Statistics and Econometrics for Finance

Abdulkader Aljandali

Multivariate Methods and Forecasting with IBM® SPSS® Statistics

Springer

Statistics and Econometrics for Finance

More information about this series at http://www.springer.com/series/10377

This is the second part of a two-part guide to quantitative analysis using the IBM SPSS Statistics software package. This volume focuses on multivariate analysis, forecasting techniques and research methods.

Abdulkader Aljandali

Multivariate Methods and Forecasting with IBM® SPSS® Statistics

Springer

Abdulkader Aljandali
Accounting, Finance and Economics Department
Regent's University London
London, UK

ISSN 2199-093X ISSN 2199-0948 (electronic)
Statistics and Econometrics for Finance
ISBN 978-3-319-56480-7 ISBN 978-3-319-56481-4 (eBook)
DOI 10.1007/978-3-319-56481-4

Library of Congress Control Number: 2017939132

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

IBM SPSS Statistics is an integrated family of products that addresses the entire analytical process, from planning to data collection to analysis, reporting and deployment. It offers a powerful set of statistical and information analysis systems that runs on a wide variety of personal computers. As such, IBM SPSS (previously known as SPSS) is extensively used in industry, commerce, banking and local and national government education. Just a small subset of users of the package in the UK includes the major clearing banks, the BBC, British Gas, British Airways, British Telecom, Eurotunnel, GlaxoSmithKline, London Underground, the NHS, BAE Systems, Royal Dutch Shell, Unilever and W.H. Smith & Son.

In fact, all UK universities and the vast majority of universities worldwide use IBM SPSS Statistics for teaching and research. It is certainly an advantage for a student in the UK to have knowledge of the package since it obviates the need for an employer to provide in-house training. There is no text at present that is specifically aimed at the undergraduate market in business studies and associated subjects such as finance, marketing and economics. Such subjects tend to have the largest numbers of enrolled students in many institutions, particularly in the former polytechnic sector. The author is not going to adopt an explicitly mathematical approach, but rather will stress the applicability of various statistical techniques to various problem-solving scenarios.

IBM SPSS Statistics offers all the benefits of the Windows environment as analysts can have many windows of different types open at once, enabling simultaneous working with raw data and results. Further, users may learn the logic of the program by choosing an analysis rather than having to learn the IBM SPSS command language. The last thing wanted by students new to statistical methodology is simultaneously to have to learn a command language. There are many varieties of tabular output available, and the user may customise output using IBM SPSS script.

This book builds on a previous publication, *Quantitative Analysis and IBM SPSS Statistics: A Guide for Business and Finance* (Springer, 2016), which provided a gentle introduction to the IBM SPSS Statistics software for both students and

professionals. This book is aimed at those who have had exposure to the program and intend to take their knowledge further. This text is more advanced compared to the one above-mentioned and will be beneficial to students in their final year of undergraduate study, master's students, researchers and professionals working in the areas of practical business forecasting or market research data analysis. This text would doubtlessly be more sympathetic to the readership than the manuals supplied by IBM SPSS Inc.

London, UK
June 10th 2017

Abdulkader Mahmoud Aljandali

Introduction

This is the second part of a two-part guide to the IBM SPSS Statistics computer package for business, finance and marketing students. This, the second part of the guide, introduces multivariate regression, logistic regression, Box-Jenkins methodology alongside other multivariate and forecasting methods. Although the emphasis is on applications of the IBM SPSS Statistics software, there is a need for the user to be aware of the statistical assumptions and rationale that underpin correct and meaningful application of the techniques that are available in the package. Therefore, such assumptions are discussed, and methods of assessing their validity are described.

This, the second part of the IBM SPSS Statistics guide, is itself divided into three sections. The first chapter of **Part I** introduces multivariate regression and the assumptions that underpin it. The chapter discusses the multicollinearity and residual problems. Two-variable regression and correlation are illustrated, and the assumptions underlying the regression method are stressed. Logistic and dummy regression models in addition to functional forms of regression are the subject matter of Chap. 2. The Box-Jenkins methodology, stationarity of data and various steps that lead to the generation of mean equations are introduced in Chap. 3. The practical utility of time series methods is discussed. Exponential smoothing and naïve models Chap. 4 conclude Part I. **Part II** introduces multivariate methods such as factor analysis (Chap. 5) discriminant analysis (Chap. 6) and multidimensional scaling (Chap. 7). This part concludes with a chapter on the hierarchical log-linear analysis model (Chap. 8).

Part III comprises chapters that introduce popular concepts usually taught under research methods. Testing for dependence using the chi square test is discussed in Chap. 9, while applications on parametric and non-parametric tests are made available to the reader in Chap. 10. Parametric methods make more rigid assumptions about the distributional form of the gathered data than do non-parametric

methods. However, it must be recognised that parametric methods are more powerful when the assumptions underlying them are met. This book concludes by a review of the concept of constant and real prices in business and the effect it might have on the recording of data over time (Chap. 11).

Contents

Part III Research Methods

List of Figures

List of Tables

Part I
Forecasting Models

Chapter 1
Multivariate Regression

More often than not, regression models involve more than one *independent* (*predictor* or *regressor*) variable. It is hard to think of any *dependent* variable (Y) in business applications that might be determined by just one factor. For example, forecasting methods are commonly applied to series such as inflation rates, unemployment, exchange rates, and population numbers etc. that have complex relationships with determining variables. This chapter introduces the multivariate linear regression model. This model may be regarded as a *descriptive* tool, by which the linear dependence of one variable on others is summarised. It may also be regarded as an *inferential* tool, via which the relationships in a population are evaluated from the examination of sample data. The question of inference may be conveniently grouped into two general classes; estimation and hypothesis testing. The multivariate linear regression model thus plays a crucial role in examining the relationships between variables and producing forecasts.

The multivariate linear regression model may be written in algebraic form as:

$$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + ... + e_i \quad \text{for } i = 1, 2, ..., n$$

where Y_i represents the values of a dependent variable, X_{1i}, X_{2i}, ... are the values of a set of independent variables, n is the number of gathered observations and e_i represents the model error. The b_i are called *regression coefficients* whose numerical values are to be determined. There are a series of assumptions that underpin the use of the above model. Violation of these assumptions may lead to incorrect inferences being made, so good research will spend some time in testing and evaluating these assumptions which are described in the next section.

A. Aljandali, *Multivariate Methods and Forecasting with IBM® SPSS® Statistics*,
Statistics and Econometrics for Finance, DOI 10.1007/978-3-319-56481-4_1

1.1 The Assumptions Underlying Regression

1.1.1 Multicollinearity

When the regressor variables are truly not independent or are rather correlated with each other (i.e. they display redundant information), *multicollinearity* is said to exist. It is imperative that users of multivariate regression not only understand the effects of multicollinearity, but learn to diagnose it. Consider the fictitious data in the table overleaf. If we regress Y on X_1, we will find (Table 1.1):

$$\hat{Y} = 0.195 + 0.582X_1$$

with a coefficient of determination of 63.4%. The positive gradient of X_1 makes sense, since as Y increases, so too does X_1. Note in the table that X_2 is treble X_1 save for the last datum point. These latter two supposedly independent variables are in fact strongly correlated or *multicollinear*. Regressing Y on X_1 and X_2 we now obtain:

$$\hat{Y} = -3.977 + 15.166\,X_1 - 4.498\,X_2.$$

The gradient of X_1 has increased nearly 26-fold to a value of 15.166. Further the gradient of X_2 is negative, which does not make sense. A negative gradient would infer that in general as Y increases then X_2 decreases. **Multicollinearity may result in the regression coefficients being mis-specified in magnitude and/or in sign.** Given that the regression coefficients are gradients reflecting rates of change (a particularly important concept in Economics), inferences about the regression coefficients may become unreliable in this situation, even though the coefficient of determination may be high. Note that it is possible that the regression errors may be very small, yet the regression coefficients are estimated poorly.

A further point is that the standard error of the sample regression coefficients can be inflated by multicollinearity. There are estimation procedures designed to combat multicollinearity – procedures designed to eliminate model instability and to reduce the variances of the regression coefficients. One alternative for reducing multicollinearity, but remaining within standard least squares estimation, is to try transformations on the regressor variables. For example, in the case of two regressors, X_1 and X_2 which are highly correlated, defining a variable $Z = X_1 + X_2$ might

Table 1.1 Fictious data

Y	X_1	X_2
1	3.8	11.4
2	3.2	9.6
3	4.0	12.0
4	4.5	13.5
5	8.6	27.0

produce an effective result. BE CAUTIOUS, however, in forming new variables like Z that do not make sense in the context of the problem at hand. For example, the researcher should be extremely reluctant to sum variables that are measured in different units. A second alternative is simply to remove one of a correlated pair of regressor variables.

The technique of *ridge regression* is probably the most popular estimating technique for combatting multicollinearity. Ridge regression falls into the category of *biased estimation techniques*. The method is based on the idea that least squares yields unbiased estimates and indeed enjoys minimum variance of all linear unbiased estimators, but there is no upper bound on the variance of the estimators and the presence of multicollinearity tends to produce large variances. As a result, a huge price is paid for the unbiased property inherent on ordinary least squares. Biased estimation is used to attain a substantial reduction in variance with an accompanied increase in stability of the regression coefficients. The coefficients do become biased and simply put, the reduction in variance should be of greater magnitude than the bias induced in the estimators. Ridge regression is available in IBM SPSS Statistics, but the syntax has to be written to run the procedure.

To detect potential multicollinearity, the user could run Pearsonian correlations between every pair of regressor variables and test for significance. An alternative is to run the regression analysis and generate *variance inflation factors* (*VIF*). **The VIF's represent the inflation in variance that each sample regression coefficient experiences above the ideal i.e. above what would be experienced if there was no correlation between the regressor variables**. As a rule of thumb, a VIF above 10 is a cause for concern. VIF's are available in the IBM SPSS Statistics regression procedure and can be generated by selecting *Collinearity diagnostics* under the *statistics* tab as shown in Fig. 1.1.

1.1.2 Homoscedasticity of the Residuals

Residuals (or *errors* or *disturbances*) in a regression analysis are simply the differences between the observed and predicted values of Y:

$$\text{Residual } (\text{or } e_i) = Y_i - \widehat{Y}_i, \text{for } i = 1, 2, 3, ..., n$$

and n is the number of observations.

The linear regression model requires that the residuals in the population should have zero mean and constant spread or variance about the regression line. If this is the case, then we should expect constant variance of the residuals about the regression line in our gathered sample. The property of constant residual variance is known as *homoscedasticity*. In practice, this assumption is often violated. It is almost endemic that as numbers become larger in an investigation, variation about the trend or fitted model becomes larger. The property of non-constant residual

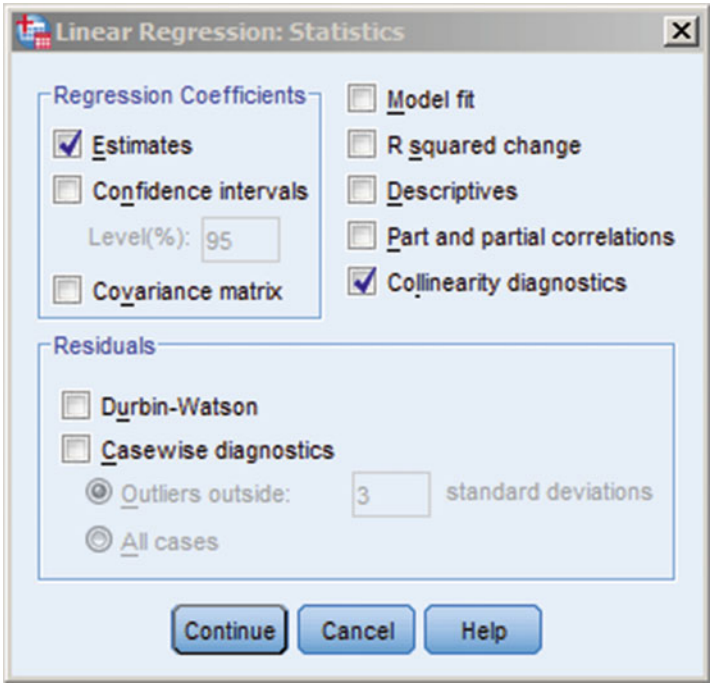


Fig. 1.1 Linear regression: statistics

variance is called *heteroscedasticity*. If the homoscedasticity assumption is violated, the regression coefficients are still unbiased estimates of their population counterparts, but they are no longer fully *efficient*. This means that other unbiased estimates exist that have smaller sample variances. The formula used for computing the variances of the sample regression coefficients may be nowhere near correct if the variance of the residuals is not constant.

Figures 1.2 and 1.3 are examples of a diagram that is commonly used to detect whether or not the homoscedasticity assumption is met. These diagrams plot the standardized values of the regression residuals on the vertical axis against the standardized predicted values. Both variables are standardized (zero mean, variance of one), so that the magnitude of the raw data values does not come into play. In Fig. 1.2, the spread or variance of the residuals seems pretty constant as we move from the left to the right of the graph. In Fig. 1.3, however, this is not the case. Figure 1.2 suggests homoscedasticity of the residuals; Fig. 1.3 suggests heteroscedasticity. If plots such as these indicate a marked curved trend, then it is likely that a linear regression model is inappropriate for the data. It should be noted in passing that such plots may reveal one or more unusually large residuals. Such points are called *outliers*, which may represent data input errors or they may reflect special cases that should be investigated further. Outliers are individual data points that do not fit the trend set by the balance of the data. One rule-of-thumb definition

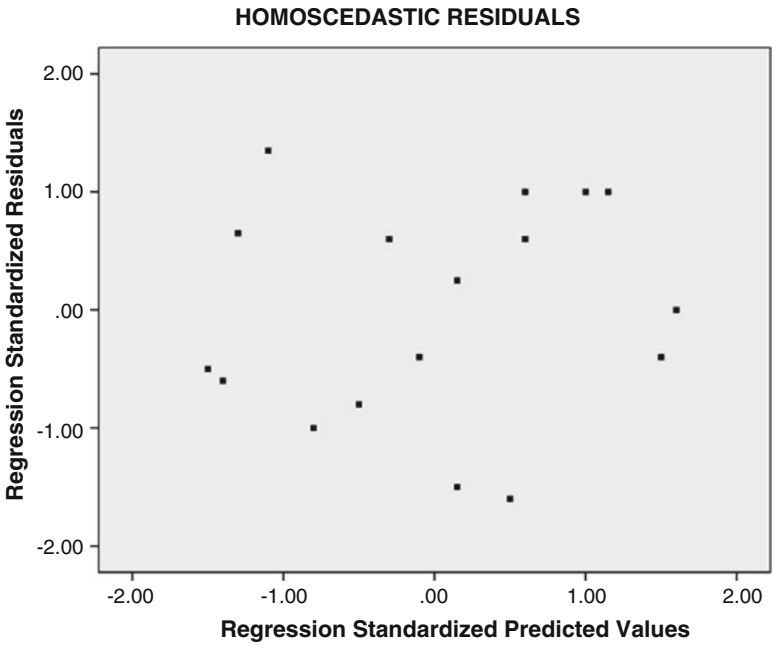


Fig. 1.2 Homoscedastic residuals

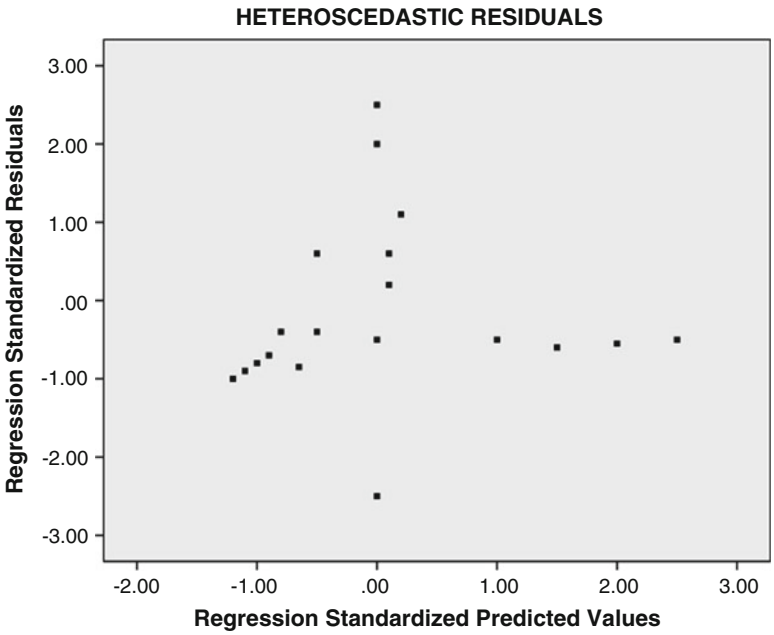


Fig. 1.3 Heteroscedastic residuals

of an outlier is any residual that lies more than three standardized units from the regression line (although ±2 standardized units does appear in the literature). Remember that the vertical axis in Figs. 1.2 and 1.3 has been standardized, so such plots offer a visual method of detecting outliers more than three standardized units from the regression line.

There are available statistical tests to obtain quantitative measures of model inadequacies due to violation of the homoscedasticity assumption. If heteroscedasticity is suspected, then efforts should be made to counter the problem. One method is to transform the X and Y variables. For example, suppose in the bivariate case that the standard deviation of the residuals increases in direct proportion to the values of X. Such heteroscedasticity may be eliminated by dividing each value of Y by the corresponding value of X. A regression equation is then calculated using 1/X as the independent variable (in place of X) and Y/X as the dependent variable (in place of Y). For multivariate problems, the situation is more complex if heteroscedasticity is present. The method of *weighted least squares* (*WLS*) has been devised to tackle the problem by calculating the absolute value of the unstandardized residuals. WLS is available in IBM SPSS Statistics under the *Linear Regression* screen.

1.1.3 Normality of the Residuals

The linear regression model requires that the residuals should be normally distributed. This is required for hypothesis testing and confidence intervals. Small departures from normality do not affect the regression model greatly, but gross departures are potentially more serious. Furthermore, if the residuals come from a distribution with thicker or heavier tails than the normal, the least squares fit may be sensitive to a small subset of the data. Heavy-tailed residual distributions often involve outliers that "pull" the least squares line in their direction.

The normality assumption may be examined graphically via a histogram or normal probability plot of the residuals. A formal statistical test is available via the Shapiro-Wilks test available in the IBM SPSS Statistics Explore routine. These were all discussed in Chap. 7 (Bivariate Correlation and Regression) in Quantitative Analysis with IBM SPSS Statistics: A Guide for Business and Finance.

1.1.4 Independence of the Residuals

Some applications of regression involve dependent and regressor variables that have a natural, sequential order over time. Such *time series models* are common in economics, business and some fields of engineering. Linear regression models assume that the residuals are *independent* or *uncorrelated*. Such an assumption for time series data is often not applicable, for example the value of an economic

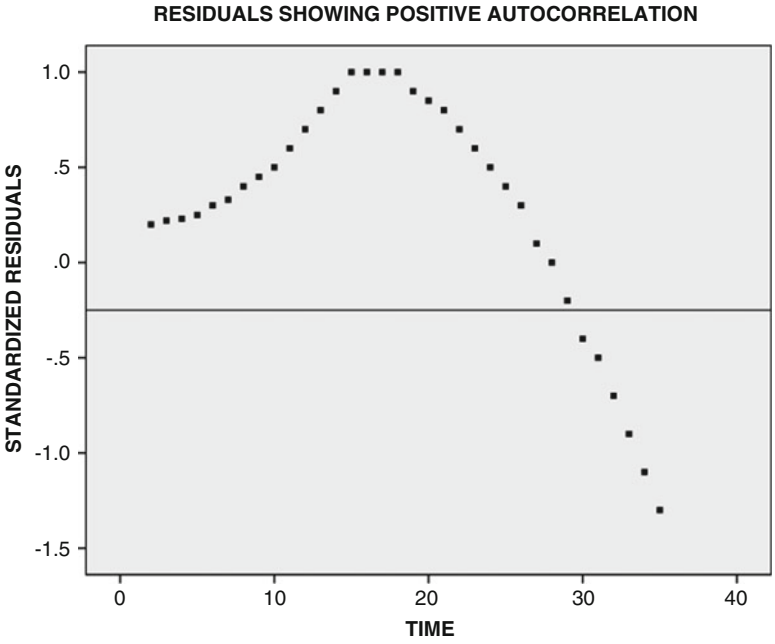


Fig. 1.4 Positively correlated residuals

variable in 1 month is often related to its value in the previous month(s). Variables such as seasonal influences and consumer population size are essentially synonymous with the variable time. Such a lack of independence is called *temporal autocorrelation*. It may be noted that autocorrelation may occur over space. For example, crop output in one region may be correlated with such outputs in neighbouring regions due to similar growing conditions. This is *spatial autocorrelation*.

Figure 1.4 exhibits positive temporal autocorrelation of the residuals; Fig. 1.5 shows negative temporal autocorrelation, which is uncommon in economic or business-related data. Such plots offer visual inspection about the presence or otherwise of autocorrelation.

In the case of positive autocorrelation, unusually large numbers of residuals with the same sign tend to cluster together. Negative autocorrelation is suggested by rapid changes in sign of the residuals. Positive autocorrelation causes estimates of the population variance of the residuals to be substantial underestimates. Such estimates are central in various test of hypotheses and in the estimation of confidence intervals, so it is in these areas that autocorrelation causes difficulties. In the instance of positive autocorrelation, confidence intervals for the population regression coefficients are deflated. One approach to overcoming the problem of autocorrelation is the *Cochrane-Orcutt procedure* discussed later in this chapter.

There are formal statistical tests for the detection of autocorrelation. Perhaps the best known is the *Durbin-Watson* (D-W) *test*, which is available in IBM SPSS

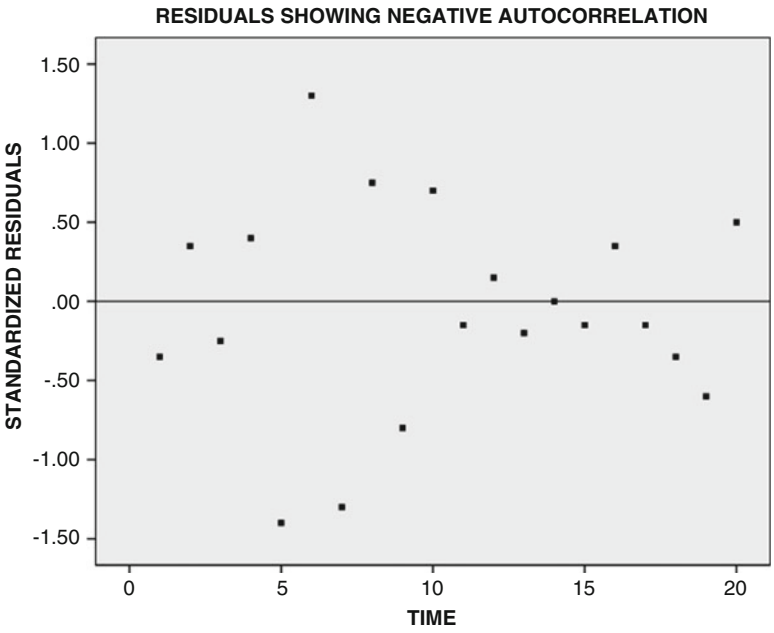


Fig. 1.5 Negatively correlated residuals

Statistics. The D-W test is usually applied to detect positive autocorrelation, since most regression problems involving time series data have the potential to exhibit this phenomenon. The population autocorrelation is denoted by ρ and the D-W test examines the null hypothesis that $\rho = 0$ versus the alternative that $\rho > 0$.

Denoting the residual at time t as e_t, the D-W test statistic is:

$$d = \frac{\sum (e_t - e_{t-1})^2}{\sum e_t^2}$$

where the e_t, t = 1, 2, ..., n are the residuals from an ordinary least squares analysis. Statistical tables are available for the significance of the d statistic. These tables contain a lower bound for d, denoted by d_L and an upper bound for d, denoted by d_U. The decision procedure is as follows:

- If $d < d_L$ reject H_0: $\rho = 0$ in favour of H_1
- If $d > d_U$, do not reject H_0: $\rho = 0$
- If $d_L < d < d_U$, then the test is inconclusive.

Clearly, small values of d suggest that H_0: $\rho = 0$ should be rejected because positive autocorrelation indicates that successive residual terms are of similar magnitude and the differences in the residuals will consequently be small.

As stated, situations where negative autocorrelation is present are rare. However, if a test is required for negative autocorrelation, one can use the statistic (d), where d

is the D-W statistic previously defined. The decision rule for H_0: $\rho = 0$ versus H_1: $\rho > 0$ is the same as that used for testing for positive autocorrelation.

1.2 Selecting the Regression Equation

We are trying to establish a linear regression equation for a dependent or response variable Y in terms of "independent" or predictor variables $X_1, X_2, X_3, \ldots$ etc. Two opposing criteria for selecting a relevant equation are usually involved. They are:

1. To make the equation useful for predictive purposes, we should want our model to include as many X_i as possible so that reliable fitted values can be determined.
2. Due to the costs involved in obtaining information on a large number of X_i and subsequently monitoring them, we would like the equation to include as few X_i as possible.

The compromise between these extremes is usually what is called *selecting the best regression equation*. There is no unique statistical procedure for doing this and personal judgment will be a necessary part of any of the statistical methods involved. There are several procedures available for selecting the best regression equation. They do not necessarily lead to the same solution when applied to the same problem, although for many problems they will achieve the same answer.

The **forward selection** procedure inserts the X_i until the regression equation is satisfactory. As each X_i is entered into the regression equation, the coefficient of determination (called the *multiple correlation coefficient* in the multivariate case) is computed. The first X_i entered into the regression is the one that will generate the largest coefficient of determination. The second X_i entered will be the variable that engenders the largest increase in the coefficient of determination etc. At some stage, the increase in the value of the coefficient of determination will not be significant. (A partial F test is used to examine whether a particular variable has taken up a significant amount of variation over that removed by variables already in the equation).

The **backward selection** procedure is an attempt to achieve a similar conclusion working from the other direction (i.e. to remove variables until the regression is satisfactory). A regression containing all the X_i is initially computed. A (partial F) test is conducted for every X_i as though it were the last variable to enter the regression equation. Variables are eliminated if they do not explain significant amounts of variation in Y. In spite of its name, the **stepwise selection** procedure is, in fact, an improved version of the forward selection procedure. The improvement involves the re-examination at every stage of the regression of the variables incorporated into the model at previous stages.

A variable which may have been the best single variable to enter at an earlier stage may, at a later stage, be superfluous because of the relationship between it and other variables now in the regression.

A judgment is made on the contribution made by each variable as though it had been the most recent variable entered, irrespective of its actual point of entry into the model. Any variable that provides a non-significant contribution is removed from the model. This process is continued until no more variables are admitted to the equation and no more are rejected. Stepwise is thus a combination of forward and backward selection. Lastly, there is the **enter selection** procedure. This process simply enters all the X_i in one block. Naturally, this latter method provides no information about the relative importance or otherwise of the X_i.

1.3 Multivariate Regression in IBM SPSS Statistics

A study was undertaken concerning workloads in 17 hospitals at various sites around England. The data are contained in the IBM SPSS Statistics file NHS.SAV associated with this chapter and the file includes the following regressor variables:

X_1 – mean daily patient load,
X_2 – monthly X-ray exposures,
X_3 – monthly occupied bed days,
X_4 – eligible population in the area (000's) and
X_5 – average length of patient's stay (days).

The dependent variable (Y) is the number of hours per month devoted to patient care. The researchers hypothesised that the regressor variables above result in the need for manpower in a hospital installation. Multivariate regression was applied using the stepwise selection procedure as a precursor to running the multivariate regression; it is worthwhile to examine the Pearsonian correlations between the five regressor variables to indicate whether multicollinearity may be a problem. As indicated in *Part 1*, correlation routines are accessed by clicking:

```
Analyze
    Correlate
        Bivariate ...
```

from the IBM SPSS Statistics Data Editor. This gives rise to the *Bivariate correlations dialogue box*, in which X_1 to X_5 inclusive are entered into the 'Variables' box. The results of Fig. 1.6 are produced. The results show that several pairs of the X_i are significantly correlated and may, therefore, create a multicollinearity problem should such pairs appear together in the final regression model.

For example X_1 and X_2, $r = 0.908$, X_1 and X_4, $r = 0.935$ are two such pairs whose correlations are significantly different from zero. For convenience, IBM SPSS Statistics places stars besides significant correlations. Noting these results, or at least the significant ones, we proceed with our multivariate regression which is accessed via:

Correlations

		Mean daily patient load	monthly x-ray exposures	monthly occupied bed days	eligible population in the area	mean length of patient's stay	monthly no. of hours devoted to care
Mean daily patient load	Pearson Correlation	1	.908**	.999**	.935**	.567*	.983**
	Sig. (2-tailed)		.000	.000	.000	.018	.000
	N	17	17	17	17	17	17
monthly x-ray exposures	Pearson Correlation	.908**	1	.920**	.914**	.353	.942**
	Sig. (2-tailed)	.000		.000	.000	.165	.000
	N	17	17	17	17	17	17
monthly occupied bed days	Pearson Correlation	.999**	.920**	1	.931**	.561*	.988**
	Sig. (2-tailed)	.000	.000		.000	.019	.000
	N	17	17	17	17	17	17
eligible population in the area	Pearson Correlation	.935**	.914**	.931**	1	.349	.941**
	Sig. (2-tailed)	.000	.000	.000		.170	.000
	N	17	17	17	17	17	17
mean length of patient's stay	Pearson Correlation	.567*	.353	.561*	.349	1	.486*
	Sig. (2-tailed)	.018	.165	.019	.170		.048
	N	17	17	17	17	17	17
monthly no. of hours devoted to care	Pearson Correlation	.983**	.942**	.988**	.941**	.486*	1
	Sig. (2-tailed)	.000	.000	.000	.000	.048	
	N	17	17	17	17	17	17

**.Correlation is significant at the 0.01 level (2-tailed).

*.Correlation is significant at the 0.05 level (2-tailed).

Fig. 1.6 Correlations between the regressor variables

```
Analyze
     Regression
          Linear...
```

The *Linear Regression dialogue box* of Fig. 1.7 is generated. (In passing, the user might note the button in the bottom left hand corner labelled WLS>>. This refers to weighted least squares mentioned previously). The Y variable is entered into the 'Dependent' box and the five X_i enter the 'Independent(s)' box. In the 'Method' box, we select the stepwise procedure from the four options. Click the Statistics... button to generate the *Linear Regression: Statistics dialogue box* of Fig. 1.8. Selecting collinearity diagnostics is always a sensible option. Also selected in Fig. 1.8 are confidence intervals for the population regression coefficients, the Durbin-Watson statistic to test for autocorrelation and a request to print out outliers which a three standard deviations or more from the regression.

It should be noted that autocorrelation is not so much a consideration here, as the hospital data are recorded over neither time nor space (the D-W option is mentioned for information). Click the Continue button to return to Fig. 1.7.

Click the Plots button to produce the *Linear Regression: Plots* dialogue box of Fig. 1.9. I have selected to construct a plot of the standardized residuals (*ZRESID) against the standardized predicted values (*ZPRED), in order to assess the homoscedasticity assumption. Also requested is a histogram of the residuals to assess the normality assumption. Click the Continue button to return to Fig. 1.7.

Click the Save button to produce the *Linear Regression: Save dialogue box* of Fig. 1.10. I have chosen to save the standardized and unstandardized residuals and

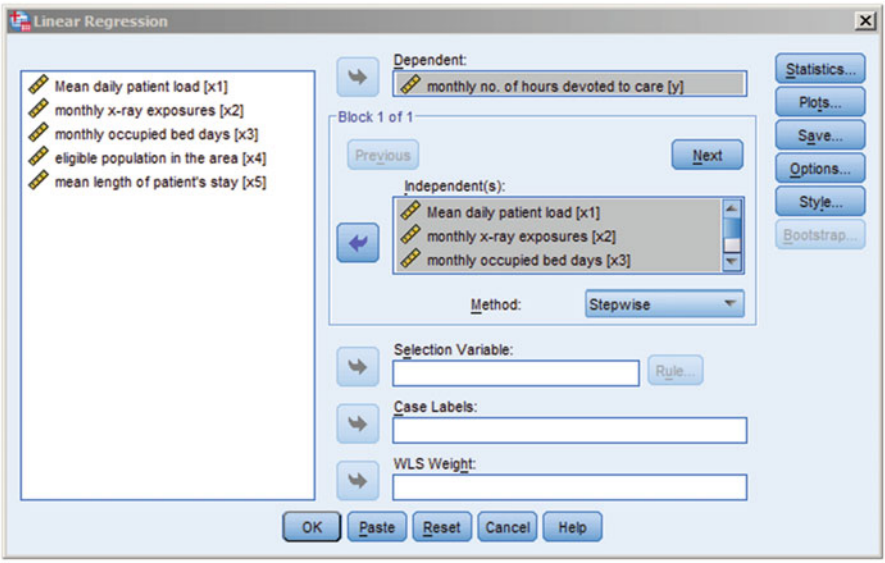


Fig. 1.7 The linear regression dialogue box

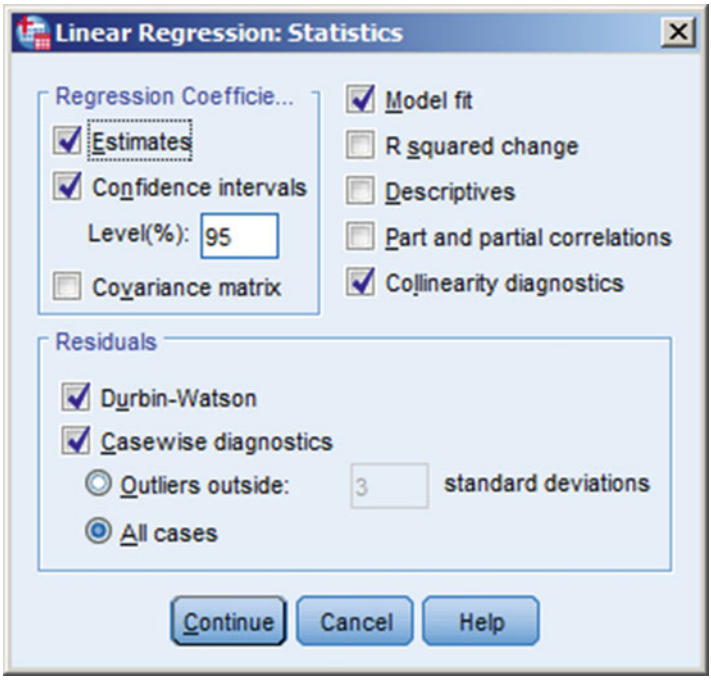


Fig. 1.8 The linear regression: statistics dialogue box

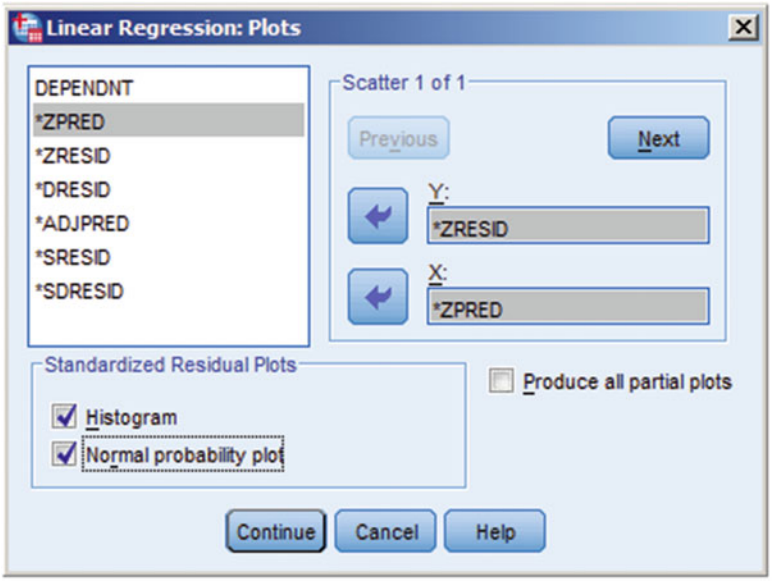


Fig. 1.9 The linear regression: plots dialogue box

predicted values. These four new variables will be added to the existing IBM SPSS Statistics data file. Click the Continue button to return to Fig. 1.7 and then the OK button to operationalise.

Figure 1.11 illustrates part of the IBM SPSS Statistics output for the hospital data. Under the heading 'Variables Entered/Removed', we see that the stepwise procedure entered X_3 and the first step and then X_2 at the second step. Therefore, X_1, X_4 and X_5 do not make significant contributions to explaining variation in Y. Hence, the monthly man-hours is primarily determined by the monthly occupied bed days and the monthly X-ray exposures. Under the heading 'Model Summary', we see that the multiple regression coefficient for a model involving just these two regressors is a healthy 98.4%. Despite the significant correlation between X_3 and X_2 evident in Fig. 1.6, the variance inflation factors (VIF) found under the heading 'Coefficients' are 6.475 <10.

Our rule of thumb suggests that multicollinearity is not a problem here. In passing, the D-W statistic is found under the heading 'Model Summary' and here has numerical value 2.618. From statistical tables, with two regressors in our model and $n = 17$ readings, $d_L = 1.015$ and $d_U = 1.536$. Since our observed value of $d > d_U$, we do not reject the hypothesis of zero autocorrelation. However, it may be recalled that autocorrelation is not a consideration here.

The equation of regression is found under the heading 'Coefficients' and is:

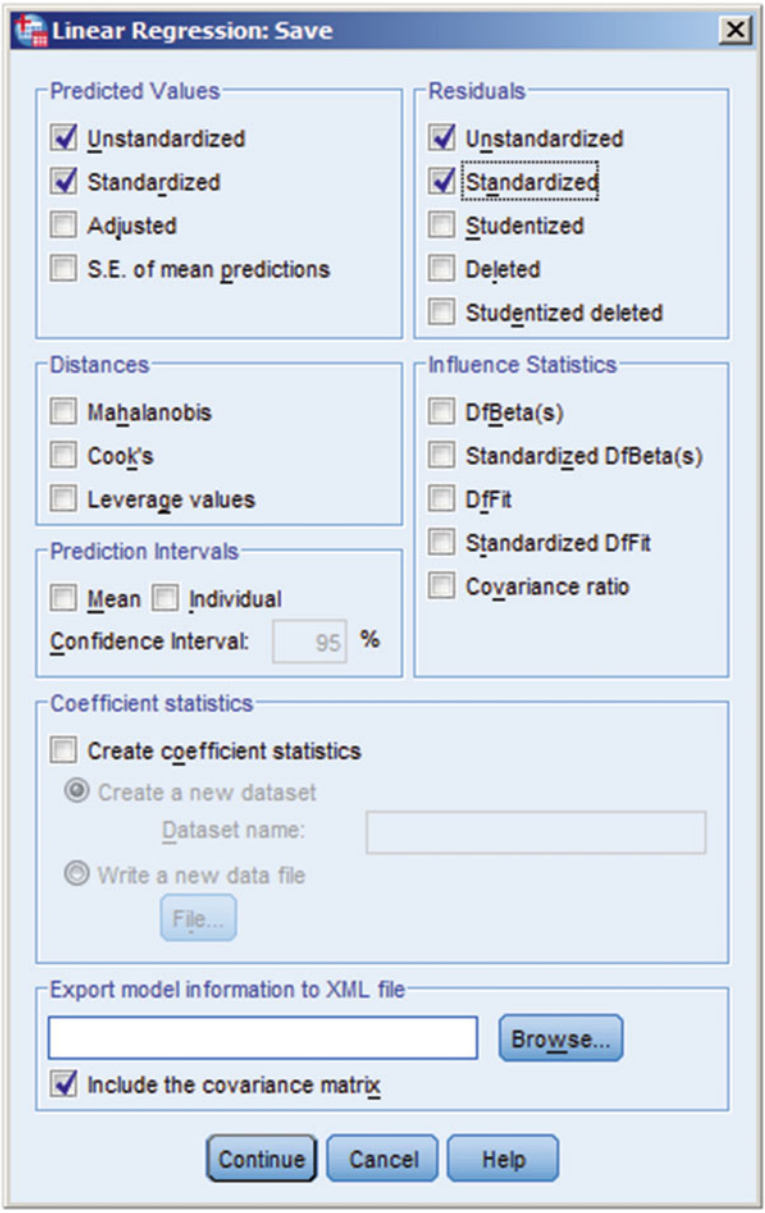


Fig. 1.10 The linear regression: save dialogue box

$$\hat{Y} = 54.154 + 0.9X_3 + 0.061X_2.$$

Both gradients are positive, as expected. Our sample gradients are denoted by b_i and based on the values of the b_i, we may wish to make inferences about their population equivalents. In particular, it is possible to test the null hypothesis:

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	monthly occupied bed days		Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	monthly x-ray exposures		Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a. Dependent Variable: monthly no. of hours devoted to care

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.988[a]	.976	.975	982.59984	
2	.992[b]	.984	.981	850.80206	2.618

a. Predictors: (Constant), monthly occupied bed days

b. Predictors: (Constant), monthly occupied bed days, monthly x-ray exposures

c. Dependent Variable: monthly no. of hours devoted to care

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Conllinearity Statistics	
		B	Std. Error	Beta	t	sig.	Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	-38.190	352.658		-.108	.915	-789.862	713.482		
	monthly occupied bed days	1.123	.045	.988	24.956	.000	1.027	1.219	1.000	1.000
2	(Constant)	54.154	307.671		.176	.863	-605.734	714.042		
	monthly occupied bed days	.900	.099	.792	9.073	.000	.687	1.113	.154	6.475
	monthly x-ray exposures	.061	.025	.214	2.451	.028	.008	.114	.154	6.475

a. Dependent Variable: monthly no. of hours devoted to care

Fig. 1.11 Part of the output from the stepwise regression procedure

H_0: a particular population gradient $\beta = 0$.

The test statistic for this is: b/standard error of which is distributed as a t statistic. For example, from the 'Coefficients' heading of Fig. 1.11, we may wish to test:

H_0: the population gradient of X_3 is zero, i.e. $\beta_3 = 0$, versus
H_1: $\beta_3 \neq 0$.

From Fig. 1.11, the test statistic has a numerical value $0.9/0.099 = 9.09$ as shown and using the above formula. This test statistic is highly significant ($p = 0.000$), so we reject the null hypothesis in favour of the alternative hypothesis. A similar conclusion is reached for the population gradient of X_2, which supports the stepwise procedure of introducing these two regressors. Under the headings 'Lower Bound' and 'Upper Bound' (for model 2), we see that a 95% confidence interval for β_3 is:

$$P(0.687 < \beta_3 < 1.113) = 0.95$$

and as expected from the hypothesis test, the value of zero for this gradient is not included in the above interval.

Figure 1.12 is a histogram of the residuals. It suggests that the residuals may not be normally distributed as is required, but the deviation from normality is not so

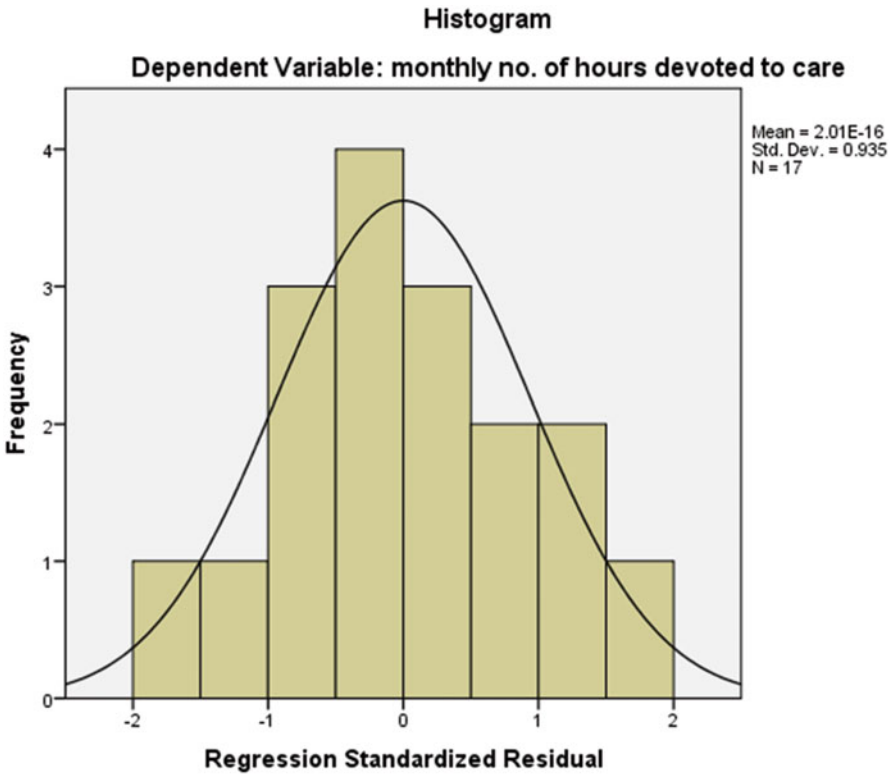


Fig. 1.12 A histogram of the regression residuals

marked as to be a problem. Figure 1.13 is the plot of the standardized residuals against the standardized predicted values. The residuals appear to exhibit a reasonably constant spread from left to right of this diagram, so the homoscedasticity assumption seems to be satisfactory. Figure 1.14 plots the residuals case by case and we note here (as well as in Fig. 1.13) that there are no outliers beyond ± 3 standard deviations from the regression.

Note that four new variables have been added to the data file. PRE_1 are the unstandardized predicted values of Y, RES_1 are the unstandardized residuals, ZPR_1 are the standardized predicted values and ZRE_1 are the standardized residuals. It is possible to construct the multiple line chart of Fig. 1.15, by plotting PRE_1 against Y. (The underscore XXX_1 means that this refers to the first regression run during this session). This diagram reinforces the general adequacy of the fit, which we expected from such a high multiple correlation coefficient.

It may be noted that if the *enter* procedure is used (i.e. all five regressors are entered en bloc), the equation of regression is:

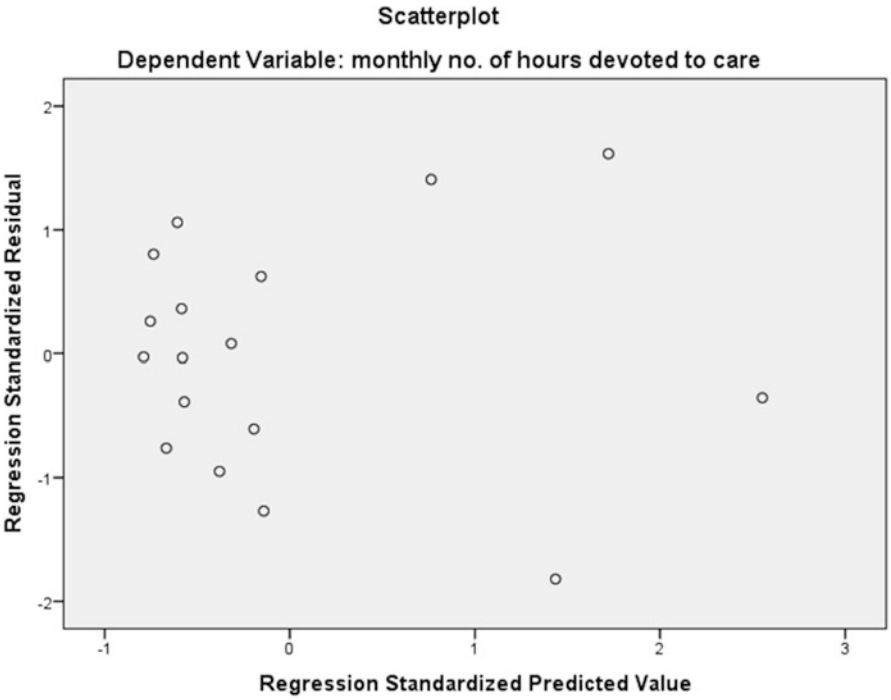


Fig. 1.13 A plot of standardized residuals against predicted values

$$\widehat{Y} = 2418.461 - 73.31X_1 - .00.8X_2 + 3.44X_3 + 10.757X_4 - 161.622X_5.$$

This model is unacceptable from the multicollinearity standpoint, in that the VIF's associated with X_1, X_3 and X_4 were found to be respectively 9542.0, 9992.7 and 23.3. One cannot consider models with all or combinations of these three variables included. The gradients associated with X_1 (mean daily patient load) and X_4 (eligible population in the area) in this model are negative, which makes little sense in the context of this problem and the sign of these gradients is doubtless due to multicollinearity.

1.4 The Cochrane-Orcutt Procedure for Tackling Autocorrelation

The effects of autocorrelation have already been alluded to. Ordinary least squares estimates of the regression coefficients are no longer minimum variance estimates. Hence, hypothesis tests and confidence intervals like those shown in the previous subsection become unreliable. To overcome the problem the researcher may turn to a model that specifically incorporates the autocorrelation structure.

Casewise Diagnostics[a]

Case Number	Std. Residual	monthly no. of hours devoted to care	Predicted Value	Residual
1	-.027	1566.52	1589.6956	-23.17558
2	-.761	1696.82	2344.5300	-647.71001
3	.260	2033.15	1811.8949	221.25510
4	.803	2603.62	1920.8191	682.80089
5	1.060	3611.37	2709.5706	901.79943
6	-.387	2613.27	2942.7598	-329.48980
7	-.034	2854.17	2883.4002	-29.23018
8	.362	3160.55	2852.5735	307.97653
9	-.949	3305.58	4112.7629	-807.18293
10	-1.270	4503.93	5584.4198	-1080.48985
11	.080	4571.89	4503.5598	68.33021
12	-.606	4741.40	5257.2593	-515.85932
13	.622	6026.52	5497.0180	529.50202
14	1.407	12343.81	11146.6829	1197.12709
15	-1.820	13732.17	15280.8198	-1548.64982
16	1.615	18414.94	17040.5555	1374.38454
17	-.354	21854.45	22155.8383	-301.38832

a. Dependent Variable: monthly no. of hours devoted to care

Fig. 1.14 A case by case analysis of the standardized residuals

Autocorrelation may result from the omission of significant regressors. If these regressors can be identified and incorporated in the model, the autocorrelation problem may be eliminated. However, there may be a real time dependency in the residuals, so the researcher has to turn to a number of estimation procedures available for this purpose. A widely used method for tackling autocorrelation is due to *Cochrane and Orcutt* (C-O).

