

패널 및 경시적 자료에서 결측치 처리 방법

이근백

성균관대학교 통계학과

2024년 9월 26일

Email: keunbaik@skku.edu

경시적 자료분석 R 활용



- 경북대학교 통계학과 학사
- 경북대학교 통계학과 석사
- 미국 University of Florida 통계학 박사
- 현) 성균관대학교 통계학과 교수

- 연속형 경시적 자료와 병주형 경시적 자료로 나누어 각 자료를 분석하기 위한 통계적 모형 소개
- R 코드를 이용한 다양한 형태의 자료분석 제시 및 실제 자료에 적용한 분석 결과를 각 장에 수록
- 통계학과 박사과정의 연구에 도움이 되도록 이론적 내용 일부 포함

- | | |
|-----------------------|----------------------|
| 1장 서론 | 8장 경시적 이진자료분석을 위한 |
| 2장 기본개념 | 가능도 방법 |
| 3장 탐색적 자료분석 | 9장 과다산포 및 상관성이 있는 자료 |
| 4장 연속형 경시적 자료분석을 위한 | 10장 일반화 선형혼합모형 |
| 선형모형 | 11장 전이모형 |
| 5장 연속형 경시적 자료분석을 위한 | 12장 주변모형, 임의효과와 결측치 |
| 선형혼합모형 | 전이모형의 비교 |
| 6장 범주형 자료분석을 위한 추정방법적 | 13장 경시적 자료에서의 결측치 |
| 7장 주변모형 | 14장 다변량 경시적 자료분석 |

◀ ◻ ▶ ◀ ◻ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡ ≡ ↺ 🔍 ↻

1 서론

2 결측치 처리

- Missing Data Mechanism
- Ignorable missingness
- MNAR

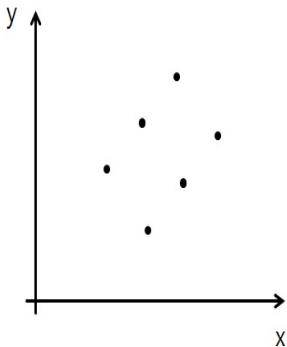
- 경시적 자료(longitudinal data)
 - 각 개체(subject)에서 일정 시간 동안 반복 측정된 자료이다.
 - 다른 개체에서 나온 결과들은 독립이며, 같은 개체에서 나온 결과들은 서로 상관관계가 존재한다.
 - 평균이 시간에 따라 개별적으로 어떻게 변화하는지 추정하는 것과 이 변화가 개체들 사이의 이질성(heterogeneity)에 영향을 미치는 요소들을 조사하는 것과 관련이 있다.
- 경시적 연구의 예: 시계열 연구, 패널분석 연구, 전향적 연구 (임상시험)

경시적 자료의 특성

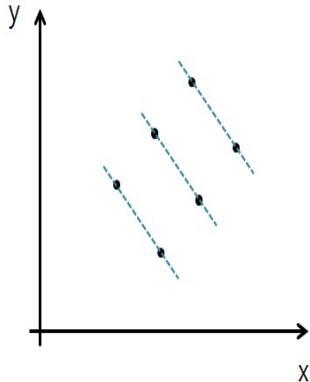
- 횡단면 자료(cross-sectional data)
 - 각 개체에서 한 시점의 응답변수 값만 관측되는 자료
 - 응답변수의 값들은 서로 독립임을 가정함
- 경시적 자료
 - 각 개체들이 일정 시간 동안 반복 측정된 자료
 - 개체 내에서 반복 측정된 자료들은 일반적으로 양의 상관관계를 가지며 이로 인하여 분석이 복잡하게 됨
 - 개체 내에서 반복된 자료는 누락으로 인하여 불균형이 일어나고 이것으로 인하여 분석이 복잡하게 됨

경시적 자료분석의 장점

- 경시적 연구는 개별적인 개체의 패턴에 대한 정보를 제공함
- 동일한 변수를 같은 개체에서 반복적으로 측정하므로 측정의 신뢰도가 높음



(a) Marginal correlation



(b) Conditional correlation

Figure: Marginal vs. conditional correlations

경시적 자료분석을 위한 모형

- 연속형 경시적 자료(Continuous longitudinal data)분석
 - 선형모형(Linear models)
 - 선형혼합모형(Linear mixed models)
- 범주형 경시적 자료(Categorical longitudinal data)분석
 - 주변모형(Marginal models)
 - Generalized estimating equation(GEE)을 이용한 방법
 - 주변화 모형(Marginalized models)
 - 일반화선형혼합모형(Generalized Linear mixed models)
- 결측치(missing data)

경시적 자료

- Notation

- y_{ij} : i 번째 개체의 j 번째 응답변수 ($i = 1, \dots, m; j = 1, \dots, n$)
- $y_i = (y_{i1}, \dots, y_{in})^T$: i 번째 응답벡터
- $x_{ij} = (1, x_{ij1}, \dots, x_{ij,p-1})^T$: 설명변수 벡터
예: $x_{ij} = (1, Gender_i, Age_{ij})^T$

결측자료의 특성

- 경시적 연구에서는 필연적으로 결측자료(missing data)가 발생할 수 있다.
- 결측의 이유(결측자료 매커니즘: missing data mechanism (MDM))는 자료분석에서 고려되어야 한다.
- 결측치는 정보의 손실을 의미하고, 잠재적으로 자료분석에서 심각한 편향을 가져올 수 있다(Daniels와 Hogan, 2008).

