Daeyoung Lim
Grad School
Year of 2010, late

1. 다음과 같은 다중선형회귀모형

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

을 고려하자. 여기서, Y는 독립변수를, $X_1, \ldots, X_p$는 설명변수들을, 그리고 $\beta_0, \beta_1, \ldots, \beta_p$는 회귀계수들을 의미하며 ϵ은 평균이 0, 분산이 σ^2인 오차항을 의미한다.

(a) 이 선형회귀모형을 적합한 후 얻어진 예측값을 $\widehat{Y}$으로 나타낼 때, $\widehat{Y}$의 기댓값 및 분산을 구하시오.

(b) 이 회귀모형의 잔차(residual)를 $e = Y - \widehat{Y}$으로 나타낼 때, e의 기댓값 및 분산을 구하시오.

(c) 잔차제곱합(residual sum of squares)의 기댓값을 구하시오.

Solution:

(a) 행렬꼴로 바꾸면 쉽다.

$$\begin{aligned}
\mathrm{E}\left(\widehat{Y}\right) &= \mathrm{E}\left(\mathbf{X}\widehat{\beta}\right) && (1)\\
&= \mathbf{X}\beta && (2)\\
\mathrm{Var}\left(\widehat{Y}\right) &= \mathrm{Var}\left(\mathbf{X}\widehat{\beta}\right) && (3)\\
&= \mathbf{X}\mathrm{Var}\left(\widehat{\beta}\right)\mathbf{X}' && (4)\\
&= \sigma^2 \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}' && (5)
\end{aligned}$$

(b) 이미 구해놨으니 더 쉽다.

$$\begin{aligned}
\mathrm{E}\left(Y - \widehat{Y}\right) &= \mathbf{X}\beta - \mathbf{X}\beta && (6)\\
&= 0 && (7)\\
\mathrm{Var}\left(Y - \widehat{Y}\right) &= \mathrm{Var}\left(\left(\mathbf{I} - \mathbf{H}\right)Y\right) && (8)\\
&= \left(\mathbf{I} - \mathbf{H}\right)\sigma^2\mathbf{I}\left(\mathbf{I} - \mathbf{H}\right)' && (9)\\
&= \sigma^2\left(\mathbf{I} - \mathbf{H}\right) && (10)
\end{aligned}$$

(c) 잔차제곱합의 기댓값은 다음과 같다.

$$\begin{aligned}
\left(Y-\mathbf{H}Y\right)'\left(Y-\mathbf{H}Y\right) &= Y'\left(\mathbf{I}-\mathbf{H}\right)Y && (11)\\
&= \left(Y-\mathbf{X}\beta\right)'\left(\mathbf{I}-\mathbf{H}\right)\left(Y-\mathbf{X}\beta\right) && (12)\\
&= \epsilon'\left(\mathbf{I}-\mathbf{H}\right)\epsilon && (13)\\
\mathrm{E}\left(\epsilon'\left(\mathbf{I}-\mathbf{H}\right)\epsilon\right) &= \mathrm{Tr}\left(\left(\mathbf{I}-\mathbf{H}\right)\mathrm{E}\left(\epsilon\epsilon'\right)\right) && (14)\\
&= \sigma^2\left(\mathrm{Tr}\left(\mathbf{I}\right)-\mathrm{Tr}\left(\mathbf{H}\right)\right) && (15)\\
&= \sigma^2\left(n-p-1\right) && (16)
\end{aligned}$$

여기서 중요한 것은 (11)에서 (12)로 넘어갈 수 있느냐 없느냐이다. (12)를 전개하면 (11)을 제외한 나머지 항들은 모두 소거되어 없어진다.

2. Suppose that the conditional distribution of X given that $P = p$ has a binomial distribution with parameters 5 and p, $X \mid P = p \sim \mathrm{Bin}\left(5, p\right)$ and the marginal distribution of P is a uniform distribution on $(0, 1)$, $P \sim \mathrm{Unif}\left(0, 1\right)$. We would like to calculate the correlation coefficient between X and P.

   (a) Compute variance of X.

   (b) Compute covariance of X and P.

   (c) Compute $\mathrm{Cor}\left(X, P\right)$.

