An Observation About Subsection 6.6: On Average Training MSE Underestimates Validation MSE

The validating mean squared error is often used to evaluate the performance of a predictor. The expected value of the validating mean squared error, taken over the distribution of training data y and validating data y_v is shown in equation (6.65) on page 285:

$$\mathrm{E}_{y,y_v}(\mathrm{MSE}_v) = \sigma^2 + \frac{1}{N}\sum_i \mathrm{Var}(\hat{y}_{v,i}) + \mathrm{bias}^2. \tag{6.65}$$

On the other hand, the expected value of the training mean squared error, taken over the distribution of training data y is shown in equation (6.66):

$$\mathrm{E}_y(\mathrm{MSE}_t) = \sigma^2 + \frac{1}{N}\sum_i \mathrm{Var}(\hat{y}_i) + \mathrm{bias}^2 - \frac{2}{N}\sum_i \mathrm{Cov}(y_i, \hat{y}_i), \tag{6.66}$$

indicating that the training mean squared error

$$\mathrm{MSE}_t = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

is a poor estimate of (6.65).

The predicted values in training and validating data are $\hat{y}_i = \hat{f}(x_i)$ and $\hat{y}_{v,i} = \hat{f}(x_{v,i})$, a function of known covariates x_i and $x_{v,i}$, respectively. When these covariates do not take the same values in training and validating data, the difference between (6.65) and (6.66) is a function of not only $\frac{2}{N}\sum_i \mathrm{Cov}(y_i, \hat{y}_i)$ but also of the second and third terms of (6.65) and (6.66). When predictions are evaluated at the same values of the covariates in the training and validating data, $E(y_i) = E(y_{v,i})$, $\hat{y}_{v,i} = \hat{y}_i$ and $\frac{1}{N}\sum_i \mathrm{Var}(\hat{y}_{v,i}) = \frac{1}{N}\sum_i \mathrm{Var}(\hat{y}_i)$. The last two equalities hold when $\hat{f}(x_i) = x_i'\hat{b}$. This is not made explicit in the book. Then it follows that, exactly,

$$\mathrm{E}_y(\mathrm{MSE}_t) = \mathrm{E}_{y,y_v}(\mathrm{MSE}_v) - \frac{2}{N}\sum_i \mathrm{Cov}(y_i, \hat{y}_i). \tag{6.67}$$

Evaluation of training and validating MSE at the same value of the covariates is less restrictive that may seem at first glance. When the objective is to obtain a measure of the validating MSE committing all the records as training data, when the choice of value of the covariates is arbitrary, it is reasonable to perform the calculations using the covariates avilable of the training data.

References