Journal of Statistical Planning and Inference 139 (2009) 3430 – 3441

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

A new alternative to the standard *F* test for clustered data

P. Lahiri[a,*], Yan Li[b,c]

[a] *Joint Program for Survey Methodology, University of Maryland, College Park 20742, USA*
[b] *Biostatistics Branch, National Cancer Institute, NIH 20852, USA*
[c] *Department of Mathematics, University of Texas, Arlington 76019, USA*

ARTICLE INFO

Article history:
Received 2 June 2008
Received in revised form
23 March 2009
Accepted 24 March 2009
Available online 2 April 2009

Keywords:
Clustered data
Intra-cluster correlation
Standard *F* test
Generalized least square test
Fuller–Battese transformation

ABSTRACT

The data collection process and the inherent population structure are the main causes for clustered data. The observations in a given cluster are correlated, and the magnitude of such correlation is often measured by the intra-cluster correlation coefficient. The intra-cluster correlation can lead to an inflated size of the standard *F* test in a linear model. In this paper, we propose a solution to this problem. Unlike previous adjustments, our method does not require estimation of the intra-class correlation, which is problematic especially when the number of clusters is small. Our simulation results show that the new method outperforms the existing methods.

1. Introduction

Linear models, in the form of either a regression model or an analysis of variance model, are widely used in various scientific investigations. In a linear model, the *F* test is used to test a linear hypothesis. For example, we can use an *F* test to test the relationship between a response variable and a number of predictor variables in a multiple regression model (e.g., when we test a hypothesis to find out if money spent on advertisement is related to the overall sales of a product in a business survey). We can also use *F* test to test the effect of a factor on the response variable in an ANOVA model (e.g., when we test the effect of different data collection methods on family expenditures in a consumer expenditure survey).

The analysts often encounter clustered data. In a given cluster, observations are generally correlated simply because of the similarity of the units in the cluster. For example, in a large scale household survey, blocks of households are often selected. The households within a given block are often similar with respect to various socio-economic characteristics leading to correlated observations. In an interviewer administered survey, observations collected by the same interviewer tend to be correlated. In health studies, observations are often collected on the same subject over time leading to correlated observations.

Wu et al. (1988) showed that the standard *F* test for a linear model leads to an inflated size for clustered data. The magnitude of such inflation increases with the increase of the intra-cluster correlation, a measure of the within cluster homogeneity. In other words, with the increase of the intra-cluster correlation the false significance rate increases. Wu et al. (1988) gave a partial explanation for this phenomenon. They argued that for correlated data the standard *F* statistic no longer follows the central *F* distribution under the null hypothesis.

Researchers have considered different procedures to test hypothesis in linear models in the presence of the intra-cluster correlation. For known intra-cluster correlation, Wu et al. (1988) adjusted the standard *F* statistic so that, under the null hypothesis and the correlated model, the adjusted statistic has approximately the same mean as that of the *F* statistics with the usual degrees

* Corresponding author. Tel.: +1 3013145903; fax: +1 3013147912.
E-mail addresses: plahiri@survey.umd.edu (P. Lahiri), lisherry@mail.nih.gov (Y. Li).

0378-3758/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2009.03.019

of freedom. The adjustment factor involves the intra-cluster correlation, which is then estimated using a two-stage procedure. The method rests on the assumption that the adjusted test statistic approximately follows central F distribution with the degrees of freedom same as those of the standard F statistic so that the usual F test can be applied using the adjusted F statistic.

Rao et al. (1993) proposed a two-stage generalized least square (GLS) test. The method involves making the observations uncorrelated, using a transformation previously proposed by Fuller and Battese (1973), and then applying the standard F test on the transformed observations. This is indeed an interesting idea. Like the test proposed by Wu et al. (1988), this test statistic depends on the intra-cluster correlation, but unlike the Wu–Holt–Holmes test statistics, the Rao–Sutradhar–Yue test statistic follows an exact central F distribution under the null hypothesis when the intra-cluster correlation is known. For this case, the Rao–Sutradhar–Yue test is indeed an F test based on the generalized least square under the original model and it is the likelihood ratio test. In fact, the test is uniformly most powerful among invariant tests of the general linear hypothesis. The final test statistic is obtained when the unknown intra-cluster correlation is replaced by an estimator obtained by Henderson's (1953) well-known method of fitting constants.

The Rao–Sutradhar–Yue method turned out to be the most promising test among all tests considered in their paper, in terms of both controlling size and gaining the power. However, we note that the estimation of the intra-cluster correlation is problematic, especially when the number of clusters available to estimate it is small or when the true intra-cluster correlation is small (but different from zero). For a given data set, the estimate of the intra-cluster correlation could be zero. When this happens, the Rao–Sutradhar–Yue method reduces to the standard F test and all the benefits of adjustments are lost.

Survey researchers frequently encounter situations with a small number of clusters. For example, only 9 female interviewers were employed to interview 489 respondents in a study conducted in 1958 at a large nonunionized oil refinery in Canada (Kish, 1962). The intra-interviewer correlation estimates appear to depend on the survey items. Attitude items and complex factual items are considered more sensitive to the intra-interviewer correlation than simple factual items are (Collins and Butcher, 1982; Feather, 1973; Fellegi, 1964; Gray, 1956; Hansen et al., 1961). According to Groves (1989), intra-interviewer correlation estimates above 0.1 are seldom observed. See Schnell and Kreuter (2005) for further discussions on this issue.

As a remedy, one can either consider a better estimator of the intra-cluster correlation or consider a test, which does not depend on the intra-cluster correlation. Even if one uses a more sophisticated variance component estimator (e.g., REML, see Jiang, 1996), the problem persists, especially when the number of clusters is small or the true intra-cluster correlation is small. Thus, in this paper we consider the other approach, which avoids the estimation of the intra-cluster correlation. In order to make the observations independent, we employ a transformation that is different from the Fuller-Battese transformation used by Rao et al. (1993). Following Rao et al. (1993), we then apply the standard F test. The advantage of our test over the Rao–Sutradhar–Yue or the Wu–Holt–Holmes test is that our test statistic is free of the unknown intra-cluster correlation. Of course, this good feature of our test is achieved at a cost of losing a few degrees of freedom. However, this does not seem to affect the efficiency of our test when the number of clusters is small. In any case, for a given data if the intra-cluster correlation estimate turns out to be zero, our approach offers a test which is different from the standard F test.

