Solutions for "Statistical Machine Learning with R: 100 Problems" (Mathematical Part)

Joe Suzuki

For the program, please refer to the solution outline (R program).html.

Chapter 2: Linear Regression

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

   (a) Let $\bar{x} := \frac{1}{N}\sum_{i=1}^N x_i$, $\bar{y} := \frac{1}{N}\sum_{i=1}^N y_i$, then,

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

$$\Longleftrightarrow N\beta_0 + \beta_1 \sum_{i=1}^N x_i = \sum_{i=1}^N y_i \Longleftrightarrow \beta_0 + \beta_1 \cdot \frac{1}{N}\sum_{i=1}^N x_i = \frac{1}{N}\sum_{i=1}^N y_i \Longleftrightarrow \beta_0 + \beta_1 \bar{x} = \bar{y}$$

   (b) From $\beta_0 = \bar{y} - \beta_1 \bar{x}$,

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

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

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

   Since $x_1, \cdots, x_N$ are not all equal, i.e., $\sum_{i=1}^N (x_i - \bar{x})^2 \neq 0$,

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

2. The slope of l $(\hat{\beta}_1)$ is

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

and the intercept ($\hat{\beta}_0$) is determined by

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

Considering $x_i - \bar{x} \mapsto x_i$, $y_i - \bar{y} \mapsto y_i$ $(i = 1, \ldots, N)$, the slope of l', $\hat{\beta}_1$, is

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

At this time, since $\bar{x} = \bar{y} = 0$, the intercept of l' is 0 (passes through the origin). Once $\hat{\beta}_1$ is obtained, the intercept $\hat{\beta}_0$ of l can be obtained using $\hat{\beta}_1$ and from (a'),

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

(a) For any $z \in \mathbb{R}^m$,

$$\begin{aligned} &Az = B^\top Bz = 0 \Rightarrow z^\top B^\top Bz = 0 \Rightarrow (Bz)^\top Bz = 0 \Rightarrow \|Bz\|^2 = 0 \Rightarrow Bz = 0 \\ &Bz = 0 \Rightarrow B^\top Bz = 0 \Rightarrow Az = 0 \end{aligned}$$

Thus,

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

(b) From (a), the kernels of linear mappings by A and B are equal. Also, by Proposition 4 (dimension theorem), the sum of the dimensions of the image and kernel of both A and B is m. Therefore, the dimensions of the images of A and B are equal, and by Proposition 4, the ranks of A and B are also equal.

5. Let $X \in \mathbb{R}^{N \times (p+1)}$ be a matrix where the first column is all ones.

(a) When $N < p + 1$, by Proposition 3,

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

Note that $X^\top X \in \mathbb{R}^{(p+1)\times(p+1)}$ is a square matrix. By Proposition 1, $X^\top X$ does not have an inverse matrix.

(b) When $N \ge p + 1$ and there are two identical columns in X, by Proposition 3,

$$\operatorname{rank}(X^\top X) \le \operatorname{rank}(X) < p + 1$$

Therefore, for the same reason as (a), $X^\top X$ does not have an inverse matrix.

6. (a) For $j = 0, 1, \ldots, p$,

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

The partial derivative with respect to β_j is

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

On the other hand, the j-th component of $X^\top y$ is $\sum_{i=1}^{N} x_{i,j}y_i$, the (j,k) component of $X^\top X$ is $\sum_{i=1}^{N} x_{i,j}x_{i,k}$, and the j-th component of $X^\top X\beta$ is $\sum_{k=0}^{p}\sum_{i=1}^{N} x_{i,j}x_{i,k}\beta_k$. Therefore, the j-th component of $-X^\top y + X^\top X\beta$ is

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

which matches $\frac{\partial L}{\partial \beta_j}$, hence the statement is proved.

(b) From the calculation in (a), $\beta \in \mathbb{R}^{p+1}$ such that $\frac{\partial L}{\partial \beta_j} = 0$ for all j satisfies

$$-X^\top y + X^\top X\beta = 0$$

Assuming that $X^\top X$ has an inverse matrix, the desired $\hat{\beta}$ is

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

7. (a) Given the condition $y = X\beta + \varepsilon$, substituting this into Proposition 11, we get

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

(b) Since $\varepsilon \sim N\left(0, \sigma^2 I\right)$, the mean of $\varepsilon \in \mathbb{R}^N$ is 0, so even when multiplied by the constant matrix $\left(X^\top X\right)^{-1} X^\top$, the mean of $\left(X^\top X\right)^{-1} X^\top \varepsilon$ is also 0. Therefore, from (a), $\mathbb{E}[\hat{\beta}] = \beta$, hence the statement is proved.

(c) From (a),

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

Here, we used the covariance matrix of ε, $\mathbb{E}\left[\varepsilon\varepsilon^\top\right] = \sigma^2 I$.

8. Let $H = X\left(X^\top X\right)^{-1} X^\top \in \mathbb{R}^{N\times N}$, and $\hat{y} = X\hat{\beta}$, then

(a) $H^2 = X\left(X^\top X\right)^{-1} X^\top X \left(X^\top X\right)^{-1} X^\top = X\left(X^\top X\right)^{-1} X^\top = H$ implies that $H^2 = H$.

(b) $(I-H)^2 = I - 2H + H^2 = I - 2H + H = I - H$ implies that $(I-H)^2 = I - H$.

(c) $HX = X\left(X^\top X\right)^{-1} X^\top X = X$ implies that $HX = X$.

(d) By Proposition 11, $\hat{y} = X\hat{\beta} = X\left(X^\top X\right)^{-1} X^\top y = Hy$ implies that $\hat{y} = Hy$.

(e) $y - \hat{y} = y - Hy = (I-H)(X\beta + \varepsilon) = (X - HX)\beta + (I-H)\varepsilon = (X-X)\beta + (I-H)\varepsilon = (I-H)\varepsilon$ implies that $y - \hat{y} = (I-H)\varepsilon$. The first equality uses (d), the next equality uses (1.12), and the penultimate equality uses (c).

