nag_simple_linear_regression (g02cac)

1. Purpose

nag_simple_linear_regression (g02cac) performs a simple linear regression with or without a constant term. The data is optionally weighted.

2. Specification

```
#include <nag.h>
#include <nagg02.h>

void nag_simple_linear_regression(Nag_SumSquare mean, Integer n,
   double x[], double y[], double wt[], double *a, double *b, double *a_serr,
   double *b_serr, double *rsq, double *rss, double *df, NagError *fail)
```

3. Description

This function fits a straight line model of the form,

$$E(y) = a + bx,$$

where $E(y)$ is the expected value of the variable y, to the data points

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n),$$

such that

$$y_i = a + bx_i + e_i, i = 1, 2, \ldots, n(n > 2).$$

where the e_i values are independent random errors. The ith data point may have an associated weight w_i, these may be used either in the situation when var $(\varepsilon_i) = \sigma^2/w_i$ or if observations have to be removed from the regression by having zero weight or have been observed with frequency w_i.

The regression coefficient, b, and the regression constant, a are estimated by minimizing

$$\sum_{i=1}^{n} w_i e_i^2,$$

if the weights option is not selected then $w_i = 1.0$.

The following statistics are computed:
the estimate of regression constant $\hat{a} = \bar{y} - \hat{b}\bar{x}$,

the estimate of regression coefficient $\hat{b} = \dfrac{\sum w_i(x_i - \bar{x})(y_i - \bar{y})}{\sum w_i(x_i - \bar{x})^2}$,

the residual sum of squares $rss = \sum w_i(y_i - \hat{y}_i)^2$,
where the weighted means $\bar{x}$ and $\bar{y}$ are

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i} \quad \text{and} \quad \bar{y} = \frac{\sum w_i y_i}{\sum w_i}.$$

The number of degrees of freedom associated with rss is

$$df = \sum w_i - 2 \quad \text{where } \mathbf{mean} = \mathbf{Nag\_AboutMean}$$
$$df = \sum w_i - 1 \quad \text{where } \mathbf{mean} = \mathbf{Nag\_AboutZero}.$$

Note: the weights should be scaled to give the correct degrees of freedom in the case var $(\varepsilon_i) = \sigma^2/w_i$.
The R^2 value or coefficient of determination

$$R^2 = \frac{\sum w_i(\hat{y}_i - \bar{y}_i)^2}{\sum w_i(y_i - \bar{y})^2} = \frac{\sum w_i(y_i - \bar{y})^2 - rss}{\sum w_i(y_i - \bar{y})^2}.$$

This measures the proportion of the total variation about the mean $\bar{y}$ that can be explained by the regression.

The standard error for the regression constant $\hat{a}$

$$a\_serr = \sqrt{\frac{rss}{df}\left(\frac{1}{\sum w_i} + \frac{(\bar{x})^2}{\sum w_i(x_i - \bar{x})^2}\right)} = \sqrt{\frac{rss}{df}\frac{1}{\sum w_i}\frac{\sum w_i x_i^2}{\sum w_i(x_i - \bar{x})^2}}.$$

The standard error for the regression coefficient $\hat{b}$

$$b\_serr = \sqrt{\frac{rss}{df \sum w_i(x_i - \bar{x})^2}}.$$

Similar formulae can be derived for the case when the line goes through the origin, that is $a = 0$.

4. Parameters

mean

Input: indicates whether nag_simple_linear_regression is to include a constant term in the regression.
If **mean** = **Nag_AboutMean**, the regression constant a is included.
If **mean** = **Nag_AboutZero**, the regression constant a is not included, i.e., $a = 0$
Constraint: **mean** = **Nag_AboutMean** or **Nag_AboutZero**.

n

Input: the number of observations, n.
Constraint: if **mean** = **Nag_AboutMean** **n** $\geq$ **2**. If **mean** = **Nag_AboutZero** **n** $\geq$ **1**.

x[n]

Input: the values of the independent variable with the ith value stored in $x[i-1]$ for $i = 1,\ldots,n$.
Constraint: all the values of x must not be identical.

y[n]

Input: the values of the dependent variable with the ith value stored in $y[i-1]$ for $i = 1,\ldots,n$.
Constraint: all the values of y must not be identical.

wt[n]

Input: if weighted estimates are required then **wt** must contain the weights to be used in the weighted regression. Otherwise **wt** need not be defined and may be set to the null pointer **NULL**, i.e.(double *)0.
Usually **wt**$[i-1]$ will be an integral value corresponding to the number of observations associated with the ith data point, or zero if the ith data point is to be ignored. The sum of the weights therefore represents the effective total number of observations used to create the regression line.
If **wt** = **NULL**, then the effective number of observations is n.
Constraint: **wt** = **NULL** or **wt**$[i-1] \geq 0.0$, for $i = 1,\ldots,n$.

a

Output: If **mean** = **Nag_AboutMean** then **a** is the regression constant $\hat{a}$, otherwise **a** is set to zero.

b

Output: the regression coefficient $\hat{b}$.

a_serr

Output: the standard error of the regression constant $\hat{a}$.

b_serr

Output: the standard error of the regression coefficient $\hat{b}$.

rsq

Output: the coefficient of determination, R^2.

rss

Output: the sum of squares of the residuals about the regression.

df

Output: the degrees of freedom associated with the residual sum of squares.

fail

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

5. Error Indications and Warnings

NE_BAD_PARAM

On entry, parameter **mean** had an illegal value.

