7.1

Derive the estimate of in-sample error (7.24).

It suffices to show that

$$\sum_{i=1}^{N}\mathrm{Cov}(\hat{y}_i, y_i) = d\sigma_\epsilon^2.$$

Note that for a linear fit, we have $\hat{y} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$, so

$$\begin{aligned}\mathrm{Cov}(\hat{y}, y) &= \mathrm{Cov}(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y, y)\\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathrm{Cov}(y, y)\\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sigma_\epsilon^2.\end{aligned}$$

Therefore, by *cyclic property* of trace operator,

$$\begin{aligned}\sum_{i=1}^{N}\mathrm{Cov}(\hat{y}_i, y_i) &= \mathrm{trace}(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\sigma_\epsilon^2\\ &= \mathrm{trace}(\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1})\sigma_\epsilon^2\\ &= \mathrm{trace}(\mathbf{I}_d)\sigma_\epsilon^2\\ &= d\sigma_\epsilon^2.\end{aligned}$$

7.3

Let $\hat{\boldsymbol{f}} = \mathbf{S}\mathbf{y}$ be a linear smoothing of $\mathbf{y}$.

（a）

If S_{ii} is the ith diagonal element of $\mathbf{S}$, show that for $\mathbf{S}$ arising from least squares projections and cubic smoothing splines, the cross-validated residual can be written as

$$y_i - \hat{f}^{-i}(x_i) = \frac{y_i - \hat{f}(x_i)}{1 - S_{ii}}. \quad (1)$$

（b）

Use this result to show that $|y_i - \hat{f}^{-i}(x_i)| \geq |y_i - \hat{f}(x_i)|$.

（c）

Find general conditions on any smoother $\mathbf{S}$ to make result (1) hold.

Without loss of generality, we assume

$$\mathbf{S} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{\Omega})^{-1}\mathbf{X}^T.$$

For least squares we have $\lambda = 0$, and for cubic smoothing we have $\lambda \geq 0$. See Chapters 3 & 5 in the text for more details.

（a）

We have

$$\begin{aligned} S_{ii} &= x_i^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{\Omega})^{-1}x_i, \\ \hat{f}(x_i) &= x_i^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{\Omega})^{-1}\mathbf{X}^T\mathbf{y}. \end{aligned}$$

Let $\mathbf{X}_{-i}$ and $\mathbf{y}_{-i}$ be the corresponding results with x_i removed, then we have

$$\begin{aligned} \hat{f}^{-i}(x_i) &= x_i^T(\mathbf{X}_{-i}^T\mathbf{X}_{-i} + \lambda\mathbf{\Omega})^{-1}\mathbf{X}_{-i}^T\mathbf{y}_{-i} \\ &= x_i^T(\mathbf{X}^T\mathbf{X} - x_i x_i^T + \lambda\mathbf{\Omega})^{-1}(\mathbf{X}^T\mathbf{y} - x_i y_i). \quad (2) \end{aligned}$$

Let $\mathbf{A} = (\mathbf{X}^T\mathbf{X} + \lambda\Omega)$, by Woodbury matrix identity, we have

$$(\mathbf{A} - x_i x_i^T)^{-1} = \mathbf{A}^{-1} + \frac{\mathbf{A}^{-1}x_i x_i^T\mathbf{A}^{-1}}{1 - x_i^T\mathbf{A}^{-1}x_i}.$$

