The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.6.1103
28(6), 1103–1111

Relative Error Prediction via Penalized Regression

Seok-Oh Jeong[a] · Seo-Eun Lee[a] · Key-Il Shin[a,1]

[a]Department of Statistics, Hankuk University of Foreign Studies

(Received August 12, 2015; Revised September 25, 2015; Accepted October 5, 2015)

Abstract

This paper presents a new prediction method based on relative error incorporated with a penalized regression. The proposed method consists of fully data-driven procedures that is fast, simple, and easy to implement. An example of real data analysis and some simulation results were given to prove that the proposed approach works in practice.

Keywords: prediction, relative error, penalized regression, LASSO

1. 서론

설명변수가 제공하는 보조정보를 이용해 반응변수의 값을 예측하는 데에는 일반적으로 평균제곱오차(mean squared error)를 최소화시키는 것을 목표로 하는 예측방법을 사용한다. 또한 최근에 각광을 받고 있는 벌점회귀(penalized regression) 역시 평균제곱오차에 대한 최적화 문제에 모형의 복잡성에 대한 벌점을 추가로 고려한 것이다. 그러나 자료에 이상점이 포함되어 있거나 오차항의 분포가 심각하게 비대칭인 경우 평균제곱오차에 기반한 예측은 문제의 소지가 있다. 이러한 문제점을 보완하기 위해 평균제곱오차 대신 평균제곱상대오차(mean squared relative error)에 기반한 예측 기법이 연구되어 왔다. Park과 Stefanski (1998)에서 처음으로 상대오차에 기반한 예측방법에 대한 체계적인 연구 결과가 제시되었으며, Park과 Shin (2006)에 의해 상대오차에 기반한 시계열자료 분석 기법이 연구된 바 있다. Jones 등 (2008)은 평균제곱상대오차와 비모수적 회귀함수 추정기법을 결합한 예측 방법을 연구하였고, Jeong과 Shin (2008)은 이 논문의 방법론이 가진 문제점을 보완한 새로운 비모수적 예측법을 제안하였다. 이후 Hwang과 Shin (2008)은 이 방법론을 소지역 추정(small area estimation)에까지 확장하였다.

본 논문에서는 평균제곱상대오차를 벌점회귀에 적용한 새로운 방법을 제안하고, 모의실험 및 실제 자료 분석 결과를 통해 제안된 방법론의 유효성을 실증하였다. 본 논문의 구성은 다음과 같다. 2절에서는 논문에 사용된 상대오차의 개념과 벌점회귀를 결합한 방법론을 소개하였다. 3절에서 제안된 방법론의 유효성을 모의실험을 통해 살펴보았고, 4절에서는 한국교통연구원의 일일교통량 자료에 제안된 방법론을 적용한 결과를 소개하였다. 끝으로 5절에서 결론 및 전망을 제시하였다.

Seok-Oh Jeong's work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012R1A1 A2006112). Key-Il Shin's work was supported by Hankuk University of Foreign Studies Research Fund of 2015.

[1]Corresponding author: Department of Statistics, Hankuk University of Foreign Studies, 81 Oedae-ro, Mohyeon-myeon, Cheoin-gu, Yongin-si, Gyeonggi-do 17035, Korea. E-mail: keyshin@hufs.ac.kr

2. 상대오차 벌점회귀 예측법

예측 대상 변수를 Y, 예측량(predictor)을 $\tilde{Y}$, 예측에 사용된 보조정보를 포함한 공변량을 $\mathbf{x} = (x_1, x_2, \ldots, x_p)$로 나타내도록 하자. 보통, 공변량 $\mathbf{x}$의 값이 주어졌을 때 Y값에 대한 예측의 정확도를 나타내는 기준으로 예측 오차 $Y - \tilde{Y}$의 크기인 MSE(mean squared error)

$$\mathrm{MSE}\left(Y, \tilde{Y}|\mathbf{x}\right) := E\left[\left(Y - \tilde{Y}\right)^2 \Big| \mathbf{x}\right]$$

