Chapter 3
Statistical Methods

Paul C. Taylor
University of Hertfordshire

28th March 2001

3.1 Introduction

- Generalized Linear Models
- Special Topics in Regression Modelling
- Classical Multivariate Analysis
- Summary

3.2 Generalized Linear Models

- Regression
- Analysis of Variance
- Log-linear Models
- Logistic Regression
- Analysis of Survival Data

The fitting of generalized linear models is currently the most frequently applied statistical technique. Generalized linear models are used to described the relationship between the mean, sometimes called the *trend*, of one variable and the values taken by several other variables.

3.2.1 Regression

How is a variable, y, related to one, or more, other variables, $x_1, x_2, \ldots, x_n$?

Names for y:

response; *dependent variable*; *output*.

Names for the x_i's:

regressors; *explanatory variables*; *independent variables*; *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 *(general) linear model* is

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

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 model is *linear* in the β's.

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

(A weighted sum of the β's.)

The main reasons for the use of the linear model.

- The maximum likelihood estimators of the β's are the same as the least squares estimators; see Section 2.4 of Chapter 2.

- Explicit formulae and rapid, reliable numerical methods for finding the least squares estimators of the β's.

- Many problems can be framed as general linear models. For example,

$$\mathrm{E}(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 \ , \qquad (3.3)$$

  can be converted by setting $x_3 = x_1 x_2$, $x_4 = x_1^2$ and $x_5 = x_2^2$.

- 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.

Non-linear Regression

Two examples are:

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

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

where the ε's and m are as in (3.1).

Problems

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 over-come 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.

Generalized Linear Models

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.

2. Instead of the expected value of the output being a linear function of the β's, we have

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

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

There is a reliable general algorithm for fitting generalized linear models.

Generalized Additive Models

Generalized additive models are a generalization of generalized linear models.

The generalization is that $g\left(\mathsf{E}(y_j)\right)$ need not be a linear function of a set of β's, but has the form

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

where the s_i's are arbitrary, usually smooth, functions.

An example of the model produced using a type of scatterplot smoother is shown in Figure 3.1.

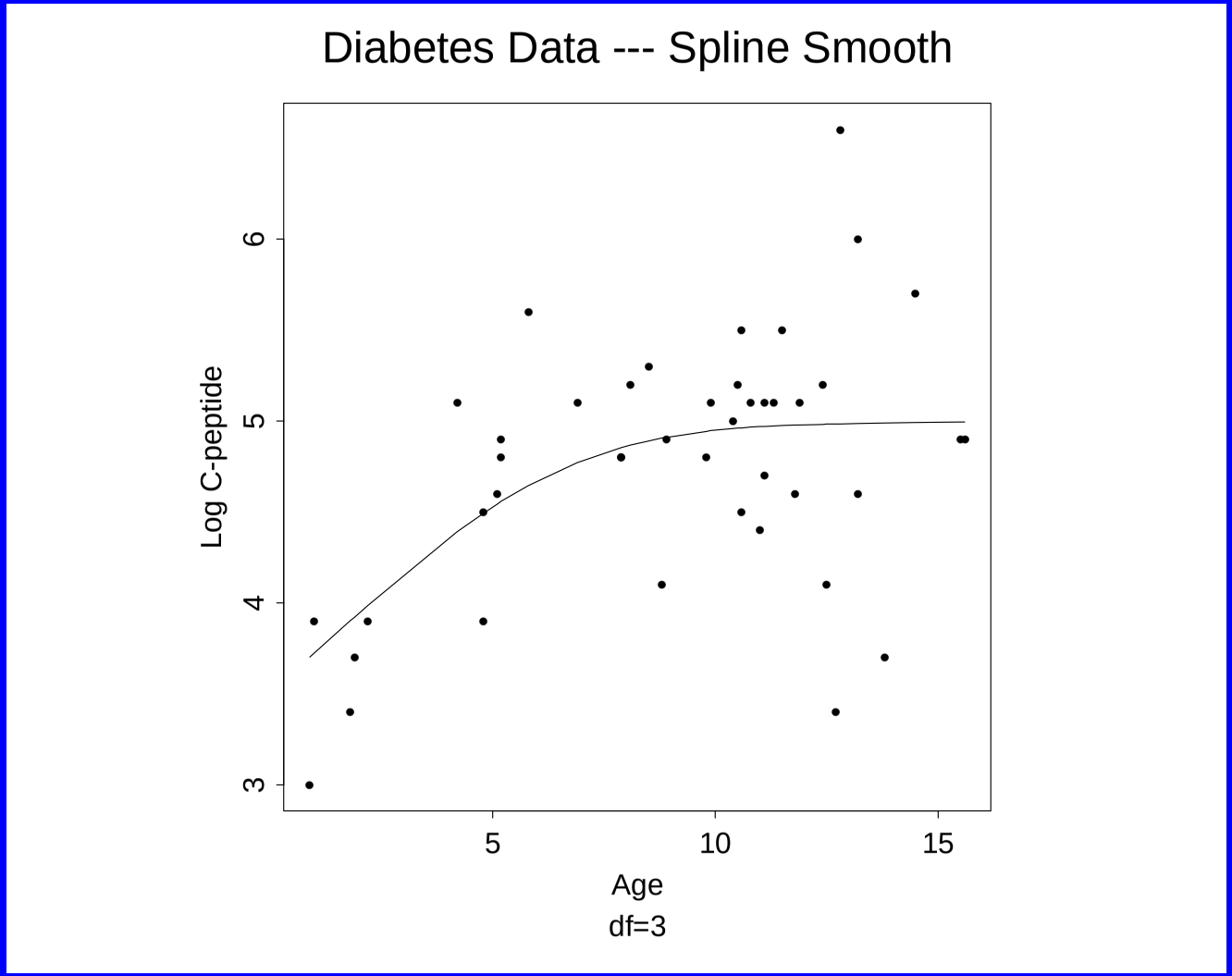
Diabetes Data --- Spline Smooth
Log C-peptide
3
4
5
6
5
10
15
Age
df=3

