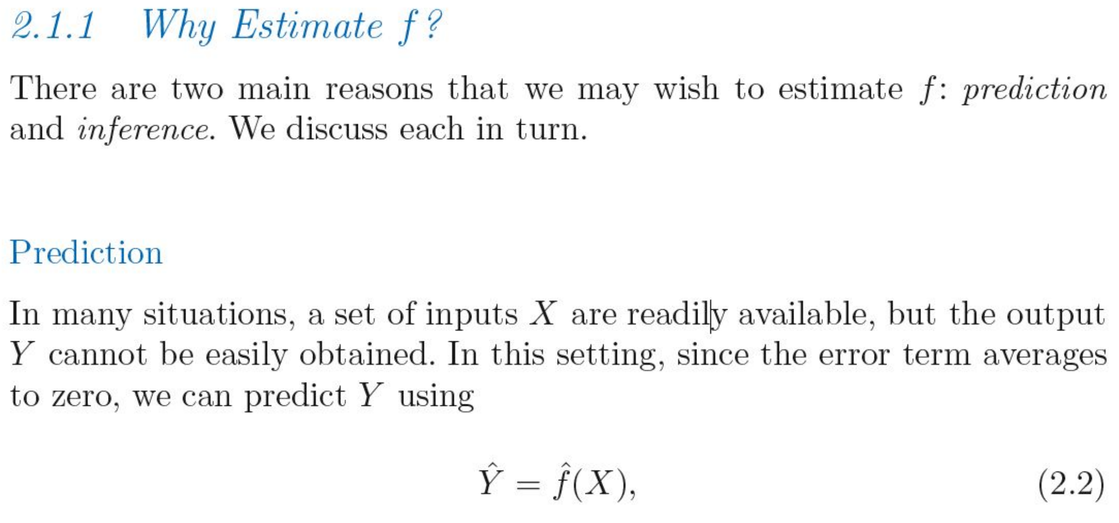
Chapter 2: What is Statistical Learning



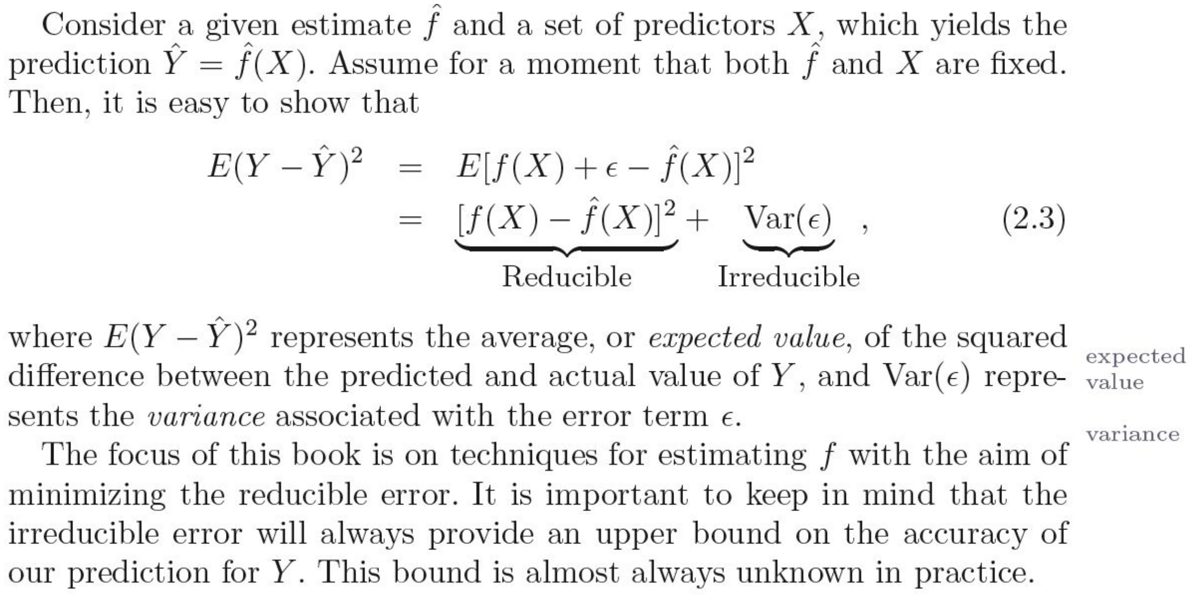
2.1.1 *Why Estimate f?*

There are two main reasons that we may wish to estimate f: *prediction* and *inference*. We discuss each in turn.

