

This article was written for *CTI Maths & Stats*, **9**, (1998), 17–20.

My experience of using S-Plus for teaching, and Introduction to S-Plus for Generalized Linear Modelling worksheets

P.M.E.Altham, Statistical Laboratory, University of Cambridge CB2 1SB
Director of Studies for the M.Phil. in Statistical Science
(commencing October 1, 1998)

P.M.E.Altham@statslab.cam.ac.uk

<http://www.statslab.cam.ac.uk/~pat/>

How I got started

I began to learn the ‘statistical system’ S-Plus in the summer of 1992, and then was able to use it for teaching, admittedly with more enthusiasm than expertise, in the academic year 1992-3.

Although I came to S-Plus having been quite a happy and confident user of such statistical computing packages as Glim and Genstat, and having had a nodding acquaintance with SPSS and BMDP, I have to admit that I found it quite hard to ‘get my mind around’ the S-Plus syntax at first, probably because of its object-orientated nature. For example, after Glim or Genstat the S-Plus way of doing an ordinary linear regression seems cumbersome and heavy-handed. Those of us who can understand the Glim directives

```
$yvar y$fit x$d e r$
```

will initially be irritated at the length of the equivalent S-Plus commands

```
lm.first <- lm(y~x)
summary(lm.first,cor=F)
lm.first$residuals
```

Now

```
lm.first
```

is our S-Plus *object*, and will remain as such until explicitly deleted. We have put the results of our linear regression of y on x into this particular object through the *assignment operator*

```
<-
```

which should be read as an arrow pointing from right to left. (The symbol ‘ \leftarrow ’, which is one keystroke less, has the same interpretation, but suffers from *looking* symmetrical, although it is interpreted as asymmetrical by S-Plus.)

We can now get excellent diagnostic plots for this linear model (residuals, qqplots etc) simply by

```
plot(lm.first,ask=T)
```

A small data-set, and some graphics

For example, consider the small data set given in Table 1, below. This was published in The Daily Telegraph on February 23, 1998 and shows the ‘Medals Table’, being the final standings of 22 countries in the Nagano Winter Olympics. The final column on this table was not published in the Telegraph; I have added it for a bit of fun. It records whether or not a country is ‘English-speaking’, and was included at the suggestion of the students to make the UK look slightly less miserable in the Chernoff faces example!

Table 1.

	Gold	Silver	Bronze	Eng-sp
Germany	12	9	8	0
Norway	10	10	5	0
Russia	9	6	3	0
Canada	6	5	4	1
USA	6	3	4	1
Holland	5	4	2	0
Japan	5	1	4	0
Austria	3	5	9	0
S.Korea	3	1	2	0
Italy	2	6	2	0
Finland	2	4	6	0
Switzer.	2	2	3	0
France	2	1	5	0
CzechR	1	1	1	0
Bulgaria	1	0	0	0
Sweden	0	2	1	0
Denmark	0	1	0	0
Ukraine	0	1	0	0
Belarus	0	0	2	0
Kazakh.	0	0	2	0
Australia	0	0	1	1
Belgium	0	0	1	0
UK	0	0	1	1

Note: 2 Golds, 2 Silvers and 2 Bronzes were awarded, respectively in the 2-man bob-sleigh, the men’s super-G, and the 4-man bobsleigh.

It is rather odd to use linear regression for this dataset, but we do so to illustrate some of the standard diagnostic plots for linear models, regressing the number of Gold medals on the number of Silver, Bronze medals, for which the fitted regression equation is

$$Gold = -0.0012(.6400) + 0.9166(.1855)Silver + 0.1849(.2225)Bronze.$$

As we might expect for such data, the errors are rather non-normal.

