베이즈데이터분석 / 이재용 교수

07강

정규모형

한국방송통신대학교

목차

목차

정규모형와 켤레사전분포

모형

$x_1, \cdots, x_n | \mu, \sigma^2 \overset{i.i.d.}{\sim} N(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 = \frac{1}{\lambda} > 0$는 둘 다 모르는 값이다.

사전분포

$$\mu | \sigma^2 \sim N(\mu_0, \frac{\lambda^{-1}}{k_0})$$

$$\sigma^2 \sim \text{Inv} - Ga(\frac{v_0}{2}, \frac{v_0}{2}\sigma_0^2) = \text{Inv} - \boldsymbol{X}^2(v_0, \sigma_0^2)$$

혹은 $\lambda = \frac{1}{\sigma^2} \sim Ga(\frac{v_0}{2}, \frac{v_0}{2}\sigma_0^2) = \boldsymbol{X}^2(v_0, \sigma_0^2)$,

여기서, $\mu_0 \in \mathbb{R}, k_0, v_0, \sigma_0^2 > 0$.

정규모형와 켤레사전분포

$$\begin{cases} x_1, \cdots, x_n \mid \mu, \sigma^2 \overset{iid}{\sim} N(\mu, \sigma^2 = \tfrac{1}{\lambda}) \\ \lambda \sim Ga\left(\frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_0^2\right) \\ \mu \mid \lambda \sim N\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right) \end{cases} , \quad \sigma^2 = \frac{1}{\lambda}$$

$$\pi(\underline{\mu}, \underline{\lambda} \mid \underset{(x_1,\cdots,x_n)}{x}) \propto \pi(\mu, \lambda) \times \prod_{i=1}^{n} f(x_i \mid \mu, \lambda)$$

$$= Ga\left(\lambda \mid \frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_0^2\right) N\left(\mu \mid \mu_0, \frac{\sigma^2}{\kappa_0}\right) \times \prod_{i=1}^{n} N(x_i \mid \mu, \sigma^2)$$

$$= \frac{\left(\frac{\nu_0}{2}\sigma_0^2\right)^{\frac{\nu_0}{2}}}{\Gamma\left(\frac{\nu_0}{2}\right)} \cdot \lambda^{\frac{\nu_0}{2}-1} \cdot e^{-\frac{\nu_0}{2}\sigma_0^2 \cdot \lambda} \times \frac{\lambda^{\frac{1}{2}} \cdot \sqrt{\kappa_0}}{\sqrt{2\pi}} e^{-\frac{\kappa_0 \lambda}{2} \cdot (\mu - \mu_0)^2}$$

$$\times \prod_{i=1}^{n} \frac{\lambda^{+\frac{1}{2}}}{\sqrt{2\pi}} e^{-\frac{\lambda}{2}(x_i - \mu)^2}$$

$$\propto \lambda^{\frac{\overset{\nu_n}{\nu_0 + n} + 1}{2} - 1} \times e^{-\frac{\lambda}{2}\left[\nu_0 \cdot \sigma_0^2 + \kappa_0(\mu - \mu_0)^2 + (n-1)s^2 + n(\bar{x} - \mu)^2\right]}$$

$$\sum (x_i - \mu)^2 = \sum (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2$$
$$= (n-1) \cdot s^2 + n(\bar{x} - \mu)^2$$

$$s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

정규모형와 켤레사전분포

$$\pi(\mu,\lambda|x) \propto \lambda^{\frac{\nu_n+1}{2}-1}\cdot e^{-\frac{\lambda}{2}\left[k_n(\mu-\mu_n)^2+\nu_n\cdot\sigma_n^2\right]}$$

$$\mu_n = \frac{n\cdot\bar{x}+k_0\cdot\mu_0}{n+k_0}$$

$$k_n = n+k_0$$

$$\nu_n = n+\nu_0$$

$$\sigma_n^2 = \frac{1}{\nu_n}\left[\frac{n\cdot k_0}{n+k_0}(\bar{x}-\mu_0)^2+(n-1)s^2+\nu_0\cdot\sigma_0^2\right]$$

$$\pi(\mu|\lambda,x) \propto e^{-\frac{\lambda\cdot k_n}{2}(\mu-\mu_n)^2}$$

$$\propto N\left(\mu \,\middle|\, \mu_n, \frac{1}{\lambda\cdot k_n}=\frac{\sigma^2}{k_n}\right)$$

