4.1 Model Assessment and Selection

Model Assessment and Selection

In this section, we discuss some of the most important concepts that arise in selecting a statistical model for a specific data set.

It is important to note that there are two separate goals that we might have in mind:

> (a) **Model selection:** is estimating the performance of different models to choose the best one.

> (b) **Model assessment:** having chosen a final model, estimating its test error on new data.

Measuring the Quality of Fit

> One of the most commonly used measures of quality of fit is the **mean squared error** (MSE), given by
>
> $$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2,$$
>
> where $\hat{f}(x_i)$ is the prediction that $\hat{f}$ gives for the ith observation.

The MSE defined above is the training MSE, which means that it is computed using the **training data** that was used to fit the model.

We are more interested in the accuracy of the predictions that we obtain when we apply our method to previously unseen **test data**.

More precisely, we are interested if $\hat{f}(x_0) \approx y_0$, where (x_0, y_0) is a previously unseen test observation not used to train the model. If we had a large number of test observations we would compute the average:

$$\mathrm{Ave}(\hat{f}(x_0) - y_0)^2.$$

Example: Polynomial fitting

Assume simulated data with $n = 20$ of the form $y_i = x_i + \epsilon_i$, where $\epsilon_i \sim N(0, 0.25^2)$.
The following model is "correct" for every $k \geq 1$, if $\beta_0 = \beta_2 = \ldots = \beta_k = 0$ and $\beta_1 = 1$.

Let us now increase the level of flexibility and fit a polynomial of order $5, 10$ and 15 to this simulated data. That is:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_k x_i^k + \epsilon_i.$$

We can see from the figure that with increasing order the polynomials fit the data more closely. However, we observe that the polynomials estimate the true f poorly because they are too wiggly.