를 이용하는 것이 일반적이며, 이 기준 하에 최적의 예측 방법은

$$\tilde{Y}^*_{\mathrm{MSE}} := \underset{\tilde{Y}}{\operatorname{argmin}} \mathrm{MSE}\left(Y, \tilde{Y}|\mathbf{x}\right) = E(Y|\mathbf{x})$$

를 사용하는 것임이 잘 알려져 있다. 즉, 조건부 기대값 $E(Y|\mathbf{x})$에 대해 선형회귀모형

$$E(Y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{2.1}$$

을 가정하고 계수벡터 $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$를 데이터를 이용해 추정하면 $\mathbf{x}$의 값이 주어졌을 때의 $\tilde{Y}^*_{\mathrm{MSE}}$ 값을 얻을 수 있게 된다. 그러나 공변량 $\mathbf{x}$의 차원이 높은 경우 위 모형이 복잡해지면서 오히려 예측력이 떨어지는 문제가 발생한다. 이러한 문제점을 해결하기 위한 방안으로 공변량 차원 증가에 따라 증가하는 모형의 복잡성에 대해 벌점을 부여하는 방법이 널리 사용된다. 벌점 부여 방식에 따라 다양한 종류의 방안을 고려할 수 있는데 본 논문에서는 LASSO(least absolute shrinkage and selection operator)를 고려한다. 즉, 관측 자료 $\{(\mathbf{x}_i, Y_i), i = 1, 2, \ldots, n\}$에 대해

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_p x_{pi})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2.2}$$

를 최소로 하는 계수 β_j, $j = 0, 1, 2, \ldots, p$를 찾아 예측에 이용하는 방법이다. 벌점모수 $\lambda > 0$는 보통 교차확인법(cross-validation) 등을 이용해 예측 성능을 극대화화는 것을 목표로 미리 선택된다. 즉, 주어진 λ에 대해 식 (2.2)를 최소로 하는 β_j, $j = 0, 1, 2, \ldots, p$를 구하고, 이를 식 (2.1)의 $E(Y|\mathbf{x})$에 대입해 LASSO 방법에 의한 예측량을 얻게 된다. 식 (2.2)의 두 번째 항의 벌점 부분을 ℓ_1-타입이 아닌 다른 형태의 벌점으로 교체하면 다양한 형태의 벌점회귀 방법이 정의되며 방법마다 나름의 장단점이 있음이 알려져 있다. 이상의 내용에 대한 체계적인 설명은 Tibshirani (1996), Buelmann과 van de Geer (2011) 등을 참조하기 바란다.

그러나 자료에 포함된 Y의 관측값에 이상점이 포함되어 있거나 오차의 의미를 규모에 따라 상대적으로 계산할 필요가 있을 때와 같이 상대오차(relative error) $(Y - \tilde{Y})/Y$에 기반한 추론이 현실적으로 더 타당한 경우가 있다. 이러한 경우 평균제곱상대오차(mean squared relative error; MSRE)를 이용하면

$$\tilde{Y}^*_{\mathrm{MSRE}} := \underset{\tilde{Y}}{\operatorname{argmin}} E\left[\left(\frac{Y - \tilde{Y}}{Y}\right)^2 \Big| \mathbf{x}\right] = \frac{E(Y^{-1}|\mathbf{x})}{E(Y^{-2}|\mathbf{x})}$$

와 같은 예측량을 얻게 되는데, 예측대상 변수 Y가 양(positive)인 경우 부등식 $\tilde{Y}^*_{\mathrm{MSRE}} \le \tilde{Y}^*_{\mathrm{MSE}}$가 성립함을 쉽게 증명할 수 있다. 즉 상대오차에 기반한 예측은 소위 축소(shrinkage) 성질을 갖게 된다는 것이다. 이와 관련한 자세한 논의는 Park과 Stefanski (1998)와 Jeong과 Shin (2008)을 참조하기 바란다.

