Regresja na kwantylach

Jerzy Mycielski

2008

Regresja dla wartości oczekiwanej

- W przypadku regresji dla wartości oczekiwanej modelujemy zależność między wartością oczekiwaną a zmiennymi objaśniającymi
- Hiperpłaszczyzna regresji y na $\mathbf{x}$ dla wartości oczekiwanej jest warunkową wartością oczekiwaną

$$\mathrm{E}\left(y|\,\mathbf{x}\right)=\mu\left(\mathbf{x}\right)$$

- Intuicyjnie, hiperpłaszczyzna regresji podaje zależność między oczekiwaną wielkością y dla znanego $\mathbf{x}$

Regresja dla kwantyli

Bezwarunkowa

- Kwantyl τ jest taką wielkością która jest większa lub równa od y z prawdopodbieństwem τ.
- Bezwarunkowy kwantyl τ będziemy więc definiować jako takie ξ, dla którego

$$F\left(\xi_\tau\right) = \tau.$$

- Dla znanej postaci dystrybuanty F kwantyl τ można policzyć jako

$$\xi_\tau = F_y^{-1}\left(\tau\right).$$

- W przypadku regresji na kwantylach modelujemy zależność między wielkością kwantyla τ a wielkością zmiennych niezależnych.
- Warunkowy kwantyl τ można więc zdefiniować jako

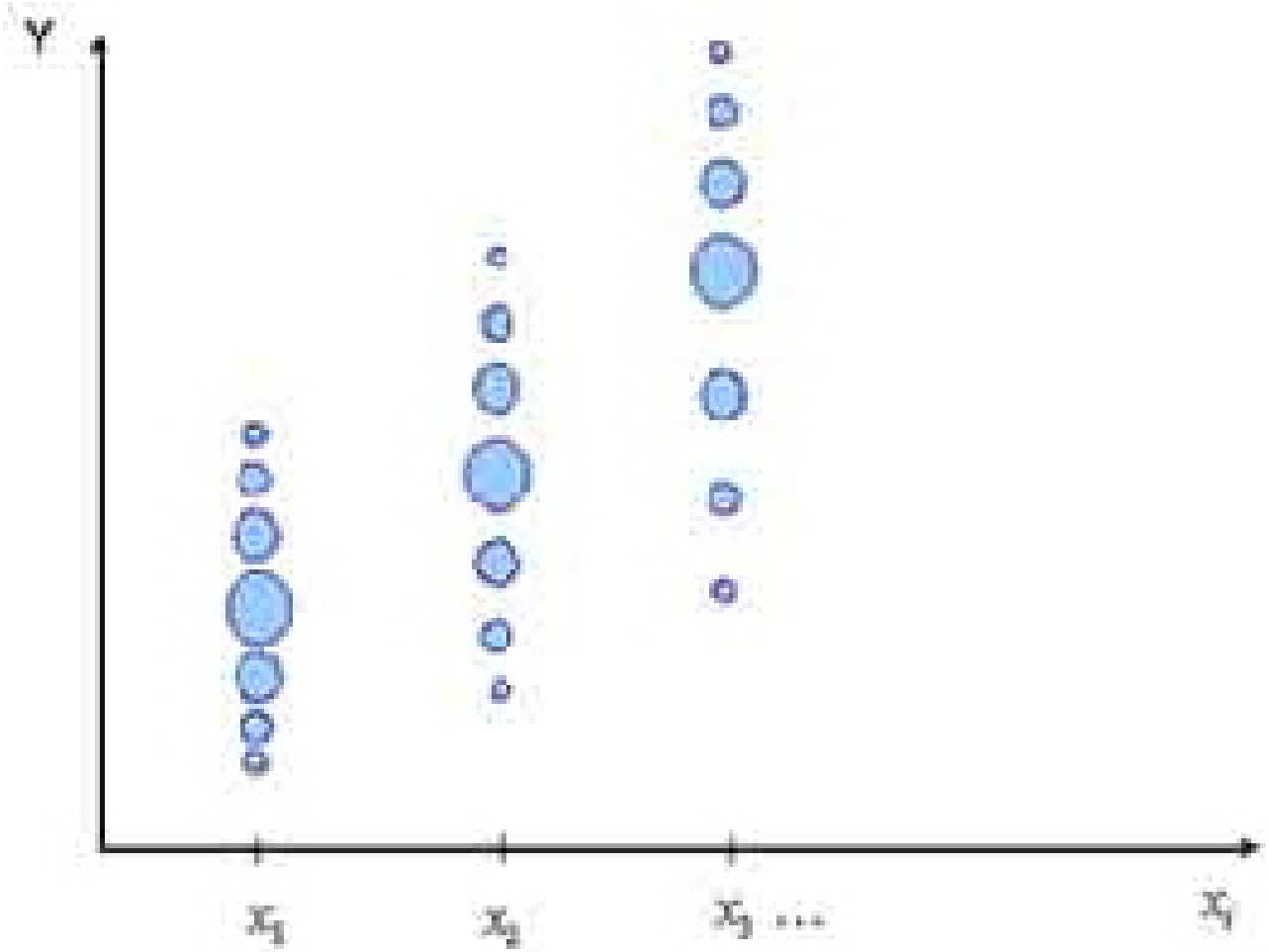
$$F_y\left(\xi_\tau \middle| \mathbf{x}\right) = \tau.$$

