The Korean Communications
in Statistics Vol. 9, No. 2, 2002
pp. 459-471

Comparison of Confidence Intervals on Variance Component In a Simple Linear Regression Model with Unbalanced Nested Error Structure

Dong Joon Park[1], Sun-Young Park[2] and Man-Ho Han[3]

Abstract

In applications using a linear regression model with nested error structure, one might be interested in making inferences concerning variance components. This article proposes approximate confidence intervals on the variance component of the primary level in a simple linear regression model with an unbalanced nested error structure. The intervals are compared using computer simulation and recommendations are provided for selecting an appropriate interval.

Keywords : mixed model; inference; least squares

1. Introduction

One of the reasons that statisticians use regression analysis is to find out linear associationship between a response variable and predictor variables. They often make inferences regarding parameters and a variance in a regression model. In applications using a linear regression model with nested error structure, one might be interested in making inferences concerning variance components in the model. The simple linear regression model with an unbalanced nested error structure includes two variance components; one in the primary level and the other in the secondary level of the model.

This article proposes three approximate confidence intervals on the variance component of the primary level in the model. The model is explained in Section 2. The distributional property of error sums of squares is obtained in Section 3. The confidence intervals on the variance component of the primary level in the model are proposed in Section 4. A simulation study is performed to compare the proposed intervals and recommendations are given for

1) Associate Professor, Division of Mathematical Sciences, Pukyong National University Nam-Gu Daeyeon 3-Dong, Pusan, Korea, 608-737 E-mail: djpark@pknu.ac.kr

2) 3) Graduate Student, Division of Mathematical Sciences, Pukyong National University Nam-Gu Daeyeon 3-Dong, Pusan, South Korea, 608-737

selecting appropriate intervals in Section 5. The proposed intervals are applied to a numerical example in Section 6.

2. A Simple Linear Regression Model With Unbalanced Nested Error Structure

The simple linear regression model with an unbalanced nested error structure is written as

$$Y_{ij} = \mu + \beta X_{ij} + A_i + E_{ij} \tag{2.1}$$

$$i = 1, \ldots, I; \; j = 1, \ldots, J_i$$

where Y_{ij} is the j th observation in the i th primary level, μ and β are unknown constants, X_{ij} is a fixed predictor variable, 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, $I > 2$, $J_i \geq 1$, and $J_i > 1$ for at least one value of i. A_i is an error term associated with the first-stage sampling unit and E_{ij} is an error term associated with the second-stage sampling unit. Model (2.1) is unbalanced since the number of observations in cells are not all equal. This error structure yields response variables that are correlated. That is,

$$Cov(Y_{ij}, Y_{i'j'}) = \begin{array}{ll} \sigma_A^2 + \sigma_E^2 & \text{if } i = i', \; j = j'; \\ \sigma_A^2 & \text{if } i = i', \; j \neq j'; \\ 0 & \text{if } i \neq i'. \end{array} \tag{2.2}$$

In order to form confidence intervals on linear functions of the variance components, an appropriate set of sums of squares is needed. One possible partitioning of model (2.1) is shown in Table 1.

TABLE 1. ANOVA for Model (2.1)

SV	DF	SS
Mean	1	$J_{.} \overline{Y}_{..}^{2}$
Covariate after mean	1	$\hat{\beta}_L^{2}(S_{uxxa} + S_{uxxe})$
Primary units adjusted for regression	$I - 1$	$R_{WB} + R_L$
Residual	$J_{.} - I - 1$	R_T
Total	$J_{.}$	$\Sigma_{i=1}^{I} \Sigma_{j=1}^{J_i} Y_{ij}^2$

The notation in Table 1 is defined as $J_{.} = \Sigma_{i=1}^{I} J_i$, $\overline{X}_{i.} = \Sigma_{j=1}^{J_i} X_{ij} / J_i$, $\overline{Y}_{i.} = \Sigma_{j=1}^{J_i} Y_{ij} / J_i$, $\overline{X}_{..} = \Sigma_{i=1}^{I} \overline{X}_{i.} J_i / J_{.}$, $\overline{Y}_{..} = \Sigma_{i=1}^{I} \overline{Y}_{i.} J_i / J_{.}$,
$S_{wxxa} = \Sigma_{i=1}^{I}(\overline{X}_{i.} - \overline{X}_{..})^2 J_i$, $S_{wxxe} = \Sigma_{i=1}^{I}\Sigma_{j=1}^{J_i}(X_{ij} - \overline{X}_{i.})^2$,
$S_{wxya} = \Sigma_{i=1}^{I}(\overline{X}_{i.} - \overline{X}_{..})(\overline{Y}_{i.} - \overline{Y}_{..})J_i$, $S_{wyya} = \Sigma_{i=1}^{I}(\overline{Y}_{i.} - \overline{Y}_{..})^2 J_i$,
$S_{wxye} = \Sigma_{i=1}^{I}\Sigma_{j=1}^{J_i}(X_{ij} - \overline{X}_{i.})(Y_{ij} - \overline{Y}_{i.})$, $S_{wyya} = \Sigma_{i=1}^{I}(\overline{Y}_{i.} - \overline{Y}_{..})^2 J_i$,
$S_{wyye} = \Sigma_{i=1}^{I}\Sigma_{j=1}^{J_i}(Y_{ij} - \overline{Y}_{i.})^2$, $\hat{\beta}_{WB} = S_{wxya}/S_{wxxa}$,
$\hat{\beta}_L = (S_{wxya} + S_{wxye})/(S_{wxxa} + S_{wxxe})$, $\hat{\beta}_T = S_{wxye}/S_{wxxe}$, $R_{WB} = S_{wyya} - \hat{\beta}_{WB}^2 S_{wxxa}$,
$R_L = \hat{\beta}_{WB}^2 S_{wxxa} + \hat{\beta}_T^2 S_{wxxe} - \hat{\beta}_L^2 (S_{wxxa} + S_{wxxe})$, and $R_T = S_{wyye} - \hat{\beta}_T^2 S_{wxxe}$.
Model (2.1) is written in matrix notation,