본 논문에서는 식 (2.2)의 첫 번째 항의 제곱오차를 상대제곱오차로 대체한

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_p x_{pi}}{Y_i}\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2.3}$$

를 최소로 하는 계수 β_j, $j = 0, 1, 2, \ldots, p$를 찾아 예측량 구축에 이용하는 상대오차 벌점회귀 방법을 제안한다. 이 방법은 상대오차 기반 예측 방법이 가진 장점과 LASSO-타입 벌점회귀 방법이 가진 안정적인 예측 성능을 동시에 갖춘 방법으로서, 부가적으로 변수 선택(variable selection)의 결과까지 얻을 수 있게 해 준다. 또한 기존의 LASSO 알고리즘을 활용해 바로 구현할 수 있다는 점도 중요한 장점이다. 식 (2.3)에서 적합도 부분인 첫 번째 항을 다시 정리하면

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_p x_{pi}}{Y_i}\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(1 - \beta_0 \cdot \frac{1}{Y_i} - \beta_1 \cdot \frac{x_{1i}}{Y_i} - \beta_2 \cdot \frac{x_{2i}}{Y_i} - \cdots - \beta_p \cdot \frac{x_{pi}}{Y_i}\right)^2 \end{aligned} \tag{2.4}$$

가 된다. 따라서 β_j, $j = 0, 1, 2, \ldots, p$에 대한 최적화 문제의 관점에서 보면 식 (2.3)의 최적화 문제는 식 (2.2)의 것과 동일한 종류의 최적화 문제임을 알 수 있다. 다만, 식 (2.3)의 벌점모수 λ의 선택을 위해 교차확인법 등을 적용할 때, 일반적 LASSO 방식은 평균제곱오차를 이용해 예측력을 평가하지만 본 논문에서 제안하는 상대오차 벌점회귀의 경우에는 반드시 상대오차를 이용해야 전반적인 추정 절차 및 최적화의 흐름이 부합하는 벌점모수가 선택될 수 있음에 유의해야 할 것이다. 또한 벌점 부분을 다른 종류의 벌점으로 대체하기만 하면 다른 종류의 벌점방법을 이용한 상대오차 벌점회귀 방법으로 쉽게 확장할 수 있으나 이는 향후 연구과제로 남겨둔다.

3. 모의실험

본 논문에서 제안한 상대오차 벌점회귀 방법의 성능을 살펴보기 위해 아래의 식 (3.1)과 같은 모형 하에서 모의실험을 수행했다.

$$\begin{aligned} y &= \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{20} x_{20} + \varepsilon, \\ \varepsilon &= Z - 24, \quad Z \sim \text{Gamma}(3, 8), \\ x_i &\sim \text{Gamma}(2, 20), \quad i = 1, 4, 5, \ldots, 20, \\ x_2 &= x_1 \cdot V, \quad x_3 = x_2 \cdot V, \quad V \sim \text{Exponential}(1), \\ \beta_j &= \begin{cases} j, & j = 1, 2, 3, 4, 5, \\ 0, & j = 6, 7, \ldots, 20. \end{cases} \end{aligned} \tag{3.1}$$

회귀계수의 상당 부분을 0으로 설정한 것은 변수 선택의 기능을 확인하기 위한 것이고, x_1을 이용해 x_2와 x_3가 생성되도록 한 것은 일정 수준의 다중공선성이 있는 자료에 대해 제안한 방법이 잘 작동함을 보여주기 위한 설정이다. 본 모의실험에서 사용된 표본의 크기는 $n = 100$, 반복 수는 $M = 1{,}000$이다.

