第4章　線型回帰モデル

4.1　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 $x_1, x_2, \ldots, x_n$ は確率変数ではなく与えられた定数.
(2) $\mathsf{E}[\epsilon_j] = 0\,(j = 1, 2, \ldots, n)$.
(3) $\mathsf{E}[\epsilon_j \epsilon_\ell] = 0\,(j \neq \ell)$.
(4) $\mathsf{Var}[\epsilon_j] = \sigma^2$. ただし分散 $\sigma^2\,(\sigma > 0)$ は未知とする.

未知の母数 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,

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

なる記号を導入する. 以下では $Q_{xx} \neq 0$ と仮定する.

関数 h を変形すれば

$$\begin{aligned} h(\alpha, \beta) &= Q_{xx}\left\{\beta - \frac{Q_{xy}}{Q_{xx}}\right\}^2 + n\{\overline{y} - \alpha - \beta\overline{x}\}^2 \\ &\quad + \frac{Q_{xx}Q_{yy} - Q_{xy}^2}{Q_{xx}} \end{aligned} \tag{4.1}$$

eq:2-1

と書ける. ここで

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

とおく. 上の式より $h(\alpha, \beta)$ は $\alpha = \widehat{\alpha}$, $\beta = \widehat{\beta}$ のときに最小値を取る~~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

$$e_j := y_j - (\widehat{\alpha} + \widehat{\beta}x_j) \qquad (j = 1, 2, \ldots, n) \tag{4.2}$$

eq:2-1a

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.

pro:2-1

命題 4.1. (1) $\mathsf{E}[\widehat{\beta}] = \beta^*$; $\mathsf{Var}[\widehat{\beta}] = \dfrac{\sigma^2}{Q_{xx}}$.

(2) $\mathsf{E}[\widehat{\alpha}] = \alpha^*$; $\mathsf{Var}[\widehat{\alpha}] = \sigma^2\left\{\dfrac{1}{n} + \dfrac{\overline{x}^2}{Q_{xx}}\right\}$.

(3) $\mathsf{Cov}[\widehat{\alpha}, \widehat{\beta}] = -\dfrac{\overline{x}\sigma^2}{Q_{xx}}$.

(4) $\mathsf{E}[e_j] = 0$; $\sum_{j=1}^{n}\mathsf{Var}[e_j] = (n-2)\sigma^2\,(j = 1, 2, \ldots, n)$.

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

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

とおく. すると

$$\begin{aligned} \epsilon_j - \overline{\epsilon} &= y_j + \alpha^* + \beta^* x_j - \{\overline{y} + \alpha^* + \beta^*\overline{x}\} \\ &= y_j - \overline{y} - \beta^*(x_j - \overline{x}), \end{aligned} \tag{4.3}$$

eq:2-1b

$$\sum_{j=1}^{n}(x_j - \overline{x})\overline{\epsilon} = 0 \tag{4.4}$$

eq:2-1c

となる．(4.3) の両辺に $x_j - \overline{x}$ をかけて j について和を取ると

$$\sum_{j=1}^{n}(x_j-\overline{x})(\epsilon_j-\overline{\epsilon}) = \sum_{j=1}^{n}(x_j-\overline{x})(y_j-\overline{y}) - \beta^* \sum_{j=1}^{n}(x_j-\overline{x})^2 = Q_{xy} - \beta^* Q_{xx}$$
$$= Q_{xx}\{\widehat{\beta}-\beta^*\}$$

を得る．この式の右辺に (4.4) を代入すると これより

$$\widehat{\beta}-\beta^* = \frac{\sum_{j=1}^{n}(x_j-\overline{x})\epsilon_j}{Q_{xx}} \tag{4.5}$$ eq:2-2

と得る．書ける．(4.5) と $\mathsf{E}[\epsilon_j]=0$ に注意すると

$$\mathsf{E}[\widehat{\beta}-\beta^*] = 0 \tag{4.6}$$ eq:2-2a

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

$$\mathsf{Var}[\widehat{\beta}] = \mathsf{E}[(\widehat{\beta}-\beta^*)^2] = \frac{\sum_{j=1}^{n}(x_j-\overline{x})^2\mathsf{E}[\epsilon_j^2]}{Q_{xx}^2} = \frac{\sigma^2}{Q_{xx}} \tag{4.7}$$ eq:2-3

