최대가능도추정법과 선형모형

서울시립대 통계학과 이용희

FALL 2019

1 서론

1.1 가능도함수와 그 성질

확률변수 Y 가 확률밀도함수 $f(y;\boldsymbol{\theta})$를 따른다고 하자. 모수 $\boldsymbol{\theta}$ 에 대한 가능도함수(likelihood function) $L(\boldsymbol{\theta})$와 로그가능도함수 $\ell(\boldsymbol{\theta})$는 다음과 같이 정의한다.

$$L(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta};y) = f(y;\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(Y=y), \quad \ell(\boldsymbol{\theta}) \equiv \ell(\boldsymbol{\theta};y) = \log L(\boldsymbol{\theta};y)$$

위에서 가능도함수를 $L(\boldsymbol{\theta};y)$로 표시한 이유는 가능도 함수는 모수 $\boldsymbol{\theta}$ 의 함수이며, 이는 확률변수의 관측값 y 가 있는 경우 얻을 수 있다는 것을 강조하기 위해서이다.

로그가능도함수를 모수 $\boldsymbol{\theta}$로 한 번 미분한 도함수(greadient)를 스코어함수(score function) $s(\boldsymbol{\theta})$ 로 아래와 같이 정의한다. 또한 두 번 미분한 헤시안(hessian)의 음수를 관측피셔정보(observed Fisher information) $\boldsymbol{J}(\boldsymbol{\theta})$ 라고 정의한다.

$$s(\boldsymbol{\theta}) \equiv s(\boldsymbol{\theta};y) = \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta};y), \quad \boldsymbol{J}(\boldsymbol{\theta}) \equiv \boldsymbol{J}(\boldsymbol{\theta};y) = -\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \ell(\boldsymbol{\theta};y)$$

위의 식에서 만약 모수벡터 $\boldsymbol{\theta}$의 차원이 p라면 $s(\boldsymbol{\theta})$는 $p \times 1$ 벡터이고 $\boldsymbol{J}(\boldsymbol{\theta};y)$는 $p \times p$행렬이다.

로그가능도함수는 다음의 두 가지 중요한 방정식을 만족한다.

$$E\left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta};y) \right\} = 0 \tag{1}$$

$$E\left\{ \left[\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta};y) \right] \left[\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta};y) \right]^t \right\} + E\left\{ \frac{\partial^2}{\partial \boldsymbol{\theta}^2} \ell(\boldsymbol{\theta};y) \right\} = 0 \tag{2}$$

식 (1)와 식 (2) 으로부터 다음과 같은 식이 유도되며

$$E[s(\boldsymbol{\theta};y)] = 0, \quad E[s(\boldsymbol{\theta};y)s^t(\boldsymbol{\theta};y)] = E[\boldsymbol{J}(\boldsymbol{\theta};y)]$$

다음과 같은 공식이 주어진다.

$$
\begin{aligned}
Var[s(\boldsymbol{\theta}; y)] &= E[s(\boldsymbol{\theta}; y)s^t(\boldsymbol{\theta}; y)] - \{E[s(\boldsymbol{\theta}; y)]E[s(\boldsymbol{\theta}; y)]^t\} \\
&= E[s(\boldsymbol{\theta}; y)s^t(\boldsymbol{\theta}; y)] - 0 \\
&= -E\left[\frac{\partial^2}{\partial \boldsymbol{\theta}^2}\log f(y; \boldsymbol{\theta})\right] \\
&= E[\boldsymbol{J}(\boldsymbol{\theta}; y)] \\
&\equiv \boldsymbol{I}(\boldsymbol{\theta})
\end{aligned}
$$

위의 식에서 스코어함수의 분산을 피셔정보(Fisher information)이라고 부르며 $\boldsymbol{I}(\boldsymbol{\theta})$로 표기한다. 여기서 주의할 점은 관측피셔정보 $\boldsymbol{J}(\boldsymbol{\theta}; y)$ 는 관측값 y가 있어야 계산이 되는 함수이지만 피셔정보 $\boldsymbol{I}(\boldsymbol{\theta})$는 $\boldsymbol{J}(\boldsymbol{\theta}; y)$의 기대값이기 때문에 모수 $\boldsymbol{\theta}$만의 함수이다.

