응용통계연구 제 6권 1호
1993년 3월 pp.1 - 11

S-PLUS 의 소개 및 SAS 와의 그래픽 비교

김 성 수[1], 한 경 수[1]

요 약

통계자료분석에서 그래픽에 의한 분석방법은 컴퓨터의 급속한 발전과 더불어 많은 기법이 개발되어 왔으며, 이는 자료에 내포된 정보 및 통계적인 모형선택방법에 유용하게 이용될 수 있다. 개인용 컴퓨터에서 그래픽 자료분석방법에 다양하게 응용될 수 있는 소프트웨어로서 S-PLUS(version 2.0)를 소개하고, 그래픽기능의 관점에서 SAS/GRAPH(version 6.04)와 비교, 논의하고자한다.

1. S-PLUS 의 소개

S-PLUS 는 일종의 프로그램언어로서 자료처리 및 그래픽 분야에 탁월한 기능을 가지고 있다. S-PLUS 의 기능을 크게 구분하면 다음의 네가지로 나눌 수 있다. 첫째, 자료처리 - 자료의 구성, 소팅, 결합 등, 둘째, 자료분석 - 자료를 분석하기 위해 필요한 수치적 계산 및 분석 결과 제공, 셋째, 언어의 기능 - 대화형 프로그래밍언어로서 함수문을 쉽게 작성할 수 있으며, C 언어 및 Fortran 언어와 interface 가 가능, 넷째, 그래픽 기능 - 대화형 그래픽에 의한 자료분석의 기능 및 분석결과의 그래픽처리 기능 등. 간단히 말하면 S-PLUS 는 그래픽처리에 의한 자료 분석 시스템으로서 object-oriented 된 프로그래밍 언어라고 말할 수 있다.

S-PLUS 를 사용하기 위한 요구사항을 살펴보면 PC 386이상으로서 coprocessor 설치, 2MB RAM 이상, hard disk 는 20 MB 정도, PC-DOS 3.1 이상, VGA, EGA, Hercules 와 같은 그래픽 장비가 지원되어야 한다.

1.1. 자료처리 방법

시작과 끝

S-PLUS 의 시작은 DOS prompt 에서 splus 를 치면 > 이 나타난다. > 는

[1] (560-756) 전북 전주시 덕진구 덕진동 전북대학교 통계학과

S-PLUS 의 언어를 읽어들일 준비가 되어 있다는 표시이다. S-PLUS 를 끝내려면 q() 를 치면된다.

자료의 입력 및 처리

자료의 입력은 scan 명령을 사용하며, 자체 내장함수문, 예를 들어 seq(-1, 1, by=.1), rnorm(10,-5,2.5) 등, 을 이용하여 자료를 구성할 수 있다.

S-PLUS 는 대화형으로 처리하지만 프로그램이 화일로 구성되어 있는 경우의 수행은 source 명령을 사용한다. 자료처리결과를 화면상에서 보지않고 화일로 저장하려면 sink 명령을 사용한다.

수치계산

S-PLUS 의 수치계산은 산술식, 행렬연산, 복소수연산, 미분, 연립방정식의 해, 다항식의 해, 확률, random number 발생 등의 다양한 계산을 수행하며, 계산은 double precision 으로 이루어진다.

FUNCTION 문

S-PLUS 는 자주 쓰이는 계산 또는 사용자가 작성한 프로그램을 function 문을 사용하여 이용할 수 있도록 해준다. 또한 function 문은 자체 function 문을 반복적으로 이용할 수 있다.

그래픽 출력

S-PLUS 그래픽은 다음과 같은 세가지 hard copy 장비 - Postscript laser printer, Hewlett-Packard Laserjet II printer, Hewlett-Packard HP-GL plotter - 를 지원한다. 그래프를 ASCII 화일로 받으면 일반 dot printer를 이용할 수 있다. 참고로 화면에 띄운 그래프를 직접 화일로 받으려면 dev.print 문을 이용한다.