$$\boldsymbol{y} = \boldsymbol{X}\underline{\boldsymbol{\alpha}} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e} \tag{2.3}$$

where $\boldsymbol{y}$ is a $J_{.} \times 1$ vector of observations, $\boldsymbol{X}$ is a $J_{.} \times 2$ matrix of known values with a column of 1's in the first column and a column of X_{ij}'s in the second column, $\underline{\boldsymbol{\alpha}}$ is a 2×1 vector of parameters with μ and β as elements, $\boldsymbol{Z}$ is a $J_{.} \times I$ design matrix with 0's and 1's, i.e. $\boldsymbol{Z} = \oplus_{i=1}^{I} \boldsymbol{1}_{J_i \times 1}$, $\boldsymbol{u}$ is an $I \times 1$ vector of random effects, and $\boldsymbol{e}$ is a $J_{.} \times 1$ vector of random error terms. By the assumptions in (2.1) the response variables have a multivariate normal distribution

$$\boldsymbol{y} \sim N(\boldsymbol{X}\underline{\boldsymbol{\alpha}}, \ \boldsymbol{V}) \tag{2.4}$$

where $\boldsymbol{V} = \sigma_A^2 \boldsymbol{Z}\boldsymbol{Z}' + \sigma_E^2 \boldsymbol{D}_{J_{.}}$ and $\boldsymbol{D}_{J_{.}}$ is a $J_{.} \times J_{.}$ identity matrix. In order to define unweighted sums of squares, the vector of means of response variables of primary level and associated variance matrix are needed. These are defined in matrix notation as $\boldsymbol{M}\boldsymbol{y} = [\overline{Y}_{1.}, \overline{Y}_{2.}, \ldots, \overline{Y}_{I.}]' = \boldsymbol{y}_M$ where

$$\boldsymbol{M} = \oplus_{i=1}^{I}[J_i^{-1} \boldsymbol{1}'_{J_i \times 1}]$$

and $\overline{Y}_{i.}$ is the mean of response variables of the ith primary level. The expectation and variance of vector of means of response variables of the primary level are $E(\boldsymbol{y}_M) = \boldsymbol{X}_M \underline{\boldsymbol{\alpha}}$ where $\boldsymbol{X}_M = \boldsymbol{M}\boldsymbol{X}$ and $V(\boldsymbol{y}_M) = \sigma_A^2 \boldsymbol{D}_I + \sigma_E^2 \boldsymbol{M}\boldsymbol{M}'$ since $\boldsymbol{M}\boldsymbol{M}' = diag[J_i^{-1}]$ and $\boldsymbol{M}\boldsymbol{Z} = \boldsymbol{D}_I$ where $\boldsymbol{D}_I$ is an $I \times I$ identity matrix. Thus, the means of response variables of primary level have a multivariate normal distribution

$$\boldsymbol{y}_M \sim N(\boldsymbol{X}_M \underline{\boldsymbol{\alpha}}, \ \boldsymbol{V}_M) \tag{2.5}$$

where $\boldsymbol{V}_M = \sigma_A^2 \boldsymbol{D}_I + \sigma_E^2 \boldsymbol{M} \boldsymbol{M}'$.

3. Distributional Property of Error Sums of Squares

In this section we report distributional results used to derive confidence intervals. Four regression coefficient estimators are considered. Consider weighted between regression coefficient estimator $\hat{\beta}_{WB}$ that is obtained from least squares regression of $\overline{Y}_{i.}$ on $\overline{X}_{i.}$ with weight J_i for each primary level i and $\hat{\beta}_{WB}$ is written as $\hat{\beta}_{WB} = S_{wxya} / S_{wxxa}$. The weighted between regression coefficient estimator is the second element of the vector $(\boldsymbol{X}_M' \boldsymbol{W} \boldsymbol{X}_M)^{-1} \boldsymbol{X}_M' \boldsymbol{W} \boldsymbol{y}_M$ where $\boldsymbol{W} = diag[J_i]$. The error sum of squares R_{WB} associated with this regression model is $R_{WB} = S_{wyya} - \hat{\beta}_{WB}^2 S_{wxxa} = \boldsymbol{y}'_M \boldsymbol{A}_W \boldsymbol{y}_M$ where $\boldsymbol{A}_W = \boldsymbol{W} - \boldsymbol{W} \boldsymbol{X}_M (\boldsymbol{X}'_M \boldsymbol{W} \boldsymbol{X}_M)^{-1} \boldsymbol{X}'_M \boldsymbol{W}$.