Figure 3.1은 모의실험에서 상대오차 벌점회귀 방법(RE-LASSO)과 일반적인 벌점회귀 방법(LASSO)을 적용해 추정된 회귀계수 값의 분포를 각각 상자그림(boxplot)으로 정리한 것이다. 이 그림에서 보듯 LASSO 방법과 RE-LASSO 방법 모두 모의실험 설계에서 설정한 0이 아닌 회귀계수 β_j, $j = 1, 2, 3, 4, 5$의 값을 잘 추정하고 있으며, 변수 선택 측면에서 볼 때 두 방법 모두 β_j, $j \geq 6$를 0으로 축소추정하는 변수 선택의 성능도 우수함을 알 수 있다. 즉 본 논문에서 제안한 상대오차 벌점회귀 방법의 회귀계수 추정 및 변수선택 성능이 기존의 LASSO 방법에 비해 결코 떨어지지 않음을 확인할 수 있다.

본 논문의 강조점인 예측력을 비교하기 위해 비교통계량으로 아래와 같이 정의된 RMSE(root mean squared error)와 RMSRE(root mean squared relative error)를 계산해 LASSO 방법과 본 논문에서

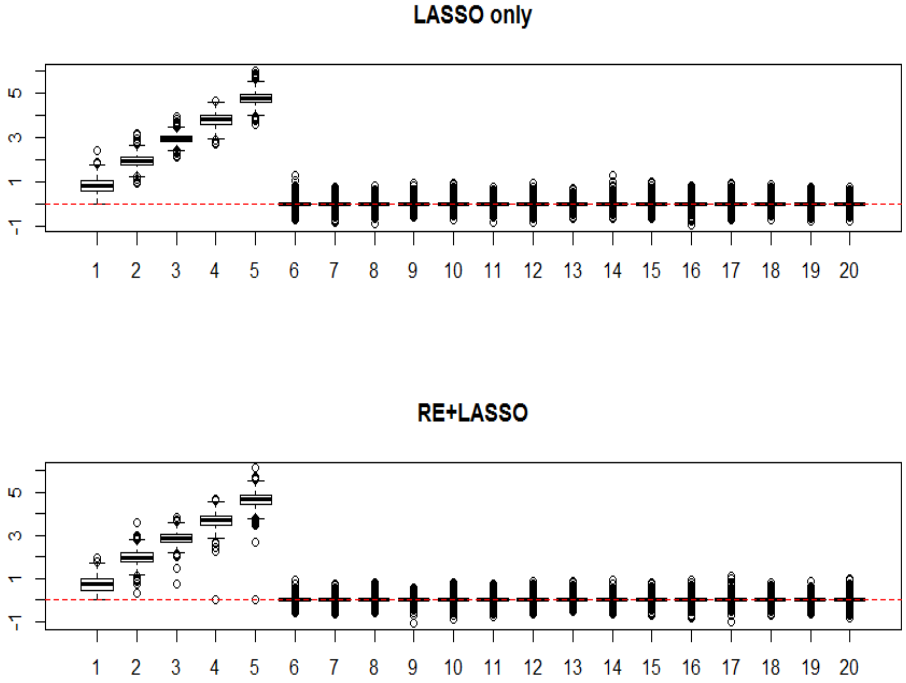

Figure 3.1. The simulation results of the regression coefficient estimates. The numbers labeled in the horizontal line are the index for the regression coefficients.

Table 3.1. Comparison of the prediction errors: The averages (and the s.e. in the parentheses) of logarithms of RMSE and RMSRE from 1,000 Monte Carlo simulation

	log(RMSE)	log(RMSRE)
LASSO	4.310(0.531)	−1.502(0.002)
RE-LASSO	4.417(0.623)	−1.856(0.001)

RMSE = root mean squared error; RMSRE = root mean squared relative error
LASSO = least absolute shrinkage and selection operator; RE-LASSO = relative error LASSO

제안한 상대오차-LASSO(RE-LASSO) 방법의 예측 성능을 비교하였다.

$$\text{RMSE} = \left\{ \frac{1}{n} \sum_{i=1}^{n} (\tilde{y}_i - y_i)^2 \right\}^{\frac{1}{2}}, \qquad \text{RMSRE} = \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\tilde{y}_i - y_i}{y_i} \right)^2 \right\}^{\frac{1}{2}}$$

