A S-PLUS: An introductory session

The following session is intended to introduce to you some features of the S-PLUS environment by using them. Many features of the system will be unfamiliar and puzzling at first, but this will soon disappear.

`login:`
`...`
`> ls -a`
`> ls -a .Data`
	Login, start your windowing system (ask a demonstrator if you need help), and check that your working directory has a subdirectory `.Data`, which in turn contains the files `.First`, `.Last` and possibly `.Audit`.

	You should also have the file `morley.data` in your working directory. If not, see a demonstrator. If you have, proceed.
`> Splus -e`	Start `Splus` with the inbuilt command line editor enabled.

	The S-PLUS program begins, with a banner.
	(Within S-PLUS the prompt on the left hand side will not be shown to avoid confusion.)
`help.start()`	Start the help facility. You should briefly explore the features of this facility with the mouse. Standard X window conventions apply.
	Iconify the help window and move on to the next part.
`motif()`	Turn on the graphics window. You may need to re-position and re-size to make it convenient to work with both windows.

`x <- rnorm(50)` `y <- rnorm(x)`	Generate two pseudo random normal vectors of $x-$ and $y-$coordinates.
`hull <- chull(x, y)`	Find their convex hull in the plane.

`plot(x, y)`
`polygon(x[hull], y[hull],`
`  dens=15)`
	Plot the points in the plane, and mark in their convex hull.

`objects()`	See which S-PLUS objects are now in the `.Data` directory.
`rm(x,y,hull)`	Remove objects no longer needed. (Clean up).

`x <- 1:20`	Make $x = (1, 2, \ldots, 20)$
`w <- 1 + sqrt(x)/2`	A 'weight' vector of standard deviations.

`dummy <- data.frame(x=x,`
`  y= x + rnorm(x)*w)`
`dummy`
	Make a *data frame* of two columns, x and y, and look at it.

`fm <- lm(y~x, data=dummy)` `summary(fm)`	Fit a simple linear regression of y on x and look at the analysis.

`fm1 <- lm(y~x, data=dummy,`
`  weight=1/w^2)`
`summary(fm1)`
	Since we know the standard deviations, we can do a weighted regression.

`lrf <- loess(y~x, dummy)`	Make a nonparametric local regression function.
`attach(dummy)`	Make the columns in the data frame visible as variables.

```
plot(x, y)
```
Standard point plot.

```
lines(x, fitted(lrf))
```
Add in the local regression.

```
abline(0, 1, lty=3)
```
The true regression line: (intercept 0, slope 1).

```
abline(coef(fm))
```
Unweighted regression line.

```
abline(coef(fm1),lty=4)
```
Weighted regression line.

At any time you can make a hardcopy of the graphics window by clicking on the `Graph` section of the window and selecting the `Print` option.

```
detach()
```
Remove data frame from the search list.

```
plot(fitted(fm),
  resid(fm),
  xlab="Fitted values",
  ylab="Residuals", main=
  "Residuals vs Fitted")
```
A standard regression diagnostic plot to check for heteroscedasticity. Can you see it?

```
qqnorm(resid(fm), main=
  "Residuals Rankit Plot")
```
A normal scores plot to check for skewness, kurtosis and outliers. (Not very useful here.)

```
rm(fm,fm1,lrf,x,dummy)
```
Clean up again.

The next section will look at data from the classical experiment of Michaelson and Morley to measure the speed of light.

```
!more morley.data
```
Optional. Temporarily interrupt S-PLUS and look at the file. This is a standard way to escape to the operating system.

```
mm <- read.table(
  "morley.data")
mm
```
Read in the Michaelson and Morley data as a data frame, and look at it. There are five experiments (col. `Expt`) and each has 20 runs (col. `Run`) and `sl` is the recorded speed of light, suitably coded.

```
attach(mm, 1)
objects()
```
Place `mm` on the top of the search list, (position 1).

```
Expt <- factor(Expt)
Run <- factor(Run)
```
Change `Expt` and `Run` into factors.

```
detach(1, save="mm")
attach(mm)
```
Save the changes and make the data frame visible at position 2 (the default).

```
plot(Expt,Speed, main=
  "Speed of Light Data",
  xlab="Experiment No.")
```
Compare the five experiments with simple boxplots.

```
fm <- aov(Speed~Run+Expt
  data=mm)
summary(fm)
```
Analyse as a randomized block, with ‘runs’ and ‘experiments’ as factors.

```
fm0 <- update(fm,
  .~.-Run)
anova(fm0,fm)
```
Fit the sub-model omitting ‘runs’, and compare using a formal analysis of variance.

```
detach()
rm(fm, fm0)
```
Clean up before moving on.

