CONFIDENCE INTERVALS ON THE AMONG GROUP VARIANCE COMPONENT IN A REGRESSION MODEL WITH AN UNBALANCED ONE-FOLD NESTED ERROR STRUCTURE

Dong Joon Park[1]

In this article we consider the problem of constructing confidence intervals for a linear regression model with nested error structure. A popular approach is the likelihood-based method employed by PROC MIXED of SAS. In this paper, we examine the ability of MIXED to produce confidence intervals that maintain the stated confidence coefficient. Our results suggest the intervals for the regression coefficients work well, but the intervals for the variance component associated with the primary level cannot be recommended. Accordingly, we propose alternative methods for constructing confidence intervals on the primary level variance component. Computer simulation is used to compare the proposed methods. A numerical example and SAS code are provided to demonstrate the methods.

Key words: Mixed model; Generalized confidence interval; PROC MIXED

1. INTRODUCTION

In applications using a linear regression model with nested error structure, one might consider inferences for variance components of the model. A regression model with an unbalanced one-fold nested error structure includes two variance components-one in the primary level and one in the secondary level. This article considers confidence intervals for the variance component of the primary level. The model is described in Section 2 and confidence intervals for the primary level variance component are presented in Section 3. A simulation study for comparing the proposed intervals and conclusions are summarized in Section 4.

2. A REGRESSION MODEL WITH AN UNBALANCED ONE-FOLD NESTED ERROR STRUCTURE

The regression model with an unbalanced one-fold nested error structure is written as

[1] 부경대학교 자연과학대학 수리과학부 통계학전공 부교수, 부산광역시 남구 대연3동 599-1

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

where Y_{ij} is the j th random observation in the i th primary level, $\mu, \beta_1, \ldots, \beta_k$ are unknown constants, $X_{1ij}, \ldots, X_{kij}$ are fixed predictor variables, and A_i and E_{ij} are jointly independent normal random variables with zero means and variances σ_A^2 and σ_E^2, respectively. Further, $n_i \geq 1, \sigma_E^2$, $n_i > 1$ for at least one value of $i, n_i > 1$, and $n = \Sigma_{i=1}^{g} n_i$. Model (2.1) is written in matrix notation as

$$\underline{Y} = X\underline{\beta} + B\underline{U} + \underline{E} \qquad (2.2)$$

where $\underline{Y}$ is a $n \times 1$ random vector of observations, X is a $n \times (k+1)$ matrix of known values with a column of 1's in the first column, $\underline{\beta}$ is a $(k+1) \times 1$ vector with elements $\mu, \beta_1, \ldots, \beta_k$, $B = \oplus_{i=1}^{g} 1_{J_{i \cdot 1}}$ is a $n \times g$ design matrix, $\underline{U}$ is a $g \times 1$ vector of random effects, and $\underline{E}$ is a $n \times 1$ vector of random error terms. Under the distributional assumptions of (2.1), $\underline{Y}$ has a multivariate normal distribution with mean $X\underline{\beta}$ and covariance matrix $\sigma_A^2 BB' + \sigma_E^2 I_n$ where I_n is an $n \times n$ identity matrix. It is also assumed $s =$ rank$(X^*) -$ rank(X) and $r = n -$ rank(X^*) are both positive where $X^* = (X, BB')$ is the horizontal concatenation of matrices X and BB'.

3. CONFIDENCE INTERVALS ON σ_A^2

Two methods for constructing a confidence interval for σ_A^2 are presented in this section. El-Bassiouni(1) and Eubank, Seely, and Lee(2) proposed two mean squares that can be used to construct confidence intervals for σ_A^2. These mean squares were earlier proposed by Olsen, Seely, and Birkes(3). Let $W = FBB'F$ where $F = X^*(X^{*'}X^*)^+ X^{*'} - X(X'X)^+ X'$ and $^+$ denotes a Moore-Penrose inverse. Let $d_1, d_2, \ldots, d_m$ denote the distinct positive eigenvalues of W, and r_l be the multiplicity of d_l for $l = 1, \ldots, m$. Note that rank$(W) = \sum_l r_l = s$. Now define $Z = F\underline{Y}$ and $S_M^2 = Z' W^+ Z / s$. Finally, define $S_E^2 = \underline{Y}'[I_n - X^*(X^{*'}X^*)^+ X^{*'}]\underline{Y}/r$. The following results are verified in the previously cited references:

$$rS_E^2/\sigma_E^2 \sim \chi_r^2, \tag{3.1a}$$

$$sS_M^2/\sigma_A^2 \sim \chi_S^2 \text{ whenever } \sigma_E^2 = 0, \text{ and} \tag{3.1b}$$

$$S_M^2 \text{ and } S_E^2 \text{ areindependent} \tag{3.1c}$$

Further, under certain conditions $sS_M^2/E(S_M^2)$ has a limiting chi-squared distribution where $E(S_M^2) = \sigma_A^2 + \sigma_E^2/h$ and $h = s(\Sigma_l r_l / d_l)^{-1}$ is the harmonic mean of the eigenvalues. In particular,

$$S_M^2/E(S_M^2) \sim \Sigma_l c_l(\rho)\chi_{r_l}^2/s, \tag{3.2a}$$

$$c_l(\rho) = \frac{1-\rho+d_l\rho}{d_l(1-\rho)/h + d_l\rho}, \tag{3.2b}$$

$$\rho = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_E^2}, \tag{3.2c}$$

and the $\chi_{r_l}^2$ are mutually independent. If all $c_l(\rho) \rightarrow 1$, then $S_M^2/E(S_M^2)$ has a limiting χ_s^2/s distribution. This occurs, for example, when $\rho \rightarrow 1$, or if all $d_l \rightarrow d$.
A confidence interval for $\sigma_A^2 = E(S_M^2) - \sigma_E^2/h$ can be based on S_M^2 and S_E^2 using a method proposed by Ting, Burdick, Graybill, Jeyaratnam and Lu (4). Although this method requires two independent mean squares that have scaled chi-squared distributions, S_M^2 and S_E^2 closely mimic these conditions. The Ting et al. $1-2\alpha$ two-sided confidence interval for σ_A^2 is

$$\begin{aligned}&[S_M^2 - \frac{1}{h}S_E^2 - (G_1^2S_M^4 + \frac{1}{h^2}H_2^2S_E^4 + \frac{1}{h}G_{12}S_M^2S_E^2)^{\frac{1}{2}};\\ &S_M^2 - \frac{1}{h}S_E^2 + (H_1^2S_M^4 + \frac{1}{h^2}G_2^2S_E^4 + \frac{1}{h}H_{12}S_M^2S_E^{2)}{}^{\frac{1}{2}}]\end{aligned} \tag{3.3}$$

where $G_1 = 1 - 1/F_{1-\alpha:s,\infty}$, $H_2 = 1/F_{\alpha:r,\infty} - 1$, $G_{12} = [(F_{1-\alpha:s,r} - 1)^2 - G_1^2F_{1-\alpha:s,r}^2 - H_2^2]/F_{1-\alpha:s,r}$, $H_1 = 1/F_{\alpha:s,\infty} - 1$, $G_2 = 1 - 1/F_{1-\alpha:r,\infty}$, $H_{12} = [(1 - F_{\alpha:s,r})^2 - H_1^2F_{\alpha:s,r}^2 - G_2^2]/F_{\alpha:s,r}$, and $F_{\delta:df1,df2}$ is the F-percentile with degrees of freedom df_1 and df_2 with area δ to the left. Negative lower bounds are increased to zero. This method is referred to as TINGM method. Given the distributional assumptions of model (2.1), interval (3.3) is expected to perform well for large values of ρ

(σ_E^2 is small relative to σ_A^2).
Another important result provided by Olsen et al. (3) is that

$$Q_l = Z' E_l Z \sim (\sigma_E^2 + d_l \sigma_A^2)\chi_{r_l}^{\,2} \quad l=1,\ldots,m, \qquad (3.4)$$

where E_l is the orthogonal projection operator of the eigenspace of d_l. The variables S_E^2 , $Q_1,\ldots, Q_m$ are mutually independent and so it follows

$$\sum_{l=1}^{m} \frac{Q_l}{\sigma_E^2 + d_l \sigma_A^2} = U \qquad (3.5)$$

has a chi-squared distribution with s degrees of freedom. This result can be used to form a generalized confidence interval for σ_A^2 in the following manner. Tsui and Weerahandi(5) introduced the concept of generalized inference for testing hypotheses and constructing confidence intervals in situations where exact methods do not exist. Application of the method requires a generalized pivotal quantity that satisfies several conditions. We use (3.5) as our GPQ and define R as the solution for σ_A^2 in the non-linear equation

