*Biostatistics* (2003), **4**, 3, *pp.* 371–383
*Printed in Great Britain*

# Scaled marginal models for multiple continuous outcomes

JASON ROY
*Center for Statistical Sciences, Box G-H, Brown University, Providence, RI 02912, USA*
jroy@stat.brown.edu

XIHONG LIN
*Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109, USA*

LOUISE M. RYAN
*Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115, USA*

## SUMMARY

In studies that involve multivariate outcomes it is often of interest to test for a common exposure effect. For example, our research is motivated by a study of neurocognitive performance in a cohort of HIV-infected women. The goal is to determine whether highly active antiretroviral therapy affects different aspects of neurocognitive functioning to the same degree and if so, to test for the treatment effect using a more powerful one-degree-of-freedom global test. Since multivariate continuous outcomes are likely to be measured on different scales, such a common exposure effect has not been well defined. We propose the use of a scaled marginal model for testing and estimating this global effect when the outcomes are all continuous. A key feature of the model is that the effect of exposure is represented by a common effect size and hence has a well-understood, practical interpretation. Estimating equations are proposed to estimate the regression coefficients and the outcome-specific scale parameters, where the correct specification of the within-subject correlation is not required. These estimating equations can be solved by repeatedly calling standard generalized estimating equations software such as SAS PROC GENMOD. To test whether the assumption of a common exposure effect is reasonable, we propose the use of an estimating-equation-based score-type test. We study the asymptotic efficiency loss of the proposed estimators, and show that they generally have high efficiency compared to the maximum likelihood estimators. The proposed method is applied to the HIV data.

*Keywords*: Asymptotic relative efficiency; Effect size; Estimating equations; Global effect; Multivariate response; Score test.

## 1. INTRODUCTION

A goal of many studies is to assess the effect of an exposure on multiple related outcomes. For example, the effect of prenatal exposure of some agent on multiple measures of birth defects may be of interest in teratology studies (Sammel and Ryan, 1996); in reproductive health studies, interest may lie in the effect of pesticide exposure on multiple measures of semen quality (Lin *et al.*, 2000). Often interest is in not only in whether the exposure affects the outcomes, but also whether the exposure affects each outcome to the same degree (i.e. whether there is a common exposure effect). If so, one would be interested in developing a more powerful one-degree-of-freedom global test for the common effect and to estimate it.

For example, in this paper we consider the effect of highly active antiretroviral therapy (HAART) on several measures of neurocognitive functioning in HIV-infected women. The study consisted of 125 women, 55 of whom were on HAART therapy and 70 of whom were not (Cohen *et al.*, 2001). HIV infection can affect both cognitive and psychomotor ability. Therefore, neurocognitive performance was measured using three instruments: Color Trail Making (CTM), Controlled Oral Word Association (COWAT) and Grooved Pegboard (GPB). COWAT is designed to measure verbal fluency; CTM and GPB measure psychomotor speed; all three measure information processing to some extent. Interest is in whether HAART treatment enhances neurocognitive functioning, which was measured by performance on the three tasks. Whether or not HAART affects the different aspects of neurocognitive functioning, e.g. psychomotor speed and verbal fluency, to the same degree is also an important substantive question. This knowledge would provide insight not only into the disease process, but also into the specific benefits of treatment. This implies a need for both tests of global exposure effects (i.e. does the exposure affect the outcomes?) and common exposure effects (i.e. does the exposure affect all of the outcomes to the same degree?). If the assumption on the common exposure effect is plausible, one would be interested in testing for the common exposure effect using a more powerful one-degree-of-freedom test and estimating it. In addition, because women were not randomized to treatment, adjustment for confounders in a regression setting is crucial.

Several approaches have been proposed for testing in problems involving multiple outcomes. Traditional techniques such as Hotelling's $T^2$ statistic, multivariate analysis of variance (MANOVA) and Bonferroni adjustment of univariate tests (Wichern and Johnson, 2002) are unsatisfactory for a variety of reasons, including that there is a lack of flexibility in the types of hypotheses that can be tested and that these tests are often inefficient (O'Brien, 1984). O'Brien (1984) and Pocock *et al.* (1987) proposed global tests that better address the questions of interest (e.g. by allowing the alternative hypothesis to have treatment effects in the same direction for all outcomes), and tend to be more efficient. However, these techniques cannot be easily adapted to observational studies where there is a need to adjust for confounding in a regression setting. In addition, these approaches do not allow one to test for a common effect of exposure.

Estimating and testing for a common exposure effect on multiple continuous outcomes poses a special difficulty in that the outcomes are typically measured on different scales. Recently, Sammel and Ryan (1996) proposed a latent variable regression model where it is assumed that the outcomes all measure some underlying latent variable. The effect of exposure on this latent variable is then of interest, and can be thought of as the common effect. However, because the marginal mean and variance depend on common parameters, the latent variable model is highly sensitive to model misspecification (Sammel and Ryan, 2002). Further, the interpretation of the hypothetical latent variable is not always clear. To address these issues, Sammel *et al.* (1999) proposed a multivariate linear mixed model (MLMM) that disentangles the mean and variance parameters. Lin *et al.* (2000) extended the MLMM by allowing a more flexible covariance structure. In both cases the fact that each outcome is measured on a different scale was addressed by scaling the outcomes by the standard deviation of the error term, which is the residual term not accounted for by the random effects. A common exposure effect on these scaled outcomes can be estimated and tested. However, a disadvantage of this approach is that the correlation between the outcomes must be correctly specified. In addition, the interpretation of the common effect is not attractive because the outcomes are not scaled by their total standard deviations, but by part of the total standard deviations. Hence it does not have the conventional effect size interpretation.

To address these concerns, we propose a scaled marginal model for estimating a common exposure effect on multiple continuous outcomes. One of the advantages of this approach is that the correlation between outcomes does not have to be correctly specified in order to obtain consistent estimates. In addition, the exposure effect has the attractive interpretation of being a common effect size. In Section 2 we specify the model. In Section 3 we describe the estimation procedure. Tests of common exposure are presented in Sections 3.2 and 3.3. We study the asymptotic efficiency of the estimators in Section 4. The application to the HIV data is presented in Section 5. Finally, there is a discussion in Section 6.

