USE OF SAS TO ANALYZE EXPERIMENTS CONDUCTED OVER TIME WITH MULTIVARIATE RESPONSES AND AUTOCORRELATED ERRORS

Walter W. Stroup, University of Nebraska-Lincoln

ABSTRACT

An experimental procedure common in applied biological research involves the use of a simple experimental design to observe responses on multiple variables, often correlated, at several intervals over time. The purpose of such experiments is generally to evaluate treatment effects on changes in the vector of responses over time. The method for doing this is given in such multivariate texts as Morrison (1975) and a method for handling this in SAS is given in Courtright (1978).

An aspect of the data arising from this type of experiment which is not as commonly addressed is the problem of serial correlation among successive observations from the same plot over time. Bjornsson (1978) describes the phenomenon, which occurred in a series of grassland experiments conducted over a 17-year period in Iceland, and proposes a solution to the problem using a transformation of the data and the design matrix based on estimates of the autoregressive parameters. However, his procedure is limited to the univariate case. It is not clear how to handle a similar problem in the multivariate case, or how to use SAS to implement a multivariate analysis such as this.

The purpose of this paper is to show how such an experiment can be analyzed and specifically how SAS can be employed to perform the required calculations. Although the methods described here can be extended to a variety of experimental designs and to the general autoregressive model of lag $p$ (AR($p$)), for the sake of clarity the discussion will be confined to the randomization complete block design and to an AR(1) model. The randomized block is by far the most common agronomic design and as Bjornsson (1978) points out, the degree of autoregression in such experiments is typically strongest for observations which are contiguous in time and fall off increasingly as the lag increases. Also for the purposes of illustration, a data set collected between 1975 and 1978 for a Range Management experiment at the University of Nebraska-Lincoln will be used. The purpose of this experiment was to investigate the changes in botanical composition which result over time from the repeated application of several weed control treatments.

BACKGROUND THEORY AND RESULTS

For the sake of illustration, assume that we wish to analyze an experiment in which $v$ treatments have been applied to plots in each of $r$ blocks. Observations are taken on each of $s$ response variables at each of $n$ time intervals (e.g., years). Let the model for such an experiment be

\[ Y_{ijkt} = \mu_{kt} + b_{ikt} + e_{jkt} + Z_{ijkt} \]

where $Y_{ijkt}$ is the observation on the $i^{th}$ block ($i = 1, 2, \ldots, r$), $j^{th}$ treatment ($j = 1, 2, \ldots, v$), $k^{th}$ response variable ($k = 1, 2, \ldots, s$), and $t^{th}$ time interval ($t = 1, 2, \ldots, n$).

INTRODUCTION

One of the more common experimental procedures in agronomic research is to employ a simple experimental design, such as a randomized complete block, and take contemporaneous observations on several inherently correlated variables at regular time intervals, such as years. The objective of such experiments is manifold: the researcher wishes to measure the relationships among the response variables, to evaluate the effect of treatments on these response variables, and to determine if changes in the effects of these treatments relative to one another occur over time. It is common practice to analyze the experiment as a split-plot, with treatment as the whole plot effect and time as the split-plot effect. This analysis is typically applied to each variable separately. The procedure is well documented in many statistical methods texts, e.g., Steel and Torrie (1980). An elaboration of this analysis is to recognize the multivariate nature of the responses and to perform a multivariate analysis of variance on the vector of response variables. Specific treatment, time, and variable comparisons, or interaction effects among any two or all three of these can be accomplished using contrasts.

The method for doing this is given in such multivariate texts as Morrison (1975) and a method for handling this in SAS is given in Courtright (1978).
and

\( \mu_{kt} = \text{overall mean for variable } k \text{ and time } t \)

\( b_{ikt} = \text{effect of block } i \text{ for variable } k \text{ at time } t \)

\( r_{jkt} = \text{effect of treatment } j \text{ for variable } k \text{ at time } t \)

\( Z_{ijkl} = \text{residual for the } ijth \text{ block-treatment combination (plot) for variable } k \text{ at time } t. \)

Let \( \mathbf{Z}_{ijt} = (Z_{ij1t}, Z_{ij2t}, \ldots, Z_{ijst}) \) be the vector of residuals at time \( t \) for the \( ijth \) plot.

The first-order autoregressive model assumes that

\[ \mathbf{Z}_{ijt} = \mathbf{P} \mathbf{Z}_{ij,t-1} + \mathbf{e}_{ijt}, \]

where \( \mathbf{P} \) is an \( s \times s \) matrix of autoregressive parameters and the \( \mathbf{e}_{ijt}'s \) are \( \text{NID}(0, I) \). This model is described in Fuller (1975, pp. 70-75).

