A Macro for Conditional Ordinary Least Squares and Weighted Least Squares Analysis of Repeated Measures Data.

by Margaret Burchinal, University of North Carolina at Chapel Hill

Abstract
WLS_RM is a PROC MATRIX macro that performs various types of analyses of repeated measures data that cannot be easily done with PROC GLM, even with the REPEATED statement. These methods for analyzing repeated measures data often will be more powerful than the tests performed by PROC GLM when the model assumptions are met. The four types of analyses conducted by WLS_RM involve fitting a truncated polynomial growth curve (PGC) model through ordinary least squares (OLS) estimation, weighted least squares (WLS) estimation, and conditional OLS estimation. The order of the PGC fit to the data is either specified by the user or, under conditional OLS estimation, is determined through series of top-down analyses of the data. WLS_RM provides point estimates of the parameters, estimated variance-covariance matrices for the transformed response variables and for the estimated parameters, and the results of the tests of the main effects and interactions of the within-subject and between-subject factors.

Introduction
While repeated observations of the same experimental units are commonly incorporated into designed studies, selection of the most appropriate method for the analyses of these data has been controversial. A variety of methods have been advanced to analyze repeated measures data. These approaches can be classified along at least two dimensions. Firstly, they involve either univariate or multivariate tests of the hypotheses regarding the within-subjects factor. Secondly, parameters are either estimated through ordinary least squares (OLS) or weighted least squares (WLS). The multivariate approaches often are regarded as superior statistically to the univariate approaches when the number of experimental units (n) with complete data on all p occasions is sufficiently large. Under these circumstances, the multivariate methods range from being nearly as powerful as the univariate methods to being more powerful when methods that maintain the nominal type I error rate are compared (Davidson, 1972).

Three multivariate approaches have been advocated: the OLS, the WLS, and the conditional OLS. The circumstances under which each method provides the most powerful tests of hypotheses concerning the within-subjects factor while being delineated by Grizzle and Allen (1969). The OLS models are preferred when the sphericity assumption is met or when p is small or n is very large so that fitting the (p-1)-th order polynomial growth curve (PGC) model does not result in multivariate tests lacking in power. If none of these conditions is met, then the WLS or conditional OLS models generally provide more powerful multivariate tests while making more stringent assumptions of the data than do the OLS models. Multivariate analyses involving unweighted least squares estimates can be performed easily in Version 5 using PROC GLM with the REPEATED statement if a rank p design matrix for the within-factor is desired. The truncated OLS, the multivariate WLS, and the conditional OLS analyses can be done with a new macro, WLS_RM.

The multivariate models for repeated measures data are special cases of the PGC model. Potthoff and Roy (1964) developed the PGC model as a generalization of the general linear model,

$$\mathbf{I} = \mathbf{A} \mathbf{B} + \mathbf{E}$$  \hspace{1cm} (1)

to describe the polynomial growth curves

$$\mathbf{I} = \mathbf{A} \mathbf{B} \mathbf{A} + \mathbf{E}$$  \hspace{1cm} (2)

In this model, \( \mathbf{I} \) is an \((n \times p)\) matrix of p observations on each of n experimental units on the same response variable. \( \mathbf{A} \) is an \((m \times r)\) between-subjects design matrix, \( \mathbf{B} \) is a \((r \times q)\) matrix of unknown parameters (where the rows of \( \mathbf{B} \) correspond to the effects of the between-subject factors and the columns of \( \mathbf{B} \) to the within-subjects effects), \( \mathbf{A} \) is a \((q \times p)\) within-subjects design matrix where \( q < p \) (often, for computational ease, \( \mathbf{A} \) consists of the first q vectors of orthogonal polynomial coefficients that correspond to the times of assessment of the dependent measure), and \( \mathbf{E} \) is an \((n \times p)\) random error matrix such that the \( i \)-th row contains the p error terms for the \( i \)-th individual and for all individuals.

Potthoff and Roy showed that (2) could be reduced to (1) through transforming \( \mathbf{I} \).

$$\mathbf{I} = \mathbf{G} \mathbf{A} ( A \mathbf{G} A^T)$$  \hspace{1cm} (3)
for some arbitrary nonstochastic, $p \times p$ symmetric, positive definite matrix $G$. The resulting estimator of $B$, $b$, would be

$$b = (X'X)^{-1} X'y$$

and the sample was sufficiently large to substitute the consistent estimator, $\hat{G}$, for $G$ in the WLS estimate of $B$.

Rao (1966) suggested an alternative method for estimating $B$. An analysis of covariance model (ANCOVA) was suggested that would permit estimation of $B$ through OLS by including in the model the $(p-q)$ covariates that are a basis of the set of linear functions of the columns of $Y$ given by

$$Y \sim N_p(I, \Sigma), \quad i=1,2,\ldots,r,$$

and that the sample was sufficiently large to substitute the consistent estimator, $\hat{G}$, for $G$ in the WLS estimate of $B$.

