A Simulation Study on Intracluster Correlation

Ken W. Li

Department of Information and Communications Technology, Hong Kong Institute of Vocational Education (Tsing Yi), Hong Kong

***Abstract*— The study of the common intracluster correlation in simple linear regression is well developed ([1] and [2]). For the situation involving various intracluster correlation coefficients, the issues become more complicated. The prime objective of this study is to compare the loss of efficiency in using the intracluster correlation structure of common form to that in Toeplitz form for simple linear regression using generalised least squares.**

***Index Terms*—Generalised Least Squares, Intracluster correlation, Toeplitz form.**

I. Introduction

For the simple linear regression model, observations are usually assumed to be statistically independent. Often, however, in practice there may be physical constraints which lead to the violation of this assumption. For example, in studying human populations for genetic traits, one may select units out of families and then members in each family unit is observed. This would result in clusters in the sample. In repeated measures regression analysis, several pairs of observations (Y, X) are generally taken on the same subject. In the analysis, all the observations on all the subjects are considered. In this case, all pairs of observations are not statistically independent.

Suppose that a sample of N elements arises from a two-stage sampling scheme. At the first stage of sampling, k clusters are drawn and at the second stage n_i elements are drawn from the i^{th} sampled cluster $\left(N = \sum_{i=1}^{k} n_i \right)$. For each sampled element, we observe a dependent variable **Y** and an independent variable **X**. For the i^{th} cluster, $(i = 1, 2, \ldots, k)$, let Y_{ij} be the j^{th} observed value of the dependent variable and X_{ij} the value of the independent variable corresponding Y_{ij}, where $j = 1, 2, \ldots, n_i$. Hence, (X_{i1}, Y_{i1}), (X_{i2}, Y_{i2}), … , (X_{in_i}, Y_{in_i}) are the n_i pairs of observations in the i^{th} cluster. Let

$$\underset{\sim}{Y}_i = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \cdot \\ \cdot \\ \cdot \\ Y_{in_i} \end{bmatrix} \text{ and } \underset{\sim}{X}_i = \begin{bmatrix} X_{i1} \\ X_{i2} \\ \cdot \\ \cdot \\ \cdot \\ X_{in_i} \end{bmatrix}.$$

Consider the repeated measures regression model in which several pairs of observations (Y, X) are taken on each k subjects,

$$Y_{ij} = \beta_0 + \beta_1 Z_{ij} + e_{ij}, \; j = 1,2,\ldots,n_i \text{ and } i = 1,2,\ldots,k \qquad (1)$$

where $Z_{ij} = \left(X_{ij} - \overline{X}_{..}\right)$ is the deviation of X_{ij} from the overall mean

$$\overline{X}_{..} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}.$$

We assume that the errors e_{ij} are normally distributed with mean 0 and common variance σ^2. Further, the covariance between any two errors is given by $f(\rho)\sigma^2$ where f is a real-valued function. That is,

$$e_{ij} \sim N\left(0, \sigma^2\right)$$

$$\text{and } Cov(e_{ij}, e_{ij}^{'}) = \begin{cases} f(\rho)\sigma^2 & \text{if } i = i^{'} \\ 0 & \text{if } i \neq i^{'} \end{cases}.$$

In the standard regression model, ordinary least squares (OLS) estimation is used to estimate the slope parameter β_1 and the intercept term β_0. For the repeated measures regression model, the assumption of uncorrelated errors is no longer satisfied, and it may be misleading to examine the relationship of Y on X using OLS. Therefore, it is recommended to use generalized least squares (GLS). Although the correlation coefficient ρ is not usually known in practice, it can be estimated and a GLS performed. Some have termed this a pseudo-generalized least squares method. Our primary concern is not to estimate ρ but to compare the loss of efficiency between two distinct error structures in the above model based on GLS estimation. The first error structure involves the common intracluster correlation. The results for this error structure are well investigated by [1] and [2].

Manuscript accepted April 2, 2006.

Ken W. Li is a lecturer in the Department of Information and Communications Technology, Hong Kong Institute of Vocational Education (Tsing Yi), Room C440, 20 Tsing Yi Road, Tsing Yi Island, New Terrorties, Hong Kong (phone: 852-2436-8588; fax:852-2436-8526; e-mail: kenli@vtc.edu.hk).

