ESL Ex 7.1: Derive the estimate of in-sample error (7.24)

$$E_y(\mathrm{Err}_{in}) = E_y(\overline{err}) + 2\frac{d}{N}\sigma_\varepsilon^2$$

meaning that $\sum_{i=1}^{N} \mathrm{Cov}(\hat{y}_i, y_i) = d\,\sigma^2_\varepsilon$

Requirements: $Y = f(X) + \varepsilon$ additive model

$E(\varepsilon) = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2$

and in addition

"a linear fit with d inputs or basis functions"

We know a linear fit: $\underset{N\times 1}{Y} = \underset{N\times d}{X}\underset{d\times 1}{\beta} + \underset{N\times 1}{\varepsilon}$ with quadr. loss $\Rightarrow \hat{\beta} = (X^TX)^{-1}X^TY$

Then a full vector of predictions $\hat{Y} = X\hat{\beta} = \underbrace{X(X^TX)^{-1}X^T}_{H} Y$

our hat-matrix – this is a socalled "linear smoother"

trick: if we find $\mathrm{Cov}(\underset{N\times 1}{\hat{Y}}, \underset{N\times 1}{Y})$ this is an $N\times N$ matrix and

the trace of this matrix will give us $\sum_{i=1}^{N} \mathrm{Cov}(\hat{y}_i, y_i)$.

The def of $\mathrm{Cov}(\hat{Y}, Y) = E((\hat{Y} - E(\hat{Y}))(Y - E(Y))^T)$ and it can be shown (rather straightforward) ⊛ that $\mathrm{Cov}(HY, Y) = H\,\underbrace{\mathrm{Cov}(Y, Y)}_{\sigma^2_\varepsilon \cdot I}$

Thus

$$\sum_{i=1}^{N} \mathrm{Cov}(\hat{y}_i, y_i) = \mathrm{tr}(\mathrm{Cov}(\hat{Y}, Y)) = \mathrm{tr}(H \cdot \sigma^2_\varepsilon I)$$

$$= \sigma^2_\varepsilon\,\mathrm{tr}(H) = \sigma^2_\varepsilon\,\mathrm{tr}(X(X^TX)^{-1}X^T)$$

using rules for trace: $tr(AB) = tr(BA)$

$$= \sigma_\varepsilon^2 \, tr(\underbrace{X^T X (X^T X)^{-1}}_{I_{d \times d}}) = \underline{\underline{d \cdot \sigma_\varepsilon^2}}$$

[This is very similar to the solution for Ex 7.5 where $\hat{Y} = SY$
and also Ex 7.6 where k-NN also can be written as $\hat{Y} = SY$]

$$\circledast \quad Cov(HY, Y) = E\left((HY - E(HY))(Y - E(Y))^T \right) = E\left(HYY^T - HYE(Y)^T - HE(Y)Y^T + HE(Y)E(Y)^T H^T \right)$$

$$= H \; E\left((YY^T - YE(Y)^T - E(Y)Y^T + E(Y)E(Y)^T) \right)$$

$$= H \; Cov(Y, Y)$$