Sankhyā : The Indian Journal of Statistics
2006, Volume 68, Part 3, pp. 392-408

Testing Treatment Effects in Two-Way Linear Models: Additive or Full Model?

Bin Cheng
Columbia University, USA
Jun Shao
University of Wisconsin-Madison, USA

Abstract

Under a two-way analysis of variance/covariance model, we consider the problem of testing the main treatment effect (a fixed effect of primary interest) when the interaction between the treatment and the other factor (which is either fixed or random) is practically negligible but not exactly zero. Although the theory for analysis of variance/covariance is well-developed (at least for the fixed effects models), practitioners are not clear on whether the test based on additive model (assuming no interaction) or the test based on full model (including interaction terms) should be adopted. The use of additive model is motivated by a possible gain in the power of the test. On the other hand, the use of full model addresses the concern of having an inflated size of the test when the interaction is not exactly zero. Under balanced fixed effects models, we show that the test based on additive model has correct size even if the additive model is wrong but its power may be very low in the presence of a small interaction effect; contrary to common beliefs, in many practical situations the gain in power by using the additive approach is not substantial even if the additive model is correct. Under unbalanced fixed effects models or balanced/unbalanced mixed effects models, the test based on additive model generally has an inflated size unless the additive model is correct.

AMS (2000) *subject classification.* Primary 62J10.
Keywords and phrases. Analysis of variance/covariance, mixed effects, unbalanced models, interaction effect, exact tests, size and power.

1 Introduction

Consider a two-way linear model such as the classical two-way analysis of variance (ANOVA) model where one factor is the main focus of the study (which will be referred to as the main treatment factor) and the other factor is not of primary interest such as a block effect (which will be referred to as the secondary factor). In a multicenter clinical trial, for example, the primary interest is the effect of a drug treatment and the secondary factor is the center. If there is a large interaction between the main treatment and the secondary factor, then the assessment and interpretation of the main treatment effect is difficult. On the other hand, in many applications the interaction effect is reasonably small and can be ignored although it is not exactly 0, especially when the secondary factor is a blocking factor and/or the study is carefully designed and performed. The interaction effect can be assessed using some statistical tests (e.g., Searle, 1971; Gail and Simon, 1985; Cheng and Shao, 2005). In multicenter clinical trials, assessing the treatment-by-center interaction is usually the first step in statistical analysis (International Conference on Harmonization (ICH) Guideline E9, 1998). The main issue addressed in this article is how to test the main treatment effect after we conclude that the interaction effect is negligible.

Since the interaction effect may not be exactly 0, we may include the interaction terms in the ANOVA model for testing the main treatment effect. This is referred to as the *full model* approach. An alternative is to use an *additive model*, i.e., the ANOVA model without any interaction term. This is referred to as the additive model approach, which is clearly justified if the interaction effect is exactly 0 (i.e., the additive model is a correct model). Even if the interaction effect is not exactly 0, some practitioners think that the additive model approach is more efficient; for example, it is stated in ICH (1998) that "the routine inclusion of interaction terms in the model reduces the efficiency of the test for the main effects". On the other hand, there is a concern of having an inflated testing size under the additive model approach when interaction effect is not exactly 0. In multicenter clinical trials, which approach should be used to test the drug treatment effect is a controversial issue. More discussions on the design and analysis in multicenter clinical trials can be found in Fleiss (1986), Källén (1997), Senn (1998), Jones et al. (1998), Gould (1998), Lin (1999), and Gallo (2000).

Does the test based on the additive model approach have an inflated size when interaction effect is not exactly 0? This question is not answered in the existing literature even though the theory of two-way linear model is well-developed. For the traditional *balanced* two-way linear models where the effects of both factors and their interaction are treated as fixed effects, our answer is no (Section 2), i.e., the test based on additive model still has the right size even if the additive model is not correct. In this case, the comparison of the two approaches should be based on the power. Contrary to many people's expectation, our result shows that the test based on additive model is not substantially more powerful than the test based on full model when the interaction is exactly or nearly 0. In fact, we show that the test based on additive model is not unbiased (in the sense that its power may be lower than the size) and its power can be much lower than the other test when there is a small but nonzero interaction. When the model is *unbalanced*, however, we show that whether the size of the test based on additive model is inflated depends on which null hypothesis is used.

Mixed effects linear models have received a great deal of attention in recent years. When there is a large number of centers in a multicenter clinical trial, centers are often treated as random effects. In Section 3 we consider mixed effects models where the main treatment effect is fixed but the secondary factor and interaction effects are random. The result is different from that in the fixed effects model: the test based on additive model has an inflated size when the additive model is wrong, regardless of whether the model is balanced or unbalanced.

In some applications, demographic variables and baseline characteristics may be useful covariates included in a two-way analysis of covariance (ANCOVA) model. In Section 4, we extend our results to ANCOVA models. A summary of our main findings is given in the last section.

2 Two-Way Fixed Effects ANOVA Models

Let y_{ijk} denote the kth observation of the ith level of the main treatment factor and the jth level of the secondary factor. In this section, we consider the following two-way analysis of variance (ANOVA) model with interaction:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \quad i = 1, \ldots, I, \qquad j = 1, \ldots, J, \\ k = 1, \ldots, n_{ij}, \tag{2.1}$$

where ε_{ijk} are independent and identically distributed normal random variables with mean 0 and variance σ^2, and μ, α_i, β_j, γ_{ij}, and σ^2 are unknown parameters. The treatment effect α_i, secondary effect β_j, and interaction γ_{ij} are unknown but fixed and satisfy the usual restrictions $\bar{\alpha}. = 0$, $\bar{\beta}. = 0$, $\bar{\gamma}_{i\cdot} = 0$, and $\bar{\gamma}_{\cdot j} = 0$, where, for any given variable x, $\bar{x}$ denotes an average and a dot is used in the subscript to denote averaging over the indicated subscript, e.g., $\bar{x}. = I^{-1} \sum_{i=1}^{I} x_i$ and $\bar{x}_{i\cdot} = J^{-1} \sum_{j=1}^{J} x_{ij}$.

We consider the following hypotheses for the main treatment effect:

$$\begin{aligned} H_0 &: \alpha_1 = \cdots = \alpha_I = 0 \quad \text{versus} \\ H_1 &: \text{some } \alpha_i\text{'s are not } 0. \end{aligned} \tag{2.2}$$

When we know that the interaction effect is practically negligible, it is tempting to use the following additive model:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}, \quad i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, n_{ij}. \tag{2.3}$$

The use of the additive model (2.3) is clearly suggested in ICH (1998), with the hope that it increases the power of testing hypotheses (2.2), compared with the use of the full model (2.1). However, the additive model (2.3) is a wrong model unless $\gamma_{ij} = 0$ for all i and j and, in practice, treatment-by-center interaction is rarely zero. Hence, we address the following two issues in this section:

1. Does the test based on model (2.3) have a wrong size under the full model (2.1) with nonzero γ-terms?
2. Is the test based on model (2.3) more powerful than that based on the full model (2.1)?

