경영과학의 응용
제2권, 1985년 8월

SAS, SPSS^X, BMDP를 중심으로 한 대형통계 패케지

한국과학기술원
응용수학과(理博) 金 炳 千

1. 서 론

컴퓨터의 발전은 통계학 발전에 크게 공헌을 하여 왔다. 1960대 초부터 컴퓨터를 이용한 통계 프로그램들이 개발되고 이에따른 Algorithm들이 발전됨에 따라 통계학 이론에서 해결할 수 없는 문제들이 컴퓨터를 이용하여 하나 둘씩 풀어져 나갔다. 또한 최근 1980년부터 퍼스날 컴퓨터가 대중에 대량 보급됨에 따라, 연구소의 컴퓨터실에서 해결해야만 했던 데이터 분석들이 실험실과 가정으로 옮겨지기 시작하였다.

통계 페케지란 과연 무엇인가? 60년 초 부터 데이타를 분석하는데 필요했던 프로그램들을 모아서 하나의 프로그램으로서 손쉽게 통계적 데이터를 분석하는 커다란 프로그램을 말한다. 페케지는 사용자가 손쉽게 쓸 수 있도록 되어 있어야 하며 프로그램 작성에서나, 데이터 처리, 화일의 관리, 보고서 작성들을 쉽게 처리할 수 있어야 한다.

그러므로 통계 페케지들은 고도의 통계학 지식이나 컴퓨터의 프로그램 지식없이 자기나름대로의 고유한 형태로서 쉽게 데이타 분석의 결과를 얻어 낼 수가 있어야 한다. 이러한 면에서 보면 페케지의 언어는 고급 언어인 High level 언어이다. 그러면 과연 페케지의 평가는 어떻게 할 것인가? 페케지의 평가는 사용자나 사용목적에 따라 다르겠지만 대체로 사용하는 편의성, 계산의 정확성, 문서화 정도, 다양한 자료 구조, 출력 형태, 그래픽(Graphic), 컴퓨터 기종에 대한 관리성 및 관리의 용이성을 들 수 있다.

통계 페케지들은 어떠한 것들이 있을까? 1983년에 통계 및 데이타 분석을 위한 페케지들은 119개가 된다고 Neffendor et al(1984)등이 발표 하였다. 이 숫자는 범용컴퓨터에 쓰이는 통계 페케지들 뿐만아니라, 퍼스날 컴퓨터용 페케지들도 포함되어 있다. 현재 많이 쓰이고 있는 대표적인 통계 페케지들은 BMDP, GLIM, MINITAB, IMSL Library, P-STAT, SAS, SPSS 등을 들 수가 있다. 그러나 이 통계 페케지들 중 가장 많이 사용되는 페케지를 든다면 다음 세가지를 대표적으로 들 수 있다.

SAS(Statistical Analysis System), SPSS^X (Statistical Package for the Social Sciences) BMDP(Biomedical Computer Programs). 이 세 페케지 중 SAS와 SPSS^{X}는 정보의 저장 및 정정, 데이터의 수정변환, 프로그래밍. 통계적 분석 화일처리 능력, Graph 등 각종 기능을 골고루 갖춘 매우 우수한 통계 페케지이고 BMDP는 SAS나 SPSS가 갖추지 못한 특수한 통계 프로그램들을 갖추고 있기 때문에 많은 통계 학자들과 사회학자들, 생물학자들이 특별한 관심사들을 갖고 있는 통계 페케지이다. 그러면 이 세가지 페케지들에 관한 특성을 찾아보기로 하자.

2. SAS (Statistical Analysis System)

1966년부터 University of North Carolina에서 개발되어, 1985년 초에는 Version 5를 내놓게 되었고 또 퍼스날 컴퓨터용(IBM-XT 와

IBM-AT) SAS를 발표하였다. 현재 국내에서는 한국과학기술원, 한국개발연구원, 경제기획원 조사통계국, 농수산부, 서울대학교, 고려대학교, 계명대학교, 대한항공, 금성사, 호남정유, 제일모직, 삼성전자, 현대 건설 등 13곳에서 사용하고 있다. SAS 페케지는 통계적 자료처리 분석을 밑받침 해주는 정보의 보관 및 편성, 화일의 관리, 자료의 변환, 보고서 작성,그래프 기능들이 잘 되어 있는 강력한 통계 자료분석용 페케지이다.