Prediction

In many situations, a set of inputs X are readily available, but the output Y cannot be easily obtained. In this setting, since the error term averages to zero, we can predict Y using

$$\hat{Y} = \hat{f}(X), \tag{2.2}$$



Consider a given estimate $\hat{f}$ and a set of predictors X, which yields the prediction $\hat{Y} = \hat{f}(X)$. Assume for a moment that both $\hat{f}$ and X are fixed. Then, it is easy to show that

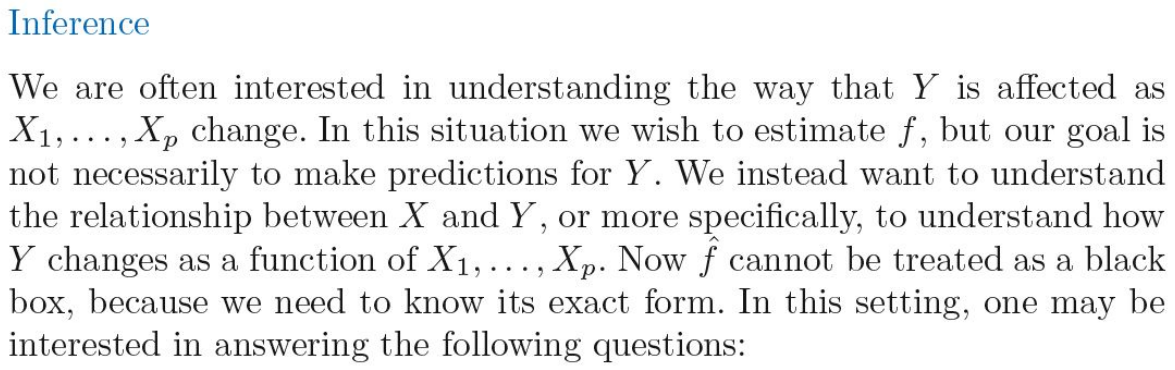
$$\begin{aligned} E(Y - \hat{Y})^2 &= E[f(X) + \epsilon - \hat{f}(X)]^2 \\ &= \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}} , \end{aligned} \tag{2.3}$$

where $E(Y - \hat{Y})^2$ represents the average, or *expected value*, of the squared difference between the predicted and actual value of Y, and $\text{Var}(\epsilon)$ represents the *variance* associated with the error term ϵ.

expected value

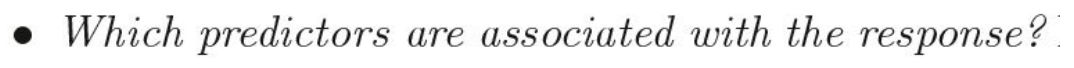
variance

The focus of this book is on techniques for estimating f with the aim of minimizing the reducible error. It is important to keep in mind that the irreducible error will always provide an upper bound on the accuracy of our prediction for Y. This bound is almost always unknown in practice.