However, the second error structure in Toeplitz form is not easy to handle. The results involve many algebraic computations and the matrices are complicated. This paper attempts to use reasonable notation to develop and simplify the results for this error structure and to compare the efficiencies based on the two structures. In the matrix notation, the model is,

$$\underset{\sim}{Y} = Z\underset{\sim}{\beta} + \underset{\sim}{e}$$

where $\underset{\sim}{e} \sim N\left(\underset{\sim N}{0}, V\right)$

and $\underset{\sim}{Y} = \begin{bmatrix} \underset{\sim}{Y}_1 \\ \underset{\sim}{Y}_2 \\ . \\ . \\ . \\ \underset{\sim}{Y}_k \end{bmatrix}$

$$\underset{\sim}{Z} = \begin{bmatrix} \underset{\sim n1}{1} & \underset{\sim}{Z}_1 \\ \underset{\sim n2}{1} & \underset{\sim}{Z}_2 \\ . & . \\ . & . \\ . & . \\ \underset{\sim nk}{1} & \underset{\sim}{Z}_k \end{bmatrix}$$

$$\underset{\sim i}{Z} = \left(\underset{\sim i}{X} - \overline{X}_{..}\underset{\sim n_i}{1}\right)$$

$$\underset{\sim}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$\underset{\sim}{e} = \begin{bmatrix} \underset{\sim 1}{e} \\ \underset{\sim 2}{e} \\ . \\ . \\ . \\ \underset{\sim k}{e} \end{bmatrix}.$$

Also, $\underset{\sim N}{0}$ is a $N \times 1$ vectors of zeros and $\underset{\sim n_i}{1}$ is a $n_i \times 1$ vectors of ones. It is assumed that observations from different clusters are uncorrelated but those in the same cluster are correlated. The covariance matrix is of the form

$$V = \begin{bmatrix} V_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & V_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & . & 0 & 0 & 0 \\ 0 & 0 & 0 & . & 0 & 0 \\ 0 & 0 & 0 & 0 & . & 0 \\ 0 & 0 & 0 & 0 & 0 & V_k \end{bmatrix}$$

and the inverse of V is

$$V^{-1} = \begin{bmatrix} V_1^{-1} & 0 & 0 & 0 & 0 & 0 \\ 0 & V_2^{-1} & 0 & 0 & 0 & 0 \\ 0 & 0 & . & 0 & 0 & 0 \\ 0 & 0 & 0 & . & 0 & 0 \\ 0 & 0 & 0 & 0 & . & 0 \\ 0 & 0 & 0 & 0 & 0 & V_k^{-1} \end{bmatrix}.$$

II. Error Structure

A. Common Intracluster Correlation

For the first structure considered, observations in the same cluster are correlated with a common intracluster correlation ρ. For the types of applications envisioned, ρ is assumed to be non-negative. The covariance matrix for i^{th} cluster given by [1] is

$$V_i = \sigma^2 \begin{bmatrix} 1 & \rho & . & . & . & \rho \\ \rho & 1 & . & . & . & \rho \\ . & & & & & \\ . & & & & & \\ . & & & & & \\ \rho & \rho & . & . & . & 1 \end{bmatrix}, \ \rho \geq 0 \text{ and } i = 1, 2, \ldots, k$$

$$= \sigma^2\left[(1-\rho)I_{n_i} + \rho J_{n_i}\right]$$

where I_{n_i} is a $n_i \times n_i$ identity matrix

and $J_{n_i} = \underset{\sim n_i}{1}\ \underset{\sim}{1}'_{n_i}$ is a $n_i \times n_i$ matrix of ones.

The inverse of V_i can be expressed as

$$V_i^{-1} = \left\{\sigma^2\left[(1-\rho)I_{n_i} + \rho J_{n_i}\right]\right\}^{-1}$$

$$= \frac{1}{\sigma^2(1-\rho)}\left[I_{n_i} - \rho\frac{J_{n_i}}{(1+n_i-\rho)}\right]$$

