Communications of the Korean Statistical Society Vol. 15, No. 3, 2008, pp. 333–342

공분산분석 모형에서의 변수선택 정리†

윤상후[1], 박정수[2]

요 약

회귀모형에서의 변수선택에 관한 정리를 공분산분석 모형으로 확장하였다. 공분산분석 모형에서 몇개의 회귀변수를 제거한 축소모형을 세우는 경우에 추정량의 변화를 알아본 결과, 회귀계수 뿐만아니라 분산분석계수도 추정량의 편차는 증가하지만 분산은 감소하며, 어떤 경우에는 평균제곱오차도 감소한다는 결론을 얻었다.

주요용어: 과소적합; 반양정치 행렬; 일반화 역행렬; 축소모형; 추정가능한 함수; 평균제곱오차; 회귀계수.

1. 회귀모형에서의 변수선택 정리

회귀분석 모형에서 여러개의 설명변수들로부터 적절한 변수들을 선택하여 모형을 적합시키는 것에 대한 이론적 근거를 알기 위해 먼저 다음과 같은 두가지 모형, 즉 완전모형(full model)과 축소모형(reduced model)을 고려해 보자.

$$[\text{완전모형}] \quad \mathbf{y} = \mathbf{X}_p\beta_p + \mathbf{X}_r\beta_r + \varepsilon, \tag{1.1}$$

$$[\text{축소모형}] \quad \mathbf{y} = \mathbf{X}_p\beta_p + \varepsilon, \tag{1.2}$$

여기서 ε은 평균 0, 분산 $\sigma^2\mathbf{I}$ 라고 가정한다. 이제 $\hat{\beta}_p$와 $\hat{\beta}_r$를 완전모형에서의 β_p와 β_r의 최소제곱추정량, $\tilde{\beta}_p$를 축소모형에서의 β_p의 최소제곱추정량이라고 하자. 완전모형이 참모형임에도 불구하고 축소모형을 적합시킨다면 (즉 모형을 잘못 수립해서 과소적합 했을 때) 이들 추정량에 대해 다음과 같은 현상이 일어남이 알려져있다 (Hocking, 1976).

정리 1.1

1. $\tilde{\beta}_p$는 일반적으로 불편추정량이 아니다. 단 예외적으로 $\beta_r = 0$이거나 $X'_pX_r = 0$ 이면 $\tilde{\beta}_p$는 불편추정량이다.

2. 행렬 $\mathrm{Var}(\hat{\beta}_p) - \mathrm{Var}(\tilde{\beta}_p)$는 양반정치(positive semi-definite) 행렬이다.

† 본 연구는 한국학술진흥재단의 2005년 지역대학우수과학자 연구비 지원에 의해 수행되었음 (KRF-2005-202-C00072).

1) (500-767) 광주광역시 북구 용봉동 300, 전남대학교 통계학과, 박사과정.
2) (500-767) 광주광역시 북구 용봉동 300, 전남대학교 통계학과, 교수. 교신저자: jspark@jnu.ac.kr

3. 행렬 $\mathrm{Var}(\hat{\beta}_r) - \beta_r \beta'_r$이 양반정치 행렬이면, $\mathrm{MSE}(\hat{\beta}_p) - \mathrm{MSE}(\tilde{\beta}_p)$도 양반정치 행렬이다.

4. $\mathbf{x}$의 어떤 특정한 값 $\mathbf{x_0}$에 대해서 $\mathrm{Var}(\mathbf{x}_0'\hat{\beta}) - \mathrm{Var}(\mathbf{x}_{0p}'\tilde{\beta}_p) \geq 0$이다.

5. 축소모형에서의 σ^2의 추정량 $\tilde{\sigma}^2$은 상향편의(biased upward) 되었다.

이 정리의 의미는 축소모형을 적합시키면 회귀계수의 추정치의 편차(bias) 는 증가하지만 분산은 감소하며, 어떤 경우에는 평균제곱오차(mean squared error)도 감소한다는 것이다. 이는 변수제거의 이론적 근거가 되는 중요한 성질로서 $\mathrm{Var}(\hat{\beta}_p) - \beta_r \beta'_r$이 양반정치인 경우에는 완전모형보다 축소모형이 평균제곱오차의 기준에서 더 바람직하다는 것이다. 정리 1.1은 여러 교재에 재수록 되어 있다 (예: 박성현, 1980; Rencher, 2000; Ravishanker과 Dey, 2001; Seber과 Lee, 2003).