결측치를 처리하기 위한 간단한 방법들

- 완전 개체(complete case: CC)을 이용한 방법
- 마지막 관찰값(last observation carried forward: LOCF)을 이용한 방법
 - CC나 LOCF는 심각한 효율성 상실을 가져올 수 있다.
- 다중 대체 (multiple imputation) 방법 (Schafer and Graham, 2002)
- 미완성 (incomplete) 자료를 이용한 분석

Full data model

- y_{ij} = 개체 i 의 j 번째 응답변수; x_{ij} 는 해당 응답변수에 상응하는 설명변수 ($i = 1, \dots, m; j = 1, \dots, n$).
- $y_i = (y_{i1}, \dots, y_{in})^T$ 는 개체 i 의 완비자료(complete data) 응답변수 벡터이다.
 y_i 는 관측된 반응변수 자료($y_i^{(o)}$)와 결측된 반응변수 자료($y_i^{(m)}$)으로 구분된다:

$$y_i = (y_i^{(o)}, y_i^{(m)})$$

- $R_i = (R_{i1}, \dots, R_{in})^T$ 는 관측여부를 나타내는 지시변수 벡터이다.

$$R_{ij} = \begin{cases} 1, & y_{ij} \text{가 관측된 경우;} \\ 0, & y_{ij} \text{가 결측된 경우.} \end{cases}$$

Full data model

- $F_i = (Y_i, R_i, X_i)$ 에서

For full data: $P(y_i, r_i | x_i, \omega)$,

For data response model: $P(y_i | x_i, \theta)$.

이 둘의 관계는 다음과 같다.

$$P(y_i | x_i, \theta(\omega)) = \sum_{r_i \in R} P(y_i, r_i | x_i, \omega),$$

여기서 R 은 R_i 의 표본공간이다.

- Y_i 와 R_i 의 pdf는

$$\begin{aligned} P(y_i, r_i | x_i, \omega) &= P(y_i^{(o)}, y_i^{(m)}, r_i | x_i, \omega) \\ &= P(y_i^{(m)} | y_i^{(o)}, r_i, x_i, \omega) P(y_i^{(o)}, r_i | x_i, \omega). \end{aligned}$$

따라서 untestable assumptions가 필요함.

결측자료 메커니즘 (Little과 Rubin, 2002)

- Full data model을 다음과 같이 분해할 수 있다.

$$P(y_i, r_i | \omega) = P(y_i | \theta(\omega)) P(r_i | y_i, \psi(\omega)),$$

여기서

$$P(r_i | y_i, \psi(\omega)) = P(r_i | y_i^{(o)}, y_i^{(m)}, \psi(\omega))$$

를 결측자료 메커니즘이다.

결측자료 메커니즘

- 완전임의결측(missing completely at random, MCAR): 결측이 발생할 확률이 응답변수에 의존하지 않는 경우

$$P(R_i|y_i, x_i, \psi) = P(R_i|x_i, \psi).$$

- 예를 들면, 2년에 걸쳐서 2번 학생들의 성적을 조사하는 연구에서 첫해에는 1000명의 학생에 대한 성적을 조사하였고, 두 번째 해에는 예산의 삭감으로 인하여 500명의 학생에 대한 성적만을 조사하는 경우 500명의 선택이 임의로 이루어졌다면, 이 결측치는 MCAR이다.
- 일반적인 GEE를 이용한 방법을 사용할 수 있다.

결측자료 메커니즘

- 임의결측(missing at random, MAR): 결측이 발생할 확률이 관측값($y_i^{(o)}$)에만 의존하고, 결측치에는 의존하는 않은 경우

$$P(R_i|y_i, x_i, \psi) = P(R_i|y_i^{(o)}, x_i, \psi).$$

- 예를 들면, 앞서 제시한 2년간 2번의 성적조사에서 첫 번째 성적이 낮은 학생들이 두 번째 성적을 위한 테스트에 덜 참여하는 경향있다. 이것은 첫 번째 성적이 낮은 학생들이 두 번째 테스트에서 결시하게 되어 결측될 확률이 높다는 의미이다.
- MAR하에서 응답변수의 모형과 결측치의 모형의 모수들이 완전히 분리가 될 때, MAR 가정은 무시가능성(ignorability)이라고 한다. 일반적으로 대부분의 통계 패키지는 이 무시가능성 하에서 분석되고 있다. 일반적인 GEE는 MAR하에서는 사용될 수 없다.
- MAR하에서 결측치를 대체하는 방법이 있다. 예를 들면, 다중대체(multiple imputation) 방법이 있다.

