Chapter g02 – Regression Analysis

1. Scope of the Chapter

This chapter covers the fitting of linear regression models, along with the examination of the fit of the model and the calculation of correlation coefficients.

2. Background

2.1. Aims of Regression Modelling

Regression analysis is the study of the relationship between one specific random variable, the *dependent* or *response variable*, and one or more known variables, called the *independent variables* or *explanatory variables*. 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. This regression model will involve 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:

- to obtain best estimates of the unknown regression parameters;
- to test hypotheses about these parameters;
- to determine the adequacy of the assumed model;
- to verify the set of relevant assumptions.

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 or mean term),

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

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.

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

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

In the *exponential regression* model, the equation takes 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. 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), viz

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

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

$$y_i = \beta_0 x_{0i} + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + e_i, \quad i = 1, 2, \ldots, n$$

where e_i are unknown random errors with $E(e_i) = 0$ and $\mathrm{var}(e_i) = \sigma^2$, σ^2 being a constant, then the estimates of the regression parameters are calculated by minimizing

$$\sum_{i=1}^{n} {e_i}^2 \tag{6}$$

with respect to $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$. 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

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

is minimised.

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 value 0 or 1. So in the example there are four dummy variables x_j, for $j = 1, 2, 3, 4$, such that $x_{ij} = 1$ if the ith observation received the jth treatment, and $x_{ij} = 0$ otherwise, along with a variable for the mean, x_0: $x_{i0} = 1$ 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

$$x_{i4} = x_{i0} - x_{i1} - x_{i2} - x_{i3},$$

for $i = 1, 2, \ldots 7$. 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 is known as *intrinsic aliasing*. In the example above no matter how many times each treatment is replicated (other than 0) the same 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.

2.5. 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 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. However, due to the way the residuals are calculated they 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, $\hat{y} = Hy$ and hence the vector of residuals, r, is $r = (I - H)y$, I being the identity matrix. The variance-covariance matrix of the residuals is then $(I - H)\sigma^2$. The diagonal elements of H, h_{ii} can therefore be used to standardize the residuals. The value h_{ii} is a measure of the effect of the independent variables for the ith observation on the fitted model and is known as the leverage. Several forms of standardized residuals and measures of influence can be computed.

As $\hat{y} = Hy$ (H is idempotent); the variance of $\hat{y}_i$ is $h_{ii}\sigma^2$ for observations with equal variance σ^2. This variance can be used to compute a confidence interval for the fitted model. If a confidence interval for a predicted observation is required the variance $(1 + h_{ii})\sigma^2$ should be used.

2.6. 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 be 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

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

2.7. Updating the Model

In order to select a suitable model it is often useful to add or remove an independent variable from the model. One approach to model selection, *forward selection*, involves adding one variable at time, the variable chosen being the one that leads to the greatest inprovement in the fit of the model. An alternative approach is to fit a full model and then drop the variable that makes least contribution to the fit of the model, this is known as *backward elimination*.

In addition to adding or removing variables from the regression model it is sometimes useful to add or drop an observation from a model. If the data is being recorded in sequence or if the data is too large to be input in one go then the model can be fitted by adding one observation at a time until all data has been included. The parameter estimates can be examined at intervals to see if there is any significant change in the model.

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 (6) 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,

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

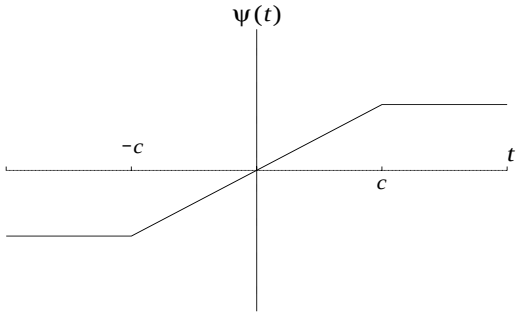

Figure 1

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).

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 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.

