

METHODS OF PARAMETER SELECTION IN EXPLORATORY MODEL ANALYSES

BU-1537 -M

October, 2000

**Walter T. Federer
and
Russell D. Wolfinger**

Keywords: none provided.

Abstract: Exploratory model selection is a real possibility with the use of present day computer software. Here three methods of selecting model blocking parameters are presented. Method 1 selects parameters from a fixed model analysis such as may be obtained from computer software such as a SAS/GLM procedure, for example. Method 2 uses REML (restricted maximum likelihood) solutions for the variance components for each of the blocking parameters. Method 3 uses BLUP-like solutions for blocking parameter effects using some such computer software as the ASA/MIXED procedure. Several modifications are discussed. An eight-row by seven-column designed experiment for seven treatments is used to illustrate the method of selecting row and column regressors and interactions of regressions.


METHODS OF PARAMETER SELECTION IN EXPLORATORY MODEL ANALYSES

by

Walter T. Federer
434 Warren Hall, Cornell University, Ithaca, NY 14853

and

Russell D. Wolfinger
SAS Institute, Inc. B52, SAS Campus Drive, Cary NC 27513

1537
BU--M

October 2000

In the Technical Report Series of the Department of Biometrics, Cornell University,
Ithaca, NY 14853

ABSTRACT

Exploratory model selection is a real possibility with the use of present day computer software. Here three methods of selecting model blocking parameters are presented. Method 1 selects parameters from a fixed model analysis such as may be obtained from computer software such as a SAS/GLM procedure, for example. Method 2 uses REML (restricted maximum likelihood) solutions for the variance components for each of the blocking parameters. Method 3 uses BLUP-like solutions for blocking parameter effects using some such computer software as the SAS/MIXED procedure. Several modifications are discussed. An eight-row by seven-column designed experiment for seven treatments is used to illustrate the method of selecting row and column regressors and interactions of regressions.

INTRODUCTION

We discuss three methods of model selection in exploratory model selection analyses. One of the methods selects model parameters using a fixed model approach. The properties of the method used here have been examined by Bozivich, Bancroft, and Hartley (1956). The other two methods select parameters from a mixed model approach. The properties of these two methods are unknown but it is suggested that an investigation of their properties and of a comparison of the three methods should be undertaken. If the problem is not resolvable theoretically, then it is suggested they be compared via simulations. The discussion herein is couched in terms of an experiment designed as an eight-row by seven-column layout for seven treatments (Federer and Schlotfeldt, 1954). The response variable is height of tobacco plants. Orthogonal polynomial regressions of row effects, of column effects, and of interactions of row and column regressions will be used to describe the spatial variation in the experiment. Since the regressions are

functions of random effects, rows and columns, they themselves are considered to be random effects

METHOD 1

A fixed effect analysis is performed on the data. For the above experiment, one could consider a variety of response models to determine which model accounts for the spatial variation present in the experiment. Some of these models written in the SAS/GLM procedure format are:

$$\text{Height} = \text{row column treatment} \quad (1)$$

$$\text{Height} = r_1 r_2 r_3 r_4 r_5 r_6 r_7 c_1 c_2 c_3 c_4 c_5 c_6 \text{ treatment} \quad (2)$$

A r_i is the i th orthogonal polynomial regressor of row effects and a c_j is the j th orthogonal polynomial regressor of column effects. Note the two models are identical except for formulation and that the regressions are functions of the row or column effects. Since the spatial variation may not be in the same direction as the row-column orientation, it may be necessary to use interactions of regressions to explain the spatial variation. Two such models are:

$$\begin{aligned} \text{Height} = & \text{row column } r_1*c_1 r_1*c_2 r_1*c_3 r_1*c_4 r_2*c_1 r_2*c_2 r_2*c_3 r_2*c_4 \\ & r_3*c_1 r_3*c_2 r_3*c_3 r_3*c_4 r_4*c_1 r_4*c_2 r_4*c_3 r_4*c_4 \text{ treatment} \end{aligned} \quad (3)$$

$$\begin{aligned} \text{Height} = & r_1 r_2 r_3 r_4 r_5 r_6 r_7 c_1 c_2 c_3 c_4 c_5 c_6 r_1*c_1 r_1*c_2 r_1*c_3 r_1*c_4 \\ & r_2*c_1 r_2*c_2 r_2*c_3 r_2*c_4 r_3*c_1 r_3*c_2 r_3*c_3 r_3*c_4 r_4*c_1 r_4*c_2 r_4*c_3 \\ & r_4*c_4 \text{ treatment} \end{aligned} \quad (4)$$

Here the interaction terms of interest are denoted by r_i*c_j , $i, j = 1, 2, 3, 4$. The more spotty the spatial variation in the experiment the higher degree of interaction regressions will be needed to account for this.

