

---

# Mixed Normal & Nonnormal Data with Missing Values, Ignoring the Missingness Mechanism

## General Location Model \_ with unrestricted & restricted condition

ESC 2024 Fall Session 4주차  
발제자: 권도현, 지민구



# Intro : Mixture Distribution

---

**혼합 분포**: 두 개 이상의 여러 개의 하위 분포가 결합된 형태. 본래 여러 그룹에 속했던 데이터를 한 곳에 모았거나, 아니면 데이터를 수집했더니 그 데이터가 여러 개의 그룹으로 나뉘어질 수 있는 상황

$$\text{형태: } f(x) = w_1 * N(\mu_1, \sigma_1) + w_2 * N(\mu_2, \sigma_2)$$

이 두 그룹이 선형결합되었다고 본다면 분산은 각각 하위 분포의 분산에 가중치가 곱한 걸 더해줄 것이다.  
그러나 혼합 분포 가정에서는 두 하위 분포가 mix 되었다고 보기에, 각각의 하위 그룹이 전체 평균과 가지는 차이를 분산에 반영함으로써 전체 분포의 분산을 구성한다. → 단순한 선형 결합과 다름

$$\sigma_{all}^2 = w_1(\mu_1 - \mu_{all})^2 + w_2(\mu_2 - \mu_{all})^2$$

# Mixture Distribution – Example

---

한 아파트 단지에서 11월 한 달간의 난방 사용 실태를 조사하는데, 특정 동들이 햇빛을 잘 받지 못해서 월별 난방 사용 시간이 다른 동들에 비해 월등히 높은 상황이다. 그리고 그러한 세대가 전체의 30%라고 한다면..

하위 그룹  $X_1 \sim N(160, 60^2)$

하위 그룹  $X_2 \sim N(70, 20^2)$

그렇다면 평균은 선형 결합되었을 때와 비슷하게  $0.3*160 + 0.7* 70 = 97$  과 같이 계산되겠지만,  
분산은 다음과 같이 계산된다.

따라서 혼합 분포에서의 분산은 단순 선형 결합을 가정했을 때의 분산에 추가적인 **평균 차이로 인한 분산을 더해준다는 개념**을 가지고 있다.

# Mixture Distribution – Connection to GLOM

---

혼합분포는 사실 하위 분포가 서로 다른 분포를 따를 때도 적용 가능하다. (예:  $X_1$ 은 정규 분포,  $X_2$ 는 포아송 분포를 따름) 또, 하위 분포의 개수도 3개 이상일 수도 있다.

실생활에서 우리는 전체 데이터가 여러 개의 하위 분포로 나뉘는 것을 많이 볼 수 있고, 이는 저번 주에 다룬 범주형 변수의 관점에서 바라볼 수도 있을 것이다. 그리고, 범주형 변수가 2개 이상이라면 이렇게 나뉘어지는 하위 분포의 개수는 더욱 많을 것이다.

→ 따라서 우리는 단순히 데이터가 혼합된 형태의 하위 분포보다는, 범주형 변수의 특성을 고려해 더 세부적인 그룹으로 나누어서 데이터를 바라보는 모델이 필요할 것이다.

# General Location Model (GLOM)

일반 위치 모델 : 연속형 변수와 범주형 변수가 혼합된 데이터, 다시 말해 연속형 변수가 범주형 변수로 나뉘어지는 Cell 요소 안의 값으로 들어가는 형태의 데이터에서 쓰일 수 있는 모델이다.

기본적인 구조: V개의 범주형 변수의 확률 분포 + 각 범주별 존재하는 K개의 연속형 변수의 조건부 분포

교과서에서의 설명: 연속형 변수  $x_i$ 와 범주형 변수  $w_i$ 에 대해 그 분포 ( $w_i$ )를,  $w_i$ 의 주변 분포 (marginal distribution) 과  $w_i$ 가 주어졌을 때  $x_i$ 의 조건부 분포로 구성한다.



아까의 예시를 조금 더 확장해보자면..

아파트 단지내 난방 이용 실태를 아파트 방향과 집 평수 별로 조사해본다고 한다면

$x_i$  = (난방 사용 시간, 난방 사용 금액)

$w_i$ =(아파트 방향, 집 평수) \_ 여기서 아파트 방향: 남향, 동향, 북향 / 집 평수: 35평, 45평, 60평

→ 2-way contingency table,  $3 \times 3 = 9$ 개의 수준 (cell)

정의에 적용하면 → 예를 들어 (남향, 35평) 이라는 특정  $w_{ic}$ 가 주어졌을 때의  $x_i$  = (난방 사용 일수, 난방 사용 금액)

$w_i$	35평	45평	60평
남향	$x_i$	$x_i$	$x_i$
동향	$x_i$	$x_i$	$x_i$
북향	$x_i$	$x_i$	$x_i$

# General Location Model – 주요 특징

---

이 예시를 통해 일반 위치 모델의 주요 특징을 살펴본다면

먼저 모두 9개의  $w_i$ 는 독립항등분포 (i.i.d) 가정을 따르고 각 cell에 속할  $\pi_1$ 부터  $\pi_9$  까지 총 9개의 서로 다른 확률이 만들어진다. 그리고 아파트 내에서 9개 외에 다른 수준에 속할 수는 없으므로 확률의 총합은 1이 된다.

$$\Pr(w_1 = \text{남향}, 35\text{평}) = \pi_1$$

$$\Pr(w_2 = \text{남향}, 45\text{평}) = \pi_2$$

...

$$\Pr(w_9 = \text{북향}, 60\text{평}) = \pi_9, \sum \pi = 1$$

그리고 특정  $w_{ic}$ 가 주어졌을 때의 조사된  $x_i = (\text{난방 사용 시간}, \text{난방 사용 금액})$  은 다변량 정규 분포를 따른다는 것이 일반 위치모델의 기본적인 가정이다.

→  $(x_i \mid w_i = U_c) \sim N_2(\mu_c, \Omega)$  (여기서는 K=2). 여기서 모든  $x_i$ 는 같은 공분산 행렬을 가지는 것으로 가정한다.

$$\text{모수의 개수: } C-1 + KC + \frac{1}{2}K(K+1) \quad \theta = (\Pi, \Gamma, \Omega)$$

$C-1$  = 확률에 관한 모수의 개수:  $C$ 개의 셀에 대해 자기 자신을 제외한  $C-1$  개 셀에 대한 확률만 알면 자기 자신은 자동적으로 정해지므로  $C-1$  개의 셀 확률만 알면 된다.

$KC$  = 평균에 관한 모수의 수 : 총  $C$ 개 cell에서 각각 연속 변수의 개수  $K$ 개 만큼의 평균이 구해지므로 총  $KC$ 개의 평균이 구해진다.

$\frac{1}{2}K(K+1)$  : 공분산 행렬에서 나오는 모수의 수:  $K * K$  공분산행렬에서 주대각선은  $K$ 개의 분산으로 이루어지고, 나머지 원소는 공분산에 해당하므로 다음과 같이 계산된다.

# General Location Model – 주요 장점

---

데이터가 일반 위치 모형을 따른다고 가정했을 때의 주요 특징은

- i) 범주형 변수  $w$  가 없다면 손쉽게 연속형 변수만 존재하는 다변량 정규 모형으로 축소가 가능하고 이는 11.2에서 배운 다변량 정규 데이터에 해당하는 알고리즘으로 결측치를 해결해볼 수 있다는 것을 의미한다.
- ii) 연속형 변수  $x$  가 없다면 범주형 변수만 존재하므로 Contingency Table 을 구성할 수 있고, 결측치가 있는 경우 partially classified contingency table을 적용할 수 있다.
- iii) 모든 셀에서 공분산 행렬이 같다고 추정됨으로써, 해석이 보다 용이해지고 필요한 매개변수의 수가 줄어들고, 과적합 문제를 방지할 수도 있다.
- iv) 범주형 변수  $Y$ 가 0과 1 사이 값을 갖는 binary variable일 때  $P(Y=1)$  은  $\frac{e^L}{1+e^L}$  과 같이 구해질 수 있고, 이때  $L$ 은 다른 연속형 변수들의 선형 결합으로 표현된다. (Details in 14.4)

# General Location Model – Complete Data

---

Loglikelihood:

$$\begin{aligned}\ell(\Gamma, \Omega, \Pi) &= \sum_{i=1}^n \ln f(x_i | w_i, \Gamma, \Omega) + \sum_{i=1}^n \ln f(w_i | \Pi) \\ &= h(\Omega) - \frac{1}{2} \text{tr} \left( \Omega^{-1} \sum_{i=1}^n x_i^T x_i \right) + \text{tr} \Omega^{-1} \Gamma \left( \sum_{i=1}^n w_i^T x_i \right) \\ &\quad + \sum_{c=1}^C \left[ \left( \sum_{i=1}^n w_{ic} \right) \left( \ln \pi_c - \frac{1}{2} \mu_c^T \Omega^{-1} \mu_c^T \right) \right],\end{aligned}$$