(f) $\|y-\hat{y}\|^2 = (y-\hat{y})^\top(y-\hat{y}) = \{(I-H)\varepsilon\}^\top(I-H)\varepsilon = \varepsilon^\top(I-H)^\top(I-H)\varepsilon = \varepsilon^\top(I-H)^2\varepsilon = \varepsilon^\top(I-H)\varepsilon$ implies that $\|y-\hat{y}\|^2 = \varepsilon^\top(I-H)\varepsilon$. Here, the second equality uses (e), the penultimate equality uses the linearity of transposition and

$$H^\top = \left\{X\left(X^\top X\right)^{-1} X^\top\right\}^\top = X\left\{\left(X^\top X\right)^{-1}\right\}^\top X^\top = X\left\{\left(X^\top X\right)^\top\right\}^{-1} X^\top = X\left(X^\top X\right)^{-1} X^\top = H$$

The final transformation relies on (b).

9. (a) Let $H := X(X^\top X)^{-1}X^\top$. By Proposition 3 and $\text{rank}(X) = p+1$,

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

Meanwhile, from (c) in the previous problem, $HX = X$ implies

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

Thus,

$$\text{rank}(HX) \le \min\{\text{rank}(H), \text{rank}(X)\} \le \text{rank}(H)$$

Therefore, $\text{rank}(H) \ge p+1$. Consequently, $\text{rank}(H) = p+1$, and by Proposition 4, the dimension of the image of H is $p+1$.

(b) From (c) in the previous problem: $HX = X$, the column vectors of X are eigenvectors of H with eigenvalue 1. Moreover, since $\text{rank}(X) = p+1$, the column vectors of X are linearly independent. Therefore, the column vectors of X form a basis for the eigenspace of H with eigenvalue 1, and its dimension is $p+1$. Here, the eigenspace of H with eigenvalue 1 is the kernel of H. Thus, by Proposition 4 and $\text{rank}(H) = p+1$, the dimension of the kernel is $N-p-1$, so the eigenspace of H with eigenvalue 0 is $N-p-1$ dimensions.

(c) For any $x \in \mathbb{R}^{p+1}$,

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

Thus, the eigenspace of H with eigenvalue 1 and the eigenspace of $I-H$ with eigenvalue 0 are equal, so the dimension of the eigenspace of $I-H$ with eigenvalue 0 is $p+1$. Furthermore,

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

Thus, the eigenspace of H with eigenvalue 0 and the eigenspace of $I-H$ with eigenvalue 1 are equal, so the dimension of the eigenspace of $I-H$ with eigenvalue 1 is $N-p-1$.

10. (a) Assume $\varepsilon \sim N\left(0, \sigma^2 I\right)$. Let $v = P\varepsilon$, then $\varepsilon = P^{-1}v = P^\top v$,

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

Here, $P(I-H)P^\top$ becomes a diagonal matrix with N eigenvalues as components. In particular, from the previous problem, $I-H$ has $N-p-1$ eigenvalues of 1 and $p+1$ eigenvalues of 0. Therefore,

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

(b) $\mathbb{E}[vv^\top] = P\mathbb{E}[\varepsilon\varepsilon^\top]P^\top = P\sigma^2 IP^\top = \sigma^2 IPP^\top = \sigma^2 I$

(c) Let $V=[v_1,\ldots,v_N]$, where $v_i \sim N\left(0,\sigma^2\right)$ for $i=1,\ldots,N$. Then, $Z_i = \frac{V_i}{\sigma}$ are mutually independent and each follows $N(0,1)$. Therefore,

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

From (a),

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

11. (a)

$$\mathbb{E}\left[(\hat{\beta}-\beta)(y-\hat{y})^\top\right] = \mathbb{E}\left[\left(X^\top X\right)^{-1} X^\top \varepsilon\varepsilon^\top (I-H)\right]$$
$$= \left(X^\top X\right)^{-1} X^\top \mathbb{E}\left[\varepsilon\varepsilon^\top\right](I-H) = \sigma^2\left\{\left(X^\top X\right)^{-1} X^\top - \left(X^\top X\right)^{-1} X^\top H\right\}$$
$$= \sigma^2\left\{\left(X^\top X\right)^{-1} X^\top - \left(X^\top X\right)^{-1} X^\top X\left(X^\top X\right)^{-1} X^\top\right\} = 0$$

(b) For $i=0,1,\cdots,p$, $\left(\hat{\beta}_i-\beta_i\right)/\left(\sqrt{B_i}\sigma\right)$ is a function of $\hat{\beta}-\beta$, and RSS is a function of $y-\hat{y}$. Since $\hat{\beta}-\beta$ and $y-\hat{y}$ both follow a normal distribution, and from (a), their covariance matrix is 0, the two are independent. Therefore, the statement is proved.

(c) For $i=0,1,\cdots,p$,

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

Since $\hat{\beta} \sim N\left(\beta, \sigma^2\left(X^\top X\right)^{-1}\right)$,

$$\frac{\hat{\beta}_i-\beta_i}{\sqrt{B_i}\sigma} \sim N(0,1)$$

Combining with 10.(c),

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

(d) Let $\sum_i = \sum_{i=1}^N$.

$$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix}$$

For this,

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

Therefore,

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

Thus, the statement is proved.

14. (a) From the equality $HX = X$ shown in 8.(c), since the first column vector of X has all its elements equal to 1, each column vector of W, with all elements equal to $1/N$, becomes an eigenvector of H with eigenvalue 1. Therefore, $HW = W$ holds. Furthermore,

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

Here, 8.(a) $H^2 = H$ was used.

(b)

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

Thus, the statement is proved.

(c)

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

To show the independence of $RSS = \|(I - H)\varepsilon\|^2$, it suffices to show the independence of $(I - H)\varepsilon$ and $(H - W)\varepsilon$, which both follow a normal distribution. Their covariance matrices are,

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

Thus, the independence in the statement is proved.

(d) Using (a),

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

Thus, the statement is proved.

15. (a) For the i-th component $\hat{y}_i - \bar{y}_i$ of $\hat{y} - \bar{y}$,

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

Thus,

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

Here, $\bar{x} \in \mathbb{R}^N$ is a column vector with all elements equal to $\frac{1}{N}\sum_{i=1}^N x_i$.

(b) Using (a),

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

(c) From (b),

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

Meanwhile, the sample correlation coefficient $\hat{r}$ is

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

Therefore,

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

This is equal to R^2.

