응용통계연구 제 19권 2호, 2006년, pp. 349-360

PROC MIXED를 활용한 혼합모형의 신뢰구간추정*

박동준[1)]

요 약

SAS의 PROC MIXED를 사용하면 일반적인 ANOVA 추정량뿐만 아니라 더 많은 장점을 갖는 제한최대우도추정법 또는 최대우도추정법으로 모수들을 추론할 수 있다. 혼합모형에 속하는 불균형중첩오차구조를 갖는 선형회귀모형에서 랜덤효과와 관련된 그룹간 분산의 신뢰구간과 고정효과에 해당되는 회귀계수들에 대한 신뢰구간을 구하기 위하여 세 가지 크기를 갖는 표본에 대하여 PROC MIXED를 사용하였다. 모의실험을 실행한 결과, 대표본인 경우에는 모수들의 신뢰구간을 구하기 위하여 PROC MIXED를 활용할 수 있지만, 소표본인 경우에는 PROC MIXED를 사용할 경우, 그룹간 분산의 신뢰구간과 회귀계수 가운데 절편항의 신뢰구간은 주어진 신뢰계수를 지키지 못하는 것을 보인다.

주요용어: PROC MIXED, 제한최대우도추정법, 혼합모형, 신뢰구간

1. 서론

불균형자료를 포함하는 선형모형에 나타나는 분산을 추정하기 위한 시도는 Henderson(1953)으로부터 시작되었다. Henderson은 랜덤모형 또는 고정효과와 랜덤효과를 모두 갖는 혼합모형의 불균형자료들의 분산을 추정하기 위하여 Henderson의 Method I, II, III의 세 가지 방법을 사용하였는데 모두 ANOVA 방법을 활용한 것으로 Searle et al.(1992)에서 그 자세한 추정방법과 예제들을 설명하고 있다.

Milliken and Johnson(2002)은 고정효과와 랜덤효과를 갖춘 혼합모형을 설계구조와 처리효과의 구조의 발생 형태에 따라 세 가지로 분류한다(Section 13.1 introduction). 일반적으로 통계 연구자들은 이러한 혼합모형에 나타난 고정효과들과 랜덤효과들 또는 그들의 선형함수에 대한 추정이나 검정에 대하여 관심을 갖게 된다. 이러한 혼합모형의 자료구조가 균형인 경우에는 혼합모형에 대한 분산분석표 안에 고정효과와 랜덤효과들의 평균제곱들이 카이제곱분포(a scaled chi-squared distribution)의 형태로 표현이 가능하게 되어 그 효과들에 대한 모수들의 구간추정이나 검정이 비교적 쉬웠다(Burdick and Graybill 1992). 그러나 불균형자료구조를 갖는 혼합모형인 경우에는 관심의 대상이 되는 고정효과들이나 랜덤효과들에 대한 추정이나 검정에서 ANOVA방법을 사용하여 손으로 계산하는 것이 매우 어려

* 이 논문은 2005학년도 부경대학교 기성회 학술연구비에 의하여 연구되었음(과제번호: PK-2005-014)

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

웠다. 그러한 경우에는 관측된 표본자료들을 이용한 최대우도추정법(Maximum Likelihood Estimation)이나 제한최대우도추정법(Restricted Maximum Likelihood Estimation)을 근거로 한 통계패키지를 사용하여 그들의 추정을 하게 된다. 더구나 이러한 혼합모형에서 반응변수들의 공분산은 서로 상관된 자료구조를 갖게 되어 추정이 쉽지 않았으나 SAS의 PROC MIXED를 활용하여 쉽게 추정과 검정을 할 수 있게 되었다(Brown and Prescott, 1999; Verbeke and Molenberghs, 1997).

SAS의 PROC MIXED는 PROC ANOVA나 PROC GLM보다 더 일반화 된 PROCEDURE로서 세 PROCEDURE 모두 CLASS, MODEL, CONTRAST, ESTIMATE, LSMEANS 문을 사용하여 일반적인 선형모형에 대한 추정과 검정에 활용할 수 있지만 PROC MIXED에서는 RANDOM 문을 사용하여 랜덤효과와 고정효과가 있는 혼합모형에서 랜덤효과에 대한 추론을 할 수 있다. 특히, 랜덤효과에 대한 분산을 추정할 때 PROC GLM은 분산에 대한 ANOVA 추정량을 계산하지만, 제한최대우도추정법이나 최대우도추정법을 사용할 수 있는 PROC MIXED는 혼합모형의 랜덤효과의 분산의 추정에 있어서 ANOVA 추정량 보다 더 선호되는 제한최대우도추정량이나 최대우도추정량을 계산한다(Searle et al., 1992). 또한 반복측정이 있는 경우는 PROC MIXED에 REPEATED문을 사용하여 공분산구조를 살펴볼 수 있다. 이 소고에서는 불균형중첩오차구조를 갖는 중회귀모형의 랜덤효과와 관련된 분산들의 신뢰구간과 회귀계수에 대한 신뢰구간을 구할 때, 소표본의 경우 PROC MIXED에서 계산되는 신뢰구간의 문제점을 모의실험을 통하여 지적하고자 한다.

