碩 士 學 位 論 文

고차원 선형모형에서 shrinkage 방법을 이용한 변수 선택 방법

高麗大學校 大學院

情報統計學科

周 아 림

2015年 6月

全 秀 榮 敎 授 指 導

碩 士 學 位 論 文

고차원 선형모형에서 shrinkage 방법을 이용한 변수 선택 방법

이 論文을 統計學 碩士學位 論文으로 提出함.

2015年 6月

高 麗 大 學 校 大 學 院

情 報 統 計 學 科

周 아 림 (印)

周 아 림의 統計學 碩士學位論文

審査를 完了함.

2015年 6月

委員長 (印)

委 員 洪 勝 萬 (印)

委 員 (印)

고차원 선형모형에서 shrinkage 방법을 이용한 변수 선택방법

요 약 문

이 논문은 고차원 선형모형에서 사용이 용이한 변수 선택방법에 대한 연구로, 여러 가지 shrinkage 방법에 대해 알아볼 것이다. n이 p보다 큰 경우 고전적으로 쓰이는 변수 선택방법인 전진적 선택방법, 후진 제거방법, 단계적 선택방법이 사용 가능하지만, n이 p보다 작은 경우의 고차원 선형모형에서는 사용하는 데에 어려움이 있다. 반면에 regularization 또는 penalized regression 방법이라고도 불리는 shrinkage 방법들은 고차원 선형모형에서 사용하기 편리하고 그 결과가 믿을만하다. 본 논문은 그 중 Ridge, LASSO, LARS, Elastic net, Adaptive LASSO 등 다섯 가지 shrinkage 방법의 특징을 간단히 살펴보고, 실제 자료인 전립선(prostate)자료와 한국 경제 수치자료에 적용해 본다. 또한 다섯 가지 shrinkage 방법의 RSS, AIC, BIC 값으로 비교하여 어떤 방법이 고차원 선형모형에서 가장 적합한지 알아본다.

목 차

표 목 차

그 림 목 차

제 1 장 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. 이것은 변수를 0으로 만드는 것과 같다. 과학적인 연구에서 정확하게 관계있는 변수를 찾는 것은 중요한 문제이기 때문에 잘못된 모형은 눈에 띄게 잘못된 결과를 가져다 줄 수 있다. 더 많은 변수를 포함하는 모형은 반응변수를 확률적으로 더 많이 설명할 수 있게 된다. 반면에, 과적합(overfitting)은 추정된 모형에 영향력이 없는 변수를 다수 포함하게 되어 예측력이 낮아지고 결과의 신뢰성이 떨어지게 된다.

고차원에서의 변수선택 문제는 다른 연구 분야보다 더 연구할 것이 많은 분야이다. 고차원 문제는 생물정보학, 경제학, 약학, 유전학 등의 많은 과학 분야에서 떠오르고 있다. 고차원의 특징은 모형이 매우 복잡하며 추정, 예측력, 설명력을 구하는 것이 상당히 어렵다는 것이다(Mallick and Yi, 2013). 그러므로 고전적인 통계방법은 계산적으로 실행이 불가능하거나 모형이 정확히 추정되지 않거나 둘 다의 문제를 가지고 있다. 많은 방법들이 고차원 선형모형에서 변수선택을 다루기 위해 개발되어 왔는데, 최근 새로운 shrinkage 방법들이 많이 제안되었다.

고차원 선형모형에서 변수선택 문제를 해결하기 위해 전부터 Hoerl and Kennard(1970), Tibshirani(1996), Efron et al.(2004), Zou and Hastie(2005) 그리고 Zou(2006) 등이 제안하였으며, 이러한 방법들은 유의하지 않은 변수의 계수들을 0에 가까운 수로 줄여주어 모형에서 제거하는 방법들이다. 그 중 첫 번째로 Hoerl and Kennard(1970)에 의해 제안된 Ridge가 있는데, 이 방법은 계수를 0 가까이로는 만들지만 정확히 0으로 줄이지는 못한다. Tibshirani(1996)에 의해 제안된 LASSO 방법은 계수를 정확히 0으로 만들어 계수의 개수를 줄여 줄 수 있다. Efron et al.(2004)에 의해 제안된 LARS는 전진적 선택방법을 조금 변형한 것이며 LASSO와 비슷한 결과 값을 가지지만 속도는 훨씬 빠르다. Zou and 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 그룹화된 자료에서 Elastic net과 Adaptive LASSO는 결과가 좋게 나오는데, Adaptive LASSO의 경우 각 계수마다 가중치(weight)를 따로 구해줘야 하므로 추정치를 구하는데 걸리는 시간이 다소 길다는 단점이 있다.

본 연구에서는 실제자료를 이용하여 고차원 선형모형에서의 다양한 변수선택방법들을 비교해 본다. 분석에 첫 번째로 사용된 실제 자료는 전립선(prostate)자료로 R에 내장되어 있으며 전립선 암과 특이항원의 관계에 대한 자료이다. 여러 의학적 수치가 기술되어 있으며 8개의 설명변수와 1개의 반응변수로 가장 영향력이 큰 변수들을 알아보

고 영향력이 거의 없는 변수는 어떤 것들이 있는지 알아본다. 두 번째 자료는 제조업의 여러 수치들을 설명변수로 두고 통합수치를 반응변수로 둔 경제자료로 제조업 업황실적 BSI, 제조업 생산지수, 제조업 가동률지수이다. 모두 small-n-large-p 형태로, 설명변수의 개수가 반응변수의 개수보다 크다. 분석을 위해 요즘 가장 경제적으로 이슈가 되고 있는 소비자 물가지수를 반응변수로 두고 설명변수로 한국 대표 경제지표 중에서 중요한 지표를 임의로 몇 개를 선택하였다. 소비자 물가지수가 어떤 경제지표에 영향을 크게 받는지 알아보고, 어떤 방법이 가장 결과가 좋은지 알아본다. 본 논문에서는 각각의 방법을 비교하고 여러 가지 자료에 따라 결과 값이 어떻게 달라질지 RSS, AIC, BIC를 이용하여 알아본다.

제 2 장에서는 고전적인 모형 선택방법인 전진적 선택방법, 후진 제거방법, 단계적 선택방법에 대해 알아보고, 본 논문에서 사용하게 될 모형 선택의 기준 3가지 RSS, AIC, BIC에 대해 알아본다. 제 3 장에서는 고차원이 어떤 것인지, 특징은 무엇인지에 대해 서술하고, 고차원 문제에서 사용할 수 있는 shrinkage 방법인 Ridge, LASSO, LARS, Elastic net, Adaptive LASSO에 대해 간단히 요약한다. 제 4 장에서는 실제 자료를 이용한 실증분석으로 전립선(prostate)자료와 최근의 경제 지표들을 이용하여 각 shrinkage 방법들을 비교한다.

제 2 장 고전적 모형 선택방법

주어진 자료에서 유의한 변수를 잘 선택해 모형을 만들어야 잘 적합된 모형이라 할 수 있다. 여러 모형 선택방법 중에 shrinkage 방법을 알아보기 전에 전부터 널리 사용되던 고전적 모형 선택방법과 예전부터 지금까지 널리 쓰이는 모형 선택의 기준에 대해 알아본다.

2.1 고전적 모형 선택방법

고전적 통계방법에서 흔히 사용되는 모형 선택방법은 전진적 선택방법, 후진 제거방법, 단계적 선택방법 등이 있다. 이 방법들은 각 단계에서 예측변수를 더하거나 제거하며 그 결과에 따라 모형을 선택한다.

2.1.1 전진적 선택방법

전진적 선택방법은 설명변수를 하나 선정하여 먼저 단순회귀모형에서부터 분석을 시작한다. 그 다음으로 나머지 변수 중 하나의 변수를 추가로 선정하여 설명변수가 2개인 회귀모형에 대하여 자료를 적합시킨다. 이러한 과정을 반복하여 변수를 하나씩 증가시키면서 자료에 가장 잘 적합한 모형을 찾아내는 방법이다. 이 방법은 이해하기 쉽고 변수의 개수가 많은 경우에도 사용할 수 있다는 장점이 있어 많이 사용되고 있다. 하지만 최적의 모형을 찾지 못할 수도 있으며 변수 값의 조그마한 변동에도 그 결과가 크게 달라지는 등 안정성이 약하다.

2.1.2 후진 제거방법

후진 제거방법은 전체 p개 변수들의 집합에서 각 단계마다 반응변수와 연관성이 적은 변수들을 사용자가 정한 함수의 값이 더 이상 향상되지 않을 때까지 제거해 나가는 방법이다. 이 방법은 중요한 변수가 모형에서 제외될 가능성이 적으므로 비교적 안전한 방법이라 할 수 있다. 그러나 한번 제외된 변수는 다시 선택되지 못한다는 단점이 있다(이재은, 2012).

