Modern Regression – the Wider Statistical Context

John Maindonald

December 30, 2011

Contents

A Warning Note:
A big computer, a complex algorithm and a long time does not equal science.
[Robert Gentleman, SSC 2003, Halifax, June 2003]

The R system will be used, extensively, for the computations of the later chapters. Readers who wish to follow the calculations as well as the accompanying discussion should therefore start by skimming through Appendices A, B and C. Those who do not already have some reasonable level of familiarity with this material will be advised to work carefully through it before proceeding too far into the text.

1 Code for the figures

Code for each of the figures in this chapter is provided as a function, included in the R package *modregR*. Additionally, it has a number of datasets that will be used in the course.

Assuming that the package *modregR* has been installed, use the command `library()` to attach it, thus:

```r
library(modregR)
```

Figure 1.1 plots `depression` in a lawn against roller `weight`, for lawn rollers of different weights. A regression line has been added.

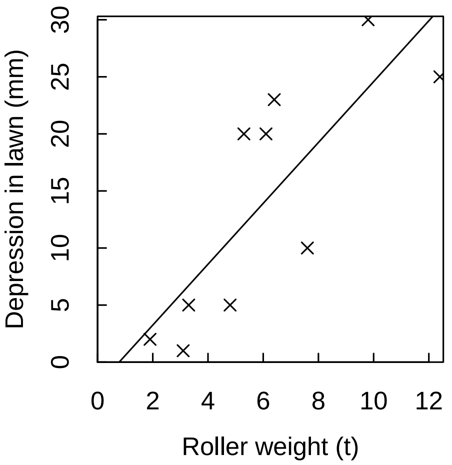

Figure 1:
Depression in lawn (mm) versus roller weight (t).

```r
## Simplified version of graph
plot(depression ~ weight,
     data=roller)
abline(lm(depression ~ weight,
          data=roller))
abline(roller.obj)
```

The discussion that follows will assume that that *DAAG* package is attached, i.e., that we have entered:

```r
library(DAAG)
```

A good way to proceed, for making sense of the R code, is to start by running the function. Then look inside the function to see what code is used. Readers who are not totally at home with R may be wise to take as given figures where the code has some modest degree of complication.

In a few cases, including `g1_1()`, the function is designed so that it can be used with any comparable dataset. Those functions that are specific to particular datasets allow, in most instances, parameter settings that change the appearance of the graph.

The function `g1_1()` has code for Figure 1.1. The function `f1_1()`) has simpler and less general code for the same purpose. The function `g1_2()` has code for Figure 1.2, and so on. For Figure 1.1, type `g1_1()`. For Figure 1.2, type `g1_2()`, and so on.

The following is the code for `f1_1`:

```
f1.1 <- function(){
## Simplified version of graph
plot(depression ~ weight,
     data=roller)
abline(lm(depression ~ weight,
          data=roller))
abline(roller.obj)
}
```

Here, the function serves only as a wrapper for the code. To run this code, enter following the command prompt:
`f1_1()`

To run the more detailed code that generated Figure 1, type

```
xlim <- c(0, max(roller$weight)*1.01)
ylim <- c(0, max(roller$depression)*1.01)
g1_1(xlab="Roller weight (t)",
     ylab = "Depression in lawn (mm)",
     xlim=xlim, ylim=ylim,
     xaxs="i", yaxs="i", pch=4)
```

Here are details (type `print(g1_1)` or `g1_1` to display this code):

```
g1_1 <- function(form=depression~weight,
                 data=roller, ...){
  roller.obj <- lm(form, data=data)
  plot(form, data=data, ...)
  abline(roller.obj)
}
library(DAAG)
xlim <- c(0, max(roller$weight)*1.01)
ylim <- c(0, max(roller$depression)*1.01)
g1_1(xlab="Roller weight (t)",
     ylab = "Depression in lawn (mm)",
     xlim=xlim, ylim=ylim,
     xaxs="i", yaxs="i", pch=4)
```

Named arguments `form` (a formula object) and `data` (a data frame) are used in calls both to `lm()` and to `plot()`. The `...` mechanism allows further arguments that, supplied in a call to `g1_1()`, pass through to the `...` argument in each of the two calls to `plot()`.

Note that the code has been written so that, for use with different data, it is only necessary to change the function arguments. Most figures that appear later in this chapter are so specific to the particular data that it has not seemed useful to make them similarly general.

1.1 A lattice plot

Note, finally, the possibility of using the *lattice* function `xyplot()`:

```
xlim <- with(roller, c(0, max(weight)*1.01))
ylim <- with(roller, c(0, max(depression)*1.01))
gph <- xyplot(depression ~ weight, data=roller,
              xlim=xlim, ylim=ylim, axs="i",
              type=c("p","r"))
print(gph)
```

In `type=c("p","r")`, `"p"` gives points, while `"r"` adds a regression line. Use `"l"` (lines) to join up points.

Lattice functions create graphics objects that can then be stored or updated (use `update()`) or printed, as required.

2 A Note on Least Squares

Several of the examples in this chapter will use the least squares methodology. A description of the least squares criterion may then be appropriate. Its use requires very specific assumptions

– assumptions that are too readily forgotten ot ignored. In some of the examples, and in some of the later chapters of this text, it will be necessary to move beyond least squares to the use of other model fitting criteria.

Figure 2 has annotation that shows the details of the vleast squares fitting criterion. It shows, as vertical lines, the residuals that are squared and summed to give the residual sum of squares when fitting a regression line. At the same time, it demonstrates the creation of a graph where the code is much more complicated and intricate than Figure 1.

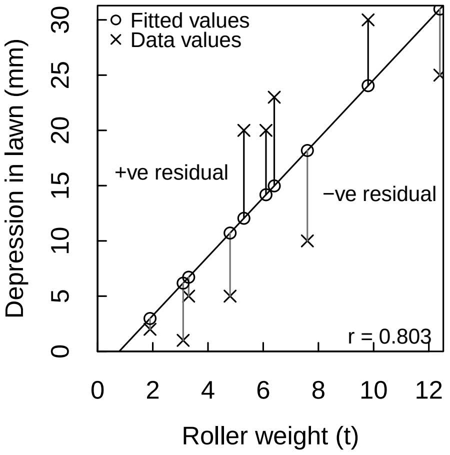

Figure 2: The fitted line is designed to minimize the sum of squares of residuals, i.e., the sum of squared lengths of the vertical lines, joining x's to o's, that are shown on the graph. For the code, examine the contents of the function `g1_2()`.

```
## To see the (not very elegant)
## code, type:
print(g1_2)
```

Use of least squares is justified (when it can be justified) by the appeal to a more fundamental criteria. In particular, if certain assumptions apply, it can be justified by an appeal to the maximum likelihood principle. Least squares may be, and commonly is, inappropriate:

- for time series data;
- for data where there are likely to be multiple levels of variation – e.g., between lawns, and between different testing positions within a lawn;
- for 0/1 data;
- for count data.

Unmodified least squares (ordinary least squares, or OLS) is inappropriate for data where different obseervations have very different accuracies.

3 The Uses of Data Analyses

The use that is in mind will commonly have large implicatiosn for the analysis. More specifically, key issues for any study are:

1. What is the aim – why undertake this investigation?
2. Is the aim realistic?
3. What is the intended use of results?

4. How widely is it hoped that results will apply?
5. What limitations, from the manner of collection or from the incompleteness of the information, may constrain the intended use?

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

3.1 Possible purposes

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

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

It is then important to ask which of these apply.

The purpose commonly has strong implications for the decision on how to handle the data. Data mining projects often focus, initially, on predictive accuracy. Questions of what interpretation can be placed on model parameters do however often arise, often as an afterthought to the main analysis. It is therefore important to be aware of the potential for misinterpretation.

Is an hypothesis essential?

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

3.2 Data Issues

- Source/target issues – are source/target differences large enough to matter?
  As examples, consider
  - A business has data on Christmas 2010 sales patterns. Will Christmas 2011 sales follow the same patterns?
  - Data in the forensic glass data set `fgl` in the *MASS* package was collected prior to 1987. Its relevance to a 2011 crime scene is, at the very least, doubtful.

  Source/target issues are inevitably tied up with wider issues of validity. Validity is, necessarily, validity for an intended purpose.
- Validity
  Many types of superannuation payments do not, for Australian Tax Department purposes, count as taxable "income". Various other payments are excluded. In a study of social and health correlates of income, ATO taxable income would not be a valid measure.

- Accuracy – are the measurements accurate?
  Any measure of income is likely to be compromised, to an extent, because some individuals will hide some sources of income. Here, it is accuracy for some intended purpose that is compromised.

- Relevance – are the data relevant to the questions of interest?
  It is pretty obviously not useful to collect information on numbers of storks, for purposes of predicting numbers of babies. What information is it then useful to collect? Examples of what might be useful include: marriage rates, cohabitation rates, age at first marriage, household income, women's education, use of contraceptives, family planning intentions?

- How many variables are there? How many observations? Are the observations structured in any way, e.g., multiple observations per individual, for a number of individuals?

