NAG C Library Chapter Introduction

g02 – Correlation and Regression Analysis

Contents

1 Scope of the Chapter

This chapter is concerned with two techniques – correlation analysis and regression modelling – both of which are concerned with determining the inter-relationships among two or more variables.

Other chapters of the NAG C Library which cover similar problems are Chapter e02 and Chapter e04. Chapter e02 routines may be used to fit linear models by criteria other than least-squares, and also for polynomial regression; Chapter e04 routines may be used to fit nonlinear models and linearly constrained linear models.

2 Background to the Problems

2.1 Correlation

2.1.1 Aims of correlation analysis

Correlation analysis provides a single summary statistic – the correlation coefficient – describing the strength of the **association** between two variables. The most common types of association which are investigated by correlation analysis are linear relationships, and there are a number of forms of linear correlation coefficients for use with different types of data.

2.1.2 Correlation coefficients

The (Pearson) product-moment correlation coefficients measure a linear relationship, while Kendall's tau and Spearman's rank order correlation coefficients measure monotonicity only. All three coefficients range from -1.0 to $+1.0$. A coefficient of zero always indicates that no **linear** relationship exists; a $+1.0$ coefficient implies a 'perfect' positive relationship (i.e., an increase in one variable is always associated with a corresponding increase in the other variable); and a coefficient of -1.0 indicates a 'perfect' negative relationship (i.e., an increase in one variable is always associated with a corresponding decrease in the other variable).

Consider the bivariate scattergrams in Figure 1: (a) and (b) show strictly linear functions for which the values of the product-moment correlation coefficient, and (since a linear function is also monotonic) both Kendall's tau and Spearman's rank order coefficients, would be $+1.0$ and -1.0 respectively. However, though the relationships in figures (c) and (d) are respectively monotonically increasing and monotonically decreasing, for which both Kendall's and Spearman's non-parametric coefficients would be $+1.0$ (in (c)) and -1.0 (in (d)), the functions are nonlinear so that the product-moment coefficients would not take such 'perfect' extreme values. There is no obvious relationship between the variables in figure (e), so all three coefficients would assume values close to zero, while in figure (f) though there is an obvious parabolic relationship between the two variables, it would not be detected by any of the correlation coefficients which would again take values near to zero; it is important therefore to examine scattergrams as well as the correlation coefficients.

In order to decide which type of correlation is the most appropriate, it is necessary to appreciate the different groups into which variables may be classified. Variables are generally divided into four types of scales: the nominal scale, the ordinal scale, the interval scale, and the ratio scale. The nominal scale is used only to categorise data; for each category a name, perhaps numeric, is assigned so that two different categories will be identified by distinct names. The ordinal scale, as well as categorising the observations, orders the categories. Each category is assigned a distinct identifying symbol, in such a way that the order of the symbols corresponds to the order of the categories. (The most common system for ordinal variables is to assign numerical identifiers to the categories, though if they have previously been assigned alphabetic characters, these may be transformed to a numerical system by any convenient method which preserves the ordering of the categories.) The interval scale not only categorises and orders the observations, but also quantifies the comparison between categories; this necessitates a common unit of measurement and an arbitrary zero-point. Finally, the ratio scale is similar to the interval scale, except that it has an **absolute** (as opposed to **arbitrary**) zero-point.

For a more complete discussion of these four types of scales, and some examples, the user is referred to Churchman and Ratoosh (1959) and Hays (1970).