2.1.3 단계적 선택방법

앞에서 논의한 두 가지 선택방법으로 선정된 회귀모형이 최상의 모형이라고는 할 수 없다. 왜냐하면 전진적 선택방법에서는 이미 채택된 변수와 추가될 변수와의 모형에 대한 결정계수 R^2값을 최대화 하는데 중점을 두고 있기 때문이다. 설명변수들이 두개인 ${}_kC_2$가지의 모형을 모두 고려하였을 때 최대의 R^2값을 갖고 있는 모형과 전진선택

법의 두 번째 단계에서 선정된 모형과 반드시 일치하지 않는다. 또한 새로운 변수를 삽입시켜 모형을 설정한 후에 모든 회귀계수의 검정을 해보면 이미 전 단계에서 선택된 변수의 회귀계수가 통계적으로 유의하지 않은 경우가 발생할 수 있으며, 이런 경우에 선택된 모형은 적절하다고 판단하기 어렵다. 후진 제거방법도 마찬가지의 이유 때문에 적절한 방법이 될 수 없는데 이런 단점을 보완한 방법이 단계적 선택방법이다. 이 방법은 전진적 선택방법과 동일하게 새로이 추가할 변수를 선택하는데, 설정된 후의 회귀모형식에 있는 모든 설명변수들의 회귀계수를 검정하여 유의하지 못한 회귀계수에 대한 변수는 삭제하고 다시 전 단계로 되돌아가 분석한다. 그러므로 단계적 선택방법은 전진적 선택방법과 후진 제거방법의 복합적인 방법이라 설명할 수 있다(김창주, 2013).

예측을 위한 모형 선택의 우선적인 기준은 의미 있고(meaningful), 설명력 있고(interpretable), 간결(parsimonious)해야 한다는 것이다. 하지만, 위의 3가지 고전적 모형 선택방법은 하나이상의 모형 선택 기준에 미치지 못하는 것으로 알려져 있다. 더구나 고차원 선형모형에서는 위의 방법들을 사용할 수 없는 경우가 많다(Mallick and Yi, 2013). 본 연구에서는 이러한 단점을 극복하기 위해 shrinkage 방법에 대해 알아볼 것이다.

2.2 모형 선택의 기준

모형을 선택하는 데에는 여러 가지 기준이 있다. 본 논문에서는 많은 기준 중에서 RSS, AIC, BIC 세 가지 기준을 두고 모형을 선택할 것이다.

2.2.1 RSS

RSS(Residual Sum of Squares)는 잔차제곱합으로 관측치 y와 추정된 값과의 편차제곱합이며, RSS 값이 작을수록 잔차의 값이 작아지는 것을 뜻한다. 따라서 여러 모형 중에서 RSS 값이 작은 모형을 좋은 모형이라고 할 수 있다. RSS는 식 (1)과 같다(김병천, 2000).

$$RSS = \sum_{i=1}^{n}\sum_{j=1}^{n}(y_{ij}-\widehat{y_{ij}})^2 = (\boldsymbol{y}-\hat{\boldsymbol{y}})^t(\boldsymbol{y}-\hat{\boldsymbol{y}}) \qquad (1)$$

여기서, $\boldsymbol{y}^t=(y_1,...,y_n)$이다.

2.2.2 AIC

회귀에서 변수 선택은 모형 선택의 문제와 동일하게 볼 수 있다. Akaike(1974)에 의해 제안된 AIC(Akaike Information Criteria)는 적합도와 간결성 사이의 상충을 잘 조절하려 한 모형 선택 기준이다. n개의 자료 수와 p개의 설명변수를 갖는 회귀식에 대한 AIC는 식 (2)와 같다.

$$\text{AIC} = n*\ln(\text{RSS}) + 2p \qquad (2)$$

AIC는 값이 작을수록 선호된다. 단일 모형에 대한 AIC의 수치적인 값은 그다지 의미를 지니지 않는다. 그러나 AIC는 적합도와 간결성의 두 기준을 기초로 모형들의 순위를 부여할 수 있다. AIC 값이 2 정도의 차이가 나지 않는 모형들은 모형의 질을 동일하게 취급할 수 있다. 그러나 AIC 값에서 큰 차이를 보이면 모형의 질이 유의한 차이가 있음을 나타낸다. 더 작은 AIC를 갖는 모형을 선택해야 한다.

AIC의 큰 장점은 내포되지 않은 모형들을 비교할 수 있다는 것이다. 임의의 한 모형이 더 큰 모형의 특수한 경우로서 얻어질 수 있다면 그 모형은 내포되었다고 한다. 이러한 경우, 만약 변수 X_1, X_2, X_3, X_4, X_5가 있다고 한다면, (X_1, X_2, X_3)에 기반한 모형과 (X_4, X_5)에 기반한 모형의 성능을 비교하려할 때 F-검정을 수행할 수 없다는 문제점이 발생한다. 따라서 이러한 두 변수의 집합을 비교·선택하는 것이 문제의 본질이다.

하지만, AIC는 F-검정으로는 할 수 없었던 이러한 비교를 가능하게 한다.

AIC로 모형을 비교하려면, 데이터에 결측값이 없어야 하고, 동일한 관측값에 대해서 계산되어야 한다. 어떤 변수에서 많은 결측값이 발생하게 되면 AIC는 효율적이지 못하다. 왜냐하면 결측값을 갖는 관측치를 제거하고 계산되기 때문이다.

2.2.3 BIC

AIC에 대한 여러 가지 수정된 방법들이 제안되었다. 하나의 유명한 방법은 Schwarz(1978)에 의해 제안된 베이즈 정보기준(BIC: Bayes Information Criteria)이다. BIC는 식 (3)과 같은 식을 가진다.

$$BIC = n*\ln(RSS) + p*\ln(n) \quad (3)$$

AIC와 BIC의 차이는 p에 대한 penalty 값의 정도이다. n>8인 경우에 BIC 기준이 더 큰 penalty를 주게 된다. 이는 AIC가 과적합되는(큰 p선호) 경향을 제어하게 된다(김기영 외 3인, 2012).

제 3 장 고차원 선형모형에서의 모형 선택방법

선형회귀는 설명변수가 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$로 주어지고 반응변수를 $\boldsymbol{y}=(y_1, \cdots, y_n)^t$라 할 때, 식 (4)와 같이 나타난다.

$$\hat{\boldsymbol{y}} = \widehat{\beta_0} + \boldsymbol{x}_1\widehat{\beta_1} + \ldots + \boldsymbol{x}_p\widehat{\beta_p} \tag{4}$$

모형적합 절차는 먼저 계수들을 $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)$인 벡터형태로 만든다. 이러한 벡터형태를 이용하여 OLS 추정치는 잔차제곱합을 최소화함으로 얻을 수 있다. 모형의 질을 평가하는 기준은 환경에 따라 달라질 수 있다. 예측력이 낮으면 모형에 의존하기 어렵다. 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. OLS는 예측력과 설명력 모두에 좋지 않다. 특히 고차원 선형회귀의 경우 설명력이 떨어지며 다중공선성이 있을 경우 불안하게 나타난다. penalization 기술은 OLS보다 뛰어나다고 제안된다.

고차원 문제는 고전적인 모형 선택방법들이 고차원 선형 모형에서 필요로 하는 변수들보다 많은 변수를 선택하는 과적합 문제를 일으키는 경향이 있다. 과적합 문제는 고차원 모형에서 심각한 편향을 발생시킬 수 있는데, 이를 피하고 더 좋은 예측력을 가지는 다양한 모형 선택방법들을 본 논문에서 알아볼 것이다(Zou and Hastie, 2005).

