The Korean Journal of Applied Statistics (2020) DOI: http://dx.doi.org/10.5351/KJAS.2020.33.1.001
33(1), 1–10

A procedure for simultaneous variable selection, variable transformation and outlier identification in linear regression

Han Son Seo[a] · Min Yoon[b,1]

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

(Received October 2, 2019; Revised December 3, 2019; Accepted December 18, 2019)

Abstract

We propose a unified approach to variable selection, transformation and outliers in the linear model. The procedure includes a sequential method for outlier detection and a least trimmed squares estimator for variable transformation. It uses all possible subsets regressions for model selection. Some real data analyses and the simulation results are provided to show the efficiency of the methods in the context of the correct variable selection and the fitness of the estimated model.

Keywords: linear regression, outliers, response transformation, variable selection

1. 서론

회귀분석에 관련된 문제들은 서로 복합적으로 연관되어 있어서 각 문제에 대한 개별적인 해결 방법 보다 여러 문제들을 동시에 고려하는 통합적 방법이 필요하다. 회귀분석의 대표적 주제인 변수선택, 이상치 탐지, 변수변환에 관련하여 복수의 주제를 동시에 다룬 방법들이 제안되고 있다.

선형회귀모형에서 비정규성이나 이분산성등을 해결하기 위해 고려되는 변수변환에서 변환추정 과정은 이상치의 영향을 받는다 (Sakia, 1992). 이상치에 대비한 변수변환 절차는 강건추정량을 사용하거나 이상치를 제거하는 방법등이 있다. Atkinson (1986)은 변수변환과정에서 이상치에 의한 수렁현상(swamping phenomenon)을 방지하기 위해 두 단계 추정절차를 제안하였으며 Parker (1988)는 강건추정량인 L1 추정량을 사용하여 변수변환 값을 추정하였다. Atkinson과 Riani (2000)는 Box-Cox 변환에서 스코어 검정(score test)을 통한 이상치 탐지 과정을 제시하였고 Cheng (2005)은 변수변환 과정에서 이상치의 영향력을 제거하기 위하여 절사우도추정법(trimmed likelihood estimation)으로 변환모수를 추정하였다. Seo 등 (2012)은 절사우도추정법 과정에서 이상치의 탐지를 위해 순차적 방법을 적용하였다.

회귀모형의 변수선택과정에서 최소제곱추정법을 사용하는 경우 이상치 문제를 반드시 고려하여야 한다. Wisnowski 등 (2003)은 강건회귀모형 추정을 통한 변수선택법을 제안하였으며 McCann과 Welsch

This work was supported by the Pukyong National University Research Fund in 2016.
[1]Corresponding author: Department of Applied Mathematics, Pukyong National University, 45, Yongso-ro, Nam-Gu, Busan 48513, Korea. E-mail: myoon@pknu.ac.kr

(2007)는 이상치에 해당하는 가변수가 포함된 확대행렬(augmented matrix)에 least angle regressions (LARS)를 적용하여 이상치 영향이 제거된 변수선택 절차를 제시하였다. Kim 등 (2008)은 잠재적 이상치를 탐지한 후 평균이동 이상치 모형에서 가능한 모든 회귀모형(all possible subset regressions)을 비교하여 변수를 선택하였다. Dupuis와 Victoria-Feser (2011)는 가중 M-추정량에 의한 모형 추정과 검정통계량으로 변수선택을 결정하는 fast robust forward selection (FRFS) 방법을 제안하였으며 Dupuis와 Victoria-Feser (2013)는 FRFS 방법의 계산 부담을 줄이기 위하여 variance inflation factor (VIF) 회귀를 응용한 강건 신속 변수선택법을 제안하였다. Seo (2018)는 VIF 회귀법에서 강건추정치 대신 잠재적인 이상치를 탐지하여 이를 분석에서 제외하고 streamwise 절차를 적용하여 변수선택과정의 신속성을 높이는 방법을 제시하였다. Seo (2019)는 이상치가 제거된 데이터로 모형을 적합하기 위해 이상치 탐지와 변수선택이 동시에 수행되는 절차를 제안하였다.

회귀분석에서 변수변환과 변수선택은 이상치뿐만 아니라 서로의 결과에 영향을 받게 되어 두 문제를 동시에 해결하는 접근법이 바람직하다. Yeo (2005)는 선형회귀모형에서 반응변수에 대한 변환과 모형에 참여한 설명변수가 상이한 모형들을 비교하는 문제에서 비내포모형(non-nested model)들 간 검정의 검정수준을 붓스트랩으로 조절하는 절차를 제시하였다. Gottardo와 Raftery (2009)는 반응변수뿐만 아니라 설명변수까지 포함하는 변수변환 문제를 다루면서 베이지안 접근법에 기반하여 각 변수들의 효과에 대한 확률적 계산을 통하여 변수를 선택하고 t-검정을 통해 이상치에 대비하는 절차를 제시하였다.

