American Journal of Biostatistics 1 (1): 32-41, 2010
ISSN 1948-9889

Simulation Study for Evaluating a New Three Stages Procedure for Selecting the Right Multivariate Regression Model

Ali Hussein AL-Marshadi
Department of Statistics, King Abdul Aziz University, Jeddah, Saudi Arabia

Abstract: Problem statement: This article considers the analysis of multivariate regression experiment that is used frequently in variety of applications research such as psychiatric epidemiologic studies. Our study concerned with multivariate regression model in which the responses were correlated in particular ways for both standard and non-standard multivariate model structures. Our objective is to find reliable procedure that can be used to guide the selection of the best multivariate regression model that has the right covariance structure and in the same time has the right multivariate model structure for both standard and non-standard multivariate model structures. **Approach:** In this study, we were proposing and evaluating a new three stages procedure that could be used to guide the selection of the best multivariate regression model that has the right covariance structure and in the same time has the right multivariate model structure using bootstrap simulation procedure. **Results:** The simulation results indicated that the performance of the new procedure in identifying the right multivariate regression model that has the right covariance structure and in the same time the right multivariate model structure from both standard and non-standard multivariate model structures was excellent overall. **Conclusion/Recommendations:** We recommended using the new procedure as stander tools to guide the selection of the best multivariate regression model that has the right covariance structure and in the same time has the right multivariate model structure.

Key words: Multivariate linear regression, information criteria, bootstrap technique, MCB procedure

INTRODUCTION

The multivariate linear regression model composed of multiple correlated dependent variables for each subject, in addition to a set of predictor variables. Multivariate linear regression allows researchers to fit a single model for each response, taking into account the correlation among the multiple responses on a given subject. The basic assumptions of multivariate regression are multivariate normality of the residuals, homogenous variances of residuals conditional on predictors, common covariance structure across observations and independent observations. When these assumptions are satisfied, the coefficients will be unbiased, the least-squares estimates will have minimum variance and the relationships among the coefficients will reflect the relationships among the predictors. In general, this is what REG procedure of the SAS System is set up to do. When we deal with the multivariate linear regression model a companion to the estimation problem is the model selection problem, which consists of choosing an appropriate model from a class of candidate models to characterize the data under study. The covariance structures of the observed multiple responses makes multivariate linear regression data analysis different from univariate multiple linear regression data in term of the prediction of an individual response component given some or all of the remaining components (McCullagh, 2006).

Although the MIXED procedure of the SAS System is used as tools for fitting mixed effects and repeated measures models, it is also a very useful tool for fitting multivariate regression. The most advantages of using the MIXED procedure instead of stander multivariate procedure are MIXED uses observations have incomplete responses, Mixed has the ability to deal with non-stander (e.g., multiple design) multivariate models and MIXED enables researchers to fit correlated error model with different covariance structure. The MIXED procedure of the SAS System has different selections for modeling the covariance structure. The MIXED procedure of the SAS System can be used to develop either Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML) estimates in order to complete the analysis of the multivariate regression, where REML estimation are generally preferred to ML. A lot of effort is usually needed to decide what the suitable covariance structure of the data is at the beginning of the statistical analysis. Statisticians often use information criteria such as AIC

(Akaike, 1974), BIC (Schwarz, 1978), CAIC (Bozdogan, 1987), HQIC (Hannan and Quinn, 1979) to guide the selection of the covariance structure in mixed models (AL-Marshadi, 2007; Keselman *et al*., 1999; Littell *et al*., 2000; Singer, 1998) Many studies have investigated performance of those information criteria in selection of the covariance structure considering repeated measures models (Keselman *et al*., 1999; Ferron *et al*., 2002; Gomez *et al*., 2005; Guerin and Stroup, 2000). One study compared the following, AIC, BIC, AICC and RIC to select stander multivariate regression model with the stander covariance structure (Azari *et al*., 2006). A simulation study by Beal (2005) compared different information criteria methods in SAS for selecting the right multiple linear regression models for large sample size. Another research by Seghouane (2006) had developed and compared a new small sample model selection criterion for multivariate regression models to other known criterion with the stander covariance structure. A new paper proposed new criterion deals with selection of variables in multivariate linear regression models with fewer observations than the dimension by using Akaike's Information Criterion (AIC) (Yamamura *et al*., 2008). A psychiatric epidemiologic study proposed multivariate linear regression as the preferred method when the multiple informant outcome data are continuous using MIXED procedure in SAS (Goldwasser and Fitzmaurice, 2001). In our resent simulation study, we were concerned with multivariate regression model in which the responses were correlated in particular ways for both standard and non-standard multivariate model structures. Our goal in that study was evaluating six model selection criteria in SAS using MIXED procedure to guide the selection of the best multivariate regression model that has the right covariance structure and in the same time has the right multivariate model structure for both standard and non-standard multivariate model structures. The study indicated that the percentages of identifying the right multivariate regression model from both standard and non-standard multivariate model structures were very low overall, except for specific models that involve indicator variable (AL-Marshadi, 2009b).

In the current study we are still concerned with the multivariate regression model in which the responses were correlated in particular ways for both standard and non-standard multivariate model structures. Our objective is to find reliable procedure that can be used to guide the selection of the best multivariate regression model that has the right covariance structure and in the same time has the right multivariate model structure for both standard and non-standard multivariate model structures using MIXED procedure in SAS.

