The Korean Journal of Applied Statistics
2021, Vol. 34, No. 6, 937–944

DOI: https://doi.org/10.5351/KJAS.2021.34.6.937
Print ISSN1225-066X / Online ISSN 2383-5818

Variable selection in partial linear regression using the least angle regression

Han Son Seo[a], Min Yoon[1,b], Hakbae Lee[c]

[a]Department of Applied Statistics, Konkuk University;
[b]Department of Applied Mathematics, Pukyong National University;
[c]Department of Applied Statistics, Yonsei University

Abstract

The problem of selecting variables is addressed in partial linear regression. Model selection for partial linear models is not easy since it involves nonparametric estimation such as smoothing parameter selection and estimation for linear explanatory variables. In this work, several approaches for variable selection are proposed using a fast forward selection algorithm, least angle regression (LARS). The proposed procedures use t-test, all possible regressions comparisons or stepwise selection process with variables selected by LARS. An example based on real data and a simulation study on the performance of the suggested procedures are presented.

Keywords: least angle regression, partial linear models, sequential selection, variable selection

1. 서론

모형추정에서 변수선택은 과대적합이나 과소적합을 방지하기 위한 중요한 절차 중 하나이다. 선형회귀모형의 경우 모형에 필요한 최소한의 설명변수는 결정계수, 수정결정계수, C_p, AIC (Akaike, 1973, 1974), BIC (Schwarz, 1978)등 모형적합성 척도에 의한 모형비교를 통해 선택되거나 전진선택법, 후진제거법, 가능한 모든회귀법, 단계별 회귀법등 검정을 통해 선택된다. 최근에는 변수의 수가 많은 경우를 대비하여 다양한 벌점 회귀에 의한 방법들이 제안되고 있다. 본 연구는 부분선형모형에서 모형의 선형성에 관련된 변수선택 문제를 다룬다. 부분선형모형은 선형모형의 모형유연성이 충분하지 않을 경우 다차원의 가법모형과 함께 선형모형의 대안으로 사용된다. 부분선형모형에서 변수선택방법은 모형의 함수부분을 추정 한 후 AIC나 교차타당성(cross-validation)등의 기준을 적용하는 방법 (Chen과 Chen, 1991; Härdle 등, 2000)과 다양한 벌점회귀(penalized regressions)를 사용하는 방법 (Fan과 Li, 2004; Bunea, 2004; Bunea와 Wegkamp, 2004; Fan과 Peng 2004; Xie와 Huang, 2009) 등이 있으며 최근에는 범함수 공변량(functional covariates)이 있는 부분선형모형에서 변수선택을 위하여 벌점최소제곱추정량을 사용하는 방법이 제시 되었다 (Aneiros 등, 2015). 순차적 방법이나 가능한 모든 회귀법(all possible subsets regression)은 변수의 개수가 많은 경우 과도한 계산량을 초래하기 때문에 일반적으로 사용되지 않는다. 본 연구에서는 벌점 추정량을 사용하는 대신 주요변수나 주요모형을 우선 선별한 후 이들에 대해 순차적인 변수 선택법을 적용하거나 가능한 모형 전부가 고려되는 방법을 제안한다. 이 과정에서 다수의 모형을 고려할 때 발생할 수 있는 계산량의 부담을 줄이기 위하여 변수들이 least angle regression (LARS)에 의해 선택되는 순서에 따라 주요변수 또는 모형을 선별하기로 한다.

This paper was written as part of Konkuk University's research support program for its faculty on sabbatical leave in 2021.

[1] Corresponding author: Department of Applied Mathematics, Pukyong National University, 45, Yongso-ro, Nam-Gu, Busan 48513, Korea. E-mail: myoon@pknu.ac.kr

Published 31 December 2021 / journal homepage: http://kjas.or.kr

2장에서는 기존의 t-검정에 의한 변수 선택법을 설명하고 부분선형모형에서 LARS를 이용한 변수선택 방법을 제안한다. 3장에서는 예제와 모의실험을 통하여 제안된 방법들의 효율성을 비교하며 4장에서는 연구 결과를 요약한다.