Since it is desirable to use as few regressors as possible to explain the variation, some method of selecting which parameters to retain in the model is needed. Such a method has been provided by Bozovich, Bancroft, and Hartley (1956). A fixed effect F-test of each parameter is performed. Any F-value exceeding the tabulated F-value at the 25% level is retained in the model. Those not exceeding this value are omitted from the model. Federer, Crossa, and Franco (1998) give the rule as:

Step 1. Obtain the row regressions, r_i , $i = 1, 2, \dots, r - 1$ for the r rows, and the column regressions, c_j , $j = 1, 2, \dots, c - 1$ for the c columns. Compute the sum of squares attributable to each regression. All regressions whose F-values are smaller than the tabulated F-value at the 25% level are relegated to the residual category. The remaining r_i and c_j are retained as blocking variables.

Step 2. Determine which interactions r_i*c_j are to be considered. Suppose that $i, j = 1, 2, 3, 4$ is the range under consideration. Together with the r_i and c_j retained in Step 1,

compute the sum of squares attributable to each $r_i \cdot c_j$ interaction. The rule in Step 1 is applied to each of the interactions to determine whether to omit or to retain them as blocking variables.

Step 3. Using the r_i , the c_j , and the $r_i \cdot c_j$ parameters retained in Steps 1 and 2, check to determine if any more should be omitted. Then use the remaining parameters in the response model to obtain an analysis of variance (ANOVA), the treatment fixed model and mixed model means, and standard errors.

Using the above rule, Federer, Crossa, and Franco (1998) found the following response model for the eight-row by seven-column design:

$$\text{Height} = c_1 c_2 c_3 c_5 r_1 r_2 r_3 r_5 r_6 r_7 r_1 \cdot c_1 r_1 \cdot c_2 r_1 \cdot c_4 r_2 \cdot c_3 r_2 \cdot c_4 r_3 \cdot c_2 \text{ treatment} \quad (5)$$

The residual mean square for equation (1) was 7,352 with 36 degrees of freedom whereas it was 4,204 with 33 degrees of freedom for equation (5).

Since only c_4 , c_6 , and r_4 were omitted, the following model is suggested as appropriate for this data set:

$$\text{Height} = \text{row column } r_1 \cdot c_1 r_1 \cdot c_2 r_1 \cdot c_4 r_2 \cdot c_3 r_2 \cdot c_4 r_3 \cdot c_2 \text{ treatment} \quad (6)$$

Equation (6) is a form of additive main effects and multiplicative interaction model (AMMI) whereas the usual form is to use principal components for the interactions. For a fixed effect analysis, the residual mean square is 4,204 for equation (5) and is 4,418 for equation (6).

Equation (6) is the response model to be used in a mixed model analysis with random blocking effects. It is suggested that the following RANDOM statements be used for the SAS/MIXED analysis and REML solutions (SAS's default option):

```
RANDOM row column;
RANDOM selected  $r_i \cdot c_j$  parameters/type = toep(1);
```

The statement "/type = toep(1)" is used to pool all interactions into one source of variation and compute one variance component for all $r_i \cdot c_j$ parameters in the statement. This should improve the computational stability and make the use of REML more appropriate.

METHOD 2

A second method for selecting parameters in a response model is given below:

Step 1. The row category is selected first as rows are orthogonal to treatments and to columns. Then, the SAS/MIXED procedure is used with r_i parameters as random variables. A REML variance component solution is obtained for each r_i . Any parameter

with a zero solution of the variance component is omitted from the model. The remaining parameters are retained as blocking variables.

Step 2. Using the selected parameters from Step 1 and the parameters for the column effects, obtain the REML solutions for column parameter variance components. Use the following RANDOM statements:

```
RANDOM c1 c2 c3 ... cc-1;
RANDOM selected ri parameters from Step 1/type = toep(1);
```

Omit all c_j parameters which have a zero REML solution for the variance component.

Step 3. Using the selected r_i and c_j parameters from Steps 1 and 2, obtain the REML solutions for the r_i*c_j , $i, j = 1, 2, 3, 4$, say, variance components. Omit all r_i*c_j which have a zero solution for their variance components. Use the following RANDOM statements:

```
RANDOM selected ri/type = toep(1);
RANDOM selected cj/type = toep(1);
RANDOM r1*c1 r1*c2 r1*c3 r1*c4 r1*c1 r2*c2 r2*c3 r2*c4 r3*c1 r3*c2
      r3*c3 r3*c4 r4*c1 r4*c2 r4*c3 r4*c4;
```

Step 4. Using the selected r_i , c_j , and r_i*c_j parameters obtain the adjusted treatment effects, means, and standard errors for the following RANDOM statements:

```
RANDOM selected ri/type = toep(1);
RANDOM selected cj/type = toep(1);
RANDOM selected ri*cj/type = toep(1);
```

One need not use the "/type = toep(1)" part of the RANDOM statement but the stability of the computations should be improved if it is used. Also, using more than one degree of freedom mean squares should make the use of the REML procedure more appropriate. The three RANDOM statements were used as it is likely that the pooled r_i have a different variance than the pooled c_j or the pooled r_i*c_j . They could all be pooled into one group if desired but this is considered to be inappropriate in general.