1.2 통계분석 방법

S-PLUS가 제공하는 통계분석기법은 언어의 기능과 그래픽기법을 응용하여 이루어진 것으로 다음과 같은 통계분석방법을 제공하고 있다.

- 이산형 및 연속형자료에 대한 가설검정 및 추론
- 상관관계분석
- 분산분석
- 시계열분석
- 회귀분석
- 생존분석
- 비모수분석

이들 분석방법들은 통계이론에 근거한 분석결과를 제공해 주며, 분석 결과들을 다양한 그래프를 통하여 해석할 수 있도록 해준다.

2. S-PLUS 와 SAS 가 제공하는 통계그래프의 기능 및 비교

통계자료분석을 위해서는 사용자가 편리하게 사용할 수 있도록 통계그래프기능이 충분히 제공되어야 한다. 이는 기본적인 단변량분석에서부터 다변량자료분석에 이르기까지 다양한 내용을 포함하고 있어야한다. 본 장에서는 S-PLUS 와 SAS/GRAPH 가 제공하는 기본 통계그래프를 소개, 비교하고자한다.

2.1. S-PLUS 그래픽기능

S-PLUS 는 그래픽종류의 다양한 function 문을 제공한다. 사용자는 이를 이용하여 원시자료뿐만 아니라, 자료처리, 통계분석결과 후의 자료를 그래픽을 이용하여 쉽게 검정해볼 수 있다. S-PLUS 그래픽기능의 특징은 사용이 간편하고, 실시간처리가 가능하며 또한 다이내믹 그래픽 기능을 포함하고 있다고 말할 수 있다.

그래픽을 화면에 띄우기 위해서는 먼저 Device 를 지정해야 하며, 가능한 Device 종류는 VGA, EGA, Hercules 이다. VGA graphic device 를 띄우려면 "> vga()"를 치면 된다. vga() 를 띄운 뒤에 그래픽 명령문 - 예를 들어 plot(car.miles) - 을 실행하면 화면에 그림을 그린다. 다시 S-PLUS prompt 로 돌아가려면 임의의 key 를 치면 된다. graphicsmode() 는 이전에 그린 그림을 화면에 다시 보여주며, 그래픽 device를 없애려면 dev.off() 를 치면 된다.

2.2 SAS/GRAPH의 소개

SAS/GRAPH 는 자료를 여러가지 색을 이용한 plot, chart, map, slide 등의 형태로 모니터나 hardcopy 출력장치 (예: 도트프린터, 레이저 프린터, 플로터,필름레코더)에 그림을 그려줄 수 있는 컴퓨터 그래픽 시스템이다. 이를 사용하기 위한 요구사항을 살펴보면 PC 286 이상으로서 coprocessor 는 추천사항이고, 640KB RAM이상, hard disk는 BASE/STAT/GRAPH 를 포함하여 15MB 정도이면 된다. 그래프에 필요한 자료는 SAS 시스템의 다른 모듈 (예: SAS/BASE,SAS/STAT,SAS/IML...) 이나 SAS/GRAPH 에서 제공하는 데이타 step들로 이루어지며 프로그램은 다음과 같은 형태로 이루어진다.

```
GOPTIONS   ( 그래프를 그리는데 필요한 선택사양들 )
           device = ? ( 예: VGA, LQ800, PS, HP7475 )
           colors = ? ( 예: red,blue,yellow )
            .
            .
DATA       ( SAS data set 구성 )
           input = ? ( 기본 데이터나 그래픽 자료 입력 )
            .
PROC       ( SAS 가 제공하는 그래픽 프로그램들 )
           GPLOT, GCHART, GMAP, GCONTOUR, G3D . . .
            .
            .
```

2.3. S-PLUS 와 SAS 가 제공하는 통계그래프

표 1 은 이들 소프트웨어가 제공하는 기본적인 통계그래프이다.

2.4. S-PLUS 의 특이한 그래픽 기능

S-PLUS 가 제공하는 기본 통계그래프 중에서 histogram, pie chart, boxplot 등과 같은 단순한 기능은 생략하고, S-PLUS 의 특징적인 그래픽 기능에 대하여 소개하면 다음과 같다.

