응용통계연구 (2012)
25(2), 341–350
DOI: http://dx.doi.org/10.5351/KJAS.2012.25.2.341

복합 분위수 회귀에 대한 붓스트랩 방법의 응용

서강민[1] · 방성완[2] · 전명식[3]

[1]고려대학교 통계학과, [2]육군사관학교 수학과, [3]고려대학교 통계학과

(2012년 2월 29일 접수, 2012년 4월 6일 수정, 2012년 4월 9일 채택)

요 약

선형 회귀모형에서 오차항들이 서로 독립이고 동일한 분포를 따른다고 가정할 경우, (회귀계수의 강건한 추정을 위하여) 모든 분위수 함수의 회귀계수가 동일한 값을 갖는다는 사실에 근거한 복합 분위수 회귀(composite quantile regression) 방법을 고려할 수 있다. 본 논문에서는 복합 분위수 회귀에서 사용되는 분위수의 개수를 선택하기 위해 붓스트랩 방법의 가능성을 검토하였다. 또한, 분위수 회귀와 복합 분위수 회귀의 성능을 비교하기 위해 붓스트랩 방법을 이용하여 신뢰구간을 구축하고, 이들의 포함확률과 평균길이를 비교하였다. 이러한 모의실험을 통하여 복합 분위수 회귀의 우월성과 통계적 추론에 있어서 붓스트랩 방법의 유용성을 확인하였다.

주요용어: 분위수 회귀, 복합분위수 회귀, 붓스트랩.

1. 서론

p차원 설명변수 $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$와 1차원 반응변수 $y_i \in R$로 이루어진 크기가 n인 (훈련)자료 $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$에 대하여 다음의 선형 회귀모형

$$y_i = \boldsymbol{x_i}'\boldsymbol{\beta} + \epsilon_i, \quad i = 1, \ldots, n \tag{1.1}$$

을 고려하자. 여기서 $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$는 회귀계수 벡터이고, $\{\epsilon_i\}_{i=1}^n$는 평균이 0인 분포 F로부터의 서로 독립인 오차항이다. 편의상, 설명변수들은 중심화 되었다고 가정하자. 회귀모형 (1.1)에서 관심이 되는 것은 회귀계수 벡터 $\boldsymbol{\beta}$에 대한 추정이며, 일반적으로 제곱 손실함수를 사용한 최소제곱추정법(least square estimation; LSE)이 계산상의 이점에 힘입어 보편적으로 널리 이용되고 있다. 그러나 최소제곱추정법은 이상치가 존재하거나 꼬리가 두꺼운 오차항의 분포에서는 추정의 효율이 떨어진다. 반면에 절대 손실함수를 일반화한 check 손실함수를 사용하는 분위수 회귀(quantile regression; QR)는 이상치에 강건한 특성을 지니고 있다. 따라서 최소제곱추정법이 적절하지 않을 경우, 이에 대한 대안으로 Koenker과 Bassett (1978)에 의해 제안된 분위수 회귀를 고려할 수 있다.

선형 회귀모형 (1.1)에서 반응변수 y에 대한 $100\tau\%$ 조건부 분위수 함수(conditional quantile function of $y|\boldsymbol{x}$) $q_\tau(y|\boldsymbol{x})$는

$$q_\tau(y|\boldsymbol{x}) = \sum_{j=1}^{p} x_j\beta_j + b_\tau, \quad \text{단} \;\; b_\tau = F^{-1}(\tau) \tag{1.2}$$

이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(No. 2010-0009204)이며 제1저자의 석사학위논문을 바탕으로 추가 연구하여 작성한 것입니다.
[3]교신저자: (136-701) 서울시 성북구 안암동 5가, 고려대학교 통계학과, 교수. E-mail: jhun@korea.ac.kr

와 같이 정의되며, 회귀계수 벡터 $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$는 check 손실함수 $\rho_\tau(t) = t(\tau - I(t < 0))$를 이용하여

$$\left(\hat{b}_\tau, \hat{\boldsymbol{\beta}}\right)^{QR} = \arg\min_{b,\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_\tau \left(y_i - b - \boldsymbol{x}_i'\boldsymbol{\beta}\right) \tag{1.3}$$

와 같이 추정된다. 선형 회귀모형에서 오차항들이 서로 독립이고 동일한 분포를 따를 때, 모든 τ값에 대한 조건부 분위수 함수 $q_\tau(y|\boldsymbol{x})$는 식 (1.2)에서처럼 동일한 회귀계수의 값을 갖는다. 이러한 사실에 기초하여 서로 다른 $\tau_k\,(k = 1, 2, \ldots, K)$에 대한 여러 개의 분위수 함수를 동시적으로 추정하는 복합 분위수 회귀(composite quantile regression; CQR)가 회귀계수의 효율적인 추정을 위해 제안되었다 (Koenker, 1984; Zou와 Yuan, 2008). 그러나 복합 분위수 회귀에서는 회귀계수의 추정에 적합한 분위수의 개수 K가 오차항의 분포 형태에 따라 달라지므로 K의 선택이 중요한 문제이다 (Bradic 등, 2010).