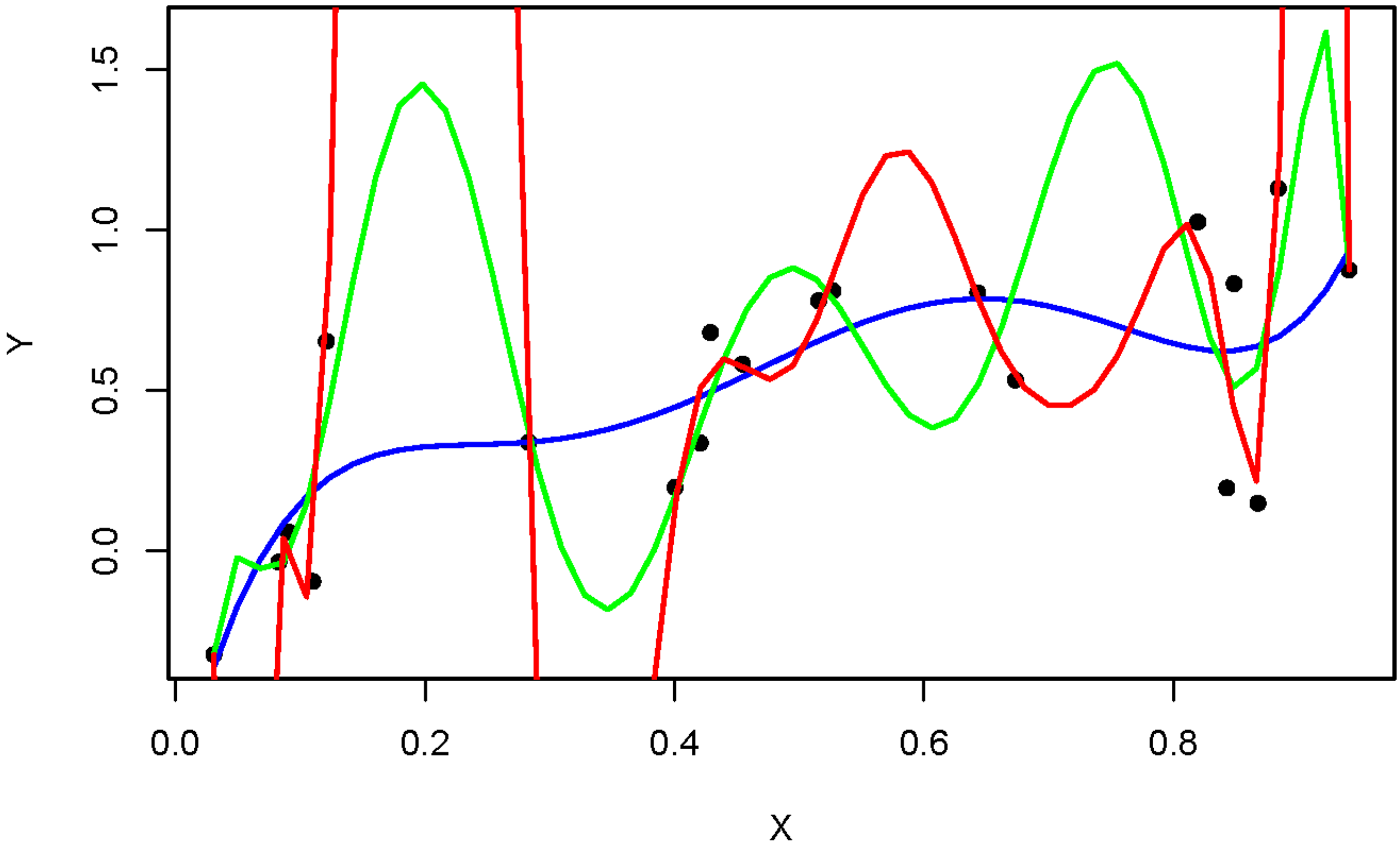


Figure 4.1.1: Fitted polynomial curves for polynomial of order $k = 5$ (blue), 10 (green) and 15 (red).

The figure below depicts the average training MSE (thick blue line) as a function of the model flexibility. The training MSE declines monotonically as flexibility increases.

Since we know the true f we can also compute the test MSE as a function of flexibility. The test MSE initially declines. However, at some point, the test MSE levels off and then starts to increase again.

When a given model yields a small training MSE but a large test MSE, we are said to be **overfitting the data**.

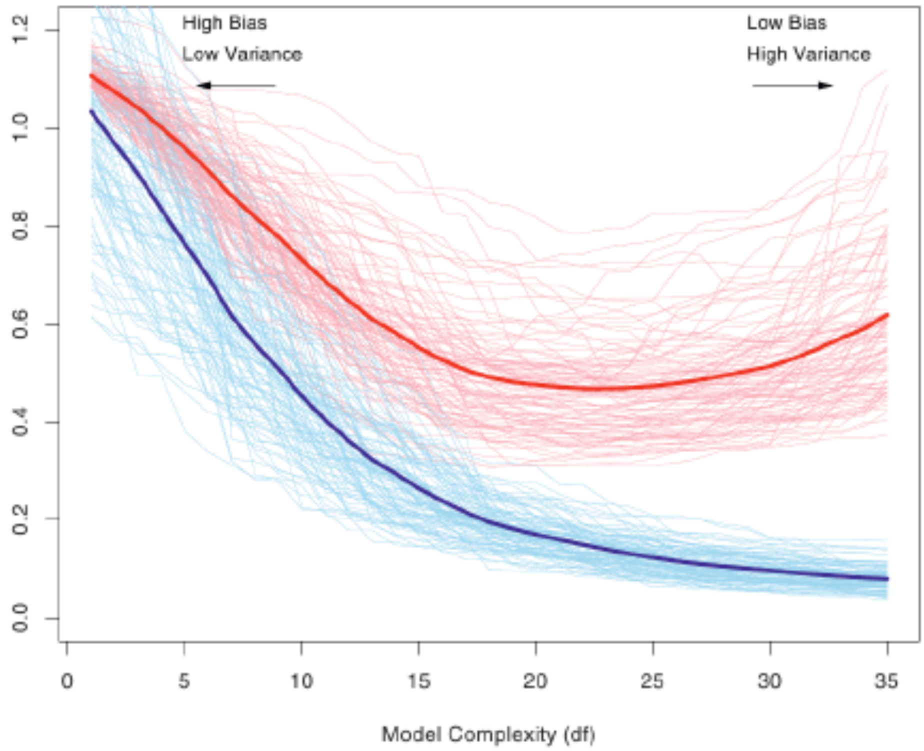


Figure 4.1.2: Training MSE (thin blue lines), test MSE (thin red lines) and their averages (thick lines).

