Improved Confidence Intervals on Total Variance in a Regression Model with Unbalanced Nested Error Structure

박동준[1], 이수진[2]

요 약

불균형중첩오차구조를 갖는 단순선형회귀모형에서 나타나는 두 분산의 합에 대한 신뢰구간을 구하기 위하여 Ting et al.(1990) 방법과 Graybill and Wang(1980) 방법과 Tsui and Weerahandi(1989)가 제안한 일반화 축량(generalized pivotal quantity)방법을 이용한 두 가지 방법 등 모두 네 가지 신뢰구간을 제안한다. 신뢰구간의 적절성을 판단하기 위하여 여러 가지 불균형 설계에 대하여 SAS/IML로 시뮬레이션을 실행하고 신뢰계수와 신뢰구간의 평균 길이를 비교한다. 불균형중첩오차구조를 갖는 단순선형회귀모형의 두 분산의 합에 대한 네 가지 신뢰구간들이 주샘플링 단위의 변화에 따라 어느 방법이 적절한 신뢰구간을 구축하는지 추천하고, 실제 예제를 적용하여 시뮬레이션의 결과와 일관성이 있는지를 확인한다.

주요용어 : 분산성분, 혼합모형, 구간추정

1. 불균형중첩오차구조를 갖는 단순선형회귀모형

통계 연구자들이 자료를 수집하는 과정에서 만나는 불균형중첩오차구조를 갖는 단순선형회귀모형은 아래와 같이 쓸 때

$$Y_{ij} = \mu + \beta X_{ij} + A_i + E_{ij} \qquad (1.1)$$
$$i = 1, \ldots, I\,;\; j = 1, \ldots, J_i$$

여기서, Y_{ij}는 i번째 수준의 j번째 반응변수값, μ와 β는 미지의 상수, X_{ij}는 i번째 수준의 j번째 예측변수이고, A_i는 주 샘플링 단위와 관련된 오차항이고 E_{ij}는 부 샘플링 단위와 관련된 오차항으로서 A_i와 E_{ij}는 평균이 0이고 분산이 각각 σ_A^2과 σ_E^2인 서로 독립인 정규 확률변수이다. $I \geq 2$이고 $J_i \geq 1$으로서 적어도 하나의 i에 대해 $J_i > 1$이다. 우리는 모형(1.1)에서 나타나는 두 분산의 합 $\gamma = \sigma_A^2 + \sigma_E^2$에 대한 여러 가지 신뢰구간을 제안한 다음, 시뮬레이션을 실행하여 어느 신뢰구간이 가장 적절한지를 살펴보고, 실제 예제에 적용하여 신뢰구간을 구한 후, 시뮬레이션에서 구한 결과와 실제 예제의 결과가 일관성이 있는지 확인한다.

2. 분산의 합에 대한 신뢰구간

1) 608-737 부산광역시 남구 대연 3동 599-1번지 부경대학교 자연과학대학 통계학전공 부교수
djpark@pknu.ac.kr

2) 608-737 부산광역시 남구 대연 3동 599-1번지 부경대학교 자연과학대학 통계학전공 석사과정

모형 (1.1)을 행렬의 형태로 고쳐 쓰면 다음과 같다.

$$\boldsymbol{y} = \boldsymbol{X}\underline{\alpha} + \boldsymbol{B}\boldsymbol{u} + \boldsymbol{e} \tag{2.1}$$

여기서, $\boldsymbol{y}$는 $J \times 1$ 관찰값 벡터, $\boldsymbol{X}$는 첫 번째 열의 값이 1이고 X_{ij} 값들이 두 번째 열을 구성하는 $J \times 2$ 행렬, $\underline{\alpha}$는 모수인 μ와 β를 원소로 하는 2×1벡터, $\boldsymbol{B}$는 0과 1의 값을 갖는 $J \times I$ 설계행렬 즉, $\boldsymbol{B} = \oplus_{i=1}^{I} \underline{1}_{J_i \times 1}$, $\boldsymbol{u}$는 랜덤효과를 나타내는 A_i를 원소로 갖는 $I \times 1$벡터, $\boldsymbol{e}$는 랜덤오차항인 E_{ij}를 원소로 갖는 $J_{\cdot} \times 1$벡터이고 $J_{\cdot} = \Sigma J_i$이다.