첫 번째 방정식 (1)는 다음과 같이 적분과 미분의 교환에 의해 증명할 수 있다.

$$
\begin{aligned}
0 &= \frac{\partial}{\partial \boldsymbol{\theta}} \int f(y; \boldsymbol{\theta})\ dy \\
&= \int \frac{\partial}{\partial \boldsymbol{\theta}} f(y; \boldsymbol{\theta})\ dy \\
&= \int \frac{\frac{\partial}{\partial \boldsymbol{\theta}} f(y; \boldsymbol{\theta})}{f(y; \boldsymbol{\theta})} f(y; \boldsymbol{\theta})\ dy \\
&= \int \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}; y) f(y; \boldsymbol{\theta})\ dy \\
&= E\left\{\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}; y)\right\}
\end{aligned}
$$

두 번째 방정식 (2)는 아래와 같이 증명할 수 있다.

$$
\begin{aligned}
0 &= \frac{\partial}{\partial \boldsymbol{\theta}} \int \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}; y) f(y; \boldsymbol{\theta})\ dy \\
&= \int \frac{\partial}{\partial \boldsymbol{\theta}} \left\{ f(y; \boldsymbol{\theta}) \left[\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}; y)\right]^t \right\}\ dy \\
&= \int \left\{ \left[\frac{\partial}{\partial \boldsymbol{\theta}} f(y; \boldsymbol{\theta})\right] \left[\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}; y)\right]^t + f(y; \boldsymbol{\theta}) \frac{\partial}{\partial \boldsymbol{\theta}} \left[\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}; y)\right]^t \right\}\ dy \\
&= \int \left\{ \left[\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}; y)\right] \left[\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}; y)\right]^t f(y; \boldsymbol{\theta}) + \left[\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \ell(\boldsymbol{\theta}; y)\right] f(y; \boldsymbol{\theta}) \right\}\ dy \\
&= E\left\{ \left[\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}; y)\right] \left[\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}; y)\right]^t \right\} + E\left\{\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \ell(\boldsymbol{\theta}; y)\right\}
\end{aligned}
$$

1.2 독립표본

표본 $Y_1, Y_2, \ldots, Y_n$ 가 분포 $f(y_i; \boldsymbol{\theta})$에서 독립적으로 얻어졌고 그 관측값이 각각 $y_1, y_2, \ldots, y_n$이라고 하면 표본으로 부터 계산된 가능도함수 $L_n(\boldsymbol{\theta})$ 은 다음과 같다.

$$L_n(\boldsymbol{\theta}) = L_n(\boldsymbol{\theta}; \boldsymbol{y}) = P_{\boldsymbol{\theta}}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n) = \prod_{i=1}^{n} f(y_i; \boldsymbol{\theta}) = \prod_{i=1}^{n} L(\boldsymbol{\theta}; y_i)$$

또한 표본에 대한 로그가능도함수 ℓ_n 은 다음과 같다.

$$\ell_n(\boldsymbol{\theta}) = \ell_n(\boldsymbol{\theta}; \boldsymbol{y}) = \log L_n(\boldsymbol{\theta}) = \log \prod_{i=1}^{n} f(y_i; \boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(y_i; \boldsymbol{\theta}) = \sum_{i=1}^{n} \ell(\boldsymbol{\theta}; y_i) \tag{3}$$

표본에 의한 로그 가능도함수 $\ell_n(\boldsymbol{\theta})$를 미분한 값, 즉 표본에 의한 스코어 함수 $s_n(\boldsymbol{\theta})$는 다음과 같이 정의한다.

$$s_n(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ell_n(\boldsymbol{\theta}; \boldsymbol{y})$$

n개의 표본에 대한 관측피셔정보 $\boldsymbol{J}_n(\boldsymbol{\theta})$와 피셔정보 $\boldsymbol{I}_n(\boldsymbol{\theta})$도 한 개의 확률 변수 경우와 유사하게 다음과 같이 정의된다.

$$\boldsymbol{I}_n(\boldsymbol{\theta}) = E\left[\boldsymbol{J}_n(\boldsymbol{\theta}; \boldsymbol{y})\right] = E\left[-\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^t} \ell_n(\boldsymbol{\theta}; \boldsymbol{y})\right]$$

1.3 최대가능도추정법

모수 $\boldsymbol{\theta}$ 에 대한 최대가능도 추정량(Maximum Likelihood Estimator;MLE) $\hat{\boldsymbol{\theta}}$는 가능도 함수를 최대로 하는 값으로 정의된다.

