Daeyoung Lim

Grad School

2015, early

1. (X, Y)가 이변량 정규분포 $\mathcal{N}_2\left(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho\right)$를 따른다고 하자. 즉, (X, Y)의 결합확률분포는

$$p(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\frac{(x-\mu_1)}{\sigma_1}\frac{(y-\mu_2)}{\sigma_2} + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right\}\right] \tag{1}$$

이다.

(a) (10점) $Y = y$로 주어졌을 때, X의 조건부 분포가

$$\mathcal{N}_1\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y-\mu_2), \sigma_1^2\left(1-\rho^2\right)\right) \tag{2}$$

이 됨을 보여라. 단, $\mathcal{N}_1\left(\mu, \sigma^2\right)$은 평균이 μ이고 표준편차가 σ인 일변량 정규분포를 나타낸다. (힌트: Y의 주변확률분포는 $\mathcal{N}_1\left(\mu_2, \sigma_2^2\right)$이다.)

(b) (5점) Y를 사용할 때, X에 대한 best predictor를 쓰시오.

(c) (10점) $\mathrm{Var}(X), \mathrm{E}(\mathrm{Var}(X \mid Y))$ 및 $\mathrm{Var}(\mathrm{E}(X \mid Y))$를 구하고 이들 간의 관계식을 쓰시오.

Solution:

(a) 먼저 일반적인 다변량 정규분포에서 해보자. 즉 어떤 random vector $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$를 때를 때 다음의 파티션을 생각할 수 있다. (모든 다변량 정규분포들의 조건부 분포들은 정규분포라는 정리가 있다. 증명은 생략한다.)

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 & \boldsymbol{\mu}_2 \end{bmatrix}' \tag{3}$$

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 \end{bmatrix}' \tag{4}$$

그리고 공분산행렬도 다음처럼 구획할 수 있다.

$$\begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \tag{5}$$

이때 다음을 정의하자. $\mathbf{z} = \mathbf{x}_1 + \mathbf{A}\mathbf{x}_2$이고 여기서 $\mathbf{A} = -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$. 그러면 다음을 알 수 있다.

$$\begin{aligned} \mathrm{Cov}(\mathbf{z}, \mathbf{x}_2) &= \mathrm{Cov}(\mathbf{x}_1, \mathbf{x}_2) + \mathrm{Cov}(\mathbf{A}\mathbf{x}_2, \mathbf{x}_2) & (6) \\ &= \boldsymbol{\Sigma}_{12} + \mathbf{A}\mathrm{Var}(\mathbf{x}_2) & (7) \\ &= \boldsymbol{\Sigma}_{12} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{22} & (8) \\ &= 0 & (9) \end{aligned}$$

그러므로 $\mathbf{z}$와 $\mathbf{x}_2$는 uncorrelated이며, 다변량 정규분포는 uncorrelated가 독립임을 의미한다. $\mathrm{E}(\mathbf{z}) = \boldsymbol{\mu}_1 + \mathbf{A}\boldsymbol{\mu}_2$임은 자명하며, 따라서 다음을 유도할 수 있다.

$$\begin{aligned}
\mathrm{E}(\mathbf{x}_1 \mid \mathbf{x}_2) &= \mathrm{E}(\mathbf{z} - \mathbf{A}\mathbf{x}_2 \mid \mathbf{x}_2) && (10)\\
&= \mathrm{E}(\mathbf{z} \mid \mathbf{x}_2) - \mathrm{E}(\mathbf{A}\mathbf{x}_2 \mid \mathbf{x}_2) && (11)\\
&= \mathrm{E}(\mathbf{z}) - \mathbf{A}\mathbf{x}_2 && (12)\\
&= \boldsymbol{\mu}_1 + \mathbf{A}(\boldsymbol{\mu}_2 - \mathbf{x}_2) && (13)\\
&= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) && (14)
\end{aligned}$$

그리고 조건부 공분산 행렬은 다음과 같다.

$$\begin{aligned}
\mathrm{Var}(\mathbf{x}_1 \mid \mathbf{x}_2) &= \mathrm{Var}(\mathbf{z} - \mathbf{A}\mathbf{x}_2 \mid \mathbf{x}_2) && (15)\\
&= \mathrm{Var}(\mathbf{z} \mid \mathbf{x}_2) + \mathrm{Var}(\mathbf{A}\mathbf{x}_2 \mid \mathbf{x}_2) - \mathbf{A}\mathrm{Cov}(\mathbf{z}, -\mathbf{x}_2) - \mathrm{Cov}(\mathbf{z}, -\mathbf{x}_2)\mathbf{A}' && (16)\\
&= \mathrm{Var}(\mathbf{z} \mid \mathbf{x}_2) && (17)\\
&= \mathrm{Var}(\mathbf{z}) && (18)
\end{aligned}$$