처음 SAS가 나왔을때는 IBM 기종에 맞추어 PL/I 언어를 사용하여 IBM 및 IBM유사 기종에 국한되었지만 최근 80년대부터는 VAX(VMS system에서만 가능) 기종, MV기종, PRIMOS 기종 등에도 확장 보급되기 시작하였다.

SAS는 1975년부터 SAS 사용자 그룹을 만들어 매년 SUGI(SAS User Group International) 회합을 갖고 많은 통계학자와 사용자들로부터 새로운 조언을 받아 통계 페케지로서는 제일 빠른 시일에 높은 수준에 도달하게 되었다. SAS 페케지를 몇가지로 나누어 보면 다음과 같이 나누어 볼 수 있다.

a. SAS/Basics
b. SAS/Statistics
c. SAS/Graph
d. SAS/ETS
e. SAS/OR

SAS/Basics는 SAS 페케지를 이용하는데 기본적으로 알아야 하는 부분이다. SAS/Basics에는 자료 저장과 편성, 자료 변환, 프로그래밍, 보고서 작성, 통계 분석, 화일 작성 등을 해 주고 있으며 Manual은 921 페이지에 달하는 방대한 양이다. 이 Basics은 나머지 부분인 Statistics, Graph, ETS, OR을 쓰는데 기본적으로 알아야 하는 부분이다. 여기에 쓰이는 SAS 언어는 SAS만이 갖고 있는 언어라고 말하기 보다는 FORTRAN, COBOL 보다도 더 높은 고급 언어라는 편이 낫다.

이러한 관계로 SAS를 다 이해 한다고 하는 것은 고급 프로그래머이고 통계학 전공을 하지 않은 이상 어렵다. 그러나 이중 10% 밖에 안되는 SAS Introductory Guide만을 읽고 배움으로서 처음 사용하려는 초보자들도 쉽게 데이타를 처리하여 통계 분석을 할 수가 있다. 이와같이 고급 언어인 SAS 언어를 배움으로서 복잡하게 구성이 된 화일과 자료를 System 구성의 지식없이도 손쉽게 다룰 수 있으며, 최근 미국에서는 프로그래머를 채용할 때 일부러 SAS 언어만 알고 있는 경험자를 채용하기도 한다.

Basic 안에는 서술적 통계 처리를 할 수 있는 Correlation, Frequeucy, Mean, Summry 등을 다루며 보고서 처리를 할 수 있는 Calendar, Chart, Plot, Print 등이 있다.

함수에는 Fortran 언어가 갖고 있는 함수 이외에 여러 분포의 random number, 여러 분포의 확률 값과 그 역 함수값 등 다양하게 다루고 있으며, 될수 있는 한 사용자에게 편리하게 사용하게 하였다. 예를 들면 변수 X1, …, X12의 합과 Y와의 최소값을 LEAST라 하면 이 SAS 프로그램은 LEAST=MIN(SUM(OF X1-X12), Y) ; 와 같이 간단하게 나타낼 수가 있다. 이밖에 SAS Utility 프로그램에는 Tape나 디스크 안에 있는 화일의 정보를 알수 있는 CONTENTS등 System 프로그램을 보다 쉽게 사용할 수 있도록 하였다. 이 중에 중요한 부분은 SAS 에서 BMDP 페케지를 사용할 수가 있다는 점이다. 뒷 부분에서 설명하겠지만 데이타들은 SAS 에서 만들고 그 데이타들을 BMDP에서 이용할 수가 있다.

특히 SAS가 다른 통계 페케지와 달리 통계분석 면에서 특징을 갖고 있다고 하면 회귀분석 방법 과정 중 GLM(Generalized Linear Model)이다. GLM은 포괄적으로 회귀분석 뿐만아니라 분산분석에서도 샘플 사이즈가 다른 경우와 Missing Cell인 경우에도 통계량을 얻을수 있다. 특히 데이타가 없는 Missing Cell의 경우 SPSS^{X}와 BMDP에서는 처리를 할 수가 없다. 이 GLM에서는 Normal Equations $X'X = X'Y$ 를 푸는 과정에서 역행렬을 얻을 때 Sweep Operator 를 씀으로서 rank (X)가 Full

Rank 뿐아니라 Full Rank가 아닌 경우에도 A Generalized Inverse*를 구할 수가 있다. 또한 Contrast, Multivariate 경우 Multiple Comparison of means 등을 Option으로 처리하여 통계량을 구할 수 있는 장점과 SAS만이 갖고 있는 Type Ⅰ, Ⅱ, Ⅲ, Ⅵ 같은 특수한 검정은 그 나름대로 각광을 받고 있다.

