NAG 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 Library which cover similar problems are Chapters E02 and 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 nonparametric 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, you are referred to Churchman and Ratoosh (1959) and Hays (1970).

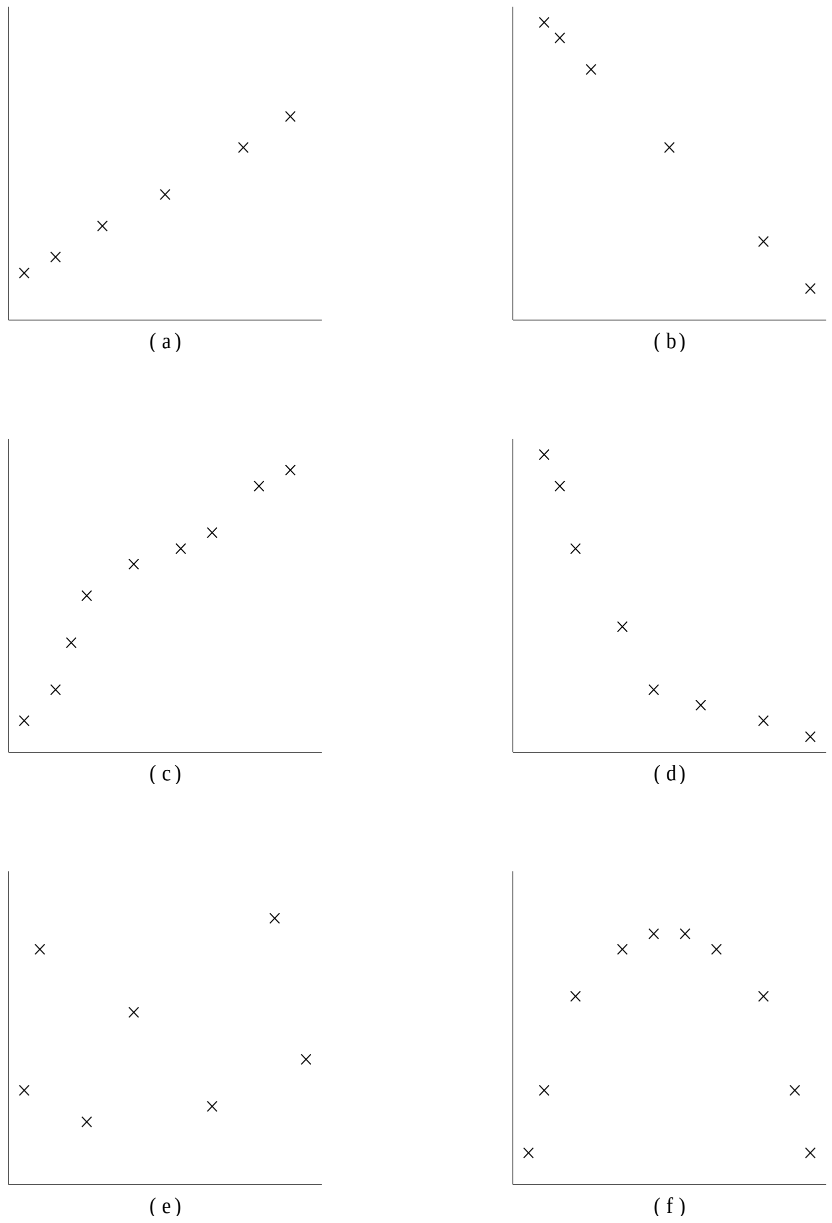


Figure 1

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 nonparametric 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 nonparametric 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

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

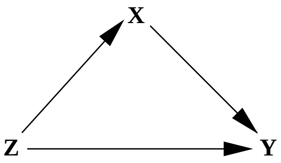
when tested against the null hypothesis that there is no correlation between the two variables; the coefficient should therefore be divided by this standard deviation and tested against the standard Normal distribution, $\mathrm{N}(0,1)$.

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 you.

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 known 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

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

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 standardizing $\Sigma_{y|x}$.

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

$$C = \tau^2\left(A^{\mathrm{T}}A\right)^{-1}$$

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

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

and

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

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.

2.1.5 Missing values

When there are missing values in the data these may be handled in one of two ways. Firstly, if a case contains a missing observation for any variable, then that case is omitted in its entirety from all calculations; this may be termed **casewise** treatment of missing data. Secondly, if a case contains a missing observation for any variable, then the case is omitted from only those calculations involving the variable for which the value is missing; this may be called **pairwise** treatment of missing data. Pairwise deletion of missing data has the advantage of using as much of the data as possible in the computation of each coefficient. In extreme circumstances, however, it can have the disadvantage of producing coefficients which are based on a different number of cases, and even on different selections of cases or samples; furthermore, the 'correlation' matrices formed in this way need not necessarily be positive semidefinite, a requirement for a correlation matrix. Casewise deletion of missing data generally causes fewer cases to be used in the calculation of the coefficients than does pairwise deletion. How great this difference is will obviously depend on the distribution of the missing data, both among cases and among variables.

Pairwise treatment does therefore use more information from the sample, but should not be used without careful consideration of the location of the missing observations in the data matrix, and the consequent effect of processing the missing data in that fashion.

2.1.6 Nearest Correlation Matrix

A correlation matrix is, by definition, a symmetric, positive semidefinite matrix with unit diagonals and all elements in the range $[-1, 1]$.