2. 변수선택 방법

반응변수와 설명변수간 관계에서 선형적 성분과 비모수적 성분을 함께 고려한 부분선형모형은 다음과 같이 정의된다.

$$Y_i = X_i^T\boldsymbol{\beta} + m(Z_i) + \varepsilon_i, \quad i = 1, \ldots, n. \tag{2.1}$$

여기서 $X_i = (x_{i1}, \ldots, x_{ip})^T$ 와 $Z_i = (z_{i1}, \ldots, z_{id})^T$ 는 설명변수 벡터, $\boldsymbol{\beta} = (\beta_1, \ldots \beta_p)^T$ 는 알려지지 않은 모수 벡터이다. m은 알려지지 않은 비모수함수이며 ε_i는 임의의 오차항을 나타내고 $E(\varepsilon_i) = 0$, $E(\varepsilon_i^2) = \sigma^2$, 그리고 $E(\varepsilon_i|X_i, Z_i) = 0$이다. 부분선형모형에서 비모수함수 $m(Z_i)$를 $m(Z_i) = E(Y_i|Z_i) - E(X_i^T\boldsymbol{\beta}|Z_i)$으로 대체하면 모형 (2.1)은 다음과 같은 선형모형으로 표현될 수 있다.

$$Y_i - E(Y_i|Z_i) = (X_i - E(X_i|Z_i))^T\boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n. \tag{2.2}$$

따라서 $\widetilde{Y}_i$와 $\widetilde{X}_i$를 Z_i에 의하여 조정된 부분잔차(partial residual)라고 할 때 $\widetilde{Y}_i$와 $\widetilde{X}_i$간의 근사 선형모형을 고려할 수 있다.

$$\widetilde{Y}_i = \widetilde{X}_i^T\boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n. \tag{2.3}$$

모형 (2.3)에서 최소제곱 추정량은 $\hat{\boldsymbol{\beta}}_{LS} = (\widetilde{X}^T\widetilde{X})^{-1}\widetilde{X}^T\widetilde{Y}$이 되고 여기서 $\widetilde{X}_j = \sum_{i=1}^n \omega_i(T_j)X_i$, $\widetilde{Y}_j = \sum_{i=1}^n Y_j - \omega_i(T_j)Y_i$이며 $\omega_i(z)$는 양의 가중함수이다. 최소제곱추정량 $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$에 의해 모수가 추정되었을 때 $m(z)$의 비모수 추정통계량은 $\hat{m}(t) = \sum_{i=1}^n \omega_i(t)(Y_i - X_i^T\hat{\boldsymbol{\beta}}_{LS})$이 된다.

본 연구에서는 최소제곱추정량을 이용한 변수선택 방법들을 제시한다. 최소제곱추정량에 의해 부분선형모형을 추정할 때 부분잔차나 비모수 함수등은 Epanechnikov 커널에 의한 커널추정법에 의해 계산되며 대역폭(bandwidth)은 교차타당성에 의해 추정된 최적값을 적용한다. 변수선택 문제에서 유의변수를 제대로 선택하기 위해서는 적절한 선택기준과 과정이 중요하다. 지금까지 제시된 변수선택기준들은 제 각각 장단점을 갖고 있어서 최적의 기준을 찾은 것이 쉽지 않다. 본 연구에서는 모형설명력 척도로 수정-R^2을 기본적으로 사용한다. 최적모형을 찾기 위한 과정으로 모형들을 비교 할 때, 각 모형의 수정-R^2 값을 비교할 수도 있고 수정-R^2가 이미 모형에 참여한 변수의 크기가 반영된 척도이지만 변수의 크기에 따른 모형의 개선정도를 고려하여 변수 크기가 한 개 증가할 때 수정-R^2값이 일정값 이상 향상 되었을 때 추가 변수 모형이 유의한 것으로 판정할 수도 있다.