In general form, the mixed effects linear model can be written as (McCulloch and Searle, 2001; Littell *et al*., 1996):

$$Y = X\beta + ZU + e \quad (1)$$

Where:
β = p×1 vector of fixed effects
U = q×1 vector of random effects
e = n×1 vector of residuals
X = n×p design matrix for fixed effects
Z = n×p design matrix for random effects

$$U\sim N(0,G),\ e\sim N(0,R),$$
$$Y\sim N(X\beta,V) \text{ and } V = ZGZ'+R$$

When V is known, the Best Linear Unbiased Estimators (BLUE) of estimable functions $h'\beta$ of the fixed effects in (1) are given by:

$$h'\hat{\beta} = h'(X'V^{-1}X)^{-}X'V^{-1}Y \quad (2)$$

With:

$$\text{var}(h'\hat{\beta}) = h'(X'V^{-1}X)^{-}h \quad (3)$$

In most applications V is unknown. Therefore, it is estimated from the data where estimators based on (2) are not generally BLUE. Various procedures were proposed for testing hypotheses on fixed effects in mixed models with unknown V, most of which assume that V is estimated by the REML method (Fai and Cornelius, 1996; Giesbrecht and Burns, 1985; Kenward and Rogers, 1997). Standard error estimates based on (3) are biased downwards when V replaced by its estimate (Kacker and Harville, 1984). Fixed effects are estimated based on (2), with V replaced by a plug-in REML estimate. Null hypotheses of the form $H_0 : h'\beta = 0$ are tested by:

$$F = \frac{\hat{\beta}'h[h'(X'\hat{V}^{-1}X)^{-}h]^{-1}h'\hat{\beta}}{\text{rank}(h)} \sim F_{(\text{rank}(h),\upsilon)} \quad (4)$$

when rank(h)>1in general, the test statistics in (4) only have approximate F-distribution. The approximate denominator degree of freedom υ of F-distribution can be determined using one of the four different methods implemented in MIXED procedure of SAS. The four methods of the approximations are residual method, containment method (this is the default in MIXED), extended Satterthwaite (1941) method of Giesbrecht and Burns (1985); Fai and Cornelius (1996) and Kenward-Roger method. Kenward and Roger (1997)

found good performance of their method across a number of designs. Also, Guerin and Stroup (2000) recommended using the Kenward-Roger method as standard operating procedure. Therefore, Kenward-Roger method was considered in this study for approximating the denominator degrees of freedom.

MATERIALS AND METHODS

The following model reflects the standard multivariate linear model:

$$Y = X\beta+e \quad (5)$$

Where:

- Y = n×r matrix of r response variables measured on n subjects
- X = n×p design matrix for explanatory variables
- β = p×r matrix of regression coefficients
- e = n×r matrix of residuals whose rows are iid normal, i.e., the rows of e~N(0,Σ)

In the standard multivariate linear regression model the response distribution is $Y\sim N(X\beta I_n,\otimes\Sigma)$, where the parameter space consists of the coefficient matrix β of order p×r plus the covariance matrix, $\Sigma\in PD_r$, the cone of symmetric positive semi-definite r-matrices. This multivariate regression setting is called the standard multivariate regression model because both components of the parameter space are unconstrained. In the following we give an example showing the format of the standard multivariate linear model (REG format) and its relationship to the MIXED format. The example considers the format with two response variables and one explanatory variable in addition to an intercept term for three subjects:

$$\begin{bmatrix} y1_1 & y2_1 \\ y1_2 & y2_2 \\ y1_3 & y2_3 \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{bmatrix} \begin{bmatrix} \beta_{01} & \beta_{02} \\ \beta_{11} & \beta_{12} \end{bmatrix} + \begin{bmatrix} e1_1 & e2_1 \\ e1_2 & e2_2 \\ e1_3 & e2_3 \end{bmatrix}$$

To use the MIXED format, we need to write Y, β and e as vectors and rearrange X accordingly as follow:

$$\begin{bmatrix} y1_1 \\ y2_1 \\ y1_2 \\ y2_2 \\ y1_3 \\ y2_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & x_1 & 0 \\ 0 & 1 & 0 & x_1 \\ 1 & 0 & x_2 & 0 \\ 0 & 1 & 0 & x_2 \\ 1 & 0 & x_3 & 0 \\ 0 & 1 & 0 & x_3 \end{bmatrix} \begin{bmatrix} \beta_{01} \\ \beta_{02} \\ \beta_{11} \\ \beta_{12} \end{bmatrix} + \begin{bmatrix} e1_1 \\ e2_1 \\ e1_2 \\ e2_2 \\ e1_3 \\ e2_3 \end{bmatrix}$$

The MIXED format in matrix notation is $\tilde{Y} = \tilde{X}\tilde{\beta} + \tilde{e}$, which is a special case of the mixed model (1) (Singer, 1998). The situation of interest in this study is one in which the responses are correlated in particular ways. Different covariance matrix structures of ∑ were used to simulate correlated error models for the simulated study data.

In this study we are proposing and evaluating a new three stages procedure which could be used to guide the selection of the best multivariate regression model that has the right covariance structure and in the same time has the right multivariate model structure. The new procedure is based on three stages using the MIXED procedure. At the first stage, the right multivariate model structure will be determined with a new model selection criterion using the bootstrap simulation procedure. The new model selection criterion will be called Variance Information Criterion (VARIC). At the second stage, the right covariance structure will be determined with the model selection criteria available in MIXED procedure using the bootstrap simulation procedure considering the right multivariate model structure that was determined in the first stage. At the third stage, the right multivariate model structure and right covariance structure that determined in the first and second stage will be used to fit the right multivariate model using MIXED procedure. Also, the study involves comparing the performance of the model selection criteria that are available in MIXED procedure using the bootstrap simulation procedure to determine the right covariance structure.

