제주교육대학교 논문집 (2001)
제30집 pp. 327~336

2段階 方法의 기저집합에서 공선성 문제

金 鍾 雨*

〈 목 차 〉

1. 서 론

선형회귀분석(linear regression analysis)은 자료의 특성을 분석하기 위하여 가장 널리 사용되는 방법이다. 그러나 선형회귀분석에 의한 회귀추정량은 이상치(outlier)에 매우 민감하므로, 이를 극복하기 위한 로버스트(robust) 추정량에 대한 방안은 여러 가지가 제안되고 있다. 본 연구는 로버스트 추정량들 중에 2단계 방법의 기저집합사용에서 발생 가능한 공선성(collinearities)의 영향력을 평가하고자 한다.

다중선형회귀분석에서 자료가 공선성을 갖고 있으면, 하나의 예측변수가 다른 하나 이상의 변수와 또는 여러 변수들과 근사적인 종속관계(approximate linear dependence)를 갖고 있는 경우이다. 이러한 자료에서 회귀추정량의 사용은 Montgomery와 Peck(1982)이 지적하는 바와 같이 원시자료(source data)의 조절에 의하여 공선성 효과를 감소시켜야 한다. 그러나, 매우 로버스트한 방법으로 알려진

* 제주교육대학교 컴퓨터교육과 교수

2단계방법은 이상치를 배제시킨 기본집합을 얻기 위하여, 기본적으로 1단계에서 기저집합(basis set)을 사용하고 있다. Atkinson(1986), Fung(1993), Woodruff와 Rocke(1994), Hadi(1992), Hadi와 Simonoff(1993), 김종우(1997)는 기저집합의 선택이 회귀추정량에 미치는 영향이 매우 크다는 연구물을 보이고 있다. 최적의 기저집합 사용만으로 로버스트 회귀추정량을 구하는 방법을 제시한 Rousseeuw(1984)는 $p+1$ (여기서, p는 rank) 크기의 기저집합을 얻기 위하여 nCp회의 반복을 사용하는 알고리즘을 제시하고 있다.

이러한 결과는 2단계방법의 사용에서 기저집합의 영향은 매우 크게 받게되는 것을 뜻하며, 기저집합은 분석하고자 하는 자료의 부분집합으로서 당연히 주어진 자료에 영향을 받게 된다. 그러므로 주어진 자료가 공선성을 띄게 되면 기저집합에 미치는 공선성의 영향력은 더욱 커지므로 이에 따른 이상치의 식별은 어려워진다.

2. 공선성이 있는 자료

선형회귀모형(linear regression model)을 다음과 같이 설정하자.

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon},$$

여기서 $\boldsymbol{Y} = (y_1, y_2, \ldots, y_n)^T$는 반응변수인 $(n \times 1)$ 벡터이다.

$\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^T$는 $\boldsymbol{x}_i^T = (x_{i1}, x_{i2}, \ldots, x_{ip})$를 행으로 갖는 설명변수인 $(n \times p)$ 행렬이다 (단, $p < n$).

$\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$는 모수인 $(p \times 1)$ 벡터이다.

$\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I})$인 $(n \times 1)$ 벡터이다.

이때, 선형회귀모형에 따른 $\boldsymbol{\beta}$의 불편추정량은

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \qquad (1)$$

이다. 따라서 $\boldsymbol{Y}$는 $N(\boldsymbol{X\beta}, \boldsymbol{I}\sigma^2)$의 분포를 따른다.

잔차 $\boldsymbol{e}$는

$$\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

$$= (I - H) Y$$

$$= (I - H)(X\beta + \varepsilon)$$

$$= (I - H) \varepsilon$$

이다. 여기서 $H = X(X^T X)^{-1} X^T$로서 「해트(hat)」 행렬이다.

