SAS PROC GLM and PROC MIXED

for Recovering Inter-Effect Information

Walter T. Federer*
Biometrics Unit
Cornell University
337 Warren Hall
Ithaca, NY 14853-7801
biometrics@cornell.edu

Russell D. Wolfinger
SAS Institute Inc. R-52
SAS Campus Drive
Cary, NC 27513
sasrdw@sas.com

BU-1330-M March 1996

* Liberty Hyde Bailey Professor Emeritus of Biological Statistics.

SAS PROC GLM and PROC MIXED

for Recovering Inter-Effect Information

Abstract

It appears that computer programming to recover various types of design effect information is not well understood. In light of the recent development of computer software for performing needed calculations and new statistical procedures, a description of SAS programs is given for experimenters who require such analyses. In the present paper, we describe programs for recovering interblock, interrow, intercolumn, intergradient, and interregression information from incomplete block and lattice rectangle-designed experiments.

1. Introduction

When analyzing the results from experiments designed as incomplete blocks or as lattice rectangles, the information obtained by recovery of interblock or interrow-intercolumn information should be utilized. Ignoring this type of information is akin to ignoring whole plot information in split plot designs. Further, it has been demonstrated that the expected mean squared error of means is smaller in the recovery analyses than in the standard intrablock or intrarow-intracolumn analyses. Also, the resulting means are less affected by random variation among the incomplete blocks or among rows and columns.

For certain types of spatial variation, standard textbook analyses may be inappropriate. An example is the differential gradient method within incomplete blocks (or rows) as described by Federer (1996). Other examples include row and column polynomial regressions for row-column designs and the same regressions applied to incomplete blocks in a rectangular array. These regressions vary from complete block to complete block and are considered to be random effects.

This paper shows how to program PROC GLM (SAS Institute Inc., 1989) and PROC MIXED (SAS Institute Inc., 1996) to perform the calculations for recovering the above types of information. Since these calculations are easily accomplished, experimenters can routinely perform them instead of using the less efficient intra-effect analyses.

2. Recovering Interblock Information

We utilize PROC GLM to obtain an ANOVA and intrablock means, and PROC MIXED to compute means adjusted for both intrablock and interblock information. We present a portion of the programs related to class and model. The first part including data, infile, and input is omitted here.

The first step is to construct a SAS data set, which for this example is named BLOCK. SAS data sets are rectangular arrays, with columns corresponding to variables and rows to observations. The data set named BLOCK is assumed to contain the following variables: Y (the response), T (the treatment), R (the complete block or replicate), B (the incomplete block within each complete block).

A PROC GLM program to obtain an ANOVA and intrablock treatment means is as follows:

```
proc glm data=block;
  class t r b;
  model y=t r b(r);
  random r b(r)
  lsmeans t;
run;
```

The PROC GLM statement invokes the procedure and the DATA=option specifies the analysis data set to be BLOCK. The CLASS statement lists which variables are to be treated as classification variables (as opposed to quantitative variables). Dummy indicator variables are created for each separate level of the classification variables.

The MODEL statement specifies the response variable Y and the fixed effects T, R, and B(R), the last one denoting the nesting of B within R. The RANDOM statement requests that R and B(R) be considered random effects in PROC GLM's construction of expected mean squares, although they are still considered as fixed effects in the ANOVA and in the construction of the adjusted means.

The LSMEANS statement requests least-squares means of the treatment effect T, which are the intrablock adjusted means.

Output from this analysis includes a standard ANOVA, Type I Tests (nested elimination of effects in the order in the model), Type III Tests (elimination of all other effects), expected mean squares for class random variables, and the estimated intrablock means.

To recover interblock information we use PROC MIXED as follows:

```
proc mixed data=block;
  class t r b;
  model y=t;
  random r b(r);
  lsmeans t;
run;
```

Note that only the fixed effects are retained in the MODEL statement. This analysis treats R and B(R) as random effects and estimates their variance components (along with that associated with the residual error) using restricted maximum likelihood (REML). The test for T as well as the printed least-squares means make use of the estimated variance components and thereby recover interblock information.

Analysts should be aware that standard textbook analyses use ANOVA rather than REML solutions for the variance components, and unless the REML and ANOVA solutions are equal, the adjusted treatment means will be different. However, the differences will usually be small. For a description of the output for this example and the ones to follow, see Federer (1995).

