Introduction to S-Plus

Written by: Derek Chung

Center for Social Science Computation & Research
145 Savery Hall
University of Washington
Seattle WA 98195 U.S.A.
(206)543-8110

December 98

http://*julius*.csscr.washington.edu/pdf/splus.pdf

INTRODUCTION TO S–PLUS

S-plus is a statistical package with an interactive environment for data analysis and graphics. S-plus is also a programming language for data analysis. Compared to SPSS and SAS, the biggest advantage of S-plus is its flexibility in doing complex computations.

Importing Data

S-plus treats data sets, scalars, vectors, matrices, functions, subroutines, even analysis outputs as objects. Virtually, you can do anything to the objects, such as exporting objects, importing objects, assigning objects, deleting objects, and getting a single attribute from an object for further analyses. In this section, we'll look at data entry in S-plus.

There are several ways that data can be imported to S-plus. You can:

- Enter a scalar, a vector, or a matrix
- Read a data file
- Import an existing S-plus data object.

Enter a scalar, a vector, or a matrix

Enter data using the **underscore character "_"** or "**<-**", with the object name on the left and the values on the right. The name must start with a letter and may contain letters, digits, and periods. S-plus is case-sensitive; x and X refer to two different objects.

Scalar:

```
> score_75
```

Assigns the value 75 to the scalar object named score.

Vector:

```
> score_c(23,13,14,15,18,43)
```

The function `c()` "collects" the values and stores them in the vector **score.**

```
> names_c("John," "Mike," "Susan," Bill," "Steve," "Jenny")
> score.1_matrix(score, ncol=2)
```

Matrices: The function **matrix** () reads data (by column) into a matrix.

```
> score.1
    [,1]  [,2]
[1,]  23    15
       [2,]13 18
[3,]  14    43
```

```
> score.2_matrix (score, ncol=2,
 byrow=T)
> score.2
    [,1]  [,2]
[1,]  23    13
[2,]  14    15
[3,]  18    43
```

```
> sex_c(1,1,0,1,1,0)
> x_cbind(score, names, sex)
```

cbind() (column bind)

binds together vectors and matrices columnwise into a new matrix.

```
> x
      score  names    sex
[1,]   "23"  "John"   "1"
[2,]   "13"  "Mike"   "1"
[3,]   "14"  "Susan"  "0"
[4,]   "15"  "Bill"   "1"
[5,]   "18"  "Steve"  "1"
[6,]   "43"  "Jenny"  "0
```

Read a Data File

Two S-Plus functions can be used to read data from an outside data file (ASCII file).

```
> mydata_scan("filename")
```

If the data is not space delimited, use the "**sep**" subcommand.

```
> mydata_scan("filename", sep=",")
```

Data is comma delimited.

```
> mydata_read. table("filename", sep="," header=T)
```

Data file contains a header (variable names) and the data is comma delimited. **Scan**() cannot be used when a data file contains a header.

Computations and Statistics

The interactive environment of S-plus lets you instantly do computations and descriptive statistics. Moreover, S-plus' object manipulations make your further computations easy. Table 1 shows some basic operators and functions for descriptive statistics.

Table 1: Arithmetic operators, logical operators, and descriptive statistics.

Arithmetic Operators		Logical Operators		Descriptive Statistics	
+	Addition	<	Less than	max()	Maximum
-	Subtraction	>	Greater than	min()	Minimum
*	Multiplication	==	Equal to	range()	Range
/	Division	!	Not	mean()	Average
%//%	Integer divide	<=	Less than or equal to	median()	Median
%%	Module function	>=	Greater than or equal to	var()	Variance
%*%	Matrix multiply	!=	Not equal to	cor()	Pearson's correlation

Some Real Analyses.

Table 2 contains data with boiling point (in Fahrenheit) and barometric pressure (inches of mercury) for eleven locations in the Alps and in Scotland. We are interested in using the measures of barometric pressure to predict the boiling points. A simple regression model is suitable for this prediction.

Table 2: Data of boiling point and barometric pressure.

Boiling point	Barometric pressure
194.5	20.79
210.7	29.04
201.3	24.02
212.2	30.06
197.9	22.40
199.4	23.35
209.5	28.49
201.4	24.02
194.3	20.79
200.9	23.89
198.4	22.67

```
> dataset_read.table ("mydata.txt", header=T)
> model_lm(boiling~pressure, data=dataset)
> summary(model)

Call: lm(formula=boiling~ pressure, data = dataset)

Residuals:
   Min       1Q      Median     3Q      Max
-0.4147    -0.23    0.06443   0.1553 0.4643
```

Coefficients:	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	154.6734	0.7327	211.0863	0.0000
pressure	1.9260	0.0297	64.9110	0.0000

```
Residual standard error:  0.3039 on 9 degrees of freedom
Multiple R-Squared:  0.9979
F-statistic:  4213 on 1 and 9 degrees of freedom, p-value is 2.467e-13

Correlation of Coefficients:
(Intercept)
 pressure  -0.9922
> #
> anova(model)
Analysis of Variance Table
Response: boiling

Terms added sequentially (first to last)
         Df    Sum of Sq   Mean Sq   F Value   Pr(F)
pressure  1     389.0744   389.0744  4213.442   2.466916e-13
Residuals 9     0.8311     0.0923
```

Graphics - The Fantastic Part

To look at graphics interactively, you need to open a graphics window. If you use the mainframe at UW, open your Mead account from MS-Windows in the computers with PG Xware installed, or use X terminals. Both ways of running X windows need to set up your DISPLAY environment under the UNIX shell. (For more information, see a CSSCR consultant.) There is also a version of S-plus for Windows available at CSSCR.