In practice, rather than having a true correlation matrix, you may find that you have a matrix of pairwise correlations. This usually occurs in the presence of missing values, when the missing values are treated in a pairwise fashion as discussed in Section 2.1.5. Matrices constructed in this way may not be not positive semidefinite, and therefore are not a valid correlation matrix. However, a valid correlation matrix can be calculated that is in some sense 'close' to the original.

Given an $n \times n$ matrix, G, there are a number of available ways of computing the 'nearest' correlation matrix, Σ to G:

(a) Frobenius Norm

Find Σ such that

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \left(s_{ij} - \sigma_{ij}\right)^2$$

is minimized.

Where S is the symmetric matrix defined as $S = \frac{1}{2}(G + G^{\mathrm{T}})$ and s_{ij} and σ_{ij} denotes the elements of S and Σ respectively.

A weighted Frobenius norm can also be used. The term being summed across therefore becomes $w_i w_j \left(s_{ij} - \sigma_{ij}\right)^2$ if row and column weights are being used or $w_{ij}\left(s_{ij} - \sigma_{ij}\right)^2$ when element-wise weights are used.

(b) Factor Loading Method

This method is similar to (a) in that it finds a Σ that is closest to S in the Frobenius norm. However, it also ensures that Σ has a k-factor structure, that is Σ can be written as

$$\Sigma = XX^{\mathrm{T}} + \operatorname{diag}\left(I - XX^{\mathrm{T}}\right)$$

where I is the identity matrix and X has n rows and k columns.

X is often referred to as the factor loading matrix. This problem primarily arises when a factor model $\xi = X\eta + D\epsilon$ is used to describe a multivariate time series or collateralized debt obligations. In this model $\eta \in \mathbb{R}^k$ and $\xi \in \mathbb{R}^n$ are vectors of independent random variables having zero mean and unit variance, with η and ϵ independent of each other, and $X \in \mathbb{R}^{n \times k}$ with $D \in \mathbb{R}^{n \times n}$ diagonal. In the case of modelling debt obligations ξ can, for example, model the equity returns of n different companies of a portfolio where η describes k factors influencing all companies, in contrast to the elements of ϵ having only an effect on the equity of the corresponding company. With this model the complex behaviour of a portfolio, with potentially thousands of equities, is captured by looking at the major factors driving the behaviour.

The number of factors usually chosen is a lot smaller than n, perhaps between 1 and 10, yielding a large reduction in the complexity. The number of the factors, k, which yields a matrix X such that $\|G - XX^{\mathrm{T}} + \operatorname{diag}(I - XX^{\mathrm{T}})\|_F$ is within a required tolerance can also be determined, by experimenting with the input k and comparing the norms.

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 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:

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

along with a variable for the mean x_0:

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

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

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

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

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 dataset 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.

2.2.3 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 which will improve 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. Depending on how the model is fit, 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 to aid in the choice of model.

(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.

2.3 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

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

where $\beta_1, \beta_2, \ldots, \beta_p$ and β_0 are the unknown parameters. Multiple linear regression models test include factors are sometimes known as **General Linear (Regression) Models**.

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

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

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).

2.3.1 Fitting the regression model – least squares estimation

One 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

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

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

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

such that

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

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

$$\sum_{i=1}^{n} e_i^2. \qquad (7)$$

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).

2.3.2 Computational methods for least squares regression

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 an n by n orthogonal matrix. If R is of full rank then $\hat{\beta}$ is the solution to

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

where $c = Q^{\mathrm{T}} y$ 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,

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

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

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

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 $(X^{\mathrm{T}}X)$, (or a scaled version), and then solving the equations

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

2.3.3 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 $\mathrm{se}\left(\hat{\beta}_k\right)$ then the t-statistic is

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

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 discernible 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.

2.3.4 Ridge regression

When data on predictor variables x are multicollinear, **ridge regression** models provide an alternative to variable selection in the multiple regression model. In the ridge regression case, parameter estimates in the linear model are found by penalised least squares:

$$\sum_{i=1}^{n}\left[\left(\sum_{j=1}^{p}x_{ij}\hat{\beta}_j\right) - y_i\right]^2 + h\sum_{j=1}^{p}\hat{\beta}_j^2, \quad h \in \mathbb{R}^{+},$$

where the value of the ridge parameter h controls the trade-off between the goodness-of-fit and smoothness of a solution.

2.4 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,

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

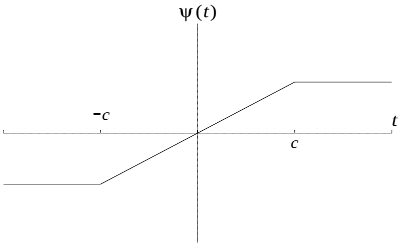


Figure 3

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.

2.5 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 nonlinear 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.

2.6 Linear Mixed Effects Regression

In a standard linear model the independent (or explanatory) variables are assumed to take the same set of values for all units in the population of interest. This type of variable is called *fixed*. In contrast, an independent variable that fluctuates over the different units is said to be *random*. Modelling a variable as *fixed* allows conclusions to be drawn only about the particular set of values observed. Modelling a variable as *random* allows the results to be generalized to the different levels that may have been observed. In general, if the effects of the levels of a variable are thought of as being drawn from a probability distribution of such effects then the variable is *random*. If the levels are not a sample of possible levels then the variable is *fixed*. In practice many qualitative variables can be considered as having *fixed effects* and most blocking, sampling design, control and repeated measures as having *random effects*.

