Limitations of Ordinary Least Squares Models in Analyzing Repeated Measures Data

CARLOS UGRINOWITSCH[1,3], GILBERT W. FELLINGHAM[2], and MARK D. RICARD[4]

[1]Human Performance Research Center and [2]Department of Statistics, Brigham Young University, Provo, UT; [3]Department of Sport, School of Physical Education and Sport, University of Sao Paulo, Sao Paulo, BRAZIL; and [4]Biomechanics Laboratory, Western Michigan University, Kalamazoo, MI

ABSTRACT

UGRINOWITSCH, C., G. W. FELLINGHAM, and M. D. RICARD. Limitations of Ordinary Least Squares Models in Analyzing Repeated Measures Data. *Med. Sci. Sports. Exerc.*, Vol. 36, No. 12, pp. 2144–2148, 2004. **Purpose:** To a) introduce and present the advantages of linear mixed models using generalized least squares (GLS) when analyzing repeated measures data; and b) show how model misspecification and an inappropriate analysis using repeated measures ANOVA with ordinary least squares (OLS) methodology can negatively impact the probability of occurrence of Type I error. **Methods:**The effects of three strength-training groups were simulated. Strength gains had two slope conditions: null (no gain), and moderate (moderate gain). Ten subjects were hypothetically measured at five time points, and the correlation between measurements within a subject was modeled as compound symmetric (CS), autoregressive lag 1 (AR(1)), and random coefficients (RC). A thousand data sets were generated for each correlation structure. Then, each was analyzed four times—once using OLS, and three times using GLS, assuming the following variance/covariance structures: CS, AR(1), and RC. **Results:** OLS produced substantially inflated probabilities of Type I errors when the variance/covariance structure of the data set was not CS. The RC model was less affected by the actual variance/covariance structure of the data set, and gave good estimates across all conditions. **Conclusions:** Using OLS to analyze repeated measures data is inappropriate when the covariance structure is not known to be CS. Random coefficients growth curve models may be useful when the variance/covariance structure of the data set is unknown. **Key Words:** MIXED MODELS, COVARIANCE STRUCTURE, RANDOM COEFFICIENTS, POWER, TYPE I ERROR

When an experiment is conducted and the variable of interest is measured from the same subject repeatedly over time, the measurements are almost always correlated in some way. To account for this correlation, repeated measures statistical analysis tools may be used to make inferences (13). The results obtained from these statistical analyses are very important to explain long-term adaptations to physiological changes caused by training, aging, growth, or development. Therefore, the selection of the most appropriate statistical method to analyze data sets with multiple measurements on a single experimental unit (subject) is critical for correct interpretation of results.

Address for correspondence: Mark D. Ricard, Western Michigan University, HPER department, 1903 West Michigan Avenue, Kalamazoo, MI 49008; Email: Mark.Ricard@wmich.edu.
Submitted for publication February 2004.
Accepted for publication July 2004.

0195-9131/04/3612-2144
MEDICINE & SCIENCE IN SPORTS & EXERCISE®

DOI: 10.1249/01.MSS.0000147580.40591.75

Ordinary least squares (OLS) computational methods are commonly used to test hypotheses of differences among factor-level means in repeated measures data, and are available in a variety of commercial statistical software packages, generally under the rubric of general linear model (GLM). However, the OLS methodology implemented in these programs has important assumptions and limitations that can directly affect both the computation of F-tests, and the estimation of means and standard errors (SE) on repeated measures data (10).

One of the most important assumptions for appropriate F-tests when using OLS in repeated measures data is that there is a constant correlation among multiple measurements within a subject (5). This assumption would not be true if measurements taken closer in time were more highly correlated than those taken farther apart in time. This type of correlation structure is likely in situations involving human performance. Thus, the assumption of constant correlation for measurements within a subject may not be true in many cases. Some statistical software packages use the Greenhouse-Geisser adjustment of the F-test when the assumption of constant correlation within subjects' measurements may not be appropriate. This adjustment produces more conservative F-values, but does not improve the estimates of means and SE (5). Some authors have suggested the use of repeated measures MANOVA instead of univariate repeated measures ANOVA when the assumption of constant covariance is not met. This is a reasonable approach if there are no missing data. MANOVA requires that the data for all individuals be available for exactly the same time points to estimate the covariance matrix. Since these conditions are often not attained in a typical experiment, MANOVA is a less attractive option.

