응용통계연구 (2010)
23(1), 207–218

일반화가법모형에서 축소방법의 적용연구

기승도[1] · 강기훈[2]

[1]한국외국어대학교 정보통계학과, 보험연구원; [2]한국외국어대학교 정보통계학과

(2010년 1월 접수, 2010년 2월 채택)

요 약

일반화가법모형은 기존 선형회귀모형의 문제점을 대부분 해결한 통계모형이지만 의미있는 독립변수의 수를 줄이는 방법이 적용되지 않을 경우 과대적합 문제가 발생할 수 있다. 그러므로 일반화가법모형에서 변수 축소방법을 적용하는 연구가 필요하다. 회귀분석에서 변수 축소방법으로 최근에는 Lasso 계열의 접근법이 연구되고 있다. 본 연구에서는 활용성이 높은 통계모형인 일반화가법모형에 Lasso 계열의 모형 중에서 Group Lasso와 Elastic net 모형을 적용하는 방법을 제시하고 이들의 해를 구하는 절차를 제안하였다. 그리고 제안된 방법을 모의실험과 실제자료인 회계년도 2005년 자동차보험 자료에 적용을 통해 비교하여 보았다. 그 결과 본 논문에서 제안한 Group Lasso와 Elastic net을 이용하여 변수 축소를 통한 일반화가법모형이 기존의 방법보다 더 나은 결과를 제공하는 것으로 분석되었다.

주요용어: 가법모형, Lasso, Group lasso, Elastic net.

1. 서론

정규분포를 가정한 전통적인 회귀모형(또는 선형모형)은 종속변수에 대한 분포 가정을 다양화 한 일반화선형모형으로 발전하였고, 종속변수와 설명변수의 관계가 선형이 아닌 비모수적 함수인 경우까지 확장되었다. 비모수적 회귀모형에서 차원이 높은 경우에 발생하는 차원의 저주(Curse of Dimensionality) 문제를 해결하기 위한 일환으로 개발된 가법모형(Additive Model)은 종속변수를 각 개별 설명변수들만의 비모수적 함수의 합으로 표현하는 방법이다. 즉, 이 방법은 종속변수를 설명변수의 가법(Additive) 관계로 설명할 수 있다는 것을 상정한 모델이다. 따라서, 가법모형은 독립변수가 p개이고, 자료의 수가 n개인 경우에 다음과 같은 식으로 표현할 수 있다.

$$y_i = \sum_{k=1}^{p} f_k(x_{ik}) + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{1.1}$$

여기서, y_i는 종속변수의 i번째 관측치이며, x_{ik}는 k번째 독립변수의 i번째 관측치, f_k는 k번째 독립변수에 의존하는 회귀함수를 나타낸다. ε_i는 오차항으로 x와 독립이고, $E(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$이다.

가법모형의 경우에도 선형모형에서처럼 정규 확률변수와는 다른 여러가지 형태의 종속변수에 대해 적용할 수 있는 방법이 필요하였는데 이러한 필요에 따라 개발된 것이 일반화가법모형(Generalized Additive Model; GAM)이다. 일반화가법모형은 Hastie와 Tibshirani (1986)가 기존의 일반화선형모형과 가법모형을 결합하여 처음으로 개발·제시하였다. 일반화가법모형에서 종속변수에 대한 분포가정 측면

이 논문은 2009년 한국외국어대학교 학술연구비 지원에 의해 이루어졌음.
[2]교신저자: (449-791) 경기도 용인시 처인구 모현면, 한국외국어대학교 정보통계학과, 부교수.
E-mail: khkang@hufs.ac.kr

은 일반화선형모형과 동일하고, 모형식은 다음과 같이 표현할 수 있다.

$$G(\mu_i) = \sum_{k=1}^{p} f_k(x_{ik}), \tag{1.2}$$

여기서, G는 연결함수(link function)이고, μ_i는 $E(y_i)$이며, y_i는 지수족(exponential family) 분포를 따른다.

일반화가법모형은 기존 선형회귀모형의 문제점을 대부분 해결한 통계모형이지만 의미있는 독립변수의 수를 줄이는 방법이 적용되지 않을 경우 과대적합(overfitting) 문제가 발생할 수 있다. 그러므로 일반화가법모형에서 변수 축소방법(shrinkage method)을 적용하는 연구가 필요하다. 최근 변수축소 분야에서는 Lasso 계열의 연구를 기존모형에 적용하는 방법에 대한 연구가 지속되고 있다. 예를 들어 Lasso를 로지스틱 회귀분석에 적용한 연구로는 Lokhorst (1999), Roth (2004), Shevade와 Keerthi (2003), Genkin 등 (2007)이 있고, 다항 로지스틱 회귀분석에 확대 적용하는 연구로는 Krishnapuram 등 (2005)이 있다. 그러나, Lasso는 독립변수가 가변수(dummy variable) 등을 사용한 범주형인 경우에는 바로 적용하기 어려운 단점이 있었고, 이러한 것을 해결한 방법이 Yuan과 Lin (2006)이 제시한 Group Lasso이다. 이런 연구결과를 바탕으로 Kim 등 (2006)은 Group Lasso를 로지스틱 회귀분석에 적용하는 방법을 연구하였고, 모델 적용상의 penalized constrained 최적화 문제를 해결하기 위하여 gradient descent algorithm을 제시하였다. Meier 등 (2008)은 Group Lasso를 로지스틱 회귀분석에 적용하면서 발생하는 penalized constrained 문제를 더욱 쉽게 해결할 수 있는 방법을 제안하였다. Meier 등 (2008)이 제시한 방법은 자료의 양(n)과 독립변수의 수(p)가 많은 경우에 효율적으로 해를 찾을 수 있는 방법이다.

