Ch.6 Estimation(추정)

6.1. ESTIMATION USING MOMENTS

Similar to how we use $\bar{x} = (x_1 + \cdots + x_n)/n$ to estimate the population mean $\mu = E(X_i)$, it is natural to use the sample moment $\hat{m}_r$

$$\hat{m}_r = \frac{1}{n}(x_1^r + \cdots + x_n^r)$$

to estimate the population moment $m_r = E[X_i^r]$.

Generally, "method of moments" refers to the method of utilizing a function(g) of moments $(\hat{m}_1, \hat{m}_2, \cdots, \hat{m}_k)$ during the estimation of η

$$\hat{\eta} = g(\hat{m}_1, \cdots, \hat{m}_k), \quad \hat{m}_r = \frac{1}{n}(x_1^r + \cdots + x_n^r) \quad (r = 1, \cdots, k)$$

The statistic $\hat{\eta} = g(\hat{m}_1, \cdots, \hat{m}_k)$ that is used in the estimation of η, is called the "method of moments estimator(MME)".

Ex.6.1.1. MME of population variance and standard deviation.

Let $X_1, \cdots, X_n$ denote the random samples from a population with variance σ^2 $(0 \le \sigma^2 \le +\infty)$.

$$\sigma^2 = \mathrm{Var}[X] = E[X^2] - E[X]^2 = m_2 - m_1^2, \quad \sigma = \sqrt{m_2 - m_1}$$

Hence the MME of population variance and standard deviation are:

$$\widehat{\sigma^2}^{MME} = \hat{m}_2 - (\hat{m}_1)^2 = \frac{1}{n}\sum X_i^2 - \left(\frac{1}{n}\sum X_i\right)^2 = \frac{1}{n}\sum (X_i - \bar{X})^2$$

$$\widehat{\sigma}^{MME} = \sqrt{\hat{m}_2 - (\hat{m}_1)^2} = \sqrt{\frac{1}{n}\sum (X_i - \bar{X})^2}$$

Ex.6.1.2. Non-uniqueness property of MME.

Let $X_1, \cdots, X_n$ denote the random samples from Poisson(λ) $(\lambda > 0)$. Population mean and variance are both λ,

$$\lambda = E[X_i] = m_1, \quad \lambda = \mathrm{Var}(X) = m_2 - m_1^2$$

Hence, depending on the choice of function to estimate λ,

$$\hat{\lambda}_1^{MME} = \hat{m}_1 = \bar{X}, \quad \hat{\lambda}_2^{MME} = \hat{m}_2 - (\hat{m}_1)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

we can get two different MMEs

As one can see from ⟨Ex.6.1.2⟩, when the population is assumed to follow a specific distribution, functional relationships among the moments, and consequently several method-of-moments may be defined for a single parameter.

Even when the population is multidimensional, MME based on sample moments can be constructed to correspond[대응] to the joint moments of the population. For instance, in the bivariate case, let $(X_1, Y_1)^t, \cdots, (X_n, Y_n)^t$ denote the random samples. Instead of $m_{r,s} = E[X_i^r Y_i^s]$ we can substitute it with

$$\hat{m}_{r,s} = \frac{1}{n}\sum_{i=1}^{n} X_i^r Y_i^s$$

to get the MME.

Ex.6.1.3. MME of Correlation Coefficient

let $(X_1, Y_1)^t, \cdots, (X_n, Y_n)^t$ denote the random samples from the bivariate distribution with mean μ_1, μ_2 variance σ_1^2, σ_2^2 $(0 < \sigma_1^2 < +\infty, 0 < \sigma_2^2 < +\infty)$ and correlation coefficient ρ $(-1 < \rho < 1)$. If we express ρ with joint moments,

$$\rho = \frac{\mathrm{Cov}(X_1, Y_1)}{\sqrt{\mathrm{Var}(X_1)\,\mathrm{Var}(Y_1)}} = \frac{E[X_1Y_1] - E[X_1]E[Y_1]}{\sqrt{E[(X_1)^2] - E[X_1]^2}\sqrt{E[(Y_1)^2] - E[Y_1]^2}}$$

If we map the sample joint moments to the population joint moments, the MME of ρ is:

$$\hat{\rho}^{MME} = \frac{\frac{1}{n}\sum_1^n X_iY_i - \bar{X}\bar{Y}}{\sqrt{\frac{1}{n}\sum_1^n (X_i^2 - (\bar{X})^2)}\sqrt{\frac{1}{n}\sum_1^n (Y_i^2 - (\bar{Y})^2)}}$$

$$= \frac{\sum_1^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n}\sum_1^n (X_i^2 - (\bar{X})^2)}\sqrt{\frac{1}{n}\sum_1^n (Y_i^2 - (\bar{Y})^2)}}$$

We see the MME of population correlation coefficient is sample correlation coefficient.

⟨Theorem 6.1.1. Convergence in Probability of MME⟩

Consider estimating a parameter $\eta = g(m_1, \cdots, m_k)$, which can be expressed as a function of moments $m_1, \cdots, m_k$, using random samples $X_1, \cdots, X_n$. If the moments can be defined as real numbers[존재] and the function g is continuous,

$$\hat{\eta}_n^{MME} = g\left(\sum_{i=1}^n X_i/n, \cdots, \sum_{i=1}^n X_i^k/n\right)$$

Converges in probability to η as the sample size grow. i.e.,

$$p\lim_{n\to\infty} \hat{\eta}_n^{MME} = \eta$$

(proof)

From ⟨Ex. 5.2.4⟩ - convergence in probability of the vector of moments - and ⟨Theorem 5.2.5⟩ - Theorem on continuous functions and convergence in probability, ⟨Theorem 6.1.1⟩ is evident.

존재) 기대값의 정의와 ⟨정리 1.6.2⟩로 부터 이는 $E(|X_1|^k) < +\infty$ 인 경우를 뜻한다.

Generally, "estimation" refers to "guessing" the value of a population parameter, and the statistics used in estimation are called "estimators". Let $X_1, \cdots, X_n$ denote the random samples from a population. we often use " $\hat{}$ " over the parameter we wish to estimate, say η, to denote an estimator:

$$\hat{\eta}(X_1, \cdots, X_n) \quad \text{or} \quad \hat{\eta}$$

or $\hat{\eta}_n$ to emphasize the sample size.

Typically, we suppose that $f(x;\theta)$, pdf that depends on the parameter $\theta \in \Omega$, is the model of the population. Here, the target of estimation is parameter θ or $\eta = \eta(\theta)$, a function of θ. Regardless of the distribution type, "consistency" refers to the property of an estimator $\hat{\eta}_n$ which converges in probability to $\eta = \eta(\theta)$ as the sample size (n) increases. Consistency is is a basic property required for estimators.

Definition: Consistency of Estimators [86]

$$\hat{\eta}_n \xrightarrow[n\to\infty]{P_\theta} \eta(\theta) \quad \forall \theta \in \Omega \iff \lim_{n\to\infty} P_\theta(|\hat{\eta}_n - \eta(\theta)| \geq \epsilon) = 0 \quad \forall \epsilon > 0, \forall \theta \in \Omega$$

[86] 확률이나 기대값을 나타낼 때 모집단 분포가 모수 $\theta \in \Omega$에 따라 정해지는 것을 나타내기 위하여 $P(X \in A)$, $E(X)$, $\cdots$ 대신 $P_\theta(X \in A)$, $E_\theta(X \in A)$, $\cdots$ 와 같은 표현을 사용한다.

Ex 6.1.4.

If we express σ^2 $(0 \leq \sigma^2 < +\infty)$ in terms of moments,

$$\sigma^2 = g(m_1, m_2) = m_2 - m_1^2$$

Since g is continuous, by ⟨Theorem 6.1.1⟩, the MME from ⟨Ex.6.1.1⟩

$$\hat{\sigma}_n^{2\,MME} = \frac{1}{n}\sum_i^n (x_i - \bar{x})^2$$

is consistent. Meanwhile, if we express sample variance with these moments,

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n}{n-1}\hat{\sigma}_n^{2\,MME}$$

Hence, sample variance is also a consistent estimator of σ^2, which is identical to the results from ⟨Ex. 5.2.5⟩.

⟨Theorem 6.1.1⟩ is the proof of the cosistency of MME in the one-dimension case. However, it is also evident that the theorem holds for the multivariate case. For instance, the sample correlation coefficient, which is the MME of ρ(population correlation coefficient), is consistent as shown in ⟨Ex.5.3.6⟩. Therefore, we can initially consider MME as our estimator when we are searching for a consistent estimator.

Also, it is known that MMEs have the following limiting distribution. We can prove the following using the CLT in the multivariate case and ⟨Theorem 5.3.3⟩

⟨Theorem 6.1.2. Limiting Distribution of MME⟩

For the MME estimator of $\eta = g(m_1, \cdots, m_K)$:

$$\hat{\eta}_n = g\left(\sum_i^n x_i/n, \cdots, \sum_i^n x_i^K/n\right)$$

? why not k ?

If $E[X_i^{2K}] < +\infty$ and the first derivative of $g(m) = g(m_1, \cdots, m_K)$, $\dot{g}(m)$, is continuous,

$$\sqrt{n}(\hat{\eta}_n - \eta) \xrightarrow[n\to\infty]{d} Z, \quad Z \sim (0, \sigma^2), \quad \sigma^2 = (\dot{g}(m))^t \Sigma \dot{g}(m), \quad \Sigma = (m_{r+s} - m_r \cdot m_s)$$

6.2 MAXIMUM LIKELIHOOD ESTIMATION

The process of estimation is guessing the true parameters in a population distribution using random samples $X_1, \cdots, X_n$, given that the population distribution is one of the pdf $f(x;\theta)$, $\theta \in \Omega$. Now, the probability that the observed values $x = (x_1, \cdots, x_n)^t$ were generated from a model with parameter θ is:

$$pdf_X(x;\theta)\,|\Delta x| = \prod_{i=1}^{n} f(x_i;\theta)\,|\Delta x_1| \cdots |\Delta x_n|$$

Hence, by comparing the probabilites that x was generated from a model with parameters θ vs θ', we know that:

$$\prod_{i=1}^{n} f(x_i;\theta) > \prod_{i=1}^{n} f(x_i;\theta') \quad \text{or} \quad \prod_{i=1}^{n} f(x_i;\theta) < \prod_{i=1}^{n} f(x_i;\theta')$$

and depending on the ^ inequality above, we can predict whether x was generated from a model with parameter θ or θ'. In the comparison above, our main concern is how $\prod_{i=1}^{n} f(x_i;\theta)$ changes depending on $\theta \in \Omega$. Thus, we fix the observation $x = (x_1, \cdots, x_n)^t$ and define $\prod_{i=1}^{n} f_X(x_i;\theta)$ as a function of θ:

$$L(\theta;x) = \prod_{i=1}^{n} f(x;\theta), \quad \theta \in \Omega$$

We call $L(\theta;x)$ the "likelihood function" and the "maximum likelihood estimation" is finding the parameter $\theta \in \Omega$ that maximizes $L(\theta;x)$ and estimating it as the population parameter. $\hat{\theta}(x) = \underset{\theta \in \Omega}{\operatorname{argmax}}\, L(\theta;x)$.

Definition: Likelihood Function and Maximum Likelihood Estimation.

(a) Likelihood Function: Function of $\theta \in \Omega$ with fixed observations $x = (x_1, \cdots, x_n)^t$.

$$L(\theta;x) = \prod_{i=1}^{n} f_X(x_i;\theta), \quad \theta \in \Omega$$

(b) Maximum Likelihood Estimation: we call $\hat{\theta}^{MLE}(x)$ the maximum likelihood estimator[87] of θ if it satisfies the following:

$$L(\hat{\theta}^{MLE}(x);x) = \max_{\theta \in \Omega} L(\theta;x)$$

[87] Estimator $\hat{\theta}(x)$ is ^the observed value of $\hat{\theta} = \hat{\theta}(X)$ for the corresponding observations x. The definition of MLE supposes that there exists a $\theta \in \Omega$ that maximizes the likelihood function.

For each observation in $x = (x_1, \cdots, x_n)^t$, if $\hat{\theta}^{MLE}(x)$ can be defined, the statistic $\hat{\theta}^{MLE} = \hat{\theta}^{MLE}(X)$ is called the "maximum likelihood estimator" (MLE).

Meanwhile, to find the maximum of $L(\theta;x)$, we oftentimes need to find the derivative of $L(\theta;x)$. The derivative of summation is much easier to calculate than the derivative of a product. Thus, we often use $\log(\cdot)$ to transfor $L(\theta;x)$.

$$\ell(\theta;x) = \log L(\theta;x) = \sum_{i=1}^{n} \log f(x_i;\theta)$$

which is called the "log likelihood function". Here, we sometimes exclude $x = (x_1, \cdots, x_n)^t$ and simple denote it as $L(\theta)$ or $\ell(\theta)$.

★ Ex. 6.2.1 Maximum Likelihood Estimation in Poisson Distribution

Let $X_1, \cdots, X_n$ be random samples from Poisson(λ) $|\lambda \geq 0|$[88]. What is the MLE of λ?

$$f_X(x;\theta) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \cdots$$

$$\prod_{i=1}^{n} f_X(x_i;\theta) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = e^{-n\lambda}\lambda^{\sum x_i} \Big/ \prod_{i=1}^{n} x_i! = L(\lambda;x)$$

$$\ell(\theta) = \log L(\theta;x) = -n\lambda + \sum x_i \log\lambda - \log\prod x_i!$$
$$= -n\lambda + \sum x_i \log(\lambda) - \sum \log x_i$$

$$\frac{\partial \ell}{\partial \theta} = -n + \frac{\sum x_i}{\lambda}$$

$$\frac{\partial^2 \ell}{\partial \theta^2} = -\frac{\sum x_i}{\lambda^2} < 0 \quad \forall\, \lambda \in \Omega = (\lambda \geq 0) \quad * \text{ If } \lambda = 0 \;\; \ddot{\ell}(\lambda=0) \text{ is undefined.}$$

$$\frac{\partial \ell}{\partial \theta}\Big|_{\hat{\theta}^{MLE}} = 0 \Rightarrow -n + \frac{\sum x_i}{\hat{\lambda}^{MLE}} = 0 \iff \hat{\lambda}^{MLE} = \frac{1}{n}\sum x_i = \bar{x}$$

[Textbook Solution]

$\dot{\ell}(\lambda) = -n + n\bar{x}/\lambda$, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

$\ddot{\ell}(\lambda) = -n\bar{x}/\lambda^2$, $\lambda > 0$

(i) $\bar{x} > 0$ 이면 $\max_{\lambda > 0} L(\lambda) = L(\bar{x})$

(ii) $\bar{x} = 0$ 이면 $L(\lambda) = -n\lambda - \log(x_1! \cdots x_n!)$ $\therefore \sup_{\lambda > 0} L(\lambda) = \lim_{\lambda \to 0} L(\lambda)$

그런데 가능도함수는 $\lambda = 0$ 에서 정의되어 있고 연속함수이므로, 어느 경우에나

$$\sup_{\lambda > 0} L(\lambda; x_1, \cdots, x_n) = L(\bar{x}; x_1, \cdots, x_n)$$

$$\therefore \max_{\lambda \geq 0} L(\lambda; x_1, \cdots, x_n) = L(\bar{x}; x_1, \cdots, x_n) \quad \forall\, x = (x_1, \cdots, x_n)^t$$

그러므로 λ의 최대가능도 추정량은 $\hat{\lambda}^{MLE} = \bar{x}$ 로 주어진다.

* Poisson($\lambda = 0$): $\begin{cases} 1 & \text{if } x = 0 \\ 0 & \text{if } x \neq 0 \end{cases}$

$\lambda > 0$ 일 때와 $\lambda = 0$ 일 때로 구분 필요. | $\bar{x} \neq 0$, $\bar{x} = 0$ 인 경우로 구분

$\lambda = 0 \Rightarrow \bar{x} = 0$ (always)

[88] 이러한 표현은 모수공간이 $\Omega = \{\lambda : \lambda \geq 0\}$ 임을 뜻하는 것이다. 한편 $\lambda = 0$ 인 경우의 포아송분포 Poisson(λ)는 상수 0에 집중된 분포를 뜻하고, 이러한 분포가 모형에 포함되어 있으므로 관측결과에 관계 없이 최대가능도 추정값이 존재하게 되고 최대가능도 추정량을 정의할 수 있는 것이다.

When we look into the increase/decrease of the log likelihood function $\ell(\theta;x) = \log L(\theta;x)$, we utilize its derivatives as shown in ⟨Ex. 6.2.1⟩. In particular, the equation $\dot{\ell}(\theta;x) = 0$ is called the likelihood equation, and among the roots ($\hat{\theta}(x)$) of the equation $\dot{\ell}(\theta;x)$, we can find the MLE.

〈Theorem 6.2.1〉

If the function $\ell(\theta)$ that is defined on an open interval Ω_0 on the real line ($\Omega_0 \subset \mathbb{R}$) is double differentiable, continuous and:

$$\ddot{\ell}(\theta) < 0, \quad \forall \theta \in \Omega_0, \quad \dot{\ell}(\hat{\theta}) = 0, \quad \hat{\theta} \in \Omega_0$$

then, $\max_{\theta \in \Omega_0} \ell(\theta) = \ell(\hat{\theta})$

(proof)

From Taylor Series in Appendix I,

$$\ell(\theta) = \ell(\hat{\theta}) + \dot{\ell}(\hat{\theta})(\theta-\hat{\theta}) + \frac{1}{2}\ddot{\ell}(\hat{\theta})(\theta-\hat{\theta})^2, \quad \exists \theta^* \in \Omega_0$$
$$= \ell(\hat{\theta}) + \frac{1}{2}\ddot{\ell}(\hat{\theta})(\theta-\hat{\theta})^2$$
$$\le \ell(\hat{\theta}) \quad (\because \ddot{\ell}(\hat{\theta}) < 0) \quad \forall \theta \in \Omega_0$$

Ex. 6.2.2. Maximum Likelihood Estimation in Binomial Distribution.