- Wynika z tego, że

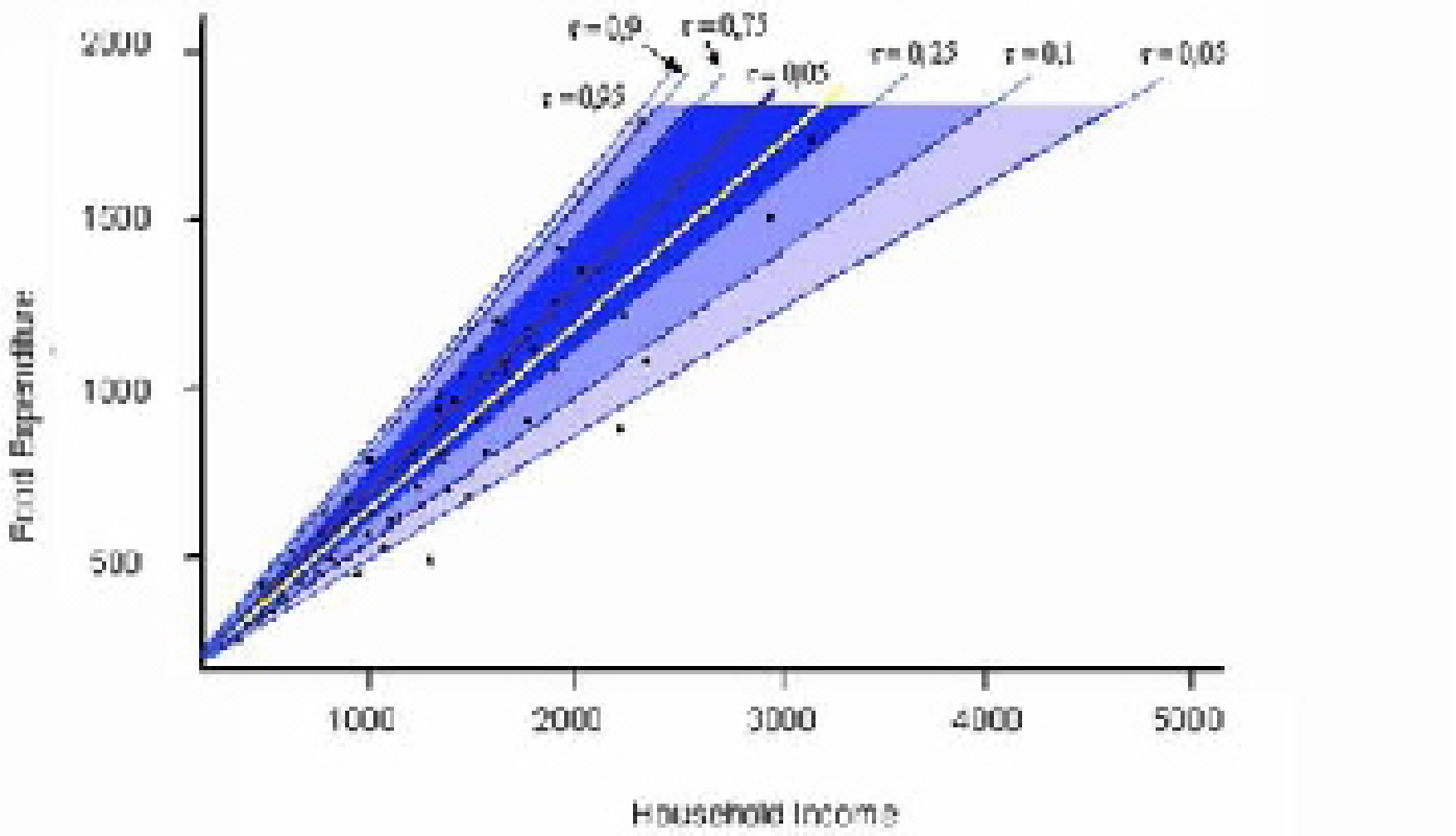
$$\xi_\tau\left(\mathbf{x}\right) = F_y^{-1}\left(\tau \middle| \mathbf{x}\right)$$

- Zauważmy, że dla każdego kwantyla, postać funkcji regresji może wyglądać nieco inaczej