3.1 능형회귀(Ridge regression)

능형회귀는 예측변수들이 상당히 다중공선적일 때 사용될 수 있는 추정법이다. 이 방법을 사용하여 추정한 회귀계수의 능형추정량들은 편향되어 있으나 일반적으로 OLS 추정량보다 더 작은 최소제곱오차를 가지는 경향이 있는 것으로 알려져 있다(Hoerl and Kennard, 1970). 회귀계수에 대한 능형추정량은 약간 변형된 형태의 정규방정식을 통하여 얻을 수 있다. 표준화된 회귀모형이 식 (5)와 같이 주어졌다고 가정하자.

$$\hat{Y} = \beta_0 + \beta_1 \widehat{X_1} + \beta_2 \widehat{X_2} + \cdots + \beta_p \widehat{X_p} \tag{5}$$

능형회귀계수를 추정하기 위한 방정식은 식 (6)과 같다

$$\begin{array}{ccccccccc} (1+k)\beta_1 & + & r_{12} & \beta_2 & + & \cdots & + & r_{1p} & \beta_p & = & r_{1y} \\ r_{21} & \beta_1 & + & (1+k)\beta_2 & + & \cdots & + & r_{2p} & \beta_p & = & r_{2y} \\ & \vdots & & \vdots & & \vdots & & \vdots & & & \\ r_{p1} & \beta_1 & + & r_{p2} & \beta_2 & + & \cdots & + (1+k) & \beta_p & = & r_{py} \end{array} \tag{6}$$

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

능형회귀가 OLS와 다른 점은 k에 있다. k=0이면 $\hat{\beta}$은 OLS 추정치가 된다. 이 때 모수 k를 편향모수(bias parameter) 혹은 능형모수(Ridge parameter), 조절모수(tuning parameter)라 부르며 k가 0으로부터 증가하면 추정치의 편의(편향)도 증가하게 된다. 반면, 전체 분산은 식 (7)과 같이 k의 감소함수이다.

$$Total\ Variance(k) = \sum_{j=1}^{p} Var(\widehat{\beta_j}(k)) = \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^2}. \tag{7}$$

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 편향 크게 증가시키지 않으면서 전체 분산을 감소시키는 적절한 k를 찾는 것이다.

Hoerl and Kennard(1970)은 데이터의 작은 변화에 대하여 능형추정치가 안정적인

값을 취하는 적절한 양수 k가 존재함을 보였다. 김기영 외 3인(2012)은 [0,1] 사이의 범위에 있는 k에 대하여 먼저 능형추정값 $\hat{\beta}_1, \cdots, \hat{\beta}_p$을 계산하고, 그 결과들을 k에 대해 플롯한 다음 추정값의 안정성의 관점에서 적절한 k 값을 취할 수 있음을 보여 주었다.

Ridge의 목표함수는 식 (8)과 같다.

$$Q(\boldsymbol{\beta}) = (\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{8}$$

3.2 LASSO(Least Absolute Shrinkage and Selection Operator)

LASSO는 Tibshirani(1996)에 의해 제안된 기법이다. 먼저 $\boldsymbol{x}_i^t = (x_{i1}, \ldots, x_{ip})$가 설명변수이고 y_i가 반응변수인 경우, 자료를 $(\boldsymbol{x}_i, y_i)$로 표기한다. 보통 회귀설정에서는 관측치들이 독립적이거나 $\boldsymbol{x}_i$가 주어진 상태에서 조건적으로 독립적인 것으로 가정한다. 그리고 $\boldsymbol{x}_i$가 표준화되어있어 $\sum_i \boldsymbol{x}_i/N = 0, \sum_i \boldsymbol{x}_i^2/N = 1$로 된다.

$\hat{\boldsymbol{\beta}}$을 식 (9)과 같이 벡터형태로 나타낸다.

$$\hat{\boldsymbol{\beta}} = (\widehat{\beta_1}, \ldots, \widehat{\beta_p})^t \tag{9}$$

LASSO 추정치 $\hat{\beta}$은 식 (10)과 같이 정의된다.

$$(\hat{\beta}) = \arg \min_{\beta} \sum_{i=1}^{p} (y_i - \sum_j \beta_j \boldsymbol{x}_i)^2 \tag{10}$$

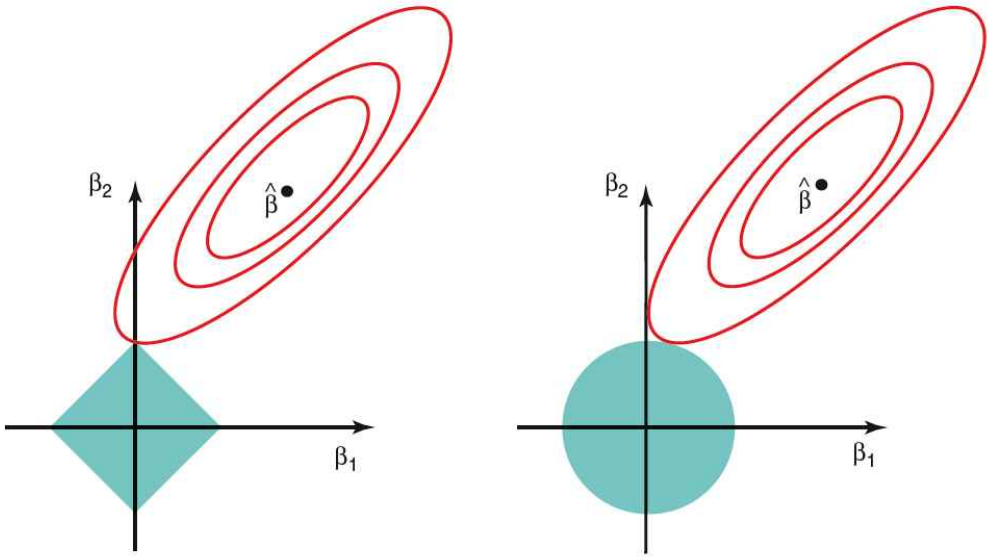
여기서 $\sum_j |\beta_j| \leq t$를 따른다. t(≥0)는 조절모수(tuning parameter)이고, 모수 t를 LASSO 추정치에 적용시켜 shrinkage의 양을 조절한다. $\hat{\beta}_j^o$를 완전 최소제곱(full least squares)추정치로 두고 $t_0 = \sum |\widehat{\beta_j^0}|$로 두면, t($\leq t_0$)의 값은 0으로 가며 shrinkage가 일어난다. 그리고 몇 개의 계수들이 정확하게 0의 값을 갖는다. 예를 들어 $t = t_0/2$라면, 효과는 사이즈 p/2의 best subset을 찾는 것과 유사하게 된다.

LASSO의 목표함수는 Ridge에서 β에 절대값을 적용한 것이다. 그 식은 식 (11)와 같다.

$$Q(\boldsymbol{\beta}) = (\boldsymbol{y} - X\boldsymbol{\beta})^T (\boldsymbol{y} - X\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{11}$$

LASSO가 0으로 가는 이유를 쉽게 이해하려면 <그림 1>을 보면 알 수 있다. <그림 1>은 p가 2인 경우인데, 위 방정식의 타원형 등고선이 보여진다. 그 중심은 OLS 추정치이며, 제약(constraint)지역은 마름모꼴로 나타나 있다. 이 LASSO의 해는 마름모와 먼저 만나는 등고선 위치에 있고, 이것이 코너에서 만나면 계수가 0이 된다. 그와 반대로 Ridge regression은 원의 형태로 코너에서 만날 수 없기 때문에 0의 계수를 가지

는 상황은 발생하지 않는다.



<그림 1> 왼쪽이 LASSO, 오른쪽이 Ridge 형태의 제약함수와 추정치의 등고선을 볼 수 있다(James et al., 2013).

3.3 LARS(Least Angle Regression)

Efron et al.(2004)에 의해 제안된 LARS 알고리즘은 수리적으로 간단하고 빠르게 계산이 가능한 stagewise 절차의 다른 형태이므로 매우 유용하고 간결한 모형 알고리즘이다. LASSO를 간단한 수정으로 이행할 수 있으며 회귀계수 절대값의 합을 제약조건으로 하는 OLS의 유용한 버전이다. 변수가 p개일 때, 오직 p개의 스텝이 요구된다. LARS 절차는 다음과 같다.

(1) 전진선택법과 같이 계수를 0으로 만들며 시작한다.
(2) 반응변수와 가장 관계가 높은 설명변수(x_1)를 찾는다.
(3) 다른 설명변수에서 가장 관계가 높은 설명변수(x_2)를 찾을 때까지 스텝을 계속 진행한다.
(4) 계속해서 x_1을 따르는 대신, LARS는 가장 높은 관계인 x_3을 찾을 때까지 이미 뽑힌 두 개의 변수들 사이의 예측방향이 등각이 되도록 진행한다.
(5) LARS는 least angle direction을 따라 다음 변수가 뽑힐 때까지 뽑힌 변수들 사이에서 등각을 이어나간다.

LARS 추정치는 $\hat{\boldsymbol{y}} = X\hat{\boldsymbol{\beta}}$이며 연속적인 스텝에서 모형에 변수를 하나씩 더하고 k 스텝 후에 0이 아닌 변수들이 k개가 되어 최종 모형은 k개의 모수를 가지게 된다. 즉 오직 선택된 변수의 개수만큼만 계산되기 때문에 LARS는 시간이 절약되는 것이 특징이다. 여기서 $\hat{\boldsymbol{\beta}}$는 반복을 통한 LASSO 추정치를 이용하여 구할 수 있다.