For many years, psychologists had (and still do) performed analysis of variance on repeated measures designs (observations on the same subject at different points in time) while ignoring correlations. In fact, their selected error terms were such that correlations did not appear in their F-ratios, and the standard literature of psychology claimed that this occurred when the errors were equi-correlated. Hyunh and Feldt (1970) and Rouanet and Lepine (1970) showed that the equi-correlation assumption for the errors was not necessary for the F-ratios to be free of the correlation term, and that a more general condition sufficed, namely that all differences should have equal variances. In an extension to the multivariate case, Thomas (1983) used the same transformation as used in this paper, and gave an alternative representation of the covariance matrix in terms of the generalized inverse of a matrix derived from the same transformation. The application to regression used in this paper extends the previous work to a set of clusters of different sizes.

In Section 2, we introduce some notation and state the problem of interest. In Section 3, we briefly review the testing procedures proposed by Wu et al. (1988) and Rao et al. (1993). We present our new test in Section 4. In Section 5, we present results from a simulation study. It appears that our test is most effective in controlling the test size under a variety of simulation conditions.

2. Two motivating examples

We begin this section with two motivating examples:

Example 1. Suppose we are interested in examining the relationship between two variables, say the hospital stay and the number of beds in the hospital. One can consider the following simple linear regression model:

$$Y_{ij} = \beta x_i + \varepsilon_{ij},$$

where Y_{ij} is the number of days for the jth patient in the ith hospital; x_i is the number of beds available in the ith hospital; ε_{ij}'s are pure errors not explained by the regression model

$$i = 1, \ldots, I; \quad j = 1, \ldots, n_i.$$

In order to understand if the above linear regression explains the relationship, one can consider the following test:

$$H_o : \beta = 0 \quad \text{vs.} \quad H_1 : \beta \neq 0.$$

Under the assumption that ε_{ij} are independently and identically distributed (iid) with $\varepsilon_{ij} \sim N(0, \sigma^2)$, where σ^2 is known, we can test the above hypothesis by the standard z test given by

$$\text{Reject } H_o \text{ if } \left| \frac{\hat{\beta}}{se(\hat{\beta})} \right| > z_{\alpha/2},$$

where

$$\hat{\beta} = \frac{\sum_{i=1}^{I} \sum_{j=1}^{n_i} x_i Y_{ij}}{\sum_{i=1}^{I} n_i x_i^2}, \quad \text{the ordinary least square estimator (OLS) of } \beta;$$
$$se(\hat{\beta}) = \frac{\sigma}{\sqrt{\sum_{i=1}^{I} n_i x_i^2}}, \quad \text{the standard error of } \hat{\beta};$$

and $z_{\alpha/2}$ is the upper $(\alpha/2)\times 100$ percentile point of the standard normal deviate Z.

One may argue that the assumption of independence for ε_{ij}'s is not justified since the observations for the same hospital are likely to be correlated. Consider the following simple correlated model:

$$Var(\varepsilon_{ij}) = \sigma^2;$$
$$Cov(\varepsilon_{ij}, \varepsilon_{ij'}) = \sigma^2 \rho;$$
$$i = 1, \ldots, I; \;\; j, j' = 1, \ldots, n_i; \;\; j \neq j',$$

where ρ is the intra-hospital correlation measuring the internal homogeneity of observations within a hospital. Let us examine the impact of the intra-hospital correlation on the size of the standard z test. The size of the test under the correlated model is given by

$$P_C\left[\left|\frac{\hat{\beta}}{se(\hat{\beta})}\right| > z_{\alpha/2} | H_0\right] = P_C\left[\left|\frac{\hat{\beta}}{se_C(\hat{\beta})}\frac{se_C(\hat{\beta})}{se(\hat{\beta})}\right| > z_{\alpha/2} | H_0\right] = P_C\left[\left|\frac{\hat{\beta}}{se_C(\hat{\beta})}\right| > \frac{se(\hat{\beta})}{se_C(\hat{\beta})} z_{\alpha/2} | H_0\right] = P_C\left[|Z| > \frac{se(\hat{\beta})}{se_C(\hat{\beta})} z_{\alpha/2} | H_0\right],$$

where

$$\frac{se(\hat{\beta})}{se_C(\hat{\beta})} = \frac{1}{\sqrt{1 + (\bar{n}_a - 1)\rho}},$$
$$\bar{n}_a = \sum_{i=1}^{I} a_i n_i, \quad \text{a weighted average sample size per hospital,}$$
$$a_i = \frac{n_i x_i^2}{\sum_{i=1}^{I} n_i x_i^2}.$$

In the above, P_C is a probability under the correlated model. Since $\rho > 0$, we have $se(\hat{\beta})/se_C(\hat{\beta}) z_{\alpha/2} < z_{\alpha/2}$ and hence

$$P_C\left[\left|\frac{\hat{\beta}}{se(\hat{\beta})}\right| > z_{\alpha/2} | H_0\right] \geqslant \alpha.$$

Thus, we will see an inflation of the nominal size. The magnitude of this inflation depends on both ρ and $\bar{n}_a$. It is interesting to note that even if ρ is small we can still see considerable inflation if the average hospital sample size is large.

In practice, for the uncorrelated linear model, σ^2 is unknown and is estimated by

$$MSE = \frac{\sum_{i=1}^{I} \sum_{j=1}^{n_i} e_{ij}^2}{\sum_{i=1}^{I} n_i - 1},$$

where $e_{ij} = y_{ij} - \hat{\beta} x_i$, the residual from the linear regression. In this case, the test reduces to the standard t test:

$$\text{Reject } H_o \text{ if } \left| \frac{\hat{\beta}}{se(\hat{\beta})} \right| > t_{\alpha/2; \sum n_i - 1},$$

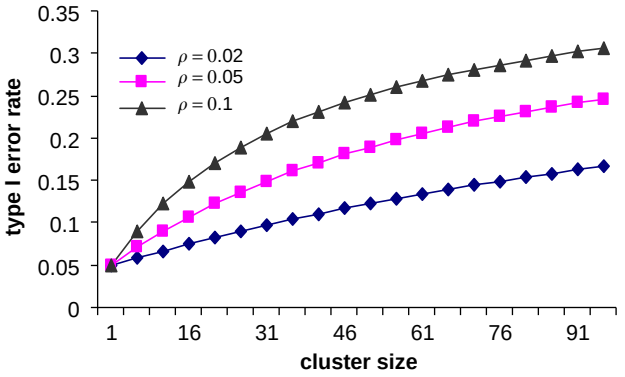

Fig. 1. Actual sizes of the standard F tests of $H_0 : \beta = 0$ for different cluster sizes and intra-cluster correlations ρ (nominal level = 0.05).

where

$$se(\hat{\beta}) = \sqrt{\frac{MSE}{\sum_{i=1}^{I} n_i x_i^2}},$$

and $t_{\alpha/2;\sum n_i - 1}$ is the upper $(\alpha/2)\times 100$ percentile point of the central t distribution with $\sum n_i - 1$ degrees of freedom. The above test is equivalent to the standard F test given by

$$\text{Reject } H_o \text{ if } F = \frac{MSR}{MSE} > F_{\alpha;1,\sum n_i - 1},$$

where MSR is the mean squared due to the regression in the ANOVA table and $F_{\alpha;1,\sum n_i - 1}$ is the upper 100α percentile point of the central F distribution with 1 and $\sum n_i - 1$ as the numerator and denominator degrees of freedom, respectively.

