Journal of the Korean Data Analysis Society (October 2015)
Vol. 17, No. 5 (B), pp. 2391-2402

고차원 선형모형에서 벌점화우도 방법을 이용한 변수 선택방법 연구*

주아림[1], 전수영[2]

요 약

선형회귀모형에서 표본 크기 n이 공변량의 수 p보다 큰 경우 여러 가지 변수 선택방법 중 일반적으로 전진적 선택방법, 후진적 제거방법, 단계적 선택방법 등이 사용된다. 하지만, p가 n에 비해 매우 큰 경우인 고차원문제를 가지는 선형모형에서는 이러한 방법들을 사용하는 데에 어려움이 있다. 이러한 변수선택의 어려운 점을 극복하기 위해 유의하지 않은 변수들을 제거하는 다양한 regularization 또는 shrinkage 방법이라고도 불리는 벌점화우도(penalized likelihood) 방법들이 고차원 선형모형에서 제안되었다. 본 논문은 고차원 선형모형에서 사용이 용이한 변수 선택방법에 대한 연구로, 여러 가지 벌점화우도 방법들에 대해 알아보고 그 응용성을 알아본다. 그 중에서 ridge, LASSO, LARS, elastic net, adaptive LASSO 등 다섯 가지 벌점화우도 방법의 특징을 간단히 살펴보고, 실제 자료인 전립선(prostate) 자료와 한국 경제 수치자료에 적용해 본다. RSS, AIC, BIC를 평가기준으로 다섯 가지 벌점화우도 방법의 실 자료 분석 결과 고차원 선형모형에서 가장 적합한 방법은 LARS 알고리즘이었다.

주요용어 : 변수선택방법, 고차원, 능형, LASSO, LARS.

1. 서론

선형모형은 반응변수에 독립변수들이 어떤 영향력을 미치는지에 대해 조사하는 통계 모형으로 널리 사용된다. 실제 자료를 사용해보면 반응변수에 영향을 미치는 설명변수는 굉장히 작고, 대부분의 설명변수들은 영향을 작게 미치거나 아예 미치지 않는다. 따라서 이러한 유의하지 않는 변수들을 모형에 포함시키지 않도록 하기 위해 계수들을 0으로 만들어 축소된 모형을 사용할 수 있다. 선형모형분석에서 관련 없는 변수를 선택한 모형은 잘못된 통계적 추론을 가져다 줄 수 있으므로, 설명력이 높은 유의한 변수를 정확하게 찾는 것은 매우 중요한 문제이다. 이러한 변수들을 더 많이 포함하는 모형은 반응변수를 확률적으로 더 많이 설명할 수 있게 된다. 반면에 과적합(overfitting)으로 추정된 모형은 영향력이 없는 변수를 다수 포함하게 되어 예측력이 낮아지고 결과의 신뢰성이 떨어지게 된다.

선형회귀모형에서 공변량의 수 p가 표본 크기 n에 비해 매우 클 경우 이를 고차원문제라고 한다. 고차원에서의 변수선택 문제가 생물정보학, 경제학, 약학, 유전학 등의 많은 과학 분야에서 떠오르고 있다. 고차원의 특징은 모형이 매우 복잡하며 추정, 예측력, 설명력을 구하는 것이 상당히 어렵기 때문에, 고전적인 통계방법은 계산적으로 실행이 불가능하고 모형이 정확히 추정되지 않는 문제를 가지고 있다(Mallick, Yi, 2013).

---

*이 논문은 제1저자 주아림(2015)의 석사학위논문의 일부를 발췌·수정한 것입니다.

[1]339-700 세종특별자치시 세종로 2511, 고려대학교 과학기술대학 정보통계학과 석사.
E-mail : jooalim@korea.ac.kr
[2](교신저자) 339-700 세종특별자치시 세종로 2511, 고려대학교 과학기술대학 응용통계학과 부교수.
E-mail : scheon@korea.ac.kr
[접수 2015년 9월 20일; 수정 2015년 10월 17일; 게재확정 2015년 10월 20일]

고차원 선형모형에서 변수선택의 어려운 점을 극복하기 위해 유의하지 않은 변수의 계수들을 0에 가까운 수로 줄여주어 모형에서 제거하는 다양한 벌점화우도(penalized likelihood) 방법들이 제안되었다. Hoerl, Kennard(1970)에 의해 제안된 능형(ridge) 방법은 계수를 0 가까이로는 만들지만 정확히 0으로 줄이지는 못한다. 하지만, Tibshirani(1996)에 의해 제안된 least absolute shrinkage and selection operator (LASSO) 방법은 계수를 정확히 0으로 만들어 계수의 수를 줄여 줄 수 있다. Efron et al.(2004)에 의해 제안된 least angle regression(LARS)은 전진적 선택방법을 조금 변형한 것이며 LASSO와 비슷한 결과 값을 가지지만 속도는 훨씬 빠르다. Zou, Hastie(2005)에 의해 제안된 elastic net과 Zou(2006)에 의해 제안된 adaptive LASSO는 약간의 차이는 있지만 LASSO를 활용하여 만들어진 LASSO의 응용버전이라 할 수 있다. 다중공선성이 강할 때와 그룹화 된 자료에서 elastic net과 adaptive LASSO는 결과가 좋게 나오는데, adaptive LASSO의 경우 각 계수마다 가중치를 따로 구해줘야 하므로 추정치를 구하는데 걸리는 시간이 다소 길다는 단점이 있다. 이와 관련하여 Jung(2007), Sun et al.(2007), Oh, Lee(2013), Lee, Kwon(2015) 등의 연구가 있다.