Unweighted between regression coefficient estimator considering primary level's means and their unweighted mean is used as an alternative of between regression coefficient estimator. Unweighted between regression coefficient estimator $\hat{\beta}_{UB}$ is obtained from the least squares regression of $\overline{Y}_{i.}$ on $\overline{X}_{i.}$ and $\hat{\beta}_{UB}$ is written as $\hat{\beta}_{UB} = S_{uxya} / S_{uxxa}$ where $S_{uxya} = \Sigma_{i=1}^{I} (\overline{X}_{i.} - \overline{\overline{X}}_{..})(\overline{Y}_{i.} - \overline{\overline{Y}}_{..})$ $S_{uxxa} = \Sigma_{i=1}^{I} (\overline{X}_{i.} - \overline{\overline{X}}_{..})^2$, $\overline{\overline{X}}_{..} = \Sigma_{i=1}^{I} \overline{X}_{i.} / I$ and $\overline{\overline{Y}}_{..} = \Sigma_{i=1}^{I} \overline{Y}_{i.} / I$. The unweighted between regression coefficient estimator is the second element of the vector $(\boldsymbol{X}_M' \boldsymbol{X}_M)^{-1} \boldsymbol{X}_M' \boldsymbol{y}_M$. The error sum of squares R_{UB} associated with this regression model is

$R_{UB} = S_{uyya} - \hat{\beta}_{UB}^2 S_{uxxa} = \boldsymbol{y}'_M \boldsymbol{A}_U \boldsymbol{y}_M$ where $\boldsymbol{A}_U = \boldsymbol{D}_I - \boldsymbol{X}_M (\boldsymbol{X}'_M \boldsymbol{X}_M)^{-1} \boldsymbol{X}'_M$.

The within regression coefficient estimator $\hat{\beta}_T = S_{wxye} / S_{wxxe}$ is obtained from the least squares regression of Y_{ij} on X_{ij} and the grouping variables. The point estimator $\hat{\beta}_T$ is the second element of the vector $(\boldsymbol{X}^{*'} \boldsymbol{X}^*)^- \boldsymbol{X}^{*'} \boldsymbol{y}$ where $\boldsymbol{X}^* = [\boldsymbol{X}\ \boldsymbol{Z}]$ and $(\boldsymbol{X}^{*'} \boldsymbol{X}^*)^-$ is a generalized inverse of $\boldsymbol{X}^{*'} \boldsymbol{X}^*$. The error sum of squares R_T associated with this regression model is $R_T = S_{wyye} - \hat{\beta}_T^2 S_{wxxe} = \boldsymbol{y}' \boldsymbol{T} \boldsymbol{y}$ where $\boldsymbol{T} = \boldsymbol{D}_{J.} - \boldsymbol{X}^* (\boldsymbol{X}^{*'} \boldsymbol{X}^*)^- \boldsymbol{X}^{*'}$.

Finally, the total regression coefficient estimator $\hat{\beta}_L = (S_{wxya} + S_{wxye})/(S_{wxxa} + S_{wxxe})$ is obtained from the least squares regression of Y_{ij} on X_{ij}. The point estimator $\hat{\beta}_L$ is the second element of the vector $(\boldsymbol{X}' \boldsymbol{X})^{-1} \boldsymbol{X}' \boldsymbol{y}$. The error sum of squares R_L associated with this regression model is $R_L = (S_{wyya} + S_{wyye}) - \hat{\beta}_L^{\,2}(S_{wxxa} + S_{wxxe}) - R_{WB} - R_T = \boldsymbol{y}'(\boldsymbol{L} - \boldsymbol{M}'\boldsymbol{A}_W\boldsymbol{M} - \boldsymbol{T})\boldsymbol{y}$ where $\boldsymbol{L} = \boldsymbol{D}_{J.} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$.

Theorem 1. R_T/σ_E^2 is a chi-squared random variable with $J_. - I - 1$ degrees of freedom.

Proof. Notice that $\boldsymbol{T}$ is idempotent. It can be shown that $\boldsymbol{X}^*(\boldsymbol{X}^{*'}\boldsymbol{X}^*)^-\boldsymbol{X}^{*'}\boldsymbol{X} = \boldsymbol{X}$ and $\boldsymbol{X}^*(\boldsymbol{X}^{*'}\boldsymbol{X}^*)^-\boldsymbol{X}^{*'}\boldsymbol{Z} = \boldsymbol{Z}$ by Theorem 7.1 in Searle (1987, p. 218). Therefore, as may be easily verified, $\boldsymbol{T}\boldsymbol{X} = \boldsymbol{0}$ and $\boldsymbol{T}\boldsymbol{Z} = \boldsymbol{0}$. It follows that $E(R_T) = E(\boldsymbol{y}'\boldsymbol{T}\boldsymbol{y}) = tr(\boldsymbol{T}\boldsymbol{V}) + \underline{\boldsymbol{\alpha}}'\boldsymbol{X}'\boldsymbol{T}\boldsymbol{X}\underline{\boldsymbol{\alpha}} = (J_. - I - 1)\sigma_E^2$. The distribution of R_T is determined by writing $R_T/\sigma_E^2 = \boldsymbol{y}'(\boldsymbol{T}/\sigma_E^2)\boldsymbol{y}$ and noting $(\boldsymbol{T}/\sigma_E^2)\boldsymbol{V} = \boldsymbol{T}(\sigma_A^2\boldsymbol{Z}\boldsymbol{Z}' + \sigma_E^2\boldsymbol{D}_{J.})/\sigma_E^2 = \boldsymbol{T}$. By Theorem 2 in Searle (1971, p. 57) R_T/σ_E^2 is a chi-squared random variable with $J_. - I - 1$ degrees of freedom.