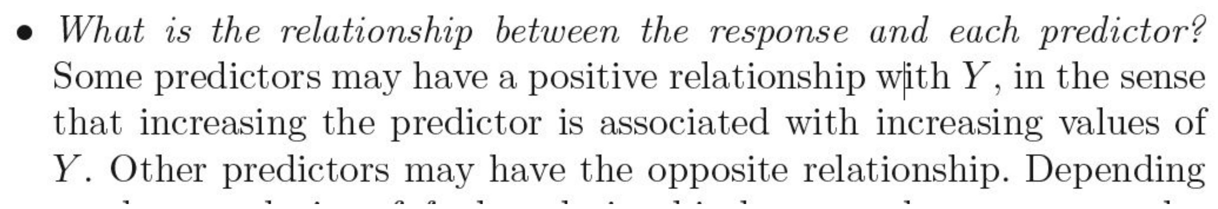


Inference

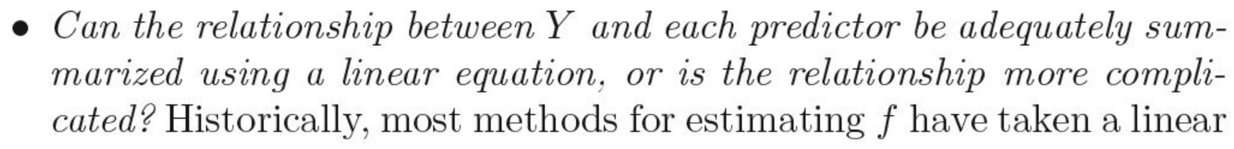
We are often interested in understanding the way that Y is affected as $X_1, \ldots, X_p$ change. In this situation we wish to estimate f, but our goal is not necessarily to make predictions for Y. We instead want to understand the relationship between X and Y, or more specifically, to understand how Y changes as a function of $X_1, \ldots, X_p$. Now $\hat{f}$ cannot be treated as a black box, because we need to know its exact form. In this setting, one may be interested in answering the following questions:



- *Which predictors are associated with the response?*



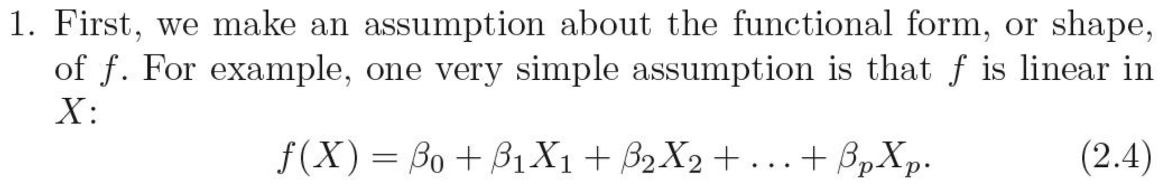
- *What is the relationship between the response and each predictor?* Some predictors may have a positive relationship with Y, in the sense that increasing the predictor is associated with increasing values of Y. Other predictors may have the opposite relationship. Depending



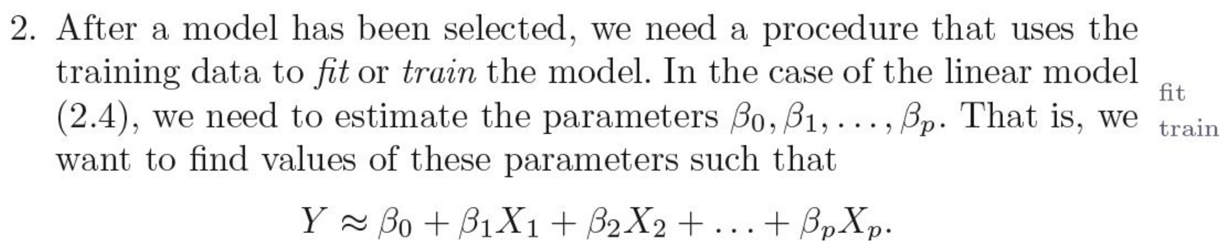
- *Can the relationship between* Y *and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?* Historically, most methods for estimating f have taken a linear

How to Estimate?

Parametric Methods: Have a functional form (Linear, Logistic Regression Example where f has some linear equation or polynomial regression where f is a polynomial)



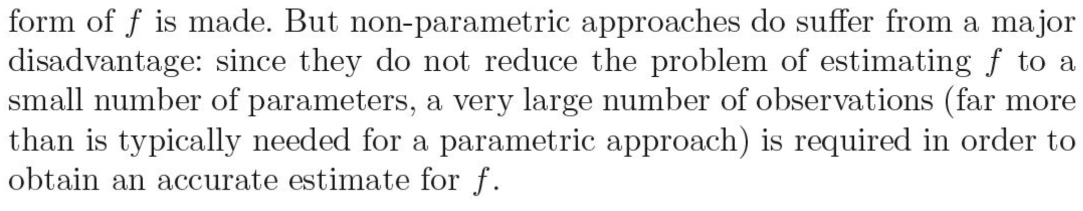
1. First, we make an assumption about the functional form, or shape, of f. For example, one very simple assumption is that f is linear in X:
$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p. \qquad (2.4)$$