The first stage of the procedure concerns with evaluating our new model selection criterion in terms of its ability to identify the appropriate multivariate model structure with the help of the bootstrap simulation procedure where the corresponding value of the Variance Information Criterion (VARIC) calculated as the variance of the absolute value for the residual of the multivariate regression model when the multivariate model was fitted with the unstructured covariance structure using MIXED procedure. We may use the new information criterion to guide the selection of the right model structure such as selecting the model structure with the smallest value of the new information criterion.

The bootstrap simulation procedure for the first and second stage involves using the bootstrap technique (Efron, 1983; 1986) and the Multiple Comparisons with the Best (MCB) procedure (Hsu, 1984) as tools to help the information criterion in identifying the right multivariate model structure in the first stage and

identifying the right covariance structure in the second stage. The idea of the new approach can be justified and applied in a very general context, one which includes the selection of the right multivariate model structure and the selection of the right covariance structure (AL-Marshadi, 2007; 2009a). The idea of using the bootstrap to improve the performance of a model selection rule was introduced by Efron (1983; 1986) and is extensively discussed by Efron and Tibshirani (1993). Recent studies applied the bootstrap technique with different approaches to select the best model in different context (AL-Marshadi, 2007; 2009a; Uraibi *et al*., 2009).

In the context of multivariate regression models, (5), the algorithm for using parametric bootstrap in our bootstrap simulation procedure for the selection of the right multivariate model structure in the first stage can be outlined as follows:

Let the observation vector O_i is defined as follows:

$$O_i = \left[y_{i1} \quad ... \quad y_{i\,r} \quad x_{i1} \quad ... \quad x_{i\,p-1} \right]`$$

where i = 1,2,…,n.

- Generate the bootstrap sample on case-by-case using the observed data (original sample) i.e., based on resampling from $(O_1,O_2,\ldots,O_n)$. The bootstrap sample size is taken to be the same as the size of the observed sample (i.e., n). The properties of the bootstrap when the bootstrap sample size is equal to the original sample size are discussed by Efron and Tibshirani (1993)
- Fit all the class of candidate multivariate regression model structures, which we would like to select the right multivariate model structure from, to the bootstrap data with the unstructured covariance structure, thereby obtaining the bootstrap value of the new information criteria VARIC* for each multivariate model structure
- Repeat the first and the second steps (W) times
- Statisticians often use the information criteria in MIXED procedure into guide the selection of the true model structure such as selecting the model structure with the smallest value of the information criteria (Keselman *et al*., 1999; Littell *et al*., 2000; Singer, 1998). We will follow the same rule in our information criteria, but we have the advantage that our information criterion has (W) replication values result of the bootstrapping of the observed data (from the first three steps). To make use of this advantage, we propose using MCB procedure (Hsu, 1984) to pick the winners (i.e., selecting the best set of models or single model if possible), when we consider the bootstrap replicates of the information criteria, that is produced by each of the model structure, as groups. The value of W = 10 was used for this study as suggested in our pervious simulation study (AL-Marshadi, 2009a). The general linear mixed effects model approach was used to pick the winners using MCB procedure in MIXED procedure in order to accommodating the violation of the equal variances assumption that was exist in the analysis of this study as suggested in (AL-Marshadi, 2008).

The simulation setup of the experiment is described below:

There are seven correlated response variables (y_1, y_2, y_3, y_4, y_5, y_6 and y_7) which are related to two predictor variables (x_1 and x_2) with five different multivariate model structures and seven different covariance structures for the seven correlated response variables. The multivariate model structures of the simulated experiment are described as follow:

- The first multivariate model structure is a standard multivariate model structure which fits seven intercepts (one for level of responses), seven slopes for x_1 and seven slops for x_2 (plus the elements of the covariance matrix of the multiple responses) as follow:

$$\begin{aligned} y_1 &= \beta_{01} + \beta_{11}x_1 + \beta_{21}x_2 + e \\ y_2 &= \beta_{02} + \beta_{12}x_1 + \beta_{22}x_2 + e \\ y_3 &= \beta_{03} + \beta_{13}x_1 + \beta_{23}x_2 + e \\ y_4 &= \beta_{04} + \beta_{14}x_1 + \beta_{24}x_2 + e \\ y_5 &= \beta_{05} + \beta_{15}x_1 + \beta_{25}x_2 + e \\ y_6 &= \beta_{06} + \beta_{16}x_1 + \beta_{26}x_2 + e \\ y_7 &= \beta_{07} + \beta_{17}x_1 + \beta_{27}x_2 + e \end{aligned}$$

- The second multivariate model structure is a standard multivariate model structure which fits seven intercepts (one for level of responses) and seven slopes for x_1 (plus the elements of the covariance matrix of the multiple responses) as follow:

$$\begin{aligned} y_1 &= \beta_{01} + \beta_{11}x_1 + e \\ y_2 &= \beta_{02} + \beta_{12}x_1 + e \\ y_3 &= \beta_{03} + \beta_{13}x_1 + e \\ y_4 &= \beta_{04} + \beta_{14}x_1 + e \\ y_5 &= \beta_{05} + \beta_{15}x_1 + e \\ y_6 &= \beta_{06} + \beta_{16}x_1 + e \\ y_7 &= \beta_{07} + \beta_{17}x_1 + e \end{aligned}$$