본 논문에서는 모형오차(model error; ME)의 붓스트랩 추정치를 기준으로 복합 분위수 회귀의 회귀계수 추정에 적절한 분위수 개수 K를 선택하는 방법을 제안하였다. 나아가, 붓스트랩 방법을 활용한 신뢰구간을 분위수 회귀와 복합 분위수 회귀에 적용하고 신뢰구간의 포함확률과 길이를 통해 두 방법론을 비교하였다. 본 논문의 2장에서는 복합 분위수 회귀에 대해 소개하고, 추정량의 점근적 분포와 분위수 개수 K에 따른 상대효율을 알아보았다. 3장에서는 복합 분위수 회귀에서 사용되는 분위수 개수 K의 선택과 회귀계수의 신뢰구간 구축을 위한 붓스트랩 방법의 활용방안을 제안하고 모의실험을 실시하였다. 마지막으로 4장에서는 제안된 방법에 대한 활용 가능성을 설명하는 결론을 도출했다.

2. 복합 분위수 회귀

선형 회귀모형 (1.1)에서와 같이 오차항들이 서로 독립이며 동일한 분포를 따르면, 식 (1.2)의 분위수 함수 $q_\tau(y|\boldsymbol{x})$에 대한 회귀계수의 값은 모든 분위수 $\tau\,(0 < \tau < 1)$에서 동일하다. 이러한 사실에 근거하여 회귀계수의 효율적인 추정을 위하여 복합 분위수 회귀(CQR)가 제안되었으며, 회귀계수 벡터 $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$는

$$\left(\hat{b}_1, \ldots, \hat{b}_K, \hat{\boldsymbol{\beta}}\right)^{CQR} = \arg\min_{b_1,\ldots,b_K,\boldsymbol{\beta}} \sum_{k=1}^{K} \left\{ \sum_{i=1}^{n} \rho_{\tau_k} \left(y_i - b_k - \boldsymbol{x}_i'\boldsymbol{\beta}\right) \right\} \tag{2.1}$$

와 같이 추정된다 (Koenker, 1984). 여기서는 Zou와 Yuan (2008)에 의해 제시된 것과 마찬가지로 K개의 분위수를

$$\tau_k = \frac{k}{K+1}, \quad (k = 1, 2, \ldots, K) \tag{2.2}$$

와 같이 동일한 간격이 되도록 선정하였다. 복합 분위수 회귀의 추정식 (2.1)은 여러 개의 분위수 함수를 동시에 고려하여 회귀계수를 추정함으로써, 하나의 함수만을 고려하는 최소제곱추정법 또는 분위수 회귀의 추정식 (1.3)보다 효율적인 추정의 결과를 얻을 수 있다. Zou와 Yuan (2008)은 다음의 정칙조건

(1) $p \times p$ 양정치행렬 C는 $\lim_{n\to\infty} (1/n)\, X'X = C$이다. 단, X는 설계행렬(design matrix)이다.

(2) 랜덤오차 ϵ이 누적분포함수 F와 확률밀도함수 f를 가질 때, p차원 벡터 $\boldsymbol{u}$에 대하여 $\lim_{n\to\infty} (1/n) \sum_{i=1}^{n} \int_0^{u_0 + x'\boldsymbol{u}} \sqrt{n}[F(a + t/\sqrt{n}) - F(a)]dt = (1/2) f(a)(u_0, \boldsymbol{u}') \left(\begin{smallmatrix} 1 & 0 \\ 0 & C \end{smallmatrix}\right) (u_0, \boldsymbol{u}')$이 성립한다.

표 **2.1.** K에 따른 점근적 상대효율의 변화

오차항의 분포	K					
	1	3	5	9	19	99
$N(0,3)$	0.637	0.856	0.906	0.935	0.949	0.955
$\chi^2(3)$	0.848	1.355	1.529	1.665	1.757	1.816
$t(3)$	1.621	1.868	1.894	1.901	1.901	1.900
$0.9N(0,1)+0.1N(0,25)$	1.832	2.341	2.418	2.437	2.421	2.412

을 만족할 때, 회귀계수 벡터에 대한 복합 분위수 회귀의 추정량 $\hat{\boldsymbol{\beta}}^{CQR}$의 분포가

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}^{CQR}-\boldsymbol{\beta}\right)\xrightarrow{d}N_p\left(\mathbf{0},\Sigma_{CQR}\right) \tag{2.3}$$

와 같이 정규 근사됨을 보였다. 이 때, $\hat{\boldsymbol{\beta}}^{CQR}$의 공분산 행렬 Σ_{CQR}은

$$\Sigma_{CQR}=C^{-1}\frac{\sum_{k,k'=1}^{K}\min(\tau_k,\tau_{k'})(1-\max(\tau_k,\tau_{k'}))}{\left(\sum_{k=1}^{K}f(b_{\tau_k})\right)^2} \tag{2.4}$$

이다. 공분산 행렬의 식 (2.4)로 부터 복합 분위수 회귀는 오차항의 분포 형태와 그에 따른 분위수의 개수 K에 따라 추정의 효율이 달라짐을 알 수 있다. 따라서 회귀계수의 추정에 앞서 주어진 자료에 적절한 K를 선택하는 것이 중요한 문제가 된다 (Zou와 Yuan, 2008).