For the general linear regression model, Wu et al. (1988) noted that, under the correlated model, the null distribution of the F statistics is no longer the F distribution. They also observed that the size will be inflated under the correlated model, but a formal proof for such inflation for the unknown σ case seems difficult and is not available. Fig. 1 plots the actual sizes of the standard z test of $H_0 : \beta = 0$ at the nominal 5% level with varying cluster sizes and intra-cluster correlations (ρ) under the correlated model. It can be seen that the size increases with the increase of ρ and/or cluster size.

Example 2. Suppose we are interested in testing if there are any significant differences among I different questionnaires in terms of time taken to complete them. To this end n_0 respondents are randomly assigned to each questionnaire, and the time taken to complete it is noted. In this situation, it is a common practice to consider the following ANOVA model:

$$y_{ij} = \mu_i + \varepsilon_{ij},$$

where y_{ij} is the time taken to complete the ith questionnaire type by the jth respondent; μ_i is the true mean time taken for the ith questionnaire type; ε_{ij}'s are the pure error terms.

Under the assumption that ε_{ij}'s are iid $N(0, \sigma^2)$, we can test the hypothesis, i.e.,

$$H_0 : \mu_1 = \cdots = \mu_I \quad \text{vs.} \quad H_1 : \text{ at least one } \mu_i \text{ is different,}$$

by the following standard ANOVA F test:

$$\text{Reject } H_0 \text{ if } F = \frac{MSTR}{MSE} > F_{\alpha/2;I-1,I(n_0-1)},$$

where $MSTR$ and MSE are the mean squared due to the treatment (questionnaire) and error, respectively. The above test can suffer from an inflated size if the assumption of independence for the error terms is not valid, say, when the questionnaires are administered by interviewers. The observations are likely to be correlated for the same interviewer and thus a correlated model for the error terms seems reasonable.

Consider a situation when J interviewers administer one type of questionnaire. Suppose we consider equal workload for the J interviewers available for this study. Thus, each interviewer conducts $n_0^* = n_0/J$ interviews. We want to study how the interviewer workload, or equivalently, the number of interviewers for questionnaire type, and the intra-interviewer correlation affect the size of the above F test. We answer the question through a small simulation. Let y_{ijk} denote the time taken by the kth respondent

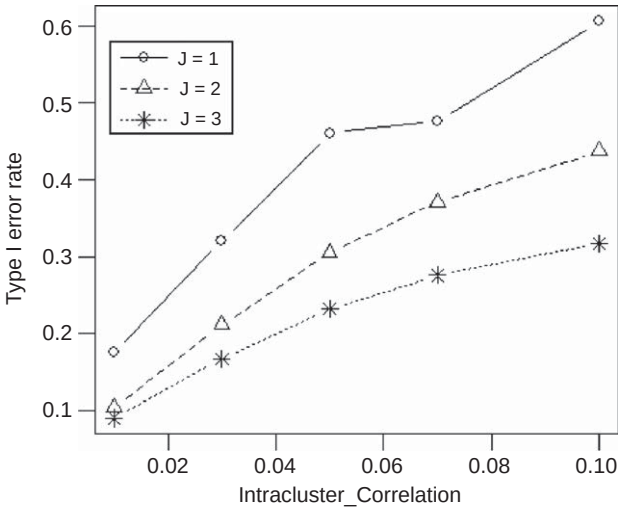

Fig. 2. Actual sizes of the standard F test of $H_0 : \mu_1 = \mu_2$ for different intra-cluster correlations and the number of interviewers per group (J) (nominal level = 0.05).

to complete questionnaire type i administered by interviewer $j (i = 1, \ldots, I; j = 1, \ldots, J; k = 1, \ldots, n_0^*)$. The following model may be appropriate in capturing the intra-interviewer correlation:

$$y_{ijk} = \mu_i + v_{ij} + \varepsilon_{ijk} \quad \text{for } i = 1, \cdots, I; \; j = 1, \ldots, J; \; k = 1, \ldots, n_0^*. \tag{1}$$

We assume that v_{ij}'s are iid $N(0, \sigma_v^2)$, ε_{ijk} are iid $N(0, \sigma_\varepsilon^2)$, and v_{ij}'s and ε_{ijk}'s are independent. Define $n = I \times J \times n_0^*$, $\sigma^2 = \sigma_v^2 + \sigma_\varepsilon^2$, and $\rho = \sigma_v^2/\sigma^2$. We first specify μ_i for $i = 1, 2, \ldots, I$. For this simulation, we specify $\sigma_\varepsilon^2 = 2$, $I = 2$, $n = 200$, $\mu_1 = \mu_2 = 20$, and vary the values of $J (= \{1, 2, 3\})$ and $\rho (= \{.01, .03, .05, .07, .1\})$. For a given simulation condition, we generate $\{y_{ijk}; k = 1, \ldots, n_0^*\}$ from model (1). The simulated data $\{(y_{ijk}, \mu_i); i = 1, 2, \ldots, I; j = 1, 2, \ldots, J; k = 1, 2, \ldots, n_0^*\}$ are used to compute the standard F statistic to test $H_0 : \mu_1 = \cdots = \mu_I$. For each simulation condition, we conduct 1000 simulation runs to approximate actual type I error rates. Fig. 2 shows the type I error rates of the standard F tests of $H_0 : \mu_1 = \mu_2$ at the nominal 5% level with varying J and ρ. It is clear that that type I errors are all inflated (> 0.05). Smaller J (or larger interviewer workload, i.e., the average number of respondents assigned to an interviewer) and/or larger ρ cause larger type I error rate inflation. When $\rho = 0.1$ and $J = 1$, type I error is even greater than 60%.

3. Statement of the problem and review of different existing tests

Let $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})'$ be a vector of n_i observations from the ith cluster on a response variable y, and let $(x_{i1l}, x_{i2l}, \ldots, x_{in_i l}), l = 1, \ldots, k-1$ be the associated values of $k-1$ predictor variables: $x_1, x_2, \ldots, x_{k-1}$.

