Chapter 3
Statistical Methods

Paul Taylor
University of Hertfordshire, United Kingdom

3.1. Introduction

This chapter describes a collection of statistical methods. The emphasis is upon what the methods are used to do and upon how to interpret the results. There is little on how to calculate the results, because the algorithms required have already been included in statistical packages for computers and this is how the calculations are performed in practice.

Examples, including computer output and my interpretation of what the output means, are given for some of the more widely used techniques. Other techniques are simply described, in terms of what they can be used to do, along with references to more detailed descriptions of these methods. Presenting examples for all the techniques would have made the chapter far too long.

Section 3.2 describes the most widely used statistical technique, namely regression analysis. Regression analysis is widely used, because there are so many statistical problems that can be presented as finding out how to predict the value of a variable from the values of other variables.

The techniques presented in Section 3.3 are regression analysis techniques for use in specific situations, which arise in practice and are not easy extensions of methods in Section 3.2. In fact, the techniques in Section 3.3.2 and Section 3.3.3 are part of an area of current research in statistical methodology.

Despite being around quite a long time (since the late 1970s or earlier, in most cases) the multivariate analysis techniques of Section 3.4 do not seem to be used as much as they might. When they are used, they are often used inappropriately. It seems likely that these techniques will start to be used more and more, because they are useful and the misconceptions about their use will gradually be eliminated. In particular, with the increase in automatic data collection, the

multivariate methods which aim to reduce the number of variables in a data set, discarding uninformative variables, ought to become more important.

3.2. Generalized Linear Models

The fitting of generalized linear models is currently the most frequently applied statistical technique. Generalized linear models are used to describe the relationship between the mean, sometimes called the *trend*, of one variable and the values taken by several other variables. Modelling this type of relationship is sometimes called *regression*. Regression, including alternatives to generalized linear modelling, is described in Section 3.2.1.

Fitting models is not the whole story in modelling; having fitted several plausible models to a set of data, we often want to select one of these models as being the most appropriate. An objective method for choosing between different models is called *analysis of variance*. Analysis of variance is presented in Section 3.2.2.

Within the generalized linear models, there is a subset of models called *linear models*. Sections 3.2.1 and 3.2.2 concentrate on linear models, because these are the most commonly used of the generalized linear models. Log-linear models and logistic regression models are two other heavily used types of generalized linear model. Section 3.2.3 describes log-linear modelling; Section 3.2.4 describes logistic regression.

Section 3.2.5 is about the analysis of survival data. The models used in the analysis of survival data are not generalized linear models. The reason that they are included here is that the techniques used for fitting generalized linear models can also be applied to the fitting of models in the analysis of survival data.

3.2.1 Regression

Regression analysis is the process of determining how a variable, y, is related to one, or more, other variables, $x_1, x_2, \ldots, x_n$. The y variable is usually called the *response* by statisticians; the x_i's are usually called the *regressors* or simply the *explanatory variables*. Some people call the response the *dependent variable* and the regressors the *independent variables*. People with an electrical-engineering background often refer to the response as the *output* and the regressors as the *inputs*. Here, we will use the terms output and inputs. Common reasons for doing a regression analysis include:

- the output is expensive to measure, but the inputs are not, and so cheap predictions of the output are sought;
- the values of the inputs are known earlier than the output is, and a working prediction of the output is required;
- we can control the values of the inputs, we believe there is a causal link between the inputs and the output, and so we want to know what values of the inputs should be chosen to obtain a particular target value for the output;

- it is believed that there is a causal link between some of the inputs and the output, and we wish to identify which inputs are related to the output.

The most widely used form of regression model is the *(general) linear model.* The linear model is usually written as

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_n x_{nj} + \varepsilon_j \quad j = 1, 2, \ldots, m \tag{3.1}$$

where the ε_j's are independently and identically distributed as $\mathcal{N}(0, \sigma^2)$ and m is the number of data points.

The linear model is called *linear*, because the expected value of y_j is a linear function, or weighted sum, of the β's. In other words, we can write the expected value of y_j as

$$\mathrm{E}(y_j) = \beta_0 + \sum_{i=1}^{n} \beta_i x_{ij}. \tag{3.2}$$

The main reasons for the use of the widespread use of the linear model are given below.

- The maximum likelihood estimators of the β's are the same as the least squares estimators; see Section 2.4 of Chapter 2.
- There are explicit formulae for the least squares estimators of the β's. This means that the estimates of the β's can be obtained without electronic computers, so the use of the linear model was common in agricultural research in the 1920s and the chemical industry in the 1930s (in the UK, at least). There are also rapid and reliable numerical methods for obtaining these estimates using a computer; see chapter 3 of [511].
- There is a large class of regression problems that can be converted into the form of the general linear model. For example, a polynomial relationship, such as
$$\mathrm{E}(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2, \tag{3.3}$$
can be converted to the form in (3.2) by setting $x_3 = x_1 x_2$, $x_4 = x_1^2$ and $x_5 = x_2^2$. (This conversion would not usually be explicitly carried out.)
For a wide range of examples of the use of the linear model see, for example [157].
- Even when the linear model is not strictly appropriate, there is often a way to transform the output and/or the inputs, so that a linear model can provide useful information.

Another major attraction of the linear model is a technical point: it is possible to determine the statistical properties of virtually any object that we might derive from the linear model. In other words, we can perform statistical inference for the linear model. In particular, we can use hypothesis testing to compare different linear models (see Section 3.2.2) and we can obtain interval estimates for predictions and for the β's.

The linear model is the main type of model used in regression. We will return to the linear model in Section 3.2.2, but now we will look at other types of regression models.

Non-linear Regression A non-linear regression model differs from a linear model in that the expected output in not a weighted sum of the inputs. An example of a non-linear model is the *allometric* model,

$$y_j = \beta_0 x_{1j}^{\beta_1} + \varepsilon_j \quad j = 1, 2, \ldots, m \tag{3.4}$$

where the ε's and m are as in (3.1). This model is known to represent the relationship between the weight of part of a plant, such as its root system, and the weight of the whole plant. Another example is the *Mitscherlich* model for relating crop yield to the amount of fertiliser applied to the crop,

$$y_j = \beta_0 \left(1 - e^{-\beta_1(x_{1j}+\beta_2)}\right) + \varepsilon_j \quad j = 1, 2, \ldots, m. \tag{3.5}$$

Here, y_j is the crop yield for the jth plot in a field, x_{1j} is the amount of fertiliser applied to the jth plot in the field, β_0 represents an upper limit on crop yield, β_2 represents the amount of fertiliser already available in the soil, while β_1 is related to how quickly the upper limit for yield is reached.

The drawbacks to using non-linear regression boil down to the fact that there is no general algebraic solution to finding the least squares estimators of the β's. (As for the linear model, least squares estimation will produce the maximum likelihood estimators, because the ε's have identical, independent normal distributions.) The problems that arise because of the lack of a general algebraic solution are as follows.

1. Estimation is carried out using iterative methods which require good choices of starting values, might not converge, might converge to a local optimum rather than the global optimum, and will require human intervention to overcome these difficulties.
2. The statistical properties of the estimates and predictions from the model are not known, so we cannot perform statistical inference for non-linear regression.
   (This is not strictly true; we could perform statistical inference using techniques such as the bootstrap, which is explained in Section 2.6.3 of Chapter 2.)

Non-linear models are usually used when the form of the relationship between the output and the inputs is known and the primary aim of the study is determination of the β's. In this case, the β's represent fundamental biological, chemical or physical constants and are of interest for their own sake. An example of this is β_0 in the Mitscherlich model at (3.5). If prediction of the output is the primary aim, then it is usually possible to use a linear model to make quite good predictions for some range of interest.

An introduction to non-linear modelling can be found in chapter 10 of [157].

Generalized Linear Models The linear model is very flexible and can be used in many situations, even when it is not strictly valid. Despite this, there is a stage where the linear model cannot do what we need. The non-linear model

above is one way to expand the possibilities of regression. The *generalized linear model* offers a much greater expansion of the problems that can be addressed by regression. The definitive reference for this topic is [368].

The generalization is in two parts.

1. The distribution of the output does not have to be the normal, but can be any of the distributions in the exponential family; see, for example [368, p. 28].
2. Instead of the expected value of the output being a linear function of the β's, we have

$$g\left(\mathrm{E}(y_j)\right) = \beta_0 + \sum_{i=1}^{n} \beta_i x_{ij} \tag{3.6}$$

where $g(\cdot)$ is a monotone differentiable function. The function $g(\cdot)$ is called the *link* function.

So, using this generalization we can branch out to consider output distributions such as the binomial or the Poisson (see Chapter 2), or the gamma [1], as well as the normal. It is also possible to restrict the range of values predicted for the expected output, to capture a feature such as an asymptote. Some non-linear models can be framed as generalized linear models. The allometric model at (3.4) can, by choosing a log link function (and a normal output distribution).

A major step forward made by [368] was the development of a general algorithm for fitting generalized linear models. The algorithm used by [368] is iterative, but has natural starting values and is based on repeated application of least squares estimation. The generalized linear model also gives us a common method for making statistical inferences, which we will use in Sections 3.2.3 and 3.2.4.

Examples of widely used generalized linear models are given in Sections 3.2.3, 3.2.4 and 3.2.5.

Generalized Additive Models Generalized additive models are a generalization of generalized linear models. The generalization is that $g\left(\mathrm{E}(y_j)\right)$ need not be a linear function of a set of β's, but has the form

$$g\left(\mathrm{E}(y_j)\right) = \beta_0 + \sum_{i=1}^{n} s_i(x_{ij}) \tag{3.7}$$

where the s_i's are arbitrary, usually smooth, functions. In particular, the scatterplot smoothers that have been one focus of statistical research for the last fifteen

[1] The gamma distribution is a continuous distribution that is not symmetric about its population mean, and only takes positive values. The exponential distribution of Section 2.2.16 in Chapter 2 is a special case of the gamma distribution. The gamma distribution with parameters n and λ, written $\Gamma(n, \lambda)$, is the distribution of the sum of n independent exponential random variables, all with rate parameter λ. So, the $\Gamma(1, \lambda)$ distribution is the same as the exponential distribution with parameter λ.

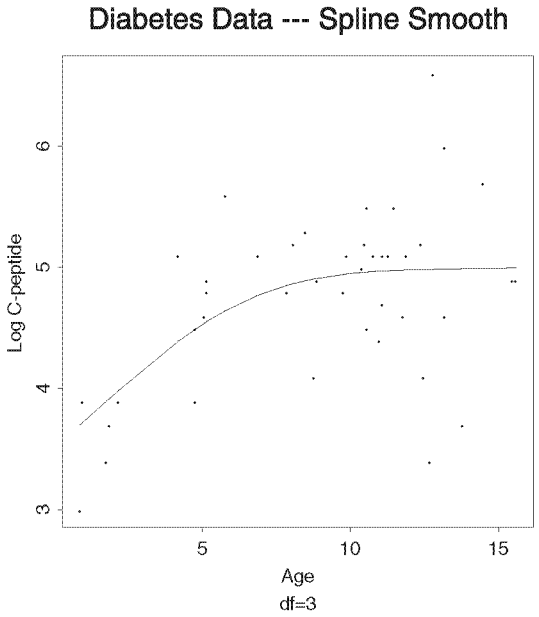

Fig. 3.1. An example of a regression model fitted using one type of scatterplot smoother; the form of the relationship between the output and the input is allowed to vary over the range of the input. The specific type of smoother used is called a *cubic B-spline* with 3 degrees of freedom.

years can be used. An example of the model produced using a type of scatterplot smoother is shown in Figure 3.1. All that has been specified about the shape of the fitted curve in Figure 3.1 is that it has to be smooth (has a continuous second derivative) and must not change direction frequently, that is it must not contain high frequency oscillations (it must not be 'wiggly'). This means that we do not have to specify the functional form of the relationship between the output and the inputs, but can use local regression methods to identify the functional form.

Methods for fitting generalized additive models exist and are generally reliable. The main drawback is that the framework of statistical inference that is available for generalized linear models has not yet been developed for generalized additive models. Despite this drawback, generalized additive models can be fitted by several of the major statistical packages already. The definitive reference for this topic is [249].

3.2.2 Analysis of Variance

The *analysis of variance*, or ANOVA, is primarily a method of identifying which of the β's in a linear model are non-zero. This technique was developed for the analysis of agricultural field experiments, but is now used quite generally.

Table 3.1. Yields of turnips (in kg) from an experiment carried out at Sonning on a farm owned by The University of Reading. The experiment had a *randomized block* design, with four blocks, each of size sixteen. There are sixteen different treatment combinations, labelled *A–R*, each appearing once in each block. Each treatment is a combination of turnip variety (Barkant or Marco), sowing date (21/8/90 or 28/8/90) and seed density (1, 2, 4 or 8 kg/ha).

Treatments				Blocks			
Variety	Date	Density	Label	I	II	III	IV
Barkant	21/8/90	1 kg/ha	A	2.7	1.4	1.2	3.8
		2 kg/ha	B	7.3	3.8	3.0	1.2
		4 kg/ha	C	6.5	4.6	4.7	0.8
		8 kg/ha	D	8.2	4.0	6.0	2.5
	28/8/90	1 kg/ha	E	4.4	0.4	6.5	3.1
		2 kg/ha	F	2.6	7.1	7.0	3.2
		4 kg/ha	G	24.0	14.9	14.6	2.6
		8 kg/ha	H	12.2	18.9	15.6	9.9
Marco	21/8/90	1 kg/ha	J	1.2	1.3	1.5	1.0
		2 kg/ha	K	2.2	2.0	2.1	2.5
		4 kg/ha	L	2.2	6.2	5.7	0.6
		8 kg/ha	M	4.0	2.8	10.8	3.1
	28/8/90	1 kg/ha	N	2.5	1.6	1.3	0.3
		2 kg/ha	P	5.5	1.2	2.0	0.9
		4 kg/ha	Q	4.7	13.2	9.0	2.9
		8 kg/ha	R	14.9	13.3	9.3	3.6

Example 3.1 (Turnips for Winter Fodder). The data in Table 3.1 are from an experiment to investigate the growth of turnips. These types of turnip would be grown to provide food for farm animals in winter. The turnips were harvested and weighed by staff and students of the Departments of Agriculture and Applied Statistics of The University of Reading, in October, 1990.

The blocks correspond to four different pieces of land. These pieces of land have different characteristics and so these four areas are expected to produce different amounts of turnips.

The four areas of land are all the same size and have each been split into sixteen identical plots; this gives a total of sixty-four plots. Each plot has a treatment applied to it; there are sixteen treatments. Each of the sixteen treatments is a combination of three properties: the variety (type) of turnip sown (*Barkant* or *Marco*); when the seed was sown (*21/8/90* or *28/8/90*); how much seed was sown (*1*, *2*, *4* or *8* kg/ha).

The following linear model

$$y_j = \beta_0 + \beta_B x_{Bj} + \beta_C x_{Cj} + \cdots + \beta_R x_{Rj} + \beta_{\mathrm{II}} x_{\mathrm{II},j} + \beta_{\mathrm{III}} x_{\mathrm{III},j} + \beta_{\mathrm{IV}} x_{\mathrm{IV},j} + \varepsilon_j \quad j = 1, 2, \ldots, 64 \tag{3.8}$$

or an equivalent one could be fitted to these data. The inputs take the values 0 or 1 and are usually called *dummy* or *indicator* variables. The value of x_{Cj}

would be 1 if treatment C (Barkant, sown on 21/8/90 at 4 kg/ha) were applied to plot j, but zero otherwise. Similarly, the value of $x_{\mathrm{III},j}$ would be 1 if plot j were in block III, but zero otherwise.

On first sight, (3.8) should also include a β_A and a β_{I}. Indeed, many statisticians would write this model with terms in β_A and β_{I}. If β_A and β_{I} are included, then the model is over-parameterised. Over-parameterised means that there are more parameters (β's) than are necessary to describe the structure in the expected values. For the turnip experiment we need a total of nineteen parameters: a baseline value; fifteen parameters to specify the **differences** between the expected yields for the sixteen treatments; three parameters to specify the **differences** between the expected yields for the four blocks. If β_A and β_{I} are included, then there are twenty-one parameters, which is two more than necessary.

The choice of model specified by (3.8) implies the following interpretations for the β's. The baseline value, β_0, is the expected turnip yield for a plot in block I that has received treatment A. The expected yield for a plot in block I that has received treatment B is $\beta_0 + \beta_B$, that is, β_B is the difference between the expected yields for treatment B and treatment A; similarly for treatments C to R. The expected yield for a plot in block II that has received treatment A is $\beta_0 + \beta_{\mathrm{II}}$, so β_{II} is the difference in expected yields for block II and block I; similarly for blocks III and IV. Finally, the expected yield for a plot that is neither in block I nor in receipt of treatment A, such as block III with treatment G, would be the baseline, plus the expected increase for using treatment G rather than treatment A, plus the increase for being in block III rather than block I. Algebraically, this is $\beta_0 + \beta_G + \beta_{\mathrm{III}}$.

The first question that we would try to answer about these data is

Does a change in treatment produce a change in the turnip yield?

which is equivalent to asking

Are any of β_B, β_C, ..., β_R *non-zero?*

which is the sort of question that can be answered using ANOVA.