In the univariate case, the procedure for analyzing such a data set is to estimate the autoregression parameter (denoted \( \hat{p} \)) and to transform the model using the following transformation matrix

\[ T = \begin{bmatrix}
(1 - \hat{p}^2)Y_{ij1} & 0 & 0 & \cdots & 0 \\
\hat{p} & 1 & 0 & \cdots & 0 \\
0 & \hat{p} & 1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
0 & \cdots & \hat{p} & 1 \\
\end{bmatrix} \]

where \( \hat{p} \) is the estimate of \( p \). If we assume to univariate version of model (1) to be

\[ Y_{ijt} = \mu_t + b_{it} + r_{jkt} + Z_{ijt}, \]

the estimation of the parameters \( \mu_t, b_{it}, \) and \( r_{jkt} \) is accomplished using the method of generalized least squares on the transformed model, where the observations are transformed as

\[ Y_{ij1} = (1 - \hat{p}^2)^{\frac{1}{2}} Y_{ij1}, \]

\[ Y_{ijt} = -\hat{p}Y_{ij1,t-1} + Y_{ijt}; t = 2, 3, \ldots, n. \]

and the design matrix is transformed analogously.

The method for estimating \( \hat{p} \) and transforming the model is given in Fuller (1975, pp. 419-429) and a specific application of the transformation for univariate agronomic data is given in Bjornsson (1978).

In the multivariate case, the mechanics of the transformation are more tedious but the idea is a straightforward extension of the analysis for the univariate case. First, the matrix of autoregressive parameters \( \mathbf{P} \) must be estimated. Denote the estimate of \( \mathbf{P} \) as \( \hat{\mathbf{P}} \). Let \( \mathbf{P}_{ij} \) be the estimate of \( \mathbf{P} \) for the \( ijth \) plot. Fuller (1975) gives the form of \( \mathbf{P}_{ij} \) as

\[
\hat{\mathbf{P}}_{ij} = \mathbf{P}_{ij1}^{-1} \mathbf{P}_{ij2}^{-1} \cdots \mathbf{P}_{ijt}^{-1}
\]

where \( \mathbf{P}_{ij} = (Z_{ij1t-1}, Z_{ij2t-1}, \ldots, Z_{ijst-1}) \).

For the purpose of this paper we shall use an estimate of \( \mathbf{P} \) of the form

\[
\tilde{\mathbf{P}}_{ij} = \sum_{i=1}^{r} \sum_{j=1}^{v} \tilde{P}_{ij}/rv
\]

This is a heuristic estimator and may not be the optimal estimate of \( \mathbf{P} \). This question has yet to be resolved.

The matrix \( \tilde{\mathbf{P}} \) can be used to form a multivariate extension of the transformation matrix given in (3). Letting \( \mathbf{y}'_{ijt} = (Y_{ij1t}, Y_{ij2t}, \ldots, Y_{ijst}) \), then the specific of this transformation for the AR(1) case is

\[ \mathbf{Y}_{ij1} = A\mathbf{y}_{ij1}, \]

\[ \mathbf{Y}_{ijt} = -\hat{p}\mathbf{Y}_{ij,t-1} + \mathbf{y}_{ijt}; t = 2, 3, \ldots, n. \]

where \( A \) is an \( s \times s \) matrix such that \( A\mathbf{A} = I_s - \hat{p}\mathbf{I} \).

This transformation can be expressed in matrix form in the following way. Let \( y'_{ijt} = (y_{ij1t}, \ldots, y'_{ijrv}). \) Define the multivariate AR(1) transformation as

\[
\mathbf{T} = \begin{bmatrix}
A \otimes I_{rv} & 0 & 0 & 0 \\
-\mathbf{P} \otimes I_{rv} & I_{svr} & 0 & 0 \\
0 & -\mathbf{P} \otimes I_{rv} & I_{svr} & 0 \\
\vdots & \vdots & \ddots & \ddots \\
0 & \cdots & \hat{p} & 1 \\
\end{bmatrix}
\]

where \( I_q \) is a \( q \)-dimensional identity matrix, \( \otimes \) denotes the Kronecker direct product, and \( O \) denotes a \( (svr) \times (svr) \) matrix of zeroes. Then the transformation is given by

\[ \mathbf{Ty} = \mathbf{TX} + \mathbf{z}, \]

where \( y' = (y'_{ij1}, y'_{ij2}, \ldots, y'_{ijrv}) \) and \( X, \mu, \) and \( z \) represent the corresponding design matrix, vector of mean, block, and treatment effect parameters, and vector of residuals respectively. Note that the \( X \) matrix is block diagonal with each block corresponding to the matrix implied by the experimental design.

