한국통계학회논문집
2010, 17권, 4호, 541–550

L_1-회귀추정량의 붕괴점 향상을 위한 알고리즘

김부용[1,a]

[a]숙명여자대학교 통계학과

요 약

L_1-회귀추정량이 수직이상점에 대해서는 매우 로버스트하지만 지렛점에 대해서는 전혀 로버스트하지 않다는 사실은 잘 알려져 있다. 본 논문에서는 수직이상점은 물론 지렛점에 대해서도 로버스트한 L_1-회귀추정을 위한 알고리즘을 제안한다. MCD 또는 MVE-추정량에 바탕을 둔 로버스트거리를 기준으로 지렛점들을 식별하고, 식별된 지렛점들의 영향력을 적절히 감소시키기 위한 가중치를 결정한다. 가중치에 의해 변환된 자료에 선형척도변환 기법에 바탕을 둔 선형계획 알고리즘을 적용함으로써 L_1-회귀추정량의 붕괴점을 향상시킨다. 다양한 형태와 규모의 자료에 대한 모의실험 결과, 제안된 알고리즘에 의한 L_1-회귀추정량의 붕괴점이 크게 향상되는 것으로 나타났다.

주요용어: L_1-회귀추정량, 수직이상점, 지렛점, 로버스트추정, 붕괴점.

1. 서론

회귀분석 자료에 이상점들이 포함되는 경우는 흔히 볼 수 있다. 특히, 고객관계관리를 위한 데이터마이닝 분야에서와 같이 철저한 자료수집 계획이나 통제가 이루어지지 않는 상황에서 얻어진 자료에는 수직이상점은 물론 지렛점이 다수 포함될 가능성이 높다. 이러한 회귀이상점들이 자료에 포함된 경우에 최소자승추정에 바탕을 둔 전통적인 회귀분석을 실행하면 그 분석결과는 크게 왜곡될 수밖에 없으므로 이상점에 대한 적절한 대책을 강구해야 하는데, 그중의 한 가지가 로버스트 회귀분석을 채택하는 것이다. 수직이상점에 대해 로버스트한 추정법으로서 L_1-회귀추정(최소절대치추정)이 오래전부터 사용되어 왔는데, 최근에는 데이터마이닝을 위한 신경망분석에서 반응변수가 정규분포를 따르지 않는 경우에 그 선행 단계에서 입력변수를 선정하기 위한 도구로서 L_1-회귀추정이 사용되기도 한다.

L_1-회귀추정량은 선형회귀모형 $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ ($\boldsymbol{y}$는 반응변수 n-벡터, X는 계수가 $p(< n)$인 설명변수 $n \times p$ 행렬, $\boldsymbol{\beta}$는 회귀계수 p-벡터 그리고 $\boldsymbol{\epsilon}$는 오차 n-벡터임)에서 오차 벡터의 L_1-norm인 목적함수 $S(\boldsymbol{\beta}) = \|\boldsymbol{\epsilon}\|_1$을 최소화하는 $\boldsymbol{\beta}$로 정의된다. 특히 오차의 분포가 Laplace분포인 경우에는 L_1-추정량이 최우추정량과 일치한다는 특징을 가지고 있다. 그런데 L_1-회귀추정량은 어떤 함수형태로 표현될 수 없기 때문에 알고리즘을 통해서만 추정치를 구할 수 있다. L_1-추정치를 구하기 위해서는 Barrodale과 Roberts (1973)나 Sherali 등 (1988)의 알고리즘을 사용할 수 있다. 한편 L_1-회귀추정량의 표본분포에 관한 연구로는 Rosenberg와 Carlson (1977)과 Dielman과 Pfaffenberger (1982)가 있는데, 오차가 Laplace분포나 Cauchy분포와 같이 꼬리가 두터운 분포를 따르는 경우에 L_1-추정량은 최소자승추정량(L_2-추정량)보다 효율성이 높다는 사실을 밝혀냈다. L_1-추정량의 다양한 통계적 특성에 관해서는 Blattberg와 Sargent (1971), Pfaffenberger와 Dinkel (1978), Bloomfield와 Steiger (1980) 등이 연구하였으며, Chen과 Wu (1993)은 L_1-추정량의 일치성을 위한 필요조건을 제시하였다. Koenker (1987)와 Dielman과 Pfaffenberger (1992)는 Bassett와 Koenker (1978)가 밝혀낸 L_1-추정량의 근사적 분포를 바

[1] (140-742) 서울시 용산구 청파동 2가, 숙명여자대학교 통계학과, 교수. E-mail: buykim@sm.ac.kr

탕으로 한 가설검정법을 제시하였으며, Kim (2004)은 붓스트랩기법에 바탕을 둔 가설검정을 제안하였다. 그리고 Dielman (2005)은 L_1-회귀에 관련된 통계적 추론들을 종합적으로 소개하였다. 한편, L_1-추정량은 L_2-추정량에 비해 수직이상점에 대한 붕괴점(breakdown point: Montgomery 등 (2006) 참조)이 높은 것으로 잘 알려져 있는데, Kim (1995)은 수직이상점에 대한 L_1-추정량의 붕괴점이 $(n-p)/2n$로서 다른 로버스트 추정량인 LMS-추정량이나 LTS-추정량의 붕괴점 만큼 높다는 사실을 입증하였다. 이러한 이론적 발전에 힘입어 L_1-회귀분석이 널리 활용되고 있다.