본 연구에서는 실제자료를 이용하여 고차원 선형모형에서의 다양한 변수선택방법들을 비교해 본다. 첫 번째 분석에 사용된 자료는 R에 내장되어 있으며 전립선암과 특이항원의 관계에 대한 전립선(prostate)자료이다. 여러 의학적 수치가 기술되어 있으며 8개의 설명변수와 1개의 반응변수로 가장 영향력이 큰 변수들을 알아보고 영향력이 거의 없는 변수는 어떤 것들이 있는지 알아본다. 두 번째 자료는 두 개의 경제지표자료로, 먼저 제조업의 여러 수치들을 설명변수로 두고 제조업 업황실적 BSI 수치를 반응변수로 둔 자료를 고려하며, 다음으로 요즘 가장 경제적으로 이슈가 되고 있는 소비자 물가지수를 반응변수로 두고 설명변수로 한국 대표 경제지표 중에서 중요한 지표를 임의로 몇 개를 선택한 자료를 고려하였다. 제조업 업황실적 BSI 및 소비자 물가지수가 어떤 경제지표에 영향을 크게 받는지 알아보고, 어떤 방법이 가장 결과가 좋은지 알아본다. 모두 small-n-large-p 형태로 설명변수의 개수가 반응변수의 개수보다 크다. 본 논문에서는 각각의 방법을 비교하고 여러 가지 자료에 따라 결과 값이 어떻게 달라질지 RSS, AIC, BIC를 이용하여 알아본다.

2장에서는 고전적인 모형 선택방법인 전진적 선택방법, 후진적 제거방법, 단계적 선택방법에 대해 알아보고, 본 논문에서 사용하게 될 모형 선택의 기준 3가지 RSS, AIC, BIC에 대해 알아본 다. 3장에서는 고차원이 어떤 것인지, 특징은 무엇인지에 대해 서술하고, 고차원 문제에서 사용할 수 있는 벌점화우도 방법인 ridge, LASSO, LARS, elastic net, adaptive LASSO에 대해 간단히 요약한다. 4장에서는 실증분석으로 전립선(prostate)자료와 최근의 경제 지표들을 이용하여 각 벌점화우도 방법들을 비교한다.

2. 고전적 모형 선택방법

주어진 자료에서 유의한 변수를 잘 선택해 모형을 만들어야 잘 적합된 모형이라 할 수 있다. 여러 모형 선택방법 중 오래 전부터 널리 사용되던 고전적 모형 선택방법으로는 전진적 선택방법, 후진적 제거방법, 단계적 선택방법 등이 있다. 이러한 방법들은 각 단계에서 예측변수를 더하거나 제거하며 그 결과에 따라 모형을 선택한다. 이중 후진 제거방법은 중요한 변수가 모형에서 제외될 가능성이 적으므로 비교적 안전한 방법이라 할 수 있으나 한번 제외된 변수는 다시 선택되지 못한다는 단점이 있다. 또한 단계적 선택방법은 전진적 선택방법과 후진 제거방법의 단점을 보완한 복합적인 방법이라 할 수 있다(Kim et al., 2012).

예측을 위한 모형 선택의 우선적인 기준은 의미있고(meaningful), 설명력있고(interpretable), 간결(parsimonious)해야 한다는 것이다. 하지만, 위의 3가지 고전적 모형 선택방법은 하나이상의 모형 선

택 기준에 미치지 못하는 것으로 알려져 있다. 더구나 고차원 선형모형에서는 위의 방법들을 사용할 수 없는 경우가 많다(Mallick, Yi, 2013). 본 연구에서는 3장에서 이러한 단점을 극복하기 위해 벌점화우도 방법에 대해 알아볼 것이다.

본 논문에서는 모형을 선택하는 여러 기준 중에서 RSS(residual sum of squares), AIC(Akaike information criteria), BIC(Bayes information criteria) 세 가지 기준을 두고 모형을 선택한다. RSS는 잔차제곱합으로 관측치와 추정치와의 편차 제곱합이며, RSS 값이 작을수록 잔차의 값이 작아진다(Kim, 2000). Akaike(1974)에 의해 제안된 AIC는 적합도와 간결성 사이의 상충을 잘 조절하려 한 모형 선택 기준으로 값이 작을수록 선호된다. AIC는 결측치를 제거하고 계산되기 때문에 많은 결측값이 발생하게 되면 효율적이지 못하다. Schwarz(1978)에 의해 제안된 BIC(베이즈 정보기준)는 AIC의 과적합되는(큰 p선호) 경향을 수정한 방법이다(Kim et al., 2012).

3. 고차원 선형모형에서의 모형 선택방법

우선 반응변수 y는 $n\times 1$벡터, 설명변수 X는 $n\times p$벡터, 모수 β는 $p\times 1$벡터, 오차항 ϵ는 $n\times 1$벡터로 $\epsilon_i \sim i.i.d.\ N(0,\sigma^2)$인 다음과 같은 선형회귀모형을 고려한다.

$$y = X\beta + \epsilon.$$

일반적으로 예측력이 낮은 추정모형은 사용하지 않는 것이 좋다. 전형적으로 모형은 정확하고 간결하며 설명력이 좋아야 한다. 특히 설명변수가 많을수록 간결성(parsimony)이 강조된다. 선형회귀모형에서 설명변수의 개수인 p가 반응변수 n보다 훨씬 클 때, 이것을 고차원(high dimensional) 문제라 일컫는다. 일반적으로 OLS는 예측력과 설명력 모두 좋지 않다. 특히 고차원 선형회귀의 경우 설명력이 떨어지며 다중공선성이 있을 경우 불안정한 추정치를 제공한다. 고차원 문제는 고전적인 모형 선택방법들이 고차원 선형 모형에서 필요로 하는 변수들보다 많은 변수를 선택하는 과적합 문제를 일으키는 경향을 일반적으로 말한다. 과적합 문제는 고차원 모형에서 심각한 편향을 발생시킬 수 있는데, 이를 피하고 더 좋은 예측력을 가지는 다양한 모형 선택방법들을 고려해야 한다(Zou, Hastie, 2005).

