한국경영과학회/대한산업공학회 2003 춘계공동학술대회
2003년 5월16일-17일, 한동대학교(포항)

Regression by Least Absolute Value Method with L1-constraint on Parameters

고영현, 전치혁
포항공과대학교 산업공학과
(Department of Industrial Engineering, POSTECH)

Abstract

OLS로 알려진 기존의 추정 방법은 변수수의 증가에 따라 다중공선성(Multicollinearity)의 문제와 더불어 해석력(interpretability)이 떨어지는 문제를 가지게 된다. 본 연구에서는 파라미터의 절대값의 크기(L1-Norm)에 제약을 줌으로써 이와 같은 OLS의 문제를 해결할 수 있는 동시에, 잔차의 제곱합 대신 절대오차를 사용하는 Least Absolute Value(LAV) 방법을 사용함으로써 이상치에 로버스트한 결과를 주는 방법론을 제안한다. 또한, 본 연구에서 제안하는 방법이 선형계획법에 의해 모델링 될 수 있는 특성으로 인해 제약조건이 있는 이차 형태의 최적화 문제보다 수행 속도면에서 뛰어난 결과를 주는 것을 수치예제을 통해 보인다.

1. 연구배경

다음과 일반적인 회귀 모형을 생각해 보자.

$$y_i = \alpha + \mathbf{x}_i^T\boldsymbol{\beta} + \varepsilon_i, \quad i = 1,2,...N \qquad (1)$$

일반적으로 이러한 문제를 해결하기 위해 제곱오차합(sum of squrare error)를 최소화 하기 위해 Ordinary Least Square (OLS) 추정 방법이 주로 사용된다. 그러나 OLS 방법의 경우, 독립변수 x의 개수가 증가하면, 독립변수들 사이의 강한 상관 관계로 인한 다중공선성(multi-collinearity)이 존재할 수 있으며, 따라서 OLS로 추정된 회귀계수의 분산이 커져 추정 회귀식의 예측력(prediction accuracy)이 떨어지는 문제점이 있다[2,3]. OLS 방법론을 이용한 추정 회귀계수의 문제점을 개선하기 위한 방법으로, Ridge Regression(RR), 주성분회귀(principal component regression : PCR), PLS(partial least square)등의 회귀 분석 자체의 변형에 의한 방법이 있을 수 있다. 이들 연구는 많은 분야에서, 다중공선성의 문제를 해결하여 좋은 예측 결과를 보여주는 것으로 알려져 있다[2,3]. 하지만, 이들 방법의 문제점은 독립 변수의 개수가 증가할 경우 변수에 대한 해석력(interpretability)이 떨어진다는데 있다. 즉, 많은 독립 변수 중 어떤 변수가 중요한 역할을 하는지에 대한 판단이 어려워진다는 점이다. 이를 해결하기 위한 방법으로는, 단계별 변수 선택법(stepwise variable selection)처럼 많은 독립 변수 중에 모델에 적합한 일부 변수를 선택하여 모델을 구성하는 방법이 가장 대표적인 방법이라 할 수 있다[5].

RR, PCR, PLS 방법은 추정 회귀계수의 크기를 축소(shrinkage)하여 추정이 보다 안정적(stable)하도록 만드는 방법으로 해석될 수 있으며, 변수 선택 방법은 중요한 변수만을 선택하여 모델을 안정적으로 만드는 방법이다. Tibshirani(1996)은 이 둘의 방법을 결합한 Least absolute shrinkage and selection operator(Lasso) 방법을 개발하였다[6]. 하지만 이 방법은 회귀계수의 추정을 위해 2차계획법(quadratic programming)을 사용하여야 하는 단점이 존재한다.

본 연구에서는 Lasso방법의 2-Norm

손실함수(loss function)를 1-Norm 손실함수로 변형함으로 2차계획법을 1차계획법 변형하여 보다 빠른 추정이 가능하도록 하는 동시에 보다 로버스트한 회귀 계수 추정 방법을 제안한다.