여러 가지 통계패키지가 가능하지만 가장 보편적으로 대학에서 쉽게 접근하여 사용할 수 있는 SAS의 PROCEDURE들 가운데 PROC MIXED를 사용하여 혼합모형 내에 있는 랜덤효과에 대한 분산들과 고정효과들의 구간추정을 활용하는 방법을 알아보고 그 문제점들을 파악하려고 한다(Littell et al., 1996; SAS Online Doc Version 9.1). 2절에서는 PROC MIXED를 적용할 구체적인 불균형중첩오차구조를 갖는 중회귀모형을 상술하였다. 3절에서는 그 모형의 모수들의 추정값들과 신뢰구간을 구하기 위하여 PROC MIXED문에서 구체적으로 사용해야 하는 option들을 적고 실제로 계산되는 과정을 설명하였다. 4절에서는 소표본과 대표본, 그리고 그 중간크기에 해당하는 표본에 대하여 모의실험의 절차를 설명한 후, 실행한 결과를 그래프로 제시하고 5절에서 결론을 맺는다.

2. 혼합모형의 행렬표현

혼합모형은 일반적인 선형모형을 일반화한 것으로서 행렬의 구조로 표현하면 다음과 같다.

$$\boldsymbol{y} = \boldsymbol{X}\underline{\beta} + \boldsymbol{B}\underline{\gamma} + \underline{\epsilon} \tag{2.1}$$

여기서, $\boldsymbol{y}$는 관측자료들의 벡터, $\underline{\beta}$는 설계행렬 $\boldsymbol{X}$와 관련된 고정효과 모수들의 벡터, $\underline{\gamma}$는 설계행렬 $\boldsymbol{B}$와 관련된 랜덤효과 모수들의 벡터, $\underline{\epsilon}$은 오차항들의 벡터이다. PROC MIXED를 사용하면 식 (2.1)에서 고정효과 모수들에 대한 구간추정과 랜덤효과들의 분산에 대한 구간추정을 할 수 있다. 예를 들어 식 (2.1)을 예측변수가 2개, g개의 군(cluster; 그룹), 각각

의 군에 n_i개의 불균형자료를 갖는 경우 일반적인 식으로 표현하면 다음과 같다.

$$Y_{ij} = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + A_i + E_{ij} \quad (2.2)$$
$$i = 1, \ldots, g; j = 1, \ldots, n_i$$

여기서, Y_{ij}는 i번째 그룹의 j번째 관측값이고, $\beta_0, \beta_1, \beta_2$는 회귀계수, X_{1ij}, X_{2ij}는 예측변수, A_i는 랜덤효과로서 i번째 그룹과 관련된 오차항이고, E_{ij}는 i번째 그룹 내의 j번째 관측값과 관련된 오차항으로서 A_i와 E_{ij}는 서로 독립이고, 평균이 0이고, 각각의 분산이 σ_A^2와 σ_E^2인 정규확률변수이며, $g > 2, n_i \geq 1$이고, 적어도 하나의 i에 대해서는 $n_i > 1$이고, $n = \sum_i n_i$이다. 식 (2.2)에서 β_0항부터 $\beta_2 X_{2ij}$까지는 고정효과들이고, A_i는 랜덤효과이므로 식 (2.2)는 혼합모형이 된다. 이 모형은 불균형중첩오차구조를 갖고 2개의 독립변수를 갖는 중회귀모형이라고 부른다. 이 모형을 식 (2.1)의 행렬의 형태로 적으면 $\boldsymbol{y}$는 크기가 $n \times 1$인 벡터, $\boldsymbol{X}$는 $n \times 3$인 행렬, $\underline{\beta}$는 회귀계수 $\beta_0, \beta_1, \beta_2$로 구성된 3×1인 벡터, $\boldsymbol{B}$는 $n \times g$인 행렬, $\underline{\gamma}$는 A_i를 원소로 갖는 $g \times 1$인 벡터로서 $\underline{\gamma} \sim N(\underline{0}, \sigma_A^2 \boldsymbol{I}_g)$의 분포를 하고 $\underline{\epsilon}$은 E_{ij}를 원소로 갖는 $n \times 1$인 벡터로서 $\underline{\epsilon} \sim N(\underline{0}, \sigma_E^2 \boldsymbol{I}_n)$의 분포를 하고 $\underline{\gamma}$와 $\underline{\epsilon}$은 서로 독립이다.

식 (2.2)에 대한 자료의 구조를 설명하기 위하여 Milliken and Johnson(2002, pp. 591-592)의 체중감량에 영향을 미치는 다이어트 방법의 연구에 관한 예제를 활용할 수 있다. 연구를 시작할 때 측정한 각 사람들의 최초 체중(X_{1ij})과 체지방의 양(X_{2ij})을 각각 예측변수라 하고 여러 다이어트 방법을 시작한 뒤 4 개월 후의 체중(Y_{ij})을 반응변수라고 하자. 그리고 랜덤효과 A_i를 여러 가지 다이어트 방법 가운데서 선택된 세 가지 방법의 효과로 가정하고, E_{ij}는 i번째 그룹 내의 j번째 오차항이라고 가정하면, A_i와 E_{ij}는 서로 독립이고, 평균이 0이고, 각각의 분산이 σ_A^2와 σ_E^2인 정규확률변수로 가정한다. 세 가지 다이어트 방법을 적용하기 위하여 각각 3명, 5명, 10명을 선발하였다면 $g = 3, n = 18$인 불균형중첩오차구조를 갖는 중회귀모형이 되고 표 2.1은 선택된 각 사람들의 자료를 포함한다.

