NAG Toolbox for MATLAB

g02jb

1 Purpose

g02jb fits a linear mixed effects regression model using maximum likelihood (ML).

2 Syntax

```
[nff, nrf, df, ml, b, se, gamma, warn, ifail] = g02jb(n, nvpr, levels,
yvid, fvid, rvid, svid, cwid, vpr, dat, fint, rint, lb, gamma, maxit,
tol, 'ncol', ncol, 'nfv', nfv, 'nrv', nrv)
```

3 Description

g02jb fits a model of the form:

$$y = X\beta + Z\nu + \epsilon$$

where y is a vector of n observations on the dependent variable,

X is a known n by p design matrix for the fixed independent variables,

β is a vector of length p of unknown *fixed effects*,

Z is a known n by q design matrix for the random independent variables,

ν is a vector of length q of unknown *random effects*;

and ϵ is a vector of length n of unknown random errors.

Both ν and ϵ are assumed to have a Gaussian distribution with expectation zero and

$$\mathrm{Var}\begin{bmatrix}\nu\\ \epsilon\end{bmatrix} = \begin{bmatrix}G & 0\\ 0 & R\end{bmatrix}$$

where $R = \sigma_R^2 I$, I is the $n \times n$ identity matrix and G is a diagonal matrix. It is assumed that the random variables, Z, can be subdivided into $g \leq q$ groups with each group being identically distributed with expectations zero and variance σ_i^2. The diagonal elements of matrix G therefore take one of the values $\left\{\sigma_i^2 : i = 1, \ldots, g\right\}$, depending on which group the associated random variable belongs to.

The model therefore contains three sets of unknowns, the fixed effects, β, the random effects ν and a vector of $g+1$ variance components, γ, where $\gamma = \left\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_{g-1}^2, \sigma_g^2, \sigma_R^2\right\}$. Rather than working directly with γ, g02jb uses an iterative process to estimate $\gamma^* = \left\{\sigma_1^2/\sigma_R^2, \sigma_2^2/\sigma_R^2, \ldots, \sigma_{g-1}^2/\sigma_R^2, \sigma_g^2/\sigma_R^2, 1\right\}$. Due to the iterative nature of the estimation a set of initial values, γ_0, for γ^* is required. g02jb allows these initial values either to be supplied by you or calculated from the data using the minimum variance quadratic unbiased estimators (MIVQUE0) suggested by Rao 1972.

g02jb fits the model using a quasi-Newton algorithm to maximize the log-likelihood function:

$$-2l_R = \log(|V|) + (n)\log\left(r'V^{-1}r\right) + \log(2\pi/n)$$

where

$$V = ZGZ' + R, \qquad r = y - Xb \qquad \text{and} \qquad b = \left(X'V^{-1}X\right)^{-1}X'V^{-1}y.$$

Once the final estimates for γ^* have been obtained, the value of σ_R^2 is given by:

$$\sigma_R^2 = \left(r'V^{-1}r\right)/(n-p).$$

Case weights, W_c, can be incorporated into the model by replacing $X'X$ and $Z'Z$ with $X'W_cX$ and $Z'W_cZ$ respectively, for a diagonal weight matrix W_c.

The log-likelihood, l_R, is calculated using the sweep algorithm detailed in Wolfinger *et al.* 1994.

4 References

Goodnight J H 1979 A Tutorial on the SWEEP operator *The American Statistician* **33 (3)** 149–158

Harville D A 1977 Maximum likelihood approaches to variance component estimation and to related problems *JASA* **72** 320–340

Rao C R 1972 Estimation of variance and covariance components in a linear model *J. Am. Stat. Assoc.* **67** 112–115

Stroup W W 1989 Predictable functions and prediction space in the mixed model procedure *Applications of Mixed Models in Agriculture and Related Disciplines* **Southern Cooperative Series Bulletin No. 343** 39–48

Wolfinger R, Tobias R and Sall J 1994 Computing Gaussian Likelihoods and Their Derivatives for General Linear Mixed Models *SIAM Sci. Statist. Comput.* **15** 1294–1310

5 Parameters

5.1 Compulsory Input Parameters

1: **n – int32 scalar**

n, the number of observations.

