HSS8005

Multivariate Statistics Stream

Semester 2, 2019–20

Dr. Lee Fawcett
School of Mathematics, Statistics & Physics

Multivariate Data Analysis, often abbreviated to MVA, provides the foundations of some of the buzz phrases being used in modern data applications. Some of the methods we will study over the next two weeks are at the forefront of methods in *Data Science*, *Big Data Analytics*, *Machine Learning* and even *Artificial Intelligence*. In this stream of HSS8005, we will look at what MVA is all about and why it is an essential part of the modern data analyst's toolkit.

As you might imagine, the mathematics, statistics and computing underpinning some of these methods is extremely complex. However, most methods are now available to use in standard software packages. As such, they are being routinely used by researchers in all disciplines, including those in the Humanities and Social Sciences.

We will aim to take the complexity out of the methodology and focus on analysing computer output to investigate hidden structure in multivariate datasets. We will also extend upon what has already been covered in this module to look at the role of MVA in model-building and prediction.

This stream is organised into three chapters:

- In Chapter 1 we will consider the motivation behind multivariate methods for analysis and how such methods might differ to (and why they might be needed over) classical approaches
- In Chapter 2 we consider methods for teasing out hidden structure in large datasets. In particular, we will focus on *Principal Components Analysis* and *Cluster Analysis*
- In Chapter 3 we return to something you were introduced to in the first part of the module: Regression analysis. Specifically, we examine *Multiple Regression* as a way of modelling multivariate datasets

Delivery of material will be organised around a variety of learning environments.

The course will begin with three hour-long lectures covering material in each of chapters 1, 2 and 3 (roughly one hour per chapter). These lectures will then be followed by an hour-long computer practical, and an hour-long seminar, for material in each of chapters 2 and 3. We will then finish this stream with an hour-long concluding lecture.

Date	Time	Place	Activity	Topics
Mon 16 Mar	12–1	Henry Daysh 6.16	Lecture	Chapter 1
Tue 17 Mar	4–6	Henry Daysh 6.16	Lecture	Chapters 2 & 3
Wed 18 Mar	10–11	Barbara Strang Chart	Practical	Chapter 2
Thu 19 Mar	*10–11*	*Barbara Strang G.33*		
	or	*or*	Seminar	Chapter 2
	4–5	*Henry Daysh 6.19*		
Tue 24 Mar	2–3	Barbara Strang 2.40	Practical	Chapter 3
Wed 25 Mar	*10–11*	*Medical School Fell*		
	or	*or*	Seminar	Chapter 3
	4–5	*Henry Daysh 6.16*		
Thu 26 Mar	4–5	Henry Daysh 6.19	Lecture	Chapters 1–3

Contents

1 Introduction **1**

1.1 Background 1

1.2 Why 'multivariate'? 2

1.3 Wanted and unwanted variability 2

1.4 Applications of MVA 3

1.4.1 London 2012 heptathlon data 4

1.4.2 Wine sales data 4

1.4.3 Microarray dataset 4

1.5 First steps 4

1.5.1 A look at the heptathlon dataset 4

1.5.2 An example of a hidden outlier 7

1.5.3 How to proceed? 8

2 Exploratory data analysis **9**

2.1 Simple methods for summarising multivariate datasets 9

2.1.1 Scatterplot matrices 9

2.1.2 Correlation matrices 12

2.1.3 Three dimensional (3-D) scatterplots 14

2.1.4 Other methods 14

2.2 Principal Components Analysis 15

2.2.1 PCA of the heptathlon dataset 15

2.2.2 PCA of the microarray dataset 21

3 Modelling **23**

3.1 Introduction 23

3.2 Simple exploration 23

3.3 Building a linear model 26
3.3.1 Fitting the full model 26
3.3.2 Backwards elimination 28
3.3.3 Using your model 29
3.3.4 Checking assumptions 30
3.4 Further topics in regression 31
3.4.1 Non-linear regression 31
3.4.2 Analysing a binary response: logistic regression 33
3.5 Bringing it all together 35

4 Practical work **37**
4.1 Exploratory multivariate data analysis in SPSS 38
4.2 Modelling multivariate data in SPSS 43

Chapter 1

Introduction

1.1 Background

With today's data collection systems and databases, a wealth of information is available to analysts at most organisations. For example, *Amazon* is a world leader in collecting, storing, processing and analysing personal information from you and every other customer as a means of determining how customers are spending their money (see the recent BBC documentary: *Amazon: What They Know About Us*, first shown on 17th February 2020). The company uses *predictive analytics* for targeted marketing to increase customer satisfaction and build company loyalty. Although 'Big Data' has helped *Amazon* evolve into a giant among online retailers, what the company knows about you may feel a bit like stalking!

Of course, collection of 'Big Data' is not just the hobby of analysts at large retail companies. In the race to secure victory in the 2016 US general election, both parties used 'Big Data Analytics', including methods from multivariate data analysis, to micro-target voters. Information was so precise that many tools allowed campaign managers to target individual voters with powerful and highly customized messages, often through social media. Of course, the same was true of all parties in the recent UK general election.

Researchers now have a wealth of data at their fingertips, often more than they know what to do with. In my case, I work with road safety engineers in city planning teams to analyse traffic flows on roads. Some local authorities that we work with – such as the *Camara de Municipal de Lisboa* in Lisbon, Portugal – have software systems supported by third parties such as *TomTom*, giving them access to huge amounts of data through personal in-car satellite navigation systems. This allows us to predict traffic flow on a very fine scale – for example, in a particular residential zone during Lisbon's Friday evening rush hour – enabling systems to be put in place to deal with congestion, or perhaps road traffic accidents, proactively. This is all possible through techniques in multivariate data analysis, using *TomTom*'s data on driver charateristics and habits.

Once we have our data, the first step is usually to do some sort of *Exploratory Data Analysis*. In other words, use graphical and numerical tools to explore our datasets to see what they look like, and if they're hiding any patterns. We will look at methods for

exploring large multivariate datasets in Chapter 2.

We might then want to know how to *use* our data, and perhaps exploit the patterns exposed in our exploratory analysis. In Chapter 3, we will consider multiple regression techniques for *modelling* multivariate datasets, and using these models to predict new and future events.

1.2 Why 'multivariate'?

If you could solve all problems by taking a single measurement on a system, the world would be a much simpler place. However, complex systems often require multiple measurements, on many variables, to understand them.

One simple example is predicting the weather. To adequately predict the temperature for a particular day, meteorologists require information on pressure, wind conditions, humidity, season and so on. These data enable them to build a model, which in turn allows them to estimate, or predict, what the future may hold.

Multivariate Data Analysis, often abbreviated to MVA, is the investigation of many variables, simultaneously, in order to understand the relationships that might exist between them. An MVA might be as simple as looking at the relationship between two variables through a scatterplot, right up to analysing millions of variables together.

An MVA can be highly graphical, although the usual output graphs might not be those you are used to seeing. For example, you might be interested in the association between height and weight for a cohort of students; this could easily be explored through a 2-D scatterplot (and would likely reveal a positive correlation). Throw a third variable into the mix – perhaps shoe size – and we're into the realms of 3-D scatterplots (not beyond the capability of most modern statistical software programmes). Graphically displaying the results for more than three variables requires some ingenuity, since we live in a 3-D world! Luckily, techniques have been developed to help us do just that.

1.3 Wanted and unwanted variability

An MVA is a study of variability and its sources. Variability may be defined as 'Wanted' or 'Unwanted'.

Wanted variability might include changing the settings on a piece of machinery to assess the impact it has on a process. In other words, the experimenter is directly responsible for this level of variability, and we want to see the effects it might have on something.

Unwanted variability results from our inability to control something, and it is usually random in nature. For example, you might be assessing the performance of a thermostat in an oven: in repeated experiments, it is very unlikely that the oven will take exactly the same time to reach the desired temperature.

A statistical *model* is a summary of your best knowledge of a system at the time of investigation. A good model can be used to forecast future events with a high level of confidence, and can be validated against new data coming online. A multivariate model is able to show the influence of both types of variability ('Wanted' and 'Unwanted') on a system, perhaps so that improvements can be made.

1.4 Applications of MVA

An MVA can be applied to any dataset for which you have measurements on more than one variable. MVA has recently found widespread use in the following areas:

- Food and drinks industry: particularly in new product development
- Business intelligence: predicting changes in dynamic markets or better product placement
- Petrochemical industry: early fault detection, and gasoline blending and optimisation
- Agricultural analysis: for example, real-time analysis of protein and moisture in wheat, barley and other crops
- Transport systems: crash forecasting, for proactive implementation of road safety interventions
- Social sciences: investigations into the influence of multiple stimulii on various patient groups, or how demographics relate to various social issues
- Consumer science: isolating demographics where product lines are better-suited, or formulating better products to meet consumer needs
- Genetics: where very large datasets are generated and the search for new proteins/genomes is critical, for example to guide best practice in cancer treatments

We will demonstrate the methods in this stream through direct application to some very interesting datasets. Well, *I* think they're interesting! In the assessment for the stream, you will get the opportunity to critique your own intersting multivariate analyses (or you can choose from a range of applications and datasets provided). Some of the following datasets are small, at least by 'big data' standards, so we can more easily get a handle on the nuts-and-bolts of what's going on in our MVA. However, one is extremely large; as we discuss, without performing an appropriate MVA it would be very difficult to detect any patterns or structure in this dataset.

1.4.1 London 2012 heptathlon data

The heptathlon at the 2012 Olympic Games in London involved athletes competing in 7 track and field events. These events can be divided into three types: (i) running events (200m, 800m and 100m hurdles); (ii) jumping events (high jump and long jump); (iii) throwing events (shot put and javelin throw). We consider the 29 athletes who completed all events and were not disqualified.

1.4.2 Wine sales data

This dataset contains information on the selling price of individual bottles of wine, sold by the UK wine subscription club *Naked Wines*. For each bottle of wine in the dataset we also have information on several other variables, including the age of the wine; the amount of rainfall observed during the wine's grape-growing season; the average day-time temperature during the wine's grape-growing season; and whether or not the wine was produced in New Zealand.

1.4.3 Microarray dataset

Gene expression is the process by which genetic instructions are used to synthesise gene products in a cell. Usually these prducts are proteins, which then go on to perform essential functions, for example, as enzymes or hormones. Expression is usually measured in terms of the amount of gene product which is being synthesised. The microarray data is a large dataset, collected by the National Cancer Institute (NCI), which contains measurements on the expression levels of 6830 genes from 64 samples taken from malignant tumours in different patients.

1.5 First steps

1.5.1 A look at the heptathlon dataset

Let's take a look at the London 2012 heptathlon dataset (Table 1.1 overleaf). This dataset is by no means 'big'. Indeed, we can fit the entire raw dataset on a single page! But it's definitely multivariate: we have information for all athletes on more than one variable. If we only had information on their times in the 100 metres hurdles – and nothing else – then our dataset would *not* be multivariate; it would be *univariate*. Or if we were only interested in the athletes' performances in the 100 metres hurdles – and nothing else, so we disregarded data on all the other variables – then we would not perform a multivariate analysis, our analysis would be *univariate*.