본 논문에서는 반응변수 변환을 고려하면서 이상치에 강건한 변수선택 방법을 제안한다. 제안된 방법은 반응변수 변환 모형에서 순차적으로 이상치를 탐지, 제외하며 최적의 변수선택은 가능한 모든 회귀모형의 모형적합성을 비교하여 수행한다. 2장에서는 본 연구에서 제안한 변수선택과정에 관여된 변수변환 과정, 이상치 탐지방법, 최적변수 선택절차를 설명한다. 3장에서는 모의실험과 실제 데이터에 적용된 결과에 의해 제안된 방법과 변수선택과정이 생략된 방법의 효율성을 비교하고 4장에서는 연구결과를 요약한다.

2. 이상치 제거와 반응변수 변환을 통한 변수선택법

본 연구에서 제시하는 이상치와 반응변수 변환을 고려한 변수선택 방법의 절차는 대략적으로 다음과 같다. 한 개의 특정 설명변수에 의한 1-변수 모형을 설정하고 이 모형을 기반으로 이상치를 고려한 반응변수 변환 값을 추정한다. 변수변환을 추정하는 과정에서 탐지된 이상치를 제거한 후 변수변환된 모형을 적합하여 모형의 적합도를 계산한다. 모든 1-변수 모형에 대해 이와 같은 과정을 반복하여 최적의 1-변수 모형을 결정하며 동일한 과정을 k-변수 모형, $k = 2, \ldots, p$에 적용하여 최적의 k-변수 모형들을 결정한다. 최종적으로 간결성 원리(rule of parsimony) 등을 고려하여 p개의 모형 중 가장 적합한 모형을 선택한다.

본 연구에서 제안하는 이상치, 변수변환, 변수선택을 동시에 고려하는 절차에서 주요 부분은 이상치의 영향력이 배제된 반응변수 변환을 추정하는 과정이다. 반응변수 변환은 멱 변환(power transformation)으로 한정하며 Box-Cox 변환 (Box와 Cox, 1964)을 사용한다.

$$Y(\lambda) = \begin{cases} \dfrac{Y^{\lambda}}{\lambda}, & \lambda \neq 0, \\ \log Y, & \lambda = 0, \end{cases}$$

Box-Cox 변환에서 모형변환계수 λ는 주로 최대우도추정법에 의해 추정되지만 최대우도추정법은 이상치에 영향을 받는다는 것이 잘 알려져 있다. 최대우도법추정법에서 이상치의 영향력을 배제하기 위한 추정방법은 이상치에 해당하는 관찰치를 제외하고 계산하는 최대절사우도법(maximum trimmed likeli-

hood estimation) (Hadi와 Luceno, 1997)이다. 우도계산에 포함되는 관찰치 집단을 M, 크기를 q라고 하고 $\hat{\beta}_q$는 변환계수가 λ일 때 β의 최대절사우도추정량이라고 할 때 최대절사우도추정량의 목표함수는 $L_q(\hat{\beta}_q) = \sum_{i \in M} \ell(\hat{\beta}_q; y_{(i)}(\lambda))$이 되며 $L_q(\hat{\beta}_q)$는 다음과 같이 분산에 대한 제곱 추정량과 반비례의 관계가 된다.

$$L_q\left(\hat{\beta}_q\right) \propto \frac{(q-p)}{\sum_{i \in M} e_i^2(\lambda)} = \frac{1}{\hat{\sigma}_q^2(\lambda)},$$

여기서 $\hat{\sigma}_q^2(\lambda) = \sum_{i \in M} e_i^2(\lambda)/(q-p)$, $e_i(\lambda) = y_i(\lambda) - x_i^T \hat{\beta}_q$, $i = 1, \ldots, n$이다. 따라서 변환계수 λ의 최대절사우도통계량은 변환계수가 λ일 때 σ^2의 최대절사우도추정량인 $\hat{\sigma_q^2}(\lambda) = \sum_{i \in M} e_i^2(\lambda)/(q-p)$을 최소화하는 추정치와 일치한다.