Parameters:

$$\hat{\Pi} = n^{-1} \sum w_i,$$

$$\hat{\Gamma} = \left( \sum x_i^T w_i \right) \left( \sum w_i^T w_i \right)^{-1},$$

$$\hat{\Omega} = n^{-1} \sum (x_i - w_i \hat{\Gamma})^T (x_i - w_i \hat{\Gamma}),$$

# General Location Model – Complete Data Loglikelihood

---

$$\ell(\Gamma, \Omega, \Pi) = \sum_{i=1}^n \ln f(x_i | w_i, \Gamma, \Omega) + \sum_{i=1}^n \ln f(w_i | \Pi)$$

$w_i$ 에 대해 주어진  $x_i$ 의 조건부  
분포에 대한 로그 가능도의 합  
(main part)

특정한 범주 조합  $w_i$ 가 나올 확  
률에 로그를 취한 것의 합

# General Location Model – Complete Data Loglikelihood

$$= h(\Omega) - \frac{1}{2} \text{tr} \left( \Omega^{-1} \sum_{i=1}^n x_i^T x_i \right) + \text{tr} \Omega^{-1} \Gamma \left( \sum_{i=1}^n w_i^T x_i \right)$$
$$+ \sum_{c=1}^C \left[ \left( \sum_{i=1}^n w_{ic} \right) \left( \ln \pi_c - \frac{1}{2} \mu_c^T \Omega^{-1} \mu_c \right) \right],$$

$$h(\Omega) = -\frac{1}{2}n \{ K \ln(2\pi) + \ln |\Omega| \}$$

로그 가능도에서 **다면량 정규 분포의 상수항을** 나타냄  
→ 차원 K와 사이즈 n을 고려하고,  $\Omega$ 의 행렬식을 통해  
공분산 행렬의 확산 정도를 보여준다.

$$\ell(\Gamma, \Omega, \Pi) = \sum_{i=1}^n \ln f(x_i | w_i, \Gamma, \Omega) + \sum_{i=1}^n \ln f(w_i | \Pi)$$

- 각 범주 조합  $w_i$ 에 대한  $x_i$ 의 합에  $\Omega$ 의 역행렬과 평균벡터를 써움 → 평균을 조절
- 마찬가지로  $\text{tr}()$ 을 통해 스칼라로 변환

- $x_i$ 의 **변동성** (분산, 공분산의 합)에 대한 정보를 제공
- $\Omega$ 의 역행렬: 변동성이 큰 데이터가 전체 데이터에 미치는 영향을 완화
- 대각합: 벡터를 스칼라로 변환해 전체 모델과의 통일성 부여

## General Location Model – Complete Data Loglikelihood

---

$$= h(\Omega) - \frac{1}{2} \text{tr} \left( \Omega^{-1} \sum_{i=1}^n x_i^T x_i \right) + \text{tr} \Omega^{-1} \Gamma \left( \sum_{i=1}^n w_i^T x_i \right)$$
$$+ \sum_{c=1}^C \left[ \left( \sum_{i=1}^n w_{ic} \right) \left( \ln \pi_c - \frac{1}{2} \mu_c^T \Omega^{-1} \mu_c^T \right) \right],$$



$\left( \sum_{i=1}^n w_{ic} \right)$  각 범주 조합이 관측된 빈도와 그 셀 확률을 고려하는 항  
 $\left( \sum_{i=1}^n w_{ic} \right)$  : 특정 셀  $c$ 에 속하는 샘플의 수를 벡터로 표현. 거기다 각 셀 확률의 로그 가능도를 곱하고, 마지막 항으로 평균벡터에 대한 변동성 반영  
→ 전체적으로 각 범주 조합이 전체 데이터에 미치는 영향을 보여줌

$$\ell(\Gamma, \Omega, \Pi) = \sum_{i=1}^n \ln f(x_i | w_i, \Gamma, \Omega) + \sum_{i=1}^n \ln f(w_i | \Pi)$$

# General Location Model – with Missing Values

---

이제  $x_i = \{\text{난방 사용 시간, 난방 사용 금액}\}$  과  $w_i = \{\text{아파트 방향, 평수}\}$ 에서 결측이 발생했다고 하자.

주요 Notation:

$x_i$ 에서의 결측:  $x_{(0),i} = \text{관측된 연속 변수 벡터}, x_{(1),i} = \text{결측된 연속 변수 벡터}$

예:  $x_i = \{183\text{시간, ?}\}$ 에서  $x_{(0),i} = \{183 \text{시간}\}, x_{(1),i} = \{?\}$ 으로 분리해준다는 개념

$w_i$ 의 결측 : 연속형 변수와 다르게  $S_i$ 로 나타내서 결측이 발생한 변수에서 모든 가능한 조합을 표현해준다.

예:  $w_i = \{\text{동향, ?}\}$ 에서  $S_i = \{(\text{동향}, 35\text{평}), (\text{동향}, 45\text{평}), (\text{동향}, 60\text{평})\}$

$x_i$ 에서 결측이 발생하면 이전 데이터에서도 숫자로 나오므로 **조건부 평균을** 추정할 수 있고,  $w_i$ 에서 결측이 발생한다면 해당 범주가 발생한 확률밖에 계산할 수가 없으므로 **조건부 확률을** 추정할 수 있다.

# EM Algorithm in General Location Model – E step

---

E 스텝에서 기본적으로 이해해야 하는 통계량은 다음과 같다.

$$T_{1i}^{(t)} = E \left( x_i^T x_i | x_{(0),i}, S_i, \theta^{(t)} \right),$$

먼저  $T_{1i}^{(t)}$  는 관측된 연속 변수의 자기 교차곱을 나타내며 행렬곱을 하면 다음과 같다. (손으로 계산)

$$T_{2i}^{(t)} = E \left( w_i^T x_i | x_{(0),i}, S_i, \theta^{(t)} \right),$$

$$T_{3i}^{(t)} = E \left( w_i | x_{(0),i}, S_i, \theta^{(t)} \right).$$

따라서 주대각선의 원소와 그 외의 원소는 각각 자기 자신간의 분산과 서로 다른 변수들끼리의 공분산을 추정하는데 활용된다.

$T_{2i}^{(t)}$  는 관측된 범주 조합에 대한 연속 변수의 합에 대한 추정이고,  $w_i$  중 하나가 누락되면 각각의 범주에 속할 확률  $\omega_{ic}$  가 추후의 과정에 의해 계산될 수 있다. 그리고 이것이  $T_{1i}^{(t)}$ 의 계산에 활용된다.

$T_{3i}^{(t)}$ 는  $w_i$  에 관한 정보, 즉 특정  $w_i$  가 누락된 경우에 각 범주 조합에 속할 확률 벡터이다.

예: (남향) 인 것은 아는데 평수를 모를 때: 평수의 종류에 따라 0.3 과 같이 나타난다. 그리고 이는  $T_{2i}^{(t)}$ 의 계산에 활용된다.  
0.5  
0.3  
0.2

# EM Algorithm in General Location Model – E step

---

위에서 제시한  $(0.5, 0.3, 0.2)$ 와 같은 확률은 어떤 과정을 통해서 구해지는가? ( $T_{3i}^{(t)}$ 의 계산)

먼저 notation을 다시 한번 정의하자면

$$\omega_{ic} = \Pr(w_i = U_c | x_{(0),i}, S_i, \theta^{(t)})$$

이고 이는 관측된 연속 변수  $x_{(0),i}$ 에 대해 특정 세대가 범주 조합  $c$ 에 속할 조건부 사후 확률 (conditional posterior probability)이다.

이러한 각각의  $w_{ic}$ 를 얻기 위해서는 먼저 해당 범주 조합  $c$ 에서의  $\delta_{ic}$ 를 구해주어야 한다. 그 식은 다음과 같은데,

$$\delta_{ic} = x_{(0)i} \Omega_{(0),i}^{-1} \mu_{(0),i}^T - \frac{1}{2} \mu_{(0)i} \Omega_{(0),i}^{-1} \mu_{(0),i}^T + \ln(\pi_c)$$

여기서 첫 번째 항은 관측된 연속변수와 각 범주에서의 조건부 평균의 상관관계에  $\Omega^{-1}$ 로 가중치를 조절한 것.

