Missing analysis tutorial

ESC 2024 Summer Session 1주차

Contents

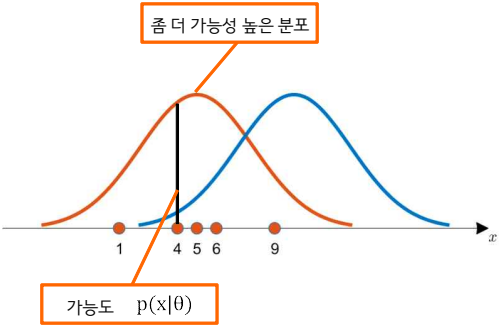
1

Background

Missing data analysis

Introduction - Likelihood (MLE)

LIkelihood는 확률 분포의 모수가, 어떤 확률 변수의 표집값과 일관되는 정도를 나타내는 값



MLE (Maximum Likelihood Estimation) 은 가능도의 곱이 가장크게 나오는 모수를 추정값으로 갖고 가는 것이다.
해당 값을 가져가는 것이 가장 그럴 듯 하므로,

$$\mathrm{L}(\theta|\mathrm{x}) = \log \mathrm{P}(\mathrm{x}|\theta) = \sum log\ P(x|\theta)$$

log를 이용하여 계산 이점을 가져감
해당 식 L이 최대가 되는θ 를 구하는 것이 목표

Introduction - MAP

MAP는 베이지안 관점에서 구하는 모수 추정값이다. 정확히는 베이지안 관점에서는 모수 또한 확률 변수이기에 모수에 대한 분포를 찾는 것이다. 로 보는 것이 정확하다.

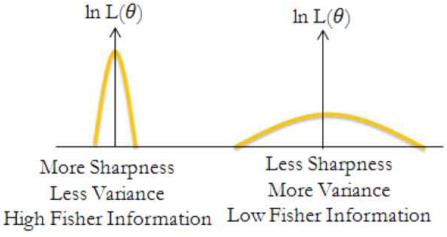
$$\hat{\theta}_{MAP} := \arg\max p(\theta|D) = \arg\max \frac{p(D|\theta)p(\theta)}{p(D)}$$
$$= \arg\max p(D|\theta)p(\theta) = \arg\max[\log p(D|\theta) + \log p(\theta)]$$

모수 또한 하나의 확률변수이기에 확률분포를 갖고 있다. 앞서 배운 MLE와의 차이점이라고 한다면 마지막 부분에서 log p(theta)가 붙는다는 것이다.

모수에 대한 분포 (prior distribution)은 다양하게 잡기는 하는데, 중요하게 봐야할 점은 우리의 관점을 넣을 수 있다는 것이다. 가령, 우리가 사전 분포가 정규분포를 따르는지 감마 분포를 따르는지 개인의 감각을 모수 추정에 넣을 수 있다는 것이다.

그렇기 때문에 이는 빈도주의보다 더 좋은 결과를 얻을 수 있지만 반대로 직관이 맞지 않게 된다면 잘못된 추정을 해버릴 수가 있다.

fisher information & multivariate norm-dist



첫 번째, 그림은 하나의 값을 중심으로 잘 모여있고,
두 번째, 그림은 하나의 값을 중심으로 퍼져있는 모습이다.
파라미터를 추정하려는 입장에서는 첫 번째 그래프가 더 좋은 모습이라고 판단할 것이다.
이를 수식적으로 표현한 것이 fisher information이라고 보면 된다.

fisher information & multivariate norm-dist

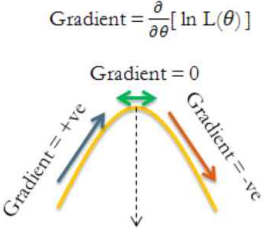


Figure: The gradient of log likelihood function is called ***score***

해당 기울기를 score function이라고 한다.

