nag_anova_factorial (g04cac)

1. Purpose

nag_anova_factorial (g04cac) computes an analysis of variance table and treatment means for a complete factorial design.

2. Specification

```
#include <nag.h>
#include <nagg04.h>

void nag_anova_factorial(Integer n, double y[], Integer nfac, Integer lfac[],
         Integer nblock, Integer inter, Integer irdf, Integer *mterm,
         double **table, double **tmean, Integer *maxt, double **e,
         Integer **imean, double **semean, double **bmean, double r[],
         NagError *fail)
```

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.,

$$\begin{array}{lll} T_1P_1 & T_2P_1 & T_3P_1 \\ T_1P_2 & T_2P_2 & T_3P_2 \end{array}$$

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 parameter estimates. For the example here these are:

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

where $\hat{}$ 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.

nag_anova_factorial 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 ANOVA table is then computed along with standard errors for the difference in treatment means.

The data for nag_anova_factorial 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 ANOVA table 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. Parameters

n

Input: the number of observations.

Constraints:

n ≥ 4.

n must be a multiple of **nblock** if **nblock** > 1.

n must be a multiple of the number of treatment combinations, that is a multiple of $\prod_{i=1}^{k}$ **lfac**$[i-1]$.

y[n]

Input: the number of observations in standard order, see Section 3.

nfac

Input: the number of factors, k.
Constraint: **nfac** ≥ 1.

lfac[nfac]

Input: **lfac**$[i-1]$ must contain the number of levels for the ith factor, $i = 1, 2, \ldots, k$.

Constraint: **lfac**$[i-1] \geq 2$ for $i = 1, 2, \ldots, k$.

nblock

Input: 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 $\prod_{i=1}^{k}$ **lfac**$[i-1]$.

inter

Input: 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**.

irdf

Input: 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 6.

Constraint: **irdf** ≥ 0.

mterm

Output: the number of terms in the analysis of variance table, see Section 6.

The number of treatment effects is **mterm** − 3.

table

Output: A pointer which points to **mterm** × 5 memory locations, allocated internally. Viewing this memory as a two dimensional **mterm** × 5 array, the first **mterm** 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 **mterm**th row corresponds to the total sum of squares for **y** and the (**mterm**−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.

tmean

Output: A pointer pointing to **maxt** memory locations, allocated internally. It contains 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.

maxt

Output: the number of treatment means that have been computed, see Section 6.

e

Output: a pointer pointing to **maxt** memory locations, allocated internally. It contains the estimated effects in the same order as for the means in **tmean**.

imean

Output: a pointer pointing to **mterm** memory locations, allocated internally. It indicates the position of the effect means in **tmean**. The effect means corresponding to the first treatment effect in the ANOVA table are stored in **tmean**[0] up to **tmean**[**imean**[0]−1]. Other effect means corresponding to the ith treatment effect, $i = 2, 3, \ldots$,**mterm**−3, are stored in **tmean**[**imean**[$i-2$]] up to **tmean**[**imean**[$i-1$] − 1].

semean

Output: a pointer pointing to **mterm** memory locations, allocated internally. It contains the standard error of the difference between means corresponding to the ith treatment effect in the ANOVA table.

bmean

Output: A pointer pointing to **nblock** + 1 memory locations, allocated internally. **bmean**[0] contains the grand mean, if **nblock** > 1, **bmean**[1] up to **bmean**[**nblock**] contain the block means.

r[n]

Output: the residuals.

fail

The NAG error parameter, see the Essential Introduction to the NAG C Library.

Note: If nag_anova_factorial is to be called repeatedly then the memory allocated to **table**, **tmean**, **e**, **imean**, **semean**, and **bmean** must be freed between calls. Users are advised to call nag_anova_factorial_free (g04czc) to achieve this.