Theorem 2. If $\sigma_A^2 = 0$, then R_{WB}/σ_E^2 is a chi-squared random variable with $I - 2$ degrees of freedom.

Proof. Notice that $\boldsymbol{A}_W\boldsymbol{V}_M = \sigma_A^2\boldsymbol{W} - \sigma_A^2\boldsymbol{W}\boldsymbol{X}_M(\boldsymbol{X}_M'\boldsymbol{W}\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M'\boldsymbol{W} + \sigma_E^2\boldsymbol{D}_I - \sigma_E^2\boldsymbol{W}\boldsymbol{X}_M(\boldsymbol{X}_M'\boldsymbol{W}\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M'$ and $tr(\boldsymbol{W}\boldsymbol{X}_M(\boldsymbol{X}_M'\boldsymbol{W}\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M'\boldsymbol{W}) = k_1$. It follows that $E(R_{WB}) = E(\boldsymbol{y}_M'\boldsymbol{A}_W\boldsymbol{y}_M) = tr(\boldsymbol{A}_W\boldsymbol{V}_M) + \underline{\boldsymbol{\alpha}}'\boldsymbol{X}_M'\boldsymbol{A}_W\boldsymbol{X}_M\underline{\boldsymbol{\alpha}} = (J_. - k_1)\sigma_A^2 + (I - 2)\sigma_E^2$ since $\boldsymbol{A}_W\boldsymbol{X}_M = \boldsymbol{0}$. The distribution of R_{WB} is determined by writing $R_{WB}/\sigma_E^2 = \boldsymbol{y}_M'(\boldsymbol{A}_W/\sigma_E^2)\boldsymbol{y}_M$ and noting $(\boldsymbol{A}_W/\sigma_E^2)\boldsymbol{V}_M = (\sigma_A^2/\sigma_E^2)\boldsymbol{A}_W + \boldsymbol{A}_W\boldsymbol{M}\boldsymbol{M}' = (\sigma_A^2/\sigma_E^2)\boldsymbol{A}_W + \boldsymbol{D}_I - \boldsymbol{W}\boldsymbol{X}_M(\boldsymbol{X}_M'\boldsymbol{W}\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M'$ since $\boldsymbol{W}\boldsymbol{M}\boldsymbol{M}' = \boldsymbol{D}_I$. Note that $\boldsymbol{D}_I - \boldsymbol{W}\boldsymbol{X}_M(\boldsymbol{X}_M'\boldsymbol{W}\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M'$ is idempotent. It follows that R_{WB}/σ_E^2 is a chi-squared random variable with $I - 2$ degrees of freedom if $\sigma_A^2 = 0$.

Theorem 3. If $\sigma_E^2 = 0$, then R_{UB}/σ_A^2 is a chi-squared random variable with $I - 2$ degrees of freedom.

Proof. Notice that $\boldsymbol{A}_U\boldsymbol{V}_M = \sigma_A^2(\boldsymbol{D}_I - \boldsymbol{X}_M(\boldsymbol{X}_M'\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M') + \sigma_E^2(\boldsymbol{M}\boldsymbol{M}' - \boldsymbol{X}_M(\boldsymbol{X}_M'\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M'\boldsymbol{M}\boldsymbol{M}')$, $tr(\boldsymbol{M}\boldsymbol{M}') = \Sigma_i(1/J_i)$, $tr(\boldsymbol{X}_M(\boldsymbol{X}_M'\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M'\boldsymbol{M}\boldsymbol{M}') = k_2$ where $k_2 = \Sigma_{i=1}^I\Sigma_{k=1}^I(\overline{X}_{i.} - \overline{X}_{k.})^2/J_k/(IS_{uxxa})$ and $\boldsymbol{A}_U\boldsymbol{X}_M = \boldsymbol{0}$. It follows that $E(R_{UB}) = E(\boldsymbol{y}_M'\boldsymbol{A}_U\boldsymbol{y}_M) = tr(\boldsymbol{A}_U\boldsymbol{V}_M) + \underline{\boldsymbol{\alpha}}'\boldsymbol{X}_M'\boldsymbol{A}_U\boldsymbol{X}_M\underline{\boldsymbol{\alpha}}$

$= (I - 2)\sigma_A^2 + (\Sigma_i(1/J_i) - k_2)\sigma_E^2$. The distribution of R_{UB} is determined by writing $R_{UB}/\sigma_A^2 = \boldsymbol{y}_M'(\boldsymbol{A}_U/\sigma_A^2)\boldsymbol{y}_M$ and noting $(\boldsymbol{A}_U/\sigma_A^2)\boldsymbol{V}_M = \boldsymbol{A}_U + (\sigma_E^2/\sigma_A^2)\boldsymbol{A}_U\boldsymbol{M}\boldsymbol{M}'$. Note that $\boldsymbol{A}_U$ is idempotent. Thus R_{UB}/σ_A^2 is a chi-squared random variable with $I - 2$ degrees of freedom if $\sigma_E^2 = 0$.