본 논문에서는 정리 1.1을 공분산분석(analysis of covariance) 모형으로 확장하였다. 즉, 공분산분석 모형에서 몇개의 회귀변수를 제거한 축소모형을 세우는 경우에 추정량의 변화를 알아본 결과, 회귀계수 뿐만아니라 분산분석계수도 추정량의 편차는 증가하지만 분산은 감소하며, 어떤 경우에는 평균제곱오차도 감소한다는 결론을 얻었다.

변수선택과 관련된 국내 논문으로는 김진흠과 김민호 (2004), 윤영주와 송문섭 (2005), 홍종선 등 (2005), Choi (2006), 박종선 (2007) 등이 있고, 국외의 최근 흥미로운 연구는 George (2000), Kadane와 Lazar (2004), Yuan과 Lin (2005), Gurka (2006), Claeskens 등 (2006) 이 있다.

2. 공분산분석 모형과 추정량의 성질

공분산분석(analysis of covaraince: ANCOVA)이란 실험의 정확도를 높이기 위해 처리에 의해 설명되지 않는 부분을 공변량을 이용하여 추가로 설명함으로서 실험의 정확도를 높이는 분석 방법이다. 이 모형은 분산분석모형과 회귀모형이 결합된 형태로 분산분석모형은 일반적으로 변수가 완전 순위(full rank)가 아닌 반면 회귀모형의 변수는 완전 순위이다.

이제 $k\,(= w + q)$개의 설명변수에 대하여 n개의 데이터가 있다고 할 때 공분산분석 모형을 행렬을 사용하여 표현하면,

$$\mathbf{y} = \mathbf{X}\tau + \mathbf{Z}\beta + \varepsilon \tag{2.1}$$

이다. 여기서 $\mathbf{X}$는 완전 순위가 아닌(less than full rank) $n \times w$ 행렬이고, τ는 $(w \times 1)$ 벡터, $\mathbf{Z}$는 완전 순위인 $n \times q$ 행렬이고 $(q = p + r + 1)$, β는 $(q \times 1)$ 벡터, $\mathbf{y}$는 $\mathbf{y} = (Y_1, Y_2, \ldots, Y_n)'$인 $(n \times 1)$ 벡터이고, 분산분석의 변수 $\mathbf{X}$의 각 열은 회귀모형의 변수 $\mathbf{Z}$의 각 열과 선형 독립이다. 또한 오차항 ε은 평균이 0이고 공분산 행렬이 $\sigma^2 \mathbf{I}_n$이다.

이 모형에서 회귀계수 β와 분산분석의 계수 τ의 최소제곱추정량은,

$$\hat{\beta} = [\mathbf{Z}'\mathbf{Q}\mathbf{Z}]^{-1}\mathbf{Z}'\mathbf{Q}\mathbf{y}, \tag{2.2}$$

$$\hat{\tau} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'(\mathbf{y} - \mathbf{Z}\hat{\beta}), \tag{2.3}$$

이다 (Ravishanker과 Dey, 2001). 여기서 $\mathbf{Q}$는

$$\mathbf{Q} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}' \tag{2.4}$$

인 대칭멱등 행렬(symmetric idempotent matrix)이다. 또 $(\mathbf{X}'\mathbf{X})^{-}$는 $\mathbf{X}'\mathbf{X}$의 일반화 역행렬(generalized inverse matrix)이며 $\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X} = \mathbf{X}'\mathbf{X}$를 만족한다.

추정량의 특성을 알아보면, 먼저 $\hat{\beta}$의 기대값은

$$E(\hat{\beta}) = [\mathbf{Z}'\mathbf{Q}\mathbf{Z}]^{-1}\mathbf{Z}'\mathbf{Q}(\mathbf{X}\tau + \mathbf{Z}\hat{\beta}) \ = \beta \tag{2.5}$$

