HSS8005: Stream A – Multivariate Techniques

Dr. Lee Fawcett

Semester 2, 2020/21

Course introduction

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
- 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.

Course introduction

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.

Course introduction

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 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**
- 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

Course introduction

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

Date	Time	Place	Activity	Topics
Mon 15 Mar	1–2	Zoom live	Lecture	Chapter 1
Tue 16 Mar	10–12	Zoom live	Lecture	Chapters 2 & 3
Wed 17 Mar	12–1	Zoom live	Practical	Chapter 2
Thu 18 Mar*	10–11	Zoom live	Seminar	Chapter 2
Tue 23 Mar	12:30–1:30	Zoom live	Practical	Chapter 3
Wed 24 Mar	10–11	Zoom live	Seminar	Chapter 3
Wed 24 Mar**	1–2	Zoom live	Lecture	Chapters 1–3

**This Zoom seminar is likely to be replaced with private directed reading*

***Date/time of concluding lecture TBC*

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:

- *Amazon*: World leader in collecting, storing, processing and analysing personal information from you and every other customer
- 2020 US election campaign
- 2019 UK General Election
- My own experience: impact work with road safety engineers

Multivariate Analysis – sometimes **two steps**:

- Exploratory Data Analysis
- Modelling

1.1 Background

With today's data collection systems and databases, a wealth of information is available to analysts at most organisations:

- *Amazon*: World leader in collecting, storing, processing and analysing personal information from you and every other customer
- 2020 US election campaign
- 2019 UK General Election
- My own experience: impact work with road safety engineers

Multivariate Analysis – sometimes **two steps**:

- Exploratory Data Analysis
- Modelling

1.1 Background

With today's data collection systems and databases, a wealth of information is available to analysts at most organisations:

- *Amazon*: World leader in collecting, storing, processing and analysing personal information from you and every other customer
- 2020 US election campaign
- 2019 UK General Election
- My own experience: impact work with road safety engineers

Multivariate Analysis – sometimes **two steps**:

- Exploratory Data Analysis
- Modelling

1.1 Background

With today's data collection systems and databases, a wealth of information is available to analysts at most organisations:

- *Amazon*: World leader in collecting, storing, processing and analysing personal information from you and every other customer
- 2020 US election campaign
- 2019 UK General Election
- My own experience: impact work with road safety engineers

Multivariate Analysis – sometimes **two steps**:

- Exploratory Data Analysis
- Modelling

1.1 Background

With today's data collection systems and databases, a wealth of information is available to analysts at most organisations:

- *Amazon*: World leader in collecting, storing, processing and analysing personal information from you and every other customer
- 2020 US election campaign
- 2019 UK General Election
- My own experience: impact work with road safety engineers

Multivariate Analysis – sometimes **two steps**:

- Exploratory Data Analysis
- Modelling

1.1 Background

With today's data collection systems and databases, a wealth of information is available to analysts at most organisations:

- *Amazon*: World leader in collecting, storing, processing and analysing personal information from you and every other customer
- 2020 US election campaign
- 2019 UK General Election
- My own experience: impact work with road safety engineers

Multivariate Analysis – sometimes **two steps**:

- Exploratory Data Analysis
- Modelling

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.

Example: **Predicting the weather**. Meteorologists require information on pressure, wind conditions, humidity, season, ...

1.2 Why 'multivariate'?

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**.

1.2 Why 'multivariate'?

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.

Throw a third variable into the mix – perhaps shoe size – and we're into the realms of 3-D scatterplots.

Graphically displaying the results for more than three variables **requires some ingenuity**, since we live in a 3-D world!

1.2 Why 'multivariate'?

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.

Throw a third variable into the mix – perhaps shoe size – and we're into the realms of 3-D scatterplots.

Graphically displaying the results for more than three variables **requires some ingenuity**, since we live in a 3-D world!

1.2 Why 'multivariate'?

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.

Throw a third variable into the mix – perhaps shoe size – and we're into the realms of 3-D scatterplots.

Graphically displaying the results for more than three variables **requires some ingenuity**, since we live in a 3-D world!

1.2 Why 'multivariate'?

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.

Throw a third variable into the mix – perhaps shoe size – and we're into the realms of 3-D scatterplots.

Graphically displaying the results for more than three variables **requires some ingenuity**, since we live in a 3-D world!

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
- **Unwanted variability** results from our inability to control something, and it is usually random in nature

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.

1.3 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
- **Unwanted variability** results from our inability to control something, and it is usually random in nature

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.

1.3 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
- **Unwanted variability** results from our inability to control something, and it is usually random in nature

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.

1.3 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
- **Unwanted variability** results from our inability to control something, and it is usually random in nature

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.

1.3 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
- **Unwanted variability** results from our inability to control something, and it is usually random in nature

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:

- **Business intelligence**: predicting changes in dynamic markets or better product placement
- **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
- **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

1.4 Applications of MVA

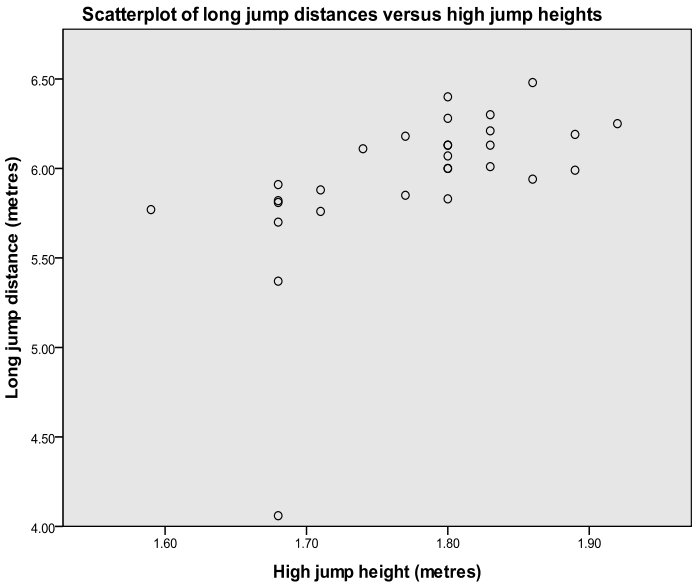
Some datasets used in this course:

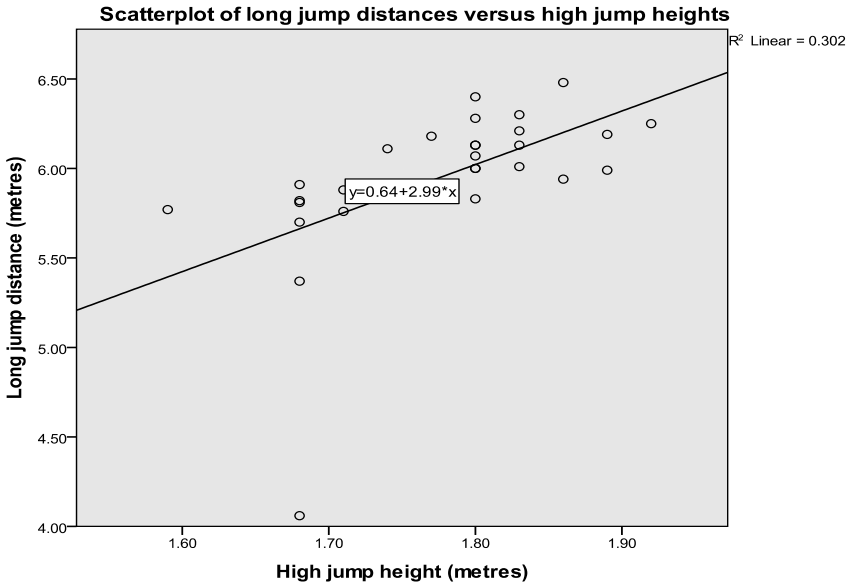
- London 2012 heptathlon data
- Wine sales data
- Microarray dataset
- US arrests dataset
- Harry Potter
- Attitudes

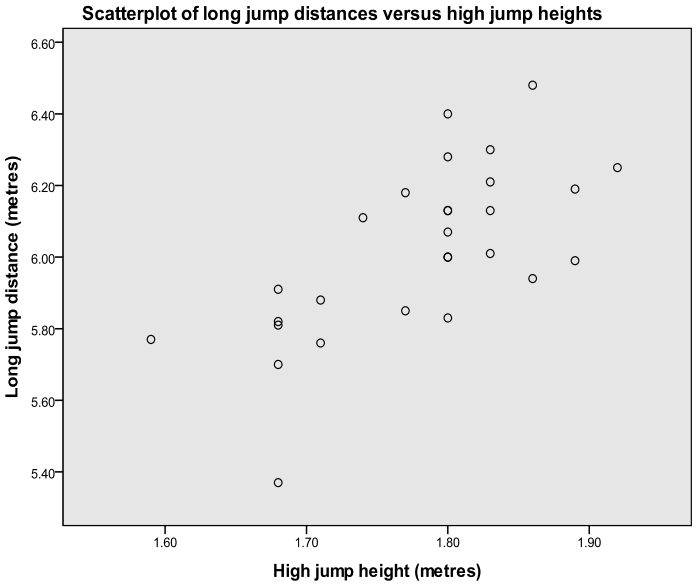
1.5 First steps: A look at the heptathlon dataset

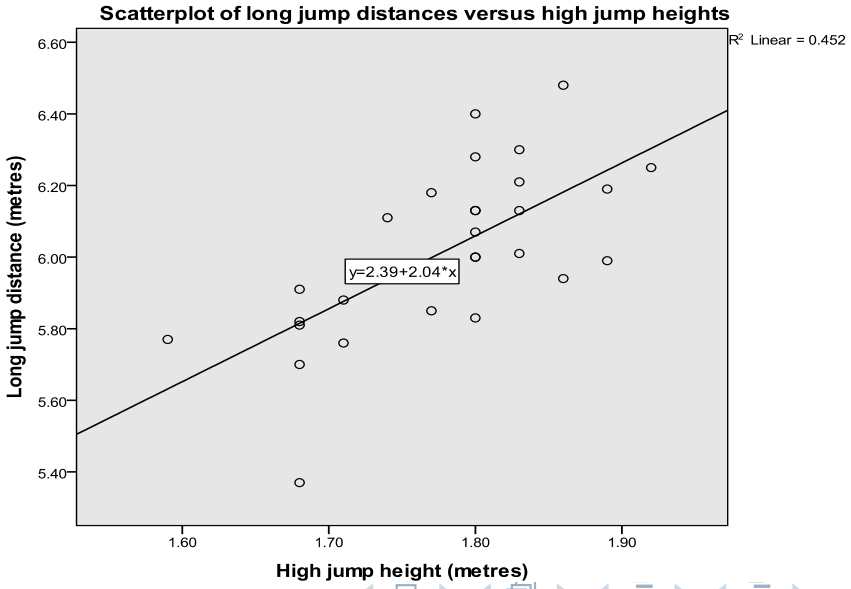
Name	100M hurdles	High jump	Shot put	200M	Long jump	Javelin	800M
Jessica Ennis	12.54	1.86	14.28	22.83	6.48	47.49	128.65
Lilli Schwarzkopf	13.26	1.83	14.77	24.77	6.30	51.73	130.50
Austra Skujyte	14.00	1.92	17.31	25.43	6.25	51.13	140.59
Antoinette Nana Djimou Ida	12.96	1.80	14.26	24.72	6.13	55.87	133.62
Jessica Zelinka	12.65	1.68	14.81	23.32	5.91	45.75	129.15
Kristina Savitskaya	13.37	1.83	14.77	24.46	6.21	43.70	132.27
Laura Ikauniece	13.71	1.83	12.64	24.16	6.13	51.27	132.13
Hanna Melnychenko	13.32	1.80	12.96	24.09	6.40	43.86	132.90
Brianne Theisen	13.30	1.83	12.89	24.35	6.01	46.47	129.27
Dafne Schippers	13.48	1.80	13.67	22.83	6.28	36.63	135.52
Nadine Broersen	13.64	1.86	13.57	25.13	5.94	51.98	136.98
Jessica Samuelsson	13.58	1.77	14.18	24.25	6.18	42.02	131.31
Katarina Johnson-Thompson	13.48	1.89	11.32	23.73	6.19	38.37	130.76
Sharon Day	13.57	1.77	14.28	24.36	5.85	43.90	131.31
Yana Maksimava	13.97	1.89	14.09	25.43	5.99	42.33	133.37
Eliska Klucinova	14.01	1.80	12.93	25.00	6.13	45.65	136.08
Ellen Sprunger	13.35	1.71	12.62	23.59	5.88	45.63	137.54
Olga Kurban	13.46	1.80	13.71	23.88	5.83	40.36	139.82
Marisa De Aniceto	13.74	1.71	13.09	25.26	5.76	51.98	136.20
Gyorgyi Farkas	14.33	1.80	13.55	25.72	6.07	46.52	137.83
Grit Sadeiko	13.50	1.74	12.43	24.25	6.11	44.12	143.01
Sofia Ifadidou	13.82	1.68	12.96	25.91	5.81	56.96	142.03
Ivona Dadic	14.58	1.80	12.19	24.29	6.00	41.82	135.90
Sarah Cowley	13.95	1.80	12.37	25.60	6.00	41.90	139.01
Louise Hazel	13.48	1.59	12.81	24.48	5.77	47.38	138.78
Ida Marcussen	14.08	1.68	14.26	25.15	5.82	42.26	133.62
Chantae McMillan	13.49	1.68	14.92	25.25	5.37	49.78	160.55
Julia Machtig	14.54	1.68	14.99	25.38	4.06	44.40	140.34
Irina Karpova	14.21	1.68	11.68	25.42	5.70	35.75	148.93

1.5 First steps: A look at the heptathlon dataset









1.5 First steps: A look at the heptathlon dataset

Initial analysis

- Long jump distance $= 0.64 + 2.99 \times$ High jump height $+ \epsilon$
- $R^2 = 0.302$
- Significant slope (p-value $= 0.2\%$)

Removal of Julia Machtig

- Long jump distance $= 2.39 + 2.04 \times$ High jump height $+ \epsilon$
- $R^2 = 0.452$
- Significant slope (p-value $= 0.5\%$)
- 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:

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

1.5 First steps: A look at the heptathlon dataset

Initial analysis

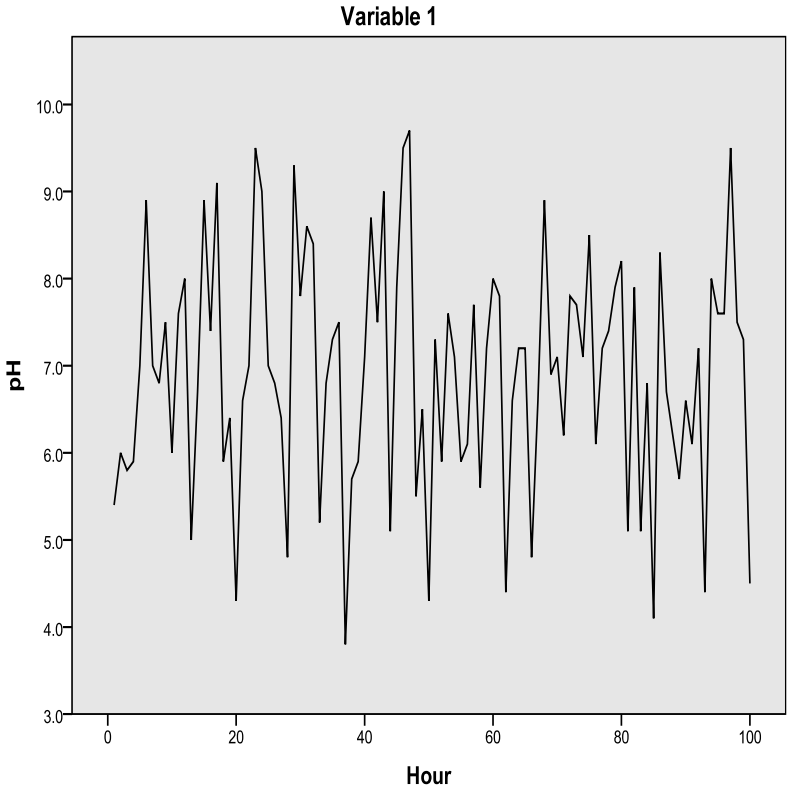
- Long jump distance $= 0.64 + 2.99 \times$ High jump height $+ \epsilon$
- $R^2 = 0.302$
- Significant slope (p-value $= 0.2\%$)

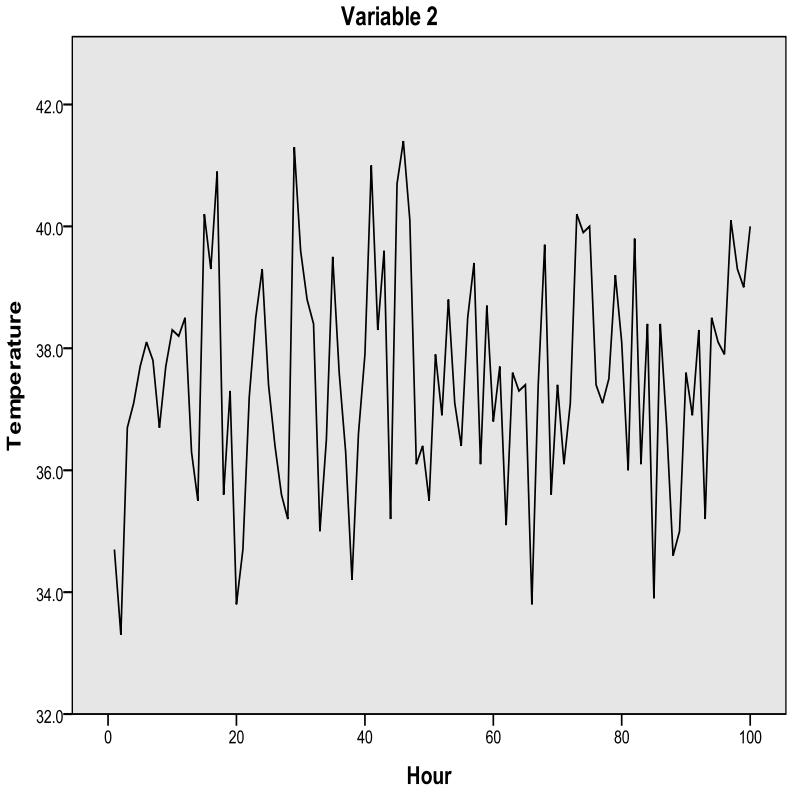
Removal of Julia Machtig

- Long jump distance $= 2.39 + 2.04 \times$ High jump height $+ \epsilon$
- $R^2 = 0.452$
- Significant slope (p-value $= 0.5\%$)
- 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:

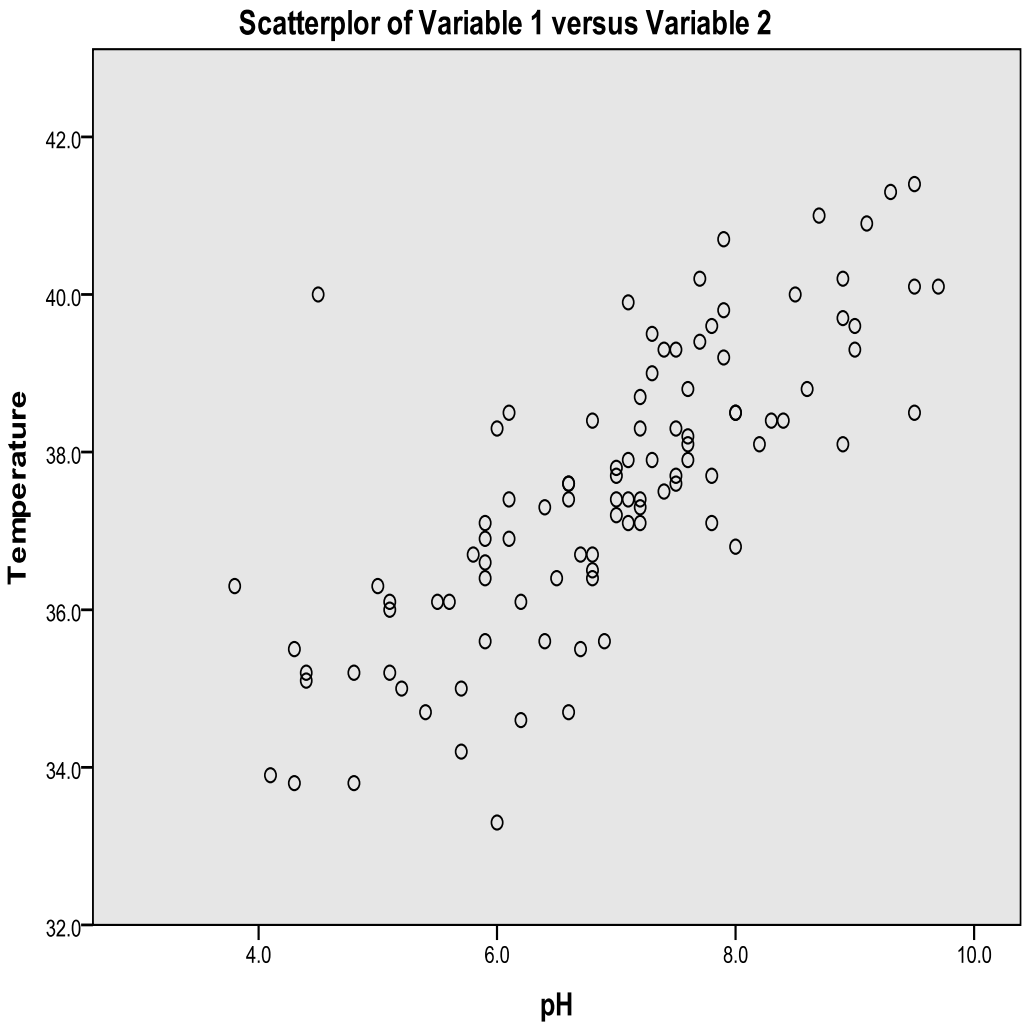
$$\text{Long jump distance} = 2.39 + 2.04 \times 1.75 = 5.96 \text{ metres.}$$

1.5.2 An example of a hidden outlier





1.5.2 An example of a hidden outlier



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?

If there *are* 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**.

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?

If there *are* 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**.

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?

If there *are* 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**.

Chapter 2

Exploration

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!), **plot it**!

- Univariate datasets: histograms, bar charts/pie charts, ...
- Bivariate datasets: histograms for the 'margins', **scatterplots** for the dependence
- Multivariate datasets...?

2.1 Simple methods for summarising multivariate datasets

Before embarking on any rigorous analysis of a multivariate dataset (or *any* dataset, for that matter!), **plot it**!

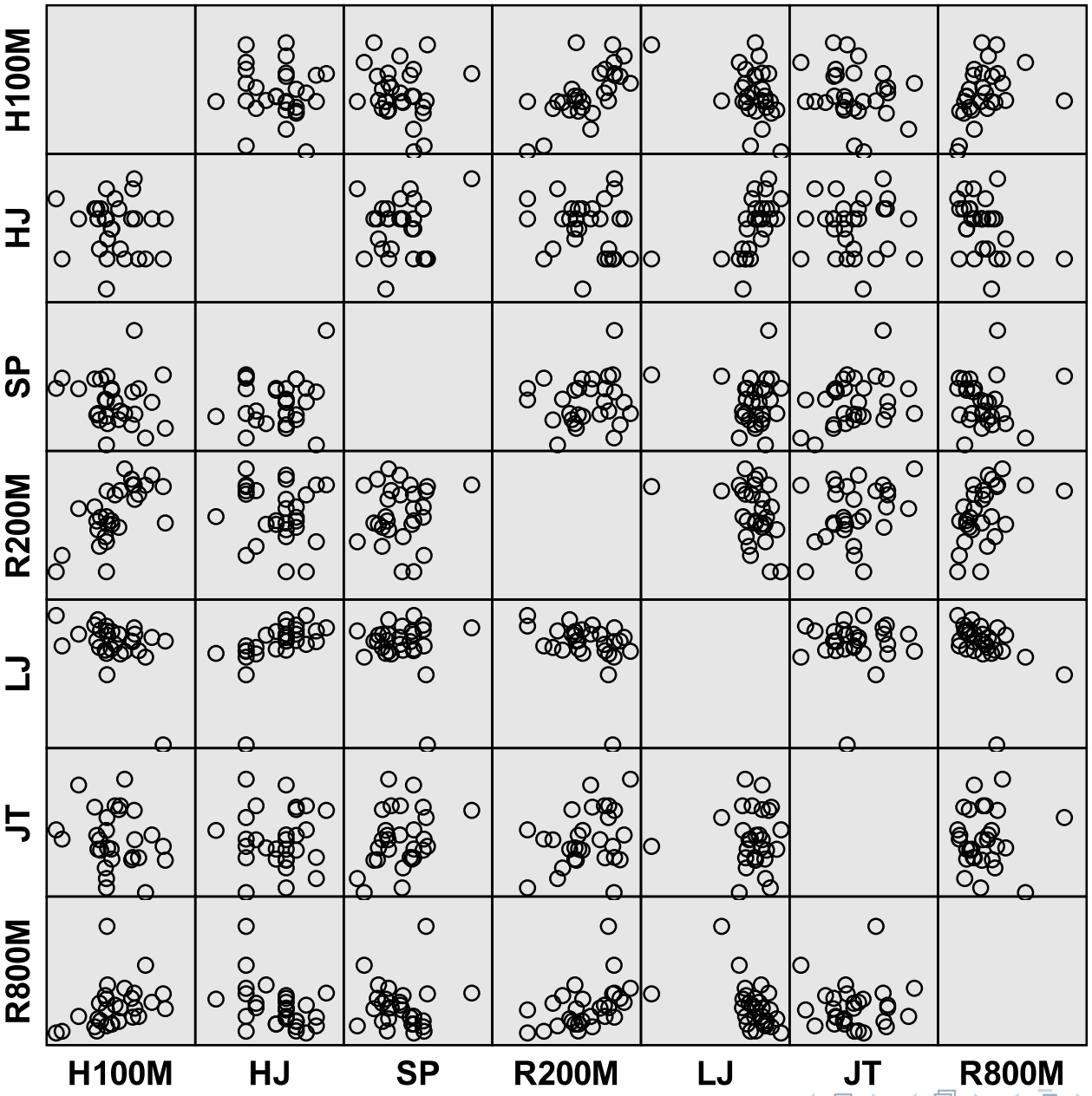
- Univariate datasets: histograms, bar charts/pie charts, ...
- Bivariate datasets: histograms for the 'margins', **scatterplots** for the dependence
- Multivariate datasets...?

2.1 Simple methods for summarising multivariate datasets

Before embarking on any rigorous analysis of a multivariate dataset (or *any* dataset, for that matter!), **plot it**!

- Univariate datasets: histograms, bar charts/pie charts, ...
- Bivariate datasets: histograms for the 'margins', **scatterplots** for the dependence
- Multivariate datasets...?

2.1.1 Scatterplot matrices



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?

We might expect that athletes who are good at one running event might also be good at another – hence the positive association between times in the 200 metres and 800 metres, as their performances might be on par in the two events.

Again, we might expect athletes who are good at running might also be good at jumping (perhaps leg length is key here?). However, we would get a negative association between performances, as being good at running would correspond to a small time, but being good at jumping would correspond to a large distance.

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?

  We might expect that athletes who are good at one running event might also be good at another – hence the positive association between times in the 200 metres and 800 metres, as their performances might be on par in the two events.

  Again, we might expect athletes who are good at running might also be good at jumping (perhaps leg length is key here?). However, we would get a negative association between performances, as being good at running would correspond to a small time, but being good at jumping would correspond to a large distance.

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?

  We might expect that athletes who are good at one running event might also be good at another – hence the positive association between times in the 200 metres and 800 metres, as their performances might be on par in the two events.

  Again, we might expect athletes who are good at running might also be good at jumping (perhaps leg length is key here?). However, we would get a negative association between performances, as being good at running would correspond to a small time, but being good at jumping would correspond to a large distance.

Questions

- What can you say about performances in the shot put and the javelin throw?

  Positive correlation – as distances thrown in one event increase, generally they also increase in the other event.

- What can you say about performances in shot put and the 200 metres?

  Not much going on here – points just seem randomly scattered, so no real (noticeable) association.

Questions

- What can you say about performances in the shot put and the javelin throw?

  Positive correlation – as distances thrown in one event increase, generally they also increase in the other event.

- What can you say about performances in shot put and the 200 metres?

  Not much going on here – points just seem randomly scattered, so no real (noticeable) association.

Questions

- What can you say about performances in the shot put and the javelin throw?

  Positive correlation – as distances thrown in one event increase, generally they also increase in the other event.

- What can you say about performances in shot put and the 200 metres?

  Not much going on here – points just seem randomly scattered, so no real (noticeable) association.

Questions

- What can you say about performances in the shot put and the javelin throw?

  Positive correlation – as distances thrown in one event increase, generally they also increase in the other event.

- What can you say about performances in shot put and the 200 metres?

  Not much going on here – points just seem randomly scattered, so no real (noticeable) association.

Questions

- Why is the leading diagonal of cells empty?

  These cells correspond to correlations between each variable and itself – we wouldn't be interested in that!

- Why are plots repeated above/below the leading diagonal?

  They show the same thing. For example, the cell in (row 2, column 1) shows the association between performances in the high jump and the 100 metres hurdles. The cell in (row 1, column 2) shows the association between performances in the 100 metres hurdles and the high jump! Basically, we don't need the plots above the leading diagonal.

Questions

- Why is the leading diagonal of cells empty?

  These cells correspond to correlations between each variable and itself – we wouldn't be interested in that!

- Why are plots repeated above/below the leading diagonal?

  They show the same thing. For example, the cell in (row 2, column 1) shows the association between performances in the high jump and the 100 metres hurdles. The cell in (row 1, column 2) shows the association between performances in the 100 metres hurdles and the high jump! Basically, we don't need the plots above the leading diagonal.

Questions

- Why is the leading diagonal of cells empty?

  These cells correspond to correlations between each variable and itself – we wouldn't be interested in that!

- Why are plots repeated above/below the leading diagonal?

  They show the same thing. For example, the cell in (row 2, column 1) shows the association between performances in the high jump and the 100 metres hurdles. The cell in (row 1, column 2) shows the association between performances in the 100 metres hurdles and the high jump! Basically, we don't need the plots above the leading diagonal.

Questions

- Why is the leading diagonal of cells empty?

  These cells correspond to correlations between each variable and itself – we wouldn't be interested in that!

- Why are plots repeated above/below the leading diagonal?

  They show the same thing. For example, the cell in (row 2, column 1) shows the association between performances in the high jump and the 100 metres hurdles. The cell in (row 1, column 2) shows the association between performances in the 100 metres hurdles and the high jump! Basically, we don't need the plots above the leading diagonal.

2.1.2 Correlation matrices

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).

Questions

- What does the correlation matrix tell you about performances in the long jump and the high jump?
  - We have a positive correlation of 0.550 – as long jump distances increase, so too do high jump heights
  - What's more, this is a statistically significant correlation as the p-value is 0.2%
  - We usually want to know if we have significance at the 5% level or the 1% level – here, we have significance at the 1% level (as 0.2% is less than 1%; if the p-value lies between 1% and 5%, we say we have significance at the 5% level)

Questions

- What does the correlation matrix tell you about performances in the long jump and the high jump?
  - We have a positive correlation of 0.550 – as long jump distances increase, so too do high jump heights
  - What's more, this is a statistically significant correlation as the p-value is 0.2%
  - We usually want to know if we have significance at the 5% level or the 1% level – here, we have significance at the 1% level (as 0.2% is less than 1%; if the p-value lies between 1% and 5%, we say we have significance at the 5% level)

Questions

- What about the dependence between performances in the 800 metres and the long jump?
  - We have a negative correlation of -0.449 – as long jump distances increase, speeds in the 800 metres decrease
  - What's more, this is a statistically significant correlation as the p-value is 1.5%
  - We have significance at the 5% level

Questions

- What about the dependence between performances in the 800 metres and the long jump?
  - We have a negative correlation of –0.449 – as long jump distances increase, speeds in the 800 metres decrease
  - What's more, this is a statistically significant correlation as the p-value is 1.5%
  - We have significance at the 5% level

Questions

- What about the dependence between performances in the shot put and 200 metres?
  - We have a positive correlation of 0.108 – very close to zero
  - This is *not* statistically significant as the p-value is 57.6%
  - Recall that a p-value less than 5% is usually required for statistical significance

Questions

- What about the dependence between performances in the shot put and 200 metres?
  - We have a positive correlation of 0.108 – very close to zero
  - This is *not* statistically significant as the p-value is 57.6%
  - Recall that a p-value less than 5% is usually required for statistical significance

Questions

- Why do all cells along the leading diagonal have correlation coefficients of exactly 1?

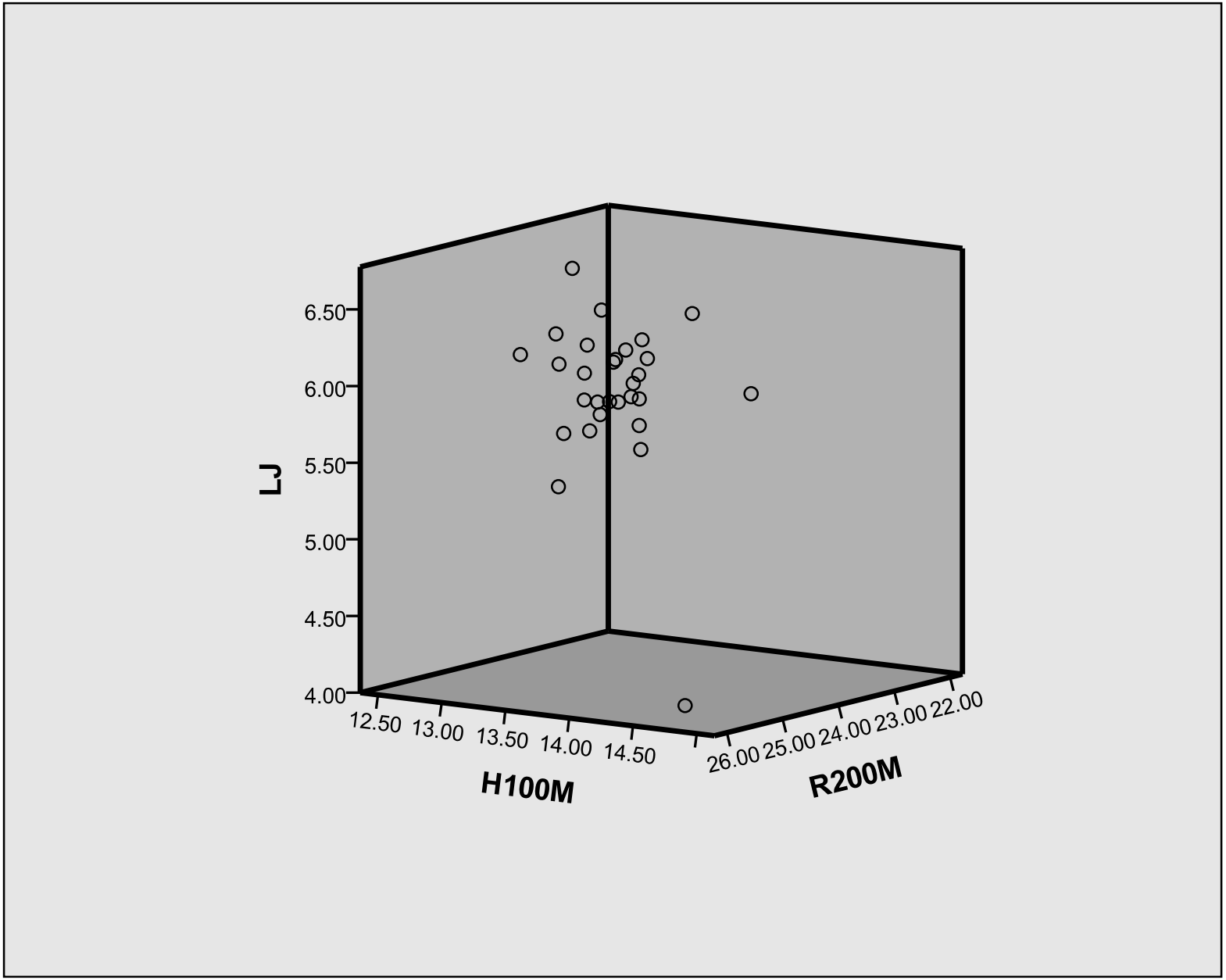
  These cells correspond to associations between each variable and itself – obviously, each variable is perfectly correlated with itself, hence a correlation of 1!

Questions

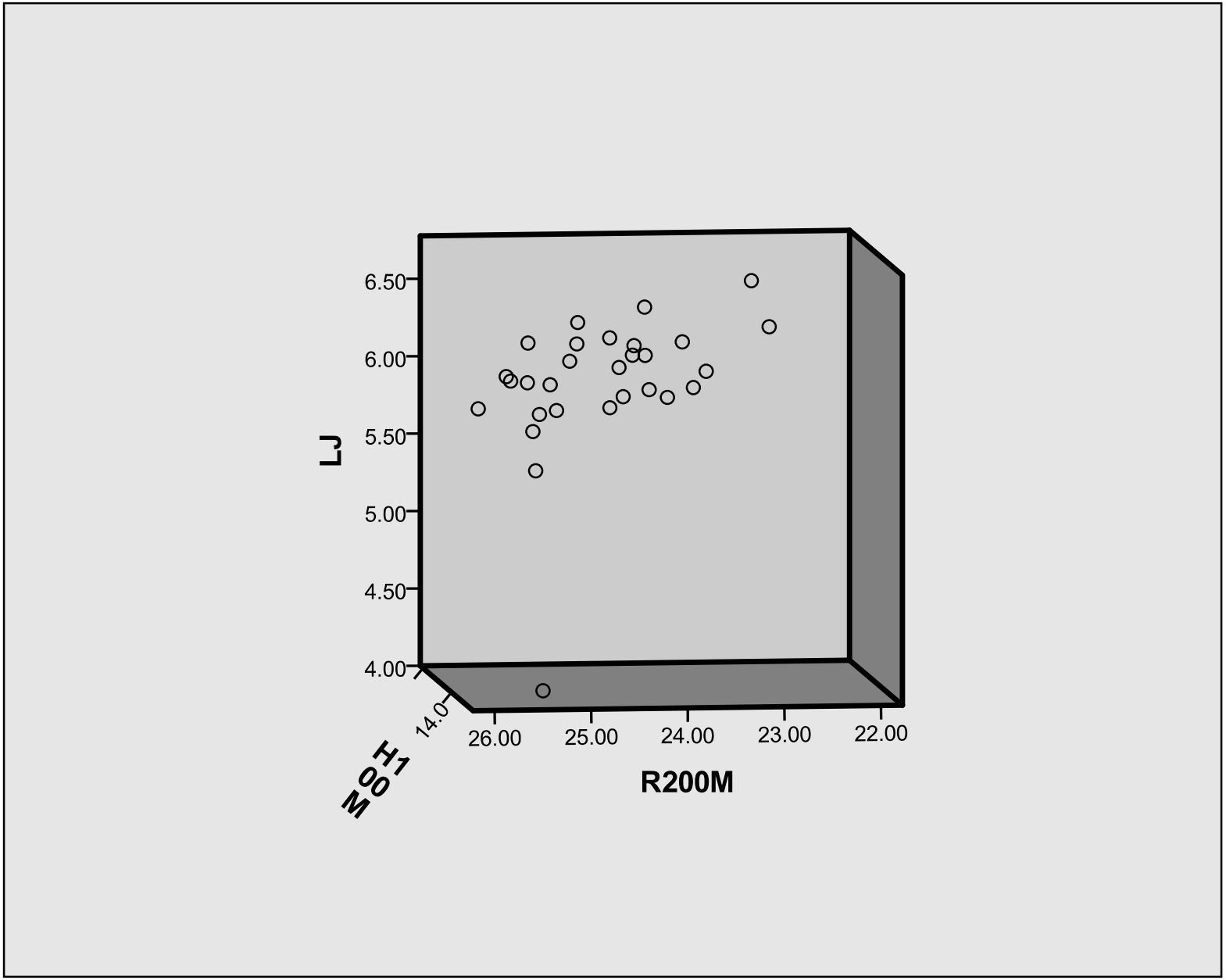
- Why do all cells along the leading diagonal have correlation coefficients of exactly 1?

  These cells correspond to associations between each variable and itself – obviously, each variable is perfectly correlated with itself, hence a correlation of 1!

2.1.3 3-D Scatterplots



2.1.3 3-D Scatterplots



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

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

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**"
- 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

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**"
- 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

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**"
- 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

2.2 Principal Components Analysis

- The main aim is to find new variables which are linear combinations of the original variables
- The new variables are such that a few of them will explain most of the variability in the data, hence effectively **reducing the number of variables**
- Unlike **stepwise regression**, PCA includes information on *all* variables through these new variables

2.2 Principal Components Analysis

- The main aim is to find new variables which are linear combinations of the original variables
- The new variables are such that a few of them will explain most of the variability in the data, hence effectively **reducing the number of variables**
- Unlike **stepwise regression**, PCA includes information on *all* variables through these new variables

2.2 Principal Components Analysis

- The main aim is to find new variables which are linear combinations of the original variables
- The new variables are such that a few of them will explain most of the variability in the data, hence effectively **reducing the number of variables**
- Unlike **stepwise regression**, PCA includes information on *all* variables through these new variables

2.2.1 PCA of the heptathlon dataset

Recall we have seven variables, which we will now label:

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

2.2.1 PCA of the heptathlon dataset

The aim of a PCA is to **construct new variables**:

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

2.2.1 PCA of the heptathlon dataset

Why would we want to do all of this?

- 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 on 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.

2.2.1 PCA of the heptathlon dataset

Why would we want to do all of this?

- 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 on 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.

2.2.1 PCA of the heptathlon dataset

Why would we want to do all of this?

- 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 on 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.

2.2.1 PCA of the heptathlon dataset

Why would we want to do all of this?

- 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 on 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.

2.2.1 PCA of the heptathlon dataset

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 $X_1, X_2, \ldots, X_7$.

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.

2.2.1 PCA of the heptathlon dataset

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 $X_1, X_2, \ldots, X_7$.

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.

2.2.1 PCA of the heptathlon dataset

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 $X_1, X_2, \ldots, X_7$.

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.

2.2.1 PCA of the heptathlon dataset

- In the jumping events large numbers are indicative of good performances
- 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

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

2.2.1 PCA of the heptathlon dataset

- In the jumping events large numbers are indicative of good performances
- 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

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

2.2.1 PCA of the heptathlon dataset

- In the jumping events large numbers are indicative of good performances
- 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

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

2.2.1 PCA of the heptathlon dataset

- In the jumping events large numbers are indicative of good performances
- 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

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

2.2.1 PCA of the heptathlon dataset

- In the jumping events large numbers are indicative of good performances
- 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

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

2.2.1 Interpretation of SPSS output

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.

2.2.1 Interpretation of SPSS output

From Figure 2.4:

$$
\begin{aligned}
Y_1 &= 0.275X_1^* + 0.220X_2 - 0.001X_3 + 0.291X_4^* + 0.301X_5 - 0.027X_6 + 0.289X_7^*, \\
Y_2 &= 0.153X_1^* + 0.139X_2 + 0.515X_3 - 0.223X_4^* + 0.019X_5 + 0.562X_6 + 0.008X_7^*, \\
Y_3 &= 0.512X_1^* - 0.616X_2 - 0.042X_3 + 0.372X_4^* - 0.238X_5 + 0.132X_6 - 0.133X_7^*,
\end{aligned}
$$

But what do these components tell us?

2.2.1 Interpretation of SPSS output

From Figure 2.4:

$$
\begin{aligned}
Y_1 &= 0.275X_1^* + 0.220X_2 - 0.001X_3 + 0.291X_4^* + 0.301X_5 - 0.027X_6 + 0.289X_7^*, \\
Y_2 &= 0.153X_1^* + 0.139X_2 + 0.515X_3 - 0.223X_4^* + 0.019X_5 + 0.562X_6 + 0.008X_7^*, \\
Y_3 &= 0.512X_1^* - 0.616X_2 - 0.042X_3 + 0.372X_4^* - 0.238X_5 + 0.132X_6 - 0.133X_7^*,
\end{aligned}
$$

But what do these components tell us?

2.2.1 Interpretation of SPSS output

- Each component is a function of all of our original X variables, and in each component the X variables have a 'coefficient'
- 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 are both very small relative to the coefficients for the running/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**

2.2.1 Interpretation of SPSS output

- Each component is a function of all of our original X variables, and in each component the X variables have a 'coefficient'
- 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 are both very small relative to the coefficients for the running/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**

2.2.1 Interpretation of SPSS output

- **Second principal component:**
  - In terms of the coefficients, it is clear that the throwing events dominate this principal component, both having similar coefficients that are larger than the others (relatively speaking)
  - Therefore, this principal component might be interpreted as an average measure of ability in the throwing events

2.2.1 Interpretation of SPSS output

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared ..	
	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

Component	Extraction Sums ..
	Cumulative %

2.2.1 Interpretation of SPSS output

Figure 2.5 also tells us how much of the variability in our original multivariate dataset is captured by our first three principal components:

- First PC (Y_1): 37.168%
- Second PC (Y_2): 21.204%
- Third PC (Y_3): 16.366%
- So 74.739% of the total variation in our original dataset being captured by the first three principal components. Pretty good!

2.2.1 Calculation of principal component scores

We can now calculate principal component **scores** for each athlete.

For example, for Jessica Ennis we have

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

2.2.1 Calculation of principal component scores

We can now calculate principal component **scores** for each athlete.

For example, for Jessica Ennis we have

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

2.2.1 Calculation of principal component scores

We also have

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

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).

$$
\begin{aligned}
Y_1 &= 0.275 \times 0.37 + 0.220 \times 1.68 - 0.001 \times 11.68 + 0.291 \times 0.49 + 0.301 \times 5.70 \\
&\quad -0.027 \times 35.75 + 0.289 \times 11.62 \\
&= 4.71089 \quad \text{and}
\end{aligned}
$$

$$
\begin{aligned}
Y_2 &= 0.153 \times 0.37 + 0.139 \times 1.68 + 0.515 \times 11.68 - 0.223 \times 0.49 + 0.019 \times 5.70 \\
&\quad +0.562 \times 35.75 + 0.008 \times 11.62 \\
&= 26.48882.
\end{aligned}
$$

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).

$$\begin{aligned} Y_1 &= 0.275 \times 0.37 + 0.220 \times 1.68 - 0.001 \times 11.68 + 0.291 \times 0.49 + 0.301 \times 5.70 \\ &\qquad -0.027 \times 35.75 + 0.289 \times 11.62 \\ &= 4.71089 \quad \text{and} \end{aligned}$$

$$\begin{aligned} Y_2 &= 0.153 \times 0.37 + 0.139 \times 1.68 + 0.515 \times 11.68 - 0.223 \times 0.49 + 0.019 \times 5.70 \\ &\qquad +0.562 \times 35.75 + 0.008 \times 11.62 \\ &= 26.48882. \end{aligned}$$

2.2.1 Calculation of principal component scores

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.

2.2.1 Calculation of principal component scores

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.

2.2.1 Calculation of principal component scores

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.

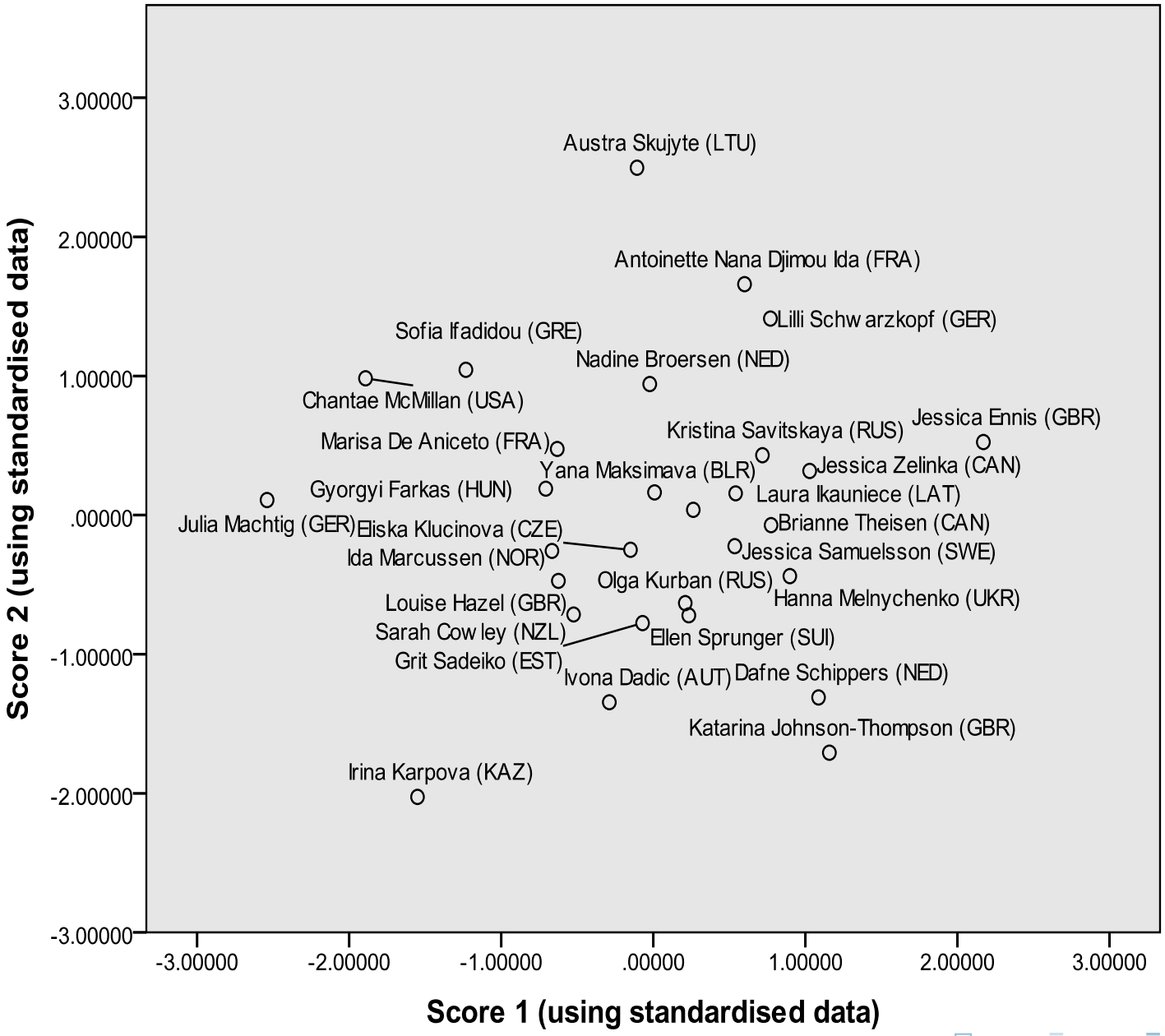
- We could have easily recorded times in the 800 metres running event in milliseconds, giving us even larger values on variable X_7
- If a certain athlete performed particularly well in the 800 metres, but not so well in all the other events, her principal component scores may still be very high

2.2.1 Calculation of principal component scores

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.

2.2.1 Calculation of principal component scores



2.2.2 PCA of the microarray dataset

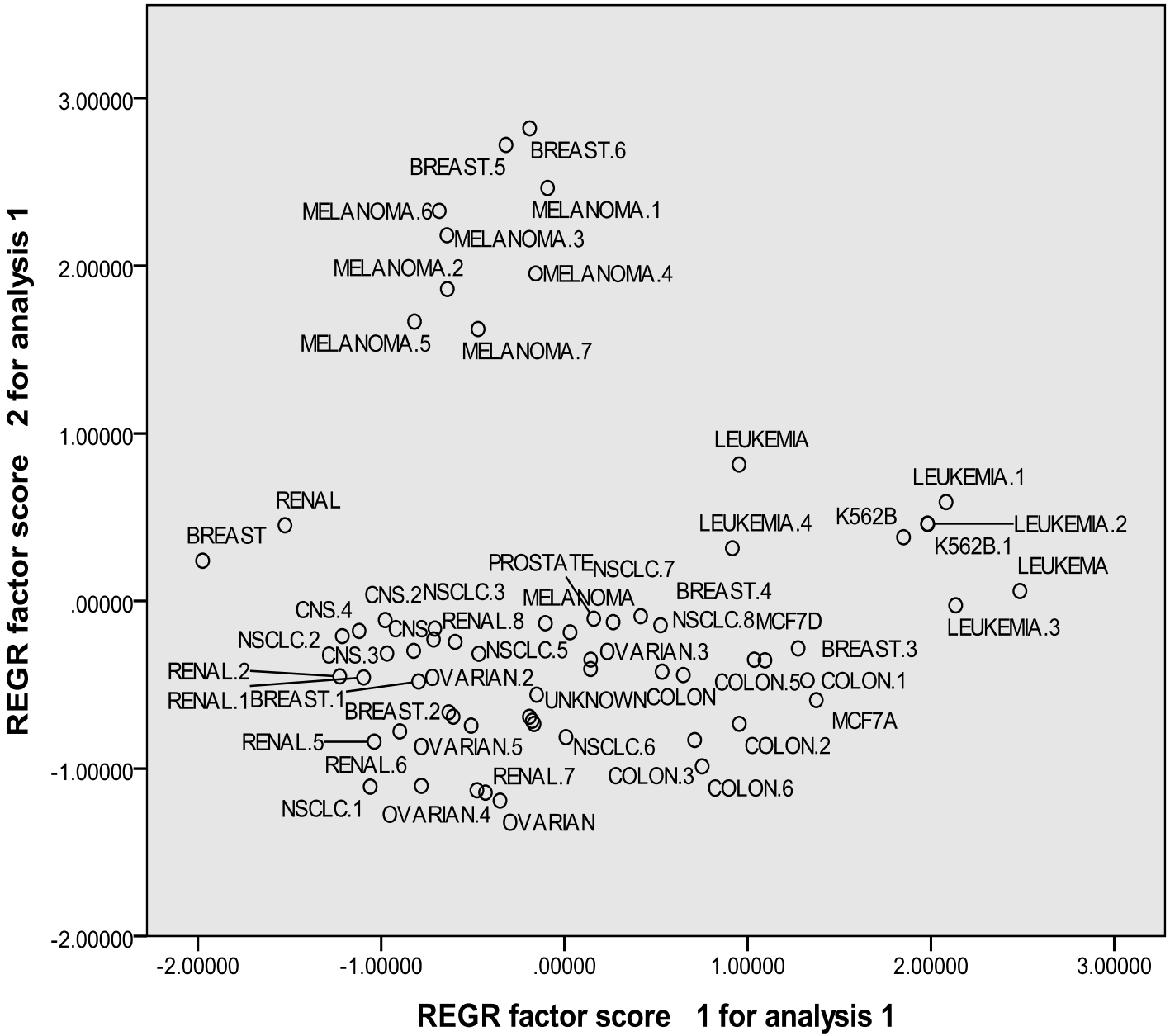
To illustrate the power of PCA in the analysis of 'big data', we return to the motivating example presented in Section 1.4.3.

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 really is a **big dataset**!

- Scatterplot matrix: 23 million plots
- No simple way of visualising these data

2.2.2 PCA of the microarray dataset



Chapter 3

Modelling

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.

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.

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.

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.

3.1 Introduction

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.

3.1 Introduction

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.

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.

3.2 Simple exploration

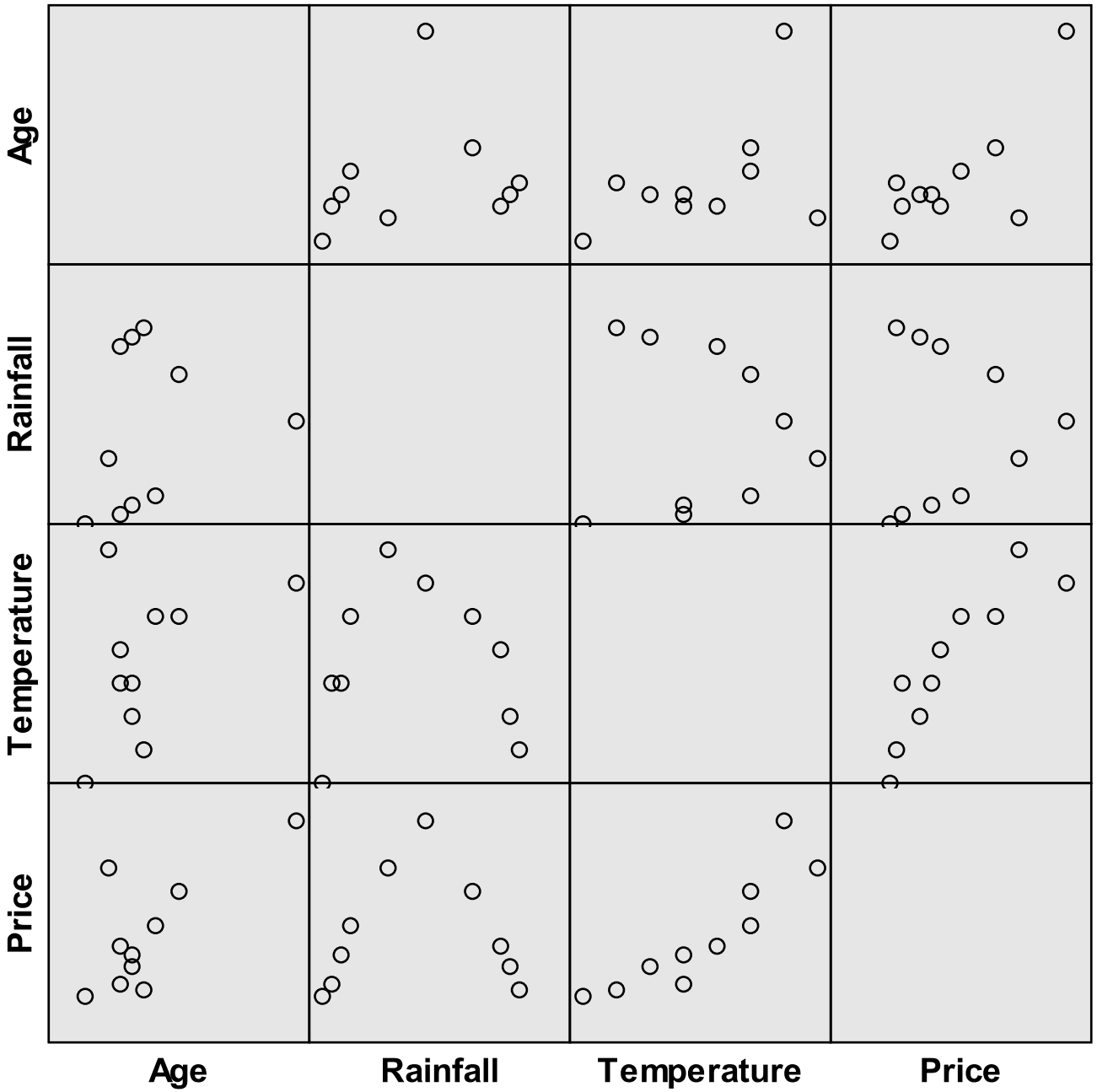
The first step of any multiple regression analysis is **simple exploration**, usually using the methods discussed in Section 2.1.

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.

3.2 Simple exploration



3.2 Simple exploration

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

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

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

All variables in the wine sales dataset are represented:

- The **response variable** (selling price) is on the left-hand-side
- The **predictor variables** (age, rainfall, temperature and New Zealand indicator) are 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.

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

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

All variables in the wine sales dataset are represented:

- The **response variable** (selling price) is on the left-hand-side
- The **predictor variables** (age, rainfall, temperature and New Zealand indicator) are 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.

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

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

All variables in the wine sales dataset are represented:

- The **response variable** (selling price) is on the left-hand-side
- The **predictor variables** (age, rainfall, temperature and New Zealand indicator) are 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.

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

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

All variables in the wine sales dataset are represented:

- The **response variable** (selling price) is on the left-hand-side
- The **predictor variables** (age, rainfall, temperature and New Zealand indicator) are 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.