Figure 1 below shows the plots of Gold against each of Silver and Bronze, with some of the country labels given, together with the qqplot and the Cook's distances for the linear regression of Gold on Silver and Bronze. It was obtained by

```
par(mfrow=c(2,2),mar=rep(6,4))
plot(Silver,Gold); identify(Silver,Gold,country)
plot(Bronze,Gold); identify(Bronze,Gold,country)
plot(lm(Gold~Silver + Bronze ),ask=T)
```

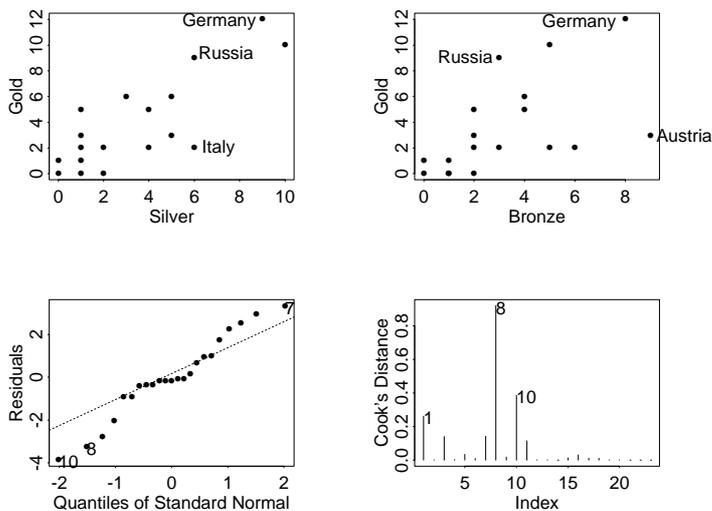


Figure 2 below shows a contour plot of the ‘z-variable’ Gold against the (x, y) variables Bronze and Silver. It was constructed by the following sequence of commands, which themselves follow the pattern of an example in Venables and Ripley (1994, p69).

```
data _ read.table("medals",header=T) ; attach(data)
z _ Gold; x _ Bronze ; y _ Silver ; country _ row.names(data)
topo.loess _ loess(z~ x*y)
topo.mar _ list(x=seq(0,11,0.5),y= seq(0,11,0.5))
topo.lop _ predict(topo.loess,expand.grid(topo.mar))
par(pty = "s")
contour(topo.mar$x,topo.mar$y,topo.lop,xlab="bronze", ylab ="silver",
levels=seq(0,12,2),cex=0.7)
points(x,y)
identify(x,y,country)
```

(Note that the plotting symbols that you see on the screen as the result of ‘points()’ may be replaced by a different symbol when you get a printout of the plot.)

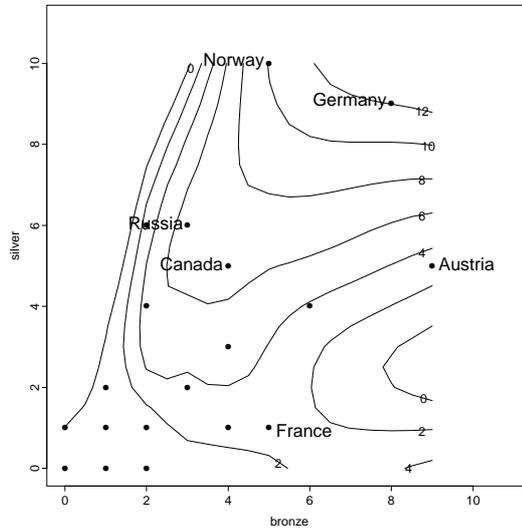
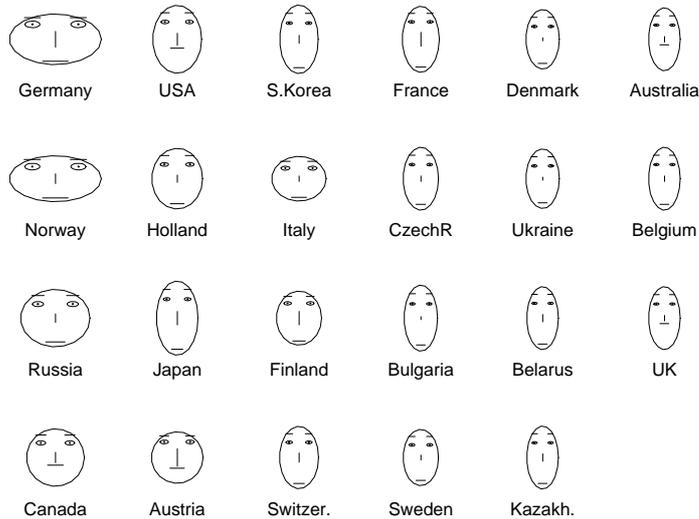