- The third multivariate model structure is a standard multivariate model structure which fits seven intercepts (one for level of responses) and seven slops for x_2 (plus the elements of the covariance matrix of the multiple responses) as follow:

$$
\begin{aligned}
y_1 &= \beta_{01} + \beta_{11}x_2 + e\\
y_2 &= \beta_{02} + \beta_{12}x_2 + e\\
y_3 &= \beta_{03} + \beta_{13}x_2 + e\\
y_4 &= \beta_{04} + \beta_{14}x_2 + e\\
y_5 &= \beta_{05} + \beta_{15}x_2 + e\\
y_6 &= \beta_{06} + \beta_{16}x_2 + e\\
y_7 &= \beta_{07} + \beta_{17}x_2 + e
\end{aligned}
$$

- The forth multivariate model structure is a non-standard multivariate model structure which is called "multiple design". It allows each response variable to have a different set of explanatory variables as follow:

$$
\begin{aligned}
y_1 &= \beta_{01} + \beta_{11}x_1 + e\\
y_2 &= \beta_{02} + \beta_{12}x_1 + e\\
y_3 &= \beta_{03} + \beta_{13}x_2 + e\\
y_4 &= \beta_{04} + \beta_{14}x_1 + \beta_{24}x_2 + e\\
y_5 &= \beta_{05} + \beta_{15}x_1 + \beta_{25}x_2 + e\\
y_6 &= \beta_{06} + \beta_{16}x_1 + \beta_{26}x_2 + e\\
y_7 &= \beta_{07} + \beta_{17}x_1 + \beta_{27}x_2 + e
\end{aligned}
$$

- The fifth multivariate model structure is also a non-standard multivariate model structure which is called "multiple design". It allows each response variable to have a different set of explanatory variables as follow:

$$
\begin{aligned}
y_1 &= \beta_{11}x_2 + e\\
y_2 &= \beta_{12}x_1 + e\\
y_3 &= \beta_{03} + \beta_{13}x_1 + \beta_{23}x_2 + e\\
y_4 &= \beta_{14}x_1 + e\\
y_5 &= \beta_{05} + \beta_{15}x_1 + \beta_{25}x_2 + e\\
y_6 &= \beta_{06} + \beta_{16}x_1 + \beta_{26}x_2 + e\\
y_7 &= \beta_{27}x_2 + e
\end{aligned}
$$

MIXED procedure is a very useful tool for fitting multivariate regression in which users find five model selection criteria available, which give users tools can be used to select an appropriate covariance structure for a multivariate regression model (Littell *et al*., 1996). The five model selection criteria are:

- Akaike (1974) Information Criterion (AIC)
- Schwarz (1978) Bayesian Information Criterion (BIC)
- Bozdogan (1987) Corrected Akaike Information Criterion (CAIC)
- Hannan and Quinn (1979) Information Criterion (HQIC)
- Hurvich and Tsai (1989) the Akaike's Information Corrected Criterion (AICC)

The second stage of the procedure concerns with comparing the five information criteria available in MIXED procedure in terms of their ability to identify the right covariance structure with the help of the bootstrap simulation procedure considering the right multivariate model structure that was determined in the first stage. The multivariate model structures that were considered in the first stage involve both standard and non-standard multivariate model structures i.e., from multivariate model structures with both "single design" such as the first three model structures and "multiple design" such as the last two model structures that were considered in this study. The algorithm for using the bootstrap simulation procedure for the selection of the right covariance structure in the second stage was applied in similar way as the one explained in the first stage. Seven covariance structures were considered in the second stage. The seven covariance structures were Independent Errors (VC), Compound Symmetry (CS), Heterogeneous Compound Symmetry (CSH), First-Order Autoregressive (AR(1)), Heterogeneous First-Order Autoregressive (ARH(1)), Banded Main Diagonal (UN(1)) and Unstructured (UN).

The multivariate regression analyses for the first multivariate model structure design can be implemented by the following example SAS code (Singer, 1998):

```
PROC MIXED DATA = one;
CLASS time;
MODEL y = time time*x1 time*x2/noint notest ddfm = kr;
REPEATED time / type = UN subject = subject;
```

The multivariate regression analyses for the second multivariate model structure design can be implemented by the following example SAS code (Singer, 1998):

```
PROC MIXED DATA = one;
CLASS time;
MODEL y = time time* x1 / noint notest ddfm = kr;
REPEATED time / type = UN subject = subject;
```

The multivariate regression analyses for the third multivariate model structure design can be implemented by the following example SAS code (Singer, 1998):

```
PROC MIXED DATA = one;
CLASS time;
MODEL y = time time*x2 / noint notest ddfm = kr;
REPEATED time / type = UN subject = subject;
```

Note: The class variable "time" in the first, second and third structure is used to identify the multiple responses.

The multivariate regression analyses for the forth multivariate model structure design can be implemented by the following example SAS code (Singer, 1998):

```
PROC MIXED DATA = one;
CLASS time;
MODEL y = time1 time1*x1 time2 time2*x1 time3 time3*x2 time4 time4*x1 time4*x2 time5 time5*x1 time5*x2 time6 time6*x1 time6*x2 time7 time7*x1 time7*x2 / noint notest ddfm = kr;
REPEATED time / type = UN subject = subject;
```

The multivariate regression analyses for the fifth multivariate model structure design can be implemented by the following example SAS code (Singer, 1998):

```
PROC MIXED DATA = one;
CLASS time;
MODEL y = time1*x2 time2*x1 time3 time3*x1 time3*x2 time4*x1 time5 time5*x1 time5*x2 time6*x1 time6*x2 time7*x2 / noint notest ddfm = kr;
REPEATED time / type = UN subject = subject;
```

Note: The variable "time" is replaced in the forth and fifth structure by individual 0-1 dummy variables, one for each responses variable.

