The Korean Journal of Applied Statistics (2018) DOI: http://dx.doi.org/10.5351/KJAS.2018.31.4.463
31(4), 463–473

Fast robust variable selection using VIF regression in large datasets

Han Son Seo[a,1]

[a]Department of Applied Statistics, Konkuk University

(Received May 2, 2018; Revised June 5, 2018; Accepted June 11, 2018)

Abstract

Variable selection algorithms for linear regression models of large data are considered. Many algorithms are proposed focusing on the speed and the robustness of algorithms. Among them variance inflation factor (VIF) regression is fast and accurate due to the use of a streamwise regression approach. But a VIF regression is susceptible to outliers because it estimates a model by a least-square method. A robust criterion using a weighted estimator has been proposed for the robustness of algorithm; in addition, a robust VIF regression has also been proposed for the same purpose. In this article a fast and robust variable selection method is suggested via a VIF regression with detecting and removing potential outliers. A simulation study and an analysis of a dataset are conducted to compare the suggested method with other methods.

Keywords: large dataset, linear regression, stagewise regression, variable selection

1. 서론

많은 개수의 관찰치와 변수를 가진 대형 데이터를 다루어야 하는 경우가 여러 분야에서 점점 일반화 되고 있다. 대형데이터에 대한 통계분석에서는 변수선택이 중요문제 중 하나이다. 변수 선택의 여러 방법들은 가능한 모든 변수 집단을 모두 고려하는 방법과 순차적으로 변수를 선택하는 방법으로 분류될 수 있다. 가능한 모든 조합의 설명변수들을 고려하고 Akaike information criteria, Bayesian information criteria, Mallows's C_p, cross-validation 등의 특정기준을 적용하여 최적 모형을 찾는 방법들은 대형 데이터의 경우 과도한 계산량으로 인해 사용하기가 어렵거나 불가능하다. 순차적인 변수선택 방법은 고려하는 각 모형의 단계에서 설명변수들이 모형에 새롭게 추가되거나 제거됨으로써 모든 변수의 조합을 고려하는 방법들에 비해 계산량이 감소하지만 단계적 회귀(stepwise regression)와 같이 설명변수들이 단계마다 반복하여 고려됨으로써 대형데이터에서는 여전히 계산량이 문제가 된다. 이에 따라 최근에는 순차적 절차를 따르면서 각 독립변수의 모형진입을 한번만 고려하여 계산 속도를 향상시킨 방법들이 제안되고 있다. 이러한 접근법을 적용한 방법중 대표적인 것은 Zhou 등 (2006)이 제안한 streamwise 변수선택법이며 information-investing 또는 α-investing과 같은 streamwise regression 알고리즘으로 각 변수의 모형 진입 또는 제거를 순차적으로 결정한다.

This paper was supported by the Konkuk University 2017.
[1]Department of Applied Statistics, Konkuk University, 120, Neungdong-ro, Gwangjin-Gu, Seoul 05029, Korea. E-mail: hsseo@konkuk.ac.kr

회귀모형에서 streamwise 변수 선택법은 모형평가 과정과 변수선택 과정으로 구분될 수 있다. Lin 등 (2011)과 Dupuis와 Victoria-Feser (2011, 2013)이 제안한 분산팽창계수(variance inflation factor; VIF) 회귀법은 모형평가과정에서 stagewise 회귀를 설정하여 변수들의 주변상관(marginal correlation)에 의해서 각 단계의 모형을 평가한다. 따라서 Lasso와 least angle regression (LARS)처럼 변수들의 다중공선성에 의한 추정치의 편기가 발생하지만 추정된 VIF로 이를 수정한다. VIF회귀법에서는 stagewise 회귀를 통해 고려된 변수의 선택여부를 α-investing 알고리즘으로 결정한다.

변수선택 문제에서 고려되는 또 하나의 측면은 방법의 강건성(robustness)이다. 변수선택방법에서 모형추정은 주로 최소제곱추정량을 기반으로 수행되기 때문에 이상치에 영향을 받게된다. 이를 해결하기 위하여 기존 모형평가측도들의 강건형 형태들이 제안되었으나 대형데이터에 적용하기에는 계산상의 부담이 크다. Dupuis와 Victoria-Feser (2011)는 가중 M-추정량을 사용하여 모형을 추정하고, 이에 따른 강건 검정통계량에 의하여 변수를 선택하는 fast robust forward selection (FRFS) 방법을 제안하였다. FRFS는 변수가 모형에 선택되어도 다시 모든 변수들이 재 고려되는 과정을 반복하여 설명변수가 많은 경우 계산량이 커진다. 이상치에 대비한 강건 변수선택법을 대형데이터에 적용하는 경우 방법의 신속성이 여전히 중요하다. Dupuis와 Victoria-Feser (2013)는 그들이 제안한 FRFS 방법의 속도를 높이기 위하여 부표본(subsampling)을 사용할 수 있지만 대형데이터의 경우 계산량이 여전히 문제가 될 수 있음을 지적하면서 VIF회귀를 응용한 강건 신속 변수선택법을 제안하였다. 이 방법은 VIF회귀 과정에서 강건추정법으로 모형을 추정하여 다중공선성과 이상치에 대비하고 변수선택에서는 FRFS에서 사용한 t-검정통계량과 유사한 강건검정통계량을 사용한다. Dupuis와 Victoria-Feser (2013)는 VIF회귀법, FRFS, 강건 VIF회귀법들을 이행속도와 변수선택의 유효성 측면에서 비교하였다.