따라서

$$\begin{aligned}
\mathrm{Var}(\mathbf{x}_1 \mid \mathbf{x}_2) &= \mathrm{Var}(\mathbf{z}) && (19)\\
&= \mathrm{Var}(\mathbf{x}_1 + \mathbf{A}\mathbf{x}_2) && (20)\\
&= \mathrm{Var}(\mathbf{x}_1) + \mathbf{A}\mathrm{Var}(\mathbf{x}_2)\mathbf{A}' + \mathbf{A}\mathrm{Cov}(\mathbf{x}_1, \mathbf{x}_2) + \mathrm{Cov}(\mathbf{x}_2, \mathbf{x}_1)\mathbf{A}' && (21)\\
&= \boldsymbol{\Sigma}_{11} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{22}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} - 2\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} && (22)\\
&= \boldsymbol{\Sigma}_{11} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} - 2\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} && (23)\\
&= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} && (24)
\end{aligned}$$

조건부 공분산행렬의 꼴을 $\boldsymbol{\Sigma}$의 $\boldsymbol{\Sigma}_{22}$에 대한 *Schur complement*라고 한다.

이 문제에서처럼 이변량으로 축소하면 $\mathbf{X} = \begin{bmatrix} X & Y \end{bmatrix}'$, $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 & \mu_2 \end{bmatrix}'$, 그리고 공분산행렬이 다음과 같아진다.

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \tag{25}$$

그러므로

$$\mathrm{E}\left(X \,|\, Y=y\right) = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}\left(y-\mu_2\right) \tag{26}$$

$$\mathrm{Var}\left(X \,|\, Y=y\right) = \sigma_1^2 - \rho^2\sigma_1^2 \tag{27}$$

$$= \sigma_1^2\left(1-\rho^2\right) \tag{28}$$

(b) MSE의 관점에서 Y의 정보를 알고 있을 때, 즉 Y가 주어졌을 때 MSE를 최소화하는 Y의 함수를 찾고자 한다. 즉

$$g^* = \underset{g}{\operatorname{argmin}}\, \mathrm{E}\left(\left(X - g\left(Y\right)\right)^2 \middle| Y\right) \tag{29}$$

그럴 때 $g^* = \mathrm{E}\left(X \,|\, Y\right)$가 된다. 따라서 위와 같은 이변량 정규분포라면 조건부 평균이 가장 좋은 예측값이다.

$$g^* = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}\left(Y-\mu_2\right) \tag{30}$$

(c) *Variance decomposition*이다. 조건부 두 개 더하면 X의 분산이 나온다.

$$\mathrm{E}\left(\mathrm{Var}\left(X \,|\, Y\right)\right) = \sigma_1^2\left(1-\rho^2\right) \tag{31}$$

$$\mathrm{Var}\left(\mathrm{E}\left(X \,|\, Y\right)\right) = \rho^2\frac{\sigma_1^2}{\sigma_2^2}\mathrm{Var}\left(Y\right) \tag{32}$$

$$= \sigma_1^2\rho^2 \tag{33}$$

$$\mathrm{Var}\left(X\right) = \mathrm{E}\left(\mathrm{Var}\left(X \,|\, Y\right)\right) + \mathrm{Var}\left(\mathrm{E}\left(X \,|\, Y\right)\right) \tag{34}$$

2. Let $X_1, \ldots, X_n$ be a random sample from the distribution with the following probability density function,

$$f\left(x;\theta\right) = \begin{cases} \dfrac{4y^3}{\theta^4}, & 0 \le y \le \theta,\ \theta > 0 \\ 0, & \text{elsewhere} \end{cases}$$

(a) Find the MLE $\widehat{\theta}_{1n}$ of θ and compute the MSE (mean squared error) of $\widehat{\theta}_{1n}$.

(b) Find an unbiased estimator of $\widehat{\theta}_{2n}$ of θ based on the sample average $\overline{X}_n$.

(c) Which one would you like better between $\widehat{\theta}_{1n}$ and $\widehat{\theta}_{2n}$ as a point estimator of θ? Give your reasoning.

