Statistica Sinica **12**(2002), 1207-1222

EFFECTS OF COVARIANCE MISSPECIFICATION IN A LATENT VARIABLE MODEL FOR MULTIPLE OUTCOMES

Mary Dupuis Sammel and Louise M. Ryan

University of Pennsylvania School of Medicine
and Harvard School of Public Health

Abstract: Sammel and Ryan (1996) developed a latent variable model that allows for covariate effects on multiple continuous outcomes. While the approach provides an effective tool for data reduction and global test for covariate effects, it makes strong assumptions about the covariance among the outcomes. In addition, some parameters are common to both the mean and variance suggesting that robustness could be a problem. This manuscript evaluates model misspecification on tests of exposure effects derived from the latent variable model. We develop a robust score test which is valid under misspecified variance assumptions and compare it to one based on Generalized Estimating Equations (GEE) (Liang and Zeger (1986)), under varying assumptions on the true model. Both models have similar loss in power under variance misspecification while the estimated global effect of the covariate is more biased towards the null for the GEE model than the LV model. As the variance/scale of the outcomes increases, the performance of the LV model improves. As for asymptotic comparisons, test performance depends upon the amount of variability and correlation among the outcomes. The LV model test is superior when the data are highly correlated, $\rho > 0.3$, and with large variance. When uncorrelated outcomes are incorporated, the GEE model is superior, except when only the correlated outcomes are impacted by the exposure.

Key words and phrases: Factor analysis, generalized estimating equations, global tests.

1. Introduction

In many applied settings, it is of interest to assess the effect of covariates on multiple outcomes. In the study of birth defects for instance, appropriately combining multiple outcomes may provide more power to test an effect than focusing on a single endpoint (Holmes et al. (1987)). Sammel and Ryan (1996) developed a model for multiple continuous outcomes that formalizes the idea of performing a factor analysis and then modeling the estimated factors as a function of the exposure of interest. While we proposed simultaneous estimation of model parameters, methods which estimate the two sets of parameters separately (Two Stage Factor Analysis-TSFA) are commonly used for validation and hypothesis testing of measurement scales and indices (Streiner and Norman

(1995, Chap.10)). Some recent examples include an evaluation of risk factors influencing menopausal symptoms (Freeman, et al. (2000)), and tests of neurobehavior (Heyer (1996)). The model also allows for adjustment with respect to other covariates and can be thought of as an extension of the random effects (RE) model of Laird and Ware (1982) (see also Harville (1977)). Tests for covariate effects on the latent variable provide a global test for multiple outcomes. While the approach is appealing, it involves strong assumptions about the covariance of the observed data.

Other approaches have been suggested for testing covariate effects with respect to multiple outcomes, including the generalized least squares methods of O'Brien (1984) and Laska, Tang, and Meisner (1992). Pocock, Geller, and Tsiatis (1987) extended these methods to combinations of test statistics based on arbitrary types of data. Normal models evaluating mean effects in the presence of correlated outcomes have been proposed for this framework (Random Effects–RE models; Laird and Ware (1982); Harville (1977)). Extensions of these methods based on Generalized Estimating Equations (GEE) (Liang and Zeger (1986); Zeger and Liang (1986); Lefkopoulou and Ryan (1993); Legler, Lefkopoulou, and Ryan (1995); Bull (1998)) focus on testing the mean structure of the data, while the covariance is treated as a nuisance. These GEE-based tests are made robust to covariance misspecification by the use of an empirical adjustment to the estimated variances of the parameters.

In this paper we assess the impact of misspecification of the latent structure on a test for covariate effects based on the latent variable model. We first derive a global score test for the effect of a covariate on multiple outcomes using the latent variable model, as well as a robust version, which we compare to a robust test based upon a GEE model. We also look at two other approaches, namely the ad hoc two-stage factor analysis described above and a random effects model. Comparisons between the various tests are based on analytic considerations as well as simulations. We illustrate the robust score tests for subsets of outcomes comparing healthy control infants to those exposed in utero to anticonvulsant medications (Holmes et al. (1994)). We conclude with an example and some practical guidelines for the use of latent variable models.

1.1. A test based on a latent variable model

Sammel and Ryan (1996) propose a two-stage model to incorporate covariate effects on multiple outcomes. At the first stage, a set of M continuous outcomes is modeled as a function of unobservable latent variables, as well as other covariates. At the second stage, the latent variables are modeled as a function of exposure or other covariates of interest. More precisely, suppose the observed data for individual i are $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iM})^T$, and let $\boldsymbol{x}_i$ represent a $P \times 1$ dimensional vector of covariates. Also, $\boldsymbol{b}_i$ represents a $Q \times 1$ vector of unobserved latent variables for individual i. Then the conditional model for $\boldsymbol{y}_i$, given the latent variables $\boldsymbol{b}_i$, can be written as

$$\boldsymbol{y}_i = \left(\boldsymbol{x}_i^T \otimes \boldsymbol{I}_M\right)\boldsymbol{\alpha} + \left(\boldsymbol{b}_i^T \otimes \boldsymbol{I}_Q\right)\boldsymbol{\lambda} = \boldsymbol{X}_i\boldsymbol{\alpha} + \boldsymbol{B}_i\boldsymbol{\lambda} + \boldsymbol{e}_i, \tag{1}$$

where $\otimes$ represents the Kronecker product (Rogers (1980), p.12).

