NAG Toolbox for MATLAB

g04ca

1 Purpose

g04ca computes an analysis of variance table and treatment means for a complete factorial design.

2 Syntax

```
[table, itotal, tmean, e, imean, semean, bmean, r, ifail] = g04ca(y,
lfac, nblock, inter, irdf, mterm, maxt, 'n', n, 'nfac', nfac)
```

3 Description

An experiment consists of a collection of units, or plots, to which a number of treatments are applied. In a factorial experiment the effects of several different sets of conditions are compared, e.g., three different temperatures, T_1, T_2 and T_3, and two different pressures, P_1 and P_2. The conditions are known as factors and the different values the conditions take are known as levels. In a factorial experiment the experimental treatments are the combinations of all the different levels of all factors, e.g.,

$$T_1P_1, \quad T_2P_1, \quad T_3P_1$$

$$T_1P_2, \quad T_2P_2, \quad T_3P_2$$

The effect of a factor averaged over all other factors is known as a main effect, and the effect of a combination of some of the factors averaged over all other factors is known as an interaction. This can be represented by a linear model. In the above example if the response was y_{ijk} for the kth replicate of the ith level of T and the jth level of P the linear model would be

$$y_{ijk} = \mu + t_i + p_j + \gamma_{ij} + e_{ijk}$$

where μ is the overall mean, t_i is the main effect of T, p_j is the main effect of P, γ_{ij} is the $T \times P$ interaction and e_{ijk} is the random error term. In order to find unique estimates constraints are placed on the parameters estimates. For the example here these are:

$$\sum_{i=1}^{3} \hat{t}_i = 0,$$
$$\sum_{j=1}^{2} \hat{p}_j = 0,$$
$$\sum_{i=1}^{3} \hat{\gamma}_{ij} = 0, \quad \text{for } j = 1, 2 \text{ and}$$
$$\sum_{j=1}^{2} \hat{\gamma}_{ij} = 0, \quad \text{for } i = 1, 2, 3,$$

where ˆ denotes the estimate.

If there is variation in the experimental conditions (e.g., in an experiment on the production of a material different batches of raw material may be used, or the experiment may be carried out on different days), then plots that are similar are grouped together into blocks. For a balanced complete factorial experiment all the treatment combinations occur the same number of times in each block.

g04ca computes the analysis of variance (ANOVA) table by sequentially computing the totals and means for an effect from the residuals computed when previous effects have been removed. The effect sum of squares is the sum of squared totals divided by the number of observations per total. The means are then subtracted from the residuals to compute a new set of residuals. At the same time the means for the original data are computed. When all effects are removed the residual sum of squares is computed from the residuals. Given the sums of squares an ANOVAtable is then computed along with standard errors for the difference in treatment means.

The data for g04ca has to be in standard order given by the order of the factors. Let there be k factors, $f_1, f_2, \ldots, f_k$ in that order with levels $l_1, l_2, \ldots, l_k$ respectively. Standard order requires the levels of factor f_1 are in order $1, 2, \ldots, l_1$ and within each level of f_1 the levels of f_2 are in order $1, 2, \ldots, l_2$ and so on.

For an experiment with blocks the data is for block 1 then for block 2, etc. Within each block the data must be arranged so that the levels of factor f_1 are in order $1, 2, \ldots, l_1$ and within each level of f_1 the levels of f_2 are in order $1, 2, \ldots, l_2$ and so on. Any within block replication of treatment combinations must occur within the levels of f_k.

The ANOVAtable is given in the following order. For a complete factorial experiment the first row is for blocks, if present, then the main effects of the factors in their order, e.g., f_1 followed by f_2 etc. These are then followed by all the two factor interactions then all the three factor interactions etc., the last two rows being for the residual and total sums of squares. The interactions are arranged in lexical order for the given factor order. For example, for the three factor interactions for a five factor experiment the 10 interactions would be in the following order:

$$f_1 f_2 f_3$$
$$f_1 f_2 f_4$$
$$f_1 f_2 f_5$$
$$f_1 f_3 f_4$$
$$f_1 f_3 f_5$$
$$f_1 f_4 f_5$$
$$f_2 f_3 f_4$$
$$f_2 f_3 f_5$$
$$f_2 f_4 f_5$$
$$f_3 f_4 f_5$$

4 References

Cochran W G and Cox G M 1957 *Experimental Designs* Wiley

Davis O L 1978 *The Design and Analysis of Industrial Experiments* Longman

John J A and Quenouille M H 1977 *Experiments: Design and Analysis* Griffin

5 Parameters

5.1 Compulsory Input Parameters

1: **y**(**n**) – **double array**

The observations in standard order, see Section 3.

2: **lfac**(**nfac**) – **int32 array**

lfac(i) must contain the number of levels for the ith factor, for $i = 1, 2, \ldots, k$.

Constraint: **lfac**$(i) \geq 2$, for $i = 1, 2, \ldots, k$.