Figure 3.1

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.

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.

Example 27 Turnips for Winter Fodder. The data in Table 3.1 are from an experiment to investigate the growth of turnips. These types of turnips 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.

Table 3.1

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

The following linear model

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

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.

On first sight, (3.8) should also include a β_A and a β_I, but we do not need them.

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),

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

The estimate of β_i is $\widehat{\beta}_i$.

Fitted values

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

Residuals

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

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

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

The key facts about S^2 is 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. Note that omitting one of the inputs from the model is equivalent to forcing the corresponding β to be zero.

Example 28 Turnips for Winter Fodder continued. Let Ω_1 to be the model at (3.8), and Ω_0 to be the following model

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

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

Table 3.2

	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

	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		

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.

Table 3.4

	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		

3.2.3 Log-linear Models

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 one of three health centres for the treatment of breast cancer;

age the age of the patient when her breast cancer was diagnosed;

survived whether the patient survived for at least three years from diagnosis;

appear appearance of the patient's tumour—either *malignant* or *benign*;

inflam amount of inflammation of the tumour—either *minimal* or *greater*.

Table 3.7

			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

For these data, the output is the number of patients in each cell.

The model is

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

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

Table 3.8

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

To summarise this model, I would construct its conditional independence graph and present tables corresponding to the interactions.
Tables are in the book.
The conditional independence graph is shown in Figure 3.2.