It is well known that, under the general full model (2.1) with usual restrictions, testing hypotheses (2.2) corresponds to the type III analysis (Speed and Hocking, 1976). Test statistics can be formed using the $R(\ )$ notation as described in Searle (1971) and Speed and Hocking (1976). Define $R(\mu, \alpha, \beta, \gamma)$ to be the reduction in the total sum of squares due to fitting model (2.1), $R(\mu, \alpha, \beta)$ to be the reduction in the total sum of squares due to fitting model (2.3), $R(\mu, \beta, \gamma)$ to be the reduction in the total sum of squares due to fitting model (2.1) with $\alpha_i = 0$, $R(\mu, \beta)$ to be the reduction in the total sum of squares due to fitting model (2.3) with $\alpha_i = 0$, $R(\gamma|\mu, \alpha, \beta) = R(\mu, \alpha, \beta, \gamma) - R(\mu, \alpha, \beta)$, $R(\alpha|\mu, \beta, \gamma) = R(\mu, \alpha, \beta, \gamma) - R(\mu, \beta, \gamma)$, and $R(\alpha|\mu, \beta) = R(\mu, \alpha, \beta) - R(\mu, \beta)$.

Under the full model approach, the type III test rejects H_0 in (2.2) if and only if

$$F_{\mathrm{A}} = \frac{R(\alpha|\mu, \beta, \gamma)/(I-1)}{\mathrm{SSE}/(N - IJ)} > F_{I-1, N-IJ, \alpha}, \tag{2.4}$$

where N is the total number of observations, $F_{m,l,a}$ is the $(1-a)$th quantile of the F-distribution with degrees of freedom (m, l), a is a given level of significance, and

$$\mathrm{SSE} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}}(y_{ijk} - \bar{y}_{ij\cdot})^2.$$

Under the additive model approach, we reject H_0 in (2.2) if and only if

$$\tilde{F}_{\mathrm{A}} = \frac{R(\alpha|\mu,\beta)/(I-1)}{[\mathrm{SSE} + R(\gamma|\mu,\alpha,\beta)]/(N-I-J+1)} > F_{I-1,N-I-J+1,\alpha}. \tag{2.5}$$

When the model is balanced in the sense that $n_{ij} = n$ for all i and j,

$$R(\gamma|\mu,\alpha,\beta) = \mathrm{SSAB} = n\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots})^2$$

and

$$R(\alpha|\mu,\beta) = \mathrm{SSA} = nJ\sum_{i=1}^{I}(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots})^2.$$

The test statistic $\tilde{F}_{\mathrm{A}}$ in (2.5), however, is more comparable with the type II test statistic $\frac{R(\alpha|\mu,\beta)/(I-1)}{\mathrm{SSE}/(N-IJ)}$ under the full model approach, which is intended for testing the following null hypothesis (Searle, 1971, p. 308):

$$H_0 : \sum_{j=1}^{J}\left(n_{ij} - \frac{n_{ij}^2}{n_{\cdot j}}\right)(\alpha_i + \gamma_{ij}) = \sum_{l\neq i}\left(\sum_{j=1}^{J}\frac{n_{ij}n_{lj}}{n_{\cdot j}}\right)(\alpha_l + \gamma_{lj}), \quad i = 1,\ldots,I-1. \tag{2.6}$$

The null hypothesis in (2.6) is the same as the null hypothesis in (2.2) when the model is balanced; otherwise, they may be different. In multicenter clinical trials, usually the null hypothesis in (2.2) is tested, not the null hypothesis in (2.6).

The following result concerns the size and the power of the test rule (2.5) when model (2.3) is wrong.

Let $\boldsymbol{\gamma} = (\gamma_{11},\ldots,\gamma_{I1},\ldots,\gamma_{1J},\ldots,\gamma_{IJ})'$, $\|\ \|$ be the Euclidean norm, $\mathbf{A}' = (\mathbf{A}_1',\cdots,\mathbf{A}_J')$ with

$$\mathbf{A}_j = \begin{pmatrix} n_{1j} - \frac{n_{1j}^2}{n_{\cdot j}} & -\frac{n_{2j}n_{1j}}{n_{\cdot j}} & \cdots & -\frac{n_{(I-1)j}n_{1j}}{n_{\cdot j}} \\ -\frac{n_{1j}n_{2j}}{n_{\cdot j}} & n_{2j} - \frac{n_{2j}^2}{n_{\cdot j}} & \cdots & -\frac{n_{(I-1)j}n_{2j}}{n_{\cdot j}} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{n_{1j}n_{(I-1)j}}{n_{\cdot j}} & -\frac{n_{2j}n_{(I-1)j}}{n_{\cdot j}} & \cdots & n_{(I-1)j} - \frac{n_{(I-1)j}^2}{n_{\cdot j}} \\ -\frac{n_{1j}n_{Ij}}{n_{\cdot j}} & -\frac{n_{2j}n_{Ij}}{n_{\cdot j}} & \cdots & -\frac{n_{(I-1)j}n_{Ij}}{n_{\cdot j}} \end{pmatrix},$$

$\mathbf{\Lambda} = \mathrm{diag}(n_{11}^{-1}, \ldots, n_{I1}^{-1}, \ldots, n_{1J}^{-1}, \ldots, n_{IJ}^{-1})$,

$$\mathbf{L} = \begin{pmatrix} \mathbf{J}'_{J-1} \\ -\mathbf{I}_{J-1} \end{pmatrix} \otimes \begin{pmatrix} \mathbf{J}'_{I-1} \\ -\mathbf{I}_{I-1} \end{pmatrix}.$$

Further, suppose that $\mathbf{I}_a$ denotes the identity matrix of order a, $\mathbf{J}_b$ denotes the b-vector of ones, and $\otimes$ denotes the Kronecker product for matrices.

THEOREM 2.1. *Assume model (2.1).*

(i) For testing the null hypothesis in (2.6), the test rule (2.5) has exactly size α.

