『統計的機械学習の数理 100 問 with R』の解答例 (数理)

鈴木 譲

プログラムに関しては、略解 (R プログラム).html を参照してください。

第 1 章 線形回帰

1. $S := \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2$

   (a) $\bar{x} := \frac{1}{N}\sum_{i=1}^{N} x_i$, $\bar{y} := \frac{1}{N}\sum_{i=1}^{N} y_i$ とおくと、

   $$\frac{\partial S}{\partial \beta_0} = -2\sum_{i=1}^{N} (y_i - \beta_0 - \beta, x_i) = 0 \Longleftrightarrow \sum_{i=1}^{N} (\beta_0 + \beta_1 x_i) = \sum_{i=1}^{N} y_i$$
   $$\Longleftrightarrow N\beta_0 + \beta_1 \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i \Longleftrightarrow \beta_0 + \beta_1 \cdot \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{N}\sum_{i=1}^{N} y_i \Longleftrightarrow \beta_0 + \beta_1 \bar{x} = \bar{y}$$

   を得る。

   (b) $\beta_0 = \bar{y} - \beta_1 \bar{x}$ より、

   $$\frac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^{N} x_i (y_i - \beta_0 - \beta_1 x_i) = 0 \Longleftrightarrow \beta_0 \sum_{i=1}^{N} x_i + \beta_1 \sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} x_i y_i$$
   $$\Longleftrightarrow (\bar{y} - \beta_1 \bar{x}) \sum_{i=1}^{N} x_i + \beta_1 \sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} x_i y_i \Longleftrightarrow N\bar{x}\bar{y} - \beta_1 N\bar{x}^2 + \beta_1 \sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} x_i y_i$$
   $$\Longleftrightarrow \beta_1 \left(\sum_{i=1}^{N} x_i^2 - N\bar{x}^2 \right) = \sum_{i=1}^{N} x_i y_i - N\bar{x}\bar{y} \Longleftrightarrow \beta_1 \sum_{i=1}^{N} (x_i - \bar{x})^2 = \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$$

   $x_1, \cdots, x_N$ がすべて等しくない、つまり $\sum_{i=1}^{N} (x_i - \bar{x})^2 \neq 0$ であるから、

   $$\beta_1 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

   となる。

2. l の傾き $(\hat{\beta}_1)$ は

   $$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

   で、切片 $(\hat{\beta}_0)$ は、

   $$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

により定まる。$x_i - \bar{x} \mapsto x_i, y_i - \bar{y} \mapsto y_i$ $(i = 1, \ldots, N)$ を考えると、l' の傾き $\hat{\beta}_1$ は、

$$\hat{\beta}_1 = \frac{\sum_{i=1}^N x_i y_i}{\sum_{i=1}^N x_i^2}$$

となる。このとき、$\bar{x} = \bar{y} = 0$ より l' の切片は 0 となる (原点を通る)。また、$\hat{\beta}_1$ が求まってから l の切片 $\hat{\beta}_0$ を求めるには $\hat{\beta}_1$ と (a') から切片 $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ を求めればよい。

(a) 任意の $z \in \mathbb{R}^m$ に対して

$$Az = B^\top Bz = 0 \Rightarrow z^\top B^\top Bz = 0 \Rightarrow (Bz)^\top Bz = 0 \Rightarrow \|Bz\|^2 = 0 \Rightarrow Bz = 0$$
$$Bz = 0 \Rightarrow B^\top Bz = 0 \Rightarrow Az = 0$$

となる。よって、

$$Az = 0 \Leftrightarrow Bz = 0$$

(b) (a) より A, B による線形写像の核が等しい。また、命題 4(次元定理) により、A, B はともに像の次元と核の次元の和が m になる。よって、A, B の像の次元は等しく、命題 4 より A, B の階数も等しい。

5. $X \in \mathbb{R}^{N \times (p+1)}$ を最初の列がすべて 1 の行列として、

(a) $N < p + 1$ のとき命題 3 より、

$$\operatorname{rank}\left(X^\top X\right) \leq \operatorname{rank}(X) = \min\{N, p+1\} = N < p + 1.$$

$X^\top X \in \mathbb{R}^{(p+1)\times(p+1)}$ が正方行列であることに注意すると、命題 1 より $X^\top X$ は逆行列をもたない。

(b) $N \geq p + 1$ であって、X のある 2 列が等しいとき、命題 3 より、

$$\operatorname{rank}\left(X^\top X\right) \leq \operatorname{rank}(X) < p + 1$$

よって (a) と同様の理由で X'X は逆行列をもたない。

6. (a) $j = 0, 1, \ldots, p$ に対して、

$$L = \frac{1}{2}\sum_{i=1}^N \left(y_i - \sum_{k=0}^p x_{i,k}\beta_k\right)^2 = \frac{1}{2}\sum_{i=1}^N \left(y_i - \sum_{k\neq j} x_{i,k}\beta_k - x_{i,j}\beta_j\right)^2$$

が成り立つ。これの β_j での偏微分は

$$\frac{\partial L}{\partial \beta_j} = \frac{1}{2}\sum_{i=1}^N \left\{2x_{i,j}^2\beta_j - 2x_{i,j}\left(y_i - \sum_{k\neq j} x_{i,k}\beta_k\right)\right\} = -\sum_{i=1}^N x_{i,j}y_i + \sum_{i=1}^N \left(x_{i,j}^2\beta_j + x_{i,j}\sum_{k\neq j} x_{i,k}\beta_k\right)$$
$$= -\sum_{i=1}^N x_{i,j}y_i + \sum_{k=0}^p \sum_{i=1}^N x_{i,j}x_{i,k}\beta_k$$

となる。一方、$X^\top y$ の第 j 成分は、$\sum_{i=1}^N x_{i,j}y_i$, $X^\top X$ の第 (j,k) 成分は $\sum_{i=1}^N x_{i,j}x_{i,k}$, $X^\top X\beta$ の第 j 成分は $\sum_{k=0}^p \sum_{i=1}^N x_{i,j}x_{i,k}\beta_k$ であるから、$-X^\top y + X^\top x\beta$ の 第 j 成分は

$$-\sum_{i=1}^N x_{i,j}y_i + \sum_{k=0}^p \sum_{i=1}^N x_{i,j}x_{i,k}\beta_k$$

となり、$\frac{\partial L}{\partial \beta_j}$ と一致することから、題意は示された。

(b) (a) の計算により $\frac{\partial L}{\partial \beta_j} = 0$ なる $\beta \in \mathbb{R}^{p+1}$ は、

$$-X^\top y + X^\top X\beta = \emptyset$$

をみたせばよい。$X^\top X$ に逆行列が存在することを仮定すれば、求める $\hat{\beta}$ は

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

となる。

7. (a) 条件より $y = X\beta + \varepsilon$ となる。これを命題 11 に代入すると、

$$\begin{aligned}\hat{\beta} &= \left(X^\top X\right)^{-1} X^\top y = \left(X^\top X\right)^{-1} X^\top (X\beta + \varepsilon) = \left(X^\top X\right)^{-1} X^\top X\beta + \left(X^\top x\right)^{-1} X^\top \varepsilon \\ &= \beta + \left(X^\top X\right)^{-1} X^\top \varepsilon\end{aligned}$$

を得る。

(b) $\varepsilon \sim N\left(0, \sigma^2 I\right)$ より、$\varepsilon \in \mathbb{R}^N$ の平均は 0 であるので、定数の行列 $\left(X^\top X\right)^{-1} X^\top$ がかかっていても、$\left(X^\top X\right)^{-1} X^\top \varepsilon$の 平均も 0 となる。よって、(a) より $\mathbb{E}[\hat{\beta}] = \beta$ であるから、題意は示された。

(c) (a) より、

$$\begin{aligned}&E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^\top\right] = E\left[\left(X^\top X\right)^{-1} X^\top \varepsilon \left\{\left(X^\top X\right)^{-1} X^\top \varepsilon\right\}^\top\right] \\ =&E\left[\left(X^\top X\right)^{-1} X^\top \varepsilon\varepsilon^\top X \left(X^\top X\right)^{-1}\right] = \left(X^\top X\right)^{-1} X^\top E\left[\varepsilon\varepsilon^\top\right] X \left(X^\top X\right)^{-1} \\ =&\left(X^\top X\right)^{-1} X^\top \sigma^2 I X \left(X^\top X\right)^{-1} = \sigma^2 \left(X^\top X\right)^{-1} X^\top X \left(X^\top X\right)^{-1} = \sigma^2 \left(X^\top X\right)^{-1}\end{aligned}$$

ここで、ε の共分散行列が $E\left[\varepsilon\varepsilon^\top\right] = \sigma^2$ となることを用いた。

8. $H = X\left(X^\top X\right)^{-1} X^\top \in \mathbb{R}^{N\times N}$, $\hat{y} = X\hat{\beta}$ とおくと、

(a) $H^2 = X\left(X^\top X\right)^{-1} X^\top X \left(X^\top X\right)^{-1} X^\top = X\left(X^\top X\right)^{-1} X^\top = H$ より、$H^2 = H$ が成立する。

(b) $(I-H)^2 = I - 2H + H^2 = I - 2H + H = I - H$ より、$(I-H)^2 = I - H$ が成立する。

(c) $HX = X\left(X^\top X\right)^{-1} X^\top = X$ より、$HX = X$ が成立する。

(d) 命題 11 より、$\hat{y} = x\hat{\beta} = x\left(X^\top x\right)^{-1} X^\top y = Hy$ となり、$\hat{y} = Hy$ が成立する。

(e) $y - \hat{y} = y - Hy = (I-H)(X\beta+\varepsilon) = (X - HX)\beta + (I-H)\varepsilon = (X-X)\beta + (I-H)\varepsilon = (I-H)\varepsilon$ となり、$y - \hat{y} = (I-H)\varepsilon$ が成立する。ただし、最初の等号は (d) を、次の等号は (1.12) を、最後から 2 番目の等号は (c) を用いた。

(f) $\|y - \hat{y}\|^2 = (y-\hat{y})^\top (y-\hat{y}) = \{(I-H)\varepsilon\}^\top (I-H)\varepsilon = \varepsilon^\top (I-H)^\top (I-H)\varepsilon = \varepsilon^T (I-H)^2\varepsilon = \varepsilon^\top (I-H)\varepsilon$ が成立する。ただし、2 番目の等号は (e) を、最後から 2 番目の等号は転置の線形性と

$$\begin{aligned}H^\top &= \left\{X\left(X^\top X\right)^{-1} X^\top\right\}^\top = X\left\{\left(X^\top X\right)^{-1}\right\}^\top X^\top = X\left\{\left(X^\top X\right)^\top\right\}^{-1} X^\top \\ &= X\left(X^\top X\right)^{-1} X^\top = H\end{aligned}$$

から、最後の変形は (b) によった。したがって、$\|y - \hat{y}\|^2 = \varepsilon^\top (I-H)\varepsilon$ が成立する。

9. (a) $H := X(X^\top X)^{-1}X^\top$ とおく。命題 3 と、$\mathrm{rank}(X) = p+1$ より、

$$\mathrm{rank}(H) \leq \min\{\mathrm{rank}(X(X^\top X)^{-1}), \mathrm{rank}(X^\top)\} \leq \mathrm{rank}(X^\top) = \mathrm{rank}(X) = p+1$$

となる。一方、前問の (c) より $HX = X$ を用いて

$$\mathrm{rank}(HX) = \mathrm{rank}(X) = p+1$$

より、

$$\mathrm{rank}(HX) \leq \min\{\mathrm{rank}(H), \mathrm{rank}(X)\} \leq \mathrm{rank}(H)$$

したがって、$\mathrm{rank}(H) \geq p+1$ が成立する。以上より、$\mathrm{rank}(H) = p+1$ だから、命題 4 より H の像の次元は $p+1$ である。

(b) 前問 (c): $HX = X$ より、X の各列ベクトルは、H の固有値 1 のベクトルである。また、$\mathrm{rank}(x) = p+1$ より、X の各列ベクトルは 1 次独立である。よって、X の各列ベクトルは H の固有値 1 の固有空間の基底となるので、その次元は $p+1$ である。ここで、H の固有値の固有空間は、H の核と等しい。よって、命題 4 と $\mathrm{rank}(H) = p+1$ を用いると、核の次元は $N-p-1$ であるから、固有値が 0 の固有空間は $N-p-1$ 次である。

(c) 任意の $x \in \mathbb{R}^{p+1}$ に対して、

$$(I-H)x = 0 \Leftrightarrow Hx = x$$

である。よって、H の固有値 1 の固有空間と $I-H$ の固有値 0 の固有空間が等しいため、$I-H$ の固有値 0 の固有空間の次元は $p+1$ である。また、

$$(I-H)x = x \Leftrightarrow Hx = 0$$

である。よって H の固有値 0 の固有空間と $I-H$ の固有値 1 の固有空間が等しいため、$I-H$ の固有値 1 の固有空間の次元は $N-p-1$ である。

10. (a) $\varepsilon \sim N\left(0, \sigma^2 I\right)$ とする。$v = P\varepsilon$ とおくと、$\varepsilon = P^{-1}v = P^\top v$ とでき、

$$\varepsilon^\top (I-H)\varepsilon = v^\top P(I-H)P^\top v$$

ここで、$P(I-H)P^\top$ は N 個の固有値を成分にもつ対角行列となる。特に、前問の結果から、$I-H$ は固有値 1 が $N-p-1$ 個、固有値 0 が $p+1$ 個あることが示される。したがって、

$$v^\top P(I-H)P^\top v = \sum_{i=1}^{N-p-1} v_i^2$$

が得られる。

(b) $E\left[vv^\top\right] = PE\left[\varepsilon\varepsilon^\top\right]P^\top = P\sigma^2 IP^\top = \sigma^2 IPP^\top = \sigma^2 I$ となる。

(c) $V := [v_1, \ldots, v_N]$, $v_i \sim N\left(0, \sigma^2\right)$, $i = 1, \ldots, N$ とすると、$Z_i = \frac{V_i}{\sigma}$ は互いに独立であって、それぞれ $N(0,1)$ に従う。 よって、

$$\sum_{i=1}^{N-p-1} \frac{v_i^2}{\sigma^2} \sim \chi^2_{N-p-1}$$

であって、(a) より

$$\frac{RSS}{\sigma^2} \sim \chi^2_{N-p-1}$$

11. (a)

$$E\left[(\hat{\beta}-\beta)(y-\hat{y})^T\right] = E\left[\left(X^TX\right)^{-1}X^T\epsilon\epsilon^T(I-H)\right] = \left(X^TX\right)^{-1}X^TE\left[\epsilon\epsilon^T\right](I-H)$$
$$= \left(X^TX\right)^{-1}X^T\sigma^2I(I-H) = \sigma^2\left\{\left(X^TX\right)^{-1}X^T - \left(X^TX\right)^{-1}X^TH\right\}$$
$$= \sigma^2\left\{\left(X^TX\right)^{-1}X^T - \left(X^TX\right)^{-1}X^TX\left(X^TX\right)^{-1}X^T\right\} = 0$$