This is how the ANOVA works. Recall, the general linear model of (3.1),

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_n x_{nj} + \varepsilon_j \quad j = 1, 2, \ldots, m$$

where the ε_j's are independently and identically distributed as $\mathcal{N}(0, \sigma^2)$ and m is the number of data points. Estimates of the β's are found by least squares estimation, which happens to be the same as maximum likelihood estimation for this model. Writing the estimate of β_i as $\hat{\beta}_i$, we can define the *fitted values*

$$\hat{y}_j = \hat{\beta}_0 + \sum_{i=1}^{n} \hat{\beta}_i x_{ij}. \tag{3.9}$$

The fitted values are what the fitted model would predict the output to be for the inputs in the set of data. The residuals are the differences between the observed output values and the fitted values,

$$r_j = y_j - \hat{y}_j. \tag{3.10}$$

The size of the residuals is related to the size of σ^2, the variance of the ε_j's. It turns out that we can estimate σ^2 by

$$S^2 = \frac{\sum_{j=1}^{m}(y_j - \hat{y}_j)^2}{m - (n+1)} \tag{3.11}$$

assuming that the model is not over-parameterised. The numerator of the right-hand side of (3.11) is called the *residual sum of squares*, while the denominator is called the *residual degrees of freedom.*

The key facts about S^2 that allow us to compare different linear models are:

- if the fitted model is adequate ('the right one'), then S^2 is a good estimate of σ^2;
- if the fitted model includes redundant terms (that is includes some β's that are really zero), then S^2 is still a good estimate of σ^2;
- if the fitted model does not include one or more inputs that it ought to, then S^2 will tend to be larger than the true value of σ^2.

So if we omit a useful input from our model, the estimate of σ^2 will shoot up, whereas if we omit a redundant input from our model, the estimate of σ^2 should not change much. The use of this concept is described in detail below. Note that omitting one of the inputs from the model is equivalent to forcing the corresponding β to be zero.

Suppose we have fitted a particular linear model, Ω_1, to a set of data. We can estimate the variance, σ^2, of the ε_j's, as the ratio

$$S_1^2 = \frac{RSS_1}{\nu_1}, \tag{3.12}$$

where RSS_1 and ν_1 are the residual sum of squares and the residual degrees of freedom, respectively, for model Ω_1.

As long as Ω_1 includes the inputs that are associated with the output, S_1^2 will be a reasonable estimate of σ^2. In particular, it is a valid estimate even if some (or all) of the β's are zero.

If we want to investigate whether certain β's are zero, then we can fit another model, Ω_0, in which those β's and the corresponding inputs are absent. We will be able to obtain another estimate of the variance

$$S_0^2 = \frac{RSS_0}{\nu_0}, \tag{3.13}$$

where RSS_0 and ν_0 are the residual sum of squares and the residual degrees of freedom, respectively, for model Ω_0.

Now, if the β's that are absent from Ω_0 really are zero, then S_0^2 and S_1^2 are both valid estimates of σ^2 and should be approximately equal. On the other hand, if some of the β's that are absent from Ω_0 are not really zero, then S_1^2 is still a valid estimate of σ^2, but S_0^2 will tend to be bigger than σ^2. The further the absent β's are from zero, the bigger S_0^2 is.

In principle then, we can compare S_0^2 with S_1^2. If S_0^2 is big relative to S_1^2, then some of the absent β's are non-zero and Ω_0 is an inadequate model. If S_0^2 and S_1^2 are similar, then this suggests that Ω_0 is a satisfactory model.

In fact we can make a much better comparison by forming a third estimate of σ^2. We calculate the extra sum of squares

$$ESS = RSS_0 - RSS_1, \tag{3.14}$$

the extra degrees of freedom

$$\nu_E = \nu_0 - \nu_1 \tag{3.15}$$

and our third estimate of σ^2 as

$$S_E^2 = \frac{ESS}{\nu_E}. \tag{3.16}$$

This third estimate, S_E^2, shares some of the properties of S_0^2 in that it is an estimate of σ^2 if Ω_0 is an adequate model, but will be higher than σ^2 if some of the absent β's are non-zero. The advantage of S_E^2 over S_0^2 is that it is much more sensitive to the absent β's being non-zero than S_0^2 is. (In addition, S_E^2 and S_1^2 are statistically independent, which simplifies the statistical theory required.)

We can carry out a hypothesis test of whether Ω_0 is adequate (versus the alternative that it is not, which implies that we would prefer to use Ω_1) by forming the following test statistic

$$F = \frac{S_E^2}{S_1^2} = \frac{ESS/\nu_E}{RSS_1/\nu_1} \tag{3.17}$$

which has various names, including the 'F-statistic', the 'F-ratio' and the 'variance ratio'.

If Ω_0 is an adequate model for the data, then F should be approximately 1 and will have an F-distribution[2] with parameters, ν_E and ν_1; if Ω_0 is not adequate then F should be large.

So we can carry out a formal hypothesis test of the null hypothesis that Ω_0 is the correct model versus the alternative hypothesis that Ω_1 is the correct model, by forming the F-statistic and comparing it with the appropriate F-distribution; we reject the null hypothesis if the F-statistic appears to be too big to come from the appropriate F-distribution. A good rule-of-thumb is that an F-statistic of over 4 will lead to rejection of Ω_0 at the 5% level of significance.

[2] The F-distribution is a well-known continuous distribution. Most statistical packages can generate the associated probabilities, which are tabulated in many elementary statistics books. If $X_1, X_2, \ldots, X_{\nu_1}$ and $W_1, W_2, \ldots, W_{\nu_2}$ are statistically independent standard normal random variables, then the ratio

$$\frac{(X_1^2 + X_2^2 + \cdots + X_{\nu_1}^2)/\nu_1}{(W_1^2 + W_2^2 + \cdots + W_{\nu_2}^2)/\nu_2}$$

will have an F-distribution with parameters (*degrees of freedom*) ν_1 and ν_2.

Table 3.2. ANOVA table for fitting the model at (3.18) to the data in Table 3.1.

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
block	3	163.737	54.57891	2.278016	0.08867543
Residuals	60	1437.538	23.95897		

Table 3.3. ANOVA table for fitting the model at (3.8) to the data in Table 3.1.

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
block	3	163.737	54.57891	5.690430	0.002163810
treat	15	1005.927	67.06182	6.991906	0.000000171
Residuals	45	431.611	9.59135		

Example 3.2 (Turnips for Winter Fodder continued). To make this idea more concrete, we can take Ω_1 to be the model at (3.8) on page 75, and Ω_0 to be the following model

$$y_j = \beta_0 + \beta_{\mathrm{II}} x_{\mathrm{II},j} + \beta_{\mathrm{III}} x_{\mathrm{III},j} + \beta_{\mathrm{IV}} x_{\mathrm{IV},j} + \varepsilon_j \quad j = 1, 2, \ldots, 64. \tag{3.18}$$

So here, Ω_0 is the special case of Ω_1 in which all of β_B, β_C, ..., β_R are zero.

The ANOVA tables for models Ω_0 and Ω_1 are shown in Tables 3.2 and 3.3. These tables are exactly how they appear in the output from a statistical package. The column labelled 'Df' contains the degrees of freedom, 'Sum of Sq' is the sum of squares column. The entries in the 'Mean Sq' column are obtained by dividing the entries in the sum of squares column by their corresponding degrees of freedom. The label, 'Mean Sq', is an abbreviation of *mean square*. We come back to the meanings of the other columns below.

The numbers that we are interested in are in the row labelled 'Residuals' in each table. We can read the 'Residuals' rows of these two tables, to obtain

$$RSS_0 = 1437.5, \quad \nu_0 = 60 \quad \text{and} \quad s_0^2 = \frac{1437.5}{60} = 23.96$$

(from Table 3.2) and

$$RSS_1 = 431.6, \quad \nu_1 = 45 \quad \text{and} \quad s_1^2 = \frac{431.6}{45} = 9.59$$

(from Table 3.3). These estimates of σ^2 are quite different, which suggests that one or more of the inputs that were omitted from Ω_0 were necessary after all. The formal statistical hypothesis testing procedure to determine whether Ω_0 is adequate is completed below.

From here, we could find the extra sum of squares and extra degrees of freedom, then calculate the F-statistic and compare it with the relevant F-distribution. We do not need to do this calculation ourselves, as the computer

does it for us in producing Table 3.3. The extra sum of squares and degrees of freedom are presented in the row labelled 'treat' in Table 3.3. We can read the 'treat' row of Table 3.3 to get

$$ESS = RSS_0 - RSS_1 = 1437.5 - 431.6 = 1005.9,$$

$$\nu_E = \nu_0 - \nu_1 = 60 - 45 = 15$$

and

$$s_E^2 = \frac{1005.9}{15} = 67.06.$$

Notice that the extra degrees of freedom, ν_E, is the same as the number of additional inputs in Ω_1 compared to Ω_0.

The procedure for obtaining the F-statistic is quite automatic; the column labelled 'F Value' is simply the 'Mean Sq' column divided by the residual mean square. So, we find the F-statistic is

$$\frac{s_E^2}{s_1^2} = \frac{67.06}{9.59} = 6.99$$

and the significance probability ("p-value") can be read from the final column, labelled 'Pr(F)'. The p-value is less than $\frac{1}{10{,}000}$ of one percent.

Recall (from Section 2.4.1 of Chapter 2) that, the p-value is the (conditional) probability of the test-statistic being so extreme, if the null hypothesis is true. Our null hypothesis is that Ω_0 is an adequate description of the variation in the turnip yields. So, the p-value that we have obtained tells us that **if** Ω_0 is an adequate model, **then** the probability of getting an F-statistic as high as 6.99 is less than $\frac{1}{10{,}000}\%$; in other words, getting an F-statistic of 6.99 is amazing if Ω_0 is the right model. On the other hand, if Ω_0 is inadequate, then we would expect the F-statistic to be inflated. The rational choice between these alternatives is to decide that Ω_0 is not a good model for the turnip yields, because the observed F-statistic is nothing like what it ought to be if Ω_0 were adequate. This tells me that model Ω_0 is inadequate, so I should use model Ω_1.

Having chosen Ω_1, in preference to Ω_0, I conclude that the different treatments produce systematic differences in the turnip yield. I draw this conclusion because the difference between Ω_0 and Ω_1 is that Ω_1 allows for treatment to alter yield, but Ω_0 does not.

Some might wonder where the 'block' row in the ANOVA tables has come from. The 'block' row in Table 3.3 comes from working out the extra sums of squares and degrees of freedom for comparing the model at (3.18) with the model at (3.19).

$$y_j = \beta_0 + \varepsilon_j \quad j = 1, 2, \ldots, 64. \tag{3.19}$$

The test for 'blocks' indicates that the yield is systematically different for different blocks, but we knew this already, so the result of this test is of no interest.

It might be thought that that is the end of the analysis for the turnip data. In some ways it is, as we have established that the different treatments result in

Table 3.4. ANOVA table that would be produced in the real-world analysis of the data in Table 3.1.

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
block	3	163.7367	54.5789	5.69043	0.0021638
variety	1	83.9514	83.9514	8.75282	0.0049136
sowing	1	233.7077	233.7077	24.36650	0.0000114
density	3	470.3780	156.7927	16.34730	0.0000003
variety:sowing	1	36.4514	36.4514	3.80045	0.0574875
variety:density	3	8.6467	2.8822	0.30050	0.8248459
sowing:density	3	154.7930	51.5977	5.37960	0.0029884
variety:sowing:density	3	17.9992	5.9997	0.62554	0.6022439
Residuals	45	431.6108	9.5914		

different yields. If I had not been illustrating the technique of ANOVA, I would have approached the analysis differently. I would be trying to find out what it was about the treatments that made a difference to the yield. I would investigate this by looking at the ANOVA for a sequence of nested models.

Table 3.4 shows the ANOVA that would usually be produced for the turnip data. Notice that the 'block' and 'Residuals' rows are the same as in Table 3.3. The basic difference between Tables 3.3 and 3.4 is that the treatment information is broken down into its constituent parts in Table 3.4.

The 'variety' row in Table 3.4 corresponds to the extra sum of squares and degrees of freedom for comparing the model at (3.18) with

$$\begin{aligned} y_j = \beta_0 + \beta_{\text{marco}} x_{\text{marco},j} \\ + \beta_{\text{II}} x_{\text{II},j} + \beta_{\text{III}} x_{\text{III},j} + \beta_{\text{IV}} x_{\text{IV},j} + \varepsilon_j \quad j = 1, 2, \ldots, 64 \end{aligned} \tag{3.20}$$

which allows for differences in yield between the two varieties. The variable x_{marco} is an indicator variable, like those used in (3.8). Similarly, the 'sowing' row corresponds to the extra sum of squares and degrees of freedom for comparing the model at (3.20) with one that allows for difference between sowing dates too. The remaining rows in Table 3.4 all correspond to the introduction of new inputs to make the model more complex. All the F-statistics are worked out using the estimate of variance obtained from the 'Residuals' row of the ANOVA.

The rows with labels such as 'variety:sowing' are the extra sums of squares and degrees of freedom corresponding to including inputs that are products of other inputs. This sort of term is called an *interaction*. As an example, take the inputs corresponding to *variety*, *sowing* and *variety:sowing*. These are, respectively:

- x_{marco}, the indicator variable that *Marco* turnips have been used;
- x_{28}, the indicator variable that sowing date was 28/8/90;
- x_{m28}, the indicator variable that the turnips were of type *Marco* and were sown on 28/8/90—$x_{\text{m28},j} = x_{\text{marco},j} \times x_{28,j}$.

Table 3.5. Variety means for the data in Table 3.1.

Barkant	Marco
6.522	4.231

Table 3.6. Sowing date by density means for the data in Table 3.1.

	Density			
Date	1kg/ha	2kg/ha	4kg/ha	8kg/ha
21/8/90	1.7625	3.0125	3.9125	5.1750
28/8/90	2.5125	3.6875	10.7375	12.2125

The reasons for using this sort of input are illustrated in the discussion of Table 3.6, below. Note that various other statistical packages use '.', '*' or '×', instead of ':' to denote interactions.

Based on the significance probabilities (the column labelled 'Pr(F)') in Table 3.4 most statisticians would identify *variety*, *sowing*, *density* and *sowing:density* as the terms that make a difference to turnip yield. The presence of the *sowing:density* interaction in the model means that we cannot consider *sowing* or *density* in isolation. Instead, we have to look at the combinations of *sowing* and *density*. There are no interactions involving *variety* in our set of terms related to turnip yield, so we can think of it in isolation.

Tables 3.5 and 3.6 show the relevant patterns: *Barkant* produces about 2.3 kg more per plot than *Marco* does regardless of sowing date and density; for the two lowest sowing densities, delaying the sowing by one week adds about 0.7 kg to the yield per plot, but for the two highest sowing densities it adds about 7 kg per plot. For the statistician this is close to the end of the analysis. The fact that the increase in yield due to delaying the sowing by a week is different for different sowing densities is the reason for the inclusion of the *sowing:density* interaction. This interaction could only be omitted from the model if the increase due to delaying sowing was the same regardless of the sowing density. There are three inputs corresponding to the *sowing:density* interaction. These inputs are calculated by forming the products of each of the three inputs corresponding to *density* with the input corresponding to *sowing*.

In this section, we have seen how the ANOVA can be used to identify which inputs are related to the output and which are not. The basic idea is to produce different estimates of the variance, σ^2, based on different assumptions, and to compare these estimates in order to decide which assumptions are true. The ANOVA procedure is only valid if the models being compared are nested, in other words, one model is a special case of the other.

3.2.3 Log-linear Models

Log-linear modelling is a way of investigating the relationships between categorical variables. Categorical variables are non-numeric variables and are sometimes called *qualitative* variables; the set of values that a categorical variable can take are called its *categories* or *levels*. There are two types of categorical variables: *nominal* variables have no ordering to their categories; the categories of an *ordinal* variable have a natural ordering, such as *none–mild–moderate–severe*. The data shown in Table 3.7 show the sort of problem attacked by log-linear modelling. There are five categorical variables displayed in Table 3.7:

centre is one of three health centres for the treatment of breast cancer;
age is the age of the patient when her breast cancer was diagnosed;
survived is whether the patient survived for at least three years from the date of diagnosis;
appear the visual appearance of the patient's tumour—either *malignant* or *benign*;
inflam the amount of inflammation of the tumour—either *minimal* or *greater*.

The combination of *appear* and *inflam* represents the state of the tumour, or the disease progression. The entries in the tables are numbers of patients in each group. For these data, the output is the number of patients in each cell.

A log-linear model is a type of generalized linear model; the output, Y_i, is assumed to have a Poisson distribution with expected value μ_i. The (natural) logarithm of μ_i is assumed to be linear in the β's. So the model is

$$y_j \sim \text{Pois}(\mu_j) \quad \text{and} \quad \log(\mu_j) = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_n x_{nj}. \qquad (3.21)$$

Since all the variables of interest are categorical, we need to use indicator variables as inputs in the same way as in (3.8).

The aim in log-linear modelling is to identify associations between the categorical variables. Associations correspond to interaction terms in the model. So, our problem becomes that of finding out which β's are zero; this is the same problem that we used the ANOVA to solve for linear models in Section 3.2.2. For the log-linear model, and indeed all generalized linear models, we consider a quantity called *deviance* when we are trying to compare two models. Deviance is analogous to the sums of squares in an ANOVA. (In fact, if the generalized linear model happens to be a linear model, the deviance and the sum of squares are the same thing.) We can identify which β's are zero by considering an analysis of deviance table, such as that in Table 3.8. As in Section 3.2.2, this table is in exactly the form that a statistical package produced.

The analysis of deviance works in a similar way to the ANOVA. Here we will consider how to interpret the results of analysis of deviance. No attempt will be made to explain the underlying statistical theory; the relevant theory and derivations can be found in [368, chapter 6].

The change in deviance for adding the interaction between *inflam* and *appear* is 95.4381 (from Table 3.8). The change in deviance is the counterpart of the extra

Table 3.7. Incidence of breast cancer at three different health centres. The data are taken from [391].