이에 본 연구에서는 일반화가법모형의 활용성 및 Lasso 계열의 변수 축소방법의 유용성에 주목하고, 일반화가법모형에 Lasso 계열의 변수 축소방법을 적용하여 효율을 파악하고자 한다. 이를 위해 Lasso 계열의 방법들 중 유용한 Group Lasso 및 Elastic net 등을 일반화가법모형의 변수 축소에 적용하는 방법을 살펴볼 것이다. 일반화가법모형에서 변수 축소방법과 관련해서는 아직까지 Lasso 계열을 접목시킨 결과는 없는 것으로 파악된다.

본 논문은 총 5절로 구성되어 있다. 2절에서는 변수 축소방법에서 Group Lasso와 Elastic net에 관해 서술하였으며, 3절에서는 본 논문에서 고려하는 일반화가법모형에서 변수 축소방법에 대해 설명하였다. 4절에서는 모의실험과 실제 자동차보험 자료를 이용하여 본 논문에서의 접근방법을 이용한 실증분석 결과를 제시하였으며, 5절은 분석한 결과를 바탕으로 결론을 도출하고 토의하였다.

2. 변수 축소방법

2.1. Group Lasso 모형

Lasso 모형은 변수 축소의 측면에서 기존 Ridge Regression이 의미 없는 변수에 대한 계수값을 작게 하지만 완전히 0으로 하지 못하는 문제점을 해소하였다. 그러나 독립변수간 상관관계가 높은 Group 변수인 경우 Lasso 모형은 예측력이 떨어지는 단점이 있다. Yuan과 Lin (2006), Zhao 등 (2006)은 Lasso 모형의 문제를 해결하는 방법으로 Group Lasso 모형을 제시하였다.

Yuan과 Lin (2006)의 Group Lasso를 설명하기 위한 Group 자료(범주형 자료)에 대한 회귀 모형식은 다음과 같다.

$$\mathbf{y} = \sum_{j=1}^{J} \mathbf{X}_j \boldsymbol{\beta}_j + \boldsymbol{\varepsilon}, \tag{2.1}$$

여기서, $\mathbf{y}$는 $n \times 1$ 벡터이고, $j = 1, \ldots, J$에 대해 $\mathbf{X}_j$는 j번째 요인에 대응되는 $n \times p_j$ 행렬이며, $\boldsymbol{\beta}_j$는 크기가 p_j인 계수벡터, $\boldsymbol{\varepsilon}$은 $n \times n$ 단위행렬 I에 대해 $N(0, \sigma^2 I)$을 따른다. 편의상 절편을 생략하기 위해 종속변수와 독립변수에 대해 각각 평균을 빼주었다고 하자.

벡터 $\boldsymbol{\eta} \in R^d$, $d \geq 1$와 $d \times d$ 양정치행렬(positive definite matrix) $\mathbf{K}$에 대해, $\|\boldsymbol{\eta}\|_K = (\boldsymbol{\eta}^T \mathbf{K} \boldsymbol{\eta})^{1/2}$로 나타내기로 하자. 편의상, $\|\boldsymbol{\eta}\| = \|\boldsymbol{\eta}\|_{I_d}$라고 하면, 양정치행렬 $\mathbf{K}_1, \ldots, \mathbf{K}_J$가 주어졌을 때, Group 자료(범주형 자료)의 모형식에서 각 범주의 계수 추정치를 구하는 방법인 Group Lasso 모형식은 다음 식 (2.2)와 같다.

$$\hat{\boldsymbol{\beta}}_{GLasso}(\lambda) = \text{argmin} \left(\frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{J} \|\boldsymbol{\beta}_j\|_{K_j} \right). \tag{2.2}$$

식 (2.2)에서 각 그룹에 포함된 인자의 수가 하나로 동일하다면, 즉, $p_1 = p_2 = \cdots = p_J = 1$이면, Group Lasso와 Lasso는 동일한 모형이다. Group Lasso의 벌점(penalty) 부분은 Lasso의 Ł$_1$ 벌점과 Ridge regression의 Ł$_2$ 벌점의 중간형태이다. 따라서 Group Lasso는 Lasso보다 더 포괄적인 모형이라 할 수 있다. $J = 2$인 경우에 Group Lasso의 해인 계수 추정치들의 기하적인 의미에 대해서는 Yuan과 Lin (2006)의 그림 1을 참조하기 바란다.

Bakin (1999)은 식 (2.2)의 $\hat{\boldsymbol{\beta}}_{GLasso}(\lambda)$를 구하는 방법으로 순차적으로 최적해를 구하는 알고리즘을 제안하였다. Yuan과 Lin (2006)도 Group Lasso의 해를 구할 수 있는 알고리즘을 제안하였는데, 이 알고리즘은 Fu (1998)가 제안한 shooting 알고리즘을 확장한 것으로 Karush-Kuhn-Tucker 조건을 이용하여 최적해가 되기 위한 필요충분 조건을 유도한 결과로 산출된 것이다.

2.2. Elastic net 모형

Elastic net은 Zou와 Hastie (2005)가 제시한 Lasso계열의 모형으로 Group Lasso 및 Hierachical Lasso (Zhou와 Zhu, 2007)와 마찬가지로 Lasso의 L_1과 Ridge regression의 L_2 벌점을 결합한 것이다. Elastic net은 Lasso와 같이 자동적으로 변수 축소가 가능하고 상관관계가 높은 Group 자료에도 적용하여 해를 구할 수 있다. 이러한 Elastic net 모형의 의미를 파악하기 위하여 앞에서와 같이 다음과 같은 자료 모형을 가정해보자. 자료가 n개의 관측값과 J개의 독립변수로 구성되어 있다고 하자. 즉, $\mathbf{y} = (y_1, \ldots, y_n)^T$를 종속변수라고 하고, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_J)$, $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^T$, $j = 1, \ldots, J$라 하자. 그리고 자료는 위치(location)와 산포(scale)에 대해 표준화된 것이라고 가정하자. 즉, $\sum_{i=1}^{n} y_i = 0$, $\sum_{i=1}^{n} x_{ij} = 0$ 그리고 $\sum_{i=1}^{n} x_{ij}^2 = 1$, $j = 1, \ldots, J$이다.