방법론의 특성 상 RMSE 측면에서는 LASSO 방법이 RE-LASSO 방법보다 우수하고 RMSRE 측면에서는 반대일 것을 예상할 수 있는데, Table 3.1의 결과를 살펴보면 역시 예상대로임을 확인할 수 있다. 다만 RMSE 측면에서 RE-LASSO 방법의 예측성능이 LASSO 방법에 비해 다소 떨어지지만 표준오차를 고려할 때 통계적으로 유의한 차이는 아닌 것으로 판단된다. 반면 RMSRE 측면에서는 RE-LASSO의 예측성능이 LASSO 방법에 비해 탁월하게 우수함을 확인할 수 있다. 이는 모의실험의 자료생성 과정에서 오차항을 감마분포를 이용해 생성했기 때문에 다수의 Y값이 이상치인 점과 오차항 분포의 비대칭성이 반영된 것 때문으로 보인다. 이러한 분포적 특성은 실제 자료에서 흔히 나타나는 현상이

Table 4.1. Comparison of the variable selection results for the whole data

LASSO		RE-LASSO	
건물의 총 면적	0.0002	셔틀버스 운행횟수	0.8580
건물의 주차용량	1.8155	건물의 주차용량	0.1251
주차요금 여부	170.4077	총 고용인 수	0.1378
총 고용인 수	0.2524		

LASSO = least absolute shrinkage and selection operator; RE-LASSO = relative error LASSO

기 때문에 RE-LASSO 방법이 현장에서 활용도가 매우 높을 것으로 기대하게 하는 결과이다. 물론 본 모의실험의 설정과 달리 오차항의 분포가 대칭성을 보이거나 이상점이 발생할 가능성이 없는 경우에는 일반적인 LASSO 방법에 의한 예측력이 더 좋을 수 있음에 유의할 필요가 있다. 이상의 내용은 다음 절의 실제 자료 분석에서 추가로 논의하였다.

4. 실제 자료 분석 - 교통연구원 자료

이 절에서는 실제 자료 분석을 통해 상대오차 벌점회귀 방법의 예측 성능을 살펴보았다. 사용된 자료는 한국교통연구원의 차량통행량 자료이다. 지역마다 주요 건물의 일평균 차량통행량(Y)을 예측함에 있어 상대오차 벌점회귀 방법을 적용해 예측에 유의미한 변수가 어떤 것이 있는 지, 또한 LASSO 방법과 예측 분석 결과가 어떤 차이를 보이는지 살펴보았다. 우선 전체 자료에 대한 분석을 시행하였고, 도시 규모별로 차량통행량이 다른 특성을 가질 것으로 예상되므로 인구 100만을 기준으로 한 규모별 분석을 실시하였다.

4.1. 전체 자료 분석

자료에 LASSO 회귀 방법과 상대오차 벌점회귀 방법을 각각 적용하여 얻은 회귀계수 추정 결과를 Table 4.1에 제시하였다. 벌점회귀에 의해 0으로 추정된 회귀계수는 표에 제시하지 않았다. 두 방법 모두 교통량 예측에 유의한 변수로 '건물의 주차용량'과 '총 고용인 수'를 선택한 것을 확인할 수 있다. '총 고용인 수'가 일평균 차량통행량에 결정적 영향을 미치는 건물을 이용하는 (유동)인구 규모를 대표한다는 점에서 납득이 되는 결과이다. '건물의 주차용량'이 선택된 것은 건물의 주차 편의성이 좋을 수록 차량통행량이 많아진다는 것으로 분석할 수 있다. 다만 LASSO의 경우 '주차요금 여부'의 계수 추정치가 양의 값인 170.4077으로, 주차요금을 받는 건물이 그렇지 않는 건물보다 평균적으로 170대 가량 차량통행량이 많은 것으로 예측하게 됨을 의미한다. 이는 통행량이 많은 건물의 경우 주차요금을 받게 되는 경향성이 강한 때문으로 보인다. 그러나 통상적으로 주차요금을 받으면 차량통행을 억제하는 효과가 있을 것으로 기대할 수도 있는 바 분석 결과를 활용해야 하는 입장에서는 다소 혼란스러운 결과일 수 있다. 반면에 RE-LASSO의 경우 '셔틀버스 운행횟수'가 차량통행량 예측에 유의한 변수로 선택되었는데 이것은 LASSO가 선택했던 '건물의 총 면적'에 해당하는 건물 규모에 관련한 변수로 해석된다.