			State of Tumour			
			Minimal Inflammation		Greater Inflammation	
Centre	Age	Survived	Malignant Appearance	Benign Appearance	Malignant Appearance	Benign Appearance
Tokyo	Under 50	No	9	7	4	3
		Yes	26	68	25	9
	50–69	No	9	9	11	2
		Yes	20	46	18	5
	70 or over	No	2	3	1	0
		Yes	1	6	5	1
Boston	Under 50	No	6	7	6	0
		Yes	11	24	4	0
	50–69	No	8	20	3	2
		Yes	18	58	10	3
	70 or over	No	9	18	3	0
		Yes	15	26	1	1
Glamorgan	Under 50	No	16	7	3	0
		Yes	16	20	8	1
	50–69	No	14	12	3	0
		Yes	27	39	10	4
	70 or over	No	3	7	3	0
		Yes	12	11	4	1

sum of squares. For the ANOVA, we take the extra sum of squares and divide by its degrees of freedom to obtain an estimate of the variance, which we compare with another estimate of variance (by forming the F-statistic and finding its p-value). For the analysis of deviance we do not need to find an estimate of variance, because the approximate distribution of the change in deviance is known.

If the *inflam:appear* interaction is superfluous, then the change in deviance will have a distribution that is approximately chi-squared[3], with one degree of freedom. The number of degrees of freedom is read from the same row of Table 3.8 as the change in deviance. So, by comparing 95.4381 with the χ^2_1 distribution and noting that 95.4381 is incredibly high for an observation from that distribution, we can say that the change in deviance is massively higher than it should be if the

[3] The chi-squared distribution is a well-known continuous distribution. Most statistical packages can generate the associated probabilities, which are tabulated in many elementary statistics books. If $X_1, X_2, \ldots, X_\nu$ are statistically independent standard normal random variables, then the sum

$$(X_1^2 + X_2^2 + \cdots + X_\nu^2)$$

will have an chi-squared distribution with ν degrees of freedom. The chi-squared distribution with ν degrees of freedom is often written χ^2_ν.

Table 3.8. An analysis of deviance table for fitting log-linear models to the data in Table 3.7.

```
Terms added sequentially (first to last)
                Df Deviance Resid. Df Resid. Dev   Pr(Chi)
          NULL                    71   860.0076
        centre  2   9.3619         69   850.6457 0.0092701
           age  2 105.5350         67   745.1107 0.0000000
      survived  1 160.6009         66   584.5097 0.0000000
        inflam  1 291.1986         65   293.3111 0.0000000
        appear  1   7.5727         64   285.7384 0.0059258
    centre:age  4  76.9628         60   208.7756 0.0000000
centre:survived 2  11.2698         58   197.5058 0.0035711
  centre:inflam 2  23.2484         56   174.2574 0.0000089
  centre:appear 2  13.3323         54   160.9251 0.0012733
   age:survived 2   3.5257         52   157.3995 0.1715588
     age:inflam 2   0.2930         50   157.1065 0.8637359
     age:appear 2   1.2082         48   155.8983 0.5465675
survived:inflam 1   0.9645         47   154.9338 0.3260609
survived:appear 1   9.6709         46   145.2629 0.0018721
  inflam:appear 1  95.4381         45    49.8248 0.0000000
```

inflam:appear interaction is redundant, therefore we can say that the interaction between *inflam* and *appear* ought to be included in the model. (The significance probability is presented in the final column of the table and turns out to be less than $5 \times 10^{-6}\%$, that is, zero to seven decimal places.)

Applying this procedure to each row of the analysis of deviance table leads me to conclude that the following interactions can be omitted from the model: `age:survived`, `age:inflam`, `age:appear` and `survived:inflam`. Strictly, we should be more careful about this, as the changes in deviance vary with the order in which the terms are included. It turns out in this particular case that this quick method leads to the answer that I would have got by a more careful approach.

The model that I would select contains all five main effects, all the two-way interactions involving *centre*, the interaction between *appear* and *inflam*, and the interaction between *appear* and *survived*. This means that: all the categorical variables are associated with *centre*; *appear* and *inflam* are associated; *appear* and *survived* are associated. To summarise this model, I would construct its conditional independence graph and present tables corresponding to the interactions.

The conditional independence graph is shown in Figure 3.2. There is an arc linking each pair of variables that appear in an interaction in the chosen model. Using the graph in Figure 3.2, I can say things like

> *'If I know the values of* centre *and* appear, *then knowing the values of* age *and* inflam *tells me nothing more about* survived.*'*

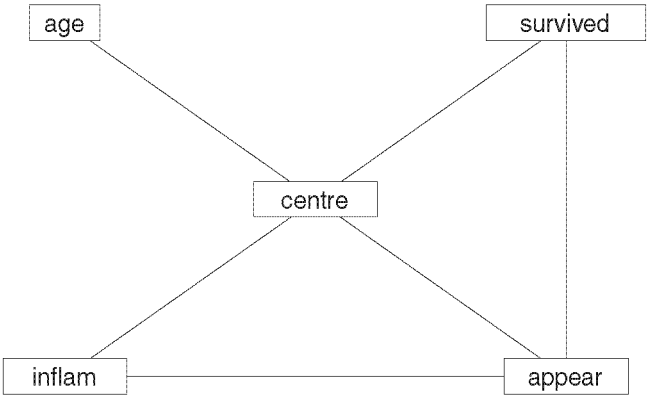

Fig. 3.2. The conditional indepence graph of the model selected for the data in Table 3.7.

because the routes from *age* and *inflam* to *survived* all pass through either *centre* or *appear*.

The frequency tables corresponding to interactions in the model are given in Tables 3.9, 3.10, 3.11, 3.12, 3.13 and 3.14. From these tables we can make statements such as

> *'The three-year survival rate in Tokyo is much bigger than in Boston or Glamorgan.'*

based on Table 3.12. By only looking at the tables corresponding to interactions that are in the selected model, we can be confident that the patterns we see in these tables are statistically significant and not simply due to chance.

The log-linear model gives us a systematic way to investigate the relationships between categorical variables. If there is an interaction in a log-linear model, this implies that the variables involved in the interaction are not independent, in other words they are related. In the particular example considered here, one of the variables, *survived*, can be considered to be an output, so we are primarily interested in associations with that variable. There is no requirement for one of the categorical variables to be considered an output. On the other hand, if one of the variables is an output, then we might prefer to model it directly; if the output is binary (has exactly two categories), as *survived* is, we can use *logistic regression* to model it directly. We take up logistic regression in Section 3.2.4. A more detailed discussion of log-linear models can be found in [368, chapter 6].

3.2.4 Logistic Regression

An alternative to using a log-linear model to analyse the breast cancer data in Table 3.7 (on page 84) would be to use logistic regression.

Table 3.9. Frequency table corresponding to the *age* by *centre* interaction for the data in Table 3.7.

	Tokyo	Boston	Glamorgan
Under 50	151	58	71
50–69	120	122	109
70 or over	19	73	41

Table 3.10. Frequency table corresponding to the *inflam* by *centre* interaction for the data in Table 3.7.

	Tokyo	Boston	Glamorgan
Minimal	206	220	184
Greater	84	33	37

Table 3.11. Frequency table corresponding to the *appear* by *centre* interaction for the data in Table 3.7.

	Tokyo	Boston	Glamorgan
Malignant	131	94	119
Benign	159	159	102

Table 3.12. Frequency table corresponding to the *survived* by *centre* interaction for the data in Table 3.7.

	Tokyo	Boston	Glamorgan
No	60	82	68
Yes	230	171	153

Table 3.13. Frequency table corresponding to the *survived* by *appear* interaction for the data in Table 3.7.

	Malignant	Benign
No	113	97
Yes	231	323

Table 3.14. Frequency table corresponding to the *inflam* by *appear* interaction for the data in Table 3.7.

	Malignant	Benign
Minimal	222	388
Greater	122	32

In logistic regression, the output is the number of successes out of a number of trials, each trial resulting in either a success or failure. For the breast cancer data, we can regard each patient as a 'trial', with success corresponding to the patient surviving for three years.

Instead of the 72 frequencies given in Table 3.7, we would have 36 pairs, each pair being the number surviving to three years and the total (surviving plus dying) number of patients. The first pair would be the 26 survivors out of the 35 patients who were in Tokyo, under 50 years old, and had a malignant tumour with minimal inflammation. Alternatively, the output could simply be given as the number of successes, either 0 or 1, for each of the 764 patients involved in the study. This is a simpler approach and is easy to do in the statistical package that I have available, so we will do this. This approach is also more stable numerically.

The model that we will fit is $P(y_j = 0) = 1 - p_j$, $P(y_j = 1) = p_j$ and

$$\log\left(\frac{p_j}{1-p_j}\right) = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_n x_{nj}. \tag{3.22}$$

Again, the inputs here will be indicators for the breast cancer data, but this is not generally true; there is no reason why any of the inputs should not be quantitative. The function $\log(p/(1-p))$ is often written $\text{logit}(p)$ by statisticians and referred to as the *logit* (of p). There are various reasons for modelling the logit, rather than the probability of success. Most of these reasons are mathematical, the simplest being that:

- using the logit prevents us from predicting success probabilities that are not between 0 and 1;
- if success has a higher probability than failure, then the logit is positive, otherwise it is negative;
- swapping the labels between success and failure simply changes the sign of the logit.

There are deeper mathematical reasons for using the logits; there are also statistical justifications based on the idea that some patients will be more robust than others—the technical name for this is the *tolerance distribution* of the patients. These deeper mathematical and statistical reasons are discussed by [123] and [368].

For a binary output, like this, the expected value is just the probability of the output being 1; as a formula this is $p_j = \mu_j$. Thus, (3.22) does conform to the form of the generalized linear model given at (3.6).

To investigate the relationship between three-year survival and the other variables we fit a logistic regression model that has 36 β's (as there are 36 possible combinations of *centre*, *age*, *inflam* and *appear*) and try to identify those that could be assumed to be zero. As with log-linear modelling, this decision is made using an analysis of deviance table. The relevant table is shown in Table 3.15.

Looking at the significance probabilities in Table 3.15 tells us that we do not need any interaction terms at all, nor do we need *inflam*, as long as we have the three main effects, *centre*, *age* and *appear*. This can be deduced because

Table 3.15. An analysis of deviance table for fitting a logistic regression to the data in Table 3.7.

	Df	Deviance	Resid. Df	Resid. Dev	Pr(Chi)
NULL			763	898.5279	
centre	2	11.26979	761	887.2582	0.0035711
age	2	3.52566	759	883.7325	0.1715588
appear	1	9.69100	758	874.0415	0.0018517
inflam	1	0.00653	757	874.0350	0.9356046
centre:age	4	7.42101	753	866.6140	0.1152433
centre:appear	2	1.08077	751	865.5332	0.5825254
centre:inflam	2	3.39128	749	862.1419	0.1834814
age:appear	2	2.33029	747	859.8116	0.3118773
age:inflam	2	0.06318	745	859.7484	0.9689052
appear:inflam	1	0.24812	744	859.5003	0.6184041
centre:age:appear	4	2.04635	740	857.4540	0.7272344
centre:age:inflam	4	7.04411	736	850.4099	0.1335756
centre:appear:inflam	2	5.07840	734	845.3315	0.0789294
age:appear:inflam	2	4.34374	732	840.9877	0.1139642
centre:age:appear:inflam	3	0.01535	729	840.9724	0.9994964

all the significance probabilities from *inflam* to the bottom of the table are relatively large. Therefore, the changes in deviance produced by adding these terms conform to what we would expect to see if these terms have no impact upon the output. Hence, we discard these terms, because we make the assumption that each term has no impact upon the output, unless we obtain evidence that it does. The evidence that we seek is a change in deviance that is not consistent with that term having no impact on the output; Table 3.15 indicates that adding *appear* to the set of inputs produces such a change in deviance—the significance probability for *appear* is small.

From here, we should check whether we need to keep *age* in the model, if *appear* and *centre* are in the model, and whether we need to keep *centre*, if *appear* and *age* are in the model. (We know that *appear* is needed, because the order of fitting in Table 3.15 adds *appear* to the model after *centre* and *age*.) Not surprisingly, it turns out that *age* is superfluous. Notice how the selected model matches the conditional independence graph in Figure 3.2 on page 86; the node for *survived* is linked to two other nodes, *centre* and *appear*.

The fitted model is simple enough in this case for the parameter estimates to be included here; they are shown in the form that a statistical package would present them in Table 3.16. Using the estimates given in Table 3.16, the fitted model is

$$\text{logit}(p_j) = 1.080257 - 0.6589141x_{Bj} - 0.4944846x_{Gj} + 0.5157151x_{aj}. \quad (3.23)$$

Here x_{Bj} is 1 if patient j was treated at the Boston centre, and 0 otherwise, that is, x_{Bj} is the indicator variable for *centre* being Boston (which is the second centre, hence the labelling in Table 3.16). Similarly, x_{Gj} is the indicator for *centre*

Table 3.16. Parameter estimates for the logistic regression model selected for the data in Table 3.7, in the form that a statistical package would present them.

```
Coefficients:
 (Intercept)     centre2     centre3     appear
    1.080257 -0.6589141 -0.4944846 0.5157151
```

being Glamorgan and x_{aj} is the indicator for *appear* being Benign. So, for a patient in Tokyo with a tumour that appears benign, the estimated logit of the probability of survival to three years is

$$1.080257 - 0 - 0 + 0.5157151 = 1.595972$$

so, solving $\log\left(\frac{p}{1-p}\right) = 1.595972$ for p, the estimated three-year survival probability is

$$\frac{1}{1 + e^{-1.595972}} = 0.8314546.$$

As we have two different ways to analyse the data in Table 3.7, how should we decide whether to use log-linear modelling or logistic regression? The choice really depends on our aims. If the overriding interest is in modelling patient survival, then the logistic regression is more natural. On the other hand the logistic regression does not tell us that *centre* and *age* are related, because this does not help in the prediction of survival. The log-linear model tells us about relationships between all the variables, not just which ones are related to *survived*.

Other differences that determine whether to use logistic regression or log-linear modelling include:

- the output has to be binary for a logistic regression, but could take, for example, one of the four values *none*, *mild*, *moderate* or *severe*, in a log-linear model;
- the variables in a log-linear model have to be categorical, but need not be in a logistic regression, so the patient's age in the breast cancer data had to be converted from the actual age in years to the three categories for use in the log-linear model.

For an extensive description of logistic regression and other ways of analysing binary data see [123].

3.2.5 Analysis of Survival Data

Analysis of survival data is not strictly within the framework of generalized linear models. This sort of analysis can be carried out using the same numerical techniques as generalized linear models, so this is a natural place to describe it.

Survival data are data concerning how long it takes for a particular event to happen. In many medical applications the event is death of a patient with

an illness, and so we are analysing the patient's survival time. In industrial applications the event is often failure of a component in a machine.

The output in this sort of problem is the survival time. As with all the other problems that we have seen in this section, the task is to fit a regression model to describe the relationship between the output and some inputs. In the medical context, the inputs are usually qualities of the patient, such as age and sex, or are determined by the treatment given to the patient.

There are two main characteristics of survival data that make them different from the data in the regression problems that we have already seen.

Censoring For many studies, the event has not necessarily happened by the end of the study period. So, for some patients in a medical trial, we might know that the patient was still alive after five years, say, but do not know when the patient died. This sort of observation would be called a *censored* observation. So, if the output is censored, we do not know the value of the output, but we do have some information about it.

Time Dependent Inputs Since collecting survival data entails waiting until the event happens, it is possible for the inputs to change in value during the wait. This could happen in a study of heart disease in which it is likely that the patient has been given medical advice to, for example, change diet or stop smoking. If the patient takes the advice, then things such as the patient's weight and blood cholesterol level are likely to change; so there would be a difficulty in deciding which value of weight to use if we wanted to model survival time using weight as an input.

These difficulties lead us to concentrate upon the *hazard* function and the *survivor* function for this sort of data. The survivor function is the probability of survival time being greater than time t. So, if the output is y, then the survivor function is

$$S(t) = P(y \geq t) = 1 - F(t)$$

where $F(t)$ is the cumulative distribution function of the output. The hazard function is the probability density of the output at time t conditional upon survival to time t, that is,

$$h(t) = f(t)/S(t),$$

where $f(t)$ is the (marginal) probability density of the output.

The hazard function indicates how likely failure is at time t, given that failure has not occurred before time t. So, for example, suppose the output is the time at which your car breaks down, measured from its last service. I would expect the hazard function to be high for the first few days after servicing (break-down due to a mistake during servicing), then it would drop, as the car (having survived the first few days) would be in good condition; after that I would anticipate a gradual rise in hazard as the condition of the car deteriorates over time, assuming no maintenance is carried out whilst the car is still running unless a service is scheduled. Notice that the hazard of break-down at, say, two years after servicing would be very high for most types of car, but the probability density would be low, because most cars would not survive to two years; the hazard is only looking

at the cars that have survived to time t and indicating the probability of their imminent failure. Most modelling of survival data is done using a *proportional-hazards model*. A proportional-hazards model assumes that the hazard function is of the form

$$h(t) = \lambda(t) \exp\{G(\boldsymbol{x}, \boldsymbol{\beta})\}, \tag{3.24}$$

where $\boldsymbol{\beta}$ is a vector of parameters to be estimated, G is a function of arbitrary, but known, form, and $\lambda(t)$ is a hazard function in its own right. The multiplier of $\lambda(t)$ is written in exponential form, because the multiplier must be greater than zero. The $\lambda(t)$ hazard function is called the *baseline hazard*. This is called a *proportional-hazard* model, because the hazard functions for two different patients have a constant ratio for different values of t (as long as the inputs are not time-dependent).