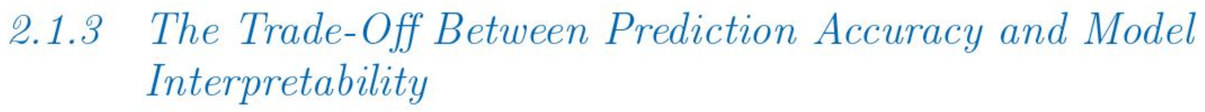
2. After a model has been selected, we need a procedure that uses the training data to *fit* or *train* the model. In the case of the linear model (2.4), we need to estimate the parameters $\beta_0, \beta_1, \ldots, \beta_p$. That is, we want to find values of these parameters such that
$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p.$$

fit
train

Non-Parametric Methods: Do not assume a particular functional form f. They get an estimate of f. A thin-plate spline is used to estimate f



form of f is made. But non-parametric approaches do suffer from a major disadvantage: since they do not reduce the problem of estimating f to a small number of parameters, a very large number of observations (far more than is typically needed for a parametric approach) is required in order to obtain an accurate estimate for f.



2.1.3 The Trade-Off Between Prediction Accuracy and Model Interpretability

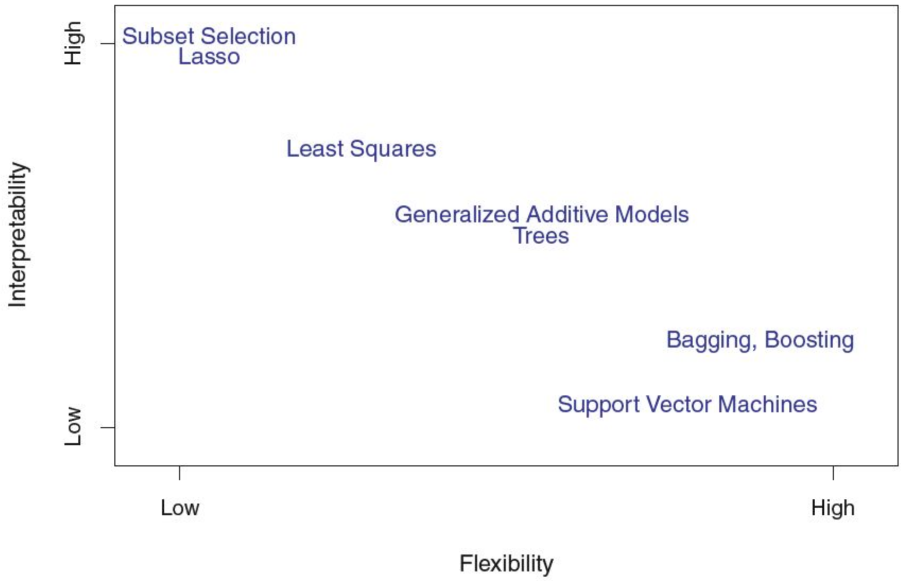


FIGURE 2.7. *A representation of the tradeoff between flexibility and interpretability, using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases.*

Some models are flexible (think spline plate), some are restrictive (Linear regression, as it HAS to be linear in coefficients). Then why would one want to use restrictive models?
When inference is the goal, restrictive is interpretable.

In contrast, very flexible approaches, such as the splines discussed in Chapter 7 and displayed in Figures 2.5 and 2.6, and the boosting methods discussed in Chapter 8, can lead to such complicated estimates of f that it is difficult to understand how any individual predictor is associated with the response.