Solution:

(a) 또 베이지언 문제다. 개인적인 생각에 이번 연도 문제는 최태련 교수님이 출제하신 것 같다. 보면 알 수 있다. 아무튼 이 문제는 분산의 분해에 대해 묻고 있는 것이다.

$$\mathrm{Var}\left(X\right) = \mathrm{Var}\left(\mathrm{E}\left(X \mid P\right)\right) + \mathrm{E}\left(\mathrm{Var}\left(X \mid P\right)\right) \qquad (17)$$

따라서 $\mathrm{E}(X \mid P) = 5P$, $\mathrm{Var}(X \mid P) = 5P(1-P)$이므로

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{Var}(5P) + \mathrm{E}(5P(1-P)) && (18)\\
&= 25\mathrm{Var}(P) + 5\left(\mathrm{E}\left(P - P^2\right)\right) && (19)\\
&= \frac{25}{12} + 5 \times \left(\frac{1}{2} - \left(\frac{1}{12} + \left(\frac{1}{2}\right)^2\right)\right) && (20)\\
&= \frac{35}{12} && (21)
\end{aligned}$$

(b) 분산에 이어 기댓값에서도 비슷한 법칙이 있다. *The law of double expectation*이라 하여 조건부를 걸어준 것을 두번 기댓값을 취하면 조건부 없는 것의 기댓값과 같아진다는 정리이다.

$$\mathrm{E}(XP) = \mathrm{E}(\mathrm{E}(XP \mid P)) \tag{22}$$

베이지언 문제에서 무엇의 기댓값을 물어본다거나 하면 (a)에서 나온 '분산의 분해(variance decomposition)'와 '이중기댓값의 정리'를 잘 쓸 줄 알아야 한다.

$$\begin{aligned}
\mathrm{Cov}(X, P) &= \mathrm{E}(XP) - \mathrm{E}(X)\,\mathrm{E}(P) && (23)\\
&= \mathrm{E}(\mathrm{E}(XP \mid P)) - \mathrm{E}(X)\,\mathrm{E}(P) && (24)\\
&= \mathrm{E}(P \cdot \mathrm{E}(X \mid P)) - \mathrm{E}(\mathrm{E}(X \mid P))\,\mathrm{E}(P) && (25)\\
&= \mathrm{E}\left(5P^2\right) - \mathrm{E}(5P) \times \frac{1}{2} && (26)\\
&= 5\left(\frac{1}{12} + \left(\frac{1}{2}\right)^2\right) - \frac{5}{4} && (27)\\
&= \frac{5}{12} && (28)
\end{aligned}$$

(c) 위에서 필요한 요소를 다 구했다.

$$\begin{aligned}
\mathrm{Cor}(X, P) &= \mathrm{Cov}(X, P) \Big/ \left(\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(P)}\right) && (29)\\
&= \frac{5}{12} \Big/ \left(\sqrt{\frac{35}{12} \cdot \frac{1}{12}}\right) && (30)\\
&= \frac{\sqrt{35}}{7} && (31)
\end{aligned}$$

3. Let $X_1, X_2, \ldots, X_n \quad (n > 2)$ be a random sample from the following density

$$f(x;\theta) = \begin{cases} \theta x^{\theta-1}, & 0 < x < 1,\ 0 < \theta < \infty \\ 0, & \text{elsewhere} \end{cases}$$

   (a) Find the maximum likelihood estimator (MLE) $\widehat{\theta}$ of θ.

   (b) Compare variacne of $\widehat{\theta}$ with the Cramér-Rao lower bound.

Solution:

(a) 위 분포는 $\mathrm{Be}(\theta, 1)$이다. 이 분포의 MLE를 여러 번 구해봤다면 $\left(-\sum \ln X_i\right)^{-1}$이라는 것을 알겠지만 그래도 구해보자.

$$L\left(\theta \mid \{X_i\}_{i=1}^{n}\right) = \theta^n \left(\prod_{i=1}^{n} X_i\right)^{\theta-1} \tag{32}$$

$$\ell\left(\theta \mid \{X_i\}_{i=1}^{n}\right) = n \ln\theta + (\theta-1)\sum_{i=1}^{n} \ln X_i \tag{33}$$

$$\ell'\left(\theta \mid \{X_i\}_{i=1}^{n}\right) = \frac{n}{\theta} + \sum_{i=1}^{n} \ln X_i \tag{34}$$

$$\widehat{\theta}^{\mathrm{MLE}} = n \bigg/ \left(-\sum_{i=1}^{n} \ln X_i\right) \tag{35}$$

베타 분포를 따르는 X에 로그를 취한 $\ln X$는 취하는 값의 범위가 $(-\infty, 0]$이기 때문에 $-\ln X$로 뒤집어주면 감마분포를 따르게 된다. 따라서 그것의 역수꼴인 MLE는 *inverse-gamma distribution*을 따른다.