표 2.1: 식 (2.2)를 적용할 수 있는 자료구조의 예시

관측값	다이어트 방법								
	$A(i=1)$			$B(i=2)$			$C(i=3)$		
	최초체중	최초체지방	4개월 후 체중	최초체중	최초체지방	4개월 후 체중	최초체중	최초체지방	4개월 후 체중
	X_{1ij}	X_{2ij}	Y_{ij}	X_{1ij}	X_{2ij}	Y_{ij}	X_{1ij}	X_{2ij}	Y_{ij}
1	175	27.1	174	172	27.8	162	165	25.5	165
2	185	32.4	173	157	26.8	162	168	29.5	165
3	172	30.5	172	158	26.8	156	159	24.4	154
4				161	30.2	166	171	30.7	162
5				178	31.1	160	174	30.1	167
6							152	23.0	151
7							173	28.2	164
8							173	29.2	162
9							216	31.5	201
10							235	31.7	226

3. 구간추정을 계산하기 위한 PROC MIXED의 적용

식 (2.2)와 같은 혼합모형에서 추정하고자 하는 모수로서는 고정효과에 속하는 회귀계수 $\beta_0, \beta_1, \beta_2$와 랜덤효과 A_i의 분산인 σ_A^2과 오차항 E_{ij}의 분산인 σ_E^2등이 있다. 특히, 분산 σ_E^2을 추정하기 위하여 Park 과 Burdick(2003)이 제안한 오차평균제곱 $S_E^2 = \boldsymbol{y}'[\boldsymbol{I}_n - \boldsymbol{X}^*(\boldsymbol{X}^{*\prime}\boldsymbol{X}^*)^+\boldsymbol{X}^{*\prime}]\boldsymbol{y}/(n-g-k)$을 사용할 수 있다. 여기서 $\boldsymbol{X}^* = [\boldsymbol{X}, \boldsymbol{BB}']$이고 k는 예측변수의 개수이며 $(n-g-k)S_E^2/\sigma_E^2$은 자유도 $n-g-k$를 갖는 카이제곱분포를 한다. 그리고 Park 과 Burdick(2003)은 분산 σ_A^2의 신뢰구간을 계산하기 위하여 통계량 $\boldsymbol{W} = \boldsymbol{FBB}'\boldsymbol{F}$와 벡터 $\boldsymbol{z} = \boldsymbol{Fy}$와 평균제곱 $S_M^2 = \boldsymbol{z}'\boldsymbol{W}^+\boldsymbol{z}/(g-1)$을 정의하였다. 여기서 $\boldsymbol{F} = \boldsymbol{X}^*(\boldsymbol{X}^{*\prime}\boldsymbol{X}^*)^+ - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^+\boldsymbol{X}'$이고 $+$는 Moore-Penrose inverse를 의미한다. 그리고 S_M^2의 기대평균제곱은 $E(S_M^2) = \sigma_A^2 + \sigma_E^2/h$로서 h는 통계량 $\boldsymbol{W}$에 대한 서로 다른 양의 값을 갖는 고유값 d_l의 조화평균이다.

3.1. σ_E^2의 구간추정

이제 σ_E^2의 구간추정을 하려고 한다. $(n-g-k)S_E^2/\sigma_E^2 \sim \chi^2_{n-g-k}$의 분포적 성질을 이용하면 σ_E^2에 대한 신뢰구간을 구할 수 있다. 즉, i번째 그룹 내의 j번째 관측값과 관련된 오차항 E_{ij}의 분산인 σ_E^2에 대한 정확한 $100(1-\alpha)\%$ 신뢰구간은 다음과 같이 구할 수 있다.

$$\left[\frac{(n-g-k)S_E^2}{\chi^2_{(n-g-k;\alpha/2)}}; \frac{(n-g-k)S_E^2}{\chi^2_{(n-g-k;1-\alpha/2)}}\right] \tag{3.1}$$

여기서 $\chi^2_{(n_1,n_2:\alpha)}$는 자유도 n_1과 n_2를 갖는 χ^2분포에서 오른쪽 끝 부분의 면적이 α인 χ^2값을 의미한다. 그러나 PROC MIXED에서는 식 (3.1)을 이용하여 σ_E^2의 신뢰구간을 계산하지 않고 Wald 통계량을 사용하여 σ_E^2의 신뢰구간을 구하는데 그 계산 방법은 3.2절에서 σ_A^2의 신뢰구간을 구할 때 같이 설명한다.

3.2. σ_A^2의 구간추정

랜덤효과인 A_i의 분산인 σ_A^2의 신뢰구간을 구하기 위하여 S_M^2은 $\sigma_E^2 = 0$일 때 $(g-1)S_M^2/\sigma_A^2 \sim \chi^2_{g-1}$인 성질과 S_M^2과 S_E^2이 서로 독립인 성질을 이용하여 Park and Burdick(2003)은 σ_A^2의 근사적인 $100(1-\alpha)\%$ 신뢰구간으로서 Ting et al.(1990) 방법과 Weerahandi(1993)가 제안한 일반화 신뢰구간 방법을 이용하여 두 가지 신뢰구간을 제안하였다.

그러나 PROC MIXED에서는 랜덤효과의 분산인 σ_A^2과 오차항 E_{ij}의 분산인 σ_E^2의 신뢰구간을 구하기 위하여 PROC MIXED 문장 다음에 CL 이란 옵션을 쓰면 제한최대우도추정방법으로 계산된 Wald 통계량을 사용하여 σ_A^2과 σ_E^2의 신뢰구간을 구할 수 있다. 표 3.1의 PROC MIXED문에 분산 σ_A^2과 σ_E^2의 불편추정량의 값과 그들의 점근분산값과 공분산의 값과 분산 σ_A^2과 σ_E^2의 90% 신뢰구간을 계산하기 위해 ASYCOV CL ALPHA=.10 과 default로 사용되는 제한최대우도추정법을 강조하기 위하여 METHOD=REML로 적는다.

