The Korean Communications
in Statistics Vol. 5, No. 1, 1998
pp. 167-176

반복측정 자료를 분석하기 위한 통계패키지의 고찰[1)]

최은숙[2)], 박태성[3)], 문경미[4)]

요 약

최근 들어 반복측정 자료에 대한 관심이 늘어나면서 이러한 자료를 분석하기 위한 통계패키지가 많이 개발되어 사용되고 있다. 본 논문에서는 연속형 반복측정 자료를 분석할 수 있는 통계모형들을 간단하게 개괄해보고 SAS, BMDP, S-PLUS, SPSS, MINITAB과 같은 통계 패키지 중에서 이러한 통계모형들을 다룰 수 있는 프로그램들을 정리해보고 그 특징들을 고찰해보았다. 그 중에서 특히 SAS의 PROC MIXED와 BMDP의 5V를 구체적으로 살펴보았다.

1. 서 론

반복측정 자료(repeated measures data)란 같은 실험 대상(experimental unit)이나 개체(subject)로부터 실험조건이나 처리(treatment)를 달리하여 반복적으로 관찰하여 얻은 자료를 의미한다 (박용규, 송혜향, 1991). 이중에서 특별히 여러 시점(time)에 걸쳐 반복적으로 측정하여 얻은 자료를 경시적 자료(longitudinal data)라고 부르기도 한다. 본 논문에서는 편의상 반복측정 자료와 경시적 자료를 구분없이 사용하였다.

반복측정자료는 한 개체로부터 반복적으로 관찰되었기 때문에 이 자료들은 서로 상관되어 있다. 또한 여러 번 관측하다 보면 결측치(missing observation)나 무응답값(nonresponse)이 자주 발생하게 된다. 이러한 특징들 때문에 반복측정자료 분석을 위한 특별한 형태의 통계모형이 필요하게 된다. 본 논문에서는 먼저 연속형 반복측정자료를 분석할 수 있는 통계모형들을 간단하게 정리해 본 후에 최근에 개발된 통계프로그램들을 살펴보고 그 특성들을 비교해보았다.

그 동안 통계프로그램과 관련된 몇몇 비교 연구결과가 발표되었는데 패키지의 선택과 활용에 관한 연구(김병천, 1987)와 EDA기능에 관한 패키지 비교 연구(허명회, 정진환, 1990)와 시계열 분석방법에 관한 패키지 비교 연구(김수화 등, 1994)가 있었으며 최근에는 공정관리를 위한 통계패키지의 비교에 관한 연구가 발표되었다(조신섭, 신봉섭, 1997).

본 논문의 구성은 다음과 같다. 2절에서는 반복측정자료를 분석하기 위한 통계모형들을 일변량 반복측정 분산분석 모형과 다변량 분산분석 모형 및 공분산 모형으로 분류하여 간단하게 소개하였다. 3절에서는 통계패키지 중에서 SAS, BMDP, SPSS, S-PLUS, MINITAB를 중심으로 반복측정 자료를 처리할 수 있는 기능을 갖춘 프로그램들을 개괄해보고 그 특징을 비교해보았다. 마지

1)이 연구는 1994년도 한국과학재단 연구비지원에 의한 결과임(과제번호: 94-0701-01-01-3)
2)(449-791) 경기도 용인시 모현면 왕산리 한국외국어대학교 통계학과
3)(449-791) 경기도 용인시 모현면 왕산리 한국외국어대학교 통계학과
4)(449-791) 경기도 용인시 모현면 왕산리 한국외국어대학교 통계학과

막 절에서는 실제 응용을 위한 제언을 하였다.