(b) $i = 0, 1, \cdots, p$ に対して、$\left(\hat{\beta}_i - \beta_i\right) / \left(\sqrt{B_i}\sigma\right)$ は $\hat{\beta} - \beta$ の関数であり、RSS は $y - \hat{y}$ の関数であり、$\hat{\beta} - \beta$ と $y - \hat{y}$ はともに正規分布に従うことと、(a) より共分散行列が 0 になることから、この 2 つは独立である。 よって、題意は示された。

(c) $i = 0, 1, \cdots, p$ に対して、

$$\frac{\hat{\beta}_i - \beta_i}{SE\left(\hat{\beta}_i\right)} = \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}\sqrt{B_i}} = \frac{\hat{\beta}_i - \beta_i}{\sqrt{\frac{RSS}{N-p-1}}\sqrt{B_i}} = \frac{\hat{\beta}_i - \beta_i}{\sqrt{\frac{RSS}{\sigma^2}}}\sqrt{B_i}\sigma = \frac{\hat{\beta}_i - \beta_i}{\sqrt{B_i}\sigma} \Big/ \sqrt{\frac{\frac{RSS}{\sigma^2}}{N-p-1}}$$

このとき、$\hat{\beta} \sim N\left(\beta, \sigma^2\left(X^TX\right)^{-1}\right)$ より、$\dfrac{\hat{\beta}_i - \beta_i}{\sqrt{B_i}\sigma} \sim N(0,1)$ であり、10.(c) も合わせて考えると、$\dfrac{\hat{\beta}_i - \beta_i}{SE\left(\hat{\beta}_i\right)} \sim t_{N-p-1}$ となることがわかる。

(d) 以下、$\sum_i = \sum_{i=1}^{N}$ のように表す。$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix}$ に対して、

$$X^TX = \begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_N \end{pmatrix}\begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix} = \begin{pmatrix} N & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix} = N\begin{pmatrix} 1 & \bar{x} \\ \bar{x} & \frac{1}{N}\sum_i x_i^2 \end{pmatrix}$$

したがって、

$$\left(X^TX\right)^{-1} = \frac{1}{N}\frac{1}{\frac{1}{N}\sum_i x_i^2 - (\bar{x})^2}\begin{pmatrix} \frac{1}{N}\sum_i x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} = \frac{1}{\sum_i (x_i - \bar{x})^2}\begin{pmatrix} \frac{1}{N}\sum_i x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}$$

となる。 ゆえに、題意は示された。

14. (a) 8. (c) で示した等式 $HX = X$ について、X の第 0 列ベクトルは成分がすべて 1 となっていることから、すべての成分 が $1/N$ であるような W の各列ベクトルは、H の固有値 1 の固有ベクトルになっている。したがって、$HW = W$ が成り立つことがわかる。さらに、

$$(I-H)(H-W) = H - W - H^2 + HW = H - W - H + W = 0$$

となる。ただし、8. (a) の $H^2 = H$ を用いた。

(b)

$$ESS = \|\hat{y} - \bar{y}\|^2 = \|Hy - Wy\|^2 = \|(H-W)y\|^2,$$
$$TSS = \|y - \bar{y}\|^2 = \|Iy - Wy\|^2 = \|(I-W)y\|^2$$

よって、題意は示された。

(c)

$$ESS = \|(H-W)y\|^2 = \|(H-W)X\beta + (H-W)\epsilon\|^2$$

$RSS = \|(I-H)\epsilon\|^2$ との独立性を示すためには、ともに正規分布にしたがう $(I-H)\epsilon$ と $(H-W)\epsilon$ の独立性を示せばよく、これらの共分散行列は、(a) より、

$$\begin{aligned} E\left[(I-H)\epsilon\{(H-W)\epsilon\}^T\right] &= E\left[(I-H)\epsilon\epsilon^T(H-W)\right] = (I-H)E\left[\epsilon\epsilon^T\right](H-W) \\ &= (I-H)\sigma^2 I(H-W) = \sigma^2(I-H)(H-W) = 0 \end{aligned}$$

となるので、題意の独立性が示された。

(d) (a) を用いると、

$$\begin{aligned} &\|(I-W)y\|^2 = \|(I-H)y + (H-W)y\|^2 \\ =&\|(I-H)y\|^2 + \|(H-W)y\|^2 + 2\{(I-H)y\}^T(H-W)y \\ =&\|(I-H)y\|^2 + \|(H-W)y\|^2 \quad + 2y^T(I-H)(H-W)y = \|(I-H)y\|^2 + \|(H-W)y\|^2 \end{aligned}$$

よって、題意は示された。

15. (a) $\hat{y} - \bar{y}$ の第 i 成分 $\hat{y}_i - \bar{y}_i$ について、

$$\hat{y}_i - \bar{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i - \hat{\beta}_0 - \hat{\beta}_1 \bar{x} = \hat{\beta}_1 (x_i - \bar{x})$$

が成り立つ。したがって、

$$\hat{y} - \bar{y} = \hat{\beta}_1 (x - \bar{x})$$

となる。 ただし、上式における $\bar{x} \in \mathbb{R}^N$ は、全成分が $\frac{1}{N}\sum_{i=1}^N x_i$ であるような列ベクトルである。

(b) (a) を用いると、

$$R^2 = \frac{ESS}{TSS} = \frac{\|\hat{y} - \bar{y}\|^2}{\|y - \bar{y}\|^2} = \frac{\hat{\beta}_1^2 \|x - \bar{x}\|^2}{\|y - \bar{y}\|^2}$$

が成り立つことがわかる。

(c) (b) より、

$$R^2 = \left\{ \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \right\}^2 \frac{\sum_i (x_i - \bar{x})^2}{\sum_i (y_i - \bar{y})^2}$$

一方で、標本相関係数 $\hat{r}$ は、

$$\hat{r} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

であるから、

$$\hat{r}^2 = \frac{\{\sum_i (x_i - \bar{x})(y_i - \bar{y})\}^2}{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2} = \left\{ \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \right\}^2 \frac{\sum_i (x_i - \bar{x})^2}{\sum_i (y_i - \bar{y})^2}$$

これは、R^2 に一致している。

17. (a) x_* が定数であることに注意する。

$$E[x * \hat{\beta}] = x_* E[\hat{\beta}]$$

より、7.(c) も合わせて考えると、

$$V\left[x_*\hat{\beta}\right] = E\left[\left\{x_*(\hat{\beta}-\beta)\right\}^T x_*(\hat{\beta}-\beta)\right] = x_* V(\hat{\beta}) {x_*}^T = \sigma^2 x_* \left(X^T X\right)^{-1} x_*^T$$

が成り立つ。

(b)

$$\frac{x_*\hat{\beta}-x_*\beta}{SE\left(x_*\hat{\beta}\right)} = \frac{x_*\hat{\beta}-x_*\beta}{\hat{\sigma}\sqrt{x_*\left(X^TX\right)^{-1}x_*^T}} = \frac{x_*\hat{\beta}-x_*\beta}{\sqrt{\frac{RSS}{N-p-1}}\sqrt{x_*\left(X^TX\right)^{-1}x_*^T}}$$

$$= \frac{x_*\hat{\beta}-x_*\beta}{\sqrt{\frac{RSS}{N-p-1}}\sqrt{x_*\left(X^TX\right)^{-1}x_*^T}} = \frac{x_*(\hat{\beta}-\beta)}{\sigma\sqrt{x_*\left(X^TX\right)^{-1}x_*^T}} \Big/ \sqrt{\frac{RSS/\sigma^2}{N-p-1}}$$

このとき、$\hat{\beta} \sim N\left(\beta, \sigma^2\left(X^TX\right)^{-1}\right)$ より、(分子) $\sim N(0,1)$ と なる。 一方で、11.(b) より、$RSS/\sigma^2 \sim \chi^2_{N-p-1}$、さらにこれ ら 2 つは独立であることも示したので、

$$\frac{x_*\hat{\beta}-x_*\beta}{SE\left(x_*\hat{\beta}\right)} \sim t_{N-p-1}$$

となることがわかる。

(c)

$$V\left[x_*\hat{\beta}-y_*\right] = \sigma^2 x_*\left(X^TX\right)^{-1}x_*^T + \sigma^2 = \sigma^2\left\{1 + x_*\left(X^TX\right)^{-1}x_*^T\right\}$$

より、

$$\frac{x_*\hat{\beta}-y_*}{\hat{\sigma}\sqrt{1+x_*\left(X^TX\right)^{-1}x_*^T}} = \frac{x_*\hat{\beta}-y_*}{\hat{\sigma}\sqrt{1+x_*\left(X^TX\right)^{-1}x_*^T}} \Big/ \sqrt{\frac{RSS/\sigma^2}{N-p-1}}$$

の、(分子) $\sim N(0,1), RSS/\sigma^2 \sim \chi^2_{N-p-1}$ より、

$$\frac{x_*\hat{\beta}-y_*}{\hat{\sigma}\sqrt{1+x_*\left(X^TX\right)^{-1}x_*^T}} \sim t_{N-p-1}$$

を得る。

第 2 章 分類

19. $f(y) = \dfrac{1}{1+e^{-y(\beta_0+x^T\beta)}}$ について、

$$f(-1) = \frac{1}{1+e^{-(-1)(\beta_0+x^T\beta)}} = \frac{1}{1+e^{(\beta_0+x^T\beta)}} = P(Y=-1)$$

$$f(1) = \frac{1}{1+e^{-1(\beta_0+x^T\beta)}} = \frac{1}{1+e^{-(\beta_0+x^T\beta)}} = \frac{e^{\beta_0+x^T\beta}}{e^{\beta_0+x^T\beta}+1} = P(Y=1)$$

よって、題意は示された。

20. $f(x)=\dfrac{1}{1+e^{-(\beta_0+x\beta)}}$ に対して、$f'(x)=\dfrac{\beta e^{-(\beta_0+x\beta)}}{\left\{1+e^{-(\beta_0+x\beta)}\right\}^2}$

$$f''(x)=\frac{-\beta^2 e^{-(\beta_0+x\beta)}\left\{1+e^{-(\beta_0+x\beta)}\right\}+2\beta^2 e^{-2(\beta_0+x\beta)}}{\left\{1+e^{-(\beta_0+x\beta)}\right\}^3}=\frac{\beta^2 e^{-(\beta_0+x\beta)}\left\{-1+e^{-(\beta_0+x\beta)}\right\}}{\left\{1+e^{-(\beta_0+x\beta)}\right\}^3}$$

となり、$\beta>0$ より、任意の $x\in\mathbb{R}$ に対して $f'(x)>0, x<-\beta_0/\beta$ で $f''(x)>0, x>-\beta_0/\beta$ で $f''(x)<0$, となる。したがって、$f(x)$ は、任意の $x\in\mathbb{R}$ で単調増加し、$x<-\beta_0/\beta$ で下に凸、 $x>-\beta_0/\beta$ で上に凸となる。題意の処理を行った結果を略解 (R プログラム第 2 章) に示した。これは $\beta_0=0$ として、 β のをそれぞれ $0, 0.2, 0.5, 1, 2, 10$ としたときの $y=f(x)$ のグラフを出力したものであるが、β のが大きくなればなるほど、$x=0$ 付近で $y=-1$ から $(x,y)=(0,0)$ を通って、$y=1$ 付近により鋭敏に変化していることがわかる。

21. $\beta_0\in\mathbb{R}$, $\beta=\begin{pmatrix}\beta_1\\ \vdots\\ \beta_p\end{pmatrix}\in\mathbb{R}$, $(x_i,y_i)\in\mathbb{R}^p\times\{-1,1\}$, $i=1,\cdots N$, $x_i=\begin{pmatrix}x_{i1}\\ \vdots\\ x_{ip}\end{pmatrix}$,

$$X=\begin{pmatrix}1 & x_1^T\\ \vdots & \vdots\\ 1 & x_N^T\end{pmatrix}=\begin{pmatrix}1 & x_{11} & \cdots & x_{1p}\\ \vdots & \vdots & \ddots & \vdots\\ 1 & x_{N1} & \cdots & X_{Np}\end{pmatrix}\in\mathbb{R}^{N\times(p+1)}$$

($x_{i0}=0$ とみなす) に対して、

$$l\left(\beta_0,\beta\right)=\sum_{i=1}^{N}\log\left\{1+e^{-y_i\left(\beta_0+x_i^T\beta\right)}\right\}=\sum_{i=1}^{N}\log\left[1+\exp\left\{-y_i\sum_{k=0}^{p}\left(x_{ik}\beta_k\right)\right\}\right]$$

となる。このとき、$j=0,1,\cdots$ に対して、

$$\frac{\partial l\left(\beta_0,\beta\right)}{\partial\beta_j}=\sum_{i=1}^{n}\frac{-\left(x_{ij}y_i\right)\exp\left\{-y_i\sum_{k=0}^{p}\left(x_{ik}\beta_k\right)\right\}}{1+\exp\left\{-y_i\sum_{k=0}^{p}\left(x_{ik}\beta_k\right)\right\}}=-\sum_{i=1}^{n}\frac{y_iv_i}{1+v_i}x_{ij}$$

が成り立つ。 ただし、$i=1,2,\cdots,N$ に対して、$v_i=\exp\left\{-y_i\sum_{k=0}^{p}\left(x_{ik}\beta_k\right)\right\}$ とした。したがって、

$$\nabla l\left(\beta_0,\beta\right)=\begin{pmatrix}\frac{\partial l(\beta_0,\beta)}{\partial\beta_0}\\ \vdots\\ \frac{\partial l(\beta_0,\beta)}{\partial\beta_p}\end{pmatrix}\in\mathbb{R}^{p+1}$$

は、$u=\begin{pmatrix}\frac{y_1v_1}{1+v_1}\\ \vdots\\ \frac{y_Nv_N}{1+v_N}\end{pmatrix}$ を用いて、$\nabla l\left(\beta_0,\beta\right)=-X^Tu$ と書ける。 さらに、$i=1,2,\cdots,j=0,1,\cdots,p$ に対して、

$$\frac{\partial v_i}{\partial\beta_j}=-y_ix_{ij}v_i$$

が成り立つことにも注意すると、$j,k=0,1,\cdots,p$ に対して,

$$\begin{aligned}\frac{\partial^2 l(\beta_0,\beta)}{\partial\beta_j\beta_k} &= -\frac{\partial}{\partial\beta_k}\sum_{i=1}^{n}\frac{y_i v_i}{1+v_i}x_{ij} = -\sum_{i=1}^{n}x_{ij}\frac{\partial}{\partial\beta_k}\frac{y_i v_i}{1+v_i}\\ &= -\sum_{i=1}^{n}x_{ij}\frac{\left(-y_i^2 x_{ik}v_i\right)(1+v_i)-(y_i v_i)\left(-y_i x_{ik}v_i\right)}{(1+v_i)^2}\\ &= -\sum_{i=1}^{n}x_{ij}\frac{(-x_{ik}v_i)(1+v_i)+x_{ik}v_i^2}{(1+v_i)^2} = \sum_{i=1}^{n}x_{ij}x_{ik}\frac{v_i}{(1+v_i)^2}\end{aligned}$$