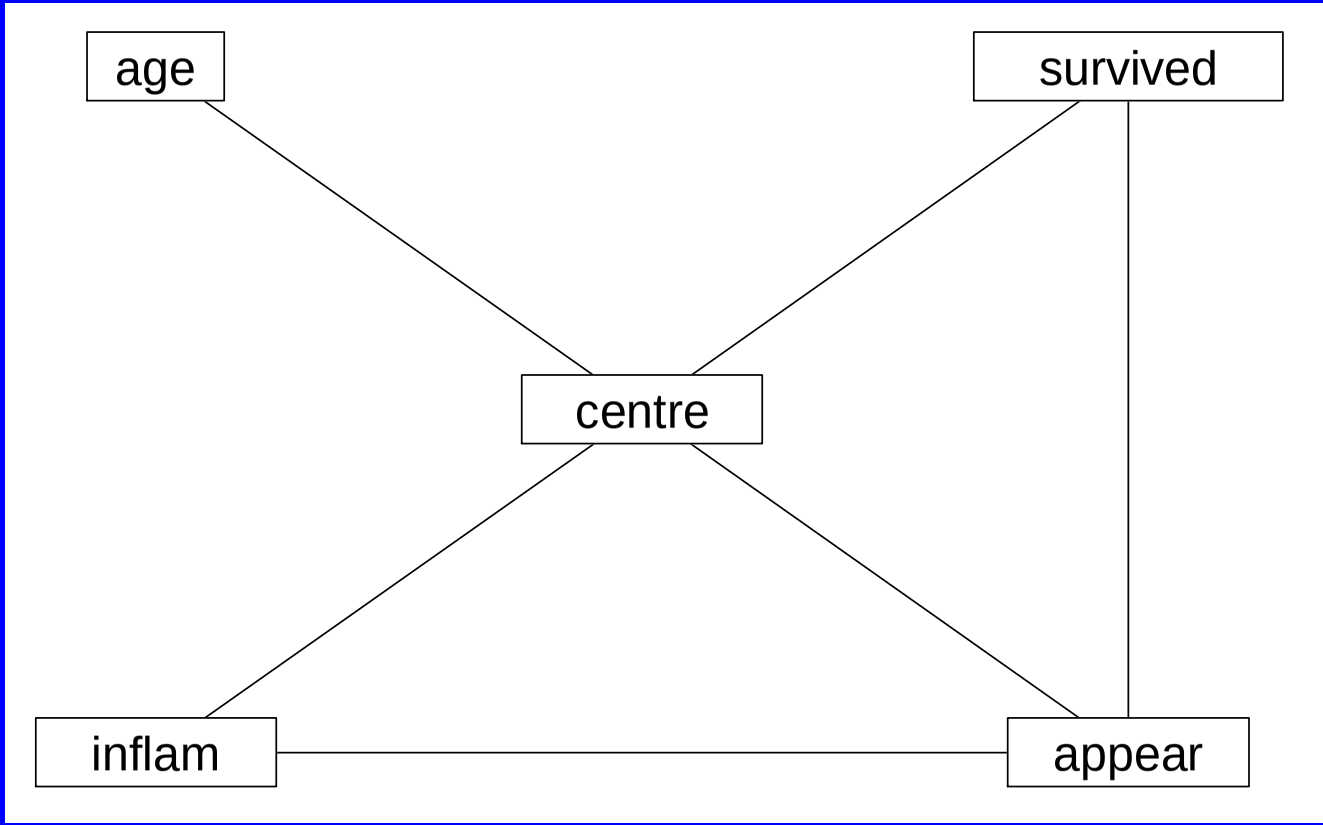

Figure 3.2

3.2.4 Logistic Regression

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.

The output would simply be given as number of successes, either 0 or 1, for each of the 764 patients involved in the study.

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

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

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.

Table 3.15

	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

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.

Table 3.16

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

Using the estimates given in Table 3.16, the fitted model is

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

3.2.5 Analysis of Survival Data

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.

We will skip this topic.

3.3 Special Topics in Regression Modelling

- Multivariate Analysis of Variance
- Repeated Measures Data
- Random Effects Models

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.

3.3.1 Multivariate Analysis of Variance

Model

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

where the $\boldsymbol{\varepsilon}_j$'s are independently and identically distributed as $\mathcal{N}^c(0, \boldsymbol{\Sigma})$ 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 β'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

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

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 model. This matrix is defined to be

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

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'.

3.3.3 Random Effects Models

Overdispersion

In a logistic regression we might replace (3.22) with

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

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.

Hierarchical models

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) with

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

where b_I, b_{II}, b_{III} and b_{IV} are independently and identically distributed as $\mathcal{N}(0, \sigma_b^2)$.

3.4 Classical Multivariate Analysis

- Principal Components Analysis
- Correspondence Analysis
- Multidimensional Scaling
- Cluster Analysis and Mixture Decomposition
- Latent Variable and Covariance Structure Models

3.4.1 Principal Components Analysis

Principal components analysis is a way of transforming a set of n-dimensional vector observations, $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$, into another set of n-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).

This will allow dimensionality reduction, so that we can do things like:

- 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 n dimensions.

