한국통계학회논문집 제3권 2호
1996년 8월 pp. 29-36

Confidence Intervals on Variance Components in Two Stage Regression Model

Dong Joon Park[1)]

Abstract

In regression model with nested error structure interval estimations about variability on different stages are proposed. This article derives an approximate confidence interval on the variance in the first stage and an exact confidence interval on the variance in the second stage in two stage regression model. The approximate confidence interval is based on Ting et al. (1990) method. Computer simulation is provided to show that the approximate confidence interval maintains the stated confidence coefficient.

1. Introduction

Regression model has been used to describe the relationship between the response and predictor variables. Exact representation of regression model is not possible because of random errors associated with factors not included in the model. In classical regression model, these errors are assumed to be uncorrelated and normally distributed with zero mean and constant variance. This article considers the multiple regression model where the responses are correlated. In particular, we consider two stage regression model, i.e., the multiple regression model with one-fold nested error structure. This model could be regarded as a single-factor covariance model with multiple concomitant variables. This model is appropriate to the data collected using two stage cluster designs. This model includes two error terms. One is assiciated with the first-stage sampling unit and the other with the second-stage sampling unit. These two error terms are independent and normally distributed with zero means and constant variances. However, this error structure gives correlated response variables.

Aitken and Longford(1986) showed ignoring the nesting structure is not appropriate to estimate regression coefficients. Park and Burdick(1993) proposed the confidence intervals on the variance components in simple regression model with one-fold nested error structure. This paper extends the work done by Park and Burdick. Yu and Burdick(1995) compared confidence

1) Full-time Lecturer, Department of Applied Mathematics, National Fisheries University of Pusan, Pusan, 608-737, Korea.

intervals on variance components based on restricted maximum likelihood estimators with confidence intervals on variance components using Ting et al. (1990) method in regression model with (Q-1) fold nested error structure. Tsubaki et al. (1995) proposed methods to estimate regression coefficients.

This article derives the confidence intervals on variance components associated with primary and secondary sampling units in two stage regression model. Section 2 describes two stage regression model specification with matrix notation and defines quadratic forms of the model that are independently chi-squared random variables. Section 3 proposes confidence intervals on variance components using Ting et al. (1990) method which requires independently distributed chi-squared random variables. Section 4 shows methods and results of simulations with regard to the confidence intervals on variance components in the model. Finally, section 5 draws conclusions.

2. Two stage regression model

The two stage regression model is written as

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 X_{h1} + \ldots + \beta_{p_1} X_{hp_1} + \delta_i \\ &\quad + \gamma_1 X_{ij1} + \ldots + \gamma_{p_2} X_{ijp_2} + \varepsilon_{ij} \qquad (2.1) \\ h &= 1, \cdots, \lambda_{p_1}\,;\ i = 1, \cdots, l_1\,;\ j = 1, \cdots, l_2 \end{aligned}$$

where Y_{ij} is the jth observation in the i th cell(group), β_0 is an intercept term, $\beta_1, \ldots, \beta_{p_1}$ are unknown parameters associated with primary units, $X_{h1}, \ldots, X_{hp_1}$are fixed predictor variables in the primary unit, $\gamma_1, \ldots, \gamma_{p_2}$ are unknown parameters associated with secondary units, $X_{ij1}, \ldots, X_{ijp_2}$ are fixed predictor variables in the secondary unit, δ_i is a random error term in the primary unit, ε_{ij} is a random error term in the secondary unit, δ_i and ε_{ij} are jointly independent normal random variables with zero means and variances σ_δ^2 and σ_ε^2, respectively. The index l_1 is the number of different combinations(cells) of levels among X_{ij}'s, i.e., $l_1 = \lambda_1 \times \lambda_2 \times \ldots \times \lambda_{p_1}$ and l_2 is the number of repetitions within an i th cell. We consider the balanced case where l_2's are same for all i's. Since β's, γ 's, X_{ij}'s, and X_{ijk}'s are fixed, and δ_i and ε_{ij} are random, model (2.1) is a mixed model.