이 된다. 위의 계산에서 $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X} = \mathbf{X}$라는 일반화 역행렬의 성질에 의해서 $\mathbf{Q}\mathbf{X}\tau = 0$를 이용하였다. 또한 $\hat{\tau}$의 기대값은

$$E(\hat{\tau}) = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}\tau$$

이다. 여기서 $(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}$ 부분이 제거되지 않아 그 만큼의 편의가 생길 뿐 아니라 유일하지도 않다. 그런데 τ의 어떠한 추정가능한 함수(estimable function) $\mathbf{c}'\tau$에 대해서, 그 기대값을 구하면, $\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X} = \mathbf{c}'$임을 이용하여,

$$E(\mathbf{c}'\hat{\tau}) = \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}\tau = \mathbf{c}'\tau \tag{2.6}$$

이 된다. 한편 $\hat{\beta}$의 분산은

$$\begin{aligned} \mathrm{Var}(\hat{\beta}) &= [\mathbf{Z}'\mathbf{Q}\mathbf{Z}]^{-1}\mathbf{Z}'\mathbf{Q}\mathbf{Q}'\mathbf{Z}[\mathbf{Z}'\mathbf{Q}\mathbf{Z}]^{-1}\sigma^2 \\ &= [\mathbf{Z}'\mathbf{Q}\mathbf{Z}]^{-1}\sigma^2 \end{aligned} \tag{2.7}$$

이다. 그리고 $\mathbf{c}'\hat{\tau}$의 분산은,

$$\begin{aligned} \mathrm{Var}(\mathbf{c}'\hat{\tau}) &= \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}' \ \mathrm{Var}(\mathbf{y} - \mathbf{Z}\hat{\beta}) \ \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{c} \\ &= \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}' \ \mathrm{Var}[(\mathbf{I} - \mathbf{W}\mathbf{Q})\mathbf{y}] \ \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{c} \\ &= \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}' \ [\mathbf{I} + \mathbf{W}] \ \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{c}\sigma^2, \end{aligned} \tag{2.8}$$

여기서 대칭행렬 $\mathbf{W}$는

$$\mathbf{W} = \mathbf{Z}(\mathbf{Z}'\mathbf{Q}\mathbf{Z})^{-1}\mathbf{Z}' \tag{2.9}$$

이다. 위의 계산에서 $\mathbf{W}\mathbf{Q}\mathbf{Q}'\mathbf{W}' = \mathbf{W}$와 $\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{W}\mathbf{Q}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{c} = 0$라는 사실이 이용되었다.

3. 축소된 공분산분석 모형에서의 추정량

행렬 $\mathbf{Z}$를 $\mathbf{Z} = (\mathbf{Z}_p, \mathbf{Z}_r)$ 형태로 분할하여 공분산분석 모형 (2.1)을 다시 표현하면

$$[\text{완전모형}] \ \ \mathbf{y} = \mathbf{X}\tau + \mathbf{Z}_p\beta_p + \mathbf{Z}_r\beta_r + \varepsilon. \tag{3.1}$$

이 된다. β_p와 β_r의 최소제곱추정량을 각각 $\hat{\beta}_p$와 $\hat{\beta}_r$로 표현하자. 이제 모델 (3.1)에서 $\mathbf{Z}_r$을 제거한 축소모형을 적어보면 다음과 같다.

$$[\text{축소모형}] \quad \mathbf{y} = \mathbf{X}\tau + \mathbf{Z}_p\beta_p + \varepsilon. \tag{3.2}$$

이 축소모형에서 얻은 β_p와 τ의 최소제곱추정량을 각각 $\tilde{\beta}_p$와 $\tilde{\tau}$라고 표시하면,

$$\begin{aligned} \tilde{\beta}_p &= (\mathbf{Z'}_p\mathbf{Q}\mathbf{Z}_p)^{-1}\mathbf{Z'}_p\mathbf{Q}\mathbf{y}, & (3.3)\\ \tilde{\tau} &= (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'(\mathbf{y} - \mathbf{Z_p}\tilde{\beta}_p) & (3.4) \end{aligned}$$

이다.

이제 완전모형 (3.1)이 참임에도 불구하고 축소모형 (3.2)를 적합시켰을 때, $\tilde{\beta}_p$와 $\tilde{\tau}$의 특성을 알아보자.