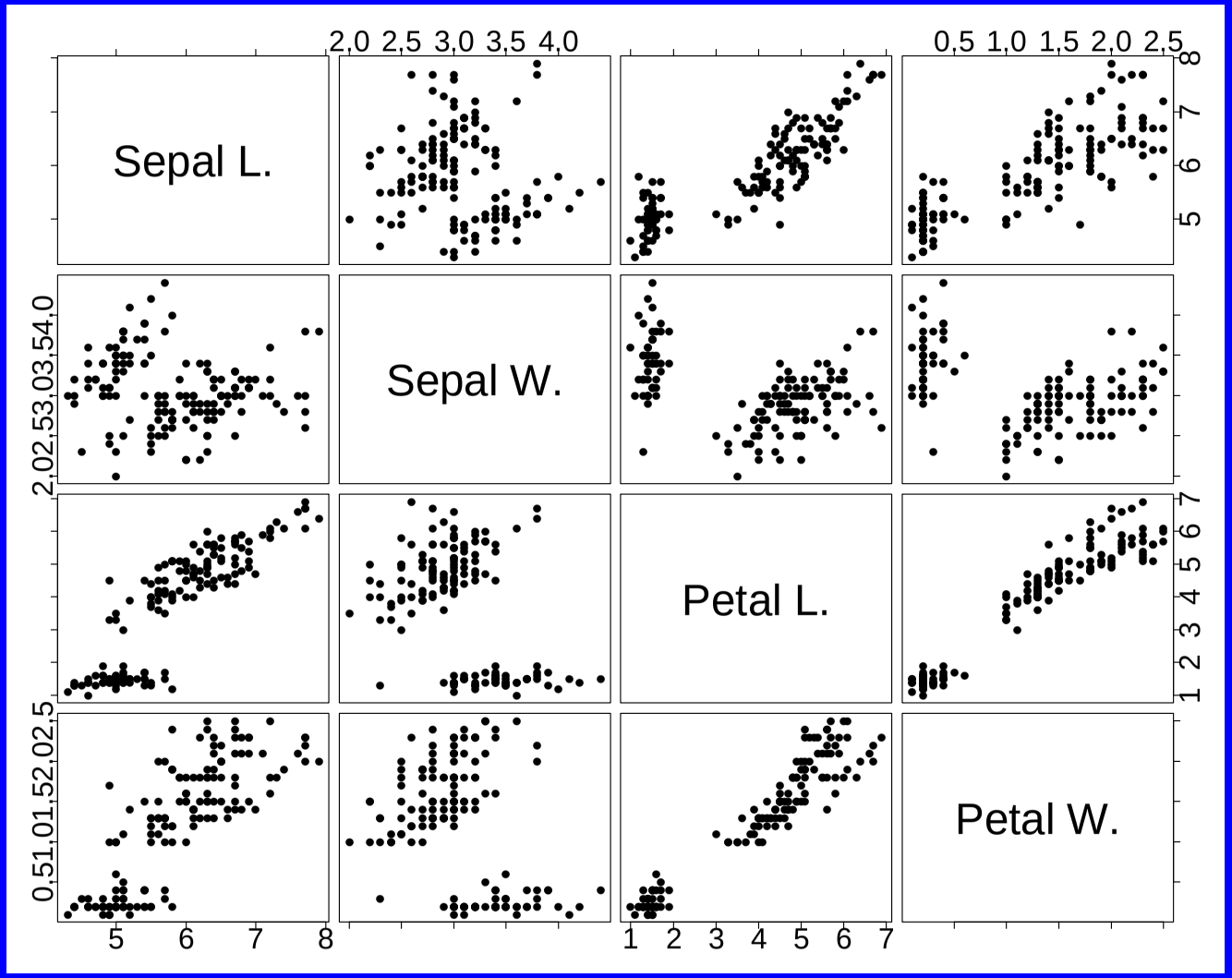

Figure 3.3
Fisher's Iris Data (collected by Anderson)

The main idea behind principal components analysis is that high information corresponds to high variance.

So, if we wanted to reduce the x's to a single dimension we would transform x to

$$y = a^\top x \ ,$$

choosing a so that y has the largest variance possible.

It turns out that a should be the eigenvector corresponding to the largest eigenvalue of the variance (covariance) matrix of 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 n dimensions.

Estimate of Σ is

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

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

- The eigenvalues of S are

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

- The eigenvectors of S corresponding to $\lambda_1, \lambda_2, \ldots, \lambda_n$ are $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$, respectively.

  The vectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ are called the *principal axes*. ($\boldsymbol{e}_1$ is the first principal axis, etc.)

- The $(n \times n)$ 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,

$$\underset{(n\times 1)}{\boldsymbol{y}} = \underbrace{\begin{bmatrix} \boldsymbol{e}_1^\top \\ \boldsymbol{e}_2^\top \\ \vdots \\ \boldsymbol{e}_n^\top \end{bmatrix}}_{(n\times n)} \underset{(n\times 1)}{\boldsymbol{x}} = E^\top \boldsymbol{x}$$

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.

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

$$\underset{(n\times n)}{\begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n \end{bmatrix}} .$$

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,

$$\sum_{i=1}^{n} \lambda_i = \sum_{i=1}^{n} s_i^2 \ ,$$

where s_i^2 is the sample variance of x_i.