Theorem 4. R_{WB}/σ_E^2 and R_T/σ_E^2 are independent and R_{UB}/σ_A^2 and R_T/σ_E^2 are independent.

Proof. Notice that $\boldsymbol{M}'\boldsymbol{A}_W\boldsymbol{M}(\sigma_A^2\boldsymbol{Z}\boldsymbol{Z}' + \sigma_E^2\boldsymbol{D}_{J.})\boldsymbol{T} = \sigma_A^2\boldsymbol{M}'\boldsymbol{A}_W\boldsymbol{M}\boldsymbol{Z}\boldsymbol{Z}'\boldsymbol{T} + \sigma_E^2\boldsymbol{M}'\boldsymbol{A}_W\boldsymbol{M}\boldsymbol{T} = \boldsymbol{0}$ from the results in Theorem 2. Accordingly R_{WB}/σ_E^2 and R_T/σ_E^2 are independent. Note that $\boldsymbol{M}'\boldsymbol{A}_U\boldsymbol{M}(\sigma_A^2\boldsymbol{Z}\boldsymbol{Z}' + \sigma_E^2\boldsymbol{D}_{J.})\boldsymbol{T} = \sigma_A^2\boldsymbol{M}'\boldsymbol{A}_U\boldsymbol{M}\boldsymbol{Z}\boldsymbol{Z}'\boldsymbol{T} + \sigma_E^2\boldsymbol{M}'\boldsymbol{A}_U\boldsymbol{M}\boldsymbol{T}$ $= \boldsymbol{0}$ using $\boldsymbol{M}\boldsymbol{T} = \boldsymbol{M}\boldsymbol{M}'\boldsymbol{Z}'\boldsymbol{T} = \boldsymbol{0}$ since $\boldsymbol{M} = \boldsymbol{M}\boldsymbol{M}'\boldsymbol{Z}'$ and $\boldsymbol{Z}'\boldsymbol{T} = \boldsymbol{0}$.
Thus R_{UB}/σ_A^2 and R_T/σ_E^2 are independent.

Olsen et al.(1976), Thomas and Hultquist(1978), and El-Bassiouni(1994) used spectal decomposition method to obtain following statistics.
They proposed a statistic $SSM = \boldsymbol{U}'\boldsymbol{U}$ which is asymptotically chi-squared distributed. In particular, $\boldsymbol{U}'\boldsymbol{U}/(\sigma_A^2 + \sigma_E^2/\lambda_H) \rightarrow \chi^2_{(I-1)}$ as $\sigma_E^2 \rightarrow 0$ where $\boldsymbol{U} = \boldsymbol{C}^+\boldsymbol{Z}'(\boldsymbol{D}_{J.} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{y}$, $\boldsymbol{C}^+$ is the Moore-Penrose inverse of $\boldsymbol{C}$, $\boldsymbol{C} = \boldsymbol{Z}'(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{Z}$, λ_H is the harmonic mean of positive eigenvalues, λ_i, of $\boldsymbol{C}$, $\lambda_H = \Sigma_i r_i/(\Sigma_{i=1}^I r_i/\lambda_i)$, and r_i is the multiplicity of positive eigenvalue λ_i. Thus, $E(\boldsymbol{U}'\boldsymbol{U}) \doteq (I - 1)(\sigma_A^2 + \sigma_E^2/\lambda_H)$. It was also shown that $\boldsymbol{U}'\boldsymbol{U}/(\sigma_A^2 + \sigma_E^2/\lambda_H)$ and R_T/σ_E^2 are independent.

If the covariate values within each group are same, this proposed statistics becomes the error sum of squares associated with unweighted between regression coefficient and the total regression coefficient estimator reduces to the weighted between regression coefficient estimator. That is, if $X_{ij} = X_i$ for all j, then $SSM = R_{UB}$ and $\hat{\beta}_L = \hat{\beta}_{WB}$. If group means of the covariate values are all same, i.e., $\overline{X}_{i.} = \overline{X}_{..} = \overline{\overline{X}}_{..}$ for all i, then $\boldsymbol{X}_M$ is linearly dependent and $\hat{\beta}_{WB}$ and $\hat{\beta}_{UB}$ are not defined.

4. Confidence Intervals on σ_A^2

The expected mean squares are summarized using the distributional property of error sums of squares.

$$\begin{aligned} E(S^2_{WB}) &= c_1\sigma_A^2 + \sigma_E^2 = \theta_{WB}, && (4.1a)\\ E(S^2_{UB}) &= \sigma_A^2 + c_2\sigma_E^2 = \theta_{UB}, \quad \text{and} && (4.1b)\\ E(S^2_T) &= \sigma_E^2 = \theta_T, && (4.1c) \end{aligned}$$

where $S^2_{WB} = R_{WB}/(I-2)$, $S^2_{UB} = R_{UB}/(I-2)$, $S^2_T = R_T/(J_. - I - 1)$, $c_1 = (J_. - k_1)/(I-2)$, and $c_2 = (\Sigma_i(1/J_i) - k_2)/(I-2)$. The mean square errors, S^2_{WB} and S^2_{UB} are independent of S^2_T and they are exactly chi-squared distributed depending on cases where $\sigma_A^2 = 0$ and $\sigma_E^2 = 0$.

