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

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

とする. 各 y_j $(j = 1, 2, \ldots, n)$ は

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

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. $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ 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 $x_1, x_2, \ldots, x_n$ は確率変数ではなく, 与えられた定数.
(2) $\mathbb{E}[\epsilon_j] = 0$ $(j = 1, 2, \ldots, n)$.
(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,

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

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,

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

と書ける.

$$\widehat{\alpha} := \overline{y} - \widehat{\beta}\overline{x}; \qquad \widehat{\beta} := \frac{Q_{xy}}{Q_{xx}}$$

とおく. 上の式より, $h(\alpha, \beta)$ は $\alpha = \widehat{\alpha}$, $\beta = \widehat{\beta}$ 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 $y = \widehat{\alpha} + \widehat{\beta}x$ を引くことができる. 回帰直線上の点 $(x_j, \widehat{\alpha} + \widehat{\beta}x_j)\,(j = 1, 2, \ldots, n)$ と観測 (x_j, y_j) との差

$$e_j := y_j - (\widehat{\alpha} + \widehat{\beta}x_j) \qquad (j = 1, 2, \ldots, n)$$

を**残差**といい,

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

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

$$\widehat{\sigma}^2 = \frac{1}{n-2}\mathsf{RSS}$$

で推定することができる.

命題 1.1 (1) $\mathbb{E}[\widehat{\beta}] = \beta^*$; $\mathbb{V}[\widehat{\beta}] = \dfrac{\sigma^2}{Q_{xx}}$.
(2) $\mathbb{E}[\widehat{\alpha}] = \alpha^*$; $\mathbb{V}[\widehat{\alpha}] = \sigma^2\left\{\dfrac{1}{n} + \dfrac{\overline{x}^2}{Q_{xx}}\right\}$.
(3) $\mathbb{COV}[\widehat{\alpha}, \widehat{\beta}] = -\dfrac{\overline{x}\sigma^2}{Q_{xx}}$.
(4) $\mathbb{E}[e_j] = 0$; $\mathbb{V}[e_j] = \sigma^2\,(j = 1, 2, \ldots, n)$.

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

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

と表現できることを示す. ただし $a_j, b_j\,(j = 1, 2, \ldots, n)$ は定数である. $\mathbb{E}[\epsilon_j] = 0$, $\mathbb{V}[\epsilon_j] = \sigma^2$ と $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ 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

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

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 $e_j\,(j = 1, 2, \ldots, n)$ を

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

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 $\epsilon_j\,(j = 1, 2, \ldots, n)$ について

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

とおく. すると

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

となる. これより

$$\widehat{\beta} - \beta^* = \frac{\sum_{j=1}^{n}(x_j - \overline{x})\epsilon_j}{Q_{xx}} \tag{1.2}$$

と書ける. (1.2) より

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

(2) の証明: また,

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

より

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

(3) の証明: 同様にして,

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

(4) の証明:

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

と表さるので, $\mathbb{E}[e_j] = 0$. また,

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

と書ける. (1.2) をから,

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

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) $$\begin{pmatrix}\widehat{\alpha}\\ \widehat{\beta}\end{pmatrix} \sim N_2\left(\begin{pmatrix}\alpha^*\\ \beta^*\end{pmatrix}, \frac{\sigma^2}{Q_{xx}}\begin{pmatrix}\frac{Q_{xx}}{n}+\overline{x}^2 & -\overline{x}\\ -\overline{x} & 1\end{pmatrix}\right),$$

(2) $$\frac{(n-2)\widehat{\sigma}^2}{\sigma^2} = \frac{\mathsf{RSS}}{\sigma^2} \sim \chi^2_{n-2},$$

(3) $\widehat{\sigma}^2$ と $(\widehat{\alpha}, \widehat{\beta})$ 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. $\mathbb{COV}[e_j, \widehat{\alpha}] = \mathbb{COV}[e_j, \widehat{\beta}] = 0$ より, $\widehat{\sigma}^2$ と $(\widehat{\alpha}, \widehat{\beta})$ 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

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

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

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

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

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

と表される. これは

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon} \tag{1.6}$$

と書ける.

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

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

を最小化することにより得られる. 関数 h は

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

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, $\boldsymbol{X}^\top\boldsymbol{X}$ は正則と仮定する. ここで,

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$$

とおくと

$$h(\boldsymbol{\beta}) \quad = \quad (\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})^\top(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}) + (\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top\boldsymbol{X}^\top\boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \tag{1.7}$$

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

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\epsilon}$$

と書けるので,

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

さらに,

$$\mathsf{RSS} := (\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})^\top(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \boldsymbol{y}^\top\left\{\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\right\}\boldsymbol{y}$$

とする. トレース, 期待値およびベキ等行列の性質より

$$\begin{aligned}\mathbb{E}[\mathsf{RSS}] &= \mathrm{Tr}\left[\left\{\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\right\}\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top]\right] &(1.8)\\ &= \sigma^2\mathrm{Tr}\left[\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\right] &(1.9)\\ &= (n-d-1)\sigma^2. &(1.10)\end{aligned}$$

よって, $n \geq d+2$ のとき, σ^2 の不偏推定量は

$$\widehat{\sigma}^2 = \frac{1}{n-d-1}\mathsf{RSS}.$$

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

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

$$\mathbb{E}[\boldsymbol{Cy}] = \boldsymbol{\beta}^* \Longleftrightarrow \boldsymbol{CX} = \boldsymbol{I}_{d+1}$$

をみたさなければならない. さらに,

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

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

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

しかし

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

よって

$$\begin{aligned}\mathbb{COV}[\boldsymbol{Cy}] &= \sigma^2\boldsymbol{CC}^\top = \sigma^2\boldsymbol{C}^*(\boldsymbol{C}^*)^\top + \sigma^2(\boldsymbol{C}-\boldsymbol{C}^*)(\boldsymbol{C}-\boldsymbol{C}^*)^\top \\ &\gtrsim \sigma^2\boldsymbol{C}^*(\boldsymbol{C}^*)^\top = \mathbb{COV}[\widehat{\boldsymbol{\beta}}].\end{aligned}$$

ただし, 正方行列 A, B に対して,

$$\boldsymbol{A} \gtrsim \boldsymbol{B} \Longleftrightarrow \boldsymbol{A}-\boldsymbol{B} \gtrsim \boldsymbol{0} \Longleftrightarrow \boldsymbol{A}-\boldsymbol{B} \text{ は非負定値}$$

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 $\gtrsim$ 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. □

命題 1.4 $n \geq d+2$ とする. $\boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$ を仮定する. このとき, 以下のことが成り立つ.
(1) $\widehat{\boldsymbol{\beta}} \sim N_{d+1}(\boldsymbol{\beta}^*, \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1})$.
(2) $\dfrac{(n-d-1)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-d-1}$.
(3) $\widehat{\boldsymbol{\beta}}$ と $\widehat{\sigma}^2$ は独立.

証明　$\boldsymbol{z} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*)/\sigma$ とし,

$$U = (\boldsymbol{X}^\top \boldsymbol{X})^{-1/2}\boldsymbol{X}^\top \boldsymbol{z}; \qquad V = \boldsymbol{z}^\top \{\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top\}\boldsymbol{z}$$

とおく. ただし, 正則な行列 $\boldsymbol{A}$ に対して, $\boldsymbol{A}^{-1} = \boldsymbol{B}\boldsymbol{B}^\top$ をみたす行列 $\boldsymbol{B}$ を $\boldsymbol{A}^{-1/2}$ [1]と記した. すると, 命題の主張は
$(1a)$ $U \sim N_{d+1}(\boldsymbol{0}, \boldsymbol{I}_{d+1})$,
$(2a)$ $V \sim \chi^2_{n-d-1}$,
$(3a)$ U と V は独立,
という形に簡略化される. $\boldsymbol{X}$ は $n \times (d+1)$ 行列で, ランクが $d+1$(フルランク) だから

$$\boldsymbol{X} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{O}^\top$$

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

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

より

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

これらより

$$U = \boldsymbol{O}\boldsymbol{P}^\top \boldsymbol{z}; \qquad V = \boldsymbol{z}^\top(\boldsymbol{I}_n - \boldsymbol{P}\boldsymbol{P}^\top)\boldsymbol{z}$$

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

$$V = \boldsymbol{z}^\top\boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{z} = (\boldsymbol{Q}^\top\boldsymbol{z})^\top(\boldsymbol{Q}^\top\boldsymbol{z})$$

と表される.

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

であるから, $U = \boldsymbol{P}^\top\boldsymbol{z}$ と $V = \boldsymbol{z}^\top\boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{z}$ は独立で, $U \sim N_{d+1}(\boldsymbol{0}, \boldsymbol{I}_{d+1})$, $V \sim \chi^2_{n-d-1}$ 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 $x_{1j}, x_{2j}, \ldots, x_{dj}$ の個数 d を増やすにつれて, 線型モデル

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

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,

$$\mathbb{E}[\mathrm{Tr}\,\{(\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}^*)^\top (\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}^*)\}] \leq \boldsymbol{C}\frac{\sigma^2 d}{n}$$

と評価できる. ただし, $\boldsymbol{C}$ は n, d に依らない定数である.

以下では, $n \geq d+2$ とし, $\boldsymbol{X}$ 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

$$\mathsf{R}_d^2 = 1 - \frac{\sum_{j=1}^n (y_j - \widehat{y_j})^2}{\sum_{j=1}^n (y_j - \overline{y})^2}; \qquad \widehat{y_j} = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1j} + \widehat{\beta}_2 x_{2j} + \cdots + \widehat{\beta}_d x_{dj}$$

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, $\sum_{j=1}^n (y_j - \widehat{y}_j)^2$ と $\sum_{j=1}^n (y_j - \overline{y})^2$ 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

$$\mathsf{R}_d^{*2} = 1 - \frac{\sum_{j=1}^n (y_j - \widehat{y}_j)^2/(n-d-1)}{\sum_{j=1}^n (y_j - \overline{y})^2/(n-1)}$$

を自由度調整済み決定係数という. これを書き直せば

$$\mathsf{R}_d^{2*} = 1 - \frac{n-1}{n-d-1}(1 - \mathsf{R}_d^2)$$

となる. この形からわかるように, d が大きくなると, $1 - \mathsf{R}_d^2$ 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 $\boldsymbol{\epsilon}$ とは独立な確率ベクトル $\tilde{\boldsymbol{\epsilon}} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$ に基づく将来の観測を

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

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

$$\begin{aligned}
\mathsf{MSE}(\widehat{\boldsymbol{y}}) &= \mathbb{E}\big[(\tilde{\boldsymbol{y}} - \widehat{\boldsymbol{y}})^\top (\tilde{\boldsymbol{y}} - \widehat{\boldsymbol{y}}\big] \\
&= \mathbb{E}\big[\{\tilde{\boldsymbol{\epsilon}} - \boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\}^\top \{\tilde{\boldsymbol{\epsilon}} - \boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\}\big] \\
&= n\sigma^2 + \mathbb{E}\big[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^\top \boldsymbol{X}^\top \boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\big] \\
&= n\sigma^2 + \mathrm{Tr}\left\{\boldsymbol{X}^\top \boldsymbol{X}\mathbb{E}[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^\top]\right\} \\
&= n\sigma^2 + (d+1)\sigma^2 = (n+d+1)\sigma^2
\end{aligned}$$

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

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

となるので, 予測誤差は

$$\mathsf{MSE}(\widehat{\boldsymbol{y}}) = \mathbb{E}\big[\mathsf{RSS}_d + 2(d+1)\sigma^2\big]$$

となるので, σ^2 が既知の場合, $\mathsf{RSS}_d + 2(d+1)\sigma^2$ 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

$$(x_{j1}, x_{j2}, \ldots, x_{jd}, \ldots, x_{jK}) \qquad (1 \le d \le K)$$

とする.

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

としたとき, 分散 σ^2 の推定量を

$$\widehat{\sigma}_F^2 = \frac{\boldsymbol{y}^\top \{\boldsymbol{I}_n - \boldsymbol{X}_F(\boldsymbol{X}_F^\top \boldsymbol{X}_F)^{-1}\boldsymbol{X}_F^\top\}\boldsymbol{y}}{n-K-1}$$

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

$$\mathsf{C}_p := \frac{\mathsf{RSS}_d}{\widehat{\sigma}_F^2} + 2(d+1)$$

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

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

で定義する.

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

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

□

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

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

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

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

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

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

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.

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

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

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

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, $\mathrm{AI}(\boldsymbol{\beta}^*,\,\sigma^2)$ を計算してもると

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

であることと命題 1.4 に注意し, $\tilde{\boldsymbol{y}}$ に関して積分[2]する.

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

[2]実際には, $\tilde{\boldsymbol{\epsilon}}\sim N_n(\boldsymbol{0},\,\sigma\boldsymbol{I}_n)$ 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

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

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

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

と書けることがわかる.

次に, 対数尤度関数を計算する.

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

に注意して,

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

となる. これを (1.11) に代入すれば,

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

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 $\mathsf{AI}(\boldsymbol{\beta}^*,\,\sigma^2)$ の不偏推定量になる.

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

なので, 近似推定量として,

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

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 $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)$ から j 番目 $(j = 1, 2, \ldots, n)$ のデータ $(\boldsymbol{x}_j, y_j)$ を除いた残りのデータからの $\boldsymbol{\beta}^*$ の推定量を考える.

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

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

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

を計算する. この操作を繰り返し, 予測誤差

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

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.

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

証明　命題の主張を証明するために,

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

を示せばよい.

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

に注意する. 等式

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

より

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

を得る. これらより

$$
\begin{aligned}
\boldsymbol{x}_1^\top\widehat{\boldsymbol{\beta}^*} &= \boldsymbol{x}_1^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}\\
&= \boldsymbol{x}_1^\top\left\{A^{-1}-\frac{A^{-1}\boldsymbol{x}_1\boldsymbol{x}_1^\top A^{-1}}{1+\boldsymbol{x}_1^\top A^{-1}\boldsymbol{x}_1}\right\}(\boldsymbol{x}_1;\,(\boldsymbol{X}^{(1)})^\top)\begin{pmatrix} y_1\\ \boldsymbol{y}^{(1)}\end{pmatrix}\\
&= \boldsymbol{x}_1^\top\left\{A^{-1}-\frac{A^{-1}\boldsymbol{x}_1\boldsymbol{x}_1^\top A^{-1}}{1+\boldsymbol{x}_1^\top A^{-1}\boldsymbol{x}_1}\right\}\{\boldsymbol{x}_1y_1+(\boldsymbol{X}^{(1)})^\top\boldsymbol{y}^{(1)}\}\\
&= \boldsymbol{x}_1\Bigg\{A^{-1}\boldsymbol{x}_1y_1-\frac{A^{-1}\boldsymbol{x}_1\boldsymbol{x}_1^\top A^{-1}\boldsymbol{x}_1y_1}{1+\boldsymbol{x}_1A^{-1}\boldsymbol{x}_1}+A^{-1}(\boldsymbol{X}^{(1)})^\top\boldsymbol{y}^{(1)}\\
&\qquad -\frac{A^{-1}\boldsymbol{x}_1\boldsymbol{x}_1^\top A^{-1}(\boldsymbol{X}^{(1)})^\top\boldsymbol{y}^{(1)}}{1+\boldsymbol{x}_1A^{-1}\boldsymbol{x}_1}\Bigg\}\\
&= \boldsymbol{x}_1^\top A^{-1}\boldsymbol{x}_1y_1-\frac{(\boldsymbol{x}_1^\top A^{-1}\boldsymbol{x}_1)^2y_1}{1+\boldsymbol{x}_1^\top A^{-1}\boldsymbol{x}_1}+\boldsymbol{x}_1^\top A^{-1}(\boldsymbol{X}^{(1)})^\top\boldsymbol{y}^{(1)}\\
&\qquad -\frac{\boldsymbol{x}_1^\top A^{-1}\boldsymbol{x}_1\boldsymbol{x}_1^\top A^{-1}(\boldsymbol{X}^{(1)})^\top\boldsymbol{y}^{(1)}}{1+\boldsymbol{x}_1^\top A^{-1}\boldsymbol{x}_1}\\
&= \boldsymbol{x}_1^\top A^{-1}\boldsymbol{x}_1y_1-\frac{(\boldsymbol{x}_1^\top A^{-1}\boldsymbol{x}_1)^2y_1}{1+\boldsymbol{x}_1^\top A^{-1}\boldsymbol{x}_1}+\boldsymbol{x}_1^\top\widehat{\boldsymbol{\beta}^*}^{(1)}-\frac{\boldsymbol{x}_1^\top A^{-1}\boldsymbol{x}_1\boldsymbol{x}_1^\top\widehat{\boldsymbol{\beta}^*}^{(1)}}{1+\boldsymbol{x}_1^\top A^{-1}\boldsymbol{x}_1}\\
&= \frac{\boldsymbol{x}_1^\top A^{-1}\boldsymbol{x}_1y_1+\boldsymbol{x}_1^\top\widehat{\boldsymbol{\beta}^*}^{(1)}}{1+\boldsymbol{x}_1^\top A^{-1}\boldsymbol{x}_1} \qquad (1.15)
\end{aligned}
$$

を得る. (1.14) と (1.15) より

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

を得る. □

3.5 BIC

線型回帰モデル

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

において, $(d+2)$ 個の未知母数 $\boldsymbol{\beta}^*$, σ^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 $\boldsymbol{y}$ の周辺分布

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

で与えられる. これは σ^2 と $(d+1)$ 個の回帰係数 $\beta_0^*, \beta_1^*, \ldots, \beta_d^*$ 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 $-2\log f_{\pi_{d+2}}(\boldsymbol{y})$ 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

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

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

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

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

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

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, $n \to \infty$ 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, $\boldsymbol{\theta} \in \Theta$ を p 次元の未知母数とし, $\boldsymbol{W}$ を n 次元確率変数とする. ただし $\Theta \subset \mathbb{R}^p$ を母数空間とする. $\boldsymbol{\theta}$ を与えたときの $\boldsymbol{W}$ の条件付き確率密度関数を

$$\boldsymbol{W}|\,\boldsymbol{\theta} \sim\ f(\boldsymbol{w}|\,\boldsymbol{\theta})$$

とし, $\boldsymbol{\theta}$ の事前分布を

$$\boldsymbol{\theta}\ \sim\ \pi(\boldsymbol{\theta})$$

とする. 対数尤度

$$\ell(\boldsymbol{\theta}|\,\boldsymbol{w}) = \log f(\boldsymbol{w}|\,\boldsymbol{\theta})$$

を $\boldsymbol{\theta}$ の最尤推定量[3]$\widehat{\boldsymbol{\theta}}$ のまわりで Taylor 展開する.

$$\begin{aligned}\ell(\boldsymbol{\theta}|\,\boldsymbol{w}) \quad &\approx \quad \ell(\widehat{\boldsymbol{\theta}}|\,\boldsymbol{w}) + \left(\frac{\partial}{\partial\boldsymbol{\theta}}\ell(\boldsymbol{\theta}|\,\boldsymbol{w})\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}\right)^{\top}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}) \\ &\qquad +\frac{1}{2}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})^{\top}\left(\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\top}}\ell(\boldsymbol{\theta}|\,\boldsymbol{w})\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}\right)(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})\end{aligned}$$

[3]すなわち, 存在すれば,

$$\widehat{\theta} \in \mathrm{argmax}_{\theta\in\Theta}\ell(\boldsymbol{\theta}|\,\boldsymbol{w})$$

で定義する.

と近似できる. ただし

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

である. さらに

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

とおいたとき, $\widehat{I}(\boldsymbol{X})$ 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

$$\frac{\partial}{\partial\boldsymbol{\theta}}\ell(\boldsymbol{\theta}|\,\boldsymbol{w})\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = \boldsymbol{0}$$

である. よって

$$\ell(\boldsymbol{\theta}|\,\boldsymbol{w}) \;\approx\; \ell(\widehat{\boldsymbol{\theta}}|\,\boldsymbol{w}) - \frac{n}{2}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})^\top\widehat{I}(\boldsymbol{w})(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})$$

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

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

と近似できると仮定すると, Bayes 的周辺分布は

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

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「≈」の意味は以下である．$a \in \mathbb{R}$ と $\{a_n\}_{n=1}^{\infty}$ に対して，

$$a \approx a_n \Longleftrightarrow \lim_{n\to\infty} \frac{a}{a_n} = 1$$

である．したがって

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

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

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

は $\log n$ に比較して無視できるので

$$\mathsf{BIC} := -2\log f(\boldsymbol{w}|\,\widehat{\boldsymbol{\theta}}) + p\log n$$

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

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

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らている．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． □