이제, 최소제곱추정법에 대한 복합 분위수 회귀의 점근적 상대효율(asymptotic relative efficiency; ARE)을 분위수 개수 K의 변화에 따라 확인하고자 한다. 보편적으로 널리 사용되는 최소제곱추정법(LSE)의 추정량 $\hat{\boldsymbol{\beta}}^{LSE}$의 점근적 분포는

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}^{LSE}-\boldsymbol{\beta}\right)\xrightarrow{d}N_p\left(\mathbf{0},\Sigma_{LSE}\right),\quad \text{단 } \Sigma_{LSE}=\sigma^2C^{-1},\ \mathrm{Var}(\epsilon)=\sigma^2 \tag{2.5}$$

이다. 따라서 공분산 행렬 Σ_{LSE}과 Σ_{CQR}의 대각합(trace)의 비, 즉 총분산의 비를 점근적 상대효율의 척도로 사용하였다. 다양한 오차항의 분포에 대한 점근적 상대효율을 살펴보기 위하여 $N(0,3)$ 분포, 비대칭분포의 예로 $\chi^2\,(df=3)$ 분포, 두꺼운 꼬리를 가지는 $t\,(df=3)$ 분포, 그리고 혼합된 분포의 예로 혼합정규분포 $0.9N(0,1)+0.1N(0,25)$를 사용하였으며 그 결과는 표 2.1에 주어져 있다.

표 2.1을 통해 오차항이 $N(0,3)$ 분포를 따르는 경우에는 예상대로 최소제곱추정법의 효율이 복합 분위수 회귀에 비해 좋은 것을 알 수 있다. 그러나 $N(0,3)$ 분포와 $\chi^2(3)$ 분포에서 $K=1$인 경우를 제외한 모든 오차항의 분포에서 복합 분위수 회귀의 상대효율이 월등히 좋은 것을 확인 할 수 있다. 일반적으로 표 2.1에서 고려된 오차항의 분포에 대하여 K가 증가할수록 처음부분에서는 최소제곱추정량에 대한 복합 분위수 회귀 추정량의 상대효율이 커지지만 충분히 큰 K를 지나서부터는 효율의 증가가 완만해지는 것을 확인할 수 있다. 특히 복합 분위수 회귀에서 최적의 K값은 항상 단조 증가하는 것이 아니며, 오차항의 형태에 따라 다른 것을 알 수 있다. 그러므로 복합 분위수 회귀를 회귀계수의 추정에 사용하기에 앞서 적절한 분위수의 개수 K를 선택하는 것은 중요한 문제로 남게 된다.

3. 붓스트랩의 활용

붓스트랩의 활용은 적절한 조건하에서 $(\hat{\boldsymbol{\beta}}_K^{CQR}-\boldsymbol{\beta})$의 분포가 붓스트랩 분포 $(\hat{\boldsymbol{\beta}}_K^{*}-\hat{\boldsymbol{\beta}}_K^{CQR})$로 근사될 수 있다는 점에서 정당화 될 수 있다 (Kocherginsky 등, 2005). 여기서 $\hat{\boldsymbol{\beta}}_K^{*}$는 분위수 개수 K를 사용

한 복합 분위수 회귀의 추정량 $\hat{\boldsymbol{\beta}}_K^{CQR}$에 대한 붓스트랩 표본으로부터 구한 회귀계수 벡터이며, $(\hat{\boldsymbol{\beta}}_K^* - \hat{\boldsymbol{\beta}}_K^{CQR})$의 근사분포는 크기가 n인 원자료 $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$로 부터 독립적으로 복원임의 추출된 충분히 큰 B개의 붓스트랩 표본 $\{\boldsymbol{x}_h^*, y_h^*\}_{h=1}^n$으로부터 구할 수 있다.

3.1. 붓스트랩을 활용한 분위수 개수 K의 선택

복합 분위수 회귀의 효율은 표 2.1에서 살펴본 바와 같이 사용되는 분위수 개수 K의 선택에 따라 달라지므로, 주어진 자료에 적합한 최적의 K값을 찾는 것은 복합 분위수 회귀에서 중요한 문제이다. 본 논문에서는 붓스트랩 방법을 사용하여 모형오차(model error; ME)

$$\mathrm{ME} = E\left[\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \Sigma_{\boldsymbol{x}} \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right] \tag{3.1}$$

를 추정하고, 이를 기준으로 적절한 분위수의 개수 K를 선택하는 알고리즘을 다음과 같이 제안한다. 여기서 $\Sigma_{\boldsymbol{x}}$는 설명변수 $\boldsymbol{x} \in R^p$의 공분산 행렬을 나타낸다.

단계 1. 주어진 자료 $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$으로부터 b번째 붓스트랩 표본 $\{\boldsymbol{x}_h^*, y_h^*\}_{h=1}^n$을 복원임의 추출한다.

단계 2. 단계 1에서 추출한 b번째 붓스트랩 표본 $\{\boldsymbol{x}_h^*, y_h^*\}_{h=1}^n$으로부터 분위수 개수 K에 대한 복합 분위수 회귀의 추정량 $\hat{\boldsymbol{\beta}}_K^{CQR}$의 붓스트랩 추정치 $\hat{\boldsymbol{\beta}}_K^*(b)\,(K = 1, 2, \ldots, q)$를 계산한다.

