The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.2.201
28(2), 201–210

Assessing Correlation between Two Variables in Repeated Measurements using Mixed Effect Models

Kyunghwa Han[a] · Inkyung Jung[b,1]

[a]Yonsei Biomedical Research Center, Department of Radiology,
Research Institute of Radiological Science, Yonsei University College of Medicine
[b]Department of Biostatistics and Medical Informatics, Yonsei University College of Medicine

(Received March 13, 2015; Revised April 2, 2015; Accepted April 6, 2015)

Abstract

Repeated measurements on each variables of interest often arise in bioscience or medical research. We need to account for correlations among repeated measurements to assess the correlation between two variables in the presence of replication. This paper reviews methods to estimate a correlation coefficient between two variables in repeated measurements using the variance-covariance matrix of linear mixed effect models. We analyze acoustic radiation force impulse imaging (ARFI) data to assess correlation between three shear wave velocity (SWV) measurements in liver or spleen and spleen length by ultrasonography. We present how to obtain parameter estimates for the variance-covariance matrix and correlations in mixed effects models using PROC MIXED in SAS.

Keywords: Linear mixed model, repeated measures, correlation, PROC MIXED.

1. 서론

생명과학 또는 의학 연구에서는 반복 측정된 변수들 간의 상관관계를 보고자 하는 경우가 발생한다. 예를 들어 영상의학 분야에서는 크론병 환자의 염증성 장질환에 대한 점수를 contrast-enhanced T1-weighted sequences magnetic resonance enterography(CE-MRE)와 diffusion-weighted imaging(DWI-MRE)로 측정하여 두 결과 간의 상관관계를 보고자 하는 경우가 있는데, 각 환자의 소장의 여러 분절(segment)에 대해 측정되고 분절의 개수가 환자마다 상이하므로 이를 고려하여 상관분석을 해야 한다. 또 다른 예로는 그레이브스 병 환자의 갑상선 기능 검사 항목인 혈중 T3, T4, 갑상선 자극 호르몬 등과 다검출 전산화 단층 촬영(multiple detector computed tomography; MDCT)을 이용한 이미지 측정치인 안외근(extraocular muscles; EOM)의 부피 간의 상관관계를 보고자 하는 경우에서, 갑상선 기능 검사 항목은 환자마다 고유한 값이지만 안외근 부피는 양안에서 측정되어 환자마다 두 개의 값을 가지므로 반복 측정됨을 고려해야 한다.

위와 같이 반복 측정된 변수들 간의 상관관계를 분석하는 데에 기본적으로 고려해 볼 수 있는 방법은 반복 측정됨을 고려하지 않고 피어슨이나 스피어만의 상관계수를 구하는 것이다. 이는 대상자 수

[1]Corresponding author: Department of Biostatistics and Medical Informatics, Yonsei University College of Medicine, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 120-752, Korea. E-mail: ijung@yuhs.ac

가 아닌 전체 측정치의 개수를 고려하므로 자유도가 높아져서 부적절한 제1종 오류를 발생시키게 된다 (Bland와 Altman, 1994). 이러한 문제를 해결하기 위해 대상자 별 반복 측정치들의 평균값에 대한 상관분석을 시행할 수 있으나, 이 역시 대상자마다 반복 측정된 횟수가 다른 경우에 이를 고려하기 어렵고 실제 상관관계를 과소 추정할 수 있어 적절하지 않다 (Lam 등, 1999). Bland와 Altman (1995a)은 대상자 간 변이를 제외하는 부분상관계수를 제안하였고, Bland와 Altman (1995b)는 각 대상자 별 반복측정 횟수를 가중치로 고려하는 가중상관계수를 제안함으로써 대상자 별로 반복측정 횟수가 다른 점을 해결하였지만, 변수들의 반복측정 횟수는 모든 대상자에게서 동일하다는 가정을 필요로 하고 환자 별 반복 측정치들의 평균값을 이용하기 때문에 상관관계가 과소 추정되는 경향이 있다.

선형혼합모형을 이용하여 반복 측정된 변수들 간의 상관계수를 구하는 방법은 Lam 등 (1999)이 제안하고 이어서 Hamlett 등 (2003)이 SAS의 MIXED procedure를 이용하는 방법을 제시하였다. Lam 등 (1999)은 변수들의 반복측정 시점이 동일한 경우에 대해서만 다루었고, Hamlett 등 (2003, 2004)은 반복측정 시점이 동일하지 않은 경우에 대해서까지 확장하였다. 그리고 Roy (2006)는 상관구조를 좀 더 다양하게 고려하거나 혼합모형에 임의 절편과 임의 계수를 고려하는 모형에 적용하는 것으로 확장하였다.

본 논문에서는 반복 측정된 변수들 간의 상관계수를 선형혼합모형을 이용하여 추정하는 방법을 소개하고 실제 자료에 적용해 보았다. 2절에서는 선형혼합모형을 소개하고 3절에서는 제안된 방법을 영상의학 자료에 적용하여 분석하고 사용된 SAS 프로그램을 실었다. 4절에서는 간단한 모의실험을 통하여 선형혼합모형을 이용한 방법과 다른 방법들과의 차이를 살펴보고, 5절에서는 결론을 맺는다.