Perhaps we only had data on the jumping events. Then our dataset would be *bivariate*, consisting of athletes' performances in both the high jump and the long jump. A bivari-

Name	Nationality	100M hurdles	High jump	Shot put	200M	Long jump	Javelin	800M
Jessica Ennis	(GBR)	12.54	1.86	14.28	22.83	6.48	47.49	128.65
Lilli Schwarzkopf	(GER)	13.26	1.83	14.77	24.77	6.30	51.73	130.50
Austra Skujyte	(LTU)	14.00	1.92	17.31	25.43	6.25	51.13	140.59
Antoinette Nana Djimou Ida	(FRA)	12.96	1.80	14.26	24.72	6.13	55.87	133.62
Jessica Zelinka	(CAN)	12.65	1.68	14.81	23.32	5.91	45.75	129.15
Kristina Savitskaya	(RUS)	13.37	1.83	14.77	24.46	6.21	43.70	132.27
Laura Ikauniece	(LAT)	13.71	1.83	12.64	24.16	6.13	51.27	132.13
Hanna Melnychenko	(UKR)	13.32	1.80	12.96	24.09	6.40	43.86	132.90
Brianne Theisen	(CAN)	13.30	1.83	12.89	24.35	6.01	46.47	129.27
Dafne Schippers	(NED)	13.48	1.80	13.67	22.83	6.28	36.63	135.52
Nadine Broersen	(NED)	13.64	1.86	13.57	25.13	5.94	51.98	136.98
Jessica Samuelsson	(SWE)	13.58	1.77	14.18	24.25	6.18	42.02	131.31
Katarina Johnson-Thompson	(GBR)	13.48	1.89	11.32	23.73	6.19	38.37	130.76
Sharon Day	(USA)	13.57	1.77	14.28	24.36	5.85	43.90	131.31
Yana Maksimava	(BLR)	13.97	1.89	14.09	25.43	5.99	42.33	133.37
Eliska Klucinova	(CZE)	14.01	1.80	12.93	25.00	6.13	45.65	136.08
Ellen Sprunger	(SUI)	13.35	1.71	12.62	23.59	5.88	45.63	137.54
Olga Kurban	(RUS)	13.46	1.80	13.71	23.88	5.83	40.36	139.82
Marisa De Aniceto	(FRA)	13.74	1.71	13.09	25.26	5.76	51.98	136.20
Gyorgyi Farkas	(HUN)	14.33	1.80	13.55	25.72	6.07	46.52	137.83
Grit Sadeiko	(EST)	13.50	1.74	12.43	24.25	6.11	44.12	143.01
Sofia Ifadidou	(GRE)	13.82	1.68	12.96	25.91	5.81	56.96	142.03
Ivona Dadic	(AUT)	14.58	1.80	12.19	24.29	6.00	41.82	135.90
Sarah Cowley	(NZL)	13.95	1.80	12.37	25.60	6.00	41.90	139.01
Louise Hazel	(GBR)	13.48	1.59	12.81	24.48	5.77	47.38	138.78
Ida Marcussen	(NOR)	14.08	1.68	14.26	25.15	5.82	42.26	133.62
Chantae McMillan	(USA)	13.49	1.68	14.92	25.25	5.37	49.78	160.55
Julia Machtig	(GER)	14.54	1.68	14.99	25.38	4.06	44.40	140.34
Irina Karpova	(KAZ)	14.21	1.68	11.68	25.42	5.70	35.75	148.93

Table 1.1: London 2012 Heptathlon dataset.

ate dataset is the simplest form of multivariate dataset, where we have information on multiple (but only two!) variables. This could then lead to a bivariate analysis.

The first stage of this bivariate analysis might consist of simple *exploration*, perhaps via a scatterplot and even the calculation of the correlation coefficient. The scatterplot of long jump performances against high jump performances for all 29 athletes can be seen in the top left-hand plot in Figure 1.1, as produced by SPSS. There seems to be a positive association between distances jumped in both events (as we might expect!), and this is supported by a significant correlation coefficient of 0.55 (p-value 0.002). However, from this exploratory analysis we also see that there seems to be an outlier, with one athlete jumping an unsually short distance in the long jump given her cleared height in the high jump.

Other standard features of an initial *exploration* of our bivariate data might include summary statistics for both variables – perhaps means and standard deviations, or medians and inter-quartile ranges. All of this helps to build a picture of what's going on in our dataset.

The next stage of our bivariate analysis might include *modelling* our bivariate data. The exploratory stage of the analysis revealed a significant association between performances in the two jumping events, so we might now attempt to *model* this association. As you know, we might do this through a simple linear regression model, providing the equation of the 'best' line of best fit in the original scatterplot. Doing this in SPSS, we get a

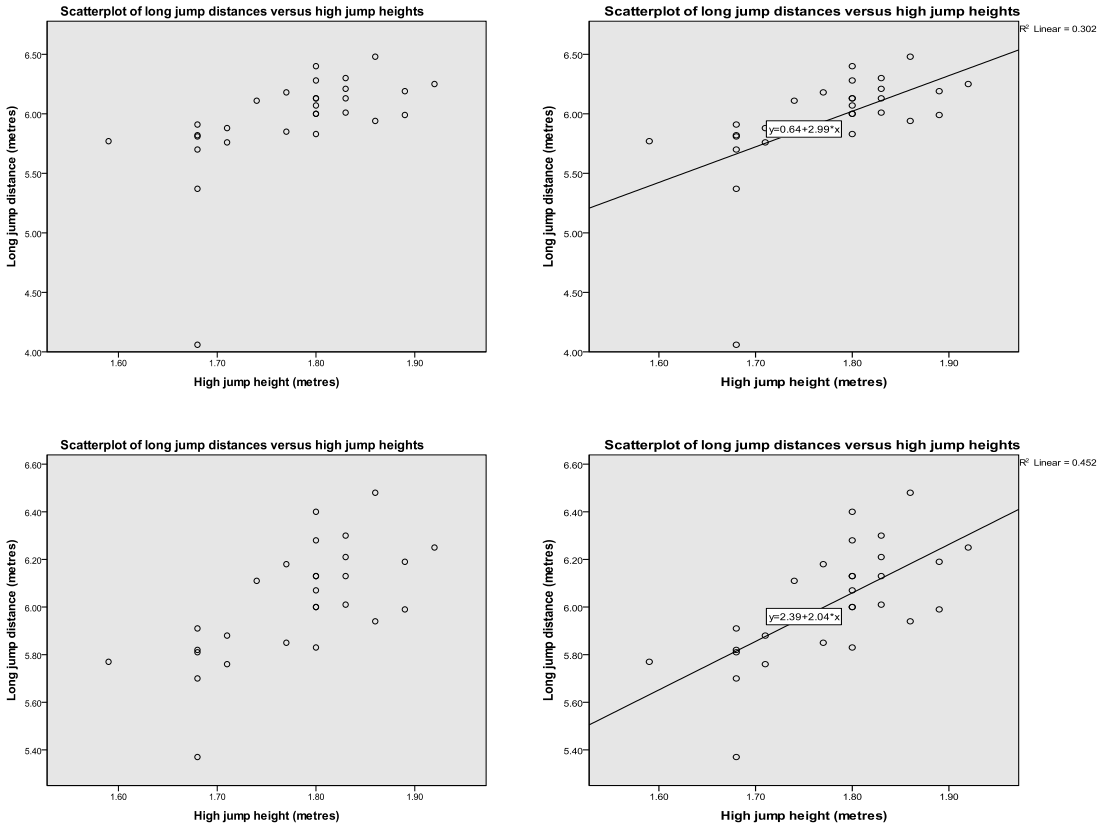

Figure 1.1: Top row: scatterplot of long jump distance versus high jump height (left), and scatterplot with regression line superimposed (right). Bottom row: scatterplot and scatterplot including regression line, after removing the apparent outlier (Julia Machtig).

model of the form

$$\text{Long jump distance} = 0.64 + 2.99 \times \text{High jump height} + \epsilon,$$

where (as always!) ϵ represents our 'noise', or error, around the fitted regression line, and $\epsilon \sim N(0, \sigma^2)$. The p-values for the fitted slope term can be obtained in SPSS as 0.002. This regression line can be seen in the top right-hand plot in Figure 1.1, superimposed onto the original scatterplot. From this modelling exercise, we can see that:

- High jump height is a (very!) significant predictor of long jump distance, since the p-value for the slope is 0.2%
- For every one metre increase in high jump height, there is approximately a three metre increase in long jump distance
- Only around 30% of the variation in long jump distances is explained by performances in the high jump, as the R^2 value is 0.302

Remember the apparent outlier detected in our earlier exploratory analysis? This relates to the performance of Julia Machtig of Germany. If we remove her from the analysis entirely, we get the scatterplots shown in the bottom row of Figure 1.1. SPSS now

computes the correlation coefficient to be 0.672 (p-value 0.000), indicating a stronger positive association than before. The updated regression equation becomes

$$\text{Long jump distance} = 2.39 + 2.04 \times \text{High jump height} + \epsilon, \tag{1.1}$$

and the fitted slope term now has a p-value of 0.005. Now we can see that, after removing this outlier:

- High jump height remains a very significant predictor of long jump distance, since the p-value of the slope is 0.5%
- For every one metre increase in high jump height, there is approximately a two metre increase in long jump distance
- Now around 45% of the variation in long jump distances is explained by performances in the high jump, as the R^2 value is 0.452. This is conisderably better than before – likely due to the removal of the outlier

Suppose an athlete obtained a height of 1.75 metres in the high jump but was then injured and unable to take part in the long jump. We could use our fitted regression model in Equation (1.1) to estimate what her long jump distance would have been:

$$\text{Long jump distance} = 2.39 + 2.04 \times 1.75 = 5.96 \text{ metres.}$$

So, that's an example of a bivariate analysis! What about all the other information available to us, relating to the other five events in the heptathlon? How could we include all of this data in the analysis too?

1.5.2 An example of a hidden outlier

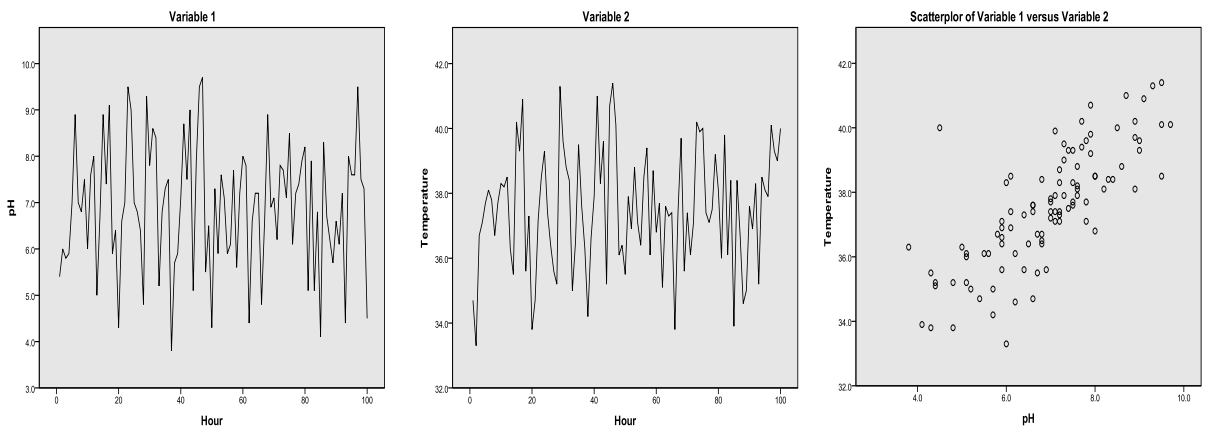

Figure 1.2: Simple univariate and multivariate control charts.

Suppose that you're the production supervisor at a company. Your process has been running smoothly until, suddenly, alarms are sounding, and you need to decide what to do!

Your production relies on observations on two main variables – temperature and acidity – both of which are recorded once per hour. When you look back at the univariate plots of temperature and acidity, all seems fine – the processes seem 'in control', and there are no obvious anomalies (see the first two plots in Figure 1.2). But the alarms were sounding for a reason!

A multivariate analysis can help to uncover problems in our data, problems that univariate analyses might not identify. This becomes apparent when we investigate the relationship between temperature and acidity through a simple scatterplot (third plot in Figure 1.2). Notice now, when we plot temperature against acidity, we see an apparent outlier – perhaps the alarms were sounding because, *given the relatively low pH reading*, the corresponding temperature reading was abnormally high! Dependencies between variables are lost when we explore variables individually and neglect to explore the relationships between them. Now imagine you had data on ten variables – there could be complex bivariate/trivariate/multivariate dependencies hidden in the data that we would need to somehow explore!