3: **nblock** – **int32 scalar**

The number of blocks. If there are no blocks, set **nblock** $= 0$ or 1.

Constraints:

nblock ≥ 0;
if **nblock** ≥ 2, **n**/**nblock** must be a multiple of the number of treatment combinations, that is a multiple of $\prod_{i=1}^{k} \mathbf{lfac}(i)$.

4: **inter – int32 scalar**

The maximum number of factors in an interaction term. If no interaction terms are to be computed, set **inter** $= 0$ or 1.

Constraint: $0 \leq$ **inter** $\leq$ **nfac**.

5: **irdf – int32 scalar**

The adjustment to the residual and total degrees of freedom. The total degrees of freedom are set to **n** $-$ **irdf** and the residual degrees of freedom adjusted accordingly. For examples of the use of **irdf** see Section 8.

Constraint: **irdf** ≥ 0.

6: **mterm – int32 scalar**

the maximum number of terms in the analysis of variance table, see Section 8.

Constraint: **mterm** must be large enough to contain the terms specified by **nfac**, **inter** and **nblock**. If the function exits with **ifail** ≥ 2, the required minimum value of **mterm** is returned in **itotal**.

7: **maxt – int32 scalar**

the maximum number of treatment means to be computed, see Section 8. If the value of **maxt** is too small for the required analysis then the minimum number is returned in **imean**(1).

Constraint: **maxt** must be large enough for the number of means specified by **lfac** and **inter**; if **inter** $=$ **nfac** then **maxt** $\geq \prod_{i=1}^{k}(\mathbf{lfac}(i)+1)-1$.

5.2 Optional Input Parameters

1: **n – int32 scalar**

Default: The dimension of the arrays **y**, **r**. (An error is raised if these dimensions are not equal.)

the number of observations.

Constraints:

- **n** ≥ 4;
- if **nblock** > 1, **n** must be a multiple of **nblock**;
- **n** must be a multiple of the number of treatment combinations, that is a multiple of $\prod_{i=1}^{k}\mathbf{lfac}(i)$.

2: **nfac – int32 scalar**

Default: The dimension of the array **lfac**.

k, the number of factors.

Constraint: **nfac** ≥ 1.

5.3 Input Parameters Omitted from the MATLAB Interface

iwk

5.4 Output Parameters

1: **table(mterm,5) – double array**

The first **itotal** rows of **table** contain the analysis of variance table. The first column contains the degrees of freedom, the second column contains the sum of squares, the third column (except for the row corresponding to the total sum of squares) contains the mean squares, i.e., the sums of squares divided by the degrees of freedom, and the fourth and fifth columns contain the F ratio and significance level, respectively (except for rows corresponding to the total sum of squares, and the residual sum of squares). All other cells of the table are set to zero.

The first row corresponds to the blocks and is set to zero if there are no blocks. The **itotal**th row corresponds to the total sum of squares for **y** and the $(\textbf{itotal} - 1)$th row corresponds to the residual sum of squares. The central rows of the table correspond to the main effects followed by the interaction if specified by **inter**. The main effects are in the order specified by **lfac** and the interactions are in lexical order, see Section 3.

2: **itotal – int32 scalar**

The row in **table** corresponding to the total sum of squares. The number of treatment effects is $\textbf{itotal} - 3$.

3: **tmean(maxt) – double array**

The treatment means. The position of the means for an effect is given by the index in **imean**. For a given effect the means are in standard order, see Section 3.

4: **e(maxt) – double array**

The estimated effects in the same order as for the means in **tmean**.

5: **imean(mterm) – int32 array**

Indicates the position of the effect means in **tmean**. The effect means corresponding to the first treatment effect in the ANOVAtable are stored in $\textbf{tmean}(1)$ up to $\textbf{tmean}(\textbf{imean}(1))$. Other effect means corresponding to the ith treatment effect, $i = 1, 2, \ldots, \textbf{itotal} - 3$, are stored in $\textbf{tmean}(\textbf{imean}(i - 1) + 1)$ up to $\textbf{tmean}(\textbf{imean}(i))$.

6: **semean(mterm) – double array**

The standard error of the difference between means corresponding to the ith treatment effect in the ANOVAtable.

7: **bmean(nblock + 1) – double array**

$\textbf{bmean}(1)$ contains the grand mean, if $\textbf{nblock} > 1$, $\textbf{bmean}(2)$ up to $\textbf{bmean}(\textbf{nblock} + 1)$ contain the block means.

8: **r(n) – double array**

The residuals.

9: **ifail – int32 scalar**

0 unless the function detects an error (see Section 6).

