Stepwise Regression: nagdmc_stepwise_reg

Purpose

nagdmc_stepwise_reg computes a forward stepwise regression on p variables using Clarke's sweep algorithm.

Declaration

```
#include <nagdmc.h>
void nagdmc_stepwise_reg(long rec1, long nvar, long nrec, long dblk, double data[],
                         long nxvar, long xvar[], long yvar, long iwts, int xtype[],
                         double Fin, double Fout, double eps, double b[],
                         double se[], double *R2, double *rms, long *df,
                         double model[], int printlevel, const char file[],
                         int *info);
```

Parameters

1: **rec1** – `long` *Input*

On entry: the index in the data of the first data record used in the analysis.

Constraint: **rec1** ≥ 0.

2: **nvar** – `long` *Input*

On entry: the number of variables in the data.

Constraint: **nvar** > 1.

3: **nrec** – `long` *Input*

On entry: the number of consecutive records, beginning at **rec1**, used in the analysis.

Constraint: **nrec** > 1.

4: **dblk** – `long` *Input*

On entry: the total number of records in the data block.

Constraint: **dblk** $\geq$ **rec1** $+$ **nrec**.

5: **data**[**dblk** $*$ **nvar**] – `double` *Input*

On entry: the data values for the jth variable (for $j = 0, 1, \ldots,$ **nvar**-1) are stored in **data**[$i*$**nvar**$+j$], for $i = 0, 1, \ldots,$ **dblk** $- 1$.

6: **nxvar** – `long` *Input*

On entry: the number of independent variables. If **nxvar** $= 0$ then all variables in the data, excluding **yvar** and **iwts**, are treated as independent variables.

Constraint: $0 \leq$ **nxvar** $<$ **nvar**.

7: **xvar**[**nxvar**] – `long` *Input*

On entry: the indices indicating the position in **data** in which values of the independent variables are stored. If **nxvar** $= 0$ then **xvar** must be 0, and the indices of independent variables are given by $j = 0, 1, \ldots,$ **nvar** $- 1$; $j \neq$ **yvar** and $i \neq$ **iwts**.

Constraints: if **nxvar** > 0, $0 \leq$ **xvar**$[i]$ $<$ **nvar**, for $i = 0, 1, \ldots,$ **nxvar** $- 1$; otherwise **xvar** must be 0.

8: **yvar** – `long` *Input*

On entry: the index in **data** in which values of the dependent variable are stored.

Constraints: $0 \leq$ **yvar** $<$ **nvar**; if **nxvar** > 0, **yvar** $\neq$ **xvar**$[i]$, for $i = 0, 1, \ldots,$ **nxvar** $- 1$.

9: **iwts** – `long` *Input*

On entry: if **iwts** $= -1$, no weights are used; otherwise **iwts** is the index in **data** in which the weights are stored.

Constraints: $-1 \leq$ **iwts** $<$ **nvar**; **iwts** $\neq$ **yvar**; and if **nxvar** > 0, **iwts** $\neq$ **xvar**$[i]$, for $i = 0, 1, \ldots,$ **nxvar** $- 1$.

10: **xtype**[$p-1$] – `int` *Input*

On entry: set **xtype**[j] $= 2$ to force the jth independent variable to be selected for a model; otherwise **xtype**[j] must be set equal to zero, for $j = 0, 1, \ldots, p-2$, where $p-1$ is the number of independent variables in the analysis.
On exit: the value of **xtype**[j] indicates the status of the jth independent variable in the data for a model:
xtype[j] $= 0$; omitted
xtype[j] $= 1$; selected
xtype[j] $= 2$; forced selection

Constraint: **xtype**[j] $\in \{0, 2\}$, for $j = 0, 1, \ldots, p-2$.

11: **Fin** – `double` *Input*

On entry: the value of the F-test which the coefficient of a variable must exceed for that variable to be included in a model. Typical values for **Fin** lie in the interval $[2, 4]$.

Constraint: **Fin** > 0.0.

Suggested value: **Fin** $= 2$.

12: **Fout** – `double` *Input*

On entry: the variable in a model corresponding to the coefficent with the lowest F-test value is removed from the model if its coefficent's F-test is less than **Fout**. The value of **Fout** is usually set equal to the value of **Fin**; a value less than **Fin** is occasionally prefered.

Constraint: $0.0 <$ **Fout** $\leq$ **Fin**.

Suggested value: **Fout** $= 2$.

13: **eps** – `double` *Input*

On entry: the tolerance for detecting collinearities between variables when adding or removing a variable from a model. Variables deemed to be collinear are excluded from the final model.

Constraint: **eps** > 0.0.

Suggested value: **eps** $= 1.0 \times 10^{-6}$.