The conventional choice for G, for patient j, is

$$G(\boldsymbol{x}_j, \boldsymbol{\beta}) = \beta_0 + \sum_{i=1}^{n} \beta_i x_{ij}. \tag{3.25}$$

This choice of G means that the effect (on hazard) of each input is multiplicative. So, for example, if x_{1j} were an indicator for whether patient j smoked, and $\beta_1 = 1$, then we would be able to say that smoking increased the hazard by a factor of $\exp(1) = 2.718$, that is, the hazard for a smoker would be 2.718 times that for a non-smoker. The simple interpretation of the β's in this model and the plausibility (to statisticians) of the effects being multiplicative has made the fitting of proportional-hazard models the primary method of analysing survival data.

The choice of G in (3.25) has another massive advantage. If we know, or are prepared to assume, the baseline hazard, $\lambda(t)$, then it turns out that the log likelihood can be converted to the same form as that of a generalized linear model, with a Poisson output and a log link function; this is the same form as the log-linear model in (3.21) on page 83. The log-linear model implied by this conversion of the log likelihood has the same inputs and β's as in (3.25), but the output is the indicator w_j, which is 1 if the jth survival time is not censored, and 0 if it is. This means that there is a method for estimating the β's, which is available in any proprietary statistical package that can fit generalized linear models; all the major statistical packages can do this.

As an example of this fitting, we could assume that the baseline hazard was

$$\lambda(t) = \alpha t^{\alpha - 1},$$

which would mean that the survival times had a distribution called the *Weibull* distribution. So, we would start by assuming $\alpha = 1$ and then construct our artificial log-linear model. We would then fit the log-linear model and obtain estimates of the β's based on the assumption that $\alpha = 1$. We then assume that the estimates of the β's are in fact the true values, and find the maximum likelihood estimate of α under this assumption. We then assume the estimate of α is the true value of α, and construct another log-linear model to obtain updated

estimates of the β's, which we then use to update the estimate of α. This process continues until convergence of the estimates of α and the β's (which will happen and will be at the global optimum; see, for example [368, chapter 13]).

In the industrial context, it is reasonable to believe that the baseline hazard is known. Experiments to determine its form could be carried out in most industrial contexts, because the values of the inputs are likely to be controllable. In the medical context, it would be neither ethical nor practical to carry out an experiment to determine the form of the baseline hazard function. To overcome this problem, we can use *Cox's proportional-hazards model*, which was introduced by [128].

Cox's proportional-hazards model allows the baseline hazard to be arbitrary. The key idea for this model is to base the estimation of the β's upon the likelihood derived from knowing the order in which patients died (or were censored), rather than the likelihood derived from the actual survival (or censoring) times. This discards information, but the information it discards is related to the baseline hazard, and the resulting likelihood does not depend upon the baseline hazard. We have to accept this loss of information, if we do not know and are unwilling to assume what form the baseline hazard takes. Again, it turns out that the likelihood for Cox's proportional-hazards model can be converted into that of an artificial log-linear model (which is not the same as the one referred to already). Fitting this artificial log-linear model produces estimates for the β's amongst other things. Cox's proportional-hazards model is used in virtually all medical analyses of survival data, sometimes regardless of whether it is appropriate.

Details of the likelihoods and references to even more detailed discussions of this sort of analysis can be found in [368, chapter 13].

3.3. Special Topics in Regression Modelling

The topics in this section are special in the sense that they are extensions to the basic idea of regression modelling. The techniques have been developed in response to methods of data collection in which the usual assumptions of regression modelling are not justified.

Multivariate analysis of variance concerns problems in which the output is a vector, rather than a single number. One way to analyse this sort of data would be to model each element of the output separately, but this ignores the possible relationship between different elements of the output vector. In other words, this analysis would be based on the assumption that the different elements of the output vector are not related to one another. Multivariate analysis of variance is a form of analysis that does allow for different elements of the output vector to be correlated to one another. This technique is outlined in Section 3.3.1.

Repeated measures data arise when there is a wish to model an output over time. A natural way to perform such a study is to measure the output and the inputs for the same set of individuals at several different times. This is called taking repeated measurements. It is unrealistic to assume (with no evidence) that the output measurements taken on a particular individual at different times are

unrelated. Various ways of coping with the relationship between output recorded on the same individual are presented in Section 3.3.2.

Random effects modelling is described in Section 3.3.3 and also as one approach to analysing repeated measures data in Section 3.3.2. Amongst other things, this technique allows (the coefficients of) the fitted regression model to vary between different individuals.

3.3.1 Multivariate Analysis of Variance

In Section 3.2.2, we saw how to test whether particular β's of the linear model of (3.1) on page 71 were zero. The technique used was called analysis of variance, or ANOVA.

Sometimes the output in a data set is a vector of variables rather than a single variable. We will see a special case of this in Section 3.3.2, where the vector of variables consists of the same quantity, blood pressure say, at several different times, recorded on the same individual. The output variables need not be the same quantity; they could be something like heights and weights for a set of children.

Given this sort of data, we might be able to analyse it using a multivariate linear model, which is

$$\underset{(c\times 1)}{\boldsymbol{y}_j} = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_n x_{nj} + \boldsymbol{\varepsilon}_j \quad j = 1, 2, \ldots, m \tag{3.26}$$

where the $\boldsymbol{\varepsilon}_j$'s are independently and identically distributed as[4] $\mathcal{N}^c(\mathbf{0}, \Sigma)$ and m is the number of data points. The '$(c \times 1)$' under '$\boldsymbol{y}_j$' indicates the dimensions of the vector, in this case c rows and 1 column; the $\boldsymbol{\beta}$'s are also $(c \times 1)$ vectors.

This model can be fitted in exactly the same way as a linear model (by least squares estimation). One way to do this fitting would be to fit a linear model to each of the c dimensions of the output, one-at-a-time.

Having fitted the model, we can obtain fitted values

$$\hat{\boldsymbol{y}}_j = \hat{\beta}_0 + \sum_{i=1}^{n} \hat{\boldsymbol{\beta}}_i x_{ij} \quad j = 1, 2, \ldots, m$$

and hence residuals

$$\boldsymbol{y}_j - \hat{\boldsymbol{y}}_j \quad j = 1, 2, \ldots, m.$$

The analogue of the residual sum of squares from the (univariate) linear model is the matrix of residual sums of squares and products for the multivariate linear

[4] The notation $\mathcal{N}^c(\boldsymbol{\mu}, \Sigma)$ denotes the c-dimensional normal distribution. If $\boldsymbol{w} \sim \mathcal{N}^c(\boldsymbol{\mu}, \Sigma)$, then $\boldsymbol{w}$ is a $(c \times 1)$ vector, whose elements are each normally distributed. The expected values of the elements of $\boldsymbol{w}$ are the corresponding elements of $\boldsymbol{\mu}$, which is also a $(c \times 1)$ vector. The variance of element i of $\boldsymbol{w}$ is element (i, i) of the $(c \times c)$ variance matrix, Σ; the covariance between elements i and k of $\boldsymbol{w}$ is element (i, k) of Σ.

model. This matrix is defined to be

$$R = \sum_{j=1}^{m} (\boldsymbol{y}_j - \hat{\boldsymbol{y}}_j)(\boldsymbol{y}_j - \hat{\boldsymbol{y}}_j)^\top.$$

The R matrix has the residual sum of squares for each of the c dimensions stored on its leading diagonal. The off-diagonal elements are the residual sums of cross-products for pairs of dimensions.

The residual degrees of freedom is exactly the same as it would be for a univariate linear model with the same set of inputs, namely m minus the number of linearly independent inputs. If we call the residual degrees of freedom ν, then we can estimate Σ by $(1/\nu)R$.

If we wish to compare two nested linear models, to determine whether certain $\boldsymbol{\beta}$'s are equal to the zero vector, $\mathbf{0}$, then we can construct an extra sums of squares and products matrix in the same way as we constructed the extra sum of squares in Section 3.2.2 on page 78. In other words, if we have model Ω_0 which is a special case of model Ω_1, and the residual sums of squares and products matrices for these models are R_0 and R_1, respectively, then the extra sums of squares and products matrix is

$$R_E = R_0 - R_1.$$

Thus, in the same way that we can form an ANOVA table, which shows the effect of introducing new inputs into the model, we can build a multivariate ANOVA, or *MANOVA*. Instead of the sums of squares of the ANOVA table, the MANOVA table contains sums of squares and products matrices.

The only sensible choices for the test statistic to compare models Ω_0 and Ω_1 are functions of the eigenvalues of the matrix

$$R_1^{-1} R_E,$$

where R_1^{-1} is the matrix inverse of R_1. By 'sensible', we mean that the test statistic does not change if the co-ordinate origin (for the output vectors) is changed, nor if the output variables are rescaled; this is demonstrated in [21]. There are four commonly used test statistics. If the dimensionality of the output is $c = 1$, then these statistics are all the same as the F-statistic of the ANOVA. Further, they are all equivalent to one another if the extra degrees of freedom is 1.

The names of the four commonly used test statistics are: *Roy's greatest root*; the *Lawley-Hotelling trace*; the *Pillai trace*; *Wilks' lambda*. Discussions of which one to use under what circumstances can be found in [326, section 13.2] and [112, section 8.3.3]. Most statistical packages that can perform MANOVA will produce these four statistics and the corresponding significance probabilities.

Having performed a MANOVA and discovered some $\boldsymbol{\beta}$'s are not equal to $\mathbf{0}$, we usually want to say which of the inputs is related to which element of the output, and how. If the input is a quantitative variable, then it is straightforward to identify the relationship, simply by considering the corresponding $\boldsymbol{\beta}$. If the input is one of a set of indicator variables corresponding to a factor, then we need

to consider the set of β's corresponding to the factor. This is harder, because there are too many different contributions in the β's for someone to grasp. In this case, we would use a technique called *canonical variates analysis*, to display the mean output for each level of the factor. Canonical variates analysis will not be described here, but descriptions of this technique can be found in many textbooks on multivariate analysis, for example [326] and [112]. One way to think of canonical variates analysis is that it is the same as principal components analysis (see Section 3.4.1) applied to the mean outputs for each distinct level of the factor. Though this is not strictly true, it conveys the basic idea, which is that of trying to focus on the differences between the means.

3.3.2 Repeated Measures Data

Repeated measures data are generated when the output variable is observed at several points in time, on the same individuals. Usually, the covariates are also observed at the same time points as the output; so the inputs are time-dependent too. Thus, as in Section 3.3.1 the output is a vector of measurements. In principle, we can simply apply the techniques of Section 3.3.1 to analyse repeated measures data. Instead, we usually try to use the fact that we have the same set of variables (output and inputs) at several times, rather than a collection of different variables making up a vector output.

Repeated measures data are often called *longitudinal data*, especially in the social sciences. The term *cross-sectional* is often used to mean 'not longitudinal'.

There are several ways to analyse this sort of data. As well as using the MANOVA, another straightforward approach is to regress the outputs against time for each individual and analyse particular parameter estimates, the gradient say, as though they were outputs. If the output variable were measured just twice, then we could analyse the difference between the two measurements, using the techniques that we saw in Section 3.2.

The three most common ways to analyse longitudinal data, at least in the statistical research literature, are *marginal* modelling, *subject-specific* modelling and *transition* modelling.

Marginal modelling is what we have been considering in the regression modelling that we have already seen. We have a model, such as (3.26) on page 94. The primary complication is that it is unrealistic to assume that there is no correlation between outputs from the same individual. We could make no assumptions about the correlation structure, as we did for the MANOVA, but usually there is reason to believe that the variance matrix has a particular form. For example, if the individuals are (adult) patients, then we might expect the correlation between the outputs in weeks 1 and 2 to be the same as that between outputs in weeks 5 and 6 (because the pairs have the same time lapse between them), with the correlation between outputs for weeks 2 and 5 being smaller.

Subject-specific modelling is regression modelling in which we allow some of the parameters to vary between individuals, or groups of individuals. An

example of this sort of model is

$$y_{jt} = \beta_0 + U_j + \beta_1 x_{1jt} + \beta_2 x_{2jt} + \cdots + \beta_n x_{njt} + \varepsilon_{jt} \quad j = 1, 2, \ldots, m, \quad (3.27)$$

where t indexes the observation time, the ε_{jt}'s are independently and identically distributed as $\mathcal{N}(0, \sigma^2)$, the U_j's are independently and identically distributed as $\mathcal{N}(0, \sigma_U^2)$ and m is the number of individuals.

The model in (3.27) is called a *random-intercept* model, because it is equivalent to a standard model with the intercept (β_0) replaced by an intercept for each individual ($\beta_0 + U_j$). This random-intercept model does not have a correlation structure to specify, as was required for the marginal model. Instead, the relationship between outputs from the same individual (at different times) is assumed to be represented by the U_j's being constant across time, but different between individuals. If U_j happens to be high, then all the outputs for individual j will be expected to be higher than the average across all individuals.

Other subject-specific models allow the other β's, not just the intercept, to vary between individuals, or allow things like a random intercept for, say, the clinic an individual attends, or the district an individual lives or works in.

Subject-specific models are a type of model known as random effects models; see Section 3.3.3.

Transition models allow the output to depend on outputs at previous time points. Thus, the relationship between outputs from the same individual is allowed for by allowing outputs from the past to be used as inputs for the output in the present. A simple example of a transition model is

$$y_{jt} = \alpha y_{j(t-1)} + \beta_0 + \beta_1 x_{1jt} + \beta_2 x_{2jt} + \cdots + \beta_n x_{njt} + \varepsilon_{jt} \quad j = 1, 2, \ldots, m, \quad (3.28)$$

where t indexes the observation time, the ε_{jt}'s are independently and identically distributed as $\mathcal{N}(0, \sigma^2)$ and m is the number of individuals.

As an example, suppose y_{jt} is a measurement of the progression of a disease. The model at (3.28) implies that the progression of the disease at time t is related to the progression at the previous time point $(t-1)$ and to the values of a set of inputs at time t.

The dependence on $y_{j(t-1)}$ can be far more complicated than in (3.28). For example, a much more general dependence would be to allow the β's to depend on $y_{j(t-1)}$. This would be difficult if the output were a continuous quantity, but might even be considered natural if the output were binary, as in the models in Section 3.2.4.

Descriptions of these three types of models, along with details on how to fit them using maximum likelihood estimation, practical examples of the their use and bibliographic notes, can be found in [155].

The choice of which of these types of models to use boils down to what the aims of the study are and what is believed about the underlying mechanism of change in the output. The implications of these choices, and hence reasons for

choosing one type of model rather than another are also given in [155], but, as an example, a marginal model is appropriate when you want to make public health statements like

Reduce your blood pressure by eating less salt.

which apply on average. This sort of public health advice is not necessarily appropriate to everybody in a population, but if everyone followed the advice then the *average* blood pressure across the population would be reduced; for some people eating less salt might make no difference to their blood pressure and there might even be individuals whose blood pressure actually increases. If you wanted to give health advice on an individual basis, then you would want to use a subject-specific model and estimate random effects, such as the U_j's in (3.27), for each individual. This would allow you give different advice to different people, each person getting advice appropriate to themselves; the advice would vary because people vary.

3.3.3 Random Effects Models

We have seen the use of random effects for longitudinal data in the subject-specific model described in Section 3.3.2. The random effects in the subject-specific model are used as a way to cope with the fact that outputs from one individual are likely to be more similar to one another than outputs from different individuals are.

There are several other situations in which random effects could be included in a model. Some of the main uses of random effects in modelling are given below.

Overdispersion Overdispersion is the phenomenon of the observed variability of the output being greater than the fitted model predicts. This can be detected for some generalized linear models; for example, overdispersion can be detected for models in which the output has a binomial distribution. The reason why overdispersion can be detected for a binomial output is that the variance is a function of the expected value for the binomial distribution. Thus, the model happens to predict the variance as well as the expected value, so we can compare the sample variance with that predicted by the model.

There are two main ways in which overdispersion might arise, assuming that we have identified the correct output distribution. One way is that we failed to record a variable that ought to have been included as an input. The reasons for not recording the missing input might include: time or expense of recording it; an incorrect belief that it was not relevant to the study; we do not actually know how to measure it. The second way is that the expected output for an individual is not entirely determined by the inputs, and that there is variation in the expected output between different individuals who have the same values for the inputs.

In either case, the solution is to allow a random effect into the linear part of the model. In a logistic regression we might replace (3.22) from page 88 with

$$\text{logit}(p_j) = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_n x_{nj} + U_j, \tag{3.29}$$

where the U_j's are independently and identically distributed as $\mathcal{N}(0, \sigma_U^2)$. We can think of U_j as representing either the effect of the missing input on p_j or simply as random variation in the success probabilities for individuals that have the same values for the input variables.

The topic of overdispersion is discussed in detail by [123, chapter 6].

Hierarchical Models We have seen a hierarchical model already. Equation (3.8) on page 75 has three parameters, β_{II}, β_{III} and β_{IV}, that concern differences between groups of individuals, in this case the suitability of different parts of a field for growing turnips. The turnip experiment of Example 3.1 features four groups of plots. Plots in the same group are physically close to one another, so they are assumed to have similar characteristics to one another. The different groups correspond to different pieces of land, so we allow for the possibility that plots in different groups might have different characteristics (in terms of suitability for growing turnips).

This allocation of individuals into groups, so as to put similar individuals in the same group as one another, is called *blocking*. The groups are referred to as *blocks*.

The hierarchy which gives its name to hierarchical models is, for the turnip experiment, made up of two levels, the upper level being the four blocks and the lower level being the plots within the blocks. It takes little imagination to envisage the extension of the hierarchy to a third level, different farms say, so that there are groups of blocks corresponding to different farms and groups of plots corresponding to blocks. To move away from the agricultural context, a medical hierarchy could be patients within wards within hospitals within health trusts; an educational hierarchy might be student within class within school within education authority.

