Statistical Perspectives on Data Mining

John Maindonald

September 3, 2008

Contents

Part I
Major Themes & Overview

Maindonald, J.H. (2006) comments, from a somewhat different perspective, on a number of the issues that are raised below.

1 A 'typical' data mining problem?

Aa a prelude to discussing of the nature and philosophy of data mining, I will take a quick look at a classification example that is intended to help illustrate some of the important issue. As is common in many of the examples that are the stock-in-trade of the data mining literature, the interest is in prediction rather than interpretation of model parameter of estimates.

1.1 Example – Forensic glass identification:

	WinF	WinNF	Veh	Con	Tabl	Head	Class'n error
Window float ('WinF': 70)	63	6	1	0	0	0	0.10
Window non-float ('WinNF': 76)	11	59	1	2	2	1	0.22
Vehicle window ('Veh': 17)	6	4	7	0	0	0	0.59
Containers ('Con': 13)	0	2	0	10	0	1	0.23
Tableware ('Tabl': 9)	0	2	0	0	7	0	0.22
Headlamps ('Head': 29)	1	3	0	0	0	25	0.14

The data consist of 214 rows × 10 columns.

Notice that:

- There are six different types of glass fragments. Window float and non-float glass (70 and 76 items respectively) are much more strongly represented than any other type of glass.
- A classification error is given – this is from use of the random forests algorithm – this will be described in due course. The overall error rate was 20.1%.

The information that is given about these data is rather scant. Assuming these samples really are representative, we can classify them with much more confidence than is the case for other glass types.

Questions, for any use of the results (e.g., to identify glass on a suspect)

How/when were data generated? (1987)

- Are the samples truly representative of the various categories of glass? (To make this judgement, we need to know how data were obtained.)

  Are they relevant to current forensic use? (Glass manufacturing processes and materials may have changed since 1987.)

  What are the prior probabilities? (Would you expect to find headlamp glass on the suspect's clothing?)

These data are probably not a good basis for making judgements about glass fragments found, in 2008, on a suspect's clothing. Too much is likely to have changed since 1987. We'd want data that are a better match with the glass fragments that one might currently expect to find. We can then generalize with confidence, from the sample from which results have been obtained to some wider population.

In practice, that may be an almost impossible ask. We may have to be content with data that are from a population that is a less than perfect match to the population to which results are to be applied.

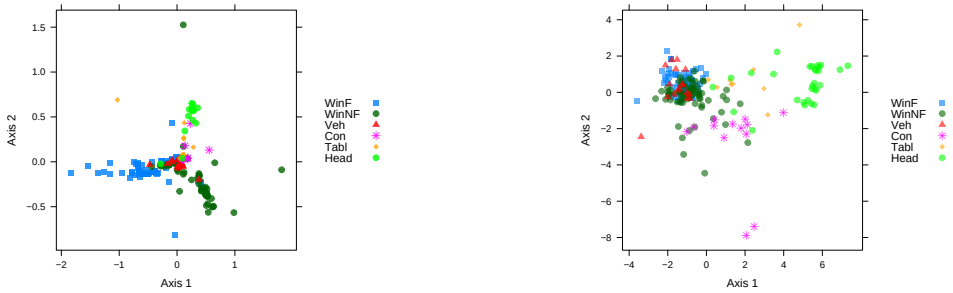

Figure 1: The two panels give two alternative two-dimensional views of the extent to which the respective classification algorithms have separated the points into the six groups.

Picturing the results: Of course, a two-dimensional view be too simplistic. With six types of glass, there are five dimensions in which one glass can be distinguished from other glasses. A two-dimensional view may lose more information than is reasonable. Checks on the adequacy of the two-dimensional view will be considered in a later chapter.

1.2 Issues for data miners to consider

Issues that need careful attention in any practical data mining project include:

- Results are required that apply to the population for which predictions are required? What are the implications for data collection and/or choice of data?
- There are many different methods/algorithms. How should the analyst choose between them? What are good ways to assess the performance of one or other algorithm?
- Often, the analyst would like to know which data columns (variables, or features) were important for the classification. Could some of them be omitted without loss?
- The analyst may want to attach an interpretation to one or more coefficients? Does the risk of heart attack increase with the amount that a person smokes?
- Above, I jumped directly into fitting a classification model, with no preliminary scrutiny of the data. This can be risky. What sorts of preliminary scrutiny can be used to identify problems with the data, or issues that ought to be addressed?
- I offered a two-dimensional summary of the results, allowing some insight into the classification result. What can be learned from such a plot? What other investigations might give useful insight on the analysis results?

Thus far, I have focused on classification. There are however other tools that the analyst will from time to time want to call into use, or that in some contexts may be more useful than classification. Easily the strongest candidate is regression. This will be needed whenever the outome is an outcome variable or measure, rather than a classification.

In fact, I will take the view that classification is a species of regression – a regression where the outcome is a classification rather than an outcome values for a continuous variable. It will become apparent that the two methodologies have important common features, as well as important differences. Where the focus is on features that they have in common, it makes sense to consider them together in the same discussion. When the differences seem more important than the common features, they will be considered together.

1.3 Statistics versus data mining – what is the connection?

I will start by giving my definition of statistics. It is the science of data collection and data analysis. Some, including perhaps some statisticians, will think this too broad. However that may be, it is the definition that underpins my thinking.

This advertises itself as a text about data mining, albeit from a statistician's perspective. I have given a definition of statistics. What however is data mining? Is it distinct from statistics? The previous section has, by way of an example, made some suggestions. It will become clear that I do not have a simple definition for data mining. Rather I will identify it by describing the territory that I judge it to have staked out for itself.

2 What is data mining?

This section will explore some of the issues that have influenced the development of data mining as a tradition of data analysis. Perhaps the most important is technological change.

2.1 Technological change

Advances in the past several decades have brought a variety of changes that affect the collection, manipulation and analysis of data (this list is taken from Maindonald , 2005):

- Large datasets that have been created by automation of data collection, and by the merging of existing databases, bring new challenges. The challenge may be to obtain forms of data summary that are suitable for analysis, and/or to handle the sheer bulk of the data. Or, as in the analysis of genomic expression array or other data where the number of outcome measures is large, the data may require substantial adaptation of existing analysis methods.
- There are new types of data, derived for example from documents, images and web pages.
- New data analysis methodologies often allow analyses that make better use of the data, more directly attuned to the questions of scientific interest, than was readily possible 15 years ago.
- Advances in statistical methodology have widened the gap between those whose statistical knowledge has not advanced much in the past decade, and those professionals who are fully au fait with modern methods.
- New statistical "meta-analysis" approaches that combine data from multiple studies into a single analysis may allow the detection of patterns that were not apparent from the individual studies. They may resolve some discrepancies between the separate analyses, while raising further questions. Note however that meta-analysis typically has complications that make automation hazardous.

2.2 A definition of data mining

Daryl Pregibon's definition of data mining as "Statistics at scale and speed" may be as apt as any. Scale and speed create, inevitably, a large demand for automation. The skill lies in knowing what to automate, when to call on the skill of the human expert, and in the use of tabular and graphical summaries that will assist the judgment of skilled data analysts or call attention to features of the data that might not otherwise be obvious. The demand for scale, speed and automation has created many opportunities for researchers from a computer science tradition to take a lead role.

Data mining, and indeed all data analysis, draws both from statistics and from computing.

- Statistics contributes: models, the distinction between signal and noise, attention to issues of generalization, well-tested modeling approaches, and a long tradition of experience in the analysis of data.

- Computing has contributed the means for managing data, for automating large parts of computations, for maintaining an audit of all steps in an analysis, and some novel algorithms and algorithmic approaches.

Comments in Witten and Frank (2000), with respect to machine learning. seem relevant also to data mining:

> In truth, you should not look for a dividing line between machine learning and statistics, for there is a continuum, and a multidimensional one at that, of data analysis techniques. . . . Right from the beginning, when constructing and refining the initial data set, standard statistical methods apply: visualisation of data, selection of attributes, discarding of outliers, and so on. Most learning algorithms use statistical tests . . . (p.26).[1]

There are then several alternative names for disciplines, or traditions, that operate in the same general arena as statistics – including especially machine learning and data mining. Another name that has soem currency is *analytics*, witness Davenport and Harris's text that *Competing on Analytics*. I will use *data analysis* as a name for activites that attract one or more of these names.

2.3 Linkages into statistics

There is an extensive statistical theory that offers important insights on all data analysis. This module will be unable to use this theory to any substantial extent; there is not time to develop the necessary theoretical tools. Instead

- This text will often fall back on a relatively informal ideas-based approach that makes little explicit use of mathematical formulae.
- Empirical approaches, e.g., to assessing accuracy, will be emphasized at the expense of modeling approaches that rely more heavily on statistical theory.

Good ways to move on from this module include getting up to speed in statistics, and working closely with experienced statisticians. This module, and other modules in this course, will give hints on areas of statistical theory that it will be useful to master, and should help motivate the theoretical content of any subsequent study of statistics.

Ideas of population and sample are crucial. These can be treated in a formal theoretical way. I will however adopt a less formal approach, using practical examples as motivation.

Statistical theory, as it affects practical data analysis, is currently developing very rapidly. This is a result of a synergy between new theoretical developments, and the computational power (software and hardware) of modern computer systems. The R system is one product of this synergy.

3 Planning, Context, Interpretation and Generalization

3.1 Questions to ask

Key issues for any study are:

1. Why am I undertaking this investigation?
2. What is the intended use of results?
3. What limitations, arising from the manner of collection or from the incompleteness of the information, may constrain that intended use?

When the analysis is complete, a key question will be: "What is the relevance of these results?"

[1] Be careful, though, what you do with outliers! Unless demonstrably erroneaous, they should, although perhaps omitted from the main analysis, be reported and included in graphs. In some analyses the interest may be in a small number of points that lie away from the main body of the data.

3.2 Purposes of a data analysis exercise

The following is a (perhaps incomplete) list of the purposes that a data analysis may aim to serve:

1. Data collection and summarization may be an end in itself. A business needs to have accurate accounts just so that it can know whether it is making a profit.
2. Prediction; i.e., the aim is to make statements that generalize beyond the circumstances that generated the particular data that are under study.
3. Understanding – the elucidation of pattern. To be of interest, the pattern must usually be relevant beyond the immediate data in which it was found, i.e., generalization is an issue here also.

3.2.1 Generalization

Most (all?) data mining analyses involve an element of generalization. In predictive modeling, generalization is an explicit concern. The nature of the generalization will typically have large implications for the investigations that are to be undertaken, of a kind that this module will explore.

3.2.2 Is an hypothesis essential?

The hypothesis testing approach to inference, while in wide use in some areas of statistical application, seems relatively uncommon in the data mining literature. Certainly, it offers a means for making statements that apply beyond the specific data used to generate and/or test them. It is not however always the best or most appropriate approach for this purpose.

3.2.3 Example –the different uses of Australian Bureau of Statistics data

Note the variety of uses of data that are collected by the the Australian Bureau of Statistics. By explicit use of samples, or (less often) census data, statements will be made that apply to one or other Australian population – to humans, sheep, farms, or whatever. Results may be used directly to allocate resources, e.g., the distribution of GST revenue to states. They are also a resource that will be used by researchers (statisticians, data miners) to find that patterns that will guide decision-making. As those decisions will affect the future, the interest is in those patterns that can be expected to persist into the future, i.e., there is a predictive element.

3.2.4 Exercises:

Set out aims for analysis for the studies that have generated the following data:

The forest cover type data set, available from the web site noted in connection with Blackard (1998). See the file **covtype.info** for details of these data.

The data set `ant111b` that gives yield of corn for each of four blocks at each of eight sites on the island of Antigua in the Caribbean, in a single year.[2]

The data set on tinting of car windows (`tinting` (also in `DAAG`).

The attitudes to science data set (`science`,*DAAG*).

Data on diet-disease associations, with the food frequency questionnaire as the diet measurement instrument.

Data on diet-genotype associations, with SNP (single nucleotide polymorphism) information for each of a number of positions on the chromosome used to indicate genotype.

[2]These data are included in the `DAAG` package for R. Several of the data sets that appear in illustrative examples in these notes are from *DAAG*.

Studies and/or associated data sets that may be encountered in remaining modules of the course.

Part II
Populations, Samples & Sample Statistics

4 Populations and Samples

The available data rarely comprises a total population. At best, it is likely to be a sample, preferably a random sample in which all population values appear with equal probability, from the population.

This is likely to be true even if the sample comprises all the data that were available at the time. Results will typically be applied in some new context, later in time, where the available data have changed. At best, changes between the original data and the later point in time for which predictions will be made will be rather similar to changes between one sample and another. This is however a best case scenario. Commonly there will be changes similar to those between one sample and another, plus systematic changes in time.

Thus a bank will have, in principle at least, complete information on financial transactions with current customers. As a guide to future financial transactions for those same customers, this is a sample of customer behavior that may or may not be a good guide to future transactions.

4.1 Why are continuous distributions important in data mining?

The stock-in-trade of data mining is classification. Why should we interested in characterizing and comparing continuous distributions?

- Commonly some explanatory variables will be continuous. Before fitting a classification model, it is desirable to do exploratory analyses that compare the groups with respect to both discrete and continuous variables.
- Categories will in some instances be formed by discretizing a continuous variable. Where possible, comparisons on the continuous scale should precede or accompany the discrete comparisons.
- For each category A, suppose that p_A is the probability, assessed independently of the data for an observation, that an observation belongs to category A. Many classification algorithms model $\log(p_A/(1-p_A))$, as a function of the explanatory factors and variables. The distribution of $\log(p_A/(1-p_A))$ is then of interest.
- Regression with a continuous dependent variable is an important methodology in its own right.

4.2 Empirical vs theoretical distributions

Consider now data that give the heights of 118 female students attending a first year statistics class at the University of Adelaide. Figure 2 plots a histogram and overlays it with a density plot. (The parameter setting `prob=TRUE` for the histogram is needed so that the units on the vertical scale are the same both for the histogram and for the density plot.) Vertical bars above the x-axis give the positions of the actual points. The function `na.omit()` omits missing values.

The vertical scale is chosen so that multiplying the height of each rectangle by the width of its base (5cm in each case) gives an estimate of the proportion of data values in that range. The same scale is used for the density plots, except that the density now changes continuously. It estimates, at each point, the proportion of values per unit interval.

```
> library(MASS)             # MASS has the survey data set
> library(lattice)
> heights <- na.omit(survey[survey$Sex=="Female", "Height"])
> ## NB: The vertical scale for the histogram must be a density scale.
```

```
> ## for consistency with the density plot.
> hst <- histogram(heights, type="density", breaks=seq(from=145, to=185, by=5),
+                  panel=function(x, ...){
+                    panel.histogram(x, col="gray90", ...)
+                    panel.densityplot(x, plot.points="", ...)
+                    panel.rug(x, ..., start=0.01, end=0.045)
+                    xval <- pretty(x, n=40)
+                    den <- dnorm(xval, mean=mean(x), sd=sd(x))
+                    panel.lines(xval, den, col="gray40", lty=2)
+                  },
+                 xlab=paste("Heights (cm) of female 1st year\n",
+                            "Adelaide University statistics students"))
>     print(hst)
```

The data set `survey` is included with the `MASS` package.

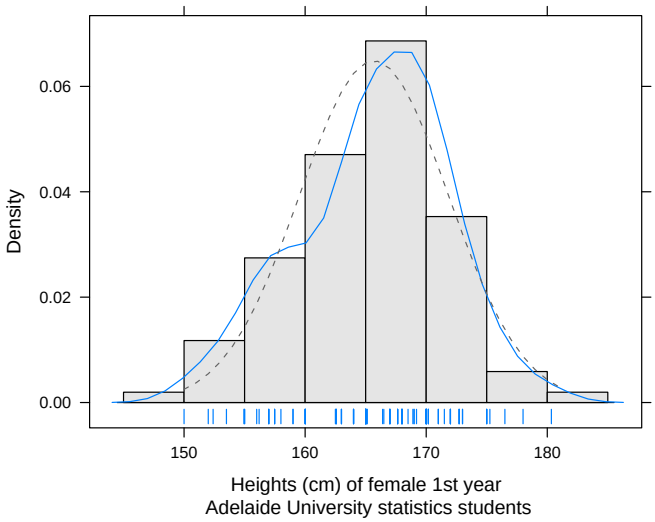

Figure 2: Vertical bars along the axis show the heights of 118 female students in a first year statistics class at the University of Adelaide. Alternative summaries of the distribution are a histogram, the overlaid density plot (solid curve), and a fitted normal curve (dashed).

Observe that:

- The vertical bars that show the distribution of data values are not very informative.
- The top of each histogram estimates the relative number of points (students) per unit along the x-axis, within the class boundaries of that histogram. That estimate changes suddenly at the class boundaries; this is an unsatisfactory feature of the histogram.
- The density curves give smooth estimates of the relative number of points (students) per unit along the x-axis. This is much preferable. However there is still an issue of the choice of bandwith for the smoother. This corresponds to the need to choose, for the histogram, the class width.
- The solid curve is a density estimate that makes very limited assumptions about the population density. The appearance of the curve will, depending on the sample size, be quite strongly affected by sampling variation. Try repeating the plot with random samples of size 102 (the same size as the sample of Adelaide students) from the normal distribution.
- The dashed curve makes the strong assumption that the population distribution is normal.

Histograms are almost never used for formal inference, i.e., for reaching conclusions about the population from which the data have come. For reaching conclusions about the population from which the data have come there are two common approaches:

1. Resampling approaches work with the actual data values.
2. Reasoning may proceed as though the population distribution is normal, i.e., use the dashed density curve.

Proceeding as though the population distribution is normal is fine provided that

- Inferences will be based on the sample mean.
- The population distribution is not too far from normal. (NB: Greater deviations from normality can be tolerated for larger sample sizes).

There are many different possible probability distributions. As has been hinted, the normal distribution often has a special role. Before explaining the reason for that role,

4.3 Theoretical probability distributions and their parameters

Models that are commonly used for population distributions include the normal (heights and weights, preferably on a logarithmic scale), exponential (lifetimes of components, where the probability of failure is unchanged over time), uniform, binomial (number of female children in a family of size N), and Poisson (failures in some fixed time interval, where the probability of failure is unchanged over time). Even if none of these is the correct distribution, one of them may be a reasonable starting point for investigation.

4.3.1 Mathematical definition

A probability distribution on the real line is a measure that defines, for all x_1 and x_2 in the support of X

$$\Pr[x_1 < X \leq x_2].$$

In a discrete population, each value has a probability (or probability mass) associated with it. In a continuous population, each value x has an associated density $f(x)$, such that for any two values a and b in the support of $f()$,

$$\Pr[a < x <= b] = \int_a^b f(x)dx$$

4.3.2 Density Curves and Cumulative Distribution Functions

These may be defined either by a density function, or by a cumulative distribution curve.

The following plots the density of a normal distribution with a mean of 0 and SD=1:

```
> curve(dnorm(x), from = -3, to = 3)
```

Why were the limits for the curve taken to be -3 and 3?

The height of the curve is the probability density. For a small interval of width h including the point, the probability is

$$h \times \text{normal density}$$

The area under the curve between $x = x_1$ and $x = x_2$ is the probability that the random variable X will lie between $x = x_1$ and $x = x_2$.

Cumulative probability curves The following plots the cumulative probability curve of a normal distribution with a mean of 0 and SD=1 (these are the defaults):

```
> curve(pnorm(x), from = -3, to = 3)
```

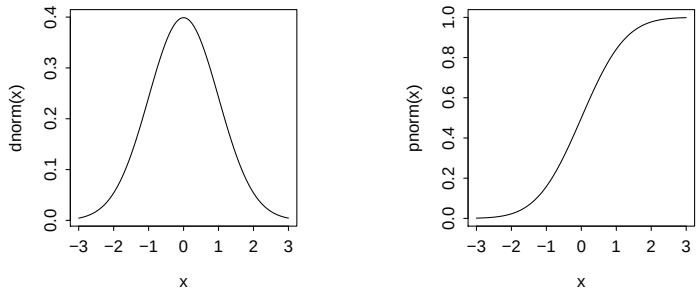

Figure 3: Normal density curve, with cumulative distribution function alongside.

The ordinates of the cumulative density curve give the cumulative probabilities, i.e., the height of the curve at x is $\Pr[X \leq x]$. It follows that

$$\Pr[x_1 < X \leq x_2 = \Pr[X \leq x_2 - \Pr[X \leq x_1.$$

Thus, suppose that X has a normal distribution with a mean of 0 and a standard deviation equal to 1. The probability that X is between -1 and 1 can be calculated as:

```
> pnorm(1) - pnorm(-1)

[1] 0.6826895
```

4.3.3 The mean and variance of a population

See Section 17 for the definition of the expectation of a random variable. The population mean is

$$\mu = \mathrm{E}[X] = \int x f(x) dx$$

while the variance is

$$\sigma^2 = \mathrm{E}[(X - \mu)^2] = \int (x - \mu)^2 f(x) dx$$

4.4 Samples from a Population – R functions

Unless stated otherwise, "sample" will mean "simple random sample".