Example – the different uses of Australian Bureau of Statistics data

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

Exercises:

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

> The forest cover type data set, available from the web site noted in connection with Blackard (1998).[1]
>
> The data set `ant111b`, discussed in Subsection 5.3, 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.
>
> The data set on tinting of car windows (`tinting`, *DAAG* package) .
>
> 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.

[1] See `http://kdd.ics.uci.edu/databases/covertype/covertype.data.html`.

3.3 Data-based Generalization

Most (all?) analyses involve an element of generalization. The aim is to make statements that have relevance beyond the specific data that are analysed. In predictive modeling, generalization is an explicit concern. The nature of the generalization will commonly have major implications for the investigations that are to be undertaken.

The term "generalization" is used here, where theoretical statisticians commonly use the word "inference". A first requirement is that relevant data should be available – relevant in the sense that they can in principle support the intended generalization. Depending on the nature of the investigation and on the resources put into data collection, available data may or may not ensure such relevance. Where an experimental or survey or other data collection protocol can achieve what is required, this is clearly a preferred approach.

Commonly data are used that call for a large element of judgement about relevance to the intended use. Data may for example have been collected a decade or more ago, where the interest is in what is happening now in 2011? What limitations does this place on the use of the data? Is is reasonable to regard the data that we have, the "source" data, as a sample from the "target" population that is of interest? This will depend on the what is under study. Data from a 20-year old study on how the freezing point of water changes with the amount of dissolved sodium chloride and with vapour pressure should be just as relevant now as when the study was conducted. Data that are used to assess credit-worthiness of loan applicants may lose its relevance after a lapse of a year or two, or sooner in a time of financial crisis.

For what purposes might the data that are displayed in Figure 1 be useful? They do have relevance for the particular lawn where the data were collected, at the time of collection. They have very limited relevance to other lawns. They do indicate, what might anyway be expected, that heavier rollers give larger depressions. The data are consistent with the assumption of a linear relationship.

4 Models

All analyses involve assumptions. Models are designed to make assumptions explicit.

4.1 Independence and other assumptions

The main part of this text will focus on methods that make quite strong assumptions. Notably, it will mostly be assumed that observations are independent. The independence assumption is an assumption about the process that generated the data, about the way that it should be modeled.

Consider aircraft maintenance. If the same mechanic services two aircraft engines at the same time using replacement parts from the same batch, this greatly increases the chances that the same mistake will be made on the two engines, or the same faulty part used. Maintenance faults are then not independent. Independence is not the harmless assumption that it is often made out to be!

A model has both fixed and random parts. Both need to be explicitly specified. Under the independence assumption, the random part consists of independent perturbations about the fixed part. It may be assumed, additionally, that the perturbations are generated by a normal distribution, usually with the same variance for all observations.

Least squares, as implemented by R's `lm()` function, is easily the most widely used methodology for regression calculations. It requires (unless weights can be specified), for meaningful results, the further assumption that perturbations about the fixed part should be unbiased and have the same variance. A further assunption is that deviations are normal, with mean zero.

Later chapters will move beyond least squares to contexts where an alternative methodology is required. There will be occasional limited forays to examine the consequences of lack of independence. For the most part, however, independence will be assumed.

Examples will help make clear the importance of some of the key assumptions. They are intended to help make clear the context for the remaining chapters. It will become apparent that assumptions that are reasonable for one use of results may for some different use of results be misleading.

4.1.1 Is the iid assumption reasonable for the lawn roller data?

The least squares analysis assumes that data values are, at least to a first approximation, independent and identically distributed (iid). Whether the values really are independent depends on how data were collected. If for example rollers were tested starting with the lightest, with successively heavy rollers tested at successive locations across the lawn, it is entirely possible that observations could be influenced by a trend in soil properties across the lawn.

4.2 Key questions for the use of models

Key questions are:

- Modeling and analysis
  - Which model?
  - Do we want to make predictions? Or is the interest in getting parameter estimates that are interpretable?
  - How will model performance be measured?
  - How close can we get to measuring the performance that matters?
- Interpretation
  - The task is easier if the aim is prediction, rather than interpretation of model parameters.
  - Can model parameters be interpreted in scientifically meaningful ways?
    [This is a minefield, with huge scope for getting it wrong.]

More detailed comments will now follow on some of the issues raised above.

The choice of method: Note the use of the word “method”, not algorithm. Algorithms specify a sequence of computational steps. Something more than an algorithm is needed, if results are to have some use that generalizes beyond the specific data used.

There are many different methods. How should the analyst choose between them? What are good ways to assess the performance of one or other algorithm? A credible measure of model performance is needed, evaluated on test data that closely reflects the context in which the model will be applied. (Consider the forensic glass data. A test set derived by splitting the sample data in two will not reflect a 2011 application context.)

Which are the important variables? Often, the analyst would like to know which data columns (variables, or features) were important for, e.g., a classification. Could some of them be omitted without loss?

The analyst may wish to attach an interpretation to one or more coefficients? Does the risk of heart attack increase with the amount that a person smokes?

For a meaningful interpretation of model parameters, it is necessary to be sure that:

- All major variables or factors that affect the outcome have been accounted for.
- Those variables and factors operate, at least to a first order of approximation, independently.

In some cases, a different but equivalent choice of parameters will be more meaningful. For working with the **??** data (*DAAG*), the parameters `dist` and `climb` are clearly not independent, making their coefficients difficult to interpret. It is better to work with `log(dist)` and `log(dist/climb)`, which are likely to be very nearly independent.

Rosenbaum (2002) suggests approaches that are often useful in the attempt to give meaningful interpretations to coefficients that are derived from observational data.

4.3 Accuracy assessment

Primarily, the accuracy assessment methods that are demonstrated here assume that the target population is essentially the same as the source population from which the data have been obtained. Even for this limited purpose, there is serious scope for getting answers that can be grossly optimistic.

In practice there will often be differences between source and target populations. This has consequences that require comment:

- Many methods work by starting with a simple form of initial model, which is successively refined. Too much refinement (over-fitting) will lead to a model with reduced predictive power when applied to new data. It is necessary to know when to stop!
- A model that is a good fit for the source population is likely to be sub-optimal for the target population. If the target differs markedly from the source, the fit may be very sub-optimal.
- Accuracy estimates for the source are likely to over-estimate accuracy for the target.
- Most often, we have accuracy comparisons for the source population. What we'd really like are accuracy comparisons for the target. The model that appears best for the source population may not be the model that performs best on the target.
- Where predictive modeling of a comparable type is being carried out repeatedly, the analyst should keep a record of the comparison between after-the-event model performance and predicted performance, e.g., from cross-validation on the original data.
- Often, predictions are successively made ahead in time. If a long enough data series is available, time series methods may be apprropriate. In effect, past changes from one time to the next are used as a guide to likely future changes.

Even where a model does take account of smooth patterns of change over time, do not put too much faith in it. A lesson from the recent financial crisis is, surely, that it is unwise to put much faith in any financial model that does not allow for occasional "shocks". The warnings in Taleb (2004) merit attention.

Three types of outcome will be considered – continuous, binary data (two categories), and count data. These require separate comment:

Continuous outcome data: For regression with a continuous outcome, normal theory accuracy estimates can, if the independent normal error assumptions are not too badly wrong, work quite well. Note however that:

- If the model is selected from a wide class of models, or if there is extensive variable selection (e.g., select the best 3 explanatory variables out of 10), the model is likely to be over-fitted and the accuracy estimates may be grossly optimistic.
- If observations are not independent, accuracy estimates may again be wrong, usually optimistic. The situation can in special cases be rescued by choosing a more realistic model for the "error". Some of the possibilities are:
  - For data that are collected over time, models are available that can account for the likely sequential correlation.
  - Variation is often multi-layered – variation between different countries, variation between humans in an individual country, variation between clinical assessment made on the same human, and so on. Again modeling approaches are available that can account for such different sources of variation.
  - Spatial models are another possibility.

Empirical methods for accuracy assessment can in principal be adapted for use where there is a complex error structure. This does however require a clear understanding of the theoretical issues, and may not be straightforward.

Binary data and count data: Here, the theoretical accuracy estimates that are available for certain of the methods rely on asymptotic approximations. For 'algorithmic' methods, including tree-based methods and random forests, theoretical results have limited relevance. Accuracy assessment almost inevitably relies on empirical methods.

The empirical methods can if used correctly cope with the effects of model and variable selection.

4.4 When are rough and ready methods enough?

Rough and ready methods are fine, if they do the job. How does one know whether the job is done?

- Watch for source/target (eg, 2008/2009) differences.
- Allow for effects of model and variable selection.
- For interpretation of individual model parameters, know the traps.
- Are there dependence issues (time series, . . .)?

Rough and ready methods may yield useful clues that make a start on gaining understanding in new areas. Greater finesse will almost inevitably be needed to make further progress once the low-hanging fruit have been harvested.

4.5 Automation

As much as is reasonable, it makes sense to automate. Attention can then be focused on those aspects of the investigation that are not susceptible to automation.

Analysis methodology can be effectively automated, to a smaller or larger extent, once the analysis has been run a number of times with similar data, and results validated. Even then, it is necessary to be open to new insights, or new wrinkles that emerge in the course of the analysis. Note also the importance of scientific insights that may suggest the form of model.