표 3.1: REML를 이용하여 분산 $\widehat{\sigma_A^2}$과 $\widehat{\sigma_E^2}$을 계산하기 위한 PROC MIXED문의 예시

```
PROC MIXED DATA=EXAMPLE ASYCOV CL ALPHA=.10 METHOD=REML;
CLASSES COL4;
MODEL COL1 = COL2 COL3;
RANDOM COL4;
```

표 3.2: 표 3.1을 1회 실행시켰을 때 분산 $\widehat{\sigma_A^2}$과 $\widehat{\sigma_E^2}$의 추정값과 $\widehat{\sigma_A^2}$과 $\widehat{\sigma_E^2}$의 점근분산과 공분산의 값

The Mixed Procedure

Covariance Parameter Estimates

Cov Parm	Estimate	Alpha	Lower	Upper
COL4	0.2221	0.1	0.04152	2304415
Residual	1.8361	0.1	1.0791	3.9591

Asymptotic Covariance Matrix of Estimates

Row	Cov Parm	CovP1	CovP2
1	COL4	0.2878	-0.05766
2	Residual	-0.05766	0.4942

표 3.1의 실행한 결과, 두 분산의 추정값은 COL4 와 Residual에 각각 나타나는데 $\widehat{\sigma_A^2} = 0.2221, \widehat{\sigma_E^2} = 1.8361$로 계산되었다. 여기서 σ_A^2의 90% 신뢰구간에 대한 계산은 Wald 통계량을 이용하여 다음 식으로부터 계산된다.

$$\left[\frac{\nu \times \widehat{\sigma_A^2}}{\chi^2_{(\nu,1-\alpha/2)}} : \frac{\nu \times \widehat{\sigma_A^2}}{\chi^2_{(\nu,\alpha/2)}}\right] \tag{3.2}$$

여기서 $\chi^2_{(n:\alpha)}$는 자유도 n을 갖는 χ^2분포에서 오른쪽 끝 면적이 α인 χ^2값을 의미한다. 즉, σ_A^2의 90% 신뢰구간의 경우, Wald 통계량인 $Z = \widehat{\sigma_A^2}/S.E.(\widehat{\sigma_A^2})$으로 계산되고 식 (3.2)의 자유도는 $\nu = 2Z^2$이므로 표 3.2의 계산 결과를 대입하면 $\nu = 2 \times (0.2221/\sqrt{0.2878})^2 = 0.3427964558$이 되므로, χ^2값을 계산하는 SAS의 CINV 함수를 이용하면 $\chi^2_{(0.3427964558,0.95)} = 1.8342502618$과 $\chi^2_{(0.3427964558,0.05)} = 3.2831959E-8$이 계산되고 이 값을 식 (3.2)에 대입하여 계산하면 표 3.2의 COL4의 Lower와 Upper에 나타난 바와 같이 σ_A^2의 90% 신뢰구간은 [0.04152 : 2304415] 이 된다.

σ_E^2의 신뢰구간을 구하기 위해서는 식 (3.2)에서 $\widehat{\sigma_A^2}$대신 $\widehat{\sigma_E^2}$을 대입하여 계산한다. 즉, $\nu = 2 \times (\widehat{\sigma_E^2}/S.E.(\widehat{\sigma_E^2}))^2 = 2 \times (1.8361/\sqrt{0.4942})^2 = 13.6433$이 되고 SAS 의 CINV 함수를 사용하여 $\chi^2_{(13.6433,0.05)} = 23.2144$와 $\chi^2_{(13.6433,0.95)} = 6.32713$으로 계산되므로 σ_E^2의 90% 신뢰구간은 [1.07909 : 3.952922]로 계산되고 이 값은 표 3.2의 Residual의 Lower와 Upper의 값과 각각 소수 셋째 자리와 둘째 자리까지 일치한다.

3.3. β_i의 구간추정

식 (2.2)의 회귀계수에 대한 신뢰구간을 계산하기 위하여 표 3.3의 PROC MIXED문 아래의 MODEL 문장의 옵션에서 ALPHA = 0.1 CL S 를 사용하면 회귀계수 β_i들의 점추정값과 회귀계수들의 90% 신뢰구간을 계산할 수 있다.

표 3.3: 제한최대우도추정법을 이용하여 $\widehat{\beta}_i$들을 계산하기 위한 PROC MIXED의 예시

```
PROC MIXED DATA=EXAMPLE NOCLPRINT NOITPRINT NOINFO;
CLASSES COL4;
MODEL COL1 = COL2 COL3 / ALPHA = 0.1 CL S;
RANDOM COL4;
```

표 3.4: 표 3.3을 1회 실행시켰을 때 $\widehat{\beta}_i$들의 추정값과 β_i들의 90% 신뢰구간

Solution for Fixed Effects								
Effect	Estimate	Standard Error	DF	tValue	Pr>|t|	Alpha	Lower	Upper
Intercept	1.7151	0.9271	2	1.85	0.2055	0.1	-0.9920	4.4223
COL2	0.9968	1.1412	13	0.87	0.3983	0.1	-1.0242	3.0178
COL3	2.7700	1.0632	13	2.61	0.0218	0.1	0.8872	4.6527

표 3.3을 1회 실행한 결과 β_1의 추정값은 0.9968로 계산되고 β_1의 90% 신뢰구간은 다음 식으로 계산된다.

$$\left[\widehat{\beta_1} - t_{(\hat{\nu}:\alpha/2)} \times S.E.(\widehat{\beta_1}) : \widehat{\beta_1} + t_{(\hat{\nu}:\alpha/2)} \times S.E.(\widehat{\beta_1})\right] \tag{3.3}$$