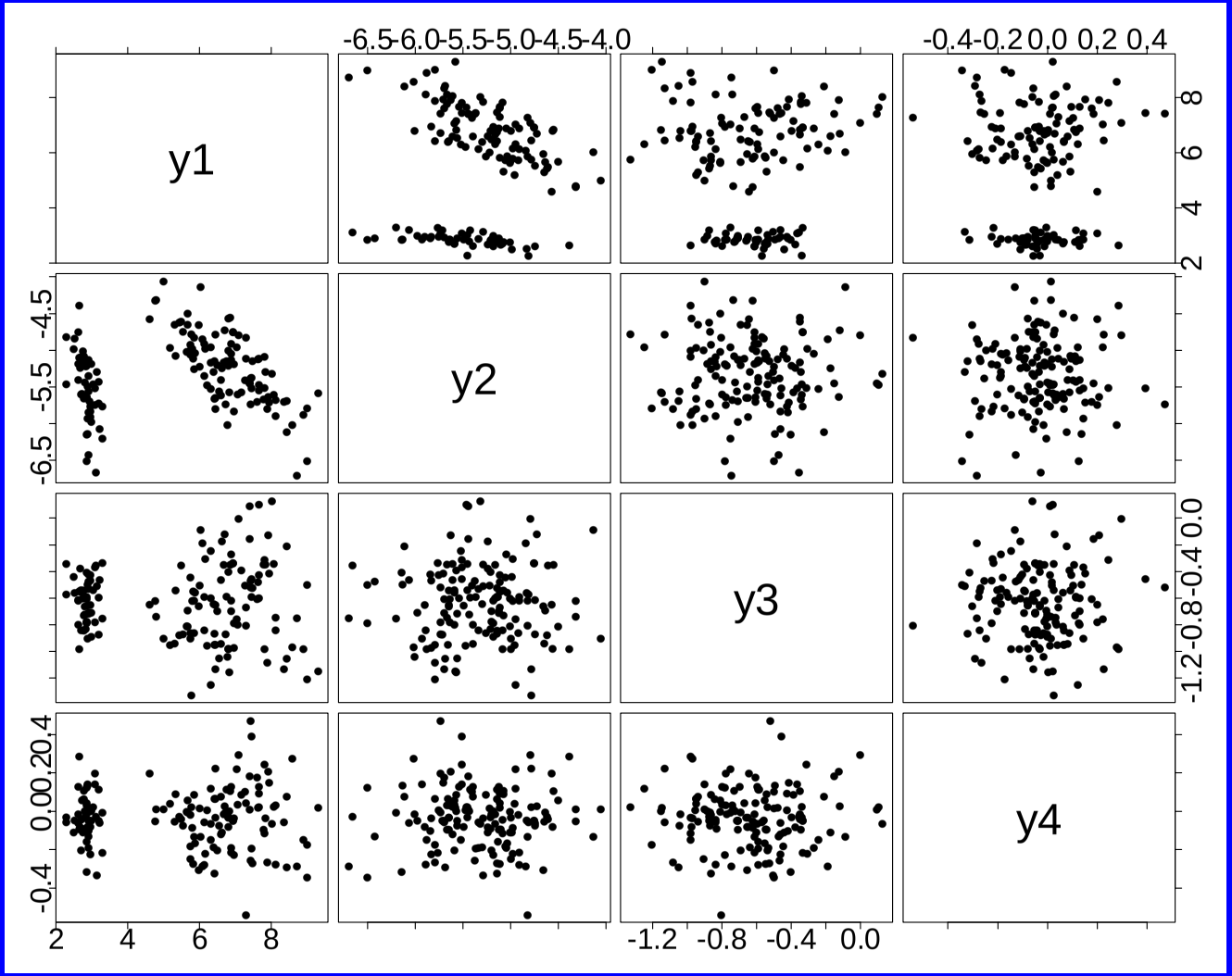

Figure 3.4
Principal component score for *Fisher's Iris Data*. Compare with Figure 3.3

Effective Dimensionality

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

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

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

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

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 approximately

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

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.