1.5.3 How to proceed?

The first step of any multivariate analysis should be *exploration*. Are there any dependencies between our variables? Are there any outliers, or groupings in our data, that cannot be seen by looking at each of the variables individually? As we will see in Chapter 2, there are some simple techniques at our disposal through common Statistics packages (e.g. SPSS) to fully explore our data. Of course, if there are any patterns in our data, we might then want to exploit these patterns. That will lead us on to what is often the second step of any multivariate analysis: *modelling*. We will consider some simple techniques for modelling multivariate data in Chapter 3.

Chapter 2

Exploratory data analysis

In this Chapter, we will consider simple methods to summarise multivariate datasets, before looking at the most commonly-used tool for exploring multivariate datasets: *Principal Components Analysis*.

2.1 Simple methods for summarising multivariate datasets

Before embarking on any rigorous analysis of a multivariate dataset (or *any* dataset, for that matter!), it is always a good idea to begin by plotting it to get a feel for its structure. When we have univariate datasets, we can simply plot a histogram (for continuous data), or a bar chart/pie chart (for categorical data). For bivariate datasets, we might also consider such plots to investigate the structure of each variable individually, but to assess the extent of dependence between our two variables we could plot a simple scatterplot of one variable against the other (provided the variables are paired in some way). This is exactly what we did to investigate the relationship between long jump and high jump performances in the heptathlon dataset in Section 1.5.1 (see the plots in Figure 1.1).

However, what if we have a multivariate dataset with more than two dimensions? For example, how could perform an initial exploration of the full heptathlon dataset?

2.1.1 Scatterplot matrices

The plot in Figure 2.1 shows a scatterplot matrix for the heptathlon dataset, as produced in SPSS. You will see how to produce such a plot for yourself in the computer practical session. The scatterplot matrix in Figure 2.1:

- Is useful for investigating (pairwise) relationships between all seven variables
- Displays variable names on both axes, giving a row and a column for each variable
- Has blank plots down the leading diagonal of the matrix

- Has repeated plots above/below this leading diagonal
- Can be thought of as a 2-D projection of points in 7-D space

How do we read this scatterplot matrix, and what does it tell us? Suppose we want to know if there is any association between performances in the long jump and high jump. Select 'LJ' along one edge of the matrix, and 'HJ' along the other. Moving into the matrix from these edges, we can then locate the cell of the matrix that displays the scatterplot of long jump distances and high jump heights. This is exactly the same as the plot shown in Figure 1.1, and reveals a positive association between performances in these two events (with a potential outlier).

But now, we have scatterplots for all possible pairwise combinations of heptathlon events! For example, we can see that there is a positive association between times in the 200 metres and 800 metres, and a negative association between times in the 800 metres and distances in the long jump.

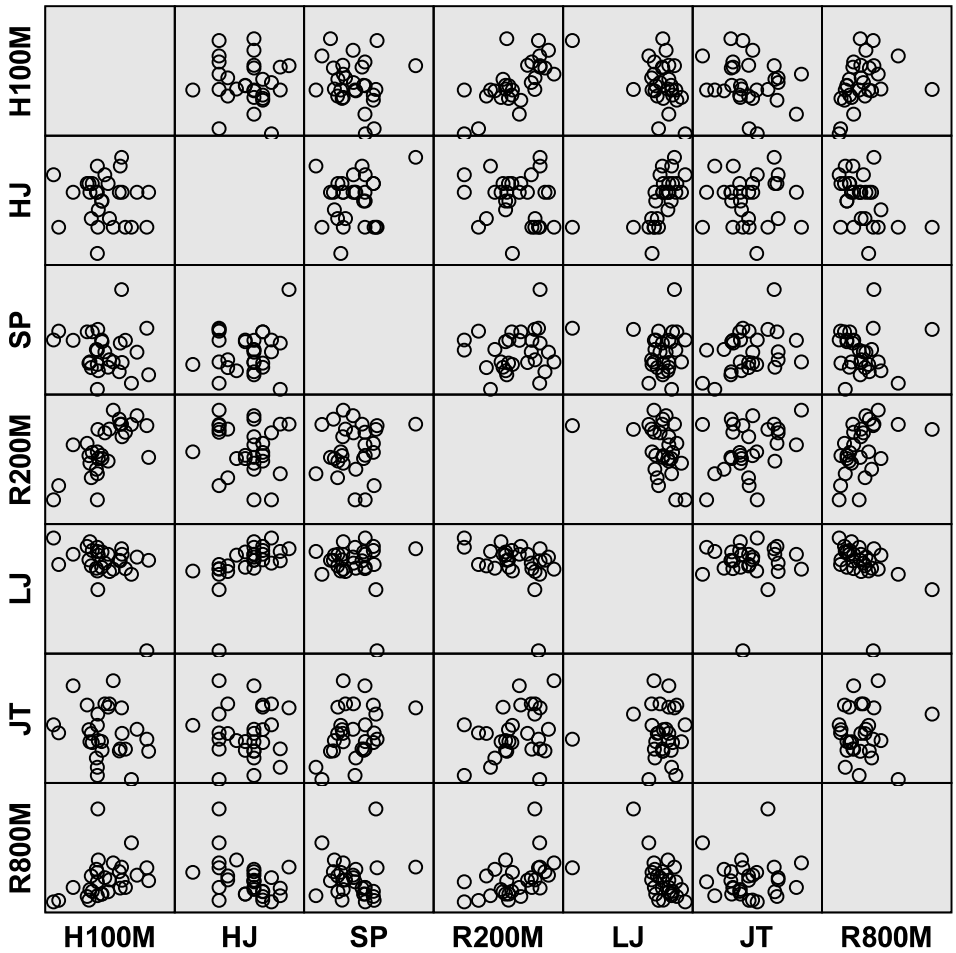

Figure 2.1: Scatterplot matrix for the heptathlon dataset.

Questions

- Why do you think there are positive associations between times in the 200 metres and 800 metres? Why *negative* associations between performances in the 800 metres and the long jump?
- What can you say about performances in the shot put and the javelin throw?
- What can you say about performances in shot put and the 200 metres?
- Why is the leading diagonal of cells empty?
- Why are plots repeated above/below the leading diagonal?

✎

2.1.2 Correlation matrices

We might want to quantify the associations displayed in a scatterplot matrix, such as that shown in Figure 2.1. We can do this in SPSS by producing a correlation matrix; you will see how to construct this for yourself in the computer practical session. The correlation matrix for five of the variables in the heptathlon dataset is shown in Figure 2.2.

To understand what this matrix is telling us, let's just focus on a single pair of variables – perhaps performances in the 200 metres and 800 metres. Remember, in the scatterplot matrix in Section 2.1.1 we identified a fairly strong, positive association between performances in these two events.

- The correlation coefficient is 0.461
- This correlation coefficient has been 'starred', indicating that it is statistically significant; in fact, the p-value is 1.2%, indicating significance at the 5% level (and nearly at the 1% level)
- This is based on the performances of all N = 29 athletes

Correlations

		HJ	SP	R200M	LJ	R800M
HJ	Pearson Correlation	1	.157	-.135	.550**	-.428*
	Sig. (2-tailed)		.415	.484	.002	.021
	N	29	29	29	29	29
SP	Pearson Correlation	.157	1	.108	-.117	-.002
	Sig. (2-tailed)	.415		.576	.545	.990
	N	29	29	29	29	29
R200M	Pearson Correlation	-.135	.108	1	-.378*	.461*
	Sig. (2-tailed)	.484	.576		.043	.012
	N	29	29	29	29	29
LJ	Pearson Correlation	.550**	-.117	-.378*	1	-.449*
	Sig. (2-tailed)	.002	.545	.043		.015
	N	29	29	29	29	29
R800M	Pearson Correlation	-.428*	-.002	.461*	-.449*	1
	Sig. (2-tailed)	.021	.990	.012	.015	
	N	29	29	29	29	29

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

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

Figure 2.2: Correlation matrix for five variables in the heptathlon dataset, as produced in SPSS.

Questions

- What does the correlation matrix tell you about performances in the long jump and the high jump?
- What about the dependence between performances in the 800 metres and the long jump?
- What about the shot put and the 200 metres?
- Why do all cells along the leading diagonal have correlation coefficients of exactly 1?

✎

2.1.3 Three dimensional (3-D) scatterplots

Both scatterplot matrices and correlation matrices give us some insight into the pairwise dependencies between variables in our multivariate dataset. We can attempt to view three-way dependencies through the production of a 3-D scatterplot in SPSS. Such a plot is shown in Figure 2.3, for performances in the 100 metres hurdles, the 200 metres and the long jump.

Shown in its original state (first plot in Figure 2.3), it is difficult to appreciate the value of such a plot. Indeed, these plots can be quite tricky to interpret. However, within SPSS you can interact with these plots: spinning them around reveals the pairwise dependencies that we can identify in a scatterplot matrix; within a 3-D scatterplot all pairwise dependencies are captured in a single graphic. The second plot in Figure 2.3 has been rotated slightly to reveal the dependence we identified earlier between performances in the 200 metres and the long jump (see the correlation matrix in Figure 2.2).

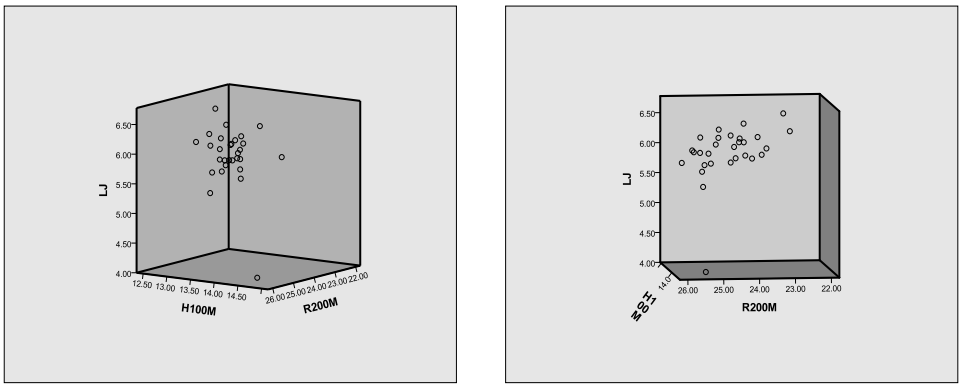

Figure 2.3: 3-D scatterplot for three variables in the heptathlon dataset, as produced in SPSS. Right: plot rotated slightly to focus attention on the dependence between the long jump and the 200 metres.

2.1.4 Other methods

There are a variety of other methods available for summarising multivariate datasets both graphically and numerically. These include:

- *Star symbol plots*: Each individual is represented by a star, these stars having the same number of rays as we have variables. The length of the rays are proportional to the measurements made on each variable.

- *Mahalanobis distance*: Put simply, this is the distance between two points in multivariate space. It is a numerical summary often used to detect outliers in multivariate space. We could do this graphically for our bivariate example in Section 1.5.2 (Figure 1.2), but trying to do so when we have more than three variables (and so producing a single scatterplot is no longer possible) is much more difficult. The Mahalanobis distance is very useful in such instances.

2.2 Principal Components Analysis

Principal Components Analysis (PCA) is the analysis of variability in a multivariate dataset. PCA is often referred to as the "workhorse of multivariate methods" as it is the most commonly used tool for exploring large, multivariate datasets. PCA provides some powerful graphical tools for exploratory analysis, and enables datasets with many variables to be displayed graphically in 2-D space. Thus, PCA is referred to as a *dimension reduction* technique.