Following Wu et al. (1988) and Rao et al. (1993), we consider a regression model with an error structure that allows for the intra-cluster correlation of the residual errors:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2}$$

where

$$\mathbf{y} = \begin{pmatrix} \mathbf{y_1} \\ \mathbf{y_2} \\ \vdots \\ \mathbf{y_I} \end{pmatrix} \quad \text{with} \quad \mathbf{y}_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in_i} \end{pmatrix},$$

$$\boldsymbol{\epsilon} = \begin{pmatrix} \boldsymbol{\epsilon_1} \\ \boldsymbol{\epsilon_2} \\ \vdots \\ \boldsymbol{\epsilon_I} \end{pmatrix} \quad \text{with} \quad \boldsymbol{\epsilon_i} = \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{in_i} \end{pmatrix},$$

$$\mathbf{X} = \begin{pmatrix} \mathbf{X_1} \\ \mathbf{X_2} \\ \vdots \\ \mathbf{X_I} \end{pmatrix}$$

with $\mathbf{X_i}$ denoting the $n_i \times k$ matrix with rows $\mathbf{x}'_{ij} = (x_{ij0}, x_{ij1}, \ldots, x_{ij,k-1})$ and $x_{ij0} = 1$, $\boldsymbol{\beta}$ is a $k \times 1$ column vector of regression parameters. Furthermore, $\boldsymbol{\epsilon} \sim N_n(0, \sigma^2\mathbf{V})$, where $\mathbf{V}$ has the block-diagonal form $\oplus_1^I \mathbf{V}_i$ with

$$\mathbf{V}_i = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}_{n_i \times n_i} = (1-\rho)\mathbf{I}_{n_i} + \rho \mathbf{1}_{n_i}\mathbf{1}'_{n_i}$$

for the ith cluster and $\rho = \sigma_v^2/\sigma^2$, the common intra-cluster correlation coefficient. Such a model was considered in Campbell (1977), Scott and Holt (1982), and others.

Suppose that the hypothesis of interest is

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{b},$$

where $\mathbf{C}$ is a known $q \times k$ matrix of rank q ($< k$), and $\mathbf{b}$ is a known $q \times 1$ vector. The standard F test of H_0 is based on

$$F_s = \frac{(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{b})'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{b})/q}{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})/(n-k)},$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, the ordinary least square estimator of $\boldsymbol{\beta}$ that does not account for the variance covariance structure of the error terms. The test is carried out as follows:

$$\text{Reject } H_0 \text{ if } F_s \geqslant F_{\alpha,q,n-k},$$

where $F_{\alpha,q,n-k}$ is the upper 100α percentile point of the central F distribution. Under model (2), the F_s statistic does not generally have an F distribution under the null hypothesis and the F test is distorted. As noted by Rao et al. (1993) and Wu et al. (1988), this F test can lead to highly inflated type I error rate (size) as ρ increases under model (2).

Wu et al. (1988) proposed a new F_A statistic with a correction multiplication factor $h(\rho)$ that accounts for the effect of the intra-cluster correlation on the standard F statistic:

$$F_A(\rho) = F_s \times h(\rho),$$

where

$$h(\rho) = \frac{[n - \mathrm{tr}(\mathbf{PV})]/(n-k)}{\mathrm{tr}(\mathbf{P}_C\mathbf{V})/q},$$
$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}', \quad \mathbf{P}_C = \mathbf{X}_C(\mathbf{X}_C'\mathbf{X}_C)^{-1}\mathbf{X}_C', \quad \mathbf{X}_C = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}',$$

and tr denotes the trace operator. Wu et al. (1988) motivated the correction multiplier $h(\rho)$ as follows. Under model (2), the numerator and the denominator of F_s are approximated, respectively, by

$$[\mathrm{tr}(\mathbf{P}_C\mathbf{V})/q](\chi_q^2/q) \quad \text{and} \quad [n - \mathrm{tr}(\mathbf{PV})]/(n-k)[\chi_{n-k}^2/(n-k)].$$

Therefore,

$$F_s \approx \frac{[\mathrm{tr}(\mathbf{P}_C\mathbf{V})/q][\chi_q^2/q]}{[n - \mathrm{tr}(\mathbf{PV})]/(n-k)[\chi_{n-k}^2/(n-k)]} = h^{-1}(\rho)F_A(\rho).$$

The statistic F_A is obtained by adjusting F_s with a correction multiplier $h(\rho)$, and the distribution of F_A is then approximated by the F distribution with q and $n-k$ df. Their simulation study showed that F_A performed much better than F_s in controlling the size.

Rao et al. (1993) proposed a two-step generalized least square F statistic obtained by first estimating ρ, and then substituting the estimate into the GLS test statistic when ρ is known. Following Fuller and Battese (1973), they first transformed model (2) into a standard regression model with iid errors.

Define

$$\alpha_i = 1 - \sqrt{\frac{1-\rho}{1+(n_i-1)\rho}},$$
$$y_{ij}^* = y_{ij} - \alpha_i\bar{y}_{i.} \quad \text{and} \quad \bar{y}_{i.} = \sum_j y_{ij}/n_i,$$
$$\mathbf{x}_{ij}^* = \mathbf{x}_{ij} - \alpha_i\bar{\mathbf{x}}_{i.} \quad \text{and} \quad \bar{\mathbf{x}}_{i.} = \sum_j \mathbf{x}_{ij}/n_i.$$

The transformed model may then be written as

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \mathbf{u}^*, \tag{3}$$

where $\mathbf{u}^* \sim N_n(\mathbf{0}, \sigma_u^2\mathbf{I}_n)$ and $\sigma_u^2 = \sigma^2\rho$. It now follows that standard OLS methods for inference on $\boldsymbol{\beta}$ may be applied to the transformed data. In particular, the standard F test can be employed on the transformed data and the GLS F test is based on

$$F_{GLS}(\rho) = \frac{(\mathbf{C}\boldsymbol{\beta}^* - \mathbf{b})'(\mathbf{X}_C^{*\prime}\mathbf{X}_C^*)^{-1}(\mathbf{C}\boldsymbol{\beta}^* - \mathbf{b})/q}{(\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}^*)'(\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}^*)/(n-k)},$$

where $\boldsymbol{\beta}^* = (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{y}^*$ and $\mathbf{X}_C^* = \mathbf{X}^*(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{C}'$. Under the null hypothesis, the GLS F test, $F_{GLS}(\rho)$, has an exact F distribution, and it is the likelihood ratio test when ρ is known.

Note that the proposed $F_A(\rho)$ statistic and $F_{GLS}(\rho)$ statistic are both dependent on the unknown intra-cluster correlation ρ. In practice, ρ is rarely known and must be estimated. Rao et al. (1993) used the well-known Henderson (1953) method of fitting constants to obtain $\hat{\rho}$, an estimator of ρ. Under certain regularity conditions, the estimator is consistent when the number of clusters is large. In our simulation, we encounter problem with this estimator, especially when ρ is small and/or the number of clusters is small, a situation often encountered in practice as noted in Section 1.