The Bias-Variance Trade-Off

If we assume that $Y = f(X) + \varepsilon$ where $\boldsymbol{E}(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma_\varepsilon^2$, we can derive an expression for the expected test MSE, for a given value x_0 as follows

$$\begin{aligned}\boldsymbol{E}(y_0 - \hat{f}(x_0))^2 &= \sigma_\varepsilon^2 + (\boldsymbol{E}(\hat{f}(x_0)) - f(x_0))^2 + \boldsymbol{E}(\hat{f}(x_0) - \boldsymbol{E}(\hat{f}(x_0)))^2 \\ &= \sigma_\varepsilon^2 + \mathrm{Bias}^2(\hat{f}(x_0)) + \mathrm{Var}(\hat{f}(x_0)).\end{aligned}$$

This equation tells us that to minimise the expected test error we need to select a model that simultaneously achieves **low variance** and **low bias**.

Variance refers here to the amount by which $\hat{f}$ would change if we estimated it using a different training data set. If a method has high variance, then small changes in the training data can result in substantial changes in $\hat{f}$.

Refer again to our polynomial example. We can see that the polynomial of order 15 is following the observations very closely. It has high variance because changing any one of these data points may cause the estimate $\hat{f}$ to vary considerably. On the other hand, the linear fit has small variance, because changing any single observation will likely cause only a tiny shift in the position of the line.

Bias refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simplier model.

Statistical Decision Theory

Let $X \in \mathbb{R}^p$ be a real valued random input vector and $Y \in \mathbb{R}$ a real valued random output variable, with joint density function $p(x, y)$.

We look for a function $f(X)$ for predicting Y given values of the input X.

The theory requires a **loss function** $L(Y, f(X))$ for penalizing errors in prediction.

Examples of loss functions

> **Squared error loss:** $L(y_i, f(x_i)) = (y_i - f(x_i))^2$ is one of the most commonly used loss functions.

> **Deviance loss:** $L(y_i, f(x_i)) = -\ell(\beta; x_i, y_i)$ where it is implicit that the parameter β determines the regression function f. By definition of the deviance, this is only correct up to a constant. But constants are irrelevant when minimising the loss.

> **0-1 loss:** In a classification problem (such as logistic regression), the response y_i is categorical, and $f(x_i) \in \{1, \dots, K\}$. Then misclassification loss is $L(y_i, f(x_i)) = \mathbf{1}_{\{y_i \neq f(x_i)\}}$.

> **Exponential loss:** Supposes $f(x_i) \in \mathbb{R}$ and $y_i \in \{-1, +1\}$, and $L(y_i, f(x_i)) = \exp(-y_i f(x_i))$.

The aim is to find ("learn") the function f which minimises

$$\boldsymbol{E}_{(X,Y)}[L(Y, f(X))] = \iint L(y, f(x))p(x, y)dxdy.$$

The marginal $p(x) = \int p(x, y)dy$ of the law $p(x, y)$ then expresses the modelling focus: what regions of $\mathbb{R}^d$ are to be modelled well by f? For instance, if predictions on an interval $[-1, +1]$ are equally important, then a valid assumption is that $p(x)$ is the uniform distribution on $[-1, +1]$; if prediction around 0 are more important, one may decide for $p(x) = \mathcal{N}(x; 0, 1)$.

For example, for $L(Y, f(X)) = (Y - f(X))^2$ we need to minimize

$$E_{(X,Y)}(Y - f(X))^2 = \iint (y - f(x))^2 p(x, y)dxdy = \iint (y - f(x))^2 p(y|x)p(x)dydx$$

$$= E_X E_{Y|X}([Y - f(X)]^2 | X).$$

It suffices to minimize this criterion pointwise:

$$f(x) = \text{argmin}_c E_{Y|X}([Y - c]^2|X = x).$$

The solution is

$$f(x) = E(Y|X = x)$$

the conditional expectation, also known as the **regression function**.

Definitions of Errors

Assume now a training set $\mathcal{T} = (x_1, y_1), \ldots, (x_N, y_N)$ drawn from $p(x, y)$, and that a function $\hat{f}(x)$ has been fitted.

The **Training Error**

$$\overline{\text{err}}(\mathcal{T}) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{f}(x_i))$$

quantifies the discrepancy between fitted value $\hat{y}_i = \hat{f}(x_i)$ and the response y_i.