결측자료 메커니즘

- 비임의결측(missing not at random, MNAR): 결측이 발생할 확률이 결측치 ($y_i^{(m)}$)에 의존하는 경우

$$P(R_i|y_i, x_i, \psi) = P(R_i|y_i^{(o)}, y^{(m)}, x_i, \psi).$$

즉, $y_i^{(o)}$ 가 주어진 상태에서 결측값은 $y^{(m)}$ 에만 의존한다.

$$P(R_i|y_i^{(o)}, y^{(m)}, x_i, \psi) \neq P(R_i|y_i^{(o)}, y^{(m)'} , x_i, \psi)$$

여기서 $y^{(m)} \neq y^{(m)'}.$

- 예를 들면, 금연연구에서 이번 달에 흡연하여 금연클리닉에 나타나지 않은 경우가 발생한다. 이 경우 이번 달의 응답변수로 인하여 이번 달에 결측치가 발생한 것이다. 이 경우 이 응답변수는 MNAR이다.
- MNAR하에서 결측치를 위한 모형이 따로 필요하게 된다.

무시가능 결측하에서 자료분석(MAR 경우)

- 다음의 2가지 조건을 만족하면 결측자료 메커니즘을 무시가능 결측 (ignorable missingness: IM)이라고 한다.
 - 결측자료 메커니즘이 MAR이다.
 - 완전자료모수 (full data parameter) ω 는 $\omega = (\theta, \psi)$ 로 분해될 수 있다. 여기서 θ 는 $P(y_i|\theta)$ 의 모수이며, ψ 는 $P(r_i|y_i, \psi)$ 의 모수이다. 즉, θ 와 ψ 는 서로 별개(disjoint)이다.
- θ 의 추론은 관측된 자료의 가능도 함수에 비례하며, ψ 에는 의존하지 않는다.

$$P(y_i^{(o)}, r_i | \theta, \psi) = P(y_i^{(o)} | \theta) P(r_i | y_i^{(o)}, \psi).$$

따라서 무시가능 결측 조건 하에서 θ 의 추론을 하기 위해서 결측자료 메커니즘 $P(r_i | y_i^{(o)}, \psi)$ 은 무시할 수 있다.

- θ 의 추론은 관측된 자료 응답모형(observed-data response model)에 비례하며, 그 모형은 다음과 같다.

$$P(y_i^{(o)} | \theta) = \int P(y_i^{(o)}, y_i^{(m)} | \theta) dy_i^{(m)}.$$

무시가능 결측하에서 자료분석(MAR 경우)

- 결측치 $y_i^{(m)}$ 은 완전자료 응답모형만을 이용하여 관측값 $y_i^{(o)}$ 로부터 대체 (imputation) 또는 외삽법(extrapolation)으로 대체될 수 있다.

$$\begin{aligned}
 P(y_i^{(m)} | y_i^{(o)}, r_i, \omega) &= \frac{P(y_i^{(m)}, y_i^{(o)} | \omega)}{P(y_i^{(o)}, r_i | \omega)} = \frac{P(r_i | y_i^{(o)}, y_i^{(m)}, \psi) P(y_i^{(o)}, y_i^{(m)} | \theta)}{P(r_i | y_i^{(o)}, \psi) P(y_i^{(o)} | \theta)} \\
 &= \frac{P(y_i^{(o)}, y_i^{(m)} | \theta)}{P(y_i^{(o)} | \theta)} \quad (\text{by separation of } \omega = (\theta, \psi)) \\
 &= P(y_i^{(m)} | y_i^{(o)}, \theta). \quad (\text{by MAR})
 \end{aligned}$$

무시가능 결측하에서 자료분석(MAR 경우)

- 무시가능 결측 하에서 경시적 자료분석
 - 가능도(likelihood)에 기초한 방법: 자료모형(data model)과 결측모형이 분리되므로 자료모형만을 이용하여 모수를 추정한다. 이외에도 EM-알고리즘을 이용한 방법, 다중대체(multiple imputation)방법이 있다.
 - 역확률 가중(inverse probability weighting: IPW)을 이용한 GEE (IPW-GEE) 방법: GEE 추정량의 편향을 보정하기 위한 자료의 관측확률의 역수를 가중값으로 이용하고, 그 개체에 가중치를 부여하여 관측된 자료를 확대하여 모수를 추정한다.