The main aim is to find new variables which are linear combinations of the original variables, and which are uncorrelated with each other. The new variables are such that a few of them will explain most of the variability in the data (hopefully!), hence effectively reducing the number of variables. Unlike stepwise regression (see Chapter 3), PCA includes information on *all* variables through these new variables. Another advantage is that there are no assumptions made about the original variables; no model is assumed, and the method is completely general!

2.2.1 PCA of the heptathlon dataset

The easiest way to explain how PCA works is through direct application to a multivariate dataset. Again, we'll use the heptathlon dataset. Recall we have seven variables, which we will now label:

$$\begin{aligned}
X_1 &= \text{Time in 100 metre hurdles} \\
X_2 &= \text{Height in high jump} \\
X_3 &= \text{Distance in shot put} \\
X_4 &= \text{Time in 200 metres} \\
X_5 &= \text{Distance in long jump} \\
X_6 &= \text{Distance thrown in javelin} \\
X_7 &= \text{Time in 800 metres}
\end{aligned}$$

The aim of a PCA is to construct new variables:

$$\begin{aligned}
Y_1 &= a_{1,1}X_1 + a_{2,1}X_2 + a_{3,1}X_3 + a_{4,1}X_4 + a_{5,1}X_5 + a_{6,1}X_6 + a_{7,1}X_7 \\
Y_2 &= a_{1,2}X_1 + a_{2,2}X_2 + a_{3,2}X_3 + a_{4,2}X_4 + a_{5,2}X_5 + a_{6,2}X_6 + a_{7,2}X_7 \\
Y_3 &= a_{1,3}X_1 + a_{2,3}X_2 + a_{3,3}X_3 + a_{4,3}X_4 + a_{5,3}X_5 + a_{6,3}X_6 + a_{7,3}X_7 \\
Y_4 &= a_{1,4}X_1 + a_{2,4}X_2 + a_{3,4}X_3 + a_{4,4}X_4 + a_{5,4}X_5 + a_{6,4}X_6 + a_{7,4}X_7 \\
\vdots \quad & \quad \vdots
\end{aligned}$$

Notice that each of the Y's is a bit like a regression model, where they're all functions of all of our original X variables. In order to find the Y's, we need to obtain estimates of all the 'a's. We'll not worry about how to estimate the 'a's in this course, as SPSS will take care of the mathematics behind this - we will just focus on interpreting the output. But why would we want to do all of this anyway?

Y_1 is known as the *first principal component* in our analysis. Notice it contains information on all of our original variables X_1 through to X_7. Y_2 is the *second principal component*, and it too contains information on all of our original variables. Y_3 is the third principal component, and so and so forth. Each of our principal components is a function of all of our original X variables, and so each one will capture the overall variability of our multivariate dataset.

In fact, the mathematics underpinning the estimation of the 'a's is such that

$$\text{Var}(Y_1) > \text{Var}(Y_2) > \text{Var}(Y_3) > \cdots ,$$

and it's often the case that the first two principal components, considered together, capture most of the variability present in our original multivariate dataset.

Thus, our aim is to work with, perhaps, just Y_1 and Y_2, rather than $X_1, X_2, \ldots, X_7$, as both Y_1 and Y_2 will likely capture most of the variability in our multivariate dataset *and* then we've 'reduced the dimension' of our analysis from 7-D to just 2-D. And we can very easily explore bivariate datasets, as we discussed in Sections 1.5 and 2.1.

Note that in the jumping events, large numbers are indicative of good performances, but in the running events, small numbers are indicative of good performances. It is good practice, when performing a PCA, to re-scale variables so that the 'direction' is interpreted in the same way for all variables. Thus, from this point forwards, we now work with a new version of the heptathlon dataset in which the running events have been adjusted so that big numbers indicate good performances. Specifically, for each athlete j we now have

$$\begin{aligned} X^*_{1_j} &= \max(X_1) - X_{1_j}, \\ X^*_{4_j} &= \max(X_4) - X_{4_j} \quad \text{and} \\ X^*_{7_j} &= \max(X_7) - X_{7_j}. \end{aligned}$$

Interpretation of SPSS output

Figures 2.4 and 2.5 shows some SPSS output as a result of performing a PCA on the heptathlon data. You will see how to run this analysis for yourself in the computer practical session.

Figure 2.4 displays the 'a' values in our construction of the principal components Y. For example, we have:

$$\begin{aligned} Y_1 &= 0.275X^*_1 + 0.220X_2 - 0.001X_3 + 0.291X^*_4 + 0.301X_5 - 0.027X_6 + 0.289X^*_7, \\ Y_2 &= 0.153X^*_1 + 0.139X_2 + 0.515X_3 - 0.223X^*_4 + 0.019X_5 + 0.562X_6 + 0.008X^*_7, \\ Y_3 &= 0.512X^*_1 - 0.616X_2 - 0.042X_3 + 0.372X^*_4 - 0.238X_5 + 0.132X_6 - 0.133X^*_7, \end{aligned}$$

where X^*_1, X^*_4 and X^*_7 are defined as above, and $X_1, \ldots, X_7$ are as defined on page 15. What do these principal components tell us?

Component Score Coefficient Matrix

	Component		
	1	2	3
H100M	.275	.153	.512
HJ	.220	.139	-.616
SP	-.001	.515	.042
R200M	.291	-.223	.372
LJ	.301	.019	-.238
JT	-.027	.562	.132
R800M	.289	.008	-.133

Extraction Method: Principal Component Analysis.
Component Scores.

Figure 2.4: Some SPSS output for a PCA of the heptathlon dataset. Shown here are the principal components score coefficients.

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared ..	
Component	Total	% of Variance	Cumulative %	Total	% of Variance
1	2.602	37.168	37.168	2.602	37.168
2	1.484	21.204	58.373	1.484	21.204
3	1.146	16.366	74.739	1.146	16.366
4	.788	11.258	85.996		
5	.557	7.957	93.954		
6	.277	3.960	97.914		
7	.146	2.086	100.000		

Total Variance Explained

	Extraction Sums ..
Component	Cumulative %
1	37.168
2	58.373
3	74.739
4	
5	
6	
7	

Extraction Method: Principal Component Analysis.

Figure 2.5: Further SPSS output for a PCA of the heptathlon dataset. Of primary interest here is the percentage of variance explained by each principal component.

- Each component is a function of all of our original X variables, and in each component the X variables have a 'coefficient'. For example, in the *first principal component*, the coefficient for the 100 metres hurdles (X_1) is 0.275 and the coefficient for the shot put (X_3) is –0.001.

- First principal component:

  - The coefficients for the throwing events (shot put and javelin) are both very small relative to the coefficients for the running and jumping events
  - The coefficients of the running and jumping events are all very similar
  - Therefore, this principal component might be interpreted as an average measure of ability in the running and jumping events

- Second principal component:

  - ✎

  - ✎

Figure 2.5 also tells us how much of the variability in our original multivariate dataset is captured by our first three principal components: 37.168%, 21.204% and 16.366% for Y_1, Y_2 and Y_3 respectively, giving 74.739% of the total variation in our original dataset being captured by the first three principal components. This is pretty good – having defined three new variables, all of which capture information from our original seven variables, we have captured almost three-quarters of the total variability in our original dataset. But three variables is surely more manageable than seven!

Calculation of principal component scores

We can now calculate principal component *scores* for each athlete. For example, for Jessica Ennis, we have

$$\begin{aligned} Y_1 &= 0.275X_1^* + 0.220X_2 - 0.001X_3 + 0.291X_4^* + 0.301X_5 - 0.027X_6 + 0.289X_7^* \\ &= 0.275 \times 2.04 + 0.220 \times 1.86 - 0.001 \times 14.28 + 0.291 \times 3.08 + 0.301 \times 6.48 \\ &\qquad -0.027 \times 47.49 + 0.289 \times 31.90 \\ &= 11.73955. \end{aligned}$$

We also have

$$\begin{aligned} Y_2 &= 0.153X_1^* + 0.139X_2 + 0.515X_3 - 0.223X_4^* + 0.019X_5 + 0.562X_6 + 0.008X_7^* \\ &= 0.153 \times 2.04 + 0.139 \times 1.86 + 0.515 \times 14.28 - 0.223 \times 3.08 + 0.019 \times 6.48 \\ &\qquad +0.562 \times 47.49 + 0.008 \times 31.90 \\ &= 34.30572. \end{aligned}$$

We'd expect Jessica Ennis's first principal component score to be high, relative to other athletes' scores, as Table 1.1 reveals she seemed to perform well in the running and jumping events (recall our interpretation of the first principal component on page 17). However, is a score of 11.73955 high, relative to the scores of other athletes? And what about a score of 34.30572 on the second principal component?

Question

Find the scores on the first two principal components for Irina Karpova of Kazakhstan. Note that for this athlete, $X_1^* = 0.37$, $X_4^* = 0.49$ and $X_7^* = 11.62$ (you will find values for X_2, X_3, X_5 and X_6 in Table 1.1 on page 5).

✎

It can be useful to plot the scores for each athlete on the first principal component against those on the second principal component. This often reveals groupings or outliers in our dataset. However, the scores calculated for Jessica Ennis and Irina Karpova on the previous page are hugely influenced by the size of observations on different variables; not all of our variables are reported on the same scale. This is obvious, since we have variables corresponding to performances in different types of event – which are measured completely differently (heights, distances and times).

We could have easily recorded times in the 800 metres running event in milliseconds, giving us even larger values on variable X_7. Then, if a certain athlete performed particularly well in the 800 metres, but not so well in all the other events, her principal component scores may still be very high as the overall component might have been 'swamped' by this single strong performance.

Thus, SPSS calculates principal component scores on the *standardised* outcomes for each variable – basically, it subtracts the mean and divides by the standard deviation for performances on each variable, and uses these *standardised* data when calculating the principal component scores for each athlete.

Figure 2.6 shows a plot of scores on the first principal component against those on the second, for each of the 29 athletes in the 2012 heptathlon. These scores have been calculated using standardised data. Some notes:

- Jessica Ennis's score on the first principal component is now around 1, using standardised data (compared to 11.7 as calculated earlier using the raw data). Her score on the second principal component is now around 0.3, compared to 34.3 as found earlier.
- There are no clear 'groupings' in the dataset – scores for all athletes are concentrated in pretty much the same area.
- Points on the outskirts of the cluster might correspond to athletes with extreme performances – some could be outliers.
- Interesting example: Austra Skujyte has an average score on the first principal component, but a high score on the second. This indicates her performances across the running and jumping events were average, but relatively speaking her performance in the throwing events was very good.
- Interesting example: Katarina Johnson-Thompson has a relatively low score on the second principal component given her score on the first principal component. This is indicative of good performances in the running and jumping events for this athlete, relative to her performances in the throwing events.
- Interesting example: Julia Machtig has a relatively low score on the first principal component given her score on the second principal component. This is indicative of good performances in the throwing events running for this athlete, relative to her performances in the running and jumping events.