4. A new test

In order to make the observations independent but at the same time to avoid the estimation of intra-cluster correlation, we employ a transformation that is different from the Fuller–Battese transformation used by Rao et al. (1993). In a multivariate setting, the same transformation was used earlier by Thomas (1983).

Under model (2), let $\mathbf{P}_i$ be an $(n_i - 1) \times n_i$ matrix such that

$$\begin{pmatrix} n_i^{-1/2}\mathbf{1}_{n_i}' \\ \mathbf{P}_i \end{pmatrix}$$

is orthogonal, $i = 1, 2, \ldots, I$. For example, if $n_i = 5$ then one may take

$$\mathbf{P}_i = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} & 0 & 0 \\ \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{12}} & -\frac{3}{\sqrt{12}} & 0 \\ \frac{1}{\sqrt{20}} & \frac{1}{\sqrt{20}} & \frac{1}{\sqrt{20}} & \frac{1}{\sqrt{20}} & -\frac{4}{\sqrt{20}} \end{pmatrix}.$$

For $i = 1, 2, \ldots, I$, let

$$\mathbf{y}_i^{new} = \mathbf{P}_i\mathbf{y}_i \quad \text{and} \quad \mathbf{X}_i^{new} = \mathbf{P}_i\mathbf{X}_i.$$

For example, for $n_i = 5$

$$\mathbf{y}_i^{new} = (\mathbf{y}_{i2}^{new}, \mathbf{y}_{i3}^{new}, \mathbf{y}_{i4}^{new}, \mathbf{y}_{i5}^{new})'$$

and

$$\mathbf{y}_i = (y_{i1}, y_{i2}, y_{i3}, y_{i4}, y_{i5})'.$$

Then, with $\mathbf{P}_i$ chosen above,

$$y_{i2}^{new} = \sqrt{2}\left[\frac{y_{i1}+y_{i2}}{2} - y_{i2}\right], y_{i3}^{new} = \sqrt{\frac{3}{2}}\left[\frac{y_{i1}+y_{i2}+y_{i3}}{3} - y_{i3}\right],$$
$$y_{i4}^{new} = \sqrt{\frac{4}{3}}\left[\frac{y_{i1}+y_{i2}+y_{i3}+y_{i4}}{4} - y_{i4}\right], y_{i5}^{new} = \sqrt{\frac{5}{4}}\left[\frac{y_{i1}+y_{i2}+y_{i3}+y_{i4}+y_{i5}}{5} - y_{i5}\right].$$

The transformed model may then be written as

$$\mathbf{y}^{new} = \mathbf{X}^{new}\boldsymbol{\beta} + \mathbf{u}^{new},$$

where

$$\mathbf{y}^{new}=\begin{pmatrix}\mathbf{y}_1^{new}\\ \mathbf{y}_2^{new}\\ \vdots\\ \mathbf{y}_I^{new}\end{pmatrix} \quad \text{with} \quad \mathbf{y}_i^{new}=\begin{pmatrix}y_{i1}^{new}\\ y_{i2}^{new}\\ \vdots\\ y_{in_i}^{new}\end{pmatrix},$$

$$\mathbf{X}^{new}=\begin{pmatrix}\mathbf{X}_1^{new}\\ \mathbf{X}_2^{new}\\ \vdots\\ \mathbf{X}_I^{new}\end{pmatrix} \quad \text{and} \quad \mathbf{u}^{new}=\begin{pmatrix}\mathbf{u}_1^{new}\\ \mathbf{u}_2^{new}\\ \vdots\\ \mathbf{u}_I^{new}\end{pmatrix}.$$

It follows that

$$Var(\mathbf{u}_i^{new})=\sigma^2\mathbf{P}_i\mathbf{V}_i\mathbf{P}_i'=\sigma^2\mathbf{P}_i[(1-\rho)\mathbf{I}_{n_i}+\rho\mathbf{1}_{n_i}\mathbf{1}_{n_i}']\mathbf{P}_i'=\sigma^2(1-\rho)\mathbf{I}_{n_i-1}.$$

Also,

$$Cov(\mathbf{u}_i^{new},\mathbf{u}_j^{new})=\mathbf{0} \quad \text{for} \quad i\neq j,$$

so that

$$Cov(\mathbf{y}_i^{new},\mathbf{y}_j^{new})=\mathbf{0} \quad \text{for} \quad i\neq j.$$

The standard F test is then applied to the transformed data to test $H_0:\mathbf{C}\boldsymbol{\beta}=\mathbf{b}$. The new test is based on

$$F_{new}=\frac{(\mathbf{C}\boldsymbol{\beta}^{new}-\mathbf{b})'(\mathbf{X}_C^{new\prime}\mathbf{X}_C^{new})^{-1}(\mathbf{C}\boldsymbol{\beta}^{new}-\mathbf{b})/q}{(\mathbf{y}^{new}-\mathbf{X}^{new}\boldsymbol{\beta}^{new})'(\mathbf{y}^{new}-\mathbf{X}^{new}\boldsymbol{\beta}^{enw})/(n-I-k)},$$

where $\boldsymbol{\beta}^{new}=(\mathbf{X}^{new\prime}\mathbf{X}^{new})^{-1}\mathbf{X}^{new\prime}\mathbf{y}^{new}$ and $\mathbf{X}_C^{new}=\mathbf{X}^{new}(\mathbf{X}^{new\prime}\mathbf{X}^{new})^{-1}\mathbf{C}'$. Under the null hypothesis, F_{new} has an exact F distribution with q and $n-I-k$ df. It is interesting to note that F_{new} does not depend on ρ, unlike $F_A(\rho)$ and $F_{GLS}(\rho)$. Of course, this good feature of our test is achieved at a cost of sacrificing I degrees of freedom. However, our simulation shows that this does not seem to affect the efficiency of our test when the number of clusters I is small.

5. A Monte Carlo simulation

In this section, we revisit the simulation experiments conducted by Wu et al. (1988) and Rao et al. (1993). They considered the following nested error regression model with two covariates, x and z, and $n_i=n_0$:

$$y_{ij}=\beta_0+\beta_1x_{ij}+\beta_2z_{ij}+v_i+u_{ij}, \quad \text{for } i=1,2,\ldots,I;\ \ j=1,2,\ldots,n_0, \tag{4}$$

where v_i's are iid $N(0,\sigma_v^2)$, u_{ij}'s are iid $N(0,\sigma_u^2)$, and $\sigma^2=\sigma_v^2+\sigma_u^2$.