음이 아닌 고정된 값 λ_1과 λ_2에 대해, 단순한 (naive) Elastic net의 목적함수 식은 다음과 같이 정의된다.

$$L(\lambda_1, \lambda_2, \boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_2 \|\boldsymbol{\beta}\|^2 + \lambda_1 \|\boldsymbol{\beta}\|_1, \tag{2.3}$$

여기서, $\|\boldsymbol{\beta}\|^2 = \sum_{j=1}^{J} \beta_j^2$, $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{J} |\beta_j|$이다.

단순한(naive) Elastic net의 해($\hat{\boldsymbol{\beta}}$)는 식 (2.3)을 최소화하여 얻을 수 있다. 즉,

$$\hat{\boldsymbol{\beta}}_{Enet} = \text{argmin}\{L(\lambda_1, \lambda_2, \boldsymbol{\beta})\} \tag{2.4}$$

그런데 식 (2.3)에서 $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$라고 하면, 식 (2.4)의 최적화 문제는 적당한 t에 대해 $(1 - \alpha)\|\boldsymbol{\beta}\|_1 + \alpha\|\boldsymbol{\beta}\|^2 \leq t$라는 제약조건 하에서 다음의 최적화 문제와 동일하게 된다.

$$\hat{\boldsymbol{\beta}}_{Enet} = \text{argmin}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2,$$

여기서 $(1-\alpha)\|\boldsymbol{\beta}\|_1 + \alpha\|\boldsymbol{\beta}\|^2$가 Elastic net의 벌점에 해당된다. 이것은 Lasso와 Ridge regression의 벌점의 볼록합(convex combination)임을 쉽게 알 수 있다. 따라서 $\alpha = 1$이면 단순한 Elastic net은 Ridge regression이 되고, $\alpha = 0$이면 Lasso가 된다. 그러므로 Elastic net은 $0 < \alpha < 1$인 경우에 대한 것이라 볼 수 있다. Elastic net의 해를 구하는 절차와 그룹핑 효과에 대한 것은 Zou와 Hastie (2005)의 논문을 참조하기 바란다.

3. 일반화가법모형에 축소방법 적용

3.1. 모형 제안

Group Lasso와 Elastic net을 일반화가법모형에 적용하기 위해 모형 (1.1)과 (1.2)를 가정하자. 독립변수가 Group 자료라면 본 연구에서 제시한 모형을 Group 형태로 확장하면 된다. 따라서 본 연구에서는 독립변수를 가장 기본이 되는 연속형자료로 가정하였다. 본 연구에서는 독립변수를 기저(basis)로 변환하여 나타낸 일반화가법모형에 Group Lasso와 Elastic net을 적용하여 기저로 구성된 독립변수의 계수 값이 0인 구성요소와 0이 아닌 구성요소를 찾아내는 방법을 제시하고자 한다. 즉, 일반화가법모형에 Group Lasso 모형과 Elastic net 모형을 적용할 수 있는 방법과 이들 방법이 적용되는 새로운 모형을 제시하는 것이다. 따라서 본 연구의 목적에 부합된 모형을 제시하기 위해서는 일반화가법모형의 함수로 된 독립변수를 기저로 전환한 선형 모형식으로 변환하여야 한다.

함수로 된 독립변수를 선형식으로 전환하기 위해 cubic spline basis를 이용한다. 일반화가법모형의 $\sum_{k=1}^{p} f_k(x_{ik})$에서 미지의 함수 부분을 다음의 식 (3.1)과 같이 나타낼 수 있다.

$$f_k(x_{ik}) = \sum_{r=1}^{q} b_r(x_{ik})\beta_{kr}, \tag{3.1}$$

여기서 q는 독립변수 함수를 쪼개어 설명하기 위한 매듭(knot)의 수이고, $b_r(x)$는 독립변수 함수 $f_k(x)$를 선형으로 변환하는 r번째 기저함수이다. β_{kr}은 선형으로 변환된 기저의 계수에 해당된다.

이 식 (3.1)을 식 (1.2)의 일반화가법모형에 대입하면, 독립변수가 선형식으로 변환된 다음 식 (3.2)와 같은 기저를 포함한 식으로 변환할 수 있다.

$$G(\mu_i) = \sum_{k=1}^{p}\sum_{r=1}^{q} b_r(x_{ik})\beta_{kr} \tag{3.2}$$

표기를 위하여 $\boldsymbol{b}_{ik} = (b_1(x_{ik}), b_2(x_{ik}), \ldots, b_q(x_{ik}))^t$, $\boldsymbol{\beta}_k = (\beta_{k1}, \beta_{k2}, \ldots, \beta_{kq})^t$, $i = 1, \ldots, n$, $k = 1, \ldots, p$라 하자. 그리고, $i = 1, \ldots, n$에 대하여 $\boldsymbol{G}^{-1}(t_i)$를 $(G^{-1}(t_1), G^{-1}(t_2), \ldots, G^{-1}(t_n))^t$라 나타내기로 하면 기저를 포함한 일반화가법모형의 변환식에 Group Lasso를 적용한 모형식은 다음 식 (3.3)과 같이 된다.

$$\hat{\boldsymbol{\beta}}^{GAM}_{GLasso}(\lambda) = \text{argmin}\left(\frac{1}{2}\left\|\mathbf{y} - \boldsymbol{G}^{-1}\left(\sum_{k=1}^{p}\boldsymbol{b}_{ik}^t\boldsymbol{\beta}_k\right)\right\|^2 + \lambda\sum_{k=1}^{p}\|\boldsymbol{\beta}_k\|_{K_k}\right). \tag{3.3}$$