In a general linear regression model, defined by

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

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

X is an n by p design matrix of independent variables,

β is a vector of p unknown parameters,

and ϵ is a vector of n, independent and identically distributed, unknown errors, with $\epsilon\tilde{}N(0,\sigma^2)$,

there are p *fixed effects* (the β) and a single *random effect* (the error term ϵ).

An extension to the general linear regression model that allows for additional *random effects* is the linear mixed effects regression model, (sometimes called the variance components model). One parameterisation of a linear mixed effects model is

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

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

X is an n by p design matrix of *fixed* independent variables,

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

Z is an n by q design matrix of *random* independent variables,

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

ϵ is a vector of length n of unknown random errors,

and ν and ϵ are normally distributed with expectation zero and variance / covariance matrix defined by

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

The routines currently available in this chapter are restricted to cases where $R = \sigma_R^2 I$, I is the $n \times n$ identity matrix and G is a diagonal matrix. Given this restriction the random variables, Z, can be subdivided into $g \leq q$ groups containing one or more variables. The variables in the ith group are identically distributed with expectation zero and variance σ_i^2. 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, with $\gamma = \left\{\sigma_1^2, \sigma_2^2, \ldots, , , \sigma_{g-1}^2, \sigma_g^2, \sigma_R^2\right\}$. Rather than work directly with γ and the full likelihood function, γ is replaced by $\gamma^* = \left\{\sigma_1^2/\sigma_R^2, \sigma_2^2/\sigma_R^2, \ldots, \sigma_{g-1}^2/\sigma_R^2, \sigma_g^2/\sigma_R^2, 1\right\}$ and the profiled likelihood function is used instead.

The model parameters are estimated using an iterative method based on maximizing either the restricted (profiled) likelihood function or the (profiled) likelihood functions. Fitting the model via restricted maximum likelihood involves maximizing the function

$$-2l_R = \log(|V|) + (n-p)\log\left(r^{\mathrm{T}}V^{-1}r\right) + \log\left|X^{\mathrm{T}}V^{-1}X\right| + (n-p)(1+\log(2\pi/(n-p))) + (n-p).$$

Whereas fitting the model via maximum likelihood involves maximizing

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

In both cases

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

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

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

Case weights, W_c, can be incorporated into the model by replacing $X^{\mathrm{T}}X$ and $Z^{\mathrm{T}}Z$ with $X^{\mathrm{T}}W_cX$ and $Z^{\mathrm{T}}W_cZ$ respectively, for a diagonal weight matrix W_c.

2.7 Quantile Regression

Quantile regression is related to least squares regression in that both are interested in studying the relationship between a response variable and one or more independent or explanatory variables. However, whereas least squares regression is concerned with modelling the conditional mean of the dependent variable, quantile regression models the conditional τth quantile of the dependent variable, for some value of $\tau \in (0,1)$. So, for example, $\tau = 0.5$ would be the median.

Throughout this section we will be making use of the following definitions:

(a) If Z is a real valued random variable with distribution function F and density function f, such that

$$F(\alpha) = P(Z \leq \alpha) = \int_{-\infty}^{\alpha} f(z)\mathrm{dz}$$

then the τth quantile, α, can be defined as

$$\alpha = F^{-1}(\tau) = \inf\{z : F(z) \geq \tau\}, \tau \in (0,1).$$

(b) $I(L)$ denotes an indicator function taking the value 1 if the logical expression L is true and 0 otherwise, e.g., $I(z<0) = 1$ if $z<0$ and 0 if $z \geq 0$.

(c) y denotes a vector of n observations on the dependent (or response) variable, $y = \{y_i : i = 1,2,\ldots,n\}$.

(d) X denotes an $n \times p$ matrix of explanatory or independent variables, often referred to as the design matrix, and x_i denotes a column vector of length p which holds the ith row of X.

2.7.1 Finding a sample quantile as an optimization problem

Consider the piecewise linear loss function

$$\rho_\tau(z) = z(\tau - I(z<0))$$

The minimum of the expectation

$$E(\rho_\tau(z-\alpha)) = (\tau-1)\int_{-\infty}^{\alpha}(z-\alpha)f(z)\mathrm{dz} + \tau\int_{\alpha}^{\infty}(z-\alpha)f(z)\mathrm{dz}$$

can be obtained by using the integral rule of Leibnitz to differentiate with respect to z and then setting the result to zero, giving

$$(1-\tau)\int_{-\infty}^{\alpha} f(z)\mathrm{d}z - \int_{\alpha}^{\infty} f(z)\mathrm{d}z = F(\alpha) - \tau = 0$$

hence $\alpha = F^{-1}(\tau)$ when the solution is unique. If the solution is not unique then there exists a range of quantiles, each of which is equally valid. Taking the smallest value of such a range ensures that the empirical quantile function is left-continuous. Therefore obtaining the τth quantile of a distribution F can be achieved by minimizing the expected value of the loss function ρ_τ.

This idea of obtaining the quantile by solving an optimization problem can be extended to finding the τth sample quantile. Given a vector of n observed values, y, from some distribution the empirical distribution function, $F_n(\alpha) = n^{-1}\sum_{i=1}^{n} I(y_i \leq \alpha)$ provides an estimate of the unknown distribution function F giving an expected loss of

$$E(\rho_\tau(y-\alpha)) = n^{-1}\sum_{i=1}^{n}\rho_\tau(y_i-\alpha)$$

and therefore the problem of finding the τth sample quantile, $\hat{\alpha}(\tau)$, can be expressed as finding the solution to the problem

$$\underset{\alpha\in\mathbb{R}}{\text{minimize}}\sum_{i=1}^{n}\rho_\tau(y_i-\alpha)$$

effectively replacing the operation of sorting, usually required when obtaining a sample quantile, with an optimization.