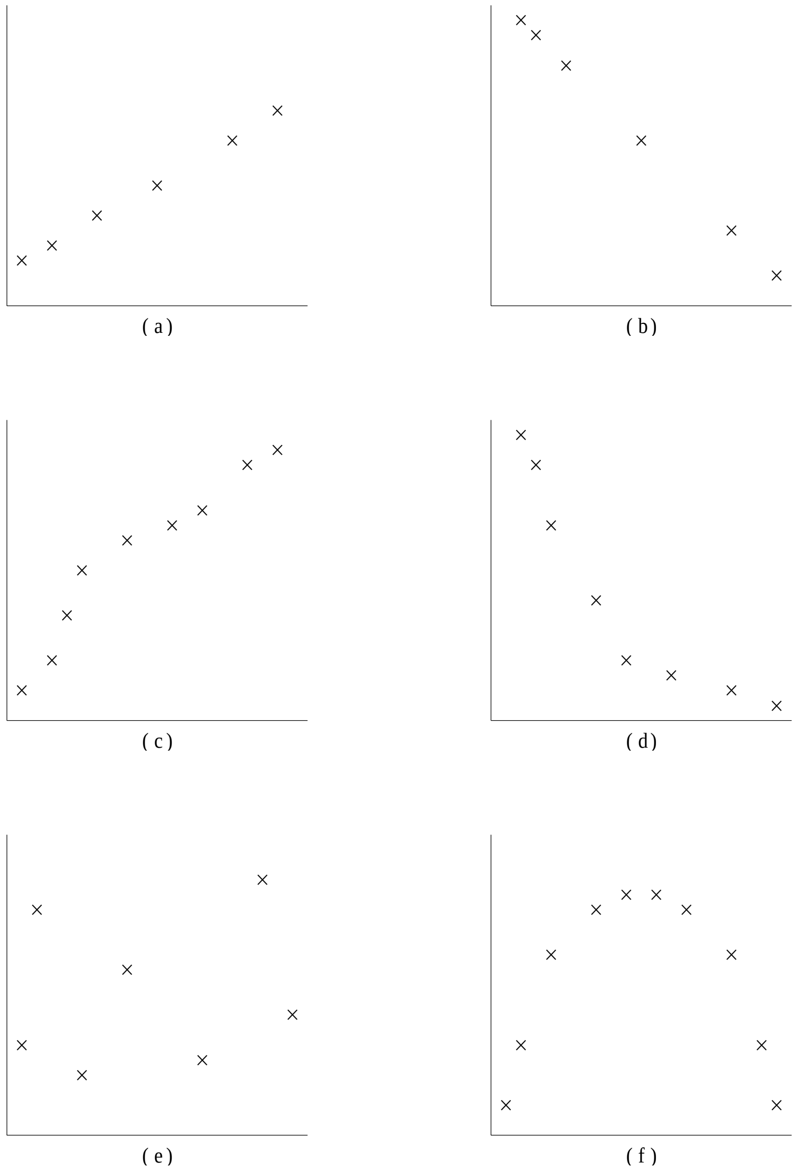


Figure 1

Product-moment correlation coefficients are used with variables which are interval (or ratio) scales; these coefficients measure the amount of spread about the linear least-squares equation. For a product-moment correlation coefficient, r, based on n pairs of observations, testing against the null hypothesis that there is no correlation between the two variables, the statistic

$$r\sqrt{\frac{n-2}{1-r^2}}$$

has a Student's t-distribution with $n-2$ degrees of freedom; its significance can be tested accordingly.

Ranked and ordinal scale data are generally analysed by non-parametric methods – usually either Spearman's or Kendall's tau rank-order correlation coefficients, which, as their names suggest, operate solely on the ranks, or relative orders, of the data values. Interval or ratio scale variables may also be validly analysed by non-parametric methods, but such techniques are statistically less powerful than a product-moment method. For a Spearman rank-order correlation coefficient, R, based on n pairs of observations, testing against the null hypothesis that there is no correlation between the two variables, for large samples the statistic

$$R\sqrt{\frac{n-2}{1-R^2}}$$

has approximately a Student's t-distribution with $n-2$ degrees of freedom, and may be treated accordingly. (This is similar to the product-moment correlation coefficient, r, see above.) Kendall's tau coefficient, based on n pairs of observations, has, for large samples, an approximately Normal distribution with mean zero and standard deviation

$$\sqrt{\frac{4n+10}{9n(n-1)}}$$

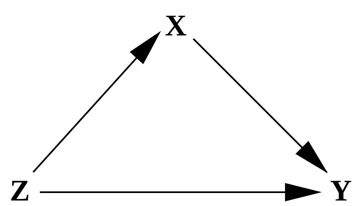
when tested against the null hypothesis that there is no correlation between the two variables; the coefficient should therefore be divided by this standard deviation and tested against the standard Normal distribution, N(0,1).

When the number of ordinal categories a variable takes is large, and the number of ties is relatively small, Spearman's rank-order correlation coefficients have advantages over Kendall's tau; conversely, when the number of categories is small, or there are a large number of ties, Kendall's tau is usually preferred. Thus when the ordinal scale is more or less continuous, Spearman's rank-order coefficients are preferred, whereas Kendall's tau is used when the data is grouped into a smaller number of categories; both measures do however include corrections for the occurrence of ties, and the basic concepts underlying the two coefficients are quite similar. The absolute value of Kendall's tau coefficient tends to be slightly smaller than Spearman's coefficient for the same set of data.

There is no authoritative dictum on the selection of correlation coefficients – particularly on the advisability of using correlations with ordinal data. This is a matter of discretion for the user.

2.1.3 Partial Correlation

The correlation coefficients described above measure the association between two variables ignoring any other variables in the system. Suppose there are three variables X, Y and Z as shown in the path diagram below.



The association between Y and Z is made up of the direct association between Y and Z and the association caused by the path through X, that is the association of both Y and Z with the third variable X. For example if Z and Y were cholesterol level and blood pressure and X were age since both blood pressure and cholesterol level may increase with age the correlation between blood pressure and cholesterol level eliminating the effect of age is required.

The correlation between two variables eliminating the effect of a third variable is know as the partial correlation. If ρ_{zy}, ρ_{zx} and ρ_{xy} represent the correlations between x, y and z then the partial correlation between Z and Y given X is

$$\frac{\rho_{zy}-\rho_{zx}\rho_{xy}}{\sqrt{(1-\rho_{zx}^2)(1-\rho_{xy}^2)}}.$$

The partial correlation is then estimated by using product-moment correlation coefficients.