Consider the simple linear regression model with *first order autocorrelated error*s. This phrase simply means that the residuals at time t are autocorrelated with the residuals at time $(t - 1)$ i.e.

$$e_t = \rho e_{t-1} + a_t$$

where e_t is the error at time t, a_t is an error at time t that is independent of e_t and ρ is a measure of the autocorrelation (ρ lying between -1 and $+1$ inclusive). Both e_t and a_t are assumed to have means of zero. Assume a simple regression model of the form:

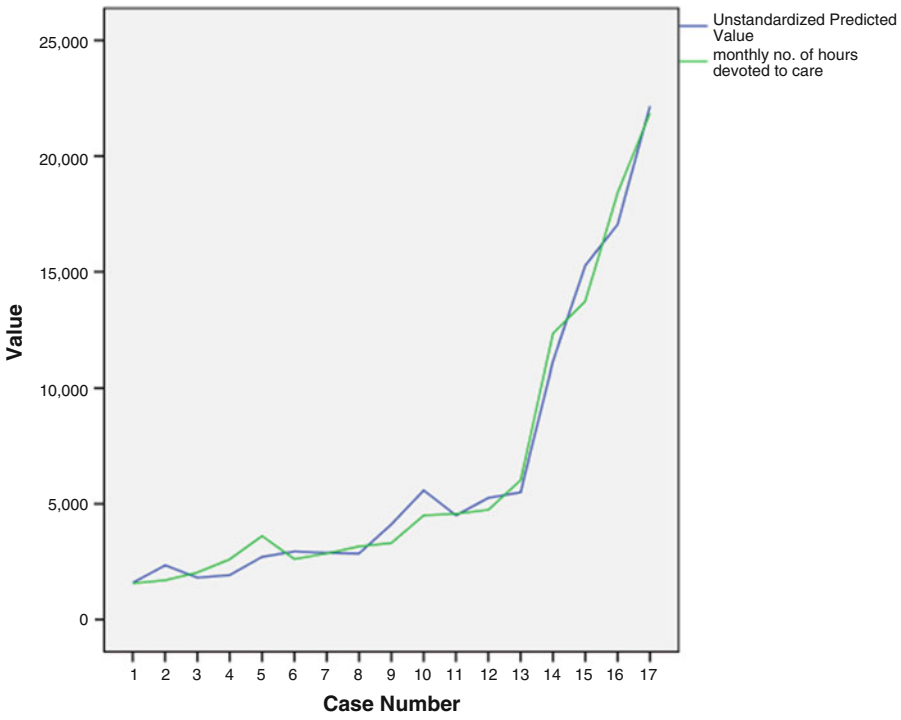


Fig. 1.15 A plot of observed versus predicted values

$$y_t = b_0 + b_1 x_t + e_t \ldots\ldots \tag{1.1}$$

The logic behind the C-O procedure is to transform the response variable so that:

$$\mathbf{Y}_t = y_t - \rho\, y_{t-1}$$

and using Eq. (1.1),

$$\begin{aligned} \mathbf{Y_t} &= b_0 + b_1 x_t + e_t - \rho(b_0 + b_1 x_{t-1} + e_{t-1}) \\ \mathbf{Y_t} &= b_0(1-\rho) + b_1(x_t - \rho x_{t-1}) + e_t - \rho e_{t-1} \\ \mathbf{Y_t} &= \beta_0 + \beta_1 \mathbf{X_t} + a_t \ldots\ldots \end{aligned} \tag{1.2}$$

where $\beta_0 = b_0(1-\rho)$, $\beta_1 = b_1$, $\mathbf{X_t} = x_t - \rho x_{t-1}$ and $a_t = e_t - \rho e_{t-1}$. Note that the error terms a_t in this (*reparametrized*) model of Eq. (1.2) are independent variables. Essentially speaking, Eq. (1.1) does not take into account of autocorrelation; Eq. (1.2) does. By transforming the regressor and response variables, we have generated a model that satisfies the usual regression assumptions and ordinary least squares may be applied to Eq. (1.2). The value of the autocorrelation

coefficient, ρ, is estimated by regressing e_t on e_{t-1}. The estimated value of ρ is simply the gradient of this latter regression.

The data in the IBM SPSS Statistics file COLGATE.SAV involve the percentage market share (MKTSHARE) for a particular brand of toothpaste (Y_t) and the dollar selling price (PRICEPND) per pound (X_t) for 20 consecutive months. We wish to build a regression model relating share of market in period t to the selling price in the same period.

Using ordinary least squares, the fitted model is:

$$\text{MKTSHARE} = 38.91 - 24.29^{*}\ \text{PRICEPND},$$

but the residual plot of Fig. 1.16 is suggestive of positive temporal autocorrelation. This is confirmed by the Durbin-Watson statistic $d = 1.136$ which is compared with the 5% critical levels for $n = 20$, of $d_L = 1.20$ and $d_U = 1.41$.

The C-O procedure is, therefore, used to estimate the model parameters. In recent versions of IBM SPSS Statistics, this is achieved by writing IBM SPSS syntax. The form of the appropriate syntax is:

```
AREG VARIABLES = dependent series name WITH independent series
names
```

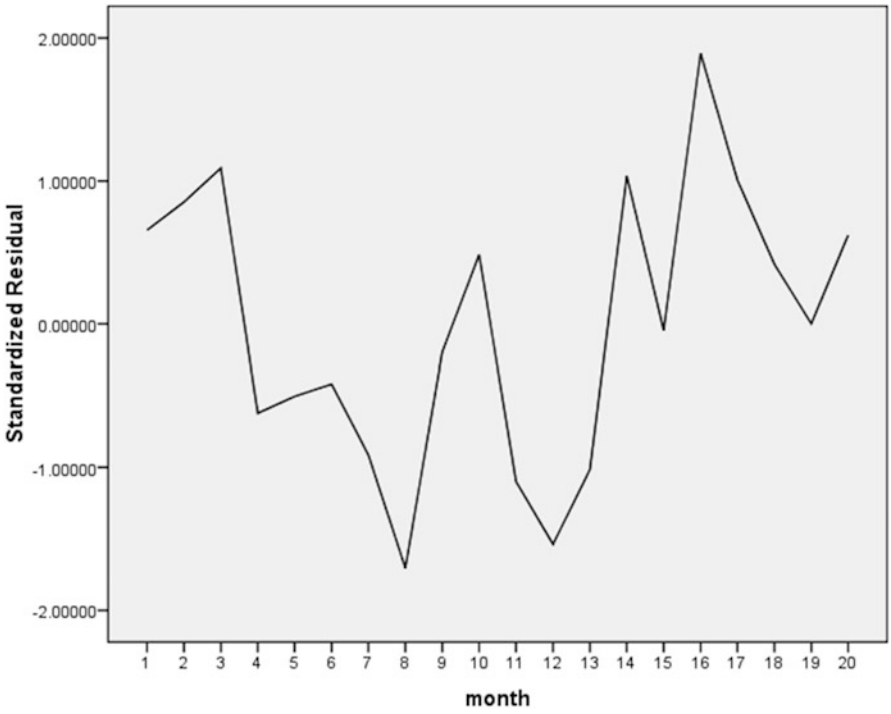


Fig. 1.16 A plot of the regression residuals over time

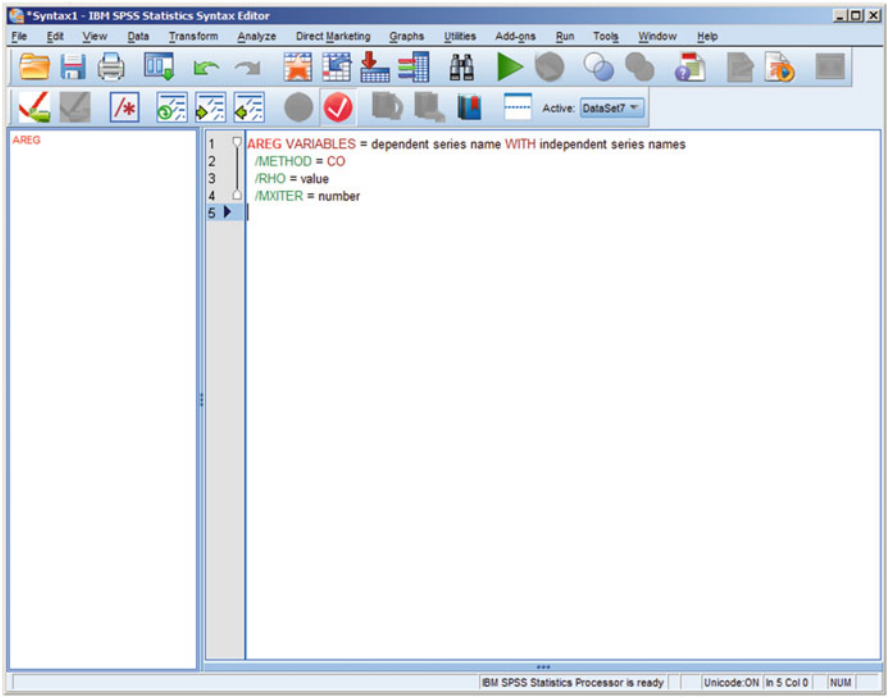


Fig. 1.17 The SPSS syntax editor

```
/METHOD = CO
/RHO = value
/MXITER = number
```

The user may specify the numerical value to be used as the initial ρ value during the iteration procedure. The default is $\rho = 0$ if this command is omitted. MXITER allows the user to specify the maximum number of iterations the procedure is allowed to cycle through in calculating estimates. The default is 10 iterations if this command is omitted.

To type in syntax, the user has to access the IBM SPSS Syntax Editor of Fig. 1.17, which is achieved via:

```
File
   New
     Syntax
```

And type

```
AREG VARIABLES = mktshare WITH pricepnd
```

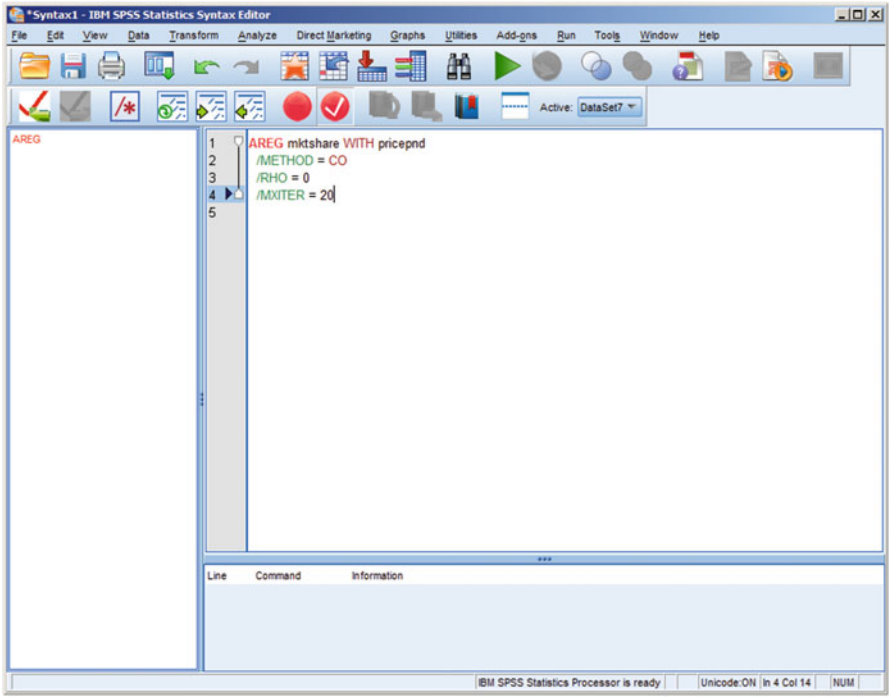


Fig. 1.18 The C-O procedure in IBM SPSS syntax

```
/METHOD = CO
/RHO = 0
/MXITER = 20
```

Then click on

```
Run
  All
```

As shown in Fig. 1.18. This will generates the output of Fig. 1.19 in the SPSS Viewer.

The estimated value of ρ is 0.413. The equation of regression obtained from the C-O procedure is:

$$\text{MKTSHARE} = 38.75 - 24.12\text{PRICEPND},$$

and the estimate of the gradient differs only slightly from that obtained via least squares and presented earlier. The Durbin-Watson statistic for the above

Auto correlation Coefficient

Rho (AR1)	Std.Error
.439	.218

The Prais-Winsten estimation method is used.

Model Fit Summary

R	R Square	Adjusted R Square	Std.Error of the Estimate	Durbin-Watson
.957	.916	.906	.398	1.909

The Prais-Winsten estimation method is used.

ANOVA

	Sum of Squares	df	Mean Square
Regression	29.413	1	29.413
Residual	2.698	17	.159

The Prais-Winsten estimation method is used.

Regression Coefficients

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std.Error	Beta	t	Sig
selling price ($)	-23.573	1.731	-.957	-13.614	.000
(Constant)	38.318	1.486		25.785	.000

The Prais-Winsten estimation method is used.

Fig. 1.19 Output from the Cochrane-Orcutt procedure

transformed model is d = 1.89 and comparing this with the appropriate table values we fail to reject the null hypothesis that the residuals are uncorrelated. Therefore, the Cochrane-Orcutt method has eliminated the original autocorrelation problem.

Chapter 2
Other Useful Topics in Regression

Once the user has grasped the fundamentals of multivariate linear regression, there are several related techniques that have application in business-orientated or socio-economic fields and which are described in this and the next chapter. Problems involving a *binary* (or *dichotomous*) response variable are common in any field where the researcher wishes to predict the presence or absence of a characteristic or outcome based upon a set of predictor variables. An example that is often cited is the presence or otherwise of coronary heart disease (binary, response variable) being dependent on smoking habits, diet, alcohol use and levels of exercise. Whether a person is accepted for a bank loan could depend on their income, employment history, monthly out-goings, amount of other loans etc. Problems involving a binary response variable may be examined by *binary logistic regression* as described in the first section of this chapter (2.1). The section that follows (2.2) briefly describes the method of *Multinomial Logistic Regression* which is similar to Logistic Regression, save that it is more general in that the response variable is not restricted to two categories.

The variables incorporated as predictors into our regression studies so far are *quantitative*; they have a well-defined scale of measurement. Sometimes, it is necessary to use *qualitative* or *categorical* variables as predictors in the regression. Examples of such categorical variables are employment status (employed or not), industrial shift worked (day, evening or night), sex (male or female) etc. Such variables have no natural scale of measurement. They are represented in statistical analyses by codes (e.g. male coded as '0' and female coded as '1') as discussed in the first volume of this guide. It is perfectly reasonable to postulate that a response variable might be related to a mix of quantitative and qualitative variables. For example, an employee's present salary might depend on age, gender, previous experience and level of education. Such problems are examined by *Dummy Regression* illustrated in the Sect. 2.3 while Sect. 2.4 discusses functional forms of regression models.

A. Aljandali, *Multivariate Methods and Forecasting with IBM® SPSS® Statistics*,
Statistics and Econometrics for Finance, DOI 10.1007/978-3-319-56481-4_2

2.1 Binary Logistic Regression

The standard regression model assumes that the dependent variable, Y, is measured quantitatively. The independent (or regressor) variables, X_i, may be measured quantitatively or qualitatively. (A dummy regressor is an example of a variable that is measured qualitatively). Binary logistic models apply to situations where the dependent variable is dichotomous in nature, taking a 0 or 1 value. For example, the dependent variable, Y, could be whether or not a person is unemployed ("employed" = 1, "unemployed" = 0). The regressors could include X_1 the average national wage rate, X_2 the individual's education, X_3 the national unemployment rate, X_4 family income etc. The question arises as to how we handle models involving dichotomous dependent variables.

2.1.1 The Linear Probability Model (LPM)

To fix ideas, consider the following simple model:

$$\widehat{Y} = \beta_1 + \beta_2 X$$

where X is family income (£ 000's) and Y is dichotomous, such that Y = 1 if the family owns a house and Y = 0 if the family does not own a house. Models such as the above which express the dichotomous Y as a linear function of the regressor variable(s) X are called *linear probability models*. However, there are problems with the assumptions that underpin regression when applying ordinary least squares to linear probability models.

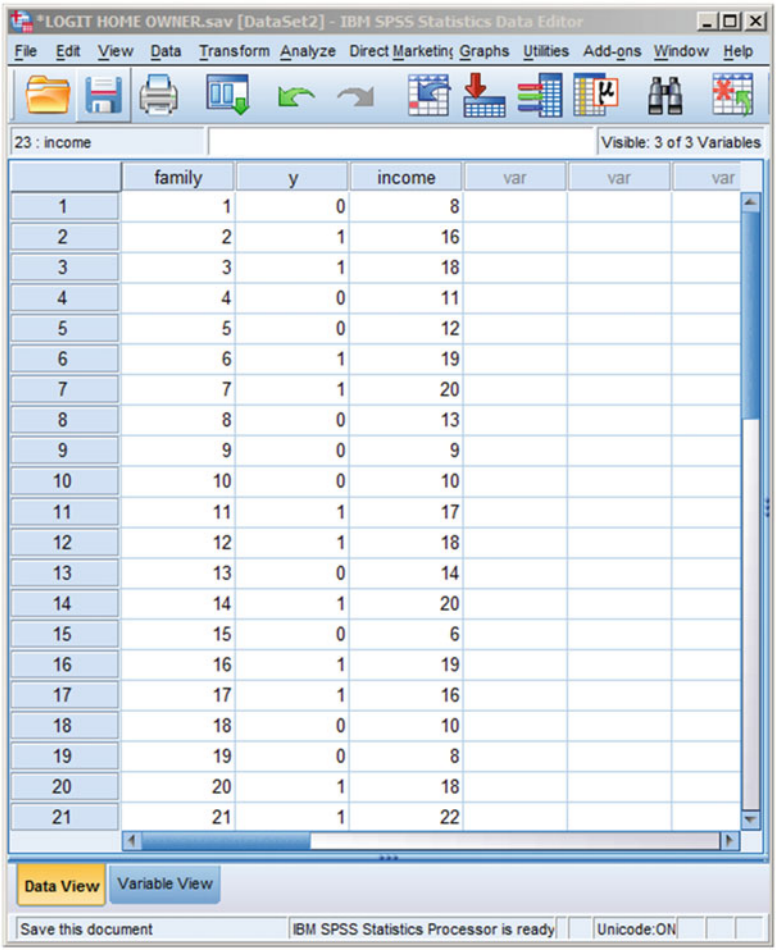
(a) The residuals are not normally distributed. To see this:

$$\text{Residual} = \text{Y} - \widehat{Y} = \text{Y} - \beta_1 - \beta_2 X$$
$$\text{When Y} = 1, \text{Residual} = 1 - \beta_1 - \beta_2 X$$
$$\text{When Y} = 0, \text{Residual} = -\beta_1 - \beta_2 X.$$

Consequently, the residuals cannot follow the normal distribution. (In fact, they are binomially distributed).

(b) It can no longer be maintained that the residuals are homoscedastic. It can be shown that the variance of the residuals depends on the value taken by X and is thus not homoscedastic.

(c) Consider the data in Fig. 2.1. Suppose a variable Y pertaining to home ownership is defined as above. When regression is applied to this LPM, we would obtain a result that:



	family	y	income	var	var	var
1	1	0	8			
2	2	1	16			
3	3	1	18			
4	4	0	11			
5	5	0	12			
6	6	1	19			
7	7	1	20			
8	8	0	13			
9	9	0	9			
10	10	0	10			
11	11	1	17			
12	12	1	18			
13	13	0	14			
14	14	1	20			
15	15	0	6			
16	16	1	19			
17	17	1	16			
18	18	0	10			
19	19	0	8			
20	20	1	18			
21	21	1	22			

Fig. 2.1 Home ownership and income (£ 000's)

$$\widehat{Y} = -0.874 + 0.098\ (\text{INCOME})$$

LPM's want to treat $\widehat{Y}$ like a probability. For example if for a particular income level, $\widehat{Y} = 0.93$, then we would guess that that family would be a home owner since the obtained result is closer to $Y = 1$ than it is to $Y = 0$. However and continuing this theme, if a family had an income of £12,000 (i.e. $X = 12$ in the LPM), then the predicted value of Y would be negative i.e. we would have negative probabilities. Indeed, it is possible to have an income level that coincides with a probability of home ownership in excess of 1. Consequently, the linear probability model is not recommended when the dependent variable is dichotomous.

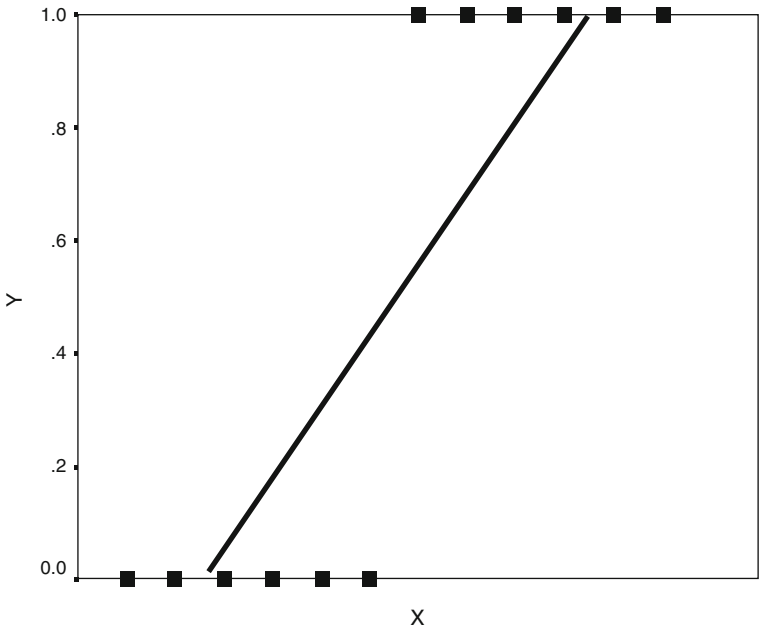


Fig. 2.2 Regression line when Y is dichotomous

(d) The value of the coefficient of determination as a measure of goodness of fit becomes questionable. Corresponding to a given value of income (X), Y is either 0 or 1. Therefore, all values of Y will either lie along the X-axis or along the line corresponding to $Y = 1$ (see Fig. 2.2). Consequently, no linear probability model is expected to fit such a scatter well. The coefficient of determination is likely to be much lower than 100% for such models (even if the model is constrained to lie between $Y = 0$ and $Y = 1$).

There are ways to overcome some of the problems associated with the linear probability model. However, there remains a fundamental problem that is not very attractive because the model assumes that Y (or probability) increases linearly with X. This implies that the impact of X remains constant throughout. Thus, in the home ownership example, we find that as X increases by a unit (£1000), the probability of home ownership increases by 0.098. This is the case whether income is £8000, £80,000 or £800,000. This seems patently unrealistic. At a very low income, a family will not own a house. At a sufficiently high income say X^*, people will be most likely to own a house. Beyond X^*, income will have little effect on the probability of owning a home. Thus at both ends of the income distribution, the probability of owning a home will be virtually unaffected by a small increase in X. The probability of owning a home is nonlinearly related to income.

2.1.2 The Logit Model

Now consider the following representation for home ownership, in which P represents the probability that a family owns a home i.e. P(Y = 1):

$$P = \frac{1}{1 + \exp - (\beta_1 + \beta_2 X)} \ldots\ldots \tag{2.1}$$

in which exp. − (X) = e^X. Equation (2.1) is called *the logistic distribution function*, which is plotted below.

As shown in Fig. 2.3, Eq. (2.1) permits P to range only between 0 and 1, thus solving one of the problems associated with the linear probability model. If P is the probability of owning a home, then (1 − P) is the probability of not owning a home and:

$$\begin{aligned} {}_{1-P} &= 1 - \frac{1}{1 + \exp - (\beta_1 + \beta_2 X)} = \frac{1 + \exp - (\beta_1 + \beta_2 X) - 1}{1 + \exp - (\beta_1 + \beta_2 X)} \\ &= \frac{\exp - (\beta_1 + \beta_2 X)}{1 + \exp - (\beta_1 + \beta_2 X)} = \frac{{}^{1}/_{\exp(\beta_1 + \beta_2 X)}}{1 + {}^{1}/_{\exp(\beta_1 + \beta_2 X)}} = \frac{{}^{1}/_{\exp(\beta_1 + \beta_2 X)}}{\exp(\beta_1 + \beta_2 X) + {}^{1}/_{\exp(\beta_1 + \beta_2 X)}} \\ &= \frac{1}{1 + \exp(\beta_1 + \beta_2 X)} \ldots\ldots \end{aligned} \tag{2.2}$$

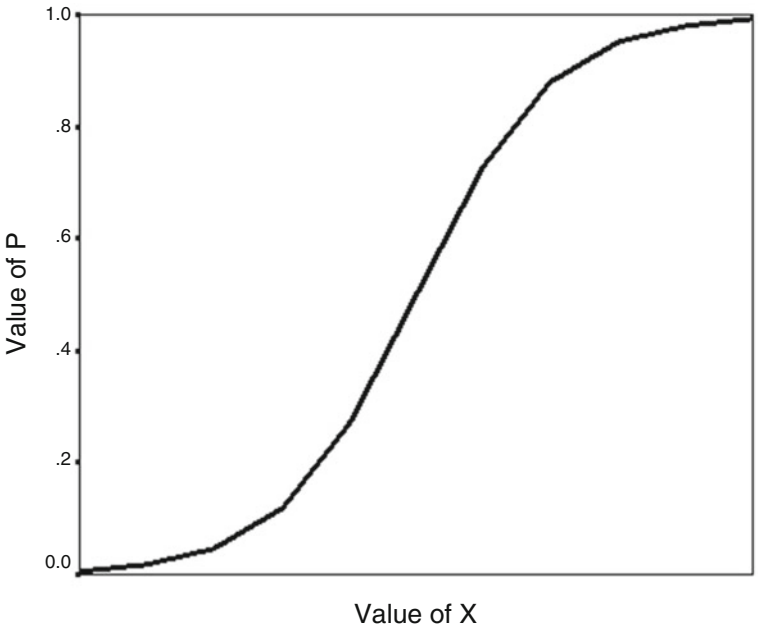


Fig. 2.3 A plot of the logistic distribution function

Therefore, using Eqs. (2.1) and (2.2), we can write:

$$\frac{P}{1-P} = \frac{1}{1+\exp-(\beta_1+\beta_2 X)}.[1+\exp(\beta_1+\beta_2 X)]$$
$$\frac{P}{1-P} = \frac{1}{[\exp(\beta_1+\beta_2 X)+1]/\exp(\beta_1+\beta_2 X)}.[1+\exp(\beta_1+\beta_2 X)]$$
$$\frac{P}{1-P} = \exp(\beta_1+\beta_2 X)$$

and taking natural logarithms (i.e. base e):

$$\ln\left(\frac{P}{1-P}\right) = \ln\left[\exp(\beta_1+\beta_2 X)\right]$$
$$\ln\left(\frac{P}{1-p}\right) = \beta_1+\beta_2 X \ldots.. \tag{2.3}$$

because $\ln(e^X) = X\ln e = X$.

The left hand side of Eq. (2.3) is called the *logit* and the whole equation is called the *logit model*. The left hand side is the logarithm of the probability that a family owns a home against the probability that it does not. This is called the *logarithm of the odds ratio*. Naturally the logit model of Eq. (2.3) may be extended to the multivariate case:

$$\ln\left(\frac{P}{1-p}\right) = \beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\ldots\ldots$$

2.1.3 Applying the Logit Model

The logit model of Eq. (2.3), where X is income (in £000's), was applied to the data in Fig. 2.1. (Computer packages use a method called "maximum likelihood" to generate the logit coefficients). The resultant model was:

$$\ln\left(\frac{\widehat{P}}{1-\widehat{P}}\right) = -1.6587+0.0792(\text{INCOME})\ldots.. \tag{2.4}$$

The first family in Fig. 2.1 had an income of £8000 (X = 8). Inserting this value of X into Eq. (2.4):

$$\ln\left(\frac{\widehat{P}}{1-\widehat{P}}\right) = -1.0251, \text{whereby} \frac{\widehat{P}}{1-\widehat{P}} = e^{-1.0251} = 0.3588.$$

$$\text{Hence}, \widehat{P} = 0.3588 - 0.3588\widehat{P}$$
$$1.3588\widehat{P} = 0.3588$$
$$\widehat{P} = 0.2641.$$

The logit model estimates that there is a probability of 0.2641 that this family owns its home. It is possible to compute the change in probability of owning a home associated with a one unit (£1000) increase in income for this family who currently earn £8000. The change in probability is given by:

$$\widehat{\beta}_2.\widehat{P}\left(1-\widehat{P}\right) = (\mathbf{0.0792})^*(\mathbf{0.2641})^*(\mathbf{0.7359}) = \mathbf{0.0139}.$$

If this family's income increases by £1000, there is an extra 1.39% chance that they will become a house owner. This extra probability is not constant, but varies with income level. The former was a disadvantage of the linear probability model.

2.1.4 The Logistic Model in IBM SPSS Statistics

An early, classic application of the logit model was in examining the choice of fertiliser used by Philippine farmers. The data are in the IBM SPSS Statistics data file called FERTILISER.SAV. The dependent variable to be explained is FERUSE – a binary variable equal to one if fertiliser is used and equal to zero otherwise. The explanatory variables are:

- CREDIT – the amount of credit (per hectare) held by the farmer,
- DMARKET – the distance of the farm to the nearest market,
- HOURMEET – no. of hours the farmer spent with an agricultural expert,
- IRSTAT – a dummy variable = 1 if irrigation is used, = 0 otherwise and
- OWNER – a dummy variable = 1 if the farmer owns the land, = 0 otherwise.

(There is an extra variable in this file called QFER, which records the amount of fertiliser used if the farmer indeed uses it). Four hundred and ninety one farms were examined. Binary logistic regression is accessed via:

```
Analyze
   Regression
      Binary logistic...
```

which generates the *Logistic Regression* dialogue box of Fig. 2.4.

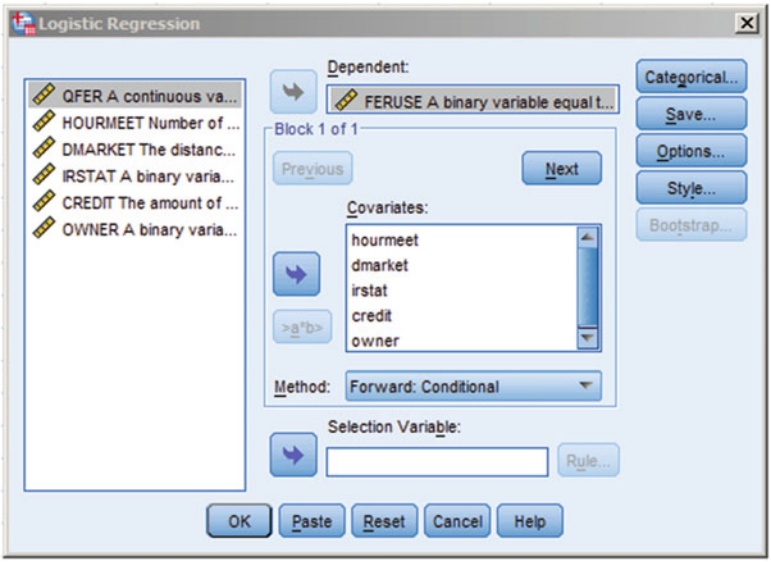


Fig. 2.4 The logistic regression dialogue box

The binary variable FERUSE is the Dependent variable and the five independent variables above are called Covariates in the context of logistic regression. Note that in the 'Method' box, the user can choose the Enter method (all independent variables entered simultaneously, Forward selection or Backward removal).

Clicking the Save... button generates the *Logistic Regression: Save* dialogue box of Fig. 2.5. This dialogue box permits the user to save the probabilities of group membership. If that probability is in excess of 0.5, the associated case is classified as being a member of the group that is coded as '1'; if that probability is less than 0.5, the case is deemed to be a member of the group coded as '0'. Standardized and unstandardized residuals can also be added to the active data file.

Clicking the Options... button in Fig. 2.4 produces the dialogue box of Fig. 2.6, where the Hosmer-Lemeshow test of model adequacy should be selected. This is discussed below. Note that a cut-off probability (Classification cutoff) of 0.5 is selected in Fig. 2.6: if the probability is above 0.5 group '1' membership is predicted and vice versa. Upon clicking the Continue and OK buttons, the results from the logistic regression are generated. Figure 2.7 presents the results for the first six farmers in the data file.

Based on the logistic analysis using the five covariates listed on page 33, the first farmer has a probability of 0.23519 of being in the group coded as '1' i.e. a fertilizer user. This probability appears under the heading PRE_1. Since this probability is less than 0.5, the logistic model predicts that the farmer will not be a fertilizer user and therefore allocates him to the group coded as '0', shown under the heading PGR_1. Examination of the FERUSE variable shows that he is indeed a group '0' member i.e. a non-fertilizer user. The second framer has a probability of 0.22654 of

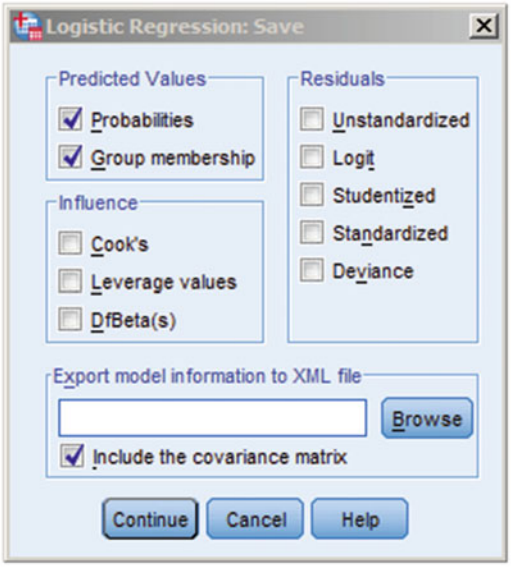


Fig. 2.5 The logistic regression: save dialogue box

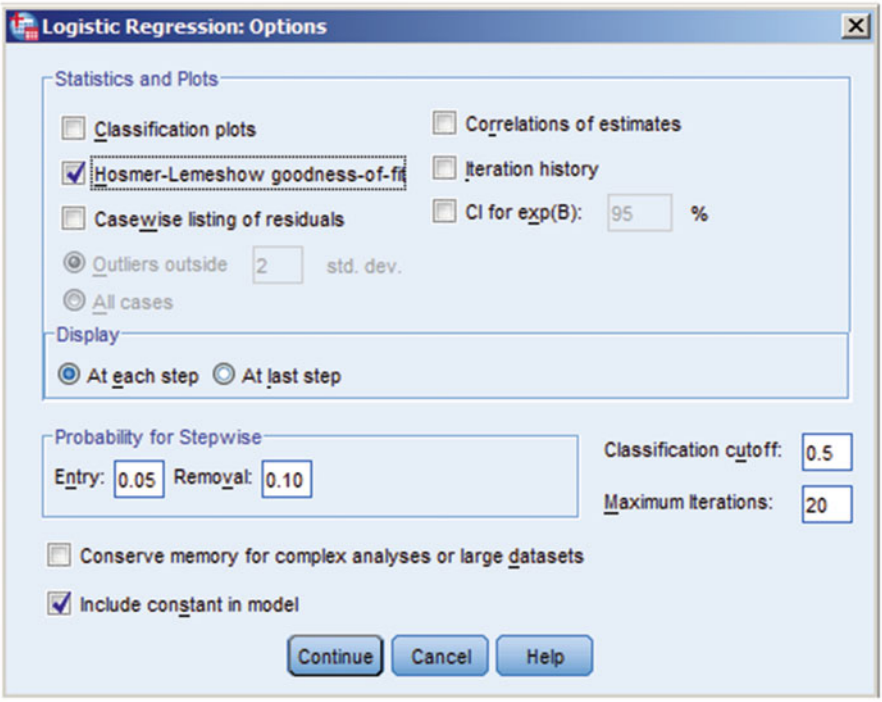
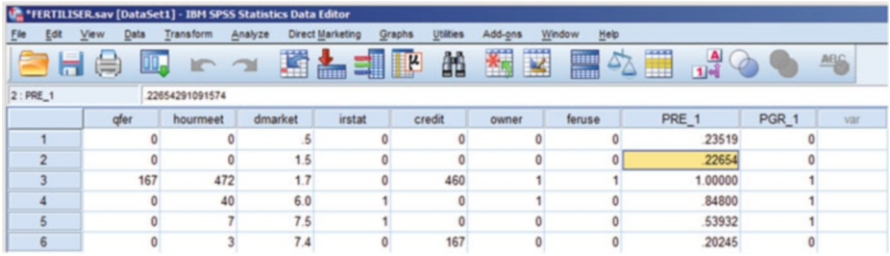


Fig. 2.6 The logistic regression: options dialogue box



	qfer	hourmeet	dmarket	irstat	credit	owner	feruse	PRE_1	PGR_1	var
1	0	0	.5	0	0	0	0	.23519	0	
2	0	0	1.5	0	0	0	0	.22654	0	
3	167	472	1.7	0	460	1	1	1.00000	1	
4	0	40	6.0	1	0	1	0	.84800	1	
5	0	7	7.5	1	0	0	0	.53932	1	
6	0	3	7.4	0	167	0	0	.20245	0	

Fig. 2.7 The first six cases in the active data file

being a fertilizer user and he is correctly classified. However, the fourth framer has a probability of 0.84800 of being a fertilizer and he is incorrectly classified under the heading PGR_1 as a group '1' member whereas he is a FREUSE = '0' member.

Figure 2.8 presents part of the output presented in the IBM SPSS STATISTICS Viewer. Via forward entry, the covariates are entered in the order of their importance. At step 1, the most important determinant of fertilizer usage is whether or not the farmer uses irrigation (IRSTAT). If the farmer was forward-thinking in using irrigation, he was also forward-thinking in applying chemical fertilizer. At step five, all the covariates are in the model and are significant (at $p < 0.05$). Note that the frictional force of distance has a negative impact on fertilizer use as exemplified by the negative coefficient attached to the variable DMARKET.

From Fig. 3.7, the equation of the logistic model is:

$$\ln\left(\frac{\mathrm{P}}{1-\mathrm{p}}\right) = -1.155 + 0.557(\mathrm{OWNER}) + 1.480(\mathrm{IRSTAT}) + \ldots..$$
$$+ 0.0004(\mathrm{CREDIT}),$$

from which the probabilities of group membership may be computed. For example, if for one particular farmer, HOURMEET = 30, DMARKET = 6, CREDIT = 200, IRSTAT = 1 and OWNER = 1, then from the above equation:

$$\ln\left(\frac{P}{1-p}\right) = 1.507,$$
$$\frac{P}{1-P} = e^{1.507} = 4.513$$
$$P = 4.513 - 4.513P$$
$$5.513P = 4.513$$
$$\text{whereby } P = 0.819.$$

There is an over 80% chance that this particular farmer is a fertiliser user. Assessment of the logistic model's forecasting adequacy may be made by

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	irstat	1.601	.196	66.517	1	.000	4.957
	Constant	-.998	.144	48.002	1	.000	.369
Step 2[b]	hourmeet	.034	.013	7.040	1	.008	1.035
	irstat	1.540	.199	59.715	1	.000	4.664
	Constant	-1.084	.148	53.360	1	.000	.338
Step 3[c]	hourmeet	.031	.013	6.241	1	.012	1.032
	irstat	1.489	.201	54.733	1	.000	4.432
	credit	.000	.000	5.401	1	.020	1.000
	Constant	-1.158	.153	57.273	1	.000	.314
Step 4[d]	hourmeet	.030	.013	5.683	1	.017	1.030
	irstat	1.521	.204	55.681	1	.000	4.579
	credit	.000	.000	6.314	1	.012	1.000
	owner	.522	.207	6.373	1	.012	1.686
	Constant	-1.392	.183	57.695	1	.000	.249
Step 5[e]	hourmeet	.028	.012	5.362	1	.021	1.029
	dmarket	-.049	.022	4.695	1	.030	.952
	irstat	1.480	.205	52.067	1	.000	4.394
	credit	.000	.000	7.031	1	.008	1.000
	owner	.557	.209	7.109	1	.008	1.745
	Constant	-1.155	.208	30.897	1	.000	.315

a. Variable(s) entered on step 1: irstat.

b. Variable(s) entered on step 2: hourmeet.

c. Variable(s) entered on step 3: credit.

d. Variable(s) entered on step 4: owner.

e. Variable(s) entered on step 5: dmarket.

Fig. 2.8 Variables in the final logistic model

examining the 'Classification Table' of Fig. 3.8 and which is part of the output in the IBM SPSS Statistics Viewer. At step 5 in the above table, 266 farmers are in the dependent variable = 0 category i.e. the FERUSE = 0 group – they do not use fertilizer. One hundred and eighty three of these farmers were predicted by the logistic model to have a probability of fertilizer use below the cut-off probability of 0.5. Hence, 183 (68.80%) of the farmers in the FERUSE = 0 group were predicted correctly. Therefore, 31.20% of the FERUSE = 0 group were incorrectly classified by the logistic model (Figs. 2.9 and 2.10).

Similarly, there were 225 farmers observed to be in the dependent variable = 1 category i.e. FERUSE = 1.159 (70.67%) farmers had probabilities above the cut-off point of 0.5 and were consequently correctly classified. Overall, 342 farmers (183 + 159) have been correctly classified into their FERUSE = 0 or FERUSE = 1 groups. This is an overall success rate of 342 out of 491 farmers or 69.7%.

Classification Table[a]

Observed			Predicted		
			FERUSE A binary variable equal to 1 if fertiliser is used- 0 otherwise		Percentage Correct
			0	1	
Step 1	FERUSE A binary variable equal to 1 if fertiliser is used- 0 otherwise	0	179	87	67.3
		1	66	159	70.7
	Overall Percentage				68.8
Step 2	FERUSE A binary variable equal to 1 if fertiliser is used- 0 otherwise	0	177	89	66.5
		1	63	162	72.0
	Overall Percentage				69.0
Step 3	FERUSE A binary variable equal to 1 if fertiliser is used- 0 otherwise	0	176	90	66.2
		1	61	164	72.9
	Overall Percentage				69.2
Step 4	FERUSE A binary variable equal to 1 if fertiliser is used- 0 otherwise	0	175	91	65.8
		1	61	164	72.9
	Overall Percentage				69.0
Step 5	FERUSE A binary variable equal to 1 if fertiliser is used- 0 otherwise	0	183	83	68.8
		1	66	159	70.7
	Overall Percentage				69.7

a. The cut value is .500

Fig. 2.9 The classification table associated with logistic regression

Fig. 2.10 The Hosmer-Lemeshow test

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	.000	0	.
2	11.786	5	.038
3	27.492	6	.000
4	18.258	7	.011
5	7.156	8	.520

Beside the Classification Table, another assessment of the adequacy of the logit model is the Hosmer-Lemeshow (HL) goodness of fit test. The HL test has as its null H_0: the model adequately predicts group membership and the null is rejected if the associated level of significance is less than 5% or 0.05. In the above example, it is found that HL = 7.156 with significance 0.520, so the null would not be rejected and the logistic model deemed an adequate representation for the data.

2.1.5 A Financial Application of the Logistic Model

The logistic model was first in Finance to predict the probability that a given firm will be a merger target. The code '1' is used if a firm was a merger target and '0' if it was not. The subsequent logistic model can presented as follows:

$$\ln\left(\frac{P}{1-p}\right) = \beta_1 + \beta_2(\text{PAYOUT}) + \beta_3(\text{TURNOV}) + \beta_4(\text{SIZE}) + \beta_5(\text{LEV}) + \beta_6(\text{VOL})$$

where:

PAYOUT = payout ratio (dividend/earnings),
TURNOV = asset turnover (sales/total asset),
SIZE = market value of equity,
LEV = leverage ratio (long-term debt/total assets) and
VOL = trading volume in the year of acquisition.

β_2, β_4 and β_5**are expected** to be negative and β_6 to be positive while β_3 to be positive or negative. Based on a sample of 24 merged firms (coded as '1') and 43 non-merged firms (coded as '0'), the results shown in Table 2.1 were obtained:

The estimated coefficients had the expected signs and all but two were statistically significantly different from zero. The results, for example, illustrate that the higher the turnover and the larger the size, the lower are the odds of the firm being a takeover target. On the other hand, the higher the trading volume, the greater the odds of being a merger candidate, for high-volume firms may imply lower acquisition transaction costs due to marketability. Based on these analyses, we conclude that one of the important factors affecting the firm's attractiveness is the inability of managers to generate sales per unit of assets. Moreover, low turnover must be accompanied by any one or a combination of low payout, low financial leverage, high trading volume and smallness in aggregate market value in order to produce a high probability of merger.

Table 2.1 Logistic estimate results for the Dietrich and Sorenson study

Variable	Coefficient	Standard error	t-value
PAYOUT	−0.74	0.29	−2.51**
TURNOV	−11.64	3.86	−3.01**
SIZE	−5.74	2.39	−2.40**
LEV	−1.33	0.97	−1.37
VOL	2.55	1.58	1.62
Intercept	−10.84	3.40	−3.20**

** Significant at $p < 0.01$

2.2 Multinomial Logistic Regression

Now we consider situations in which the response variable has more than the two categories of the binary case. Variables with more than two categories are called *polychotomous* rather than dichotomous. Such situations are analysed by multinomial logistic regression which is very similar to the binary logistic regression of the previous section, save that it is more general since the dependent or response variable is not restricted to two categories. For example, a survey of opinions about a proposed road improvement scheme could produce responses (Y) of "against" (code '0', say), "undecided" (code '1') or "in favour" (code '2'). Multinomial logistic regression could be used to see if Y might depend on the resident's proximity to the road (X_1), the resident's age (X_2), whether or not the resident has children (a categorical variable X_3) etc. As another example, in a Marketing scenario, a firm might be examining consumer attitudes towards several types of product packaging (the polychotomous response variable). Such attitudes may well depend on consumers' income levels, their age, purchase purpose etc. I do not propose presenting an example on multinomial regression in IBM SPSS Statistics as it is so similar to the binary example, plus it does require a deeper statistical knowledge for the fullest use of the method.

Suffice it to say, that multinomial logistic regression is accessed via:

```
Analyze
    Regression
        Multinomial logistic...
```

which gives ride to the *Multinomial Logistic Regression* dialogue box of Fig. 2.11. In the example of attitudes to the road scheme, OPINION is the dependent variable. The dialogue box of Fig. 2.11 requires 'Factors' and/or 'Covariates'. Covariates are simply the quantitative variables (continuous measurement) in the analysis, such as LOCATION the distance of the resident from the road. Factors are categorical variables like CHILDREN – whether or not the resident has children. Such a categorical variable might be coded as '0' and '1'. The analysis proceeds as in the binary case. Predicted probabilities, group membership, raw and standardized residuals and goodness of fit statistics may be generated and/or saved.

2.3 Dummy Regression

Dummy variables are most commonly used when a researcher wants to insert nominal scale categorical variables into a regression equation. A set of dummy variables is created by treating each category of a categorical variable as a separate variable and assisting arbitrary scores for all cases depending on their presence or absence in each category. Suppose that an engineer wishes to relate the effective

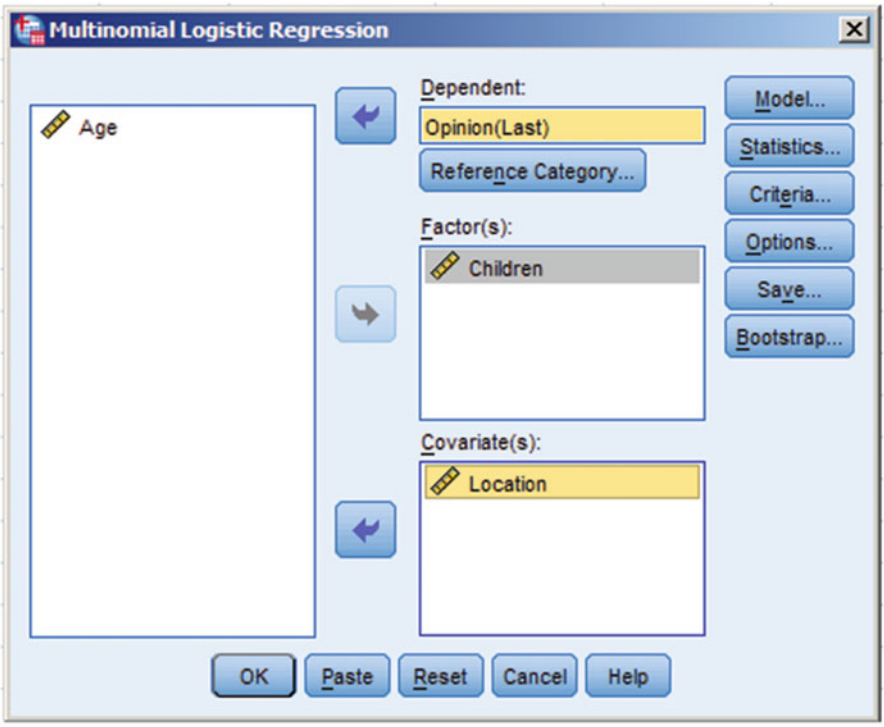


Fig. 2.11 The multinomial logistic regression dialogue box

life time (Y) of a cutting tool used on a lathe to lathe speed in revolutions per minute (X_1) and the type of cutting tool used (X_2). This second variable is categorical and has two levels, tool types A and B. We may invoke a *dummy variable* X_2 with the codes:

$X_2 = 0$ if the observation is from tool type A
$X_2 = 1$ if the observation is from tool type B

The choice of '0' and '1' to identify the levels of qualitative variable is arbitrary. We thus have a model of the form:

$$Y = b_0 + b_1X_1 + b_2X_2 + e$$

Where e represents the error term. To interpret the coefficients in the model, consider first tool type A, for which $X_2 = 0$. The regression model becomes:

$$Y = b_0 + b_1X_1 + e$$

The relationship between tool life and lathe speed for tool type A is a straight line with intercept b_0 and gradient b_1. For tool type B, $X_2 = 1$ and the regression model is:

$$Y = b_0 + b_1X_1 + b_2 + e$$
$$Y = (b_0 + b_2) + b_1X_1 + +e$$

That is, for tool type B, the relationship between tool life and lathe speed is a straight line with gradient b_1, but with intercept $(b_0 + b_2)$. These two responses describe two parallel regression lines with different intercepts.

One may generalise this approach to qualitative factors with any number of levels. Suppose there were three tool types, then two dummy variables are necessary:

X_2	X_3	
0	0	If the observation is from tool type A
1	0	If the observation is from tool type B
0	1	If the observation is from tool type C

And the regression model is:

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + e$$

Generally, a qualitative categorical variable with L levels is represented by $(L - 1)$ dummy variables, each taking on the values 0 and 1.

