A Resampling Method on Pivotal Estimating Functions

Kun Nie

Biostat 277,Winter 2004

March 17, 2004

Outline

- Introduction

- A General Resampling Method

- Examples

  - Quantile Regression

  -Rank Regression

  -Simulation Study

- Discussions

Introduction

Def 1. M-estimate (P.J., Bickel, K.A., Docksum): Suppose $i.i.d. X_1, \cdots, X_n$ are distributed to P_θ. Write $P = \{P_\theta : \theta \in \Theta\}$, where Θ is an open set in R. Let $\rho : X \times \Theta \to R$ where

$$D(\theta, \theta_0) = E_{\theta_0}(\rho(X_1, \theta) - \rho(X_1, \theta_0))$$

is uniquely minimized at θ_0. Let $\widehat{\theta_n}$ be the minimum contrast estimate such that

$$\widehat{\theta_n} = argmin \frac{1}{n} \sum_{i=1}^{n} \rho(X_i, \theta).$$

Suppose $\psi = \frac{\partial \rho}{\partial \theta}$ is well defined, then

$$S_X(\theta) \hat{=} \frac{1}{n} \sum_{i=1}^{n} \psi(X_i, \theta) = 0 \qquad (1)$$

when $\theta = \widehat{\theta_n}$. $\widehat{\theta_n}$ is an M-estimate.

Discussions on M-estimate:

- Solutions to equation (1) are called M-estimate. We even do not require that $\widehat{\theta_n}$ is a minimum contrast.

- If ψ is differentiable, then under certain conditions the distribution of $\widehat{\theta_n}$ is approximately normal,

  - the asymptotic mean is θ_0

  - the asymptotic variance is

$$(E_P(\frac{\partial \psi}{\partial \theta})(X_1, \theta(P)))^{-1} \times var(\psi(X_1, \theta(P))) \times (E_P^{'}(\frac{\partial \psi}{\partial \theta})(X_1, \theta(P)))^{-1}$$

- However, under a semi-parametric model setting the estimating function ψ is not smooth, and it is difficult to calculate the asymptotic variance of $\widehat{\theta_n}$ with the above formula.

Example (Koenker and Bassett, 1978)

Model: $Y_i = \beta' z_i + \varepsilon_i$

where ε_i's are assumed to be independent but may not be identically distributed. The distribution of ε_i is not specified. The median of ε_i is 0. A commonly used estimating function S for β is:

$$n^{-1/2} \sum_{i=1}^{n} z_i (I(Y_i - \beta' z_i \leq 0) - 1/2).$$

S is not continuous.

Q: How to make inference on β?

A New Resampling Method

Suppose that the distribution (or limit distribution) of random vector $S_X(\beta_0)$ can be generated by a $p \times 1$ random vector U, whose distribution is completely known or can be estimated consistently. Parzen et. al. proposed the following procedure:

For $j = 1, \cdots, M$,

- Step 1: generate random sample u_j from U

- Step 2: solve the equation $S_X(\beta) = u_j$ and get a solution β_{u_j}

When M is large (e.g., 1000), the empirical distribution of β_U can be obtained.

Theorem 1. Let n be the sample size for X. If there exist a sequence of constants c_n and a nonsingular matraix A such that,

A1:

$$sup(\frac{\|S_X(\beta) - S_X(\beta^*) - An^{1/2}(\beta - \beta^*)\|}{1 + n^{1/2}\|\beta - \beta^*\|}) \to 0$$

almost surely, where β, β^* are in $U(\beta_0, c_n)$. Furthermore, for $\|\beta - \beta_0\| \geq c_n$,

A2:

$$inf\|S_X(\beta)\| = \gamma_n \to \infty$$

Then the asymptotic conditional distribution of $n^{1/2}(\tilde{\beta} - \beta_U)$ given X is asymptotically identical to the asymptotic distribution of $n^{1/2}(\widehat{\beta} - \beta_0)$, where $\tilde{\beta}$ is a realization of $\widehat{\beta}$ after observing X. More specifically, they are asymptotically distributed as $-A^{-1}U$.