The matrix $\boldsymbol{\lambda}$ contains the factor loadings which associate the Q-dimensional latent vector $\boldsymbol{b}_i$ with the observed data $\boldsymbol{y}_i$. Assume $\boldsymbol{e}_i \sim N(0, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi} = \text{diag}\left\{\sigma_m^2\right\}_{m=1,\ldots,M}$. That is, conditional on the latent data, $\boldsymbol{b}_i$, the outcomes are independent. The second stage of the model can then be written as

$$\boldsymbol{b}_i = \boldsymbol{Z}_i\boldsymbol{\theta} + \boldsymbol{\delta}_i, \tag{2}$$

where $\boldsymbol{\delta}_i \sim \text{N}(\boldsymbol{0}, \boldsymbol{I}_Q)$, and $\boldsymbol{Z}_i$ reflects exposures and other covariates of interest. These covariates differ from the covariates $\boldsymbol{x}_i$, as they are the subset to be tested. For simplification we consider only a single latent factor, $Q = 1$, and scalar covariate z, therefore models (1) and (2) imply the following marginal model for $\boldsymbol{y}_i$:

$$f(\boldsymbol{y}_i|\boldsymbol{\alpha}, \boldsymbol{\lambda}, \boldsymbol{\Psi}, \theta) \sim \text{N}\left(\boldsymbol{X}_i\boldsymbol{\alpha} + \boldsymbol{\lambda} z_i\theta, \boldsymbol{\lambda}\boldsymbol{\lambda}^T + \boldsymbol{\Psi}\right). \tag{3}$$

The question to be addressed in this paper is how tests based on (3) perform when the model has been misspecified, in particular, when the assumed marginal covariance of $\boldsymbol{y}_i$ is wrong. The test of primary interest is for the null hypothesis of no exposure effect: $H_o\colon\ \theta = 0$. Under the marginal log-likelihood for model (3), the efficient score is

$$S_{lv}(\theta) = \frac{\partial}{\partial\theta} l(\theta, \boldsymbol{\alpha}, \boldsymbol{\lambda}, \boldsymbol{\Psi}) = \sum_{i=1}^{n} z_i\boldsymbol{\lambda}^T\left(\boldsymbol{\lambda}\boldsymbol{\lambda}^T + \boldsymbol{\Psi}\right)^{-1}\left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\alpha} - \boldsymbol{\lambda} z_i\theta\right).$$

A score test for the hypothesis is thus

$$T_{lv_m} = S_{lv}(0)^T V_{11} S_{lv}(0) \mid_{\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}, \boldsymbol{\Psi}=\hat{\boldsymbol{\Psi}}, \boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}}, \tag{4}$$

where V_{11} is the $1,1$ element of the inverse of the Fisher information matrix, $I(\boldsymbol{\zeta}^*)^{-1}$, see (11) in Appendix A, evaluated under the null hypothesis. In addition, a "robust" version of this test can be computed as

$$T_{lv_r} = S_{lv}(0)^T \boldsymbol{\Sigma}_l^{-1} S_{lv}(0) \mid_{\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}, \boldsymbol{\Psi}=\hat{\boldsymbol{\Psi}}, \boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}}, \tag{5}$$

$$\boldsymbol{\Sigma}_l = V_{11}^{-1}\sum_{i=1}^{n}\left\{z_i\boldsymbol{\lambda}^T\left(\boldsymbol{\lambda}\boldsymbol{\lambda}^T + \boldsymbol{\Psi}\right)^{-1}\text{Var}\,(\boldsymbol{y})\left(\boldsymbol{\lambda}\boldsymbol{\lambda}^T + \boldsymbol{\Psi}\right)^{-1}\boldsymbol{\lambda} z_i\right\} V_{11}^{-1}.$$

Here Var $(\boldsymbol{y})$ is replaced by the moment estimator $\sum\left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\alpha}\right)\left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\alpha}\right)^T$ to yield a consistent estimator of Σ_l. In addition, we consider an adjustment to the model-based test, T_{lv_m} where this test statistic is divided by a constant, for a better approximation to a χ^2 distribution. The constant reflects the deviation of the assumed model variance from the true variance structure (Rotnitzky and Jewell (1990)).

1.2. A test based on GEEs

Lefkopoulou and Ryan (1993) derive a test for multiple outcomes based on the moments

$$\mathrm{E}\left(\boldsymbol{y}_i\right) = \boldsymbol{X}_i\boldsymbol{\alpha} + \mathbf{1}z_i\theta \text{ and } \mathrm{Var}\ \left(\boldsymbol{y}_i\right) = \boldsymbol{A} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{R}\boldsymbol{\Sigma}^{1/2}, \tag{6}$$

where $\mathbf{1}$ is an $M\times 1$ vector of 1's, $\boldsymbol{R}=(1-\rho)\boldsymbol{I}+\rho\mathbf{1}\mathbf{1}^T$ is an exchangeable correlation matrix and $\boldsymbol{\Sigma}=\mathrm{diag}\left(\epsilon_m^2\right)$; $m=1,\ldots,M$, is the diagonal matrix of the elements of variance of $\boldsymbol{y}_i$. Typically, a common ϵ^2 is assumed for all outcomes in the GEE model. However, this may not be appropriate unless the outcomes are repeated measures. The corresponding generalized estimating equation for θ is

$$\boldsymbol{S}_g\left(\theta\right) = \sum_{i=1}^{n} z_i^T\mathbf{1}^T\boldsymbol{\Sigma}^{-1/2}\boldsymbol{R}^{-1}\boldsymbol{\Sigma}^{-1/2}\left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\alpha} - \mathbf{1}z_i\theta\right) = 0.$$

In practice, ρ and $\boldsymbol{\Sigma}$ are estimated using the method of moments. To construct a test of $H_o : \theta = 0$, an empirical estimator of the variance is used which is robust to covariance misspecification (Liang and Zeger (1986)). The resulting "robust" test is then

$$T_{gee} = \boldsymbol{S}_g^T\left(0\right)\boldsymbol{\Sigma}_g^{-1}\boldsymbol{S}_g^T\left(0\right)|_{\boldsymbol{\Sigma}=\hat{\boldsymbol{\Sigma}},\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}}, \tag{7}$$

where $\boldsymbol{\Sigma}_g = h_1^{-1}g\left(\rho,\boldsymbol{\Sigma}\right)h_1^{-1}$ for h_1 the 1,1 element of the inverse information matrix, assuming the model is correctly specified and

$$g\left(\rho,\boldsymbol{\Sigma}\right) = \sum_{i=1}^{n} z_i^T\mathbf{1}^T\boldsymbol{\Sigma}^{-1/2}\boldsymbol{R}^{-1}\boldsymbol{\Sigma}^{-1/2}\mathrm{Var}\ \left(\boldsymbol{y}_i\right)\boldsymbol{\Sigma}^{-1/2}\boldsymbol{R}^{-1}\boldsymbol{\Sigma}^{-1/2}\mathbf{1}z_i.$$

This variance of the test reduces when the distribution of the data is correctly specified.