(a) It is well known that the best linear unbiased estimator (BLUE) for $\underset{\sim}{\beta}$ is the generalized least squares estimator ([1])

$$\hat{\underset{\sim}{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$$

$$= \left(Z'V^{-1}Z\right)^{-1}\left(Z'V^{-1}\underset{\sim}{Y}\right)$$

$$= \left(\sum_{i=1}^{k} Z_i'V_i^{-1}Z_i\right)^{-1}\left(\sum_{i=1}^{k} Z_i'V_i^{-1}\underset{\sim}{Y}_i\right)$$

(b) The variance-covariance matrix of $\hat{\beta}_0^{(1)}$ and $\hat{\beta}_1^{(1)}$ is given by

$$\text{Var}\begin{bmatrix} \hat{\beta}_0^{(1)} \\ \hat{\beta}_1^{(1)} \end{bmatrix} = \left(Z'V^{-1}Z\right)^{-1}$$

$$= \left(\sum_{i=1}^{k} Z_i'V_i^{-1}Z_i\right)^{-1}.$$

(c) Variance Inflation Factor

Even if the OLS estimator is reasonably efficient, the adjustment of OLS inferences for the clustering effect can still be made. In general the covariance matrix of $\hat{\underset{\sim}{\beta}}^{(1)}$ in (1) is given by

$$C = \sigma^2 (Z'Z)^{-1}(Z'VZ)(Z'Z)^{-1}$$
$$= \sigma^2 (Z'Z)^{-1} D$$

where $D = (Z'VZ)(Z'Z)^{-1}$

when $\rho = 0$, $(Z'VZ)(Z'Z)^{-1} = 1$

$$C = \sigma^2 (Z'Z)^{-1}$$

which is equivalent to C_0, the covariance matrix of $\hat{\beta}_0$ under OLS model. D represents the inflation that is needed to adjust the effect of incorrectly omitting the intracluster correlation from the model and is conditional on the observed X. From [1],

$$D = I + (M - I)\rho$$

where $M = \left(\sum_{i=1}^{k} n_i Z'_{B_i} Z_{B_i} \right) (Z'Z)^{-1}$

$$Z_{B_i} = \begin{bmatrix} 1 & \overline{Z}_{i.} \\ 1 & \overline{Z}_{i.} \\ . & . \\ . & . \\ . & . \\ 1 & \overline{Z}_{i.} \end{bmatrix}$$

Since $\overline{Z}_{..} = 0$, D is also known as the "misclassification effect".

B. Intracluster Correlation in Toeplitz Form

If observations are collected over time or space, adjacent observations on a subject may be more highly correlated than observations further apart. In this case, it would seem more appropriate to assume that observations in the same cluster are correlated with coefficient, ρ^{ℓ}, $\ell = 0, 1, \ldots, (n_i - 1)$. For this structure, the correlation at any two points depends on the distance between the points involved. The covariance matrix for the i^{th} cluster is

$$V_i = \sigma^2 \begin{bmatrix} 1 & \rho & . & . & . & \rho^{n_i - 1} \\ \rho & 1 & . & . & . & \rho^{n_i - 2} \\ \rho^2 & \rho & . & . & . & \rho^{n_i - 3} \\ & & & & & \\ & & . & . & . & \\ \rho^{n_i - 1} & \rho^{n_i - 2} & & & & 1 \end{bmatrix}$$

It can be shown ([3]) that the inverse of V_i for the above i^{th} cluster is:

$$V_i^{-1} = \sigma^2 \begin{bmatrix} a_1 & b_1 & 0 & . & . & . & 0 & 0 \\ & a_2 & b_2 & . & . & . & 0 & 0 \\ & & a_3 & . & . & . & 0 & 0 \\ & & & . & . & . & . & . \\ & & & & & & a_{n_i - 1} & b_{n_i - 1} \\ & & & & & & & a_{n_i} \end{bmatrix}$$

where $a_1 = a_{n_i} = \dfrac{1}{1 - \rho^2}$

$$a_i = \frac{1 + \rho^2}{1 - \rho^2}, i = 1, 2, \ldots, n_i - 1,$$

and $b_i = -\dfrac{\rho}{1 - \rho^2}$.