여기서 $t_{(n:\alpha)}$ 는 자유도 n을 갖는 t분포에서 오른쪽 끝 면적이 α인 t값을 의미하고, $S.E.(\widehat{\beta_1})$은 $\widehat{\beta_1}$의 표준오차이다. t값을 구하기 위하여 SAS의 $TINV$함수를 사용하면 $TINV(0.95:13) = 1.770933396$이 계산되는데 여기서 자유도가 13인 이유는 표 3.3에서 사용한 모형의 그룹의 크기가 각각 $3, 5, 10$ 이었기 때문에 관측값의 합은 $n = \sum_{i=1}^{3} n_i = 18$이 된다. 그러므로 t값을 계산하기 위한 자유도는 $n - rank(\boldsymbol{XB}) = 18 - 5 = 13$으로 계산된다. (3.3)의 공식에 따라 β_1의 90% 신뢰구간은 $0.9968 \pm t(13:0.05) \times 1.1412 = 0.9968 \pm (1.770933396) \times 1.1412$을 계산하여 표 3.4의 COL2의 Lower 와 Upper에 $[-1.0242 : 3.0178]$가 된다.

4. 모의실험의 실행 및 결과

PROC MIXED가 포함된 표 3.1과 표 3.3을 이용하여 SAS/IML로서 코딩한 매크로 프로그램을 모의실험을 하기 위해서는 표 3.1과 표 3.3 이전의 문장에 독립변수 X_{1ij}와 X_{2ij}를

랜덤으로 발생시켜서 임의의 행렬안의 제 2 열(COL2)과 3 열(COL3) 에 저장한다. 그리고 회귀계수 $\beta_0, \beta_1, \beta_2$의 임의의 상수값을 부여한 후, SAS의 정규확률변수의 난수발생함수인 RANNOR를 이용하여 각각 $N(0, \sigma_A^2)$과 $N(0, \sigma_E^2)$의 가정에 따라 발생시킨 값인 랜덤효과 A_i와 오차항 E_{ij}를 생성하고, 식 (2.2)의 우변과 같이 모두 합하여 종속변수 Y_{ij}를 생성한다. 그리고 그 결과를 임의의 행렬의 제 1열(COL1)에 저장한다. 이 때 그 임의의 행렬의 제 4열(COL4)에는 각 그룹을 표시하는 그룹의 번호를 저장한다. 이러한 방법으로 생성된 난수들을 사용하여 표 3.1과 표 3.3이 포함된 매크로 프로그램으로 모의실험을 실행시킨다.

유도한 신뢰구간들은 짧아야 바람직하지만 유도된 신뢰구간들은 우선적으로 주어진 신뢰계수를 유지해야 한다. 불균형중첩오차구조를 갖는 단순회귀모형에 대해서 표 4.1과 같이 그룹의 크기 $g = 3$이고 관측값의 크기 $n = 18$인 소표본인 경우와 $g = 30$이고 $n = 173$인 대표본인 경우와 그들의 대략 중간크기에 해당하는 $g = 10$이고 $n = 53$인 세 가지 패턴에 대하여 모의실험을 실행시킨다. $\rho = \sigma_A^2/(\sigma_A^2 + \sigma_E^2)$이라하면, 일반성을 잃지 않고 $\sigma_A^2 = 1 - \sigma_E^2$으로 적을 수 있으므로, $\rho = \sigma_A^2$이 되고, $1 - \rho = \sigma_E^2$이 된다. 모의실험 실행시 ρ의 값을 $0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$ 까지 변화시키되 각각의 ρ값에 대해서 2000번씩 반복하여 모의실험을 실행하였다. 각 2000번의 모의실험을 실행한 다음, 주어진 모수를 포함하는 신뢰구간들의 개수를 2000으로 나누어 신뢰계수를 계산하였다.

표 4.1: 모의실험에 사용된 g와 n_i의 값

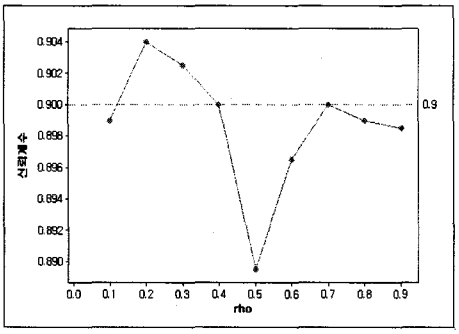
패턴	g	n_i	n
1(소표본)	3	3 5 10	18
2 (패턴 1 과 3 사이 크기의 표본)	10	1 1 1 5 5 5 5 10 10 10	53
3(대표본)	30	1 1 1 3 3 3 3 5 5 5 5 5 6 6 6 6 6 6 7 7 7 7 7 8 8 8 8 10 10 10	173

그리고 신뢰구간의 평균길이는 계산된 신뢰구간들의 상한에서 하한을 감한 모든 신뢰구간들의 길이의 평균값이 된다. 이항분포에 대한 정규근사를 사용하면 신뢰계수가 0.9 일 때 2000번의 모의실험 실행에서 추정된 신뢰계수가 0.887 보다 작을 기회는 2.5%보다 작다.