Example 1: Heteroscedastic Quantile Regression

Model: $Y_i = \beta^{'} z_i + \varepsilon_i$

where ε_i's are assumed to be independent but may not be identically distributed. $\beta_o^{'} z_i$ is the 100τth percentile of Y_i. The distribution of ε_i is not specified. The estimating function for β is:

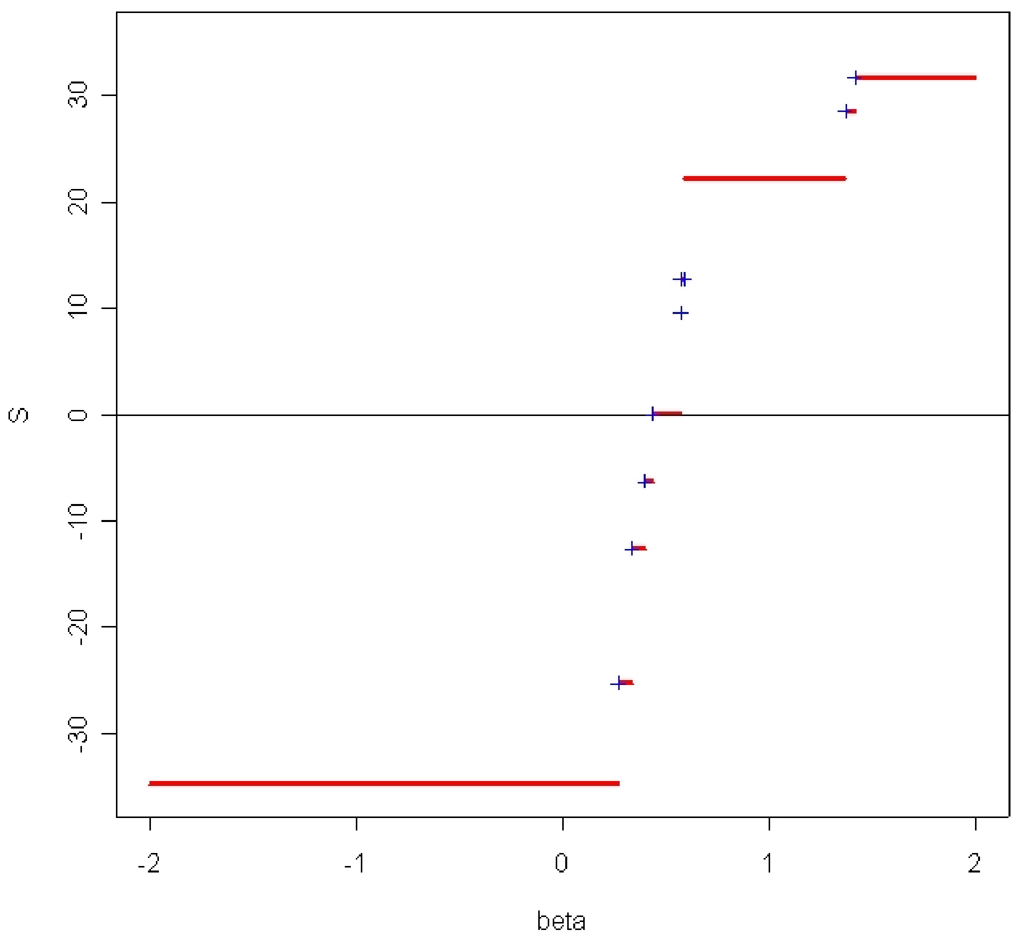
$$S_X = n^{-1/2} \sum_{i=1}^{n} z_i (I(Y_i - \beta^{'} z_i \leq 0) - \tau). \quad (2)$$

To solve the equation $S_X(\beta) = 0$, we turn to solve the following minimizing problem (Bassett and Koenker, 1982):

$$\rho = -\sum_{i=1}^{n} (Y_i - \beta^{'} z_i)(I(Y_i - \beta^{'} z_i \leq 0) - \tau) \quad (3)$$