Constraint: $\mathbf{n} \geq 1$.

2: **nvpr – int32 scalar**

If $\mathbf{rint} = 1$ and $\mathbf{svid} \neq 0$, **nvpr** is the number of variance components being estimated $- 2$, $(g - 1)$, else $\mathbf{nvpr} = g$.

If $\mathbf{nrv} = 0$, **nvpr** is not referenced.

Constraint: if $\mathbf{nrv} \neq 0$, $1 \leq \mathbf{nvpr} \leq \mathbf{nrv}$.

3: **levels(ncol) – int32 array**

$\mathbf{levels}(i)$ contains the number of levels associated with the ith variable of the data matrix **dat**. If this variable is continuous or binary (i.e., only takes the values zero or one) then $\mathbf{levels}(i)$ should be 1; if the variable is discrete then $\mathbf{levels}(i)$ is the number of levels associated with it and $\mathbf{dat}(j, i)$ is assumed to take the values 1 to $\mathbf{levels}(i)$, for $j = 1, 2, \ldots, \mathbf{n}$.

Constraint: $\mathbf{levels}(i) \geq 1$, for $i = 1, 2, \ldots, \mathbf{ncol}$.

4: **yvid – int32 scalar**

The column of **dat** holding the dependent, y, variable.

Constraint: $1 \leq \mathbf{yvid} \leq \mathbf{ncol}$.

5: **fvid(nfv) – int32 array**

The columns of the data matrix **dat** holding the fixed independent variables with $\mathbf{fvid}(i)$ holding the column number corresponding to the ith fixed variable.

Constraint: $1 \leq \mathbf{fvid}(i) \leq \mathbf{ncol}$, for $i = 1, 2, \ldots, \mathbf{nfv}$.

6: **rvid(nrv) – int32 array**

The columns of the data matrix **dat** holding the random independent variables with $\mathbf{rvid}(i)$ holding the column number corresponding to the ith random variable.

Constraint: $1 \leq \mathbf{rvid}(i) \leq \mathbf{ncol}$, for $i = 1, 2, \ldots, \mathbf{nrv}$.

7: **svid – int32 scalar**

The column of **dat** holding the subject variable.

If **svid** $= 0$, no subject variable is used.

Specifying a subject variable is equivalent to specifying the interaction between that variable and all of the random-effects. Letting the notation $Z_1 \times Z_S$ denote the interaction between variables Z_1 and Z_S, fitting a model with **rint** $= 0$, random-effects $Z_1 + Z_2$ and subject variable Z_S is equivalent to fitting a model with random-effects $Z_1 \times Z_S + Z_2 \times Z_S$ and no subject variable. If **rint** $= 1$ the model is equivalent to fitting $Z_S + Z_1 \times Z_S + Z_2 \times Z_S$ and no subject variable.

Constraint: $0 \leq$ **svid** $\leq$ **ncol**.

8: **cwid – int32 scalar**

The column of **dat** holding the case weights.

If **cwid** $= 0$, no weights are used.

Constraint: $0 \leq$ **cwid** $\leq$ **ncol**.

9: **vpr**(**nrv**) **– int32 array**

vpr(i) holds a flag indicating the variance of the ith random variable. The variance of the ith random variable is σ_j^2, where $j =$ **vpr**$(i) + 1$ if **rint** $= 0$ and **svid** $\neq 0$ and $j =$ **vpr**(i) otherwise. Random variables with the same value of j are assumed to be taken from the same distribution.

Constraint: $1 \leq$ **vpr**$(i) \leq$ **nvpr**, for $i = 1, 2, \ldots,$ **nrv**.

10: **dat**(**lddat**,**ncol**) **– double array**

lddat, the first dimension of the array, must be at least **n**.

Array containing all of the data. For the ith observation:

dat$(i,$ **yvid**$)$ holds the dependent variable, y.

If **cwid** $\neq 0$, **dat**$(i,$ **cwid**$)$ holds the case weights.