Fisher information(I)는 다음과 같이 쓸 수 있다.

$$I(\theta) = E_{x\sim p_\theta}[(s(\theta)-0)(s(\theta)-0)^T]$$
$$= E_{x\sim p_\theta}[(\sum_{i=1}^n \nabla_\theta \log p_\theta(x))(\sum_{i=1}^n \nabla_\theta \log p_\theta(x))^T]$$
$$= nE_{x_1\sim p_\theta}[\nabla_\theta \log p_\theta(x_1)\nabla_\theta \log p_\theta(x_1)^T]$$
$$= nI_1(\theta)$$

즉, fisher information은 어떻게 보면 기울기에 대한 분산으로 가운데에 몰려있는 우도함수가 더 정보가 많다는 것을 의미한다고 볼 수 있다. 이것 외에 해당 내용이 중요한 내용은 정규근사에 있다.

cramer-rao lower bound에 의해

$$Var(\hat{\theta}) \geq \frac{1}{nI(\theta)}$$

다음과 같은 성질을 알 수 있고, 여기에 CLT를 첨가하여 우리가 알고 있는 정규근사를 해보면

$$\sqrt{n}(\hat{\theta}-\theta) \to N(0, \frac{1}{I(\theta)})$$

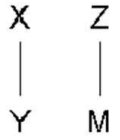
다음과 같이 변할 수 있다. 해당 내용을 통해 분포를 근사할 수 있다.

2

Missing data pattern

Missing data analysis

MAR, MCAR, MNAR



(a) MCAR

(b) MAR

(c) MNAR

MAR과 MCAR MNAR의 차이에 대해 알아보자

MAR (Missing At Random), MCAR (Missing At Completely Random), MNAR (Missing Not At Random)

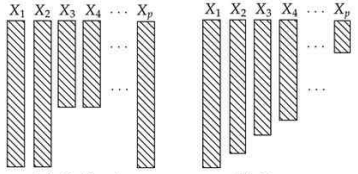
다음 그림에서 살펴보면, 관측이 다 되는 데이터 X, 관측이 될 수도 있고 아닐수도 있는 데이터 Y, 관측 여부를 알려주는 데이터 M, M에 대한 latent variable Z가 있다고 하자.

MCAR는 결측의 여부가 X, Y와는 관계성이 없는 것을 나타내면,

MAR은 결측의 여부가 X와만 관계성이 있다는 것

MNAR은 결측의 여부가 X, Y 모두와 관계성이 있는 경우를 나타낸다.

Missing data pattern



Multivariate pattern

각 변수들이 독립적으로 결측이 발생할 확률이 존재한다는 것이다. 즉, 결측 데이터가 특정 변수나 관측치에 집중되지 않고 전체 데이터셋에 걸쳐 불규칙하게 발생하는 것을 말한다. 다소 일반적인 형태로 볼 수 있다.

Monotone Pattern

한 개의 변수에 결측이 일어나게 된다면, 다른 변수에 결측이 될 가능성이 큰 경우다. 이는 종종 시간 순서가 있는 패널 데이터에서 나타나는 형태이다. 가령 예를 들면 사람의 맥박을 반복적으로 측정한다고 보자. 이렇게 된다면 특정 한 사람이 사망한 이후에는 결측이 일어날 것이고, 이는 다른 사람에게도 일어날 가능성이 크다.

General Pattern

이는 진짜 랜덤하게 발생하는 경우이다. 특정 패턴이나 순서 규칙이 없는 경우. 이것은 거의 분석이 불가능한 경우이다.

File matching pattern

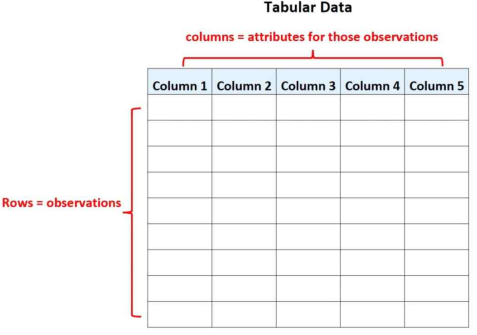
다음과 같이 결측이 같이 일어나고, 측정이 같이 되는 그러한 패턴이라고 보면 된다. 다음 그림에서 보면 X2와 X4가 같이 관측이 되며, X3과 Xp가 같이 관측됨을 알 수 있다.

3

Bayesian ML inference

Missing data analysis

MAR



$(n \times p)$ 형태의 missing이 존재하는 tabular dataset이 있다고 가정해보자.
이때, n은 샘플 수이며, p는 feature의 수라고 생각하면 된다.

각각의 값을 $Y = (y_{ij})$이라고 표현을 해보자.

즉, i번째 행에는 $(y_{i1}, y_{i2}, \ldots, y_{ip})$의 값이 존재하는 것이다.