선형모형에서 변수선택의 가장 기본적인 방법은 Bonferonni 부등식을 이용한 t-검정이다. 식 (2.3)에서 계산되는 최소제곱추정량은 근사적으로 정규분포를 따르게 되어 (Härdle 등, 2000) 귀무가설 $\mathrm{H}_0 : \beta_i = 0$에 대하여 최소제곱추정량 $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$이 적용되는 일반적인 t–통계량에 의해 t-검정을 수행할 수 있다. Bonferonni 부등식을 이용한 t-검정 절차를 "TO-절차" 라고 표기하기로 하고 그 과정을 요약하면 다음과 같다.

(1) t-검정(TO-절차)

- $E(Y_i|Z_i)$, $E(X_{ki}|Z_i)$을 커널추정법으로 추정한 후 $\widetilde{Y}_i$, $\widetilde{X}_i$을 계산한다.
- $\widetilde{Y}_i$, $\widetilde{X}_i$에 대한 선형모형을 설정하고 최소제곱 추정법으로 $\hat{\boldsymbol{\beta}}_{LS}$을 추정 한 후 Bonferroni 한계에 의한 t-검정을 수행하여 유의적인 변수를 선택한다.

t-검정을 이용한 "TO-절차"는 다수의 연구결과에서 알려졌듯이 위음성(false negative)등이 발생할 경향이 있어서 이에 대한 대안으로 개별적인 모형을 고려하는 변수선택 과정이 제안된다. 본 연구에서는 개별적인 모형을 고려하는 변수선택 방법 대신 일부 모형만 선별하여 비교하는 절차를 제안한다. 제안되는 방법들은 설명변수들이 LARS에 의해 선택되는 순서를 기준으로 일부 모형을 선별한다. 주요 변수 선택에 사용되는 LARS 알고리즘은 다음과 같은 과정으로 수행된다.

LARS에서는 반응변수와 상관관계가 가장 큰 설명변수를 첫 번째 변수로 선택한다. 현재 단계에서 LARS 추정치에 관여된 설명변수의 집합을 A라고 하면 A는 현 단계모형의 잔차와 상관관계가 큰 변수들로 구성된다. 변수집합 A에 상응하는 현재 단계 LARS 추정치를 $\hat{\mu}_{\mathrm{A}}$라고 하면, 다음 단계의 LARS 추정치는 $\mu(\gamma) = \hat{\mu}_{\mathrm{A}} + \gamma u_{\mathrm{A}}$, $\gamma > 0$가 되며 여기서 u_A는 A에 있는 각 변수들 벡터와 동일한 각을 유지하는 벡터이다. 즉 다음 단계의 LARS 추정치는 기존의 추정치에서 동일 각 벡터의 방향으로 $\hat{\gamma}$만큼 이동한 값이 된다. 이때 $\hat{\gamma}$는 A에 포함되지 않은 어떤 변수와 새로운 추정치 $\mu(\gamma)$의 잔차 간 상관관계 정도와 A에 포함된 변수들과 $\mu(\gamma)$의 잔차간 상관관계 정도가 동일하도록 결정된다. LARS를 활용한 부분모형 비교방법들 LT-절차, LA-절차, LAT-절차, LS-절차를 다음과 같이 제안한다.

(2) LARS + t-검정과 가능한 모든 회귀모형(LT-절차)

가능한 모든 회귀모형(all possible regression)을 고려한 변수선택은 과도한 계산량이 필요하다. LT-절차는 이중 일부분의 모형만을 고려하며, 고려대상은 LARS와 t-검정을 통하여 선정한다. LT-절차에서 변수를 선택하는 과정은 다음과 같다.