3.1. 능형회귀(Ridge regression)

능형회귀는 예측변수들이 상당히 다중공선적일 때 사용될 수 있는 추정법이다. 능형회귀 방법을 사용하여 추정한 회귀계수의 추정량들은 편향되어 있으나 일반적으로 OLS 추정량보다 더 작은 최소제곱오차를 가지는 경향이 있는 것으로 알려져 있다(Hoerl, Kennard, 1970). 능형회귀의 목표함수는 $Q(\beta) = (y-X\beta)^T(y-X\beta) + \lambda\sum_{j=1}^{p}\beta_j$로, 회귀계수에 대한 능형추정량은 다음과 같이 약간 변형된 형태의 정규방정식을 통하여 얻을 수 있다.

$$\begin{array}{rcrcrcrcl}
(1+k)\beta_1 & + & r_{12}\ \beta_2 & + & \cdots & + & r_{1p}\ \beta_p & = & r_{1y} \\
r_{21}\ \beta_1 & + & (1+k)\beta_2 & + & \cdots & + & r_{2p}\ \beta_p & = & r_{2y} \\
\vdots & & \vdots & & \vdots & & \vdots & & \\
r_{p1}\ \beta_1 & + & r_{p2}\ \beta_2 & + & \cdots & + & (1+k)\ \beta_p & = & r_{py}
\end{array}$$

여기서 $r_{ij}(i,j=1,\cdots,p)$는 i번째 예측변수와 j번째 예측변수 사이의 상관계수이며 r_{iy}는 i번째 예측변수와 반응변수 y사이의 상관계수이다. 위 식의 해, $\hat{\beta}_1, \cdots, \hat{\beta}_p$는 능형회귀계수의 추정치가 된다.

능형회귀가 OLS와 다른 점은 k에 있다. $k=0$이면 $\hat{\beta}$은 OLS 추정치가 된다. 이 때 모수 k를 편향(bias)모수 혹은 능형(ridge)모수, 조절(tuning)모수라 부르며 k가 0으로부터 증가하면 추정치의 편의(편향)도 증가하게 된다. 반면, 전체 분산은 다음과 같이 k의 감소함수가 된다.

$$Total\ Variance(k) = \sum_{j=1}^{p} Var(\hat{\beta}_j(k)) = \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^2}.$$

이는 회귀계수에 대한 OLS 추정량의 전체분산에서 작은 값을 갖는 고유치 λ_j의 영향을 보여주고 있다. k를 무한히 계속 증가시키면 회귀계수 추정치는 모두 0으로 접근하는 경향이 있다. 능형회귀의 아이디어는 편향이 크게 증가시키지 않으면서 전체 분산을 감소시키는 적절한 k를 찾는 것이다. Hoerl, Kennard(1970)는 데이터의 작은 변화에 대하여 능형추정치가 안정적인 값을 취하는 적절한 양수 k가 존재함을 보였다. Kim et al.(2012)은 [0,1] 사이의 범위에 있는 k에 대하여 먼저 능형추정치를 계산하고, 그 결과들을 k에 대해 그래프를 그려 추정치의 안정성 관점에서 적절한 k 값을 취할 수 있음을 보여 주었다.

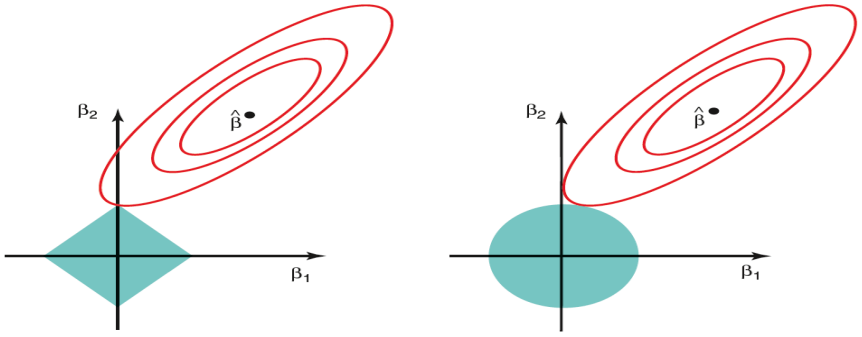
하지만, 능형회귀는 OLS보다 예측 수행은 좋으나 간명한(parsimonious) 모형을 찾지 못하고 항상 모든 모형에 예측변수를 포함하는 단점이 있다.

3.2. LASSO(Least Absolute Shrinkage and Selection Operator)

LASSO는 Tibshirani(1996)에 의해 제안된 기법이다. LASSO 추정치는 다음과 같이 정의된다.

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{p} (y_i - \sum_j \beta_j x_i)^2.$$

여기서 $\sum_j |\beta_j| \le t$를 따른다. $t(\ge 0)$는 조절모수이고, 모수 t를 LASSO 추정치에 적용시켜 shrinkage의 양을 조절한다. $\hat{\beta}_j^o$을 완전 최소제곱추정치로 두고 $t_0 = \sum |\widehat{\beta_j^0}|$로 두면, $t(\le t_0)$의 값은 0으로 가며 shrinkage가 일어난다. 이때 p개의 계수들 중에서 일부는 정확하게 0의 값을 갖는다. 예를 들어 $t = t_0/2$라면, 효과는 $p/2$ 크기의 최적의 계수들(best subset)을 찾는 것과 유사하게 된다. LASSO의 목표함수는 다음과 같이 능형회귀에서 β에 절대값을 적용한 것과 같다.