We now look at some more graphical features: contour and 3−dimensional perspective plots.

```
x <- seq(-pi, pi, len=50)
y <- x
```
x is a vector of 50 equally spaced values in $-\pi \leq x \leq \pi$. y is the same.

```
f <- outer(x, y,
  function(x,y)
    cos(y)/(1+x^2))
```
f is a square matrix, with rows and columns indexed by x and y respectively, of values of the function $\cos(y)/(1+x^2)$.

```
oldpar <- par()
par(pty="s")
```
Save the plotting parameters and set the plotting region to "square".

```
contour(x, y, f)
contour(x, y, f,
  nint=15, add=T)
```
Make a contour map of f; add in more lines for more detail.

```
fa <- (f-t(f))/2
```
`fa` is the "asymmetric part" of f. (`t()` is transpose).

```
contour(x, y, fa, nint=15)
```
Make a contour,. . .

```
par(oldpar)
```
. . . and restore the old graphics parameters.

```
persp(x, y, f)
persp(x, y, fa)
image(x, y, f)
image(x, y, fa)
```
Make some pretty perspective and high density image plots, (of which you can get hardcopies if you wish)

```
objects(); rm(x,y,f,fa)
```
and clean up before moving on.

```
th <- seq(-pi, pi,
  len=100)
z <- exp(1i*th)
```
S-PLUS can do complex arithmetic, also. `1i` is used for the complex number i

```
par(pty="s")
plot(z, type="l")
```
Plotting complex arguments means plot imaginary versus real parts. This should be a circle.

```
w <- rnorm(100) +
  rnorm(100)*1i
```
Suppose we want to sample points within the unit circle. One method would be to take complex numbers with standard normal real and imaginary parts. . .

```
w <- ifelse(Mod(w) > 1,
  1/w, w)
```
and to map any outside the circle onto their reciprocal.

```
plot(w, xlim=c(-1,1),
  ylim=c(-1,1), pch="+",
  xlab="x", ylab="y")
lines(z)
```
All points are inside the unit circle, but the distribution is not uniform.

```
w <- sqrt(runif(100))*
  exp(2*pi*runif(100)*1i)
plot(w, xlim=c(-1,1),
  ylim=c(-1,1), pch="+",
  xlab="x", ylab="y")
lines(z)
```
The second method uses the uniform distribution. The points should now look more evenly spaced over the disc.

```
rm(th,w,z)
```
Clean up again.

```
par(oldpar)
```
Restore standard graphics parameters.

```
butterfly()
```
An old favourite. Take a hardcopy if you wish.

```
rm(oldpar)
```
Clean up again.

```
q()
```
Quit the S-PLUS program. . .

```
>
```
. . . and return to UNIX.

B The Inbuilt Command Line Editor in S-PLUS

B.1 Preliminaries

The August 1991 release of S-PLUS has inbuilt command line editor that allows recall, editing and re-submission of prior commands.

To use it, start the S-PLUS program with

```
$ Splus -e
```

Inside the editor either emacs or vi conventions are available, according to the shell environment variable `S_CLEDITOR`. To get the emacs conventions use (in `csh` and variants)

```
$ setenv S_CLEDITOR emacs
```

and for the vi conventions to apply, put vi instead of emacs. This statement would normally be included in your `.login` file (or equivalent) and would then be done automatically at login time. To avoid forgetting to include the `-e` a handy alias for your `.cshrc` file is, say

```
alias S+ 'Splus -e'
```

after which `S+` is the command to start S-PLUS with command line editor.

The usual typographical conventions apply: `^M` means ***Hold the*** `Control` ***down while you press the*** `m` ***key***, but `Esc m` means ***First press the*** `Esc` ***key and then the*** `m` ***key***. Note that case is significant after `Esc`.

B.2 Editing Actions

The S-PLUS program keeps a history of the commands you type, including the error lines, and commands in your history may be recalled, changed if necessary, and re-submitted as new commands. In *emacs* style command line editing any straight typing you do while in this editing phase causes the characters to be inserted in the command you are editing, displacing any characters to the right of the cursor. In *vi* mode character insertion mode is started by `Esc i` or `Esc a`, characters are typed and insertion mode is finished by typing a further `Esc` .

Pressing the `Return` command at any time causes the command to be re-submitted.

Other editing actions are summarised in the following table.

Unfortunately it does not seem to be possible to bind the motion keys, for example, to the arrow keys, which is something of a nuisance.

B.3 Command Line Editor Summary

1. Command recall and vertical motion	emacs style	vi style
Go to the previous command (backwards in the history)	`^P`	`Esc k`
Go to the next command (forwards in the history)	`^N`	`Esc j`
Find the last command with the `text` string in it	`^R text`	`Esc ? text`

2. Horizontal motion of the cursor		
Go to the beginning of the command	`^A`	`Esc ^`
Go to the end of the line	`^E`	`Esc $`
Go back one word	`Esc b`	`Esc b`
Go forward one word	`Esc f`	`Esc w`
Go back one character	`^B`	`Esc h`
Go forward one character	`^F`	`Esc l`

3. Editing and re-submission		
Insert `text` at the cursor	text	`Esc i` text `Esc`
Append `text` after cursor	`^F`text	`Esc a` text `Esc`
Delete the previous character (left of the cursor)	`Delete`	`Esc shift-x`
Delete the character under the cursor	`^D`	`Esc x`
Delete rest of the word under the cursor, and 'save' it	`Esc d`	`Esc dw`
Delete from cursor to end of command, and 'save' it	`^K`	`Esc shift-d`
Insert (yank) the last 'saved' text here	`^Y`	`Esc shift-y`
Transpose the character under the cursor with the next	`^T`	`Esc xp`
Change the rest of the word to lower case	`Esc l`	
Change the rest of the word to capitals (upper case)	`Esc c`	
Re-submit the command to S-PLUS	`Return`	`Return`

NOTE: With `vi` style commands the `Esc` need only be issued before the first recall command, and to terminate insert and append commands, as is usual in `vi`.

The final `Return` terminates the command line editing sequence for commands of either style.

C Exercises

C.1 The cloud point data

Source: Draper & Smith, *Applied Regression Analysis*, p. 162
Category: Polynomial regression. Simple plots.

Description

The cloud point of a liquid is a measure of the degree of crystalization in a stock that can be measured by the refractive index. It has been suggested that the percentage of I_8 in the base stock is an excellent predictor of cloud point using the second or third order model:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + E, \qquad E \sim \mathrm{N}(0, \sigma^2)$$

Data

The following data was collected on stocks with known percentage of I_8:

I_8%	Cloud Point	I_8%	Cloud Point	I_8%	Cloud Point	I_8%	Cloud Point
0	21.9	2	26.1	5	28.9	8	31.4
0	22.1	3	26.8	6	29.8	8	31.5
0	22.8	3	27.3	6	30.0	9	31.8
1	24.5	4	28.2	6	30.3	10	33.1
2	26.0	4	28.5	7	30.4		

The data may be read from file `cloud.data` in a form suitable to construct a data frame.

Suggested analysis

Fit polynomial regression models using the `lm()` function, and choose the degree carefully.