Olsen et al.(1976), Eubank et al.(2001)과 El-Bassiouni(1994)가 제안한 제곱합을 이용하여 Park and Burdick(2003)은 근사적으로 χ^2_{df}분포를 따르는 다음의 통계량을 제안했다. 행렬 w를 $W = FBB'F$ 이라 하고 $F = X^*(X^{*'}X^*)^+X^{*'} - X(X'X)^+X'$ 이고, $X^* = [X, BB']$이다. 행렬 W의 서로 다른 고유값을 $d_1, d_2, \ldots, d_m$; $l = 1, 2, \ldots, m$ 이면 $rank(W) = I - 1 = n_1$임을 보였다. 벡터 $\boldsymbol{z}$를 $\boldsymbol{z} = \boldsymbol{F}\boldsymbol{y}$로 정의하면 행렬 W와 벡터 $\boldsymbol{z}$를 이용한 평균 제곱을 $S_M^2 = [\boldsymbol{z}'\boldsymbol{W}^+\boldsymbol{z}]/n_1$ 이라 쓰고, 모형 (1.1)의 분산분석과정에서 나타나는 오차 평균 제곱을 S_E^2 을 $S_E^2 = y'[D_J - X^*(X^{*'}X^*)^-X^*]\,y/n_2$ 이라 할 때 $n_2 = J_{\cdot} - I - 1$이다. 그리고 Eubank et al. 로부터 다음이 보여졌다.

$$\frac{n_2 S_E^2}{\sigma_E^2} \sim \chi^2_{n_2} \tag{2.2a}$$

$$\frac{n_1 S_M^2}{\sigma_A^2 + \frac{1}{h}\sigma_E^2} \approx \chi^2_{n_1} \text{ 이고 } \sigma_E^2 = 0 \text{ 이면 } \frac{n_1 S_M^2}{\sigma_A^2} \sim \chi^2_{n_1} \tag{2.2b}$$

$$S_M^2 \text{과 } S_E^2 \text{은 서로 독립} \tag{2.2c}$$

$$E(S_M^2) = \sigma_A^2 + \frac{1}{h}\sigma_E^2 \tag{2.2d}$$

여기서, h는 행렬 W의 고유값들의 조화평균이다. 두 분산의 합 $\gamma = \sigma_A^2 + \sigma_E^2$에 대한 신뢰구간을 구하기 위하여 기대 평균제곱을 다음과 같이 쓴다.

$$\begin{aligned} E(S_M^2) &= \sigma_A^2 + \frac{1}{h}\sigma_E^2 = \theta_M \\ E(S_E^2) &= \sigma_E^2 = \theta_E \end{aligned} \tag{2.3}$$

그러면 식 (2.3)으로부터 γ는 $\gamma = \sigma_A^2 + \sigma_E^2 = \theta_M + (1 - 1/h)\theta_E$로 쓸 수 있다.

γ에 대한 신뢰구간을 구하기 위하여 Ting et al.(1990)방법에 적용한 결과 γ에 대한 $100(1-\alpha)\%$ 양쪽 신뢰구간은 다음 식으로 구해진다.

$$\begin{aligned} &[S_M^2 + (1 - \frac{1}{h})S_E^2 - \{K_1^2 S_M^4 + (1 - \frac{1}{h})^2 K_2^2 S_E^4 + (1 - \frac{1}{h})^2 K_{12} S_M^2 S_E^2\}^{\frac{1}{2}}; \\ &S_M^2 + (1 - \frac{1}{h})S_E^2 + \{L_1^2 S_M^4 + (1 - \frac{1}{h})^2 L_2^2 S_E^4 + (1 - \frac{1}{h})^2 L_{12} S_M^2 S_E^2\}^{\frac{1}{2}}] \end{aligned} \tag{2.4}$$

여기서, $F_1 = F_{(\alpha/2:n_1,n_2)}$, $F_2 = F_{(1-\alpha/2:n_1,n_2)}$, $K_1 = 1-1/F_{(1-\alpha/2:n_1,\infty)}$, $K_2 = 1/F_{(\alpha/2:n_2,\infty)}-1$, $K_{12} = [(F_1-1)^2 - L_1^2F_1^2 - L_2^2]/F_1$, $L_1 = 1/F_{(\alpha/2:n_1,\infty)}-1$, $L_2 = 1-1/F_{(1-\alpha/2:n_2,\infty)}$, $L_{12} = [(1-F_1)^2 - H_1^2F_1^2 - H_2^2]/F_1$ 이다. 식 (2.4)를 TING 방법이라 하자.