마찬가지 방법으로 일반화가법모형 변환식에 Elastic net을 적용한 모형식은 다음 식 (3.4)와 같다.

$$\hat{\boldsymbol{\beta}}^{GAM}_{Enet}(\lambda_1, \lambda_2) = \text{argmin}\left(\left\|\mathbf{y} - \boldsymbol{G}^{-1}\left(\sum_{k=1}^{p}\boldsymbol{b}_{ik}^t\boldsymbol{\beta}_k\right)\right\|^2 + \lambda_2\sum_{k=1}^{p}\|\boldsymbol{\beta}_k\|^2 + \lambda_1\sum_{k=1}^{p}\|\boldsymbol{\beta}_k\|_1\right). \tag{3.4}$$

3.2. 알고리즘

Group Lasso와 Elastic net을 적용한 일반화가법모형의 해를 구하는 방법도 기존의 일반화가법모형의 해를 구하는 방법과 유사한 순차적인 방법(iterative method)을 사용한다. 이 최적해들은 벌점항의 비중에 해당하는 λ값의 변화에 따라 달라진다. λ가 매우 크면 해가 Group Lasso 모형식에서 벌점의 비중이 커져서 거의 모든 변수의 계수값이 0이 되고, λ값이 매우 작으면 거의 모든 변수에 계수값이 계산된다. 따라서 본 연구에서는 λ값의 변화 정도에 따라 계산된 계수값으로 추정한 종속변수 값과 모형식 적합에 사용된 종속변수 관측값의 차이가 최소화 되도록 하는 순차적 방법으로 해를 구하였다. 그 절차는 다음과 같다.

Group Lasso GAM의 해를 구하는 절차

단계 1: 기저로 변환된 일반화가법모형에 Group Lasso에서 해를 구하는 방법을 적용하여 독립변수의 1차 계수값을 추정한다.

단계 2: 자료에 Group Lasso를 적용하기 위해 필요한 λ의 범위 $(\lambda_{\min}, \lambda_{\max})$를 구한다.

단계 3: λ값을 변화시키면서 해 $\hat{\boldsymbol{\beta}}$을 구한다.

단계 4: 단계 3에서 계산된 값을 적용하여 $\mathbf{y}$의 추정값 $\hat{\mathbf{y}}$을 산출하고 차이 $\hat{\mathbf{y}} - \mathbf{y}$를 구한다.

단계 5: 단계 4에서 계산된 차이를 비교하여 최소가 되는 모수 추정값 $\hat{\boldsymbol{\beta}}$를 선택한다.

Elastic net GAM의 해를 구하는 절차

단계 1: 기저로 변환된 일반화가법모형에 Elastic net의 해를 구하는 순차적 방법을 적용하여 독립변수의 1차 계수값을 추정한다.

단계 2: 단계 1의 방법을 λ_1과 λ_2에 따른 벌점의 변화에 따라 독립변수의 계수를 산출한다. Zho와 Hastie (2005)의 Lemma 1을 이용할 수 있다.

단계 3: 단계 2에서 구한 해를 적용하여 $\mathbf{y}$의 추정값 $\hat{\mathbf{y}}$을 구하고 $\hat{\mathbf{y}} - \mathbf{y}$를 구한다.

단계 4: 단계 3에서 구한 차이 값이 최소인 독립변수의 계수 추정값 $\hat{\boldsymbol{\beta}}$을 선택한다.

4. 자료분석

4.1. 모의실험

3장에서 제안된 변수 축소방법의 효능을 파악하기 위해 모의실험을 수행하였다. 본 연구에서 가상자료를 활용한 시뮬레이션은 두가지 목표하에 실시되었다. 첫째는 제안된 Group Lasso GAM, Elastic net GAM, 기존의 GAM의 모형 적합도를 확인해보는 것이다. 이를 위해서 가상자료는 훈련자료(Training Data)와 검증자료(Test Data)의 두 가지로 생성하였다. 생성된 훈련자료에 Group Lasso GAM, Elastic net GAM, 기존 GAM을 적합시켜서 각 변수의 계수를 산출하였다. 산출된 계수값으로 검증자료(Test Data)에 적합시킨 이후, 적합된 추정값과 실제값을 비교하여 Group Lasso GAM, Elastic net GAM, 기존 GAM의 적합률을 비교하여 보았다.

둘째로는 Group Lasso GAM, Elastic net GAM, 기존의 GAM 에서 간결률(parsimonious rate)측면에서 어떤 모형이 나은지 확인해보았다. 여기서, 간결률은 유의한 독립변수의 개수에 대한 측도로 간결률이 높을수록 많은 독립변수의 계수가 0으로 되어 간단한 모형이 된다는 의미이다. 즉, 비교되는 모형들이 동일한 적합률이거나 적합률이 다소 떨어지더라도 간결률이 높은 모형이 더 간단한 모형이라고 간

주할 수 있다. 따라서 Group Lasso GAM, Elastic net GAM, 기존 GAM이 이러한 측면에서 어떤 모형이 더 장점을 가지는지 가상자료를 사용하여 확인해보았다. 즉, 3가지 모형을 검증자료들(test data, 약 100개의 자료)에 적합시킨 후, 각 모형별 간결률을 검증자료별로 산출하였다. 모형 비교는 검증자료들의 모형별 간결률의 평균값으로 모형별 차이를 비교하는 방법을 사용하였다. 간결률은 모형이 적용된 독립변수 계수값 중에서 값이 0인 계수값의 비율로 계산하였다.