Lewis and van Knipperberg preferred the ANCOVA model to the unweighted and weighted models. The unweighted approach was rejected as "conceptually unsatisfying" since no assumptions were made about higher order terms in testing each polynomial coefficient. They stated that the ANCOVA method provides more powerful tests than the unweighted method when the means of the transformed variables that are used as covariates are actually close to zero in the population and the reduction in the standard error of the adjusted estimate is sufficient to compensate for the loss of the degrees of freedom" (p. 190-191). The ANCOVA method was preferred to the WLS method because the substitution of the sample covariance matrix for the population covariance matrix depends on asymptotic arguments. For finite samples, the degrees of freedom will be too large and, thus, the estimated covariances too small with this approach. Lewis and van Knipperberg concluded by suggesting that the most appropriate multivariate approach to analyzing repeated measures data would involve two steps. First, the order of the within-subjects factor would be determined through a step-down analysis of the transformed variables or through specification of $q$ based on theoretical or previous empirical work. Then, the parameters, predicted values, standard errors would be estimated under the selected multivariate ANCOVA model.

However, recent work has strongly suggested that selection of the number of covariates on the basis of the data results in grossly underestimated standard errors (Kenward, 1985). Kenward found in a simulation study that when the number of covariates is large and the residual degrees of freedom are small the estimated standard errors were severely biased when selection of covariates was based on the data. The OLS and the ANCOVA models result in unbiased estimates of standard errors when the selected covariates for the ANCOVA models was not based on the data. Thus, the ANCOVA model could provide more precise parameter estimates without underestimating the standard errors when all covariates are included in the model, the order of the model is selected a priori, and when the conditions described by Lewis and van Knipperberg were met.

**The WLS_RE Macro**

This PROC MATRIX macro, WLS_RE, allows you to perform this kind of analysis as well as the step-down analysis described by Lewis and van Knipperberg.
Knipperberg and the WLS analysis. You must supply the \((n \times p)\) matrix of response variables and the \((n \times r)\) between-subjects design matrix. The general linear hypotheses specified by the contrast matrices are tested through multivariate tests of the interactional effects of the within*between factors and the main effects of the within factor and through univariate tests of the effects of the between-subjects factors. The macro computes three multivariate tests statistics, the Hotelling-Lawley Trace, Pillai's Trace, and Wilks· Lamda, and their associated approximate \(F\) statistics when multivariate tests are performed. The estimated covariance matrix of the transformed dependent measures, the matrix of parameter estimates, the estimated covariance matrix of the estimated parameters, and statistical test results are reported in the output. You must either specify matrices that contain the dependent and independent measures as \(\mathbf{Y}\) and \(\mathbf{X}\) respectively in the invocation of the macro, or you must have specified matrices called \(\mathbf{Y}\) and \(\mathbf{X}\) respectively prior to invoking the macro. The DUPLICATE option must be in effect when the macro is invoked. The default contrast matrices, \(\mathbf{L}\) and \(\mathbf{M}\), assume that a cell-mean between-subjects design matrix was provided and that the between-subjects design was orthogonal. The default contrast matrices only will provide the \((r-1)\) rank test of the between factor averaged across time, the \((q-1)\) rank test of the within factor averaged across group, and the \((r-1) (q-1)\) rank test of the within*within interaction.

Other \(\mathbf{L}\) and \(\mathbf{M}\) matrices can be provided. The macro will accept a maximum of nine contrasts among the between-subjects factors. \(\mathbf{L} \cdot L_5 \cdot \mathbf{M}_9\). You must specify the number of \(\mathbf{L}\) contrast matrices that are being used to test hypotheses about the between factor(s) if the number is greater than two. For example, if \(L_3 \cdot L_4 \cdot \mathbf{M}_1\) and \(\mathbf{M}_4\) would have to be specified when WLS_RM was called. Only one within-subjects factor is permitted. The contrast matrices called \(\mathbf{L}_1\) and \(\mathbf{M}_1\) are required to be averages used to test the main effects of time and of the between-subjects factors, respectively.

The model that is fit to the data by default is the ANCOVA model in which the \((q-1)\) higher-order within contrasts are used as covariates. You must specify the desired order of the PSC model fit by WLS_RM by stating \(\mathbf{Q}=q\) the desired order of the within design matrix. The three other models can be requested are a \((q-1)\)-th degree WLS model in which the user specifies \(\mathbf{Q} = 3\) and \(\mathbf{Q} = q\), the truncated OLS model in which the user requests \(\mathbf{Q} = 4\) and \(\mathbf{Q} = q\), and the ANCOVA model in which the complexity of the model is determined through the analysis described by Lewis and van Knipperberg. Then, you would request \(\mathbf{Q} = 2\) and would be provided with a listing of the step-down tests performed as well as the results of the omnibus multivariate and univariate tests. The complexity of the model is determined by identifying the highest order nonsignificant within*between interaction as well as determining the highest-order significant term in the tests of the main effects of the within-subjects factors.