2. 혼합모형

2.1. 선형혼합모형

i번째 대상자에서 j번째 반복 측정된 두 변수 U와 W를 (U_{ij}, W_{ij})라 하고, (U_{ij}, W_{ij})는 평균이 (μ_U, μ_W)이고 분산-공분산 행렬이 Σ인 이변량 정규 분포를 따른다고 가정하면, Σ는 다음과 같이 표현할 수 있다.

$$\Sigma = \begin{pmatrix} \sigma_U^2 & \sigma_{UW} \\ \sigma_{UW} & \sigma_W^2 \end{pmatrix},$$

여기서 σ_U^2과 σ_W^2은 U와 W의 분산이고 σ_{UW}는 U와 W의 공분산이다. $\sigma_{UW} = \sigma_U \sigma_W \rho_{UW}$로 표현할 수 있고 ρ_{UW}가 우리가 구하고자 하는 U와 W의 상관 계수이다.

i번째 대상자의 관측치를 $Y_i = (U_{i1}, W_{i1}, U_{i2}, W_{i2}, \ldots, U_{in_i}, W_{in_i})'$라 하고 X_i를 고정효과에 대한 계획행렬(design matrix), Z_i를 임의효과에 대한 계획행렬이라 하면 식 (2.1)과 같은 선형혼합모형을 고려해 볼 수 있다.

$$Y_i = X_i\beta + Z_i\gamma_i + \varepsilon_i, \tag{2.1}$$

여기서 β는 고정효과의 회귀계수의 벡터, γ_i는 임의효과의 벡터이고 ε_i는 임의오차의 벡터를 나타낸다. γ_i는 평균벡터가 0이고 공분산 행렬이 G인 다변량 정규분포를, ε_i는 평균벡터가 0이고 공분산 행렬이 R_i인 다변량 정규분포를 따르며 이 둘은 서로 독립이라고 가정한다. 이러한 가정에 따르면 식 (2.1)에서 $E(Y_i) = X_i\beta$이고 $\mathrm{Var}(Y_i) = Z_i G Z_i^{'} + R_i$로 표현된다.

우리의 관심은 U와 W임으로 위 모형 식에서 계획행렬인 X_i는 절편을 포함하여 세 개의 열과 $2n_i$개의

행을 가지는 행렬로, β는 세 개의 모수 $\beta_0, \beta_1, \beta_2$를 원소로 가지는 벡터로 다음과 같이 표현할 수 있다.

$$X_i = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

여기서 X_i는 완전계수(full rank)가 아니므로 $\beta_2 = 0$으로 설정하여 β_0와 β_1이 추정 가능하도록 한다. 그러면 U의 평균은 $\beta_0 + \beta_1, W$의 평균은 β_0로 추정할 수 있다.

임의효과 γ_i와 그에 대한 계획행렬 Z_i는 우리가 가정하는 Y_i의 분산-공분산 행렬의 형태에 따라 적절히 정의할 수 있고, 2.2절과 2.3절에서 설명한다.

2.2. 반복 측정이 동시에 된 경우

두 변수 U, W의 반복 측정이 동시에 된 경우에는 동일한 시점에서 측정된 변수들 간 상관관계와 서로 다른 시점에서 측정된 변수들 간 상관관계가 다르다고 생각할 수 있다. 서로 다른 시점(j와 $j', j \neq j'$)에 측정된 두 변수의 상관계수는 $\mathrm{Corr}(U_{ij}, U_{ij'}) = \rho_U$, $\mathrm{Corr}(W_{ij}, W_{ij'}) = \rho_W$, $\mathrm{Corr}(U_{ij}, W_{ij'}) = \rho_{UW}\delta$와 같이 표현할 수 있다. 서로 다른 시점에 측정된 변수들 간의 상관관계는 동일한 시점에서 측정된 경우보다 작을 것으로 기대되기 때문에 δ는 1보다 작다고 가정한다. 각 대상자 별로 n_i번씩 반복 측정된 경우 i번째 대상자에 대한 U와 W의 분산-공분산 행렬은 다음과 같이 나타낼 수 있다.

