한국통계학회논문집 제4권 1호
1997년 4월 pp. 1-10

Nonnegative Garrote 에서의 영향관측값 검출[1)]

안 병 진[2)]

요 약

Breiman(1995)에 의하여 제안된 nonnegative garrote을 Mallows C_p의 확장된 개념인 C_H의 관점에서 최소자승법, 능형회귀, 축소추정량, 변수선택등의 방법과 비교하였다. 또한 C_H를 각 관측값이 기여한 양으로 분해하여 nonnegative garrote의 경우 영향관측값을 검출할 수 있는 한가지 방법을 다루었다.

1. 서 론

회귀분석에서 설명변수들간에 다공선성이 존재하면 최소자승법에 의하여 추정되는 회귀계수들은 신뢰할 수 없게 된다. 이러한 문제를 극복하기 위하여 능형회귀(Hoerl 과 Kennard, 1970), 주성분회귀(Lott, 1973)등 다양한 편의 추정량들이 제안되었다. 모형이 많은 설명변수를 포함하고 있는 경우 좀더 단순한 모형을 얻기 위한 변수선택방법에 대하여도 많은 연구가 이루어져 왔다. Breiman(1995)은 자료의 조그마한 변화에도 지나치게 결과가 바뀌게 되는 변수 선택방법보다는 안정적이고 능형회귀보다는 단순한 모형을 유도할 수 있는 nonnegative garrote(NG)를 제안하였다. 이 논문에서는 Mallows(1973) C_p의 확장된 개념인 C_H의 관점에서 NG를 최소자승추정량, 능형회귀, 축소추정량(Stein, 1960), 변수선택등 기존의 방법과 비교 검토하고자 한다.

한편 회귀분석을 위한 자료에서는 다공선성과 영향관측값이 동시에 존재할 수 있으며(Lawrence와 Marsh, 1984) 영향관측값인 경우 소수의 자료가 분석에 커다란 영향을 줄 수 있다. Walker 와 Birch(1988)는 능형회귀에서 영향관측값의 검출방법에 대하여 다루었으며, Leger와 Altman(1993)은 변수 선택을 할때 관측값을 하나씩 교대로 제거시켜 가면서 관측값을 제거한후 선택된 모형과 제거하기 전에 선택된 모형에서의 예측값을 비교해 봄으로써 영향관측값을 검출하는 방법에 대하여 다루었다. 이 논문에서는 NG를 적용할때 영향관측값의 검출방법에 대하여 다루고자 한다.

1) 이 논문은 1996년도 건국대학교 학술진흥연구비에 의해 연구되었음.
2) (143-701) 서울 광진구 모진동 93-1, 건국대학교 응용통계학과 교수.

2. 비교하고자 하는 추정량

다음과 같이 n개의 자료를 갖는 중회귀모형을 고려해 보자.

$$y = X\beta + \varepsilon \tag{2.1}$$

여기서 $y=(y_1, y_2, \cdots, y_n)^T$는 $n\times 1$ 반응변수벡터이며 $X=(X_1, X_2, \cdots, X_k)$는 설명변수값으로 이루어진 $n\times k$ 행렬이고, y와 X는 표준화 되었다. β는 모회귀계수 벡터이고 ε는 $n\times 1$ 벡터이며 $E(\varepsilon)=0$ 이고 $Var(\varepsilon)=\sigma^2 I$ 이다. 이때 최소자승법에 의해서 구한 β의 추정량은

$$\hat{\beta}_{LSE} = (X^T X)^{-1} X^T y \tag{2.2}$$

이므로 각 측정값에서의 추정값벡터 $\hat{y}_{LSE}=(\widehat{y_1}, \widehat{y_2}, \cdots, \widehat{y_n})^T$는

$$\hat{y}_{LSE} = H_{LSE}\ y \tag{2.3}$$

이 된다. 여기서 $H_{LSE}=X(X^T X)^{-1}X^T$이다.

Breiman(1995)에 의하여 제안된 NG추정량은

$$\tilde{\beta}_{Garrote} = B\,\hat{\beta}_{LSE} \tag{2.4}$$

의 형태를 갖는다. 여기서 $B=diag(b_1, b_2, \cdots, b_k)$, $b_i \geq 0$ $i=1,2,\cdots k$ 이다. 이때 대각행렬 B는 $trace(B)\leq s$ 라는 제약조건하에서

$$D=(y-X\tilde{\beta}_{Garrote})^T(y-X\tilde{\beta}_{Garrote}) \tag{2.5}$$

를 최소화 함으로써 구한다. 모수 s값이 작아질수록 더 많은 b_i값이 영이 되어지고 최소자승추정량에 비하여 (2.4)식에 주어진 추정량 벡터의 원소들의 절대값이 작아지게 된다.