The file TOOLLIFE.SAV contains data about the lifetimes of the cutting tools (LIFETIME), the lathe (SPEED) and the type of cutting tool (TYPE). Figure 2.12 presents a scatterplot of the lifetimes of the two types of tool. This is created by selecting:

```
Graphs
   Legacy dialogs
        Scatter/Dot
           Simple Scatter
```

In the Scatterplot dialogue box, we want the two tool types plotted with different symbols, so under the heading 'Set Markers By' enter the variable TYPE. Inspection of the scatter diagram indicates that two different regression lines are required to model adequately these data, with the intercepts depending on the type of tool used. The dummy variable is coded as before, $X_2 = 0$ if it is a type A tool and $X_2 = 1$ if it is type B. The equation of regression is obtained in the usual manner:

```
Analyse
        Regression
               Linear
```

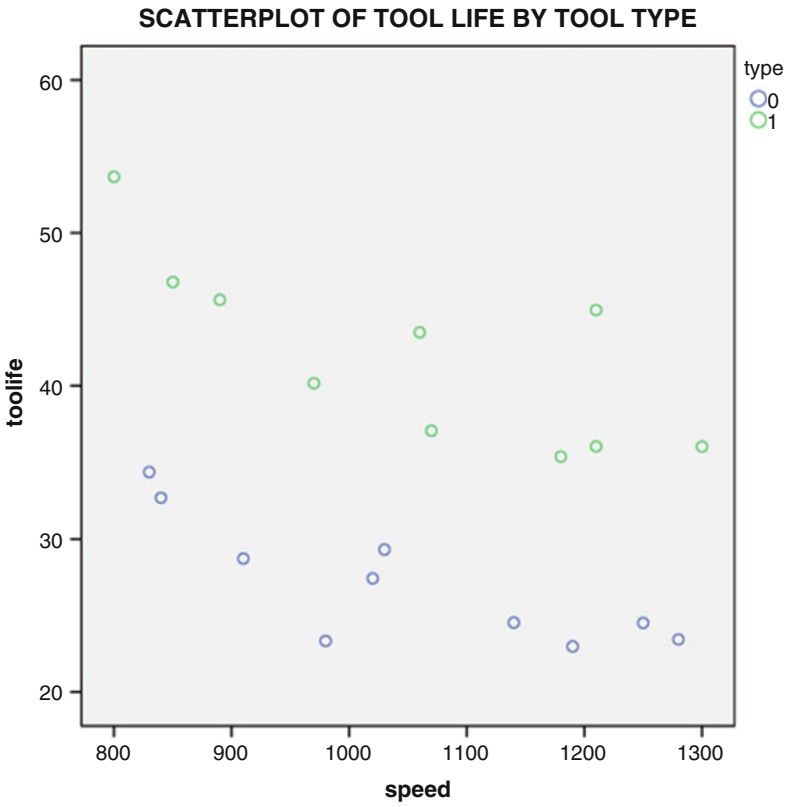


Fig. 2.12 Scatterplot of tool life by tool type

The least squares fit using the Enter selection method is shown in Fig. 2.13 and is:

$$\overbrace{TOOLLIFE} = 52.146 - 0.024^{*}\ \text{SPEED} + 14.957^{*}\text{TYPE},$$

With $r^2 = 0.900$. The t statistics show that both regression coefficients are significantly different from zero. The negative coefficient associated with the variable SPEED makes sense, in that we expect TOOLLIFE to decrease as SPEED increases. The parameter (14.957) associated with TYPE is the change in mean TOOLLIFE resulting from a change from tool type A to tool type B. A 95% confidence interval could be selected for the equivalent population coefficient and it would be found to be:

$$11.879 < \beta_2 < 18.035$$

Model Summary

Model	R	R Square	Adjusted R Square	Std.Error of the Estimate
1	.941[a]	.885	.871	3.261

a. Predictors: (Constant), type, speed

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std.Error	Beta	t	Sig.
1	(Constant)	52.416	4.917		10.659	.000
	speed	-.024	.005	-.433	-5.257	.000
	type	14.957	1.459	.845	10.254	.000

a. Dependent Variable: toolife

Fig. 2.13 Part of the output from dummy regression

A plot of the residuals against the fitted values (both standardized) is shown in Fig. 2.14. These two variables were saved as part of the dummy regression procedure and the scatterplot constructed. The type B residuals in Fig. 2.14 exhibit slightly more scatter than those of type A, implying that there may be a mild inequality of variance problem. The normal probability plot revealed no model inadequacy in this respect.

Since two different regression lines are used to model the relationship between tool life and lathe speed, we could initially fit two separate straight line models instead of a single model with a dummy variable. However, the single-model approach is preferred because the analyst has only one equation to work with instead of two, a much simpler practical result. Furthermore, since both straight lines are assumed to have the same gradient, it makes sense to combine the data from both types to produce a single estimate of this common parameter. This approach also gives one estimate of the common residual variance, σ^2.

Suppose that we expect the regression lines relating tool life to lathe speed to differ in both intercept and gradient. It is possible to model this situation with a single regression equation by using dummy variables. The model is:

$$\text{TOOLLIFE} = b_0 + b_1{}^* \text{ SPEED} + b_2{}^* \text{ TYPE} + b_3{}^* \text{ SPEED}{}^* \text{ TYPE}$$

We observe that a cross product term between lathe speed and the dummy variable (SPEED*TYPE) has been added to the model. To interpret the coefficient b_3 for this model, first consider a type A tool for which TYPE = 0, then the above model becomes:

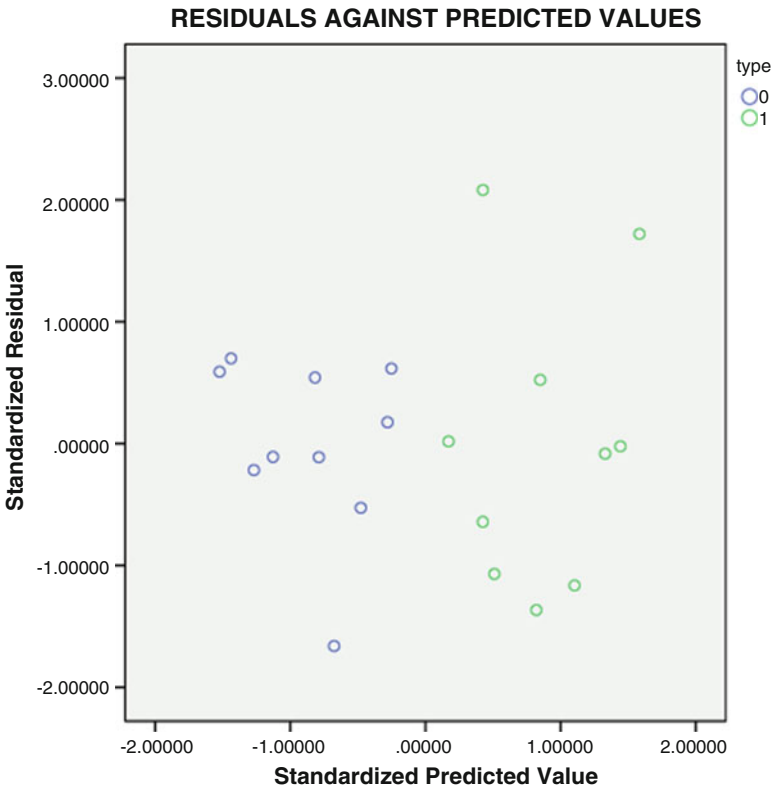


Fig. 2.14 A plot of residuals against predicted values

$$\text{TOOLLIFE} = b_0 + b_1{}^* \text{ SPEED},$$

A line with intercept b_0 and gradient b_1. For tool type B, TYPE = 1 and our model becomes:

$$\text{TOOLLIFE} = (b_0 + b_2) + (b_1 + b_3)^* \text{ SPEED}$$

A line with intercept now of (b_0+b_2) and gradient now of (b_1 +b_3). Hence, the parameter b_2 reflects the change in intercept associated with changing from tool type A to tool type B and b_3 indicates the change in gradient associated with changing from tool type A to tool type B. Fitting this model is equivalent to fitting two separate regression equations. An advantage to the use of the dummy variable is that tests of hypotheses may be performed directly. For example, to test whether two regression lines have the same intercept but possibly different gradients, then by reference to the above equation, we should examine:

$$H_0 : \beta_2 = 0$$
$$H_1 : \beta_2 \neq 0$$

To test that the two regression lines have a common gradient, but possibly different intercepts, the hypotheses are:

$$H_0 : \beta_3 = 0$$
$$H_1 : \beta_3 \neq 0$$

It is simple matter to compute the cross product term (variable name CROSSPRO, say) = SPEED.TYPE via:

```
Transform
    Compute
```

Which generate the Compute Variable dialogue box of Fig. 2.15. Operationalising, the new variable CROSSPRO is added to the working file as shown in Fig. 2.16. Part of the results of running the dummy regression with the cross product term are shown in Fig. 2.17.

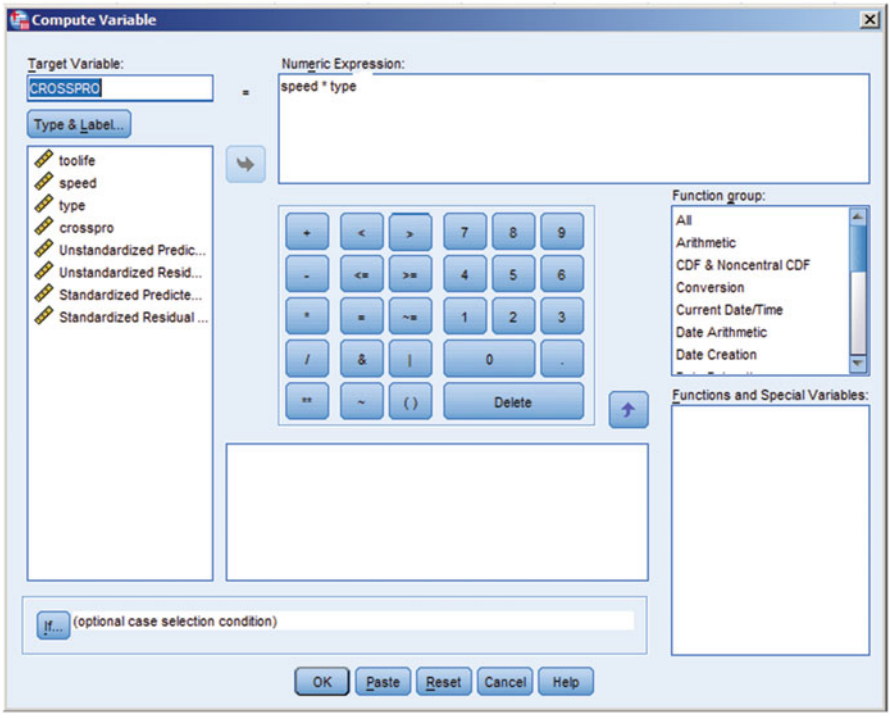
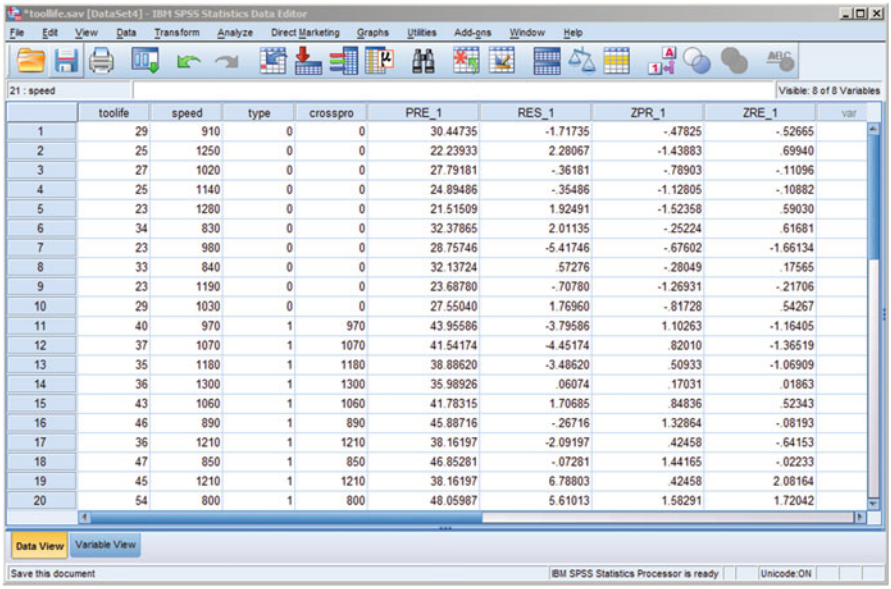


Fig. 2.15 Computation pf the cross-product term



	toolife	speed	type	crosspro	PRE_1	RES_1	ZPR_1	ZRE_1	var
1	29	910	0	0	30.44735	-1.71735	-.47825	-.52665	
2	25	1250	0	0	22.23933	2.28067	-1.43883	.69940	
3	27	1020	0	0	27.79181	-.36181	-.78903	-.11096	
4	25	1140	0	0	24.89486	-.35486	-1.12805	-.10882	
5	23	1280	0	0	21.51509	1.92491	-1.52358	.59030	
6	34	830	0	0	32.37865	2.01135	-.25224	.61681	
7	23	980	0	0	28.75746	-5.41746	-.67602	-1.66134	
8	33	840	0	0	32.13724	.57276	-.28049	.17565	
9	23	1190	0	0	23.68780	-.70780	-1.26931	-.21706	
10	29	1030	0	0	27.55040	1.76960	-.81728	.54267	
11	40	970	1	970	43.95586	-3.79586	1.10263	-1.16405	
12	37	1070	1	1070	41.54174	-4.45174	.82010	-1.36519	
13	35	1180	1	1180	38.88620	-3.48620	.50933	-1.06909	
14	36	1300	1	1300	35.98926	.06074	.17031	.01863	
15	43	1060	1	1060	41.78315	1.70685	.84836	.52343	
16	46	890	1	890	45.88716	-.26716	1.32864	-.08193	
17	36	1210	1	1210	38.16197	-2.09197	.42458	-.64153	
18	47	850	1	850	46.85281	-.07281	1.44165	-.02233	
19	45	1210	1	1210	38.16197	6.78803	.42458	2.08164	
20	54	800	1	800	48.05987	5.61013	1.58291	1.72042	

Fig. 2.16 The new data file

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.942[a]	.888	.867	3.314

a. Predictors: (Constant), crosspro, speed, type

Coefficients[a]

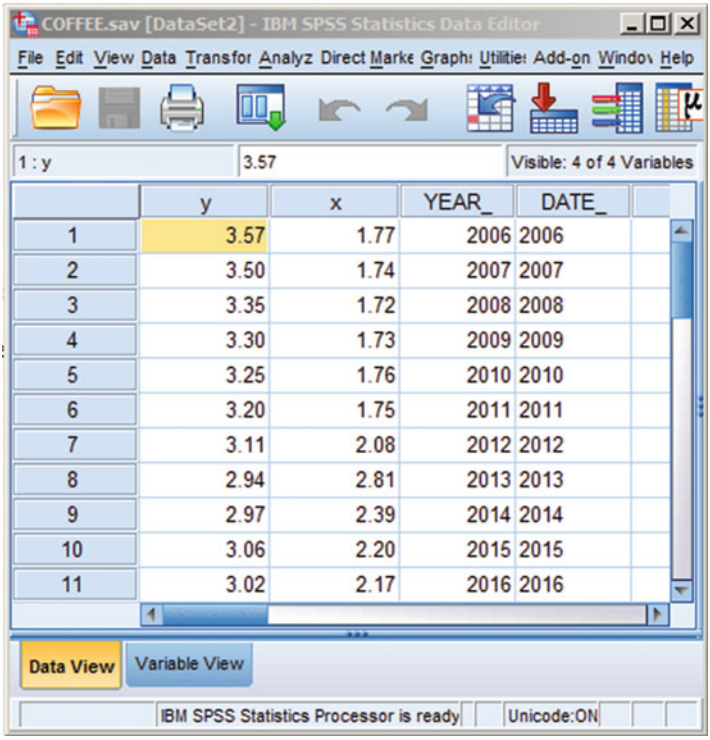
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std.Error	Beta			Lower Bound	Upper Bound
1	(Constant)	48.908	7.176		6.815	.000	33.695	64.121
	speed	-.021	.007	-.373	-3.066	.007	-.035	-.006
	type	21.642	9.927	1.222	2.180	.045	.598	42.687
	crosspro	-.006	.009	-.388	-.681	.506	-.026	.013

a. Dependent Variable: toolife

Fig. 2.17 Part of the output for dummy regression with a cross product term

2.4 Functional Forms of Regression Models

Several classes of model occur in finance, business and economics that are not linear in form. Such models can, however, be transformed into linear ones and ordinary least squares (OLS) applied. To understand these, it is necessary to remind ourselves of the following laws of logarithms that are valid regardless of the base used:



	y	x	YEAR_	DATE_
1	3.57	1.77	2006	2006
2	3.50	1.74	2007	2007
3	3.35	1.72	2008	2008
4	3.30	1.73	2009	2009
5	3.25	1.76	2010	2010
6	3.20	1.75	2011	2011
7	3.11	2.08	2012	2012
8	2.94	2.81	2013	2013
9	2.97	2.39	2014	2014
10	3.06	2.20	2015	2015
11	3.02	2.17	2016	2016

Fig. 2.18 Raw data

(a) $\log(XY) = \log X + \log Y$
(b) $\log(X/Y) = \log X - \log Y$
(c) $\log(X)^n = n.\log X$

In the above, X and Y are assumed to be positive and n is some constant.

Consider the data in Fig. 2.18 below which presents average annual coffee consumption (Y; cups per day) in the London area in relation to average annual retail price (X; £ per lb.).

The data are in the file COFFEE.SAV. These data were subjected to bivariate regression analysis and the results obtained are presented below in Fig. 2.19.

The two variable regression model is thus of the form $\widehat{Y} = 4.171 - 0.48\,x$ which indicates that if the average retail price increases by a pound, then the average consumption of coffee would reduce by nearly half a cup. The coefficient of determination indicated that about 66% of the variation in average daily coffee consumption is explained by a linear relationship with the retail price of coffee.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std.Error	Beta	t	Sig.
1	(Constant)	4.171	.233		17.935	.000
	x	-.480	.114	-.814	-4.206	.002

a. Dependent Variable: y

Fig. 2.19 Bivariate regression results

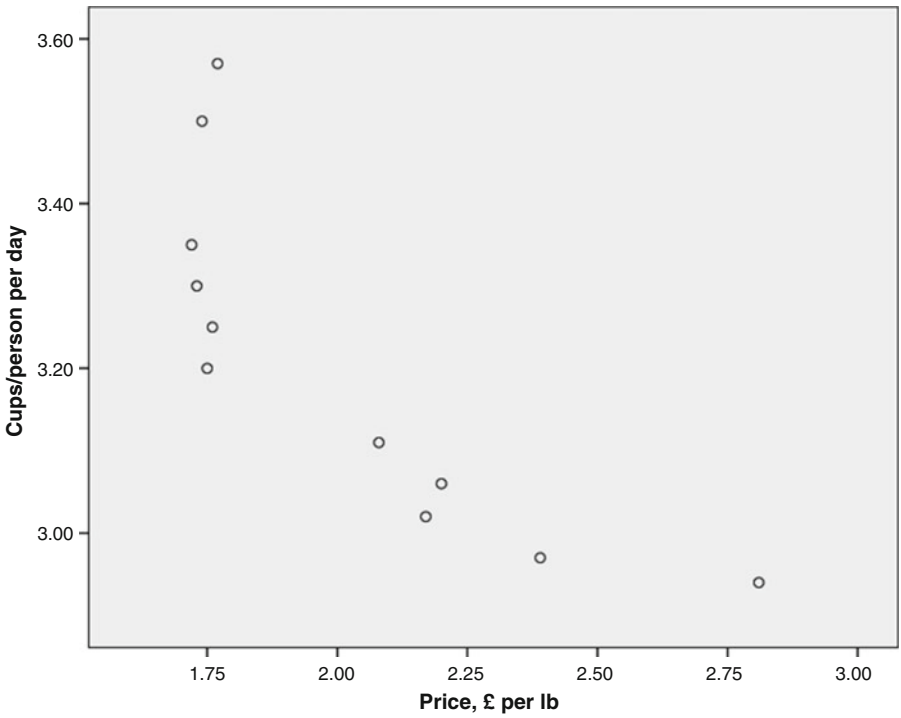


Fig. 2.20 A plot of average annual coffee consumption against average price

2.4.1 The Power Model

Figure 2.20 casts grave doubt as to whether the relationship between these two variables is a linear one. In fact, a more suitable model for the raw data in Fig. 2.18

$$\widehat{Y} = \beta_1 X^{\beta_2} \ldots\ldots \tag{2.5}$$

in which β_1 and β_2 are parameters to be estimated. Eq. (2.5) is referred to as **a power model**. Eq. (2.5) may be expressed in a linear form by the simple expedient of taking natural logarithms (i.e. base e) to derive:

$$\ln Y = \ln \beta_1 + \beta_2 \ln X \ldots\ldots \tag{2.6}$$

The fact that Eq. (2.6) is in linear form may be shown by letting $Y^* = \ln Y$, $\alpha = \ln \beta_1$ and $X^* = \ln X$ to establish that:

$$Y^* = \alpha + \beta_2 X^* \ldots\ldots \tag{2.7}$$

Ordinary least squares may now be used to estimate α and β_2 in Eq. (2.7). Knowing the value of α, we can readily find β_1 because $\alpha = \ln \beta_1 \Rightarrow \beta_1 = e^{\alpha}$. The linear Eq. (2.6) is known as **a log-log model**, a **double-log model** or perhaps most commonly, **a log-linear model**.

How can we assess if Eq. (2.5) is an adequate representation of the data in Fig. 2.18? If it is and from Eq. (2.6), a plot of ln$Y$ against ln$X$ should be linear or closely so. Figure 2.21 presents this plot, from which it is seen that the plot is not perfectly linear, but it seems to be more linear than the plot in Fig. 2.20.

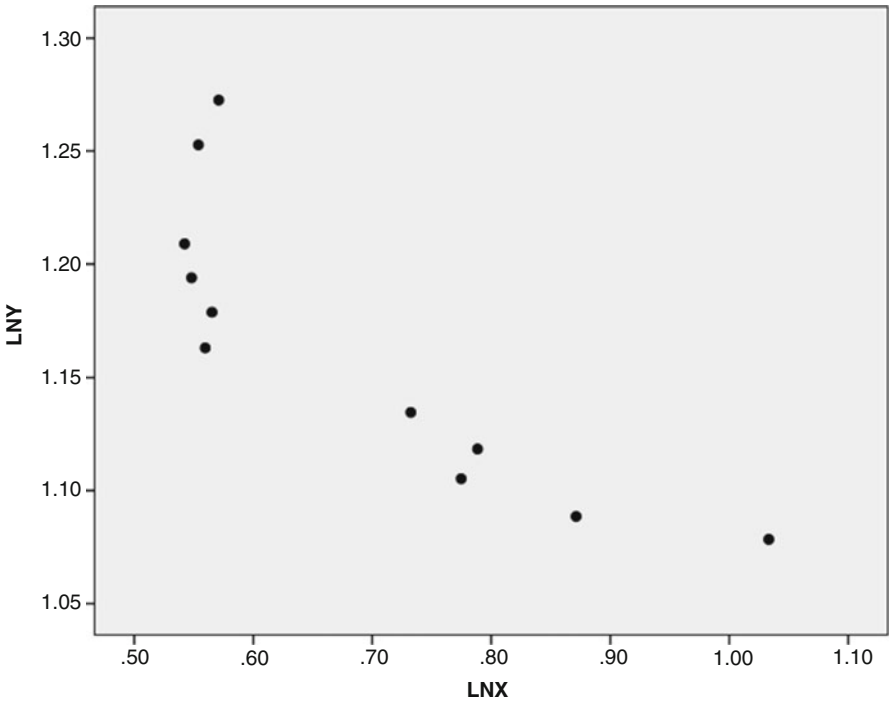


Fig. 2.21 A plot of lnY against lnX

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	1.390	.049		28.489	.000
	LNX	-.331	.069	-.847	-4.771	.001

a. Dependent Variable: LNY

Fig. 2.22 Results of regressing lnY against lnX

Regressing ln$Y$ against ln$X$ by means of OLS, the results in Fig. 2.22 are obtained:

The coefficient of determination of the model in Fig. 2.22 is 71.7%. *However, this is not directly comparable with the coefficient of determination obtained for the bivariate linear model in* Fig. 2.19*, since the models are different*. The power model of Eq. (2.5) may well be a more adequate representation of the raw data than is the purely linear model presented in Fig. 2.19.

A very attractive feature of the log-linear model is that the slope coefficient (or gradient) β_2 in Eq. (2.6) measures **the elasticity** of the variable Y with respect to variable X. The elasticity of Y with respect to X is defined as the percentage change in Y for a given (small) percentage change in X. If Y represents the quantity of a commodity demanded and X is its unit price, then β_2 measures what is called **the price elasticity of demand**:

$$\textbf{Price elasticity of demand} = \frac{\textbf{\%change in quantity demanded}}{\textbf{\%change in price}}$$

From Fig. 2.22, the results corresponding to Eq. (2.6) are:

$$\text{Ln}Y = 1.39 - 0.331\ \text{Ln}X \ldots\ldots \tag{2.8}$$

whereby the price elasticity coefficient is about −0.33. This implies that in the face of a 1% increase in the price of coffee, demand for coffee (as measured by cups of coffee consumed) decreases on average by 0.33%. When the price elasticity is less that 1 in absolute terms, we say that the demand for coffee is price inelastic; if the price elasticity exceeds 1, we say that the demand for coffee is price elastic.

The results in Eq. (2.8) may be used to compute the parameters of the power model of Eq. (2.5). Comparing Eq. (2.8) with Eq. (2.6), we find that $\ln\beta_1 = 1.39$ which implies that $\beta_1 = e^{1.39} = 4.014$ and that $\beta_2 = -0.331$. Inserting these values back into the original power model of Eq. (2.5), we establish that:

$$\widehat{Y} = 4.014X^{-0.331}$$

This numerical result may be obtained directly in IBM SPSS Statistics by using the 'Curve Estimation' procedure, which is part of that package's Regression routine.

The idea of taking a logarithmic transformation may be extended. Consider the compound interest formula and for simplicity let us suppose annual compounding for t years:

$$FV = PV(1 + r)^t \ldots\ldots \tag{2.9}$$

in which FV and PV are respectively future and present values of an investment and r is the annual interest rate. Taking logarithms:

$$\ln FV = \ln PV + t \ln (1 + r),$$

which is in linear form. This is called **a semilog model**, because only one variable (FV) appears in logarithmic form. Again, this model has an important property. Here, the slope coefficient (or gradient) represents **the rate of growth of Y**. If the gradient is negative, we have **a rate of decay**. Research has shown that the U.S GDP (in billions of dollars, constant prices) between 1998 and 2015 inclusive is well approximated by a semilog model. It has been established that:

$$\ln GDP = 9.164 + 0.039t$$

which suggests that GDP grew at a rate of 3.9% over this period. Note that in 1998, t = 0, so we estimate that $\ln GDP = 9.164 \Rightarrow \text{GDP} = e^{9.164} = 9547$ billion dollars.

2.4.2 The Reciprocal Model

Models of the following type are known as **reciprocal models:**

$$Y = \beta_1 + \beta_2\left(\frac{1}{X}\right) \ldots\ldots \tag{2.10}$$

in which β_1 and β_2 are parameters to be estimated. The model has the feature that as X increases indefinitely, the term $\beta_2\left(\frac{1}{X}\right)$ approaches zero**,** and Y therefore approaches the limit (or **asymptotic value**) β_1. Reciprocal models have built in them an asymptote or limit value that the dependent variable will take when the value of the X variable increases indefinitely.

The shape of the reciprocal model is shown in Fig. 2.23 in which the asymptote is represented by the horizontal line. In Fig. 2.23, β_1 and β_2 are both positive, non-zero. One application of the reciprocal model is the average fixed cost of production (Y) against levels of output (X). As X increases, Y decreases to a finite limit.

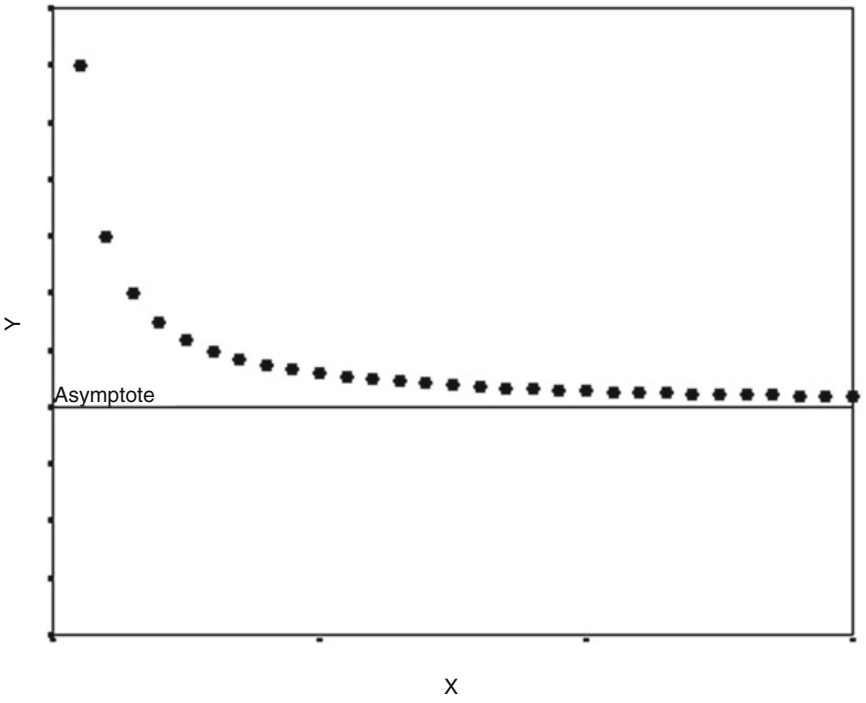


Fig. 2.23 The reciprocal model with asymptote

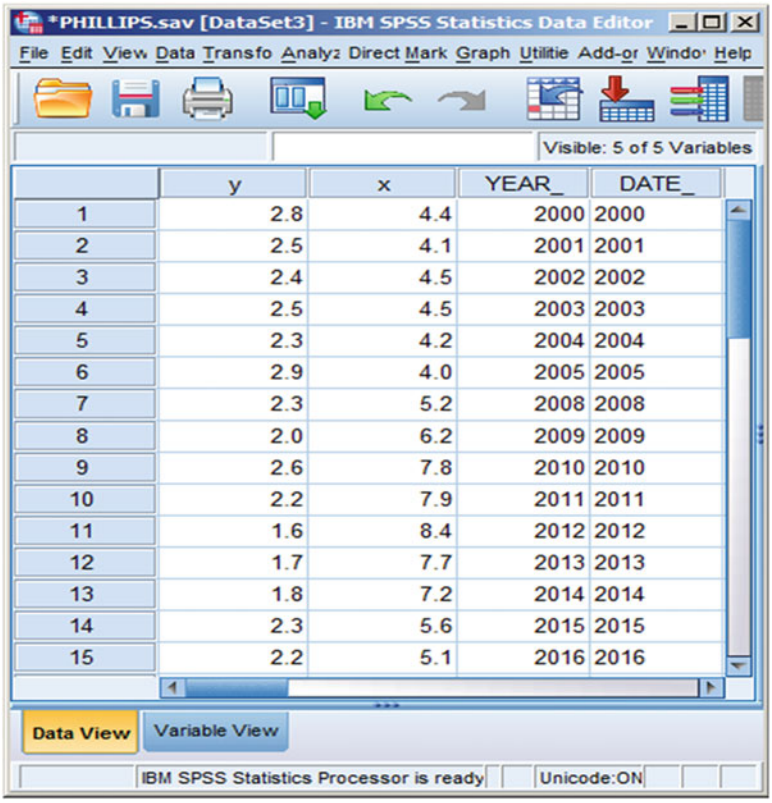
An important application of the reciprocal model is the **Phillips curve**. This is based on empirical observation by the economist A.W. Phillips of the relationship between the level of unemployment and the year to year increase or rate of change in wage rates. (Note that by inference, the rate of change in wages will impact on the rate of change in commodity prices or inflation). Figure 2.24 presents U.K. annual increases in wage rates (Y%) against unemployment (X%) from 2000 to 2016 inclusive. The data are available on the book webpage under the file PHILLIPS.SAV.

Upon regressing Y against 1/X, the results below are obtained (Fig. 2.25):

The reciprocal model obtained is thus:

$$\widehat{Y} = 1.232 + 5.63\ (1/\mathrm{X}) \ldots . \tag{2.11}$$

Inherent in Eq. (2.11) is that as the unemployment rate increases, the % increase in wages declines. Like any equation, Eq. (2.11) cuts the X-axis when Y = 0, i.e. 1.232/5.63 = 1/X = 0.2188 ⇒ X = 4.6%. This is the rate of unemployment consistent with no change in wage rates, which theoretically means stable prices. This point is called **the non-accelerating inflation rate of unemployment** (NAIRU) or **the natural rate of unemployment**. The fact that the β_1 is positive



	y	x	YEAR_	DATE_
1	2.8	4.4	2000	2000
2	2.5	4.1	2001	2001
3	2.4	4.5	2002	2002
4	2.5	4.5	2003	2003
5	2.3	4.2	2004	2004
6	2.9	4.0	2005	2005
7	2.3	5.2	2008	2008
8	2.0	6.2	2009	2009
9	2.6	7.8	2010	2010
10	2.2	7.9	2011	2011
11	1.6	8.4	2012	2012
12	1.7	7.7	2013	2013
13	1.8	7.2	2014	2014
14	2.3	5.6	2015	2015
15	2.2	5.1	2016	2016

Fig. 2.24 UK increases in wage rates and unemployment

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	1.232	.292		4.221	.001
	RECIP	5.630	1.531	.714	3.678	.003

a. Dependent Variable: % increase in wages

Fig. 2.25 Regression of increases in wage rates against the reciprocal of unemployment

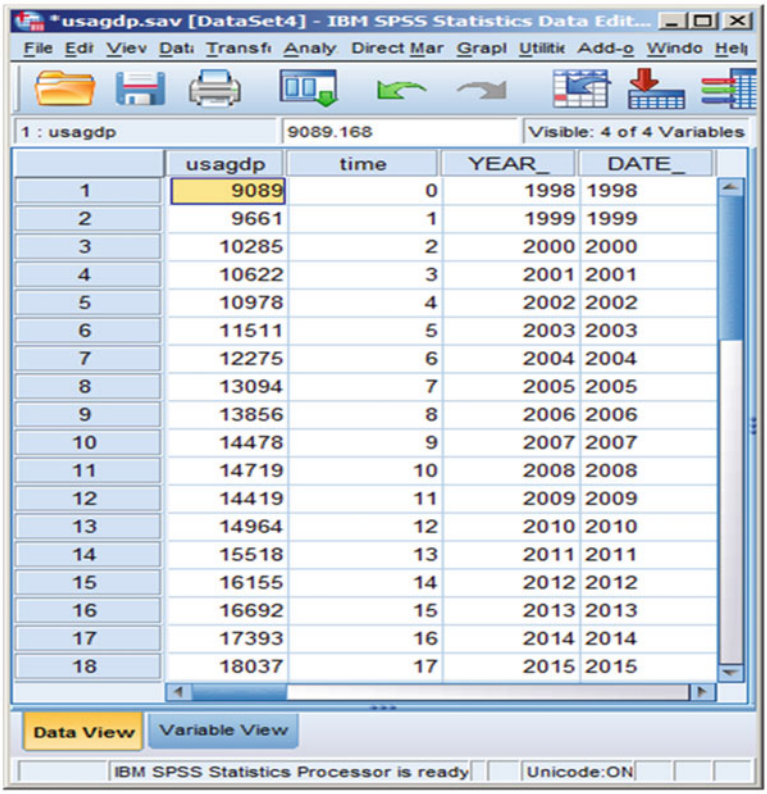
means that in theory the wage rate could not decease. As the % unemployed (X) increases indefinitely, the percentage decrease in wage rates approaches the asymptote or limit of 1.232% and will, therefore, not be worse than this figure per year. The figure of 1.232% is called the **wage floor.**

2.4.3 The Linear Trend Model

Instead of the semilog model where ln$Y$ is regressed against time t, some researchers have advocated the linear trend model wherein Y is regressed against time t:

$$Y = \beta_1 + \beta_2 t \ldots\ldots \tag{2.12}$$

By **trend**, we mean any sustained upward or downward movement in the behaviour of a variable. If the slope coefficient or gradient β_2 is positive, there is an upward trend in Y, whereas if β_2 is negative there is a downward trend in Y. The data in Fig. 2.26 present the GDP of the United States in current billions of dollars, between 1972 and 1991 inclusive. The source is *The Economic Report of the President*, January 1993 and the figures are in the file GDPUSA.SAV on the file server.



	usagdp	time	YEAR_	DATE_
1	9089	0	1998	1998
2	9661	1	1999	1999
3	10285	2	2000	2000
4	10622	3	2001	2001
5	10978	4	2002	2002
6	11511	5	2003	2003
7	12275	6	2004	2004
8	13094	7	2005	2005
9	13856	8	2006	2006
10	14478	9	2007	2007
11	14719	10	2008	2008
12	14419	11	2009	2009
13	14964	12	2010	2010
14	15518	13	2011	2011
15	16155	14	2012	2012
16	16692	15	2013	2013
17	17393	16	2014	2014
18	18037	17	2015	2015

Fig. 2.26 United States GDP, 1972–1991

(The U.S. GDP data of the semilog section were at constant prices). Figure 2.27 presents a plot of these data over time. Note that the time variable, t, in Eq.)2.12) is given values from 0 to 17 inclusive. Figure 2.27 suggests that the data exhibit a reasonable trend with a positive gradient. The results of regressing GDP against time were:

so the derived linear trend model is (Fig. 2.28):

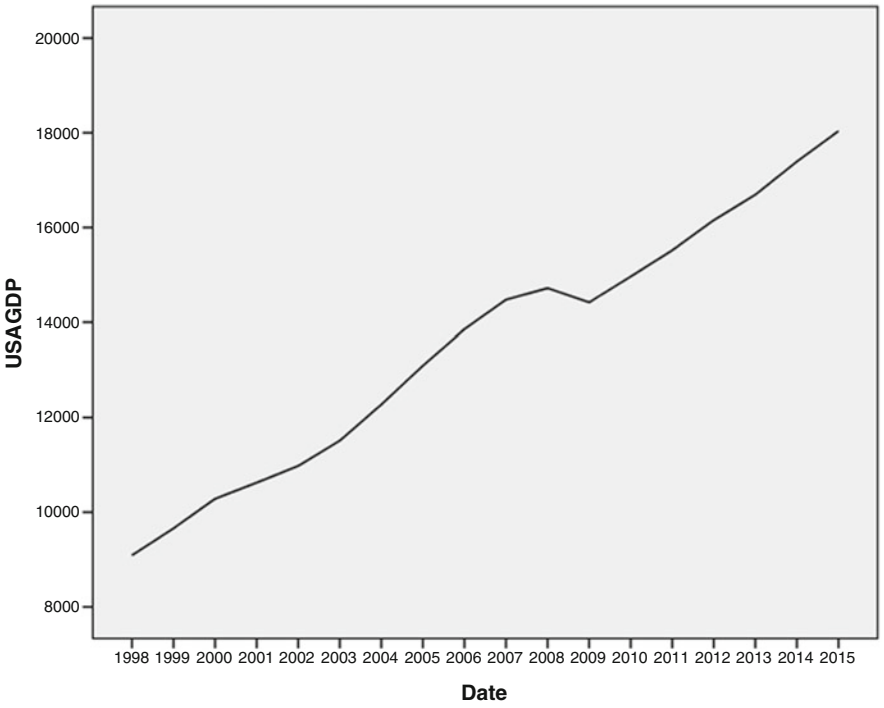


Fig. 2.27 A plot of U.S.A. GDP over time

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	9212.284	147.713		62.366	.000
	time	509.293	14.833	.993	34.335	.000

a. Dependent Variable: usagdp

Fig. 2.28 Regression results for GDP against t

$$\widehat{Y} = 9212.284 + 509.293t$$

Note that an important point is choice between the semilog and linear trend model depends on whether one is interested in the rate of growth (semilog) or absolute growth (linear trend model). Again recall that it is not possible to compare the coefficients of determination between such competing models.

Chapter 3
The Box-Jenkins Methodology

The Box-Jenkins approach to time series modelling consists of extracting predictable movements (or patterns) from the observed data through a series of iterations. The univariate Box-Jenkins method is purely a forecasting tool; no explanation is offered in that there are no regressor-type variables. The Box-Jenkins approach follows a three phase procedure:

- **Model identification:** a particular category of Box-Jenkins (B-J) model is identified by using various statistics computed from an analysis of the historical data.
- **Model estimation and verification:** once identified, the "best model" is estimated such that the fitted values come as close as possible to capturing the pattern exhibited by the actual data.
- **Forecasting:** the final model is used to forecast the time series and to develop confidence intervals that measure the uncertainty associated with the forecast.

3.1 The Property of Stationarity

Time series data (denoted by Y_t) consist of readings on a variable taken at equally intervals of time. How would one compute the mean of a time series of a specified length? Calculating the mean of a sequence of observations might appear to be a trivial problem, as we would just sum all readings and divide by their number. However, if the series is steadily increasing overtime, i.e. exhibits a trend and we make decisions based on this mean, we would certainly not, for example, want to use this parameter as a forecast of the future level of the series. We would also not use the overall mean to make inferences (e.g. as the centre of confidence intervals) at time periods at the beginning or end the series. If we regard our gathered series as but one example of all possible series that could be generated by the same mechanism, we are further faced with the problem of estimating the mean for

A. Aljandali, *Multivariate Methods and Forecasting with IBM® SPSS® Statistics*,
Statistics and Econometrics for Finance, DOI 10.1007/978-3-319-56481-4_3

each time period, as we have a sample only of one item. It is similarly impossible to estimate the variance at any one time period.

The observed value of a series at particular time should be viewed as a random value; that is if a new set of data could be obtained under similar conditions, we would not obtain the identical numerical value. Let us measure at equal intervals the thickness of wire made on a continuous extraction machine. Such a list of measurements can be interpreted as a *realization* of wire thickness. If we were repeatedly to stop the process, service the machine and to restart the process to obtain new wires under similar machine conditions, we would be able to obtain new realizations from the same *stochastic* process. These realizations could be used to calculate the mean thickness of the wire after 1 min, 2 min etc. The term *stochastic* simply means "random" and the term *process* should be interpreted as the mechanism generating data. The problem is that in most situations, we can obtain only one realization. We cannot, for example, stop the economy, go back to some arbitrary point and then restart the economic process to observe a new realization. With a single realization, we cannot estimate with any precision the mean at each time period t and it is impossible to estimate the variance and autocorrelations. Therefore, to estimate the mean, variance and autocorrelation parameters of a stochastic process based on a single realization, the time series analyst must impose restrictions on how the data can be gathered.

A series that measures the cumulative effect of something is called *integrated*. Most of the probability theory of time series is concerned with integrated series that are *stationary*. **Broadly speaking, a time series is said to be stationary if there is no systematic change in mean (no trend) over time, if there is no systematic change in variance and if period variations have been removed.**

3.1.1 Trend Differencing

The assumption of no trend returns us to the problem posed at the start of this section. If there is no trend in the series, we might be willing to assume that the mean is constant for each time period and that the observed value at each time period is representative of that mean. The second condition above refers to constant variance. The variance of a series expresses the degree of variation about the assumed constant mean and as such gives a measure of uncertainty around this mean. If the variance is not constant over time, but say increases, it would be incorrect to believe that we can express the uncertainty around a forecasted mean level with a variance based on all the data. Most business and economic time series are *non-stationary*. Time series analysis often requires one to turn a non-stationary series into a stationary one in order to apply various aspects of statistical theory.

The first stage in any time series analysis should be to plot the available observations against time. This is often a very valuable part of any data analysis, since qualitative features such as trend, seasonality and outliers will usually be visible if present in the data. Consider Fig. 3.1, which is a plot of a company's

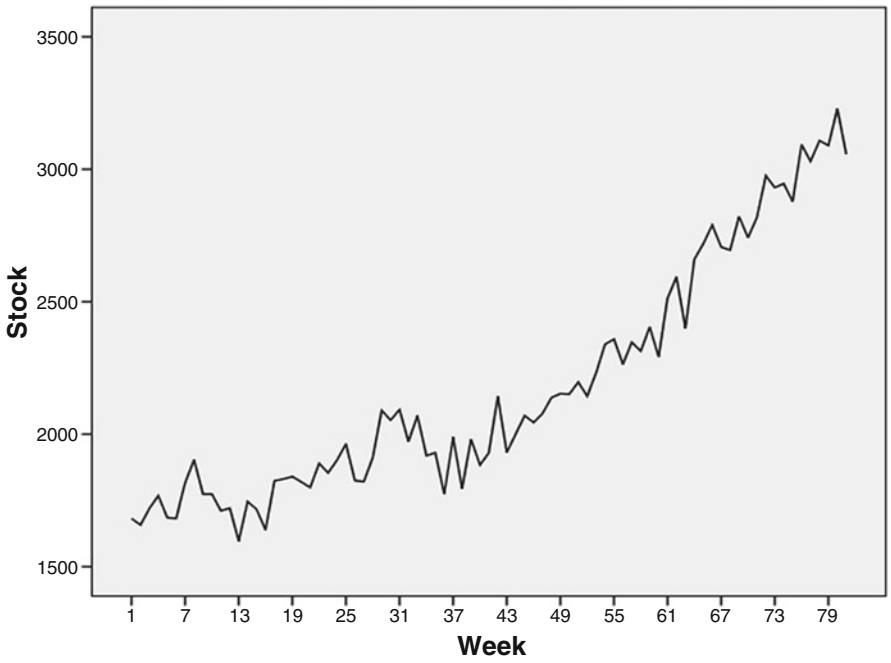


Fig. 3.1 Stock levels over time

inventory levels over 81 consecutive weeks (data file STOCK.SAV). A visual inspection clearly evidences that there is a trend in the data. The time series is not stationary. To achieve stationarity, the trend has to be eliminated.

Most economic time series are characterized by movements along a trend time such as in Fig. 3.1. Although there is a general understanding of what a trend is, it is difficult to give a more precise definition of the term trend than "any systematic change in the level of a time series". The difficulty in defining a trend stems from the fact that what looks like a change in the level in a short series of observations may turn out not to be a trend when a longer series becomes available, but rather be part of a cyclical movement.

Box and Jenkins advocated that an integrated time series can have the trend removed by the method of *differencing*. The method of differencing consists of subtracting the values of the observations from one another in some prescribed time-dependent order. For example, a *first order difference transformation* is defined as the difference between the values of two adjacent observations; second order differencing consists of taking differences of the differenced series; and so on.

Consider the series 1, 3, 5, 7, 9 and 11 which exhibits a constant increase (trend) of two units from one observation to the next. We now take the *first order differences*:

$$3-1=2$$
$$5-3=2$$
$$7-5=2$$
$$9-7=2$$
$$11-9=2$$

By taking the first order differences of a series with a linear trend, the trend disappears. Let us apply the method to a series with a non-linear trend: 1, 6, 15, 28, 45, 66 and 91. *The first order differences* are 5, 9, 13, 17, 21 and 25. This differenced series possesses a linear trend with a constant increase of 4. Therefore, by taking the differences of the differences (i.e. *second order differences*), we would obtain a trend-free series. Second order differences, in fact, remove a quadratic trend; third order differencing removes a cubic trend. It is rare for economic time series to involve more than second order differencing. Note that every time that we difference a series, we lose an observation. Due to random fluctuations in the data, such neat results as above cannot always be obtained. However and as said, for many economic time series, first or second order differencing will be sufficient to remove the trend component (called a *detrended* series), so that further analysis can proceed. Note that once the trend has been removed, further differencing will continue to produce a series without a trend. However, each additional differencing results is one additional data point being lost. Therefore, such *overdifferencing* will needlessly complicate the model and should be avoided.

3.1.2 Seasonal Differencing

A lot of economic time series evidence seasonal patterns that make the time series non-stationary. Many monthly or quarterly series will exhibit effects which have a high degree of regularity. The adjustment procedure now to be employed is called *seasonal differencing*, in contrast with consecutive differencing discussed in the last subsection. This involves taking differences among the detrended observations spaced at four-period intervals i.e. if a quarterly pattern is evident, compute the differences between the first quarter value of each successive year and similarly the differences between the second, third and fourth quarters of successive years. Season differencing of order 1 indicates that we are taking the first differences among the same quarters in different years. The seasonal adjustment just described is said to involve a *span* of 4 periods. A span of 4 implies that a lag of 4 periods is used in the seasonal differencing operation.

3.1.3 Homoscedasticity of the Data

The process of differencing attempts to produce stationarity when there is a trend. When the variance of a time series is thought not to be a constant over time, there are several data transformations available. Two of the transformations commonly used are the logarithmic and the square root transform. The logarithmic is particularly effective when (i) the variance of the series is proportional to the mean level of the series or (ii) the mean level of the series increases or decreases at a constant percentage. In that logs and roots of negative values are unreal, **such transforms must precede any differencing that may be required**.

3.1.4 Producing a Stationary Time Series in IBM SPSS Statistics

The problem of non-constant variance in Fig. 3.1 does not appear to be a problem here. Therefore, no transformation will be used on the Y_t. If a log or root transform of Y_t is deemed necessary, then the new variable may be computed via:

```
Transform
    Compute Variable ...
```

Some order of differencing is needed for the inventory data of Fig. 3.1. To generate a new time series containing differenced data, from the Data Editor select:

```
Transform
    Create Time Series ...
```

Which produces the *Create Time Series dialogue* box of Fig. 3.2. The default is differencing of order 1. Entering the variable STOCK into the 'New variable' box generates a default name of STOCK_1 for the first differences, as shown in Fig. 3.2. This default may be changed. In the box labelled 'Name' type in a new name (say FIRSTDIF) and click the Change button which has now become black. Our dialogue box now looks like Fig. 3.3. Click the OK button to operationalize and the new variable FIRSTDIF is added to the active file as shown in Fig. 3.4. Differencing causes the first reading for FIRSTDIF to be missing. The rest of the first order differences are $657 - 682 = -25$, $720 - 657 = 63$ etc. Figure 3.5 is a plot if the first differences of the variable STOCK over time and the trend does appear to have been removed and we regard FIRSTDIF to be stationary. (There is a formal statistical test available called the Dicky-Fuller test to examine if a series is stationary. This is not available in IBM SPSS Statistics, so it is omitted here).