Another limitation of OLS in the univariate setting is also related to missing values. Algorithms for the computation of variance components using OLS are not optimal when data are missing, even if the assumptions about the covariance structure are correct (12).

In the late 1970s a new methodology called linear mixed models, which uses an estimation algorithm called generalized least squares (GLS), and is designed to deal with correlated data, was developed (4,6–8). These ideas have been implemented into a number of commercially available software packages, one of which is SAS® Proc Mixed. With the development of this methodology, it is now possible to model a rich variety of covariance structures, deal effectively with missing data, (1,2,10) and handle nonconstant measurement time points (3).

Basically, GLS estimates are appropriately weighted based on the covariance structure of the data. The estimates of the variance components are acquired by maximizing either the full likelihood (maximum likelihood, or ML) or a restricted version of the likelihood (restricted or residual maximum likelihood, or REML). Thus, linear mixed models based on GLS are designed to deal with repeated measures data, as well as other types of correlated data.

Linear mixed models also have the ability to fit a variety of growth curve models, and thus compare the rate of growth, or slopes, among curves (11,12). Instead of comparing means across treatments and time points, and dealing with complex interpretation of the model interaction terms, growth curves test for differences in initial values and in rates of growth of each factor's levels, allowing for a more straightforward interpretation of data (14). A particular implementation of the growth curve model is called the random coefficients growth curve, where each subject's parameters are considered a random draw from a population of such parameters. This implementation is also known as multilevel models, or hierarchical linear models.

Thus, the purposes of this study were to introduce and present the advantages of linear mixed models and GLS in the analysis of repeated measures data, to show how model misspecification and inappropriate analysis using ordinary least squares (OLS) can directly impact the probability of occurrence of Type I errors depending upon the variance/covariance structure of the original data set, and to compare the performance of these models in analyzing incomplete (unbalanced) data sets.

METHODS

Data generation. The data sets used in the study were simulated to mirror the training effects of three groups of 10 subjects that participated in a strength-training program. Data generation and analysis were performed using SAS® Proc IML to generate the data and SAS® Proc Mixed to analyze the data. A Visual Basic wrapper controlled program flow. The simulated experiment included three treatment groups called high-resistance low-repetition (HR), low-resistance high-repetition (LR), and control (C). Data were simulated as if subjects were measured at five equally spaced time points, labeled 0, 1, 2, 3, and 4. The response variable represented the 1RM results obtained in a biceps curl exercise. All groups had the same initial average 1RM values (75 lb). Then, strength gains could follow two possible patterns: null, or moderate. The null condition had slopes (strength gains) equal to zero for all three treatment groups. The moderate condition had slopes of 1.5, 1, and 0 $lb \cdot wk^{-1}$ for the HR, LR, and C groups, respectively. Observations were obtained using the fixed effects described in the preceding paragraph, and then adding random variation using three different variance/covariance structures.

The first structure we chose to use was compound symmetric (CS). This structure has constant variances across time and constant covariances between measurements. The particular values we used for this structure, shown in Figure 1, seemed to us to be representative of values that would be expected from an experiment of the type we describe. Data with this structure can be appropriately analyzed using OLS when the data set is complete and balanced. Thus, we should see no differences between an OLS and a GLS analysis of a balanced data set of this type. The CS structure is also appealing because it is relatively simple, requiring the estimation of only two variance components.

The second covariance structure we used is called autoregressive lag 1 (AR(1)) and exhibits a pattern of higher correlation between data points taken closer in time. The actual values we used for this structure are shown in Figure 2. We would not expect an OLS and GLS analysis of this type of data to yield similar results. This is an appealing structure to consider, as measurements closer in time do tend to be more highly correlated in human performance

$$\begin{vmatrix} 102.5 & 83 & 83 & 83 & 83 \\ 83 & 102.5 & 83 & 83 & 83 \\ 83 & 83 & 102.5 & 83 & 83 \\ 83 & 83 & 83 & 102.5 & 83 \\ 83 & 83 & 83 & 83 & 102.5 \end{vmatrix}$$

FIGURE 1—Compound symmetric variance/covariance structure used to generate simulation data. Main diagonal shows the variances of subject measurements at times 0 to 4. Off-diagonal shows the covariances of subject measurements. These covariances are by position, so the 1,2 element of the matrix is the covariance between the measurement at time 0 and the measurement at time 1.

