韓國數學教育學會誌<數學教育>
1984. 3. Vol. XXII, No. 2.

Journal of the Korea Society of Mathematical Education
Mar. 1984. Vol. XXII, No. 2.

線型 回歸 模型의 Adaptive L_p-norm 推定値

李 昌 信

1. 서 론

최근에 들어 robust한 통계적 추론에 대해 많은 연구가 진행되고 있다. 이는 통계적 모형에 대한 기본적 가정이 자료의 분석결과에 많은 영향을 미치기 때문이다.

최소 자승 추정법을 사용하는 사람들은 측량 오차가 정규분포를 한다고 믿고 있으나 심히 벗어난 관측치는 최소자승 추정치를 쓸모 없게 만든다. 이러한 영향은 회귀 문제에서 더욱 심각하다.

분포 자유(derstribution-free) 검정법의 출현으로 모집단이 정규 분포를 따를 때에도 표본 평균보다 과히 나쁘지 않고, 두터운 꼬리를 갖는 분포를 따를 때에는 훨씬 높은 효율을 갖는 추정치를 유도할 수 있다.

위치 모수의 robust추정치로서 가장 많이 사용되어 온 것은 α-trimmed 평균과 Winsorized 평균이다.

Huber [6]는 추정치의 최대 분산을 최소화 시키는 M-추정법을 발표하여 robust 추정법 연구가 본격화 되었다.

Robust 추정법이란 기초 분포의 작은 변화에도 민감한 고전적인 방법에 비해 높은 효율을 갖고, 정규 분포를 할 때에도 효율이 거의 떨어지지 않는 추정법이다.

Adaptive 방법은 주로 Hogg [5]에 의하여 소개되었는데 이 방법에서는 자료에서 얻어진 정보를 통해서 선택된 모형에 기초를 두어 통계적 추론을 하게 된다.

이 논문에서는 2장에서 위치 모수와 회귀 모수의 robust 추정치의 고찰을, 3장에서는 adaptive procedure에 대한 고찰과 Hogg가 제안한 분포 꼬리의 길이에 대한 측도를 소개하고, 이를 사용하여 adaptive L_p-추정치를 Monte Carlo 연구로 분석해 보았다.

다중 회귀 분석에서 adaptive L_p-추정치는 잔차 분포가 긴 꼬리를 가질 때는 p값이 작을수록 robust하다. 여러 잔차분포에서 최소 자승 추정치에 비해 그 효율이 대체로 높고, 정규 분포일 때도 과히 떨어지지 않았다.

2. Robust 추정치에 대한 고찰

다음과 같은 선형 회귀 모형을 생각해 보자.

$$Y_i=\sum_{j=1}^{k}X_{ij}\beta_j+\epsilon_i,\ \ i=1,\ \ldots,\ n.$$

여기서 X_{ij}는 기지의 상수이며, β_j들은 추정해야 할 회귀 모수들이고, ϵ_i는 서로 독립이고 같은 분포를 갖는 임의 오차이다. 최소 자승 추정치나 임의 오차가 정규 분포를 따를 때의 최우 추정치는

$$\sum_{i=1}^{n}(Y_i-\sum_{j=1}^{k}X_{ij}\beta_j)^2$$

을 최소로 하는 값으로 결정된다.

그러나 최소 자승 추정법은 정규 분포란 가정에 예민하고 최우 추정치는 경우마다 달라지므로 모집단의 분포에 둔감하면서도 효율이 떨어지지 않는 추정법을 연구하게 된다.

2.1. 위치 모수의 추정치

Huber [6]의 M-추정치는 최소 자승 추정법의 자승 함수를 변형시켜서 위치 모수의 추정치 T를

$$\sum_{i=1}^{n}\rho(X_i-T)$$

를 최소화하는 값으로 정의했다. 여기서 ρ는 비상수 함수이며 일반적으로 척도 불변인 추정치

는 방정식

$$\sum_{i=1}^{n}\Psi[(X_i-T)/S]=0$$

의 해가 된다. 여기서 s는 산포의 robust 추정치이며, Ψ는 ρ의 도함수이다.

Huber, Hampel, Andrew는 Ψ함수를 제안, 수정하였다.

