Statistical Perspectives on Data Mining

John Maindonald

July 28, 2006

Abstract

This document identifies statistical issues that can be and commonly are important for data mining problems. As far as possible, it will avoid the technical language of mathematical statistics.

Key issues for any data analysis are:

1. Why are we 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?

Finally, when results are presented, the data analyst should be well placed to answer the question: "What is the relevance of these results?"

Part I discusses statistical issues and ideas. Examples, most of them taken from published data, highlight the importance of these issues. Remaining parts of the document summarize important results and concepts from classical statistics. More recent data mining methodologies supplement rather than displace the classical methodologies. Models used in data mining often build on or incoporate classical models.

The source population for the data is rarely, for observational data, identical to the target population to which results will be applied. A model that is effective in describing the processes that generated the data rarely provides a totally accurate description for the processes that will apply when use is made of model results. This has large implications for the assessment of predictive accuracy.

Multi-level modeling (Part IV), where the relevant data are available, is sometimes helpful in the search for a way around this difficulty, at least to the extent of taking account of different sources of variation, and the consequent dependence of predictive accuracy on the nature of the intended generalization. The need to tune predictive accuracy calculations to the target population has not received much attention in the data mining literature.

There are 11 accompanying sets of laboratory exercises that use the R system for computing and graphics.

Contents

Part I
Statistical Issues and Ideas

1 Preliminaries

1.1 Statistics versus data mining – what is the connection?

Technological advance 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 application area specialists whose statistical knowledge has not advanced much in the past decade, and those professionals who are fully au fait with modern methods.
- New statistical technologies that combine data from multiple studies in a single analysis may allow the detection of patterns that were not apparent from the individual studies. They may resolve apparent discrepancies between results from the separate analyses.

Daryl Pregibon's has defined data mining as "Statistics at scale and speed". This may be as apt as any definition that is available. 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. Arguably, this does not give enough weight to areas of investigation in which researchers from a computer science tradition have taken a lead.

More cynical definitions are possible!

> A combination of large databases and bad statistics.
>
> Statistics plus marketing.

I do not accept distinctions that focus on the nature of the data analysis enterprise. Statistical issues have a very wide relevance.

Comments in Witten and Frank (2000), in respect of 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).

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.

The issues noted in this module are important for all data analysts, whether they call themselves statisticians or data miners.

1.2 Statistical theory

There is an extensive statistical theory that offers commentary on issues of model choice and the properties of estimators. This module will be unable to use this theory to any substantial extent, thus forcing a relatively informal ideas-based approach that makes little explicit use of mathematical formulae. 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 as motivation examples of a type that commonly appear in practice.

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.

1.3 What is the purpose of these analyses?:

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?"

Commonly, it is important that results generalize beyond the circumstances that generated the particular data that are under study. In other words, the purpose is in some sense predictive. This module has a particular focus on implications that arise from the demand for predictive generalization.

The great majority of data mining analyses (all?) 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.

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.

Whatever the approach to inference, it is necessary to have a clear view of the purpose of the study. This may be a search for patterns (exploratory data analysis), or a formal data analysis that is intended to answer a specific research question. The following is a (perhaps incomplete) list of the purposes that a data analysis may be intended to serve:

1. Data collection and summarization is an end in itself.
2. Understanding – the elucidation of pattern.
3. Prediction; i.e., the aim is to make statements that generalize beyond the circumstances that generated the particular data that are under study.

Data collection and summarization are ends in themselves for much of the work of the Australian Bureau of Statistics. By explicit use of samples, or (occasionally) based on census data, statements will be made that apply to one or other Australian population – to humans, sheep, farms, or whatever. The data may be used directly to allocate resources, e.g., the distribution of GST revenue to states. It is 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.

Even for item 2, generalization is an issue. The intention is to make statements that will apply to other similar data sets.

Examples: 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.[1]

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.

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

2 Statistical Issues and Ideas

2.1 Population and sample

The ideas of population and sample are crucial. These are more subtle than is obvious on the surface.

2.1.1 Populations

In statistical theory a population is a set of values, with an associated probability measure. Here, it will be sufficient to consider two types of distribution — discrete (e.g., 0, 1, 2, ...) or continuous (e.g., any value on the real line). 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 $f()$,

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

In making inferences, it is necessary to distinguish two populations (they are too easily assumed identical):

The source population from which data have been drawn

The target population (if any) to which results will be applied.

In both cases we typically have, not the whole population, but a sample. Most classical statistical theory, and most use of models in data mining, assumes that samples, both from the source population and from the target population, are drawn according to simple random sampling. For observational data, this may be unrealistic. [There can also be complex sampling designs where randomization is used, but according to more complex schemes.]

2.1.2 Samples

In a simple random sample, often referred to as a random sample, each element of the population has the same probability of inclusion. Various modifications to simple random sampling schemes are available; these are important in sample surveys and in experimental design.

Given the population distribution, the statistical properties of random samples or repeated random samples can be derived theoretically.

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

2.1.3 Mean, variance and other sample statistics

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.

2.2 Models

Data analysis proceeds by a process of fitting models. The chosen model is designed to reflect, at least empirically, the processes that generated the data. Any use of analysis results, whether called data mining or statistical analysis, implicitly assumes a model that relates the data to the population to which the data is thought to have relevance. This may in principle be, and in practice often is, different from the model used for analysis. In order to discuss the practical implications of this point, use of some technical statistical apparatus is inescapable.

What then is a model? Why is the idea of a model important for data analysis? What is the role of models? The following are relevant:

> Maindonald & Braun (2003), Chapter 3.
>
> Breiman (2001), and its following discussion, is a good read, notwithstanding my judgment that Breiman's main thesis is nonsense!

Questions that will be addressed include:

1. What is a model? Can there be "algorithmic" models? (Models can be algorithmic only in the sense that they have an algorithmic motivation. All models, however motivated, can be use algorithmically. This may not not be a good idea! When does it make sense?)

2. Is model-free inference (whether formal or informal) possible? (No, but ...)

3. Can any search for patterns in data be model-free? (Not really, but simplistic assumptions can sometimes gain the investigator quite a bit of leverage. We'd like to distinguish problems where simplistic assumptions will yield useful results from problems where they will mostly turn up uninteresting and/or specious patterns.)