결측치가 있는 데이터에서 모든 Y가 $f_Y(y|\theta)$에서 샘플링 된다는 것을 전제하였을 때,

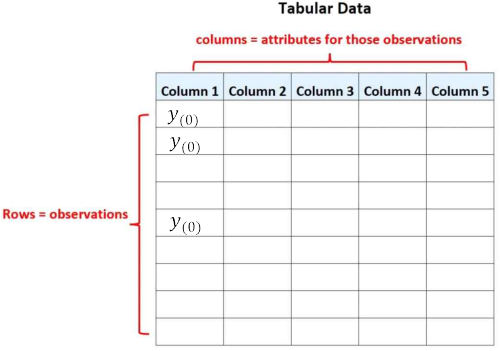
우리의 목표는 관측한 값들로 적절한 θ를 추정하는 것이다.

우리가 관측한 y값을 $\tilde{y}$로 표현하여 우도함수를 만들면,

$L_Y(\theta|\tilde{y}) = f_Y(\tilde{y}|\theta)$

앞에서 설명한 Likelihood의 개념을 생각해보면 알 수 있다.

MAR



이제 결측치라는 것을 인지하면서 모수에 대한 Likelihood를 생각해보자.

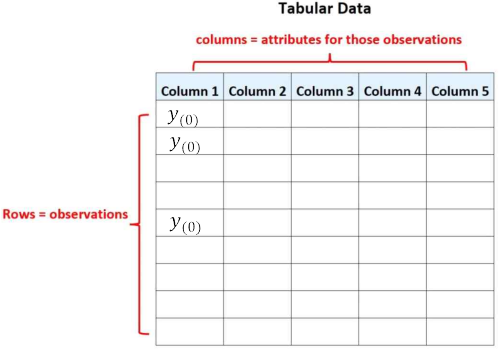
y값들 중에서 관측이 되지 않은 값들을 $y_{(0)}$

y값들 중에서 관측이 된 값들을 $\tilde{y}_{(1)}$

$$L_{ign}(\theta|\tilde{y}_{(1)}) = f_Y(\tilde{y}_{(1)}|\theta) = \int f_Y(\tilde{y}_{(1)}, y_{(0)}|\theta)dy_{(0)}$$

다음과 같이 표현할 수 있다. 해당 식을 그럴싸해 보이지만, 여기서의 문제점은 결측값이 발생한 결측 메커니즘에 대한 가정이 없다는 점입니다. 해당 매커니즘을 어떻게 넣어줄까

MAR



이제 결측치라는 것을 인지하면서 모수에 대한 Likelihood를 생각해보자.

y값들 중에서 관측이 되지 않은 값들을 $y_{(0)}$

y값들 중에서 관측이 된 값들을 $\tilde{y}_{(1)}$

$$L_{ign}(\theta|\tilde{y}_{(1)}) = f_Y(\tilde{y}_{(1)}|\theta) = \int f_Y(\tilde{y}_{(1)}, y_{(0)}|\theta)dy_{(0)}$$

다음과 같이 표현할 수 있다. 해당 식을 그럴싸해 보이지만, 여기서의 문제점은 결측값이 발생한 결측 메커니즘에 대한 가정이 없다는 점입니다. 해당 매커니즘을 어떻게 넣어줄까

MAR - indicator matrix

Tabular Data

columns = attributes for those observations

Column 1	Column 2	Column 3	Column 4	Column 5
0				
0				
1				
1				
0				

Rows = observations

indicator matirx $R = (r_{ij})$를 도입한다.

$r_{ij} = 1$이라면, 관측이 된 것이고, $r_{ij} = 0$이면 관측이 되지 않는 것이라고 판단한다.

해당 indicator matrix과 데이터 관계성이 있다고 판단을 한다.

$f_{R|Y}(R|Y, \phi)$가 존재한다고 생각하는 것이다.

(여기서 ϕ는 $f_{R|Y}$의 unknown parameter이다.)

$$L_{full}(\theta, \phi|\tilde{y}_{(1)}, \tilde{r}) = \int f_{R,Y}(\tilde{y}_{(1)}, \tilde{r}, y_{(0)}|\theta, \phi)dy_{(0)}$$
$$= \int f_Y(\tilde{y}_{(1)}, y_{(0)}|\theta, \phi) f_{R|Y}(\tilde{r}|\tilde{y}_{(1)}, y_{(0)}, \theta, \phi)dy_{(0)}$$

앞부분은 우리가 알 수 있는 것이니, 뒷 부분을 어떻게 구할 것이냐가 중요한 핵심이라고 보면 된다.