역확률 가중(IPW)을 이용한 GEE (IPW-GEE) 방법

- IPW-GEE는 주로 단조 결측자료(monotone missing data) 패턴에서 사용되며, 그 식은 다음과 같다.

$$\text{GEE:} \quad \sum_{i=1}^m D_i^T V_i^{-1} (y_i - \mu_i(\beta)) = 0,$$

$$\text{IPW-GEE:} \quad \sum_{i=1}^m D_i^T V_i^{-1} R_i W_i (y_i - \mu_i(\beta)) = 0,$$

여기서 small

$$y_i = (y_{i1}, \dots, y_{in_i})^T, \quad \mu_i(\beta) = (\mu_{i1}(\beta), \dots, \mu_{in_i}(\beta))^T,$$

$$D_i = \frac{\partial \mu_i}{\partial \beta} = (D_i(j, k)), \quad D_i(j, k) = \frac{\partial \mu_{ij}}{\partial \beta_k}, \quad j = 1, \dots, n_i, k = 0, \dots, p-1,$$

$$V_i = V_i(\beta, \alpha, \phi) = C_i^{1/2} R_i(\alpha) C_i^{1/2},$$

$$R_i = \text{diag}\{R_{i1}, \dots, R_{in_i}\}, \quad W_i = \text{diag}\{w_{i1}, \dots, w_{in_i}\},$$

$$w_{ij} = 1/P(R_{ij} = 1 | y_i^{(o)}) \text{ (가중치)}.$$

정신분열증자료 분석(Gibbons과 Hedeker, 1994)

- 경시적 순서화 자료를 경시적 이진자료로 변환하여 분석하였다.
IMPS가 4이상으로 응답하면 1로, 나머지는 0으로 변환하여 분석하였다.
- $y_{ij} = i$ 번째 개체의 j 번째 이진 응답변수이다($j = 1, \dots, 4$).
- 총 381명의 환자 중 293명은 항정신성 약물을 투여받은 그룹이며, 93명은 위약을 받은 그룹이다.
- 설명변수는 항정신성 약물의 복용여부(복용한 경우, $TxDrug_{ij} = 1$; 복용을 안 한 경우, $TxDrug_{ij} = 0$), 변환한 주차($SWeek_{ij} = \sqrt{Week_{ij}}$)를 고려한다.
- 결측치는 MAR을 가정한다.

정신분열증자료 분석(Gibbons과 Hedeker, 1994)

- 가정된 모형:

$$\begin{aligned}\text{logit}P(y_{ij} = 1|x_{ij}) &= \beta_0 + \beta_1 SWeek_{ij} + \beta_2 TxDrug_i + \beta_3 SWeek_{ij} \times TxDrug_i, \\ \text{logit}P(R_{ij} = 1) &= \gamma_0 + \gamma_1 TxDrug_i + \gamma_2 SWeek_{ij} \\ &+ \gamma_3 y_{i,j-1} + \gamma_4 y_{i,j-2} 1_{(j>2)} + \gamma_5 y_{i,j-3} 1_{j>3}.\end{aligned}$$

- IPW-GEE를 적합한 모형:

$$\begin{aligned}\text{logit}\hat{P}(y_{ij} = 1|x_{ij}) &= 3.173^* - 1.169^* SWeek_{ij} - 0.359 TxDrug_i \\ &- 0.248 SWeek_{ij} \times TxDrug_i, \\ \text{logit}\hat{P}(R_{ij} = 1) &= 6.913^* + 0.847^* TxDrug_i - 2.893^* SWeek_{ij} \\ &+ 0.753^* y_{i,j-1} - 0.674 y_{i,j-2} 1_{(j>2)} + 1.728^* y_{i,j-3} 1_{j>3}.\end{aligned}$$

- GEE를 적합한 모형:

$$\begin{aligned}\text{logit}\hat{P}(y_{ij} = 1|x_{ij}) &= 3.107^* - 1.100^* SWeek_{ij} - 0.258 TxDrug_i \\ &- 0.345 SWeek_{ij} \times TxDrug_i.\end{aligned}$$

IPW-GEE의 해석

- IPW-GEE와 GEE의 모수의 적합된 값들이 다를 수 있다.
- 항정신성 약물($TxDrug_i = 1$)을 투여한 그룹일수록, 주 차가 적을수록 중도탈락이 덜 일어남을 알 수 있다. 또한 lag가 3 차이 나는 응답변수의 값이 클수록 중도탈락이 덜 일어남을 알 수 있다.
- 항정신성 약물을 투여한 그룹에서 변환한 주 차($\sqrt{Week_{ij}}$)가 1 증가함에 따라 로그 오즈는 -1.417 ($= -1.169 - 0.248$) 변함을 알 수 있다. 또한 위약을 투여한 그룹에서 로그 오즈는 -1.169 변함을 알 수 있다.