## 2. The scaled marginal model for multiple continuous outcomes

### 2.1 *Review of the scaled linear mixed model*

Suppose that $M$ continuous outcomes $y_i = (y_{i1}, \dots, y_{iM})^T$, an exposure variable $w_i$ and a vector of covariates $x_i$ are observed for the $i$th of $n$ subjects. If $y_i$ was a vector of repeated measurements of the same outcome (as in longitudinal studies), then an attractive analytic approach would be to fit a linear mixed model. However, when $y_i$ represents a vector of different response variables measured at a common time, the traditional mixed model approach is not satisfactory. This is largely because the $M$ responses are likely measured on different scales, which makes defining a common exposure effect within this framework difficult. As a result, Lin *et al.* (2000) proposed the following scaled linear mixed model for the $j$th outcome $y_{ij}$ $(j = 1, \dots, M)$

$$\frac{y_{ij}}{\tau_j} = x_i^T \beta_j + w_i \alpha + z_{ij}^T b_i + e_{ij}, \tag{2.1}$$

where $z_{ij}$ is a design vector, $b_i$ is a vector of random effects following $N\{0, D(\theta)\}$, $\tau_j$ is the standard deviation of $\{y_{ij}|x_i, w_i, b_i\}$, and the $e_{ij}$ are errors distributed as $N(0, 1)$. Like linear mixed models for longitudinal data, within-subject correlation is modeled using random effects. The primary difference is model (2.1) first scales the responses. This model is attractive because it enables the estimation of a global treatment effect on the scaled outcomes. Therefore, one can easily test for an exposure effect using a one-degree-of-freedom test (i.e. test whether $\alpha = 0$). In addition, the mean is not overly restrictive in that the effect of $x_i$ on $y_{ij}$ is not constrained to be constant across outcomes $j$.

However, there are several drawbacks of this approach. First, notice that var$(y_{ij}|b_i) = \tau_j^2$ and the marginal variance var$(y_{ij}) = \tau_j^2 z_{ij}^T D(\theta) z_{ij} + \tau_j^2$. Therefore, the outcomes are standardized by the conditional standard deviations given $b_i$, as opposed to the marginal standard deviations of $y_{ij}$. In other words, the outcomes are standardized by part of the total standard deviation. This is unattractive because the regression parameters do not have effect size interpretations in the conventional way and would cause difficulties in explaining the results to practitioners. Second, depending on the form of $z_{ij}^T b_i$, the marginal variance of $y_{ij}/\tau_j$, which is equal to var$(y_{ij}/\tau_j) = z_{ij}^T D(\theta) z_{ij} + 1$, may not be constant across outcomes. Hence this does not fulfill the goal of standardization—to make the standardized outcomes have a common variance 1. Finally, consistency of the estimators of $\beta$, $\alpha$ and $\tau$ requires the covariance matrix $V_i$ of $y_i = (y_{i1}, \dots, y_{iM})^T$ to be correctly specified. As a result, estimation of the regression parameters might not be robust to misspecification of the correlation matrix. This is because if $V_i$ is misspecified, equation (7) of Lin *et al.* (2000) (the score equations for $\tau^2$) does not have zero expectation. Note that this property is different from that of standard linear mixed models, where misspecification of the correlation structure only affects the efficiency but not the consistency of the regression parameter estimators.

### 2.2 *The scaled marginal model*

We are interested in a model that keeps some of the attractive features of the scaled linear mixed model, while addressing the shortcomings. First, scaling by the total standard deviation of $y_{ij}$, as opposed to the standard deviation of $y_{ij}$ given $b_i$, is more attractive and practically meaningful. This is because (1) the mean parameters will have conventional effect size interpretations and will be easy to explain to practitioners; (2) the variance of the scaled responses will be constant. Second, we are interested in an estimation procedure that is robust to misspecification of the within-subject correlation matrix.

First, consider the following scaled marginal model with heterogeneous exposure effects

$$\frac{\mathrm{E}(y_{ij}|x_i, w_i)}{\sigma_j} = x_i^T \beta_j + w_i \alpha_j, \tag{2.2}$$

for $j = 1, \ldots, M$, where $\mathrm{var}(y_{ij}|x_i, w_i) = \sigma_j^2$. Note that model (2.2) makes no assumption on the joint distribution of $y_{ij}$ and no assumption on the correlation among the $y_{ij}$. The estimated exposure effects $\alpha_j$ and the covariate effects $\beta_j$ should be similar to those from fitting separate regression models for individual outcomes. As described previously, a major challenge is to determine whether or not the exposure $w_i$ affects the $M$ outcomes to the same degree. We define this common effect in terms of $w_i$ having the same 'effect size' on each outcome. Therefore, we are interested in testing the hypothesis that $\mathrm{H}_0$: $\alpha_j = \alpha$ $(j = 1, \ldots, M)$. In Section 3.2 we derive the estimating-equation-based score-type test of this hypothesis.

If the null hypothesis is correct, then we can estimate a common exposure effect using the following scaled marginal model:

$$\frac{\mathrm{E}(y_{ij}|x_i, w_i)}{\sigma_j} = x_i^T \beta_j + w_i \alpha. \tag{2.3}$$

The parameters $\beta_j$ still represent the effect of the covariates on each outcome. The parameter $\alpha$ is now the common exposure effect. The primary benefit of model (2.3) is the effect of exposure can be tested using a more powerful one-degree-of-freedom test. When the common effect assumption is valid, this test will clearly be more powerful than an $M$-degree-of-freedom test, where heterogeneous effects were assumed. In addition, because we scaled by the standard deviations of $\{y_i|x_i, w_i\}$, the parameter $\alpha$ has a meaningful interpretation in that it represents common effect size. We would like to note that, as pointed out by the Associate Editor, the meanings of the scale parameters $\sigma_j$ are affected by the choices of covariates $x_i$ included in the models. In other words, the scale parameters are conditional on the selected set of the covariates $x_i$.