MAR - sufficient conditions for Bayesian inference

MAR은 Missing at random이다. Missing at random은 결측이 관측된 데이터에 의존하지만, 관측되지 않은 데이터와는 무관한 경우를 말한다. 즉, 관측데이터로 충분히 설명이 되는 상황이라는 것이다.

가령, 설문 조사에서 남성이 여성보다 특정 문항 답변을 자주 빠뜨린다고 가정을 해보자. 결측은 성별(관측된 데이터)에 의존하지만, 설문 문항의 답변 내용(관측되지 않는 데이터)에는 의존하지 않습니다.

베이지안적 관점에서 우리가 알고 싶은 모수 θ에 대한 추정을 어떻게 해야할까?

MAR의 정의는 다음과 같다.

$$f_{R|Y}(R = \tilde{r}|Y_{(1)} = \tilde{y}_{(1)}, Y_{(0)} = y_{(0)}, \phi)$$
$$= f_{R|Y}(R = \tilde{r}|Y_{(1)} = \tilde{y}_{(1)}, Y_{(0)} = y^*_{(0)}, \phi) \text{ for all } y_{(0)}, y^*_{(0)} \text{ and } \phi$$

다음 증명에서 볼 수 있듯이, 완전 관측된 데이터 $Y_{(0)}$이 어떻게 된다고 하더라도, R의 확률값이 안변한다는 것을 알 수 있다.

우리는 이러한 MAR 정의를 이용하여 어떻게 베이지안적 관점에서 모수를 추정하는 것인지 볼 것이다.

즉, 어떻게 어떻게 θ에 대한 posterior distribution을 구하게 되는지를 보게 된다는 것이다.

MAR - sufficient conditions for Bayesian inference

proof

$\pi(\theta, \phi) = \pi_1(\theta)\pi_2(\phi)$

둘은 서로 priori independent이기 때문에 다음과 같이 분리 할 수 있다.

$L_{full}(\theta, \phi|\tilde{y}_{(1)}, \tilde{r}) = L_{ign}(\theta|\tilde{y}_{(1)}) \times L_{rest}(\phi|\tilde{y}_{(1)}, \tilde{r})$ for all θ, ϕ

$L_{full}(\theta, \phi|\tilde{y}_{(1)}, \tilde{r}) = \int f_Y(\tilde{y}_{(1)}, y_{(0)}|\theta) f_{R|Y}(\tilde{r}|\tilde{y}_{(1)}, y_{(0)}, \phi) dy_{(0)}$

$= \int f_Y(\tilde{y}_{(1)}, y_{(0)}|\theta) dy_{(0)} \int f_{R|Y}(\tilde{r}|\tilde{y}_{(1)}, y_{(0)}, \phi) dy_{(0)}$

$= L_{ign}(\theta|\tilde{y}_{(1)}) L_{rest}(\phi|\tilde{y}_{(1)}, \tilde{r})$

고로, posterior distribution은

$p(\theta, \phi|\tilde{y}_{(1)}, \tilde{r}) \propto \pi(\theta, \phi) \times L_{full}(\theta, \phi|\tilde{y}_{(1)}, \tilde{r})$

$= [\pi_1(\theta) \times L_{ign}(\theta|\tilde{y}_{(1)}] \times [\pi_2(\phi) \times L_{rest}(\phi|\tilde{y}_{(1)}, \tilde{r})]$ for all θ, ϕ

고로,

$p(\theta, \phi|\tilde{y}_{(1)}, \tilde{r}) \propto \pi_1(\theta) \times L_{ign}(\theta|\tilde{y}_{(1)})$ for all θ

인 것을 구할 수 있다.

베이지안적 관점에서는 θ는 고정된 상수가 아닌, 확률 변수이다.

표본이 충분히 크고, θ에 대한 공간이 제한되지 않는다면, θ의 posterior distribution은 normal approximation에 의해

$(\theta|data) \sim N(\hat{\theta}, I^{-1}_{ign(\theta\theta)})$

다음과 같이 표현할 수 있다.

여기서, $I_{ign(\theta\theta)}$는 likelihood에 θ에 대해 두 번 미분한 것이다.