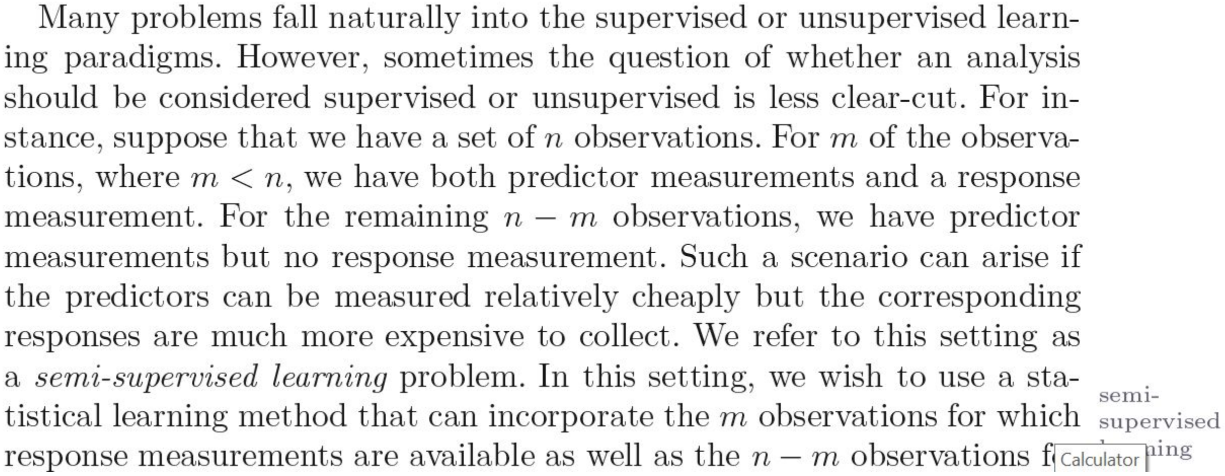
More on this in page 26-27 of the book.

ods. In some settings, however, we are only interested in prediction, and the interpretability of the predictive model is simply not of interest. For instance, if we seek to develop an algorithm to predict the price of a stock, our sole requirement for the algorithm is that it predict accurately—interpretability is not a concern. In this setting, we might expect that it will be best to use the most flexible model available. Surprisingly, this is not always the case! We will often obtain more accurate predictions using a less flexible method. This phenomenon, which may seem counterintuitive

(It is because they might overfit)

2.1.4 Supervised Versus Unsupervised Learning

We know it. A new info below



Many problems fall naturally into the supervised or unsupervised learning paradigms. However, sometimes the question of whether an analysis should be considered supervised or unsupervised is less clear-cut. For instance, suppose that we have a set of n observations. For m of the observations, where $m < n$, we have both predictor measurements and a response measurement. For the remaining $n - m$ observations, we have predictor measurements but no response measurement. Such a scenario can arise if the predictors can be measured relatively cheaply but the corresponding responses are much more expensive to collect. We refer to this setting as a *semi-supervised learning* problem. In this setting, we wish to use a statistical learning method that can incorporate the m observations for which response measurements are available as well as the $n - m$ observations f

semi-supervised ning

Beyond the scope of the book (semi-supervised)

2.1.5 Regression Versus Classification Problems

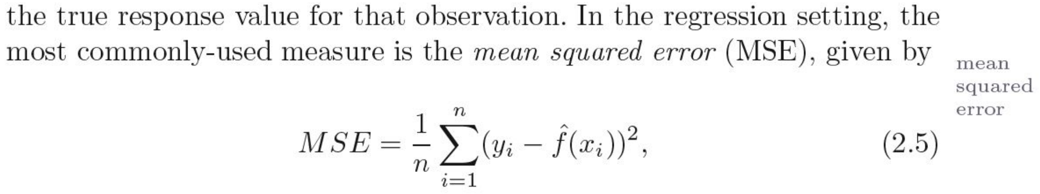
Variables can be numerical (quantitative) or categorical (qualitative). We use regression for former and classification for latter type of response. Distinction is not always crisp. Logistic Regression is used in classification. However, it has probability measures which are continuous numerical values, so it is regression as well. Whether the predictors are numerical or categorical is less important, provided they are “appropriately” coded before fitting.