두 번째 항은 해당 범주 조합에서의 연속변수의 평균 벡터에 마찬가지로  $\Omega^{-1}$ 로 가중치를 조절한 것 → 변수 분산 UP, 가중치 DOWN

세 번째 항은 지금까지 관측된 데이터에서 특정 범주  $c$ 에 속할 확률에 로그를 취해준 값.

# EM Algorithm in General Location Model – E step

---

구해진  $\delta_{ic}$ 는 같은 결측된 범주 유형 내에서 가능한 범주에 대해 모든  $\delta_{ic}$  를 같은 방식으로 계산하여, softmax 함수를 취해주 어 0과 1 사이의 값을 갖게 하고, 확률의 성질을 부여한다.

예를 들어, (북향,35평) 일 때의  $\omega_{ic}$  를 추정하는 데에는 해당 범주에 대한  $\delta_{ic}$  를 계산해 지수함수를 취한 것을 나머지 범주 (북향,45평), (북향,60평)에 대한  $\delta_{ic}$ 를 계산해 각각에 지수함수를 취해 합한 것으로 나누어주니 확률의 성질을 갖게 된다. → 따라서  $\omega$ 를 구하기 위한 합리적인 방법이다.

$$\frac{\exp(\delta_{i,"\text{북향},35\text{평}"})}{\exp(\delta_{i,"\text{북향},35\text{평}"}) + \exp(\delta_{i,"\text{북향},45\text{평}"}) + \exp(\delta_{i,"\text{북향},60\text{평}"})} = \omega_{i,"\text{북향},35\text{평}"}$$

이와 같이 모든 범주 조합에 대해  $\omega$ 를 구할 수 있고 이것이 E step에서의  $T_{3i}^{(t)}$  의 계산에 활용된다.

# EM Algorithm in General Location Model – E step

---

$T_{2i}^{(t)}$ 의 계산은 다음과 같은데, 일단 notation부터 다루자면

$$E(w_{ic}x_{ij}|x_{(0),i}, S_i, \theta^{(t)}) = \begin{cases} \omega_{ic}\hat{x}_{ij}^{(c)}, & \text{if } x_{ij} \text{ is missing,} \\ \omega_{ic}x_{ij}, & \text{if } x_{ij} \text{ is observed.} \end{cases}$$

이전 스텝에서 구해진 특정 범주에 속할 확률  $\omega_{ic}$  와 연속 변수 곱의 기댓값인데, 이때 관측되었으면 그대로 적어준다.

연속 변수  $x_i$ 와  $x_j$  중  $x_i$  가 관측되지 않았다고 한다면, 관측되지 않은 값인  $\hat{x}_i$  에 대해서는 해당 범주 조합  $c$  내에서 추정된  $x_i$  의 평균 벡터, 다른 관측된 연속 변수  $x_j$ , 그리고  $x_j$ 와의 공분산행렬을 사용해 다음과 같이 계산된다.

$$\hat{x}_{\text{난방 사용 시간}(c)} = \mu_{\text{난방 사용 시간}(c)} + \Sigma_{\text{난방 사용 시간}, \text{난방 사용 금액}(c)} \Sigma_{\text{난방 사용 금액}}^{-1} (\text{난방 사용 금액} - \mu_{\text{난방 사용 금액}(c)})$$

[남향, 35평]이라는 범주 조합에서 난방 사용 시간의 평균을 알고, 해당 조합에서의 난방 사용 시간, 금액 간의 공분산, 그리고 관측된 난방 사용 금액과 실제 [남향, 35평]에서의 난방 사용금액의 평균의 차이 등을 구할 수 있으므로 [남향, 35평]의 난방 사용 시간 조건부 평균을 구해줄 수 있다.

총 분산의 역수는 전과 마찬가지로 가중치를 반영하는 역할을 한다.

# EM Algorithm in General Location Model – E step

---

$$E(x_{ij}x_{ik} \mid x_{(0),i}, S_i, \theta^{(t)}) = \sum_{c \in S_i} \omega_{ic} E(x_{ij}x_{ik} \mid x_{(0),i}, w_i = U_c, \theta^{(t)})$$

from  $T_{3i}^{(t)}$       from  $T_{2i}^{(t)}$

$x_{ij}, x_{ik}$  both observed;  
 $x_{ij}$  missing,  $x_{ik}$  observed;  
 $x_{ik}$  missing,  $x_{ij}$  observed;  
 $x_{ij}, x_{ik}$  both missing.

$$= \begin{cases} x_{ij}x_{ik}, & \\ x_{ik} \sum_{c \in S_i} \omega_{ic} \hat{x}_{ij}^{(c)}, & \\ x_{ij} \sum_{c \in S_i} \omega_{ic} \hat{x}_{ik}^{(c)}, & \\ \sigma_{jk \cdot (0),i} + \sum_{c \in S_i} \omega_{ic} \hat{x}_{ij}^{(c)} \hat{x}_{ik}^{(c)}, & \end{cases}$$

마지막으로  $T_{1i}^{(t)}$  은 연속 변수간의 자기 교차곱의 기댓값을 나타내며 이는 지금까지 계산된 통계량을 기반으로 계산된다.

$x_{ij}, x_{ik}$ : 각각 (난방 사용 시간, 난방 사용 금액) 으로 생각할 수 있고 두 개 모두 결측된 상황에는 조건부 공분산을 고려해주어야 한다.

여기서 조건부 공분산은 각 변수값이 범주에서의 평균에서 벗어난 정도를 알려주는 것이기 때문에 두 변수값이 모두 결측되어 추정치를 사용해야 하는 마지막의 경우에는 특히 유용하다.

$$E(\text{난방 사용 시간} \times \text{난방 사용 금액}) = \sum_c \omega_{ic} (\hat{x}_{\text{난방 사용 시간}(c)} \cdot \hat{x}_{\text{난방 사용 금액}(c)} + \text{Cov}(\text{난방 사용 시간}, \text{난방 사용 금액} \mid \text{범주 조합 } c))$$

## EM Algorithm in General Location Model – M step

---

$$\Pi^{(t+1)} = n^{-1} \sum_{i=1}^n T_{3i}^{(t)},$$

$$\Gamma^{(t+1)} = H^{-1} \left( \sum_{i=1}^n T_{2i}^{(t)} \right),$$

$$\Omega^{(t+1)} = n^{-1} \left[ \sum_{i=1}^n T_{1i}^{(t)} - \left( \sum_{i=1}^n T_{2i}^{(t)} \right)^T H^{-1} \left( \sum_{i=1}^n T_{2i}^{(t)} \right) \right],$$

M 스텝은 다음과 같은데, 각각 확률, 평균 벡터, 그리고 공분산에 대한 모수를 그 전 시점으로부터 업데이트하는 과정을 거친다. 우리가 E 스텝에서 구한  $T_{3i}^{(t)}$ ,  $T_{2i}^{(t)}$ ,  $T_{1i}^{(t)}$ 를 차례대로 사용하는 것을 볼 수 있다.

# General Location Model – Data Augmentation

---

Data Augmentation \_ I step (Imputation) 과 P step (Posterior)로 구성되는데 이는 EM 알고리즘의 두 스텝과 대응된다.  
따라서 DA에서의 데이터 역시 연속형 변수와 범주형 변수 모두에서 결측이 발생할 수 있고, 우선 다음과 같이 정보가 없는 사전 분포 (noninformative prior distribution) 을 가정한다.

$$p(\Pi, \Gamma, \Omega) = \prod_{c=1}^C \pi_c^{-1/2} |\Omega|^{-(K+1)/2}$$

디리클레 분포에서 보았듯이 각 범주의 확률에  $\frac{-1}{2}$ 를 씌워주어 각 범주에 대한 중립성을 유지하고, 연속 변수의 차원인 K가 커지면 작아지는 공분산 행렬의 역행렬로 변동성을 줄여준다.

이제 I Step은 I1 과 I2라는 2개의 substep으로 구성되는데, 우선 I1에서는 범주형 변수의 대체를 다루고, I2에서는 연속변수의 대체를 다룬다.

I1 → Contingency table 의 특정 cell 에 개체  $i$ 를 지금까지 추정된  $\theta^t$ 를 기반으로 주어진  $\omega_{ic}$ 를 이용해 채워 넣는다.