- TO-절차에서와 같이 $\widetilde{Y}_i$, $\widetilde{X}_i$를 계산한다.
- $\widetilde{Y}_i$, $\widetilde{X}_i$에 의하여 LARS 추정을 수행한 후 그 결과에 따라 출현한 순서대로 $\widetilde{X}_i$를 정렬 한다, $\{\widetilde{X}_1, \widetilde{X}_2, \ldots, \widetilde{X}_p\}$.
- TO-절차를 수행한 결과 유의하게 판정된 변수 중에서 $\{\widetilde{X}_1, \widetilde{X}_2, \ldots, \widetilde{X}_p\}$의 순서상 가장 마지막에 해당하는 변수를 $\widetilde{X}_r$이라고 하면, $\{\widetilde{X}_1, \widetilde{X}_2, \ldots, \widetilde{X}_r\}$변수에 대하여 가능한 모든 회귀모형식 비교를 실행하여 최적의 변수를 선택한다.

(2) LARS와 가법모형(LA-절차)

LA-절차는 McCann과 Welsch (2007)가 선형모형에서 이상치가 존재할 때 강건한 모형선택을 위해 제시한 방법과 유사하게 수행된다. LARS 알고리즘에 의하여 변수의 중요도 순서를 정한 후 이에 의한 p개의 모형을 비교하여 최적변수를 선택하는 방법이다.

- LT-절차와 유사하게 $\widetilde{Y}_i$와 $\widetilde{X}_i$, $i = 1, 2, \ldots, p$에 대하여 LARS 알고리즘을 수행하여 변수 출현 순서에 의한 수열 $\{\widetilde{X}_1, \widetilde{X}_2, \ldots, \widetilde{X}_p\}$을 생성한다.
- 이 수열에 기반한 p개의 가법모형을 고려한다. : 모형1 = $\{\widetilde{X}_1\}$, 모형2= $\{\widetilde{X}_1, \widetilde{X}_2\}$, ..., 모형$p$ = $\{\widetilde{X}_1, \widetilde{X}_2, \ldots, \widetilde{X}_p\}$.
- p개의 모형에 대하여 수정-R^2 등 적합성을 나타내는 수치를 계산한 후 최적모형을 선택한다.

(2) LARS + 단계적 선택(LS-절차)

LS-검정은 독립변수를 단계별로 한 개씩 늘려가는 순차적인 방법이다. k개의 변수가 포함된 k 단계 최적 변수군을 M_k라고 하자. $(k+1)$ 단계 최적변수군 M_{k+1}은 M_k에 새로운 변수를 한 개 더 포함시켜 완성한다. M_k에 새롭게 포함될 후보변수들(M_k^c)에 대해 한 개씩 순서대로 포함 여부를 검정한다. k 단계에서 새롭게 포함될 후보변수들의 검정순서를 나타내는 수열을 $S_k = \{\widetilde{X}_1, \widetilde{X}_2, \ldots, \widetilde{X}_{p-k}\}$라고 할 때 k 단계에서 수행되는 검정과정은 다음과 같다.

- 현재 단계의 최적변수군인 M_k에 S_k의 변수 $\widetilde{X}_{(i)}$를 한 개 더 추가한 변수군을 $R_k^i (R_k^i = M_k \bigcup \{\widetilde{X}_{(i)}\})$라고 하자. R_k^1부터 시작하여, 추가한 변수군이 더 유의하다는 결과를 얻을 때 까지 S_k의 순서대로 R_k^i 변수군 모형과 M_k 변수군 모형을 비교하는 적합성 검정을 수행한다.

  (a) 만약 $\widetilde{X}_{(q)}$가 새로운 변수로 판정되면, 즉 R_k^q가 유의적인 변수군 이라고 하면 $k+1$ 단계의 최적변수군 M_{k+1}과 검정 순서 수열 S_{k+1}를 각각 $M_{k+1} = R_k^q = M_k \bigcup \{\widetilde{X}_{(q)}\}$, $S_{k+1} = \{\widetilde{X}_{(1)}, \widetilde{X}_{(2)}, \ldots, \widetilde{X}_{(q-1)}, \widetilde{X}_{(q+1)}, \ldots, \widetilde{X}_{(p-k)}\}$로 지정하고 위 단계를 반복한다.