먼저 $\tilde{\beta}_p$의 기대값을 계산하면 (이때 기대치는 완전모형 (3.1)에 대해서 취해짐을 주의한다), $\mathbf{Q}\mathbf{X}\tau = 0$임을 이용하여,

$$\begin{aligned} E(\tilde{\beta}_p) &= (\mathbf{Z'}_p\mathbf{Q}\mathbf{Z}_p)^{-1}\mathbf{Z'}_p\mathbf{Q}(\mathbf{X}\tau + \mathbf{Z}\beta)\\ &= (\mathbf{Z'}_p\mathbf{Q}\mathbf{Z}_p)^{-1}\mathbf{Z'}_p\mathbf{Q}(\mathbf{Z}_p\beta_p + \mathbf{Z}_r\beta_r)\\ &= \beta_p + \mathbf{D}\beta_r, \qquad (3.5) \end{aligned}$$

여기서

$$\mathbf{D} = (\mathbf{Z'}_p\mathbf{Q}\mathbf{Z}_p)^{-1}\mathbf{Z'}_p\mathbf{Q}\mathbf{Z}_r \tag{3.6}$$

이다. 즉 $\mathbf{D}\beta_r$ 만큼의 편의가 있음을 알 수 있다. 그리고 $\tilde{\beta}_p$의 분산은

$$\begin{aligned} \mathrm{Var}(\tilde{\beta}_p) &= (\mathbf{Z'}_p\mathbf{Q}\mathbf{Z}_p)^{-1}\mathbf{Z'}_p\mathbf{Q}\mathbf{Q}'\mathbf{Z_p}(\mathbf{Z'}_p\mathbf{Q}\mathbf{Z}_p)^{-1}\sigma^2\\ &= (\mathbf{Z'}_p\mathbf{Q}\mathbf{Z}_p)^{-1}\sigma^2 \qquad (3.7) \end{aligned}$$

이고, 평균제곱오차는

$$\mathrm{MSE}(\tilde{\beta}_p) = (\mathbf{Z'}_p\mathbf{Q}\mathbf{Z}_p)^{-1}\sigma^2 + \mathbf{D}\beta_r\beta_r^{'}\mathbf{D}' \tag{3.8}$$

이다.

다음으로 축소모형을 택함(model misspecification)으로인한 분산분석계수의 추정량에서의 변화를 확인해보자. 먼저 τ의 어떠한 추정가능한 함수 $\mathbf{c}'\tau$의 추정량 $\mathbf{c}'\tilde{\tau}$의 기대값을 계산하면, $E(\mathbf{y}) = \mathbf{X}\tau + \mathbf{Z}_p\beta_p + \mathbf{Z}_r\beta_r$을 이용하여,

$$\begin{aligned} E(\mathbf{c}'\tilde{\tau}) &= \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'E(\mathbf{y}) - \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Z_p}\,E(\tilde{\beta}_p)\\ &= \mathbf{c}'\tau - \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'(\mathbf{Z}_p\mathbf{D} - \mathbf{Z}_r)\beta_r \qquad (3.9) \end{aligned}$$

이다. 또한 $\mathbf{c}'\tilde{\tau}$의 분산은 (2.8)에서와 비슷한 방법으로

$$\begin{aligned} \mathrm{Var}(\mathbf{c}'\tilde{\tau}) &= \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\,\mathrm{Var}(\mathbf{y} - \mathbf{Z_p}\tilde{\beta}_p\,)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{c}\\ &= \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'(\mathbf{I} + \mathbf{W}_p)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{c}\sigma^2 \qquad (3.10) \end{aligned}$$

이 된다. 여기서, 대칭행렬 $\mathbf{W_p}$는

$$\mathbf{W_p} = \mathbf{Z_p}(\mathbf{Z'}_p\mathbf{Q}\mathbf{Z_p})^{-1}\mathbf{Z'}_p \tag{3.11}$$

이다. $\mathbf{c}'\tilde{\tau}$의 평균제곱오차는 (3.9)와 (3.10)로부터

$$\begin{aligned}\mathrm{MSE}(\mathbf{c}'\tilde{\tau}) &= \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'(\mathbf{I}+\mathbf{W}_p)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{c}\sigma^2 \\ &\quad +\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'(\mathbf{Z_p}\mathbf{D}-\mathbf{Z_r})\beta_r\beta_r'(\mathbf{Z_p}\mathbf{D}-\mathbf{Z_r})'\,\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{c} \quad (3.12)\end{aligned}$$

이다.

4. 변수선택 정리의 공분산분석 모형으로 확장

앞 절들에서 얻은 결과를 이용하여 정리 1.1을 공분산분석 모형으로 확장시킨 다음과 같은 정리를 얻을 수 있다.