최대절사우도추정량에서 절사되는 관찰치인 이상치의 정의에 따라 추정량의 성격이 달라진다. Cheng (2005)은 Box-Cox 변환의 변환모수 λ를 추정할 때 낮은 우도의 관찰치를 이상치라고 정의하여 다수의 부표집(subsampling) 과정을 통해 최대절사우도추정량을 계산하였다. 이 방법은 사전에 이상치의 크기가 결정되어야 하며 모든 크기의 이상치에 대하여 추정과정을 적용할 경우 상당한 계산량이 필요하게 된다. 이러한 점을 개선하기 위하여 Seo 등 (2012)은 최대절사우도추정량을 적용할 때 모형의 적합성에 벗어나는 관찰치를 이상치로 정의하고 사전에 이상치 크기를 결정할 필요 없이 각 모형마다 이상치를 탐지, 제거하는 방법을 제안하였다. 이상치 탐지를 위하여 사용되는 방법은 Hadi와 Simonoff (1993)가 제안한 순차적 방법이다. Hadi와 Simonoff (1993)이 제안한 순차적 이상치탐지법은 기초 정상치군에서 시작하여 각 단계별로 이상치군의 크기를 줄여가면서 이상치군에 대한 최종적인 이상치 여부를 결정한다. Hadi-Simonoff 절차의 첫 단계에서는 정상치군만으로 모형을 추정한 후 정상치군과 이상치군에 속한 관찰치에 내적 스튜던트화 잔차(internally studentized residual)를 계산한다. 다음 단계에서는 계산된 내적 스튜던트화 잔차의 순서 통계량에 대한 t-검정을 통하여 순위 순서통계량에 속하는 관찰치를 최종이상치군으로 판단하거나 그렇지 않을 경우 정상치군의 숫자를 한 개 늘려서 앞선 절차를 반복한다. Hadi-Siminoff 방법과 같은 순차적인 방법은 기초정상치군이 전체 절차의 정확성에 매우 중요하게 작용하며 이와 관련하여 Hadi와 Siminoff는 두 가지 기초군 선정 방법을 제시하고 있다.

본 연구에서 제안하는 변수변환 추정과정은 Seo 등 (2012)과 유사하게 순차적 이상치탐지법과 최대절사우도추정법을 적용하여 다음과 같은 절차로 수행된다.

단계 0. 현 단계에서의 반응변수 변환추정값을 $\hat{\lambda}_{pr}$라고 하고 이에 대한 절사제곱추정량을 $S^2(\hat{\lambda}_{pr})$이라고 하자.

단계 1. $y(\hat{\lambda}_{pr})$로 반응변수를 변환한 후 다음과 같은 과정으로 이상치를 탐지한다.

1-1. 현 단계의 정상치군을 M_{pr}이라고 할 때 M_{pr}만으로 모형을 적합시킨 후 다음과 같이 내적 스튜던트화 찬차 d_i를 계산한다.

$$d_i = \begin{cases} \dfrac{y_i\left(\hat{\lambda}_{pr}\right) - x^T \hat{\beta}_M}{\hat{\sigma}_M \sqrt{1 - x_i^T \left(X_M^T X_M\right)^{-1} x_i}}, & \text{if } i \in M_{pr}, \\ \dfrac{y_i\left(\hat{\lambda}_{pr}\right) - x^T \hat{\beta}_M}{\hat{\sigma}_M \sqrt{1 + x_i^T \left(X_M^T X_M\right)^{-1} x_i}}, & \text{if } i \notin M_{pr}. \end{cases}$$

1-2. $|d|_{(j)}$를 $|d_i|$의 크기순 j번째 순서통계량이라고 할 때

(a) $|d|_{(s+1)} \geq t_{(\alpha/2(s+1),s-k)}$이면 마지막 $(n-s)$ 순서 통계량에 해당하는 관찰치를 모두 이상치라고 판단하고 검정을 마친다.

(b) 만약 $|d|_{(s-1)} < t_{(\alpha/2(s+1),s-k)}$이면 $|d|_{(s+1)}$에 해당하는 관찰치를 정상치군에 포함시킨 후 위의 이상치 탐지 과정을 반복하고 $n = s+1$인 경우에는 이상치에 해당하는 관찰치가 없다고 판정한다.

단계 2. 이상치 검정 결과 탐지된 정상치군을 M이라고 하고 크기를 r이라고 할 때 M만을 사용해 변환계수의 잠정적인 추정값 $\hat{\lambda}^{TP}$를 구한다.

단계 3. $\hat{\lambda}^{TP}$에 따라 반응변수를 변환시킨 후 정상치 M만을 가지고 β의 최소제곱추정량(least squares estimator; LSE)인 $\hat{\beta}_M$를 구하고 이를 이용하여 다음과 같이 $S^2(\hat{\lambda}^{TP})$를 계산한다.

$$S^2\left(\hat{\lambda}^{TP}\right) = \sum_{i \in M} \frac{e_i^2\left(\hat{\lambda}^{TP}\right)}{(r-p)}, \quad \text{여기서} \;\; e_i\left(\hat{\lambda}^{TP}\right) = y_i\left(\hat{\lambda}^{TP}\right) - x_i^T \hat{\beta}_M.$$

단계 4. $S^2(\hat{\lambda}^{TP})$와 $S^2(\hat{\lambda}_{pr})$를 비교하여 반응변수 변환값을 조절한다.