MNAR하에서 완전자료 모형

- 완전자료모형은 외삽법(extrapolation) 분해에 의해서 다음과 같이 분해될 수 있다

$$\begin{aligned}P(y_i, r_i | \omega) &= P(y_i^{(o)}, r_i | \omega_o) P(y_i^{(m)} | y_i^{(o)}, r_i, \omega_E) \\ &= \text{observed data model} \times \text{extrapolation model}\end{aligned}$$

여기서 $\omega = (\omega_o, \omega_E)$ 이며, $P(y_i^{(o)}, r_i | \omega_o)$ 는 인식가능하다. 그리고 $P(y_i^{(m)} | y_i^{(o)}, r_i, \omega_E)$ 는 일반적으로 인식이 불가능하다.

MNAR하에서 완전자료 모형

- 민감모수(sensitivity parameter): 아래의 조건을 만족하는 재모수화된 $\xi(\omega) = (\xi_S, \xi_M)$ 이 존재한다면, ξ_S 를 민감모수(sensitivity parameter)라고 한다.
 - ξ_S 는 ω_E 의 일정하지 않은 함수(nonconstant function)이다.
 - 관측자료 가능도 $P(y_i^{(o)}, r_i | \xi_S, \xi_M)$ 는 ξ_S 의 함수로서 상수이다(constant as a function of ξ_S).
 - 고정된 ξ_S 에 대해서 관측자료 가능도는 ξ_M 의 일정하지 않은 함수이다.
- 의미: ξ_S 는 외삽법 분해의 의한 분포를 구분하고(index), 모든 ξ_S 의 값에 대해서 모형의 적합은 동일하다. 그리고 ξ_S 가 고정된 상태에서 완전 자료 모형(full data model)은 식별 가능하다(identified).
- 민감모수는 관측된 자료에 의해서 식별 가능하지 않다 (cannot be identified by observed data).

MNAR하에서 완전자료 모형

- 응답변수 벡터(y_i)와 결측여부를 나타내는 지시변수 벡터(R_i)의 결합분포인 완전자료모형을 3가지 모형으로 분해한다.

- 선택모형(selection model, SM)(Kenward, 1998; Little과 Rubin, 1987, 2002):

$$P(y_i, R_i) = P(R_i|y_i)P(y_i).$$

- 선택모형 $P(R_i|y_i)$ 의 분포의 가정에 의해서 관심 모형인 $P(y_i)$ 의 모수추정이 민감하게 반응한다.
- 선택모형에서 민감도 분석을 반드시 시행할 필요가 있다.
- 패턴 혼합모형(pattern mixture model, PMM)(Little, 1993):

$$P(y_i, R_i) = P(y_i|R_i)P(R_i).$$

- 결측자료 패턴 별로 다른 분포를 기술할 수 있다.

- 이 분포는 종종 관측된 자료로 식별할 수 없는 모수를 명시할 수 있고, 여기서 식별 불가능한 모수에 대한 민감도 분석을 수행할 수 있다.
- 공유모수모형(shared parameter model, SPM)(Ten Have, 1998):

$$P(y_i, R_i) = \int P(y_i|b_i)P(R_i|b_i)f(b_i)db_i.$$

- 일반적으로 기저의 결측자료 메카니즘을 이해하기가 어렵다.
- 적분을 요구한다.
- MNAR하에서의 GEE: MAR하에서 IPW-GEE를 이용한 방법과 유사한 방법을 이용하여 분석할 수 있다.

선택모형

$$P(y_i, r_i | \omega) = P(r_i | y_i, \psi) P(y_i | \theta).$$

$$Y \sim F(y; \theta),$$

$$R|Y = y \sim \text{Ber}(\pi(y)),$$

여기서 $g(\pi(y)) = \psi_0 + \psi_1 y$

- $F(y; \theta)$ 가 특별한 모수적 모형 (예를 들면, 정규분포)일때 위의 모형은 전적으로 인식가능하다.
- 이 경우에 민감모수의 의해서 일반적으로 모수화 할 수 없다. 즉, 민감도 분석을 할 수 없다.
- 만일 $P(y^{(o)})$ 또는 $P(y)$ 가 비모수적이라면, 민감도 분석을 할 수 있다.

선택모형

만약 각 개체 별로 2번의 응답변수 $Y_i = (Y_{i1}, Y_{i2})$ 를 가정하자. R_i 는 Y_{i2} 의 관측여부를 나타내는 지시변수이다. 이 경우 MDM은 다음과 같다.

$$\text{logit}P(R_i = 1|y_i, \psi) = \psi_0 + \psi_1 y_{i1} + \psi_2 y_{i2},$$

여기서 $R_i = 1$ 은 Y_{i2} 가 관측된 경우이다.