The simulation study: A simulation study of multivariate regression data was conducted to evaluate the new three stages procedure in terms of the percentage of number of times that it identified the best multivariate regression model that has the right covariance structure and in the same time has the right multivariate model structure.

Correlated multivariate normal data were generated according to MIXED format model. There were 35 scenarios to generate data involving five multivariate regression model structures and seven covariance structures with one setting of covariance matrix parameter values for each covariance structure and sample sizes 40 (n = 40 subjects). The 7 settings of covariance matrix parameter values are given in Table 1. For each scenario, we simulated 150 datasets. SAS code was written to generate the datasets according to the described setup using the SAS®9.1.3 package (SAS Institute Inc., 2008). We will consider the case when we have 12 subjects as an example to explain the process of generating the datasets. A 12 7×1 vector of standard normal random deviates were generated using SAS's NORMAL function. Denoted the vector:

$$\varepsilon_i = \left[\varepsilon_{1i}\varepsilon_{2i}\varepsilon_{3i}\varepsilon_{4i}\varepsilon_{5i}\varepsilon_{6i}\varepsilon_{7i}\right]'$$

where, $i = 1,2,3,...,12$. Note that the 12 represents the 12 subjects and the 7 represents the 7 levels of time effect within each subject. Then the 12 7×1 vectors of residuals for model (5) were calculated as:

$$e_i = \Sigma^{\frac{1}{2}}\varepsilon_i \; ; \; i = 1,2,3,...,12$$

Where:

$\Sigma^{\frac{1}{2}}$ = The Cholevsky decomposition of Σ
Σ = The covariance matrix of multiple response variable

Therefore, the vector e_i is defined as the rows of the residuals matrix, e, such that $e \sim N(0,\Sigma)$. The fixed portion of the model, $X\beta$, is added to the residuals matrix, e, according to the model structure to give the vector of response, Y. The first explanatory variable was considered as indicator variable with two levels and the second explanatory variable was considered as random variable generated from normal distribution with mean equal 30 and variance equal to 5. Each one of the 150 generated data sets was fitted to all the possible combination of the selected model structures and covariance structures for the two set of model structures and covariance structures mentioned before. Then each one of the information criteria was calculated according to the process of the new three stages procedure in order to identify the best multivariate regression model that has the right covariance structure and in the same time the right multivariate model structure.

The 7 settings of the covariance matrix are given in Table 1 which can be categorized to seven covariance structures. The first one, (Setting No. 1) represents Compound Symmetry (CS) covariance structure.

Table 1: The setting of seven covariance matrix structures used in the simulations

Setting no.	Covariance matrix
1.	$\begin{bmatrix} 16 & 12.8 & 12.8 & 12.8 & 12.8 & 12.8 & 12.8 \\ 12.8 & 16 & 12.8 & 12.8 & 12.8 & 12.8 & 12.8 \\ 12.8 & 12.8 & 16 & 12.8 & 12.8 & 12.8 & 12.8 \\ 12.8 & 12.8 & 12.8 & 16 & 12.8 & 12.8 & 12.8 \\ 12.8 & 12.8 & 12.8 & 12.8 & 16 & 12.8 & 12.8 \\ 12.8 & 12.8 & 12.8 & 12.8 & 12.8 & 16 & 12.8 \\ 12.8 & 12.8 & 12.8 & 12.8 & 12.8 & 12.8 & 16 \end{bmatrix}$
2.	$\begin{bmatrix} 16 & 12.8 & 10.24 & 8.192 & 6.5536 & 5.24288 & 4.194304 \\ 12.8 & 16 & 12.8 & 10.24 & 8.192 & 6.5536 & 5.24288 \\ 10.24 & 12.8 & 16 & 12.8 & 10.24 & 8.192 & 6.5536 \\ 8.192 & 10.24 & 12.8 & 16 & 12.8 & 10.24 & 8.192 \\ 6.5536 & 8.192 & 10.24 & 12.8 & 16 & 12.8 & 10.24 \\ 5.24288 & 6.5536 & 8.192 & 10.24 & 12.8 & 16 & 12.8 \\ 4.194304 & 5.24288 & 6.5536 & 8.192 & 10.24 & 12.8 & 16 \end{bmatrix}$
3.	$\begin{bmatrix} 4 & 4.8 & 5.12 & 5.12 & 4.9152 & 4.58752 & 4.194304 \\ 4.8 & 9 & 9.6 & 9.6 & 9.216 & 8.6016 & 7.86432 \\ 5.12 & 9.6 & 16 & 16 & 15.36 & 14.336 & 13.1072 \\ 5.12 & 9.6 & 16 & 25 & 24 & 22.4 & 20.48 \\ 4.9152 & 9.216 & 15.36 & 24 & 36 & 33.6 & 30.72 \\ 4.58752 & 8.6016 & 14.336 & 22.4 & 33.6 & 49 & 44.8 \\ 4.194304 & 7.86432 & 13.1072 & 20.48 & 30.72 & 44.8 & 64 \end{bmatrix}$
4.	$\begin{bmatrix} 4 & 4.8 & 6.4 & 8 & 9.6 & 11.2 & 12.8 \\ 4.8 & 9 & 9.6 & 12 & 14.4 & 16.8 & 19.2 \\ 6.4 & 9.6 & 16 & 16 & 19.2 & 22.4 & 25.6 \\ 8 & 12 & 16 & 25 & 24 & 28 & 32 \\ 9.6 & 14.4 & 19.2 & 24 & 36 & 33.6 & 38.4 \\ 11.2 & 16.8 & 22.4 & 28 & 33.6 & 49 & 44.8 \\ 12.8 & 19.2 & 25.6 & 32 & 38.4 & 44.8 & 64 \end{bmatrix}$
5.	$\begin{bmatrix} 16 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 16 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 16 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 16 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 16 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 16 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 16 \end{bmatrix}$
6.	$\begin{bmatrix} 4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 9 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 16 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 25 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 36 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 49 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 64 \end{bmatrix}$
7.	$\begin{bmatrix} 4 & 2.4 & 4.8 & 8 & 8.4 & 7 & 4.96 \\ 2.4 & 9 & 2.4 & 1.5 & 2.7 & 7.35 & 10.8 \\ 4.8 & 2.4 & 16 & 3.4 & 10.08 & 15.4 & 6.48 \\ 8 & 1.5 & 3.4 & 25 & 18.9 & 16.45 & 9.2 \\ 8.4 & 2.7 & 10.08 & 18.9 & 36 & 4.62 & 22.56 \\ 7 & 7.35 & 15.4 & 16.45 & 4.62 & 49 & 16.24 \\ 4.96 & 10.8 & 6.48 & 9.2 & 22.56 & 16.24 & 64 \end{bmatrix}$

