혼합모형의 구간추정을 위한 PROC MIXED의 활용

박 동 준[1)]

요 약

SAS의 PROC MIXED는 ANOVA 추정량보다 더 다양한 잔차최대우도추정법 또는 최대우도추정법으로 모수들을 추론할 수 있다. 혼합모형에 속하는 불균형중첩오차구조를 갖는 선형회귀모형에서 랜덤효과에 해당되는 그룹간의 분산과 고정효과에 해당되는 회귀계수들에 대한 신뢰구간을 구하기 위하여 대표본인 경우와 소표본인 경우에 대하여 PROC MIXED를 사용한다. 시뮬레이션을 실행한 결과, 대표본인 경우에는 모수들의 신뢰구간을 구하기 위하여 PROC MIXED를 활용할 수 있지만, 소표본인 경우에는 PROC MIXED를 사용할 경우, 그룹간 분산과 회귀계수 가운데 하나인 절편항에 대한 신뢰구간은 시뮬레이터된 신뢰계수가 명시한 신뢰계수를 지키지 못하는 것을 보인다.

KEY WORDS: Restricted Maximum Likelihood, PROC MIXED, 혼합모형, 구간추정

1. 서 론

잔차최대우도추정법이나 최대우도추정법을 사용할 수 있는 PROC MIXED[2]는 고정효과와 랜덤효과를 함께 포함하는 혼합모형에서 고정효과와 랜덤효과와 관련된 모수들의 추론에 사용될 수 있다. 이 소고에서는 불균형중첩오차구조를 갖는 단순회귀모형의 랜덤효과에 나타나는 분산들과 회귀계수에 대한 신뢰구간을 구할 때, 대표본과 소표본의 경우 PROC MIXED에서 계산되는 신뢰구간의 문제점을 시뮬레이션을 통하여 지적하고자 한다.

2절에서는 PROC MIXED를 적용할 구체적인 불균형중첩오차구조를 갖는 단순회귀모형을 상술하였다. 3절에서는 그 모형의 모수들의 추정량들과 신뢰구간을 구하기 위하여 PROC MIXED 문에서 구체적으로 사용해야 하는 option들을 적고 실제로 계산되는 과정을 설명하였다. 4절에서는 대표본과 소표본에 대하여 시뮬레이션을 실행한 결과를 그래프로 제시하고, 마지막으로 5절에서 결론을 맺는다.

2. 불균형중첩오차구조를 갖는 회귀모형의 행렬형태로 표현

불균형중첩오차구조를 갖고 k개의 독립변수를 갖는 단순회귀모형은 다음과 같이 적는다.

$$Y_{ij} = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \cdots + \beta_k X_{kij} + A_i + E_{ij} \qquad (2.1)$$
$$i = 1, \ldots, g;\ j = 1, \ldots, n_i$$

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

여기서, Y_{ij}는 i 번째 그룹의 j 번째 관찰값이고, $\beta_0, \beta_1, ..., \beta_k$는 회귀계수, $X_{1ij}, ..., X_{kij}$는 고정된 예측변수, A_i는 i 번째 그룹과 관련된 오차항이고, E_{ij}는 i 번째 그룹내의 j 번째 관찰값들과 관련된 오차항으로서 A_i와 E_{ij}는 서로 독립이고, 평균이 0 이고, 각각의 분산이 σ_A^2 와 σ_E^2 인 정규확률변수이며, $I > 2$, $n_i \geq 1$, 적어도 하나의 i 에 대해서는 $n_i > 1$ 이고, $n = \Sigma_i n_i$ 이다. 식(2.1)에서 β_0 항부터 $\beta_k X_{kij}$ 까지는 고정효과들이고, A_i 는 랜덤효과이므로 식(2.1)은 혼합모형이 된다. 이 식에서 나타나는 모수로는 회귀계수 β_i들과 분산 σ_A^2 과 σ_E^2 이 있다.