- 완전히 기술된 모수적 선택모형(fully specified parametric selection model)은 모든 모수들이 인식가능 하며, 이 경우 민감모수가 존재하지 않으므로 민감도 분석을 할 수 없다.
- 반모수적 선택모형(semiparametric selection model)에서는 MDM은 모수적인 모형을 관측자료의 분포는 비모수적으로 가정한다. 이 경우 MDM을 위한 모형에서 인식이 가능하지 않은 민감모수가 존재한다. 이 경우 민감도 분석을 수행할 수 있다.

패턴 혼합모형

$$\begin{aligned}
 P(y_i, r_i | x_i, \omega) &= P(y_i | r_i, x_i, \omega) P(r_i | x_i, \omega), \\
 \Rightarrow P(y_i | x_i, \omega) &= \sum_{r \in \mathbb{R}} P(y_i | r_i, x_i, \omega) P(r_i | x_i, \omega),
 \end{aligned}$$

여기서 $\mathbb{R}$ 는 R_i 의 표본공간이다.

$Y_{i1}, \dots, Y_{iJ}$ 는 i 번째 개체의 완전자료 응답변수이며, $S_i \in \{1, \dots, J\}$ 는 결측자료의 패턴을 의미한다. 결측치들이 단조 결측자료(monotone missing data) 패턴을 가정한다. 그러면 $S_i = \sum_{j=1}^J R_{ij}$ 이다. $P_k(y_{ij} | y_{i1}, \dots, y_{ij-1})$ 는 $S = k$ 의 결측자료 패턴에서 조건부 확률이다. 완전 응답변수 $y_i = (y_{i1}, \dots, y_{iJ})$ 의 $S = k$ 의 결측자료 패턴에서 결합확률은 다음과 같다.

$$P_k(y_i) = P_k(y_{i1}, \dots, y_{ik}) P_k(y_{i,k+1}, \dots, y_{iJ}),$$

여기서 $P_k(y_{i1}, \dots, y_{ik})$ 는 인식이 가능하고, $P_k(y_{i,k+1}, \dots, y_{iJ})$ 는 인식이 불가능하다.

패턴 혼합모형

$J = 4$ 인 경우,

	$j = 2$	$j = 3$	$j = 4$
$S = 1$	?	?	?
$S = 2$	$P_2(y_2 y_1)$	?	?
$S = 3$	$P_3(y_2 y_1)$	$P_3(y_{i3} y_{i1}, y_{i2})$	?
$S = 4$	$P_4(y_2 y_1)$	$P_4(y_3 y_1, y_2)$	$P_4(y_4 y_1, y_2, y_3)$

여기서 ?에 해당되는 확률은 인식이 불가능한 조건부 확률들이다.

패턴 혼합모형에서 identification strategies

- MAR for discrete-time pattern mixture models under monotone dropout
- Interior family constraints
 - Complete case missing value constraints
 - Nearest-neighbor or adjacent-pattern constraints
 - Available case missing value constraints
- Non-future dependence missing value restrictions

Complete case missing value constraints

	$j = 2$	$j = 3$	$j = 4$
$S = 1$	$P_{\geq 2}(y_{i2} y_{i1})$	$P_{\geq 3}(y_{i3} y_{i1}, y_{i2})$	$P_4(y_{i4} y_{i1}, y_{i2}, y_{i3})$
$S = 2$	$P_2(y_{i2} y_{i1})$	$P_{\geq 3}(y_{i3} y_{i1}, y_{i2})$	$P_4(y_{i4} y_{i1}, y_{i2}, y_{i3})$
$S = 3$	$P_3(y_{i2} y_{i1})$	$P_3(y_{i3} y_{i1}, y_{i2})$	$P_4(y_{i4} y_{i1}, y_{i2}, y_{i3})$
$S = 4$	$P_4(y_{i2} y_{i1})$	$P_4(y_{i3} y_{i1}, y_{i2})$	$P_4(y_{i4} y_{i1}, y_{i2}, y_{i3})$

$P_1(y_{i2}|y_{i1}) = P_{\geq 2}(y_{i2}|y_{i1})$ 을 위해서 $S \geq 2$ 의 결측자료 패턴의 모든 개체를 이용할 수 있다.

성장 호르몬 연구(Growth Hormone Study)

- 성인들의 근육강화에 대한 4개의 그룹으로 이루어진 경시적 연구
- 160명을 4개의 그룹으로 임의로 배정하였다.
 - Growth hormone (G)
 - Growth hormone + exercise (EG)
 - Placebo (P)
 - Placebo + exercise (EP)
- 응답변수: Quadriceps strength (QS) recorded at baseline, 3 months, 6 months
- 각 그룹 별로 QS의 평균을 비교한다.