(a) 만약 $S^2(\hat{\lambda}^{TP}) > S^2(\hat{\lambda}_{pr})$이면 $\hat{\lambda}_{pr}$을 λ의 최종 추정값으로 결정하고 절차를 끝낸다.

(b) $S^2(\hat{\lambda}^{TP}) \leq S^2(\hat{\lambda}_{pr})$이면 $\hat{\lambda}^{TP}$를 새로운 $\hat{\lambda}_{pr}$값으로 대체한 후 단계1에서부터 위 과정을 반복한다.

이상치 탐지과정에서 최초로 사용되는 기초 정상치군은 최소제곱추정치를 이용한 점진적 생성법으로 선정한다. 이 방법은 모든 관찰치를 회귀모형에 적합시킨 후 잔차가 가장 작은 p개의 관찰치를 선택하고 이 데이터를 기반으로 추정된 모형에서 $|d_i|$가 가장 작은 $(p+1)$개의 관찰치를 선택한다. 같은 과정을 반복하여 최종적으로 크기가 $\text{int}[(n+p-1)/2]$인 기초 정상치군을 선정한다.

각 독립변수들의 조합에 대해 제안된 과정을 통해 이상치와 변수변환을 고려한 모형이 추정되면 모형간 비교를 통해 변수선택이 완성된다. 일반적으로 변수선택은 가능한 모든 변수조합을 고려하는 방법과 순차적 선택방법을 통해 수행된다. 전자의 경우에는 주로 Akaike information criterion (AIC), Bayesian information criterion (BIC), Mallows's C_p, Cross-validation 등의 기준을 적용하고 후자의 방법은 검정을 통해 최적모형을 결정한다 (Zhou 등, 2006). 검정을 통해 변수를 선택할 경우 변수변환과 이상치 제거등이 고려된 상황에서는 비내포 모형, 상이한 관찰치 크기의 문제로 인하여 모형 비교에 적절한 검정통계량을 찾는 것이 쉽지 않다. 따라서 특정기준에 의해 최적모형을 결정하는 절차를 따른다고 할 때 다양한 기준이 적용될 수 있으나 본 연구에서는 상이한 데이터 크기를 고려하는 수정-R^2를 사용하여 모형을 선택한다.

3. 예제와 모의실험

3.1. 모의실험

본 연구에서 제안된 변수선택 방법의 성취도를 평가하고 변수변환을 고려하지 않은 변수선택법과 비교하기 위하여 모의실험을 수행한다. 실험에 사용되는 관찰치중 정상관찰치는 Cheng (2005), Seo 등 (2012), Dupuis와 Victoria-Feser (2013)의 연구에서 사용된 모형과 유사하게 생성되며 데이터에 포함되는 비정상치는 이상치와 지레점(leverage point)으로 구분되어 다양한 비율로 설계된다. 실험에 사용되는 데이터 추출모형은 다음과 같다. 설명변수의 크기를 p, 실제 모형에 포함되는 변수의 크기를 k라고 할 때 정상 관찰치에 해당하는 p개의 설명변수는 범위 $(0, 6)$의 균등분포에서 추출하며 반응변수 Y는

k개 설명변수의 선형모형에서 계산된 값을 변환변수 λ에 따른 Box-Cox 변환 $Y^{(\lambda)}XI_k + \varepsilon$의 역함수를 적용하여 생성한다. Box-Cox 변환은 반응변수에 대하여 단조증가함수이므로 역함수가 존재한다. 이상치에 해당하는 관찰치는 정상치와 동일하게 설명변수를 추출하고 반응변수는 평균 4, 분산 0.5의 정규분포에서 임의 추출 한 후 Box-Cox 변환 역함수를 적용한다. 지레점 관찰치는 p개의 설명변수가 균등분포 대신 다변량 정규분포 $X_i^T \sim MN((14+p)J_{p-1}, 0.5I_{p-1})$에서 추출 생성된다. 모형에 참가하는 k개의 설명변수 외 $(p-k)$개의 노이즈 변수 중 k개의 변수는 모형에 포함되는 변수와 일정한 상관관계를 갖도록 다음과 같이 추출한다.

$$X_{k+\ell} = X_\ell + 3e_{k+\ell}, \quad \ell = 1, 2, \ldots, 2k, \; e_{k+1}, e_{k+2}, \ldots, e_{3k} \overset{\text{iid}}{\sim} N(0,1)$$