$$\begin{vmatrix} 83.00 & 66.40 & 53.12 & 42.50 & 34.00 \\ 66.40 & 83.00 & 66.40 & 53.12 & 42.50 \\ 53.12 & 66.40 & 83.00 & 66.40 & 53.12 \\ 42.50 & 53.12 & 66.40 & 83.00 & 66.40 \\ 34.00 & 42.50 & 53.12 & 66.40 & 83.00 \end{vmatrix}$$

FIGURE 2—Auto-regressive lag (1) variance/covariance structure used to generated simulation data. Rows and columns as in Figure 1.

trials. In addition, AR(1) requires the estimation of only two variance components.

Finally, we generated data using a random coefficients (RC) growth curve model. In this model, an individual's intercept and slope are considered to be random draws from a distribution of intercepts and slopes. We used the slope/intercept covariance structure of variance of intercepts = 100, variance of slopes = 7, and covariance between intercepts and slopes = −16 along with an independent random error of $\sigma^2 = 13$. This yielded the within-subject covariance structure shown in Figure 3. Although slightly more complicated than the other two structures (requiring the estimation of four rather than two variance components), it is appealing because it is so broadly applicable. We also had some indication from previous work that this structure might be less sensitive to misspecification, and thus more robust in our analyses.

The determinants of the variance/covariance matrices presented in Figures 1, 2, and 3 were 0.6282×10^8, 0.6616×10^8, and 0.6028×10^8, respectively, so these structures account for roughly the same variability across models.

The GLS algorithm as implemented in SAS® allows the practitioner to model a rich variety of variance/covariance structures. The interested reader should refer to the SAS® help files to find available structures. We selected these three variance/covariance structures for the following reasons: a) the compound symmetric structure is assumed in univariate repeated measures routines available in commercial software using OLS; b) the AR(1) structure accurately reflects strength increments due to strength training (and other expected changes across time in human performance) (9); and c) random coefficients is a general model that could have wide applicability in human performance literature.

$$\begin{vmatrix} 113 & 84 & 68 & 52 & 36 \\ 84 & 88 & 66 & 57 & 48 \\ 68 & 66 & 77 & 62 & 60 \\ 52 & 57 & 62 & 80 & 72 \\ 36 & 48 & 60 & 72 & 97 \end{vmatrix}$$

FIGURE 3—Random coefficients variance/covariance structure used to generate simulation data. Rows and columns as in Figure 1.

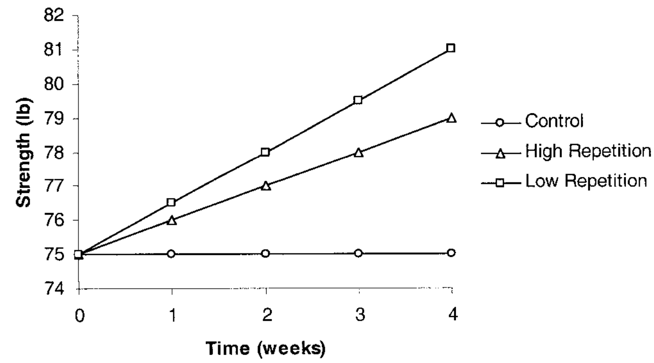

FIGURE 4—Hypothetical rate of strength increments in training groups in the moderate slope difference condition.

Independent random errors can be manipulated to assume any covariance structure by premultiplying the independent variates by some square-root decomposition of the covariance structure. We used a Cholesky decomposition, and used independent random variates generated by SAS® Proc IML.

Data Analysis. The data were generated with two possible differences of slopes (null and moderate), and three possible variance/covariance structures (CS, AR(1), and RC) for six different types of data sets.

There were 1000 data sets generated in each one of the conditions. These data sets were analyzed using OLS, using GLS, assuming a CS covariance structure, using GLS assuming an AR(1) covariance structure, and using GLS assuming a random coefficients model. Figure 4 gives a pictorial view of the hypothetical data situation in the moderate slope difference condition. All data sets were analyzed using a model that included individual fixed effects for treatment (trt), time, and the treatment by time (trt*time) interaction. Since a significant trt*time interaction would be indicative of a significant difference among the slopes of the lines of the different treatment conditions, we focused our analysis on that term.