(a) LASSO (b) Ridge

Figure 1. A graphical illustration of properties of two different penalty functions. The eclipses represent contours of the objective functions. The square and round shapes represent the lasso and ridge constraint, respectively (James et al., 2013).

$$Q(\beta) = (y - X\beta)^t (y - X\beta) + \lambda \sum_{j=1}^{p} |\beta_j|.$$

Figure 1은 p가 2인 경우인데, 타원형 등고선의 중심은 OLS 추정치이며, 제약(constraint)지역은 마름모꼴로 나타나 있다. 이 LASSO의 해는 마름모와 먼저 만나는 등고선 위치에 있고, 이것이 코너에서 만나면 계수가 0이 되기 때문에 LASSO 추정치가 0으로 갈 수 있다. 이와 반대로 능형회귀는 원의 형태로 코너에서 만날 수 없기 때문에 0의 계수를 가지는 상황은 발생하지 않는다.

하지만, LASSO 방법은 여러 장점에도 불구하고 표본크기보다 더 많은 예측변수를 선택할 수 없으며, 변수들 간에 그룹구조를 가지고 있을시 그룹 내에 오직 하나의 변수만을 선택하며, 예측변수끼리 상관성이 높을 때 추정치의 안정성이 좋지 않은 단점이 있다.

3.3. Adaptive LASSO

LASSO는 하나의 조절모수를 사용함으로써 고차원에서 부적절한 모수나 over-shrinkage 모수를 포함할 수 있기 때문에, Zou(2006)는 계수마다 다른 가중치를 주는 adaptive LASSO를 제안하였다. 알려진 가중치 벡터 조건 w_j를 이용한 adaptive LASSO는 다음과 같다.

$$\arg\min_{\beta} \|\boldsymbol{y} - \sum_{j=1}^{p} \boldsymbol{x_j}\beta_j\|^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j|.$$

LASSO는 가중치를 통해 중요한 변수의 shrinkage를 낮게 주고, 중요하지 않은 변수의 shrinkage는 크게 주는 방법이다. 하지만, 이 방법은 많은 조절모수들을 선택해서 사용해야함으로 모수 추정에 있어 컴퓨팅 속도가 매우 길다는 단점이 있다.

3.4. Elastic net

Zou, Hastie(2005)는 변수들간에 그룹화가 형성이 될 때 문제가 있는 LASSO방법을 개선시키기 위해 naive elastic net과 이를 보완한 elastic net을 제안했다. 즉, LASSO는 $p > n$인 경우 over shrinkage하는 경향이 있고, 변수의 그룹 내 상관관계가 큰 경우 그룹으로부터 오직 하나의 변수만을 뽑아내는 단점이 있어 Zou, Hastie(2005)는 naive elastic net을 제안하였다.

$$\hat{\beta} = \arg\min_{\beta} L(\lambda_1, \lambda_2, \beta) = \arg\min_{\beta} \left[|y - X\beta|^2 + \lambda_2 |\beta|^2 + \lambda_1 |\beta|_1 \right], \; |\beta|^2 = \sum_{j=1}^{p} \beta_j^2, \; |\beta|_1 = \sum_{j=1}^{p} |\beta_j|.$$

Elastic net 추정치 $\hat{\beta}$은 penalty 모수가 (λ_1, λ_2)이며 주어진 augmented 자료가 (y^*, X^*) 일 때, LASSO 타입으로 추정치를 구할 수 있다.

$$\hat{\beta}(elastic \; = t) = \sqrt{1+\lambda_2}\,\hat{\beta}^* = (1+\lambda_2)\hat{\beta}(naive \; elastic \; = t).$$

where $\hat{\beta}^* = \arg\min_{\beta^*} \|y^* - X^*\beta^*\|^2 + \frac{\lambda_1}{\sqrt{1+\lambda_2}} |\beta^*|_1$

Elastic net 계수는 재척도화된 naive elastic net의 계수이다. 척도변환은 naive elastic의 변수선택 성질을 보존하고 shrinkage의 가장 단순한 방법으로 LASSO의 안정화 버전이라 할 수 있다.

자동 변수 선택방법인 naive elastic net은 LASSO의 한계를 극복할 수 있으나 그리 만족스러운

결과를 주지는 못하여 naive라 불린다. 회귀 예측력 설정에서 정확한 벌점화(penalization) 방법은 잔차와 분산의 교환을 통해 좋은 예측력을 갖는 방법을 말한다. naive elastic net 추정치는 먼저 ridge 회귀계수를 찾아 λ_2를 고정시키고, 다음으로 LASSO 방법을 통해 shrinkage를 한다. 여기서shrinkage를 두 번 시행하게 되는데, 이러한 두 번의 shrinkage 과정은 LASSO나 ridge shrinkage와 비교하여 불필요한 여분의 잔차를 가지게 되어 분산을 줄이는 것에 도움이 되지 않는다. 따라서 변수들간에 그룹화가 형성이 되어 있지 않을 때에 변수 선택에 문제가 있다.