2. 2. 1-Norm Loss Function Approach to the Lasso

데이터가 다음 형태로 존재한다고 할 때, $(\mathbf{x}_i,\ y_i),\ i=1,2,..N$, 단 $\mathbf{x}_i=(x_{1i},\ x_{2i},\cdots\ x_{1p})^T$, $\hat{\boldsymbol{\beta}}=(\hat{\beta}_1,\hat{\beta}_2,...,\hat{\beta}_p)^T$ 라 하면, Lasso의 추정치는 다음과 같이 정의 된다.

$$(\hat{\alpha},\hat{\boldsymbol{\beta}})=\arg\min\left\{\sum_{i=1}^{N}(y_i-\alpha-\sum_j\beta_j x_{ij})^2\right\} \quad \text{subject to}\sum_j|\beta_j|\le t \qquad (2)$$

여기서 t 는 축소의 정도를 조절하는 파라미터가 된다. 보통 많이 알려진 분산 축소 기법인 능형회귀분석(Ridge Regression : RR)의 경우는 식(2)의 제약조건이 아래와 같이 회귀계수의 이차항의 합이 제한조건이 되는 문제로 정의된다 [2,3].

$$\sum_j\beta_j^2\le t \qquad (3)$$

Lasso 추정 방법은 전통적인 능형회귀 방법의 제약조건을 식(2)와 같이 변형시킨 방법이라 할 수 있고, 이 제약 조건의 특성상 t가 줄어듦에 따라 중요하지 않은 변수들의 회귀계수들부터 차례로 0을 만들게 되는 특성을 지닌다. $t_0=\sum_j|\beta_j^{OLS}|$라 하면, $t\ge t_0$ 인 경우엔 정확히 $\boldsymbol{\beta}^{Lasso}=\boldsymbol{\beta}^{OLS}$ 가 되며, t가 t_0 보다 작아짐에 따라 변수선택 및 축소의 영향이 반영되는 알고리즘이라 할 수 있다. 식(2)에서 알 수 있듯이 Lasso 추정치를 구하기 문제는 선형 제한조건이 있는 이차계획법 문제가 된다. 이 문제의 경우, 목적식의 Hessian 행렬의 크기가 $p\times p$ 가 되고, 따라서 이 문제를 푸는데 $O(p^3)$의 계산을 요구하게 된다. 즉, 변수의 개수가 증가하게 됨에 따라 많은 시간을 요하게 된다.

본 연구에서는 목적식의 제곱 오차를 절대 오차로 바꾸는 간단한 아이디어를 통해 문제를 제한 조건이 있는 선형 계획법 문제로 변형할 수 있음을 보인다. 절대 오차를 사용하면 식(2)는 다음처럼 변형될 수 있다.

$$(\hat{\alpha},\hat{\boldsymbol{\beta}})=\arg\min\left\{\sum_{i=1}^{N}|y_i-\alpha-\sum_j\beta_j x_{ij}|\right\} \quad \text{subject to}\sum_j|\beta_j|\le t \qquad (4)$$

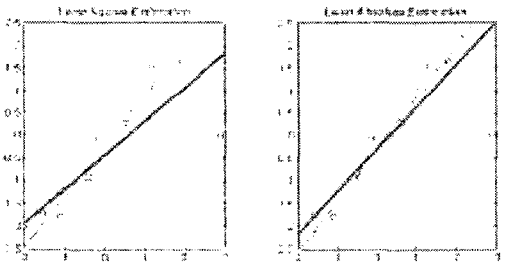
또한 절대값 부호를 없애기 위해, $u_i,v_i,\beta_j^+,\beta_j^-\ge 0$ 의 인공변수(artifitial variable)을 도입하여 $u_i+v_i=|y_i-\alpha-\sum_j\beta_j x_{ij}|$, $|\beta_j|=\beta_j^++\beta_j^-$ 로 치환할 수 있고, 식 (4)는 식 (5)와 같은 선형 계획법 문제로 변형될 수 있다[4].

$$\begin{aligned}&(\hat{\alpha},\hat{\boldsymbol{\beta}})=\arg\min\left\{\sum_{i=1}^{N}(u_i+v_i)\right\}\\ &\text{s.t}\ \sum_j(\beta_j^++\beta_j^-)\le t\\ &u_i-v_i+\alpha+\sum_j(\beta_j^+-\beta_j^-)x_{ij}=y_i\ \ for\ 1\le i\le N\\ &u_i,v_i,\beta_j^+,\beta_j^-\ge 0\end{aligned} \qquad (5)$$

본 연구에서는 이와 같은 최적화 문제를 1차 손실함수를 사용함으로 1-Norm Lasso라 부르기로 한다.