2. 반복측정 자료 분석을 위한 통계 모형

반복 측정자료를 분석하기 위한 통계모형은 크게 세 가지로 분류할 수 있다. 첫 번째 모형은 보통의 일변량 분산분석(ANOVA) 모형을 확장하여 반복측정 자료를 분석할 수 있도록 확장한 반복측정 분산분석(repeated measures ANOVA) 모형이다. 두 번째 모형은 각 개체로부터 얻어진 반응변수 벡터가 서로 독립인 다변량 정규분포를 따른다는 가정을 기초로 하는 다변량 분산분석(MANOVA) 모형이다. 이 모형은 반응변수의 공분산행렬의 구조에 대하여 아무런 가정을 하지 않으며 모든 원소들을 모수로 간주하여 추정하게 된다. 이에 비해 세 번째 모형인 구조적 공분산 모형(structured covariance models)은 반응변수의 공분산행렬의 여러 구조적 형태를 고려할 수 있는 모형으로 가장 일반적인 모형이다. 반복측정 분산분석 모형과 다변량 분산분석 모형은 널리 알려져 있는 고전적인 방법이므로 간단하게 정리해보고 구조적 공분산 모형을 자세히 정리해보았다.

2.1 반복측정 분산분석(repeated measures ANOVA) 모형

이 방법은 보통의 일변량 분산분석 모형에 개체의 효과를 나타내는 랜덤효과를 추가한 모형이다. 이 모형에서는 오차항과 랜덤효과항이 서로 독립이며 각각 정규분포를 따른다는 가정이 추가로 필요하게 된다. 이 추가된 랜럼항에 의하여 같은 개체에서 관찰된 종속변수들간에 상관관계가 존재하게 되며 다른 개체에서 관측된 종속변수들 간에는 독립 관계가 성립하게 된다. 그러나 이 모형은 같은 개체에서 얻어진 종속변수들 간의 상관관계를 나타내는 공분산행렬이 구형성(sphericity or compound symmetry)의 형태가 된다고 하는 강한 가정을 내포하고 있다. 이 가정은 종속변수들 간의 상관관계가 시점에 상관없이 항상 동일하다는 의미이다.

예를 들어, 그룹 h에서 i번째 개체로부터 j번째 얻은 반응변수를 y_{hij}라고 표시하면 반복측정 분산분석 모형은 아래와 같이 정의할 수 있다.

$$y_{hjk} = \mu + \alpha_h + \beta_{hi} + \gamma_j + \delta_{hj} + e_{hij},$$

$h=1,2,\cdots,g$ $j=1,2,\cdots,n_h$ $k=1,2,\cdots,t$. 여기서 μ는 전체평균이고 α_h는 h번째 그룹의 효과를 나타내며 β_{hi}는 h번째 집단 내에서의 i번째 개체의 랜덤효과를 나타내는 확률변수로 정규분포를 따르고, γ_j는 j번째 시점의 효과를 나타내고 δ_{hj}는 h번째 그룹의 효과와 j번째 시점 효과간의 교호작용을 나타낸다. 마지막 항 e_{hij}는 오차항으로 역시 β_{hi}와 마찬가지로 확률변수이며 정규분포를 따른다고 가정한다.

이 모형은 보통의 분산분석 모형에서 개체효과항만을 랜덤으로 선언하면 되기 때문에 대부분의 통계패키지에 있는 분산분석용 프로그램을 이용하여 분석할 수 있다. 이 모형에 대한 자세한 설명은 박용규, 송혜향(1991) 또는 Winer(1971)를 참조하기 바란다.

2.2 다변량 분산분석(MANOVA) 모형

다변량 분산분석 모형은 일변량 분산분석 모형을 다변량으로 확장한 모형이다. 이 모형은 다변량 정규분포를 따르는 서로 독립인 종속변수들로 구성된 행렬 Y의 기대값이 공변량 행렬 X과 모수행렬 B의 곱으로 표시되는 다음과 같은 형태를 따른다.

$$Y = XB + E$$

여기서 E는 오차항 행렬을 나타낸다. 모수행렬 B에 대한 추정과 검정에 관하여 많은 고전적인 연구가 있으며 특히 모수 효과를 보기 위한 검정통계량으로 다음의 대표적인 4가지 통계량이 사용된다 (Johnson and Wichern, 1992).

① Wilks' Λ ② Pillai's Trace ③ Hotelling-Lawley Trace ④ Roy's maximum Root

이들 통계량은 표본의 크기가 큰 경우 즉 오차 자유도가 큰 경우에는 근사적으로 동일한 분포를 따르게 된다. 이들 4가지 통계량은 모두 오차제곱합 행렬(Error Sum of Squares Matrix)의 특성치로부터 쉽게 계산할 수 있다. 만약 비교하는 그룹의 수가 2개인 경우에는 이들 통계량은 Hotelling T^2 통계량과 동일한 검정결과를 제공하게 된다.