고, 잔차 e의 평균과 분산은 다음과 같다.

$$E[e] = E[(I - H) \varepsilon] \tag{2}$$

$$= (I - H) E[\varepsilon]$$

$$= 0$$

$$Var[e] = Var[(I - H) \varepsilon] \tag{3}$$

$$= (I - H) Var[\varepsilon] (I - H)$$

$$= (I - H) \sigma^2$$

따라서, 잔차 e는 $N(0, (I - H) \sigma^2)$의 분포를 따르므로,

$$e_i / \sigma \sqrt{1 - h_{ii}} \sim N(0, 1)$$

이다. σ^2의 최소제곱추정량 s^2은

$$s^2 = e^T e / (n - p)$$

이다.

회귀 추정을 위한 식 (1),(2),(3)의 사용은 주어진 행렬에 공선성이 존재할 때, $(X^T X)^{-1}$는 존재할 수 없으므로 회귀추정은 불가능하게 된다. 따라서, 어떤 자료가 공선성이 있는지에 대한 검사가 필요하며, 일반적으로 상관계수, 고유치, 고유벡터, VIF, 스펙트랄 등을 조사하여 주어진 변수를 점검한다. 이러한 검사 방법을 Mason 과 Gunst(1985)는 변수간의 공선성은 자료 행렬이

$$X = [X_i, X_i^*]$$

여기서 X_i는 X의 i번째 열이고,

X_i^*는 X에서 X_i를 제외한 나머지행렬

이라면, X_i를 종속변수로 하고 나머지 독립변수에 대하여 회귀방정식을 적합시킬 때 얻어지는 결정계수 R_i^2를 구하면,

$$R_i^2 = \frac{X_i' X_i^* (X_i^{*'} X_i^*)^{-1} X_i^{*'} X_i}{X_i^{*'} X_i - (\sum_j x_{ij})^2 / n} \tag{4}$$

이다. 여기서 X_i의 특징값이 크게 증가하면 R_i^2는 1로 근접하게 되므로 i번째 예측변수는 다른 변수와 공선성관계에 놓이게 된다. 이러한 방법으로 또다른 j 번째 변수도 공선성을 유발시킬 수 있고, i번째 변수와 j 번째 변수 사이에도 공선성이 발생하여 $X_i - X_j \dashrightarrow 0$이 되도록 할 수 있음을 보이고 있다. 이러한 공선성은 세력점(leverage point)에 의하여 공선성이 유발될 수 있음을 보이고 있는 것으로, 이들을 인식하기 위한 개선 방법으로 Stewart(1987)는 근사공선성(near collinearity)을 찾기 위하여 공선성 지표(collinearity index)의 사용을 제안했으며, Hadi와 Velleman(1987)은 이 지표를 사용하여 공선성과 영향력 관측점의 관계를 조사했다.

3. 기저집합에서 공선성의 문제

분석하고자 하는 자료에서 공선성의 유무를 사실상 분석하기가 쉽지 않을뿐더러 이상치를 식별하려는 2단계방법의 기본집합 사용은 시작 행렬을 기저집합으로 할 때, 이러한 위험에 노출을 더욱 극대화시킬 수 있다. 왜냐하면 다중이상치 식별을 위하여 2단계방법은 시작행렬을 이상치가 없으리라 예상되는 기저집합을 최적의 기본집합으로 구성하고 자료 집합을 기본집합과 나머지집합으로 분리하여, 기본집합을 사용하여 나머지집합에서 이상치를 식별하는 방법을 사용하고 있기 때문이다. 이러한 문제점은 기저행렬을 사용하는 엄준근외 2인(1995)의 ELMS(extended least median of squares), Hadi와 Simonoff(1993), Rousseeuw(1984)의 LMS(least median of squares) 등에서 모두 나타나게 된다.

