응용통계연구 제 17권 1호, 2004년, pp. 145-151

PROC MIXED가 제시하는 분산의 합의 신뢰구간의 문제점

박동준[1)]

요 약

SAS의 PROC MIXED procedure는 다양한 형태의 혼합모형에 적합한 자료를 분석하고, 그 자료들이 채집된 모집단의 모수들에 관한 통계적 추론을 하는데 사용된다. 그러나 혼합모형에 해당되는 불균형중첩오차구조를 갖는 선형회귀모형안에 나타나는 두 개의 분산의 합에 대한 신뢰구간을 구할 때 PROC MIXED의 REML 추정량으로부터 계산되는 신뢰구간은 신뢰계수를 지키지 못한다는 것을 시뮬레이션을 통하여 보인다.

주요용어: PROC MIXED, 신뢰구간, 혼합모형, 분산성분

1. 서 론

혼합모형은 고정효과와 랜덤효과를 함께 포함하는 선형모형으로서, 일반적인 선형모형에서 고정효과와 함께 나타나는 모수들은 설명변수들과 관련된 모수를 의미하고, 랜덤효과와 함께 나타나는 모수들은 자료들 사이의 변동을 설명하는 분산 또는 공분산들을 의미한다. PROC MIXED는 이러한 혼합모형에서 나타나는 고정효과와 랜덤효과와 관련된 모수들의 추론에 사용될 수 있고, 표준적인 선형모형을 다루는 PROC GLM의 대안으로 사용될 수 있다. 식(2.2)의 불균형중첩오차구조를 갖는 단순회귀모형은 두 개의 분산을 갖는 랜덤효과항들과 회귀계수에 해당되는 고정효과를 포함하는 혼합모형에 해당하므로 식(2.2)의 두 분산의 합의 구간추정을 위하여 PROC MIXED를 이용할 수 있다. 이 소고에서는 불균형중첩오차구조를 갖는 단순회귀모형의 랜덤효과항에 나타나는 두 개의 분산의 합에 대한 신뢰구간을 구할 때, PROC MIXED에서 계산되는 신뢰구간의 문제점을 시뮬레이션을 통하여 지적하고자 한다.

2절에서는 PROC MIXED를 사용할 수 있는 일반적인 혼합모형의 구조를 설명하고, 특히 PROC MIXED를 적용할 구체적인 불균형중첩오차구조를 갖는 단순회귀모형을 상술하였다. 3절에서는 그 모형의 분산들의 추정량들과 그 추정량들의 분산들과 공분산을 구하기 위하여 PROC MIXED문에서 구체적으로 사용해야 하는 option들을 적고 설명하였다. 그리고 분산의 합에 대한 신뢰구간을 구하기 위하여 PROC MIXED로부터 계산된 분산의 합의 추정량과 그 추정량들의 점근추정오차를 고려한 정규근사식을 적었다. 4절에서는 시뮬레이션의 방법과 PROC MIXED를 이용한 분산의 합의 신뢰구간이, 특히, 소표본에서는 명시한(stated) 신뢰계수를 잘 지키지 못하는 시뮬레이션의 결과를 보였다. 5절에서 시뮬레이션의 결과에 대한 결론을 맺으면서 이러한 PROC MIXED의 난점을 극복하기 위한 하나의 대안으로서 Ting et al.(1990) 방법과 Khuri et al.(1998)의 일반화 p 값을 이용한 방법을 제안하였다.

1) (608-737) 부산광역시 남구 대연 3동 599-1, 부경대학교 자연과학대학 수리과학부, 부교수
E-mail: djpark@pknu.ac.kr

2. PROC MIXED에서 사용하는 혼합모형의 구조

PROC MIXED에서 사용하는 혼합모형은 표준의 선형모형을 일반화한 것으로서 다음과 같이 적는다.

$$\underline{Y} = X\underline{\beta} + Z\underline{\gamma} + \underline{\epsilon}. \tag{2.1}$$

여기서, $\underline{Y}$는 관찰자료들의 벡타, $\underline{\beta}$는 고정효과로서 설계행렬 X와 관련된 미지의 모수들의 벡타, $\underline{\gamma}$는 설계행렬 Z와 관련된 랜덤효과들의 벡타, $\underline{\epsilon}$은 오차항들을 포함하는 벡타이다.