$I_{ign(\theta\theta)} = -D_{\theta\theta}(L_{ign}(\theta|\tilde{y}_{(1)}))|_{\theta=\hat{\theta}}$이다.

3

Frequentist ML inference

MAR - sufficient conditions for Frequentist ML inference

frequentist의 입장에서는표본을 통해 얻는 $\hat{\theta}$의 분포에 대해 관심이 많다. 이것은 그냥 단순하게 전에 구한 식에서 θ와 $\hat{\theta}$를 뒤집어서 생각해도 괜찮다.

즉, $(\hat{\theta}|data) \sim N(\theta, I^{-1}_{ign(\theta\theta)})$로 본다는 것이다.

$\hat{\theta}$에 대한 표본 분포를 살펴보기 위해서는 여러가지의 표본이 필요하다. 이 부분이 Frequentist와 Bayesian의 차이점이라고 볼 수 있다.

Frequentist는 여러 개의 샘플을 기반으로 우리가 알고 싶은 모수를 추정하고 싶어 한다. 하지만 Bayesian은 우리가 선험적으로 알고 있는 prior distribution을 기반으로 하기 때문에, 데이터의 개수가 적어도 prior distribution의 영향이 클 뿐이다.

이를 위해 frequentist의 입장에서는 여러 개의 샘플이 필요하다. 그렇기에 MAR보다 조금 더 가장 가정을 한다.

(해당 패턴을 MAAR이라 하자. Missing at always random)

$f_{R|Y}(R = r|Y_{(1)} = y_{(1)}, Y_{(0)} = y_{(0)}, \phi)$

$= f_{R|Y}(R = r|Y_{(1)} = y_{(1)}, Y_{(0)} = y^*_{(0)}, \phi)$ for all $r, y_{(1)}, y_{(0)}, y^*_{(0)}, \phi$

그러니까 해당 내용은 n개의 샘플 테이블이 전부 다 MAR이라는 것이다.

$$\begin{pmatrix} \theta - \hat{\theta} \\ \phi - \hat{\phi} \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix} \begin{pmatrix} I_{\mathrm{full}(\theta\theta)} & I_{\mathrm{full}(\theta\phi)} \\ I_{\mathrm{full}(\theta\phi)} & I_{\mathrm{full}(\phi\phi)} \end{pmatrix}^{-1} \right]$$

frequentist의 입장에서 θ와 ϕ를 추정하는 것에 대해 보면 다음과 같다.

여기서 $(\hat{\theta}, \hat{\phi})$는 (θ, ϕ)의 ML estimate이다.

$I_{full(\theta\theta)} = -D_{\theta\theta}(\log L_{full}(\theta, \phi|\tilde{y}_{(1)}, \tilde{r})|_{\theta=\hat{\theta},\phi=\hat{\phi}}$

$I_{full(\theta\phi)} = -D_{\theta\phi}(\log L_{full}(\theta, \phi|\tilde{y}_{(1)}, \tilde{r}))|_{\theta=\hat{\theta},\phi=\hat{\phi}}$

$I_{full(\phi\phi)} = -D_{\phi\phi}(\log L_{full}(\theta, \phi|\tilde{y}_{(1)}, \tilde{r})|_{\theta=\hat{\theta},\phi=\hat{\phi}}$

Example MAR & MAAR

우리가 두 그룹의 평균 차이를 검정하고자 한다고 하자.

$(y_i, x_i), i = 1, ..., n$이 있고, 이때, y는 outcome이고, $x_i = j$는 그룹을 나타낸다고 하자. (간단하게 1, 2 그룹 2개가 있다고 해보자.)

결측치가 존재하지 않을 때 검정을 생각해보면,

$(y_i|x_i = j, \theta) \sim_{ind} N(\mu_j, \sigma^2)$을 따르게 된다고 보며, $\theta = (\mu_1, \mu_2, \sigma^2)$

두 그룹의 평균차이를 $\delta = \mu_2 - \mu_1$이라고 해보면

$t = ((\bar{y}_2 - \bar{y}_1) - \delta)/(s\sqrt{1/n_1 + 1/n_2})$

다음과 같이 t 값을 구할 수 있다.