성장 호르몬 연구(Growth Hormone Study)

자료요약

Treatment	k	n_k	Month		
			0	6	12
EP	1	7	65 (32)		
	2	2	87 (52)	86 (51)	
	3	31	65 (24)	81 (25)	73 (21)
	All	40	66 (26)	82 (26)	73 (21)
EG	1	12	58 (26)		
	2	4	57 (15)	68 (26)	
	3	22	78 (24)	90 (32)	88 (32)
	All	38	69 (25)	87 (32)	88 (32)

성장 호르몬 연구

- EG와 EP 그룹 간의 평균을 비교하고자 함.
- 패턴 혼합모형을 이용한 분석
 - MAR
 - MNAR at fixed points (민감도 분석)
- 자료구조
 - (Y_1, Y_2, Y_3) =quad strength measures at month 0, 6, 12
 - Z = 그룹 (0=EP, 1=EG)
 - (R_1, R_2, R_3) =response indicators (1=observed, 0=missing)
 - $S = \sum_j R_j$ =missing data pattern

성장 호르몬 연구

- $P_s(y_1)$ 은 인식가능하다 ($s = 1, 2, 3$).
- $P_s(y_2|y_1)$ 과 $P_s(y_3|y_1, y_2)$ 은 다음과 같다.

	$j = 2$	$j = 3$
$S = 1$	$P_1(y_2 y_1)$	$P_1(y_3 y_1, y_2)$
$S = 2$	$P_2(y_2 y_1)$	$P_2(y_3 y_1, y_2)$
$S = 3$	$P_3(y_2 y_1)$	$P_3(y_3 y_1, y_2)$

Non-identified components in red.

Model specification for $Y|S$

- $Y_1|S$ 의 모형:

$$Y_1|S = 1 \sim N(\mu_1, \sigma_1^2),$$

$$Y_1|S = 2 \sim N(\mu_2, \sigma_2^2),$$

$$Y_1|S = 3 \sim N(\mu_3, \sigma_3^2).$$

- $Y_2|Y_1, S$ 의 모형:

$$Y_2|Y_1, S = 1 \sim N(\alpha_{2.1}^{(0)} + \beta_{2.1}Y_1, \sigma_{2.1}^2),$$

$$Y_2|Y_1, S \geq 2 \sim N(\alpha_{2.1}^{(1)} + \beta_{2.1}Y_1, \sigma_{2.1}^2).$$

- $Y_3|Y_1, Y_2, S$ 의 모형:

$$Y_3|Y_1, Y_2, S = 1, 2 \sim N(\alpha_{3.21}^{(0)} + \beta_{3.21}Y_1 + \gamma_{3.21}Y_2, \sigma_{3.21}^2),$$

$$Y_3|Y_1, Y_2, S = 3 \sim N(\alpha_{3.21}^{(1)} + \beta_{3.21}Y_1 + \gamma_{3.21}Y_2, \sigma_{3.21}^2).$$

Model for S

패턴 지시변수(S)는 다항분포를 가진다.

$$S \sim \text{Multi}(\phi),$$

여기서 $\phi = (\phi_1, \phi_2, \phi_3)$ 이며 $\phi_1 + \phi_2 + \phi_3 = 1$.

Parameterize departures from MAR

- Model parameters in terms of (ξ_M, ξ_S) :

$$\begin{aligned}\xi_S &= \alpha_{2 \cdot 1}^{(0)}, \alpha_{3 \cdot 21}^{(0)} \\ \xi_M &= \begin{cases} \mu_1, \sigma_1^2, \mu_2, \sigma_2^2, \mu_3, \sigma_3^2, \\ \alpha_{2 \cdot 1}^{(1)}, \beta_{2 \cdot 1}, \sigma_{2 \cdot 1}^2, \\ \alpha_{3 \cdot 21}^{(1)}, \beta_{3 \cdot 21}, \gamma_{3 \cdot 21}, \sigma_{3 \cdot 21}^2, \\ \phi_1, \phi_2, \phi_3. \end{cases}\end{aligned}$$

- 인식가능한 모수와 비인식 모수, 그리고 민감모수를 관련을 보기 위해서 함수 h 를 도입한다.
- $\alpha_{2 \cdot 1}^{(0)}$ 와 $\alpha_{3 \cdot 21}^{(0)}$ 에 대해서,

$$\begin{aligned}\alpha_{2 \cdot 1}^{(0)} &= h_1(\xi_M, \Delta) = \alpha_{2 \cdot 1}^{(1)} + \Delta_1, \\ \alpha_{3 \cdot 21}^{(0)} &= \alpha_{3 \cdot 21}^{(1)} + \Delta_2.\end{aligned}$$

MAR induced by setting $\Delta_1 = \Delta_2 = 0$.

Interpretation of sensitivity parameter

Assume $\Delta_1 = \Delta_2 = \Delta$.

$$\Delta_1 = E(Y_2|Y_1, S = 1) - E(Y_2|Y_1, S \geq 2),$$

$$\Delta_2 = E(Y_3|Y_1, Y_2, S = 1, 2) - E(Y_3|Y_1, Y_2, S = 3)$$

Full-data mean at $j = 2$ is

$$E_{\Delta}(Y_2) = \sum_{s=1}^3 \phi_s E_{\Delta}(Y_2|S = s).$$

Effect of departures from MAR quantified by

$$E_{\Delta}(Y_2) - E_{\Delta=0}(Y_2) = \phi_1 \Delta.$$

Hence, the contribution of the sensitivity parameter to the shift in μ_2 is proportional to the fraction dropping out at $S = 1$.