위와 같은 혼합모형의 모수들을 구간추정하기 위하여 PROC MIXED를 사용할 수 있다. PROC MIXED에서 사용하는 혼합모형은 표준의 선형모형을 일반화한 것으로서 다음과 같이 적는다.

$$\underline{y} = \boldsymbol{X}\underline{\beta} + \boldsymbol{B}\underline{\gamma} + \underline{\epsilon} \tag{2.2}$$

여기서, $\underline{y}$는 관찰자료들의 벡타, $\underline{\beta}$ 는 고정효과인 설계행렬 $\boldsymbol{X}$ 와 관련된 미지의 모수들의 벡타, $\underline{\gamma}$ 는 설계행렬 $\boldsymbol{B}$ 와 관련된 랜덤효과들의 벡타, $\underline{\epsilon}$ 은 오차항들을 포함하는 벡타이다. 식(2.1)을 식(2.2)의 행렬의 형태로 적으면 $\underline{y}$는 크기가 $n \times 1$인 벡타, $\boldsymbol{X}$ 는 $n \times (k+1)$ 인 행렬, $\underline{\beta}$ 는 $\beta_0, \beta_1, ..., \beta_k$ 를 원소로 갖는 $(k+1) \times 1$ 인 벡타, $\boldsymbol{B}$ 는 $n \times g$ 인 행렬, $\underline{\gamma}$ 는 A_i 를 원소로 갖는 $g \times 1$ 인 벡타로서 $\underline{\gamma} \sim N(\underline{0}, \sigma_A^2 \boldsymbol{I}_g)$, $\underline{\epsilon}$ 은 E_{ij} 를 원소로 갖는 $n \times 1$ 인 벡타로서 $\underline{\epsilon} \sim N(\underline{0}, \sigma_E^2 \boldsymbol{I}_n)$이고, $\underline{\gamma}$ 와 $\underline{\epsilon}$ 는 서로 독립이다.

3. PROC MIXED를 이용한 분산과 회귀계수의 구간추정

Park and Burdick(2003)에서 모수들의 신뢰구간을 유도하는데 필요한 통계량 $\boldsymbol{W}$ 와 $\boldsymbol{F}$ 를 제안하였다. 즉, $\boldsymbol{W} = \boldsymbol{F}\boldsymbol{B}\boldsymbol{B}'\boldsymbol{F}$, 그리고 $\underline{z} = \boldsymbol{F}\underline{y}$, 여기서 $\boldsymbol{F} = \boldsymbol{X}^*(\boldsymbol{X}^{*\prime}\boldsymbol{X}^*)^+ - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^+\boldsymbol{X}'$, $\boldsymbol{X}^* = [\boldsymbol{X}, \boldsymbol{B}\boldsymbol{B}']$, $^+$ 는 Moore-Penrose inverse를 의미한다. 이 때 평균제곱 $S_E^2 = \underline{y}'[\boldsymbol{I}_n - \boldsymbol{X}^*(\boldsymbol{X}^{*\prime}\boldsymbol{X}^*)^+\boldsymbol{X}^{*\prime}]\underline{y} / (n-g-k)$가 정의되면 $(n-g-k)S_E^2/\sigma_E^2$ 은 자유도 $n-g-k$ 를 갖는 카이제곱분포를 한다. 그러므로 i 번째 그룹내의 j 번째 관찰값들과 관련된 오차항 E_{ij}의 분산인 σ_E^2 에 대한 $100(1-\alpha)\%$ 정확한(exact) 신뢰구간은 다음과 같이 구할 수 있다.

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

다음으로 i 번째 그룹과 관련된 오차항 A_i 의 분산인 σ_A^2 의 신뢰구간을 구하기 위하여 $S_M^2 = \underline{z}'\boldsymbol{W}^+\underline{z}$ 이고 $\sigma_E^2 = 0$ 일 때 $(g-1)S_M^2/\sigma_A^2 \sim \chi_{g-1}^2$ 인 성질과 S_M^2 과 S_E^2이 서로 독립인 성질을 이용하여 Park and Burdick(2003)은 σ_A^2 의 $100(1-\alpha)\%$ 근사적인(approximate) 신뢰구간을 제안하였다.