  (b) S_k에 포함된 모든 변수에 대하여 R_k^i, $i = 1, 2, \ldots, p-k$가 비유의적이라고 하면 M_k가 최종적으로 최적변수군으로 판정된다.

  위 절차에서 1단계 변수군 M_1과 수열 S_1은 $M_0 = \emptyset$과 LARS에 따른 변수순서 수열 $S_0 = \{\widetilde{X}_{(1)}, \widetilde{X}_{(2)}, \ldots, \widetilde{X}_{(p)}\}$을 기초로 t-검정 또는 F-검정에 의해 결정한다.

3. 예제와 모의실험

3.1. 모의실험

변수선택 방법들을 비교하기 위하여 식 (2.1)의 부분선형모형을 기반으로 가상 데이터를 생성한다. 모수 벡터 $\boldsymbol{\beta}$는 첫 번째 두 번째 다섯 번째 요소가 각각 3, 1.5, 2이고 나머지는 0인 $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T = (3, 1.5, 0, 0, 2, 0, \ldots, 0)^T$로 지정되어 모형의 유의변수는 X_1, X_2, X_5가 된다. 설명변수 X는 평균 0, 공분산행렬 $(\rho^{|j-k|})_{jk}$인 정규분포에서 생성되며 오차항 ε은 표준정규분포에서 생성된다. 비모수 함수 $m(z)$는 함수 $m(z) = (z^2 - 1/3) + \sqrt{12}(z - 1/2)$에 따라 생성되며 비모수 함수의 설명변수 Z는 범위 (0, 1)의 균등분포에서 생성된다. 데이터 크기 n과 설명변수의 개수 p는 $(n, p) \in \{(30, 10), (50, 30), (100, 70), (200, 150)\}$와 같은 네 가지 경우가 설정되며 각 경우에 있어서 설명변수 X의 공분산을 조절하는 ρ 값을 0, 0.5, 0.9로 지정한다. 각 경우마다 실험 횟수는 총 500 번이며 동일한 설명변수 값을 반복 사용하지 않는다. LARS는 Efron 등 (2004)에서 제시된 조율모수 수식이 반영된 R 패키지 lars의 함수들에 의해 수행되었다. 각 방법의 검정력은 세 개의 척도 P_1, P_2, P_3에 의해 계산된다. P_1은 세 개의 유의변수를 정확하게 찾은 비율이고, P_2는 적어도 한 개 이상의 유의변수를 찾은 비율이며 $(1 - P_2)$는 일종의 가면현상(masking phenomenon)이 발생한 비율이 된다. P_3는 선택된 설명변수 중에 비유의적 변수가 포함된 비율, 일종의 수렁현상(swamping phenomenon) 이 발생한 비율이다. Table 1은 모의실험을 통하여 변수선택 방법들을 비교한 결과를 보여준다.

대부분의 경우에 있어서 모든 방법들의 P_2값은 1이거나 1에 가깝다. 따라서 모든 방법들이 유의한 변수를 전부 찾지 못하는 경우는 거의 없다고 할 수 있다. 모든 방법들에 있어서, 변수간의 상관정도를 표시하는 ρ 값이 커질수록 P_1과 P_3에 근거한 변수선택의 정확도가 대체로 낮아지고 있으나 상대적으로 TO-절차가 ρ에 대하여 가장 강건한 결과를 보여주고 있다.