2.7.2 From least squares to quantile regression

Given the vector y it is a well known result that the sample mean, $\hat{y}$, solves the least squares problem

$$\underset{\mu\in\mathbb{R}}{\text{minimize}}\sum_{i=1}^{n}(y_i-\mu)^2.$$

This result leads to least squares regression where, given design matrix X and defining the conditional mean of y as $\mu(X) = X\beta$, an estimate of β is obtained from the solution to

$$\underset{\beta\in\mathbb{R}^p}{\text{minimize}}\sum_{i=1}^{n}\left(y_i - x_i^{\mathrm{T}}\beta\right)^2.$$

Quantile regression can be derived in a similar manner by specifying the τth conditional quantile as $Q_y(\tau|X) = X\beta(\tau)$ and estimating $\beta(\tau)$ as the solution to

$$\underset{\beta\in\mathbb{R}^p}{\text{minimize}}\sum_{i=1}^{n}\rho_\tau\left(y_i - x_i^{\mathrm{T}}\beta\right). \qquad (10)$$

2.7.3 Quantile regression as a linear programming problem

By introducing $2n$ slack variables, $u = \{u_i : i = 1,2,\ldots,n\}$ and $v = \{u_i : i = 1,2,\ldots,n\}$, the quantile regression minimization problem, (10), can be expressed as a linear programming (LP) problem, with primal and associated dual formulations

(a) Primal form

$$\underset{(u,v,\beta)\in\mathbb{R}_+^n\times\mathbb{R}_+^n\times\mathbb{R}^p}{\text{minimize}}\ \tau e^{\mathrm{T}}u + (1-\tau)e^{\mathrm{T}}v \quad \text{subject to} \quad y = X\beta + u - v \qquad (11)$$

where e is a vector of length n, where each element is 1.

If r_i denotes the ith residual, $r_i = y_i - x_i^{\mathrm{T}}\beta$, then the slack variables, (u,v), can be thought as corresponding to the absolute value of the positive and negative residuals respectively with

$$u_i = \begin{cases} r_i \text{ if } r_i > 0 \\ 0 \text{ otherwise} \end{cases} \qquad v_i = \begin{cases} -r_i \text{ if } r_i < 0 \\ 0 \text{ otherwise} \end{cases}$$

(b) Dual form

The dual formulation of (11) is given by

$$\underset{d}{\text{maximize}} y^{\mathrm{T}} d \qquad \text{subject to} \qquad X^{\mathrm{T}} d = 0, d \in [\tau - 1, \tau]^n$$

which, on setting $a = d + (1 - \tau)e$, is equivalent to

$$\underset{a}{\text{maximize}} y^{\mathrm{T}} a \qquad \text{subject to} \qquad X^{\mathrm{T}} a = (1 - \tau) X^{\mathrm{T}} e, a \in [0, 1]^n \tag{12}$$

(c) Canonical form

Linear programming problems are often described in a standard way, called the canonical form. The canonical form of an LP problem is

$$\underset{z}{\text{minimize}} c^{\mathrm{T}} z \qquad \text{subject to} \qquad l_l \leq \begin{Bmatrix} z \\ Az \end{Bmatrix} \leq l_u.$$

Letting 0_p denote a vector of p zeros $\pm\infty_p$ denote a vector of p arbitarily small or large values, $I_{n \times n}$ denote the $n \times n$ identity matrix, $c = \{a, b\}$ denote the row vector constructed by concatenating the elements of vector b to the elements of vector a and $C = [A, B]$ denote the matrix constructed by concatenating the columns of matrix B onto the columns of matrix A then setting

$$\begin{array}{ll} c^{\mathrm{T}} = \{0_p, \tau e^{\mathrm{T}}, (1 - \tau) e^{\mathrm{T}}\} & z^{\mathrm{T}} = \{\beta^{\mathrm{T}}, u^{\mathrm{T}}, v^{\mathrm{T}}\} \\ A = [X, I_{n \times n}, -I_{n \times n}] & b = y \\ l_u = \{+\infty_p, \infty_n, \infty_n, y\} & l_l = \{-\infty_p, 0_n, 0_n, y\} \end{array}$$

gives the quantile regression LP problem as described in (11).

Once expressed as an LP problem the parameter estimates $\hat{\beta}(\tau)$ can be obtained in a number of ways, for example via the inertia-controlling method of Gill and Murray (1978) (see E04MFF/E04MFA), the simplex method or an interior point method as used by G02QFF and G02QGF.

2.7.4 Estimation of the covariance matrix

Koenker (2005) shows that the limiting covariance matrix of $\sqrt{n}\left(\hat{\beta}(\tau) - \beta(\tau)\right)$ is of the form of a Huber Sandwich. Therefore, under the assumption of Normally distributed errors

$$\sqrt{n}\left(\hat{\beta}(\tau) - \beta(\tau)\right) \sim N\left(0, \tau(1 - \tau) H_n(\tau)^{-1} J_n H_n(\tau)^{-1}\right) \tag{13}$$

where

$$J_n = n^{-1} \sum_{i=1}^{n} x_i x_i^{\mathrm{T}}$$
$$H_n(\tau) = \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} x_i x_i^{\mathrm{T}} f_i\left(Q_{y_i}(\tau | x_i)\right)$$

and $f_i\left(Q_{y_i}(\tau | x_i)\right)$ denotes the conditional density of the response y evaluated at the τth conditional quantile.