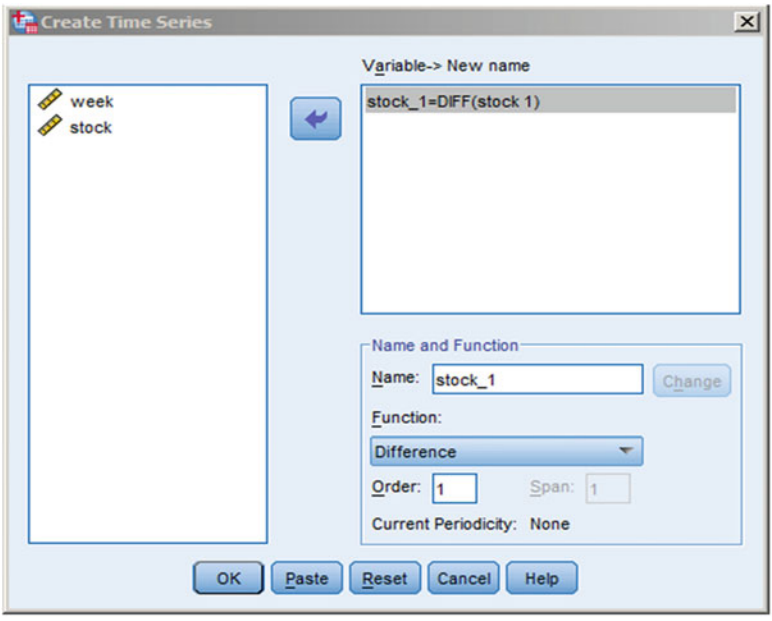


Fig. 3.2 The create time series dialogue box

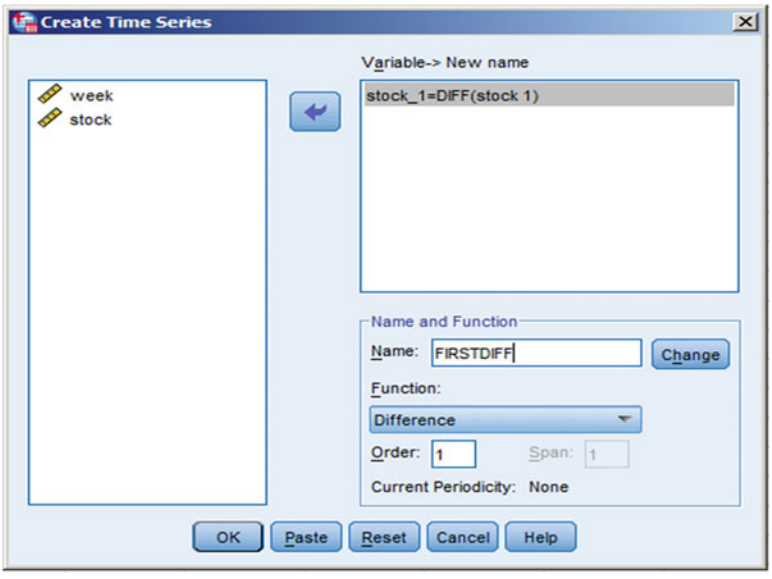


Fig. 3.3 The default variable name change

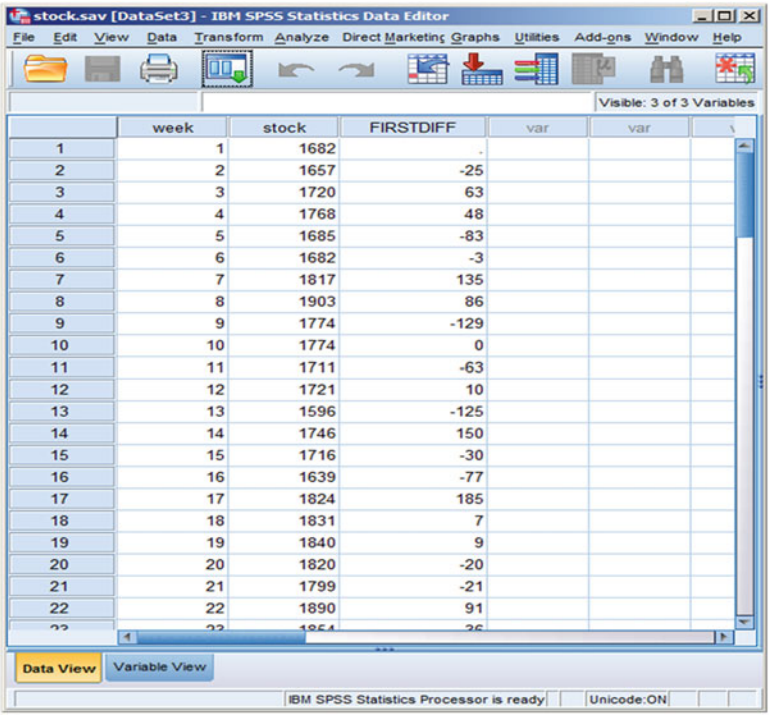


Fig. 3.4 The variable FIRSTDIF added to the active file

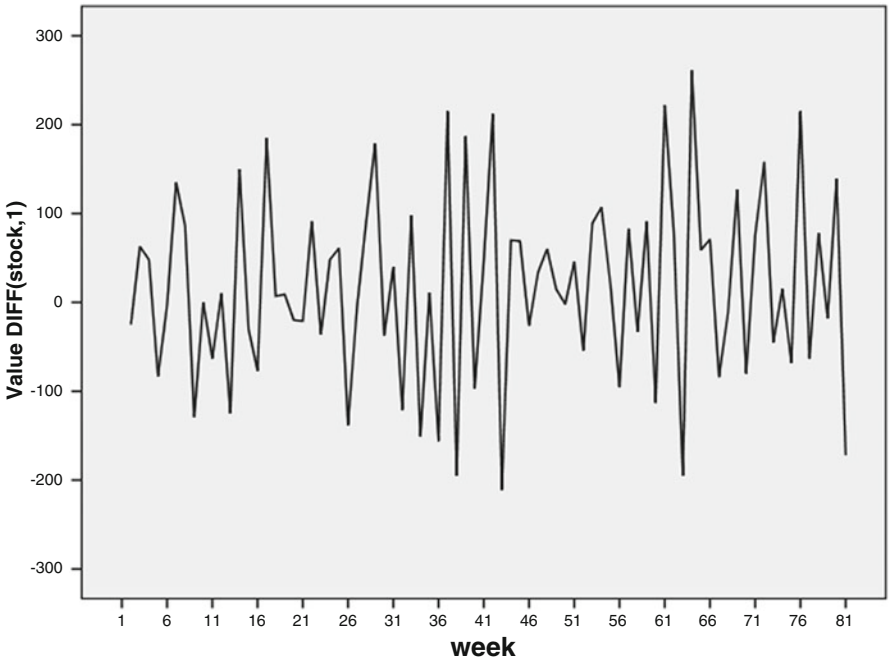


Fig. 3.5 A plot of first differences of the variable STOCK

3.2 The ARIMA Model

B-J models are known as Auto Regressive Integrated Moving Average (ARIMA). The methods used to solve the parameters of ARIMA models require quite a lot of computation, so for practical use, software is needed. The methods used in identifying, estimating and diagnosing ARIMA models are quite evolved.

The ARIMA procedure is carried out on stationary data. The notation Z_t is used for the stationary data at time t, whereas Y_t is the non-stationary datum value at that time. The ARIMA process considers linear models of the form:

$$Z_t = \mu + \varnothing_1 Z_{t-1} + \varnothing_2 Z_{t-2} + \ldots - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \ldots + e_t$$

where Z_t, Z_{t-1} are stationary data points; e_t,e_{t-1} are present and past forecast errors and $\mu, \varnothing_1, \varnothing_2 \ldots, \theta_1, \theta_2 \ldots$are parameters of the model to be estimated.

If a successful model involved only $\varnothing_1$i.e. was of the form:

$$Z_t = \mu + \varnothing_1 Z_{t-1} + e_t$$

The series is said to be governed by a first order autoregressive process, written AR(1). $\varnothing_1$ is called *the autoregressive parameter* and the model above, describes the effect of a unit change in Z_{t-1} on Z_t. Similarly the model:

$$Z_t = \mu + \varnothing_1 Z_{t-1} + \varnothing_2 Z_{t-2} + \ldots + \varnothing_p Z_{t-p} + e_t$$

Is called a p-order autoregressive process, written as AR(p). The sum of the coefficients $\varnothing_i$, i = 1, 2, …, p of an autoregressive process must always be less than unity.

If a successful model only involved θ_1 i.e. was of the form:

$$Z_t = \mu - \theta_1 e_{t-1} + e_t$$

Then the time series is said to be governed by a first order moving average process, written as MA(1). θ_1 is called *the moving average parameter*. Similarly, the model:

$$Z_t = \mu - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \ldots - \quad \theta_q e_{t-q} + e_t$$

is called a q-order moving average model written as MA(q).

Models involving both autoregressive and moving average processes are called *mixed models*. If a mixed model contained an autoregressive process of order 1 and a moving average process of order 2, the n the model is written as ARIMA(1,2) and would be of the form:

$$Z_t = \mu + \varnothing_1 Z_{t-1} - \theta_1 e_{t-1} - \theta_2 e_{t-2} + e_t$$

When differencing has been used to generate stationarity, the model is said to be *integrated* and is written as ARIMA (p,d,q). The middle parameter d is simply the number of times that the series had to be differenced before stationarity was achieved. If the (stationary) Z_t in the above equation had to be differenced twice before stationarity was achieved, then that model would be written as ARIMA (1,2,2).

3.3 Autocorrelation

To identify the model that best describes the time series under consideration, two sets of statistics are used: autocorrelations (AC) and partial autocorrelations (PAC). Both measure how much interdependence there is among the observations and take values that range between ± 1, depending on the pattern of the relationship. If, for example, values of the time series that are above the mean value of the series are immediately followed by values that are below the mean, then both the AC and PAC will be negative. This is said to be *negative autocorrelation*.

3.3.1 ACF

AC's provide us with a numerical measure of the relationship of specific values of a time series to other values in the time series. That is, they measure the relationship of a variable to itself over time. AC's are normally computed for different time lags. For example, given n readings Z_1, Z_2, ..., Z_n, we can form $n - 1$ pairs of observations (Z_1, Z_2), ((Z_2, Z_3) ... ((Z_{n-1}, Z_n). Regarding the first observation in each pair as one variable and the second observation as the second variable, we can compute the Pearsonian correlation coefficient at, in the example of the data in Fig. 3.1, a lag of 1 week. This measures correlation between successive readings and is called the *first order autocorrelation coefficient*. Similarly, we could compute the correlation between observations at a distance k apart, which is called the *kth. order autocorrelation coefficient*.

For example, consider the data:

$$51, 52, 54, 60, 55, 61, 62, 66, 60, 62, 66 \ldots$$

The first order autocorrelation coefficient is calculated using the standard formula for the Pearsonian coefficient, involving the pairs:

$$(51, 52)\ (52, 54)\ (54, 60)\ (60, 55) \ldots$$

The second order autocorrelation coefficient would be computed using the pairs:

$$(51, 54)\ (52, 60)\ (54, 55)\ (60, 61)\ldots$$

We use the notation that r_k is the (auto) correlation between Z_t and Z_{t-k} so r_4 is the (auto) correlation between Z_t and Z_{t-4}. When the AC's are computed for lag 1, lag 2, lag 3 and so and are graphed (r_k against k), the result is called the *sample autocorrelation function* (*ACF*) or *correlogram*. This graph is useful for determining whether a series is stationary and for identifying a tentative ARIMA model. (By default, IBM SPSS Statistics produces and ACF up to lag of 16).

If a series is non-stationary by virtue of having an upwards trend, then readings a few lags apart will be autocorrelated. If the data are stationary, however, the autocorrelations should all be zero (indicative of random error). This should be a characteristic of the ACF for stationary data. To test whether or not the autocorrelation coefficient is statistically equal to zero, we use, for large samples the t statistic – and meaningful economic time series should involve large samples. When the number of readings is reasonably large and to test the hypothesis that the population autocorrelation coefficient (ρ_k) at lag k is zero, i.e.:

$$H_0 : \rho_k = 0 \text{ against} H_1 : \rho_k \neq 0,$$

We adopt *Bartlett's method*. Bartlett derived that if the above null hypothesis is true, then the sample autocorrelations at lag k, r_k, will be closely normally distributed with zero mean and a variance of:

$$\text{Var}\,(r_k) = n^{-1}\{\ 1 + 2\left(r_1^2 + r_2^2 + r_3^2 + \ldots + r_{k-1}^2\right)\}.$$

The test statistic is:

$$\frac{r_k}{SD\ of\ r_k}$$

Which is distributed as the test statistic with n − 2 degrees of freedom. Given that the t distribution is asymptotically normal, the boundaries of the critical region for the above test are usually taken at ±1.96 (±2).

For example, suppose for a given set of eight readings, that the autocorrelations at lags 1 and 2 were respectively $r_1 = -0.412$ and $r_2 = -0.343$ and that we wished to test if the second order autocorrelation coefficient was significantly different from zero, i.e.:

$$H_0 : \rho_2 = 0 \text{ against} H_1 : \rho_2 \neq 0$$

We may compute that:

$$\text{Var}\,(r_2) = 8^{-1}\left\{1 + 2(-0.412)^2\right\} = 0.1674 \text{ therefore SD of} r_2 = 0.4092$$

The test statistic under H_0 becomes:

$$(-0.343 - 0)/0.4092 = -0.838$$

Which is well distant from the critical region boundary for t ($v = 6$). We, therefore, fail to reject H_0 and conclude that ρ_k is zero. It may also be shown that the first order autocorrelation coefficient is also not significant.

The curved lines of Fig. 3.8 (see page 73) represent 95% confidence limits for the autocorrelation coefficients, r_k, based on Bartlett's variance formula given by the Var(r_k) equation. These serve as indications as to with autocorrelations are statistically significant. The first seven such coefficients in Fig. 3.8 exhibit this characteristic – once more indicative that the non-differenced data are not stationary.

It should be noted that if the researcher is sure that the time series data are stationary, then the (r_k) in Bartlett's variance formula are in (theory) zero. This leads to Quenouille's formula for the variance of the r_k in the instance of stationary data, that:

$$\text{Var}(r_k) = n^{-1}$$

Most computer packages have both Bartlett's and Quenouille's variance formulae as available options.

It can be shown that the autocorrelation for an AR(1) model will in theory be:

$$r_k = \varnothing_1^k$$

Whenever the series is stationary, it may be shown that the sum of the AR coefficients:

$$\varnothing_1 + \varnothing_2 + \varnothing_3 \ldots$$

Will be less than one. In the case of an AR(1) model, this implies that $\varnothing_1$ will be less than one, so the AC's will be decreasing in absolute value as the lag increases, i.e. $\varnothing_1 > \varnothing_2^2 > \varnothing_3^2 > \ldots > \varnothing_k^2$, which simply says that the relationship weakens as we go back over time. Further the autocorrelations decline fairly rapidly.

It can be shown that the autocorrelation coefficients for a moving average process of order 1, MA (1), in theory are:

$$r_k = \frac{-\theta_1}{1 + \theta_1^2} \, for \; k = 1$$

$$r_k = 0 \, for \; k = 2$$

3.3.2 PACF

Partial autocorrelation coefficients PAC's are closely related to the AC's. They also take on values between −1 and 1. A diagram of PAC's against the lag k is called the *partial autocorrelation function (PACF)*. A partial autocorrelation is the measure of the relationship between two variables when the effect of other variables has been removed or held constant. With temporal data, r_{kk} is the partial autocorrelation between Z_t and Z_{t-k} when the effect of the intervening variables Z_{t-1}, Z_{t-2},. . ., Z_{t-k+1} has been removed. This adjustment is to see if the correlation between Z_t and Z_{t-k} is due to the intervening variables or if indeed there is something else causing the relationship. As is discussed in the next section, the behavior of the PAC's along with the AC's for a stationary time series is used to identify a tentative ARIMA model.

The formula for the partial autocorrelation coefficient is quite complex, but numerical values are computed by available statistical packages. It was shown by Quenouille that:

$$\text{Var}(r_{kk}) = n^{-1},$$

So it is possible to examine the hypothesis, such as:

$$H_0 : \rho_{kk} = 0 \text{ versus} H_1 : \rho_{kk} \neq 0$$

For example, suppose that $r_{33} = -0.0318$ based on eight readings. This is the correlation between Z_t and Z_{t-3} when the effects of Z_{t-1} and Z_{t-2} have been removed. The test statistic is again distributed as the t statistic:

$$\frac{r_{kk}}{SD\ of\ r_{kk}}$$

So $-0.318/0.354 = -0.898$ which is not statistically significant. We, therefore, fail to reject the hypothesis that ρ_{kk} is zero. Again, for large n the boundaries of the critical region are usually taken at ± 1.96 (± 2).

If the data are stationary, then the partial autocorrelations should, in theory, be zero. The partial autocorrelation coefficients for the (non-differenced) inventory data of Fig. 3.1 are plotted on the partial autocorrelation function of Fig. 3.9. The two horizontal lines again represent the 95% confidence interval. Although not statistically significant, the PAC's fail to die out, indicating that the data are not stationary.

3.3.3 Patterns of the ACF and PACF

It is possible to use the ACF and PACF to recognise patterns that characterise moving average (MA), autocorrelation (AR) and mixed (ARMA) models, when the assumption of stationarity has been satisfied. It should be appreciated that we are focusing on theoretical models, but this does facilitate recognition of similar patterns in actual time series data. By comparing actual ACF's and PACF's to the theoretical patterns, we shall be able to identify the specific type of B-J model that will adequately represent the data.

There are general guidelines:

- If the autocorrelations decay and the partial autocorrelations have spikes, the process can be captured by an AR model, where the order equals the number of significant spikes. The ACF should show exponentially declining values.
- If the partial autocorrelations decay and the autocorrelations have spikes, the process is best captured by an MA model, where the order equals the number of significant spikes. The PACF should show exponentially declining values.
- If both the autocorrelation and partial autocorrelations are characterized by irregular patterns on the ACF and PACF, the process is best captured by an ARMA model, where the order equals the number of significant spikes. It may be necessary to invoke several cycles of the identification-estimation-diagnosis process.

AC patterns for moving average models are among the easiest to recognise in practice.

3.3.4 Applying an ARIMA Model

In this section, we apply an ARIMA model to the inventory data that were graphed in Fig. 3.1. Previous analysis (Fig. 3.5) has suggested that first order differencing induced stationarity. The parameter d in the ARIMA (p,d,q) model thus take a value of unity. To estimate appropriate values for the remaining parameters p and q, we plot the ACF and PACF for the first order differenced data. These plots are presented respectively in Figs. 3.8 and 3.9. The ACF shows that the only significant coefficient is associated with a one period lag. (The value of this autocorrelation coefficient is shown to be − 0.484 later in this section). The PACF has a significant spike at a lag of 1.

We now need to refer to the characteristics of theoretical models cited earlier. The fact that arguments could be made for any of the three types of Box-Jenkins models indicates the difficulty and indeed sometimes subjective nature of the fitting process. Here, arguments could be made for ARIMA (1, 1,0), ARIMA (0,1,1) and ARIMA (1,1,1) models, so all will be fitted. The ARIMA (1,1,0) is the same as AR (1) with first order differencing; the ARIMA (0,1,1) is the same as MA (1) with first

order differencing. Obviously, there needs to be some method for evaluating competing models and this will be the subject matter of the next section.

Reconsider the inventory data in the file STOCK.SAV. First order differencing is thought to have produced stationarity, so the parameter d = 1 in the ARIMA (p,d,q) models. The ACF and PACF need to be generated to suggest suitable values for the parameters p and q. In the IBM SPSS Data Editor, click:

```
Analyze
    Forecasting
       Autocorrelations
```

To produce the Autocorrelations dialogue box of Fig. 3.6.

Remember that is the stationary variable FIRSTDIF for which we want ACF and PACF plots so this variable is entered into the 'Variables' box. Both Autocorrelations and partial autocorrelations are the default. Click the Options button to produce the *Autocorrelations: Options dialogue box* of Fig. 3.7 and select Bartlett's approximation so that we may assess the significance of any correlation coefficients generated. The default is that coefficients are generated up to 16 lags, which is sufficient for most practical purposes (Fig. 3.8).

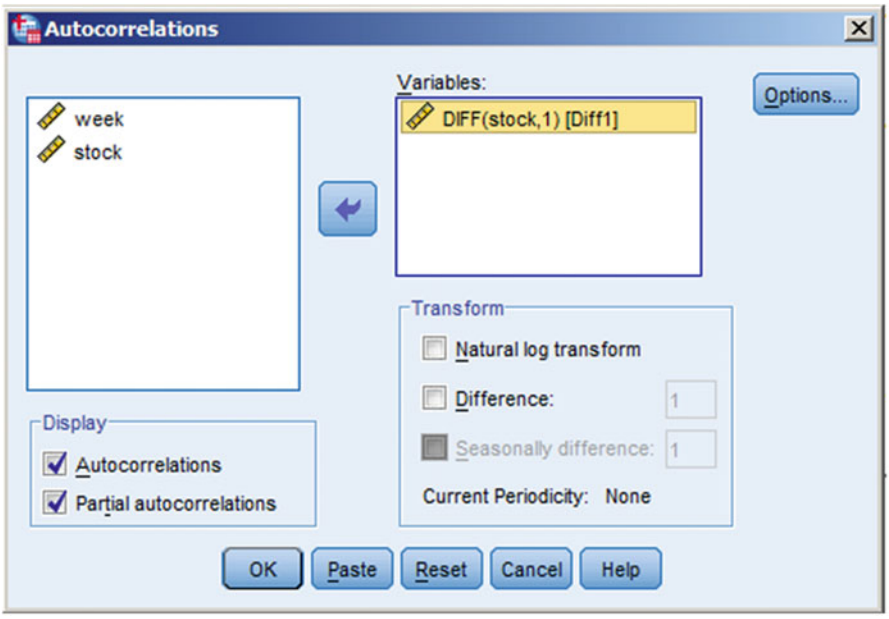


Fig. 3.6 The autocorrelations dialogue box

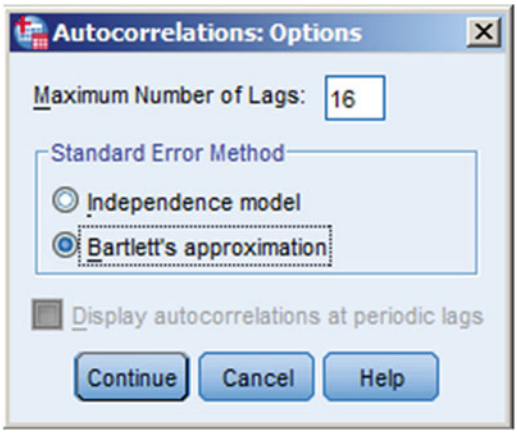


Fig. 3.7 The autocorrelations: options dialogue box

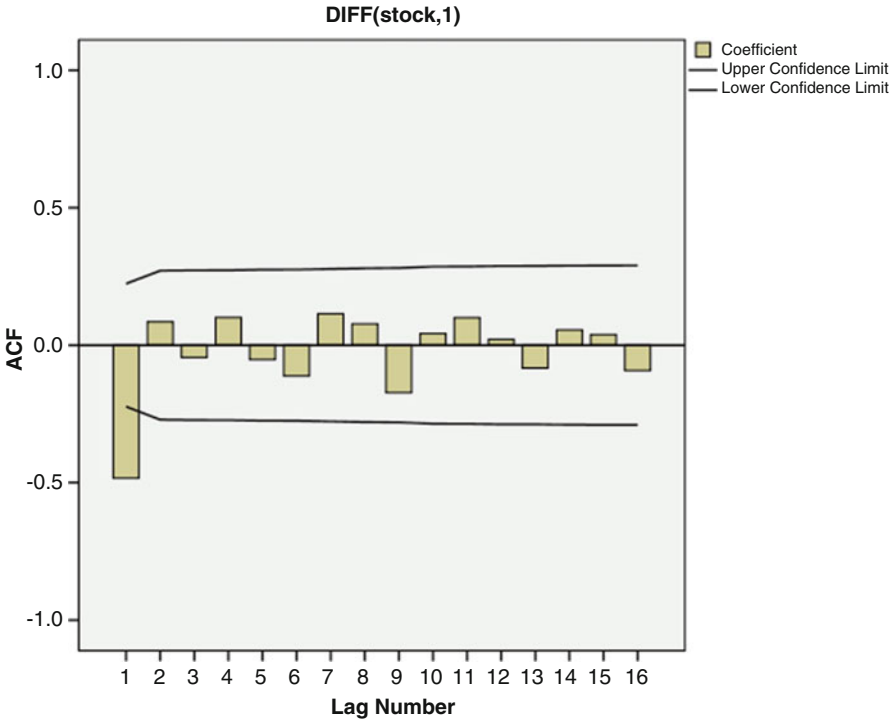


Fig. 3.8 The ACF plot

The ACF shows that the only significant coefficient is associated with a one period lag (The value of the autocorrelation coefficient is − 0.484 at this lag, as shown in the IBM SPSS output that is generated prior to the last two figures). The

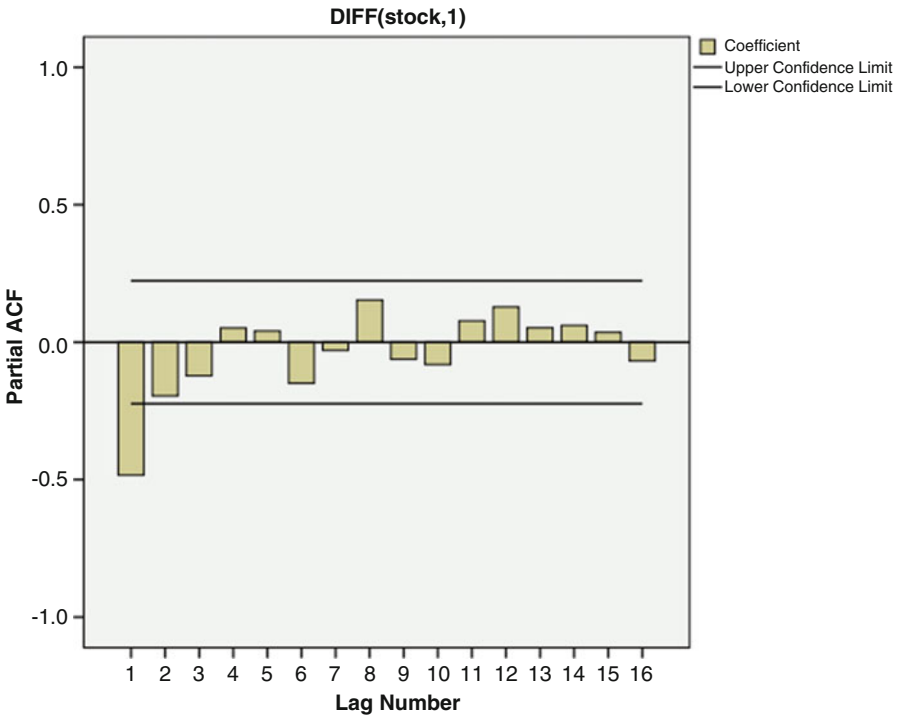


Fig. 3.9 The PACF plot

PACF has a significant spike at lag 1. We now need to refer to the characteristics of theoretical ARIMA model and compare them with our observed ACF and PCF plots. Arguments could be made for ARIMA (1,1,0), ARIMA (0,1,1) and ARIMA (1,1,1). Many text books contain theoretical ACF and PACF plots for a variety of ARIMA (p,d,q) models. This is a typical situation when several competing models are feasible. All potential models should be fitted and the next section suggests criteria by which a "best" model may be selected (Fig. 3.9).

3.4 ARIMA Models in IBM SPSS Statistics

This section illustrates just the fitting of an ARIMA (1,1,1) model. To fit ARIMA models, click:

```
Analyse
    Forecasting
       Create Models ...
```

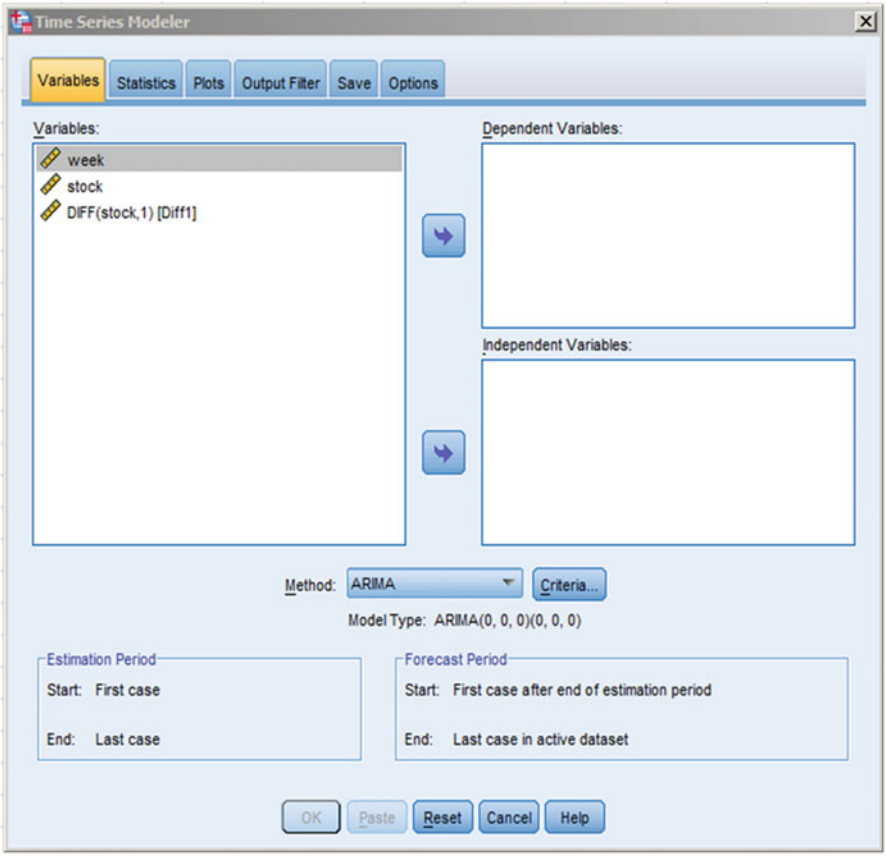


Fig. 3.10 The ARIMA dialogue box

Which produces the ARIMA dialogue box of Fig. 3.10. STOCK is the variable that we are examining and we set the parameters p = 1, d = 1 and q = 1 under criteria as shown in Fig. 3.11. Alternatively and to generate the same results, we could select FIRSTDIF as the variable to be considered and set p = 1, d = 0 and q = 1, since first differencing has already been accomplished. Click the Save button to produce the *ARIMA: Save dialogue box* of Fig. 3.12. Here, we select the 'Add to file' option to save such as the predicted values in our working file. I have also chosen to use this.

ARIMA model to forecast the stock level for week 82 as shown in Fig. 3.12. This forecast will also be added to the working file.

Amongst the IBM SPSS output is the *Akaike Information Criterion (AIC)* which takes into account how well the derived model fits the observed time series. The "best" model under this criterion is the one with the lowest AIC value. Also produced is the *Schwartz Bayesian Criterion (SBC)* which performs a similar task

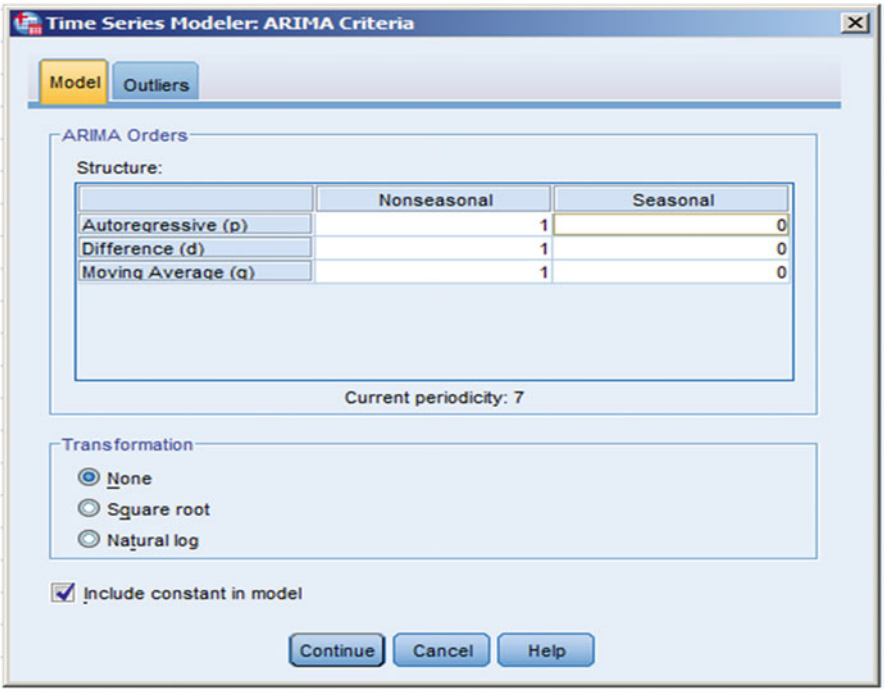


Fig. 3.11 The ARIMA criteria dialogue box

to the AIC and the model with the minimum SBC is sought. The AIC is generally for autoregressive models and the SBC is the more general criterion. In the IBM SPSS output, Standard Error refers to the standard error (or standard deviation) of the residuals. Again, the smallest value of this measure is sought. The square of this figure is the variance of the residuals (called *residual variance* in IBM SPSS). The ARIMA (1,1,1) has a standard error of residuals of 92.0712. In the IBM SPSS output and under the heading 'Variables in the Equation', we find that our ARIMA (1,1,1) model is:

$$Z_t = 18.2664 - 0.2106\, Z_{t-1} - 0.4023\, e_{t-1}.$$

However, study of the coefficients shows that $\varnothing_1 = -0.2106$ (referred to as AR1 in the output) is not significantly different from zero ($t = -1.045$, $p = 0.299$). Also, the moving average term $\theta_1 = -0.4023$ (referred to as MA in the output) is not significantly different from zero ($t = 2.148$, $p = 0.035$). Therefore, one could refer to the ARIMA (0,1,1) which has lower standard error, AIC and SBC than does the ARIMA (1,1,0) model.

Considering the ARIMA (0,1,1) model, it is found that:

$$Z_t = 18.2016 - 0.5455\, e_{t-1} \tag{3.1}$$

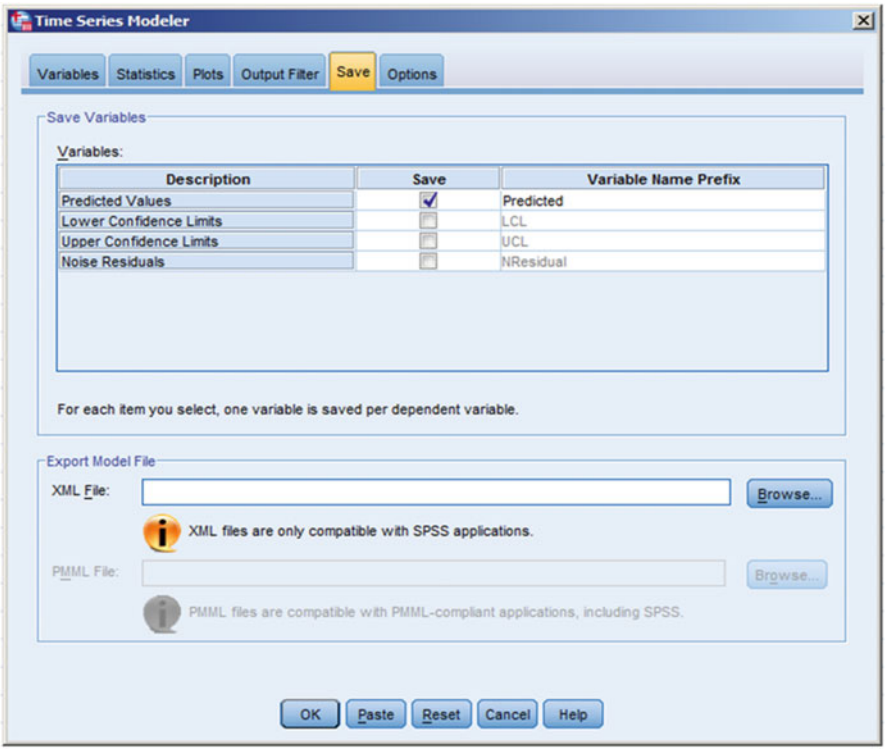


Fig. 3.12 The ARIMA save dialogue box

And the MA1 coefficient is statistically significant from zero (t = 5.720, p = 0.000). As one fits a variety of ARIMA models, the predicted values are given the IBM SPSS variables names FIT_1, FIT_2 etc. Figure 3.13 presents a plot of the observed and predicted stock levels for the ARIMA (1,1,1) model. Note that the model fails to capture turning points on time, which is a common characteristic of B-J models in general. It should also be noted that in practice, competing B-J models may have very little to choose between them in terms of the standard error, AIC and SBC. Identifying the appropriate order of a mixed model, for example, can be difficult.

IBM SPSS Statistics was used to fit these competing models. It will be noted that when fitting the ARIMA (0,1,1) model, forecasted values (FIT_2 – fitted values fir model 2), residuals or errors (ERR_2 – errors for model 2), along with lower and upper 95% confidence levels (LCL_2 and UCL_2) were computed and saved. The original inventory levels are for 81 weeks. The ARIMA models were asked to generate forecasts of the stock levels for weeks 82 and 83.

Equation 3.1 is used to generate these forecasts. Remembering that first order differencing was invoked, $Z_t = Y_t - Y_{t-1}$, so to forecast the stock level at week 82:

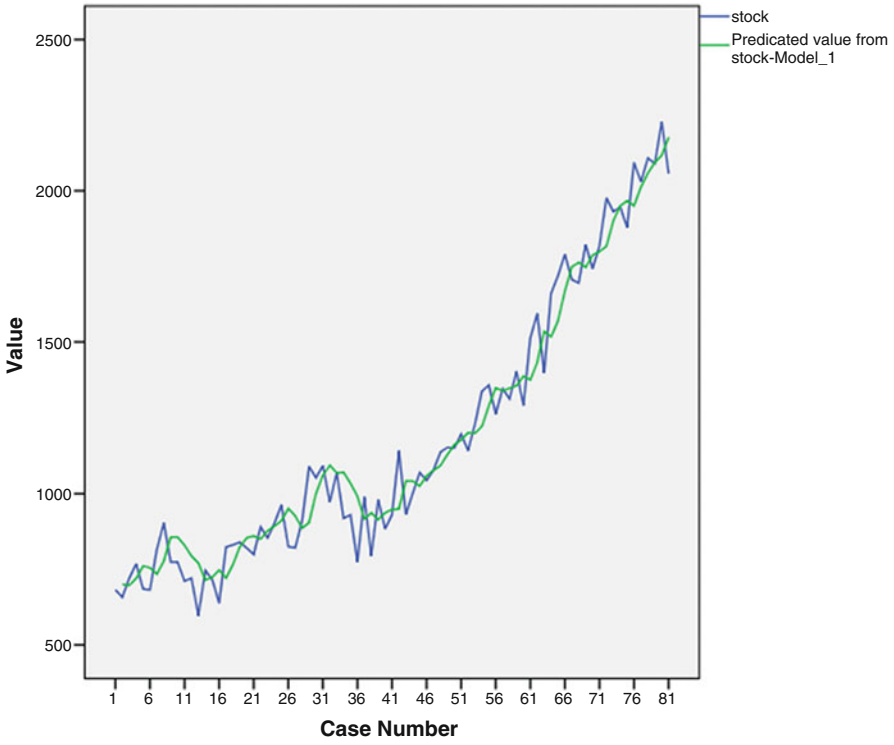


Fig. 3.13 Observed and predicted stock levels

$$Y_{82} - Y_{81} = 18.2016 - 0.5455\ \varepsilon_{81}$$

The stock level at week 81 (Y_{81}) was 2057 units and the error in forecast generated by this ARIMA model at this week (ε_{81}) was − 125.1968 (i.e. this ARIMA model overestimated the stock level at week 81), whereby:

$$Y_{82} = 2057 + 18.2016 - (0.5455)\ (-125.1968) = 2143.496$$

Similarly,

$$Y_{83} - Y_{82} = 18.2016 - 0.5455\ \varepsilon_{82}$$

We have no error value corresponding to week 82 (ε_{82}), so this is treated as zero and:

$$Y_{83} = 2057 + 18.2016 - 0 = 2161.698$$

Figure 3.13 is a plot of the actual and forecasted stock levels from the ARIMA (0,1,1) model.

The Box-Jenkins method is just one procedure available for forecasting. There can be no doubt that ARIMA models can be constructed to fit a wide variety of patterns and this can be done with minimum effort as long as a computer is available. Like all other time series models, ARIMA models suffer limitations. They generally fail to capture turning points on time and they provide the decision maker with little explanation. For example, they do not provide information on the potential impact of policies such as pricing actions or advertising programmes. However, multivariate Box-Jenkins models partially overcome these problems. As in the present example, competing ARIMA models may have little to choose between them. Identifying the appropriate order of a mixed model, for example can be difficult.

In order to clarify the choice between different univariate time series models, there have been several 'competitions' to compare the forecasting accuracy of different methods. However, the results of these competitions have not always been consistent. Given the different analysts and data sets used this is perhaps not surprising. The Box-Jenkins approach has not been consistently the best. Regression methods do rather better on average than univariate models, but again, this is not consistently the case.

A final point is that, although there is an advantage in being able to choose from the broad class of ARIMA models, there are also dangers in that considerable experience is need to interpret the ACF and PACF and other indicators. Moreover, when variation in a series is dominated by trend and seasonality, the effectiveness of the fitted ARIMA model is mainly determined by the differencing procedure employed rather than by the identification of the autocorrelation and/or moving average structures of the differenced series. In some situations, a large expenditure of time and effort can be justified and then the Box-Jenkins approach is worth considering. However, for routine sales forecasting, simple methods are more likely to be understood by managers and workers who have to utilize or implement the results.

Whilst noting that the Box-Jenkins approach has been one of the most influential developments in time series analysis, Box-Jenkins models are only worth considering when the following conditions are satisfied:

– The analyst is competent to implement it
– The objectives justify the complexity and
– The variation in the series is not dominated by trend and seasonality

Chapter 4
Exponential Smoothing and Naïve Models

4.1 Exponential Smoothing Models

Exponential smoothing models are amongst the most widely used time series models in the fields of economics, finance and general business analysis. The essence of these models is that new forecasts are derived by adjusting the previous forecast to reflect its forecast error. In this way, the forecaster can continually revise the forecast based on previous experiences. Exponential smoothing models have the advantage of requiring retention of only a limited amount of data. They can also be created with simple spreadsheet programmes. However, such is their importance that they are part of general statistical software, such as IBM SPSS Statistics.

The simplest model is the *single parameter exponential smoothing model*. Here, the forecast for the next and all subsequent periods is determined by adjusting the current-period forecast by a proportion of the difference between the forecast and the actual value. If recent forecasts have proven accurate, it seems reasonable to base subsequent forecasts on these estimates. Conversely, if recent predictions have been subject to large errors, new forecasts should take this into consideration.

In symbols, the single parameter model may be written as:

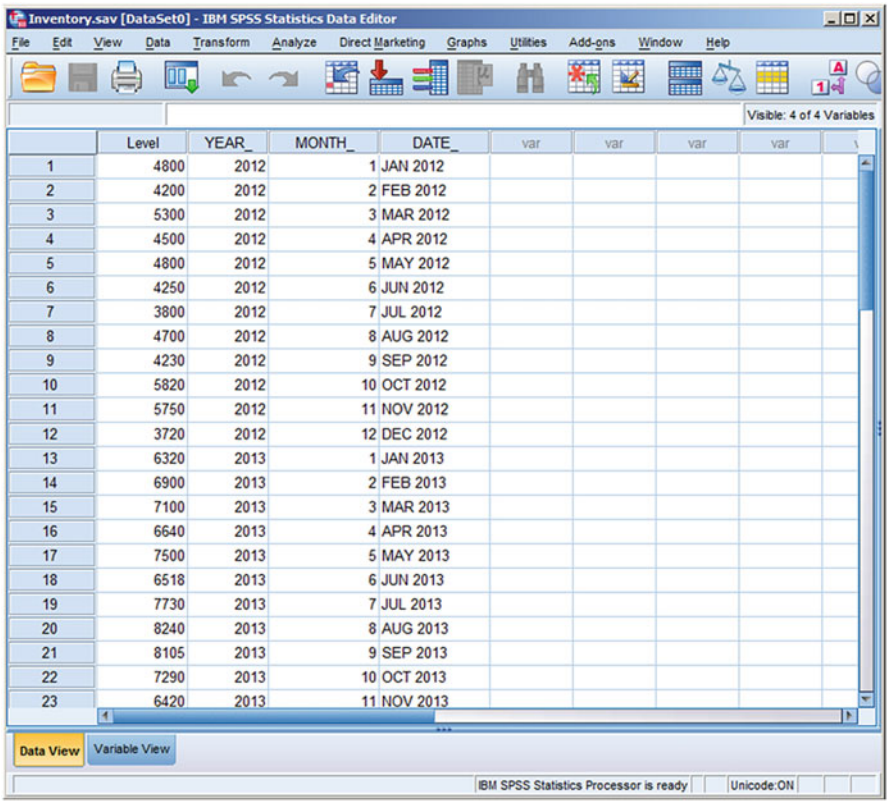
$$\widehat{Y}_{t+1} = \widehat{Y}_t + \alpha\left(Y_t - \widehat{Y}_t\right) \ldots\ldots \tag{4.1}$$

where $\widehat{Y}_t$ is the forecasted value of a variable at time t, Y_t is the observed value of that variable at time t and α is the smoothing parameter that has to be estimated and $0 \leq \alpha \leq 1$.

Consider the data in Fig. 4.1 (INVENTORY.SAV) which represent a company's monthly stock levels for a particular product (variable name LEVEL) from January 2012 to December 2014 inclusive. The data are graphed in Fig. 4.2.

The single parameter exponential smoothing model of Eq. (4.1) is fitted by clicking:

A. Aljandali, *Multivariate Methods and Forecasting with IBM® SPSS® Statistics*,
Statistics and Econometrics for Finance, DOI 10.1007/978-3-319-56481-4_4



	Level	YEAR_	MONTH_	DATE_	var	var	var	var
1	4800	2012	1	JAN 2012				
2	4200	2012	2	FEB 2012				
3	5300	2012	3	MAR 2012				
4	4500	2012	4	APR 2012				
5	4800	2012	5	MAY 2012				
6	4250	2012	6	JUN 2012				
7	3800	2012	7	JUL 2012				
8	4700	2012	8	AUG 2012				
9	4230	2012	9	SEP 2012				
10	5820	2012	10	OCT 2012				
11	5750	2012	11	NOV 2012				
12	3720	2012	12	DEC 2012				
13	6320	2013	1	JAN 2013				
14	6900	2013	2	FEB 2013				
15	7100	2013	3	MAR 2013				
16	6640	2013	4	APR 2013				
17	7500	2013	5	MAY 2013				
18	6518	2013	6	JUN 2013				
19	7730	2013	7	JUL 2013				
20	8240	2013	8	AUG 2013				
21	8105	2013	9	SEP 2013				
22	7290	2013	10	OCT 2013				
23	6420	2013	11	NOV 2013				

Fig. 4.1 A company's monthly stock levels over time

```
Analyze
   Forecasting
      Create Time Series
              Method: Exponential Smoothing
```

which generates the Exponential Smoothing dialogue box of Fig. 4.3. Clicking the Criteria button will open the dialogue box in Fig. 4.4, under which various model types are listed and which are summarised below:

- **Simple.** This model is appropriate for series in which there is no trend or seasonality. Its only smoothing parameter is level. Simple exponential smoothing is most similar to an ARIMA model with zero orders of autoregression, one order of differencing, one order of moving average, and no constant.
- **Holt's linear trend.** This model is appropriate for series in which there is a linear trend and no seasonality. Its smoothing parameters are level and trend,

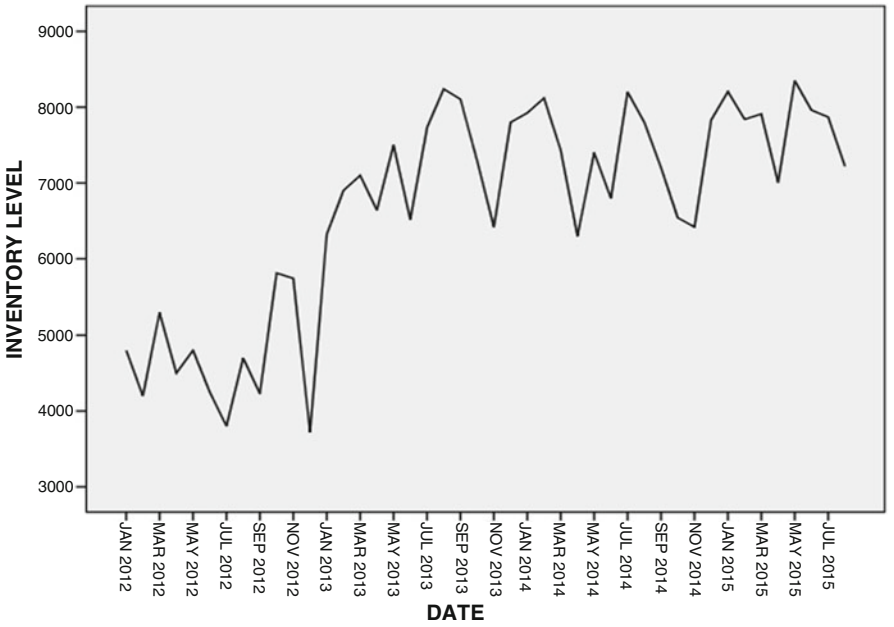


Fig. 4.2 A plot of stock levels over time

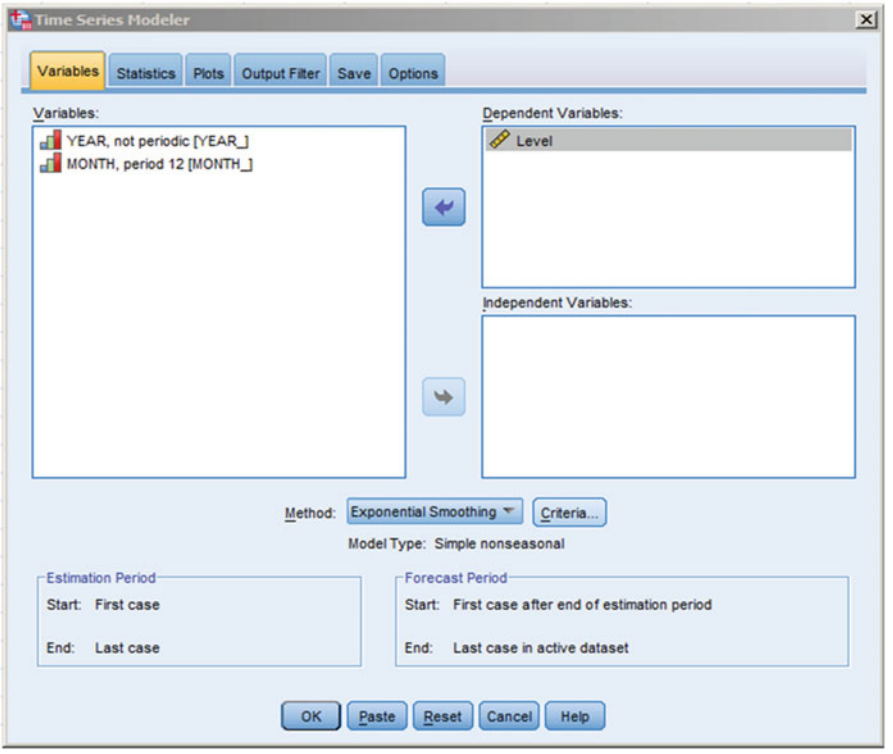


Fig. 4.3 The exponential smoothing dialogue box

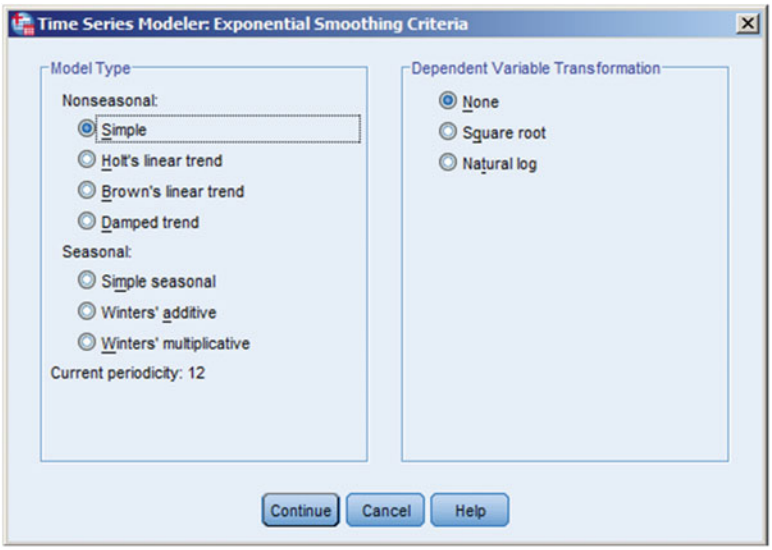


Fig. 4.4 The exponential smoothing: parameters dialogue box

which are not constrained by each other's values. Holt's model is more general than Brown's model but may take longer to compute for large series. Holt's exponential smoothing is most similar to an ARIMA model with zero orders of autoregression, two orders of differencing, and two orders of moving average.