이러한 다변량 분산분석 모형을 확장하여 반복측정 자료를 분석할 수 있도록 확장한 모형으로 프로파일 분석 (profile analysis) 모형과 성장곡선모형(growth curve model) (Potthoff and Roy, 1964) 등이 있다. 이 모형에 대한 추정과 검정은 다변량 분산분석 모형의 방법과 동일하다.

이 모형들은 반복측정 분산분석 모형과는 달리 공분산행렬의 구형성에 관한 가정을 필요로 하지 않으며 오차 행벡터들의 다변량 정규 분포성(multivariate normal distribution)에 대한 가정만을 필요로 한다. 그러나 관측된 개체의 수가 시점의 수보다 적으면 검정통계량의 분포를 구하는데 어려움이 발생할 수 있다.

2.3 구조적 공분산 모형

지금까지 반복측정 분산분석 모형과 다변량 분산분석모형을 살펴보았다. 이제 살펴볼 모형은 좀 더 일반적인 형태의 모형으로 다양한 형태의 공분산행렬을 사용할 수 있는 장점을 갖고 있다. $i(=1,\cdots,n)$번째 개체로부터 시점 $j(=1,\cdots,t)$에서 관측된 종속변수를 y_{ij}라고 표시하고 y_{i1}부터 y_{it}로 이루어진 벡터를 y_i라고 표시하면 다변량 선형모형은

$$y_i = X_i\beta + e_i$$

으로 표현될 수 있다. 여기서 X_i는 txp 공변량 행렬(covariate matrix)이고 β는 px1 회귀모수 벡터이고 e_i는 tx1 오차항벡터로 다변량 정규분포를 따른다고 가정하자. 즉 $e_i \sim i.i.d.\ N_t(0,\Sigma)$이다. 이 모형에 대하여 Jennrich and Schluchter(1986)는 여러 형태의 구조적 공분산행렬을 사용하여 최대우도추정법으로 모수벡터 β를 추정하는 방법을 소개했다. β의 MLE를 구하기 위해서는 대개 반복적인 알고리즘이 필요하게 된다. Jennrich and Schluchter(1986)는 Newton-Raphson알고리즘과 Fisher scoring 알고리즘과 EM 알고리즘을 제안

하였다. 이 모형의 Σ는 여러 가지 구조를 가질 수 있다. <표 1>에는 반복측정 자료의 분석에 사용될 수 있는 여러 형태의 공분산행렬의 구조를 정리해놓았다. <표 1> 이외의 다른 형태의 공분산행렬은 Jennrich and Schluchter(1986)를 참조하기 바란다.

<표 1> $t=3$ 인 경우의 공분산행렬 구조의 예

구조 형태				
독립	복합대칭 (CS)	AR-1	비구조	랜덤
$\sigma^2\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}$	$\sigma^2\begin{bmatrix}1&\rho&\rho\\\rho&1&\rho\\\rho&\rho&1\end{bmatrix}$	$\sigma^2\begin{bmatrix}1&\rho&\rho^2\\\rho&1&\rho\\\rho^2&\rho&1\end{bmatrix}$	$\begin{bmatrix}\sigma_{11}&\sigma_{12}&\sigma_{13}\\\sigma_{12}&\sigma_{22}&\sigma_{23}\\\sigma_{13}&\sigma_{32}&\sigma_{33}\end{bmatrix}$	$Z\sigma_b^2 Z' + \sigma^2 I$