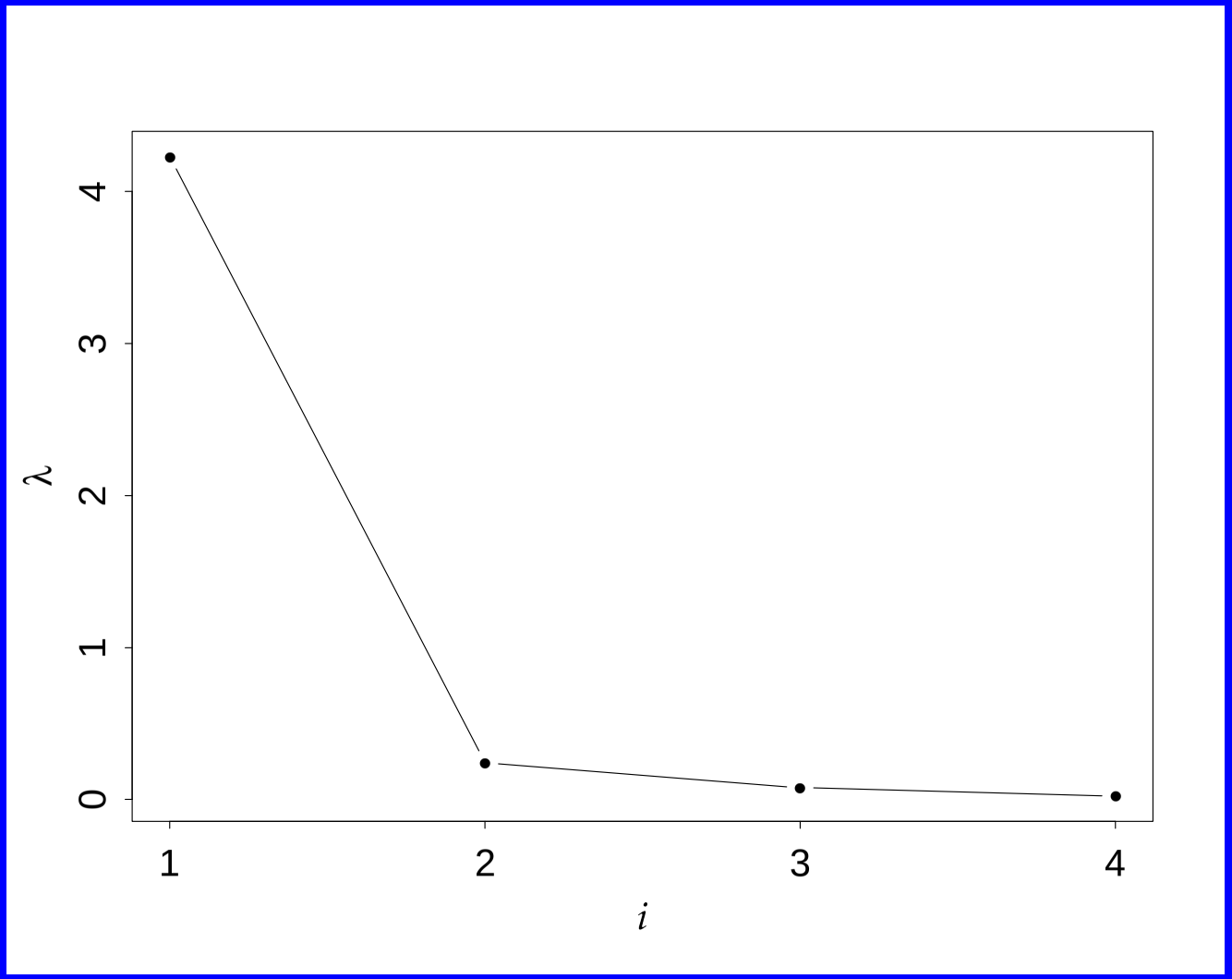

Figure 3.5
We look for the elbow; in this case we only need the first component.

Normalising

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.

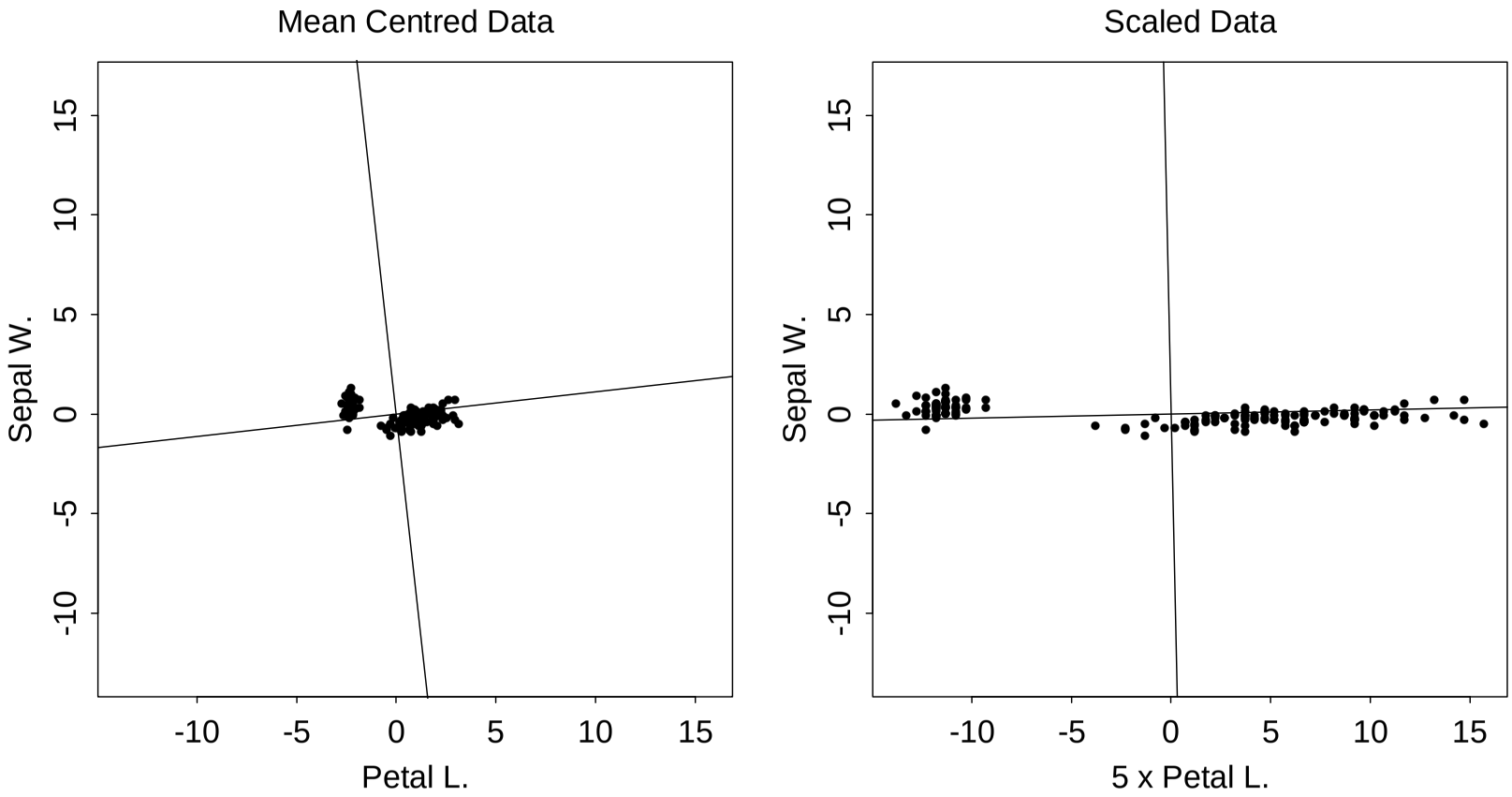

Figure 3.6 If we don’t normalise.