We allow the $y_{ij}$ to be correlated by assuming in the estimation procedure a working within-subject correlation matrix and allow the working correlation matrix to be misspecified. Therefore, this model is nearly as general as the case where models are fitted separately for each response, but accounts for correlation within subject (see Section 3.1) and allows one to easily test for a global exposure effect using a one-degree-of-freedom test. It is convenient to write model (2.3) in matrix notation. Let $\sigma^2 = (\sigma_1^2, \ldots, \sigma_M^2)$, $\Psi = \mathrm{diag}(\sigma^2)$, $y_i^* = \Psi^{-1/2} y_i$, $X_i = (x_i^T \otimes \mathrm{I}, w_i 1_M)$ and $\gamma = (\beta_1^T, \ldots, \beta_M^T, \alpha)^T$. Then, model (2.3) can be succinctly written as

$$\mathrm{E}(y_i^*|X_i) = X_i \gamma.$$

## 3. INFERENCE IN THE SCALED MARGINAL MODEL

### 3.1 *Estimation*

For estimation we focus on model (2.3), which assumes a common exposure effect. We do this because (i) estimation assuming heterogeneous effects in model (2.2) only requires minor changes to the below algorithm and (ii) the score test for common exposure (Section 3.2) only requires fitting the common exposure model. We propose estimating $\delta = (\sigma^{2^T}, \gamma^T)^T$ by solving the following set of estimating equations:

$$\sum_{i=1}^{n} X_i^T \mathrm{R}^{-1}(y_i^* - X_i\gamma) = 0, \tag{3.1}$$

$$\sum_{i=1}^{n} \left\{ \frac{y_{ij}}{\sigma_j} \left( \frac{y_{ij}}{\sigma_j} - X_i\gamma \right) - 1 \right\} = 0, \; j = 1, \ldots, M, \tag{3.2}$$

where $\mathrm{R} = \mathrm{R}(\theta)$ is a working correlation matrix and depends on a vector of parameters $\theta$, which can be estimated using the method of moments (Liang and Zeger, 1986). Possible choices of a working correlation matrix include independence, exchangeable and factor analytic. Notice that these estimating equations are unbiased even if R is misspecified. This is different from the score equations in scaled linear mixed models (Lin *et al.*, 2000), where the consistency of both the estimated regression coefficients and the estimated marginal variances requires the correlation matrix among the outcomes to be correctly specified.

The estimating equation (3.1) was constructed in such a way that (1) it is asymptotically fully efficient under independence ($\mathrm{R} = I$); (2) the estimating equations for $\gamma$ (for known $\sigma^2$) are fully efficient when R is correctly specified, even when the outcomes within each subject are correlated. However, within-subject correlation is not accounted for in the estimating equation (3.2) for $\sigma^2$. This is due to the robustness consideration. If one puts the working correlation matrix R in (3.2), the estimator of $\sigma^2$ would likely be biased if R is misspecified. The resulting estimator of $\sigma^2$ is hence robust to misspecification of the working correlation matrix, but is likely to be less efficient compared to the maximum likelihood counterpart when R is correctly specified. It should be noted that the estimating equations (3.2) are equivalent to the score equations for $\sigma^2$ assuming the responses are independent and normally distributed. We study in Section 4 the asymptotic efficiency of the estimators of $\gamma$ and $\sigma^2$.

Denote the estimating equations by $U(\delta) = \sum_{1=1}^{n} U_i(\delta) = 0$, where $U_i = (U_{1i}^T, U_{2i}^T)^T$, $U_{1i}$ is equation (3.2) and $U_{2i}$ is equation (3.1). These estimating equations can be solved using a modified Gauss–Seidel algorithm (Lange, 1999). This algorithm can be implemented by alternating between a generalized estimating equation (GEE) routine (Liang and Zeger, 1986) using existing software for $\gamma$, and a Newton–Raphson algorithm for $\sigma^2$. Specifically, we first set initial values for $\sigma^2$ using the sample variances of $y_i$. Next, estimate $\gamma$ as

$$\hat{\gamma}_{\text{new}} = \left( \sum_{i=1}^{n} X_i^T \mathrm{R}^{-1} X_i \right)^{-1} \sum_{i=1}^{n} X_i^T \mathrm{R}^{-1} y_i^*.$$

This can be done using standard software by first calculating $y_i^*$ (using the current estimate of $\sigma$) and then fitting a GEE model with identity link and GEE variance parameter fixed at one (e.g. SAS PROC GENMOD with the NOSCALE option). This GEE routine will also update the estimate of $\theta$ using a moment estimator. Given the current estimates of $\theta$ and $\gamma$, update the estimate of $\sigma^2$ using Newton–Raphson. Specifically, the updated estimate of $\sigma^2$ is found by iterating

$$\sigma^2_{\text{new}} = \sigma^2_{\text{old}} + \left[ \sum_i \{\Psi^{-1} + (1/2)\mathrm{diag}(X_i\gamma)\Psi^{-1}\mathrm{diag}(X_i\gamma)\} \right]^{-1} \sum_i \{\Psi^{-1/2}\mathrm{diag}(y_i)(y_i^* - X_i\gamma) - 1_M\}$$

until convergence, where $\Psi$ is a function of $\sigma^2_{\text{old}}$. The above steps are repeated until convergence.

Denote by $\tilde{\delta}$ the solution of $\sum U_i(\delta) = 0$. The variance of $\tilde{\delta}$ is estimated using the sandwich method as $\mathrm{var}(\tilde{\delta}) = (\tilde{\mathcal{I}})^{-1}$, where $\tilde{\mathcal{I}} = \mathrm{H}(\tilde{\delta})^T \{\sum_i U_i(\tilde{\delta}) U_i(\tilde{\delta})^T\}^{-1} \mathrm{H}(\tilde{\delta})$. Here

$$\mathrm{H}(\delta) = \mathrm{E}\left\{-\frac{\partial U(\delta)}{\partial \delta^T}\right\} = \begin{pmatrix} \mathrm{H}_{11} & \mathrm{H}_{12} \\ \mathrm{H}_{21} & \mathrm{H}_{22} \end{pmatrix},$$

where

$$\mathrm{H}_{11} = \Psi^{-1} + (1/n) \textstyle\sum_i \frac{1}{2} \mathrm{diag}(X_i \gamma) \Psi^{-1} \mathrm{diag}(X_i \gamma), \qquad \mathrm{H}_{12} = \frac{1}{n} \sum \mathrm{diag}(X_i \gamma) X_i,$$

$$\mathrm{H}_{21} = \tfrac{1}{2} \textstyle\sum_i X_i^T \mathrm{R}^{-1} \Psi^{-1} \mathrm{diag}(X_i \gamma) \qquad \mathrm{H}_{22} = \sum_i X_i^T \mathrm{R}^{-1} X_i.$$