정리 4.1

1. 행렬 $\mathrm{Var}(\hat{\beta}_p) - \mathrm{Var}(\tilde{\beta}_p)$는 양반정치 행렬이다.
2. $\mathrm{Var}(\mathbf{c}'\hat{\tau}) - \mathrm{Var}(\mathbf{c}'\tilde{\tau}) \geq 0$이다.
3. 만약 행렬 $\mathrm{Var}(\hat{\beta}_r) - \beta_r\beta_r'$가 양반정치 행렬이면, (i) $\mathrm{MSE}(\hat{\beta}_p) - \mathrm{MSE}(\tilde{\beta}_p)$도 양반정치 행렬이고, (ii) $\mathrm{MSE}(c'\hat{\tau}) - \mathrm{MSE}(c'\tilde{\tau}) \geq 0$이다.
4. $\tilde{\sigma}^2$은 상향 편의되었다.

이 정리의 증명을 위해서 미리 몇가지 사실을 알아두자. 먼저 분할행렬의 역행렬 공식을 이용하여

$$\begin{aligned}\mathrm{Var}(\hat{\beta}) &= [\mathbf{Z}'\mathbf{Q}\mathbf{Z}]^{-1}\sigma^2 \\ &= \begin{bmatrix}\mathbf{Z}'_p\mathbf{Q}\mathbf{Z}_p & \mathbf{Z}'_p\mathbf{Q}\mathbf{Z}_r \\ \mathbf{Z}'_r\mathbf{Q}\mathbf{Z}_p & \mathbf{Z}'_r\mathbf{Q}\mathbf{Z}_r\end{bmatrix}^{-1}, \qquad \sigma^2 = \begin{bmatrix}\mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22}\end{bmatrix}\sigma^2. \quad (4.1)\end{aligned}$$

여기서, (3.6)의 $\mathbf{D}$를 이용하여 표현하면,

$$\begin{aligned}\mathbf{A}_{11} &= (\mathbf{Z}'_p\mathbf{Q}\mathbf{Z}_p)^{-1} + \mathbf{D}\mathbf{A}_{22}\mathbf{D}', \quad (4.2)\\ \mathbf{A}_{12} &= -\mathbf{D}\mathbf{A}_{22}, \\ \mathbf{A}_{21} &= -\mathbf{A}_{22}\mathbf{D}', \quad (4.3)\\ \mathbf{A}_{22} &= (\mathbf{Z}'_r\mathbf{Q}\mathbf{Z}_r - \mathbf{D}'\mathbf{Z}'_p\mathbf{Q}\mathbf{Z}_r)^{-1}.\end{aligned}$$

이제 (2.9)의 $\mathbf{W} = \mathbf{Z}(\mathbf{Z}'\mathbf{Q}\mathbf{Z})^{-1}\mathbf{Z}'$를 역행렬 공식 (4.1)와 (4.2)를 이용해 다시 쓰면

$$\mathbf{W} = \mathbf{Z_p}(\mathbf{Z'}_p\mathbf{Q}\mathbf{Z_p})^{-1}\mathbf{Z'}_p + \mathbf{Z_p}\mathbf{D}\mathbf{A}_{22}\mathbf{D}'\mathbf{Z'}_p + \mathbf{Z_r}\mathbf{A}_{21}\mathbf{Z'}_p + \mathbf{Z_p}\mathbf{A}_{12}\mathbf{Z'}_r + \mathbf{Z_r}\mathbf{A}_{22}\mathbf{Z'}_r \quad (4.4)$$

이다. $\mathbf{W}-\mathbf{W}_p$를 (4.3)와 (4.4)을 이용해 표현하면

$$\begin{aligned}\mathbf{W}-\mathbf{W}_p &= \mathbf{Z_p}\mathbf{D}\mathbf{A}_{22}\mathbf{D}'\mathbf{Z}'_p - 2\mathbf{Z_r}\mathbf{A}_{22}\mathbf{D}'\mathbf{Z}'_p + \mathbf{Z}_r\mathbf{A}_{22}\mathbf{Z}'_r \\ &= (\mathbf{Z}_p\mathbf{D}-\mathbf{Z}_r)\mathbf{A}_{22}(\mathbf{Z}_p\mathbf{D}-\mathbf{Z}_r)' \qquad (4.5)\end{aligned}$$

이다. 그런데 $\mathbf{A}_{22}$는 $\mathrm{Var}(\hat{\beta}_r)/\sigma^2$이므로 양반정치 행렬이여서 $\mathbf{W}-\mathbf{W}_p$도 양반정치 행렬이다.