The model (2.1) is written in matrix notation,

$$\underline{Y} = ZX_1\underline{\beta}+X_2\underline{\gamma}+Z\underline{\delta}+\underline{\varepsilon} \tag{2.2.1}$$

$$= Z\underline{U}+X_2\underline{\gamma}+\underline{\varepsilon} \tag{2.2.2}$$

$$= X\underline{\alpha}+\underline{\xi} \ , \tag{2.2.3}$$

where

$$\underline{U}= X_1\underline{\beta} + \underline{\delta}, \quad X=(ZX_1 \ \ X_2), \quad \underline{\alpha}=\begin{pmatrix}\underline{\beta}\\ \underline{\gamma}\end{pmatrix}, \quad \text{and} \quad \underline{\xi}= Z\underline{\delta} + \underline{\varepsilon},$$

where $\underline{Y}$ is an $l_1l_2\times 1$ vector of observations, Z is an $l_1l_2\times l_1$ design matrix with 0's and 1's, i.e., $Z=\bigoplus_{i=1}^{l_1}\mathbb{1}_{l_2}$ where $\mathbb{1}_{l_2}$ is an $l_2\times 1$ column vector of 1's and $\oplus$ is the direct sum operator, X_1 is an $l_1\times(p_1+1)$ matrix of known values with a column of 1's in the first column and p_1 columns of X_{ij}'s from the second column to the p_1th column, $\underline{\beta}$ is a $(p_1+1)\times 1$ vector of parameters associated with X_{ij}'s, X_2 is an $l_1l_2\times p_2$ matrix of known values with p_2 columns of X_{ijk}'s from the first column to the p_2th column, $\underline{\gamma}$ is a $p_2\times 1$ vector of parameters associated with X_{ijk}'s, $\underline{\delta}$ is an $l_1\times 1$ vector of random error terms, and $\underline{\varepsilon}$ is an $l_1l_2\times 1$ vector of random error terms. In particular,

$$\underline{Y}=\begin{pmatrix} Y_{11}\\ Y_{12}\\ \vdots\\ Y_{1l_2}\\ Y_{21}\\ Y_{22}\\ \vdots\\ Y_{2l_2}\\ \vdots\\ Y_{l_11}\\ Y_{l_12}\\ \vdots\\ Y_{l_1l_2}\end{pmatrix}, \quad Z=\begin{pmatrix} 10\cdots0\\ 10\cdots0\\ \vdots\\ 10\cdots0\\ 01\cdots0\\ 01\cdots0\\ \vdots\\ 01\cdots0\\ \vdots\\ 00\cdots1\\ 00\cdots1\\ \vdots\\ 00\cdots1\end{pmatrix}, \quad X_2=\begin{pmatrix} X_{111} & X_{112}\cdots X_{11p_2}\\ X_{121} & X_{122}\cdots X_{12p_2}\\ & \vdots\\ X_{1l_21} & X_{1l_22}\cdots X_{1l_2p_2}\\ X_{211} & X_{212}\cdots X_{21p_2}\\ X_{221} & X_{222}\cdots X_{22p_2}\\ & \vdots\\ X_{2l_21} & X_{2l_22}\cdots X_{2l_2p_2}\\ & \vdots\\ X_{l_111} & X_{l_112}\cdots X_{l_11p_2}\\ X_{l_121} & X_{l_122}\cdots X_{l_12p_2}\\ & \vdots\\ X_{l_1l_21} & X_{l_1l_22}\cdots X_{l_1l_2p_2}\end{pmatrix}, \quad \underline{\varepsilon}=\begin{pmatrix} \varepsilon_{11}\\ \varepsilon_{12}\\ \vdots\\ \varepsilon_{1l_2}\\ \varepsilon_{21}\\ \varepsilon_{22}\\ \vdots\\ \varepsilon_{2l_2}\\ \vdots\\ \varepsilon_{l_11}\\ \varepsilon_{l_12}\\ \vdots\\ \varepsilon_{l_1l_2}\end{pmatrix},$$

$$X_1 = \begin{pmatrix} 1 & X_{11} & X_{12} \cdots & X_{1p_1} \\ 1 & X_{11} & X_{12} \cdots & X_{2p_1} \\ & & \vdots & \\ 1 & X_{\lambda_1 1} & X_{\lambda_2 2} \cdots & X_{\lambda_{p_1} p_1} \end{pmatrix}, \qquad \underline{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p_1} \end{pmatrix}, \qquad \underline{\gamma} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_{p_2} \end{pmatrix}, \qquad \underline{\delta} = \begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_{l_1} \end{pmatrix}.$$