2. Comparison of Test Statistics

2.1. Data generating models

In this section we describe several data generating models which will be the basis for evaluating and comparing the different tests. The first true data generating model (DGM 1) assumes the data originate from a single factor latent variable model as described in (3). Data generating model 2 assumes a common exposure effect and a constant correlation among the outcomes, as under (6).

A more general model is considered as model 3, which assumes an arbitrary covariance structure and a common exposure effect on all the outcomes:

$$\boldsymbol{y}_i \sim \mathrm{N}\left(\boldsymbol{X}_i\boldsymbol{\alpha} + \mathbf{1}z_i\theta, \begin{bmatrix} \boldsymbol{A} & 0 \\ 0 & \boldsymbol{I} \end{bmatrix}\right). \tag{8}$$

We consider several correlation structures which imply a subset of independent outcomes with covariance $\boldsymbol{I}$, while the correlated subset, $\boldsymbol{A}$, has an exchangeable form as at (6). Data for model 4 will have a similar structure to the variance described above in (8), but will have a subset of the outcomes whose means are not impacted by the exposure. For example, assume that the distribution of the observed outcomes is

$$\boldsymbol{y}_i \sim \mathrm{N}\left(\begin{bmatrix} \boldsymbol{X}_i\boldsymbol{\alpha} + \mathbf{1}z_i\theta \\ \boldsymbol{X}_i\boldsymbol{\alpha} \end{bmatrix}, \begin{bmatrix} \boldsymbol{A} & 0 \\ 0 & \boldsymbol{I} \end{bmatrix}\right). \tag{9}$$

For all the data generating models above, we examine situations where the marginal variance has a constant or non-constant scale.

2.2. Model performance−simulation approach

The proposed data generating models 3, 6, 8, and 9, were evaluated using a simulation approach. Simulations were based on 2500 datasets (Table 1), each of size 100, with 6 outcomes per subject for evaluating the size of the tests under the null hypothesis, and 1000 datasets for the power computations (Tables 2 and 3). We evaluate the various tests under the null hypothesis of no exposure effect (Table 1), and explore a global exposure effect of -0.16 in Table 2. In Table 3 the global exposure is assumed to be -0.80 with larger variance/scale. As for the marginal variance structure, for homogeneous outcomes we take $\boldsymbol{\sigma}^2 = (0.1, 0.1, 0.1, 0.1, 0.1, 0.1)^T$; for heterogeneous or non-constant scale we use $\boldsymbol{\sigma}^2 = (0.4, 0.4, 0.4, 0.1, 0.1, 0.1)^T$. When the LV model is correct we assume $\boldsymbol{\lambda} = (0.3, 0.3, 0.3, 0.1, 0.1, 0.1)$, which corresponds to an average correlation under model 2 of 0.37. These parameter values were selected to be similar to those observed from the birth defects example presented in our earlier work (Sammel and Ryan (1996)). The assumed models are presented as follows: (1) latent variable with model based variance (LV-SCORE); (2) latent variable with robust on sandwich variance (LV-SCORES); (3) latent variable with adjusted model-based variance (LV-SCOREA) (Rotnitzky and Jewell (1990)); (4) two-stage factor analysis (TSFA); (5) Generalized Estimating Equation (GEE) with exchangeable correlation; and (6) random effects model (RE). Model 5 is equivalent to model 6 except that it uses the robust variance adjustment. All data simulations were done using SAS IML, and analyses were conducted using SAS Proc Mixed and an IML macro written to fit the latent variable model.

Table 1. Bias and Power of Tests Under True Global Exposure Effect θ=0.

	Homogeneous Variance		Heterogeneous Variance	
Assumed Model	Bias	Rejection Rate	Bias	Rejection Rate
True data generating model (DGM) is latent variable.				
LV–SCORE	-0.001	0.051	0.000	0.055
LV–SCORES		0.049		0054
LV-SCOREA		0.050		0.056
TSFA	-0.001	0.044	0.000	0.046
GEE	-0.001	0.058	0.000	0.057
RE	0.001	0.051	0.000	0.061
DGM is compound symmetry with ρ=0.3.				
LV–SCORE	-0.003	0.042	0.002	0.046
LV–SCORES		0.043		0.045
LV–SCOREA		0.042		0.050
TSFA	-0.001	0.050	0.001	0.045
GEE	0.001	0.057	-0.001	0.054
RE	-0.001	0.053	-0.001	0.045
DGM is independent subset with ρ=0.3.				
LV–SCORE	-0.001	0.054	-0.001	0.057
LV–SCORES		0.054		0.058
LV–SCOREA		0.054		0.058
TSFA	-0.001	0.050	-0.001	0.047
GEE	0.000	0.052	0.000	0.052
RE	-0.001	0.061	0.001	0.054

Table 1 describes the test size or validity of the various tests under the assumption of no exposure effect, $\theta = 0$. The rejection rates are within sampling error of the nominal level, 0.05, and thus all proposed tests are valid.