위의 모형 중에서 '랜덤' 모형은 $y_i = X_i\beta + e_i$ 중에 오차항이 또 다른 랜덤효과항을 포함하고 있는 경우로 오차항이 $e_i = Z_i b_i + U_i$로 표현되는 경우이다. 여기서 Z_i는 $t_i \times k$ 인 행렬이고, b_i는 평균이 0 이고 분산이 $\sigma_b^2 I$인 정규분포를 따르고, U_i는 평균이 0 이고 분산이 $\sigma^2 I$ 인 정규분포를 따른다고 가정한다. 그리고, b_i와 $\sigma^2 I$ 는 서로 독립이다. 따라서 y_i의 분산은

$$Var\ (y_i) = Var(Z_i b_i) + Var(U_i) = Z_i \sigma_b^2 Z_i' + \sigma^2 I$$

이 된다. 이 모형은 모형 안에 랜덤항이 포함되어 개체내의 변동을 효과적으로 설명할 수 있다. 이 모형을

$$y_i = X_i\beta + Z_i b_i + U_i$$

으로 다시 표현할 수 있으며 고정효과항과 랜덤효과항을 동시에 포함하고 있으므로 혼합모형(mixed model)이라고 부른다. 이러한 형태의 혼합모형은 Laird and Ware(1982)가 제안하였다.

3. 반복 측정자료를 분석하기 위한 통계 프로그램의 고찰

이제 통계패키지 중에서 SAS, BMDP, S-PLUS, SPSS, MINITAB을 중심으로 2절에서 살펴본 반복측정자료를 분석하기 위한 여러 형태의 통계 모형을 다룰 수 있는 프로그램들을 간략히 정리해 보고 그 특성을 살펴보자.

3.1 SAS

SAS에서 연속형 반복측정 자료를 처리할 수 있는 통계 프로그램은 PROC ANOVA, PROC GLM, PROC MIXED가 있다.

3.1.1 PROC ANOVA

PROC ANOVA는 분산분석의 가장 기본적이고 효율적인 프로그램이다. 그러나 이 프로그램은 처리별로 같은 수의 측정치가 있는 경우(균형 자료인 경우)의 분산분석에만 이용될 수 있다. 이 프로그램을 이용하여 반복측정 분산분석 모형을 추정할 수 있다. 그러나 만약 자료가 불균형 자료라면 정확한 분석이 불가능하므로 좀 더 일반적 분석을 할 수 있는 프로그램인 PROC GLM을 사용해야한다.

3.1.2 PROC GLM

PROC GLM은 반복측정 분산분석과 다변량 분산분석을 모두 처리할 수 있는 프로그램이다. 이 프로그램 안에 MANOVA나 REPEATED에서의 NOU 옵션 등으로 다변량 분석을 할 수 있다. 다변량 검정통계량인 Wilk's Λ 통계량, Lawley-Hotelling 검정통계량, Pillai's 검정통계량과 Roy's 최대근 검정통계량의 값을 함께 출력시킨다. 다변량 분석에서는 결측치가 생긴 경우 결측치를 포함하는 개체에 대하여서는 그 개체로부터 관측된 자료까지 제거한 후에 분석을 한다. 그러나 반복측정 분산분석 모형을 사용한 경우에는 결측치가 존재하더라도 나머지 관측된 자료를 이용하여 분석한다. 다변량 분산분석 모형은 기본적으로 공분산행렬이 구형성 가정을 만족한다는 가정 하에서 분석을 실시한다. GLM은 공분산행렬의 구형성 가정에 대한 검정을 실시할 수 있다. 분산이 구형성을 만족하지 않으면 정확한 분석을 할 수 없다. 이런 경우 모든 분산의 형태에 따라 분석할 수 있는 프로그램이 PROC MIXED이다.

3.1.3 PROC MIXED

PROC MIXED을 이용하여 2절에서 소개한 공분산행렬 모형을 추정할 수 있다. 특히 고정효과와 랜덤효과를 모두 포함하고 있는 혼합모형을 다룰 수 있으며 공분산행렬의 구조를 지정해 줄 수 있다. 즉 구형성이 아닌 경우에도 AR(1), RANDOM, UNST 등의 공분산 구조 형태를 분석할 수 있다. 또한 옵션문을 이용하여 최대우도추정법과 제한된 최대우도추정법(restricted maximum likelihood, REML) 중에서 원하는 추정법을 선택할 수 있는 기능도 갖고 있다. 이 프로그램은 반복측정 자료의 분석 이외에도 다양한 형태의 실험계획 모형을 처리할 수 있는 기능을 갖고 있는 프로그램이다.