5. Error Indications and Warnings

NE_INT_ARG_LT
On entry, **n** must not be less than 4: **n** = ⟨*value*⟩.
On entry, **nfac** must not be less than 1: **nfac** = ⟨*value*⟩.
On entry, **nblock** must not be less than 0: **nblock** = ⟨*value*⟩.
On entry, **inter** must not be less than 0: **inter** = ⟨*value*⟩.
On entry, **irdf** must not be less than 0: **irdf** = ⟨*value*⟩.

NE_2_INT_ARG_GT
On entry, **inter** = ⟨*value*⟩ while **nfac** = ⟨*value*⟩.
These parameters must satisfy **inter** ≤ **nfac**.

NE_INTARR
On entry, **lfac**[⟨*value*⟩] = ⟨*value*⟩.
Constraint: **lfac**[$i-1$] ≥ 2 for $i = 1, 2, \ldots,$**nfac**.

NE_INT_2
On entry, **nblock** = ⟨*value*⟩, **n** = ⟨*value*⟩.
Constraint: **n** must be a multiple of **nblock**, when **nblock** > 1.

NE_PLOT_TREAT
The number of plots per block is not a multiple of the number of treatment combinations.

NE_ARRAY_CONSTANT
On entry, the elements of the array **y** are constant.

NE_G04CA_RES_DF
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.

NE_ALLOC_FAIL
Memory allocation failed.

6. Further Comments

The number of rows in the ANOVA table 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 ANOVA table 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, nag_regress_confid_interval (g02cbc) or nag_regsn_mult_linear (g02dac). 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 nag_anova_random (g04bbc) or a prior call to nag_anova_factorial if the data is reordered between calls. The residuals from the preliminary analysis are then input to nag_anova_factorial. 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.

6.1. 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.

6.2. References

Cochran W G and Cox G M (1957) *Experimental Designs* Wiley.
Davis O L (ed.) (1978) *The Design and Analysis of Industrial Experiments* Longman.
John J A and Quenouille M H (1977) *Experiments: Design and Analysis* Griffin.

7. See Also

nag_regress_confid_interval (g02cbc)
nag_regsn_mult_linear (g02dac)
nag_anova_factorial_free (g04czc)

8. Example

The data, given by John and Quenouille (1977), is for the yield of turnips for a factorial experiment with two factors, the amount of phosphate with 6 levels and the amount of liming with 3 levels. The design was replicated in 3 blocks. The data is input and the analysis of variance computed. The analysis of variance table and tables of means with their standard errors are printed.