Table 2 summarizes the 1000 simulation runs for each model where the true global exposure effect is $\theta = -0.16$. For each assumed model the bias in this global exposure is estimated, and the rejection rate or power is reported. The global exposure estimate for the latent variable model is $\sum_{m=1}^{M} \lambda_m \theta / M$, and we observe that the tests (1, 2, or 3) are all equivalent. For the GEE model the global exposure is an average over all the outcomes, $z_i\theta$. Each model is fit assuming homogeneous and heterogeneous scale for the observed outcomes. Of note is the fact that the latent variable models with different variance assumptions perform similarly under all scenarios, as does the TSFA model. Under heterogeneity of variance, models are also similar but have less power overall. The GEE model has slightly less power when the true data structure is LV, while the LV model discriminates similarly when the data have a constant correlation. When there is a subset of independent outcomes, The LV model performs poorly. However, when the uncorrelated outcomes are unaffected by the exposure, the LV model is preferred. As for the amount of bias in the estimation of the global exposure, we see that the naive two-stage factor analysis consistently underestimates the true exposure effect by 30 to 70 percent. The GEE model underestimates the global mean by 22.6 percent when the true model is LV with homogeneous variance. However, the bias for the LV model is only 1 percent when the true model generating the data is exchangeable.

Table 2. Bias and Power of Tests Under True Global Exposure Effect $\theta = -0.16$.

	Homogeneous Variance		Heterogeneous Variance	
Assumed Model	Bias (%)	Rejection Rate	Bias (%)	Rejection Rate
True data generating model (DGM) is latent variable.				
LV–SCORE	-0.004(2.4)	0.908	-0.002 (0.2)	0.764
LV–SCORES		0.912		0.764
LV–SCOREA		0.903		0.748
TSFA	-0.048 (29.7)	0.906	-0.078 (49.1)	0.760
GEE	-0.037 (23.3)	0.825	-0.066 (41.6)	0.616
RE	-0.018 (11.5)	0.895	-0.052 (32.7)	0.732
DGM is compound symmetry with ρ=0.37.				
LV–SCORE	0.002 (1.1)	0.712	0.019 (11.6)	0.706
LV–SCORES		0.714		0.704
LV–SCOREA		0.706		0.687
TSFA	-0.065 (41.2)	0.712	-0.068 (42.4)	0.703
GEE	-0.001 (0.9)	0.789	-0.002 (1.2)	0.886
RE	-0.001 (0.1)	0.768	-0.002 (1.3)	0.833
DGM is independent subset with ρ=0.37.				
	Homogeneous Variance		Heterogeneous Variance	
Assumed Model	Bias (%)	Rejection Rate	Bias (%)	Rejection Rate
LV–SCORE	-0.060 (37.2)	0.541	-0.059 (36.8)	0.385
LV–SCORES		0.542		0.382
LV–SCOREA		0.544		0.379
TSFA	-0.098 (61.5)	0.534	-0.111 (69.9)	0.379
GEE	-0.001 (0.4)	0.982	0.001 (0.6)	0.995
RE	-0.001 (0.4)	0.979	-0.001 (0.4)	0.994
DGM is independent subset with correlated outcomes affected.				
LV–SCORE	-0.005 (2.6)	0.971	-0.003 (2.0)	0.812
LV–SCORES		0.970		0.811
LV–SCOREA		0.967		0.810
TSFA	-0.056 (34.8)	0.964	-0.080 (49.7)	0.806
GEE	-0.062 (38.9)	0.703	-0.116 (72.5)	0.260
RE	-0.053 (33.5)	0.771	-0.101 (63.4)	0.371

The LV model performs consistently better when the variance of the outcomes is larger, $\boldsymbol{\sigma}^2 = (5,5,5,5,5,5)^T$ for homogeneous models, and $\boldsymbol{\sigma}^2 = (20,20,20,5,5,5)^T$ for heterogeneous or non-constant scale (Table 3). In this situation the factor loadings are $(0.4,0.4,0.4,2,2,2)^T$ corresponding to a correlation of $\rho = 0.12$. All models have similar power except when independent subset of outcomes are present where the LV model is unable to combine the information appropriately for testing. Bias in estimation of the exposure effect is similar under misspecification with the exception of the TSFA model, which underestimates the effect severely (80−90 percent).

Table 3. Bias and Power of Tests Under True Global Exposure Effect θ=-0.80.

	Homogeneous Variance		Heterogeneous Variance	
Assumed Model	Bias (%)	Rejection Rate	Bias (%)	Rejection Rate
True data generating model (DGM) is latent variable.				
LV–SCORE	-0.004(0.6)	0.748	-0.010 (1.2)	0.791
LV–SCORES		0.748		0.789
LV–SCOREA		0.736		0.796
TSFA	-0.709 (88.7)	0.748	-0.713 (89.1)	0.789
GEE	-0.054 (6.8)	0.728	0.320 (40.0)	0.791
RE	-0.022 (2.8)	0.754	0.303 (37.9)	0.786
DGM is compound symmetry with ρ=0.12.				
LV–SCORE	0.001 (0.1)	0.986	-0.069 (8.6)	0.924
LV–SCORES		0.988		0.924
LV–SCOREA		0.983		0.928
TSFA	-0.660 (82.5)	0.985	-0.069 (86.0)	0.921
GEE	0.001 (0.1)	0.992	-0.066 (8.2)	0.969
RE	-0.001 (0.1)	0.992	-0.065 (8.2)	0.966
DGM is independent subset with ρ=0.12.				
LV–SCORE	-0.239 (29.8)	0.775	-0.166 (20.7)	0.714
LV–SCORES		0.777		0.714
LV–SCOREA		0.777		0.707
TSFA	-0.714 (89.2)	0.772	-0.725 (90.7)	0.736
GEE	-0.003 (0.3)	1.000	-0.073 (9.2)	0.997
RE	-0.002 (0.3)	1.000	-0.073 (9.2)	0.997
DGM is independent subset with correlated outcomes affected.				
LV–SCORE	-0.003 (0.4)	0.999	-0.025 (3.1)	0.972
LV–SCORES		0.999		0.972
LV–SCOREA		0.998		0.972
TSFA	-0.660 (82.5)	0.999	-0.693 (86.6)	0.966
GEE	-0.235 (29.4)	0.949	-0.174 (21.8)	0.969
RE	-0.200 (24.9)	0.971	-0.168 (21.0)	0.973

2.3. Model performance−asymptotic relative efficiency

For comparisons of efficiency, we consider the robust Wald tests for the GEE model (7) compared to that of the latent variable model (5) score test, both with robust variance estimates. First, we assume an arbitrary data generating model which will take more specific forms presently, then consider the asymptotic variance of the two normally distributed tests. The tests both have the form $T_j = \mathbf{\Sigma}_j^{-1/2}\boldsymbol{S}_j$, $j = 1$ indicating the LV model and $j = 2$ the GEE model. For these tests to be comparable, they must have the same size under an arbitrary model. This condition will be satisfied if the asymptotic mean is 0 and the test is standardized to have variance 1. Because of the empirical variance used to construct both tests, our tests have this desired property.