Let $X_1, \cdots, X_n$ be random samples from Bernoulli(p) ($0 \le p \le 1$).
Find the MLE of p

$$f_X(x;p) = p^x(1-p)^{1-x}, \quad x = 0, 1$$

$$L(\theta;x) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

$$l(\theta;x) = \log L(\theta;x) = \sum x_i \log(p) + (n-\sum x_i)\log(1-p)$$

$$\dot{l}(\theta;x) = \sum x_i/p - (n-\sum x_i)/(1-p)$$
$$= \frac{(\sum x_i)(1-p) - (n-\sum x_i)\cdot p}{p(1-p)} = \frac{\sum x_i - p\sum x_i - np + p\sum x_i}{p(1-p)} = \frac{\sum x_i - np}{p(1-p)} = \frac{n(\bar{x}-p)}{p(1-p)}$$

$$\ddot{l}(\theta;x) = -\frac{\sum x_i}{p^2} - \frac{n-\sum x_i}{(1-p)^2} < 0 \quad \forall p: 0 < p < 1$$

$$\dot{l}(\theta=\bar{x};x) = \frac{n(\bar{x}-\bar{x})}{p(1-p)} = 0 \quad \text{if } p \neq 0, 1$$

(ii) if $\bar{x} = 0$

$$l(\theta;x) = \log L(\theta;x) = \sum x_i \log(p) + (n-\sum x_i)\log(1-p)$$
$$= n\bar{x}\log(p) + n(1-\bar{x})\log(1-p)$$
$$= n\log(1-p)$$

$$\sup_{0<p<1} n\log(1-p) = \lim_{p\to 0} n\log(1-p)$$

~~$\hat{p}^{MLE} = 0 = \bar{x}$~~

(iii) if $\bar{x} = 1$

$$l(\theta;x) = \log L(\theta;x) = n\bar{x}\log(p) + n(1-\bar{x})\log(1-p)$$
$$= n\log(p)$$

$$\sup_{0<p<1} l(\theta;x) = \sup_{0<p<1} n\log p = \lim_{p\to 1} n\log p$$ ~~$\hat{p} = 1 = \bar{x}$~~

가능도함수는 $p=0, p=1$ 에서 정의됨 로그가능도 함수는 No!

가능도함수는 $p=0$, $p=1$ 에서 정의되고 편측함수이므로, 어느 경우에나

$\sup_{0<p<1} L(p; x_1, \cdots, x_n) > L(\hat{p}; x_1, \cdots, x_n)$ $\lim_{p\to 1} l(\theta;x) = l(\theta=1;x)$

또한 $p=0$ 인 경우는 $x_1 = \cdots = x_n = 0$ 이고, $p=1$ 인 경우는 $x_1 = \cdots = x_n = 1$ 이므로

$\therefore \max_{0\le p\le 1} L(p; x_1, \cdots, x_n) = L(\hat{p}; x_1, \cdots, x_n) \quad \forall x = (x_1, \cdots, x_n)^t$

그러므로 p의 최대가능도 추정량은 $\hat{p}^{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$ 로 주어진다.

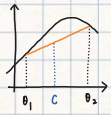
In 〈Theorem 6.2.1〉, the condition was: $\ddot{\ell}(\theta) < 0 \quad \forall \theta \in \Omega$. A sufficient (충분조건) for $\ell(\theta)$ to be a "strictly concave function", i.e.,

$$\ell(a\theta_1 + (1-a)\theta_2) > a\,\ell(\theta_1) + (1-a)\,\ell(\theta_2) \quad \forall a: 0 < a < 1$$
$$\forall \theta_i \in \Omega_0 \ (i=1,2) \ (\theta_1 \neq \theta_2)$$

* in contrasts, convex function: $\ell(a\theta_1+(1-a)\theta_2) < a\,\ell(\theta_1) + (1-a)\,\ell(\theta_2)$

i.e. This is a condition stating that the value of the function at an interior point of a line segment is greater than the corresponding convex combination of the function values at the two endpoints.

Concave function



$$a = \frac{c-\theta_1}{\theta_2-\theta_1}, \quad 1-a = 1-\frac{c-\theta_1}{\theta_2-\theta_1} = \frac{\theta_2-\theta_1-c+\theta_1}{\theta_2-\theta_1} = \frac{\theta_2-c}{\theta_2-\theta_1}$$

If the log likelihood function satisfies the above condition, finding the root of the first derivative $\dot{\ell}(\theta)$ gives us the mle of the likelihood function by theorem 6.2.1. Yet oftentimes, finding the root can be difficult. In those cases, the following theorem on concave functions come in handy.

〈Theorem 6.2.2〉 Meaning of 0 variance

If the function $\ell(\theta)$ defined on an open interval $\Omega_0 \subset \mathbb{R}$ is double differentiable, its second derivative is continuous and:

$$\ddot{\ell}(\theta) < 0 \quad \forall \theta \in \Omega_0 \quad \text{and} \quad \lim_{\theta \to \partial(\Omega_0)} \ell(\theta_0) = -\infty$$ [89]

the root, $\hat{\theta}$, of equation $\dot{\ell}(\theta) = 0, \theta \in \Omega_0$ exists and is unique.

(proof)

We will prove the theorem given that $\Omega_0 = (-\infty, +\infty)$. In this case,

$$\lim_{\theta\to-\infty} \ell(\theta) = -\infty, \quad \lim_{\theta\to+\infty} \ell(\theta) = -\infty$$

* This is a Corollary to the EVT (Extreme Value Theorem)

and Since $\ell(\theta)$ is continuous, $\ell(\theta)$ has a maximum. i.e.,

$$\ell(\hat{\theta}) = \max_{\theta\in\Omega_0} \ell(\theta), \quad \hat{\theta} \in \Omega_0$$

[89] 여기에서 $\partial(\Omega_0)$는 수직선 위의 구간 Ω_0의 경계 (boundary)를 나타내는 기호이고, $\theta \to \partial(\Omega_0)$가 뜻하는 것은 유한구간의 경우에는 그 끝점으로 가까이 가고, 무한구간의 경우에는 무한히 커지거나 작아지는 것을 뜻한다.

5.3.4 Maximum-Minimum Theorem *Let $I := [a, b]$ be a closed bounded interval and let $f : I \to \mathbb{R}$ be continuous on I. Then f has an absolute maximum and an absolute minimum on I.*

Such $\hat{\theta}$ exists. If $\ell(\theta)$ is maximum at points $\hat{\theta}_1$ and $\hat{\theta}_2$, suppose:

$$\ell(\hat{\theta}_1) = \ell(\hat{\theta}_2) = \max_{\theta\in\Omega_0} \ell(\theta), \quad \hat{\theta}_1 < \hat{\theta}_2$$

However, since $\ell(\theta)$ is concave,

$$\ell\left[\tfrac{1}{2}\theta_1+\tfrac{1}{2}\theta_2\right) > \tfrac{1}{2}\ell(\theta_1)+\tfrac{1}{2}\ell(\theta_2) = \max_{\theta\in\Omega_0}\ell(\theta),\ \tfrac{1}{2}\hat{\theta}_1+\tfrac{1}{2}\theta_2\in\Omega_0.$$

a Contradiction arises. Therefore $\ell(\theta)$ is maximum only at a Single point $\hat{\theta}$. It is evident that $\dot{\ell}(\hat{\theta})=0$ and thus $\hat{\theta}$ is the unique root of $\ell(\theta)$.

If finding the exact root is difficult, we can use the first order Taylor series approximation[90]: $0=\dot{\ell}(\theta')\simeq\dot{\ell}(\theta^0)+\ddot{\ell}(\theta^0)(\hat{\theta}^1-\theta^0)$

we can repeatedly apply this equation to find the root, and for the initial input $\hat{\theta}^0$, MME is often used.

[90] 이러한 반복법은 뉴턴-랩슨 (Newton-Raphson) 방법이라고 하며 많은 패키지에 함수화 되어 있다.

Definition: Solving likelihood equation via one-step iteration.

$$\hat{\theta}^{r+1}=\hat{\theta}^r+\left[-\ddot{\ell}(\hat{\theta}^r)\right]^{-1}\dot{\ell}(\hat{\theta}^r),\ (r=0,1,\cdots),\ \hat{\theta}^0=\hat{\theta}^{MME}$$
$$\Leftrightarrow \hat{\theta}^r+\left[-\bar{\ddot{\ell}}(\hat{\theta}^r)\right]^{-1}\bar{\dot{\ell}}(\hat{\theta}^r)$$

Ex.6.2.3. Maximum Likelihood Estimation in Logistic Distribution.

Let $X_1,\cdots,X_n$ be the random samples from Logistic distribution $L(\theta,1)$ $(-\infty<\theta<+\infty)$ in ⟨Ex.4.1.5⟩. Find an algorithm that calculates the mle of θ.

(Solution) The pdf of $L(\theta,1)$ distribution is:

$$pdf(x;\theta)=\frac{e^{x-\theta}}{(1+e^{x-\theta})^2}=\frac{e^{-x+\theta}}{(1+e^{-x+\theta})^2},\ -\infty<x<+\infty$$

The log likelihood function and its derivatives are:

$$\ell(\theta)=n\bar{x}-n\theta-2\sum_{i=1}^{n}\log(1+e^{x_i-\theta})=-n\bar{x}+n\theta-2\sum_{i=1}^{n}\log(1+e^{-x_i+\theta})$$

$$\dot{\ell}(\theta)=-n-2\sum_{i=1}^{n}\frac{e^{x_i+\theta}}{1+e^{x_i-\theta}}=n-2\sum_{i=1}^{n}\frac{e^{-x_i+\theta}}{1+e^{-x_i+\theta}}$$

$$\ddot{\ell}(\theta)=-2\sum\frac{e^{-x_i+\theta}}{1+e^{-x_i+\theta}}-\frac{(e^{-x_i+\theta})^2}{(1+e^{-x_i+\theta})^2}=-2\sum_{i=1}^{n}\frac{e^{-x_i+\theta}}{(1+e^{-x_i+\theta})^2}<0\ \forall\theta:-\infty<\theta<+\infty$$

$\therefore \ddot{\ell}(\theta)<0\ \forall\theta:-\infty<\theta<+\infty$

$\lim_{\theta\to-\infty}\ell(\theta)=-\infty$ and $\lim_{\theta\to+\infty}\ell(\theta)=-\infty$

$\therefore$ By ⟨Theorem 6.2.2⟩, the mle of θ is given as the root of

$$\dot{\ell}(\theta)=n-2\sum_{i=1}^{n}e^{-x_i+\theta}/(1+e^{-x_i+\theta})=0$$

Meanwhile, the mean of $L(\theta,1)$ is θ. Thus $\hat{\theta}^{MME}=\bar{x}$. We can use the following iterative process to find $\hat{\theta}^{MLE}$.

$$\hat{\theta}^{r+1}=\hat{\theta}^r+\left[-\ddot{\ell}(\hat{\theta}^r)\right]^{-1}\cdot\dot{\ell}(\hat{\theta}^r),\ (r=0,1,\cdots)\quad \hat{\theta}^0=\bar{x}$$

☆☆
* Logistic distribution 의 대칭성을 활용하여 $\theta\to-\theta$ 로 변환한 이유
$\lim_{\theta\to-\infty}\ell(\theta)=+\infty$ 임을 보이기 쉬워서. 반대로, θ 일 때는 $\lim_{\theta\to+\infty}\ell(\theta)=-\infty$

⟨Theorem 6.2.3. 1-1 transformation of parameter & maximum likelihood estimation⟩

If the mle of θ exists, let $g(\theta)$ be a 1-1 transformation of θ such that

$$\eta=g(\theta)$$

The mle of $\hat{\eta}^{MLE}$ is: $\hat{\eta}^{MLE}=g(\hat{\theta}^{MLE})$

(proof)
For a normal distribution $N(0,\sigma^2)$, the pdf is invariant whether the population variance σ^2 or the population standard deviation $\sigma=\sqrt{\sigma^2}$ is taken as the parameter; Similarly, the pdf of a random sample is invariant under a one-to-one reparameterization of the parameter. Meaning, for $\eta=g(\theta),\theta\in\Omega$, a 1-1 transformation:

$$pdf(x_1,\cdots,x_n;\theta)=pdf(x_1,\cdots,x_n;g^{-1}(\eta)),\ \eta\in g(\Omega)$$

Thus, if we take $\eta=g(\theta)$ as the parameter, the likelihood function will be:

$$L(g^{-1}(\eta);x_1,\cdots,x_n)=pdf(x_1,\cdots,x_n;g^{-1}(\eta))$$

Meanwhile,

$$\max_{\eta\in g(\Omega)}L(g^{-1}(\eta);x_1,\cdots,x_n)=\max_{\theta\in\Omega_0}L(\theta;x_1,\cdots,x_n)=L(\hat{\theta}^{MLE})=L(g^{-1}(g(\hat{\theta}^{MLE})))$$

$$\Rightarrow \max_{\eta\in g(\Omega)}L(g^{-1}(\eta);x_1,\cdots,x_n)=L(g^{-1}(g(\hat{\theta}^{MLE})))$$

$$\therefore \hat{\eta}^{MLE}=g(\hat{\theta}^{MLE})\quad\square$$

☆ Ex. 6.2.4. Maximum Likelihood Estimation in Exponential Distribution.

Let $X_1,\cdots,X_n$ be random samples from $Exp(\theta)$ $(0<\theta<+\infty)$. Find the mle of θ.

$f_X(x)=\frac{1}{\theta}e^{-x/\theta}$

$\log f_X(x)=-\log(\theta)-\frac{x}{\theta}$

$L(\theta)=\sum\log f_X(x_i)=-n\log(\theta)-\frac{n\bar{x}}{\theta},\ \bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$

$\dot{L}(\theta)=-\frac{n}{\theta}+\frac{n\bar{x}}{\theta^2}$

$\ddot{L}(\theta)=\frac{n}{\theta^2}-2\frac{n\bar{x}}{\theta^3}=\frac{n}{\theta^2}\left(1-\frac{2\bar{x}}{\theta}\right)$ ← 이 로그가능도함수는 순단목함수가 아니기 때문에 ⟨정리 6.2.2⟩를 적용할 수 없다.

let $\beta=\frac{1}{\theta}\Rightarrow f_X(x)=\beta e^{-\beta x}$

$\log f_X(x)=\log\beta-\beta x$

$L(\beta;x)=\sum\log f_X(x)=n\log\beta-n\beta\bar{x}$

$\dot{L}(\theta)=n/\beta-n\bar{x}$

$\ddot{L}(\theta)=-n/\beta^2<0\ \forall\beta:(0<\beta<+\infty)$ + $\lim_{\beta\to0}L(\beta)=-\infty$, $\lim_{\beta\to\infty}L(\beta)=-\infty$

$\therefore \hat{\beta}^{MLE}$ exists. $\dot{L}(\theta)=n/\beta-n\bar{x}=0\ \therefore 1/\beta-\bar{x}=0\Leftrightarrow\beta=\frac{1}{\bar{x}}$

$L(\theta)$ is a concave function. Apply theorem 6.2.2.

$\hat{\theta}^{MLE}=g(\hat{\beta}^{MLE})=\frac{1}{\hat{\beta}^{MLE}}=\bar{x}\quad\therefore\hat{\theta}^{MLE}=\bar{x},\ \bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$

As we have seen from 〈Ex. 6.2.4〉, if the log likelihood function becomes a concave function through 1-1 transformation of the parameter, it becomes easier to find the mle. In the next theorem, we will introduce the maximum likelihood estimation of a specific type of probability model that has the this property.

〈Theorem 6.2.4. Maximum Likelihood Estimation in Exponential Family〉

If the pdf is:

$$pdf(x;\eta) = \exp\{\eta T(x) - A(\eta) + S(x)\}, \quad x \in \mathcal{X}, \ \eta \in N$$

and $X_1, X_2, \cdots, X_n$ are random samples from a population[91] satisfying (i), (ii) and (iii), and if the likelihood function

$$\dot{A}(\eta) = \frac{1}{n}\sum_{i=1}^{n} T(x_i) \ \ (\text{or}) \ \ E_\eta[T(X_1)] = \frac{1}{n}\sum_{i=1}^{n} T(x_i), \quad \eta \in N$$

has the root $\hat{\eta} = \hat{\eta}(x_1, \cdots, x_n)$, $\hat{\eta}$ is the mle of η

(i) Support of the distribution $\mathcal{X} = \{x : pdf(x;\eta) > 0\}$ does not depend on the parameter(s)

(ii) parameter space is an open interval in $\mathbb{R}$

(iii) $Var(T(X_i)) > 0 \Leftrightarrow T(x) \neq$ constant

(proof) The key to this proof is the following properties of cgf

$$cgf_{T(X_i);\eta}(s) = \log mgf_{T(X_i);\eta}(s) = \log E_\eta[e^{sT(X_i)}] = A(s+\eta) - A(\eta)$$

$-\epsilon < s < \epsilon \ (\exists \epsilon > 0)$?

$$* \ E_\eta[e^{sT(x_i)}] = \int_{-\infty}^{\infty} \exp\{(s+\eta)T(x_i) - A(\eta) + S(x)\}\,dx$$

$$= \int_{-\infty}^{\infty} \exp\{A(\eta+s) - A(\eta)\} \exp\{(s+\eta)T(x) - A(\eta+s) + S(x)\}\,dx$$

$$= \exp\{A(\eta+s) - A(\eta)\}, \quad s+\eta \in N, \ \eta \in N, \ |s| < \epsilon \ (\exists \epsilon > 0)^*$$

$$E[T(X_i)] = \frac{\partial}{\partial s} cgf_{T(X_i);\eta}(s)\Big|_{s=0} = \frac{\partial}{\partial s}(A(\eta+s) + A(\eta))\Big|_{s=0} = \dot{A}(\eta)$$

$$Var[T(X_i)] = \frac{\partial^2}{\partial s^2} cgf_{T(X_i);\eta}(s)\Big|_{s=0} = \frac{\partial^2}{\partial s^2}(A(\eta+s) + A(\eta))\Big|_{s=0} = \ddot{A}(\eta) \quad \eta \in N.$$

* 여기에서 s의 범위가 원점이 포함된 열린구간으로 정해질 수 있는 것은 조건 (ii) 때문이다.

From condition (iii) and 〈Theorem 1.6.4〉, $\therefore \ddot{A}(\eta) > 0 \ \ \forall \eta \in N \ \ (\because \ddot{A}(\eta) = Var(T(x_i)))$

Thus, for the log-likelihood function:

$$\ell(\eta) = \eta\sum_{i=1}^{n} T(x_i) - nA(\eta) + \sum_{i=1}^{n} S(x_i)$$

$$\ddot{\ell}(\eta) = -n\ddot{A}(\eta) < 0 \quad \forall \eta \in N$$

$\therefore \ell(\eta)$ is a strictly concave function

Thus, by 〈Theorem 6.2.1〉, if the root $\hat{\eta} = \hat{\eta}(x_1, \cdots, x_n)$ of the likelihood function exists,* $\hat{\eta}$ is the mle of η.