3.2 BMDP

BMDP는 다른 통계패키지 보다 더 오랜 역사를 갖고 있으며 주로 의학 분야에서 많이 사용되는 모형들을 처리할 수 있도록 개발되었다. 최근에는 모든 형태의 통계모형들을 다룰 수 있도록 확장되었다. 그러나 워낙 이 프로그램이 오래 전에 FORTRAN으로 작성되었기 때문에 그래픽 기능은 떨어진다. 이 BMDP패키지에는 반복측정 자료를 분석할 수 있는 여러 프로그램들을 포함되어 있다. 그 중에서 3V, 4V, 5V, 8V 등을 살펴보도록 하자.

3.2.1 BMDP 3V

3V는 고정효과 항과 랜덤효과항이 함께 존재하는 혼합모형에서 최대우도추정법과 제한된 최대

우도추정법을 사용하여 모수를 추정한다. 여기서는 특별한 옵션을 사용하지 않고도 불균형 자료를 분석할 수 있다. 만약 균형자료라면 비음 분산 성분(nonnegative variance component)을 추정할 때 고전적 분산분석에서 얻어진 평균과 분산의 추정치와 REML 추정치가 일치하게 된다. 이와 같이 혼합모형에서의 분산분석을 할 수 있는 기능을 갖고 있는 프로그램이 3V이다.

3.2.2 BMDP 4V

4V는 반복측정 자료 모형과 split-plot 모형에 대하여 일변량 분석과 다변량 분석을 모두 할 수 있는 기능을 갖춘 프로그램이다. 즉 반복측정 자료에 대하여 반복측정 분산분석과 다변량 분석을 할 수 있다. 그러나 4V는 결측치가 있는 경우 그 결측치를 포함하는 개체의 모든 자료를 빼고 분석하므로 정확한 분석을 할 수는 없다.

3.2.3 BMDP 5V

5V는 공분산행렬을 포함하는 불균형 반복 자료를 분석할 수 있는 프로그램이다. SAS의 MIXED와 가장 근접한 프로그램이며 SAS의 MIXED보다도 훨씬 오래 전에 개발된 프로그램이다. 이 프로그램도 역시 SAS의 MIXED와 마찬가지로 최대우도추정법과 제한된 최대우도추정법을 모두 사용할 수 있다. 5V는 2V, 3V, 4V에서 다룰 수 있는 대부분의 통계 모형들을 처리할 수 있다. 또한 이 프로그램은 결측치로 인해 자료가 불완전한 경우에도 사용할 수 있다. 즉 결측치를 갖는 개체에서 얻은 다른 자료들을 분석에 포함시킨다. 그러므로 결측치가 있는 경우에도 그 개체의 자료를 버리지 않기 때문에 같은 최대우도추정법을 사용하는 다른 V-series(2V, 3V, 4V, 8V 등)에서 나온 결과와 조금 다른 결과가 출력된다. 이 프로그램은 주로 반복측정 모형을 위한 프로그램이지만 요인분석(factorial design model)에도 적용할 수 있다.

3.2.4 BMDP 8V

8V는 자료가 균형자료일 때만 분산 분석을 실시할 수 있는 제한된 기능을 갖고 있는 프로그램이다. 만약 자료가 균형자료가 아니라면 3V, 4V, 5V 등을 사용하여 분석하는 것이 바람직하다.

3.3 S-PLUS

S-PLUS는 SAS나 BMDP처럼 반복측정 자료를 분석할 수 있는 다양한 프로그램을 갖추고 있지 못하다. 그렇지만 여러 형태의 옵션문을 사용하여 적절한 분석을 할 수 있다. S-PLUS에서 반복측정 자료를 분석할 수 있는 프로그램은 일반적인 일변량 분산분석 모형을 다루는 aov와 다변량 분산분석 모형을 다루는 manov이다. aov를 이용하여 반복측정 분산분석 모형을 적합할 수 있다. aov는 균형자료와 불균형자료를 모두 처리할 수 있으며 자료가 불균형자료인 경우에는 특정의 옵션을 사용하여 분석할 수 있다. manov는 일반적인 다변량 분산분석 모형을 다룰 수 있도록 개발된 프로그램이다. 따라서 manova를 이용하여 반복측정자료에 대한 분석을 할 수 있다. 그러나 이 때에 공분산행렬은 특정한 구조를 명시할 수 없는 비구조적 형태이기 때문에 특정 형태의 공분산행렬을 사용하는 공분산행렬 모형의 분석은 불가능하다. 그리고 만약 자료에 결측치가 있는 경우에는 옵션문을 사용하여 결측치가 포함된 열의 자료를 포함하여 분석할 수도 있고 제외시킨 후에 분석할 수도 있다. 즉 'na'라는 옵션을 사용하면 결측치가 포함된 열의 자료를 포함하