설명변수의 크기 p가 커짐에 따라 LA-절차와 LS-절차의 P_1과 P_3 값은 대체로 개선되고 있으며 ρ값이 커질수록 개선의 정도도 크다. 반면에 t-검정에 기반한 TO-절차와 LT-절차의 경우 $p = 10$ 정도로 설명변수의 크기가 작은 경우 잘 작동하지만 설명변수의 크기 p가 커짐에 따라 급격하게 정확성이 낮아지며 $p = 150$에서는 다소 회복된 결과를 보이고 있다. t-검정에 기반한 TO-절차와 LT-절차를 비교하면, $\rho = 0$ 일 때를 제외하고 대부분의 경우 LT-절차가 TO-절차보다 고려하는 모형이 더 많음에도 불구하고 변수 간 공선성이 커지면 정확도가 더 낮아서 LT-절차에서 적용되는 변수선택의 기준값이 제 역할을 하지 못하는 것을 암시하고 있다. 다만 LT-절차의 경우 $P_3 = 0$ 이거나 매우 작아서 변수선택의 수렁화 현상이 발생하는 경우가 거의 없음을 알 수 있다.

본 연구에서 고려한 방법들 중 설명변수의 크기가 작을 때($p = 10$) 에는 t-검정에 기반한 방법들이 효과적이다. 특히 ρ가 작을 때에는 LT-절차가, ρ가 클 때에는 TO-절차가 가장 정확도가 높다. LA-절차와 LS-절차는

Table 1: Summary of simulation results

		$\rho = 0.0$			$\rho = 0.5$			$\rho = 0.8$		
		P_1	P_2	P_3	P_1	P_2	P_3	P_1	P_2	P_3
$n = 30$ $p = 10$	TO	0.896	1.000	0.104	0.900	0.900	0.100	0.864	1.000	0.066
	LT	0.998	1.000	0.000	0.984	0.984	0.000	0.278	1.000	0.002
	LA	0.946	1.000	0.048	0.848	0.848	0.138	0.502	1.000	0.384
	LS	0.962	1.000	0.036	0.866	0.866	0.124	0.540	1.000	0.340
$n = 50$ $p = 30$	TO	0.624	1.000	0.376	0.682	0.682	0.316	0.534	1.000	0.364
	LT	0.628	1.000	0.000	0.274	0.274	0.000	0.000	0.996	0.006
	LA	0.974	1.000	0.026	0.948	0.984	0.046	0.612	1.000	0.332
	LS	0.990	1.000	0.010	0.954	0.954	0.042	0.630	1.000	0.310
$n = 100$ $p = 70$	TO	0.555	1.000	0.445	0.535	0.535	0.465	0.455	1.000	0.515
	LT	0.565	1.000	0.000	0.235	0.235	0.000	0.000	1.000	0.000
	LA	1.000	1.000	0.000	0.990	0.990	0.010	0.720	1.000	0.265
	LS	1.000	1.000	0.000	0.990	0.990	0.010	0.750	1.000	0.230
$n = 200$ $p = 150$	TO	0.705	1.000	0.295	0.680	0.680	0.032	0.675	1.000	0.325
	LT	0.720	1.000	0.000	0.270	0.270	0.000	0.000	1.000	0.000
	LA	1.000	1.000	0.000	1.000	1.000	0.000	0.920	1.000	0.080
	LS	1.000	1.000	0.000	1.000	1.000	0.000	0.930	1.000	0.070

* $P_1 = Pr$ (an exactly correct selection), $P_2 = Pr$ (at least one correct variable is selected), $P_3 = Pr$ (an incorrect variable is selected).

t-검정에 기반한 TO-절차와 LT-절차 보다 대부분의 경우 정확도가 높고 설명변수의 크기에 잘 대응하고 있으며 변수간의 상관관계에 대해서도 $\rho = 0.5$ 정도 까지는 정확성이 잘 유지되고 있다. LA-절차와 LS-절차간에는 큰 차이가 없으나 본 연구에서 제시한 LS-절차가 LA-절차보다 좀 더 정확성이 높은 결과를 보여준다.