일반적인 최적화 문제에서, 다음과 같은 $L_2:E(Y-f(x))^2$ 처럼 2차 손실 함수를 최소화하는 방법을 LS(least square) 방법이라 부르고, $L_1:E|Y-f(x)|$ 처럼 1차 손실 함수를

사용하는 방법을 LAV(least absolute value) 방법이라 부른다. 보통 2차 손실함수는 미분 가능하다는 장점 때문에 계산상 이점과 통계적 해석이 용이하다는 많은 장점을 지닌다. 반면에 1차 손실 함수는 미분이 불가능하므로 분석의 어려움을 이유로 2차 손실함수에 비해 널리 사용되고 있지는 않지만 이상치에 로버스트한 추정 결과를 준다는 큰 장점을 가진다. 예를 들어, 이차 손실 함수를 이용해 떤 점 $X=x$ 에서의 Y 값을 추정하는 문제는 $\mathrm{argmin}_c E_{Y|x}([Y-c]^2 \mid X=x)$ 로 정의 되고 그 해는 그 해는 $\hat{c}=E(Y \mid X=x)$ 이지만 1차 손실 함수를 사용한 경우 문제는 $\mathrm{argmin}_c E_{Y|x}(|Y-c| \mid X=x)$ 로 정의되고 그 해는 $\hat{c}=median(Y \mid X=x)$ 으로 된다. 즉, 1차 손실함수를 적용한 경우, 중간값이 가지는 성질과 유사하게 이상치에 로버스터한 회귀계수를 얻어낼 수 있다. [그림 1]은 이상치가 존재할 때 OLS 및 LAV의 방법이 이상치에 영향을 받는 정도를 보여준다. [그림 1]에서도 알 수 있듯이, 이상치가 존재할 경우 LAV의 방법이 OLS 방법보다 적게 영향을 받고 있음을 알 수 있다. 따라서 본 연구에서 제안하는 방법은 계산상으로 빠르면서, 로버스트한 예측 결과를 주는 2가지 장점을 가질 것으로 기대되고 4장의 수치예제를 통해 이를 보이도록 한다.



(a) OLS (b) LAV

[그림 1] 이상치에 의한 OLS 방법과 LAV 추정 회귀계수의 변화

3. 조절 파라미터 t의 선택

$\sum_j |\beta_j| \le t$ 의 제약식에서 알 수 있듯이 t 가 충분히 큰 값이라면 회귀계수에 대한 제약이 전혀 없는 LAV의 추정 방법과 동일한 결과를 준다는 것을 알 수 있다. 따라서, β^{LAV}를 LAV 방법에 의해 추정된 회귀 계수라 할 때, $t_0=\sum_j |\beta_j^{LAV}|$ 를 조절 파라미터의 최대값으로 설정할 수 있고, t 의 값이 줄어듬에 따라 회귀계수의 축소와 더불어 변수선택이 일어나게 된다. 따라서 $0<t\le t_0$ 의 구간에서 t를 변화시켜가며, 교차 타당성(cross validation) 방법에 의해 최적의 t 를 선택해 낸다. 교차 타당성(cross validation) 방법은 모델을 형성하기 위한 데이터를 k개로 나누어 k 개의 집단중에 k -1개를 모델 구축에 사용하고 나머지 한 개의 집단을 타당성 검정에 사용하는 것을 k번 반복하는 방법으로 아래와 같이 계산될 수 있다.

$$CUMPRESS=\frac{\sum_{i}^{k}\sum_{j}^{m} L(y_{ij},\hat{y}_{ij})}{N} \qquad (6)$$

여기서 m은 각 집단의 데이터의 개수, $\hat{y}_{i.}$ 은 전체 k 개의 데이터 집단 중 i 번째 데이터 집단을 제외하고 모델을 만들었을 때의 예측치의 값을 의미하고 본 연구에서는 k 를 10으로 하는 10-fold 교차타당성 방법을 사용한다[3].

4. 수치예제

본 장에서는 본 연구에서 제안하는 1-Norm lasso의 성능을 비교하기 위해 모형예측력과 수행속도 측면의 두 가지로 분석하려 한다.

4.1 모형 예측력 비교

본 절에서는 모형 예측력을 비교하기 위해 다음과 같은 모형을 생각한다.

$$y_i = \mathbf{x}_i^T\boldsymbol{\beta} + \varepsilon_i, \qquad i = 1,2,...N$$

단, $\boldsymbol{\beta} = (3,1.5,0,0,2,0,0,0,)^T$ 로 설정하고, 잔차의 분포 및 데이터의 개수, 그리고 독립변수들간의 상관관계에 대해서는 다음과 같은 상황을 고려하도록 한다.