In many cases the levels in the hierarchy consist of samples of groups taken from populations of groups. In the turnip experiment, the growth of the turnips is affected by the different blocks, but the effects (the β's) for each block are likely to be different in different years. So we could think of the β's for each block as coming from a population of β's for blocks. If we did this, then we could replace the model in (3.8) on page 75 with

$$\begin{aligned} y_j = {} & \beta_0 + \beta_B x_{Bj} + \beta_C x_{Cj} + \cdots + \beta_R x_{Rj} \\ & + b_{\text{I}} x_{\text{I},j} + b_{\text{II}} x_{\text{II},j} + b_{\text{III}} x_{\text{III},j} + b_{\text{IV}} x_{\text{IV},j} + \varepsilon_j \quad j = 1, 2, \ldots, 64 \end{aligned} \tag{3.30}$$

where b_{I}, b_{II}, b_{III} and b_{IV} are independently and identically distributed, each being $\mathcal{N}(0, \sigma_b^2)$. This is how random effects, like the b's can be used in hierarchical models. We can do the same thing in both the medical and educational scenarios given above.

We would not use random effects to replace β_B to β_R, because there is an implicit belief in this experiment that these are constants that will be the same in the future, that is, they are not random. In addition, they can be attributed to specific actions by the person growing the turnips, whereas the block effects are attributable to random fluctuations in the quality of the land, the weather, whether the field was flooded by the river, and so on.

For the turnip data using a random effects model like (3.30) makes no difference to our conclusions about the effects of the different treatments. This is because the turnip data were generated using a carefully designed experiment; one of the goals of the design used was to eliminate any dependence of the treatment comparisons upon σ_b^2. In other studies, a random effects model might be essential. There are two common reasons for requiring a random effects model. First, we actually want to estimate variances such as σ_b^2, either for their own sake, or because a treatment has been applied to the groups in a particular level, rather than to individuals in the lowest level. Secondly, the data may have been generated by a study in which we had less control over the values taken by the inputs than we had for the turnip experiment. This can lead to treatment effects whose estimates do depend on the variance of random effects, which we would want to take into account, or it might lead to situations in which the β's cannot be estimated without including random effects in the model. Examples of both these possibilities are given by [368, chapter 14].

3.4. Classical Multivariate Analysis

The techniques in this section are used for analysing samples that are in the form of observations that are vectors. They all have their roots in linear algebra and geometry, which is perhaps why these diverse techniques are grouped together under the heading of *multivariate analysis*.

Principal components analysis is a method of transforming the vector observations into a new set of vector observations. The goal of this transformation is to concentrate the information about the differences between observations into a small number of dimensions. This allows most of the structure of the sample to be seen by plotting a small number of the new variables against one another. A fairly detailed description of principal components analysis is given in Section 3.4.1.

The relationships between the rows and columns of a table of counts can be highlighted using correspondence analysis, which is described in Section 3.4.2. Despite having different aims from principal components analysis, it turns out that the mathematical derivation of correspondence analysis is very closely related to that of principal components analysis. Many of the techniques regarded as *multivariate analysis* are related to one another mathematically, which is another reason for techniques with disparate aims being grouped together under this heading.

There are situations in which the information available is in the form of the dissimilarities (or distances) between pairs of individuals, rather than the vectors

corresponding to each individual. Multidimensional scaling is a set of techniques for constructing the set of vectors from the set of dissimilarities. Multidimensional scaling is described, along with a contrived example, in Section 3.4.3. The number of situations in which this technique is useful in practice is surprisingly high.

Cluster analysis is a collection of techniques for creating groups of objects. The groups that are created are called clusters. The individuals within a cluster are similar in some sense. An outline of the various different types of cluster analysis is given in Section 3.4.4, along with a short description of *mixture decomposition*. Mixture decomposition is similar to cluster analysis, but not the same. The difference is given in Section 3.4.4.

The relationships between the elements of the observation vectors can be investigated by the techniques in Section 3.4.5. This is a different emphasis from the other multivariate techniques presented here, as those techniques are trying to achieve goals in spite of the relationships between elements of the observation vectors. In other words, the structure being studied in Section 3.4.5 is regarded as a nuisance in most other multivariate analysis.

The main omission from this set of multivariate techniques is anything for the analysis of data in which we know that there is a group structure (and we know what that structure is, as well as knowing it exists). Discrimination, or *pattern recognition*, or *supervised classification*, is a way to model this sort of data, and is addressed in (parts of) Chapters 7, 8 and 9; for a statistical view of this topic see, for example [241]. Canonical variates analysis is a way to obtain a diagram highlighting the groups structure; see, for example [326] and [112].

3.4.1 Principal Components Analysis

Principal components analysis is a way of transforming a set of c-dimensional vector observations, $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$, into another set of c-dimensional vectors, $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_m$. The $\boldsymbol{y}$'s have the property that most of their information content is stored in the first few dimensions (features).

The idea is that this will allow us to reduce the data to a smaller number of dimensions, with low information loss, simply by discarding some of the elements of the $\boldsymbol{y}$'s. Activities that become possible after the dimensionality reduction include:

- obtaining (informative) graphical displays of the data in 2-D;
- carrying out computer intensive methods on reduced data;
- gaining insight into the structure of the data, which was not apparent in c dimensions.

One way that people display c-dimensional data, with the hope of identifying structure is the pairwise scatterplot. We will look at this method and see what its weaknesses are before going on to look at how principal components analysis overcomes these problems.

Figure 3.3 shows a pairwise scatterplot for a data set with $c = 4$ dimensions and $m = 150$ individuals. The individuals are iris flowers; the four dimensions

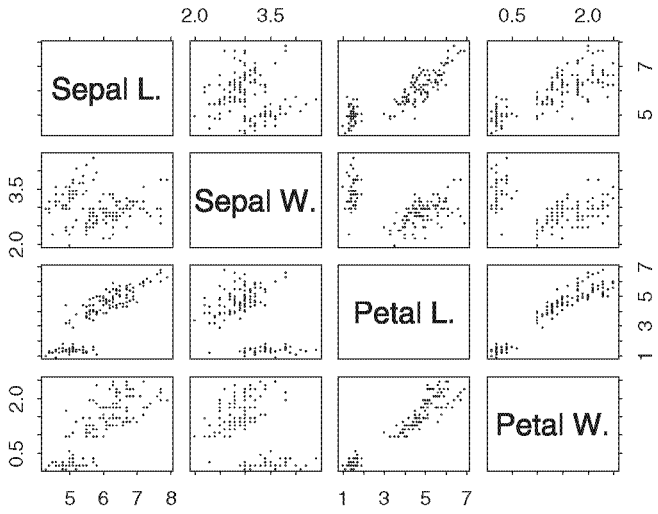

Fig. 3.3. Pairwise scatterplots for Fisher's Iris Data.

correspond to the sepal lengths and widths, and the petal lengths and widths of the flowers. It is known that there are three different species of iris represented in this set of data, but we are not using this information in the analysis. These data are very famous and are usually known as *Fisher's Iris Data*, even though they were collected by Anderson and described in [20], because they were used by Fisher in [180] (the first paper on linear discriminant analysis).

The pairwise scatterplot is constructed by plotting each variable against every other variable. The idea is to have all information available simultaneously. The main structure that can be identified in Figure 3.3 is that there is a distinct group of individuals with low values for both *Petal W.* and *Petal L.* (petal width and length).

The drawbacks with pairwise scatterplots are as follows.

- A high number of dimensions leads to a very complicated picture.
- It is often difficult to locate the same observation in several different panels (plots).
- There is an implicit assumption that the structure within the data is related to pairs of features. There might be other 2-D projections that are clearer.

There are ways to improve upon pairwise scatterplots using dynamic graphics to look at 3-D projections (this is called *spinning*) or to allow highlighting of individuals in all the panels simultaneously (this is called *brushing*). It is also possible to use symbols or *glyphs*, such as *Chernoff's faces* introduced by [115], as described in [326, pages 35–40]. Even though they are improvements, none of these techniques is particularly easy to use when searching for structure and they are all time consuming, requiring a human 'expert' to interpret them.

We need a method that will regularly reveal structure in a multivariate data set. Principal components analysis is such a method.

The main idea behind principal components analysis is that high information corresponds to high variance. So, if we wanted to reduce the $\boldsymbol{x}$'s to a single dimension we would transform $\boldsymbol{x}$ to $y = \boldsymbol{a}^\top \boldsymbol{x}$, choosing $\boldsymbol{a}$ so that y has the largest variance possible. (The length of $\boldsymbol{a}$ has to be constrained to be a constant, as otherwise the variance of y could be increased simply by increasing the length of $\boldsymbol{a}$ rather than changing its direction.) An alternative, but equivalent, approach is to find the line that is closest to the $\boldsymbol{x}$'s; here we measure the distance from the $\boldsymbol{x}$'s to a line by finding the squared perpendicular distance from each $\boldsymbol{x}_j$ to the line, then summing these squared distances for all the $\boldsymbol{x}_j$'s. Both approaches give the same answer; we will use the variance approach because this makes it simpler to write down the procedure for carrying out principal components analysis.

It is possible to show that the direction of maximum variance is parallel to the eigenvector corresponding to the largest eigenvalue of the variance (covariance) matrix of $\boldsymbol{x}$, Σ. It is also possible to show that of all the directions orthogonal to the direction of highest variance, the (second) highest variance is in the direction parallel to the eigenvector of the second largest eigenvalue of Σ. These results extend all the way to c dimensions. The eigenvectors of Σ can be used to construct a new set of axes, which correspond to a rotation of the original axes. The variance parallel to the first axis will be highest, the variance parallel to the second axis will be second highest, and so on.

In practice, we do not know Σ, so we use the sample variance (covariance), S, which is defined to be

$$\underset{(c\times c)}{S} = \frac{1}{m-1}\sum_{j=1}^{m}(\boldsymbol{x}_j - \overline{\boldsymbol{x}})(\boldsymbol{x}_j - \overline{\boldsymbol{x}})^\top, \tag{3.31}$$

where $\overline{\boldsymbol{x}} = \frac{1}{m}\sum_j \boldsymbol{x}_j$.

To specify the principal components, we need some more notation.

- The eigenvalues of S are

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_c \geq 0.$$

- The eigenvectors of S corresponding to $\lambda_1, \lambda_2, \ldots, \lambda_c$ are $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_c$, respectively.
  The vectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_c$ are called the *principal axes.* ($\boldsymbol{e}_1$ is the first principal axis, etc.)
- The $(c \times c)$ matrix whose ith column is $\boldsymbol{e}_i$ will be denoted as E.

The principal axes (can be and) are chosen so that they are of length 1 and are orthogonal (perpendicular). Algebraically, this means that

$$\boldsymbol{e}_i^\top \boldsymbol{e}_{i'} = \begin{cases} 1 & \text{if } i = i' \\ 0 & \text{if } i \neq i' \end{cases}. \tag{3.32}$$

The vector $\boldsymbol{y}$ defined as,

$$\underset{(c\times 1)}{\boldsymbol{y}} = \underset{(c\times c)}{\begin{bmatrix} \boldsymbol{e}_1^\top \\ \boldsymbol{e}_2^\top \\ \vdots \\ \boldsymbol{e}_c^\top \end{bmatrix}} \underset{(c\times 1)}{\boldsymbol{x}} = E^\top \boldsymbol{x}$$

is called the vector of *principal component scores* of $\boldsymbol{x}$. The ith principal component score of $\boldsymbol{x}$ is $y_i = \boldsymbol{e}_i^\top \boldsymbol{x}$; sometimes the principal component scores are referred to as the principal components.

The principal component scores have two interesting properties.

1. The elements of $\boldsymbol{y}$ are uncorrelated and the sample variance of the ith principal component score is λ_i. In other words the sample variance matrix of $\boldsymbol{y}$ is
$$\underset{(c\times c)}{\begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_c \end{bmatrix}}.$$
So an alternative way to view principal components analysis is as a method for obtaining a set of uncorrelated variables.
2. The sum of the sample variances for the principal components is equal to the sum of the sample variances for the elements of $\boldsymbol{x}$. That is,
$$\sum_{i=1}^{c} \lambda_i = \sum_{i=1}^{c} s_i^2,$$
where s_i^2 is the sample variance of x_i.

The principal components scores are plotted in Figure 3.4. Compare Figure 3.4 with Figure 3.3. In Figure 3.3 every panel exhibits two clusters and all four variables carry information about the two clusters. In Figure 3.4, on the other hand most of the information about these two clusters is concentrated in the first principal component, y_1. The other three principal components do not appear to carry any structure. One might wish to argue that the plots of y_3 and y_4 against y_2 suggest two overlapping clusters separated by the line $y_2 = -5.25$; this indicates that y_2 carries some structural information. The clear message of Figure 3.4 is that the structural information for these data is stored in at most two principal components, so we can achieve a dimensionality reduction from four dimensions down to two dimensions, with no important loss of information.

There are two more issues to consider before we finish our look at principal components analysis.

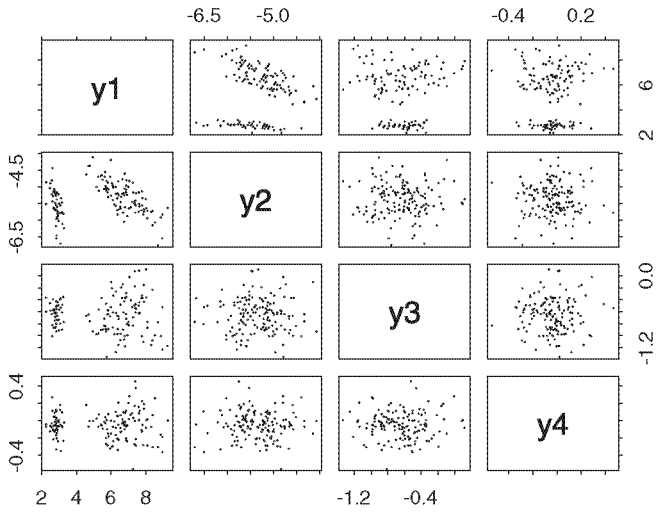

Fig. 3.4. Plot of the principal components for Fisher's Iris Data.

- How many of the principal components are needed to get a good representation of the data? That is, what is the effective dimensionality of the data?
- Should we normalise the data before carrying out principal components analysis?

Effective Dimensionality There are three main ways of determining how many principal components are required to obtain an adequate representation of the data.

1. **The proportion of variance accounted for** Take the first r principal components and add up their variances. Divide by the sum of all the variances, to give

$$\frac{\sum_{i=1}^{r} \lambda_i}{\sum_{i=1}^{c} \lambda_i}$$

which is called the *proportion of variance accounted for by the first r principal components.*

Usually, projections accounting for over 75% of the total variance are considered to be good. Thus, a 2-D picture will be considered a reasonable representation if

$$\frac{\lambda_1 + \lambda_2}{\sum_{i=1}^{c} \lambda_i} > 0.75.$$

2. **The size of *important* variance** The idea here is to consider the variance if all directions were equally important. In this case the variances would be

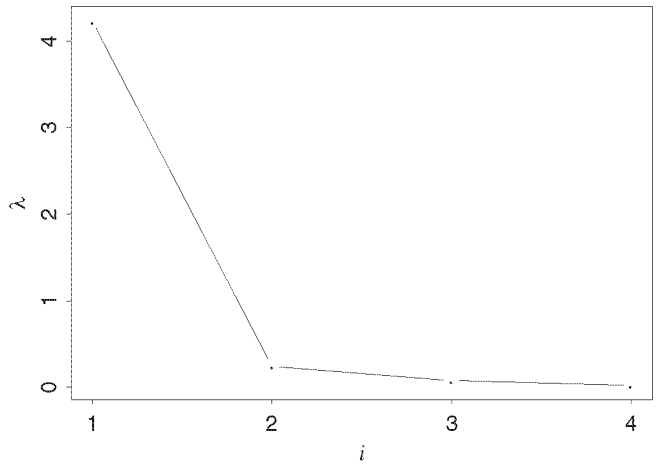

Fig. 3.5. A scree diagram for Fisher's Iris Data.

approximately

$$\overline{\lambda} = \frac{1}{c}\sum_{i=1}^{c} \lambda_i.$$

The argument runs

If $\lambda_i < \overline{\lambda}$, then the ith principal direction is less interesting than average.

and this leads us to discard principal components that have sample variances below $\overline{\lambda}$.

3. **Scree diagram** A scree diagram is an index plot of the principal component variances. In other words it is a plot of λ_i against i. An example of a scree diagram, for the Iris Data, is shown in Figure 3.5.
The idea is that where the curve flattens out is where the variation is just random fluctuation with no structure. So we look for the elbow; in this case we only need the first component.

These methods are all ad-hoc, rely on judgement and have no theoretical justification for their use. The first two methods have the convenient trait of being precisely defined, so that these methods could be performed by computer. Though they are not supported by any theory, these methods work well in practice. My personal preference is to use the scree diagram and the proportion of variance methods, in tandem.