### 3.2 *The score test for common exposure*

An assumption of model (2.3) is that the exposure effect is the same for all outcomes. To check this assumption, we are interested in testing the null hypothesis that there is a common exposure effect using the heterogeneous exposure model (2.2), i.e. $H_0$: $\alpha_j = \alpha$, $(j = 1, \dots, M)$. We propose an estimating-equation-based score-type test statistic. An advantage of this score test is that one only needs to fit the common exposure model (2.3), and does not need to fit the heterogeneous exposure effect model (2.2). We first rewrite the null hypothesis in terms of two nested models by reparametrizing the model. Specifically model (2.2) is equivalent to

$$\mathrm{E}(Y_{ij}/\sigma_j) = x_i^T \beta_j + w_i \eta_1 + w_i \mathrm{I}(j > 1) \eta_j, \tag{3.3}$$

where $\mathrm{I}(\cdot)$ is an indicator function and we arbitrarily set $\alpha_1$ as the baseline (i.e. $\eta_1 = \alpha_1$ and $\eta_j = \alpha_j - \alpha_1$ for $j > 1$). We can then reformulate the null hypothesis as $H_0$: $\eta_j = 0$, $j = 2, \dots, M$. Let $\gamma^0 = (\beta^T, \eta_1)^T$, $\delta^0 = (\sigma^2, \gamma^{0T})^T$ and $\eta = (\eta_1, \dots, \eta_M)^T$. Then, we can partition the estimating functions as $U(\delta) = (U^{1T}, U^{2T})^T$, where $U^1$ is the estimating function for $\delta^0$ and $U^2$ is the estimating function for $\eta_2, \dots, \eta_M$, i.e.

$$U^1 = \begin{pmatrix} \sum n^{-1} \{\Psi^{-1/2} \mathrm{diag}(y_i)(y_i^* - X_i \gamma^0 - w_i \Delta_1 \eta) - 1_M\} \\ \sum X_i^T \mathrm{R}^{-1} (y_i^* - X_i \gamma^0 - w_i \Delta_1 \eta) \end{pmatrix}$$
$$U^2 = \sum w_i \Lambda \mathrm{R}^{-1} (y_i^* - X_i \gamma^0 - w_i \Delta_1 \eta),$$

where $\Delta_j$ is the $M \times M$ identity matrix except the $j$th diagonal element is zero and $\Lambda$ is an $(M-1) \times M$ matrix which is the identity matrix with the first row deleted. Define

$$\mathrm{A} = \mathrm{E}\left(\frac{\partial U^2}{\partial \delta^{0T}}\right) = -\left\{\frac{1}{2} \sum w_i \Lambda \mathrm{R}^{-1} \Psi^{-1} \mathrm{diag}(X_i \gamma^0 + w_i \Delta_1 \eta),\ \sum w_i \Lambda \mathrm{R}^{-1} X_i\right\}$$

and $\mathrm{G} = \sum U_i(\delta) U_i(\delta)^T$. Also, note that $\mathrm{E}\left(-\frac{\partial U^1}{\partial \delta^{0T}}\right) = \mathrm{H}(\delta^0)$, where H was defined in Section 3.1. Then, following Breslow (1990), it can be shown that the score test of $H_0 : \eta_2 = \cdots = \eta_M = 0$ is

$$S = \{U^2(\tilde{\delta}^{0T})\}^T \Sigma^{-1}(\tilde{\delta^0}) U^2(\tilde{\delta}^{0T}),$$

where $\Sigma = \mathrm{G}_{22} - \mathrm{A}\mathrm{H}^{-1}\mathrm{G}_{12} - \mathrm{G}_{21}(\mathrm{H}^{-1})^T \mathrm{A}^T + \mathrm{A}\mathrm{H}^{-1}\mathrm{G}_{11}(\mathrm{H}^{-1})^T \mathrm{A}^T$, $\mathrm{G}_{11}, \mathrm{G}_{12}, \mathrm{G}_{21}$ and $\mathrm{G}_{22}$ are the corresponding submatrices of G and all of the matrices are evaluated at $\tilde{\delta}^0$. The score statistic $S$ asymptotically follows a $\chi^2$ distribution with $M - 1$ degrees of freedom under $H_0$.

### 3.3 *The Wald test for common exposure*

Alternatively, a Wald test for common exposure can be constructed by fitting the heterogeneous exposure model (2.2) and estimating $\alpha$, or equivalently by fitting the heterogeneous exposure model (3.3) and estimating $\eta$. Estimates of $\eta$ can easily be found using the algorithm described in Section 3.1 by changing the definition of the design matrix. Alternatively, estimates of $\eta$ can be obtained by estimating the vector $\alpha$ from model (2.2) and transforming these estimates to $\tilde{\eta}$. The true covariance matrix of $\tilde{\eta}$ has a sandwich form similar to that given at the end of Section 3.1. Calculations of the Wald statistic hence involve the same amount of additional programming as is required to calculate the score statistic, and it further requires fitting the heterogeneous exposure model (3.3). Therefore, the score test for the common exposure effect is more convenient to do than the Wald test.

However, we can calculate a naive Wald test for the common exposure effect by fitting the heterogeneous exposure model (3.3) and obtaining naive estimates of the covariance matrix of $\tilde{\eta}$ directly from standard GEE software at convergence. We call this estimated covariance matrix 'naive' because it assumes $\sigma^2$ is known and ignores the variability associated with estimation of $\sigma^2$. The Wald test based on this naive variance is $\tilde{\eta}^{*T}\{\widetilde{\mathrm{var}}(\tilde{\eta}^*)\}^{-1}\tilde{\eta}^*$, where $\eta^* = (\eta_2, \dots, \eta_M)^T$ and $\widetilde{\mathrm{var}}(\tilde{\eta}^*)$ denotes the 'naive' covariance of $\eta$. This naive Wald test is easier to compute than the score test, because it can be output directly from standard software packages, whereas the score test requires additional code for finding $\Sigma$ (see Section 3.2). Unfortunately, because this test is using the naive standard errors by ignoring the variability in estimating $\sigma^2$, it is likely to be asymptotically biased. However, this bias might be small in practice in some settings. We will compare in Section 5 the score test and the naive Wald test in the analysis of the HIV data.