그러나 L_1-회귀추정량은 수직이상점에 대해서는 매우 로버스트하지만 지렛점에 대해서는 전혀 로버스트하지 않기 때문에 붕괴점은 $1/n$에 불과하다. 그러므로 계획된 실험에서 얻은 자료와 같이 지렛점은 포함되지 않고 수직이상점만 포함된 자료에 L_1-회귀분석이 제한적으로 사용될 수밖에 없다. 따라서 L_1-회귀분석의 보다 폭넓은 활용을 위해서는 L_1-추정량의 붕괴점을 향상시킬 필요가 있다. 즉, 지렛점에 대해서도 로버스트한 L_1-추정치를 구할 수 있어야 한다. 본 연구에서는 회귀자료에 포함된 지렛점들을 식별하고, 그들의 영향력을 적절히 감소시키기 위한 가중치를 결정하며, 가중치에 의해 변환된 자료에 선형척도변환(linear scaling transformation) 기법에 의한 선형계획 알고리즘을 적용하여 L_1-회귀추정량의 붕괴점을 향상시키고자 한다. 그리고 제안된 알고리즘의 로버스트 성질을 평가하기 위하여, 다양한 형태와 규모의 자료에 대한 몬테칼로 모의실험을 실행하여 제안된 알고리즘에 의한 L_1-회귀추정량의 붕괴점이 향상되는지 확인하고자 한다.

2. 지렛점의 식별

L_1-추정량에 대한 지렛점의 영향을 적절히 조정하기 위해서는 우선 자료에 포함된 지렛점을 정확히 식별해야 한다. 지렛점을 식별하기 위해서는 Hadi (1994)의 방법 등을 활용할 수 있는데, 이러한 식별방법들은 다변량자료에서 이상점을 식별하는 문제와 회귀자료에서 지렛점을 식별하는 문제를 동일하게 취급하며, 대부분 Mahalanobis-제곱거리(MSD)인 $\mathrm{MSD}_i = (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})^T \hat{\Sigma}^{-1} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})$ $(i = 1, \ldots, n$, $\boldsymbol{x}_i^T$는 행렬 X의 i번째 행, $\hat{\boldsymbol{\mu}}$은 위치모수 추정량 벡터, $\hat{\Sigma}$는 산포모수 추정량 행렬임)을 활용한다. 그런데 MSD의 계산을 위하여 위치모수와 산포모수를 각각 전통적인 방식, $\hat{\boldsymbol{\mu}} = n^{-1} \sum_{i=1}^{n} \boldsymbol{x}_i$, $\hat{\Sigma} = (n-1)^{-1} \sum_{i=1}^{n} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})^T$에 의해 추정하는 경우, $\hat{\boldsymbol{\mu}}$과 $\hat{\Sigma}$은 로버스트 추정량이 아니기 때문에 식별과정에 가림현상(masking effect)이나 붙음현상(swamping effect)이 발생할 수 있고, 따라서 MSD를 바탕으로 얻은 지렛점 식별 결과는 정확하지 않을 수 있다. 그러므로 위치모수와 산포모수의 로버스트추정량인 MCD(minimum covariance determinant)-추정량이나 MVE(minimum volume ellipsoid)-추정량에 바탕을 둔 로버스트제곱거리(RSD; robust squared distance)의 분포를 바탕으로 지렛점을 식별할 것을 제안한다.

2.1. MCD/MVE-추정량

Rousseeuw (1985)가 제시한 MCD-추정량 $(\boldsymbol{\mu}_J, \Sigma_J)$은 다음과 같이 정의된다.

$$\Sigma_J = \frac{1}{h} \sum_{i \in J} (\boldsymbol{x}_i - \boldsymbol{\mu}_J)(\boldsymbol{x}_i - \boldsymbol{\mu}_J)^T, \quad \boldsymbol{\mu}_J = \frac{1}{h} \sum_{i \in J} \boldsymbol{x}_i,$$

여기서 $h = [(n+p+1)/2]$ ($[\cdot]$는 최대정수 함수임)이고, J는 h개의 원소로 구성된 지수집합인데 원소의 수가 h인 모든 부분집합 H에 대해 $|\Sigma_J| \le |\Sigma_H|$을 만족시키는 집합이다. 즉, MCD-추정량은 사전에 결정된 크기의 관찰치 부분집합들 중에서 공분산행렬의 행렬식이 최소가 되는 부분집합에서의 위치모수와 산포모수의 추정량이다. 한편, MVE-추정량 $(\boldsymbol{\mu}^*, \Sigma^*)$은 다음과 같이 정의된다. 즉, $\boldsymbol{\mu}^*$는 h개의 관찰치를 포함하는 타원체들 중에서 부피가 최소인 타원체의 중심이고, 공분산 추정량 Σ^*는 $\boldsymbol{\mu}^*$에 대응하

는 공분산에 일치추정량을 위한 인자를 곱한 것으로 얻어진다. MCD-추정량과 MVE-추정량의 붕괴점은 $h = [(n + p + 1)/2]$일 때 $[(n - p + 1)/2]/n$인 것으로 알려져 있는데, 이는 로버스트 추정량들이 가질 수 있는 붕괴점의 최대치에 해당된다. 한편, MCD-추정치나 MVE-추정치를 구하기 위해서는 많은 분량의 공분산행렬의 행렬식이나 타원체의 부피를 계산해야 하기 때문에 막대한 계산이 요구된다. 특히, 규모가 큰 자료에서는 이러한 추정량들의 계산효율성이 심각하게 낮아지는 문제가 발생하는데, 계산효율성이 개선된 Woodruff와 Rocke (1994)와 Rousseeuw와 Driessen (1999)의 알고리즘을 적용할 수 있다.