* $\exists \hat{\eta}$ s.t. $\dot{\ell}(\hat{\eta}) = \sum_{i=1}^{n} T(x_i) - n\dot{A}(\hat{\eta}) = 0$

* 91) 이러한 조건을 만족하는 〈정리 6.2.4〉에 주어진 형태의 확률밀도함수의 집합을 단일모수 지수족 (single parameter exponential family of pdf's) 이라고 한다.

Definition: 1-1 Parameter transformation in Exponential family and Maximum Likelihood Estimation.

If $pdf(x;\theta) = \exp\{g(\theta)T(x) - A(g(\theta)) + S(x)\}$, $x \in \mathcal{X}$, $\theta \in \Omega$

and reparameterization $\eta = g(\theta)$ satisfies the conditions in 〈Theorem 6.2.4〉, the likelihood function using θ is:

$$E_\theta[T(X_i)] = \frac{1}{n}\sum_{i=1}^{n} T(x_i), \quad \theta \in \Omega.$$

$$\| \ \dot{A}(\eta)$$

$$\rightarrow \dot{\ell}(\eta) = \sum T(x_i) - n\dot{A}(\eta) = 0$$

$$\Rightarrow \sum T(x_i) - nE_\eta[T(X_i)] = 0$$

$$\Rightarrow \frac{1}{n}\sum T(x_i) = E_\eta[T(X_i)] = E_\theta[T(X_i)]$$

Ex. 6.2.5. Examples of exponential family that Satisfies the conditions in 〈Theorem. 6.2.4.〉

(a) Bernoulli distribution model Bernoulli(p), $0 < p < 1$:

$$f(x) = p^x(1-p)^{1-x} = \exp\{x\log p + (1-x)\log(1-p)\}$$

$$= \exp\left\{x\log\left(\frac{p}{1-p}\right) + \log(1-p)\right\} \ (x = 0, 1)$$

let $\eta = \log\left(\frac{p}{1-p}\right) \Leftrightarrow e^\eta = \frac{p}{1-p} \Leftrightarrow e^\eta(1-p) = p \Leftrightarrow e^\eta - e^\eta p = p$

$\Leftrightarrow e^\eta = e^\eta p + p \Leftrightarrow e^\eta = p(1+e^\eta)$

$\therefore p = e^\eta/(1+e^\eta)$

$T(x) = x$, $A(\eta) = -\log(1-p)$

$$E[T(X)] = \dot{A}(\eta) = \frac{\partial}{\partial\eta} - \log\left(\frac{1}{1+e^\eta}\right) = \frac{\partial}{\partial\eta}\log(1+e^\eta) = \frac{e^\eta}{1+e^\eta}$$

$$E[T(X)] = \frac{e^\eta}{1+e^\eta} = p \quad \therefore \frac{1}{n}\sum X_i = \hat{p}$$

(b) Poisson Distribution Model Poisson(λ) ($\lambda > 0$):

$$f_X(x) = \frac{e^{-\lambda}\lambda^x}{x!} = \exp\{-\lambda + x\log\lambda - \log x!\} \quad x = 0, 1, 2, \cdots$$

$T(x) = x$, $\eta = \log\lambda$, $A(\eta) = \lambda$, $S(x) = -\log x!$

exponential family: $\exp\{\eta T(x) - A(\eta) + S(x)\}$

$Var(T(x)) = \lambda \neq 0$ ✓✓

Support of distribution does not depend on the parameter. ✓✓

parameter space is an open interval $\Omega_0 \subset \mathbb{R}$ ✓✓

$$E[T(X)] = \dot{A}(\eta) = \frac{\partial}{\partial\eta}e^\eta = e^\eta = \lambda$$

$$\therefore \hat{\lambda}^{MLE} = \frac{1}{n}\sum X_i$$

(c) Geometric Distribution Model Geo(p) $0<p<1$:

$$f(x;p) = (1-p)^{x-1}\cdot p = \exp\{(x-1)\log(1-p) + \log p\}$$
$$= \exp\{x\log(1-p) + \log(\frac{p}{1-p})\} \quad x = 1, 2, \cdots$$

$T(x)=x$, $\eta = \log(1-p)$, $p = 1-e^{\eta}$, $A(\eta) = -\log(\frac{p}{1-p}) = \log(\frac{1-p}{p}) = \log(\frac{e^{\eta}}{1-e^{\eta}})$

(i) support of distribution does not depend on the parameters.
(ii) Parameter space Ω_0 is an open interval in $\mathbb{R}$ (수직선)
(iii) $Var(T(X)) = Var(X) = (1-p)/p^2 \neq 0$ $\because p: 0<p<1$

All 3 conditions are satisfied.

$$E[T(X)] = \dot{A}(\eta) = \frac{\partial}{\partial\eta}\log(\frac{e^{\eta}}{1-e^{\eta}}) = \left[\frac{e^{\eta}}{1-e^{\eta}} + \frac{(e^{\eta})^2}{(1-e^{\eta})^2}\right] \Big/ \left(\frac{e^{\eta}}{1-e^{\eta}}\right)$$
$$= \left[\frac{e^{\eta} - (e^{\eta})^2 + (e^{\eta})^2}{(1-e^{\eta})^2}\right] \Big/ \left[\frac{e^{\eta}}{1-e^{\eta}}\right]$$
$$= (1-e^{\eta})^{-1} = \frac{1}{p}$$
$$\therefore \frac{1}{\bar{x}} = \hat{p}^{MLE}, \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

* 추가조건 : $\frac{1}{n}\sum_{i=1}^{n} x_i = E_p(X_i) = 1/p$, $0<p<1$

이로부터 $\frac{1}{n}\sum_{i=1}^{n} x_i > 1$ 인 경우에 $\hat{p}^{MLE} = 1/\bar{x}$

(d) Exponential Distribution Model Exp(θ), $\theta>0$:

$$f_X(x;\theta) = \frac{1}{\theta}e^{-x/\theta}, \quad 0<x<+\infty \quad \sim Gamma(1,\theta) \quad \alpha\beta^2 = \theta^2$$
$$= \exp\{-\log\theta - x/\theta\}$$

$T(x) = -x$ $\eta = \theta^{-1}$, $A(\eta) = \log(\theta) = -\log(\eta)$

(i) Support of the distribution $\mathcal{X} := \{x : 0<x\}$ does not depend on the parameter
(ii) parameter space Ω_0 is an open interval in $\mathbb{R}$
(iii) $Var(T(X)) = Var(-X) = Var(X) = \theta^2 \neq 0$

$$E[T(X)] = \dot{A}(\eta) = -\frac{1}{\eta} = -\theta$$
$$\therefore -\frac{1}{n}\sum_{i=1}^{n} x_i = -\hat{\theta}^{MLE} \Rightarrow \hat{\theta}^{MLE} = \bar{x}$$

(e) Pareto Distribution Model Pareto(1,θ), $\theta>0$:

$$f(x;\theta) = \theta x^{-\theta-1}, \quad x>1$$
$$= \exp\{\log\theta - (\theta+1)\log x\}$$
$$= \exp\{\log\theta - \theta\log x - \log x\}$$

$T(x) = \log x$, $\eta = -\theta$, $A(\eta) = -\log\theta = -\log(-\eta)$

(i) Support of the distribution $\mathcal{X} := \{x : x>1\}$ does not depend on the parameter
(ii) parameter space $\Omega_0 := \{\theta : \theta>0\}$ is an open interval in $\mathbb{R}$
(iii) $Var(T(X)) = Var(\log X) \neq 0$ let $y = \log x$, $x = e^y$ $|J_{x^{-1}}| = |\frac{dx}{dy}| = |e^y| = |x|$
$\therefore pdf_Y(y) = \theta(e^y)^{-\theta-1}\cdot e^y = \theta e^{-\theta y} \sim Exp(1/\theta)$
$\therefore Var(Y) = \theta^{-2}$

* log of Pareto(1,θ) is Exp($1/\theta$)

$$E[T(X)] = \dot{A}(\eta) = \frac{\partial}{\partial\eta} - \log(-\eta) = -\frac{-1}{-\eta} = -\frac{1}{\eta} = \frac{1}{\theta}$$

$$\therefore \overline{\log X} = (\hat{\theta}^{MLE})^{-1} \Rightarrow \hat{\theta}^{MLE} = (\overline{\log X})^{-1}, \quad \overline{\log X} = \frac{1}{n}\sum_{i=1}^{n}\log x_i$$

추가로 : "한편 이 경우에 $X_i^{-\theta} \sim U(0,1)$, $\theta\log X \sim Exp(1)$, $E_\theta(-\log X_i) = -1/\theta$ ...

So far, we have looked into differentiable likelihood functions. However, there are cases in which differentiation is not possible. In these cases, we must find the mle through investigating the increase or decrease of likelihood functions. Also, we will look into examples where the mle is not unique.

★ Ex. 6.2.6. Maximum Likelihood Estimation in Double Exponential Distribution.

Let $X_1, \cdots, X_n (n=2m+1)$ be random samples from the Double Exponential distribution $DE(\theta,1)$ $(-\infty<\theta<+\infty)$. Find the mle of θ

$$f_X(x;\theta) = \frac{1}{2}e^{-|x-\theta|}, \quad -\infty<x<+\infty$$

$$\log f_X(x) = -\log 2 - |x-\theta|$$
$$\ell(\theta) = \sum_{i=1}^{n}\log f_X(x_i) = -n\log 2 - \sum_{i=1}^{n}|x_i-\theta|$$

let r be a number such that $x_{(r)} \le \theta \le x_{(r+1)}$

$$\ell(\theta) = -n\log 2 - \left[\sum_{i=1}^{r}(\theta - x_{(i)}) + \sum_{i=r+1}^{n}(x_{(i)} - \theta)\right]$$
$$= (n-2r)\theta + \sum_{i=1}^{r}x_{(i)} - \sum_{i=r+1}^{n}x_{(i)}$$

Suppose that $\hat{\theta} = x_{(r)}$
$$\ell(\hat{\theta}) = (n-2r)x_{(r)} + \sum_{i=1}^{r}x_{(i)} - \sum_{i=r+1}^{n}x_{(i)}$$
if $r=m+1$, $\ell(\hat{\theta}) = (2m+1-2m-2)x_{(m+1)} + \sum_{i=1}^{m+1}x_{(i)} - \sum_{i=m+2}^{2m+1}x_{(i)}$
$$= -x_{(m+1)} + \sum_{i=1}^{m+1}x_{(i)} - \sum_{i=m+2}^{2m+1}x_{(i)}$$
if $r=m$, $\ell(\hat{\theta}) = (2m+1-2m)x_{(m)} + \sum_{i=1}^{m}x_{(i)} - \sum_{i=m+1}^{2m+1}x_{(i)}$
$$= x_m + \sum_{i=1}^{m}x_{(i)} - \sum_{i=m+1}^{2m+1}x_{(i)}$$
$$\therefore \ell(\hat{x}_{(m+1)}) > \ell(\hat{x}_{(m)})$$

Ex. 6.2.7. Maximum Likelihood Estimation in uniform distribution

Let $X_1, \cdots, X_n$ be random samples from $U[0,\theta]$ $(0<\theta<+\infty)$

$$f_X(x) = \frac{1}{\theta}I_{[0,\theta]}(x)$$

$$\log f_X(x) = -\log\theta + \log I_{[0,\theta]}(x)$$
$$\ell(\theta) = \sum_{i=1}^{n}\log f_X(x) = -n\log\theta + \sum_{i=1}^{n}\log I_{[0,\theta]}(x_i)$$
$$= -n\log\theta + \sum\log I_{[x_{(i)},\infty)}^{(\theta)}$$
$$= -n\log\theta + \log I_{[x_{(n)},\infty)}^{(\theta)}$$

← 여기서 문제는 $I_{[x_{(i)},\infty)}^{(\theta)} = 0$ 인 경우 log 값이 정의가 되지 않는다. 따라서 교재에서는 log를 사용하지 않는다.

$$\hat{\theta}^{MLE} = \underset{\theta\in\Omega_0}{\operatorname{argmax}}\,\ell(\theta) = x_{(n)}$$

[Textbook] : $pdf(x_1,\cdots,x_n;\theta) = \prod_{i=1}^{n}\frac{1}{\theta}I_{[0,\theta]}(x_i) = \theta^{-n}I_{[0\le\min x_i,\ \max x_i\le\theta]}$

$L(\theta) = \theta^{-n}I_{[x_{(n)},+\infty]}(\theta)$
$\therefore \hat{\theta} = x_{(n)}$

Ex. 6.2.8. When mle is not unique.

Let $X_1, \cdots, X_n$ be random samples from $U[\theta-1, \theta+1]$ $(-\infty < \theta < +\infty)$

$$f_X(x) = \frac{1}{(\theta+1)-(\theta-1)} I_{[\theta-1, \theta+1]}(x) = \frac{1}{2} I_{[\theta-1, \theta+1]}(x)$$

$$L(\theta; x) = \prod_{i=1}^{n} f_X(x_i; \theta) = 2^{-n} I_{[\theta-1 \le x_{(1)},\ x_{(n)} \le \theta+1]}$$

$$= 2^{-n} I_{[x_{(n)}+1,\ x_{(1)}-1]}(\theta)$$

$$\therefore \hat{\theta}^{MLE} = \underset{\theta \in \Omega_\theta}{\operatorname{argmax}}\ 2^{-n} I_{[x_{(n)}+1,\ x_{(1)}-1]}(\theta) = \left\{ \theta : x_{(n)}+1 \le \theta \le x_{(1)}-1 \right\}$$

For instance, $a(X_{(n)}+1) + (1-a)(X_{(1)}+1)$, $0 \le a \le 1$ are all mle of θ.

In this case, other criteria, such as efficiency or interpretability, should be additionally considered.

6.3. MAXIMUM LIKELIHOOD ESTIMATION IN THE MULTIVARIATE CASE.

Let's think about the maximum likelihood estimation of parameters that are multidimensional vector such as $(\mu, \sigma^2)^t$ in $N(\mu, \sigma^2)$

Definition: Maximum likelihood estimation of multidimensional parameters.

Let $x = (x_1, \cdots x_n)^t$ be the observation of random samples $X_1, \cdots, X_n$ from one of the models with pdf $f(x;\theta)$, $\theta = (\theta_1, \cdots, \theta_k)^t \in \Omega$

(a) Likelihood / log likelihood functions: Function of parameter vector $\theta = (\theta_1, \cdots, \theta_k)^t \in \Omega_\theta$

$$L(\theta; x) = \prod_{i=1}^{n} f(x_i; \theta),\ \theta = (\theta_1, \cdots, \theta_k)^t \in \Omega$$

$$\ell(\theta; x) = \log L(\theta; x) = \sum_{i=1}^{n} \log f(x_i; \theta)$$

(b) Maximum Likelihood Estimation of parameter vector: $\hat{\theta}^{MLE}$, which maximizes the likelihood function is called the mle of $\theta = (\theta_1, \cdots, \theta_k)^t$.

$$L(\hat{\theta}^{MLE}(x); x) = \max_{\theta \in \Omega} L(\theta; x),\ \hat{\theta}^{MLE}(x) \in \Omega$$

(c) Maximum Likelihood Estimation of each element in a parameter vector: Each element in $\hat{\theta}^{MLE}(x)$ is an mle of its corresponding element in $\theta = (\theta_1, \cdots, \theta_k)^t$. i.e.,

$$\hat{\theta}^{MLE}(x) = (\hat{\theta}_1^{MLE}(x), \cdots, \hat{\theta}_k^{MLE}(x))$$

The process of Maximum Likelihood Estimation of multidimensional parameter is similar to that of finding the maximum of multivariate function. You can find the maximum of each variable Step-by-Step or use the likelihood equation just like the single parameter case.

Ex. 6.3.1. Maximum Likelihood Estimation in DE distribution

Let $X_1, \cdots, X_n$ $(n = 2m+1)$ be random samples from $DE(\mu, \sigma)$ $(-\infty < \mu < +\infty,\ \sigma > 0)$. Find the mle of $\theta = (\mu, \sigma)^t$.

The pdf of $DE(\mu, \sigma)$ is: $pdf(x;\theta) = \frac{1}{2\sigma} e^{-|(x-\mu)/\sigma|}$, $-\infty < x < +\infty$, $\theta = (\mu, \sigma)^t$

$$\log f_X(x) = -\log 2\sigma - |(x-\mu)/\sigma|$$

$$\ell(\theta; x) = -n \log 2\sigma - \sum_{i=1}^{n} |(x_{(i)} - \mu)/\sigma|$$

From ⟨Ex. 6.2.6⟩, we know that the log likelihood function achieves its maximum at $\hat{\mu} = x_{(m+1)}$, the median, for fixed σ.