본 논문에서는 신속하게 수행되고 이상치에 강건한 변수선택 방법으로써, VIF회귀법에서 강건추정치를 사용하는 대신 잠재적인 이상치를 탐지하여 분석에서 제외하는 사전심사(pre-screening)과정을 제안한다. 또한 변수선택과정의 신속성을 위해 streamwise 절차와 간편한 이상치 탐지 방법을 적용한다.

2장에서는 VIF회귀법, 강건 VIF회귀법을 소개하고 새로운 강건, 신속 변수선택과정을 제안한다. 3장에서는 모의실험과 실제 데이터를 통하여 제안된 방법과 VIF회귀, 강건 VIF회귀 간 효율성을 비교하고 4장에서는 연구결과를 요약한다.

2. VIF회귀를 이용한 변수 선택법

2.1. 변수선택법의 모형추정

대형 데이터의 변수선택법 중 강건성과 신속성 측면에서 제안된 VIF회귀, FRFS, 강건 VIF회귀의 모형추정과정에 대하여 소개 한다.

다단계 변수선택법에서 현 단계 모형에 포함된 변수의 집합을 M이라고 하고 M에 의한 설명변수 설계행렬을 X_M으로 표시할 때 기존 모형에 또 다른 변수 $x^* \notin M$가 포함될 여부를 판정하기 위하여 다음과 같은 모형을 고려한다.

$$y = X_M^T \beta_M + \beta^* x^* + \varepsilon, \quad \varepsilon \sim N\left(0, \sigma^2 I\right) \tag{2.1}$$

β^*의 최소제곱추정치(least squares estimator; LSE)를 $\hat{\beta}^*$라 하고, $l^2 = x^{*\prime}(I - X_M(X_M^T X_M)^{-1} X_M^T)x^*$으로 표기하면 $\hat{\beta}^* \sim N(\beta^*, l^{-2}\sigma^2)$가 되며 변수 x^*의 유의성에 대한 가설 $\mathrm{H}_0 : \beta^* = 0$은 검정통계량 $t^{(\mathrm{se})} = \hat{\beta}^* / \mathrm{sd}(\hat{\beta}^*)$으로 판정할 수 있다.

대형데이터의 변수선택 문제에서 신속성에 중점을 둔 대표적인 방법은 VIF회귀이다. VIF회귀에서는 변수선택절차의 신속성을 위하여 stepwise 회귀모형인 식 (2.1) 대신 회귀모형 $y = X_M^T \beta_M + \varepsilon$의 잔차

인 r와 x^*에 의한 다음과 같은 stagewise 회귀모형을 고려한다.

$$r = \delta x^* + \bar{\varepsilon}, \quad \bar{\varepsilon} \sim N\left(0, \bar{\sigma}^2 I\right) \tag{2.2}$$

$\hat{\delta}$를 δ의 LSE라고 하면 식 (2.1) 아래에서 $\hat{\delta} = l^2\hat{\beta}^*$이며 가설 $\mathrm{H}_0 : \delta = 0$에 대한 검정통계량 $t^{(\mathrm{sa})} = \hat{\delta}/\widehat{\mathrm{sd}(\hat{\delta})}$와 $t^{(\mathrm{se})}$는 $t^{(\mathrm{sa})} \approx |l| \cdot t^{(\mathrm{se})}$과 같은 관계를 갖는다.

따라서 $l^2 = 1$, 즉 x^*와 X_M이 직교관계에 있으면 $t^{(\mathrm{sa})}$-검정은 $t^{(\mathrm{se})}$-검정과 동일하지만 만약 $l^2 < 1$, 즉 x^*가 X_M에 상관관계를 갖게 되면 식 (2.2)에서 가정한 오차항의 독립성이 지켜지지 않게 되어 $t^{(\mathrm{sa})}$-검정은 변수 x^*의 유의성 검정에 관련하여 편기를 갖게 된다. VIF회귀는 stepwise 회귀에 의한 검정 결과와 비슷하도록 매 단계의 검정에서 편기를 수정한다.

식 (2.1)의 가정에서 $\hat{\delta} \sim N(l^2\beta^*, l^2\sigma^2)$이므로 $\hat{\delta}/|l|\sigma \sim N(0,1)$이 된다. 따라서 σ와 l에 대한 적절한 추정치를 사용하면 $\mathrm{H}_0 : \beta^* = 0$에 대한 검정통계량으로써 $t^{(\mathrm{vp})} = \hat{\delta}/|l|\hat{\sigma}$을 사용할 수 있다. Lin 등 (2011)에서는 식 (2.1)에서 $\beta^* = 0$을 가정한 모형에서 계산된 평균제곱오차의 제곱근(root mean squared error; RMSE)을 $\hat{\sigma}$으로 사용하였다. 이 추정치는 모형식 (2.1)이나 모형식 (2.2)에서 계산되는 RMSE 보다 과적합의 가능성이나 편기가 작다 (Foster와 Stine, 2008). l은 $1/l^2$이 x^*의 X_M에 대한 분산팽창계수에 해당하므로 정확한 계산이 가능하지만 대형데이터의 경우 계산량을 고려하여 전체 데이터에서 추출한 임의의 표본에서 계산된 분산팽창계수를 사용한다. Lin 등 (2011)에서 제시한 VIF회귀의 모형평가 과정을 요약하면 다음과 같다.

회귀모형 $y = X_M^T\beta_M + \varepsilon$에서 LSE에 의한 회귀모형을 추정한 후 잔차 r과 RMSE, $\hat{\sigma^2} = \|r\|/(n - |M| - 1)$을 계산하며 회귀모형 $r = \delta x^* + \bar{\varepsilon}$에서 최소제곱추정법으로 $\hat{\delta}$을 계산한다. 전체 데이터에서 크기 m의 표본을 추출하고 표본자료(S)만을 가지고 회귀모형 $x^* = X_M^T\beta_M + \varepsilon$을 추정한 후 분산팽창계수 $\mathrm{VIF}_S = 1/(1 - R_S^2)$를 계산한다. 변수에 대한 검정통계량 $t^{(\mathrm{vp})} = \hat{\delta}\sqrt{\mathrm{VIF}_S}/\hat{\sigma}$을 사용한다.