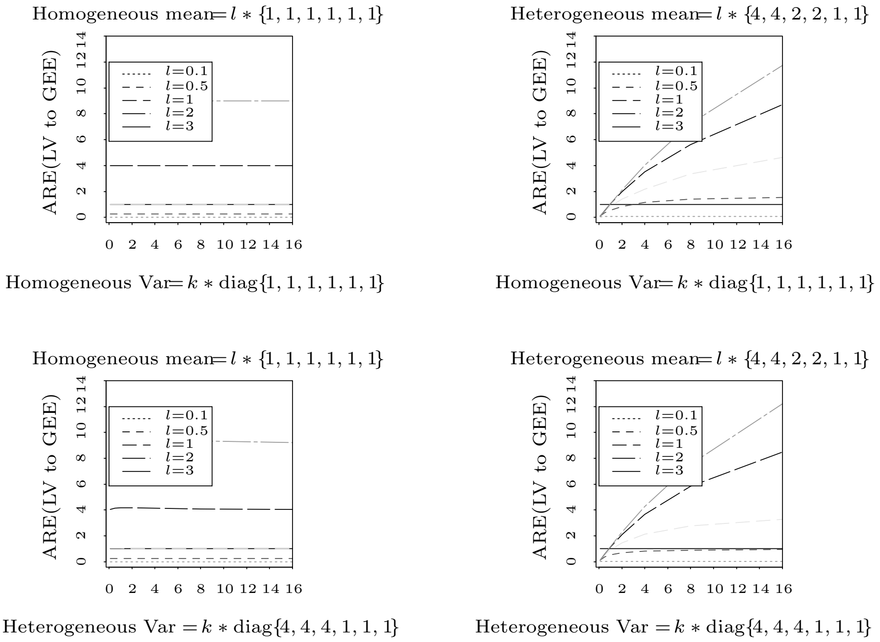
To compare the power of the two tests, we calculate Pitman's Asymptotic Relative Efficiency (ARE) under various data generating models indexed by the parameter θ. Suppose that at an arbitrary point θ in the alternative space, a test statistic T_j satisfies $\sqrt{n}\,(\mathrm{T}_j - \mu_j(\theta)) \xrightarrow{\mathcal{L}} \mathrm{N}\left[0, \sigma_j^2(\theta)\right]$. Then, using the results of Serfling ((1980), p.316), one can show that the relative efficiency of two such tests is:

$$ARE(T_1, T_2) = \frac{\mathbf{\Delta}_2}{\mathbf{\Delta}_1}, \text{ where } \mathbf{\Delta}_j = \lim_{n\to\infty}\left[\left.\frac{\sigma_j\,(\theta)}{\mu_j'\,(\theta)}\right|_{\theta=\theta_o}\right]^2.$$

Derivations of the specific forms for the asymptotic mean and variance of the tests, T_j, involve the asymptotic limits (in probability) of the estimated parameters, $\hat{\boldsymbol{\lambda}}$, $\hat{\boldsymbol{\Psi}}$, $\hat{\rho}$, $\hat{\mathbf{\Sigma}}$, given the particular data generating model. These limiting quantities, denoted by $\boldsymbol{\lambda}_o$, $\boldsymbol{\Psi}_o$, ρ_o, $\mathbf{\Sigma}_o$, have been previously described (Sammel (1995)).

In this section we evaluate Asymptotic Relative Efficiencies (AREs) under several different data generating models. Since the ARE comparisons are influenced by the amount of variability and strength of association among the outcomes, we illustrate the comparison under a variety of values, k. We specify the variance taking homogeneous as Var $= k * I\,(m)$ and heterogeneity as Var $= k * Diag\,(4,4,4,1,1,1)$ for $k = 0.1, 0.5, 1, 4, 9, 16$.

Figure 1 illustrates the ARE for the latent variable robust score test (5) relative to the GEE global score test (7). In addition to the homogeneous versus heterogeneous variance, we can also illustrate the impact when the exposure on the mean is homogeneous or heterogeneous. This is done via the factor loading vector $\boldsymbol{\lambda}$, where we assume $\boldsymbol{\lambda} = l * (1,1,1,1,1,1)^T$ for homogeneity, and $\boldsymbol{\lambda} = l * (4,4,2,2,1,1)^T$ for heterogeneity. The strength of association among the outcomes is controlled by l, where $l = (0.1, 0.5, 1, 2, 3)$.



Homogeneous mean= $l * \{1, 1, 1, 1, 1, 1\}$

Heterogeneous mean= $l * \{4, 4, 2, 2, 1, 1\}$

Homogeneous Var= $k * \text{diag}\{1, 1, 1, 1, 1, 1\}$

Homogeneous Var= $k * \text{diag}\{1, 1, 1, 1, 1, 1\}$

Homogeneous mean= $l * \{1, 1, 1, 1, 1, 1\}$

Heterogeneous mean= $l * \{4, 4, 2, 2, 1, 1\}$

Heterogeneous Var $= k * \text{diag}\{4, 4, 4, 1, 1, 1\}$

Heterogeneous Var $= k * \text{diag}\{4, 4, 4, 1, 1, 1\}$

Figure 1. ARE comparisons when true model is latent variable

When the exposure or mean effect is homogeneous, the ARE remains constant as k increases, and the values are only slightly larger when the variance is heterogeneous. The ARE is less than 1 for weakly correlated outcomes, $\boldsymbol{\lambda} = 0.1$, and 0.5; equivalent when $\boldsymbol{\lambda} = 1$; and greater than 1 for values $\boldsymbol{\lambda} > 1$. When the exposure effect is heterogeneous, the ARE is slightly stronger when the outcomes have homogeneous variance, and increases linearly as the variance increases. The latent variable test is superior except when the variability of the outcomes is small, $k < 4$ for homogeneous outcomes and $k < 9$ under scale heterogeneity.