## 4. Asymptotic efficiency

Our parameter estimator is estimating-equation-based. While it is more robust than the maximum likelihood estimator (MLE), it might be less efficient than the MLE when the likelihood function and the correlation matrix are correctly specified. We consider in this section the asymptotic efficiency of our estimator of $\gamma$ and $\sigma^2$ relative to the MLE when the likelihood function and the correlation matrix are correctly specified. Specifically, in our asymptotic efficiency study, we further assume the outcomes follow model (2.3) and are normally distributed with the correlation matrix correctly specified. It follows that

$$Y_i \sim N(\Psi^{1/2} X_i \gamma,\ \Psi^{1/2} R_{\mathrm{T}}(\theta) \Psi^{1/2}), \tag{4.1}$$

where $R_{\mathrm{T}}$ is the true correlation matrix. Assuming $\theta$ is known, some calculations show that the information matrix of the MLE of $(\hat{\gamma}, \widehat{\sigma^2})$ is

$$\hat{\mathcal{I}} = \begin{pmatrix} \hat{\mathcal{I}}_{11} & \hat{\mathcal{I}}_{12} \\ \hat{\mathcal{I}}_{12}^T & \hat{\mathcal{I}}_{22} \end{pmatrix},$$

where

$$\hat{\mathcal{I}}_{11} = \sum_{i=1}^{n} X_i^T R_{\mathrm{T}}^{-1} X_i$$

$$\hat{\mathcal{I}}_{12}[., j] = \frac{1}{2\hat{\sigma}_j^2} \sum_{i=1}^{n} X_i^T R_{\mathrm{T}}^{-1} (\mathrm{I} - \Delta_j) X_i \hat{\gamma}$$

$$\hat{\mathcal{I}}_{22}[j, k] = \frac{n}{4\hat{\sigma}_j^4} \mathrm{I}(j = k) + \sum_{i=1}^{n} \frac{1}{4\hat{\sigma}_j^2 \hat{\sigma}_k^2} [\mathrm{tr}\{(\mathrm{I} - \Delta_j) R_{\mathrm{T}}^{-1} (\mathrm{I} - \Delta_k) R_{\mathrm{T}}\} + \hat{\gamma}^T X_i^T (\mathrm{I} - \Delta_j) R_{\mathrm{T}}^{-1} (\mathrm{I} - \Delta_k) X_i \hat{\gamma}]$$

for $j, k = 1, \dots, M$. The notation $\hat{\mathcal{I}}_{12}[., j]$ refers to the $j$th column of $\hat{\mathcal{I}}_{12}$. The asymptotic relative efficiency (ARE) for a particular vector of parameters $\phi$ is

$$\mathrm{ARE}(\tilde{\phi}, \hat{\phi}) = \left( \lim_{n \to \infty} \frac{|\mathrm{cov}(\hat{\phi})|}{|\mathrm{cov}(\tilde{\phi})|} \right)^{1/q}$$

where $\phi$ is $\beta$, $\alpha$ or $\sigma^2$ and $q$ is the dimension of $\phi$. The term $\mathrm{cov}(\hat{\phi})$ is just the appropriate submatrix of $\hat{\mathcal{I}}^{-1}$, whereas $\mathrm{cov}(\tilde{\phi})$ is the submatrix of the sandwich estimator $\tilde{\mathcal{I}}^{-1}$.

The AREs of $\tilde{\gamma}$ and $\tilde{\sigma}^2$ compared to $\hat{\gamma}$ and $\hat{\sigma}^2$ are of interest. First note that when there is no correlation between the outcomes (i.e. $R_{\mathrm{T}} = \mathrm{I}$) and working independence is assumed, the estimator $\tilde{\delta}$ is fully efficient. This can be seen by noticing that the estimating equations $U(\delta)$ are equivalent to the score equations from the true model (4.1), when $R_{\mathrm{T}} = R = \mathrm{I}$. However, when there is correlation between the outcomes, $\tilde{\delta}$ will tend to be less efficient.

We study the AREs numerically for the following simple special case:

$$\mathrm{E}(y_{ij}/\sigma_j) = \beta_{0j} + x_i \beta_{1j} + w_i \alpha$$

where $w_i$ is a binary exposure variable with half of the subjects being exposed and half not, $x_i$ is a confounder and $\mathrm{corr}(y_i) = R_{\mathrm{T}}$. The covariate $x_i$ was generated from $N(3, 1)$ if $w_i = 0$ and from $N(1, 1)$ if $w_i = 1$. This was done to create imbalance in the data—mimicking real-world observational studies. For simplicity we assumed $\beta_{0j} = \beta_0$ and $\beta_{1j} = \beta_1$ for all $j$. We calculated the AREs for a variety of values of $\gamma$ and $\theta$. It can be easily shown that the ARE of $(\gamma, \sigma^2)$ does not depend on $\sigma^2$. We assumed the number of outcomes was $M = 3$, and the true correlation matrix was either exchangeable or factor analytic. For simplicity, we used the following factor analytic structure:

$$R_{\mathrm{T}} = \begin{pmatrix} 1 & \theta & \theta \\ \theta & 1 & 1-\theta \\ \theta & 1-\theta & 1 \end{pmatrix},$$

so that $R_{\mathrm{T}}$ only depends on a single parameter. We calculate the asymptotic relative efficiency of our estimator when the correlation matrix is misspecified and is correctly specified.

The results are given in Table 1. The efficiency of our estimating-equation-based estimators tend to be quite high compared to the MLEs when the correlation is small ($\theta = 0.3$), especially for the estimator of the exposure effect $\alpha$. The efficiency loss is larger when the correlation increases ($\theta = 0.7$), with approximately 20% efficiency loss in some cases. When the true correlation is exchangeable, misspecifying the correlation matrix has little effect on efficiency. However, when the true correlation is factor analytic, correctly specifying R results in efficiency gains, particularly for $\alpha$. The efficiency of the estimator of $\sigma^2$ is also quite high compared to its maximum likelihood counterpart, although its loss of efficiency is a little higher than that of the estimator of $\alpha$. This is because the estimating equations of $\sigma^2$ ignore correlation between outcomes.