4. What is the point of the insistence, common among statisticians, that models have both "fixed" and "random" components? (It has mostly to do with the way that results from the model will be used. If the "random" component(s) are wrongly specified, false inferences may be drawn. Note that what is for one purpose "fixed" may for another purpose be "random")

5. How do we judge whether a model has worked or failed? (We challenge it, and observe how it responds. Which challenges will be effective, and which are so undemanding as to be useless?)

6. Which models are likely, in one or other context, to be effective? Which are likely to be ineffective? (A measure of effectiveness is needed. We'd like to know the effectiveness of the model, e.g., the accuracy of predictions, when the model is used in in practice. The best we can do may be to assess its effectiveness for the same data used to generate the model. This is NOT a simple matter testing the accuracy of the model with the same data that were used to fit the model; some greater subtlety is necessary.)

7. When is a data set "large" and/or "complex" for the purpose of the inferences, formal or informal, that are intended? (I will argue that the statistical modeling framework is necessary in order to give meaningful answers to this question. It is not at all the case that large size and high complexity from a database perspective is the same as large size and high complexity from a statistical modeling perspective.)

8. To what extent do "large" and/or "complex" data sets (now in a statistical sense) raise issues that are different from those that arise with small data sets?

2.2.1 The fixed and random components of models:

Statistical models have both "fixed effects" components, and random components. Models have both a fixed term and a random term. Most of the models considered in this course have the form:

$$y_i = f(x_i; \theta) + \epsilon_i, i = 1, 2, \ldots, n$$

where $f(x_i; \theta)$ is the fixed term, and ϵ_i is the random term. A further common (and commonly unjustified!) assumption is that the x_i are independently and identically distributed (iid), most commonly with a normal distribution. This will be referred to as the iid assumption.

It can be important to get both types of component right. Standard normal theory models that assume iid errors can be seriously limiting, and can lead to seriously misleading inferences. The later discussion will use data for which a multi-level model is appropriate to illustrate how accuracies may be different for different predictions, and how use of an iid errors model in such a case gives predictive accuracy estimates that are likely to be wrong, irrespective of the intended predictions.

Multi-level models offer a relatively simple form of escape from iid assumptions, and are a good starting point for demonstrating the implications of non-iid error structure. There are many practical contexts where they are effective – this relatively simple form of escape from iid assumptions is all that is needed. They have just enough complexity to provide a context in which to demonstrate the problems that may arise with an uncritical use of models that assume iid errors, or that proceed as though errors are iid. Other common types of non-iid model are:

1. Time series models, where it is common to model a sequential correlation structure;

2. Repeated measures models, which compare the times series (the "profiles") of different "individuals";

3. Models for spatial variation, which may model a spatial correlation structure.

There will not be time to do more than note these other types of model. Comments that I will make on contexts where a multi-level model is appropriate will however have broad relevance to these other non-iid contexts. We will illustrate consequences that may result from ignoring the correlation structure, which in practice usually means that we implicitly assume an iid structure. To what extent may it nevertheless be possible to make valid inferences or find patterns that are likely to provide a useful basis for further investigation?

2.3 Measurement error effects:

Measurement error in the explanatory variables of a model can have major implications for inference. An example will be given in Subsection 3.3.

2.4 Inference

Statistics works from samples, and aims to make inferences regarding the population from which the sample has been taken. There is no single universally agreed methodological principle — rather there are several different principles that are widely used. These all use the same probability theory, but use it differently. Differences between competent professionals typically relate more to the methodological principles to be followed than to the conclusions reached.

A simple intuitive approach assumes that the sample is the population in miniature. This view makes best sense for large populations. It must of course be possible to regard the sample as a random sample.

More formal methodologies are, broadly, of two types. One type of methodology is based around ideas of likelihood – given the model, what is the probability of obtaining a sample similar to that observed? Parameter values are chosen to maximize the likelihood. This methodology does not allow statements about the probability that a parameter has a particular value or range of values. (If such statements are made, there is an implicit use of a specific, but unstated, Bayesian prior, as used in Bayesian inference!)

Another type of methodology uses Bayes' theorem. The statistical model gives the probability of the data given the parameter values, i.e., the likelihood. Additionally, the prior distribution of the parameter values must be specified. If there is no hard information on this, then (this is where controversy arises), some plausible form of distribution must be assumed. Bayes' theorem is then used to derive the (posterior) probability distribution of the parameter values given the data. Statements about the probability that a parameter has a particular value or range of values are central to Bayesiam inference.

Theory that will be mentioned but not covered in any detail may include: minimum variance unbiased estimation in the context of linear least squares theory with a possibly arbitrary variance-covariance matrix, and the minimum squared error criterion.

(Important ideas that cannot be covered include: sufficiency, Fisher information, efficiency, maximum likelihood, asymptotic properties of the log-likelihood ratio, other asymptotic theory, and applications of Bayes theorem.)

2.4.1 Maximum Likelihood Estimators and Least Squares

Models have both a fixed term and a random term. Most of the models considered in this course have the form:

$$y_i = f(x_i; \theta) + \epsilon_i, i = 1, 2, \ldots, n$$

where $f(x_i; \theta)$ is the fixed term, and ϵ_i is the random term.

If the x_i are independently and identically distributed (i.i.d.) normal, then the least squares estimator of θ is the same as the maximum likelihood estimator. This is not true in general. The practical consequence is that the least squares estimator can be far from optimal.

2.5 Predictive Accuracy

Statistics has, as a major aim, the development of methods that distinguish what is "real" from effects or patterns that may well be due to chance. If this is the emphasis, then attention to predictive validation questions makes a lot of sense. When the results from the analysis are used to make predictions, do they check out?

2.5.1 Source and target populations

Here we discuss in more detail the questions:

- what is the population from which the data were derived, i.e., what is the *source* population. and how were the data sampled?
- what is the population to which results will be applied, i.e., what is the *target* population, and how will that population, in practice, be sampled?