모형(2.1)에서 행렬 X를 나누어서 $X=(X_p, X_r)$로 하고 다시 표현하면

$$y = X_p\beta_p + X_r\beta_r + \varepsilon \tag{2.6}$$

이 되는데, X_p는 $n\times p$ 행렬이고 X_r은 $n\times(k-p)$행렬이다. β_p와 β_r도 X_p와 X_r에 따라서 나누어져 있다. 이때 (2.6)식에 주어진 모형에서 X_r을 제거하고 $y=X_p\beta_p+\varepsilon$ 만을 적합시켜 얻은 변수선택에 의한 β_p의 최소자승추정량은 $\tilde{\beta}_{sub} = (X_p^T X_p)^{-1} X_p^T\, y$ 이므로 $H_p = X_p(X_p^T X_p)^{-1}X_p^T$라 하면 추정값벡터는

$$\tilde{y}_{sub} = H_p y \tag{2.7}$$

이 된다. 이때 $Z=(I-H_p)X_r$ 이라 하면 Weisberg(1981)에 의하여

$$\begin{aligned} H_r &= Z(Z^TZ)^{-1}Z^T \\ &= (I-H_p)X_r(X_r^T(I-H_p)X_r)^{-1}X_r^T\ (I-H_p) \end{aligned} \tag{2.8}$$

이 되어 $H_{LSE}=H_p+H_r$이 된다.

모형 (2.1)에 있는 행렬 X는 정칙값분해 (Singular Value Decomposition)에 의하여 다음과 같이 분해되어 질수 있다.

$$X = UDV^T \tag{2.9}$$

여기서 $U^TU=V^TV=I$, $D=diag(\theta_1^{\frac{1}{2}}, \theta_2^{\frac{1}{2}}, \cdots, \theta_k^{\frac{1}{2}})$이며 $\theta_1, \theta_2, \cdots, \theta_k$는 행렬 X^TX의 고유치이다. $n\times k$ 행렬 $U=(u_{ij})$의 열들은 XX^T의 고유벡터이며, $k\times k$ 행렬 V의 열들은 X^TX의 고유벡터이다.

위에서 언급한 추정량이외에도 능형회귀, 축소추정량(Stein, 1960)등을 포함하는 추정량들을 행렬 H에 따라 하나의 추정량군으로

$$\tilde{y} = Hy \tag{2.10}$$

의 형태로 표현할 수 있다. 여기서 $\tilde{y}=(\tilde{y}_1, \tilde{y}_2, \cdots, \tilde{y}_n)^T$은 각 측정값에서의 추정값벡터이며, $H=(h_{ij})$ $i,j=1,2,\cdots,n$이다. 비교하고자 하는 추정량들을 행렬 H에 따라 정칙값분해를 이용하여 표현하면 표 2.1과 같다. 표 2.1의 k^*와 c는 능형회귀와 축소추정량에서의 축소모수이다.

표 2.1 비교하고자 하는 추정량들의 행렬 H

추 정 량	행 렬 H
최소자승법	$X(X^TX)^{-1}X^T = UU^T$
능형회귀	$X(X^TX+k^*I)^{-1}X^T = UD(D^2+k^*I)^{-1}DU^T$
축소추정량	$cX(X^TX)^{-1}X^T = cUU^T$, $0<c\leq 1$
변수선택	$X_p(X_p^TX_p)^{-1}X_p^T = UU^T-H_r$
NG	$XB(X^TX)^{-1}X^T = UBU^T$

비교하고자 하는 추정량들의 i-번째 관측값에서의 추정된 값 $\tilde{y}_i$는 다음과 같이

$$\tilde{y}_i = \sum_{j=1}^{n} h_{ij}\, y_j \tag{2.10}$$

이며, $\partial \tilde{y}_i / \partial y_i = h_{ii}$ 이므로 행렬 H의 대각원소는 최소자승법에서와 마찬가지로 다른 추정량의 경우에도 중요한 역할을 한다. 각 추정량의 경우 행렬 H의 대각원소와 $\hat{y}_i$ 및 $\tilde{y}_i$의 분산을 정리해 보면 표2.2 와 같다.