이러한 모의실험의 목적에 맞도록 자료를 만들기 위해, 독립변수간 다양한 상관관계를 가정하고, 모든 독립변수의 계수들이 0이 아닌 경우와 계수들 중 일부가 0인 경우를 가정하였다. 우선, 자료생성을 위하여 다음과 같은 로짓모형을 고려하였다.

$$\text{logit}(\mathbf{p}_X) = \mathbf{X}\boldsymbol{\beta} + \sigma\boldsymbol{\varepsilon}. \tag{4.1}$$

0과 1로 구성된 이진자료(binary data) 종속변수 $\mathbf{y}$의 생성은 우선 식 (4.1)의 우변 값을 생성하고 이에 따라 계산되는 개별 확률 p가 0.5보다 작으면 0, 크면 1로 하였다. 본 연구의 목적 중 하나는 일반화가법모형의 개선에 있으며, 일반화가법모형의 종속변수를 설명할 수 있는 지수족 분포들 중에서 가장 기본이 되는 것이 이항분포이기 때문에 이를 분석대상으로 하여 분석과정 및 결과를 제시하면 향후 다른 분포로 확장이 충분히 쉽게 이루어질 것으로 판단된다. 이러한 이유로 본 연구의 분석은 종속변수에 대한 분포가정을 이항분포인 경우로 한정하여 실시하였다. $\mathbf{X}$는 변수의 수가 8개 또는 15개($p = 8, 15$)로 구성되어 있는 디자인행렬(design matrix)이다. 오차항 $\boldsymbol{\varepsilon}$은 $N(0, \sigma^2 I)$를 따르고 $\sigma = 2$를 이용하였다.

독립변수간 상관관계는 높은 수준에서 낮은 수준까지 4가지로 설정하였다. 즉, 독립변수간 상관계수(ρ)가 0.5, 0.3, 0.1 및 $0.5^{|i-j|}$인 경우 4가지를 사용하였다. 매듭(knot)의 수는 11개로 하고 위치는 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0 으로 설정하였다. 계수 β에 대한 가정으로는 다음의 세 가지 경우를 고려하였다.

- 모형 (I) : $p = 8$, $\boldsymbol{\beta} = (3.0, 1.5, 1.2, 0.4, 0.2, 2.0, 1.3, 3.2)$.
- 모형 (II) : $p = 8$, $\boldsymbol{\beta} = (3.0, 0.0, 1.2, 0.4, 0.0, 2.0, 1.3, 0.0)$.
- 모형 (III) : $p = 15$, $\boldsymbol{\beta} = (3.0, 1.5, 1.2, 0.4, 0.2, 2.0, 1.3, 3.2, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0)$.

훈련자료로는 앞서 제시된 조건별로 표본크기 100인 1개의 자료집합을 생성하였고, 검증자료는 자료의 적합률을 확인할 필요가 있으므로, 제시된 조건별로 표본크기 100인 자료집합을 총 100회 생성하였다. 본 연구에서 제시한 모형들을 비교 분석하기 위하여 통계언어 R을 이용하여 프로그램화 하고 분석하였다.

모형 (I)에 대해 본 연구에서 제시한 2가지 모형과 기존 GAM의 적합률의 평균을 계산한 결과를 표 4.1에 나타내었다. 여기서 적합률은 종속변수인 이진자료 값을 제대로 분류한 비율에 해당된다. 기존 GAM 모형의 적합률은 다른 모형들 보다 대체적으로 독립변수들간의 상관관계가 낮을수록 높은 것으로 나타났다. 그런데 상관관계가 높은 경우에는 기존 GAM과 Elastic net GAM, Group Lasso GAM의 적합률에 차이가 없는 것으로 분석되었다. 독립변수간 상관관계가 더 커지면 본 연구에서 제안한 Elastic net GAM, Group Lasso GAM 모델의 적합률이 더 높아질 것으로 예상된다. 상관관계가 낮은 경우에 기존 GAM의 적합률이 높은 것은 기존 GAM의 경우 모든 변수를 버리지 않고 포함하고 있고, 다른 두 가지 방법은 상관관계가 높지 않은 일부 변수를 제외함으로써 정보의 손실이 있기 때문이라 생각된다. 그리고 상관관계가 높을수록 Elastic net GAM과 Group Lasso GAM의 적합률이 기존 GAM보다 높은 것은 이들 두 방법이 기존 GAM보다 상관관계가 높은 변수들을 효과적으로 제어할 수 있는 모형이기 때문으로 판단된다.

표 **4.1.** 모형 (I)의 경우 적합률: 괄호 안은 중앙값, *는 5%, **는 1% 유의수준에서 기존 GAM과 유의한 차이 표시

구분	$\rho = 0.5$	$\rho = 0.3$	$\rho = 0.1$	$\rho = 0.5^{\|i-j\|}$
Group Lasso	0.8289	0.7896**	0.7298**	0.7468**
GAM	(0.8300)	(0.7900)	(0.7400)	(0.7500)
기존	0.8335	0.8132	0.7798	0.7955
GAM	(0.8300)	(0.8100)	(0.7800)	(0.7900)
Elastic net	0.8385	0.7960**	0.7247**	0.7503**
GAM	(0.8400)	(0.8000)	(0.7300)	(0.7500)

표 **4.2.** 모형 (II)의 경우 적합률: 괄호 안은 중앙값, *는 5%, **는 1% 유의수준에서 기존 GAM과 유의한 차이 표시

구분	$\rho = 0.5$	$\rho = 0.3$	$\rho = 0.1$	$\rho = 0.5^{\|i-j\|}$
Group Lasso	0.7782	0.7532	0.6869**	0.7103**
GAM	(0.7800)	(0.7500)	(0.6900)	(0.7200)
기존	0.7760	0.7607	0.7215	0.7407
GAM	(0.7800)	(0.7600)	(0.7300)	(0.7400)
Elastic net	0.7920**	0.7553	0.6796**	0.7067**
GAM	(0.7900)	(0.7600)	(0.6800)	(0.7200)