대형데이터의 변수선택법에서 극단치의 영향력을 제한하기 위해 강건추정량을 사용할 수 있다. Dupuis 등 (2011)은 변수선택을 위한 모형설정과정에서 가중 M-추정치 사용을 제안하였다.

강건모형추정은 가중치 w_i, $i = 1, \ldots, n$에 대하여 식 (2.3)의 해를 계산한 가중최소자승법으로 β를 추정한다.

$$\sum_{i=1}^{n} w_i(r_i; c) r_i x_i = 0, \tag{2.3}$$

가중치 w_i는 표준화 잔차 $r_i = (y_i - x_i^T\beta)/\sigma$에 의해 결정되며 Dupuis와 Victoria-Feser (2011)은 다음과 같은 Tukey의 biweight를 사용하였다.

$$w_i(r_i; c) = \begin{cases} \left(\left(\dfrac{r_i}{c}\right)^2 - 1\right)^2, & |r_i| \le c, \\ 0, & |r_i| > c, \end{cases} \tag{2.4}$$

$\hat{\beta}$에 의해 $\hat{\sigma}$이 계산되고 이에 따라 r_i이 계산됨으로 식 (2.3)을 만족하는 해를 찾는 것은 간단하지 않다. 따라서 저차원의 모형에 의해 식 (2.3)을 만족하는 추정치 $\hat{\beta}^0$을 계산한 후 이를 기반으로 $\hat{\sigma}^0$, $\hat{r}_i^0$을 추정하고 이에 따라 가중치 w_i^0를 계산한다. 가중치 w_i^0, $i = 1, \ldots, n$으로 구성된 벡터를 w^0라고 표기할 때 고정된 가중치 w^0에 대해 $X^w = \mathrm{diag}(\sqrt{w^0}X)$, $y^w = \mathrm{diag}(\sqrt{w^0})y$라고 하면 식 (2.3)의 해는 $\hat{\beta} = ((X^w)^T X^w)^{-1}(X^w)^T y^w$이 된다. Dupuis와 Victoria-Feser (2011)가 $\hat{\beta}^0$, $\hat{\sigma}^0$을 추정하고 가중치 w_i^0을 계산하기 위하여 제안한 모형추정의 사전단계는 다음과 같다.

각 독립변수에 대한 p개의 단순선형모형 $y = \beta_{01} + \beta_1 x_1 + \varepsilon_1, \ldots, y = \beta_{0p} + \beta_p x_p + \varepsilon_p$을 설정한 후 각 모형에 대하여 식 (2.3)을 적용하여 절편추정치 $\hat{\beta}_{01}, \ldots, \hat{\beta}_{0p}$와 기울기 추정치 $\hat{\beta}_1, \ldots, \hat{\beta}_p$를 계산한다. p개의 각 선형모형 추정식에서 $\hat{\sigma}_j = \text{MAD}(y_i - \hat{\beta}_{0j} - \hat{\beta}_j x_{ij})$, $r_{ij} = (y_i - \hat{\beta}_{0j} - \hat{\beta}_j x_{ij})/\hat{\sigma}_j$, $j = 1, \ldots, p$을 계산하고 이에 따라 가중치 w_{ij}를 계산한다. $X_b^w = [1 \ \sqrt{w_{i1}} x_{i1} \ \cdots \ \sqrt{w_{ip}} x_{ip}]$, $X_b^{w^2} = [1 \ w_{i1} x_{i1} \ \cdots \ w_{ip} x_{ip}]$, $i = 1, \ldots, n$이라고 하면 β의 잠정 추정치 $\hat{\beta}_0 = ((X_b^w)^T X_b^w)^{-1} (X_b^{w^2})^T y$을 계산한다. 잠정 추정치 $\hat{\beta}_0$에 의하여 $\hat{\sigma}^0 = \text{MAD}(y_i - x_i^T \hat{\beta}^0)$, $r_i^0 = (y_i - x_i^T \hat{\beta}^0)/\hat{\sigma}^0$을 계산하고 다시 한번 가중치 w_i^0를 구한 후 이것을 기반으로 β의 추정치 $\hat{\beta} = ((X^w)^T X^w)^{-1} (X^w)^T y^w$을 계산한다.

Dupuis와 Victoria-Feser (2011)는 기존의 모형에 포함되지 않은 각 변수와 y간의 가중부분상관계수(weighted partial correlation)를 계산하여 가장 큰 상관관계에 있는 변수에 대한 검정을 통해 포함여부를 결정하였으며 각 단계의 모형추정 과정에서 앞서 설명한 것처럼 p개의 단순선형모형을 이용한 방법(FRFS-Marginal)외에 식 (2.3)의 해를 직접적으로 계산하여 M-추정치를 계산하는 방법(FRFS-Full)을 제안하였다.