나머지 $(p-2k)$개의 노이즈 변수는 정규분포에서 추출한다.

$$X_i = e_i, \quad i = 3k+1, \ldots, p, \; e_{3k+1}, e_{3k+2}, \ldots, e_p \overset{\text{iid}}{\sim} N(0,1)$$

모의실험은 이상치가 없는 경우, 이상치가 존재하는 경우, 지레점이 포함된 경우, 이상치이면서 지레점인 관찰치가 포함된 경우로 나누어 수행된다. 실험의 횟수는 총 100번이고 데이터의 크기는 $n = 40$이다. 설명변수의 크기는 $p = 5$이며 이중 실제모형에 포함되는 변수의 크기는 $k = 2$로 고정하고 그들 간의 상관관계를 나타내는 값 $\theta = 0.1$로 고정한다. 모의실험 데이터의 모수인 변환모수 λ는 -1, -0.5, 0, 0.5, 1이고 이상치 또는 지렛점의 비율은 각각 5%와 10%으로 지정하여 실험한다.

각 방법의 효율성은 세 개의 척도 P_1, P_2, P_3에 의해 계산된다. P_1은 실제모형에 포함된 변수들 집단을 정확하게 찾은 비율이고, P_2는 적어도 한 개 이상의 변수를 찾은 비율이어서 일종의 가면현상(masking phenomenon)이 발생한 비율은 $(1-P_2)$가 된다. P_3은 선택된 변수들 중에 실제모형에 포함되지 않은 변수가 포함된 비율, 즉 수렁현상이 발생한 비율이다. 본 연구에서 제안한 변수변환이 포함된 변수선택방법은 결과를 나타내는 표에서 TR로 표기하며 참고로 비교하는 변수변환 과정이 생략된 방법은 NO-TR로 표기한다.

Table 3.1의 결과에 의하면, 변수변환이 필요하지 않은 $\lambda = 1$인 경우를 제외하고 실험의 모든 경우에 있어서 변수변환이 생략된 방법은 효율성이 매우 낮은 것을 알 수 있다. 다만 $\lambda = 1$인 경우 NO-TR 방법은 작은 차이이지만 변수변환을 수행한 TR 방법보다 높은 효율성을 보이고 있다. 변수변환을 수행한 TR 방법은 이상치와 지레점에 관련된 네 가지 경우 모두에 있어서 매우 정확하게 변수변환 모수를 추정하고 방법의 효율성도 높다. 특히, 지레점만 존재하는 데이터의 경우에는 거의 완벽하게 정확한 변수선택 결과를 보여주는데 이는 지레점이 이상치가 아닌 경우 모형의 적합성이 높아지는데서 기인한 것이다. 이상치의 비율이 5%에서 10%로 높아지면 전반적으로 TR 방법의 효율성이 낮아지며 지레점인 동시에 이상치가 포함된 경우 네 가지 경우중 상대적으로 효율성이 가장 낮다.

3.2. Stackloss 데이터

Stackloss 데이터는 3개의 독립변수 air flow (X_1), inlet temperature (X_2), concentration of acid (X_3)와 반응변수 stack loss (Y)에 대한 21개 관찰치로 구성 되어있으며 이 데이터에 관련된 많은 연구가 수행되었다. 일반적으로 관찰치 1, 3, 4, 21을 이상치로 판정된다 (Brownlee, 1965; Daniel과 Wood, 1980). 변수선택에 관련하여 설명변수 X_3는 1차 선형 모델에서 포함되지 않아야 하며 관찰치 1, 3, 4 및 21은 이상치라고 분석된다 (Kim 등, 2008; Hoeting 등, 1996; Seo, 2019). 변수변환을 고려한 분석에서 Atkinson (1982)은 세 변수를 모두 고려하고 모든 관찰치를 사용 하였을 때 $\lambda = 0.3$, 관찰치 21을 제거 한 경우에는 $\lambda = 0.48$로 추정하였다. Carroll과 Ruppert (1985)는 pseudo-MLE 방법을 사용한 경우 $\lambda = 0.49$, 이상치는 2, 4, 21로 판정하였으며 그들이 제안한 BIT(1.5) 방법에 의하면 모수에 따라