이 소고에서 사용한 불균형중첩오차구조를 갖는 단순회귀모형은 PROC MIXED에서 사용하는 식(2.1)의 혼합모형에 속하고, 구체적으로 표현하면 식(2.2)와 같이 적는다.

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + A_i + E_{ij}, \quad i = 1, \cdots, I; j = 1, \cdots, J_i. \tag{2.2}$$

여기서, Y_{ij}는 i번째 주샘플링단위의 j번째 관찰값이고, β_0와 β_1은 미지의 상수, X_{ij}는 고정된 예측변수, A_i와 E_{ij}는 서로 독립이고, 평균이 0 이고, 각각의 분산이 σ_A^2와 σ_E^2인 정규확률변수이며, $I > 2$, $J_i \geq 1$, 적어도 하나의 i에 대해서는 $J_i > 1$ 이어야 한다. 식(2.2)를 식(2.1)의 행렬의 형태로 적으면 $\underline{Y}$는 크기가 $J_. \times 1$인 벡타, 여기서 $J_. = \sum_{i=1}^{I} J_i$, X는 $J_. \times 2$인 행렬, $\underline{\beta}$는 β_0와 β_1을 원소로 갖는 2×1 인 벡타, Z는 $J_. \times I$인 행렬, $\underline{\gamma}$는 A_i를 원소로 갖는 $I \times 1$인 벡타, $\underline{\epsilon}$은 E_{ij}를 원소로 갖는 $J_. \times 1$ 인 벡타가 된다. 식(2.2)의 A_i와 E_{ij}는 각각 주샘플링 단위와 부샘플링 단위와 관련된 오차항이다. 이러한 오차구조들은 서로 관련(correlated)되므로 오차의 구조는 식(2.3)과 같다.

$$Cov(Y_{ij}, Y_{i'j'}) = \begin{cases} \sigma_A^2 + \sigma_E^2, & \text{if } i = i', j = j'; \\ \sigma_A^2, & \text{if } i = i', j \neq j'; \\ 0, & \text{if } i \neq i'. \end{cases} \tag{2.3}$$

3. PROC MIXED의 활용

일반선형모형에 대한 분산(Variance Component)의 선형결합이나 분산의 비율에 대한 신뢰구간에 관한 구체적인 내용은 Burdick과 Graybill(1992)을 참고할 수 있다. 식(2.2)와 같이 J_i가 서로 다른 불균형중첩오차구조를 갖는 단순회귀모형에서 반응변수 Y_{ij}의 총분산(Total Variance)에 대한 통계적 추론을 할 경우, 반응변수에 대한 총분산을 두 분산의 합 $\gamma = \sigma_A^2 + \sigma_E^2$이라고 하자. 이때 γ에 대한 신뢰구간은 잔차최대우도(residual(or restricted) maximum likelihood)를 이용하여 분산들의 추정량과 그들의 분산을 구하는 PROC MIXED를 사용하여 신뢰구간을 손쉽게 구할 수 있다.

그림 4.1은 PROC MIXED를 사용하여 γ에 대한 신뢰구간을 구하기 위한 SAS/IML로 프로그래밍한 매크로프로그램의 일부분이다. 첫째 줄의 PROC MIXED의 문장 이전에는 식(2.2)의 가정에 따라 SAS의 정규확률변수들의 난수발생함수인 RANNOR를 이용하여 평균이 0 이고, 분산이 각각 σ_A^2과 σ_E^2인 정규확률변수 A_i와 E_{ij} 들을 생성한다. 즉, ρ를 $\rho = \sigma_A^2/(\sigma_A^2 + \sigma_E^2)$라고 할 때 일반성을 잃지 않고, $\sigma_A^2 = 1 - \sigma_E^2$ 라고 적을 수 있으므로

$\rho = \sigma_A^2$와 $1-\rho = \sigma_E^2$가 된다. 따라서 정규확률변수들인 $A_i \sim N(0,\rho)$와 $E_{ij} \sim N(0,1-\rho)$는 RANNOR를 이용하여 생성된다. 유사한 방법을 이용하여 난수로 생성된 X_{ij}값들과 임의의 상수인 β_0와 β_1을 식(2.2)에 대입하면 반응변수인 Y_{ij} 들이 발생된다. 그리고 이렇게 생성된 모든 확률변수들을 포함한 DATA SET 이름을 EXAMPLE로 놓은 다음, Y_{ij} 값들을 그림 4.1의 셋째 줄의 Y에, X_{ij} 값들을 셋째 줄의 X에, A_i값들을 둘째 줄과 넷째 줄의 A에 대입하여 사용한다.