(ii) For testing the null hypothesis in (2.2), the size of the test rule (2.5) is at least α and strict inequality holds when

$$F_{I-1,N-I-J+1,\alpha} < \rho = \lim_{\|\boldsymbol{\gamma}\|\to\infty} \frac{\boldsymbol{\gamma}'\mathbf{A}(\mathbf{A}'\mathbf{\Lambda}\mathbf{A})^{-1}\mathbf{A}'\boldsymbol{\gamma}/(I-1)}{\boldsymbol{\gamma}'\mathbf{L}(\mathbf{L}'\mathbf{\Lambda}\mathbf{L})^{-1}\mathbf{L}'\boldsymbol{\gamma}/(N-I-J+1)}, \tag{2.7}$$

(iii) When the model is balanced, H_0 in (2.6) is the same as H_0 in (2.2) and, hence, the test rule (2.5) has exactly size α.

(iv) In any case, the test rule (2.5) is not unbiased in the sense that its power may be lower than α.

PROOF. From Searle (1971, formulas (60) and (69) in §7.2), $R(\alpha|\mu,\beta) = \bar{\mathbf{y}}'\mathbf{A}\mathbf{T}\mathbf{A}'\bar{\mathbf{y}}$, where

$$\bar{\mathbf{y}} = (\bar{y}_{11.}, \ldots, \bar{y}_{I1.}, \ldots, \bar{y}_{1J.}, \ldots, \bar{y}_{IJ.})'$$

is the vector of cell sample means and $\mathbf{T}^{-1} = \mathrm{Var}(\mathbf{A}'\bar{\mathbf{y}})/\sigma^2 = \mathbf{A}'\mathbf{\Lambda}\mathbf{A}$. Hence,

$$R(\alpha|\mu,\beta) = \bar{\mathbf{y}}'\mathbf{A}(\mathbf{A}'\mathbf{\Lambda}\mathbf{A})^{-1}\mathbf{A}'\bar{\mathbf{y}}. \tag{2.8}$$

Similarly, by Searle (1971, formula (93) in §7.2),

$$R(\gamma|\mu,\alpha,\beta) = \bar{\mathbf{y}}'\mathbf{L}(\mathbf{L}'\mathbf{\Lambda}\mathbf{L})^{-1}\mathbf{L}'\bar{\mathbf{y}}. \tag{2.9}$$

(i) It follows from formulas (2.8) and (2.9) that, under H_0 in (2.6), the numerator and denominator of $\tilde{F}_{\mathrm{A}}$ in (2.5) are independently distributed as a central chi-square with degrees of freedom $I-1$ and a noncentral chi-square with degrees of freedom $N - I - J + 1$ and noncentrality parameter $\delta = \boldsymbol{\gamma}'\mathbf{L}(\mathbf{L}'\boldsymbol{\Lambda}\mathbf{L})^{-1}\mathbf{L}'\boldsymbol{\gamma}/\sigma^2$, respectively. Then,

$$\begin{aligned}
&\sup_{H_0 \text{ in (2.6) holds}} P\left(\tilde{F}_{\mathrm{A}} > F_{I-1,N-I-J+1,\alpha}\right) \\
&\quad = \sup_{\delta} P_\delta\left(\tilde{F}_{\mathrm{A}}^{-1} < F^{-1}_{I-1,N-I-J+1,\alpha}\right) \\
&\quad = P_{\delta=0}\left(\tilde{F}_{\mathrm{A}}^{-1} < F^{-1}_{I-1,N-I-J+1,\alpha}\right) \\
&\quad = \alpha,
\end{aligned}$$

where the first equality follows from the fact that $\tilde{F}_{\mathrm{A}}^{-1}$ has the noncentral F-distribution with degrees of freedom $(N - I - J + 1, I - 1)$ and noncentrality parameter δ, the second equality follows from the monotone property of the noncentral F-distribution, and the last equality follows from $F^{-1}_{I-1,N-I-J+1,\alpha} = F_{N-I-J+1,I-1,1-\alpha}$. Thus, the test rule (2.5) has size α for testing (2.6), even if model (2.3) is wrong.

(ii) Under H_0 in (2.2), the numerator and denominator of $\tilde{F}_{\mathrm{A}}$ in (2.5) are independently distributed as noncentral chi-square with noncentrality parameters $\lambda = \boldsymbol{\gamma}'\mathbf{A}(\mathbf{A}'\boldsymbol{\Lambda}\mathbf{A})^{-1}\mathbf{A}'\boldsymbol{\gamma}/\sigma^2$ and δ, respectively. When $\lambda = \delta = 0$, $P(\tilde{F}_{\mathrm{A}} > F_{I-1,N-I-J+1,\alpha}) = \alpha$. Hence, the size of test rule (2.5) for testing (2.2) is at least α. Let W be the numerator of $\tilde{F}_{\mathrm{A}}$. Then $E(W) = \sigma^2[1 + \lambda/(I - 1)]$ and $\mathrm{Var}(W) = \sigma^4[2 + 4\lambda/(I - 1)]/(I - 1)$. Using Chebyshev's inequality, we can show that W/λ converges in probability to $\sigma^2/(I - 1)$ as $\lambda \to \infty$. Similarly, we can show that the denominator of $\tilde{F}_{\mathrm{A}}$ divided by δ converges in probability to $\sigma^2/(N - I - J + 1)$ as $\delta \to \infty$. Consequently, $\tilde{F}_{\mathrm{A}}$ converges in probability to ρ in (2.7) as $\|\boldsymbol{\gamma}\| \to \infty$. If $\rho > F_{I-1,N-I-J+1,\alpha}$, then $P(\tilde{F}_{\mathrm{A}} > F_{I-1,N-I-J+1,\alpha}) \to 1$ and, therefore, the size of test rule (2.5) for testing (2.2) is strictly larger than α.

(iii) The result in this part is a consequence of (i), since the null hypotheses in (2.2) and (2.6) are the same when the model is balanced.

(iv) In general, the numerator and denominator of $\tilde{F}_{\mathrm{A}}$ in (2.5) are independently distributed as noncentral chi-square with noncentrality parameters d and δ, respectively, where $d = E(\bar{\mathbf{y}}')\mathbf{A}(\mathbf{A}'\boldsymbol{\Lambda}\mathbf{A})^{-1}\mathbf{A}'E(\bar{\mathbf{y}})/\sigma^2$. Thus, the power of the test is a continuous function of δ and d. Denote this function by $\phi(d, \delta)$. From the monotone property of the noncentral F-distribution, $\phi(0, \delta) < \phi(0, 0) = \alpha$ when $\delta > 0$. Since ϕ is continuous, there exists a $d > 0$ so that $\phi(d, \delta) < \alpha$. This shows that test rule (2.5) is not unbiased. □

Thus, using an additive model when it is in fact wrong does not inflate the size for testing (2.6) or testing (2.2) when the model is balanced. The problem of the test rule (2.5), however, is that its power may be lower than α (not unbiased). In the unbalanced case, the test rule in (2.5) has size α for testing (2.6), but may have an inflated size for testing (2.2).