ただし、途中で $y_i\in\{-1,1\}$ より、$y_i^2=1$ となることを用いた。 このとき、W を、第 (i,i) 成分が $v_i/(1+v_i)^2$ となるような N 次の対角行列、すなわち

$$W=\begin{bmatrix}\frac{v_1}{(1+v_1)^2} & 0 & \cdots & 0\\ 0 & \frac{v_2}{(1+v_2)^2} & \ddots & \vdots\\ \vdots & \ddots & \ddots & 0\\ 0 & \cdots & 0 & \frac{v_N}{(1+v_N)^2}\end{bmatrix}\in\mathbb{R}^{N\times N}$$

とすると、$WX\in\mathbb{R}^{N\times p+1}$ の第 (i,k) 成分 $(i=1,\cdots,N,k=0,1,\cdots,p)$ は、

$$\frac{v_i}{(1+v_i)^2}x_{ik}$$

となるから、$X^TWX\in\mathbb{R}^{(p+1)\times(p+1)}$ の第 (j,k) 成分 $(j,k=0,1,\cdots,p)$ は、

$$\sum_{i=1}^{N}x_{ji}\frac{v_i}{(1+v_i)^2}x_{ik} = \sum_{i=1}^{n}x_{ij}x_{ik}\frac{v_i}{(1+v_i)^2} = \frac{\partial^2 l(\beta_0,\beta)}{\partial\beta_j\beta_k}$$

が成り立つ。したがって、求める 2 階微分 $\nabla^2 l(\beta_0,\beta)$ は、

$$\nabla^2 l(\beta_0,\beta) = X^TWX$$

と表される。ここで、任意の $i=1,\cdots,N$ について、 $v_i>0$ より、 $v_i/(1+v_i)^2>0$ となるので、$U\in\mathbb{R}^{N\times N}$ の各成分を、W の各成分の平方根としたもの、すなわち

$$U=\begin{bmatrix}\sqrt{\frac{v_1}{(1+v_1)^2}} & 0 & \cdots & 0\\ 0 & \sqrt{\frac{v_2}{(1+v_2)^2}} & \ddots & \vdots\\ \vdots & \ddots & \ddots & 0\\ 0 & \cdots & 0 & \sqrt{\frac{v_N}{(1+v_N)^2}}\end{bmatrix}$$

とすれば、 $W=U^TU$ と書けるので、$\nabla^2 l(\beta_0,\beta)=X^T\left(U^TU\right)X=(UX)^TUX$ となる。したがって、 命題 10 の 1. ならば 3. が成り立つを用いれば、$\nabla^2 l(\beta_0,\beta)$ は非負定値行列で あることがわかるので、$l(\beta_0,\beta)$ は凸である。

23. 更新規則を $\beta_{\text{old}},\beta_{\text{new}},u,W,X$ を用いて書き直すと、

$$\beta_{\text{new}} \leftarrow \beta_{\text{old}} + \left(X^TWX\right)^{-1}X^Tu$$

となる。このとき、右辺を変形していくと、

$$\begin{aligned}&\beta_{\text{old}} + \left(X^T W X\right)^{-1} X^T u = \left(X^T W X\right)^{-1} X^T W X \beta_{\text{old}} + \left(X^T W X\right)^{-1} X^T u \\ &= \left(X^T W X\right)^{-1} X^T \left(W X \beta_{\text{old}} + u\right) = \left(X^T W X\right)^{-1} X^T W \left(X \beta_{\text{old}} + W^{-1} u\right)\end{aligned}$$

したがって、$z = X\beta_{\text{old}} + W^{-1}u$ とすれば、更新規則は

$$\beta_{\text{new}} \leftarrow \left(X^T W X\right)^{-1} X^T W z$$

と表される。

25. 尤度 $\displaystyle\prod_{i=1}^{N} \frac{1}{1+\exp\left\{-y_i\left(\beta_0+\beta^T x_i\right)\right\}}$ の最大化を考えるとき、任意の $i = 1, \cdots, N$ に対して $y_i\left(\beta_0 + \beta^T x_i\right) \geq 0$ が成り立っているならば、 任意の β_0, β を固定した とき、たとえば $\beta_0 \leftarrow 2\beta_0, \beta \leftarrow 2\beta$ と置き換えることで、 (2.1) の指数部分をより小さくすることができる。したがって、 最大値

$$\max_{\beta_0,\beta} \prod_{i=1}^{N} \frac{1}{1+\exp\left\{-y_i\left(\beta_0+\beta^T x_i\right)\right\}}$$

は存在しないから、 仮定を満足しているもとでは、 ロジスティック回帰の最尤推定のパラメータを見出すことはできない。

26. 正答率は $(39+42)/100 = 0.81$ となった。

27.

$$S_{k,l} = \left\{ x \in \mathbb{R}^p \mid \frac{\pi_k f_k(x)}{\sum_{j=1}^K \pi_j f_j(x)} = \frac{\pi_l f_l(x)}{\sum_{j=1}^K \pi_j f_j(x)} \right\}$$

$$f_k(x) = \frac{1}{\sqrt{(2\pi)^p \det \Sigma_k}} \exp\left\{ -\frac{1}{2}\left(x-\mu_k\right)^T \Sigma_k^{-1}\left(x-\mu_k\right) \right\}$$

(a) $\pi_k = \pi_l$ を仮定すると、

$$\begin{aligned}&\frac{\pi_k f_k(x)}{\sum_{j=1}^K \pi_j f_j(x)} = \frac{\pi_l f_l(x)}{\sum_{j=1}^K \pi_j f_j(x)} \\ \Longleftrightarrow &f_k(x) = f_l(x) \\ \Longleftrightarrow &\frac{1}{\sqrt{(2\pi)^p \det \Sigma_k}} \exp\left\{ -\frac{1}{2}\left(x-\mu_k\right)^T \Sigma_k^{-1}\left(x-\mu_k\right) \right\} \\ &= \frac{1}{\sqrt{(2\pi)^p \det \Sigma_l}} \exp\left\{ -\frac{1}{2}\left(x-\mu_l\right)^T \Sigma_l^{-1}\left(x-\mu_l\right) \right\} \\ \Longleftrightarrow &\sqrt{\frac{\det \Sigma_k}{\det \Sigma_l}} = \exp\left\{ \frac{1}{2}\left\{ -\left(x-\mu_k\right)^T \Sigma_k^{-1}\left(x-\mu_k\right) + \left(x-\mu_l\right)^T \Sigma_l^{-1}\left(x-\mu_l\right) \right\} \right.\end{aligned}$$

上式で両辺の対数を取っても必要十分なので、

$$\log \frac{\det \Sigma_k}{\det \Sigma_l} = -\left(x-\mu_k\right)^T \Sigma_k^{-1}\left(x-\mu_k\right) + \left(x-\mu_l\right)^T \Sigma_l^{-1}\left(x-\mu_l\right)$$

よって、 $S_{k,l}$ が題意の形で与えられることが示された。

(b) $\Sigma_k = \Sigma_l = \Sigma$ を仮定すると、

$$\begin{aligned}
&\frac{\pi_k f_k(x)}{\sum_{j=1}^K \pi_j f_j(x)} = \frac{\pi_l f_l(x)}{\sum_{j=1}^K \pi_j f_j(x)} \\
\Longleftrightarrow &\pi_k f_k(x) = \pi_l f_l(x) \\
\Longleftrightarrow &\pi_k \exp\left\{-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)\right\} = \pi_l \exp\left\{-\frac{1}{2}(x-\mu_l)^T \Sigma^{-1}(x-\mu_l)\right\} \\
\Longleftrightarrow &\frac{\pi_k}{\pi_l} = \exp\left\{\frac{1}{2}\left\{-(x-\mu_l)^T \Sigma^{-1}(x-\mu_l) + (x-\mu_k)^T \Sigma^{-1}(x-\mu_k)\right\}\right\}
\end{aligned}$$

上式で両辺の対数を取っても必要十分なので、

$$\log\frac{\pi_k}{\pi_l} = \frac{1}{2}\left\{-(x-\mu_l)^T \Sigma^{-1}(x-\mu_l) + (x-\mu_k)^T \Sigma^{-1}(x-\mu_k)\right\}$$

ここで、Σ が共分散行列であるから対称行列となるので、

$$\Sigma\Sigma^{-1} = I \Longleftrightarrow \left(\Sigma\Sigma^{-1}\right)^T = I \Longleftrightarrow \left(\Sigma^{-1}\right)^T \Sigma^T = I \Longleftrightarrow \left(\Sigma^{-1}\right)^T \Sigma = I$$

したがって、$\Sigma^{-1}\,(=(s_{ij})$ としておく $)$ も対称行列となることに注意すると、

$$\begin{aligned}
\log\frac{\pi_k}{\pi_l} &= \frac{1}{2}\sum_{i,j} s_{ij}\left\{(x_i-\mu_{ki})(x_j-\mu_{kj}) - (x_i-\mu_{li})(x_j-\mu_{lj})\right\} \\
&= \frac{1}{2}\sum_{i,j} s_{ij}\left\{(x_i-\mu_{ki})(x_j-\mu_{kj}) - (x_i-\mu_{li})(x_j-\mu_{lj})\right\} \\
&= \frac{1}{2}\sum_{i,j} s_{ij}\left\{x_i(-\mu_{kj}+\mu_{lj}) + x_j(-\mu_{ki}+\mu_{li}) + \mu_{ki}\mu_{kj} - \mu_{li}\mu_{lj}\right\} \\
&= \frac{1}{2}\sum_{i,j} s_{ij}\left\{2x_i(-\mu_{kj}+\mu_{lj}) + (\mu_{ki}\mu_{kj} - \mu_{li}\mu_{lj})\right\} \\
&= (\mu_l-\mu_k)^T \Sigma^{-1} x + \frac{1}{2}(\mu_k-\mu_l)^T \Sigma^{-1}(\mu_k-\mu_l)\}
\end{aligned}$$

$$(\mu_k-\mu_l)^T \Sigma^{-1} x - \frac{1}{2}(\mu_k-\mu_l)^T \Sigma^{-1}(\mu_k-\mu_l) + \log\frac{\pi_k}{\pi_l} = 0$$

から、求める a, b は

$$\begin{aligned}
a &= \left\{(\mu_k-\mu_l)^T \Sigma^{-1}\right\}^T = \Sigma^{-1}(\mu_k-\mu_l) \\
b &= -\frac{1}{2}(\mu_k-\mu_l)^T \Sigma^{-1}(\mu_k-\mu_l) + \log\frac{\pi_k}{\pi_l}
\end{aligned}$$

となる。

(c) (b) で得た $\Sigma_k = \Sigma_l$ の場合の平面の式で、さらに $\pi_k = \pi_l$ とすれば、

$$(\mu_k-\mu_l)^T \Sigma^{-1} x - \tfrac{1}{2}(\mu_k-\mu_l)^T \Sigma^{-1}(\mu_k-\mu_l) = 0 \Leftrightarrow (\mu_k-\mu_l)^T \Sigma^{-1}\left(x - \tfrac{\mu_k-\mu_l}{2}\right) = 0$$

となるので、境界は平面 $x = (\mu_k-\mu_l)/2$ となる。

第 3 章 リサンプリング

32. $(A+UCV)\left(A^{-1}-A^{-1}U\left(C^{-1}+VA^{-1}U\right)^{-1}VA^{-1}\right)=I$ を示せば十分。

$$
\begin{aligned}
&(A+UCV)\left(A^{-1}-A^{-1}U\left(C^{-1}+VA^{-1}U\right)^{-1}VA^{-1}\right)\\
=&I+UCVA^{-1}-U\left(C^{-1}+VA^{-1}U\right)^{-1}VA^{-1}-UCVA^{-1}U\left(C^{-1}+VA^{-1}U\right)^{-1}VA^{-1}\\
=&I+UCVA^{-1}-UC\cdot C^{-1}\cdot\left(C^{-1}+VA^{-1}U\right)^{-1}VA^{-1}\\
&-UC\cdot VA^{-1}U\cdot\left(C^{-1}+VA^{-1}U\right)^{-1}VA^{-1}\\
=&I+UCVA^{-1}-UC\cdot\left(C^{-1}+VA^{-1}U\right)\cdot\left(C^{-1}+VA^{-1}U\right)^{-1}VA^{-1}\\
=&I+UCVA^{-1}-UCVA^{-1}=I
\end{aligned}
$$

よって、題意は示された。

33. (a)

$$
X=\begin{pmatrix}x_1\\ \vdots\\ x_N\end{pmatrix}\in\mathbb{R}^{N\times(p+1)}
$$

に対して、

$$
X^TX=\sum_{i=1}^{N}x_i^Tx_i=\sum_{i\in S}x_i^Tx_i+\sum_{i\notin S}x_i^Tx_i=X_S^TX_S+X_{-S}^TX_{-S}
$$

が成り立つことに注意する。 32. で示した等式で、

$$
A=X^TX,\ U=X_S^T,\ V=-X_S,\ C=I
$$

とすれば、

$$
\begin{aligned}
\left\{X^TX-X_S^TX_S\right\}^{-1}&=\left(X^TX\right)^{-1}+\left(X^TX\right)^{-1}X_S^T\left\{I-X_S\left(X^TX\right)^{-1}X_S^T\right\}^{-1}X_S\left(X^TX\right)^{-1}\\
\left(X_{-s}^TX_{-s}\right)^{-1}&=\left(X^TX\right)^{-1}+\left(X^TX\right)^{-1}X_S^T\left(I-H_S\right)^{-1}X_S\left(X^TX\right)^{-1}
\end{aligned}
$$