If **svid** $\neq 0$, **dat**$(i,$ **svid**$)$ holds the subject variable.

The remaining columns hold the values of the independent variables.

Constraints:

if **cwid** $\neq 0$, **dat**$(i,$ **cwid**$) \geq 0$;
if **levels**$(j) \neq 1$, **dat**$(i,j) > 0$, **dat**$(i,j) \leq$ **levels**(j).

11: **fint – int32 scalar**

Flag indicating whether a fixed intercept is included (**fint** $= 1$).

Constraint: **fint** $= 0$ or 1.

12: **rint – int32 scalar**

Flag indicating whether a random intercept is included (**rint** $= 1$).

If **svid** $= 0$, **rint** is not referenced.

Constraint: **rint** $= 0$ or 1.

13: **lb – int32 scalar**

the size of the array **b**.

Constraint: $\mathbf{lb} \geq \mathbf{fint} + \sum_{i=1}^{\mathbf{nfv}} \max(\mathbf{levels}(\mathbf{fvid}(i)) - 1, 1) + L_S \times \left(\mathbf{rint} + \sum_{i=1}^{\mathbf{nrv}} \mathbf{levels}(\mathbf{rvid}(i)) \right)$ where

$L_S = \mathbf{levels}(\mathbf{svid})$ if $\mathbf{svid} \neq 0$ and 1 otherwise

14: **gamma**$(\mathbf{nvpr} + \mathbf{2})$ **– double array**

Holds the initial values of the variance components, γ_0, with $\mathbf{gamma}(i)$ the initial value for σ_i^2/σ_R^2, for $i = 1, 2, \ldots, g$. If $\mathbf{rint} = 1$ and $\mathbf{svid} \neq 0$, $g = \mathbf{nvpr} + 1$, else $g = \mathbf{nvpr}$.

If $\mathbf{gamma}(1) = -1$, the remaining elements of **gamma** are ignored and the initial values for the variance components are estimated from the data using MIVQUE0.

Constraint: $\mathbf{gamma}(1) = -1$ or $\mathbf{gamma}(i) \geq 0$, for $i = 1, 2, \ldots, g$.

15: **maxit – int32 scalar**

The maximum number of iterations.

maxit < 0

The default value of 100 is used.

maxit $= 0$

The parameter estimates (β, ν) and corresponding standard errors are calculated based on the value of γ_0 supplied in **gamma**.

16: **tol – double scalar**

The tolerance used to assess convergence. If $\mathbf{tol} = 0$, the default value of $\epsilon^{0.7}$ is used, where ϵ is the ***machine precision***.

5.2 Optional Input Parameters

1: **ncol – int32 scalar**

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

the number of columns in the data matrix, **dat**.

Constraint: $\mathbf{ncol} \geq 2$.

2: **nfv – int32 scalar**

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

the number of independent variables in the model which are to be treated as being fixed.

Constraint: $0 \leq \mathbf{nfv} < \mathbf{ncol}$.

3: **nrv – int32 scalar**

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

the number of independent variables in the model which are to be treated as being random.

Constraint: $0 \leq \mathbf{nrv} < \mathbf{ncol}$.

5.3 Input Parameters Omitted from the MATLAB Interface

lddat

5.4 Output Parameters

1: **nff – int32 scalar**

The number of fixed effects estimated (i.e., the number of columns, p, in the design matrix X).

2: **nrf – int32 scalar**

The number of random effects estimated (i.e., the number of columns, q, in the design matrix Z).

3: **df – int32 scalar**

The degrees of freedom.

4: **ml – double scalar**

$-2l_R(\hat{\gamma})$ where l_R is the log of the maximum likelihood calculated at $\hat{\gamma}$, the estimated variance components returned in **gamma**.

5: **b(lb) – double array**

The parameter estimates, (β, ν), with the first **nff** elements of **b** containing the fixed effect parameter estimates, β and the next **nrf** elements of **b** containing the random effect parameter estimates, ν.