최근 IBM-XT와 IBM-AT用으로 나온 SAS PC는 SAS의 Basic Statistics와 IML이 있다. 이 중 SAS/IML은 데이타의 입력이 행렬식이고 어떤 행렬식의 Operation도 가능하도록 되어 있다. 이는 마치 범용컴퓨터에서 쓰이는 SAS 중 Proc Matrix와 같은 역할을 한다. 퍼스날 컴퓨터용 SAS는 300K 이상의 Memory 사이즈와 Hard Disk를 요구하고 있다.

3. SPSS^X(Statistical Package for the Social Sciences)

SPSS^{X}는 통계 페케지 중 국내에서 제일 많이 쓰이고 있는 SPSS의 최신 Version이다. SPSS는 1965년 Stanfod 대학에서 개발하기 시작하여 계속 수정 보완하여 SPSS Release 9까지 개발되었지만 SAS가 다양하게 보급 됨에 따라 1977년부터 새로운 통계 페케지 개발을 서두르게 되었다. 1982년 7월 IBM/OS에 테스트를 하면서 SPSS^{X}를 내놓게 되었으며 1983년 4월 VAX의기종의Version을 만들면서SPSS^{X}의 보급이 다양화하게 되었다. 국내에서는 이 새로운 SPSS^{X}가 보급 되어지고 있는 곳은 몇군데 안되고 대부분 SPSS가 보급 되어져 있다. SPSS에 관해서는 이 동우 교수(1984)가 SPSS에 관해 언급한 바가 있으므로 새로 변모한 SPSS^{X}에 관해 논해 보고자 한다.

SPSS^{X}는 SPSS가 갖고 있는 통계적 방법은 대부분 그대로 놔두고 SPSS의 최대 약점인 화일 관리에 보다더 확장을 하여 데이타 관리면에 많은 수정 보완을 하였다. 예를 들면 제일 먼저 변한 것은 코딩에서 Format가 Free 로 바뀌었다. SPSS에서는 Columm 1과 Columm 16부터 프로그램 Coding을 하였지만 SPSS^{X}에서는 이러한 제약 조건을 두지 않고 사용할 수가 있다.

다음 SPSS에서는 Data Input 과정에서 꼭 Fixed Format로 주어져야만 했지만 SPSS^{X}에서는 Free format으로도 데이타를 읽을 수가 있다는 점이다. 〈표 1〉 에서는 SPSS 에서 SPSS^{X}로 바뀐 명령어를 요약했고 〈표 2〉에서는 없어진 명령어를 요약하였다. 대부분의 통계

〈표 1〉 바뀐 명령어

SPSS 명령어	SPSS^{X}명령어
ADD CASES	ADD FILES
ADD DATA LIST	MATCH FILES
ADD SUBFILES	ADD FILES
ADD VARIABLES	MACH FILES
DELETE VAR	GET or SAVE
END INPUT DATA	END DATA
GET FILE	GET
INITIALIZE	LEAVE
INPUT FORMAT	DATA LIST
INPUT MEDIUM	FILE HANDLE
KEEP-VARS	GET or SAVE
LAG	LAG function
LIST FILEINFO	DISPLAY
LIST CASES	LIST
MERGE FILE SS	MATCH FILES
PAGESIZE	SET
PRINT BACK	SET
RAW OUTPUT UNIT	PROCEDURE OUTPUT
READ INPUT DATA	BEGIN DATA
READ MATRIX	INPUT MATRIX
REORDER VARS	GET or SAVE
RUN NAME NAME	TITLE
RUN SUBFILES	SPILT FILE
SAVE FILE	SAVE
SEED	SET
TASK NAME	SUBTITLE
VARIABLE LIST	DATA LIST
WRITE FILEINFO	EXPORT
NEW REGRESSION	REGRESSION
WRITE CASES	WRITE or EXPORT

*A가 X의 a generalized inverse라 함은 XAX=X를 만족하는 행렬을 말한다.