Lastly, the times of observation must be provided if they were not equally-spaced as \(\mathbf{TIME}=n\) a row vector containing the times. You may provide labels for variables and tests if desired. Labels should correspond to the \(\mathbf{L}\) matrices that specify the contrasts of interest on the between-subjects factors. For example, including \(\mathbf{LABEL2} = \mathbf{SEX}\) would result in a label of \(\mathbf{SEX}\) in the header whenever the general linear hypothesis, \(\mathbf{LM}\), being tested involved \(\mathbf{L}_2\).

In summary, this PEOC MATRIX macro allows you to perform a more powerful analysis of repeated measures data by employing the ANCOVA model rather than with the OLS model when the \(Y_1\) and \(Y_2\) variables are correlated and the other model assumptions are not. As such, it is a complement to the analysis possible with Version 5 PEOC GLM.

References


The Code

```plaintext
OPTIONS NODATE;
%macro WLS; H=M,
   A = %GPOL (TIME);
   A1 = A1, (1);, *DETERMINING TYPE OF ANALYSIS REQUESTED BY USER:
   IF TYPE=1 THEN DO;
   ANCFLAG = 0;
   YSTAR = Y * A1;
   LINK OLS;
   ELSE IF TYPE=2 THEN DO;
   ANCFLAG = 1;
   YSTAR = Y - A;
   LINK OLS;
   END;
   ELSE IF TYPE=3 THEN DO;
   ANCFLAG = 0;
   YSTAR = Y;
   LINK WLS;
   END;
   ELSE IF TYPE=4 THEN DO;
   ANCFLAG = 0;
   YSTAR = Y;
   LINK OLS;
   END;
   *****TEST WHETHER COVARIATES DIFFER SIGNIFICANTLY FROM ZERO;
   IF TYPE=1 THEN DO;
   M = 1; (Q); L = J, (P-Q);, *NOTE:
   M = K2; L = L1; PRINT M L;
   DFH = NROW (L);
   *****TEST OVERALL GROUP*TIME INTERACTION;
   IF R > 1 THEN DO;
   DO DFM = 1 TO NURL;
   *****TEST OVERALL TIME EFFECT;
   IF Q > 1 THEN DO;
   NOTE; END;
   *****TEST OVERALL EFFECT - MULTIVARIATE TESTS FOR;
   B = Z; F = Null; PRINT B F;
   END;
```


IF TYPE=1 THEN L = L || J. (DFM-PQ-0); LINK MANOVA; END;

**TEST BETWEEN-SUBJECTS EFFECTS:**

IF R > 1 THEN DO:

**NOTE **

WHOLE TEST OF EFFECT - UNIVARIATE TEST OF:
NOTE LABEL ANOVA EFFECT -

~M=N; L=DPMT; PRINT ML; DFM = NWG. (L); IF TYPE=1 THEN L = L || J. (DFM-PQ-0); LINK MANOVA; END;

STOP

**-----------------------------------------------**

MANOVA:

DFM = NWG. (L) & NWST((E') & LST((E') & INV(L) & XPL(I) & XST((E') & L = B = A; E = NWST((E' & (SIGMA(DFM)) = 1;

**COMPUTE MULTIVARIATE TEST STATISTICS:**

IF NCOL(H) > 1 THEN DO;

PB = NCOL(H);

IF DFM<PE THEN SE=DFM; ELSE SE=PE;

ME = -5 # (DFE-FP-DP) - 1; DFMT = SE # (2#PE + SE + I);

DFMT = 2 # (SE # ME + 1);

HTRACE = TRACE(INV(B) & H);

F (DFEHT / HTRACE) / (SE / DFMT);

PROB = 1 - PROBF (DFMT, DFEHT);

NOTE **

MODELLING-LAWLEY TRACE********

NAME = "DP-ERROR" "DF-ERROR" "PROB" "F";

LABELS = HTRACE | DFM | DEBUT | F | PROB;

PRINT LABELS COLNAME = NAME ROWNAME=NUll;

DFPMT = SE # (2#ME + SE + 1);

PILTRACE = TRACE(HAT*(INV(PH))); F (DFPMT / DFPMT) / (SE / PILTRACE);

PROB = 1 - PROBF (DFPMT, DFPMT);

NOTE **

WILKES-BARLETT TRACEn********

NAME = "DP-ERROR" "DF-ERROR" "PROB" "F";

LABELS = PILTRACE | DFPMT | DEBUT | F | PROB;

PRINT LABELS COLNAME = NAME ROWNAME=NUll;

DE = PE + DFM - PE;