Normalising Consider the two scatterplots in Figure 3.6. The scatterplot on the left shows two of the variables from Fisher's iris data. The scatterplot on the

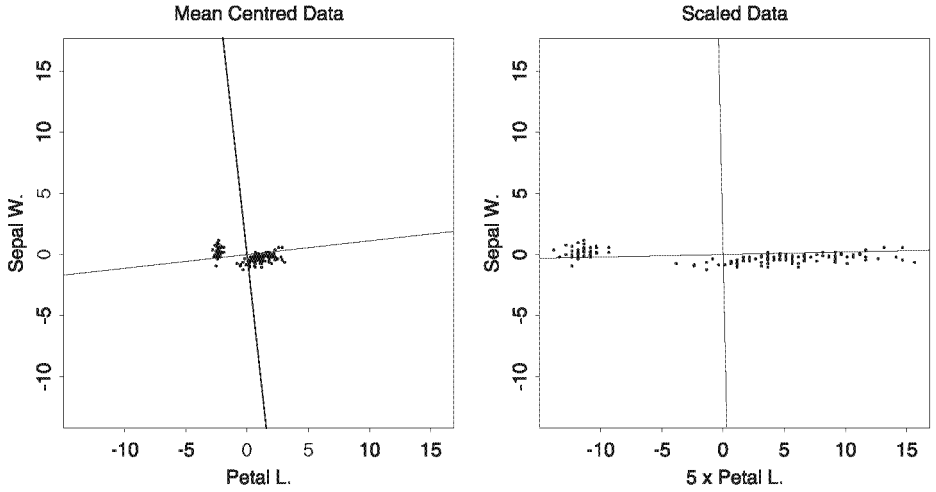

Fig. 3.6. The effect of non-uniform dilation on the principal axes.

right shows the same two variables but with one of them, *Petal L.*, multiplied by 5. The variables have been centred (had the mean value subtracted from each observation), so that the point being illustrated will be easier to see in the diagrams. Principal components analysis was carried out separately for each pair of variables. The (directions of) the principal axes are shown, as two lines passing through $(0, 0)$. The effect of multiplying *Petal L.* by 5 can be seen in the change of direction of the principal axes; the first principal axis moves closer to the *Petal L.* axis. As we increase the dilation in the *Petal L.* direction the first principal axis tends to become parallel to the axis of the first variable.

This is important, because it means the results of principal components analysis depend upon the measurement units (centimetres, inches, feet, stones, pounds, kilograms, etc.). Usually we want to identify structure, regardless of the particular choice of measurement scale, so we normalise the data.

The data can be normalised by carrying out the following steps.

- Centre each variable. In other words, subtract the mean of each variable to give
$$\overset{\circ}{\boldsymbol{x}}_j = \boldsymbol{x}_j - \overline{\boldsymbol{x}}.$$
- Divide each element of $\overset{\circ}{\boldsymbol{x}}_j$ by its standard deviation; as a formula this means calculate
$$z_{ij} = \frac{\overset{\circ}{x}_{ij}}{s_i},$$
where s_i is the sample standard deviation of x_i.

The result is to produce a $\boldsymbol{z}_j$ for each $\boldsymbol{x}_j$, with the property that the sample variance matrix of $\boldsymbol{z}$ is the same as the correlation matrix for $\boldsymbol{x}$.

The only reason that I can think of for not normalising is if the measurements of each element of $\boldsymbol{x}$ are in the same units, **and** that the scale is relevant; spatial data would be an example of a type of data for which normalising would not usually be appropriate.

So, in general, many statisticians would recommend normalising and applying principal components analysis to the $\boldsymbol{z}_j$'s rather than the $\boldsymbol{x}_j$'s matrix.

Interpretation The final part of a principal components analysis is to inspect the eigenvectors in the hope of identifying a meaning for the (important) principal components. For the normalised *Iris Data* the eigenvector matrix (E) is

$$\begin{pmatrix} 0.521 & 0.377 & 0.720 & 0.261 \\ -0.269 & 0.923 & -0.244 & -0.124 \\ 0.580 & 0.024 & -0.142 & -0.801 \\ 0.565 & 0.067 & -0.634 & 0.524 \end{pmatrix}$$

and the eigenvalues (the λ's) are

$$(2.92, 0.91, 0.15, 0.02)$$

so, for the normalised data, the first two principal components should be adequate, with the third and fourth components containing very little variation.

The first eigenvector (column 1 of E) suggests that the first principal component is large for large values of variables 1, 3, and 4, and for small values of variable 2, using the original (unnormalised) scale. (On the normalised scale variable 2 takes negative values when the first principal component is large; negative values on the normalised scale correspond to values that are below the mean on the unnormalised scale.) The sort of meaningful interpretation that can be made here is that y_1 is large for irises with long, thin sepals and big petals.

The second principal component essentially measures the size of the sepals (variables 1 and 2); it is dominated by variable 2 (sepal width); there is essentially no contribution from the petal dimensions.

Interpretation of the other two principal components is pointless, as they contain very little information.

Principal Components and Neural Networks In Chapter 8, it is stated that principal components analysis can be performed using neural network and Hebbian learning. Here we will demonstrate how principal components analysis is related to Hebbian learning.

In the neural network context, the network produces an output, o, from a set of inputs, $\boldsymbol{x}$, using a set of weights $\boldsymbol{w}$. The relationship between these three components is

$$o = \boldsymbol{w}^\top \boldsymbol{x}.$$

Earlier in this chapter, we considered the problem of finding $\boldsymbol{a}$ so as to maximise the variance of y, where

$$y = \boldsymbol{a}^\top \boldsymbol{x}.$$

Clearly, if we set $\boldsymbol{w} = \boldsymbol{a}$ then the neural network output, o, will be the first principal component score of $\boldsymbol{x}$, namely y.

We can, without any loss in generality, assume that the (input) variables have been centred (see 'Normalising' above). In this case, the (sample) variance of y will be

$$\frac{1}{m-1}\sum_{j=1}^{m} y_j^2.$$

So, for the neural network to perform principal components analysis, it must have the goal of maximising

$$\sum_{j=1}^{m} o_j^2 = \sum_{j=1}^{m} (\boldsymbol{w}^\top \boldsymbol{x}_j)^2.$$

It is apparent that we can increase the size of this function by making the elements of $\boldsymbol{w}$ bigger, and we can make it as large as we like. Therefore, we need a constraint on $\boldsymbol{w}$; the standard constraint is $\boldsymbol{w}^\top \boldsymbol{w} = 1$.

Putting all these bits together, we conclude that a neural network that learns how to maximise

$$\sum_{j=1}^{m} o_j^2$$

subject to

$$\boldsymbol{w}^\top \boldsymbol{w} = 1$$

will produce the first principal component score on its output.

To solve this problem we form the Lagrangian

$$L = \sum_{j=1}^{m} o_j^2 + \lambda(1 - \boldsymbol{w}^\top \boldsymbol{w}),$$

differentiate it to give

$$\frac{\partial L}{\partial \boldsymbol{w}} = \sum_{j=1}^{m} 2o_j \boldsymbol{x}_j - 2\lambda \boldsymbol{w}$$

and then try to determine what λ is when $\frac{\partial L}{\partial \boldsymbol{w}} = 0$.

If $\frac{\partial L}{\partial \boldsymbol{w}} = 0$, then

$$\lambda \boldsymbol{w} = \sum_{j=1}^{m} o_j \boldsymbol{x}$$

so

$$\lambda \boldsymbol{w}^\top \boldsymbol{w} = \sum_{j=1}^{m} o_j \boldsymbol{w}^\top \boldsymbol{x}_j = \sum_{j=1}^{m} o_j^2$$

and hence (because we have the constraint $\boldsymbol{w}^\top \boldsymbol{w} = 1$)

$$\lambda = \sum_{j=1}^{m} o_j^2.$$

We can now substitute this value into the expression for $\frac{\partial L}{\partial \boldsymbol{w}}$, giving

$$\frac{1}{2}\frac{\partial L}{\partial \boldsymbol{w}} = \sum_{j=1}^{m} o_j \boldsymbol{x}_j - \sum_{j=1}^{m} o_j^2 \boldsymbol{w} = \sum_{j=1}^{m} o_j(\boldsymbol{x}_j - o_j \boldsymbol{w}).$$

Comparing this with the modified version of the plain Hebbian rule given in Chapter 8

$$\Delta w_k^q = \eta o^q (x_k^q - o^q w_k^q)$$

we can see that the modified Hebbian rule is performing a gradient descent for the problem of finding the first principal component.

3.4.2 Correspondence Analysis

Correspondence analysis is a way to represent the structure within *incidence matrices*. Incidence matrices are also called *two-way contingency tables*. The definitive text on this subject in English is [225], but there has been a great deal of work in this area by French data theorists. An example of a (5×4) incidence matrix, with marginal totals, is shown in Table 3.17. The aim is to produce a picture that shows which groups of staff have which smoking habits. There is an implicit assumption that smoking category and staff group are related.

The first step towards this goal is to transform the incidence matrix into something that is related more directly to association between the two variables. Having seen that log-linear modelling in Section 3.2.3 is concerned with association of categorical variables, we might attempt to find some quantity based on the log-linear model for our transformation of the incidence matrix. We do not do this, but we do something very similar. We convert the incidence matrix to a different measure of distance (chi-squared distance), which is based on the

Table 3.17. Incidence of smoking amongst five different types of staff. The data are taken from [225, page 55].

	Smoking Category				
Staff Group	None	Light	Medium	Heavy	Total
Senior Managers	4	2	3	2	11
Junior Managers	4	3	7	4	18
Senior Employees	25	10	12	4	51
Junior Employees	18	24	33	13	88
Secretaries	10	6	7	2	25
Total	61	45	62	25	193

traditional test for association between two categorical variables, the chi-squared test of association. This is the test that was used before log-linear models were developed and is still used if either you do not have a computer handy or you do not want to introduce your clients/students to log-linear modelling.

Notation Denote the incidence matrix as $\underset{(m\times n)}{X}$.

The row totals are

$$X_{j+} = \sum_{i=1}^{n} X_{ji} \quad \text{for } j = 1, \ldots, m,$$

the column totals are

$$X_{+i} = \sum_{j=1}^{m} X_{ji} \quad \text{for } i = 1, \ldots, n$$

and the grand total is

$$X_{++} = \sum_{j=1}^{m} X_{j+} = \sum_{i=1}^{n} X_{+i}.$$

From these totals we can calculate the *expected values* under the assumption that there is no association between the row variable and the column variable—smoking and staff group in the example. The expected values are

$$E_{ji} = \frac{X_{j+}X_{+i}}{X_{++}}.$$

So in the smoking example $E_{23} = \frac{18\times 62}{193}$.

Chi-squared test of association For the smoking example, we can perform a statistical test of the hypothesis

H_0: P(Being in smoking category i) is the same for each staff group.

This boils down to saying that each row is generated by the same smoking distribution, or that rows (and columns) are proportional to one another. This is the same as saying smoking category and staff group are unrelated. A test statistic for this test is Pearson's chi-squared statistic

$$\chi^2 = \sum_{j=1}^{m}\sum_{i=1}^{n}\left\{\frac{(X_{ji} - E_{ji})^2}{E_{ji}}\right\}. \tag{3.33}$$

The bigger χ^2, the greater the evidence against H_0.

Chi-squared distance It turns out that χ^2 can be decomposed in a way that will allow us to highlight the patterns in the incidence matrix. This decomposition arises from consideration of row profiles.

The *row profiles* of an incidence matrix are defined to be the vectors

$$\boldsymbol{p}_j = (p_{j1}, p_{j2}, \ldots, p_{jn})^\top \quad \text{for } j = 1, \ldots, m,$$

where $p_{ji} = X_{ji}/X_{j+}$.
The *mean row profile* of an incidence matrix is the vector

$$\overline{\boldsymbol{p}} = (\overline{p}_1, \overline{p}_2, \ldots, \overline{p}_n)^\top,$$

where $\overline{p}_i = X_{+i}/X_{++}$.
The main structure of interest in the incidence matrix is captured by how the row profiles differ from the mean row profile. For the data in Table 3.17, these differences correspond to the ways in which the proportion of each type of smoker for a particular type of worker vary relative to the average proportions.
We could measure how a row profile differed from the mean row profile by forming the squared (euclidean) distance between them, namely

$$(\boldsymbol{p}_j - \overline{\boldsymbol{p}})^\top (\boldsymbol{p}_j - \overline{\boldsymbol{p}}).$$

In correspondence analysis, we used chi-squared distance instead of the squared distance. The chi-squared distance is

$$d_j^2 = (\boldsymbol{p}_j - \overline{\boldsymbol{p}})^\top D_{\overline{\boldsymbol{p}}}^{-1} (\boldsymbol{p}_j - \overline{\boldsymbol{p}}), \tag{3.34}$$

where $D_{\overline{\boldsymbol{p}}}$ is a diagonal matrix with the elements of $\overline{\boldsymbol{p}}$ along the leading diagonal. In other words,

$$D_{\overline{\boldsymbol{p}}} = \begin{bmatrix} \overline{p}_1 & 0 & \cdots & 0 \\ 0 & \overline{p}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \overline{p}_n \end{bmatrix}$$

and (consequently)

$$D_{\overline{\boldsymbol{p}}}^{-1} = \begin{bmatrix} 1/\overline{p}_1 & 0 & \cdots & 0 \\ 0 & 1/\overline{p}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\overline{p}_n \end{bmatrix}.$$

The chi-squared distance is related to the chi-squared statistic through the formula

$$\chi^2 = \sum_{j=1}^{m} X_{j+} d_j^2.$$

The quantity $\chi_j^2 = X_{j+} d_j^2$ is called the *weighted chi-squared distance* of $\boldsymbol{p}_j$ from $\overline{\boldsymbol{p}}$.

Given the relationship between the chi-squared statistic and the weighted chi-squared distances of the row profiles, we can transform the incidence matrix, X, into a new matrix, Z say, which has the following properties.

Table 3.18. The Z matrix for the incidence matrix given in Table 3.17.

	Smoking Category			
Staff Group	None	Light	Medium	Heavy
Senior Managers	0.020	-0.025	-0.020	0.035
Junior Managers	-0.051	-0.042	0.036	0.079
Senior Employees	0.159	-0.039	-0.078	-0.073
Junior Employees	-0.134	0.055	0.064	0.034
Secretaries	0.054	0.005	-0.026	-0.050

- The squared length of the vector made from the jth row of Z is χ^2_j / X_{++}.
- The direction of this vector indicates how the jth row profile differs from the mean row profile.
- If a particular row profile happens to be the same as the mean row profile, then the vector made from the jth row of Z will consist of zeroes. (In other words, the mean profile corresponds to the co-ordinate origin.)

The (j, i)th element of the Z matrix is

$$Z_{ji} = \sqrt{(X_{j+}/X_{++})}(p_{ji} - \overline{p}_i)\sqrt{\frac{1}{\overline{p}_i}}. \tag{3.35}$$

An alternative, but equivalent, formulation of Z_{ji}, which is more obviously related to the chi-squared statistic of (3.33), is

$$Z_{ji} = \frac{1}{\sqrt{X_{++}}} \times \frac{(X_{ji} - E_{ji})}{\sqrt{E_{ji}}}. \tag{3.36}$$

The formulation in (3.36) also illustrates that we could consider column profiles instead of row profiles, and we would arrive at the same Z matrix. The Z matrix for the data in Table 3.17 is shown in Table 3.18.

We could try to interpret the Z matrix directly, by looking for entries that are large in absolute value. For these data the chi-squared test of association indicates that observed association between type of worker and amount of smoking is not statistically significant (the significance probability, or p-value, is 17%). As a result most of the values in the Z matrix are quite small. There are two relatively high values, 0.159 and -0.134 in the first column. These indicate that the *senior employees* group contains more non-smokers than average and that the *junior employees* group contains fewer.

If the Z matrix were more complicated, then we would seek to draw a diagram to display the structure of the Z matrix. We could apply the same methods as we used for principal components analysis; for principal components analysis we searched for directions of high variance, but for correspondence analysis we seek directions of high weighted chi-squared distance. The matrix

$$Z^\top Z$$

is analogous to the sample variance matrix in principal components analysis. The diagonal elements of $Z^\top Z$ sum to χ^2/X_{++} and indicate how much of the chi-squared statistic is related to each of the four smoking categories; these elements are the analogues of the sample variances for the untransformed variables in principal components analysis. The off-diagonal elements are analogues of the sample covariances.

Having met principal components analysis, we know that the direction of highest weighted chi-squared distance is found by performing an eigen decomposition of $Z^\top Z$. The eigenvectors of $Z^\top Z$ for the data in Table 3.17 are

$$\begin{pmatrix} -0.8087001 & 0.17127755 & 0.0246170 & 0.5621941 \\ 0.1756411 & -0.68056865 & -0.5223178 & 0.4828671 \\ 0.4069601 & -0.04167443 & 0.7151246 & 0.5667835 \\ 0.3867013 & 0.71116353 & -0.4638695 & 0.3599079 \end{pmatrix}$$

the corresponding eigenvalues are

$$(0.0748, 0.0100, 0.0004, 0.0000)$$

and so we can see that most of the chi-squared statistic is accounted for by the first eigenvector. (The eigenvalues sum to the sum of the diagonal elements of $Z^\top Z$, in the same way that the eigenvalues in principal components analysis sum to the sum of the diagonal elements of the sample variance matrix. Here, the sum of the eigenvalues is χ^2/X_{++}.) The first two eigenvectors account for virtually all of the chi-squared statistic.

To convert Z to scores corresponding to the eigenvectors (in the same way that we converted to the principal component scores in principal components analysis) we simply multiply Z by the eigenvector matrix, V say, to get the matrix of row scores, G say. That is we find

$$G = ZV.$$

For the incidence matrix in Table 3.17 we find that G is

$$\begin{pmatrix} -0.01570127 & 0.046251973 & -0.016945718 & 2.775558 \times 10^{-17} \\ 0.07908382 & 0.074303261 & 0.010293294 & -6.245005 \times 10^{-17} \\ -0.19564528 & 0.005479739 & 0.002650324 & -5.204170 \times 10^{-17} \\ 0.15730008 & -0.038991402 & -0.002231942 & -1.387779 \times 10^{-17} \\ -0.07237356 & -0.028400773 & 0.002908443 & -1.387779 \times 10^{-17} \end{pmatrix}.$$

As we discovered that the first two components accounted for most of the chi-squared statistic, we can display the relevant structure of Table 3.17 by plotting the first two columns of G against each other. In addition, we plot the first two columns of V on the same diagram. The resulting plot is shown in Figure 3.7. The technical name for this sort of diagram is a *biplot*; see [222] for uses, interpretation and generalisations of this sort of diagram.