Therefore, (2) becomes

$$\begin{aligned} \hat{f}^{-1}(x_i) &= x_i^T\left(\mathbf{A}^{-1} + \frac{\mathbf{A}^{-1}x_i x_i^T\mathbf{A}^{-1}}{1 - x_i^T\mathbf{A}^{-1}x_i}\right)(\mathbf{X}^T\mathbf{y} - x_i y_i) \\ &= \left(x_i^T\mathbf{A}^{-1} + \frac{S_{ii}x_i^T\mathbf{A}^{-1}}{1 - S_{ii}}\right)(\mathbf{X}^T\mathbf{y} - x_i y_i) \\ &= x_i^T\mathbf{A}^{-1}\mathbf{X}^T\mathbf{y} - x_i^T\mathbf{A}^{-1}x_i y_i + \frac{S_{ii}x_i^T\mathbf{A}^{-1}\mathbf{X}^T\mathbf{y}}{1 - S_{ii}} - \frac{S_{ii}x_i^T\mathbf{A}^{-1}x_i y_i}{1 - S_{ii}} \\ &= \hat{f}(x_i) - y_i S_{ii} + \frac{S_{ii}\hat{f}(x_i)}{1 - S_{ii}} - \frac{y_i S_{ii}^2}{1 - S_{ii}} \\ &= \frac{\hat{f}(x_i) - y_i S_{ii}}{1 - S_{ii}}. \end{aligned}$$

Therefore by simple algebra we have (1).

（b）

Note that $\mathbf{S} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{\Omega})^{-1}\mathbf{X}^T$ is positive-semidefinite and has eigen-decomposition

$$\mathbf{S} = \sum_{k=1}^{N} \rho_k(\lambda)\mathbf{u}_k\mathbf{u}_k^T.$$

See Section 5.4.1 in the text for more details. Therefore, we know that $\mathbf{SS} \preceq \mathbf{S}$, so that

$$0 \le (S^2)_{ii} = \sum_{k \ne i} S_{ik}^2 + S_{ii}^2 \le S_{ii}$$

from which we know $0 \le S_{ii} \le 1$.

By (1) we have $|y_i - \hat{f}^{-i}(x_i)| \ge |y_i - \hat{f}(x_i)|$.

（c）

For general linear smoother $\hat{\boldsymbol{f}} = \mathbf{S}\mathbf{y}$, if $\mathbf{S}$ only depends on $\mathbf{X}$ and other tuning parameters (i.e., independent of $\mathbf{y}$), (1) still holds.

To see that, note that if we replace y_i with $\hat{f}^{-i}(x_i)$ (obtained by (2)) in $\mathbf{y}$ and denote the new vector by $\mathbf{y}'$, $\mathbf{S}$ is not changed. Thus we have

$$\begin{aligned}
\hat{f}^{-i}(x_i) &= (\mathbf{S}\mathbf{y}')_i \\
&= \sum_{i \ne j} S_{ij}\mathbf{y}'_j + S_{ii}\hat{f}^{-i}(x_i) \\
&= \hat{f}(x_i) - S_{ii}y_i + S_{ii}\hat{f}^{-i}(x_i),
\end{aligned}$$

therefore we obtain (1).

7.4

Consider the in-sample prediction error (7.18) and the training error $\overline{\text{err}}$ in the case of squared-error loss:

$$\begin{aligned}
\text{Err}_{\text{in}} &= \frac{1}{N}\sum_{i=1}^{N} E_{Y^0}(Y_i^0 - \hat{f}(x_i))^2 \\
\overline{\text{err}} &= \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{f}(x_i))^2.
\end{aligned}$$

Add and subtract $f(x_i)$ and $E\hat{f}(x_i)$ in each expression and expand. Hence establish that the average optimism in the training error is

$$\frac{2}{N}\sum_{i=1}^{N}\text{Cov}(\hat{y}_i, y_i),$$

as given in (7.21).

We start with $\mathrm{Err_{in}}$. Let's denote $\hat{y}_i = \hat{f}(x_i)$ and write

$$Y_i^0 - \hat{f}(x_i) = Y_i^0 - f(x_i) + f(x_i) - E\hat{y}_i + E\hat{y}_i - \hat{y}_i$$

so that

$$\begin{aligned}
\mathrm{Err_{in}} &= \frac{1}{N}\sum_{i=1}^{N} E_{Y^0}\big(Y_i^0 - f(x_i) + f(x_i) - E\hat{y}_i + E\hat{y}_i - \hat{y}_i\big)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N} A_i + B_i + C_i + D_i + E_i + F_i,
\end{aligned}$$

