

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

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:

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

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

$$E_y(\text{MSE}_t) = \sigma^2 + \frac{1}{N} \sum_i \text{Var}(\hat{y}_i) + \text{bias}^2 - \frac{2}{N} \sum_i \text{Cov}(y_i, \hat{y}_i). \quad (6.66)$$

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. These covariates are not necessarily the same. With a large number of records the second terms on the right hand side of these expressions will approach a common value. Specifically, if the covariates in the training and validating data take the same values,  $\frac{1}{N} \sum_i \text{Var}(\hat{y}_{v,i}) = \frac{1}{N} \sum_i \text{Var}(\hat{y}_i)$ . This is not made explicit in the book. Then it follows that

$$E_y(\text{MSE}_t) = E_{y,y_v}(\text{MSE}_v) - \frac{2}{N} \sum_i \text{Cov}(y_i, \hat{y}_i). \quad (6.67)$$

## References