For our purposes, the interpretation of Figure 3.7 is fairly easy. The different categories of staff are all close to the origin; this indicates that their row profiles

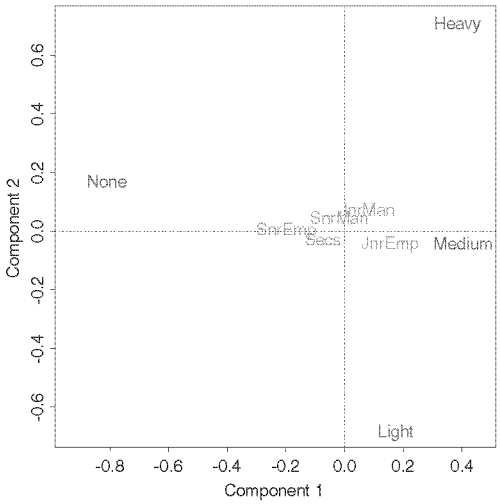

Fig. 3.7. Biplot of the first two components produced by applying correspondence analysis to the data in Table 3.17.

are all similar to the mean row profile, in other words, the differences in smoking habits are not large between different types of staff. We know that the patterns that we see in this diagram are not statistically significant, so perhaps we should leave the interpretation as it stands; for illustration of how to interpret the diagram, we will over interpret it. The *senior employees* group is displaced from the origin in the direction of the *none* category for smoking. This indicates that the proportion of non-smokers is relatively high for senior employees. The *junior employees* group, on the other hand, is shown as being in the opposite direction from the *none* category, indicating a lower proportion of non-smokers than the mean profile. The shortfall in non-smokers for the *junior employees* is taken up by an excess of *medium* smokers, which can be seen from the fact that *junior employees* and *medium* are plotted in the same direction as one another (relative to the origin).

3.4.3 Multidimensional Scaling

Multidimensional scaling is the process of converting a set of pairwise dissimilarities for a set of points, into a set of co-ordinates for the points. The key problem here is that the locations of the points in space, or the *configuration* of the points, is unknown. The aim is to find a configuration in which the pairwise distances between points approximate the dissimilarities as closely as possible.

At first sight, it is difficult to see how the pairwise dissimilarities might be known if the configuration is not known. Examples of dissimilarities could be: the price of an airline ticket between pairs of cities; the difference between the

aggregate scores of pairs of teams in a sports league, the aggregation being across the matches between that pair of teams; road distances between towns (as opposed to straight-line distances); a coefficient indicating how different the artefacts found in pairs of tombs within a graveyard are. For the airline ticket example, the idea would be to construct a map in which the distances correspond to cost of travel between the cities; we might wish to compare this map with a standard map showing the geographical locations of the cities, but we are not trying to reconstruct the geographical map. Similarly, with road distances, we would be trying to construct a map in which distance corresponds to how far one has to drive a car to travel between the towns.

Two types of multidimensional scaling will be described here. *Classical scaling* assumes that the dissimilarities are euclidean distances (that is, distances like those that we meet in everyday life, with the same geometric properties). *Ordinal scaling* assumes merely that the dissimilarities are in the same order as the distances between points in the configuration; if the dissimilarity between two points, A and B, is 0.5 and the dissimilarity between A and a third point, C, is 1, then this says that the distance between A and C in the configuration should be bigger than the distance from A to B, but not necessarily twice as big.

Classical Scaling Classical scaling is also known as *metric scaling* and as *principal co-ordinates analysis*. The name 'metric' scaling is used because the dissimilarities are assumed to be distances—or in mathematical terms the measure of dissimilarity is the *euclidean metric*. The name 'principal co-ordinates analysis' is used because there is a link between this technique and principal components analysis. The name 'classical' is used because it was the first widely used method of multidimensional scaling, and pre-dates the availability of electronic computers.

To derive a configuration, we can start off by imagining that we knew the configuration and working out what the pairwise distances would be, and then trying to reverse the process. So, let X be the matrix whose rows specify the positions of the m objects in the configuration. That is

$$\underset{(m\times c)}{X} = \begin{bmatrix} \boldsymbol{x}_1^\top \\ \boldsymbol{x}_2^\top \\ \vdots \\ \boldsymbol{x}_m^\top \end{bmatrix},$$

where $\boldsymbol{x}_j$ is the $(c \times 1)$ vector representing the position of the jth object in c-dimensional space; if $c = 2$ then all the objects lie on a plane, if $c = 3$ then they occupy a 3-D space, just as we do.

This configuration would lead to a matrix of pairwise distances, D say,

$$D = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1m} \\ d_{21} & d_{22} & \cdots & d_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ d_{m1} & d_{m2} & \cdots & d_{mm} \end{bmatrix},$$

where $d_{k\ell}$ is the distance between objects k and ℓ. The elements of D are obtained from X using the following relationship:

$$d_{k\ell}^2 = \boldsymbol{x}_k^\top \boldsymbol{x}_k + \boldsymbol{x}_\ell^\top \boldsymbol{x}_\ell - 2\boldsymbol{x}_k^\top \boldsymbol{x}_\ell. \tag{3.37}$$

What we want to do is go from D to X. A step on the way is the matrix

$$B = XX^\top = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mm} \end{bmatrix}$$

because it is easy to construct an appropriate configuration given B, and we can rewrite (3.37) in terms of the elements of B as in

$$d_{k\ell}^2 = b_{kk} + b_{\ell\ell} - 2b_{k\ell}. \tag{3.38}$$

If we make an assumption, then (3.38) can be rearranged to give the $b_{k\ell}$ in terms of the $d_{k\ell}$'s. The assumption that we need to make is that the origin of the co-ordinate system is at the mean of the configuration; this is equivalent to assuming that the columns of X have been centred. Specifying where the co-ordinate origin is cannot alter the pairwise distances, so we can make this assumption, without loss of generality. The point of this assumption is that it implies that

$$\mathbf{1}_m^\top X = \mathbf{0}_c^\top,$$

where $\mathbf{1}_m$ denotes the $(m \times 1)$ vector whose elements are all equal to 1, and $\mathbf{0}_c$ is a similar $(c \times 1)$ vector of zeroes. So,

$$\mathbf{1}_m^\top B = \mathbf{0}_c^\top X^\top = \mathbf{0}_m^\top,$$

and

$$B\mathbf{1}_m = \mathbf{0}_m.$$

This is equivalent to saying that the rows and columns of B sum to zero. In other words,

$$\sum_{k=1}^{m} b_{k\ell} = \sum_{\ell=1}^{m} b_{k\ell} = 0.$$

This result allows us to sum (3.38) over ℓ to get

$$\begin{aligned} \sum_{\ell=1}^{m} d_{k\ell}^2 &= mb_{kk} + \sum_{\ell=1}^{m} b_{\ell\ell} - (2 \times 0) \\ &= mb_{kk} + \sum_{\ell=1}^{m} b_{\ell\ell}, \end{aligned}$$

which can in turn be summed over k, to give

$$\sum_{k=1}^{m}\sum_{\ell=1}^{m} d_{k\ell}^2 = \sum_{k=1}^{m} mb_{kk} + m\sum_{\ell=1}^{m} b_{\ell\ell} = 2m\sum_{k=1}^{m} b_{kk}.$$

Thus, we have a formula for $\sum_{k=1}^{m} b_{kk}$ (or, equivalently, $\sum_{\ell=1}^{m} b_{\ell\ell}$) in terms of the $d_{k\ell}$'s, which we can use to obtain a formula for b_{kk} in terms of the $d_{k\ell}$'s, which we can substitute into (3.38) to give a formula for $b_{k\ell}$ in terms of the $d_{k\ell}$'s.

If we obtain all these formulae, we find that the route from D to B is to calculate the matrix

$$D^* = \begin{bmatrix} d_{11}^2 & d_{12}^2 & \cdots & d_{1m}^2 \\ d_{21}^2 & d_{22}^2 & \cdots & d_{2m}^2 \\ \vdots & \vdots & \ddots & \vdots \\ d_{m1}^2 & d_{m2}^2 & \cdots & d_{mm}^2 \end{bmatrix},$$

then use the formula

$$B = -\frac{1}{2}\left(I_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^\top\right) D^* \left(I_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^\top\right)$$

to find B.

To get a configuration that has the same pairwise distances as in D, we need to find a matrix, X, which satisfies

$$XX^\top = B$$

and any such matrix will do. The choice made in classical scaling is the matrix

$$X^* = \underset{(m\times c)}{E} \begin{bmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\lambda_c} \end{bmatrix}, \tag{3.39}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_c > 0$ are the non-zero eigenvalues of B and the ith column of E is $\boldsymbol{e}_i$, the eigenvector corresponding to λ_i. These eigenvectors can be, and are, chosen to satisfy (3.32) on page 103, that is, to be *orthonormal*.

This method of choosing X from B is called the *spectral decomposition* of B. The eigenvalues of B will all be greater than or equal to zero, as long as the distances in D are euclidean. Here, we can regard euclidean as meaning just like distances in the physical world that we live in. If any eigenvalue turns out to be negative, then we can either just ignore the problem, or use ordinal scaling (see below) instead.

The main advantage of using the spectral decomposition of B to find a configuration is that X^* turns out to be the matrix containing the principal components scores that would be obtained if we knew X and performed principal components analysis upon it. This is why classical scaling is sometimes called principal co-ordinates analysis. This means that if we want, for example, a 2-D map in which the distances between points in the configuration most closely approximate the distances in D, then all we need do is plot the first two columns of X^* against one another. This is equivalent to plotting the first two principal components against one another.

The results of applying classical scaling to British road distances are shown in Figure 3.8. These road distances correspond to the routes recommended by the

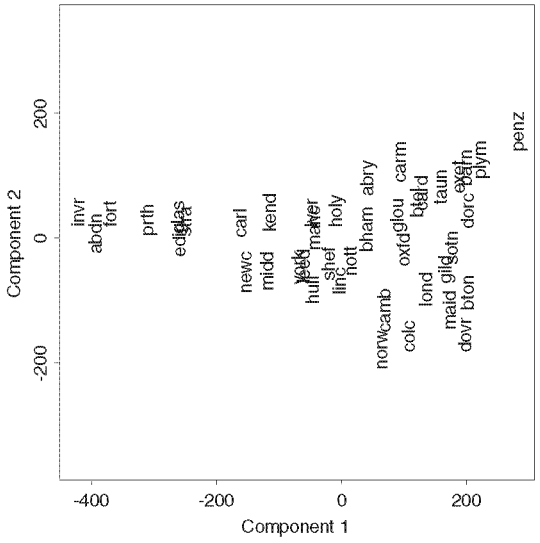

Fig. 3.8. A 2-D map of the island of Great Britain, in which the distances are approximations to the road distances between the towns and cities labelled. The labels are rotated so as to reduce the amount of overlap between labels.

Automobile Association; these recommended routes are intended to give the minimum travelling time, not the the minimum journey distance. An effect of this, that is visible in Figure 3.8 is that the towns and cities have lined up in positions related to the motorway (high-speed road) network. I can see the places on the M6 motorway, Birmingham (*bham*), Manchester (*manc*), Kendal (*kend*), Carlise (*carl*), and its continuation the A74 which goes to Glasgow (*glas*), all appear along a straight line in the configuration. Similarly, the M1 and A1 routes up the eastern side of England can be seen by the alignment of Nottingham (*nott*), Sheffield (*shef*), Leeds (*leed*), York (*york*), Middlesborough (*midd*) and Newcastle (*newc*); the line from Gloucester (*glou*) to Plymouth (*plym*) corresponds to the M5. The map also features distortions from the geographical map such as the position of Holyhead (*holy*), which appears to be much closer to Liverpool (*lver*) and Manchester than it really is, and the position of Cornish peninsula (the part ending at Penzance, *penz*) is further from Carmarthen (*carm*) than it is physically. These distortions are due to features such as large water bodies (between Carmarthen and Penzance) making the road distance much bigger than the straight-line distance, or because a place (Holyhead) is isolated. The isolation of Holyhead results in there being just one fast road to it, and most routes from the South of England (places with high component 1 values) to Holyhead go through Birmingham; this means that most of the distances to Holyhead are the same as the distance to Birmingham plus a constant amount (because most of the places are in the South of England). The best compromise in terms of getting

most distances approximately right is to place Holyhead north of Birmingham, that is, in the same area as Manchester and Liverpool.

Ordinal Scaling Ordinal scaling is used for the same purposes at classical scaling, but for dissimilarities that are not metric, that is, they are not what we would think of as distances. Ordinal scaling is sometimes called *non-metric scaling*, because the dissimilarities are not metric. Some people call it *Shepard-Kruskal scaling*, because Shepard and Kruskal are the names of two pioneers of ordinal scaling.

In ordinal scaling, we seek a configuration in which the pairwise distances between points have the same rank order as the corresponding dissimilarities. So, if $\delta_{k\ell}$ is the dissimilarity between points k and ℓ, and $d_{k\ell}$ is the distance between the same points in the derived configuration, then we seek a configuration in which

$$d_{k\ell} \leq d_{ab}$$

if

$$\delta_{k\ell} \leq \delta_{ab}.$$

If we have such a configuration to hand, then all is well and the problem is solved. If we do not have such a configuration, or no such configuration exists, then we need a way to measure how close a configuration is to having this property. The quantity used to measure how close a configuration is to having this property is called *STRESS*, which is zero if the configuration has this property and greater than zero if it does not.

The definition of STRESS is omitted in the interests of brevity; see [326, page 117] for the definition. Basically, the STRESS measures the size of the smallest change in the $d_{k\ell}$'s required to obtained distances with the correct ordering. A set of fitted distances, $\hat{d}_{k\ell}$'s, say, that are in the correct order is created. The STRESS is

$$\sum_k \sum_\ell \left(d_{k\ell} - \hat{d}_{k\ell}\right)^2.$$

The part of the definition that is omitted here is how to obtain the $\hat{d}_{k\ell}$'s, which are themselves the result of an optimisation procedure. Whilst the $d_{k\ell}$'s are distances between points in a configuration, the $\hat{d}_{k\ell}$'s used in calculating the STRESS do not generally correspond to any valid configuration; the STRESS is purely a measure of how far the distances are from being in the correct order.

The principle of using STRESS was introduced by [324]. Given that we have STRESS, we can think of it as a function of the co-ordinates of the points in the configuration. We can then start from some arbitrary configuration and iteratively improve it. A practical steepest descent algorithm is given by [325]; obtaining the derivative of STRESS is non-trivial.

Armed with STRESS and the algorithm for minimising it, ordinal scaling boils down to:

- choose a dimensionality, c say;
- choose a c-dimensional configuration;
- update the configuration iteratively, so as to minimise the value of STRESS.

Often classical scaling is applied directly to the dissimilarities, or some transformation of them, to obtain a starting configuration. This is believed to both shorten computing time and increase the chance of finding the globally optimal configuration. Sometimes a higher dimensionality than is required, $c + 1$ say, is used during the minimisation. The aim of this is to allow the configuration to change in ways that would not be allowed in c dimensions. For example, in a 2-D configuration a point might be on the right of the configuration, when its optimal position is on the left. In this situation it is possible that moving the point to the left is prevented because small movements to the left increase the STRESS. If the same configuration is then embedded in a 3-D space, then it is sometimes possible for the move from the right to the left to be made by using a route that comes out of the 2-D plane. These strategies generally work well, but they are not guaranteed to produce the global optimum; if it is vital that the globally optimal configuration be found, then the techniques described in Chapter 10 might be used.

Choice of c is sometimes part of the problem. If this is so, then a range of values of c are tried. The minimised STRESS values are then plotted against c and this gives a diagram like a scree diagram (see Figure 3.5), in which the 'elbow' is sought. In this procedure, it is usual to start with the highest value of c under consideration. The optimum configuration for c dimensions would be used to give a starting configuration for $c - 1$ dimensions, either by simply discarding the cth dimension from the co-ordinates of the configuration, or by some transformation such as principal components analysis (see Section 3.4.1).

3.4.4 Cluster Analysis and Mixture Decomposition

Cluster analysis and mixture decomposition are both techniques to do with identification of concentrations of individuals in a space.

Cluster Analysis Cluster analysis is used to identify groups of individuals in a sample. The groups are not pre-defined, nor, usually, is the number of groups. The groups that are identified are referred to as *clusters*. There are two major types of clustering, *hierarchical* and *non-hierarchical*; within hierarchical clustering, there are two main approaches, *agglomerative* and *divisive*.

In hierarchical clustering, we generate m different partitions of the sample into clusters, where m is the number of individuals in the sample. One of these partitions corresponds to a single cluster made up of all m individuals in the sample, while at the opposite extreme there is a partition corresponding to m clusters, each made up of just one individual. Between these extremes there is a partition with 2 clusters, one with 3 clusters, and so on up to a partition with $m - 1$ clusters. The key characteristic of these partitions, which makes them hierarchical, is that the partition with r clusters can be used to produce

the partition with $r-1$ clusters by merging two clusters, and it can also be used to produce the partition with $r+1$ clusters by splitting a cluster into two.

As we know both the top layer and the bottom layer of the hierarchy, there are two natural approaches to finding the intervening layers. We could start with m clusters, each containing one individual, and merge a pair of clusters to get $m-1$ clusters, and continue successively merging pairs of clusters; this approach is called agglomerative clustering. Alternatively, we could start with a single cluster, split it into two, then split one of the new clusters to give a total of three clusters, and so on; this approach is called divisive clustering.

Agglomerative clustering has been preferred traditionally, because the number of partitions considered in building the hierarchy is much smaller than for divisive clustering; the number of partitions considered is cubic in m for agglomerative, but exponential in m for divisive.

In order to decide which clusters to merge, or which clusters to split, we need a way to measure distance between clusters. The usual methods for measuring distance between clusters are given below.