FRFS는 강건 변수선택법이지만 대규모 데이터에서 수행속도가 매우 느리다. 이를 개선 하기 위하여 Dupuis와 Victoria-Feser (2013)는 강건모형추정과 VIF 선택법을 결합한 변수선택과정을 제안하였다. Dupuis와 Victoria-Feser (2013)에서 사용한 가중치 함수는 다음과 같은 Huber 가중함수이다.

$$w_i(r_i; c) = \min\left\{1, \frac{1}{\|r_i\|}\right\}. \tag{2.5}$$

식 (2.1)에서 기존의 설명변수집단 M을 확정할 때 사용된 가중치벡터를 w_M^0이라고 하고 추가 여부를 고려하는 변수 $x^* \notin M$만의 주변모형에서 식 (2.3)에 의해 계산된 가중치 벡터를 $w*$라고 할 때 $X_M^w = \text{diag}(\sqrt{w_M^0} X_M)$, $x^{*w} = \text{diag}(\sqrt{w^*}) x^*$, $y^w = \text{diag}(w_M^0) y$으로 정의하자. 식 (2.1)에서 β^*에 대한 강건추정치 $\hat{\beta}^{*w}$는 다음과 같다.

$$\hat{\beta}^{*w} = \left(x^{*w^T} x^{*w} - x^{*w^T} X_M^w \left(X_M^{w^T} X_M^w \right)^{-1} X_M^{w^T} x^{*w} \right)^{-1} \left(x^{*w^T} x^{*w} \right) \left(x^{*w^T} x^{*w} \right)^{-1} x^{*w^T} r_M^w,$$

여기서 r_M^w은 $y^w = X_M^{w^T} \beta_M^w + \varepsilon$ 모형의 잔차이다. 회귀모형 $r_M^w = \delta^w x^{*w} + \tilde{\varepsilon}$에서 $\hat{\delta}^w$에 대한 가중추정치를 $\hat{\delta}^w$라고 하고 $x^{*w} = X_M^{w^T} \beta_M^w + \varepsilon$ 모형의 분산팽창계수를 VIF^w라고 한다면, $\hat{\delta}^w = \hat{\beta}^{*w}/\sqrt{\text{VIF}^w}$가 된다. 따라서 $\hat{\sigma}^2 \approx \hat{\sigma}^2$가 되는 적절한 분산의 추정치를 계산한다면 새로운 변수 x^*의 유의성에 대한 강건검정통계량 T_w는 다음과 같이 정의될 수 있다. $T_w = \hat{\beta}^{*w}/\sqrt{\text{Var}(\hat{\beta}^{*w})} \approx \sqrt{\text{VIF}^w}\hat{\delta}^w/\sqrt{\hat{\sigma}^2(\sum_{i=1}^n x_i^{*w^2})^{-1}\kappa_c^{-1}}$ 여기서 κ_c는 $\widehat{\text{Var}(\hat{\beta}^{*w})}$에 포함되는 상수이며 가중치함수 식 (2.5)의 c값에 의해 결정된다. Dupuis와 Victoria-Feser (2013)에서는 $\hat{\sigma}^2 = \text{MAD}(r_M^w - x^{*w}(x^{*w^T} x^{*w})^{-1} x^{*w^T} r_M^w)$을 사용하였으며 계산량을 고려하여 VIF^w는 전체데이터에서 추출된 임의의 표본에 의하여 계산하였다.

2.2. 변수선택법의 검정절차

변수선택법은 모형추정과정과 변수선택과정을 통해 수행된다. Streamwise 회귀는 절차의 신속성을 위하여 stagewise 회귀와 같이 각 단계에서 한 개의 변수만을 고려하여 모형화하고 해당 변수의 추가여부를 검정하는 절차로 구성된다. Zhou 등 (2006)은 streamwise 회귀에서 변수 추가 여부를 위해 information-investing과 α-investing을 이용한 알고리즘을 제안하였으며 VIF회귀와 강건 VIF회귀에서는 α-investing 알고리즘을 사용하여 FRFS보다 신속하게 변수선택과정을 수행한다.

information-investing 알고리즘과 α-investing 알고리즘의 체계는, 주어진 자원(wealth)을 가지고 매 단계 검정에서 일정한 자원을 투자(investment)하여 그 결과에 따라 자원이 증가하거나 감소한다고 할 때 자원이 소진할 때까지 투자 또는 검정을 수행하는 것이다. 이 때 매 단계의 투자환경, 검정여건은 단계별로 달라진다. α-investing 알고리즘은 일종의 일종오류(Type I error; α) 허용치를 자원으로 사용하며 각 단계에서 검정이 기각되어 변수가 선택되면 자원이 증가하고 검정이 채택되면 자원이 감소한다. 매 단계에서 수행되는 검정의 유의수준은 이전 단계들의 검정여부와 검정의 회수에 따라 조정된다. 유의수준은 검정회수가 많아질수록 유의수준이 작아져서 많은 수의 변수가 포함되는 과 적합 모형을 경계한다. 전 단계에서 검정이 기각되어 변수가 선택되면 다음단계에서 유의수준은 커져서 변수 선택에 보다 더 허용적인 검정을 수행하게 되며 전 단계에서 검정이 채택되어 변수가 선택되지 않을 경우 다음 단계의 유의수준은 작게 설정되어 변수 선택이 어려워지는 환경이 된다. 구체적인 α-investing 알고리즘은 다음과 같이 요약할 수 있다.

최초 자원을 w_0라고 하고 i번째 검정에서 자원을 w_i라고 표기하면, i번째 검정은 유의수준 $\alpha = w_i/(1-i-h)$으로 수행된다. 여기서 h는 가장 최근에 귀무가설이 기각된 검정 순서이다. 유의수준 α_i에 의한 i번째 검정 결과 귀무가설이 기각되면 대상 변수가 모형에 포함되고 $i+1$번째 검정의 자원이 $w_{i+1} = w_i + w_\Delta$로 증가되어 적용되는 유의수준도 커진다. 반면에 i번째 검정에서 귀무가설이 채택되면 대상 변수를 모형에 포함시키지 않고 다음의 $i+1$번째 자원은 $w_{i+1} = w_i - \alpha_i/(1-\alpha_i)$로 감소하고 유의수준도 따라서 감소하게 된다. 위 과정은 최초 자원 w_0으로 시작하여 자원이 고갈될 때 까지 반복 수행한다.