표 **4.3.** 모형 (III)의 경우 적합률: 괄호 안은 중앙값, *는 5%, **는 1% 유의수준에서 기존 GAM과 유의한 차이 표시

구분	$\rho = 0.5$	$\rho = 0.3$	$\rho = 0.1$	$\rho = 0.5^{\|i-j\|}$
Group Lasso	0.7857**	0.7483*	0.6876	0.7004*
GAM	(0.8000)	(0.7500)	(0.6900)	(0.7100)
기존	0.7576	0.7308	0.6944	0.7149
GAM	(0.7600)	(0.7300)	(0.6900)	(0.7200)
Elastic net	0.7861**	0.7404	0.6804**	0.6921**
GAM	(0.8000)	(0.7400)	(0.6850)	(0.6900)

모형 (II)의 적합률 분석결과는 표 4.2와 같다. 변수간 상관관계가 높은 (상관관계 0.5)경우, 모든 변수의 계수가 존재하는 앞의 분석결과와 유사한 것으로 분석되었다. 상관관계가 낮은 경우에는 기존 GAM이 Group Lasso GAM이나 Elastic net GAM보다 적합률이 더 큰 것으로 분석되었다.

모형 (III)의 경우 적합률 분석결과는 표 4.3과 같다. 이 경우는 $p = 15$로 변수의 수가 많고, 변수중 많은 β 값이 0인 경우로 분석결과는 상관관계가 아주 낮은 경우(0.1 또는 $0.5^{|i-j|}$)를 제외하고는 Elastic net GAM과 Group Lasso GAM이 기존 GAM보다 더 적합률이 높은 것으로 분석되었다. 이러한 분석결과는 통계분석에서 Elastic net과 Group Lasso 등 Lasso 계열의 축소방법이 변수가 많은 경우에 더 장점을 가진 모형이라는 점이 일반화가법모형에도 나타난 결과로 판단된다.

다음으로 모형의 간결(parsimonious) 정도를 측정하기 위해서, 본 연구에서는 간결률(parsimonious rate)이라는 계산기준을 사용하였다. 간결률은 모든 독립변수를 사용하여 모형에 적합시킨 결과 추정된 독립변수들 중에서 계수값이 0인 개수를 전체 독립변수의 개수로 나눈 비율을 의미한다. 예를 들어, 8개의 독립변수를 사용하였고 이 중에서 추정된 계수값이 0인 경우가 4개라면 간결률은 50%(= 4/8)가 된다. 이것은 간결률이 클수록 모형이 더 단순해진다는 것을 의미한다.

각 모형에 대해 간결률을 구한 결과를 표 4.4에 나타내었다. 기존 GAM은 변수 축소과정이 없기 때문에 간결률이 모든 모형에서 0이고 따라서, 기존 GAM은 독립변수 중에서 의미있는 변수를 찾아내는 측면에서 약점이 있는 모형인 것으로 판단된다. 독립변수를 많이 사용할수록 적합률이 높아지는 것이 통계모형의 특성이므로, 앞에서 살펴본 바와 같이 기존 GAM이 Elastic net GAM이나 Group Lasso

표 **4.4.** 각 모형에서 계산된 간결률

모형	구분	$\rho = 0.5$	$\rho = 0.3$	$\rho = 0.1$	$\rho = 0.5^{\lvert i-j \rvert}$
(I)	Group Lasso GAM	0.2263	0.1788	0.1488	0.1700
	Elastic net GAM	0.1400	0.0863	0.1113	0.0938
(II)	Group Lasso GAM	0.3475	0.2960	0.2488	0.2400
	Elastic net GAM	0.1838	0.1538	0.1888	0.1550
(III)	Group Lasso GAM	0.4980	0.3740	0.2887	0.3013
	Elastic net GAM	0.1393	0.1140	0.1853	0.1367

GAM에 비하여 적합률이 더 높았던 이유이기도 하다. 따라서 모든 독립변수를 포함시켜 자료의 적합률을 높이는 방법인 기존 GAM모형은 모형의 예측력(적합률) 분석 이외의 변수 축소와 같은 통계분석에서는 통계모형으로서 활용성에 다소 제약이 있을 수 있고, 이는 기존 GAM모형에서 유의한 독립변수를 찾아내는 변수 축소방법의 연구가 의미가 크다는 것을 뜻한다.

또한 기존 GAM 모형의 경우 간결률이 0임에도 불구하고, 즉 정보 손실이 전혀 없음에도 불구하고 독립변수간 상관관계가 높은 경우에는 본 연구에서 소개한 Elastic net GAM이나 Group Lasso GAM보다 적합률이 떨어진다는 점은 의미있는 발견이라고 판단된다. 실제 자료에서는 일반적으로 독립변수간 상관관계가 존재하므로, 변수 축소 과정이 생략된 기존 GAM보다는 변수 축소 과정이 포함된 Elastic net GAM이나 Group Lasso GAM이 더 유용한 분석도구가 될 수 있다는 것을 의미한다. 즉, 실제자료 분석에서는 적합률 측면과 간결률 측면을 모두 고려하여야 하므로 기존 GAM보다는 변수 축소 과정이 포함된 Elastic net GAM이나 Group Lasso GAM이 더 나은 모형이 될 것으로 판단된다.