がわかる．

(2) の証明: また，$\widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x}$ と (4.5) から

$$\widehat{\alpha}-\alpha^* = \overline{y} - \widehat{\beta}\overline{x} - \alpha^* - \widehat{\beta}\overline{x} - \alpha^* = \alpha^* + \beta^*\overline{x} + \overline{\epsilon}$$
$$= \overline{\epsilon} - \overline{x}(\widehat{\beta}-\beta^*) = \sum_{j=1}^{n}\left\{\frac{1}{n} - \frac{(x_j-\overline{x})\overline{x}}{Q_{xx}}\right\}\epsilon_j \tag{4.8}$$ eq:2-2b

がわかる．再度 $\mathsf{E}[\epsilon_j]=0\,(j=1,\,2,\,\ldots,\,n)$ と (4.8) から より

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

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

$$\mathsf{Var}[\widehat{\alpha}] = \mathsf{E}[(\widehat{\alpha}-\alpha^*)^2] = \sum_{j=1}^{n}\left\{\frac{1}{n} - \frac{(x_j-\overline{x})\overline{x}}{Q_{xx}}\right\}^2 \mathsf{E}[\epsilon_j^2]$$
$$= \sigma^2\left\{\frac{1}{n} + \frac{\overline{x}^2}{Q_{xx}}\right\} \tag{4.9}$$ eq:2-2c

がわかる．

(3) の証明: 同様に，(4.5) と (4.8) から

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

がわかる.

(4) の証明: (4.2) から

$$\begin{aligned} e_j &= y_i - \{\widehat{\alpha} + \widehat{\beta}x_j\} = y_j - \{\overline{y} - \widehat{\beta}\overline{x} + \widehat{\beta}x_j\} \\ &= y_j - \overline{y} - \widehat{\beta}(x_j - \overline{x}) = (\epsilon_j - \overline{\epsilon}) - (x_j - \overline{x})(\widehat{\beta} - \beta^*) \end{aligned} \tag{4.10}$$

eq:2-2d

と表さる. よって, $\mathsf{E}[\epsilon_j] = 0\,(j = 1, 2, \ldots, n$, $\mathsf{E}[\overline{\epsilon}] = 0$ と (4.7) からので

$$\mathsf{E}[e_j] = 0$$

がわかる. つぎに, (4.10) の最右辺を展開して, j について和と期待値を取ると

$$\begin{aligned} \sum_{j=1}^{n} \mathsf{E}[e_j^2] &= \sum_{j=1}^{n} \mathsf{E}[(\epsilon_j - \overline{\epsilon})^2] - 2\sum_{j=1}^{n}(x_j - \overline{x})\mathsf{E}[(\epsilon_j - \overline{\epsilon})(\widehat{\beta} - \beta^*)] \\ &\quad + \sum_{j=1}^{n}(x_j - \overline{x})^2\mathsf{E}[(\widehat{\beta} - \beta^*)^2] \end{aligned} \tag{4.11}$$

eq:2-4

と書ける. (4.5) から

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

となる. これらの 2 つの式と (4.7) を (4.11) に代入すれば

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

を得る. □

pro:2-2 **命題 4.2.** 正規性を仮定する. $n \geq 3$ のとき

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

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

(3) $\widehat{\sigma}^2$ と $(\widehat{\alpha}, \widehat{\beta})$ は独立.

Proof. (1), (2) は命題 4.1 と正規性の仮定より直ちにわかる. $\mathsf{Cov}[e_j, \widehat{\alpha}] = \mathsf{Cov}[e_j, \widehat{\beta}] = 0$ より $\widehat{\sigma}^2$ と $(\widehat{\alpha}, \widehat{\beta})$ は独立 □

4.2 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

$$y_j = \beta_0^* + \beta_1^* x_{j1} + \beta_2^* x_{j2} + \cdots + \beta_d^* x_{jd} + \epsilon_j \qquad (j = 1, 2, \ldots, n) \quad (4.12)$$ eq:2-5

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 $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ は独立同一分布に従う確率変数列 (誤差項) であり. $\mathsf{Var}[\epsilon_j] = \sigma^2\,(j = 1, 2, \ldots, n)$ とする. ただし $\sigma^2\,(\sigma > 0)$ は未知とする. (4.12) を重回帰モデルという.