2.10. Correlation Analysis

Correlation analysis aims to measure the strength of the association between pairs of variables. In correlation analysis all variables are treated equally, in contrast to regression analysis in which one variable is considered the response variable and other variables are the explanatory variables.

The correlation between two variables is measured by a correlation coefficient, r, with $-1 \leq r \leq 1$. If $r = 0$ then there is no association between the two variables and the can be considered as independent. If $r = 1$ then the is a perfect positive association between the two variables and if $r = -1$ there is a perfect negative association between the two variables, that is if one variable

increases the other decreases. The form of the association measured depends on the correlation coefficient.

There are two main types of correlation coefficient, the product-moment correlation coefficient and nonparametric or rank correlation coefficients. The product-moment correlation coefficient, sometimes called the Pearson product-moment correlation coefficient, measures the linear association between two variables, that is it measures the extent to which one variable increases or decreases linearly with the other. It is equivalent to the regression coefficient of one variable on the other when both variables have been standardised to have unit variance.

There are two common rank based correlation coefficients, Kendall's tau and Spearman's rho. They both measure monotonic association. This is the extent to which one variable simply increases or decreases with the other without specifying the form of the increase or decrease. In choosing between the two coefficients, Kendall's tau is often preferred when the data take a small number of different values and there are a large number of ties, otherwise Spearman's rho is used. In comparing rank correlation coefficients to the product-moment correlation coefficient, the rank correlation coefficients are able to detect correlation when the relationship is basically nonlinear but monotonic rather than just linear and rank correlation coefficients also have the advantage of being less effected by extreme values or outliers in the data.

It should be noted that it is possible for there to be a nonlinear relationship between two variables such that both the product-moment and rank correlation coefficients are small. To be sure of detecting the relationships between variables the correlation analysis should be combined with examining plots of the data and, if required, applying suitable transformations to the data.

To test the significance of correlation coefficients t-tests can be used. For the product-moment correlation coefficient the statistic,

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

computed from a sample of size n, has a Student's t-distribution with $n-2$ degrees of freedom under the null hypothesis that there is no correlation between the two variables. For Spearman's rho the distribution of the statistic is only approximate for large samples.

2.11. 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:

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

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 is 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

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

where $w(t)$, $u(t)$ and $v(t)$ are functions such that they return value 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.

3. References

Atkinson A C (1986) *Plots, Transformations, and Regression* Oxford University Press.
Cook R D and Weisberg S (1986) *Residuals and Influence in Regression* Chapman and Hall.
Draper N R and Smith H (1966) *Applied Regression Analysis* Wiley.
Hammarling S (1985) The Singular Value Decomposition in Multivariate Statistics *ACM Signum Newsletter* **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.
Huber P J (1981) *Robust Statistics* Wiley.
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.

4. Available Functions

Compute product-moment correlation coefficients	`g02bxc`
Compute Kendall's tau and Spearman's rho rank correlation coefficients	`g02brc`
Robust estimate of covariance matrix	`g02hkc`
Least-squares regression	
Fit a sample linear regression	`g02cac`
Fit a sample linear regression and compute confidence intervals for model values	`g02cbc`
Fit a general linear regression model	`g02dac`
Add or drop an observation from a regression model	`g02dcc`
Add a variable to a regression model	`g02dec`
Drop a variable from a regression model	`g02dfc`
Compute the parameter estimates for an updated model	`g02ddc`
Fit the regression model to a new dependent variable	`g02dgc`
Compute parameter estimates using given constraints	`g02dkc`
Compute an estimable function	`g02dnc`
Compute standardized residuals and measures of influence	`g02fac`
Generalized linear model	
Normal errors	`g02gac`
Binomial errors	`g02gbc`
Poisson errors	`g02gcc`
Gamma errors	`g02gdc`
Compute the parameter estimates for given constraints	`g02gkc`
Compute an estimable function	`g02gnc`
Robust regression	
Bounded influence regression (M estimation)	`g02hac`