2.2. RSD0에 의한 지렛점 식별

MCD-추정량과 MVE-추정량을 $(\boldsymbol{\mu}_J^*, \Sigma_J^*)$로 표기하고 MSD에 적용하면 다음과 같은 RSD,

$$\mathrm{RSD}_i = (\boldsymbol{x}_i - \boldsymbol{\mu}_J^*)^T \Sigma_J^{*-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_J^*)$$

가 정의된다. 지렛점을 식별하기 위하여 RSD의 계층적 군집화를 적용할 수 있지만, 데이터마이닝 분야에서와 같이 방대한 자료에 계층적 군집화를 적용하면 계산효율성에 심각한 문제가 발생한다. 그러므로 RSD의 분포를 바탕으로 한 지렛점 식별을 고려할 수 있다. RSD의 분포를 바탕으로 지렛점을 식별하기 위해서는 지렛점의 판정을 위한 경계치를 결정해야 하는데, Hardin과 Rocke (2004)가 제시한 RSD의 근사적 분포로부터 경계치를 구할 수 있다. 즉, RSD의 분포,

$$\frac{d(a - p + 1)}{pa}\mathrm{RSD} \sim F(p,\, a - p + 1)$$

(여기서 $d = nP[\chi^2(p + 2) < \chi^2_{h/n}(p)]/h$이며, a의 계산은 Hardin과 Rocke (2004)을 참조)로부터 경계치 $\kappa = paF_{1-\alpha}(p, a - p + 1)/\{d(a - p + 1)\}$을 구하며, κ보다 큰 RSD에 대응하는 관찰치를 지렛점으로 식별한다.

3. 붕괴점이 향상된 L_1-추정 알고리즘

L_1-추정량이 지렛점에 의해 많은 영향을 받으므로 L_1-추정량의 붕괴점을 향상시키기 위해서는 지렛점에 대해서도 로버스트한 알고리즘을 개발해야 할 필요가 있다. 따라서 본 논문에서는 식별된 지렛점에 적절한 가중치를 부여하고, L_1-추정을 위한 Sherali 등 (1988)의 알고리즘을 수정하여, 수직이상점과 지렛점에 대해 로버스트한 추정치를 얻을 수 있는 알고리즘을 제안하고자 한다. 지렛점들의 영향력을 적절한 수준으로 줄이기 위해서 지렛점에 해당되는 관찰치에는 정상점보다 작은 크기의 가중치를 부여하되, 경계치 κ을 초과하는 RSD에 대응하는 관찰치에 대해서는 (3.1)과 같이 해당 RSD의 크기에 반비례하는 가중치 q_i를 부여한다.

$$q_i = \begin{cases} 1, & \text{for } i \in A, \\ \dfrac{\kappa}{\mathrm{RSD}_i}, & \text{for } i \notin A, \end{cases} \tag{3.1}$$

여기서 $A = \{i \mid \mathrm{RSD}_i \le \kappa\}$는 지렛점이 아닌 정상점들의 지수집합이다. 가중치를 적용하여 설명변수 행렬을 다음과 같이 변환시킬 수 있으며,

$$X_\alpha = Q\left(X - \boldsymbol{\ell}_n \boldsymbol{\mu}_J^{*T}\right) + \boldsymbol{\ell}_n \boldsymbol{\mu}_J^{*T}, \quad Q = \mathrm{diag}\,[q_i]$$

(단, $\boldsymbol{\ell}_n$은 단위벡터임), 변환된 자료 X_α에 L_1-추정 알고리즘을 적용함으로써 L_1-회귀추정량의 붕괴점을 향상시키고자 한다. L_1-추정을 위한 알고리즘들은 목적함수인 $S(\boldsymbol{\beta})$을 최소화하기 위하여 주로 선형계획법을 채택한다. 즉, 다음과 같은 선형계획문제,

$$\text{maximize}\left\{\boldsymbol{\ell}_n^T\boldsymbol{e}^+ + \boldsymbol{\ell}_n^T\boldsymbol{e}^- : X_\alpha\hat{\boldsymbol{\beta}} + I_n\boldsymbol{e}^+ - I_n\boldsymbol{e}^- = \boldsymbol{y},\ \boldsymbol{e}^+ \geq \mathbf{0},\ \boldsymbol{e}^- \geq \mathbf{0}\right\} \tag{3.2}$$

(여기서 벡터 $\boldsymbol{e}^+$와 $\boldsymbol{e}^-$의 원소는 각각 잔차의 양과 음의 편차를 의미함)을 해결함으로써 L_1-추정치를 구할 수 있다. 문제 (3.2)는 일반적인 심플렉스방법에 의해 해결할 수도 있지만, 문제해결에 필요한 계산의 양을 줄이기 위한 다양한 알고리즘들이 개발되었는데 Barrodale과 Roberts (1973), Armstrong 등 (1979), Gentle 등 (1987), Coleman과 Li (1992) 등이 대표적인 것들이다. 그런데 이러한 알고리즘들은 자료의 규모가 클 경우에 막대한 양의 계산이 요구된다는 단점을 극복할 수가 없다. 그래서 Sherali 등 (1988)은 선형척도변환 기법을 적용하여 계산효율성이 개선된 알고리즘을 제안하였는데, 대규모 자료에 적용할 경우에 기존의 알고리즘들보다 계산효율성이 상대적으로 높다는 사실을 입증하였다. 따라서 이 알고리즘을 바탕으로 L_1-추정량의 붕괴점을 향상시킬 수 있는 알고리즘(HBL_1)을 개발하고자 한다.

