu.ar)

Universidad de San Andres

Septiembre de 2006

Outline

- 1 Regresion
- 2 Regresion por cuantiles
- 3 Estimacion e Inferencia
- 4 Lecturas y Software

Una digresion

Quantile regression:

- ¿Regresion cuantil?
- ¿Regresion cuantilica?
- ¿Retroceso del quantile? (Algoritmo no convergente)

quantile regression = regresion por cuantiles

Esperanzas condicionales y regresion

Consideremos un modelo lineal simple:

$$y = x'\beta + u$$

con $E(u|x) = 0$. De modo que:

$$E(y|x) = x'\beta$$

$x'\beta$ es una *funcion de regresion*, es la esperanza condicional de y en x .

Trivialmente:

$$\frac{\partial E(y|x)}{\partial x_k} = \beta_k$$

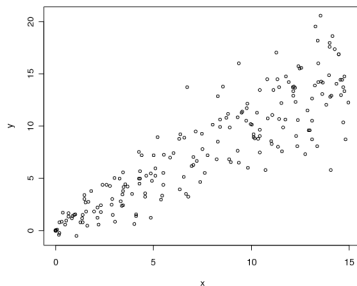
β mide como cambios marginales en x afectan a $E(y|x)$.

Si u es independiente de x , tambien es cierto que:

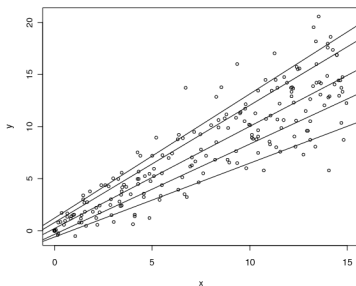
$$\frac{\partial y}{\partial x} = \beta$$

- La independencia es lo que le da sentido a la idea de 'alterar x dejando u constante'.
- El efecto de x sobre $E(y|x)$ resume tambien el efecto de x sobre y .

Consideremos el siguiente caso de *heterocedasticidad*:

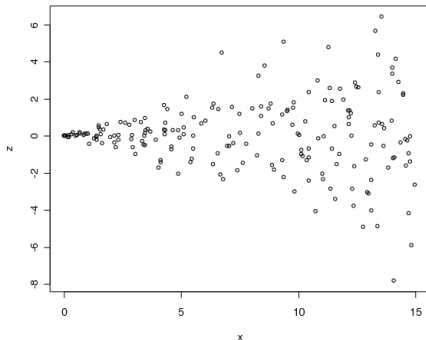


- ¿Cual es el efecto de x sobre $E(y|x)$?
- ¿Cual es el efecto de x sobre y ?



- El efecto de x es mayor 'arriba'. El efecto no es *homogeneo*.
- $\beta = \partial E(y|x)/\partial x$ *no* resume el ejemplo de x sobre y .

- ¿Es posible que x no afecte a $E(y|x)$ pero que si a y ?



x no afecta a $E(y|x)$
pero si afecta a y .

Regresion por cuantiles

Intenta modelar el efecto de x sobre **toda** la distribucion de y .

Cuantiles condicionales y quantile regression

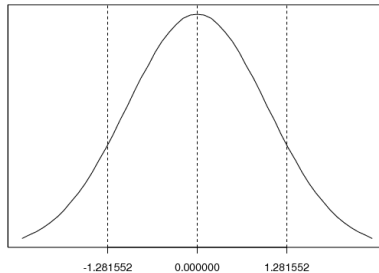
$Z \sim F(z)$ continua y monotona.

El τ -esimo cuantil de Z es un numero $Q_Z(\tau)$ que satisface:

$$F(Q_Z(\tau)) = \tau$$

o sea, el τ -esimo cuantil es un numero del soporte de la distribucion tal que la probabilidad de que ocurran valores menores es τ .