<그림 2>는 설명변수 $\boldsymbol{X} = (x_1, x_2)$가 2개인 경우의 알고리즘을 나타낸 것이다. $\hat{\mu}$을 추정치라 둘 때 $\boldsymbol{c}(\hat{\boldsymbol{\mu}})$을 잔차 추정치라 두고 식 (12)으로 나타낼 수 있다.

$$\hat{\boldsymbol{c}} = \boldsymbol{c}(\hat{\boldsymbol{\mu}}) = X'(y - \hat{\boldsymbol{\mu}}) \tag{12}$$

즉 $\hat{c_i}$은 변수 x_i와 잔차벡터의 관계를 부분적으로 나타낸다. 잔차 관계는 x_1, x_2에 의해 걸쳐진 선형공간 $\mathcal{L}(X)$에서 y의 추정치인 $\overline{\boldsymbol{y}_2}$에 의존한다.

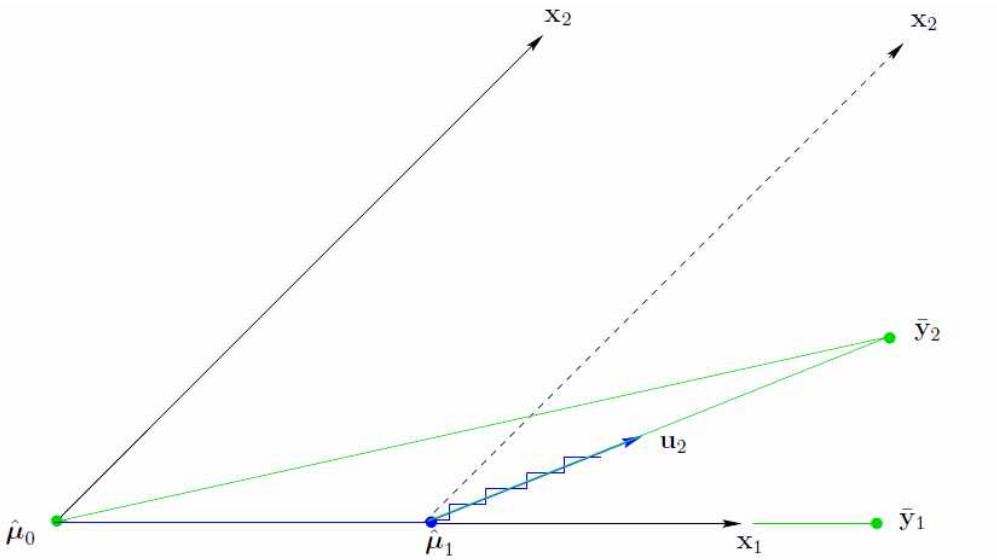
$$\boldsymbol{c}(\hat{\boldsymbol{\mu}}) = X'(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) = X'(\overline{\boldsymbol{y}}_2 - \hat{\boldsymbol{\mu}}) \tag{13}$$

이 알고리즘은 반응변수를 표준화했으므로 $\hat{\boldsymbol{\mu}}_0 = 0$에서 시작한다. <그림 2>에서 보면 $(\overline{\boldsymbol{y}}_2 - \widehat{\boldsymbol{\mu}_0})$에 의해 만들어진 각이 x_1, x_2에 의해 만들어진 각보다 더 작다. LARS는 x_1

의 방향에서 $\hat{\mu}_0$을 찾는다.

$$\hat{\mu}_1 = \hat{\mu}_0 + \hat{\gamma}_1 x_1 \tag{14}$$

$\hat{\gamma}_1$은 ε과 같은 값으로, 매우 많은 경우의 과정을 반복한다. 고전적 변수 선택방법인 전진적 선택방법은 $\mathcal{L}(x_1)$에서 $\bar{y}_1$와 같은 $\hat{\mu}_1$을 만들기에 충분히 큰 $\hat{\gamma}_1$으로 늘린다. LARS는 x_1과 x_2 관계와 같은 $\bar{y}_2 - \hat{\mu}$를 만드는 값인 $\hat{\gamma}_1$의 중간값(intermediate value)을 사용한다. 즉 이등분각을 사용하여 $c_1(\hat{\mu}_1) = c_2(\hat{\mu}_1)$를 성립시킨다.

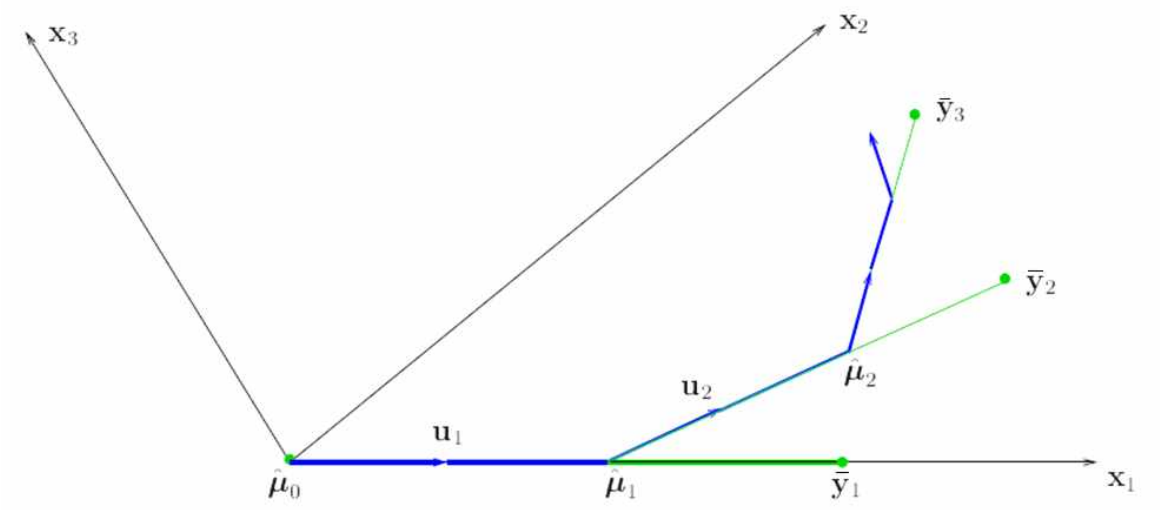


<그림 2> 설명변수가 2개일 경우의 LARS 알고리즘(Efron et al., 2004)

u_2를 2등분선에 따라 놓인 단위벡터로 두면, LARS 추정치는 식 (15)와 같이 된다.

$$\hat{\mu}_2 = \hat{\mu}_1 + \hat{\gamma}_2 u_2 \tag{15}$$

$\hat{\gamma}_2$는 설명변수가 2개인 상태에서 $\hat{\mu}_2 = \bar{y}_2$를 만드는 것으로 선택한다. 설명변수가 2개보다 많을 때는 $\hat{\gamma}_2$은 더 작아지며 <그림 3>과 같이 다른 방향을 따르게 된다.



<그림 3> 설명변수가 3개 이상일 경우의 LARS 알고리즘(Efron et al., 2004)

3.4 Elastic net

Zou and Hastie(2005)는 Naive elastic net을 제안하고 또한 그것을 보완한 Elastic net을 제안했다. LASSO는 p>n인 경우 over shrinkage 하는 경향이 있고, 변수의 그룹 내에 상관관계가 큰 경우 그룹으로부터 오직 하나의 변수만을 뽑아내는 단점이 있는데 이것을 보완하기 위해 Zou and Hastie(2005)는 먼저 Naive elastic net을 제안하였다. 자료의 관측치가 n이고 변수가 p개라 할 때, $\boldsymbol{y} = (y_1, \ldots, y_n)^T$는 반응변수가 되며 j=1,…, p 일 때, $\boldsymbol{x}_j = (x_1, \cdots, x_n)^t$ 이면 모형행렬(model matrix) $\boldsymbol{X} = (\boldsymbol{x}_1 | \ldots | \boldsymbol{x}_p)$가 된다. 설명변수와 반응변수를 표준화하면 식 (16)과 같은 식이 성립된다.

$$\sum_{i=1}^{n} y_i = 0, \ \sum_{i=1}^{n} x_{ij} = 0, \ \sum_{i=1}^{n} x_{ij}^2 = 1, \ \text{for } j = 1, 2, \ldots, p \tag{16}$$

$\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$, $|\boldsymbol{\beta}|^2 = \sum_{j=1}^{p} \beta_j^2$, $|\boldsymbol{\beta}|_1 = \sum_{j=1}^{p} |\beta_j|$일 때, 양수인 λ_1과 λ_2를 고정시키기 위한 Naive elastic net 기준을 식 (17)과 같이 정의한다.

$$L(\lambda_1, \lambda_2, \boldsymbol{\beta}) = |\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}|^2 + \lambda_2 |\boldsymbol{\beta}|^2 + \lambda_1 |\boldsymbol{\beta}|_1 \tag{17}$$

식 (18)는 Naive elastic net 추정치 $\hat{\boldsymbol{\beta}}$인 식 (17)을 최소화한 것이다.