여 분석하라는 의미이며 'null'이라는 옵션을 사용하면 결측치가 포함된 열의 자료를 제외시키라는 의미이다 .

3.4 SPSS

SPSS도 S-PLUS와 마찬가지로 일변량 분산분석과 다변량 분산분석이 가능하나 그 이외의 일반적인 공분산행렬 모형의 분석은 불가능하다. SPSS에는 일변량 분산분석을 위한 ANOVA와 다변량 분산분석을 위한 MANOVA가 있다. ANOVA 모형의 명령문으로는 일반적인 요인분석 ANOVA, 단순 요인분석 ANOVA, 및 일원배치 ANOVA등이 있다.

SPSS에서는 일변량이나 다변량적 접근방법을 사용하여 광범위한 분석을 할 수 있는데, 메뉴에서 ANOVA모형을 선택한 후에 옵션문인 일변량 또는 다변량을 선택한다. 특히, 반복측정자료를 분석하기 위해 ANOVA 모형을 선택한 후에 옵션문인 Repeated measures를 사용하면 된다.

SPSS도 역시 S-PLUS와 마찬가지로 옵션문을 사용하여 균형자료와 불균형자료 모두 분석할 수 있는 특징을 갖고 있다. 즉 자료에 결측치가 있는 경우에 결측치가 포함된 관측치를 포함하여 분석하고자 할 때는 INCLUDE라는 옵션을 사용하면 된다. 여기서 옵션을 선택하지 않을 경우에는 자동적으로 결측치가 포함된 열의 자료를 제외시켜 분석한다.

3.5 MINITAB

MINITAB에는 반복측정 자료를 처리할 수 있는 통계 프로그램으로 ANOVA, MANOVA, GLM이 포함되어 있다. 이 프로그램들은 모두가 결측치가 생긴 경우에 결측치를 포함하는 개체의 나머지 모든 자료를 제거한 후 분석하는 것이 특징이다. 공분산행렬 모형을 다룰 수 있는 기능을 갖춘 프로그램은 현재까지는 개발되지 못한 상태이다.

3.5.1 ANOVA

ANOVA에서는 보통의 분산분석 모형에서 사용하는 일원배치(one-way)모형과 다원배치(multi-way)모형 등을 다룰 수 있다. 일원배치 모형은 불균형 자료를 분석할 수 있으나 다원배치 모형은 균형자료만을 처리할 수 있다. 그러므로 결측치를 포함하는 자료들은 GLM을 사용하여 분석하는 것이 바람직하다. 그러나 균형자료인 경우에는 ANOVA를 사용하는 것이 GLM을 사용하는 것보다 더 빠르며 컴퓨터 공간도 덜 사용하기 때문에 보다 효율적이다.

3.5.2 MANOVA

MANOVA는 ANOVA를 다변량 자료로 확장한 형태로서 일반적인 다변량 분산분석 모형을 처리할 수 있는 기능을 갖고 있다. 또한 검정통계량으로 Wilk's Λ 통계량, Lawley-Hotelling 검정통계량, Pillai's 검정통계량과 Roy's 최대근 검정통계량의 값을 구할 수 있다. 또한 MANOVA에서의 그룹의 크기가 두 개인 경우에는 Hotelling T^2 통계량도 구할 수 있다. 이 프로그램을 이용하여 2절에서 다룬 반복측정 자료를 분석하기 위한 다변량 분산분석 모형을 적합할 수 있다. 그러나 결측치가 생긴 경우에 결측치를 포함하는 열의 자료를 모두 제거한 후에 분석하는 단점을 갖

고 있다.