표 4.4에서 보면 간결률은 Group Lasso GAM, Elastic net GAM, 기존 GAM의 순서인 것으로 파악된다. 모형 (I)은 실제 모든 변수에 계수값이 존재하는 경우이고, 모형 (II)는 자료를 생성할때 독립변수의 계수가 0인 비율이 약 38% (= 3/8)인 점을 감안하면, Group Lasso GAM과 Elastic net GAM의 경우 조금 부족한 점은 있으나 변수 축소에 효과적인 것으로 보인다. 특히 변수간 상관관계가 0.5인 경우의 Group Lasso GAM의 간결률은 0.3475로 모의실험 가정에서 0의 비율과 유사한 것으로 나타났다. 모형 (III)의 경우에도 모형 (II)와 유사하게 독립변수간 상관관계가 클수록 간결률이 크고, 독립변수간 상관관계가 작을수록 간결률이 작은 것으로 분석되었다. 모형 (III)은 전체 독립변수 중에서 약 47%의 계수가 0인 것으로 가정하여 생성된 것이다. 이러한 자료생성가정에 가장 부합된 간결률을 보이고 있는 것은 3가지 모형 중에서 Group Lasso GAM인 것으로 나타났다. 특히 독립변수간 상관관계가 0.5경우의 Group Lasso GAM의 간결률은 0.4980으로 자료생성 가정과 매우 유사하다. 반면에 Elastic net GAM은 Group Lasso GAM에 비하여 간결률이 자료생성 가정과 차이가 많이 나는 것으로 나타났다. 이러한 결과는 Elastic net GAM이 Group Lasso GAM에 비하여 자료 손실이 적다는 것을 의미하고, 앞의 적합률 분석에서 독립변수간 상관관계가 높은 경우에 Elastic net GAM의 적합률이 Group Lasso GAM보다 높은 결과와 밀접하게 관계가 있는 것으로 판단된다.

4.2. 자동차보험 통계활용

본 소절에서는 본 연구에서 제시한 변수 축소방법을 실제 자료에 활용하기 위해 회계년도 2005년(FY2005)의 우리나라 자동차보험 통계자료에 적용하여 보았다. 본 통계자료는 업계 전체자료 중 일부를 표본 추출한 것이다. 즉, 보험개발원에 집적된 FY2005의 자동차보험산업 전체 자료 중에서 임의표본추출하여, 분석에 사용될 약 50만개 자료집합을 만들었다. 자료는 사고년도 기준(Accident Year Basis), 역년도 기준(Calendar Year Basis), 증권년도 기준(Policy Year Basis)의 3가지 통계추출 기준

표 **4.5.** 자동차보험 통계 분석자료의 변수 특성

구분	세부분류
연령	기명피보험자 연령, 18세부터 98세까지 1세 단위
가입경력	운전경력(자동차보험 가입경력)에 따라 1년 단위
할인할증	40% 부터 200%의 범위내에서 10% 단위
자동차연식	1975년식부터 2005년식까지 1년 단위
사고유무(종속변수)	사고가 있으면 '1', 없으면 '0'

표 **4.6.** 자동차보험 통계의 독립변수간 상관관계

	나이	연식	할인할증	가입경력
나이	1.000	−0.072	−0.289	0.344
연식		1.000	0.047	−0.065
할인할증			1.000	−0.573
가입경력				1.000

표 **4.7.** 자동차보험 자료의 분석결과 : **는 1% 유의수준에서 기존 GAM과 유의한 차이 표시

구분	평균 적합률	평균 간결률
Group Lasso GAM	0.8700**	0.5225
기존 GAM	0.8526	0.0000
Elastic net GAM	0.8682**	0.9275

중에서 계약자료와 사고자료가 가장 일치하는 증권년도기준으로 하였다. 분석도구인 R 프로그램의 처리 용량 관계로 약 50만건 자료에서 임의추출 방법으로 100건씩 추출한 자료를 분석에 사용하였다. 이러한 실제자료의 분석을 보다 정밀하게 하기 위하여 매듭의 개수를 15개로 확장하여 분석을 실시하였다.

통계추출 대상 보험계약은 개인용 및 플러스개인용으로 하였다. 개인용 및 플러스 개인용이 전체 자동차보험에서 차지하는 비율이 평균유효대수 기준으로 70.4%로 매우 높고, 보험회사들이 가장 관심을 가지고 있는 보험종목이기 때문이다. 그리고 현재 자동차보험에서 운영되고 있는 모든 담보를 분석대상으로 하였다. 개인용 자동차보험에서 운영되고 있는 담보는 대인배상Ⅰ, 대인배상Ⅱ, 대물배상, 자기신체사고, 자기차량손해, 무보험차상해가 있다. 종속변수는 자동차보험 사고가 증권별로 있으면 '1', 없으면 '0'인 이진자료이다. 앞의 기준으로 추출된 자료 중 논리적 모순이 있는 자료를 제거하는 방법으로 자료를 정리하였다. 예를 들면, '보험료가 없는 자료', '대인배상Ⅰ의 보험금이 1억원을 초과하는 자료' 등을 분석 자료에서 배제하거나 논리에 맞도록 수정하였다. 이상의 분석에 사용된 자동차보험 실제자료의 변수 특성을 요약하면 표 4.5와 같다.

본 분석에 사용된 자동차보험 통계의 상관관계를 분석하여 보면 표 4.6과 같다. 상관계수가 가입경력과 할인할증의 경우는 매우 높은 −0.578이고, 가입경력과 연령은 0.344, 나이와 할인할증은 −0.289인 것으로 나타났다. 자동차연식의 경우에는 다른 독립변수와 상관관계가 낮은 것으로 나타났다. 본 분석의 상관관계 수준은 모의실험의 상관관계별 분석에서, 상관관계 0.5와 $0.5^{|i-j|}$의 사이 정도로 판단된다.

4.1절의 모의실험 분석결과를 보면, 독립변수간 상관관계가 0.5와 같이 높은 경우에 Elastic net GAM과 Group Lasso GAM의 적합률이 높은 것으로 분석되었다. 실제 자동차보험 자료에 Elastic net GAM과 Group Lasso GAM, 그리고 기존 GAM을 적용해본 결과, 표 4.7에 의하면 분석결과가 모의실험의 결과와 일치하는 것으로 나타났다. 즉, 본 연구에서 제시한 Group Lasso GAM과 Elastic net GAM이 기존 GAM보다 적합률이 높은 것으로 분석되었다. 간결률 측면을 보면, Elastic net GAM,

Group Lasso GAM, 기존 GAM의 순서로 큰 것으로 분석되었다. 실제 자료의 경우는 독립변수간 상관관계가 높은 경우가 흔히 있다. 이러한 특성을 볼 때, 변수 축소방법인 Elastic net과 Group Lasso를 적용한 Elastic net GAM 및 Group Lasso GAM이 유용한 분석 모델이 될 것으로 판단된다.