17. (a) Note that x_* is a constant.

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

Considering also 7.(c),

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

(b)

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

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

Since $\hat{\beta} \sim N\left(\beta, \sigma^2\left(X^\top X\right)^{-1}\right)$, the numerator $\sim N(0,1)$. On the other hand, by 11.(b), $RSS/\sigma^2 \sim \chi^2_{N-p-1}$, and since these two are also shown to be independent,

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

(c)

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

Therefore,

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

Since the numerator $\sim N(0,1)$ and $RSS/\sigma^2 \sim \chi^2_{N-p-1}$, we obtain

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

Chapter 3: Classification

19. For $f(y) = \dfrac{1}{1+e^{-y(\beta_0 + x^T\beta)}}$,

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

Thus, the proposition is shown.

20. For $f(x) = \dfrac{1}{1+e^{-(\beta_0+x\beta)}}$, $f'(x) = \dfrac{\beta e^{-(\beta_0+x\beta)}}{\left\{1+e^{-(\beta_0+x\beta)}\right\}^2}$

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

Since $\beta > 0$, for any $x \in \mathbb{R}$, $f'(x) > 0$, $x < -\beta_0/\beta$ has $f''(x) > 0$, and $x > -\beta_0/\beta$ has $f''(x) < 0$. Therefore, $f(x)$ is monotonically increasing for any $x \in \mathbb{R}$, convex down for $x < -\beta_0/\beta$, and convex up for $x > -\beta_0/\beta$. The results of the implementation of the proposition are shown in the outline (R program Chapter 2). This is a graph of $y = f(x)$ where $\beta_0 = 0$ and β is $0, 0.2, 0.5, 1, 2, 10$. It shows that as β increases, y changes more sharply around $x = 0$, transitioning from $y = -1$ to $(x, y) = (0, 0)$ to $y = 1$.

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

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

(with $x_{i0} = 0$ assumed),

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

Then, for $j = 0, 1, \cdots,$

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

holds. Here, for $i = 1, 2, \cdots, N$, $v_i = \exp\left\{-y_i \sum_{k=0}^{p}(x_{ik}\beta_k)\right\}$ Thus,

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

can be written as $u = \begin{pmatrix} \frac{y_1 v_1}{1+v_1} \\ \vdots \\ \frac{y_N v_N}{1+v_N} \end{pmatrix}$ so that $\nabla l(\beta_0, \beta) = -X^T u$ Additionally, for $i = 1, 2, \cdots, j = 0, 1, \cdots, p$,

$$\frac{\partial v_i}{\partial \beta_j} = -y_i x_{ij} v_i$$

holds, so for $j, k = 0, 1, \cdots, p$,

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

Here, since $y_i \in \{-1, 1\}$, we have $y_i^2 = 1$. At this point, let W be an N-dimensional diagonal matrix with (i, i) element $v_i / (1 + v_i)^2$,

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

Then, the (i,k) element $(i=1,\cdots,N, k=0,1,\cdots,p)$ of $WX \in \mathbb{R}^{N\times(p+1)}$ is

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

Therefore, the (j,k) element $(j,k=0,1,\cdots,p)$ of $X^TWX \in \mathbb{R}^{(p+1)\times(p+1)}$ is

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

Thus, the desired second derivative $\nabla^2 l\left(\beta_0,\beta\right)$ is

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

Here, for any $i=1,\cdots,N$, since $v_i>0$, $v_i/\left(1+v_i\right)^2>0$, we can define $U\in\mathbb{R}^{N\times N}$ with each element being the square root of each element of W,

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

Then, since $W=U^TU$, we have $\nabla^2 l\left(\beta_0,\beta\right) = X^T\left(U^TU\right)X = (UX)^TUX$. Thus, using Proposition 10.1, $\nabla^2 l\left(\beta_0,\beta\right)$ is a non-negative definite matrix, and therefore $l\left(\beta_0,\beta\right)$ is convex.

23. Rewriting the update rule using $\beta_{\text{old}},\beta_{\text{new}},u,W,X$,

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

then,

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

Thus, letting $z = X\beta_{\text{old}}+W^{-1}u$, the update rule is

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

25. When considering the maximization of the likelihood $\prod_{i=1}^{N}\frac{1}{1+\exp\left\{-y_i\left(\beta_0+\beta^Tx_i\right)\right\}}$, if $y_i\left(\beta_0+\beta^Tx_i\right)\geq 0$ holds for any $i=1,\cdots,N$, then for any fixed β_0,β, for instance by substituting $\beta_0\leftarrow 2\beta_0, \beta\leftarrow 2\beta$, the exponent part in (2.1) can be made smaller. Therefore, the maximum value

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

does not exist, and under this assumption, the parameters for logistic regression cannot be estimated by maximum likelihood.

26. The accuracy rate was $(39+42)/100 = 0.81$.

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) Assuming $\pi_k = \pi_l$,

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

Taking the logarithm of both sides of the above equation,

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

Thus, $S_{k,l}$ is given in the form as required.

(b) Assuming $\Sigma_k = \Sigma_l = \Sigma$,

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

Taking the logarithm of both sides of the above equation,

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

Here, since Σ is a covariance matrix, it is a symmetric matrix,

$$\begin{aligned} \Sigma\Sigma^{-1} = I &\iff \left(\Sigma\Sigma^{-1}\right)^T = I \\ &\iff \left(\Sigma^{-1}\right)^T \Sigma^T = I \\ &\iff \left(\Sigma^{-1}\right)^T \Sigma = I \end{aligned}$$

Thus, noting that Σ^{-1} (denoted as (s_{ij})) is also a symmetric matrix,

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

Therefore,

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

The desired a, b are

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

(c) For the plane equation obtained in (b) where $\Sigma_k=\Sigma_l$ and $\pi_k=\pi_l$,

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

Thus, the boundary is the plane $x=\left(\mu_k-\mu_l\right)/2$.

Chapter 4: Resampling

32. It is sufficient to show

$$\left(A+UCV\right)\left(A^{-1}-A^{-1}U\left(C^{-1}+VA^{-1}U\right)^{-1}VA^{-1}\right)=I$$

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

Thus, the proposition is shown.

