NAG Toolbox

nag_anova_dummyvars (g04ea)

1 Purpose

nag_anova_dummyvars (g04ea) computes orthogonal polynomial or dummy variables for a factor or classification variable.

2 Syntax

```
[x, rep, ifail] = nag_anova_dummyvars(typ, levels, ifact, v, 'n', n)
[x, rep, ifail] = g04ea(typ, levels, ifact, v, 'n', n)
```

3 Description

In the analysis of an experimental design using a general linear model the factors or classification variables that specify the design have to be coded as dummy variables. nag_anova_dummyvars (g04ea) computes dummy variables that can then be used in the fitting of the general linear model using nag_correg_linregm_fit (g02da).

If the factor of length n has k levels then the simplest representation is to define k dummy variables, X_j such that $X_j = 1$ if the factor is at level j and 0 otherwise for $j = 1, 2, \ldots, k$. However, there is usually a mean included in the model and the sum of the dummy variables will be aliased with the mean. To avoid the extra redundant argument $k - 1$ dummy variables can be defined as the contrasts between one level of the factor, the reference level, and the remaining levels. If the reference level is the first level then the dummy variables can be defined as $X_j = 1$ if the factor is at level j and 0 otherwise, for $j = 2, 3, \ldots, k$. Alternatively, the last level can be used as the reference level.

A second way of defining the $k - 1$ dummy variables is to use a Helmert matrix in which levels $2, 3, \ldots, k$ are compared with the average effect of the previous levels. For example if $k = 4$ then the contrasts would be:

$$\begin{array}{cccc} 1 & -1 & -1 & -1 \\ 2 & 1 & -1 & -1 \\ 3 & 0 & 2 & -1 \\ 4 & 0 & 0 & 3 \end{array}$$

Thus variable j, for $j = 1, 2, \ldots, k - 1$ is given by

$X_j = -1$ if factor is at level less than $j + 1$

$X_j = \sum_{i=1}^{j} r_i / r_{j+1}$ if factor is at level $j + 1$

$X_j = 0$ if factor is at level greater than $j + 1$

where r_j is the number of replicates of level j.

If the factor can be considered as a set of values from an underlying continuous variable then the factor can be represented by a set of $k - 1$ orthogonal polynomials representing the linear, quadratic etc. effects of the underlying variable. The orthogonal polynomial is computed using Forsythe's algorithm (Forsythe (1957), see also Cooper (1968)). The values of the underlying continuous variable represented by the factor levels have to be supplied to the function.

The orthogonal polynomials are standardized so that the sum of squares for each dummy variable is one. For the other methods integer (± 1) representations are retained except that in the Helmert representation the code of level $j + 1$ in dummy variable j will be a fraction.

4 References

Cooper B E (1968) Algorithm AS 10. The use of orthogonal polynomials *Appl. Statist.* **17** 283–287

Forsythe G E (1957) Generation and use of orthogonal polynomials for data fitting with a digital computer *J. Soc. Indust. Appl. Math.* **5** 74–88

5 Parameters

5.1 Compulsory Input Parameters

1: **typ** – CHARACTER(1)

The type of dummy variable to be computed.

If **typ** = 'P', an orthogonal Polynomial representation is computed.

If **typ** = 'H', a Helmert matrix representation is computed.

If **typ** = 'F', the contrasts relative to the First level are computed.

If **typ** = 'L', the contrasts relative to the Last level are computed.

If **typ** = 'C', a Complete set of dummy variables is computed.

Constraint: **typ** = 'P', 'H', 'F', 'L' or 'C'.

2: **levels** – INTEGER

k, the number of levels of the factor.

Constraint: **levels** ≥ 2.

3: **ifact**(**n**) – INTEGER array

The n values of the factor.

Constraint: $1 \leq$ **ifact**$(i) \leq$ **levels**, for $i = 1, 2, \ldots, n$.

4: **v**(:) – REAL (KIND=nag_wp) array

The dimension of the array **v** must be at least **levels** if **typ** = 'P', and at least 1 otherwise

If **typ** = 'P', the k distinct values of the underlying variable for which the orthogonal polynomial is to be computed.