Figure 3 below shows Chernoff's faces for the 22 countries, constructed by using all 4 of the above columns, via the commands

```
a _ as.matrix(data)
faces(a, labels=country, nrow=4, ncol=6)
```

This (default) setting of 'faces()' uses the four variables Gold, Silver, Bronze and Eng-sp in the order 1=area of face, 2= shape of face, 3= length of nose, 4= location of mouth. (We could have drawn the faces using a total of 15 different variables, had they been available.)



Why should we go to the trouble of learning S-Plus?

Of course to compare Glim with S-Plus is like comparing a bicycle with a Rolls Royce, and there will be some tasks (like finding one's way round a small city like Cambridge) for which a bicycle will be a far cheaper and more effective means of transport than a large, expensive and powerful car. Where S-Plus will come into its own is in its breadth and versatility, its ease of handling every conceivable variety of statistical problem, its excellent facilities for matrix manipulation and its beautiful graphics. Needless to say, the graduate students that I was 'teaching' tended to pick it up far more quickly than I did. Most of these students had only rather modest previous experience at undergraduate level of statistical computing, for example with Glim or Minitab. In addition to using S-Plus for teaching statistical inference in Cambridge, I also had a very enjoyable 2 weeks in August 1997 teaching Generalized Linear Modelling via S-Plus at the International Summer School at the University of Jyväskylä, Finland. Thus a significant proportion of all the students I have taught did not have English as their first language. This fact is relevant for a rather 'wordy' language like S-Plus, and for example a Swedish computer keyboard did at first give us problems: it did not at first recognise the crucial symbol \sim .

Sources of help in learning

All my students have taught me a lot, over the years, both in terms of finding my way around S-Plus and of course much wider problems. But at the beginning I was trying to learn mainly from books and notes. I sorely missed an S-Plus equivalent of the book by Aitkin et al (1989) from which I had learned Glim in order to teach it, some years earlier. But in these circumstances one finds that a patient and saintly colleague close at hand is far more helpful than scores of text-books or fancy on-line 'help' systems. When one is quite stuck, perhaps because one has provoked some of the incomprehensible S-Plus 'warning messages', which are such a trial to the beginner, it is a great blessing to be able to turn to intelligent and patient human being. Such a person is Dr Richard J. Gibbens, our Royal Society University Research Fellow, who generously shares his S-Plus expertise. See

<http://www.statslab.cam.ac.uk/~richard/>

for Richard's lecture notes on 'Case Studies in S-Plus', a course which he gives to our graduate students.

The first text-book I used was 'Statistical Models in S', edited by J.M. Chambers and T.J. Hastie (1992). This was undoubtedly helpful, but it somehow seemed to assume that the reader had an awful lot of *time*. For example, I remember thrashing about in total exasperation to try to find exactly what the output for estimates of factor effects was actually telling me. The problem here is that S-Plus uses the Helmert parametrisation rather than a more conventional one such as $\theta_1 = 0$, which is the 'corner-point' constraint familiar to Glim and Genstat users. This problem can easily be fixed by use of an appropriate 'options()' command, but you could waste some time before you realised this. My

S-Plus teaching was made immeasurably easier by the appearance at the end of 1994 of the excellent text book ‘Modern Applied Statistics with S-Plus’ by W.N.Venables and B.D.Ripley; among its many helpful features is the substantial library of data-sets available as

`library(MASS)`

Venables and Ripley’s book, and the 1997 second edition of this book, have been an invaluable source to me, and are very good value for money: the first edition was under 40 pounds for 462 pages, and the second edition is 35 pounds for 548 pages. However both this book, and the Chambers and Hastie book mentioned above, can be quite daunting to a beginner by virtue of their length, depth, and level of detail.