III. Loss of Efficiency

The efficiency of the estimators based on the common intracluster correlation error structure and the intracluster correlation structure in Toeplitz form, will be compared. For fixed values of ρ $(\rho = 0.01, 0.05, 0.1, 0.2, 0.3, 0.5, 0.9)$, one may calculate the ratio of $\mathrm{Var}\left(\underset{\sim}{C}' \hat{\underset{\sim}{\beta}}^{(2)} \right)$ to $\mathrm{Var}\left(\underset{\sim}{C}' \hat{\underset{\sim}{\beta}}^{(1)} \right)$ for an arbitrary coefficient vector $\underset{\sim}{C}$, where $\hat{\underset{\sim}{\beta}}^{(2)}$ denotes the GLE of $\underset{\sim}{\beta}$ with respect to the intracluster correlation in Toeplitz form and $\hat{\underset{\sim}{\beta}}^{(1)}$ denotes the GLE of $\underset{\sim}{\beta}$ with respect to common intracluster correlation. The following lemma given by [4] will be used in the efficiency comparison.

Lemma: Suppose A and B are any $n \times n$ matrices such that A^{-1} and B^{-1} exist. Define g to be real-valued function from n-Euclidean space without the zero vector by

$$g\left(\underset{\sim}{x} \right) = \frac{\underset{\sim}{x}' A^{-1} \underset{\sim}{x}}{\underset{\sim}{x}' B^{-1} \underset{\sim}{x}}, \quad x \in R^n - \{0\}.$$

Then, $\lambda_1 \le g\left(\underset{\sim}{x} \right) \le \lambda_n$

where λ_1 and λ_n are the smallest and largest eigenvalues respectively for the matrix BA ([4]). It follows that

$$\lambda_1 \le e\left(\underset{\sim}{C} \right) \le \lambda_n$$

where $e\left(\underset{\sim}{C} \right)$ is the efficiency of $\hat{\underset{\sim}{\beta}}^{(2)}$ and $\hat{\underset{\sim}{\beta}}^{(1)}$ defined by

$$e\left(\underset{\sim}{C} \right) = \frac{\mathrm{Var}\left(\underset{\sim}{C}' \hat{\underset{\sim}{\beta}}^{(2)} \right)}{\mathrm{Var}\left(\underset{\sim}{C}' \hat{\underset{\sim}{\beta}}^{(1)} \right)}.$$

$$= \frac{\sigma^2 \underset{\sim}{C}' Z' \left(V^{(2)}\right)^{-1} Z \underset{\sim}{C}}{\sigma^2 \underset{\sim}{C}' Z' \left(V^{(1)}\right)^{-1} Z \underset{\sim}{C}}$$

where $\left(V^{(2)}\right)^{-1}$ is the inverse of the second error structure and $\left(V^{(1)}\right)^{-1}$ respects to the first error structure.

In particular, λ_1, λ_n is the smallest and largest eigenvalue of $V^{(1)}.V^{(2)}$ respectively.

IV. Simulation Study

In order to compare the loss of efficiency between two different error structures, a simulation study was conducted using various correlation coefficients, $\rho = 0.01, 0.05, 0.1, 0.2, 0.3, 0.5, 0.9$, different k (the number of clusters) and n_i (the cluster size). For illustrations, $k = 5$ and $k = 10$ were selected with different values of n_i over three different intervals: $[2,6]$, $[2,10]$ and $[2,22]$.

For given $k \geq 1$, $n_i \geq 1$ and $0 \leq \rho \leq 1$, matrices V_i (an $n_i \times n_i$ covariance matrix in the i^{th} cluster corresponding to the second error structure), U_i (an $n_i \times n_i$ covariance matrix in the i^{th} cluster corresponding to the first error structure), V (an $N \times N$ covariance matrix for the second error structure) and U (an $N \times N$ covariance matrix for the first error structure) were calculated. The eigenvalues of the product UV were obtained to yield the upper and lower bounds. Finally, the loss of efficiency was evaluated. Table I shows the run with unequal cluster sizes.

