2000 Proceedings of the Autumn Conference, Korean Statistical Society

이변량 반복측정자료에서 가중일치상관계수의 추정

강 보 경[1], 김 규 성[2]

요 약

이변량 반복측정자료에서 Chinchilli 등(1996)이 제안한 가중일치상관계수는 두 변수의 일치성을 나타내는 측도이다. 기존에 제안된 가중일치상관계수 추정법은 변동효과 및 측정오차의 분산성분을 각각 최소제곱법으로 비편향 추정하여 구하는 것이다. 본 연구에서는 반복측정자료의 주변 우도함수를 설정한 후, 우도함수에 기초한 분산성분을 구하여 가중일치상관계수를 추정하는 방법을 제안한다. 이때, 각 분산성분은 유사/의사 우도함수 및 사후 분포에서 반복시행을 통하여 구해진다.

주요용어 : 가중일치상관계수; 사후분포; 이변량 혼합모형; 유사/의사 우도함수;

1. 서론

어떤 실험에서 동일한 개체에 두 가지 처리를 실시하는 문제를 고려해보자. 일반적인 관심은 두 처리간에 차이가 있는지 여부이겠지만, 경우에 따라서는 두 처리 방법의 일치성에 관심이 모아질 때도 있다. 예를 들어 두 명의 임상병리학자가 환자들의 정신상태를 일정기간 진단하여 환자들을 몇 가지의 범주로 분류할 때, 두 임상병리학자의 분류결과의 일치여부가 중요한 관심사일 수 있다. 이러한 경우 두 처리 결과의 일치도를 말해줄 수 있는 객관적인 판단기준이 필요해진다.

두 연속형 반응변수의 관련성은 피어슨 상관계수, 쌍체 t-검정, 변동계수, 혹은 급내상관계수 등을 이용하여 판단할 수 있으나 두 변수의 일치도를 말하기에는 부족함이 있다. (Lin, 1989; Zar, 1996). 두 변수가 일치하는 상황은 두 변수의 관측치의 분포가 원점을 지나며 기울기가 1이 되는 경우이다. 따라서 관측치가 이 직선으로부터 얼마나 벗어났는지 측정함으로써 두 변수의 일치성을 표현할 수 있다. Lin(1989)은 이러한 생각에 기초하여 두 연속형 변수 (X, Y)의 일치성을 측정하는 일치상관계수(Concordance Correlation Coefficient) ρ_c를 제안하고,

$$\rho_c = \frac{2\sigma_{XY}}{\sigma_{XX} + \sigma_{YY} + (\mu_X - \mu_Y)^2} \tag{1}$$

ρ_c의 점근적 성질을 규명하였다. 여기서 μ_X, (μ_Y)는 변수 X (Y)의 평균, σ_{XX}, (σ_{YY})는 변수 X(Y)의 분산이며 σ_{XY}는 변수 X, Y의 공분산이다. 일치상관계수는 피어슨 상관계수와는 달리 두 변수의 분산, 공분산은 물론 두 변수의 평균도 연관성 판정에 반영하는 장점이 있다.

반복측정자료(repeated measured data)는 동일 실험단위를 대상으로 시간 또는 공간을 달리하여 반복 측정한 자료이다. 한 예로 의학 및 보건학 분야에서 많이 행해지는 다시점 연구

1) (130-743) 서울특별시 동대문구 전농동 90, 서울시립대학교 전산통계학과, 대학원
2) (130-743) 서울특별시 동대문구 전농동 90, 서울시립대학교 전산통계학과, 조교수.
E-mail : kskim@uoscc.uos.ac.kr

(longitudinal study)에서는 동일 실험단위를 대상으로 시간의 경과에 따라 반복 측정함으로써 시간의 경과에 따른 변화를 연구하는 것이다. Chinchilli 등(1996)은 일치상관계수를 반복측정 자료에 적용하여 가중일치상관계수(Weighted Concordance Correlation Coefficient)를 제안하고, 가중일치상관계수를 추정하는 방법을 소개하였다. 이들이 제안한 가중일치상관계수 추정 방법은 각 분산성분을 독립적으로 추정하여 대입하는 것인데, 이 때 분산성분 추정에 이들이 이용한 방법은 적률추정법이었다. 일반적인 혼합모형(mixed model)에서 적률추정법은 계산이 간단하여 사용하기 쉬운 장점이 있는 반면, 동시에 저효율 등의 단점이 알려져 있다. 따라서 보다 효율적인 분산성분 추정법을 이용한다면 개선된 가중일치상관계수 추정법을 구할 수 있을 것이다.