표 2.2에서 h_i^*는 (2.3)식에 있는 행렬 H_{LSE}의 i-번째 대각원소이며 $h_{r,i}$는 (2.8)식에 있는 행렬 H_r의 대각원소값이다. 편의추정량들의 경우 행렬 H의 대각원소값과 $Var(\tilde{y}_i)$는 최소자승법에 비하여 작아지며 축소모수값에 따라 변화함을 알 수 있다. 표2.1에 있는 행렬 H와 (2.3)식의 H_{LSE}의 곱은 $HH_{LSE} = H$이므로

$$Cov(\hat{y}_i, \tilde{y}_i) = h_{ii}\sigma^2 \tag{2.11}$$

이 된다.

표 2.2 행렬 H의 대각원소와 $Var(\hat{y}_i)$ 및 $Var(\tilde{y}_i)$

추 정 량	h_{ii}	$Var(\hat{y}_i)/\sigma^2$ 및 $Var(\tilde{y}_i)/\sigma^2$
최소자승법	$\sum_j u_{ij}^2 = h_i^*$	$\sum_j u_{ij}^2$
능형회귀	$\sum_j (\frac{\theta_j}{\theta_j + k^*}) u_{ij}^2$	$\sum_j (\frac{\theta_j}{\theta_j + k^*})^2 u_{ij}^2$
축소추정량	$c\sum_j u_{ij}^2$	$c^2 \sum_j u_{ij}^2$
변수선택	$h_i^* - h_{r,i}$	$h_i^* - h_{r,i}$
NG	$\sum_j b_j u_{ij}^2$	$\sum_j b_j^2 u_{ij}^2$

3. 판단기준 C_H의 분해

회귀분석의 결과는 소수의 측정값에 의하여 크게 영향받을 수 있다. 이러한 영향관측값을 검출할 때 흔히 쓰이는 방법이 자료를 교대로 제거해 가면서 관심을 갖고 있는 통계량의 변화를 살펴보는 방법이다. NG와 같은 편의추정량의 경우에는 최소자승법과는 달리 자료를 제거할 때마다 전체를 다시 계산하여야 하므로 계산비용이 많이 발생하게 된다. 예를 들어 NG의 경우 i-번째 관측값을 제거했을 때 회귀계수의 추정값을 구하기 위해서는 주어진 s값에 따라 (2.5)식을 최소화하는 행렬 B를 매번 구해야 하므로 계산비용이 많이 발생하게 된다.

Weisberg(1981)는 변수선택의 판단기준으로 널리 사용되어지는 Mallows C_p통계량을 각각의 관측값이 기여한 양으로 분해함으로써 변수선택에서의 영향관측값을 검출하고자 하였다. 같은 방법을 NG에 적용하여 Mallows C_p의 확장된 개념인 C_H를 각각의 관측값이 기여한 양으로 분해함으로써 축소모수를 결정하는 하나의 판단기준인 C_H에 영향을 주는 관측값을 찾는 방법을 다루고자 한다.

추정된 값 $\tilde{y}$가 $E(y)$에 얼마나 잘 적합한가를 알아보기 위하여 $\tilde{y}$의 평균제곱오차 (mean squares error)를 구해보면

$$\begin{aligned} MSE(\tilde{y}) &= E(\tilde{y}-X\beta)^T(\tilde{y}-X\beta) \\ &= trace(H^2)\sigma^2+\beta^T X^T(I-H)^2 X\beta \end{aligned} \tag{3.1}$$

이 된다.

RSS_{LSE}를 최소자승법의 경우 잔차자승합이라 하고 RSS를 표 2.1에 있는 추정량들의 잔차자승합이라 하면

$$\begin{aligned} RSS &= (y-Hy)^T(y-Hy) \\ &= RSS_{LSE}+y^T(H^2-2H+H_{LSE})y \\ &= RSS_{LSE}+\sum_{i=1}^{n}(\tilde{y}_i-\hat{y}_i)^2 \end{aligned} \tag{3.2}$$

이 되며, (3.2)식에 기대값을 취하면

$$E(RSS) = trace(I-H)^2\sigma^2+\beta^T X^T(I-H)^2 X\beta \tag{3.3}$$

이 된다. (3.3)식을 (3.1)식에 대입하면

$$MSE(\tilde{y}) = E(RSS)-n\sigma^2+2trace(H)\sigma^2 \tag{3.4}$$

을 얻는다. 이때 σ^2대신 $\hat{\sigma}^2 = y^T(I-H_{LSE})y/(n-k)$를 대입하고 $E(RSS)$의 자리에 RSS를 대입하면 $MSE(\tilde{y})/\sigma^2$의 추정량 C_H가 다음과 같이 주어진다.