증명: (정리 4.1)

1. (4.1)와 (4.2)로부터

$$\mathrm{Var}(\hat{\beta}_p) = \mathbf{A}_{11}\sigma^2 = (\mathbf{Z}'_p\mathbf{Q}\mathbf{Z}_p)^{-1}\sigma^2 + \mathbf{D}\mathbf{A}_{22}\mathbf{D}'\sigma^2 \qquad (4.6)$$

이고, (3.7)으로부터 $\mathrm{Var}(\tilde{\beta}_p) = (\mathbf{Z}'_p\mathbf{Q}\mathbf{Z}_p)^{-1}\sigma^2$이므로,

$$\mathrm{Var}(\hat{\beta}_p) - \mathrm{Var}(\tilde{\beta}_p) = \mathbf{D}A_{22}\mathbf{D}'\sigma^2$$

이다. 그런데 $\mathbf{A}_{22}$가 양반정치 행렬이므로 $\mathrm{Var}(\hat{\beta}_p) - \mathrm{Var}(\tilde{\beta}_p)$도 양반정치 행렬이다.

2. (2.8)과 (3.10)로부터

$$\mathrm{Var}(\mathbf{c}'\hat{\tau}) - \mathrm{Var}(\mathbf{c}'\tilde{\tau}) = \mathbf{c}'(\mathbf{X}'\mathbf{X})^-\mathbf{X}'(\mathbf{W}-\mathbf{W}_p)\mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{c}\sigma^2$$

이다. $\mathbf{W}-\mathbf{W}_p$가 양반정치 행렬임은 (4.5)에서 보여졌으므로 $\mathrm{Var}(\mathbf{c}'\hat{\tau})-\mathrm{Var}(\mathbf{c}'\tilde{\tau}) \geq 0$이다.

3. (i) (2.5)과 (4.6)으로부터,

$$\mathrm{MSE}(\hat{\beta}_p) = \mathrm{Var}(\hat{\beta}_p) = (\mathbf{Z}'_p\mathbf{Q}\mathbf{Z}_p)^{-1}\sigma^2 + \mathbf{D}A_{22}\mathbf{D}'\sigma^2$$

이므로 (3.8)를 빼면,

$$\mathrm{MSE}(\hat{\beta}_p) - \mathrm{MSE}(\tilde{\beta}_p) = \mathbf{D}(A_{22}\sigma^2 - \beta_r\beta_r')\mathbf{D}' \qquad (4.7)$$

이다. $\mathrm{Var}(\hat{\beta}_r) = A_{22}\sigma^2$이므로, $\mathrm{Var}(\hat{\beta}_r) - \beta_r\beta_r'$가 양반정치 행렬이면 $\mathrm{MSE}(\hat{\beta}_p)$ $-\mathrm{MSE}(\tilde{\beta}_p)$도 양반정치 행렬이다.

(ii) (2.8)과 (3.12), (4.5)으로부터

$$\begin{aligned}&\mathrm{MSE}(\mathbf{c}'\hat{\tau}) - \mathrm{MSE}(\mathbf{c}'\tilde{\tau}) \\ &\quad= \mathbf{c}'(\mathbf{X}'\mathbf{X})^-\mathbf{X}' \\ &\qquad\times[(\mathbf{W}-\mathbf{W}_p)\sigma^2 - (\mathbf{Z_p}\mathbf{D}-\mathbf{Z_r})\beta_r\beta_r'(\mathbf{Z_p}\mathbf{D}-\mathbf{Z_r})']\mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{c} \\ &\quad= \mathbf{c}'(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\,(\mathbf{Z_p}\mathbf{D}-\mathbf{Z_r})(\mathbf{A}_{22}\sigma^2 - \beta_r\beta_r')(\mathbf{Z_p}\mathbf{D}-\mathbf{Z_r})'\,\mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{c}\end{aligned}$$

이다. 따라서 $\mathrm{Var}(\hat{\beta}_r) - \beta_r\beta_r'$이 양반정치 행렬이면 $\mathrm{MSE}(c'\hat{\tau}) - \mathrm{MSE}(c'\tilde{\tau}) \geq 0$ 이다.