Evaluation when GEE is correct is presented in Figure 2. The LV test has better performance when the outcomes have moderate correlation, homogeneous variance, and the ARE increases linearly with increasing variance. The LV test is most efficient for strongly correlated outcomes, $\rho = 0.6$, when $k = 2$ or higher for homogeneous outcomes. When the outcomes are moderately correlated, $\rho = 0.3$, the GEE test does slightly better when outcomes have heterogeneous variance, except for $k \geq 8$ and $k \geq 11$.

When a subset of the outcomes are independent, the latent variable model does not perform well, results are depicted in Figure 3. Under this set of assumptions, the latent variable model is superior only when the set of correlated outcomes has a very strong correlation. As illustrated in the figure, the ARE surpasses 1 when $\rho = 0.6$ and the variability is moderate to large, $k = 8$ for homogeneous outcomes, and $k = 6$ when outcomes have heterogeneous variance.

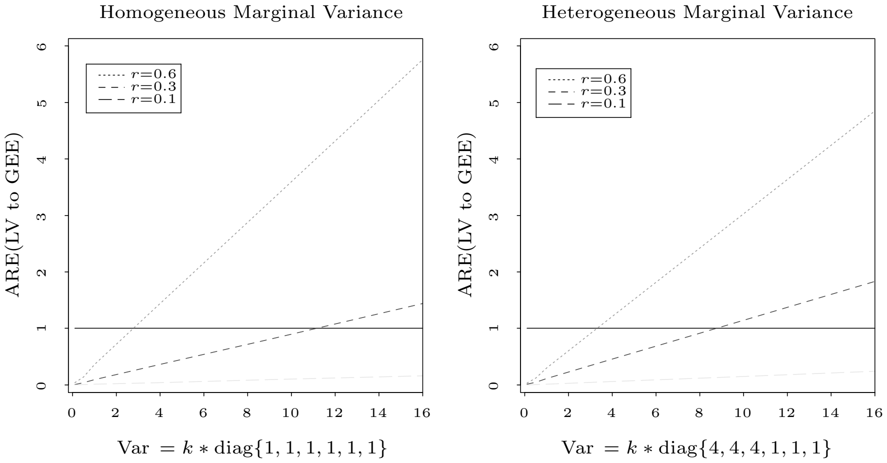


Figure 2. ARE comparisons when true variance has constant correlation

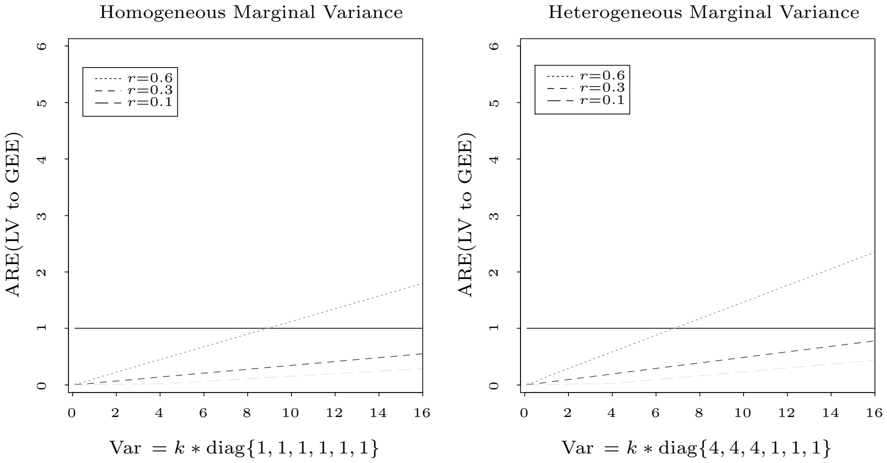


Figure 3. ARE comparisons when true model has independent subset

3. Example

Holmes, Harvey, Brown and Khoshbin (1994) discuss an observational cohort study of infants born at Brigham and Women's Hospital in Boston, MA. Our example considers two subgroups from this study, infants whose mothers are epileptics who took medications during their pregnancy (exposed subjects), and control infants, whose mothers were randomly chosen from those who gave birth at the same hospital at the same time as the exposed mothers. Various outcomes were assessed on the infants including weight, measures of size, and a variety of cranial and limb measurements. Variables include (in order) bitemporal (side to side) head diameter, nose length, finger length, weight, anterior-posterior (front to back) head diameter, and upper lip width. For consideration in the latent variable model, the best grouping of variables would be of those which are moderately correlated (conceptually) and have been implicated in previous literature to be influenced by the exposure. Items which are too strongly associated with one another would dominate the latent variable score. If outcomes are uncorrelated with one another, then GEE methods are more appropriate. In our example, we anticipate that the various size measurements would meet model requirements.

Table 4 shows some summary statistics for the 628 infants, for a subset of continuous measurements that will be used to illustrate the tests developed in the paper. Means and standard errors are given for control and exposed infants, along with the estimated exposure effect based on a linear regression model that also adjusts for gender and gestational age. Exposure to anticonvulsant medications resulted in a decrease in all outcomes of interest except for upper lip width. Wider upper lip has been associated with the characteristic “anticonvulsant face” which reflect subtle abnormalities of exposure. Pearson correlation coefficients among the outcomes are also presented. All outcomes are positively correlated, while lip width is relatively uncorrelated with the other outcomes. For the remaining outcomes, correlations range between 0.19 and 0.60.