그림 4.1의 첫째 줄의 PROC MIXED문 다음의 DATA=EXAMPLE을 적음으로써 생성된 자료들을 불러서 사용할 수 있다. 이어서 ASYCOV를 사용하면 관찰된 Fisher information 행렬의 역행렬인 σ_A^2와 σ_E^2의 추정량에 대한 분산 및 공분산값이 포함된 점근공분산행렬이 구해진다. 그리고 γ의 신뢰구간을 구하는데 직접적으로 필요하지 않은 분류에 관한 내용(Class Level Information)과 반복 계산된 내력(Iteration History)과 모형의 정보(Model Information)들은 NOCLPRINT NOITPRINT NOINFO를 적음으로써 모두 출력화면에서 제외시킨다. 둘째 줄의 CLASS문에는 분류변수(Classification Variable)를 사용하여야 하므로 식(2.2)에서 주샘플링단위에 해당되는 A를 적는다. 셋째 줄의 MODEL문에는 "반응변수=고정효과의 변수"를 적어야 하므로 $Y = X$를 적는다. 넷째 줄의 RANDOM문에는 랜덤효과에 해당되는 변수를 적어야 하므로 A를 적는다. SAS의 Version 7과 8에서는 PROC MIXED를 실행하였을 때 자동적(Default)으로 나타나는 값들을 출력화면에 보이지 않도록 하기 위하여, 그림 4.1의 다섯째 줄에 ODS(Output Delivery System)문에 EXCLUDE를 사용함으로써 분산들의 추정값(COVPARMS)과 모형에 대한 적합통계량값(FITSTATISTICS)과 고정효과 X의 TYPE 3 검정(TESTS3)과 분산들의 추정량에 대한 분산들과 공분산의 점근추정값(ASYCOV)들을 배제시킨다. 그 대신 γ에 대한 정규근사 계산 때 쓰기 위하여 분산들의 추정값과 분산추정량들의 분산과 공분산의 점근추정값들을 골라내기 위한 방법으로, 여섯째 줄과 일곱째 줄에 각각 COVEST 와 COVMATRIX 라는 이름을 주어서 ODS OUTPUT COVPARMS = COVEST와 ODS OUTPUT ASYCOV = COVMATRIX 로 적는다. 그림 4.1의 마지막 줄에 다음에는 매크로프로그램을 1000회 시뮬레이션 실행하여 생성

```
.
.
PROC MIXED DATA=EXAMPLE ASYCOV NOCLPRINT NOITPRINT NOINFO;
CLASS A;
MODEL Y = X;
RANDOM A;
ODS LISTING EXCLUDE COVPARMS FITSTATISTICS TESTS3 ASYCOV;
ODS OUTPUT COVPARMS = COVEST;
ODS OUTPUT ASYCOV = COVMATRIX;
.
.
```

그림 4.1: γ의 신뢰구간을 구하기 위하여 매크로 내부에서 사용된 PROC MIXED의 예제

된 분산들의 추정값들과 그들의 점근추정값들을 식(3.1)에 대입하여 γ에 대한 $100(1-\alpha)\%$ 신뢰구간을 구한다.

$$[\ \hat{\gamma} - Z_{\alpha/2} \cdot SE(\hat{\gamma})\ :\ \hat{\gamma} + Z_{\alpha/2} \cdot SE(\hat{\gamma})\ ] \tag{3.1}$$

여기서 $\hat{\gamma} = \hat{\sigma}_A^2 + \hat{\sigma}_E^2$, $SE(\hat{\gamma}) = \sqrt{V(\hat{\sigma}_A^2) + V(\hat{\sigma}_E^2) + 2Cov(\hat{\sigma}_A^2, \hat{\sigma}_E^2)}$, $Z_{\alpha/2}$는 표준정규분포의 제 $100(1-\alpha/2)$ 백분위수값이다.