I2 → 채워진 범주형 변수와  $\theta^t$ 를 기반으로 결측된  $x_{(1),i}$  들을 채워 넣는다. → 이 과정으로 Complete dataset  $Y^{(t)}$  생성

# General Location Model – Data Augmentation

$$\theta = (\Pi, \Gamma, \Omega)$$

다음 P Step 에서는 I step으로부터 채워진 완전한 데이터셋  $Y^{(t)}$ 에서 파라미터들을 새롭게 갱신하는 과정을 거친다. 이때 범주형 변수의 확률 파라미터  $\pi^{t+1}$ 은 다음과 같은 과정으로 업데이트 되는데, 이는 디리클레 분포를 따른다

$$p(\Pi | \{Y^{(t)}\}) = \prod_{c=1}^C \pi_c^{n_c^{(t)} - 1/2} \quad \text{왜 } \frac{1}{2} \text{ 인가?}$$

그리고 연속형 변수의 공분산 행렬  $\Omega^{t+1}$ 은 다음과 같이 Inverse-Wishart 분포를 따르고 그로부터 계산된다.

$$(\Omega | \Pi^{(t+1)}, Y^{(t)}) \sim \text{inv-Wishart}(S^{(t)}, n - C)$$

마지막으로 각 범주의 평균 벡터  $\mu^{t+1}$ 는 지금까지 주어진 정보들과  $\mu^t$ 의 사후분포를 이용해서 추출되고, 이는 연속형 변수의 개수인  $K$  차원의 다변량 정규분포를 따른다.

$$(\mu_c | \Pi^{(t+1)}, \Omega^{(t+1)}, Y^{(t)}) \sim N_K \left( \bar{y}_c^{(t)}, \Omega^{(t+1)} / n_c^{(t)} \right)$$

각 범주에 속하는 관측치의  
수에 따라 분산을 조정하는  
역할을 한다

셀  $c$ 에서 대체된 연속형 변수의 평균

# 3

## The General Location Model with Parameter Constraints

---

1. Introduction
2. Restricted Models for the Cell Means
3. Loglinear Models for the Cell Probabilities
4. Modification to the Algorithms of Previous Sections to  
Accommodate Parameter Restrictions

## 14.3 The General Location Model with Parameter Constraints

---

앞서 언급한 General Location Model의 경우

1. distinct mean vector  $\mu_c$  for each cell  $c$
2. cell probability  $\pi_c$  (물론  $\sum \pi_c = 1$ 은 생각하지 않는 경우에 말이다!)

에 대해 어떠한 제약(constraints, restrictions)도 걸려있지 않았다.

그럼 이제부터는 parameters에 제약이 걸린 경우에 어떤 형태의 분석이 필요할까?에 대해  
다뤄보도록 하겠다.

그러면 우선 어떤 형태로 제약을 걸 수 있을까?

우선  $\mu_c$ 에 대해서는 ANOVA-like restriction을 줄 것이고

$\pi_c$ 에 대해서는 restricted loglinear model을 사용할 것이다.

그래서... 이 제약들이 의미하는게 뭘까?

## 14.3.2 Restricted Models for the Cell Means

---

**ANOVA**는 데이터 내 그룹 간 변동성과 그룹 내 변동성을 비교하여 그룹 간 차이를 분석하는 방법이다.

**ANOVA-like restriction**이란 각 셀  $c$ 에서의 평균값들에 특정한 제약을 두고 분석하는 방식을 의미한다.

$\mu_c$ 는 각 셀의 평균을 의미한다. 이 셀은 당연하게도 **범주형 변수들의 조합**에 의해 정의된다. 이 때

$\mu_c$ 가 단순히 각각의 셀에서 독립적으로 추정되지 않고, **셀들 간의 차이를 설명할 수 있는 구조적 제약**이 존재할 수 있다. 이 때 범주형 변수들 사이의 효과(**main effect**든 **interaction effect**든)가 결국 **continuous RV**의 평균에 어떤 영향을 미치는지 설명하고자 하는 것이 ANOVA-like restriction이다. 즉, 범주형 변수들이 발생시키는 효과로 제약을 준다는 것인데, 결국 ANOVA-like restriction은 **연속형 변수들의 평균이 범주형 변수들에 의해 어떻게 영향을 받는지를** 구조적으로 모델링한다고 말할 수 있겠다.

Given that  $w_i = U_c$ ,

$$(x_i | w_i = U_c, \theta) \sim_{\text{ind}} N_K(\mu_c, \Omega)$$

연속형 RV와 범주형 RV의 관계

## 14.3.2 Restricted Models for the Cell Means

---

우선 일반적인 경우를 생각해 보자. (이해를 위한 정말 간단한 모델을 가져와 봤다)

$X$  = 연봉

$Y_1$  = 직업(판사, 학생)

$Y_2$  = 성별(남자, 여자)



이라고 생각해보자.

일반적인 경우에서 우리는 단순히 남성/여성과 학생/판사의 연봉 평균을 각각 독립적으로 추정할 수 있다. 그러면, 각각의 범주형 변수 조합(성별과 직업)에 따라 독립적인 평균값을 가지게 될 것이다. 즉, 범주형 변수들이 서로 별개의 영향을 미치는 것처럼 계산될 수 있겠다.

자 그럼 이제 범주형 변수들이 독립적으로 연봉에 영향을 미치는 것이 아니라, 구조적으로 제약된 방식으로 영향을 미친다고 가정해보자. 이 구조적 제약은 **주효과(main effect)**와 **상호작용 효과(interaction effect)**를 반영하는 방식다. 예를 들어:

- 성별은 연봉에 영향을 주지만, 그 차이는 고정된 상수로 반영된다.
- 직업은 연봉에 더 큰 영향을 미치고, 그 차이는 직업 유형에 따른 상수로 반영된다.
- 성별과 직업의 상호작용도 존재한다.

여기서 ANOVA-like 제약 적용을 해보면 다음과 같이 연봉의 평균을 표현할 수 있다.

$$\mu_c(\text{연봉의 평균}) = \mu(\text{기본값}) + \beta_1(\text{성별효과}) + \beta_2(\text{직업효과}) + \beta_3(\text{상호작용효과})$$

## 14.3.2 Restricted Models for the Cell Means

---

for  $u \leq C$ ,  $z_i : 1 \times u$  vector of design variables for unit i

$z_i = w_i A$ ,  $A = C \times u$  matrix that represents the chosen model

$f(x_i | w_i, \theta) \sim N_k(z_i B, \Omega)$ ,  $B = u \times k$  matrix of unknown parameters

$\theta = (B, \Omega, \Gamma)$ ,  $E[x_i | w_i, \theta] = w_i A B$  so that  $\Gamma = AB$

In the model of Section 14.2  $A = C \times C$  identity matrix

우선

$u \leq C$ 인 이유를 생각해보자.

C가 의미하는 것은 결국 contingency table에서 발생 가능한 모든 cell의 개수이다.

예를 들어 직업과 성별이 연봉에 영향을 미친다고 했을 때 A에 담긴 정보는 '각 범주형 변수가 어떻게 평균에 영향을 미치는가?'이다.

즉,

1. 성별이 남자라면?
2. 직업이 판사라면?
3. 성별과 직업의 interaction이 존재한다면?
4. 성별도 직업도 큰 영향을 주지 않는다면?

대부분 4.에 대한 고려는 안 할 것이므로 이 부분은 생략하자. 결국 이렇게 총 4가지 case가 존재하는데,  $C=4$ 고 당연히  $u$ 는 이거보단 적어야 할 것이다. 그럼 이걸 설명하는 행렬이 곧  $A$ 라고 생각하면 된다.

여기서  $w_i$ 를 곱해준게  $z_i$ 니까  $z_i$ 는 결국 우리가 알던 원래 data에 제약이 이렇게 들어있다를 보여주는 것이다.

그럼 B는 무엇을 의미하는가?  $z_i B$ 는 곧  $x_i$ 들의 평균(기댓값)이다. 즉, A가 주는 정보가 '이렇게 제약이 걸려있다'였다면 얼마나 제약이 걸려있는지를 추정하는게 우리의 목표이다. 14.2에서는 특별한 제약이 없었다. 이말은 무엇인가? 범주형 변수들이 서로 독립적으로 연속형 변수의 평균에 영향을 미친다라는 뜻이다. 따라서 A는 identity matrix인 것이다.

## 14.3.3 Loglinear Models for the Cell Probabilities

---

### Log-Linear Model: Additive Effect Model of Cell Count

앞서 우리는 범주형 변수들이 독립적으로 연속형 변수들에 영향을 미치는 것이 아니라, 구조적으로 제약된 방식으로 영향을 미치는 경우를 생각해 봤고, 이 구조적 제약은 **주효과(main effect)**와 **상호작용 효과(interaction effect)**를 반영한다고 했다. 그럼 당연히 cell probability도 이전과는 다른 형태로 표현될 것이라고 생각할 수 있다.

제약이 없는 경우를 다시 떠올려 보면, 그냥 단순하게  $\pi_c$ 로 cell probability를 표현해도 아무런 문제가 없었다. 그러나, 이제는 제약이 존재하니까... 각  $\pi_c$ 의 조합으로 나타날 수 있는 효과들을 고려해야 한다. 이말은 즉, 파라미터가 이전보다 늘어난다는 것이다! 앞선 예시를 다시 생각해보자.