3.5. LARS(Least Angle Regression)

LASSO와 이와 연관된 추정량을 얻기 위해 Efron et al.(2004)에 의해 제안된 LARS 알고리즘은 수리적으로 간단하고 빠르게 계산이 가능한 stagewise 절차의 다른 형태로, 컴퓨팅 속도가 매우 빠르게 개선한 유용하고 간결한 모형 알고리즘이다. LASSO를 간단한 수정으로 이행할 수 있으며 회귀계수들의 절대값의 합을 제약조건으로 하는 OLS의 유용한 버전이다. 변수가 p개일 때, 오직 p개의 단계가 요구된다.

LARS 절차는 먼저 전진선택법과 같이 모든 계수를 0으로 만들며 시작하며 반응변수와 가장 관계가 높은 설명변수 x_1을 찾는다. 다음으로 x_1을 제외한 설명변수에서 가장 관계가 높은 설명변수 x_2를 찾을 때까지 계속 진행한다. 다음으로 가장 높은 관계인 x_3를 찾을 때까지 이미 뽑힌 두 개의 변수들(x_1,x_2) 사이의 예측방향이 등각이 되도록 진행한다. 이와 같이 LARS는 최소 각 방향에 따라 다음 변수가 뽑힐 때까지 뽑힌 변수들 사이에서 등각을 이어나간다. LARS 추정치는 단계가 진행됨에 따라 모형에 변수를 하나씩 더하고 k 단계 후에는 0이 아닌 변수들이 k개가 되어 최종 모형은 k개의 모수를 가지게 된다. 즉 오직 선택된 변수의 개수만큼만 계산되기 때문에 LARS는 시간이 절약되는 것이 특징이다.

4. 실증분석

4.1. 전립선(prostate)자료

전립선자료는 Stamey et al.(1989)에 의해 연구된 자료로, 전립선 적출술을 받은 남성 97명의 의학 측정자료와 전립선 특이항원 양 사이의 상관관계를 검사한 자료이다. large-p-small-n 형태의 자료에서의 적합도 중요하지만, 우선 large-n-small-p에서 적합이 잘 되어야 하고, 또한 본 연구의 결과와 선행 연구의 결과를 비교하기 위해 이 자료를 사용하며 전립선 자료의 요인으로 8가지를 고려한다. 즉, log(암의 크기)(lcavol), log(전립선 무게)(lweight), 나이(age), log(전립선 비대종양의 양)(lbph), 암이 정낭에 침범한 확률(svi), log(세균 침투량)(lcp), 글리손점수(gleason), 글리손 4점 또는 5점의 백분율(pgg45)을 고려한다.

글리손 점수는 조직검사에서 암의 악성도를 숫자로 표현한 점수이며 6점 이하면 낮은 악성도를 띄는 암이다. 더빗-왓슨 통계량이 1.507로 자료는 독립성을 가지고 있다. 8개의 요인들과 반응변수를 표준화시킨 후, 반응변수인 log(전립선 특이항원)에 선형모형으로 적합시켰다. 본 논문의 실증분석에 있는 모든 예제는 R package(버전 3.2.0)에서 monomvn, lars, elasticnet, parcor 등을 이용하여 ridge, LASSO, LARS, adaptive LASSO, elastic net의 결과를 비교하였다. 본 분석에 사용된 자료는 전체 자료 중 임의로 선택된 자료로 $n=80, p=8$의 형태이고 더빈-왓슨 통계량은 1.718, p-value는 0.904이다.

다섯 가지 방법 모두 적용한 결과 Table 1을 보면, 암의 크기(lcavol)가 가장 영향력이 큰 것으로

나타났다. 다음으로 암이 정낭에 침범할 확률(svi), 전립선 무게(lweight), 전립선 비대종양의 양(lbph) 순 등이었다. 가장 영향력이 작은 변수는 나이(age)이다. 그 외에 세균침투, 글리손 점수, 글리손 4점 또는 5점의 백분율은 전립선 특이항원과 크게 관계가 없는 것으로 보인다. ridge를 제외하고 LASSO와 LARS는 4개, elastic net은 6개, adaptive LASSO는 4개의 계수를 가지는 것으로 나타났다. 참고로 Tibshirani(1996)는 LASSO를 이용하여 lcavol, svi, lweight 3가지 변수를 선택하였는데, 위 결과와 같이 전립선 특이항원에 가장 영향을 많이 끼치는 것은 암의 크기였다.

모형 선택기준인 RSS, AIC, BIC를 종합적으로 보았을 때 다섯 가지 방법 중 LARS가 가장 좋은 결과를 주었다. Ridge는 8개의 요인을 모두 사용하여 모형을 만듦으로 RSS값이 작지만 간결성을 충족하지 못하고, 속도가 느린 것을 알 수 있다. LASSO와 LARS는 앞에서 언급한 것과 같이 같은 값을 가지나 속도 면에서 LARS가 빠른 것을 알 수 있다. Elastic net은 언급된 방법 중 결과가 가장 좋지 않았다. Adaptive LASSO는 조절변수의 가중치를 매번 구해야 하므로 속도가 매우 느리게 나타난다. 따라서 LARS에 의해 추정된 다음의 모형이 가장 적절한 모형이라 할 수 있다.

$$\hat{y} = 0.522lcavol + 0.069lweight + 0.028lbph + 0.164svi.$$

Table 1. Comparison of various estimates in prostate data

Variable	Ridge	LASSO	LARS	Elastic net	Adaptive LASSO
lcavol	0.565	0.522	0.522	0.457	0.605
lweight	0.149	0.069	0.069	0.113	0.054
age	-0.116	0	0	0	0
lbph	0.135	0.028	0.028	0	0.071
svi	0.263	0.164	0.164	0.221	0.201
lcp	-0.091	0	0	0.122	0
gleason	0.045	0	0	0.037	0
pgg45	0.087	0	0	0.041	0
RSS	26.092	28.123	28.123	30.185	28.662
AIC	334.377	337.467	337.467	348.513	341.490
BIC	357.550	355.490	355.490	371.686	362.087
CPU(minutes)	12.146	1.804	1.704	2.441	1379.677