2.2 Assessing Model Accuracy

Different models are suitable for different types of datasets. No single best universal model. Selecting best one is challenging task of statistical learning.

Regression Setting:

2.2.1 Measuring the Quality of Fit

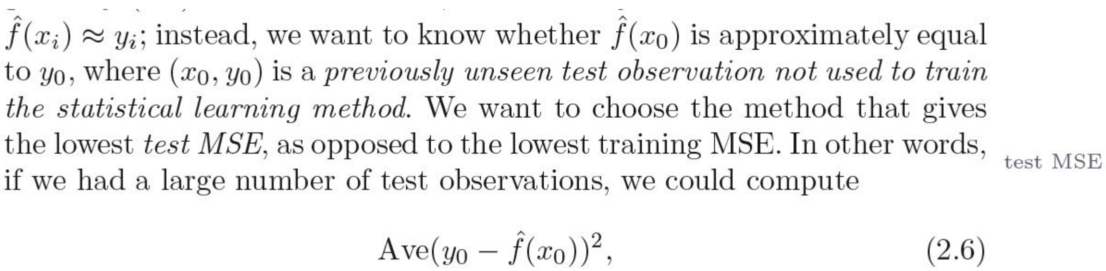


the true response value for that observation. In the regression setting, the most commonly-used measure is the *mean squared error* (MSE), given by

mean squared error

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2, \qquad (2.5)$$

We could measure this for training data {xi,yi}. But we are more interested in future, hence we measure it on test data.



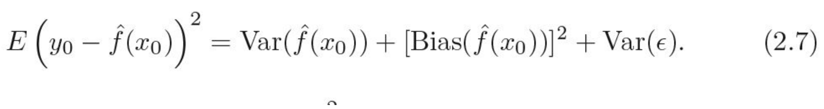
$\hat{f}(x_i) \approx y_i$; instead, we want to know whether $\hat{f}(x_0)$ is approximately equal to y_0, where (x_0, y_0) is a *previously unseen test observation not used to train the statistical learning method*. We want to choose the method that gives the lowest *test MSE*, as opposed to the lowest training MSE. In other words, if we had a large number of test observations, we could compute

test MSE

$$\mathrm{Ave}(y_0 - \hat{f}(x_0))^2, \qquad (2.6)$$

Lot of story that I know and then the section suggests cross validation

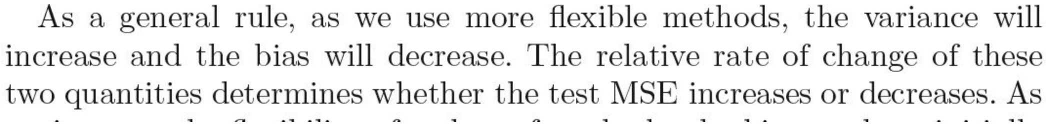
2.2.2 The Bias-Variance Trade-Off



$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon). \qquad (2.7)$$

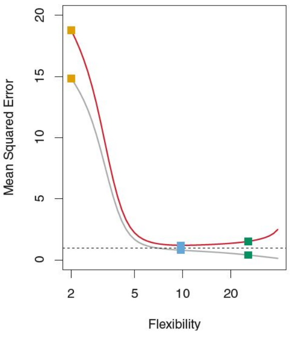
E(y0-fhat(xo))^2 represents expected test MSE for test point x0, if we train a model repeatedly using many sets of training samples and use those models to estimate x0.
We need to choose a statistical learning method that would lower variance and bias both. Both are non-negative, so the expected error sq can never be less than the third term

Variance in fhat means, extent to which estimate f would change if we use a different training set. Ideally it shouldn't. If a learning method has high variance, then small changes in training data would change estimate by large extent (Imagine, the curve is so wiggly that changing any one point will change the curve). More flexible statistical methods have high variance.
Bias refers to modeling a complicated phenomenon in a simple way. For example: Linear Regression assumes that there's linear relationship between points, but it is unlikely. Less flexible models have high bias. If the relationship between points in non-linear, no matter how many points we add (increasing number of training samples), the bias won't be reduced. Recall that variance reduces if we add more training samples.