1. 성별이 남자라면?
2. 직업이 판사라면?
3. 성별과 직업의 interaction이 존재한다면?

위와 같이 범주형 변수들이 가질 수 있는 효과를 생각해볼 수 있었는데, 그럼 이를 효과적으로 나타내는 모델은 무엇이 있을까? 여기서 사용 되는 것이 바로 log-linear model이다.

### 14.3.3 Loglinear Models for the Cell Probabilities

---

suppose the cells are formed by a joint classification of  $V = 3$  categorical variables  $Y_1, Y_2$ , and  $Y_3$  with  $I_1, I_2, I_3$  levels, then  $C = I_1 \times I_2 \times I_3$

we modify the notation so that  $\pi_{jkl}$  is the probability that

$Y_1 = j, Y_2 = k, Y_3 = l$

for  $j = 1, \dots, I_1, k = 1, \dots, I_2, l = 1, \dots, I_3$

The log-linear models are obtained by :

$$\ln \pi_{jkl} = \alpha + \alpha_j^{(1)} + \alpha_k^{(2)} + \alpha_l^{(3)} + \alpha_{jk}^{(12)} + \alpha_{kl}^{(23)} + \alpha_{jl}^{(13)} + \alpha_{jkl}^{(123)}$$

즉 이런식으로 합의 형태로 선형적으로 표현이 가능하다. parameter의 수가 늘어났다고 말한 부분이 여기서 바로 직관적인 이해가 가능한데, 결국  $\pi_{jkl}$ 이  $\alpha$ 로 표현되기 때문이다. 예시를 이렇게 들어서 그렇지 실제로는 이보다 더 많은 경우의 수가 존재할 것이기 때문에, 고려해야 할 부분들이 참 많다고 할 수 있다.

그러나 log-linear model의 또다른 장점을 이를 이용해서 parameter의 개수를 줄일 수 있다는 것이다. 아니 방금 parameter의 개수가 늘어났다고 그러지 않았나? 맞다. 그러니까 줄이는 것이다. 아까 연봉 예시를 생각해보자. 다만 이번에는, 성별 하나만 의거해서 연봉에는 영향을 주지 않는다고 생각해보자. 그러면 해당 효과에 의거한  $\alpha$ 를 0으로 부여하면서 점차 parameter를 소거해 나갈 수 있다.

## 14.3.4 Modifications to the Algorithms of Previous Sections to Accommodate Parameter Restrictions

---

그럼 이제부터는 주어진 제약을 기반으로 EM 알고리즘을 수정하는 단계에 대해 알아보자.

For a general  $V$  – way table with  $C = \prod_{j=1}^V I_j$  cells  
let  $\alpha \neq 0$  terms in the log-linear model, and  $\pi_c(\alpha)$  for the constrained probability of a unit falling in cell  $c = 1, \dots, C$

이제 이러한 조건을 가지고 incomplete data를 여기에 fitting하여 missing을 추정하는 EM 알고리즘을 알아보자.

앞서 General Location Model에서 언급된 것처럼 모든 random variables은 서로 독립이라는 가정하에서 얻어진  $\alpha^{(0)}, \Omega^{(0)}, B^{(0)}$ 이라는 초기값을 생각해보자.

또한

$\Gamma^{(0)} = AB^{(0)}$ 도 놓치면 안 되겠다.

이 때

$A$ 는 known matrix이고  $\pi_c^{(0)} = \pi_c(\alpha^{(0)}), c = 1, \dots, C$ 이다.

지금 우리는 incomplete data를 다루고 있고 일부 값이 누락된(**missing data가 존재하는**) 상황이다.

이때 **complete data**에서 사용했던 충분통계량을 어떻게 활용할 수 있을지 생각해볼 수 있다. 여기서 주목해야 할 점은 restriction이 적용되길 했지만, 이 모델도 exponential family models로 볼 수 있다는 것이다. 따라서 우리가 사용해야하는 통계량은 general location model의 경우와 크게 다르지 않다. 우리가 사용해야할 통계량은 **minimum sufficient statistic**이다.

for  $u \leq C$ ,  $z_i : 1 \times u$  vector of design variables for unit  $i$   
 $z_i = w_i A$ ,  $A = C \times u$  matrix that represents the chosen model  
 $f(x_i | w_i, \theta) \sim N_k(z_i B, \Omega)$ ,  $B = u \times k$  matrix of unknown parameters  
 $\theta = (B, \Omega, \Gamma)$ ,  $E[x_i | w_i, \theta] = w_i A B$  so that  $\Gamma = AB$   
In the model of Section 14.2  $A = C \times C$  identity matrix

## 14.3.4 Modifications to the Algorithms of Previous Sections to Accommodate Parameter Restrictions

---

**minimum sufficient statistic**란 그럼 무엇일까? Incomplete data 중에서 우리가 사용 할 수 있는 것들로만 모아서 만든 통계량을 의미한다. Incomplete data를 처리할 때, 모든 데이터를 완전하게 관측할 수 없으므로, 관측 가능한 정보만을 사용해서 파라미터를 추정하는데, **minimal sufficient statistics**는 우리가 가지고 있는 정보를 효율적으로 요약해서 사용할 수 있게 합니다. 즉, 관측된 데이터에서 파라미터를 추정하는 데 필요한 최소한의 정보를 요약한다고 생각하면 된다.

$$\begin{aligned} \sum x_i^T x_i \\ \sum w_i^T w_i A = \sum w_i^T z_i \end{aligned}$$

linear combinations of the counts  $\sum w_i$

determined by the margins fitted in the log-linear model

**log-linear 모델에 fit되었다는 것의 의미:** 결국 범주형 변수들이 조합된 셀에 대한 확률(즉,  $\pi_c$ )이 **log-linear 형태**로 설명된다. 데이터가 불완전하므로 모든 사람들의 성별과 직업 정보가 완벽히 관측되지 않을 수 있음. 각 unit이 특정 셀에 속할 확률이 **log-linear 모델**에 의해 추정된다는 의미이다.

marginalization의 의미는, log-linear 모델을 적용하면서 각 범주형 변수의 조합에 대한 셀 확률을 계산하는 과정에서, 특정 변수들에 대한 효과를 고려한 합산을 통해 해당 셀에 속 할 확률을 구하는 방식이다.

## 14.3.4 Modifications to the Algorithms of Previous Sections to Accommodate Parameter Restrictions

$$\sum x_i^T x_i$$

$$\sum w_i^T w_i A = \sum w_i^T z_i$$

linear combinations of the counts  $\sum w_i$

determined by the margins fitted in the log-linear model

사용되는 통계량의 전체적인 모습은 비슷하다.

$$\sum w_i^T x_i \rightarrow \sum w_i^T w_i A$$

$\sum w_i$ 가 log-linear model에 fitted 되었다는거 말고는 말이다.

그럼 순수 궁금증이 하나 떠오른다. 왜  $\sum x_i^T x_i$ 는 변하지 않고 가만히 있을까?

이유는 간단하다. 우리가 처음에 제약을 어디에 걸었는지 생각해보자.

바로 Cell Mean과 Cell Probabilities이다. 모두 '범주형 변수'에만 걸려있다.

General Location Model과 다른 점은, 결국 이렇게 범주형 변수에 걸려있는 제약들이 연속형 변수에 영향을 미칠 때 발생한다. 즉, 범주형 변수와 연속형 변수 간의 관계에서만 나타난다는 것이다.

그러나  $\sum x_i^T x_i$  자체는 그저 연속형 변수  $x_i$  자체에 대한 정보이다. 이는 **관측한 연속형 변수의 상관 구조**를 반영하는 부분이고, 따라서 제약이 걸리지 않은 부분이다. 이런 이유 때문에 A 행렬을 적용하지 않고 원래의 충분통계량을 그대로 사용한다고 볼 수 있다.