重回帰モデル (4.12) は

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

と表される. これは

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon} \quad (4.13)$$ eq:2-6

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

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

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

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

とおくと

$$h(\boldsymbol{\beta}) = (\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^\top (\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \mathbf{X}^\top\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \quad (4.14)$$ eq:2-7

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.

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

と書けるので

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

さらに

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

とする. トレース, 期待値およびベキ等行列の性質より

$$\begin{aligned} \mathsf{E}[\mathsf{RSS}] &= \mathrm{Tr}\left[\left\{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right\}\mathsf{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top]\right] && (4.15) \\ &= \sigma^2 \mathrm{Tr}\left[\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right] && (4.16) \\ &= (n - d - 1)\sigma^2. && (4.17) \end{aligned}$$ eq:2-7

よって $n \geq d + 2$ のとき σ^2 の不偏推定量は

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

thm:2-1 **定理 4.3.** (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.

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

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

をみたさなければならない. さらに

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

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,

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

しかし

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

よって

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

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

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

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 $\gtrsim$ 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. □

pro:2-3

命題 4.4. $n \geq d+2$ 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) $\widehat{\boldsymbol{\beta}} \sim \mathsf{N}_{d+1}(\boldsymbol{\beta}^*, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$.
(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.

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

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

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 $\mathbf{A}^{-1/2}$ [1]と記した. すると命題の主張は
(1a) $U \sim \mathsf{N}_{d+1}(\mathbf{0}, \mathbf{I}_{d+1})$,
(2a) $V \sim \chi^2_{n-d-1}$,
(3a) U と V は独立,

[1] これは一意ではないことに注意せよ.

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

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

と分解できる. ただし $\mathbf{P}$ は $n \times (d+1)$ 行列で $\mathbf{P}^\top\mathbf{P} = \mathbf{I}_{d+1}$, $\mathbf{\Lambda}$ 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

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

より

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

これらより

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

と書けることがわかる. ここで適当な $n \times (n-d-1)$ 行列 $\mathbf{Q}$ 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

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

と表される.

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

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

4.3 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

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

と評価できる. ただし $\mathbf{C}$ 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.

4.3.1 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.

4.3.2 Mallows C_p

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

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

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

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

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

$$\mathsf{E}[\mathsf{RSS}_d] = \mathsf{E}\big[\boldsymbol{y}^\top\big\{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\big\}\boldsymbol{y}\big] = (n-d-1)\sigma^2 \qquad (4.18)$$

eq:2-7a

となるので, 予測誤差は

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

となる. よって 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

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

とする.

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

としたとき分散 σ^2 の推定量を

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

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.

問 4.1. $\boldsymbol{y} \sim \mathsf{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ 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) $\mathsf{E}[\mathbf{Q}\boldsymbol{y}]$ を求めよ.
(3) 式 (4.18) を証明せよ.
$\boldsymbol{y}$ の分布の仮定より $\mathsf{E}[\boldsymbol{y}] = \mathbf{X}\boldsymbol{\beta}$ と $\mathsf{Var}[\boldsymbol{y}] = \sigma^2\mathbf{I}_n$ となることと注意 A.41(2) を使うとよい.

4.3.3 AIC

df:2-1

定義 4.5. (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

$$\mathsf{KL}(\mathsf{p}, \mathsf{q}) = \int \log\left(\frac{\mathsf{p}}{\mathsf{q}}\right) \mathsf{p}\, d\mathsf{m}$$

で定義する.

re:2-1 **注意 4.6.** $x\log x \geq x-1\,(x>0)$ に注意して

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

□

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

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

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

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

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

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

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.

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

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

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

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 $\mathsf{AI}(\boldsymbol{\beta}^*, \sigma^2)$ を計算してみると

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

であることと命題 4.4 に注意し $\tilde{\boldsymbol{y}}$ に関して積分[2]する.