It is commonly, implicitly if not explicitly, assumed that the population to which results will be applied or to which results are relevant is the same as the population used to derive the data, and that the sampling mechanism is the same in the two cases. Moreover it is assumed that both samples are obtained using an iid sampling mechanism, that they are in this sense *random* samples. Alternatively, predictions may be applied to the whole of the target population.

In a sampling context, the iid assumption is that sample items appear in the sample, independently between items, with a probability that is proportional to the frequency with which they appear in the population. Important questions are:

- Is the iid assumption, in the absence of an explicit randomization mechanism, justified?
- If wrong, is it wrong in ways that are likely to affect the validity of results? Is there some obviously preferable non-iid model (e.g., a time series or multi-level model) that can be used to give more valid results?

2.5.2 Measures of model performance

Measures of model performance are required in order to choose between models, and in order to give an assessment of how the model is likely to perform in practice.

Note first theoretical measures. R^2 is widely used, and may have some limited usefulness for comparing between models, but is almost useful as an absolute measure of performance. Other more satisfactory theoretical measures that may be used to compare models include AIC, BIC and the Schwartz criterion. Discussion of these criteria is beyond the scope of this document. Additionally, various goodness of fit and diagnostic criteria can be useful sources of insight that may help explain why the model performs as it does, suggest how the model might be improved, and draw attention to weaknesses in the model. Again, judgements are necessary on the relevance of one or other criterion, for the intended use of model results.

Often predictive accuracy, appropriately measured, is most pertinent. For normal theory linear models, and various other models, pertinent theoretical measures are available, though with the limitation that they relate to the population from which the data were derived. For classification models, however, classification accuracy is often the important criterion, perhaps with different costs assigned to different possibilities for mis-classification. The relevant *confusion* matrix, giving for each category the probability of assignent to each of the available categories, must then be estimated empirically. The following indicates the range of methodologies:

1. The confusion matrix may be estimated from the data used to derive the model. This *resubstitution* estimate can be hoplessly biased, and should not be used.
2. A *training* set may be used to derive the model, and a *test* set used to test the model. Ideally, the test data should reflect the conditions under which the model will be used in practice, i.e., they should be from the target population, allowing a genuine external assessment of accuracy.
3. More commonly, the data are randomly or arbitrarily split into a training and test set, with the training data used to derive the model, and the test data used to assess accuracy. As the training and test data are from the same orginal sample, the estimate is for the source population.
4. In item 3, the training (set I) and test (set II) data were derived from the same original sample. It therefore makes sense to swap the training and test sets (II/I in place of I/II), use II to fit the model, and I to test the model. Predictions are then available for all data, but always with the data used to test the model distinct from the data used to fit the model. An overall accuracy is then available that combines the I/II and the II/I results. This method is 2-fold cross-validation. The folds are I/II and II/I. The estimate is unbiased, but relates to the source population.
5. For use of cross-validation methodology, it is more usual to split the data into k parts, where a common choice is $k = 10$. Each part is then left out in turn; the model is derived using the remainder of the data, and predictions are made for the part that is left out. This is done for all k parts. Predictions are, finally, available for all the data, and the confusion matrix can be estimated. The estimate is unbiased, but relates to the source population.

Ideas of training and test set, and the use of cross-validation, can be important when theoretical assumptions fail, or when the relevant distributional theory is unavailable. As always, it may be necessary to adapt results to the specific inferences that are required. This can be particularly important when algorithms are treated as black boxes.

2.5.3 Source and target populations – some further comments

To what target population do these results apply?

In medical contexts, the question may be: "To what target population do these results apply?" It is assumed that results will be useful to someone somewhere, but who?

Thus, given results from a trial of a new drug, do they apply to the wider population of those who have the same disease symptoms as the patients in the trial? Patients will have been screened for suitability for the treatment. At best, they apply only to patients who pass the same screening criteria. If the trial was conducted in Australia, do they apply also in China, with patients whose lifestyle and eating habits are very different?

What are the appropriate test data?

(taken from my draft paper)

Here we assume that there is an identifiable target population, though perhaps lying somewhere in the future. The following four alternatives are an attempt at classifying the possible ways in which the target population may relate to the source population:

1. The data used to develop the model are, to a close approximation, a random sample from the population to which predictions will be applied. If this can be assumed, a simple use of a resampling method will give an estimate of the score function that is unbiased with respect to the population that is the target for predictions.

2. Test data are available that are from the target population, with a sampling mechanism that reflects the intended use of the model. The test data can then be used to derive a realistic estimate of predictive accuracy.

3. The sampling mechanism for the target data differs from the mechanism that yielded the data in 1, or yielded the test data in 2. However, there is a model that predicts how predictive accuracy will change with the change in sampling mechanism. Thus, in the "attitudes to science" data, the predictive accuracy for the mean of a new class depends on the number in the class.

4. The connection between the population from which the data have been sampled and the target population may be weak or tenuous. It may be so tenuous that a confident prediction of the score function for the target population is impossible. In other words, a realistic test set and associated sampling mechanism may not be available. An informed guess may be the best that is available.

These four possibilities are not completely distinct; they overlap at the boundaries.

Borderline Cases:

There may be a test set that is a plausible proxy for the target population. Examples are:

- The evaluation of algorithms for gene prediction.
  [Ideally, they should be effective in finding new genes!]

- Algorithms for finding protein homologies.

Alternatively, there may be very limited information on the source of variation that is most relevant to the predictions that are of interest. Maindonald, Waddell and Petry (2001) brought together data, on codling moth response to disinfestation with methyl bromide, for seven different varieties of cherry and for two seasons. The data were extensive, allowing a relatively accurate estimate of within season effects. Evidence on between season effects was of course very inaccurate, but did indicate that this source of variation was of a similar order of magnitude to the within season effects that the data were

designed to investigate. As the interest is in using experimental results to predict the consequences of a disinfestation protocol in a future season, this is an important limitation.

Informed guesses are more widely used than data analysts may admit. Commonly, predictions are required for a future time, but the only accuracy estimates are for prediction for another observation at the same time. Alternatively, as with the methyl bromide disinfestation data, information on temporal effects may be very limited. Questions that should be asked include:

- Is there any relevant past experience with comparable predictions?
- Is the accuracy for prediction in time likely to be similar to that for prediction in space, or smaller, or larger?