**E Step:**

$$T_{1i}^{(t)} = E \left( x_i^T x_i | x_{(0),i}, S_i, \theta^{(t)} \right),$$

$$T_{2i}^{(t)} = E \left( w_i^T x_i | x_{(0),i}, S_i, \theta^{(t)} \right),$$

$$T_{3i}^{(t)} = E \left( w_i | x_{(0),i}, S_i, \theta^{(t)} \right).$$

↑

제약이 없었을 때

General Location Model의

E-step

## 14.3.4 Modifications to the Algorithms of Previous Sections to Accommodate Parameter Restrictions

---

General Location Model과 동일하다. 다만? 앞서 언급했던 것 처럼 충분통계량만 살짝 수정됐다.

$$\begin{aligned}T_{1i}^{(t)} &= E(x_i^T x_i \mid x_{(0),i} S_i, \theta^{(t)}) \\T_{2i}^{(t)} &= E(w_i^T w_i A \mid x_{(0),i} S_i, \theta^{(t)}) \\T_{3i}^{(t)} &= E(w_i \mid x_{(0),i} S_i, \theta^{(t)})\end{aligned}$$

이제 이걸 바탕으로

$$\begin{aligned}\sum T_{1i}^{(t)} \\ \sum T_{2i}^{(t)} \\ \sum T_{3i}^{(t)}\end{aligned}$$

를 계산하면 그만이다! 즉 E-step과 I-step에서는 General Location Model과 큰 차이가 없다.

**E Step:**

$$T_{1i}^{(t)} = E\left(x_i^T x_i \mid x_{(0),i}, S_i, \theta^{(t)}\right),$$

$$T_{2i}^{(t)} = E\left(w_i^T x_i \mid x_{(0),i}, S_i, \theta^{(t)}\right),$$

$$T_{3i}^{(t)} = E\left(w_i \mid x_{(0),i}, S_i, \theta^{(t)}\right).$$

↑

제약이 없었을 때

General Location Model의  
E-step

## 14.3.4 Modifications to the Algorithms of Previous Sections to Accommodate Parameter Restrictions

---

그러나 M-step과 P-step의 경우 방법이 약간 달라진다.

잊지말자 우선 우리가 update해야 할 parameters는 다음과 같다.

$$\theta = (B, \Gamma, \Omega)$$

우선 log-linear model의 parameter인  $\alpha$ 에 대해서는,  $\sum T_{3i}^{(t)}$ 을 통해 계산된 cell frequencies(빈도수)를 바탕으로 multway table을 만들어야한다. 이 때 빈도수는 부분적인(fractional) 빈도수로 처리해야 한다.

이게 무슨말인가? 우리는 incomplete한 case를 다루고 있기 때문에, data가 정확히 어느 범주(category 나아가 cell)에 속할지 확신할 수 없다. 즉, data가 여러 범주 중 하나에 속할 가능성이 있기 때문에, 한 범주에 완전히 속한다고 보지 않고 여러 범주에 분할된 확률(fractional probability)로 분배되는 것이다.

아까의 예시를 다시 생각해보자. 우리는 missing이 있는 data가 남자인지 여자인지, 판사인지 학생인지 확실하게 알 수 없다. 예를 들어, 한 사람이 **남자일 확률 0.6, 여자일 확률 0.4**이고, **교수일 확률 0.5, 학생일 확률 0.5**라고 하면, 이 사람은 각 범주에 부분적으로 분배된다. 이렇게 말이다.

- 남자면서 교수:  $0.6 * 0.5 = 0.3$
- 남자면서 학생:  $0.6 * 0.5 = 0.3$
- 여자면서 교수:  $0.4 * 0.5 = 0.2$
- 여자면서 학생:  $0.4 * 0.5 = 0.2$

이 값들이  $\sum T_{3i}^{(t)}$ 에 저장이 되고, 이를 바탕으로 우리는 log-linear model의 parameter인  $\alpha$ 를 추정하는 것이다.

$$\begin{aligned}T_{1i}^{(t)} &= E(x_i^T x_i | x_{(0),i} S_{i,\theta}^{(t)}) & \sum T_{1i}^{(t)} \\T_{2i}^{(t)} &= E(w_i^T w_i A | x_{(0),i} S_{i,\theta}^{(t)}) \rightarrow & \sum T_{2i}^{(t)} \\T_{3i}^{(t)} &= E(w_i | x_{(0),i} S_{i,\theta}^{(t)}) & \sum T_{3i}^{(t)}\end{aligned}$$

## 14.3.4 Modifications to the Algorithms of Previous Sections to Accommodate Parameter Restrictions

---

만약 explicit한 estimates가 존재하지 않는 상황이라면, **IPF(Iterative Proportional Fitting)**를 적용  
해서  $\alpha$ 를 update하면 되는데, 그러면 EM 알고리즘에서 ECM 알고리즘으로 바뀐다.(저번주 복습!!!)

[IPF를 이용한 ECM 알고리즘의 구축]

ML에서 가장 흔히 쓰이는 방법론—> IPF (Iterative Proportional Fitting)

IPF는 테이블의 margin에 대한 정보를 맞추기 위해 비례적인 조정(proportional adjustment)을 가함.  
이때 테이블의 margin은 최소 충분 통계량이다.

EM의 M step을 IPF로 대체하여 ECM 알고리즘으로 로그선형모델의 모수를 추정한다.

예를 들어 남자 교수에 대한 데이터가 부족하거나 추정이 어렵다면, 남자 교수 카테고리의 비율을 다른 정보(다른 범주의 확률이나 전체 데이터를 보정)를 이용하여 다시 맞춰 놓고 추정을 진행한다는 것이다.

남자 교수 **0.3**, 남자 학생 **0.3**, 여자 교수 0.2, 여자 학생 0.2에서

남자 교수 **0.25**, 남자 학생 **0.35**, 여자 교수 0.2, 여자 학생 0.2으로 조정할 수 있다는 이야기이다.

Bayesian의 경우(DA)도 비슷하게 **Bayesian IPF**를 사용하면 된다.

## 14.3.4 Modifications to the Algorithms of Previous Sections to Accommodate Parameter Restrictions

---

Complete data가 주어졌을 때 ML estimates는 다음과 같이 계산된다.

$$\hat{B} = (\sum z_i^T z_i)^{-1} \sum z_i^T x_i$$
$$\hat{\Omega} = (\sum x_i - z_i \hat{B})^T (x_i - z_i \hat{B})$$

여기서 이제  $z_i = w_i A$ 를 대입하고,

$$\sum x_i^T x_i \rightarrow \sum T_{1i}^{(t)}$$
$$\sum w_i^T x_i \rightarrow \sum T_{2i}^{(t)}$$
$$\sum w_i^T w_i \rightarrow D^{(t)}$$

(where  $D^{(t)}$  is a matrix with elements of  $\sum T_{3i}^{(t)}$  on the diagonal and zeros elsewhere,  
앞의 H matrix와 비슷한 것으로 보면 됨)

로 바꿔주고 계산을 진행하면? M-step에서 다음과 같은 estimates를 계산 가능하다.

$$\hat{B}^{(t+1)} = (A^T D^{(t)} A)^{-1} A^T \left( \sum T_{2i}^{(t)} \right)$$

$$\Gamma^{(t+1)} = A \hat{B}^{(t+1)}$$

$$\Omega^{(t+1)} = n^{-1} \left[ \sum_{i=1}^n T_{1i}^{(t)} - \left( \sum_{i=1}^n T_{2i}^{(t)} \right)^T A (A^T D^{(t)} A)^{-1} A^T \left( \sum_{i=1}^n T_{2i}^{(t)} \right) \right]$$

## 14.3.4 Modifications to the Algorithms of Previous Sections to Accommodate Parameter Restrictions

---

아까 제약이 없는 경우에는  $A$ 가  $C \times C$ 의 identity matrix가 된다는 사실을 기억 하는가? 제약이 없으면 결국 이 estimates도

$$\begin{aligned}\Pi^{(t+1)} &= n^{-1} \sum_{i=1}^n T_{3i}^{(t)}, \\ \Gamma^{(t+1)} &= H^{-1} \left( \sum_{i=1}^n T_{2i}^{(t)} \right), \\ \Omega^{(t+1)} &= n^{-1} \left[ \sum_{i=1}^n T_{1i}^{(t)} - \left( \sum_{i=1}^n T_{2i}^{(t)} \right)^T H^{-1} \left( \sum_{i=1}^n T_{2i}^{(t)} \right) \right]\end{aligned}$$

$$\begin{aligned}\hat{B}^{(t+1)} &= (A^T D^{(t)} A)^{-1} A^T \left( \sum T_{2i}^{(t)} \right) \\ \Gamma^{(t+1)} &= A \hat{B}^{(t+1)} \\ \Omega^{(t+1)} &= n^{-1} \left[ \sum_{i=1}^n T_{1i}^{(t)} - \left( \sum_{i=1}^n T_{2i}^{(t)} \right)^T A (A^T D^{(t)} A)^{-1} A^T \left( \sum_{i=1}^n T_{2i}^{(t)} \right) \right]\end{aligned}$$

즉, 처음 우리가 다뤘던 모델과 같아진다. 여기서도 summation의 첨자를 생각해보면  $x_i$ 의 variance-covariance matrix인  $\Omega$ 만 1부터 n까지 적혀있음을 확인 가능하다. 즉, 오롯이  $x_i$ 와 관련이 있다면, 제약에 영향을 받지 않음을 다시 한 번 알 수 있다.