C.2 The Janka hardness data

Source: E. J. Williams: *Regression Analysis*, Wiley, 1959.
Category: Polynomial regression. Transformations.

Description

The Janka hardness is an important structural property of Australian timbers, which is difficult to measure. It is, however, related to the density of the timber, which is relatively easy to measure. A low degree polynomial regression is suggested as appropriate.

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + E$$

where Y is the hardness and x the density.

Data

The following data comes from samples of 36 Australian Eucalypt hardwoods.

D	H	D	H	D	H	D	H	D	H	D	H
24.7	484	30.3	587	39.4	1210	42.9	1270	53.4	1880	59.8	1940
24.8	427	32.7	704	39.9	989	45.8	1180	56.0	1980	66.0	3260
27.3	413	35.6	979	40.3	1160	46.9	1400	56.5	1820	67.4	2700
28.4	517	38.5	914	40.6	1010	48.2	1760	57.3	2020	68.8	2890
28.4	549	38.8	1070	40.7	1100	51.5	1710	57.6	1980	69.1	2740
29.0	648	39.3	1020	40.7	1130	51.5	2010	59.2	2310	69.1	3140

The data may be read as a data frame from file `janka.data`.

Suggested analysis

Fit polynomial regression models, choosing the degree carefully. Examine the residuals and see if the data has any obvious outliers or heteroscedasticity. Check to see what effect a square root or log transformation has on the residual pattern when plotted against fitted values.

More advanced: Consider a quasi-likelihood model with variance proportional to the mean and a square root link.

C.3 The Tuggeranong house price data

Source: Dr Ray Correll, Personal communication
Category: Multiple regression, coplots.

Description

Before buying a house in Tuggeranong in February, 1987, a cautious potential householder collected some data on houses on the market. The data for 20 such houses is shown in the table and is available as the file `house.dat`. The variables collected are, in order, price, total floor area, block area, number of main rooms, age of house and whether or not the house was centrally heated.

Data

The data is given in Table 5 and is available as the file `house.data`.

Suggested analysis

Explore the data with `coplot()` using `Age` and `CentHeat` as conditioning variables.

Choose a multiple regression model carefully and check for outliers, (that is, for “bargains” and “rip-offs”).

Price ($000s)	Floor (m^2)	Block (m^2)	Rooms	Age (years)	Cent. Heat.
52.00	111.0	830	5	6.2	no
54.75	128.0	710	5	7.5	no
57.50	101.0	1000	5	4.2	no
57.50	131.0	690	6	8.8	no
59.75	93.0	900	5	1.9	yes
62.50	112.0	640	6	5.2	no
64.75	137.6	700	6	6.6	yes
67.25	148.5	740	6	2.3	no
67.50	113.5	660	6	6.1	no
69.75	152.0	645	7	9.2	no
70.00	121.5	730	5	4.3	yes
75.50	141.0	730	7	4.3	no
77.50	124.0	670	6	1.0	yes
77.50	153.5	795	7	7.0	yes
81.25	149.0	900	6	3.6	yes
82.50	135.0	810	6	1.7	yes
86.25	162.0	930	6	1.2	yes
87.50	145.0	825	6	0.0	yes
88.00	172.0	950	7	2.3	yes
92.00	170.5	870	7	0.7	yes

Table 5: The Tuggeranong house price data

C.4 Yorke Penninsula wheat yield data

Source: K. W. Morris (private communication)
Category: Multiple regression.

Description

The annual yield of wheat in a marginal wheat growing district on the Yorke Penninsula, South Australia, together with the rainfall for the three growing months, for the years 1931–1955. The year itself is potentially a surrogate predictor to allow for improvements in varieties and farm practice. Yield is in bushels per acre, and rainfall is in inches.

Data

The data is given in Table 6 and may be read as a data frame from file `sawheat.data`.

Year	Rain0	Rain1	Rain2	Yield	Year	Rain0	Rain1	Rain2	Yield
1931	.05	1.61	3.52	.31	1944	3.30	4.19	2.11	4.60
1932	1.15	.60	3.46	.00	1945	.44	3.41	1.55	.35
1933	2.22	4.94	3.06	5.47	1946	.50	3.26	1.20	.00
1934	1.19	11.26	4.91	16.73	1947	.18	1.52	1.80	.00
1935	1.40	10.95	4.23	10.54	1948	.80	3.25	3.55	2.98
1936	2.96	4.96	.11	5.89	1949	7.08	5.93	.93	11.89
1937	2.68	.67	2.17	.03	1950	2.54	4.71	2.51	6.56
1938	3.66	8.49	11.95	16.03	1951	1.08	3.37	4.02	1.30
1939	5.15	3.60	2.18	6.57	1952	.22	3.24	4.93	.03
1940	6.44	2.69	1.37	8.43	1953	.55	1.78	1.97	.00
1941	2.01	6.88	.92	8.68	1954	1.65	3.22	1.65	3.09
1942	.73	3.30	3.97	2.49	1955	.72	3.42	3.31	2.72
1943	2.52	1.93	1.16	.98					

Table 6: Yorke Penninsula wheat yield data

Suggested analysis

Fit a multiple regression model and check the results, via residual plots especially.

C.5 The Iowa wheat yield data

Source: CAED Report, 1964. Quoted in Draper & Smith.
Category: Multiple regression; diagnostics.

Description

The data gives the pre-season and three growing months' precipitation, the mean temperatures for the three growing months and harvest month, the year, and the yield of wheat for the USA state of Iowa, for the years 1930–1962.

Data

The data is given in Table 7 and may be read as a data frame from file `iowheat.data`.

Suggested analysis

Fit a multiple regression model and select carefully the predictors. Work either by backward elimination of forward selection. Examine the residuals by plotting them in turn against each predictor variable.

Consider the effect of adding quadratic terms in the predictors.

It is interesting to compare this set of data with the Yorke Penninsula data for a similar period.