In the case where $\sigma_A^2 \to 0$, S^2_{WB} and S^2_T should be used to construct confidence intervals on σ_A^2. The variance component σ_A^2 can be represented by functions of expected mean squares in (4.1a) and (4.1c), $\sigma_A^2 = (\theta_{WB} - \theta_T)/c_1$. An approximate confidence interval on σ_A^2 can be constructed using the method of Ting et al.(1990). In particular, the $1-2\alpha$ two-sided confidence interval for this form of σ_A^2 is

$$\begin{aligned} \frac{1}{c_1}[&(S^2_{WB} - S^2_T) - (G_1^2 S^4_{WB} + G_2^2 S^4_T + G_{12}S^2_{WB}S^2_T)^{\frac{1}{2}};\\ &(S^2_{WB} - S^2_T) + (H_1^2 S^4_{WB} + H_2^2 S^4_T + H_{12}S^2_{WB}S^2_T)^{\frac{1}{2}}] \end{aligned} \qquad (4.2)$$

where $F_1 = F_{(\alpha:I-2,J_.-I-1)}$, $F_2 = F_{(1-\alpha:I-2,J_.-I-1)}$, $G_1 = 1 - 1/F_{(\alpha:I-2,\infty)}$, $G_2 = 1/F_{(1-\alpha:J_.-I-1,\infty)} - 1$,
$G_{12} = [(F_1-1)^2 - G_1^2F_1^2 - G_2^2]/F_1$, $H_1 = 1/F_{(1-\alpha:I-2,\infty)} - 1$,
$H_2 = 1 - 1/F_{(\alpha:J_.-I-1,\infty)}$, $H_{12} = [(1-F_2)^2 - H_1^2F_2^2 - H_2^2]/F_2$, and $F_{(\delta:n_1,n_2)}$ is the F-value for n_1 and n_2 degrees of freedom with δ area to the right. Since $\sigma_A^2 > 0$, any negative bound is defined to be zero. Interval (4.2) is referred to as TINGW method.

Another approach is adapting generalized p-values method proposed by Khuri et al.(1998) to construct an approximate confidence interval on σ_A^2. It was shown in Section 3 that $(I-1)S^2_M/(\sigma_A^2 + \sigma_E^2/\lambda_H)$ is chi-squared distributed with $(I-1)$ degrees of freedom as σ_E^2 approaches zero, $(J_. - I - 1)S_T^2/\sigma_E^2 \sim \chi^2_{(J_. - I - 1)}$, and they are independent where $S_M^2 = SSM/(I - 1)$. Thus, using this property, the estimators of σ_E^2 are obtained by $(J_. - I - 1)s_T^2 / U_1$ where s_T^2 is an observed value of S_T^2 and U_1 has a chi-squared distribution with $(J_. - I - 1)$ degrees of freedom. The estimators of σ_{AE}^2 are obtained by $(I - 1)s_M^2 / U_2$ where $\sigma_{AE}^2 = \sigma_A^2 + \sigma_E^2 / \lambda_H$, s_M^2 is an observed value of S_M^2, and U_2 has a chi-squared distribution with $(I - 1)$ degrees of freedom. Thus, a generalized pivotal quantity σ_A^2 can be represented as $\sigma_A^2 \doteq (I - 1)s_M^2 / U_2 - (J_. - I - 1)s_T^2 / [\lambda_H U_1]$. Accordingly, an approximate $1 - 2\alpha$ two-sided confidence interval for this form of σ_A^2 is

$$[C_\alpha \ ; \ C_{1-\alpha}] \tag{4.3}$$

where C_α is the αth percentile of the distribution constructed by the generalized pivotal quantity. Interval (4.3) is referred to as GPQ method.

When σ_E^2 approaches zero, S^2_{UB} and S_T^2 can be used and σ_A^2 is represented $\sigma_A^2 = \theta_{UB} - c_2\theta_T$ from (4.1b) and (4.1c). The Ting et al. $1 - 2\alpha$ two-sided confidence interval for this form of σ_A^2 is

$$\begin{aligned}[S^2_{UB} &- c_2S_T^2 - (G_1^2S^4_{UB} + c_2^2G_2^2S_T^4 + c_2G_{12}S^2_{UB}S_T^2)^{\frac{1}{2}}; \\ S^2_{UB} &- c_2S_T^2 + (H_1^2S^4_{UB} + c_2^2H_2^2S_T^4 + c_2H_{12}S^2_{UB}S_T^2)^{\frac{1}{2}}].\end{aligned} \tag{4.4}$$

Interval (4.4) is referred to as TINGU method. Table 2 summarizes the methods proposed in this section.

If $I = 3$, then $c_2 = 1/c_1$ and $c_2\boldsymbol{A}_W = \boldsymbol{A}_U$. Thus $S^2_{WB}/c_1 = c_2S^2_{WB} = S^2_{UB}$ and TINGW and TINGU methods are same.

5. Simulation Study

The methods proposed in Section 4 are now compared using simulation study. The criteria for analyzing the performance of the methods are; 1) their ability to maintain stated confidence coefficient, and 2) 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. Four unbalanced patterns were selected for simulation study and are shown in Table 3.