- **Minimum distance** or *single-link* defines the distance between two clusters as the minimum distance between an individual in the first cluster and an individual in the second cluster.
- **Maximum distance** or *complete-link* defines the distance between two clusters as the maximum distance between an individual in the first cluster and an individual in the second cluster.
- **Average distance** defines the distance between two clusters as the mean distance between an individual in the first cluster and an individual in the second cluster, taken over all such pairs of individuals.
- **Centroid distance** defines the distance between two clusters as the squared distance between the mean vectors (that is, the centroids) of the two clusters.
- **Sum of squared deviations** defines the distance between two clusters as the sum of the squared distances of individuals from the joint centroid of the the two clusters minus the sum of the squared distances of individuals from their separate cluster means.

These different methods of measuring distance between clusters lead to different characteristics in the clusters. This means that you ought to choose a method that is appropriate, rather than trying as many as you can. The characteristics of the various methods, and hence how to choose between them, are described and discussed in [326, section 3.1] and [235, section 7.3.1].

The results of a hierarchical cluster analysis are almost always presented as a *dendrogram*, such as that in Figure 3.9. The dendrogram in Figure 3.9 is the result of applying complete linkage clustering (agglomerative clustering where the distance between clusters is *maximum distance*, also called *complete-link*, as defined above) to a set of nine individuals, which are labelled 1 to 9. If we consider the partition giving two clusters, we can see that one cluster consists of individuals 1–4 and the other consists of 5–9. We can see this by following the two vertical lines which are linked at the top of the diagram, down to the labels for the individuals at the bottom of the diagram. Similarly, the partition

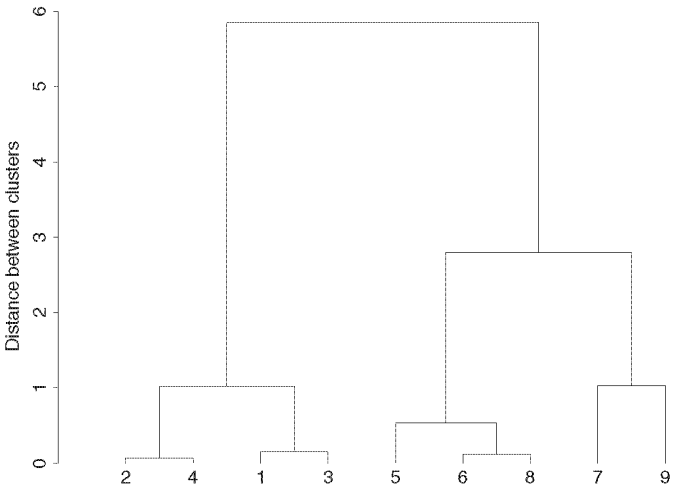

Fig. 3.9. An example of a dendrogram.

corresponding to three clusters gives the clusters as: 1–4; 5, 6 and 8; 7 and 9. We can see this by looking for a height where there are three vertical lines (around about 2 on the vertical scale) and tracing the three vertical lines down to the individuals. The vertical scale indicates the distance between the two clusters being merged. So, for example, in the two-cluster partition, the clusters are just under 6 units apart, while for the three-cluster partition the two closest clusters are just under 3 units apart.

Non-hierarchical clustering is essentially trying to partition the sample so as to optimize some measure of clustering. The choice of measure of clustering is usually based on properties of sums of squares and products matrices, like those met in Section 3.3.1, because the aim in the MANOVA is to measure differences between groups. The main difficulty here is that there are too many different ways to partition the sample for us to try them all, unless the sample is very small (around about $m = 10$ or smaller). Thus our only way, in general, of guaranteeing that the global optimum is achieved is to use a method such as branch-and-bound, as used by [510], for example.

One of the best known non-hierarchical clustering methods is the *k-means* method of [362]. The k-means method starts with k clusters and allows each individual to be moved from its current cluster to another cluster. Individuals are moved between clusters until it becomes impossible to improve the measure of clustering. An algorithm for performing this method of clustering is given by [248], which uses the sum of the squared distances of individuals from their cluster centroids as the measure of clustering (the smaller the better). There is no guarantee that the global optimum will be achieved.

A description of the various criteria that are used and the optimisation algorithms available can be found in [235, section 7.4].

Mixture Decomposition Mixture decomposition is related to cluster analysis in that it is used to identify concentrations of individuals. The basic difference between cluster analysis and mixture decomposition is that there is an underlying statistical model in mixture decomposition, whereas there is no such model in cluster analysis. The probability density that has generated the sample data is assumed to be a mixture of several underlying distributions. So we have

$$f(\boldsymbol{x}) = \sum_{k=1}^{K} w_k f_k(\boldsymbol{x}; \boldsymbol{\theta}_k),$$

where K is the number of underlying distributions, the f_k's are the densities of the underlying distributions, the $\boldsymbol{\theta}_k$'s are the parameters of the underlying distributions, the w_k's are positive and sum to one, and f is the density from which the sample has been generated.

The reason for considering such a model is that we want to summarise the mechanism that generated the sample by stating which probability distribution produced the sample. The problem overcome by using a mixture distribution is that f might not correspond to a standard distribution with well known properties. The f_k's are chosen so that they do correspond to standard distributions, usually multivariate normal distributions.

We can consider this sort of model either because we believe that there are K sub-populations mixed together to give the population from which our sample has been drawn, or simply because it is a pragmatic way of summarising a non-standard distribution.

The $\boldsymbol{\theta}_k$'s and the w_k's can be estimated using iterative methods. Descriptions of these methods and examples of the use of mixture decomposition can be found in [235, chapter 3].

3.4.5 Latent Variable and Covariance Structure Models

A *latent* variable is a variable that is not included in our set of data, either because it cannot be measured directly, or because it was not thought of, but the latent variable is related to the variables that are included in the data set. A latent variable is often of interest because it is believed that the latent variable causes changes in the values of several of the variables that we have observed; this underlying cause of variation induces a covariance structure between the variables that it changes.

The (k, ℓ)th off-diagonal element of the variance-covariance matrix of variables x_1 to x_n contains the covariance between the kth and ℓth variable. The diagonal elements of the matrix contain the variances of the variables. In this section we examine ways to model the structure of such a matrix. There are various reasons for wanting to do this, but an important one is simply that it

provides a way to reduce the number of parameters from $n(n+1)/2$ to something more manageable. This can result in a model which is easier to understand and also one which has improved predictive power.

Several such methods for modelling variance-covariance matrices have been developed, with differing complexity and power. All of them have convenient graphical (in the mathematical sense) representations, in which the nodes of a graph represent the variables and edges connecting the nodes represent various kinds of relationships between the variables (the nature of the relationship depending upon the type of graph in question). We shall begin with the simplest case—that of *path analysis*—which goes back to the work of [548, 549].

Path Analysis In path analysis the aim is to decompose the relationships between the observed variables into *causal paths*, and attribute strengths to each of these paths. To start with, we shall assume that

(a) the variables are known to have a *weak causal order*. This means that we can order the variables such that a given variable may (but need not) be a cause of variables later than itself in the list, but cannot be a cause of variables earlier than itself in the list; and
(b) *causal closure* holds. This means that the covariation between two of the measured variables is due to the causal effects of one on the other or to the effects on both of them of some other variable(s) in the model.

Of course, in order to make this rigorous we really ought to define what we mean by *cause*. Since this has occupied philosophers since the dawn of time, we shall content ourselves with the following working definition: *x is a cause of y if and only if the value of y can be changed by manipulating x and x alone*. Note that the effect of x on y may be *direct* or it may be *indirect*. In the latter case, the effect arises because x affects some *intermediate* variable z which in turn affects y.

If a unit change in x directly induces a change of c units in y then we will write $y = cx$. Conventionally, if unstandardised variables are used then the coefficient is called an *effect coefficient*, written as c, while if standardised variables are used the coefficient is called a *path coefficient*, written as p. Note the distinction between these and regression coefficients (the β's): a regression coefficient simply tells us the expected difference in y between two groups that happen to have a unit difference on x. This may be direct causal, indirect causal, or it may be due to some selection or other mechanism. We shall see examples of the difference between path (denoted p) and regression (denoted b) coefficients below. To begin with, however, Figure 3.10 shows four possible patterns of causal relationships between three variables x, y, and z when they have the weak causal order $x > y > z$ (the four examples in Figure 3.10 do not exhaust the possibilities). The arrows indicate the direction of causation.

In Figure 3.10(a), x causes y, which in turn causes z. If y is fixed (by, for example, selecting only those cases which have a given value of y), then x and z are conditionally independent (this follows from assumption (b) above). In

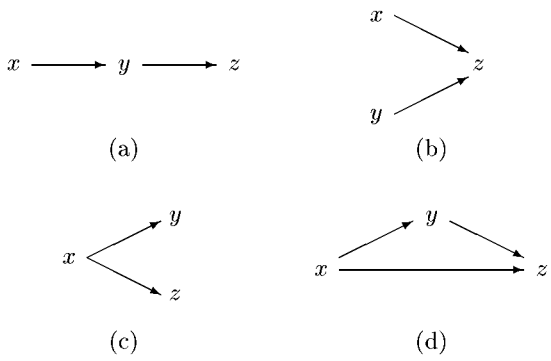

Fig. 3.10. Four of the possible patterns of causal relationships between three variables, x, y and z.

Figure 3.10(b), x and y separately have a causal effect on z, but x and y are marginally independent (there is no arrow linking them). Note, however, that they may not be conditionally independent given values of z. In Figure 3.10(c), x has a causal effect on y and a causal effect on z. Here y and z are conditionally independent given x, but if x is uncontrolled then they may have an induced covariance.

Figure 3.10(d) is more interesting than the others, so we will develop it in a little more detail, with illustrative numbers. First, however, note that it is clear from the description of the first three models that the notion of conditional independence has an important role to play in disentangling relationships between variables—we will have more to say about this below. Secondly, note that, given the two assumptions above, one can estimate the path or effect coefficients between n variables by a set of $(n-1)$ multiple regression analyses. That is, starting with (one of) the variable(s) at the end of the weak causal order, one can construct a regression model to predict it in terms of all the variables which precede it. This is repeated for all variables in turn, working back towards the beginning of the order. Given this modelling process, it is reasonable that the resulting structures should be called *univariate recursive regression models.*

In Figure 3.10(d) the path coefficients for the pairs (y, z) and (x, z) are estimated by regressing z on x and y. That is, the model $z = p_{(z,y)} \cdot y + p_{(z,x)} \cdot x$. Suppose that, if we do this, we obtain the regression (and here also path) coefficients $p_{(z,y)} = -0.2$ and $p_{(z,x)} = -0.3$. Similarly, we can regress y on x to obtain the path coefficient for the pair (x, y)—yielding 0.8, say. Finally, we can also regress z just on x. This regression coefficient will not be a path coefficient, since it will include direct effects of x on z and also indirect effects via y. Decomposing the overall effect of x on z in this way, we find that the regression coefficient is $-0.46 = (-0.3) + (0.8)(-0.2)$. Here the first term is the direct effect and the second term the indirect via y.

Finally, in this model, let us look at the covariation between y and z. This does not simply arise from the regression effect of y on z. There is also a 'spurious' effect arising from the effect of x on both y and z, so that the overall covariation is given by $(-0.2) + (0.8)(-0.3) = -0.44$.

This sort of process yields a decomposition of the overall covariation between any two variables into direct causal relationships, indirect causal relationships, and noncausal relationships (induced by the effect of other variables on the two in question).

The set of variables in a path model can be divided into two groups: those which have no predecessors, called *exogenous* variables, and those which are predicted by others in the model, called *endogenous* variables. Denoting the former by $\boldsymbol{z}$ and the latter by $\boldsymbol{y}$, we can express a path model as

$$\boldsymbol{y} = C\boldsymbol{y} + D\boldsymbol{z} + \boldsymbol{e}$$

where C and D are matrices of regression coefficients and $\boldsymbol{e}$ is a vector of random error terms. Note that assumption (a) means that there are no causal loops.

Up until this point things have been straightforward, but this ease of manipulation and interpretation has been bought by assumptions (a) and (b), which are fairly drastic. Often it is not so easy to assert a causal order between variables (straightforward for *income* and *expenditure*, but not so straightforward for *depression* and *quality of life*). This means that one would sometimes like to relax things—replacing the arrows between the nodes representing the variables by an undirected edge. Likewise, the assumption that the model is complete, that all indirect links between two variables have been included, is often hard to support. To cope with problems such as these, path models have been extended. One very important extension is to include *latent* variables.

Latent Variables A latent variable is an unmeasured (and typically unmeasurable) variable which accounts for and explains the observed correlations between the *manifest* or measured variables. That is, the observed variables are postulated to have high correlations with each other precisely because they have high correlations with the unmeasured latent variable. Latent variables can be considered as convenient mathematical fictions to explain the relationships between the manifest variables, or attempts can be made to interpret them in real terms. Social class, intelligence, and ambition are examples of latent variables which arise in the social and behavioural sciences. One of the earliest types of latent variable models was the *factor analysis model* (originally developed as a model for intelligence in the early decades of the twentieth century, when a 'general' intelligence factor was postulated). In mathematical terms, the manifest variables $\boldsymbol{x}$ are to be explained in terms of a number of latent variables $\boldsymbol{y}$, assumed to follow a distribution $\mathcal{N}(\boldsymbol{0}, I)$, so that we have

$$\boldsymbol{x} = \boldsymbol{\mu} + \Gamma\boldsymbol{y} + \boldsymbol{e}$$

where Γ is a matrix of *factor loadings*, and $\boldsymbol{e}$ is a random vector typically assumed to follow a distribution $\mathcal{N}(\boldsymbol{0}, \Theta)$. If the covariance matrix of the manifest

variables is Σ, then we have

$$\Sigma = \Gamma\Gamma^\top + \Theta.$$

(We note parenthetically here that factor loading matrices of the form $\Gamma M^\top$, where M is a non-singular orthogonal matrix, will also satisfy this expression. This means that it is the space spanned by the factors which is unique, and not the individual factors themselves. This ambiguity is probably one of the reasons that factor analysis acquired a reputation of not being an entirely respectable technique, a reputation which is no longer deserved—the method is now well-understood and can shed valuable light on data structures.)

Nowadays the most common method of estimating the factor loadings is probably via maximum likelihood.

Linear Structural Relational (LISREL) Models In recent decades, the basic factor analysis model has been substantially developed into the general class of linear structural relational (or LISREL) models; see, for example [74, 161]. These often involve multiple latent variables, and either hypothesise or set out to discover relationships between the latent variables themselves, as well as between the latent and the manifest variables. The applicability of such models is immediately seen when one recognises that no measurements are made without error (especially in the human sciences, where these methods were originally developed and continue to be heavily applied). This means that the measurement actually observed is, in fact, merely a manifest version of an unobserved latent variable, so that a more elaborate (and correct) model will consist of a *measurement model* describing the relationships between the measured and latent variables and a *structural model* describing the relationships between the latent variables themselves.

Examples of important kinds of LISREL models are *multitrait multimethod* models and *multiple indicator multiple cause* models. In the former, several hypothesised traits are measured by several methods. The model is used to tease apart effects due to the trait being measured and effects due to the method of measurement used. In the latter, unobserved latent variables are influenced by several causes and the effects are observed on several indicators. This is proposed by [174] as a suitable model for quality of life measurements, where some of the measured variables are indicative of poor quality of life while others cause poor quality of life.

Recent Developments Up until this point, the methods we have been describing have their origins in structural relationships between the variables. In the last decade or so, however, an alternative type of model has attracted considerable interest. Models of this type go under various names—conditional independence graphs, Bayesian belief networks, graphical models, and so on; see, for example [540, 165, 337]. They are described in more detail in Chapter 4. They differ from the models above because they are based on modelling the joint probability

distribution of the observed variables in terms of conditional independence relationships. Such models have been developed in greatest depth for multivariate normal data and for categorical data—Figure 3.2 showed a conditional independence graph for a categorical data set. With multivariate normal data, the inverse of the variance-covariance matrix has a particularly useful interpretation: the off-diagonal elements of the matrix give the conditional covariances between the variables. (The conditional covariances are often called *partial covariances.*) In particular, it follows that if a given off-diagonal element is zero then the two variables in question are independent given the other variables in the matrix. In terms of a graphical representation, this corresponds to the two variables not being connected by an edge in the graph. Note that, unlike the structural models introduced above, conditional independence graphs are essentially undirected—though the relationships between undirected graphs and directed graphs is now well understood.

To develop a conditional independence model, two sorts of information are required. The first is the topology of the graph—the edges linking the nodes in the pictorial representation and the conditional independence structure in the mathematical formalism. The second is the numerical detail of the conditional dependencies: if variable A depends on variables B, C, and D, what does the conditional distribution of A look like for each combination of the levels of B, C, and D? If sufficient human expertise is available, then, of course both of these can be obtained by interrogation. Otherwise data analytic methods are required. (In fact, conditional independence models have their origins in the Bayesian school of inference, so that it is common to find prior expertise and data integrated in developing the models.) Extracting the numeric details of the conditional independencies from data is, in principle, straightforward enough. This is not quite so true for constructing the model's basic structure since the space of possible models is often vast. To do this effectively requires a large data set. This is currently a hot research topic.

3.5. Conclusion

The techniques presented in this chapter do not form anything like an exhaustive list of useful statistical methods. These techniques were chosen because they are either widely used or ought to be widely used. The regression techniques are widely used, though there is some reluctance amongst researchers to make the jump from linear models to generalized linear models.

The multivariate analysis techniques ought to be used more than they are. One of the main obstacles to the adoption of these techniques may be that their roots are in linear algebra.

I feel the techniques presented in this chapter, and their extensions, will remain or become the most widely used statistical techniques. This is why they were chosen for this chapter.