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

と書ける. 命題 4.4 から

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

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,

$$\mathsf{AI}(\boldsymbol{\beta}^*, \sigma^2) = \mathsf{E}[n\log(2\pi\widehat{\sigma}^2(\boldsymbol{y}))] + \frac{(n+d+1)(n-d-1)}{n-d-3} \qquad (4.19)$$ eq:2-11

と書けることがわかる.

次に, 対数尤度関数を計算する.

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

に注意して

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

となる. これを (4.19) に代入すれば

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

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

$$\mathsf{AIC} := -2\log \mathsf{p}(\boldsymbol{y}|\, \mathbf{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{y}), \widehat{\sigma}^2(\boldsymbol{y})) + 2(d+2) \qquad (4.20)$$ eq:2-12

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, (4.20) 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.

4.3.4 交差検証法

観測 $(\mathbf{X}_1, y_1), \ldots, (\mathbf{X}_2, y_2), \ldots, (\mathbf{X}_n, y_n)$ から j 番目 $(j = 1, 2, \ldots, n)$ 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.

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

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.

pro:2-4 **命題 4.7.** CV は以下のように表現できる.

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

Proof. 命題の主張を証明するために

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

を示せばよい.

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

に注意する. 等式

$$(\mathbf{X}^\top\mathbf{X})^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\boldsymbol{x}_1\boldsymbol{x}_1^\top\mathbf{A}^{-1}}{1+\boldsymbol{x}_1^\top\mathbf{A}^{-1}\boldsymbol{x}_1} \tag{4.21}$$ eq:2-13

より

$$\begin{aligned}1-\boldsymbol{x}_1^\top(\mathbf{X}^\top\mathbf{X})^{-1}\boldsymbol{x}_1 &= 1-\boldsymbol{x}_1^\top\mathbf{A}^{-1}\boldsymbol{x}_1 + \frac{\boldsymbol{x}_1^\top\mathbf{A}^{-1}\boldsymbol{x}_1\boldsymbol{x}_1^\top\mathbf{A}^{-1}\boldsymbol{x}_1}{1+\boldsymbol{x}_1^\top\mathbf{A}^{-1}\boldsymbol{x}_1}\\ &= \frac{1-(\boldsymbol{x}_1^\top\mathbf{A}^{-1}\boldsymbol{x}_1)^2+(\boldsymbol{x}_1^\top\mathbf{A}^{-1}\boldsymbol{x}_1)^2}{1+\boldsymbol{x}_1^\top\mathbf{A}^{-1}\boldsymbol{x}_1}\\ &= \frac{1}{1+\boldsymbol{x}_1^\top\mathbf{A}^{-1}\boldsymbol{x}_1} \qquad (4.22)\end{aligned}$$ eq:2-14

