The Weight Function in BIRQ Estimator for the AR(1) Model with Additive Outliers

Byoung Cheol Jung[1] and Sang Moon Han[2]

요 약

In this study, we investigate the effects of the weight function in the bounded influence regression quantile (BIRQ) estimator for the AR(1) model with additive outliers. In order to down-weight the outliers of X-axis, the Mallows' (1973) weight function has been commonly used in the BIRQ estimator. However, in our Monte Carlo study, the BIRQ estimator using the Tukey's bisquare weight function shows less MSE and bias than that of using the Mallows' weight function or Huber's weight function.

Keywords : 로버스트추정, Regression quantile estimator, Weight Function, AR(1).

1. 서론

Fox (1972)가 1차 자기상관을 갖는 시계열 모형 (AR(1))에서 이상치의 종류를 IO 이상치 (innovation outlier)와 AO 이상치 (additive outlier, AO)로 구분하여 연구한 이래로 Denby와 Martin (1979), de Jongh 과 de Wet (1985) 및 한상문과 정병철 (2004) 등 많은 연구자들이 이상치가 존재하는 AR(1) 모형에서 로버스트 추정방법에 대하여 연구하고 있다. 특히 L-추정량에 관한 연구는 de Jongh과 de Wet (1985) 및 한상문과 정병철 (2004)을 들 수 있다. de Jongh과 de Wet (1985)은 Koenker와 Bassett (1978)이 제안한 회귀 분위수(Regression Quantile, RQ) 추정방법을 이상치가 존재하는 AR(1) 모형에 제안하였다. 더불어 그들은 X축의 이상치에 대한 비중강하(down-weight)의 방법으로 Mallows (1973) 가중치 함수를 이용한 유계영향(bounded-influence) RQ (BIRQ) 추정방법을 제안하였다.

본 논문에서는 가산적 이상치 (AO)가 존재하는 AR(1) 모형에서 de Jongh과 de Wet (1985)이 제안한 BIRQ 추정량에서 비중강하에 사용되는 가중치의 효과에 대하여 연구하고자 한다. X축의 이상치에 대한 비중강하의 방법으로는 통상적으로 Mallows (1973)가 제안한 가중치 함수가 사용된다. 하지만 이와 같은 가중치 함수가 추정량에 미치는 연구는 거의 되어 있지 않다. 본 연구에서는 Mallows 가중치 함수 이외에 전통적인 M-추정법에서 주로 사용되는 Huber 가중치 함수, Tukey 가중치 함수 및 Rousseeuw와 Leroy (1987)에 의하여 회귀모형에서 GM-추정량을 구할 때 사용된 가중치 함수 등 여러 가중치 함수들을 이용하여 BIRQ 추정량을 유도하고 각 가중치에 따른 추정량의 효율성에 대하여 연구하고자 한다.

1) Contract Professor, Department of Statistics, Sungshin University, Seoul 136-742, Korea.

2) Professor, Department of Statistics, University of Seoul, Seoul 136-701, Korea.

2. BIRQ 추정량

다음과 같은 가산적 이상치 (additive outlier, AO)를 갖는 1차 자기상관을 갖는 모형을 고려해보자.

$$X_i = \mu + \rho X_{i-1} + \epsilon_i \tag{1}$$

$$Y_t = X_t + \nu_t \tag{2}$$

여기서 μ와 ρ은 각각 상수항과 1차 자기상관계수를 나타내는 모수이다. 또한 $|\rho| < 1$이며 ϵ_t는 $N(0, \sigma_\epsilon^2)$를 따르는 IID 오차항을 나타낸다. 더불어 ν_t는 다음과 같이 "0에서 퇴행"과 "평균이 0인 확률분포"와의 오염된 분포를 따른다고 가정한다.