Consider testing (2.2), which is usually the main focus in multicenter clinical trials. From the proof of Theorem 2.1, if $\rho > F_{I-1,N-I-J+1,\alpha}$, the size of test rule (2.5) can be arbitrarily close to 1; if $\rho = 0$ (which holds when the model is balanced), then the size of test rule (2.5) is α. When $0 < \rho \leq F_{I-1,N-I-J+1,\alpha}$, an explicit form of the size of test rule (2.5) is difficult to obtain. However, the following example indicates that the size may still be larger than α. We consider the case where $I = 2$, $J = 3$, $(n_{11}, n_{21}, n_{12}, n_{22}, n_{13}, n_{23}) = (9, 8, 8, 9, 5, 4)$, and $(\gamma_{11}, \gamma_{21}, \gamma_{12}, \gamma_{22}, \gamma_{13}, \gamma_{23}) = (0.5, -0.5, 0.3, -0.3, -0.8, 0.8)$. At level $\alpha = 0.05$, $\rho = 3.61 < 4.09 = F_{I-1,N-I-J+1,\alpha} = F_{1,39,0.05}$, but our simulation based on 100,000 runs shows that the size of the test rule (2.5) is at least 0.1089.

We now compare the power of the two tests F_{A} and $\tilde{F}_{\mathrm{A}}$ in the balanced case, since in the balanced case they have the same size regardless of whether the additive model is correct or not. Note that $\tilde{F}_{\mathrm{A}}$ has a larger denominator degree of freedom than does F_{A}, which leads to the impression of having more power. For some values of δ and d (defined in the proof of Theorem 2.1), we computed the ratio of the power of $\tilde{F}_{\mathrm{A}}$ over the power of F_{A}. The results for $I = 2$ and some values of n and J are shown in Figure 1. The following are our findings.

1. When $\delta = 0$, which is the most favourable case for $\tilde{F}_{\mathrm{A}}$, the power ratio is larger than 1, but is between 1 and 1.10 when $(n, J) = (2, 3)$, between 1 and 1.05 when $(n, J) = (2, 5)$, and close to 1 in all cases where n is larger than or equal to 5.
2. The power of $\tilde{F}_{\mathrm{A}}$ decreases as δ increases (as expected) and the decrease is sharper for smaller d values.
3. The difference between F_{A} and $\tilde{F}_{\mathrm{A}}$ diminishes quickly when any of n, J, and d is not very small.

Thus, we conclude that (at least in the balanced case with $I = 2$), using the additive model does not have a substantial power increase over using the full model as expected (ICH, 1998). Although its size is not inflated, the test based on additive model is not recommended because its power can be substantially low even for a small value of nonzero δ.

3 Two-Way Mixed Effects ANOVA Models

Consider the two-way mixed effects model

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \quad i = 1, \ldots, I, j = 1, \ldots, J, k = 1, \ldots, n_{ij},$$
$$\beta_j \sim N(0, \sigma^2_\beta), \quad \gamma_{ij} \sim N(0, \sigma^2_\gamma), \quad \varepsilon_{ijk} \sim N(0, \sigma^2), \tag{3.1}$$
$$\beta_j\text{'s}, \ \gamma_{ij}\text{'s, and } \varepsilon_{ijk}\text{'s are independent},$$

where μ is an unknown parameter and α_i's are fixed treatment effects satisfying $\bar{\alpha}. = 0$. The hypotheses for treatment effects are given by (2.2), regardless of whether the model is balanced or not.

Although we consider problems parallel to Section 2, there are some differences between the results under fixed effects and mixed effects models.

When the model is balanced ($n_{ij} = n$ for all i and j), Test rule (2.4) under the full model approach has a size arbitrarily close to 1, because, under H_0 in (2.2), $F_{\mathrm{A}}/(1+n\sigma^2_\gamma/\sigma^2)$ has the central F-distribution with degrees of freedom $(I-1, N-IJ)$ and σ^2_γ/σ^2 can be arbitrarily large. Thus, a commonly used unbiased test of size α under the full (balanced) model approach rejects H_0 if and only if

$$\frac{\mathrm{SSA}/(I-1)}{\mathrm{SSAB}/(I-1)(J-1)} > F_{I-1,(I-1)(J-1),\alpha}.$$

When the mixed effects model (3.1) is unbalanced, it is well known that testing hypotheses (2.2) under the full model approach is difficult, because none of $R(\alpha|\mu)$, $R(\alpha|\mu,\beta)$, and $R(\alpha|\mu,\beta,\gamma)$ is chi-square distributed, and none of them is independent of $R(\gamma|\mu,\alpha,\beta)$. Gallo and Khuri (1990) derived a size α test for hypotheses (2.2) based on the earlier work of Khuri and Littell (1987). However, the test is complicated because it involves determination of a sequence of orthogonal matrices and computation of the maximum eigenvalue of some matrix.

Thus, under unbalanced mixed effects models, one may be tempted to apply the additive model (model (3.1) with $\sigma^2_\gamma = 0$) approach which uses test rule (2.5) for testing (2.2) for its simplicity, regardless of whether the model is balanced or not. Unlike the result in Section 2, however, the following result shows that test rule (2.5) has an inflated size when σ^2_γ is not exactly 0 unless there is no replicates ($n_{ij} = 1$ for all i and j, in which case the full model is essentially the same as the additive model).

THEOREM 3.1. *Assume model (3.1).*

(i) The size of test rule (2.5) *for testing* (2.2) *is*

$$\sup_{\delta \geq 0} P\left(h(\delta) > F_{I-1,N-I-J+1,\alpha} \right), \tag{3.2}$$

where

$$h(\delta) = \frac{\mathbf{e}'(\mathbf{\Lambda} + \delta \mathbf{I}_{IJ})^{1/2}\mathbf{A}(\mathbf{A}'\mathbf{\Lambda}\mathbf{A})^{-1}\mathbf{A}'(\mathbf{\Lambda} + \delta \mathbf{I}_{IJ})^{1/2}\mathbf{e}/(I-1)}{[\chi^2_{N-IJ} + \mathbf{e}'(\mathbf{\Lambda}+\delta\mathbf{I}_{IJ})^{1/2}\mathbf{L}(\mathbf{L}'\mathbf{\Lambda}\mathbf{L})^{-1}\mathbf{L}'(\mathbf{\Lambda}+\delta\mathbf{I}_{IJ})^{1/2}\mathbf{e}]/(N-I-J+1)}, \tag{3.3}$$

χ^2_m *denotes the central chi-square random variable with degree of freedom* m, $\delta = \sigma^2_\gamma/\sigma^2$, $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}_{IJ})$ *and is independent of* χ^2_{N-IJ}, *and* $\mathbf{\Lambda}$, $\mathbf{A}$ *and* $\mathbf{L}$ *are defined in* (2.8) *and* (2.9).