4. 축소모형에서 σ^2의 추정량은

$$\tilde{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{X}\tilde{\tau} - \mathbf{Z}_p\tilde{\beta}_p)'(\mathbf{y} - \mathbf{X}\tilde{\tau} - \mathbf{Z}_p\tilde{\beta}_p)}{n - w - p - 1} \tag{4.8}$$

이다. $\mathbf{B} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}' + \mathbf{Q}\mathbf{W}_p\mathbf{Q}$에 대하여,

$$\mathbf{y} - \mathbf{X}\tilde{\tau} - \mathbf{Z}_p\tilde{\beta}_p = (\mathbf{I} - \mathbf{B})\mathbf{y}$$

이므로, (4.8)의 분모는 SSE $= \mathbf{y}'(\mathbf{I} - \mathbf{B})'(\mathbf{I} - \mathbf{B})\mathbf{y}$이다. $E(\mathbf{y}) = \mathbf{X}\tau + \mathbf{Z}\beta$이고 $\mathrm{Var}(\mathbf{y}) = \sigma^2 I$을 이용하여,

$$E(\mathrm{SSE}) = tr[(\mathbf{I} - \mathbf{B})'(\mathbf{I} - \mathbf{B})]\sigma^2 + (\mathbf{X}\tau + \mathbf{Z}\beta)'(\mathbf{I} - \mathbf{B})'(\mathbf{I} - \mathbf{B})(\mathbf{X}\tau + \mathbf{Z}\beta)$$

이다. $\mathbf{I} - \mathbf{B} = \mathbf{Q}(\mathbf{I} - \mathbf{W}_p\mathbf{Q})$이고, $\mathbf{Q}$가 대칭멱등 행렬임을 이용하여,

$$\begin{aligned} tr[(\mathbf{I} - \mathbf{B})'(\mathbf{I} - \mathbf{B})] &= tr(\mathbf{Q}) - 2\,tr(\mathbf{W}_p\mathbf{Q}) + tr(\mathbf{W}_p\mathbf{Q}\mathbf{W}_p\mathbf{Q}) \\ &= (n - w) - 2(p + 1) + p + 1 \\ &= n - w - p - 1 \end{aligned}$$

이다. 따라서

$$E(\tilde{\sigma}^2) = \sigma^2 + \frac{(\mathbf{X}\tau + \mathbf{Z}\beta)'(\mathbf{I} - \mathbf{B})'(\mathbf{I} - \mathbf{B})(\mathbf{X}\tau + \mathbf{Z}\beta)}{n - w - p - 1} \tag{4.9}$$

이다. 그런데 (4.9)의 두번째 항의 분자는 항상 0 이상이므로 $\tilde{\sigma}^2$은 상향 편의(biased upward) 되었다.

□

한편 $\mathbf{Q}\mathbf{X}\tau = 0$와 $(\mathbf{I} - \mathbf{W}_p\mathbf{Q})$이 멱등행렬임을 이용하여 (4.9)의 두번째 항의 분자를 다시 표현하면

$$\beta'\mathbf{Z}'\mathbf{Q}(\mathbf{I} - \mathbf{W}_p\mathbf{Q})\mathbf{Z}\beta \tag{4.10}$$

이 된다. 여기서 행렬 $\mathbf{Z}$ 를 $(\mathbf{Z}_p, \mathbf{Z}_r)$ 형태로 분할하여 정리하면 (4.10)은

$$\beta_r'\mathbf{Z}_r'\mathbf{Q}(\mathbf{I} - \mathbf{W}_p\mathbf{Q})\mathbf{Z}_r\beta_r \quad \text{또는} \quad \beta_r'\mathbf{Z}_r'(\mathbf{I} - \mathbf{B})\mathbf{Z}_r\beta_r$$

이 된다. 그래서 (4.9)는

$$E(\tilde{\sigma}^2) = \sigma^2 + \frac{\beta_r'\mathbf{Z}_r'(\mathbf{I} - \mathbf{B})\mathbf{Z}_r\beta_r}{n - w - p - 1} \tag{4.11}$$

으로 쓰여진다.