또 하나의 대안으로 S_M^2과 S_E^2을 Graybill과 Wang(1980) 방법에 적용하면 γ에 대한 $100(1-\alpha)\%$ 양쪽 신뢰구간은 다음 식으로 구해진다.

$$[S_M^2 + (1-\frac{1}{h})S_E^2 - \{G_1^2S_M^4 + (1-\frac{1}{h})^2G_2^2S_E^4\}^{\frac{1}{2}};$$
$$S_M^2 + (1-\frac{1}{h})S_E^2 + \{H_1^2S_M^4 + (1-\frac{1}{h})^2H_2^2S_E^4\}^{\frac{1}{2}}] \tag{2.5}$$

여기서, $G_1 = 1-1/F_{(1-\alpha/2;n_1,\infty)}$, $G_2 = 1-1/F_{(1-\alpha/2;n_2,\infty)}$, $H_1 = 1/F_{(\alpha/2;n_1,\infty)}-1$, $H_2 = 1/F_{(\alpha/2;n_2,\infty)}-1$. 식 (2.5)를 GRW 방법이라 하자.

이제 Tsui and Weerahandi(1989)가 제안한 일반화 p값을 활용한 일반화 축량 방법을 이용한다. 식 (2.2a)로부터 σ_E^2의 추정값은 $\hat{\sigma}_E^2 = n_2 s_E^2/U_E^*$ 으로 계산되는데 여기서, s_E^2은 오차평균 제곱 S_E^2의 관찰값들로서 σ_E^2의 구체적인 값이 주어진다면 $s_E^2 = \sigma_E^2\chi_{n_2}^2/n_2$ 으로 계산되고, 일반화축량 $U_E^* \sim \chi_{n_2}^2$ 로부터 생성되어진다. 식 (2.3)의 θ_M의 추정값은 $\hat{\theta}_M \fallingdotseq n_1 s_M^2/U_M^*$ 으로 계산되는데 여기서 s_M^2은 S_M^2의 관찰값들로서 θ_M의 구체적인 값이 주어지면 $s_M^2 \fallingdotseq \theta_M\chi_{n_1}^2/n_1$ 으로 계산되고, 일반화축량 $U_M^* \sim \chi_{n_1}^2$ 분포로부터 생성된다. 그러므로 γ의 추정값은 $\hat{\gamma} = \hat{\sigma}_A^2 + \hat{\sigma}_E^2 = \hat{\theta}_M + (1-1/h)\hat{\theta}_E = n_1 s_M^2/U_M^* + (1-1/h)n_2 s_E^2/U_E^*$ 로 구해진다. 이렇게 계산된 γ의 값들로 형성된 확률분포를 G라고 하자. γ에 대한 $100(1-\alpha)\%$ 양쪽 신뢰구간은 다음 식으로부터 구할 수 있다.

$$[G_{\alpha/2} \quad ; \quad G_{1-\alpha/2}] \tag{2.6}$$

여기서 $G_{\alpha/2}$와 $G_{1-\alpha/2}$는 각각 확률분포 G의 제 $\alpha/2$ 백분위수 값과 제 $1-\alpha/2$ 백분위수 값이다. 식 (2.6)을 GEN1 방법이라 하자.

Olsen et al. 은 $Q_l = \boldsymbol{z}'\boldsymbol{E}_l\boldsymbol{z} \sim (\sigma_E^2 + d_l\sigma_A^2)\chi_{r_l}^2$; $l = 1,...,m$ 임을 보였다. 여기서 $\boldsymbol{E}_l$은 행렬 W의 고유값 d_l로 구성된 고유공간의 직교사영 연산자이고 $S_E^2, Q_1, ..., Q_m$들은 서로 독립임이 보여졌다. 그러므로 $\Sigma Q_l/[d_l(\gamma-\sigma_E^2)+\sigma_E^2] = \chi_{n_1}^2$ 로 쓸 수 있고, 일반화 축량 방법을 적용하여 $\Sigma q_l/[d_l(\hat{\gamma}-n_2 s_E^2/U_E^*)+n_2 s_E^2/U_E^*] = U_M^*$ 로 적는다. 이 식에서 q_l은 Q_l의 관찰값들로서 σ_E^2과 σ_A^2의 구체적인 값들이 주어진다면 $q_l = (\sigma_E^2 + d_l\sigma_A^2)\chi_{r_l}^2$ 로 계산되고, 생성된 관찰값들 q_l과 관찰값 s_E^2과 일반화 축량값들 U_E^* 와 U_M^*를 대입하여 계산한 γ의 값들로 형성된 확률분포를 R이라고 하자. 그러면 일반화 축량을 이용한 γ에 대한 $100(1-\alpha)\%$ 양쪽 신뢰구간은 다음 식으로부터 구할 수 있다.

$$[R_{\alpha/2} \quad ; \quad R_{1-\alpha/2}] \tag{2.7}$$

여기서 $R_{\alpha/2}$와 $R_{1-\alpha/2}$는 각각 확률분포 R의 제 $\alpha/2$ 백분위수 값과 제 $1-\alpha/2$ 백분위수 값이다. 식 (2.7)을 GEN2 방법이라고 하자.