The **Generalisation or Test error**

$$\text{Err}_{\mathcal{T}} = E_{(X,Y)|\mathcal{T}}[L(Y, \hat{f}(X))]$$

$$\text{Err}_{\mathcal{T}} = E[L(Y, f(X))|\mathcal{T}] = \int L(y, \hat{f}(x))p(x, y)dxdy$$

is obtained by averaging the loss over all new data pairs drawn from $p(x, y)$, where the model fit based on $\mathcal{T}$ is left fixed. Small $\text{Err}_{\mathcal{T}}$ is desirable since then $\hat{f}$ generalises well and yields small loss for new data. Note that the generalisation error $\text{Err}_{\mathcal{T}}$ does depend on the training data. Here the test error refers to the error for this specific training set.

The **Expected error**

$$\text{Err} = E_{\mathcal{T}}[\text{Err}_{\mathcal{T}}]$$

$$\text{Err} = E[\text{Err}_{\mathcal{T}}] = \int \text{Err}_{\mathcal{T}} p(\mathcal{T})d\mathcal{T}$$

averages the generalisation error over all training sets $\mathcal{T}$ of size N, assuming they are drawn from $p(x, y)$. If the expected error is small, then the general approach taken in fitting $\hat{f}$ is right; the quality of an individual $\hat{f}$ however still depends on the training set. In applications, one is interested in $\text{Err}_{\mathcal{T}}$, which describes the quality of the presently fitted model. Estimates for $\text{Err}_{\mathcal{T}}$ can only be obtained by estimating Err.

Optimism of the Training Error Rate

The **in-sample error**

$$\text{Err}_{\text{in}} = \frac{1}{N} \sum_{i=1}^{N} E_{\mathbf{Y}}[L(Y_i, \hat{f}(x_i))|\hat{f}]$$

$$\mathrm{Err}_{\mathrm{in}}(\mathcal{T})=\frac{1}{N}\sum_{i=1}^{N}\int L(y,\hat{f}(x_i))p(y|x=x_i)dy$$

calculates the expected loss when the response is resampled at each location x_i, with $\hat{f}$ left unchanged. This is effectively the generalisation error, conditional on the new data (X,Y) arriving at the old training coordinates $x_1,\ldots,x_N$. The best guess for Err is typically given by $\mathrm{Err}_{\mathrm{in}}(\mathcal{T})$.

Note that $\hat{f}$ has been chosen to minimise the loss $\overline{\mathrm{err}}$ at the training data. Resampling the data, but keeping $\hat{f}$ constant, will hence typically *increase* the loss. The difference

$$\mathrm{op}(\mathcal{T})=\mathrm{Err}_{\mathrm{in}}(\mathcal{T})-\overline{\mathrm{err}}(\mathcal{T})$$

is called the **optimism** of the training error.

Observe that the optimism depends on the fitted regression function $\hat{f}$, which in turn depends on the responses y_i at x_i. Taking the average over the responses y_i (drawn from $p(y|x=x_i)$) and thus the average over the possible $\hat{f}$ defines the **average optimism** $E_{\mathbf{y}}[\mathrm{op}]$.

Where the locations x_i are fixed according to $\mathcal{T}$ and the averages are taken over resampled y_i.

For most loss functions including "squared" and "0-1", the average optimism is:

$$E_{\mathbf{y}}[\mathrm{op}]=\frac{2}{N}\sum_{i=1}^{N}\mathrm{Cov}(\hat{y}_i,y_i)=\frac{2}{N}\mathrm{tr}(\mathrm{Cov}(\boldsymbol{H}\boldsymbol{y},\boldsymbol{y}))=\frac{2\sigma^2}{N}\mathrm{tr}(\boldsymbol{H})$$

We hence estimate:

$$\mathrm{Err}(\mathcal{T})\approx\mathrm{Errin}(\mathcal{T})=\overline{\mathrm{err}}(\mathcal{T})+\mathrm{op}(\mathcal{T})\approx\overline{\mathrm{err}}(\mathcal{T})+E_{\mathbf{y}}[\mathrm{op}]=\overline{\mathrm{err}}(\mathcal{T})+\frac{2d\sigma^2}{N}$$

where $d=\mathrm{tr}(\boldsymbol{H})$ denotes the effective degrees of freedom.

Note that here we have

$$\hat{y}=\boldsymbol{H}y.$$

Then the effective number of parameters is defined as $\mathrm{tr}(\boldsymbol{H})$. The effective number of parameters is also known as the effective degrees of freedom.

The effective degrees of freedom of the fit can be defined in various ways to implement goodness-of-fit tests, cross-validation, and other statistical inference procedures. Here one can distinguish between regression effective degrees of freedom and residual effective degrees of freedom. The effective degrees of freedom above are the regression effective degrees of freedom. The residual effective degrees of freedom are defined by replacing $\boldsymbol{H}$ with $\boldsymbol{I}-\boldsymbol{H}$.

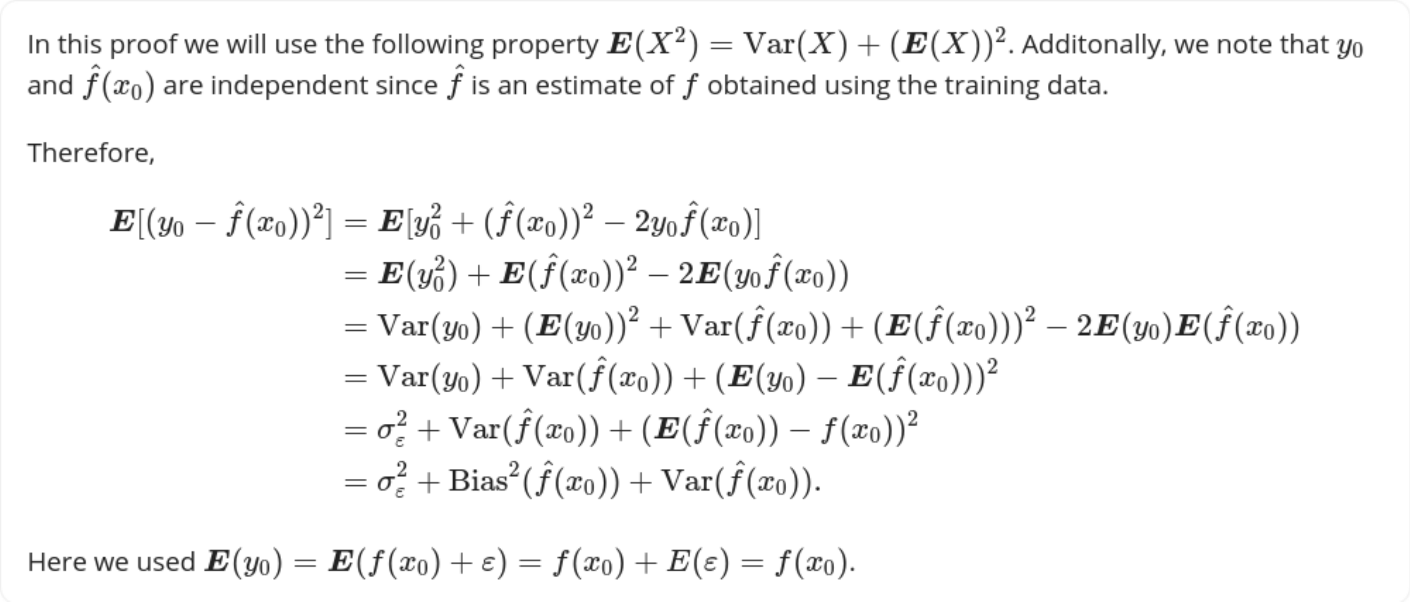
Activity: The Bias Variance Trade-Off

Question *Submitted Feb 7th 2024 at 3:46:30 pm*

If we assume that $Y = f(X) + \varepsilon$ where $\boldsymbol{E}(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma_\varepsilon^2$, show that the expected test MSE, for a given value x_0 can be decomposed as follows:

$$\boldsymbol{E}(y_0 - \hat{f}(x_0))^2 = \sigma_\varepsilon^2 + \mathrm{Bias}^2(\hat{f}(x_0)) + \mathrm{Var}(\hat{f}(x_0)).$$

Solution



In this proof we will use the following property $\boldsymbol{E}(X^2) = \mathrm{Var}(X) + (\boldsymbol{E}(X))^2$. Additonally, we note that y_0 and $\hat{f}(x_0)$ are independent since $\hat{f}$ is an estimate of f obtained using the training data.