ただし、$H_S=X_S\left(X^TX\right)^{-1}X_S^T$ である。 よって、題意は示さ れた。

(b)

$$
\begin{aligned}
\hat{\beta}_{-S}&=\left(X_{-S}^TX_{-S}\right)^{-1}X_{-S}^Ty_{-S}\\
&=\left\{\left(X^TX\right)^{-1}+\left(X^TX\right)^{-1}X_S^T\left(I-H_S\right)^{-1}X_S\left(X^TX\right)^{-1}\right\}\left(X^Ty-X_S^Ty_S\right)\\
&=\hat{\beta}-\left(X^TX\right)^{-1}X_S^Ty_S+\left(X^TX\right)^{-1}X_S^T\left(I-H_S\right)^{-1}\left(X_S\hat{\beta}-H_Sy_S\right)\\
&=\hat{\beta}-\left(X^TX\right)^{-1}X_S^T\left(I-H_S\right)^{-1}\left\{\left(I-H_S\right)y_S-X_S\hat{\beta}+H_Sy_S\right\}\\
&=\hat{\beta}-\left(X^TX\right)^{-1}X_S^T\left(I-H_S\right)^{-1}\left(y_S-X_S\hat{\beta}\right)\\
&=\hat{\beta}-\left(X^TX\right)^{-1}X_S^T\left(I-H_S\right)^{-1}\left(y_S-\hat{y}\right)\\
&=\hat{\beta}-\left(X^TX\right)^{-1}X_S^T\left(I-H_S\right)^{-1}e_S
\end{aligned}
$$

よって、題意は示された。

34.

$$\begin{aligned} y_S - X_S\hat{\beta}_{-S} &= y_S - X_S\left\{\hat{\beta} - \left(X^TX\right)^{-1}X_S^T\left(I - H_S\right)^{-1}e_S\right\} \\ &= y_S - X_S\hat{\beta} + X_S\left(X^TX\right)^{-1}X_S^T\left(I - H_S\right)^{-1}e_S \\ &= e_S + H_S\left(I - H_S\right)^{-1}e_S \\ &= \left(I - H_S\right)\left(I - H_S\right)^{-1}e_S + H_S\left(I - H_S\right)^{-1}e_S \\ &= \left(I - H_S\right)^{-1}e_S \end{aligned}$$

したがって、 CV の全グループの二乗誤差の和は、

$$\sum_S \left\|y_S - X_S\hat{\beta}_{-S}\right\|^2 = \sum_S \left\|\left(I - H_S\right)^{-1}e_S\right\|^2$$

と書くことができる。よって、題意は示された。

35. 二乗誤差の和は一致しており、cv.fast の実行時間 のほうが短いことがわかる。

39. 最初の 3 種類のデータは. $j = 1, 2, 3$ に対して、第 1 変数を第 3-第 4 変数に回帰したときの切片および 2 個の傾きの推定値を求め、その推定値の標準偏差を評価したものである。

第 4 章 情報量規準

40. (a)

$$\begin{aligned} \max_{\beta\in\mathbb{R}^{p+1}} l &= \max_{\beta\in\mathbb{R}^{p+1}} \sum_{i=1}^N \log f\left(y_i \mid x_i, \beta\right) = \max_{\beta\in\mathbb{R}^{p+1}} \sum_{i=1}^N \left\{-\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{\|y_i - x_i\beta\|^2}{2\sigma^2}\right\} \\ &= -\frac{N}{2}\log\left(2\pi\sigma^2\right) - \min_{\beta\in\mathbb{R}^{p+1}} \sum_{i=1}^N \left\{\frac{\|y_i - x_i\beta\|^2}{2\sigma^2}\right\} \\ &= -\frac{N}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\min_{\beta\in\mathbb{R}^{p+1}} \|y - X\beta\|^2 \end{aligned}$$

したがって、$\sigma^2 > 0$ が既知であるとき、β について l を最大化することは、$\|y - X\beta\|^2$ を最小化することと同値となる、

(b) l を σ^2 で偏微分すると、

$$\frac{\partial l}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\left(\sigma^2\right)^2}\|y - X\beta\|^2$$

このとき、$\dfrac{\partial l}{\partial \sigma^2} = 0$ とするような、σ^2 の最尤推定量 $\hat{\sigma}^2$ は、$\hat{\beta} = \left(X^TX\right)^{-1}X^Ty$ を用いて、

$$\hat{\sigma}^2 = \frac{1}{N}\|y - X\hat{\beta}\|^2$$

で与えられる。

(c) 任意の $x > 0$ に対して、 $\log x \leq x - 1$ となることを用いると、任意の $\mathbb{R}$ 上の確率密度関数 f, g に対して、

$$\begin{aligned} &\int_{-\infty}^{\infty} f(x)\log\frac{f(x)}{g(x)}dx = -\int_{-\infty}^{\infty} f(x)\log\frac{g(x)}{f(x)} \\ \geq &-\int_{-\infty}^{\infty} f(x)\left\{\frac{g(x)}{f(x)} - 1\right\}dx = -\int_{-\infty}^{\infty}\{g(x) - f(x)\}dx = -(1-1) = 0 \end{aligned}$$

最後に、f, g は確率密度関数であることを用いた。したがって、題意の不等式が成り立つ。

41. (a)

$$
\begin{aligned}
f^N(y \mid x, \beta) =& \prod_{i=1}^{N} f\left(y_i \mid x_i, \beta\right) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2\right\} \\
\frac{\partial f^N(y \mid x, \beta)}{\partial \beta_k} =& \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \frac{1}{2\sigma^2}\left(\sum_{i=1}^{N} 2x_{ik}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)\right) \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2\right\} \\
=& \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \frac{1}{\sigma^2}\left(\sum_{i=1}^{N} x_{ik}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)\right) \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2\right\} \\
=& f^N(y \mid x, \beta)\sum_{i=1}^{N}\frac{x_{ik}}{\sigma^2}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right) \\
\frac{\partial l}{\partial \beta_k} =& \sum_{i=1}^{N}\frac{\partial}{\partial \beta_k}\log f\left(y_i \mid x_i, \beta\right) \\
=& \sum_{i=1}^{N}\frac{\partial}{\partial \beta_k}\left\{-\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2\right\} \\
=& \sum_{i=1}^{N}\frac{\partial}{\partial \beta_k}\left\{-\frac{1}{2\sigma^2}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2\right\} = \sum_{i=1}^{N}\left\{\frac{x_{ik}}{\sigma^2}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)\right\} \\
=& \frac{\partial f^N(y \mid x, \beta)/\partial \beta_k}{f^N(y \mid x, \beta)}
\end{aligned}
$$

したがって、

$$\nabla l = \frac{\nabla f^N(y \mid x, \beta)}{f^N(y \mid x, \beta)}$$

(b) $f^N(y \mid x, \beta)$ は同時確率密度関数なので、

$$\int f^N(y \mid x, \beta)dy = 1$$

が成り立つ。このとき、 β による微分と y による積分の順序が 可換であると仮定すると、 上式の両辺を β で偏微分して、

$$\int \nabla f^N(y \mid x, \beta)dy = 0$$

が成立する。

(c)

$$E[\nabla l] = \int \frac{\nabla f^N(y \mid x, \beta)}{f^N(y \mid x, \beta)} f^N(y \mid x, \beta)dy = \int \nabla f^N(y \mid x, \beta)dy = 0$$

(d) (c) で得た等式の両辺を β で偏微分すると、

$$\begin{aligned}0 =&\nabla(E[\nabla l]) = \nabla \int (\nabla l) f^N(y \mid x, \beta) dy = \int \nabla \left\{ (\nabla l) f^N(y \mid x, \beta) \right\} dy \\ =&\int \left(\nabla^2 l\right) f^N(y \mid x, \beta) dy + \int (\nabla l) \nabla f^N(y \mid x, \beta) dy \\ =&E\left[\nabla^2 l\right] + \int (\nabla l)^2 f^N(y \mid x, \beta) dy = E\left[\nabla^2 l\right] + E\left[(\nabla l)^2\right]\end{aligned}$$

よって、題意は示された。したがって、(d) より、

$$\frac{1}{N} E\left[(\nabla l)^2\right] = -\frac{1}{N} E\left[\nabla^2 l\right]$$

が成り立つ.

42. (a) β の不偏推定量 $\tilde{\beta}$ について、

$$\int \tilde{\beta}_i f^N(y \mid x, \beta) dy = \beta_i$$

が成り立つので、この両辺を β_j で偏微分すると.

$$\int \tilde{\beta}_i \frac{\partial}{\partial \beta_j} f^N(y \mid x, \beta) dy = \int \tilde{\beta}_i f^N(y \mid x, \beta) (\nabla l) dy = \begin{cases} 1, & (i = j) \\ 0 & (i \neq j) \end{cases}$$

これを、共分散行列の形で表すと、

$$E\left[\tilde{\beta} (\nabla l)^T\right] = \int \tilde{\beta} \left\{ \frac{\nabla f^N(y \mid x, \beta)}{f^N(y \mid x, \beta)} \right\}^T f^N(y \mid x, \beta) dy = I$$

したがって、$E\left[(\tilde{\beta} - \beta)(\nabla l)^T\right] = I$。最後に、$E[\nabla l] = 0$ であることを用いた。

(b) 求める共分散行列は、

$$\begin{bmatrix} E\left[(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)^T\right] & E\left[(\nabla l)(\tilde{\beta} - \beta)^T\right] \\ E\left[(\tilde{\beta} - \beta)(\nabla l)^T\right] & E\left[(\nabla l)^2\right] \end{bmatrix} = \begin{bmatrix} V(\tilde{\beta}) & I \\ I & NJ \end{bmatrix}$$

(c)

$$\begin{bmatrix} V(\tilde{\beta}) - (NJ)^{-1} & 0 \\ 0 & NJ \end{bmatrix} = \begin{bmatrix} I & -(NJ)^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} V(\tilde{\beta}) & I \\ I & NJ \end{bmatrix} \begin{bmatrix} I & 0 \\ -(NJ)^{-1} & I \end{bmatrix}$$

これらは、非負定値行列となるので、任意の $x, y \in \mathbb{R}^{p+1}$ に対して、$z = [x, y]^T$ として、

$$z^T \begin{bmatrix} V(\tilde{\beta}) - (NJ)^{-1} & 0 \\ 0 & NJ \end{bmatrix} z = x^T \left\{ V(\tilde{\beta}) - (NJ)^{-1} \right\} x + y^T (NJ) y \geq 0$$

が成り立つ。このとき、$y = 0$ でも成り立つので、$V(\tilde{\beta}) - (NJ)^{-1}$ は非負定値である。したがって、Cramer-Rao の不等式の成立が示された。

43. (a) $E\left[(\tilde{\beta} - \beta)(\nabla l)^T\right] = I \in \mathbb{R}^{(p+1)\times(p+1)}$ の、両辺のトレースを取って、

$$p + 1 = \mathrm{tr}\left\{ E\left[(\tilde{\beta} - \beta)(\nabla l)^T\right] \right\} = \mathrm{tr}\left\{ E\left[(\nabla l)^T (\tilde{\beta} - \beta)\right] \right\} = E\left[(\tilde{\beta} - \beta)^T (\nabla l)\right]$$

よって、題意は示された。

(b)

$$E\left[\left\|X\left(X^TX\right)^{-1}\nabla l\right\|^2\right] = \mathrm{tr}\left\{E\left[(\nabla l)^T\left(X^TX\right)^{-1}X^TX\left(X^TX\right)^{-1}(\nabla l)\right]\right\}$$
$$= \mathrm{tr}\left\{E\left[(\nabla l)^T\left(X^TX\right)^{-1}(\nabla l)\right]\right\} = \mathrm{tr}\left\{E\left[\left(X^TX\right)^{-1}(\nabla l)(\nabla l)^T\right]\right\}$$
$$= \mathrm{tr}\left\{\left(X^TX\right)^{-1}E\left[(\nabla l)(\nabla l)^T\right]\right\} = \mathrm{tr}\left\{\left(X^TX\right)^{-1}\frac{1}{\sigma^2}X^TX\right\} = \frac{1}{\sigma^2}\,\mathrm{tr}\left\{I_{p+1}\right\} = \frac{p+1}{\sigma^2}$$

(c)

$$\left\{E\left[(\tilde\beta-\beta)^T\nabla l\right]\right\}^2 = \left\{E\left[(\tilde\beta-\beta)^TX^TX\left(X^TX\right)^{-1}\nabla l\right]\right\}^2$$
$$\le E\left[\left\|(\tilde\beta-\beta)^TX^T\right\|^2\right]E\left[\left\|X\left(X^TX\right)^{-1}\nabla l\right\|^2\right] = E\left[\|X(\tilde\beta-\beta)\|^2\right]E\left[\left\|X\left(X^TX\right)^{-1}\nabla l\right\|^2\right]$$

ただし、2 行目から 3 行目への変形で、Schwartz の不等式を用いた。したがって、

$$(p+1)^2 \le \frac{p+1}{\sigma^2}E\left[\|X(\tilde\beta-\beta)\|^2\right]E\left[\|X(\tilde\beta-\beta)\|^2\right] \ge (p+1)\sigma^2$$

44. (a)

$$\log f(u\mid x,\gamma) = -\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}(u-x\gamma)^2(u-x\gamma)^2$$
$$\{(u-x\beta)-x(\gamma-\beta)\}^2 = (u-x\beta)^2 - 2(\gamma-\beta)^Tx^T(u-x\beta) + (\gamma-\beta)^Tx^Tx(\gamma-\beta)$$

より、

$$\log f(u\mid x,\gamma)$$
$$= -\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}(u-x\beta)^2 + \frac{1}{\sigma^2}(\gamma-\beta)^Tx^T(u-x\beta) - \frac{1}{2\sigma^2}(\gamma-\beta)^Tx^Tx(\gamma-\beta)$$

となる。これを、 $(x,u) = (x_1,z_1),\cdots,(x_N,z_N)$ として和を取ると、

$$-\sum_{i=1}^N \log f\left(z_i\mid x_i,\gamma\right)$$
$$= \frac{N}{2}\log 2\pi\sigma^2 + \frac{1}{2\sigma^2}\sum_{i=1}^N\left(z_i-x_i\beta\right)^2 - \frac{1}{\sigma^2}\sum_{i=1}^N(\gamma-\beta)^Tx_i^T\left(z_i-x_i\beta\right) + \frac{1}{2\sigma^2}\sum_{i=1}^N(\gamma-\beta)^Tx^Tx(\gamma-\beta)$$
$$= \frac{N}{2}\log 2\pi\sigma^2 + \frac{1}{2\sigma^2}\|z-X\beta\|^2 \quad - \frac{1}{\sigma^2}(\gamma-\beta)^TX^T(z-X\beta) + \frac{1}{2\sigma^2}(\gamma-\beta)X^TX(\gamma-\beta)$$

となる。よって、示された。

(b) $E[z-X\beta] = 0$,

$$E\left[\|z-X\beta\|^2\right] - E_Z\left[\sum_{i=1}^N\log f\left(z_i\mid x_i,\gamma\right)\right]$$
$$= \frac{N}{2}\log 2\pi\sigma^2 + \frac{1}{2\sigma^2}N\sigma^2 + \frac{1}{2\sigma^2}\|X(\gamma-\beta)\|^2 = \frac{N}{2}\log 2\pi\sigma^2e + \frac{1}{2\sigma^2}\|X(\gamma-\beta)\|^2$$