33. (a) For

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

$$X^T X = \sum_{i=1}^{N} x_i^T x_i = \sum_{i \in S} x_i^T x_i + \sum_{i \notin S} x_i^T x_i = X_S^T X_S + X_{-S}^T X_{-S}$$

Noting the equation shown in 32,

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

we have,

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

where $H_S = X_S \left(X^T X\right)^{-1} X_S^T$. Thus, the proposition is shown.

(b)

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

Thus, the proposition is shown.

34.

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

Therefore, the sum of squared errors for all CV groups can be written as,

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

Thus, the proposition is shown.

35. The sum of squared errors matches, and the execution time of cv.fast is shorter.

39. The first three types of data are. For $j = 1, 2, 3$, the intercept and two slope estimates when the first variable is regressed on the third and fourth variables are obtained, and the standard deviations of those estimates are evaluated.

Chapter 5 Information Criteria

40. (a)

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

Thus, when $\sigma^2 > 0$ is known, maximizing l with respect to β is equivalent to minimizing $\|y - X\beta\|^2$.

(b) Differentiating l with respect to σ^2 gives,

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

Setting $\dfrac{\partial l}{\partial \sigma^2} = 0$ gives the maximum likelihood estimate $\hat{\sigma}^2$ as

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

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

(c) Using the fact that for any $x > 0$, $\log x \leq x - 1$, for any probability density functions f and g on $\mathbb{R}$,

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

Finally, using the fact that f and g are probability density functions, we have shown the desired inequality.

41. (a)

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

Therefore,

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

(b) Since $f^N(y \mid x, \beta)$ is a joint probability density function, we have

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

Assuming that differentiation with respect to β and integration with respect to y can be interchanged, differentiating both sides with respect to β gives,

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

(c)

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

(d) Differentiating both sides of the equation obtained in (c) with respect to β gives,

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

Thus, the desired result is shown. Therefore, from (d),

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

is satisfied.

42. (a) For an unbiased estimator $\tilde{\beta}$ of β,

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

is satisfied, so differentiating both sides with respect to β_j gives,

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

Expressing this in the form of a covariance matrix,

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

Therefore, $E\left[(\tilde{\beta} - \beta)(\nabla l)^T \right] = I$. Finally, using the fact that $E[\nabla l] = 0$,

(b) The desired covariance matrix is,

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

(c)

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

These are non-negative definite matrices, so for any $x, y \in \mathbb{R}^{p+1}$, let $z = [x, y]^T$,

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

This is satisfied even when $y = 0$, so $V(\tilde{\beta}) - (NJ)^{-1}$ is non-negative definite. Thus, the validity of the Cramer-Rao inequality is shown.

43. (a) Taking the trace of both sides of $E\left[(\tilde{\beta}-\beta)(\nabla l)^T\right] = I \in \mathbb{R}^{(p+1)\times(p+1)}$,

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

Thus, the desired result is shown.

(b)

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

(c)

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

Using Schwartz's inequality from the second to the third line. Therefore,

$$\begin{aligned} (p+1)^2 &\leq \frac{p+1}{\sigma^2}E\left[\|X(\tilde{\beta}-\beta)\|^2\right] \\ E\left[\|X(\tilde{\beta}-\beta)\|^2\right] &\geq (p+1)\sigma^2 \end{aligned}$$

44. (a)

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

Thus,

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

Taking the sum over $(x, u) = (x_1, z_1), \cdots, (x_N, z_N)$, we get

$$\begin{aligned}&-\sum_{i=1}^{N}\log f(z_i \mid x_i, \gamma)\\ &= \frac{N}{2}\log 2\pi\sigma^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{N}(z_i - x_i\beta)^2 - \frac{1}{\sigma^2}\sum_{i=1}^{N}(\gamma-\beta)^T x_i^T (z_i - x_i\beta) + \frac{1}{2\sigma^2}\sum_{i=1}^{N}(\gamma-\beta)^T x^T x(\gamma-\beta)\\ &= \frac{N}{2}\log 2\pi\sigma^2 + \frac{1}{2\sigma^2}\|z - X\beta\|^2 - \frac{1}{\sigma^2}(\gamma-\beta)^T X^T (z - X\beta) + \frac{1}{2\sigma^2}(\gamma-\beta) X^T X(\gamma-\beta)\end{aligned}$$

Thus, the desired result is shown.

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

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

(c) The value of (4.8) is given by

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

The sum of the KL information is,

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

Taking $\gamma = \hat{\beta}$, obtained by the least squares method,

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

Thus, the average minimum value of (b) is

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

which is achieved by the least squares method.

(d)

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

Since the second term and beyond do not depend on k, the desired result is shown.

45.

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

(a) Using the Maclaurin series expansion of $\log(x+1)$,

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

we get,

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

(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)

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

For each j, the coefficient of $m^{n-j}\prod_{i=1}^{j}(m+2(i-1))$ is 1 for n. Therefore, the coefficient of the nth term is

$$\sum_{j=0}^{n}(-1)^j\left(\begin{array}{c}n\\j\end{array}\right)=\sum_{j=0}^{n}(-1)^j1^{n-j}\left(\begin{array}{c}n\\j\end{array}\right)=(-1+1)^n=0$$

(e) For each j, the coefficient of the $n-1$ term of $m^{n-j}\prod_{i=1}^{j}(m+2(i-1))$ is

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

Therefore, the coefficient of the $n-1$ term is

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

(f) For $n \geq 3$, from (d) and (e),

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

By setting

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

we get

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

Thus,

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

Thus, the desired result is shown.

Chapter 6 Sparse Estimation

49.

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

Differentiating with respect to β gives

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

Thus, for $\dfrac{\partial L}{\partial \beta} = 0$,

$$\left(X^T X + N\lambda I\right)\beta = X^T y$$

For a solution $\beta = \hat{\beta}$ to exist, $X^T X + N\lambda I$ must be invertible.

Assuming $\lambda > 0$, since $X^T X \in \mathbb{R}^{p\times p}$ is non-negative definite, all eigenvalues $\mu_1, \cdots, \mu_p$ of $X^T X$ are non-negative. Thus, the characteristic polynomial of $X^T X + N\lambda I$ is $\varphi(t)$,

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

Since $N\lambda > 0$, all roots, i.e., eigenvalues of $X^T X + N\lambda I$ are non-negative.