- **Brown's linear trend.** This model is appropriate for series in which there is a linear trend and no seasonality. Its smoothing parameters are level and trend, which are assumed to be equal. Brown's model is therefore a special case of Holt's model. Brown's exponential smoothing is most similar to an ARIMA model with zero orders of autoregression, two orders of differencing, and two orders of moving average, with the coefficient for the second order of moving average equal to the square of one-half of the coefficient for the first order.
- **Damped trend.** This model is appropriate for series with a linear trend that is dying out and with no seasonality. Its smoothing parameters are level, trend, and damping trend. Damped exponential smoothing is most similar to an ARIMA model with 1 order of autoregression, 1 order of differencing, and 2 orders of moving average.
- **Simple seasonal.** This model is appropriate for series with no trend and a seasonal effect that is constant over time. Its smoothing parameters are level and season. Simple seasonal exponential smoothing is most similar to an ARIMA model with zero orders of autoregression, one order of differencing, one order of seasonal differencing, and orders 1, p, and $p + 1$ of moving average, where p is the number of periods in a seasonal interval (for monthly data, $p = 12$).

- **Winters' additive.** This model is appropriate for series with a linear trend and a seasonal effect that does not depend on the level of the series. Its smoothing parameters are level, trend, and season. Winters' additive exponential smoothing is most similar to an ARIMA model with zero orders of autoregression, one order of differencing, one order of seasonal differencing, and $p + 1$ orders of moving average, where p is the number of periods in a seasonal interval (for monthly data, $p = 12$).
- **Winters' multiplicative.** This model is appropriate for series with a linear trend and a seasonal effect that depends on the level of the series. Its smoothing parameters are level, trend, and season. Winters' multiplicative exponential smoothing is not similar to any ARIMA model.

The variable name LEVEL is entered in the appropriate box and the Simple model is chosen (i.e. single parameter). Clicking the Save button in Fig. 4.5 generates the Exponential Smoothing: Save dialogue box.

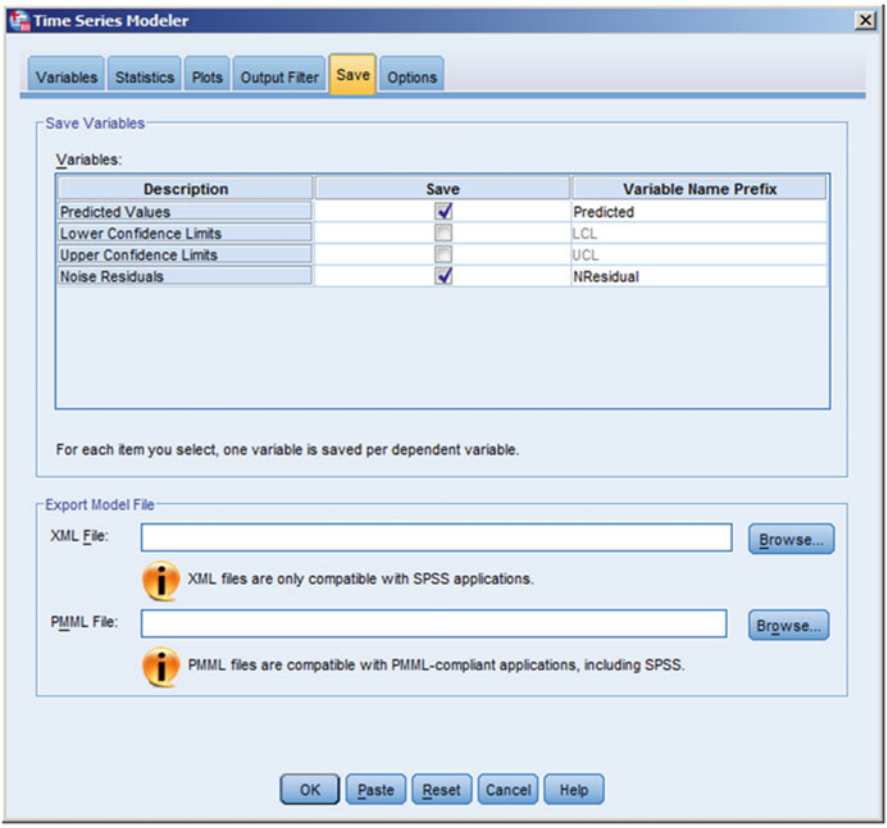


Fig. 4.5 The exponential smoothing: save dialogue box

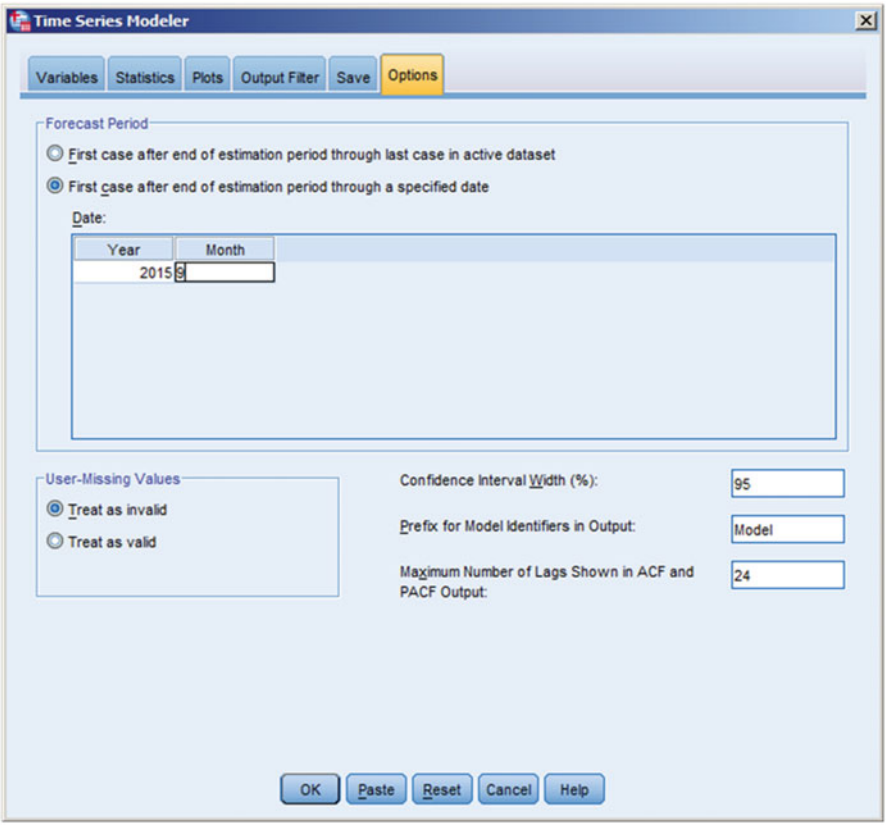
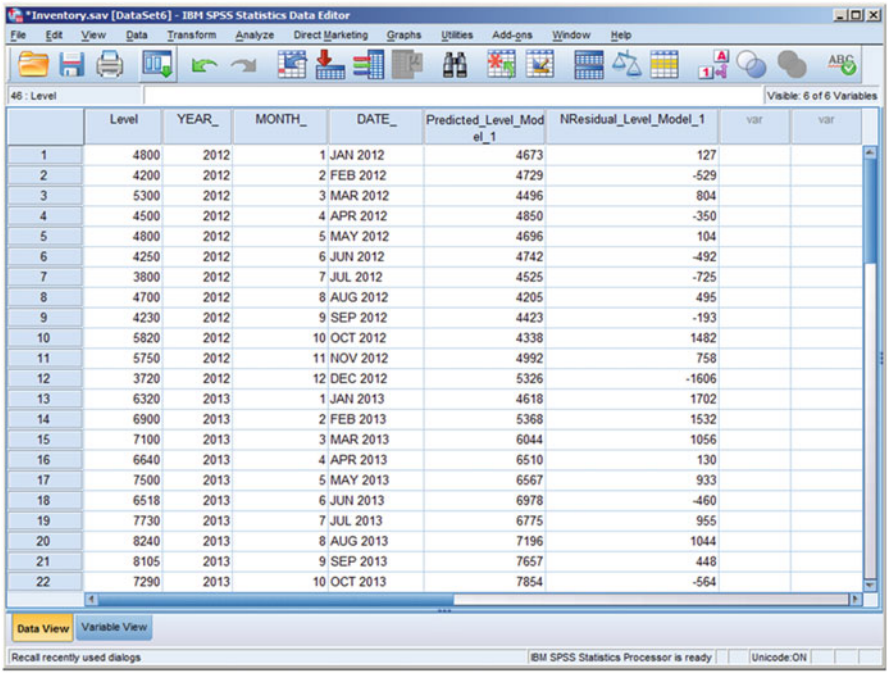


Fig. 4.6 The exponential smoothing: options dialogue box

The dialogue box of Fig. 4.5 permits the user to add the forecasted values from the optimal single parameter smoothing model (variable name Predicted) and the associated residuals or errors (variable name NResidual) to the data file. It also permits forecasting and in Fig. 4.6, a forecast for September 2005 has been requested. The latter is also added to the active data file. Upon running the procedure, the results of Fig. 4.7 are produced. The predicted value for September 2015 is 7593 units of stock. The single parameter model can be very accurate, but only for short term forecasting. Also, it does not fit the data well if there is a trend or seasonality present. However, there are other parameters that take these factors into account. A parameter γ is used if a trend is present and a parameter δ is used if seasonality is present. (The latter requires a minimum of four seasons of data). Figure 4.8 plots the observed and forecasted stock level data and it is evident that although our model is optimal in terms of the single parameter, it probably require their parameters to be incorporated.



	Level	YEAR_	MONTH_	DATE_	Predicted_Level_Model_1	NResidual_Level_Model_1	var	var
1	4800	2012	1	JAN 2012	4673	127		
2	4200	2012	2	FEB 2012	4729	-529		
3	5300	2012	3	MAR 2012	4496	804		
4	4500	2012	4	APR 2012	4850	-350		
5	4800	2012	5	MAY 2012	4696	104		
6	4250	2012	6	JUN 2012	4742	-492		
7	3800	2012	7	JUL 2012	4525	-725		
8	4700	2012	8	AUG 2012	4205	495		
9	4230	2012	9	SEP 2012	4423	-193		
10	5820	2012	10	OCT 2012	4338	1482		
11	5750	2012	11	NOV 2012	4992	758		
12	3720	2012	12	DEC 2012	5326	-1606		
13	6320	2013	1	JAN 2013	4618	1702		
14	6900	2013	2	FEB 2013	5368	1532		
15	7100	2013	3	MAR 2013	6044	1056		
16	6640	2013	4	APR 2013	6510	130		
17	7500	2013	5	MAY 2013	6567	933		
18	6518	2013	6	JUN 2013	6978	-460		
19	7730	2013	7	JUL 2013	6775	955		
20	8240	2013	8	AUG 2013	7196	1044		
21	8105	2013	9	SEP 2013	7657	448		
22	7290	2013	10	OCT 2013	7854	-564		

Fig. 4.7 The active data file with forecasted and predicted values, plus residuals

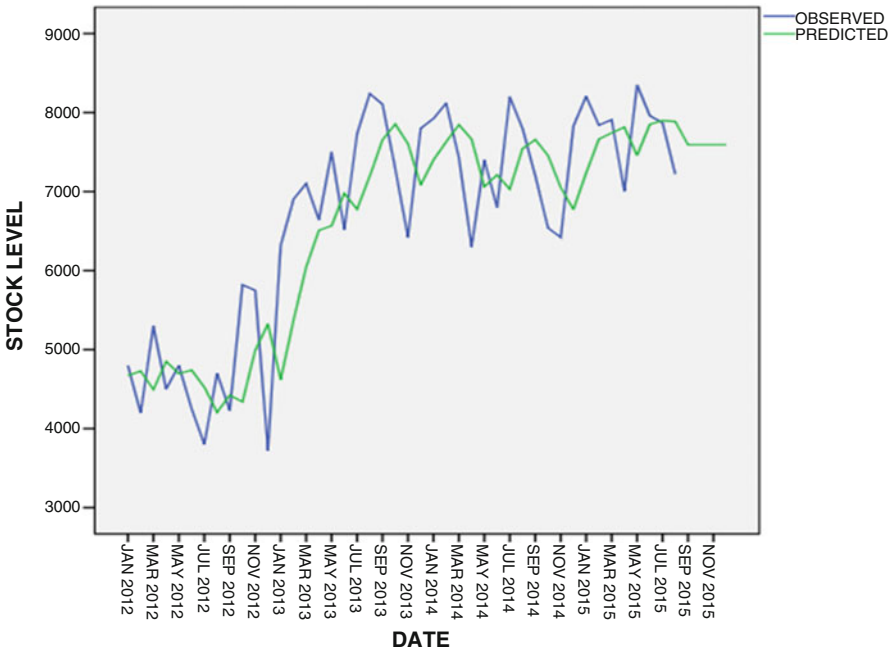


Fig. 4.8 A plot of observed and predicted stock levels

4.2 The Naïve Models

Models that have become known as Naïve 1 and Naïve 2 can be included under the heading of time series models. The Naïve 1 or no change model assumes that a forecast of the series at a particular period equals the actual value at the last period available i.e. $\widehat{Y}_{t+1} = Y_t$ for annual data. This simply says that the forecast for 2017 should equal the actual value for 2016. For monthly data, the subscript changes to $\widehat{Y}_{t+12} = Y_t$; this says that the forecast for June 2017 should equal the actual value for June 2016. For quarterly data, $\widehat{Y}_{t+4} = Y_t$, which says that the forecast for Q1 2017 should equal the actual value for Q1 2016. The Naïve 1 model is often included in forecasting studies since it acts as a yardstick with which other models, like ARIMA or the exponential smoothing class of models may be compared. Importantly, there is the suggestion that this model outperforms more formal forecasting methods in many cases.

The Naïve 2 model assumes that the growth rate in the previous period applies to the generation of forecasts for the current period. For annual data, the model is:

$$\widehat{Y}_{t+1} = Y_t\left[1 + \frac{Y_t - Y_{t-1}}{Y_{t-1}}\right] \ldots\ldots \tag{4.2}$$

For example, if $Y_{2016} = 80$ and $Y_{2015} = 60$, then the quantity $\left[\frac{Y_t - Y_{t-1}}{Y_{t-1}}\right] = 0.33$, indicating a growth rate of 33% from 2015 to 2016. Consequently, the forecast for 2017 would equal the value for 2016 plus this growth rate. The subscripts for the Naïve 2 model change as previously described for annual and quarterly data. The observation recorded one time period ago is called a *lag of 1 time period*. Lagged values are readily computed in IBM SPSS Statistics via clicking:

```
Transform
    Compute Variable ...
```

And then using the inbuilt lag function. The form of the lag function in IBM SPSS Statistics is lag(variable name, no. of cases). Returning to the inventory data, if we wish to lag by a year, the lag command is LAG (LEVEL, 12) as shown in the Compute Variable dialogue box of Fig. 4.9:

The lagged values have been given the variable name LAG12 in Fig. 4.9. Operationalising the lag command produces the results of Fig. 4.10:

LAG12 for January 2013 is equal to the observed value for January 2012. Here, the value of the variable LAG12 thus represents the forecasted value for January 2013 under the Naïve 1 model. The residuals (variable name NResidual) are simply LEVEL – LAG12 and these may be derived via the Compute Variable dialogue box if Fig. 4.9. The Target Variable in Fig. 4.9 would be RESID; the Numeric Expression would be LEVEL – LAG12, which generates the results of Fig. 4.11:

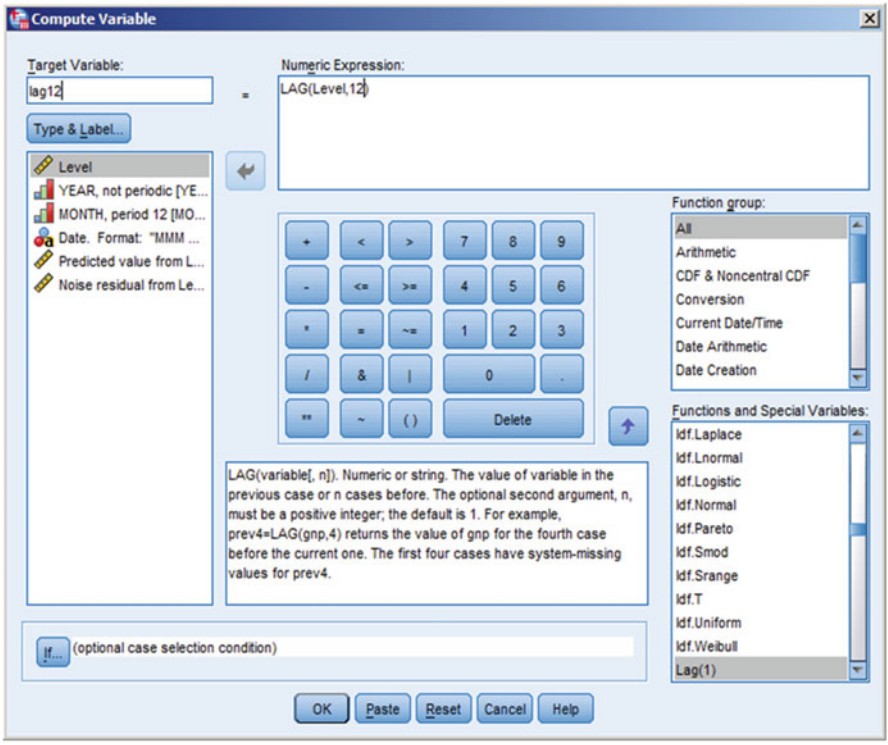


Fig. 4.9 The compute variable dialogue box

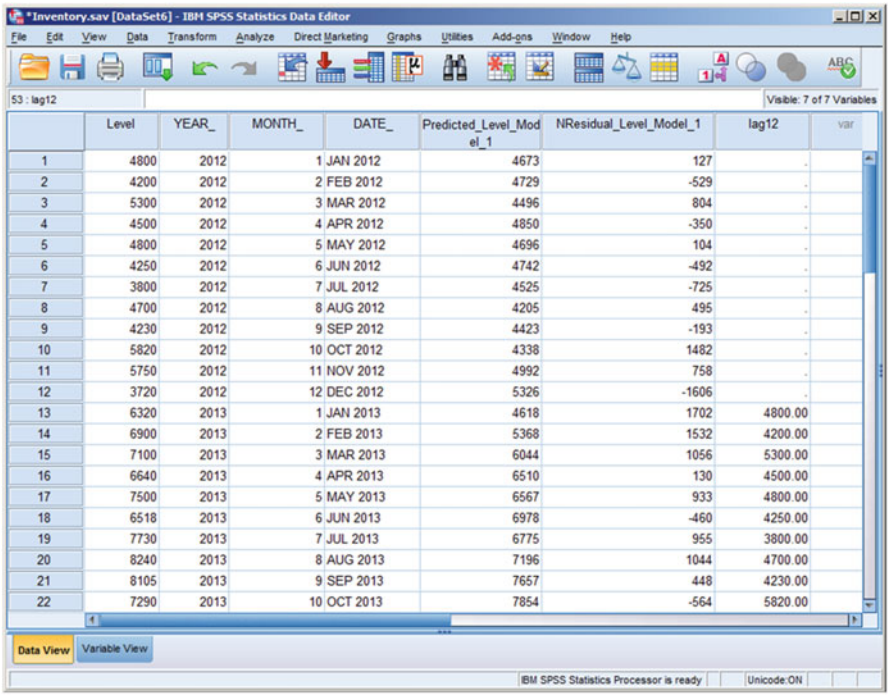
Our data file has insufficient readings to run the Naïve 2 model, since for monthly data and from Eq. (4.2),

$$\widehat{Y}_{t+12} = Y_t\left[1 + \frac{Y_t - Y_{t-12}}{Y_{t-12}}\right].$$

Therefore, consider the data in the file EARNINGS.SAV on the file server, which contains an estimation of UK monthly earnings (variable name EARNINGS) from tourism (£million) from January 2014 to April 2017 inclusive. To apply the Naïve 2 model, we need to create variables lagged 12 months and lagged 24 months. For example and from Eq. (4.2), the forecasted value for May 2017 will equal:

$$\widehat{Y}\ May_{2017} = Y\,May_{2016}\quad\left[1 + \frac{Y\ May_{2017} - Y\ May_{2016}}{Y\ May_{2016}}\right]$$

Two variables that I have named LAG12 and LAG24 have to be created in Fig. 4.12 which indicates that there is a large amount of data loss when applying the Naïve 2 model to monthly data. From Fig. 4.12, we shall lose all data prior to



	Level	YEAR_	MONTH_	DATE_	Predicted_Level_Model_1	NResidual_Level_Model_1	lag12	var
1	4800	2012	1	JAN 2012	4673	127	.	
2	4200	2012	2	FEB 2012	4729	-529	.	
3	5300	2012	3	MAR 2012	4496	804	.	
4	4500	2012	4	APR 2012	4850	-350	.	
5	4800	2012	5	MAY 2012	4696	104	.	
6	4250	2012	6	JUN 2012	4742	-492	.	
7	3800	2012	7	JUL 2012	4525	-725	.	
8	4700	2012	8	AUG 2012	4205	495	.	
9	4230	2012	9	SEP 2012	4423	-193	.	
10	5820	2012	10	OCT 2012	4338	1482	.	
11	5750	2012	11	NOV 2012	4992	758	.	
12	3720	2012	12	DEC 2012	5326	-1606	.	
13	6320	2013	1	JAN 2013	4618	1702	4800.00	
14	6900	2013	2	FEB 2013	5368	1532	4200.00	
15	7100	2013	3	MAR 2013	6044	1056	5300.00	
16	6640	2013	4	APR 2013	6510	130	4500.00	
17	7500	2013	5	MAY 2013	6567	933	4800.00	
18	6518	2013	6	JUN 2013	6978	-460	4250.00	
19	7730	2013	7	JUL 2013	6775	955	3800.00	
20	8240	2013	8	AUG 2013	7196	1044	4700.00	
21	8105	2013	9	SEP 2013	7657	448	4230.00	
22	7290	2013	10	OCT 2013	7854	-564	5820.00	

Fig. 4.10 Lagged values of the variable LEVEL

January 2016. We can now compute the forecasted values from the Naïve 2 model. I have called the forecasted values YHAT, which is given by:

$$YHAT = LAG12\left[1 + \frac{LAG12 - LAG24}{LAG24}\right]$$

and the residuals may be computed as:

$$RESID = EARNINGS - YHAT$$

as shown in Fig. 4.13. The reader can plot the two forecasts, Naives 1 and 2 on a graph (Fig. 4.15) by following the instructions in Fig. 4.14.

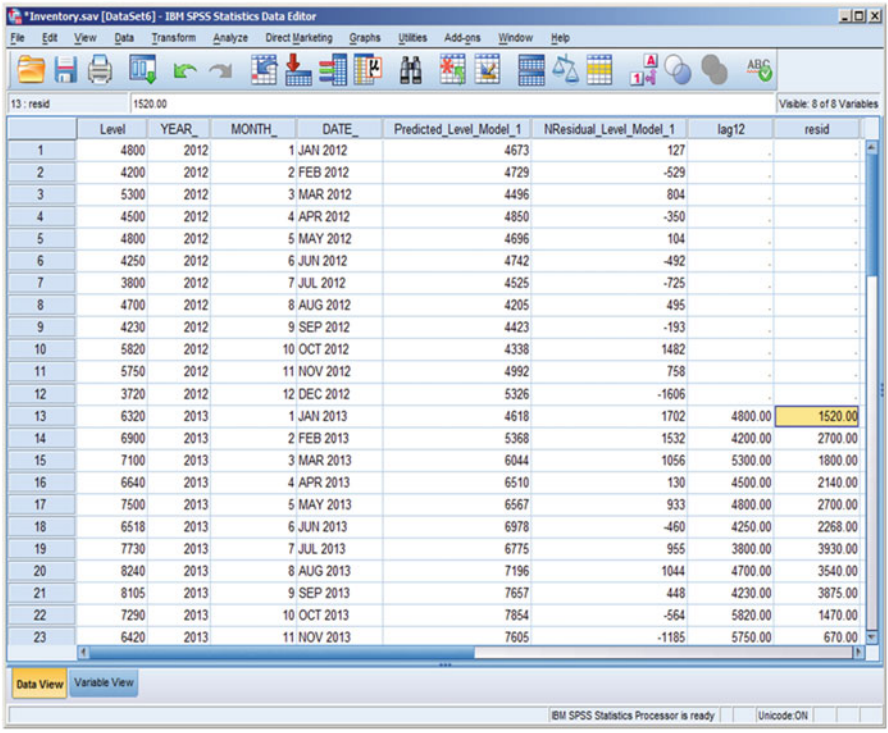


Fig. 4.11 Computation of the residuals from the Naïve 1 model

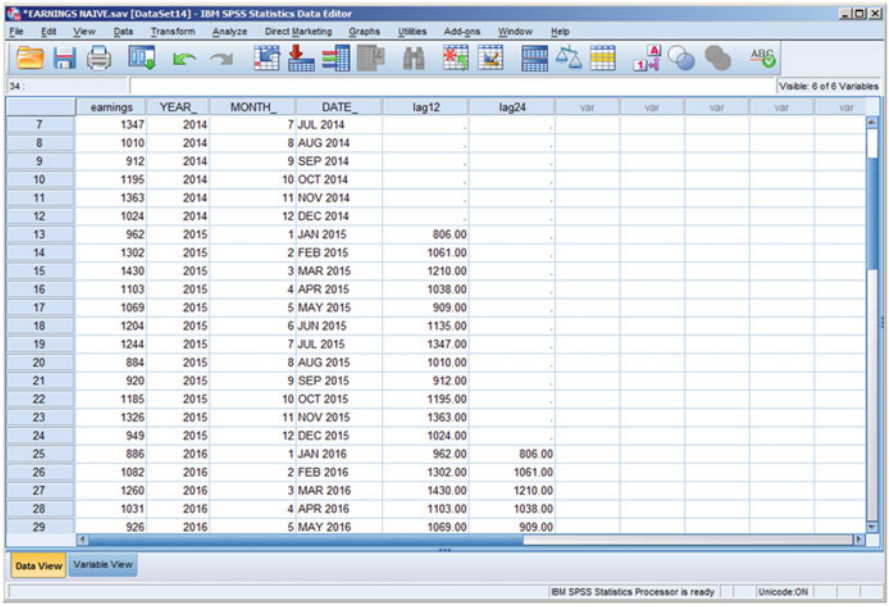
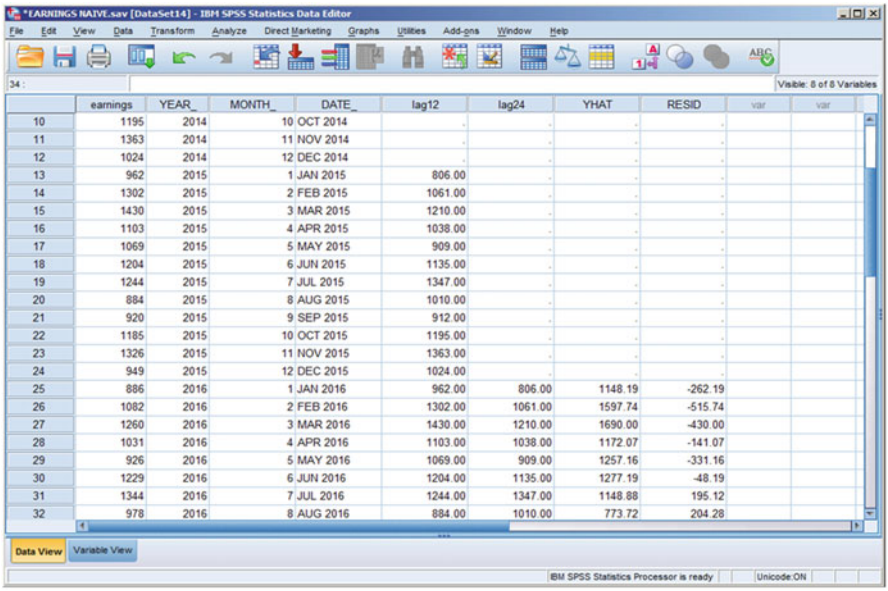


Fig. 4.12 Creation of LAG12 and LAG24



	earnings	YEAR_	MONTH_	DATE_	lag12	lag24	YHAT	RESID	var	var
10	1195	2014	10	OCT 2014	.	.	.	.		
11	1363	2014	11	NOV 2014	.	.	.	.		
12	1024	2014	12	DEC 2014	.	.	.	.		
13	962	2015	1	JAN 2015	806.00	.	.	.		
14	1302	2015	2	FEB 2015	1061.00	.	.	.		
15	1430	2015	3	MAR 2015	1210.00	.	.	.		
16	1103	2015	4	APR 2015	1038.00	.	.	.		
17	1069	2015	5	MAY 2015	909.00	.	.	.		
18	1204	2015	6	JUN 2015	1135.00	.	.	.		
19	1244	2015	7	JUL 2015	1347.00	.	.	.		
20	884	2015	8	AUG 2015	1010.00	.	.	.		
21	920	2015	9	SEP 2015	912.00	.	.	.		
22	1185	2015	10	OCT 2015	1195.00	.	.	.		
23	1326	2015	11	NOV 2015	1363.00	.	.	.		
24	949	2015	12	DEC 2015	1024.00	.	.	.		
25	886	2016	1	JAN 2016	962.00	806.00	1148.19	-262.19		
26	1082	2016	2	FEB 2016	1302.00	1061.00	1597.74	-515.74		
27	1260	2016	3	MAR 2016	1430.00	1210.00	1690.00	-430.00		
28	1031	2016	4	APR 2016	1103.00	1038.00	1172.07	-141.07		
29	926	2016	5	MAY 2016	1069.00	909.00	1257.16	-331.16		
30	1229	2016	6	JUN 2016	1204.00	1135.00	1277.19	-48.19		
31	1344	2016	7	JUL 2016	1244.00	1347.00	1148.88	195.12		
32	978	2016	8	AUG 2016	884.00	1010.00	773.72	204.28		

Fig. 4.13 Forecasted and residual values from the Naïve 2 model

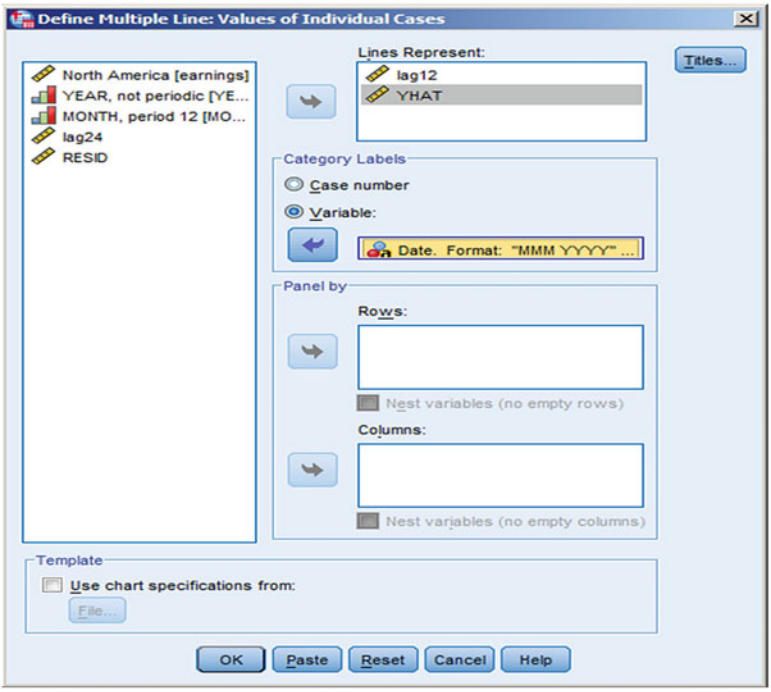


Fig. 4.14 Define graphs with multiple lines, Naïve 1 and 2

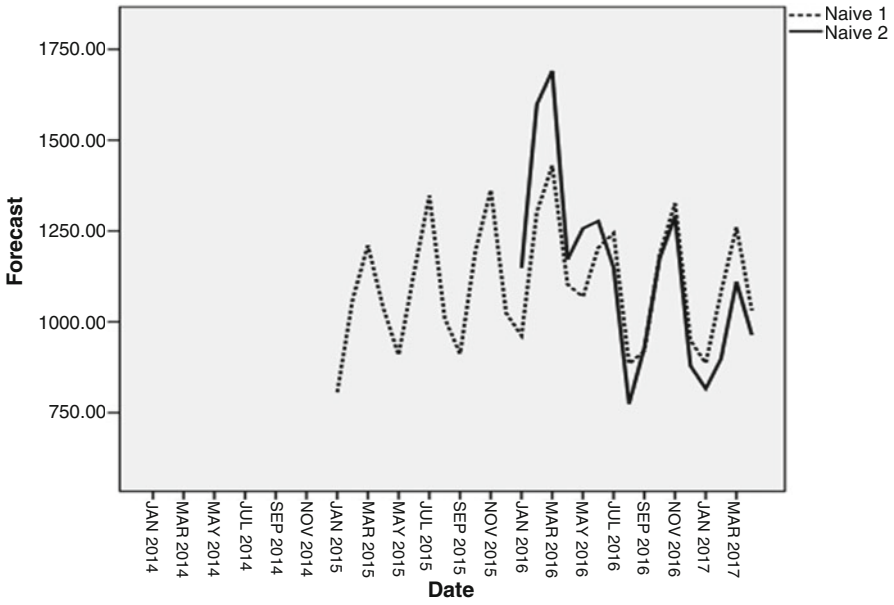


Fig. 4.15 Forecasts generated using the Naïve 1 and 2 models

Part II
Multivariate Methods

Chapter 5
Factor Analysis

The objective of factor analysis is to extract underlying "factors" that explain the correlations among a set of variables. It essentially summarises a large number of variables with a smaller set of derived variables. For example, in a survey, many variables relating to consumption of products might be expressed as a function of just three factors related to 'product quality', 'product utility' and 'product price'. Factor analysis seeks to identify these underlying (and not directly observable) dimensions.

The major assumption of factor analysis is that observed correlations between variables can be used to explain complex phenomena. For example, in terms of product evaluation, suppose that consumers had to rate a household product on a series of variables, amongst which were 'value for money', 'long lasting', 'better than rival products' and 'child proof'. A 1 to 5 rating scale could be employed to measure these variables. If it is found that scores on these four variables are highly inter-correlated, it could be due to their sharing and reflecting the factor of 'product quality'. While it is possible that all the variables in a study significantly contribute to a specific factor, the researcher hopes that it is only a subset for the purpose of interpreting the factor. We require factors to be meaningful, simple and interpretable.

In general, let $X_1, X_2, \ldots, X_k$ be k variables that have been measured in a study over a reasonable large number of cases. Let F_j be the jth factor that underlies the data set, then we may say that:

$$F_j = W_{j1}X_1 + W_{j2}X_2 + \ldots + W_{jk}X_k,$$

Where the W_{ji} are weights known as *factor score coefficients*, *facto weights* or *factor loadings*. The larger is a particular coefficient, the more associated variable contributes to that factor. Collecting together all the variables that contribute most to the factor (i.e. they possess the highest factor score coefficients) will hopefully enable one to label or name the factor.

A. Aljandali, *Multivariate Methods and Forecasting with IBM® SPSS® Statistics*,
Statistics and Econometrics for Finance, DOI 10.1007/978-3-319-56481-4_5

5.1 The Correlation Matrix

The derivation of the inter-correlations between all the variables in a study is a useful first step in factor analysis. During the factor analysis procedure, IBM SPSS Statistics computes the inter-correlations between all pairs of variables. Note that the level of measurement should be such that the Pearsonian correlation coefficient is an acceptable summary statistic. If the correlations between variables are small, it is unlikely that they share common factors.

To illustrate the link between variable inter-correlations and factor analysis, I shall use a data file that contains aspects of educational provision in 10 London boroughs in the early 2000s. The appropriate data file is LONDON EDUCATION. SAV. The seven variables examined are:

X_1 – pupil-teacher ratio in primary schools,
X_2 – expenditure/1000 persons on primary school teachers,
X_3 – expenditure/1000 persons on secondary school teachers,
X_4 – expenditure/1000 persons on non-teaching staff,
X_5 – administrative costs/1000 persons,
X_6 – net expenditure/1000 persons on secondary education and.
X_7 – net expenditure/1000 persons on tertiary education.

It may be recalled that the IBM SPSS Correlations procedure is accessed by:

```
Analyse
      Correlate
          Bivariate...
```

The inter-correlations between the seven study variables are shown in Fig. 5.1. Note that X_1 and X_2 are significantly correlated with each other; so too are the pairs X_3 with X_6 and X_4 with X_7. Possibly these three significant inter-correlations indicate that each pair of variables share common factors. It should be noted that the user may request the correlation matrix of Fig. 5.1 during the IBM SPSS Factor Analysis procedure, rather than via the IBM SPSS Correlations procedure above.

5.2 The Terminology and Logic of Factor Analysis

By definition, the correlation matrix of Fig. 5.1 has a leading diagonal (top left to bottom right) of 1's. If the off-leading diagonal elements were zero, then there would be no inter-correlations to suggest any underlying factors. A matrix with leading diagonal elements of 1 and off-diagonal elements of zero is called an identity matrix. If the correlation matrix is an identity matrix then the application of factor analysis is inappropriate. Bartlett's test of *sphericity* tests:

Correlations

		x1	x2	x3	x4	x5	x6	x7
x1	Pearson Correlation	1	-.863**	-.460	-.243	-.096	-.162	-.434
	Sig. (2-tailed)		.001	.181	.499	.792	.655	.210
	N	10	10	10	10	10	10	10
x2	Pearson Correlation	-.863**	1	.750*	-.021	.177	-.151	.222
	Sig. (2-tailed)	.001		.012	.954	.626	.677	.538
	N	10	10	10	10	10	10	10
x3	Pearson Correlation	-.460	.750*	1	-.091	.120	-.622	-.182
	Sig. (2-tailed)	.181	.012		.802	.742	.055	.615
	N	10	10	10	10	10	10	10
x4	Pearson Correlation	-.243	-.021	-.091	1	.362	-.179	.552
	Sig. (2-tailed)	.499	.954	.802		.304	.620	.098
	N	10	10	10	10	10	10	10
x5	Pearson Correlation	-.096	.177	.120	.362	1	-.101	-.225
	Sig. (2-tailed)	.792	.626	.742	.304		.782	.532
	N	10	10	10	10	10	10	10
x6	Pearson Correlation	-.162	-.151	-.622	-.179	-.101	1	.083
	Sig. (2-tailed)	.655	.677	.055	.620	.782		.819
	N	10	10	10	10	10	10	10
x7	Pearson Correlation	-.434	.222	-.182	.552	-.225	.083	1
	Sig. (2-tailed)	.210	.538	.615	.098	.532	.819	
	N	10	10	10	10	10	10	10

**.Correlation is significant at the 0.01 level (2-tailed).

*.Correlation is significant at the 0.05 level (2-tailed).

Fig. 5.1 Inter-correlations between study variables

$$H_0 : \text{the correlation matrix is an identity matrix,}$$

and has an associated statistic that is closely distributed as chi-square. If the significance associated with Bartlett's statistic is less than 0.05, then we reject the above null hypothesis and may fruitfully continue with factor analysis. Similarly available is the *Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy*. As a rule of thumb, if the KMO statistic is greater than or equal to 0.6, then the user may proceed with factor analysis.

Prior to application of factor analysis, all the study variables are standardized to have zero mean and unit variance. In the example of the London boroughs' educational data, there are thus 7 units of variance in the study, 7 X_i each having a variance of 1. As discussed in the previous section, linear functions of the study variables, here X_1 to X_7,are derived by the factor analysis procedure. The first factor that is derived (or *extracted*) is the combination that accounts for the largest amount of the 7 units of variance in the data set. The second factor accounts for the second largest amount of variance in the sample and so on. Technically, it is possible to extract as many factors as there are variables in the study.

As part of the output, each study variable has associated with it a quantity called the *communality*. The communality of each variable is the proportion (expressed in decimal form) of variance in each variable that is explained by all the factors thus far derived. For example, if a study generated three underlying factors and the communality associated with a particular variable is 0.96, then 96% of the variance in that variable is explained by the three extracted factors.

The amount of variance in the entire sample that is explained by each factor is referred to as an *eigenvalue* in the context of factor analysis. For example, if the first extracted factor in a study has an eigenvalue of 8.0 and there are 20 variables in that study (i.e. a total of 20 units of variance), then this first factor explains $8/20 = 40.0\%$ of the variation in the whole sample. A conventional criterion is that only factors with eigenvalues in excess of 1 are significant. (Factors with eigenvalues or explained variances of less than 1 are no better than a single variable). At this point, it is best to examine these measures for the London borough's data.

The IBM SPSS Factor Analysis procedure is accessed via:

```
Analyse
   Dimension Reduction
      Factor...
```

Which gives rise to the *Factor Analysis dialogue box* of Fig. 5.2. The variables X_1 to X_7 inclusive are entered into the 'Variables' box. At this point, simply click the OK button to operationalise, but in later sections, use will be made of the buttons at the bottom of the dialogue box of Fig. 5.2. In Fig. 5.3, under the heading

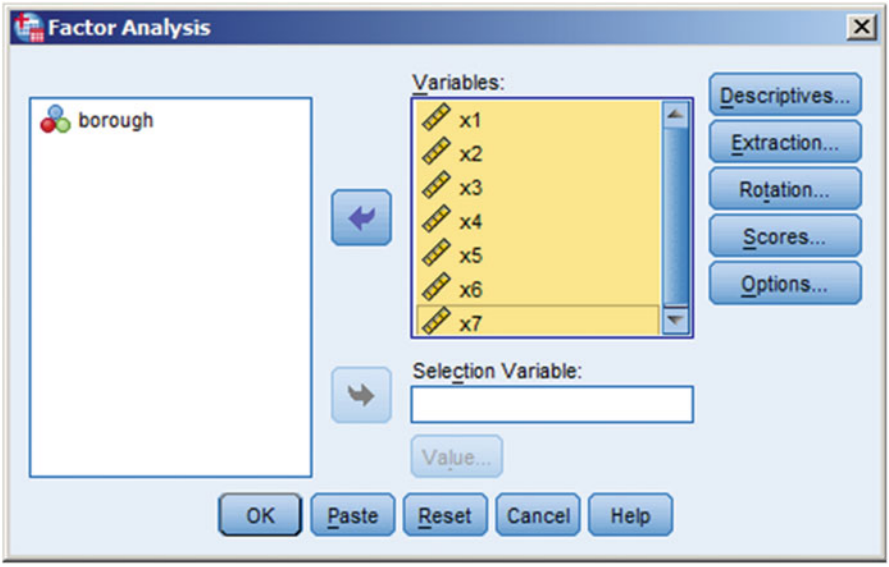


Fig. 5.2 The factor analysis dialogue box

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.587	36.951	36.951	2.587	36.951	36.951
2	1.764	25.205	62.157	1.764	25.205	62.157
3	1.358	19.405	81.561	1.358	19.405	81.561
4	1.035	14.789	96.350	1.035	14.789	96.350
5	.216	3.093	99.443			
6	.033	.478	99.921			
7	.006	.079	100.000			

Extraction Method: Principal Component Analysis.

Fig. 5.3 The eigenvalues associated with the factor extraction

Fig. 5.4 The communalities associated with the study variables

Communalities

	Initial	Extraction
x1	1.000	.969
x2	1.000	.979
x3	1.000	.973
x4	1.000	.939
x5	1.000	.968
x6	1.000	.983
x7	1.000	.934

Extraction Method: Principal Component Analysis.

'Extraction of sum of squared loadings', (the only heading to be considered here) we find the eigenvalues given under the 'Total' column. There are four factors with eigenvalues in excess of unity.

The first factor has an eigenvalue of 2.587, which represents $2.587/7 = 36.957\%$ of the variation in the whole sample. The second factor has an eigenvalue of 1.764, representing $1.764/7 = 25.2\%$ of the variation in the whole sample. Hence the first two factors explain a cumulative total of $36.957\% + 25.2\% = 62.157\%$ of the variation in the sample. All four of the significant extracted factors explain a cumulative total of 81.704% of all the variation in the sample, which is a healthy total. About 18% of the variance in the sample is not explained by the variables being members of the four extracted factors.

Figure 5.4 presents the communalities (under the heading 'Extraction') associated with the seven study variables. 96.9% of the variation in X_1 is explained by the four, extracted factors.

5.3 Rotation and the Naming of Factors

Although there are four factors underlying our study, at present we do not know what these factors represent. We need to name them. Part of the IBM SPSS Statistics output will be present the *factor loadings* mentioned in the introduction. These loadings are the correlations between each variable and the factor under consideration. Variables with a large loading on a factor are closely related to that factor. In our present example, each variable will have three loadings – one for each extracted factor. To name a factor, we select only those variables that have a high loading on that factor and use the names of these variables to derive an overall phrase of English to represent them. The factors should be named parsimoniously. Often, however, it can be difficult to decide whether a variable has a high enough loading on a particular factor, or indeed to which factor it may be appropriately ascribed.

Consider Fig. 5.5, which shows the loadings of four variables on two extracted factors. Note the negative loadings are possible; a variable may be negatively correlated with a factor. To simplify allocating the variables to the factors, we may if desired, *rotate* the axes of Fig. 5.5. Imagine rotating the axes through 45° clockwise. Then the first variable would have a high loading on factor 2, but low loading on factor 1. If the axes are rotated so as to preserve the right angle between them, the rotation is called *orthogonal*; if the user is happy for a rotation to produce axes that are not mutually perpendicular, then the rotation is said to be *oblique*.

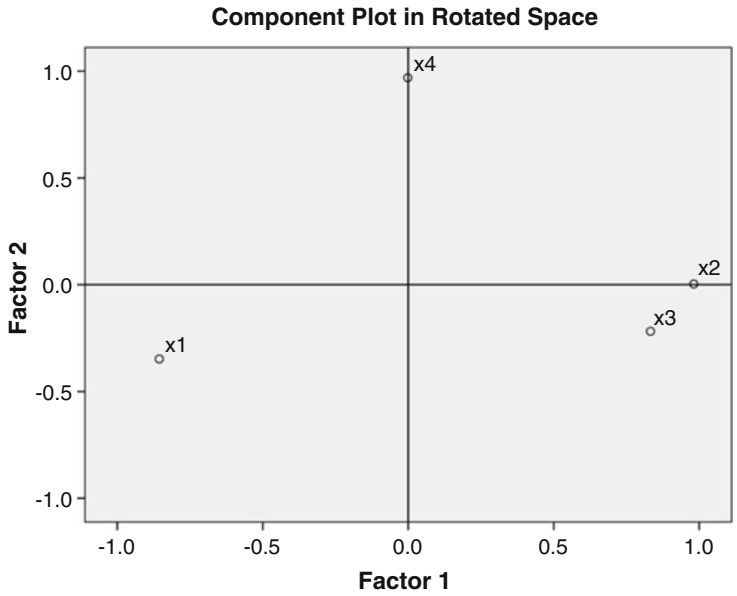


Fig. 5.5 Loadings of four variables on two factors

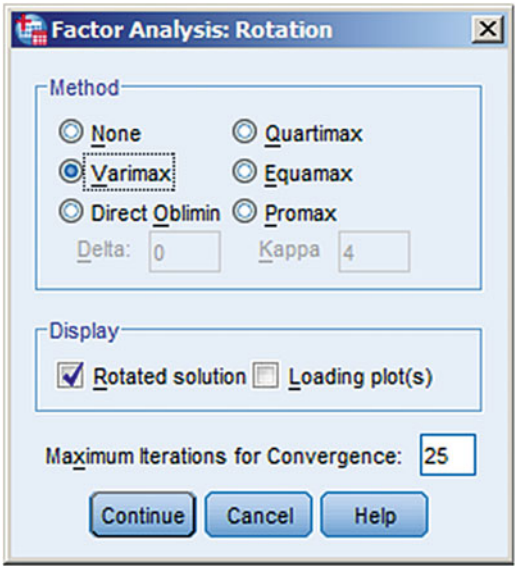


Fig. 5.6 The factor analysis: rotation dialogue box

There are several methods of rotation available in IBM SPSS Statistics. Perhaps the most commonly used is the *varimax* rotation. This rotation minimises the number of variables that have high loadings on each factor and like all rotations, it simplifies the interpretation of the factors. The use of a rotation should be encouraged as it invariably assists interpretation.