$$\hat{\boldsymbol{\theta}}_{MLE} = \arg\max_{\boldsymbol{\theta}} L_n(\boldsymbol{\theta})$$

많은 경우 가능도 함수를 최대화하는 값을 구하기 어려우므로 가능도 함수의 로그 함수, 즉 로그가능도함수를 최대로 하는 값으로 최대가능도 추정량을 구한다.

$$\hat{\boldsymbol{\theta}}_{MLE} = \arg\max_{\boldsymbol{\theta}} \ell_n(\boldsymbol{\theta})$$

만약 로그가능도 함수가 모수 $\boldsymbol{\theta}$에 대하며 미분가능한 함수이면 최대가능도 추정량은 다음과 같은 방정식에 의하여 구할 수 있다.

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ell_n(\boldsymbol{\theta}; \boldsymbol{y}) = s_n(\boldsymbol{\theta}) = \boldsymbol{0}$$

최대가능도 추정량은 적당한 조건하에서 다음과 같은 점근적 성질(Asymptotical properties)을 가진다.

- $\hat{\boldsymbol{\theta}}_{MLE}$는 모수의 참값 $\boldsymbol{\theta}_0$로 확률적 수렴한다.

$$\hat{\boldsymbol{\theta}}_{MLE} \to_p \boldsymbol{\theta}_0 \quad \text{as } n \to \infty$$

- 최대가능도추정량 $\hat{\boldsymbol{\theta}}_{MLE}$는 점근적으로 정규분포를 따른다.

$$\hat{\boldsymbol{\theta}}_{MLE} \sim_d N(\boldsymbol{\theta}_0, \boldsymbol{I}_n^{-1}(\boldsymbol{\theta}_0))$$

2 선형모형

반응변수가 y이고 $p-1$개의 독립변수 $(x_1, x_2, \cdots, x_{p-1})$가 있다고 가정하고 표본의 크기 n인 자료가 얻어지면 선형회귀식을 행렬로 다음과 같이 표현할 수 있다.

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{i,p-1} + e_i \\ &= \boldsymbol{x}_i^t \boldsymbol{\beta} + e_i, \quad i = 1, 2, \cdots, n \end{aligned}$$

위의 식을 다시 표현하면 다음과 같이 쓸 수 있다. 이와 같은 회귀모형을 선형중회귀모형이라 부르며. 각 개체에 대한 모형의 방정식을 하나의 식으로 표현하면 다음과 같다.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ 1 & x_{21} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{n,p-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

위의 선형모형(linear model)을 벡터와 행렬을 이용하여 표시하면 다음과 같다.

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \tag{4}$$

여기서 $\boldsymbol{y}$ 는 $n \times 1$ 반응변수 벡터, $\boldsymbol{X}$ 는 p 개의 독립변수로 이루어진 $n \times p$ 계획행렬(design matrix)이다. 모수 벡터 $\boldsymbol{\beta}$ 는 $p \times 1$ 벡터로 각 독립변수에 대한 회귀계수 벡터이다. $\boldsymbol{e}$는 $n \times 1$ 오차벡터이다.

여기서 회귀분석의 오차항의 가정을 살펴보면 오차항이 서로 독립이고 동일한 분산을 갖는다. 즉, 오차항은 다음의 분포를 따른다. 즉, $\boldsymbol{e} \sim (0, \sigma^2 \boldsymbol{I})$. 관측값 벡터 $\boldsymbol{y}$의 평균과 분산을 보면

$$E(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}, \quad Var(\boldsymbol{y}|\boldsymbol{X}) = \sigma^2 \boldsymbol{I}$$

여기서 오차항이 정규분포를 따른다면 ($\boldsymbol{e} \sim N(0, \sigma^2 \boldsymbol{I})$ 관측값 벡터 $\boldsymbol{y}$ 또한 정규분포를 따른다

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$$

또한 X 가 완전계수(full rank) 행렬이라고 가정하자.