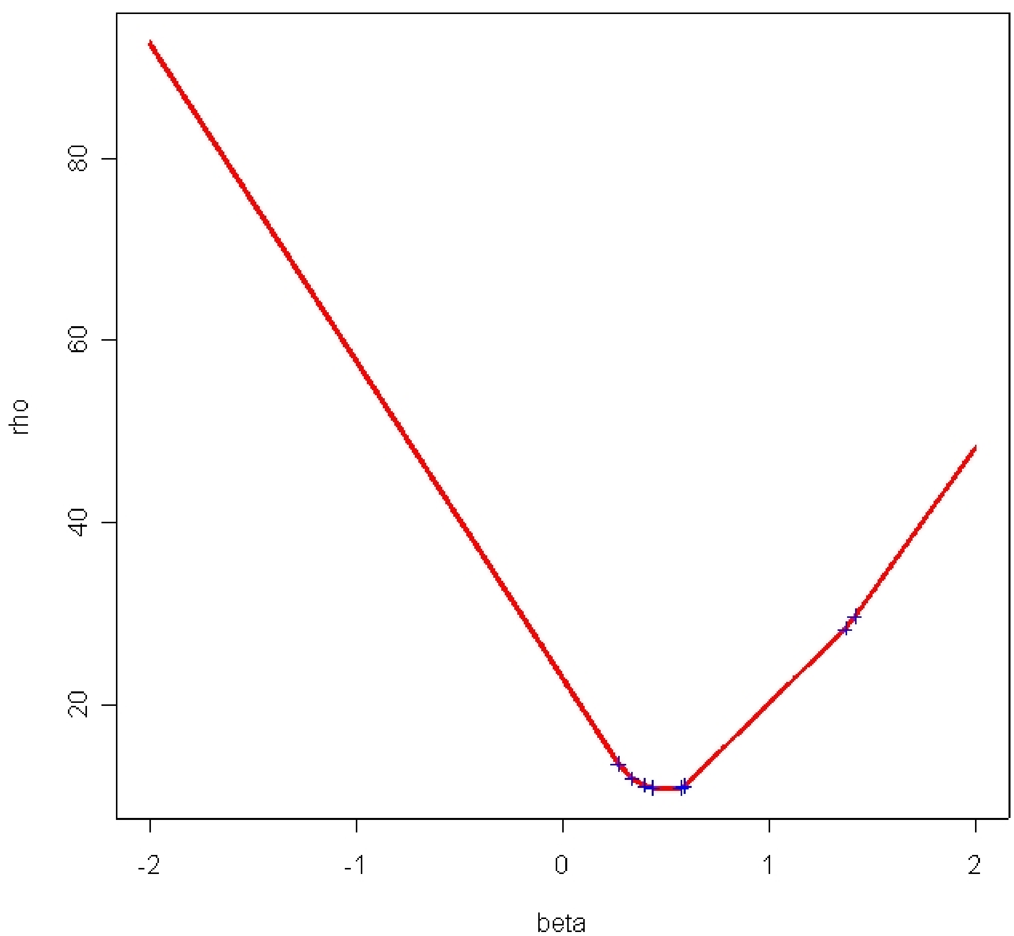
(Quantile regression model in S-Plus/R/STATA)

Illustration: By setting $z = (1, 1, 1, 2, 2, 2, 3, 3, 3, 4)$, and $y_i = 0.5z_i + \epsilon_i$, where ϵ_i are $i.i.d.\ N(0, 1)$, we got the following plot of $S_X(\beta)$ when $\tau = 0.5$:



Note: $y = (1.41, 0.57, 2.52, 0.87, 2.74, 0.80, 1.76, 1.71, 0.81, 1.35)$ in this example

Plot of $\rho(\beta)$



Note: ρ is continuous, nonnegative and convex.