Cuantiles de la distribucion normal estandar



Recordar que el modelo simple de regresion puede ser visto como

$$E(y|x) = x'\beta$$

En forma analogia, el modelo de regresion para el τ -esimo cuantil de la distribucion de y condicional en x sera:

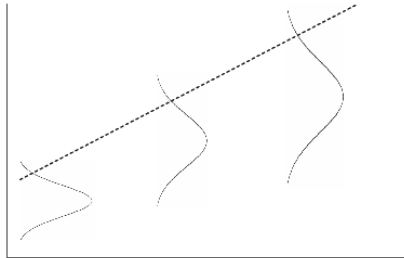
$$Q_{y|x}(\tau) = x'\beta(\tau)$$

Con :

$$\frac{\partial Q_{y|x}(\tau)}{\partial x} = \beta(\tau)$$

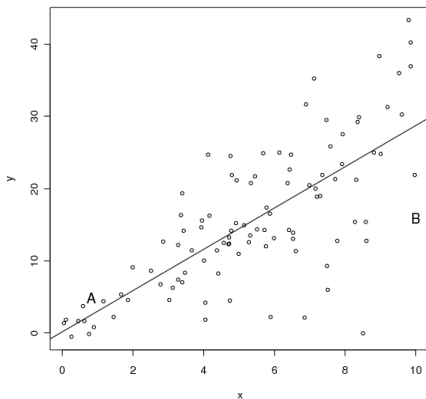
Estamos permitiendo que el efecto de x sobre y sea distinto en distintos lugares de la distribucion de y dado x

Ejemplo: $Q_{y|x}(0.75) = \beta_0(0.75) + \beta_1(0.75) x$



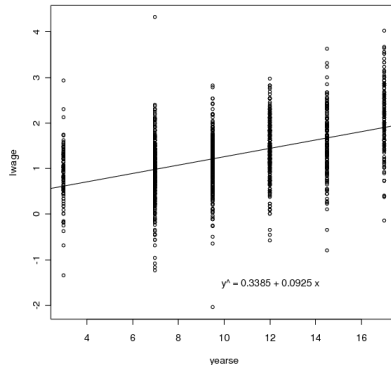
La recta une los cuantiles 0.75 de cada distribucion *condicional*.

Importante: el metodo estima rectas para distintos lugares de la distribucion *condicional* (y no de la no condicional!!!!)



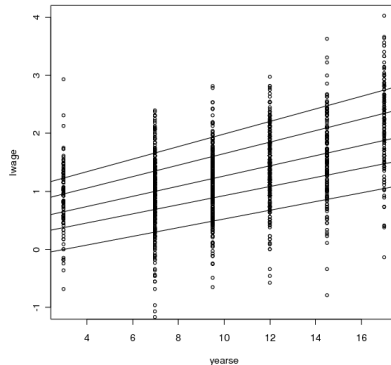
- 'A' esta arriba de la distribucion condicional y abajo en la no condicional.
- 'B' esta en el medio de la no condicional, y muy abajo en la condicional.

Ejemplo: Retornos a la Educacion



- El retorno 'medio' es 0.0925.

Estimacion por quantile regression:

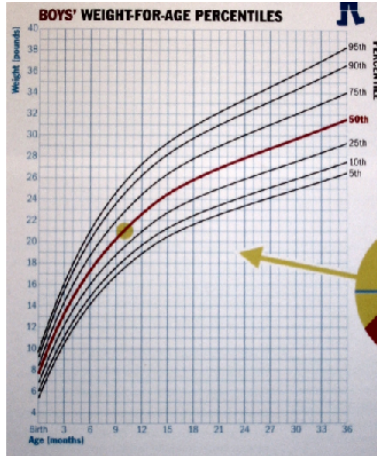


- El retorno es superior en los niveles mas altos.

Resumen de resultados:

Cuantiles	intercepto	educacion
0.10	-0.2260	0.0753
0.25	0.1420	0.0787
0.50	0.3841	0.0881
0.75	0.6580	0.0992
0.90	0.9055	0.1083
Retorno medio (MCO)	0.3385	0.0925

Ejemplo: Growth-Size Charts



Ejemplo: Factores Limitantes en Ecologia

Cade, B., Terrell, J. y Schroeder, R., 1999, Estimating the effects of limiting factors with regression quantiles, *Ecology*, 80(1), 311/323.