◆잔차의 분포 : 표준편차 (2 or 4)의 정규분포 (Normal distribution) 또는 파라미터 (1.5 또는 3)의 라플라스 분포 (Laplace distribution)

◆ 데이터의 개수 : 30 또는 100

◆ 독립변수들간의 상관관계 : $\rho_{ij} = \rho^{|i-j|}$, 단, $\rho = 0.5$ 또는 $\rho = 0.9$. 여기서 ρ_{ij} 는 i 와 j 번째 독립 변수간의 상관 관계를 의미한다.

본 연구에서 잔차의 분포를 정규 분포 혹은 라플라스 분포로 한 것은 라플라스 분포가 Kurtosis가 더 커 잔차가 정규 분포인 경우 보다 이상 데이터를 포함할 가능성이 크기 때문에 이상치를 포함한 경우를 설명하기 위함이다. 참고로 라플라스 분포의 파라미터는 정규분포의 표준편차(2 또는 4)와 유사한 값을 가지도록 설정하였다.

본 연구에서는 위와 같은 설정의 모든 조합(16가지)에 대하여 50번의 반복 실험을 거쳐 다음처럼 실제 회귀계수 $\boldsymbol{\beta}$ 와 각 방법에 의해 추정된 회귀계수 $\hat{\boldsymbol{\beta}}$ 과의 제곱근 평균 제곱 오차 (RMSE)를 기준으로 각 방법을 평가하도록 한다.

$$\mathrm{RMSE} = \sqrt{\frac{1}{50}\sum_{i=1}^{50}(\beta_i - \hat{\beta}_i)^2} \qquad (7)$$