From (2.2.3), the variance-covariance matrix of $\underline{Y}$ is

$$Var(\underline{Y}) = \sigma_\delta^2 ZZ' + \sigma_\varepsilon^2 I_{l_1 l_2}, \tag{2.3}$$

since $\underline{\delta} \sim N(\underline{0}, \sigma_\delta^2 I_{l_1})$ and $\underline{\varepsilon} \sim N(\underline{0}, \sigma_\varepsilon^2 I_{l_1 l_2})$ where I_{l_1} is an $l_1 \times l_1$ identity matrix. From the assumptions in (2.1) and equation (2.3),

$$\underline{Y} \sim N(X\underline{\alpha}, \ \sigma_\delta^2 ZZ' + \sigma_\varepsilon^2 I_{l_1 l_2}). \tag{2.4}$$

The regression sums of squares of model (2.1) are now investigated. The reductions in sums of squares of the model are attributable to fitting the primary and secondary fixed variables and are expressed into the quadratic forms. Let $G_1 = (X^{*'} X^*)^{-1}$ and $G_2 = (\overline{X}_2' \overline{X}_2)^{-1}$ where $X^* = (X_1 \ X_2^*), X_2^* = \frac{Z'}{l_2} X_2, \overline{X}_2 = WX_2,$ and $W = I_{l_1 l_2} - ZZ'/l_2$. Define $H_1 = X^* G_1 X^{*'}$ and $H_2 = \overline{X}_2 G_2 \overline{X}_2'$. Now consider the quadratic forms $R_1 = \underline{Y}' \frac{Z}{l_2} (I_{l_1} - H_1) \frac{Z'}{l_2} \underline{Y}$ and $R_2 = \underline{Y}' W' (I_{l_1 l_2} - H_2) W \underline{Y}$. The quadratic form R_1 is determined by computing the regression of $\overline{Y}_{i.}$ on X_{ij} and $\overline{X}_{i.k}$ where $\overline{Y}_{i.} = \Sigma_{j=1}^{l_2} Y_{ij} / l_2$ and $\overline{X}_{i.k} = \Sigma_{j=1}^{l_2} X_{ijk} / l_2$. The quadratic form R_2 is calculated by the regression of Y_{ij} on the secondary fixed variables, X_{ijk}, and grouping variables. Under the distributional assumptions in (2.1), the quadratic forms $R_1 / (\sigma_\delta^2 + \frac{\sigma_\varepsilon^2}{l_2})$ and $R_2 / \sigma_\varepsilon^2$ are chi-squared random variables with $l_1 - p_1 - p_2 - 1$ and $l_1 l_2 - l_1 - p_2$ degrees of freedom, respectively. In addtion, the quadratic forms $R_1 / (\sigma_\delta^2 + \frac{\sigma_\varepsilon^2}{l_2})$ and $R_2 / \sigma_\varepsilon^2$ are independent(see Park(1996)). That is,

$$\frac{R_1}{\sigma_\delta^2 + \frac{\sigma_\varepsilon^2}{l_2}} \sim \chi^2_{l_1 - p_1 - p_2 - 1} \tag{2.5}$$

and

$$\frac{R_2}{\sigma_\varepsilon^2} \sim \chi^2_{l_1 l_2 - l_1 - p_2}. \tag{2.6}$$