As a general rule, as we use more flexible methods, the variance will increase and the bias will decrease. The relative rate of change of these two quantities determines whether the test MSE increases or decreases. As

Initially bias will reduce very fast, but after sometime, no reduction in bias, but variance increases. This is the point where we should stop. See figure below and the blue square is optimum



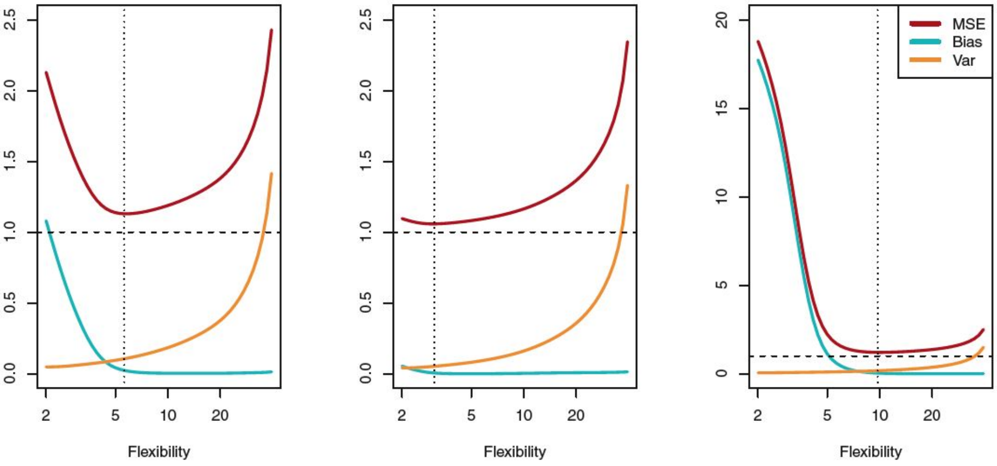
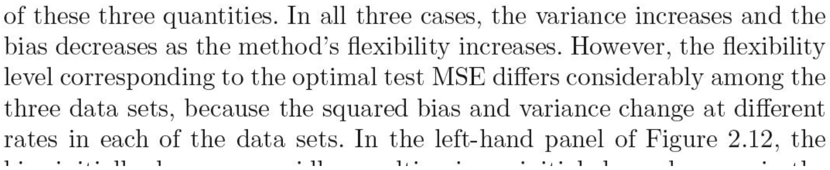


FIGURE 2.12. *Squared bias (blue curve), variance (orange curve), Var(ε) (dashed line), and test MSE (red curve) for the three data sets in Figures 2.9–2.11. The vertical dotted line indicates the flexibility level corresponding to the smallest test MSE.*

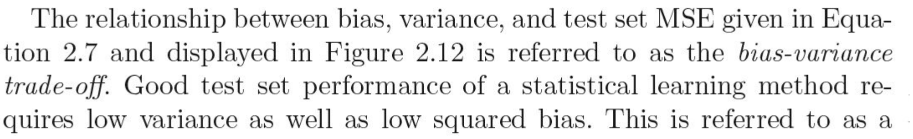


of these three quantities. In all three cases, the variance increases and the bias decreases as the method's flexibility increases. However, the flexibility level corresponding to the optimal test MSE differs considerably among the three data sets, because the squared bias and variance change at different rates in each of the data sets. In the left-hand panel of Figure 2.12, the

Important point: In the rightmost, there's a dramatic decline in bias, only because the true f is highly non-linear. So, looking at this curve (experimenting), one can figure out which ML method is suitable!

The center curve bias decreases very slowly, so it the dataset must be linear.