Of course, ballpark estimates can be badly astray.

Over-fitting

Models that are optimal for the source population are unlikely to be optimal for the target population. Indeed a simpler model that is sub-optimal for the source population will often do a better job. As a model is better tuned to the quirks of the source population, its tuning for the target population deteriorates. See Hand (2006) for further discussion of this issue, with examples.

2.6 Further reading

Chapter 4 of Senn (2003) is a relatively non-technical discussion of approaches to statistical inference. See Young and Smith (2005) for a more technical account.

3 Data Analysis and Interpretation Issues

3.1 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.

3.2 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.

3.2.1 Unequal subgroup weights – an example

Figure 3.2.1 relates to data collected in an experiment on the use of painkillers.[2]. Notice that the overall comparison (average for baclofen versus average for no baclofen) goes in a different direction 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.[3] 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.

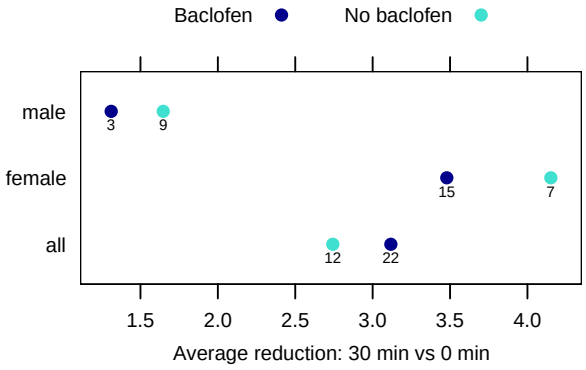

Figure 1: Does baclofen, following operation (additional to earlier painkiller), reduce pain? Subgroup numbers, shown below each point in the graph, weight the overall averages when sex is ignored.

The overall averages in Figure 3.2.1 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.

3.2.2 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

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

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

(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!

3.2.3 Simpson's paradox

In multi-way tables, such weighting effects 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.

3.3 Measurement error effects

Errors in explanatory variables, if they are sufficiently extreme, have two effects:

1. Estimates of effects will be reduced, relative to the true effects.
2. Effects, reduced or not, are hard to detect.

The attempt to use food frequency questionnnaires (FFQs) or food diaries, in studies that are designed to detect diet-disease associations, provides a telling and interesting case study. A recent major study with biomarkers has demonstrated large person-specific biases in standard dietary intake measurement "instruments" (diaries or questionnaires). These biases severely complicate the finding of a relationship between such measures and health outcomes. Not only is there an error that varies from recording time to recording time, for an individual. There is also a person-specific bias, that can be substantially larger than the random occasion to occasion error. See Schatzkin et al (2003) and the power point presentation Carroll (2004).

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+
(The AARP results will be available within the next few months)	

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.

3.4 Further examples and discussion

3.4.1 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.

3.4.2 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 1: 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[4] 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:

	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 2: 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?

3.4.3 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 3: 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 4: 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.

3.4.4 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 as the outcome. Risk is expressed as a probability density. Thus these analyses have probability density as the outcome variable.

[4] 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

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 5 shows the estimated ratio of risk relative to the baseline line, i.e., to the risk from all other factors.

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 5: 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.

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).

3.4.5 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 **?**.

3.4.6 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 6 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 6: 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?

3.4.7 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. Overall, the conclusion now 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.

Lawlor et al (2004) discuss why there is agreement for most outcomes, but not for CHD. Other studies have shown that for CHD, childhood socio-economic indicators are important as predictors of CHD, independently of adult socio-economic status (SES), behavioural and physiological risk factors. That is not true for the other outcomes considered. Additionally, the use of HRT is "strongly socially patterned". These are just the circumstances that can be expected to lead to confounding (Simpson's paradox type effects), as already discussed.

(The argument has the following character. There will be a group of individuals who had low childhood SES, but high adult SES. Their low childhood SES may both lead to low use of HRT and consequent lowered risk of CHD. In the analysis, the only adjustment is for their high adult SES. This leads to over-correction for SES. The benefit that arises from non-use of HRT is wrongly ascribed, in the analysis and its associated interpretation, to their high adult SES.)

If Lawlor et al (2004)'s account is correct, these investigations highlight the importance of accounting properly for socio-economic effects. When studying an outcome of interest from an observational study, it is important to ask whether the simpler type of model that can account for breast cancer risk is adequate, or whether the situation that pertains to CHD risk is more likely.

3.4.8 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.

3.5 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.

3.6 Variable selection and other multiplicity effects

Model coefficients and estimates can be susceptible to huge biases when there is substantial variable selection that is designed to improve discrimination between subgroups of the data. This is an especial issue for the analysis of microarray and other genomic data. 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 (unpub. manuscript) 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.

4 Data Analysis System Use and Development Strategies

In our course material, the R system is a particular focus of attention, for several reasons. Major aspects of its development have had the character of a co-operative interaction research project. The associated statistical computing issues have spawned an extensive statistical literature. It is now the dominant system used for analyses that are presented in statistical conference papers. It has become

a preferred environment for academic statistical software development, with inroads also into the data mining and machine learning communities. It has is now a de facto standard, in quality of code, range of abilities, and integration into a common language framework.

Papers and books that may warrant attention include Chapter 2 of Maindonald & Braun (2003), Chambers (2000), Maindonald (2004b) and Maindonald (2004a). (All these have a strong focus towards S-PLUS or R, and to systems that interface to R or take R as a point of departure for further development.)

Part II
Populations, Distributions and Samples

5 Populations

5.1 Probability distributions

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.