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} L(\lambda_1, \lambda_2, \boldsymbol{\beta}) \tag{18}$$

이 절차를 penalized least square 방법이라 한다.

자동 변수 선택방법인 Naive elastic net은 LASSO의 한계를 극복할 수 있으나 그리 만족스러운 결과를 주지는 못하여 Naive라 불린다.

회귀 예측력 설정에서 정확한 penalization 방법은 잔차와 분산의 교환을 통해 좋은 예측력을 가진다. Naive elastic net 추정치를 구하는 두 가지 단계가 있는데 Ridge 회귀계수를 찾아 λ_2를 고정시키고 LASSO 해결법을 통해 LASSO와 같은 shrinkage를 하는 것이다. 이 방법에서 shrinkage를 2번 행하게 된다. 이 Double shrinkage는 LASSO나 Ridge shrinkage와 비교하여 불필요한 여분의 잔차를 가지게 되며 분산을

줄이는 것에 도움이 되지 않는다.

Elastic net 추정치는 주어진 자료가$(\boldsymbol{y},\boldsymbol{X})$이고 penalty 모수가 (λ_1,λ_2)이며 augmented 자료가 $(\boldsymbol{y}^*,\boldsymbol{X}^*)$ 일 때, LASSO 타입으로 풀이가 식 (19)과 같이 가능하다.

$$\hat{\beta}^* = \arg\min_{\beta^*} \|\boldsymbol{y}^* - \boldsymbol{X}^*\beta^*\|^2 + \frac{\lambda_1}{\sqrt{1+\lambda_2}}|\beta^*|_1 \tag{19}$$

Elastic net(corrected)추정치 $\hat{\beta}$는 식 (20)과 같이 정의된다.

$$\hat{\beta} = \sqrt{1+\lambda_2}\,\hat{\beta}^* \tag{20}$$

Naive elastic net의 추정치는 $\hat{\beta} = 1/\sqrt{(1+\lambda_2)}\,\hat{\beta}^*$이므로 다음과 같은 식 (21)가 성립할 수 있다.

$$\hat{\beta} = (1+\lambda_2)\hat{\beta} \tag{21}$$

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

3.5 Adaptive LASSO

Tibshirani(1996)에 의해 제안된 LASSO는 오라클 절차(oracle procedure)로 사용이 불가능하다(Zou, 2006). 그러나 점근적 설정은 소위 불공평하다고 일컫는데, 이는 penalty가 계수마다 같은 값으로 할당되기 때문이다. 그러므로 계수마다 다른 가중치(weight)를 주는 것을 고려할 수 있다. 중요한 변수의 shrinkage를 낮게 주고, 중요하지 않은 변수의 shrinkage 크게 주는 방법이다. Weighted LASSO를 고려해 보면 다음 식 (22)과 같다.

$$\arg\min_{\beta} \|\boldsymbol{y} - \sum_{j=1}^{p} \boldsymbol{x}_j \beta_j\|^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j| \tag{22}$$

여기서 조건 w_j는 알려진 가중치 벡터이다. 식 (22)에서 가중치가 자료에 의존적이고 교묘하게 선택됐다는 것을 알 수 있으며 weighted LASSO는 큰 오라클 성질(oracle properties)을 가질 수 있다. 이 새로운 방법들을 Zou(2006)가 제안하였고, Adaptive LASSO라 부른다. Adaptive LASSO 추정치 계산으로 LASSO를 풀며 효율적인 알고리즘을 사용할 수 있다.

제 4 장 실증분석

4.1 전립선(prostate)자료

전립선(prostate)자료는 Stamey et al.(1989)에 의해 연구된 자료로, 전립선 적출술을 받은 남성 97명의 의학 측정자료와 전립선 특이항원 양 사이의 상관관계를 검사한 것이다. large-p-small-n 형태의 자료에서의 적합도 중요하지만, 우선 large-n-small-p에서 적합이 잘 되어야 하고 본 논문의 결과와 선행 연구의 결과를 비교하기 위해 이 자료를 사용한다. 전립선 자료의 요인은 다음과 같다.

· log(암의 크기) (lcavol)
· log(전립선 무게) (lweight)
· 나이(age)
· log(전립선 비대종양의 양) (lbph)
· 암이 정낭에 침범한 확률 (svi)
· log(세균 침투량) (lcp)
· 글리손(gleason)점수 (gleason)
· 글리손 4점 또는 5점의 백분율 (pgg45)

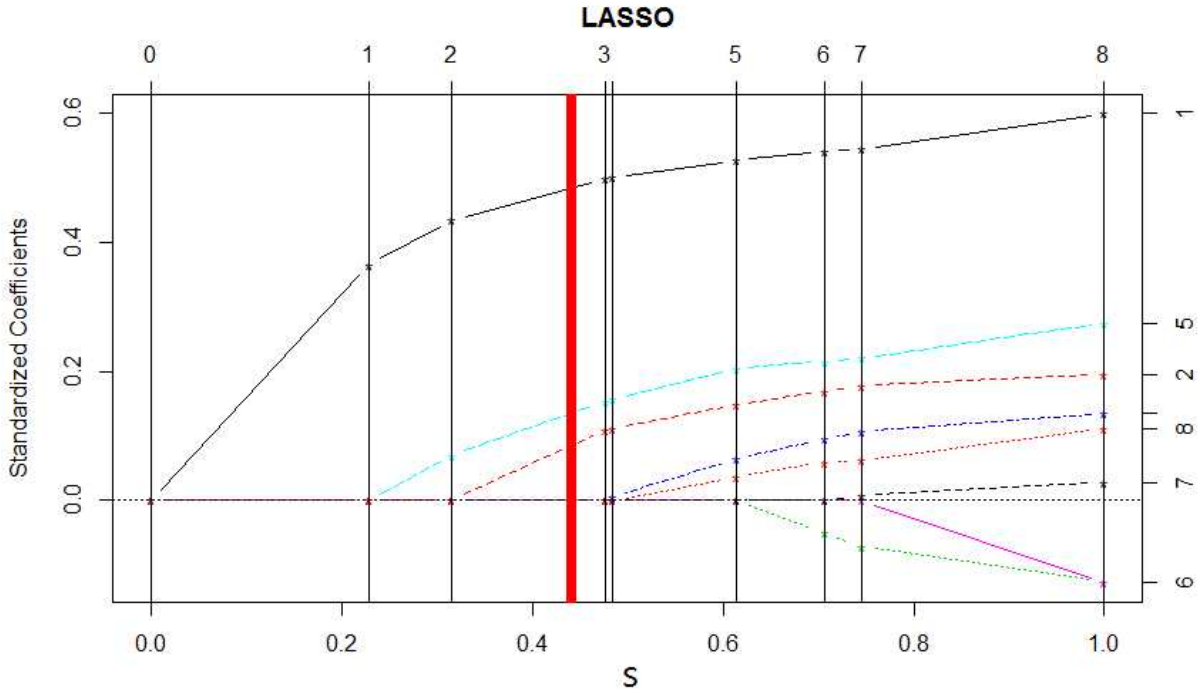
글리손 점수는 조직검사에서 암의 악성도를 숫자로 표현한 점수이며 6점 이하면 낮은 악성도를 띄는 암이다. 더빗-왓슨 통계량이 1.507로 자료는 독립성을 가지고 있다. 8개의 요인들과 반응변수를 표준화 시킨 후, 반응변수인 log(전립선 특이항원)에 선형모형으로 적합시켰다. R package에서 monomvn, lars, elasticnet, parcor 등을 이용하여 Ridge, LASSO, LARS, Adaptive LASSO, Elastic net의 결과를 비교하였다.

Tibshirani(1996)에 의한 LASSO 결과는 일반화 교차검증(generalized cross-validation)에 의해 $\hat{s}\,(=t/\sum|\beta_j^0|)$를 0.44로 고정하였을 때 <표 1>과 같이 나타나며 s는 scaled LASSO parameter이다.

<표 1> LASSO 결과

LASSO 결과(계수)	
lcavol	0.56
lweight	0.1
svi	0.16

LASSO에 의해 암의 크기(lcavol), 암이 정낭에 침범할 확률(svi), 전립선 무게(lweight) 3가지 변수가 선택되었는데, 암의 크기(lcavol) 계수가 가장 큰 것으로 보아, 전립선 특이항원에 가장 영향을 많이 끼치는 것은 암의 크기로 보인다. 그 다음은 암이 정낭에 침범할 확률, 전립선 무게 순이다. 변수가 선택된 순서는 다음 <그림 3>과 같다.



<그림 4> LASSO 계수가 선택되는 순서를 알 수 있으며, 계수의 크기가 어떻게 변하는지 알 수 있다. 빨간 선이 s가 0.44일 때이며, 3가지 변수가 선택되었다.