In general, let a set of variables be partitioned into two groups Y and X with n_y variables in Y and n_x variables in X and let the variance-covariance matrix of all n_y+n_x variables be partitioned into

$$\begin{bmatrix}\Sigma_{xx} & \Sigma_{yx}\\ \Sigma_{xy} & \Sigma_{yy}\end{bmatrix}.$$

Then the variance-covariance of Y conditional on fixed values of the X variables is given by

$$\Sigma_{y|x} = \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}.$$

The partial correlation matrix is then computed by standardising $\Sigma_{y|x}$.

2.1.4 Robust estimation of correlation coefficients

The product-moment correlation coefficient can be greatly affected by the presence of a few extreme observations or outliers. There are robust estimation procedures which aim to decrease the effect of extreme values.

Mathematically these methods can be described as follows. A robust estimate of the variance-covariance matrix, C, can be written as

$$C = \tau^2(A^TA)^{-1}$$

where τ^2 is a correction factor to give an unbiased estimator if the data is Normal and A is a lower triangular matrix. Let x_i be the vector of values for the ith observation and let $z_i = A(x_i - \theta)$, θ being a robust estimate of location, then θ and A are found as solutions to

$$\frac{1}{n}\sum_{i=1}^{n} w(\|z_i\|_2)z_i = 0$$

and

$$\frac{1}{n}\sum_{i=1}^{n} w(\|z_i\|_2)z_iz_i^T - v(\|z_i\|_2)I = 0,$$

where $w(t)$, $u(t)$ and $v(t)$ are functions such that they return a value of 1 for reasonable values of t and decreasing values for large t. The correlation matrix can then be calculated from the variance-covariance matrix. If w, u, and v returned 1 for all values then the product-moment correlation coefficient would be calculated.

2.2 Regression

2.2.1 Aims of regression modelling

In regression analysis the relationship between one specific random variable, the **dependent** or **response variable**, and one or more known variables, called the **independent variables** or **covariates**, is studied. This relationship is represented by a mathematical model, or an equation, which associates the dependent variable with the independent variables, together with a set of relevant assumptions. The independent variables are related to the dependent variable by a function, called the **regression function**, which involves a set of unknown **parameters**. Values of the parameters which give the best fit for a given set of data are obtained; these values are known as the **estimates** of the parameters.

The reasons for using a regression model are twofold. The first is to obtain a **description** of the relationship between the variables as an indicator of possible causality. The second reason is to **predict** the value of the dependent variable from a set of values of the independent variables. Accordingly, the most usual statistical problems involved in regression analysis are:

(i) to obtain best estimates of the unknown regression parameters;

(ii) to test hypotheses about these parameters;

(iii) to determine the adequacy of the assumed model; and

(iv) to verify the set of relevant assumptions.

2.2.2 Linear regression models

When the regression model is linear in the parameters (but not necessarily in the independent variables), then the regression model is said to be linear; otherwise the model is classified as nonlinear.

The most elementary form of regression model is the **simple linear regression** of the dependent variable, Y, on a single independent variable, x, which takes the form

$$E(Y) = \beta_0 + \beta_1 x \tag{1}$$

where $E(Y)$ is the expected or average value of Y and β_0 and β_1 are the parameters whose values are to be estimated, or, if the regression is required to pass through the origin (i.e., no constant term),

$$E(Y) = \beta_1 x \tag{2}$$

where β_1 is the only unknown parameter.

An extension of this is **multiple linear regression** in which the dependent variable, Y, is regressed on the p ($p > 1$) independent variables, $x_1, x_2, \ldots, x_p$, which takes the form

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \tag{3}$$

where $\beta_1, \beta_2, \ldots, \beta_p$ and β_0 are the unknown parameters.

A special case of multiple linear regression is **polynomial linear regression**, in which the p independent variables are in fact powers of the same single variable x (i.e., $x_j = x^j$, for $j = 1, 2, \ldots, p$).

In this case, the model defined by (3) becomes

$$E(Y) = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_p x^p. \tag{4}$$

There are a great variety of **nonlinear regression models**; one of the most common is **exponential regression**, in which the equation may take the form

$$E(Y) = a + be^{cx}. \tag{5}$$

It should be noted that equation (4) represents a **linear** regression, since even though the equation is not linear in the independent variable, x, it is linear in the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$, whereas the regression model of equation (5) is **nonlinear**, as it is nonlinear in the parameters (a, b and c).

2.2.3 Fitting the regression model – least-squares estimation

The method used to determine values for the parameters is, based on a given set of data, to minimize the sums of squares of the differences between the observed values of the dependent variable and the values predicted by the regression equation for that set of data – hence the term **least-squares** estimation. For example, if a regression model of the type given by equation (3), namely

$$E(Y) = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p,$$

where $x_0 = 1$ for all observations, is to be fitted to the n data points

$$\begin{array}{c} (x_{01}, x_{11}, x_{21}, \ldots, x_{p1}, y_1) \\ (x_{02}, x_{12}, x_{22}, \ldots, x_{p2}, y_2) \\ \vdots \\ (x_{0n}, x_{1n}, x_{2n}, \ldots, x_{pn}, y_n) \end{array} \tag{6}$$

such that

$$y_i = \beta_0 x_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + e_i, \quad i = 1, 2, \ldots, n$$

where e_i are unknown independent random errors with $E(e_i) = 0$ and $\mathrm{var}(e_i) = \sigma^2$, σ^2 being a constant, then the method used is to calculate the estimates of the regression parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ by minimizing

$$\sum_{i=1}^{n} e_i^2. \tag{7}$$

If the errors do not have constant variance, i.e.,

$$\mathrm{var}(e_i) = \sigma_i^2 = \frac{\sigma^2}{w_i}$$

then **weighted least-squares** estimation is used in which

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

is minimized. For a more complete discussion of these least-squares regression methods, and details of the mathematical techniques used, see Draper and Smith (1985) or Kendall and Stuart (1973).