Table 3.1. Model selection results

		$\lambda = -1$		$\lambda = -0.5$		$\lambda = 0$		$\lambda = 0.5$		$\lambda = 1$	
		TR	NO-TR	TR	NO-TR	TR	NO-TR	TR	NO-TR	TR	NO-TR
	Mean	−0.97		−0.49		−0.01		0.49		0.97	
	(SD)	(0.14)		(0.08)		(0.03)		(0.07)		(0.13)	
(a)	P1	0.85	0.00	0.88	0.00	0.99	0.08	0.91	0.03	0.91	0.98
	P2	1.00	0.76	1.00	0.87	1.00	0.99	1.00	1.00	1.00	1.00
	P3	0.15	1.00	0.12	1.00	0.01	0.88	0.09	0.97	0.09	0.02
	Mean	−0.94		−0.49		0.01		0.47		0.94	
	(SD)	(0.15)		(0.09)		(0.03)		(0.09)		(0.19)	
(b)	P1	0.82	0.00	0.84	0.00	0.95	0.07	0.82	0.06	0.79	0.95
	P2	1.00	0.78	1.00	0.79	1.00	0.99	1.00	1.00	1.00	1.00
	P3	0.18	1.00	0.16	1.00	0.05	0.86	0.18	0.94	0.21	0.05
	Mean	−0.99		−0.49		0.00		0.50		1.00	
	(SD)	(0.05)		(0.02)		(0.01)		(0.02)		(0.05)	
(c)	P1	1.00	0.00	1.00	0.00	0.99	0.08	1.00	0.07	1.00	1.00
	P2	1.00	0.70	1.00	0.76	1.00	1.00	1.00	1.00	1.00	1.00
	P3	0.00	1.00	0.00	1.00	0.00	0.86	0.00	0.93	0.00	0.00
	Mean	−0.94		−0.47		0.00		0.48		0.91	
	(SD)	(0.14)		(0.08)		(0.03)		(0.08)		(0.18)	
(d)	P1	0.57	0.00	0.50	0.01	0.67	0.06	0.51	0.09	0.53	0.59
	P2	1.00	0.78	1.00	0.83	1.00	1.00	1.00	1.00	1.00	1.00
	P3	0.43	1.00	0.50	0.99	0.33	0.91	0.49	0.91	0.47	0.41
	Mean	−0.92		−0.47		−0.01		0.45		0.92	
	(SD)	(0.16)		(0.08)		(0.03)		(0.08)		(0.16)	
(e)	P1	0.76	0.00	0.77	0.00	0.95	0.08	0.77	0.17	0.77	0.93
	P2	1.00	0.80	1.00	0.80	1.00	0.95	1.00	1.00	1.00	1.00
	P3	0.24	1.00	0.23	1.00	0.05	0.88	0.23	0.83	0.23	0.07
	Mean	−0.99		−0.50		0.00		0.49		0.99	
	(SD)	(0.05)		(0.02)		(0.01)		(0.02)		(0.05)	
(f)	P1	1.00	0.00	1.00	0.00	0.98	0.07	1.00	0.02	0.98	0.97
	P2	1.00	0.67	1.00	0.72	1.00	0.98	1.00	1.00	1.00	1.00
	P3	0.00	1.00	0.00	1.00	0.01	0.89	0.00	0.97	0.02	0.00
	Mean	−0.93		−0.47		0.00		0.48		0.94	
	(SD)	(0.18)		(0.08)		(0.04)		(0.08)		(0.21)	
(g)	P1	0.41	0.00	0.39	0.00	0.43	0.08	0.37	0.08	0.33	0.40
	P2	1.00	0.80	1.00	0.84	1.00	0.98	1.00	1.00	1.00	1.00
	P3	0.59	1.00	0.60	0.99	0.57	0.90	0.63	0.91	0.66	0.60

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}. Simulated data have $n = 40$ cases with $p = 5$ including $k = 2$ target regressors, and θ, correlation among target regressors, is 0.85. Data were either not contaminated, had high leverage, outliers, or outlying response and high leverage with proportion of 5% or 10%: (a) No contamination, (b) 5% outliers, (c) 5% high leverage, (d) 5% outliers & 5% high leverage, (e) 10% outliers, (f) 10% high leverage, (g) 10% outliers & 10% high leverage. TR = suggested method; NO-TR = a variable selection method without variable transformation procedure.

Table 3.2. Variable selection results using the suggested method for the Stackloss data

Variables	$\hat{\lambda}$	Outliers	Adj-R^2
X_1	0.46	4, 21	0.9645*
X_2	0.11		0.7487
X_3	−0.51		0.2214
X_1, X_2	0.55	4, 21	0.9744**
X_1, X_3	0.45	4, 21	0.9622
X_2, X_3	−0.02		0.7534
X_1, X_2, X_3	0.50	4, 21	0.9722*

*: the best model among k-models; **: the best model overall.