Following Wu et al. (1988) and Rao et al. (1993), we generate (x_{ij},z_{ij}) from a bivariate normal distribution with additional random effects components that allow for intra-cluster correlations ρ_x and ρ_z on both x and z:

$$x_{ij}=\mu_x+v_{xi}+u_{xij}, \quad z_{ij}=\mu_z+v_{zi}+u_{zij}, \tag{5}$$

where v_{xi} iid $N(0,\sigma_{vx}^2)$, v_{zi} iid $N(0,\sigma_{vz}^2)$, u_{xi} iid $N(0,\sigma_{ux}^2)$, u_{zi} iid $N(0,\sigma_{uz}^2)$, $\rho_x=\sigma_{vx}^2/\sigma_x^2$, $\rho_z=\sigma_{vz}^2/\sigma_z^2$, $\sigma_x^2=\sigma_{vx}^2+\sigma_{ux}^2$, and $\sigma_z^2=\sigma_{vz}^2+\sigma_{uz}^2$. Furthermore, v_{xi} and v_{zi} are correlated with covariance σ_{vxz}, and u_{xij} and u_{zij} are correlated with covariance σ_{uxz}. Let

$$\rho_{xz}=\sigma_{vxz}/\sigma_x\sigma_z \quad \text{and} \quad corr(x,z)=\sigma_{xz}/\sigma_x\sigma_z,$$

where $\sigma_{xz}=\sigma_{vxz}+\sigma_{uxz}$ and $corr(x,z)$ denotes the correlation between x_{ij} and z_{ij}. The parameters σ_{vx}^2, σ_{vxz}, σ_{ux}^2, σ_{uxz}, etc. are chosen to satisfy $\sigma_x^2=\sigma_z^2=20$, $\rho_x=0.1$, $\rho_z=0.5$, $\rho_{xz}=0$, and $corr(x,z)=-0.33$. We fix $I=10$, $\mu_x=100$, $\mu_z=200$, $\beta_0=10$ and $\sigma^2=10$. Both Wu et al. (1988) and Rao et al. (1993) used these parameter values in their simulation study.

For given (x_{ij},z_{ij}), we generate y_{ij} from model (4) with selected ρ and n_0 given in Tables 2 and 3. The simulated data $\{(y_{ij},x_{ij},z_{ij}),j=1,2,\ldots,n_0;i=1,2,\ldots,I\}$ are used to compute the test statistics. For each simulation condition, given in Tables 2 and 3, we consider 1000 simulation runs in order to obtain size estimates for two different null hypotheses: (1) $H_0:\beta_1=0$ and (2) $H_0:\beta_2=0$. Tables 2 and 3 report simulated sizes of the standard F statistic, F_s, the GLS F statistic with known ρ, $F_{GLS}(\rho)$, the GLS F statistic with estimated $\hat{\rho}$, $F_{GLS}(\hat{\rho})$, and new statistic F_{new}. We have not included $F_A(\hat{\rho})$, because it has been shown to perform inferior to $F_{GLS}(\hat{\rho})$ by Rao et al. (1993). We use the nominal level at $\alpha=0.05$ for each simulation condition. All the simulated test sizes are subject to Monte Carlo simulation errors and thus one must apply caution in interpreting results presented in Tables 2 and 3. For 1000 simulation runs, the standard error of a simulated test size $\hat{\alpha}$ is given by $\sqrt{\hat{\alpha}(1-\hat{\alpha})/1000}$. Since most of the simulated test sizes are reasonably close to 5%, we compute the simulation error only for $\hat{\alpha}=0.05$ and consider a simulated test

Table 1
The proportions of zero intra-cluster correlation estimates for different simulation conditions given in Table 3.

ρ	$I \times n_0$		
	10×15	6×25	3×50
0.0	0.703	0.733	0.865
0.05	0.305	0.302	0.501
0.1	0.119	0.131	0.304
0.3	0.003	0.008	0.105
0.5	0.001	0.002	0.049

Table 2
Simulated sizes (%) of F_s, $F_{GLS}(\rho)$, $F_{GLS}(\hat{\rho})$ and F_{new} tests of (1) $H_0 : \beta_1 = 0$ and (2) $H_0 : \beta_2 = 0$ (nominal level = 0.05, $I = 10$).

n_0	ρ	$H_0 : \beta_1 = 0$				$H_0 : \beta_2 = 0$			
		F_{new}	$F_{GLS}(\rho)$	$F_{GLS}(\hat{\rho})$	F_s	F_{new}	$F_{GLS}(\rho)$	$F_{GLS}(\hat{\rho})$	F_s
10	0.0	0.057	0.051	0.052	0.051	0.043	0.053	0.050	0.053
	0.1	0.048	0.053	0.061	**0.068**[a]	0.057	0.055	**0.068**[a]	**0.112**[a]
	0.3	0.043	0.048	0.048	**0.086**[a]	0.061	0.054	**0.081**[a]	**0.212**[a]
	0.5	0.060	0.061	0.063	**0.109**[a]	0.042	0.053	0.061	**0.322**[a]
5	0.0	0.034	0.051	0.054	0.051	0.051	0.039	0.038	0.039
	0.1	0.047	0.047	0.058	0.061	0.048	0.044	0.062	**0.076**[a]
	0.3	0.058	0.042	0.056	**0.097**[a]	0.045	0.056	**0.088**[a]	**0.145**[a]
	0.5	0.054	0.052	**0.067**[a]	**0.153**[a]	0.055	0.050	**0.076**[a]	**0.188**[a]
2	0.0	0.052	0.049	**0.069**[a]	0.049	0.051	0.037	0.058	0.037
	0.1	0.049	0.058	**0.078**[a]	0.064	0.056	0.054	**0.076**[a]	0.055
	0.3	0.039	0.045	**0.069**[a]	0.056	0.043	0.047	**0.064**[a]	0.047
	0.5	0.042	0.049	**0.074**[a]	0.060	0.058	0.057	**0.092**[a]	0.062

[a]All the simulated test sizes where the nominal level is not maintained (> 0.064) are bolded.

Table 3
Simulated sizes (%) of F_s, $F_{GLS}(\rho)$, $F_{GLS}(\hat{\rho})$ and F_{new} tests of (1) $H_0 : \beta_1 = 0$ and (2) $H_0 : \beta_2 = 0$ (nominal level = 0.05, $I \times n_0 = 150$).