Table 4. Summary Statistics

	Exposed		Control			
	n=176		n=452			
Outcome (cm)	Mean	(SE)	Mean	(SE)	$\hat{\beta}^*$	p-value*
Head-bt	9.10	(0.059)	9.40	(0.024)	-0.212	<0.001
Nose length	1.94	(0.010)	1.98	(0.015)	-0.031	0.104
Finger length	2.86	(0.021)	2.95	(0.017)	-0.045	0.103
Weight(kg)	3.28	(0.005)	3.41	(0.026)	-0.004	0.917
Head-ap	11.67	(0.055)	11.74	(0.027)	-0.007	0.880
Upper lip width	0.91	(0.009)	0.89	(0.006)	0.017	0.113

*Adjusted for gender and gestational age using linear regression.

Correlation Matrix

	Head-bt	Nose	Finger	Weight	Head-ap	Lip
Head-bt	1.000					
Nose length	0.249	1.000				
Finger length	0.275	0.055	1.000			
Weight(kg)	0.596	0.250	0.464	1.000		
Head-ap	0.542	0.189	0.213	0.542	1.000	
Upper lip width	0.029	0.039	0.100	0.179	0.100	1.000

Both the latent variable and GEE models are fit to subsets of the outcomes in Table 5. Estimates of the global exposure effect are presented, as well as p-values for the two tests, one assuming the covariance of the model is correct (model-based tests) described for the LV model as (4) or as the random effects (RE) model. The other (robust) tests use the sandwich variance estimate to protect against variance misspecification are described as (5) for the LV model and as (7) for the GEE. The final column reflects the assumption of independence for the assumed variance. Restricted maximum likelihood estimates (REML) of variance parameters have been used in this example (Harville (1977)). The subset of outcomes presented in Model 1 represent a set of correlated outcomes each with a modest exposure effect. The estimate of the exposure effect is slightly larger for the LV model, however, conclusions about the exposure effect are similar for the GEE model under both the independence and compound symmetry variance assumptions. In model 2 weight and anterior/posterior head diameter (head-ap) are added to the model. These outcomes are strongly correlated with one another and the other outcomes in the model, while mean levels do not differ between the two exposure groups. The result of adding these two outcomes to the model is a decrease in the global exposure effect for both models, and a parallel decrease in the significance level of the tests. The robust test GEE model is influenced the least by the addition of these outcomes. The model with the working assumption of independence is impacted the least by the addition of these outcomes. This result is consistent with both the simulation and asymptotic comparisons of the previous sections.

Table 5. Estimation and tests for global exposure effect

	Latent	GEE	GEE
	Variable	CS	Independence
Model 1: Head-bt, Nose length and Finger length			
Global Exposure estimate	-0.093	-0.050	-0.052
Model based test p-value	<0.001	0.001	<0.001
Robust test p-value	<0.001	0.002	<0.001
Model 2: Head-bt, Nose length, Finger length, Weight and Head-ap			
Global Exposure estimate	-0.052	-0.037	-0.044
Model based test p-value	0.025	0.028	0.001
Robust test p-value	0.053	0.016	0.010
Model 3: Head-bt, Nose length, Finger length, and -1*(Upper lip)			
Global Exposure estimate	-0.073	-0.031	-0.032
Model based test p-value	<0.001	0.001	0.001
Robust test p-value	<0.001	0.002	0.001

Model 3 incorporates the uncorrelated outcome upper lip width to the subset of outcomes in Model 1. For consistency in the estimate of exposure effect we have included the negative of the lip measurement, -1*(uplip), to the model. The addition of this outcome does not significantly effect the tests of significance when compared to Model 1, but the effect estimates for both models has decreased. Surprisingly, in this situation the latent variable model test has not been effected, although the computational results indicate it is less efficient.

4. Conclusions

This paper has focused on the comparison of several approaches to testing and estimating covariate effects on multiple outcomes. The first approach was based on a latent variable model and assumes that the covariates of interest affect outcomes through an underlying latent structure, in a method similar to factor analysis. The other tests, generalized estimating equations and random effects test assume an average effect of exposure over all outcomes.

We derived the distribution of the tests under general assumptions for the distribution of the outcomes, then compared the relative performance of the tests under various true models. The latent variable model and the GEE model have similar detriments in power under misspecification. Bias of the global exposure effect is more severe under the GEE model. The latent variable model is inferior when the variability of the outcomes is very small, or the correlation among the outcomes is weak. Care in modeling is warranted when there are uncorrelated subsets of outcomes.

The latent variable model is efficient even when the true correlation structure is exchangeable, when the variability of the outcomes is large, $k \geq 8$, and when that variability/scaling is heterogeneous, and for modestly correlated outcomes, $\rho \geq 0.3$. If uncorrelated outcomes are included, the latent variable model outperforms the GEE only when the outcomes are highly correlated, $\rho \geq 0.6$ and variance moderate to large, $k \geq 6$, or when only the correlated outcomes are impacted by the covariate. Our findings suggest that when used in the right setting, the latent variable model can provide a powerful and robust approach to the analysis of multiple outcome data.

In addition to testing covariates, the latent variable model allows for estimation of the latent outcome, which is an overall summary measure or ranking of subjects, i.e., severity score for the severity of birth defects in our example. This score gives a relative ranking of the subjects where the effect of the covariates influences the relative ranking.

Acknowledgements

This research was supported in part by NIEHS grant ES07142 and NCI grant CA48061. Our thanks to Dr. Lewis Holmes for the use of his data, and to Ms. Carrie Wager and Dr. Anita Weber for programming assistance.