The automation of aicraft control offers interesting parallels. A huge research and engineering investment is required to allow it to work effectively, and to maintain it. An investment of this magnitude makes economic sense only for very large aircraft such as the A380 – aircraft that in fact would be very difficult or impossible to fly without extensive automation. Such an investment would be hard to justify for a Fokker Friendship.

Note also that automation is most feasible in applications where mistakes can be tolerated, where it is not necessary to be consistently correct.

4.6 The 'Learning from Data' Perspective

Model fitting involves a process that can be described as "learning from data". The terminology seems particularly apt when resampling methods have a substantial part in choosing and comparing models, and in assessing accuracy or other performance measures.

'Statistical Learning' is a name for methods that allow automatic choice of model. It is important to keep in mind conditions that apply to the use of such methods. A brief note on terminology now follows.

Modeling (or a machine?) that learns from the data

In the modeling that is in view here, the data have a substantial role in determining the form of model. The demand for such models arises, in part, from the size of the data sets (many observations, and/or many variables) that are now commonly available.

- The (computing) machine *learns* from data
- Use *training* data to train the machine or software
- Use *validation* data to tune the model
- Use *test* data to check on model performance.

This use of language borrows extensively from the *machine learning* literature.

Cross-validation extends the training/test approach to make more complete use of the data. The downside is that it is less widely applicable than a pure training/test approach.

In the applications of ideas of statistical or machine learning that are prominent in the data mining literature, the aim is usually prediction rather than the obtaining of interpretable model parameters. Hence the name "predictive modeling".

5 Examples that Highlight Key Issues

The example that follow in this section are designed to highlight issues that were raised above, and others besides. They do this without getting into detailed model fitting, such as will be described in later chapters.

5.1 A clearly inappropriate use of least squares modeling

Figure 3 plots historical measurements of the speed of light, as given at `http://en.wikipedia.org/wiki/Speed_of_light` (Oct 15, 2011).

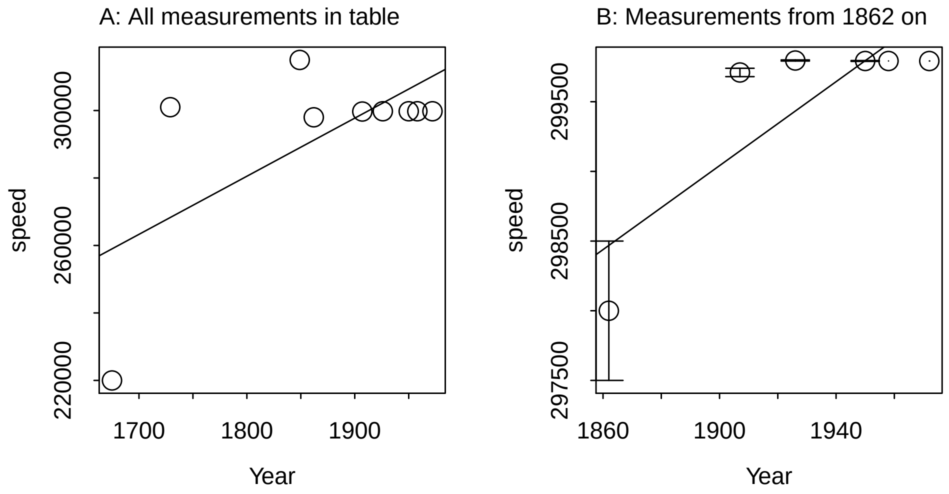

Figure 3: Historical measurements of the speed of light – speed versus year. Panel A shows all measurements. Panel B shows those measurements (1862 or later) for which error bounds are provided. The lines are fitted ignoring dramatic reductions, with time, in the errors. In panel B, error limits for the two final measurements are so small relative to the *y*-limits that they do not appear.

Pretty clearly, both lines are nonsense. Notice also that in panel A the 1675 measurement, almost certainly the least accurate, has the largest leverage in determining the line. In Panel B the 1862 measurement, which has the largest error, has also the largest leverage.

The following stores the data in a data frame:

```
cvalues <- data.frame(
  Year = c(1675, 1729, 1849, 1862, 1907, 1926, 1950, 1958, 1972),
  speed = c(220000, 301000, 315000, 298000, 299710, 299796, 299792.5,
              299792.50, 299792.4562),
  error = c(NA, NA, NA, 500, 30, 4, 3, 0.1, 0.00111)
)
```

Code for a simplified version of Panel A is:

```
plot(speed ~ Year, data=cvalues)
obj <- lm(speed ~ Year, data=cvalues)
abline(obj)
```

Code for Panel B is:

```
subdata <- subset(cvalues, Year>=1862)
ylim <- with(subdata, range(c(speed-error, speed+error), na.rm=TRUE))
plot(speed ~ Year, data=subdata, ylim=ylim)
obj <- lm(speed ~ Year, data=subdata)
abline(obj)
with(subdata, arrows(Year, speed-error, Year, speed+error,
                     angle=90, length=0.1, code=3))
## For the graph as displayed, type
g1_3(data=subset(cvalues, Year>=1862), errors=TRUE)
```

It is in principle possible to do an analysis that attempts to adjust for the widely different error limits. For example, it might be assumed that these error limits are roughly proportional to standard errors under repeated sampling. The measurement errors are, unfortunately, not sufficienty comparable for this to make sense.[2] In several instances, the measurement lies outside of the error limits for the previous measurement, indicating that the error estimates are not accounting for intrinsic biases in the measurement methods.

The regression line enterprise was fatally flawed from the beginning. It mixes older highly inaccurate measurements, with unknown biases, with modern highly accurate measurements where there is much less potential for bias. As noted, it gives the largest 'leverages' in determining the regression line to the oldest and least reliable measurement(s). The regression calculations are just about as misguided as it is possible for such calculations to be!

5.2 Variation between workers, given wages in successive years

Contrast the lawn roller data with the `Wages` dataset (*Ecdat* package). Observations are available, for each worker, at each of 7 successive time points. One possible approach is to calculate an intercept and slope for each worker, and to model these as functions of the explanatory variables.

The following makes the data available, and adds a column that identifies individuals:

```
library(Ecdat)
data(Wages)
Wages$ID <- rep(1:595, each=7)
library(lattice)      # Will use xyplot() from lattice
```

[2] If estimates are unbiased, theory suggests use of weights that are proportional to the inverses of the squares of the errors. Results from such a (flawed) analysis do indeed show no detectable trend with time.

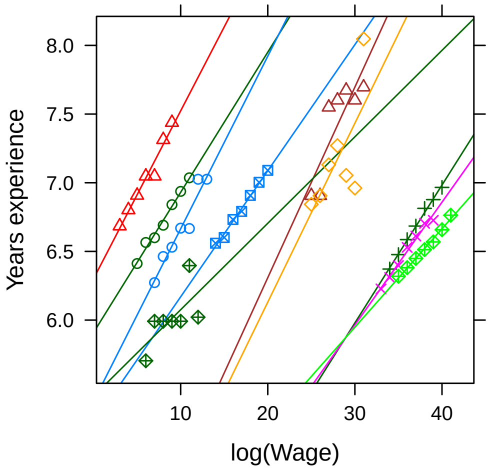

Figure 4:
Wages for 10 randomly chosen individuals, versus years of experience.

```
## Simplified code
set.seed(31)  # gives plot shown
choose10 <- sample(1:595,10)
which10 <- Wages$ID %in% choose10
gph <- xyplot(lwage ~ exp,
              groups=ID,
              data=Wages,
              subset=which10,
              pch=1:10,
              type=c("p","r"))
print(gph)
## For the graph as shown, type
## print(g1_4(seed=31))
```

Plotting the data for all 595 individuals would give a graph that is too crowded to be useful. Some idea of what the data looks like can be gained by taking a random sample of 10 individuals, and plotting their wages (or `log(Wages)`) against years of experience. The graph should identify the different individuals, and perhaps add a separate regression line for each individual. With `lattice` graphics all that is needed is to create a column `ID` that identifies individuals, then adding the argument `groups=ID` when calling `xyplot()`

Figure 4 shows the resulting graph. As one might expect, there is huge variation between workers. Other variables in the model will no doubt account for some of this (how much?).

5.3 Variation within a site, and variation between sites

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

Table 1: 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).

The second column of Table 1 has means of corn yields in blocks of land on the island of Antigua. In comparing yields from different blocks, there are two sorts of comparison. Blocks on the same site should be relatively similar, while blocks in different sites should be relatively more different, as Figure 5 suggests.

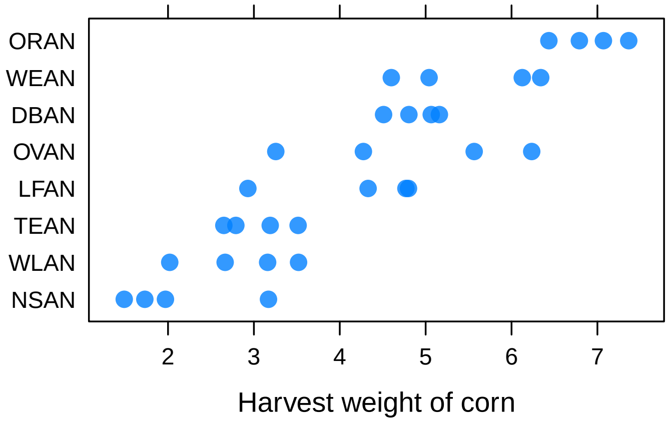

Figure 5: Corn yields are shown for each of 4 parcels of land on each of 8 sites on the Caribbean island of Antigua. Sites have been ordered according to the mean yields.