2.2.4 Regression models and designed experiments

One application of regression models is in the analysis of experiments. In this case the model relates the dependent variable to qualitative independent variables known as **factors**. Factors may take a number of different values known as **levels**. For example, in an experiment in which one of four different treatments is applied, the model will have one factor with four levels. Each level of the factor can be represented by a dummy variable taking the values 0 or 1. So in the example there are four dummy variables x_j, for $j = 1, 2, 3, 4$ such that:

$$\begin{aligned} x_{ij} &= 1 \text{ if the } i\text{th observation received the } j\text{th treatment} \\ &= 0 \text{ otherwise,} \end{aligned}$$

along with a variable for the mean x_0:

$$x_{i0} = 1 \text{ for all } i.$$

If there were 7 observations the data would be:

Treatment	Y	x_0	x_1	x_2	x_3	x_4
1	y_1	1	1	0	0	0
2	y_2	1	0	1	0	0
2	y_3	1	0	1	0	0
3	y_4	1	0	0	1	0
3	y_5	1	0	0	1	0
4	y_6	1	0	0	0	1
4	y_7	1	0	0	0	1

Models which include factors are sometimes known as **General Linear (Regression) Models**. When dummy variables are used it is common for the model not to be of full rank. In the case above, the model would not be of full rank because

$$x_{i4} = x_{i0} - x_{i1} - x_{i2} - x_{i3}, \quad i = 1, 2, \ldots, 7.$$

This means that the effect of x_4 cannot be distinguished from the combined effect of x_0, x_1, x_2 and x_3. This is known as **aliasing**. In this situation, the aliasing can be deduced from the experimental design and as a result the model to be fitted; in such situations it is known as intrinsic aliasing. In the example above no matter how many times each treatment is replicated (other than 0) the aliasing will still be present. If the aliasing is due to a particular data set to which the model is to be fitted then it is known as extrinsic aliasing. If in the example above observation 1 was missing then the x_1 term would also be aliased. In general intrinsic aliasing may be overcome by changing the model, e.g., remove x_0 or x_1 from the model, or by introducing constraints on the parameters, e.g., $\beta_1 + \beta_2 + \beta_3 + \beta_4 = 0$.

If aliasing is present then there will no longer be a unique set of least-squares estimates for the parameters of the model but the fitted values will still have a unique estimate. Some linear functions of the parameters will also have unique estimates; these are known as **estimable functions**. In the example given above the functions $(\beta_0 + \beta_1)$ and $(\beta_2 - \beta_3)$ are both estimable.

2.2.5 Selecting the regression model

In many situations there are several possible independent variables, not all of which may be needed in the model. In order to select a suitable set of independent variables, two basic approaches can be used.

(a) All possible regressions

In this case all the possible combinations of independent variables are fitted and the one considered the best selected. To choose the best, two conflicting criteria have to be balanced. One is the fit of the model as measured by the residual sum of squares. This will decrease as more variables are added to the model. The second criterion is the desire to have a model with a small number of significant terms. To aid in the choice of model, statistics such as R^2, which gives the proportion of variation explained by the model, and C_p, which tries to balance the size of the residual sum of squares against the number of terms in the model, can be used.

(b) Stepwise model building

In stepwise model building the regression model is constructed recursively, adding or deleting the independent variables one at a time. When the model is built up the procedure is known as forward selection. The first step is to choose the single variable which is the best predictor. The second independent variable to be added to the regression equation is that which provides the best fit in conjunction with the first variable. Further variables are then added in this recursive fashion, adding at each step the optimum variable, given the other variables already in the equation. Alternatively, backward elimination can be used. This is when all variables are added and then the variables dropped one at a time, the variable dropped being the one which has the least effect on the fit of the model at that stage. There are also hybrid techniques which combine forward selection with backward elimination.