TABLE 3. Unbalanced Patterns Used in Simulation

Pattern	I	J_i
1	3	3 5 10
2	5	1 3 5 7 10
3	7	1 2 4 6 8 10
4	10	1 1 1 5 5 5 5 10 10 10

Let $\rho = \sigma_A^2 / (\sigma_A^2 + \sigma_E^2)$. Without loss of generality $\sigma_A^2 = 1 - \sigma_E^2$ so that $\rho = \sigma_A^2$ and $1 - \rho = \sigma_E^2$. A_i and E_{ij} are independently generated from normal populations with zero means and variance ρ and $1 - \rho$, respectively, using RANNOR routines of SAS. Values of μ and β are respectively varied from -3 to 3 in increments of 1 so that 49 different combinations of μ and β are used. Any fixed values of X_{ij}'s are given. Then Y_{ij}'s are calculated according to model (2.1) and R_{WB}, R_T, SSM, and R_{UB} are computed as shown in Section 3. Simulated values for S^2_{WB}, S^2_T, S^2_M, and S^2_{UB} are substituted into appropriate formula and the intervals are computed. Values of ρ are varied from 0.001 to 0.999 in increments of 0.1. Each value of ρ is simulated 2000 times for each pattern. Two-sided intervals are computed based on equal tailed F-values.

Confidence coefficients are determined by counting the number of the intervals that contain σ_A^2. 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 2000 replications will be less than 0.8866. The average lengths of the two-sided confidence intervals are also calculated.

Table 4 and 5 present the results of the simulation for stated 90% confidence intervals on σ_A^2. The numbers in the body of Table 4 and 5 respectively report range of simulated confidence coefficients and average interval lengths and minimum and maximum values for the range as ρ ranges from 0.001 to 0.999. Different combinations of μ and β do not change the trend of simulation results and the change of minimum values of stated confidence coefficients is at most 0.012.

TABLE 4. 90% Range of Simulated Confidence Coefficients

Pattern	1			2		
ρ	TINGW	GPQ	TINGU	TINGW	GPQ	TINGU
0.001	0.9005	0.9035	0.9005	0.8905	0.894	0.87
0.1	0.908	0.9045	0.908	0.8925	0.894	0.883
0.2	0.898	0.9085	0.898	0.8875	0.8955	0.884
0.3	0.893	0.904	0.893	0.89	0.896	0.898
0.4	0.9095	0.905	0.9095	0.8845	0.8985	0.907
0.5	0.9015	0.905	0.9015	0.898	0.8985	0.905
0.6	0.897	0.905	0.897	0.88	0.897	0.9045
0.7	0.899	0.9065	0.899	0.8655	0.896	0.8905
0.8	0.898	0.907	0.898	0.87	0.897	0.893
0.9	0.8975	0.905	0.8975	0.8635	0.896	0.8935
0.999	0.902	0.905	0.902	0.872	0.896	0.8925
MAX	0.9095	0.9085	0.9095	0.898	0.8985	0.907
MIN	0.893	0.9035	0.893	0.8635	0.894	0.87
Pattern	3			4		
0.001	0.9	0.908	0.854	0.897	0.899	0.8135
0.1	0.8965	0.9095	0.866	0.901	0.8995	0.853
0.2	0.8955	0.907	0.888	0.8845	0.8985	0.868
0.3	0.8865	0.906	0.8955	0.8885	0.9015	0.8765
0.4	0.863	0.9065	0.883	0.869	0.905	0.8915
0.5	0.871	0.904	0.884	0.862	0.902	0.884
0.6	0.865	0.905	0.882	0.8715	0.903	0.891
0.7	0.8645	0.9055	0.895	0.862	0.9025	0.913
0.8	0.8735	0.901	0.907	0.857	0.902	0.898
0.9	0.858	0.9005	0.895	0.841	0.902	0.899
0.999	0.8685	0.8995	0.9045	0.856	0.9	0.904
MAX	0.9	0.9095	0.907	0.901	0.905	0.913
MIN	0.858	0.8995	0.854	0.841	0.8985	0.8135

Simulation results are consistent with our study since TINGW method improves as ρ approaches zero while TINGU method performs well as ρ is closed to one across all four patterns. Three methods generally maintain stated confidence coefficients across all values of ρ for patterns 1. However, only GPQ method keeps the stated confidence coefficients for all ρ values of four patterns. The average interval lengths of three methods generate wider intervals as ρ increases for all four patterns. For smaller ρ value, say $\rho \leq 0.1$, in pattern 3 and 4, TINGW method has shortest interval lengths. For other values of ρ in four patterns, GPQ method has shortest interval length.