Resampling Procedure:

for $j = 1, ..., M$,

- Step 1: generate $\xi_1, \cdots, \xi_n \sim Bernoulli(\tau)$,

- Step 2:$u_j = n^{-1/2} \sum_1^n z_i(\xi_i - \tau)$

- Step 3: Solve equation $S_X(\beta) = u_j$ and get the solution β_{u_j} by:

  -Let $(Y_{n+1}, z_{n+1}) = (N, n^{1/2}u/\tau)$, where N is a large number s.t. $I(Y_{n+1} - \beta' z_{n+1} \leq 0)$ is always 0.

  -Solve $S_X^* = n^{-1/2} \sum_{i=1}^{n+1} z_i(I(Y_i - \beta' z_i \leq 0) - \tau) = 0$ equivalently.

Thus we get the empirical distribution of β_{u_j}.

Example 2: Rank Regression

Again, assume that $Y_i = \beta^{'} z_i + \epsilon_i$, but ϵ_i's are $i.i.d.$, and β does not include the intercept term. The estimating function $S_X(\beta)$ based on ranks is:

$$S_X = \sum_{i=1}^{n} (z_i - \overline{z})\phi(R(Y_i - \beta^{'} z_i)), \qquad (4)$$

where

-ϕ is an increasing function

-R is the rank function for $\{Y_1 - \beta^{'} z_1, \cdots, Y_n - \beta^{'} z_n\}$.

Then $\widehat{\beta}$, the solution to $S_X(\beta) = 0$, is a minimizer of the following function:

$$\rho = \sum_{i=1}^{n} \phi(R(Y_i - \beta^{'} z_i))(Y_i - \beta^{'} z_i - \bar{Y} + \beta^{'} \overline{z}) (5)$$

An efficient program called RREGRESSION is available to minimize ρ.

Resampling Procedure

for $j = 1, ..., M$,

- Step 1: generate $(\eta_1, \cdots, \eta_n)$ from random permutation of $(1, \cdots, n)$,

- Step 2:$u_j = \sum_1^n (z_i - \bar{z})\phi(\eta_i)$

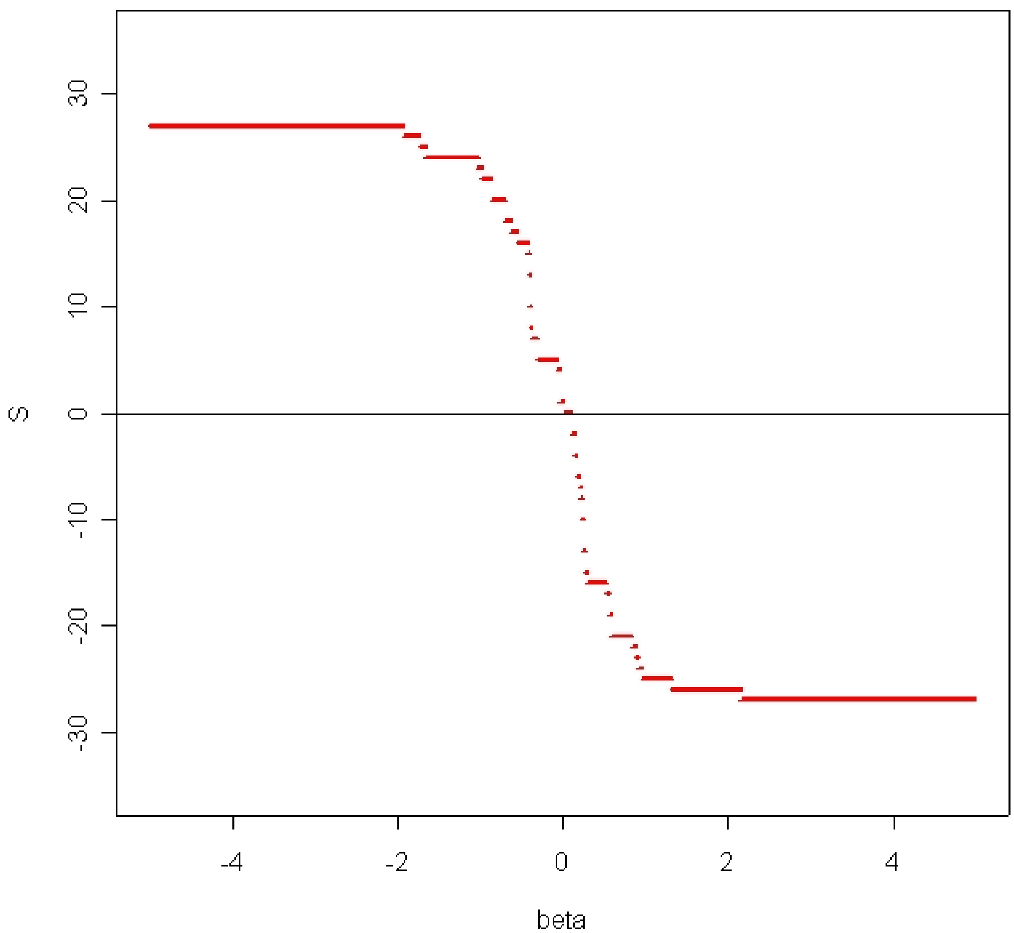
- Step 3: Solve equation $S_X(\beta) = u_j$ and get the solution β_{u_j} by:

  -Let $(Y_{n+1}, z_{n+1}) = (N, \bar{z} - (n+1)u/[n(\phi(n+1) - \bar{\phi}))]$, where N is a large number s.t. $R(Y_{n+1} - \beta' z_{n+1}$ is always $n + 1$.

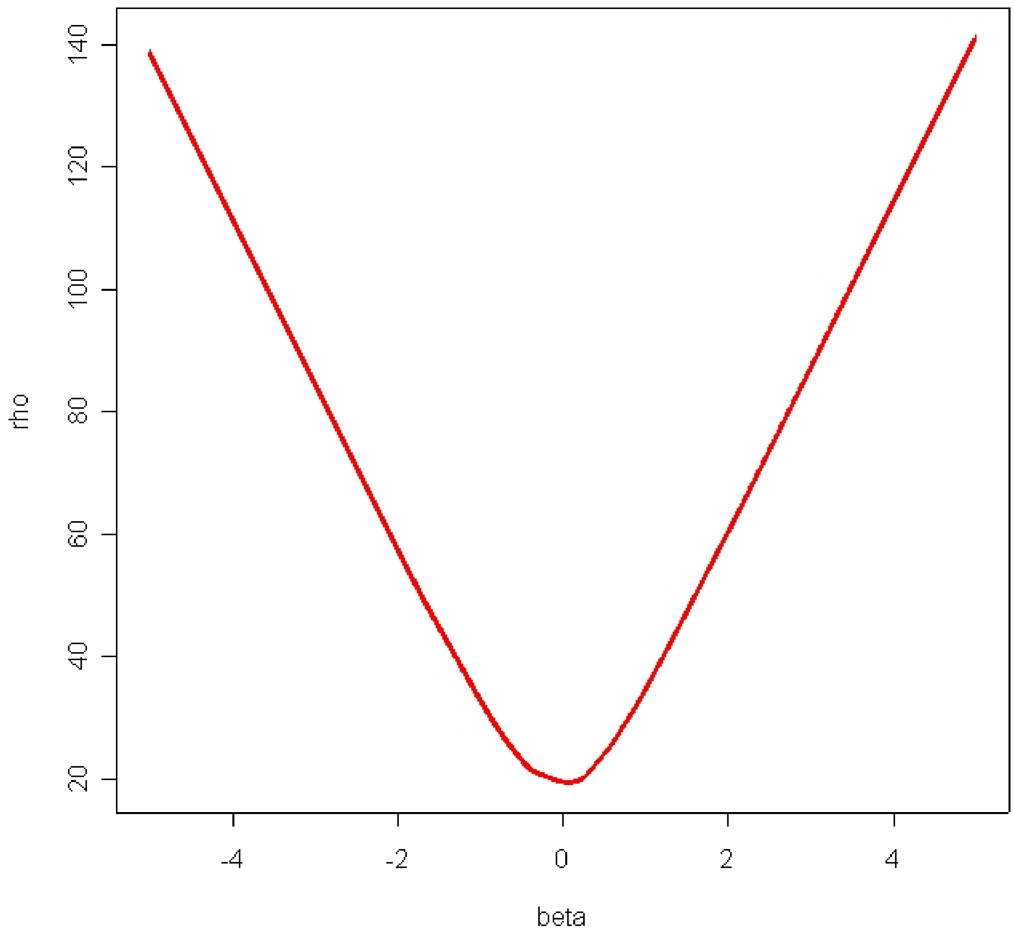
  -Solve $S_X^* = \sum_{i=1}^{n+1} (z_i - \bar{z})\phi(R(Y_i - \beta' z_i)) = 0$ equivalently.