2.2.6 Examining the fit of the model

Having fitted a model two questions need to be asked: first, ‘are all the terms in the model needed?’ and second, ‘is there some systematic lack of fit?’. To answer the first question either confidence intervals can be computed for the parameters or t-tests can be calculated to test hypotheses about the regression parameters – for example, whether the value of the parameter, β_k, is significantly different from a specified value, b_k (often zero). If the estimate of β_k is $\hat{\beta}_k$ and its standard error is $\mathrm{se}(\hat{\beta}_k)$ then the t-statistic is

$$\frac{\hat{\beta}_k - b_k}{\sqrt{\mathrm{se}(\hat{\beta}_k)}}.$$

It should be noted that both the tests and the confidence intervals may not be independent. Alternatively F-tests based on the residual sums of squares for different models can also be used to test the significance of terms in the model. If model 1, giving residual sum of squares RSS_1 with degrees of freedom ν_1, is a sub-model of model 2, giving residual sum of squares RSS_2 with degrees of freedom ν_2, i.e., all terms in model 1 are also in model 2, then to test if the extra terms in model 2 are needed the F-statistic

$$F = \frac{(RSS_1 - RSS_2)/(\nu_1 - \nu_2)}{RSS_2/\nu_2}$$

may be used. These tests and confidence intervals require the additional assumption that the errors, e_i, are Normally distributed.

To check for systematic lack of fit the residuals, $r_i = y_i - \hat{y}_i$, where $\hat{y}_i$ is the fitted value, should be examined. If the model is correct then they should be random with no discernable pattern. Due to the way they are calculated the residuals do not have constant variance. Now the vector of fitted values can be written as a linear combination of the vector of observations of the dependent variable, y, $\hat{y} = Hy$. The variance-covariance matrix of the residuals is then $(I - H)\sigma^2$, I being the identity matrix. The diagonal elements of H, h_{ii}, can therefore be used to standardize the residuals. The h_{ii} are a measure of the effect of the ith observation on the fitted model and are sometimes known as **leverages**.

If the observations were taken serially the residuals may also be used to test the assumption of the independence of the e_i and hence the independence of the observations.

2.2.7 Computational methods

Let X be the n by p matrix of independent variables and y be the vector of values for the dependent variable. To find the least-squares estimates of the vector of parameters, $\hat{\beta}$, the QR decomposition of X is found, i.e.,

$$X = QR^*$$

where $R^* = \begin{pmatrix} R \\ 0 \end{pmatrix}$, R being a p by p upper triangular matrix, and Q is a n by n orthogonal matrix. If R is of full rank then $\hat{\beta}$ is the solution to

$$R\hat{\beta} = c_1$$

where $c = Q^T y$ and c_1 is the first p rows of c. If R is not of full rank, a solution is obtained by means of a singular value decomposition (SVD) of R,

$$R = Q_* \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} P^T,$$

where D is a k by k diagonal matrix with non-zero diagonal elements, k being the rank of R, and Q_* and P are p by p orthogonal matrices. This gives the solution

$$\hat{\beta} = P_1 D^{-1} Q_{*_1}^T c_1,$$

P_1 being the first k columns of P and Q_{*_1} being the first k columns of Q_*.

This will be only one of the possible solutions. Other estimates may be obtained by applying constraints to the parameters. If weighted regression with a vector of weights w is required then both X and y are premultiplied by $w^{1/2}$.

The method described above will, in general, be more accurate than methods based on forming $(X^T X)$, (or a scaled version), and then solving the equations

$$(X^T X)\hat{\beta} = X^T y.$$

2.2.8 Robust estimation

Least-squares regression can be greatly affected by a small number of unusual, atypical, or extreme observations. To protect against such occurrences, robust regression methods have been developed. These methods aim to give less weight to an observation which seems to be out of line with the rest of the data given the model under consideration. That is to seek to bound the influence. For a discussion of influence in regression, see Hampel *et al.* (1986) and Huber (1981).

There are two ways in which an observation for a regression model can be considered atypical. The values of the independent variables for the observation may be atypical or the residual from the model may be large.

The first problem of atypical values of the independent variables can be tackled by calculating weights for each observation which reflect how atypical it is, i.e., a strongly atypical observation would have a low weight. There are several ways of finding suitable weights; some are discussed in Hampel *et al.* (1986).

The second problem is tackled by bounding the contribution of the individual e_i to the criterion to be minimized. When minimizing (7) a set of linear equations is formed, the solution of which gives the least-squares estimates. The equations are

$$\sum_{i=1}^{n} e_i x_{ij} = 0 \quad j = 0, 1, \ldots, k.$$

These equations are replaced by

$$\sum_{i=1}^{n} \psi(e_i/\sigma) x_{ij} = 0 \quad j = 0, 1, \ldots, k, \tag{8}$$

where σ^2 is the variance of the e_i, and ψ is a suitable function which down weights large values of the standardized residuals e_i/σ. There are several suggested forms for ψ, one of which is Huber's function,

$$\psi(t) = \begin{cases} -c, & t < c \\ t, & |t| \leq c \\ c, & t > c \end{cases} \tag{9}$$

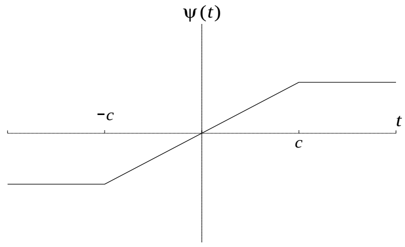


Figure 2

The solution to (8) gives the M-estimates of the regression coefficients. The weights can be included in (8) to protect against both types of extreme value. The parameter σ can be estimated by the median absolute deviations of the residuals or as a solution to, in the unweighted case,

$$\sum_{i=1}^{n} \chi(e_i/\hat{\sigma}) = (n-k)\beta,$$

where χ is a suitable function and β is a constant chosen to make the estimate unbiased. χ is often chosen to be $\psi^2/2$ where ψ is given in (9). Another form of robust regression is to minimize the sum of absolute deviations, i.e.,

$$\sum_{i=1}^{n} |e_i|.$$

For details of robust regression, see Hampel *et al.* (1986) and Huber (1981).