Therefore,

$$\begin{aligned}
\boldsymbol{E}[(y_0 - \hat{f}(x_0))^2] &= \boldsymbol{E}[y_0^2 + (\hat{f}(x_0))^2 - 2y_0\hat{f}(x_0)] \\
&= \boldsymbol{E}(y_0^2) + \boldsymbol{E}(\hat{f}(x_0))^2 - 2\boldsymbol{E}(y_0\hat{f}(x_0)) \\
&= \mathrm{Var}(y_0) + (\boldsymbol{E}(y_0))^2 + \mathrm{Var}(\hat{f}(x_0)) + (\boldsymbol{E}(\hat{f}(x_0)))^2 - 2\boldsymbol{E}(y_0)\boldsymbol{E}(\hat{f}(x_0)) \\
&= \mathrm{Var}(y_0) + \mathrm{Var}(\hat{f}(x_0)) + (\boldsymbol{E}(y_0) - \boldsymbol{E}(\hat{f}(x_0)))^2 \\
&= \sigma_\varepsilon^2 + \mathrm{Var}(\hat{f}(x_0)) + (\boldsymbol{E}(\hat{f}(x_0)) - f(x_0))^2 \\
&= \sigma_\varepsilon^2 + \mathrm{Bias}^2(\hat{f}(x_0)) + \mathrm{Var}(\hat{f}(x_0)).
\end{aligned}$$

Here we used $\boldsymbol{E}(y_0) = \boldsymbol{E}(f(x_0) + \varepsilon) = f(x_0) + E(\varepsilon) = f(x_0)$.

Activity: Optimism of the Training Error Rate

Question *Submitted Feb 7th 2024 at 3:49:37 pm*

Note that the in sample error can be rewritten as

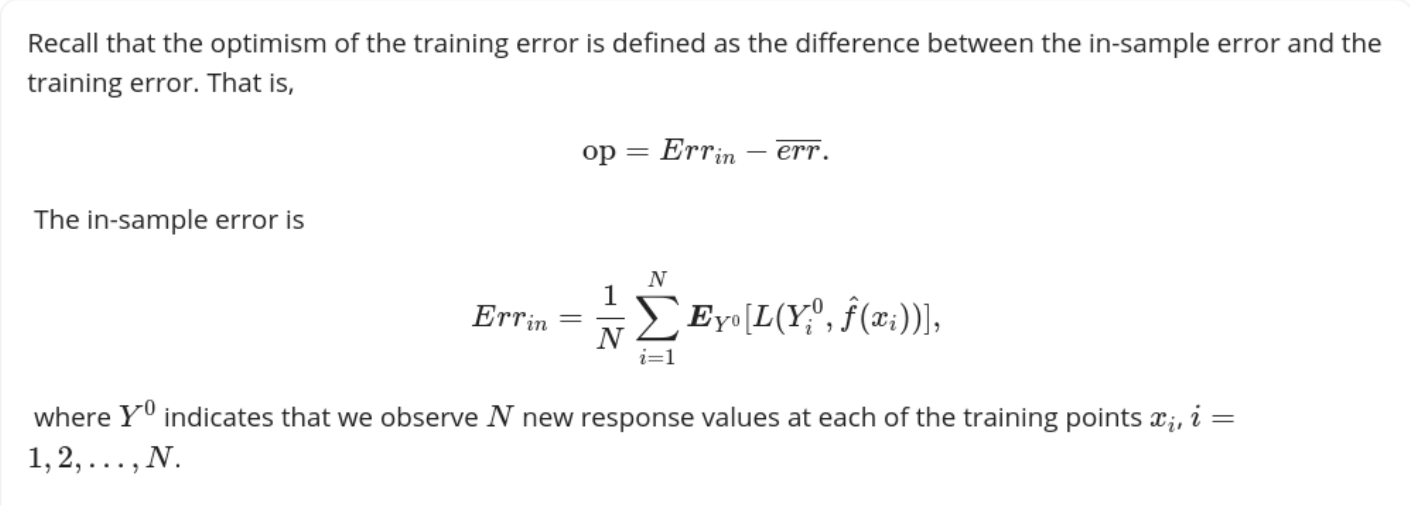
$$Err_{in} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{E}_{Y^0}[L(Y_i^0, \hat{f}(x_i))],$$

where Y^0 notation indicates that we observe N new response values at each of the training points x_i , $i = 1, 2, \ldots, N$. Show that the average optimism of the training error is

$$\boldsymbol{E}_y[op] = \frac{2}{N}\sum_{i=1}^{N} \operatorname{Cov}(\hat{y}_i, y_i)$$

for the squared error loss function.