단계 3. 단계 2에서 계산된 붓스트랩 추정치 $\hat{\boldsymbol{\beta}}_K^*(b)$를 이용하여 분위수 개수 K에 대한 b번째 붓스트랩 모형오차

$$\mathrm{ME}_K^*(b) = \left(\hat{\boldsymbol{\beta}}_K^*(b) - \hat{\boldsymbol{\beta}}_K^{CQR}\right)' \hat{\Sigma}_X \left(\hat{\boldsymbol{\beta}}_K^*(b) - \hat{\boldsymbol{\beta}}_K^{CQR}\right), \quad (K = 1, 2, \ldots, q)$$

를 계산한다.

단계 4. 단계 1∼단계 3을 독립적으로 $B\,(b = 1, \ldots, B)$번 반복하여 분위수 개수 K에 대한 모형오차를 $\mathrm{ME}_K^*(b)\,(b = 1, \ldots, B)$의 평균

$$\mathrm{ME}_K^* = \frac{1}{B}\sum_{b=1}^{B} \mathrm{ME}_K^*(b), \quad (K = 1, 2, \ldots, q)$$

으로 추정한다.

단계 5. $K = 1, 2, \ldots, q$에 대해 계산된 모형오차의 추정치 ME_K^*을 비교하여 가장 작은 값을 가지는 K를 복합 분위수 회귀에 적절한 분위수의 개수로 선택한다.

앞에서 제시한 분위수 개수 K를 선택하는 붓스트랩 방법의 성능을 살펴보기 위해 식 (1.1)의 선형 회귀모형 $y = \boldsymbol{x}'\boldsymbol{\beta} + \epsilon$을 이용한 모의실험을 시행하였다. 여기서 $\boldsymbol{\beta} = (3, 1.5, 2)'$이고, 설명변수 $\boldsymbol{x} = (x_1, x_2, x_3)'$는 다변량 정규분포 $N(0, \Sigma_{\boldsymbol{x}})$을 따르며, 설명변수의 공분산 행렬은 $(\Sigma_{\boldsymbol{x}})_{i,j} = \rho^{|i-j|}$, $1 \le i, j \le 3$에서 $\rho = 0.5$를 이용하였다. 표본의 크기 $n = 100$과 붓스트랩 표본추출의 횟수 $B = 100$을 사용하였으며, 다양한 형태의 분포에서의 성능을 확인하기 위해 오차항의 분포는 2장에서 사용된 $N(0, 3)$ 분포, $\chi^2\,(df = 3)$ 분포, $t\,(df = 3)$ 분포, 그리고 $0.9N(0, 1) + 0.1N(0, 25)$ 분포를 사용하였다. 알고리즘의 평균적인 성능을 확인하기 위하여 100회의 독립반복 시행을 하였으며, 이를 통해 구한 모형오차 ME_K^*의 평균은 그림 3.1에 나타나 있다.

그림 3.1에서 점선으로 표시된 붓스트랩을 통한 모형오차의 추정치가 실선으로 표시된 실제 모형오차를 비교적 잘 추정하고 있음을 볼 수 있다. 실험에 사용된 모든 오차항의 분포에서 K가 커짐에 따라 실제