α-investing 알고리즘은 순차적 검정에서 Bonferroni 규칙을 적용할 때 나타나는 유의수준의 보수적 경향성을 지양하고 일종의 오류기각 비율인 mFDR (False Discovery Rate)을 조절할 수 있다 (Foster와 Stine, 2008). 구체적으로 mFDR은 $E(F)/(E(R)+\eta)$으로 정의되며, 여기서 F는 전체 검정에서 일종오류가 발생한 횟수이고 R은 귀무가설을 기각한 총 횟수이며 η는 조절상수이다.

2.3. 이상치 탐지를 통한 강건 변수선택법

자료에 극단치가 존재할 때 FRFS와 강건 VIF회귀처럼 강건통계량을 사용하는 대신 극단치를 탐지하여 제거하는 접근법을 사용할 수 있다. 대형 자료의 경우에도 변수선택문제에서 이른바 pre-screening 방법이 시도된다. Fan과 Lv (2008)는 고차원 데이터의 변수선택 문제에서 모형에 포함될 가능성이 낮은 변수들을 일차적으로 탐지, 제거하는 방법을 제시하였다. 본 논문에서 제안하는 신속하고 강건한 변수선택 방법은 대형자료에 존재하는 극단치를 제거한 후 VIF회귀를 수행하는 것이다.

일반적인 VIF회귀는 강건 VIF회귀나 FRFS 방법보다 신속하게 수행된다. 강건 VIF회귀나 FRFS가 반복적으로 가중치와 모형을 추정하여야 하는 반면 VIF회귀 방법은 한 번의 추정 과정만 수행한다. Dupuis와 Victoria-Feser (2013)에서 VIF회귀방법, FRFS, 강건 VIF회귀 방법들 간의 속도비교에서 VIF회귀의 신속성이 확인된다.

본 논문에서 제시하는 방법은 VIF회귀 방법에서 VIF추정을 위해 추출되는 표본의 대상을 이상치가 제거된 데이터로 제한하는 것이다. 극단치 탐지는 각 단계에서 모형에 포함되는 변수의 구성에 따라 결과가 달라질 수 있다 (Hadi와 Simonoff, 1993). 각 단계에서 극단치 탐지과정을 수행하면 정확성이 높아지는 반면 대형 자료의 경우 계산량이 과도하게 커진다. 본 연구에서는 모든 변수가 포함된 모형에서 각 관찰치별 t-검정을 통해 표본의 대상이 되는 데이터 집단을 구성하고 이 집단을 전 단계의 표본대상으로 사용한다. 따라서 기존의 VIF회귀에 한 번의 모형추정 과정이 추가되므로 계산량은 큰 변화가 없다.

본 연구에서 제안한 방법은 속도와 강건성 측면에서 다른 방법들과 비교될 수 있으나 본 연구에서 제

안한 방법과 계산량에 큰 차이가 없는 VIF회귀방법이 강건 VIF회귀방법이나 FRFS 방법보다 신속하므로 (Dupuis와 Victoria-Feser, 2013) 방법들 간 속도 비교는 생략한다. 또한 강건성 측면에서 강건 VIF회귀방법이 FRFS 방법보다 우수하므로 (Dupuis와 Victoria-Feser, 2013) 본 연구에서는 이상치제거 VIF회귀방법, VIF회귀방법, 강건 VIF회귀방법간 강건성을 비교하기로 한다.

3. 예제와 모의실험

3.1. 모의실험

본 연구에서 제안된 변수선택방법과 기존의 VIF회귀, 강건 VIF회귀의 강건성을 비교하기 위하여 모의실험을 수행한다. 모의실험은 Dupuis와 Victoria-Feser (2013)에서 수행된 실험과유사하게 설계된다. 실험에 사용되는 데이터 추출모형은 다음과 같다.

설명변수의 크기를 p라고 하고 실제 모형에 포함되는 변수의 크기를 k라고 한다면 처음의 k개 변수는 다변량 정규분포에서 추출한다.

$$X_1, X_2, \ldots, X_k \sim \text{MVN}(\mathbf{0}, \Sigma_1), \quad \Sigma_1(\text{diagonal} = 1, \text{off-diagonal} = \theta)$$

$(p-k)$개의 노이즈 변수중 $2k$개의 변수는 모형에 포함되는 변수와 일정한 상관관계를 갖도록 추출한다.

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

나머지 $(p-3k)$개의 노이즈 변수는 정규분포에서 추출한다.

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

독립변수 y는 추출된 데이터의 선형모형에서 계산된다.

$$y = X_1 + X_2 + \cdots + X_k + \sigma\varepsilon, \quad \varepsilon \sim N(0,1)$$

실험의 횟수는 총 200번이고 각 데이터의 크기는 $n = 1000$이며 설명변수의 크기는 $p = 100$과 $p = 1000$이다. 실제모형에 포함되는 변수의 크기는 $k = 5$로 고정하고 그들 간의 상관관계를 나타내는 $\theta = 0.1$과 $\theta = 0.85$로 지정하며 VIF 추정을 위해 추출하는 표본의 크기는 $m = 200$이다. 강건 VIF회귀에서 사용할 가중치 함수는 $c = 1.345$ 값의 식 (2.5)를 사용하며 변수선택 과정으로 적용할 α-investing 알고리즘의 최초자원은 0.5, 자원 증가분은 0.05로 지정한다.