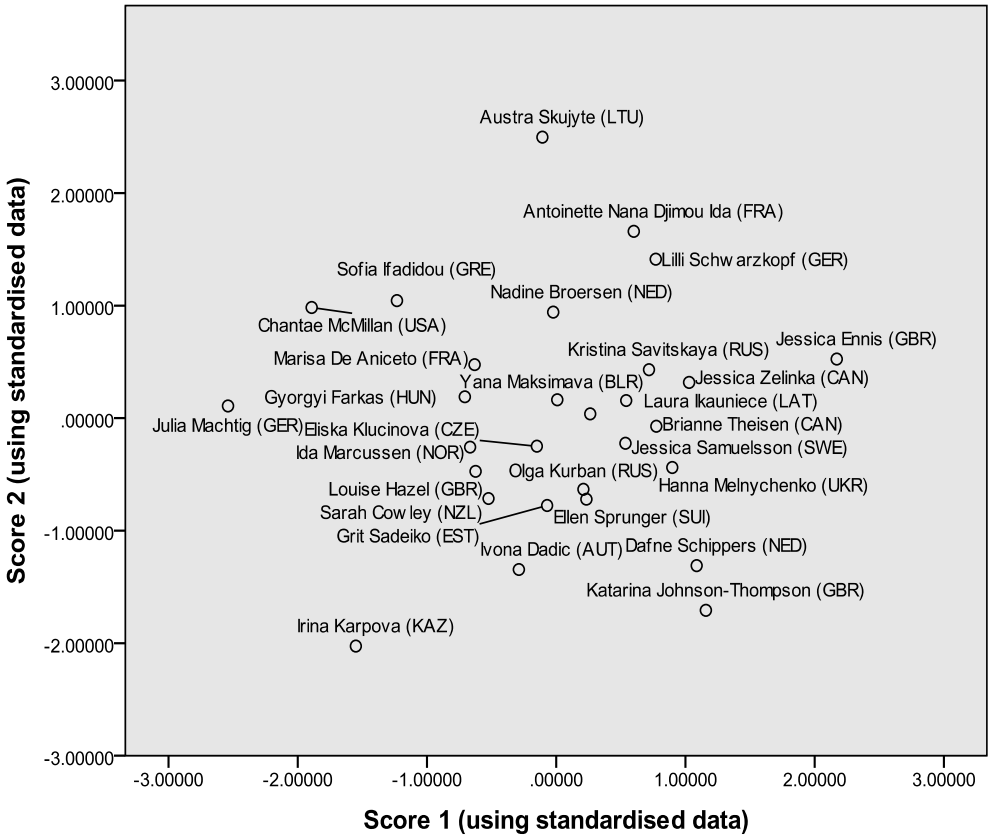

Figure 2.6: Score plot for the heptathlon data, as produced in SPSS.

2.2.2 PCA of the microarray dataset

To illustrate the power of PCA in the analysis of 'big data', we return to the motivating example presented in Section 1.4.3. Recall that the microrarray data consist of measurements on the expression levels of 6,830 genes from 64 samples taken from malignant tumours in different patients.

This is a much bigger dataset than the heptathlon dataset: there are 64 rows (one for each patient/tumour) and 6,830 columns (one for each gene). Obviously, it would be impractical to employ any of the methods presented in Section 2.1: imagine examining scatterplots or correlation coefficients for all possible pairs from 6,830 variables! There would be over 23 million plots and correlation coefficients to examine! The only way to examine such a dataset is through a PCA.

Figure 2.7 shows a score plot for the first two principal components from a PCA of the microarray data in SPSS. The analysis took SPSS around 72 minutes to complete on my laptop! What can we learn from this plot?

- Tissue samples from tumours of the same type tend to be located close to each other: given only information on the gene expressions of each tumour, the analysis confirms that tumours of the same type have similar genetic compositions.
- There are perhaps three distinct clusters of points in the score plot, suggesting different genetic compositions for certain groups of tumours. For example, most

melanomas seem very different to tumours of other type, as do leukaemia. This could indicate very different treatment regimes for these types of cancer, relative to other cancers.

- Interestingly, the breast and renal tumours towards the far left of the plot were both rare types of tumours – the PCA has perhaps detected this through different genetic compositions in these two samples.

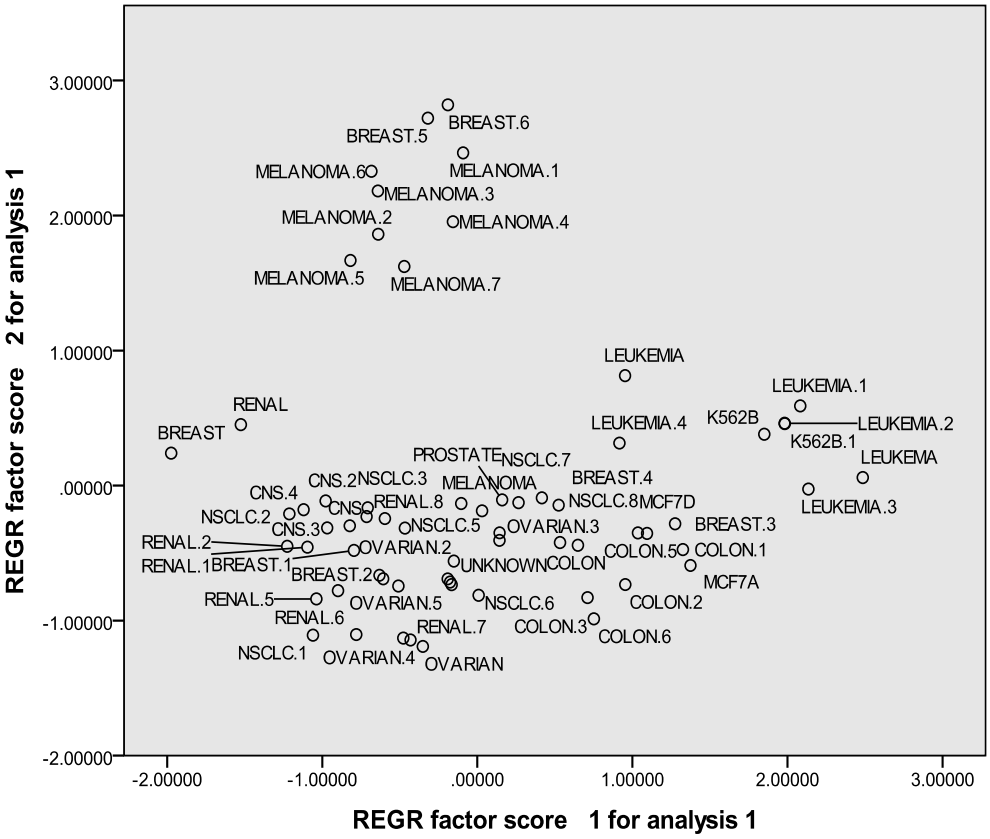

Figure 2.7: Score plot for the microarray data, as produced in SPSS.

Chapter 3

Modelling

In this Chapter, we will consider *multiple linear regression* as our main technique for modelling a multivariate dataset. This can be considered a direct extension of the simple linear regression model you considered in the first block of this module.

3.1 Introduction

The aim of the material in Chapter 2 was *exploration* of a multivariate dataset. Through various graphical and numerical summaries we considered simple ways of exploring a multivariate dataset, although these mainly pairwise methods are difficult (or impossible) to utilise in very large datasets (as we discussed for the microarray dataset, which consists of 6,830 variables!). Principal Components Analysis can then be used to reduce the dimensionality of our multivariate dataset, capturing most of the variability in our dataset in just a few variables, enabling us to more easily display our data. The score plot can then be used to explore groupings in our dataset or to more easily detect outliers.

In this Chapter, we will consider the problem of *modelling* our multivariate dataset. Now, one of the variables in our multivariate dataset is identified as the *response* or *dependent* variable, and the others are identified as the *predictor* or *independent* variables. Although exploration of the multivariate dataset is still of interest here, the primary goal is to build a model that enables accurate prediction of the response variable, given information on the predictor variables. We will demonstrate the method of multiple linear regression through direct application to the wine sales dataset, first discussed in Section 1.4.2.

3.2 Simple exploration

The first step of any multiple regression analysis is simple exploration, usually using the methods discussed in Section 2.1. Indeed, the first step of *any* statistical analysis should be simple exploration! Before you embark upon anything vaguely sophisticated, it is important to get a feel for your dataset. This will enable you to perform simple sanity

checks of any output associated with your analysis – having a feel for your dataset gives you a certain degree of intuition regarding your analysis, and whether the conclusions you draw are sensible.

Figure 3.1 shows a scatterplot matrix of the wine sales data, and Figure 3.2 shows the correlation matrix. It seems reasonable to assume that 'Price' would be our response variable ($£Y$) and the others our predictor variables; we might want to see how the age of the wine (X_1 years) and the amount of rainfall / average temperature during the grape-growing season (X_2mm / X_3°C, respectively) influences the price. We might also want to investigate whether the *indicator variable* – whether or not the wine was produced in New Zealand ($X_4 = 0$ or 1) – has any impact on the selling price of these wines.

Looking at the bottom row of plots in Figure 3.1, and the correlations in Figure 3.2, building a statistical model seems justified; there is clearly a relationship between Y and each of X_1, X_2, X_3 and X_4.

- Y versus X_1: As the age of wine increases, so too does the selling price. The older (perhaps 'vintage') wines are more expensive than the younger wines, and the relationship is fairly linear (that is, the points in the scatterplot seem to follow a straight line). This is supported by a statistically significant correlation coefficient of 0.733.
- Y versus X_2: There is clearly a relationship between observed rainfall during the grape-growing season and selling price, although this relationship is fairly complex. The selling price increases with rainfall, almost linearly, to a point; then the direction changes and selling price decreases. Perhaps there is an 'optimal' moisture level, in terms of the quality of the grapes produced, before the quality begins to deteriorate. Notice that the correlation coefficient here is *not* statistically significant – this is to be expected, however, as the correlation coefficient is designed to detect simple linear associations only.
- Y versus X_3: As the average temperature during the grape-growing season increases, so too does the selling price. Perhaps warmer weather contributes to better quality grapes, producing better (and hence more expensive) wines? Again, the relationship is reasonably linear. This relationship is supported by a statistically significant correlation coefficient of 0.887.
- Y versus X_4: The scatterplot for wine price versus the New Zealand indicator is not shown, as the indicator is binary and does not lead to an appealing plot. However, the correlation coefficient is statistically significant, and at 0.885 suggests that wines produced in New Zealand are significantly more expensive than wines from all other countries in the sample.

The only other significant association appears to be between X_3 and X_4, suggesting that wines from New Zealand are associated with grapes harvested from a warmer growing season. Notice that, unlike with the heptathlon dataset analysed in Chapter 2, this simple exploratory analysis will be followed by fitting a statistical model rather than a principal components analysis. This is because, unlike the heptathlon dataset, the wine dataset has a clear response variable (with several predictor variables).

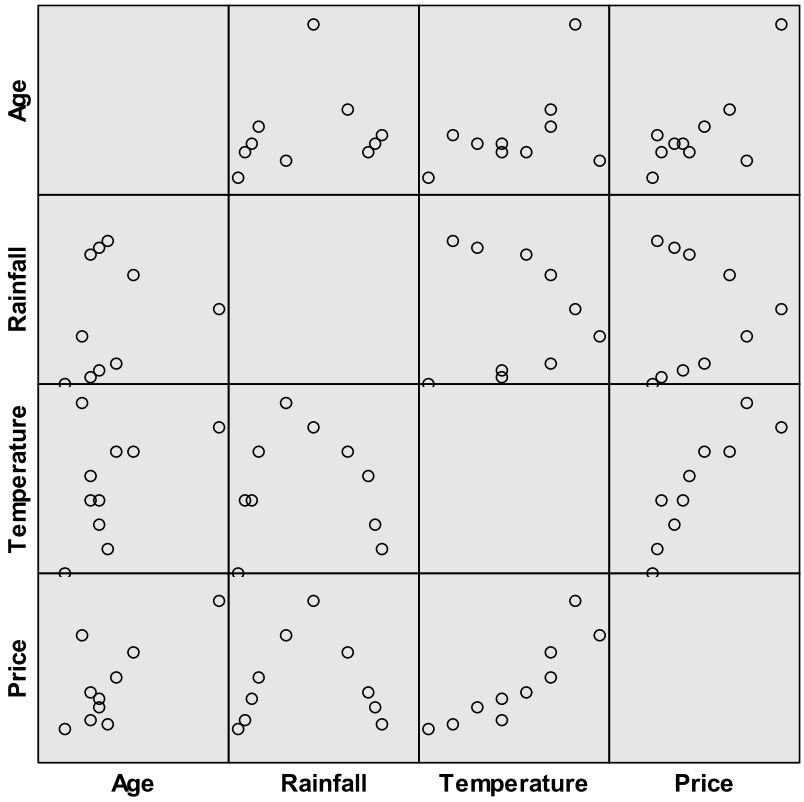

Figure 3.1: Scatterplot matrix for the wine sales dataset.

Correlations

		Age	Rainfall	Temperature	NZ	Price
Age	Pearson Correlation	1	.215	.475	.560	.733*
	Sig. (2-tailed)		.552	.165	.092	.016
	N	10	10	10	10	10
Rainfall	Pearson Correlation	.215	1	-.024	.109	.069
	Sig. (2-tailed)	.552		.948	.765	.849
	N	10	10	10	10	10
Temperature	Pearson Correlation	.475	-.024	1	.746*	.887**
	Sig. (2-tailed)	.165	.948		.013	.001
	N	10	10	10	10	10
NZ	Pearson Correlation	.560	.109	.746*	1	.885**
	Sig. (2-tailed)	.092	.765	.013		.001
	N	10	10	10	10	10
Price	Pearson Correlation	.733*	.069	.887**	.885**	1
	Sig. (2-tailed)	.016	.849	.001	.001	
	N	10	10	10	10	10

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

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

Figure 3.2: Correlation matrix for the wine sales dataset, as produced in SPSS.

3.3 Building a linear model

We now build on methods introduced in the first part of this module. Specifically, we construct a *multiple linear regression model* of the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon. \tag{3.1}$$

Notice that all variables in the wine sales dataset are represented in this model, with the response variable (selling price) being on the left-hand-side of Equation (3.1), and all of the predictor variables (age, rainfall, temperature and New Zealand indicator) being on the right-hand-side. Each of the predictor variables has a coefficient, labelled β_1 through to β_4, and there is also a 'constant' or 'intercept' term β_0. And of course, the usual 'error' term ϵ.

The aim now is to estimate values for the *regression coefficients*, which could help us decide whether or not the associated variables are important in the model. For example, if we find that the age coefficient β_1 is not significantly different from zero, then the associated predictor variable, X_1, would 'drop out' of the model and our conclusion would be that age is not a significant predictor of the wine's selling price. Fitting the model given by Equation (3.1) is known as fitting the 'full model', as this model includes *all* of our available predictor variables.

3.3.1 Fitting the full model

You will see how to fit the model in Equation (3.1) in the associated computer practical session; the SPSS output for this is shown in Figure 3.3. This reveals the fitted model:

$$Y = -11.6 + 0.626X_1 - 0.016X_2 + 1.035X_3 + 3.779X_4 + \epsilon. \tag{3.2}$$

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-11.600	8.464		-1.370	.229
	Age	.626	.212	.313	2.960	.032
	Rainfall	-.016	.052	-.027	-.303	.774
	Temperature	1.035	.293	.463	3.538	.017
	NZ	3.779	1.414	.367	2.672	.044

a. Dependent Variable: Price

Figure 3.3: Regression output for the wine sales data, as produced in SPSS.

Question

What does Equation (3.2), and the associated SPSS output in Figure 3.3, tell us?

Although not shown in Figure 3.3, SPSS gives a 'model summary', interpreted as a measure of performance of the multiple linear regression model. For the full model as shown in Equation (3.2), the model summary is $R^2 = 0.964$, which tells us that 96.4% of the variability in wine price is explained by the four predictor variables. The closer the value of R^2 is to 100%, the better the model, so we're not doing too badly here!

However, this model is rather complex, including information on four predictor variables. Are all of these variables necessary in the model? Recall from our discussion above that the rainfall coefficient was not statistically significant, indicating that X_2 might not be needed in the model.

3.3.2 Backwards elimination

We now remove rainfall from the regression model as its coefficient was not significantly different from zero, indicating that rainfall is not a significant predictor of price. Although our simple exploratory analysis in Section 3.2 revealed a strong pattern between rainfall and price, you might recall that this pattern was not simple, and perhaps not linear. Unfortunately the model form given by Equation (3.1) can only be used to capture linear associations between the response and predictor variables. Thus, the output in Figure 3.4 shows our updated model, now reduced from the full model, as fitted in SPSS.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-13.657	4.654		-2.934	.026
	Age	.613	.191	.307	3.215	.018
	Temperature	1.052	.265	.470	3.966	.007
	NZ	3.731	1.295	.363	2.882	.028

a. Dependent Variable: Price

Figure 3.4: Regression output for the wine sales data, as produced in SPSS. This is a reduced model, having removed the rainfall variable.

Some things to note:

- All remaining variables in the model (X_1, X_3 and X_4) seem to be important predictors of the selling price of wine, as the p-values associated with their regression coefficients (β_1, β_3 and β_4) are all less than 5% (1.8%, 0.7% and 2.8%, respectively). Thus the wine's age, the average temperature during the grape-growing season and whether or not the wine was produced in New Zealand are all significant (linear) predictors of the selling price.
- Of the remaining variables in the model, X_3 is the most significant, having the smallest p-value.
- Though not shown here, the model summary from SPSS is $R^2 = 0.963$, telling us that 96.3% of the variability in wine price is explained by our three remaining predictor variables. Notice this is a *very* slight reduction on the R^2 value obtained from our full model in Section 3.3.1; a small price to pay, perhaps, for simplifying our model and now only requiring information on three variables rather than four!
- Our fitted model is now:

$$Y = -13.657 + 0.613X_1 + 1.052X_3 + 3.731X_4 + \epsilon. \tag{3.3}$$

- As all remaining predictor variables are important predictors of selling price, the model in Equation (3.3) is often termed the *final model*.

In this Section we have demonstrated the process of *backwards elimination*. After having fitted the full model (as we did in Section 3.3.1), we then eliminate variables to see if they are actually needed in the model. How do we know which variables to eliminate?

1. Fit the full model, and make a note of the R^2 value.
2. Look for variables with insignificant coefficients – that is, variables for which the coefficients have p-values greater than 5%. In the wine sales analysis, there was only one candidate for elimination: the rainfall variable was the only variable whose coefficient had a p-value greater than 5%. But what if there was more than one apparently unimportant predictor variable?
3. Remove only one variable at a time: of those variables with p-values greater than 5%, choose the one which is least significant (i.e. the one with the greatest p-value). Remove this variable, and re-fit the multiple linear regression model in SPSS, making a note of the new R^2 value. How has the R^2 changed? If it has increased, or only decreased very slightly, the removal of the variable you selected was worthwhile. If it has decreased considerably, make a decision as to whether or not the variable should be removed. Perhaps it was marginally significant, with a p-value close to 5%?
4. Repeat steps 2 and 3 until all remaining predictor variables are significant and you are happy with the R^2 model summary. You have now completed the process of backwards elimination, and have obtained the final model.

You will see a more complicated example of the process of backwards elimination in the computer practical session.

3.3.3 Using your model

We will now see how the final model can be put to good use.

Question

During the growing season in 2014, the Marlborough region of New Zealand experienced a total of 115mm of rainfall, and the average afternoon temperature was 17°C. How much might we expect a bottle of wine to sell for this year, in the UK?

✎

3.3.4 Checking assumptions

As already discussed, the R^2 model summary can be used to check the overall quality of our model. However, there are various assumptions underpinning a multiple linear regression model, as set out in Equation (3.1), that must also be checked – especially if we are to trust calculations such as that in Section 3.3.3. The main assumption is that the model error term, denoted by ϵ in Equation (3.1) and throughout, is Normally distributed with constant variance. This is a rather complex concept, but one that can be quite easy to check in SPSS. You will see how to do this in the computer practical; for our wine analysis, based on the final model obtained in Section 3 3.2, we can obtain the plots shown in Figure 3.5.

The left-hand plot in Figure 3.5 shows a chart of the calculated residuals ϵ. If our assumption that they follow a Normal distribution is valid, then we would expect to see the chart resembling the classic 'bell-shaped curve' of the Normal distribution. We can see that this is questionable for our wine sales analysis, although such charts are difficult to interpret for such small sample sizes (we only have 10 wines in our sample!).

The right-hand plot in Figure 3.5 checks for constant variance (often termed 'homoscedasticity'). Basically, in this plot you're looking for random scatter, with no pattern or structure. This appears to be the case here, although again this can be difficult to interpret with such a small sample.

If the assumptions underpinning the model were found to be invalid for our data, then the the validity of the model itself – and hence our use of the model – might be brought into question.

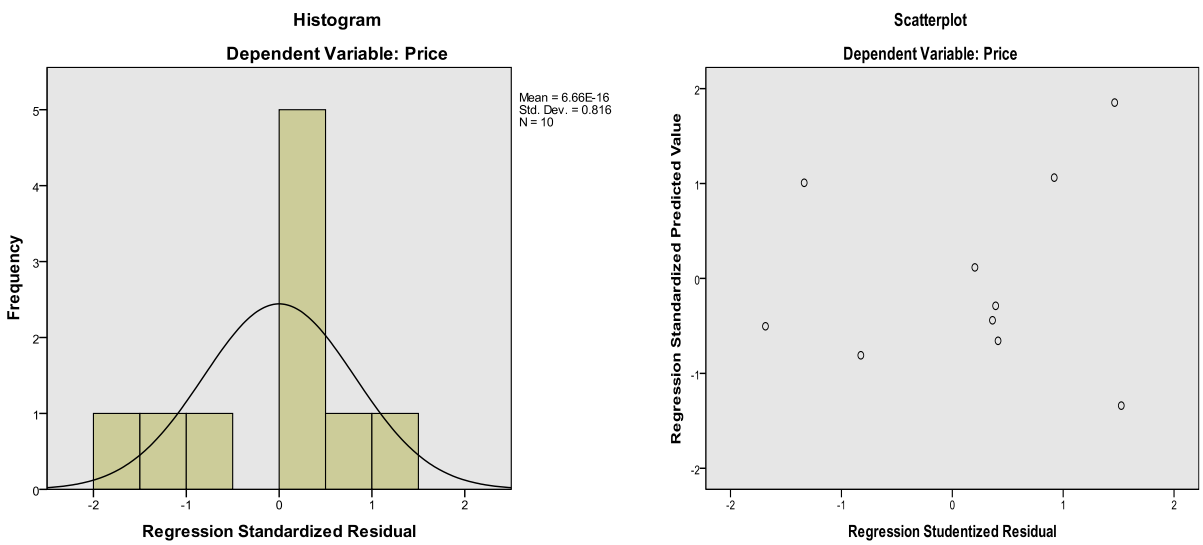

Figure 3.5: Checking assumptions of the multiple linear regression model for the wine sales data. Left: histogram of the model residuals, compared to a standard Normal distribution; right: plot of residuals versus predicted values, used to detect heteroscedasticity.

3.4 Further topics in regression

3.4.1 Non-linear regression

In our multiple regression analysis in Section 3.3, the rainfall variable was found not to be a significant (linear) predictor of wine selling price. However, as the scatterplot matrix in Figure 3.1 revealed, there was a clear (bot non-linear) association between rainfall and price. This is an example of when a non-linear regression might be appropriate. You will see how to perform a non-linear regression in SPSS in the computer practical session.

Let us focus on this association, and disregard information on the other predictor variables (X_1, X_3 and X_4). The simplest form of non-linear regression model is a quadratic regression, the SPSS output for which is shown in Figure 3.6 for our wine sales data.

- The fitted model is now:

$$Y = -1658.13 + 28.967X_2 - 0.125X_2^2 + \epsilon. \tag{3.4}$$

- Notice that the price of wine now depends on rainfall, but also (rainfall)^2, hence the name 'quadratic' regression used here.
- Both the rainfall and (rainfall)^2 variables are important predictors of price, as both of their coefficients are extremely significant (the associated p-values are 0.000).
- Although not shown in Figure 3.6, the model summary for this quadratic regression is $R^2 = 0.974$ – surpassing that for our multiple regression models in Section 3.3. Thus, 97.4% of the variability in wine price is explained by rainfall and (rainfall)^2.
- Figure 3.7 confirms that our model fits well – we see the quadratic regression curve superimposed on the scatterplot of price versus rainfall (as previously shown in Figure 3.1).

Question

Suppose the total rainfall at a vineyard over the grape-growing season was 115mm. How much might we expect a bottle of wine to sell for?

✎

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-1658.130	102.620		-16.158	.000
	Rainfall	28.967	1.784	49.256	16.235	.000
	RainfallSq	-.125	.008	-49.197	-16.216	.000

a. Dependent Variable: Price

Figure 3.6: SPSS output for a quadratic regression linking wine selling price to rainfall accumulations observed during the grape-growing season.

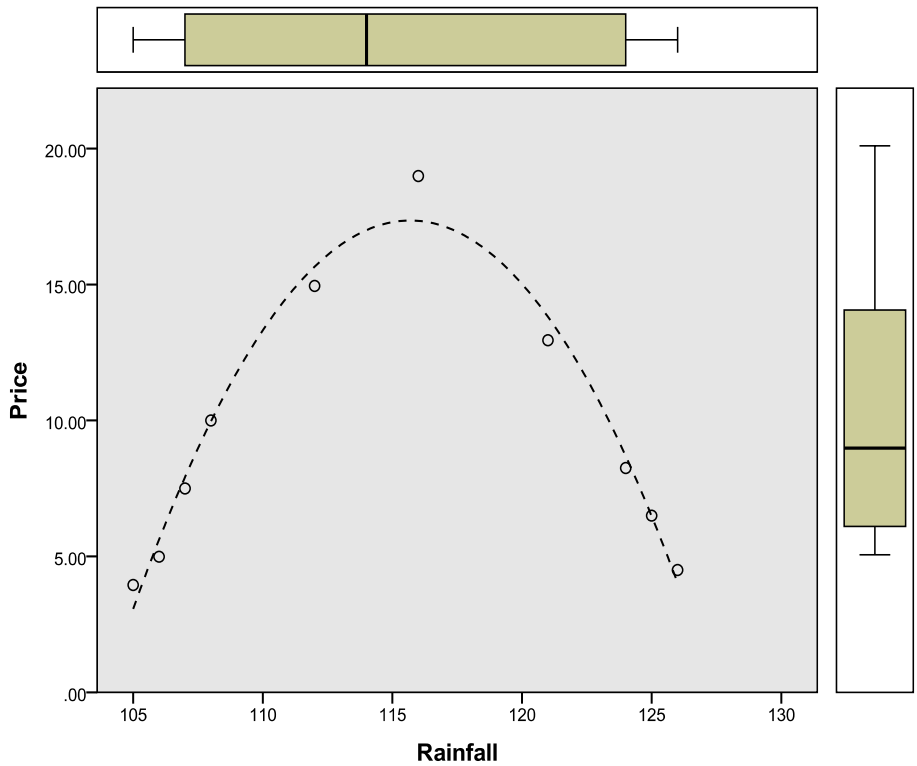

Figure 3.7: Scatterplot of wine selling price against rainfall accumulations observed during the grape-growing season, with fitted quadratic regression curve, as given by Equation (3.4), superimposed.

3.4.2 Analysing a binary response: logistic regression

Suppose the marketing team at *Mars* are interested in the ability of cinema–goers to recall their brand using *product placement* during a film. Further, they think there might be a relationship between the chance of someone being able to recall their brand and the time in the film at which the product placement occurred. To investigate, 25 volunteers took part in a marketing experiment. Initially, the volunteers knew nothing about the aims of the experiment; after they each watched a film of length $2\frac{1}{4}$ hours, they were asked if they could recall the brand that had been "placed" in their film. The product placement for each volunteer happened at a different time during the film. The results are shown in Table 3.1 below, where X is the time, from the start of the film, at which the product placement occurred and Y takes the value 1 if the volunteer *could* recall the brand, and 0 if they *could not*.

Volunteer	1	2	3	4	5	6	7	8	9	10	11	12	13
X (minutes)	10	15	20	25	30	35	40	45	50	55	60	65	70
Y	0	0	0	0	0	0	1	0	0	1	0	1	0
Volunteer	14	15	16	17	18	19	20	21	22	23	24	25	
X (minutes)	75	80	85	90	95	100	105	110	115	120	125	130	
Y	1	1	1	1	0	1	1	1	0	1	1	1	

Table 3.1: Product placement dataset.

Question

Comment on the relationship between X and Y as shown by the data in Table 3.1.

Notice that the response variable here – whether a volunteer can/cannot recall the *Mars* brand – is not like the response variable in the wine sales example (price of a bottle of wine). The response variable is now *binary* – i.e. it can only take one of two values; we use 0 for 'no' and 1 for 'yes'. We have already thought about how to perform a regression analysis when one of the *predictor* variables is binary (see Section 3.3), but not when the main *response* variable takes this form.

We cannot perform a standard regression of Y on X here: our response variable takes the value 0 or 1, whereas the output from a standard simple linear regression would not be restricted to take just these values. When our response variable is binary, we usually perform a *binary logistic regression*. You will see how to conduct such an analysis, in SPSS, in the computer practical session. Shown in Figure 3.8 is the output from SPSS for such an analysis as applied to the data in Table 3.1.

- A binary logistic regression ensures the model returns values between 0 and 1; so not exactly on the same scale as the response variable itself (which is entirely binary), but returning values on a restricted scale. In fact, the values returned from a binary logistic regression model are *probabilities*; specifically, the model form for our product placement example is

$$P(Y = 1) = \frac{e^{-2.99+0.45X}}{1 + e^{-2.99+0.45X}}. \tag{3.5}$$

- The precise form of Equation (3.5) ensures the output lies between 0 and 1.
- Notice that the time coefficient is significant, since its p-value is less than 5% (in fact, it's 1%). This suggests that the time at which the product placement takes place in the movie does indeed impact upon a volunteer's ability to recall the brand. As the coefficient is positive, this suggests that as the time into the film increases, so too does a person's likelihood of being able to recall the brand (as you might expect?).

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Time	.045	.017	6.618	1	.010	1.046
	Constant	-2.991	1.282	5.443	1	.020	.050

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

Figure 3.8: SPSS output for a quadratic regression linking wine selling price to rainfall accumulations observed during the grape-growing season.

Question

Suppose 100 people watch a movie with a product placement appearing at 92 minutes. How many of these people might we expect to be able to recall the brand?

✎

3.5 Bringing it all together

The techniques discussed in Section 3.4 can, of course, be incorporated into a multiple regression model. For example, we might have a binary response, such as that discussed in Section 3.4.2, that might depend on several variables – some of which might reveal a non-linear (perhaps quadratic) association with the response.

Chapter 4

Practical work

In this Chapter, we will revisit some of the examples from Chapters 2 and 3, looking at how to implement the methods in SPSS and how to interpret the associated output.

The best way to access the course datasets is to:

- Visit the dedicated webpage for this Stream:

  `www.mas.ncl.ac.uk/~nlf8/teaching/hss8005`

- Click on the Practicals tab
- Right-click on the dataset you're interested in
- Save the file somewhere convenient in your filespace
- Import the data into SPSS: within SPSS, select File⟶Import Data⟶CSV Data; find the dataset in the folder in which you saved it, and select Open; then follow the on-screen instructions within SPSS
- Alternatively, the datasets can also be accessed through Blackboard

4.1 Exploratory multivariate data analysis in SPSS

You should work through as many questions in this Section as you can during the first computer practical session. Any questions you do not complete should be finished in your own time. Lots of guidance is included with the first question, less so as you progress through the Section.

1. Let's consider the construction of simple **scatterplot matrices** in SPSS. Import the `heptathlon` dataset that we studied in lectures, and take a look around the dataset. You should have 29 rows in the dataset (one row for each athlete), and 8 columns (one for the athlete's name, and then one containing each athlete's performance across the seven events in the heptathlon).

   To construct a scatterplot matrix:

   (i) Click `Graphs`⟶`Chart Builder`⟶`OK`

   (ii) From the `Gallery` menu, choose `Scatter/Dot`

   (iii) You should see eight types of plot appear; hovering over each one should reveal a text box, explaining what each plot is. Select the `Scatterplot Matrix` by double-clicking on the second box of the bottom row

   (iv) An empty chart should now appear in the chart preview box. From the `Variables` list, you should now drag-and-drop all of the variables associated with the heptathlon into the `Scattermatrix` box underneath the empty chart. Don't include the names of the athletes, just the seven variables containing the heptathlon performances. Note that you can select multiple variables all at once by selecting the first one in the `Variables` list, and then selecting the last one with the shift key held down you can then move the whole block of variables across to the `Scattermatrix` box in the empty chart

   (v) Click `OK`

   You should now have produced the scatterplot matrix shown in Figure 2.1 of the lecture notes (page 10).

   Questions:

   (a) Look along the top row of plots. What do these plots show? What is the nature of the association between performances in the 100 metres hurdles and the 800 metres? What about the relationship between performances in the 100 metres hurdles and the long jump?

   (b) How are performances in the shot put and the javelin throw associated?

   Keep the heptathlon dataset open in SPSS, you will need it for the next question!

2. We will now consider the construction of **correlation matrices** in SPSS. Once again, we will use the heptathlon dataset as an example.

> **To construct a correlation matrix:**
>
> (i) Select `Analyze`⟶`Correlate`⟶`Bivariate`
>
> (ii) Select the 100 metres hurdles, and then select the 800 metres with the shift key held down. Then move all variables across together by clicking the arrow
>
> (iii) Click `OK`

You should now have produced the correlation matrix shown in Figure 2.2 of the lecture notes (page 12).

Questions:

(a) Can you explain how the correlation matrix supports your answers in questions 1(a) and 1(b)?

(b) Using the correlation matrix, can you quickly identify the pairs of variables that are significantly associated? What about those that are not? Which pair of variables reveals the most significant association?

(c) Why do all the cells down the diagonal of the matrix contain '1'?

Keep the heptathlon dataset open in SPSS, you will need it for the next question!

3. SPSS can be used to produce **3-D scatterplots**. We will demonstrate their construction via the heptathlon dataset, so keep this dataset open in SPSS!

> **To produce a 3-D scatterplot:**
>
> (i) Select `Graphs`⟶`Chart Builder`⟶`OK`
>
> (ii) Select `Scatter/Dot` in the `Gallery` menu
>
> (iii) Find `Simple 3-D Scatter` (first row, 3rd plot), and double-click
>
> (iv) Drag-and-drop three variables of your choosing into the `X-Axis? Y-Axis?` and `Z-Axis?` boxes of the preview chart
>
> (v) Click `OK`

You should now have produced a 3-D scatterplot in keeping with that in Figure 2.3 of the lecture notes (page 14).

Question:

(a) What can you say about the plot you have just produced? Are the relationships between the three variables you selected clear, as the plot stands?

You might be finding it rather challenging to answer the question in part (a). If so, why not try spinning your plot?

Spinning a 3-D scatterplot:

(i) Double-click on your 3-D scatterplot so it opens in the `Chart Editor` window

(ii) Select `Edit`⟶`3-D Rotation`

(iii) Now click anywhere on your plot and drag the plot round in any direction you like!

Question:

(b) Rotate your 3-D scatterplot so you can easily see the association between a single pair of variables you have selected. You should see that you have a recreation of the corresponding plot in the scatterplot matrix in question 1(a).

4. In Section 2.2 of the lecture notes we considered **Principal Components Analysis** (PCA) as a way of exploring multivariate datasets. Recall from page 16 of the notes that, for easier interpretation of the principal components, we should make sure the direction of each variable has the same interpretation. In the notes, we re-scaled data on the running events to make high values correspond to better performances.

   Import the `heptathlon2` dataset into SPSS. Pay particular attention to the data in the columns for the running events.

Performing a Principal Components Analysis:

(i) Click on `Analyze`⟶`Dimension Reduction`⟶`Factor`

(ii) Highlight all seven heptathlon events and move them all over into the Variables box in one go (see question 2 for how to do this)

(iii) Select `Extraction`, and make sure `Correlation matrix` is selected before you press `Continue`

(iv) Select `Scores`, then select `Save as Variables`. Also make sure `Regression` is selected in the `Method` box, and that you check the box that says `Display factor score coefficient matrix`. Press `Continue`

(v) Click `OK`

There should now be lots of output printed to the screen! Let's now go through it all.

Questions:

(a) First of all, the principal components. In your SPSS output, look for the `Component Score Coefficient Matrix`. The first column gives you the coefficients of your first principal component, so you should see that

$$Y_1 = 0.275X_1^* + 0.220X_2 - 0.001X_3 + 0.291X_4^* + 0.301X_5 - 0.027X_6 + 0.289X_7^*,$$

where the heptathlon variables have labels $X_1, X_2, \ldots$ for the 100 metres hurdles, the high jump, ..., and the stars indicate that performances in the running events have been re-scaled so that high values indicate better performances.

What is your second principal component? And the third? Compare your answers to the material at the bottom of page 16 of the lecture notes.

(b) Recall that the aim of a PCA is to reduce the dimension of a multivariate dataset, perhaps so that it can be viewed in two dimensions using standard techniques. We would like most of the variability across our multivariate dataset to be explained by just a few of the principal components (which are functions of all of the original variables in our dataset).

Look for the Table entitled `Total Variance Explained`. What percentage of the variability in the heptathlon dataset is explained by the first two principal components?

Constructing a PCA score plot:

Finally, let's produce a score plot. Given the options you selected to perform the PCA, you should notice that the (standardised) scores for each athlete, on each of the principal components, have been added to the original data spreadsheet in the columns to the right-hand-side of the heptathlon dataset.

(i) Select `Graphs`⟶`Chart Builder`⟶`OK`

(ii) From the `Gallery` menu choose `Scatter/Dot`, and double-click the `Simple scatter` option

(iii) Select '`REGR factor score 1`' for the `X-Axis` and '`REGR factor score 2`' for the `Y-Axis`

(iv) Before clicking `OK`, select `Groups/Point ID label` and select `Point ID label`. In the chart preview you should see that a `Point Label Variable?` box has appeared; drag and drop the variable corresponding to the Athlete's names into the `Point Label Variable?` box, before clicking `OK`.

You should now have reproduced the plot in Figure 2.6 of the lecture notes (page 21). Consider the comments on page 20 of the lecture notes for how we might interpret this plot.

4. This is an **Open Question**. Import the `USArrests` dataset into SPSS. Take a look around the dataset. You should see arrests, per 100,000 residents, for assault, murder and rape in each of the 50 US states in 1983. Also given is the percent of the population living in urban areas, and a 'region' identifier. You should use exploratory data analysis techniques to investigate this multivariate dataset.

   Hint: If you perform a PCA following the steps in question 4 (page 40), when you select `Extraction` *make sure '*`Fixed number of factors`*' is checked, and enter 2 in the '*`Factors to extract:`* box.*

4.2 Modelling multivariate data in SPSS

You should work through as many questions in this Section as you can during the second computer practical session. Any questions you do not complete should be finished in your own time.

1. In Section 3.3 of these notes we considered **multiple linear regression** as a way of modelling multivariate datasets. Let's consider an example.

   Suppose you work as an advertising manager for a company that markets do-it-yourself (DIY) products through a nationwide chain of retail outlets. The company advertises mainly on a local basis, via newspapers, hoardings and the exterior of buses. You have been asked to assess the relationship between the amount spent on advertising per quarter and the corresponding level of sales achieved.

   After giving it some thought, you realise that many factors could influence sales: as well as the amount spent on advertising, the number of households in the catchment area of a particular shop could affect sales; so too could the number of competitor shops within this catchment area; even the weather in the region may have an impact on DIY activity!

   Clearly it is impossible to include every variable which may show some relationship with sales, so you decide as a starting to point to include

   1. Advertising expenditure
   2. Number of households within a five mile radius
   3. Number of competitor shops within a five mile radius

   These data can be opened by importing the `diy` dataset from the course webpage.

   Questions:

   (a) Perform a simple exploratory analysis of the DIY dataset using scatterplot matrices and correlation matrices, as covered for the heptathlon data in Section 4.1.

   (b) What can you say about the associations between the response variable (`Sales`) and the other variables?

   Fitting the 'full' multiple linear regression model:

   (i) Select `Analyze`⟶`Regression`⟶`Linear`

   (ii) Enter the response variable (`Sales`) as the `Dependent`

   (iii) Enter `Advertising`, `Households` and `CompetingStores` as the `Independent(s)`

   (iv) Press `OK`

   You have now fitted the 'full' multiple linear regression model. Let's consider the output.

Questions:

(c) Locate the `Coefficients` box in the SPSS output. Hopefully, you can see that your regression equation is

$$\begin{aligned}\text{Sales} &= -0.803 - 19.766 \cdot \text{Advertising} + 0.036 \cdot \text{Households} \\ &\quad +4.374 \cdot \text{CompetingStores} + \epsilon.\end{aligned}$$

(d) Using the information in the `Coefficients` box, which of the predictor variables appears most important (smallest p-value)? Which appears least important (largest p-value)? Which variables do you think might not be needed in the model (p-values greater than 5%, or 0.05)?

(e) Locate the `Model Summary` box in the SPSS output. What proportion of the variability in sales is explained by the three predictor variables?

Backwards elimination

(i) You should have noticed that *all three* of the predictor variables might be removed from the model, as they *all* have p-values greater than 5%. However, we only remove one variable at a time – the one which is *least* important: in this case, `Advertising`.

(ii) Repeat steps (i)-(iv) above for fitting the full model, but now do not include the `Advertising` variable

(iii) Look at your new `Coefficients` box. Do any of the remaining variables have p-values greater than 5%? Which is the *least* important? Remove the variable you think is not significant, and least important in the model, and re-fit; continue this process until all remaining variables are important in the model, giving you your 'final model'

Questions:

(f) After completing your process of backwards elimination, you should find that the only remaining important predictor in the model is `CompetingStores`, with regression equation

$$\text{Sales} = 12.959 + 5.235\text{CompetingStores} + \epsilon.$$

What is the p-value associated with the variable `CompetingStores`? How would we interpret the coefficient of 5.235?

(g) Locate the new `Model Summary` box. What proportion of the variability in sales is explained by the number of competing stores, and how has this changed from that of the full model?

(h) Let's suppose that, for a proposed new store, there will be eight competing stores within a five mile radius. What level of sales might we expect at our new store?

Checking assumptions

(i) Repeat the regression analysis above, using `CompetingStores` as the only predictor variable; i.e. `Analyze`⟶`Regression`⟶`Linear`. However, before you click `OK` in SPSS, select `Plots`

(ii) Insert `*ZPRED` in the `Y:` box and `*SRESID` in the `X:` box

(iii) Also check the box for `Histogram` in the option for `Standardized Residual Plots`

(iv) Hit `Continue`, and then `OK`

You should now have plots for the DIY dataset similar to those in Figure 3.5 of the lecture notes for the wine sales data (page 30).

Question:

(i) How would you interpret the plots you have just produced? *Hint: See Section 3.3.4 for some help here*

2. Import the `wine` dataset into SPSS. We will use this to demonstrate a simple **non-linear regression**, as discussed in Section 3.4.1.

   Questions:

   (a) Following the instructions for the analysis of the DIY dataset in question 1, you should be able to reproduce the multiple linear regression example discussed in Section 3.3 of these notes.

   (b) You might recall, from Section 3.4.1, the non-linear association between the total rainfall accumulation during the grape-growing season and the selling price of wine. In fact, in Section 3.4.1 we performed a *quadratic regression* to capture this complex relationship.

   Fitting a quadratic regression model:

   (i) Recall that a quadratic regression model fits the response variable to data on the predictor variable and the square of the predictor variable. In SPSS, we therefore need to create a new variable – say (Rainfall)^2. Thus, select `Transform`$\longrightarrow$`Compute Variable`

   (ii) In the `Target Variable` box, type `RainfalSq`

   (iii) Move the `Rainfall` variable into the `Numeric Expression` box

   (iv) From the on-screen calculator, select the `**` key and also enter the number 2

   (v) Now, in the `Numeric Expression` box, you should have `Rainfall**2`; in other words, (Rainfall)^2

   (vi) If you now look in your SPSS data window, you should see that there is now a new column of data called `RainfallSq`, containing the squares of the original `Rainfall` values

   (vii) Now select `Analyze`$\longrightarrow$`Regression`$\longrightarrow$`Linear`

   (viii) Move `Price` into the `Dependent` box, and `Rainfall` and `RainfallSq` into the `Independent(s)` box

   (ix) Select `OK`

   You should now have re-created the analysis discussed in Section 3.4.1 of the lecture notes.

   Questions:

   (c) Locate the `Coefficients` table in your SPSS output. What is your quadratic regression equation? Are the Rainfall and (Rainfall)^2 variables significant?

   (d) We have a model summary of $R^2 = 0.974$. What does this mean?

   (e) Can you perform a full multiple linear regression analysis on the wine sales data, to include this non-linear relationship between wine sales and Rainfall?

3. Import the dataset `potter` into SPSS. Take a look around the data. You should see that there are 30 rows, each corresponding to a participant in a recent survey investigating whether or not self-declared Harry Potter fans 'identify' with the main protagonist in the franchise.

   The column labelled 'Potter' is the main response variable, taking 1 if the survey participant said they *did* identify with Harry Potter, and 0 otherwise. Column 2 shows the Age of the participant, and column 3 takes the value 0 if the participant was male, and 1 if they were female.

   Notice that the main response variable is binary, so we will perform a **binary logistic regression** in SPSS.

   Fitting a binary logistic regression model

   (i) Select `Analyze`⟶`Regression`⟶`Binary Logistic`

   (ii) Declare `Potter` as the `Dependent` variable

   (iii) Both `Age` and `Sex` should be entered into the box entitled `Block 1 of 1`

   (iv) Press `OK`

   Questions:

   (a) In your SPSS output, locate the box `Variables in the Equation`

   (b) Notice that the p-value for the Sex variable is 0.084, or 8.4%. This is greater than 5%, and so this variable should be removed from the model; the p-value for the Age variable *is* less than 5%, and so this variable *does* appear important in the model

   (c) Follow the steps above for performing a binary logistic regression in SPSS, but now without the Sex variable. You should see that your regression equation is now
   $$P(\text{Harry Potter} = 1) = \frac{\exp\{4.045 - 0.259 \cdot \text{Age}\}}{1 + \exp\{4.045 - 0.259 \cdot \text{Age}\}}.$$
   How would you interpret the Age coefficient in this model?

   (d) Suppose a new 19-year-old Harry Potter fan is recruited to the study. How likely is it that he will identify with Harry Potter?

4. This is an **Open Question**. Import the `attitudes` dataset into SPSS. Take a look around the dataset. You should see a multivariate dataset consisting of seven observations on 30 cases.

   From a survey of the clerical employees of a large financial services organisation, the data are aggregated from the questionnaires of the approximately 35 employees for each of 30 (randomly selected) departments. The numbers give the percentage of favourable responses to seven questions in each department. You should use multiple linear regression techniques to explore this multivariate dataset.