Thus, if $x_{(1)} < \cdots < x_{(n)}$ denotes the order statistics of the observations,

$$\max_{-\infty < \mu < +\infty} l(\mu, \hat{\sigma}) = l(\hat{\mu}, \sigma),\ \hat{\mu} = x_{(m+1)}\ (n = 2m+1)$$

If we investigate the increase/decrease of $l(\hat{\mu}, \sigma)$,

$$\frac{\partial}{\partial \sigma} l(\hat{\mu}, \sigma) = -\frac{n}{\sigma} + \frac{1}{\sigma^2} \sum_{i=1}^{n} |(x_{(i)} - \mu)| > 0 \iff \sigma < \frac{1}{n} \sum_{i=1}^{n} |x_i - \hat{\mu}|$$

↑ $l(\hat{\mu}, \sigma)$에서 증가하는 구간으로 재확

$\therefore\ l(\hat{\mu},\sigma)$ achieves maximum at $\hat{\sigma}=\frac{1}{n}\sum_{i=1}^{n}|x_i-\hat{\mu}|$

$$\therefore \max_{\sigma>0}\max_{-\infty<\mu<+\infty} l(\mu,\sigma)=\max_{\sigma>0} l(\hat{\mu},\sigma)=l(\hat{\mu},\hat{\sigma}),$$

$$\hat{\mu}=x_{(m+1)},\quad \hat{\sigma}=\frac{1}{n}\sum_{i=1}^{n}|x_i-x_{(m+1)}|$$

$$\therefore \hat{\theta}^{MLE}=(\hat{\mu}^{MLE},\hat{\sigma}^{MLE})^t=\left(x_{(m+1)},\ \frac{1}{n}\sum_{i=1}^{n}|x_i-x_{(m+1)}|\right)^t \text{ a.}$$

Ex.6.3.2. Maximum Likelihood Estimation in Normal Distribution

Let $X_1,\cdots,X_n$ be random samples from the normal distribution $N(\mu,\sigma^2)$ $(-\infty<\mu<+\infty,\ \sigma>0)$. Find the mle of $\theta=(\mu,\sigma^2)^t$.

$$f_X(x;\theta)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\}$$

$$\log f_X(x;\theta)=-\frac{1}{2}\log 2\pi-\log\sigma-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}$$

$$l(\theta;x)=\sum_{i=1}^{n}\log f_X(x_i;\theta)=-\frac{n}{2}\log 2\pi-n\log\sigma-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2}$$

$$\frac{\partial}{\partial\mu}l(\theta;x)=-\frac{1}{2}\cdot-2\cdot\sum_{i=1}^{n}\frac{(x_i-\mu)}{\sigma}=\sum_{i=1}^{n}(x_i-\mu)/\sigma>0$$

$$\Rightarrow n\bar{x}-n\mu>0$$

$$\Rightarrow \bar{x}>\mu \quad \therefore \hat{\mu}^{MLE}=\bar{x}.$$

$$\therefore \max_{-\infty<\mu<+\infty} l(\mu,\sigma;x)=l(\hat{\mu},\sigma;x)$$

$$\frac{\partial}{\partial\sigma}l(\hat{\mu},\sigma;x)=-\frac{n}{\sigma}+\sum_{i=1}^{n}\frac{(x_i-\hat{\mu})^2}{\sigma^3}>0$$

$\hat{\sigma}^{MLE}$ 구할때 $\Rightarrow -n\sigma^2+\sum_{i=1}^{n}(x_i-\hat{\mu})^2>0$

$\hat{\sigma}^{2\,MLE}=(\hat{\sigma}^{MLE})^2$ 이지만 이 경우는 다음에 나오지 때문에 $\frac{\partial}{\partial\sigma^2}$가 적절.

$$\therefore \hat{\sigma}^{2\,MLE}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2>\sigma^2$$

$$\therefore \hat{\theta}^{MLE}=(\hat{\mu}^{MLE},\hat{\sigma}^{2\,MLE})=\left(\frac{1}{n}\Sigma x_i,\ \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)$$

[Textbook]

(…) $l(\mu,\sigma^2)=-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2-\frac{n}{2}\log\sigma^2-\frac{n}{2}\log 2\pi$

$$\sum_{i=1}^{n}(x_i-\mu)^2=\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\mu)$$

$$\therefore \max_{-\infty<\mu<+\infty} l(\mu,\sigma^2)=l(\hat{\mu},\sigma^2),\quad \hat{\mu}=\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

let $\theta_2=\sigma^2$ and investigate the increase/decrease of $l(\hat{\mu},\sigma^2)=l(\hat{\mu},\theta_2)$

$$\frac{\partial}{\partial\theta_2}l(\hat{\mu},\theta_2)=\frac{1}{2\theta_2^2}\sum_{i=1}^{n}(x_i-\bar{x})^2-\frac{n}{2\theta_2^2}>0 \Leftrightarrow \sigma^2=\theta_2<\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$\therefore \max_{\sigma^2>0}\max_{-\infty<\mu<+\infty} l(\mu,\sigma^2)=\max_{-\infty<\mu<+\infty} l(\mu,\sigma^2)=l(\hat{\mu},\hat{\sigma}^2),\ \hat{\mu}=\frac{1}{n}\Sigma x_i,\ \hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$\therefore \hat{\theta}^{MLE}=(\hat{\mu}^{MLE},\hat{\sigma}^{2\,MLE})=\left(\bar{x},\ \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)$$

⟨Theorem 6.3.1: Function of Parameters and Maximum Likelihood Estimation⟩

If the mle of θ exists and $\eta=(\eta_1,\eta_2)^t=(g_1(\theta),g_2(\theta))^t$, $\theta\in\Omega$ is a one-to-one transformation, the mle of $\eta_1=g_1(\theta)$ is:

$$\hat{\eta}_1^{MLE}=g_1(\hat{\theta}^{MLE})$$

(proof) From ⟨theorem 6.2.3⟩, we know that the mle of one-to-one transformation

$$\eta=(\eta_1,\eta_2)^t=(g_1(\theta),g_2(\theta))^t=g(\theta),\ \theta\in\Omega \text{ is}$$

$$\hat{\eta}^{MLE}=(\hat{\eta}_1^{MLE},\hat{\eta}_2^{MLE})^t=(g_1(\hat{\theta}^{MLE}),g_2(\hat{\theta}^{MLE}))^t$$

Meanwhile, $\eta_1=g_1(\theta)$ is the first element of the multidimensional parameter $\eta=(\eta_1,\eta_2)^t=(g_1(\theta),g_2(\theta))^t$ and from the definition of each element's mle, the mle of $\eta_1=g_1(\theta)$ is the first element of $\hat{\eta}^{MLE}$:

$$\hat{\eta}_1^{MLE}=g_1(\hat{\theta}^{MLE}).$$

When we find the mle of $\eta_1=g(\theta)$ using ⟨theorem 6.3.1⟩, we must find the other function $\eta_2=g_2(\theta)$ so that

$$\eta=(\eta_1,\eta_2)^t=(g_1(\theta),g_2(\theta))^t,\ \theta\in\Omega$$

is a one-to-one transformation.

Ex. 6.3.3.

*Signum

(a) Let us define the function sgn(μ) as:

$$\mathrm{sgn}(\mu):=\begin{cases}+1, & \mu>0\\ 0, & \mu=0\\ -1, & \mu<0\end{cases}$$

Then the function $\eta=(|\mu|,\mathrm{Sig}(\mu),\sigma^2)^t$ is a one-to-one transformation of $\theta=(\mu,\sigma^2)^t$. Hence the mle of η, using n random samples, is:

$$\hat{\eta}^{MLE}=(|\bar{x}|,\mathrm{Sig}(\bar{x}),\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2)^t$$

From above, the mle of $|\mu|$ is $\hat{\eta}_1^{MLE}=|\bar{x}|$

* Key takeaway: Sig(x), |x| is one-to-one with x

(b) For the parameter $\sigma^2=p(1-p)$ in Bernoulli(p) $(0\le p\le 1)$,

$$\eta=(p(1-p),\ \mathrm{sgn}(p-1/2))^t$$

is a one-to-one transformation of p. Hence, the mle of $\eta_1=p(1-p)$ using n random samples, is

$$\hat{\eta}_1^{MLE}=\hat{\sigma}^{2\,MLE}=\hat{p}^{MLE}(1-\hat{p}^{MLE}),\quad \hat{p}^{MLE}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

In the multidimensional case, similar to the one-dimensional case, we can find the mle among the roots $\hat{\theta}(x)$ of the likelihood equation:

$$\dot{l}(\theta; x) = (\partial l(\theta;x)/\partial\theta_1, \cdots, \partial l(\theta;x)/\partial\theta_K)^t = 0$$

Relating to the above, the following theorems are the expansion of 〈Theorem 6.2.1〉 and 〈Theorem 6.2.2〉 to the multidimensional cases.

〈Theorem 6.3.2〉

If the function $l(\theta)$ defined on the open box $\Omega_0 = (a_1,b_1)\times\cdots\times(a_K,b_K) \subset \mathbb{R}^K$ is double differentiable (partial differentiation; 편미분) and its double derivative[92] is continuous and:

$$c^t \ddot{l}(\theta) c < 0, \quad \forall c \neq 0, \quad \forall \theta \in \Omega_0 \text{ and } \dot{l}(\hat{\theta}) = 0, \hat{\theta} \in \Omega_0.$$

Then, $\max_{\theta\in\Omega_0} l(\theta) = l(\hat{\theta})$

〈Theorem 6.3.3〉

If the function $l(\theta)$ defined on the open box $\Omega_0 = (a_1,b_1)\times\cdots\times(a_K,b_K) \subset \mathbb{R}^K$ is double differentiable (partial differentiation; 편미분) and its double derivative is continuous and:

$$c^t \ddot{l}(\theta) c < 0, \quad \forall c \neq 0, \quad \forall \theta \in \Omega_0 \text{ and } \lim_{\theta\to\partial(\Omega_0)} l(\theta) = -\infty$$[93]

The root, $\hat{\theta}$, of the equation $\dot{l}(\theta)=0$ exists and is unique.

The following theorem's proof will be omitted as it is identical to the one-dimension case.

〈Theorem 6.3.4. Maximum Likelihood Estimation in the Multi-parameter Exponential Family〉

Let the distribution's $pdf(x;\eta) = \exp\left\{\sum_{i=1}^n \eta_i T_i(x) - A(\eta) + S(x)\right\}$, $x \in \mathcal{X}$, $\eta = (\eta_1,\cdots,\eta_K)^t \in N$

and conditions (i),(ii) and (iii) be satisfied[94]

$X_1,\cdots,X_n$ are random samples from this distribution

If the likelihood equation[95] $\dot{A}(\eta) = \frac{1}{n}\sum_{i=1}^n T(x_i) \to E_\eta[T(x_i)] = \frac{1}{n}\sum_{i=1}^n T(x_i)$, $\eta \in N$ has the root $\hat{\eta} = \hat{\eta}(x_1,\cdots,x_n)$, $\hat{\eta}$ is the mle of η.

(i) The support of the distribution, $\mathcal{X}$, is invariant to the parameter.
(ii) Parameter space is inclusive of an open-box $(a_1,b_1)\times\cdots\times(a_K,b_K) \subset R^K$
(iii) $c^t T(x) = c_1T_1(x) + \cdots + c_KT_K(x)$ is not a constant. i.e.,
$*c_i \neq 0, \cdots, c_K \neq 0$ $\quad \mathrm{Var}(c^t T(x_i)) > 0 \ \forall c \in \mathbb{R}^K : c \neq 0$

In many cases when 〈Theorem 6.3.4〉 is utilized, similar to the one-dimension case, the parameter η used in the expression of exponential families is a one-to-one transformation of $\theta \in \Omega$:

$$\eta = g(\theta) = (g(\theta_1), \cdots, g(\theta_K))^t$$

In other words, the pdf in 〈Theorem 6.3.4〉 is

$$pdf(x;\theta) = \exp\left\{\sum_{i=1}^n g_j(\theta)T_j(x) - A(g(\theta)) + S(x)\right\}, \ x \in \mathcal{X}, \ \theta \in \Omega$$

Hence, when the conditions given in 〈Theorem 6.3.4〉 are satisfied, by 〈Theorem 6.3.3〉 – a theorem on one-to-one transformation of parameters –, the mle of θ is given as the root of the equation:

$$E_\theta[T(x_i)] = \frac{1}{n}\sum_{i=1}^n T(x_i), \ \theta \in \Omega$$

92) $\ddot{l}(\theta) = \left(\frac{\partial^2}{\partial\theta_i\partial\theta_j} l(\theta)\right)$ 은 헤시안 행렬이라고 불리우는 이차편도함수들의 행렬을 뜻한다.

93) 여기에서 $\partial(\Omega_0)$는 열린구간다면체 Ω_0의 경계 (boundary)를 나타내는 기호이고, $\theta \to \partial(\Omega_0)$가 뜻하는 것은 θ의 각성분이 유한구간의 끝점으로 가거나 또는 무한히 커지거나 작아지는 것을 뜻한다.

94) 이러한 조건을 만족하는 〈정리 6.2.4〉에 주어진 형태의 확률밀도함수의 집합을 다중모수지수족 (multi-parameter exponential family of pdfs) 이라고 한다.

Definition: one-to-one transformation in exponential family and likelihood functions.

If the population pdf is:

$$pdf(x;\theta) = \exp\left\{g(\theta)^t T(x) - A(g(\theta)) + S(x)\right\}, \ x \in \mathcal{X}, \ \theta \in \Omega$$

and $g(\theta)$ is reparameterized to $\eta = g(\theta)$ and the conditions (i),(ii) & (iii) are satisfied from 〈Theorem 6.3.4〉, the likelihood equation of θ is:

$$E_\theta[T_1(X_i)] = \frac{1}{n}\sum_{i=1}^n T_1(x_i), \ \cdots, \ E_\theta[T_K(X_i)] = \frac{1}{n}\sum_{i=1}^n T_K(x_i), \ \theta \in \Omega.$$

95) 여기에서 $\dot{A}(\eta) = (\partial A/\partial\eta_1, \cdots, \partial A/\partial\eta_K)^t$, $T(x) = (T_1(x), \cdots, T_K(x))^t$

Ex. 6.3.4. Normal Distribution $N(\mu,\sigma^2)$, $-\infty < \mu < +\infty$, $\sigma^2 > 0$ as a multi-parameter exponential family.

$$f(x;\theta) = (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\}, \ -\infty < x < +\infty, \ \theta = (\mu,\sigma^2)^t$$

let $\eta_1 = \frac{\mu}{\sigma^2}$, $\eta_2 = -\frac{1}{2\sigma^2} \Rightarrow \mu = -\frac{1}{2}\frac{\eta_1}{\eta_2}$, $\sigma^2 = -\frac{1}{2\eta_2}$

$$f(x;\theta) = \exp\left\{-\frac{1}{2}\log(2\sigma^2) - \frac{1}{2}\log(\pi) - \frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\}$$

$$= \exp\left\{\frac{1}{2}\log(-\eta_2) - \frac{1}{2}\log(\pi) - \frac{1}{2}\frac{\left(x+\frac{1}{2}\frac{\eta_1}{\eta_2}\right)^2}{-\frac{1}{2\eta_2}}\right\}$$

$$= \exp\left\{\frac{1}{2}\log(-\eta_2) - \frac{1}{2}\log(\pi) + \eta_2\left(x^2 + x\frac{\eta_1}{\eta_2} + \frac{\eta_1^2}{4\eta_2^2}\right)\right\}$$

$$= \exp\left\{\frac{1}{2}\log(-\eta_2) - \frac{1}{2}\log(\pi) + \eta_2 x^2 + \eta_1 x + \frac{\eta_1^2}{4\eta_2}\right\}$$

$$\therefore A(\eta) = -\frac{1}{2}\log(-\eta_2) - \frac{\eta_1^2}{4\eta_2}$$

Conditions (i), (ii) and (iii) are Satisfied.

log likelihood function $\ell(\eta_1,\eta_2;x) = \sum_{i=1}^{n} \log f_X(x_i)$

$$= \sum_{i=1}^{n} \frac{1}{2}\log(-\eta_2) - \frac{1}{2}\log(\pi) + \eta_2 x^2 + \eta_1 x + \frac{\eta_1^2}{4\eta_2}$$

$$= -\frac{n}{2}\log(\pi) + \eta_2 \sum_{i=1}^{n} x_i^2 + \eta_1 \sum_{i=1}^{n} x_i - n A(\eta)$$

$$\frac{\partial}{\partial \eta_1}\ell(\theta;x) = \sum x_i - n\cdot\frac{\partial}{\partial \eta_1}A(\eta) \quad * \ \frac{\partial}{\partial \eta_1}A(\eta) = E_\eta[T_1(x)] = E_\theta[T_1(x)] = \mu$$

Expectation does not change even after parameter transformation.

$$\therefore \left.\frac{\partial}{\partial \eta_1}\ell(\theta;x)\right|_{\eta_1=\hat{\eta}_1} = 0$$

$$\Rightarrow \sum x_i - n\frac{\partial}{\partial \eta_1}A(\eta)\Big|_{\eta_1=\hat{\eta}_1} \Rightarrow \sum x_i - n E_{\hat{\eta}}[T_1(x)] \quad * \ T_1(x)=x,\ T_2(x)=x^2$$

$$\Rightarrow \sum x_i - n E_{\hat{\theta}}[T_1(x)] \Rightarrow \sum x_i - n\hat{\mu} = 0$$

$$\therefore \hat{\mu}^{MLE} = \frac{1}{n}\sum x_i = \bar{x}$$

$$\frac{\partial}{\partial \eta_2}\ell(\theta;x) = \sum x_i^2 - n\frac{\partial}{\partial \eta_2}A(\eta_2) = \sum x_i^2 - n E_\eta[T_2(x)] = \sum x_i^2 - n E_\theta[T_2(x)] = \sum x_i^2 - n(\sigma^2+\mu^2)$$

$$\left.\frac{\partial}{\partial \eta_2}\ell(\theta;x)\right|_{\eta_2=\hat{\eta}_2} = 0 \Leftrightarrow \sum x_i^2 - n E_{\hat{\theta}}[T_2(x)] = \sum x_i^2 - n(\hat{\sigma}^2+\hat{\mu}^2) = 0 \quad \therefore \hat{\sigma}^2 = \frac{1}{n}\sum x_i^2 - \hat{\mu}^2$$

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

[Textbook]

1…) 이항분포 크기 n인 랜덤표본에 기초한 로그가능도함수

$$\ell(\eta_1,\eta_2) = \eta_1\sum_{i=1}^{n} x_i + \eta_2\sum_{i=1}^{n} x_i^2 - n A(\eta)$$

로 주어지고, 모수 $\theta=(\mu,\sigma^2)^t$를 이용하여 가능도 방정식을 나타내면,

$$E_\theta[X_i] = \frac{1}{n}\sum_{i=1}^{n} x_i,\ E_\theta[x_i^2] = \frac{1}{n}\sum_{i=1}^{n} x_i^2,\ \theta=(\mu,\sigma^2)^t,\ -\infty<\mu<+\infty,\ \sigma^2>0$$

따라서 μ, σ^2의 최대가능도 추정값을 각각 $\hat{\mu}, \hat{\sigma}^2$ 이라고 하면

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x},\ \hat{\sigma}^2 + (\hat{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

$$\therefore \hat{\mu}^{MLE} = \bar{x},\ \hat{\sigma}^{2\,MLE} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

Ex. 6.3.5. Bivariate Normal Distribution Model

The pdf of bivariate normal distribution $N(\mu_1,\mu_2;\sigma_1^2,\sigma_2^2,\rho)$ $(-\infty<\mu_j<+\infty,\ \sigma_j^2>0\ (j=1,2),\ -1<\rho<1)$ is:

$$|\det(2\pi\Sigma)|^{-1/2}\exp\left\{-\frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu)\right\},\quad \Sigma = \begin{bmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{bmatrix}$$

$$\det(\Sigma) = \sigma_1^2\sigma_2^2 - \rho^2\sigma_1^2\sigma_2^2 = \sigma_1^2\sigma_2^2(1-\rho^2)$$

$$\therefore |\det(2\pi\Sigma)|^{-1/2} = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}$$

$$(x-\mu)^t\Sigma^{-1}(x-\mu) = \begin{bmatrix}x' & y'\end{bmatrix}\begin{bmatrix}\sigma_2^2 & -\rho\sigma_1\sigma_2\\ -\rho\sigma_1\sigma_2 & \sigma_1^2\end{bmatrix}\begin{bmatrix}x'\\ y'\end{bmatrix}\cdot\frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)},\quad x'=x-\mu_1,\ y'=y-\mu_2$$

$$= \left(\sigma_2^2x'^2 - 2\rho\sigma_1\sigma_2\cdot x'y' + \sigma_1^2y'^2\right)\frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)}$$

$$= \frac{1}{1-\rho^2}\left\{\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right\} = Q$$

$$\therefore \text{pdf}\,f(x,y;\theta) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2}Q\right\}$$