Code is:

```
## Simplified code
library(lattice); library(DAAG)
Site <- with(ant111b, reorder(site, harvwt, FUN=mean))
gph <- dotplot(Site ~ harvwt, data=ant111b)
print(gph)
## For graph as shown, type:
## print(g1_5())
```

In making a prediction for a new site, variation between sites comes into play, with a large increase in the standard error of the prediction. The iid assumption may be reasonable for site means. For individual parcel yields over all 8 sites, it is clearly wrong. Variation between parcels on different sites is much greater than variation between parcels within a site would suggest.

An easy way to get an appropriate standard error for prediction for the average of 4 parcels at a new site is to work with site means. For any other prediction of this type (e.g., for a single parcel at a new site), a simple form of multi-level form of analysis is required.[3] Multi-level models will not be further considered in this text. The absence of any further explicit mention does not however mean that they are unimportant!

5.4 Record times for athletic races – continuous outcome data

Figure 6 plots, against distance, world record times as at October 2006 for athletic track and road races:

The range of distances and times is huge, from 100m in 9.6sec to 292.2km in 24h. Distances thus vary by a factor of almost 3000, and times by a factor of 9000. As a result, differences from the fitted line that are relatively large, as much as 15% of the estiamted time for some points, are not obvious.

A simplified version of Figure 6, without the arrows and associated labeling, can be obtained as follows.[4]

[3] All this assumes that predictions are restricted to use of the data in this dataset. It is entirely possible, however, that covariate information – perhaps soil type and historic rainfall – could be collected that would explain much of the site to site variation.

[4] To print the graph with annotation as shown, type:

```
gph <- g1_6()
```

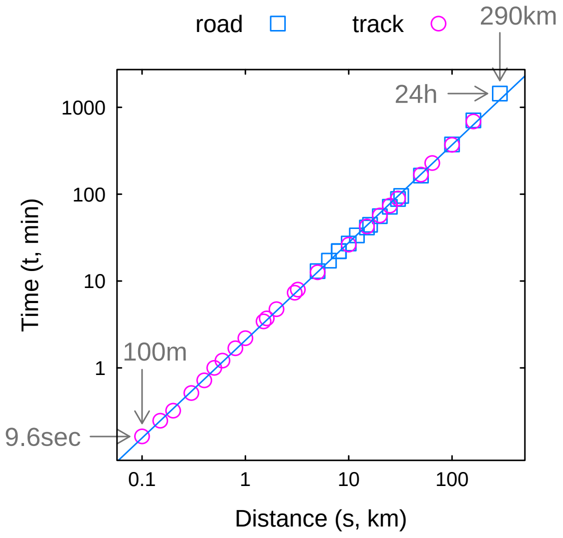

Figure 6: Record times versus distances, with both scales logarithmic, for track and road athletic races. With a ratio of largest to smallest time that is ~3000, differences from the line have to be large to be visually obvious. The plot of residuals shows that, for the longest race, the difference from the line is >15%. (Differences on a scale of natural logarithms, if small, are a little less than fractional differences.) The next figure shows a clear systematic pattern in the residuals.

First create a graphics object that has the points:

```
library(DAAG)
library(lattice)
scaleset <- list(log = 10, tck = -0.4,
                 x = list(at = 10^c((-1):2)),
                 y = list(at = 10^(0:3)))
gph0 <- xyplot(Time ~ Distance, groups = roadORtrack,
               data = worldRecords,
               scales = scaleset)
gph01 <- update(gph0, xlab = "Distance (s, km)",
                ylab = "Time (t, min)",
                auto.key = list(columns = 2))
```

Here, `update()` adds labeling and a key to an initial graph. Below, the layering features of the *latticeExtra* package will be used to add a line to the plot region.

Now create a graphics object that has one line for all data, then using the function `as.layer()` (*latticeExtra* package) to add the contents of its plot region to the graph that has the points:

```
library(latticeExtra)
gph2 <- xyplot(Time ~ Distance, data = worldRecords,
               scales = list(log = 10), type="r")
gph12 <- gph01+as.layer(gph2)
print(gph12)
```

`as.layer(gph2)` creates a layer for 'adding' to the plot region.

Figure 7 examines the residuals from the line in Figure6. Systematic deviations are now very obvious. In the lower panel, a smooth curve has been fitted to the residuals. A *statistical learning* approach has been used to determine the smoothness of the curve.

The following generates the residuals (`resid1`) that are plotted in the upper panel:

```
wr.lm <- lm(log(Time) ~ log(Distance),
            data = worldRecords)
resid1 <- resid(wr.lm)
```

```
trellis.par.set(clip=list(panel="off",strip="on"))
print(gph, position = c(0.05, 0, 1, 0.95))
trellis.par.set(clip=list(panel="on",strip="on"))
```

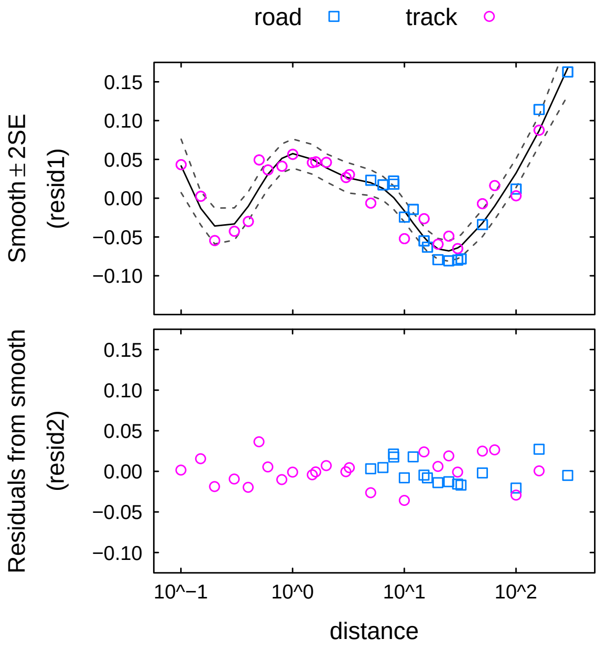

Figure 7:
The upper panel shows residuals (1) from the straight line fit in Figure 6. These show a clear pattern, and are clearly not independent. Fitting a smooth curve to the residuals (1) then yields the residuals (2), shown in the lower panel. The curve is designed to leave, as far as possible, residuals (2) that reflect independent random variation. The independence assumption is especially crucial for the 95% pointwise confidence limits about the curve that are shown in the upper panel.

The following fits a smooth curve to the residuals:

```
library(mgcv)
wr.gam <- gam(resid1 ~ s(log(Distance)),
              data = worldRecords)
hat.gam <- predict(wr.gam, se.fit = TRUE)
wrgamdata <- with(worldRecords,
                  data.frame(distance=Distance,
                             roadORtrack=roadORtrack,
                             resid1=resid1,
                             resid2=resid(wr.gam),
                             hat=hat.gam$fit,
                             se=hat.gam$se.fit))
ord <- with(wrgamdata, order(distance))
wrgamdata <- wrgamdata[ord,]
```

The gam() function (*mgcv*) fits a smooth curve, with automatic choice of smoothing parameter.

Simplified code that plots the figure in the upper panel is:

```
parset = simpleTheme(pch = c(0, 1), lty = c(2, 1, 2),
                     col.line = c("gray30", "black", "gray30"))
ylim <- c(-0.125, 0.175)
gph0 <- xyplot(resid1 ~ distance, groups = roadORtrack,
               data = wrgamdata, ylim = ylim, xlab = "",
               scales = list(x = list(log = 10)),
               auto.key = list(columns = 2))
gph1 <- xyplot(I(hat-2*se) + hat + I(hat+2*se) ~ distance,
               data = wrgamdata, outer = FALSE, ylim = ylim,
               scales = list(tck = -0.4, x = list(log = 10)),
               type = "l", lty=c(2,1,2))
library(latticeExtra)
gph01 <- update(gph0+as.layer(gph1), par.settings=parset,
               ylab = expression(atop(Smooth %+-% 2 * SE,"(resid1)")))
```

```
print(gph01)
```

Code for the lower panel, which plots the residuals (`resid2`) from the smooth, is:

```
ylab2 <- expression(atop("Residuals from smooth","(resid2)"))
gph2 <- xyplot(resid2 ~ distance, groups = roadORtrack,
               data = wrgamdata, ylim = ylim, ylab=ylab2,
               scales = list(tck = -0.4, x = list(log = 10)),
               type = "p", par.settings = parset)
print(gph2)
```

To reproduce Figure 7 with a layout close to that shown, type:

```
gphs <- g1_7()
print(gphs$upper, position=c(0, 0.45, 1, 1))
print(gphs$lower, position=c(0, 0, 1, 0.55) , newpage = FALSE)
## NB: If necessary, adjust the above position arguments
```

Questions that may be asked include:

- Will the pattern be the same in 2030?
- Is it consistent across geographical regions?
- Does it partly reflect greater attention paid to some distances?
- So why/when the smooth, rather than the line?