본 연구에서는 이 점에 착안하여 보다 효율적인 가중일치상관계수 추정법을 제안하고자 한다. 일반적인 혼합모형에서 실험개체 내 분산이 다를 때 분산성분을 추정하는 방법은 Lin 등(1997)에 의하여 소개되었다. 이 방법에서는 변량 효과 및 측정오차의 분산성분에 분포가정을 하여 반복측정자료의 주변 우도함수를 설정한 후, 주변우도함수를 최대화하는 분산성분을 찾는 것이다. 이때 주변우도함수의 형태를 구체적으로 유도하기가 어렵기 때문에, 각 분산성분에 대한 유사/의사 우도함수 및 사후분포를 유도하여 이에 근거한 근을 반복적으로 구하게 된다. 본 연구에서는 Lin 등이 제안한 방법을 가중일치상관계수의 분산성분을 추정하는데 이용할 것이다. 제2절에서는 Chinchilli 등이 제안한 가중일치상관계수를 소개하고 적률추정법을 이용한 가중일치상관계수 추정법을 소개한다. 제3절에서는 이변량 반복측정자료의 주변 우도함수를 설정하여 각 분산성분을 추정한 후, 가중일치상관계수를 추정하는 방법을 제안하며, 제4절에서는 연구 결과를 요약하고, 향 후 연구과제를 토의한다.

2. 최소제곱법을 이용한 가중일치상관계수의 추정

이변량 반복측정 실험에서 실험개체 $i(i=1,\ldots,n)$에 대한 변수 X와 Y의 자료 벡터를 각각 $\mathbf{X}_i=[x_{i1},x_{i2},\cdots,x_{im_{X_i}}]'$와 $\mathbf{Y}_i=[y_{i1},y_{i2},\cdots,y_{im_{Y_i}}]'$라 하자. 여기서 m_{X_i}와 m_{Y_i}는 각 실험개체 i에서의 변수별 반복수이며, 쌍으로 관측된 경우는 반복수가 동일하며, $m_{X_i}=m_{Y_i}$, 쌍으로 관측되지 않았을 때는 반복수가 동일하지 않을 수도 있다. 이 자료에 대하여 다음과 같은 혼합모형을 고려하자. (Chinchilli 등 ;1996).

$$\mathbf{X}_i=\mathbf{P}_{\mathbf{X}_i}(\boldsymbol{\beta}_{\mathbf{X}}+\mathbf{u}_{\mathbf{X}_i})+\boldsymbol{\varepsilon}_{\mathbf{X}_i}\quad,\qquad \mathbf{Y}_i=\mathbf{P}_{\mathbf{Y}_i}(\boldsymbol{\beta}_{\mathbf{Y}}+\mathbf{u}_{\mathbf{Y}_i})+\boldsymbol{\varepsilon}_{\mathbf{Y}_i} \tag{2}$$