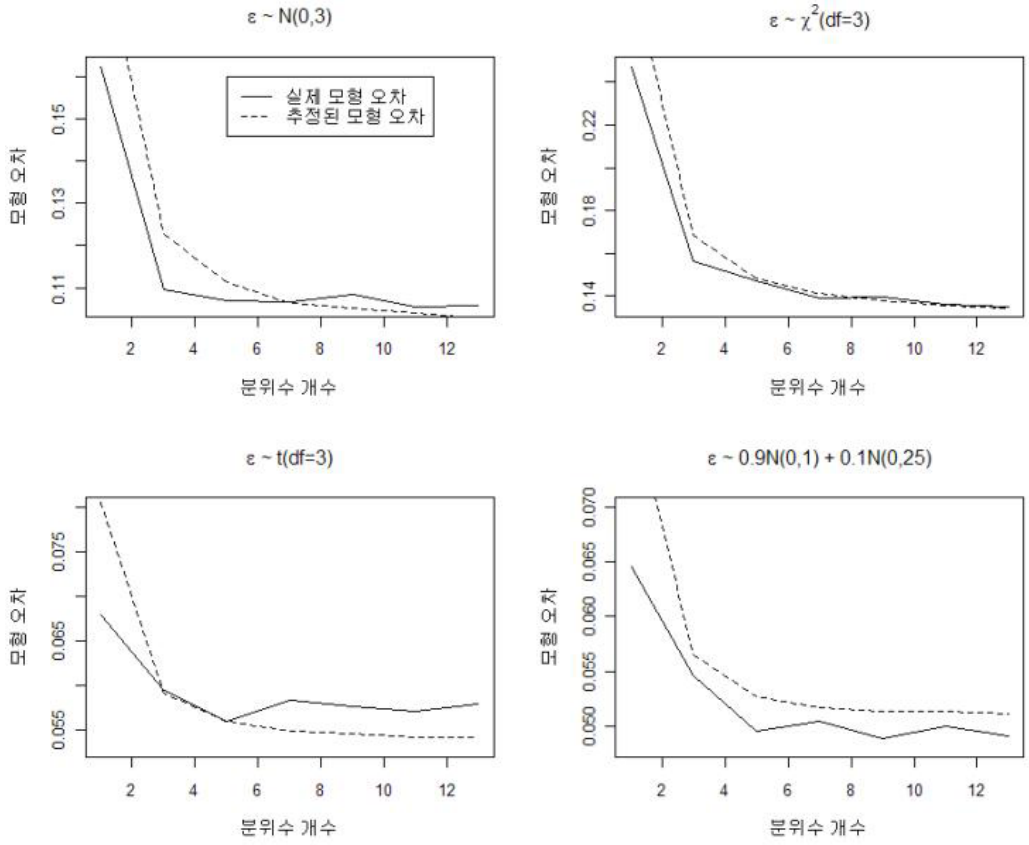

그림 3.1. K의 변화에 따른 모형오차의 변화

모형오차가 처음에는 급격히 그리고 차츰 완만하게 감소하는 경향을 확인할 수 있으며, 붓스트랩 추정치 또한 이 형태를 잘 추정하고 있는 것을 알 수 있다. 이런 모형오차의 변화는 표 2.1의 결과와도 일치한다. 따라서 오차항의 모분포를 모르는 실제상황에서 복합 분위수 회귀 방법으로 회귀계수를 추정하고자 할 때, 임의로 분위수 개수 K를 정하기에 앞서 붓스트랩 방법을 통해 모형오차를 추정하고 이에 근거하여 모형오차를 가장 작게 하거나 모형오차의 크기의 변화량이 작아지는 지점의 K를 복합 분위수 회귀에 적절한 분위수 개수로 선택할 수 있을 것이다.

3.2. 붓스트랩을 활용한 신뢰구간의 구축

분위수 회귀에서 신뢰구간을 구축하는 방법으로는 추정량의 점근적 정규분포의 공분산 행렬을 커널(kernel)을 이용하여 추정하는 방법 (Chen과 Wei, 2005)과 순위검정(rank test)을 역으로 이용하는 순위점수 방법 (Koenker, 1994)을 들 수 있다. 그러나 이러한 방법들은 계산이 매우 복잡하며, 회귀계수의 추정량에 대한 공분산 행렬을 효율적으로 추정하지 못한다는 단점이 지적되었다 (Kocherginsky 등, 2005). 본 논문에서는 붓스트랩을 통해 공분산 행렬을 추정한 후 점근적 정규성을 이용하여 복합 분위수 회귀의 추정량에 대한 신뢰구간을 구축하는 방법을 다음과 같이 제안하고자 한다. 여기서 최적의 분위수 개수 K는 3.1장의 알고리즘을 통해 선택되었다고 가정한다.

표 **3.1.** $N(0,3)$분포하에서 QR과 CQR 신뢰구간의 포함확률과 평균길이

p	β	$\rho=0$				$\rho=0.5$			
		$n=50$		$n=100$		$n=50$		$n=100$	
		QR	CQR	QR	CQR	QR	CQR	QR	CQR
3	β_1	0.964	0.938	0.968	0.952	0.968	0.950	0.974	0.958
		(1.420)	(1.096)	(0.966)	(0.747)	(1.737)	(1.337)	(1.177)	(0.915)
	β_2	0.942	0.936	0.956	0.948	0.940	0.930	0.958	0.958
		(1.422)	(1.096)	(0.953)	(0.747)	(1.744)	(1.343)	(1.167)	(0.918)
	β_3	0.958	0.918	0.962	0.960	0.954	0.924	0.964	0.956
		(1.445)	(1.099)	(0.970)	(0.753)	(1.770)	(1.343)	(1.190)	(0.926)
5	β_1	0.966	0.952	0.964	0.942	0.964	0.948	0.960	0.954
		(1.524)	(1.172)	(1.003)	(0.770)	(1.964)	(1.512)	(1.295)	(0.996)
	β_2	0.978	0.952	0.972	0.964	0.974	0.948	0.976	0.960
		(1.522)	(1.179)	(0.996)	(0.770)	(1.972)	(1.526)	(1.281)	(0.993)
	β_3	0.972	0.950	0.946	0.926	0.974	0.954	0.940	0.936
		(1.503)	(1.153)	(0.995)	(0.769)	(1.952)	(1.497)	(1.282)	(0.991)
	β_4	0.968	0.938	0.954	0.936	0.968	0.944	0.956	0.944
		(1.515)	(1.164)	(1.001)	(0.775)	(1.952)	(1.500)	(1.290)	(1.002)
	β_5	0.962	0.950	0.972	0.958	0.960	0.940	0.966	0.956
		(1.515)	(1.161)	(0.998)	(0.774)	(1.950)	(1.495)	(1.287)	(1.001)

괄호 안의 숫자는 신뢰구간의 평균길이를 나타냄

단계 1. 주어진 자료 $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$으로부터 b번째 붓스트랩 표본 $\{\boldsymbol{x}_h^*, y_h^*\}_{h=1}^n$을 복원임의 추출한다.