5. 결론

회귀모형에서 축소모형을 적합시키면 회귀계수의 추정치의 편차는 증가하지만 분산은 감소한다는 정리가 알려져 있다. 이는 변수제거의 이론적 근거가 되는 중요한 성질로서 어떤 조건이 만족되는 경우에는 완전모형보다 축소모형이 평균제곱오차의 기준에서 더 바람직하다는 것이다. 본 논문에서는 이 정리를 공분산분석(analysis of covariance: ANCOVA) 모형으로 확장하였다.

공분산분석이란 실험의 정확도를 높이기 위해 처리에 의해 설명되지 않는 부분을 공변량을 이용하여 추가로 설명함으로서 실험의 정확도를 높이는 분석 방법이다. 이 모형은 분산분석모형과 회귀모형이 결합된 형태로 분산분석모형은 일반적으로 변수가 완전 순위(full rank)가 아닌 반면 회귀모형의 변수는 완전 순위이다.

공분산분석 모형에서 몇개의 회귀변수를 제거한 축소모형을 세우는 경우에 추정량의 변화를 알아 보았다. 회귀계수 뿐만아니라 분산분석계수도 추정량의 편차는 증가하지만 분산은 감소하며, 어떤 경우에는 평균제곱오차도 감소한다는 결론을 얻었다. 따라서 편차와 분산 간의 균형을 이루는, 또는 평균제곱오차를 줄이는 적절한 모형을 찾아야 할 것이다.

참고문헌

김진흠, 김민호 (2004). 변수선택 편향이 없는 회귀나무를 만들기 위한 알고리즘, <응용통계연구>, **17**, 459–473.

박성현 (1980). <회귀분석>, 민영사, 서울.

박종선 (2007). Variable selection in sliced inverse regression using generalized eigenvalue problem with penalties, <한국통계학회논문집>, **14**, 215–227.

윤영주, 송문섭 (2005). Variable selection via penalized regression, <한국통계학회논문집>, **12**, 615–624.

홍종선, 함주형, 김호일 (2005). 수정 결정계수를 사용한 로지스틱 회귀모형에서의 변수선택법, <응용통계연구>, **18**, 435–443.

Choi, S. H. (2006). Interval regression models using variable selection, <한국통계학회논문집>, **13**, 125–134.

Claeskens, G., Croux, C. and Van Kerckhoven, J. (2006). Variable selection for logistic regression using a prediction-focused information criterion, *Biometrics*, **62**, 972–979.

George, E. I. (2000). The variable selection problem, *Journal of the American Statistical Association*, **95**, 1304–1308.

Gurka, M. J. (2006). Selecting the best linear mixed model under REML, *The American Statistician*, **60**, 19–26.

Hocking, R. R. (1976). The Biometrics invited paper. The analysis and selection of variables in linear regression, *Biometrics*, **32**, 1–49.

Kadane, J. B. and Lazar, N. A. (2004). Methods and criteria for model selection, *Journal of the American Statistical Association*, **99**, 279–290.

Ravishanker, N. and Dey, D. K. (2001). *A First Course in Linear Model Theory*, Chapman & Hall/CRC, New York.

Rencher, A. C. (2000). *Linear Models in Statistics*, John Wiley & Sons, New York.

Seber, G. A. F. and Lee, A. J. (2003). *Linear Regression Analysis*, 2^{nd} edition. John Wiley & Sons, New York.

Yuan, M. and Lin, Y. (2005). Efficient empirical Bayes variable selection and estimation in linear models, *Journal of the American Statistical Association*, **100**, 1215–1225.

[2008년 2월 접수, 2008년 3월 채택]

Variable Selection Theorem for the Analysis of Covariance Model†

Sang-Hoo Yoon[1], Jeong-Soo Park[2]

Abstract

Variable selection theorem in the linear regression model is extended to the analysis of covariance model. When some of regression variables are omitted from the model, it reduces the variance of the estimators but introduces bias. Thus an appropriate balance between a biased model and one with large variances is recommended.

Keywords: Estimable function; generalized inverse; mean squared error; positive semi-definite matrix; reduced model.

† This paper was supported by The Korea Research Foundation, 2005 (KRF-2005-202-C00072).

1) Ph.D. student, Department of Statistics, Chonnam National University, 300 Yongbong-dong, Gwangju 500-767.

2) Professor, Department of Statistics, Chonnam National University, 300 Yongbong-dong, Gwangju 500-767. Correspondence: jspark@jnu.ac.kr