If **typ** $\neq$ 'P', **v** is not referenced.

Constraint: if **typ** = 'P', the k values of **v** must be distinct.

5.2 Optional Input Parameters

1: **n** – INTEGER

Default: the dimension of the array **ifact**.

n, the number of observations for which the dummy variables are to be computed.

Constraint: **n** $\geq$ **levels**.

5.3 Output Parameters

1: **x**$(ldx, :)$ – REAL (KIND=nag_wp) array

The first dimension of the array **x** will be **n**.

The second dimension of the array **x** will be **levels** $- 1$ if **typ** = 'P', 'H', 'F' or 'L' and at least **levels** if **typ** = 'C'.

The n by k^* matrix of dummy variables, where $k^* = k - 1$ if **typ** = 'P', 'H', 'F' or 'L' and $k^* = k$ if **typ** = 'C'.

2: **rep**(**levels**) – REAL (KIND=nag_wp) array

The number of replications for each level of the factor, r_i, for $i = 1, 2, \ldots, k$.

3: **ifail** – INTEGER

ifail $= 0$ unless the function detects an error (see Section 5).

6 Error Indicators and Warnings

Errors or warnings detected by the function:

ifail $= 1$

On entry, **levels** < 2,
or **n** $<$ **levels**,
or $ldx <$ **n**,
or **typ** $\neq$ 'P', 'H', 'F', 'L' or 'C'.

ifail $= 2$

On entry, a value of **ifact** is not in the range $1 \leq$ **ifact**$(i) \leq$ **levels**, for $i = 1, 2, \ldots, n$,
or **typ** = 'P' and not all values of **v** are distinct,
or not all levels are represented in **ifact**.

ifail $= 3$ (*warning*)

An orthogonal polynomial has all values zero. This will be due to some values of **v** being very close together. Note this can only occur if **typ** = 'P'.

ifail $= -99$

An unexpected error has been triggered by this routine. Please contact NAG.

ifail $= -399$

Your licence key may have expired or may not have been installed correctly.

ifail $= -999$

Dynamic memory allocation failed.

7 Accuracy

The computations are stable.

8 Further Comments

Other functions for fitting polynomials can be found in Chapter E02.

9 Example

Data are read in from an experiment with four treatments and three observations per treatment with the treatment coded as a factor. nag_anova_dummyvars (g04ea) is used to compute the required dummy variables and the model is then fitted by nag_correg_linregm_fit (g02da).

9.1 Program Text

```
     function g04ea_example

fprintf('g04ea example results\n\n');

typ    = 'C';
n1 = nag_int(1);
levels = 4*n1;
ifact  = [ n1;    4;    2;    3;    4;    2;
             4;    1;    3;    1;    3;    2];
y      = [ 33.63 39.62 38.18 41.46 38.02 35.83 ...
           35.99 36.58 42.92 37.80 40.43 37.89];

% Calculate dummy variables
v      = [0];
[x, rep, ifail] = g04ea( ...
                          typ, levels, ifact, v);

m      = levels;
isx    = ones(m,1,nag_int_name);
mean_p = 'M';
ip     = nag_int(m+1);

% Fit general linear regression model
[rss, idf, b, se, covar, res, h, q, svd, irank, p, wk, ifail] = ...
  g02da(mean_p, x, isx, ip, y);

% Display results
if svd
  fprintf('Model not of full rank, rank = %4d\n\n', irank);
end
fprintf('Residual sum of squares = %12.3e\n', rss);
fprintf('Degrees of freedom      = %4d\n', idf);
fprintf('\nVariable   Parameter estimate   Standard error\n\n');
ivar = double([1:ip]');
fprintf('%6d%20.4e%17.4e\n',[ivar b se]');
```

9.2 Program Results

```
     g04ea example results

Model not of full rank, rank =    4

Residual sum of squares =    2.223e+01
Degrees of freedom      =    8

Variable   Parameter estimate   Standard error

     1          3.0557e+01       3.8494e-01
     2          5.4467e+00       8.3896e-01
     3          6.7433e+00       8.3896e-01
     4          1.1047e+01       8.3896e-01
     5          7.3200e+00       8.3896e-01
```