Conversely, if $X^T X + N\lambda I$ is invertible, then for any $i = 1, \cdots, p$,

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

Since μ_i is an eigenvalue of $X^T X$ and $X \in \mathbb{R}^{N\times p}$ is arbitrary, μ_i can take any non-negative value. Thus, for (49.1) to always hold, $\lambda > 0$ is necessary. Therefore, the desired result is shown.

50. (a)

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

For (50.1) to hold for $x > x_0$,

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

is necessary. For (50.1) to hold for $x < x_0$,

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

is necessary. Thus, z must be greater than or equal to the left derivative of f at $x = x_0$ and less than or equal to the right derivative. Since f is differentiable at $x = x_0$, $z = f'(x_0)$ is necessary. Conversely, when $z = f'(x_0)$, since f is a convex function, (50.1) holds. Thus, the desired result is shown.

(b) For $zx \leq |x|$ to hold,

$$\begin{cases} x = 0 \text{ always holds} \\ x > 0 \text{ needs } z \leq 1 \\ x < 0 \text{ needs } z \geq -1 \end{cases}$$

Thus, for (50.2) to hold, $|z| \leq 1$ is necessary. Conversely, if $|z| \leq 1$,

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

holds, so (50.2) is satisfied. Therefore, the desired result is shown.

(c) i. When $x_0 < 0$, the subdifferential is $\{-1\}$

ii. When $x_0 = 0$,

$$f(x) \geq f\left(x_0\right) + z\left(x - x_0\right) \Leftrightarrow |x| \geq zx$$

so the subdifferential is $[-1, 1]$.

iii. When $x_0 > 0$, the subdifferential is $\{1\}$

(d) For $f(x) = x^2 - 3x + |x|$,

$$f(x) = \left\{ \begin{array}{ll} x^2 - 2x & x \geq 0 \\ x^2 - 4x & x < 0 \end{array} \right.$$

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

Thus, this $f(x)$ has a minimum at $x = 1$. Next, for $f(x) = x^2 + x + 2|x|$,

$$f(x) = \left\{ \begin{array}{ll} x^2 + 3x & x \geq 0, \\ x^2 - x & x < 0 \end{array} \right.$$

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

Thus, this $f(x)$ has a minimum at $x = 0$.

51. $\mathcal{S}_\lambda(x)$ can be written using the sign function

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

as

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

52.

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

The subdifferential is

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

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

Thus,

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

Using $\mathcal{S}_\lambda(x)$, we can write $\beta=S_\lambda(z)$.

55. The function cv.glmnet performs 10-fold cross-validation to determine the optimal value of λ for performing Lasso. The glmnet function takes the response variable, explanatory variables, and λ as arguments and performs Lasso. The selected variables are V3, V4, V5.

56. (a) Differentiating S with respect to β_1 and substituting $(\beta_1,\beta_2)=\left(\hat{\beta}_1,\hat{\beta}_2\right)$, we get $0=\sum_{i=1}^N -2x_{i,1}\left(y_i-\hat{\beta}_1x_{i,1}-\hat{\beta}_2x_{i,2}\right)$. Thus,

$$\sum_{i=1}^N x_{i,1}\left(y_i-\hat{\beta}_1x_{i,1}-\hat{\beta}_2x_{i,2}\right)=0$$

Similarly, differentiating with respect to β_2, we get

$$\sum_{i=1}^N x_{i,2}\left(y_i-\hat{\beta}_1x_{i,1}-\hat{\beta}_2x_{i,2}\right)=0$$

Then,

$$\begin{aligned} & y_i-\beta_1x_{i,1}-\beta_2x_{i,2} \\ =& y_i-\hat{y}_i+\hat{y}_i-\beta_1x_{i,1}-\beta_2x_{i,2} \\ =& y_i-\hat{y}_i+\hat{\beta}_1x_{i,1}+\hat{\beta}_2x_{i,2}-\beta_1x_{i,1}-\beta_2x_{i,2} \\ =& y_i-\hat{y}_i-\left(\beta_1-\hat{\beta}_1\right)x_{i,1}-\left(\beta_2-\hat{\beta}_2\right)x_{i,2} \end{aligned}$$

Moreover,

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

Expanding the desired sum,

$$\begin{aligned} & \sum_{i=1}^N \left(y_i-\beta_1x_{i,1}-\beta_2x_{i,2}\right)^2 \\ =& \sum_{i=1}^N \left[\left(y_i-\hat{y}_i\right)-\left\{\left(\beta_1-\hat{\beta}_1\right)x_{i,1}+\left(\beta_2-\hat{\beta}_2\right)x_{i,2}\right\}\right]^2 \\ =& \sum_{i=1}^N \left(y_i-\hat{y}_i\right)^2+\sum_{i=1}^N\left\{\left(\beta_1-\hat{\beta}_1\right)x_{i,1}+\left(\beta_2-\hat{\beta}_2\right)x_{i,2}\right\}^2 \\ =& \left(\beta_1-\hat{\beta}_1\right)^2\sum_{i=1}^N x_{i,1}^2+2\left(\beta_1-\hat{\beta}_1\right)\left(\beta_2-\hat{\beta}_2\right)\sum_{i=1}^N x_{i,1}x_{i,2}+\left(\beta_2-\hat{\beta}_2\right)^2\sum_{i=1}^N x_{i,2}^2+\sum_{i=1}^N\left(y_i-\hat{y}_i\right)^2 \end{aligned}$$

(b) Let A$(1,0)$, B$(0,1)$, C$(-1,0)$, D$(0,-1)$. The range of $\left(\hat{\beta}_1, \hat{\beta}_2\right)$ is obtained by excluding the portions that touch the four sides (excluding vertices) when a circle centered at $\left(\hat{\beta}_1, \hat{\beta}_2\right)$ and a square touch the four sides of $\left(\hat{\beta}_1, \hat{\beta}_2\right)$.
For the side AB, the range is the portion above AB, excluding the portion between the lines AD and BC. Similar consideration for the other three sides gives the range of $\left(\hat{\beta}_1, \hat{\beta}_2\right)$.