Textbook η_1, η_2 are incorrect

let $\eta_1 = \frac{\mu_1}{(1-\rho^2)\sigma_1^2}$, $\eta_2 = \frac{\mu_2}{(1-\rho^2)\sigma_2^2}$, $\eta_3 = \frac{\rho}{(1-\rho^2)\sigma_1\sigma_2}$,

$\eta_4 = -\frac{1}{2(1-\rho^2)\sigma_1^2}$, $\eta_5 = -\frac{1}{2(1-\rho^2)\sigma_2^2}$

$$Q = \frac{1}{1-\rho^2}\left\{\left(\frac{x-\mu_1}{\sigma}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right\}$$

$$= \frac{1}{1-\rho^2}\left\{\frac{1}{\sigma_1^2}(x^2 - 2x\mu_1 + \mu_1) - 2\frac{\rho}{\sigma_1\sigma_2}(xy - x\mu_2 - y\mu_1 + \mu_1\mu_2) + \frac{1}{\sigma_2^2}(y^2 - 2y\mu_2 + \mu_2^2)\right\}$$

$$\therefore -\frac{1}{2}Q = \left[-\frac{1}{2(1-\rho^2)}\left\{-\frac{2\mu_1}{\sigma_1^2} - \frac{2\rho\mu_2}{\sigma_1\sigma_2}\right\}\right]x +$$

$$\left[-\frac{1}{2(1-\rho^2)}\left\{-\frac{2\mu_2}{\sigma_2^2} - \frac{2\rho\mu_1}{\sigma_1\sigma_2}\right\}\right]y +$$

$$\frac{\rho}{(1-\rho^2)\sigma_1\sigma_2}xy - \frac{1}{2(1-\rho^2)\sigma_1^2}x^2 - \frac{1}{2(1-\rho^2)\sigma_2^2}y^2 + A(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2)$$

$$= \eta_1 x + \eta_2 y + \eta_3 xy + \eta_4 x^2 + \eta_5 y^2 + A(\eta)$$

$$\therefore \text{pdf}\,f(x;\theta) = \exp\left\{\eta_1 x + \eta_2 y + \eta_3 xy + \eta_4 x^2 + \eta_5 y^2 + A(\eta)\right\}$$

The range of parameters are: $-\infty<\eta_j<+\infty$ $(j=1,2,3)$, $\eta_j<0$ $(j=4,5)$ and we see that the conditions in ⟨Theorem 6.3.2⟩ are satisfied. Using the parameter $\theta=(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)^t$, the likelihood equation based on n random samples is:

$$E_\theta[X_i] = \frac{1}{n}\sum_{i=1}^{n} x_i,\quad E_\theta[Y_i] = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$E_\theta[X_iY_i] = \frac{1}{n}\sum_{i=1}^{n} x_iy_i,\ E_\theta[X_i^2] = \frac{1}{n}\sum x_i^2,\ E_\theta[Y_i^2] = \frac{1}{n}\sum y_i^2$$

The if we denote the mle of $\mu_j, \sigma_j^2\ (j=1,2), \rho$ as $\hat{\mu}_j, \hat{\sigma}_j\ (j=1,2), \hat{\rho}$ respectively,

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x},\quad \hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$$

$$\hat{\rho}\hat{\sigma}_1\hat{\sigma}_2 + \hat{\mu}_1\hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} x_iy_i,\quad \hat{\sigma}_1^2 + \hat{\mu}_1^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2,\ \hat{\sigma}_2 + \hat{\mu}_2^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2$$

$$\therefore \hat{\rho} = \frac{\frac{1}{n}\sum x_iy_i - \left[\frac{1}{n}\sum_{i=1}^{n} x_i\right]\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right]}{\sqrt{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]\left[\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2\right]}}$$

Ex. 6.3.6. Multinomial Distribution Model $\text{Multi}(n,(p_1,\cdots,p_K)^t)$ $p_1+\cdots+p_K=1$ $p_j>0$, $j=1,2,\cdots,K$

$X=(X_1,\cdots,X_K)^t \sim \text{Multi}(n,(p_1,\cdots,p_K)^t)$

The pdf of X is:

$$f(x;p) = \binom{n}{x_1,\cdots,x_K}p_1^{x_1}\cdots p_K^{x_K},\quad x_1+\cdots+x_n=n$$

$$x_j=0,\cdots,n\ (j=1,\cdots,K),\quad p=(p_1,\cdots,p_K)^t$$

Here, $X_1+\cdots+X_K=n$, $p_1+\cdots+p_K=1$. To eliminate such dependency, Let

$$Y_j=X_j,\ \mu_j=p_j\ (j=1,\cdots,r),\ Y=(Y_1,\cdots,Y_r)^t,\ \mu=(\mu_1,\cdots,\mu_r)^t,\ r=K-1$$

and

$$\eta_j = \log(p_j/p_K)\quad (j=1,\cdots,r)$$

with η as our parameter, the pdf of $Y=(Y_1,\cdots,Y_r)^t$ is:

$$pdf(y;\eta)=\exp\{\eta_1 y_1+\cdots+\eta_r y_r - n\,A(\eta)+S(y)\}$$

$$A(\eta)=\log(1+e^{\eta_1}+\cdots+e^{\eta_r}),\quad S(y)=\log\binom{n}{y_1\cdots y_r x_K},\quad x_K=n-y_1-\cdots-y_r$$

the parameter space is $-\infty<\eta_j<+\infty$ $(j=1,\cdots,r)$ and we see that the conditions in ⟨Theorem 6.3.4⟩ are satisfied.

Hence, if we write the likelihood equation using parameter $p=(p_1,\cdots,p_K)^t$,

$$E_p[Y_1]=y_1,\ \cdots,\ E_p[Y_r]=y_r,\quad 0<p_j\ (j=1,2,\cdots,r),\quad p_1+\cdots+p_r<1)$$

Now if we denote the mle of $p=(p_1,\cdots,p_K)^t$ as $(\hat{p}_1,\cdots,\hat{p}_K)^t$,

$$n\hat{p}_1=y_1,\ \cdots,\ n\hat{p}_r=y_r,\quad 0<\hat{p}_j\ (j=1,\cdots,r),\quad \hat{p}_1+\cdots+\hat{p}_r<1$$

Therefore, as ⟨Ex.6.2.5⟩-(a), if $0<x_j/n<1$ $(j=1,\cdots,K)$,

$$\hat{p}_j=x_j/n\quad (j=1,\cdots,K)$$

6.4. ASYMPTOTIC PROPERTY OF MLE

Consistency is a basic property required for an estimator. In addition to consistency, if we were to know the limiting distribution of the estimator, we could make further statistical inferences, such as the confidence interval around a parameter. In this section we will cover the consistency of mle and the derivation of the limiting distribution of mle using the properties of the log likelihood function.

Ex. 6.4.1 Consistency of MLE and its limiting distribution.

(a) Bernoulli Distribution Model. Bernoulli(p), $0\le p\le 1$:

the mle of p using size n random samples $X_1,\cdots,X_n$ is:

$$\hat{p}=\frac{1}{n}\sum_{i=1}^{n}x_i,\quad f_X(x)=p^x(1-p)^{1-x}$$

$$E[X]=\sum_{x=0,1}x\,p^x(1-p)^{1-x}=p$$

$$\frac{1}{n}\sum x_i \xrightarrow[n\to\infty]{p\ \text{parameter}} E[X]=p,\ \forall p:0\le p\le 1\quad (\because LLN)$$

$$E[X^2]=\sum_{x=0,1}x^2p^x(1-p)^{1-x}=p$$

$$\therefore Var(X)=p-p^2=p(1-p)$$

$$\lim_{n\to\infty}P_p(|\hat{p}_n^{MLE}-p|\ge\epsilon)=0\quad \forall\epsilon>0,\ \forall p:0\le p\le 1$$

$$\frac{\sqrt{n}(\bar{x}-\mu)}{p(1-p)}\xrightarrow[n\to\infty]{d}N(0,1)\ \because CLT$$

$$\Leftrightarrow \sqrt{n}(\hat{p}-p)\xrightarrow[n\to\infty]{d}N(0,p(1-p))\quad 96)$$

96) 이러한 표현은 다음과 같은 표현을 간략하게 나타낸 것이다.

$$\sqrt{n}(\hat{p}^{MLE}-p)\xrightarrow[n\to\infty]{d}Z,\quad Z\sim N(0,p(1-p))$$

(b) Poisson Distribution Model, Poisson(λ), $\lambda\ge 0$:

$$\hat{\lambda}^{MLE}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

By the LLN, $\frac{1}{n}\sum X_i\xrightarrow[n\to\infty]{p}E[X]=\lambda\quad\forall\lambda\ge 0$

$$\Leftrightarrow \forall\epsilon>0,\ \lim_{n\to\infty}P_\lambda(|\hat{\lambda}_n-\lambda|\ge\epsilon)=0$$

$$\Leftrightarrow \forall\epsilon>0,\ \exists K(\epsilon)\ s.t.\ \text{if } n>K(\epsilon),\ P(|\hat{\lambda}_n-\lambda|\ge\epsilon)=0$$

By CLT, $\sqrt{n}(\bar{x}-\mu)\xrightarrow[n\to\infty]{d}N(0,Var(x))\overset{d}{=}N(0,\lambda)$

$$\therefore \sqrt{n}(\hat{\lambda}^{MLE}-\lambda)\xrightarrow[n\to\infty]{d}N(0,\lambda)\quad\forall\lambda\ge 0$$

(c) Exponential Distribution Model. Exp(θ), $\theta>0$:

$$f_X(x;\theta)=\frac{1}{\theta}e^{-x/\theta}$$

$$\log f_X(x;\theta)=-\log\theta-x/\theta,\quad \ell(\theta)=-n\log\theta-\frac{n\bar{x}}{\theta}$$

$$\dot{\ell}(\theta)=-\frac{n}{\theta}+\frac{n\bar{x}}{\theta^2}>0\Leftrightarrow\bar{x}>\theta\ \therefore\hat{\theta}^{MLE}=\bar{x}$$

$$\therefore\hat{\theta}^{MLE}=\bar{x}\quad\therefore\hat{\theta}^{MLE}\xrightarrow[n\to\infty]{P_\theta}E[X]=\theta\quad\forall\theta>0$$

$$\sqrt{n}(\bar{x}-\mu)\xrightarrow[n\to\infty]{d}N(0,\theta^2)$$

$$\therefore\sqrt{n}(\hat{\theta}^{MLE}-\theta)\xrightarrow[n\to\infty]{d}N(0,\theta^2)\quad(\because CLT)$$

(d) Normal Distribution Model $N(\mu,\sigma^2)$, $-\infty<\mu<+\infty$, $\sigma^2>0$:

$\hat{\mu}^{MLE}=\frac{1}{n}\sum_{i=1}^{n}X_i$, $\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2=\frac{1}{n}\sum_{1}^{n}X_i^2-\left(\frac{1}{n}\sum X_i\right)^2$

$\hat{\mu}_n\xrightarrow[n\to\infty]{P_\theta}E[X]=\mu$, $\hat{\sigma}_n^2\xrightarrow[n\to\infty]{P_\theta}E[X^2]-E[X]^2=(\sigma^2+\mu^2)-\mu^2=\sigma^2$ $\quad\forall\mu:-\infty<\mu<+\infty$, $\forall\sigma^2:\sigma^2>0$

$\sqrt{n}(\bar{X}-\mu)\xrightarrow[n\to\infty]{d}N(0,\sigma^2)\Leftrightarrow\sqrt{n}(\hat{\mu}_n-\mu)\xrightarrow[n\to\infty]{d}N(0,\sigma^2)$

분포수렴

$\sqrt{n}(\bar{X}-\mu)\sim N(0,\sigma^2)\Leftrightarrow\sqrt{n}(\hat{\mu}_n-\mu)\sim N(0,\sigma^2)$ ∵ (Theorem 4.2.4)

$n\hat{\sigma}^2/\sigma^2\sim\chi^2(n-1)$ ∵ Theorem 4.2.a.

From ⟨Ex. 5.3.2a⟩, $\sqrt{n}(\hat{\sigma}_n^{2\,MLE}-\sigma^2)\xrightarrow[n\to\infty]{d}N(0,(\rho_4+2)\sigma^4)$ ☆☆

(e) Bivariate Normal Distribution Model.

$N(\mu_1,\mu_2;\sigma_1^2,\sigma_2^2,\rho)$ $(-\infty<\mu_j<+\infty,\ \sigma_j^2>0\ (j=1,2),\ -1<\rho<1)$

$\hat{\mu}_1^{MLE}=\frac{1}{n}\sum_{i=1}^{n}X_i$, $\hat{\mu}_2=\frac{1}{n}\sum_{i=1}^{n}Y_i$, $\sigma_1^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$, $\sigma_2^2=\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})^2$

$\hat{\rho}=\dfrac{\frac{1}{n}\sum X_iY_i-\left(\frac{1}{n}\sum X_i\right)\left(\frac{1}{n}\sum Y_i\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$

$\sqrt{n}(\hat{\mu}_1^{MLE}-\mu_1)\xrightarrow[n\to\infty]{d}N(0,\sigma_1^2)$, $\sqrt{n}(\hat{\mu}_2^{MLE}-\mu_2)\xrightarrow[n\to\infty]{d}N(0,\sigma_2^2)$

$\sqrt{n}(\hat{\sigma}_1^{2\,MLE}-\sigma_1^2)\xrightarrow[n\to\infty]{d}N(0,(\rho_4+2)\sigma_1^4)$, $\sqrt{n}(\hat{\sigma}_2^{2\,MLE}-\sigma_2^2)\xrightarrow[n\to\infty]{d}N(0,(\rho_4+2)\sigma_2^4)$

$\sqrt{n}(\hat{\rho}^{MLE}-\rho)\xrightarrow[n\to\infty]{d}N(0,(1-\rho^2)^2)$

In ⟨Ex. 6.4.1⟩, the 'mle's are cosistent and asymptotically normal. However, in all the examples that we have seen, the mle was equivalent to MME and the properties of MLEs were derived from that of MME. Now we will investigate whether the MLE has these aforementioned properties generally.

First, we will intuitively look into what kind properties are needed for a MLE to be consistent. The log likelihood function of n samples is:

$$\ell_n(\theta)=\sum_{i=1}^{n}\log f_X(x_i)$$

It is the sum of n iid random variables, thus we will express it's mean as:

$$\bar{\ell}_n(\theta)=\frac{1}{n}\sum_{1}^{n}\log f_X(x_i)$$

Also, under the condition that the MLE is unique, we will express the MLE as:[97]

$$\hat{\theta}_n=\underset{\theta\in\Omega}{\operatorname{argmax}}\,\ell_n(\theta)=\underset{\theta\in\Omega}{\operatorname{argmax}}\,\bar{\ell}_n(\theta)$$

97) 여기에서 $\underset{\theta\in\Omega}{\operatorname{argmax}}\,\bar{\ell}_n(\theta)$ 는 'argument of maximizing" 의 약자로서 $\bar{\ell}_n(\theta)$를 최대로 하는 $\theta\in\Omega$ 가 값을 뜻하는 기호이다. 이 기호가 명확한 의미를 가지려면 최대로 하는 $\theta\in\Omega$의 값이 유일하게 존재하여야 하는 것이다.

Supposing that the expectation exists, if the true parameter value is θ°, from the Law of Large Numbers,

$$\bar{\ell}_n(\theta)=\frac{1}{n}\sum_{i=1}^{n}\log f(x_i;\theta)\xrightarrow[n\to\infty]{P_\theta}E_{\theta^\circ}[\log f(x_i;\theta)]$$

From this, we can make the following conjecture:

$$Q:\ \hat{\theta}_n=\underset{\theta\in\Omega}{\operatorname{argmax}}\,\bar{\ell}_n(\theta)\xrightarrow[n]{P_{\theta^\circ}}\underset{\theta\in\Omega}{\operatorname{argmax}}\,E_{\theta^\circ}[\log f(x_i;\theta)]\ ?$$

From the following theorem, we know that $\underset{\theta\in\Omega}{\operatorname{argmax}}\,E_{\theta^\circ}[\log f(x_i;\theta)]=\theta^\circ$. Thus the aforementioned conjecture is precisely guessing the convergence of probability of $\hat{\theta}_n$ to θ°.

⟨Theorem 6.4.1. Kullback-Leibler Divergence⟩[98]

Suppose that the pdf $f_X(x;\theta)$ of random variable X satisfies the following two conditions.

(R0) (Identifiability)[99] $f(\cdot;\theta)=f(\cdot;\theta^\circ)\Rightarrow\theta=\theta^\circ$

(R1) (Common support; 공통의 support) $\{x:f(x;\theta)>0\}=\mathcal{X}$ does not depend on $\theta\in\Omega$

Let: $KL(\theta,\theta^\circ)=-E_{\theta^\circ}\left[\log\left\{f(x;\theta)/f(x;\theta^\circ)\right\}\right]$

The following holds:

(i) $KL(\theta,\theta^\circ)\geq 0\quad\forall\theta\in\Omega$

(ii) $KL(\theta,\theta^\circ)=0\quad\forall\theta=\theta^\circ$

$\log(t)=\log(1)+\frac{1}{t}\Big|_{t=1}(t-1)-\frac{1}{2t^2}\Big|_{t=1}(t^*-1)^2$

$=(t-1)-\frac{1}{2}(t^*-1)^2$

$\Rightarrow\log(t)\leq t-1\Leftrightarrow-\log t\geq-t+1$

(proof) For all $t\in\mathbb{R}$, the inequality $-\log t\geq -t+1$ is true.*

From this inequality, the following holds:

$$-\log\left\{f(x;\theta)/f(x;\theta^\circ)\right\}\geq-f(x;\theta)/f(x;\theta^\circ)+1$$

Take the expectation on both sides when the parameters of the pdf are θ° and use condition (R1):

$$KL(\theta,\theta^\circ)=E_{\theta^\circ}\left[-\log\left\{f(x;\theta)/f(x;\theta^\circ)\right\}\right]\geq E_{\theta^\circ}\left[-f(x;\theta)/f(x;\theta^\circ)+1\right]^*=0$$

$E_{\theta^\circ}\left[-f(x;\theta)/f(x;\theta^\circ)+1\right]$

$=\int_{x\in S_X}\left(-f(x;\theta)/f(x;\theta^\circ)\right)\cdot f(x;\theta^\circ)\,dx+1$

$=-1+1=0$

Thus (i) is satisfied. Meanwhile, the equality is satisfied when $t=1$, i.e., $f(x;\theta)=f(x;\theta^\circ)$. Hence (ii) is also satisfied.