(ii) The size in (3.2) *is at least* α, *and it is larger than* α *if*

$$P\left(\frac{\mathbf{e}'\mathbf{A}(\mathbf{A}'\mathbf{\Lambda}\mathbf{A})^{-1}\mathbf{A}'\mathbf{e}/(I-1)}{\mathbf{e}'\mathbf{L}(\mathbf{L}'\mathbf{\Lambda}\mathbf{L})^{-1}\mathbf{L}'\mathbf{e}/(N-I-J+1)} > F_{I-1,N-I-J+1,\alpha} \right) > \alpha, \tag{3.4}$$

which is further implied by

$$\frac{(I-1)(J-1)\max\{n_{ij}\}}{(N-I-J+1)\min\{n_{ij}\}} F_{I-1,N-I-J+1,\alpha} < F_{I-1,(I-1)(J-1),\alpha}, \tag{3.5}$$

where $F_{m,l}$ *denotes a random variable having the central F-distribution with degrees of freedom* (m, l).

(iii) When the model is balanced ($n_{ij} = n$ *for all* i *and* j), *the size of test rule (2.5) is*

$$\tilde{\alpha} = P\left(F_{I-1,(I-1)(J-1)} > \frac{(I-1)(J-1)}{N-I-J+1} F_{I-1,N-I-J+1,\alpha} \right),$$

which is larger than α *when* $n > 1$ *and* $\sigma^2_\gamma > 0$, *and equal to* α *when* $n = 1$.

(iv) When $\sigma^2_\gamma = 0$ *(i.e., the additive model is correct), test rule* (2.5) *has size* α.

PROOF.

(i) Under null hypothesis (2.2), $\bar{\mathbf{y}} \sim N(\mu \mathbf{J}_{IJ}, \sigma^2 \mathbf{\Lambda} + \sigma_\gamma^2 \mathbf{I}_{IJ} + \sigma_\beta^2 \mathbf{I}_J \otimes \mathbf{J}_I \mathbf{J}_I')$. Since $\mathbf{A}_j' \mathbf{J}_I = \mathbf{0}$ for $j = 1, \ldots, J$, $\mathbf{A}'\mathbf{J}_{IJ} = (\mathbf{A}_1'\mathbf{J}_I, \ldots, \mathbf{A}_J'\mathbf{J}_I) = \mathbf{0}$ and $\mathbf{A}'(\mathbf{I}_J \otimes \mathbf{J}_I\mathbf{J}_I') = (\mathbf{A}_1'\mathbf{J}_I\mathbf{J}_I', \ldots, \mathbf{A}_J'\mathbf{J}_I\mathbf{J}_I') = \mathbf{0}$. From $E(\mathbf{A}'\bar{\mathbf{y}}) = \mu \mathbf{A}'\mathbf{J}_{IJ} = \mathbf{0}$ and $\mathrm{Var}(\mathbf{A}'\bar{\mathbf{y}}) = \sigma^2 \mathbf{A}'\mathbf{\Lambda}\mathbf{A} + \sigma_\gamma^2 \mathbf{A}'\mathbf{A}$,

$$\mathbf{A}'\bar{\mathbf{y}} \sim N(\mathbf{0}, \sigma^2 \mathbf{A}'\mathbf{\Lambda}\mathbf{A} + \sigma_\gamma^2 \mathbf{A}'\mathbf{A}).$$

Similarly, since $(\mathbf{J}_{I-1}, -\mathbf{I}_{I-1})\mathbf{J}_I = \mathbf{0}$, we have

$$\mathbf{L}'\mathbf{J}_{IJ} = \mathbf{L}'(\mathbf{J}_J \otimes \mathbf{J}_I) = ((\mathbf{J}_{J-1}, -\mathbf{I}_{J-1})\mathbf{J}_J) \otimes ((\mathbf{J}_{I-1}, -\mathbf{I}_{I-1})\mathbf{J}_I) = \mathbf{0}$$
$$\text{and } \mathbf{L}'(\mathbf{I}_J \otimes \mathbf{J}_I) = ((\mathbf{J}_{J-1}, -\mathbf{I}_{J-1})((\mathbf{J}_{I-1}, -\mathbf{I}_{I-1})\mathbf{J}_I) = \mathbf{0},$$

which imply

$$\mathbf{L}'\bar{\mathbf{y}} \sim N(\mathbf{0}, \sigma^2 \mathbf{L}'\mathbf{\Lambda}\mathbf{L} + \sigma_\gamma^2 \mathbf{L}'\mathbf{L}).$$

Let $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}_{IJ})$ and be independent of $\bar{\mathbf{y}}$. By checking the means and covariance matrix, we conclude that the joint distribution of $\mathbf{A}'\bar{\mathbf{y}}$ and $\mathbf{L}'\bar{\mathbf{y}}$ is the same as that of $\sigma \mathbf{A}'(\mathbf{\Lambda} + \delta \mathbf{I})^{1/2}\mathbf{e}$ and $\sigma \mathbf{L}'(\mathbf{\Lambda} + \delta \mathbf{I})^{1/2}\mathbf{e}$. Since $\mathrm{SSE} \sim \sigma^2 \chi^2_{N-IJ}$ and is independent of $\bar{\mathbf{y}}$, the distribution of $h(\delta)$ in (3.3) is the same as that of

$$\frac{\bar{\mathbf{y}}'\mathbf{A}(\mathbf{A}'\mathbf{\Lambda}\mathbf{A})^{-1}\mathbf{A}'\bar{\mathbf{y}}/(I-1)}{[\mathrm{SSE} + \bar{\mathbf{y}}'\mathbf{L}(\mathbf{L}'\mathbf{\Lambda}\mathbf{L})^{-1}\mathbf{L}'\bar{\mathbf{y}}]/(N-I-J+1)}$$

Then the result follows from (2.8)-(2.9).