I-step 같은 경우  $\Omega^{(t+1)}$ 의 첫 번째 draw가 자유도가  $n - u$ 로 조정된 inverse-Wishart distribution을 따른다. 그리고 나면  $B^{(t+1)}$ 는 방금 우리가 본  $\hat{B}^{(t+1)}$ 가 centered되고 variance-covariance matrix 가  $\Omega^{(t+1)}$ 인 MVN에서 draw 가능 하다.

# 4

## Regression Problems Involving Mixtures of Continuous and Categorical Variables

---

1. Normal Linear Regression with Missing Continuous or Categorical Covariates
2. Logistic Regression with Missing Continuous or Categorical Covariates

## 14.4.1 Normal Linear Regression Involving Missing Continuous or Categorical Variables

---

자 그럼 이제 14.2의 내용과 14.3의 내용을 바탕으로 missing data가 존재하는 linear regression을 해결하는 알고리즘을 생각해보자.

conditional distribution of a continuous variable (say

$X_1$ ) given the other variables

:

$$(X_1 | X_2, \dots, X_k, Y_1, \dots, Y_V, \theta) \sim N(\beta_{c0}(\theta) + \sum_{j=2}^K \beta_j(\theta) X_j, \sigma^2(\theta))$$

where  $c$  is the cell of the contingency table formed by  $(Y_1, \dots, Y_V)$

**Property 6.1** Let  $g(\theta)$  be a function of the parameter  $\theta$ . If  $\hat{\theta}$  is an ML estimate of  $\theta$ , then  $g(\hat{\theta})$  is an ML estimate of  $g(\theta)$ .

**Property 6.1B** Let  $g(\theta)$  be a function of the parameter  $\theta$ , and let  $\theta^{(d)}$  be the  $d$ th draw from the posterior distribution of  $\theta$ ,  $d = 1, \dots, D$ . Then,  $g(\theta^{(d)})$  is a draw from the posterior distribution of  $g(\theta)$ .

이러한 property에 의해 위의 모델의 ML estimate는 쉽게 구할 수 있다. 이유는 간단하다. Property 6.1, 6.1B가 말하는건 결국 MLE의 invariance property와 관련된 내용이다. missing이 존재하는 data 도 결국 앞서 언급한 것처럼 complete data의 linear combination으로 표현 가능하다. 즉, complete data의 함수로 표현이 가능하다. 그럼 이제 이 property를 이용해서 regression parameters인  $\{\beta_{c0}(\theta), \beta_j(\theta), \sigma^2\}$ 에서  $\theta$ 만 ML estimates인  $\hat{\theta}$ 나 draws인  $\theta^{(d)}$ 로 바꿔주기만 하면 된다.

여기에 MANOVA-restriction을 추가로 걸어주면 기댓값 부분인  $\{\beta_{c0}\}$ 에 restriction이 생길 것이고, 이 때는 linear이 아닌 다른 형태의 regression도 도출이 가능하다.

갑자기 나온 MANOVA는 별게 아니고 ANOVA의 Multivariate version이다. 아까의 예시에서  $X_2 =$  근무시간이라는 새로운 변수가 추가되었다고 보면 된다.

# General Location Model – 주요 장점

---

데이터가 일반 위치 모형을 따른다고 가정했을 때의 주요 특징은

- i) 범주형 변수  $w$  가 없다면 손쉽게 연속형 변수만 존재하는 다변량 정규 모형으로 축소가 가능하고 이는 11.2에서 배운 다변량 정규 데이터에 해당하는 알고리즘으로 결측치를 해결해볼 수 있다는 것을 의미한다.
- ii) 연속형 변수  $x$  가 없다면 범주형 변수만 존재하므로 Contingency Table 을 구성할 수 있고, 결측치가 있는 경우 partially classified contingency table을 적용할 수 있다.
- iii) 모든 셀에서 공분산 행렬이 같다고 추정됨으로써, 해석이 보다 용이해지고 필요한 매개변수의 수가 줄어들고, 과적합 문제를 방지할 수도 있다.
- iv) 범주형 변수  $Y$ 가 0과 1 사이 값을 갖는 binary variable일 때  $P(Y=1)$  은  $\frac{e^L}{1+e^L}$  과 같이 구해질 수 있고, 이때  $L$ 은 다른 연속형 변수들의 선형 결합으로 표현된다. (Details in 14.4)

## 14.4.1 Normal Linear Regression Involving Missing Continuous or Categorical Variables

---

특별한 케이스 하나를 생각해보자.  $X$ 라는 single continuous outcome만 존재하고, regressor(independent variable)은 모두 categorical variable인 경우를 생각해보자.

이러한 경우, EM 알고리즘의 M-step은 weighted linear regression으로 표현된다.

그럼 weight을 어떤걸 쓴다는 것일까? 바로 probabilities that an incomplete unit  $i$  belongs to each of the set of possible cells  $S_i$ 라고 나와 있다.  $S_i$ 가 뭐였더라...

바로 각각의 데이터 포인트가 어느 셀(cell)에 속할 확률이다. 어라? 이런 흐름 어디서 봤던 것 같은데....

남자면서 교수:  $0.6 * 0.5 = 0.3$

남자면서 학생:  $0.6 * 0.5 = 0.3$

여자면서 교수:  $0.4 * 0.5 = 0.2$

여자면서 학생:  $0.4 * 0.5 = 0.2$

아 이거였다! 즉, 앞서 언급한 fractional entries가 가중으로 들어가는 weight linear regression model이 되는 것이다.

Ibrahim (1990), Horton and Laird (1998), Ibrahim et al. (1999), 그리고 Schluchter와 Jackson (1989)의 연구는 모두 EM 알고리즘을 이용해 불완전한 범주형 데이터와 관련된 문제를 다룬다. 특히 일반화 선형 모델(GLM)에서 불완전한 범주형 공변량(covariate)이 있는 경우, 이 연구들은 가중치를 사용한 선형 회귀 방법을 적용하여 이러한 문제를 해결한다.

Ibrahim(1990): EM 알고리즘에서 가중치를 사용하는 개념을 도입하여 범주형 공변량이 불완전한 상황에서도 추정을 가능하게 하는 방법을 제시

Horton과 Laird(1998), Ibrahim et al.(1999) 등은 이 연구를 확장하여 일반화 선형 모델뿐만 아니라 다양한 통계 모델에 이를 적용

Schluchter와 Jackson(1989)은 이러한 방법을 생존 분석에 적용하여, 생존 시간이 불완전한 범주형 공변량과 연관된 경우에도 EM 알고리즘을 사용해 데이터를 분석할 수 있도록 함.

즉, 이 연구들은 범주형 공변량에 결측값이 있는 경우 이를 다루기 위해 각 범주에 속할 확률에 기반한 가중치를 통해 선형 회귀 분석을 진행할 수 있도록 EM 알고리즘을 수정하는 방법들을 제시하고 있다고 볼 수 있겠다.

## 14.4.2 Logistic Regression Involving Missing Continuous or Categorical Variables

---

마지막으로 binary한 값을 가지는 categorical variable을 dependent variable로 설정한 regression 을 살펴보자.

$Y_1$ 을 binary variable이라고 할 때 general location model을 통해 다음을 알 수 있다

conditional distribution of  $Y_1$  given  $(Y_2, \dots, Y_V), (X_1, \dots, X_k)$  is Bernoulli with

$$\text{logit}(\Pr(Y_1 = 1) | Y_2, \dots, Y_V, X_1, \dots, X_K, \theta) = \gamma_{d0}(\theta) + \sum_{j=1}^K \gamma_{dj}(\theta) X_j$$

where  $d$  indexes the cell defined by the values of  $(Y_2, \dots, Y_V)$  and  $\theta$  represents the location parameters.

근데 Logit은 뭐지...

$$\log(\text{odds ratio}) = \log\left(\frac{p(y=1|x)}{1-p(y=1|x)}\right) = w_i^0 + w^T X_i$$

왜 logit transformation을 사용할까?