C.6 The gasoline yield data

Source: *Estimate gasoline yields from crudes*
by Nilon H. Prater, Petroleum Refiner, **35**, #5.
Category: Analysis of variance, covariance, and multiple regression.
Modern regression.

Description

The data gives the gasoline yield as a percent of crude oil, say y, and four independent variables which may influence yield. These are

x_1**:** The crude oil gravity, in ^{0}API,

x_2**:** The crude oil vapour pressure,

x_3**:** The crude oil 10% point, ASTM,

x_4**:** The gasoline end point.

The data comes as 10 separate samples, and within each sample the values for x_1, x_2, and x_3 are constant.

Data

The data is shown in Table 8, and is available as the file `oil.data` in a form suitable for constructing a data frame.

Suggested analysis

Using `EndPt` as a covariate, check to see if differences between samples can be accounted for by regression models on the other predictors.

More advanced: Fit a two stratum ANOVA model using between and within samples as the two strata.

C.7 The Michaelson and Morley speed of light data

Source: Weekes: A Genstat Primer.
Category: Analysis of Variance.

Description

The classical data of Michaelson and Morley on the speed of light. The data consists of five experiments, each consisting of 20 consecutive "runs". The response is the speed of light measurement, suitably coded. The data is here viewed as a randomized block experiment with *experiment* and *run* as the factors. *run* may also be considered a quantitative variate to account for linear (or polynomial) changes in the measurement over the course of a single experiment.

Data

The data is given in Table 9 and may be read as a data frame from file `morley.data` in a form suitable for constructing a data frame.

Suggested analysis

Using an single classification ANOVA model check for differences between experiments and summarise your conclusions.

C.8 The rat genotype data

Source: Quoted in Scheffe, H.: *The Analysis of Variance*
Category: Unbalanced double classification.

Description

Data from a foster feeding experiment with rat mothers and litters of four different genotypes: A, F, I and J. The measurement is the litter weight gain after a trial feeding period.

Data

The data is given in Table 10 and may be read as a data frame from file `genotype.data`.

Suggested analysis

Fit a double classificaiton model. Check for interaction using both a formal analysis and graphically using `interaction.plot()`. Test the main effects and summarise.

C.9 Fisher's sugar beet data

Source: R. A. Fisher, ***Design of Experiments***.
Category: Analysis of variance and covariance.

Description

A classical 3×2^3 randomized block experiment in four blocks of size 24. The response is the total weight of sugarbeet roots off the plot, but this is accompanied by the number of roots measured. The suggestion is that number of roots should be a covariate to allow for varying plot size.

The factors are ***Variety***. (3 levels, a, b and c), and N, P and K each at 2 levels, present or absent.

Data

The data is given in Table 11 and may be read from the file `sugar.data` in a form suitable to construct a data frame.

Suggested analysis

Analyse the data as a randomised block experiment with `Wt` as the response and `No` as a covariate. Prune the model of all unnecessary interaction terms and summarise.

C.10 A barley split plot field trial

Source: Unknown. Traditional data.
Category: Multistratum analysis of variance.

Description

An experiment involving barley varieties and manure (nitrogen) was conducted in 6 blocks of 3 whole plots.

Each whole plot was divided into 4 subplots. Three varieties of barley were used in the experiment with one variety being sown in each whole plot, while the four levels of manure (0, 0.01, 0.02, and 0.04 tons per acre) were used, one level in each of the four subplots of each whole plot. In the above table V_i denotes the ith variety and N_j denotes the jth level of nitrogen.

Data

The data is given in Table 12 and may be read as a data frame from file `barley.data`.

Suggested analysis

Analyse as a split plot field experiment and summarise.

C.11 The snail mortality data

Source: Zoology Department, The University of Adelaide.
Category: Generalized Linear Modelling.

Description

Groups of 20 snails were held for periods of 1, 2, 3 or 4 weeks (exposure) in carefully controlled conditions of temperature (3 levels) and relative humidity (4 levels). There were two species of snail, A and B, and the experiment was designed as a $4 \times 3 \times 4 \times 2$ completely randomized design. At the end of the exposure time the snails were tested to see if they had survived; this process itself is fatal for the animals. The object of the exercise was to model the probability of survival in terms of the stimulus variables, and in particular to test for differences between species.

The data is unusual in that in most cases fatalities during the experiment were fairly small.

Data

The data is given in Table 13 and may be read as a data frame from file `snails.data`.

Suggested analysis

The data is intersting in that although it has many extremely small cell counts there is every indication that some of the likelihood ratio large sample theory is quite safe.

Fit Binomial models to the data with either a logit or a probit link. Show that a model with parallel regressions on `Temp`, `Humid`, `Exposure` and `Exposure`^{2} for each species (in the logit/probit scale) is reasonable, and summarise.

C.12 The Kalythos blindness data

Source: S. D. Silvey: ***Statistical Inference***. (Fictitious?)
Category: Generalized linear modelling

Description

On the Greek island of Kalythos the male inhabitants suffer from a congenital eye disease, the effects of which become more marked with increasing age. Samples of islander males of various ages were tested for blindness and the results recorded.

Data

The data is given in Table 14 and may be read as a data frame from file `kalythos.data`.

Suggested analysis

Using an logit or probit model estimate the LD50, that is, the age at which the probability of blindness is $p = \frac{1}{2}$, together the standard error. Check how different the logit and probit models are in this respect.

C.13 The Stormer viscometer calibration data

Source: E. J. Williams: ***Regression Analysis***, Wiley, 1959
Category: Nonlinear regression, special regression.

Description

The stormer viscometer measures the viscosity of a fluid by measuring the time taken for an inner cylinder in the mechanism to perform a fixed number of revolutions in response to an actuating weight. The viscometer is calibrated by measuring the time taken with varying weights while the mechanism is suspended in fluids of accurately known viscosity. The data comes from such a calibration, and theoretical considerations suggest a nonlinear relationship between time, weight and viscosity of the form

$$T_i = \frac{\beta v_i}{w_i - \theta} + E_i$$

where β and θ are unknown parameters to be estimated.