3. Confidence Intervals on σ_δ^2 and σ_ε^2

Define $S_\delta^2 = R_1 / n_1$ and $S_\varepsilon^2 = R_2 / n_2$, where $n_1 = l_1 - p_1 - p_2 - 1$ and $n_2 = l_1 l_2 - l_1 - p_2$. Using (2.5) and (2.6), the expected mean squares are

$$E(S_\delta^2) = \sigma_\delta^2 + \frac{\sigma_\varepsilon^2}{l_2} = \theta_\delta \tag{3.1}$$

$$E(S_\varepsilon^2) = \sigma_\varepsilon^2 = \theta_\varepsilon. \tag{3.2}$$

Since $R_2 / \sigma_\varepsilon^2 \sim \chi^2_{n_2}$, an exact confidence interval on σ_ε^2 exists. This exact $(1-2\alpha)$ two-sided confidence interval on σ_ε^2 is

$$\left[\frac{S_\varepsilon^2}{F_{\alpha : n_2, \infty}} \; ; \; \frac{S_\varepsilon^2}{F_{1-\alpha : n_2, \infty}} \right], \tag{3.3}$$

where $F_{\delta : v_1, v_2}$ is the $(1 - \delta)$ th percentile F-value with v_1 and v_2 degrees of freedom.

The variance component σ_δ^2 is represented by the mean squares in (3.1) and (3.2). From (3.1) and (3.2),

$$\sigma_\delta^2 = \theta_\delta - \frac{\theta_\varepsilon}{l_2} . \tag{3.4}$$

Confidence intervals on σ_δ^2 can be constructed using the method of Ting et al.(1990). The $(1 - 2\alpha)$ two-sided confidence interval on σ_δ^2 using (3.4) is

$$\left[S_\delta^2 - \frac{S_\varepsilon^2}{l_2} - \left(U_1^2 S_\delta^4 + U_2^2 \frac{S_\varepsilon^4}{l_2^2} + U_{12} S_\delta^2 \frac{S_\varepsilon^2}{l_2} \right)^{\frac{1}{2}} ; \right. \quad (3.5)$$
$$\left. S_\delta^2 - \frac{S_\varepsilon^2}{l_2} + \left(V_1^2 S_\delta^4 + V_2^2 \frac{S_\varepsilon^4}{l_2^2} + V_{12} S_\delta^2 \frac{S_\varepsilon^2}{l_2} \right)^{\frac{1}{2}} \right],$$

where $U_1 = 1 - 1/F_{\alpha:n_1,\infty}$, $U_2 = 1/F_{1-\alpha:n_2,\infty} - 1$,

$U_{12} = [\ (F_{\alpha:n_1,n_2} - 1)^2 - U_1^2 F_{\alpha:n_1,n_2}^2 - U_2^2\ ] \ / F_{\alpha:n_1,n_2}$,

$V_1 = 1/F_{1-\alpha:n_1,\infty} - 1$, $V_2 = 1 - 1/F_{\alpha:n_2,\infty}$,

$V_{12} = [\ (1 - F_{1-\alpha:n_1,n_2})^2 - V_1^2 F_{1-\alpha,n_1,n_2}^2 - V_2^2] \ / F_{1-\alpha:n_1,n_2}$. Since $\sigma_\delta^2 > 0$, any negative bound is defined to be zero.

4. Simulation Study

Computer simulation was performed to compare the stated confidence coefficient and expected interval lengths. The criteria for analyzing the performance for the method are their ability to maintain stated confidence coefficients and the average length of two-sided confidence intervals. Although shorter average lengths are preferable, it is necessary that the methods should maintain the stated confidence coefficient.