3.2. Ragweed pollen level 자료

제안된 변수선택방법을 돼지풀 꽃가루오염도 자료(Ragweed pollen level data)에 적용한다. 이 데이터는 1993년 미시간에서 수집된 87일 간의 돼지풀 꽃가루 오염도와 관련 정보로 구성되어 있다. 반응변수는 꽃가루 오염도 수준이며 설명변수는 심각한 낙수여부(X_1), 온도(X_2), 풍속(X_3)이다. 설명변수는 표준화하여 사용하고 반응변수는 왜도가 심하므로 제곱근 변환 후 사용한다. Ni 등 (2009)에서 언급된 것처럼 돼지풀 시즌이 시작되는 날로 부터 경과일수(Z) 는 꽃가루 오염도 수준과 비선형적 관계를 갖고 있으므로 꽃가루 오염도 수준에 대하여 Z의 비모수적 함수가 포함되는 부분선형모형을 고려할 수 있다.

부분선형모형의 변수선택 문제를 위하여 선형관계의 설명변수는 x_1, x_2, x_3 외에 제곱항 x_2^2, x_3^2 교차항 x_1x_2, x_1x_3, x_2x_3 등 여덟 개의 변수를 고려한다. 여러 변수선택 방법이 적용된 결과는 Table 2 에 요약되어 있다. 모든 방법이 주효과 변수만을 선택하였으며 t-검정에 기반한 TO-절차, LT-절차는 두 개의 변수를, LA-절차와 LS-절차는 세 개의 변수를 선택하였다. Ni 등 (2009)에서도 x_1, x_2, x_3가 선택된 부분선형모델을 최종 모형으로 추정하여 LA-절차와 LS-절차의 결과와 일치하고 있다.

4. 결론

부분선형모형에서 변수선택을 할 때, 기존의 방법들이 부분선형의 함수부분을 포함한 모형 전체를 대상으로 벌점추정량을 사용하거나 선정기준을 적용하는 것에 반해 본 연구에서 제안된 방법은 부분잔차에 의한 선형모형을 기반으로 수행된다. 이러한 방법은 선형부분에 관여된 변수 선택을 먼저 수행한 후 선택된 변수와 함께

Table 2: Selected variables by applying selection methods to ragweed pollen level data

Method	Selected variables'
TO	x_2, x_3
LT	x_1, x_2
LA	x_1, x_2, x_3
LS	x_1, x_2, x_3

함수부분을 추정하여 부분선형모형을 완성하는 방식이다. 이 과정에서 선형변수 선택과 함수추정이 동시에 수행되지 않음으로 해서 모형의 설명력이 정확하지 않을 수 있으나 우선 수행된 변수선택이 일종의 최종적인 변수선택에 앞선 예비적인 과정으로 간주할 수 있으며 실험결과 변수선택 후 함수부분 추정에 의한 최종적인 변수 선택 과정과 이를 생략한 과정은 큰 차이가 없음을 알 수 있다.

본 연구에서는 부분선형모형에서 변수선택 방법으로 기존의 t-검정과 LARS를 이용한 변수선택과정을 제안하였다. 변수선택 과정에서 비교 대상이 될 모형을 압축하기 위하여 LARS의 결과를 활용하며 LARS와 더불어 t-검정 수행(LT-절차), 가법모형 비교(LA-절차), 순차적 검정(LS-절차)등이 각각 제안되었다. 기존의 t-검정에 의한 변수선택방법과 LARS를 활용한 방법들을 비교한 결과, 기존의 t-검정은 변수크기가 작은 경우, 특히 변수간 상관관계가 높을 때 정확성이 높다. LARS를 이용한 방법들 중 LT-절차는 변수크기가 작고 변수간 상관관계가 높지 않을 때 유용하다. 변수크기가 클 경우에는 LA-절차, LS-절차등이 t-검정을 활용한 방법들보다 더 효과적이다.