2단계방법에서 초기로 설정하는 기본집합은 rank + 1인 기저집합(basis set)으로 가법적 이상치의 포함을 억제하기 위하여 로버스트한 추정을 사용하며, 염준근외 2인(1995)에서 제시된 ELMS를 응용한 방법은 김종우(1997)에서 밝혀진바 같이 일반적인 모의실험에서 다른 방법에 비하여 효과적임을 보이고 있으므로 ELMS 알고리즘을 살펴보면,

ELMS 알고리즘

<u>1 단계</u>. 기저집합 J의 초기 설정. 단, J의 크기는 $k(=p+1)$이다.

OLS를 사용한 잔차에서 절대 표준화잔차 $|r_i|$를 구하고 이를 오름차순으로 정렬하여 작은 크기 순으로 k개를 선택한다.

$$|r_{i_1}|_{1:n} \le |r_{i_2}|_{2:n} \le , \ldots , \le |r_{i_n}|_{n:n}$$

여기서 $r_{i_j} = e_i/\sqrt{1-h_{ii}}$ 단, $i_j \in \{\, i \mid i = 1, 2, \ldots, n \,\}$. (5)

$$e_i = y_i - \boldsymbol{x}_i \hat{\boldsymbol{\beta}} ,$$

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y},$$

$$h_{ii} = \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i .$$

일 때, 기저집합의 초기 설정은 $J = \{\, i_1, i_2, \ldots, i_k \,\}$ 이다.

<u>2 단계</u>. 기저집합의 선택.

기저집합 J에서 $(k-1)$개의 원을 취하고 나머지집합 J^c에서 1개의 원을 취하여 새로운 기저집합 $J_{(i)j}$를 구성한다. 이들 중에서 각 기저집합들 중에 최소잔차중위수를 갖는 기저집합을 선택한다.

$$\underset{\hat{\boldsymbol{\beta}}_{J_{(i)j}}}{\text{Minimize}} \ \underset{J_{(i)j}}{\text{med}}\, e_i^2$$

여기서 $J_{(i)j} = J_{(i)} \bigcup \{ j \}$.

$$J_{(i)} = J - \{ i \}, \quad i \in J \ (= \{ i_1, i_2, \ldots, i_k \}),$$

$$j \in J^c \ (= \{ i_{k+1}, i_{k+2}, \ldots, i_{n-k} \}),$$

$$e_i = y_i - \boldsymbol{x_i}\ \hat{\boldsymbol{\beta}}_{J(i)}\ ,$$

$$\hat{\boldsymbol{\beta}}_{J(i)} = (\ \boldsymbol{X}^T_{J(i)}\boldsymbol{X}_{J(i)})^{-1}\boldsymbol{X}^T_{J(i)}\ \boldsymbol{Y}_{J(i)}\ ,$$

이다. 공선성을 갖고 있는 기저집합은 식 (5)에서 이상치 판별을 위한 s^2을 사용한 i번째 관측치의 표준화잔차(studentized residual) r_i는

$$r_i = e_i / s\sqrt{1 - h_{ii}}$$

여기서 h_{ii}는 해트행렬 $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$의 대각선상의 원

이고, 공선성의 성질에 따라 해트행렬의 구성요소인 $\boldsymbol{X}^T\boldsymbol{X}$는 singular 행렬이 되므로 역행렬이 존재 할 수 없으므로 이 알고리즘은 실행이 중지될 것이다. 따라서, 이를 극복하기 위한 근사 역행렬의 사용이 제시되고 있다.(김종우,1997)

4. 모의실험 및 실증분석

1) 모의실험의 설계 및 절차

여러 가지 2단계방법들 중에 이상치에 매우 로버스트한 김종우(1997)의 ELMS 방법을 사용하여 기저집합을 구하고, 단순 선형회귀 모형과 다중 선형회귀 모형으로 나누어 실시한다. 모의실험을 위한 난수 발생과 프로그램의 실행은 SAS/IML을 사용하고, 공선성 발생의 정도를 det 함수와 식(4)를 사용한 VIF를 사용한다.

기저집합을 사용하는 이상치 식별 방법에서의 문제점은 모집합에는 공선성이 있으나 기저집합에 공선성이 없는 경우와 모집합에는 공선성이 없으나 기저집합에 공선성이 있는 경우로 나누어 볼 수 있다.