3. Recovery of Interrow and Intercolumn Information

The SAS data set for this example is assumed to be named ROW_COL and to contain the following variables: Y (the response), T(the treatment), R (the replicate), B (the row, nested within replicates), and C (the column, nested within replicates).

A PROC GLM program for obtaining an ANOVA and intrarow-intracolumn adjusted treatment means is as follows:

```
proc glm data=row_col
  class t r b c;
  model y=t r b(r) c(r);
  random r b(r) c(r);
  lsmeans t;
run;
```

This analysis extends the previous one by the addition of the column effect C(R). The output is the same as in the previous analysis, and the LSMEANS statement computes the intrarow-intercolumn adjusted treatment means.

A PROC MIXED program for obtaining treatment means adjusted for interrow and intercolumn information is obtained as follows:

```
proc mixed data=row_col;
  class t r b c;
  model y=t;
  random r b(r) c(r);
  lsmeans t;
run;
```

REML is used to compute the estimates for the variance components corresponding to R, B(R), C(R), and the residual. The LSMEANS statement automatically computes the interrow-intercolumn adjusted treatment means.

4. Recovery of Interblock and Intergradient Information

When gradients appear within incomplete blocks or within each row (column) of a lattice rectangle designed experiment, the above analyses will be inappropriate (Federer, 1996). If the gradients are considered to be random effects, then intergradient information as well as interblock (row) information will need to be recovered.

The nature of the gradients may take various forms. If they follow a polynomial regression format, then the polynomial regression effects should be included in the SAS data set along with the other design variables.

For this example we assume a linear form for the gradient. The data set GRADIENT is assumed to contain the following variables: Y (the response), R (the replicate), B (the block within replicate), and G (the gradient within B).

A PROC GLM program to compute the intrablock and intragradient analysis is as follows:

```
proc glm data=gradient;
  class r t b;
  model y=r t b(r) g g*r g*b(r);
  random r b(r);
  lsmeans t;
run;
```

Note that the gradient variable G is not a class variables, and it enters the model as a main effect and as an interaction with replications and blocks. (The symbol * is used to indicate interaction.) The particular model denoted by the above equation assumes that there is a single linear gradient (G) for the entire experiment, that there are differential linear gradients (G*R) within *each* complete block, and that there are differential regressions (G*B(R)) within *each* incomplete block. For model one of Federer (1996), omit the terms "g" and "g*r" from the MODEL statement, for the second model omit "g", and for the third model omit "g*r". The statement in MODEL above is not one of the models considered by Federer (1996). This presentation demonstrates flexibility in obtaining an appropriate analysis for an experiment. Although G*R and G*B(R) are considered to be random effects; they were not placed in the RANDOM statement because PROC GLM allows only classification effects there.

A PROC MIXED program for obtaining treatment means adjusted for interblock and intergradient information and using REML solutions for the variance components is obtained as follows:

```
proc mixed data=gradient;
  class r t b;
  model y=t g;
  random r b(r) g*r g*b(r);
  lsmeans t;
run;
```

Note that G*R and G*B(R) are now considered to be random effects and G a fixed effect.

5. Recovery of Interregression Information

If there are differential gradients among the rows and among the columns within complete blocks and interactions of these regressions, it will be necessary to obtain different analyses than those described previously. This analysis is an alternative to the standard textbook analyses for incomplete block and lattice rectangle designed experiments.

Suppose that linear and quadratic regressions for rows and columns, along with their interactions, are considered to be appropriate for controlling the particular type of spatial variation present in the experiment. Then linear and quadratic effects BL and BQ (for rows) and CL and CQ (for columns) should be present in the input SAS data set, along with same variables used in previous examples. The data set for this example is called REGRESS.

A PROC GLM program to obtain an ANOVA for the intraregression analysis and the intra-regression treatment means is as follows:

```
proc glm data=regress;
  class t r;
  model y=t r bl*r bq*r cl*r cq*r bl*cl*r bl*cq*r bq*cl*r bq*cq*r;
  random r;
  lsmeans t;
run;
```

The MODEL statement creates all of the linear and quadratic regression effects and their interactions. For this particular model, there are eight regression effects in each complete block, resulting in 8r degrees of freedom for the pooled regressions mean square, where r is the number of replications (equal to the number of levels of R). The resulting ANOVA has r degrees of freedom for each regression term sum of squares. If any of the regression terms are deemed inappropriate, they may be omitted in the MODEL statement or more may be added.