Table I. Unequal cluster sizes

Case	k	n_i, $i = 1, 2, \ldots, k$
(i)	5	2, 3, 4, 5, 6
(ii)	10	2, 2, 2, 3, 2, 4, 2, 5, 2, 6
(iii)	5	2, 4, 6, 8, 10
(iv)	10	2, 2, 2, 4, 2, 6, 2, 8, 2, 10
(v)	5	2, 7, 12, 17, 22
(vi)	10	2, 2, 2, 7, 2, 12, 2, 17, 2, 22

Simulation was also run for equal cluster size, $n_i = n = 2, 3, 4, 5, 10, 20$ (6 cases) with different values of k. In order to obtain the eigenvalues of UV we only need to compute the eigenvalue corresponding to the i^{th} block of UV, $i = 1, 2, \ldots, k$. This is expected since U and V are symmetric and so the product UV is symmetric. More precisely, we consider

$$U = \begin{bmatrix} U_1 & & & 0 \\ & U_2 & & \\ & & \ddots & \\ 0 & & & U_k \end{bmatrix},$$

$$V = \begin{bmatrix} V_1 & & & 0 \\ & V_2 & & \\ & & \ddots & \\ 0 & & & V_k \end{bmatrix}$$

and get,

$$UV = \begin{bmatrix} U_1V_1 & & & 0 \\ & U_2V_2 & & \\ & & \ddots & \\ 0 & & & U_kV_k \end{bmatrix}.$$

Since the eigenvalues of the matrix UV only depend on the cluster size n_i, the number of clusters k and the arrangement of different V_i can be discarded. For example, repeated eigenvalues are observed in the case of equal cluster size. In fact, we run the program for $k = 1$ with $n_i = 2, \ldots, 20$ (total of 19 cases) we find out that we only need to calculate the eigenvalues corresponding to the V_i matrix with largest cluster size.

V. Result and Discussion

A. Upper bound on efficiency for combined cases

Fig. 1(a) and 1(b) present upper bound for equal cluster size and unequal cluster size respectively. We investigated that the upper bound on efficiency only depends on cluster size n_i. It is independent of the number of clusters k. It is quite obvious in Fig. 1(b) that the graphs are upper bound on efficiency for Cases (i) and (ii), Cases (iii) and (iv), and Cases (v) and (vi) respectively. For $k = 5$ or 10, the same graph is obtained but as n_i increases, the bound increases. For both plots, we know that when cluster size is large and ρ approaches 1, the efficiency on upper bound is infinite.

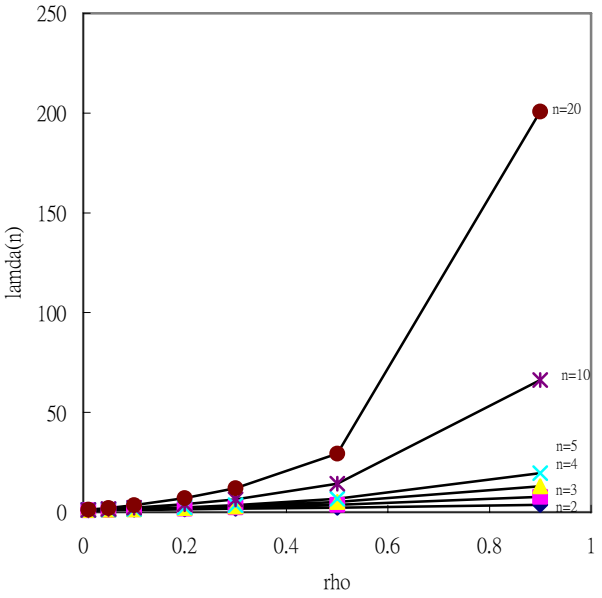

Fig. 1(a). Upper bound on efficiency (equal cluster size)

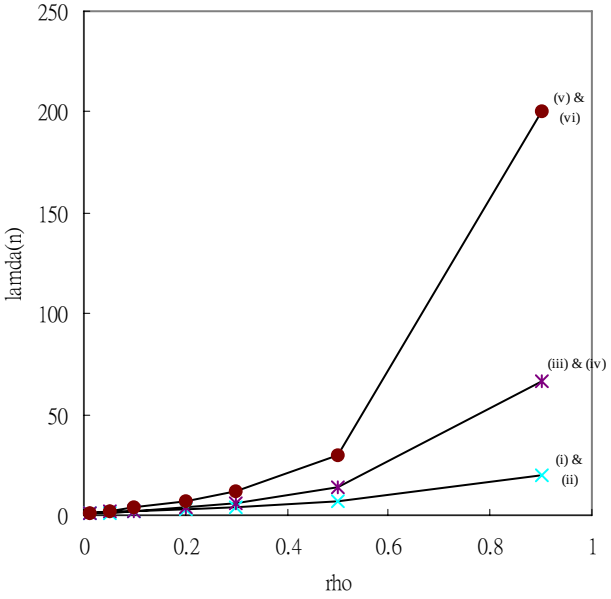

Fig. 1(b). Upper bound on efficiency (unequal cluster size)

B. Lower bound on efficiency for combined cases

Fig. 2(a) and Fig. 2(b) present lower bound for equal cluster size and unequal cluster size respectively. We got the same pattern from the above plots, the lower bound on efficiency does not depend on cluster size n_i very much. It does not depend on the number of clusters k. As ρ increases, the lower bound decreases. As ρ approaches 1, the efficiency tends to zero.