단계 2. 단계 1에서 추출한 b번째 붓스트랩 표본 $\{\boldsymbol{x}_h^*, y_h^*\}_{h=1}^n$으로부터 복합 분위수 회귀의 추정량 $\hat{\beta}_{K,j}^{CQR}$의 붓스트랩 추정치 $\hat{\beta}_{K,j}^*(b)\,(j=1,2,\ldots,p)$를 계산한다.

단계 3. 단계 1∼단계 2를 독립적으로 $B\,(b=1,\ldots,B)$번 반복하여 $\hat{\beta}_{K,j}^{CQR}$의 표준오차를

$$\mathrm{se}\left(\hat{\beta}_{K,j}^{CQR}\right)=\sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\beta}_{K,j}^*(b)-\hat{\beta}_{K,j}^*(\cdot)\right)^2},\quad (j=1,2,\ldots,p)$$

으로 추정한다. 여기서 $\hat{\beta}_{K,j}^*(\cdot)=(1/B)\sum_{b=1}^{B}\hat{\beta}_{K,j}^*(b)$이다.

단계 4. 단계 1∼단계 3의 과정을 통해 계산한 $\mathrm{se}(\hat{\beta}_{K,j}^{CQR})$을 사용하여 붓스트랩 신뢰구간

$$I_\alpha(X:\beta_j)=\left[\hat{\beta}_{K,j}^{CQR}-z_{\frac{\alpha}{2}}\mathrm{se}\left(\hat{\beta}_{K,j}^{CQR}\right),\,\hat{\beta}_{K,j}^{CQR}+z_{\frac{\alpha}{2}}\mathrm{se}\left(\hat{\beta}_{K,j}^{CQR}\right)\right],\quad (j=1,2,\ldots,p)$$

을 구한다. 여기서 $z_{\alpha/2}$는 표준정규분포의 $(100\times\alpha/2)\%$ 분위수이다.

앞에서 제시한 붓스트랩 방법을 이용하여 신뢰구간을 구축하고 분위수 회귀(QR)와 복합 분위수 회귀(CQR)의 성능을 비교하기 위하여, 3.1절에서와 동일한 모형 및 실험과정을 이용하여 모의실험을 시행하였다. 여기서는 $\boldsymbol{\beta}=(3,1.5,2)'$과 $\boldsymbol{\beta}=(3,1.5,2,0.5,4)'$을 고려하였으며, p차원 설명변수 $\boldsymbol{x}\sim N(\mathbf{0},\Sigma_{\boldsymbol{x}})$의 공분산 행렬 $(\Sigma_{\boldsymbol{x}})_{i,j}=\rho^{|i-j|}$, $1\le i,j\le p$에서 $\rho=0,0.5$를 이용하였다. 표본크기는 $n=50,100$을 사용하였으며, 복합 분위수 회귀의 경우 분위수 개수는 $K=9$로 고정하였고 분위수 회귀에서는 $\tau=0.5$을 이용하여 중위수 함수를 추정하였다. 실험에 사용된 명목포함확률은 0.95이며, 추정량의 표준오차를 구하기 위해 사용한 붓스트랩 표본추출 횟수는 $B=50$이다. 다양한 분포 하에서의 성능을 확인하기 위하여 오차항의 분포는 앞의 모의실험과 마찬가지로 $N(0,3)$ 분포, $\chi^2\,(df=3)$ 분포,

표 **3.2.** $\chi^2(3)$ 분포하에서 QR과 CQR 신뢰구간의 포함확률과 평균길이

p	β	$\rho=0$				$\rho=0.5$			
		$n=50$		$n=100$		$n=50$		$n=100$	
		QR	CQR	QR	CQR	QR	CQR	QR	CQR
3	β_1	0.954	0.944	0.966	0.970	0.950	0.946	0.974	0.972
		(1.725)	(1.294)	(1.154)	(0.829)	(2.116)	(1.585)	(1.411)	(1.017)
	β_2	0.954	0.952	0.954	0.944	0.952	0.948	0.950	0.958
		(1.745)	(1.293)	(1.160)	(0.845)	(2.148)	(1.592)	(1.427)	(1.034)
	β_3	0.970	0.962	0.962	0.968	0.970	0.962	0.964	0.966
		(1.732)	(1.303)	(1.167)	(0.844)	(2.119)	(1.595)	(1.429)	(1.032)
5	β_1	0.966	0.968	0.952	0.960	0.966	0.970	0.956	0.956
		(1.843)	(1.420)	(1.182)	(0.879)	(2.373)	(1.829)	(1.524)	(1.132)
	β_2	0.970	0.968	0.962	0.964	0.966	0.962	0.956	0.970
		(1.850)	(1.421)	(1.163)	(0.868)	(2.388)	(1.835)	(1.502)	(1.122)
	β_3	0.972	0.952	0.968	0.964	0.960	0.954	0.960	0.968
		(1.809)	(1.392)	(1.176)	(0.870)	(2.340)	(1.798)	(1.515)	(1.121)
	β_4	0.986	0.980	0.970	0.964	0.982	0.980	0.962	0.972
		(1.861)	(1.428)	(1.185)	(0.881)	(2.400)	(1.845)	(1.525)	(1.135)
	β_5	0.968	0.972	0.956	0.956	0.972	0.962	0.966	0.964
		(1.838)	(1.419)	(1.170)	(0.871)	(2.373)	(1.827)	(1.501)	(1.117)