$$\pi(\lambda|\mu,x) \propto Ga\left(\lambda \,\middle|\, \frac{\nu_n+1}{2}, \frac{1}{2}\left[k_n(\mu-\mu_n)^2+\nu_n\cdot\sigma_n^2\right]\right)$$

$$\pi(\lambda|x) = \int \pi(\mu,\lambda|x)\,d\mu$$

$$= \lambda^{\frac{\nu_n+1}{2}-1}\cdot e^{-\frac{\lambda}{2}\nu_n\cdot\sigma_n^2} \times \boxed{\int e^{-\frac{\lambda k_n}{2}(\mu-\mu_n)^2}\,d\mu}$$

$$\sqrt{2\pi}\cdot\frac{1}{\sqrt{\lambda\cdot k_n}}$$

$$\propto \lambda^{\frac{\nu_n}{2}-1}\cdot e^{-\frac{\nu_n\cdot\sigma_n^2}{2}\cdot\lambda}$$

$$\propto Ga\left(\lambda \,\middle|\, \frac{\nu_n}{2}, \frac{\nu_n}{2}\sigma_n^2\right)$$

$$\pi(\mu|x) = \int \pi(\mu,\lambda|x)\,d\lambda$$

$$= \cdots$$

사후분포

조건부사후분포 $\pi(\mu|\lambda\ \ ,x)$

$$\mu|\lambda,x \sim N(\mu_n, \frac{1}{k_n \cdot \lambda})$$

조건부사후분포 $\pi(\lambda|\mu,x)$

$$\lambda|\mu,x \sim Ga(\frac{v_n+1}{2}, \frac{1}{2}[k_n(\mu-\mu_n)^2 + \sigma_n^2 \cdot v_n\ ])$$

모형

$x_1,\cdots,x_n|\mu,\sigma^2 \overset{iid}{\sim} N(\mu,\sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0$는 둘다모르는값이다.

사전분포

$\mu|\sigma^2 \sim N(\mu_0, \frac{\sigma^2}{k_0})$

$\sigma^2 \sim \text{Inv}-Ga(\frac{v_0}{2}, \frac{v_0}{2}\sigma_0^2) = Inv-\boldsymbol{\chi}^2(v_0, \sigma_0^2)$

혹은 $\lambda = \frac{1}{\sigma^2} \sim Ga\left(\frac{v_0}{2}, \frac{v_0}{2}\sigma_0^2\right) = \boldsymbol{\chi}^2(v_0, \sigma_0^2)$,

여기서, $\mu_0 \in \mathbb{R}, k_0, v_0, \sigma_0^2 > 0$.

사후분포

주변 사후분포 $\pi(\lambda|x)$

$$\lambda|x\sim Ga(\frac{v_n}{2},\frac{v_n}{2}\sigma_n^2).$$

주변 사후분포 $\pi(\mu|x)$

$$\mu|x\sim t_{v_n}(\mu_n,\frac{\sigma_n^2}{k_n}).$$

모형

$x_1,\cdots,x_n|\mu,\sigma^2 \overset{iid}{\sim} N(\mu,\sigma^2), \mu\in\mathbb{R}, \sigma^2>0$는 둘다모르는값이다.

사전분포

$\mu|\sigma^2\sim N(\mu_0,\frac{\sigma^2}{k_0})$

$\sigma^2\sim \mathrm{Inv}-Ga(\frac{v_0}{2},\frac{v_0}{2}\sigma_0^2) = Inv-\chi^2(v_0,\sigma_0^2)$

혹은 $\lambda = \frac{1}{\sigma^2}\sim Ga\left(\frac{v_0}{2},\frac{v_0}{2}\sigma_0^2\right) = \chi^2(v_0,\sigma_0^2)$,

여기서, $\mu_0\in\mathbb{R}$, $k_0, v_0, \sigma_0^2>0$.

사후분포

사후분포의 파라미터들은 다음과 같이 정의된다.

$$\mu_n = \frac{k_0 \cdot \mu_0 + n \cdot \bar{x}}{k_0 + n}$$

$$k_n = k_0 + n$$

$$v_n = v_0 + n$$

$$\sigma_n^2 = \frac{1}{v_n}\left[\frac{k_0 \cdot n}{k_n}(\bar{x} - \mu_0)^2 + (n-1)s^2 + v_0 \cdot \sigma_0^2\right].$$

모형

$x_1, \cdots, x_n | \mu, \sigma^2 \overset{iid}{\sim} N(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0$는 둘다 모르는 값이다.