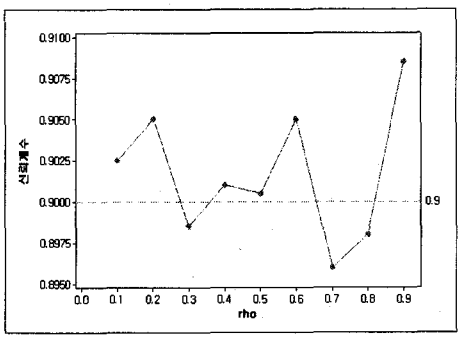
그림 4.1은 두 개의 분산 σ_E^2과 σ_A^2에 대한 90% 신뢰구간을 구하기 위하여 ρ의 각 값에 따라 2000회씩 모의실험을 실행하였을 때 계산된 신뢰계수를 나타낸다. 그림 4.2는 세 개의 회귀계수 β_0와 β_1과 β_2에 대한 90% 신뢰구간을 구하기 위하여 ρ의 각 값에 따라 2000회씩 모의실험을 실행하였을 때 계산된 신뢰계수를 나타낸다.

그림 4.1(a), 그림 4.1(b), 그림 4.1(c)로부터 σ_E^2의 90% 신뢰구간을 계산했을 때 신뢰계수는 ρ가 0.1부터 0.9까지 변하더라도 0.887보다 모두 크며 주어진 신뢰계수 0.9에 매우 가까운 값을 유지하는 것을 보인다. 그러므로 표본의 크기에 관계없이 PROC MIXED에서 계산되는 σ_E^2의 신뢰구간은 주어진 신뢰계수를 유지함을 알 수 있다.

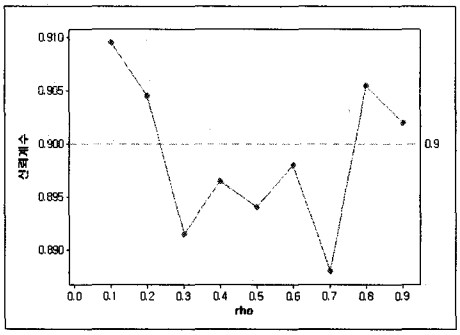
그림 4.1(d), 그림 4.1(e), 그림 4.1(f)는 σ_A^2의 90%신뢰구간을 계산했을 때 신뢰계수를



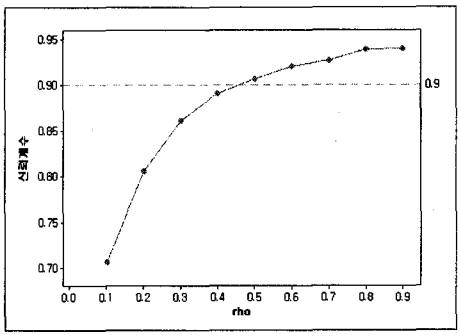
(a) 소표본(패턴1)에서 σ_E^2의 90%신뢰구간을 계산했을 때 신뢰계수의 변화



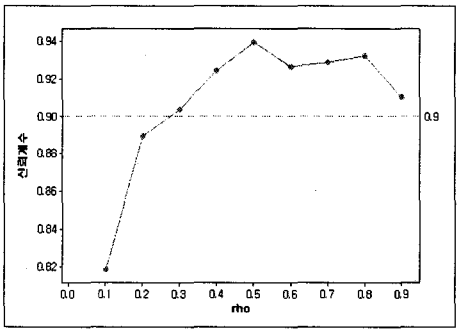
(b) 패턴2에서 σ_E^2의 90%신뢰구간을 계산했을 때 신뢰계수의 변화



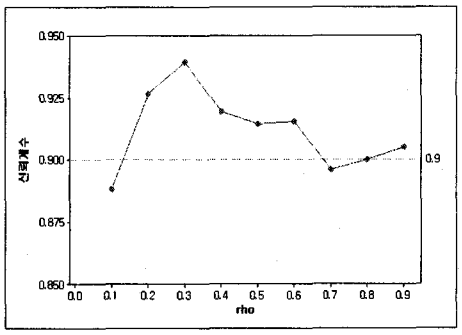
(c) 대표본(패턴3)에서 σ_E^2의 90%신뢰구간을 계산했을 때 신뢰계수의 변화



(d) 소표본(패턴1)에서 σ_A^2의 90%신뢰구간을 계산했을 때 신뢰계수의 변화



(e) 패턴2에서 σ_A^2의 90%신뢰구간을 계산했을 때 신뢰계수의 변화

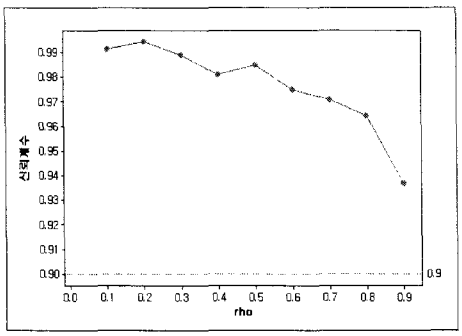


(f) 대표본(패턴3)에서 σ_A^2의 90%신뢰구간을 계산했을 때 신뢰계수의 변화

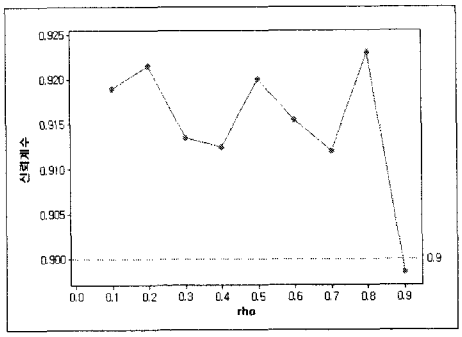
그림 4.1: σ_E^2과 σ_A^2에 대한 신뢰계수의 변화

보인다. 그림 4.1(d)와 그림 4.1(e)로부터 PROC MIXED는 표본의 크기가 작을 때, 즉, 패턴1과 패턴2에서는 각각 $\rho \leq 0.3$이나 $\rho \leq 0.2$일 때 신뢰계수들이 주어진 신뢰계수 0.9보다