또 다른 robust 추정치로서 많이 사용되어지는 L-추정치는 α-trimmed 평균과 Winsorized 평균이 있다.

α-trimmed 평균은 다음과 같이 정의된다.

$$m(\alpha)=\left(\frac{1}{h}\right)\sum_{i=g+1}^{n-g}X_{(i)},$$

여기서 $g=n\alpha$, $h=n-2g=n-2n\alpha$가 된다. 즉 최고 최저 $g=n\alpha$개의 관측치를 버린 후의 평균이 된다. 예상되는 분포의 꼬리가 정규 분포보다 길고 Cauchy 분포보다 짧으면 n에 따라 α값으로 $\frac{1}{4}$ 또는 $\frac{1}{3}$을, 이중 지수 분포 정도라면 $\alpha=\frac{1}{8}$정도를, 또 정규분포와 이중 지수 분포 사이라면 α값으로 약 $\frac{1}{5}$을 택한다. 어느 경우에서나 매우 robust하며 쉽게 통계적 결론을 얻을 수 있다.

2.2. 회귀 모수의 추정치

2.2.1. M-추정치

robust 추정치로서 가장 널리 사용되고 있는 회귀 모수의 M-추정치는 다음을 최소화시킨다.

$$\sum_{i=1}^{n}\rho\left(\frac{\gamma_i}{s}\right),\ \gamma_i=Y_i-\sum_{j=1}^{k}X_{ij}\beta_j$$

여기서 s는 γ의 표본 산포의 추정치이다.

오차 분포가 긴 꼬리를 가질때는 Huber, Hampel, Andrew의 Ψ함수를 사용하여 훨씬 바람직한 β_j의 추정치를 얻을 수 있다.

2.2.2. R-추정치

R-추정치는 순위를 이용한 추정치로서 Hodges 등이 발표한 위치 모수 추정법이 Adichie에 의해 회귀 모수 추정법으로 확장되고, Jaeckel 등에 의해 다중 회귀 모형의 추정 문제도 가능하게 되었다.

2.2.3. L_p-추정치

L_p-추정치란 ρ와 Ψ를 다음과 같이 정의하여서

$$\rho(t)=|t|^p$$

$$\Psi(t)=p|t|^{p-1}\mathrm{sign}(t)$$

$$\sum_{i=1}^{n}|Y_i-\sum_{j=1}^{k}X_{ij}\beta_j|^p$$

즉, 잔차의 p승을 최소로 하는 β값이다.

Forsythe[4]는 $N(0,1)$과 $N(\mu,4)$가 0%, 2.5%, 5%, 7.5%, 10% contaminated되어 있을 때 β_1, β_2의 최소 자승 추정치와 p에 1.25, 1.50, 1.75를 대입한 경우의 제곱 평균 오차를 비교하였다.

대칭인 경우 contamination이 증가할수록 작은 p값이 바람직하며, $p=1.5$가 최선의 절충안으로 제안되었다. 최소 자승 추정치에 비해 제일 나쁠 때가 95% 효율을 갖고 contamination이 증가하면 훨씬 높은 효율을 가졌다.

skew된 경우 최소 자승 추정치는 정규분포에서 멀어질수록 효율이 크게 떨어지나, $1<p<2$인 경우는 최소 자승 추정치가 이상적일 때도 효율이 크게 떨어지지 않는다.

Ekblom[2]은 $N(0,1)$과 Laplace, Cauchy분포와 $N(0,1)$과 $N(0,4)$, $N(0,1)$과 $N(0,9)$, $N(0,1)$과 $N(4,1)$의 contaminated 정규 분포와 $\chi^2(6)$에 대해 $L_2, L_{2.5}, L_{1.25}, L_1, L_{0.75}$ 추정치와 Huber의 M-추정치를 단순 회귀 경우와 다중 회귀 경우에 대해 각각 비교 연구하였다. 그 결과 순수한 정규 분포일 때를 제외하고는 Huber의 방법이 L_2보다 우수하였다. 또한 $L_{1.25}$는 $L_{1.5}$보다 성공적이고 $L_{0.75}$는 Cauchy에서 퍽 효과적이었다. 꼬리가 매우 길거나 심히 치우친 분포에는 $p<1$를 사용하는 것이 효과적이다.