The second one, (Setting No. 2) represents First-Order Autoregressive (AR(1)) covariance structure. The third one, (Setting No. 3) represents Heterogeneous First-Order Autoregressive (ARH(1)) covariance structure. The fourth one, (Setting No. 4) represents Heterogeneous Compound Symmetry (CSH) covariance structure. The fifth one, (Setting No. 5) represents Independent Errors (VC) covariance structure. The sixth one, (Setting No. 6) represents Banded Main Diagonal (UN(1)) covariance structure. The seventh one, (Setting No. 7) represents Unstructured (UN) covariance structure.

RESULTS

The simulation results indicated that the procedure in the first stage selects the right multivariate model structure as member of the best subset hundred percent of the times from the class of candidate multivariate model structures for each of the covariance structures with the new Variance Information Criterion (VARIC). Table 2 summarizes results of the percentage of number of times that the procedure in the first stage selects the right multivariate model structure alone from the class of candidate multivariate model structures for each of the covariance structures with the new Variance Information Criterion (VARIC). Table 2 indicate that the new Variance Information Criterion (VARIC) showed very good performance in identifying the right multivariate model structure in the first stage except for specific models with Unstructured (UN) covariance structure, banded main diagonal (UN(1)) covariance structure and independent errors (VC) covariance structure that have slightly low percentage comparing to others which could be related to conversion difficulty.

The simulation results indicated that the procedure in the second stage selects the right covariance structure as member of the best subset hundred percent of the

Table 2: The percentage of number of times that the procedure in the first stage selects the right multivariate model structure alone from a class of candidate multivariate model structures for each one of the covariance structures with the new information criterion (where nominal Type I error = 0.05)

Covariance structure	Model structure	The Variance Information Criterion (VARIC) (%)
1	1	100.0000
1	2	100.0000
1	3	97.9020
1	4	96.6670
1	5	93.3330
2	1	100.0000
2	2	100.0000
2	3	98.6580
2	4	97.3333
2	5	93.3330
3	1	100.0000
3	2	100.0000
3	3	98.9470
3	4	100.0000
3	5	96.6670
4	1	100.0000
4	2	100.0000
4	3	98.0000
4	4	98.0000
4	5	94.0000
5	1	100.0000
5	2	100.0000
5	3	94.3090
5	4	100.0000
5	5	88.6670
6	1	100.0000
6	2	100.0000
6	3	75.1820
6	4	100.0000
6	5	96.6670
7	1	100.0000
7	2	99.3330
7	3	97.8570
7	4	90.6670
7	5	78.6660

times from the class of candidate covariance structures for each multivariate model structure determined in the first stage as the right model structure with the five criteria. Table 3 present the percentage of number of times that the procedure in the second stage selects the right covariance structure alone from the class of candidate covariance structures for each multivariate model structure determined in the first stage as the right model structure with the five criteria. Table 3 indicate that although all the criteria showed very good performance in identifying the right covariance structure in the second stage, CAIC and BIC criteria have the best performance overall.

Table 3: The Percentage of number of times that the procedure in the second stage selects the right covariance structure alone from a class of candidate covariance structures for each multivariate model structure determined in the first stage as the right model structure for with the five criteria (where nominal Type I error = 0.05)

