

第1章 線型モデル

1 線型単回帰モデルと最小 2 乗推定量

線型単回帰モデル (simple linear regression model) を考える. $n \in \mathbb{N}$ とし, n 個の観測の組を

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

とする. 各 y_j ($j = 1, 2, \dots, n$) は

$$y_j = \alpha^* + \beta^* x_j + \epsilon_j \quad (j = 1, 2, \dots, n)$$

なる線型構造を持って分布しているとする. ここで, $\alpha^* \in \mathbb{R}$ を y 切片項, $\beta^* \in \mathbb{R}$ を回帰係数と呼ぶ. これらは未知の母数 (パラメータ) と仮定する. y_j を従属変数 (応答変数), x_j を独立変数 (説明変数) と呼ぶ. $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ は独立同一分布に従う確率変数列である.

この節において, 単回帰モデルに以下の仮定 (1) ~ (4) をおくことにする.

- (1) 説明変数 x_1, x_2, \dots, x_n は確率変数ではなく, 与えられた定数.
- (2) $\mathbb{E}[\epsilon_j] = 0$ ($j = 1, 2, \dots, n$).
- (3) $\mathbb{E}[\epsilon_j \epsilon_\ell] = 0$ ($j \neq \ell$).
- (4) $\mathbb{V}[\epsilon_j] = \sigma^2$. ただし, 分散 σ^2 ($\sigma > 0$) は未知とする.

未知の母数 α^*, β^* を推定するために, 最小 2 乗法を用いる. 推定量導出のために,

$$h(\alpha, \beta) := \sum_{j=1}^n \{y_j - (\alpha + \beta x_j)\}^2$$

を考える. α, β に関して h の最小化問題を考える. 最小化問題の解 (存在すれば) を最小 2 乗推定量ということにする.

簡単のために,

$$\begin{aligned}\bar{y} &= \frac{1}{n} \sum_{j=1}^n y_j; & \bar{x} &= \frac{1}{n} \sum_{j=1}^n x_j; \\ Q_{xy} &= \sum_{j=1}^n (x_j - \bar{x})(y_j - \bar{y}); \\ Q_{xx} &= \sum_{j=1}^n (x_j - \bar{x})^2; & Q_{yy} &= \sum_{j=1}^n (y_j - \bar{y})^2\end{aligned}$$

なる記号を導入する. $Q_{xx} \neq 0$ と以下では仮定.

関数 h を変形すれば,

$$\begin{aligned}h(\alpha, \beta) &= Q_{xx} \left\{ \beta - \frac{Q_{xy}}{Q_{xx}} \right\}^2 + n \{ \bar{y} - \alpha - \beta \bar{x} \}^2 \\ &\quad + \frac{Q_{xx} Q_{yy} - Q_{xy}^2}{Q_{xx}}\end{aligned}\tag{1.1}$$

と書ける.

$$\hat{\alpha} := \bar{y} - \hat{\beta} \bar{x}; \quad \hat{\beta} := \frac{Q_{xy}}{Q_{xx}}$$

とおく. 上の式より, $h(\alpha, \beta)$ は $\alpha = \hat{\alpha}$, $\beta = \hat{\beta}$ のときに最小値をとる.

最小 2 乗推定量 (値) を用いて, 回帰直線 $y = \hat{\alpha} + \hat{\beta}x$ を引くことができる. 回帰直線上の点 $(x_j, \hat{\alpha} + \hat{\beta}x_j)$ ($j = 1, 2, \dots, n$) と観測 (x_j, y_j) との差

$$e_j := y_j - (\hat{\alpha} + \hat{\beta}x_j) \quad (j = 1, 2, \dots, n)$$

を残差といい,

$$\text{RSS} := \sum_{j=1}^n e_j^2 = \sum_{j=1}^n \{y_j - (\hat{\alpha} + \hat{\beta}x_j)\}^2$$

を残差平方和という. $n \geq 3$ のとき, 未知の分散 σ^2 を

$$\hat{\sigma}^2 = \frac{1}{n-2} \text{RSS}$$

で推定することができる.

命題 1.1 (1) $\mathbb{E}[\hat{\beta}] = \beta^*$; $\mathbb{V}[\hat{\beta}] = \frac{\sigma^2}{Q_{xx}}$.

(2) $\mathbb{E}[\hat{\alpha}] = \alpha^*$; $\mathbb{V}[\hat{\alpha}] = \sigma^2 \left\{ \frac{1}{n} + \frac{\bar{x}^2}{Q_{xx}} \right\}$.

(3) $\text{COV}[\hat{\alpha}, \hat{\beta}] = -\frac{\bar{x}\sigma^2}{Q}$.

(4) $\mathbb{E}[e_j] = 0$; $\mathbb{V}[e_j] = \sigma^2 (j = 1, 2, \dots, n)$.

証明

証明の方針 $\hat{\alpha} - \alpha^*$, $\hat{\beta} - \beta^*$ を $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ の線型結合で表現できることに注意する. すなわち

$$\hat{\alpha} - \alpha^* = \sum_{j=1}^n a_j \epsilon_j, \quad \hat{\beta} - \beta^* = \sum_{j=1}^n b_j \epsilon_j$$

と表現できることを示す. ただし a_j, b_j ($j = 1, 2, \dots, n$) は定数である. $\mathbb{E}[\epsilon_j] = 0$, $\mathbb{V}[\epsilon_j] = \sigma^2$ と $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ は互いに独立であることに注意して, 期待値と分散の性質を利用すると

$$\begin{aligned} \mathbb{E}\left[\sum_{j=1}^n a_j \epsilon_j\right] &= 0, & \mathbb{V}\left[\sum_{j=1}^n a_j \epsilon_j\right] &= \sum_{j=1}^n a_j^2 \sigma^2, \\ \mathbb{E}\left[\sum_{j=1}^n b_j \epsilon_j\right] &= 0, & \mathbb{V}\left[\sum_{j=1}^n b_j \epsilon_j\right] &= \sum_{j=1}^n b_j^2 \sigma^2, \end{aligned}$$

となることがわかる. あとは $\sum_{j=1}^n a_j^2$, $\sum_{j=1}^n b_j^2$ の計算をすればよい. つぎに e_j ($j = 1, 2, \dots, n$) を

$$e_j = (\epsilon_j - \bar{\epsilon}) - (x_j - \bar{x})(\hat{\beta} - \beta^*)$$

と表現して, (1) – (3) の結果と分散の性質を用いて計算する.

(1) の証明: 誤差項 ϵ_j ($j = 1, 2, \dots, n$) について

$$\bar{\epsilon} := \frac{1}{n} \sum_{j=1}^n \epsilon_j$$

とおく. すると

$$\begin{aligned}y_j - \bar{y} - \beta^*(x_j - \bar{x}) &= \epsilon_j - \bar{\epsilon}, \\ \sum_{j=1}^n (x_j - \bar{x})\bar{\epsilon} &= 0\end{aligned}$$

となる. これより

$$\hat{\beta} - \beta^* = \frac{\sum_{j=1}^n (x_j - \bar{x})\epsilon_j}{Q_{xx}} \quad (1.2)$$

と書ける. (1.2) より

$$\begin{aligned}\mathbb{E}[\hat{\beta} - \beta^*] &= 0, \\ \mathbb{V}[\hat{\beta}] &= \mathbb{E}[(\hat{\beta} - \beta^*)^2] = \frac{\sum_{j=1}^n (x_j - \bar{x})^2 \mathbb{E}[\epsilon_j^2]}{Q_{xx}^2} = \frac{\sigma^2}{Q_{xx}}.\end{aligned} \quad (1.3)$$

(2) の証明: また,

$$\hat{\alpha} - \alpha^* = \bar{\epsilon} - \bar{x}(\hat{\beta} - \beta^*) = \sum_{j=1}^n \left\{ \frac{1}{n} - \frac{(x_j - \bar{x})\bar{x}}{Q_{xx}} \right\} \epsilon_j$$

より

$$\begin{aligned}\mathbb{E}[\hat{\alpha} - \alpha^*] &= 0, \\ \mathbb{V}[\hat{\alpha}] &= \mathbb{E}[(\hat{\alpha} - \alpha^*)^2] = \sum_{j=1}^n \left\{ \frac{1}{n} - \frac{(x_j - \bar{x})\bar{x}}{Q_{xx}} \right\}^2 \mathbb{E}[\epsilon_j^2] = \sigma^2 \left\{ \frac{1}{n} + \frac{\bar{x}^2}{Q_{xx}} \right\}.\end{aligned}$$

(3) の証明: 同様にして,

$$\begin{aligned}\text{COV}[\hat{\alpha}, \hat{\beta}] &= \mathbb{E}[(\hat{\alpha} - \alpha^*)(\hat{\beta} - \beta^*)] \\ &= \frac{1}{Q_{xx}} \sum_{j=1}^n (x_j - \bar{x}) \left\{ \frac{1}{n} - \frac{(x_j - \bar{x})\bar{x}}{Q_{xx}} \right\} \mathbb{E}[\epsilon_j^2] = -\frac{\bar{x}\sigma^2}{Q_{xx}}.\end{aligned}$$

(4) の証明:

$$e_j = y_j - \bar{y} - \hat{\beta}(x_j - \bar{x}) = (\epsilon_j - \bar{\epsilon}) - (x_j - \bar{x})(\hat{\beta} - \beta^*)$$

と表さるので, $\mathbb{E}[e_j] = 0$. また,

$$\begin{aligned} \sum_{j=1}^n \mathbb{E}[e_j^2] &= \sum_{j=1}^n \mathbb{E}[(\epsilon_j - \bar{\epsilon})^2] - 2 \sum_{j=1}^n (x_j - \bar{x}) \mathbb{E}[(\epsilon_j - \bar{\epsilon})(\hat{\beta} - \beta^*)] \\ &\quad + \sum_{j=1}^n (x_j - \bar{x})^2 \mathbb{E}[(\hat{\beta} - \beta^*)^2] \end{aligned} \quad (1.4)$$

と書ける. (1.2) をから,

$$\begin{aligned} \sum_{j=1}^n \mathbb{E}[(\epsilon_j - \bar{\epsilon})^2] &= (n-1)\sigma^2, \\ \sum_{j=1}^n (x_j - \bar{x}) \mathbb{E}[(\epsilon_j - \bar{\epsilon})(\hat{\beta} - \beta^*)] &= \sum_{j=1}^n (x_j - \bar{x}) \mathbb{E}[\epsilon_j(\hat{\beta} - \beta^*)] \\ &= \sum_{j=1}^n (x_j - \bar{x}) \mathbb{E} \left[\frac{\epsilon_j \sum_{\ell=1}^n (x_\ell - \bar{x}) \epsilon_\ell}{Q_{xx}} \right] \\ &= \sigma^2 \end{aligned}$$

となる. これらの 2 つの式と (1.3) を (1.4) に代入すれば,

$$\sum_{j=1}^n \mathbb{E}[e_j^2] = (n-2)\sigma^2$$

を得る. □

命題 1.2 正規性を仮定する. $n \geq 3$ のとき,

- (1) $\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \alpha^* \\ \beta^* \end{pmatrix}, \frac{\sigma^2}{Q_{xx}} \begin{pmatrix} \frac{Q_{xx}}{n} + \bar{x}^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \right),$
- (2) $\frac{(n-2)\hat{\sigma}^2}{\sigma^2} = \frac{\text{RSS}}{\sigma^2} \sim \chi_{n-2}^2,$
- (3) $\hat{\sigma}^2$ と $(\hat{\alpha}, \hat{\beta})$ は独立.

証明 (1), (2) は命題 1.1 と正規性の仮定より直ちにわかる. $\text{COV}[e_j, \hat{\alpha}] = \text{COV}[e_j, \hat{\beta}] = 0$ より, $\hat{\sigma}^2$ と $(\hat{\alpha}, \hat{\beta})$ は独立 □

2 線型重回帰モデル

$d, n \in \mathbb{N}$ とし, n 個の観測の組を

$$(y_j, x_{j1}, x_{j2}, \dots, x_{jd}) \quad (j = 1, 2, \dots, n)$$

とする. ただし, $x_{j\ell}$ ($j = 1, 2, \dots, n; \ell = 1, 2, \dots, d$) 変量は定数とする. さらに, 変数間には

$$y_j = \beta_0^* + \beta_1^* x_{j1} + \beta_2^* x_{j2} + \dots + \beta_d^* x_{jd} + \epsilon_j \quad (j = 1, 2, \dots, n) \quad (1.5)$$

なるモデルを仮定する. ただし, β_0^* は y 切片項, $\beta_1^*, \beta_2^*, \dots, \beta_d^*$ は重回帰係数という. これらは未知とする. また, u_1, u_2, \dots, u_n は独立同一分布に従う確率変数列 (誤差項) であり, $\forall[\epsilon_j] = \sigma^2$ ($j = 1, 2, \dots, n$) とする. ただし, σ^2 ($\sigma > 0$) は未知とする. (1.5) を重回帰モデルという.

重回帰モデル (1.5) は

$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}_{=\mathbf{y}} = \underbrace{\begin{pmatrix} 1 & x_{11} & \cdots & x_{1d} \\ 1 & x_{21} & \cdots & x_{2d} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nd} \end{pmatrix}}_{=\mathbf{X}} \underbrace{\begin{pmatrix} \beta_0^* \\ \beta_1^* \\ \vdots \\ \beta_d^* \end{pmatrix}}_{=\boldsymbol{\beta}^*} + \underbrace{\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}}_{=\boldsymbol{\epsilon}}$$

と表される. これは

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon} \quad (1.6)$$

と書ける.

回帰係数ベクトル $\boldsymbol{\beta}^*$ の最小 2 乗推定量は

$$h(\boldsymbol{\beta}) = h(\beta_0, \beta_1, \dots, \beta_d) = \sum_{j=1}^n \{y_j - (\beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \dots + \beta_d x_{jd})\}^2$$

を最小化することにより得られる. 関数 h は

$$h(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

と書ける. ただし, 行列 A に対して, A^\top はその転置である.

以後、 $\mathbf{X}^\top \mathbf{X}$ は正則と仮定する。ここで、

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

とおくと

$$h(\beta) = (\mathbf{y} - \mathbf{X}\hat{\beta})^\top (\mathbf{y} - \mathbf{X}\hat{\beta}) + (\hat{\beta} - \beta)^\top \mathbf{X}^\top \mathbf{X} (\hat{\beta} - \beta) \quad (1.7)$$

と書ける。このことより、 $h(\beta)$ は $\beta = \hat{\beta}$ で最小となる。よって、 $\hat{\beta}$ は β^* の最小 2 乗推定量である。

$$\hat{\beta} - \beta^* = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \epsilon$$

と書けるので、

$$\begin{aligned} \mathbb{E}[\hat{\beta}] &= \beta^*, \\ \text{COV}[\hat{\beta}] &= \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}. \end{aligned}$$

さらに、

$$\text{RSS} := (\mathbf{y} - \mathbf{X}\hat{\beta})^\top (\mathbf{y} - \mathbf{X}\hat{\beta}) = \mathbf{y}^\top \{ \mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \} \mathbf{y}$$

とする。トレース、期待値およびベキ等行列の性質より

$$\mathbb{E}[\text{RSS}] = \text{Tr} \left[\{ \mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \} \mathbb{E}[\epsilon \epsilon^\top] \right] \quad (1.8)$$

$$= \sigma^2 \text{Tr} \left[\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \right] \quad (1.9)$$

$$= (n - d - 1) \sigma^2. \quad (1.10)$$

よって、 $n \geq d + 2$ のとき、 σ^2 の不偏推定量は

$$\hat{\sigma}^2 = \frac{1}{n - d - 1} \text{RSS}.$$

定理 1.3 (Gauss-Markov の定理) 最小 2 乗推定量 $\hat{\beta}$ は最良線型不偏推定量 (best linear unbiased estimator, BLUE) である。

証明 任意の線型推定量は $(d+1) \times n$ 行列 C を用いて Cy と表される. これが不偏推定量となるためには,

$$\mathbb{E}[Cy] = \beta^* \iff CX = I_{d+1}$$

をみたさなければならない. さらに,

$$\text{COV}[Cy] = \mathbb{E}[(Cy - \beta^*)(Cy - \beta^*)^\top] = C\mathbb{E}[uu^\top]C^\top = \sigma^2 CC^\top.$$

最小 2 乗推定量は $C^* := (X^\top X)^{-1}X^\top$ に対応しており,

$$\begin{aligned} CC^\top &= (C - C^* + C^*)(C - C^* + C^*)^\top \\ &= (C - C^*)(C - C^*)^\top + C^*(C^*)^\top \\ &\quad + (C - C^*)(C^*)^\top + C^*(C - C^*)^\top. \end{aligned}$$

しかし

$$(C - C^*)(C^*)^\top = \{C - (X^\top X)^{-1}X^\top\}X = CX - I_{d+1} = 0.$$

よって

$$\begin{aligned} \text{COV}[Cy] &= \sigma^2 CC^\top = \sigma^2 C^*(C^*)^\top + \sigma^2 (C - C^*)(C - C^*)^\top \\ &\succcurlyeq \sigma^2 C^*(C^*)^\top = \text{COV}[\hat{\beta}]. \end{aligned}$$

ただし, 正方行列 A, B に対して,

$$A \succcurlyeq B \iff A - B \succcurlyeq 0 \iff A - B \text{ は非負定値}$$

とした. したがって, 最小 2 乗推定量は共分散行列を最小 (上の \succcurlyeq の意味で最小) となるので, 線型不偏推定量の族の中で最良である. \square

命題 1.4 $n \geq d+2$ とする. $\epsilon \sim N_n(0, \sigma^2 I_n)$ を仮定する. このとき, 以下のことが成り立つ.

- (1) $\hat{\beta} \sim N_{d+1}(\beta^*, \sigma^2 (X^\top X)^{-1})$.
- (2) $\frac{(n-d-1)\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-d-1}^2$.
- (3) $\hat{\beta}$ と $\hat{\sigma}^2$ は独立.

証明 $z = (y - X\beta^*)/\sigma$ とし,

$$U = (X^\top X)^{-1/2} X^\top z; \quad V = z^\top \{I_n - X(X^\top X)^{-1} X^\top\} z$$

とおく. ただし, 正則な行列 A に対して, $A^{-1} = BB^\top$ をみたす行列 B を $A^{-1/2}$ ¹ と記した. すると, 命題の主張は

$$(1a) U \sim N_{d+1}(\mathbf{0}, I_{d+1}),$$

$$(2a) V \sim \chi_{n-d-1}^2,$$

$$(3a) U \text{ と } V \text{ は独立},$$

という形に簡略化される. X は $n \times (d+1)$ 行列で, ランクが $d+1$ (フルランク) だから

$$X = P\Lambda O^\top$$

と分解できる. ただし, P は $n \times (d+1)$ 行列で $P^\top P = I_{d+1}$, Λ は正の成分をもつ $d+1$ 次の対角行列で, O は $d+1$ 次の直交行列である. このとき,

$$(X^\top X)^{-1} = \{O\Lambda P^\top P\Lambda O^\top\}^{-1} = \{O\Lambda^2 O^\top\}^{-1} = O\Lambda^{-2} O^\top$$

より

$$(X^\top X)^{-1/2} = O\Lambda^{-1} O^\top.$$

これらより

$$U = OP^\top z; \quad V = z^\top (I_n - PP^\top) z$$

と書けることがわかる. ここで, 適当な $n \times (n-d-1)$ 行列 Q をうまくとり, $n \times n$ 行列 $H := (P : Q)$ が n 次直交行列になるようにする. すると, $I_n = HH^\top = PP^\top + QQ^\top$ であることから, $I_n - PP^\top = QQ^\top$. よって,

$$V = z^\top QQ^\top z = (Q^\top z)^\top (Q^\top z)$$

と表される.

$$H^\top z = \begin{pmatrix} P^\top z \\ Q^\top z \end{pmatrix} \sim N_n(\mathbf{0}, I_n)$$

であるから, $U = P^\top z$ と $V = z^\top QQ^\top z$ は独立で, $U \sim N_{d+1}(\mathbf{0}, I_{d+1})$, $V \sim \chi_{n-d-1}^2$ となる. □

¹これは一意ではないことに注意せよ.

3 変数選択の規準

説明変数 $x_{1j}, x_{2j}, \dots, x_{dj}$ の個数 d を増やすにつれて、線型モデル

$$y_j = \beta_0^* + \beta_1^* x_{1j} + \beta_2^* x_{2j} + \dots + \beta_d^* x_{dj} + \epsilon_j \quad (j = 1, 2, \dots, n)$$

によるデータに対する説明力は増していき、観測値に対するモデルの適合度は高くなる。しかし、 d を増やしていくにつれて、未知母数である回帰係数の推定量の推定誤差は増していく。あとの章でわかるように、

$$\mathbb{E}[\text{Tr}\{(\mathbf{X}\hat{\beta} - \mathbf{X}\beta^*)^\top(\mathbf{X}\hat{\beta} - \mathbf{X}\beta^*)\}] \leq C \frac{\sigma^2 d}{n}$$

と評価できる。ただし、 C は n, d に依らない定数である。

以下では、 $n \geq d + 1$ とし、 \mathbf{X} はフルランクとする。

3.1 自由度調整済み決定係数

決定係数

$$R_d^2 = 1 - \frac{\sum_{j=1}^n (y_j - \hat{y}_j)^2}{\sum_{j=1}^n (y_j - \bar{y})^2}; \quad \hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 x_{1j} + \hat{\beta}_2 x_{2j} + \dots + \hat{\beta}_d x_{dj}$$

の値は説明変数の個数 d を増やせば、1 に近づいていく。そこで、 $\sum_{j=1}^n (y_j - \hat{y}_j)^2$ と $\sum_{j=1}^n (y_j - \bar{y})^2$ をそれらの自由度 $n - d - 1, n - 1$ で割ったもので置き換えたもの

$$R_d^{*2} = 1 - \frac{\sum_{j=1}^n (y_j - \hat{y}_j)^2 / (n - d - 1)}{\sum_{j=1}^n (y_j - \bar{y})^2 / (n - 1)}$$

を自由度調整済み決定係数という。これを書き直せば

$$R_d^{*2} = 1 - \frac{n-1}{n-d-1} (1 - R_d^2)$$

となる。この形からわかるように、 d が大きくなると、 $1 - R_d^2$ は小さくなるものの、分母の $n - d - 1$ が小さくなるので、 R_d^{*2} は必ずしも 1 には近づかない。 R_d^{*2} を最大にする説明変数の組 (d の選択) を選択すればよい。

3.2 Mallows C_p

モデルの誤差項 ϵ とは独立な確率ベクトル $\tilde{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ に基づく将来の観測を

$$\tilde{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta}^* + \tilde{\epsilon}$$

とする. 未来の観測 $\tilde{\mathbf{y}}$ を $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ で予測するとき, その平均 2 乗誤差は

$$\begin{aligned} \text{MSE}(\hat{\mathbf{y}}) &= \mathbb{E}[(\tilde{\mathbf{y}} - \hat{\mathbf{y}})^\top (\tilde{\mathbf{y}} - \hat{\mathbf{y}})] \\ &= \mathbb{E}[\{\tilde{\epsilon} - \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\}^\top \{\tilde{\epsilon} - \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\}] \\ &= n\sigma^2 + \mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^\top \mathbf{X}^\top \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)] \\ &= n\sigma^2 + \text{Tr}\{\mathbf{X}^\top \mathbf{X} \mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^\top]\} \\ &= n\sigma^2 + (d+1)\sigma^2 = (n+d+1)\sigma^2 \end{aligned}$$

となる. 一方, 残差平方和 $\text{RSS}_d = (\mathbf{y} - \hat{\mathbf{y}})^\top (\mathbf{y} - \hat{\mathbf{y}})$ の期待値は

$$\mathbb{E}[\text{RSS}_d] = \mathbb{E}[\mathbf{y}^\top \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top\} \mathbf{y}] = (n-d-1)\sigma^2$$

となるので, 予測誤差は

$$\text{MSE}(\hat{\mathbf{y}}) = \mathbb{E}[\text{RSS}_d + 2(d+1)\sigma^2]$$

となるので, σ^2 が既知の場合, $\text{RSS}_d + 2(d+1)\sigma^2$ が予測誤差の不偏推定量になっていることが分かる.

候補となる最大のモデルの説明変数の組を

$$(x_{j_1}, x_{j_2}, \dots, x_{j_d}, \dots, x_{j_K}) \quad (1 \leq d \leq K)$$

とする.

$$\mathbf{X}_F = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1d} & \dots & x_{1K} \\ x_{21} & x_{22} & \dots & x_{2d} & \dots & x_{2K} \\ \vdots & & \vdots & & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nd} & \dots & x_{nK} \end{pmatrix}$$

としたとき, 分散 σ^2 の推定量を

$$\hat{\sigma}_F^2 = \mathbf{y}^\top \{\mathbf{I}_n - \mathbf{X}_F(\mathbf{X}_F^\top \mathbf{X}_F)^{-1} \mathbf{X}_F^\top\} \mathbf{y}$$

とする. σ^2 の推定量 $\hat{\sigma}_F^2$ で割ったもの

$$C_p := \frac{\text{RSS}_d}{\hat{\sigma}_F^2} + 2(d+1)$$

を Mallows の C_p 規準といい, これを最小にする変数の組を選べばよい. これは

$$(\text{モデルの適合度}) + 2 \times (\text{モデルの母数の個数})$$

という形をしており, 第 2 項はモデルの複雑さに対する罰則項として機能する.

3.3 AIC

定義 1.5 (Kullback-Leibler 情報量) f, g を Lebesgue 測度 μ に関する確率密度関数とする. Kullback-Leibler 情報量を

$$\text{KL}(f, g) = \int \log \left(\frac{f}{g} \right) f d\mu$$

で定義する.

注意 1.6 $x \log x \geq x - 1$ ($x > 0$) に注意して,

$$\begin{aligned} \text{KL}(f, g) &= \int \log \left(\frac{f}{g} \right) f d\mu = \int \frac{f}{g} \log \left(\frac{f}{g} \right) g d\mu \\ &\geq \int \left(\frac{f}{g} - 1 \right) g d\mu = 0. \end{aligned}$$

□

以下では, $n \geq d+4$ とする. $\mathbf{y} = \mathbf{X}\beta^* + \epsilon$ の確率密度関数を $f(\mathbf{y} | \mathbf{X}\beta^*, \sigma^2)$ とし, 将来の変数 $\tilde{\mathbf{y}} = \mathbf{X}\beta^* + \tilde{\epsilon}$ の確率密度関数を $f(\tilde{\mathbf{y}} | \mathbf{X}\beta^*, \sigma^2)$ と書くことにする. β^*, σ^2 の推定量 $\hat{\beta}(\mathbf{y}), \hat{\sigma}^2(\mathbf{y})$ を $f(\tilde{\mathbf{y}} | \mathbf{X}\beta^*, \sigma^2)$ に代入したもの

$$f(\tilde{\mathbf{y}} | \mathbf{X}\hat{\beta}(\mathbf{y}), \hat{\sigma}^2(\mathbf{y}))$$

を推定されたモデルの分布とし, これで将来の分布 $f(\tilde{\mathbf{y}} | \mathbf{X}\beta^*, \sigma^2)$ を予測するとき, それらの分布間の Kullback-Leibler 情報量

$$\begin{aligned} &\text{KL}(f(\cdot | \mathbf{X}\beta^*, \sigma^2), f(\cdot | \mathbf{X}\hat{\beta}(\mathbf{y}), \hat{\sigma}^2(\mathbf{y}))) \\ &= \int \cdots \int \log \left\{ \frac{f(\tilde{\mathbf{y}} | \mathbf{X}\beta^*, \sigma^2)}{f(\tilde{\mathbf{y}} | \mathbf{X}\hat{\beta}(\mathbf{y}), \hat{\sigma}^2(\mathbf{y}))} \right\} \cdot f(\tilde{\mathbf{y}} | \mathbf{X}\beta^*, \sigma^2) d\tilde{\mathbf{y}} \end{aligned}$$

で測る。以後、ルベーク測度 $d\mu(\tilde{\mathbf{y}})$ を $d\tilde{\mathbf{y}}$ と書くことにする。これは \mathbf{y} に依存するランダムな量なので、 \mathbf{y} に関して期待値を取ったもの

$$\mathbb{E} \left[\text{KL}(f(\cdot | \mathbf{X}\boldsymbol{\beta}^*, \sigma^2), f(\cdot | \mathbf{X}\hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\sigma}^2(\mathbf{y}))) \right]$$

を考える。すると、この関数が Mallows の C_p 規準で考えたところの平均 2 乗予測誤差に対応している。

$$\begin{aligned} & \mathbb{E} \left[\text{KL}(f(\cdot | \mathbf{X}\boldsymbol{\beta}^*, \sigma^2), f(\cdot | \mathbf{X}\hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\sigma}^2(\mathbf{y}))) \right] \\ &= \mathbb{E} \left[\int \cdots \int \log\{f(\tilde{\mathbf{y}} | \mathbf{X}\boldsymbol{\beta}^*, \sigma^2)\} f(\tilde{\mathbf{y}} | \mathbf{X}\boldsymbol{\beta}^*, \sigma^2) d\tilde{\mathbf{y}} \right] \\ & \quad - \mathbb{E} \left[\int \cdots \int \log\{f(\tilde{\mathbf{y}} | \mathbf{X}\hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\sigma}^2(\mathbf{y}))\} f(\tilde{\mathbf{y}} | \mathbf{X}\boldsymbol{\beta}^*, \sigma^2) d\tilde{\mathbf{y}} \right] \end{aligned}$$

と書き直せる。右辺の第 1 項目は推定されたモデルの分布に無関係なので、後者を 2 倍したものを

$$\text{AI}(\boldsymbol{\beta}^*, \sigma^2) := -2\mathbb{E} \left[\int \cdots \int \log\{f(\tilde{\mathbf{y}} | \mathbf{X}\hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\sigma}^2(\mathbf{y}))\} f(\tilde{\mathbf{y}} | \mathbf{X}\boldsymbol{\beta}^*, \sigma^2) d\tilde{\mathbf{y}} \right]$$

とおく。これを赤池情報量と呼ぶ。これの (漸近) 不偏推定量が AIC 規準となる。具体的に、 $\text{AI}(\boldsymbol{\beta}^*, \sigma^2)$ を計算してみると

$$-2\log f(\tilde{\mathbf{y}} | \mathbf{X}\hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\sigma}^2(\mathbf{y})) = n \log(2\pi\hat{\sigma}^2(\mathbf{y})) + \frac{(\tilde{\mathbf{y}} - \mathbf{X}\hat{\boldsymbol{\beta}}(\mathbf{y}))^\top (\tilde{\mathbf{y}} - \mathbf{X}\hat{\boldsymbol{\beta}}(\mathbf{y}))}{\hat{\sigma}^2(\mathbf{y})}$$

であることと命題 1.4 に注意し、 $\tilde{\mathbf{y}}$ に関して積分²する。

$$\begin{aligned} & \int \cdots \int \left\{ -2\log f(\tilde{\mathbf{y}} | \mathbf{X}\hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\sigma}^2(\mathbf{y})) \right\} f(\tilde{\mathbf{y}} | \mathbf{X}\boldsymbol{\beta}^*, \sigma^2) d\tilde{\mathbf{y}} \\ &= \int \cdots \int \left\{ n \log(2\pi\hat{\sigma}^2(\mathbf{y})) + \frac{\{\tilde{\boldsymbol{\epsilon}} + \mathbf{X}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}(\mathbf{y}))\}^\top \{\tilde{\boldsymbol{\epsilon}} + \mathbf{X}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}(\mathbf{y}))\}}{\hat{\sigma}^2(\mathbf{y})} \right\} \\ & \quad \times f(\tilde{\boldsymbol{\epsilon}} | \mathbf{0}, \sigma^2) d\tilde{\boldsymbol{\epsilon}} \quad \left(\tilde{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta}^* + \tilde{\boldsymbol{\epsilon}} \text{ と変換} \right) \\ &= n \log(2\pi\hat{\sigma}^2(\mathbf{y})) + \frac{n\sigma^2 + (\hat{\boldsymbol{\beta}}(\mathbf{y}) - \boldsymbol{\beta}^*)^\top \mathbf{X}^\top \mathbf{X} (\hat{\boldsymbol{\beta}}(\mathbf{y}) - \boldsymbol{\beta}^*)}{\hat{\sigma}^2(\mathbf{y})} \end{aligned}$$

²実際には、 $\tilde{\boldsymbol{\epsilon}} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ に関して積分していることに注意せよ。

と書ける. 命題 1.4 から

$$\begin{aligned}\mathbb{E}[(\widehat{\beta}(\mathbf{y}) - \beta^*)^\top \mathbf{X}^\top \mathbf{X} (\widehat{\beta}(\mathbf{y}) - \beta^*)] &= (d+1)\sigma^2, \\ \mathbb{E}\left[\frac{\sigma^2}{\widehat{\sigma}^2(\mathbf{y})}\right] &= (n-d-1)\mathbb{E}\left[\frac{1}{\chi_{n-d-1}^2}\right] = \frac{n-d-1}{n-d-3}\end{aligned}$$

である. さらに, $\widehat{\beta}(\mathbf{y})$ と $\widehat{\sigma}^2(\mathbf{y})$ の独立性であることに注意すれば,

$$\text{AI}(\beta^*, \sigma^2) = \mathbb{E}[n \log(2\pi\widehat{\sigma}^2(\mathbf{y}))] + \frac{(n+d+1)(n-d-1)}{n-d-3} \quad (1.11)$$

と書けることがわかる.

次に, 対数尤度関数を計算する.

$$\begin{aligned}\widehat{\sigma}^2(\mathbf{y}) &= \frac{1}{n-d-1} \mathbf{y}^\top \{ \mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \} \mathbf{y} \\ &= \frac{1}{n-d-1} (\mathbf{y} - \mathbf{X}\widehat{\beta}(\mathbf{y}))^\top (\mathbf{y} - \mathbf{X}\widehat{\beta}(\mathbf{y}))\end{aligned}$$

に注意して,

$$\begin{aligned}\mathbb{E}\left[-\log f((\mathbf{y} | \mathbf{X}\widehat{\beta}(\mathbf{y}), \widehat{\sigma}^2(\mathbf{y}))\right] \\ &= \mathbb{E}\left[n \log(2\pi\widehat{\sigma}^2(\mathbf{y})) + \frac{(\mathbf{y} - \mathbf{X}\widehat{\beta}(\mathbf{y}))^\top (\mathbf{y} - \mathbf{X}\widehat{\beta}(\mathbf{y}))}{\widehat{\sigma}^2(\mathbf{y})}\right] \\ &= \mathbb{E}[n \log(2\pi\widehat{\sigma}^2(\mathbf{y}))] + (n-d-1)\end{aligned}$$

となる. これを (1.11) に代入すれば,

$$\text{AI}(\beta^*, \sigma^2) = \mathbb{E}\left[-2 \log f(\mathbf{y} | \mathbf{X}\widehat{\beta}(\mathbf{y}), \widehat{\sigma}^2(\mathbf{y})) + 2(d+2) \frac{n-d-1}{n-d-3}\right]$$

と書けることがわかる. すなわち, 上の式の右辺の期待値記号の中身が $\text{AI}(\beta^*, \sigma^2)$ の不偏推定量になる.

$$\lim_{n \rightarrow \infty} \frac{n-d-1}{n-d-3} = 1$$

なので, 近似推定量として,

$$\text{AIC} := -2 \log f(\mathbf{y} | \mathbf{X}\widehat{\beta}(\mathbf{y}), \widehat{\sigma}^2(\mathbf{y})) + 2(d+2) \quad (1.12)$$

が得られる. これを AIC 規準という. AIC を最小化する変数の組を選べばよい. Mallows の C_p 規準と同様, (1.12) の第 1 項目はデータの適合度, 第 2 項目はモデルの複雑さに対する罰則と解釈できる.

3.4 交差検証法

観測 $(x_1, y_1), \dots, (x_2, y_2), \dots, (x_n, y_n)$ から j 番目 ($j = 1, 2, \dots, n$) のデータ (x_j, y_j) を除いた残りのデータからの β^* の推定量を考える.

$$\begin{aligned}\widehat{\beta}^{(j)} &= \{(\mathbf{X}^{(j)})^\top \mathbf{X}^{(j)}\}^{-1} (\mathbf{X}^{(j)})^\top \mathbf{y}^{(j)}, \\ \mathbf{y}^{(j)} &= (y_1, \dots, y_{j-1}, y_{j+1}, \dots, y_n)^\top, \\ (\mathbf{X}^{(j)})^\top &= (\mathbf{x}_1, \dots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \dots, \mathbf{x}_n).\end{aligned}$$

を考える. y_j の予測量 $\widehat{y}_j = \mathbf{x}_j^\top \widehat{\beta}^{(j)}$ を構成し, y_j に対する予測誤差

$$\{y_j - \mathbf{x}_j^\top \widehat{\beta}^{(j)}\}^2$$

を計算する. この操作を繰り返し, 予測誤差

$$\text{CV} := \frac{1}{n} \sum_{j=1}^n \{y_j - \mathbf{x}_j^\top \widehat{\beta}^{(j)}\}^2$$

を得る. これを交差検証法 (クロス・バリデーション) といい, CV を最小にする説明変数の組を選ぶ.

命題 1.7 CV は以下のように表現できる.

$$\text{CV} = \frac{1}{n} \sum_{j=1}^n \left\{ \frac{y_j - \mathbf{x}_j^\top \widehat{\beta}^*}{1 - \mathbf{x}_j^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_j} \right\}^2.$$

証明 命題の主張を証明するために,

$$\frac{y_1 - \mathbf{x}_1^\top \widehat{\beta}}{1 - \mathbf{x}_1^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_1} = y_1 - \mathbf{x}_1^\top \widehat{\beta}^{(1)}$$

を示せばよい.

$$\begin{aligned}\mathbf{X}^\top &= (\mathbf{x}_1; (\mathbf{X}^{(1)})^\top), \\ \mathbf{X}^\top \mathbf{X} &= \mathbf{x}_1 \mathbf{x}_1^\top + (\mathbf{X}^{(1)})^\top \mathbf{X}^{(1)} =: \mathbf{x}_1 \mathbf{x}_1^\top + A\end{aligned}$$

に注意する. 等式

$$(\mathbf{X}^\top \mathbf{X})^{-1} = A^{-1} - \frac{A^{-1} \mathbf{x}_1 \mathbf{x}_1^\top A^{-1}}{1 + \mathbf{x}_1^\top A^{-1} \mathbf{x}_1} \quad (1.13)$$

より

$$\begin{aligned}
1 - \mathbf{x}_1^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_1 &= 1 - \mathbf{x}_1^\top A^{-1} \mathbf{x}_1 + \frac{\mathbf{x}_1^\top A^{-1} \mathbf{x}_1 \mathbf{x}_1^\top A^{-1} \mathbf{x}_1}{1 + \mathbf{x}_1^\top A^{-1} \mathbf{x}_1} \\
&= \frac{1 - (\mathbf{x}_1^\top A^{-1} \mathbf{x}_1)^2 + (\mathbf{x}_1^\top A^{-1} \mathbf{x}_1)^2}{1 + \mathbf{x}_1^\top A^{-1} \mathbf{x}_1} \\
&= \frac{1}{1 + \mathbf{x}_1^\top A^{-1} \mathbf{x}_1} \tag{1.14}
\end{aligned}$$

を得る. これらより

$$\begin{aligned}
\mathbf{x}_1^\top \widehat{\boldsymbol{\beta}}^* &= \mathbf{x}_1^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \\
&= \mathbf{x}_1^\top \left\{ A^{-1} - \frac{A^{-1} \mathbf{x}_1 \mathbf{x}_1^\top A^{-1}}{1 + \mathbf{x}_1^\top A^{-1} \mathbf{x}_1} \right\} (\mathbf{x}_1; (\mathbf{X}^{(1)})^\top) \begin{pmatrix} y_1 \\ \mathbf{y}^{(1)} \end{pmatrix} \\
&= \mathbf{x}_1^\top \left\{ A^{-1} - \frac{A^{-1} \mathbf{x}_1 \mathbf{x}_1^\top A^{-1}}{1 + \mathbf{x}_1^\top A^{-1} \mathbf{x}_1} \right\} \{ \mathbf{x}_1 y_1 + (\mathbf{X}^{(1)})^\top \mathbf{y}^{(1)} \} \\
&= \mathbf{x}_1 \left\{ A^{-1} \mathbf{x}_1 y_1 - \frac{A^{-1} \mathbf{x}_1 \mathbf{x}_1^\top A^{-1} \mathbf{x}_1 y_1}{1 + \mathbf{x}_1^\top A^{-1} \mathbf{x}_1} + A^{-1} (\mathbf{X}^{(1)})^\top \mathbf{y}^{(1)} \right. \\
&\quad \left. - \frac{A^{-1} \mathbf{x}_1 \mathbf{x}_1^\top A^{-1} (\mathbf{X}^{(1)})^\top \mathbf{y}^{(1)}}{1 + \mathbf{x}_1^\top A^{-1} \mathbf{x}_1} \right\} \\
&= \mathbf{x}_1^\top A^{-1} \mathbf{x}_1 y_1 - \frac{(\mathbf{x}_1^\top A^{-1} \mathbf{x}_1)^2 y_1}{1 + \mathbf{x}_1^\top A^{-1} \mathbf{x}_1} + \mathbf{x}_1^\top A^{-1} (\mathbf{X}^{(1)})^\top \mathbf{y}^{(1)} \\
&\quad - \frac{\mathbf{x}_1^\top A^{-1} \mathbf{x}_1 \mathbf{x}_1^\top A^{-1} (\mathbf{X}^{(1)})^\top \mathbf{y}^{(1)}}{1 + \mathbf{x}_1^\top A^{-1} \mathbf{x}_1} \\
&= \mathbf{x}_1^\top A^{-1} \mathbf{x}_1 y_1 - \frac{(\mathbf{x}_1^\top A^{-1} \mathbf{x}_1)^2 y_1}{1 + \mathbf{x}_1^\top A^{-1} \mathbf{x}_1} + \mathbf{x}_1^\top \widehat{\boldsymbol{\beta}}^{*(1)} - \frac{\mathbf{x}_1^\top A^{-1} \mathbf{x}_1 \mathbf{x}_1^\top \widehat{\boldsymbol{\beta}}^{*(1)}}{1 + \mathbf{x}_1^\top A^{-1} \mathbf{x}_1} \\
&= \frac{\mathbf{x}_1^\top A^{-1} \mathbf{x}_1 y_1 + \mathbf{x}_1^\top \widehat{\boldsymbol{\beta}}^{*(1)}}{1 + \mathbf{x}_1^\top A^{-1} \mathbf{x}_1} \tag{1.15}
\end{aligned}$$

を得る. (1.14) と (1.15) より

$$\begin{aligned}
 \frac{y_1 - \mathbf{x}_1^\top \widehat{\boldsymbol{\beta}}^{*(1)}}{1 - \mathbf{x}_1^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_1} &= (1 + \mathbf{x}_1^\top A^{-1} \mathbf{x}_1) \left\{ y_1 - \frac{\mathbf{x}_1^\top A^{-1} \mathbf{x}_1 y_1 + \mathbf{x}_1^\top \widehat{\boldsymbol{\beta}}^{*(1)}}{1 + \mathbf{x}_1^\top A^{-1} \mathbf{x}_1} \right\} \\
 &= (1 + \mathbf{x}_1^\top A^{-1} \mathbf{x}_1) \left\{ \frac{y_1 - \mathbf{x}_1^\top \widehat{\boldsymbol{\beta}}^{*(1)}}{1 + \mathbf{x}_1^\top A^{-1} \mathbf{x}_1} \right\} \\
 &= y_1 - \mathbf{x}_1^\top \widehat{\boldsymbol{\beta}}^{*(1)}
 \end{aligned}$$

を得る. □

3.5 BIC

線型回帰モデル

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}, \quad \text{COV}[\boldsymbol{\epsilon}] = \sigma^2 \mathbf{I}_n$$

において, $(d+2)$ 個の未知母数 $\boldsymbol{\beta}^*, \sigma^2$ に正則な事前分布 $\pi_{d+2}(\boldsymbol{\beta}^*, \sigma^2)$ を仮定する. ここで, 「正則」とは

$$\int \cdots \int \pi_{d+2}(\boldsymbol{\beta}^*, \sigma^2) d\boldsymbol{\beta}^* d\sigma^2 = 1$$

が成立していることである. すると \mathbf{y} の周辺分布

$$f_{\pi_{d+2}}(\mathbf{y}) = \int \cdots \int f(\mathbf{y} | \mathbf{X}\boldsymbol{\beta}^*, \sigma^2) \pi_{d+2}(\boldsymbol{\beta}^*, \sigma^2) d\boldsymbol{\beta}^* d\sigma^2$$

与えられる. これは σ^2 と $(d+1)$ 個の回帰係数 $\beta_0^*, \beta_1^*, \dots, \beta_d^*$ に事前分布を想定した Bayes 的周辺尤度である. これを最大にする説明変数の組 (Bayes 的周辺尤度 $-2 \log f_{\pi_{d+2}}(\mathbf{y})$ を最小にする説明変数の組) を選択する. 基準となるモデルを定め, 比較するモデルとの周辺尤度の比を Bayes 因子 (Bayes factor) と呼ぶ.

例えば, 最も簡単なモデル

$$y_j = \beta_0^* + \epsilon_j \quad (j = 1, 2, \dots, n) \tag{1.16}$$

を考え, 2 個の未知母数 β_0^*, σ^2 に事前分布 $\pi_2(\beta_0^*, \sigma^2)$ を想定した Bayes 的周辺尤度

$$f_{\pi_2}(\mathbf{y}) = \int \int f(\mathbf{y} | \beta_0^*, \sigma^2) \pi_2(\beta_0^*, \sigma^2) d\beta_0^* d\sigma^2$$

を考える。ただし、 $f(\mathbf{y}|\beta_0^*, \sigma^2)$ はモデル (1.16) の確率密度関数である。これらの比

$$\frac{f_{\pi_{d+2}}(\mathbf{y})}{f_{\pi_2}(\mathbf{y})}$$

を考える。この値が $1/2$ を超えていれば、 d 個の説明変数は Bayes 的な意味を持つと解釈できる。これが Bayes 因子であり、この値を最大にする説明変数の組を選択すればよい。

Bayes 的周辺尤度や Bayes 因子の問題点は、これらが事前分布の取り方に依存していることである。ここでは、 $n \rightarrow \infty$ とすることにより、その極限を求める。

そのために、Laplace 近似を用いる。以下の議論は数学的な厳密性にかけるものであることに注意せよ。章末の補遺を参照のこと。一般に、 θ を p 次元の未知母数とし、 X を n 次元確率変数とし、 θ を与えたときの X の条件付き確率密度関数を

$$X|\theta \sim f(x|\theta)$$

とし、 θ の事前分布を

$$\theta \sim \pi(\theta)$$

とする。対数尤度

$$\ell(\theta|x) = \log f(x|\theta)$$

を θ の最尤推定量³ $\hat{\theta}$ のまわりで Taylor 展開する。

$$\begin{aligned} \ell(\theta|x) &\approx \ell(\hat{\theta}|x) + \left(\frac{\partial}{\partial \theta} \ell(\theta|x) \Big|_{\theta=\hat{\theta}} \right)^\top (\hat{\theta} - \theta) \\ &\quad + \frac{1}{2} (\hat{\theta} - \theta)^\top \left(\frac{\partial^2}{\partial \theta \partial \theta^\top} \ell(\theta|x) \Big|_{\theta=\hat{\theta}} \right) (\hat{\theta} - \theta) \end{aligned}$$

と近似できる。ただし、

$$\begin{aligned} \frac{\partial}{\partial \theta} \ell(\theta|x) &= \left(\frac{\partial}{\partial \theta_1} \ell(\theta|x), \dots, \frac{\partial}{\partial \theta_p} \ell(\theta|x) \right)^\top \\ \frac{\partial^2}{\partial \theta \partial \theta^\top} \ell(\theta|x) &= \left(\frac{\partial^2}{\partial \theta_j \partial \theta_k} \ell(\theta|x) \right)_{j, k=1, 2, \dots, p} \\ \theta &= (\theta_1, \theta_2, \dots, \theta_p)^\top \end{aligned}$$

³すなわち、存在すれば、

$$\hat{\theta} \in \operatorname{argmax}_{\theta \in \Theta} \ell(\theta|x)$$

で定義する。

である。さらに,

$$\widehat{I}(\mathbf{x}) = -\frac{1}{n} \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \ell(\boldsymbol{\theta} | \mathbf{x}) \Big|_{\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}}$$

とおいたとき, $\widehat{I}(\mathbf{X})$ はある正定値行列に確率収束すると仮定する. $\widehat{\boldsymbol{\theta}}$ は $\boldsymbol{\theta}$ の最尤推定量だから,

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta} | \mathbf{x}) \Big|_{\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}} = \mathbf{0}$$

である. よって,

$$\ell(\boldsymbol{\theta} | \mathbf{x}) \approx \ell(\widehat{\boldsymbol{\theta}} | \mathbf{x}) - \frac{n}{2} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top \widehat{I}(\mathbf{x}) (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$$

と近似できる. $\pi(\boldsymbol{\theta})$ は $\boldsymbol{\theta}$ に関して滑らかな関数で

$$\pi(\boldsymbol{\theta}) \approx \pi(\widehat{\boldsymbol{\theta}})$$

と近似できると仮定すると, Bayes 的周辺分布は

$$\begin{aligned} f_\pi(\mathbf{x}) &= \int \cdots \int f(\mathbf{x} | \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \int \cdots \int \exp\{\ell(\boldsymbol{\theta} | \mathbf{x})\} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &\approx f(\mathbf{x} | \widehat{\boldsymbol{\theta}}) \frac{(2\pi)^{p/2}}{|n\widehat{I}(\mathbf{x})|^{1/2}} \pi(\widehat{\boldsymbol{\theta}}) \\ &\quad \times \int \cdots \int \frac{|n\widehat{I}(\mathbf{x})|^{1/2}}{(2\pi)^{p/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^\top n\widehat{I}(\mathbf{x})(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})\right\} d\boldsymbol{\theta} \\ &= f(\mathbf{x} | \widehat{\boldsymbol{\theta}}) \frac{(2\pi)^{p/2}}{|n\widehat{I}(\mathbf{x})|^{1/2}} \pi(\widehat{\boldsymbol{\theta}}) \end{aligned}$$

で近似できることが知られている. これを Laplace 近似という. ここで「 \approx 」の意味は以下である. $a \in \mathbb{R}$ と $\{a_n\}_{n=1}^\infty$ に対して,

$$a \approx a_n \iff \lim_{n \rightarrow \infty} \frac{a}{a_n} = 1$$

である. したがって,

$$-2 \log f_\pi(\mathbf{x}) \approx -2 \log f(\mathbf{x} | \widehat{\boldsymbol{\theta}}) + p \log n + p \log \left(\frac{|\widehat{I}(\mathbf{x})|^{1/(2p)}}{2\pi} \right) - 2 \log \pi(\widehat{\boldsymbol{\theta}})$$

と書くことができる. n が大きいとき, 定式の右辺の 3, 4 項目

$$p \log \left(\frac{|\widehat{I}(\mathbf{x})|^{1/(2p)}}{2\pi} \right) - 2 \log \pi(\widehat{\boldsymbol{\theta}})$$

は $\log n$ に比較して無視できるので,

$$\text{BIC} := -2 \log f(\mathbf{x} | \widehat{\boldsymbol{\theta}}) + p \log n$$

なる近似式を得る. これを Schwarz の Bayes 情報量規準という.
線型回帰モデルに適用してみると

$$\text{BIC} = -2 \log f(\mathbf{y} | \mathbf{X}\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2) + (d+2) \log n$$

を得る.

注意 1.8 AIC と比較すると, モデルの複雑さへの罰則項が異なる. すなわち, AIC は $2(d+2)$ で, BIC は $(d+2) \log n$ である. これらの違いから以下の性質が知られている. BIC は真のモデルの選択に対する一致性を持つ. 一方, AIC はその性質を持たない. しかし, AIC は予測誤差を最小にするモデルを選択するが, BIC はそのような性質を持たない. \square

4 補遺 : Laplace 近似について

命題 1.9 連続関数 $h, g: \mathbb{R}^d \rightarrow \mathbb{R}$ と $\mathbf{x}_0 \in \mathbb{R}^d$ に以下の条件を仮定する.

(1) h は \mathbf{x}_0 において最小値を取り,

$$h(\mathbf{x}) > h(\mathbf{x}_0) \quad (\mathbf{x} \neq \mathbf{x}_0).$$

(2) h は \mathbf{x}_0 において, 2 回連続微分可能で,

$$\begin{aligned} \frac{\partial}{\partial \mathbf{x}} f(\mathbf{x}) \Big|_{\mathbf{x}=\mathbf{x}_0} &= \mathbf{0}, \\ I(\mathbf{x}_0) &= \left(\frac{\partial^2}{\partial x_j \partial x_k} h(\mathbf{x}) \right)_{1 \leq j, k \leq d} \Big|_{\mathbf{x}=\mathbf{x}_0} \quad \text{は正定値,} \\ \mathbf{x} &= (x_1, x_2, \dots, x_d)^\top. \end{aligned}$$

(3) ある $\delta > 0$ と $\epsilon > 0$ が存在して,

$$\lim_{\lambda \rightarrow \infty} e^{\lambda h(\mathbf{x}_0) - \epsilon} g(\mathbf{x}_0) \int_{|\mathbf{x} - \mathbf{x}_0|_2 > \lambda} \exp\{-\lambda h(\mathbf{x})\} g(\mathbf{x}) d\mathbf{x} = 0.$$

このとき,

$$\lim_{\lambda \rightarrow \infty} e^{\lambda h(\mathbf{x}_0)} \left(\frac{\lambda}{2\pi} \right)^{d/2} \int_{\mathbb{R}^d} e^{-\lambda h(\mathbf{x})} g(\mathbf{x}) d\mathbf{x} = g(\mathbf{x}_0) |I(\mathbf{x}_0)|^{1/2}$$

が成立する.

証明 まず,

$$h(\mathbf{x}) - h(\mathbf{x}_0) = \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^\top I(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)(1 + o(1)) \quad (\mathbf{x} \rightarrow \mathbf{x}_0).$$

これと g の連続性より, $\forall \epsilon > 0$ に対して, $\delta > 0$ がうまく取れて, $|\mathbf{x} - \mathbf{x}_0|_2 \leq \delta$ ならば,

$$\begin{aligned} (1 - \epsilon) \left\{ \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^\top I(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) \right\} &\leq h(\mathbf{x}) - h(\mathbf{x}_0) \\ &\leq (1 + \epsilon) \left\{ \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^\top I(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) \right\}, \\ (1 - \epsilon)g(\mathbf{x}_0) &\leq g(\mathbf{x}) \leq (1 + \epsilon)g(\mathbf{x}_0) \end{aligned}$$

となる. これから

$$\begin{aligned} &\lambda^{d/2} \int_{|\mathbf{x} - \mathbf{x}_0|_2 \leq \delta} \exp\left\{-\frac{\lambda}{2}(1 \pm \epsilon)(\mathbf{x} - \mathbf{x}_0)^\top I(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)\right\} (1 \pm \epsilon)g(\mathbf{x}_0) d\mathbf{x} \\ &= \int_{|\mathbf{y}|_2 \leq \sqrt{\lambda}\delta} \exp\left\{-\frac{1}{2}(1 \pm \epsilon)\mathbf{y}^\top I(\mathbf{x}_0)\mathbf{y}\right\} (1 \pm \epsilon)g(\mathbf{x}_0) d\mathbf{y} \\ &\longrightarrow (1 \pm \epsilon)^{-d/2+1} (2\pi)^{d/2} \sqrt{|I(\mathbf{x}_0)|} g(\mathbf{x}_0) \quad (\lambda \rightarrow \infty). \end{aligned}$$

よって,

$$\begin{aligned} &(1 - \epsilon)^{-d/2+1} g(\mathbf{x}_0) (2\pi)^{d/2} \sqrt{|I(\mathbf{x}_0)|} \\ &\leq \liminf_{\lambda \rightarrow \infty} \lambda^{d/2} \int_{|\mathbf{x} - \mathbf{x}_0|_2 \leq \delta} \exp\{-\lambda h(\mathbf{x})\} g(\mathbf{x}) d\mathbf{x} \\ &\leq \limsup_{\lambda \rightarrow \infty} \lambda^{d/2} \int_{|\mathbf{x} - \mathbf{x}_0|_2 \leq \delta} \exp\{-\lambda h(\mathbf{x})\} g(\mathbf{x}) d\mathbf{x} \\ &\leq (1 + \epsilon)^{-d/2+1} g(\mathbf{x}_0) (2\pi)^{d/2} \sqrt{|I(\mathbf{x}_0)|}. \end{aligned} \tag{1.17}$$

次に、容易にわかるように、

$$\int_{|\mathbf{x}-\mathbf{x}_0|_2>\delta} \exp\{-\lambda h(\mathbf{x})\} g(\mathbf{x}) d\mathbf{x} = 0. \quad (1.18)$$

よって、 \mathbf{x}_0 の近くの評価 (1.17) と遠方での評価 (1.18) と $\epsilon >$ は任意の取ったことより、

$$\lim_{\lambda \rightarrow \infty} \lambda^{d/2} \int_{\mathbb{R}^d} \exp\{-\lambda\{h(\mathbf{x}) - h(\mathbf{x}_0)\}\} g(\mathbf{x}) d\mathbf{x} = g(\mathbf{x}_0)(2\pi)^{d/2}|I(\mathbf{x}_0)|^{1/2}.$$

したがって、

$$\lim_{\lambda \rightarrow \infty} e^{\lambda h(\mathbf{x}_0)} \left(\frac{\lambda}{2\pi}\right)^{d/2} \int_{\mathbb{R}^d} \exp\{-\lambda h(\mathbf{x})\} g(\mathbf{x}) d\mathbf{x} = g(\mathbf{x}_0)|I(\mathbf{x}_0)|^{1/2}.$$

□