괄호 안의 숫자는 신뢰구간의 평균길이를 나타냄

표 **3.3.** $t(3)$ 분포하에서 QR과 CQR 신뢰구간의 포함확률과 평균길이

p	β	$\rho=0$				$\rho=0.5$			
		$n=50$		$n=100$		$n=50$		$n=100$	
		QR	CQR	QR	CQR	QR	CQR	QR	CQR
3	β_1	0.958	0.948	0.954	0.952	0.956	0.958	0.954	0.944
		(0.993)	(0.833)	(0.627)	(0.542)	(1.216)	(1.020)	(0.765)	(0.664)
	β_2	0.968	0.968	0.954	0.956	0.966	0.960	0.960	0.948
		(0.994)	(0.835)	(0.624)	(0.542)	(1.213)	(1.021)	(0.763)	(0.665)
	β_3	0.972	0.966	0.966	0.950	0.962	0.966	0.962	0.948
		(0.985)	(0.829)	(0.623)	(0.543)	(1.211)	(1.020)	(0.761)	(0.666)
5	β_1	0.986	0.974	0.974	0.960	0.980	0.968	0.968	0.966
		(1.085)	(0.898)	(0.675)	(0.569)	(1.397)	(1.159)	(0.871)	(0.737)
	β_2	0.974	0.966	0.968	0.962	0.974	0.960	0.972	0.960
		(1.082)	(0.903)	(0.675)	(0.565)	(1.398)	(1.167)	(0.868)	(0.729)
	β_3	0.968	0.964	0.960	0.964	0.968	0.960	0.960	0.962
		(1.064)	(0.884)	(0.669)	(0.564)	(1.375)	(1.143)	(1.143)	(0.729)
	β_4	0.970	0.950	0.970	0.962	0.964	0.948	0.974	0.960
		(1.094)	(0.908)	(0.672)	(0.561)	(1.404)	(1.165)	(0.870)	(0.724)
	β_5	0.986	0.978	0.974	0.966	0.978	0.970	0.968	0.972
		(1.052)	(0.876)	(0.667)	(0.562)	(1.359)	(1.133)	(0.858)	(0.725)

괄호 안의 숫자는 신뢰구간의 평균길이를 나타냄

$t\,(df\,=3)$ 분포, 그리고 $0.9N(0,1)+0.1N(0,25)$ 분포를 사용하였다. 각각의 오차항에 대한 500회의 독립반복시행을 통하여 포함확률을 추정하고 신뢰구간의 평균길이를 구하였으며, 이와 같은 실험의 결과는 표 3.1∼표 3.4에 요약되었다.

표 **3.4.** $0.9N(0,1)+0.1N(0,25)$ 분포하에서 QR과 CQR 신뢰구간의 포함확률과 평균길이

p	β	$\rho=0$				$\rho=0.5$			
		$n=50$		$n=100$		$n=50$		$n=100$	
		QR	CQR	QR	CQR	QR	CQR	QR	CQR
3	β_1	0.966	0.960	0.956	0.938	0.960	0.952	0.952	0.938
		(0.950)	(0.785)	(0.616)	(0.512)	(1.163)	(0.960)	(0.754)	(0.627)
	β_2	0.958	0.962	0.942	0.954	0.974	0.952	0.954	0.960
		(0.943)	(0.786)	(0.611)	(0.511)	(1.161)	(0.966)	(0.750)	(0.627)
	β_3	0.958	0.956	0.946	0.940	0.964	0.960	0.940	0.936
		(0.945)	(0.795)	(0.602)	(0.506)	(1.168)	(0.980)	(0.739)	(0.618)
5	β_1	0.980	0.970	0.924	0.960	0.984	0.974	0.968	0.952
		(1.009)	(0.837)	(0.537)	(0.527)	(1.310)	(1.085)	(0.825)	(0.681)
	β_2	0.982	0.974	0.930	0.958	0.978	0.964	0.968	0.962
		(1.028)	(0.843)	(0.539)	(0.527)	(1.337)	(1.096)	(0.831)	(0.680)
	β_3	0.978	0.974	0.946	0.956	0.976	0.964	0.976	0.952
		(1.016)	(0.845)	(0.531)	(0.521)	(1.318)	(1.097)	(0.818)	(0.674)
	β_4	0.984	0.962	0.926	0.966	0.980	0.960	0.960	0.952
		(1.034)	(0.864)	(0.547)	(0.531)	(1.327)	(1.104)	(0.839)	(0.683)
	β_5	0.974	0.960	0.934	0.946	0.972	0.966	0.966	0.944
		(1.020)	(0.850)	(0.542)	(0.531)	(1.311)	(1.093)	(0.831)	(0.684)