Clearly the smooth curve (line, with 'corrections' from the line) would be useful to race organizers who wished to estimate the time at which a race winner could be expected to appear. For this purpose, i.e., for interpolating between existing data values, the assumption of independent "errors" about a smooth curve works well.

The methodology generalizes, allowing the fitting of curves and surfaces, in principle in an arbitrary number of dimensions. These notes will use the abilities in the R package *mgcv*, documented in detail in Wood (2006). The ability to fit such curves and surfaces automatically is remarkable, relative to what was available a decade ago.

5.5 Global average temperature – a time series

Figure 8 shows anomalies, in hundredths of a degree C, from the 1951-1980 global average as baseline. Anomalies are plotted against year. The grey curve is a plot of average anomaly up to that point in time. Observe that 1976 was the last year in which the global temperature fell below the average to that time.[5]

Code for Figure 8, but omitting the arrow, is:

```
anom <- loti[, "J.D"]
num <- seq(along=anom)
AVtodate <- cumsum(anom)/num
yr <- loti$Year
plot(anom ~ yr, xlab="Year",
     ylab=expression("Difference from 1951-1980 ("*
     degree*"C)"))
```

[5] These are the GISS (Goddard Institute for Space Studies) Land-Ocean Temperature Index (LOTI) data, from `http://data.giss.nasa.gov/gistemp/tabledata/GLB.Ts+dSST.txt`. They are available in R image format from `http://www.maths.anu.edu.au/~johnm/sldm/data-RData`

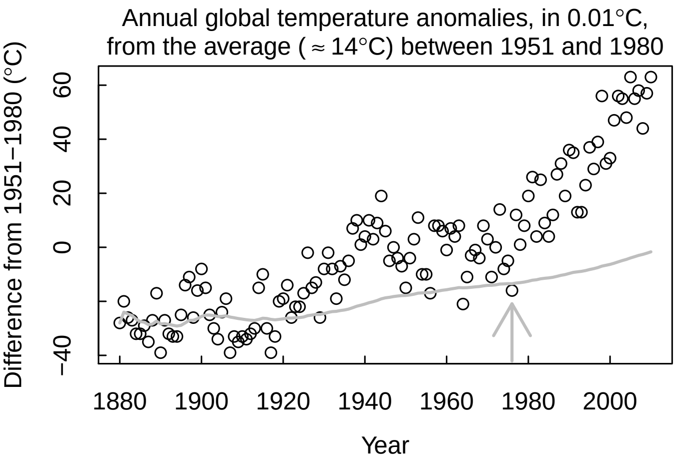

Figure 8: Data are anomalies, in 0.01 °C, from the global average temperature (~14^oC) between 1951 and 1980. The grey curve shows, at each time, the average anomaly up to that time. For years after 1976, global temperatures were in each case above the average up to that year.

```
lines(AVtodate ~ yr, col="gray", lwd=2)
## For the graph as shown, including the arrow, type:
## g1_8()
```

Observe that there have been 34 successive years (from 1977 to 2010 inclusive) in which the global average fell below the average up to that year. Now assume that variations from the underlying trend are independent between years (they are not!). Then under the hypothesis that global temperature is varying randomly about a common mean, the probability of this happening is $2^{-32} = 2.32 \times 10^{-10}$. An argument along these lines was offered by a speaker on the Australian ABC Science Show on April 3 2011.[6]

Mean temperatures are quite strongly correlated from one year to the next. A model that has independent random variation about a trend, whether a trend that is flat or one that steadily moves upwards, is just wrong. ENSO and other climatic oscillation patterns, and perhaps sunspot patterns, lead to local trends, sometimes lasting for as much as a decade, above or below the long-term trend.

Use of least squares to fit time series trend lines can be seriously misleading!

The LOTI data that are shown in Figure 8 are presented by NASA as a fair representation of global mean trends. The GISS LOTI data, because they include Arctic temperatures, may be preferable to the HadCRUT series from the UK; see however `http://www.cru.uea.ac.uk/cru/data/temperature/#faq`. If one fits a line to the HadCRUT data for the cherry-picked years 1998-2008, it shows a relatively flat trend. If one assumes independent variation about the trend and a confidence interval for the slope of the line, a zero slope is well within the limits. What does this prove? As with the specious argument that compared yearly means with the mean to date, pretty much nothing! The confidence limit calculations rely on seriously flawed assumptions.

The above arguments trivialise and misrepresent what is a serious issue for our times, and are the stuff of junk science. Relatively respectable forms of modeling, using standard time series modeling, are of course possible. Note however that statistical evidence is inevitably part only of the story. Results from statistical analysis of historical and current data must

[6] The speaker's argument was even more flawed – he was using monthly data. The global average had fallen below the long-term average for 192 successive months. $2^{-192} = \ldots$

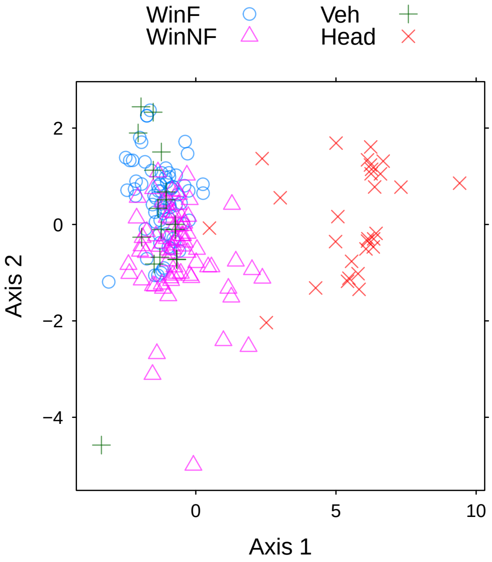

Figure 9:
Visual representation of the classification rule, given by *linear discriminant analysis*, for the forensic glass data. A three-dimensional pattern of separation between categories has been collapsed down to two dimensions. Some categories may therefore be better distinguished than is evident from this figure. Data were collected prior to 1987, and probably much earlier.

be interpreted using whatever light can be shed by the modeling of the physical processes that are involved.[7]

5.6 *Categorical with >2 categories – forensic glass identification:

Now consider a classification example. intended to help illustrate some of the important issues. As is common in many of the examples that are the stock-in-trade of the data mining literature, the interest is in prediction rather than interpretation of model parameter estimates.

The example relates to glass fragments that were collected in the course of forensic work. Numbers of pieces of glass of each of the different types are given:

Window float (70)	Window non-float (76)	Vehicle window (17)
Containers (13)	Tableware (9)	Headlamps (29)

Variables are %'s of Na, Mg, . . . , plus refractive index. In all there are 214 rows of data (observations) × 10 columns (variables).

The aim is to find a rule that predicts the type of any new piece of glass. Figure 9 is a visual summary of the result from the use of a simple form of classification methodology, with the name *linear discriminant analysis*. In order to simplify the plot, the two least frequent glass types – Tabl and Con – have been omitted:

Code for Figure 9 is:

```
library(MASS)
usefgl <- droplevels(subset(fgl, !(fgl$type %in% c("Tabl", "Con"))))
fgl.hat <- predict(lda(type~., data=usefgl))
parset <- simpleTheme(pch=1:4, alpha=0.6)
gph <- xyplot(fgl.hat$x[,2] ~ fgl.hat$x[,1], groups=usefgl$type)
```

[7] As I understand the evidence, modeling of the physical processes establishes that increases in anthropogenic CO_2 and other greenhouse gases in the atmosphere must lead to temperature increases, of a magnitude similar to what has been observed. Other drivers (including changes in solar activity) account, currently, for a relatively minor part of temperature changes.

```
gph <- update(gph, auto.key=list(columns=2),
              xlab="Axis 1", ylab="Axis 2",
              aspect=1, scales=list(tck=0.4), par.settings=parset)
print(gph)
## For the graph as shown, type:
## gph <- g1_9(); print(gph)
```

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

How/when were data generated? (prior to 1987)

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

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

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

These data are not a good basis for making judgements about glass fragments found, in 2011, on a suspect’s clothing. Too much is likely to have changed since 1987. The *source* population is unlikely to be a good match for a 2011 population as target.

We’d want data that are a better match with the glass fragments that one might currently expect to find. We can then generalize with confidence, from the sample from which results have been obtained to some wider population.

There is another way to proceed. The detailed chemical characteristics of the glass found on the suspect can be compared with those of the glass found at the crime scene. One can then ask whether the match is so close that the two sets of fragments are likely to be from the same glass object. This frames the question in a manner that requires much more detailed data, and a different analysis strategy.

In practice, we often have to be content with data that are from a population that is a less than perfect match to the population to which results are to be applied. Any available checks can and should however be applied to investigate the closeness of the source/target match.

Analysis results: The overall classification error, from use of the random forests algorithm that will be described later, 20.1%. The confidence with which different glass types can be classified does however vary greatly from one type to another. For glass types other than Window float and Window non-float, the evidence is rather scanty.