Figure 5.2 presented the *Factor Analysis dialogue box*. At the bottom of this box, click the Rotation button to generate the factor analysis; Rotation dialogue box of Fig. 5.6. Of the available options, Varimax has been chosen in this dialogue box. Upon returning to the dialogue box of Fig. 5.2 the output of Fig. 5.7 is produced.

At the top of Fig. 5.5 under the heading 'Component Matrix' are the unrotated factor loadings for the seven variables on the three factors. (The factors are referred to as "components" here). Beneath this are the factor loadings after the Varimax rotation has been employed. It is these latter loadings that are used for the purpose of naming the four factors.

Variables X_1 and X_2 have the highest loadings on the first factor. Referring back to the list of variables on page 98, this is a factor that stresses boroughs with low pupil-teacher ratios (X_1 and note that it is negative loading) tending to spend relatively high amounts per 1000 inhabitants on primary school teachers (X_2). We could subjectively name this factor as 'level of provision of teaching resources in primary schools'. Recall that in the original correlation matrix of Fig. 5.1, these two variables were significantly and strongly negatively correlated. A majority of the variance in these two variables (as per their communalities) is doubtless explained by their being members of factor 1.

Turning to the second factor, variables X_6, X_3 and X_5 have the highest loadings. (Where to draw the line is a subjective matter that should be made within the

Fig. 5.7 Unrotated and rotated factor loadings

Component Matrix[a]

	Component			
	1	2	3	4
x1	-.849	-.334	.317	-.192
x2	.937	-.076	-.282	.124
x3	.791	-.568	-.005	-.155
x4	.279	.616	.690	-.077
x5	.256	-.083	.637	.700
x6	-.344	.505	-.545	.560
x7	.315	.827	-.019	-.387

Extraction Method: Principal Component Analysis.

a. 4 components extracted.

Rotated Component Matrix[a]

	Component			
	1	2	3	4
x1	-.929	-.286	.151	-.038
x2	.974	.005	.167	.058
x3	.669	-.233	.684	.054
x4	-.040	.875	.130	.392
x5	.092	.024	.042	.978
x6	.030	-.065	-.988	-.046
x7	.233	.873	-.139	-.314

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 7 iterations.

context of what makes sense in the study at hand). This factor reflects boroughs with low net expenditure on secondary education as a whole, having relatively high expenditure on secondary teaching staff and relatively high administrative costs. This second factor could be labelled '*high expenditure on administrative and teaching staff at the expense of other aspects of educational provision*'. The third factor shows that variables X_4 and X_7 have the highest loadings. This factor reflects boroughs with relatively high expenditures on non-teaching staff and tertiary education. It could be labelled '*expenditure on ancillary staff and tertiary education at the expense of primary and secondary education*'.

It must be stressed that the naming of factors is a subjective exercise and an elements of common sense is involved within the problem context.

5.4 Factor Scores in IBM SPSS Statistics

A useful, final phase of factor analysis is to obtain *factor scores*. These reflect how much a particular case (here, the boroughs) possess the characteristic of an extracted factor. The factor scores may be saved in the working file for further analysis. There are three types of factor score in IBM SPSS Statistics, the *Anderson-Rubin*, *Bartlett* and *regression* factor scores. All of these have a zero mean; in addition the Anderson-Rubin factor scores have a variance of one. The larger the factor score in a positive/negative sense, the more/less that case possess the characteristic of that factor. Factor scores are generated for each case on each extracted factor.

At the bottom of the Factor Analysis dialogue box of Fig. 5.2, click the Scores button to produce the *Factor Analysis: Factor Scores dialogue box* of Fig. 5.8. The Anderson-Rubin method of generating factor scores has been selected and the resultant scores are to be added to the working file, as shown in Fig. 5.9. On the first factor, boroughs numbered 1, 6 and 13 have relatively high scores; they tend to possess the characteristic of that factor. Boroughs numbered 3, 12 and 16 have relatively large negative scores, indicating that they do not possess the characteristics of the first factor. On factor two, borough 5 has a large positive score and borough 11 a relatively large negative score. Further analysis could involve recoding the factor scores into 'high', 'medium' and 'low', e.g. by assigning one third of the boroughs into each group. If one knew which political party ran each

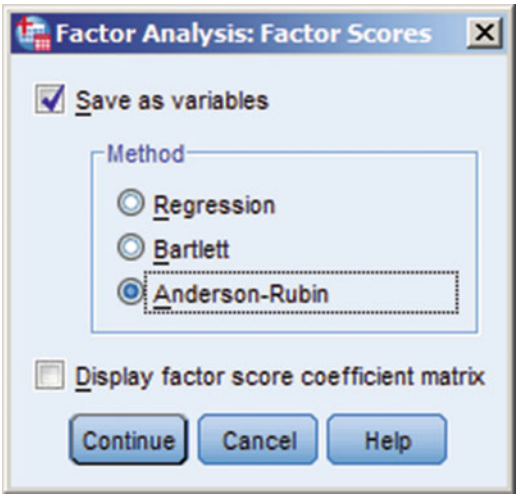


Fig. 5.8 The factor analysis: factor scores dialogue box

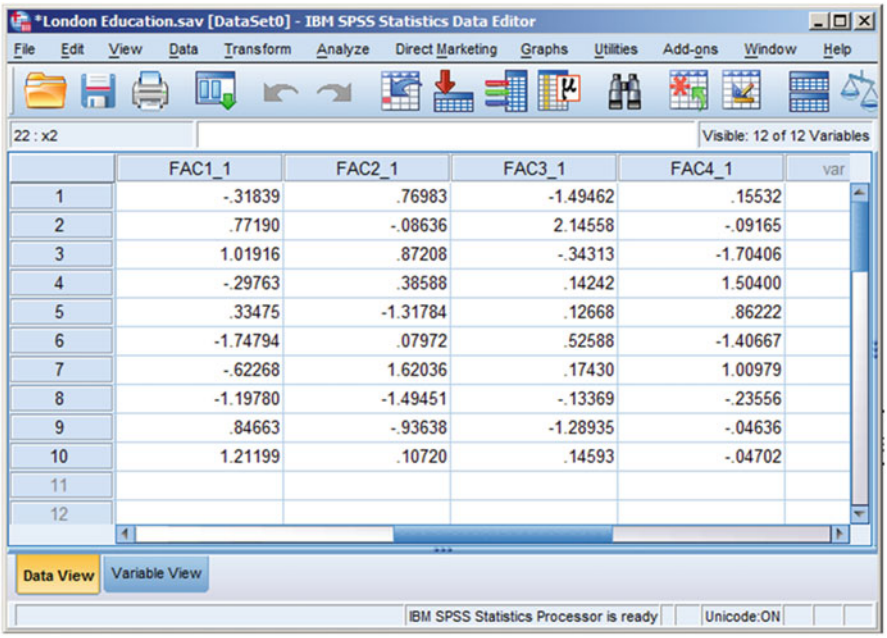


Fig. 5.9 Factor scores added to the active file

borough, it would be possible to form a contingency table and to test whether the magnitude of the factor scores depended or not on the political party involved. Contingency analysis and the ci-square are covered in the first volume of this guide.

Chapter 6
Discriminant Analysis

Based on a collection of variables, such as annual income, age, marital status etc., discriminant analysis seeks to distinguish among several mutually exclusive groups, such as good and bad credit risks. The available data are values for cases whose group membership is known i.e. individuals who have already proven to be good or bad credit risks. Discriminant analysis enables us to:

- Identify which of the collected variables are important for distinguishing among groups and
- Develop a procedure for predicting group membership for new cases whose group membership is presently undetermined.

6.1 The Methodology of Discriminant Analysis

Discriminant analysis produces linear combinations of the independent (or *predictor*) variables and uses them as a basis for classifying cases into one of a series of mutually exclusive groups. For discriminant analysis to be "optimal" in the sense that the probability of a misclassification is minimized, the variables should be samples from normal populations. However, there is evidence that even in the case of dichotomous predictor variables (e.g. of the "yes/no" type), discriminant analysis often performs adequately.

In discriminant analysis (and indeed other multivariate statistical procedures like factor analysis), the emphasis is on analysing the variables together. By considering variables together, we are able to incorporate important information about their relationships. In discriminant analysis, a linear combination of the predictor variables is formed (called a linear discriminant function) and serves as a basis for assigning cases to groups. The linear discriminant function is:

A. Aljandali, *Multivariate Methods and Forecasting with IBM® SPSS® Statistics*,
Statistics and Econometrics for Finance, DOI 10.1007/978-3-319-56481-4_6

$$D = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p,$$

Where the X_i are values of the predictor variables and the b_i are coefficients estimated from the data. If this linear function is to distinguish between groups of good and bad credit risks for example, these groups should differ in their D scores, which are referred to as *discriminant scores*. Hence, the b_i are computed so that the values of the discriminant function differ as much as possible between the groups. The b_i are called the *discriminant function coefficients*. IBM SPSS Statistics reports the b_i in standardized form i.e. all the predictor variables are initially standardized to have a zero mean and unit variance. Using the D scores, IBM SPSS Statistics computes the probabilities of each case belonging to the various groups in the study. It should be noted that only one discriminant function is needed to distinguish between two groups, two discriminate functions to distinguish between three groups etc.

Discriminant analysis produces three statistics that assess the adequacy of any discrimination achieved. The *square of the canonical correlation* represents the proportion of total variance in the discriminant function scores explained by differences between the groups. It is akin to the coefficient of determination in regression. *Eigenvalues* are also computed. (Here, they represent the ratio of the between groups sum of squares and the within groups sum of squares). Large eigenvalues indicate "good" linear discriminant functions. Finally, Wilks' Lambda (λ) is the proportion of total variance in the discriminant scores explained by differences among groups. It might be noted that:

$$(\text{Canonical correlation})^2 + \lambda = 1$$

6.2 Discriminant Analysis in IBM SPSS Statistics

The data in the IBM SPSS file LIBRARY.SAV relate to library provision in London's outer boroughs. The variable POPN is the population of each borough, coded as '0' or '1' according to whether the population is respectively below or above the mean outer London population figure. The remaining variables are:

X_1 – no. of library staff,
X_2 – number of library points,
X_3 – reference books (000's),
X_4 – total books held (000's),
X_5 – no. of serials subscribed to,
X_6 – expenditure per 1000 inhabitants on books,
X_7 – expenditure per 1000 inhabitants on newspapers and.
X_8 – total expenditure per 1000 inhabitants.

The variable POPN represents the groups in this example. There is however, no need for discriminant analysis to be restricted to just two groups. X_1 and X_8 are the

predictor variables. Discriminant is used here to see if aspects of library provision as exemplified by the predictor variables, can be used to estimate whether or not a borough has a relatively large or small population. Conversely, one can examine variables that discriminate library provision on boroughs of relatively low and high population.

The IBM SPSS Discriminant Analysis procedure is accessed via:

```
Analyse
   Classify
      Discriminant...
```

Giving rise to the *Discriminant Analysis dialogue box* of Fig. 6.1. The 'Grouping Variable' is POPN and the 'Independents' are X_1 to X_8. Note that the variable POPN has two question marks besides it, since IBM SPSS requires the codes used for the groups to be stated. Click the Define Range button to generate the *Discriminant Analysis: Define Ranges dialogue box* of Fig. 6.2, in which the codes of 0 and 1 are entered as appropriate.

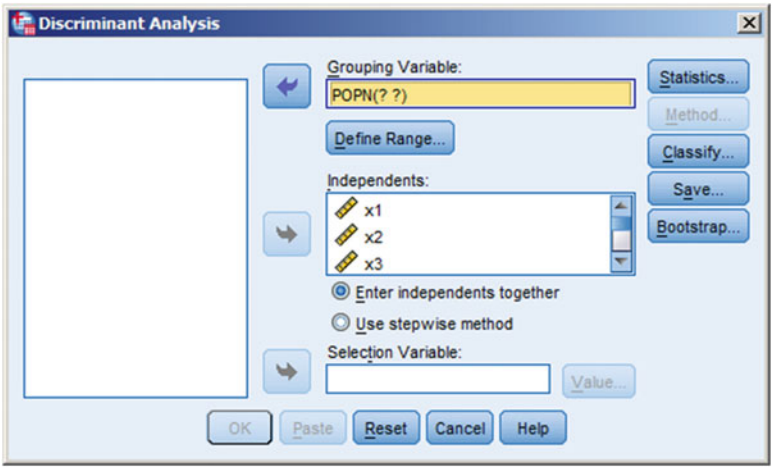


Fig. 6.1 The discriminant analysis dialogue box *x*

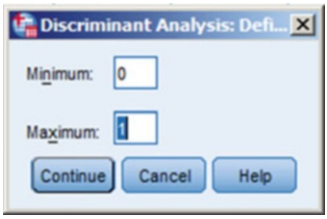


Fig. 6.2 The discriminant analysis: define ranges dialogue box

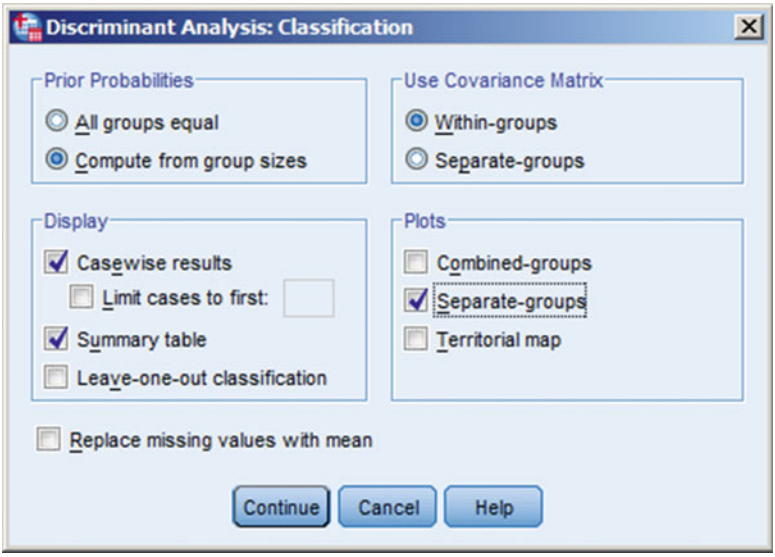


Fig. 6.3 The discriminant analysis: classification dialogue box

Returning to the *Discriminant Analysis dialogue box* of Fig. 6.1, permits some control over the output displayed. Clicking the Classify button at the bottom of this dialogue box gives rise the *Discriminant Analysis: Classification dialogue box of* Fig. 6.3. Under the heading 'Prior probabilities', I have selected the option that the pre-analysis (prior) probabilities of group membership should be computed from the size of the groups and determine group memberships. These values are used in the process of classifying cases. Under the heading 'Display', the choice "casewise results" will present codes for the actual group membership and the code for the predicted group. Probabilities of group membership and the discriminant scores for each case in the analysis. Also under this heading, selection of the "summary table" will show the numbers of cases correctly and incorrectly classified by the discriminant analysis. This table is called the *confusion matrix*. Under the heading 'Plots', I have chosen "separate groups", which will present a histogram of the discriminant scores for all groups (here two groups) in the study.

6.3 Results of Applying the IBM SPSS Discriminant Procedure

Figures 6.4, 6.5 and 6.6 present part of the results of running discriminant analysis on our library provision data. The output is in several parts.

The canonical correlation indicates that a proportion of $(.931)^2$ or .866 of the variance in the discriminant scores is explained by differences between the groups,

Summary of Canonical Discriminant Functions

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	6.477[a]	100.0	100.0	.931

a. First 1 canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.134	22.130	8	.005

Standardized Canonical Discriminant Function Coefficients

	Function 1
x1	-1.313
x2	-0.39
x3	1.101
x4	-.733
x5	.145
x6	.641
x7	.127
x8	.886

Fig. 6.4 IBM SPSS output from discriminant analysis

or if you prefer, Wilks' Lambda indicates that a proportion of .134 is not explained. The chi square statistic (which is based on Wilks' Lambda) tests the hypothesis that in the populations from which the samples are drawn, there is no difference in the group mean scores on D – the discriminant scores. Adopting the conventional significance level of 0.05, we here find that this hypothesis is rejected, so our discriminant analysis may be regarded as successful. Under the heading 'Functions at Group Centroids' we see that the mean discriminant score for the low population group is 3.237, whereas the mean for the high population group is − 1.766. If new

Structure Matrix

	Function 1
x1	-.495
x4	-.442
x2	-.351
x3	-.150
x6	.139
x8	.106
x5	-.031
x7	-.008

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function

Functions at Group Centroids

POPN	Function 1
0	3.237
1	-1.766

Unstandardized canonical discriminant functions evaluated at group means

Fig. 6.4 (continued)

Prior Probabilities for Groups

POPN	Prior	Cases Used in Analysis	
		Unweighted	Weighted
0	.353	6	6.000
1	.647	11	11.000
Total	1.000	17	17.000

Casewise Statistics

	Case Number	Actual Group	Highest Group					Second Highest Group			Discriminant Scores
			Predicted Group	P(D>d | G=g)			Squared Mahalanobis Distance to Centroid	Group	P(G=g | D=d)	Squared Mahalanobis Distance to Centroid	Function 1
				p	df	P(G=g | D=d)					
Original	1	0	0	.122	1	1.000	2.396	1	.000	42.907	4.785
	2	0	0	.555	1	1.000	.348	1	.000	19.467	2.647
	3	0	0	.342	1	1.000	.903	1	.000	35.435	4.187
	4	0	0	.394	1	1.000	.726	1	.000	34.272	4.089
	5	0	0	.255	1	.998	1.296	1	.002	14.932	2.099
	6	0	0	.105	1	.978	2.630	1	.022	11.429	1.615
	7	1	1	.676	1	1.000	.175	0	.000	21.011	-1.347
	8	1	1	.354	1	1.000	.859	0	.000	35.156	-2.692
	9	1	1	.894	1	1.000	.018	0	.000	26.369	-1.898
	10	1	1	.198	1	.999	1.655	0	.001	13.809	-.479
	11	1	1	.564	1	1.000	.332	0	.000	19.588	-1.189
	12	1	1	.931	1	1.000	.007	0	.000	25.891	-1.852
	13	1	1	.613	1	1.000	.255	0	.000	30.334	-2.271
	14	1	1	.244	1	.999	1.356	0	.001	14.729	-.601
	15	1	1	.888	1	1.000	.020	0	.000	26.452	-1.906
	16	1	1	.162	1	1.000	1.959	0	.000	40.986	-3.165
	17	1	1	.799	1	1.000	.065	0	.000	27.638	-2.020

Classification Results[a]

		POPN	Predicted Group Membership		Total
			0	1	
Original	Count	0	6	0	6
		1	0	11	11
	%	0	100.0	.0	100.0
		1	.0	100.0	100.0

a. 100.0% of original grouped cases correctly classified.

Fig. 6.4 (continued)

cases have a discriminant score near − 1.766, they would be classified as high population boroughs.

The 'Standardized Canonical Discriminant Function Coefficients' give us the equation of the linear discriminant function that is used to generate the discriminant scores and which could be used to forecast group membership of new cases. The linear discriminant function is:

$$D = -1.313X_1 - .039X_2 + \ldots + .886X_8.$$

It is tempting to interpret the magnitude of the above coefficients as indicators of the relative importance of the variables. However, since the variables are correlated, it is not possible to assess the importance of an individual variable. The value of a coefficient depends on the other variables included in the equation. A better way to assess the contribution of a variable to the discrimination process is to examine the correlations between the values (D) of the discriminant function and the original variables. These correlations are given in the 'Structure Matrix' and go

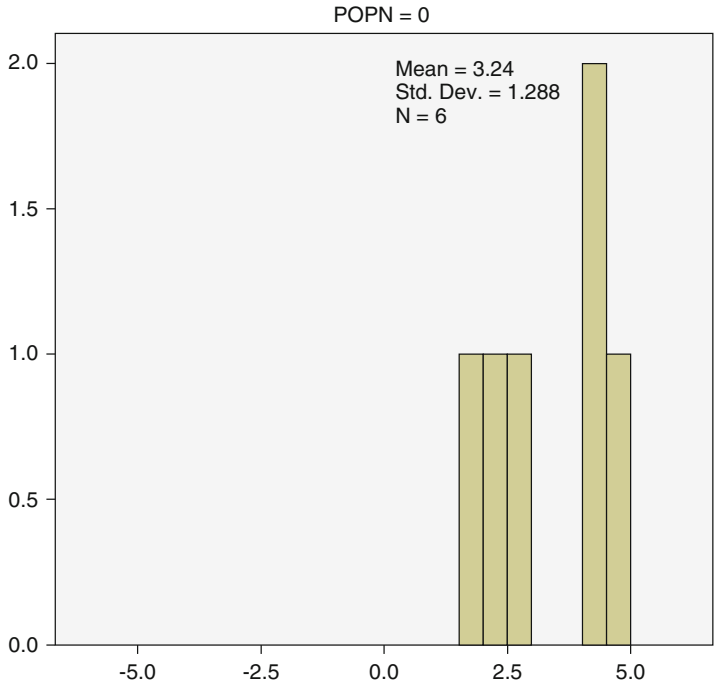


Fig. 6.5 Histogram of discriminant scores for the low population group

under the awesome title of "pooled within-groups correlations between discriminating variables and canonical discriminant functions". The correlations in the 'Structure Matrix' are ordered in terms of absolute magnitude. The negative sign for X_1 indicates that small discriminant scores are associated with high numbers of library staff or if you prefer, large scores are associated with small numbers of such staff. The signs are arbitrary. The most important variables that discriminate between the two groups of POPN are X_1 – number of library staff, X_4 – total number of books held and X_2 – number of library points.

The prior probability for group membership for POPN = 0 is 6/17 = 0.353. The casewise statistics present actual and predicted group membership. The (squared) Mahalanobis distance is a measure of how much a case's values on the independent variables differ from the average of all cases. A large Mahalanobis distance identifies a case as having extreme values on one or more of the independent variables. For example, although borough number 16 is not misclassified by the discriminant analysis, it has the largest Mahalanobis distance of 2.923. Examination of the data file suggests that this due to a particularly high reading on X_4.

At the end of the results in 6.4 is the Confusion Matrix presented under the heading 'Classification Results'. This confirms that none of the 17 boroughs, was

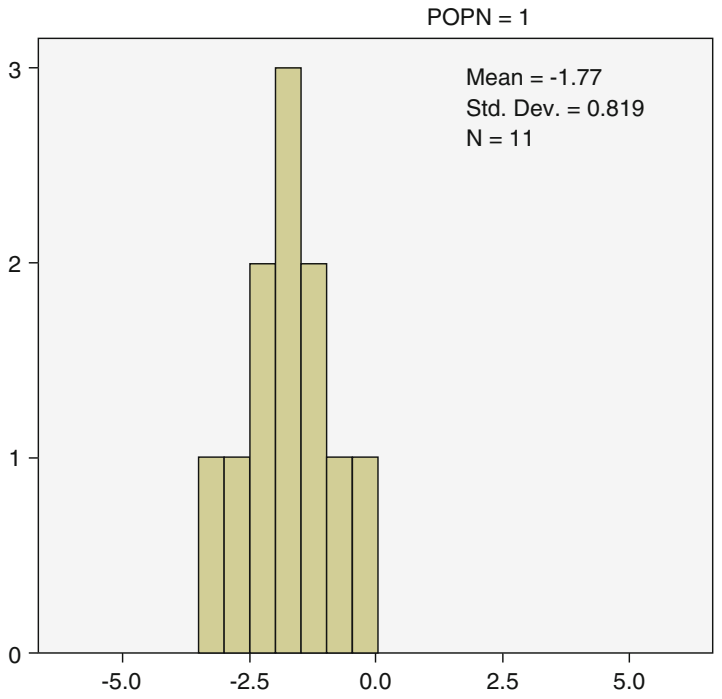


Fig. 6.6 Histogram of discriminant scores for the high population group

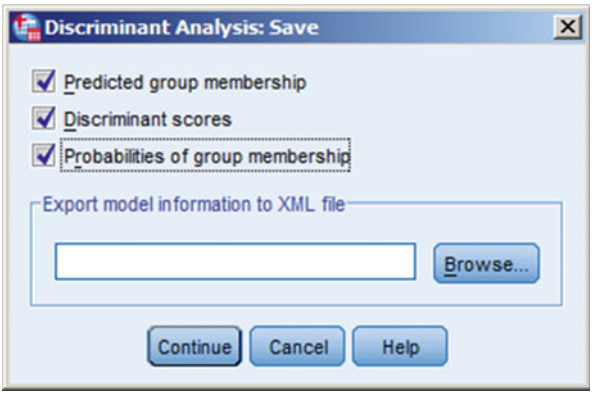
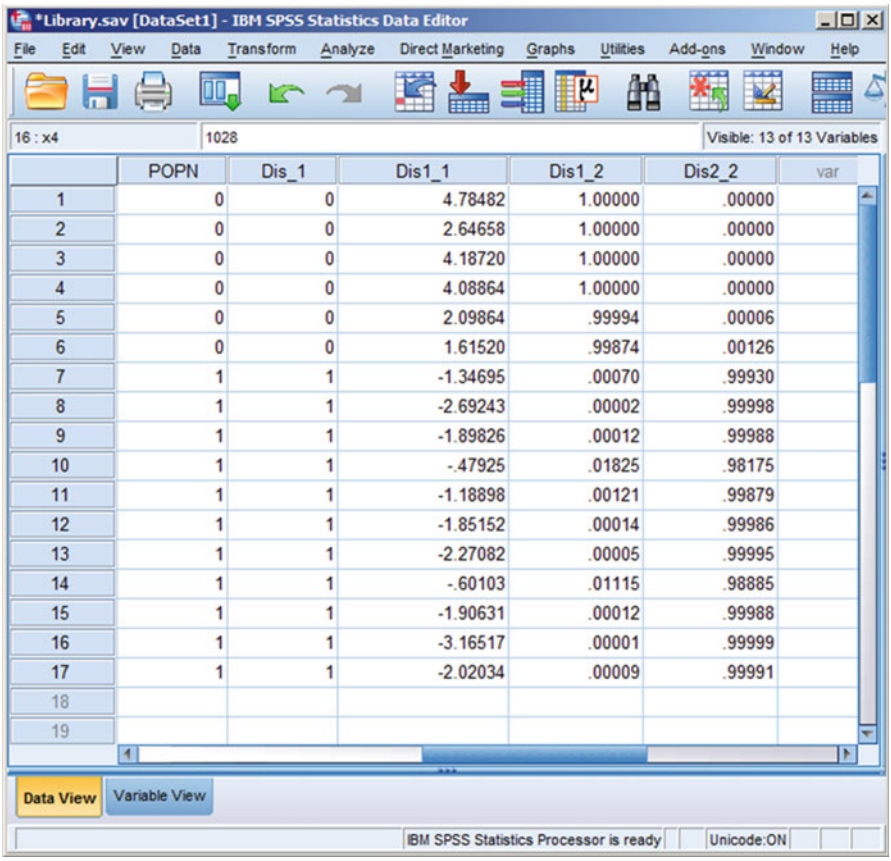


Fig. 6.7 The discriminant analysis: save dialogue box

misclassified. 6.5 and 6.6 present histograms of the discriminant function scores for the two population groups. (The mean discriminant scores are rounded to two decimal places here).

Clicking the Save button on the *Discriminant Analysis dialogue box* of Fig. 6.1 produces the *Discriminant Analysis: Save dialogue box* of Fig. 6.7. This permits the



	POPN	Dis_1	Dis1_1	Dis1_2	Dis2_2	var
1	0	0	4.78482	1.00000	.00000	
2	0	0	2.64658	1.00000	.00000	
3	0	0	4.18720	1.00000	.00000	
4	0	0	4.08864	1.00000	.00000	
5	0	0	2.09864	.99994	.00006	
6	0	0	1.61520	.99874	.00126	
7	1	1	-1.34695	.00070	.99930	
8	1	1	-2.69243	.00002	.99998	
9	1	1	-1.89826	.00012	.99988	
10	1	1	-.47925	.01825	.98175	
11	1	1	-1.18898	.00121	.99879	
12	1	1	-1.85152	.00014	.99986	
13	1	1	-2.27082	.00005	.99995	
14	1	1	-.60103	.01115	.98885	
15	1	1	-1.90631	.00012	.99988	
16	1	1	-3.16517	.00001	.99999	
17	1	1	-2.02034	.00009	.99991	
18						
19						

Fig. 6.8 Results of discriminant analysis added to the working file

user to save the predicted group membership (IBM SPSS default name DIS_1), the discriminant scores (DIS1_1) and the probabilities of group membership in the active file, as shown in Fig. 6.8. The probabilities of group membership have the default IBM SPSS variable names DIS1_2 for the first group (POPN =0) and DIS2_2 for the second group (POPN =1). Of course these latter probabilities must sum to one.

Chapter 7
Multidimension Scaling (MDS)

Suppose you had a map of the locations of several towns. It would be a simple matter to construct a table (or matrix) of distances between them. Now consider the reverse problem, where you are given the matrix of distances between the towns and are asked to reproduce the map. Geometric techniques are available for this purpose, but considerably more effort would be needed. Essentially, MDS is a method for solving this reverse problem. However, typical applications of MDS are more complex than this simple problem would suggest. Firstly, data usually contain error or noise. Secondly, it is seldom known in advance whether a two-dimensional map will suffice or whether a map using three, four or even more dimensions is required.

Generally, MDS is a set of mathematical techniques that enables the researcher to reveal the "hidden structure" or "underlying dimensionality" in a data set. For example, in one well known application of MDS, respondents in a national survey were asked to evaluate actual or potential candidates for the U.S. presidency. The respondents were not directed as to the criteria by which they should make their judgements; it was left to each respondent to make their rating based on any factors they wished.

How similarly did the respondents view the candidates? What identifiable features could be discerned in these evaluations of the candidates? Can we understand what led respondents to make their decisions? MDS helped answer these questions by plotting the political candidates on a two dimensional map. The closer were candidates in this spatial representation, the more similar they were perceived by the respondents and vice versa. The two axes (called *dimensions*) of the diagram were labelled 'partisanship' and 'ideology' by the researchers. By finding key differences between political candidates at opposite ends of each dimension, the researchers could attempt to develop indicators of variables that could be measures in future elections.

MDS uses *proximities* between the study objects as input. A proximity is a number which indicates how similar or dissimilar two objects are thought to be. All the proximities in a study constitute a (dis)similarity matrix. The chief output of

A. Aljandali, *Multivariate Methods and Forecasting with IBM® SPSS® Statistics*,
Statistics and Econometrics for Finance, DOI 10.1007/978-3-319-56481-4_7

MDS is a map as mentioned above, consisting of a geometric representation of the study objects. Each point on the map corresponds to one of the study objects. The geometric structure of the map reflects the "hidden structure" in the data and usually makes the data easier to comprehend. The larger the dissimilarity (or the smaller the similarity) between two study objects as shown by their proximity value, the further apart will be these objects on the spatial map.

As an example of the spatial map generated by MDS, consider Fig. 7.1. Five hypothetical products are involved, a, b, c, d and e. Proximity measures were obtained by asking respondents to rank each pair of products in terms of how similar they are perceived to be. There being five products, there are ten (n[n − 1]/2 = 5.4/2 = 10) pairs, so the ranks are from 1 (say 'most familiar') to 10 ('most dissimilar'). In marketing applications of MDS, such a ranking task is often performed by asking respondents to sort cards on which each pair is marked.

Figure 7.1 relates to one consumer's rankings. The two dimensions are the "underlying dimensions" that the respondent is deemed to have used to perform the ranking exercise. Let us assume that the researcher is able to label dimension 1 as say 'price' and dimension 2 as 'quality'. (The method by which the dimensions are labelled is discussed later). Products a and c are perceived similarly in terms of 'price' in that they have high, similar scores on this dimension. However, the latter two products are perceived quite differently in terms of 'quality'. None of the five

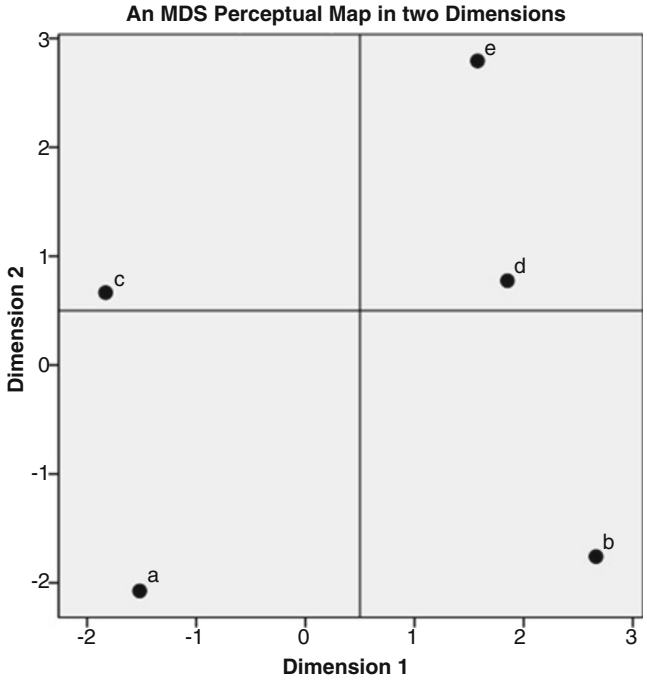


Fig. 7.1 A hypothetical MDS perceptual map

products are close to each other in Fig. 7.1. Therefore, this respondent perceives the products quite differently in terms of these two factors. Should any of the products have been close to each other on the spatial map, then they could be regarded as substitutes. The number of pertinent dimensions to be used in a particular study may be known to the market researcher from past experience. There are statistics produced in MDS and which may be used to compare models of competing dimensionality.

The main thrust of MDS in the field of marketing has been to examine relationships between brands of particular product group. Specifically, MDS has been used to derive (i) consumer perceptions of the similarity of brands and (ii) consumer preferences for brands. In this sense, MDS is an extension of the one-dimensional attitude scales like the semantic differential. Instead of positioning attitudes about brands on one-dimensional scales, MDS positions brands in an n-dimensional space, where n is the minimum underlying dimensionality of the relationship.

About 34% of marketing-orientated businesses in the U.S.A. use MDS. Research has shown that business applications of the technique have covered:

- identification of the salient product attributes perceived by buyers,
- the combinations of product attributes most preferred,
- the products that are viewed as substitutes for each other and those that are differentiated from each other,
- the viable segments that exist in a market and
- those "holes" in the market that can support a new product venture.

It has been suggested that potential marketing applications would involve product life-cycle analysis, market segmentation, vendor evaluation, advertising evaluation, test marketing, salesperson/store image and brand switching research.

7.1 Types of MDS Model and Rationale of MDS

MDS may use just one *matrix of proximities* as input. For example, if we have judgements from one consumer about the (dis)similarity between pairs of bottled beers, we have one matrix. On the other hand, if we have judgements from many car drivers about the dis(similarities) between pairs of automobiles, then we have many proximity matrices – one for each driver. If the proximities are measured on a nominal (rare) or ordinal scale (for example, ranks), we have *nonmetric MDS models*. If the proximities involve ratio or interval measurement we have metric MDS models. The spatial map that is derived from MDS is usually two-or-three-dimensional Euclidean space, but it may have more dimensions.

The mathematics underlying MDS is highly complex and varies across different types of MDS model. However, brief aspects may be discussed here. The observed proximity between object i and object j is denoted by δ_{ij}. In order to generate the spatial map, these observed proximities are transformed into scaled distances,

denoted by d_{ij}. The magnitude of the d_{ij} should reflect the magnitude of the observed δ_{ij}. We may say that:

$$d_{ij} = f(\delta_{ij}),$$

Where f is a function of some specified type and is discussed below. The discrepancy between $f(\delta_{ij})$ and d_{ij} is simply:

$$f(\delta_{ij}) - d_{ij}$$

The sum of the squares of these discrepancies is:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} [f(\delta_{ij}) - d_{ij}]2$$

Next, we divide by a scale factor so as to measure the squared discrepancies relative to a sensible measuring stick. The scale factor most commonly used is:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}2$$

Finally, the square root is taken of the result and is called the *f-stress*. The larger is the f-stress, the worse the scaling process reflects the original observed proximities. If f-stress equals zero, then $f(\delta_{ij}) = d_{ij}$. It should be noted that other stress measures are referred to later on, especially one called *S-stress*. They tend to differ only if the nature of the scale factor used.

The essential feature of MDS is that we seek that function f which generates scaled differences, but which has minimum f-stress over all possible functions. Methods for deriving the function f vary according to the type of MDS model being employed.

7.2 Methods for Obtaining Proximities

In marketing applications of MDS, a simple method for deriving proximity measures is to have consumers sort the study objects accordingly to perceived similarity. The typical instruction is to place the study objects into mutually exclusive categories so that objects in the same category are more similar to each other than those in other categories. A matrix of proximities among the objects can be derived for the consumer group, by counting the number of times each pair of study objects is placed into the same category. Similarly, ranking the degrees of similarity between pairs of study objects is common in marketing, as per the hypothetical example presented in the introduction.

A very common way to elicit proximities from data that are not proximities (i.e. inappropriate for MDS in their original form) is to compute some measure of (dis)similarity between the rows (or columns) of a table. For example, the rows of

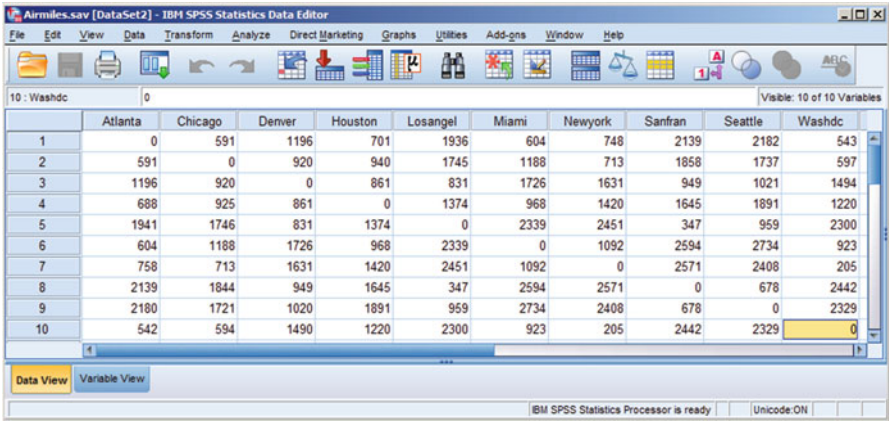
the original table might be various countries and the columns could be measures such as GNP, energy consumption or unemployment. The most common way to derive proximities is to compute correlations between the countries. Sometimes the proximities could be frequencies, like the number of telephone calls between cities, travel volume or any other form of transaction.

7.3 The Basics of MDS in IBM SPSS Statistics: Flying Mileages

An example commonly used to illustrate metric MDS is the flying mileages between Ten American Cities. The relevant data are shown in Fig. 7.2 (AIRMILES.SAV). For example, the flying distance between Atlanta and Denver is 1196 miles. The cities are the study objects and the mileages are the proximities. Note that all the diagonal elements are zero. Figure 7.3 shows the IBM SPSS *Multidimensional Scaling dialogue box* which accesses the MDS procedure. This dialogue box is accessed via:

```
Analyze
   Scale
     Multidimensional scaling (ALSCAL)...
```

The ten cities are the 'Variables' under study. Here, the data are already in the form of a square, symmetric proximity matrix. In such a square, symmetric matrix, the rows and columns represent the same items – here the ten cities. By clicking the Model. . .and Options. . .buttons, we obtain the *MDS: model dialogue box* and the *MDS: options dialogue box* of Figs. 7.4 and 7.5 respectively.



	Atlanta	Chicago	Denver	Houston	Losangel	Miami	Newyork	Sanfran	Seattle	Washdc
1	0	591	1196	701	1936	604	748	2139	2182	543
2	591	0	920	940	1745	1188	713	1858	1737	597
3	1196	920	0	861	831	1726	1631	949	1021	1494
4	688	925	861	0	1374	968	1420	1645	1891	1220
5	1941	1746	831	1374	0	2339	2451	347	959	2300
6	604	1188	1726	968	2339	0	1092	2594	2734	923
7	758	713	1631	1420	2451	1092	0	2571	2408	205
8	2139	1844	949	1645	347	2594	2571	0	678	2442
9	2180	1721	1020	1891	959	2734	2408	678	0	2329
10	542	594	1490	1220	2300	923	205	2442	2329	0

Fig. 7.2 Airmiles data

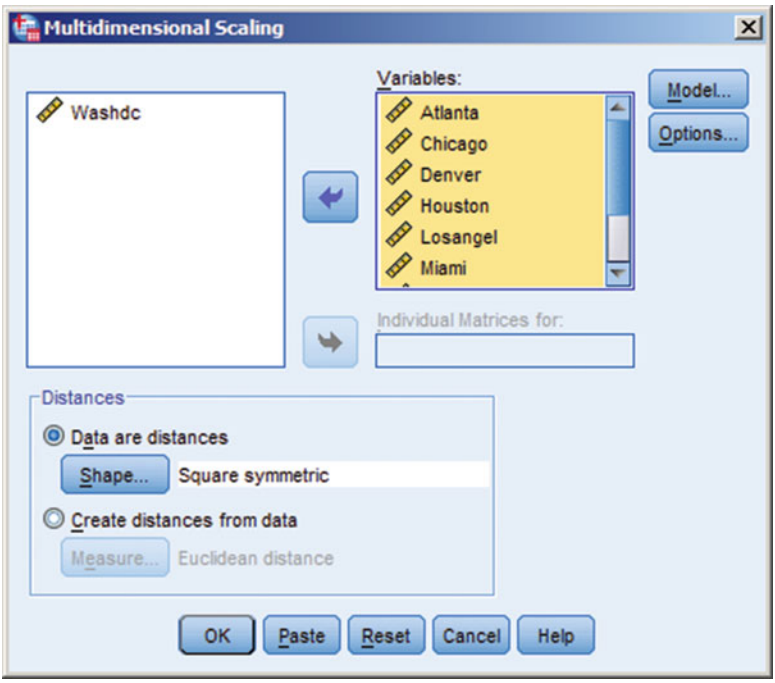


Fig. 7.3 The MDS dialogue box: data format

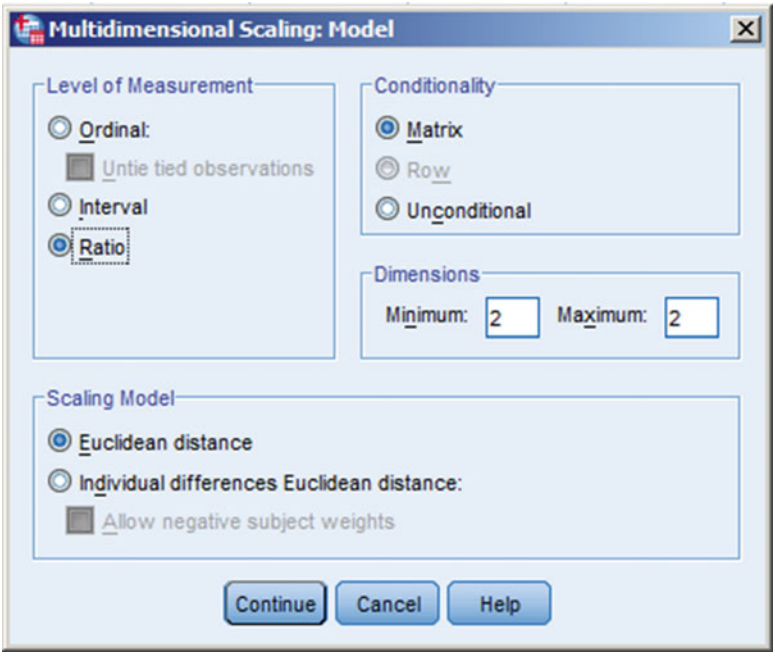


Fig. 7.4 The MDS: model dialogue box

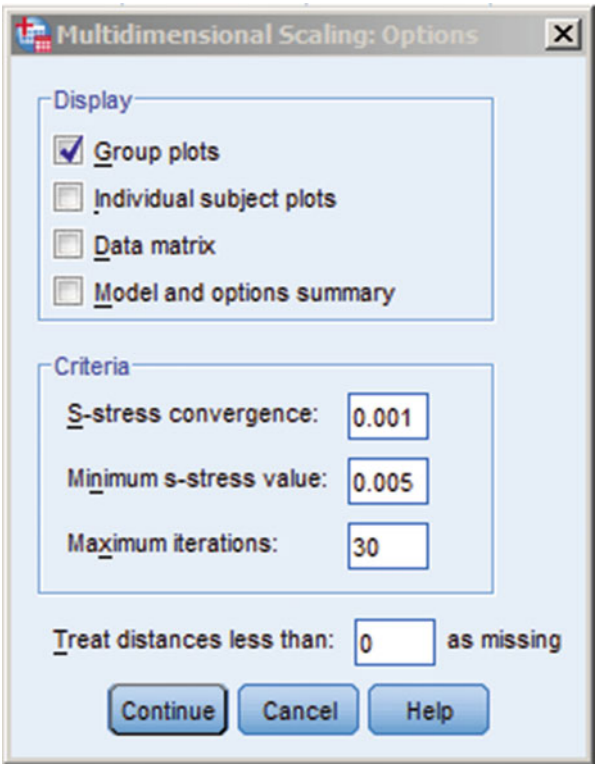


Fig. 7.5 The MDS: options dialogue box

The level of measurement in the current example is ratio (as selected in Fig. 7.4); the default is ordinal. In the dialogue box of Fig. 7.5, we have requested group plots, which are presented in Figs. 7.6 and 7.7.

Figures 7.6 and 7.7 were generated in IBM SPSS because we selected group plots in the dialogue box of Fig. 7.5. In Fig. 7.6, cities that are similar (i.e. short flying distances, here) are represented by points that are close together, with the converse being true. The orientation of Fig. 7.6 is arbitrary. The central point about MDS is that the distances between the points in Fig. 7.6 (referred as d_{ij}, representing the distance between cities i and j) should correspond to the proximities (referred to as δ_{ij}, representing here the air miles between cities I and j). A good way to examine this correspondence is via scatterplot between d_{ij} and δ_{ij}. This is produced in Fig. 7.8 and is called *Shepherd diagram*. Each point in Fig. 7.8 corresponds to one pair (i,j) of study objects. The horizontal axis contains the δ_{ij}, which have been standardised, so their units have changed.

The (standardised) d_{ij} are plotted on the vertical axis. The standardisation used in this Shepherd diagram generates a zero vertical axis intercept. Such a good, clean-cut pattern as Fig. 7.8 is uncommon in practice. Indeed, Fig. 7.8 exhibits a virtually perfect fit i.e. no scatter because the data have essentially no error, because we have

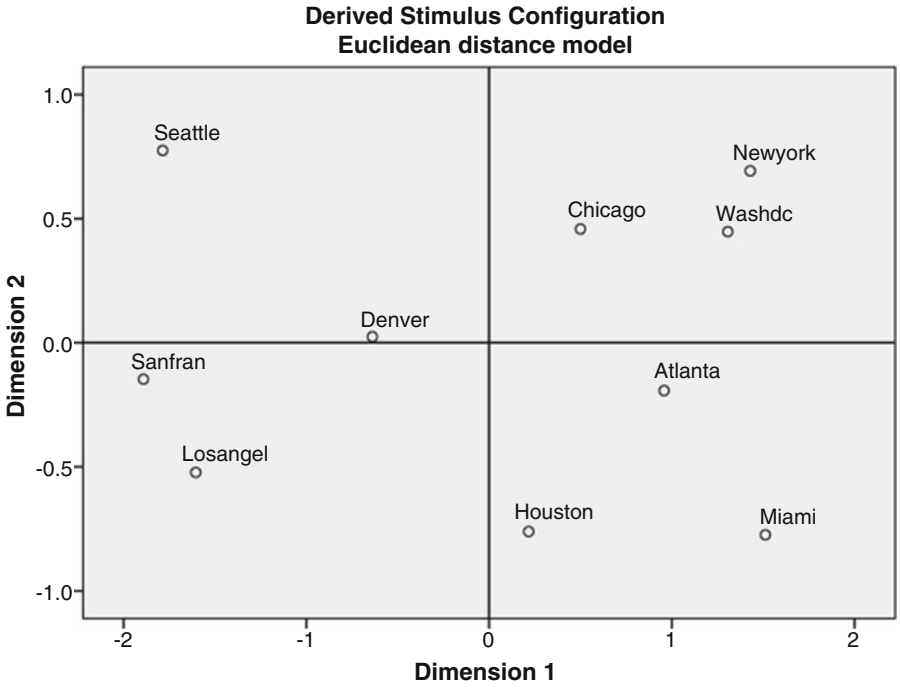


Fig. 7.6 MDS plot of intercity flying mileages

properly assumed that the data are at the ratio level of measurement and we have correctly used an MDS model with two and only two relevant dimensions.

It remains to label the two axes in Fig. 7.6. This is a subjective matter. The most common way of interpreting the dimensions in such as Fig. 7.8 is to look at the study objects opposite extremes of the two axes. In this simple example, the vertical axis would be labelled 'north-south' as Seattle and Miami are the most northerly and southerly towns in the analysis. Similarly, the horizontal axis would be labelled 'west-east'. Figure 7.7 presents the non-graphical output from the IBM SPSS Statistics multidimensional scaling routine.

This MDS run terminated after one iteration. IBM SPSS Statistics reports *Young's S-stress statistic* as 0.00469. Like f-stress, S-stress is a measure of fit ranging from 1 (worst possible fit) and 0 (perfect fit). This statistic measures the correspondence between the squared distances δ_{ij}^2 and the squared scaled distances d_{ij}^2 produced by MDS. Kruskal's stress index of 0.00443 measures the correspondence between the just δ_{ij} and the d_{ij} and again we seek values close to zero. The squared correlation (RSQ) is the coefficient of determination between the δ_{ij} and the d_{ij} and from Fig. 7.7 it was apparent that an RSQ value close to unity would result.

The following rule of thumb has been offered for the interpretation of the value of S-stress:

Iteration history for the 2 dimensional solution (in squared distances)

Young's S-stress formula 1 is used.

Iteration	S-stress	Improvement
1	.00469	

Iterations stopped because
S-stress is less than .005000

Stress and squared correlation (RSQ) in distances

RSQ values are the proportion of variance of the scaled data (disparities) in the partition (row, matrix, or entire data) which is accounted for by their corresponding distances. Stress values are Kruskal's stress formula 1.