を得る. これらより

$$\begin{aligned}\boldsymbol{x}_1^\top\widehat{\boldsymbol{\beta}} &= \boldsymbol{x}_1^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y}\\ &= \boldsymbol{x}_1^\top\left\{\mathbf{A}^{-1}-\frac{\mathbf{A}^{-1}\boldsymbol{x}_1\boldsymbol{x}_1^\top\mathbf{A}^{-1}}{1+\boldsymbol{x}_1^\top\mathbf{A}^{-1}\boldsymbol{x}_1}\right\}(\boldsymbol{x}_1;\,(\mathbf{X}^{(1)})^\top)\begin{pmatrix}y_1\\ \boldsymbol{y}^{(1)}\end{pmatrix}\\ &= \boldsymbol{x}_1^\top\left\{\mathbf{A}^{-1}-\frac{\mathbf{A}^{-1}\boldsymbol{x}_1\boldsymbol{x}_1^\top\mathbf{A}^{-1}}{1+\boldsymbol{x}_1^\top\mathbf{A}^{-1}\boldsymbol{x}_1}\right\}\{\boldsymbol{x}_1y_1+(\mathbf{X}^{(1)})^\top\boldsymbol{y}^{(1)}\}\\ &= \boldsymbol{x}_1\Bigg\{\mathbf{A}^{-1}\boldsymbol{x}_1y_1-\frac{\mathbf{A}^{-1}\boldsymbol{x}_1\boldsymbol{x}_1^\top\mathbf{A}^{-1}\boldsymbol{x}_1y_1}{1+\boldsymbol{x}_1\mathbf{A}^{-1}\boldsymbol{x}_1}+\mathbf{A}^{-1}(\mathbf{X}^{(1)})^\top\boldsymbol{y}^{(1)}\\ &\qquad -\frac{\mathbf{A}^{-1}\boldsymbol{x}_1\boldsymbol{x}_1^\top\mathbf{A}^{-1}(\mathbf{X}^{(1)})^\top\boldsymbol{y}^{(1)}}{1+\boldsymbol{x}_1\mathbf{A}^{-1}\boldsymbol{x}_1}\Bigg\}\\ &= \boldsymbol{x}_1^\top\mathbf{A}^{-1}\boldsymbol{x}_1y_1-\frac{(\boldsymbol{x}_1^\top\mathbf{A}^{-1}\boldsymbol{x}_1)^2y_1}{1+\boldsymbol{x}_1^\top\mathbf{A}^{-1}\boldsymbol{x}_1}+\boldsymbol{x}_1^\top\mathbf{A}^{-1}(\mathbf{X}^{(1)})^\top\boldsymbol{y}^{(1)}\\ &\qquad -\frac{\boldsymbol{x}_1^\top\mathbf{A}^{-1}\boldsymbol{x}_1\boldsymbol{x}_1^\top\mathbf{A}^{-1}(\mathbf{X}^{(1)})^\top\boldsymbol{y}^{(1)}}{1+\boldsymbol{x}_1^\top\mathbf{A}^{-1}\boldsymbol{x}_1}\\ &= \boldsymbol{x}_1^\top\mathbf{A}^{-1}\boldsymbol{x}_1y_1-\frac{(\boldsymbol{x}_1^\top\mathbf{A}^{-1}\boldsymbol{x}_1)^2y_1}{1+\boldsymbol{x}_1^\top\mathbf{A}^{-1}\boldsymbol{x}_1}+\boldsymbol{x}_1^\top\widehat{\boldsymbol{\beta}}^{(1)}-\frac{\boldsymbol{x}_1^\top\mathbf{A}^{-1}\boldsymbol{x}_1\boldsymbol{x}_1^\top\widehat{\boldsymbol{\beta}}^{(1)}}{1+\boldsymbol{x}_1^\top\mathbf{A}^{-1}\boldsymbol{x}_1}\\ &= \frac{\boldsymbol{x}_1^\top\mathbf{A}^{-1}\boldsymbol{x}_1y_1+\boldsymbol{x}_1^\top\widehat{\boldsymbol{\beta}}^{(1)}}{1+\boldsymbol{x}_1^\top\mathbf{A}^{-1}\boldsymbol{x}_1} \qquad (4.23)\end{aligned}$$ eq:2-15

を得る. (4.22) と (4.23) より

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

を得る. □

4.3.5 BIC

線型回帰モデル

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

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

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

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 $\boldsymbol{y}$ の周辺分布

$$\mathsf{p}_{\pi_{d+2}}(\boldsymbol{y}) = \int \cdots \int \mathsf{p}(\boldsymbol{y} | \boldsymbol{X}\boldsymbol{\beta}, \sigma^2) \pi_{d+2}(\boldsymbol{\beta}, \sigma^2)\, d\mathsf{m}_d(\boldsymbol{\beta})\, d\mathsf{m}(\sigma^2)$$

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 $-2 \log \mathsf{p}_{\pi_{d+2}}(\boldsymbol{y})$ 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

$$y_j = \beta_0^* + \epsilon_j \qquad (j = 1, 2, \ldots, n) \tag{4.24}$$

eq:2-16

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

$$\mathsf{p}_{\pi_2}(\boldsymbol{y}) = \int \int \mathsf{p}(\boldsymbol{y} | \beta_0, \sigma^2) \pi_2(\beta_0, \sigma^2)\, d\mathsf{m}(\beta_0)\, d\mathsf{m}(\sigma^2)$$

を考える. ただし, $\mathsf{p}(\boldsymbol{y} | \beta_0^*, \sigma^2)$ はモデル (4.24) の確率密度関数である. これらの比

$$\frac{\mathsf{p}_{\pi_{d+2}}(\boldsymbol{y})}{\mathsf{p}_{\pi_2}(\boldsymbol{y})}$$