The classification error rates were:

	Classification error
Window float (’WinF’: 70)	0.10
Window non-float (’WinNF’: 76)	0.22
Vehicle window (’Veh’: 17)	0.59
Containers (’Con’: 13)	0.23
Tableware (’Tabl’: 9)	0.22
Headlamps (’Head’: 29)	0.14

6 Data Exploration Prior to Analysis

Before plotting any graphs, one wants to know what data the columns hold. Commonly, columns will be one of:

- numeric, with enough distinct values that the data can be treated as continuous
- numeric, with a small number of values that code for unordered or ordered categories
- character
- factor – which is a common way to store character data. What is stored are integers 1, 2, ... Associated with the factor (as an "attribute") is a table that translates 1 to the first factor level, 2 to the second level, and so on.

Before we do the analyses that will be described, it is helpful to have basic information on the columns in the data, including information on relationships between explanatory variables. The *rattle* GUI is very helpful in this respect. If you load a data frame into *rattle*, it will display basic information on each column.

Basically, we'd like to ensure, if we can, that:

- all columns have a distribution that is reasonably well spread out over the whole range of values, i.e., we want to avoid having most values squashed together at one end of the range, with a small number of very small or very large values occupying the remaining part of the range
- relationships between columns (which, except for the relationship with the outcome variable we prefer to be weak) are roughly linear.

The dataset `mammals` *MASS* is an extreme example of data that cries out for transformation. Figure 10A shows the scatterplot for the raw data, while Figure 10B shows the scatterplot for the logged data.

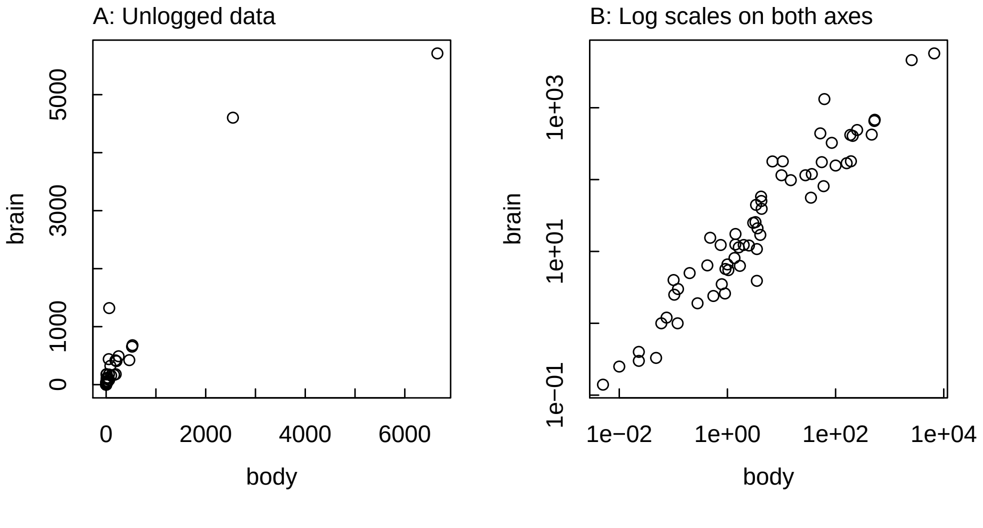

Figure 10: Brain weight (g) versus Body weight (kg), for 62 species of mammal. Panel A shows the unlogged data, while Panel B uses log scales, for both axes. Notice that the scales are labeled in the original (unlogged) units.

Where values are concentrated at one end of the range, the small number (perhaps one or two) of values that lie at the other end of the range will, in a straight line regression with that column as the only explanatory variable, be a leverage point. When it is one explanatory variable among several, those values will have an overly large say in determining the coefficient for that variable.

The commonest situation is where positive (or non-zero) values are squashed together in the lower part of the range, with a tail out to the right. The distribution is then described as skewed to the right. Often, in these circumstances, a logarithmic transformation will remove much or all of the skew. Where transformations can be used to ensure that values in all columns are reasonably spread out over the whole of their range, it will then often turn out that relationships between variables are approximately linear.
Code for Figure 10 is:

```
## Panel A
plot(brain ~ body,
     data=mammals, pty="s")
## Panel B
plot(brain ~ body,
     data=mammals, log="xy", pty="s")
```

For graphs side by side as in Figure 10, type:

```
gph <- g1_10()
print(gph)
```

In most cases, the skewness of the distributions will be less extreme than in Figure 10. There are various types of plots, in addition to simple scatterplots, that can be used to investigate such distributional issues as skewness and the presence of outliers. A key tool, in much of the later discussion, will be the scatterplot matrix. Figure 11 is a scatterplot matrix representation of Figure 10B.

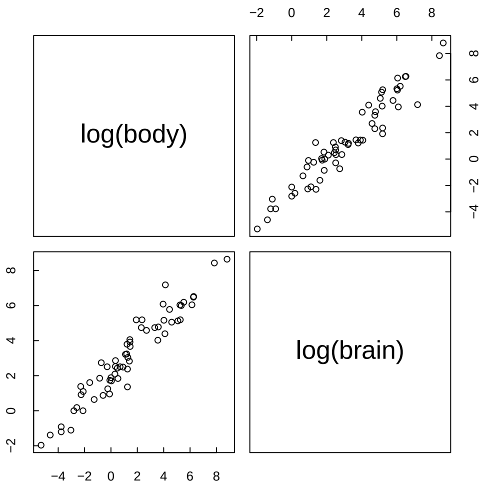

Figure 11:
Scatterplot matrix representation of Figure 10.

```
## Code
pairs(log(mammals),
      labels=c("log(body)",
               "log(brain)"))
## Alternatively, type:
## g1_11()
```

7 Weighting Effects in Summary Data

The discussion that follows highlights weighting effects, of which all data analysts ought to be aware, that are important for their own sake. At the same time it highlights issues that can and commonly arise in the attempt to move beyond prediction to accounting for the contribution of individual model terms for predicted values.

The fitting of a model to data yields a form of data summary. When a model is fitted that generates the same form of data summary, the same spurious effects will appear.

In tables of counts, the effect described here is known as the Yule-Simpson effect. In genetics, it has the name epistasis.

7.1 Simple examples

Consider first a simple fictitious example.

Bias from addition over unequally weighted sub-categories

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).

Several further examples, of this same general character, appear below.

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.

7.2 The UCB Admissions Data

Data are admission frequencies, by sex, for the six largest departments at the University of California at Berkeley in 1973. For a reference to a web page that has the details; see the belp page for `UCBAdmissions`. Type

```
help(UCBAdmissions)      # Get details of the data
example(UCBAdmissions)   # Example code gives tabular and graphical
                         # (using mosaicplot) summaries of the data.
```

Note the margins of the table:

```
str(UCBAdmissions)
```

```
table [1:2, 1:2, 1:6] 512 313 89 19 353 207 17 8 120 205 ...
- attr(*, "dimnames")=List of 3
 ..$ Admit : chr [1:2] "Admitted" "Rejected"
 ..$ Gender: chr [1:2] "Male" "Female"
 ..$ Dept  : chr [1:6] "A" "B" "C" "D" ...
```

Observe that `UCBAdmissions` is a 3-way table, with margins `Admit`, `Gender` and `Dept`.

Here, we will calculate overall admission rates separately for males and females, admission rates by department separately for males and females, and for each department the number of males and females applying, as a proportion of the total number of the relevant `Gender`. The reasoning is that, if different genders have different departmental preferences, overall admission rates for males will be biased towards admission rates for departments that are popular with males, while overall admission rates for females will be biased towards admission rates for departments that are popular with females.

Two functions that will be important for the calculations are `margin.table()` and `prop.table()`.

- The following are the overall admission rates:

```
alltab <- margin.table(UCBAdmissions, margin=c(1,2))
alltab
```

```
          Gender
Admit      Male Female
  Admitted 1198    557
  Rejected 1493   1278
```

Now calculate, for each Gender (margin 2), the proportions admitted and rejected. We require a table with the margin `Gender` (=2). Proportions are calculated across the elements of the remaining margin of the table, which is `Admit`. The proportion admitted provides all the needed information. Hence the restriction to row 1 (`[1, ]`).

```
round(prop.table(alltab, margin=2)[1, ], 3)
```

```
 Male Female
0.445  0.304
```

- The following are the admission rates for the different departments. We require a table with margins `Gender` (=2) and `Dept` (=3). Proportions are calculated across the elements of the remaining margin of the table. We require only the proportion admitted. Hence the restriction to row 1 (`[1, , ]`), which at the same time gives a compact table:

```
round(prop.table(UCBAdmissions, margin=2:3)[1, , ], 3)
```

```
        Dept
Gender       A    B     C     D     E     F
  Male   0.621 0.63 0.369 0.331 0.277 0.059
  Female 0.824 0.68 0.341 0.349 0.239 0.070
```

- Now calculate the numbers of males and females applying to each department. The margins that we require in the table are `Gender` (=2) and now `Dept` (=3).

```
(totbydept <- margin.table(UCBAdmissions, margin=c(2,3)))
```

```
        Dept
Gender     A   B   C   D   E   F
  Male   825 560 325 417 191 373
  Female 108  25 593 375 393 341
```

  Proportions will now be calculated for the margin `Dept` of the table `totbydept` just obtained:

```
round(prop.table(totbydept, margin=1), 3)
```

```
        Dept
Gender       A     B     C     D     E     F
  Male   0.307 0.208 0.121 0.155 0.071 0.139
  Female 0.059 0.014 0.323 0.204 0.214 0.186
```

Do the data provide evidence, across the University as a whole, of sex-based discrimination? A relatively small proportion of females (5.9%) applied to department A where admission rates were relatively high, while a high proportion (32.3% and 21.4% respectively) applied to departments C and E where admission rates were relatively low. The very high number of males applying to departments A and B has biased the male rates towards the relatively high admission rates in those departments, while the relatively high number of females applying to departments C, D and F biased the overall female rates towards the low admission rates in those departments. The overall bias arose because males favored departments where there were a relatively larger numbers of places.