3.3 Building a linear model

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 Building a linear model

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 Building a linear model

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

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

This gives:

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

3.3.1 Fitting the full model

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

This gives:

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

Question

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

- The age coefficient is positive; as the age of wine increases, so too does the price. Notice that this coefficient is statistically significant, as the associated p-value is less than 5% (it is 3.2%)
- The rainfall coefficient is negative, although it is quite close to zero. In fact, this coefficient is *not* statistically significant as the p-value is *not* less than 5% (it is 77.4%)
- The temperature coefficient is positive; as the average daytime temperature during the grape-growing season increases, so too does the selling price. This coefficient is statistically significant, as the associated p-value is less than 5% (it is 1.7%)
- The New Zealand coefficient is statistically significant as the p-value is less than 5% (it is 4.4%)

Question

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

- The age coefficient is positive; as the age of wine increases, so too does the price. Notice that this coefficient is statistically significant, as the associated p-value is less than 5% (it is 3.2%)
- The rainfall coefficient is negative, although it is quite close to zero. In fact, this coefficient is *not* statistically significant as the p-value is *not* less than 5% (it is 77.4%)
- The temperature coefficient is positive; as the average daytime temperature during the grape-growing season increases, so too does the selling price. This coefficient is statistically significant, as the associated p-value is less than 5% (it is 1.7%)
- The New Zealand coefficient is statistically significant as the p-value is less than 5% (it is 4.4%)

Question

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

- The age coefficient is positive; as the age of wine increases, so too does the price. Notice that this coefficient is statistically significant, as the associated p-value is less than 5% (it is 3.2%)
- The rainfall coefficient is negative, although it is quite close to zero. In fact, this coefficient is *not* statistically significant as the p-value is *not* less than 5% (it is 77.4%)
- The temperature coefficient is positive; as the average daytime temperature during the grape-growing season increases, so too does the selling price. This coefficient is statistically significant, as the associated p-value is less than 5% (it is 1.7%)
- The New Zealand coefficient is statistically significant as the p-value is less than 5% (it is 4.4%)

Question

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

- The age coefficient is positive; as the age of wine increases, so too does the price. Notice that this coefficient is statistically significant, as the associated p-value is less than 5% (it is 3.2%)
- The rainfall coefficient is negative, although it is quite close to zero. In fact, this coefficient is *not* statistically significant as the p-value is *not* less than 5% (it is 77.4%)
- The temperature coefficient is positive; as the average daytime temperature during the grape-growing season increases, so too does the selling price. This coefficient is statistically significant, as the associated p-value is less than 5% (it is 1.7%)
- The New Zealand coefficient is statistically significant as the p-value is less than 5% (it is 4.4%)

3.3.1 Fitting the full model

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 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!

3.3.1 Fitting the full model

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 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!

3.3.1 Fitting the full model

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.

3.3.1 Fitting the full model

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.

3.3.1 Fitting the full model

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

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	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

3.3.2 Backwards elimination

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	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

3.3.2 Backwards elimination

- 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 are all less than 5%
- 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

3.3.2 Backwards elimination

- 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 are all less than 5%
- 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

3.3.2 Backwards elimination

- 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 are all less than 5%
- 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

3.3.2 Backwards elimination

- 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 are all less than 5%
- 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

3.3.2 Backwards elimination

- Our fitted model is now:

$$Y = -13.657 + 0.613X_1 + 1.052X_3 + 3.731X_4 + \epsilon.$$

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

3.3.2 Backwards elimination

- Our fitted model is now:

$$Y = -13.657 + 0.613X_1 + 1.052X_3 + 3.731X_4 + \epsilon.$$

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

3.3.2 Backwards elimination

In this Section we have demonstrated the process of **backwards elimination**.

After having fitted the **full model** we then eliminate variables to see if they are actually needed in the model.

How do we know which variables to eliminate?

3.3.2 Backwards elimination

In this Section we have demonstrated the process of **backwards elimination**.

After having fitted the **full model** we then eliminate variables to see if they are actually needed in the model.

How do we know which variables to eliminate?

3.3.2 Backwards elimination

In this Section we have demonstrated the process of **backwards elimination**.

After having fitted the **full model** we then eliminate variables to see if they are actually needed in the model.

How do we know which variables to eliminate?

3.3.2 Backwards elimination

1. Fit the **full model**, and make a note of the R^2 value
2. Look for variables with **insignificant coefficients** – coefficients that have p-values greater than 5%. What if there was more than one such coefficient?
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 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
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 obtained the **final model**

3.3.2 Backwards elimination

1. Fit the **full model**, and make a note of the R^2 value
2. Look for variables with **insignificant coefficients** – coefficients that have p-values greater than 5%. What if there was more than one such coefficient?
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 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
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 obtained the **final model**

3.3.2 Backwards elimination

1. Fit the **full model**, and make a note of the R^2 value
2. Look for variables with **insignificant coefficients** – coefficients that have p-values greater than 5%. What if there was more than one such coefficient?
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 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
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 obtained the **final model**

3.3.2 Backwards elimination

1. Fit the **full model**, and make a note of the R^2 value
2. Look for variables with **insignificant coefficients** – coefficients that have p-values greater than 5%. What if there was more than one such coefficient?
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 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
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 obtained the **final model**

3.3.2 Backwards elimination

1. Fit the **full model**, and make a note of the R^2 value
2. Look for variables with **insignificant coefficients** – coefficients that have p-values greater than 5%. What if there was more than one such coefficient?
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 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
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 obtained the **final model**

3.3.3 Using your model

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?

We have

$$Y = -13.657 + 0.613X_1 + 1.052X_3 + 3.731X_4 + \epsilon.$$

We have $X_1 = 6$, $X_2 = 115$, $X_3 = 17$ and $X_4 = 1$, although X_2 is no longer needed. Subbing into the regression equation:

$$Y = -13.657 + 0.613 \times 6 + 1.052 \times 17 + 3.731 \times 1 = 11.636,$$

in other words, around £11.64 per bottle.

3.3.3 Using your model

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?

We have

$$Y = -13.657 + 0.613X_1 + 1.052X_3 + 3.731X_4 + \epsilon.$$

We have $X_1 = 6$, $X_2 = 115$, $X_3 = 17$ and $X_4 = 1$, although X_2 is no longer needed. Subbing into the regression equation:

$$Y = -13.657 + 0.613 \times 6 + 1.052 \times 17 + 3.731 \times 1 = 11.636,$$

in other words, around £11.64 per bottle.

3.3.3 Using your model

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?

We have

$$Y = -13.657 + 0.613X_1 + 1.052X_3 + 3.731X_4 + \epsilon.$$

We have $X_1 = 6$, $X_2 = 115$, $X_3 = 17$ and $X_4 = 1$, although X_2 is no longer needed. Subbing into the regression equation:

$$Y = -13.657 + 0.613 \times 6 + 1.052 \times 17 + 3.731 \times 1 = 11.636,$$

in other words, around £11.64 per bottle.

3.3.3 Using your model

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?

We have

$$Y = -13.657 + 0.613X_1 + 1.052X_3 + 3.731X_4 + \epsilon.$$

We have $X_1 = 6$, $X_2 = 115$, $X_3 = 17$ and $X_4 = 1$, although X_2 is no longer needed. Subbing into the regression equation:

$$Y = -13.657 + 0.613 \times 6 + 1.052 \times 17 + 3.731 \times 1 = 11.636,$$

in other words, around £11.64 per bottle.

3.3.4 Checking assumptions

There are various **assumptions** underpinning a multiple linear regression model that must 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, is **Normally distributed** with **constant variance** (*homoscedasticity*).

This is a rather complex concept, but one that can be quite easy to check in SPSS.

3.3.4 Checking assumptions

There are various **assumptions** underpinning a multiple linear regression model that must 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, is **Normally distributed** with **constant variance** (*homoscedasticity*).

This is a rather complex concept, but one that can be quite easy to check in SPSS.

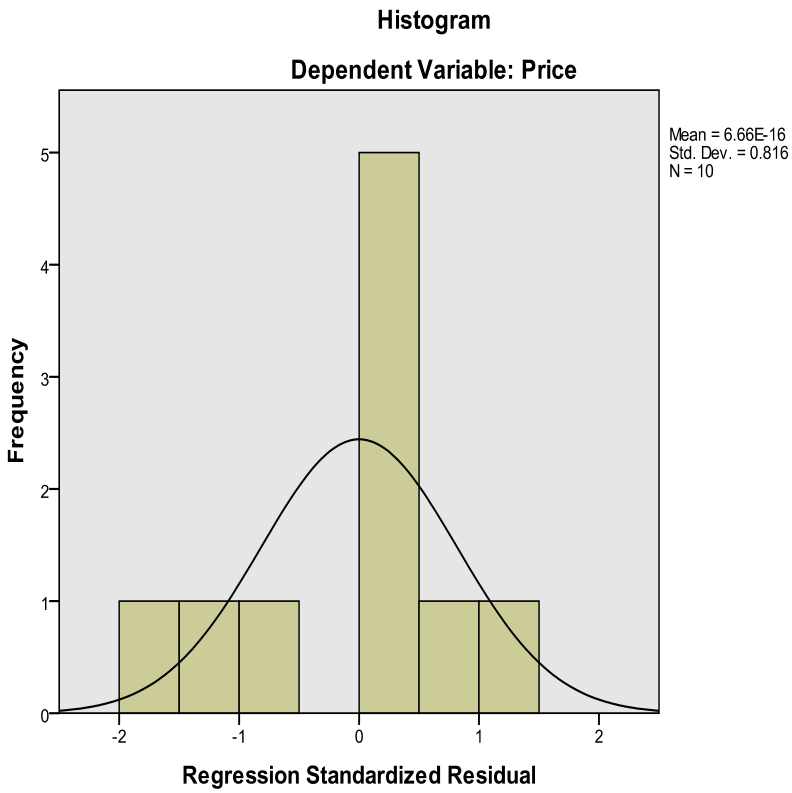
3.3.4 Checking assumptions

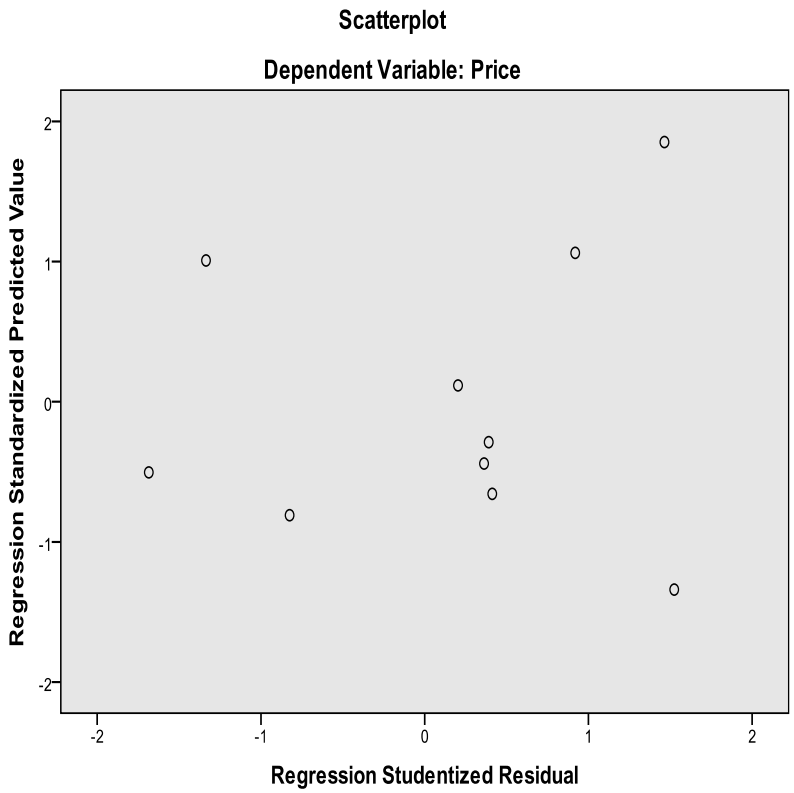
There are various **assumptions** underpinning a multiple linear regression model that must be checked – especially if we are to trust calculations such as that in Section 3.3.3.