(c) When the unit circle centered at the origin is considered instead of a square, the range can be written as

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

Chapter 8 Nonlinear Regression

57. (a) $L = \sum_{i=1}^{N}\left(y_i - \sum_{j=0}^{p}\beta_j x_i^j\right)^2$ For this,

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

we set $L = \|y - X\beta\|^2$. Therefore, $\hat{\beta}$ that minimizes L can be written as

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

under the assumption that $X^T X$ is regular. Here, since $\operatorname{rank} X = \operatorname{rank} X^T X$, the condition for $X^T X \in \mathbb{R}^{(p+1)\times(p+1)}$ to be regular is $\operatorname{rank} X = p+1$. Assuming that among $x_1, \cdots, x_N$, there are $p+1$ distinct values, let them be $x_{(1)}, \cdots, x_{(p+1)}$, and consider the matrix

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

From the Vandermonde inequality,

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

we see that $\operatorname{rank} X' = p+1$, hence $\operatorname{rank} X = p+1$. Conversely, when there are p or fewer distinct values among $x_1, \cdots, x_N$, $\operatorname{rank} X < p+1$, so $X^T X$ is not regular. Therefore, the condition for $\beta_0, \beta_1, \cdots, \beta_p$ to be uniquely determined is that there are at least $p+1$ distinct values among $x_1, \cdots, x_N$. Under this condition, the solution is

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

(b) Similarly to (a),

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

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

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \in \mathbb{R}^N \ , \ X = \begin{pmatrix} 1 & f_1\left(x_1\right) & f_2\left(x_1\right) & \cdots & f_p\left(x_1\right) \\ 1 & f_1\left(x_2\right) & f_2\left(x_2\right) & \cdots & f_p\left(x_2\right) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & f_1\left(x_N\right) & f_2\left(x_N\right) & \cdots & f_p\left(x_N\right) \end{pmatrix} \in \mathbb{R}^{N \times (p+1)} \ , \ \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \in \mathbb{R}^{p+1}$$

$(f_0(\cdot) = 1)$ then, $L = \|y - X\beta\|^2$ and the $\beta = \hat{\beta}$ that minimizes L, assuming $X^T X$ is regular, is

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

The condition for $X^T X \in \mathbb{R}^{(p+1) \times (p+1)}$ to be regular is $N \geq p+1$ and the $p+1$ column vectors of X being linearly independent.

58. (a) For each $i = 1, \cdots, k+1$ and the cubic polynomials f_{i-1}, f_i,

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

holds, so

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

then,

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

so,

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

hence, $\varepsilon_j = \delta_j$ $(j = 0, 1, 2)$ holds. Therefore,

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

so, the required γ_i can be set as $\gamma_i = \varepsilon_3 - \delta_3$. Thus, the statement is proven.

(b) There exist $K + 4$ constants $\beta_1, \beta_2, \cdots, \beta_{K+4}$ such that

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

holds, i.e., for any $i = 0, 1, \cdots, K$ and any $x \in [\alpha_i, \alpha_{i+1}]$,

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

holds. For $i = 0$, in $x \in [\alpha_0, \alpha_1]$, since $f(x) = f_0(x)$, there exist unique $\beta_1, \beta_2, \beta_3, \beta_4$ such that

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

so $(*)$ holds. For $i = 1$, in $x \in [\alpha_1, \alpha_2]$, since $f_1(x) = f_0(x) + \gamma_1 (x - \alpha_1)^3$, by setting $\beta_5 = \gamma_1$, $(*)$ holds. Once $\beta_1, \cdots, \beta_{i+4}$ are determined, i.e., the coefficients $\beta_1, \beta_2, \cdots, \beta_{i+4}$ for $f(x)$ in $x \in [\alpha_0, \alpha_{i+1}]$ are determined,

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

by setting $\beta_{i+5} = \gamma_{i+1}$, for $x \in [\alpha_{i+1}, \alpha_{i+2}]$,

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

holds. Hence, by considering up to $i = K - 1$, we can determine $\beta_1, \cdots, \beta_{K+4}$ that satisfy the required equation. Thus, the statement is proven.

60. (a)

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

To show that the natural cubic spline curve $g(x)$ becomes a straight line for $x \geq \alpha_K$ and that the first and second derivatives match at the boundary $x = \alpha_K$,

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

is required. Therefore,

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

hence, $6 \sum_{i=3}^{K+1} \gamma_i = 0$, therefore,

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

(b) For $j = 1, \cdots, K-1$ and $x \geq \alpha_K$,

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

so for $j = 1, \cdots, K-2$,

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

is true. Simplifying the numerator of equation (60.1),

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

hence,

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

(c) First, for any $x \leq \alpha_1$,

$$d_j(x) = 0$$

thus, for any $x \leq \alpha_1$,

$$g(x) = \gamma_1 + \gamma_2 x + \sum_{j=3}^{K} \gamma_j \left\{d_{j-2}(x) - d_{K-1}(x)\right\} = \gamma_1 + \gamma_2 x$$

is a linear function of x. Next, for any $x \geq \alpha_K$,

$$g(x) = \gamma_1 + \gamma_2 x + \sum_{j=3}^{K} \gamma_j h_j(x)$$

since $h_j(x)$ is at most a linear polynomial for $j = 3, \cdots, K$, $g(x)$ is also linear. Thus, the statement is proven.

62. (a) For a natural cubic spline function g, $g''(x_1) = g''(x_N) = 0$ and the third derivatives being constant (denoted γ_i) in each interval $[x_i, x_{i+1}]$,

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

(b) Under the assumption of $\int_{x_1}^{x_N} g''(x)h''(x)dx$,

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

Lastly, from 60.(c), since $g''(x) = 0$ for $x \notin [x_1, x_N]$. Thus, the statement is proven.

(c) For a natural cubic spline function g, if $g(x_i) = f(x_i)$ for $i = 1, \cdots, N$, then setting $h(x) = f(x) - g(x)$, $h(x_i) = 0$ for $i = 1, \cdots, N$. Combining this with the inequality in (b),

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

64. The required smoothing spline function g minimizes

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

over all $f : \mathbb{R} \to \mathbb{R}$. Then, the first term of $RSS(g, \lambda)$ is

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

and the second term is

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

thus,

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

Differentiating both sides by γ and simplifying,

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

Thus, the first part of the statement is proven.