$$V_i = \mathrm{Cov}\begin{pmatrix} U_{i1} \\ W_{i1} \\ U_{i2} \\ W_{i2} \\ \vdots \\ U_{in_i} \\ W_{in_i} \end{pmatrix} = \begin{pmatrix} \sigma_U^2 & \sigma_{UW} & \sigma_U^2\rho_U & \sigma_{UW}\delta & \cdots & \sigma_U^2\rho_U & \sigma_{UW}\delta \\ \sigma_{UW} & \sigma_W^2 & \sigma_{UW}\delta & \sigma_W^2\rho_W & \cdots & \sigma_{UW}\delta & \sigma_W^2\rho_W \\ \sigma_U^2\rho_U & \sigma_{UW}\delta & \sigma_U^2 & \sigma_{UW} & \vdots & \sigma_U^2\rho_U & \sigma_{UW}\delta \\ \sigma_{UW}\delta & \sigma_W^2\rho_W & \sigma_{UW} & \sigma_W^2 & \vdots & \sigma_{UW}\delta & \sigma_W^2\rho_W \\ \vdots & \vdots & \cdots & \cdots & \ddots & \vdots & \vdots \\ \sigma_U^2\rho_U & \sigma_{UW}\delta & \sigma_U^2\rho_U & \sigma_{UW}\delta & \cdots & \sigma_U^2 & \sigma_{UW} \\ \sigma_{UW}\delta & \sigma_W^2\rho_W & \sigma_{UW}\delta & \sigma_W^2\rho_W & \cdots & \sigma_{UW} & \sigma_W^2 \end{pmatrix}.$$

식 (2.1)의 Y_i의 분산을 추정하기 위해 G와 R_i를 추정해야 하는데, 여기서는 서로 다른 두 시점의 U와 W 간의 상관관계를 나타내는 행렬과 동일한 시점에서의 상관관계를 블록으로 나타내는 행렬로 나누어 V_i를 다시 표현해보면 아래와 같다.

$$V_i = \begin{pmatrix} \sigma_U^2\rho_U & \sigma_{UW}\delta & \sigma_U^2\rho_U & \sigma_{UW}\delta & \cdots & \sigma_U^2\rho_U & \sigma_{UW}\delta \\ \sigma_{UW}\delta & \sigma_W^2\rho_W & \sigma_{UW}\delta & \sigma_W^2\rho_W & \cdots & \sigma_{UW}\delta & \sigma_W^2\rho_W \\ \sigma_U^2\rho_U & \sigma_{UW}\delta & \sigma_U^2\rho_U & \sigma_{UW}\delta & \vdots & \sigma_U^2\rho_U & \sigma_{UW}\delta \\ \sigma_{UW}\delta & \sigma_W^2\rho_W & \sigma_{UW} & \sigma_W^2\rho_W & \vdots & \sigma_{UW}\delta & \sigma_W^2\rho_W \\ \vdots & \vdots & \cdots & \cdots & \ddots & \vdots & \vdots \\ \sigma_U^2\rho_U & \sigma_{UW}\delta & \sigma_U^2\rho_U & \sigma_{UW}\delta & \cdots & \sigma_U^2\rho_U & \sigma_{UW}\delta \\ \sigma_{UW}\delta & \sigma_W^2\rho_W & \sigma_{UW}\delta & \sigma_W^2\rho_W & \cdots & \sigma_{UW}\delta & \sigma_W^2\rho_W \end{pmatrix} + \begin{pmatrix} \eta & \xi & 0 & 0 & \cdots & 0 & 0 \\ \xi & \nu & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \eta & \xi & \vdots & 0 & 0 \\ 0 & 0 & \xi & \nu & \vdots & 0 & 0 \\ \vdots & \vdots & \cdots & \cdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \eta & \xi \\ 0 & 0 & 0 & 0 & \cdots & \xi & \nu \end{pmatrix}, \quad (2.2)$$

여기서 $\eta = \sigma_U^2(1-\rho_U)$, $\nu = \sigma_W^2(1-\rho_W)$, $\xi = \sigma_{UW}(1-\delta)$이다. 식 (2.2)에서 첫 번째 행렬은 $Z_i G Z_i'$를, 두 번째 행렬은 R_i를 의미하고, 따라서 Z_i, G, γ_i, R_i를 다음과 같이 표현할 수 있다.

$$Z_i = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad G = \begin{pmatrix} \sigma_U^2\rho_U & \sigma_{UW}\delta \\ \sigma_{UW}\delta & \sigma_W^2\rho_W \end{pmatrix}, \qquad \gamma_i = \begin{pmatrix} \gamma_{1i} \\ \gamma_{2i} \end{pmatrix}, \qquad R_i = I_{n_i} \otimes \begin{pmatrix} \eta & \xi \\ \xi & \nu \end{pmatrix},$$

여기서 I_{n_i}는 n_i차원의 항등행렬(identity matrix)이다.

2.3. 반복 측정이 동시에 되지 않은 경우

U와 W의 상관 계수는 측정 시점에 상관없이 $\mathrm{Corr}(U_{ij}, W_{ij'}) = \rho_{UW}$로 동일하다고 가정하고 이 때 각 대상자 별로 n_i번씩 반복 측정된 경우 i번째 대상자에 대한 U와 W의 분산-공분산 행렬은 다음과 같이 나타낼 수 있다.

$$V_i = \mathrm{Cov}\begin{pmatrix} U_{i1} \\ W_{i1} \\ U_{i2} \\ W_{i2} \\ \vdots \\ U_{in_i} \\ W_{in_i} \end{pmatrix} = \begin{pmatrix} \sigma_U^2 & \sigma_{UW} & \sigma_U^2\rho_U & \sigma_{UW} & \cdots & \sigma_U^2\rho_U & \sigma_{UW} \\ \sigma_{UW} & \sigma_W^2 & \sigma_{UW} & \sigma_W^2\rho_W & \cdots & \sigma_{UW} & \sigma_W^2\rho_W \\ \sigma_U^2\rho_U & \sigma_{UW} & \sigma_U^2 & \sigma_{UW} & \vdots & \sigma_U^2\rho_U & \sigma_{UW} \\ \sigma_{UW} & \sigma_W^2\rho_W & \sigma_{UW} & \sigma_W^2 & \vdots & \sigma_{UW} & \sigma_W^2\rho_W \\ \vdots & \vdots & \cdots & \cdots & \ddots & \vdots & \vdots \\ \sigma_U^2\rho_U & \sigma_{UW} & \sigma_U^2\rho_U & \sigma_{UW} & \cdots & \sigma_U^2 & \sigma_{UW} \\ \sigma_{UW} & \sigma_W^2\rho_W & \sigma_{UW} & \sigma_W^2\rho_W & \cdots & \sigma_{UW} & \sigma_W^2 \end{pmatrix}.$$

따라서 식 (2.2)와 같은 형태로 다시 나타내면 식 (2.3)과 같다.

$$V_i = \begin{pmatrix} \sigma_U^2\rho_U & \sigma_{UW} & \sigma_U^2\rho_U & \sigma_{UW} & \cdots & \sigma_U^2\rho_U & \sigma_{UW} \\ \sigma_{UW} & \sigma_W^2\rho_W & \sigma_{UW} & \sigma_W^2\rho_W & \cdots & \sigma_{UW} & \sigma_W^2\rho_W \\ \sigma_U^2\rho_U & \sigma_{UW} & \sigma_U^2\rho_U & \sigma_{UW} & \vdots & \sigma_U^2\rho_U & \sigma_{UW} \\ \sigma_{UW} & \sigma_W^2\rho_W & \sigma_{UW} & \sigma_W^2\rho_W & \vdots & \sigma_{UW} & \sigma_W^2\rho_W \\ \vdots & \vdots & \cdots & \cdots & \ddots & \vdots & \vdots \\ \sigma_U^2\rho_U & \sigma_{UW} & \sigma_U^2\rho_U & \sigma_{UW} & \cdots & \sigma_U^2\rho_U & \sigma_{UW} \\ \sigma_{UW} & \sigma_W^2\rho_W & \sigma_{UW} & \sigma_W^2\rho_W & \cdots & \sigma_{UW} & \sigma_W^2\rho_W \end{pmatrix} + \begin{pmatrix} \eta & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \nu & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \eta & 0 & \vdots & 0 & 0 \\ 0 & 0 & 0 & \nu & \vdots & 0 & 0 \\ \vdots & \vdots & \cdots & \cdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \eta & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & \nu \end{pmatrix}, \tag{2.3}$$

여기서 $\eta = \sigma_U^2(1-\rho_U)$, $\nu = \sigma_W^2(1-\rho_W)$이다. 식 (2.2)와 비교해보면 $\xi = 0$이 되어 V_i의 각 블록의 비대각 원소가 상수인 σ_{UW}로 일정함을 알 수 있다. Z_i, G, γ_i는 2.2절에서와 같고 $R_i = I_{n_i} \otimes \left(\begin{smallmatrix} \eta & 0 \\ 0 & \nu \end{smallmatrix}\right)$로 표현할 수 있다.

3. 실제 자료 분석

3.1. ARFI 자료 소개

본 자료는 선천성 담도 폐쇄증 환자인 소아청소년 32명을 대상으로 간 섬유화 진단을 위해 비침습적 섬

Table 3.1. Acoustic radiation force impulse imaging (ARFI) data

ID	Liver 1	Liver 2	Liver 3	Spleen 1	Spleen 2	Spleen 3	Spleen length
1	1.36	1.58	1.50	2.20	3.50	0.84	5.50
2	1.69	1.44	1.29	2.45	2.99	2.14	5.70
3	1.65	1.93	1.59	2.93	2.62	3.42	6.00
4	2.93	2.88	2.95	2.31	2.20	2.31	6.40
5	3.32	3.10	3.54	3.42	2.55	3.22	5.90
				⋮			

Liver 1, Liver 2, Liver 3: Three ARFI SWV measurements in liver.
Spleen1, Spleen 2, Spleen 3: Three ARFI SWV measurements in spleen.

Table 3.2. ARFI data in the format for use of SAS PROC MIXED

ID	Vtype	Replicates	Response12	Response13
1	1	1	1.36	1.36
1	2	1	2.20	5.50
1	1	2	1.58	1.58
1	2	2	3.50	-
1	1	3	1.50	1.50
1	2	3	0.84	-
2	1	1	1.69	1.69
2	2	1	2.45	5.70
2	1	2	1.44	1.44
2	2	2	2.99	-
2	1	3	1.29	1.29
2	2	3	2.14	-

Response12: SWV measurement in liver or spleen.
Response13: SWV measurement in liver or spleen length.

유화 검사인 고속 음향 복사력 임펄스(acoustic radiation force impulse imaging; ARFI)를 시행한 결과이다. 간과 비장에서 전단파 속도(shear wave velocity; SWV, m/s)를 각각 세 번씩 반복 측정하였고, 복부 초음파를 이용하여 환자의 비장 길이(cm)를 측정하였다 (Table 3.1). 본 예제의 연 구 목적은 간과 비장에서의 SWV 간, SWV와 비장 길이 간의 상관성을 보는 것이다. 여기서 간과 비장에서의 SWV 간 상관관계를 보고자 하는 경우에는 세 번씩 동일하게 반복 측정 되었으므로 2.2절의 방법을 이용할 수 있고, 비장 길이는 한 번만 측정되므로 SWV와의 상관관계를 볼 때는 2.3절의 방법을 이용하여야 한다.

G와 R_i는 SAS의 PROC MIXED를 이용하여 구현할 수 있는데, 이를 사용하기 위해서는 자료의 각 행이 각 측정치를 의미하도록 일변량 형태의 데이터 셋이 필요하다. Table 3.1의 자료를 SAS의 PROC MIXED에서 활용 가능하도록 바꾸면 Table 3.2와 같다. 여기서 ID는 각 대상자 고유 번호를, Vtype은 변수의 종류를 의미하며 이 자료에서 1은 간, 2는 비장을 의미한다. Replicates는 대상자 내에서 반복 측정 단위를, Response는 각 변수의 측정치를 의미한다. PROC MIXED의 RANDOM, REPEATED statements에 VType 변수를 사용하여 U와 W 간의 상관관계를 모형화하여 G와 R_i를 추정한다.

3.2. SWV 간의 상관분석

우선 반복 측정됨을 고려하지 않고 피어슨의 상관계수를 구하면 $\rho = 0.2071$ $(p = 0.0429)$이고, 반복 측

Table 3.3. Estimated variance-covariance matrix for the APRI data for ♯1

	Liver 1	Spleen 1	Liver 2	Spleen 2	Liver 3	Spleen 3
Liver 1	1.2749	0.1754	1.0170	0.1692	1.0170	0.1692
Spleen 1	0.1754	0.5582	0.1692	0.4118	0.1692	0.4118
Liver 2	1.0170	0.1692	1.2749	0.1754	1.0170	0.1692
Spleen 2	0.1692	0.4118	0.1754	0.5582	0.1692	0.4118
Liver 3	1.0170	0.1692	1.0170	0.1692	1.2749	0.1754
Spleen 3	0.1692	0.4118	0.1692	0.4118	0.1754	0.5582

Table 3.4. Estimated correlation matrix for the APRI data for ID ♯1

	Liver 1	Spleen 1	Liver 2	Spleen 2	Liver 3	Spleen 3
Liver 1	1.0000	0.2079	0.7977	0.2005	0.7977	0.2005
Spleen 1	0.2079	1.0000	0.2005	0.7378	0.2005	0.7378
Liver 2	0.7977	0.2005	1.0000	0.2079	0.7977	0.2005
Spleen 2	0.2005	0.7378	0.2079	1.0000	0.2005	0.7378
Liver 3	0.7977	0.2005	0.7977	0.2005	1.0000	0.2079
Spleen 3	0.2005	0.7378	0.2005	0.7378	0.2079	1.0000

정된 값들의 평균을 이용하여 상관계수를 구하면 $\rho = 0.2403$ ($p = 0.1853$)이다. 2.2절에서 제안된 반복측정됨을 고려하는 방법을 이용하여 간과 비장에서 측정된 SWV 간의 상관관계를 분석하기 위한 SAS code는 아래와 같다.

```
proc mixed;
   class ID Vtype Replicates ;
   model Response12 = Vtype / s ddfm=kr ;
   repeated Vtype / type = un subject = Replicates(ID)  r rcorr ;
   random Vtype / type = un subject = ID g gcorr v vcorr ;
run;
```

모수 추정은 PROC MIXED에서 기본적으로 사용하는 제한된 최대우도추정법(restricted maximum likelihood estimation method)을 이용하였다. DDFM = kr은 고정효과에 대한 유의성 검정 시 자유도 추정을 Kenward와 Roger (1997)가 제안한 방법을 이용하여 자유도의 분모를 계산한다는 것이다. REPEATED statement의 r, RANDOM statement의 g, v는 각각 R_i, G, V_i를, rcorr, gcorr, vcorr은 이들의 상관계수 행렬을 의미하며, SAS의 결과 창에는 SUBJECT = effect에 기반하여 첫 번째 대상자에 대한 분산-공분산 행렬 또는 상관계수 행렬이 출력된다. 각 옵션에 숫자를 지정하여 다른 대상자에 대한 각 행렬을 출력할 수도 있다 (SAS Institute, Inc., 2008).

Table 3.3은 ID가 1인 대상자의 분산-공분산 행렬(V_i)이 추정된 결과이고, Table 3.4는 이에 대한 상관계수 행렬이다. Table 3.3에서는 간과 비장에서의 SWV에 대한 분산(σ_U^2, σ_W^2)이 각각 1.2749, 0.5582로 추정됨을 알 수 있다. Table 3.4에서 같은 시점에서의 두 부위에서의 SWV 간 상관계수(ρ_{UW})는 0.2079로, 서로 다른 시점에 측정된 SWV 간 상관계수는 간에서 $\hat{\rho}_U = 0.7977$, 비장에서 $\hat{\rho}_W = 0.7378$로 추정됨을 알 수 있다. 서로 다른 시점에서의 간과 비장의 SWV 간 상관계수는 $\hat{\rho}_{UW}\hat{\delta} = 0.2005$임으로 δ는 0.9644(= 0.2005/0.2079)로 추정할 수 있다. 선형혼합모형에서의 고정효과의 회귀계수는 $\hat{\beta}_0 = 3.1527$ ($p < 0.0001$), $\hat{\beta}_1 = -0.7029$ ($p = 0.0011$)로 추정되었다. 따라서 간에서의

Table 3.5. Estimated variance-covariance matrix for the APRI data for ID ♯1

	SWV in Liver 1	Spleen length	SWV in Liver 2	SWV in Liver 3
SWV in Liver 1	1.2749	−0.1919	1.0169	1.0169
Spleen length	−0.1919	9.6730	−0.1919	−0.1919
SWV in Liver 2	1.0169	−0.1919	1.2749	1.0585
SWV in Liver 3	1.0169	−0.1919	1.0169	1.2749

Table 3.6. Estimated correlation matrix for the APRI data for ID ♯1

	SWV in Liver 1	Spleen length	SWV in Liver 2	SWV in Liver 3
SWV in Liver 1	1.0000	−0.0547	0.7977	0.7977
Spleen length	−0.0547	1.0000	−0.0547	−0.0547
SWV in Liver 2	0.7977	−0.0547	1.0000	0.7977
SWV in Liver 3	0.7977	−0.0547	0.7977	1.0000

SWV의 평균은 $3.1527 - 0.7029 = 2.4498$로, 비장에서의 SWV의 평균은 3.1527로 추정되고, 두 평균치가 같다고 할 수 있는지는 β_1에 대한 유의성 검정을 통해 알 수 있고 그 결과 $p = 0.0011$임으로 유의수준 5%하에서 유의한 차이를 보인다고 해석할 수 있다.

3.3. SWV와 비장 길이 간의 상관 분석

반복 측정됨을 고려하지 않고 피어슨의 상관계수를 구하면 $\hat{\rho} = -0.0559\,(p = 0.6010)$이고, 반복 측정된 값들의 평균을 이용하여 상관계수를 구해보면 $\hat{\rho} = -0.0600\,(p = 0.7530)$이다. 2.3절에서 제안된 반복 측정됨을 고려하는 방법을 이용하여 간에서 측정된 SWV와 비장 길이 간의 상관관계를 분석하기 위한 SAS code는 아래와 같다.

```
proc mixed;
   class ID Vtype Replicates ;
   model Response13 = Vtype / s ddfm=kr ;
   repeated Vtype / subject = Replicates(ID)  r rcorr ;
   random Vtype / type = un subject = ID g gcorr v vcorr ;
run;
```

3.2절과 다른 점은 REPEATED statement에서 type = un을 사용하지 않은 것이다. SAS에서는 기본적으로 분산성분(variance component) 구조를 가정하므로 위 프로그램을 이용하면 식 (2.3)에서와 같이 R_i의 비대각 원소가 0으로 가정되는 선형혼합모형을 적합하게 된다. Table 3.5는 ID가 1인 대상자의 분산-공분산 행렬(V_i)의 추정결과이고, Table 3.6은 이에 대한 상관계수 행렬이다. Table 3.5에서는 간에서의 SWV와 비장 길이에 대한 분산(σ_U^2, σ_W^2)이 각각 1.2749, 9.6730으로 추정됨을 알 수 있다. Table 3.6에서 간에서의 SWV와 비장 길이 간의 상관계수(ρ_{UW})는 −0.0547, 서로 다른 시점에 측정된 간에서의 SWV 간 상관계수(ρ_U)는 0.7977로 추정됨을 알 수 있다. 2.3절에서 보았듯이 SWV와 비장 길이 간의 상관계수는 SWV가 측정된 시점과 상관없이 모두 −0.0547로 동일함을 확인할 수 있다. 선형혼합모형에서의 회귀계수는 $\hat{\beta}_0 = 9.7319 (p < 0.0001)$, $\hat{\beta}_1 = -6.9221 (p < 0.0001)$로 추정되었다. 따라서 간에서의 SWV의 평균은 $9.7319 - 6.9221 = 2.4498$로, 비장 길이의 평균은 9.7319로 추정되고, 두 평균치는 유의수준 5% 하에서 유의한 차이를 보인다고 해석할 수 있다.