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.

See the book for an interpretation for *Fisher's Iris Data*.

3.4.2 Correspondence Analysis

Correspondence is a way to represent the structure within *incidence matrices*. Incidence matrices are also called *two-way contingency tables*.

An example of a (5×4) incidence matrix, with marginal totals is shown in Table 3.17.

Table 3.17

	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

Two Stages

- Transform the values in a way that relates to a test for association between rows and columns (chi-squared test).

- Use a dimensionality reduction method to allow us to draw a picture of the relationships between rows and columns in 2-D.

Details are like principal components analysis mathematically; see the book.

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.

Examples of dissimilarities could be:

- the price of an airline ticket between pairs of cities;
- 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.

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.

The derivation of the method used to obtain the configuration is given in the book.

The results of applying classical scaling to British road distances are shown in Figure 3.7. These road distances correspond to the routes recommended by the *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 3.7 is that the towns and cities have lined up in positions related to the motorway network.

- 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.

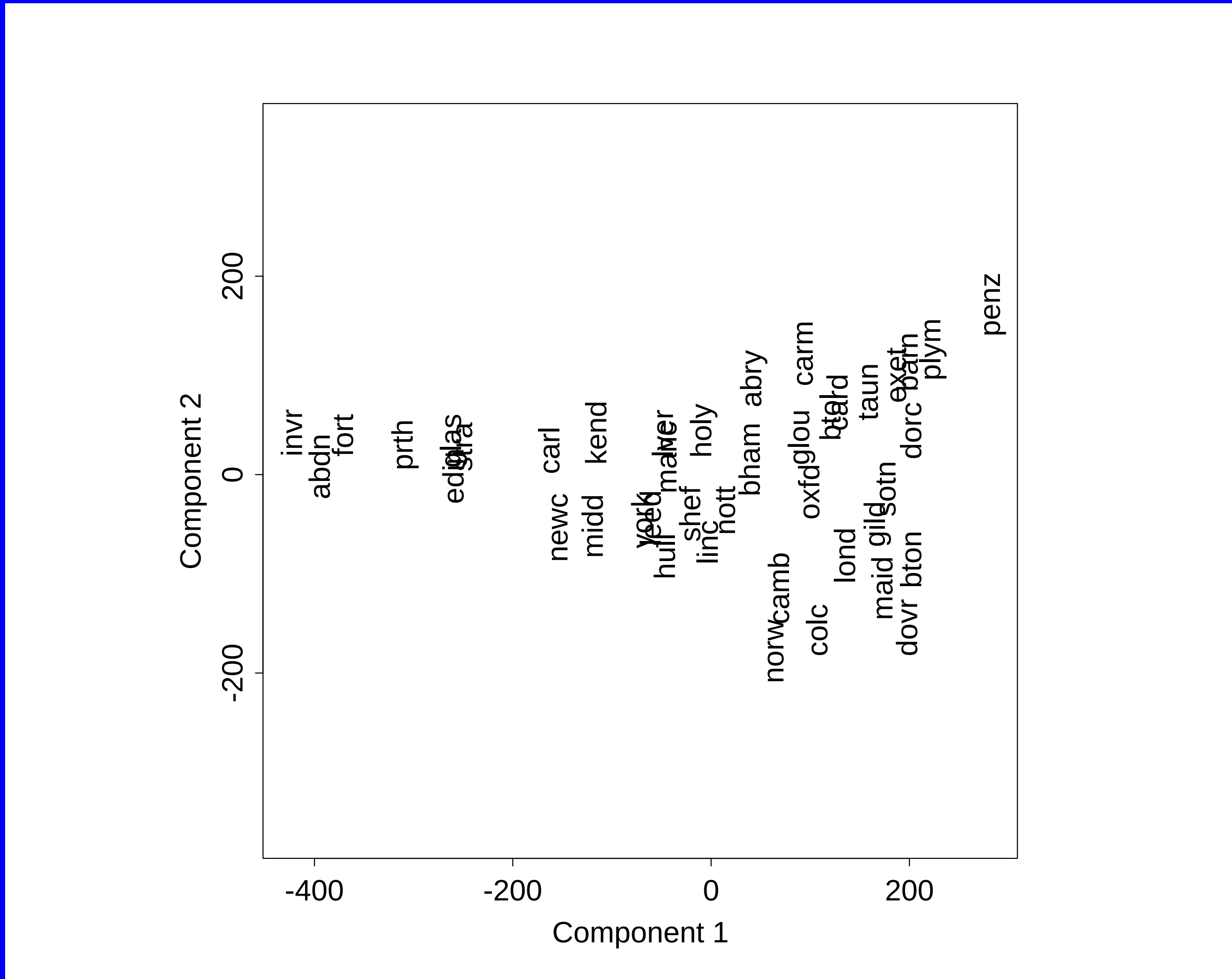
Component 1
Component 2
-400
-200
0
200
invr
abdn
fort
prth
edin
glas
stra
carl
newc
midd
kend
york
leed
hull
manc
lver
shef
linc
holy
nott
bham
abry
norw
camb
colc
oxfd
glou
carm
btel
card
lond
taun
maid
gild
sotn
exet
bath
dovr
bton
dorc
plym
penz

Figure 3.7

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

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

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*.