Solution



Recall that the optimism of the training error is defined as the difference between the in-sample error and the training error. That is,

$$\text{op} = Err_{in} - \overline{err}.$$

The in-sample error is

$$Err_{in} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{E}_{Y^0}[L(Y_i^0, \hat{f}(x_i))],$$

where Y^0 indicates that we observe N new response values at each of the training points x_i, $i = 1, 2, \ldots, N$.

Therefore,

$$\begin{aligned}
\boldsymbol{E}y(\mathrm{op}) &= \boldsymbol{E}_y[Err_{in} - \overline{err}] = \boldsymbol{E}_y(Err_{in}) - \boldsymbol{E}_y(\overline{err}) \\
&= \boldsymbol{E}_y[\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{E}_{Y^0}[L(Y_i^0, \hat{f}(x_i))]] - \boldsymbol{E}_y[\frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}(x_i))] \\
&= \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{E}_y\boldsymbol{E}_{Y^0}[(Y_i^0 - \hat{y}_i)^2] - \boldsymbol{E}_y[(y_i - \hat{y}_i)^2]) \\
&= \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{E}_y\boldsymbol{E}_{Y^0}[(Y_i^0)^2 - 2Y_i^0\hat{y}_i + \hat{y}_i^2] - \boldsymbol{E}_y[y_i^2 - 2y_i\hat{y}_i + \hat{y}_i^2]) \\
&= \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{E}_y\boldsymbol{E}_{Y^0}[(Y_i^0)^2] - 2\boldsymbol{E}_y\boldsymbol{E}_{Y^0}[Y_i^0\hat{y}_i] + \boldsymbol{E}_y\boldsymbol{E}_{Y^0}[\hat{y}_i^2] \\
&- \boldsymbol{E}_y[y_i^2] + 2\boldsymbol{E}_y[y_i\hat{y}_i] - \boldsymbol{E}_y[\hat{y}_i^2]) \\
&= \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{E}_{Y^0}[(Y_i^0)^2] - 2\boldsymbol{E}_y[\hat{y}_i]\boldsymbol{E}_{Y^0}[Y_i^0] + \boldsymbol{E}_y[\hat{y}_i^2] - \boldsymbol{E}_y[y_i^2] + 2\boldsymbol{E}_y[y_i\hat{y}_i] - \boldsymbol{E}_y[\hat{y}_i^2]).
\end{aligned}$$

Since distributional properties of Y^0 and y are the same $\boldsymbol{E}_{Y^0}[(Y_i^0)^2] = \boldsymbol{E}_y(y_i^2)$ and

$$\begin{aligned}
\boldsymbol{E}_y(\mathrm{op}) &= \frac{1}{N}\sum_{i=1}^{N}(2\boldsymbol{E}_y[y_i\hat{y}_i] - 2\boldsymbol{E}_y[\hat{y}_i]\boldsymbol{E}_y[y_i]) \\
&= \frac{2}{N}\sum_{i=1}^{N}\mathrm{Cov}(y_i, \hat{y}_i).
\end{aligned}$$

Additional Activity

Question 1 *Submitted Feb 7th 2024 at 3:50:00 pm*

Model Assessment in the sense of this section is:

○ estimating the performance of different models in order to choose the best one;

○ having chosen a final model, estimating its training error on the dataset used for fitting the model;

● having chosen a final model, estimating its test error on new data.

Question 2 *Submitted Feb 7th 2024 at 3:50:39 pm*

Which of the following statements are true about the Bias-Variance Tradeoff:

☑ in order to minimize the expected test error we need to select a model that simultaniously achieves low bias and low variance;

☑ variance in the bias-variance trade-off sense refers to the amount by which $\hat{f}$ would change if we estimated it using a different training data set;

☑ bias refers to the error that is introduced by approximating a real-life problem by a much simplier model.