(b) 피셔 정보량(*Fisher information*)을 구해야 한다.

$$\ln f(X \mid \theta) = \ln\theta + (\theta-1)\ln X \tag{36}$$

$$\frac{d}{d\theta} \ln f(X \mid \theta) = \frac{1}{\theta} + \ln X \tag{37}$$

$$\frac{d^2}{d\theta^2} \ln f(X \mid \theta) = -\frac{1}{\theta^2} \tag{38}$$

$$\mathcal{I}(\theta) = -\mathrm{E}_X\left(\frac{d^2}{d\theta^2} \ln f(X \mid \theta)\right) \tag{39}$$

$$= \frac{1}{\theta^2} \tag{40}$$

여기서 주의해야 할 점은 MLE는 비편향 추정량이 아니라는 것이다. *Cramér-Rao bound*는

추정량이 비편향인가 아닌가에 따라 바뀐다.

$$\mathrm{CB}\left(\widehat{\theta}\right) = \begin{cases} \mathrm{Var}\left(\widehat{\theta}\right) \geq \dfrac{1}{n\mathcal{I}\left(\theta\right)}, & \text{if } \widehat{\theta} \text{ is unbiased} \\ \mathrm{Var}\left(\widehat{\theta}\right) \geq \dfrac{d}{d\theta}\mathrm{E}\left(\widehat{\theta}\right) \Big/ \left(n\mathcal{I}\left(\theta\right)\right), & \text{if } \widehat{\theta} \text{ is biased} \end{cases} \tag{41}$$

우리가 구한 MLE는 분포가 다음과 같이 구해진다.

$$-\ln X_i \sim \mathrm{Exp}\left(\theta\right) \tag{42}$$

$$-\sum_{i=1}^{n} \ln X_i \sim \mathrm{Ga}\left(n, \theta\right) \tag{43}$$

$$1 \Bigg/ \left(-\sum_{i=1}^{n} \ln X_i\right) \sim \mathrm{InvGam}\left(n, \theta\right) \tag{44}$$

$$n \Bigg/ \left(-\sum_{i=1}^{n} \ln X_i\right) \sim \mathrm{InvGam}\left(n, n\theta\right) \tag{45}$$

$$\mathrm{E}\left(\widehat{\theta}^{\mathrm{MLE}}\right) = \frac{n}{n-1}\theta \tag{46}$$

$$\mathrm{Var}\left(\widehat{\theta}^{\mathrm{MLE}}\right) = \frac{n^2\theta^2}{\left(n-1\right)^2\left(n-2\right)} \tag{47}$$

이다. *Inverse gamma distribution*이 생소한 사람은 위키피디아에 검색해보길 권한다. 쉽게는 감마 확률변수에 역수를 취하면 inverse-gamma 확률변수가 된다. (여기서 모든 변수는 rate-parameter이다.) 아무튼 위에서 보는 것처럼 MLE는 편향되어 있으므로 CB 중 아래 써야 하며, 기댓값을 모수에 대해 미분하면 $n/\left(n-1\right)$이 되고 고로 *Cramér-Rao bound*는 다음과 같다.

$$\mathrm{Var}\left(\widehat{\theta}^{\mathrm{MLE}}\right) \geq \frac{\theta^2}{n-1} \tag{48}$$

4. Let $X_1, \ldots, X_n$ be a random sample from the following probability density function(pdf),

$$f\left(x;\theta\right) = \theta/x^2, \quad 0 < \theta \leq x < \infty.$$

   1. Find a sufficient statistic for θ.
   2. Find the maximum likelihood estimator (MLE) of θ.
   3. Find the method of moments estimator (MME) of θ.

Solution:

(a) Support가 모수에 의존하므로 가능도함수를 쓸 때 이를 명시해주어야 한다.

$$L\left(\theta \mid \{X_i\}_{i=1}^{n}\right) = \prod_{i=1}^{n} \frac{\theta}{X_i^2} \cdot I_{(\theta,\infty)}\left(X_i\right) \tag{49}$$

$$= \theta^2 \left(\prod_{i=1}^{n} X_i^2\right)^{-1} \cdot I_{\left(0,X_{(1)}\right)}\left(\theta\right) \tag{50}$$

*Fisher-Neyman factorization theorem*에 따르면 $X_{(1)}$이 θ에 대한 최소충분통계량이다(minimal sufficient statistic). 이 문제의 의도된 바는 최소충분통계량을 구하는 것이지만 그냥 충분통계량만 구하라고 하면 $X_1, \ldots, X_n$도 되고 $X_{(1)}, \ldots, X_{(n)}$도 된다. 문제가 사실 잘못 됐다.