More generally, the asymptotic covariance matrix for $\hat{\beta}(\tau_1), \hat{\beta}(\tau_1), \ldots, \hat{\beta}(\tau_n)$ has blocks defined by

$$\operatorname{cov}\left(\sqrt{n}\left(\hat{\beta}(\tau_i) - \beta(\tau_i)\right), \sqrt{n}\left(\hat{\beta}(\tau_j) - \beta(\tau_j)\right)\right) = \left(\min(\tau_i, \tau_j) - \tau_i \tau_j\right) H_n(\tau_i)^{-1} J_n H_n(\tau_j)^{-1} \tag{14}$$

Under the assumption of independent, identically distributed (iid) errors, (13) simplifies to

$$\sqrt{n}\left(\hat{\beta}(\tau) - \beta(\tau)\right) \sim N\left(0, \tau(1 - \tau) s(\tau)^2 \left(X^{\mathrm{T}} X\right)^{-1}\right)$$

where $s(\tau)$ is the sparsity function, given by

$$s(\tau) = \frac{1}{f(F^{-1}(\tau))}$$

a similar simplification occurs with (14).

In cases where the assumption of iid errors does not hold, Powell (1991) suggests using a kernel estimator of the form

$$\hat{H}_n(\tau) = (nc_n)^{-1}\sum_{i=1}^{n} K\left(\frac{y_i - x_i^{\mathrm{T}}\hat{\beta}(\tau)}{c_n}\right) x_i x_i^{\mathrm{T}}$$

for some bandwidth parameter c_n satisfying $\lim_{n\to\infty} c_n \to 0$ and $\lim_{n\to\infty} \sqrt{n}c_n \to \infty$ and Hendricks and Koenker (1991) suggest a method based on an extension of the idea of sparsity.

Rather than use an asymptotic estimate of the covariance matrix, it is also possible to use bootstrapping. Roughly speaking the original data is resampled and a set of parameter estimates obtained from each new sample. A sample covariance matrix is then constructed from the resulting matrix of parameter estimates.

2.8 Latent Variable Methods

Regression by means of projections to latent structures also known as partial least squares, is a latent variable linear model suited to data for which:

the number of x-variables is high compared to the number of observations;

x-variables and/or y-variables are multicollinear.

Latent variables are linear combinations of x-variables that explain variance in x and y-variables. These latent variables, known as factors, are extracted iteratively from the data. A choice of the number of factors to include in a model can be made by considering diagnostic statistics such as the variable influence on projections (VIP).

2.9 LARS, LASSO and Forward Stagewise Regression

Least Angle Regression (LARS), Least Absolute Shrinkage Selection Operator (LASSO) and forward stagewise regression are three closely related regression techniques. Of the three, only LASSO has an easily accessible mathematical description suitable for being summarised here. A full description of the all three methods and the relationship between them can be found in Efron *et al.* (2004) and the references there in.

Given a vector of n observed values, $y = \{y_i : i = 1, 2, \ldots, n\}$ and an $n \times p$ design matrix X, where the jth column of X, denoted x_j, is a vector of length n representing the jth independent variable x_j, standardized such that $\sum_{i=1}^{n} x_{ij} = 0$, and $\sum_{i=1}^{n} x_{ij}^2 = 1$ and a set of model parameters β to be estimated from the observed values, the LASSO model of Tibshirani (1996) is given by

$$\underset{\alpha,\beta\in\mathbb{R}^p}{\text{minimize}} \left\| y - \alpha - X^{\mathrm{T}}\beta \right\|^2 \quad \text{subject to} \quad \|\beta\|_1 \le t \tag{15}$$

for a given value of t, where $\alpha = \bar{y} = n^{-1}\sum_{i=1}^{n} y_i$. The positive LASSO model is the same as the standard LASSO model, given above, with the added constraint that

$$\beta_j \ge 0, \quad j = 1, 2, \ldots, p.$$

Rather than solve (15) for a given value of t, Efron *et al.* (2004) defined an algorithm that returns a full solution path for all possible values of t. It turns out that this path is piecewise linear with a finite number of pieces, denoted K, corresponding to K sets of parameter estimates.

3 Recommendations on Choice and Use of Available Routines

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

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

G02BAF computes the product-moment correlation coefficients.

G02BTF updates the sample sums of squares and cross-products and deviations from the means by the addition/deletion of a (weighted) observation.

G02BUF computes the sample sums of squares and cross-products deviations from the means (optionally weighted). The output from multiple calls to G02BUF can be combined via a call to G02BZF, allowing large datasets to be summarised across multiple processing units.

G02BTF updates the sample sums of squares and cross-products and deviations from the means by the addition/deletion of a (weighted) observation.

G02BWF 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.

G02BXF computes both the (optionally weighted) covariance matrix and the (optionally weighted) correlation matrix. These are returned in two-dimensional arrays. (Note that G02BTF and G02BUF can be used to compute the sums of squares from zero.)

G02BGF can be used to calculate the correlation coefficients for a subset of variables in the data matrix.

3.1.2 Product-moment correlation with missing values

If there are missing values then G02BUF and G02BXF, as described above, will allow casewise deletion by you giving the observation zero weight (compared with unit weight for an otherwise unweighted computation).