Illustration

Plot of $S_X(\beta)$

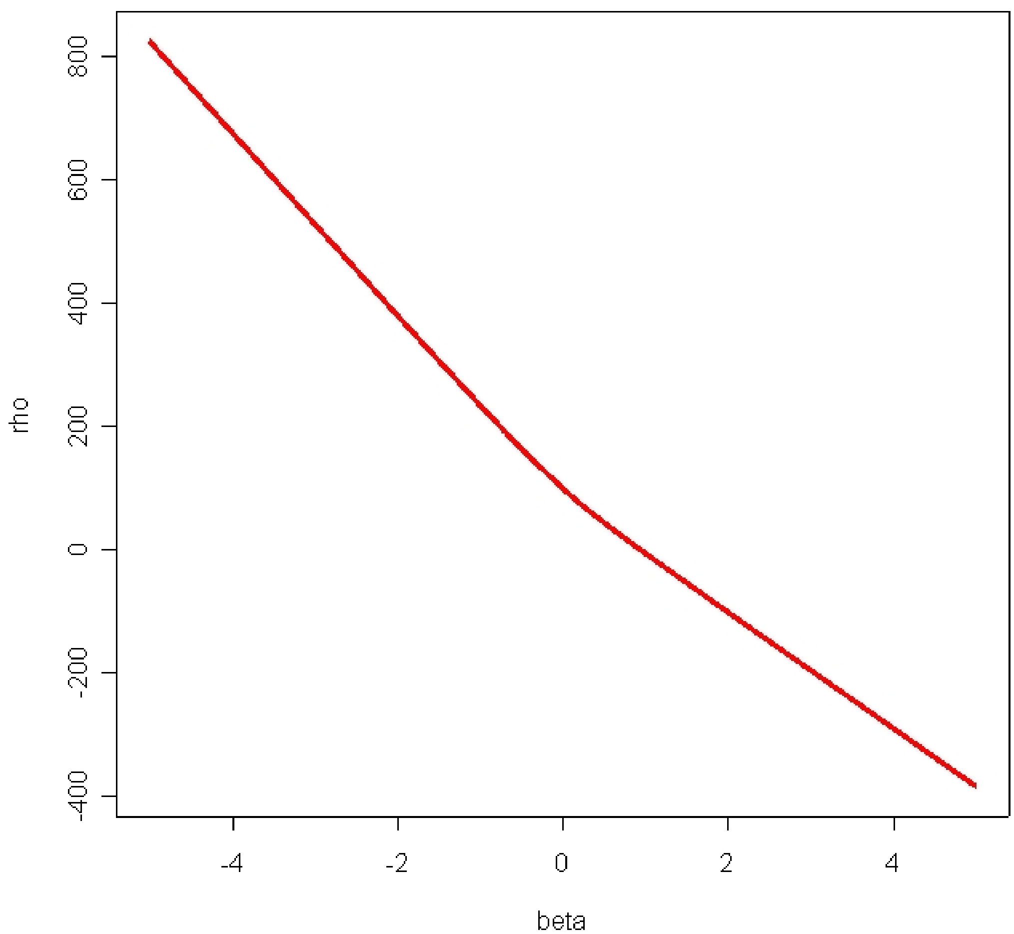


Plot of $\rho(\beta)$



Note: ρ is continuous, nonnegative and convex.