## 5. APPLICATION TO THE HIV DATA

We illustrate the methods with a study of neurocognitive performance in HIV-infected women. The study involved 125 HIV-seropositive women from the HER study (Smith *et al.*, 1997)—an epidemiologic study of HIV-infected women. Women were given a neurocognitive exam every six months after their CD4 count fell below 100 ($10^6$ cells/l). Neurocognitive data at baseline and at their most recent visit (about 3 years since the baseline) were used in the analysis. The goal was to determine whether HAART, which has been shown to reduce viral load and improve immune functioning, enhances neurocognitive performance (Cohen *et al.*, 2001).

Table 1. *Asymptotic relative efficiency of the scaled marginal model parameter estimates with respect to the MLEs. Upper entry is for exchangeable $R_{\mathrm{T}}$; lower entry is factor analytic $R_{\mathrm{T}}$*

| | | | Working $R$ | | | | | |
|---|---|---|---|---|---|---|---|---|
| True parameter values | | | Independence | | Exchangeable | | Factor analytic | |
| $\theta$ | $\beta$ | $\alpha$ | $\alpha$ | $\sigma^2$ | $\alpha$ | $\sigma^2$ | $\alpha$ | $\sigma^2$ |
| 0.3 | (1,1) | 0 | 1 | 0.94 | 1 | 0.94 | 1 | 0.94 |
| 0.3 | (1,1) | 0 | 0.97 | 0.87 | 0.98 | 0.87 | 1 | 0.88 |
| 0.3 | (1,1) | 3 | 0.95 | 0.94 | 0.96 | 0.94 | 0.95 | 0.94 |
| 0.3 | (1,1) | 3 | 0.89 | 0.87 | 0.89 | 0.88 | 0.92 | 0.88 |
| 0.3 | (3,3) | 0 | 1 | 0.95 | 1 | 0.95 | 1 | 0.95 |
| 0.3 | (3,3) | 0 | 0.97 | 0.88 | 0.97 | 0.88 | 1 | 0.89 |
| 0.3 | (3,3) | 3 | 0.96 | 0.95 | 0.96 | 0.95 | 0.96 | 0.95 |
| 0.3 | (3,3) | 3 | 0.89 | 0.88 | 0.89 | 0.88 | 0.92 | 0.89 |
| 0.7 | (1,1) | 0 | 1 | 0.78 | 1 | 0.78 | 1 | 0.78 |
| 0.7 | (1,1) | 0 | 0.90 | 0.80 | 0.90 | 0.80 | 1 | 0.81 |
| 0.7 | (1,1) | 3 | 0.84 | 0.78 | 0.84 | 0.78 | 0.84 | 0.78 |
| 0.7 | (1,1) | 3 | 0.81 | 0.81 | 0.81 | 0.81 | 0.86 | 0.80 |
| 0.7 | (3,3) | 0 | 1 | 0.79 | 1 | 0.79 | 1 | 0.79 |
| 0.7 | (3,3) | 0 | 0.90 | 0.83 | 0.90 | 0.83 | 1 | 0.82 |
| 0.7 | (3,3) | 3 | 0.84 | 0.79 | 0.84 | 0.79 | 0.84 | 0.79 |
| 0.7 | (3,3) | 3 | 0.81 | 0.83 | 0.80 | 0.83 | 0.86 | 0.83 |

Three tasks were used to measure neurocognitive performance: Color Trail Making 1 total time (CTM), Controlled Word Association Test (COWAT) and Grooved Pegboard total time (GPB). CTM and GPB are the time it takes to complete the corresponding tasks (shorter time indicates better performance). COWAT is a score that measures verbal fluency, and lower values indicate better performance. These tasks measure neurocognitive functioning from different perspectives. For example, COWAT is a measure of verbal fluency; CTM and GPB measure psychomotor speed and information processing. The outcome variables of interest are changes in these three measures from baseline. Specifically, we define the three outcome variables as $y_1$ = CTM (evaluation 2) − CTM (evaluation 1), $y_2$ = COWAT (evaluation 2) − COWAT (evaluation 1) and $y_3$ = GPB (evaluation 2) − GPB (evaluation 1). A negative value of each of the three outcomes indicates an improvement in neurocognitive performance for that measure. One subject that had an extreme outlier for $y_1$ was removed from the analyses.

Table 2 gives the means and standard deviations of the three outcomes for each treatment group (HAART and non-HAART). For all three measures, neurocognitive performance has declined for the non-HAART group and improved for the HAART group. Because the three outcomes measure different aspects of neurocognitive functioning, one important question is whether HAART has a common effect on the three measures. This is difficult to tell from the raw data, because the outcomes are measured on different scales. The (pooled) estimated standard deviations are 21.6, 5.6 and 35.9 for the three outcomes, respectively. Once we scale by these sample standard deviations, we see that a common effect size of HAART may be quite plausible (mean differences of $-0.44$, $-0.47$ and $-0.44$, respectively). Figure 1 presents box plots of the standardized outcomes, stratified by treatment group.

These preliminary analyses ignore potential confounders, as well as the correlation between the

Table 2. *Summary statistics for outcomes by treatment status. $y_1$ is color trail (eval. 2 − eval. 1); $y_2$ is COWAT (eval. 2 − eval. 1); $y_3$ is grooved peg board (eval. 2 − eval. 1)*

| | Non-HAART ($n$ = 69) | | HAART ($n$ = 55) | | |
|---|---|---|---|---|---|
| Outcome | Mean | SD | Mean | SD | Mean diff. |
| $y_1$ | 3.9 | 22.6 | −5.6 | 19.1 | −9.5 |
| $y_2$ | 1.5 | 4.9 | −1.1 | 6.1 | −2.6 |
| $y_3$ | 6.6 | 38.8 | −9.2 | 30.0 | −15.8 |
| $y_1/\text{sd}(y_1)$ | 0.18 | 1.0 | −0.26 | 0.89 | −0.44 |
| $y_2/\text{sd}(y_2)$ | 0.27 | 0.88 | −0.20 | 1.1 | −0.47 |
| $y_3/\text{sd}(y_3)$ | 0.18 | 1.1 | −0.26 | 0.84 | −0.44 |