4.2. 경제지표 자료

다음으로 한국은행 경제통계시스템에서 발췌한 2010년도부터 2014년도까지 한국 경제지표 자료를 가지고 분석을 하였다. 한국은행 경제통계시스템에서는 한국은행 및 타 기관 작성 통계수치를 이용자가 빠르고 편리하게 열람할 수 있도록 그래프와 보도자료 등을 제공한다. 최근 한국뿐만 아니라 전 세계적으로 청년실업이 심각하며 빈곤층은 계속해서 늘고 있을 정도로 경제상황이 좋지 않다. 이러한 경제상황에서 몇 가지 반응변수를 설정하여 어떤 변수가 반응변수에 크게 영향을 주며 유의한지에 대해 알아본다.

제조업 부문에서 부분적으로 제조업 업황실적 BSI를 반응변수로 하고 세부사항을 나누어 설명변수로 두어 small-n-large-p의 월별자료를 이용하였다. 이 분석을 통해 전반적으로 한국의 제조업에 가장 큰 영향을 주는 제조업은 어떤 것인지 알아보고 영향을 적게 주는 제조업은 어떤 것인지 알아본다. 또한, 한국경제통계시스템의 100대 통계지표에서 월별자료가 있는 것과 분류할 수 있는 지표들을 분류하여 설명변수를 늘린 후, 최근 큰 이슈인 소비자 물가지수를 반응변수로 둔 자료도 살펴본다. 이 100대 통계지표는 한국경제에 가장 큰 영향을 주고받는 지표들로 한국의 경제상황을

한 눈에 알아볼 수 있다. 모든 요인자료들은 계절적 영향을 없애기 위해 전기대비 증감 지수자료를 사용하고 표준화하였다.

1) 제조업 업황실적 BSI 자료

BSI(business survey index)는 기업가의 현재 기업경영상황에 대한 판단과 향후 전망을 조사하여 경기 동향을 파악하고 경기를 전망하기 위해 작성되고 있으며, 각 업체의 응답을 아래와 같은 공식에 따라 지수화한 것이다.

$$\text{업종별 } BSI = \frac{(\text{긍정적인 응답업체수} - \text{부정적인 응답업체수})}{\text{전체 응답업체수}} \times 100 + 100.$$

BSI가 기준치인 100인 경우 긍정적인 응답업체수와 부정적인 응답업체수가 같음을 의미하며, 100 이상인 경우에는 긍정 응답업체수가 부정 응답업체수보다 많음을, 100이하인 경우에는 그 반대임을 나타낸다. 산업별 BSI의 공식은 다음과 같다.

$$\text{산업(제조업,비제조업)별 } BSI = \sum_{i=1}^{n} w_i BSI_i.$$

여기서 w_i는 각 업종별 GDP 비중, BSI_i는 각 업종별 BSI를 나타낸다(Park, Park, 2015).

설명변수는 식료품, 음료, 섬유, 의복모피 등 제조업 부문의 28가지이고, 반응변수는 그 28가지의 수치를 통합하여 나타낸 수치이다. 기간은 2014년 1년 동안을 12개월로 나타낸 것으로 12개의 반응변수를 가지고 자료는 $n = 12, p = 28$ 형태이다. Table 2를 보면 ridge를 제외하고 LASSO와 LARS는 4개, elastic net은 5개, adaptive LASSO는 2개의 계수를 가진다.

RSS, AIC, BIC를 종합적으로 본 결과, 이 역시 LASSO와 LARS의 결과가 가장 좋게 나타났고, 속도는 LARS가 가장 빨랐다(Table 2). 설명변수간에 그룹화가 형성이 되어 있지 않기 때문에 elastic net의 결과가 가장 좋지 못하였고, 또한 계수를 2개를 뽑은 adaptive LASSO보다 RSS값도 낮은 것으로 나타났다. Adaptive LASSO는 조절변수의 가중치를 매번 구해야 하므로 이번 자료에서 역시 속도가 가장 느린 것으로 나타났다. LARS에 의해 추정된 모형은

$$\hat{y} = 0.029ec23 + 0.264ec24 + 0.075ec25 + 0.515ec26.$$

이며, 반응변수에 가장 크게 영향을 주는 변수는 경공업 업황실적 BSI(ec26), 대기업(ec24), 중소기업(ec25) 순으로 나타났다. 즉, 분석결과 제조업의 업황실적에 경공업, 대기업, 중소기업 순으로 영향을 끼치는 것으로 나타나, 한국 제조업의 업황실적을 크게 좌지우지하는 요인이 이 3가지임을 알 수 있다. 방법마다 각각 결과차이가 나타나지만 대체적으로 비슷하게 변수계수를 설정한다.