Mouse 를 이용한 자료의 표시

Plot 에 원하는 글자를 붙이거나, 하나의 plot 에 두개 이상의 문자나 line 으로 자료를 표시한 경우에 각각을 구분하는 설명을 더하고자 할 때는 text문이나, legend 문을 이용한다. Mouse 를 이용하여 관심있는 관찰값의 번호를 표시할 때는 identify 문을 이용하며, 원하는 위치에 문자 라벨을 붙일 때는 locator 문을 이용한다. 또한 문자라벨과 자료를 연결시키고자 할 때는 다음과 같은 표현

```
> locator(n=2, type="l")
```

에서 mouse 를 이용하여 연결한다.

DENSITY PLOT

연속적인 자료에 대한 density 를 보기 위해 histogram 을 그리면 상대적인 빈도수를 보여준다. 확률밀도함수를 보기위해 density 문을 이용하면 연속적인 자료에 대한 smooth 한 curve 를 볼 수 있다.

다이내믹 그래픽 기능

다이내믹 그래픽 방법은 컴퓨터 그래픽 터미널에서 실시간(real time)으로 자료를분석하는 기법을 말한다. Brushing 은 다이내믹 그래픽 방법을 이용하여 다변량 자료를 분석하는 방법으로서 mouse 로 조정되는 사각형을 이용하여 scatterplot matrices에서 임의의 한 점을 밝게하면 다른 그림에서도 똑같은 점이 밝게 빛나는 것을 말한다.

표 1. S-PLUS 와 SAS 가 제공하는 통계그래프

기 능		S-PLUS	SAS/GRAPH
변수들의 그래픽표시	Histogram	0	0
	Pie Chart	0	0
	Bar Chart	0	0
	Box Plot	0	0
	Multiple Box Plot	0	0
	Dot Chart	0	
	Stem-and-leaf Plot	0	0
	Normal prob. Plot	0	0
	Quantile-quantile Plot	0	0
	Survival Curve	0	
	Time Series Plot	0	0
	Scatter Plot	0	0
	Density Plot (pdf plot)	0	
	Pairwise Scatter Plot	0	0
	Star Plot	0	0
	Chernoff's Face	0	
삼차원 자료의 입체적 표시	등고선표시	0	0
	입체모형표시 (회전가능포함)	0	0
	영상표시	0	0
	삼차원 scatter plot 표현		0
	삼차원 빈도수표현		0
다이내믹 그래프 기능	Brush (Mouse를 이용하여 다변량자료의 scatter plot에서 관심있는 관찰점들을 동시에 반짝거리게 함으로써 자료의구조 분석을 용이하게 하는 기능)	0	
	Spin (Mouse를 이용하여 삼차원 축을 중심으로 회전시켜가며 자료의 구조를 파악하는기능)	0	

Spin 은 다차원 자료를 삼차원으로 projection 시켜 회전시키므로써 자료의 구조를 파악하는 기법을 말한다. Brushing 과 Spin 은 직접적인 분석방법으로서 그래픽터미널에서 작동하므로 이를 사용하기 위해서는 VGA, EGA, Hercules 를 이용해야 한다. Brush 문을 사용하면 다차원 자료의 scatterplot matrix 와 spin 기능을 수행할 수 있는 그림이 화면에 나타나며 mouse를 이용하여 원하는 기능을 수행할 수 있다. Spin 기능은 brush 문에서도 수행할 수 있지만 화면 전체에서 수행할 수 있도록 하여준다.

3. 그래픽의 응용

3.1 그래픽의 다양한 표현

통계자료분석에서 다양한 분석의 제공을 위해서는 그래픽의 모양을 자유롭게 표현할 수 있어야할 것이다. 그래픽의 응용관점에서 S-PLUS와 SAS/GRAPH 를 다음과 같은 몇가지 관점에서 살펴보고자 한다.