$p+1$개의 모수를 모아놓은 모수벡터는 $\boldsymbol{\theta} = (\boldsymbol{\beta}^t, \sigma^2)^t$이다. 여기서 편의상 오차항의 분산을 $\tau = \sigma^2$ 로 표시하하고자 한다, 즉 $\boldsymbol{\theta} = (\boldsymbol{\beta}^t, \tau)^t$

2.1 최소제곱 추정

위의 선형 모형 가정하에서, 최소제곱 추정량 $\hat{\boldsymbol{\beta}}$ (least square estimator)는 다음과 같이 오차제곱합(Error Sum of Sqaures) SSE 를 최소로 하는 추정량이다.

$$\begin{aligned} SSE(\boldsymbol{\beta}) &= \sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^t\boldsymbol{\beta})^2 \\ &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^t(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ \hat{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta}} SSE(\boldsymbol{\beta}) \end{aligned}$$

따라서 $\hat{\boldsymbol{\beta}}$는 오차제곱합을 최소로 하는 계수 벡터이며 최소제곱 추정량은 다음과 같이 주어진다.

$$\hat{\boldsymbol{\beta}}_{LS} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}$$

오차제곱합에 최소제곱 추정량을 사용하면 이를 잔차제곱합(Residual Sum of Squares)라고 하며 이를 $SSE(\hat{\boldsymbol{\beta}})$로 표시한다.

$$SSE(\hat{\boldsymbol{\beta}}) = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^t(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$$

오차항의 분산에 대한 추정은 정규분포 가정을 오차항에 대한 정규분포를 가정하고 다음과 같은 잔차제곱합의 분포에 대한 결과를 이용하면

$$SSE(\hat{\boldsymbol{\beta}}) \sim \sigma^2\chi^2(n-p)$$

오차항의 분산 σ^2의 불편추정량 S^2을 구할 수 있다.

$$S^2 = \frac{SSE(\hat{\boldsymbol{\beta}})}{(n-p)} = \frac{\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^t\hat{\boldsymbol{\beta}})^2}{(n-p)}$$

즉,

$$E(S^2) = \sigma^2$$

여기서 자유도 $n-p$은 자료의 개수 n에서 절편을 포함한 회귀계수의 개수 p를 뺀 수이다.

2.2 가능도함수

선형모형 (4)에 대한 가능도 함수는 다음과 같이 주어진다.

$$\begin{aligned} L_n(\boldsymbol{\theta}; \boldsymbol{y}) &= L(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) \\ &= \prod_{i=1}^{n} f(y_i) \\ &= \prod_{i=1}^{n}(2\pi\sigma^2)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2}(y_i - \boldsymbol{x}_i^t\boldsymbol{\beta})^2\right] \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^t(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right] \end{aligned}$$

또한 분산에 대한 모수를 $\tau = \sigma^2$ 과 같이 쓰면 로그 가능도함수는 다음과 같다.

$$
\begin{aligned}
\ell_n(\boldsymbol{\theta}; \boldsymbol{y}) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{(\boldsymbol{y}-\boldsymbol{X\beta})^t(\boldsymbol{y}-\boldsymbol{X\beta})}{2\sigma^2} \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\tau - \frac{(\boldsymbol{y}-\boldsymbol{X\beta})^t(\boldsymbol{y}-\boldsymbol{X\beta})}{2\tau}
\end{aligned}
$$

이제 로그가능도함수로부터 구할 수 있는 스코어함수 $s(\boldsymbol{\theta}; y)$ 와 그에 대한 관측 피셔정보 $\boldsymbol{J}_n(\boldsymbol{\theta}; \boldsymbol{y})$ 은 다음과 같이 주어진다.

$$
\begin{aligned}
s(\boldsymbol{\theta}; \boldsymbol{y}) &= \frac{\partial}{\partial\boldsymbol{\theta}}\ell_n(\boldsymbol{\theta}; \boldsymbol{y}) \\
&= \begin{bmatrix} \frac{\partial}{\partial\boldsymbol{\beta}}\ell_n(\boldsymbol{\theta}; \boldsymbol{y}) \\ \frac{\partial}{\partial\tau}\ell_n(\boldsymbol{\theta}; \boldsymbol{y}) \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{X}^t(\boldsymbol{y}-\boldsymbol{X\beta})/\tau \\ -\frac{n}{2\tau} + \frac{(\boldsymbol{y}-\boldsymbol{X\beta})^t(\boldsymbol{y}-\boldsymbol{X\beta})}{2\tau^2} \end{bmatrix} \\
\boldsymbol{J}_n(\boldsymbol{\theta}; \boldsymbol{y}) &= -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^t}\ell_n(\boldsymbol{\theta}; y) \\
&= -\begin{bmatrix} \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^t}\ell_n(\boldsymbol{\theta}; y) & \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\tau^t}\ell_n(\boldsymbol{\theta}; y) \\ \frac{\partial^2}{\partial\tau\partial\boldsymbol{\beta}^t}\ell_n(\boldsymbol{\theta}; y) & \frac{\partial^2}{\partial\tau\partial\tau^2}\ell_n(\boldsymbol{\theta}; y) \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{X}^t\boldsymbol{X}/\tau & -\boldsymbol{X}^t(\boldsymbol{y}-\boldsymbol{X\beta})/\tau^2 \\ -\boldsymbol{X}^t(\boldsymbol{y}-\boldsymbol{X\beta})/\tau^2 & -\frac{n}{2\tau^2} + \frac{(\boldsymbol{y}-\boldsymbol{X\beta})^t(\boldsymbol{y}-\boldsymbol{X\beta})}{\tau^3} \end{bmatrix}
\end{aligned}
$$