A generalized least squares solution for the vector \( \mathbf{y} \) can be obtained from (6) and is as follows:

\[
\mathbf{\hat{y}} = (X'\mathbf{TX})^{-1}X'\mathbf{TY}
\]
Note that equation (8) is similar to the Aitken estimator given in Press (1972, pp. 225–227). Press also gives the standard error for \( \hat{\beta} \) and a method for testing various hypotheses concerning \( \hat{\beta} \).

IMPLEMENTING THE ANALYSIS USING SAS

Options Available to the User

The estimation of \( \hat{\beta} \) described above can be performed using SAS. The procedure would involve the following steps.

1. Use PROC GLM to perform an analysis of the untransformed data using model (1). Output the vector of residuals \( \mathbf{z} \).
2. Use PROC MATRIX to estimate \( \mathbf{P} \) according to equations (4) and (5).
3. The matrix \( \mathbf{A} \) can be computed using the function \( \text{HALF}(\mathbf{I}(\mathbf{S}) - \mathbf{P}' \mathbf{P}) \) in PROC MATRIX.
4. Form the \( \mathbf{T} \) matrix and compute \( \mathbf{r}_y, \mathbf{r}_x \) using PROC MATRIX. Output the matrices \( \mathbf{r}_y \) and \( \mathbf{r}_x \).
5. Estimate \( \hat{\beta} \) using PROC GLM using the vector of transformed observations and the transformed design matrix output in step 4.

In practice, however, this procedure is not recommended since the size of the matrices will be prohibitively large. For example, in a Range Management study conducted at the University of Nebraska–Lincoln, the effects of 6 treatment variables on changes in 4 correlated response variables over a 4-year period were studied. The study involved 4 blocks in a randomized complete block design. Using the above procedure would require a \( 384 \times 384 \) \( \mathbf{T} \) matrix and a \( 384 \times 144 \) \( \mathbf{X} \) matrix which is well beyond the capacity of SAS. Obviously a more practical approach is required.

There are two approaches which might be used which will be distinguished by the labels "standard contrast approach" and "autoregressive contrast approach". The "standard contrast approach" is a mild extension of the "contrast approach" described by Courtright (1978). The researcher could define a priori contrasts according to some biologically meaningful rationale on both the contemporaneous response variables and over time as well as the interaction contrasts which follow from the response variable and time contrasts. These contrasts could then be analyzed separately. Since this method has been illustrated by Courtright, no further discussion is required in this paper. The "autoregressive contrast approach", which will be demonstrated below, involves a combination of contrasts defined on the contemporaneous response variables, estimation of the autoregressive parameters, transformation of the data and design matrix using the \( \mathbf{T} \) matrix given in (3), and analysis of each transformed response contrast using the general linear model procedure. In many agronomic experiments the "autoregressive contrast approach" is preferable for two reasons: first, there may not exist a theoretically reasonable or informative set of contrasts defined over time, and, second, this approach involves explicit estimation of the autoregressive structure of the observations which may in itself be of biological interest.

For the purposes of illustration, data from the Range Management study mentioned earlier will be used. In this experiment several weed control treatments were applied to the same plots over a period of four years. The percent composition of four species groups—annual grasses, perennial grasses, perennial forbes, and sagewort, an undesirable weed—was measured annually. Percent composition of other species were also measured, but are not of interest in this analysis. The purpose of the experiment was to determine the effect of treatments on changes in species composition over time—i.e. was there a year-by-treatment interaction? Although the experiment involved a total of six treatments and four blocks, to keep this illustration compact only three treatments and three blocks will be used in this example. Note that negative correlation is inherent among the response variables since an increase in one species would require a corresponding decrease in one or more of the other species within the same plot.

Programming the Analysis

The first step of the analysis is to create response variables consisting of contrasts defined on the species groups. In this experiment the following contrasts were deemed appropriate: sagewort versus the average of the other species groups, annual grasses versus the average of the two perennial groups, and perennial grasses versus perennial forbes. The necessary control cards for this step are given in Table 1.

The above three response variable contrasts are then analyzed using PROC GLM using a standard split-plot-over-time model. Although this analysis is not corrected for autocorrelation, the purpose of this step is to output residuals so that autoregressive parameters can be estimated. This is accomplished by sorting the data set of residuals by block and treatment and estimating the first-order autocorrelation coefficient among these residuals for each block-treatment combination.