그러나 PROC MIXED에서는 랜덤효과의 분산인 σ_A^2 과 오차항 E_{ij}의 분산인 σ_E^2 의 신뢰구간을 구하기 위하여 PROC MIXED 문장 다음에 CL 이란 옵션을 쓰면 Restricted Maximum Likelihood 으로 계산된 Wald 통계량을 사용하여 σ_A^2 과 σ_E^2 의 신뢰구간을 구할 수 있다. 우선 식(2.1)을 PROC MIXED에 활용하기 위하여 독립변수의 수가 두 개인 경우를 생각하면 (3.2)와 같이 쓸 수 있다.

$$Y_{ij} = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + A_i + E_{ij} \qquad (3.2)$$
$$i = 1,...,g;\ j = 1,...,n_i$$

즉, <표 3.1>의 이전의 문장에 SAS/IML을 이용하여 독립변수 X_{1ij} 와 X_{2ij} 를 랜덤으로 발생시켜서 임의의 행렬안의 COL2 와 COL3 에 저장하고, 회귀계수 β_0, β_1, β_2 의 임의의 상수값을 부여한 다음, 그룹간의 오차인 A_i 와 E_{ij} 를 식(2.1)의 가정에 따라 SAS의 RANNOR 함수를 이용하여 각각 $N(0,\sigma_A^2)$ 과 $N(0,\sigma_E^2)$로부터 발생시킨 후, 그 값들을 모두 합하여 종속변수 Y_{ij} 를 생성하고, 그 결과를 행렬의 COL1에 저장한다. 이 때 행렬의 COL4에는 각 그룹을 표시하기 위하여 그룹의 번호를 저장한다. 따라서 <표 3.1>의 PROC MIXED문에 분산 σ_A^2 과 σ_E^2 의 불편추정량의 값을 계산하고 90% 신뢰구간을 계산하기 위해 ALPHA = 0.1 과 CL 과 METHOD = REML로 적었다.

<표 3.1> REML를 이용하여 분산 σ_A^2 과 σ_E^2 을 계산하기 위한 PROC MIXED의 예

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

<표 3.2> <표 3.1>을 1회 실행시켰을 때 분산 σ_A^2 과 σ_E^2 의 추정값과 σ_A^2 과 σ_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 3.2>를 실행한 결과, 두 분산의 추정값은 $\hat{\sigma}_A^2 = 0.2221$, $\hat{\sigma}_E^2 = 1.8361$ 로 계산되었다. 여기서 σ_A^2 의 90% 신뢰구간에 대한 계산은 Wald 통계량을 이용하여 다음 식으로부터 계산된다.

$$[\ \frac{\nu\hat{\sigma_A^2}}{\chi^2_{\nu,1-\alpha/2}}\ :\ \frac{\nu\hat{\sigma_A^2}}{\chi^2_{\nu,\alpha/2}}\ ] \qquad (3.3)$$

즉, σ_A^2 의 90% 신뢰구간의 경우, Wald 통계량인 $Z=\hat{\sigma^2}/S.E.(\hat{\sigma^2})$ 으로 계산되고 식(3.3)의 $\nu=2Z^2$ 이므로 $\nu=2*(0.2221/\sqrt{0.2878})^2=0.3427964558$이고, 카이제곱분포의 확률값을 계산하는 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 (3.3)에 대입하여 계산하면 <표 3.2>의 COL4의 Lower와 Upper에 나타난 바와 같이 σ_A^2 의 90% 신뢰구간은 대략적으로 [0.04152 : 2304415]이 되고, 이와 유사한 방법으로 σ_E^2의 90% 신뢰구간도 [1.0791 : 3.9591]으로 계산된다.

식(2.1)의 회귀계수에 대한 신뢰구간을 계산하기 위하여 PROC MIXED 아래의 MODEL 문장의 옵션에서 ALPHA = 0.1 CL S 를 사용하면 회귀계수 β_i 들의 점추정값과 회귀계수들의 t-type의 90% 신뢰구간을 계산할 수 있다.