3. Adaptive Procedure

Adaptive란, 모형을 선택하기 위해 자료를 수집하고, 자료에서 얻어진 정보를 통해서 선택된 모형에 기초를 두어 통계적 추론을 하는 방법을 가리킨다.

앞에서 소개된 α-trimmed 평균에서 α는 표본을 관측하기 전에 미리 고정된 값이었으나 Jaeckel은 $\hat{\alpha}$가 표본에 의존하는 $m(\hat{\alpha})$를 제안하였고, Hogg는 추정치를 표본첨도 k에 따라 선택하였다.

Hogg와 Fisher, Randles는 꼬리의 길이에 대한 측도로서 다음의 통계치를 사용하였다.

$$Q=\frac{\bar{U}(0.05)-\bar{L}(0.05)}{\bar{U}(0.5)-\bar{L}(0.5)}$$

여기서 $\bar{U}(\beta)$, $\bar{L}(\beta)$는 최고 최저 $n\beta$개의 순서 통계량의 평균이다. Q는 순서 통계량의 선형 함수의 비가 되며, 그 수렴 정도는 첨도 k보다 우수하다.

3.2. Adaptive L_p-norm 추정치의 제안

3.2.1. Adaptive L_p-norm 추정방법

Hogg [5]가 제안한 Adaptive 회귀 분석법의 기본 단계에 따른 adaptive L_p-norm 추정법을 써 보기로 한다.

1. 이상점에 큰 영향을 받지 않는 적당한 추정치를 찾는다. 여기서는 최소 자승 추정치를 사용한다.

2. 잔차를 계산하고 Q를 사용하여 꼬리의 길이를 결정하고 잔차의 분포를 그 길이에 따라 분류한다.

우선 각 분포에 대해 표본 크기 20인 pseudo-random variable을 생성하여 Q값을 구했다. 이것을 100번 시행하여 구한 평균과 표준 편차는 표 1과 같다.

표 1. Simulation을 통한 Q값

분 포	uniform	정규	이중지수	Cauchy
평균(MQ)	1.95	2.44	2.98	4.99
표준편차(σQ)	0.07	0.14	0.31	2.6

각 분포에서 평균은 0, 분산은 1이며, Cauchy에서는

$$f(x)=\frac{1}{\pi\beta\{1+[(x-\alpha)/\beta]^2\}}$$

에서 $\alpha=0$, $\beta=1$로 하였다.

표 2. Q의 범위와 p값

Q범위	꼬리길이	p값
$Q<2$	짧을때	$p=3$
$2<Q<2.6$	중간정도일때	$p=2$
$2.6<Q<3.2$	길때	$p=1.5$
$3.2<Q$	매우 길때	$p=1$

그 결과 Q에 따른 p값을 표 2와 같이 택하였다.

3. 택해진 p값에 따라 추정치를 재계산한다. 즉,

$$\sum_{i=1}^{n}|Y_i-\sum_{j=1}^{k}X_{ij}\beta_j|^p$$

을 최소로 하는 값을 찾는다.

3.2.2. Algorithm

L_p-norm 추정치는 비선형 추정치이므로 수치해를 반복법으로 구하는 것이 일반적인 방법이다.

초기값으로는 최소자승 추정치를 사용하고 L_p-추정치는 가중 최소자승 추정 반복법으로 구할 수 있다.

잔차와 가중치를

$$\gamma_i=Y_i-\sum_{j=1}^{k}X_{ij}\beta_j,\quad W_i=\frac{1}{|\gamma_i|^{2-p}}$$

로 놓으면, L_p-norm 추정치는

$$\sum_{i=1}^{n}W_i|\gamma_i|^2$$

을 최소화하는 값이 되며, 단 γ_i가 10^{-10}보다 작을 때는 10^{-10}으로 계산하였다.

Fletcher, Grant, Hebden [3]과 Ramsay [7]의 algorithm을 이용하면 훨씬 가속적으로 수렴된다.