For matrix
Stress = .00443 RSQ = .99993

Configuration derived in 2 dimensions

Stimulus Coordinates

Stimulus Number	Stimulus Name	Dimension 1	Dimension 2
1	Atlanta	.9574	-.1929
2	Chicago	.5002	.4584
3	Denver	-.6384	.0241
4	Houston	.2167	-.7607
5	Losangel	-1.6049	-.5224
6	Miami	1.5118	-.7741
7	Newyork	1.4290	.6923
8	Sanfran	-1.8916	-.1469
9	Seattle	-1.7863	.7749
10	Washdc	1.3061	.4474

Fig. 7.7 IBM SPSS statistics output for the airmiles data (AIRMILES.SAV)

- If S-stress $\leq .25$, very good fit
- If $.25 <$ S-stress $\leq .5$, good fit,
- If $.05 <$ S-stress $\leq .10$, relatively good fit and
- If S-stress $> .1$, poor fit.

It might be noted that if we used MDS not on the actual flying miles in Fig. 8.2, but rather on the ranks of these mileages, then we would be performing a non-metric MDS.

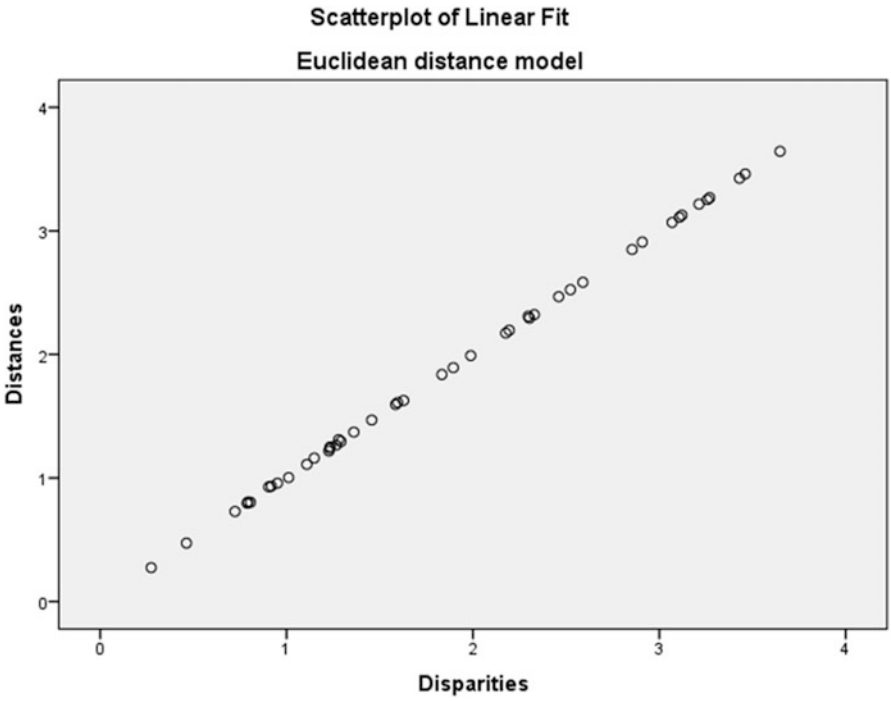


Fig. 7.8 Scatterplot of raw data versus distances

7.4 An Example of Nonmetric MDS in IBM SPSS Statistics: Perceptions of Car Models

An example of nonmetric MDS often cited in marketing is the measurement of consumers' perceptions of similarity and their preferences for 11 competing American car models. Consumers were asked to rank order the degrees of similarity between all 55 (n[n − 1]/2 = 11*10/2 = 55) combination of cars. The combinations were presented on cards. The data are in the file CARS.SAV.

Being ranks, the proximity data are ordinal, so this option must be selected from the *MDS: Model dialogue box*. The results for one particular consumer are presented in Figs. 7.9 and 7.10. By examining the location of the car makes relative to the axes, dimension 1 was labelled 'high luxuriousness – low luxuriousness' from left to right. Dimension 2 was labelled 'high sportiness-low sportiness' from top to bottom of this axis. This spatial map represents the perceptual space of this particular consumer. The positioning of car makes relative to each seems to yield competitive segments, for example, Mercedes and BMW are similarly perceived by this individual. The stress and RSQ measures in Fig. 7.10 are indicative of the goodness of fit derived from this model. In studies such as this, it is common to include an extra hypothetical product into the study objects. Here, it would be "my

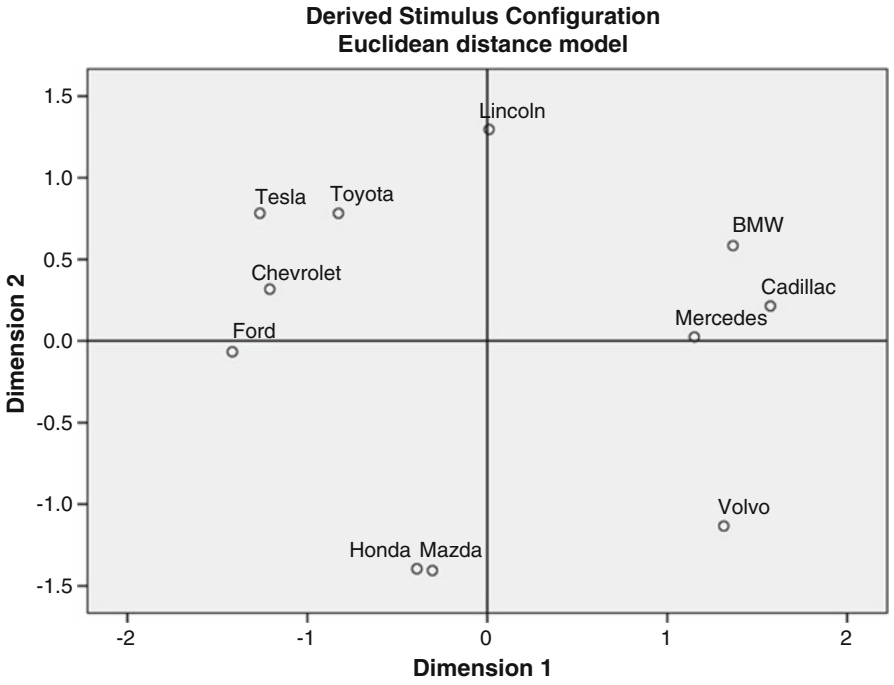


Fig. 7.9 MDS map for a consumer's perceptions of car makes

ideal car". These are called ideal points in marketing and would be included in the ranking procedure. For example, a consumer who likes big, luxury cars would probably have this ideal point near to the Lincoln and Cadillac. When each consumer's ideal point is positioned in space, the market researcher can look for clustering of ideal points which can be used to predict market shares.

7.5 Methods of Computing Proximities

There are several competing methods for computing proximities according to whether the data are interval, frequencies or binary. Remember that proximity is a measure of how similar or dissimilar two objects are. The default for interval data is the Euclidean distance as used with the airlines example. In the *MDS: Model dialogue box* there is the option to create measures from the data. Selecting this option generates the *MDS: Create Measure dialogue box* of Fig. 7.11.

Suppose a study involves two variables X and Y with the following values:

$$\begin{array}{lccccc} X_i: & 4 & 7 & 9 & 4 & 3 \\ Y_i: & 3 & 7 & 12 & 3 & 5 \end{array}$$

Iteration history for the 2 dimensional solution (in squared distances)

Young's S-stress formula 1 is used.

Iteration	S-stress	Improvement
1	.43329	
2	.38047	.05283
3	.35926	.02121
4	.35171	.00755
5	.34759	.00413
6	.34476	.00283
7	.34263	.00213
8	.34099	.00165
9	.33979	.00120
10	.33938	.00041

Iterations stopped because
S-stress improvement is less than .001000

Stress and squared correlation (RSQ) in distances

RSQ values are the proportion of variance of the scaled data (disparities) in the partition (row, matrix, or entire data) which is accounted for by their corresponding distances. Stress values are Kruskal's stress formula 1.

For matrix
Stress = .26623 RSQ = .54648

Stimulus Coordinates

Stimulus Number	Stimulus Name	Dimension 1	2
1	Toyota	-.8272	.7825
2	Ford	-1.4174	-.0664
3	Honda	-.3912	-1.3957
4	Chevrole	-1.2091	.3175
5	Tesla	-1.2639	.7827
6	Mercedes	1.1508	.0248
7	Volvo	1.3144	-1.1339
8	Cadillac	1.5733	.2141
9	BMW	1.3654	.5842
10	Lincoln	.0099	1.2964
11	Mazda	-.3051	-1.4062

Fig. 7.10 Output for MDS of car make similarities

The following proximity measures are available for interval data:

(a) *Euclidean distance* which is the default for interval data:

$$\text{Dist}\,(X, Y) = \left\{(4-3)^2 + (7-7)^2 + \ldots + (3-5)^2\right\}^{0.5} = (15)^{0.5} = 3.87$$

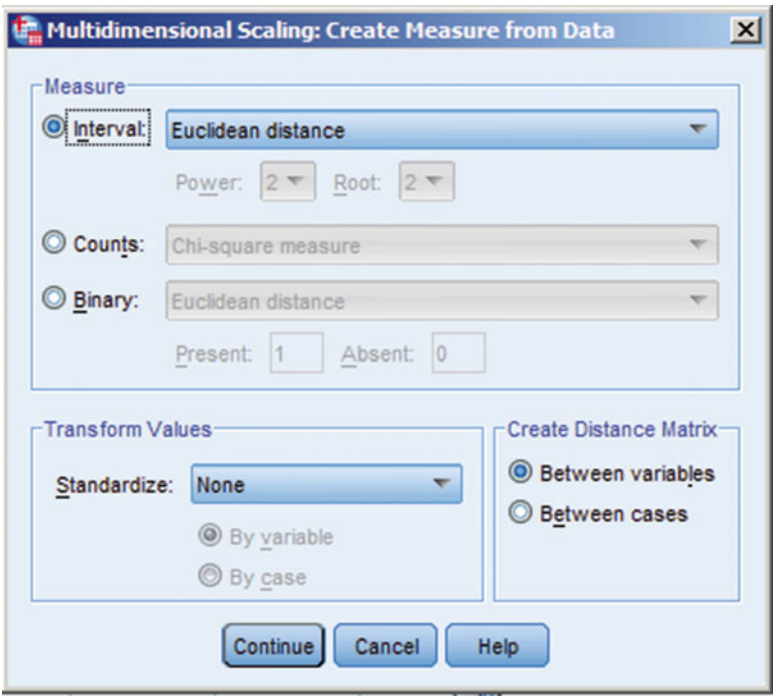


Fig. 7.11 The MDS: create measure dialogue box

(b) *Squared Euclidean distance:*

$$Dist(X, Y) = (4-3)^2 + (7-7)^2 + \cdots + (3-5)^2 = 15$$

(c) *City block or Manhattan distance*, which uses absolute values:

$$\text{Dist}\ (\text{X}, \text{Y}) = |\ 4-3\ | + |\ 7-7\ | + \ldots + |\ 3-5\ | = 7$$

Figure 7.12 is an MDS map of the inter-city airmiles data employing the city block distance measure. The map naturally suggests that this is a totally inappropriate way of conceiving of distance in this instance.

(a) The *Chebychev distance measure* is simply the maximum absolute difference between the values for the items:

$$\text{Dist}(\text{X}, \text{Y}) = |\ 9-12\ | = 3$$

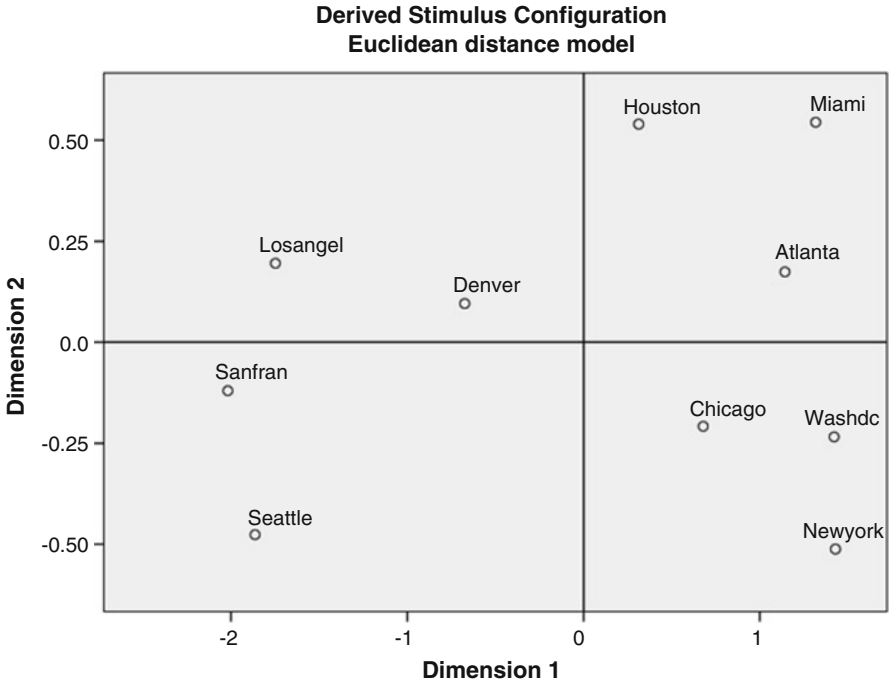


Fig. 7.12 MDS plot of intercity flying mileages using Manhattan distances

(b) The Minkowski distance measure between two items in the pth. root of the sum of the absolute differences to the pth. power between the values for the items:

$$\text{Dist}(\text{X}, \text{Y}) = \left\{ \sum\nolimits_i |X_i - Y_i|^p \right\}^{1-p}$$

And the user may specify a value for p. The default is p = 2.

7.6 Weighted Multidimensional Scaling in IBM SPSS, INDSCAL

The analyses so far have involved one matrix of proximities. This simplest form of MDS is called classical multidimensional scaling (CMDS). CMDS was extended to allow for more than one matrix of proximities. This is naturally important, as researchers will tend to have more than one such matrix, especially in the marketing context, where panels of consumers are questioned about similarities. Replicated multidimensional scaling (RMDS) was the first development that permitted more than one matrix. However, RMDS is of limited use in the marketing context,

because it assumes that all data matrices are the same, save for error. The matrices are replicates of each other with no systematic differences. Most market research studies assume and search for interpersonal differences in consumer attitudes and cognition. Under WMDS, several matrices can be assumed to differ from each other without the assumption that such differences are solely due to error. It should be noted that the WMDS model is also known as *INDSCAL (individual scaling Euclidean distance model)*.

Suppose sampled consumers were individually asked to judge how similar were pairs of competing retail stores were similar. There is a total of 15 stores in the study. The name of each was written on a card. The experimenter selected one store – called the standard – and asked the respondent to select that store from the remaining 14 stores which was most similar to the standard. The selected store is removed and the respondent now asked which of the remaining 13 stores is most similar to the standard. This method is repeated until all the 14 non-standard stores were exhausted. The store that is deemed most similar to the standard is ranked as 1, the next most similar as 2 etc.

Ideally, this task should be repeated until all stores have acted as the standard, but in a market research survey, this may not be possible. The standard is conventionally entered as a zero in the data matrix for WMDS. If the task has been completed will all 15 stores acting as the standard, the input data matrix for the first respondent would be like:

	Shop 1	Shop2	Shop3		Shop14	Shop15
Comparison 1	0	2	1	. . .	10	11
Comparison 2	7	5	8	. . .	0	2
	. . .	. . .	. . .	. . .	. . .	. . .
	. . .	. . .	. . .	. . .	. . .	. . .
Comparison 15	1	3	0	. . .	9	10

At the first comparison, store number 1 was the standard and store 3 was perceived as being most similar to it. Store 2 was the next most similar. At the second comparison, store 14 acted as the standard.

Suppose (for simplicity) that three consumers have performed the above task with all 15 stores acting as the standard. The data input data matrix has 45 rows (3 respondents X 15 standards) and 15 columns (no. of stores). Our data are no longer square symmetric as in previous analyses. In the *MDS dialogue box* of Fig. 7.3, we turn to the 'Distances' box and the option:

- Data are distances

And by clicking the Shape. . .button, we derive the Multidimensional Scaling: Shape of Data dialogue box of Fig. 7.13. Here, we choose:

- Rectangular

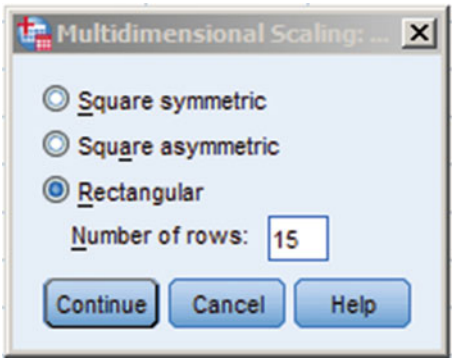


Fig. 7.13 The multidimensional scaling: shape of data dialogue box

And the number of rows per respondent is 15. IBM SPSS Statistics now knows that with 45 rows, there are three respondents.

Next we turn to the *MDS: Model dialogue box* of Fig. 7.4. Our ranked data are ordinal. Under the heading 'Scaling Model' box of Fig. 7.4, INDSCAL is accessed by choosing the option:

- Individual Differences Euclidean Model

We now turn the 'Conditionality' box of Fig. 7.4. When the values in a row are ranked only relative to other values in the same row, we say that the data are *now conditional*. We thus select:

- Row

Suppose that the marketing practitioner is satisfied that consumers perceive similarities (and dissimilarities) between these stores in terms of three dimensions only (say, 'price', 'quality' and 'range of choice'). We select this option under the heading 'Dimensions' in the dialogue box. The *MDS: Model dialogue box* should now appear as in Fig. 7.14.

The MDS map would now be in three dimensions. Where INDSCAL differs from both classical and replicated MDS in that weights are produced for each subject. These weights measure the importance of each dimension to the subject. If an individual's weight on a particular dimension is large, near to its maximum of 1, then that dimension is relatively important; if the weight for a dimension is near its minimum of 0, then that dimension is relatively unimportant.

The *weirdness index* (WI) helps to interpret these weights. WI varies between 0 and 1 and indicates how unusual or (weird) each respondent's weights are relative to the mean weights of all respondents.

– A respondent with weights proportional to the mean weights has WI of zero. Such a respondent is a typical respondent.
– A respondent with one large weight and many low weights has a WI close to one. Such a respondent is dissimilar to the rest.

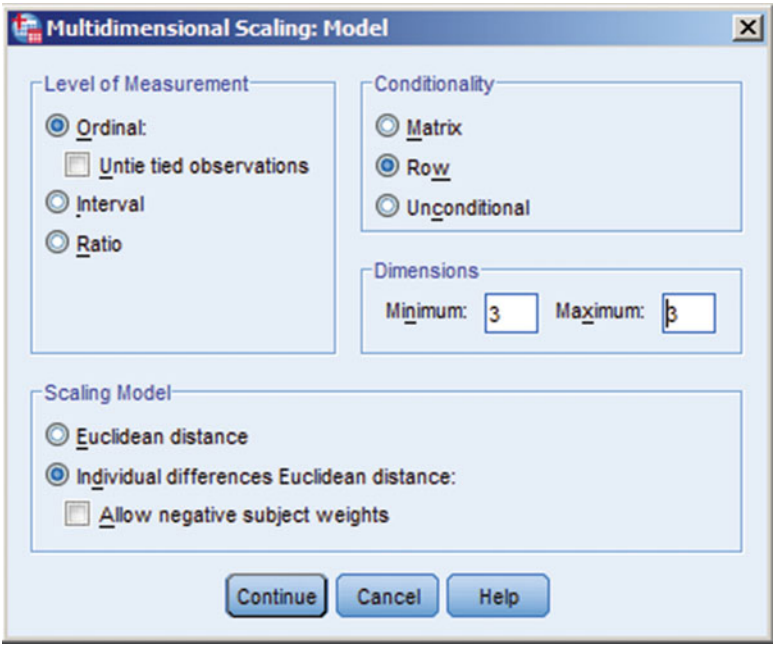


Fig. 7.14 The MDS: model dialogue box for the store perception data

Chapter 8
Hierchical Log-linear Analysis

We have already seen in Part 1 of this guide how to examine two-way classifications in contingency tables via the chi-square statistic. As further variables are added to cross-classification tables, the number of cells rapidly increases and it is difficult to unravel the associations between the variables by examining only the cell entries. Although the traditional chi-square approach often provides an insight about the relationships among variables, it:

- Confuses the relationship between a pair of categorical variables with the relationship when other variables are present,
- Does not allow for the simultaneous examination of these pairwise relationships and
- Ignores the possibility of three-variable and higher-order interactions among the variables.

Hierarchical log-linear analysis is deigned to solve these more complex problems.

8.1 The Logic and Terminology of Log-linear Analysis

One objective of log-linear analysis is to assess if variables are associated (or dependent) or whether they are not associated (or independent). For simplicity, the logic and terminology of log-linear analysis will be introduced in the context of only two variables. It should be appreciated, however, that the prime utility of log-linear analysis lies in the multi-variate context, which is illustrated later. The underlying logic of two-way contingency table analysis involves the multiplication law of probability under the null hypothesis that the two variables are independent. If two variables A and B are independent, then elementary probability theory states that P(A) and (b) = P(A).P(B). Log-linear analysis attempts to use the logic of multiple regression models which are additive. However, the additive property can

A. Aljandali, *Multivariate Methods and Forecasting with IBM® SPSS® Statistics*,
Statistics and Econometrics for Finance, DOI 10.1007/978-3-319-56481-4_8

be ascribed to the above probability law by taking natural logarithms (base e) i.e. lnP (A and B) = lnP(A) + lnP(B).

The frequency table overleaf was derived from a study that examined factors that influenced the amount of effort put into comparing alternatives by consumers in the retail environment. Two factors that might influence consumer effort are the degree of interest in the purchase (IBM SPSS variable name INTEREST) in the purchase and the level of risk (RISK) attached to the purchase by the consumer. Before considering the influences or otherwise of these two variables upon consumer effort, is it possible that they themselves are associated? For example, do consumers who regard this particular purchase of 'high interest', therefore also regard it as one of 'high risk'? Both variables in the above table have been coded as 1 for "high", 2 for "medium" and 3 for "low". A chi-square statistic of 3.0462, df = 4, p = 0.5501 indicates that we do not reject the hypothesis that INTEREST and RISK are themselves independent.

Interest				
Risk	1	2	3	Total
1	13	12	12	37
2	20	17	14	51
3	17	19	26	62
Total	50	48	52	150

Using log-linear model, the number of cases in each cell of the above table can be expressed as the sum of the "effects" of RISK and INTEREST and any interaction between the two. This is simply saying that the observed frequencies in any cell are due to scores ("effects") on the two variables and any interaction between them. To obtain a log-linear model, the natural logs of the above cell frequencies are used:

Interest				
Risk	1	2	3	Mean
1	2.565	2.485	2.485	2.512
2	2.998	2.833	2.639	2.823
3	2.833	2.944	3.258	3.012
Total	2.799	2.754	2.794	2.782

In general for a two-way situation, the log of the observed frequency in the ith row and the jth column is given by:

$$\mathrm{LnF_{ij}} = \mu + \vartheta_i^A + \vartheta_j^B + \vartheta_{ij}^{AB} \ldots \tag{8.1}$$

Where $\mathrm{F_{ij}}$ are the observed frequencies in the cell, ϑ_i^A is the effect of the ith category of variable A, ϑ_j^B is the effect of the jth category of variable B and ϑ_{ij}^{AB} is the interaction effect for the ith category of variable A and the jth category of variable B. The quantity μ is the grand mean of the logs of all the observed

frequencies. Equation 8.1 is called the *saturated model*, in that it contains all single variable effects and all possible interactions.

Using Eq. 8.1, the log-linear model for the first cell (top left hand corner) is:

$$\text{Ln}(13) = \mu + \vartheta^{RISK=1} + \vartheta^{INTEREST=1} + \vartheta^{RISK=1}_{INTEREST=1} \cdots \tag{8.2}$$

Thirteen is the observed frequency in this cell. The term μ is the grand mean of the entire table, here 2.782. The ϑ parameters represent the increments or decrements from μ for particular combinations of values of row and column variables. Each category of the row and column variables has an associated ϑ. $\vartheta^{RISK=1}$ indicates the "effect" of being in the first (or "low") category of risk. This effect is computed as:

$$\begin{aligned}\vartheta^{(RISK=1)} &= \text{mean of logs in the "RISK} = 1\text{" cells} - \text{grand mean} \\ &= 2.512 - 2.782 = -.270\end{aligned}$$

The ϑ parameter for one category of a variable is just the mean log of the frequencies in a particular category minus the grand mean, so another example:

$$\begin{aligned}\vartheta^{(INTEREST=1)} &= \text{mean of the logs in the "INTEREST} = 1\text{" cells} - \text{grand mean} \\ &= 2.799 - 2.782 = .017\end{aligned}$$

Positive values of ϑ occur when the mean number of cases in a row or column is larger than the overall grand mean.

Consider the interaction effect in Eq. 8.1. Rearranging:

$$\begin{aligned}\vartheta^{RISK=1}_{INTEREST=1} &= \text{Ln}(\text{F}_{\text{ij}}) - (\mu + \vartheta^{\text{RISK}=1} + \vartheta^{\text{INTEREST}=1}), \\ &= \ln(13) \times -(2.782 - 0.270 + 0.017) = 0.036\end{aligned}$$

Hence, it is possible to compute the ϑ parameters for the ain effects and their interactions:

Main effects:

$$\begin{aligned}\vartheta^{RISK=1} &= 2.512 - 2.782 = -0.270 \\ \vartheta^{RISK=2} &= 2.823 - 2.782 = 0.041 \\ \vartheta^{RISK=3} &= 30.12 - 2.782 = 0.230 \\ \vartheta^{INTEREST=1} &= 2.799 - 2.782 = 0.017 \\ \vartheta^{INTEREST=2} &= 2.754 - 2.782 = -0.028 \\ \vartheta^{INTEREST=3} &= 2.794 - 2.782 = 0.012\end{aligned}$$

Interaction effects:

$$\vartheta^{RISK=1}_{INTEREST=1} = 2.565-(2.782-0.270+0.017) = 0.036$$
$$\vartheta^{RISK=1}_{INTEREST=2} = 2.485-(2.872-0.270-0.028) = 0.001$$
$$\vartheta^{RISK=1}_{INTEREST=3} = 2.485-(2.782-0.0270+0.012) = -0.039$$
$$\vartheta^{RISK=2}_{INTEREST=1} = 2.998-(2.782+0.041+0.017) = 0.158$$
$$\vartheta^{RISK=2}_{INTEREST=2} = 2.833-(2.782+0.041-0.028) = 0.038$$
$$\vartheta^{RISK=2}_{INTEREST=3} = 2.639-(2.782+0.041+0.012) = -0.196$$
$$\vartheta^{RISK=3}_{INTEREST=1} = 2.833-(2.782+0.230+0.017) = -0.196$$
$$\vartheta^{RISK=3}_{INTEREST=2} = 2.944-(2.782+0.0230-0.028) = -0.040$$
$$\vartheta^{RISK=3}_{INTEREST=3} = 3.258-(2.782+0.230+0.012) = 0.234$$

It will be found that knowing some of the ϑ, we automatically know others. For example, allowing for decimal rounding error:

$$\vartheta^{RISK=1} + \vartheta^{RISK=2} + \vartheta^{RISK=3} = 0$$

Similarly for the categories of INTEREST.

Also for the interactions, there are zero sums, such as:

$\sum \vartheta^{RISK=1}$ for INTEREST = 1, 2 and 3 is zero and
$\sum \vartheta^{INTEREST=2}$ for RISK = 1, 2 and 3 is zero.

IBM SPSS Statistics reports the minimum number of ϑ values that are sufficient to derive the rest.

8.2 IBM SPSS Statistics Commands for the Saturated Model

I shall use the data in the contingency table of the previous section, involving the variables INTEREST and RISK to illustrate aspects of log-linear analysis in IBM SPSS. The data are to be found in the IBM SPSS file CONSUMER.SAV. It is necessary to understand application and interpretation of the saturated model or Eq. 8.1, before performing more advanced, multi-variate log-linear analyses. The IBM SPSS hierarchical loginear procedure is accessed by clicking:

```
Analyse
     Loglinear
        Model Selection...
```

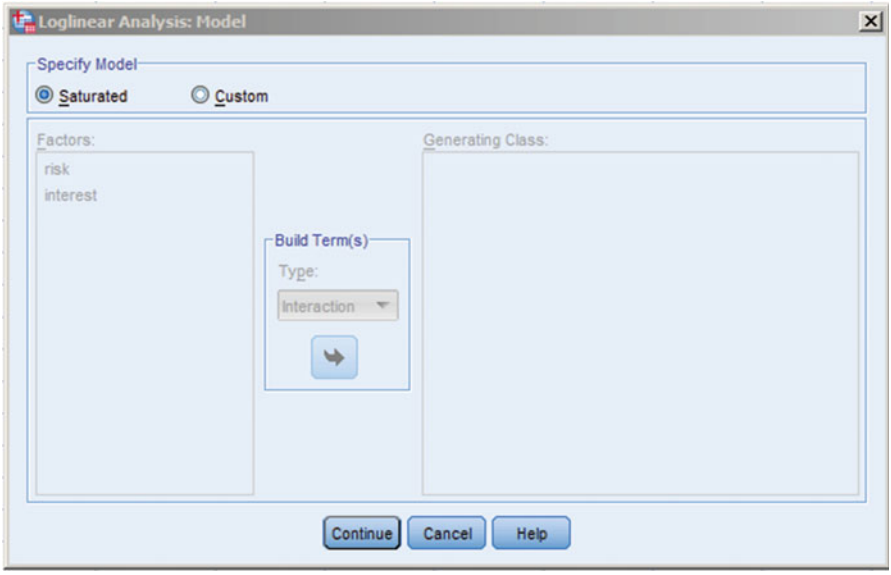


Fig. 8.1 The loglinear analysis: model dialogue box

Which produces the *Model Selection Loglinear Analysis dialogue box* of 8.1. The 'Factors' in this analysis are just the two variables RISK and INTEREST. The Define Ranges button will become active and must be clicked. It is necessary to define the minimum code (here '1') and the maximum code (here '3') employed for the factors. Under the heading 'Model Building', the main effects and interaction terms are entered into the model in a single step when the saturated model is being analysed. Clicking the Model button generates the *Loglinear Analysis: Model dialogue box* of Fig. 8.1 and it will be seen that the saturated model is the default. Clicking the Options button on the dialogue box of Fig. 8.2 produces the *Loglinear Analysis: Options dialogue box* of Fig. 8.3. In this latter dialogue box, I have requested the frequencies and residuals to be displayed. Under the heading 'Display for Saturated Model', I have requested the parameter estimates, whose numerical values were computed at the bottom of page 142. Also in this dialogue box and under the heading 'Model Criteria' I have set the value of a quantity called "*delta*" to be zero, rather than its default value of 0.5. There are problems in all forms of contingency table analysis if any of the cells contain zero frequencies (e.g. ln0 does not exist). Therefore and by default, IBM SPSS Statistics adds 0.5 (and calls this quantity "delta") to each cell frequency in case this problem arises. However, we know here that this will be unnecessary, hence the user's recalibration of delta in this instance.

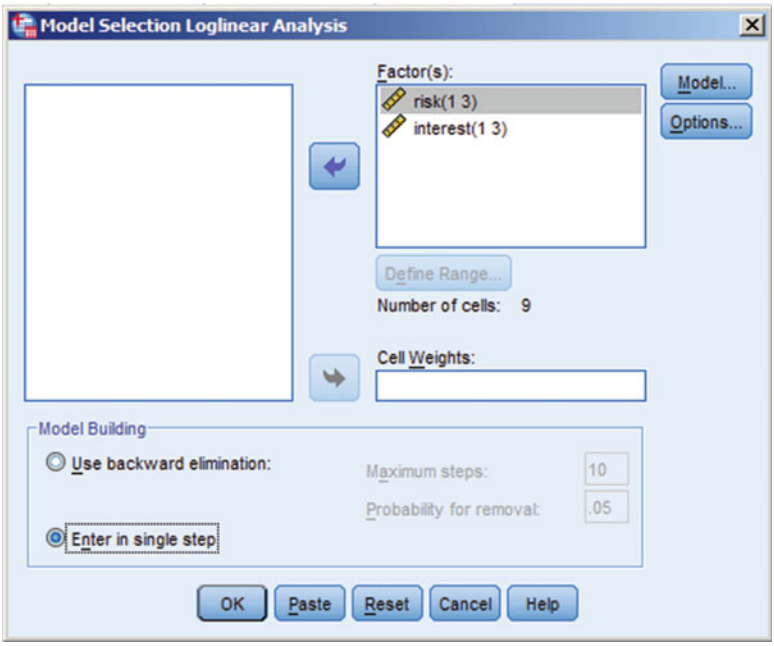
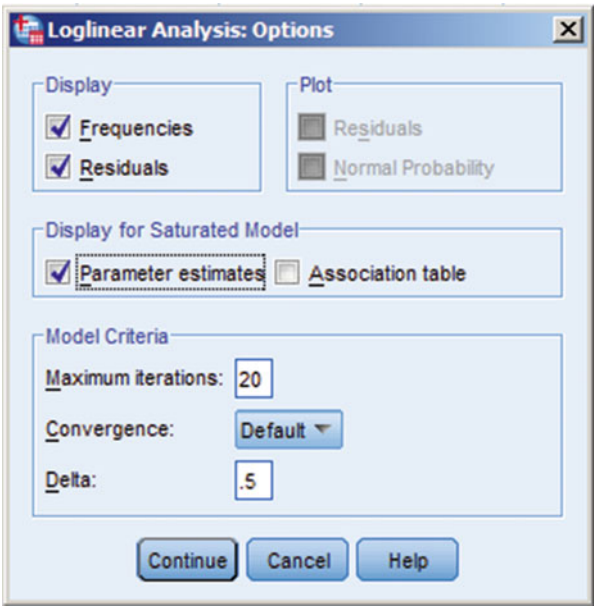


Fig. 8.2 The model selection loglinear analysis dialogue box

Fig. 8.3 The loglinear analysis: options dialogue box



The IBM SPSS output for the log-linear analysis of the saturated model is presented in Fig. 8.4. The observed (OBS) frequencies are simply those presented on page 142 and which could be readily generated from the IBM SPSS Crosstabs procedure. Recall that the saturated log-linear model includes all possible main effects and interactions. The saturated model reproduces the observed cell frequencies exactly, so the expected (EXP) and observed cell counts are equal. We shall later consider models that do not exactly reproduce cell counts.

A test of the hypothesis that a particular model adequately fits the observed data can be based on the familiar Pearsonian chi-square statistic. An alternative statistic is the *likelihood ratio chi-square*, but for large samples, the two are equivalent. The advantage of the likelihood-ratio chi-square (as will be seen later) is that it can be partitioned into interpretable parts. The saturated model always has chi-square statistics of zero ($p = 1.0$). This naturally means that the saturated model adequately fits the data, since chi-square is not the critical region; it always fits perfectly. (Of course the residuals and their standardised counterparts are zero for the saturated model).

In Fig. 8.4 and under the heading 'Estimates for Parameters' are the values of the ϑ coefficients computed in Sect. 8.2. The notation RISK*INTEREST refers to the interaction between these two factors. The parameters for this interaction correspond to the order of presentation of the OBS and EXP frequencies. Parameter 1 corresponds to RISK = 1, INTEREST = 1; parameter 2 to RISK = 1, INTEREST = 2. There is no need to report the parameter value of RISK = 1, INTEREST = 3 in Fig. 8.4, since these three parameters sum to zero. Parameter 3 for this interaction is for RISK = 2, INTEREST = 1; parameter 4 is for RISK = 2, INTEREST = 2. Again there is no need for the third parameter to be reported. The parameters for RISK = 3, INTEREST = 1, 2 and 3 may be derived from those already obtained. Turning to the main effects, parameters for RISK = 1 and RISK = 2 (and similarly for INTEREST) are reported and the third such parameter may be derived knowing that their sum is zero.

The Z-values in Fig. 8.4 are the ϑ divided by their standard deviations (here called *standard error*, Std Err) and they are thus standard normally distributed. Adopting a conventional significance level of 5%, any parameter coefficient beyond $\pm$ 1.96 suggests rejection of H_0: that the particular ϑ value is zero. The Z-value for low RISK is negative and significant. This suggests that the total of 37 cases in the category RISK = 1 is significantly smaller than the total of cases in the other two RISK categories. Ninety five percentage confidence intervals for the parameters ϑ are also presented in Fig. 8.4. Other aspects of the log-linear output are discussed in ensuing sections.

Cell Counts and Residuals

risk	interest	Observed Count[a]	Observed %	Expected Count	Expected %	Residuals	Std. Residuals
1	1	13.500	9.0%	13.500	9.0%	.000	.000
	2	12.500	8.3%	12.500	8.3%	.000	.000
	3	12.500	8.3%	12.500	8.3%	.000	.000
2	1	20.500	13.7%	20.500	13.7%	.000	.000
	2	17.500	11.7%	17.500	11.7%	.000	.000
	3	14.500	9.7%	14.500	9.7%	.000	.000
3	1	17.500	11.7%	17.500	11.7%	.000	.000
	2	19.500	13.0%	19.500	13.0%	.000	.000
	3	26.500	17.7%	26.500	17.7%	.000	.000

a. For saturated models, .500 has been added to all observed cells.

Goodness-of-Fit Tests

	Chi-Square	df	Sig.
Likelihood Ratio	.000	0	.
Pearson	.000	0	.

Parameter Estimates

Effect	Parameter	Estimate	Std. Error	Z	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
risk*interest	1	.036	.174	.208	.835	-.306	.378
	2	.002	.177	.009	.993	-.346	.349
	3	.153	.159	.961	.337	-.159	.466
	4	.037	.163	.228	.820	-.283	.357
risk	1	-.262	.125	-2.103	.035	-.506	-.018
	2	.039	.115	.336	.737	-.187	.265
interest	1	.015	.115	.129	.897	-.213	.243
	2	-.027	.118	-.232	.816	-.258	.203

Fig. 8.4 IBM SPSS output for the saturated model

8.3 The Independence Model

Representing an observed frequency table by a saturated log-linear model does not result in a simple, particularly meaningful description of the relationship between variables, especially when there are more than two of them. Possibly, parameters with small values could be removed from the saturated model to create simpler models. To illustrate the general procedure for fitting a model that does not contain all possible parameters (called an *unsaturated model*), consider the familiar independence hypothesis for the two-way contingency table. If variables are independent, they can be represented by a log-linear model that does not have an interaction

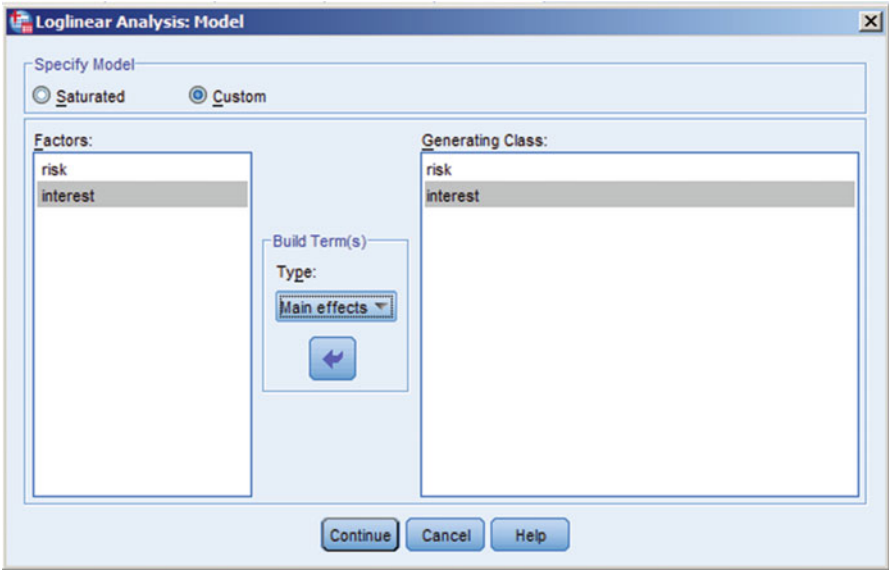


Fig. 8.5 The loglinear analysis: model dialogue box for main effects only

term. For example, if variables A and B are independent and with reference to Eq. 8.1, then the independence model may be written as:

$$\mathrm{Ln}\left(\widehat{F_{ij}}\right) = \mu + + \vartheta_i^A + \vartheta_i^B + \ldots \tag{8.3}$$

Being an unsaturated model, the F_{ij} are no longer exact, but are estimated, hence the "hat" (^) notation. The unsaturated model of Eq. 8.3 consists of only main effects, so we do not want the interaction term of the saturated model. The difference this time is in the selection of options in the Loglinear Analysis: model dialogue box of Fig. 8.1. At the top of this dialogue box we wish to customise our model rather than select the default saturated model, so select the option 'Custom'. In the box titled 'Build term (s)', select the option 'Main effects' rather than the default 'Interactions'. The variables RISK and INTEREST are entered into the 'Generating Class' box and the dialogue box appears as in Fig. 8.5. The IBM SPSS output from this unsaturated model is presented in Fig. 8.6.

The Pearsonian chi-square statistic in Fig. 8.6 is the same as that stated on page 141 and which would be derived from the IBM SPSS Crosstabs procedure. Recall that the chi-square statistics in Fig. 8.6 test the hypothesis that our model (here without interaction effects) fits the data adequately. As the significance levels are in excess of 0.05, we fail to reject this hypothesis and the independence model is not rejected. Hence, RISK and INTEREST are independent and there is no need for the interaction term.

Cell Counts and Residuals

risk	interest	Observed Count	Observed %	Expected Count	Expected %	Residuals	Std. Residuals
1	1	13.000	8.7%	12.333	8.2%	.667	.190
	2	12.000	8.0%	11.840	7.9%	.160	.046
	3	12.000	8.0%	12.827	8.6%	-.827	-.231
2	1	20.000	13.3%	17.000	11.3%	3.000	.728
	2	17.000	11.3%	16.320	10.9%	.680	.168
	3	14.000	9.3%	17.680	11.8%	-3.680	-.875
3	1	17.000	11.3%	20.667	13.8%	-3.667	-.807
	2	19.000	12.7%	19.840	13.2%	-.840	-.189
	3	26.000	17.3%	21.493	14.3%	4.507	.972

Goodness-of-Fit Tests

	Chi-Square	df	Sig.
Likelihood Ratio	3.060	4	.548
Pearson	3.046	4	.550

Fig. 8.6 IBM SPSS output for the unsaturated model

The residuals also indicate the adequacy or otherwise of the model fit. They should be small in value and exhibit no discernible pattern. To this end, it is easiest to examine the standardised residuals and values beyond 1.96 are extreme. There are no such residuals in Fig. 8.6. The standardised residuals should be normally distributed if the model is adequate. Derivation of a normal probability plot is an option in the *Loglinear Analysis: Options dialogue box* of Figs. 8.3 and 8.7 presents this diagram, which suggests no serious departure from normality. The user could plot the standardised residuals against the observed or expected frequencies and there should be no discernible pattern present if the residuals are random.

8.4 Hierarchical Models

The interaction between RISK and INTEREST in the previous sections is called a *2-way interaction*. Higher order interactions may be part of the study. For example, if a consumer regarded a particular purchase as interesting and shopped under a high degree of time pressure, then (s)he might attach a high element of risk to the buying process. This would be a *3-way interaction*. In a hierarchical model, if a term exists for the interaction of a set of variables, then there must be lower-order terms for all possible combinations for these variables. For a three variable model,

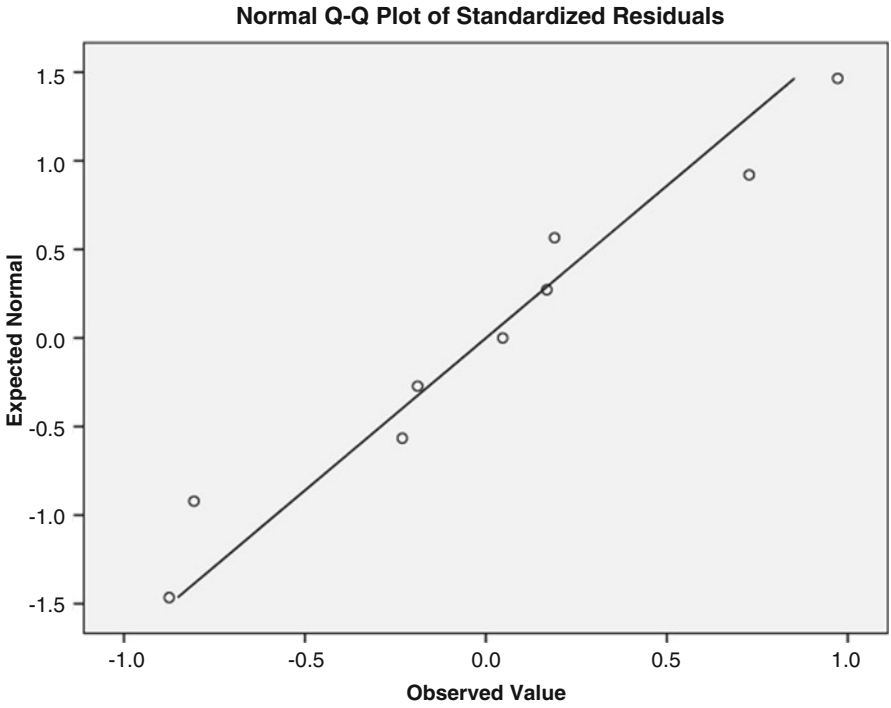


Fig. 8.7 A normal probability plot of residuals from the unsaturated model

if the user includes the 3-way interaction ϑ^{ABC}, then the terms ϑ^{A}, ϑ^{B}, ϑ^{C}, ϑ^{AB}, ϑ^{AC} and ϑ^{BC} must also be included. This 3-way interaction is used when the researcher believes that the variables A, B and C are mutually dependent in some way. To describe a hierarchical model, one customises the model by simply entering all the pertinent variables into the box entitled 'Generating Class' in the *Loglinear Analysis: Model dialogue box* and under the heading 'Build Term(s)' choose the default 'Interactions'. In the above 3-way interaction model, A* B*C would imply that ϑ parameters for A*B, A*C, B*C, A, B and C will be computed and reported as well.

However, many different models are possible for a set of variables. The selected model should fit the data, the substantially interpretable and as simple (parsimonious) as possible. For example high-order interactions are often difficult to interpret. In order to derive a parsimonious model, one could fit the saturated model and examine standardised values for the ϑ parameters. Another strategy is to test systematically the contribution to the model made by terms of a particular order. For example, one could fit a model with interaction terms then a model with fixed effects only. The change in the chi-square statistic between the two models is attributable to the interaction terms.

The hierarchical procedure is illustrated by introducing two further variables into the previous study and which are found in the file CONSUMER.SAV. These

two variables were also deemed in the study to influence the extent of consumer comparison shopping i.e. the effort consumers place in the purchase process. The first variable is the time pressure (IBM SPSS variable named TIME) under which consumers shop and the second variable is the status (STATUS) that consumers attach to the particular purchase. The latter two variables are coded from '1' to '3' exactly as for RISK and INTEREST. We shall employ hierarchical log-linear analysis to establish any significant relationships between combinations of RISK, INTEREST, TIME and STATUS that determine the observed frequencies in each class of these four factors.

Partitioning the chi-square statistic can be a useful first step in identifying a "best", parsimonious model between the study variables. With four variables, we have 81 ($3 \times 3 \times 3 \times 3$) cells in our frequency table, so it is necessary to let delta equal to its default value of 0.5, in that there may well be zero frequencies in some cells. Setting delta is performed in the *Loglinear Analysis: Options dialogue box*. If we run the saturated model, it is now necessary to discuss that part of the IBM SPSS output obtained in Fig. 8.8. Under the heading 'K-way and higher order effects, the first line is a test of the hypothesis that the fourth order interaction is zero (there being no higher order interaction). Both the likelihood ratio and Pearson chi-square statistics reject these hypotheses ($p = 0.0$). The second line tests the hypothesis that third AND fourth order interactions are zero. Again both test to reject this hypothesis ($P < 0.05$). We may use the fact that chi-square is additive and subtractive in terms of its numerical value and degrees of freedom. The chi-square for fourth order interactions is 0.000 with $df = 16$; that for third AND fourth order interactions is 0.000, $df = 48$. Therefore, the chi-square statistic for the hypothesis that just third interactions are zero is $0.000 - 0.000 = 0.000$ with $df = 48 - 16 = 32$. Test results for such individual effects are presented under the heading 'K-way effects' in Fig. 8.8. The likelihood ratio results suggest that main effects ($K = 1$) and second order interactions are significantly different from zero; the Pearsonian results suggest that main effects, second and third order interactions are significantly different from zero. Both approaches agree that any fourth order interaction is not significant.

The results in Fig. 8.8 under the heading 'Partial Associations' provide an indication of the collective importance of effects of various orders. They do not, however, test individual terms. That is, although the overall hypothesis of second order interaction effects are zero may be rejected, that does not mean that every third order effect is present. One approach is to fit two models differing only in the presence of the effect to be tested and test the differences in the two chi-square statistics for significance. This difference in chi-square statistics is called the partial chi-square. *Partial chi-squares* are based on the likelihood ratio chi-square.

Overall, we know that second order interaction effects are significantly different from zero. Examination of the partial chi-squares shows that of the possible 2-way

Goodness-of-Fit Tests

	Chi-Square	df	Sig.
Likelihood Ratio	.000	0	.
Pearson	.000	0	.

K-Way and Higher-Order Effects

	K	dt	Likelihood Ratio		Pearson		Number of Iterations
			Chi-Square	Sig.	Chi-Square	Sig.	
K-way and Higher Order Effects[a]	1	80	668.799	.000	1290.720	.000	0
	2	72	655.656	.000	1225.013	.000	2
	3	48	.000	1.000	.000	1.000	3
	4	16	.000	1.000	.000	1.000	2
K-way Effects[b]	1	8	13.144	.107	65.707	.000	0
	2	24	655.656	.000	1225.013	.000	0
	3	32	.000	1.000	.000	1.000	0
	4	16	.000	1.000	.000	1.000	0

df used for these tests NOT been adjusted for structural or sampling zeros. Tests using these df may be conservative.

a. Tests that k-way and higher order effects are zero.

b. Tests that k-way effects are zero.