그러나 선형척도변환 기법을 적용하더라도 관찰치의 수가 많은 경우에는 문제 (3.2)의 제약행렬의 차수가 매우 커지므로 계산효율성이 떨어지게 된다. 그러므로 다음과 같은 쌍대선형계획문제,

$$\text{maximize}\left\{\boldsymbol{y}^T\boldsymbol{\xi} : X_\alpha^T\boldsymbol{\xi} = \mathbf{0},\ -\boldsymbol{\ell}_n \leq \boldsymbol{\xi} \leq \boldsymbol{\ell}_n\right\} \tag{3.3}$$

을 통하여 계산의 양을 줄이면서 최적해를 구하는 시도를 할 수 있다. 그런데 문제 (3.3)를 해결하기 위한 알고리즘의 매 반복에서는 쌍대변수에 가해진 제약 $\boldsymbol{\ell}_n \leq \boldsymbol{\xi} \leq \boldsymbol{\ell}_n$ 때문에 가능해가 다원체(polytope)의 경계에 매우 근접한 값으로 얻어질 수 있다. 그런 경우에는 다음 반복에서 최신화된 가능해가 목적함수 $\boldsymbol{y}^T\boldsymbol{\xi}$을 충분히 증가시키지 못하는 문제를 야기한다. 그러므로 현재의 가능해를 새로운 공간에서의 가능해로 변환시키되, 그 공간에서의 가능해는 제약식의 경계로부터 상당히 떨어진 위치에 있도록 변환하여 다음 반복에서는 가능해가 충분한 거리를 이동할 수 있도록 할 필요가 있다. 따라서 다음과 같은 특수한 선형변환,

$$\boldsymbol{\zeta} = D^{-1}\boldsymbol{\xi},$$
$$D = \text{diag}[v_i], \quad v_i = \begin{cases} 1 - \xi_i, & \text{if } \xi_i \geq 0, \\ 1 + \xi_i, & \text{if } \xi_i < 0 \end{cases}$$

을 통하여 문제 (3.3)를 (3.4)로 변환하게 되는데,

$$\text{maximize}\left\{(D\boldsymbol{y})^T\boldsymbol{\zeta} : (DX_\alpha)^T\boldsymbol{\zeta} = \mathbf{0},\ -\boldsymbol{\ell}_n \leq D\boldsymbol{\zeta} \leq \boldsymbol{\ell}_n\right\} \tag{3.4}$$

어느 경우든 동일한 최적해가 얻어지며 문제 (3.4)에서는 알고리즘의 실행 반복수가 상당히 감소되어 계산효율성이 향상될 수 있다. 한편 문제 (3.4)를 위한 알고리즘의 t-번째 반복에서, 영공간 $(D^{(t)}X_\alpha)^T\boldsymbol{\zeta}^{(t)} = 0$에서의 목적함수 $D^{(t)}\boldsymbol{y}$의 사영은

$$\boldsymbol{h}^{(t)} = \left[I - D^{(t)}X_\alpha\left\{\left(D^{(t)}X_\alpha\right)^T D^{(t)}X_\alpha\right\}^{-1}\left(D^{(t)}X_\alpha\right)^T\right]D^{(t)}\boldsymbol{y}$$

이며, 목적함수의 증가를 가져오는 최신화된 해는 사영된 경사(gradient)를 따라서 적당한 거리를 이동하면 얻을 수 있다. 이동거리를 δ라 하면 최신화된 해는 $\boldsymbol{\zeta}^{(t+1)} = \boldsymbol{\zeta}^{(t)} + \delta\boldsymbol{h}^{(t)}$에 의해 얻어지며, 이를 선

형변환 $\boldsymbol{\zeta} = D^{-1}\boldsymbol{\xi}$에 의해 ξ-공간으로 역변환 시키면, 문제 (3.3)에서의 최신화된 해는 $\boldsymbol{\xi}^{(t+1)} = \boldsymbol{\xi}^{(t)} + \delta D^{(t)}\boldsymbol{h}^{(t)}$가 된다. 이러한 반복과정을 알고리즘의 종료기준이 충족될 때까지 실행하면 최적해를 얻게 된다. 표현을 정형화하기 위하여, 사영된 경사를 $D^{(t)}\boldsymbol{h}^{(t)} = \boldsymbol{g}^{(t)}$로, 이동거리를 $\delta = \lambda/\phi^{(t)}$ $(0 < \lambda < 1)$로 표기하면, 문제 (3.3)의 최적해는 다음과 같이 표현된다.

$$\boldsymbol{\xi}^{(t+1)} = \boldsymbol{\xi}^{(t)} + \frac{\lambda}{\phi^{(t)}}\boldsymbol{g}^{(t)}.$$

그런데 이와 같은 최적해는 문제 (3.3)의 가능해 영역 안에 위치해야 하기 때문에 각 반복에서 $-1 \le \xi_i^{(t)} + g_i^{(t)}/\phi^{(t)} \le 1$의 관계가 성립되어야 하며, 따라서 $\phi^{(t)}$는 다음과 같이 결정되어야 한다.

$$\phi^{(t)} = \max_i \left[\max \left\{ \frac{g_i^{(t)}}{1-\xi_i^{(t)}}, \frac{-g_i^{(t)}}{1+\xi_i^{(t)}} \right\} \right].$$