14: **b**[p] – `double` *Output*

On exit: the parameter estimates. **b**[0] is the mean parameter. **b**[i] is the coefficient of the ith variable included in the model, for $i = 1, 2, \ldots, p-1$. If **nxvar** > 0 then the order the independent variables are added to the model is defined by **xvar**, otherwise the order is defined by indices in the data.

15: **se**[p] – `double` *Output*

On exit: the standard errors of the parameters in **b**.

16: **R2** – `double *` *Output*

On exit: the R^2-value for the fitted model.

17: **rms** – `double *` *Output*

On exit: the residual mean square for the fitted model.

18: **df** – `long *` *Output*

On exit: the degrees of freedom for the for the residual mean square.

19: **model**[$(3 * p * (p+1))/2 +$ **nvar** $+ 14$] – `double` *Output*

On exit: if not 0, information on the fitted model for use in the functions described in 'See Also'.

20: **printlevel** – `int` *Input*

On entry: controls the information printed by the function. If **printlevel** is set equal to zero, no information is printed; otherwise **printlevel** must be set equal to one, and information is printed to a location determined by the value of **file**.

Constraint: **printlevel** $\in \{0, 1\}$.

21: **file**[] – `char` *Input*

On entry: if **printlevel** = 1, the value of **file** determines the destination of the printed summary information; otherwise **file** is not referenced. If **file** is equal to 0, information is printed to the screen of the terminal used to call **nagdmc_stepwise_reg**; otherwise information is printed to the ASCII file named **file**.

Constraint: if **printlevel** = 1 and **file** $\neq 0$, **nagdmc_stepwise_reg** returns an error if a file named **file** cannot be opened for writing.

22: **info** – `int *` *Output*

On exit: **info** gives information on the success of the function call:

-1:

0: the function successfully completed its task..one or more of the variance terms were zero when computing the correlation matrix

i; $i = 1, 2, \ldots, 4, 6, 7, \ldots, 13, 20, 21$: the specification of the ith formal parameter was incorrect.

55: a negative value for a weight was found.

99: the function failed to allocate enough memory.

100: an internal error occurred during the execution of the function.

Notation

nrec	the number of data records, n.
nxvar	determines the number of independent variables, $p-1$.
xvar	the independent variables that take the values in X.
yvar	the dependent variable that takes the values in y.
iwts	if **iwts** ≥ 0, **iwts** is the index in the data that defines the weights, W.
Fin	the minimum value, α_1, of partial F-tests for entry into a model.
Fout	the minimum value, α_2, of partial F-tests for removal from a model.

Description

The general multiple linear regression model is defined by

$$y = X\beta + \varepsilon,$$

where y is a vector of n observations on the dependent variable,
X is an n by p matrix of the independent variables of column rank k,
β is a vector of length p of unknown parameters,
and ε is a vector of length n of unknown, Normally distributed, random errors such that var $\varepsilon = V\sigma^2$, where V is a known diagonal matrix of size n.

In most linear regression models the first term is taken as a mean term or an intercept, i.e., $X_{i,1} = 1$, for $i = 1, 2, \ldots, n$; this is assumed in the NAG DMC.

The least squares estimates $\hat{\beta}$ of the parameters β minimise $(y - X\beta)^T(y - X\beta)$ while the weighted least squares estimates minimise $(y - X\beta)^T W(y - X\beta)$.

The goal of stepwise regression is to determine a model by selecting statistically significant variables from the $p-1$ available variables (the intercept is always included in a model). The forward stepwise method used to select variables incorporates an entry step and a removal step.

In the entry step the entry statistic is computed for each variable eligible for entry in the model. If no variable has a partial F-test value which exceeds the specified critical value, α_1, then the process is terminated; otherwise the variable with the largest value on the entry statistic is entered into the model.

In the removal step the removal statistic is computed for each variable eligible to be removed from the model. If no variable has a partial F-test value which is less than a critical value, α_2, then the process is terminated; otherwise the variable with the smallest value on the removal statistic is removed from the model.

The forward stepwise procedure begins by performing forward entry for all independent variables in a model. At any subsequent step where two or more variables have been selected for entry into the model, the entry procedure followed by the removal procedure is performed. The forward stepwise selection procedure halts when neither procedure can be performed.

References and Further Reading

Clarke M R B (1981) Algorithm AS 178: the Gauss-Jordan sweep operator with detection of collinearity *Applied Statistics* **31** 166-169.

Dempster A P (1969) *Elements of Continuous Multivariate Analysis* Addison-Wesley.

Draper N R and Smith H (1985) *Applied Regression Analysis* (2nd Edition) Wiley.

See Also

nagdmc_basic_reg	simplified version of **nagdmc_linear_reg** using a restricted set of parameters.
nagdmc_extr	computes fitted values, residuals and leverages for a regression.
nagdmc_linear_reg	linear model with Normal errors.
nagdmc_predict_reg	computes predictions given a fitted regression model.
stepwise_reg_ex.c	the example calling program.