Robust regressions using least absolute deviations can be computed using routines in Chapter e02.

2.2.9 Generalized linear models

Generalized linear models are an extension of the general linear regression model discussed above. They allow a wide range of models to be fitted. These included certain non-linear regression models, logistic and probit regression models for binary data, and log-linear models for contingency tables. A generalized linear model consists of three basic components:

(a) A suitable distribution for the dependent variable Y. The following distributions are common:

  (i) Normal

  (ii) binomial

  (iii) Poisson

  (iv) gamma

  In addition to the obvious uses of models with these distributions it should be noted that the Poisson distribution can be used in the analysis of contingency tables while the gamma distribution can be used to model variance components. The effect of the choice of the distribution is to define the relationship between the expected value of Y, $E(Y) = \mu$, and its variance and so a generalized linear model with one of the above distributions may be used in a wider context when that relationship holds.

(b) A linear model $\eta = \sum \beta_j x_j$, η is known as a **linear predictor**.

(c) A link function $g(\cdot)$ between the expected value of Y and the **linear predictor**, $g(\mu) = \eta$. The following link functions are available:

  For the binomial distribution ϵ, observing y out of t:

  (i) logistic link: $\eta = \log\left(\frac{\mu}{t-\mu}\right)$;

  (ii) probit link: $\eta = \Phi^{-1}\left(\frac{\mu}{t}\right)$;

(iii) complementary log-log: $\eta = \log\left(-\log\left(1 - \frac{\mu}{t}\right)\right)$.

For the Normal, Poisson, and gamma distributions:

(i) exponent link: $\eta = \mu^a$, for a constant a;

(ii) identity link: $\eta = \mu$;

(iii) log link: $\eta = \log\mu$;

(iv) square root link: $\eta = \sqrt{\mu}$;

(v) reciprocal link: $\eta = \frac{1}{\mu}$.

For each distribution there is a **canonical link**. For the canonical link there exist sufficient statistics for the parameters. The canonical links are:

(i) Normal – identity;

(ii) binomial – logistic;

(iii) Poisson – logarithmic;

(iv) gamma – reciprocal.

For the general linear regression model described above the three components are:

(i) Distribution – Normal;

(ii) Linear model – $\sum\beta_j x_j$;

(iii) Link – identity.

The model is fitted by **maximum likelihood**; this is equivalent to least-squares in the case of the Normal distribution. The residual sums of squares used in regression models is generalized to the concept of **deviance**. The deviance is the logarithm of the ratio of the likelihood of the model to the full model in which $\hat{\mu}_i = y_i$, where $\hat{\mu}_i$ is the estimated value of μ_i. For the Normal distribution the deviance is the residual sum of squares. Except for the case of the Normal distribution with the identity link, the χ^2 and F-tests based on the deviance are only approximate; also the estimates of the parameters will only be approximately Normally distributed. Thus only approximate z- or t-tests may be performed on the parameter values and approximate confidence intervals computed.

The estimates are found by using an **iterative weighted least-squares** procedure. This is equivalent to the Fisher scoring method in which the Hessian matrix used in the Newton–Raphson method is replaced by its expected value. In the case of canonical links the Fisher scoring method and the Newton–Raphson method are identical. Starting values for the iterative procedure are obtained by replacing the μ_i by y_i in the appropriate equations.

3 Recommendations on Choice and Use of Available Functions

3.1 Correlation

3.1.1 Product-moment correlation

Let SS_x be the sum of squares of deviations from the mean, $\bar{x}$, for the variable x for a sample of size n, i.e.,

$$SS_x = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

and let SC_{xy} be the cross-products of deviations from the means, $\bar{x}$ and $\bar{y}$, for the variables x and y for a sample of size n, i.e.,

$$SC_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

Then the sample covariance of x and y is

$$\mathrm{cov}(x,y) = \frac{SC_{xy}}{(n-1)}$$

and the product-moment correlation coefficient is

$$r = \frac{\mathrm{cov}(x,y)}{\sqrt{\mathrm{var}(x)\,\mathrm{var}(y)}} = \frac{SC_{xy}}{\sqrt{SS_x SS_y}}.$$

nag_sum_sqs (g02buc) computes the sample sums of squares and cross-products deviations from the means (optionally weighted). nag_sum_sqs_update (g02btc) updates the sample sums of squares and cross-products and deviations from the means by the addition/deletion of a (weighted) observation. nag_cov_to_corr (g02bwc) computes the product-moment correlation coefficients from the sample sums of squares and cross-products of deviations from the means. The three routines compute only the upper triangle of the correlation matrix which is stored in a one-dimensional array in packed form. nag_corr_cov (g02bxc) computes both the (optionally weighted) covariance matrix and the (optionally weighted) correlation matrix. These are returned in two-dimensional arrays. (Note that nag_sum_sqs_update (g02btc) and nag_sum_sqs (g02buc) can be used to compute the sums of squares from zero.)