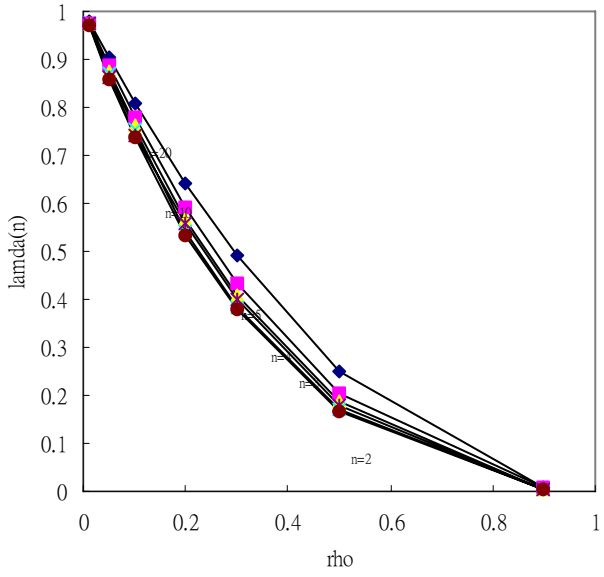

Fig. 2(a). Lower bound on efficiency (equal cluster size)

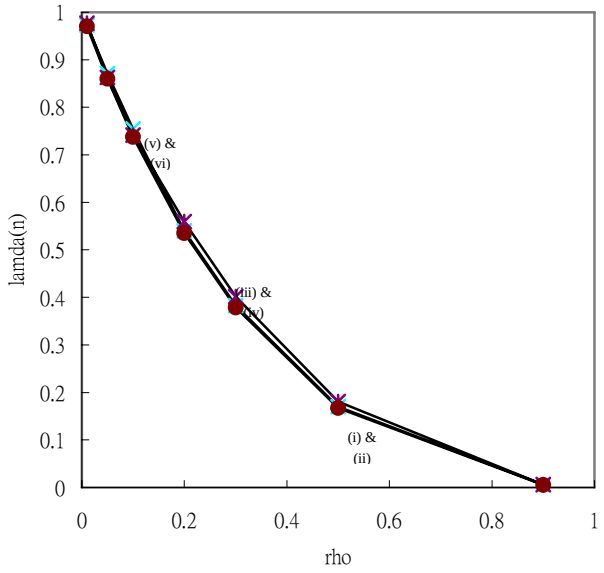

Fig. 2(b). Lower bound on efficiency (unequal cluster size)

C. Upper and Lower bound on efficiency

Fig. 3(a) and Fig. 3(b) present upper and lower bounds for $n_i = n = 2$ (equal cluster size) as well as Cases (i) and (ii) (unequal cluster size) respectively. By judging the efficiencies on upper and lower bounds in the same plot as shown in Fig. 3(a) and Fig. 3(b), the upper bound increases dramatically with ρ increases. The lower bound decreases gradually with ρ decreases, it approaches 0 and ρ goes to 1. For larger n_i, Cases (iii) and (iv), and Cases (v) and (vi), we obtained similar

pattern of plots but larger upper bound and smaller lower bound. Thus we confirm that we only need to consider the largest i^{th} cluster of UV and obtain the bounds on efficiency. It is not necessary to know all eigenvalues of UV and the bounds are not affected by k or different arrangements of n_i or even the total number of observations, N.