For $j = 3$,

$$E_{\Delta}(Y_3) - E_{\Delta=0}(Y_3) = \Delta\{(\phi_1 + \phi_2) + \phi_1\beta_{3\cdot 21}\}.$$

Sensitivity analysis

- Specify a range $\mathbb{D}$ for Δ .
- Relies on contextual interpretation.
- Example: $\mathbb{D} = [-20, 0]$
 - Conditional mean for dropouts at t is 0 to 20 points lower than for non-dropouts.
- Equivalent to repeating analysis over a set of point of Δ . item Yields contour plots that can be useful.

Growth hormone data analysis

Define $\theta = E(Y_3|Z = 1) - E(Y_3|Z = 0)$.

- Sensitivity analysis: Contour of θ .

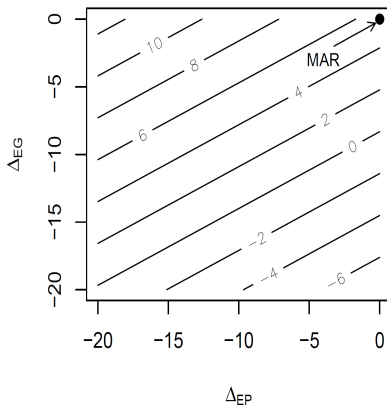


Figure: Contour plot of θ

Growth hormone data analysis

Mean (S.D) for QS at each time point, stratified by treatment group.

MVN=multivariate normal distribution assumed for joint distribution of responses; MAR=missing at random constraints; MNAR= Δ assumed different by treatment group.

Trt	Month	MVN	Pattern Mixture Models	
			MAR	MNAR
EP	0	65 (4.2)	66 (4.6)	66 (4.6)
	6	81 (4.4)	82 (4.7)	80 (4.9)
	12	73 (3.7)	73 (4.0)	69 (4.6)
EG	0	69 (4.2)	69 (4.4)	69 (4.4)
	6	81 (6.0)	81 (6.4)	78 (6.8)
	12	78 (6.3)	78 (6.7)	72 (8.1)
Diff. at 12 month		5.7 (7.3)	5.4 (7.8)	2.6 (9.3)

Growth hormone data analysis

- MAR analysis shows some evidence that GH increases QS (mena=5.7, SD=7.3).
- MNAR analysis shows smaller effect than MAR analysis (mean=2.6, SD=9.3).
- MNAR assumption has greater effect on EG arm.
 - Mean at month 12 changes from 78 to 72.
 - Largely due to higher dropout rate.

Summary: Growth hormone data analysis

- General PMM has large number of sensitivity parameters.
 - Dimension needs to be reduced.
 - Use qualitative considerations.
- Must decide on a range $\mathbb{D}$ for sensitivity parameters.
 - Choice can be subjective.
 - Can calibrate using variance (no SE) from MAR model.
 - Want to ensure $Y^{(m)}$ does not get extrapolated beyond reasonable range.
- Use contour plots.

Reference I

- 이근백 (2022). 경시적 자료분석 R활용. 자유아카데미.
- Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 125-134.
- Daniels, M.J. and Hogan, J.W. (2008). *Missing Data In Longitudinal Studies: Strategies for Bayesian modeling and Sensitivity Analysis*. Chapman & Hall/CRC, Boca Raton, FL.
- Diggle, P.J., Heagerty, P., Liang, K.-Y. and Zeger, S.L. (2002). *Analysis of Longitudinal Data*. Oxford, Clarendon Press.
- Gibbons, R. D. and Hedeker, D. (1994). Application of random-effects probit regression models. *Journal of Consulting and Clinical Psychology*, **62**, 285-296.
- Kenward, M.G. (1998). Selection models for repeated measurements with nonrandom dropout: an illustration of sensitivity. *Statistics in Medicine*, 17, 2723-2732.

Reference II

- Liang, K.-Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13-22.
- Little, R.J. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data, 2nd Ed.*, Wiley: New York.
- Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, **57**, 120-125.
- Patterson, H.D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545-554.
- Ten Have T. R., Kunselman A. R., Pulkstenis E. P., Landis J R. (1998). Mixed effect logistic regression models for longitudinal binary response data with informative drop out. *Biometrics*, **54**, 367-383.
- Tierny, L. and Kadane, J.B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81**, 82-86.

통계는 의미없는 숫자처럼 보이지만 사람에 관한 것이다.
'스토리가 있는 통계학'
by Andrew Vickers

Thank you!