한편, 이와 같은 반복과정을 통하여 구한 해가 최적해인지 여부는 다음과 같이 확인된다. t-번째 반복에서 만약 $\boldsymbol{g}^{(t)} = \mathbf{0}$이면 $\boldsymbol{\xi}^{(t)}$는 당연히 최적해이기 때문에 $\boldsymbol{g}^{(t)} \neq \mathbf{0}$인 경우만을 살펴보기로 한다. 사영된 경사,

$$\boldsymbol{g}^{(t)} = D^{(t)^2}\boldsymbol{y} - D^{(t)^2}X_\alpha \left\{ \left(D^{(t)}X_\alpha\right)^T D^{(t)}X_\alpha \right\}^{-1} \left(D^{(t)}X_\alpha\right)^T D^{(t)}\boldsymbol{y}$$

은

$$\left\{ \boldsymbol{y} - \left(D^{(t)^2}\right)^{-1}\boldsymbol{g}^{(t)} \right\}^T = \boldsymbol{y}^T D^{(t)^2} X_\alpha \left\{ X_\alpha^T D^{(t)^2} X_\alpha \right\}^{-1} X_\alpha^T \tag{3.5}$$

와 같이 표현할 수 있으며, $\boldsymbol{\xi}^{(t)}$의 가능해 성질을 표현하는 식 $X_\alpha^T(\boldsymbol{\xi}^{(t+1)} - \boldsymbol{\xi}^{(t)}) = \mathbf{0}$으로부터 다음과 같은 관계식으로 얻을 수 있다.

$$\boldsymbol{y}^T D^{(t)^2} X_\alpha \left\{ X_\alpha^T D^{(t)^2} X_\alpha \right\}^{-1} X_\alpha^T \left(\boldsymbol{\xi}^{(t+1)} - \boldsymbol{\xi}^{(t)}\right) = \mathbf{0} \tag{3.6}$$

식 (3.5)와 (3.6)을 연결시키면

$$\begin{aligned} \boldsymbol{y}^T\left(\boldsymbol{\xi}^{(t+1)} - \boldsymbol{\xi}^{(t)}\right) &= \boldsymbol{g}^{(t)^T}\left(D^{(t)^2}\right)^{-1}\left(\boldsymbol{\xi}^{(t+1)} - \boldsymbol{\xi}^{(t)}\right) \\ &= \frac{\lambda}{\phi^{(t)}}\boldsymbol{g}^{(t)^T}\left(D^{(t)^2}\right)^{-1}\boldsymbol{g}^{(t)} \\ &> 0 \end{aligned} \tag{3.7}$$

이므로 $\boldsymbol{y}^T\boldsymbol{\xi}^{(t+1)} > \boldsymbol{y}^T\boldsymbol{\xi}^{(t)}$임을 확인할 수 있다. 그런데 수열 $\{\boldsymbol{y}^T\boldsymbol{\xi}^{(t)}\}$는 약한 쌍대성이라는 성질에 따라 상단제약 되므로 결국 $\{\boldsymbol{y}^T\boldsymbol{\xi}^{(t)}\}$는 수렴하게 된다. 따라서 알고리즘을 통하여 구한 해는 최적해라고 할 수 있다. 그리고 식 (3.7)로부터 알고리즘의 종료기준을 $\|\boldsymbol{g}^{(t)}\|_\infty < \eta$ (단, $\|\cdot\|_\infty$은 L_∞-norm, η는 공차를 의미하는데 Sherali 등 (1988)은 $\eta = 10^{-6}$, $\lambda = 0.97$을 제안하였음)로 설정하는 것이 타당함을 알 수 있다.

위와 같은 반복과정을 통하여 최적해 $\boldsymbol{\xi}^{(t)}$을 구하면 L_1-추정치 $\hat{\boldsymbol{\beta}}$은 다음과 같이 구한다. 즉, (3.2)에서의 $\hat{\boldsymbol{\beta}}$은 (3.4)의 제약식 $(DX_\alpha)^T\boldsymbol{\zeta} = \mathbf{0}$에 대응하는 쌍대변수이므로 $\hat{\boldsymbol{\beta}}^T(DX_\alpha)^T = (D\boldsymbol{y})^T$의 관계가 성립하며, $(DX_\alpha)^T$가 완전계수를 갖는다는 가정에 의해 L_1-추정치는

$$\hat{\boldsymbol{\beta}} = \left(X_\alpha^T D^{(t)^2} X_\alpha\right)^{-1} X_\alpha^T D^{(t)^2}\boldsymbol{y}$$

와 같이 구한다. 위와 같은 반복과정을 자세히 기술하면 다음과 같다.

알고리즘: HBL_1

단계 1) 로버스트제곱거리 $\text{RSD}_i = (\boldsymbol{x}_i - \boldsymbol{\mu}_J^*)^T \Sigma_J^{*-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_J^*)$ (단, $\boldsymbol{\mu}_J^*, \Sigma_J^*$는 MCD/MVE -추정치)를 계산하고, RSD_i의 경계치 $\kappa = paF_{1-\alpha}(p, a-p+1)/\{d(a-p+1)\}$을 적용하여 지렛점을 식별한다.

단계 2) 가중치행렬 $Q = \text{diag}[q_i]$, $q_i = 1$ for $i \in A$, $q_i = \kappa/\text{RSD}_i$ for $i \notin A$ (단, A는 정상점들의 지수집합)을 구성하고, 행렬 X를 $X_\alpha = Q(X - \boldsymbol{\ell}_n \boldsymbol{\mu}_J^{*T}) + \boldsymbol{\ell}_n \boldsymbol{\mu}_J^{*T}$ 와 같이 변환한다.