(ii) The size in (3.2) is at least $P(h(0) > F_{I-1,N-I-J+1,\alpha})$, which is equal to α (see the proof of (iv)). The size in (3.2) is also at least $P(h(\infty) > F_{I-1,N-I-J+1,\alpha})$. Since

$$h(\infty) \sim \frac{\mathbf{e}'\mathbf{A}(\mathbf{A}'\mathbf{\Lambda}\mathbf{A})^{-1}\mathbf{A}'\mathbf{e}/(I-1)}{\mathbf{e}'\mathbf{L}(\mathbf{L}'\mathbf{\Lambda}\mathbf{L})^{-1}\mathbf{L}'\mathbf{e}/(N-I-J+1)},$$

(3.4) implies that the size in (3.2) is larger than α. Finally,

$$\frac{\mathbf{e}'\mathbf{A}(\mathbf{A}'\mathbf{\Lambda}\mathbf{A})^{-1}\mathbf{A}'\mathbf{e}/(I-1)}{\mathbf{e}'\mathbf{L}(\mathbf{L}'\mathbf{\Lambda}\mathbf{L})^{-1}\mathbf{L}'\mathbf{e}/(N-I-J+1)}$$
$$\geq \frac{\min\{n_{ij}\}\mathbf{e}'\mathbf{\Lambda}^{1/2}\mathbf{A}(\mathbf{A}'\mathbf{\Lambda}\mathbf{A})^{-1}\mathbf{A}'\mathbf{\Lambda}^{1/2}\mathbf{e}/(I-1)}{\max\{n_{ij}\}\mathbf{e}'\mathbf{\Lambda}^{1/2}\mathbf{L}(\mathbf{L}'\mathbf{\Lambda}\mathbf{L})^{-1}\mathbf{L}'\mathbf{\Lambda}^{1/2}\mathbf{e}/(N-I-J+1)},$$

which has the same distribution as

$$\frac{(N-I-J+1)\min\{n_{ij}\}}{(I-1)(J-1)\max\{n_{ij}\}}F_{I-1,(I-1)(J-1)},$$

since $\mathbf{L}'\mathbf{\Lambda}\mathbf{A} = \mathbf{0}$ so that $\mathbf{e}'\mathbf{\Lambda}^{1/2}\mathbf{A}(\mathbf{A}'\mathbf{\Lambda}\mathbf{A})^{-1}\mathbf{A}'\mathbf{\Lambda}^{1/2}\mathbf{e}$ and $\mathbf{e}'\mathbf{\Lambda}^{1/2}\mathbf{L}(\mathbf{L}'\mathbf{\Lambda}\mathbf{L})^{-1}\mathbf{L}'\mathbf{\Lambda}^{1/2}\mathbf{e}$ are independent chi-square random variables. Thus, (3.4) is implied by (3.5).

(iii) When the model is balanced,

$$h(\delta) = \frac{(1+n\delta)\chi^2_{I-1}/(I-1)}{[\chi^2_{N-IJ} + (1+n\delta)\chi^2_{(I-1)(J-1)}]/(N-I-J+1)},$$

where χ^2_{I-1}, χ^2_{N-IJ}, $\chi^2_{(I-1)(J-1)}$ are independent. Note that $h(\delta)$ is an increasing function of δ with $h(0) \sim F_{I-1,N-I-J+1}$ and

$$h(\infty) \sim \frac{N-I-J+1}{(I-1)(J-1)}F_{I-1,(I-1)(J-1)}.$$

Then, the size of rule (2.5)

$$\begin{aligned}
&= \sup_{\delta} P\left(h(\delta) > F_{I-1,N-I-J+1,\alpha}\right)\\
&= P\left(h(\infty) > F_{I-1,N-I-J+1,\alpha}\right)\\
&= P\left(\frac{N-I-J+1}{(I-1)(J-1)}F_{I-1,(I-1)(J-1)} > F_{I-1,N-I-J+1,\alpha}\right)\\
&= \tilde{\alpha}.
\end{aligned}$$

Since $h(\delta)$ is increasing in δ, $\tilde{\alpha} > \alpha$ when $\sigma^2_\gamma > 0$ and $n > 1$, and $\tilde{\alpha} = \alpha$ when $N-I-J+1 = (I-1)(J-1)$, i.e., $n = 1$ (no replicates).

(iv) Since $\mathbf{L}'\mathbf{\Lambda}\mathbf{A} = \mathbf{0}$, the numerator and denominator of $h(\delta)$ are independent when $\delta = 0$ and, therefore, $h(0) \sim F_{I-1,N-I-J+1}$. Hence, the size in (3.2) $= \alpha$ when $\sigma^2_\gamma = 0$. □

Hence, the use of additive model may inflate the size for testing (2.2) when the additive model is wrong. This is different from the result in Theorem 2.1 where misusing the additive model does not inflate the size.

In the unbalanced case, the size in (3.2) does not have an explicit form, because the numerator and denominator of $h(\delta)$ in (3.3) are not independent when $\delta > 0$. Both (3.4) and (3.5) are sufficient conditions under which the size of test rule (2.5) is inflated (they are satisfied when the model is balanced unless $n = 1$). Condition (3.5) is easier to check. For example, when $I = 2, J = 3$ and $(n_{11}, n_{21}, n_{12}, n_{22}, n_{13}, n_{23}) = (9, 8, 8, 9, 5, 4)$, (3.5) holds. However, we still do not know the magnitude of the size inflation. A simulation of 100,000 runs shows that the size of test rule (2.5) with $\alpha = 0.05$ is at least 0.7089. To see that condition (3.5) is only sufficient, we consider the case where $I = 2, J = 4$ and all n_{ij}'s are equal to 1 except that $n_{24} = 2$. In this case, condition (3.5) does not hold but the size of test rule (2.5) with $\alpha = 0.05$ is at least 0.0966 based on a simulation of 100,000 runs.

When the additive model is correct $\delta = 0$, it is expected that test rule (2.5) has size α. We provide a proof of such a result (Theorem 3.1(iv)) that cannot be found in the literature.

4 Two-Way Models With Covariates

In some applications, there are covariates such as demographic variables and baseline characteristics. Including covariates that are related to the response variable reduces error variability and, hence, increases the power of various tests.

We consider the following popular two-way analysis of covariance (ANCOVA) model:

$$y_{ijk} = \mu+\alpha_i+\beta_j+\gamma_{ij}+\boldsymbol{\eta}'\mathbf{z}_{ijk}+\varepsilon_{ijk}, \quad i = 1, \ldots, I, j = 1, \ldots, J, k = 1, \ldots, n_{ij}, \tag{4.1}$$

where $\mathbf{z}_{ijk}$'s are q-dimensional covariate vectors, $\boldsymbol{\eta}$ is a q-dimensional unknown parameter vector, and ε_{ijk}'s are random errors having a normal distribution $N(0, \sigma^2)$. The assumptions on α_i's β_j's, and γ_{ij}'s in model (4.1) are the same as those under the fixed effects model (2.1) or under mixed effects model (3.1).