(1) 그래픽의 기본요소인 선그리기, 면그리기, 글씨쓰기, 색칠하기 등이 자유롭고 다양한가.
(2) 자료를 나타내기 위한 표현의 방법 - 예를 들어 선의 종류, 두께, 색깔, 임의의 문자 표현 등이 자유로운가.
(3) 그래프위에 글씨와 같은 정보를 삽입할 수 있는가, 또한 글씨의 크기를 자유롭게 조정할 수 있는가.
(4) 그래프의 축에 관한 정보를 임의로 줄 수 있는가, 또한 축의 모양을 자유롭게 바꿀 수 있는가.
(5) 자료의 정보를 주는 그래프를 기존의 그래프에 덧붙일 수 있는가.
(6) 범위가 다른 자료를 한 그림에 표시할 수 있는가.
(7) 한 화면 위에 여러개의 그림을 동시에 그릴 수 있는가, 크기의 조정은 가능한가.

위에 열거한 여러가지 관점중에서 1,2,3의 관점에서 보면, SAS/GRAPH가 더 다양하고 정교한 그림을 제공하여 주며, 5의 관점에서는 S-PLUS가 사용하기 간편하고, 4,6,7의 관점에서는 둘다 자유롭다고 할 수 있겠다.

3.2 통계자료분석의 그래픽응용

통계그래프에서 가장 중요한 문제는 그래픽을 이용한 자료의 분석기능이며, 이를 위한 다양하고 풍부한 기능이 제공되어야 한다. 여기에서는 자료분석에 대한 그래픽 기능에 대하여 간단한 예로 단순회귀모형에서 회귀진단 결과의 index plot 을 그리기 위한 응용 예를 보이고자 한다.

3.2.1. S-PLUS의 예

```
# read data
vga()                              # start a color graphic device
sample <- scan("reg.dat")          # reg.dat : see the following sas program
sample <- matrix(sample,ncol=3,byrow=T)
id <- sample[,1]
x <- sample[,2]
y <- sample[,3]
sample.reg <- lsfit(x,y,intercept=T)    # regression analysis
sample.diag <- ls.diag(sample.reg)     # regression diagnostics
#-------------------------
# scatterplot of (x,y)-data
#-------------------------
plot(x,y)
title(main="Scatterplot of X,Y")
abline(lsfit(x,y),lty=2) # fitted line of regression
b <- sample.reg$coef
b <- signif(b,4)          # signif : significant number is 4
if (b[2] < 0 ) ft <- paste("Fitted line of y = ",b[1],b[2],"x",sep="")
else ft <- paste("Fitted line of y = ",b[1],"+",b[2],"x",sep="")
text(locator(1),ft)
readline()                # press return-key
#--------------------------------
# calculate regression diagnostics
#--------------------------------
p <- sample.diag$hat
st.res <- sample.diag$std.res
std.res <- sample.diag$stud.res
cook <- sample.diag$cooks
dfit <- sample.diag$dfits
#--------------------------------
# calculate 95 % C.I. of yhat
#--------------------------------
res <- sample.reg$residuals
sse <- t(res) %*% res
n <- nrow(sample)
mse <- sse / (n - 2)
m.x <- mean(x)
sxx <- var(x) * (n-1)
ci  <- sqrt(mse*(1/n + (x - m.x)^2/sxx))
yhat <- y - res
zu <- qnorm(0.975)
yupper <- yhat + zu*ci
ylower <- yhat - zu*ci
#--------------------------------
# plotting reg. diagnostics and 95 % C.I.
#--------------------------------
par(mfrow=c(2,2))          # plotting 2 by 2 figures
plot(id,p,main="HAT MATRIX",xlab="number",ylab="Pii")
identify(id,p)
plot(id,std.res,main="STUDENTIZED RES",xlab="number",ylab="r*i")
identify(id,std.res)
```

```
plot(id,cook,main="COOK Distance",xlab="number",ylab="Cii")
identify(id,cook)
xorder <- order(x)
xo <- x[xorder]          # sorted version of x
yuo <- yupper[xorder]    # sorted version of yupper
ylo <- ylower[xorder]     # sorted version of ylower
yho <- yhat[xorder]      # sorted version of yhat
matplot(xo,cbind(yuo,yho,ylo),type="l",lty=c(2,1,2),col=c(6,1,6),
       main="95 % C.I. of YHAT",xlab="x_value",ylab="yhat")
```