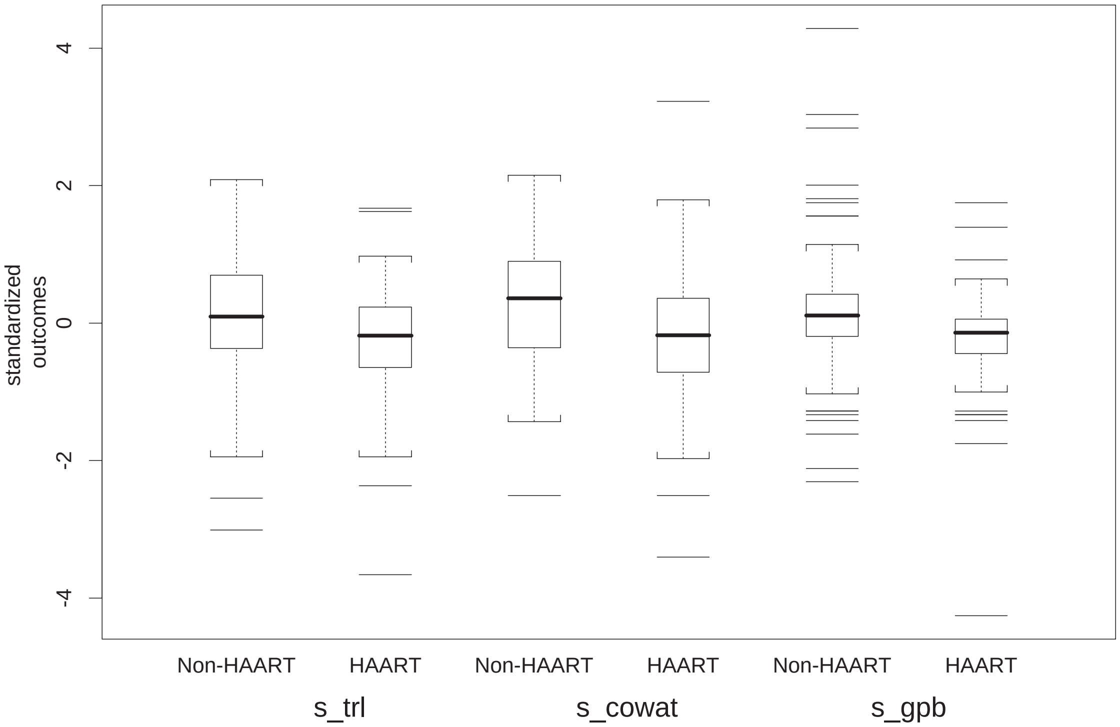

Fig. 1. Box plots of the standardized outcome variables stratified by treatment status. The sample standard deviations were used to standardize each outcome. The standardized outcomes are color trail (s_trl), COWAT (s_cowat) and grooved pegboard (s_gpb).

outcomes. We therefore apply our marginal scaled model to these data. To determine whether a common treatment effect was reasonable, we first fitted model (2.2) with $w$ = HAART (1 = HAART, 0 = non-HAART) and the following covariates: age at baseline (years); depression severity as measured by the Center for Epidemiology Scale (CESD); and CD4 count at baseline. We also considered other covariates, such as change in CD4 count, illicit drug use and plasma viral load, but these had little impact in the model. We therefore did not include these variables in the model. The individual treatment effects $\alpha_j$ were estimated as −0.47 (SE = 0.18), −0.52 (SE = 0.22) and −0.47 (SE = 0.17). The score test statistic for common exposure effect was 0.06 (d.f. = 2, $p$-value = 0.97). Therefore, there is no evidence that a

Table 3. *Parameter estimates and estimated standard errors from application to the HIV data*

| | | Estimate | SE | Naive SE |
|---|---|---|---|---|
| Intercept | $\beta_{01}$ | −0.57 | 0.77 | 0.78 |
| | $\beta_{02}$ | 1.52 | 0.67 | 0.67 |
| | $\beta_{03}$ | −0.25 | 0.75 | 0.73 |
| Age | $\beta_{11}$ | 0.012 | 0.014 | 0.014 |
| | $\beta_{12}$ | −0.031 | 0.014 | 0.014 |
| | $\beta_{13}$ | 0.008 | 0.013 | 0.013 |
| CESD | $\beta_{31}$ | 0.014 | 0.009 | 0.009 |
| | $\beta_{32}$ | 0.002 | 0.008 | 0.008 |
| | $\beta_{33}$ | 0.008 | 0.010 | 0.009 |
| CD4 (40,100) | $\beta_{41}$ | −0.48 | 0.27 | 0.27 |
| | $\beta_{42}$ | −0.10 | 0.23 | 0.23 |
| | $\beta_{43}$ | −0.12 | 0.18 | 0.18 |
| CD4 (<40) | $\beta_{51}$ | −0.18 | 0.23 | 0.23 |
| | $\beta_{52}$ | −0.17 | 0.21 | 0.21 |
| | $\beta_{53}$ | −0.21 | 0.17 | 0.17 |
| HAART | $\alpha$ | −0.49 | 0.13 | 0.14 |
| | $\sigma_1^2$ | 411 | 71.6 | |
| | $\sigma_2^2$ | 28.6 | 4.6 | |
| | $\sigma_3^2$ | 1206 | 293 | |

common treatment effect is not a reasonable assumption. In addition, we calculated the naive Wald test using the naive standard errors to test for the common exposure effect. The value of this test statistic was 0.06 ($p$-value = 0.97), which was consistent with the score test.