Partial Associations

Effect	df	Partial Chi-Square	Sig.	Number of Iterations
risk*time*status	8	.000	1.000	2
risk*time*interest	8	.000	1.000	2
risk*status*interest	8	.000	1.000	2
time*status*interest	8	.000	1.000	2
risk*time	4	320.112	.000	3
risk*status	4	.000	1.000	3
time*status	4	.000	1.000	3
risk*interest	4	.000	1.000	3
time*interest	4	.000	1.000	3
status*interest	4	326.363	.000	2
risk	2	6.412	.041	2
time	2	6.412	.041	2
status	2	.160	.923	2
interest	2	.160	.923	2

Fig. 8.8 IBM SPSS for the 4-way loglinear model

effects, RISK*TIME and INTEREST*TIME only are significant. These two pairs contribute most of the overall significance. None of the 3-way combinations are significantly different from zero. Hence, it appears that the study can be represented parsimoniously by a model that includes just the above 2-way interactions and being hierarchical, the individual variable main effects (RISK, TIME and INTEREST) must be included too.

8.5 Backward Elimination

The above process is useful for showing the logical of hierarchical log-linear analysis and an approach to obtaining a significant and at the same time, parsimonious model. However, as in regression analysis, another way to arrive at a "best" model is by using variable selection procedures. Forward selection adds effects to the model, while backwards elimination starts off with the saturated model and removes effects that do not satisfy the criterion for remaining in the model. Since backward elimination appears to be the better procedure for model selection, this is the approach available in IBM SPSS.

The initial model for backward elimination need to be saturated. Indeed, preceding analysis suggests that we need never consider a possible 4-way interaction. To operationalise backward elimination, the user simply selects this option in the Model Selection *Loglinear Analysis dialogue box*. Every step in the elimination process is reported, but in most instances, we are only interested in the k-way effects that remain in the final step as shown in Fig. 8.9. As expected, the 2-way interactions RISK*TIME and INTEREST*TIME remain.

The chi-square goodness of fit statistics indicate that we do not reject the hypothesis that this model is an adequate fir for the data. For brevity, just the Z-scores associated with the ϑ coefficients for the RISK*TIME interaction are reported. All the Z-scores are statistically significant and study of their sign in the study context would explain why this 2-way interaction is significant. A plot of the residuals should be generated for this model to assess further its adequacy.

Parameter Estimates

Effect	Parameter	Estimate	Std. Error	Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
risk*time*status*interest*	1	1.508	.548	2.753	.006	.434	2.581
	2	-.752	.590	-1.275	.202	-1.907	.404
	3	-.752	.590	-1.275	.202	-1.907	.404
	4	1.494	.548	2.728	.006	.421	2.568
	5	-.752	.589	-1.277	.202	-1.907	.403
	6	.389	.609	.638	.523	-.805	1.583
	7	.389	.609	.638	.523	-.805	1.583
	8	-.731	.589	-1.240	.215	-1.886	.424
	9	-.752	.589	-1.277	.202	-1.907	.403
	10	.389	.609	.638	.523	-.805	1.583
	11	.389	.609	.638	.523	-.805	1.583
	12	-.731	.589	-1.240	.215	-1.886	.424
	13	1.587	.547	2.904	.004	-.516	2.659
	14	-.802	.589	-1.361	.173	-1.957	.353
	15	-.802	.589	-1.361	.173	-1.957	.353
	16	1.565	.547	2.862	.004	.493	2.637
risk*time*status	1	.007	.407	.018	.986	-.791	.806
	2	-.006	.407	-.014	.988	-.804	.792
	3	-.023	.426	-.053	.957	-.858	.813
	4	-.001	.426	-.003	.998	-.837	.834
	5	-.023	.426	-.053	.957	-.858	.813

Backward Elimination Statistics

Step Summary

Step[a]		Effects	Chi-Square[c]	df	Sig.	Number of Iterations
0	Generating Class[b]	risk*time*status*interest	.000	0	.	
	Deleted Effect 1	risk*time*status*interest	.000	16	1.000	2
1	Generating Class[b]	risk*time*status, risk*time*interest, risk*status*interest, time*status*interest	.000	16	1.000	
	Deleted Effect 1	risk*time*status	.000	8	1.000	2
2	Generating Class[b]	risk*time*interest, risk*status*interest, time*status*interest	.000	24	1.000	
	Deleted Effect 1	risk*time*interest	.000	8	1.000	3
3	Generating Class[b]	risk*status*interest, time*status*interest, risk*time	.000	32	1.000	
	Deleted Effect 1	risk*status*interest	.000	8	1.000	2
4	Generating Class[b]	time*status*interest risk*time,	.000	48	1.000	

Goodness-of-Fit Tests

	Chi-Square	df	Sig.
Likelihood Ratio	3.060	64	1.000
Pearson	3.046	64	1.000

Fig. 8.9 Part of the results from backward elimination

Part III
Research Methods

Chapter 9
Testing for Dependence

9.1 Introduction

There are many business-related research projects in which a major focus would be whether two, non-metric variables are dependent upon each other or not. For example, in market research analyses, it is common to test whether factors such as consumer attitudes (e.g. classified as "favorable", "unfavorable" or "neutral"), consumer behavior or expenditure patterns depend on a series of socio-economic variables such as age, income, family structure, religion, gender etc. In a financial context, we may wish to examine whether today's movements in shares (classified as "up", "down" or "stationary") depend upon past movements in exchange rates (classified as "up", "down" or "stationary"). In such instances, we have recourse to **Pearson's chi-squared (χ^2) test of independence**, which may be applied even at the nominal level of measurement.

As an illustration, banks and financial institutions have become increasingly aware that customers' perceptions of various forms of service quality (SERVQUAL) have an impact on the organization's ability to retain/expand its business. This statement is all but true given the daily challenges banks are presented with because of the continuous technology improvements. A recent study shows that customers are more likely to use internet banking rather than go physically to their respective branch. Consequently, the SERVQUAL literature has been increasing rapidly over the last two decades, both in the commercial and academic sectors.

A survey has been conducted to assess customers' satisfaction with a particular bank's credit facilities. Levels of satisfaction (IBM SPSS variable name **LEVELSAT**) are coded as 1 "dissatisfied", 2 "neutral/undecided" and 3 "satisfied". Research is conducted to see if levels of satisfaction depend on the customer's affluence, measured in terms of how much the customer deposits per month (**DEPOSIT**), which is coded as 1 "less than £500″, 2 "£500 up to £2500″ and

A. Aljandali, *Multivariate Methods and Forecasting with IBM® SPSS® Statistics*,
Statistics and Econometrics for Finance, DOI 10.1007/978-3-319-56481-4_9

Table 9.1 Contingency table

	Level of satisfaction			
Monthly deposit	Dissatisfied	Neutral/undecided	Satisfied	Total
Less than £1500	560	111	78	749
£1500 up to £5000	216	490	107	813
More than £5000	274	306	100	680
Total	1050	907	285	2242

3 "more than £2500″. Three thousand four hundred and two customers responded to a survey and the data are available in the file SERVQUAL.SAV

The first part of the analysis of (in) dependency involves generating what is called a **crosstabulation** or **contingency table** of the results, shown below (Table 9.1):

Out of a total of 2242 customers, 560 deposited less than £1500 per month and were "dissatisfied" with the bank's credit facilities. Five hundred and sixty is called the **observed frequency**. Generally, the contingency table suggests that the more an individual deposits the more he is likely to be satisfied with the credit facilities on offer. This would imply that **LEVELSAT** and **DEPOSIT** are **mutually dependent**.

This chi-squared test has the null hypothesis H_0: the two study variables, levels of satisfaction and amount of monthly deposit, are **independent**. In most research, we will want to reject the null in favour of the alternative that a dependency exists.

Remember that all statistical tests are conducted under the assumption that the null is true. Consider the top left hand corner cell in the above contingency table. By the multiplication law of probability and via the statistical independence assumed under the null:

P(customer deposits < £1500 AND is dissatisfied) =
P(customer deposits < £1500).P(customer is dissatisfied) =
(749 / 2242) * (1050/2242) using the associated row and column totals in the contingency table.

Consequently, under the null hypothesis, the **expected frequency** of customers who deposit < £1500 and who are dissatisfied with the credit facilities is (749 / 2242) * (1050/2242) * 2242 = 346.5 – it seems that this expected frequency (E_i) derived under the assumption that the null is true, is quite dissimilar to the observed frequency (O_i) of 560, suggesting that we may have to reject the null. We need to compute the expected frequencies for all nine cells in the contingency table and compare them with their associated observed frequencies before we make a final conclusion about the null. From the above arithmetic it may be noted that there is a quick way to compute the expected frequencies via:

$$E_i = \frac{(rowtotal)(columntotal)}{n}$$

where n is the total number of observations in the contingency table, here 3042.

The further apart are the nine observed frequencies and their expected counterparts, the less likely is the null to be true. The differences between the O_i and the E_i form the basis for the chi squared test statistic which is:

$$\sum_i \frac{(O_i - E_i)^2}{E_i} \in \chi^2$$

where the number of degrees of freedom is given by $\nu =$ (no. of rows $-$ 1).(no. of columns $-$ 1), so here, $\nu = 4$.

9.2 Chi-Square in IBM SPSS Statistics

Naturally, IBM SPSS Statistics performs all of the above arithmetic and also generates the contingency table and expected values if desired. It is always sensible to select the expected values as an option in order to specify the nature of any dependency that may exist. In IBM SPSS Statistics, click:

```
Analyze
   Descriptive Statistics
      Crosstabs
```

which produces the Crosstabs dialogue box on Fig. 9.1.

It makes no difference (save for presentation) which variable constitutes the rows of the contingency table and which forms the columns. Click the Statistics... button in the above dialogue box. There is a wealth of options here, but just the Chi-Square statistic has been selected as per Fig. 9.2. Click the Cells... button on the Crosstabs dialogue box to obtain the output in Fig. 9.3. And I have selected both the observed and expected frequencies to be presented in the contingency table. Note also that **unstandardized residuals** have been requested. These are simply the residuals $=$ Oi $-$ E_i.

The **standardized residuals** permit the user the better method for determining why any dependency exists, since they remove the problem of the magnitudes of the frequencies involved e.g. if $O_i = 10{,}000$ and $E_i = 9400$, the unstandardized residual is 600, but the standardized residual will probably be low.

Upon running the Crosstabs routine, the following results are obtained (see below)

deposit * levelsat Crosstabulation

			levelsat			
			Dissatisfied	Neutral/undecided	Satisified	Total
deposit	Less than £ 1500	Count	560	111	78	749
		Expected Count	350.8	303.0	95.2	749.0
		Residual	209.2	-192.0	-17.2	
		Standardized Residual	11.2	-11.0	-1.8	
	£ 1500 to £ 5000	Count	216	490	107	813
		Expected Count	380.8	328.9	103.3	819.0
		Residual	-164.8	161.1	3.7	
		Standardized Residual	-8.4	8.9	4	
	More than £ 2500	Count	274	306	100	680
		Expected Count	318.5	275.1	86.4	680.0
		Residual	-44.5	30.9	13.6	
		Standardized Residual	-2.5	1.9	1.5	
Total		Count	1050	907	285	2242
		Expected Count	1050.0	907.0	285.0	2242.0

Chi-Square Tests

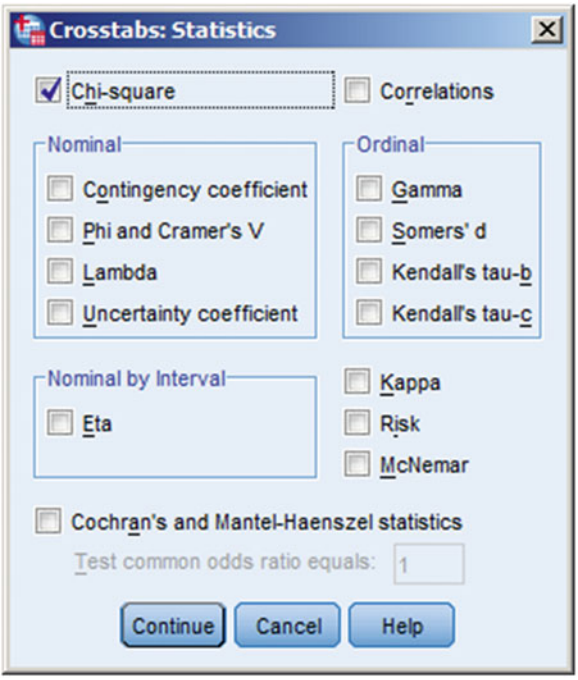
	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	411.705[a]	4	.000
Likelihood Ratio	434.982	4	.000
Linear-by-Linear Association	118.330	1	.000
N of Valid Cases	2242		

a. 0 cells (0.0%) have expected count less than 5. the minimum expected count is 86.44.

Fig. 9.1 The Crosstabs dialogue box

The value of $\chi^2$is 411.705 with an associated significance level of 0.000 to three decimal places. If the significance level is less than 0.05, we reject the null hypothesis. Consequently, we reject the null here and conclude that satisfaction levels depend upon the size of the customers' monthly deposits. Examination of the standardized residuals indicates why such a dependency exists. For example in the top left hand cell, the unstandardized residual is $560 - 350.8 = 209.2$ which has a standardized value of 11.2. Therefore, we **observe** significantly more dissatisfied customers with monthly deposits less than £1500 than we would **expect** if the null

Fig. 9.2 The Crosstabs: statistics dialogue box



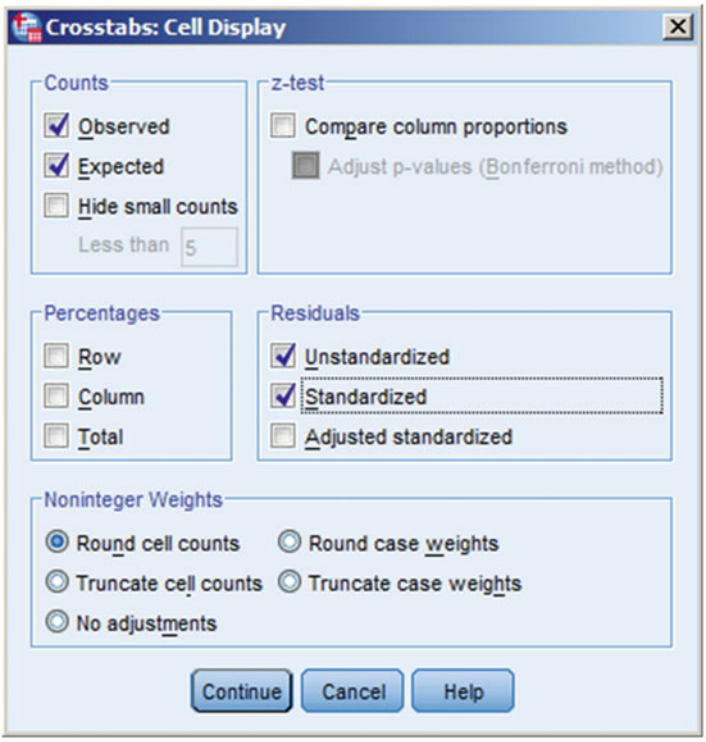


Fig. 9.3 The Crosstabs: cell display dialogue box

deposit * levelsat Crosstabulation

			levelsat			
			Dissatisified	Neutral/undecided	Satisified	Total
deposit	Less than £ 1500	Count	560	111	78	749
		Expected Count	258.5	223.3	267.1	749.0
		Residual	301.5	-112.3	-189.1	
		Std. Residual	18.7	-7.5	-11.6	
	£ 1500 to £ 2500	Count	216	490	107	813
		Expected Count	280.6	242.4	290.0	813.0
		Residual	-64.6	247.6	-183.0	
		Std. Residual	-3.9	15.9	-10.7	
	More than £ 2500	Count	274	306	900	1480
		Expected Count	510.8	441.3	527.9	1480.0
		Residual	-236.8	-135.3	372.1	
		Std. Residual	-10.5	-6.4	16.2	
Total		Count	1050	907	1085	3042
		Expected Count	1050.0	907.0	1085.0	3042.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1338.803[a]	4	.000
Likelihood Ratio	1276.677	4	.000
Linear-by-Linear Association	838.882	1	.000
N of Valid Cases	3042		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 223.32.

Fig. 9.4 A Crosstabulation of deposits and levels of satisfaction (Note: If there are three or more study variables, it is best not to use the chi-squared test of independence. There is a method called log-linear analysis which is available in IBM SPSS Statistics (please refer to Chap. 8))

was true. Alternatively, examine the top right hand corner cell in which the standardized residual is −1.8. We **observe** significantly less satisfied customers depositing less than £1500 per month than we would **expect** if the null was correct. Further examination of the remaining cells in the table will further amplify these conclusions (Fig. 9.4).

Chapter 10
Testing for Differences Between Groups

10.1 Introduction

The purpose of this chapter is to present methods for examining whether mutually exclusive groups of objects or people differ in some respect. For example, in marketing studies, one might wish to see if different consumer age groups have different expenditure patterns for a particular product or service. One might wish to see if competing investment portfolios differ in their net returns, whether different currencies react to economic downturns in different ways over time or if different types of advertising campaign generate different consumer reactions. All that is required to apply the methods below is two or more mutually exclusive groups and some measurable characteristic possessed by all members of each group and which has the potential to generate differences between the groups.

The following data are percentage changes in price for samples of three types of shares, between the close of trading on a Monday and the close of trading the following Tuesday (Table 10.1):

We shall test the Differences between groups:

X that the percentage changes in the populations of the three types of shares (i.e. the groups) are equal. The alternative hypothesis is that one or more of the groups differs from the rest.

- If the data (i.e. the % changes in share prices) are deemed to be drawn from non-normal populations, then the ***nonparametric Kruskal-Wallis*** test should be used.
- If the data are believed to be drawn from normal populations with equal variance, then ***the parametric one-way analysis of variance*** **(ANOVA)** should be used.

Remember that parametric tests are the more powerful as long as the assumptions underlying them are met. Data assumptions need to be tested before deciding which of the above two tests to use. (Note that neither of the above two tests requires equal sample sizes, which happens to be the case in Table 10.1).

A. Aljandali, *Multivariate Methods and Forecasting with IBM® SPSS® Statistics*,
Statistics and Econometrics for Finance, DOI 10.1007/978-3-319-56481-4_10

Table 10.1 Types of shares * % change in shares

Shares in		
Metals (%)	Aerospace industries (%)	Banks & building societies (%)
0.3	−0.6	−3.1
−1.7	1.8	0.0
−2.0	1.4	2.8
2.6	−2.7	1.6
0.1	−2.0	1.5

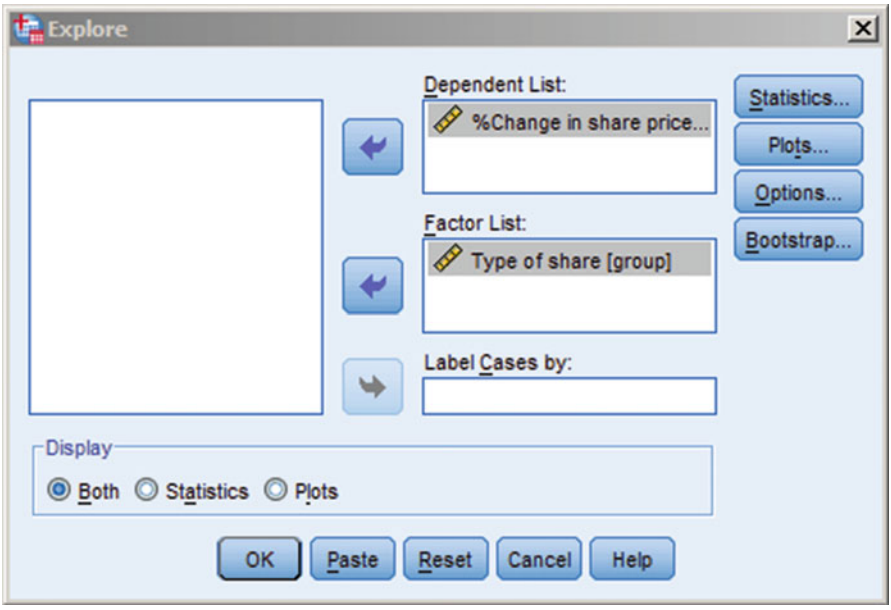


Fig. 10.1 The explore dialogue box

10.2 Testing for Population Normality and Equal Variances

To test the normality and equal variance assumptions via IBM SPSS Statistics, open the data file and click:

```
Analyze
   Descriptive Statistics
      Explore
```

Which will give rise to Fig. 10.1. The dependent variable is called **CHANGE** (i.e. the % changes in share prices) and the factor list should contain the variable name **GROUP** (i.e. the three type of share). Click the PLOTS. . . button and select:

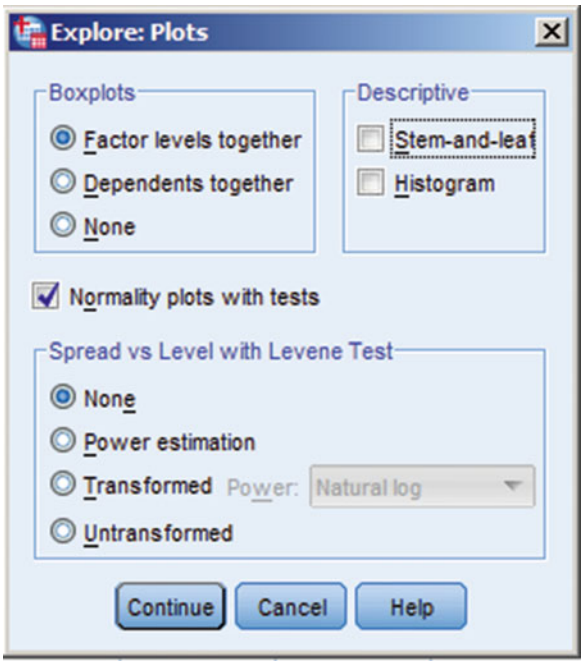


Fig. 10.2 The explore: plots dialogue box

(i) Normality plots with tests and
(ii) 'Spread versus Level' with Levene test for untransformed data

These choices are shown in Fig. 10.2:

Normality is tested via the **Shapiro-Wilk** statistic. The null hypothesis of the Shapiro-Wilk test is that the particular sample is drawn from a normal population. The null is rejected if the significance level $p < 0.05$. As shown in the IBM SPSS output overleaf, the significance levels associated with each sample are – ***shares in metals*** ($p = 0.517$), ***shares in aerospace industries*** ($p = 0.466$) and ***shares in banks/building societies*** ($p = 0.392$), so we conclude that all three samples have been drawn from normally distributed populations (Figs. 10.3 and 10.4).

Test of homogeneity of variance					
		Levene statistic	df1	df2	Sig.
Change %change in share price	Based on mean	.124	2	12	.885
	Based on median	.047	2	12	.954
	Based on median and with adjusted df	.047	2	8.884	.955
	Based on trimmed mean	.106	2	12	.900

Tests of Normality

	Type of share	Kolmogorov-Smirnov[a] Statistic	df	Sig.	Shapiro-Wilk Statistic	df	Sig.
% Change in share price	Metals	.206	5	.200*	.918	5	.517
	aerospace industries	.219	5	.200*	.910	5	.466
	banks	.260	5	.200*	.897	5	.392

*. This is a lower bound of the true significance.

a.Lilliefors Significance Correction

Fig. 10.3 Test of normality output

		Levene Statistic	df1	df2	Sig.
change %Change in share price	Based on Mean	.124	2	12	.885
	Based on Median	.047	2	12	.954
	Based on Median and with adjusted df	.047	2	8.884	.955
	Based on trimmed mean	.106	2	12	.900

Fig. 10.4 Test of homogeneity of Variance output

The **Levene test** examines the null hypothesis that the three samples have been drawn from populations with equal variance. (This latter property is called **homogeneity of variance**). The Levene statistic 'based on the mean' has significance $p = 0.885$ as shown below, so we fail to reject the null. Hence, both assumptions underlying the one-way ANOVA are met. Note that among the various diagrams produced by the IBM SPSS Statistics Explore routine is the boxplot in Fig. 10.5.

10.3 The One-Way Analysis of Variance (ANOVA)

Having tested the two assumptions underlying the test, we apply the one-way ANOVA in IBM SPSS Statistics via:

```
Analyze
   Compare Means
      One-way ANOVA
```

and **CHANGE** is the dependent variable and **GROUP** is the factor. The null hypothesis is that the population means from which the three samples have been drawn are equal. The alternative hypothesis is that one or more of these means differ. (Note that the null is tested by the F statistic and evidence exists that the

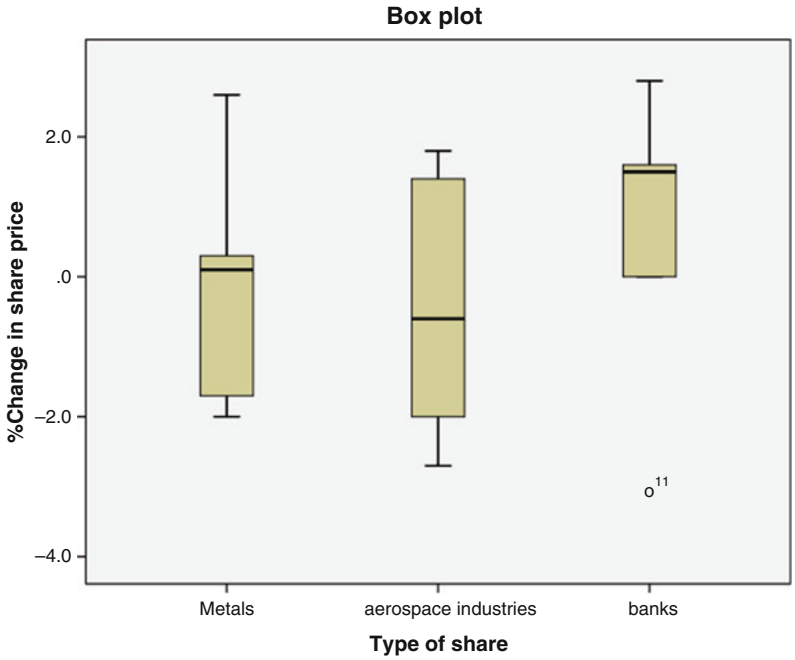


Fig. 10.5 Box plots of type of share * % change in price

F statistic is particularly sensitive to departures from normality, which is why we checked this assumption).

If the alternative hypothesis is selected by the test, then we shall need to know exactly which pair(s) of samples has/have caused the null to be rejected. This is conducted by means of a **multiple comparisons procedure.** There are several such procedures available but perhaps the most widely used is due to **Scheffé**. Click the Post Hoc. . . button at the bottom of the one-way ANOVA dialogue box (Fig. 10.6) and select Scheffé from the alternatives presented under the heading 'equal variances assumed' as shown in Fig. 10.7.

As shown in the output of Fig. 10.8, the one-way ANOVA F statistic has value 0.304 ($p = 0.743$) so we fail to reject the null and conclude that the three population mean % changes in the three groups of share prices are equal.

If the null had been rejected, we would have used Scheffé's multiple comparisons procedure to locate the differences. The results from this procedure appear under the heading 'Post Hoc tests' in the SPSS output. Given that here we have failed to reject the null, it is no surprise to see below that changes in no pairs of shares are significantly different from zero. For example, the difference in sample means for changes in metal shares and aerospace shares is 0.2800, which is not significantly different from zero ($p = 0.977$); the difference in sample means for changes in metal shares and bank/building society shares is −0.7000 ($p = 0.866$). No pair exhibits a significant difference.

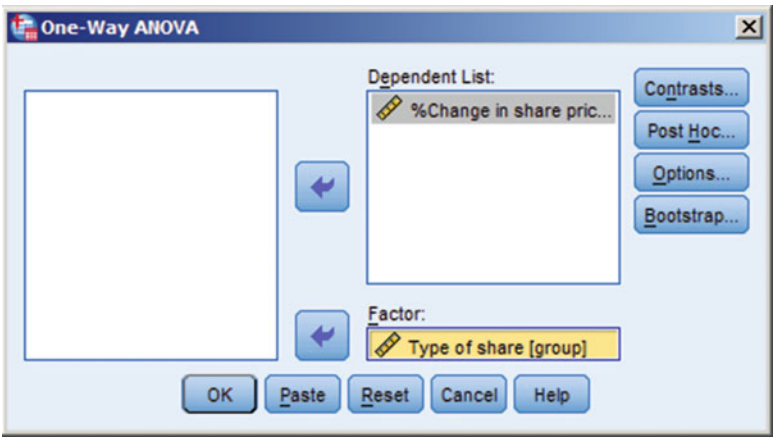


Fig. 10.6 The one-way ANOVA box

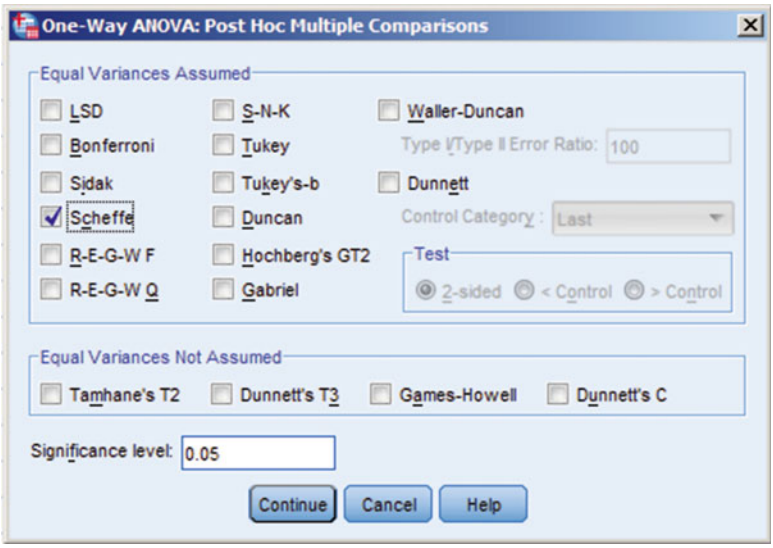


Fig. 10.7 The one-way ANOVA: post hoc multiple comparisons box

10.4 The Kruskal-Wallis Test

It must be appreciated that the Kruskal-Wallis (KW) test is the less powerful here, since the assumptions underlying the one-way ANOVA were met. However, this will not always be the case, so the test is illustrated below. The null hypothesis is that there is no overall difference between the $k = 3$ populations from which the samples have been drawn. Failure to reject the null indicates that the population means are equal. Rejection of the null implies that the three populations differ in some respect; it may not be the means. The KW test is accessed via:

ANOVA

% Change in share price

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2.548	2	1.274	.304	.743
Within Groups	50.312	12	4.193		
Total	52.860	14			

Post Hoc Tests

Multiple Comparisons

Dependent Variable: %Change in share price

Scheffe

(I) Type of share	(J) Type of share	Mean Difference (I-J)	Std.Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Metals	aerospace industries	.2800	1.2950	.977	-3.330	3.890
	banks	-.7000	1.2950	.866	-4.310	2.910
aerospace industries	Metals	-.2800	1.2950	.977	-3.890	3.330
	banks	-.9800	1.2950	.756	-4.590	2.630
banks	Metals	.7000	1.2950	.866	-2.910	4.310
	aerospace industries	.9800	1.2950	.756	-2.630	4.590

Fig. 10.8 Part of the IBM SPSS statistics output: ANOVA & multiple comparisons

```
Analyze
   Nonparametric tests
      k independent samples
```

and **CHANGE** is 'test variable' and **GROUP** is the 'grouping variable'. IBM SPSS ranks the sample data – i.e. all 15 readings independently of group membership – from low to high. The logic is that if the three groups are equal, then they will all receive similar ranked values; if one group has % share price changes that are much lower than the % changes in the other two groups of shares, then that former group will receive more low ranked values. The mean of the ranks allocated to each sample is part of the Kruskal-Wallis output, and they are shown in Fig. 10.9:

The lowest ranks have been allocated to changes in shares for the aerospace industries. However, as shown below, the KW test statistic has value 0.666 ($p = 0.717$), so we fail to reject the null. There are no overall differences between the three populations from which the samples have been drawn, so their mean % changes in share prices are considered to be equal. Had we rejected the null, we would have had recourse to the multiple comparisons procedure associated with the Kruskal-Wallis test.

In this Kruskal-Wallis test, the ranks from 1 to 15 were allocated to the total of 15 share price changes. The sum of the first n integers is $n(n + 1)/2$, so a total of $15.(16)/2 = 120$ ranking points were allocated to the three groups. Under the equality required of the null, each group should have received $120/3 = 40$ ranking points.

Kruskal-Wallis Test

Ranks

	Type of share	N	Mean Rank
%Change in share price	Metals	5	7.90
	aerospace industries	5	6.90
	banks	5	9.20
	Total	15	

Test Statistics[a,b]

	%Change in share price
Chi-Square	.666
df	2
Asymp.Sig.	.717

a. Kruskal Wallis Test

b. Grouping Variable: Type of Share

Fig. 10.9 Kruskal-Wallis test output

With five share prices in each group, the mean rank should be $40/5 = 8$ under the null. Hence, there are marginally lower % share price change in the aerospace industry than would be expected under the null and marginally higher % changes for bank shares. However and overall, the test produces insufficient evidence to reject the null hypothesis of equality.

Chapter 11
Current and Constant Prices

11.1 HICP and RPI

Harmonized Indices of Consumer prices (HICP) are constructed in each member state of the European Union for the purpose of international comparisons of consumer price inflation. HICP's are published by *Eurostat* – The European Commission's statistical arm – and have been published monthly since March 1997. They provide cross-country comparisons of inflation on a comparable or harmonized basis. Before the introduction of the HICP, comparison of inflation rates across the European Union countries was not possible, due to the different ways countries computed consumer price indices. Also, the basket of goods covered by each country was different.

The UK continues to produce the Retail Price Index (RPI). The product coverage of the HICP and the RPI is to some extent similar. However, a number of RPI goods are not included in the HICP. These include mortgage repayments, buildings insurance, estate agents' fees, council tax, and expenditure by households on education and health care. Other excluded areas include technically difficult sectors where differences in national markets make the production of indices difficult. For example, in the fields of health and education, many goods and services are heavily subsidized by the state but the extent of such subsidies varies substantially across Member States of the European Union.

On the other hand, the HICP includes some expenditures not included in the RPI, for example personal computers and air fares. These and other items are included in the "all items" HICP since their expenditures exceed one part per thousand of consumers' total expenditure – the threshold set by *Eurostat* – above which items

A. Aljandali, *Multivariate Methods and Forecasting with IBM® SPSS® Statistics*,
Statistics and Econometrics for Finance, DOI 10.1007/978-3-319-56481-4_11

have to be included in the computation of the HICP. The HICP covers all private households, whereas the RPI excludes the top 4% in terms of their income. The RPI also excludes pensioner households who derive at least 75% of their income from state benefits.

The RPI annual rate generally exceeds the HICP annual rate. This is mainly due to the HICP's exclusion of most housing costs that are included in the RPI. These housing costs have been rising relatively rapidly over the last few years.

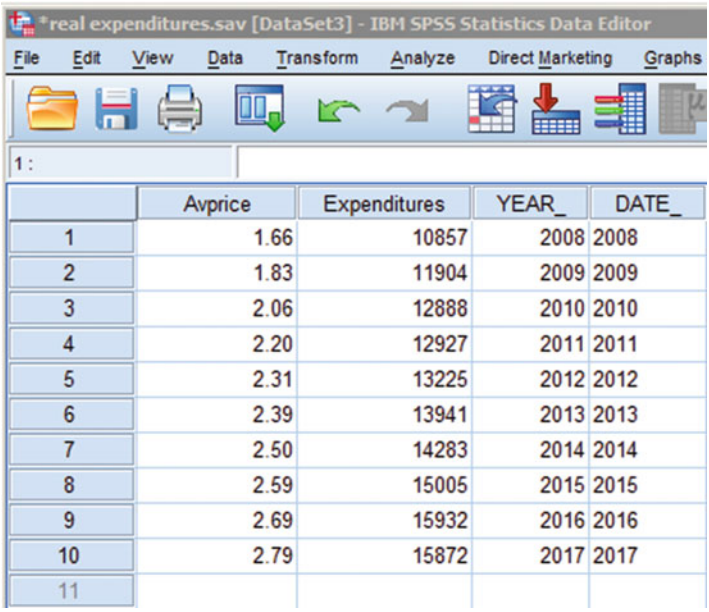
In terms of their basic usability, there is little to choose between them. Both are published each month and are subject to minimal revisions – the RPI is by convention never revised. A key advantage of the RPI is its familiarity and credibility based on a longer history. Inevitably, it will be some time before the HICP becomes widely recognised by the public. The HICP's exclusion of most elements of owner-occupier housing costs lessens its relevance for some users, but this must be weighed against the significant difficulties encountered in measuring such costs appropriately, reflected in the absence of any international consensus in this area.

Since 2003, the HICP has become known as the CPI (consumer price index). Different CPI's are available for particular sectors of the economy such as communications, recreation/culture, transport, alcoholic beverages, tobacco, clothing, footwear, hotels/cafes/restaurants, water and gas/other fuels. The original HICP was based on 1996 i.e. 1996 = 100. All series based on 1996 have been discontinued and a new base was introduced in 2005 = 100. Recently the new base changed to 2015 = 100. The UK *Office of National Statistics* makes the CPI data available in spreadsheet format. The CPI rises less quickly than the RPI because of the way it is calculated. As mentioned earlier, it also excludes housing costs and council tax.

The HICP and RPI are not the only indicators published by the Office for National Statistics in index number form. Among a variety of other indicators, you will find index numbers relating to gross domestic product, purchasing power of Sterling, retail sales (value and volume), all shares index and effective exchange rates.

11.2 Current and Constant Prices

Secondary data sources e.g. data supplied by the UK Office for National Statistics often report financial data recorded at *current prices* and at *constant prices* therefore it is important to distinguish between the two. The data file in Fig. 11.1 represents a time series estimate of UK consumers' expenditure on ciders (£ million, 2008–2017), together with the average price of a litre of cider (£).



	Avprice	Expenditures	YEAR_	DATE_
1	1.66	10857	2008	2008
2	1.83	11904	2009	2009
3	2.06	12888	2010	2010
4	2.20	12927	2011	2011
5	2.31	13225	2012	2012
6	2.39	13941	2013	2013
7	2.50	14283	2014	2014
8	2.59	15005	2015	2015
9	2.69	15932	2016	2016
10	2.79	15872	2017	2017
11				

Fig. 11.1 Current and constant prices data file

The price indices for cider in Fig. 11.1 use 2008 as the base year (written as 2008 = 100). The price index for 2009 is computed as follows:

```
Transform
   Compute
```

And by typing the formula below as Numeric expression as per Fig. 11.2:

$$\textbf{Pindex} = (\textbf{current price}/\textbf{price index})^{*}\ \textbf{100}$$

Consequently, a new variable (Pindex) will be added to the data file as shown in Fig. 11.3.

The expenditure figures in Fig. 11.1 are said to be recorded at *current prices*, which are the actual expenditures at the time of purchase. Despite the annual **rises** in expenditures, a moment's thought should suggest that this increase does not necessarily imply that consumers are drinking more cider. For example, if in 2018, cider cost £2 /l and 100 l were consumed, then expenditure that year would be £200. If in 2001, beer cost £5 /l and only 50 l were consumed, then expenditure that year would be £250. The reason that expenditure has risen is due to the increase in the price of the product, not due to an increase in consumption.

When the impact of price changes is removed from a time series, that series is said to be *deflated*. The process of deflation can be explained by asking the question "what would beer expenditure have been in 2009 if prices had not increased from

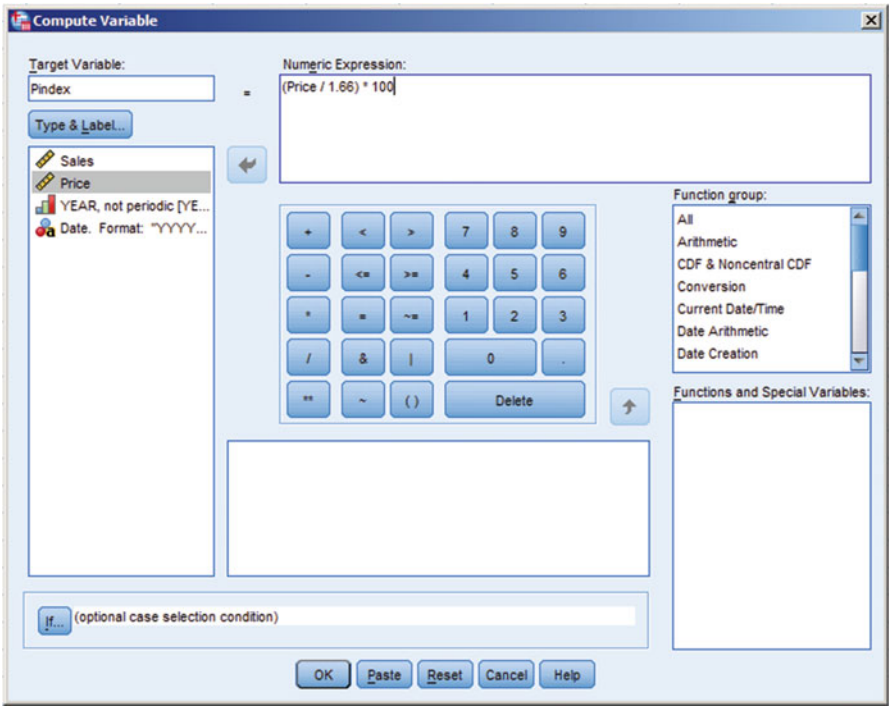
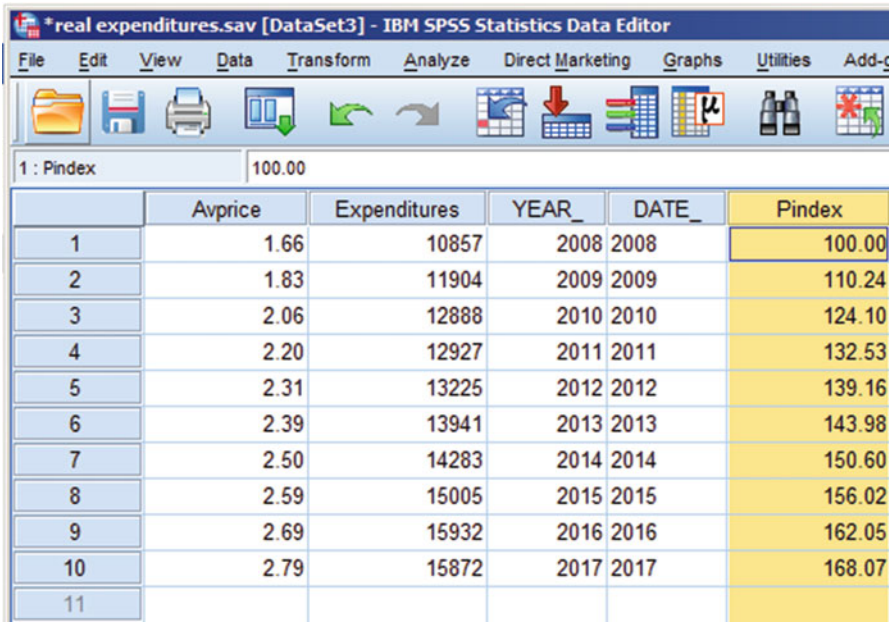


Fig. 11.2 Compute variable dialogue box – Pindex



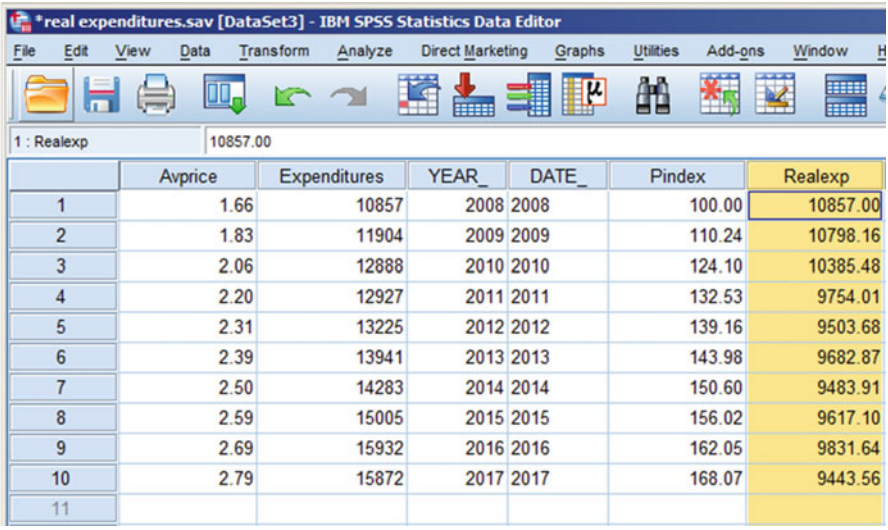
	Avprice	Expenditures	YEAR_	DATE_	Pindex
1	1.66	10857	2008	2008	100.00
2	1.83	11904	2009	2009	110.24
3	2.06	12888	2010	2010	124.10
4	2.20	12927	2011	2011	132.53
5	2.31	13225	2012	2012	139.16
6	2.39	13941	2013	2013	143.98
7	2.50	14283	2014	2014	150.60
8	2.59	15005	2015	2015	156.02
9	2.69	15932	2016	2016	162.05
10	2.79	15872	2017	2017	168.07
11					

Fig. 11.3 Price index variable added to the data file

the base year of 2008?" In 2009, (11,904/1.83) litres (millions) were consumed. At base year prices, expenditure would have been (11,904/1.83)*1.66 = £10798.2 (million) which is a reduction in expenditure when compared with 2008. This latter figure is referred to expenditure at *constant* (or *real*) prices. Similarly, consider 2010; what would have been the expenditure that year if cider prices had not risen since the base year? In 2010, (12,888/2.06) litres (millions) were consumed. At base year prices, expenditure would have been (12,888/2.06)*1.66 = £10385.5 (million) – a still further reduction. In the above instance, the constant price, computed as (12,888/2.06)*1.66, may be written as 12,888 ÷ (2.06/1.66) which indicates that the relationship between current and constant prices is:

$$\textbf{Constant price} = (\textbf{current price}/\textbf{price index})^{*}\ \textbf{100}$$

The expenditure data at real prices (variable name REALEXP) has been computed in IBM SPSS Statistics as reported in Fig. 11.4. **The expenditure data at constant and current prices are plotted overleaf in** Fig. 11.5. It should be noted how significant it is to note whether one if referring to constant or current prices. (Also note that these two values are obviously equal at the base year).



	Avprice	Expenditures	YEAR_	DATE_	Pindex	Realexp
1	1.66	10857	2008	2008	100.00	10857.00
2	1.83	11904	2009	2009	110.24	10798.16
3	2.06	12888	2010	2010	124.10	10385.48
4	2.20	12927	2011	2011	132.53	9754.01
5	2.31	13225	2012	2012	139.16	9503.68
6	2.39	13941	2013	2013	143.98	9682.87
7	2.50	14283	2014	2014	150.60	9483.91
8	2.59	15005	2015	2015	156.02	9617.10
9	2.69	15932	2016	2016	162.05	9831.64
10	2.79	15872	2017	2017	168.07	9443.56
11						

Fig. 11.4 Real expenditures variable added to the data file

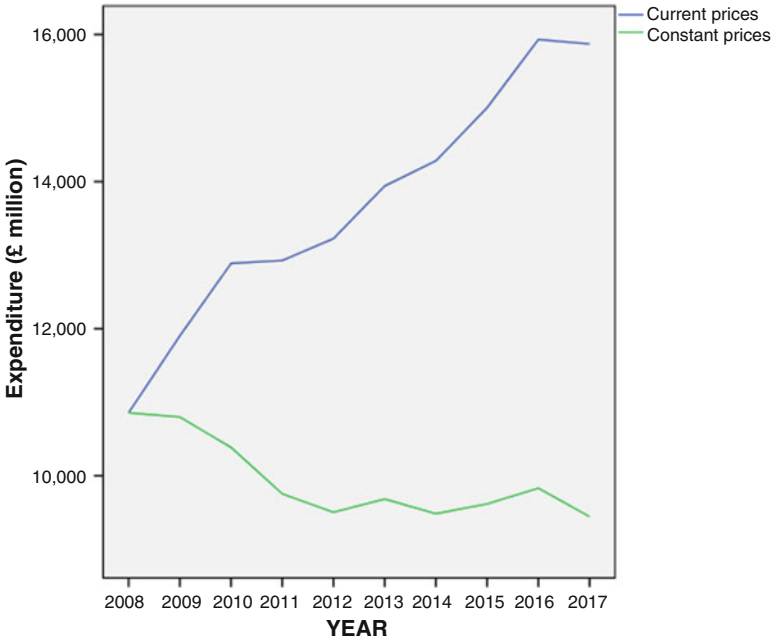


Fig. 11.5 Plot of real expenditures vs current expenditures

References

Aljandali, A. (2014). Exchange rate forecasting: Regional applications to ASEAN, CACM, MERCOSUR and SADC countries. Unpublished PhD thesis, London Metropolitan University, London.

Aljandali, A. (2016). Quantitative Analysis and IBM SPSS Statistics: A Guide for Business and Finance (1st ed.). New York: Springer.

Brace, N., Kemp, R., & Sneglar, R. (2016). *IBM SPSS for psychologists* (6th ed.). Basingstoke: Palgrave Macmillan.

Bryman, A., & Cramer, D. (2011). *Quantitative data analysis with IBM SPSS 17, 18 and 19: A guide for social scientist* (1st ed.). London: Routledge.

Charry, K., & Coussement, K. (2016). *Marketing research with IBM® SPSS statistics: A practical guide* (1st ed.). Oxon: Routledge.

Coshall, J. T. (2008). *SPSS for windows, a user's guide: Volume 2*. Unpublished manuscript, London Metropolitan University, London.

Elliott, A. C., & Woodward, W. A. (2015). *IBM SPSS by example: A practical guide to statistical data analysis* (2nd ed.). Los Angeles: SAGE.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics and sex, drugs and rock 'n' roll* (4th ed.). London: SAGE.

Macinnes, J. (2016). *An introduction to secondary data analysis with IBM SPSS statistics* (1st ed.). Los Angeles: SAGE.

Pallant, J. (2013). *SPSS survival manual: A step by step guide to data analysis using SPSS* (5th ed.). Maidenhead: Open University Press.

Salkind, N. J. (2014). *Statistics for people who (think they) hate statistics* (5th ed.). Los Angeles: SAGE.

Wagner, W. E. (2016). *Using IBM® SPSS® statistics for research methods and social science statistics* (6th ed.). Los Angeles: SAGE.

A. Aljandali, *Multivariate Methods and Forecasting with IBM® SPSS® Statistics*,
Statistics and Econometrics for Finance, DOI 10.1007/978-3-319-56481-4

Index

A. Aljandali, *Multivariate Methods and Forecasting with IBM® SPSS® Statistics*,
Statistics and Econometrics for Finance, DOI 10.1007/978-3-319-56481-4