3.5.3 GLM

MINITAB의 GLM은 SAS의 GLM과 유사한 기능을 갖춘 프로그램으로 일변량의 분산분석 모형과 다변량 분산분석 모형을 처리할 수 있는 기능을 갖고 있다. 특히 여러 개의 부명령어를 사용하여 반복측정 분산분석 모형과 다변량 분산분석 모형을 적합할 수 있는 기능을 갖고 있다. 따라서 이 GLM은 ANOVA와 MANOVA의 기능을 모두 갖춘 일반적인 형태의 프로그램으로 볼 수 있으며 균형자료와 불균형자료를 처리할 수 있다. 그러나 공분산행렬 모형을 다룰 수 있는 기능은 갖고 있지 않다.

3.6 비교

지금까지 여러 가지 통계패키지 중에서 반복측정 자료를 처리할 수 있는 프로그램들을 살펴보았다. 2절에서 살펴본 통계모형 중에서 반복측정 분산분석 모형과 다변량 분산분석 모형은 기존의 고전적인 모형들을 반복측정 자료를 분석할 수 있도록 약간 확장시킨 모형이므로 이 두 모형을 다룰 수 있는 프로그램은 많이 개발되어 있는 상태이다. 그러나 공분산행렬 모형은 현재까지 SAS의 MIXED와 BMDP의 5V에서만 다룰 수 있다. 앞으로 다른 통계패키지들도 이러한 기능을 갖춘 프로그램들을 곧 개발하리라 생각한다. <표2>은 2절에서 다룬 3가지 반복측정 자료 모형을 다룰 수 있는 프로그램들을 통계패키지 별로 분류하여 정리해 보았다.

<표 2> 반복측정 자료를 분석할 수 있는 여러 통계 프로그램들

통계 패키지	반복측정 자료 통계모형			
	1. 반복측정 분산분석		2. 다변량 분산분석	3. 공분산 행렬 모형
	균형자료	불균형 자료		
SAS	ANOVA GLM	GLM	GLM	MIXED
BMDP	4V, 8V	4V	4V	3V, 5V
SPSS	ANOVA	ANOVA	MANOVA	
S-PLUS	aov	aov	manova	
MINITAB	ANOVA GLM	GLM	GLM	

<표 3> 각 프로그램의 특징 비교

통계 패키지	프로그램	추정방법 LSE	MLE	REML	Σ 의 구조 I	CS	AR1	결측치 이용	다변량 분석	불균형 자료 처리
SAS	ANOVA	○	×	×	×	○	×	×	×	×
	GLM	○	×	×	×	○	×	○	×	○
	일변량	○	×	×	×	○	×	×	○	○
	다변량	○	×	×	×	○	×	×	○	○
	MIXED	×	○	○	○	○	○	○	×	○
BMDP	3V	×	○	×	×	○	×	×	×	○
	4V	○	×	×	×	○	×	×	○	○
	5V	×	○	○	○	○	○	○	×	○
	8V	○	×	×	×	○	×	×	×	×
SPSS	ANOVA	○	×	×	×	○	×	○	×	○
	MANOVA	○	×	×	×	○	×	*1	○	○
S-PLUS	aov	○	×	×	×	○	×	○	×	○
	manova	○	×	×	×	○	×	*2	○	○
MINITAB	ANOVA	○	×	×	×	○	×	×	×	*3
	MANOVA	○	×	×	×	○	×	×	○	○
	GLM	○	×	×						

*1: ○(옵션 INCLUDE) ×(옵션 LISTWISE)
*2: ○ ○(옵션 NA) ×(옵션 NULL)
*3: (one-way) ×(multi-way)

<표3>에서는 각 통계패키지 내의 프로그램 별로 몇 가지 기능에 대하여 비교하여 정리해 보았다. 첫 번째 부분은 각 프로그램들이 사용하는 추정방법이 최대제곱추정법(LSE)인지 최대우도추정법(MLE)인지의 여부와 제한된 최대우도추정법(REML)을 사용하는 것이 가능한 지를 표시하였다. 두 번째 부분은 공분산행렬의 구조에 대하여 정리하였고 세 번째 부분은 결측치가 있는 경우에 그 개체 열의 자료를 모두 제외하고 분석을 하는지 아니면 일부를 포함하여 분석을 하는 지를 나타내었다(제외하면 ×, 이용하는 경우는 ○). 네 번째 부분은 다변량 분산분석의 가능 유무를 나타내고 마지막으로 다섯 번째는 불균형자료의 처리 여부를 나타내었다. 즉 그룹마다 개체수가 다른 경우에 대하여 분석의 가능 여부를 나타내었다.