2) 소비자 물가지수

소비자 물가지수는 가구에서 일상생활을 영위하기 위해 구입하는 상품과 서비스의 가격변동을 측정하기 위하여 작성한 지수이다. 기준연도는 2010년이며 조사품목은 상품 및 서비스 481개 품목이다. 가중치는 2012년 전국가구(농·어가 제외) 월평균 소비지출액에서 각 품목의 소비지출액이 차지하는 비중으로 1,000분비로 산출한다. 가격조사는 서울, 부산, 대구 등 37개 도시에서 조사하며 농축수산물, 석유류, 공업제품, 전기·수도·가스, 서비스, 집세로 분류한다. 계산식은 라스파이레스 산식을 이용하며 다음 식과 같다(Kim, Kim, 2015).

$$P_L = \frac{\sum(P_i^t Q_i^{2012})}{\sum(P_i^{2012} Q_i^{2012})} = \sum S_i^{2012}(P_i^t / P_i^{2012}) \text{ where } S_i^{2012} = \frac{(P_i^{2012} Q_i^{2012})}{\sum(P_i^{2012} Q_i^{2012})}.$$

여기서 P는 가격, Q는 수량, S는 가중치, 2012는 가격 및 가중치기준시점, t는 가격조사시점, i는 품목이다. 설명변수는 100대 통계지표와 제조업산업에서 몇 가지 세분화된 월별자료로 개수는 196개이다. 2010년 1월부터 2014년 12월까지 총 60개의 자료로 $n=60, p=196$ 형태의 자료이다.

Table 2. Comparison of various estimates in BSI data

Variable	Ridge	LASSO	LARS	Elastic net	Adaptive LASSO
ec1(grocery)	-0.015	0	0	0	0
ec2(beverage)	0.043	0	0	0	0
ec3(fiber)	0.074	0	0	0	0
ec4(clothes·fur)	0.055	0	0	0	0
ec5(leather·bag·shoes)	0.035	0	0	0	0
ec6(timber·wood)	0.054	0	0	0	0
ec7(pulp·paper)	-0.035	0	0	0	0
ec8(printing·record)	-0.02	0	0	0	0
ec9(oil refining·cokes)	-0.024	0	0	0	0
ec10(chemicals)	0.092	0	0	0	0
ec11(medicals·medicine)	0.02	0	0	0	0
ec12(rubber·plastic)	-0.033	0	0	0	0
ec13(nonmetallic mineral)	-0.012	0	0	0	0
ec14(primary metal)	0.111	0	0	0.096	0
ec15(metal processing)	0.03	0	0	0	0
ec16(electronic·image)	0.093	0	0	0	0
ec17(medical·precision device)	0.058	0	0	0	0
ec18(electrical equipment)	0.058	0	0	0	0
ec19(other machineries)	0.094	0	0	0	0
ec20(vehicle)	0.002	0	0	0	0
ec21(shipbuilding·transport)	0.03	0	0	0	0
ec22(furniture)	-0.03	0	0	0	0
ec23(cigarette, other products)	0.095	0.029	0.029	0	0
ec24(medium and small firm)	0.12	0.264	0.264	0.291	0.015
ec25(heavy chemical industry)	0.104	0.075	0.075	0.132	0
ec26(light industry)	0.131	0.515	0.515	0.322	0.92
ec27(export company)	0.032	0	0	0	0
ec28(domestic company)	0.09	0	0	0.087	0
RSS	0.179	0.45	0.45	0.52	0.466
AIC	37.326	0.494	0.494	50.161	46.838
BIC	51.389	2.919	2.919	64.223	60.415
CPU(minutes)	34.173	3.0516	2.3121	8.903	196.928

Table 3에서 RSS, AIC와 BIC를 종합하여 보았을 때 LASSO의 결과가 좋으나 LARS와 크게 차이가 없는 것으로 보인다. 유의한 계수의 개수는 ridge를 제외하고 LASSO는 29개, LARS는 30개, elastic net은 35개, adaptive LASSO는 7개로 나타났다. Ridge는 RSS 값은 작으나 AIC와 BIC의 값이 상당히 높고, elastic net과 adaptive LASSO의 결과 값도 그리 좋지 못한 것으로 보인다. Elastic net과 adaptive LASSO만 비교하였을 때, elastic net이 더 좋은 모형을 나타내는데 이는 그룹화 되지 않

은 모형이므로 이러한 결과가 나타났다고 할 수 있다. 종합적으로 보았을 때 소비자 물가지수에 가장 영향력이 큰 변수는 생산자 물가지수(ec177), 근원 인플레이션지수(ec180), 예금은행대출금(ec165) 순이었다. 소비자 물가지수에 가장 영향력이 큰 것은 생산자 물가지수임에 틀림없다. 그리고 생산자 물가지수와 소비자 물가지수는 비슷한 동향을 가진다고 할 수 있다. 근원 인플레이션지수 역시 생산자 물가지수와 비슷한 동향을 가지나 예금은행대출금은 다른 동향을 가지며 소비자 물가지수에 큰 영향을 끼치는 것으로 보인다. LARS에 의한 최종모형은 다음과 같다.

$$\begin{aligned}\hat{y} = & -0.052ec6 + 0.052ec7 + 0.036ec8 - 0.033ec12 - 0.099ec13 + 0.008ec17 \\ & - 0.109ec20 + 0.021ec35 - 0.044ec43 - 0.034ec45 - 0.034ec78 - 0.023ec88 \\ & - 0.083ec96 - 0.107ec99 + 0.006ec129 - 0.083ec143 + 0.034ec146 \\ & - 0.04ec147 + 0.004ec150 + 0.04e162 - 0.116ec165 + 0.021ec168 + 0.041ec171 \\ & + 0.055ec174 - 0.04ec175 + 0.329ec177 + 0.286ec180 + 0.036ec187.\end{aligned}$$

Table 3. Comparison of various estimates in consumer price index data

Method	Ridge	LASSO	LARS	Elastic net	Adaptive LASSO
Number of selected variables	196	29	30	35	7
RSS	0.122	9.07	9	14.861	23.22
AIC	263.706	186.293	187.818	549.925	576.701
BIC	672.104	242.841	246.46	956.228	983.004
CPU(minutes)	166.206	42.174	22.504	297.038	1935.02

5. 결론

본 논문에서 ridge, LASSO, LARS, elastic net, adaptive LASSO 등 다섯 가지 벌점화우도(penalized likelihood) 방법을 이용하여 실제 자료를 고차원 선형모형에 적합하였고, RSS, AIC, BIC 값으로 다섯 가지 벌점화우도 방법들의 적합도를 비교해 보았다. 세 가지 모형 선택기준을 통해 가장 좋은 모형을 추정하는 방법으로 본 연구에서는 LARS인 것으로 나타났다.