다음으로 두 방법의 예측력 비교를 위해 전체 자료를 훈련자료(training set) 75%, 시험자료(test set) 25%로 임의 추출을 통해 나눈 후, 각 방법을 적용해 RMSE와 RMSRE를 계산하는 작업을 1,000회 반복하였다. Figure 4.1의 왼쪽 그래프는 각 방법의 RMSE 값의 분포를, 오른쪽 그래프는 RMSRE 값을 비교한 것이다. 가로축은 LASSO 방법에 의한 오차값을 나타내고, 세로축은 상대오차 벌점회귀 방법의 오차값을 나타낸 것으로 대각선보다 아래에 있는 점은 RE-LASSO 방법의 예측오차가 더 큰 경우를 의미한다. 오차값들을 시각적으로 보다 잘 비교할 수 있도록 모든 값에 로그를 취하였다. 예상대로 RMSE를 비교한 왼쪽 그래프에서는 LASSO 방법을 이용하여 계산된 RMSE 값이 상대오차 벌점회귀

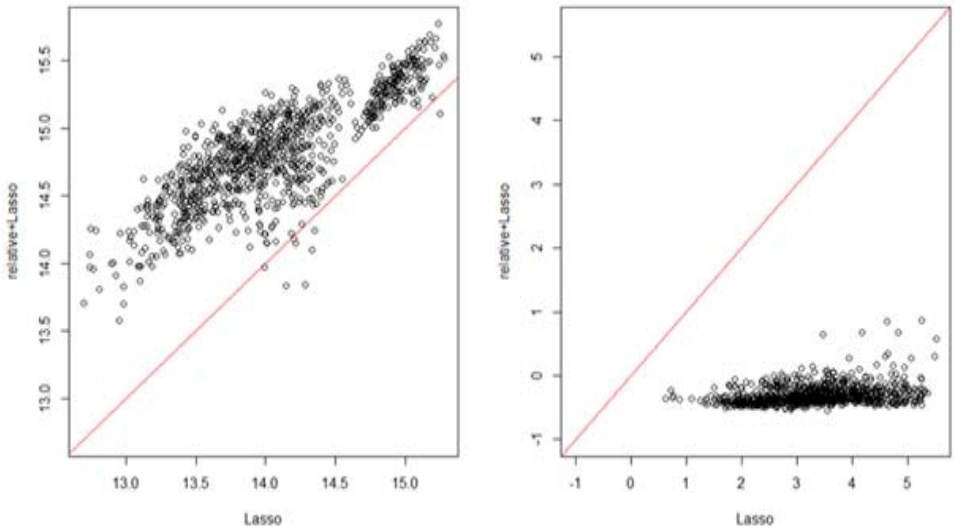

Figure 4.1. Comparison of the prediction errors for the traffic data - the whole data.

Table 4.2. Comparison of the prediction errors: the averages (and the s.e. in the parentheses) of logarithms of RMSE and RMSRE for the traffic data - the whole data

	log(RMSE)	log(RMSRE)
LASSO	14.05(0.018)	3.3521(0.030)
RE-LASSO	14.82(0.012)	−0.3056(0.005)

RMSE = root mean squared error; RMSRE = root mean squared relative error
LASSO = least absolute shrinkage and selection operator; RE-LASSO = relative error LASSO

Table 4.3. Comparison of the variable selection results for the traffic data - Population ≥ 1,000,000

LASSO 회귀계수		RE-LASSO 회귀계수	
건물의 총 면적	0.0037	건물의 총 면적	0.0003
건물의 주차용량	1.8728	건물의 주차용량	0.0839
승용차 요일제	1.2738	총 고용인 수	0.0593
총 고용인 수	2.1020		