⟨The Intuitive Background Behind the Consistency of MLE⟩

(Law of Large Numbers) $\bar{\ell}_n(\theta)=\frac{1}{n}\sum_{i=1}^{n}\log f(x_i;\theta)\xrightarrow[n\to\infty]{P_{\theta^\circ}}E_{\theta^\circ}[\log f(x_i;\theta)]$

(unique determination) ⇓ ⇓

(Intuitive Guess) $\hat{\theta}_n=\underset{\theta\in\Omega}{\operatorname{argmax}}\,\bar{\ell}_n(\theta)\xrightarrow[n\to\infty]{P_{\theta^\circ}}\theta^\circ=\operatorname{argmax}E_{\theta^\circ}[\log f(x_i;\theta)]$

98) 〈정리 6.4.1〉 에서의 $KL(\theta,\theta^0)$는 두 분포가 서로 다른 정도를 나타내는 측도로서 쿨백-라이블러 (Kullback-Leibler) 괴리도 (divergence) 라고 한다.

99) 식별 가능성의 정확한 뜻은 θ일 때의 분포와 θ^0일 때의 분포가 같으면 $\theta=\theta^0$ 인 것을 뜻한다.

To show the consistency of mle for the general case, it is well known that the following conditions must be satisfied, which are stronger conditions than the convergence in probability of the mean of log likelihood.[100]

(Uniform Convergence in Probability)

: $\sup_{|\theta-\theta^0|\le K} |\bar{l}_n(\theta) - E_{\theta^0}[\log f(X_i;\theta)]| \xrightarrow[n\to\infty]{P_{\theta^0}} 0$, $\forall K>0$

(Continuity of Expectation of log-likelihood)

$E_{\theta^0}[\log f(X_i;\theta)]$ is a continuous function of θ

In addition, it is known that if the log likelihood function satisfies the conditions in 〈Theorem 6.3.3〉, which is:

$$c^t \ddot{l}_n(\theta) c < 0 \quad \forall c \neq 0,\ \forall \theta \in \Omega \text{ and } \lim_{\theta \to \partial(\Omega)} l_n(\theta) = -\infty$$

the condition of (Uniform Convergence in Probability) is automatically satisfied.[101]

100) 참고서적 [P.J. Bickel & K.A. Doksum] Mathematical Statistics p.304-305 를 참조하시오

101) 참고서적 [D. Pollard] (1991) Econometric Theory, 7, p.186-p.199를 참조하시오

Q: Where is KL-Divergence used?

A: To show that $\operatorname{argmax}_{\theta\in\Omega} E_{\theta^0}[\log f(X;\theta)] = \theta^0$

How:

we know that $KL(\theta,\theta^0) = -E_{\theta^0}\left[\log\left\{f(X;\theta)/f(X;\theta^0)\right\}\right] \ge 0$

$-KL(\theta,\theta^0) = E_{\theta^0}[\log f(X;\theta) - \log f(X;\theta^0)] \le 0$

$= E_{\theta^0}[\log f(X;\theta) \le E_{\theta^0}[\log f(X;\theta^0)]$

$\therefore \operatorname{argmax}_{\theta\in\Omega} E_{\theta^0}[\log f(X;\theta)] = \theta^0$

Summary: 여기서 우리가 보이고 싶은 것은 mle의 일치성이다:

① LLN 을 이용: $\bar{l}_n(\theta;X) = \frac{1}{n}\sum \log f_X(X;\theta) \xrightarrow[n\to\infty]{P_{\theta^0}} E_{\theta^0}[\log f_X(X;\theta)]$ ← notation 주의. True parameter θ^0

② Argmax 쓰기: $\operatorname{argmax}_{\theta\in\Omega} \bar{l}_n(\theta;X) \xrightarrow[n\to\infty]{P_{\theta^0}} \operatorname{argmax}_{\theta\in\Omega} E_{\theta^0}[\log f_X(X;\theta)]$ * 추정.

③ KL-Divergence의 특성을 활용: $\operatorname{argmax}_{\theta\in\Omega} E_{\theta^0}[\log f_X(X;\theta)] = \theta^0$

* KL-Divergence: $-\log(t) \ge -t+1$

〈Theorem 6.4.2. Consistency of MLE〉

If

① the parameter space Ω is an open box in $\mathbb{R}^k$ ← KL-divergence conditions

② Population of random samples satisfy conditions (R0), (R1) from 〈Theorem 6.4.1〉.

③ log likelihood functin $\bar{l}_n(\theta)$, a function of n random samples, is double differentiable

④ Double derivative is continuous

⑤ log likelihood is strictly concave and has one maxima point (순요목 단봉 초고가능도)

: $c^t \ddot{l}_n(\theta) c < 0 \quad \forall c \neq 0,\ \forall \theta \in \Omega$ and $\lim_{\theta\to\partial(\Omega)} l_n(\theta) = -\infty$

⑥ Continuity of Expectation of log likelihood

$\forall \theta \in \Omega$, $E_{\theta^0}[\log f(X_i;\theta)]$ exists and is continuous for $\theta \in \Omega$

The $\hat{\theta}_n^{MLE}$ is a unique root of the likelihood function and is consitent: i.e.,

$$\dot{l}_n(\hat{\theta}_n^{MLE}) = 0,\ \hat{\theta}_n^{MLE} \xrightarrow[n\to\infty]{P_{\theta^0}},\ \forall \theta \in \Omega$$

" - 정리 6.4.2 의 조건들 "

Ex.6.4.2. Logistic Distribution Model $L(\theta,1)$, $-\infty<\theta<+\infty$

The pdf of the Logistic Distribution $L(\theta,1)$ is:

$$f(x;\theta) = \frac{e^{x-\theta}}{(1+e^{x-\theta})^2} = \frac{e^{-x+\theta}}{(1+e^{-x+\theta})^2},\quad -\infty < x < +\infty$$

$$\log f_X(x) = (x-\theta) - 2\log(1+e^{x-\theta}) = [-x+\theta) + 2\log(1+e^{-x+\theta}),\ -\infty < x < +\infty$$

$$\therefore l_n(\theta) = \sum_{i=1}^n \log f_X(x_i) = n(\bar{x}-\theta) - 2\sum_{i=1}^n \log(1+e^{x-\theta}) = n(-\bar{x}+\theta) + 2\sum_{i=1}^n \log(1+e^{-x_i+\theta})$$

In 〈Ex.6.2.3〉, we have proved the log likelihood is strictly concave and has a unique maxima. The continuity of expectation of log likelihood is also proved. Hence the log likelihood $\hat{\theta}^{MLE}$ is consistent.

① $l_n(\theta) = \sum_{i=1}^n \log f_X(x_i;\theta) = n(\bar{x}-\theta) - 2\sum_i^n \log(1+e^{x-\theta})$

$$\dot{l}_n(\theta) = -n - 2\sum_i^n \frac{-e^{x-\theta}}{1+e^{x-\theta}} = -n + 2\sum_i^n \frac{e^{x-\theta}}{1+e^{x-\theta}}$$

$$\ddot{l}_n(\theta) = 2\sum \frac{-e^{x-\theta}}{1+e^{x-\theta}} + 2\sum_i^n \frac{e^{x-\theta}}{(1+e^{x-\theta})^2} = 2\sum_i^n \frac{-e^{x-\theta} - e^{2(x-\theta)} + e^{x-\theta}}{(1+e^{x-\theta})^2} = -2\sum_i^n \frac{e^{2(x-\theta)}}{(1+e^{x-\theta})^2}$$

$\therefore \ddot{l}_n(\theta) < 0 \quad \forall \theta \in \Omega$

② $n(\bar{x}-\theta) - 2\sum_i^n \log(1+e^{x-\theta}) \le n(\bar{x}-\theta) - 2\sum_i^n \log(e^{x-\theta}) = n(\bar{x}-\theta) - 2\sum_i^n (x_i-\theta)$

$$= -n(\bar{x}-\theta)$$

$$\lim_{\theta\to-\infty} l_n(\theta) \le \lim_{\theta\to-\infty} -n(\bar{x}-\theta) = -\infty$$

$$\lim_{\theta\to+\infty} l_n(\theta) = \lim_{\theta\to+\infty} n(\bar{x}-\theta) - 2\sum_i^n \log(1+e^{x-\theta})$$

$$= \lim_{\theta\to+\infty} n(\bar{x}-\theta) - 0 = -\infty$$

$\therefore$ $\lim_{\theta\to\partial(\Omega)} l_n(\theta) = -\infty$ □

$\therefore \ddot{l}_n(\theta) < 0 \quad \forall \theta \in \Omega$, $\lim_{\theta\to\partial(\Omega)} l_n(\theta) = -\infty$

Ex. 6.4.3. Gamma Distribution Model Gamma(α,β), $\alpha>0$, $\beta>0$.

The pdf of Gamma(α,β) is: $f_X(x)=\frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}\exp\left[-x/\beta\right]I_{[0,\infty)}(x)$

$\log f_X(x) = -\log(\Gamma(\alpha)) - \alpha\log\beta + (\alpha-1)\log(x) - x/\beta$

$\ell(\theta;x)=\sum_1^n \log f_X(x_i;\theta) = -n\log(\Gamma(\alpha)) - n\alpha\log\beta + (\alpha-1)n\,\overline{\log(x)} - \frac{n\bar{x}}{\beta}$, $\overline{\log(x)}=\frac{1}{n}\sum_1^n\log(x_i)$, $\bar{x}=\frac{1}{n}\sum_1^n x_i$

$\dot{\ell}(\theta;x)=\begin{bmatrix}\partial\ell/\partial\alpha\\ \partial\ell/\partial\beta\end{bmatrix}=\begin{bmatrix}-n\Psi(\alpha)-n\log\beta+(\alpha-1)\overline{\log(x)}\\ -\frac{n}{\beta}+\frac{n\bar{x}}{\beta^2}\end{bmatrix}$

The Gamma distribution is an exponential family, thus the conditions in <Theorem 6.2.4> are satisfied.

Thus the roots of the likelihood equation, $\hat{\alpha}^{MLE}$ and $\hat{\beta}^{MLE}$, are consistent estimators.

Next, we will look into the limiting distribution of MLEs. We will find the limiting distribution using the Law of Large Numbers and the CLT. Thus, let's first intuitively think whether the MLEs can be expressed as the arithmetic mean (산술적 평균) of random variables.

Let $\hat{\theta}_n$ be the MLE and $\bar{\ell}_n(\theta)=n^{-1}\ell_n(\theta)=n^{-1}\sum_1^n\log f(X_i;\theta)$

Under suitable conditions:

① $\dot{\ell}_n(\theta)=\dot{\ell}(\hat{\theta})+\ddot{\ell}(\hat{\theta})(\theta-\hat{\theta})+R_n(\theta^*)\Big]$, $R_n(\theta^*)\xrightarrow[n\to\infty]{}0$

$\Rightarrow \dot{\ell}(\hat{\theta})=\dot{\ell}(\theta)-\ddot{\ell}(\hat{\theta})(\theta-\hat{\theta})+R_n(\theta^*)$ $\quad\exists\theta^*$ s.t. $|\theta^*-\hat{\theta}|\le|\hat{\theta}-\theta|$

$\Rightarrow \dot{\ell}(\hat{\theta})=\dot{\ell}(\theta)+\ddot{\ell}(\hat{\theta})(\hat{\theta}-\theta)+R_n(\theta^*)$

$\Rightarrow \dot{\ell}(\hat{\theta})\simeq\dot{\ell}(\theta)+\ddot{\ell}(\hat{\theta})(\hat{\theta}-\theta)$

$\Rightarrow \frac{1}{n}\dot{\ell}(\hat{\theta})\simeq\frac{1}{n}\dot{\ell}(\theta)+\frac{1}{n}\ddot{\ell}(\hat{\theta})(\hat{\theta}-\theta)$

$\Rightarrow \overline{\dot{\ell}(\hat{\theta})}\simeq\overline{\dot{\ell}(\theta)}+\overline{\ddot{\ell}(\hat{\theta})}(\hat{\theta}-\theta)$

$\Rightarrow \overline{\dot{\ell}(\hat{\theta})}-\overline{\dot{\ell}(\theta)}\simeq\overline{\ddot{\ell}(\hat{\theta})}(\hat{\theta}-\theta)$

$\Rightarrow (\hat{\theta}-\theta)\simeq\left[\overline{\ddot{\ell}(\hat{\theta})}\right]^{-1}\left[\overline{\dot{\ell}(\hat{\theta})}-\overline{\dot{\ell}(\theta)}\right]$

$\Rightarrow \sqrt{n}(\hat{\theta}-\theta)\simeq\left[\overline{\ddot{\ell}(\hat{\theta})}\right]^{-1}\sqrt{n}\left[\overline{\dot{\ell}(\hat{\theta})}-\overline{\dot{\ell}(\theta)}\right]=\left[-\overline{\ddot{\ell}(\hat{\theta})}\right]^{-1}\sqrt{n}\,\overline{\dot{\ell}(\theta)}$

② Under suitable conditions, by the Law of Large Numbers,

$-\overline{\ddot{\ell}}_n(\theta)=\frac{1}{n}\sum_1^n\left[-\frac{\partial^2}{\partial\theta^2}\log f(X_i;\theta)\right]\xrightarrow[n\to\infty]{P_\theta}E_\theta\left[-\frac{\partial^2}{\partial\theta^2}\log f(X_i;\theta)\right]$

$\therefore$ Let $I(\theta)=E\left[-\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right]$

$\Rightarrow \sqrt{n}(\hat{\theta}-\theta)\simeq[I(\theta)]^{-1}\sqrt{n}\,\bar{\dot{\ell}}_n(\theta)$

③ $\frac{\partial}{\partial\theta}\log f(X_i;\theta)$ $(i=1,2,\cdots,n)$ are iid. Thus, from CLT,

$\sqrt{n}\left(n^{-1}\sum_1^n\frac{\partial}{\partial\theta}\log f(X_i;\theta)-E_\theta\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right]\right)\xrightarrow[n\to\infty]{d}N\left(0,\ Var_\theta\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right]\right)$

$\therefore \sqrt{n}\left(\bar{\dot{\ell}}_n(\theta)-E_\theta\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right]\right)\xrightarrow[n\to\infty]{d}N\left(0,\ Var_\theta\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right]\right)$

④ Meanwhile $E_\theta\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right]=\int_{-\infty}^{\infty}\frac{f'(x;\theta)}{f(x;\theta)}f(x;\theta)\,dx=\int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}f(x;\theta)\,dx$

$=\frac{\partial}{\partial\theta}\int_{-\infty}^{\infty}f(x;\theta)\,dx=\frac{\partial}{\partial\theta}1=0$

$\therefore \sqrt{n}\left(\bar{\dot{\ell}}_n(\theta)-E_\theta\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right]\right)=\sqrt{n}\,\bar{\dot{\ell}}_n(\theta)\xrightarrow[n\to\infty]{d}N\left(0,\ Var_\theta\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right]\right)$

Combine ②, ③, ④

$\sqrt{n}(\hat{\theta}-\theta)\simeq[I(\theta)]^{-1}\sqrt{n}\,\bar{\dot{\ell}}(\theta)\xrightarrow[n\to\infty]{d}[I(\theta)]^{-1}N\left(0,\ Var_\theta\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)\right)$

$=N\left(0,\ [I(\theta)]^{-2}\cdot Var_\theta\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)\right)$

Hence, under suitable conditions, we can assume that the limiting distribution of MLE is a normal distribution. We call this property "asymptotic normality".

In this conjecture, the conditions that needs to be robust are:

- treatment of the remainder term,
- definition of variance,
- Conditions required for the exchange of the order of integration/differentiation in expectation calculation, etc.

<Basic requirements for MLEs asymptotic normality>

(R0) (Identifiability) $f(\cdot;\theta)=f(\cdot;\theta^*)\Rightarrow\theta=\theta^*$

(R1) (common support) $\{x: f(x;\theta)>0\}=\mathcal{X}$ that does not depend on $\theta\in\Omega$ 공통의 support

(R2) (open parameter space) The parameter space Ω is an open space in $\mathbb{R}^k$

(R3) (Differentiable log likelihood)
: The log likelihood function $\ell_n(\theta)=\sum_1^n\log f_X(x_i;\theta)$, which is a function of all observations $(x_1,\cdots,x_K)^t$, $x_i\in\mathcal{X}$ $(i=1,\cdots,n)$ is double differentiable, and its derivatives[192] $\dot{\ell}_n(\theta)$, $\ddot{\ell}_n(\theta)$ are continuous.

(R4) (Exchange of the order between Integration- or sum- and differentiation) (적분 또는 합과 미분의 순서 교환 가능)

Let $u(x)$ be a function of $x=(x_1,\cdots,x_n)^t$, and suppose that its expectation $E_\theta[u(X)]$ exists. Also if we denote the differentiation w.r.t θ denote ∂, ∂^2,

$$\partial_\theta^r E_\theta[u(X)]=\begin{cases}\int_{R^n}u(x)\,\partial_\theta^r(pdf(x;\theta))\,dx & \text{(continuous)}\\ \sum_{x\in R^n}u(x)\,\partial_\theta^r(pdf(x;\theta))\,dx & \text{(discrete)}\end{cases}\quad(r=1,2)$$

Here, $pdf(x;\theta)=\prod_{i=1}^n f(x_i;\theta)$

(R5) (Existence of Information number) $\forall\theta\in\Omega$

$$I(\theta)=Var_\theta\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)$$

is well defined in $\mathbb{R}^K$ $(K=1,2,\cdots)$ and its inverse (or inverse matrix) exists.

The first derivative of the log likelihood function $\dot{\ell}_n(\theta)$ is called the "Score function" and has a pivotal role in investigating the properties of MLE. The following theorem details the properties of $\dot{\ell}_n(\theta)$.