The main assumption is that the model error term, denoted by ϵ, is **Normally distributed** with **constant variance** (*homoscedasticity*).

This is a rather complex concept, but one that can be quite easy to check in SPSS.

3.3.4 Checking assumptions





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.

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.

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.

3.4.1 Non-linear regression

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.

3.4.1 Non-linear regression

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.

3.4.1 Non-linear regression

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

The fitted model is now:

$$Y = -1658.13 + 28.967X_2 - 0.125X_2^2 + \epsilon.$$

3.4.1 Non-linear regression

- Notice that the price of wine now depends on rainfall, but also $(\text{rainfall})^2$, hence the name **quadratic** regression used here
- Both the rainfall and $(\text{rainfall})^2$ variables are important predictors of price, as both of their coefficients are extremely significant (the associated p-values are 0.000)
- 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 $(\text{rainfall})^2$

3.4.1 Non-linear regression

- Notice that the price of wine now depends on rainfall, but also $(\text{rainfall})^2$, hence the name **quadratic** regression used here
- Both the rainfall and $(\text{rainfall})^2$ variables are important predictors of price, as both of their coefficients are extremely significant (the associated p-values are 0.000)
- 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 $(\text{rainfall})^2$

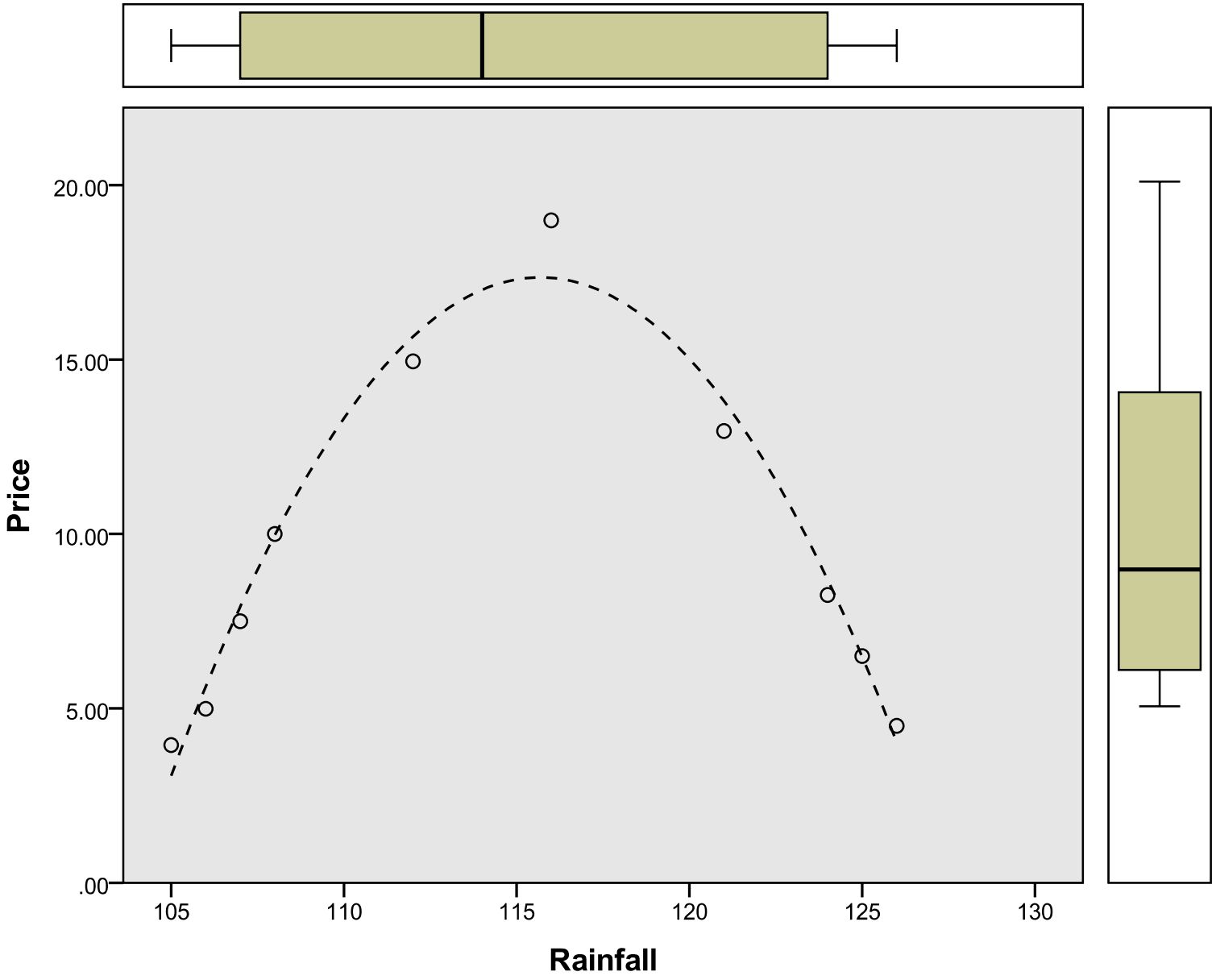
3.4.1 Non-linear regression

- Notice that the price of wine now depends on rainfall, but also $(\text{rainfall})^2$, hence the name **quadratic** regression used here
- Both the rainfall and $(\text{rainfall})^2$ variables are important predictors of price, as both of their coefficients are extremely significant (the associated p-values are 0.000)
- 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 $(\text{rainfall})^2$

3.4.1 Non-linear regression

- Notice that the price of wine now depends on rainfall, but also $(\text{rainfall})^2$, hence the name **quadratic** regression used here
- Both the rainfall and $(\text{rainfall})^2$ variables are important predictors of price, as both of their coefficients are extremely significant (the associated p-values are 0.000)
- 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 $(\text{rainfall})^2$

3.4.1 Non-linear regression



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?

Remember we have

$$Y = -1658.13 + 28.967X_2 - 0.125X_2^2 + \epsilon.$$

Subbing in $X_2 = 115$ and $X_2^2 = 125^2 = 13225$ gives

$$Y = -1658.13 + 28.967 \times 115 - 0.125 \times 13225 = £19.95.$$

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?
Remember we have

$$Y = -1658.13 + 28.967X_2 - 0.125X_2^2 + \epsilon.$$

Subbing in $X_2 = 115$ and $X_2^2 = 125^2 = 13225$ gives

$$Y = -1658.13 + 28.967 \times 115 - 0.125 \times 13225 = £19.95.$$

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?
Remember we have

$$Y = -1658.13 + 28.967X_2 - 0.125X_2^2 + \epsilon.$$

Subbing in $X_2 = 115$ and $X_2^2 = 125^2 = 13225$ gives

$$Y = -1658.13 + 28.967 \times 115 - 0.125 \times 13225 = £19.95.$$

3.4.2 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.

3.4.2 Logistic regression

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, 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*.

3.4.2 Logistic regression

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

Question

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

Most of the 1's are associated with higher values of X – indicating that people are more likely to be able to recall the brand if the product is placed later in the movie.

Question

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

Most of the 1's are associated with higher values of X – indicating that people are more likely to be able to recall the brand if the product is placed later in the movie.

3.4.2 Logistic regression

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). pause

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.

3.4.2 Logistic regression

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). pause

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.

3.4.2 Logistic regression

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**.

3.4.2 Logistic regression

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**.

3.4.2 Logistic regression

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.

3.4.2 Logistic regression

- A binary logistic regression ensures the model returns values between 0 and 1; so not on the same scale as the response variable itself, 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

$$P(Y = 1) \quad = \quad \frac{e^{-2.99+0.045X}}{1 + e^{-2.99+0.045X}}.$$

- The precise form of this Equation 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%)

3.4.2 Logistic regression

- A binary logistic regression ensures the model returns values between 0 and 1; so not on the same scale as the response variable itself, 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

$$P(Y = 1) = \frac{e^{-2.99+0.045X}}{1 + e^{-2.99+0.045X}}.$$

- The precise form of this Equation 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%)

3.4.2 Logistic regression

- A binary logistic regression ensures the model returns values between 0 and 1; so not on the same scale as the response variable itself, 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

$$P(Y = 1) = \frac{e^{-2.99+0.045X}}{1 + e^{-2.99+0.045X}}.$$

- The precise form of this Equation 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%)

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?

We have $X = 92$, giving

$$\begin{aligned} P(Y=1) &= \frac{e^{-2.99+0.045\times 92}}{1+e^{-2.99+0.045\times 92}} \\ &= 0.7595, \end{aligned}$$

so around 76 people in 100.

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?

We have $X = 92$, giving

$$\begin{aligned} P(Y = 1) &= \frac{e^{-2.99+0.045\times 92}}{1 + e^{-2.99+0.045\times 92}} \\ &= 0.7595, \end{aligned}$$

so around 76 people in 100.

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?

We have $X = 92$, giving

$$\begin{aligned} P(Y = 1) &= \frac{e^{-2.99+0.045\times 92}}{1 + e^{-2.99+0.045\times 92}} \\ &= 0.7595, \end{aligned}$$

so around 76 people in 100.