식(3.1)의 정규근사식 대신 Burdick과 Graybill(1992)의 3장에서 γ에 대한 신뢰구간을 구하기 위하여 Satterthwaite방법, Welch방법, Graybill과 Wang방법등을 제안하고 있지만, 이들 방법 모두 σ_A^2와 σ_E^2의 불편추정량들이 서로 독립이고 각각 정확한 카이제곱분포를 요구한다. 그러나 식(2.2)의 σ_E^2의 불편추정량은 정확한 카이제곱분포를 하지만, σ_A^2의 불편추정량은 근사적인 카이제곱분포의 형태를 취하므로 γ에 대한 신뢰구간을 구하기 위해서는 다른 근사식을 사용하는 것보다 식(3.1)의 정규근사를 사용하는 것이 PROC MIXED를 간편하게 활용하는 방법이 된다. PROC MIXED에 관하여 보다 자세한 내용은 SAS의 Online Doc(Version 8)에서 찾을 수 있다.

4. 시뮬레이션의 실행

유도한 신뢰구간들을 판정하는 기준은 크게 두 가지로 구분할 수 있다: 1)첫째, 시뮬레이터된 신뢰구간들은 명시된(stated) 신뢰계수$(1-\alpha)$를 유지해야 한다, 2)둘째, 양쪽 신뢰구간의 평균길이는 짧을수록 바람직하다. 짧은 신뢰구간이 바람직하지만 유도한 신뢰구간은 우선적으로 신뢰계수를 유지해야 한다. 불균형중첩오차구조를 갖는 단순회귀모형에 대해서 표 4.1과 같이 $I=3$과 $I=10$과 $I=30$의 세 가지 패턴에 대해서 시뮬레이션을 시행하였다. 3절에서 설명한 그림 4.1이 포함된 SAS/IML로 프로그래밍한 매크로를 실행시킬 때, ρ의 값을 0.001, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.999 까지 변화시키되 각각의 ρ값에 대해서 1000번씩 시뮬레이션을 실행하였다. 각 1000번의 시뮬레이션을 실행한 다음, γ를 포함하는 신뢰구간들의 개수를 1000으로 나누어 신뢰계수를 계산하였다.

표 4.1: 시뮬레이션에 사용된 I와 J_i의 값

패턴	I	J_i	$J_.$
1	3	3 5 10	18
2	10	1 1 1 5 5 5 5 10 10 10	53
3	30	1 1 1 3 3 3 3 5 5 5 5 5 6 6 6 6 6 6 7 7 7 7 7 8 8 8 8 10 10 10	173

그리고 신뢰구간의 평균길이는 계산된 신뢰구간들의 상한에서 하한을 감한 모든 신뢰구간들의 길이의 평균값이 된다. 이항분포에 대한 정규근사를 사용하면 참신뢰계수가 0.9

일 때 1000번의 시뮬레이션 실행에서 추정된 신뢰계수가 0.8814 보다 작을 기회는 2.5% 보다 작다. 표 4.2는 γ에 대한 90% 신뢰구간을 구하기 위하여 ρ값에 따라 1000회씩 시뮬레이션을 실행하였을 때 계산된 신뢰계수와 신뢰구간의 평균길이를 나타낸다.

표 4.2: PROC MIXED로 시뮬레이터된 90% 신뢰구간의 신뢰계수와 평균길이

	시뮬레이터된 신뢰계수			신뢰구간의 평균길이		
ρ	패턴 1	패턴 2	패턴 3	패턴 1	패턴 2	패턴 3
0.001	0.8850	0.9030	0.9000	1.3214671	0.6844607	0.3576012
0.1	0.8560*	0.9090	0.8950	1.4394805	0.7191657	0.3694501
0.2	0.8360*	0.8880	0.8920	1.5333086	0.7908918	0.4002304
0.3	0.8150*	0.8600*	0.8980	1.6984080	0.8778323	0.4415527
0.4	0.7890*	0.8390*	0.8930	1.7840318	0.9494793	0.4918554
0.5	0.7470*	0.8470*	0.8910	1.9412385	1.0544330	0.5541932
0.6	0.7540*	0.8410*	0.8840	2.1626564	1.1593301	0.6157282
0.7	0.6690*	0.8300*	0.8680*	2.1516335	1.2522920	0.6744327
0.8	0.7000*	0.8290*	0.8666*	2.3911152	1.3417022	0.7383816
0.9	0.6890*	0.8240*	0.8770*	2.4889468	1.4546989	0.7991264
0.999	0.6790*	0.8270*	0.8730*	2.6804925	1.5236820	0.8580632
MAX	0.8850	0.9090	0.9000	2.6804925	1.5236820	0.8580632
MIN	0.6790*	0.8240*	0.8666*	1.3214671	0.6844607	0.3576012