67. (a) Using matrices,

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

This can be rewritten as,

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

so, $\sum_{i=1}^{N} K\left(x, x_i\right)\left(y_i - \left[1, x_i\right]\beta(x)\right)^2 = (y - X\beta(x))^T W'(y - X\beta(x))$ Differentiating by β, $-2X^T W'(y - X\beta(x))$ which equals zero, $X^T W' y = X^T W' X\beta(x)$ hence,

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

So,

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

W is a diagonal matrix with $K\left(x, x_1\right), \cdots, K\left(x, x_N\right)$ as its diagonal elements.

Chapter 9 Support Vector Machines

75. (a)

$$\begin{cases} \epsilon_i = 0 & \text{if it is on the correct margin or not,} \\ 0 < \epsilon_i < 1 & \text{if it is between the margin and the boundary,} \\ \epsilon_i = 1 & \text{if it is on the boundary,} \\ \epsilon_i > 1 & \text{if it is on the opposite side of the boundary.} \end{cases}$$

(b) Suppose there exists a solution $M > 0$ such that for at least r different $i \in \{1, \cdots, n\}$, $y_i(\beta_0 + x_i\beta) < 0$. For each such i,

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

implies $\epsilon_i > 1$, and thus $\sum_{i=1}^N \epsilon_i > r$. If $\gamma \leq r$, then

$$\sum_{i=1}^N \epsilon_i \leq \gamma \leq r$$

which is a contradiction.

(c) For some $\gamma = \gamma_0$, let $(\beta, \beta_0, \epsilon_i)$ maximize M to M_0. For $\gamma = \gamma_1 > \gamma_0$, the conditions are still satisfied, and the optimal value of M is at least M_0.

76. Consider (8.24). If there exists an index $j \in \{1, \cdots, m\}$ such that $f_j(\beta) > 0$, we can increase the value of α_j to make $L(\alpha, \beta)$ arbitrarily large. On the other hand, if $f_j(\beta) \leq 0$ for all $j \in \{1, \cdots, m\}$, setting $\alpha_1 = \cdots = \alpha_m = 0$ gives the maximum value of $L(\alpha, \beta)$ as $f_0(\beta)$. Thus, (8.24) is proven. Next, for any $\alpha \in [0, \infty)^m, \beta \in \mathbb{R}^p$,

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

which implies

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

for any $\alpha \in [0, \infty)^m, \beta \in \mathbb{R}^p$. This inequality holds even when taking inf over β on the left and sup over α on the right, implying (8.25). Now,

$$(p, m) = (2, 1)$$
$$L(\alpha, \beta) = \beta_1 + \beta_2 + \alpha(\beta_1^2 + \beta_2^2 - 1)$$

Thus,

$$f_0(\beta) = \beta_1 + \beta_2$$
$$f_1(\beta) = \beta_1^2 + \beta_2^2 - 1$$
$$\alpha_1 = \alpha$$

From (8.24),

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

This is minimized at $\beta_1 = \beta_2 = -1/\sqrt{2}$, giving the minimum value $-\sqrt{2}$. Hence, the left side of (8.25) is $-\sqrt{2}$. Next,

$$\frac{\partial L}{\partial \beta_1} = \frac{\partial L}{\partial \beta_2} = 0$$

yields

$$\begin{cases} 1 + 2\alpha\beta_1 = 0 \\ 1 + 2\alpha\beta_2 = 0 \end{cases}$$

implying $\beta_1 = \beta_2 = -1/(2\alpha)$. Thus,

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

The maximum value of this is $-\sqrt{2}$ for $\alpha = -1/\sqrt{2}$. Thus, equality holds in (8.25).

77. (a) From (8.30), let $(f, x_0, x) \mapsto (f_0, \beta^*, \beta)$:

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

where $\alpha_i \ge 0$ and $f_i(\beta) \le 0$ for $i = 1, \cdots, m$.

(b) For (8.26):

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

Thus, (8.27), (8.28), (8.29) become

$$\left\{ \begin{array}{l} \beta_1^2 + \beta_2^2 - 1 \le 0, \\ \alpha(\beta_1^2 + \beta_2^2 - 1) = 0, \\ \left(\begin{array}{c} 1 \\ 1 \end{array}\right) + \alpha \left(\begin{array}{c} 2\beta_1 \\ 2\beta_2 \end{array}\right) = \left(\begin{array}{c} 0 \\ 0 \end{array}\right) \end{array} \right.$$

78. First, from (8.27), for $i = 1, \cdots, N$,

$$\begin{aligned} &y_i(\beta_0 + x_i\beta) - (1 - \epsilon_i) \ge 0, \\ &\epsilon_i \ge 0 \end{aligned}$$

Next, from (8.28), for $i = 1, \cdots, N$,

$$\begin{aligned} &\alpha_i\{y_i(\beta_0 + x_i\beta) - (1 - \epsilon_i)\} = 0, \\ &\mu_i\epsilon_i = 0 \end{aligned}$$

and from (8.29),

$$\left\{ \begin{array}{l} \frac{\partial L_P}{\partial \beta_0} = 0, \\ \frac{\partial L_P}{\partial \beta} = \mathbf{0}, \\ \frac{\partial L_P}{\partial \epsilon_i} = 0, \end{array} \right.$$

yields

$$\left\{ \begin{array}{l} \sum_{i=1}^N \alpha_i y_i = 0, \\ \beta - \sum_{i=1}^N \alpha_i y_i x_i^T = 0, \\ C - \alpha_i - \mu_i = 0 \end{array} \right.$$

79. By optimizing L_P with respect to β_0, β, from (8.32) and (8.34), L_P is written as:

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

Using (8.33),

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

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

Thus, we can construct the function L_D with Lagrange multipliers $\alpha_i, \mu_i \geq 0, i = 1, \cdots, N$:

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

Given (8.32) and (8.34),

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

By solving this dual problem, we obtain $\alpha_i, i = 1, \cdots, N$ which gives β by substituting into (8.33).