Fixed effects

If **fint** $= 1$, **b**(1) contains the estimate of the fixed intercept. Let L_i denote the number of levels associated with the ith fixed variable, that is $L_i =$ **levels**(**fvid***i*)). Define

if **fint** $= 1$, $F_1 = 2$ else if **fint** $= 0$, $F_1 = 1$;

$F_{i+1} = F_i + \max(L_i - 1, 1)$, $i \geq 1$.

Then for $i = 1, 2, \ldots,$ **nfv**:

if $L_i > 1$, **b**$(F_i + j - 2)$ contains the parameter estimate for the jth level of the ith fixed variable, for $j = 2, 3, \ldots, L_i$;

if $L_i \leq 1$, **b**(F_i) contains the parameter estimate for the ith fixed variable.

Random effects

Redefining L_i to denote the number of levels associated with the ith random variable, that is $L_i =$ **levels**(**rvid***i*)). Define

if **rint** $= 1$, $R_1 = 2$ else if **rint** $= 0$, $R_1 = 1$;
$R_{i+1} = R_i + L_i$, $i \geq 1$.

Then for $i = 1, 2, \ldots,$ **nrv**:

if **svid** $= 0$,

if $L_i > 1$, **b**$(\mathbf{nff} + R_i + j - 1)$ contains the parameter estimate for the jth level of the ith random variable, for $j = 1, 2, \ldots, L_i$;

if $L_i \leq 1$, **b**$(\mathbf{nff} + R_i)$ contains the parameter estimate for the ith random variable;

if **svid** $\neq 0$,

let L_S denote the number of levels associated with the subject variable, that is $L_S =$ **levels**(**svid**);

if $L_i > 1$, **b**$(\mathbf{nff} + (s-1)L_S + R_i + j - 1)$ contains the parameter estimate for the interaction between the sth level of the subject variable and the jth level of the ith random variable, for $s = 1, 2, \ldots, L_S$ and $j = 1, 2, \ldots, L_i$;

if $L_i \leq 1$, $\mathbf{b}(\mathbf{nff} + (s-1)L_S + R_i)$ contains the parameter estimate for the interaction between the sth level of the subject variable and the ith random variable, for $s = 1, 2, \ldots, L_S$;

if $\mathbf{rint} = 1$, $\mathbf{b}(\mathbf{nff} + 1)$ contains the estimate of the random intercept.

6: **se(lb) – double array**

The standard errors of the parameter estimates given in **b**.

7: **gamma(nvpr + 2) – double array**

gamma(i), for $i = 1, 2, \ldots, g$, holds the final estimate of σ_i^2 and **gamma**$(g+1)$ holds the final estimate for σ_R^2.

8: **warn – int32 scalar**

Is set to 1 if a variance component was estimated to be a negative value during the fitting process. Otherwise **warn** is set to 0.

If $\mathbf{warn} = 1$, the negative estimate is set to zero and the estimation process allowed to continue.

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:

ifail $= 1$

On entry, **n** < 2,
or **ncol** < 1,
or **lddat** $<$ **n**,
or **yvid** < 1 or **yvid** $>$ **ncol**,
or **cwid** < 0 or **cwid** $>$ **ncol**,
or **nfv** < 0 or **nfv** $\geq$ **ncol**,
or **fint** $\neq 0$ and **fint** $\neq 1$,
or **nrv** < 0 or **nrv** $\geq$ **ncol**,
or **nvpr** < 0 or **nvpr** $>$ **nrv** or (**nrv** > 0 and **nvpr** < 1),
or **rint** $\neq 0$ and **rint** $\neq 1$,
or **svid** < 0 or **svid** $>$ **ncol**,
or **lb** is too small.

ifail $= 2$

On entry, **levels**$(i) < 1$, for at least one i,
or **fvid**$(i) < 1$, or **fvid**$(i) >$ **ncol**, for at least one i,
or **rvid**$(i) < 1$, or **rvid**$(i) >$ **ncol**, for at least one i,
or **vpr**$(i) < 1$ or **vpr**$(i) >$ **nvpr**, for at least one i,
or at least one discrete variable in array **dat** has a value greater than that specified in **levels**,
or **gamma**$(i) < 0$, for at least one i, and **gamma**$(1) \neq -1$.

ifail $= 3$

Degrees of freedom < 1. The number of parameters exceed the effective number of observations.

ifail $= 4$

The function failed to converge to the specified tolerance in **maxit** iterations. See Section 8 for advice.