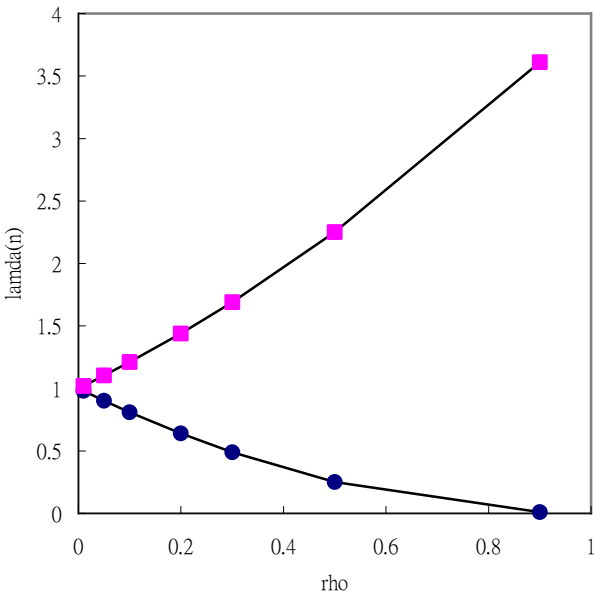

Fig. 3(a). Upper and lower bounds on efficiency for $n_i = n = 2$ (equal cluster size)

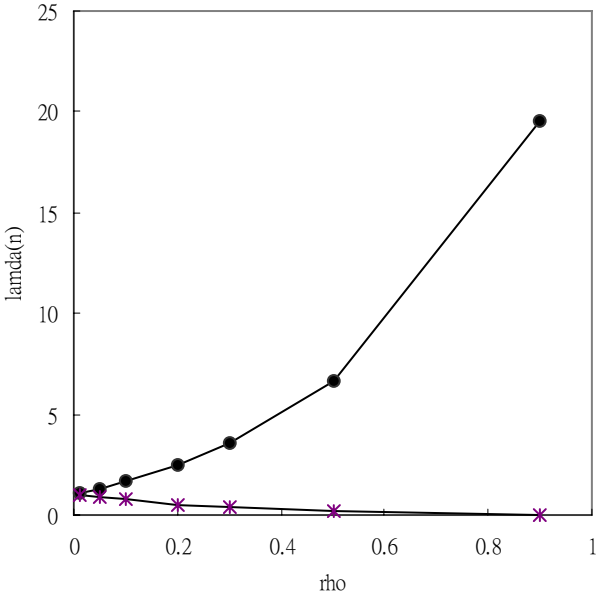

Fig. 3(b). Upper and lower bounds on efficiency for Cases (i) and (ii) (unequal cluster size)

D. Loss of efficiency for combined cases

Fig. 4(a) and Fig. 4(b) present loss of efficiency for upper and lower bounds for equal cluster size and unequal cluster size respectively. From both plots, they reveal that ρ increases, L_{MAX}, the upper bound on the loss of efficiency increases. As ρ approaches to 1, L_{MAX} is strictly less than one. This indicates that we obtain a gain in efficiency by employing the second error structure, Toeplitz form. When ρ tends to 1, the two distinct error structures are more or less with the same efficiency but the second error structure is better. There is not much inference on L_{MAX} with different values of n_i. When ρ is small, we find out that there is more gain in the efficiency of the second structure since L_{MAX} approaches 0. As n_i gets larger, L_{MAX} becomes more accurate. Moreover, when ρ approaches to zero, the loss of efficiency tends to zero. This simply means that we almost have an exact gain in our constructed error structure.

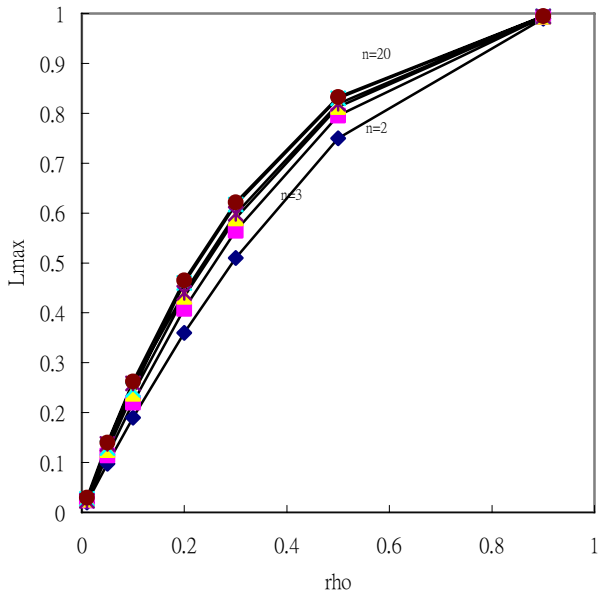

Fig. 4(a). Upper bound on the loss of efficiency (equal cluster size)