Wrong ρ in the paper



Simulation Study

(Median regression Model) 1000 samples $\{(y_i, z_i), i = 1, \cdots, 50\}$ with $\beta_0 = (0, 1, 1)$ were generated. The 1st components of z_i were all 1s and the 2nd components of z_i were Bornoulli with success probability 0.5. The 3rd component of z_i were i.i.d. standard normal. The C.I.'s of the 3rd component of β were obtained by 1000 resamplings of U. Table 1 was based on 1000 simulations.

Table 2. *Empirical coverage probabilities* (ECP) *and estimated mean lengths* (EML) *for various interval procedures*

(*a*) *Gaussian error with mean* 0 *and variance* 0·5

Confidence level		Resample		Bootstrap		STATA	
		ECP	EML	ECP	EML	ECP	EML
0·95	S	0·95	0·62	0·95	0·59	0·97	0·58
	P	0·98	0·62	0·98	0·59		
	B	0·93	0·64	0·94	0·61		
0·90	S	0·92	0·51	0·91	0·49	0·94	0·49
	P	0·95	0·51	0·94	0·49		
	B	0·90	0·53	0·89	0·51		
0·85	S	0·88	0·45	0·86	0·43	0·91	0·43
	P	0·91	0·43	0·89	0·42		
	B	0·87	0·47	0·85	0·45		

(*b*) *Lognormal error with mean $e^{0\cdot25}$ and variance $(e - e^{0\cdot5})$*

Confidence level		Resample		Bootstrap		STATA	
		ECP	EML	ECP	EML	ECP	EML
0·95	S	0·97	0·62	0·96	0·60	0·97	0·59
	P	0·98	0·62	0·98	0·60		
	B	0·95	0·65	0·93	0·63		
0·90	S	0·92	0·52	0·91	0·50	0·95	0·49
	P	0·95	0·51	0·94	0·49		
	B	0·88	0·54	0·87	0·52		
0·85	S	0·88	0·46	0·87	0·44	0·92	0·43
	P	0·91	0·44	0·89	0·42		
	B	0·85	0·47	0·82	0·46		

Note: based on 1000 simulations

(*c*) *Gaussian error with mean 0 and heteroscedastic variance*

Confidence level		Resample ECP	Resample EML	Bootstrap ECP	Bootstrap EML	STATA ECP	STATA EML
0·95	S	0·95	0·66	0·95	0·65	0·60	0·29
	P	0·97	0·65	0·95	0·64		
	B	0·94	0·66	0·93	0·64		
0·90	S	0·91	0·55	0·90	0·54	0·53	0·24
	P	0·92	0·53	0·91	0·53		
	B	0·90	0·55	0·88	0·53		
0·85	S	0·87	0·48	0·86	0·47	0·47	0·21
	P	0·87	0·47	0·87	0·46		
	B	0·85	0·48	0·83	0·47		

S, standard method; P, percentile method; B, bias correction method.

Note: based on 1000 simulations

Discussions

- The proposal is useful when the point estimate $\widehat(\beta)$ can be easily obtained but its variance is difficult to estimate by conventional method

- There is no analytical proof that the traditional bootstrap method is valid for general quantile regression model.

- When the error terms are heteroscedastic, conventional quantile regression method in STATA could be bad, while the method proposed in this paper performs well.

- The method proposed in this paper has potentials to real data analysis.

Reference

M.I. Parzen, L.J.Wei,Z. Ying, 1994, *A Resampling Method Based on Pivotal Estimating Functions*,Biometrika, Vol 81, 341-350

P.J.Bickel, K.A. Doksum, 2001, *Mathematical Statistics*, Prentice Hall

B. Efron, R.J.Tibshirani, 1993, *An Introduction to the Bootstrap*, Chapman & Hall

Keonker, R Bassett, G, 1978, *Regression Quantiles*, Economitrica 84, 33-50

Keonker, R Bassett, G, 1982, *An Empirical Quantile Function for linear Models with i.i.d. Errors* , JASA 77, 407-15