Let $\mathbf{y}$ be the vector formed by listing y_{ijk} in the order of j, i, and k. Then model (4.1) can be written in the matrix form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{Z}\boldsymbol{\eta} + \boldsymbol{\varepsilon},$$

where $\mathbf{X}$ is the usual design matrix for the two-way ANOVA model, $\mathbf{Z}$ is the design matrix containing $\mathbf{z}_{ijk}$'s, $\boldsymbol{\theta} = (\mu, \alpha_1, \cdots, \alpha_I, \beta_1, \cdots, \beta_J, \gamma_{11}, \cdots, \gamma_{IJ})'$, and $\boldsymbol{\varepsilon}$ is the error vector. The least squares estimator of $\boldsymbol{\eta}$ is

$$\hat{\boldsymbol{\eta}} = [\mathbf{Z}'(\mathbf{I} - \mathbf{P_X})\mathbf{Z}]^{-1}\mathbf{Z}'(\mathbf{I} - \mathbf{P_X})\mathbf{y},$$

where $\mathbf{P_X} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$ (see Searle, 1971, formula (11) in §8.2). Define

$$u_{ijk} = y_{ijk} - \hat{\boldsymbol{\eta}}'\mathbf{z}_{ijk}$$

and

$$\bar{\mathbf{u}} = (\bar{u}_{11\cdot}, \dots, \bar{u}_{I1\cdot}, \dots, \bar{u}_{1J\cdot}, \dots, \bar{u}_{IJ\cdot})',$$

which can be called the adjusted cell mean vector.

LEMMA 4.1. *Assume model (4.1) with either fixed effects or mixed effects. Let SSE be the error sum of squares with y_{ijk}'s replaced by u_{ijk}'s. Then $\bar{\mathbf{u}}$ and SSE are independent.*

PROOF. Let $\mathbf{P}_{(\mathbf{X},\mathbf{Z})}$ denote the projection onto the column space associated with the matrix $(\mathbf{X}, \mathbf{Z})$. Then SSE $= \boldsymbol{\varepsilon}'(\mathbf{I} - \mathbf{P}_{(\mathbf{X},\mathbf{Z})})\boldsymbol{\varepsilon}$. Note that $\hat{\boldsymbol{\eta}} = \boldsymbol{\eta}+[\mathbf{Z}'(\mathbf{I} - \mathbf{P_X})\mathbf{Z}]^{-1}\mathbf{Z}'(\mathbf{I} - \mathbf{P_X})\boldsymbol{\varepsilon}$. Then $\hat{\boldsymbol{\eta}}$ and $(\mathbf{I} - \mathbf{P}_{(\mathbf{X},\mathbf{Z})})\boldsymbol{\varepsilon}$ are independent since $\mathbf{Z}'(\mathbf{I} - \mathbf{P_X})(\mathbf{I} - \mathbf{P}_{(\mathbf{X},\mathbf{Z})}) = \mathbf{Z}'(\mathbf{I} - \mathbf{P}_{(\mathbf{X},\mathbf{Z})}) = \mathbf{0}$. Furthermore, $\bar{\varepsilon}_{ij\cdot}$ is independent of $(\mathbf{I} - \mathbf{P}_{(\mathbf{X},\mathbf{Z})})\boldsymbol{\varepsilon}$ since $(\mathbf{I} - \mathbf{P}_{(\mathbf{X},\mathbf{Z})})\mathbf{P_X} = \mathbf{P_X} - \mathbf{P}_{(\mathbf{X},\mathbf{Z})}\mathbf{P_X} = \mathbf{0}$. Then, the result follows from $\bar{u}_{ij\cdot} = \mu + \alpha_i + \beta_j + \gamma_{ij} + (\boldsymbol{\eta} - \hat{\boldsymbol{\eta}})'\bar{\mathbf{z}}_{ij\cdot} + \bar{\varepsilon}_{ij\cdot}$.

Note that $\bar{\mathbf{u}}$ is normally distributed with covariance matrix

$$\mathrm{Var}(\bar{\mathbf{u}}) = (\mathbf{V} + \boldsymbol{\Lambda})\ \sigma^2$$

for fixed effects models and

$$\mathrm{Var}(\bar{\mathbf{u}}) = (\mathbf{I}_J \otimes (\mathbf{J}_I\mathbf{J}_I'))\ \sigma_\beta^2 + \mathbf{I}_{IJ}\ \sigma_\gamma^2 + (\mathbf{V} + \boldsymbol{\Lambda})\ \sigma^2$$

for mixed effects models,

$$\mathbf{V} = \bar{\mathbf{z}}'[\mathbf{Z}'(\mathbf{I} - \mathbf{P_X})\mathbf{Z}]^{-\mathbf{1}}\bar{\mathbf{z}}$$

and $\bar{\mathbf{z}} = (\bar{\mathbf{z}}_{11\cdot}, \dots, \bar{\mathbf{z}}_{I1\cdot}, \dots, \bar{\mathbf{z}}_{1J\cdot}, \dots, \bar{\mathbf{z}}_{IJ\cdot})'$. It can be seen that results in the previous sections are derived based on the key fact that SSA, SSAB, and various R's are quadratic functions of $\bar{\mathbf{y}}$, which is independent of SSE. Under the ANCOVA model, the adjusted cell mean vector $\bar{\mathbf{u}}$ plays the same role as $\bar{\mathbf{y}}$. These results and Lemma 1 ensure that the results in Sections 2 and 3 are still applicable to the two-way ANCOVA models with the following modifications: (i) $\bar{\mathbf{y}}$ should be replaced by $\bar{\mathbf{u}}$; (ii) SSE should be defined as $\mathbf{y}'(\mathbf{I} - \mathbf{P}_{(\mathbf{X},\mathbf{Z})})\mathbf{y} = \boldsymbol{\varepsilon}'(\mathbf{I} - \mathbf{P}_{(\mathbf{X},\mathbf{Z})})\boldsymbol{\varepsilon}$; and (iii) the degree of freedom for SSE, which appears as denominator degree of freedom in some tests, should be changed from $N - IJ$ to $N - IJ - q$ due to estimation of $\boldsymbol{\eta}$. □

5 Summary

Under two-way balanced/unbalanced fixed/mixed effects ANOVA/ ANCOVA models, we address the issue of testing the main treatment effect (the factor of primary interest) when the interaction effect is practically negligible but not exactly 0. We study two ways of performing the test, the additive model approach and the full model approach. Our findings are:

1. When all effects are fixed (non-random) and the model is balanced, the test based on additive model has correct size even if the additive model is wrong, but its power may be very low in the presence of a small interaction effect. Even if the additive model is correct or nearly correct, the gain in using the additive approach may not be substantial.