5. 결 론

표 4.2에서 신뢰계수가 0.8814 보다 작은 값은 *로 표시하였다. I가 3 에서 30 으로 증가함에 따라 0.9에 가까운 신뢰계수는 1개에서 7개로 늘어났다. 그러나 표 4.2의 시뮬레이터된 신뢰계수들은 패턴 1 또는 2 와 같이 소표본인 경우 대부분의 ρ값에 대해서 0.8814 보다 작게 나타난다. 패턴 3과 같은 대표본의 경우에도 0.8814 보다 작게 나타나는 경우는 많이 줄어들었지만, $\rho \geq 0.7$인 경우 시뮬레이터된 신뢰계수가 0.8814 이하로 떨어진다. 그러므로 PROC MIXED는 $I = 30$인 경우에도 γ에 대한 바람직한 신뢰구간을 제시하지 못하는 것을 알 수 있다. 시뮬레이션의 결과를 종합하면 혼합모형에 해당되는 불균형중첩오차구조를 갖는 단순회귀모형의 두 분산의 합 γ에 대한 신뢰구간을 PROC MIXED를 이용하여 구한다면, $I = 3$, $I = 10$, $I = 30$ 일 때 각각 $\rho \geq 0.1$, $\rho \geq 0.3$, $\rho \geq 0.7$의 경우 신뢰계수를 지키지 못하는 잘못된 신뢰구간을 계산하게 된다. PROC MIXED 의 이러한 난점을 극복하기 위해서 γ에 대한 구간추정의 한 방법으로 박동준(2003)이 제시한 Ting et al.(1990) 방법 또는 Khuri et al.(1998)의 일반화 p 값(generalized p value)을 이용한 일반화축량(generalized pivotal quantity)방법을 이용하여 γ에 대한 신뢰구간을 구할 수 있다.

참고문헌

[1] Burdick, R. K. and Graybill, F. A.(1992), *Confidence Intervals on Variance Components*, Marcel Dekker, New York.

[2] Khuri, A. I., Mathew, T., and Sinha, B.(1998), *Statistical Tests for Mixed Linear Models*, John Wiley & Sons, New York.

[3] Park, D. J.(2003), Interval Estimation for Sum of Variance Components in A Simple Linear Regression Model with Unbalanced Nested Error Structure, *The Korean Communications in Statistics*, vol. 10, no. 2, 361-370.

[4] SAS Online Doc(version 8), http://v8doc.sas.com/sashtml.

[5] Ting, N., Burdick, R. K., Graybill, F. A., Jeyaratnam, S., and Lu, T.-F. C.(1990), Confidence Intervals on Linear Combinations of Variance Components, *Journal of Statistical Computation and Simulation*, vol. 35, 135-143.

[2003년 4월 접수, 2003년 9월 채택]

Misleading Confidence Interval for Sum of Variances Calculated by PROC MIXED of SAS

Dong Joon Park [1)]

ABSTRACT

PROC MIXED fits a variety of mixed models to data and enables one to use these fitted models to make statistical inferences about the data. However, the simulation study in this article shows that PROC MIXED using REML estimators provides one with a confidence interval, that does not keep the stated confidence coefficients, on sums of two variance components in the simple regression model with unbalanced nested error structure which is a mixed model.

Keywords: PROC MIXED, confidence interval, mixed model, variance component.

1) Associate Professor, Division of Mathematical Sciences, Nam-Gu, Daeyeon 3-Dong, 599-1 Pusan 608-737 Korea.

E-mail : djpark@pknu.ac.kr