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

$$\Pr[\mathrm{x}_1 < \mathrm{X} \leq \mathrm{x}_2].$$

5.2 Density Curves and Cumulative Distribution Functions

Here, attention will be restricted to continuous distributions. 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)
```

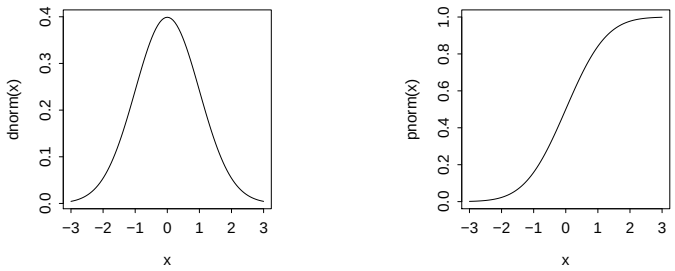

Figure 2: Normal density curve, with cumulative distribution function alongside.

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

$$\Pr[\mathrm{x}_1 < \mathrm{X} \leq \mathrm{x}_2 = \Pr[\mathrm{X} \leq \mathrm{x}_2 - \Pr[\mathrm{X} \leq \mathrm{x}_1.$$

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
```

5.3 The mean and variance of a population

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

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

while the variance is

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

6 Samples

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

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

```
> rnorm(n = 10)
> rnorm(n = 10, mean = 11, sd = 2)
> runif(n = 10)
> 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)
```

6.1 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. For displaying the sample distribution of a set of values, histograms have traditionally been the first recourse. A better plot, often, may be a density plot. This is, essentially, a smoothed version of a histogram.

Below, we plot the distribution of heights of 118 female students attending a first year statistics class at the University of Adelaide. In Figure 3 we plot a histogram and overlay 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.)

The function `na.omit()` omits missing values.

```
> library(MASS)
> y <- na.omit(survey[survey$Sex == "Female", "Height"])
> hist(y, prob = TRUE, xlab = "Heights of female students", main = "")
> lines(density(y))
```

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

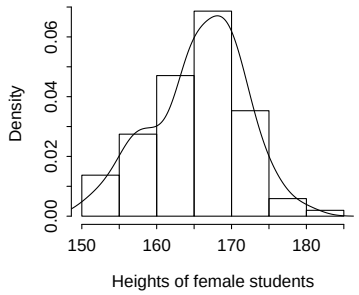

Figure 3: Histogram, with overlaid density plot, showing the distributions of heights of 118 female students in a first year statistics class at the University of Adelaide.

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.

```
> av <- mean(y)
> sdev <- sd(y)
> plot(density(y, kernel = "gaussian", width = 10), xlab = "Heights of female students",
+     main = "")
> rug(y)
> chw <- par()$cxy[1]
> chh <- par()$cxy[2]
> abline(v = av, col = "gray")
> ytop <- par()$usr[4] - 0.15 * par()$cxy[2]
> text(av, ytop, "mean", col = "gray45", xpd = TRUE)
> abline(v = av - 0.65 * sdev, col = "gray", lty = 2)
> text(av - sdev - chw, ytop - 0.85 * chh, "mean\n-SD", col = "gray40",
+     xpd = TRUE, cex = 0.8)
> abline(v = av + sdev, col = "gray", lty = 2)
> text(av + sdev + 0.65 * chw, ytop - 0.85 * chh, "mean\n+SD",
+     col = "gray40", xpd = TRUE, cex = 0.8)
```

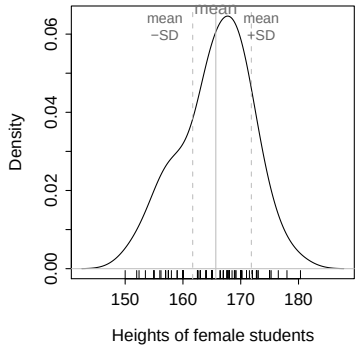

Figure 4: Density plot, now with a larger smoothing window (`width`) and with a gaussian (normal) kernel, showing the distribution of heights of 118 female students in a first year statistics class at the University of Adelaide. Marks on the horizontal axis show the actual heights. Also marked off, in 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.

6.2 The smoothness of the density plot

We can also 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 5 shows, for the Adelaide female student data, the effect of varying the bandwidth.

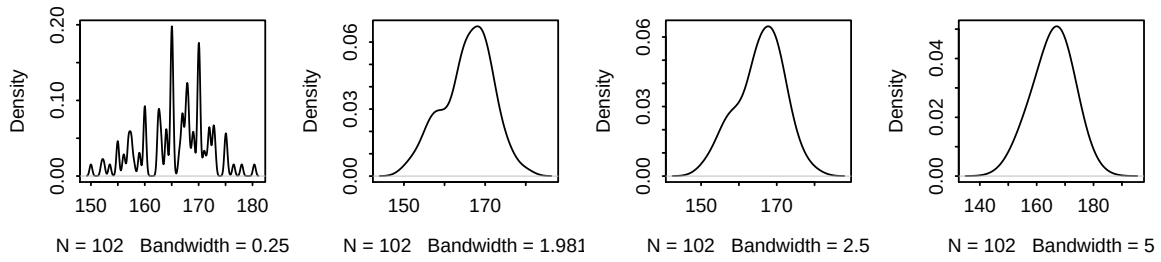

Figure 5: Density curves for Adelaide female student heights. Curves are shown for three different choices of bandwidth: 0.25, 1.98 (the default for these data), 2.5 and 5.0. The normal kernel (the default) is used in each case, so that increasing the bandwidth forces the curve closer to normal.