또한 종종 너무 많은 반복을 하게 되므로 θ의 상한을 0.8로, 하한을 -1.0으로 하고 수렴 공식은

$$\max_j\left|\frac{b_j^{(k)}-b_j^{(k-1)}}{b_j^{(k-1)}}\right|<0.001$$

로 하였다.

3.3. Adaptive L_p-추정치에 대한 Monte Carlo연구

다음과 같은 다중선형 회귀 모형에 대해 연구하였다.

$$Y_i=3+2X_{i2}+X_{i3}+\epsilon_i$$

표본크기는 100으로, 잔차 분포로는 Uniform, 정규분포, 이중지수분포, Cauchy, $N(0,1)$과 $N(0.4)$의 Contaminated정규분포, $N(0,1)$과

$N(0,9)$의 Contaminated정규 분포를, p값으로는 1, 1.25, 2, 3을 사용한다.

추정치의 상대효율을 나타내는 측도로 Moberg, Ramberg, Randles가 제안한 오차 자승의 합(TSE)의 25% trimmed 평균을 사용하였다.

각 분포에 대해 $p=1.25$인 경우와 최소 자승 추정치와 adaptive L_p-norm추정치를 구하였다.

L_1-norm추정치는 Simplex방법을 이용한 Barrodale-Roberts [1]의 L_1 program으로 해를 구했다. 가중 최소자승 추정법으로는 잘 수렴하지 않기 때문이다.

3.4. Monte Carlo연구 결과

각 잔차 분포에서 추정치의 상대 효율은 표 3과 같다.

표 3. 최소 자승 추정치와 비교한 $L_{1.25}$추정치와 adaptive L_p-norm 추정치의 상대효율

분 포	$L_{1.25}$	adaptive
uniform분포	0.562	1.144
정규분포	0.879	1.012
이중지수분포	1.450	1.377
Cauchy분포	9.521	11.092
contaminated정규 1	0.886	0.815
contaminated정규 2	1.195	1.474

결과를 요약해 보면

1. Adaptive L_p-norm 추정치의 효율은 대체로 우수하며 정규분포의 경우도 떨어지지 않았다.

2. 이중 지수 분포의 경우는 $L_{1.25}$가 약간 우수하였다.

3. Cauchy의 경우는 월등히 그 효율이 높았다.

4. contaminated 정규 1의 경우는 최소 자승 추정치에 못 미친 결과가 나타났다.

5. $p=1.25$인 경우 여러번 수렴하지 않는다. (133번중 20번) 15회 반복에도 수렴하지 않을 때에는 최종 반복에서의 해를 사용하였다.

표 4. 100회 시행 중 각 p값의 선택횟수

분포 \ p값	1	1.25	2	3
uniform	0	7	52	41
정규분포	2	19	52	27
이중지수분포	10	31	45	14
Cauchy	70	13	15	2
contaminated 정규 1	9	29	47	15
contaminated 정규 2	20	34	41	5

adaptive경우 선택한 p값들은 표 4와 같다.

L_p-norm 추정치의 초기값으로 최소 자승 추정치 대신 L_1을 사용해 비교해 보아도 좋을 것 같다.

참 고 문 헌

1. Barrodale, I. and Robert, F.D.K. (1973), "An improved algorithm for discrete, linear approximation," *SLAM* 10, 839-848.
2. Ekblom, Hakan. (1974), "Lp-methods for Robust regression," *Nordisk Tidskr. Informations behandling* (*BIT*), 14, 22-32.
3. Fletcher, R., Grant, J.A. and Hebden, M.D. (1971), "The calculation of linear Best Lp Approximations," *Computer Journal.* 14, 276-279.
4. Forsythe, A.B. (1972), "Robust estimation of straight line regression coefficients by minimizing pth power deviations," *Technometrics.* 14, 159-166.
5. Hogg, R.V. (1974), "Adaptive robust procedures, a partial review and some suggestions for future applications and theory," *Journal of American Statistical Association.* 69, 909-927.
6. Huber, P.J. (1964), "Robust estimation of Location parameter," *Annals of Mathematical Statics,* 35, 73-101.
7. Ramsay, J.O. (1975), "Solving Implicit Equations in Psychometric Data Analysis," *Psychometrika.* 40, 337-360.