NAG Toolbox

nag_univar_outlier_peirce_1var (g07ga)

1 Purpose

nag_univar_outlier_peirce_1var (g07ga) identifies outlying values using Peirce's criterion.

2 Syntax

```
[iout, niout, diff, llamb, ifail] = nag_univar_outlier_peirce_1var(p, y, ldiff, 'n', n, 'mean', mean, 'var', var)
[iout, niout, diff, llamb, ifail] = g07ga(p, y, ldiff, 'n', n, 'mean', mean, 'var', var)
```

3 Description

nag_univar_outlier_peirce_1var (g07ga) flags outlying values in data using Peirce's criterion. Let

y denote a vector of n observations (for example the residuals) obtained from a model with p parameters,

m denote the number of potential outlying values,

μ and σ^2 denote the mean and variance of y respectively,

$\tilde{y}$ denote a vector of length $n-m$ constructed by dropping the m values from y with the largest value of $|y_i - \mu|$,

$\tilde{\sigma}^2$ denote the (unknown) variance of $\tilde{y}$,

λ denote the ratio of $\tilde{\sigma}$ and σ with $\lambda = \frac{\tilde{\sigma}}{\sigma}$.

Peirce's method flags y_i as a potential outlier if $|y_i - \mu| \geq x$, where $x = \sigma^2 z$ and z is obtained from the solution of

$$R^m = \lambda^{m-n}\frac{m^m(n-m)^{n-m}}{n^n} \tag{1}$$

where

$$R = 2\exp\left(\left(\frac{z^2-1}{2}\right)(1-\Phi(z))\right) \tag{2}$$

and Φ is the cumulative distribution function for the standard Normal distribution.

As $\tilde{\sigma}^2$ is unknown an assumption is made that the relationship between $\tilde{\sigma}^2$ and σ^2, hence λ, depends only on the sum of squares of the rejected observations and the ratio estimated as

$$\lambda^2 = \frac{n-p-mz^2}{n-p-m}$$

which gives

$$z^2 = 1 + \frac{n-p-m}{m}(1-\lambda^2) \tag{3}$$

A value for the cutoff x is calculated iteratively. An initial value of $R = 0.2$ is used and a value of λ is estimated using equation (1). Equation (3) is then used to obtain an estimate of z and then equation (2) is used to get a new estimate for R. This process is then repeated until the relative change in z between consecutive iterations is $\leq \sqrt{\epsilon}$, where ϵ is ***machine precision***.

By construction, the cutoff for testing for $m+1$ potential outliers is less than the cutoff for testing for m potential outliers. Therefore Peirce's criterion is used in sequence with the existence of a single potential outlier being investigated first. If one is found, the existence of two potential outliers is investigated etc.

If one of a duplicate series of observations is flagged as an outlier, then all of them are flagged as outliers.

4 References

Gould B A (1855) On Peirce's criterion for the rejection of doubtful observations, with tables for facilitating its application *The Astronomical Journal* **45**

Peirce B (1852) Criterion for the rejection of doubtful observations *The Astronomical Journal* **45**

5 Parameters

5.1 Compulsory Input Parameters

1: **p** – INTEGER

p, the number of parameters in the model used in obtaining the y. If y is an observed set of values, as opposed to the residuals from fitting a model with p parameters, then p should be set to 1, i.e., as if a model just containing the mean had been used.

Constraint: $1 \leq \mathbf{p} \leq \mathbf{n} - 2$.

2: **y**(**n**) – REAL (KIND=nag_wp) array

y, the data being tested.

3: **ldiff** – INTEGER

The maximum number of values to be returned in arrays **diff** and **llamb**.

If **ldiff** ≤ 0, arrays **diff** and **llamb** are not referenced.

5.2 Optional Input Parameters

1: **n** – INTEGER

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

n, the number of observations.

Constraint: **n** ≥ 3.

2: **mean_p** – REAL (KIND=nag_wp)

Default: 0.0

If **var** > 0.0, **mean** must contain μ, the mean of y, otherwise **mean** is not referenced and the mean is calculated from the data supplied in **y**.

3: **var** – REAL (KIND=nag_wp)

Default: 0.0

If **var** > 0.0, **var** must contain σ^2, the variance of y, otherwise the variance is calculated from the data supplied in **y**.

5.3 Output Parameters

1: **iout**(**n**) – INTEGER array

The indices of the values in **y** sorted in descending order of the absolute difference from the mean, therefore $|\mathbf{y}(\mathbf{iout}(i-1)) - \mu| \geq |\mathbf{y}(\mathbf{iout}(i)) - \mu|$, for $i = 2, 3, \ldots, \mathbf{n}$.

2: **niout** – INTEGER

The number of potential outliers. The indices for these potential outliers are held in the first **niout** elements of **iout**. By construction there can be at most $\mathbf{n} - \mathbf{p} - 1$ values flagged as outliers.

3: **dif**(**ldiff**) – REAL (KIND=nag_wp) array

diff(i) holds $|y - \mu| - \sigma^2 z$ for observation **y**(**iout**(i)), for $i = 1, 2, \ldots, \min(\mathbf{ldiff}, \mathbf{niout} + 1, \mathbf{n} - \mathbf{p} - 1)$.

4: **llamb**(**ldiff**) – REAL (KIND=nag_wp) array

llamb(i) holds $\log\left(\lambda^2\right)$ for observation **y**(**iout**(i)), for $i = 1, 2, \ldots, \min(\mathbf{ldiff}, \mathbf{niout} + 1, \mathbf{n} - \mathbf{p} - 1)$.

5: **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 $= 1$

Constraint: $\mathbf{n} \geq 3$.

ifail $= 2$

Constraint: $1 \leq \mathbf{p} \leq \mathbf{n} - 2$.

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

Not applicable.

8 Further Comments

One problem with Peirce's algorithm as implemented in nag_univar_outlier_peirce_1var (g07ga) is the assumed relationship between σ^2, the variance using the full dataset, and $\tilde{\sigma}^2$, the variance with the potential outliers removed. In some cases, for example if the data y were the residuals from a linear regression, this assumption may not hold as the regression line may change significantly when outlying values have been dropped resulting in a radically different set of residuals. In such cases nag_univar_outlier_peirce_2var (g07gb) should be used instead.

9 Example

This example reads in a series of data and flags any potential outliers.

The dataset used is from Peirce's original paper and consists of fifteen observations on the vertical semidiameter of Venus.

9.1 Program Text

```
function g07ga_example

fprintf('g07ga example results\n\n');

y = [-0.30;  0.48;  0.63; -0.22; 0.18;
     -0.44; -0.24; -0.13; -0.05; 0.39;
      1.01;  0.06; -1.40;  0.20; 0.10];

p     = nag_int(2);
ldiff = nag_int(1);

% Get a list of potential outliers
[iout, niout, dif, llamb, ifail] = ...
  g07ga(p, y, ldiff);

% Display results
fprintf('Number of potential outliers: %2d\n',niout);
fprintf('  No.  Index    Value');
if ldiff > 0
  fprintf('       Diff    ln(lambda^2)');
end
fprintf('\n');

for i=1:niout
  fprintf(' %4d %4d %10.2f', i, iout(i), y(iout(i)));
  if i <= ldiff
    fprintf(' %10.2f %10.2f', dif(i), llamb(i));
  end
  fprintf('\n');
end
```

9.2 Program Results

```
g07ga example results

Number of potential outliers:  2
  No.  Index    Value       Diff    ln(lambda^2)
    1   13      -1.40       0.31      -0.30
    2   11       1.01
```