Ridge는 좋은 모형의 기준인 간결성을 갖출 수 없는데, 이를 보완하기 위하여 LASSO가 제안되었으며 LARS는 LASSO의 약간 수정된 버전으로 속도가 훨씬 빠르다. 또한, 다중공선성에 약한 LASSO를 보완하기 위하여 그룹화된 자료에 강한 elastic net과 계수의 수치가 큰 자료에 유용한 adaptive LASSO가 제안되었다. 하지만 그룹화되지 않은 자료와 계수의 수치가 크지 않은 자료에서는 LASSO와 LARS가 좋은 것으로 판단된다.

LASSO와 LARS는 거의 비슷한 결과 값을 갖는데, LARS의 속도가 훨씬 빠른 것을 알 수 있다. 특히 그 차이는 고차원 자료에서 두드러지게 나타났다. 결과가 가장 좋지 않은 것은 adaptive LASSO이며 그 이유는 그룹화 되어있지 않은 자료이고, 조절변수의 가중치를 변수마다 매번 구해야하는 작업 때문에 속도도 adaptive LASSO가 가장 느린 것으로 보인다.

References

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Trans Automatic Control,* 19(6), 716-723.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. (2004). Least angle regression, *Annals of Statistics,* 32, 407-451.

Hoerl, A. E., Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems, *Technometrics,* 12, 55-67.

James, R., Witten, D., Hastie, T., Tibshirani, R. (2013). *An introduction to statistical learning,* Springer.

Joo, A. (2015). *The variable selection methods using the shrinkage methods in high-dimensional linear model,* Master's Thesis, Korea University.

Jung, K.-M. (2007). A robust estimator in ridge regression, *Journal of the Korean Data Analysis Society,* 9(2), 535-543.

Kim, B. C. (2000). *Matrix algebra for statistics,* Free Academy.

Kim, K. Y., Jhun, M. S., Kang, H. C., Lee, S. K. (2012). *Regression analysis by example,* Free Academy.

Kim, B. K., Kim, D. Y. (2015). *Consumer price index in March 2015,* Statistics Korea.

Lee, S., Kwon, S. (2015). Moderately clipped LASSO for the sparse high-dimensional logistic regression models, *Journal of the Korean Data Analysis Society,* 17(3A), 1145-1154.

Mallick, H., Yi, N. (2013). Bayesian methods for high dimensional linear models, *Journal of Biometrics and Biostatistics*, S1, 005. doi, 10.4172/2155-6180.S1-005.

Oh, E. J., Lee, H. (2013). Dimension reduction and prediction for high-dimensional regression models using the graphical lasso, *Journal of the Korean Data Analysis Society,* 15(5), 2321-2332.

Park, S. B., Park, D. H. (2015). *Business survey index and economic sentiment index in April 2015,* The Bank of Korea.

Schwarz, G. (1978). Estimating the dimension of a model, *The Annals of Statistics,* 6, 461-464.

Stamey, T. A., Kabalin, J. N., McNeal, J. E., Johnstone, I. M., Freiha, F., Redwine, E. A., Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate, II, radical prostatectomy treated patients, *The Journal of Urology,* 141(5), 1076-1083.

Sun, X., Choi, H., Kwon, S. (2007). A sparse ridge estimation for the sparse logistic regression model, *Journal of the Korean Data Analysis Society*, 16(4A), 1715-1725.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO, *Journal of the Royal Statistical Society, Series B,* 58, 267-288.

Zou, H. (2006). The adaptive LASSO and its oracle properties, *Journal of the American Statistical Association,* 101, 1418-1429.

Zou, H., Hastie, T. (2005). Regularization and variable select ion via the naive net, *Journal of the Royal Statistical Society, Series B*, 67, 301-320.

The Study of Variable Selection Methods Using the Penalized Likelihood Methods in High-dimensional Linear Model

Alim Joo[1], *Sooyoung Cheon*[2]

Abstract

If the sample size is greater than the number of covariates or parameters, typically forward selection, backward elimination and stepwise selection methods are used among several variable selection methods. However, in the case of high-dimensional linear models, it is difficult to use such methods. In order to overcome this problem, penalized likelihood, also called regularization or shrinkage, methods may be used in high-dimensional linear models. This paper considers several variable selection methods used in high-dimensional linear models, and investigates the applicability of penalized likelihood methods. We briefly review ridge, LASSO, LARS, elastic net, adaptive LASSO, and apply these methods to the real data, prostate and Korea economic index data. Numerical results indicate that the best method was the LARS algorithm among five penalized likelihood methods based on RSS, AIC and BIC.

Keywords : Variable selection method, High-dimension, Ridge, LASSO, LARS.

[1]Graduate Student, Department of Applied Statistics, College of Science and Technology, Korea University, 2511 Sejong-ro, Sejong-city, 339-700, Korea. E-mail : jooalim@korea.ac.kr
[2](Corresponding Author) Associate Professor, Department of Applied Statistics, College of Science and Technology, Korea University, 2511 Sejong-ro, Sejong-city, 339-700, Korea. E-mail : scheon@korea.ac.kr
[Received 20 September 2015; Revised 17 October 2015; Accepted 20 October 2015]