3.1.2 Non-parametric correlation

nag_ken_spe_corr_coeff (g02brc) computes Kendall and/or Spearman non-parametric rank correlation coefficients. The function allows for a subset of variables to be selected and for observations to be excluded from the calculations if, for example, they contain missing values.

3.1.3 Partial correlation

nag_partial_corr (g02byc) computes a matrix of partial correlation coefficients from the correlation coefficients or variance-covariance matrix returned by nag_corr_cov (g02bxc).

3.1.4 Robust correlation

nag_robust_m_corr_user_fn (g02hlc) and nag_robust_m_corr_user_fn_no_derr (g02hmc) compute robust esimates of the variance-covariance matrix by solving the equations

$$\frac{1}{n}\sum_{i=1}^{n} w(\|z_i\|_2)z_i = 0$$

and

$$\frac{1}{n}\sum_{i=1}^{n} u(\|z_i\|_2)z_i z_i^T - v(\|z_i\|_2)I = 0,$$

as described in Section 2.1.3 for user-supplied functions w and u. Two options are available for v, either $v(t) = 1$ for all t or $v(t) = u(t)$.

nag_robust_m_corr_user_fn_no_derr (g02hmc) requires only the function w and u to be supplied while nag_robust_m_corr_user_fn (g02hlc) also requires their derivatives.

In general nag_robust_m_corr_user_fn (g02hlc) will be considerably faster than nag_robust_m_corr_user_fn_no_derr (g02hmc) and should be used if derivatives are available.

nag_robust_corr_estim (g02hkc) computes a robust variance-covariance matrix for the following functions:

$$\begin{aligned} u(t) &= a_u/t^2 \text{ if } t < a_u^2 \\ u(t) &= 1 \text{ if } a_u^2 \le t \le b_u^2 \\ u(t) &= b_u/t^2 \text{ if } t > b_u^2 \end{aligned}$$

and

$$\begin{aligned} w(t) &= 1 \text{ if } t \le c_w \\ w(t) &= c_w/t \text{ if } t > c_w \end{aligned}$$

for constants a_u, b_u and c_w.

These functions solve a minimax space problem considered by Huber (1981). The values of a_u, b_u and c_w are calculated from the fraction of gross errors; see Hampel *et al.* (1986) and Huber (1981).

To compute a correlation matrix from the variance-covariance matrix nag_cov_to_corr (g02bwc) may be used.

3.2 Regression

3.2.1 Simple linear regression

Two functions are provided for simple linear regression. The function nag_simple_linear_regression (g02cac) calculates the parameter estimates for a simple linear regression with or without a constant term. The function nag_regress_confid_interval (g02cbc) calculates fitted values, residuals and confidence intervals for both the fitted line and individual observations. This function produces the information required for various regression plots.

3.2.2 Multiple linear regression – general linear model

nag_regsn_mult_linear (g02dac)
: fits a general linear regression model using the QR method and an SVD if the model is not of full rank. The results returned include: residual sum of squares, parameter estimates, their standard errors and variance-covariance matrix, residuals and leverages. There are also several routines to modify the model fitted by nag_regsn_mult_linear (g02dac) and to aid in the interpretation of the model.

nag_regsn_mult_linear_addrem_obs (g02dcc)
: adds or deletes an observation from the model.

nag_regsn_mult_linear_upd_model (g02ddc)
: computes the parameter estimates, and their standard errors and variance-covariance matrix for a model that is modified by nag_regsn_mult_linear_addrem_obs (g02dcc), nag_regsn_mult_linear_add_var (g02dec) or nag_regsn_mult_linear_delete_var (g02dfc).

nag_regsn_mult_linear_add_var (g02dec)
: adds a new variable to a model.

nag_regsn_mult_linear_delete_var (g02dfc)
: drops a variable from a model.

nag_regsn_mult_linear_newyvar (g02dgc)
: fits the regression to a new dependent variable, i.e., keeping the same independent variables.

nag_regsn_mult_linear_tran_model (g02dkc)
: calculates the estimates of the parameters for a given set of constraints, (e.g., parameters for the levels of a factor sum to zero) for a model which is not of full rank and the SVD has been used.

nag_regsn_mult_linear_est_func (g02dnc)
: calculates the estimate of an estimable function and its standard error.

Note: nag_regsn_mult_linear_add_var (g02dec) also allows the user to initialise a model building process and then to build up the model by adding variables one at a time.

3.2.3 Selecting regression models

To aid the selection of a regression model the following routines are available.

nag_all_regsn (g02eac)
: computes the residual sums of squares for all possible regressions for a given set of dependent variables. The routine allows some variables to be forced into all regressions.

nag_cp_stat (g02ecc)
: computes the values of R^2 and C_p from the residual sums of squares as provided by nag_all_regsn (g02eac).

nag_step_regsn (g02eec)
: enables the user to fit a model by forward selection. The user may call nag_step_regsn (g02eec)

a number of times. At each call the routine will calculate the changes in the residual sum of squares from adding each of the variables not already included in the model, select the variable which gives the largest change and then if the change in residual sum of squares meets the given criterion will add it to the model.