My S-Plus worksheets

My worksheets, which are available on my homepage, have evolved from several years’ experience with teaching graduate students statistical inference via S-Plus. They are intended as a *short* introduction to S-Plus, rather along the lines of BBC tapes/booklets for ‘Get by in Russian’, ‘Get by in German’ and so forth. Just as the BBC booklets do not have a detailed description of Russian or German syntax, nor do my worksheets have a detailed description of S-Plus syntax. In these worksheets I do treat S-Plus as the means to an end, the end being statistical inference, rather than as an end in itself. To a certain extent students will pick up what they need to know about the syntax of the language by this ‘plunge-in’ method. Later on, the interested student may read the appropriate introductory chapters in Venables and Ripley’s book to get an authoritative description of the syntax. These worksheets are not aimed at professional statisticians, but at (mathematically able) graduates who may be intending to become professional statisticians. A companion set of ‘solutions’ to these worksheets is under preparation: this will describe the statistical interpretation of the analyses, together with some basic points about report preparation. For example, students find it helpful to be told at an early stage

- i) how to LaTeX a document
- ii) how to incorporate S-Plus graphs into the LaTeX report.

Remarks on practical details

Note that S-Plus is a *high level* language. One consequence of this for the unwary beginner is that she may unwittingly use the following as S-Plus objects:

`T, F, c, t, row`

and such use does provoke ‘warning messages’ since all these symbols have already been allocated a meaning within S-Plus, for example

`t(X)`

produces the transpose of the matrix X .

The on-line help system in S-Plus is very good, and contains a wealth of scholarly information on a myriad of statistical techniques. It suffers from its sheer *wordiness*: at present it fails to put the university teacher out of a job, since it makes very little use of mathematical symbols or illustrative diagrams. This will no doubt be remedied in due course.

S-Plus is expensive (though prices seem to be coming down, it seems there is now a 'student' version available for 50 pounds). Great interest is generally shown by my audience when I mention the free 'look-alike' version, R, available via the World-Wide-Web. I understand that R has a fundamental design feature, in the way it uses memory, that is greatly superior to that of S-Plus, but I have been happy to use S-Plus, although it must be said that I am not generally working with huge data-sets. The Statistical Laboratory started with only a 4-user license S-Plus mainly for the use of our graduate students, who were then following the Diploma in Mathematical Statistics course. We now have a 16-user license for our Unix workstations platform. This seems to work well for 95% of the time. (The 16-license user is a strict upper bound, since if a 17th individual tries to use S-Plus, his/her access is barred. It is a simple matter to find out who are the current users of S-Plus, so that appropriate action can be taken: occasionally an inexperienced user has forgotten to 'come out of' S-Plus, so is still using up one of the 16 'slots'.) We can also demonstrate our S-Plus computing in seminars with an appropriate projector.

In quite a short time, say about 6 months, good students are able to combine S-Plus and LaTeX with their knowledge of statistical techniques to produce Applied Project reports that look very polished and professional; these will obviously be very suitable pieces of work to show prospective employers.

REFERENCES

- Aitkin M, Anderson D, Francis B and Hinde J. *Statistical Modelling with GLIM*. Oxford: Oxford University Press 1989
- Chambers J M, and Hastie T J, eds *Statistical Models in S*. New York: Chapman and Hall. (Formerly Monterey: Wadsworth and Brooks/Cole.)1992
- Ripley B D, *Review of S-Plus for Windows version 4.0*, Maths&Stats v 8 n4 Nov 1997
- Venables W N, Ripley B D, *Modern Applied Statistics with S-Plus*, Springer-Verlag 1997 (also first edition, 1994)