Consider matrices X_1 and X_2 and the degrees of freedom in chi-squared random variables in (2.5) and (2.6). When $p_1 = 2$ and $p_2 = 3$, $l_1 \geq 7$ and $l_2 \geq 4$. Six designs are formed by taking all combinations of $l_1 = 8, 14, 20$ and $l_2 = 5, 10$ with $p_1 = 2$ and $p_2 = 3$. Let $\rho = \sigma_\delta^2/(\sigma_\delta^2 + \sigma_\varepsilon^2)$. Without loss of generality $\sigma_\delta^2 = 1 - \sigma_\varepsilon^2$ so that $\rho = \sigma_\delta^2$ and $1 - \rho = \sigma_\varepsilon^2$. Therefore, $S_\delta^2 \sim ((\rho + \frac{1-\rho}{l_2})/n_1)\chi_{n_1}^2$ and $S_\varepsilon^2 \sim ((1-\rho)/n_2)\chi_{n_2}^2$. These independent scaled chi-squared random variables can be generated by the RANGAM routine of the Statistical Analysis System(SAS). Values of ρ are varied from 0 to 1 in increments of 0.1 and simulated 1000 times for each design. Simulated values of S_δ^2 and S_ε^2 are substituted into (3.5) and the intervals are computed.

The two-sided intervals are calculated based on equal tailed F-values. Confidence coefficients are determined by counting the number of the intervals that contain σ_δ^2. Using the normal approximation to the binomial, if the true confidence coefficient is 0.90, there is less than a 2.5% chance that an estimated confidence coefficient based on 1000 replications will be less than 0.8814. The average lengths of the two-sided confidence intervals are also calculated. Table 1 reports the results of the simulation for stated 90% confidence intervals and the range of two-sided interval lengths on σ_δ^2 using (3.5) when $p_1 = 2$ and $p_2 = 3$. The proposed interval generally keep the stated confidence coefficients since all simulated confidence coefficients are bigger than 0.8814 and are not too conservative. The interval lengths get smaller as l_1 becomes bigger since it increases n_1 degrees of freedom. In addition, the interval lengths get smaller as l_2 becomes bigger since it increases n_2 degrees of freedom.

TABLE 1. Simulated Confidence Coefficients and Average Interval Lengths for 90% Two-sided Intervals on σ_δ^2

l_1	l_2		Coefficient	Length
8	5	Max	0.916	19.3902
		Min	0.890	3.7052
8	10	Max	0.911	18.7950
		Min	0.895	1.9164
14	5	Max	0.920	2.3854
		Min	0.892	0.3743
14	10	Max	0.910	2.4561
		Min	0.891	0.1848
20	5	Max	0.926	1.5090
		Min	0.896	0.2310
20	10	Max	0.915	1.5130
		Min	0.888	0.1119

5. Conclusions

This paper utilized distributional property of variance components in two stage regression model and derived confidence intervals on the variance components by use of independent quadratic forms which are chi-squared distributed. An exact confidence interval on the variability in the second stage of the model was obtained in (3.3) and an approximate confidence interval on the variability in the first stage of the model was proposed in (3.5). The simulations were performed to show that the proposed approximate confidence interval kept the stated confidence coefficients and average interval lengths changed as degrees of freedom of chi-squared random variables increased. The proposed confidence interval is recommended in two stage regression model applications.

References

[1] Aitken, M. and Longford, N. T. (1986). Statistical modelling issues in school effectiveness studies(with discussion), *Journal of the Royal Statististical Society A.*, 14A, 1-43.

[2] Park, D. J. (1996). The distributions of variance components in two stage regression model, 통계이론방법연구(forthcoming).

[3] Park, D. J. and Burdick, R. K.(1993). Confidence inrervals on the among group variance component in a simple linear regression model with a balanced one-fold nested error structure, *Communications in Statistics-theory and methods*, 22(12), 3435-3452.

[4] Searle, S. R.(1971). Linear Models, New York: John Wiley & Sons inc.

[5] Statistical Analysis System(1994). SAS User's Guide: Statistics, Cary, North Carolina: SAS Institute Inc.

[6] 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*, Vol. 35, 135-143.

[7] Tsubaki, M., Tsubaki, H, and Kusumi, M.(1995). A new estimation method for a useful class of mixed models, Proceedings of the 9th Asia Quality Management Symposium - Quality Enhancement for Global Prosperity, 275-280.

[8] Yu, Qing-Ling, and Burdick, R. K.(1995). Confidence intervals on variance components in regression models with balanced (Q-1)-fold nested error structure, *Communications in Statistics-theory and methods.*, 24(5), 1151-1167.