$$CND(\cdot\,|\gamma, \sigma^2) = (1-\gamma)\delta(\cdot) + \gamma F \tag{3}$$

γ는 오염비율을 나타내고 $\delta(\cdot)$는 0에 퇴행(degenerate)하는 값, 즉 $\Pr(\nu_i = 0) = 1-\gamma$이며 $\Pr(\nu_i \neq 0) = \gamma$이다. 이때 F는 ν_t의 오염된 부분에 해당하는 확률분포로 평균이 0이고 대칭인 분포를 나타내며 정규분포, SLACU, SLASH, CAUCHY 등의 분포가 사용될 수 있다. Fox (1972)에 의하면, 식 (2)와 같은 시계열 모형은 정규분포를 따르는 관측치 X_i에 "가산적 효과" ν_i가 더해진 형태로 구성된 시계열 모형이다. 이와 같은 모형에서 de Jongh과 de Wet (1985)은 Koenker와 Bassett (1978)이 제안한 회귀 분위수(Regression Quantile, RQ) 추정방법 및 유계영향 회귀 분위수 (BIRQ) 추정방법을 제안하였다. 먼저 본 모형에서 모수에 대한 BIRQ 추정방법에 대하여 알아보자. $w_i\ (i = 1, \cdots, n-1)$를 i번째 개체에 대한 가중치라 했을 때, BIRQ 추정량은 다음과 같은 과정을 통하여 얻어진다.

Step 1. 먼저 α^{th} 일반화 RQ 추정량 $\widehat{\mu^w}(\alpha), \widehat{\rho^w}(\alpha)$와 $(1-\alpha)^{th}$ 일반화 RQ 추정량 $\widehat{\mu^w}(1-\alpha), \widehat{\rho^w}(1-\alpha)$를 각각 유도한다. 이때 θ^{th} 일반화 RQ 추정량은 다음과 같은 식을 최소화하는 추정량으로 정의된다.

$$\min_{\mu,\rho}\left[\theta\sum_1 w_i|Y_i - \mu - \rho Y_{i-1}| + (1-\theta)\sum_2 w_i|Y_i - \mu - \rho Y_{i-1}|\right] \tag{4}$$

여기서 $\sum_1 w_i|Y_i - \mu - \rho Y_{i-1}|$은 $w_i(Y_i - \mu - \rho Y_{i-1}) > 0$인 모든 관측값들을 더하는 기호를 각각 나타낸다.

Step 2. 식 (4)에서 얻어진 RQ 추정량들을 이용하여 다음과 같은 가중치를 다시 구한다.

$$d_i^{(w)} = \begin{cases} w_i & \text{if } \widehat{\mu^{(w)}}(\alpha) + \widehat{\rho^{(w)}}(\alpha)\,Y_{i-1} \leq Y_i \leq \widehat{\mu^{(w)}}(1-\alpha) + \widehat{\rho^{(w)}}(1-\alpha)\,Y_{i-1}, \end{cases} \tag{5}$$

Step 3. μ와 ρ에 대한 BIRQ 추정량을 각각 $\widehat{\mu^{(w)}}(BIRQ, \alpha)$와 $\widehat{\rho^{(w)}}(BIRQ, \alpha)$이라 놓는다면 이는 다음과 같은 식을 최소화하는 가중최소제곱 추정량으로 구해진다.

$$\min_{\mu,\rho}\sum_{i=2}^{n} d_i^{(w)}(Y_i - \mu - \rho Y_{i-1})^2 \tag{6}$$

3. X축 방향 이상치에 대한 가중치 함수

2장에 나타난 BIRQ 추정량을 구하기 위해서는 X축 방향의 이상치에 대한 비중강하의 목적으로 사용되는 가중치 w_i가 꼭 필요하다. 본 절에서는 여러 가지 방법에 의한 가중치 함수를 정의하고자 한다.

3.1 Mallows(1973) 가중치 함수

Mallows (1973)는 X축 방향에 대한 가중치를 부여하는 방법으로 다음과 같은 과정을 이용하였다. 먼저 $Y_{(1)} \leq Y_{(2)} \leq \cdots Y_{(n-1)}$을 $Y_i, i=1,\cdots,n-1$ 들의 순서화된 값이라 하고 $z_1,\cdots,z_{n-1}$을 Y_i의 순위라 하고, $L=[\tau n]+1$ 및 $U=n-L$를 각각 정의하자. 이때 $[x]$는 x들 중에서 가장 큰 정수값을 나타내는 기호이며 τ는 알려진 상수이다. 본 연구에서는 $\tau=0.1$로 고정하였다. Mallows는 $(n-1)$개의 관측치 $Y_1,\cdots,Y_{n-1}$에 대하여 다음과 같은 가중치를 고려하였다.

$$w_i^M = \begin{cases} 1, & \text{if } L \leq z_i \leq U \\ \dfrac{Y_{(L)} - Y_{(U)}}{2Y_{i-1} - Y_{(U)} - Y_{(L)}}, & \text{if } z_i < L \\ \dfrac{Y_{(U)} - Y_{(L)}}{2Y_{i-1} - Y_{(U)} - Y_{(L)}}, & \text{if } z_i > U \end{cases} \tag{8}$$

식 (8)의 가중치를 이용하여 식 (7)에서 얻어지는 절사율 α인 BIRQ 추정량을 BIRQ(M, α)로 정의하자.