(c) (4.8) の値は、

$$-\sum_{i=1}^N\int_{-\infty}^{\infty}\left\{\log f\left(z\mid x_i,\gamma\right)\right\}f\left(z\mid x_i,\beta\right)dz$$

として書けて、KL 情報量の和は、

$$\sum_{i=1}^{N}\int_{-\infty}^{\infty} f\left(z \mid x_i,\beta\right)\log\frac{f\left(z \mid x_i,\beta\right)}{f\left(z \mid x_i,\gamma\right)}dz = E_Z\left[\sum_{i=1}^{N}\log\frac{f\left(z \mid x_i,\beta\right)}{f\left(z \mid x_i,\gamma\right)}\right] = \frac{1}{2\sigma^2}\|X(\gamma-\beta)\|^2$$

このとき、 $\gamma=\hat{\beta}$ を最小二乗法によって求めると、

$$E\left[\|X(\hat{\beta}-\beta)\|^2\right] = E\left[\mathrm{tr}\left\{(\hat{\beta}-\beta)^T X^T X(\hat{\beta}-\beta)\right\}\right] = \mathrm{tr}\left\{V[\hat{\beta}]X^T X\right\} = \mathrm{tr}\left(\sigma^2 I\right) = (p+1)\sigma^2$$

したがって、(b) の平均的な最小値は、

$$\frac{N}{2}\log 2\pi\sigma^2 e + \frac{p+1}{2}$$

であり、その最小値は最小二乗法によって実現できる。

(d)

$$\frac{N}{2}\log\left(2\pi\sigma_k^2 e\right) + \frac{k+1}{2} = \frac{1}{2}\left(N\log\sigma_k^2 + k\right) - \frac{N}{2}\log 2\pi + \frac{N+1}{2}$$

第 2 項以降は k に依らないから、題意は示された。

45.

$$E\left[U^n\right] = \prod_{i=1}^{n}(m+2(i-1))$$

(a) $\log(x+1)$ の Maclaurin 展開、

$$\log(x+1) = \sum_{i=1}^{\infty}\frac{(-x)^i}{-i} = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots$$

を用いると、

$$\begin{aligned} E\left[\log\frac{U}{m}\right] &= E\left[\log\left(\frac{U}{m}-1+1\right)\right] \\ &= E\left[\left(\frac{U}{m}-1\right) - \frac{1}{2}\left(\frac{U}{m}-1\right)^2 + \frac{1}{3}\left(\frac{U}{m}-1\right)^3 - \cdots\right] \\ &= E\left[\frac{U}{m}-1\right] - \frac{1}{2}E\left[\left(\frac{U}{m}-1\right)^2\right] + \cdots \end{aligned}$$

となる。

(b)

$$\begin{aligned} E\left[\frac{U}{m}-1\right] &= \frac{1}{m}E[U] - 1 = \frac{1}{m}\cdot m - 1 = 0, \\ E\left[\left(\frac{U}{m}-1\right)^2\right] &= \frac{1}{m^2}E\left[(U-m)^2\right] = \frac{1}{m^2}\left\{m(m+2) - 2m^2 + m^2\right\} = \frac{2}{m} \end{aligned}$$

(c)

$$\sum_{j=0}^{n}(-1)^{n-j}\left(\begin{array}{c} n \\ j \end{array}\right) = \sum_{j=0}^{n}1^j(-1)^{n-j}\left(\begin{array}{c} n \\ j \end{array}\right) = (1-1)^n = 0$$

(d)

$$E\left[(U-m)^n\right] = \sum_{j=0}^{n}(-1)^j \left(\begin{array}{c} n \\ j \end{array} \right) m^{n-j} \prod_{i=1}^{j}(m+2(i-1))$$

であり、各 j に対する $m^{n-j}\prod_{i=1}^{j}(m+2(i-1))$ の n 次の係数は 1 である。したがって、求める n 次の係数は、

$$\sum_{j=0}^{n}(-1)^j \left(\begin{array}{c} n \\ j \end{array} \right) = \sum_{j=0}^{n}(-1)^j 1^{n-j} \left(\begin{array}{c} n \\ j \end{array} \right) = (-1+1)^n = 0$$

(e) 各 j に対する $m^{n-j}\prod_{i=1}^{j}(m+2(i-1))$ の $n-1$ 次の係数は、

$$\sum_{i=1}^{j} 2(i-1) = j(j-1)$$

である。したがって、求める $n-1$ 次の係数は、

$$\begin{aligned} &\sum_{j=0}^{n}(-1)^j \left(\begin{array}{c} n \\ j \end{array} \right) j(j-1) = \sum_{j=2}^{n} \frac{n!}{(n-j)!(j-2)!}(-1)^{n-j-2} \\ =&n(n-1)\sum_{i=0}^{n-2} \left(\begin{array}{c} n-2 \\ i \end{array} \right) (-1)^{n-2-i}1^i = n(n-1)(-1+1)^{n-2} = 0 \end{aligned}$$

(f) $n \geq 3$ に対して、(d) および (e) より、

$$E\left[(U-m)^n\right] = O\left[\frac{1}{m^2}\right]$$

が成り立つことがわかる。ここで、

$$\begin{aligned} U &= \frac{N\hat{\sigma}^2(S)}{\sigma^2(S)} \sim \chi^2_{N-k(S)-1}, \\ m &= N-k(S)-1 \end{aligned}$$

とすれば、

$$\begin{aligned} &E\left[\log\frac{U}{m}\right] = E\left[\log\left(\frac{N\hat{\sigma}^2(S)}{\sigma^2(S)}/(N-k(S)-1)\right)\right] = E\left[\log\left(\frac{\hat{\sigma}^2(S)}{N-k(S)-1}/\frac{\sigma^2(S)}{N}\right)\right] \\ =&-\frac{1}{N-k(S)-1}+O\left(\frac{1}{N^2}\right) = -\frac{1}{N}-\frac{k(S)+1}{N\{N-k(S)-1\}}+O\left(\frac{1}{N^2}\right) = -\frac{1}{N}+O\left(\frac{1}{N^2}\right) \end{aligned}$$

以上より、

$$\begin{aligned} &E\left[\log\frac{\hat{\sigma}^2(s)}{\sigma^2}\right] = E\left[\log\frac{U}{N}\right] = \log\frac{m}{N}+E\left[\log\frac{U}{m}\right] \\ &= \log\left\{1-\frac{k(S)+1}{N}\right\}-\frac{1}{N}+O\left(\frac{1}{N^2}\right) = -\frac{k(S)+2}{N}+O\left(1/N^2\right) \end{aligned}$$

となる。よって、題意は示された。

第 5 章 スパース推定

49.

$$L = \frac{1}{N}\|y - X\beta\|^2 + \lambda\|\beta\|_2^2$$

を、β で偏微分すると、

$$\frac{\partial L}{\partial \beta} = -\frac{2}{N}X^T(y - X\beta) + 2\lambda\beta = \left(-\frac{2}{N}X^TX + 2\lambda I\right)\beta - \frac{2}{N}X^Ty$$

となるので、$\dfrac{\partial L}{\partial \beta} = 0$ となるためには、

$$\left(X^TX + N\lambda I\right)\beta = X^Ty$$

このとき、解 $\beta = \hat{\beta}$ が存在するためには、$X^TX + N\lambda I$ が正則であることが必要十分となる。

まず、$\lambda > 0$ であると仮定すると、$X^TX \in \mathbb{R}^{p\times p}$ が非負定値となることから、X^TX の固有値 $\mu_1, \cdots, \mu_p$ はすべて非負。したがって、$X^TX + N\lambda I$ の固有多項式を $\varphi(t)$ とすると、

$$\varphi(t) = \det\left(X^TX + N\lambda I - tI\right) = (t - \mu_1 - N\lambda)\cdots(t - \mu_p - N\lambda)$$

となるから、$N\lambda > 0$ より、これらの根、すなわち $X^TX + N\lambda I$ の固有値はすべて非負となる。

逆に、$X^TX + N\lambda I$ が正則と仮定すると、任意の $i = 1, \cdots, p$ に対して、

$$\mu_i + N\lambda > 0$$

である。μ_i は X^TX の固有値であり、$X \in \mathbb{R}^{N\times p}$ が任意ならば、μ_i は任意の非負の値を取りうる。したがって、(49.1) が常に成り立つためには、$\lambda > 0$ が必要。よって、題意は示された。

50. (a)

$$f(x) \geq f(x_0) + z(x - x_0)$$

$x > x_0$ で (50.1) が成立するためには、

$$\frac{f(x) - f(x_0)}{x - x_0} \geq z$$

が必要となる。また、$x < x_0$ で (50.1) が成立するためには、

$$\frac{f(x) - f(x_0)}{x - x_0} \leq z$$

が必要となる。したがって、z は $x = x_0$ における f の左微分以上で、なおかつ右微分以下であることが必要となり、f が $x = x_0$ で微分可能だから、$z = f'(x_0)$ が必要となる。 逆に、$z = f'(x_0)$ となるとき、 f が凸関数であることから. (50.1) が 成り立つ。よって、 題意は示された。

(b) $zx \leq |x|$ の成立のためには、$\begin{cases} x = 0 \text{ のとき} & \text{無条件に成立} \\ x > 0 \text{ のとき} & z \leq 1 \text{ が必要} \\ x < 0 \text{ のとき} & z \geq -1 \text{ が必要。} \end{cases}$ となる。したがって、(50.2) の成立には、 $|z| \leq 1$ が必要。 逆に、$|z| \leq 1$ のとき、

$$zx \leq |z||x| \leq |x|$$

となるので、(50.2) が成立する。よって、題意は示された。

(c) i. $x_0 < 0$ のとき、 (a) より劣微分は、 $\{-1\}$

ii. $x_0 = 0$ のとき、 $f(x) \geq f(x_0) + z(x - x_0) \Leftrightarrow |x| \geq zx$ となるので、(b) より劣微分は、 $[-1, 1]$

iii. $x_0 > 0$ のとき、(a) より劣微分は、$\{1\}$

(d) $f(x) = x^2 - 3x + |x|$ のとき、

$$f(x) = \begin{cases} x^2 - 2x & x \geq 0 \\ x^2 - 4x & x < 0 \end{cases}$$

$$f'(x) = \begin{cases} 2x - 2, & x > 0 \\ 2x - 3 + [-1, 1] = -3 + [-1, 1] = [-4, -2] & x = 0 \\ 2x - 4 & x < 0 \end{cases}$$

したがって、 この $f(x)$ は、 $x = 1$ で極小となる。 次に、 $f(x) = x^2 + x + 2|x|$ のとき、

$$f(x) = \begin{cases} x^2 + 3x & x \geq 0, \\ x^2 - x & x < 0 \end{cases}$$

$$f'(x) = \begin{cases} 2x + 3 & x \geq 0, \\ 2x + 1 + 2[-1, 1] = 1 + 2[-1, 1] = [-1, 3], & x = 0 \\ 2x - 1 & x < 0 \end{cases}$$

したがって、この $f(x)$ は $x = 0$ で極小となる。

51. $\mathcal{S}_\lambda(x)$ は、 符号関数

$$\mathrm{sgn}(x) = \begin{cases} -1, & (x < 0) \\ 0, & x = 0 \\ 1 & x > 0 \end{cases}$$

を用いて、

$$\mathcal{S}_\lambda(x) = \mathrm{sgn}(x) \max\{|x| - \lambda, 0\}$$

と書くことができる。

52.

$$L = \frac{1}{2N} \sum_{i=1}^{N} (y_i - x_i\beta)^2 + \lambda|\beta|$$

の劣微分を考えると、

$$\frac{\partial L}{\partial \beta} = -\frac{1}{N} \sum_{i=1}^{N} x_i (y_i - x_i\beta) + \lambda \begin{cases} 1 & (\beta > 0) \\ -1 & (\beta < 0) \\ [-1, 1] & (\beta = 0) \end{cases} = -\frac{1}{N}(z - \beta) + \lambda \begin{cases} 1 & (\beta > 0) \\ -1 & (\beta < 0) \\ [-1, 1] & (\beta = 0) \end{cases}$$

となるので、

$$\frac{\partial L}{\partial \beta} = 0 \Leftrightarrow \begin{cases} 0 = -z + \beta + \lambda & (\beta > 0) \\ 0 = -z + \beta - \lambda & (\beta < 0) \\ 0 = -z + \beta + \lambda[-1, 1] & (\beta = 0) \end{cases} \Leftrightarrow \beta = \begin{cases} z - \lambda & (z > \lambda) \\ z + \lambda & (z < -\lambda) \\ 0 \end{cases}$$

このとき、 $\mathcal{S}_\lambda(x)$ を用いて、$\beta = S_\lambda(z)$ と書ける。

55. 関数 cv.glmnet は、10-fold クロスバリデーションを行って、Lasso の処理を行うための最適な λ の値を決定することができる。関数 glmnet は、目的変数、説明変数、λ を引数とし、Lasso 処理を行う。選択された変数は V3,V4,V5 である。

56. (a) S を β_1 で偏微分して、$(\beta_1, \beta_2) = \left(\hat{\beta}_1, \hat{\beta}_2\right)$ を代入すると、$0 = \sum_{i=1}^{N} -2x_{i,1}\left(y_i - \hat{\beta}_1 x_{i,1} - \hat{\beta}_2 x_{i,2}\right)$。したがって、

$$\sum_{i=1}^{N} x_{i,1}\left(y_i - \hat{\beta}_1 x_{i,1} - \hat{\beta}_2 x_{i,2}\right) = 0$$

同様に β_2 についても偏微分していくと、

$$\sum_{i=1}^{N} x_{i,2}\left(y_i - \hat{\beta}_1 x_{i,1} - \hat{\beta}_2 x_{i,2}\right) = 0$$

を得る。そして、

$$\begin{aligned} & y_i - \beta_1 x_{i,1} - \beta_2 x_{i,2} = y_i - \hat{y}_i + \hat{y}_i - \beta_1 x_{i,1} - \beta_2 x_{i,2} \\ =& y_i - \hat{y}_i + \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2} - \beta_1 x_{i,1} - \beta_2 x_{i,2} = y_i - \hat{y}_i - \left(\beta_1 - \hat{\beta}_1\right) x_{i,1} - \left(\beta_2 - \hat{\beta}_2\right) x_{i,2} \end{aligned}$$

さらに、

$$\sum_{i=1}^{N} x_{i,1}\left(y_i - \hat{y}_i\right) = \sum_{i=1}^{N} x_{i,2}\left(y_i - \hat{y}_i\right) = 0$$