Data

The data is given in Table 15 and may be read as a data frame from file `stormer.data`.

Suggested analysis

Estimate the nonlinear regression model using the `nlr()` software. A suitable initial value may be obtained by writing the regression model in the form

$$w_i T_i = \beta v_i + \theta T_i + (w_i - \theta) E_i$$

and regressing $w_i T_i$ on v_i and T_i using ordinary linear regression.

C.14 The chlorine availability data

Source: Draper & Smith, *Applied Regression Analysis*, (adapted).
Category: Nonlinear regression

Description

The following set of industrial chemical data shows the amount of chlorine available in a certain product at various times of testing after manufacture. A nonlinear regression model for the chlorine decay of the form

$$Y = \beta_0 + \beta_1 \exp(-\theta x)$$

has been suggested on theoretical grounds, with Y the amount remaining at time x.

Data

The data is given in Table 16 and may be read as a data frame from file `chlorine.data`.

Suggested analysis

Fit the nonlinear regression model using the `nlr()` function. A simple way to find an initial value is to guess a value for θ, say $\theta = 1$ and plot Y against $\exp(-\theta x)$. Now repeatedly either double or halve θ until the plot is near to linear. (This can be done very simply with the inbuilt line editor.) Once an initial value for θ is available, initial values for the others can be got by linear regression.

Check the fitted model for suspicious data points.

C.15 The saturated steam pressure data

Source: Quoted in Draper & Smith: *Applied Regression Analysis*. . .
Category: Nonlinear regression.

Description

The data gives the temperature (^{0}C) and pressure (Pascals) in a saturated steam driven experimental device. The relationship between pressure, Y, and temperature, x, in saturated steam can be written as

$$Y = \alpha \exp\left\{\frac{\beta x}{\gamma + x}\right\} + E$$

However a more realistic model may have the experimental errors multiplicative rather than additive, in which case an analysis in the log scale using the model

$$\log Y = \log \alpha + \left\{\frac{\beta x}{\gamma + x}\right\} + E$$

may be more appropriate.

Data

The data is given in Table 17 and may be read as a data frame from file `steam.data`.

Suggested analysis

Fit both models and compare. Initial values may be got by a similar method to that employed for the Chlorine data, since again there is only one nonlinear parameter.

C.16 Count Rumford's friction data

Source: Bates & Watts: *Nonlinear Regression Analysis...*
Category: Nonlinear regression

Description

Data on the amount of heat generated by friction was obtained by Lord Rumford in 1798. A bore was fitted into a stationary cylinder and pressed against the bottom by a screw. The bore was turned by a team of horses for 30 minutes, after which Lord Rumford "suffered the thermometer to remain in its place nearly three quarters of an hour, observing and noting down, at small intervals of time, the temperature indicated by it".

Newton's law of cooling suggests a nonlinear regression model of the form

$$Y = \beta_0 + \beta_1 \exp(-\theta x)$$

where Y is the temperature and x is the time in minutes.

Data

The data is given in Table 18 and may be read as a data frame from file `rumford.data`.

Suggested analysis

This data is mainly of historical interest. Handle similarly to the Chlorine data above.

C.17 The jellyfish data

Source: *Interactive Statistics*, Ed. Don McNeil.
Category: Bivariate, two sample data.

Description

Two samples of jellyfish, from Danger Island and Salamander Bay respectively, were measured for `length` and `width`.

Data

The data is given in Table 19 and may be read as a data frame from file `jellyfish.data`.

Suggested analysis

Plot the two samples and mark in their convex hulls. Test for differences using Hotelling's T^2. (A simple way of conducting the analysis is to regress a dummy variable for *Location* on *Length* and *width* and to test the significance of both regression coefficients simultaneously.

C.18 The Archæological pottery data

Source: Tubb, A. *et al.* Archæometry, **22**, 153–171, (1980)
Category: Multivariate analysis

Description

The data arises from a chemical analysis of 26 samples of pottery found at Romano-British kiln sites in Wales, Gwent and the New Forest. The variables describe the composition, in terms of various metals, and are expressed as percentages of the oxides of the metals.

The metals are aluminium, iron, magnesium, calcium and sodium and the sites are

L: Llanederyn, C: Caldicot I: Island Thorns A: Ashley Rails

Data

The data is given in Table 20 and may be read as a data frame from file `pottery.data`.

Suggested analysis

Investigate both numerically and graphically using a simple discriminant analysis. Exhibit the four samples using the first two discriminant functions as coordinate axes. Summarise.

C.19 The Beaujolais quality data

Source: Quoted in Weekes: *A Genstat Primer*
Category: Multivariate analsysis

Description

Quality measurements for some identified samples of young Beaujolais. Extracted from Table 1 in M. G. Jackson, *et al*: Red wine quality: correlations between colour, aroma and flavour and pigment and other parameters of young Beaujolais, *Journal of Science of Food and Agriculture*, **29**, 715–727, (1978).

Data

The data is given in Table 21 and may be read as a data frame from file `beaujolais.data`.

Suggested analysis

Look at ways of exhibiting the data graphically. Consider a principal component analysis using the correlation matrix and look for any wild outliers.

C.20 The painters data of de Piles

Source: Weekes: A Genstat Primer.
Category: Multivariate Analysis: Discriminant Analysis.

Description

The data shows the subjective assessment, on a 0–20 integer scale, of 54 classical painters. The painters were assessed on four characteristics: composition, drawing, colour and expression. The data is due to the Eighteenth century art critic, de Piles.

The School to which a painter belongs is indicated by a letter code as follows:

A	Renaissance	E	Lombard
B	Mannerist	F	Sixteenth Century
C	Seicento	G	Seventeenth Century
D	Venetian	H	French

Data

The data is given in Table 22 and may be read as a data frame from file `painters.data`.