3.2 Huber와 Tukey의 가중치 함수

Holland와 Welsh (1977) 및 Hogg 등 (1988)은 회귀모형에서 회귀계수에 대한 M-추정량의 형태를 가중최소제곱 방정식의 해로 유도하였다. 이때 M-추정량을 얻기 위한 가중치는 잔차를 이용한 가중치로 주어지는데, 이와 같은 가중치를 본 모형에 응용하여 X-축의 이상치에 대한 비중강하의 목적으로 사용하고자 한다.

먼저 다음과 같은 Huber 형태의 가중치 함수를 정의해보자.

$$w_i^H = \begin{cases} 1, & \text{if } |Y_i| \leq ks \\ ks/|Y_i|, & otherwise \end{cases} \tag{9}$$

여기서 k는 조절상수를 나타내며 s는 $Y_i(i=1,\cdots,n-1)$들의 로버스트한 표준편차의 추정량을 나타낸다. 식 (9)에 나타난 가중치는 Denby와 Martin (1979)이 AR(1)모형에서 GM 추정량을 구할 때 사용한 방법과 유사하다. 본 연구에서는 조절상수 $k=1.0$을 사용하고 $s=MAD/0.6745$를 사용하였다. 여기서 $MAD=median_i\{Y_i - med(Y_i)\}$로 정의된다. 식 (9)의 가중치를 이용하여 식 (7)에서 얻어지는 절사율 α인 BIRQ 추정량을 BIRQ(H, α)로 정의하자.

다음으로 다음과 같은 Tukey 형태의 가중치 함수를 정의해보자.

$$w_i^T = \begin{cases} [1-(Y_i/ks)^2]^2, & \text{if } |Y_i| \leq ks \\ 0, & otherwise \end{cases} \tag{10}$$

식 (10)에 나타난 가중치도 Denby와 Martin (1979)이 AR(1)모형에서 GM 추정량을 구할 때 사용한 방법과 유사하다. 본 연구에서는 Denby와 Martin (1979)의 연구에서처럼 조절상수 $k = 3.9$를 사용하고 $s = MAD/0.6745$를 사용하였다. 식 (10)의 가중치를 이용하여 식 (7)에서 얻어지는 절사율 α인 BIRQ 추정량을 BIRQ(T, α)로 정의하자.

3.3 Rousseeuw의 가중치 함수

회귀모형에서 회귀계수에 대한 GM 추정량을 유도하기 위하여 Rousseeuw와 Leory (1987)는 최소부피 타원체 추정량의 개념을 이용하여 가중함수를 정의하였다. 이와 같은 가중함수를 본 모형에 적용하면 다음과 같은 가중함수를 얻을 수 있다.

$$w_i^L = \min\left[1, \{\frac{b}{(Y_i - m_Y)^2/c_Y^2}\}^{1/2}\right], \; i = 1, \cdots, n-1 \tag{11}$$

여기서 b는 $\chi^2(1, 0.95)$를 흔히 사용하고, m_Y는 $Y_i (i = 1, \cdots, n-1)$들의 중위수를 나타내며 $s = MAD/0.6745$이다. 이와 같은 가중함수 w_i^L은 Y_i와 m_Y의 마할라노비스 거리를 로버스트한 형태로 이해할 수 있다. 따라서 X축에 이상치가 존재하는 경우에는 대응되는 가중치가 작게 부여되게 된다.

4. 모의실험

본 논문에서 제안한 각 가중치들에 의하여 얻어지는 BIRQ 추정량들의 효율성을 알아보기위하여 모의실험을 실시하였다. 모의실험에서 표본수 n은 100으로 고정하였고, 상수항 μ는 0으로 고정하였으며 1차 자기상관계수 ρ은 0.5와 0.8을 각각 고려하였다. 또한 식 (3)의 가산적 오차항 부분에 나타나는 오염률 γ의 값은 0에서 0.2까지 0.05단위씩 증가시켰으며, 가산적 오차항 ν_i의 확률분포 F로는 정규분포를 사용하였다. 이때 ν_i의 분산을 9, 36, 100 및 ∞(SLASH 분포)으로 변화시켜가며 실험하였다. 더불어 BIRQ 추정량을 구하기 위하여 필요한 절사율 α의 값은 0.05, 0.1 및 0.15등을 사용하였으며 X축 방향의 이상치에 대한 비중강하의 목적으로 3장에서 제안한 4가지 가중치 함수들이 사용되었다.