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_3는 선택된 변수들 중에 실제모형에 포함되지 않은 변수가 포함된 비율, 즉 수렁현상(swamping phenomenon)이 발생한 비율이다. 비교대상의 세 가지 변수추출방법은 표와 그림에서 FastVIF, RobVIF, OutdVIF로 표기되며 각각 Lin 등 (2011)이 제안한 VIF 방법, Dupuis와 Victoria-Feser (2013)의 강건 VIF 방법, 본 연구에서 제안한 이상치 제거후 VIF 방법을 나타낸다.

Table 3.1을 보면 모든 경우에 있어서 대체로 OutdVIF 방법, RobVIF 방법, FastVIF 방법의 순서로 효율적인 것을 알 수 있다. 다만 모형참여변수간 상관관계가 낮을 때 RobVIF 방법은 5% 이상치 경우와 5% 이상치및 지레점 경우 P_2와 P_3관점에서 상대적 우위를 보이고 있으나 이는 최종 선택된 변수의 크기가 작은데서 기인한 것이다.

Table 3.1. Model selection results

		$\theta = 0.1$						$\theta = 0.85$					
		$p = 100$			$p = 1000$			$p = 100$			$p = 1000$		
		P_1	P_2	P_3	P_1	P_2	P_3	P_1	P_2	P_3	P_1	P_2	P_3
No cont.	FastVIF	0.135	1.000	0.825	0.370	1.000	0.380	0.160	1.00	0.840	0.170	1.000	0.815
	RobVIF	0.335	1.000	0.635	0.385	1.000	0.280	0.240	1.00	0.720	0.185	1.000	0.705
	OutdVIF	0.440	1.000	0.535	0.695	1.000	0.215	0.495	1.00	0.505	0.470	1.000	0.530
5% hl.	FastVIF	0.140	1.000	0.830	0.145	1.000	0.480	0.105	1.00	0.690	0.045	1.000	0.680
	RobVIF	0.270	1.000	0.590	0.105	0.965	0.330	0.215	1.00	0.720	0.090	1.000	0.700
	OutdVIF	0.455	1.000	0.530	0.665	1.000	0.210	0.350	1.00	0.650	0.375	1.000	0.625
5% outliers	FastVIF	0.000	0.035	0.065	0.000	0.000	0.050	0.000	0.99	0.780	0.000	0.995	0.535
	RobVIF	0.185	1.000	0.625	0.185	1.000	0.790	0.120	1.00	0.840	0.070	1.000	0.795
	OutdVIF	0.365	1.000	0.515	0.225	0.895	0.700	0.420	1.00	0.580	0.665	1.000	0.325
5% hl & out.	FastVIF	0.000	0.035	0.110	0.000	0.155	0.070	0.000	0.98	0.800	0.000	0.995	0.675
	RobVIF	0.085	1.000	0.870	0.055	1.000	0.780	0.115	1.00	0.855	0.040	1.000	0.830
	outdVIF	0.115	0.755	0.550	0.070	0.995	0.425	0.265	1.00	0.715	0.530	1.000	0.435

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 $n = 1000$ cases with $p = 100$ and $p = 1000$ including $k = 5$ target regressors, and θ, correlation among target regressors, is 0.1 and 0.85. Data were either not contaminated, had 5% high leverage, 5% outliers, or 5% outlying response and high leverage. FastVIF = fast VIF; RobVIF = robust VIF; OutdVIF = outlier-detect VIF.

Table 3.2. Number of selected variables in 100 random ordering

	First-order terms model			Second-order interactions model		
	FastVIF	RobVIF	OutdVIF	FastVIF	RobVIF	OutdVIF
Mean	8.90	9.70	11.60	16.00	17.20	13.20
SD	2.04	2.13	1.20	3.21	5.68	6.23
Max	13.00	13.00	13.00	23.00	32.00	21.00
Min	5.00	4.00	7.00	9.00	8.00	2.00

FastVIF = fast VIF; RobVIF = robust VIF; OutdVIF = outlier-detect VIF.

실험 데이터에서 각 변수선택방법을 사용하여 최종 추정된 모형을 크기 $n = 1000$의 또 다른 실험 데이터에 적용하여 방법의 효율성을 측정한다. 효율성의 측도는 강건측도인 절대예측오차 중위수(median absolute prediction error; MAPE)를 사용한다.

Figure 3.1의 결과를 보면 이상치가 없는 경우(no contamination)에서 $p = 1000$, $\theta = 0.1$인 경우를 제외하고 FastVIF와 OutdVIF방법의 효율성이 비슷하며 RobVIF의 MAPE값은 상대적으로 크다. 그 외의 경우에는 OutdVIF방법이 RobVIF와 대등하거나 나은 결과를 보여주고 있으며 FastVIF는 이상치에 잘 대응하지 못하고 있음을 알 수 있다.

3.2. Boston housing 자료

변수선택 방법을 실제 데이터에 적용하는 예제로서 Boston housing 자료 (Harrison과 Rubinfeld, 1978)를 사용하기로 한다. 1970년 인구조사의 결과인 Boston housing 자료는 보스턴 인근 주택가격의 중앙값과 이에 영향을 미치는 13개의 변수를 포함하고 있고 관찰치의 크기는 506개이다 ($n = 506$, $p = 13$). 각 변수에 대한 정보는 다음과 같다.