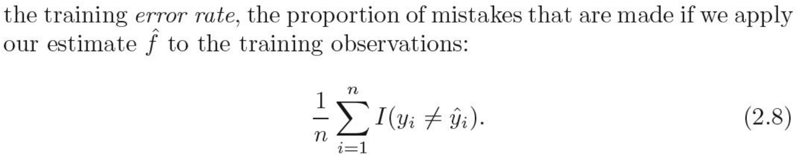
The left-most one is kind of in between



The relationship between bias, variance, and test set MSE given in Equation 2.7 and displayed in Figure 2.12 is referred to as the *bias-variance trade-off*. Good test set performance of a statistical learning method requires low variance as well as low squared bias. This is referred to as a

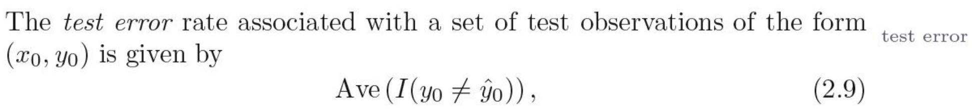
The term trade-off because it is possible to find methods with either high bias^2 (low variance) or high variance (low bias^2). Ideal (suited) method should have both low. Then test MSE is lowest.

2.2.3 The Classification Setting



the training *error rate*, the proportion of mistakes that are made if we apply our estimate $\hat{f}$ to the training observations:

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i). \tag{2.8}$$



The *test error* rate associated with a set of test observations of the form (x_0, y_0) is given by

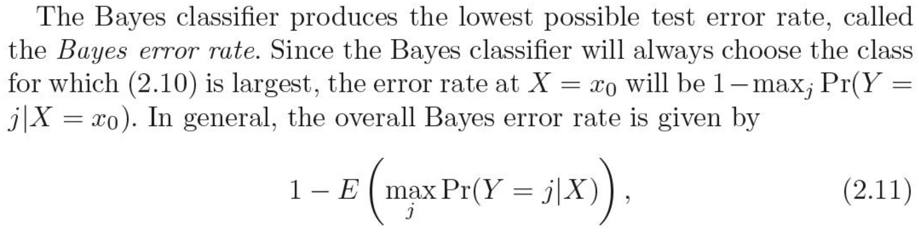
test error

$$\mathrm{Ave}\left(I(y_0 \neq \hat{y}_0)\right), \qquad (2.9)$$

The Bayes Classifier:

Select the class for which P(Y=i|X) is maximum among all i’s. In case of 2 classes, this means choose i if P(Y=i|X) > 0.5

Bayes Error rate: Analogous to irreducible error in regression

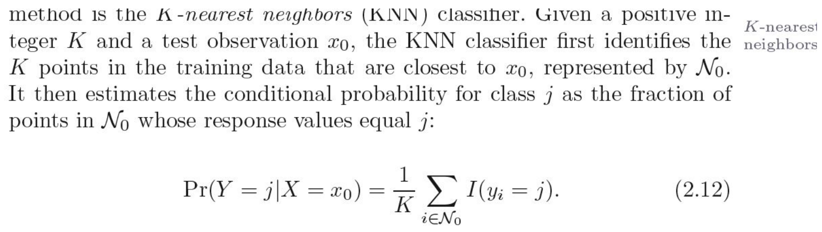


The Bayes classifier produces the lowest possible test error rate, called the *Bayes error rate*. Since the Bayes classifier will always choose the class for which (2.10) is largest, the error rate at $X = x_0$ will be $1 - \max_j \Pr(Y = j|X = x_0)$. In general, the overall Bayes error rate is given by

$$1 - E\left(\max_j \Pr(Y = j|X)\right), \qquad (2.11)$$

The k Nearest Neighbors (kNN) Classifier:

In theory, Bayes Classifier is preferred, but we don’t know always the probability distributions or conditional probability P(Y|X). In that case kNN is very useful to estimate the probability and make a classification decision.



method is the *K-nearest neighbors* (KNN) classifier. Given a positive integer K and a test observation x_0, the KNN classifier first identifies the K points in the training data that are closest to x_0, represented by $\mathcal{N}_0$. It then estimates the conditional probability for class j as the fraction of points in $\mathcal{N}_0$ whose response values equal j:

K-nearest neighbors

$$\Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j). \qquad (2.12)$$

Despite simplicity, kNN can produce surprisingly good results. Something to try out for small problems. But for big datasets, it takes very long time.