모든 ρ, γ 및 ν_i의 분산수준에 대하여 1,000번의 반복이 실시되었으며, 각 반복마다 μ와 ρ에 대한 BIRQ 추정량을 2장의 방법을 사용하여 계산하였다. 이와 같이 얻어진 1,000개의 추정량을 이용하여 각 추정량의 평균제곱오차(Mean Square Error, MSE) 및 편의(Bias)를 계산하였다.

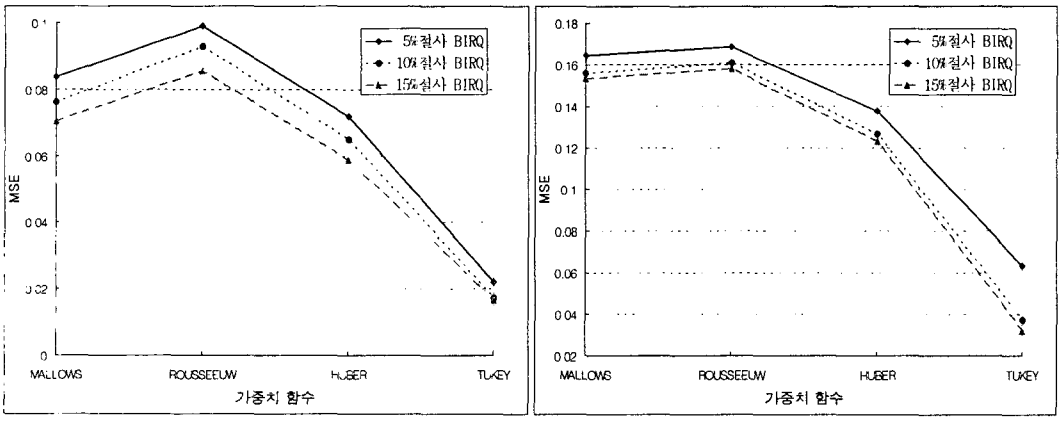
4.1 MSE 비교

다음 <그림 1>은 ν_i의 분산이 100이고 $\rho = 0.5$인 경우 오염률 (γ)이 0.1과 0.2인 경우 각 오염률에 따른 각 추정량들의 MSE를 나타낸 것이다. 이 경우 OLS 추정량은 MSE가 다른 추정량에 비하여 크게 나타나므로 그림에서 제시하지 않았다.

<그림 1>의 결과를 살펴보면, $\gamma \geqq 0.05$인 모든 경우에 Tukey의 bisquare 가중함수를 고려한 BIRQ 추정량의 MSE가 절사율 α와 관계없이 가장 작게 나타났다. 그동안 통상적으로 사용되던 Mallows 가중치 함수를 사용한 BIRQ 추정량은 자료의 오염률에 관계없이 Huber 형태의 가중치 함수를 사용한 BIRQ 추정량이나 Tukey의 가중치 함수를 이용한 BIRQ 추정량에 비하여 큰 MSE 수준을 보여주고 있다.