The R functions `rnorm()` (normal), `rexp()` (exponential), `runif()` (uniform), `rbinom()` (binomial), and `rpois()` (Poisson), all take samples from infinite distributions.

```
> rnorm(n=10)
> runif(n=10)
```

The function `sample()` takes samples from a specified finite distribution. Samples may be taken without (the default) or with replacement. In without replacement sampling, each population value can appear at most once in the sample.

In with replacement sampling, each sampled element is placed back in the population before taking the next element. This is equivalent to sampling without replacement from the infinite population obtained by specifying a uniform distribution on the sample values. Try

```
> sample(1:8, size=5)
> sample(1:8, size=5, replace=TRUE)
> sample(c(2,8,6,5,3), size=4)
> sample(c(2,8,6,5,3), size=10, replace=TRUE)
```

Bootstrap sampling is with replacement sampling from an empirical distribution.

4.5 Displaying the distribution of sample values:

Examination of a the sample distribution may allow an assessment of whether the sample is likely to have come, e.g., from a normal population distribution. There is an art to making this comparison. In the sequel, some of the different ways in which the comparion might be made will be investigated.

4.5.1 An estimated density curve

Earlier, in Figure 2, we fitted a density curve to the distribution of heights of 118 female students attending a first year statistics class at the University of Adelaide. We now continue that discussion.

First, repeat the plot with a wider smoothing window. In the figure, I've added marks on the horizontal axis that show the actual heights. Also marked off, in gray lines, are the mean, mean-SD and mean+SD.

```
> heights <- na.omit(survey[survey$Sex=="Female", "Height"])
> ## NB: The vertical scale for the histogram must be a density scale.
> ## for consistency with the density plot.
> ## bw is the bandwidth of the smoother, in x-axis units
> den <- densityplot(heights, type="density", plot.points="rug", bw=2.5,
+                    panel=function(x, ...){
+                      panel.densityplot(x, ...)
+                      xval <- pretty(x, n=40)
+                      av <- mean(x); sdev <- sd(x)
+                      panel.abline(v=av, col="gray")
+                      panel.abline(v=av-sdev, col="gray", lty=2)
+                      panel.abline(v=av+sdev, col="gray", lty=2)
+                      den <- dnorm(xval, mean=av, sd=sdev)
+                      panel.lines(xval, den, col="gray40", lty=2)
+                      ytop <- 1.02*current.panel.limits()$ylim[2]
+                      panel.text(av-sdev, ytop, pos=1,
+                                 labels="mean-SD ", col="gray40")
+                      panel.text(av+sdev, ytop, pos=1,
+                                 labels=" mean+SD", col="gray40")
+                  },
+                 xlab=paste("Heights (cm) of female 1st year\n",
+                            "Adelaide University statistics students"))
> print(den)
```

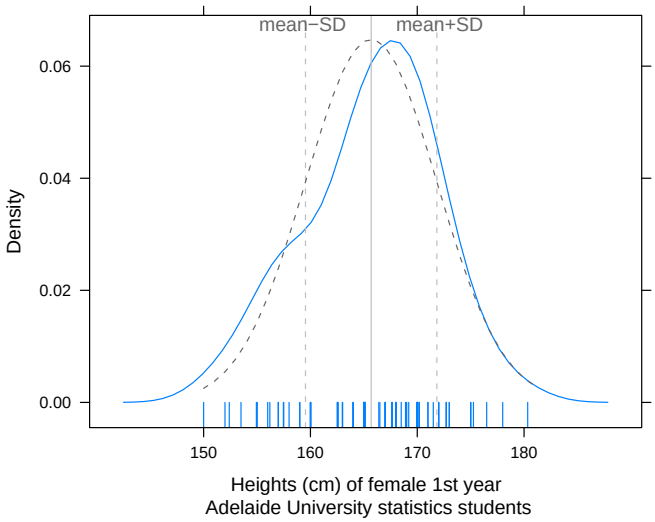

Figure 4: Density plot, now with a larger smoothing window (`bw`) and with a gaussian (normal) kernel, showing the distribution of heights of 118 female students in a first year statistics class at the University of Adelaide. A normal density curve has been added. Marks on the horizontal axis show the actual heights. Also marked off, in gray lines, are the mean, mean-SD and mean+SD.

Note: If data have a sharp lower or upper cutoff (a sharp lower cutoff at zero is common), parameters `from` and/or `to` can be set to ensure that this sharp cutoff is reflected in the fitted density.

Exercise: Draw a random sample of size 20 from an exponential distribution with `rate = 1`. Plot an estimated density curve.

4.5.2 Normal and other probability plots

Although preferable to histograms, density plots not in general an ideal tool for judging whether the sample is likely to have come from one or other theoretical distribution, most often the normal distribution. The appearance depends too much on the choice of bandwith. It lacks visual cues that can be used to identify differences from the theoretical distribution and decide whether they are important.

A much better tool is the Q-Q plot, which is a form of cumulative probability plot. Here, the focus will be on the comparison with a normal distribution, and the relevant Q-Q plot is a normal probability plot, using the function `qqnorm()`. Figure 5 shows a normal probability plot for the distribution of heights of the 118 female students in a first year statistics class at the University of Adelaide.

```
> y <- na.omit(survey[survey$Sex=="Female", "Height"])
> qqnorm(y)
```

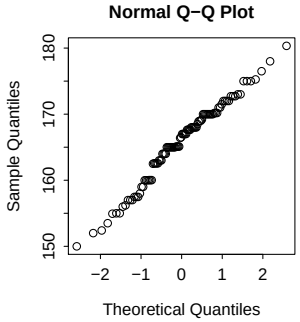

Figure 5: Normal probability plot for the distribution of heights of 118 female students in a first year statistics class at the University of Adelaide.

If data are from a normal distribution, points should lie close to a line. For a small sample size, quite large deviations from a line can be accepted. If the sample is large, points should lie close to a line. It is useful to draw repeated Q-Q plots with random samples of the same size from a normal distribution, in order to calibrate the eye. The function `qreference()` from the `DAAG` package may be useful for this purpose. For example:

```
> y <- na.omit(survey[survey$Sex=="Female", "Height"])
> qreference(y, nrep=6)
```

4.5.3 *Boxplots, and the inter-quartile range:

Another widely used measure of variability is the inter-quartile range. Boxplots, often used as summary plots to indicate the distribution of values in a sample, are drawn so that 50% of the sample lies between the upper and lower bounds of the central box. Figure 6 shows a boxplot representation of data on heights of female students in a first year statistics class at the University of Adelaide. The following code may be used to reproduce the boxplot, omitting the annotation.

```
> attach(survey)
> boxplot(Height[Sex=="Female"])
> detach(survey)
```

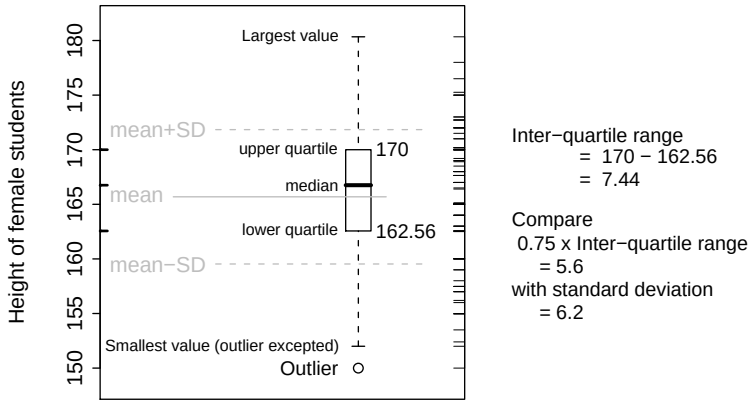

Figure 6: Boxplot, with annotation that explains boxplot features. Lines in gray show mean-SD, mean, and mean+SD. Data are heights of 118 female students in a first year statistics class at the University of Adelaide.

4.5.4 A further note on density estimation – controlling smoothness

We can control the smoothness of the density plot. There are various ways to do the smoothing. By default, with a normal "kernel", a mixture of normal densities is used.

Increasing the bandwidth makes the estimated density more like the density that is used as the kernel. Thus increasing the bandwith, with a `"gaussian"` kernel, is alright providing that the sample really is from a normal distribution. Try the following:

```
> plot(density(rnorm(50), kernel="rectangular", bw=0.5), type="l")
> plot(density(runif(50), kernel="rectangular", bw=0.5), type="l")
> plot(density(runif(50), kernel="gaussian", bw=0.5), type="l")
```

The density curve for a set of sample values lies somewhere between the theoretical distribution that is used as the kernel, and the sample distribution. Figure 7 shows, for the Adelaide female student data, the effect of varying the bandwidth.

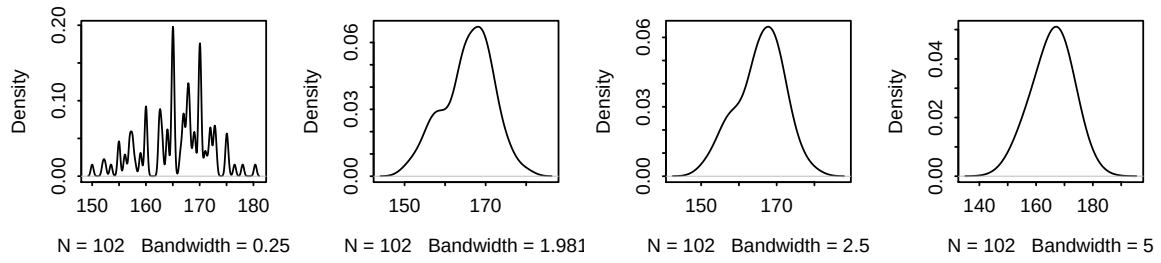

Figure 7: Density curves for Adelaide female student heights. Curves are shown for three different choices of bandwidth: 0.25, 1.98 (the default for these data), 2.5 and 5.0. The normal kernel (the default) is used in each case, so that increasing the bandwidth forces the curve closer to normal.

The default bandwith usually gives acceptable results. Experimentation with different choices of bandwidth is sometimes insightful.

5 Sample Statistics and Sampling Distributions

5.1 Variance and Standard Deviation:

In a sample, the *variance* is the average of the sum of squares of the deviations from the mean. If n is the sample size then, to correct for the fact that deviations are measured from the sample mean (rather than from the true mean), the sum of squares of deviations from the mean is usually divided by $n-1$. Thus, given sample values $x_1, x_2, \ldots, x_n$, the usual estimate of the variance σ^2 is

$$\widehat{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Why divide by $n-1$ rather than by n. A sample of one gives no information on the variance. Every value additional to the first gives one additional piece of information.

The standard deviation (SD) is the square root of the variance. The standard deviation is widely used, both in statistical theory and for descriptive purposes, as a measure of variability. The most obvious intuitive interpretations of the SD assume a normal population, or a random sample from a normal population. If data are from a normal population, then 68% of values will on average be within one standard deviation either side of the mean.

A key idea is that sample statistics have a sampling distribution – the distribution of values that would be observed from repeated random samples. This is an idea that will be illustrated in laboratory exercises.

Sample survey theory is one of several areas where there has been a strong tradition of basing all inferences on variances. This works well when inferences are mostly for means or totals and samples are large. The reason for this will become apparent below, in the discussion of the sampling distribution of the mean. There are however important small sample applications where it does not work well, and sample survey analysts are now moving away from the former relatively exclusive reliance on variance based inferences.

5.2 The Standard Error of the Mean (SEM):

The standard deviation estimates the variability for an individual sample value. This variability does not change (though the estimate will) as the sample size increases. On the other hand, the sample

mean does become less susceptible to variability as the sample size increases. If σ is the standard deviation then, for a random sample, the standard error of the mean is $\sigma/\sqrt{n}$.

Here are calculations that give, for the student heights, the mean, the standard deviation and the standard error:

```
> attach(survey)
> y <- na.omit(Height[Sex=="Female"])
> sd(y)

[1] 6.151777

> sd(y)/sqrt(length(y))

[1] 0.6091167

> detach(survey)
```

The standard error of the mean is, with a sample of 118, less than a tenth the size of the standard deviation. This result relies crucially on the i.i.d. assumption. This will be an important issue for multi-level models.

5.3 The sampling distribution of the mean:

We have just one sample, and therefore just one mean. The standard error of the mean relates to the distribution of means that might be expected if multiple samples (always of size 118) could be taken from the population that provided the sample.

It is however possible to simulate the taking of such repeated samples. As the sample distribution seems close to normal, the use of repeated samples of size 118 from a normal distribution seems reasonable. The following assumes a mean of 165.69, as for the sample, and the same SD of 6.15 as for the sample.

```
> av <- numeric(1000)
> for (i in 1:1000) av[i] <- mean(rnorm(118, mean=165.69, sd=6.15))
> plot(density(av), main="")
```

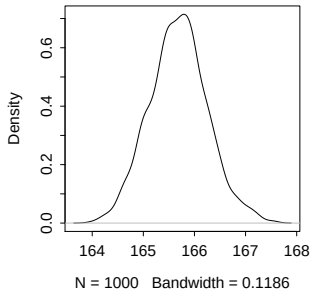

Figure 8: Simulated distribution of the mean, for samples of size 118 from a normal distribution with mean=165.7 and SD=6.15, as for the sample of UAdelaide students.

An alternative is to take repeated samples, with replacement, from the original sample itself. This is equivalent to sampling from a population in which each sample value is repeated an infinite number of times. The approach is known as "bootstrapping". This repeated sampling from the sample is just about as good an approximation as is available, if no use is made of theoretical results or approximations, to repeated sampling from the original population.

```
> av <- numeric(1000)
> for (i in 1:1000)
+      av[i] <- mean(sample(y, size=length(y), replace=TRUE))
> avdens <- density(av)
> plot(density(y), ylim=c(0, max(avdens$y)))
> lines(avdens, col="gray")
```

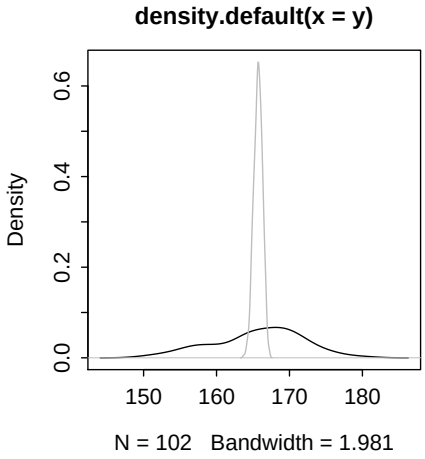

Figure 9: Simulated distribution of the mean, for repeated samples of size 118, with replacement, from the sample of University of Adelaide students.

The sampling distribution for the mean looks, in the company of the distribution of sample values. like a veritable Eiffel tower!

A practical consequence of the Central Limit Theorem is that the sampling distribution will for a sample of this size be much the same (close to a normal distribution) irrespective of the distribution from which the sample is taken, providing that the distribution is roughly symmetric and not unduly spread out in the tails. Try the following, which takes samples from a uniform distribution on the interval (0,1):

```
> par(mfrow=c(1,2))
> av <- numeric(1000)
> xval <- pretty(c(-.5, 1.5), 500)
> plot(xval, dunif(xval), type="l")
> for (i in 1:1000) av[i] <- mean(runif(n=118))
> plot(density(av))
> par(mfrow=c(1,1))
```

All statistics have sampling distributions. For example, there is a sampling distribution for the median. Unlike the distribution of the mean, this is strongly affected by the distribution from which the sample is drawn. Coefficients in linear or other models have sampling distributions.

Exercise 1: Try varying the sizes of the samples for which the averages are calculated. Even with `n` as small as 5 or 6, the distribution will be quite close to normal. Try also varying the number of samples that are taken. Taking some number of samples greater than 1000 will estimate the distribution more accurately; with fewer samples the estimate will be less accurate.

Exercise 2: Repeat, but now sampling from: (a) a uniform distribution, and (b) an exponential distribution.

6 The Assessment of Accuracy

Having trained a model, we would like to know how well the model has performed. If model A performs better than model B we will, other things being equal, prefer model A.

The discussion separates into two parts: model accuracy, and the accuracy of parameter estimates, with model accuracy usually an over-riding requirement. Accuracy of parameter estimates has additional complications, beyond those involved in assessing accuracy of model predictions.

6.1 Predictive accuracy

A first requirement is that predictions should be accurate, ideally for test data that accurately reflect the context in which the model will be used.

1. For a regression model with a continuous outcome, define the prediction error to be the difference between the model prediction and the observed value. The root mean square prediction error is then a measure of accuracy.

2. For a classification model, the percentage of correct classifications is often a suitable measure of accuracy. The deviance, or another "information" measure may by used, in some computational and theoretical contexts, as a proxy for percentage of correct classifications.

In practice predictive accuracy is commonly assessed for the same population from which the sample is derived. Assessment of the extent to which results are relevant to the target population is then a matter for separate investigation. This is a key issue, that is too often ignored.

Mechanisms that can be used for such assessment include:

- Derivation of a theoretically based estimate, e.g., for the error mean square for an `lm()` linear model.

- The training/test set approach, using a random split into training and test set.

- Cross-validation, in which each of k parts of the data become in turn the test set, with remaining data ($k-1$ out of k parts) used for training.

- Bootstrap approaches can be used in much the same way as cross-validation, c.f., the approach used by the *randomForest* package. Observations that for the time being serve as test data are said to be "out-of-bag", or OOB.

The final three methods are "resampling" methods, i.e., they rely on taking some form of sample from the one original available sample. As described here, all methods assume that observations have been sampled independently.

Laboratory Notes 4 demonstrate the use of cross-validation for assessing the predictive accuracy of a model.

6.2 Accuracy of parameter estimates

Quite stringent conditions are necessary to ensure that estimates for a regression or classifiction model will be unbiased or have negligible bias. The model must be correct. Part V illustrates, with examples, some of the issues.

Available methods are:

1. Estimates that depend heavily on distributional assumptions may be calculated from the one available sample. The standard errors, t-statistics, and related statistics that are included in the output from R's `lm()` linear modelling function have this character.

2. Bootstrap samples can be used to derive the sampling distributions of some of the statistics that may be of interest – means, means and regression coefficients. This approach does however have limitations, which can be serious. For extreme quantiles, it will fail.

3. In a limited range of circumstances, permutation methods may be used for tests of statistical significance.

As described here, all methods assume that observations have been sampled independently from the relevant population. Exact theoretically based results are available for models with iid normal errors. If the distribution is not normal results are, under relatively weak independence assumptions, valid asymptotically, i.e., it is valid in the limit as the sample size goes to infinity.

Bootstrap and permutation methods do not rely, directly, on normality assumptions. Some assumptions are however necessary if results are to be susceptible to ready interpretation. How does one interpret the result of a bootstrap version of a t-test for comparing two means, if the two distributions have a markedly different shape?

Laboratory Notes 3 demonstrate bootstrap samping and a permutation distribution approach, for the comparison of two means. It is assumed that there are no other factors that might, in part or whole, account for any difference.

6.3 Comparing two populations

Cuckoos lay eggs in the nests of other birds. The eggs are then unwittingly adopted and hatched by the host birds. Latter (1902) collected the cuckoo egg data presented in Figure 10A in order to investigate claims, made in Newton (1893-1896, p. 123), that the eggs that cuckoos lay in the nests of other birds tend to match the eggs of the host bird in size, shape and color.

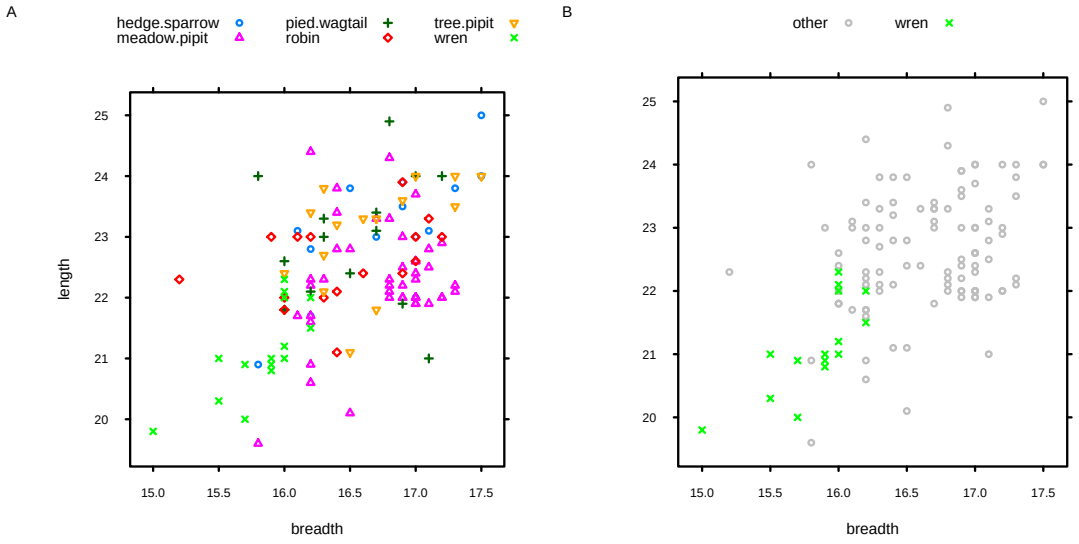

Figure 10: Length versus breadth of cuckoo eggs, identified according to the species of host bird in whose nest the eggs were laid.