$I \times n_0$	ρ	$H_0 : \beta_1 = 0$				$H_0 : \beta_2 = 0$			
		F_{new}	$F_{GLS}(\rho)$	$F_{GLS}(\hat{\rho})$	F_s	F_{new}	$F_{GLS}(\rho)$	$F_{GLS}(\hat{\rho})$	F_s
10×15	0.0	0.056	0.055	0.055	0.055	0.05	0.043	0.041	0.043
	0.05	0.044	0.044	0.055	0.06	0.059	0.055	**0.083**[a]	**0.118**[a]
	0.1	0.041	0.049	0.054	**0.076**[a]	0.046	0.062	**0.086**[a]	**0.154**[a]
	0.3	0.048	0.048	0.055	**0.173**[a]	0.063	**0.071**[a]	**0.092**[a]	**0.29**[a]
	0.5	0.042	0.04	0.043	**0.237**[a]	0.03	0.037	0.042	**0.41**[a]
6×25	0.0	0.058	0.056	0.057	0.056	0.051	0.049	0.047	0.049
	0.05	0.042	0.042	0.056	**0.073**[a]	0.045	0.038	0.055	**0.08**[a]
	0.1	0.053	0.052	**0.065**[a]	**0.106**[a]	0.049	0.047	0.059	**0.129**[a]
	0.3	0.062	0.06	0.064	**0.232**[a]	0.036	0.038	0.042	**0.268**[a]
	0.5	0.038	0.037	0.047	**0.324**[a]	0.038	0.043	0.048	**0.395**[a]
3×50	0.0	0.06	**0.069**[a]	**0.07**[a]	**0.069**[a]	0.059	0.035	0.04	0.035
	0.05	0.055	0.046	**0.082**[a]	**0.101**[a]	0.042	0.042	**0.078**[a]	**0.126**[a]
	0.1	0.049	0.05	**0.07**[a]	**0.169**[a]	0.053	0.06	**0.099**[a]	**0.263**[a]
	0.3	0.052	0.053	0.061	**0.32**[a]	0.048	0.048	**0.068**[a]	**0.455**[a]
	0.5	0.041	0.043	0.053	**0.472**[a]	0.042	0.042	0.058	**0.578**[a]

[a]All the simulated test sizes where the nominal level is not maintained (> 0.064) are bolded.

size inflated when it exceeds $0.05 + 1.96 \times \sqrt{0.05(1-0.05)/1000} = 0.064$. Following Table 4 of Graubard and Korn (1993), we use boldface for all the simulated test sizes that are inflated. We have taken this approach in order to simplify the presentation of our results.

In Table 2, we fix number of clusters $I(=10)$ and consider different cluster sizes. Thus, this table helps us to understand the effect of the total sample size $n = I \times n_0$. It is clear from Table 2 that both $F_{GLS}(\rho)$ and F_{new} tests maintain the nominal level for every parameter setting. The remaining test statistics F_s and $F_{GLS}(\hat{\rho})$ have inflated levels for many cases, especially when ρ and/or n_0 are large. When ρ is zero, most test statistics maintain the nominal level except for $F_{GLS}(\hat{\rho})$ (observed level = 0.069) when the cluster size n_0 is small. This probably results from the poor estimation of ρ with a total sample of size 20 only.

In Table 3, we consider different number of clusters (i.e., I) for a fixed total sample size at $n = 150$. The purpose is to study the effect of the number of clusters on the sizes of different tests. We have three different combinations of $I \times n_0 : 10 \times 15, 6 \times 25$, and 3×50. Again, we observe that $F_{GLS}(\rho)$ and F_{new} maintain the nominal level in most cases. The test F_s leads to the highest inflated value when $I \times n_0 = 3 \times 50$, i.e., when the cluster size is the largest. This is consistent with our observation in Example 2 of Section 2. In that example, we noted that the size of F_s increases as the interviewer workload becomes large.

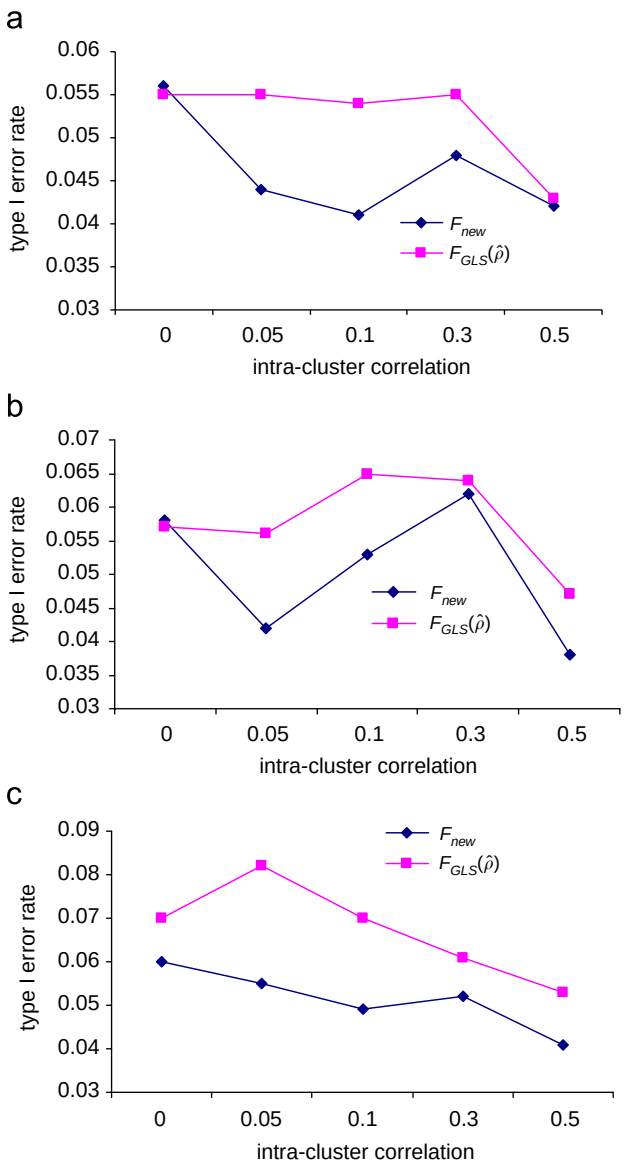

Fig. 3. Type I error rate vs. intra-cluster correlation for $F_{GLS}(\hat{\rho})$ and F_{new} tests of $H_0 : \beta_1 = 0$, nominal 5% level, $I \times n_0 = 150$. (a–c) Correspond to the number of clusters of 10, 6, and 3, respectively, for a fixed total sample size $n = 150$.

Figs. 3 and 4 display different plots of the size vs. intra-cluster correlation ρ for $F_{GLS}(\hat{\rho})$ and F_{new} tests of (1) $H_0 : \beta_1 = 0$ and (2) $H_0 : \beta_2 = 0$, respectively. These plots provide a good visual comparison of F_{new} and $F_{GLS}(\hat{\rho})$. It is clear that our new test outperforms $F_{GLS}(\hat{\rho})$ in terms of controlling the size of the test. The test $F_{GLS}(\hat{\rho})$ is doing a poor job when the number of clusters is small ($I = 3$) and ρ is not very large. This is probably because of the poor estimation of ρ (see Table 3). Note that when ρ is close to zero, the estimator of ρ does not perform very well (see Table 1). When $\rho = 0.1$ size of $F_{GLS}(\hat{\rho})$ could be as high as 0.099.