LASSO = least absolute shrinkage and selection operator; RE-LASSO = relative error LASSO

방법을 이용하여 얻은 RMSE보다 작은 경우가 1000번 중 988번이었다. 하지만 오른쪽의 MSRE를 비교한 그래프의 경우, 모든 점들이 대각선 아래에 놓여있음을 알 수 있다. 또한 Table 4.2는 위의 결과를 수치로 정리한 결과로 이상의 모든 분석 결과가 3절의 모의실험 결과와 동일한 양상을 보임을 확인할 수 있다.

4.2. 도시 규모별 자료 분석

주어진 자료를 인구가 100만 이상인 도시와 그 외의 도시로 나누어 4.1절과 같은 방법으로 분석을 시행하였다. 다만 변수선택 결과에만 의미있는 논의 사항이 도출되었고 예측력 비교 결과는 앞 절의 전체 자료를 이용한 결과와 거의 같은 양상을 보였기 때문에 변수선택에 대한 논의만 논문의 본문에 포함하였

Table 4.4. Comparison of the variable selection results for the traffic data - Population < 1,000,000

LASSO 회귀계수		RE-LASSO 회귀계수	
근방 지하철 노선 여부	351.1009	건물의 총 면적	0.0305
건물의 총 면적	1.1847	총 고용인 수	0.1813
총 고용인 수	0.8415		

LASSO = least absolute shrinkage and selection operator; RE-LASSO = relative error LASSO

다. 다음 Table 4.3은 인구가 100만 이상인 도시의 자료를 따로 분석한 후 얻은 변수 선택 결과이다. 변수 선택 결과를 보면 두 방법 모두 통행량에 유의한 영향을 미칠 것으로 예상되는 '건물의 총 면적', '건물의 주차용량' 그리고 '총 고용인 수'를 선택하고 있어 납득할 만하다. 다만 LASSO의 경우 '승용차 요일제'의 시행 여부가 유의한 양의 계수를 갖는 것으로 나타났는데 이는 승용차 요일제 채택을 통해 통행량을 억제하고자 하는 기대와는 반대되는 결과로서, 앞서 전체 자료 분석의 결과에서 '주차요금 여부'와 같은 맥락에서 이해할 수 있다. 반면에 상대오차 벌점회귀 결과를 보면 통행량에 유의할 것으로 예상되는 변수만 선택되었다는 점이 주목할 만하다. 전체 자료 분석 시 선택되었던 '셔틀버스 운행횟수'와 같이 건물의 규모나 번화가 여부를 간접적으로 나타내는 지표 대신 '건물의 총 면적'이 선택되어 분석 결과의 해석이 용이하다.

다음으로 Table 4.4는 인구 100만 미만의 도시에 대한 변수 선택 결과이다. 100만 이상의 도시 자료 분석 결과와 비교하면 LASSO의 경우 '건물의 총 면적' 대신 '근방 지하철 노선 여부'가 선택되었다. 근방에 지하철 노선 수가 많을 수록 건물이 위치한 지역이 번화하기 때문으로 해석할 수도 있으나, 인구 100만 미만의 도시에서 주변 대중교통 편의성이 건물 차량 통행량을 증가시킨다는 결과로서 역시 해석상 혼란이 있는 것은 부인할 수 없다. 그에 반해 RE-LASSO는 '건물의 총 면적'과 '총 고용인 수'를 일관되게 선택하면서 인구 100만 이상인 도시와 달리 '건물의 주차용량'은 선택하지 않았는데, 100만 미만인 도시의 경우 통행량 자체가 100만 이상인 도시에 위치한 건물에 비해 상대적으로 작기 때문에 주차 편의성이 크게 문제되지 않기 때문인 것으로 풀이된다.