4. 결 론

대부분의 통계패키지들은 고전적인 분산분석 모형과 다변량 분산분석 모형을 처리할 수 있는 프로그램을 갖고 있다. 그러나 가장 일반적인 모형인 공분산행렬 모형을 처리할 수 있는 통계 프로그램들을 많이 개발되지 않은 상태이다. 현재는 SAS의 MIXED와 BMDP의 5V 만이 이 모형을 다룰 수 있다. 앞으로 다른 통계패키지들도 이러한 기능을 갖춘 프로그램들을 곧 개발하리라

생각한다. 여러 가지 통계패키지 중에서 SAS와 BMDP가 반복측정 자료를 분석할 수 있는 가장 많은 프로그램을 보유하고 있는 상태이며 그 기능도 서로 비슷한 것으로 조사되었다.

최근에는 반복측정 자료가 단순하게 종속변수만을 관측하여 그룹의 차이를 본다든지 시간에 따른 단순한 변화를 보는 것 외에도 여러 형태의 독립변수들을 동시에 관측하여 이 독립변수가 종속변수에 미치는 영향을 알아보는 것이 주된 목적이 될 때가 많다. 이런 경우에는 독립변수와 종속변수들 간의 관계를 회귀모형으로 설명하는 것이 가장 바람직한 방법이고 이때 이러한 분석을 쉽게 할 수 있도록 하는 모형이 바로 공분산행렬 모형이다. 또한 이 모형은 관측 시간에 따라 변하는 값을 갖는 시간-종속(time-dependent)인 독립변수를 잘 처리할 수 있는 장점을 갖고 있다.

또한 반복측정 자료는 일정 기간동안 꾸준히 자료를 관측하기 때문에 결측치가 발생하는 경우가 많다. 따라서 결측치를 효과적으로 처리할 수 있는 통계 모형을 사용하는 것이 바람직하다. 즉 결측치를 포함하는 환자의 자료를 처리할 때 결측치를 제외한 나머지 자료는 분석에 포함시키는 것 바람직하다. 공분산행렬 모형 이러한 기능도 역시 갖추고 있다.

이상을 종합해 보면 시간-종속인 독립변수와 결측치 같은 반복측정 자료를 특징을 가장 잘 소화할 수 있는 일반적인 형태의 모형인 공분산행렬 모형을 사용하여 자료를 분석하는 것이 바람직하고 이러한 모형을 적합할 수 있는 프로그램인 SAS의 MIXED나 BMDP의 5V를 사용하는 것을 권하고 싶다.

참고문헌

[1] 김병천(1987). 개인용 컴퓨터에서 통계패키지의 선택과 활용, 『응용통계연구』, 제1권 1호, 75-90.

[2] 김수화, 김승희, 조신섭(1994). 통계패키지에서의 시계열 분석방법의 비교연구, 『한국통계학회 논문집』, 제1권 1호, 119-130.

[3] 박용규, 송혜향 (1991). 『반복측정 자료의 분산분석법』, 자유아카데미, 서울.

[4] 허명회, 정진환(1990). 탐색적 데이터분석(EDA) 기능에 관한 통계패키지 프로그램의 비교 검토, 『응용통계연구』, 제3권 2호, 17-25.

[5] Jennrich, R. I. and Schluchter, M. D. (1986). Unbalanced Repeated-Measures Models with Structured Covariance Matrices. *Biometrics,* **42**, 805-820.

[6] Johnson, R. A. and Wichern, D. W. (1992). *Applied Multivariate Statistical Analysis* (3rd ed.). New York: Prentice Hall.

[7] Laird, N. M., and Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics*, **38,** 963-974.

[8] Potthoff, R. F. and Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika,* **51**, 313-326.

[9] Winer, B. J. (1971). *Statistical Principles in Experimental Design*, 2d Edition, New York: McGraw-Hill Book Co.