To recover interregression information, PROC MIXED is used. Appropriate code is as follows:

```
proc mixed data=regress;
  class t r;
  model y=t;
  random r bl*r bq*r cl*r cq*r bl*cl*r bl*cq*r bq*cl*r bq*cq*r;
  lsmeans t;
run;
```

Note all terms involving the replication effect R are considered to be random effects. With this many random effects in the model, it is quite possible that the REML estimate for one or more of them may be zero, which is effectively equivalent to dropping these terms from the model. In the lattice square example (Cochran and Cox, 1957, Table 12.5), the mean squares for CQ*R and BQ*CL*R were less than the residual mean square, and hence were eliminated from the model.

6. Comments

The preceding REML analyses rely upon the assumption that the data are normally distributed. To this end, an analysis may be desired for a transformation of the data such as log or square root. Such transformations are easily accomplished while constructing the input SAS data set. The need for such a transformation can often be determined by examination of residuals and/or empirical best linear unbiased predictions of the random effects.

Searle *et al.* (1992) state that REML and ANOVA solutions for variance components are equal for all balanced sets of data. By "balanced" they mean orthogonal. For the triple lattice example X1.3 of Federer (1955), which is a partially balanced incomplete block design, REML and ANOVA solutions are identical. For the balanced lattice square example of Table 12.5 in Cochran and Cox (1957), REML and ANOVA solutions are different.

Federer (1996) shows that quite striking results are sometimes possible when using other-than-textbook analyses of responses from experiments. Using the analysis in Section 5 on the data from the balanced lattice square designed experiment of Table 12.5 in Cochran and Cox (1957), the residual error mean square was essentially halved. Using the analysis of Section 4 on these data resulted in a decrease of 16% in the residual error mean square, or the equivalent of an additional replicate. This demonstrates that models appropriate to the experimental situation should be used rather than using standard textbook analyses all the time.

We conclude with a few comments regarding the use of PROC MIXED. First, when all effects in the RANDOM statement of PROC MIXED share one or more effects in common, it is often more efficient computationally to "factor out" this common effect into the optional SUBJECT=effect. For example, the RANDOM statement from Section 4

```
random r b(r) g*r g*b(r);
```

can also be written as

```
random int b g g*b/subject=r;
```

The gain in computational efficiency grows with the number of levels of R.

When performing inference, it is often useful to construct single degree of freedom contrasts among the adjusted means. Simple differences of the means are automatically generated using the DIFF option in the LSMEANS statement of PROC MIXED, and multiple comparison adjustments accounting for simultaneous inference are available as well (refer to SAS Institute, Inc. 1996). Custom contrasts can be constructed with CONTRAST and ESTIMATE statements as demonstrated in Federer (1995) and in SAS/STAT Inst. Inc. (1959).

7. References

Cochran, W. G. and G. M. Cox (1957). *Experimental Designs*, 2nd edition. John Wiley & Sons, Inc., New York.

Federer, W. T. (1955). *Experimental Design--Theory and Application*. Macmillan Company, New York (Republished by Oxford & IBH Publishing Company, New Delhi, 1967 and 1974.)

Federer, W. T. (1995). SAS for ANOVAs and recovery of interblock, intercolumn, and intergradient information. BU-1295 in the Technical Report Series of the Biometrics Unit, Cornell University, June (Price $5).

Federer, W. T. (1996). Recovery of interblock, intergradient, and intervariety information in incomplete block and lattice rectangle designed experiments. *Biometrics* (in press).

SAS/STAT Institute Inc. (1989). SAS/STAT Users Guide, Version 6, 4th edition, Volume 2. SAS Institute Inc., Cary, NC.

SAS Institute Inc. (1996). SAS/STAT Software: Changes and Enhancements through Release 6.11. SAS Institute Inc., Cary, NC.

Searle, S. R., G. Casella, and C. E. McCulloch (1992). *Variance Components*. John Wiley & Sons, Inc., New York, Chichester, Brisbane, Toronto, Singapore.