(1) 모집합에는 공선성이 없으나 기저집합에 공선성이 발생하는 경우

이상치를 포함하고 있는 선형모형에서 공선성을 전제하지 않은 상태에서 singular 행렬의 발생 비율을 측정한다.

case 1. 단순선형모형 $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$에서 $i = 1, 2, \ldots, 30$이고 $\beta_0 = 0$, $\beta_1 = 1$일 때 $x_i \sim U(0, 15)$, $\varepsilon_i \sim N(0, 1)$에서 정상적인 관측치를 30개 구하고, 이

상치 발생을 위하여 $y_i = x_i + 4$, $x_i = k - .05(i-1)$, $k =$ 7.5(low leverage), 20(high leverage), $i = 1, 2, \ldots, l$(단, l은 이상치 수)를 사용하여 1000회의 모의실험으로 singular 행렬의 발생 비율을 측정한다.

case 2. 다중선형모형 $y_i = \beta_o + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_5 x_{5i} + \varepsilon_i$에서 $i =$ 1,2, ... ,30이고 $\beta_o = 0$, $\beta_1 = \beta_2 = \beta_3 = 1$일 때 $x_{1i} \sim U(0, 15)$, $x_{2i} \sim U(0, 15)$, ... , $x_{5i} \sim U(0, 15)$, $\varepsilon_i \sim N(0, 1)$에서 정상적인 관측치를 30개 구하고, 공선성 발생을 위하여 $y_i = x_i + 4$, $x_i = k - .05(i-1)$, $k =$ 7.5(low leverage), 20(high leverage), $i = 1, 2, \ldots, l$(단, l은 이상치 수)를 사용하여 1000회의 모의실험으로 singular 행렬의 발생 비율을 측정한다.

〈표 1〉 공선성이 없는 자료에서 모의실험

rank	outliers	회귀추정	기저집합	기본집합
1	0	0	.03	0
	0.33	0	.06	0
3	0	0	0	0
	0.33	0	.286	0
5	0	0	0	0
	0.33	0	.916	0

(주) 1. 회귀추정은 단순선형회귀추정량에 의한 공선성 발생률
2. 기저집합은 rank+1 크기에서 공선성 발생률
3. 기본집합은 $\left[\frac{n}{2}\right]$([ ]는 가우스기호) 크기에서 공선성 발생률

(2) 모집합에 공선성이 있을 때, 기저집합에 공선성이 발생하는 경우

이상치를 포함하고 있는 선형모형에서 공선성을 유발시킨 상태에서 singular 행렬의 발생 비율을 측정한다.

case 1. 다중선형모형 $y_i = \beta_o + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$에서 $i =$ 1,2,...,30이고 $\beta_o = 0$, $\beta_1 = \beta_2 = 1$일 때 $x_{1i} \sim U(0, 15)$, $x_{2i} \sim U(0, 15)$, $\varepsilon_i \sim N(0, 1)$에서 정상적인

관측치를 30개 구하고, 공선성 발생을 위하여 $y_i = x_i + 4$, $x_i = k - .05(i-1)$, $k =$ 7.5(low leverage), 20(high leverage)과 이상치의 위치를 $i = 1, 2, \ldots, l$(단, l은 이상치 수)를 사용하여 1000회의 모의실험으로 singular 행렬의 발생 비율을 측정한다.