2. When all effects are fixed and the model is unbalanced, the test based on additive model generally has an inflated size for testing hypotheses (2.2).

3. When the main treatment effects are fixed but the other effects are random so that the model is mixed-effect, the use of the additive model results in an inflated size unless the additive model is exactly correct, regardless of whether the model is balanced or not.

Based on our findings, we recommend the full model approach. The additive model approach is not suitable under mixed effects models. Under fixed effects models, the test based on additive model is not recommended because of its inflated size (under unbalanced models) and/or its low power.

Acknowledgments. We would like to thank the referees for helpful comments and suggestions.

References

Cheng, B. and Shao, J. (2006). Exact tests for treatment-by-center interaction in clinical trials. *Statist. Sinica*, to appear.

Fleiss, J.L. (1986). Analysis of data from multiclinic trials. *Controlled Clinical Trials*, **7**, 267–275.

Gail, M. and Simon, R. (1985). Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*, **41**, 361–372.

Gallo, J. and Khuri, A.I. (1990). Exact tests for the random and fixed effects in an unbalanced mixed two-way cross-classification model. *Biometrics*, **46**, 1087–1095.

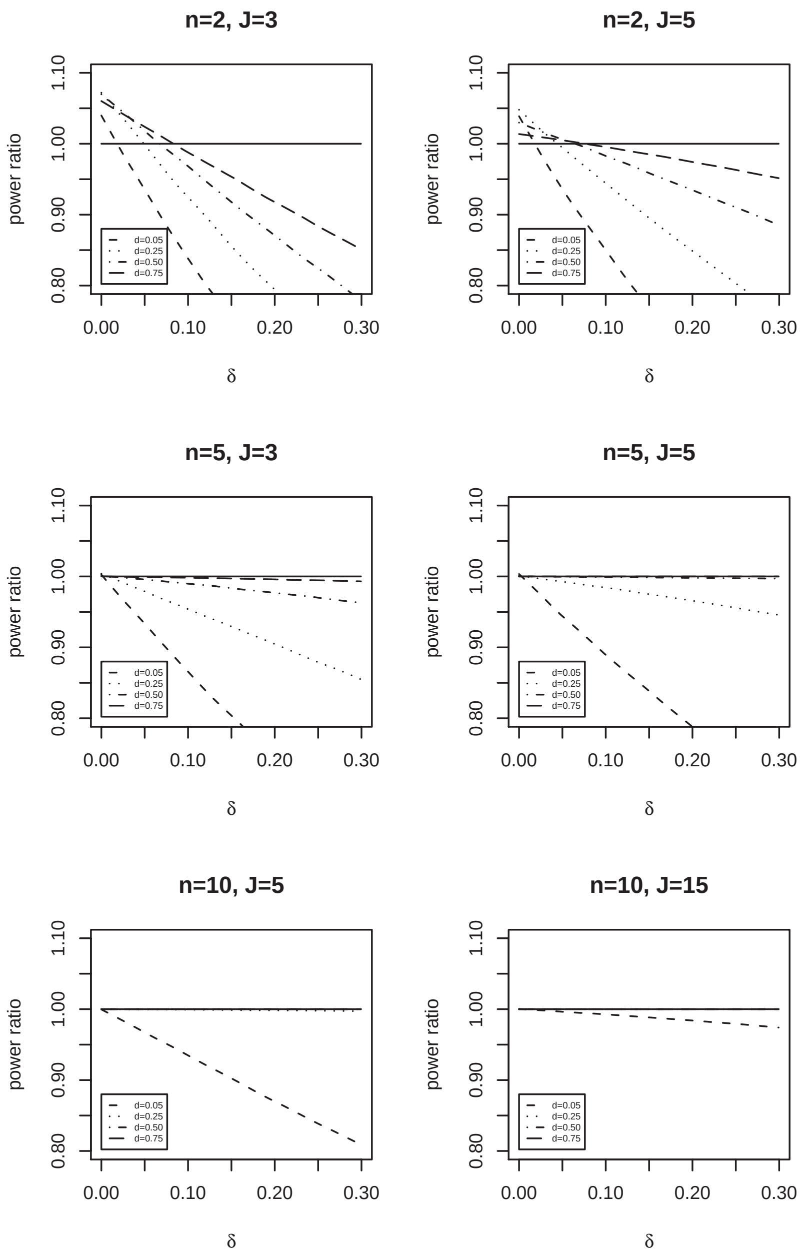

Figure 1: Ratio of the power of the test rule (2.5) over the power of the test rule (2.4), $I = 2$, $d = \sum_i \alpha_i^2/(\sigma^2 I)$, $\delta = \sum_{i,j} \gamma_{ij}^2/(\sigma^2 IJ)$

Gallo, P. (2000). Center-weighting issues in multicenter clinical trials. *J. Biopharm. Statist.*, **10**, 145–163.

Gould, A.L. (1998). Multi-centre trial analysis revisited. *Stat. Med.*, **17**, 1779–1797.

ICH (1998). *Statistical Principles for Clinical Trials.* Tripartite International Conference on Harmonization Guideline, E9. (http://www.ich.org/LOB/media/MEDIA485.pdf).

Jones, B., Teather, D., Wang, J. and Lewis, J.A. (1998). A comparison of various estimators of a treatment difference for a multi-centre clinical trial. *Stat. Med.*, **17**, 1767–1777.

Källén, A. (1997). Treatment-by-center interaction: What is the issue? *Drug Information Journal*, **31**, 927–936.

Khuri, A. I. and Littell, R.C. (1987). Exact tests for the main effects variance components in an unbalanced random two-way model. *Biometrics*, **43**, 545–560.

Lin, Z. (1999). An issue of statistical analysis in controlled multi-Centre studies: How shall we weight the centres? *Stat. Med.*, **18**, 365–373.

Searle, S.R. (1971). *Linear Models.* Wiley, New York.

Senn, S. (1998). Some controversies in planning and analysing multi-centre trials. *Stat. Med.*, **17**, 1753–1765.

Speed, F.M. and Hocking, R.R. (1976). The use of the $R(\ )$-notation with unbalanced data. *Amer. Statist.*, **30**, 30–33.

Bin Cheng
Department of Biostatistics
Mailman School of Public Health
Columbia University
722 West 168th Street, Room 640
New York, NY 10032, USA
E-mail: bc2159@columbia.edu

Jun Shao
Department of Statistics
University of Wisconsin-Madison
Medical Science Center
1300 University Avenue
Madison, WI 53706, USA
E-mail: shao@stat.wisc.edu

Paper received June 2004; revised May 2006.