본 연구에서 제시한 방법들은 부분잔차의 선형모형을 기반으로 모형을 비교하였으나 비선형 함수 추정이 포함된 부분선형모형의 설명력을 기반으로 LA- 절차나 LS-절차를 수행할 수 있다. 실제로 이와 같은 절차를 설계하여 적용한 결과 계산량이 많아지는 단점에 비해 LA-절차와 LS-절차보다 향상된 효율성을 보여주지 못했다. 본문에서 사용한 P_1, P_2, P_3 외에 단일수치를 사용한 변수선택방법의 비교를 위해 각 실험결과의 Mattews correlation coefficient (MCC)에 의한 전체 실험의 MCC 분포를 활용할 수 있다.

References

Akaike H (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd International Symposium on Information Theory*, 267–281, Budapest.

Akaike H (1974). A new look a the statistical model identification, *IEEE Transactions on Automatic Control*, **19**, 716–723.

Aneiros G, Ferraty F, and Vieu P (2015). Variable selection in partial linear regression with functional covariate, *Statistics*, **49**, 1322–1347.

Bunea F (2004). Consistent covariate selection and post model selection inference in semiparametric regression, *Annals of Statistics*, **32**, 898–927.

Bunea F and Wegkamp M (2004). Two-stage model selection procedures in partially linear regression, *The Canadian Journal of Statistics*, **32**, 105–118.

Chen, H. and Chen, K.(1991). Selection of the splined variables and convergence rates in a partial spline model, *The Canadian Journal of Statistics*, **19**, 323–339.

Efron B, Hastie T, Johnstone I, and Tibshirani R (2004). Least angle regression, *The Annals of Statistics*, **32**, 407–499.

Fan J and Li R (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis, *Journal of American Statistical Association*, **99**, 710–723.

Fan J and Peng H (2004). Nonconcave penalized likelihood with a diverging number of parameters, *Annals of*

Statistics, **32**, 928–961.

Härdle W, Liang H, and Gao J (2000). *Partially Linear Models*, Physica-Verlag, Heidelberg.

McCann L and Welsch R (2007). Robust variable selection using least angle regression and elemental set sampling, *Computational Statistics and Data Analysis*, **52**, 249–257.

Ni X, Zhang H, and Zhang D (2009). Automatic model selection for partially linear models, *Journal of Multivariate Analysis*, **100**, 2100–2111.

Schwarz G (1978). Some comments on C_p, *Technometrics*, **15**, 662–676.

Xie H and Huang J (2009). SCAD-penalized regression in high-dimensional partially linear model, *Annals of Statistics*, **37**, 673–696.

Received August 10, 2021; Revised August 16, 2021; Accepted August 16, 2021

부분선형모형에서 LARS를 이용한 변수선택

서한손[a], 윤민[1,b], 이학배[c]

[a]건국대학교 응용통계학과; [b]부경대학교 응용수학과; [c]연세대학교 응용통계학과

요 약

본 연구는 부분선형모형에서 변수선택의 문제를 다룬다. 부분선형모형은 평활화모수 추정과 같은 비모수 추정과 선형설명변수에 대한 추정의 문제를 함께 포함하고 있어 변수선택이 쉽지 않다. 본 연구에서는 빠른 전진선택법인 LARS 를 이용한 변수선택법을 제시한다. 제안된 방법은 LARS에 의하여 선별된 변수들에 대하여 t-검정, 가능한 모든 회귀모형 비교 또는 단계별 선택법을 적용한다. 제안된 방법들의 효율성을 비교하기 위하여 실제데이터에 적용한 예제와 모의실험 결과가 제시된다.

주요용어: 부분선형모형, 변수선택, 순차적 선택, LARS

이 논문은 2021학년도 건국대학교의 연구년교원 지원에 의하여 연구되었음.

[1]교신저자: (48513) 부산시 남구 용소로 45, 부경대학교 응용수학과. E-mail: myoon@pknu.ac.kr