$$C_H = \frac{RSS}{\hat{\sigma}^2} - n + 2trace(H) \tag{3.5}$$

표2.1에 있는 추정량들의 축소모수는 C_H를 최소화함으로써 구할 수 있다. 변수선택의 경우에는 (3.5)식의 C_H는 Mallows(1973) C_p와 동일하며 능형회귀의 경우는 Mallows(1973) C_L과 동일하다. 축소추정량의 경우에는 C_H를 최소화하여 구한 c가 Dempster외 2인(1977)의 STEINM이라고 불리워진 통계량과 동일하다. NG의 경우 각 s값에 따라 (2.5)식을 통하여 구한 행렬 B를 이용하여 C_H를 최소화하는 s와 행렬 B를 구할 수 있다.

Weisberg(1981)의 방법을 NG에 적용하기 위하여 (3.5)식에 있는 C_H를 각 측정값이 기여한 양으로 분해해 보기로 하자. i-번째 관측값에서의 추정된 값 $\tilde{y}_i$의 평균제곱오차는

$$MSE(\tilde{y}_i) = Var(\tilde{y}_i) + [E(\tilde{y}_i) - E(y_i)]^2 \tag{3.6}$$

이 된다. 한편 (2.1)식의 모형에서 $E(\hat{y}_i) = E(y_i) \quad i = 1, 2, \cdots, n$ 이므로

$$\begin{aligned} E(\tilde{y}_i - \hat{y}_i)^2 &= Var(\tilde{y}_i - \hat{y}_i) + [E(\tilde{y}_i) - E(\hat{y}_i)]^2 \\ &= Var(\tilde{y}_i) + Var(\hat{y}_i) - 2Cov(\tilde{y}_i, \hat{y}_i) + [E(\tilde{y}_i) - E(y_i)]^2 \end{aligned} \tag{3.7}$$

이다. (3.7)식을 (3.6)식에 대입하고, (2.11)식을 이용하면

$$MSE(\tilde{y}_i) = E(\tilde{y}_i - \hat{y}_i)^2 - h_i^* \sigma^2 + 2h_{ii}\sigma^2 \tag{3.8}$$

이 얻어진다. 이때 σ^2대신 $\hat{\sigma}^2$를 대입하고 $E(\tilde{y}_i - \hat{y}_i)^2$ 대신에 $(\tilde{y}_i - \hat{y}_i)^2$을 대입하면 $MSE(\tilde{y}_i)/\sigma^2$의 추정량 $C_{H,i}$가

$$C_{H,i} = \frac{(\tilde{y}_i - \hat{y}_i)^2}{\hat{\sigma}^2} - h_i^* + 2h_{ii} \tag{3.9}$$

로 주어진다. (3.9)식에 주어진 $C_{H,i}$는 변수선택의 경우는 Weisberg(1981)에 의한 $C_{p,i}$와 동일하고 최소자승법의 경우는 $C_{H,i} = h_i^*$가 된다. (3.2)식에 의하여 $\sum(\tilde{y}_i - \hat{y}_i)^2 = RSS - RSS_{LSE}$ 이고 $\sum_i h_i^* = k$, $\sum_i h_{ii} = trace(H)$ 이므로 $\sum C_{H,i} = C_H$가 된다. 따라서 C_H는 각 측정값이 C_H를

결정하는데 기여한 값 $C_{H,i}$의 합으로 분해되었음을 알 수 있다. NG의 경우에도 C_H를 $C_{H,i}$로 분해하여 어떤 $C_{H,i}$값이 다른 값들에 비하여 차이가 많이 나면 i-번째 관측값이 C_H를 결정하는데 많은 영향을 주었다고 볼 수 있다.

4. 적용사례

이절에서는 Walker 와 Birch(1988)의 논문에서 인용된 자료를 이용하여 앞절에서 논의된 내용을 적용해보고자 한다. 자료는 표 4.1에 주어져 있고 이 자료를 표준화한 후 여러 추정방법에 적용하였다.