괄호 안의 숫자는 신뢰구간의 평균길이를 나타냄

표 3.1∼표 3.4에서 표본의 크기가 $n=50$이고 설명변수의 차원이 $p=5$인 경우 복합 분위수 회귀에 비해 분위수 회귀의 추정된 포함확률이 명목 포함확률보다 큰 경향을 보인다. 그러나 표본크기가 $n=100$으로 커지면서 두 방법 모두 안정되지만 일반적으로 복합 분위수 회귀를 사용한 신뢰구간의 포함확률이 명목 포함확률 0.95에 더 가깝다는 것을 알 수 있다. 이와 같은 결과는 실험에 사용된 모든 분포에서 유사하게 발생한다. 나아가, 분위수 회귀와 복합 분위수 회귀의 신뢰구간에 대한 평균길이를 비교해 보면, 표본크기(n), 설명변수의 수(p), 설명변수의 상관관계(ρ)가 달라지는 모든 영역에서 복합 분위수 회귀의 신뢰구간 평균길이가 분위수 회귀의 신뢰구간 평균길이에 비해 짧게 구해지는 것을 확인할 수 있다. 이는 좁은 신뢰구간의 영역으로 유사한 정도의 포함확률을 가지게 되므로 추정의 성능이 더 우수하다고 말할 수 있고, 또한 이와 같은 추정량의 신뢰구간을 통해 더 정확하게 검정을 할 수 있음을 의미한다. 이러한 결과는 바로 복합 분위수 회귀의 추정량에 대한 분산이 더 작아서 추정의 효율이 높기 때문이며, 표 2.1에서 추정량의 점근적 분포를 통해 살펴본 효율의 비교와도 일치하는 결과이다. 이와 같이 신뢰구간의 비교를 통해 복합 분위수 회귀 방법이 분위수 회귀 방법에 비해 성능이 우수함을 확인할 수 있었다.

4. 결론

선형 회귀모형에서 오차항들이 서로 독립이고 동일한 분포를 따른다는 가정이 있는 경우, 다수의 분위수 함수를 동시에 고려하는 복합 분위수 회귀 방법을 통해 회귀계수를 효율적으로 추정할 수 있다. 본 논문에서는 복합 분위수 회귀에서 사용되는 분위수의 개수 K를 선택하는 문제에 대해 붓스트랩 방법을 제안하였다. 제안된 붓스트랩 방법을 통해 추정한 모형오차가 실제 모형오차를 잘 추정하는 것을 확인하였으며, 모형오차의 붓스트랩 추정치를 가장 작게 하는 분위수 개수 K를 복합 분위수 회귀에 적절한 분위수 개수로 선택할 수 있었다. 나아가, 분위수 회귀와 복합 분위수 회귀의 성능을 회귀계수의 붓스트랩

신뢰구간의 관점에서 비교해 보았다. 복합 분위수 회귀의 신뢰구간은 포함확률 측면에서는 분위수 회귀의 신뢰구간과 유사한 결과를 제공하지만 평균길이는 더 짧게 구해지는 것을 확인할 수 있었다. 유사한 포함확률 하에서 더 좁은 신뢰구간을 설정한다는 점에서 복합 분위수 회귀 방법이 더 효율적인 추정 결과를 도출한다고 말할 수 있다.

참고문헌

Bradic, J., Fan, J. and Wang, W. (2010). Penalized composite quasi-likelihood for ultrahigh dimensional variable selection, *Journal of the Royal Statistical Society*, **73**, 325–349.

Chen, C. and Wei, Y. (2005). Computational issues for quantile regression, *Sankhyā: The Indian Journal of Statistics*, **67**, 399–417.

Kocherginsky, M., He, X. and Mu, Y. (2005). Practical confidence intervals for regression quantiles, *Journal of Computational and Graphical Statistics*, **14**, 41–55.

Koenker, R. (1984). A note on L-estimates for linear models, *Statistics and Probability Letters*, **2**, 323–325.

Koenker, R. (1994). *Confidence Intervals for Regression Quantiles, Asymptotic Statistics*, Springer-Verlag, New York, 349–359.

Koenker, R. and Bassett, G. (1978). Regression quantiles, *Econometrica*, **46**, 33–50.

Zou, H. and Yuan, M. (2008). Composite quantile regression and the oracle model selection theory, *Annals of Statistics*, **36**, 1108–1126.

Bootstrapping Composite Quantile Regression

Kangmin Seo[1] · Sungwan Bang[2] · Myoungshic Jhun[3]

[1]Department of Statistics, Korea University
[2]Department of Mathematics, Korea Military Academy
[3]Department of Statistics, Korea University

(Received February 29, 2012; Revised April 6, 2012; Accepted April 9, 2012)

Abstract

Composite quantile regression model is considered for iid error case. Since the regression coefficients are the same across different quantiles, composite quantile regression can be used to combine the strength across multiple quantile regression models. For the composite quantile regression, bootstrap method is examined for statistical inference including the selection of the number of quantiles and confidence intervals for the regression coefficients. Feasibility of the bootstrap method is demonstrated through a simulation study.

Keywords: Quantile regression, composite quantile regression, bootstrap.

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2010-0009204).
[3]Corresponding author: Professor, Department of Statistics, Korea University, Seoul 136-701, Korea.
E-mail: jhun@korea.ac.kr