Table 3.3. Model selection results for the Minitab Tree data

Variables	TR			NO-TR	
	$\hat{\lambda}$	Outliers	Adj-R^2	Outliers	Adj-R^2
X_1	0.38		0.9552*		0.9331
X_2	−0.19		0.4093		0.3558
X_3	−0.10		−0.0275		−0.0345
X_4	−0.16		0.0085	26, 27, 28, 29, 30, 31	0.1019
X_1, X_2	0.31		0.9760**		0.9442*
X_1, X_3	0.40		0.9543		0.9352
X_1, X_4	0.31		0.9560		0.9308
X_2, X_3	−0.24		0.4053		0.3124
X_2, X_4	−0.28		0.4102		0.3120
X_3, X_4	−0.19		−0.0162		−0.0648
X_1, X_2, X_3	0.31		0.9751		0.9456
X_1, X_2, X_4	0.24		0.9758*	1, 2, 3, 18	0.9645**
X_1, X_3, X_4	0.33		0.9548		0.9328
X_2, X_3, X_4	−0.35		0.4116		0.2869
X_1, X_2, X_3, X_4	0.23		0.9754*		0.9540*

TR = suggested method; NO-TR = a variable selection method without variable transformation procedure. *: the best model among k-models; **: the best model overall.

$\lambda = 0.39$ 또는 $\lambda = 0.41$과 이상치 2, (3), 4, 21로 판정하였다. 대체로 변수변환 값은 제곱근으로 변수변환 하는 것이 동의된다. 변수변환을 고려한 변수선택을 위해 TR 방법을 적용한 결과는 Table 3.2와 같다. 반응변수를 $\sqrt{Y}$으로 변환하고 X_1(air flow), X_2(inlet temperature), 두 변수로만 모형을 설정하며 관찰치 4, 21을 이상치로 판정한다.

3.3. Minitab Tree 데이터

Minitab Tree 데이터 (Ryan 등, 1976)는 펜실바니아주 Allegheny National Forest에서 표집된 흑체리 나무에 관한 자료이다. 자료는 지상 4.5 피트에서 잰 나무의 지름(X_1)과 나무의 높이(X_2), 나무의 부피(Y)를 담은 31개의 관찰치로 구성되어 있다. 이 데이터에 대하여 Atkinson (1985)은 스코어 검정 결과 반응변수 변환이 필요하며 이상치는 없다고 판단하였다. 본 연구에서는 범위 $(0, 5)$의 균등분포에서 생성한 두 개의 가변수 (X_3, X_4)를 Minitab Tree 데이터에 포함한 후 변수선택 문제를 시도하였다. 변수변환과정이 포함된 방법(TR)과 생략된 방법(NO-TR)을 적용한 결과는 Table 3.3과 같다.

변수변환과정이 포함된 방법에 의한 최적 모형은 독립 변수 X_1, X_2를 포함하고 반응변수를 변환계수 $\hat{\lambda} = 0.31$으로 변환한 모형이며 데이터에서 이상치는 없는 것으로 판정하였고 최적모형의 수정-R^2은 0.976이다. 변수변환과정이 생략한 방법은 이상치 1, 2, 3, 18과 함께 독립변수 X_1, X_2, X_4를 선택하여 실제모형과 비교할 때 변수선택과 이상치탐지 측면에서 수렁현상이 발생하며 TR 방법에 의한 모형 보다 모형의 적합도가 낮다.

4. 결론

본 연구에서는 선형모형에서 반응변수의 변환을 고려하고 이상치의 영향력을 배제한 변수선택 과정을 제안하였다. 제안된 방법에서는 이상치 탐지를 위해 순차적인 탐지법을 적용하며 반응변수 변환값은 탐지된 이상치를 제외한 절사제곱추정법으로 추정하고 최종적인 변수선택은 가능한 모든 모형을 비교하여 수행한다. 본 연구에서 제안한 방법을 Stackloss 데이터와 Minitab Tree 데이터에 적용한 결과 기존 연구와 변수선택이 대체로 일치하였고 변수변환이 고려됨에 따라 이상치 판정에 변화가 있었으며 추정된 모형의 설명력이 향상된 것을 알 수 있었다. 다양한 이상치 설계와 변환계수 값을 가정한 모의실험은 제안된 방법이 이상치만 고려한 방법에 비해 변수선택과 모형설명력 측면에서 더 효율적임을 보여준다. 데이터 규모가 커서 가능한 모든 모형을 고려하는 것이 어려울 경우에는 Seo (2019)에서 처럼 고려해야 할 모형의 숫자를 줄이기 위해 각 특정 변수크기의 최적모형을 기반으로 다음 단계의 최적모형을 선별하는 순차적 절차를 적용할 수도 있고 기초 정상치관찰치 수를 늘린 시점에서 이상치 탐지를 시작할 수 있다.

References

Atkinson, A. C. (1985). *Plots, Transformations and Regression: An Introduction to Graphical Method of Diagnostic Regression Analysis*, Oxford University Press, Oxford.