Eggs laid in the nests of wrens are clearly much smaller, both in length and breadth, than the eggs laid in the nests of other birds. Visually, it is hard to see much distinction between eggs laid in the nests of these other species. For now, it therefore seems reasonable to examine the comparison between eggs laid in the nests of wrens, and eggs laid in the nexts of other birds, as in Figure 10B. Observe that Figure 10B tells much the same story, whether we focus on length or on breadth.

There are two types of questions that these data might be used to answer:

1. Are the apparent differences, between eggs found in the nests of wrens and eggs found in the nests of other birds, reproducible. If another sample of cuckoo eggs was collected, similarly a mix of eggs from wrens' nests and eggs from the nests of other birds, is it likely that similar differences would again be found?

2. Given a sample of cuckoo eggs is it possible to predict, with some reasonable accuracy, which eggs are from cuckoos and which from other birds?

The first question is often of interest in a data mining context, but is not usually the question of most direct interest. The second question is the one that is more commonly the focus of direct interest.

For the moment, the focus will be on the first question. I will look first at informal graphical comparisons, then moving to more formal methods.

6.3.1 Comparisons for individual variables

Figure 10B provided what is perhaps the most obvious form of graphical comparison. But might the difference between eggs laid in wren nests and eggs laid in other nests be merely a result of chance?

First, consider how we might do this separately for length. Figure 11 shows the comparison.

```
> dotwren <- dotplot(species %in% "wren" ~ length, data=cuckoos,
+                    scales=list(y=list(labels=c("Other", "Wren"))),
+                    xlab="Length (mm)")
> print(dotwren)
```

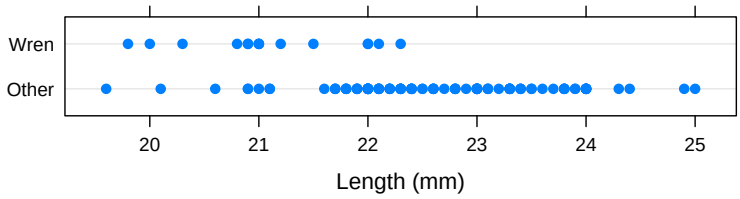

Figure 11: Dotplot comparison between lengths of eggs laid in wren nests and eggs laid in other nests.

6.3.2 A check that uses the bootstrap

The bootstrap (resampling with replacement) may be used to check whether the difference is likely to be more than noise. Repeated pairs of with replacement samples are drawn, the first member of the pair from "other" (non-wren), and the second member from eggs laid in wren nests. For each pair, calculate the difference between the means. Repeat a number of times (here 100, so that points stand out clearly, but 1000 would be better), and plot the differences. Figure 12 shows the result of one such sampling experiment.

```
> avdiff <- numeric(100)
> for(i in 1:100){
+   avs <- with(cuckoos, sapply(split(length, species %in% "wren"),
+                               function(x)mean(sample(x, replace=TRUE))))
+   avdiff[i] <- avs[1] - avs[2]  # FALSE (non-wren) minus TRUE (wren)
+ }
> dotdiff <- dotplot(~ avdiff, xlab="Length difference, non-wren - wren (mm)")
> print(dotdiff)
```

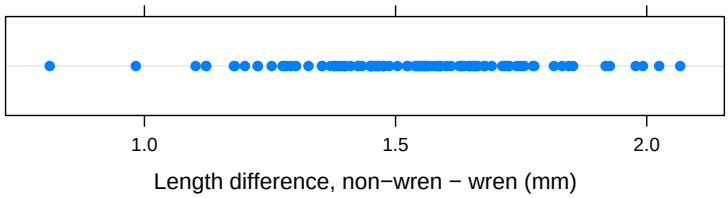

Figure 12: Differences, in successive bootstrap samples, between mean length of eggs in non-wren nests, and eggs in wren nests.

Observe that none of the differences are anywhere near zero. This is convincing evidence that the length differences are unlikely to be due to chance.

6.3.3 A check that uses simulation (the parametric bootstrap)

The difference here is that the random samples are drawn from normal distributions with the same mean and variance as in the samples.

If the variances can be assumed equal, the relevant distribution (when an infinite number of bootstrap samples are taken) can be determined theoretically, and except as a learning exercise there is not much point in such a simulation. If variances are unequal, the situation is more complicated. The standard theoretical approaches do however have simulation counterparts.

For a *t*-text comparison that allows for unequal variances, proceed thus:

```
> id <- as.numeric(with(cuckoos, species %in% "wren"))+1
> Species <- c("non-wren", "wren")[id]
> with(cuckoos, t.test(length[Species=="non-wren"],
+                      length[Species=="wren"]))

        Welch Two Sample t-test

data:  length[Species == "non-wren"] and length[Species == "wren"]
t = 7.0193, df = 21.244, p-value = 5.872e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.069984 1.970016
sample estimates:
mean of x mean of y
    22.64     21.12
```

Part III
Linear Models with an i.i.d. Error Structure

Most accounts of linear models assume that errors are independently and identically distributed (i.i.d.). That assumption is by no means necessary. In real world examples, it is often patently false. It will however be our starting point, for several reasons:

- There are a wide range of situations where the i.i.d. errors assumption is a reasonable approximation.
- It is enough to deal with one complication at a time.

7 Basic ideas of linear modeling

The base R system and the various R packages provide, between them, a huge range of model fitting abilities. In these notes, the major attention will be on the model fitting function is `lm()`, where the `lm` stands for linear model. Here, we fit a straight line, which is very obviously a linear model! This simple starting point gives little hint of the range of models that can be fitted using R's linear model `lm()` function. A later laboratory will build on the simple ideas that are presented here to present a far more expansive view of linear models.

R's implementation of linear models uses a symbolic notation Wilkinson & Rogers (1973), that gives a straightforward means for describing elaborate and intricate models.

7.1 Straight line Regression

	weight	depression
1	1.90	2.00
2	3.10	1.00
3	3.30	5.00
4	4.80	5.00
5	5.30	20.00
6	6.10	20.00
7	6.40	23.00
8	7.60	10.00
9	9.80	30.00
10	12.40	25.00

Table 1: Data showing depression in lawn (mm.), for various weights of roller (t)

A straight line regression model for the data in Table 1 can be written

$$\text{depression} = \alpha + \beta \times \text{weight} + \text{noise}.$$

Writing y in place of `depression` and x in place of `weight`, we have:

$$y = \alpha + \beta x + \varepsilon.$$

Subscripts are often used. Given observations (x_1, y_1), $(x_2, y_2), \ldots, (x_n, y_n)$, we may write

$$y_i = \alpha + \beta x_i + \varepsilon_i.$$

In standard analyses, we assume that the ε_i are independently and identically distributed as normal variables with mean 0 and variance σ^2. The $\alpha + \beta x$ term is the deterministic component of the model, and ε is the random noise. Greatest interest usually centers on the deterministic term. The R function `lm()` provides a way to estimate the slope β and the intercept α (the line is chosen so that the sum of squares of residuals is as small as possible). Given estimates (a for α and b for β), we can pass the straight line

$$\widehat{y} = a + bx$$

through the points of the scatterplot. Fitted or predicted values are calculated using the above formula, i.e.

$$\widehat{y}_1 = a + bx_1, \ \widehat{y}_2 = a + bx_2, \ldots.$$

By construction, the fitted values lie on the estimated line. The line passes through the cloud of observed values. Useful information about the noise can be gleaned from an examination of the residuals, which are the differences between the observed and fitted values,

$$e_1 = y_1 - \widehat{y}_1, \ e_2 = y_2 - \widehat{y}_2, \ldots .$$

In particular, a and b are estimated so that the sum of the squared residuals is as small as possible, i.e., the resulting fitted values are as close (in this "least squares" sense) as possible to the observed values. The residuals are shown as vertical lines, gray for negative residuals and black for positive

residuals, in Figure 13.

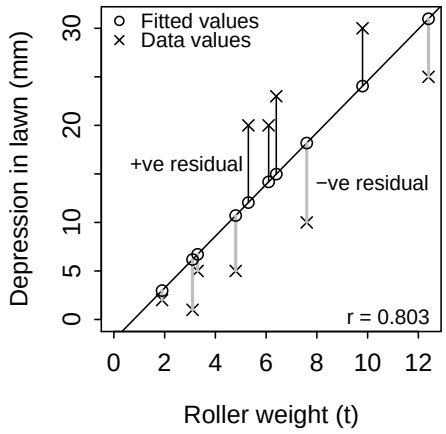

Figure 13: Lawn depression for various weights of roller, with fitted line. The fitted line is designed to minimize the sum of squares of residuals, i.e., the sum of squared lengths of the vertical lines, joining x's to o's, that are shown on the graph.

7.2 Syntax – model, graphics and table formulae:

The syntax for `lm()` models that will be demonstrated here is used right throughout the modeling functions in R, with modification as required. A very similar syntax can be used for obtaining graphs and for certain types of tables.

The following plots the data in the data frame `roller` (shown in Table 1) that is in the *DAAG* package.

```
> library(DAAG)
> plot(depression ~ weight, data=roller)
```

The formula `depression ~ weight` can be used either as a graphics formula or as a model formula. Just to see what happens, try fitting a straight line, and adding it to the above plot:

```
> lm(depression ~ weight, data=roller)

Call:
lm(formula = depression ~ weight, data = roller)

Coefficients:
(Intercept)       weight
     -2.087        2.667

> abline(lm(depression ~ weight, data=roller))
```

The different components of the model are called `terms`. In the above, there is one term only on the right, i.e., `weight`.

7.3 The technicalities of linear models

7.3.1 The model matrix – straight line regression example

The quantity that is to be minimized can be written:

$$\sum_{i=1}^{10}(y_i - a - bx_i)^2$$

Now observe how this can be written in matrix form. Set

$$\mathbf{X} = \begin{pmatrix} 1 & 1.9 \\ 1 & 3.1 \\ 1 & 3.3 \\ 1 & 4.8 \\ 1 & 5.3 \\ 1 & 6.1 \\ 1 & 6.4 \\ 1 & 7.6 \\ 1 & 9.8 \\ 1 & 12.4 \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} 2 \\ 1 \\ 5 \\ 5 \\ 20 \\ 20 \\ 23 \\ 10 \\ 30 \\ 25 \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} a \\ b \end{pmatrix}$$

Here **X** has the name "model matrix".

The quantity that is to be minimized is, then, the sum of squares of

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b} = \begin{pmatrix} 2 - (a + 1.9b) \\ 1 - (a + 3.1b) \\ 5 - (a + 3.3b) \\ 5 - (a + 4.8b) \\ 20 - (a + 5.3b) \\ 20 - (a + 6.1b) \\ 23 - (a + 6.4b) \\ 10 - (a + 7.6b) \\ 30 - (a + 9.8b) \\ 25 - (a + 12.4b) \end{pmatrix}$$

The sum of squares of elements of $\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b}$ can be written

$$\mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$$

The least squares equations can be solved using matrix arithmetic. For our purposes, it will be sufficient to use the R function `lm()` to handle the calculation:

```
> lm(depression ~ weight, data=roller)

Call:
lm(formula = depression ~ weight, data = roller)

Coefficients:
(Intercept)       weight
     -2.087        2.667
```

Both `weight` and `depression` are variables, i.e., they take values on the real line. They have, within R, class "numeric".

Recap, and Next Steps in Linear Modeling

The straight line regression model is one of the simplest possible type of linear model. We have shown how to construct the model matrix that R uses when it fits such models. Here, it had two columns only. Omission of the intercept term will give an even simpler model matrix, with just one column.

Regression calculations in which there are several explanatory variables are handled in the obvious way, by adding further columns as necessary to the model matrix. This is however just the start. There is a great deal more that can be done with model matrices, as will be demonstrated.

7.3.2 What is a linear model?

The models discussed here are linear, in the sense that predicted values are a linear combination of a finite set of basis functions. The basis functions can be nonlinear functions of the features, allowing

the modeling of systems in which there can nonlinear components that enter additively. The technical mathematical apparatus of linear models has a wider importance than linear models per se. It is a fundamental component of many of the algorithms that have been developed by machine learners, by data miners, and by statisticians.

Data that are intended for regression calculations consist of multiple observations (or instances, or realizations) of a vector $(x_1, x_2, \ldots, x_k, y)$ of real numbers, where the x_is are explanatory variables and y is the dependent variable.

Given $x_1, x_2, \ldots, x_k$, which take values on the real line, a first step (which in the simplest case maps the x_is onto themselves), is the formation of basis' functions

$$\phi_1(x_1, x_2, \ldots x_k), \phi_2(x_1, x_2, \ldots x_k), \ldots, \phi_p(x_1, x_2, \ldots x_k)$$

In the simplest case $p = k$ and $\phi_1(x_1, x_2, \ldots x_p) = x_1$, $\phi_2(x_1, x_2, \ldots x_p) = x_2$, $\ldots$, $\phi_p(x_1, x_2, \ldots x_p) = x_p$.

Then any function with values on the real line such that

$$f(x_1, x_2, \ldots x_k) = b_1\phi_1(x_1, x_2, \ldots x_k) + b_2\phi_2(x_1, x_2, \ldots x_k) + \ldots + b_p\phi_p(x_1, x_2, \ldots x_k)$$

where the elements of $\mathbf{b} = (b_1, b_2, \ldots b_p)$ are the only unknowns, specifies a linear model.

The model is linear in the values that the ϕ's take on the sample data. It is not, in general, linear in the x_i's. **Here endeth our brief excursion that has defined the term** *linear model*.

The random part of the model: The statistical output (standard errors, p-values, t-statistics) from the `lm()` function assumes that the random term is i.i.d. (independently and identically distributed) normal. Least squares estimation is them equivalent to maximising the likelihood.

What if the i.i.d. assumption is false? Depending on the context, this may or may not matter. In general, it is unwise to assume that it does not matter!

If the i.i.d. normal errors assumption is false in ways that are to some extent understood, then it may be possible to make use of functions in one or other of the R packages that are designed to facilitate the modeling of the random part of the model. Typically, these fit the model by maximising the likelihood. Note especially the R packages *nlme* and *lme4*, for handling multilevel and related models, and *arima* and related functions in the *stats* package that fit time series models.

7.3.3 Model terms, and basis functions:

In the very simple model in which depression is modeled as a linear function of `weight`, there the one term (`weight` generates two basis functions: $\phi_1(x) = 1$ and $\phi_2(x) = x$ which mapped values of `weight` into itself. (Basis functions seem an unnecessary complication, for such a simple example.)

7.3.4 Multiple Regression

In multiple regression, the model matrix has one column for the constant term (if any), plus one column for each additional explanatory variable. Thus, multiple regrssion is an easy extension of straight line regression. Further flexibility is obtained by transforming variable values, if necessary, before use of the variable in a multiple regression equation.

In the next example, there are multiple explanatory variables. We start with simple multiple linear regression model, and then look to see whether there is a case to replace the linear terms by polynomial or spline terms. Polynomial and spline terms extend the idea of "linear model", with the result that the dependence upon the variables in the model may be highly nonlinear! The `lm()` function will fit any model for which the fitted values are a linear combination of basis functions. Each basis function can in principle be an arbitrary transformation of one or more explanatory variables. "Additive models" may be better terminology.

The example will use the `hills2000` data set that is in the `DAAG` package. The row names store the names of the hillraces. For the `Caerketton` race, where the time seems anomalously small, `dist` should probably be 1.5mi not 2.5mi. The safest option may be to omit this point.

The interest is in prediction of `time` as a function of `dist` and `climb`. First examine the scatterplot matrices, for the untransformed variables, and for the log transformed variables. The pattern of relationship between the two explanatory variables – `dist` and `climb` – is much closer to linear for the log transformed data, i.e., the log transformed data are consistent with a form of parsimony that is advantageous if we hope to find a relatively simple form of model. Note also that the graphs of `log(dist)` against `log(time)` and of `log(climb)` against `log(time)` are consistent with approximately linear relationships. Thus, we will work with the logged data:

```
> library(DAAG)
> match("Caerketton", rownames(hills2000))

[1] 42

> loghills2k <- log(hills2000[-42, ])   # Omit the dubious point
> names(loghills2k) <- c("ldist", "lclimb", "ltime", "ltimef")
> loghills2k.lm <- lm(ltime ~ ldist + lclimb, data=loghills2k)
> par(mfrow=c(2,2))
> plot(loghills2k.lm)   # Diagnostic plot
> par(mfrow=c(1,1))
```

We pause at this point and look more closely at the model that has been fitted. Does `log(time)` really depend linearly on the terms `ldist` and `log(lclimb)`?

The function `termplot()` gives a graphical summary that can be highly useful. The graph is called a termplot because it shows the contributions of the different terms in the model. We use the function `mfrow()` to place the graphs side by side in a panel of one row by two columns:

```
> ## Plot the terms in the model
> if(dev.cur()==3)invisible(dev.set(2))
> par(mfrow=c(1,2))
> termplot(loghills2k.lm, col.term="gray", partial=TRUE,
+          col.res="black", smooth=panel.smooth)
> par(mfrow=c(1,1))
```

The plot shows the "partial residuals" for log(time) against log(dist) (left panel), and for log(time) against log(climb) (right panel). They are partial residuals because, for each point, the means of contributions of other terms in the model are subtracted off. The vertical scales show changes in `ltime`, about the mean of `ltime`.

The lines, which are the contributions of the individual linear terms ("effects") in this model, are shown in gray so that they do not obtrude unduly. For the lines as well as the points, the contributions of each term are shown after averaging over the contributions of all other terms. The dashed curves, which are smooth curves that are passed through the partial residuals, are the primary feature of interest in these plots. In both panels, they show clear indications of curvature.

This can be modeled, in the R context, by fitting either polynomial or spline curves. Spline curves are vastly more flexible than polynomial curves.

7.3.5 Modeling qualitative effects – a single factor

The `sugar` data frame (*DAAG* package) compares the amount of sugar obtained from an unmodified wild type plant with the amounts from three different types of genetically modified plants. In Table 2, the data are shown, with a model matrix alongside that may be used in explaining the effect of plant type (`Control`, or one of the three modified types `A` or `B` or `C`) on the yield of `sugar`.

In the model matrix in Table 2, `Control` is the baseline, and the yields for `A`, `B` and `C` are estimated as differences from this baseline. Then for each of the three treatments `A`, `B` and `C` there is an indicator variable that is 1 for that treatment, and otherwise zero. There are three basis functions that are used to account for the four levels of the factor `trt`.

The code used to fit the model is:

Sugar yield data

	weight	trt
1	82.00	Control
2	97.80	Control
3	69.90	Control
4	58.30	A
5	67.90	A
6	59.30	A
7	68.10	B
8	70.80	B
9	63.60	B
10	50.70	C
11	47.10	C
12	48.90	C

Model matrix

(Intercept)	trtA	trtB	trtC
1	0	0	0
1	0	0	0
1	0	0	0
1	1	0	0
1	1	0	0
1	1	0	0
1	0	1	0
1	0	1	0
1	0	1	0
1	0	0	1
1	0	0	1
1	0	0	1

Table 2: The data frame `sugar` is shown in the left panel. The right panel has R's default form of model matrix that is used in explaining the yield of `sugar` as a function of treatment (`trt`)

```
> library(DAAG)
> sugar.lm <- lm(weight ~ trt, data = sugar)
> summary(sugar.lm)

Call:
lm(formula = weight ~ trt, data = sugar)

Residuals:
     Min       1Q   Median       3Q      Max
-13.3333  -2.7833  -0.6167   2.1750  14.5667

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   65.367      2.236  29.229 2.03e-09
trt1          17.867      3.874   4.613  0.00173
trt2          -3.533      3.874  -0.912  0.38834
trt3           2.133      3.874   0.551  0.59685

Residual standard error: 7.747 on 8 degrees of freedom
Multiple R-squared: 0.7915,        Adjusted R-squared: 0.7133
F-statistic: 10.12 on 3 and 8 DF,  p-value: 0.004248
```

`Control` was taken as the baseline; the fitted value is 83.23, which is given as `(Intercept)`. The vales that are given for remaining treatments are differences from this baseline. Thus the fitted value (here equal to the mean) for treatment A is 83.23-21.40, that for B is 83.23-15.73, while that for C is 83.23-34.33.

The termplot summary

Again, termplots can be an excellent way to summarize results. Here is the termplot summary for the analysis of the cuckoo egg length data:

```
> termplot(sugar.lm, partial.resid = TRUE, se = TRUE)
```

The dotted lines show one standard deviation limits either side of the mean.

In the above model there was just one term, i.e., species, and hence just one graph. This one graph brings together information from the values of the six basis functions that correspond to the term `species`. The vertical scale is labeled to show deviations of egg lengths from the overall mean.

In this example the so-called "partial residuals" are the deviations from the overall mean. The dashed lines show one standard error differences in each direction from the species mean. (The standard error of the mean measures the accuracy of the mean, in the same way that the standard deviation measures the accuracy of the of an individual egg length.)

A note on factors: The names for the different values that a factor can take are the "levels".

```
> levels(bowler)
> levels(innings)
```

Internally, factors are stored as integer values. Each of the above factors has two levels. A lookup table is used to associate levels with these integer values.

Other things to try: The function `expand.grid()` can be helpful for setting up the values of the factors. We use `xtable()` to check that this gives the correct table:

```
> y <- c(10, 14, 40, 50)
> Z <- expand.grid(bowler = c("A", "B"), innings = c("one", "two"))
> xtabs(y ~ bowler + innings, data = Z)

      innings
bowler one two
     A  10  40
     B  14  50
```

Other parameterizations

1. Above we used the default "corner" parameterization, which R calls the "treatment" parameterization. There are alternatives. The most commonly used alternative parameterization is the "anova" parameterization, which R calls the "sum" parameterization. Use it thus:

   ```
   > options(contrasts = c("contr.sum", "contr.poly"))
   > model.matrix(~trt, data = sugar)

      (Intercept) trt1 trt2 trt3
   1            1    1    0    0
   2            1    1    0    0
   3            1    1    0    0
   4            1    0    1    0
   5            1    0    1    0
   6            1    0    1    0
   7            1    0    0    1
   8            1    0    0    1
   9            1    0    0    1
   10           1   -1   -1   -1
   11           1   -1   -1   -1
   12           1   -1   -1   -1
   attr(,"assign")
   [1] 0 1 1 1
   attr(,"contrasts")
   attr(,"contrasts")$trt
   [1] "contr.sum"

   > lm(weight ~ trt, data = sugar)
   ```

```
Call:
lm(formula = weight ~ trt, data = sugar)

Coefficients:
(Intercept)         trt1         trt2         trt3
     65.367       17.867       -3.533        2.133
```

These are called the "sum" contrasts (i.e., a particular form of parameterization) because they are constrained to sum to zero. The sum contrasts have been favoured in texts on analysis of variance.

2. There can be interactions between factors, or between factors and variables.

7.3.6 Grouping model matrix columns according to term

Quite generally, the basis functions $\phi_1, \phi_2, \ldots, \phi_p$ may be further categorized into groups, with one group for each term the model, thus:

$$\underbrace{\phi_1, \ldots, \phi_{m_1}}_{\text{Term1}}, \underbrace{\phi_{m_1+1}, \ldots, \phi_{m_2}}_{\text{Term2}}, \ldots$$

In the above, the basis functions for one factor formed just one termx. More generally, thare may be one group of basis functions for each of several factors. In the later discussion of spline terms, several basis functions will be required to account for each spline term in the model.

7.4 *Linear models, in the style of R, can be curvilinear models

We want to model y as a curvilinear function of x. This is straightforward, using the abilities of the *splines* package. The following uses the data frame `fruitohms` in the *DAAG* package.

First `ohms` is plotted against `juice`. The function `ns()` (*splines* package) is then used to set up the basis functions for the curve and pass a curve through these data. (There are other mechanisms, some of them more direct, but this is more insightful for present purposes.)

```
> library(DAAG)
> plot(ohms ~ juice, data = fruitohms)
> library(splines)
> fitohms <- fitted(lm(ohms ~ ns(juice, df = 3), data = fruitohms))
> points(fitohms ~ juice, data = fruitohms, col = "gray")
```

The parameter `df` (degrees of freedom) controls the smoothness of the curve. A large value for `df` allows a very flexible curve, e.g., a curve that can have multiple local maxima and minima.

The `termplot()` function offers another way to view the result. There is an option that allows, also, one standard error limits about the curve:

```
> ohms.lm <- lm(ohms ~ ns(juice, df = 3), data = fruitohms)
> termplot(ohms.lm, partial = TRUE, se = TRUE)
```

The labeling on the vertical axis shows differences from the overall mean of `ohms`. In this example the *partial* is just the difference from the overall mean.

Spline basis elements

It is insightful to extract and plot the elements of the B-spline basis. This can be done as follows:

```
> par(mfrow = c(2, 2))
> basismat <- model.matrix(ohms.lm)
> for (j in 2:5) plot(fruitohms$juice, basismat[, j])
```

The first column of the model matrix is the constant term in the model. Remaining columns are the spline basis terms. The fitted values are determined by adding a linear combination of these four curves to the constant term.

Splines in models with multiple terms

For present purposes, it will be enough to note that this is possible. Consider for example

```
> loghills2k <- log(hills2000[, ])
> names(loghills2k) <- c("ldist", "lclimb", "ltime", "ltimef")
> loghill2k.lm <- lm(ltime ~ ns(ldist, 2) + lclimb, data = loghills2k)
> par(mfrow = c(1, 2))
> termplot(loghill2k.lm, col.term = "gray", partial = TRUE, col.res = "black",
+     smooth = panel.smooth)
> par(mfrow = c(1, 1))
```

7.5 *Linear models – matrix derivations & extensions

- $\boldsymbol{y}$ (n by 1) is a vector of observed values, $\mathbf{X}$ (n by p) is model matrix, and $\boldsymbol{\beta}$ (p by 1) is a vector of coefficients.
- The model is $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, i.e. $\boldsymbol{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$ where the vector $\boldsymbol{\epsilon}$ of residuals is n by 1
- Least squares *normal* equations are
$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\boldsymbol{y}$$
(assuming $\boldsymbol{\epsilon}_i$ are iid normal, these are the maximum likelihood estimates)
- If variances are unequal, modify *normal* equations to
$$\mathbf{X}'\boldsymbol{W}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\boldsymbol{W}\boldsymbol{y}$$
where $\boldsymbol{W}$ is a diagonal matrix with elements equal to the inverses of the variances (justification is from maximum likelihood, or argue that leverage should be independent of variance)
- Assume $E[\boldsymbol{y}] = \mu = \mathbf{X}\boldsymbol{\beta}$, i.e. $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$.

7.5.1 Linear Models – general variance-covariance matrix

More generally, if $\boldsymbol{\epsilon}$ is multivariate normal with known variance-covariance matrix $\boldsymbol{\Sigma}$, then ML theory gives the equation as above with $\boldsymbol{W} = \boldsymbol{\Sigma}^{-1}$.

Two values with a high positive correlation contain, jointly, less information than two independent values. In the extreme case where the correlation is 1, the two variables carry the same information.

If the variance-covariance matrix $\boldsymbol{\Sigma}$ is not known, many different special methods are available for specific types of correlation structure that have in practice proved useful. For example, time series models typically try to account for correlations that are highest between points that are close together in time. Spatial analysis models typically allow for correlations that are a function of separation in space. Hierarchical multi-level models allow for different variance-covariance structures at each of several levels of hierarchy.

7.5.2 Least squares computational methods

A separate set of notes describes the approach, based on the QR matrix decomposition, that is used in R and in most of the R packages. Where methods that are directly based on QR are too slow, there may be a specialized method that takes advantage of structure in $\mathbf{X}$ to greatly speed up computation. Sparse least squares is an important special case. See Bates (2006); Koenker and & Ng (2003).

Part IV
Linear Classification Models & Beyond

See Ripley (1996); Venables and Ripley (2002); Maindonald & Braun (2007, Section 12.2).

The methods discussed here may be contrasted with the strongly non-parametric random forest method that uses an ensemble of trees. See Maindonald & Braun (2007, Section 11.7). A good strategy for getting started on an analysis where predictive accuracy is of primary importance is to fit a linear discriminant model with main effects only, comparing the accuracy from a random forest analysis. If the random forest analysis gives little or no improvement, the linear discriminant model may be hard to better. There is much more that can be said, but this may be a good starting strategy.

Notation and types of model

Observations are rows of a matrix $\mathbf{X}$ with p columns. The vector $\mathbf{x}$, is a row of $\mathbf{X}$, but in column vector form. The outcome is categorical, one of g classes.

Methods discussed here will all work with monotone functions of the columns of $\mathbf{X}$. By allowing columns that are non-linear monotone functions of the initial variables, additive non-linear effects can be accommodated.

For the discussion of logistic regression that now follows, $g = 2$. Logistic regression is a specific type of Generalized Linear Model, and will be introduced in this more general context.

The logistic regression model, fitted using R's `glm()` function, is closely analagous to the linear discriminant model, fitted using `lda()` with $g = 2$. Classification may be seen as regression with a categorical outcome. The differences in output between `glm()` and `lda()` reflect in part the difference between a regresson focus and a discrimination focus. They use different estimation criteria. Additionally, there are differences in output that in part reflect the different motivations of the two types of model.

The regression perspective, as implemented in `glm()`, is better suited to uses of the model where it is hoped to interpret model parameters in some meaningful manner. The classification/discrimination perspective, as implemented in `lda()` and `qda()`, has advantages if the chief interest is in prediction and predictive accuracy is of primary importance. It is often useful, with the one set of data, to complement the output from `lda()` with ouptut from `glm()`.

Section V will describe problems where the output from `glm()`, or some equivalent software, is more or less essential for the intended purpose of the analysis. Implementations that have a regression prespective are better adapted than those with a discrimination perspective for handling the problems described in Subsections 12.3, 12.5 and 12.6.

8 Generalized Linear Models

The models described here, or in the case of the airbag data an extension of such a model, are needed for handling the problems that are described in Subsections 12.3, 12.5 and 12.6. Data analysts should be aware of them, as they provide the only satistactory way to handle many of the problems for which they are designed.

Generalized linear models (GLMs) extend linear models in two ways. They allow for a more general form of expression for the expectation, and they allow various types of non-normal error terms. Logistic regression models are perhaps the most widely used GLM. In later comparisons with classification models, these will be the only models considered.

8.1 GLM models – models for E[y]

The straight line regression model has the form

$$y = \alpha + \beta x + \varepsilon$$

where, if we were especially careful, we would add subscript is to y, x, and ε. In this introductory discussion, we will consider with models where there is just one x, in order to keep the initial discussion simple.

Taking expectation on both sides of the equation used for the above straight line regression model, it follows that

$$\mathrm{E}[y] = \alpha + \beta x$$

where E is *expectation*. It is this form of the equation that is the point of departure for our discussion of generalized linear models. This class of models was first introduced in the 1970s, giving a unified theoretical and computational approach to models that had previously been treated as distinct. These models have been a powerful addition to the data analyst's armory of statistical tools.

8.1.1 Transformation of the expected value on the left

GLMs allow a transformation $f()$ to the left hand side of the regression equation, i.e., to E[y]. The result specifies a linear relation with x. In other words,

$$f(\mathrm{E}[y]) = \alpha + \beta x$$

where $f()$ is a function, which is usually called the *link* function. In the fitted model, we call $\alpha+\beta x$ the linear predictor, while E[y] is the expected value of the response. The function $f()$ transforms from the scale of the response to the scale of the linear predictor.

Some common examples of link functions are: $f(x) = x$, $f(x) = 1/x$, $f(x) = \log(x)$, and $f(x) = \log(x/(1-x))$. The last is referred to as the logit link and is the link function for logistic regression. Note that these functions are all monotonic, i.e., they increase or (in the case of $1/x$) decrease with increasing values of x.

8.1.2 Noise terms need not be normal

We may write

$$y_i = \mathrm{E}[y_i] + \varepsilon_i.$$

Here the y_i may have a distribution different from the normal distribution. The restriction is that the distribution must be from the *exponential* family. Common distributions are the binomial where y_i is the number responding out of a given total n_i, and the Poisson where y_i is a count. The y_i are assumed independent between observations, usually with different E[y_i].

Even more common may be models where the random component differs from the binomial or Poisson by having a variance that is larger than the mean. The analysis proceeds as though the distribution were binomial or Poisson, but the theoretical binomial or Poisson variance estimates are replaced by a variance that is estimated from the data. Such models are called, respectively, quasi-binomial models and quasi-Poisson models.

8.2 Generalized Linear Models – theory & computation

Here, it is convenient to recast the equations in matrix form.

- As before, we have $\boldsymbol{\mu} = \mathrm{E}[\mathbf{y}]$ (n by 1), $\mathbf{X}$ (n by p), and $\boldsymbol{\beta}$ (p by 1).
- The model is now

  $$f(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}, \quad \text{where } \mathrm{E}[\mathbf{y}] = \boldsymbol{\mu}$$

  Here, $f()$, which must be monotonic, has the name *link* function. For example,

  $$f(\boldsymbol{\mu}_i) = \log\left(\frac{\boldsymbol{\mu}_i}{N_i - \boldsymbol{\mu}_i}\right)$$

- The distribution of y_i is a function of the predicted value μ_i, independently for different observations. The different y_i are from the same exponential family, but the distributions are not identical. Commonly used exponential family distributions are the normal, binomial and Poisson.
- An extension is to the quasi-exponential family, where the variance is a constant multiple of an exponential family variance. The multiplying constant is estimated as part of the analysis. Applications for models with quasibinomial or quasipoisson errors may if anything be more extensive than for their exponential family counterparts.
- Just as for linear models, spline or other terms that model nonlinear responses can be fitted.

8.2.1 Maximum likelihood parameter estimates

- Recall that the equation is
$$f(\boldsymbol{\mu}) = E(\boldsymbol{y}) = \mathbf{X}\boldsymbol{\beta}$$
where $\boldsymbol{\mu} = E[\boldsymbol{y}]$
- Assuming a distribution from the exponential family, the maximum likelihood estimates of the parameters are given by
$$\mathbf{X}'\boldsymbol{W}\boldsymbol{\mu} = \mathbf{X}'\boldsymbol{W}\boldsymbol{y}$$
where $f(\boldsymbol{\mu}_i) = \mathbf{X}_i\boldsymbol{\beta}$
- Note that the (diagonal) element $\boldsymbol{W}_{ii}$ of $\boldsymbol{W}$ are functions both of var$[\boldsymbol{y}_i]$ and of $f(\boldsymbol{\mu}_i)$
- The ML equations must in general be solved by iteration ($\boldsymbol{\beta}$ appears on both sides of the equation.) Iteratively reweighted least squares is used, i.e. Newton-Raphson. Each iteration uses a weighted least squares calculation. As the weights are inversely proportional to the variances, they depend on the fitted values. Starting values are required to initiate calculations. The weighted least squares calculation is repeated, with new weights at each new iteration, until the fitted values converge.

8.2.2 Use and interpretation of model output

- GLMs with binomial errors are formally equivalant to discriminant models where there are two categories. The GLM framework has advantages for some problems.
- Output is in much the same form as for the `lm` models. There are additional subtleties of interpretation – a *z value* is not a *t*-statistic, though for some GLMs that yield *z value*s there are specific circumstances where it is reasonable to treat `z value`s as *t*-statistics.
[More technically, they are Wald statistics.]
- Except in special cases, the statistical properties of parameters rely on asymptotic results. Standards errors and *t*-statistics rely on first-order Taylor series approximations that, in the worst case, can fail badly. This applies, especially, to binary logistic regression.
- Predicted values are calculated on one or other of two different scales
  - Write $\hat{\boldsymbol{\mu}}$ for the estimate of E$[\boldsymbol{y}]$. This is the vector of predicted values on the scale of the response.
  - The vector of predicted values on the scale of the response is $f(\hat{\boldsymbol{\mu}}) = \mathbf{X}\hat{\boldsymbol{\beta}}$.

  Predictions on the scale of the response are, for logistic regression models, predictions of the probability that the outcome will be 1, rather than 0. If the probability is greater than 0.5, the prediction will be 1; if less than 0.5 the prediction is 0. The 0.5 cutoff can be adjusted for differences in the prior probability.
- For logistic regression models, and Poisson models with small expected values, assessments of predictive accuracy should be derived using a resampling approach, perhaps cross-validation.

9 Linear Methods for Discrimination

As before, observations are rows of a matrix $\mathbf{X}$ with p columns. The vector $\mathbf{x}$, is a row of $\mathbf{X}$, but in column vector form.

The outcome is categorical, one of g classes, where now g may be greater than 2. The matrix $\mathbf{W}$ estimates the within class variance-covariance matrix, while $\mathbf{B}$ estimates the between class variance-covariance matrix. Details of the estimators used are not immediately important. Note however that they may differ somewhat between computer programs.

The functions that will be used here are `lda()` and `qda()`, from the *MASS* package. The function `lda()` implements linear discriminant analysis, while `qda()` implements quadratic discriminant analysis. Quadratic discriminant analysis is an adaptation of linear discriminant analysis to handle data where the variance-covariance matrices of the different classes are markedly different.

An attractive feature of `lda()` is that it yields "scores" that can be plotted. Let $r = min(g-1, p)$. Recall that p is the number of columns of a version of the model matrix that lacks an initial column of ones. Then assuming that $\mathbf{X}$ has no redundant columns, there will be r sets of scores. The r sets of scores can be examined using a pairs plot. Often, most of the information is in the fist two or three dimensions. Such plots may be insightful even for data where `lda()` is inadequate as a classification tool.

9.1 `lda()` and `qda()`

The functions `lda()` and `qda()` in the *MASS* package implement a Bayesian decision theory approach.

- A prior probability π_c is assigned to the cth class ($i = 1, \ldots g$).
- The density $p(\mathbf{x}|c)$ of $\mathbf{x}$, conditional on the class c, is assumed multivariate normal, i.e., rows of $\mathbf{X}$ are sampled independently from a multivariate normal distribution.
- For linear discrimination, classes are assumed to have a common covariance matrix Σ, or more generally a common $p(\mathbf{x}|c)$. For quadratic discrimination, different $p(\mathbf{x}|c)$ are allowed for different classes.
- Use Bayes' formula to derive $p(c|\mathbf{x})$. The allocation rule that gives the largest expected accuracy chooses the class with maximal $p(c|\mathbf{x})$; this is the Bayes' rule.
- More generally, assign cost L_{ij} to allocating a case of class i to class j, and choose c to minimize $\sum_i L_{ic} p(i|\mathbf{x})$.

Note that `lda()` and `qda()` use the prior weights, if specified, as weights in combining the within class variance-covariance matrices.

Using Bayes' formula

$$\begin{aligned} p(c|\mathbf{x}) &= \frac{\pi_c p(\mathbf{x}|c)}{p(\mathbf{x})} \\ &\propto \pi_c p(\mathbf{x}|c) \end{aligned}$$

The Bayes' rule maximizes $p(c|\mathbf{x})$. For this it is sufficient, for any given $\mathbf{x}$, to maximize

$$\pi_c p(\mathbf{x}|c)$$

or, equivalently, to maximize

$$\log(\pi_c) + \log(p(\mathbf{x}|c))$$

Now assume $p(\mathbf{x}|c)$ is multivariate normal, i.e.,

$$p(\mathbf{x}|c) = (2\pi)^{\frac{p}{2}} \mid \Sigma_c \mid)^{-\frac{1}{2}} exp(-\frac{1}{2} Q_c)$$

where

$$Q_c = (\mathbf{x} - \mu_c)^T \Sigma_c^{-1} (\mathbf{x} - \mu_c)$$

Then

$$\log(\pi_c) + \log(p(\mathbf{x}|c)) = \log(\pi_c) - \frac{1}{2} Q_c + \frac{p}{2} \log(2\pi) - \frac{1}{2} \log(|\ \Sigma_c\ |)$$

Leaving off the $\log(2\pi)$ and multiplying by -2, this is equivalent to minimization of

$$Q_c + \log(|\ \Sigma_c\ |) - 2\log(\pi_c) = (\mathbf{x} - \mu_c)^T \Sigma_c^{-1} (\mathbf{x} - \mu_c) + \log(|\ \Sigma_c\ |) - 2\log(\pi_c)$$

The observation $\mathbf{x}$ is assigned to the group for which

$$(\mathbf{x} - \mu_c)^T \Sigma_c^{-1} (\mathbf{x} - \mu_c) + \log(|\ \Sigma_c\ |) - 2\log(\pi_c)$$

is smallest.

Set $\mu_c = \bar{\mathbf{x}}_c$, and replace $|\ \Sigma_c\ |$ by an estimate $\mathbf{W}_c$.

[Note that the usual estimate of the variance-covariance matrix (or matrices) is positive definite, providing that the same observations are used in calculating all elements in the variance-covariance matrix and $\mathbf{X}$ has no redundant columns.]

Then $\mathbf{x}$ is assigned to the group to which, after adjustments for possible differences in π_c and $|\ \Sigma_c\ |$, the Mahalanobis distance

$$(\mathbf{x} - \bar{\mathbf{x}}_c)^T \mathbf{W}_c^{-1} (\mathbf{x} - \bar{\mathbf{x}}_c)$$

of $\mathbf{x}$ from $\mathbf{x}_c$ is smallest.

If a common variance-covariance matrix $\mathbf{W}_c = \mathbf{W}$ can be assumed, a linear transformation is available to a space in which the Mahalanobis distance becomes a Eucliean distance. Replace $\mathbf{x}$ by

$$\mathbf{z} = (U^T)^{-1} \mathbf{x}$$

and $\bar{\mathbf{x}}_c$ by $\bar{\mathbf{z}}_c = (U^T)^{-1} \bar{\mathbf{x}}_c$ where U is an upper triangular matrix such that $U^T U = \Sigma$. Then

$$(\mathbf{x} - \mu_c)^T \mathbf{W}^{-1} (\mathbf{x} - \mu_c) = (\mathbf{z} - \bar{\mathbf{z}}_c)^T (\mathbf{z} - \bar{\mathbf{z}}_c)$$

which in the new space is the squared Euclidean distance to from $\mathbf{z}$ to $\bar{\mathbf{z}}_c$.

Note: For estimation of the posterior probabilities, the simplest approach is that described above. Thus, replace $p(c|\mathbf{x}; \theta)$ by $p(c|\mathbf{x}; \hat{\theta})$ for calculation of posterior probabilities (the 'plug-in' rule). Here, θ is the vector of parameters that must be estimated. The functions `predict.lda()` and `predict.qda()` offer the alternative estimate `method="predictive"`, which takes account of uncertainty in $p(c|\mathbf{x}; \hat{\theta})$. Note also `method="debiased"`, which may be a reasonable compromise between `method="plugin"` and `method="predictive"`

9.2 Canonical discriminant analysis

Here we assume a common variance-covariance matrix. As described above, replace $\mathbf{x}$ by

$$\mathbf{z} = \mathbf{U}^{\mathbf{T}^{-1}} \mathbf{x}$$

where $\mathbf{U}$ is an upper triangular matrix such that $\mathbf{U}^{\mathbf{T}} \mathbf{U} = \mathbf{W}$.

The between classes variance-covariance matrix becomes

$$\tilde{\mathbf{B}} = \mathbf{U}^{\mathbf{T}^{-1}} \mathbf{B} \mathbf{U}^{-1}$$

The ratio of between to within class variance of the linear combination $\alpha^T \mathbf{z}$ is then

$$\frac{\alpha^T \tilde{\mathbf{B}} \alpha}{\tilde{\alpha}^T \tilde{\alpha}}$$

The matrix $\tilde{\mathbf{B}}$ admits the principal components decomposition

$$\tilde{\mathbf{B}} = \lambda_1 \mathbf{u_1} \mathbf{u_1^T} + \lambda_2 \mathbf{u_2} \mathbf{u_2^T} + \ldots + \lambda_r \mathbf{u_r} \mathbf{u_r^T}$$

The choice $\alpha = \mathbf{u_1}$ maximizes the ratio of the between to the within group variance, a fraction λ_1 of the total. The choice $\alpha = \mathbf{u_2}$ accounts for the next largest proportion λ_2, and so on.

The vectors $\mathbf{u}_1, \ldots \mathbf{u}_r$ are known as "linear discriminants" or "canonical variates". Scores, which are conveniently centered about the mean over the data as a whole, are available on each observation for each discriminant. These locate the observations in r-dimensional space, where r is at most $\min(g-1, p)$. A simple rule is to assign observations to the group to which they are nearest, i..e., the distance d_c is smallest in a Euclidean distance sense.

For plotting in two dimensions, one takes the first two sets of discriminant scores. A point $\mathbf{z}_i$ that is represented as

$$\zeta_{i1}\mathbf{u}_1 + \zeta_{i2}\mathbf{u}_2 + ... + \zeta_{ir}\mathbf{u}_r$$

is plotted in two dinsensions as (ζ_{i1}, ζ_{i2}), or in three dinsensions as $(\zeta_{i1}, \zeta_{i2}, \zeta_{i3})$. The amounts by which the original columns of $\mathbf{x}_i$ need to be multiplied to give ζ_{i1} are given by the first column of the list element `scaling` in the `lda` object. For ζ_{i2}, the elements are those in the second column, and so on. See the example below.

As variables have been scaled so that within group variance-covariance matrix is the identity, the variance in the transformed space is the same in every direction. An equal scaled plot should therefore be used to plot the scores.

9.2.1 Linear Discriminant Analysis – Fisherian and other

Fisher's linear disciminant analysis was a version of canonical discriminant analysis that used a single discriminant axis. The more general case, where there can be as many as $r = \min(g-1, p)$ discriminant functions, is described here.

The theory underlying `lda()` assigns $\mathbf{x}$ to the class that maximizes the likelihood. This is equivalent to choosing the class c that minimizes $d_c + \log(\pi_c)$, where if the same estimates are used for $\mathbf{W}$ and $\mathbf{B}$, d_c is the distance as defined for Fisherian linear discriminant analysis. Recall that π_c is the prior probability of class c.

The output from `lda()` includes the list element `scaling`, which is a matrix with one row for each column of $\mathbf{X}$ and one column for each discriminant function that is calculated. This gives the discriminant(s) as functions of the values in the matrix $\mathbf{X}$.

9.2.2 Example – analysis of the forensic glass data

The data frame `fgl` in the *MASS* gives 10 measured physical characteristics for each of 214 glass fragments that are classified into 6 different types.

The following may help make sense of the information in the list element `scaling`.

```
library(MASS}
fgl.lda <- lda(type ~ ., data=fgl)
scores <- predict(fgl.lda, dimen=5)$x  # Default is dimen=2
## Now calculate scores from other output information
checkscores <- model.matrix(fgl.lda)[, -1] %*% fgl.lda$scaling
## Center columns about mean
checkscores <- scale(checkscores, center=TRUE, scale=FALSE)
plot(scores[,1], checkscores[,1])  # Repeat for remaining columns
## Check other output information
fgl.lda
```

93% of the information, as measured by the trace, is in the first two discriminants.

9.3 Two groups – comparison with logistic regression

Logistic regression, which can be handled using R's function `glm()`, is a special case of a Generalized Linear Model (GLM). The approach is to model $p(c|\mathbf{x};\hat{\theta})$ using a parametric model that may be the same logistic model as for linear and quadratic discriminant analysis.

In this context it is convenient to change notation slightly, and give $\mathbf{X}$ an initial column of ones. In the linear model and generalized linear model contexts, $\mathbf{X}$ has the name "model matrix".

The vector $\mathbf{x}$ is a row of $\mathbf{X}$, but in column vector form. Then if π is the probability of membershipÂăin the second group, the model assumes that

$$\log(\pi/(1-\pi) = \beta'\mathbf{x}$$

where β is a constant.

Compare logistic regression with linear discriminant analysis:

- Inference is conditional on the observed $\mathbf{x}$. A model for $p(\mathbf{x}|c)$ is not required. Results are therefore more robust against the distribution $p(\mathbf{x}|c)$.
- Parametric models with "links" other than the logit $f(\pi) = \log(\pi/(1-\pi)$ are available. Where there are sufficient data to check whether one of these other links may be more appropriate, this should be done. Or there may be previous experience with comparable data that suggests use of a link other than the logit.
- Observations can be given prior weights.
- There is no provision to adjust predictions to take account of prior probabilities, though this can be done as an add-on to the analysis.
- The fitting procedure minimizes the deviance, which is twice the difference between the log-likelihood for the model that is fitted and the loglikelihood for a 'saturated' model in which predicted values from the model equal observed values. This does not necessarily maximize predictive accuracy.
- Standard errors and Wald statistics (roughly comparable to t-statistics) are provided for parameter estimates. These are based on approximations that may fail if predicted proportions are close to 0 or 1 and/or the sample size is small.

9.4 Linear models vs highly non-parametric approaches

The linearity assumptions are restrictive, even allowing for the use of regression spline terms to model non-linear effects. It is not obvious how to choose the appropriate degree for each of a number of terms. The attempt to investigate and allow for interaction effects adds further complications. In order to make progress with the analysis, it may be expedient to rule out any but the most obvious interaction effects. These issues affect regression methods (including GLMs) as well as discriminant methods.

On a scale in which highly parametric methods lie at one end and highly non-parametric methods at the other, linear discriminant methods lie at the parametric end, and tree-based methods and random forests at the non-parametric extreme. An attraction of tree-based methods and random forests is that model choice can be pretty much automated.

9.5 Low-dimensional Graphical Representation

In linear discriminant analysis, discriminant scores in as many dimensions as seem necessary are used to classify the points. These scores can be plotted. Each pair of dimensions gives a two-dimensional projection of the data. If there are three groups and at least two explanatory variables, the two-dimensional plot is a complete summary of the analysis. Even where higher numbers of dimensions are required, two dimensions may capture most of the information. This can be checked.

With most other methods, a low-dimensional representation does not arise so directly from the analysis. The following approach, which can be used directly with random forests, can be adapted for use with other methods. The proportion of trees in which any pair of points appear together at the same node may be used as a measure of the "proximity" between that pair of points. Then, subtracting proximity from one to obtain a measure of distance, an ordination method can be used to find a representation of those points in a low-dimensional space.

Part V
Data Analysis and Interpretation Issues

Here, we draw attention to sources of bias or misleading results.

10 Sources of Bias

10.1 Data collection biases

Large biases can arise from the way that data have been collected. The Literary Digest poll that was taken prior to the US 1936 Presidential election, where Roosevelt had 62% of the vote rather than the predicted 43%, is an infamous example. The estimate of 43% was based on a sample, highly biased as it turned out, of 2.4 million!

The problems that arise can be exacerbated by more directly statistical problems, i.e., issues that it is important to note even if random samples are available. Estimates of regression coefficients, or other model parameters, cannot necessarily be taken at their face value.

10.2 Biases from omission of features (variables or factors)

Data analysis has as its end point the use of forms of data summary that will convey, fairly and succinctly, the information that is in the data. Considerable technical skill may be required to extract that information. Simple forms of data summary, which seem superficially harmless, can lead to misleading inferences.

The problem arises, often, from a combination of unbalance in the data and failure to account properly for important variables. To focus the discussion, consider observational studies of the effects of modest wine-drinking on heart disease (Jackson et al., 2005). There are a large number of factors that affect heart disease – genetic, lifestyle, diet, and so on. Any analysis of observational data that tries to account for their joint effect will inevitably be simplistic. The assumptions made about the form of the response (usually, a straight line on a suitably transformed scale) will be simplistic. Simplistic assumptions will be made about interaction effects (how does alcohol intake interact with other dietary habits?), and so on.

Some of the possibilities that it may be necessary to contemplate, for this specific example and more generally, are:

1. The issue is one of design of data collection, as well as analysis. If information has not been collected on relevant variables, the analyst cannot allow for their effect(s).

2. If the data are observational, there may be crucial variables on which it is impossible to collect information. Or there may be no good understanding of what the relevant variables are.

3. Providing the problem is understood and handled appropriately, large effects are unlikely, in large data sets, to arise from differences between sub-populations.

4. Small effects are highly likely, and should always be treated with scepticism. Small effects that are artefacts of the issues noted here show up more readily than small effects that are genuine. This is because the effects that will be noted here will almost invitably skew estimates of genuine effects, either exaggerating the effect or (just as likely) reversing the direction of its apparent effect.

10.2.1 Unequal subgroup weights – an example

Figure 10.2.1 relates to data collected in an experiment on the use of painkillers.[3]. Notice that the overall comparison (average for baclofen versus average for no baclofen) goes in a different direction

[3]Gordon, N. C. et al.(1995): "Enhancement of Morphine Analgesia by the GABAB against Baclofen". Neuroscience 69: 345-349

from the comparison for the two sexes separately.

Researchers had been looking for a difference between the two analgesic treatments, without and with baclofen. When the paper was first submitted for publication, an alert reviewer spotted that some of the treatment groups contained more women than men, and proposed a re-analysis to determine whether this accounted for the results.[4] When the data were analysed to take account of the gender effect, it turned out that the main effect was a gender effect, with a much smaller difference between treatments.

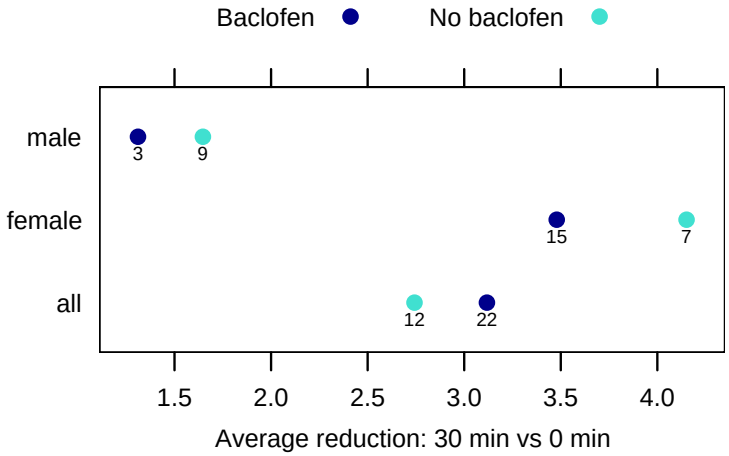

Figure 14: Does baclofen, following operation (additional to earlier painkiller), reduce pain? Subgroup numbers, shown below each point in the graph, weight the overall averages when sex is ignored.

The overall averages in Figure 10.2.1 reflect the following subgroup weighting effects:

Baclofen: 15f to 3m, i.e. $\frac{15}{18}$ to $\frac{3}{18}$ (a little less than f average)
No baclofen: 7f to 9m, i.e. $\frac{7}{16}$ to $\frac{9}{16}$ ($\approx \frac{1}{2}$-way between m & f)

This is still only part of the story. More careful investigation revealed that the response to pain has a different pattern over time. For males, the sensation of pain declined more rapidly over time.

Strategies

(i) Simple approach Calculate means for each subgroup separately.
Overall treatment effect is average of subgroup differences.
Effect of baclofen (reduction in pain score from time 0) is:

Females: 3.479 - 4.151 = -0.672 (-ve, therefore an increase)

Males: 1.311 - 1.647 = -0.336

Average over male and female = -0.5 × (0.672+0.336) = -0.504

(ii) Fit a model that accounts for sex and baclofen effects y = overall mean + sex effect + baclofen effect + interaction
(At this point, we are not including an error term).

Why specify a model?

It makes assumptions explicit. More anon!

[4] Cohen, P. 1996. Pain discriminates between the sexes. New Scientist, 2 November, p. 16.

10.2.2 Simpson's paradox

In multi-way tables, weighting effects such as have been noted lead to Simpson's paradox, known in the genetic context as epistasis. Here is a contrived example; data are admissions to a fictitious university:

	Engineering		Sociology		Total	
	Female	Male	Female	Male	Female	Male
Admit	10	30	30	15	*40*	*45*
Deny	10	30	10	5	*20*	*35*

Summing over the two separate tables is equivalent, for purposes of calculating overall admission rates, to the following:

Females: $\frac{10}{20} \times \frac{20}{60} + \frac{30}{40} \times \frac{40}{60}$ [0.33 (Eng) : 0.67 (Soc)]

Males: $\frac{30}{60} \times \frac{60}{80} + \frac{15}{20} \times \frac{20}{80}$ [0.75 (Eng) : 0.25 (Soc)]

The Overall Rates are:

- females ($\frac{2}{3}$): bias (0.33:0.67) is towards the Sociology rate (0.75)
- males ($\frac{45}{80}$): bias is (0.75:0.25) towards the Engineering rate (0.5).

For a real-life example that demonstrates this effect, see the data set `UCBAdmissions` that is supplied with the R system. Type

```
help(UCBAdmissions)     # Optional; get details of the data
example(UCBAdmissions)  # Summarize total data, and breakdown
                        # by departments
```

Several further examples, of this same general character, will be given in the next subsection.

Simpson's paradox and epistasis

In population genetics, Simpson's paradox type effects are known as epistasis. Most human societies are genetically heterogeneous. In San Francisco, any gene that is different between the European and Chinese populations will be found to be associated with the use of chopsticks! If a disease differs in frequency between the European and Chinese populations, then a naive analysis will find an association between that disease and any gene that differs in frequency between the European and Chinese populations.

Such effects are a major issues for gene/disease population association studies. It is now common to collect genetic fingerprinting data that should identify major heterogeneity. Providing such differences are accounted for, large effects that show up in large studies are likely to be real. Small effects may well be epistatic.

10.3 Model and/or variable selection bias

10.3.1 Model selection

When the model is fitted to the data used to select the model from a set of possible models, the effect is anti-conservative. Thus, standard errors will be smaller than indicated by the theory, and coefficients and t-statistics larger. Such anti-conservative estimates of standard errors and other statistics may, unless the bias is huge, nevertheless provide the useful guidance. Use of test data that are separate from data used to develop the model deals with this issue.

There is a further important issue, that use of separate test data does not address. Almost inevitably, none of the models on offer will be strictly correct. Mis-specification of the fixed effects, and to a lesser extent of the random effects, is likely to bias model estimates, at the same time inflating the error variance or variances, i.e., it may to some extent work in the opposite direction to selection effects.

10.3.2 Variable selection and other multiplicity effects

Variable selection has the same, or greater, potential for bias as model selection. This is an especial issue for the analysis of microarray and other genomic data, where a small number of gene expression measures, perhaps of the order of 5 - 20, may be selected from 10,000 or more. See Ambroise and McLachlan (2001) for a critique of papers where the authors have fallen prey to this trap. This can also be an issue for graphs that are based on the data that remain after selection.

Empirical accuracy assessments seem the only good way to address the major issues that can arise here. There are traps for data analysts who have not taken adequate account of the implications of selecting, for use in a regression or discriminant or similar analysis, a small number of variables ("features") from a much larger number. Maindonald (2003) gives a relatively elementary account of this matter, which should be accessible to non-specialists. The paper Ambroise and McLachlan (2001) is a careful examination of several examples, all concerned with the use of discriminant methods in connection with microarray data, from the literature. The same effects can arise from model tuning. Cross-validation is a key tool in this context. This, or the bootstrap, seems the only good way to allow for the skewing of results that can arise from potentially huge variable selection effects. Any model tuning and/or variable selection must be repeated at each cross-validation fold.

11 Errors in x

In the discussion so far, it has been assumed, either that the explanatory variables are measured with negligible error or that the interest is in the regression relationship given the observed values of explanatory variables.

This subsection is designed to draw attention to the effect that errors in the explanatory variables can have on regression slope. Discussion will be limited to a relatively simple "classical" errors in x model. For this model the error in x, if large, reduces the chances that the estimated slope will appear statistically significant. Additionally, it reduces the expected magnitude of the slope, i.e., the slope is attenuated. Even with just one explanatory variable x, it is not possible to estimate the magnitude of the error or consequent attenuation from the information shown in a scatterplot of y versus x. For estimating the magnitude of the error, there must be a direct comparison with values that are measured with negligible error.

The discussion will now turn to as study on the measurement of dietary intake. The error in the explanatory variable, as commonly measured, turned out to be larger and of greater consequence than most researchers had been willing to contemplate.

11.0.3 Measurement of dietary intake

The 36-page Diet History Questionnaire is a Food Frequency Questionnaire (FFQ) that was developed and evaluated at the U.S. National Cancer Institute, for use in large-scale trials that look for dietary effects on cancer and on other diseases. Given the huge scale of some of these trials, some costing US$100,000,000 or more, it has been important to have an instrument that is relatively cheap and convenient. Unfortunately, as the study that is reported in Schatzkin et al (2003) demonstrates, the FFQ seems too inaccurate to serve its intended purpose.

This FFQ queries frequency of intake over the previous year for 124 food items, asking details of portion sizes for most of them. There are supplementary questions on such matters as seasonal intake and food type. More detailed food records may be collected at specific times, which can then be used to calibrate the FFQ results. One such instrument is a 24-hour dietary recall, questioning participants on their dietary intake in the previous 24 hours. The accuracy of the 24-hour dietary recall was a further concern of the Schatzkin et al (2003) study. Doubly Labeled Water, which is a highly expensive biomarker, was used as an accurate reference instrument.

Schatzkin et al (2003) reported measurement errors where the standard deviation for estimated energy intake was seven times the standard deviation, between different individuals, of the reference. Additionally, Schatzkin et al (2003) found a bias in the relationship between FFQ and reference that further reduced the attenuation factor, to 0.04 for women and to 0.08 for men. For the relationship

between the 24 hour recalls and the reference, the attenuation factors were 0.1 for women and 0.18 for men, though these can be improved by the use of repeated 24-hour recalls. These errors were much larger than most researchers had been willing to contemplate.