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

$$\boldsymbol{X}|\,\boldsymbol{\theta} \sim \mathsf{p}(\boldsymbol{x}|\,\boldsymbol{\theta})$$

とし, $\boldsymbol{\theta}$ の事前分布を

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

とする．対数尤度

$$\ell(\boldsymbol{\theta}|\,\boldsymbol{x}) = \log \mathsf{p}(\boldsymbol{x}|\,\boldsymbol{\theta})$$

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

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

と近似できる．ただし

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

である．さらに

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

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

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

[3]すなわち, 存在すれば,

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

で定義する.

である．よって

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

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

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

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，

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

である．したがって

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

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

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

は $\log n$ に比較して無視できるので，

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

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

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

を得る．

2-2 **注意 4.8.** 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. □

4.4 補遺：Laplace 近似について

pro:2-5 **命題 4.9.** 連続関数 $h,\ g:\mathbb{R}^d \longrightarrow \mathbb{R}$ と $\boldsymbol{x}_0 \in \mathbb{R}^d$ に以下の条件を仮定する.

(1) h は $\boldsymbol{x}_0$ において最小値を取り,

$$h(\boldsymbol{x}) > h(\boldsymbol{x}_0) \qquad (\boldsymbol{x} \neq \boldsymbol{x}_0).$$

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

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

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

$$\lim_{\lambda\to\infty} e^{\lambda h(\boldsymbol{x}_0)-\epsilon} g(\boldsymbol{x}_0) \int_{|\boldsymbol{x}-\boldsymbol{x}_0|_2>\lambda} \exp\{-\lambda h(\boldsymbol{x})\} g(\boldsymbol{x})\, d\mathsf{m}_d(\boldsymbol{x}) = 0.$$

このとき

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

が成立する.

Proof. まず

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

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, $|\boldsymbol{x}-\boldsymbol{x}_0|_2 \le \delta$ ならば,

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

となる. これから

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

よって

$$\begin{aligned}
&(1-\epsilon)^{-d/2+1}g(\boldsymbol{x}_0)(2\pi)^{d/2}\sqrt{|\boldsymbol{I}(\boldsymbol{x}_0)|}\\
&\le\liminf_{\lambda\to\infty}\lambda^{d/2}\int_{|\boldsymbol{x}-\boldsymbol{x}_0|_2\le\delta}\exp\{-\lambda h(\boldsymbol{x})\}\,g(\boldsymbol{x})\,d\mathsf{m}_d(\boldsymbol{x})\\
&\le\limsup_{\lambda\to\infty}\lambda^{d/2}\int_{|\boldsymbol{x}-\boldsymbol{x}_0|_2\le\delta}\exp\{-\lambda h(\boldsymbol{x})\}\,g(\boldsymbol{x})\,d\mathsf{m}_d(\boldsymbol{x})\\
&\le(1+\epsilon)^{-d/2+1}g(\boldsymbol{x}_0)(2\pi)^{d/2}\sqrt{|\mathbf{I}(\boldsymbol{x}_0)|}.
\end{aligned}\tag{4.25}$$

eq:2-17

次に, 容易にわかるように

$$\int_{|\boldsymbol{x}-\boldsymbol{x}_0|_2>\delta}\exp\bigl\{-\lambda h(\boldsymbol{x})\bigr\}g(\boldsymbol{x})\,d\mathsf{m}_d(\boldsymbol{x})=0.\tag{4.26}$$

eq:2-18

よって, $\boldsymbol{x}_0$ の近くの評価 (4.25) と遠方での評価 (4.26) と $\epsilon>0$ は任意に取ったことより

$$\lim_{\lambda\to\infty}\lambda^{d/2}\int_{\mathbb{R}^d}\exp\bigl\{-\lambda\{h(\boldsymbol{x})-h(\boldsymbol{x}_0)\}\bigr\}g(\boldsymbol{x})\,d\mathsf{m}_d(\boldsymbol{x})=g(\boldsymbol{x}_0)(2\pi)^{d/2}|\mathbf{I}(\boldsymbol{x}_0)|^{1/2}.$$

したがって,

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

□