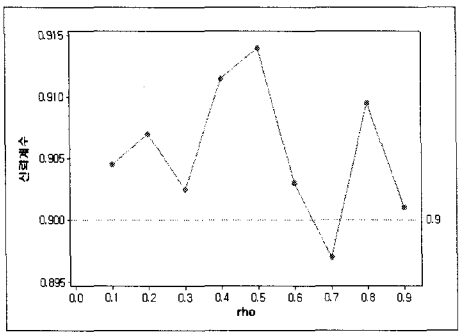
현저히 낮게 나타남을 볼 수 있다. 그러나 패턴3과 같이 그룹의 크기가 30이고 총 관측값의 크기가 173이 되면 비로소 PROC MIXED에서 계산되는 σ_A^2의 신뢰구간은 주어진 신뢰계수를 유지함을 알 수 있다.



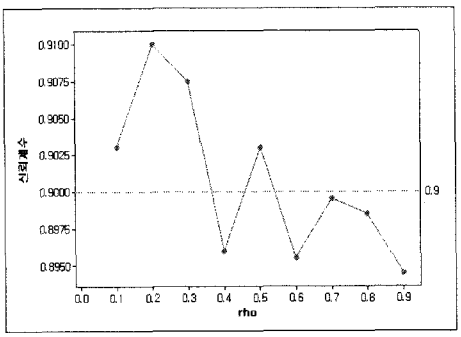
(a) 소표본(패턴1)에서 β_0의 90%신뢰구간을 계산했을 때 신뢰계수의 변화



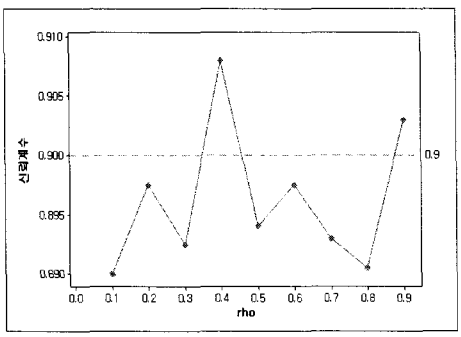
(b) 패턴2에서 β_0의 90%신뢰구간을 계산했을 때 신뢰계수의 변화



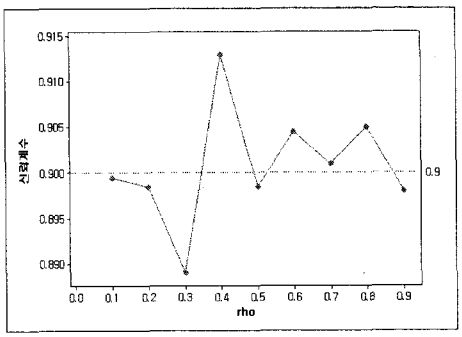
(c) 대표본(패턴3)에서 β_0의 90%신뢰구간을 계산했을 때 신뢰계수의 변화



(d) 소표본(패턴1)에서 β_1의 90%신뢰구간을 계산했을 때 신뢰계수의 변화

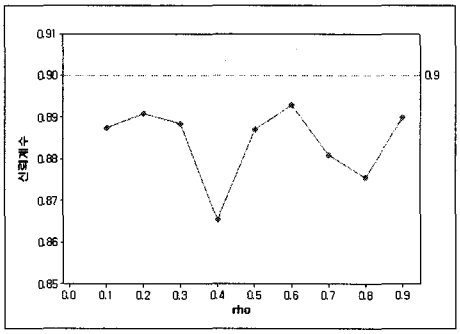


(e) 패턴2에서 β_1의 90%신뢰구간을 계산했을 때 신뢰계수의 변화

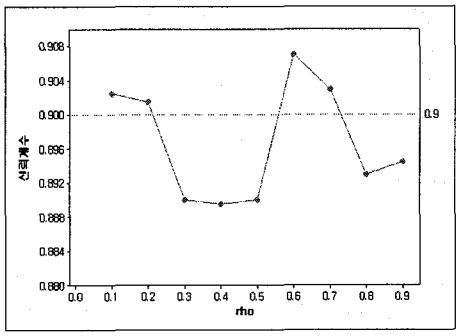


(f) 대표본(패턴3)에서 β_1의 90%신뢰구간을 계산했을 때 신뢰계수의 변화

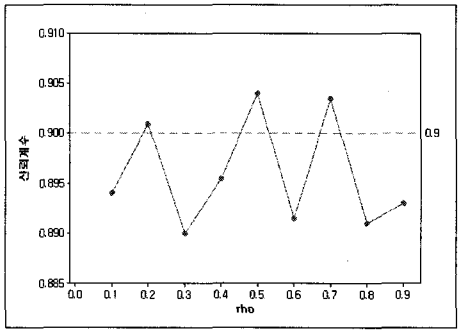
그림 4.2: $\beta_0, \beta_1, \beta_2$에 대한 신뢰계수의 변화(계속)



(g) 소표본(패턴1)에서 β_2의 90%신뢰구간을 계산했을 때 신뢰계수의 변화



(h) 패턴2에서 β_2의 90%신뢰구간을 계산했을 때 신뢰계수의 변화



(i) 대표본(패턴3)에서 β_2의 90%신뢰구간을 계산했을 때 신뢰계수의 변화

그림 4.2: $\beta_0, \beta_1, \beta_2$에 대한 신뢰계수의 변화

그림 4.2(a), 그림 4.2(b), 그림 4.2(c)는 β_0의 90%신뢰구간의 신뢰계수를 보인다. 그림 4.2(a)로부터 PROC MIXED는 패턴1에서 ρ의 값이 커짐에 따라 신뢰계수는 0.9에 가까워지지만 모든 ρ값에 대해서 신뢰계수가 0.9보다 매우 크게 나타나는 것을 볼 수 있다. 그러나 그림 4.2(b)의 패턴2와 그림 4.2(c)의 패턴3에서 신뢰계수들은 비교적 신뢰계수를 잘 유지하는 것을 볼 수 있다.