		The five criteria (%)				
Model structure	Covariance structure	AIC	BIC	CAIC	HQIC	AICC
1	1	99.3333	100	100	100.0000	99.3333
1	2	99.3333	100	100	100.0000	100.0000
1	3	100.0000	100	100	100.0000	100.0000
1	4	99.3333	100	100	100.0000	100.0000
1	5	95.9999	100	100	99.3333	98.0000
1	6	98.6667	100	100	100.0000	100.0000
1	7	100.0000	100	100	100.0000	100.0000
2	1	100.0000	100	100	100.0000	100.0000
2	2	98.6667	100	100	100.0000	100.0000
2	3	100.0000	100	100	100.0000	100.0000
2	4	99.3333	100	100	100.0000	99.3333
2	5	98.0000	100	100	100.0000	99.3333
2	6	99.3333	100	100	100.0000	100.0000
2	7	100.0000	100	100	100.0000	100.0000
3	1	100.0000	100	100	100.0000	100.0000
3	2	100.0000	100	100	100.0000	100.0000
3	3	100.0000	100	100	100.0000	100.0000
3	4	100.0000	100	100	100.0000	100.0000
3	5	98.0000	100	100	100.0000	100.0000
3	6	98.0000	100	100	100.0000	99.3333
3	7	100.0000	100	100	100.0000	100.0000
4	1	100.0000	100	100	100.0000	100.0000
4	2	99.3333	100	100	100.0000	100.0000
4	3	100.0000	100	100	100.0000	100.0000
4	4	100.0000	100	100	100.0000	100.0000
4	5	98.6667	100	100	100.0000	100.0000
4	6	99.3333	100	100	100.0000	99.3333
4	7	100.0000	100	100	100.0000	100.0000
5	1	98.6667	100	100	100.0000	100.0000
5	2	100.0000	100	100	100.0000	100.0000
5	3	100.0000	100	100	100.0000	100.0000
5	4	100.0000	100	100	100.0000	100.0000
5	5	95.3333	100	100	100.0000	97.3333
5	6	98.6667	100	100	100.0000	100.0000
5	7	100.0000	100	100	100.0000	100.0000

DISCUSSION

In our simulation, we considered multivariate regression models in which the responses were correlated in particular ways, looking at the performance of a new three stages procedure that could be used to guide the selection of the best multivariate regression model that has the right covariance structure and in the same time has the right multivariate model structure from both standard and non-standard multivariate model structures. The main result of our article is that the performance of the new three stages procedure in identifying the best multivariate regression model that has the right covariance structure and in the same time the right multivariate model structure was excellent overall in terms of identified the right multivariate model structure in the first stage then identifying the right covariance structure in the second

stage for the multivariate model structure that was determined in the first stage. Finally, the best multivariate regression model that has the right covariance structure and in the same time has the right multivariate model structure is fitted in the third stage using MIXED procedure. Hence, the new procedure is recommended to be used to guide the selection of the best multivariate regression model that has the right covariance structure and in the same time has the right multivariate model structure from both standard and non-standard multivariate model structures, taking into consideration that if the MCB procedure suggested the best subset of model structures (covariance structures) contains more than one model structure (covariance structure), then we recommend selecting the right model structure (covariance structure) as the one with a simplest structure since the careful examination of simulation results showed that in such case the others model structures (covariance structures) in the best subset are just overfitted structures. In case of the first stage of the procedure, the overfitted model structures contain the predictors of the right model, plus any additional predictors. I.e., if the best subset of model structures consists of more than model structure then they could be for example the first multivariate model structure considered as the overfitted model structure and the second multivariate model structure considered as the right model structure. In case of the second stage of the procedure, the overfitted covariance structure contains the same parameters of the right covariance structure, plus any additional parameters. I.e., if the best subset of covariance structures consists of more than one covariance structure then they could be for example Heterogeneous Compound Symmetry covariance structure as the overfitted covariance structure and compound symmetry covariance structure as the right covariance structure. Also, the careful examinations of the whole simulation results reveal that proceeding with the second stage using other than the right selected model structure that was determined in the first stage could be resulted with misleading selection for the covariance structure or sometimes with conversion problem in MIXED procedure in the second stage therefore it is important to follow the right sequences suggested in the three stage procedure to insure selecting the best multivariate regression model that has the right covariance structure and in the same time has the right multivariate model structure.

CONCLUSION

The evaluation of the new three stages procedure indicate that its performance in identifying the right multivariate regression model that has the right covariance structure and in the same time the right multivariate model structure was excellent for both standard and non-standard multivariate model structures. Therefore, the new procedure is recommended to be used as stander tools to guide the selection of the best multivariate regression model that has the right covariance structure and in the same time has the right multivariate model structure.

REFERENCES

Akaike, H., 1974. A new look at the statistical model identification. Trans. Autom. Control, 19: 716-723.

AL-Marshadi, A.H., 2007. The new approach to guide the selection of the covariance structure in mixed model. Res. J. Med. Sci., 2: 88-97. http://www.insipub.com/rjmms/2007/88-97.pdf

AL-Marshadi, A.H., 2008. A simulation study on tests of hypotheses for fixed effects in mixed models for one-way anova under the violation of the equal variances assumption of the treatment groups with and without missing data. JKAU Sci., 20: 57-68. http://www.kau.edu.sa/Files/320/Researches/52400_22707.pdf

AL-Marshadi, A.H., 2009a. Bootstrap simulation procedure applied to the selection of the multiple linear regressions. JKAU Sci., 21: 197-212. http://www.kau.edu.sa/Files/320/Researches/53951_24467.pdf

AL-Marshadi, A.H., 2009b. Comparison of model selection criteria for multivariate regression model with mixed model. J. Applied Sci. Res. (In press).

Azari, R., L. Li and C.L. Tsai, 2006. Longitudinal data model selection. Comput. Stat. Data Anal., 50: 3053-3066. DOI: 10.1016/j.csda.2005.05.009

Beal, D.J., 2005. Information criteria method in sas for multiple linear regression models. SESUG Proceeding (SA05), SESUG Inc., pp: 1-10. http://analytics.ncsu.edu/sesug/2007/SA05.pdf

Bozdogan, H., 1987. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. Psychometrika, 52: 345-370. DOI: 10.1007/BF02294361

Efron, B. and R.J. Tibshirani, 1993. Introduction to the Bootstrap. 1st Edn., Chapman and Hall, New York, ISBN: 0-412-04231-2, pp: 396.