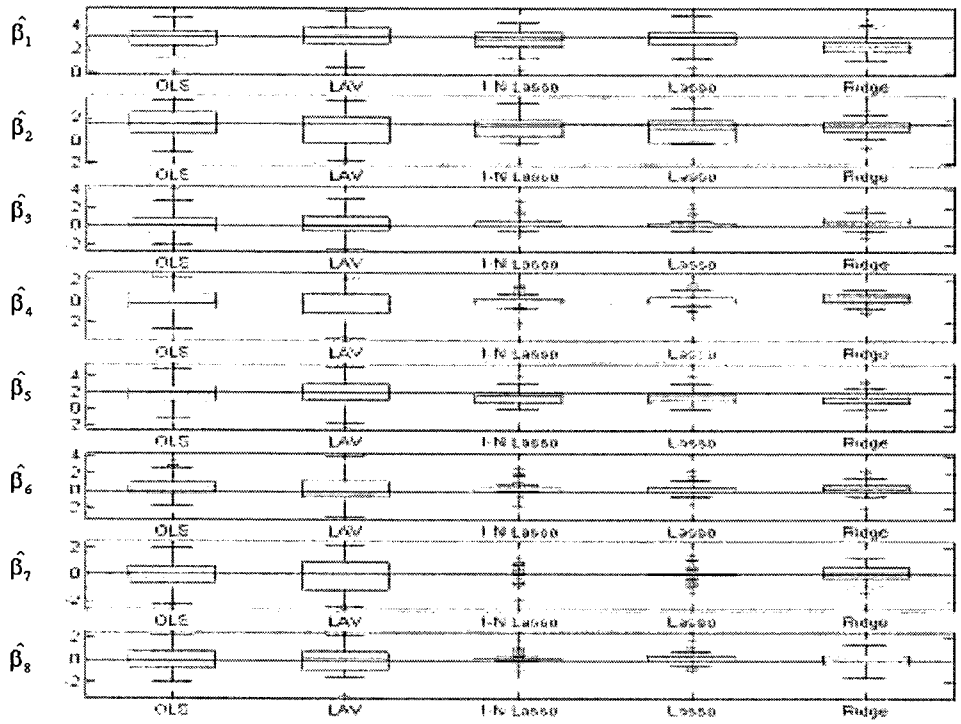
[표 1] 각 실험조건별 RMSE 비교

No	Distribution	Parameter	# of Data	Correlation	LAV	OLS	1 Norm -Lasso	Lasso	Ridge
1	Normal	2	30	0.5	0.654	0.548	0.512	**0.459***	0.468
2	Normal	4	30	0.5	1.245	1.050	0.855	**0.793***	0.837
3	Laplace	1.5	30	0.5	0.541	0.593	**0.416***	0.462	0.466
4	Laplace	3	30	0.5	1.099	1.015	0.759	**0.731***	0.765
5	Normal	2	100	0.5	0.328	0.240	0.233	**0.190***	0.249
6	Normal	4	100	0.5	0.627	0.507	0.470	**0.421***	0.458
7	Laplace	1.5	100	0.5	0.217	0.273	**0.170***	0.205	0.261
8	Laplace	3	100	0.5	0.457	0.508	**0.364***	0.399	0.505
9	Normal	2	30	0.9	1.585	1.291	0.955	0.826	**0.749***
10	Normal	4	30	0.9	3.257	2.610	1.409	1.264	**1.010***
11	Laplace	1.5	30	0.9	1.362	1.300	0.860	0.846	**0.834***
12	Laplace	3	30	0.9	2.794	2.656	1.199	1.206	**1.089***
13	Normal	2	100	0.9	0.748	0.581	0.560	**0.423***	0.506
14	Normal	4	100	0.9	1.308	1.032	0.822	**0.717***	0.743
15	Laplace	1.5	100	0.9	0.546	0.570	**0.356***	0.378	0.551
16	Laplace	3	100	0.9	1.072	1.214	**0.694***	0.749	0.797

*: 각 실험 조건별 최적 추정 방법

일반적으로 잔차의 분포가 라플라스 분포처럼 heavy tail을 가지고, 데이터의 개수가 충분할 경우, LAV의 추정방법이 OLS 보다 안정적이라고 알려져 있다 [1]. 이는 [표 1]의 (7,8,15,16)에 해당하고 이 경우 모두 LAV의 방법이 OLS 방법보다 RMSE가 작음을 확인할 수 있다. 즉, 라플라스 분포처럼 heavy tail을 가진다는 것은 이상치의 존재 가능성이 크다는 말이고, 따라서 LAV의 방법이 보다 로버스트하다는 것을 의미한다. 이와 마찬가지로, (7,8,15,16)의 경우에 있어 1-Norm Lasso의 방법이 Lasso 방법 및 Ridge 방법에 비해 정확한 결과를 보여주고 있음을 확인할 수 있다.

일반적으로 능형 회귀와 같은 분산 축소 기법들은 회귀계수의 추정에 약간의 편의(bias)를 줌으로서 추정된 회귀계수의 분산을 줄여 정확한 예측을 할 수 있도록 하는 기법이며 본 연구에서 제안하는 1-Norm Lasso와 기존의 Lasso 방법은 분산 축소와 동시에 변수 선택이 이루어진다. [그림 2]는 이와 같은 방법들에 의한 실험 16에 대한 회귀계수 추정결과를 나타낸 것으로 50번 반복 실험한 결과를 상자도표(box-plot)으로 표현한 것이다. 우선 8개의 변수에 대해 모두 RR, 1-Norm Lasso, Lasso의 방법이 불편의 추정 방법인 OLS나 LAV방법에 비해 약간의 편의는 조금 크지만 분산이 작은 것을 알 수 있으며, RR과는 달리 1-Norm Lass 및 Lasso의 경우 실제 회귀계수의 값이 0인 (3,4,6,7,8)번 변수에 대해 대부분 회귀계수의 추정치를 0으로 예측하고 있음을 알 수 있다. 즉, 이와 같은 방법은 제한 조건을 통해 불필요한 변수의 회귀계수를 0으로 만듦으로 자체적으로 변수 선택이 이루어지게 함을 확인할 수 있다.



[그림 2] 각 방법에 의한 추정회귀계수 (실험 16)

[표 2]는 1-Norm Lasso와 Lasso의 변수선택 결과를 나타낸 것이다. 1-Norm lasso의 경우 실제 모형에 포함된 3개의 변수를 정확히 선택한 경우는 Lasso와 유사하며 대체적으로 불필요한 변수를 적게 선택하였음을 알 수 있다.

[표 2] 변수 선택 결과

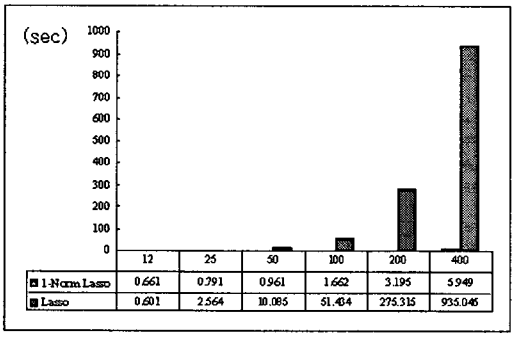
선택된 변수	1-Norm Lasso	Lasso
3개	7	8
4개	17	12
5개	22	25
6개	4	3
7개	0	2
계	50	50

4.2 수행속도 비교

본 절에서는 이차 계획법 문제를 풀어야 하는 Lasso 방법과 본 연구에서 제안하는 1-Norm Lasso의 수행속도를 비교하기 위해 다음과 같은 두가지 실험을 한다.