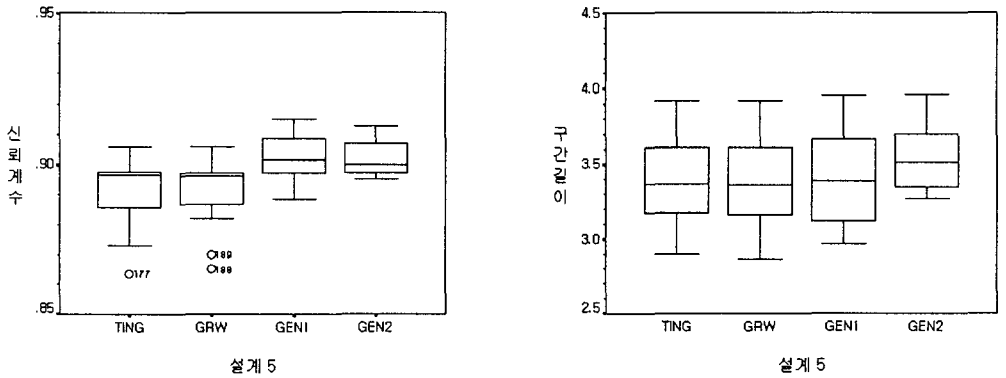
3. 시뮬레이션의 실행

신뢰구간을 비교하기 위하여 SAS/IML을 이용하여 El-Bassiouni and Seely(1996)가 사용한 일곱 가지의 설계에 세 가지(<표 3.1>의 설계번호 6, 9, 10)를 더 추가하여 모두 10가지 유형에 대한 γ의 신뢰구간을 비교한다. ρ를 총 분산에 대한 급간 분산의 비율로서 급내 상관계수 $\rho=\sigma_A^2/(\sigma_A^2+\sigma_E^2)$라 할 때 일반성을 잃지 않고 $\sigma_A^2=1-\sigma_E^2$으로 쓸 수 있고, 따라서 $\rho=\sigma_A^2$과 $1-\rho=\sigma_E^2$으로 쓴다. ρ=0.001, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.999로 모두 11개의 ρ값에 대하여 2000번씩 시뮬레이션을 실행시켜서 γ의 신뢰구간을 계산한다.

<표 3.1> 시뮬레이션에 사용된 불균형 설계

설계	표본의 크기	I	J_i	불균형 측도
1	15	3	3, 5, 7	0.887
2	15	3	1, 5, 9	0.458
3	30	3	2, 10, 18	0.458
4	30	6	1, 1, 5, 5, 9, 9	0.458
5	30	6	1, 1, 1, 1, 13, 13	0.289
6	157	6	1, 1, 2, 3, 50, 100	0.080
7	45	9	1, 1, 1, 5, 5, 5, 9, 9, 9	0.458
8	45	9	1, 1, 1, 1, 1, 1, 1, 19, 19	0.253
9	59	10	1, 1, 4, 5, 6, 6, 8, 8, 10, 10	0.524
10	55	10	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	0.621

설계번호 5의 경우에 <그림 3.1>은 γ의 명시된 신뢰계수가 90%일 때 시뮬레이터된 신뢰계수들과 신뢰구간들의 평균길이를 상자그림으로 나타낸 것이다.



<그림 3.1> γ의 명시된 신뢰계수가 90%일 때 시뮬레이터된 신뢰계수

4. 예제의 적용

γ의 네 가지 신뢰구간을 계산하기 위하여 Milliken과 Johnson(2002, p.428~429)의 자료를 사용한다. γ의 신뢰구간을 계산하기 위하여 사전점수를 설명변수 X_{ij}, 성적을 반응변수 Y_{ij}, 구체적인 교육방법으로 가르치는 교사들의 효과를 A_i, 그리고 학생 개인에 따른 오차를 E_{ij}로 가정한다. <표 3.1>의 설계번호 5를 적용하고 복원추출의 방법을 사용하여 $J_1=1$, $J_2=1$, $J_3=1$, $J_4=1$, $J_5=J_6=13$ 을 랜덤으로 선택한 결과를 <표 4.1>에 수록하였다.