2.3 최대가능도 추정량

이제 회귀계수 $\boldsymbol{\beta}$에 대한 최대가능도 추정량은 스코어함수로 부터 얻어진 방정식 $s(\boldsymbol{\theta}; \boldsymbol{y}) = \boldsymbol{0}$ 으로부터 얻어지며 다음과 같은 형태를 가진다.

$$
\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}
$$

$$
\hat{\sigma}^2 = \hat{\tau} = (\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^t(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})/n = \frac{SSE(\hat{\boldsymbol{\beta}})}{n}
$$

여기서 유의할 점은 회귀계수 $\boldsymbol{\beta}$ 의 최대가능도 추정량은 최소제곱법으로 구한 추정량과 동일하다. 따라서 $\hat{\boldsymbol{\beta}}$ 은 최소분산 불편 추정량이다. 하지만 오차항의 분산 σ^2 에 대한 최대가능도 추정량은 불편추정량이 아니다.

$$
E(\hat{\sigma}^2) = E\left[(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^t(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})/n\right] = E\left[\frac{SSE}{n}\right] \neq \sigma^2
$$

참고로 오차항의 분산 σ^2에 대한 불편추정량은 $SSE/(n-p)$이다.

최대가능도 추정량의 점근적 분포를 이용하면 다음과 같이 말할 수 있다. 오차항이 정규분포인 선형모형인 경우 아래의 분포는 점근분포가 아닌 정확한 분포이다.

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \sim N(\mathbf{0}, \boldsymbol{I}_n^{-1}(\boldsymbol{\theta}_0))$$

여기서

$$\boldsymbol{I}_n(\boldsymbol{\theta}) = E[\boldsymbol{J}(\boldsymbol{\theta}; \boldsymbol{y})] = \begin{bmatrix} \boldsymbol{X}^t\boldsymbol{X}/\tau & 0 \\ 0 & \frac{n}{2\tau^2} \end{bmatrix}$$

그리고

$$\boldsymbol{I}_n^{-1}(\boldsymbol{\theta}) = \begin{bmatrix} \tau(\boldsymbol{X}^t\boldsymbol{X})^{-1} & 0 \\ 0 & \frac{2\tau^2}{n} \end{bmatrix} = \begin{bmatrix} \sigma^2(\boldsymbol{X}^t\boldsymbol{X})^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}$$

따라서 회귀계수 $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$의 분포는 평균이 $\mathbf{0}$ 이고 공분산이 $\sigma^2(\boldsymbol{X}^t\boldsymbol{X})^{-1}$ 인 정규분포를 따른다.

여기거 주목할 점은 가능도함수에 최대가능도추정량을 대입하면 그 값이 $SSE(\hat{\boldsymbol{\beta}})$의 함수로 나타난다.

$$\begin{aligned}
L_n(\hat{\boldsymbol{\theta}}) &= L_n(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) \\
&= (2\pi\hat{\sigma}^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\hat{\sigma}^2}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^t(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\right] \\
&= (2\pi\hat{\sigma}^2)^{-\frac{n}{2}} \exp\left[-\frac{n}{2}\right] \\
&= \left(2\pi\frac{SSE(\hat{\boldsymbol{\beta}})}{n}\right)^{-\frac{n}{2}} \exp\left[-\frac{n}{2}\right] \\
l_n(\hat{\boldsymbol{\theta}}) &= l_n(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) \\
&= \text{constant} - \frac{n}{2}\log SSE(\hat{\boldsymbol{\beta}})
\end{aligned}$$

따라서 잔차제곱합 $SSE(\hat{\boldsymbol{\beta}})$ 작아지면 가능도함수는 커진다.