- Factor limitante: 'menos disponible entre los que afectan al crecimiento, supervivencia y reproduccion...'
- Ley de Liebig 'del minimo'
- Efecto de factores: no separable. Interacciones.
- Problema: como medir el efecto de un factor cuando es el unico limitante.

- Caso: 43 parcelas de 0.2 Ha (1989-1993) en Missouri (EEUU)
- y : 'produccion' de bellotas en un bosque de robles.
(Densidad o biomasa).
- x : 'forest suitability index' (porcentaje de cobertura, numero de especies, etc.).
- (y, x) muy similar al caso de educacion.
- CTS: en el 'medio' estan los casos en donde hay factors no observables (clima?) que actuan como limitantes.
- El verdadero efecto de x esta mejor medido 'arriba' de la distribucion.

y = densidad, x = calidad del bosque, u = no observables (clima?)

$$y = \beta_0 + \beta_1 x + \delta g(x)u$$

$g(.)$ es una funcion arbitraria, introduce una interaccion entre x y u .

$$\frac{\partial y}{\partial x} = \beta_1 + \delta u$$

si $\delta > 0$ el efecto de x sobre y es mayor cuanto mayor es u .

Los metodos de regresion por cuantiles permite modelar el efecto de estas interacciones sin hacer supuestos sobre la forma en la que x y u interactuan.

Estimacion

Minimos cuadrados ordinarios: $\operatorname{argmin} \sum (y_i - \hat{\alpha} - \hat{\beta}x_i)^2$

- La solucion es una recta en (X, Y) .
- La recta pasa 'por el medio' (los errores son penalizados en forma simetrica)

Minimos desvios absolutos: $\operatorname{argmin} \sum |y_i - \hat{\alpha} - \hat{\beta}x_i|$

- La solucion es una recta en (X, Y) .
- La recta pasa 'por el medio' (los errores son penalizados en forma simetrica)

Koenker y Bassett (1978): la solucion al problema pasa por penalizar **asimetricamente** los errores de estimacion.

$$\hat{\beta}(\tau) = \operatorname{argmin} \sum_{i=1}^n \rho_{\tau}(y_i - x_i' \beta)$$

con $\rho_{\tau}(z) \equiv z(\tau - I(z < 0))$, $\tau \in (0, 1)$, produce estimadores consistentes y asintoticamente normales.

- Penaliza a los errores positivos con τ y a los negativos con $1 - \tau$
- Cuando $\tau = 0.5$, minimos desvios absolutos, penalizacion simetrica

Cuestiones computacionales

El problema de estimacion puede ser reescrito como:

$$\min_{(b(\tau), u, v) \in \mathbb{R} \times \mathbb{R}_+^{2n}} \{ \tau 1_n' u + (1 - \tau) 1_n' v \mid 1_n' b(\tau) + u - v = y \}$$

$e_i \equiv x_i' b(\tau)$, 1_n un vector columna de n , u y v son variables de holgura complementaria: es un **programa lineal**.

En la practica estimamos $\hat{\beta}(\tau_i), i = 1, \dots, m$ (coeficientes para m cuantiles)

Denominemos con $\beta(\tau_i)$ a los coeficientes poblacionales y construyamos los siguientes vectores.