사전분포

$\mu | \sigma^2 \sim N(\mu_0, \frac{\sigma^2}{k_0})$

$\sigma^2 \sim \mathrm{Inv} - Ga(\frac{v_0}{2}, \frac{v_0}{2}\sigma_0^2) = Inv - \chi^2(v_0, \sigma_0^2)$

혹은 $\lambda = \frac{1}{\sigma^2} \sim Ga\left(\frac{v_0}{2}, \frac{v_0}{2}\sigma_0^2\right) = \chi^2(v_0, \sigma_0^2)$,

여기서, $\mu_0 \in \mathbb{R}$, $k_0, v_0, \sigma_0^2 > 0$.

베이즈 추정량과 신용집합

베이즈 추정량들

$$\hat{\mu}^B = E(\mu|x) = \mu_n$$

$$\hat{\sigma}^{2B} = E(\sigma^2|x) = \frac{v}{v-2} \cdot \frac{\sigma_n{}^2}{k_n}$$

$$\hat{\lambda}^B = E(\lambda|x) = \frac{1}{\sigma_0^2}$$

μ에 대한 $100(1-\alpha)\%$ 신용집합

$$\mu_n \pm t_{\frac{\alpha}{2}}(v_n) \cdot \frac{\sigma_n}{\sqrt{k_n}}.$$

주변 사후분포 $\pi(\lambda|x)$

$$\lambda|x \sim Ga\left(\frac{v_n}{2}, \frac{v_n}{2}\sigma_n^2\right).$$

주변 사후분포 $\pi(\mu|x)$

$$\mu|x \sim t_{v_n}\left(\mu_n, \frac{\sigma_n^2}{k_n}\right).$$

베이즈 추정량과 신용집합

λ에 대한 $100(1-\alpha)\%$ 신용집합

$$\left[Ga_{1-\frac{\alpha}{2}}\left(\frac{v_n}{2}, \frac{v_n}{2}\sigma_n^2\right), Ga_{\frac{\alpha}{2}}\left(\frac{v_n}{2}, \frac{v_n}{2}\sigma_n^2\right)\right]$$

주변 사후분포 $\pi(\lambda|x)$

$$\lambda|x \sim Ga\left(\frac{v_n}{2}, \frac{v_n}{2}\sigma_n^2\right).$$

주변 사후분포 $\pi(\mu|x)$

$$\mu|x \sim t_{v_n}\left(\mu_n, \frac{\sigma_n^2}{k_n}\right).$$

몬테 카를로 방법을 이용한 베이즈 추론

- $\pi(\mu,\lambda|x)$로 부터 사후 표본을 다음과 같이 추출할 수 있다. $j = 1,2,\ldots,m$에 대해서

$$\lambda_j \sim \pi(\lambda|x) = Ga(\frac{v_n}{2},\frac{v_n}{2}\sigma_n^2)$$

$$\mu_j \sim \pi\left(\mu\middle|\lambda_j,x\right) = N(\mu_n,\frac{1}{\kappa_n\lambda_j})$$

를 추출한다.

$$(\mu_1,\lambda_1),\ldots,(\mu_m,\lambda_m) \overset{i,i,d.}{\sim} \pi(\mu,\lambda|x)$$

$\sigma_j^2 = \dfrac{1}{\lambda_j}$로 변환하면

$$\left(\mu_j,\sigma_j^2\right) \overset{i,i,d.}{\sim} \pi(\mu,\sigma^2|x)$$

몬테 카를로 방법을 이용한 베이즈 추론

- (베이즈 추정량)

$$\hat{\mu}^{B} = \frac{1}{m}\sum_{j=1}^{m}\mu_j$$

$$\hat{\sigma}^{2,B} = \frac{1}{m}\sum_{j=1}^{m}\sigma_j^2$$

- (100(1 – α)% 신용구간)

$$[\mu_{(m\frac{\alpha}{2})}, \mu_{(m(1-\frac{\alpha}{2}))}]$$

$$[\sigma^2{}_{(m\frac{\alpha}{2})}, \sigma^2{}_{(m(1-\frac{\alpha}{2}))}]$$

목차

자료의 설명

- 자료의 변수는 총 10개로 다음과 같다.
  순번, 측정 일자, 가슴 둘레, 소매길이, 신장, 허리, 샅높이, 머리 둘레, 발 길이, 몸무게이다.
- 모든 길이의 단위는 센티미터이고 몸무게의 단위는 킬로그램이다.