Atkinson, A. C. (1986). Diagnostic tests for transformation, *Technometrics*, **28**, 29–37.

Atkinson, A. C. and Riani, M. (2000). *Robust Diagnostic Regression Analysis*, Springer, New York.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations (with discussion), *Journal of Royal Statistical Society, Series B*, **26**, 211–246.

Brownlee, K. A. (1965). *Statistical Theory and Methodology in Science and Engineering* (2nd ed), Wiley, New York.

Carroll, R. J. and Ruppert, D. (1988). *Transformation and Weighting in Regression* (2nd ed), Wiley, New York.

Cheng, T. C. (2005). Robust regression diagnostics with data transformations, *Computational Statistics and Data Analysis*, **49**, 875–891.

Daniel, C. and Wood, F. S. (1980). *Fitting Equations to Data: Computer Analysis of Multifactor Data*, John Wiley & Sons, New York.

Dupuis, D. J. and Victoria-Feser, M. P. (2011). Fast robust model selection in large datasets, *Journal of the American Statistical Association*, **106**, 203–212.

Dupuis, D. J. and Victoria-Feser, M. P. (2013). Robust VIF regression with application to variable selection in large data sets, *Annals of Applied Statistics*, **7**, 319–341.

Gottardo, R. and Raftery, A. (2009). Bayesian robust transformation and variable selection: a unified approach, *Canadian Journal of Statistics*, **37**, 361–380.

Hadi, A. S. and Luceno, A. (1997). Maximum trimmed likelihood estimators: a unified approach, examples, and algorithms, *Computational Statistics and Data Analysis*, **25**, 251–272.

Hadi, A. S. and Simonoff, J. S. (1993). Procedures for the identification of multiple outliers in linear models, *Journal of the American Statistical Association*, **88**, 1264–1272.

Hoeting, J., Raftery, A. E., and Madigan, D. (1996). A method for simultaneous variable selection and outlier identification in linear regression, *Computational Statistics and Data Analysis*, **22**, 251–270.

Kim, S., Park, S. H., and Krzanowski, W. J. (2008). Simultaneous variable selection and outlier identification in linear regression using the mean-shift outlier model, *Journal of Applied Statistics*, **35**, 283–291.

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

Parker, I. (1988). Transformations and influential observations in minimum sum of absolute errors regression, *Technometrics*, **30**, 215–220.

Ryan, T. A., Joiner, B. L., and Ryan, B. F. (1976). *Minitab Student Handbook*, Duxbury Press, Mass.

Sakia, R. M. (1992). The Box-Cox transformation technique: a review, *The Statistician*, **41**, 169–178.

Seo, H. S. (2018). Fast robust variable selection using VIF regression in large datasets, *The Korean Journal of Applied Statistics*, **31**, 463–473.

Seo, H. S. (2019). Unified methods for variable selection and outlier detection in linear regression, *Communications for Statistical Applications and Methods*, **26**, 575–582.

Seo, H. S., Lee, G. Y., and Yoon, M. (2012). Robust response transformation using outlier detection in regression model, *The Korean Journal of Applied Statistics*, **25**, 205–213.

Wisnowski, J. W., Simpson, J. R., Montgomery, D. C., and Runger, G. C. (2003). Resampling methods for variable selection in robust regression, *Computational Statistics and Data Analysis*, **43**, 341–355.

Yeo, I. (2005). Variable selection and transformation in linear regression models, *Statistics and Probability Letters*, **72**, 219–226.

Zhou, J., Foster, D. P., and Ungar, L. H. (2006). Streamwise feature selection, *Journal of Machine Learning Researches*, **7**, 1861–1885.

선형회귀에서 변수선택, 변수변환과 이상치 탐지의 동시적 수행을 위한 절차

서한손[a] · 윤민[b,1]

[a]건국대학교 응용통계학과, [b]부경대학교 응용수학과

(2019년 10월 2일 접수, 2019년 12월 3일 수정, 2019년 12월 18일 채택)

요 약

본 연구에서는 선형회귀모형에서 이상치와 변수변환을 고려한 변수선택 알고리즘을 다룬다. 제안된 방법은 잠재적 이상치를 탐지하여 제거한 후 변수변환 추정을 위해 최소 절사 제곱 추정법을 적용하며 가능한 모든 회귀모형을 비교하여 최종적으로 변수를 선택한다. 정확한 변수 선택과 추정된 모델의 적합도의 맥락에서 방법의 효율성을 보여주기 위해 실제 데이터 분석 및 시뮬레이션 결과가 제시된다.

주요용어: 변수변환, 변수선택, 선형회귀, 이상치

이 논문은 2016학년도 부경대학교 연구년 교수 지원사업에 의하여 연구되었음.

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