- $\hat{\beta} \equiv (\hat{\beta}(\tau_1)' \cdots \hat{\beta}(\tau_m'))'$
- $\beta \equiv (\beta(\tau_1)' \cdots \beta(\tau_m'))'$

Inferencia

Bajo el supuesto de que la muestra es independiente (pero no necesariamente idénticamente distribuida) y bajo condiciones de regularidad estándar, es posible mostrar que:

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \Lambda)$$

$\Lambda = \Lambda_{j,p}$, $j = 1, \dots, m$, $p = 1, \dots, m$ con:

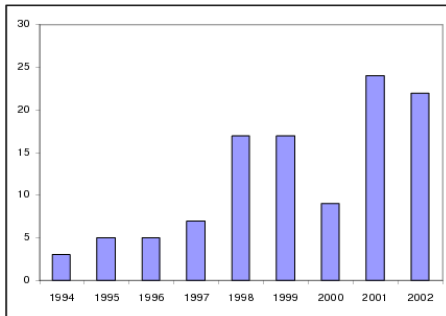
$$\Lambda_{j,p} = (\min\{\tau_j, \tau_p\} - \tau_j \tau_p) \left(E[f_{\tau_j}(0|x) x x'] \right)^{-1} E[x x'] \left(E[f_{\tau_j}(0|x) x x'] \right)^{-1}$$

en donde $f_{\tau_s}(0|x)$ es la función de densidad de y condicional en $x' \beta(\tau_s)$.

Alternativas:

- Estimar Λ , pasa por estimar $f_{\tau_s}(0|x)^{-1}$ (*sparsity*) no parametricamente.
- Metodo de inversion de test de rangos (Koenker, 1994).
- Bootstrap (Buchinsky, 1998)

Evolucion de la literatura



Fuente: Econlit

Teoria

Seminal paper: Koenker y Bassett (1978, Econometrica)

- Datos censurados: Powell (1986, JoE), Portnoy (2003, JASA).
- Selectividad: Buchinsky (2001, EE)
- Regresion no-lineal: Koenker y D'Orey (94, JoE), Lee (2003, ET).
- Datos en paneles: Koenker (2004), Lamarche (2004).
- Modelos binarios: Kordas (2003), Belluzo (2004, JBES).

- Endogeneidad y variables instrumentales: Abadie, Angrist e Imbens (2002, Econometrica), Arias, Hallock y Sosa Escudero (EE, 2001).
- Raices unitarias y series temporales (Hassan y Koenker, 2000, Econometrica)
- Algoritmos: Fitzenberger (2001, JoE), Biliass et al (2000, JoE).
- Modelos de duracion: Koenker y Goessling (2003, JASA).
- Modelos mal especificados (Angrist et al (2004), Kim y White (2002))

Aplicaciones

Seminal papers: Buchinsky (1994), Chamberlain (1994), ecuaciones de Mincer.

- Ecuaciones de Mincer (Buchinsky, 94, 98).
- Frontera estocástica: Bernini, et al. (2004), EE.
- Modelos de crecimiento: Canarella y Pollard (2004, JDE).
- Brecha de géneros: García et al. (2001)
- Asimetría y volatilidad en finanzas: Park (2002, EcJ), Bassett et al. (2002)
- Tamaño de firmas al inicio: Machado y Mata (2000, JAE).
- Efecto del tamaño de clase: Levin (2001)

- Demanda de alcohol: Manning (1995)
- Descomposicion de desigualdad: Martins y Pereira (LabEc, 2004), Machado y Mata (JAE, 2004)
- CAVIAR (conditional value at risk): Engle y Manganelli (2000).
- Pobreza: Arias y Sosa Escudero (2004).
- Willingness to pay: Belluzo (2004, JBES).
- Imputacion de rentas (Gasparini y Sosa Escudero, 2005, JIncDist).
- Brecha publico-privado (Mueller (98), Poterba y Reuben (1995)).

Libros y articulos generales

Articulos generales:

- Koenker (2000, JoE): nota historica
- Koenker y Hallock (2001, JEPersp): Articulo general
- Buchinsky (1998, JHR): Articulo general.
- Cade y Noon (2003, FrontEcology): Introduccion en ecologia.

Textos:

- Machado, Koenker y Fitzenberger (2001): *Economic Applications of Quantile Regression*, Springer-Verlag, texto de aplicaciones.
- Koenker, R. (2005): *Quantile Regression*, Cambridge University Press.

Software

- Splus y R
- Stata