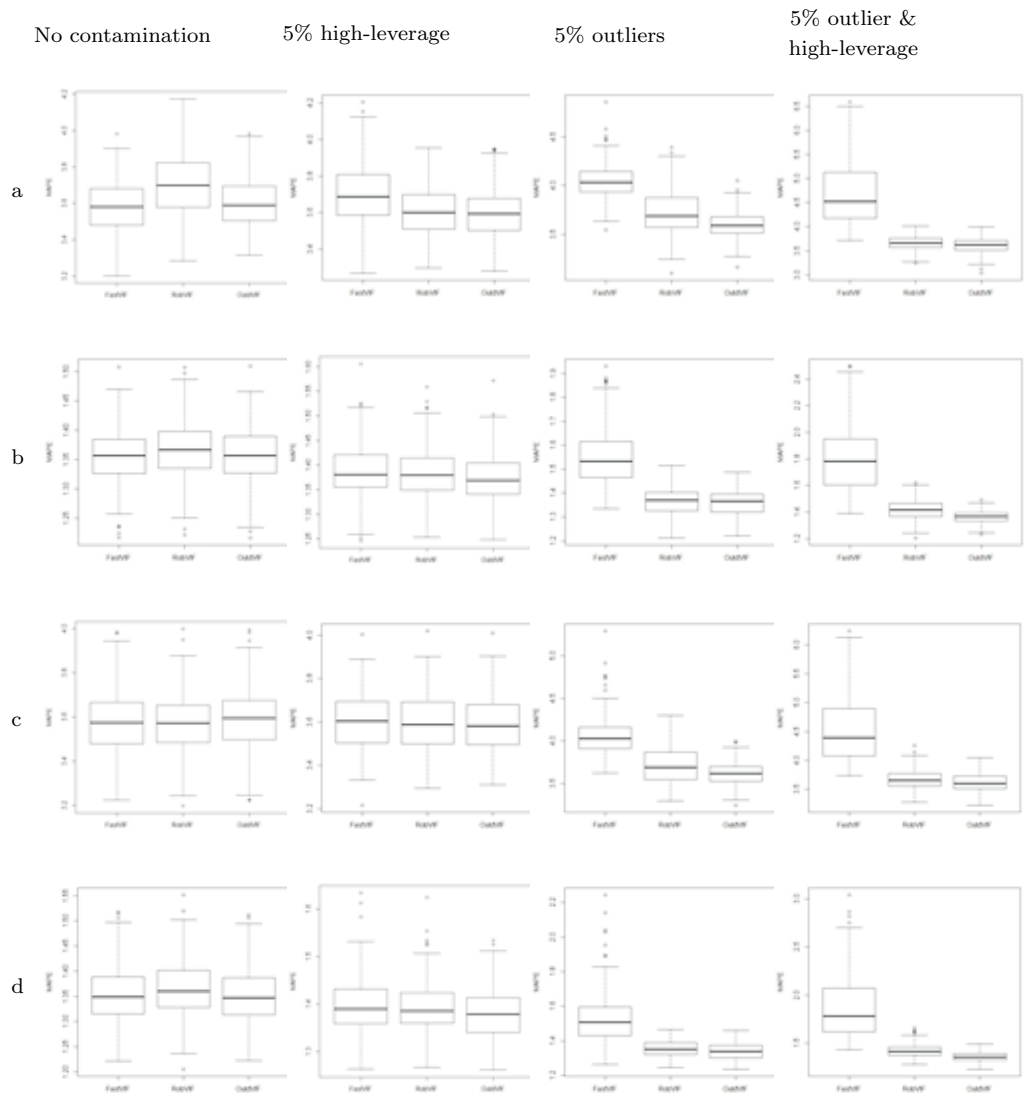


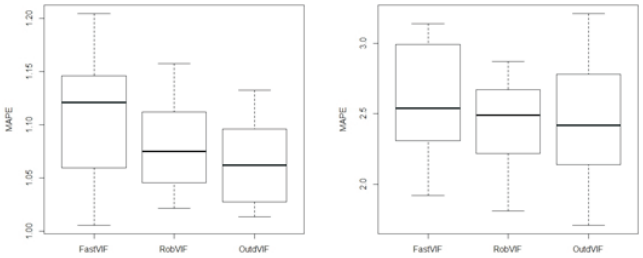
Figure 3.1. Out-of-sample mean absolute prediction errors of the models estimated by fast VIF and robust VIF regression, and outlier-detect VIF regression. Simulated data have $n = 1000$ cases with p including $k = 5$ target regressors, and θ, correlation among target regressors; a: $(p = 100, \theta = 0.1)$, b: $(p = 100, \theta = 0.85)$, c: $(p = 1000, \theta = 0.1)$, d: $(p = 1000, \theta = 0.85)$. VIF = variance inflation factor.

- AGE : 1940년 이전 건축된 자가주택의 비율
- B : $1000(\text{시별 흑인비율} - 0.63)^2$
- CHAS : 찰스강변 여부에 대한 더미변수
- CRIM : 시 별 1인당 범죄율
- DIS : 5개의 보스턴 직업센터까지의 가중 거리

Table 3.3. Number of selected times of each variable in 100 random ordering

	age	b	chas	crim	dis	indus	lstat	nox	ptratio	rad	rm	tax	zn
FastVIF	35	74	85	59	82	41	100	59	100	42	100	64	49
RobVIF	83	100	63	45	88	55	69	52	100	71	100	82	62
OutdVIF	93	98	54	97	57	100	100	78	100	83	100	97	99

FastVIF = fast VIF; RobVIF = robust VIF; OutdVIF = outlier-detect VIF.



(a) First-order only (b) With interactions

Figure 3.2. Boston housing data: Out-of-sample median absolute prediction errors of the models, in 10-fold cross-validation.