Other routines also handle missing values in the calculation of unweighted product-moment correlation coefficients. Casewise exclusion of missing values is provided by G02BBF while pairwise omission of missing values is carried out by G02BCF. These two routines calculate a correlation matrix for all the variables in the data matrix; similar output but for only a selected subset of variables is provided by routines G02BHF and G02BJF respectively. As well as providing the Pearson product-moment correlation coefficients, these routines also calculate the means and standard deviations of the variables, and the matrix of sums of squares and cross-products of deviations from the means. For all four routines you are free to select appropriate values for consideration as missing values, bearing in mind the nature of the data and the possible range of valid values. The missing values for each variable may be either different or alike and it is not necessary to specify missing values for all the variables.

3.1.3 Nonparametric correlation

There are five routines which perform nonparametric correlations, each of which is capable of producing both Spearman's rank order and Kendall's tau correlation coefficients. The basic underlying concept of both these methods is to replace each observation by its corresponding rank or order within the observations on that variable, and the correlations are then calculated using these ranks.

It is obviously more convenient to order the observations and calculate the ranks for a particular variable just once, and to store these ranks for subsequent use in calculating all coefficients involving that variable; this does however require an amount of store of the same size as the original data matrix, which in some cases might be excessive. Accordingly, some routines calculate the ranks only once, and replace the input data matrix by the matrix of ranks, which are then also made available to you on exit from the routine, while others preserve the data matrix and calculate the ranks a number of times within the routine; the ranks of the observations are not provided as output by routines which work in the latter way. If it is possible to arrange the program in such a way that the first technique can be used, then efficiency of timing is achieved with no additional storage, whereas in the second case, it is necessary to have a second matrix of the same size as the data matrix, which may not be acceptable in certain circumstances; in this case it is necessary to reach a compromise between efficiency of time and of storage, and this may well be dependent upon local conditions.

Routines G02BNF and G02BQF both calculate Kendall's tau and/or Spearman's rank order correlation coefficients taking no account of missing values; G02BNF does so by calculating the ranks of each variable only once, and replacing the data matrix by the matrix of ranks, whereas G02BQF calculates the ranks of each variable several times. Routines G02BPF and G02BRF provide the same output, but treat missing values in a 'casewise' manner (see above); G02BPF calculates the ranks of each variable only once, and overwrites the data matrix of ranks, while G02BRF determines the ranks of each variable several times. For 'pairwise' omission of missing data (see above), the routine G02BSF provides Kendall and/or Spearman coefficients.

Since G02BNF and G02BPF order the observations and calculate the ranks of each variable only once, then if there are M variables involved, there are only M separate 'ranking' operations; this should be contrasted with the method used by routines G02BQF and G02BRF which perform $M(M-1)/2+1$ similar ranking operations. These ranking operations are by far the most time-consuming parts of these nonparametric routines, so for a matrix of as few as five variables, the time taken by one of the slower routines can be expected to be at least a factor of two slower than the corresponding efficient routine; as the number of variables increases, so this relative efficiency factor increases. Only one routine, G02BSF, is provided for pairwise missing values, and this routine carries out $M(M-1)$ separate rankings; since by the very nature of the pairwise method it is necessary to treat each pair of variables separately and rank them individually, it is impossible to reduce this number of operations, and so no alternative routine is provided.

3.1.4 Partial correlation

G02BYF computes a matrix of partial correlation coefficients from the correlation coefficients or variance-covariance matrix returned by G02BXF.

3.1.5 Robust correlation

G02HLF and G02HMF compute robust estimates of the variance-covariance matrix by solving the equations

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

and

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

as described in Section 2.1.4 for user-supplied functions w and u. Two options are available for v, either $v(t) = 1$ for all t or $v(t) = u(t)$.

G02HMF requires only the function w and u to be supplied while G02HLF also requires their derivatives.

In general G02HLF will be considerably faster than G02HMF and should be used if derivatives are available.

G02HKF computes a robust variance-covariance matrix for the following functions:

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

and

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

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 G02BWF may be used.

3.1.6 Nearest correlation matrix

Four routines are provided to calculate a nearest correlation matrix. The choice of routine will depend on what definition of 'nearest' is required and whether there is any particular structure desired in the resulting correlation matrix.

G02AAF computes the nearest correlation matrix in the Frobenius norm, using the method of Qi and Sun (2006).

G02ABF uses an extension of the method implemented in G02AAF allowing for the row and column weighted Frobenius norm to be used as well as bounds on the eigenvalues of the resulting correlation matrix to be specified.

G02AEF computes the factor loading matrix, allowing a correlation matrix with a k-factor structure to be computed.

G02AJF again computes the nearest correlation matrix in the Frobenius norm, but allows for element-wise weighting as well as bounds on the eigenvalues.

3.2 Regression

3.2.1 Simple linear regression

Four routines are provided for simple linear regressions: G02CAF and G02CCF perform the simple linear regression with a constant term (equation (1) above), while G02CBF and G02CDF fit the simple linear regression with **no** constant term (equation (2) above). Two of these routines, G02CCF and G02CDF, take account of missing values, which the others do not. In these two routines, an observation is omitted if it contains a missing value for either the dependent or the independent variable; this is equivalent to both the casewise and pairwise methods, since both are identical when there are only two variables involved. Input to these routines consists of the raw data, and output includes the coefficients, their standard errors and t values for testing the significance of the coefficients; the F value for testing the overall significance of the regression is also given.

3.2.2 Ridge regression

G02KAF calculates a ridge regression, optimizing the ridge parameter according to one of four prediction error criteria.

G02KBF calculates ridge regressions for a given set of ridge parameters.

3.2.3 Polynomial regression and nonlinear regression

No routines are currently provided in this chapter for polynomial regression. If you wish to perform polynomial regressions you have three alternatives: you can use the multiple linear regression routines, G02DAF, with a set of independent variables which are in fact simply the same single variable raised to different powers, or you can use the routine G04EAF to compute orthogonal polynomials which can then be used with G02DAF, or you 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.