に注意して、 題意の総和を展開していくと、

$$\begin{aligned} & \sum_{i=1}^{N} \left(y_i - \beta_1 x_{i,1} - \beta_2 x_{i,2}\right)^2 \\ &= \sum_{i=1}^{N} \left[\left(y_i - \hat{y}_i\right) - \left\{\left(\beta_1 - \hat{\beta}_1\right) x_{i,1} + \left(\beta_2 - \hat{\beta}_2\right) x_{i,2}\right\}\right]^2 \\ &= \sum_{i=1}^{N} \left(y_i - \hat{y}_i\right)^2 + \sum_{i=1}^{N} \left\{\left(\beta_1 - \hat{\beta}_1\right) x_{i,1} + \left(\beta_2 - \hat{\beta}_2\right) x_{i,2}\right\}^2 \\ &= \left(\beta_1 - \hat{\beta}_1\right)^2 \sum_{i=1}^{N} x_{i,1}^2 + 2\left(\beta_1 - \hat{\beta}_1\right)\left(\beta_2 - \hat{\beta}_2\right) \sum_{i=1}^{N} x_{i,1} x_{i,2} + \left(\beta_2 - \hat{\beta}_2\right)^2 \sum_{i=1}^{N} x_{i,2}^2 + \sum_{i=1}^{N} \left(y_i - \hat{y}_i\right)^2 \end{aligned}$$

(b) $\mathrm{A}(1,0), \mathrm{B}(0,1), \mathrm{C}(-1,0), \mathrm{D}(0,-1)$ とする。このとき,題意の $\left(\hat{\beta}_1, \hat{\beta}_2\right)$ の範囲は、$\left(\hat{\beta}_1, \hat{\beta}_2\right)$ を中心とするような円と正方形が接するときに 4 辺 (頂点を除く) と接するようなものを 除いた部分になる。

辺 AB について、 そのような範囲は、 直線 AD と直線 BC に 挟まれた部分 (境界除く) のうち、 辺 AB より上にある部分とな る。 他 3 辺についても同様に考えると、 題意の $\left(\hat{\beta}_1, \hat{\beta}_2\right)$ の範囲が得られる。

(c) 正方形ではなく、 原点を中心とすると単位円である場合は、

$$\left\{\left(\hat{\beta}_1, 0\right); \left|\hat{\beta}_1\right| > 1\right\} \cup \left\{\left(0, \hat{\beta}_2\right); \left|\hat{\beta}_2\right| > 1\right\}$$

と書ける。

第 6 章 非線形回帰

57. (a) $L = \sum_{i=1}^{N} \left(y_i - \sum_{j=0}^{p} \beta_j x_i^j \right)^2$ に対して、

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \in \mathbb{R}^N \ , \ X = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^p \\ 1 & x_2 & x_2^2 & \cdots & x_2^p \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_N & x_N^2 & \cdots & x_N^p \end{pmatrix} \in \mathbb{R}^{N\times(p+1)} \ , \ \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \in \mathbb{R}^{p+1}$$

とおくと、 $L = \|y - X\beta\|^2$ と書ける。したがって、 L を最小化するような $\hat{\beta}$は、6.より、$X^T X$ が正則であるという仮定のもとで、

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T y$$

となる。ここで、 $\operatorname{rank} X = \operatorname{rank} X^T X$ が成り立つことより、 $X^T X \in \mathbb{R}^{(p+1)\times(p+1)}$ が正則であるための条件は、 $\operatorname{rank} X = p+1$ となることである。このとき、 x_1、$\cdots, x_N$ のうち、 相異 なるものが $p+1$ 個存在するとして、それらを $x_{(1)}, \cdots, x_{(p+1)}$ として、行列

$$X' = \begin{pmatrix} 1 & x_{(1)} & \cdots & x_{(1)}^p \\ 1 & x_{(2)} & \cdots & x_{(2)}^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{(p+1)} & \cdots & x_{(p+1)}^p \end{pmatrix} \in \mathbb{R}^{(p+1)\times(p+1)}$$

を考えると、 Vandermonde の不等式より、

$$\det X' = (-1)^{\frac{p(p+1)}{2}} \prod_{1\le i<j\le p+1} \left(x_{(i)} - x_{(j)}\right) \neq 0$$

となるので、 $\operatorname{rank} X' = p+1$ となることから、 $\operatorname{rank} X = p+1$ が成り立つ、 逆に、 $x_1, \cdots, x_N$ のうち異なるものが p 個以下のときは、$\operatorname{rank} X < p+1$ となるから、$X^T X$ は正則とならな い。 以上より、 求める $\beta_0, \beta_1, \cdots, \beta_p$ が一意に定まる条件は、$x_1, \cdots, x_N$ のうち異なるものが $p+1$ 個以上存在することであ り、そのもとでの解は、

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix} = \left(X^T X\right)^{-1} X^T y$$

である。

(b) (a) と同様に、

$$f_j := \begin{cases} \mathbb{R} \to \{0\}, & j = 0 \\ \mathbb{R} \to \mathbb{R} & j = 1, \cdots, p \end{cases}$$

$$L := \sum_{i=1}^{N} \left(y_i - \sum_{j=0}^{p} \beta_j f_j\left(x_i\right) \right)^2$$

$$y=\begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \in \mathbb{R}^N \ , \ X=\begin{pmatrix} 1 & f_1(x_1) & f_2(x_1) & \cdots & f_p(x_1) \\ 1 & f_1(x_2) & f_2(x_2) & \cdots & f_p(x_2) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & f_1(x_N) & f_2(x_N) & \cdots & f_p(x_N) \end{pmatrix} \in \mathbb{R}^{N\times(p+1)} \ ,$$

$$\beta=\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \in \mathbb{R}^{p+1}$$

$(f_0(\cdot)=1)$ とおくと、$L=\|y-X\beta\|^2$ と書けるので、L を最小化するような $\beta=\hat{\beta}$は、X^TX が正則 であるという仮定のもとで、

$$\hat{\beta}=\left(X^TX\right)^{-1}X^Ty$$

となる。そして、$X^TX\in\mathbb{R}^{(p+1)\times(p+1)}$ が正則であるための条件は、$N\geq p+1$ であり、なおかつ X の $p+1$ 個の列ベクトル が線形独立であることである。

58. (a) 各 $i=1,\cdots,k+1$ と 3 次多項式 f_{i-1},f_i に対して、

$$\begin{cases} f_{i-1}(\alpha_i)=f_i(\alpha_i) \\ f_{i-1}^{(1)}(\alpha_i)=f_i^{(1)}(\alpha_i) \\ f_{i-1}^{(2)}(\alpha_i)=f_i^{(2)}(\alpha_i) \end{cases}$$

が成り立つことから、

$$\begin{cases} f_i(x)=\sum_{j=0}^3\varepsilon_j(x-\alpha_i)^j \\ f_{i-1}(x)=\sum_{j=0}^3\delta_j(x-\alpha_i)^j \end{cases}$$

と書けば、

$$\begin{cases} f_i^1(x) &= \sum_{j=1}^3 j\varepsilon_j(x-\alpha_i)^{j-1} \\ f_i^2(x) &= \sum_{j=2}^3 j(j-1)\varepsilon_j(x-\alpha_i)^{j-2} \end{cases}$$

のようになるので、

$$\begin{cases} \varepsilon_0 &= f_i(\alpha_i)=f_{i-1}(\alpha_i)=\delta_0, \\ \varepsilon_1 &= f_i^{(1)}(\alpha_i)=f_{i-1}^{(1)}(\alpha_i)=\delta_1 \\ 2\varepsilon_2 &= f_i^{(2)}(\alpha_i)=f_{i-1}^{(2)}(\alpha_i)=2\delta_2 \end{cases}$$

となり、$\varepsilon_j=\delta_j$ $(j=0,1,2)$ が成り立つ。 したがって、

$$f_i(x)-f_{i-1}(x)=(\varepsilon_3-\delta_3)(x-\alpha_i)^3$$

となるから、 題意の γ_i は、$\gamma_i=\varepsilon_3-\delta_3$ と定めればよい。 よっ て、題意は示された。

(b) ある $K+4$ 個の定数 $\beta_1,\beta_2,\cdots,\beta_{K+4}$ が存在して、

$$f(x)=\beta_1+\beta_2x+\beta_3x^2+\beta_4x^3+\sum_{j=1}^K\beta_{j+4}(x-\alpha_j)_+^3$$

が成り立つこと、すなわち、任意の $i = 0, 1, \cdots, K$ および、任意の $x \in [\alpha_i, \alpha_{i+1}]$ に対して、

$$f(x) = \beta_1 + \beta_2 x + \beta_3 x^2 + \beta_4 x^3 + \sum_{j=1}^{i} \beta_{i+4} (x - \alpha_i)^3$$

が成り立つことを示す。 $i = 0$ のとき、$x \in [\alpha_0, \alpha_1]$ で $f(x) = f_0(x)$ となるので、ある $\beta_1, \beta_2, \beta_3, \beta_4$ がただ一つ存在して、

$$f(x) = \beta_1 + \beta_2 x + \beta_3 x^2 + \beta_4 x^3$$

となるので、$(*)$ が成り立つ。 $i = 1$ のとき、 $x \in [\alpha_1, \alpha_2]$ に対して、$f_1(x) = f_0(x) + \gamma_1 (x - \alpha_1)^3$ となるから、$\beta_5 = \gamma_1$ と定めれ ば、$(*)$ が成り立つ。そして、$\beta_1, \cdots, \beta_{i+4}$ まで定まったとき, すなわち $x \in [\alpha_0, \alpha_{i+1}]$ に対する $f(x)$ の係数 $\beta_1, \beta_2, \cdots, \beta_{i+4}$ が既に定められているとき、

$$f_{i+1}(x) = f_0(x) + \sum_{j=1}^{i+1} \gamma_j (x - \alpha_j)^3$$

から、 $\beta_{i+5} = \gamma_{i+1}$ とすれば、 $x \in [\alpha_{i+1}, \alpha_{i+2}]$ に対して、

$$f(x) = \beta_1 + \beta_2 x + \beta_3 x^2 + \beta_4 x^3 + \sum_{j=1}^{i+1} \beta_{j+4} (x - \alpha_j)^3$$

が成り立つ。よって、$i = K - 1$ の場合まで考えると、題意の等式を満たすような $\beta_1, \cdots, \beta_{K+4}$ を定めることができる。よって、題意は示された。

60. (a)

$$g(x) = \gamma_1 + \gamma_2 x + \gamma_3 \frac{(x - \alpha_1)^3}{\alpha_K - \alpha_1} + \cdots + \gamma_K \frac{(x - \alpha_{K-2})^3}{\alpha_K - \alpha_{K-2}} + \gamma_{K+1} \frac{(x - \alpha_{K-1})^3}{\alpha_K - \alpha_{K-1}}$$

と書く時、自然な 3 次のスプライン曲線 $g(x)$ が $x \geq \alpha_K$ で直線となることと、 境界 $x = \alpha_K$ で 1 次、2 次微分係数がともに 一致することから、

$$g''(\alpha_K) = 0$$

であることが必要になる。このとき、

$$g''(\alpha_K) = \sum_{i=3}^{K} 6\gamma_i \cdot \frac{\alpha_K - \alpha_1}{\alpha_K - \alpha_1} = 6 \sum_{i=3}^{K+1} \gamma_i$$

となることから、$6 \sum_{i=3}^{K+1} \gamma_i = 0$、したがって

$$\gamma_{K+1} = -\sum_{j=3}^{K} \gamma_j$$

が成り立つ。

(b) $j = 1, \cdots, K - 1$, および $x \geq \alpha_K$ について、

$$d_j(x) = \frac{(x - \alpha_j)^3 - (x - \alpha_K)^3}{\alpha_K - \alpha_j}$$

となるので、$j=1,\cdots,K-2$ に対して、

$$\begin{aligned}h_{j+2}(x) =& d_j(x)-d_{K-1}(x)\\ =& \frac{(x-\alpha_j)^3-(x-\alpha_K)^3}{\alpha_K-\alpha_j}-\frac{(x-\alpha_{K-1})^3-(x-\alpha_K)^3}{\alpha_K-\alpha_{K-1}}\\ =& \frac{\left\{(x-\alpha_j)^3-(x-\alpha_K)^3\right\}(\alpha_K-\alpha_{K-1})}{(\alpha_K-\alpha_j)(\alpha_K-\alpha_{K-1})}-\frac{\left\{(x-\alpha_{K-1})^3-(x-\alpha_K)^3\right\}(\alpha_K-\alpha_j)}{(\alpha_K-\alpha_j)(\alpha_K-\alpha_{K-1})}\end{aligned}$$

が成り立つ。このとき、(60.1) 式の分子を整理していくと、

$$\begin{aligned}&\left\{(x-\alpha_j)^3-(x-\alpha_K)^3\right\}(\alpha_K-\alpha_{K-1})-\left\{(x-\alpha_{K-1})^3-(x-\alpha_K)^3\right\}(\alpha_K-\alpha_j)\\ =&\left(-\alpha_j^3+\alpha_K^3\right)(\alpha_K-\alpha_{K-1})-\left(-\alpha_{K-1}^3+\alpha_K^3\right)(\alpha_K-\alpha_j)\\ &+3x\left\{\left(\alpha_j^2-\alpha_K^2\right)(\alpha_K-\alpha_{K-1})-\left(\alpha_{K-1}^2-\alpha_K^2\right)(\alpha_K-\alpha_j)\right\}\\ &+3x^2\left\{(-\alpha_j+\alpha_K)(\alpha_K-\alpha_{K-1})-(\alpha_K-\alpha_{K-1})(\alpha_K-\alpha_j)\right\}\\ =&-(\alpha_K-\alpha_{K-1})(\alpha_K-\alpha_j)(\alpha_{K-1}-\alpha_j)(\alpha_j+\alpha_{K-1}+\alpha_K)\\ &+3x(\alpha_K-\alpha_{K-1})(\alpha_K-\alpha_j)(\alpha_{K-1}-\alpha_j)\\ =&(\alpha_K-\alpha_{K-1})(\alpha_K-\alpha_j)(\alpha_{K-1}-\alpha_j)(3x-\alpha_j-\alpha_{K-1}-\alpha_K)\end{aligned}$$

となるので、

$$h_{j+2}(x)=(\alpha_{K-1}-\alpha_j)(3x-\alpha_j-\alpha_{K-1}-\alpha_K)$$

が成り立つ.