6 Error Indicators and Warnings

Errors or warnings detected by the function:

$\textbf{ifail} = 1$

On entry, $\textbf{n} < 4$,
or $\textbf{nfac} < 1$,
or $\textbf{nblock} < 0$,
or $\textbf{inter} < 0$,
or $\textbf{inter} > \textbf{nfac}$,
or $\textbf{irdf} < 0$.

ifail = 2

On entry, **lfac**(i) ≤ 1, for some $i = 1, 2, \ldots,$ **nfac**,
or the value of **maxt** is too small,
or the value of **mterm** is too small,
or **nblock** > 1 and **n** is not a multiple of **nblock**,
or the number of plots per block is not a multiple of the number of treatment combinations.

ifail = 3

On entry, the values of **y** are constant.

ifail = 4

There are no degrees of freedom for the residual or the residual sum of squares is zero. In either case the standard errors and F-statistics cannot be computed.

7 Accuracy

The block and treatment sums of squares are computed from the block and treatment residual totals. The residuals are updated as each effect is computed and the residual sum of squares computed directly from the residuals. This avoids any loss of accuracy in subtracting sums of squares.

8 Further Comments

The number of rows in the ANOVAtable and the number of treatment means are given by the following formulae.

Let there be k factors with levels l_i for $i = 1, 2, \ldots, k$. and let t be the maximum number of terms in an interaction then the number of rows in the ANOVAtables is:

$$\sum_{i=1}^{t} \binom{k}{i} + 3.$$

The number of treatment means is:

$$\sum_{i=1}^{t} \prod_{j \in S_i} l_j,$$

where S_i is the set of all combinations of the k factors i at a time.

To estimate missing values the Healy and Westmacott procedure or its derivatives may be used, see John and Quenouille 1977. This is an iterative procedure in which estimates of the missing values are adjusted by subtracting the corresponding values of the residuals. The new estimates are then used in the analysis of variance. This process is repeated until convergence. A suitable initial value may be the grand mean. When using this procedure **irdf** should be set to the number of missing values plus one to obtain the correct degrees of freedom for the residual sum of squares.

For analysis of covariance the residuals are obtained from an analysis of variance of both the response variable and the covariates. The residuals from the response variable are then regressed on the residuals from the covariates using, say, g02cb or g02da. The coefficients obtained from the regression can be examined for significance and used to produce an adjusted dependent variable using the original response variable and covariate. An approximate adjusted analysis of variance table can then be produced by using the adjusted dependent variable. In this case **irdf** should be set to one plus the number of fitted covariates.

For designs such as Latin squares one more of the blocking factors has to be removed in a preliminary analysis before the final analysis. This preliminary analysis can be performed using g04bb or a prior call to g04ca if the data is reordered between calls. The residuals from the preliminary analysis are then input to g04ca. In these cases **irdf** should be set to the difference between **n** and the residual degrees of freedom from preliminary analysis. Care should be taken when using this approach as there is no check on the orthogonality of the two analyses.

9 Example

```
y = [274;
     361;
     253;
     325;
     317;
     339;
     326;
     402;
     336;
     379;
     345;
     361;
     352;
     334;
     318;
     339;
     393;
     358;
     350;
     340;
     203;
     397;
     356;
     298;
     382;
     376;
     355;
     418;
     387;
     379;
     432;
     339;
     293;
     322;
     417;
     342;
     82;
     297;
     133;
     306;
     352;
     361;
     220;
     333;
     270;
     388;
     379;
     274;
     336;
     307;
     266;
     389;
     333;
     353];
lfac = [int32(6);
     int32(3)];
nblock = int32(3);
inter = int32(2);
irdf = int32(0);
mterm = int32(6);
maxt = int32(27);
[table, itotal, tmean, e, imean, semean, bmean, r, ifail] = ...
    g04ca(y, lfac, nblock, inter, irdf, mterm, maxt)
```

```
table =
   1.0e+05 *
```

```
    0.0000    0.3012    0.1506    0.0001    0.0000
    0.0001    0.7301    0.1460    0.0001    0.0000
    0.0000    0.2160    0.1080    0.0001    0.0000
    0.0001    0.3119    0.0312    0.0000    0.0000
    0.0003    0.6663    0.0196         0         0
    0.0005    2.2254         0         0         0
itotal =
           6
tmean =
  254.7778
  339.0000
  333.3333
  367.7778
  330.7778
  360.6667
  334.2778
  353.7778
  305.1111
  235.3333
  332.6667
  196.3333
  342.6667
  341.6667
  332.6667
  309.3333
  370.3333
  320.3333
  395.0000
  370.3333
  338.0000
  373.3333
  326.6667
  292.3333
  350.0000
  381.0000
  351.0000
e =
  -76.2778
    7.9444
    2.2778
   36.7222
   -0.2778
   29.6111
    3.2222
   22.7222
  -25.9444
  -22.6667
   55.1667
  -32.5000
    0.4444
  -20.0556
   19.6111
  -27.2222
   14.2778
   12.9444
   24.0000
  -20.1667
   -3.8333
   39.3333
  -26.8333
  -12.5000
  -13.8889
   -2.3889
   16.2778
imean =
           6
           9
          27
           0
           0
```

```
           0
semean =
   20.8681
   14.7560
   36.1446
         0
         0
         0
bmean =
  331.0556
  339.5556
  354.7778
  298.8333
r =
     array elided
ifail =
           0
```