3.2.4 Multiple linear regression – general linear model

G02DAF 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 G02DAF and to aid in the interpretation of the model.

G02DCF adds or deletes an observation from the model.

G02DDF computes the parameter estimates, and their standard errors and variance-covariance matrix for a model that is modified by G02DCF, G02DEF or G02DFF.

G02DEF adds a new variable to a model.

G02DFF drops a variable from a model.

G02DGF fits the regression to a new dependent variable, i.e., keeping the same independent variables.

G02DKF 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.

G02DNF calculates the estimate of an estimable function and its standard error.

Note: G02DEF also allows you to initialize a model building process and then to build up the model by adding variables one at a time.

If you wish to use methods based on forming the cross-products/correlation matrix (i.e., $(X^{\mathrm{T}}X)$ matrix) rather than the recommended use of G02DAF then the following routines should be used.

For regression through the origin (i.e., no constant) G02CHF preceded by:

G02BDF (no missing values, all variables)

G02BKF (no missing values, subset of variables)

G02BEF (casewise missing values, all variables)

G02BLF(casewise missing values, subset of variables)

G02BFF* (pairwise missing values, all variables)

G02BMF* (pairwise missing values, subset of variables)

For regression with intercept (i.e., with constant) G02CGF preceded by:

G02BAF (no missing values, all variables)

G02BGF (no missing values, subset of variables)

G02BBF (casewise missing values, all variables)

G02BHF (casewise missing values, subset of variables)

G02BCF* (pairwise missing values, all variables)

G02BJF* (pairwise missing values, subset of variables)

Note that the four routines using pairwise deletion of missing value (marked with $*$) should be used with great caution as the use of this method can lead to misleading results, particularly if a significant proportion of values are missing.

Both G02CGF and G02CHF require that the correlations/sums of squares involving the dependent variable must appear as the last row/column. Because the layout of the variables in your data array may not be arranged in this way, two routines, G02CEF and G02CFF, are provided for rearranging the rows and columns of vectors and matrices. G02CFF simply reorders the rows and columns while G02CEF forms smaller vectors and matrices from larger ones.

Output from G02CGF and G02CHF consists of the coefficients, their standard errors, R^2-values, t and F statistics.

3.2.5 Selecting regression models

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

G02EAF 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.

G02ECF computes the values of R^2 and C_p from the residual sums of squares as provided by G02EAF.

G02EEF enables you to fit a model by forward selection. You may call G02EEF 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.

G02EFF uses a full stepwise selection to choose a subset of the explanatory variables. The method repeatedly applies a forward selection step followed by a backward elimination step until neither step updates the current model.

3.2.6 Residuals

G02FAF computes the following standardized residuals and measures of influence for the residuals and leverages produced by G02DAF:

(i) Internally studentized residual;

(ii) Externally studentized residual;

(iii) Cook's D statistic;

(iv) Atkinson's T statistic.

G02FCF computes the Durbin–Watson test statistic and bounds for its significance to test for serial correlation in the errors, e_i.

3.2.7 Robust regression

For robust regression using M-estimates instead of least squares the routine G02HAF will generally be suitable. G02HAF 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 G02HDF may be used. G02HDF requires you to supply weights, if required, and also routines to calculate the ψ-function and, optionally, the χ-function. G02HBF can be used in calculating suitable weights. The routine G02HFF can be used after a call to G02HDF in order to calculate the variance-covariance estimate of the estimated regression coefficients.

For robust regression, using least absolute deviation, E02GAF can be used.

3.2.8 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.

G02GAF Normal distribution.

G02GBF binomial distribution.

G02GCF Poisson distribution.

G02GDF gamma distribution.

While G02GAF can be used to fit linear regression models (i.e., by using an identity link) this is not recommended as G02DAF will fit these models more efficiently. G02GCF can be used to fit log-linear models to contingency tables.

In addition to the routines to fit the models there is one routine to predict from the fitted model and two routines to aid interpretation when the fitted model is not of full rank, i.e., aliasing is present.

G02GPF computes a predicted value and its associated standard error based on a previously fitted generalized linear model.

G02GKF 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.

G02GNF calculates an estimate of an estimable function along with its standard error.

3.2.9 Linear mixed effects regression

There are four routines for fitting linear mixed effects regression.

G02JAF and G02JDF uses restricted maximum likelihood (REML) to fit the model.

G02JBF and G02JEF uses maximum likelihood to fit the model.

For all routines the output includes: either the maximum likelihood or restricted maximum likelihood and the fixed and random parameter estimates, along with their standard errors. Whilst it is possible to fit a hierachical model using G02JAF or G02JBF, G02JDF and G02JEF allow the model to be specified in a more intuitive way. G02JCF must be called prior to calling G02JDF or G02JEF.

As the estimates of the variance components are found using an iterative procedure initial values must be supplied for each σ. In all four routines you can either specify these initial values, or allow the routine to calculate them from the data using minimum variance quadratic unbiased estimation (MIVQUE0). Setting the maximum number of iterations to zero in any of the routines will return the corresponding likelihood, parameter estimates and standard errors based on these initial values.

3.2.10 Linear quantile regression

Two routines are provided for performing linear quantile regression, G02QFF and G02QGF. Of these, G02QFF provides a simplified interface to G02QGF, where many of the input parameters have been given default values and the amount of output available has been reduced.