TABLE 5. 90% Range of Average Interval Lengths

Pattern	1			2		
ρ	TINGW	GPQ	TINGU	TINGW	GPQ	TINGU
0.001	44.670385	4.7037761	44.670385	1.6963909	1.7565652	2.4211901
0.1	59.306305	6.203999	59.306305	2.3232264	2.1452094	2.9775824
0.2	88.336569	7.7152441	88.336569	2.9723129	2.5340456	3.5968303
0.3	107.57395	9.2208541	107.57395	3.7117547	2.9165519	4.2278956
0.4	120.93906	10.721291	120.93906	4.1990992	3.2901384	4.6446833
0.5	142.92995	12.216601	142.92995	4.8447865	3.6554658	5.2186535
0.6	168.20474	13.706938	168.20474	5.6089767	4.0127826	5.7815406
0.7	185.53583	15.192167	185.53583	6.0088214	4.3633191	6.2783312
0.8	216.90426	16.673825	216.90426	7.0531375	4.7080909	7.0924004
0.9	245.96681	18.151819	245.96681	7.5738493	5.0491171	7.6091444
0.999	246.10563	19.612088	246.10563	8.4534132	5.3857702	8.5034603
MAX	246.10563	19.612088	246.10563	8.4534132	5.3857702	8.5034603
MIN	44.670385	4.7037761	44.670385	1.6963909	1.7565652	2.4211901
Pattern	3			4		
0.001	0.8841862	1.2104748	1.4681376	0.3847124	0.7396307	0.7652631
0.1	1.3228924	1.5056313	1.8549788	0.6214983	0.9254198	0.9536777
0.2	1.8482663	1.8000139	2.327806	0.8308489	1.1062396	1.1387063
0.3	2.251486	2.0867147	2.6258684	1.0543185	1.2748048	1.3411049
0.4	2.7262211	2.3635822	3.127743	1.2549925	1.4296433	1.5096349
0.5	3.1438896	2.6303032	3.4974074	1.4559445	1.5735124	1.6897315
0.6	3.568139	2.8878889	3.7790143	1.5924883	1.7080275	1.7819612
0.7	4.0057759	3.1385151	4.1079702	1.806329	1.8377487	1.9402982
0.8	4.5275753	3.3852496	4.5825573	1.9948288	1.9655106	2.0982791
0.9	4.7874478	3.6297314	4.8973924	2.2045574	2.0937172	2.2359168
0.999	5.1810588	3.8715458	5.1928903	2.4126914	2.2214677	2.3882471
MAX	5.1810588	3.8715458	5.1928903	2.4126914	2.2214677	2.3882471
MIN	0.8841862	1.2104748	1.4681376	0.3847124	0.7396307	0.7652631

In summary, if $\rho \leq 0.1$ in pattern 3 and 4, TINGW method is recommended. For other values of ρ in four patterns, GPQ method is recommended because it keeps the stated confidence coefficient and generates shortest average interval lengths.

6. A Numerical Example

The results of the simulation study are applied to a data set. Scheffe (1959, p. 216) wrote a data set of 94 observations for seven types of starch film and the data set was reproduced with permission of the author and publisher from Industrial Statistics by Freeman(1942). The dependent variable in the data set is the breaking strength in grams and the independent variable is the thickness in 10^{-4} inch from tests of starch film. The data set was constructed by selecting three types of starch, Potato, Canna, and Wheat. Three observations are selected from Potato, five from Canna, and ten from Wheat. This data set has the form of pattern 1 in Table 3 and is used to fit the simple linear regression model of the breaking strength on the thickness of starch film assuming an unbalanced nested error structure.

The selected data set was listed in Table 6. In order to apply the methods proposed in Section 4 to the data set a SAS code was programmed and 90% confidence intervals on σ_A^2 were calculated. The resulting intervals were given in Table 7.

TABLE 6. The Data Set Used For The Example

Type	Obs.	1	2	3	4	5	6	7	8	9	10
Potato	Y	983.3	958.8	747.8							
	X	13.0	13.3	10.7							
Canna	Y	791.7	610.0	710.0	940.7	990.0					
	X	7.7	6.3	8.6	11.8	12.4					
Wheat	Y	263.7	130.8	382.9	302.5	213.3	132.1	292.0	315.5	262.4	314.4
	X	5.0	3.5	4.7	4.3	3.8	3.0	4.2	4.5	4.3	4.1

From SAS output the estimators $\hat{\sigma}_A^2$ and $\hat{\sigma}_E^2$ are computed as 8479.97 and 3063.89, respectively. Therefore, the estimate of the ratio of variance in primary unit to total variance $\hat{\rho}$ is 0.7345. GPQ should be used because it keeps the stated confidence level and generates the shortest interval length among three methods in pattern 1 of Tables 4 and 5. The calculated interval lengths in Table 7 are consistent with the simulation results in Table 5.

TABLE 7. 90% Confidence Intervals On σ_A^2

Methods	Lower Bound	Upper Bound	Length
TINGW(TINGU)	2702.9	3359262.4	3356559.5
GPQ	915.6	216887.0	215971.4

References

[1] El-Bassiouni, M. Y.(1994), Short Confidence Intervals for Variance Components, *Communications in Statistics-Theory and Methods*, vol. 23 no. 7, 1915-1933.

[2] Freeman, H. A.(1942), *Industrial Statistics*, John Wiley & Sons, New York.

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

[4] Olsen, A., Seely, J., and Birkes, D.(1976), Invariant Quadratic Unbiased Estimation for Two Variance Components, *The Annals of Statistics*, vol. 4, no. 5, 878-890.

[5] Scheffe, Henry.(1959), *The Analysis of Variance*, John Wiley & Sons, New York.

[6] Searle, S. R.(1971), *Linear Models*, John Wiley & Sons, New York.

[7] Searle, S. R.(1987), *Linear Models for Unbalanced Data*, John Wiley & Sons, New York.

[8] Thomas, J. D. and Hultquist, R. A.(1978), Interval Estimation for The Unbalanced Case of The One-way Random Effects Model, *The Annals of Statistics*, vol. 6, no. 3, 582-587.

[9] 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*, vol. 35, 135-143.

[2002년 3월 접수, 2002년 7월 채택]