Suggested analysis

Using a multivariate analysis of variance check for differences between schools. Use the likelihood ratio test. Also find the canonical $F-$statistics and discriminant functions.

Plot the painters on the first two discriminant function axes and use the school symbol as a plotting character. Mark in the convex hulls of the schools. Using `identify()` find interactively some of the painers that appear to lie towards the extremes of the plot, or who deviate considerably from their school mean.

Year	Rain0	Temp1	Rain1	Temp2	Rain2	Temp3	Rain3	Temp4	Yield
1930	17.75	60.2	5.83	69.0	1.49	77.9	2.42	74.4	34.0
1931	14.76	57.5	3.83	75.0	2.72	77.2	3.30	72.6	32.9
1932	27.99	62.3	5.17	72.0	3.12	75.8	7.10	72.2	43.0
1933	16.76	60.5	1.64	77.8	3.45	76.4	3.01	70.5	40.0
1934	11.36	69.5	3.49	77.2	3.85	79.7	2.84	73.4	23.0
1935	22.71	55.0	7.00	65.9	3.35	79.4	2.42	73.6	38.4
1936	17.91	66.2	2.85	70.1	0.51	83.4	3.48	79.2	20.0
1937	23.31	61.8	3.80	69.0	2.63	75.9	3.99	77.8	44.6
1938	18.53	59.5	4.67	69.2	4.24	76.5	3.82	75.7	46.3
1939	18.56	66.4	5.32	71.4	3.15	76.2	4.72	70.7	52.2
1940	12.45	58.4	3.56	71.3	4.57	76.7	6.44	70.7	52.3
1941	16.05	66.0	6.20	70.0	2.24	75.1	1.94	75.1	51.0
1942	27.10	59.3	5.93	69.7	4.89	74.3	3.17	72.2	59.9
1943	19.05	57.5	6.16	71.6	4.56	75.4	5.07	74.0	54.7
1944	20.79	64.6	5.88	71.7	3.73	72.6	5.88	71.8	52.0
1945	21.88	55.1	4.70	64.1	2.96	72.1	3.43	72.5	43.5
1946	20.02	56.5	6.41	69.8	2.45	73.8	3.56	68.9	56.7
1947	23.17	55.6	10.39	66.3	1.72	72.8	1.49	80.6	30.5
1948	19.15	59.2	3.42	68.6	4.14	75.0	2.54	73.9	60.5
1949	18.28	63.5	5.51	72.4	3.47	76.2	2.34	73.0	46.1
1950	18.45	59.8	5.70	68.4	4.65	69.7	2.39	67.7	48.2
1951	22.00	62.2	6.11	65.2	4.45	72.1	6.21	70.5	43.1
1952	19.05	59.6	5.40	74.2	3.84	74.7	4.78	70.0	62.2
1953	15.67	60.0	5.31	73.2	3.28	74.6	2.33	73.2	52.9
1954	15.92	55.6	6.36	72.9	1.79	77.4	7.10	72.1	53.9
1955	16.75	63.6	3.07	67.2	3.29	79.8	1.79	77.2	48.4
1956	12.34	62.4	2.56	74.7	4.51	72.7	4.42	73.0	52.8
1957	15.82	59.0	4.84	68.9	3.54	77.9	3.76	72.9	62.1
1958	15.24	62.5	3.80	66.4	7.55	70.5	2.55	73.0	66.0
1959	21.72	62.8	4.11	71.5	2.29	72.3	4.92	76.3	64.2
1960	25.08	59.7	4.43	67.4	2.76	72.6	5.36	73.2	63.2
1961	17.79	57.4	3.36	69.4	5.51	72.6	3.04	72.4	75.4
1962	26.61	66.6	3.12	69.1	6.27	71.6	4.31	72.5	76.0

Table 7: The Iowa historical wheat yield data

Sample	x_1	x_2	x_3	x_4	y	Sample	x_1	x_2	x_3	x_4	y
1	31.8	0.2	316	365	8.5	6	40.0	6.1	217	212	7.4
1	31.8	0.2	316	379	14.7	6	40.0	6.1	217	272	18.2
1	31.8	0.2	316	428	18.0	6	40.0	6.1	217	340	30.4
2	32.2	2.4	284	351	14.0	7	40.3	4.8	231	307	14.4
2	32.2	2.4	284	424	23.2	7	40.3	4.8	231	367	26.8
						7	40.3	4.8	231	395	34.9
3	32.2	5.2	236	267	10.0	8	40.8	3.5	210	218	8.0
3	32.2	5.2	236	360	24.8	8	40.8	3.5	210	273	13.1
3	32.2	5.2	236	402	31.7	8	40.8	3.5	210	347	26.6
4	38.1	1.2	274	285	5.0	9	41.3	1.8	267	235	2.8
4	38.1	1.2	274	365	17.6	9	41.3	1.8	267	275	6.4
4	38.1	1.2	274	444	32.1	9	41.3	1.8	267	358	16.1
						9	41.3	1.8	267	416	27.8
5	38.4	6.1	220	235	6.9	10	50.8	8.6	190	205	12.2
5	38.4	6.1	220	300	15.2	10	50.8	8.6	190	275	22.3
5	38.4	6.1	220	365	26.0	10	50.8	8.6	190	345	34.7
5	38.4	6.1	220	410	33.6	10	50.8	8.6	190	407	45.7

Table 8: The gasoline recovery data

Runs 1–10					Runs 11–20				
$\mathcal{E}_1$	$\mathcal{E}_2$	$\mathcal{E}_3$	$\mathcal{E}_4$	$\mathcal{E}_5$	$\mathcal{E}_1$	$\mathcal{E}_2$	$\mathcal{E}_3$	$\mathcal{E}_4$	$\mathcal{E}_5$
850	960	880	890	890	1000	830	880	910	870
740	940	880	810	840	980	790	910	920	870
900	960	880	810	780	930	810	850	890	810
1070	940	860	820	810	650	880	870	860	740
930	880	720	800	760	760	880	840	880	810
850	800	720	770	810	810	830	840	720	940
950	850	620	760	790	1000	800	850	840	950
980	880	860	740	810	1000	790	840	850	800
980	900	970	750	820	960	760	840	850	810
880	840	950	760	850	960	800	840	780	870