<그림 1> ν_i의 분산이 100이고 $\rho = 0.5$인 경우 각 오염률, 절사율 및 가중치 함수에 따른 각 추정량의 MSE

(a) 오염률(γ)이 10%인 경우 (b) 오염률(γ)이 20%인 경우

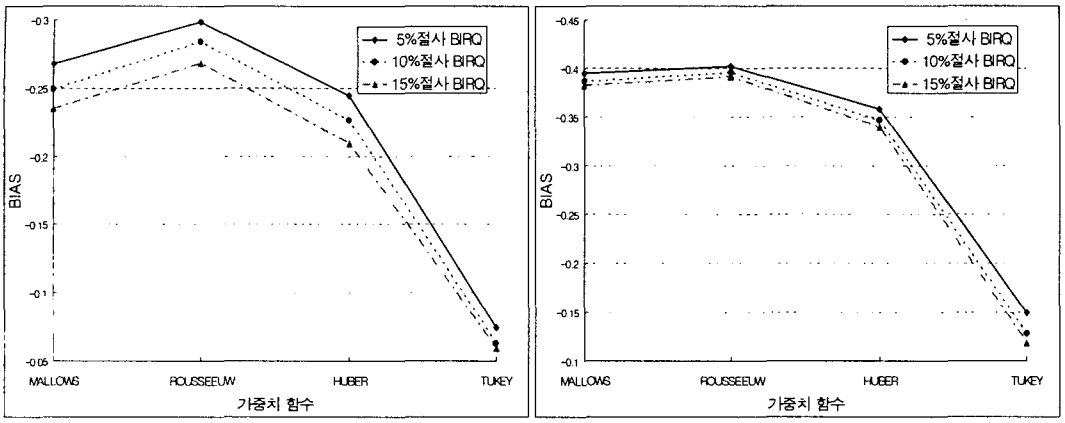


4.2 BIAS 비교

다음 <그림 2>은 ν_i의 분산이 100이고 $\rho = 0.5$인 경우 오염률 (γ)이 0.1과 0.2인 경우 각 오염률에 따른 각 추정량들의 편의를 나타낸 것이다. 먼저 각 추정량들은 모두 음(-)의 편의를 보여 참 모수 ρ를 과소추정하고 있음을 보여준다. <그림 2>에 따르면 자료의 오염률 γ값에 관계없이 Tukey의 bisquare 가중치 함수를 사용하는 BIRQ 추정량의 편의가 가장 작게 나타나고 있다. 절사율 관점에서는 대부분의 가중치 함수에서15% 절사를 사용한 BIRQ 추정량이 5%나 10% 절사를 이용한 BIRQ 추정량보다 효율적이었다. 특히 절사율 15%인 BIRQ(T, 0.15) 추정량의 편의가 가장 작게 나타나고 있다.

<그림 2> ν_i의 분산이 100이고 $\rho = 0.5$인 경우 각 오염률, 절사율 및 가중치 함수에 따른 각 추정량의 BIAS

(a) 오염률(γ)이 10%인 경우 (b) 오염률(γ)이 20%인 경우



5. 결론

본 연구에서는 가산적 이상치를 갖는 AR(1) 모형에서 자기상관계수에 대한 유계영향 회귀분위수 (BIRQ) 추정에 이용되는 가중치 함수의 효과에 대하여 모의실험을 통하여 연구하였다. 모의실험 결과, Tukey의 가중치 함수를 사용하는 BIRQ 추정량이 Mallows나 Huber의 가중치 함수를 사용하는 BIRQ 추정량에 비하여 작은 MSE와 편의를 보여 가장 효율적으로 나타났다.

참고문헌

De Jongh, P.J. and De Wet T. (1985). Trimmed Mean and Bounded Influence Estimators for the Parameters of the AR(1) Process, *Communications in Statistics -Theory and Methods*, Vol. 14, 1361-1375.

Denby, L. and Martin, R.D. (1979). Robust Estimation of the First-Order Autoregressive Parameter, *Journal of the American Statistical Association*, Vol. 74, 140-146.

Fox, A.J. (1972). Outliers in Time Series, *Journal of the Royal Statistical Society, Series B*, Vol. 34, 350-363.

Han, S.M and Jung, B.C. (2004). AR(1) 모형의 모수에 대한 L-추정법, 「응용통계연구」 Accepted.

Hogg, R.V., Bril, G.K., Han, S.M. and Yuh, L. (1988). An Argument for Adaptive Robust Estimation, Probability and Statistics, Essay in Honor of Graybill, F.A., North Holland, 135-148.

Holland, P.W. and Welsh, R.E. (1977). Robust Regression using Iteratively Reweighted Least Squares, *Communications in Statistics - Theory and Methods*, Vol. 6, 813-827.

Koenker, R. and Bassett, G. (1978). Regression Quantiles, *Econometrica*, Vol. 46, 33-50.

Mallows, C.L. (1973). Influence Functions, *Unpublished paper presented at a conference on robust regression held at Cambridge, Mass., and sponsored by the National Bureau of Economic Research.*

Rousseeu, P.J. and Leory, A.M. (1987). *Robust Regression and Outlier Detection*, New York, Wiley.