102) $\dot{\ell}_n(\theta) = \left[\frac{\partial}{\partial\theta_1}\ell_n(\theta), \cdots, \frac{\partial}{\partial\theta_K}\ell_n(\theta)\right]^t$, $\ddot{\ell}_n(\theta) = \left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell(\theta)\right]$

103) $\partial_\theta g(\theta) = \left[\frac{\partial}{\partial\theta_1}g(\theta), \cdots, \frac{\partial}{\partial\theta_K}g(\theta)\right]^t$, $\partial_\theta^2 g(\theta) = \left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}g(\theta)\right]$

〈Theorem 6.4.3. Properties of the Score function〉

If conditions (R0) - (R5) are satisfied, the following are true:

(a) $E_\theta\left\{\frac{\partial}{\partial\theta}\log f(X_i;\theta)\right\} = 0$

(b) $I(\theta) = Var_\theta\left\{\frac{\partial}{\partial\theta}\log f(X_i;\theta)\right\} = E_\theta\left\{-\frac{\partial^2}{\partial\theta^2}\log f(x_i;\theta)\right\}$

(proof)

(a)

$\int_{x\in S_X = \mathcal{X}} f_X(x;\theta)dx = 1 \Rightarrow \frac{\partial}{\partial\theta}\int_{x\in S_X} f_X(x;\theta)dx = 0$

* $f_X(x;\theta) = \exp\{\log f(x;\theta)\}$

$$\Rightarrow \frac{\partial}{\partial\theta}\int_{x\in S_X}\exp\{\log f(x;\theta)\}dx = \int_{x\in S_X}\frac{\partial}{\partial\theta}\exp\{\log f(x;\theta)\}dx$$
$$= \int_{x\in S_X}\left\{\frac{\partial}{\partial\theta}\log f(x;\theta)\right\}\exp\{\log f(x;\theta)\}dx$$
$$= \int_{x\in S_X}\left\{\frac{\partial}{\partial\theta}\log f(x;\theta)\right\}f(x;\theta)dx = E_\theta\left[\frac{\partial}{\partial\theta}\log f(x_i;\theta)\right]$$

$\therefore E_\theta\left[\frac{\partial}{\partial\theta}\log f(x;\theta)\right] = 0$

(b) $\frac{\partial^2}{\partial\theta^2}\int_{x\in S_X} f(x;\theta)dx = \int_{x\in S_X}\frac{\partial^2}{\partial\theta^2}\exp\{\log(f(x;\theta))\}dx = \int_{x\in S_X}\frac{\partial}{\partial\theta}\left\{\frac{\partial}{\partial\theta}\log f(x;\theta)\right\}\exp\{\log f(x;\theta)\}dx$

$$= \int\left\{\frac{\partial^2}{\partial\theta^2}\log f(x;\theta)\right\}\exp\{\log f(x;\theta)\} + \left\{\frac{\partial}{\partial\theta}\log f(x;\theta)\right\}^2\exp\{\log f(x;\theta)\}dx$$

$$= E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(x;\theta)\right] + E_\theta\left[\left\{\frac{\partial}{\partial\theta}\log f(x;\theta)\right\}^2\right] = 0$$

$$\Leftrightarrow E_\theta\left[\left\{\frac{\partial}{\partial\theta}\log f(x;\theta)\right\}^2\right] = -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(x;\theta)\right]$$

$$\Leftrightarrow E_\theta\left[\left\{\frac{\partial}{\partial\theta}\log f(x;\theta)\right\}^2\right] - E_\theta\left[\frac{\partial}{\partial\theta}\log f(x;\theta)\right]^2 = -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(x;\theta)\right]$$

$$\Leftrightarrow Var\left[\frac{\partial}{\partial\theta}\log f(x;\theta)\right] = -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(x;\theta)\right]$$

$$\Leftrightarrow Var\left[\frac{\partial}{\partial\theta}\log f(x;\theta)\right] = E_\theta\left[-\frac{\partial^2}{\partial\theta^2}\log f(x;\theta)\right]$$

〈Theorem 6.4.4. Asymptotic Normality of MLE〉

Suppose that the following conditions (R6) and (R7) are satisfied on top on (R0) - R(5) in a Probability Model (확률모형) with pdf $f(x_i;\theta), \theta\in\Omega$.

(R6) (Consistency) the MLE $\hat{\theta}_n^{MLE}$ is the unique root of the likelihood equation

$$\dot{\ell}_n(\theta) = \sum_{i=1}^n\frac{\partial}{\partial\theta}\log f(x_i;\theta) = 0$$

and is consistent. i.e., $\hat{\theta}_n^{MLE}\xrightarrow[n\to\infty]{P_\theta}\theta$

(R7) (Size of Remainder) The third derivative[104] of the log likelihood function is continuous and there exists $M(x_i)$ such that

$$\max_{\theta\in\Omega}\|\partial_\theta^3\log f(x_i;\theta)\| \le M(x_i),\ E[M(x_i)] < +\infty$$

If these conditions are satisfied, MLE has the following limiting distribution

$$\sqrt{n}(\hat{\theta}_n^{MLE} - \theta)\xrightarrow[n\to\infty]{d} N(0, [I(\theta)]^{-1})$$

(proof) The proof for the multivariate case is Similar. Thus, we will show the proof in one-dimension case only.

For conciseness, we will denote the derivative of the mean of log-likelihood function as:

$$\bar{\dot{\ell}}_n(\theta) = \dot{\ell}_n(\theta)/n,\quad \bar{\ddot{\ell}}_n(\theta) = \ddot{\ell}_n(\theta)/n,\quad \overline{\ell_n^{(3)}}(\theta) = \partial_\theta^3\ell_n(\theta)/n$$

$\hat{\theta}^{MLE}$: MLE of θ

We will divide the proof into four large steps.

① [First order Approximation of the likelihood equation and remainder term]

From (R6), (R7) and Taylor series expansion,

$$0 = \bar{\dot{\ell}}_n(\hat{\theta}) = \bar{\dot{\ell}}_n(\theta) + \bar{\ddot{\ell}}_n(\theta)[\hat{\theta}-\theta] + \overline{\ell_n^{(3)}}(\theta_n^*)[\hat{\theta}-\theta]^2 \quad \exists\,\theta_n^*\ s.t.\ |\theta-\theta_n^*| \le |\theta-\hat{\theta}_n|$$

② Convergence in Probability of the Remainder Term.

$$R_n = \frac{1}{2}\overline{\ell_n^{(3)}}(\theta_n^*)(\hat{\theta}_n-\theta)$$

WTS: $P_\theta(|R_n|\ge\epsilon) = 0\ \ \forall\epsilon>0$.

$$P_\theta(|R_n|\ge\epsilon) = P(|R_n|\ge\epsilon, |\overline{\ell_n^{(3)}}(\theta_n^*)|>K) + P(|R_n|\ge\epsilon, |\overline{\ell_n^{(3)}}(\theta_n^*)|\le K)$$

$$\le \underbrace{P(|\overline{\ell_n^{(3)}}(\theta_n^*)|>K)}_{①} + \underbrace{P(|R_n|/|\overline{\ell_n^{(3)}}(\theta_n^*)|\ge\epsilon/K)}_{②}$$

① $P(|\overline{\ell_n^{(3)}}(\theta_n^*)|>K) \le P(\frac{1}{n}\sum_{i=1}^n M(x_i) > K)$

$$\le E_\theta[n^{-1}\sum_{i=1}^n M(x_i)]/K = E_\theta[M(x_i)]/K$$

② $P_\theta(|R_n|)/|\overline{\ell_n^{(3)}}(\theta_n^*)| \ge \epsilon/K) = P_\theta(|\hat{\theta}_n-\theta|\ge 2\epsilon/K)$

$\lim_{n\to\infty}P_\theta(|\hat{\theta}_n-\theta|\ge\epsilon/K) = 0$ (∵ R6)

$$\therefore \lim_{n\to\infty}P(|R_n|>\epsilon) = E_\theta[M(X_i)]/K + \lim_{n\to\infty}P_\theta(|\hat{\theta}_n-\theta|\ge\epsilon/K)$$
$$= E_\theta[M(x_i)]/K \xrightarrow[K\to\infty]{} 0$$

$\therefore \lim_{n\to\infty}P(|R_n|\ge\epsilon) = 0 \quad \therefore R_n = \frac{1}{2}\overline{\ell_n^{(3)}}(\theta^*)(\hat{\theta}_n-\theta)\xrightarrow[n\to\infty]{P_\theta} 0$

---

104) 삼차편도함수는 $\frac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta_k}\ell_n(\theta)$ 을 뜻하고, 고려해야 "max" 수식 $\max_{\theta\in\Omega}\|\partial_\theta^3\log f(x;\theta)\| = \max_{1\le h,i,j\le K}\left|\frac{\partial^3}{\partial\theta_h\partial\theta_i\partial\theta_j}\log f(x_i;\theta)\right|$

③ (score of the MLE)

From steps ① and ②

$$0 = \dot{\ell}_n(\hat\theta) = \dot{\ell}_n(\theta) + \ddot{\ell}_n(\theta)(\hat\theta-\theta) + \tfrac{1}{2}\overline{\ell_n^{[3]}}(\theta^*)(\hat\theta-\theta)^2$$
$$= \dot{\ell}_n(\theta) + \left\{\ddot{\ell}_n(\theta) + \tfrac{1}{2}\overline{\ell_n^{[3]}}(\theta^*)(\hat\theta-\theta)\right\}(\hat\theta-\theta)$$

$$\Rightarrow\ -\dot{\ell}_n(\theta) = \left\{\ddot{\ell}_n(\theta) + \tfrac{1}{2}\overline{\ell_n^{[3]}}(\theta^*)(\hat\theta-\theta)\right\}(\hat\theta-\theta)$$

$$* \quad -\ddot{\ell}_n(\theta) = -\frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\theta^2}\log f(x_i;\theta) \xrightarrow[n\to\infty]{} -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(x_i;\theta)\right] = E_\theta\left[-\frac{\partial^2}{\partial\theta^2}\log f(x_i;\theta)\right] = I(\theta)$$

For large n, $-\ddot{\ell}_n(\theta) \simeq \left\{-I(\theta) + R_n\right\}(\hat\theta-\theta)$

$$\therefore\ \ddot{\ell}_n(\theta)\left\{I(\theta)-R_n\right\}^{-1} \simeq (\hat\theta-\theta)$$
$$\Rightarrow \sqrt{n}\,\ddot{\ell}(\theta)\left\{I(\theta)-R_n\right\}^{-1} \simeq \sqrt{n}(\hat\theta-\theta),\quad R_n \xrightarrow[n\to\infty]{P_\theta} 0$$

④ (Limiting Distribution of MLE)

From ⟨Theorem 6.4.3⟩ and CLT,

$$\sqrt{n}\,\dot{\ell}_n(\theta) = \sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\theta}\log f_X(x) - E_\theta\left[\frac{\partial}{\partial\theta}\log f_X(x)\right]\right\} \sim N\left[0,\ \mathrm{Var}\left[\frac{\partial}{\partial\theta}\log f_X(x)\right]\right]$$

$$* \quad E_\theta\left[\frac{\partial}{\partial\theta}\log f(x;\theta)\right] = 0,\quad \mathrm{Var}\left[\frac{\partial}{\partial\theta}\log f_X(x)\right] = I(\theta) = E_\theta\left[-\frac{\partial^2}{\partial\theta^2}\log f_X(x)\right]$$

$$\therefore \sqrt{n}\,\dot{\ell}_n(\theta) \xrightarrow[n\to\infty]{d} N[0, I(\theta)]$$

$$\Rightarrow \sqrt{n}\,\dot{\ell}_n(\theta)\left\{I(\theta)-R_n\right\}^{-1} \xrightarrow[n\to\infty]{d} I(\theta)^{-1}Z,\quad Z\sim N[0,I(\theta)]$$
$$\xrightarrow[n\to\infty]{d} N\left(0, I(\theta)^{-1}\right) \qquad \square$$

The convergence in probability of the remainder term in ⟨Theorem 6.4.4⟩, (R7), is a stronger condition than what is minimally necessary. This was for the conciseness of the proof and a weaker condition would still work.

Meanwhile, as we have seen in the last step of the proof in ⟨Theorem 6.4.4⟩, under suitable conditions,

$$\sqrt{n}(\hat\theta-\theta) = \left[I(\theta)\right]^{-1}\sqrt{n}\,\dot{\ell}_n(\theta) + o_n,\quad o_n \xrightarrow[n\to\infty]{P_\theta} 0$$

we know that the approximation above holds. Meaning, the error of estimation in MLE depends on $\sqrt{n}\,\dot{\ell}_n(\theta)$ and its precision improves as $I(\theta)$ increases. In this respect, we call $\sqrt{n}\,\dot{\ell}_n(\theta)$ the "score function" as it shows the estimation score of MLE and $I(\theta)$ the "information number(정보량)" as it shows the estimation precision of MLE.

As we have seen in ⟨Ex.6.4.2⟩ and ⟨6.4.3⟩, if we do not know the specific form(구체적인 형태) of the MLE but know that it is the root of a likelihood equation, we approximate the MLE using the "one-step iteration". We can show that the following theorem holds for the estimator calculated by the one-step iteration algorithm using the method in ⟨Theorem 6.4.4⟩.

⟨Theorem 6.4.5. Approximation of MLE using one-step iteration⟩

In a probability model that has the pdf $f(x;\theta), \theta\in\Omega$, suppose that the conditions (R0)-(R7) are satisfied, and suppose that there exists an estimator $\hat\theta_n^{(0)}$ satisfying the following condition.

(R8) $\sqrt{n}(\hat\theta_n^{(0)}-\theta)$ has a limiting distribution. Meaning, there exists a random variable Z such that $\sqrt{n}(\hat\theta_n^{(0)}-\theta) \to Z$.

In this case, using the first order approximation(일차근사), we know that $\hat\theta_n^{(1)}$ defined as:

$$\hat\theta_n^{(1)} = \hat\theta_n^{(0)} + \left[-\ddot{\ell}_n(\hat\theta_n^{(0)})\right]^{-1}\dot{\ell}_n(\hat\theta_n^{(0)})$$

has the limiting distribution:

$$\sqrt{n}(\hat\theta_n^{(1)}-\theta) \xrightarrow[n\to\infty]{d} N\left[0, \left[I(\theta)\right]^{-1}\right)$$

Ex. 6.4.4. Logistic Distribution Model $L(\theta,1)$, $-\infty<\theta<+\infty$

$$f(x;\theta) = \frac{e^{x-\theta}}{(1+e^{x-\theta})^2} = \frac{e^{-x+\theta}}{(1+e^{-x+\theta})^2},\quad -\infty<x<+\infty$$

$$\log f(x;\theta) = (x-\theta) - 2\log(1+e^{x-\theta}) = -(x-\theta) - 2\log(1+e^{-x+\theta})$$

$$\dot{\ell}_n(\theta;x) = n(\bar x-\theta) - 2\sum_{i=1}^n \log(1+e^{x-\theta}) = -n(\bar x-\theta) - 2\sum_{i=1}^n \log(1+e^{-x_i+\theta})$$

$$\dot{\ell}_n(\theta;x) = -n + 2\sum_{i=1}^n e^{x_i-\theta}/[1+e^{x_i-\theta}] = n - 2\sum_{i=1}^n e^{-x_i+\theta}/(1+e^{-x_i+\theta})$$

$$\ddot{\ell}_n(\theta;x) = -2\sum_{i=1}^n e^{x_i-\theta}/[1+e^{x_i-\theta}] + 2\sum_{i=1}^n e^{x_i-\theta}/(1+e^{x_i-\theta})^2$$
$$= -2\sum_{i=1}^n e^{x_i-\theta}/(1+e^{x_i-\theta})^2$$

$\hat\theta_n^{MLE}$ is the unique root of the likelihood equation $\dot{\ell}(\theta;x)=0$.

Let $\hat\theta_n^{(0)} = \bar x = \frac{1}{n}\sum_1^n x_i$

$$\hat\theta_n^{(1)} = \hat\theta_n^{(0)} + \left[-\ddot{\ell}_n(\hat\theta_n^{(0)})\right]^{-1}\dot{\ell}_n(\hat\theta_n^{(0)})$$

Meanwhile, we have proved in ⟨Ex. 5.4.3⟩ that the mean and variance of $L(\theta,1)$ is θ and $\pi^2/3$ respectively. Hence, by the CLT, the limiting distribution of $\hat\theta_n^{(0)}$ is:

$$\sqrt{n}(\hat\theta_n^{(0)}-\theta) \xrightarrow[n\to\infty]{d} N(0, \pi^2/3)$$

Also from ⟨Ex. 5.4.3⟩, $X_i-\theta \overset{d}{=} \log\frac{U}{1-U}$, $\frac{1}{1+e^{-(x_i-\theta)}} \overset{d}{=} U$, $U\sim U(0,1)$

Hence, the information number is:

$$I(\theta) = E_\theta\left[-\ddot{\ell}_n(\theta)\right] = 2E[U(1-U)] = 1/3$$

In this case, it is known that the other conditions in ⟨Theorem 6.4.5⟩ are satisfied. Hence the estimator $\hat\theta_n^{(1)}$ has the following limiting distribution:

$$\sqrt{n}(\hat\theta_n^{(1)}-\theta) \xrightarrow[n\to\infty]{d} N(0,3)$$

Ex. 6.4.5. Gamma Distribution Model Gamma(α, β), α>0, β>0

The pdf of Gamma(α,β) is: $\frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} I_{(x>0)}$, $\theta=(\alpha,\beta)^t$

$\log f_X(x) = -\log\Gamma(\alpha) - \alpha\log\beta + (\alpha-1)\log x - \frac{x}{\beta}$

$\ell(\theta) = -n\log\Gamma(\alpha) - n\alpha\log\beta + (\alpha-1)\sum_1^n \log x_i - \frac{1}{\beta}\sum_{i=1}^n x_i$

$\dot\ell(\theta) = \left[\frac{\partial}{\partial\alpha}\ell(\theta), \frac{\partial}{\partial\beta}\ell(\theta)\right]^t$

1) $\frac{\partial}{\partial\alpha}\ell(\theta) = -n\Psi(\alpha) - n\log\beta + \sum_1^n \log x_i$

2) $\frac{\partial}{\partial\beta}\ell(\theta) = -n\alpha\beta^{-1} + \beta^{-2}\sum_1^n x_i$

$\Psi = \frac{\partial^2}{\partial\alpha^2}\log(\Gamma(\alpha))$

$\therefore \dot\ell(\theta) = \begin{bmatrix} -n\Psi(\alpha) - n\log\beta + \sum_1^n \log x_i \\ -n\alpha\beta^{-1} + \beta^{-2}\sum_1^n x_i \end{bmatrix}$, $\ddot\ell(\theta) = \begin{bmatrix} -n\dot\Psi(\alpha) & -n\beta^{-1} \\ -n\beta^{-1} & n\alpha\beta^{-2} - 2n\bar x\cdot\beta^{-3} \end{bmatrix}$

For simplicity, we will use the following abbreviations.