$$\sum_{l=1}^{m} \frac{q_l}{r s_E^2 / W + d_l \sigma_A^2} = U \qquad (3.6)$$

where q_l and s_E^2 are observed values of Q_l and S_E^2, respectively, and $W = rS_E^2 / \sigma_E^2$ and $U = \Sigma_l Q_l / (\sigma_E^2 + d_l \sigma_A^2)$ are jointly independent observable chi-squared random variables with degrees of freedom r and s, respectively. Note the distribution of R is completely determined by the joint distribution of W and U and is free of the parameters contained in model (2.1). An approximate $100(1-\alpha)\%$ confidence interval for σ_A^2 is now defined as

$$[R_{\alpha/2} \ ; \ R_{1-\alpha/2}] \qquad (3.7)$$

where $R_{\alpha/2}$ is the percentile $\alpha/2$ of the distribution of R and $R_{1-\alpha/2}$ is percentile $1-\alpha/2$. This method is referred to as GEN method. We used simulation with 10,000 iterations to model the distribution of R in the following manner. The observed values q_l and s_E^2 for a given data set are placed in equation (3.6). A set of random variables U and W are then simulated. If $U > W\sum_l q_l/(r s_E^2)$, then we set $R = 0$ since $\sigma_A^2 \geq 0$.

If $U\leq W\sum_{l} q_l/(rs_E^2)$ then the bisection method is used to solve the non-linear equation in (3.6).

4. SIMULATION STUDY AND CONCLUSIONS

The methods proposed in Section 3 are now compared using a simulation study. The criteria for analyzing the performance of the methods are their ability to maintain stated confidence coefficient, and the average length of two-sided confidence intervals. Although shorter average interval lengths are preferable, it is necessary that the methods first maintain the stated confidence coefficient. Five unbalanced patterns selected for the study are shown in Table 1.

Table 1. Unbalanced Patterns Used in Simulations

Pattern	g	n_i
1	3	5,10,15
2	3	1,1,100
3	6	1,1,1,1,1,100
4	6	1,1,2,3,50,
5	10	1,1,4,5,6,6,8,8,10,10

Recall $\rho=\sigma_A^2/(\sigma_A^2+\sigma_E^2)$. Without loss of generality, we set $\sigma_A^2=1-\sigma_E^2$ so that $\rho=\sigma_A^2$ and $1-\rho=\sigma_E^2$. The random variables A_i and E_{ij} were independently generated from normal populations with zero means and variances ρ and $1-\rho$, respectively, using the RANNOR routine of SAS. Responses for the Y_{ij} were constructed according to model (2.1) using a fixed set of values for X_{ij}. Values of ρ were varied from 0.001 to 0.999 in increments of 0.1. Simulations of 2,000 iterations were performed for each value of ρ in each pattern. Two-sided intervals were computed for each proposed method. Confidence coefficients were determined by counting the number of the intervals that contain σ_A^2. The average lengths of the two-sided confidence intervals were also calculated. Using the normal approximation to the binomial, if the true coefficient is 0.90, there is less than a 2.5 % chance that an estimated confidence coefficient based on 2,000 replications will be less than 0.887.

Using this criterion, the GEN method maintains the stated confidence coefficients across all values of ρ for all patterns. In contrast, TINGM provides a confidence coefficient less than the stated level in cases where ρ is small in the last four patterns. This is because S_M^2/σ_A^2 has an exact chi-squared distribution only when $\sigma_E^2=0\ (\rho=1)$. The average interval lengths are comparable for the two methods. Thus, in situations where ρ is thought to be small (say $\rho\leq 0.4$), TINGM is not recommended for extremely unbalanced datasets. In any other situation, either method can be recommended.

REFERENCES

1. El-Bassiouni, M. Y(1994). Short confidence intervals for variance components, Communications in statistics-Theory and Methods, 23(7), 1915-1933.

2. Eubank, L.; Seely, J.; Lee, Y(2001). Unweighted mean squares for the general two variance component mixed model,Graybill Conference, Ft. Collins, Co. June.

3. Olsen, A; Seely, J; Birkes, D(1976). Invariant quadratic unbiased estimation for two variance components, Annals of Statistics, 4, 878-890.

4. Ting, N; Burdick, R. K.; Graybill, F. A.; Jeyaratnam, S.; Lu, T.-F. C(1990). Confidence intervals on linear combinations of variance components, Journal of Statistical Computation, 35, 135-143.

5. Tsui, K.; Weerahandi, S(1989). Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters, Journal of the American Statistical Association, 84, 602-607.