6. Concluding remarks

In this paper, we consider a new adjustment to the standard F test in the presence of the intra-cluster correlation. Unlike some previous adjustments proposed in the literature, our test does not involve estimation of the unknown intra-cluster correlation. The estimation of the intra-cluster correlation is problematic, especially when the number cluster is small and/or when the intra-cluster correlation is small, a situation often encountered in practice. Our test should be useful in such a situation. The proposed test extended the previous work that is free of intra-cluster correlation to a set of clusters of different sizes. For illustration purpose, the new test was derived under the assumption of equi-correlated error terms for a univariate mixed model. Further

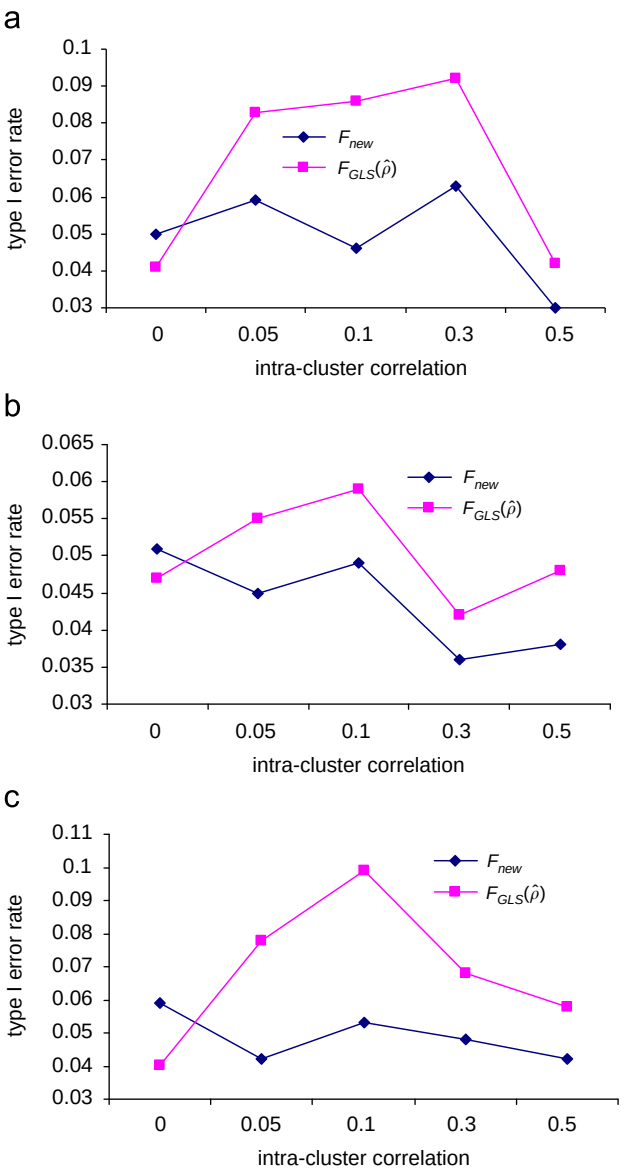

Fig. 4. Type I error rate vs. intra-cluster correlation for $F_{GLS}(\hat{\rho})$ and F_{new} tests of $H_0 : \beta_2 = 0$, nominal 5% level, $I \times n_0 = 150$. (a–c) Correspond to the number of clusters of 10, 6, and 3, respectively, for a fixed total sample size $n = 150$.

extension of this work to multivariate mixed model while loosening of the condition of equi-correlation could be a potentially good problem for future research.

Acknowledgment

The authors are grateful to an anonymous referee for his/her comments on earlier drafts, and to Professor Rahul Mukerjee for helpful discussions.

References

Campbell, C., 1977. Properties of ordinary and weighted least squares estimators for two stage samples. In: Proceedings of the Social Statistics Section, American Statistical Association, pp. 800–805.

Collins, M., Butcher, B., 1982. Interviewer and clustering effects in an attitude survey. Journal of the Market Research Society 25, 39–58.

Feather, J., 1973. A study of interviewer variance. WHO International Collaborative Study of Medical Care Utilization, Saskatchewan Study Area Reports, Series II, Monograph No. 3.

Fellegi, I.P., 1964. Response variance and its estimation. Journal of the American Statistical Association 59, 1016–1041.
Fuller, W.A., Battese, G.E., 1973. Transformations for estimation of linear models with nested error structures. Journal of the American Statistical Association 68, 626–632.
Graubard, B.I., Korn, E.L., 1993. Hypothesis testing with complex survey data: the use of classical quadratic test statistics with particular reference to regression problems. Journal of the American Statistical Association 88, 629–641.
Gray, P.G., 1956. Examples of interviewer variability taken from two sample surveys. Applied Statistics V, 73–85.
Groves, R.M., 1989. Survey Errors and Survey Costs. Wiley, New York.
Henderson, C.R., 1953. Estimation of variance and covariance components. Biometrics 9, 226–252.
Hansen, M.H., Hurwitz, W.N., Bershad, M.A., 1961. Measurement errors in censuses and surveys. Bulletin of the International Statistical Institute 38, 359–374.
Hyunh, H., Feldt, L.S., 1970. Conditions under which mean square ratios in repeated measurements designs have exact *F*-distributions. Journal of the American Statistical Association 65, 1582–1589.
Jiang, J., 1996. REML estimation: asymptotic behavior and related topics. Annals of Statistics 24, 255–286.
Kish, L., 1962. Studies of interviewer variance for attitudinal variables. Journal of the American Statistical Association 57, 92–115.
Rao, J.N.K., Sutradhar, B.C., Yue, K., 1993. Generalized least squares *F* test in regression analysis with two-stage cluster samples. Journal of the American Statistical Association 88, 1388–1391.
Rouanet, H., Lepine, D., 1970. Comparison between treatments in a repeated-measurement design: ANOVA and multivariate methods. British Journal of Mathematical and Statistical Psychology 23, 147–163.
Schnell, R., Kreuter, F., 2005. Separating interviewer and sampling point effects. Journal of Official Statistics 21, 389–410.
Scott, A.J., Holt, D., 1982. The effect of two-stage sampling on ordinary least squares methods. Journal of the American statistical Association 77, 848–854.
Thomas, D.R., 1983. Univariate repeated measures techniques applied to multivariate data. Psychometrika 48, 451–464.
Wu, C.F.J., Holt, D., Holmes, D.J., 1988. The effect of two-stage sampling on the *F* statistics. Journal of the American Statistical Association 83, 150–159.