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

6.3 Normal and other probability plots

Although preferable to histograms, density plots not 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 6 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)
```

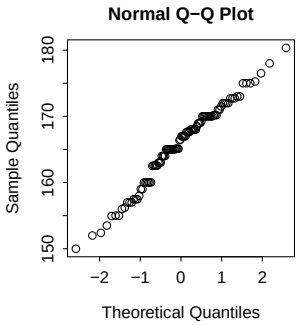

Figure 6: Normal probability plot for the distribution of heights of 118 female students in a first year statistics class at the University of Adelaide.

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)
```

6.4 *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 7 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)
```

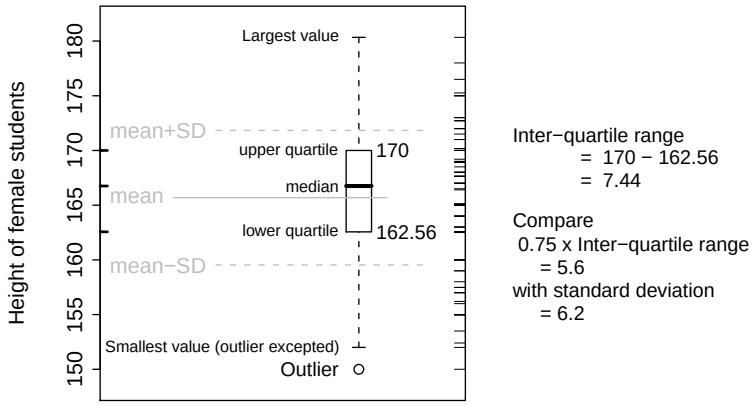

Figure 7: Boxplot, with annotation that explains boxplot features. Lines in gray show mean-SD, mean, and mean+SD. Data are heights of 118 female students in a first year statistics class at the University of Adelaide.

7 Sample Statistics – Variance and Standard Deviation:

See Subsection 5.3 for a definition of population variance (= σ^2, where σ is the standard deviation). See Subsection 2.1.3 for a definition of sample mean and sample variance. The standard deviation (SD) is the square root of the variance.

7.1 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.

7.2 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 = "")
```

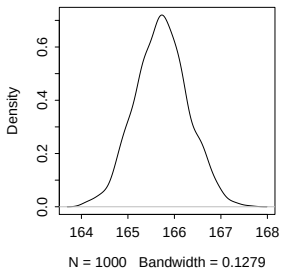

Figure 8: Simulated distribution of the mean, for samples of size 118 from a normal distribution with mean=165.7 and SD=6.15, as for the sample of UAdelaide students.

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")
```

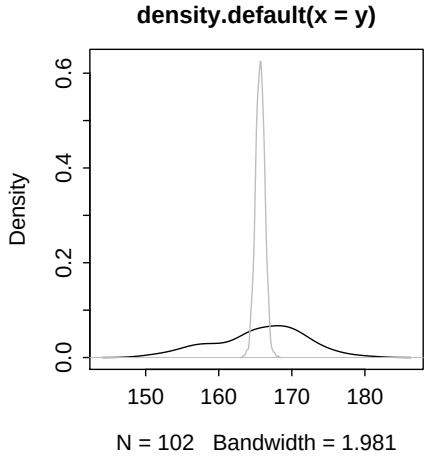

Figure 9: Simulated distribution of the mean, for repeated samples of size 118, with replacement, from the sample of UAdelaide students.

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(-0.5, 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.

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. Indeed, 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.

8 Straight Line Models in R

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.

8.1 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.

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

The following plots the data in the data frame `roller` (shown in Table 7) 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`.

8.2 Straight Line Regression – More Details

The straight line regression model has the form

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

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 10.

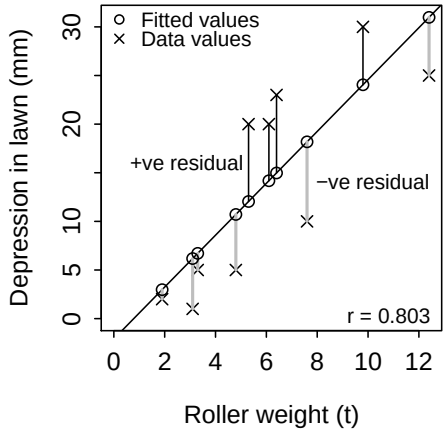

Figure 10: Lawn depression for various weights of roller, with fitted line. The fitted line is designed to minimize the sum of squares of residuals, i.e., the sum of squared lengths of the vertical lines, joining x's to o's, that are shown on the graph.

8.3 The Model Matrix

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 $\mathbf{X}$ 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".

8.4 Recap, and Next Steps in Linear Modeling

This section has discussed one of the simplest possible type of linear model. It has shown how to construct the model matrix with which R works 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 now be demonstrated.

9 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, ..., x_k, y) of real numbers, where the x_is are explanatory variables and y is the dependent variable.

Given x_1, x_2, ..., 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 $\phi_1(x_1, x_2, \ldots x_p) = x_1$, $\phi_2(x_1, x_2, \ldots x_p) = x_2$, ..., $\phi_p(x_1, x_2, \ldots x_p) = x_p$.

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.

Among the R packages that have abilities that have functions that allow this flexibility, note the *nlme* package, and especially the `lme()` function in that package. The final section of these notes will demonstrate the use of `lme()`, albeit in a very simple example of such modeling. Comparison with the more simplistic results that are obtained from use of `lm()` can be insightful.

9.1 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.)

10 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 data that will be used are a subset of the `races2000` data set that is in the `DAAG` package. To make the data available, do the following:

```
> library(DAAG)
> names(races2000)
```

```
 [1] "h"     "m"     "s"     "h0"    "m0"    "s0"    "dist"  "climb" "time"
[10] "timef" "type"
```

```
> hill2k <- races2000[races2000$type == "hill", 7:10]
```

The row names store the names of the hillraces. I have recently discovered that 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. For later reference, note the row number:

```
> match("Caerketton", rownames(hill2k))
```

```
[1] 42
```

```
> hill2k[42, "dist"]
```

```
[1] 2.5
```

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:

```
> if (dev.cur() == 2) invisible(dev.set(3))
> loghill2k <- log(hill2k[-42, ])
> names(loghill2k) <- c("ldist", "lclimb", "ltime", "ltimef")
> loghill2k.lm <- lm(ltime ~ ldist + lclimb, data = loghill2k)
> par(mfrow = c(2, 2))
> plot(loghill2k.lm)
> 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:

```
> if (dev.cur() == 3) invisible(dev.set(2))
> par(mfrow = c(1, 2))
> termplot(loghill2k.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 spline curves. This discussion will be continued below.

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 8: 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`)

11 Modeling Qualitative Effects

11.1 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 8, 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 8, `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:

```
> 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)   83.233      4.473  18.609 7.17e-08
trtA         -21.400      6.325  -3.383 0.009597
trtB         -15.733      6.325  -2.487 0.037680
trtC         -34.333      6.325  -5.428 0.000625

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.)