Table 4.1. Simulation Results

ρ_{UW}	n	replicates	Correlation using mixed effect model (SE)	Naïve Pearson correlation (SE)	Pearson correlation based on means (SE)
0.2	30	3	0.2040 (0.1422)	0.2033 (0.1415)	0.2287 (0.1704)
		5	0.2069 (0.1383)	0.2061 (0.1374)	0.2387 (0.1725)
	100	3	0.2009 (0.0809)	0.2007 (0.0808)	0.2249 (0.0955)
		5	0.2033 (0.0764)	0.2031 (0.0763)	0.2344 (0.0946)
0.6	30	3	0.5981 (0.0933)	0.5961 (0.0931)	0.6706 (0.1026)
		5	0.6002 (0.0896)	0.5978 (0.0894)	0.6916 (0.0998)
	100	3	0.5987 (0.0526)	0.5981 (0.0526)	0.6710 (0.0564)
		5	0.5995 (0.0497)	0.5988 (0.0496)	0.6905 (0.0541)
0.75	30	3	0.7463 (0.0623)	0.7437 (0.0625)	0.8378 (0.0570)
		5	0.7477 (0.0587)	0.7446 (0.0589)	0.8625 (0.0500)
	100	3	0.7484 (0.0346)	0.7476 (0.0346)	0.8393 (0.0306)
		5	0.7486 (0.0323)	0.7476 (0.0323)	0.8626 (0.0268)

4. 모의 실험

4.1. 모의 실험 방법

2절에서 소개한 방법과 피어슨의 상관계수를 활용하는 방법 간의 비교를 위해 모의 실험을 시행하였다. 3절의 ARFI 자료와 같이 반복 측정되는 변수는 두 개로 가정하였고, 식 (2.1)의 $\beta = (3, -1, 0)'$으로, 식 (2.2)의 모수들의 참값은 $\sigma_U^2 = 1.5, \sigma_W^2 = 0.5, \rho_U = 0.8, \rho_W = 0.7, \delta = 0.9, \rho_{UW} = 0.2, 0.6$ 또는 0.75로 설정하였다. 다른 모수들은 $\sigma_{UW} = \rho_{UW}\sigma_U\sigma_W, \eta = \sigma_U^2(1-\rho_U), \nu = \sigma_W^2(1-\rho_W), \xi = \sigma_{UW}(1-\delta)$임을 이용하여 계산하였다. 식 (2.2)에서 확률 변수인 γ_i, ε_i를 평균이 영벡터이고 분산-공분산 행렬이 앞에서 설정된 모수들로 구성되는 G와 R_i인 다변량 정규분포로부터 랜덤 추출하여 관측치(Y_i)를 생성하였다. 대상자 수가 30, 100인 경우에 대하여 대상자 별 반복측정 횟수를 3 또는 5로 설정하여 1000번 반복하는 모의실험을 수행하였다. 2절에서 소개한 혼합모형을 이용한 상관계수, 반복 측정됨을 고려하지 않거나 각 대상자 별 변수의 평균값을 분석 단위로 하는 피어슨의 상관계수를 추정하였다. 1000번 반복하여 얻어진 상관계수 추정치들의 평균을 구하고, 추정치들의 표준편차를 경험적 표준오차(Standard Error; SE)로 추정하였다.

4.2. 모의 실험 결과

변수 간 상관관계, 대상자 수, 대상자 별 반복측정 횟수에 따른 추정 결과가 Table 4.1에 제시되어 있다. 대상자 수나 반복측정 횟수에 상관없이 각 대상자 별 변수의 평균값을 이용하면 두 변수 간의 상관관계를 과대 추정하는 경향을 보였고 실제 상관계수가 커질수록 편향(bias)은 더 크게 나타났다. 표준오차도 다른 추정 방법에 비해 $\rho_{UW} = 0.75$인 경우에는 더 작게, 나머지 경우에는 더 크게 추정되었다. 반복 측정됨을 고려하지 않은 피어슨의 상관계수는 혼합모형을 이용하여 추정한 상관계수와 비슷하게 추정되었지만, 실제 상관계수가 커질수록 다소 과소 추정하는 경향을 보였다. 이에 비해 혼합모형을 이용하는 방법은 참값과 가장 가까운 값으로 추정하는 것으로 보인다.