<표 3.3> Restricted ML를 이용하여 회귀계수 β_i를 계산하기 위한 PROC MIXED의 예

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

<표 3.4> <표 3.3>을 1회 실행시켰을 때 β_i의 추정값과 β_i의 90% 신뢰구간

Solution for Fixed Effects								
		Standard						
Effect	Estimate	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 3.3>을 실행한 결과 β_1 의 추정값은 0.9968로 계산되고 β_1의 90% 신뢰구간은 다음 식으로 계산된다.

$$[\ \hat{\beta_1}-t_{(\hat{\nu}:\alpha/2)}S.E.(\hat{\beta_1})\ :\ \hat{\beta_1}+t_{(\hat{\nu}:\alpha/2)}S.E.(\hat{\beta_1})\ ] \qquad (3.4)$$

$TINV$ 함수를 사용하여 $TINV(0.95:13)=1.770933396$이 계산되고 (3.4)의 공식에 따라 β_1의 신뢰구간은 $0.9968\pm t(13:0.05)\times 1.1412 = 0.9968\pm(1.770933396)\times 1.1412$ 을 계산하여 <표 3.4>의 COL2의 Lower 와 Upper에 [-1.0242 : 3.0178]로 계산된다.

4. 시뮬레이션의 실행

유도한 신뢰구간들은 짧은 신뢰구간이 바람직하지만 유도된 신뢰구간은 우선적으로 신뢰계수를 유지해야 한다. 불균형중첩오차구조를 갖는 단순회귀모형에 대해서 <표 4.1>과 같이 $g=3$

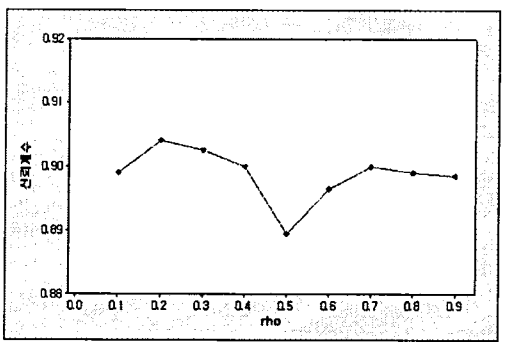
이고 $n=18$ 인 소표본인 경우와 $g=30$ 이고 $n=173$ 인 대표본의 두 가지 패턴에 대하여 <표 3.1>과 <표 3.3>을 SAS/IML로 프로그래밍한 매크로안에 넣어서 실행시킨다. $\rho=\sigma_A^2/(\sigma_A^2+\sigma_E^2)$ 이라하면, 일반성을 잃지 않고 $\sigma_A^2=1-\sigma_E^2$ 으로 적을 수 있으므로, $\rho=\sigma_A^2$ 이 되고, $1-\rho=\sigma_E^2$ 이 된다. 시뮬레이션의 실행시 ρ 의 값을 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 까지 변화시키되 각각의 ρ 값에 대해서 2000번씩 시뮬레이션을 실행하였다. 각 2000번의 시뮬레이션을 실행한 다음, 각각의 모수를 포함하는 신뢰구간들의 개수를 2000으로 나누어 신뢰계수를 계산하였다.

<표 4.1> 시뮬레이션에 사용된 g 와 n_i 의 값

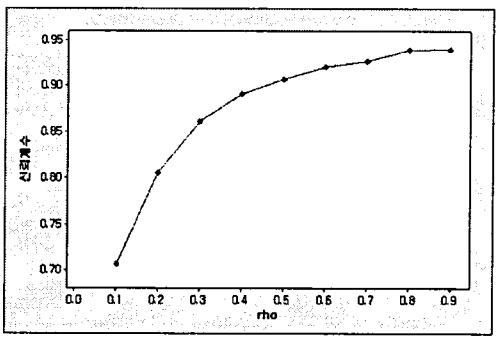
패 턴	g	n_i	n
1(소표본)	3	3 5 10	18
2(대표본)	30	1 1 1 3 3 3 3 5 5 5 5 5 6 6 6 6 6 6 7 7 7 7 7 8 8 8 8 10 10 10	173

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

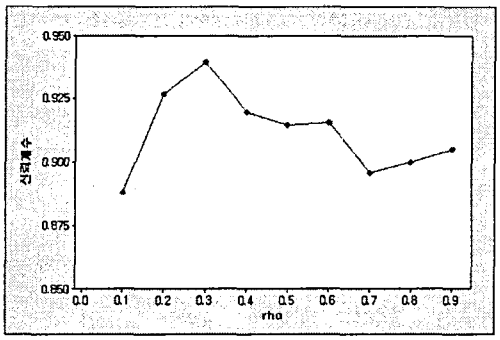
<그림 4.1>은 각 모수에 대한 90% 신뢰구간을 구하기 위하여 ρ 값에 따라 2000회씩 시뮬레이션을 실행하였을 때 계산된 신뢰계수를 나타낸다.