$(\overline{\log x}) = \frac{1}{n}\sum_{i=1}^n \log(x_i)$, $\Psi(\alpha) = \frac{\partial}{\partial\alpha}\log\Gamma(\alpha)$,

$\partial_\alpha = \frac{\partial}{\partial\alpha}$, $\partial_\beta = \frac{\partial}{\partial\beta}$, $\partial_{\alpha\alpha} = \frac{\partial^2}{\partial\alpha^2}$, $\partial_{\alpha\beta} = \frac{\partial^2}{\partial\alpha\partial\beta}$, $\partial_{\beta\beta} = \frac{\partial^2}{\partial\beta^2}$

The pdf of Gamma distributions is in the exponential family and satisfies the conditions in ⟨Theorem 6.4.4.⟩.

As we have shown in ⟨Ex.6.4.3⟩, the MLE of θ is given as the unique root of the likelihood equation $\dot\ell(\theta)=0$ and we can find its approximation using the one-step iteration algorithm and MME $=\hat\theta_n^{(0)}$ as the following:

$$\hat\theta_n^{(1)} = \hat\theta_n^{(0)} + \left[-\ddot\ell_n(\hat\theta_n^{(0)})\right]^{-1}\dot\ell_n(\hat\theta_n^{(0)}), \quad \hat\theta_n^{(0)} = \hat\theta^{MME}$$

Meanwhile, the mean and variance of the Gamma distribution is $\alpha\beta, \alpha\beta^2$ respectively. The MME of $\theta=(\alpha,\beta)^t$, $\hat\theta^{MME} = (\hat\alpha^{MME}, \hat\beta^{MME})^t$ is given as the root of the following simultaneous equation, and from ⟨Theorem 6.1.2⟩, this estimator satisfies Condition (RS) in ⟨Theorem 6.4.5⟩.

$$\begin{cases} \hat\alpha\hat\beta = \frac{1}{n}\sum_1^n X_i = \bar X \\ \hat\alpha\hat\beta^2 = n^{-1}\sum_{i=1}^n (X_i-\bar X)^2 \end{cases}$$

Meanwhile, the information number matrix and its inverse matrix is:

$I(\theta) = E_\theta[-\ddot\ell_1(\theta)] = \begin{pmatrix} \dot\Psi(\alpha) & 1/\beta \\ 1/\beta & \alpha/\beta^2 \end{pmatrix}$, $[I(\theta)]^{-1} = \frac{1}{\alpha\dot\Psi(\alpha)-1}\begin{pmatrix} \alpha & -\beta \\ -\beta & \dot\Psi(\alpha)\beta^2 \end{pmatrix}$

and for $\hat\theta_n^{(1)} = \hat\theta_n^{(0)} + \left[-\ddot\ell_n(\hat\theta_n^{(0)})\right]^{-1}\cdot\dot\ell_n(\hat\theta_n^{(0)})$

$$\sqrt{n}(\hat\theta_n^{(1)} - \theta) \xrightarrow[n\to\infty]{d} N(0, [I(\theta)]^{-1})$$

6.5. LEAST SQUARES ESTIMATION

We will learn about the "least squares estimation" which is used in the estimation of parameters in a linear regression model, often used in the investigation of functional relationship(함수관계) between random variables.

Definition: Linear Regression Model.

$$\begin{cases} Y_i = x_{i0}\beta_0 + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p + e_i \\ E[e_i]=0,\ Var[e_i]=\sigma^2,\ i=1,2,\cdots,n,\ Cov(e_i,e_j)=0\ (i\neq j) \\ -\infty<\beta_j<+\infty,\ j=0,1,\cdots,p,\ 0<\sigma^2<+\infty \end{cases}$$

In such linear regression models, the objective is to estimate the mean response under the assumption that the mean of the response variable Y_i is a linear function of the explanatory variables. Accordingly, the method of least squares proposes to use, among the assumed functions, the one that minimizes the sum of squared distances from the observed values as the estimator. * $E[Y] = \beta_0 + \beta_1 X$

In other words, least squared estimation is finding $\hat\beta = (\hat\beta_0, \hat\beta_1, \cdots, \hat\beta_p)^t$ that minimizes:

$$\sum_1^n \left\{ Y_i - (x_{i0}\beta_0 + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p)\right\}^2$$

and estimating $E[\hat Y_i]$ as the mean response when the explanatory variables are $(x_{i0}, x_{i1}, \cdots, x_{ip})$.

LSE in Linear Regression Model

$$\hat\beta^{LSE} = (\hat\beta_0^{LSE}, \cdots, \hat\beta_p^{LSE}) = \underset{\beta\in\Omega}{\operatorname{argmax}} \sum_{i=1}^n \left\{ Y_i - (x_{i0}\beta_0 + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p)\right\}^2$$

$\hat\beta^{LSE}$ is called the LSE (Least Squares Estimator) of $\beta=(\beta_0,\cdots,\beta_p)^t$, and we define the LSE of the mean response $\mu(x_0,x_1,\cdots,x_p) = E(Y|x_0,x_1,\cdots,x_p)$ as the following:

$$\hat\mu^{LSE}(x_0,\cdots,x_p) = x_0\hat\beta_0^{LSE} + x_1\hat\beta_1^{LSL} + \cdots + x_p\hat\beta_p^{LSE}$$

To express the linear regression model with vectors and matrices, let:

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix},\ X = \begin{pmatrix} x_{10} & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{np} \end{pmatrix},\ \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix},\ e = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

Now we can simplify the expression of the linear regression model and the least squares estimator as: 105

Linear Regression Models and LSE using vectors & matrices.

$$\begin{cases} Y = X\beta + e \\ E[e]=0,\ Var(e)=\sigma^2 I_n \\ \beta\in\mathbb{R}^{p+1},\ \sigma^2>0,\ \operatorname{rank}(X)=p+1 \end{cases}$$

$$\hat\beta^{LSE} = \underset{\beta\in\mathbb{R}^{p+1}}{\operatorname{argmin}} \|Y-X\beta\|^2,\quad \|Y-X\hat\beta^{LSE}\|^2 = \min_{\beta\in\mathbb{R}^{p+1}}\|Y-X\beta\|^2$$

105) 여기에서 rank(X)는 행렬 X의 계수(rank)를 뜻하고, $\underset{\beta\in\mathbb{R}^{p+1}}{\operatorname{argmin}} \|Y-X\beta\|^2$ 에서 argmin은 'argument of minimizing'의 약자이다.

〈Theorem 6.5.1. Least Squared Estimation in Linear Regression Model〉

(a) Let $\pi = X(X^tX)^{-1}X^t$ 106)

$\pi^t = \pi$, $\pi^t(I-\pi)=0$, $\pi X = X$

(b) $\|Y-X\beta\|^2 = \|\pi(Y-X\beta)\|^2 + \|(I-\pi)Y\|^2$

$X\hat{\beta}^{LSE}$, $\hat{\beta}^{LSE} = (X^TX)^{-1}X^tY$

(proof)

(a) $\pi^t = [X(X^tX)^{-1}X^t]^t = (X^t)^t[(X^tX)^t]^tX^t = X(X^tX)^{-1}X^t = \pi$

$\pi^t(I-\pi) = \pi^t - \pi^t\pi = \pi - \pi^t\pi = X(X^tX)^{-1}X^t - X(X^tX)^{-1}X^tX(X^tX)^{-1}X^t$
$= X(X^tX)^{-1}X^t - X(X^tX)^{-1}X^t = 0$

$\pi X = X(X^tX)^{-1}X^tX = X$

(b) $\|Y-X\beta\|^2 = \|Y-\pi Y+\pi Y-X\beta\|^2 = \|Y-\pi Y\|^2 + \|\pi Y-X\beta\|^2 + \underbrace{2(Y-\pi Y)^t(\pi Y-X\beta)}_{*}$

* $(Y-\pi Y)^t(\pi Y-X\beta) = Y^t(I-\pi)^t\pi Y - Y^t(I-\pi)^tX\beta$
$= 0 - Y^t(I-\pi)X\beta$
$= 0 - Y^t(X-\pi X)\beta$
$= 0 - 0$ ($\because X=\pi X$)

$\therefore \|Y-X\beta\|^2 = \|Y-\pi Y\|^2 + \|\pi Y-X\beta\|^2$

$\therefore \underset{\beta\in\mathbb{R}^{p+1}}{\operatorname{argmin}} \|Y-X\beta\|^2 = \underset{\beta\in\mathbb{R}^{p+1}}{\operatorname{argmin}} \|Y-\pi Y\|^2 + \|\pi Y-X\beta\|^2 = (X^tX)^{-1}X^tY$ □

In Linear Regression Model, we use the following mean squared error (MSE) as the estimate of error term's variance.

$$\hat{\sigma}^2 = \|Y-X\hat{\beta}^{LSE}\|^2/(n-p-1)$$

〈Theorem 6.5.2. Properties of LSE〉

The following holds for the Linear Regression Model:

$$\begin{cases} Y = X\beta + e \\ E[e]=0,\ Var(e)=\sigma^2 I_n \\ \beta\in\mathbb{R}^{p+1},\ \sigma^2>0,\ rank(X)=p+1 \end{cases}$$

(a) $E[\hat{\beta}^{LSE}] = \beta$, $Var(\hat{\beta}^{LSE}) = \sigma^2(X^TX)^{-1}$

(b) $E[\hat{\sigma}^2] = \sigma^2$

(c) If the distribution of error terms are normal, i.e., $e_i \sim N_n(0,\sigma^2 I)$

$\hat{\beta}^{LSE} \sim N(\beta, \sigma^2(X^tX)^{-1})$, $(n-p-1)\hat{\sigma}^2/\sigma^2 \sim \chi^2(n-p-1)$
and $\hat{\beta}^{LSE}$ and $\hat{\sigma}^2$ are independent.

(proof)

(a) $E[\hat{\beta}^{LSE}] = E[(X^tX)^{-1}X^tY] = (X^tX)^{-1}X^tX\beta = \beta$

$Var(\hat{\beta}^{LSE}) = Var[(X^tX)^{-1}X^tY] = (X^tX)^{-1}X^t\,Var(Y)\,X(X^tX)^{-1}$
$= (X^tX)^{-1}X^t\cdot\sigma^2 I_n\cdot X(X^tX)^{-1}$
$= \sigma^2(X^tX)^{-1}X^tX(X^tX)^{-1}$
$= \sigma^2(X^tX)^{-1}$

(b) $\hat{\sigma}^2 = \|Y-X\hat{\beta}^{LSE}\|^2/(n-p-1) = \|Y-\pi Y\|^2/(n-p-1)$
$= \|(I-\pi)Y\|^2/(n-p-1)$
$= Y^t(I-\pi)^t(I-\pi)Y/(n-p-1)$
$= Y^t(I-\pi)Y/(n-p-1)$

$(I-\pi)X\beta = 0$ $\therefore$ $(I-\pi)Y = (I-\pi)Y - (I-\pi)X\beta$
$= (I-\pi)(Y-X\beta)$
$= (I-\pi)e$

$\therefore \hat{\sigma}^2 = Y^t(I-\pi)^t(I-\pi)Y/(n-p-1)$
$= e^t(I-\pi)^t(I-\pi)e/(n-p-1)$
$= e^t(I-\pi)e/(n-p-1)$

* $e^t(I-\pi)e \in \mathbb{R}$ (상수) $\therefore e^t(I-\pi)e = trace(e^t(I-\pi)e)$
$= trace((I-\pi)ee^t)$

$E[e^t(I-\pi)e] = E[trace(e^t(I-\pi)e)] = E[trace((I-\pi)ee^t)]$

$= trace(E[(I-\pi)ee^t])$
$= trace((I-\pi)E[ee^t])$
$= trace((I-\pi)\{E[ee^t] - E[e]E[e]^t\})$
$= trace((I-\pi)Var(e))$
$= trace((I-\pi)\cdot\sigma^2)$
$= \sigma^2 trace(I-\pi)$
$= \sigma^2[trace(I) - trace(X(X^tX)^{-1}X^t)]$
$= \sigma^2[n - trace((X^tX)^{-1}X^tX)]$
$= \sigma^2[n - trace(I_{p+1})]$
$= \sigma^2(n-p-1)$

$\therefore E[\hat{\sigma}^2] = E[e^t(I-\pi)e]/(n-p-1) = \sigma^2(n-p-1)/(n-p-1) = \sigma^2$

(c) proof is detailed in 〈Theorem 4.4.6〉

If the columns of matrix X, which contains the values of explanatory variables, are orthogonal, X^tX becomes a diagonal matrix. Thus not only calculation but also the interpretation of the relationship between random variables become easier. However, if the columns of the matrix are not orthogonal, we can transform them into an orthogonal matrix so that estimation, such as the least squares estimation becomes easier.

For the explanatory random variable matrix $X \in \mathbb{R}^{n\times(p+1)}$ if $\operatorname{rank}=(p+1)$

$$X=(X_0, X_1),\ X_0 : n\times p_0,\ X_1 : n\times p_1\ (p_0+p_1=p+1)$$

$$\Pi_0 : X_0(X_0^tX_0)^{-1}X_0^t,\ \beta^t=(\beta_0^t, \beta_1^t)$$

we can express the mean response(평균반응) as the following:

$$\begin{aligned} X\beta=(X_0\ X_1)\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix} &= X_0\beta_0+X_1\beta_1 \\ &= X_0\beta_0+X_0(X_0^tX_0)^{-1}X_0^tX_1\beta_1-X_0(X_0^tX_0)^{-1}X_0^tX_1\beta_1+X_1\beta_1 \\ &= X_0(\beta_0+(X_0^tX_0)^{-1}X_0^tX_1\beta_1)+(X_1\beta_1-\Pi_0X_1\beta_1) \\ &= X_0(\beta_0+(X_0^tX_0)^{-1}X_0^tX_1\beta_1)+(I-\Pi_0)X_1\beta_1 \end{aligned}$$

Let $X_{1|0}=(I-\Pi_0)X_1$. Then, the columns of $X_{1|0}$ is orthogonal to the columns of X_0 and the mean response can be expressed as:

$$X\beta=X_0\beta_0+X_1\beta_1=X_0\gamma_0+X_{1|0}\beta_1,\ \gamma_0=(\beta_0+(X_0^tX_0)^{-1}X_0^tX_1\beta_1)$$

and insted of $X=(X_0,X_1)$, we can use $(X_0,X_{1|0})$ to express the Linear Regression Model. Using the orthogonal column vectors of $(X_0,X_{1|0})$,

for $\Pi_{0,1}=X(X^TX)^{-1}X^t$, $\Pi_0=X_0(X_0^tX_0)^{-1}X_0^t$, $\Pi_{1|0}=X_{1|0}(X_{1|0}^tX_{1|0})^{-1}X_{1|0}^t$,

the following holds:

$$\Pi_{0,1}=\Pi_0+\Pi_{1|0},\ \Pi_0^t\Pi_{1|0}=0$$

Meaning, we can decompose $\Pi_{0,1}$ as the sum of two orthogonal projection matrices. Utilizing this property, we can prove the following theorem:

〈Theorem 6.5.3. Orthogonalization of Explanatory Random Variables and Least Squares Estimation〉

In the Linear Regression Model from 〈Theorem 6.5.2〉, let:

$$X=(X_0, X_1),\ X_0 : n\times p_0,\ X_1 : n\times p_1\ (p_0+p_1=p+1)$$

$$\Pi_0=X_0(X_0^tX_0)^{-1}X_0^t,\ X_{1|0}=(I-\Pi_0)X_1,\ \Pi_{1|0}=X_{1|0}(X_{1|0}^tX_{1|0})^{-1}X_{1|0}^t$$

Then the following holds:

(a) $\hat{X\beta}^{LSE}=X_0\hat{\beta}_0^{LSE}+X_1\hat{\beta}_1=X_0\hat{\gamma}_0^{LSE}+X_{1|0}\hat{\beta}_1^{LSE}$

$$X_0\hat{\gamma}_0^{LSE}=\Pi_0Y,\ X_{1|0}\hat{\beta}_1^{LSE}=\Pi_{1|0}Y$$

$$\hat{\beta}_0^{LSE}=\hat{\gamma}_0^{LSE}-(X_0^tX_0)^{-1}X_0^tX_1\hat{\beta}_1^{LSE}$$

(b) $$\operatorname{Cov}(\hat{\gamma}_0^{LSE}, \hat{\beta}_1^{LSE})=0$$

$$E[\hat{\gamma}_0^{LSE}]=\gamma_0,\ \operatorname{Var}(\hat{\gamma}_0^{LSE})=\sigma^2(X_0^tX_0)^{-1}$$

$$E[\hat{\beta}_1^{LSE}]=\beta_1,\ \operatorname{Var}(\hat{\beta}_1^{LSE})=\sigma^2(X_{1|0}^tX_{1|0})^{-1}$$

Ex. 6.5.1. Linear Regression Model that includes the intercept (경회귀)

Suppose that the mean response in a linear regression model is:

$$E[Y_i|\ x_{i0}, x_{i1}, \cdots, x_{ip}]=\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip}\ (i=1,2,\cdots,n)$$

Let $\overline{x_{\cdot j}}=(x_{1j}+\cdots+x_{nj})/n\ (j=1,\cdots,p)$. Then the mean response can be expressed as:

$$\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip}=\gamma_0+\beta_1(x_{i1}-\overline{x_{\cdot 1}})+\cdots+\beta_p(x_{ip}-\overline{x_{\cdot p}})$$

In this case,

$$X_0=(1,\cdots,1)^t,\ X_{1|0}=(x_{i1}-\overline{x_{\cdot 1}},\cdots,x_{ip}-\overline{x_{\cdot p}})\ 1\le i\le p$$

Thus, if we let:

$$S_{jK}=\sum_{i=1}^{n}(x_{ij}-\overline{x_{\cdot j}})(x_{iK}-\overline{x_{\cdot K}}),\ S_{jY}=\sum_{i=1}^{n}(x_{ij}-\overline{x_{\cdot j}})(Y_i-\overline{Y})$$

then we know that:

$$X_{1|0}^tX_{1|0}=(S_{jK}),\ X_{1|0}^tY=(S_{jY})$$

Hence the LSE is given as:

$$\hat{\gamma}_0=\overline{Y},\ \begin{bmatrix}S_{11}&\cdots&S_{1p}\\ \vdots&&\vdots\\ S_{p1}&\cdots&S_{pp}\end{bmatrix}\begin{bmatrix}\hat{\beta}_1\\ \vdots\\ \hat{\beta}_p\end{bmatrix}=\begin{bmatrix}S_{1Y}\\ \vdots\\ S_{pY}\end{bmatrix}$$

□