본 연구에서는 앞에서 설명한 다섯 가지 shrinkage 방법에 RSS에 의한 값으로 계수를 고정하여 구한다. 분석을 하기 전, 더빈-왓슨 통계량이 1.507이고 p-value가 0.006으로 독립성 가정을 할 수 없다. 그래서 80개의 자료를 임의로 선택하여 더빈-왓슨 통계량을 1.718, p-value를 0.904로 만든다. 자료는 n=80, p=8의 형태이고, 각 방법의 속도를 추가하였다. 그 결과는 <표 2>와 같다.

종합적인 결과로 계수를 비교하여 살펴보면 이 역시 암의 크기가 가장 영향력이 큰 것으로 나타났다. 그 다음으로 암이 정낭에 침범할 확률, 전립선 무게, 전립선 비대종양(lbph)의 양 순 등이었다. 가장 영향력이 작은 변수는 나이이다. 그 외에 세균침투, 글리손 점수, 글리손 4점 또는 5점의 백분율은 전립선 특이항원과 크게 관계가 없다는 해석을 할 수 있다. Ridge를 제외하고 LASSO와 LARS는 4개, Elastic net은 6개, Adaptive LASSO는 4개의 계수를 가지는 것으로 나타났다.

<표 2> 전립선 자료결과

변수 명	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
속도(ms)	12.146	1.804	1.704	2.441	1379.677

모형 선택기준인 RSS, AIC, BIC를 종합적으로 보았을 때 LARS의 결과가 가장 좋다. Ridge는 8개의 요인을 모두 사용하여 모형을 만듦으로 RSS값이 작지만 간결성을 충족하지 못하고, 속도가 느린 것을 알 수 있다. LASSO와 LARS는 앞에서 언급한 것과 같이 거의 비슷한 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. 최종 모형은 식 (23)과 같다.

$$\hat{y} = 0.522lcavol + 0.069weight + 0.028lbph + 0.164svi \qquad (23)$$

<표 3> 전립선 자료를 이용한 LASSO, LARS와 OLS 계수 비교

변수 명	LASSO	LARS	OLS
lcavol	0.522	0.522	0.574
lweight	0.069	0.069	0.118
lbph	0.028	0.028	0.127
svi	0.164	0.164	0.256

LASSO, LARS로 변수를 선택하고 OLS와 계수를 비교해보았다. LASSO, LARS가 추정한 값이 대체로 더 작으며 각 계수 크기의 순서는 같았다.

4.2 경제수치 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 세계 경제상황이 좋지 않다. 전 세계적으로 청년실업이 심각하며 빈곤층은 계속해서 늘고 있다. 이러한 경제상황에서 몇 가지 반응변수를 설정하여 어떤 변수가 반응변수에 크게 영향을 주며 유의한지에 대해 알아본다. 설명변수의 개수인 p를 늘리기 위해 가장 흔히 사용되는 자료인 월별자료를 사용하고 small-n-large-p 자료를 만들어 제조업 부문에서 부분적으로 제조업 업황실적 BSI, 제조업 생산지수, 제조업 가동률 실적수치를 반응변수로 두고 세부사항을 나누어 설명변수로 두었다. 이 분석을 통해 전반적으로 한국의 제조업에 가장 큰 영향을 주는 제조업은 어떤 것인지 알아보고 영향을 적게 주는 제조업은 어떤 것인지 알아본다.

그리고 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.

4.2.1 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, 100 이하인 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(박성빈, 박동화, 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. 결과는 <표 4>와 같다. 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. 종합적으로 보았을 때, 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. 결과가 가장 좋지 못한 것은 Elastic net으로 계수를 2개를 뽑은 Adaptive LASSO보다 RSS값도 낮은 것으로 나타났다. Adaptive LASSO는 조절변수의 가중치를 매번 구해야 하므로 이번 자료에서 역시 속도가 가장 느린 것으로 나타났다.

<표 4> 제조업 업황실적 BSI 자료결과

변수 명	Ridge	LASSO	LARS	Elastic net	Adaptive LASSO
ec1(식료품)	-0.015	0	0	0	0
ec2(음료)	0.043	0	0	0	0
ec3(섬유)	0.074	0	0	0	0
ec4(의복모피)	0.055	0	0	0	0
ec5(가죽·가방·신발)	0.035	0	0	0	0
ec6(목재·나무)	0.054	0	0	0	0
ec7(펄프·종이)	-0.035	0	0	0	0
ec8(인쇄·기록매체복제)	-0.02	0	0	0	0
ec9(석유정제·코크스)	-0.024	0	0	0	0
ec10(화학물질·제품)	0.092	0	0	0	0
ec11(의료물질·의약품)	0.02	0	0	0	0
ec12(고무·플라스틱)	-0.033	0	0	0	0
ec13(비금속광물)	-0.012	0	0	0	0
ec14(1차 금속)	0.111	0	0	0.096	0
ec15(금속가공)	0.03	0	0	0	0
ec16(전자·영상·통신장비)	0.093	0	0	0	0
ec17(의료·정밀기기)	0.058	0	0	0	0
ec18(전기장비)	0.058	0	0	0	0
ec19(기타 기계·장비)	0.094	0	0	0	0
ec20(자동차)	0.002	0	0	0	0
ec21(조선·기타운수)	0.03	0	0	0	0
ec22(가구)	-0.03	0	0	0	0
ec23(기타제품,담배)	0.095	0.029	0.029	0	
ec24(중소기업)	0.12	0.264	0.264	0.291	0.015
ec25(중화학공업)	0.104	0.075	0.075	0.132	
ec26(경공업)	0.131	0.515	0.515	0.322	0.92
ec27(수출기업)	0.032	0	0	0	0
ec28(내수기업)	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
속도(ms)	34.173	3.0516	2.3121	8.903	196.928

최종 모형은 식 (24)와 같다.

$$\hat{y} = 0.029ec23 + 0.264ec24 + 0.075ec25 + 0.515ec26 \quad (24)$$

<표 5> 제조업 업황실적 BIS 자료를 LASSO, LARS와 OLS 계수 비교

변수 명	LASSO	LARS	OLS
ec23(기타제품,담배)	0.029	0.029	0.097
ec24(중소기업)	0.264	0.264	0.479
ec25(중화학공업)	0.075	0.075	0.2
ec26(경공업)	0.515	0.515	0.331

OLS 추정치는 고차원에서는 구할 수 없어 LASSO, LARS 방법으로 변수를 선택하고, 선택된 변수로 OLS 값을 구하여 LASSO, LARS와 OLS 추정치를 비교해 보았다. 더빈-왓슨 통계량은 1.718이고, p-value는 0.476이다. 다른 방법들과 상이한 결과가 나타났으며 추정치의 값들이 대체로 크다.

4.2.2 제조업 생산지수

제조업 생산지수는 제조업에 대한 생산활동의 수준과 그 변동을 측정하기 위해 작성하는 지수이다(박성동 외 11인, 2015). 부가가치를 가중치로 두며 기준시점 고정 가중평균법(라스파이레스 산식)을 사용하여 계산하였다.

설명변수는 식료품, 음료, 담배 등 24가지이고, 반응변수는 그 24가지의 수치를 통합하여 나타낸 수치이다. 기간은 2014년 1년 동안을 12개월로 나타낸 것으로 12개의 반응변수를 가지고 자료는 n=12, p=24의 형태이다. 결과는 <표 6>과 같다. 유의한 계수의 개수는 Ridge를 제외하고 LASSO는 6개, LARS는 6개, Elastic net 5개, Adaptive LASSO는 1개로 나타났다. 방법마다 각각 차이가 나타나지만 대체적으로 비슷하게 변수계수를 설정한다. 반응변수에 가장 크게 영향을 주는 변수는 코크스, 연탄 및 석유정제(ec55), 기타운송장비(ec67), 가구(e68) 순으로 나타났다. 이 결과는 전반적인 제조업 생산지수와 가장 비슷하게 수치변화를 가지는 것이 코크스, 연탄 및 석유정제(ec55), 기타운송장비(ec67), 가구(e68) 순이라는 것과 같다. 종합적으로 보았을 때, 유의한 계수는 이 역시 LASSO와 LARS의 결과가 가장 좋게 나타났고, 속도는 LARS가 가장 빨랐다. 이번 자료에서는 Adaptive LASSO의 결과가 가장 좋지 않다. 하나의 변수만 뽑았으며 속도도 느리게 나타났다.