7 Accuracy

The accuracy of the results can be adjusted through the use of the **tol** parameter.

8 Further Comments

Wherever possible any block structure present in the design matrix Z should be modelled through a subject variable, specified via **svid**, rather than being explicitly entered into **dat**.

g02jb uses an iterative process to fit the specified model and for some problems this process may fail to converge (see **ifail** $= 4$). If the function fails to converge then the maximum number of iterations (see **maxit**) or tolerance (see **tol**) may require increasing; try a different starting estimate in **gamma**. Alternatively, the model can be fit using restricted maximum likelihood (see g02ja) or using the noniterative MIVQUE0.

To fit the model just using MIVQUE0, the first element of **gamma** should be set to -1 and **maxit** should be set to zero.

Although the quasi-Newton algorithm used in g02jb tends to require more iterations before converging compared to the Newton–Raphson algorithm recommended by Wolfinger *et al.* 1994, it does not require the second derivatives of the likelihood function to be calculated and consequentially takes significantly less time per iteration.

9 Example

```
n = int32(24);
nvpr = int32(1);
levels = [int32(1);
     int32(4);
     int32(3);
     int32(2);
     int32(3)];
yvid = int32(1);
fvid = [int32(3);
     int32(4);
     int32(5)];
rvid = [int32(3)];
svid = int32(2);
cwid = int32(0);
vpr = [int32(1)];
dat = [56, 1, 1, 1, 1;
     50, 1, 2, 1, 1;
     39, 1, 3, 1, 1;
     30, 2, 1, 1, 1;
     36, 2, 2, 1, 1;
     33, 2, 3, 1, 1;
     32, 3, 1, 1, 1;
     31, 3, 2, 1, 1;
     15, 3, 3, 1, 1;
     30, 4, 1, 1, 1;
     35, 4, 2, 1, 1;
     17, 4, 3, 1, 1;
     41, 1, 1, 2, 1;
     36, 1, 2, 2, 2;
     35, 1, 3, 2, 3;
     25, 2, 1, 2, 1;
     28, 2, 2, 2, 2;
     30, 2, 3, 2, 3;
     24, 3, 1, 2, 1;
     27, 3, 2, 2, 2;
     19, 3, 3, 2, 3;
     25, 4, 1, 2, 1;
     30, 4, 2, 2, 2;
     18, 4, 3, 2, 3];
```

```
fint = int32(1);
rint = int32(1);
lb = int32(25);
gamma = [1;
     1;
     0];
maxit = int32(-1);
tol = 0;
[nff, nrf, df, ml, b, se, gammaOut, warn, ifail] = ...
     g02jb(n, nvpr, levels, yvid, fvid, rvid, svid, cwid, vpr, dat, fint,
rint, lb, gamma, maxit, tol)
```

```
nff =
           6
nrf =
          16
df =
          16
ml =
  141.6877
b =
   37.0000
    1.0000
  -11.0000
   -8.2500
    0.5000
    7.7500
   10.7631
    3.7276
   -1.4476
    0.3733
   -0.5269
   -3.7171
   -1.2253
    4.8125
   -5.6450
    0.5903
    0.3987
   -2.3806
   -4.5912
   -0.6009
    2.2742
   -2.8052
         0
         0
         0
se =
    4.0421
    3.0461
    3.0461
    1.8736
    2.6497
    2.6497
    3.8855
    2.6268
    2.6268
    2.6268
    3.8855
    2.6268
    2.6268
    2.6268
    3.8855
    2.6268
    2.6268
    2.6268
    3.8855
    2.6268
    2.6268
    2.6268
         0
```

```
         0
         0
gammaOut =
   46.7969
   11.5365
    7.0208
warn =
           0
ifail =
           0
```