Kwantyle



Krzywe Engla



Estymacja parametrów - regresja na wartości oczekiwanej

- W przypadku regresji dla wartości oczekiwanej standardowo stosuje się metodę najmniejszych kwadratów.
- Zakładamy, że $\mu(\mathbf{x})$ jest dana funkcją parametryczną $\mu(\mathbf{x}, \boldsymbol{\beta})$ o nieznanym wektorze parametrów $\boldsymbol{\beta}$.
- Wektor $\boldsymbol{\beta}$ znajdujemy minimalizując sumę kwadratów reszt:

$$\min_{\boldsymbol{\beta} \in \mathbf{R}^K} \sum_{i=1}^{N} \left[y_i - \mu(\mathbf{x}_i, \boldsymbol{\beta}) \right]^2$$

Estymacja parametrów - regresja na kwantylach

Mediana

- W przypadku regresji na kwantylach estymację można także sprowadzić do rozwiązywania problemu maksymalizacyjnego.
- Rozpatrzmy najprostszy przypadek szukania mediany y ($\tau = 0.5$) poprzez rozwiązywanie następującej funkcji celu

$$\min_{\xi_{0.5} \in \mathbf{R}} \sum_{i=1}^{N} |y_i - \xi_{0.5}|$$

- Warunki pierwszego rzędu dla tego problemu (dla N nieparzystego i $y_i \neq y_i$ dla $i \neq j$) będą miały następującą postać:

$$\sum_{i=1}^{N} \left[\mathbb{I}\left(y_i > \xi_{0.5}\right) - \mathbb{I}\left(y_i < \xi_{0.5}\right) \right] = 0$$

lub

$$\frac{\sum_{i=1}^{N} \mathbb{I}\left(y_i > \xi_{0.5}\right)}{\sum_{i=1}^{N} \mathbb{I}\left(y_i < \xi_{0.5}\right)} = 1$$

- Warunek pierwszego rzędu będzie więc spełniony dla $\xi_{0.5}$ równemu y_i, które jest medianą z próby.
- W ogólniejszym przypadku estymacji parametrów regresji na medianie rozwiązujemy następujący problem maksymalizacyjny:

$$\min_{\beta \in \mathbf{R}^K} \sum_{i=1}^{N} |y_i - \xi_{0.5}(\mathbf{x}, \beta)|$$

- Zauważmy, że estymator dla regresji dla mediany jest równoważny estymatorowi *LAD* (**L**east **A**bsolute **D**eviation)

Estymacja parametrów - regresja na kwantylach

- Dla dowolnego τ oszacowanie kwantyla z próby można znaleźć rozwiązując zadanie:

$$\min_{\xi_\tau \in \mathbf{R}} \sum_{i=1}^{N} \rho_\tau \left(y_i - \xi_\tau \right)$$

gdzie

$$\rho_\tau (x) = \begin{cases} \tau x & \text{dla} \quad x > 0 \\ -(1-\tau) x & \text{dla} \quad x < 0 \end{cases}$$

- Warunki pierwszego rzędu w tym przypadku będą miały postać

$$\frac{\sum_{i=1}^{N} \mathbb{I} \left(y_i > \xi_\tau \right)}{\sum_{i=1}^{N} \mathbb{I} \left(y_i < \xi_\tau \right)} = \frac{1-\tau}{\tau}$$

- Estymatorem jest odpowiedni kwantyl empiryczny.

- W ogólniejszym przypadku estymacji parametrów regresji na medianie rozwiązujemy następujący problem maksymalizacyjny:

$$\min_{\beta \in \mathbf{R}^K} \sum_{i=1}^{N} \rho_\tau \left(y_i - \xi_\tau \left(\mathbf{x}, \beta \right) \right)$$

- Rozwiązanie tego problemu maksymalizacji znajduje się metodami programowania liniowego (simplex)

Regresja na kwantylach - zalety

- Przeprowadzając regresję na szeregu kwantyli, opisujemy zależność całego rozkładu zmiennej zależnej od zmiennych niezależnych a nie tylko wartości oczekiwanej zmiennej zależnej
- Heteroskedastyczność można łatwo wykryć analizując wyniki regresji na kwantylach
- W przypadku występowania heteroskedastyczności estymacja regresji na medianie może okazać się bardziej efektywnym sposobem szukania wartości parametrów niż regresja na wartości oczekiwanej
- Kwantyl zmiennej losowej przekształconej za pomocą przekształcenia monotonicznie rosnącego $g(y)$ jest równy przekształconemu w ten sam sposób kwantylowi oryginalnej zmiennej

$$\xi_{\tau}(\mathbf{x}) = F^{-1}_{g(y)}(\tau|\,\mathbf{x}) = g\left(F_y^{-1}(\tau|\,\mathbf{x})\right)$$

- Regresja na kwantylach jest odporna na problem outlier'ów (obserwacji błędnych)

Regresja na kwantylach - trudności

- Największy problem związany z regresją na kwantylach związany jest z brakiem wzorów analitycznych na macierz wariancji i kowariancji uzykanych oszacowań.
- W związku z tym trudniejsze jest znajdywanie przedziałów ufności oraz weryfikacja hipotez statystycznych
- Oszacowania macierzy wariancji i kowarinacji można znaleźć metodą bootstrap