3.2.2 SAS/GRAPH 의 예

```
/* regression diagnostic example */
data reg;
     input id x y @@;
cards;
1  15  95  2 26 71  3 10  83  4  9 91  5 15 102  6 20 87  7 18  93  8 11 100
9   8 104 10 20 94 11  7 113 12  9 96 13 10  83 14 11 84 15 11 102 16 10 100
17 12 105 18 42 57 19 17 121 20 11 86 21 10 100
run;
proc reg data=reg;
     model y=x / influence;
     output out=regout cookd=cook h=hat rstudent=stdres; run;
goptions reset=all dev=vga;
symbol1 cv=yellow v=diamond ci=green i=rlclm95;
title   c=gold h=3 f=swiss  'Fitting Line and Confidence Band';
proc gplot data=regout;
     plot y*x=1 ; run;
symbol c=yellow v=dot i=none;
title  c=gold h=3 f=swiss 'Hat Diagonal';
proc gplot data=regout;
     plot hat*id / haxis=1 to 21 by 1 ; run;
title c=gold h=3 f=swiss 'COOK Distance';
proc gplot data=regout;
     plot cook*id / haxis=1 to 21 by 1; run;
title c=gold h=3 f=swiss 'Studentized Residual';
proc gplot data=regout;
     plot stdres*id / haxis=1 to 21 by 1; run;
```

4. 결 론

이상에서 살펴본 바와 같이 S-PLUS와 SAS/GRAPH 의 가장 큰 특징은 사용자 스스로 통계그래프의 작성을 가능하게 해준다는 점일 것이다.S-PLUS 와 SAS/GRAPH 의 특징을 간단히 살펴보면 다음과 같이 말할 수 있다.

S-PLUS 의 특징

- 통계그래프를 작성시에 C, Fortran 언어와 interface가 가능하며, 일반 C 언어와 유사한 구조를 택하고 있으므로 프로그램이 용이하다.
- MOUSE를 이용한 자료의 표시가 가능하다.
- 다이내믹 그래픽기능이 들어있어 다변량자료에 대한 시각적인 분석을 가능하게 해 준다.
- 대화형 시스템의 구조를 택하고 있으므로 그래픽을 통한 자료분석이 용이하다.

SAS/GRAPH 의 특징

- 그래픽의 표현방법(색깔, 글씨모양 등)이 다양하고 모양이 정교하다.
- SAS 의 다른 모듈과 결합하여 사용할 수 있다.
- 출력장치의 지원이 다양하다. 예를 들어 Laser printer, Plotter, Film recorder 등.

참고 문헌

[1] Becker, R. A., Chambers, J. M., and Wilks, A. R. (1988), *The New S language*, Wadsworth and Brooks/Cole Advanced Books and Software, Pacific Grove, California.

[2] Cleveland, W. S., and McGill, R. (1988), *Dynamic Graphics for Statistics*, Wadsworth and Brooks/Cole Advanced Books and Software, Pacific Grove, California.

[3] SAS Institute Inc. (1988), *SAS/GRAPH User's Guide*, Ver 6.03,ed., U.S.A.

[4] Statistical Sciences, Inc. *S-PLUS for DOS User's Manual* (1991), Seattle, Washington.

Introduction to S-PLUS and Graphical Comparison with SAS

Sung-Soo Kim[1], Kyung-Soo Han[1]

Abstract

Statistical graphics have been important new tools for data analysis and many statistical softwares are exploiting graphical methods. Among these softwares available in personal computer at low cost, we introduce S-PLUS(version 2.0). S-PLUS is an interactive graphical data analysis system and object-oriented programming language. SAS/GRAPH is another popular graphical system for displaying data in the form of color plots, charts, maps, and slides on screen and hardcopy devices. S-PLUS is compared to SAS/GRAPH(version 6.04) in viewpoints of statistical graphics.

[1] Department of Statistics, Chonbuk National University, Jeonju, 560-756