Appendix A

This appendix provides details on deriving the two tests for exposure. The asymptotic variance of $\boldsymbol{\theta}_{lv}$ may be found by inverting the expected information from the marginal log-likelihood from the latent variable model described in Section 1.1. For $\boldsymbol{\zeta}^*=(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\lambda}, \boldsymbol{\Psi})$, the expected information can be partitioned as

$$I\left(\boldsymbol{\zeta}^*\right) = \begin{bmatrix} I_{\theta\theta} & I_{\theta\alpha} & I_{\theta\lambda} & I_{\theta\sigma^2} \\ I_{\theta\alpha}^T & I_{\alpha\alpha} & I_{\alpha\lambda} & I_{\alpha\sigma^2} \\ I_{\theta\lambda}^T & I_{\alpha\lambda}^T & I_{\lambda\lambda} & I_{\lambda\sigma^2} \\ I_{\theta\sigma^2}^T & I_{\alpha\sigma^2}^T & I_{\lambda\sigma^2}^T & I_{\sigma^2\sigma^2} \end{bmatrix}, \qquad (10)$$

and $\boldsymbol{\sigma}^2 = \left(\sigma_1^2, \ldots, \sigma_M^2\right)^T$ is the vector of diagonal elements of $\boldsymbol{\Psi}$. Evaluation of specific elements of this matrix reveals that $I_{\theta\sigma^2}$ and $I_{\alpha\sigma^2}$ are zero.

For testing, we need to compute the information under the null hypothesis H_o: $\boldsymbol{\theta} = 0$, in which case, $I_{\theta\lambda}$, $I_{\alpha\lambda}$, and $I_{\sigma^2\lambda}$ are also zero. Therefore, at H_o, the information (10) has form

$$I\left(\boldsymbol{\zeta}^*\right) = \begin{bmatrix} I_{\theta\theta} & I_{\theta\alpha} & 0 & 0 \\ I_{\theta\alpha}^T & I_{\alpha\alpha} & 0 & 0 \\ 0 & 0 & I_{\lambda\lambda} & 0 \\ 0 & 0 & 0 & I_{\sigma^2\sigma^2} \end{bmatrix}. \qquad (11)$$

This is a block diagonal matrix where the mean parameters are asymptotically independent of the variance components. Hypothesis tests on the mean parameters $\boldsymbol{\mu} = (\boldsymbol{\theta}, \boldsymbol{\alpha})$ are of primary interest, in particular the null hypothesis $H_o : \boldsymbol{\theta} = 0$. Therefore, the model-based variance is V_{11} the first element of $V = I\left(\boldsymbol{\zeta}^*\right)^{-1}$.

This variance form is used to conduct a score test for the global effect of the latent variable, described at (5). For a "robust" version of the variance we define

$$W = V \left[\sum_{i=1}^{n} D_i^T \left(\boldsymbol{\lambda}\boldsymbol{\lambda}^{\boldsymbol{T}} + \boldsymbol{\Psi}\right)^{-1} \text{Var } (Y) \left(\boldsymbol{\lambda}\boldsymbol{\lambda}^T + \boldsymbol{\Psi}\right)^{-1} D_i \right] V$$

for $D = \left(\boldsymbol{\lambda}\boldsymbol{Z}_i \ \boldsymbol{X}_i\right)^T$ (Rotnitzky and Jewell (1990)). In practice a moment estimate of W is used utilizing the estimated mean model.

References

Bull, S. D. (1998). Regression models for multiple outcomes in large epidimiologic studies. *Statist. Medicine* **17**, 2179-2197.

Freeman, E. W., Grisso, J. A., Berlin, J., Sammel, M. D., Garcia-Espana, B. and Hollander, L. (2001). Symptom reports from a cohort of African American and Caucasian women in the late reproductive years. *Menopause*, **8**, 33-42.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. Amer. Statist. Assoc.* **72**, 320-340.

Heyer, N. J., Bittner, A. C. and Echeverria, D. (1996). Analyzing multivariate neurobehavioral outcomes in occupational studies: a comparison of approaches. *Neurotoxicology and Teratology* **18**, 401-406.

Holmes, L. B., Harvey, E. A., Kleiner, B. C., Leppig, K. A., Cann, C.I., Muñoz, A. and Polk, B. F. (1987). Predictive value of minor anomalies: II. Use in cohort studies to identify teratogens. *Teratology* **36**, 291-297.

Holmes, L. B., Harvey, E. A., Brown, K. S. and Khoshbin, S. (1994). Anticonvulsant teratogenesis: 1. A study design for newborn infants. *Teratology* **49**, 202-207.

Laird, N. M and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38,** 963-974.

Laska, E. M., Tang, D. I., and Meisner, M. J. (1992). Testing hypotheses about an identified treatment when there are multiple endpoints. *J. Amer. Statist. Assoc.* **87**, 825-831.

Lefkopoulou, M. and Ryan, L. M. (1993). Global tests for multiple binary outcomes. *Biometrics* **49,** 975-988.

Legler, J. M., Lefkopoulou, M. and Ryan, L. M. (1995). Efficiency and power of tests for multiple binary outcomes. *J. Amer. Statist. Assoc.* **90,** 680-693.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

Rogers, G. S. (1980). *Matrix Derivatives.* Marcel Dekker, New York.

Rotnitzky, A. and Jewell, N. P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika* **77,** 485-497.

Sammel, M. D. and Ryan, L. M. (1996). Latent variable models with fixed effects. *Biometrics* **52**, 650-663.

Sammel, M. D. (1995). Latent variable models for multiple outcomes. Doctoral thesis, Department of Biostatistics, Harvard School of Public Health.

Serfling, R. J. (1980). *Approximation Theorems for Mathematical Statistics.* John Wiley, New York.

Streiner, D. L. and Norman, G. R. (1995) *Health Measurement Scales: A Practical Guide to Their Development and Use*, 2nd Edition. Oxford University Press, Oxford.

Zeger, S. L. and Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121-130.

Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine,
432 Guardian Drive, Room 605, Philadelphia, PA 19104-6021, U.S.A.

E-mail: msammel@cceb.upenn.edu

Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Room
409, Boston, MA 02115, U.S.A.
E-mail: lryan@hsph.harvard.edu

(Received October 2000; accepted May 2002)