이렇게 하여 δ의 95% 신뢰구간을 구하게 되면,

$I_{0.95}(\delta) = \bar{y}_2 - \bar{y}_1 \pm t_{v,0.975}(s\sqrt{1/n_1 + 1/n_2})$

다음과 같이 두 평균 차이의 신뢰구간을 구할 수 있다.

결측치가 있는 상황을 살펴보자.

X는 항상 관측이 되는데, Y는 간혹가다가 결측치가 있다고 하자.

Y가 관측될 확률에 대해서는 다음과 같이 표현할 수 있다.

$$\Pr(r_i = 1|x_i, y_i, \phi) = b_1(x_i, \phi)$$

이 결측은 MAR과 MAAR를 따르며 그 이유는 결측이 결측이 될 수도 있고 아닐 수도 있는 y에는 의존하지 않고, 항상 관측되는 x에는 의존하기 때문이다.

하지만, b의 함수가 다음과 같이 되면 어떨까?

$$b(x_i, y_i, \phi) = \begin{cases} 1, \text{ if } x_i = 1 \\ 1, \text{ if } x_i = 2 \text{ and } y_i \le \phi, \\ 0, \text{ if } x_i = 2 \text{ and } y_i > \phi. \end{cases}$$

이렇게 된다면, x=2일 때는 랜덤하게 결측이 아니다. 고로 이는 MAR은 되지만, MAAR은 아니게 된다.

Example MAR & MAAR

MAR와 MAAR의 정의에 대해 다시 살펴보면,

MAAR은

$$f_{R|Y}(R = r|Y_{(1)} = y_{(1)}, Y_{(0)} = y_{(0)}, \phi)$$
$$= f_{R|Y}(R = r|Y_{(1)} = y_{(1)}, Y_{(0)} = y^*_{(0)}, \phi) \text{ for all } r, y_{(1)}, y_{(0)}, y^*_{(0)}, \phi$$

MAR은

$$f_{R|Y}(R = \tilde{r}|Y_{(1)} = \tilde{y}_{(1)}, Y_{(0)} = y_{(0)}, \phi)$$
$$= f_{R|Y}(R = \tilde{r}|Y_{(1)} = \tilde{y}_{(1)}, Y_{(0)} = y^*_{(0)}, \phi) \text{ for all } y_{(0)}, y^*_{(0)} \text{ and } \phi$$

이렇게 되어 있다.

즉, MAR입장에서는 어떠한 데이터가 해당 조건을 만족하면 되는 것인데, 이를 만족하는 데이터가 위에 상황에서도 존재한다.

하지만, MAAR 입장에서는 모든 데이터가 해당 조건을 만족해야하는 것인데, 저렇게 된다면, x=2인 상황에서는 만족을 못하게 된다.

frequentist의 입장에서 보았을 때는 $\bar{y}_2$는 μ_2에 대해서 편향된 추정치이며, 해당 편향 자체는 ϕ가 무엇인지에 따라 달라집니다. ϕ가 증가하게 되면 결측값이 줄어들기 때문에 편향이 감소할 것이고, 반대로 감소하게 되면 편향이 커지게 됩니다. 고로, Frequentist의 입장에서 보았을 때는 해당 데이터가 random하게 missing되었다라고 보는 것은 어렵습니다.

Bayesian의 입장에서 보았을 때는 θ와 ϕ 자체가 서로 priori independent를 이룬다는 가정을 갖고 있기 때문에, 우리가 궁금한 θ와 ϕ 자체는 서로 관계가 없는 사이입니다. 그렇기에 베이지안 입장에서는 이게 random하게 missing 되었다라고 볼 수 있는 것이죠

4

Weights complete-case Analysis

Cell weighting

해당 방식은 앞에서 우도함수를 이용한 분포를 구하는 방식과는 다르게 샘플링 가중치를 이용하여 분석을 하는 경우이다.

가령 우리가 쉽게 생각할 수 있는 예로는 사전조사이다.

Sample

	B1	B2	B3	Total
A1	20	40	40	100
A2	50	140	310	500
A3	100	50	50	200
A4	30	100	70	200
Total	200	330	470	1000

다음과 같이 샘플 개수가 잡혔다고 보자. 하지만, 우리가 목표로 삼았던 샘플 수는 다음과 같다고 보자.