The results reported in Schatzkin et al (2003) raise serious questions about what such studies can achieve, using instruments such as those presently available that are sufficiently cheap and convenient that they can be used in large studies. The measurement instrument and associated study design issues have multi-million dollar implications. Carroll (2004) gives an accessible summary of the issues.

This is a multi-million dollar issue. The following prospective studies that use such instruments are complete or nearly complete:

NHANES:	n = 3,145 women aged 25-50
(National Health and Nutrition Examination Survey)	
Nurses Health Study:	n = 60,000+
Pooled Project:	n = 300,000+
Norfolk (UK) study:	n = 15,000+
AARP:	n = 250,000+

Only 1 prospective study has found firm evidence suggesting a fat and breast cancer link, and 1 has found a negative link. The lack of consistent (even positive) findings led to the Women's Health Initiative Dietary Modification Study in which 60,000 women have been randomized to two groups: healthy eating and typical eating. Objections to this study are:

- Cost ($100,000,000+)
- Can Americans can really lower % fat calories from to 20%, from the current 35%
- Even if the study is successful, difficulties in measuring diet mean that we will not know what components led to the decrease in risk.

11.0.4 A simulation of the effect of measurement error

Suppose that the underlying regression relationship that is of interest is

$$y_i = \alpha + \beta z_i + \epsilon_i \ (i = 1, \ldots, n)$$

and that the measured values are

$$x_i = z_i + \eta_i$$

where

$$\text{var}[\epsilon_i] = \sigma^2; \ \text{var}[\eta_i] = \tau^2$$

Figure 15 shows the effect. If τ is 40% of the standard deviation in the x direction, i.e., $\tau = 0.4 s_z$, the effect on the slope is modest. If $\tau = 2 s_z$, the attenuation is severe.

The expected value of the attenuation in the slope is, to a close approximation

$$\lambda = \frac{1}{1 + \tau^2 / s_z^2}$$

where $s_z = \sqrt{\sum_{i=1}^{n} (z_i - \bar{z})^2}$. If $\tau = 0.4 s_z$, then $\lambda \simeq 0.86$.

Whether a reduction in slope by a factor of 0.86 is of consequence will depend on the nature of the application. Often there will be more important concerns. Very small attenuation factors (large attenuations), e.g. less than 0.1 such as were found in the Schatzkin et al (2003) study, are likely to have serious consequences for the use of analysis results.

Points to note are:

- From the data used in the panels of Figure 15, it is impossible to estimate τ, or to know the underlying z_i values. This can be determined only from an investigation that compares the x_i with an accurate, i.e., for all practical purposes error-free, determination of the z_i. The Schatzkin et al (2003) study that will be discussed below made use of a highly expensive reference instrument, too expensive for standard use, to assess and calibrate the widely used cheaper measuring instruments.

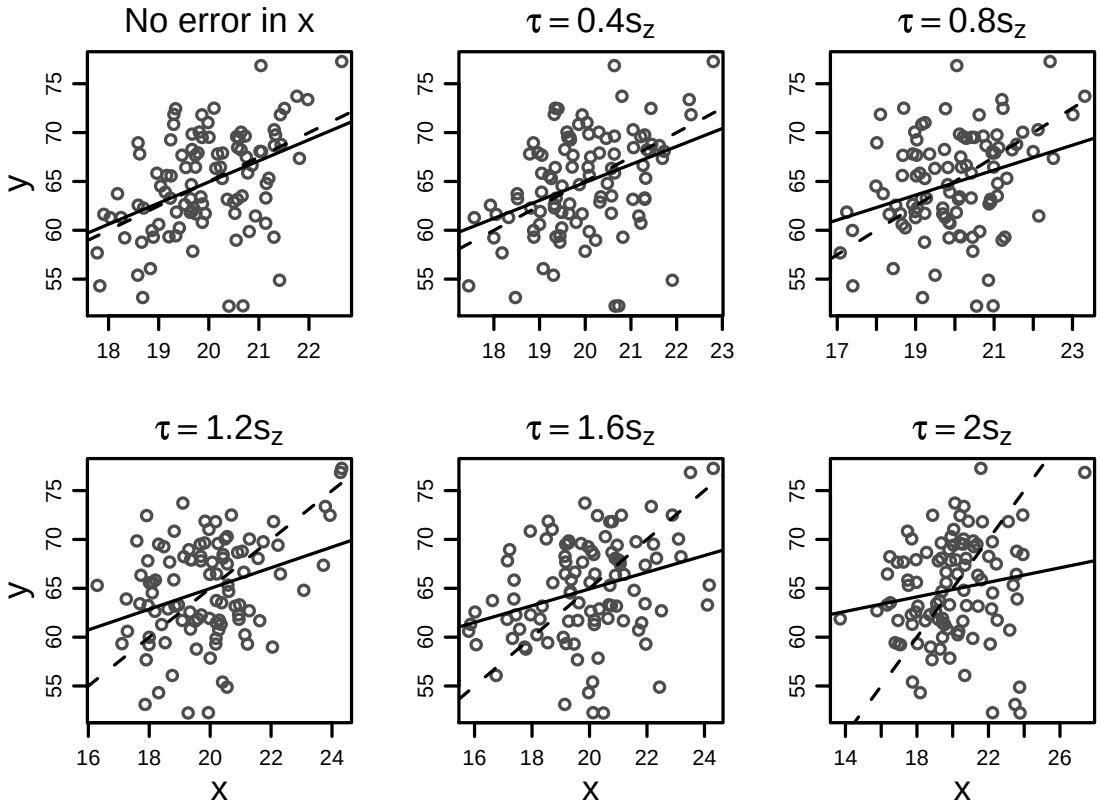

Figure 15: The fitted solid lines show how the regression line for y on x change as the error in x changes. The underlying relationship, shown with the dashed line, is in each instance $y = 15 + 2.5\ z$. Note that $s_z^2 = \sum_{i=1}^{n}(z_i - \bar{z})^2$, and that τ is the standard deviation of the independent errors that are added to the z_i.

- A test for $\beta = 0$ can be undertaken in the usual way, but with reduced power to detect an effect that may be of interest.
- The t-statistic for testing $\beta = 0$ is affected in two ways; the numerator is reduced by an expected factor of λ, while the standard error that appears in the numerator increases. Thus if $\lambda = 0.1$, the sample size required to detect a non-zero slope is inflated by more than the factor of 100 that the effect on the slope alone would suggest.

In social science, the ratio τ^2/s_z^2 has the name *reliability*. As Fuller (1987) points out, a better name is reliability ratio.

11.0.5 Errors in variables – multiple regression

Again, attention will be limited to the classical errors in x model. Where one only of several variables is measured inaccurately, its coefficient may on that account not appear statistically significant, or be severely attenuated. For remaining variables (measured without error) possible scenarios include: the coefficient suggests a relationship when there is none, or the coefficient is reversed in sign. Where several variables are measured with error, there is even more room for misleading and counter-intuitive coefficient values.

12 Further examples and discussion

12.1 Does screening reduce deaths from gastric cancer?

The issue here is that of comparing groups who may differ in respects other then the respect that is under investigation. In other words, there are likely to be hidden variables.

Patients who had surgery for gastric cancer were divided into two groups – those who had presented with cancer at a hospital or doctor's surgery, and those who had been diagnosed with cancer as a result of screening. Mortality was assessed in the 5 years following surgery:

	Mortality	Number
Unscreened Group	41.9%	352
Screened Group	28.2%	308

Table 3: Mortality in five-year period following surgery for cancer, classified according to whether patients presented with cancer, or cancer was detected by screening.

What are the possible explanations for the higher mortality in the unscreened group?

Screening may be catching cancer early, thus reducing the risk of death.

Cancers detected by screening may be at an earlier stage of development, and thus less immediately fatal.

Some cancers detected by screening may be of a less dangerous type, that progress slowly, or may never progress to become fatal.

All three effects may contribute to the difference.

Question: What are likely/possible missing variables/factors, for these data?

The appropriate approach is to identify several large groups of patients, randomly assigning groups for screening or no screening. Study participants are then followed for, e.g., the next decade. One study[5] classified 24,134 survey recipients as screened or unscreened, according as they had been screened, or not, in the previous year. It then followed them up for 40 months:

[5] used in: Inaba et al. 1999: Evaluation of a Screening Program on Reduction of Gastric Cancer Mortality in Japan: Preliminary Results from a Cohort Study. Preventive Medicine 29: 102-106

	Male		Female	
	Unscreened (n = 6,536)	Screened (n = 4,934)	Unscreened (n = 8,456)	Screened (n = 4,208)
Gastric cancer				
No. of deaths	19	8	9	4
Mortality rate	86.8	53.0	31.0	40.2
All causes				
No. of deaths	473	237	403	97
Mortality rate	2,199.0	1,593.1	1,370.7	829.4

Table 4: Mortality rates (deaths per 100,000 person years), from gastric cancer and from all causes.

Question: What are likely/possible missing variables/factors, for these data?

12.2 Cricket – Runs Per Wicket:

	1st innings		2nd innings		Overall	
	Runs	Wickets	Runs	Wickets		
Bowler A	40	4	240	6	280	10
Bowler B	70	5	50	1	120	6

Table 5: Runs per wicket for each bowler in the two innings.

The runs per wicket are:

	1st innings	2nd innings
Bowler A	10.00	40.00
Bowler B	14.00	50.00

Table 6: Runs per wicket for each bowler in the two innings.

Observe that although Bowler A does better than bowler B in each innings, his overall average is worse – 28 runs per wicket as opposed to 20.

A fair way to make the comparison is to model the effects both of bowler and of innings, using a linear model.

12.3 Alcohol consumptions and risk of coronary heart disease

Here, there are may factors for which there should be an adjustment. After adjusting for the effects of other factors. how does level of alcohol consumption affect risk of death? The method of analysis used is survival analysis, which will not be covered in this course. Think of it as an extension of the regression methodology that will be considered later in the course, with the risk of death relative to the baseline as the outcome. (Risk is expressed as a probability density; in this context it has the name “hazard” rate.)

No. of events (mortality/CHD)	All-cause mortality	Coronary heart disease
Men		
Never drink (16/43)	2.3 (1.2 – 3.8)	1.8 (1.3 – 2.5)
Special occasions (33/76)	1.4 (0.9 – 2.2)	1.1 (0.8 – 1.4)
1–2 times/month (37/93)	1.5 (1.0 – 2.2)	1.0 (0-8 – 1.3)
1–2 times/week (82/306)	1 (baseine)	1.0 (baseline)
Almost daily (52/219)	0.9 (0.7 – 1.3)	0.9 (0.8 – 1.1)
Twice a day or more (22/41)	2.5 (1.5 – 4.1)	1.1 (0.8 – 1.5)
Women		
Never drink (9/43)	1.5 (0.7 – 3.5)	1.8 (1.3 – 2.8)
Special occasions (40/127)	1.5 (0.7 – 3.5)	1.2 (0.9 – 1.5)
1–2 times/month (14/61)	1.7 (1.0 – 2.9)	1.0 (0-8 – 1.8)
1–2 times/week (26/137)	1 (baseine)	1.0 (baseline)
Almost daily (18/59)	1.3 (0.7 – 2.4)	0.8 (0.6 – 1.2)
Twice a day or more (5/7)	4.8 (1.8 – 12.7)	1.3 (0.6 – 2.8)

Table 7: Increased risk of mortality, relative to baseline, according to frequency of alcohol consumption. Factors for which adjustment was made were age, smoking, employment grade, blood cholesterol, blood pressure, body mass index, and general health as measured by a score from a questionnaire. CHD was recorded as an outcome if there was an episode of fatal or non-fatal cornonary heart disease.

Britton & Marmot (2004) report on an 11-year follow-up of a study of 10,308 London-based civil servants aged 35-55 years at baseline (33% female). Adjustments were made for age, smoking, employement grade, blood cholesterol, blood pressure, body mass index, and general health as measured by a score from a questionnaire. Table 7 shows the estimated ratio of risk relative to the baseline line, i.e., to the risk from all other factors.

Thus, it looks as though modest levels of alcohol consumption may be beneficial. However the results remain controversial. There may for example be lifestyle factors, associated with levels of alcohol consumption, for which factors such as employment have not made adequate adjustment. If such factors are correlated with frequency of drinking, this might in part explain the result. See especially Jackson et al. (2005).

Note also another source of evidence, derived from so-called Mendelian randomization studies. (Mendelian dose assignment would be a more accurate description than "Mendelian randomization".) Half of the Japanese population is homozygous or heterozygous for a non-functional variant of the gene ALDH2, making them unable to metabolise alcohol properly, with unpleasant consequences. The effect is more serious for the homozygotes than for the heterozygotes. The result is that homozygotes heavily curtail their alcohol consumption and heterozygotes curtail it to some lesser extent. The incidence of CHD closely reflects results predicted by Britton & Marmot (2004). At the same time, no association was apparent between genotype and other factors implicated in CHD. See Davey Smith & Ebrahim (2005).

12.4 Do the left-handed die young

A number of papers, in Nature, in the psychological literature and in the medical literature, have argued that left-handed people have poorer survival prospects than right-handers. It turns out that, in a large cross-sectional sample of the British population that was studied in the 1970s, the proportion of left-handers declined from around 15% for ten-year-olds to around 5% for 70-year olds. If average age at death is compared between left-handers and right-handers, left-handers will be over-represented among those dying young, and over-represented among those dying in older years. Hence the average age will be be lower for left-handers than for right-handers. Disturbingly it has been easier to get this nonsense published than to get refutations published.

Again survival analysis methods are required for a proper analysis. Once the effect noted above has been removed, there may be a small residual effect from lef-thandedness. See Bland & Altman

(2005).

12.5 Do airbags reduce risk of death in an accident

Each year the National Highway Traffic Safety Administration in the USA collects, using a random sampling method, data from all police-reported crashes in which there is a harmful event (people or property), and from which at least one vehicle is towed. The data in Table 8 are a summary of a subset of the 1997-2002 data, as reported in Meyer & Finney (2006).

Seatbelt	Airbag	Fatalities	Occupants
seatbelt	airbag	8626	4871940
none	airbag	10650	870875
seatbelt	none	7374	2902694
none	none	20550	1952211

Table 8: Number of fatalities, by use of seatbelt and presence of airbag.

Meyer & Finney (2006) conclude that on balance (over the period when their data were collected) airbags cost lives. Although their study is better than the official National Highway Traffic Safety Administration assessment of the evidence, based on accidents where there was at least one death. In order to obtain a fair comparison, it is necessary to adjust, not only for the effects of seatbelt use, but also for speed of impact. When this is done, airbags appear on balance to be dangerous, with the most serious effects in high impact accidents. Strictly, the conclusion is that, conditional on involvement in an accident that was sufficiently serious to be included in the database (at least one vehicle towed away from the scene), airbags are harmful.

Both sets of data are from accidents, and there is no way to know how many cases there were with airbags where accidents (serious enough to find their way into the database) were avoided, as opposed to the cases without aribags where accidents were avoided. Tests with dummies do not clinch the issue; they cannot indicate how often it will happen that an airbag disables a driver to an extent that they are unable to recover from an accident situation enough to avoid death or serious injury.

In ongoing debate and controversy over the use of airbags, errors have been identified in the data. Use of the corrected data do not, however, substantially change the conclusions. Further questions, additional to those noted above, have been raised. A forthcoming issue of *Chance* will take up some of these further issues. The data (the initial data an/or perhaps the corrected data) will appear in one of the sets of laboratory exercises.

Before installation of airbags was ever made mandatory, should there have been a large controlled trial in which one out of every two cars off the production line was fitted with an airbag? Would it have worked? Or would there be too much potential for driver behaviour to be influenced by whether or not there was an airbag in the car? Would it have been possible to sell the idea of such a trial to the public?

12.6 Hormone replacement therapy

Cohort and other population based studies have suggested hormone replacement therapy (HRT) reduces the risk of coronary heart disease (CHD). A large meta-analysis of what were identified as the best quality observational studies found a relative reduction in risk of 50% from any use of HRT.

A large randomized controlled trial found an increase in hazard, from use of CRT, of 1.29 (95% CI 1.02–1.63), after 5 years of follow-up. Thus, so far from reducing CHD risk, it increases the risk. The conclusion given in a 2006 ABC Health Report interview is that:

> Hormone therapy, both oestrogen combined with progesterone and oestrogen alone, increase risk of cardio vascular disease, stroke, blood clots and the hormone therapy that was combined meaning oestrogen and progesterone increase risk of breast cancer.
> [This is taken from: `http://www.abc.gov.au/rn/healthreport/stories/2006/1530042.htm`]

This was an especial puzzle because the results of the observational studies have been consistent with the results of randomized trials for other outcomes – breast cancer (increased risk for the combined oestrogen/progesterone HRT; for a 50-year old from 11 in 1000 to maybe 15 in 1000), colon cancer (reduced risk), hip fracture (reduced risk, but diet, exercise and other drugs can achieve the same or better results) and stroke (increased risk; for a 50-year old from 4 in 1000 to 6 in 1000). See the ABC web page just noted and, e.g., Rossouw et al. (2002) for further details and references.

A recent analysis by Hérnan et al. of the observational data gave the following factors by which the average risk is multiplied: These effects are assumed to add.

Years of follow-up	0 - 2	>2 - 5	>5
Multiply risk by	1.5	1.3	0.67

Years since menopause	<10	10 - 20	>20
Multiply risk by	0.89	1.24	1.65

The observational data included some individuals with long follow-up times, whereas the nature of a randomized trial (after randomization, there is a limited follow-up time) rules out long follow-up times. Moreover, in order to make up numbers, the randomized trials included many women with long times following menopause. Both these factors increase the average estimated risk for the randomized trials, relative to the observational data. The analysis will appear later this year, in a paper in the journal *Epidemiology*.

In part, the issue is that both the randomized trials and the observational studies yielded averages for populations that were heterogeneous in ways that gave different relative weights to relevant subpopulations. Earlier analyses failed to identify important relevant covariates.

12.7 Freakonomics

Several of the studies that are discussed in Leavitt and Dubner (2005), some with major public policy relevance, relied to an extent on regression methods – usually generalized linear models rather than linear models. References in the notes at the end of their book allow interested readers to pursue technical details of the statistical and other methodology. The conflation of multiple sources of insight and evidence is invariably necessary, in such studies, if conclusions are to carry conviction. Ignore the journalistic hype, obviously the responsibility of the second author, in the preamble to each chapter.

12.8 Further reading

See Rosenbaum (1999) and Rosenbaum (2002) for a comprehensive overview of issues that commonly arise in the analysis of observational data, and of approaches that may be available to handle some of the major sources of potential difficulty.

Part VI
Ordination

13 Examples, Theory and Overview

Ordination is a generic name for methods for providing a low-dimensionaL view of points in multi-dimensional space, such that "similar" objects are near each other and dissimilar objects are separated. The plot(s) from an ordination in 2 or 3 dimensions may provide useful visual clues on clusters in the data and on outliers.

Here, the discussion will turn to multi-dimensional scaling (MDS) methods where distances are given, or that start by calculating distances between points, then using the distances as the starting point for an ordination. Similarities can be transformed into distances, though often with some arbitrariness in the way that this is done.

Examples are:

1. From Australian road travel distances between cities and larger towns, can we derive a plausible "map" showing the relative geographic locations?

2. Starting with genomic data, various methods are available for calculating genomic "distances " between, e.g., different insect species. The distance measures are based on evolutionary models that aim to give distances between pairs of species that are a monotone function of the time since the two species separated.

3. Given a matrix $\mathbf{X}$ of n observations by p variables, a low-dimensional representation is required, i.e., the hope is that a major part of the information content in the data can be summarized in a small number of constructed variables. There is typically no good model, equivalent to the evolutionary models used by molecular biologists, that can be used to motivate distance calculations. There is then a large element of arbritariness in the distance measure used.

If data can be separated into known classes that should be reflected in any ordination, then the scores from classification using `lda()` may be a good basis for an ordination. Plots in 2 or perhaps 3 dimensions may then reveal additional classes and/or identify points that may be misclassified and/or are in some sense outliers. It may indicate whether the classes that formed the basis for the ordination seem real and/or the effectiveness of the discrimination method in choosing the boundaries between classes.

The function `randomForest()` is able to return "proximities" that are measures of the closeness of any pair of points. These can be turned into rough distance measures that can then form the basis for an ordination. With Support Vector Machines, decision values are available from which distance measures can be derived and used as a basis for ordination.

13.1 Distance measures

13.1.1 Euclidean distances

Treating the rows of $\mathbf{X}$ (n by p) as points in a p-dimensional space, the squared Euclidean distance d_{ij}^2 between points i and j is

$$d_{ij}^2 = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2$$

The distances satistfy the triangle inequality

$$d_{ij} \leq d_{ik} + d_{kj}$$

The columns of $\mathbf{X}$ can be arbitrarily transformed before calculating the d_{ij}. Where all elements of a column are positive, use of the logarithmic transformation is common. A logarithmic scale makes sense for biological morphometric data, and for other data that has similar characteristics. For morphometric data, the effect is to focus attention on relative changes in the various body proportions, ignoring the overall magnitude.