5. 결론

동일한 대상자에서 반복 측정된 결과들은 보통 독립적이지 않고 상관성이 존재한다. 따라서 반복 측정

된 변수들 간의 상관관계를 볼 때에도 이를 고려해야 하며 이를 위해 선형혼합모형이 유용하게 쓰일 수 있다. 본 연구에서는 변수들의 반복측정이 동시에 된 경우와 그렇지 않은 경우로 나누어 선형혼합모형을 이용한 상관계수의 추정방법을 소개하고 ARFI 자료에 적용하기 위한 SAS code 및 분석 결과를 제시함으로써 연구자들이 쉽게 활용할 수 있도록 하였다. ARFI 자료를 분석한 결과는 혼합모형을 이용한 경우가 반복 측정됨을 고려하지 않거나 각 대상자들의 반복 측정된 값들의 평균을 이용하여 분석한 결과와 값이 크게 다르지는 않으나, 혼합모형을 이용하는 방법이 반복 측정된 자료들의 상관성을 고려할 수 있으므로 더 적절하다고 생각된다. 또한 모의실험을 통하여 반복 측정된 변수들 간 상관관계를 구할 때에는 혼합모형을 이용하는 방법이 편향(bias)을 줄일 수 있는 방법임을 확인하였다.

본 연구에서 제안된 방법을 활용할 때에 주의할 점은, 선형혼합모형은 정규성을 가정하므로 이에 대한 충분한 검토가 필요하며 만약 자료가 정규분포를 따르지 않는 경우에는 로그 변환 등을 통하여 정규성을 만족하도록 한 후 선형혼합모형을 이용하여야 한다는 점이다. 이러한 변환 후에도 정규성 가정이 부적절한 경우에는, 반복측정을 고려하지 않는 경우에 스피어만의 상관 계수를 구하는 것과 같이 비모수적인 방법을 고려해볼 수 있지만 이에 대해서는 추가적인 연구가 필요하다. 또한 반복 측정된 변수들 간의 분산-공분산 행렬의 구조를 어떻게 가정하느냐에 따라 결과가 달라질 수 있으므로 주의해서 결정해야 한다.

References

Bland, J. M. and Altman, D. G. (1994). Correlation, regression, and repeated data, *British Medical Journal*, **308**, 896.

Bland, J. M. and Altman, D. G. (1995a). Calculating correlation coefficients with repeated observations: Part 1-Correlation within subjects, *British Medical Journal*, **310**, 446.

Bland, J. M. and Altman, D. G. (1995b). Calculating correlation coefficients with repeated observations: Part 2-Correlation between subjects, *British Medical Journal*, **310**, 633.

Hamlett, A., Ryan, L., Serrano-Trespalacios, P. and Wolfinger, R. (2003). Mixed models for assessing correlation in the presence of replication, *Journal of the Air & Waste Management Association*, **53**, 442–450.

Hamlett, A., Ryan, L. and Wolfinger, R. (2004). On the use of PROC MIXED to estimate correlation in the presence of repeated measures, SAS Users Group International, *Proceedings of the Statistics and Data Analysis Section*, 1–7.

Kenward, M. G. and Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood, *Biometrics*, 983–997.

Lam, M., Webb, K. A. and O'Donnell, D. E. (1999). Correlation between two variables in repeated measures, *American Statistical Association, Proceedings of the Biometric Section*, 213–218.

Roy, A. (2006). Estimating correlation coefficient between two variables with repeated observations using mixed effects model, *Biometrical Journal*, **48**, 286–301.

SAS Institute Inc. (2008). SAS/STAT 9.2 User's Guide. Cary, NC: SAS Institute Inc.

혼합모형을 이용한 반복 측정된 변수들 간의 상관분석

한경화[a] · 정인경[b,1]

[a]연세대학교 의과대학 연세의생명연구원; 영상의학교실; 방사선의과학연구소,
[b]연세대학교 의과대학 의학정보통계학과

(2015년 3월 13일 접수, 2015년 4월 2일 수정, 2015년 4월 6일 채택)

요 약

생명과학 또는 의학 연구에서는 반복 측정된 변수들 간의 상관 관계를 보고자 하는 경우가 발생한다. 반복 측정된 것을 고려하지 않으면 상관관계를 과소 추정하는 경향이 나타나므로 이를 고려해야 하며, 선형혼합모형의 분산-공분산 행렬을 이용하여 상관관계를 추정할 수 있다. 본 연구에서는 변수들의 반복 측정이 동시에 된 경우와 그렇지 않은 경우로 나누어 혼합모형을 이용한 상관계수의 추정방법을 소개한다. 고속 음향 복사력 임필스 영상(acoustic radiation force impulse imaging; ARFI)으로 간과 비장에서 각각 세 번씩 전단파 속도를 반복 측정하고 복부 초음파 검사로 비장 길이를 측정한 자료에서 전단파 속도와 비장 길이 간의 상관 관계를 분석하기 위해 본 논문에서 소개한 방법들을 적용하였고 SAS의 PROC MIXED를 이용하는 방법을 제시하였다.

주요용어: 선형혼합모형, 반복측정, 상관관계, PROC MIXED.

[1]교신저자: (120-752) 서울특별시 서대문구 연세로 50-1, 연세대학교 의과대학 의학정보통계학과.
E-mail: ijung@yuhs.ac