<표 6> 제조업 생산지수 자료결과

변수 명	Ridge	LASSO	LARS	Elastic net	Adaptive LASSO
ec46(식료품)	0.079	0.049	0.049	0.035	0
ec47(음료품)	0.011	0	0	0	0
ec48(담배)	0.025	0	0	0	0
ec49(섬유제품)	-0.039	0	0	0	0
ec50(의복,악세서리)	0.016	0	0	0	0
ec51(가죽·가방·신발)	-0.044	-0.183	-0.183	0	0
ec52(목재·나무제품)	0.042	0	0	0	0
ec53(펄프·종이제품)	0.037	0	0	0	0
ec54(인쇄, 기록매체복제)	0.001	0	0	0	0
ec55(코크스, 연탄)	0.145	0.55	0.55	0.349	0.247
ec56(화학물질)	0.077	0	0	0.039	0
ec57(의료용 물질)	-0.031	0	0	0	0
ec58(고무·플라스틱)	-0.061	0	0	0	0

ec59(비금속광물제품)	-0.023	0	0	0	0
ec60(1차금속제품)	0.086	0.067	0.067	0.039	0
ec61(금속가구제품)	0.027	0	0	0	0
ec62(전자부품)	0.067	0	0	0	0
ec63(의료,정밀기기)	0.055	0	0	0	0
ec64(전기장비)	-0.031	0	0	0	0
ec65(기타기계 및 장비)	-0.035	0	0	0	0
ec66(자동차·트레일러)	-0.066	0	0	0	0
ec67(기타운송장비)	-0.113	-0.228	-0.228	-0.196	0
ec68(가구)	-0.066	-0.113	-0.113	0	0
ec69(기타제품)	-0.025	0	0	0	0
RSS	4.808	3.45	3.45	5.275	8.005
AIC	68.845	28.864	28.864	69.957	72.962
BIC	80.967	32.258	32.258	82.079	84.599
속도(ms)	44.513	5.259	2.838	7.7829	1304.881

최종 모형은 다음 식 (25)와 같다.

$$\hat{y} = 0.049ec46 - 0.183ec51 + 0.55ec55 + 0.067ec60 - 0.228ec67 - 0.113ec68 \quad (25)$$

OLS 추정치는 고차원에서는 구할 수 없어, LASSO, LARS 방법으로 변수를 선택하고, 그 값을 OLS와 비교하기 위해 추가하였다. 더빈-왓슨 통계량은 1.5289이고, p-value는 0.6268이다. 다른 방법들과 상이한 결과가 나타났으며 추정치의 값들이 대체로 크다.

<표 7> 제조업 생산지수 자료를 이용한 LASSO, LARS와 OLS 계수 비교

변수 명	LASSO	LARS	OLS
ec46(식료품)	0.049	0.049	0.058
ec55(코크스, 연탄)	0.55	0.55	-0.409
ec51(가죽·가방·신발)	-0.183	-0.183	0.636
ec60(1차금속제품)	0.067	0.067	0.233
ec67(기타운송장비)	-0.228	-0.228	-0.265
ec68(가구)	-0.113	-0.113	-0.210

4.2.3 제조업 가동률지수

제조업 가동률지수는 제조업 품목별 가동률지수(생산실적/생산능력)에 가중치(부가가치)를 적용하여 합산한 지수이다(박성동 외 11인, 2015).

설명변수는 식료품, 음료, 담배 등 21가지이고, 반응변수는 그 21가지의 수치를 통합하여 나타낸 수치이다. 기간은 2014년 1년 동안을 12개월로 나타낸 것으로 12개의 반응변수를 가지고 자료는 n=12, p=21의 형태이다. 결과는 <표 8>과 같다. 유의한 계수의 개수는 Ridge를 제외하고 LASSO와 LARS는 6개를 뽑았고, Elastic net은 5개, Adaptive LASSO는 3개이다.

결과 값은 약간 상이하였다. 종합적으로 RSS와 AIC, BIC의 값으로 비교한 결과, LASSO와 LARS의 결과가 가장 좋게 나타났고, 속도는 LARS가 가장 빨랐다. 반응변수에 가장 크게 영향을 주는 변수는 섬유제품(ec119), 자동차 및 트레일러(ec134), 화학물질 및 화학제품(ec125) 부문 순으로 나타났다. 이 결과는 한국의 제조업 가동률지수와 가장 크게 영향을 미치는 부문이 섬유제품, 자동차 및 트레일러, 화학물질 및 화학제품 순이라는 것과 같다. 제조업 가동률 중 섬유제품업의 가동률이 가장 많은 부분을 차지할 것으로 예상된다. 다른 방법보다 Elastic net의 결과가 가장 좋지 않았는데, 선택된 계수를 보아도 다른 방법으로 선택된 계수들과 상당히 상이한 것을 볼 수 있다. 속도는 이번에도 Adaptive LASSO가 가장 느린 것으로 나타났다.

<표 8> 제조업 가동률지수 자료결과

변수 명	Ridge	LASSO	LARS	Elastic net	Adaptive LASSO
ec116(식료품)	0.037	0	0	0	0
ec117(음료품)	-0.01	0	0	0	0
ec118(담배)	0.024	0	0	0	0
ec119(섬유제품)	0.062	0.367	0.367	0.156	0.429
ec120(의복,악세서리)	-0.018	0	0	0	0
ec121(가죽·가방·신발)	0.006	0	0	0	0
ec122(목재제품)	0.036	0	0	0	0
ec123(펄프, 종이)	0.059	0	0	0.066	0
ec124(코크스, 연탄)	0.032	0	0	0	0
ec125(화학물질)	0.096	0.111	0.111	0	0.047
ec126(고무,플라스틱)	0.047	0	0	0.066	0
ec127(비금속광물제품)	0.078	0	0	0	0

ec128(제1차금속제품)	0.064	0	0	0	0
ec129(금속가공제품)	0.051	0	0	0	0
ec130(전자부품)	0.104	0.048	0.048	0	0
ec131(의료,정밀기기)	0.096	0.08	0.08	0.076	0
ec132(전기장비)	0.054	0.043	0.043	0	0
ec133(기타기계)	0.098	0	0	0	0
ec134(자동차·트레일러)	0.134	0.313	0.313	0.116	0.489
ec135(기타운송장비)	0.09	0	0	0	0
ec136(가구)	0.077	0	0	0	0
RSS	0.017	0.31	0.31	3.62	0.478
AIC	-4.648	-0.039	-0.039	59.437	33.152
BIC	6.019	3.356	3.356	70.105	43.335
속도(ms)	29.991	3.322	1.983	6.550	1290.799

최종 모형은 식 (26)과 같다.

$$\hat{y} = 0.367ec119 + 0.111ec125 + 0.048ec130 + 0.08ec131 + 0.043ec132 + 0.313ec134 \quad (26)$$

OLS 추정치는 고차원에서는 구할 수 없어, LASSO, LARS 방법으로 변수를 선택하고, 값을 비교하기 위해 추가하였다. 더빈-왓슨 통계량은 1.2507이고, p-value는 0.2532이다. 다른 방법들과 상이한 결과가 나타났으며 추정치의 값들이 대체로 크다.

<표 9> 제조업 가동률지수 자료를 이용한 LASSO, LARS와 OLS 계수 비교

변수 명	LASSO	LARS	OLS
ec119(섬유제품)	0.367	0.367	0.323
ec125(화학물질)	0.111	0.111	0.177
ec130(전자부품)	0.048	0.048	0.114
ec131(의료,정밀기기)	0.08	0.08	0.021
ec132(전기장비)	0.043	0.043	0.117
ec134(자동차·트레일러)	0.313	0.313	0.399

4.2.4 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(김보경, 김대유, 2015).

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

$$※\ S_i^{2012} = \frac{(P_i^{2012} Q_i^{2012})}{\sum(P_i^{2012} Q_i^{2012})}$$

* P:가격, Q:수량, S:가중치, 2012:가격 및 가중치기준시점, t:가격조사시점, i:품목

설명변수는 100대 통계지표와 제조업산업에서 몇 가지 세분화된 월별자료로 개수는 196개이다. 2010년 1월부터 2014년 12월까지 총 60개의 자료로 n=60, p=196 형태의 자료이다.

결과는 <표 10>과 같은데, 모든 변수를 선택하는 Ridge를 제외하고 나머지 네 개의 방법 중 계수를 0으로 채택했을 경우는 삭제하였다. RSS, AIC와 BIC를 종합하여 보았을 때 결과는 LASSO의 결과가 좋으나 LARS와 크게 차이가 없는 것으로 보인다. 유의한 계수의 개수는 Ridge를 제외하고 LASSO는 29개, LARS는 30개, Elastic net 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 나타내며 그룹화 되지 않은 모형이므로 이러한 결과가 나타났다고 할 수 있다. 종합적으로 보았을 때 소비자 물가지수에 가장 영향력이 큰 변수는 생산자 물가지수(ec177), 근원 인플레이션지수(ec176), 국고채 발행액(ec174) 순이었다. 소비자 물가지수에 가장 영향력이 큰 것은 생산자 물가지수임에 틀림없다. 그리고 생산자 물가지수와 소비자 물가지수는 비슷한 동향을 가진다고 할 수 있다. 근원 인플레이션지수와 국고채 발행액 역시 생산자 물가지수와 비슷한 동향을 가지며 소비자 물가지수에 큰 영향을 끼친다.