We therefore fitted the model assuming a common effect of HAART on the scaled outcomes. Because there are only three outcomes, we left the working correlation matrix unspecified. The results are given in Table 3. The estimate of the common effect of HAART was −0.49 (SE = 0.13). This indicates that HAART does enhance neurocognitive performance. Specifically, the change in neurocognitive performance for HAART-treated women was about a half a standard deviation better on all three outcomes than for women who did not take HAART. It may be more meaningful to translate the common effect onto the original scales by multiplying the estimate of $\alpha$ by the corresponding standard deviations. The estimated effect of HAART on the change in CTM, COWAT and GPB are −9.9, −2.6 and −17.0, respectively. Older ages at baseline were associated with better verbal fluency (COWAT), but there was no age effect on the other outcomes. Larger values of CESD (more depressed) were associated with poorer neurocognitive performance, although the estimates were not significant. The estimated effect of CD4 count at baseline was negative, indicating sicker patients showed more benefits from HAART. Again, however, these tended to not be significant. The estimated working correlation matrix was

$$\mathrm{R} = \begin{pmatrix} 1 & 0.12 & 0.29 \\ 0.12 & 1 & 0.29 \\ 0.29 & 0.29 & 1 \end{pmatrix}.$$

Table 3 also presents the naive SEs for $\gamma$ which were obtained directly from SAS PROC GENMOD. These SEs do not take into the account the fact that $\sigma^2$ was estimated. These naive SEs perform quite well in this example.

As a sensitivity analysis, we carried out an alternative analysis where instead of defining the response variable as a difference between each measure at the two time points, we defined it as the response at the most recent visit. Similar to ANCOVA models, we included the response at baseline as a covariate in $x$ and assumed working independence for the covariance matrix. The score test for a common treatment effect could not be rejected, and the estimated common effect was significant (though slightly smaller in magnitude). Because the results were similar, we do not present the findings here.

## 6. DISCUSSION

In this paper we proposed a scaled marginal model for test and estimation of global exposure effects. An attractive feature of this model is the common exposure effect has an appealing interpretation in terms of the common effect size. In addition, few restrictions are made on the coefficients of the other covariates in the model for the scaled mean. It should be noted that the scale parameters are conditional on the set of covariates included in the model, i.e. different choices of covariates affect the meanings of the scales. Hence our method requires correctly specifying the scaled mean models by including appropriate covariates. This is in the same spirit as GEEs. However, we do not make any assumption on the joint distribution of the outcomes and no assumption on the correlation matrix. As a result, our method is more robust than maximum likelihood estimation.

Our asymptotic efficiency analysis shows that in most of the situations that we considered there was not much loss of efficiency compared to the MLE; in the worst cases there was about 20% efficiency loss. Estimation of the model parameters is estimating equation based and can be easily obtained by iterating between GEE estimation for $\gamma$ using standard software (e.g. SAS PROC GENMOD) and a Newton–Raphson algorithm for estimating $\sigma^2$. The SAS macro that was used for estimation can be downloaded from `http://stat.brown.edu/AMPD`.

There are several simpler, *ad hoc*, methods that could be used to fit scaled models. One possibility is to just scale the outcomes by their sample standard deviations and fit a GEE model with a common exposure effect. This approach is problematic because the sample standard deviations estimates do not properly account for heterogeneity of the population. In addition, it would ignore the fact that the standard deviations are estimated. A slightly more complicated alternative would be to fit the model in stages. First fit a GEE model to the outcomes scaled by the sample standard deviations. Then use the residuals from this model to come up with new estimates of the standard deviations and repeat the process. Again, this approach does not account for the fact that the scale parameters are estimated. Also, the properties of the corresponding estimates are not known.

For simplicity, we assume in this paper that the scale parameters are constant across exposure groups and a common set of covariates $x_i$ is included in each outcome model. Extensions of the proposed model to relax these assumption by allowing different scale parameters for different exposure groups and different sets of covariates $x_i$ for different outcome models are straightforward with some changes of notation. We can also easily allow the exposure variable to be continuous in our model. The same estimation procedure applies. An area that warrants further research is the theoretical properties of the naive SEs and the naive Wald test. The advantage of using the naive estimates is they are easier to compute. For the HIV data, the naive SEs and the naive Wald test using the naive SEs performed well. It is worth studying whether that is true in general and under what conditions they will perform poorly.

## ACKNOWLEDGEMENTS

The authors are grateful to Dr Ron Cohen of Brown University School of Medicine for supplying the data, and the Associate Editor and two referees for helpful comments. This research was supported in part by grants from the US National Institutes of Health.

## REFERENCES

BRESLOW, N. (1990). Tests of hypotheses in overdispersed Poisson regression and other quasi-likelihood models. *Journal of the American Statistical Association* **85**, 565–571.

COHEN, R. A., BOLAND, R., PAUL, R., TAHIMA, K. T., SCHOENBAUM, E. E., CELENTANO, D. D., SCHUMAN, P., SMITH, D. K. AND CARPENTER, C. (2001). Neurocognitive performance enhanced by highly active antiretroviral therapy in HIV-infected women. *AIDS* **15**, 341–345.

LANGE, K. (1999). *Numerical Analysis for Statisticians*. New York: Springer.

LIANG, K.-Y. AND ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

LIN, X., RYAN, L., SAMMEL, M., ZHANG, D., PADUNGTOD, C. AND XU, X. (2000). A scaled linear mixed model for multiple outcomes. *Biometrics* **56**, 593–601.

O'BRIEN, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics* **40**, 1079–1087.

POCOCK, S. J., GELLER, N. L. AND TSIATIS, A. A. (1987). The analysis of multiple endpoints in clinical trials. *Biometrics* **43**, 487–498.

SAMMEL, M. D. AND RYAN, L. M. (1996). Latent variable models with fixed effects. *Biometrics* **52**, 650–663.

SAMMEL, M. D. AND RYAN, L. M. (2002). Effects of covariance misspecification in a latent variable model for multiple outcomes. *Statistica Sinica* **12**, 1207–1222.

SAMMEL, M. D., LIN, X. AND RYAN, L. (1999). Multivariate linear mixed models for multiple outcomes. *Statistics in Medicine* **18**, 2479–2492.

SMITH, D. K., WARREN, D. L., VLAHOV, D., SCHUMAN, P., STEIN, M. D., GREENBERG, B. L. AND HOLMBERG, S. D. (1997). Design and baseline participant characteristics of the Human Immunodeficiency Virus Epidemiology Research (HER) Study: a prospective cohort study of human immunodeficiency virus infection in US women. *American Journal of Epidemiology* **146**, 459–469.

WICHERN, D. W. AND JOHNSON, R. A. (2002). *Applied Multivariate Statistical Analysis*. Englewood Cliffs, NJ: Prentice Hall.

[*Received February 18, 2002; first revision August 8, 2002; second revision November 11, 2002; accepted for publication November 12, 2002*]