## 14.4.2 Logistic Regression Involving Missing Continuous or Categorical Variables

### Logistic Regression

Dependent Variable이 0 혹은 1의 값만을 가지는 binary한 경우를 생각하자.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \mathbb{E}[\epsilon_i] = 0$$

$$\epsilon = \begin{cases} 1 - \beta_0 - \beta_1 x_i & \text{if } y_i = 1 \\ -\beta_0 - \beta_1 x_i & \text{if } y_i = 0 \end{cases}$$

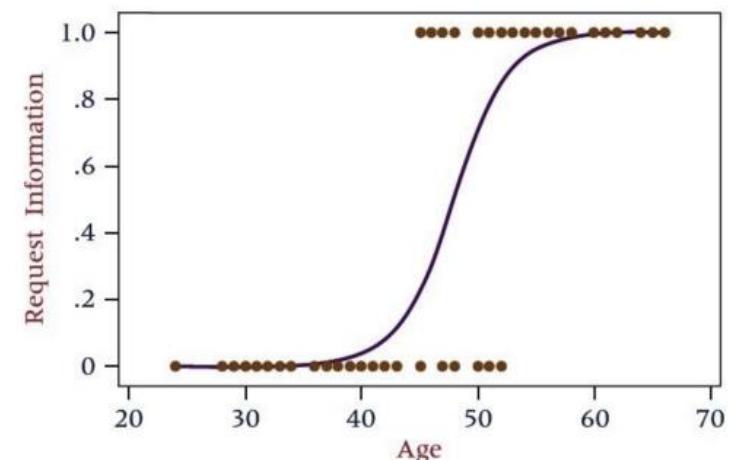
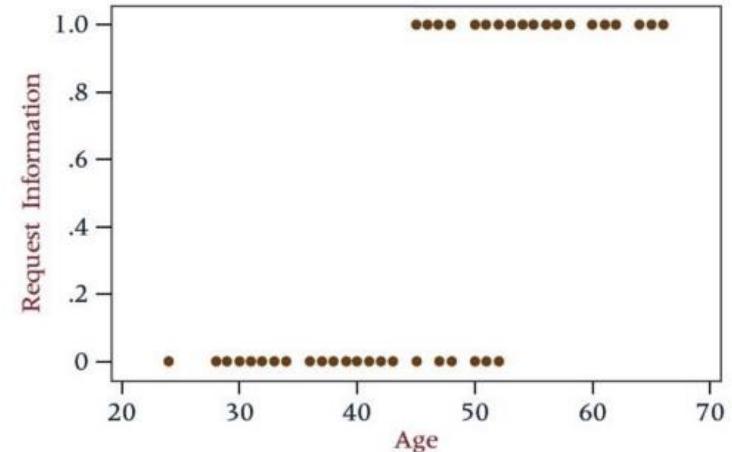
이를 통해 error term이 우선 normal distribution을 따르지 않음을 알 수 있고, 나아가

$$\mathbb{V}[\epsilon_i] = \mathbb{V}[y_i] = \mathbb{E}[(y_i - \mathbb{E}[y_i])^2] = p_i(1 - p_i)^2 + (1 - p_i)(0 - p_i)^2 = p_i(1 - p_i)$$

분산 또한 constant하지 않음을 알 수 있다.

이러한 문제가 발생하니... 다음과 같은 모델을 대신 사용하자.

$$y_i = \frac{\exp(\beta_0 + \sum_{j=1}^k \beta_j x_j)}{1 + \exp(\beta_0 + \sum_{j=1}^k \beta_j x_j)} + \epsilon_i, \quad \mathbb{E}[\epsilon_i] = 0$$



## 14.4.2 Logistic Regression Involving Missing Continuous or Categorical Variables

근데 Logit은 뭐지...

odds ratio = S로 두면 다음과 같이 표현 가능하다.

$$\log(\text{odds ratio}) = \log\left(\frac{p(y=1|x)}{1-p(y=1|x)}\right) = w_i^0 + w^T x_i$$

$$S = \frac{p}{1-p} = \exp\left(\beta_0 + \sum_{j=1}^k \beta_j x_j\right)$$

왜 logit transformation을 사용할까?

우리는 범주형 변수( $Y_1$ )가

**Bernoulli 분포**를 따른다고 가정하고 있고,  $Y_1$ 이 1일 확률을 추정하는 모델이 필요하다. 이때 logistic regression은 그 확률을 모델링하는 일반적인 방법이고, 이 과정에서 logit transformation이 사용된다.

logit transformation은 확률의 선형화를 실현해준다. 확률 값은 항상 0과 1 사이에 존재하지만, 선형 회귀는 그 값을 제한 없이 예측한다. 따라서, 확률을 예측하는 모델에서는 0과 1 사이로 값이 제한되도록 해야 한다. **logit transformation**은 확률을 선형 공간에서 다루기 쉽게 변환해준다.

이후부터는 이전과 동일하다. regression parameter인  $\{\gamma_{d0}(\theta), \gamma_{dj}(\theta)\}$ 의  $\theta$ 자리에

ML estimates인

$\hat{\theta}$ 나 draws인  $\theta^{(d)}$ 를 넣어주기만 하면 된다. 마찬가지로, 제약이 생기는 경우 다른 형태의 logistic models가 도출된다.

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{j=1}^k \beta_j x_j$$

이렇게 비교적 쉽게 linear하게 표현도 가능하다.

Maximum likelihood estimation is used to estimate the regression equation (beyond the scope of this course)

## 14.4.2 Logistic Regression Involving Missing Continuous or Categorical Variables

Multiple Imputation(MI) ch 10 복습!

### Multiple imputation

- 채워 넣는 값에도 uncertainty를 부여하고
- 채워 넣은 data를 활용한 estimate에도 uncertainty가 부여됨

	X	Y		X	Y		X	Y		X	Y
1	9	11	1	9	11	1	9	11	1	9	11
2	10	11	2	10	14	2	10	13	2	10	18
3	12	15	3	12	15	3	12	15	3	12	15
4	14	16	4	14	14	4	14	20	4	14	18
5	17	17	5	17	17	5	17	17	5	17	17
6	19	15	6	19	15	6	19	15	6	19	15
7	7	21	7	7	21	7	7	21	7	7	21
8	5	5	8	5	5	8	5	5	8	5	5
9	21	19	9	21	19	9	21	19	9	21	19
10	25	23	10	25	23	10	25	23	10	25	23

sample mean	15.3	15.4	15.9	16.2
sample variance	28	26.3	28.5	26.6
variance of	2.8	2.63	2.85	2.66
sample mean				

이렇게 구하고 난 뒤, multiple imputation의 방법론을 이용하여 **Y의 평균**과 관련된 추정치를 다음과 같이 계산할 수 있음!

평균 = 각 complete data set에서의 표본평균들의 평균

분산1 = 각 complete data set에서의 표본평균의 분산들의 평균

분산2 = 각 complete data set에서의 표본평균들의 표본분산

총분산 = 분산1 + (1+1/4) x 분산2

즉,

$$\text{평균} = (15.3 + 15.4 + 15.9 + 16.2)/4 = 15.7$$

$$\text{분산1} = (2.8 + 2.63 + 2.85 + 2.66)/4 = 2.735$$

$$\text{분산2} = \frac{(15.3 - 15.7)^2 + (15.4 - 15.7)^2 + (15.9 - 15.7)^2 + (16.2 - 15.7)^2}{4-1} = 0.18$$

$$\text{총분산} = 2.735 + (1 + \frac{1}{4})0.18 = 2.96$$

일반적으로 MI 방식에서는 결측값을 여러 번 대체하여 여러 개의 "complete-data"를 만든다. 그런 다음, 각 대체된 data에 대해 **general location model**을 맞추고 그에 따른 parameter를 추정한 후 이를 종합한다. 이때 parameter를 직접 변환하는 과정이 필요하게 된다.

하지만 여기서 책은 또 다른 방법을 하나 제안한다. 제시된 대안적 방법은, parameter의 변환 과정 없이, 각 결측값이 대체된 데이터 세트마다 **standard logistic regression**을 사용하여 모델을 바로 적용한다는 것이다. 즉, 미리 정의된 general location model에 의존하지 않고, 대체된 각각의 데이터 세트에 대해 **logistic regression**을 수행한 후, 그 결과를 모아서 최종 파라미터 추정을 진행하는 방식이다. 여기서 적용하는 모델은 다음과 같다.

$$(Y_1 | Y_2, \dots, Y_V, X_1, \dots, X_k) \sim \text{Bern}(\gamma_{d0} + \sum_{j=1}^K \gamma_{dj}(X_j))$$

일반적인 ML estimation이나 Bayes 방법에서는 데이터가 정규 분포를 따른다는 가정이 모델의 전반적인 분석과 결과에 크게 영향을 미칠 수 있지만, MI를 사용하면 이러한 가정이 덜 중요해진다. 특히 결측 정보의 비율이 적고 imputation이 적게 일어나는 경우, MI 방법은 더 효과적일 수 있다. 또한 완전한 데이터 분석을 위해 MI가 사용되므로, joint distribution을 기반으로 한 ML이나 Bayes 방법보다 더 유연하고 안정적이라고 볼 수 있다.

---

감사합니다