Table 9: The Michaelson and Morley speed of light data

Litter's Genotype	Mother's Genotype A	F	I	J
A	61.5	55.0	52.5	42.0
	68.2	42.0	61.8	54.0
	64.0	60.2	49.5	61.0
	65.0		52.7	48.2
	59.7			39.6
F	60.3	50.8	56.5	51.3
	51.7	64.7	59.0	40.5
	49.3	61.7	47.2	
	48.0	64.0	53.0	
		62.0		
I	37.0	56.3	39.7	50.0
	36.3	69.8	46.0	43.8
	68.0	67.0	61.3	54.5
				55.3
				55.7
J	59.0	59.5	45.2	44.8
	57.4	52.8	57.0	51.5
	54.0	56.0	61.4	53.0
	47.0			42.0
				54.0

Table 10: The rat genotype data

				Block 1		Block 2		Block 3		Block 4	
V	N	P	K	No	Wt	No	Wt	No	Wt	No	Wt
a	–	–	–	124	162	133	162	114	127	127	158
a	–	–	k	131	152	161	164	130	141	145	188
a	–	p	–	115	173	134	175	134	142	109	162
a	–	p	k	126	140	133	158	106	148	132	160
a	n	–	–	136	184	134	178	127	168	139	199
a	n	–	k	134	112	156	193	101	171	138	191
a	n	p	–	132	190	104	166	119	157	132	193
a	n	p	k	120	175	147	155	107	139	148	192
b	–	–	–	145	133	147	130	139	138	127	128
b	–	–	k	156	117	152	137	107	121	147	147
b	–	p	–	152	140	138	101	125	124	120	143
b	–	p	k	137	127	145	132	125	132	143	139
b	n	–	–	124	163	138	159	140	166	159	174
b	n	–	k	136	143	142	144	133	142	148	159
b	n	p	–	140	168	142	150	133	118	138	157
b	n	p	k	146	144	135	160	138	155	140	153
c	–	–	–	113	122	138	132	119	123	127	146
c	–	–	k	91	107	149	171	118	142	129	151
c	–	p	–	123	118	139	142	127	120	124	138
c	–	p	k	129	140	126	115	129	130	142	152
c	n	–	–	121	118	141	152	127	149	127	165
c	n	–	k	126	148	128	152	107	147	110	136
c	n	p	–	103	112	144	175	102	152	143	173
c	n	p	k	120	162	125	160	129	173	137	185

Table 11: Fisher's sugar beet data

Block	Variety	N_1	N_2	N_3	N_4	Block	Variety	N_1	N_2	N_3	N_4
	V_1	111	130	157	174		V_1	74	89	81	122
I	V_2	117	114	161	141	IV	V_2	64	103	132	133
	V_3	105	140	118	156		V_3	70	89	104	117
	V_1	61	91	97	100		V_1	62	90	100	116
II	V_2	70	108	126	149	V	V_2	80	82	94	126
	V_3	96	124	121	144		V_3	63	70	109	99
	V_1	68	64	112	86		V_1	53	74	118	113
III	V_2	60	102	89	96	VI	V_2	89	82	86	104
	V_3	89	129	132	124		V_3	97	99	119	121

Table 12: A split plot barley field trial

		Relative Humidity 60.0%			65.8%			70.5%			75.8%		
		Temperature			Temperature			Temperature			Temperature		
Species	Exposure	10	15	20	10	15	20	10	15	20	10	15	20
A	1	0	0	0	0	0	0	0	0	0	0	0	0
	2	0	1	1	0	1	0	0	0	0	0	0	0
	3	1	4	5	0	2	4	0	2	3	0	1	2
	4	7	7	7	4	4	7	3	3	5	2	3	3
B	1	0	0	0	0	0	0	0	0	0	0	0	0
	2	0	3	2	0	2	1	0	0	1	1	0	1
	3	7	11	11	4	5	9	2	4	6	2	3	5
	4	12	14	16	10	12	12	5	7	9	4	5	7

Table 13: The snail mortality data

Age:	20	35	45	55	70
No. tested:	50	50	50	50	50
No. blind:	6	17	26	37	44

Table 14: The Kalythos blindness data

	Weight		
Viscosity	20	50	100
14.7	35.6	17.6	
27.5	54.3	24.3	
42.0	75.6	31.4	
75.7	121.2	47.2	24.6
89.7	150.8	58.3	30.0
146.6	229.0	85.6	41.7
158.3	270.0	101.1	50.3
161.1		92.2	45.1
298.3		187.2	89.0
			86.5

Table 15: The Stormer viscometer calibration data

Weeks	Percent available	Weeks	Percent available	Weeks	Percent available
8	49, 49	20	42, 43, 42	32	40, 41
10	47, 47, 48, 48	22	40, 41, 41	34	40
12	43, 45, 46, 46	24	40, 40, 42	36	38, 41
14	43, 43, 45	26	40, 41, 41	38	40, 40
16	43, 43, 44	28	40, 41	40	39
18	45, 46	30	38, 40, 40	42	39

Table 16: The chlorine availability data.