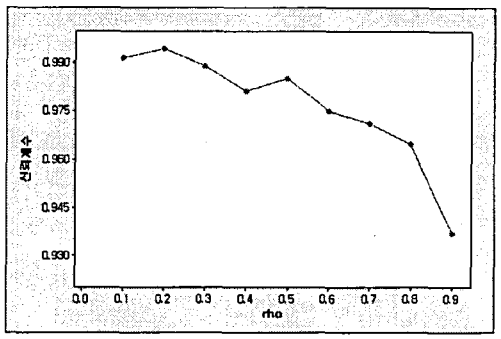
<그림 4.1a>소표본에서 σ_E^2의 90% 신뢰구간을 구했을 때 시뮬레이터된 신뢰계수의 변화



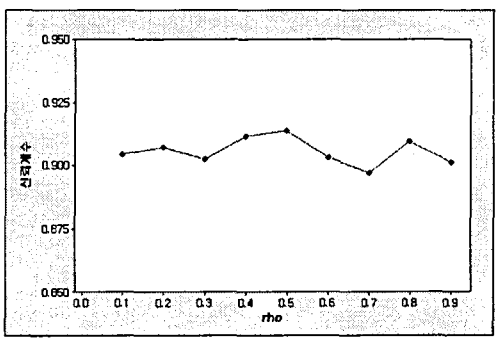
<그림 4.1b>소표본에서 σ_A^2의 90% 신뢰구간을 구했을 때 시뮬레이터된 신뢰계수의 변화



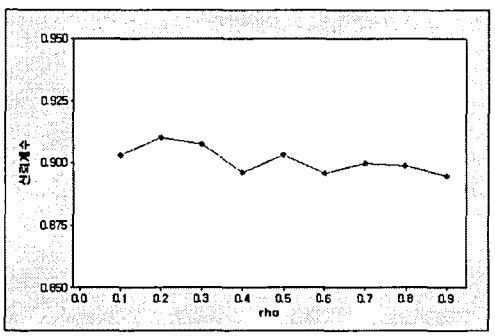
<그림 4.1c>대표본에서 σ_A^2의 90% 신뢰구간을 구했을 때 시뮬레이터된 신뢰계수의 변화



<그림 4.1d>소표본에서 β_0의 90% 신뢰구간을 구했을 때 시뮬레이터된 신뢰계수의 변화



<그림 4.1e>대표본에서 β_0의 90% 신뢰구간을 구했을 때 시뮬레이터된 신뢰계수의 변화



<그림 4.1f>소표본에서 β_1의 90% 신뢰구간을 구했을 때 시뮬레이터된 신뢰계수의 변화

5. 결 론

혼합모형의 분산에 대한 불편추정량을 구하는 잔차최대우도추정법을 사용하는 PROC MIXED를 시뮬레이션을 실행하여 그룹간의 분산 σ_A^2 과 그룹내의 분산 σ_E^2 과 고정효과에 해당되는 회귀계수들인 β_0 와 β_1 들의 신뢰구간을 계산해 보았다. 일반적으로 σ_E^2에 대한 신뢰구간과 β_1에 대한 신뢰구간은 명시한 신뢰계수를 지켰으나, 그룹간의 분산인 σ_A^2 과 고정효과인 β_0에 대한 신뢰구간들은 소표본인 경우, PROC MIXED는 명시한 신뢰계수를 제대로 지키지 못하였다. 소표본의 경우, 그룹간의 분산인 σ_A^2의 올바른 신뢰구간을 구하기 위한 하나의 대안으로 박동준(2003)을 참고할 수 있다. 그리고 향후, 고정효과인 β_0에 대한 신뢰구간을 위한 연구가 필요하다.

참고문헌

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, pp 361-370.

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