The columns may be standardized before calculating distances, i.e., scaled so that the standard deviation is one. The columns may be weighted differently. Use of an unweighted measure with all columns scaled to a standard deviation of one is equivalent to working with the unscaled columns and calculating d_{ij}^2 as

$$d_{ij}^2 = \sum_{k=1}^{p} w_{ij}(x_{ik} - x_{jk})^2$$

where $w_{ij} = (s_i s_j)^{-1}$ is the inverse of the product of the standard deviations for columns i and j. Results may depend strongly on the distance measure.

13.1.2 Non-Euclidean distance measures

Euclidean distance is one of many possible choices of distance measures, still satisfying the triangle inequality. As an example of a non-Euclidean measure, consider the Manhattan distance. This has

$$d_{ij} = \sum_{k=1}^{p} | x_{ik} - x_{jk} |$$

The Manhattan distance is the shorest distance for a journey that always proceeds along one of the co-ordinate axes. In Manhattan in New York, streets are laid out in a rectangular grid. This is then (with $k = 2$) the walking distance along one or other street. For other choices, see the help page for the function `dist()`.

The function `daisy()` in the *cluster* package offers a still wider range of possibilities, including distance measures that can be used when columns that are factor or ordinal. It has an argument `stand` that can be used to ensure standardization when distances are calculated. Unless measurements are comparable (e.g., relative growth, as measured perhaps on a logarithmic scale, for different body measurements), then it is usually desirable to standardize before using ordination methods to examine the data.

Irrespective of the method used for the calculation of the distance measure, ordination methods yield a representation in Euclidean space. Dedpending on the distance measure and the particular set of distances, an exact representation may or may not be possible.

See Gower & Legendre (1986) for a detailed discussion of the netric and Euclidean propoerties of a wide variety of similarity coefficients.

13.2 From distances to a configuration in Euclidean space

Here, we show how an **X**-matrix like representation, i.e., a representation in Euclidean space, can be recovered from a matrix of pairwise "distances" between points. The matrix that results will be written X, to distinguish it from any initial matrix $\mathbf{X}$ that has been the starting point for the calculation of distances.

The dimension (number of columns of X) may be as many as one less than the number of points. The only constraint on the "distances" is that they must satisfy the triangle inequality that was noted above, i.e., $d_{ij} \le d_{ik} + d_{kj}$.

Clearly the distances will be unaffected if the columns of X are centred so that all columns have mean 0. They will, also, be unaffected by arbitrary orthogonal rotation of the column space, i.e., replace X by $\mathsf{X}\mathbf{P}$, where $\mathbf{P}$ is an orthogonal matrix. Orthogonality implies that $\mathbf{P}^T\mathbf{P} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix.

The methodology that will now be described yields a matrix X whose columns are orthogonal, i.e., the pairwise innner product of any pair of columns is 0. Moreover, the columns can be (and are) ordered so that the successive columns explain successively larger proportion of the inter-point distances. Often, most of the information is in the first few columns.

Observe that, given X, the squared Euclidean distance between points i and j can be written

$$\begin{aligned} d_{ij}^2 &= \sum_{k=1}^{p} (\mathsf{x}_{ik} - \mathsf{x}_{jk})^2 \\ &= \sum_{k=1}^{p} \mathsf{x}_{ik}^2 + \sum_{k=1}^{p} \mathsf{x}_{jk}^2 - 2\sum_{k=1}^{p} \mathsf{x}_{ik}\mathsf{x}_{jk} \end{aligned}$$

Thus

$$d_{ij}^2 = q_{ii} + q_{jj} - 2q_{ij} \tag{1}$$

where $q_{ii} = \sum_{k=1}^{p} \mathsf{x}_{ik}^2$; $\quad q_{ij} = \sum_{k=1}^{p} \mathsf{x}_{ik}\mathsf{x}_{jk}$.

Observe that q_{ij} is the (i, j)th element of the matrix $\mathbf{Q} = \mathsf{X}\mathsf{X}'$. Thus, the matrix $\mathsf{X}\mathsf{X}'$ has all the information needed to construct distances.

Now require that columns of X are centered, i.e.

$$\sum_{i=1}^{n} \mathsf{x}_{ik} = 0, i = 1, \ldots p$$

This implies that

$$\begin{aligned}
\sum_{i=1}^{n} q_{ij} &= \sum_{i=1}^{n}(\sum_{k=1}^{p} \mathsf{x}_{ik}\mathsf{x}_{jk}) \\
&= \sum_{k=1}^{p}(\sum_{i=1}^{n} \mathsf{x}_{ik}\mathsf{x}_{jk}) \\
&= \sum_{k=1}^{p}(\mathsf{x}_{jk} \sum_{i=1}^{n} \mathsf{x}_{ik}) \\
&= 0
\end{aligned}$$

i.e., that the rows and columns of $\mathbf{Q}$ sum to zero.

13.2.1 Low-dimensional representation

It will now be shown that given distances d_{ij}, then equation 1 uniquely determines a matrix $\mathbf{Q}$ whose rows and columns sum to zero. The demand that the d_{ij} satisfy the triangle inequality is unfortunately not enough to guarantee that this matrix will be positive definite, as is required to yield a configuration that can be exactly embedded in Euclidean space.

Set $A = \sum_{i=1}^{n} q_{ii}$. Summing $d_{ij} = q_{ii} + q_{jj} - 2q_{ij}$ over i, it follows that

$$\sum_{i=1}^{n} d_{ij}^2 = A + nq_{jj} \tag{2}$$

$$\sum_{j=1}^{n} d_{ij}^2 = A + nq_{ii} \tag{3}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2 = 2nA \tag{4}$$

From equation 4

$$A = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2 \tag{5}$$

From equation 1, substituting for q_{ii} and q_{jj} from equations 2 and 3 above, and then for A from equation 5 above

$$\begin{aligned}
q_{ij} &= -\frac{1}{2}d_{ij}^2 + \frac{1}{2n}(\sum_{i=1}^{n} d_{ij}^2 + \sum_{j=1}^{n} d_{ij}^2 - 2A) \\
&= -\frac{1}{2}d_{ij}^2 + \frac{1}{2n}(\sum_{i=1}^{n} d_{ij}^2 + \sum_{j=1}^{n} d_{ij}^2 - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2)
\end{aligned}$$

Having thus recovered a symmetric matrix $\mathbf{Q}$, the spectral decomposition yields

$$\mathbf{Q} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$$

where $\mathbf{\Lambda}$ is a diagonal matrix. The diagonal elements λ_i are ordered so that

$$\lambda_1 \geq \lambda_2 \ldots \geq \lambda_n$$

As the rows and columns of $\mathbf{Q}$ sum to zero, $\mathbf{Q}$ is singular. Hence if $\mathbf{Q}$ is positive definite, as required for exact embedding in Euclidean space, $\lambda_i \geq 0$ for all λ_i and $\lambda_n = 0$.

Two important points are:

- Often, most of the information will be in the first few dimensions. We may for example be able to approximate $\mathbf{Q}$ by replacing $\mathbf{\Lambda}$ in $\mathbf{Q} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ by a version of $\mathbf{\Lambda}$ in which diagonal elements after the kth have been set to zero. If `cmdscale()` is called with `eig=TRUE`, it returns both the eigenvalue information (the λ_i) and a goodness of fit statistic, by default (assuming at least two non-zero λ_i) for the configuration with $k = 2$.
- If $\mathbf{Q}$ is not positive semidefinite, the ordination can still proceed. However one or more eigenvalues λ_i will now be negative. If relatively small, it may be safe to ignore dimensions that correspond to negative eigenvalues. It is then more than otherwise desirable to check that the ordination reproduces the distances with acceptable accuracy.

13.2.2 The connection with principal components

Let $\mathbf{X}$ be a matrix that is the basis for the calculation of Euclidean distances, after any transformations and/or weighting. Then metric p-dimensional ordination, applied to Euclidean distances between the rows of $\mathbf{X}$, yields an orthogonal transformation of the space spanned by the columns of $\mathbf{X}$. If the successive dimensions are chosen to "explain" successively larger proportions of the trace of $\mathbf{X}\mathbf{X}^T$, it is equivalent to the principal components transformation. Thus `cmdscale()` yields, by a different set of matrix manipulations, a principal components decomposition.

13.3 Non-metric scaling

These methods all start from "distances", but allow greater flexibility in their use to create an ordination. The aim is to represent the "distances" in as few dimensions as possible. AS described here, a first step is to treat the distances as Euclidean, and determine a configuration in Euclidean space. These Euclidean distances are then used as a starting point for a representation in which the requirement that these are Euclidean distances, all determined with equal accuracy, is relaxed. The methods that will be noted here are:

Sammon scaling: A configuration with distances $\tilde{d}$ is chosen to minimize a weighted squared "stress"

$$\frac{1}{\sum_{i\neq j} d_{ij}} \sum_{i\neq j} \frac{(d_{ij} - \tilde{d}_{ij})^2}{d_{ij}}$$

Kruskal's non-metric multidimensional scaling: This aims to minimize

$$\frac{\sum_{i\neq j}(\theta(d_{ij}) - \tilde{d}_{ij})^2}{\sum_{i\neq j} \tilde{d}_{ij}^2}$$

with respect to the configuration of points and an increasing function θ of the distance d_{ij}.

Often, it makes sense to give greater weight to small distances than to large distances. The distance scale should perhaps not be regarded as rigid. Larger distances may not be measured on the same Euclidean scale as shorter distances. The ordination should perhaps preserve relative rather than absolute distances.

13.4 Examples

13.4.1 Australian road distances

The distance matrix that will be used is in the matrix `audists`, in the image file **audists.Rdata**. Consider first the use of classical multi-dimensional scaling, as implemented in the function `cmdscale()`:

```
> library(DAAGxtras)
> aupoints <- cmdscale(audists)
> plot(aupoints)
> text(aupoints, labels = paste(rownames(aupoints)))
```

An alternative to `text(aupoints, labels=paste(rownames(aupoints)))`, allowing better placement of the labels, is `identify(aupoints, labels=rownames(aupoints))`. We can compare the distances in the 2-dimensional representation with the original road distances:

```
> audistfits <- as.matrix(dist(aupoints))
> misfit <- as.matrix(dist(aupoints)) - as.matrix(audists)
> for (j in 1:9) for (i in (j + 1):10) {
+     lines(aupoints[c(i, j), 1], aupoints[c(i, j), 2], col = "gray")
+     midx <- mean(aupoints[c(i, j), 1])
+     midy <- mean(aupoints[c(i, j), 2])
+     text(midx, midy, paste(round(misfit[i, j])))
+ }
> colnames(misfit) <- abbreviate(colnames(misfit), 6)
> print(round(misfit))
```

```
          Adelad Alice Brisbn Broome Cairns Canbrr Darwin Melbrn Perth Sydney
Adelaide       0   140   -792   -156    366     20     11     82   482   -273
Alice        140     0  -1085   -175    -41     76   -118    106   -26   -314
Brisbane    -792 -1085      0    198    319    -25   -233   -471   153    -56
Broome      -156  -175    198      0    527     -7      6    -65   990     70
Cairns       366   -41    319    527      0    277    -31    178     8    251
Canberra      20    76    -25     -7    277      0     -1   -241   372     -8
Darwin        11  -118   -233      6    -31     -1      0    -12    92    -58
Melbourne     82   106   -471    -65    178   -241    -12      0   301   -411
Perth        482   -26    153    990      8    372     92    301     0    271
Sydney      -273  -314    -56     70    251     -8    -58   -411   271      0
```

The graph is a tad crowded, and for detailed information it is necessary to examine the table.

It is interesting to overlay this "map" on a physical map of Australia.

```
> library(oz)
> oz()
> points(aulatlong, col = "red", pch = 16, cex = 1.5)
> comparePhysical <- function(lat = aulatlong$latitude, long = aulatlong$longitude,
+     x1 = aupoints[, 1], x2 = aupoints[, 2]) {
+     fitlat <- predict(lm(lat ~ x1 + x2))
+     fitlong <- predict(lm(long ~ x1 + x2))
+     x <- as.vector(rbind(lat, fitlat, rep(NA, 10)))
+     y <- as.vector(rbind(long, fitlong, rep(NA, 10)))
+     lines(x, y, col = 3, lwd = 2)
+ }
> comparePhysical()
```

An objection to `cmdscale()` is that it gives long distances the same weight as short distances. It is just as prepared to shift Canberra around relative to Melbourne and Sydney, as to move Perth. It makes more sense to give reduced weight to long distances, as is done by `sammon()` (*MASS*).

```
> aupoints.sam <- sammon(audists)
```

```
Initial stress        : 0.01573
stress after  10 iters: 0.00525, magic = 0.500
stress after  20 iters: 0.00525, magic = 0.500
```

```
> oz()
> points(aulatlong, col = "red", pch = 16, cex = 1.5)
> comparePhysical(x1 = aupoints.sam$points[, 1], x2 = aupoints.sam$points[,
+     2])
```

Notice how Brisbane, Sydney, Canberra and Melbourne now maintain their relative positions much better.

Now try full non-metric multi-dimensional scaling (MDS). This preserves only, as far as possible, the relative distances. A starting configuration of points is required. This might come from the configuration used by `cmdscale()`. Here, however, we use the physical distances.

```
> oz()
> points(aulatlong, col = "red", pch = 16, cex = 1.5)
> aupoints.mds <- isoMDS(audists, as.matrix(aulatlong))

initial  value 11.875074
iter   5 value 5.677228
iter  10 value 4.010654
final  value 3.902515
converged

> comparePhysical(x1 = aupoints.mds$points[, 1], x2 = aupoints.mds$points[,
+     2])
```

Notice how the distance between Sydney and Canberra has been shrunk quite severely.

13.4.2 Genetic Distances – Hasegawa's selected primate sequences

Here, matching genetic DNA or RNA or protein or other sequences are available from each of the different species. Distances are based on probabilistic genetic models that describe how gene sequences change over time. The package *ape* implements a number of alternative measures. For details see `help(dist.ape)`.

Hasegawa's sequences were selected to have as little variation in rate, along the sequence, as possible. The sequences are available from:
`http://evolution.genetics.washington.edu/book/primates.dna`. They can be read into R as:

```
> library(ape)
> webpage <- "http://evolution.genetics.washington.edu/book/primates.dna"
> primates.dna <- read.dna(con <- url(webpage))
> close(con)
```

Now calculate distances, using Kimura's F84 model, thus

```
> primates.dist <- dist.dna(primates.dna, model = "F84")
```

We now try for a two-dimensional representation, using `cmdscale()`.

```
> primates.cmd <- cmdscale(primates.dist)
> eqscplot(primates.cmd)
> rtleft <- c(4, 2, 4, 2)[unclass(cut(primates.cmd[, 1], breaks = 4))]
> text(primates.cmd[, 1], primates.cmd[, 2], row.names(primates.cmd),
+     pos = rtleft)
```

Now see how well the distances are reproduced:

```
> d <- dist(primates.cmd)
> sum((d - primates.dist)^2)/sum(primates.dist^2)

[1] 0.1977138
```

This is large enough (20%, which is a fraction of the total sum of squares) that it may be worth examining a 3-dimensional representation.

```
> library(lattice)
> primates.cmd <- cmdscale(primates.dist, k = 3)
> cloud(primates.cmd[, 3] ~ primates.cmd[, 1] * primates.cmd[,
+     2])
> d <- dist(primates.cmd)
> sum((d - primates.dist)^2)/sum(primates.dist^2)

[1] 0.1045168
```

Now repeat the above with `sammon()` and `mds()`.

```
> primates.sam <- sammon(primates.dist, k = 3)

Initial stress        : 0.11291
stress after  10 iters: 0.04061, magic = 0.461
stress after  20 iters: 0.03429, magic = 0.500
stress after  30 iters: 0.03413, magic = 0.500
stress after  40 iters: 0.03409, magic = 0.500

> eqscplot(primates.sam$points)
> rtleft <- c(4, 2, 4, 2)[unclass(cut(primates.sam$points[, 1],
+     breaks = 4))]
> text(primates.sam$points[, 1], primates.sam$points[, 2], row.names(primates.sam$points),
+     pos = rtleft)
```

There is no harm in asking for three dimensions, even if only two of them will be plotted.

```
> primates.mds <- isoMDS(primates.dist, primates.cmd, k = 3)

initial  value 19.710924
iter   5 value 14.239565
iter  10 value 11.994621
iter  15 value 11.819528
iter  15 value 11.808785
iter  15 value 11.804569
final  value 11.804569
converged

> eqscplot(primates.mds$points)
> rtleft <- c(4, 2, 4, 2)[unclass(cut(primates.mds$points[, 1],
+     breaks = 4))]
> text(primates.mds$points[, 1], primates.mds$points[, 2], row.names(primates.mds$points),
+     pos = rtleft)
```

13.4.3 Pacific rock art

Here, the the 614 features were all binary – the presence or absence of specific motifs in each of 98 Pacific sites. (Actually, there were 103 sites, but 5 were omitted because they had no motifs in common with any of the other sites.) Data are from Meredith Wilson's PhD thesis at Australian National University.

The binary measure of distance was used – the number of locations in which only one of the sites had the marking, as a proportion of the sites where one or both had the marking. Here then is the calculation of distances:

```
> pacific.dist <- dist(x = as.matrix(rockArt[-c(47, 54, 60, 63,
+     92), 28:641]), method = "binary")
> sum(pacific.dist == 1)/length(pacific.dist)
```

```
[1] 0.6311803

> plot(density(pacific.dist, to = 1))
> symmat <- as.matrix(pacific.dist)
> table(apply(symmat, 2, function(x) sum(x == 1)))

13 21 27 28 29 32 33 35 36 38 40 41 42 43 44 45 46 47 48 49 51 52 53 54 55 56
 1  1  1  1  2  1  2  1  2  2  1  2  4  3  1  3  1  2  1  1  2  2  3  2  2  2
57 58 61 62 64 65 66 67 68 69 70 71 73 75 76 77 79 81 83 84 85 90 91 92 93 94
 1  3  3  1  2  1  1  1  3  3  1  1  4  1  2  1  1  1  2  1  1  3  1  1  3  1
95 96 97
 1  3  4
```

It turns out that 63% of the distances were 1. This has interesting consequences, for the plots we now do.

```
> pacific.cmd <- cmdscale(pacific.dist)
> plot(pacific.cmd)
> pacific.sam <- sammon(pacific.dist, pacific.cmd)

Initial stress        : 0.58369
stress after  10 iters: 0.41996, magic = 0.018
stress after  20 iters: 0.22171, magic = 0.213
stress after  30 iters: 0.18573, magic = 0.098
stress after  40 iters: 0.16241, magic = 0.225
stress after  50 iters: 0.15903, magic = 0.225
stress after  60 iters: 0.15786, magic = 0.500
stress after  70 iters: 0.15734, magic = 0.500
stress after  80 iters: 0.15717, magic = 0.500
stress after  90 iters: 0.15700, magic = 0.500
stress after 100 iters: 0.15687, magic = 0.338

> plot(pacific.sam$points)
```

Part VII
*Some Further Types of Model

14 *Multilevel Models – Introductory Notions

Basic ideas of multilevel modeling will be illustrated using data on yields from packages on eight sites on the Caribbean island of Antigua. They are a summarized version of a subset of data given in Andrews and Herzberg 1985, pp.339-353.

Multilevel models break away from the assumption of independently and identically distributed observations. The dependence is however of a very specific form. Models for time series move away from those assumptions in a different way, typically allowing some form of sequential correlation.

Depending on the use that will be made of the results, it may be essential to correctly model the structure of the random part of the model. The analysis will use the abilities of the `lme()` function in the *nlme* package, though the example is one where it is easy, using modest cunning, to get the needed sums of squares from a linear model calculation. For these data, there is more than one type (or "level") of prediction or generalization, with very different accuracies for the different generalizations. The data give results for each of several packages at a number of different locations (sites). In such cases, a prediction for a new package at one of the existing locations is likely to be more accurate than

a prediction for a totally new location. Multi-level models are able to account for such differences in predictive accuracy.

The multiple levels that are in view are multiple levels in the *noise* or *error* term, and are superimposed on any effects that are predictable. For example, differences in historical average annual rainfall may partly explain location to location differences in crop yield. The error term in the prediction for a new location will account for variation that remains after taking account of differences in the rainfall.