The same procedure was applied to incomplete data sets. Twenty percent of the response values of each data set were deleted at random before these analyses were undertaken.

The proportion the 1000 runs that resulted in a significant trt*time interaction was computed. In the null-slope condition, this number is an estimate of the probability of occurrence of a Type I error for the various combinations of conditions. In the moderate slope difference condition, this

TABLE 1. Percentage of 1000 runs that rejected a null hypothesis of no difference in treatment slopes at the $\alpha = 0.05$ level.

	OLS (%)	CS-GLS (%)	AR(1)-GLS (%)	RC-GLS (%)
CS-Complete	6.0	6.0	0.0	4.8
CS-20% deleted	6.8	6.6	0.1	4.8
AR(1)-complete	30	30	4.6	5.2
AR(1)-20% deleted	24	23	8.2	7.0
RC-complete	36	36	10	6.1
RC-20% deleted	30	31	14	6.7

Data were generated using the null-slope difference condition. Columns represent the variance/covariance structure used to generate the data, and rows represent the analysis method.

* For 1000 simulations, responses in the 3–7% range are not unreasonable.

TABLE 2. Percentage of 1000 runs that rejected a null hypothesis of no difference in treatment slopes at the $\alpha = 0.05$ level.

	OLS (%)	CS-GLS (%)	AR(1)-GLS (%)	RC-GLS (%)
CS-complete	57*	57*	3.2	52*
CS-20% deleted	46*	47*	7.3	41*
AR(1)-complete	53	53	24*	24*
AR(1)-20% deleted	46	46	27	22*
RC-complete	57	57	27	18*
RC-20% deleted	50	50	29	16*

Data were generated using the moderate slope difference condition. Columns represent the variance/covariance structure used to generate the data, and rows represent the analysis method.

* Indicates a cell where null case (as shown in Table 1) is at the appropriate level.

proportion is an estimate of the power of the analyses in the various conditions.

RESULTS

The results of the simulation for the null-slope condition are found in Table 1. With 1000 simulations, the standard error of the estimated percentage rejecting in the null case is approximately 0.7%. Thus, outcomes from 3 to 7% are not unreasonable. However, outcomes greater than 15% in the null case are definitely a cause for concern. When the data are balanced, the OLS and CS-GLS analyses will return identical results. OLS performed as it should when the generated data were CS. However, when generated data were either AR(1) or RC, OLS was substantially anticonservative. AR(1)-GLS only included the nominal 5% level when the data were complete and generated using an AR(1) structure. The RC-GLS method displayed reasonable results for all underlying covariance structures and for balanced or unbalanced data. In general, OLS (and CS-GLS) tended to become more conservative with missing data, while AR(1)-GLS became more anticonservative when data were missing.

The results of the simulations for data generated under the moderate slope condition are found in Table 2. These data are indicative of the power of the various methods, but it is unreasonable to examine power in any situation where the nominal 5% level under the null condition is not maintained. When the generated data have a CS structure, there is not a significant difference in power among OLS, CS-GLS, and RC-GLS methods. ($P > 0.45$ in every case). There is a significant decrease in power from the complete data case to the unbalanced case ($P < 0.05$ in every case). When the generated data are AR(1), there is no significant difference in power between the AR(1)-GLS and RC-GLS methods ($P > 0.73$). When data are missing, the AR(1)-GLS method is significantly more powerful than the RC-GLS method. When generated data are RC, only the RC-GLS method is a reasonable candidate for analysis.

DISCUSSION

The main findings of this study were: a) knowing the variance/covariance structure of the data set is very important in selecting the appropriate variance/covariance model for data analysis; b) model misspecification can substantially affect the significance level of a statistical test; c) assuming a CS structure when that is not the case produces the highest probability of occurrence of Type I errors whether the analysis is OLS or GLS; d) linear random coefficients growth curve models have good performance characteristics with any of the variance/covariance structures we tested; and e) these findings generally apply to both complete and incomplete data sets.

The repeated measures OLS algorithm produces accurate F-tests when data are balanced, and quite reasonable results when data are unbalanced and the variance/covariance structure of the data is CS. Our results indicate that when the variance/covariance structure of the data is not CS, there is a substantial inflation of Type I error rates.