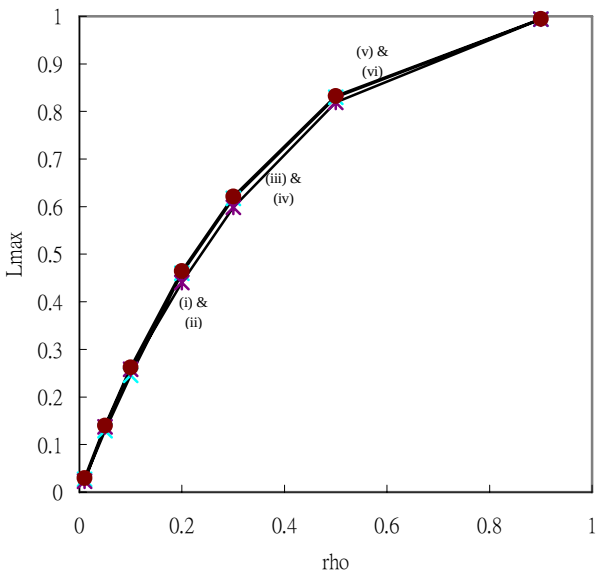

Fig. 4(b). Upper bound on the loss of efficiency (unequal cluster size)

VI. Conclusion

The idea of repeated measures regression analysis seems to be quite successful to handle inadequate observations which may be due to financial or human resources constraint in human surveys. In general, the model may deal with more than one independent variable. The principle will not change but the algebraic calculations are more tedious for multiple independent variables. Since the observations are not statistically independent, it is recommended to use GLS to obtain the best linear unbiased estimator (BLUE) of $\underset{\sim}{\beta}$. By comparing the loss of efficiency on two distinct error structures of common intracluster correlation and intracluster correlation in Toeplitz form that ρ increases, L_{MAX} increases. $1-e\left(\underset{\sim}{C}\right)$ is strictly less than 1 or $0<e\left(\underset{\sim}{C}\right)<1$, cluster size does not affect the efficiency very much.

In order to compare the GLS estimators $\hat{\underset{\sim}{\beta}}^{(1)}$ and $\hat{\underset{\sim}{\beta}}^{(2)}$ in (1), which may be equivalently written as

$Y_{ij}=\beta_0+\beta_1 X_{ij}+d_i+e_{ij}$, $i=1,2,\ldots,k$ and $j=1,2,\ldots,n_i$

where $d_i \sim \text{NID}\left(0,\sigma_{v_i}^2\right)$ and $e_{ij} \sim \text{NID}\left(0,\sigma_e^2\right)$.

The subject effect d_i represents the observations on a single subject to be high or low. We therefore generate d_i and e_{ij} as random variables from a normal distribution means 0 and variances $\sigma_{v_i}^2$ and σ_e^2 respectively. X_{ij} can also be generated from a symmetric multinormal distribution having $E\left(X_{ij}\right)=0$ and $E\left(X_{ij}^2\right)=1$, $E\left(X_{ij},X_{ij}^{'}\right)=\rho_x$, $j\neq j^{'}$, $E\left(X_{ij},X_{\ell j}^{'}\right)=0$, $i\neq\ell$. Y_{ij}'s are evaluated by putting the values of β_0, β_1, X_{ij}, d_i and e_{ij}. Then we perform some tests for comparison of estimation properties and significance test properties.

References

[1] A.J. Scott & D. Holt, "The effect of two-stage sampling on Ordinary Least Squares Method," *Journal of the American Statistical Association*, vol. 77, no. 380, 1982, pp. 848–854.

[2] A.P. Donner & G.A. Wells, "A comparison of confidence interval methods for the intracluster correlation coefficient," *Biometrics*, vol. 42, no. 2, 1986, pp. 401–412.

[3] D.F. Morrison, *Multivariate Statistical Methods.* New York: McGraw-Hill, 1976.

[4] C.S. Wong, *Linear Algebra*. New Jersey: Prentice Hall, 1976.