<표 10> 소비자 물가지수 자료결과

변수 명		Ridge	LASSO	LARS	Elastic net	Adaptive LASSO
업황실적	ec6(목재·나무)	-0.078	-0.052	-0.052	-0.061	-0.099
	ec7(펄프·종이)	0.051	0.052	0.052	0.004	0
	ec8(인쇄·기록매체복제)	0.117	0.036	0.036	0	0
	ec12(고무·플라스틱)	-0.049	-0.032	-0.033	-0.007	0
	ec13(비금속광물)	-0.072	-0.098	-0.099	-0.035	-0.028
	ec17(의료·정밀기기)	0.057	0.007	0.008	0	0
	ec20(자동차)	-0.06	-0.109	-0.109	-0.09	-0.028
	ec35(운수업)	0.089	0.02	0.021	0	0
	ec36(숙박업)	-0.042	0	0	-0.013	0
	ec43(출판,영상,)	-0.065	-0.044	-0.044	0	0
ec45(경기선행지수순환변동치)		-0.154	-0.034	-0.034	-0.004	0
생산지수	ec49(섬유제품)	0.024	0	0	0.004	0
	ec54(인쇄,기록매체복제)	-0.025	0	0	-0.024	0
	ec61(금속가구제품)	-0.022	0	0	-0.04	0
	ec64(전기장비)	-0.047	0	0	-0.009	0
출하지수	ec78(인쇄, 기록매체)	-0.024	0	0	-0.02	0
	ec85(금속가공제품)	-0.023	-0.032	-0.034	-0.042	0
	ec87(의료,정밀 기기)	-0.042	-0.022	-0.023	-0.009	0
	ec88(전기장비)	-0.044	0	0	-0.032	0
재고지수	ec94(식료품제조업)	0.003	0	0	-0.006	0
	ec96(담배제조업)	-0.055	-0.081	-0.083	0	0
	ec99(가죽, 가방, 신발)	-0.072	-0.107	-0.107	-0.16	0
ec129(금속가공업 가동률지수)		-0.03	0	0.006	-0.05	0
ec139(자동차부품 판매업지수)		-0.034	0	0	-0.021	0
도소매업지수	ec143(기계장비 도매업)	-0.063	-0.082	-0.083	-0.017	-0.004
	ec146(음식료 소매업)	0.016	0.033	0.034	0.004	0
	ec147(통신장비 소매업)	-0.061	-0.04	-0.04	-0.043	0
	ec150(문화용품 소매업)	0.089	0.005	0.004	0	0
ec162(실업률)		0.065	0.039	0.04	0.004	0
ec165(예금은행대출금)		-0.062	-0.115	-0.116	-0.139	0
ec167(예금은행 대출금리)		0.049	0	0	0.037	0
ec168(코스닥지수)		0.055	0.022	0.021	0.025	0

ec170(고객예탁금)	0.019	0	0	0.029	0
ec171(채권거래대금)	0.066	0.041	0.041	0	0
ec174(국고채발행액)	0.112	0.056	0.055	0.082	0.102
ec175(어음부도율)	-0.067	-0.039	-0.04	0	0
ec177(생산자물가지수)	0.107	0.33	0.329	0.338	0.345
ec180(근원인플레이션지수)	0.113	0.285	0.286	0.176	0.291
ec187(직접투자(자산))	0.082	0.036	0.036	0	0
ec191(순상품교역조건지수)	-0.095	0	0	-0.01	0
ec192(원/미국달러)	-0.033	0	0	-0.015	0
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
속도(ms)	166.206	42.174	22.504	297.038	1935.02

최종 모형 식은 식 (27)과 같다.

$$\begin{aligned}\hat{y} = & -0.052ec6 + 0.052ec7 + 0.036ec8 - 0.033ec12 - 0.099ec13 + 0.008ec17 \\ & - 0.109ec20 + 0.021ec35 - 0.044ec43 - 0.034ec45 - 0.034ec78 - 0.023ec88 \\ & - 0.083ec96 - 0.107ec99 + 0.006ec129 - 0.083ec143 + 0.034ec146 \\ & - 0.04ec147 + 0.004ec150 + 0.04e162 - 0.116ec165 + 0.021ec168 + 0.041ec171 \\ & + 0.055ec174 - 0.04ec175 + 0.329ec177 + 0.286ec180 + 0.036ec187\end{aligned} \quad (27)$$

OLS 추정치는 고차원에서는 구할 수 없어, LARS 방법으로 변수를 선택하고, 값을 비교하기 위해 추가하였다. 더빈-왓슨 통계량은 1.9173이고, p-value는 0.5398이다. 결과는 많이 상이하였고 대체로 추정치가 줄어들었다.

<표 11> 소비자 물가지수 자료를 이용한 LARS와 OLS 계수 비교

변수 명		LARS	OLS
업황실적	ec6(목재·나무)	-0.052	-0.056
	ec7(펄프·종이)	0.052	0.109
	ec8(인쇄·기록매체복제)	0.036	0.089
	ec12(고무·플라스틱)	-0.033	-0.039
	ec13(비금속광물)	-0.099	-0.087
	ec17(의료·정밀기기)	0.008	0.057
	ec20(자동차)	-0.109	-0.147
	ec35(운수업)	0.021	0.109

	ec43(출판,영상)	-0.044	-0.076
ec45(경기선행지수순환변동치)		-0.034	-0.178
출하지수	ec78(인쇄, 기록매체)	-0.034	0.025
	ec88(전기장비)	-0.023	-0.318
재고지수	ec96(담배제조업)	-0.083	-0.164
	ec99(가죽, 가방, 신발)	-0.107	-0.055
ec129(금속가공업 가동률지수)		0.006	0.039
도소매업지수	ec143(기계장비 도매업)	-0.083	-0.209
	ec146(음식료 소매업)	0.034	0.045
	ec147(통신장비 소매업)	-0.04	-0.066
	ec150(문화용품 소매업)	0.004	0.192
ec162(실업률)		0.04	0.177
ec165(예금은행대출금)		-0.116	-0.023
ec168(코스닥지수)		0.021	0.044
ec171(채권거래대금)		0.041	0.155
ec174(국고채발행액)		0.055	0.103
ec175(어음부도율)		-0.04	-0.132
ec177(생산자물가지수)		0.329	0.240
ec180(근원인플레이션지수)		0.286	0.309
ec187(직접투자(자산))		0.036	0.144

제 5 장 결론

본 논문에서 Ridge, LASSO, LARS, Elastic net, Adaptive LASSO 등 다섯 가지 shrinkage 방법을 이용하여 실제 자료를 고차원 선형모형에 적합하였고, RSS, AIC, BIC 값으로 다섯 가지 shrinkage 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 제안되었다. 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, 속도도 Adaptive LASSO가 가장 느렸다. 왜냐하면 조절변수의 가중치를 변수마다 매번 구해야하는 작업 때문이다.

참 고 문 헌

김기영, 전명식, 강현철, 이성건(2012). *예제를 통한 회귀분석. 자유아카데미.*

김병천(2000). *통계학을 위한 행렬대수학.* 자유아카데미.

김보경, 김대유(2015). *2015년 3월 소비자물가동향.* 통계청.

김창주(2013). *변수선택방법과 판정보류 옵션을 적용한 분류방법.* 호서대학교.

박성동, 전백근, 박병선, 박인천, 정환걸, 이복현, 한시문, 최정수, 양모승, 이무영, 문권순, 박원란(2015). *2015년 3월 산업활동동향.* 통계청.

박성빈, 박동화(2015). *2015년 4월 기업경기실사지수(BSI) 및 경제심리지수(ESI).* 한국은행.

이재은(2012). *분류분석에서의 이산화를 통한 변수선택에 관한 연구.* 성균관대학교.

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. and Tibshirani, R.(2004). *Least Angle Regression*. Annals of Statistics, 32, 407-451.

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

James, R., Witten, D., Hastie, T. and Tibshirani, R.(2013). *An Introduction to Statistical Learning*. Springer.

Mallick, H. and Yi, N.(2013). *Bayesian Methods for High Dimensional Linear Models.* Journal of Biometrics & Biostatistics, S1, 005. doi, 10.4172/2155-6180.S1-005.

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. and Yang, N.(1989). *Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. Radical prostatectomy treated patients.* Journal of Urology, 141, 5, 1076-1083.

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