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

Dividing available data into training and validating sets allows computation of sample training and validating mean squared errors. This information can be used either to compare models according to their prediction performance, or to study the predictive performance of a given model.

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, conditional on known covariates, is shown in equation (6.65) on page 285:

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

where for the ith term in the sum, $\mathrm{bias}^2(i) = \left(\mathrm{E}(y_{v,i}) - \mathrm{E}(\hat{y}_{v,i})\right)^2$.

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

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

where for the ith term in the sum, $\mathrm{bias}^2(i) = \left(\mathrm{E}(y_i) - \mathrm{E}(\hat{y}_i)\right)^2$, 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 and the choice of value of

the covariates is arbitrary, it is reasonable to perform the calculations using the covariates available of the training data.

In the least squares linear regression setting, when predictions $\hat{y}_i = x_i'\hat{b}$ are evaluated at the same values of the covariates, $X_t = X_v = X$, the second terms in (6.65) and (6.66) take the form

$$\begin{aligned}\frac{1}{N}\sum_i \mathrm{Var}(\hat{y}_i) &= \frac{1}{N}\,\mathrm{tr}\Big[\mathrm{Var}\Big(X\hat{b}|y,\sigma^2\Big)\Big] \\ &= \frac{1}{N}\,\mathrm{tr}\Big[(X'X)^{-1}X'X\Big]\sigma^2 \\ &= \frac{p}{N}\sigma^2\end{aligned}$$

where p is the number of columns of the $(N \times p)$ full column rank matrix X. In this setting, with bias $= 0$ and $\sum_i \mathrm{Cov}(y_i, \hat{y}_i) = p\sigma^2$, the training mean squared error is equal to the maximum likelihood estimator of the residual variance in the normal linear regression model (see page 267). Its expectation takes the form

$$\begin{aligned}\mathrm{E}_y(\mathrm{MSE}_t) &= \sigma^2 + \frac{p}{N}\sigma^2 - \frac{2p}{N}\sigma^2 \\ &= \frac{N-p}{N}\sigma^2.\end{aligned}$$

Covariance functions can be used to estimate validating means squared errors without splitting the data into training and validating sets as in traditional cross-validation. However, in contrast with traditional cross-validation, the estimate of validating mean squared error using covariance functions is model dependent.

Limiting behaviour of validating mean squared error.

The distribution of MSE_v as N (the length of the vector of validating records) tends to infinity is studied as follows. Consider the validating mean squared error

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

where y_i is the ith validating record and $\hat{y}_i$ its predictor. Add and subtract Xb inside the square brackets and expand the square. This yields

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^{N}\Big(y_i - x_i'\widehat{b}\Big)^2 = \frac{1}{N}\Big[(y-Xb) - \Big(X\widehat{b} - Xb\Big)\Big]'\Big[(y-Xb) - \Big(X\widehat{b} - Xb\Big)\Big] \\ &= \frac{1}{N}(y-Xb)'(y-Xb) - \frac{2}{N}(y-Xb)'X\Big(\widehat{b} - b\Big) + \frac{1}{N}\Big(\widehat{b} - b\Big)'X'X\Big(\widehat{b} - b\Big).\end{aligned}$$

As $N \to \infty$ the first term tends to σ^2 and the second tends to zero in probability (the elements of $\left(\widehat{b} - b\right)$ converge to zero due to the consistency of the least squares estimator). Therefore,

$$\mathrm{MSE}_v \underset{N\to\infty}{\sim} \sigma^2 + \frac{1}{N}\left(\widehat{b} - b\right)' X'X \left(\widehat{b} - b\right). \tag{6.68}$$

Note that this second term can also be written as

$$\frac{2}{n}(y - Xb)'X\left(\widehat{b} - b\right) = \frac{2}{n}(y - E(y))'(\widehat{y} - E(\widehat{y})),$$

whose expectation over the joint distribution of training and validating data is $\frac{2}{n}\sum_{i=1}^{n} Cov(y_i, \widehat{y}_i)$, involving the covariance between observations and their predictions. Given the model, this covariance in the validating data is zero.

Define $z'z = \left(\widehat{b} - b\right)' A \left(\widehat{b} - b\right)$, with $A = X'X\frac{1}{\sigma^2}$. Since $A \operatorname{Var}\left(\widehat{b} - b\right) = I$, an idempotent matrix, $z'z \sim \chi^2 \operatorname{tr}(I)$. It follows that

$$\mathrm{MSE}_v \underset{N\to\infty}{\sim} \sigma^2 + \frac{\sigma^2}{N}\chi^2(p). \tag{6.69}$$

If $N \to \infty$ and the number of covariates increases proportionately as $p = \alpha N$, $\alpha < 1$, MSE_v converges in probability to $\sigma^2(1 + \alpha)$. The term in α accounts for uncertainty in b.

References