The results that give the overall proportions are, for these data and depending on the intended use, an unsatisfactory and potentially misleading summary. The phenomenon that they illustrate, known as Simpson's paradox or as the Yule-Simpson effect, is discussed in Aldrich (1995); Simpson (1951).

Here, information was available on the classifying factor on which it was necessary to condition. This will not always be the case. It is always in principle possible that there is some further variable that when conditioned on can change an observed association, perhaps even reversing it. It may of course be argued, as perhaps in the present case, that the processes are well enough understood that substantial effects from further variables are unlikely.

7.3 Analysis of a substantial dataset – US accident data

The US National Highway Traffic Safety Administration (NHTSA) 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 frame `nassCDS` (*DAAG*) is derived from NHTSA data for the years 1997 – 2002.[8]

Data collection used a complex sampling scheme in which the sampling fraction differs between observations. Each point has to be multiplied by the relevant sampling fraction, to

[8] They hold a subset of the columns from a corrected version of the data analyzed in Meyer (2006). See also `help(nassCDS)`. More complete data are available from one of the web pages
`http://www.stat.colostate.edu/~meyer/airbags.htm` (SAS transport file)
or `http://www.maths.anu.edu.au/~johnm/datasets/airbags/` (R image file).

get a proper estimate of its contribution to the total number of accidents. The column `weight` (`national` = *national inflation factor* in the SAS dataset) gives the relevant multiplier.

Meyer (2006) argues that on balance (over the period when their data were collected) airbags cost lives. After adjustment both for seatbelt use and for speed of impact, airbags appear on balance to be dangerous. The apparent effects, most serious in high impact accidents, are however at the level of statistical error.

Strictly, the conclusion is that, conditional on involvement in an accident that was sufficiently serious to be included in the database, and conditioning also on `seatbelt` (seatbelt use or not) and `dvcat` (force of impact) there is a suggestion that airbags are harmful. Conditional on the airbag failing to prevent an outcome that is somewhat serious, there is a suggestion that airbags are harmful!

Meyer's analysis is open to serious challenge. Farmer (2005) argues that these data have too many uncertainties and sources of potential bias to give reliable results when analyzed using Meyer's approach. Additionally, there are other factors on which the effects of airbag use could and perhaps should be conditioned.

7.3.1 An analysis in the style of Meyer

The analysis that follows will use a subset of the data used by Meyer. The oldest vehicle was from 1953, while the oldest with an airbag was from 1986. In an analysis that does not allow for age of vehicle, results for vehicles without airbags may be biased towards results for older vehicles. If there is an adjustment for age of vehicle, vehicles that are much older than 1986 will not contribute useful information on the effectiveness of airbags. Hence vehicles older than 1986 will omitted, together with observations with weight 0, and one observation where the year of vehicle was unknown. This omits 2726 records out of the total of 26217, leaving 23491 records.

```
library(DAAG)
nassnew <- subset(nassCDS, !is.na(yearVeh) & yearVeh>=1986 & weight>0)
```

Survival rates, according to airbag use: The following estimates numbers of front seat passengers alive and dead, classified by airbag use:

```
library(DAAG)
(abtab <- xtabs(weight ~ dead + airbag, data=nassnew))
```

```
       airbag
dead       none  airbag
  alive 4357430 6614169
  dead    29897   25919
```

The function `prop.table()` can then be used to obtain the proportions in margin 1, i.e., the proportions dead broken down by airbag use:

```
round(prop.table(abtab, margin=2)["dead", ], 4)
```

```
  none airbag
0.0068 0.0039
```

```
## Alternatively, the following gives proportions alive & dead
## round(prop.table(abtab, margin=2), 4)
```

The above might be taken to suggest that the deployment of an airbag substantially reduces the risk of mortality.

```
abSBtab <- xtabs(weight ~ dead + seatbelt + airbag, data=nassnew)
## Take proportions, retain margins 2 & 3, i.e. airbag & seatbelt
round(prop.table(abSBtab, margin=2:3)["dead", , ], 4)
```

```
         airbag
seatbelt    none airbag
  none    0.0180 0.0155
  belted  0.0039 0.0021
```

The results are now much less favorable to airbags. To see why, consider:

```
margin.table(abSBtab, margin=2:3)  # Add over margin 1
```

```
         airbag
seatbelt     none   airbag
  none     916169   885635
  belted  3471158  5754453
```

In the overall table, the results without airbags are mildly skewed (~4.12:1.37) to the results for `belted`, while with airbags they are highly skewed (~57.6:8.86) to the results for `belted`.

Taking Account of Estimated Force of Impact: Now take account, additionally, of estimated force of impact (`dvcat`):

```
ASdvtab <- xtabs(weight ~ dead + seatbelt + airbag + dvcat,
                     data=nassnew)
## Use ftable to get a compact, flattened version of the table
round(ftable(prop.table(ASdvtab, margin=2:4)["dead", , , ]), 6)
```

```
                 dvcat  1-9km/h    10-24    25-39    40-54      55+
seatbelt airbag
none     none          0.000000 0.002583 0.020300 0.040323 0.204534
         airbag        0.004023 0.004873 0.010982 0.075990 0.269959
belted   none          0.000000 0.000380 0.005743 0.028141 0.139204
         airbag        0.000000 0.000195 0.003331 0.022666 0.157394
```

It will be apparent that differences between `none` and `airbag` are now below any reasonable threshold of statistical detectability.

7.3.2 More Variables Still

There are at least two other variables that may affect the risk of death. These are the year of manufacture of the vehicle, and the age of the occupant. Possibly also the year of the accident might be important, but the data do not have enough information to allow this effect to be modeled in addition to all the others.

Subsection 7.3 uses a generalized linear model (GLM) to analyse these data. This use of regression methodology gives results that are closely equivalent to those from the tabulations that appear above. The choice between different possible tabulations becomes a choice between different possible sets of explanatory factors in a regression equation.

7.4 Alternative and better approach – a ratio of ratios

Farmer (2006) presented results from a different analysis that used front seat passenger mortality, in cars without passenger airbags, as a standard against which to compare driver mortality. In the absence of any effect from airbags, the ratio of driver mortality to passenger mortality should be the same, irrespective of whether or not there was a driver airbag.

An airbag is counted as available (really, potentialy available) if a bag had at some previous time been installed. The breakdown that is based on "availability" is checking the effectiveness of installing airbags into cars.

This analysis uses a data series, intended to document all accidents in which there was at least one fatality. The url `http://www-fars.nhtsa.dot.gov/Main/index.aspx` gives access to a web-based interface to the annual datasets. Table 2 shows the results for 1998:

	DrivDead	PassDead	Dratio
Driver-abag	423	514	0.82
Driver-none	2722	2897	0.94

Table 2: Right front seat passenger death counts in 1998 compared with driver death counts, with and without driver air bags. The comparison is for vehicles that did not have passenger airbags.

Code that reproduces the table is:

```
library(xtable)
ratioTab <- pairedRatios(fars1998, compareBY="airbag-avail")
ratioTab[,3] <- round(ratioTab[,3],3)
```

One would expect factors such as seatbelt availability and velocity of impact to affect passengers and drivers similarly. The ratio of driver deaths to passenger deaths is 0.823 when airbags are fitted, against 0.940 without airbags. Airbags appear then to have reduced driver deaths from 0.940 of passenger deaths to 0.823 of passenger deaths, i.e., by a factor of 0.8755, or a reduction of 12.45%.

Figure 12 shows the results graphically.

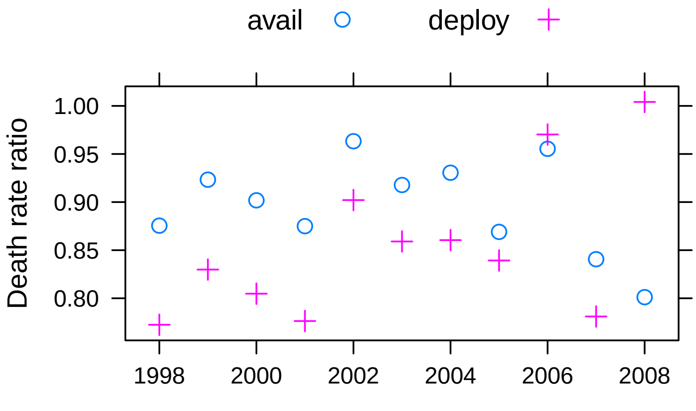

Figure 12: Driver death rate, as a fraction of passenger death rate: (a) when an airbag was not available; (b) when an air bag, even if available, did not deploy.