(c) まず、任意の $x\le\alpha_1$ に対して、

$$d_j(x)=0$$

となることに注意すると、 任意の $x\le\alpha_1$ に対して、

$$g(x)=\gamma_1+\gamma_2x+\sum_{j=3}^{K}\gamma_j\left\{d_{j-2}(x)-d_{K-1}(x)\right\}=\gamma_1+\gamma_2x$$

であり、これは x の線形の関数である。次に、任意の $x\ge\alpha_K$ に対して、

$$g(x)=\gamma_1+\gamma_2x+\sum_{j=3}^{K}\gamma_jh_j(x)$$

となるが、(b) より $j=3,\cdots,K$ に対して $h_j(x)$ は高々1 次の多項式となるから、$g(x)$ も線形となる。よって、題意は示さ れた。

62. (a) 自然な 3 次のスプライン関数 g について、 $g''(x_1)=g''(x_N)=0$ となることと、 各区間 $[x_i,x_{i+1}]$ において 3 次微分係数が一定値 (これを γ_i とする) となる

$$\begin{aligned}&\int_{x_1}^{x_N}g''(x)h''(x)dx=\left[g''(x)h'(x)\right]_{x_1}^{x_N}-\int_{x_1}^{x_N}g^{(3)}(x)h'(x)dx\\ &=0-\sum_{i=1}^{N-1}\int_{x_i}^{x_{i+1}}\gamma_ih'(x)dx=-\sum_{i=1}^{N-1}\gamma_i\left\{h\left(x_{i+1}\right)-h\left(x_i\right)\right\}\end{aligned}$$

(b) $\int_{x_1}^{x_N} g''(x)h''(x)dx$ の仮定の元で、

$$\int_{-\infty}^{\infty} \{f''(x)\}^2 dx \geq \int_{x_1}^{x_N} \{f''(x)\}^2 dx = \int_{x_1}^{x_N} \{g''(x)+h''(x)\}^2 dx$$
$$= \int_{x_1}^{x_N} \left[\{g''(x)\}^2 + \{h''(x)\}^2\right] dx + 2\int_{x_1}^{x_N} g''(x)h''(x)dx = \int_{x_1}^{x_N} \left[\{g''(x)\}^2 + \{h''(x)\}^2\right] dx$$
$$\geq \int_{x_1}^{x_N} \{g''(x)\}^2 dx = \int_{-\infty}^{\infty} \{g''(x)\}^2 dx$$

ただし、最後で 60.(c) から、$x \notin [x_1, x_N]$ で $g''(x) = 0$ となることを用いた。よって、題意は示された。

(c) 自然な 3 次スプライン関数が g が、 $i = 1, \cdots, N$ に対 して $f(x_i) = g(x_i)$ を満たすとき、$h(x) = f(x) - g(x)$ とすると、$h(x_i) = 0, i = 1, \cdots, N$ となる。このとき、(b) の不等式も合わせて考えると、$RSS(f, \lambda)$ を最小にするような任意の $f : \mathbb{R} \to \mathbb{R}$ に対して、

$$RSS(f,\lambda) = \sum_{i=1}^{N} \{y_i - f(x_i)\}^2 + \lambda \int_{-\infty}^{\infty} \{f''(x)\}^2 dx \geq \sum_{i=1}^{N} \{y_i - g(x_i)\}^2 + \lambda \int_{-\infty}^{\infty} \{g''(x)\}^2 dx$$
$$= \mathrm{RSS}(g,\lambda)$$

64. 題意の平滑化スプライン関数 g は、

$$RSS(f,\lambda) = \sum_{i=1}^{N} \{y_i - f(x_i)\}^2 + \lambda \int_{-\infty}^{\infty} \{f''(x)\}^2 dx$$

を最小にするような $f : \mathbb{R} \to \mathbb{R}$ 全体の元であるとする。このとき、$RSS(g, \lambda)$ の第 1 項は、

$$\sum_{i=1}^{N} \{y_i - g(x_i)\}^2 = \left\| \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} - \begin{pmatrix} g(x_1) \\ \vdots \\ g(x_N) \end{pmatrix} \right\|^2$$
$$= \left\| \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} - \begin{pmatrix} g_1(x_1) & \cdots & g_N(x_1) \\ \vdots & \ddots & \vdots \\ g_1(x_N) & \cdots & g_N(x_N) \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_N \end{pmatrix} \right\|^2 = \|y - G\gamma\|^2$$

であり、第 2 項は、

$$\lambda \int_{-\infty}^{\infty} \{g''(x)\}^2 dx = \lambda \int_{-\infty}^{\infty} \sum_{i=1}^{N} \gamma_i g_i''(x) \sum_{j=1}^{N} \gamma_j g_j''(x) dx$$
$$= \lambda \sum_{i=1}^{N}\sum_{j=1}^{N} \gamma_i\gamma_j \int_{-\infty}^{\infty} g_i''(x)g_j''(x)dx = \lambda \sum_{i=1}^{N} \gamma_i \sum_{j=1}^{N} \gamma_j g_{i,j}'' = \lambda\gamma^T G''\gamma$$

となるので、

$$RSS(g,\lambda) = \|y - G\gamma\|^2 + \lambda\gamma^T G''\gamma$$

の両辺を γ で微分して整理すると、

$$0 = -2G^T(y - G\gamma) + 2\lambda G''\gamma \Longrightarrow \left(G^T G + \lambda G''\right)\gamma = G^T y$$
$$\Longrightarrow \gamma = \hat{\gamma} = \left(G^T G + \lambda G''\right)^{-1} G^T y$$

よって、前半部分の題意は示された。

67. (a) 行列を用いると、

$$\sum_{i=1}^{N} K(x, x_i)\left(y_i - [1, x_i]\,\beta(x)\right)^2$$
$$= (y - X\beta(x))^T \begin{pmatrix} K(x, x_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & K(x, x_N) \end{pmatrix} (y - X\beta(x))$$

と書き換えることができる。このとき、

$$W' = \begin{pmatrix} K(x, x_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & K(x, x_N) \end{pmatrix}$$

とすると、$\sum_{i=1}^{N} K(x, x_i)\left(y_i - [1, x_i]\,\beta(x)\right)^2 = (y - X\beta(x))^T W'(y - X\beta(x))$ となる。これを β で微分すると、$-2X^T W'(y - X\beta(x))$ となるが、これが 0 に等しいとき、$X^T W' y = X^T W' X\beta(x)$ より、

$$\hat{\beta}(x) = \beta(x) = \left(X^T W' X\right)^{-1} X^T W' y$$

を得る。したがって、

$$W = W' = \begin{pmatrix} K(x, x_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & K(x, x_N) \end{pmatrix}$$

であり、すなわち W は $K(x, x_1), \cdots, K(x, x_N)$ を対角成分と する対角行列である。

第 8 章 サポートベクトルマシン

75. (a)

$$\begin{cases} \epsilon_i = 0 \text{ のときは、} & \text{手前のマージン上にあるか、もしくはそうでない場合である。} \\ 0 < \epsilon_i < 1 \text{ のときは、} & \text{境界と手前のマージンの間にある。} \\ \epsilon_i = 1 \text{ のときは、} & \text{境界上にある。} \\ \epsilon_i > 1 \text{ のときは、} & \text{境界の反対側にある。} \end{cases}$$

(b) 少なくとも r 個の異なる $i \in \{1, \cdots, n\}$ に対して、 $y_i(\beta_0 + x_i\beta) < 0$ となるときに、解 $M > 0$ が存在すると仮定すると、 その各 i に対して、

$$y_i(\beta_0 + x_i\beta) \geq M(1 - \epsilon_i)$$

を満たすには、 $\epsilon_i > 1$ が必要であり、すなわち、$\sum_{i=1}^{N} \epsilon_i > r$ が必要となる。このとき、$\gamma \leq r$ となっているならば、$\sum_{i=1}^{N} \epsilon_i \leq \gamma \leq r$ となるので、矛盾する。

(c) ある $\gamma=\gamma_0$ において M を最適値 M_0 にするような $(\beta,\beta_0,\epsilon_i)$ は、 $\gamma=\gamma_1>\gamma_0$ の場合においても条件を満たすの で、 その場合においての最適値 M は、 少なくとも M_0 以上と なる。

76. まず、(8.24) について、$f_j(\beta)>0$ となるような添字 $j\in\{1,\cdots,m\}$ が存在するとき、α_j の値を大きくすることにより、$L(\alpha,\beta)$ の値をいくらでも大きくすることができる。 一方で、任意の $j\in\{1,\cdots,m\}$ に対して $f_j(\beta)\leq 0$ となるとき、$\alpha_1=\cdots=\alpha_m=0$ とすれば、$L(\alpha,\beta)$ は最大値 $f_0(\beta)$ を取る。 よって、 (8.24) は示された。次に、任意の $\alpha\in[0,\infty)^m,\beta\in\mathbb{R}^p$ について、

$$\sup_{\alpha'\geq 0} L\left(\alpha',\beta\right)\geq L(\alpha,\beta)\geq \inf_{\beta'} L\left(\alpha,\beta'\right)$$

から、任意の $\alpha\in[0,\infty)^m,\beta\in\mathbb{R}^p$ について、

$$\sup_{\alpha'\geq 0} L\left(\alpha',\beta\right)\geq \inf_{\beta'} L\left(\alpha,\beta'\right)$$

が成り立つ。この不等式は、左辺で β について inf、右辺で α について sup を取っても成り立つので、文字を適宜置き換えることにより (8.25) が成り立つことがわかる。 そして、

$$\begin{aligned}(p,m)&=(2,1)\\ L(\alpha,\beta)&=\beta_1+\beta_2+\alpha\left(\beta_1^2+\beta_2^2-1\right)\end{aligned}$$

だから、

$$\begin{aligned}f_0(\beta)&=\beta_1+\beta_2\\ f_1(\beta)&=\beta_1^2+\beta_2^2-1\\ \alpha_1&=\alpha\end{aligned}$$

とすればよく。 (8.24) より、

$$\sup_{\alpha\geq 0} L(\alpha,\beta)=\begin{cases}\beta_1+\beta_2 & \left(\beta_1^2+\beta_2^2-1\leq 0\right)\\ \infty & \left(\beta_1^2+\beta_2^2-1>0\right)\end{cases}$$

と書けて、これは $\beta_1=\beta_2=-1/\sqrt{2}$ で最小値 $-\sqrt{2}$ を取る。 したがって、 (8.25) の左辺は $-\sqrt{2}$ となる。

次に、$\dfrac{\partial L}{\partial\beta_1}=\dfrac{\partial L}{\partial\beta_2}=0$ のとき、$\begin{cases}1+2\alpha\beta_1=0\\ 1+2\alpha\beta_2=0\end{cases}$ から、$\beta_1=\beta_2=-1/(2\alpha)$ となるので、

$$\inf_{\beta} L(\alpha,\beta)=-\frac{1}{2\alpha}-\frac{1}{2\alpha}+\alpha\left\{\left(-\frac{1}{2\alpha}\right)^2+\left(-\frac{1}{2\alpha}\right)^2-1\right\}=-\frac{1}{2\alpha}-\alpha=-\left(\alpha+\frac{1}{2\alpha}\right)$$

この値の最大値は、 相加平均と相乗平均の大小関係から、 $\alpha=1/(2\alpha)$、$\alpha<0$. すなわち $\alpha=-1/\sqrt{2}$ のときにとり、 その最大値は $-\sqrt{2}$ である。よって、 (8.25) で等号が成り立つ。

77. (a) (8.30) で、 $(f,x_0,x)\mapsto(f_0,\beta^*,\beta)$ として整理すると、

$$\begin{aligned}f_0\left(\beta^*\right)&\leq f_0(\beta)-\nabla f_0\left(\beta^*\right)^T\left(\beta-\beta^*\right)=f_0(\beta)+\sum_{i=1}^m\alpha_i\nabla f_i\left(\beta^*\right)^T\left(\beta-\beta^*\right)\\ &\leq f_0(\beta)+\sum_{i=1}^m\alpha_i\left\{f_i(\beta)-f_i\left(\beta^*\right)\right\}=f_0(\beta)+\sum_{i=1}^m\alpha_i f_i(\beta)\leq f_0(\beta)\end{aligned}$$

ただし、 最後の行の変形で $,i=1,\cdots,m$ に対して $\alpha_i \geq 0, f_i(\beta) \leq 0$ となることを用いた。

(b) (8.26) については、

$$f_0(\beta) = \beta_1 + \beta_2,$$
$$f_1(\beta) = \beta_1^2 + \beta_2^2 - 1$$

となるから、 (8.27),(8.28),(8.29) はそれぞれ、

$$\begin{cases} \beta_1^2 + \beta_2^2 - 1 \leq 0, \\ \alpha\left(\beta_1^2 + \beta_2^2 - 1\right) = 0, \\ \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \alpha \begin{pmatrix} 2\beta_1 \\ 2\beta_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \end{cases}$$

となる。

78. まず、 (8.27) より、 $i=1,\cdots,N$ について、

$$y_i\left(\beta_0 + x_i\beta\right) - (1-\epsilon_i) \geq 0$$
$$\epsilon_i \geq 0$$

が得られ、次に、(8.28) より、$i=1,\cdots,N$ について、

$$\alpha_i\left\{y_i\left(\beta_0 + x_i\beta\right) - (1-\epsilon_i)\right\} = 0$$
$$\mu_i\epsilon_i = 0$$

が得られる、そして、(8.29) は、$\begin{cases} \frac{\partial L_P}{\partial \beta_0} = 0, \\ \frac{\partial L_P}{\partial \beta} = \mathbf{0}, \\ \frac{\partial L_P}{\partial \epsilon_i} = 0, \end{cases}$ とできるので、$\begin{cases} \sum_{i=1}^N \alpha_i y_i = 0 \\ \beta - \sum_{i=1}^N \alpha_i y_i x_i^T = 0, \\ C - \alpha_i - \mu_i = 0 \end{cases}$

が得られる。

79. L_P の、 β_0, β における最適化を図るとき。 (8.32), (8.34) から、 L_P は以下のように書ける。

$$\frac{1}{2}\|\beta\|_2^2 + \sum_{i=1}^N (C - \mu_i - \alpha_i)\,\epsilon_i + \sum_{i=1}^N \alpha_i - \sum_{i=1}^N \alpha_i y_i\left(\beta_0 + x_i\beta\right) = \frac{1}{2}\|\beta\|_2^2 + \sum_{i=1}^N \alpha_i - \sum_{i=1}^N \alpha_i y_i x_i \beta$$

さらに、(8.33) を用いて、

$$\frac{1}{2}\|\beta\|_2^2 = \frac{1}{2}\left(\sum_{i=1}^N \alpha_i y_i x_i^T\right)^T \left(\sum_{i=1}^N \alpha_i y_i x_i^T\right), = \frac{1}{2}\sum_{i=1}^N\sum_{j=1}^N \alpha_i\alpha_j y_i y_j x_i x_j^T$$
$$-\sum_{i=1}^N \alpha_i y_i x_i \beta = -\sum_{i=1}^N \alpha_i y_i x_i \sum_{j=1}^N \alpha_j y_j x_j^T = -\sum_{i=1}^N\sum_{j=1}^N \alpha_i\alpha_j y_i y_j x_i x_j^T$$