그림 4.2(d), 그림 4.2(e), 그림 4.2(f)는 β_1의 90%신뢰구간의 신뢰계수를 보인다. 세 개의 패턴으로부터 구한 신뢰계수들이 ρ의 모든 값에 대해서 주어진 신뢰계수 0.9를 잘 유지함을 알 수 있다. 그림 4.2(g), 그림 4.2(h), 그림 4.2(i)는 β_2의 90%신뢰구간의 신뢰계수를 보인다. β_1과 같이 β_2의 신뢰계수들도 세 개의 패턴에 대해서 주어진 신뢰계수 0.9에 비교적 가까운 값을 보이고 있다.

5. 결론

혼합모형의 분산에 대한 불편추정량을 구할 때 제한최대우도추정법을 사용하는 PROC

MIXED를 모의실험하여 랜덤효과인 그룹간의 분산 σ_A^2과 그룹내의 분산 σ_E^2과 고정효과인 회귀계수 β_0와 β_1과 β_2들의 신뢰구간을 계산해 보았다. 일반적으로 PROC MIXED로부터 계산되는 σ_E^2에 대한 신뢰구간과 β_1과 β_2에 대한 신뢰구간은 표본의 크기와 관계없이 주어진 신뢰계수를 비교적 잘 지켰으나, 그룹간의 분산인 σ_A^2의 신뢰구간은 표본의 크기가 패턴1과 같이 소표본일 때 $\rho \leq 0.3$인 경우와 소표본 보다는 약간 큰 패턴2일 때 $\rho \leq 0.2$인 경우에 신뢰계수를 유지하지 못하였다. 소표본의 경우, 그룹간의 분산인 σ_A^2의 올바른 신뢰구간을 구하기 위한 하나의 대안으로 Park과 Burdick(2003)을 참고할 수 있다. 또한 고정효과인 β_0에 대한 신뢰구간들은 표본의 크기가 적절한 패턴2와 대표본인 패턴3에서는 ρ의 모든 값에 대해서 비교적 신뢰계수를 잘 유지하였으나 패턴1과 같이 소표본인 경우 주어진 신뢰계수보다 훨씬 큰 매우 보수적인 결과를 보이므로 향후 연구에서 여기에 대한 개선방법을 찾도록 하겠다.

참고문헌

SAS Online Doc(version 9.1), http://v8doc.sas.com/sashtml.

Brown, H. and Prescott, R. (1999). *Applied Mixed Models in Medicine*, John Wiley & Sons, Inc.

Burdick, R. K. and Graybill, F. A. (1992). *Confidence Intervals on Variance Components*, Marxel Dekker, Inc.

Henderson, C. R. (1953). Estimation of variance and covariance components, *Biometrics*, **9**, 226-252.

Littell, R. C., Milliken, G. A., Stroup, W. W., and Wolfinger, R. D. (1996). *SAS Systems for Mixed Models*, SAS Institute Inc. Cary, NC, USA.

Milliken, G. A. and Johnson, D. E. (2002). *Analysis of Messy Data Volume III: Analysis of Covariance*, Chapman & Hall/CRC

Park, D. J. and Burdick, R. K. (2003). Performance of confidence intervals in regression models with unbalanced one-fold nested error structure, *Communications in Statistics-Simulation and Computation*, **32**, 3 717-732.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992). *Variance Components*, John Wiley & Sons, Inc.

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*, **35**, 135-143.

Verbeke, G. and Molenberghs, G. (1997). *Linear Mixed Models in Practice: A SAS-oriented Approach*, Springer-Verlag, New York, Inc.

Weerahandi, S. (1993). Generalized confidence intervals, *Journal of the American Statistical Association*, **88**, 899-905.

[2005년 10월 접수, 2006년 1월 채택]

Interval Estimation in Mixed Model by Use of PROC MIXED*

Dong Joon Park[1)]

ABSTRACT

PROC MIXED in SAS can be utilized to make inferences on parameters in a mixed model by use of Restricted Maximum Likelihood Estimation Method or Maximum Likelihood Estimation Method which has more merits than ANOVA method. A regression model with unbalanced nested error structure that belongs to a mixed model is used to construct confidence intervals on variances among groups, within groups, and regression coefficients in the model. PROC MIXED is applied to three different sample sizes for simulation. As a result of the simulation study, PROC MIXED generates confidence intervals on parameters that maintain the stated confidence coefficient in a large sample size. However, it does not generate confidence intervals that maintain the stated confidence coefficient for variance components among groups and intercept in a small sample size.

Keywords: PROC MIXED, Restricted Maximum Likelihood Estimation, Mixed Model, Confidence Interval

* This work was supported by Pukyong National University Research Fund in 2005 (project number: PK-2005-014)

1) Professor, College of Natural Sciences, Pukyong National University, 599-1, Daeyeon3-Dong, Nam-Gu Busan, 608-737, Korea.
Email: djpark@pknu.ac.kr