Efron, B., 1983. Estimating the error rate of a prediction rule: Improvement on cross-validation. J. Am. Stat. Assoc., 78: 316-331. http://www.jstor.org/pss/2288636

Efron, B., 1986. How biased is the apparent error rate of a prediction rule?. J. Am. Stat. Assoc., 81: 416-470. http://www.jstor.org/pss/2289236

Fai, A.H.T. and P.L. Cornelius, 1996. Approximate f-tests of multiple degree of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiments. J. Stat. Comput. Simulat., 54: 363-378. DOI: 10.1080/00949659608811740

Ferron, J., R. Dailey and Q. Yi, 2002. Effects of misspecifying the first-level error structure in two-level models of change. Multivariate Behav. Res., 37: 379-403. DOI: 10.1207/S15327906MBR3703_4

Giesbrecht, F.G. and J.C. Burns, 1985. Two stage analysis based on a mixed model: Large sample asymptotic theory and small-sample simulation results. Biometrics, 41: 477-486. http://www.jstor.org/pss/2530872

Goldwasser, M.A. and G.M. Fitzmaurice, 2001. Multivariate linear regression analysis of childhood psychopathology using multiple informant data. Int. J. Methods Psychiat. Res., 10: 1-10. DOI:10.1002/mpr.95

Gomez, E.V., G.B. Schaalje and G.W. Fellingham, 2005. Performance of the Kenward-Roger method when the covariance structure is selected using AIC and BIC. Commun. Stat. Simul. Comput., 34: 377-392. http://cat.inist.fr/?aModele=afficheN&cpsidt=16799994

Guerin, L. and W.W. Stroup, 2000. A simulation study to evaluate PROC MIXED analysis of repeated measures data. Proceedings of the 2000 Conference on Applied Statistics in Agriculture, (ASA'00), Kansas State University, Manhattan, KS., pp: 170-203. http://www.k-state.edu/stats/agstat.conference/2000_table_of_contents.htm

Hannan, E.J. and B.G. Quinn, 1979. The Determination of the order of an autoregression. J. R. Stat. Soc., Ser. B. (Methodological), 41: 190-195. http://www.citeulike.org/user/fabiobayer/article/4647560

Uraibi, H.S., H. Midi, B.A. Talib and J.H. Yousif *et al*., 2009. Linear regression model selection based on robust bootstrapping technique. Am. J. Applied Sci., 6: 1198-1198. http://www.scipub.org/fulltext/ajas/ajas661191-1198.pdf

Hsu, J.C., 1984. Constrained simultaneous confidence intervals for multiple comparisons with the best. Ann. stat. 12: 1136-1144. http://www.jstor.org/pss/2240990

Hurvich, C.M. and C.L. Tsai, 1989. Regression and time series model selection in small samples. Biometrika, 76: 297-307. http://biomet.oxfordjournals.org/cgi/content/abstract/76/2/297

Kacker, R.N. and D.A. Harville, 1984. Approximations for standard errors of estimators of fixed and random effects in mixed linear models. J. Am. Stat. Assoc., 79: 853-862. http://www.jstor.org/pss/2288715

Kenward, M.G. and J.H. Rogers, 1997. Small sample inference for fixed effect from restricted maximum likelihood. Biometrics, 53: 983-97. http://www.jstor.org/pss/2533558

Keselman, H.J., J. Algina, R.K. Kowalchuk and R.D. Wolfinger, 1999. The analysis of repeated measurements: A comparison of mixed-model Satterthwaite F test and a nonpooled adjusted degrees of freedom multivariate test. Commun. Stat. Theor. Methods, 28: 2967-2999. http://home.cc.umanitoba.ca/~kesel/cis1999b.pdf

Littell, R.C., G.R. Milliken, W.W. Stroup, and R.D. Wolfinger, 1996. SAS system for mixed models. 1st Edn. Cary, NC: SAS Institute Inc., ISBN: 1-55544-779-1, pp: 633.

Littell, R.C., J. Pendergast and R. Natarajan, 2000. Modeling covariance structure in the analysis of repeated measures data. Stat. Med., 19: 1793-1819. DOI: 10.1002/1097-0258(20000715)19:13<1793::AID-SIM482>3.0.CO;2-Q

McCullagh, P., 2006. Structured covariance matrices in multivariate regression models. Technical reports online. http://www.stat.uchicago.edu/~pmcc/reports/similarity.pdf

McCulloch, C. and S. Searle, 2001 Generalized, Linear and Mixed Models. 1st Edn. Wiley, New York, ISBN: 0-471-19364-X, pp: 325.

SAS Institute Inc., SAS OnlineDoc 9.1.3. SAS Institute Inc., Cary, NC. http://support.sas.com/onlinedoc/913/docMainpage.jsp

Satterthwaite, F.E., 1941. Synthesis of variance. Psychometrika, 6: 309-316. DOI: 10.1007/BF02288586

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat., 6: 461-464. http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.aos/1176344136

Seghouane, A.K., 2006. Multivariate regression model selection from small samples using kullback'symmetric divergence. Signal Process., 86: 2074-2084. DOI: 10.1016/j.sigpro.2005.10.009

Singer, J.D., 1998. Using SAS PROC MIXED to fit multilevel models, hierarchical models and individual growth models. J. Educ. Behav. Stat., 24: 323-325. http://gseweb.harvard.edu/~faculty/singer/Papers/sasprocmixed.pdf

Yamamura, M., H. Yanagihara and M.S. Srivastava, 2008. Variable selection in multivariate linear regression models with fewer observations than dimension. http://www.math.sci.hiroshima-u.ac.jp/stat/TR/TR08/TR08-13.PDF