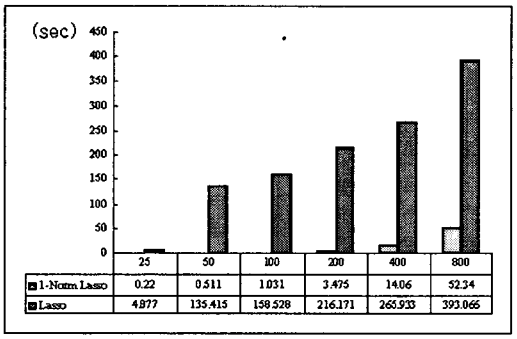
실험 1 : $N = 200, P$=[12, 25, 50, 100, 200, 400];
실험 2: P =2, N = [25 50 100 200 400 800];

실험 1은 데이터의 개수를 200개로 고정시켜 놓고 변수의 개수를 증가시켜 가며, 변수의 증가에 따른 수행속도를 비교하기 위한 것이고, 실험 2는 변수의 개수를 고정시켜 놓고 데이터의 개수를 증가시켜가며, 데이터의 개수에 따른 수행속도를 비교하기 위한 것이다.



[그림 3] 실험 1: 변수 증가에 따른 수행속도

실험 1의 결과를 보여주는 [그림 3]에서 Lasso의 경우는 수행속도가 변수 개수의 증가에 대략 이차 형태로 증가하고 있음을 알 수 있고, 본 연구에서 제안하는 1-Norm lasso의 경우 선형적으로 증가하고 있음을 알 수 있다.



[그림 4] 데이터 개수 증가에 따른 수행 속도

실험 2의 결과를 보여주는 [그림 4]에서는 Lasso의 경우 수행 속도가 1-Norm lasso보다 다소 오래 걸리기는 하지만 비교적 완만하게 증가하고 있음을 알 수 있고, 1-Norm lasso의 경우 이차적으로 증가하고 있는 것으로 보인다. 실험 1과 실험 2을 통해, 서론에서 제시했듯이 1-Norm lasso가 lasso에 비해 일반적으로 빠른 것을 알 수 있고, 특히 변수의 개수가 많을 경우에는 1-Norm lasso가 훨씬 효율적인 것을 알 수 있다. 하지만, 데이터의 개수가 아주 클 경우, 1-Norm lasso의 방법의 경우도 시간이 많이 걸리게 된다는 것을 알 수 있다. 이 경우 식(5)에서 정의된 원 문제를 쌍대문제(dual problem)으로 정의함으로써 수행속도를 단축시킬 수 있을 것으로 기대된다[4].

5. 토의 및 결론

본 연구에서 제안하는 1-Norm Lasso의 방법은 이상치가 존재하는 경우, 타 방법들에 비해 안정적인 회귀계수의 추정이 가능하다는 것을 수치예제를 통해 확인할 수 있었고, 회귀계수를 추정하는 알고리즘의 수행속도도 기존의 Lasso 방법보다 빠르다는 것을 확인할

수 있었다. 특히, 데이터의 특성에 따라 최적의 방법이 다르다는 결론을 얻을 수 있었다. 따라서 잔차 정규 확률도 분석 및 독립변수의 상관관계 분석을 통해 적절한 추정 방법이 선택될 수 있을 것이다.

감사의 글

이 논문은 2003년도 IMT2000 사업 (과제번호: 00015993)에 의하여 지원되었음.

Reference

[1] Brikes, D. and Dodge, Y., *Alternative Methods of Regressions*, John Wiley & Sons, New York, 1993.

[2] Frank, I. E. and Friedman, J. H., "A Statistical View of Some Chemometrics Regression Tools", *Technometrics*, vol. 35, pp. 109-135, 1993.

[3] Hastie *et al.*, *The Elements of Statistical Learning*, Springer, New York, 2001.

[4] Murty, K. G., *Linear Programming*, John Wiley & Sons, New York, 1983.

[5] Neter, *et al.*, *Applied Linear Statistical Models*, IRWIN, New York, 1996.

[6] Tibshirani, R., "Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society*, Series B 58(1), pp. 267-288, 1996.