표 4.1 인용된 자료

Case	X_1	X_2	X_3	X_4	X_5	X_6	Y
1	57.0	6.40	12	293.2	41.1	45.0	61.1
2	53.0	5.00	12	354.3	51.0	32.0	62.3
3	50.3	5.75	14	293.5	24.9	29.4	59.4
4	41.2	4.50	13	299.0	19.4	20.3	66.2
5	36.7	5.15	13	286.0	18.6	17.4	66.0
6	35.5	4.25	10	254.8	17.1	14.9	71.4
7	26.4	3.35	10	270.4	17.6	14.5	75.4
8	25.0	2.50	9	239.2	13.6	13.2	83.2
9	23.5	3.45	11	270.5	14.3	11.7	73.2
10	26.7	6.00	11	298.0	12.9	10.4	71.1
11	25.8	5.70	11	247.0	11.9	15.2	72.8
12	25.7	6.75	12	260.1	12.5	19.5	75.6
13	27.0	4.95	12	228.8	10.5	18.6	76.0
14	24.5	3.65	12	179.4	8.3	19.1	70.2
15	23.1	4.05	11	176.8	8.5	15.9	68.6

이 자료에 NG를 적용한 경우 모수 s값에 따라 (2.5)식을 최소화하는 행렬 B를 구한후 (3.5)식에 의하여 C_H값을 구한 결과 그림 4.1과 같으며, s=4.8일 때 C_H=4.822가 되어 가장 작아진다. s=6일 때는 최소자승법과 동일하게 C_H=6이 됨을 확인할 수 있다.

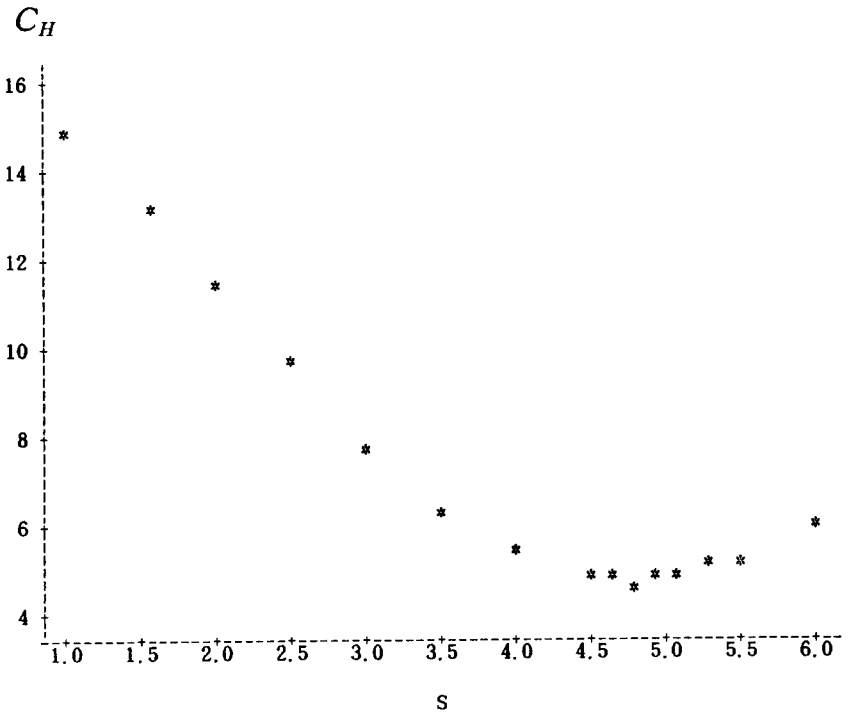

그림 4.1 s값에 따른 C_H값

다른 추정량의 경우도 C_H를 최소화함으로써 축소모수를 구할 수 있고, 이때 행렬 H의 대각원소값과 $C_{H,i}$ 값중에서 크기가 큰 순서로 3개씩만 정리해보면 표 4.2와 같다. 괄호안에 있는 값이 $C_{H,i}$값이다. 최소자승법의 경우는 행렬 H의 대각원소값과 $C_{H,i}$값이 동일하다.

$C_{H,i}$와 h_{ii}값들을 살펴보면 C_H통계량을 결정하는데 각각의 관측값이 기여하는 바가 다르다는 것을 알 수 있다. 예를 들면 NG의 경우 $C_{H,2}$=0.564 인데 반하여 $C_{H,13}$ =0.048이다. 따라서 NG의 경우에도 축소모수가 소수의 자료에 의하여 결정될 수 있다. 표 4.2에서 보는 바와 같이 영향관측값과 $C_{H,i}$의 변화량($h_i^* - C_{H,i}$)도 추정방법에 따라 달라진다. 능형회귀에서의 영향관측값 검출을 다룬 Walker와 Birch(1988)의 논문에서는 8번, 1번, 2번 관측값을 영향관측값으로 보고 있는데 C_H관점에서 보면 능형회귀에서는 2번, 1번, 6번 관측값을, NG에서는 2번, 8번, 1번 관측값을 잠재적 영향관측값으로 판단하게 된다.