IF SE = 0 THEN T = B*PE*(PE + DFM - 4) # / DE;

ELSE T = 1;

DFML = PE + DFR;

B1 = DFR - (PE + DFM +1) # / 2;

B2 = (PE + DFM -2) # / 4;

IF DE > 2 THEN DFM = B*PE - 2#B;

ELSE IF SE = 1 THEN DFM = DFR + DFM - PE;

ELSE IF SE = 2 THEN DFM = 2*(DFE+DFM-PE-1);

LAM = LAMBDA ** ((0-7) T) /

F (1-LAM) / (DFML / DFM); PROB = 1 - PROBF (DFML, DFM);;

NOTE **

WALLS LAMDA********

NAME = "SS-ERROR" "SS-ERROR" "PROB" "F";

LABELS = LAMDA | DFML | DFM | F | PROB;

PRINT LABELS COLNAME = NAME ROWNAME=NUll;

*OLS TEST OF GROUP PGC PARAMETERS

P = (V ELL T(B) # DFM) / (V ELL T(B) # DFM);

PROB = 1 - PROBF (DFM / DFM);

REGION = IF ANCOVA = 1 THEN DO:

**COMPUTE ANCOVA TEST TO DETERMINE DEGREE OF "GROUP" PGC;
**ITERATE TEST THE NEED FOR THE NEXT-HIGHEST ORDER MODEL;
IF PROB((Q-1)< .05 THEN QQ=Q-1;
ELSE QQ = 0;

PROBANC = PROBUN(Q-2, 1);

END:

NOTE **

TOP-DOWN SERIES OF TESTS OF THE HYPOTHESIS THAT*;
NOTE THE C-TH ORDER POLYNOMIAL TERM WHEN ALL HIGHER *;
NOTE OTHER TERMS ARE USED AS COVARIATES;
NOTE **

NAME = "DP-ERROR" "SS-ERROR" "PROB" "F";

LABELS = NAME ROWNAME=NUll;

NULL2 = "M " "M " "M " "M " "M " "M ";

DO WHILE (CUT < (Q-2));

CUT = CUT + 1: LINK ANCOVA; END;

644
NOTE: THE HIGHEST ORDER TERM INCLUDED IN SIGNIFICANT.*
NOTE: TEST OF THE HYPOTHESIS LR= 0 WHEN ALL HIGHER
NOTE: TERMS WERE USED AS COVARIATES.
NOTE: NAME = "PGC TERM".
PRINT QQ COLNAME = NAME ROWNAME = NULL;
END;

**** PERFORM UNIVARIATE TESTS OF BETWEEN-SUBJECTS FACTORS;
ELSE DO;

OLS:
NOTE: FITTING Q-TH OLS MODEL, COMPUTING B, *
NOTE: AND SIGMAHAT(I) AND SIGMAHAT(B). *
NOTE: **********************
XPI = INV ([X*STR(I') * X];
DFE = N - R - IP - Q;
B = XPI* X*STR(I') * YSTAR;
S = X*STR(I') * (I - I) - X * XPI * X*STR(I');
SIGMAHAT = (A * X*STR(I')) * S * A1) / DFE;
COV B = INV. [X(I1:B) X*STR(I') X(I1:B)] * SIGMAHAT(1:0,1:Q);
PRINT SIGMAHAT B COV_B;
RETURN;

WLS:
NOTE: **********************
NOTE: FITTING Q-TH WLS MODEL, COMPUTING B, *
NOTE: AND SIGMAHAT(I) AND SIGMAHAT(B).
NOTE: **********************
XPI = INV ([X*STR(I') * X];
DFE = N - R - Q;
S = X*STR(I') * (I - I) - X * XPI * X*STR(I');
SIGMAHAT = (A * X*STR(I')) * S * A1;
B = XPI* X*STR(I') * YSTAR;
SIGMAHAT = A1;
COV B = XPI* SIGMAHAT;
PRINT SIGMAHAT B COV_B;
RETURN;

ANCOVA:
NOTE: YSTAR( ,Q-CNT); IF CNT NE 0 THEN X2 = X 1 YSTAR( ,Q-CNT+1 = Q);:
XPI2 = INV ([X*STR(I') * X2];
B2 = XPI2* X*STR(I');
DFE2 = DFE-CNT;
BESID = YT - X2 * B2;
SSE = BESID*STR(I');
LB = C2 - B2;
ILXPXIL = C2 * XPI2 * C2*STR(I');
ILXPXIL = INV. [ILXPXIL];
SSR = BESID*STR(I');
F = (SSR / DFE2) / (SSR / DFE2);
PROBANC = 1 - PROBF (F, DFE2, DFE2);
PRINT LABELS COLNAME = NULL2 ROWNAME = NULL;
IF (QQ = 0 AND PROBANC < .05) THEN QQ = Q-CNT-1; RETURN;

%END;