Examples abound where the intended use of the data makes a multi-level model appropriate. Examples of two levels of variability, at least as a first approximation, include: variation between houses in the same suburb, as against variation between suburbs; variation between different clinical assessments of the same patients, as against variation between patients; variation within different branches of the same business, as against variation between different branches; variations in the bacterial count between different samples from the same lake, as opposed to variation between different subsamples of the same sample; variation between the drug prescribing practices of clinicians in a particular specialty in the same hospital, as against variation between different clinicians in different hospitals; and so on. In all these cases, the accuracy with which predictions are possible will depend on the mix of the two levels of variability that are involved. These examples can all be extended in fairly obvious ways to include more than two levels of variability.

In all the examples just mentioned, one source of variability is *nested* within the other – thus packages of land are nested within locations. Variation can also be *crossed*. For example different years may be crossed with different locations. Years are not nested in locations, nor are locations nested in years. Examples of crossed error structures are beyond the scope of the present discussion.

14.1 The Antiguan Corn Yield Data

For the version of the Antiguan corn data presented here, the hierarchy has two levels of *random effects*. Variation between packages in the same site is at the lower of the two levels, and is called level 0 in the later discussion. Variation between sites is the higher of the two levels, and is called level 1 in the later discussion. A farmer who lived close to one of the experimental sites might take data from that site as indicative of what to expect. Other farmers may think it more appropriate to regard their farms as new sites, distinct from the experimental sites, so that the issue is one of generalizing to new sites.

The analysis will use the `lme()` function in the *nlme* package, though the example is one where it is easy, using modest cunning, to get the needed sums of squares from a linear model calculation.

The data that will be analyzed are in the second column of Table 9, which has means of packages of land for the Antiguan data. In comparing yields from different packages, there are two sorts of comparison. Packages on the same site should be relatively similar, while packages in different sites should be relatively more different. The figure that was given earlier suggested that this is indeed the case.

Note: In an analysis of variance formalization, the two-level structure of variation is handled by splitting variation, as measured by the total sum of squares about the grand mean, into two parts – variation within sites, and variation between site means. The final two columns in Table 9 indicate how to calculate the relevant sums of squares and (by dividing by degrees of freedom) mean squares. The division of the sum of squares into two parts mirrors two different types of predictions that can be based on these data. First, suppose that we are interested in another package on one of these same sites. Within what range of variation would we expect its yield to lie? Second, suppose that a trial were to be carried out on some different site, not one of the original eight. What is the likely range of variation of the mean yield, i.e., how accurate is the accuracy of prediction of the yield for that new site?

The model

The model that is used is:

$$\text{yield} = \text{overall mean} + \begin{matrix}\text{site effect}\\(\text{random})\end{matrix} + \begin{matrix}\text{package effect}\\(\text{random})\end{matrix}$$

Site	Site means	Site effect		Residuals from site mean
DBAN	5.16, 4.8, 5.07, 4.51		+0.59	0.28, −0.08, 0.18, −0.38
LFAN	2.93, 4.77, 4.33, 4.8		−0.08	−1.28, 0.56, 0.12, 0.59
NSAN	1.73, 3.17, 1.49, 1.97		−2.2	−0.36, 1.08, −0.6, −0.12
ORAN	6.79, 7.37, 6.44, 7.07	(4.29)	+2.62	−0.13, 0.45, −0.48, 0.15
OVAN	3.25, 4.28, 5.56, 6.24		+0.54	−1.58, −0.56, 0.73, 1.4
TEAN	2.65, 3.19, 2.79, 3.51		−1.26	−0.39, 0.15, −0.25, 0.48
WEAN	5.04, 4.6, 6.34, 6.12		+1.23	−0.49, −0.93, 0.81, 0.6
WLAN	2.02, 2.66, 3.16, 3.52		−1.45	−0.82, −0.18, 0.32, 0.68
v		square, add, multiply by 4, divide by d.f.=7, to give ms		square, add, divide by d.f.=24, to give ms

Table 9: The leftmost column has harvest weights (`harvwt`), for the packages in each site, for the Antiguan corn data. Each of these harvest weights can be expressed as the sum of the overall mean (= 4.29), site effect (third column), and residual from the site effect (final column). This information that can be used to create the analysis of variance table. (Details of the analysis of variance approach to analysis of these data, although straightforward, get only passing mention in these notes.)

In formal mathematical language:

$$y_{ij} = \mu + \underset{\text{(site, random)}}{\alpha_i} + \underset{\text{(package, random)}}{\beta_{ij}} \qquad (i = 1, \ldots, 8; j = 1, \ldots, 4)$$

with $\text{var}[\alpha_i] = \sigma^2_L$, $\text{var}[\beta_{ij}] = \sigma^2_B$.

The quantities σ^2_L and σ^2_B are known, technically, as *variance components*. (Those who are familiar with the analysis of variance breakdown may wish to note that the variance components analysis allows inferences that are not immediately available from the breakdown of the sums of squares in the analysis of variance table.) Importantly, the variance components provide information that can help design another experiment.

14.2 The variance components

Here is how the variance components should be interpreted, for the Antiguan data:

- Variation between packages at a site is due to one source of variation only. Denote this variance by σ^2_B. The variance of the difference between two such packages is $2\sigma^2_B$
  [Both packages have the same site effect α_i, so that $\text{var}(\alpha_i)$ does not contribute to the variance of the difference.]

- Variation between sites in different plots is partly a result of variation between packages, and partly a result of additional variation between sites. In fact, if σ^2_L is the (additional) component of the variation that is due to variation between sites, the variance of the difference between two packages that are in different site is
  $$2(\sigma^2_L + \sigma^2_B)$$

- For s single package, the variance is $\sigma^2_L + \sigma^2_B$. The variance of the estimate of the site mean is a mean over the four packages at the one site, and is
  $$\sigma^2_B + \frac{\sigma^2_L}{4}$$

[Notice that while σ_L^2 is divided by four, σ_B^2 is not. This is because the site effect is the same for all four packages.]

15 *Survival models

Survival (or failure) analysis introduces features different from any of those encountered in the regression methods discussed in earlier chapters. It has been widely used for comparing the times of survival of patients suffering a potentially fatal disease who have been subject to different treatments. Computations can be handled in R using the *survival* package, written for S-PLUS by Terry Therneau, and ported to R by Thomas Lumley.

Section 12.4 discusses an example that is onveniently handled using survival models.

Other names, mostly used in non-medical contexts, are *Failure Time Analysis* and *Reliability*. Yet another term is *Event History Analysis*. The focus is on time to any event of interest, not necessarily failure. It is an elegant methodology that is too little known outside of medicine and industrial reliability testing.

Applications include:

- the failure time distributions of industrial machine components, electronic equipment, automobile components, kitchen toasters, light bulbs, businesses, etc. (failure time analysis, or reliability),
- the waiting time to germination of seeds, to marriage, to pregnancy, or to getting a first job,
- the waiting time to recurrence of an illness or other medical condition.

The outcomes are survival times, but with a twist. The methodology is able to handle data where failure (or another event of interest) has, for a proportion of the subjects, not occurred at the time of termination of the study. It is not necessary to wait till all subjects have died, or all items have failed, before undertaking the analysis! Censoring implies that information about the outcome is incomplete in some respect, but not completely missing. For example, while the exact point of failure of a component may not be known, it may be known that it did not survive more than 720 hours (= 30 days). In a clinical trial, there may for some subjects be a final time up to which they survived, but no subsequent information. Such observations are said to be right censored.

Thus, for each observation there are two items of information: a time, and censoring information. Commonly the censoring information indicates either right censoring denoted by 0, or failure denoted by 1.

Many of the same issues arise as in more classical forms of regression analysis. One important set of issues has to do with the diagnostics used to check on assumptions. Here there have been large advances in recent years. A related set of issues has to do with model choice and variable selection. There are close connections with variable selection in classical regression. Yet another set of issues has to do with the incomplete information that is available when there is censoring.

Yang & Letourneau (2005) is an interesting example of a data mining paper where survival methods could and should have been used. The methodology may be regarded as an unsatisfactory attempt to reinvent survival methods! Their methodology is tortuous and does not make the most effective use of the data.

Part VIII
Technical Mathematical Results

16 Least Squares Estimates

16.1 The mean is a least squares estimator

The `lm()` function uses the method of least square to find estimates. The following is the simplest possible example. Given sample values

$$y_1, y_2, \ldots, y_n$$

what choice of μ will minimize $\sum_{i=1}^{n}(x_i - \mu)^2$? Observe that

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\mu)^2 &= \sum_{i=1}^{n}[(x_i-\bar{x})+(\bar{x}-\mu)]^2 \\ &= \sum_{i=1}^{n}[(x_i-\bar{x})^2+2(x_i-\bar{x})(\bar{x}-\mu)+(\bar{x}-\mu)^2] \\ &= \sum_{i=1}^{n}(x_i-\bar{x})^2+2(\bar{x}-\mu)\sum_{i=1}^{n}(x_i-\bar{x})+n(\bar{x}-\mu)^2\end{aligned}$$

As

$$\sum_{i=1}^{n}(x_i-\bar{x}=\sum_{i=1}^{n}x_i-n\bar{x}=0$$

this equals

$$\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\mu)^2$$

Then $n(\bar{x}-\mu)^2 >= 0$, with equality for $\mu = \hat{\mu} = \bar{x}$.

Because $\bar{x}$ is the least squares estimator of μ, it is possible to use a linear model to calculate the mean. For this, a model is specified in which the only term is the constant term. Thus, for the female Adelaide statistics students:

```
library(MASS)
y <- na.omit(survey[survey$Sex=="Female", "Height"])
lm(y ~ 1)
```

16.2 Least squares estimates for linear models

Given the model

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

the least squares estimate $\mathbf{b}$ of β is obtained by solving the normal equation

$$\mathbf{X}'\mathbf{X}\beta = \mathbf{X}'\mathbf{y}$$

In practice it is usually best not to solve this equation directly, but to work from the QR orthogonal decomposition of $\mathbf{X}$. For details, see the references that appear on the help page for R's function `qr()`.

16.3 Beyond Least Squares – Maximum Likelihood

Least squares may not work very well for non-normal data. Typically, statisticians then appeal to the maximum likelihood principle. For normal data, with independent and identically distributed errors, maximum likelihood gives the same parameter estimates as least squares. Section 8 has brief notes on two types of model where it really is necessary to work with maximum likelihood estimates.

17 Variances of Sums and Differences

The needed results are most easily derived using expectation algebra. For present purposes, it will be adequate to define

$$\mathrm{E}[g(X)] = \int g(x)f(x)dx$$

if X is a continuous random variable with density $f(x)$ at the point x, and

$$\mathrm{E}[g(X)] = \sum g(x)\Pr(X = x)$$

where the integral or sum is taken over the support of X. The key result from expectation algebra is that, for any two random variables X and Y, $\mathrm{E}[c_1X + c_2X] = c_1\mathrm{E}[X] + c_2\mathrm{E}[Y]$. The proof, for two special cases noted above, is left as an exercise.

The variance of a random variable X with mean $\mu = \mathrm{E}[X]$ is $\mathrm{E}(y - \mu)^2$. Then

$$\mathrm{var}[X_1 + X_2] = \mathrm{var}[X_1] + \mathrm{var}[X_2] + 2\mathrm{cov}[X_1, X_2]$$

which equals $\mathrm{var}[X_1] + \mathrm{var}[X_2]$ if and only if

$$\mathrm{cov}[X_1, X_2] = \mathrm{E}[(X_1 - \mathrm{E}[X_1])(X_2 - \mathrm{E}[X_2])] = 0$$

A very similar argument shows that $\mathrm{var}[X_1 - X_2] = \mathrm{var}[X_1] + \mathrm{var}[X_2]$ if and only if $\mathrm{cov}[X_1, X_2] = 0$.

A sufficient condition for $\mathrm{cov}[X_1, X_2] = 0$ is that X_1 and X_2 are independent.

18 References

References

Ambroise, C. and McLachlan, G.J. 2001. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences USA*, **99** 6562-6566.

Bates, D. 2007. Comparing least squares calculations. *Vignette "Comparisons" accompanying the package "Matrix" for R.*

Blackard, Jock A. 1998. Comparison of Neural Networks and Discriminant Analysis in Predicting Forest Cover Types. Ph.D. dissertation. Department of Forest Sciences. Colorado State University. Fort Collins, Colorado.
[Data are available from <URL:\http://www.ics.uci.edu/~mlearn/MLRepository.html>]

Bland, M. & Altman, D. 2005. Do the left-handed die young? *Significance*, 2:166–170.

Breiman, L. 2001. Statistical modeling: the two cultures (with discussion). *Statistical Science* **16** 199- 231.
[This is a controversial paper whose major claims are, in my view and in that of at least one of the discussants, nonsense. It, and the subsequent discussion are a good read.]

Britton, A., & Marmot, M. 2004. Different measures of alcohol consumption and risk of coronary heart disease and all-cause mortality: 11-year follow-up of the Whitehall II Cohort Study. *Addiction* **99**:109–116.

Carroll, R.J. 2004. Measuring diet. Texas A & M Distinguished Lecturer series.
[Data are available from <URL:http://stat.tamu.edu/~carroll/talks.php]

Chambers, J.M. 2000. Users, Programmers, and Statistical Software. *ASA Journal of Computational and Graphical Statistics* 9:3 (September, 2000), pp. 404-422.
[Discusses issues that are of importance when software systems are used for data analysis, and how these should affect the design of statistical software systems. In the R project, John Chambers how has a number of very able statistical computing specialists involved with him in thinking through such issues, and to encoding in software the ideas that emerge.]

Cox, D.R. and Solomon, P.J. 2003. *Components of Variance.* Chapman and Hall.
[Multi-level models are, as usually formulated, components of variance models.]

Dalgaard, P. 2002. *Introductory Statistics with R.* Springer-Verlag, New York.
[This is an introductory account of the use of the R language for statistical analysis, with a slant towards biostatistical applications.]

Davey Smith, G. & Ebrahim, S. 2005. What can mendelian randomisation tell us about modifiable behavioural and environmental exposures? British Medical Journal **330**:1076 - 1079.

Fuller, W. A. 1987. *Measurement Error Models*. Wiley.

. Gower, J. C. & Legendre, P. 1986. Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification* **3:** 5-48. Jackson, R., Broad, J., Connor, J. and Wells, S. 2001. Alcohol and ischaemic heart disease: probably no free lunch. The Lancet 366: 1911-1912.

Han, J. and Kamber, M. 2001. *Data Mining: Concepts and Techniques.* Morgan Kaufmann.
[This is a widely used data mining text.]

Hand, J., Blunt, G., Kelly, M.G. and Adams, N.M. 2000. Data mining for fun and profit (with discussion). *Statistical Science* **15:** 111-131.
[This gives a statistical perspective on data mining.]

Hand, D.J. 2006. Classifier technology and the illusion of progress. *Statistical Science* **21: 15:** 1-14, and (comment) 15-34.

Hand, D., Mannila, H. and Smyth, P. 2001. *Principles of Data Mining.* MIT Press.
[While better than other treatments of statistical issues that I have seen in data mining texts, there are nevertheless serious gaps in its treatment.]

Jackson, R., Broad, J., Connor, J. and Wells, S. 2001. Alcohol and ischaemic heart disease: probably no free lunch. The Lancet 366: 1911-1912.

Koenker, R and Ng, P 2003. SparseM: A sparse matrix package for R. *Journal of Statistical Software* 8(6).

Latter, O. H., 1902. The egg of *cuculus canorus*. an inquiry into the dimensions of the cuckoo's egg and the relation of the variations to the size of the eggs of the foster-parent, with notes on coloration, &c. *Biometrika*, 1:164–176.

Leavitt, S. D. and Dubner, S. J. 2005. *Freakonomics. A Rogue Economist Explores the Hidden Side of Everything*. William Morrow.

Maindonald, J. H. 2003. The role of models in predictive validation. Invited Paper.

Maindonald, J.H. 2004a. Computation and biometry. In Modern Biometry, from Encyclopedia of Life Support Systems (EOLSS), Developed under the Auspices of the UNESCO, Eolss Publishers, Oxford, UK, `http://www.eolss.net`

Maindonald, J.H. 2004b. Statistical Computing. In Modern Biometry, from Encyclopedia of Life Support Systems (EOLSS), Developed under the Auspices of the UNESCO, Eolss Publishers, Oxford, UK, `http://www.eolss.net`.
[This has my view of major current directions in statistical computing software development.]

Maindonald, J.H. 2005. Data, science, and new computing technology. New Zealand Science Review 62: 126-128.

Maindonald, J.H. 2006. Data Mining Methodological Weaknesses and Suggested Fixes. Proceedings of Australasian Data Mining Conference (Aus06), Sydney, Nov 29-30, 2006.
`http://www.maths.anu.edu.au/~johnm/dm/ausdm06/ausdm06-jm.pdf` (paper)
`http://wwwmaths.anu.edu.au/~johnm/dm/ausdm06/ohp-ausdm06.pdf` (overheads)

MAINDONALD, J. H. AND BRAUN, W.J. 2007. *Data Analysis and Graphics Using R – An Example-Based Approach.* 2^{nd} edition, Cambridge University Press.
`<URL:http://www.maths.anu.edu.au/~johnm/r-book.html>`
[This is aimed at practicing scientists who have some modest statistical sophistication, and at statistical practitioners. It demonstrates the use of the R system for data analysis and for graphics.]

MAINDONALD, J.H., WADDELL, B.C. AND PETRY, R.J. 2001. Apple cultivar effects on codling moth (Lepidoptera: Tortricidae) egg mortality following fumigation with methyl bromide. Postharvest Biology and Technology **22** 99-110.

MEYER, M.C. AND FINNEY, T. 2005. Who wants airbags?. *Chance* **18**:3-16.

NEWTON, A. 1893-1896. Cuckoos. In A. Newton and H. Gadow, eds., *Dictionary of Birds*. A. and C. Black.

R CORE DEVELOPMENT TEAM. *An Introduction to* R. Supplied with most installations of R, and available also from CRAN sites (`<URL:http://cran.r-project.org>` gives the list of sites).

RIPLEY, B. D. 1996. *Pattern Recognition and Neural Networks*. Cambridge University Press.

ROSENBAUM, P.R. 1999. Choice as an alternative to control in observational studies. *Statistical Science* **14** 259-278, with following discussion, pp. 279-304.

ROSENBAUM, P.R. 2002. *Observational Studies*, 2nd edn. Springer-Verlag.
[This is an important recourse and source of insight for anyone who works with observational data.]

SCHATZKIN, A., KIPNIS, V., CARROLL R.J., MIDTHUNE, D., SUBAR, A.F., BINGHAM, S., SCHOELLER D.A., TROIANO, R.P. AND FREEDMAN, L.S. 2003. A comparison of a food frequency questionnaire with a 24-hour recall for use in an epidemiological cohort study: results from the biomarker-based Observing Protein and Energy Nutrition (OPEN) study. *International Journal of Epidemiology* 32: 1054-1062.

ROSSOUW, J.E., ANDERSON, G.L., PRENTICE, RL, ET AL. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial. *Journal of the American Medical Association* **2002**:288:321.

SENN, S., 2003. *Dicing with Death: Chance, Risk and Health*. Cambridge University Press.

TORGO, L. 2003. *Data Mining with R.* (Available from `<URL:http://www.liacc.up.pt/~ltorgo>`)
[This has a data mining flavour. There is a brief discussion of databases. The second of the data sets (a stock market time series) is available as a MySQL database. This may be a good way to start learning about the interface that the *RODBC* package offers to MySQL. The reliance on the R `source()` command for storage and entry of data is not a good idea, in general. Use image (**.RData**) files instead. Comments on statistical issues, and notably on the handling of missing data, suggest approaches that, while widely used in the past, are known to have serious potential problems.]

VENABLES, W. N. AND RIPLEY, B. D. 2002. *Modern Applied Statistics with* S. Springer-Verlag, 4 edition. See also R Complements to Modern Applied Statistics with S.
`http://www.stats.ox.ac.uk/pub/MASS4/`
[Note especially pp.331–341 (lda and qda) and pp.187–198 (logistic and other GLMs). In the third edition, see pap.344-354 and pp.211-226]

WILKINSON, G. N. & ROGERS, C. E. 1973. *Symbolic description of models in analysis of variance.* Applied Statistics 22: 392-399.

WITTEN, I.H. AND FRANK, E. 2000. *Data Mining. Practical machine Learning Tools and Techniques with Java Implementations.* Morgan Kaufmann.
[This is a popular data mining text.]

Wood, S. N. 2006. *Generalized Additive Models.* An Introduction with R. Chapman & Hall/CRC. [This has an elegant treatment of linear models and generalized linear models, as a lead-in to generalized additive models.]

Yang, C, & Letourneau, S.(2005) Learning to Predict Train Wheel Failures. The Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2005). Chicago, Illinois, USA. August 21-22, 2005. NRC 48130.
`iit-iti.nrc-cnrc.gc.ca/iit-publications-iti/docs/NRC-48130.pdf`

Young, G. and Smith, R. L. 2005. *Essentials of Statistical Inference*. Cambridge University Press.