5. 결론

본 논문에서는 LASSO 회귀 방법의 적합도 부분을 상대오차로 설정하여 회귀계수를 추정하는 방법을 제안하였다. 이상치가 포함되어 있거나 오차항의 분포가 비대칭인 경우, 예측력을 높임에 있어 제안된 방법이 유의함을 보이기 위해 시행한 모의실험 및 실제 자료 적용 결과를 요약하면 다음과 같다.

- 일반 LASSO 방법을 통해 얻어진 변수 선택 결과와 비교하였을 때, 제안된 방법을 통해 유사한 양상의 변수선택 결과를 얻을 수 있었다.
- 일반 LASSO 방법과 제안된 방법의 예측력을 비교한 결과, 제곱오차의 경우 두 방법의 차이가 크지 않으나 상대오차의 경우 제안된 방법이 월등히 우수하다.
- 구현이 쉽고 계산이 빠르다.

본 논문에서의 결과를 바탕으로 LASSO 외에 adaptive-LASSO 혹은 다른 종류의 벌점방법을 이용한 연구를 향후 연구과제로 수행해 볼만 하다. 또한 고차원 회귀 문제에서 피할 수 없는 다중공선성(multicollinearity) 문제를 상대오차를 사용하면 다소 완화할 수 있을 것으로 예상된다. 위 식 (2.4)의 상대오차에 의한 적합도 부분의 회귀계수에 대한 이차식 표현을 살펴보면 설사 몇몇 x변수들이 강한 상관관계를 갖더라도 Y_i값으로 나누는 과정에서 상관성이 희석되는 효과가 발생하게 됨을 예상할

수 있다. x 변수들 간에 정확히 $x_j = ax_{j'}$ (a는 상수, $j \neq j'$)의 관계가 성립하지 않는 한 상관성이 상당 수준 희석되기 때문인데, 이러한 현상에 대해 심도있게 연구할 필요가 있다.

References

Buelmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data - Methods, Theory and Applications*, Springer.

Hwang, H.-J. and Shin, K.-I. (2008). Shrinkage prediction for small area estimation, *The Korean Journal of Applied Statistics*, **21**, 109–123.

Jeong, S.-O. and Shin, K.-I. (2008). A new nonparametric method for prediction based on mean squared relative errors, *Korean Communications in Statistics*, **15**, 255–264.

Jones, M. C., Park, H., Shin, K.-I., Vines, S. K. and Jeong, S.-O. (2008). Relative error prediction via kernel regression smoothers, *Journal of Statistical Planning and Inference*, **138**, 2887–2898.

Park, H. and Shin, K.-I. (2006). A shrinked forecast in stationary process favoring percentage error, *Journal of Time Series*, **27**, 129–139.

Park, H. and Stefanski, L. A. (1998). Relative-error prediction, *Statistics and Probability Letters*, **40**, 227–236.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society B*, **21**, 279–289.

벌점회귀를 통한 상대오차 예측방법

정석오[a] · 이서은[a] · 신기일[a,1]

[a]한국외국어대학교 통계학과

(2015년 8월 12일 접수, 2015년 9월 25일 수정, 2015년 10월 5일 채택)

요 약

본 논문에서는 상대오차의 개념과 벌점회귀를 결합한 새로운 예측방법을 제시하였다. 제안된 방법은 오차항의 분포가 정규성을 크게 벗어나 있어 이상점을 포함하거나 오차항의 분포가 심각하게 비대칭인 경우에도 안정적으로 예측력이 유지할 뿐 아니라 벌점회귀를 통한 변수선택의 성능도 우수하다. 또한 개념적으로 쉽고, 계산 속도가 빠르며, 기존의 알고리즘을 활용해 구현하는 것이 매우 쉽다. 한국교통연구원의 일일 차량통행량 자료 실제 분석 및 모의실험을 통해 제안된 방법의 우수한 성질을 확인하였다.

주요용어: 예측, 상대오차, 벌점회귀, LASSO

[1]교신저자: (17035) 경기도 용인시 처인구 모현면 외대로81, 한국외국어대학교 통계학과.
E-mail: keyshin@hufs.ac.kr