(b) 이 경우에는 미분을 써서 최대가능도추정량을 구할 수 없다. 왜냐하면 정의역이 모수에 의존하기 때문이다. 하지만 가로축을 θ, 세로축을 가능도함수($L\left(\theta \mid \mathbf{X}\right)$)로 놓으면 θ가 증가함에 따라 가능도함수 역시 단조증가한다. 따라서 θ가 가장 클 때가 어디인지를 찾아야 한다. 정의역의 범위가 $0 < \theta \leq x$이므로 모든 $X_i,\ i = 1, \ldots, n$과 비교해서 θ가 작아야 하므로 그냥 가장 작은 $X_{(1)}$보다만 작으면 된다. 따라서 θ가 가장 클 때는 $X_{(1)}$이다. $\widehat{\theta}^{\mathrm{MLE}} = X_{(1)}$

(c) 이 경우 적률추정량은 존재하지 않는다. 왜냐하면 기댓값(1차 적률)이 존재하지 않기 때문이다.

$$\mathrm{E}\left(X_1\right) = \int_{\theta}^{\infty} \frac{\theta}{x}\, dx \tag{51}$$

$$= \theta \ln x\Big|_{\theta}^{\infty} \tag{52}$$

$$= \infty \tag{53}$$

전년도 기출에서인가 언급했었지만 확률공간(probability space)과 같은 유한측도공간(finite measure space)에서는 *Hölder inequality*로 인해 $1 \leq p < q$일 경우 q차 적률이 존재하면 그보다 작은 p차 적률은 무조건 존재한다고 했다. 뒤집어 말하면 q보다 작은 p차 적률이 발산할 경우 q차 적률은 자동적으로 함께 발산하게 되어 있다. 이 때문에 1차적률인 기댓값이 유한하지 않고 무한대로 발산하면 그보다 큰 모든 $n > 1$인 n차 적률은 모두 무한대로 발산한다. 고로 그 어떤 적률추정량도 존재하지 않게 된다.

5. Let $X_1, \ldots, X_n$ be a random sample from the following probability density function (pdf),

$$f\left(x;\theta\right) = \theta e^{-\theta x}, \quad 0 < x < \infty,$$

where $\theta = \theta_0$ or $\theta = \theta_1$. We assume that known fixed numbers $\theta_1 > \theta_0$.

(a) Explain the *Neyman-Pearson lemma* briefly.

(b) Explain the most powerful (MP) test briefly.

(c) Obtain the MP test for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ by the *Neyman-Pearson lemma*.

Solution:

(a) Simple과 simple 가설을 비교하는 경우 다음과 같이 가능도비가 어느 상수 k보다 작아지면 귀무가설은 기각된다.

$$\Lambda = \frac{L(\theta_0 \,|\, x)}{L(\theta_1 \,|\, x)} \le k \tag{54}$$

(b) 이때 $\Pr(\Lambda \le k \,|\, H_0) = \alpha$ (여기서 α는 신뢰수준) 이면 최강력 검정이라 부른다.

(c)

$$\Lambda = \frac{\theta_0^n \exp\left(-\theta_0 \sum_{i=1}^n X_i\right)}{\theta_1^n \exp\left(-\theta_1 \sum_{i=1}^n X_i\right)} \tag{55}$$

$$= \left(\frac{\theta_0}{\theta_1}\right)^n \exp\left((\theta_1 - \theta_0) \sum_{i=1}^n X_i\right) \le k \tag{56}$$

따라서 $2\theta \sum_{i=1}^n X_i \le 2\theta k'$일 때 귀무가설이 기각되면 최강력 검정이 된다. 여기서 2θ를 곱해준 이유는 카이제곱분포를 따르게 만들어주고자 함이다. 그냥 감마분포로 해도 무방하다.

$$\Pr\left(2\theta_0 \sum_{i=1}^n X_i \le \chi^2_{2n}(1-\alpha)\right) = \alpha \tag{57}$$