여기서 $\mathbf{P}_{\mathbf{X}_i},\mathbf{P}_{\mathbf{Y}_i}$는 각각 $m_{X_i}\times p, m_{Y_i}\times p$차원의 기지의 설계행렬, $\boldsymbol{\beta}_X$, $\boldsymbol{\beta}_Y$는 $p\times 1$차원의 미지의 고정효과벡터, $\mathbf{u}_{\mathbf{X}_i},\mathbf{u}_{\mathbf{Y}_i}$는 $p\times 1$차원의 변동효과벡터, 그리고 $\boldsymbol{\varepsilon}_{\mathbf{X}_i}$와 $\boldsymbol{\varepsilon}_{\mathbf{Y}_i}$는 측정오차벡터로서 $m_{X_i}\times 1, m_{Y_i}\times 1$ 차원을 갖는다. 실험개체 i에 대한 변동효과벡터 $\mathbf{u}_i=(\mathbf{u}_{\mathbf{X}_i}', \mathbf{u}_{\mathbf{Y}_i}')'$와 측정오차벡터 $\boldsymbol{\varepsilon}_i=[\boldsymbol{\varepsilon}_{X_i}'\boldsymbol{\varepsilon}_{Y_i}']'$는 서로 독립이라 가정하자. 그리고 실험개체간의 관련성을 표현하는 나타내는 변동효과벡터는 다음을 가정하자.

$$\mathbf{u}_i \sim \left(\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\mathbf{D}_{XX} & \mathbf{D}_{XY}\\ \mathbf{D}_{YX} & \mathbf{D}_{YY}\end{pmatrix}\right),\quad i=1,\ldots,n \tag{3}$$

또한 각 실험개체 내에서 측정오차간의 관계는 쌍으로 실험할 경우와 그렇지 않은 경우로 나누어 생각할 수 있는데, 쌍으로 관측된 실험이 아닌 경우는 실험 개체 i 내 측정오차는 서로 독립이라고 가정할 수 있으며,

$$\boldsymbol{\varepsilon}_i \sim \left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma_{XX_i} \mathbf{I}_{m_{X_i}} & \mathbf{0} \\ \mathbf{0} & \sigma_{YY_i} \mathbf{I}_{m_{Y_i}} \end{pmatrix} \right) \tag{4}$$

쌍으로 관측된 경우($m_i = m_{X_i} = m_{Y_i}$)는, 실험개체 i 내 측정오차는 쌍으로 인한 상관관계가 있다고 볼 수 있다. 즉,

$$\boldsymbol{\varepsilon}_i \sim \left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma_{XX_i} \mathbf{I}_{m_i} & \sigma_{XY_i} \mathbf{I}_{m_i} \\ \sigma_{YX_i} \mathbf{I}_{m_i} & \sigma_{YY_i} \mathbf{I}_{m_i} \end{pmatrix} \right) \tag{5}$$

위에서 언급된 반복측정자료는 반복수가 다르고 실험개체와 실험반복을 구분해야 하기 때문에 일치상관계수를 직접 적용하기는 어렵다. 이 문제를 해결하는 방법으로 Chinchilli 등(1996)은 각 개체 내에서 두 변수의 최소제곱 회귀계수의 추정치를 구하고 회귀계수 추정치의 분산을 구한 후, 이렇게 구한 n개의 분산을 가중 평균한 가중일치상관계수를 제안하였다. 즉 개체 내에서 관측치를 다음과 같이 변환한 후,

$$\mathbf{X}_i^* = (\mathbf{P}_{\mathbf{X}_i}'\mathbf{P}_{\mathbf{X}_i})^{-1}\mathbf{P}_{\mathbf{X}_i}'\mathbf{X}_i, \quad \mathbf{Y}_i^* = (\mathbf{P}_{\mathbf{Y}_i}'\mathbf{P}_{\mathbf{Y}_i})^{-1}\mathbf{P}_{\mathbf{Y}_i}'\mathbf{Y}_i \tag{6}$$

개체별 일치상관계수, $\rho_{c,i}$,를 구한다.

$$\begin{aligned} \rho_{c,i} &= \frac{1}{p}\sum_{j=1}^{p} \text{concordance correlation}\,(\mathrm{X}_{ij}^*, \mathrm{Y}_{ij}^*) \\ &= \frac{1}{p}\sum_{j=1}^{p} \frac{2(\Delta_{XY_i})_{jj}}{(\Delta_{XX_i})_{jj} + (\Delta_{YY_i})_{jj} + \{(\mu_{X_i} - \mu_{Y_i})(\mu_{X_i} - \mu_{Y_i})'\}_{jj}} \end{aligned} \tag{7}$$