Some commonly used graphics commands:

plot (`xlyl...`)

barplot (`x,names=NULL,...`) Creates a bar graph.

hist(`x, nclass= , breaks= , probability=F....`) Creates a histogram.

pie (`x, names=NULL, explode=F,...`) Creates a pie chart.

density (`XI n=50,na.rm=F'`)Creates x , y coordinates of an estimate of the probability density of the data.

boxplot (`...`) Produces side by side box plots from a number of vectors.

pairs (`matrix ....`) Produces all pair-wise scatter plots.

Graphical parameters

Arguments for titling a plot:

`main="string"` Main title, above plot.

`sub="string"` Sub-title, below plot.

e.g., `main="Score Distribution for Female Students"`

Arguments for labeling axes:

xlab= "string", ylab="string" Labels for the x and y axes.

axes Logical flag specifying `axes=F` forces S-plus to omit drawing the axes.

xlim= , ylim=

Vectors giving the min and max values for the x and y axes. S-plus may extend the axes further to produce pretty thick labels.

type="type of plot"

"p" points, "l" lines, "b" both, "o" both but overstruck

"h" high density

"n" null (sets up the axes to be filled in later)

More than one graph in a page

`> par(mfrow=c(2,3))`

This function specifies the number and arguments of graphs on a page using the **mfrow=** option. `mfrow=c(2,3)` allows six plots to appear on a page (two rows of three plots each).

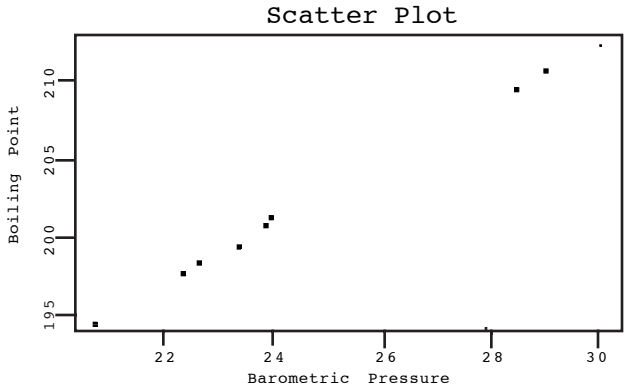
Save graphics into a postscript file

`> postscript (file- "your-file.ps")`

Note that you will use this function before issuing any graphics commands.

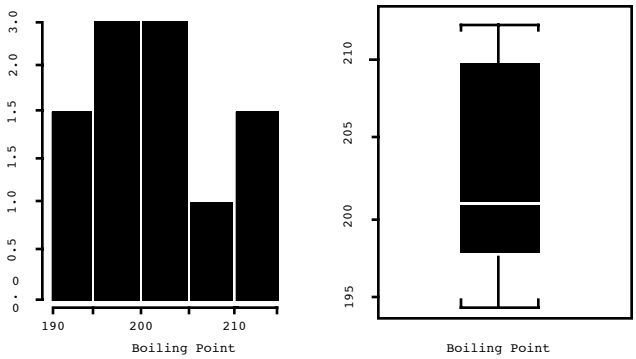
Examples

Let's go back to the previous boiling point example. People typically use plots to explore the data and diagnose the linear model. We show the plots in the following figures and associate the corresponding S-plus commands.



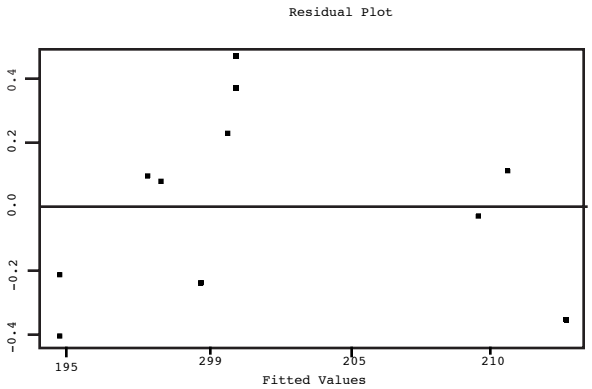
Scatter Plot (Figure 1):

```
> plot (dataset$pressure,
  dataset$boiling,
  xlab="Barometric Pressure,"
  ylab= "BoilingPoint"
  main="Scatter Plot")
```



Histogram and Boxplot (Figure 2):

```
> par(mfrow=c(1,2))
> hist(dataset$boiling,
  xlab="Boiling Point")
> boxplot (dataset$boiling,
  xlab="Boiling Point"
```



Residual Plot (Figure 3):

```
> plot(model$fitted, model$resid,
  xlab="Fitted Values",
  ylab= "Residuals",
  main= "Residual Plot")
> abline(h=0)
```

Functions – Programming Language

In S-plus you are able to create your own functions to solve problems of the same kind.

Functions are written in the form:

```
function name-function(arguments)expressions
```

where **arguments** is a list of parameters separated by commas which may be used by the function, and **expressions** are any legal S-plus expressions. The following is an example of a function to compute standard deviation in S-plus:

```
std_function(x)
{
sd_sqrt (var (x))
sd
}

> std(c(1,2,3,4,5))
> [1] 1.581139
```

Help Resources

S-plus has a very good on-line help system, where you can get detailed descriptions of any S-plus command. To activate on-line help, simply issue the ? command. MathSoft also published a series of S-plus manuals. We keep copies in CSSCR's main office (Room 145 Savery); please ask consultants for details.

dc,11/98