と書けるので、Lagrange 係数であった $\alpha_i, \mu_i \geq 0, i=1,\cdots,N$ を引数とする関数、

$$L_D = \sum_{i=1}^N \alpha_i - \frac{1}{2}\sum_{i=1}^N\sum_{j=1}^N \alpha_i\alpha_j y_i y_j x_i x_j^T$$

を構成することができる。ただし、 (8.32), (8.34) に注意すると、$\alpha_i, i=1,\cdots,N$ の動く範囲は、

$$\begin{cases} 0 \leq \alpha_i \leq C \quad (i=1,\cdots,N) \\ \sum_{i=1}^N \alpha_i y_i = 0 \end{cases}$$

となる。この双対問題を解くことによって得られた $\alpha_i, i = 1, \cdots, N$ を (8.33) に代入することにより、 β が得られる。

80.

$$y_i(\beta_0 + x_i\beta) > 1 \Rightarrow y_i(\beta_0 + x_i\beta) - (1 - \epsilon_i) > 0 (\because (8.37), (8.38)) \Rightarrow \alpha_i = 0, (\because (8.35))$$

$$0 < \alpha_i < C \Rightarrow \begin{cases} \mu_i > 0, (\because (8.34)) \\ y_i(\beta_0 + x_i\beta) - (1 - \epsilon_i) = 0, (\because (8.35)) \end{cases} \Rightarrow \epsilon_i = 0 (\because (8.36))$$

$$y_i(\beta_0 + x_i\beta) < 1 \Rightarrow \epsilon_i > 0 (\because (8.37)) \Rightarrow \mu_i = 0 (\because (8.36)) \Rightarrow \alpha_i = C (\because (8.34))$$

81. (a) $\alpha_1 = \cdots = \alpha_N = 0$ が成り立つとき、(8.33) より $\beta = \mathbf{0}$、 (8.34), (8.36) より、$\epsilon_1 = \cdots = \epsilon_N = 0$ が成り立つことがわかる。このとき、 任意の $i = 1, \cdots, N$ に対して、 $y_i(\beta_0 + x_i\beta) = 1$ が成り立つと仮定すると、

$$y_i\beta_0 = 1 > 0$$

となる。 さらに、 $y_i \in \{-1, 1\}$ となることから、$y_1 = \cdots = y_N$ が必要となる。このとき、 任意の $i = 1, \cdots, N$ に対して、 $(y_i, \beta_0) = (\pm 1, \pm 1)$ (複号同順) となる。よって、題意は示さ れた。

(b) $y_i(\beta_0 + x_i\beta) \neq 1$ のもとで $\alpha_i = C$ ならば $\epsilon_i > 0, \alpha_i = 0$ ならば $\epsilon_i = 0$ が成り立つ。このとき、

$$\epsilon_* = \min_{i=1,\cdots,N} \epsilon_i \geq 0$$

として、各 ϵ_i を、 $\epsilon_i - \epsilon_*$ に、 β_0 を $\beta_0 + y_i\epsilon_*$ とそれぞれ置き換えると、 $y_i^2 = 1$ となることに注意すれば、

$$y_i(\beta_0 + y_i\epsilon_* + x_i\beta) = y_i(\beta_0 + x_i\beta) + \epsilon_*$$

が成り立つので、 $y_i(\beta_0 + x_i\beta) - (1 - \epsilon_i)$ は、

$$y_i(\beta_0 + x_i\beta) + \epsilon_* - (1 - \epsilon_i + \epsilon_*) = y_i(\beta_0 + x_i\beta) - (1 - \epsilon_i)$$

となり、 置き換えても値は変化しない。このとき、(8.35) および、80. の 3 つ目の命題の対偶も合わせて考えると、$\alpha_i = 0, C$ いずれの場合でも、(8.37) では等号が成立する。したがって、 7 つの KKT 条件が成立する。 このとき、 置き換える前後で、

$$\epsilon_* \sum_{i=1}^{N} (C - \mu_i) = \epsilon_* \sum_{i=1}^{N} \alpha_i > 0$$

だけ減少するので、 これは L_P の最適解にはならない。

(c) (a)(b) および命題 80 より 、 $0 < \alpha_i < C$ もしくは $y_i(\beta_0 + x_i\beta) = 1$ となる添え字 i が少なくとも 1 つ存在する。前者の場合においても、 80. の 2 つの命題より、$y_i(\beta_0 + x_i\beta) = 1$ が成り立つ。 よって、 題意は示された。

82.

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i\alpha_j y_i y_j x_i x_i^T \ , \ \sum_{i=1}^{N} \alpha_i y_i = 0 \ , \ 0 \leq \alpha_i \leq C$$

を、

$$L_D = -\frac{1}{2}\alpha^T D_{mat}\alpha + d_{vec}^T\alpha \ ,\ A_{mat}\alpha \geq b_{vec} \ ,\ b_{vec} = [0, -C, \cdots, -C, 0, \cdots, 0]^T \in \mathbb{R}^m$$

(2 つ目の不等式の、 最初の meq 個については等号成立) と、

$$A_{mat} \in \mathbb{R}^{m\times N} \ ,\ meq \in \mathbb{N} \ ,\ D_{mat} \in \mathbb{R}^{N\times N} \ ,\ d_{vec} \in \mathbb{R}^N$$

に書き直すとき、

$$z = \begin{bmatrix} x_{1,1}y_1 & \cdots & x_{1,p}y_1 \\ \vdots & \ddots & \vdots \\ x_{N,1}y_N & \cdots & x_{N,p}y_N \end{bmatrix} \in \mathbb{R}^{N\times p}$$

とし、

$$m = 2N+1 \ ,\ A_{mat} = \begin{bmatrix} y_1 & \cdots & y_N \\ -1 & & \\ & \ddots & \\ 1 & & -1 \\ & \ddots & \\ & & 1 \end{bmatrix} \in \mathbb{R}^{(2N+1)\times N} \ ,\ meq = 1 \ ,\ D_{mat} = zz^T \ ,$$

$d_{vec} = [1, \cdots, 1]^T$

とおけば、題意の双対問題と同値となる。

83.

$$\begin{aligned} K(x,y) &= \left(1 + x^T y\right)^2 = (1 + x_1y_1 + x_2y_2)^2 \\ &= 1 + 2x_1y_1 + 2x_2y_2 + x_1^2y_1^2 + 2x_1x_2y_1y_2 + x_2^2y_2^2 \\ &= \left[1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, \sqrt{2}x_1x_2, x_2^2\right]\left[1, \sqrt{2}y_1, \sqrt{2}y_2, y_1^2, \sqrt{2}y_1y_2, y_2^2\right]^T \end{aligned}$$

したがって、写像 ϕ は

$$(x_1, x_2) \mapsto \left(1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, \sqrt{2}x_1x_2, x_2^2\right)$$

となる。

84. (a) 任意の $f, g, h \in V, \alpha, \beta \in \mathbb{R}$ に対して、$\langle f, g\rangle = \displaystyle\int_0^1 f(x)g(x)dx$ とおくと、

$$\begin{aligned} \langle f, g\rangle &= \int_0^1 f(x)g(x)dx = \int_0^1 g(x)f(x)dx = \langle g, f\rangle, \\ \langle \alpha f + \beta g, h\rangle &= \int_0^1 \{\alpha f(x) + \beta g(x)\}h(x)dx = \alpha\int_0^1 f(x)h(x)dx + \beta\int_0^1 g(x)h(x)dx \\ &= \alpha\langle f, h\rangle + \beta\langle g, h\rangle, \\ \langle f, f\rangle &= \int_0^1 \{f(x)\}^2 dx \geq 0 \text{ (等号成立は, } f \equiv 0 \text{ の場合のみ)} \end{aligned}$$

が成り立つ。したがって、 $\langle \cdot, \cdot \rangle$ は V の内積となる。

(b) 任意の $x, y \in \mathbb{R}^p$ に対して、 $\langle x, y \rangle = \left(1 + x^T y\right)^2$ とする とき、

$$\langle 0 \cdot x, y \rangle = 1^2 = 1 \neq 0 = 0 \cdot \langle x, y \rangle$$

となる。これは、$\langle \cdot, \cdot \rangle$ が線形性を持たないことを意味するので、V 上の内積とはならない。

教師なし学習

90. (a) 各 $j = 1, \cdots, p$ について,

$$\frac{1}{|C_k|} \sum_{i \in C_k} \sum_{i' \in C_k} (x_{i,j} - x_{i',j})^2 = 2 \sum_{i \in C_k} (x_{i,j} - \bar{x}_{k,j})^2$$

が成り立つことを示せば十分。 (90.1) の左辺について、

$$\begin{aligned}
&\frac{1}{|C_k|} \sum_{i \in C_k} \sum_{i' \in C_k} (x_{i,j} - x_{i',j})^2 \\
=&\frac{1}{|C_k|} \sum_{i \in C_k} \sum_{i' \in C_k} (x_{i,j} - \bar{x}_{k,j} + \bar{x}_{k,j} - x_{i',j})^2 \\
=&\frac{1}{|C_k|} \sum_{i \in C_k} \sum_{i' \in C_k} (x_{i,j} - \bar{x}_{k,j})^2 + \frac{1}{|C_k|} \sum_{i \in C_k} \sum_{i' \in C_k} (\bar{x}_{k,j} - x_{i',j})^2 \\
&+ \frac{2}{|C_k|} \sum_{i \in C_k} \sum_{i' \in C_k} (x_{i,j} - \bar{x}_{k,j}) (\bar{x}_{k,j} - x_{i',j})
\end{aligned}$$

第 1 項と第 2 項は等しく、 第 3 項は 0 となる。したがって、

$$\frac{1}{|C_k|} \sum_{i \in C_k} \sum_{i' \in C_k} (x_{i,j} - x_{i',j})^2 = \frac{2}{|C_k|} \sum_{i \in C_k} \sum_{i' \in C_k} (x_{i,j} - \bar{x}_{k,j})^2 = 2 \sum_{i \in C_k} (x_{i,j} - \bar{x}_{k,j})^2$$

右辺は、 (90.1) の右辺に一致している。よって、題意の等式は示された。

(b) (a) より、 スコア S は、

$$S = 2 \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{p} (x_{i,j} - \bar{x}_{k,j})^2 = 2 \sum_{k=1}^{K} \sum_{i \in C_k} \|x_i - \bar{x}_k\|^2$$

と書ける。このとき、各 $k = 1, \cdots, K$ および任意のベクトル x に対して、

$$\begin{aligned}
&\sum_{i \in C_k} \|x_i - x\|^2 = \sum_{i \in C_k} \|(x_i - \bar{x}_k) - (x - \bar{x}_k)\|^2 \\
&= \sum_{i \in C_k} \|x_i - \bar{x}_k\|^2 + \sum_{i \in C_k} \|x - \bar{x}_k\|^2 - 2 (x - \bar{x}_k)^T \sum_{i \in C_k} (x_i - \bar{x}_k) \\
&= \sum_{i \in C_k} \|x_i - \bar{x}_k\|^2 + \sum_{i \in C_k} \|x - \bar{x}_k\|^2 \leq \sum_{i \in C_k} \|x_i - \bar{x}_k\|^2
\end{aligned}$$

これは、2 つのステップを踏んでも、スコア S が増加することはないことを示している。

(c) まず、1 つ目のケースでは、 3,10 がそれぞれクラスタ 1,2 の中心になり。 $0, 6, 10$ から最も近いクラスタの中心がそれぞ れ $3, 3, 10$ となるので、 処理を継続してもその状態は変わらな い。このときのスコア S は、

$$S = 3^2 + 3^2 + 0^2 = 18$$

となる。 次に 2 つ目のケースでは、 0,8 がそれぞれクラスタ 1,2 の中心になり。0,6,10 から最も近いクラスタの中心がそれ ぞれ $0, 8, 8$ となるので、 処理を継続してもその状態は変わらな い。このときのスコア S は、

$$S = 0^2 + 2^2 + 2^2 = 8$$

となる。

93. Centroid リンケージを適用するとき、はじめに $(5, 8), (9, 0)$ が結合されて、クラスタ距離は $\sqrt{4^2 + 8^2} = 80$ となる。このとき、中心は $(7, 4)$ となり、もう一方の中心 $(0, 0)$ と結合したときのクラスタ間の距離は $\sqrt{7^2 + 4^2} = \sqrt{65}$ とな り、最初の結合よりクラスタ間の距離が小さくなるため、 樹形図の木が交差してしまう。

94. (a) $\|\phi\|^2 = 1$ の元で、$\|X\phi\|^2$ を最大化するような ϕ を考える。このとき、KKT 条件を考えると、

$$L = \|X\phi\|^2 - \gamma\left(\|\phi\|^2 - 1\right)$$

として、$\partial L/\partial\gamma = 0, \partial L/\partial\phi = 0$ となることであり、これを解 くと、

$$X^T X\phi = \gamma\phi$$

を得る。このとき、ϕ が $X^T X$、そして Σ の固有ベクトルであることが必要。

$$\|X\phi\|^2 = \phi^T X^T X\phi = \phi^T X^T X\phi = \gamma\phi^T\phi = \gamma\|\phi\|^2 = \gamma$$

これを最大化するような γ は、 $X^T X$ の固有値のうち最大もの。 このとき、$\phi = \phi_1$ は λ_1 の固有ベクトルとなるので、$\Sigma\phi_1 = \lambda_1\phi_1$ が成り立つ。

(b) 対称行列の異なる固有空間に属するベクトルは、直交する。固有値がすべて異なるので、各固有空間の次元が 1 であって、固有ベクトルどうし直交する。

97.

$$\sum_{i=1}^{N}\left\|x_i - x_i\Phi\Phi^T\right\|^2 = \sum_{i=1}^{N}\|x_i\|^2 + \sum_{i=1}^{N} x_i\Phi\Phi^T\left(x_i\Phi\Phi^T\right)^T - 2\sum_{i=1}^{N} x_i\left(x_i\Phi\Phi^T\right)^T$$

$$= \sum_{i=1}^{N}\|x_i\|^2 - \sum_{i=1}^{N} x_i\Phi\Phi^T\left(x_i\Phi\Phi^T\right)^T = \sum_{i=1}^{N}\|x_i\|^2 - \sum_{i=1}^{N} x_i\Phi\Phi^T x_i^T = \sum_{i=1}^{N}\|x_i\|^2 - \sum_{i=1}^{N}\|x_i\Phi\|^2,$$

$$\sum_{i=1}^{N}\|x_i\Phi\|^2 = \sum_{i=1}^{N}\sum_{j=1}^{m}(x_i\phi_j)^2 = \sum_{j=1}^{m}\sum_{i=1}^{N}(x_i\phi_j)^2 = \sum_{j=1}^{m}\left\|\begin{array}{c} x_1\phi_j \\ \vdots \\ x_N\phi_j \end{array}\right\|^2 = \sum_{j=1}^{m}\|X\phi_j\|^2$$