5. 결론 및 토론

본 연구에서는 변수 축소방법들 중에서 유용한 방법인 것으로 알려진 Group Lasso와 Elastic net을 일반화가법 모형에 적용하여 보았다. 이들 방법을 적용한 모형을 Elastic net GAM, Group Lasso GAM이라고 이름 붙이고, 이들 모형이 기존 GAM모형과 비교하여 어떠한 특징과 장점이 있는지 분석하여 보았다. 그리고 Elastic net GAM 및 Group Lasso GAM의 해를 구하는 절차도 제시하였다. 제시된 모형과 해를 구하는 절차에 따라 최종적으로 실제 자료에 적용하여 보았다. 자료로 제시된 통계모형을 가상자료와 실제자료로 분석하고 비교하였다. 분석결과 독립변수간 상관관계가 높은 경우에는 본 연구에서 제시한 Elastic net GAM과 Group Lasso GAM이 기존 GAM보다 적합률이 더 큰 것으로 나타났다. 반면에 독립변수간 상관간계가 낮은 경우에는 Elastic net GAM과 Group Lasso GAM이 기존 GAM보다 적합률이 더 낮은 것으로 분석되었다. 실제 자료인 자동차보험자료에 적용하여 본 결과도 두 가지 제안된 모형의 적합률이 기존 GAM 보다 높은 것으로 분석되었다. 이는 실제자료의 독립변수간 상관관계가 높기 때문에 발생한 것으로 보인다.

적합률과 간결률을 동시해 고려해 보면, 가장 좋은 모델이 적합률이 높고 간결률도 높은 경우일 것이다. 그런데 적합률과 간결률은 상호 교환(trade off) 관계에 있다고 볼 수 있으므로, 이 두 가지 면에서 모두 좋을 수는 없고 적절한 기준으로 정해야 한다. 이러한 모형선택 기준에 따라 기존 GAM, Elastic net GAM, Group Lasso GAM을 비교하여 보면, 본 연구에서 제시한 모형들이 기존 GAM모형보다 우수한 모형인 것으로 판단된다. 본 연구에서 제시한 Elastic net GAM과 Group Lasso GAM 중에서는 Group Lasso GAM이 Elastic net GAM보다 조금 더 우수한 것으로 판단된다. 이러한 결과는 본 연구의 모의실험 가정하에서 파악된 것이며, 향후 다양한 모의실험 또는 모형비교를 통해 모형간 차이 비교가 더 진행될 필요가 있다고 생각된다. 아울러, 제안된 방법의 이론적인 성질도 규명할 필요가 있다고 본다.

참고문헌

Bakin, S. (1999). Adaptive regression and model selection in data mining problems, *Ph.D. Dissertation*, The Australian National University, Canberra.

Fu, W. (1998). Penalized regressions: The Bridge versus the Lasso, *Journal of Computational and Graphical Statistics*, **7**, 397–416.

Genkin, A., Lewis, D. D. and Madigan, D. (2007). Large-scale bayesian logistic regression for text categorization, *Technometrics*, **49**, 291–304.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models (with discussion), *Statistical Science*, **1**, 297–318.

Kim, Y., Kim, J. and Kim, Y. (2006). Blockwise sparse regression, *Statistica Sinica*, **16**, 375–390.

Krishnapuram, B., Carin, L., Figueiredo, M. A. and Hartemink, A. J. (2005). Sparse multinomial logistic regression: Fast algorithms and generalization bounds, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**, 957–968.

Lokhorst, J. (1999). The Lasso and generalized linear models, *Honors Project*, University of Adelaide, Adelaide.

Meier, L., van de Geer, S. and Buhlmann, P. (2008). The Group Lasso for logistic regression, *Journal of the Royal Statistical Society*, **70**, 53–71.

Roth, V. (2004). The generalized Lasso, *IEEE Transactions on Neural Networks*, **15**, 16–28.

Shevade, S. and Keerthi, S. (2003). A simple and efficient algorithm for gene selection using sparse logistic regression, *Bioinformatics*, **19**, 2246–2253.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society*, **68**, 49–67.

Zhao, P., Rocha, G. and Yu, B. (2006). Grouped and hierarchical model selection through composite absolute penalties, *Technical Report*, University of California at Berkeley, Department of Statistics.

Zhou, N. and Zhu, J. (2007). Group variable selection via hierarchical Lasso and its oracle property, manuscript.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the Elastic net, *Journal of the Royal Statistical Society*, **67**, 301–320.

A Study on Applying Shrinkage Method in Generalized Additive Model

Seungdo Ki[1] · Kee-Hoon Kang[2]

[1]Department of Statistics, Hankuk University of Foreign Studies, Korea Insurance Research Institute
[2]Department of Statistics, Hankuk University of Foreign Studies

(Received January 2010; accepted February 2010)

Abstract

Generalized additive model(GAM) is the statistical model that resolves most of the problems existing in the traditional linear regression model. However, overfitting phenomenon can be aroused without applying any method to reduce the number of independent variables. Therefore, variable selection methods in generalized additive model are needed. Recently, Lasso related methods are popular for variable selection in regression analysis. In this research, we consider Group Lasso and Elastic net models for variable selection in GAM and propose an algorithm for finding solutions. We compare the proposed methods via Monte Carlo simulation and applying auto insurance data in the fiscal year 2005. It is shown that the proposed methods result in the better performance.

Keywords: Additive model, Lasso, Group Lasso, Elastic net.

This research was supported by the research fund of Hankuk University of Foreign Studies, 2009.
[2]Corresponding author: Associate Professor, Department of Statistics, Hankuk University of Foreign Studies, Mohyun, Cheoin-Koo, Yongin 449-791, Korea. E-mail: khkang@hufs.ac.kr