8.1. Program Text

```
/* nag_anova_factorial(g04cac) Example Program.
 *
 * Copyright 1998 Numerical Algorithms Group.
 *
 * Mark 5, 1998.
 *
 */
#include <nag.h>
#include <nag_stdlib.h>
#include <stdio.h>
#include <nagg04.h>

#define NMAX 54
#define MAXF 2
#define MAXT 27
#define MTERM 6
#define BMAX 4
#define LDT MTERM

main()
{

  double  r[NMAX], y[NMAX];
  double *bmean=0, *e=0, *semean=0,
  *table=0, *tmean=0;

  Integer c__27 = 27;
  Integer lfac[MAXF];
```

```
  Integer *imean=0;
  Integer mterm = MTERM;
  Integer nfac, irdf;
  Integer i, j, k, l, n;
  Integer num;
  Integer inter, nblock;
  Integer itotal, ntreat;

#define LFAC(I) lfac[(I)-1]
#define IWK(I) iwk[(I)-1]
#define IMEAN(I) imean[(I)-1]
#define Y(I) y[(I)-1]
#define TMEAN(I) tmean[(I)-1]
#define SEMEAN(I) semean[(I)-1]
#define R(I) r[(I)-1]
#define E(I) e[(I)-1]
#define BMEAN(I) bmean[(I)-1]
#define TABLE(I,J) table[((I)-1) * ( 5) + ((J)-1)]

  Vprintf("g04cac Example Program Results\n\n");

  /*  Skip heading in data file */
  Vscanf("%*[^\n]");

  Vscanf("%ld%ld%ld%ld%*[^\n]",&n, &nblock, &nfac, &inter);

  if (n <= NMAX && nblock <= BMAX - 1 && nfac <= MAXF)
    {
      for (j = 0; j < nfac; ++j)
        Vscanf("%ld",&lfac[j]);
      Vscanf("%*[^\n]");

      for (i = 0; i < n; ++i)
        Vscanf("%lf",&y[i]);
      Vscanf("%*[^\n]");

      irdf = 0;
      g04cac(n, y, nfac, lfac, nblock, inter, irdf, &mterm, &table,
             &tmean, &c__27, &e, &imean, &semean, &bmean, r,
             NAGERR_DEFAULT);

      itotal = mterm;
      Vprintf("\n  ANOVA table\n\n");
      Vprintf("   Source      df         SS          MS          F\
          Prob\n\n");
      k = 0;
      if (nblock > 1)
        {
          ++k;
          Vprintf("%s   "," Blocks    ");
          for (j = 1; j <= 5; ++j)
            Vprintf("%4.2f  ",TABLE(1,j));
         Vprintf("\n");
        }
      ntreat = mterm - 2 - k;
      for (i = 1; i <= ntreat; ++i)
        {
          Vprintf("%s%2i   "," Effect  ",i);
          for (j = 1; j <= 5; ++j)
            Vprintf("%4.2f  ",TABLE(k+i,j));
          Vprintf("\n");
        }
      Vprintf("%s   "," Residual  ");
      for (j = 1; j <= 3; ++j)
        Vprintf("%4.2f  ",TABLE(mterm-1,j));
      Vprintf("\n");

      Vprintf("%s   "," Total     ");
      for (j = 1; j <= 2; ++j)
        Vprintf("%4.2f  ",TABLE(mterm,j));
```

```
      Vprintf("\n");

      Vprintf("\n");
      Vprintf("  Treatment Means and Standard Errors ");
      Vprintf("\n");
      Vprintf("\n");
      k = 0;
      for (i = 0; i < ntreat; ++i)
        {
          l = imean[i];
          Vprintf("%s%2i"," Effect ",i+1);
          Vprintf("\n");

          Vprintf("\n");
          num=1;
          for (j = k; j < l; ++j)
            {
              Vprintf("%10.2f%s",tmean[j],num%8?"":"\n");
              num++;
            }
          Vprintf("\n");

          Vprintf("\n%s%10.2f\n\n"," SE of difference in means  = ",semean[i]);
          k = l;
        }
      g04czc(&table, &tmean, &e, &imean, &semean, &bmean);
      exit(EXIT_SUCCESS);
    }
  else
    {
      Vprintf( "Incorrect input value of n or nblock or nfac.\n");
      exit(EXIT_FAILURE);
    }
}
```

8.2. Program Data

```
g04cac Example Program Data
54 3 2 2  : n nblock nfac inter
6 3       : lfac

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
```

8.3. Program Results

```
g04cac Example Program Results


  ANOVA table

   Source       df       SS         MS      F   Prob

 Blocks         2.00   30118.78  15059.39  7.68  0.00
 Effect   1     5.00   73008.17  14601.63  7.45  0.00
 Effect   2     2.00   21596.33  10798.17  5.51  0.01
 Effect   3    10.00   31191.67   3119.17  1.59  0.15
 Residual      34.00   66627.89   1959.64
 Total         53.00  222542.83

  Treatment Means and Standard Errors

 Effect  1

    254.78    339.00    333.33    367.78    330.78    360.67

 SE of difference in means  =      20.87
```

```
Effect  2

   334.28    353.78    305.11

SE of difference in means  =      14.76

Effect  3

   235.33    332.67    196.33    342.67    341.67    332.67    309.33    370.33
   320.33    395.00    370.33    338.00    373.33    326.67    292.33    350.00
   381.00    351.00

SE of difference in means  =      36.14
```