- INDUS : 비소매상업지역 토지의 비율
- LSTAT : 모집단의 하위계층의 비율
- NOX : 10ppm 당 농축 일산화질소
- PTRATIO : 시별 학생/교사 비율
- RAD : 방사형 도로까지의 접근성 지수
- RM : 주택 당 평균 방의 개수
- TAX : 10,000 달러 당 재산세율
- ZN : 25,000 평방피트를 초과하는 거주지역의 비율
- MEDV : 본인 소유 주택가격의 중앙값

VIF 추정을 위해 추출하는 부표본의 크기는 $m = 50$이다. Table 3.2는 각 방법의 stagewise 회귀추정 단계에서 모형에 적용되는 변수의 순서가 결과에 미치는 영향력을 상쇄하기 위하여 변수의 순서를 100회 임의로 정하여 수행한 결과 각 방법에서 선택한 변수의 개수를 보여준다. 일차선형모형에서 OutdVIF는 FastVIF와 RobVIF 보다 많은 수의 변수를 선택하고 교호작용모형에서는 반대로 적은수의 변수를 선택하는 경향이 있음을 알 수 있다. 또한 OutdVIF는 일차선형모형에서 변수의 순서에 따라 선택하는 변수 갯수의 변동이 작으나 교호작용모형에서는 변수의 순서에 영향을 다른 방법에 비해 상대적으로 많이 받는다. Table 3.3은 일차선형모형 가정 아래 100회의 무작위 순서에 의해 변수선택법을 수행한 결과 각 변수마다 선택된 횟수를 나타낸다. 세 방법에서 공통적으로, 모든 시행에서 선택되는 변수는 ptratio, rm 이다. 다른 방법에 비해 OutdVIF는 crim, indus, zn변수를 상대적으로 더 자주 선택한다. Figure 3.2는 10-분할 CV에 의하여 MAPE를 계산한 결과이다. 일차선형모형과 교호작용모형에서 OutdVIF는 FastVIF와 RobVIF보다 높은 효율성을 보여준다. 특히 교호작용모형에서는 선택변수의 개수가 작음에도 OutdVIF의 MAPE가 작음을 알 수 있다.

대형데이터의 변수선택문제에서 많이 사용되는 또 하나의 데이터는 College data이다 (Stock과 Watson, 2007). College data는 Dupuis와 Victoria-Feser (2013)에서 자세하게 분석되었으며 본 연구에서도 College data를 적용하여 방법들을 비교한 결과 각 방법들 간의 효율성은 Boston housing data의 결과와 유사하였다.

4. 결론

본 연구에서는 대형 데이터의 변수선택 문제에서 강건성과 신속성을 고려하여 stagewise 회귀와 α-알고리즘이 수행되면서 한 번의 이상치 탐지절차만 추가되는 방법을 제안하였다. 강건성의 관점에서 단계별로 이상치를 탐지하는 방법을 고려해 볼 수 있으나 대형데이터의 경우 계산량의 문제가 발생할 수 있다. 모의실험과 예제를 통해 본 연구에서 제안한 이상치 제거를 통한 변수선택방법은 기존의 강건추정법에 의한 변수선택방법보다 더 효율적이라는 것을 알 수 있다. 변수선택의 변수절차로 적용된 α-알고리즘에서 각종 모수에 대한 최적값은 여전히 해결해야 할 문제이며 많은 변수를 포함하는 대형데이터의 변수선택 문제에 대비하여 본 연구에서 제안한 이상치에 관련된 사전선별작업을 관찰치 뿐만 아니라 설명변수에도 동시에 적용하는 이차원 사전선별 과정을 추후에 연구해 볼 수 있다.

References

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.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space, *Journal of the Royal Statistical Society. Series B*, **70**, 849–911.

Foster, D. P. and Stine, R. A. (2008). α-investing: a procedure for sequential control of expected false discoveries, *Journal of the Royal Statistical Society. Series B*, **70**, 429–444.

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.

Harrison, D. and Rubinfeld, D. L. (1978). Hedonic prices and the demand for clean air, *Journal of Environmental Economics and Management*, **5**, 81–102.

Lin, D., Foster, D. P., and Ungar, L. H. (2011). VIF regression: a fast regression algorithm for large data, *Journal of the American Statistical Association*, **106**, 232–247.

Stock, J. H. and Watson, M. W. (2007). *Introduction to Econometrics*, 2nd ed. Boston: Addison Wesley.

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

대형 데이터에서 VIF회귀를 이용한 신속 강건 변수선택법

서한손[a,1]

[a]건국대학교 응용통계학과

(2018년 5월 2일 접수, 2018년 6월 5일 수정, 2018년 6월 11일 채택)

요 약

본 연구에서는 선형회귀모형을 가정한 대형 데이터에서의 변수선택 알고리즘을 다룬다. 방법의 속도와 강건성에 주안점을 둔 여러 알고리즘들이 제안되었다. 그 중에서 streamwise 회귀 접근법을 사용한 VIF회귀는 신속하고 정확하게 수행된다. 그러나 VIF회귀는 최소제곱방법에 의해 모형이 추정되므로 이상치에 민감하다. 변수선택방법의 강건성을 높이기 위해 가중 추정치를 사용한 강건측도가 제안되었으며 강건 VIF회귀도 제안되었다. 본 연구에서는 잠재적 이상치를 탐지하여 제거한 후 VIF회귀를 수행하는, 빠르고 강건한 변수선택 방법을 제안한다. 제안된 방법은 모의실험과 데이터 분석 통해 다른 방법들과 비교된다.

주요용어: 단계별 회귀, 대형자료, 변수선택, 선형회귀

이 논문은 2017학년도 건국대학교 KU학술연구비의 지원에 의한 논문임.

[1](05029) 서울시 광진구 능동로 120, 건국대학교 응용통계학과. E-mail: hsseo@konkuk.ar.kr