- *hierarchical*
  - *agglomerative*
  - *divisive*
- *non-hierarchical*

- **Minimum distance** or *single-link*
- **Maximum distance** or *complete-link*
- **Average distance**
- **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.

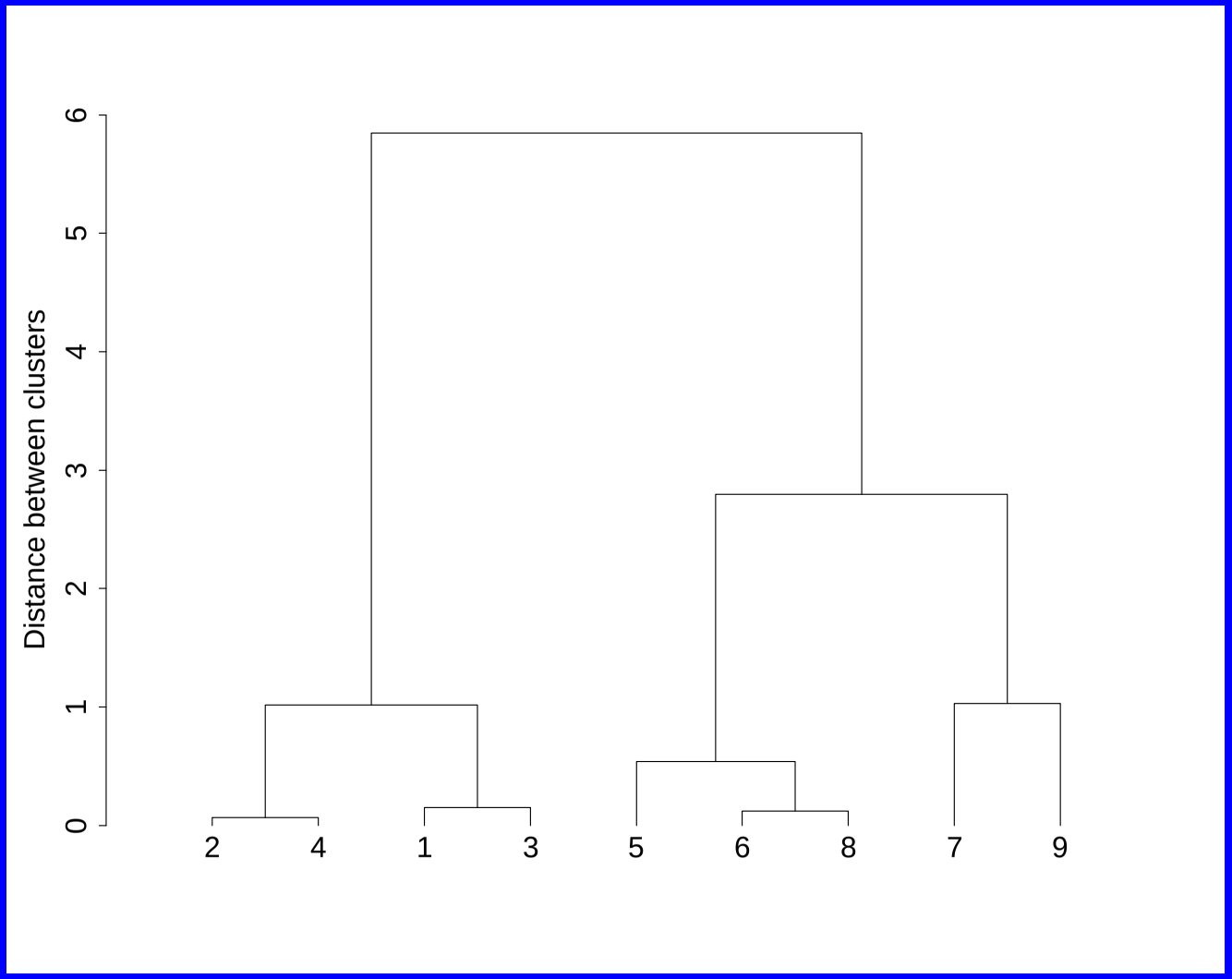

Figure 3.8
Usual way to present results of hierarchical clustering.

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¿

One of the best known non-hierarchical clustering methods is the *k-means* method.

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

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

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.

Details in one of Hand's books.

3.4.5 Latent Variable and Covariance Structure Models

I have never used the techniques in this section, so I do not conssider myself expert enough to give a presentation on them.

Not enough time to cover everything.

3.5 Summary

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.