It is disheartening to note that the repeated measures OLS analysis is the prevalent methodology used in the exercise science field. Our findings indicate that there is reason to believe that some nontrivial fraction of the literature may indicate significant treatment effects when in fact they do not exist, because of an inappropriate application of statistical methodology. Our type I error rates were in the 20–30% range when the OLS or CS-GLS methodology was used, and the data were really AR(1) or RC (Table 1). Our rates clearly overstate this problem because many researchers are using the Greenhouse–Geisser correction in an attempt to deal with the difficulty. Nonetheless, the appropriate use of GLS would undoubtedly help alleviate the problem.

Random coefficients growth curve models have been largely used in the literature outside the exercise science field (11,14). Our data indicate that this broad classification of models might offer an appropriate alternative when there is doubt concerning the true underlying variance/covariance structure of the data set. This model type had consistently good α-level performance (Table 1) and reasonably good power performance over the variety of variance/covariance structures considered, even with unbalanced data sets (Table 2).

When analyzing repeated measures data, researchers should make every effort to determine what type of underlying variance/covariance structure the data may have. Failing to spend the time to determine what type of variance/covariance structure is present could easily lead to an inappropriate data analysis. Readers are encouraged to refer to Littell et al. (10) for a review of a possible methodology to select the appropriate variance/covariance structure for data analysis. If doubt persists concerning the appropriate variance/covariance structure, then our study indicates that a random coefficients model could be used, thanks to its robustness across the structures we examined.

In conclusion, the use of OLS algorithms to analyze repeated measures data should be limited to cases where a CS covariance structure is known to be present. The linear mixed model, with its greater flexibility to model the variance/covariance structure of the data set, should be more generally used to analyze growth-type data. Random coefficients growth curve models should be considered for analysis when the variance/covariance structure of the data set is uncertain.

This study was partially funded by Coordenadoria de Aperfeiçoamento de Pessoal de Ensino Superior, Brazil.

REFERENCES

1. Catellier, D. J., and K. E. Muller. Test for Gaussian repeated measures with missing data in small samples. *Stat. Med.* 19:1101–1114, 2000.
2. Cnaan, A. V., N. M. Laird, and P. Slasor. Using the general linear mixed model to analyze unbalanced repeated measures and longitudinal data. *Stat. Med.* 16:2349–2380, 1997.
3. Daniels, M. J., and Y. D. Zhao. Modelling the random effects covariance matrix in longitudinal data. *Stat. Med.* 22:1631–1647, 2003.
4. Harville, D. A. Maximum likelihood approaches to variance component estimation and to related problems. *J. Am. Stat. Assoc.* 72:320–338, 1977.
5. Kuehl, R. O. *Design of Experiments: Statistical Principles of Research Design and Analysis.* 2nd Ed. Pacific Grove: Duxbury, 2000, 666.
6. Laird, N. M., N. Lange, and D. Stram. Maximum likelihood computations with repeated measures: application of the EM algorithm. *J. Am. Stat. Assoc.* 82:97–105, 1987.
7. Laird, N. M., and J. H. Ware. Random-effects models for longitudinal data. *Biometrics.* 38:963–974, 1982.
8. Lindstrom, M. J., and D. M. Bates. Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *J. Am. Stat. Assoc.* 83:1014–1022, 1988.
9. Littell, R. C., G. A. Miliken, W. W. Stroup, and R. D. Wolfinger. *SAS System for Mixed Models.* Cary: SAS Institute, 1996, 633.
10. Littell, R. C., J. Pendergast, and R. Natarajan. Modelling covariance structure in the analysis of repeated measures data. *Stat. Med.* 19:1793–1819, 2000.
11. Ojeda, M. M., and N. Longford. A two-level regression model for growth curves with correlated data. *Biometrical J.* 40:447–454, 1998.
12. Palta, M., and T. Cook. Some considerations in the analysis of rates of change in longitudinal studies. *Stat. Med.* 6:599–611, 1987.
13. Park, T., J.-K. Park, and C. S. Davis. Effects of covariance model assumptions on hypothesis tests for repeated measurements: analysis of ovarian hormone data and pituitary-pteryomaxillary distance data. *Stat. Med.* 20:2441–2453, 2001.
14. Zee, B. C. Growth curve model analysis for quality of life data. *Stat. Med.* 17:757–766, 1998.