여기서 $\Delta_{XX_i} = Var(\mathbf{X}_i^*), \Delta_{XY_i} = COV(\mathbf{X}_i^*, \mathbf{Y}_i^*)$ 와 $\Delta_{YY_i} = VAR(\mathbf{Y}_i^*)$는 $p \times p$행렬이고 이 행렬의 대각원소는 각각 $(\Delta_{XX_i})_{jj}, (\Delta_{XY_i})_{jj}, (\Delta_{YY_i})_{jj}$이며, $\mu_{X_i} = E(\mathbf{X}_i^*), \mu_{Y_i} = \mathrm{E}(\mathbf{Y}_i^*)$이다. 그리고 이렇게 구한 n개의 개체 내 일치상관계수를 가중 평균하여 가중일치상관계수를 정의한다.

$$\rho_c = \left(\sum_{i=1}^{n} w_i\right)^{-1}\left(\sum_{i=1}^{n} w_i \rho_{c,i}\right) \tag{8}$$

여기서 가중치 w_i는

$$w_i = \frac{1}{(1+\sigma_{XX_i})(1+\sigma_{YY_i}) - \sigma_{XY_i}^2} \tag{9}$$

이다.

위의 (8)에서 정의한 가중일치상관계수를 추정하는 방법으로 Chinchilli 등은 변동효과와 측정오차의 분산성분을 최소제곱추정치로 대치하여 구하는 방법을 제안하였다. 즉, $m_{X_i} > p$, $m_{Y_i} > p$ 일 때, σ_{XX_i}에 대한 불편추정량은 다음과 같으며, ($\hat{\sigma}_{YY_i}$도 유사하게 구함),

$$\hat{\sigma}_{XX_i} = \frac{1}{m_{X_i} - p}\mathbf{X}_i'\left\{\mathbf{I}_{m_{X_i}} - \mathbf{P}_{\mathbf{X}_i}(\mathbf{P}_{\mathbf{X}_i}'\mathbf{P}_{\mathbf{X}_i})^{-1}\mathbf{P}_{\mathbf{X}_i}'\right\}\mathbf{X}_i \tag{10}$$

이를 이용하여 구한 변동효과의 분산의 불편추정량은 다음과 같다. ($\hat{\mathbf{D}}_{\mathbf{YY}}$, $\hat{\mathbf{D}}_{\mathbf{XY}}$도 유사).

$$\hat{\mathbf{D}}_{\mathrm{XX}}=\frac{1}{\mathrm{n}-1}\sum_{i=1}^{n}(\mathbf{X}_{\mathbf{i}}^{*}-\overline{\mathbf{X}}_{\cdot}^{*})(\mathbf{X}_{\mathbf{i}}^{*}-\overline{\mathbf{X}}_{\cdot}^{*})-\frac{1}{n}\sum_{i=1}^{n}\hat{\sigma}_{XX_i}(\mathbf{P}_{\mathbf{X}_i}{}'\mathbf{P}_{\mathbf{X}_i})^{-1} \tag{11}$$

여기서 $\overline{\mathbf{X}}_{\cdot}^{*}=\sum_{i=1}^{n}\mathbf{X}_{\mathbf{i}}^{*}/\mathrm{n}$, $\overline{\mathbf{Y}}_{\cdot}^{*}=\sum_{i=1}^{n}\mathbf{Y}_{\mathbf{i}}^{*}/\mathrm{n}$. 또한 μ_X, μ_Y는 회귀계수의 일반화 최소제곱 추정치를 이용하여 추정한다.