<table>
<thead>
<tr>
<th>TABLE 1. SAS Control Cards for Generating Residuals and Estimating the Autoregressive Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATA A;</td>
</tr>
<tr>
<td>INPUT BLK YR TRT SGWRT ANN G PER G PER F;</td>
</tr>
<tr>
<td>SG_V_OTH=(3*SGWRT-ANN G-PER G-PER F)/SQR(12);</td>
</tr>
<tr>
<td>ANN V PER=(2*ANN G-PER G-PER F)/SQR(6);</td>
</tr>
<tr>
<td>PG V PF=(PER G-PER F)/SQR(2);</td>
</tr>
<tr>
<td>* VARIABLES SGWRT, ANN G, PER G, AND PER F ARE THE CORRELATED RESPONSE VARIABLES. RESPONSE VARIABLE CONTRASTS ARE NORMALIZED BY DIVIDING BY SUM OF SQUARED CONTRAST COEFFICIENTS;</td>
</tr>
<tr>
<td>CARDS;</td>
</tr>
<tr>
<td>PROC GLM; CLASSES BLK TRT YR;</td>
</tr>
<tr>
<td>MODEL SG_V_OTH ANN V PER PG V PF = BLK TRT BLK*TRT</td>
</tr>
</tbody>
</table>

662
Once the first-order autocorrelations are calculated for each block-treatment combination, the autoregressive parameter for each response variable contrast is calculated according to the formula

\[ \hat{\rho}_k = \frac{1}{\sqrt{v}} \sum_{i=1}^{n} \sum_{j=1}^{n} \hat{\rho}_{ijk} \]

where \( \hat{\rho}_{ijk} \) is the estimated first order autoregressive parameter for the ith block, jth treatment, and kth response variable contrasts. The output for the response variable contrast \( S_2V_\text{OTH} \) is summarized in Table 2. In this example, the autocorrelation coefficient has been used as a quick estimate of the autoregression parameter in the manner of Bjornsson (1978). Fuller (1975, pp. 419-429) presents a more rigorous estimation procedure. However, the autocorrelation is convenient and serves the purpose of this example. Output for the other contrasts is handled analogously to that for \( S_2V_\text{OTH} \); for the sake of brevity subsequent discussion will therefore be confined to \( S_2V_\text{OTH} \).

**TABLE 2. Summary of Output From Preliminary Analysis of Data Using Control Statements From Table 1**

<table>
<thead>
<tr>
<th>DEPENDENT VARIABLE: ( S_2V_\text{OTH} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS: 37.59</td>
</tr>
<tr>
<td>COEFFICIENT OF VARIATION: 29.55</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>TYPE 1 SS</th>
<th>F VALUE*</th>
<th>PR-F*</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLK</td>
<td>2</td>
<td>237.17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRT</td>
<td>2</td>
<td>47.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLK*TRT</td>
<td>4</td>
<td>328.81</td>
<td></td>
<td></td>
</tr>
<tr>
<td>YR</td>
<td>3</td>
<td>850.32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLK*YR</td>
<td>6</td>
<td>616.95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRT*YR</td>
<td>6</td>
<td>470.94</td>
<td>2.09</td>
<td>.1308</td>
</tr>
</tbody>
</table>

**AUTOCORRELATIONS BLK/TRT**

-0.24 -0.80 -0.67
-0.74 -0.21 -0.40
-0.47 -0.71 -0.77

**ESTIMATE OF AUTOREGRESSIVE PARAMETER FOR SUBSEQUENT ANALYSIS:** -0.55813 (using equation 9)

* Since TRT*YR effect is of paramount interest, F-value and PR-F are given only for this effect. Tests of TRT and YR main effects would require appropriate TEST statements in control cards.

At this point the data and design matrix are transformed according to the method described in equations (3) through (5). This transformation can be accomplished in SAS using PROC MATRIX. Once the transformation has been computed, the transformed data and design matrix are output and are analyzed using the general linear models procedure. The necessary control cards for these steps are given in Table 3.