where

$$\begin{aligned}
A_i &= E_{Y^0}(Y_i^0 - f(x_i))^2 \\
B_i &= E_{Y^0}(f(x_i) - E\hat{y}_i)^2 = (f(x_i) - E\hat{y}_i)^2 \\
C_i &= E_{Y^0}(E\hat{y}_i - \hat{y}_i)^2 = (E\hat{y}_i - \hat{y}_i)^2 \\
D_i &= 2E_{Y^0}(Y_i^0 - f(x_i))(f(x_i) - E\hat{y}_i) \\
E_i &= 2E_{Y^0}(Y_i^0 - f(x_i))(E\hat{y}_i - \hat{y}_i) \\
F_i &= 2E_{Y^0}(f(x_i) - E\hat{y}_i)(E\hat{y}_i - \hat{y}_i) = 2(f(x_i) - E\hat{y}_i)(E\hat{y}_i - \hat{y}_i)
\end{aligned}$$

Similarly for $\overline{\mathrm{err}}$ we have

$$y_i - \hat{f}(x_i) = y_i - f(x_i) + f(x_i) - E\hat{y}_i + E\hat{y}_i - \hat{y}_i$$

and

$$\begin{aligned}
\overline{\mathrm{err}} &= \frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i) + f(x_i) - E\hat{y}_i + E\hat{y}_i - \hat{y}_i)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N} G_i + B_i + C_i + H_i + J_i + F_i,
\end{aligned}$$

where

$$\begin{aligned}
G_i &= (y_i - f(x_i))^2 \\
H_i &= 2(y_i - f(x_i))(f(x_i) - E\hat{y}_i) \\
J_i &= 2(y_i - f(x_i))(E\hat{y}_i - \hat{y}_i).
\end{aligned}$$

Therefore, we have

$$\begin{aligned}E_{\mathbf{y}}(\text{op}) &= E_{\mathbf{y}}(\text{Err}_{\text{in}} - \overline{\text{err}})\\ &= \frac{1}{N}\sum_{i=1}^{N} E_{\mathbf{y}}[(A_i - G_i) + (D_i - H_i) + (E_i - J_i)].\end{aligned}$$

For A_i and G_i, the expectaion over $\mathbf{y}$ captures unpredictable error and thus $E_{\mathbf{y}}(A_i - G_i) = 0$. Similarly we have $E_{\mathbf{y}}D_i = E_{\mathbf{y}}H_i = E_{\mathbf{y}}E_i = 0$, and thus

$$\begin{aligned}E_{\mathbf{y}}(\text{op}) &= -\frac{2}{N}\sum_{i=1}^{N} J_i\\ &= -\frac{2}{N}\sum_{i=1}^{N} E_{\mathbf{y}}(y_i - f(x_i))(E\hat{y}_i - \hat{y}_i)\\ &= \frac{2}{N}\sum_{i=1}^{N}[E_{\mathbf{y}}(y_i\hat{y}_i) - E_{\mathbf{y}}y_i E_{\mathbf{y}}\hat{y}_i]\\ &= 2\text{Cov}(y_i, \hat{y}_i).\end{aligned}$$

7.5

For a linear smoother $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$, show that

$$\sum_{i=1}^{N}\text{Cov}(\hat{y}_i, y_i) = \text{trace}(\mathbf{S})\sigma_\epsilon^2,$$

which justifies its use as the effective number of parameters.

$$\begin{aligned}\sum_{i=1}^{N}\text{Cov}(\hat{y}_i, y_i) &= \text{trace}(\text{Cov}(\hat{\mathbf{y}}, \mathbf{y}))\\ &= \text{trace}(\text{Cov}(\mathbf{S}\mathbf{y}, \mathbf{y}))\\ &= \text{trace}(\mathbf{S}\text{Cov}(\mathbf{y}, \mathbf{y}))\\ &= \text{trace}(\mathbf{S}\text{Var}(\mathbf{y}))\\ &= \text{trace}(\mathbf{S})\sigma_\epsilon^2.\end{aligned}$$