학생 개인에 따른 오차의 분산의 추정량은 $\hat{\sigma}_E^2=8.457$으로 계산되고, 교사들의 효과에 대한 분산의 추정량은 $\hat{\sigma}_A^2=S_M^2-\hat{\sigma}_E^2/h=7.964-8.457/1.349=1.695$이다. 급내 상관계수의 추정량은 $\hat{\rho}=\hat{\sigma}_A^2/(\hat{\sigma}_A^2+\hat{\sigma}_E^2)=0.167$ 이 된다. <그림 3.1>의 신뢰계수 대한 상자그림에서 ρ가 작을 때($\rho=0.2$) TING방법과 GRW방법은 신뢰계수를 유지하지 못하므로 사용할 수 없고, 그 대신 ρ의 모든 값에서 신뢰계수가 유지되는 GEN1방법과 GEN2방법을 사용한다.

<표 4.1> Milliken과 Johnson(2002)에서 랜덤으로 선택된 자료들

관찰값	교사											
	4		5		5		8		7		9	
	Score (Y)	Pre-score (X)	Score (Y)	Pre-score (X)	Score (Y)	Pre-score (X)	Score (Y)	Pre-score (X)	Score (Y)	Pre-score (X)	Score (Y)	Pre-score (X)
1	85	76	83	65	86	87	85	67	89	93	99	88
2									94	75	99	89
3									85	75	89	78
4									84	72	95	75
5									85	63	90	76
6									84	65	91	82
7									91	78	92	90
8									88	73	92	74
9									88	69	92	75
10									89	68	90	70
11									88	62	94	92
12									84	66	90	78
13									83	66	92	89

<표 4.2> γ의 90% 신뢰구간

방법	신뢰하한	신뢰상한	구간의 길이	신뢰계수 유지여부	비고
TING	5.46	36.87	31.40	유지 못함	사용불가
GRW	5.72	37.00	31.28	유지 못함	사용불가
GEN1	5.81	36.36	30.55	유지함	추천
GEN2	8.18	39.82	31.64	유지함	추천

5. 결론

지금까지 불균형 중첩 오차구조를 갖는 단순 선형 회귀모형에서 나타나는 두 분산 σ_A^2과

σ_E^2의 합 $\gamma=\sigma_A^2+\sigma_E^2$에 대한 신뢰구간을 구하는 TING, GRW, GEN1, GEN2의 네 가지 방법을 유도하였다. 지면의 제한으로 <표 3.1>에 나타난 모든 설계에 대한 신뢰계수와 신뢰구간의 길이를 제시할 수 없지만 결론을 요약하면 주 샘플링 단위가 $I=3$일 때는 GRW방법과 GEN1방법과 GEN2방법을 사용하여 신뢰구간을 계산한 다음, 신뢰구간이 가장 짧게 나타나는 방법을 추천한다. 그러나 주 샘플링 단위가 $I\geq 6$이고 $\rho\geq 0.4$일 때는 GRW방법이나 비교적 정확한 신뢰계수를 유지하는 GEN1방법 또는 GEN2방법을 이용하여 신뢰구간이 짧은 것을 선택할 수 있지만 $\rho<0.4$일 때는 GEN1방법과 GEN2방법을 사용할 수 있다.

참 고 문 헌

[1] El-Bassiouni, M. Y. (1994), Short confidence intervals for variance components, *Communications in Statistics Theory and Method*, 23(7), 1915-1933.

[2] El-Bassiouni, M. Y. and Seely, J. F. A modified harmonic mean test procedure for variance components, *Journal of Statistical Planning and Inference*, 49, 319-326.

[3] Eubank, L., Seely, J., Lee, Y. (2001), Unweighted mean squares for the general two variance component mixed model, *Proceeding of the Graybill Conference, Ft. Collins, Co.*, June, 281-290.

[4] Graybill, F. A., Wang, C-M. (1980), Confidence intervals on nonnegative linear combinations of variances, *Journal of the American Statistical Association*, 75, 869-873.

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

[6] Olson, A., Seely, J., and Birkes, D. (1976), Invariant Quadratic Unbiased Estimation for Two Variance Components, *Annals of Statistics*, 4, 878-890.

[7] 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.

[8] 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.

[9] Tsui, K and Weerahandi, S. (1989), Generalized p-value in significance testing of hypotheses in the presence of nuisance parameters, *Journal of American Statistical Association*, 84, 602-607.