**TABLE 3. SAS Control Cards for Transforming the Data and Design Matrix and for Analysis Recounting for Autoregression**

```
DATA A;
  INPUT BLK YR TRT SGWRT ANN G PER G PER F;
  SGVOTH=(3*SGWRT-ANN G-PER G-PER F)/SQRT (12);
  DROP-SGWRT ANN G PER G PER F;
* COMPLETE ANALYSIS WOULD REQUIRE ADDITIONAL RUNS FOR REMAINING RESPONSE VARIABLE CONTRASTS. THESE WOULD BE HANDLED SIMILARLY USING CONTRASTS DEFINED IN TABLE 1;
CARDS;
  PROC MATRIX;
  FETCH DATASET;
  BLOCK=DESIGN (DATASET(,I));
  NBLK=NCOL (BLOCK);
  BLK=BLOCK (,1: OF BLK);
  FREE BLOCK;
  YEAR=DESIGN (DATASET (,2));
  DF_YR=NCOL (YEAR)-1;
  YR=YEAR (.1: DF_YR);
  FREE YEAR;
  TREATMENT=DESIGN (DATASET (,3));
  NTRT=NCOL (TREATMENT);
  TRT=TREATMENT (,1: DF_TRT);
  FREE TREATMENT;
  Y=DATASET 1(,4);
  BXT=BLK @ TRT;
  BXY=BLK @ I YR;
  TXY=TRT @ I YR;
  NOBS=NROW(Y) ;
  X=J(NOBS,l,l)ll mil TRTll BXTll BXY 11 TXY;
  SIZE=NBLK*NTRT;
  TM=T @ I (SIZE);
  T=T @ I (SIZE);
  * T IS THE FIRST ORDER AUTOREGRESSIVE TRANSFORMATION GIVEN IN EQUATION 3, USING PARAMETER ESTIMATE FROM TABLE 2. TM IS KROENECKER DIRECT PRODUCT TO YIELD FULL TRANSFORMATION MATRIX FOR ALL BLOCK-TREATMENT COMBINATIONS.
```
The output for the analysis of the transformed data is summarized in Table 4. Notice that the MSE and coefficient of variation have been substantially reduced relative to the analysis on Table 2 which assumes independent errors. More importantly, the F-value for the test of year-by-treatment interaction, computed by

\[
F(\text{YR*TRT}) = \frac{\text{SS(}T1Y1\text{)} + \text{SS(}T1Y2\text{)} + \ldots + \text{SS(}T2Y3\text{))}{6} \times \frac{\text{MSE}}{}
\]

is 4.02 with a probability of a greater F of .0194. The transformation, then, serves the purpose of accounting for the autocorrelation among the errors and thus increasing the power of the test of the year-by-treatment interaction. In the original analysis this effect would have gone undetected.

The parameter estimates which accompany the analysis of variance in the general linear model output correspond to the biased parameter estimates which would normally accompany the GLM output when the SOLUTION option is used. These can be used either to calculate least squares means and test differences among them using the LSD procedure or to calculate and test treatment, year, and treatment-by-year contrasts.

### TABLE 4. Summary of Generalized Least Squares Analysis Accounting for Autoregression

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type I SS</th>
<th>F-Value</th>
<th>P&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DEPENENT VARIABLE: SG_V_OTH</strong> (Transformed)</td>
<td></td>
<td>18.55</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>MSE:</strong></td>
<td></td>
<td>18.55</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>SOURCE</strong></th>
<th><strong>DF</strong></th>
<th><strong>F-VALUE</strong></th>
<th><strong>P&gt;F</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BLK</strong></td>
<td>2</td>
<td>815.97</td>
<td></td>
</tr>
<tr>
<td><strong>B1</strong></td>
<td>1</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td><strong>B2</strong></td>
<td>1</td>
<td>815.82</td>
<td></td>
</tr>
<tr>
<td><strong>TRT</strong></td>
<td>2</td>
<td>222.82</td>
<td></td>
</tr>
<tr>
<td><strong>T1</strong></td>
<td>1</td>
<td>52.50</td>
<td></td>
</tr>
<tr>
<td><strong>T2</strong></td>
<td>1</td>
<td>170.32</td>
<td></td>
</tr>
<tr>
<td><strong>BLK*TRT</strong></td>
<td>4</td>
<td>753.42</td>
<td></td>
</tr>
<tr>
<td><strong>B1T1</strong></td>
<td>1</td>
<td>8.09</td>
<td></td>
</tr>
<tr>
<td><strong>B1T2</strong></td>
<td>1</td>
<td>481.86</td>
<td></td>
</tr>
</tbody>
</table>

The purpose of this paper has been to show that analysis of a designed experiment with a multivariate response vector and a time series error structure is indeed possible using SAS. The procedure proposed in this paper has several advantages to the statistician working with statistically naive researchers, who might otherwise shy away from such forbidding ideas as multivariate analysis or time series. First, the vector of correlated responses is reduced to a set of intuitively appealing and easily understandable response contrasts. Second, the autocorrelations are presented in a tractable form. Finally, the generalized least squares analysis appears in the form of an ANOVA table, the interpretation of which is familiar to researchers whose statistical background may consist of no more than a single - and very applied - course in experimental design and analysis.

### REFERENCES


Courtright, J.A. 1978. Repeated Measures in SAS.
Proceedings of the Third Annual SAS Users


STATISTICS
Nonparametrics and Categorical
Data Analysis