문제

- army-physical.csv 자료를 읽어들인다.
- 자료의 변수는 몇 개이고, 개수는 몇 개인가?
- 각 변수의 요약통계량을 구하시오. 육군 키의 평균, 최대값과 최소값은 얼마인가?
- 각 변수의 히스토그램을 그리시오.
- 모든 두 변수간의 산점도를 그리시오.
- 두 개의 변수들간의 공분산과 상관계수를 구하시오.
  상관계수가 가장 큰 변수 둘은 무엇인가?

육군신체측정 정보

자료 읽기

```
army = read.csv("army-physical.csv", header=T, sep=",")
head(army)
str(army)
```

1변량 자료 탐색

```
library(dplyr)
library(ggplot2)
library(Hmisc)
dim(army)
summary(army)
hist(army)
army %>% select(height, weight, bust, sleeve) %>% hist
```

2변량 자료 탐색

```
library(GGally)
army %>% ggpairs
army %>% select(height, weight, bust, sleeve) %>% ggpairs
cov(army)
cor(army)
```

예. 대한민국 육군 남성의 키 분포

- 대한민국 육군들의 키의 평균의 사후분포를 구하고자 한다.
- 모형 : $x_1, x_2, \ldots, x_n | \mu \sim N(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0.$
- 사전분포

$$\mu | \sigma^2 \sim N(\mu_0, \frac{\sigma^2}{k_0})$$

$$\sigma^2 \sim Inv \sim \mathcal{X}^2(v_0, \sigma_0^2)$$

혹은 $\lambda = \frac{1}{\sigma^2} \sim \mathcal{X}^2(v_0, \sigma_0^2)$

$$\mu_0 = 170$$

$$\kappa_0 = 0.1$$

$$\sigma_0^2 = 3.5$$

$$v_0 = 0.1 .$$

예. 대한민국 육군 남성의 키 분포

사후분포

$$\lambda|x \sim Ga(\frac{v_n}{2}, \frac{v_n}{2}\sigma_n^2)$$

$$\mu|x \sim t_{v_n}(\mu_n, \frac{\sigma_n^2}{k_n})$$

여기서

$$\mu_n = \frac{k_0 \cdot \mu_0 + n \cdot \bar{x}}{k_0 + n}$$

$$k_n = k_0 + n$$

$$v_n = v_0 + n$$

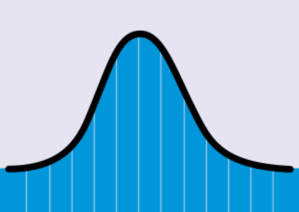
$$\sigma_n^2 = \frac{1}{v_n}[\frac{k_0 \cdot n}{k_n}(\bar{x} - \mu_0)^2 + (n-1)s^2 + v_0 \cdot \sigma_0^2].$$

예. 대한민국 육군 남성의 키 분포

문제

1. μ와 σ^2의 사후평균과 95% 신용구간을 구하시오.

목차



정규모형과 무정보사전분포(noninformative prior)

모형

$$x_1, \ldots, x_n | \mu, \sigma^2 \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$$

사전분포

$$\pi(\mu, \sigma^2) d\mu d\sigma^2 = \frac{1}{\sigma^2} d\mu d\sigma^2$$

혹은

$$\pi(\mu, \lambda) d\mu d\lambda = \frac{1}{\lambda} d\mu d\lambda$$

사후분포

완전 조건부 사후분포(full conditional posteriors)

$$\mu|\lambda, x \sim N(\bar{x}, \frac{1}{n \cdot \lambda})$$

$$\lambda|\mu, x \sim Ga(\frac{n}{2}, \frac{1}{2}(n-1) \cdot s^2).$$

주변 사후분포

$$\mu|x \sim t_{n-1}\left(\bar{x}, \frac{s^2}{n}\right), \lambda|x \sim Ga(\frac{n-1}{2}, \frac{n-1}{2} s^2)$$

μ에 대한 $100(1 - \alpha)\%$ 신용구간

$$\bar{x} \pm t_{\frac{\alpha}{2}(n-1)} \cdot \frac{s}{\sqrt{n}}$$

다음시간

08강

랜덤숫자발생