Calculations use the functions `pairedRatios()` and `yearlyratio()` (both from *modregR*). Code is:

```
matchedratio <- funratio(progress=FALSE)
dfratio <- data.frame(annual_airbag)
nam <- colnames(annual_airbag)
names(dfratio) <- c("airbag-avail"="avail", "airbag-deploy"="deploy")[nam]
dfratio$year <- as.numeric(rownames(annual_airbag))
library(lattice)
parset <- simpleTheme(pch=c(1,3), cex=1.05)
gph <- xyplot(avail+deploy ~ year, data=dfratio,
              xlab="", ylab="Death rate ratio",
              auto.key=list(columns=2),
              par.settings=parset)
print(gph)
## Alternatively, type
## print(g1_12())
```

There may of course be some effect from other factors on the driver to passenger death ratio. These can be checked – broad checks should be adequate in revealing anything of consequence.

Notice that, with the exception of 2006 and 2008, the relative death rate when the bag deployed is less than the relative death rate when an airbag was available.

7.4.1 The moral of this story is . . .

In the analyses just considered, there were many factors that could influence the risk of death – so many that the attempt to account for them all led to an analysis in which statistical noise made it impossible to reach useful conclusions about the parameter of interest. After adjusting for the effect of seatbelts, there were a number of other factors whose effects were similar to or greater than the seatbelt effect that was the chief interest of the analysis

A much simpler analysis was however possible, allowing an estimate that was pretty much unaffected by the factors that affected the death rate itself. This neatly avoided problems that, in the death count summary tables, arose from the need to account for many different factors.

Data summarisation, or an equivalent use of regression methodology, cannot deliver a clear conclusion in cases where there are many different possible choices of the regression equation, and where the conclusion depends on the equation that is chosen. If an alternative is available that can avoid or largely avoid the need to account for many different explanatory variables, it should be preferred.

7.5 Summary of continuous outcome data - an example

Unequal subgroup weights create exactly the same potential, as with binary (or categorical) outcome data, for misleading summary tables.

Unequal subgroup weights with continuous data – an example

Figure 7.5 relates to data collected in an experiment on the use of painkillers.[9]. Researchers were investigating whether bacloen, administered following operation (additional to earlier painkiller), reduced pain? Notice that the overall comparison (average for baclofen versus

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

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.[10] 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.

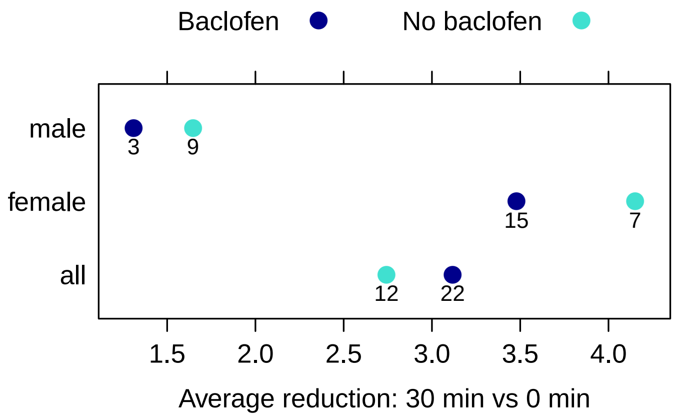

Figure 13:
Data are pain reduction scores. Averages for "all" ignore sex. Subgroup numbers (below each point) weight the "all" value for baclofen towards the high female average, and for no baclofen slightly towards the low male average.

The overall averages in Figure 7.5 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.

Code for Figure 13 is:

```r
library(lattice)
parset <- simpleTheme(cex=1.35, pch=16,
                      col=c("darkblue","turquoise"))
gabalong <- data.frame(values=unlist(gaba["30",])[-1],
                        sex=rep(c("male", "female", "all"), rep(2,3)),
                        trt=rep(c("Baclofen","No baclofen"),3))
gph <- stripplot(sex~values, groups=trt, data=gabalong,
                 par.settings=parset,
                 xlab=list("Average reduction: 30 min vs 0 min",
                 cex=1.0),
                 scales=list(cex=1.0),
                 panel=function(x,y,...){
                     panel.stripplot(x,y,...)
                     ltext(x,y,paste(c(3,9,15,7,22,12)), pos=1, cex=0.8)
                 }, auto.key=list(columns=2, points=TRUE, cex=1.0))
plot(gph)
## Alternatively, type:
```

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

```
## gph <- g.13(); print(gph)
```

Strategies

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

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

Males: 1.311 - 1.647 = -0.336

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

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

7.6 Biases from omission of features – further comments

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.

Biases from Errors in Explanatory Variables

Biases, or even spurious effects, may arise because some variables are measured with a large error. Even if there is no bias in the error, regression relationships in which those variables appear as explanatory variables may have coeficients that are biased, os show spurious effects. See Maindonald & Braun (2010, Section 6.7).

Biases in regression coefficients

These same biases can, and commonly do, affect regression coefficients. Omission of explanatory variables from a regression relationship can bias the estimates of the coefficients of the variables that remain. Estimates of coefficients are conditional on the variables that are included.

8 Exercises

1. Table 3 shows a game of cricket scenario in which bowler A does better than bowler B in each innings, but ends with a worse overall average – 28 runs per wicket as opposed to 20.

	1st innings			2nd innings			Overall		
	R	W	*RPW*	R	W	*RPW*	R	W	*RPW*
Bowler A	40	4	*10.0*	240	6	*40.0*	280	10	*298.0*
Bowler B	70	5	*14.0*	50	1	*50.0*	120	6	*20.0*

Table 3: Runs (R), wickets (W) and runs per wicket (*RPW*), 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.

Table 4 shows an alternative way to make the comparison between bowlers:

	1st innings		2nd innings		Overall	
	RPW	*Wickets*	RPW	*Wickets*	RPW	*Wickets*
Bowler A	10.0	(4)	40.0	(6)	$\frac{10+40}{2} = 25$	(10)
Bowler B	14.0	(5)	50.0	(1)	$\frac{50+14}{2} = 32$	(6)

Table 4: Runs per wicket (*RPW*) for each bowler in the two innings.

Table 4 weights the contributions (av1 and av2) of the two innings equally. What is the relative weighting in $\frac{\text{Total runs}}{\text{Total wickets}}$? [Hint: $\frac{\text{Total runs}}{\text{Total wickets}} = \frac{\text{av1}\times\text{wickets1}+\text{av2}\times\text{wickets2}}{\text{wickets1}+\text{wickets2}}$]

References

Aldrich, J., 1995. Correlations genuine and spurious in Pearson and Yule. *Statistical Science*, 10:364–376.

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>`]

Farmer, C.H. 2005. Another look at Meyer and Finney's 'Who wants airbags?'. *Chance*, 19:15–22.

Maindonald, J. H. and Braun, W.J. 2010. *Data Analysis and Graphics Using R – An Example-Based Approach*, 3^{rd} edition, Cambridge University Press.
`<URL:http://www.maths.anu.edu.au/~johnm/r-book.html>`
[This is aimed at practicing scientists who have some modest statistical sophistication, and at statistical practitioners. It demonstrates the use of the R system for data analysis and for graphics.]

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

Meyer, M.C. and Finney, T. 2005. Who wants airbags?. *Chance* **18**:3-16.

Meyer, M.C.. Commentary on "Another look at Meyer and Finney's 'who wants airbags?'". *Chance*, 19:23–24, 2006.

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.]

Simpson, E. H., 1951. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B*, 13:238–241.

Taleb, Naseem. 2004. *Fooled By Randomness: The Hidden Role Of Chance In Life And In The Markets.* Random House, 2ed.
[Has many insightful comments about the over-interpretation of phenomena in which randomness is likely to have a large role.]

Wood, S. N. 2006. *Generalized Additive Models.* An Introduction with R. Chapman & Hall/CRC.
[This has an elegant treatment of linear models and generalized linear models, as a lead-in to generalized additive models. It is an almost indispensible reference for use of the `mgcv` package for R. Refer to it for the theory of and practical use of regression splines, various types of penalized splines, thin plate splines, etc. A final chapter is devoted to Generalised Additive mixed models.]

Brief Notes on R Documentation

Official Documentation: "An Introduction to R", written by the R Development Core Team. To download an up-to-date copy, go to CRAN.

Web-based Documentation: See Documentation on the web page `http://www.r-project.org`

Note the R Wiki (`http://wiki.r-project.org/rwiki/doku.php`) and the extensive collection of help information that is listed under Other (`http://www.r-project.org/other-docs.html`).

For examples of R graphs, see `http://addictedtor.free.fr/graphiques/`.

R News: Successive issues of *R News* contain much useful information. These can be copied down from one of the CRAN sites.

Contributed Documentation: There is an extensive collection of user-written documents on R that can be accessed by going to this same mirror site, and clicking (under Documentation) on **Contributed**. See also the links that John Fox gives on the web page for his book that is noted under the reference for his book.

Books: See `http://www.R-project.org/doc/bib/R.bib` for a list that is updated regularly.