3.2.4 Residuals

nag_regsn_std_resid_influence (g02fac) computes the following standardized residuals and measures of influence for the residuals and leverages produced by nag_regsn_mult_linear (g02dac):

(i) Internally studentized residual;

(ii) Externally studentized residual;

(iii) Cook's D statistic;

(iv) Atkinson's T statistic.

nag_durbin_watson_stat (g02fcc) computes the Durbin–Watson test statistic and bounds for its significance to test for serial correlation in the errors, e_i.

3.2.5 Robust regression

For robust regression using M-estimates instead of least-squares the routine nag_robust_m_regsn_estim (g02hac) will generally be suitable. nag_robust_m_regsn_estim (g02hac) provides a choice of four ψ-functions (Huber's, Hampel's, Andrew's and Tukey's) plus two different weighting methods and the option not to use weights. If other weights or different ψ-functions are needed the routine nag_robust_m_regsn_user_fn (g02hdc) may be used. nag_robust_m_regsn_user_fn (g02hdc) requires the user to supply weights, if required, and also routines to calculate the ψ-function and, optionally, the χ-function. nag_robust_m_regsn_wts (g02hbc) can be used in calculating suitable weights. The routine nag_robust_m_regsn_param_var (g02hfc) can be used after a call to nag_robust_m_regsn_user_fn (g02hdc) in order to calculate the variance-covariance estimate of the estimated regression coefficients.

For robust regression, using least absolute deviation, nag_lone_fit (e02gac) can be used.

3.2.6 Generalized linear models

There are four routines for fitting generalized linear models. The output includes: the deviance, parameter estimates and their standard errors, fitted values, residuals and leverages. The routines are:

nag_glm_normal (g02gac) – Normal distribution

nag_glm_binomial (g02gbc) – binomial distribution

nag_glm_poisson (g02gcc) – Poisson distribution

nag_glm_gamma (g02gdc) – gamma distribution

While nag_glm_normal (g02gac) can be used to fit linear regression models (i.e., by using an identity link) this is not recomended as nag_regsn_mult_linear (g02dac) will fit these models more efficiently. nag_glm_poisson (g02gcc) can be used to fit log-linear models to contingency tables.

In addition to the routines to fit the models there are two routines to aid the interpretation of the model if a model which is not of full rank has been fitted, i.e., aliasing is present.

nag_glm_tran_model (g02gkc)
computes parameter estimates for a set of constraints, (e.g., sum of effects for a factor is zero), from the SVD solution provided by the fitting routine.

nag_glm_est_func (g02gnc)
calculates an estimate of an estimable function along with its standard error.

3.2.7 Polynomial regression and non-linear regression

No routines are currently provided in this chapter for polynomial regression. Users wishing to perform polynomial regressions do however have three alternatives: they can use the multiple linear regression routines, nag_regsn_mult_linear (g02dac), with a set of independent variables which are in fact simply the same single variable raised to different powers, or they can use the routine nag_dummy_vars (g04eac) to compute orthogonal polynomials which can then be used with nag_regsn_mult_linear (g02dac), or they can use the routines in Chapter e02 (Curve and Surface Fitting) which fit polynomials to sets of data points using the techniques of orthogonal polynomials. This latter course is to be preferred, since it is more efficient and liable to be more accurate, but in some cases more statistical information may be required than is provided by those routines, and it may be necessary to use the routines of this chapter.

More general nonlinear regression models may be fitted using the optimization routines in Chapter e04, which contains routines to minimize the function

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

where the regression parameters are the variables of the minimization problem.

4 Index

5 Functions Withdrawn or Scheduled for Withdrawal

None.

6 References

Atkinson A C (1986) *Plots, Transformations and Regressions* Clarendon Press, Oxford

Churchman C W and Ratoosh P (1959) *Measurement Definitions and Theory* Wiley

Cook R D and Weisberg S (1982) *Residuals and Influence in Regression* Chapman and Hall

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

Hammarling S (1985) The singular value decomposition in multivariate statistics *SIGNUM Newsl.* **20 (3)** 2–25

Hampel F R, Ronchetti E M, Rousseeuw P J and Stahel W A (1986) *Robust Statistics. The Approach Based on Influence Functions* Wiley

Hays W L (1970) *Statistics* Holt, Rinehart and Winston

Huber P J (1981) *Robust Statistics* Wiley

Kendall M G and Stuart A (1973) *The Advanced Theory of Statistics (Volume 2)* (3rd Edition) Griffin

McCullagh P and Nelder J A (1983) *Generalized Linear Models* Chapman and Hall

Searle S R (1971) *Linear Models* Wiley

Siegel S (1956) *Non-parametric Statistics for the Behavioral Sciences* McGraw–Hill

Weisberg S (1985) *Applied Linear Regression* Wiley