〈표 2〉 공선성이 있는 자료에서 모의실험

rank	outliers	회귀추정	기저집합	기본집합
2	0	0	.133	0
	0.33	0	.097	.001

(주) 1. 회귀추정은 단순선형회귀추정량에 의한 공선성 발생률
2. 기저집합은 rank+1 크기에서 공선성 발생률
3. 기본집합은 $\left[\frac{n}{2}\right]$([ ]는 가우스기호) 크기에서 공선성 발생률

2) 실증분석

case 1. Belsley(1984)에 의해 소개된 근사공선성을 집김하는 자료로서 회귀추정으로는 공선성이 알려지지 않으며, 단지 근사공선성에 의하여 공선성이 인식되는 자료이다. 기저집합의 사용에서 공선성은 발생되며, 2단계방법에서 기본집합의 크기가 $\left[\frac{n}{2}\right]$에서도 공선성이 발생하고 있음을 알 수 있다.

case 2. Daniel과 Wood(1980)는 비선형최소제곱을 설명하기 위한 자료로서 회귀추정에 의하여는 공선성이 알려지지 않으며, 단지 근사공선성에 의하여 공선성이 인식되는 자료이다. 기저집합의 사용에서 공선성은 발생되며, 2단계방법에서 기본집합의 크기가 $\left[\frac{n}{2}\right]$에서도 공선성이 발생하고 있음을 알 수 있다.

5. 결 론

이상치를 파악하는 것은 주어진 자료를 분석하는데 결정적인 의미를 지니고 있다. 그러나 주어진 자료의 공선성 문제는 회귀추정 자체를 불가능하게 하거나 또는 왜곡시킬 수 있다. 이상치에 매우 로버스트한 것으로 알려진 2단계방법에서 기

지집합과 기본집합의 사용은 모의실험에서 나타난바와 같이 이상치가 많이 존재할 경우에 다중회귀분석에서 크게 문제시 될 수 있음을 보이고 있다. 또한 단순회귀추정은 공선성 인식에 매우 둔함을 보이고 있다. 따라서 초기 기본집합을 구하고 직접접근방법을 사용하여 기본집합의 크기를 늘려 가면서 이상치를 식별하는 외향성 검정방법은 공선성의 문제점을 줄일 수 있음을 보이고 있다. 그러나, 실증자료에서 제시한 봐와같이 근사공선성의 문제에는 2단계방법의 효용성은 떨어짐을 보이고 있으므로 이에 대한 보완으로 근사공성성의 점검은 2단계방법에 앞서 필요하다.

❖ 참 고 문 헌 ❖

[1] 김종우(1997), "다수 이상치 식별을 위한 2단계방법에 관한 연구" 박사학위논문, 동국대학교.

[2] 임순근,박종구,김종우(1995), "다변량 자료에서 다수 이상치 인식의 절차", 품질경영학회지, 제23권, 제4호, pp. 28-41.

[3] Atkinson, A. C.(1986), "Masking Unmasked," *Biometrika*, Vol.73, No.3, pp. 533-541.

[4] Belsley,D.A.(1984), "Demeaning conditioning diagnostics through centering", *American Statistics*, Vol.38, pp. 73-93.

[5] Daniel,C. and Wood F.S.(1980). Fitting Equations to Data, 2nd ed. Wiley, New York.

[6] Hadi, A.(1992), "Identifying Multiple Outliers in Multivariate Data," *Journal of the Royal Statistical Society, Series-B*, Vol.54, No.3, pp.761-771.

[7] Hadi, A. and Simonoff, J. S.(1993), "Procedures for the Identifying of Multiple Outliers in Linear Models," *Journal of the American Statistical Association*, Vol.88, No.424, pp.1264-1272.

[8] Mason,R.L. and Gunst,R.F.(1985),"Outlier-Induced Collinearities," *Technometrics*, Vol.27,No.4,pp.401-407.

[9] Montgomery,D.C.,and Peck,E.A.(1982), Introduction to Linear Regression Analysis, NewYork:John Wiley.

[10] Rousseeuw, P. J.(1984), "Least Median of Squares Regression," *Journal of the American Statistical Association,* Vol.79, No.388, pp.871-880.

[11] Woodruff, D. L., and Rocke, D. M. (1994), "Computable Robust Estimation of Multivariate Location and Shape in High Dimension Using Compound Estimation,"*Journal of the American Statistical Association,* Vol.89, No 427, pp.888-896.