〈표 2〉 SPSS^{x} 없어진 명령어

ALLOCATE
ASSIGN MISSING
DELETE SUBFILES
FILE NAME
GET ARCHIVE
LIST ERRORS
OSIRIS VARS
SAVE ARCHIVE
SUBFILE LIST
CANCORR
GUTTMAN SCALE
TRANSFORM

처리과정의 방법들은 다시 쓰였거나 그대로 남겨 두었다.

미국 KANSAS 대학교에서 SPSS^{x}를 테스트 한 결과 첫째로 화일관리면에서 속도가 5~6배로 빨라지고 둘째로 데이타 입출력 면에서 작업과정이 많이 줄어 들어 데이타의 관리가 무척 용이해졌고 세째로 SPSS의 명령어가 다른 패케지보다 용이하지만 너무 낡은 방식이었기 때문에 불편한 점이 많았지만 SPSS^{x}에서는 효과적이고 쉬운 명령어가 채택되어 다른 패케지를 이용했던 사용자도 쉽게 SPSS^{x}를 적응할 수 있는 점이다. 또 SPSS에 없었던 새로운 분석 과정들을 넣었는데 LOGLINEAR, MATRIX, PROBIT, LISREL VI 등이다. 또한 1985년에 들어SPSS^{x}는 SAS와 데이타 교환할 수 있는 과정 등을 삽입한 Release 2.1을 내놓았다.

SPSS^{x}도 IBM-XT과 IBM-AT用SPSS/PC는 Interactive 시스템으로 적은 명령어들로 쉽고 빠르게 데이타의 분석을 할 수가 있으며데이타의 Sorl, 변환 등 빠르고 정확하게 얻을수 있다. 또 데이타의 입력도 alphanumeric, integer, decimal 등 자유자재로 할 수가 있다. 이 SPSS/PC는 MS/DOS를 쓰는 기종과 320K Memory, Hard Disk와 8087 Co-Processar를 요구하고 있으며 Work Space는 64 K 이다. 64 K의 Work Space는 곧 256 K로 증가될 것으로 알려지고 있다. 64 K의 Work space로처리할 수 있는 범위는 Cross Tab의 경우 2150 cells를 다룰 수 있으며 MEANS 경우 2-dimension으로 1450 cell을 다룰 수 있으므로 연구용이나 대학교의 실험 실습용으로는 충분한 Work space이다.

4. BMDP (Biomedical Computer Programs)

1961년 처음 UCLA 대학교에서부터 BMD라는 이름으로 시작되었으며 1968년부터 BMDP라는 패케지로 출발되었다. BMDP 패케지는 이름 그대로 Biomedical과 실험계획의 데이타를 통계적 분석을 하는 다양한 통계 패케지이다.국내에서는 현재 서울대학교, 고려대학교를 포함하여 몇몇 기관에서 사용하고 있지만 최신 version을 갖고 있는 곳은 서울대학교 뿐이다. 한국과학기술원 경우 사용자의 빈도가 낮아 더 이상 사용치 않아 구입을 하지 않고 있고 대부분 옛날 Version을 사용하고 있다. 이와같은 면은 BMDP가SPSS^{x}나 SAS 보다 데이타 처리나 화일 관리하는데 너무도 빈약하고 통계 분석에는 초보자들이 사용하기 힘든 과정이 많아 경시되는 경향이다.

그러나 BMDP는 그 차체가 갖고 있는 특수성을 고려해 보면 통계적으로 높은 차원의 데이타를 분석하는 통계적 처리 방법이 제일 많이 포함되어 있어 통계학자, 사회학자, 심리학자, 생물학자, 의학자들은 아직도 BMDP를 사용하고자 한다. 또한 Manual의 예제 어느 통계 패케지보다도 잘 설명되어 있다.

최근 80년에 들어 BMDP는 새로 Version을 통해 Data 처리와 화일 관리에 많은 신경을 쓰고 있지만 BMDP에서는 그 고유성을 잃지 않으려고 애쓰고 있다. 81년도 Version에는 40개가 넘는 통계 프로그램이 넘게 있으며 Box-Jenkins의 Time Series, Survial Analysis, Stepwise Discriminant Analysis, Derivative-free Nonlinear regression, Stepwise

logistic regression등 새로 개발되는 통계적처리 방법을 많이 다루었다.