Temp	Press	Temp	Press	Temp	Press
0	4.14	50	98.76	90	522.78
10	8.52	60	151.13	95	674.32
20	16.31	70	224.74	100	782.04
30	32.18	80	341.35	105	920.01
40	64.62	85	423.36		

Table 17: Temperature and pressure in saturated steam

Time (min.)	Temp (^{0}F)	Time (min.)	Temp (^{0}F)
4.0	126	24.0	115
5.0	125	28.0	114
7.0	123	31.0	113
12.0	120	34.0	112
14.0	119	37.5	111
16.0	118	41.0	110
20.0	116		

Table 18: The Rumford friction cooling data

Danger Island				Salamander Bay			
Width	Length	Width	Length	Width	Length	Width	Length
6.0	9.0	11.0	13.0	12.0	14.0	16.0	20.0
6.5	8.0	11.0	14.0	13.0	17.0	16.0	20.0
6.5	9.0	11.0	14.0	14.0	16.5	16.0	21.0
7.0	9.0	12.0	13.0	14.0	19.0	16.5	19.0
7.0	10.0	13.0	14.0	15.0	16.0	17.0	20.0
7.0	11.0	14.0	16.0	15.0	17.0	18.0	19.0
8.0	9.5	15.0	16.0	15.0	18.0	18.0	19.0
8.0	10.0	15.0	16.0	15.0	18.0	18.0	20.0
8.0	10.0	15.0	19.0	15.0	19.0	19.0	20.0
8.0	11.0	16.0	16.0	15.0	21.0	19.0	22.0
9.0	11.0			16.0	18.0	20.0	22.0
10.0	13.0			16.0	19.0	21.0	21.0

Table 19: The jellyfish data – Danger Island and Salamander Bay

Site	Al	Fe	Mg	Ca	Na	Site	Al	Fe	Mg	Ca	Na
L	14.4	7.00	4.30	0.15	0.51	C	11.8	5.44	3.94	0.30	0.04
L	13.8	7.08	3.43	0.12	0.17	C	11.6	5.39	3.77	0.29	0.06
L	14.6	7.09	3.88	0.13	0.20	I	18.3	1.28	0.67	0.03	0.03
L	11.5	6.37	5.64	0.16	0.14	I	15.8	2.39	0.63	0.01	0.04
L	13.8	7.06	5.34	0.20	0.20	I	18.0	1.50	0.67	0.01	0.06
L	10.9	6.26	3.47	0.17	0.22	I	18.0	1.88	0.68	0.01	0.04
L	10.1	4.26	4.26	0.20	0.18	I	20.8	1.51	0.72	0.07	0.10
L	11.6	5.78	5.91	0.18	0.16	A	17.7	1.12	0.56	0.06	0.06
L	11.1	5.49	4.52	0.29	0.30	A	18.3	1.14	0.67	0.06	0.05
L	13.4	6.92	7.23	0.28	0.20	A	16.7	0.92	0.53	0.01	0.05
L	12.4	6.13	5.69	0.22	0.54	A	14.8	2.74	0.67	0.03	0.05
L	13.1	6.64	5.51	0.31	0.24	A	19.1	1.64	0.60	0.10	0.03
L	12.7	6.69	4.45	0.20	0.22						
L	12.5	6.44	3.94	0.22	0.23						

Table 20: The pottery composition data

Label	OQ	AC	pH	TSO	Label	OQ	AC	pH	TSO
A	13.54	1.51	3.36	13.8	I	12.25	1.32	3.38	1.4
B	12.58	1.35	3.15	5.2	J	14.04	1.52	3.61	4.5
C	11.83	1.09	3.30	10.6	K	12.67	1.62	3.38	0.4
D	12.83	1.15	3.41	2.2	L	13.54	1.57	3.55	7.9
E	12.83	1.32	3.44	2.3	M	13.75	1.63	3.34	6.3
F	12.12	1.23	3.31	10.5	N	9.63	0.78	3.19	40.4
G	11.29	1.14	3.49	2.5	O	12.42	1.14	3.31	3.1
H	12.79	1.22	3.56	16.7					

Table 21: Quality measurements on young Beaujolais wine samples

	Composition	Drawing	Colour	Expression	School
Da Udine	10	8	16	3	A
Da Vinci	15	16	4	14	A
Del Piombo	8	13	16	7	A
Del Sarto	12	16	9	8	A
Fr. Penni	0	15	8	0	A
Guilio Romano	15	16	4	14	A
Michelangelo	8	17	4	8	A
Perino del Vaga	15	16	7	6	A
Perugino	4	12	10	4	A
Raphael	17	18	12	18	A
F. Zucarro	10	13	8	8	B
Fr. Salviata	13	15	8	8	B
Parmigiano	10	15	6	6	B
Primaticcio	15	14	7	10	B
T. Zucarro	13	14	10	9	B
Volterra	12	15	5	8	B
Barocci	14	15	6	10	C
Cortona	16	14	12	6	C
Josepin	10	10	6	2	C
L. Jordaens	13	12	9	6	C
Testa	11	15	0	6	C
Vanius	15	15	12	13	C
Bassano	6	8	17	0	D
Bellini	4	6	14	0	D
Giorgione	8	9	18	4	D
Murillo	6	8	15	4	D
Palma Giovane	12	9	14	6	D
Palma Vecchio	5	6	16	0	D
Pordenone	8	14	17	5	D
Tintoretto	15	14	16	4	D
Titian	12	15	18	6	D
Veronese	15	10	16	3	D
Albani	14	14	10	6	E
Caravaggio	6	6	16	0	E
Corregio	13	13	15	12	E
Domenichino	15	17	9	17	E
Guercino	18	10	10	4	E
Lanfranco	14	13	10	5	E
The Carraci	15	17	13	13	E
Durer	8	10	10	8	F
Holbein	9	10	16	13	F
Pourbus	4	15	6	6	F
Van Leyden	8	6	6	4	F
Diepenbeck	11	10	14	6	G
J. Jordaens	10	8	16	6	G
Otho Venius	13	14	10	10	G
Rembrandt	15	6	17	12	G
Rubens	18	13	17	17	G
Teniers	15	12	13	6	G
Van Dyck	15	10	17	13	G
Bourdon	10	8	8	4	H
Le Brun	16	16	8	16	H
Le Suer	15	15	4	15	H
Poussin	15	17	6	15	H

Table 22: The subjective assessment data of de Piles