Target

	B1	B2	B3	Total
A1	80	40	55	175
A2	60	150	340	550
A3	170	60	200	430
A4	55	165	125	345
Total	365	415	720	1500

Cell weighting

Cell weighting

	B1	B2	B3
A1	4.00	1.00	1.38
A2	1.20	1.07	1.10
A3	1.70	1.20	4.00
A4	1.83	1.65	1.79

다음과 같이 우리가 얻은 표본 결과에 weight 4의 가중치를 추가로 넣어 줘서 계산을 해주는 것이다.

예를 들면, 우리나라 사람들의 평균 키를 구하고 싶은데

데이터 값이 $y_{1,1}, y_{1,2}, \ldots, y_{1,20}$이렇게 된 것이다.

그렇게 되면 샘플의 평균 데이터 값은 $\bar{y}_1$이 된다.

이를 그대로 이용하지 않고, $4\bar{y}_1$으로 넣어줘서 해당 값의 가중치를 세게 넣어주는 방식이다.

Calibration

calibration 기법에 경우 우리가 benchmark information이 있는 경우에 사용을 한다.

우리가 어떠한 샘플을 갖고 있고 해당 샘플의 평균이 μ라는 것을 알고 있다고 가정을 해보자.

n개의 데이터가 존재하고 우리는 결측으로 인한 편향을 맞춰주려는 것이라고 보면 된다.

즉, $\frac{1}{n}\sum w_i x_i = \mu$에 맞춰주어야 한다는 것이다.

단순하게 맞춰줄 순 있지만, 해당 방식에는 몇몇 개의 제약조건이 달리게 된다.

먼저, weights의 합이 1이 되도록 맞춰주어야 한다. 또한 몇몇 값들이 튀는 것을 방지해주기 위해 inital weights와 비슷하게 나와야한다. 대게 inital weights는 계산의 편의를 위해 exp(1)으로 잡는다.

initial weights와 비슷하다는 것은 KL divergence를 통해 구한다.

$$Q = -\sum w_i log \frac{w_i}{d_i}$$

맞춰주어야 하는 목적함수가 있고 몇몇 개의 제약조건이 존재하기 때문에 이는 dual problem으로 해결한다.

이를 라그랑지안 식으로 표현을 하면,

$$L = -\sum w_i \log \frac{w_i}{d_i} + \lambda_1(\sum w_i - 1) + \lambda_2(\sum w_i x_i - \mu)$$

다음과 같이 된다.

이를 weights에 대해 미분을 해보면,

$$\frac{\partial L(w_i)}{\partial w_i} = -\log(w_i) + \lambda_1 + \lambda_2 x_i$$

와 같이 된다. 이를 정리해보면,

$$w_i \propto \exp(-\lambda^T x_i)$$

가 나와 해당 weights를 이용하여 적절한 가중을 해준다.

이 뒷부분의 dual problem 해결은 closed form으로 해결이 되지 않기에 보편적으로 newton-raphson이라던지 gradient descent라던지 하는 방식으로 최적값을 찾게 된다.

과제 - 예시

Q

우리는 오늘 MAR에 대해 배웠다.

$Y_i \sim exp(\theta)$ 의 값을 가지는 데이터가 있다고 해보자.

이때, $Y_{obs} = (y_1, ..., y_n)^T$

$Y_{miss} = (y_{n+1}, ..., y_p)^T$

이렇게 되어있다.

이를 통해 모든 데이터를 통해 구한 θ 추정값과 결측치만을 갖고 구한 θ 추정값에 대해서 생각해보세요

A

$f(Y|\theta) = \frac{1}{\theta^p} \exp(-\sum^p(\frac{y_i}{\theta}))$

$f(Y_{obs}|\theta) = \frac{1}{\theta^n} \exp(-\sum^n(\frac{y_i}{\theta}))$

$R = (1, ...1, 0, ..., 0)$

이렇게 됨을 알 수 있다.

ϕ를 관측될 확률이라고 보고 우리 해당 확률이 베르누이를 따른다고 하자 그렇게 된다면,

$f(R|Y, \phi) = \phi^n(1-\phi)^{(p-n)}$

$L(\theta, \phi|Y_{obs}, R) = \phi^n(1-\phi)^{(p-n)} \frac{1}{\theta^n} \exp(-\sum^n(\frac{y_i}{\theta}))$

가 될 것이고 이를 통해 θ의 추정값을 알 수 있게 된다.

$\hat{\theta} = \frac{\sum_{i=1}^n y_i}{n}$

감사합니다