특히 회귀분석에는 서로 다른 5가지의 방법을 제시해 놓은 점은 SAS나 SPSS^{x}과 대조를 이룬다. 이와같은 특수성 때문에 SAS 에서는 BMDP를 연결하여 쓸 수 있도록 해 놓았다. BMDP의 약점인 File 보관, 데이타 처리, 변환 등을 SAS를 이용해 놓고 SAS에 없는 통계적 분석을 BMDP를 불러서 쓸 수 있도록 해 놓았다. 자세한 논평은 JASA(1978) 해 놓았다.

BMDP 역시 퍼스날 컴퓨터용 BMDP를 MC 68000의 CPU를 이용하여 Unix에 내놓았으나 실패하고, 1985년부터 MS/DOS용으로 내 놓았으며 SAS, SPSS와 같이 Hard Disk와 많은 Memory를 요구하는 것은 마찬가지이다.

5. 결론

BMDP, SAS, SPSS^{x}의 비교사항을 〈표3〉와 〈표4〉에 설명하였다. 〈표3〉는 시스템에 관한 비교이고 〈표4〉는 실제로 많이 쓰이는 부분을 5가지로 나누어 비교하였다. 이와같이 BMDP는 통계적 측면에서 우수하지만 화일 관리, 데이타 처리 등에 약한 점을 들 수가 있다. SPSS와 SAS는 모든 면에서 우수하게 나타나 있는 것을 알 수 있다. 그러나 이 세 통계 페케지들은 각각 독자적인 우수성을 갖고 있고 서로 다른 환경과 서로 다른 목적에 의해 발전해 가고 있다. 앞으로도 계속되어 이 페케지들은 발전될 것이라는데 아무도 의심할 여지 없다.

〈표 3〉 **BMDP, SAS, SPSS^x 비교표**

이 름	사 용 도	사용언어	기 종	Operating System	Site	Note
BMDP	EMDA	FORTRAN	Most	Most	1500	41개의 프로그램으로 구성되어 있으며 통계적 측면에서 매우 우수함.
SAS	EMTDA	PL/I+ Assembler	IBM+ Compatible DEC-VAX PRIME DG-ECLIPSE	OS/VS VM/CMS DOS/VSE VMS/PRIMOS AOS	3400	강력하고 손쉽게 데이타관리를 할 수 있고 75개의 Procedure를 갖고 있음. Color graph도 손쉽게 처리
SPSS^x	EMTDA	FORTRAN+ Assembler	Most	Most	4000	아주 쉽게 통계적 데이타를 처리하고 많은 면을 사용자 측면에서 개발된 우수한 통계 메이커 페케지

*E-Editing M-Data Management T-Tabulation
D-Descriptive Statistics A-Statistical Analysing

〈표 4〉 **BMDP, SAS, SPSS^{x}의 사용상 비교표**

	BMDP	SAS	SPSS^x
Editing	2	4	5
Data Magement	2	5	4
Tabulation	1	4	4
Descriptive Statistic	4	5	5
Statistical Analysis	5	4	4

*5는 "매우 우수함", 1은 "매우 나쁨"의 척도

참 고 문 헌

1. 박성현, "回歸分析," 大英社, 1983.
2. 여운방, "計量分析의 電算處理," 한국개발연구원 1983.
3. 이동우, '사회과학도를 위한 SPSS의 개요와 이용', 경영과 컴퓨터, 1984
4. Kenneth N. Berte and Ivor S. Francis, "A Review of Manuals for BMDP and SPSS," JASA, 1978

5. Mervin E, Muller, "A Riew of Manuals for B-MDP and SPSS," JASA, 1978.
6. Neffendor, et al, "Software for Statistical and Survey Analysis," Computational Statictics & Data Analysis, 1984
7. Thomas A, Bubolz, "Book Review of BMDP Statatistical Software 1981," JASA, 1984
8. BMDP manual, University of California Press, 1981
9. SAS Manual, SAS, 1982
10. SPSS^{X} Manual, McGraw Hill, 1983