$$\begin{bmatrix}\widehat{\boldsymbol{\beta}_X}\\ \widehat{\boldsymbol{\beta}_Y}\end{bmatrix}=\left\{\sum_{i=1}^{n}\begin{bmatrix}\hat{\Delta}_{XX_i} & \hat{\Delta}_{XY_i}\\ \hat{\Delta}_{YX_i} & \hat{\Delta}_{YY_i}\end{bmatrix}^{-1}\right\}^{-1}\left\{\sum_{i=1}^{n}\begin{bmatrix}\hat{\Delta}_{XX_i} & \hat{\Delta}_{XY_i}\\ \hat{\Delta}_{YX_i} & \hat{\Delta}_{YY_i}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{X}_{\mathbf{i}}^{*}\\ \mathbf{Y}_{\mathbf{i}}^{*}\end{bmatrix}\right\} \tag{12}$$

이 방법은 반복측정자료 모형에 분포가정은 하지 않고 단지 2차 적률까지만 가정을 한 후, 기본적으로 최소제곱법을 이용하여 분산성분의 비편향 추정량을 구하는 방법이다. 이 방법에서 제공하는 가중일치상관계수의 형태는 완성된 형태로 수식이 유도되고, 그 형태가 알기 쉬워서 이용하기 용이한 장점이 있다. 그러나 반복측정자료를 일반 혼합모형으로 확장하여 생각하면 위에서 제안한 최소제곱법/적률법을 이용한 방법은 개선의 여지가 있을 것으로 보인다. 즉, 변동효과의 분산 및 측정오차의 분산에 적당한 분포가정을 한 후, 우도함수를 이용하여 최대우도 추정법으로 분산성분을 추정하는 방법을 고려해 볼 수 있을 것이다. 다음절에서는 이 방법을 이변량 반복측정자료에 적용하여 가중일치상관계수를 추정하는 방법을 소개한다.

3. 우도함수를 이용한 가중일치상관계수의 추정

Lin, et al.(1997)은 변동효과와 측정오차에 정규가정을 하고, 개체내 측정오차의 분산성분에는 역감마분포를 가정하여 고정 효과 및 분산성분을 추정하는 추정방정식을 제안하였다. 본 연구에서는 이들의 결과를 이변량 반복측정자료에 적용하여 분산성분 추정에 이용한다.

변동효과 $\mathbf{u}_i$와 측정오차 $\boldsymbol{\varepsilon}_i$의 평균과 분산은 앞에서와 같으며 추가로 정규분포를 가정한다. 또한 실험개체 내 측정오차를 변동분산으로 간주하여 역감마분포를 고려한다. 즉, $1/\sigma_{XX_i}$과 $1/\sigma_{YY_i}$은 서로 독립이고 다음의 분포를 따른다고 가정하자.

$$\frac{1}{\sigma_{XX_i}}\sim \mathrm{G}\left(\frac{1+2\delta_X}{\delta_X},\frac{\delta_X}{(1+\delta_X)\sigma_{XX_i^0}}\right),\ \frac{1}{\sigma_{YY_i}}\sim \mathrm{G}\left(\frac{1+2\delta_Y}{\delta_Y},\frac{\delta_Y}{(1+\delta_Y)\sigma_{YY_i^0}}\right) \tag{13}$$

단, $\delta_X\geq 0, \delta_Y\geq 0, \sigma_{XX_i^0}\geq 0, \sigma_{YY_i^0}\geq 0$이다. 또한, $\mathbf{w}_i$가 개체 내 분산에 영향을 미치는 개체수준에 대한 기지의 $s\times 1$차원벡터이고, $\boldsymbol{\eta}_X, \boldsymbol{\eta}_Y$가 $s\times 1$차원의 미지의 모수 벡터일 때, 평균 $\sigma_{XX_i^0}, \sigma_{YY_i^0}$에 대해 $\log(\sigma_{XX_i^0})=\mathbf{w}_i{}'\ \boldsymbol{\eta}_X$, $\log(\sigma_{YY_i^0})=\mathbf{w}_i{}'\ \boldsymbol{\eta}_Y$를 가정하자. $\boldsymbol{\theta}$가 $\mathbf{D}$의 분산성분과 ρ_e로 이루어진 벡터이고 $\boldsymbol{\eta}=(\boldsymbol{\eta}_X{}'.\boldsymbol{\eta}_Y{}')'$, $\boldsymbol{\delta}=(\delta_X,\delta_Y)'$, $\phi_{X_i}=1/\sigma_{XX_i}$이고, $\phi_{Y_i}=1/\sigma_{YY_i}$ 일 때, $(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\delta})$에 대한 주변 우도함수는