단계 3) 반복수 $t = 0$, 초기치 $\boldsymbol{\xi}^{(0)} = \mathbf{0}$을 지정한다.

단계 4) 대각행렬 $D^{(t)} = \text{diag}[\nu_1^{(t)}, \ldots, \nu_n^{(t)}]$, $\nu_i^{(t)} = \min\{1 + \xi_i^{(t)}, 1 - \xi_i^{(t)}\}$을 구성하고, 다음과 같이 X_α와 $\boldsymbol{y}$를 변환한다. $\tilde{X}_\alpha = D^{(t)} X_\alpha$, $\tilde{\boldsymbol{y}} = D^{(t)} \boldsymbol{y}$.

단계 5) $\boldsymbol{g}^{(t)} = D^{(t)}\{I - \tilde{X}_\alpha (\tilde{X}_\alpha^T \tilde{X}_\alpha)^{-1} \tilde{X}_\alpha^T\} \tilde{\boldsymbol{y}}$를 계산한 후, 만약 $\|\boldsymbol{g}^{(t)}\|_\infty < \eta$이면 L_1-추정치 $\hat{\boldsymbol{\beta}} = (\tilde{X}_\alpha^T \tilde{X}_\alpha)^{-1} \tilde{X}_\alpha^T \tilde{\boldsymbol{y}}$를 구하고 알고리즘을 종료한다.

단계 6) 그러나 $\|\boldsymbol{g}^{(t)}\|_\infty \geq \eta$이면 최신화 과정 $\boldsymbol{\xi}^{(t+1)} = \boldsymbol{\xi}^{(t)} + (\lambda/\phi^{(t)}) \boldsymbol{g}^{(t)}$, $\phi^{(t)} = \max[\max\{g_i^{(t)}/(1 - \xi_i^{(t)}), -g_i^{(t)}/(1 + \xi_i^{(t)})\}]$을 통해 새로운 가능해를 구하고 단계 4로 간다.

4. 알고리즘의 평가

제안된 알고리즘 HBL_1에 의한 추정이 L_2-회귀추정보다 수직이상점은 물론 지렛점에 대해서 더 로버스트한지, 그리고 전통적인 L_1-회귀추정보다 지렛점에 대해서 더 로버스트한지 확인하고, 로버스트 회귀추정으로 잘 알려진 LMS-추정이나 LTS-추정의 결과와 비교분석하기 위하여 모의실험을 실행하였다. 모의실험의 반복수는 1,000회이며 프로그램은 SAS/IML을 사용하였다.

4.1. 자료의 생성

모의실험을 위하여 다양한 규모와 특성을 갖는 자료를 생성하였다. 설명변수의 수(2 (1) 6)와 관찰치의 수(30 (10) 110)의 모든 조합에 해당하는 설명변수 행렬들을 생성하였다. 정상점들만 갖는 자료의 설명변수들의 값은 각각 정규분포$(0, 1)$로부터 난수를 생성하였으며 절편을 포함한 설명변수 행렬 $X_{n \times p}$를 구성하였다. 그리고 정규분포$(0, 1)$로부터 생성한 오차에 사전에 규정한 회귀계수 값 $\boldsymbol{\beta}_0 = [1, 1, \ldots, 1]^T$을 적용하여 반응변수 벡터 $\boldsymbol{y}$를 생성하였다. 한편, 자료에 지렛점과 수직이상점을 심기 위하여 관찰치의 일부를 다음과 같이 오염시켰다. 즉, 강도가 약한 지렛점들은 정규분포$(2, 1)$로부터, 강한 지렛점들은 정규분포$(3, 1)$로부터 생성하였으며, 지렛점의 비중을 두 가지 수준$(5\%, 10\%)$으로 설정하여 설명변수의 정상점 일부를 지렛점들로 교체하였다. 또한 수직이상점들을 자료에 심었는데 정규분포$(5, 1)$로부터 난수를 생성하여 반응변수의 정상점 일부$(5\%, 10\%)$를 수직이상점들로 교체하였다. 따라서 표 1의 시나리오와 같이 이상점들이 포함되지 않은 자료, 수직이상점들만 포함된 자료, 약한 지렛점들만 포함된 자료, 강한 지렛점들만 포함된 자료, 수직이상점들과 약한 지렛점들이 동시에 포함된 자료, 수직이상점들과 강한 지렛점들이 동시에 포함된 자료 등을 생성하였다.

4.2. 평가 결과

제안된 알고리즘이 수직이상점은 물론 지렛점에 대해서도 로버스트한지 평가하기 위하여 L_2-추정, 전통적인 L_1-추정, HBL_1-추정, LMS-추정 그리고 LTS-추정의 결과를 비교분석하였다. 추정량의 붕괴

표 1: 추정법별 ADD 및 AVE의 평균

자료의 시나리오(단위: %)			ADD					AVE				
수직 이상점	약한 지렛점	강한 지렛점	L_2	L_1	HBL_1	LMS	LTS	L_2	L_1	HBL_1	LMS	LTS
0	0	0	.1035	.1303	.1303	.2411	.2510	.0177	.0280	.0280	.0950	.1025
5	0	0	.1930	.1417	.1418	.2379	.2499	.0350	.0317	.0317	.0929	.1024
10	0	0	.2768	.1582	.1583	.2364	.2470	.0468	.0355	.0355	.0920	.1002
0	5	0	.3769	.2176	.1303	.2475	.2575	.0314	.0373	.0281	.0992	.1070
0	10	0	.4689	.3725	.1303	.2501	.2606	.0351	.0531	.0281	.0994	.1076
0	0	5	.5133	.3162	.1304	.2497	.2598	.0372	.0487	.0281	.1018	.1093
0	0	10	.5908	.5608	.1304	.2552	.2649	.0407	.0640	.0281	.1042	.1117
5	5	0	.3804	.2263	.1418	.2462	.2562	.0454	.0428	.0317	.0974	.1057
10	10	0	.5135	.4080	.1583	.2479	.2582	.0588	.0704	.0355	.0964	.1052
5	0	5	.5161	.3359	.1420	.2486	.2593	.0495	.0579	.0318	.1002	.1089
10	0	10	.6427	.5838	.1585	.2526	.2631	.0617	.0816	.0356	.1008	.1095