80.

$$\begin{aligned} &y_i(\beta_0 + x_i\beta) > 1 \Rightarrow y_i(\beta_0 + x_i\beta) - (1 - \epsilon_i) > 0 \Rightarrow \alpha_i = 0, \\ &0 < \alpha_i < C \Rightarrow \mu_i > 0 \Rightarrow y_i(\beta_0 + x_i\beta) - (1 - \epsilon_i) = 0 \Rightarrow \epsilon_i = 0, \\ &y_i(\beta_0 + x_i\beta) < 1 \Rightarrow \epsilon_i > 0 \Rightarrow \mu_i = 0 \Rightarrow \alpha_i = C \end{aligned}$$

81. (a) When $\alpha_1 = \cdots = \alpha_N = 0$, from (8.33), $\beta = \mathbf{0}$. From (8.34) and (8.36), $\epsilon_1 = \cdots = \epsilon_N = 0$. Assuming $y_i(\beta_0 + x_i\beta) = 1$ for all $i = 1, \cdots, N$,

$$y_i\beta_0 = 1 > 0$$

Since $y_i \in \{-1, 1\}$, $y_1 = \cdots = y_N$ is necessary. Thus, for all $i = 1, \cdots, N$, $(y_i, \beta_0) = (\pm 1, \pm 1)$ (same sign). Hence, the statement is proven.

(b) If $\alpha_i = C$ implies $\epsilon_i > 0$ and $\alpha_i = 0$ implies $\epsilon_i = 0$, then,

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

By replacing each ϵ_i with $\epsilon_i - \epsilon_*$ and β_0 with $\beta_0 + y_i\epsilon_*$,

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

so

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

This does not change the value. Considering (8.37) with the contrapositive of the third proposition of 80, equality holds in (8.37) for both cases $\alpha_i = 0, C$. Therefore, all seven KKT conditions hold. After replacing,

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

is reduced, which is not an optimal solution for L_P.

(c) From (a) and (b) and the proposition 80, there exists at least one index i such that $0 < \alpha_i < C$ or $y_i(\beta_0 + x_i\beta) = 1$. In the former case, from the two propositions of 80, $y_i(\beta_0 + x_i\beta) = 1$ holds. Hence, the statement is proven.

82.

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

can be rewritten as

$$L_D = -\frac{1}{2}\alpha^T D_{mat}\alpha + d_{vec}^T\alpha$$

with the constraints

$$A_{mat}\alpha \geq b_{vec}$$

where

$$b_{vec} = [0, -C, \cdots, -C, 0, \cdots, 0]^T \in \mathbb{R}^m, \quad A_{mat} \in \mathbb{R}^{m\times N}, \quad meq \in \mathbb{N}, \quad D_{mat} \in \mathbb{R}^{N\times N}, \quad d_{vec} \in \mathbb{R}^N$$

Setting

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

and

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

we obtain the same dual problem.

83.

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

Thus, the mapping ϕ is

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

84. (a) For any $f, g, h \in V$ and $\alpha, \beta \in \mathbb{R}$,

$$\langle f, g\rangle = \int_0^1 f(x)g(x)dx$$

defines the inner product:

$$\begin{aligned}
\langle f, g\rangle &= \int_0^1 f(x)g(x)dx = \int_0^1 g(x)f(x)dx = \langle g, f\rangle, \\
\langle \alpha f + \beta g, h\rangle &= \int_0^1 \{\alpha f(x) + \beta g(x)\}h(x)dx = \alpha\int_0^1 f(x)h(x)dx + \beta\int_0^1 g(x)h(x)dx = \alpha\langle f, h\rangle + \beta\langle g, h\rangle, \\
\langle f, f\rangle &= \int_0^1 \{f(x)\}^2 dx \geq 0 \text{ (equality holds if and only if } f \equiv 0)
\end{aligned}$$

Thus, $\langle\cdot,\cdot\rangle$ is an inner product on V.

(b) For any $x, y \in \mathbb{R}^p$, let $\langle x, y\rangle = (1 + x^T y)^2$. Then,

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

which shows $\langle\cdot,\cdot\rangle$ is not linear and thus not an inner product on V.

Chapter 10 Unsupervised Learning

90. (a) For each $j = 1, \cdots, p$,

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

It is sufficient to show the above. For the left-hand side of (90.1),

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

The first and second terms are equal, and the third term is zero. Thus,

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

The right-hand side matches the right-hand side of (90.1). Therefore, the desired equality is proven.

(b) From (a), the score S can be written as:

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

In this case, for each $k = 1, \cdots, K$ and any vector x,

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

This shows that even taking two steps does not increase the score S.

(c) First, in the first case, 3 and 10 are the centers of clusters 1 and 2 respectively. The nearest cluster centers for 0, 6, and 10 are 3, 3, and 10 respectively, and continuing the process will not change this state. The score S in this case is:

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

In the second case, 0 and 8 are the centers of clusters 1 and 2 respectively. The nearest cluster centers for 0, 6, and 10 are 0, 8, and 8 respectively, and continuing the process will not change this state. The score S in this case is:

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

93. When applying Centroid Linkage, initially (5,8) and (9,0) are combined, and the cluster distance is $\sqrt{4^2 + 8^2} = 80$. At this point, the center is (7,4), and the distance between this and the other center (0,0) is $\sqrt{7^2 + 4^2} = \sqrt{65}$. Since this distance is smaller than the initial cluster distance, the dendrogram tree intersects.

94. (a) Under the condition that $\|\phi\|^2 = 1$, consider maximizing $\|X\phi\|^2$. The KKT condition gives:

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

Taking the derivatives $\partial L/\partial\gamma = 0$ and $\partial L/\partial\phi = 0$, we obtain:

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

Thus, ϕ must be an eigenvector of $X^T X$ and therefore of Σ:

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

The maximum γ is the largest eigenvalue of $X^T X$. Thus, $\phi = \phi_1$ is the eigenvector corresponding to λ_1 of Σ, satisfying:

$$\Sigma\phi_1 = \lambda_1\phi_1$$

(b) Vectors belonging to different eigenspaces of a symmetric matrix are orthogonal. Since all eigenvalues are distinct, each eigenspace has dimension 1, and the eigenvectors are orthogonal.

97.

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

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

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