NE_INT_ARG_LT

On entry, **n** must not be less than 1: **n** = ⟨*value*⟩
if **mean** = **Nag_AboutZero**.
On entry, **n** must not be less than 2: **n** = ⟨*value*⟩
if **mean** = **Nag_AboutMean**.

NE_NEG_WEIGHT

On entry, at least one of the weights is negative.

NE_WT_LOW

On entry, **wt** must contain at least 1 positive element if **mean** = **Nag_AboutZero** or at least 2 positive elements if **mean** = **Nag_AboutMean**.

NE_X_OR_Y_IDEN

On entry, all elements of **x** and/or **y** are equal.

NE_SW_LOW

On entry, the sum of elements of **wt** must be greater than 1.0 if **mean** = **Nag_AboutZero** or greater than 2.0 if **mean** = **Nag_AboutMean**.

NE_ZERO_DOF_RESID

On entry, the degrees of freedom for the residual are zero, i.e., the designated number of parameters = the effective number of observations.

NW_RSS_EQ_ZERO

Residual sum of squares is zero, i.e., a perfect fit was obtained.

6. Further Comments

The time taken by the function depends on n.
The function uses a two-pass algorithm.

6.1. Accuracy

The computations are believed to be stable.

6.2. References

Draper N R and Smith H (1981) *Applied Regression Analysis.* (2nd Edn) Wiley.

7. See Also

nag_regress_confid_interval (g02cbc)

8. Example

A program to calculate regression constants, $\hat{a}$ and $\hat{b}$, the standard error of the regression constants, the regression coefficient of determination and the degrees of freedom about the regression.

8.1. Program Text

```
/* nag_simple_linear_regression(g02cac) Example Program
 *
 * Copyright 1994 Numerical Algorithms Group.
 *
 * Mark 3, 1994.
 */

#include <nag.h>
#include <stdio.h>
#include <nag_stdlib.h>
#include <nagg02.h>

#define NMAX 10
main()
{
  Nag_SumSquare mean;
  char m, w;
  Integer i, n;
  double x[NMAX], y[NMAX], wt[NMAX];
  double a, b, err_a, err_b, rsq, rss, df;
  double *wtptr;

  Vprintf("g02cac Example Program Results\n");
  /*  Skip heading in data file */
  Vscanf("%*[^\n]");
  Vscanf(" %c %c",&m, &w);
  Vscanf("%ld", &n);
  if (n>=1 && n<=NMAX)
    {
      if (m == 'M' || m == 'm')
        mean = Nag_AboutMean;
      else
        mean = Nag_AboutZero;
      if (w == 'W' || w == 'w')
        {
          wtptr = wt;
          for(i = 0; i < n; ++i)
            Vscanf("%lf%lf%lf", &x[i], &y[i], &wt[i]);
        }
      else
        {
          wtptr = (double *)0;
          for(i = 0; i < n; ++i)
            Vscanf("%lf%lf", &x[i], &y[i]);
        }

      g02cac(mean, n, x, y, wtptr, &a, &b, &err_a, &err_b, &rsq, &rss,
             &df, NAGERR_DEFAULT);


      if (mean == Nag_AboutMean)
        {
          Vprintf("\nRegression constant a = %6.4f\n\n", a);
          Vprintf("Standard error of the regression constant a = %6.4f\n\n",
                  err_a);
        }

      Vprintf("Regression coefficient b = %6.4f\n\n", b);
      Vprintf("Standard error of the regression coefficient b = %6.4f\n\n",
              err_b);

      Vprintf("The regression coefficient of determination = %6.4f\n\n", rsq);
      Vprintf("The sum of squares of the residuals about the \
regression = %6.4f\n\n", rss);
      Vprintf("Number of degrees of freedom about the \
regression = %6.4f\n\n",df);
    }
  else
```

```
    {
      Vfprintf(stderr, "n is out of range:\
 n = %-3ld\n",n);
      exit(EXIT_FAILURE);
    }
  exit(EXIT_SUCCESS);
}
```

8.2. Program Data

```
g02cac Example Program Data
m w
8
1.0  20.0  1.0
0.0  15.5  1.0
4.0  28.3  1.0
7.5  45.0  1.0
2.5  24.5  1.0
0.0  10.0  1.0
10.0 99.0  1.0
5.0  31.2  1.0
```

8.3. Program Results

```
g02cac Example Program Results

Regression constant a = 7.5982

Standard error of the regression constant a = 6.6858

Regression coefficient b = 7.0905

Standard error of the regression coefficient b = 1.3224

The regression coefficient of determination = 0.8273

The sum of squares of the residuals about the regression = 965.2454

Number of degrees of freedom about the regression = 6.0000
```