$$L(\beta, \theta, \eta, \delta) = \prod_{i=1}^{n} \int \exp\{l(\mathbf{X}_i, Y_i; \mathbf{u}_i, \phi_{X_i}, \phi_{Y_i}) + l(\mathbf{u}_i) + l(\phi_{X_i}) + l(\phi_{Y_i})\} d\mathbf{u}_i d\phi_{X_i} d\phi_{Y_i} \quad (14)$$

위의 식 (14)를 최대화하는 최대우도추정량을 직접 구하기는 쉽지 않아 보인다. 따라서 Lin 등이 일변량 혼합모형에서 이용한 방법을 위의 모형 (14)에 적용하여, 유사우도(quasi-likelihood), 의사우도(pseudo-likelihood)와 적률 방법을 이용하면, 모수 및 분산성분을 추정할 수 있다. 그리고 추정된 분산성분을 가중일치상관계수의 분산성분에 대체하여 가중일치상계수를 추정한다.

4. 토의 및 결론

이변량 반복측정실험에서 두 변수의 일치성을 판정하는 측도로써 가중일치상관계수는 효과적이다. 기존에 제안된 가중일치상관계수 추정법은 변동효과 및 측정오차의 분산성분을 각각 최소제곱법으로 비편향 추정하여 구하는 것인데 반해, 본 연구에서는 반복측정자료의 주변 우도함수를 설정한 후, 우도함수에 기초하여 가중일치상관계수를 추정하는 방법이다. 본 연구에서 제안한 방법은 변동효과 및 측정오차의 분산성분에 추가적인 분포가정을 해야 한다는 제약점이 있고 계산량이 많다는 단점은 있으나, 일반적인 혼합모형에 적용할 수 있다는 장점이 있다. 요즘의 형편으로 보아 계산 알고리즘의 단순화보다는 추정방법이 일반화가 더 중요해 보이며, 그러한 측면에서 본 연구에서 제안한 방법은 일반적인 혼합모형에까지 가중일치상관계수를 적용했다는데 의미가 있다 하겠다.

참고문헌

[1] Chinchilli, V.M., Martel, J.K., Kumanyika, S. and Lloyd, T. (1996). A weighted Concordance Correlation Coefficient for Repeated Measurement Designs. Biometrics 52, 341-353.

[2] Chinchilli, V.M., Esinhart, J.D. & Miller, W.G. (1995). Partial Likelihood Analysis of Within-Unit Variance in Repeated Measurement Experiments. Biometrics 51, 205-216.

[3] Davidian, M. and Carroll, R.J. (1987). Variance function estimation, Journal of the American Statistical Association 82, 1079-1091.

[4] Lin, L.I-K.(1989). A concordance correlation coefficient to evaluate to reproducibility. Biometrics 45, 255-268.

[5] Lin, X., Raz, J. & Harlow, S.D. (1997). Linear Mixed Models with Heterogeneous Within-Cluster Variances. Biometrics 53, 910-923.

[6] McCullagh, P. and Nelder, J.A. (1989). Generalized Linear Models, 2nd edition. Chapman and Hall, London.

[7] Rochon, J. (1996). Analyzing Bivariate Repeated Measures for Discrete and Continuous Outcome Variables. Biometrics 52, 740-750.

[8] Searle, S.R., Casella, G., and McCulloch, C.E. (1992). Variance Component. John Wiley & Sons, New York.

[9] Wedderburn, R.W.M.(1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. Biometrika 61, 439-447.

[10] Zeller, A. (1962). An efficient method of estimating seemingly unrelated regression equations and tests for aggregation bias. Journal of the American Statistical Association 57, 348-368.