Solution:

3. $X_1, \ldots, X_n$가 다음의 결합 확률밀도함수를 갖는 다항분포를 따른다고 하자.

$$f(x_1, \ldots, x_n) = \binom{n}{x_1 \cdots x_k} p_1^{x_1} \cdots p_k^{x_k}, \quad x_i = 0, \ldots, n \ (i = 1, \ldots, k), \ x_1 + \cdots + x_k = n$$

단, $p_1 + \cdots + p_k = 1$이다.

$$H_0 : p_i = p_{i0}, \ i = 1, \ldots, k \quad \text{대} \quad H_a : \text{not } H_0$$

의 가설검정에 대하여 다음에 답하시오.

(a) 가능도비(또는 일반화 가능도비) 검정의 검정통계량을 구하시오.

(b) n의 값이 클 때, 유의수준 α인 가능도비 검정의 근사적 기각역을 구하시오.

Solution:

(a) 먼저 최대가능도추정량을 구하자. 로그가능도함수는 다음과 같다.

$$\ell(p_1, \ldots, p_k \mid \{X_i\}_{i=1}^n) = \ln n! - \sum_{i=1}^k \ln X_i! + \sum_{i=1}^k X_i \ln p_i \tag{35}$$

하지만 모수를 추정하는데 모수에 제약조건이 붙어있다. 따라서 라그랑지 승수를 붙여야 한다.

$$\ell^*(p_1, \ldots, p_k \mid \{X_i\}_{i=1}^n) = \ln n! - \sum_{i=1}^k \ln X_i! + \sum_{i=1}^k X_i \ln p_i + \lambda \left(1 - \sum_{i=1}^k p_i\right) \tag{36}$$

그러므로 다음 두 편도함수를 구할 수 있다.

$$\frac{\partial}{\partial p_i} \ell^* = \frac{X_i}{p_i} - \lambda = 0 \tag{37}$$

$$\frac{\partial}{\partial \lambda} \ell^* = 1 - \sum_{i=1}^k p_i = 0 \tag{38}$$

라그랑지 승수에 대한 편도함수는 원래 제약조건을 다시 돌려준다. 따라서 (37)만을 이용해서 $X_i = \lambda p_i$이므로

$$\sum_{i=1}^k X_k = \lambda \sum_{i=1}^k p_i \tag{39}$$

따라서 $\widehat{\lambda} = n$이 된다. 다시 이를 (37)에 넣으면

$$\hat{p}_i = \frac{X_i}{n} \tag{40}$$

이 나온다. 일반화 가능도비를 구하면

$$\Lambda = \frac{L_0}{\widehat{L}} \tag{41}$$

$$= \prod_{i=1}^{k} \left(\frac{p_{i0}}{\hat{p}_i}\right)^{X_i} \tag{42}$$

$$= \prod_{i=1}^{k} n^{Xi} \left(\frac{p_{i0}}{X_i}\right)^{X_i} \tag{43}$$

$-2\ln\Lambda \xrightarrow{d} \chi^2(k-1)$라고 알려져 있다. 여기서 자유도가 $k-1$인 이유는 제약조건이 1개 있기 때문이다. 그러므로 검정통계량은 $-2\ln\Lambda$이다.

(b) 위의 근사분포를 이용해서 chi-squared distribution의 임계값을 찾으면 된다.

(c) (첨부) (a)에서 정확검정을 할 수도 있다. 즉, 관측한 값을 벡터로 $\mathbf{x}$라 나타내기로 하자. 그러면 유의확률(p-value)는 다음과 같이 brute-force로 구할 수도 있다.

$$\text{p-value} = \sum_{\{\mathbf{y} \mid f(\mathbf{y}|H_0) \le f(\mathbf{x}|H_0)\}} f(\mathbf{y}) \tag{44}$$

즉, 관측한 값보다 확률이 작아지는 모든 가능한 outcome $\mathbf{y}$를 구해서 확률값을 다 더하면 된다. 하지만 카테고리의 개수가 많아지고 관측치가 많아질수록 정확검정은 너무 고통스러워진다.

4. A regression analysis (involving 45 observations) relating a dependent variable (Y) and two independent variables resulted in the following information.

$$\hat{y} = 0.408 + 1.3387x_1 + 2x_2$$

The SSE for the above model is 49.

When two other independent variables were added to the model, the following information was provided.

$$\hat{y} = 1.2 + 3.0x_1 + 12x_2 + 4.0x_3 + 8x_4$$

This latter model's SSE is 40.

With $\alpha = 0.05$, test to determine if the two added independent variables contribute significantly to the model.

Solution: