

# 第1章 線型モデル

この章では、線型回帰モデルについて簡単に説明する。第 1.1 では、単回帰モデルにおける最小 2 乗推定量について簡単に説明する。第 1.2 では重回帰モデルを扱い、前節の結果を拡張する。第 1.3 では重回帰モデルにおける回帰係数の選択問題について概説する。自由度調整済み寄与率, Mallows  $C_p$ , 交差検証法, AIC, BIC について解説をしていく。第 1.4 節の補遺では、BIC の導出に用いた Laplace 近似の正当化の議論をまとめた。

## 1.1 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誤差項である。

この節において単回帰モデルの誤差項に以下の仮定 (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とする。
- (2)  $E[\epsilon_j] = 0$  ( $j = 1, 2, \dots, n$ ).
- (3)  $E[\epsilon_j \epsilon_\ell] = 0$  ( $j \neq \ell$ ).
- (4)  $\text{Var}[\epsilon_j] = \sigma^2$ . ただし分散  $\sigma^2$  ( $\sigma > 0$ ) は未知とする。

未知の母数  $\alpha^*, \beta^*$  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.  $\alpha, \beta$  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

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

とおく.

以下では  $Q_{xx} \neq 0$  と仮定する. 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. ここで

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

とおいた. (1.1) より  $h(\alpha, \beta)$  は  $\alpha = \hat{\alpha}, \beta = \hat{\beta}$  のときに最小値を取る.

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

$$e_j := y_j - (\hat{\alpha} + \hat{\beta}x_j) \quad (j = 1, 2, \dots, n)\tag{1.2}$$

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. さらに,  $n \geq 3$  のとき未知の分散  $\sigma^2$  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.  $n \geq 3$  とする. 最小 2 乗推定量  $(\hat{\alpha}, \hat{\beta})$  と  $\hat{\sigma}^2$  の積率について, 以下が成立する.

- (1)  $E[\hat{\beta}] = \beta^*$ ;  $\text{Var}[\hat{\beta}] = \frac{\sigma^2}{Q_{xx}}$ .
- (2)  $E[\hat{\alpha}] = \alpha^*$ ;  $\text{Var}[\hat{\alpha}] = \sigma^2 \left\{ \frac{1}{n} + \frac{\bar{x}^2}{Q_{xx}} \right\}$ .
- (3)  $\text{Cov}[\hat{\alpha}, \hat{\beta}] = -\frac{\bar{x}\sigma^2}{Q_{xx}}$ .
- (4)  $E[e_j] = 0$ ;  $\sum_{j=1}^n \text{Var}[e_j] = (n-2)\sigma^2$  ( $j = 1, 2, \dots, n$ ).

*Proof.* (1) の証明: 誤差項  $\epsilon_j$  ( $j = 1, 2, \dots, n$ ) に対して

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

とおく. すると

$$\begin{aligned} \epsilon_j - \bar{\epsilon} &= y_j + \alpha^* + \beta^* x_j - \{\bar{y} + \alpha^* + \beta^* \bar{x}\} \\ &= y_j - \bar{y} - \beta^*(x_j - \bar{x}), \end{aligned} \quad (1.3)$$

$$\sum_{j=1}^n (x_j - \bar{x}) \bar{\epsilon} = 0 \quad (1.4)$$

となる. (1.3) の両辺に  $x_j - \bar{x}$  をかけて  $j$  について和を取ると

$$\begin{aligned} \sum_{j=1}^n (x_j - \bar{x})(\epsilon_j - \bar{\epsilon}) &= \sum_{j=1}^n (x_j - \bar{x})(y_j - \bar{y}) - \beta^* \sum_{j=1}^n (x_j - \bar{x})^2 = Q_{xy} - \beta^* Q_{xx} \\ &= Q_{xx} \{\hat{\beta} - \beta^*\} \end{aligned}$$

を得る. この式の右辺に (1.4) を代入すると

$$\hat{\beta} - \beta^* = \frac{\sum_{j=1}^n (x_j - \bar{x}) \epsilon_j}{Q_{xx}} \quad (1.5)$$

と得る. (1.5) と  $E[\epsilon_j] = 0$  に注意すると

$$E[\hat{\beta} - \beta^*] = 0 \quad (1.6)$$

を得る. さらに,  $\{\epsilon_j\}_{j=1}^n$  は互いに独立で  $\text{Var}[\epsilon_j] = \sigma^2$  ( $j = 1, 2, \dots, n$ ) に注意すると

$$\text{Var}[\hat{\beta}] = E[(\hat{\beta} - \beta^*)^2] = \frac{\sum_{j=1}^n (x_j - \bar{x})^2 E[\epsilon_j^2]}{Q_{xx}^2} = \frac{\sigma^2}{Q_{xx}} \quad (1.7)$$

がわかる.

(2) の証明: また,  $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$  と (1.5) から

$$\begin{aligned} \hat{\alpha} - \alpha^* &= \bar{y} - \hat{\beta}\bar{x} - \alpha^* - \hat{\beta}\bar{x} + \alpha^* = \alpha^* + \beta^*\bar{x} + \bar{\epsilon} \\ &= \bar{\epsilon} - \bar{x}(\hat{\beta} - \beta^*) = \sum_{j=1}^n \left\{ \frac{1}{n} - \frac{(x_j - \bar{x})\bar{x}}{Q_{xx}} \right\} \epsilon_j \end{aligned} \quad (1.8)$$

がわかる. 再度  $E[\epsilon_j] = 0$  ( $j = 1, 2, \dots, n$ ) と (1.8) から

$$E[\hat{\alpha} - \alpha^*] = 0$$

を得る. さらに,  $\{\epsilon_j\}_{j=1}^n$  は互いに独立で  $\text{Var}[\epsilon_j] = \sigma^2$  ( $j = 1, 2, \dots, n$ ) から

$$\begin{aligned} \text{Var}[\hat{\alpha}] &= E[(\hat{\alpha} - \alpha^*)^2] = \sum_{j=1}^n \left\{ \frac{1}{n} - \frac{(x_j - \bar{x})\bar{x}}{Q_{xx}} \right\}^2 E[\epsilon_j^2] \\ &= \sigma^2 \left\{ \frac{1}{n} + \frac{\bar{x}^2}{Q_{xx}} \right\} \end{aligned} \quad (1.9)$$

がわかる.

(3) の証明: 同様に, (1.5) と (1.8) から

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

がわかる.

(4) の証明: (1.2) から

$$\begin{aligned} e_j &= y_j - \{\hat{\alpha} + \hat{\beta}x_j\} = y_j - \{\bar{y} - \hat{\beta}\bar{x} + \hat{\beta}x_j\} \\ &= y_j - \bar{y} - \hat{\beta}(x_j - \bar{x}) = (\epsilon_j - \bar{\epsilon}) - (x_j - \bar{x})(\hat{\beta} - \beta^*) \end{aligned} \quad (1.10)$$

と表さる. よって,  $E[\epsilon_j] = 0$  ( $j = 1, 2, \dots, n$ ),  $E[\bar{\epsilon}] = 0$  と (1.7) から

$$E[e_j] = 0$$

がわかる. つぎに, (1.10) の最右辺を展開して,  $j$  について和と期待値を取ると

$$\begin{aligned} \sum_{j=1}^n E[e_j^2] &= \sum_{j=1}^n E[(\epsilon_j - \bar{\epsilon})^2] - 2 \sum_{j=1}^n (x_j - \bar{x}) E[(\epsilon_j - \bar{\epsilon})(\hat{\beta} - \beta^*)] \\ &\quad + \sum_{j=1}^n (x_j - \bar{x})^2 E[(\hat{\beta} - \beta^*)^2] \end{aligned} \quad (1.11)$$

と書ける. (1.5) から

$$\begin{aligned}
 \sum_{j=1}^n E[(\epsilon_j - \bar{\epsilon})^2] &= \sum_{j=1}^n \epsilon_j^2 - nE[\bar{\epsilon}^2] = (n-1)\sigma^2, \\
 \sum_{j=1}^n (x_j - \bar{x})E[(\epsilon_j - \bar{\epsilon})(\hat{\beta} - \beta^*)] &= \sum_{j=1}^n (x_j - \bar{x})E[\epsilon_j(\hat{\beta} - \beta^*)] \\
 &\quad - \underbrace{\sum_{j=1}^n (x_j - \bar{x})E[\bar{\epsilon}(\hat{\beta} - \beta^*)]}_{=0} \\
 &= \sum_{j=1}^n (x_j - \bar{x})E[\epsilon_j(\hat{\beta} - \beta^*)] \\
 &= \sum_{j=1}^n (x_j - \bar{x})E\left[\frac{\epsilon_j \sum_{\ell=1}^n (x_\ell - \bar{x})\epsilon_\ell}{Q_{xx}}\right] \\
 &= \sigma^2
 \end{aligned}$$

となる. これらの 2 つの式と (1.7) を (1.11) に代入すれば

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

を得る. □

命題 1.2. 正規性を仮定すると以下のことがわかる.  $n \geq 3$  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.

*Proof.* (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  $(\hat{\alpha}, \hat{\beta})$  は独立であることがわかる. □

## 1.2 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

$$y_j = \beta_0^* + \beta_1^* x_{j1} + \beta_2^* x_{j2} + \cdots + \beta_d^* x_{jd} + \epsilon_j \quad (j = 1, 2, \dots, n) \quad (1.12)$$

なるモデルを仮定する. ただし  $\beta_0^*$  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  $\epsilon_1, \epsilon_2, \dots, \epsilon_n$  は独立同一分布に従う確率変数列 (誤差項) であり.  $\text{Var}[\epsilon_j] = \sigma^2$  ( $j = 1, 2, \dots, n$ ) とする. ただし  $\sigma^2$  ( $\sigma > 0$ ) は未知とする. (1.12) を重回帰モデルという.

重回帰モデル (1.12) 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

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon} \quad (1.13)$$

と書ける.

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

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

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

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

とおくと

$$h(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \mathbf{X}^\top \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \quad (1.14)$$

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

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\epsilon}; \quad E[\boldsymbol{\epsilon}] = \boldsymbol{\epsilon}_n; \quad \text{Var}[\boldsymbol{\epsilon}] = \sigma^2 \mathbf{I}_n$$

であること注意すると

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

となることがわかる.

さらに

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

とおく. トレース, 期待値およびベキ等行列の性質より

$$E[\text{RSS}] = \text{Tr} \left[ \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top\} E[\epsilon \epsilon^\top] \right] \quad (1.15)$$

$$= \sigma^2 \text{Tr} [\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top] \quad (1.16)$$

$$= (n - d - 1) \sigma^2 \quad (1.17)$$

となる. よって,  $n \geq d + 2$  のとき  $\sigma^2$  の不偏推定量は

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

がわかる.

**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. すなわち,  $\mathbf{C}$  は  $(d + 1) \times n$  の定数行列に対して

$$\text{Var}[\hat{\beta}] \preceq \text{Var}[\mathbf{C}\mathbf{y}]$$

となる. ただし,  $\mathbf{A} \preceq \mathbf{B} \Leftrightarrow \mathbf{A} - \mathbf{B}$  は半正定値 と定めた.

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

$$E[\mathbf{C}\mathbf{y}] = \beta^* \iff \mathbf{C}\mathbf{X} = \mathbf{I}_{d+1}$$

をみたさなければならない. さらに

$$\text{Var}[\mathbf{C}\mathbf{y}] = E[(\mathbf{C}\mathbf{y} - \beta^*)(\mathbf{C}\mathbf{y} - \beta^*)^\top] = \mathbf{C}E[\mathbf{u}\mathbf{u}^\top]\mathbf{C}^\top = \sigma^2 \mathbf{C}\mathbf{C}^\top$$

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

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

である. しかし

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

となる. よって

$$\begin{aligned} \text{Var}[\mathbf{C} \mathbf{y}] &= \sigma^2 \mathbf{C} \mathbf{C}^\top = \sigma^2 \mathbf{C}^* (\mathbf{C}^*)^\top + \sigma^2 (\mathbf{C} - \mathbf{C}^*) (\mathbf{C} - \mathbf{C}^*)^\top \\ &\succeq \sigma^2 \mathbf{C}^* (\mathbf{C}^*)^\top = \text{Var}[\hat{\boldsymbol{\beta}}] \end{aligned}$$

を得る. ただし正方行列  $\mathbf{A}, \mathbf{B}$  に対して

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

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  $\succeq$  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.  $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$  を仮定する. このとき以下のことが成り立つ.

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

*Proof.*  $\mathbf{z} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*)/\sigma$  とし

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

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

- (1a)  $\mathbf{U} \sim N_{d+1}(\mathbf{0}, \mathbf{I}_{d+1})$ ,
- (2a)  $\mathbf{V} \sim \chi_{n-d-1}^2$ ,
- (3a)  $\mathbf{U}$  と  $\mathbf{V}$  は独立

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

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

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

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

より

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

<sup>1</sup>これは一意ではないことに注意せよ.

となる. これらより

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

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

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

と表される.

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

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

### 1.3 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

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

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.

#### 1.3.1 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.

### 1.3.2 Mallows $C_p$

モデルの誤差項  $\epsilon$  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

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

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

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

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

$$E[\text{RSS}_d] = E[\mathbf{y}^\top \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top\} \mathbf{y}] = (n-d-1)\sigma^2 \quad (1.18)$$

となるので, 予測誤差は

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

となる. よって  $\sigma^2$  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

$$(x_{j1}, x_{j2}, \dots, x_{jd}, \dots, x_{jK}) \quad (1 \leq d \leq K)$$

とする. ここで

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

としたとき分散  $\sigma^2$  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.  $\sigma^2$  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.

問 1.1.  $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$  とする.  $\mathbf{X}^\top \mathbf{X}$  は正則とし,  $\mathbf{Q} := \mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$  とおく. このとき, 以下の問いに答えよ.

- (1)  $\mathbf{Q}^\top = \mathbf{Q}$  と  $\mathbf{Q}^2 = \mathbf{Q}$  を確認せよ.
- (2)  $E[\mathbf{Q}\mathbf{y}]$  を求めよ.
- (3) 式 (1.18) を証明せよ.

$\mathbf{y}$  の分布の仮定より  $E[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$  と  $\text{Var}[\mathbf{y}] = \sigma^2 \mathbf{I}_n$  となることと注意 ??(2) を使うとよい.

### 1.3.3 交差検証法

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

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

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

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

を計算する. この操作を繰り返し, 予測誤差

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

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 1.5. CV は以下のように表現できる.

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

*Proof.* 命題の主張を証明するために

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

を示せばよい.

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

に注意する. 等式

$$(\mathbf{X}^\top \mathbf{X})^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1} \mathbf{x}_1 \mathbf{x}_1^\top \mathbf{A}^{-1}}{1 + \mathbf{x}_1^\top \mathbf{A}^{-1} \mathbf{x}_1} \quad (1.19)$$

より

$$\begin{aligned} 1 - \mathbf{x}_1^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_1 &= 1 - \mathbf{x}_1^\top \mathbf{A}^{-1} \mathbf{x}_1 + \frac{\mathbf{x}_1^\top \mathbf{A}^{-1} \mathbf{x}_1 \mathbf{x}_1^\top \mathbf{A}^{-1} \mathbf{x}_1}{1 + \mathbf{x}_1^\top \mathbf{A}^{-1} \mathbf{x}_1} \\ &= \frac{1 - (\mathbf{x}_1^\top \mathbf{A}^{-1} \mathbf{x}_1)^2 + (\mathbf{x}_1^\top \mathbf{A}^{-1} \mathbf{x}_1)^2}{1 + \mathbf{x}_1^\top \mathbf{A}^{-1} \mathbf{x}_1} \\ &= \frac{1}{1 + \mathbf{x}_1^\top \mathbf{A}^{-1} \mathbf{x}_1} \end{aligned} \quad (1.20)$$

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\{ \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1} \mathbf{x}_1 \mathbf{x}_1^\top \mathbf{A}^{-1}}{1 + \mathbf{x}_1^\top \mathbf{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\{ \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1} \mathbf{x}_1 \mathbf{x}_1^\top \mathbf{A}^{-1}}{1 + \mathbf{x}_1^\top \mathbf{A}^{-1} \mathbf{x}_1} \right\} \{ \mathbf{x}_1 y_1 + (\mathbf{X}^{(1)})^\top \mathbf{y}^{(1)} \} \\
&= \mathbf{x}_1 \left\{ \mathbf{A}^{-1} \mathbf{x}_1 y_1 - \frac{\mathbf{A}^{-1} \mathbf{x}_1 \mathbf{x}_1^\top \mathbf{A}^{-1} \mathbf{x}_1 y_1}{1 + \mathbf{x}_1 \mathbf{A}^{-1} \mathbf{x}_1} + \mathbf{A}^{-1} (\mathbf{X}^{(1)})^\top \mathbf{y}^{(1)} \right. \\
&\quad \left. - \frac{\mathbf{A}^{-1} \mathbf{x}_1 \mathbf{x}_1^\top \mathbf{A}^{-1} (\mathbf{X}^{(1)})^\top \mathbf{y}^{(1)}}{1 + \mathbf{x}_1 \mathbf{A}^{-1} \mathbf{x}_1} \right\} \\
&= \mathbf{x}_1^\top \mathbf{A}^{-1} \mathbf{x}_1 y_1 - \frac{(\mathbf{x}_1^\top \mathbf{A}^{-1} \mathbf{x}_1)^2 y_1}{1 + \mathbf{x}_1^\top \mathbf{A}^{-1} \mathbf{x}_1} + \mathbf{x}_1^\top \mathbf{A}^{-1} (\mathbf{X}^{(1)})^\top \mathbf{y}^{(1)} \\
&\quad - \frac{\mathbf{x}_1^\top \mathbf{A}^{-1} \mathbf{x}_1 \mathbf{x}_1^\top \mathbf{A}^{-1} (\mathbf{X}^{(1)})^\top \mathbf{y}^{(1)}}{1 + \mathbf{x}_1^\top \mathbf{A}^{-1} \mathbf{x}_1} \\
&= \mathbf{x}_1^\top \mathbf{A}^{-1} \mathbf{x}_1 y_1 - \frac{(\mathbf{x}_1^\top \mathbf{A}^{-1} \mathbf{x}_1)^2 y_1}{1 + \mathbf{x}_1^\top \mathbf{A}^{-1} \mathbf{x}_1} + \mathbf{x}_1^\top \widehat{\boldsymbol{\beta}}^{(1)} - \frac{\mathbf{x}_1^\top \mathbf{A}^{-1} \mathbf{x}_1 \mathbf{x}_1^\top \widehat{\boldsymbol{\beta}}^{(1)}}{1 + \mathbf{x}_1^\top \mathbf{A}^{-1} \mathbf{x}_1} \\
&= \frac{\mathbf{x}_1^\top \mathbf{A}^{-1} \mathbf{x}_1 y_1 + \mathbf{x}_1^\top \widehat{\boldsymbol{\beta}}^{(1)}}{1 + \mathbf{x}_1^\top \mathbf{A}^{-1} \mathbf{x}_1} \tag{1.21}
\end{aligned}$$

を得る. (1.20) と (1.21) より

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

を得る. □

### 1.3.4 AIC

定義 1.6. (Kullback-Leibler 情報量)  $p, q$  を Lebesgue 測度  $m$  に関する確率密度関数とする. Kullback-Leibler 情報量を

$$\text{KL}(p, q) = \int \log \left( \frac{p}{q} \right) p \, dm$$

で定義する.

注意 1.7.  $x \log x \geq x - 1$  ( $x > 0$ ) に注意して

$$\begin{aligned} \text{KL}(p, q) &= \int \log \left( \frac{p}{q} \right) p \, d\mathbf{m} = \int \frac{p}{q} \log \left( \frac{p}{q} \right) q \, d\mathbf{m} \\ &\geq \int \left( \frac{p}{q} - 1 \right) q \, d\mathbf{m} = 0 \end{aligned}$$

となる. □

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

$$p(\cdot | \mathbf{X}\hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\sigma}^2(\mathbf{y}))$$

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

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

で測る. 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

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

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.

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

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

$$A(\boldsymbol{\beta}^*, \sigma^2) := -2E \left[ \int \cdots \int \log \{ p(t | \mathbf{X}\hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\sigma}^2(\mathbf{y})) \} p(\tilde{\mathbf{y}} | \mathbf{X}\boldsymbol{\beta}^*, \sigma^2) dm_n(t) \right]$$

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  $AI(\beta^*, \sigma^2)$  を計算してみると

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

であることと命題 1.4 に注意し  $t$  に関して積分<sup>2</sup>する. すると

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

と書けることがわかる. 命題 1.4 から

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

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

$$AI(\beta^*, \sigma^2) = E[n \log(2\pi\hat{\sigma}^2(\mathbf{y}))] + \frac{(n+d+1)(n-d-1)}{n-d-3} \quad (1.22)$$

と書けることがわかる.

次に, 期待最大対数尤度関数 (最大対数尤度の期待値) を計算する.

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

に注意して, 最大対数尤度の期待値を計算すると

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

となる. これを (1.22) に代入すれば

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

<sup>2</sup>実際には,  $s \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$  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  $Al(\beta^*, \sigma^2)$  の不偏推定量になる. 右辺の括弧のなかの第 2 項目について極限をとると

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

となる. したがって, 近似推定量として

$$AIC := -2 \log p(\mathbf{y} | \mathbf{X} \hat{\beta}(\mathbf{y}), \hat{\sigma}^2(\mathbf{y})) + 2(d+2) \quad (1.23)$$

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, (1.23) 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.

### 1.3.5 BIC

#### 線型回帰モデル

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

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

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

が成立していることである. ただし,  $m$  は  $\mathbb{R}$  上の Lebeague 測度である. すると  $\mathbf{y}$  の周辺分布

$$p_{\pi_{d+2}}(\mathbf{y}) = \int \cdots \int p(\mathbf{y} | \mathbf{X} \beta, \sigma^2) \pi_{d+2}(\beta, \sigma^2) dm_d(\beta) dm(\sigma^2)$$

で与えられる. これは  $\sigma^2$  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  $-2 \log p_{\pi_{d+2}}(\mathbf{y})$  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

$$y_j = \beta_0^* + \epsilon_j \quad (j = 1, 2, \dots, n) \quad (1.24)$$

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

$$p_{\pi_2}(\mathbf{y}) = \int \int p(\mathbf{y} | \beta_0, \sigma^2) \pi_2(\beta_0, \sigma^2) dm(\beta_0) dm(\sigma^2)$$

を考える. ただし,  $p(\mathbf{y}|\beta_0^*, \sigma^2)$  はモデル (1.24) の確率密度関数である. これらの比

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

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  $\theta$  を  $k$  次元の未知母数とし,  $X$  を  $n$  次元確率変数とし,  $\theta$  を与えたときの  $X$  の条件付き確率密度関数を

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

とし,  $\theta$  の事前分布を

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

とする. 対数尤度

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

を  $\theta$  の最尤推定量<sup>3</sup> $\hat{\theta}$  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

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

である. さらに

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

<sup>3</sup>すなわち, 存在すれば,

$$\hat{\theta} \in \arg \max_{\theta \in \Theta} \ell(\theta|x)$$

で定義する.

とおいたとき  $\widehat{\mathbf{I}}(X)$  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

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

である. よって

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

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

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

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

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

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

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

は  $\log n$  に比較して無視できるので,

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

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

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

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貫性を持つ. 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.  $\square$

## 1.4 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,

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

である.

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

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

となる.

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

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

となる. このとき

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

が成立する.

*Proof.* まず

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

に注意する. 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

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

となる. これから

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

となる. よって

$$\begin{aligned} &(1 - \epsilon)^{-d/2+1} g(\mathbf{x}_0) (2\pi)^{d/2} \sqrt{|\mathbf{I}(\mathbf{x}_0)|} \\ &\leq \liminf_{\lambda \rightarrow \infty} \lambda^{d/2} \int_{|\mathbf{x} - \mathbf{x}_0|_2 \leq \delta} \exp \{-\lambda h(\mathbf{x})\} g(\mathbf{x}) d\mathbf{m}_d(\mathbf{x}) \\ &\leq \limsup_{\lambda \rightarrow \infty} \lambda^{d/2} \int_{|\mathbf{x} - \mathbf{x}_0|_2 \leq \delta} \exp \{-\lambda h(\mathbf{x})\} g(\mathbf{x}) d\mathbf{m}_d(\mathbf{x}) \\ &\leq (1 + \epsilon)^{-d/2+1} g(\mathbf{x}_0) (2\pi)^{d/2} \sqrt{|\mathbf{I}(\mathbf{x}_0)|} \end{aligned} \quad (1.25)$$

がわかる. 次に, 容易にわかるように

$$\int_{|\mathbf{x} - \mathbf{x}_0|_2 > \delta} \exp \{-\lambda h(\mathbf{x})\} g(\mathbf{x}) d\mathbf{m}_d(\mathbf{x}) = 0 \quad (1.26)$$

となる. よって,  $\mathbf{x}_0$  の近くの評価 (1.25) と遠方での評価 (1.26) と  $\epsilon > 0$  は任意に取ったことより

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

となる. したがって

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

がわかる. □

## 1.5 章末注釈と参考文献

この章は [3] を参考にした.

## 1.6 演習問題

演習問題 1.1.

演習問題 1.1 の解答

□