점은 기본적으로 편의의 크기와 관련이 있기 때문에 각 추정법에 의한 추정치 $\hat{\boldsymbol{\beta}}$이 사전에 규정한 회귀계수 값 $\boldsymbol{\beta}_0$로부터 얼마나 떨어진 값을 얻게 되는지 측정하기 위하여 추정치와 규정치의 절대편차평균(ADD: $\sum_{j=1}^{p}\sum_{m=1}^{1000}|\hat{\beta}_j^{(m)}-\beta_{0j}|/1000p$)을 구하였다. 그리고 각 추정량의 효율성을 평가하기 위하여 추정치의 분산의 평균(AVE: $\sum_{j=1}^{p}\sum_{m=1}^{1000}(\hat{\beta}_j^{(m)}-\bar{\hat{\beta}}_j)^2/999p$)을 계산하였다. 각 시나리오별로 다양한 설명변수의 수와 관찰치의 수에 따라 자료를 생성하고 ADD 및 AVE를 측정하였는데, 동일한 시나리오에서는 각 추정법의 ADD 및 AVE가 설명변수의 수나 관찰치의 수에 따라 크게 다르지 않으므로 측정치들의 평균을 구하였다.

추정법별로 측정한 ADD와 AVE의 평균은 표 1에 수록되었으며, 각 추정법들의 평가결과는 다음과 같다. (1)이상점들이 포함되지 않은 자료의 경우: L_2-추정의 ADD 및 AVE가 가장 작게 나타났으며 L_1-추정과 HBL_1-추정의 ADD와 AVE는 L_2-추정의 경우보다는 약간 크게 측정되었다. 그러나 LMS-추정과 LTS-추정의 ADD와 AVE는 상대적으로 크게 나타났다. 예상대로, 이상점들이 포함되지 않은 자료의 경우에는 L_2-추정이 가장 적합한 것으로 나타났다. (2)수직이상점들만 포함된 자료의 경우: L_1-추정과 HBL_1-추정의 ADD 및 AVE가 가장 작게 측정되었으며 두 추정 사이에는 별 차이가 없었다. 특히 L_2-추정은 L_1-추정이나 HBL_1-추정보다 수직이상점에 대해 로버스트하지 않다는 사실을 확인할 수 있다. (3)지렛점들만 포함된 자료의 경우: HBL_1-추정의 ADD 및 AVE가 가장 작게 측정되었으며, L_2-추정과 L_1-추정의 ADD와 AVE는 상대적으로 크게 측정되었다. 특히 지렛점의 수준이 강하거나 지렛점의 비중이 높을수록 L_2-추정과 L_1-추정의 ADD와 AVE가 커졌다. 따라서 HBL_1-추정이 지렛점에 대해 상대적으로 로버스트하다고 할 수 있다. (4)수직이상점들과 지렛점들이 동시에 포함된 자료의 경우: 이 경우에도 HBL_1-추정의 ADD 및 AVE가 가장 작고 L_2-추정과 L_1-추정의 ADD와 AVE는 매우 크게 측정되었다. 예상과 달리 LMS-추정과 LTS-추정의 ADD와 AVE도 HBL_1-추정보다 상당히 크게 측정되었는데, 그 이유는 LMS-추정과 LTS-추정을 위한 SAS/IML의 알고리즘이 계산효율성을 높이기 위하여 근사적인 방법을 사용하기 때문인 것으로 해석된다. 전반적으로, 수직이상점들이나 지렛점들이 포함된 자료에서 HBL_1-추정의 ADD와 AVE가 가장 작게 나타났으므로 알고리즘 HBL_1에 의한 추정량의 붕괴점이 가장 높으며 효율성도 우수한 것으로 평가된다.

5. 결론

이상점들은 통계적 추론에 막대한 영향을 미치기 때문에 최소자승추정에 의한 회귀분석 결과 역시 심하게 왜곡될 수 있다. 따라서 회귀분석에 앞서 자료에 수직이상점이나 지렛점이 존재하는지를 확인

하고 적절한 로버스트 추정법을 적용해야 한다. 로버스트 회귀추정의 하나인 L_1-추정은 수직이상점에 대해서는 매우 로버스트하지만 지렛점에 대해서는 로버스트하지 않기 때문에, 훌륭한 통계적 특성을 가지고 있음에도 불구하고 널리 사용되지 않고 있다. 따라서 L_1-회귀분석의 폭넓은 활용을 위해서는 지렛점에 대해서도 로버스트한 L_1-추정알고리즘이 요구된다. 본 논문에서는 지렛점들의 영향력을 적절히 감소시키기 위하여, MCD 또는 MVE-추정량에 바탕을 둔 로버스트거리를 기준으로 가중치를 결정하고, 이 가중치에 의해 변환된 자료에 선형척도변환 기법에 바탕을 둔 선형계획 알고리즘을 적용하여 L_1-회귀추정량의 붕괴점을 향상시켰다. 다양한 형태와 규모의 자료에 대한 모의실험 결과, 수직이상점이나 지렛점이 포함된 자료에서 제안된 알고리즘 HBL_1에 의한 추정량의 붕괴점과 효율성이 L_2-추정, L_1-추정, LMS-추정 그리고 LTS-추정보다 높은 것으로 나타났다.