11.2 Two factors – two bowlers and two innings

The following table compares bowler B with bowler A, over the two innings of a cricket match:

	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

Runs per wicket calculations that are based on the run and wicket totals over both innings will be differently biased for the two bowlers. Bowler 1's average is biased towards his results for the second innings, when runs were more plentiful and wickets were harder to get. Bowler 2's average is biased towards his results for the first innings, when fewer runs were on offer and wickets were easier to get.

In place of such a calculation, we will fit an additive model that explains runs/wicket as an effect of bowler and of innings, to the 2×2 table. In the statistical terminology that is used by R, we have `bowler` and `innings`. These are categorical variables, each categorizing the data, though in different ways. It is not, for these data, meaningful to calculate an error term. The interest here is, rather, in finding a fair way to represent results from this partticular match.

The runs per wicket information is:

	one	two
A	10.00	40.00
B	14.00	50.00

According to the model, the values in the table are:

	one	two
A	μ	$\mu + \tau_2$
B	$\mu + \beta_2$	$\mu + \beta_2 + \tau_2$

Table 9: Runs per wicket for each bowler in the two innings.

We require values of μ, β_2 and τ_2 that minimize

$$(10 - \mu)^2 + (40 - \mu - \tau_2)^2 + (14 - \mu - \beta_2)^2 + (50 - \mu - \tau_2 - \beta_2)^2$$

We can put this into the same form as for the lawn roller data by setting:

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} 10 \\ 40 \\ 14 \\ 50 \end{pmatrix}$$

The sum of squares to be minimized is $(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$ where

$$\beta = \begin{pmatrix} \mu \\ \alpha_2 \\ \beta_2 \end{pmatrix}$$

Option 1: Explicitly solve the matrix equations.

Option 2: Use R's `lm()` function, first creating columns of dummy values (0s and 1s) that account for the two factors in the model.

Define `X` thus:

```
> y <- c(10, 14, 40, 50)
> X <- data.frame(bowler = c(0, 1, 0, 1), innings = c(0, 0, 1,
+     1))
> lm(y ~ bowler + innings, data = X)

Call:
lm(formula = y ~ bowler + innings, data = X)

Coefficients:
(Intercept)       bowler      innings
        8.5          7.0         33.0
```

given explicitly.

Option 3 (in practical analysis; much the preferred option): As described below, set up factors `bowler` and `innings`. The code that is called by the `lm()` function will recognize that `bowler` and `innings` are factors, and set up the appropriate basis vectors. In this instance one basis vector in the default parameterization consisting of 0s and 1s, is enough for each factor.

First create factors `bowler` and `innings`, thus:

```
> y <- c(10, 14, 40, 50)
> bowler <- factor(c("A", "B", "A", "B"))
> innings <- factor(c("one", "one", "two", "two"))
> xtabs(y ~ bowler + innings)

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

The two factors give two different classifications of the data. The factor `bowler`, indicates that the first two observations were for bowler `A`, while the second two observations were for bowler `B`. The factor `innings` indicates that the first and third observations are for innings `one`, while the second second and fourth observations are for innings `two`.

Now comes the magic of a model formula. The function `lm()` is called with the model formula `y ~ bowler + innings` as its formula argument. The function `lm()` immediately calls `model.matrix()`. This takes the terms in the formula, and from them calculates the columns of 0s and 1s that are used by the default parameterisation, thus:

```
> lm(y ~ bowler + innings)

Call:
lm(formula = y ~ bowler + innings)

Coefficients:
(Intercept)      bowlerB   inningstwo
        8.5          7.0         33.0
```

Fitted Values: Plugging the estimates of μ, β_2 and τ_2 back into the table, we have Table 10:

	one	two
A	$\hat{\mu} = 8.5$	$\hat{\mu} + \hat{\tau_2} = 8.5 + 33 = 41.5$
	(10)	(40)
B	$\hat{\mu} + \hat{\beta_2} = 8.5 + 7 = 12.5$	$\hat{\mu} + \hat{\beta_2} + \hat{\tau_2} = 8.5 + 7 + 33 = 48.5$
	(14)	(50)

Table 10: Estimates of runs per wicket for each bowler in the two innings, calculated from the model. The actual runs per wicket are given in parentheses.

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

```
> levels(bowler)

[1] "A" "B"

> levels(innings)

[1] "one" "two"
```

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
```

The model formula can now extract bowler and innings from the columns of `Z`.

```
> lm(y ~ bowler + innings, data = Z)

Call:
lm(formula = y ~ bowler + innings, data = Z)

Coefficients:
(Intercept)      bowlerB   inningstwo
        8.5          7.0         33.0
```

A further refinement would be to include `y` as a column of `Z`).

Finally, here is the matrix that R uses for its least squares calculation

```
> model.matrix(~bowler + innings, data = Z)

  (Intercept) bowlerB inningstwo
1           1       0          0
2           1       1          0
3           1       0          1
4           1       1          1
attr(,"assign")
[1] 0 1 2
```

```
attr(,"contrasts")
attr(,"contrasts")$bowler
[1] "contr.treatment"

attr(,"contrasts")$innings
[1] "contr.treatment"
```

11.3 Extensions:

1. The above is the "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(~bowler + innings, data = Z)

  (Intercept) bowler1 innings1
1           1       1        1
2           1      -1        1
3           1       1       -1
4           1      -1       -1
attr(,"assign")
[1] 0 1 2
attr(,"contrasts")
attr(,"contrasts")$bowler
[1] "contr.sum"

attr(,"contrasts")$innings
[1] "contr.sum"

> lm(y ~ bowler + innings, data = Z)

Call:
lm(formula = y ~ bowler + innings, data = Z)

Coefficients:
(Intercept)      bowler1     innings1
       28.5         -3.5        -16.5
```

   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. As just hinted, there can be interactions between factors, or between factors and variables. The following includes the term `bowler:innings`, which allows the difference between the two bowlers to be different between the two innings:

```
> options(contrasts = c("contr.treatment", "contr.poly"))
> lm(y ~ bowler + innings + bowler:innings, data = Z)

Call:
lm(formula = y ~ bowler + innings + bowler:innings, data = Z)

Coefficients:
       (Intercept)             bowlerB          inningstwo  bowlerB:inningstwo
                10                   4                  30                   6
```

3. It is easy to put in additional bowlers.