Prior to calling G02QGF the optional parameter array must be initialized by calling G02ZKF with OPTSTR set to **Initialize**. Once these arrays have been initialized G02ZLF can be called to query the value of an optional parameter.

3.2.11 Partial Least Squares (PLS)

G02LAF calculates a nonlinear, iterative PLS by using singular value decomposition.

G02LBF calculates a nonlinear, iterative PLS by using Wold's method.

G02LCF calculates parameter estimates for a given number of PLS factors.

G02LDF calculates predictions given a PLS model.

3.2.12 LARS, LASSO and Forward Stagewise Regression

Two routines for fitting a LARS, LASSO or forward stagewise regression are supplied: G02MAF and G02MBF. The difference between the two routines is in the way that the data, X and y, are supplied. The first routine, G02MAF takes X and y directly, whereas G02MBF takes the data in the form of the cross-products: $X^{\mathrm{T}}X$, $X^{\mathrm{T}}y$ and $y^{\mathrm{T}}y$. In most situations G02MAF will be the recommended routine as the full data tends to be available. However when there is a large number of observations (i.e., n is large) it might be preferable to split the data into smaller blocks and process one block at a time. In such situations G02BUF and G02BZF can be used to construct the required cross-products and G02MBF called to fit the required model.

Both G02MAF and G02MBF return K sets of parameter estimates, which, because of it's piecewise linear nature, define the full LARS, LASSO or forward stagewise regression solution path. However, parameter estimates are sometimes required at points along the solution path that differ from those returned by G02MAF and G02MBF, for example when performing a cross-validation. G02MCF will return the parameter estimates in such cases.

4 Functionality Index

Computes a correlation matrix with fixed submatrix	G02ANF
Computes the nearest correlation matrix using element-wise weights	G02AJF
Computes the nearest correlation matrix using the k-factor model	G02AEF
Computes the nearest correlation matrix using the method of Qi and Sun,	
augmented G02AAF to incorporate weights and bounds	G02ABF
unweighted	G02AAF
Correlation-like coefficients,	
all variables,	
casewise treatment of missing values	G02BEF
no missing values	G02BDF
pairwise treatment of missing values	G02BFF
subset of variables,	
casewise treatment of missing values	G02BLF
no missing values	G02BKF
pairwise treatment of missing values	G02BMF
Generalized linear models,	
binomial errors	G02GBF
computes estimable function	G02GNF
gamma errors	G02GDF
Normal errors	G02GAF
Poisson errors	G02GCF
prediction	G02GPF
transform model parameters	G02GKF
Hierarchical mixed effects regression,	
initiation	G02JCF
using maximum likelihood	G02JEF
using restricted maximum likelihood	G02JDF
Least angle regression (includes LASSO),	
Additional parameter calculation	G02MCF
Model fitting,	
Cross-product matrix	G02MBF
Raw data	G02MAF
Linear mixed effects regression,	
via maximum likelihood (ML)	G02JBF
via restricted maximum likelihood (REML)	G02JAF

Robust correlation,	
Huber's method	G02HKF
user-supplied weight function only	G02HMF
user-supplied weight function plus derivatives	G02HLF
Robust regression,	
compute weights for use with G02HDF	G02HBF
standard M-estimates	G02HAF
user-supplied weight functions	G02HDF
variance-covariance matrix following G02HDF	G02HFF
Selecting regression model,	
all possible regressions	G02EAF
forward selection	G02EEF
R^2 and C_p statistics	G02ECF
Service routines,	
for multiple linear regression,	
reorder elements from vectors and matrices	G02CFF
select elements from vectors and matrices	G02CEF
general option getting routine	G02ZLF
general option setting routine	G02ZKF
Simple linear regression,	
no intercept	G02CBF
no intercept with missing values	G02CDF
with intercept	G02CAF
with intercept and with missing values	G02CCF
Stepwise linear regression,	
Clarke's sweep algorithm	G02EFF

5 Auxiliary Routines Associated with Library Routine Parameters

G02EFH nagf_correg_linregm_fit_stepwise_sample_monfun
See the description of the parameter MONFUN in G02EFF.

G02HDZ nagf_correg_robustm_user_dummy_chi
See the description of the parameter CHI in G02HDF.

6 Routines Withdrawn or Scheduled for Withdrawal

None.

7 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

Efron B, Hastie T, Johnstone I and Tibshirani R (2004) Least Angle Regression *The Annals of Statistics (Volume 32)* **2** 407–499

Gill P E and Murray W (1978) Numerically stable methods for quadratic programming *Math. Programming* **14** 349–372

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

Hendricks W and Koenker R (1991) Hierarchical spline models for conditional quantiles and the demand for electricity *Journal of the Maerican Statistical Association* **87** 58–68

Huber P J (1981) *Robust Statistics* Wiley

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

Koenker R (2005) *Quantile Regression* Econometric Society Monographs, Cambridge University Press, New York

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

Powell J L (1991) Estimation of monotonic regression models under quantile restrictions *Nonparametric and Semiparametric Methods in Econometrics* Cambridge University Press, Cambridge

Qi H and Sun D (2006) A quadratically convergent Newton method for computing the nearest correlation matrix *SIAM J. Matrix AnalAppl* **29(2)** 360–385

Searle S R (1971) *Linear Models* Wiley

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

Tibshirani R (1996) Regression Shrinkage and Selection via the Lasso *Journal of the Royal Statistics Society, Series B (Methodological) (Volume 58)* **1** 267–288

Weisberg S (1985) *Applied Linear Regression* Wiley