참고 문헌

Armstrong, R. D., Frome, E. L. and Kung, D. S. (1979). A revised simplex algorithm for the absolute deviation curve fitting problem, *Communications in Statistics - Simulation and Computation*, **8**, 175–190.

Barrodale, I. and Roberts, F. D. K. (1973). An improved algorithm for discrete linear approximation, *SIAM Journal on Numerical Analysis*, **10**, 839–848.

Bassett, G. and Koenker, R. (1978). Asymptotic theory of least absolute error regression, *Journal of the American Statistical Association*, **73**, 618–622.

Blattberg, R. and Sargent, T. (1971). Regression with non-Gaussian stable disturbances; some sampling results, *Econometrica*, **39**, 501–510.

Bloomfield, P. and Steiger, W. (1980). Least absolute deviations curve-fitting, *SIAM Journal on Scientific Computing*, **1**, 290–301.

Chen, X. R. and Wu, Y. (1993). On a necessary condition for the consistency of the L_1-estimates in linear regression models, *Communications in Statistics - Theory and Methods*, **22**, 631–639.

Coleman, T. F. and Li, Y. (1992). A globally and quadratically convergent affine scaling method for linear problems, *Mathematical Programming*, **56**, 189–222.

Dielman, T. E. (2005). Least absolute value regression: recent contributions, *Journal of Statistical Computation and Simulation*, **75**, 263–286.

Dielman, T. E. and Pfaffenberger, R. (1982). LAV estimation in linear regression; a review, *TIMS/Studies in the Management Sciences*, **19**, 31–52.

Dielman, T. E. and Pfaffenberger, R. (1992). A further comparison of tests of hypothesis in LAV regression, *Computational Statistics & Data Analysis*, **14**, 375–384.

Gentle, J. E., Narula, S. C. and Sposito, V. A. (1987). Algorithms for unconstrained L_1 linear regression, In *Statistical Data Analysis based on the L_1-norm and Related Methods*, edited by Y. Dodge, North-Holland, 83–94.

Hadi, A. S. (1994). A modification of a method for the detection of outliers in multivariate samples, *Journal of the Royal Statistical Society*, **56**, 393–396.

Hardin, J. and Rocke, D. M. (2004). Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator, *Computational Statistics & Data Analysis*, **44**, 625–638.

Kim, B. Y. (1995). On the robustness of L_1-estimator in linear regression models, *The Korean Communications in Statistics*, **2**, 277–287.

Kim, B. Y. (2004). Resampling-based hypothesis test in L_1-regression, *The Korean Communications in Statistics*, **11**, 643–655.

Koenker, R. (1987). A comparison of asymptotic testing methods for L_1-regression, In *Statistical Data Analysis based on the L_1-norm and Related Methods*, ed. by Y. Dodge. 287–298.

Montgomery, D. C., Peck, E. A. and Vining, G. G. (2006). *Introduction to Linear Regression Analysis*, John Wiley & Sons, New Jersey.

Pfaffenberger, R. C. and Dinkel, J. J. (1978). Absolute deviations curve fitting; An alternative to least squares, In *Contributions to Survey Sampling and Applied Statistics*, edited by H. A. David, Academic Press, New York, 279–294.

Rosenberg, B. and Carson, D. (1977). A simple approximation of the sampling distribution of least absolute residuals regression estimates, *Communications in Statistics - Simulation and Computation*, **6**, 421–437.

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point, *Mathematical Statistics and Applications*, B, ed. by W. Grossmann, G. Pflug, I. Vincze, and W. Werz.

Rousseeuw, P. J. and Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator, *Technometrics*, **41**, 212–223.

Sherali, H., Skarpness, B. and Kim, B. Y. (1988). An assumption-free convergence analysis for a perturbation of the scaling algorithm for linear programs, with application to the L_1-estimation problem, *Naval Research Logistics*, **35**, 473–492.

Woodruff, D. L. and Rocke, D. M. (1994). Computable robust estimation of multivariate location and shape in high dimension using compound estimators, *Journal of the American Statistical Association*, **89**, 888–896.

2010년 3월 접수; 2010년 5월 채택

Algorithm for the L_1-Regression Estimation with High Breakdown Point

Bu-Yong Kim[1,a]

[a]Department of Statistics, Sookmyung Women's University

Abstract

The L_1-regression estimator is susceptible to the leverage points, even though it is highly robust to the vertical outliers. This article is concerned with the improvement of robustness of the L_1-estimator. To improve its robustness, in terms of the breakdown point, we attempt to dampen the influence of the leverage points by means of reducing the weights corresponding to the leverage points. In addition the algorithm employs the linear scaling transformation technique, for higher computational efficiency with the large data sets, to solve the linear programming problem of L_1-estimation. Monte Carlo simulation results indicate that the proposed algorithm yields L_1-estimates which are robust to the leverage points as well as the vertical outliers.

Keywords: L_1-estimation, vertical outlier, leverage point, robustness, breakdown point.

[1] Professor, Department of Statistics, Sookmyung Women's University, Chungpa-dong, Yongsan-ku, Seoul 140-742, Korea. E-mail: buykim@sm.ac.kr