For the data set described above, the parameters selected in Step 1 were r_1 and r_3 . The c_j parameters selected in Step 2 were c_1, c_2, c_3, c_4 , and c_5 . The eight r_i*c_j parameters selected in Step 3 were $r_1*c_1, r_1*c_2, r_1*c_3, r_1*c_4, r_2*c_2, r_2*c_3, r_3*c_2$, and r_4*c_4 . The response model obtained was:

$$\text{Height} = r_1 r_3 c_1 c_2 c_3 c_4 c_5 r_1*c_1 r_1*c_2 r_1*c_3 r_1*c_4 r_2*c_2 r_2*c_3 r_3*c_2 r_4*c_4 \text{ treatment} \quad (7)$$

When Steps 1 and 2 are combined, the following response model resulted:

$$\text{Height} = r_1 r_2 r_3 r_5 r_6 r_7 c_1 c_2 c_3 c_4 c_5 r_1*c_1 r_1*c_2 r_1*c_3 r_1*c_4$$

$$r2*c2 \ r2*c3 \ r2*c4 \ r3*c2 \ treatment \quad (8)$$

The residual mean square from PROC GLM for equation (7) was 5,542 and for equation (8) was 4,130. The latter is approximately the same as was obtained for equation (6).

METHOD 3

The exploratory parameter selection for this method proceeds as follows:

Step 1. For a PROC MIXED REML analysis, use the following RANDOM statements:

```
RANDOM r1 r2 r3 ... rr-1/solution;
RANDOM c1 c2 c3 ... cc-1/solution;
```

Omit all r_i and c_j for which the probability of a larger t-value is greater than .25.

Step 2. Using the r_i and c_j that were not omitted in Step 1, use the following RANDOM statements:

```
RANDOM selected ri/type = toep(1);
RANDOM selected cj/type = toep(1);
RANDOM all ri*cj under consideration/solution;
```

Omit all r_i*c_j for which the probability of a larger t-value is greater than .25.

Step 3. Using the selected r_i , c_j , and r_i*c_j , use a PROC MIXED analysis and the following RANDOM statements to obtain the treatment effects and means and their standard errors:

```
RANDOM selected ri/type = toep(1);
RANDOM selected cj/type = toep(1);
RANDOM selected ri*cj/type = toep(1);
```

Using the above parameter selection method, the following response model was obtained:

$$Height = r1 \ r2 \ r3 \ r5 \ c1 \ c2 \ c3 \ c5 \ r1*c2 \ r1*c4 \ r2*c3 \ r3*c2 \ treatment \quad (9)$$

The residual mean square from a PROC GLM analysis for equation (9) was 4,862.

MODIFICATION OF METHODS

Several modifications of the above three methods are possible. If the residual mean square is associated with a large enough number of degrees of freedom, the steps may be combined. The r_i , c_j , and r_i*c_j parameters to retain as blocking parameters are all selected at the same time. This was done for this data set for Method 1 and the same parameters were retained as with the above sequential method.

COMMENTS

Different response models were obtained for each of the above methods. Likewise different response models may be obtained if parameters are selected for categories simultaneously or sequentially. Possibly the reason only r_1 and r_3 were selected in the sequential form of Method 2 was the large residual variance, 30,228, associated with only a row-treatment (randomized complete block) model. Using a row-column-treatment model, the residual mean square was reduced to 7,352.

Properties of Method 1 have been investigated by Bozivich, Bancroft, and Hartley (1956). Properties of Methods 1 and 2 are unknown. It is suggested that they are likely candidates to replace Method 1 for use in exploratory model selection. The tabulated values of the F and t statistics used here was the 25% level. However, the advances in computer software is such that one could use a 21%, a 26%, or some other value for determining whether or not to use a parameter as a blocking variable. Perhaps a 22.567% value would be optimal for one of the methods. A refinement of the point at which to omit parameters may improve the properties of the method.

With regard to Method 2, a zero variance component is not involved in obtaining the adjusted treatment means. However, omitting it will change the values of the other parameter variance component solutions and most likely the standard errors. Hence, before obtaining the final form of the response model, treatment effects and means, and standard errors, they should be omitted.

Instead of using the residual mean square from a fixed effect analysis, SAS/GLM for example, one may use the REML solutions for the residual variance component from a SAS/MIXED analysis. The values of the residual mean squares will differ but not to any great extent.

In the process of investigating model selection procedures, one should bear in mind the results of the classic Box and Cox (1964) paper. Their use of curve-fitting for model selection is insightful.

LITERATUR CITED

Box, G. E. P. and D. R. Cox (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B* 26:211-252.

Bozivich, H., T. A. Bancroft, and H. O. Hartley (1956). Power of analysis of variance test procedures for incompletely specified models. *Annals of Mathematical Statistics* 27:1017-1043.

Federer, W. T., J. Crossa, and J. Franco (1998). Forms of spatial analyses with mixed model effects and exploratory model selection. BU-1406-M in the Technical Report Series of the Department of Biometrics, Cornell University, Ithaca, NY 14853.

Federer, W. T. and C. S. Schlottfeldt (1954). The use of covariance to control gradients in experiments. *Biometrics* 10:282-290.