표 4.2 행렬 H의 대각원소와 $C_{H,i}$

Case	최소자승법	능형회귀	축소추정량	변수선택	NG
1	0.770 (0.770)	0.761 (0.754)	0.689 (0.671)	0.345	0.652 (0.537)
2	0.855 (0.855)	0.834 (0.817)	0.765 (0.709)		0.663 (0.564)
3				0.275	
⋮	⋮	⋮	⋮	⋮	⋮
6	0.584 (0.584)	0.561 (0.544)	0.522 (0.461)	0.165 (0.321)	0.437
7				0.099 (0.344)	
8				0.281 (0.349)	0.351 (0.543)
⋮	⋮	⋮	⋮	⋮	⋮
C_H 축소모수	6.0	5.802 k^*= 0.031	5.373 c = 0.895	2.396 p=2	4.822 s=4.8

5. 결 론

Breiman(1995)에 의하여 제안된 NG를 Mallows C_p의 확장된 개념인 C_H의 관점에서 최소자승법, 능형회귀, 축소추정량, 변수선택등의 방법과 비교하였다. Mallows C_p는 Breiman(1992)에 의하여 지적된 바와 같이 예측오차(prediction error) 를 추정하는데 문제가 있을 수 있으므로 little bootstrap, cross-validation 등의 방법으로 축소모수를 결정하여, 모의 실험을 통하여 추정량들을 비교하는 것이 C_H의 관점에서의 단순비교보다는 훨씬 바람직할 것이다.

NG를 사용할 경우 영향관측값의 검출을 위해서 자료를 교대로 제거하면서 관심있는 통계량의 변화를 살피는 방법은 계산비용 때문에 고려하지 않았고 C_H를 최소화하도록 축소모수를 결정한 후 C_H를 각 관측값이 기여한 양으로 분해하여 $C_{H,i}$값이 다른 것에 비하여 차이가 많이 나는 것을 영향관측값이라고 보았다. 이 방법은 계산비용이 적게 들면서도 간단하다는 장점은 있지만 축소모수 역시 영향관측값에 영향을 받아 결정되어지므로 축소모수를 고정한 후 C_H를 분해하는 것은 문제가 발생할 수 있다. NG를 사용할 때 축소모수 s를 결정하기 위하여 C_H의 최소화가 아닌 다른 방법을 사용할 때의 영향관측값 검출에 대한 연구도 필요한 것으로 생각된다.

참고 문헌

[1] Breiman, L. (1995). Better subset regression using the nonnegative garrote, *Technometrics*, Vol. 37, 373-384.

[2] Breiman, L. (1992). The little boostrap and other methods for dimensionality selection in regression:The X-fixed prediction error, *Journal of the American Statistical Association*, Vol. 87, 738-754.

[3] Dempster, A.P., Schatzoff, M. and Wermuth, N. (1977). A simulation study of alternatives to ordinary least square, *Journal of the American Statistical Association*, Vol. 72, 77-91.

[4] Hoerl, A.E. and Kennard, R.W. (1970). Ridge regresion: biased estimation for nonorthogonal problems , *Technometrics*, Vol. 12, 55-67.

[5] Leger, C. and Altman, N. (1993). Assessing influence in variable seletion problems, *Journal of the American Statistical Association*, Vol. 88, 547-556.

[6] Lawrence, K.D. and Marsh, L.C. (1984). Robust ridge regression methods for predicting U.S. coal mining fatalities, *Communications in statistics*, Vol. 13, 139-149.

[7] Lott, W.F. (1973). The optimal set of principal component restrictions on a least squares regression, *Communications in Statistics*, Vol. 2, 449-464.

[8] Mallows, C.L (1973). Some comments on C_p, *Technometrics*, Vol. 15, 661-675.

[9] Stein, C.M. (1960). Multiple regression. Contribution to probability and ststistics. Essays in honor of Harold Hotelling, I.(ed.), Olkin, Stanford Univ. Press, 424-443.

[10] Walker, E. and Birch, J.B. (1988). Influence measures in ridge regression, *Technometrics*, Vol. 30, 221-227.

[11] Weisberg, S. (1981). A statistic for allocating C_p to individual cases, *Technometrics*, Vol. 23, 27-31.