```
> y3 <- c(10, 14, 24, 40, 50, 150)
> Z <- expand.grid(bowler = c("A", "B", "C"), innings = c("one",
+     "two"))
> lm(y3 ~ bowler + innings, data = Z)

Call:
lm(formula = y3 ~ bowler + innings, data = Z)

Coefficients:
(Intercept)      bowlerB      bowlerC   inningstwo
         -7            7           62           64
```

4. Further extensions will be noted in the next subsection.

11.4 The grouping of model terms

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:

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

In the above, the basis functions formed just one group. 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.

12 *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.

12.1 *Fitting Spline Terms to the Hill Race Data

We return again to the hill race data. We had

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

A spline of degree 3 (by default a cubic polynomial) seemed adequate for capturing the curvature in the partial residuals for `ldist`, while a spline of degree 4 seemed adequate for capturing the slightly more complicated pattern of curvature in the partial residuals for `lclimb`:

```
> library(splines)
> loghill2ks.lm <- lm(ltime ~ ns(ldist, 3) + ns(lclimb, 4), data = loghill2k)
```

Notice that the first plot brings together the information associated with the basis functions that are generated by `bs(ldist,3)`, while the second plot brings together the information associated with the basis functions that are generated by `bs(lclimb,4)`

Diagnostic plots: The following is a series of diagnostic plots, designed to highlight issues that it may be important to consider:

```
> if (dev.cur() == 2) invisible(dev.set(3))
> par(mfrow = c(2, 2))
> plot(loghill2ks.lm)
> par(mfrow = c(1, 1))
```

The diagnostic plots cannot possibly identify all possible problems with the fit of the models to the data. It is possible to have models where the diagnostic plots look fine, but the model is lousy. They can however be very useful in picking up some issues that commonly merit attention – outliers, non-normality in the residuals, heterogeneity of variance, and points that individually have a large effect on the fitted model.

Notice that, in the diagnostic plot, one point (row 19: 12 Trig Trog) has a huge Cook's distance. With a time of 8.3h, it is the longest of any of the races.

The following plots the contributions of the individual spline curves ("the effects"), shows the partial residuals, and passes a smooth curve (red dashes) through the partial residuals:

```
> if (dev.cur() == 3) invisible(dev.set(2))
> par(mfrow = c(1, 2))
> termplot(loghill2ks.lm, col.term = "gray", partial = TRUE, col.res = "black",
+     smooth = panel.smooth)
> par(mfrow = c(1, 1))
```

Also the fitted curve for `lclimb` is not monotonic for small values of `lclimb`. It would be desirable to constrain it to be monotonic.

*The basis functions

Use the following to inspect and plot the basis functions:

```
> bases <- model.matrix(loghill2ks.lm)
> colnames(bases)
> options(digits = 3)
> bases[1:5, ]
> par(mfrow = c(2, 2))
> for (i in 0:3) plot(loghill2k$lclimb, bases[, 5 + i])
> par(mfrow = c(1, 1))
```

The contribution of `lclimb` to the fitted values is determined as a linear combination of these four curves.

Part IV
Generalizations of Linear Models

13 Generalized Linear Models & Survival 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 3.4.4, 3.4.5, 3.4.6 and 3.4.7. 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.

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.

13.1 Generalized Linear Models

These are an extension of linear models. 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.

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.

13.1.1 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 $\mathrm{E}[y]$. The result specifies a linear relation with x. In other words,

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

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 $\mathrm{E}[y]$ 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.

13.1.2 Noise terms need not be normal

We may write

$$y = \mathrm{E}[y] + \varepsilon.$$

Here the elements of y may have a distribution different from the normal distribution. Common distributions are the binomial where y is the number responding out of a given total n, and the Poisson where y is a count.

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.

13.2 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.

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

Part V
Multi-level Modeling

14 Multi-level Models – General Comments

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.

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.

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

14.1 The Antiguan Corn Yield Data – An Example

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 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.

The data that will be analyzed are in the second column of Table 11, 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 11 indicate how to calculate the

Site	Site means	Site effect		Residuals from site mean
DBAN	5.16, 4.8, 5.07, 4.51	(4.29)	+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		+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
		square, add, multiply by 4, divide by d.f.=7, to give ms		square, add, divide by d.f.=24, to give ms

Table 11: 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.)

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?

14.2 The model

The model that is used is:

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

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 $\text{var}[\alpha_\text{i}] = \sigma^2_\text{L}$, $\text{var}[\beta_\text{ij}] = \sigma^2_\text{B}$.

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.

15 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 $\text{var}(\alpha_\text{i})$ 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

$$2(\sigma_L^2 + \sigma_B^2)$$

- For s single package, the variance is $\sigma_L^2 + \sigma_B^2$. The variance of the estimate of the site mean is a mean over the four packages at the one site, and is

$$\sigma_B^2 + \frac{\sigma_L^2}{4}$$

[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.]

Part VI
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 13 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

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

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

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

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, $\mathrm{E[c_1X + c_2X]} = \mathrm{c_1E[X] + c_2E[Y]}$. 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

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

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.

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>`
Analyses of these data, using tree-based regression, will be discussed in some detail.]

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.

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.

Lawlor, D. A., Davey Smith, D. F. and Ebrahim, S. 2004. Commentary: The hormone replacement – coronary heart disease conundrum: is this the death of observational epidemiology? International Journal of Epidemiology **33**:464–467.

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

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

Maindonald, J.H. 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. 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., *unpublished manuscript.* Predictive Validation, Algorithmic Models and Data Mining.
[This is a response both to the Breiman (2001) and to simplistic approaches to predictive validation.]

MAINDONALD, J.H. AND BRAUN, W.J. 2003. *Data Analysis and Graphics Using R – An Example-Based Approach.* Cambridge University Press.
`<URL:http://wwwmaths.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.

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).

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.]

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.]

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.