[**http://www.deeplearningbook.org/**](http://www.deeplearningbook.org/)

**Elements of Statistical Learning:**

**The outputs vary in nature among the examples:**

**Qualitative variables:** are also referred to as categorical or discrete variables as well as

Factors. typically represented numerically by codes, two classes or categories, such as “success” or “failure,” “survived” or “died.” These are often represented by a

single binary digit or bit as 0 or 1, or else by − 1 and 1. When there are more than two categories, several alternatives are available. The most useful and commonly used coding is via dummy variables .

**Quantitative Variables:** Continuous variables.

For both types of outputs it makes sense to think of using the inputs to predict the output.

This distinction in output type has led to a naming convention for the prediction tasks: **regression** when we predict quantitative outputs, and **classification** when we predict qualitative outputs.

Inputs also vary in measurement type; we can have some of each of qualitative

and quantitative input variables. These have also led to distinctions in the types of methods that are used for prediction: some methods are defined most naturally for quantitative inputs, some most naturally for qualitative and some for both.

A third variable type is **ordered categorical**, such as small, medium and large , where there is an ordering between the values, but no metric notion is appropriate (the difference between medium and small need not be the same as that between large and medium).

**Notations :**

We will typically denote an input variable by the symbol X . If X is

a vector, its components can be accessed by subscripts Xj . Quantitative

outputs will be denoted by Y , and qualitative outputs by G (for group).

We use uppercase letters such as X , Y or G when referring to the generic

aspects of a variable. Observed values are written in lowercase; hence the

i th observed value of X is written as xi (where xi is again a scalar or

vector). Matrices are represented by bold uppercase letters; for example, a

set of N input p -vectors xi, i = 1, . . . ,N would be represented by the NÅ~p

matrix X . In general, vectors will not be bold, except when they have N

components; this convention distinguishes a p -vector of inputs xi for the

2.3 Least Squares and Nearest Neighbors 11

i th observation from the N -vector xj consisting of all the observations on

variable Xj . Since all vectors are assumed to be column vectors, the i th

row of X is xTi

, the vector transpose of xi .

For the moment we can loosely state the learning task as follows: given

the value of an input vector X , make a good prediction of the output Y,

denoted by ˆ Y (pronounced “y-hat”). If Y takes values in IR then so should

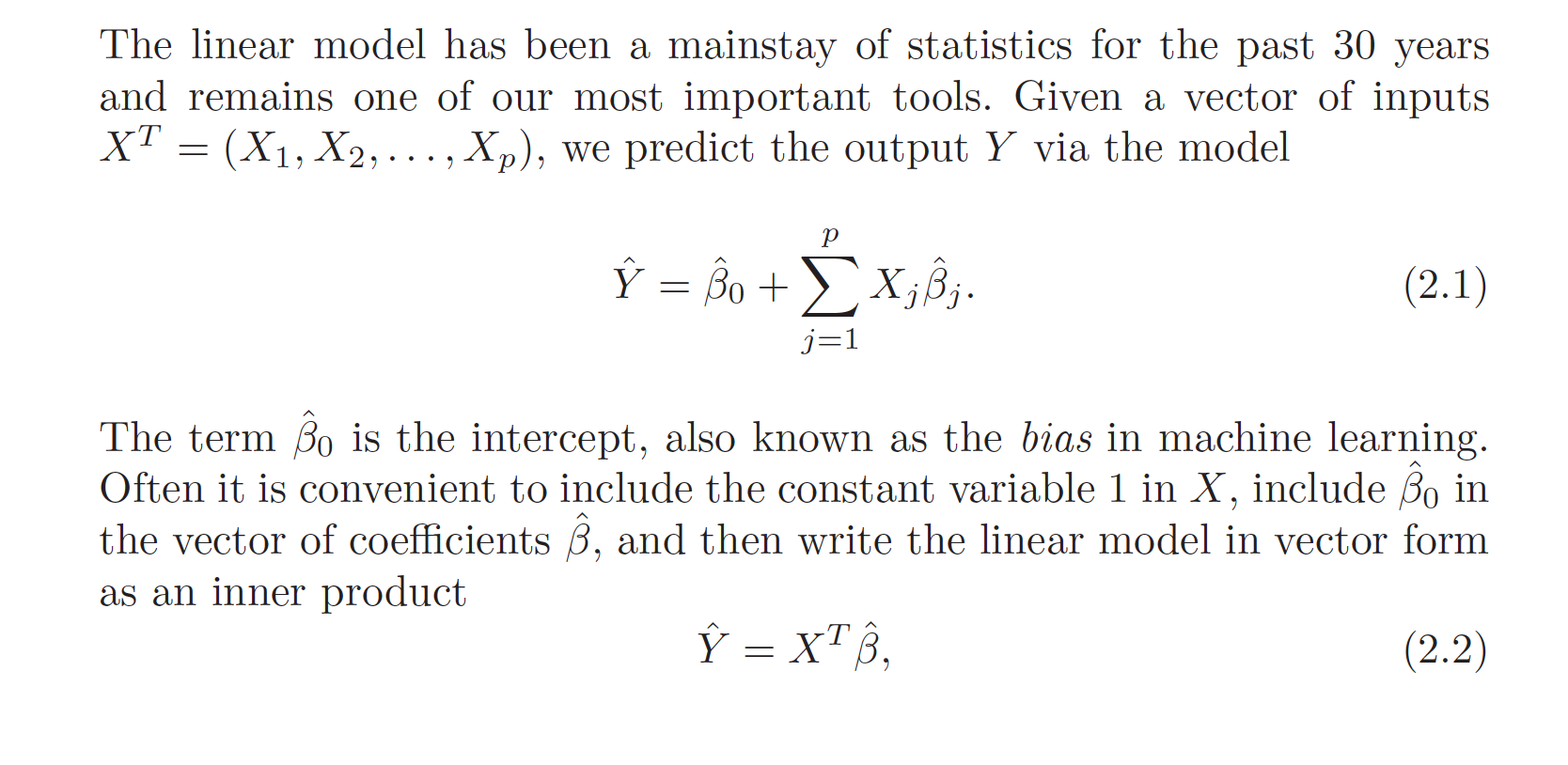
ˆ Y ; likewise for categorical outputs, ˆG should take values in the same set G associated with G .

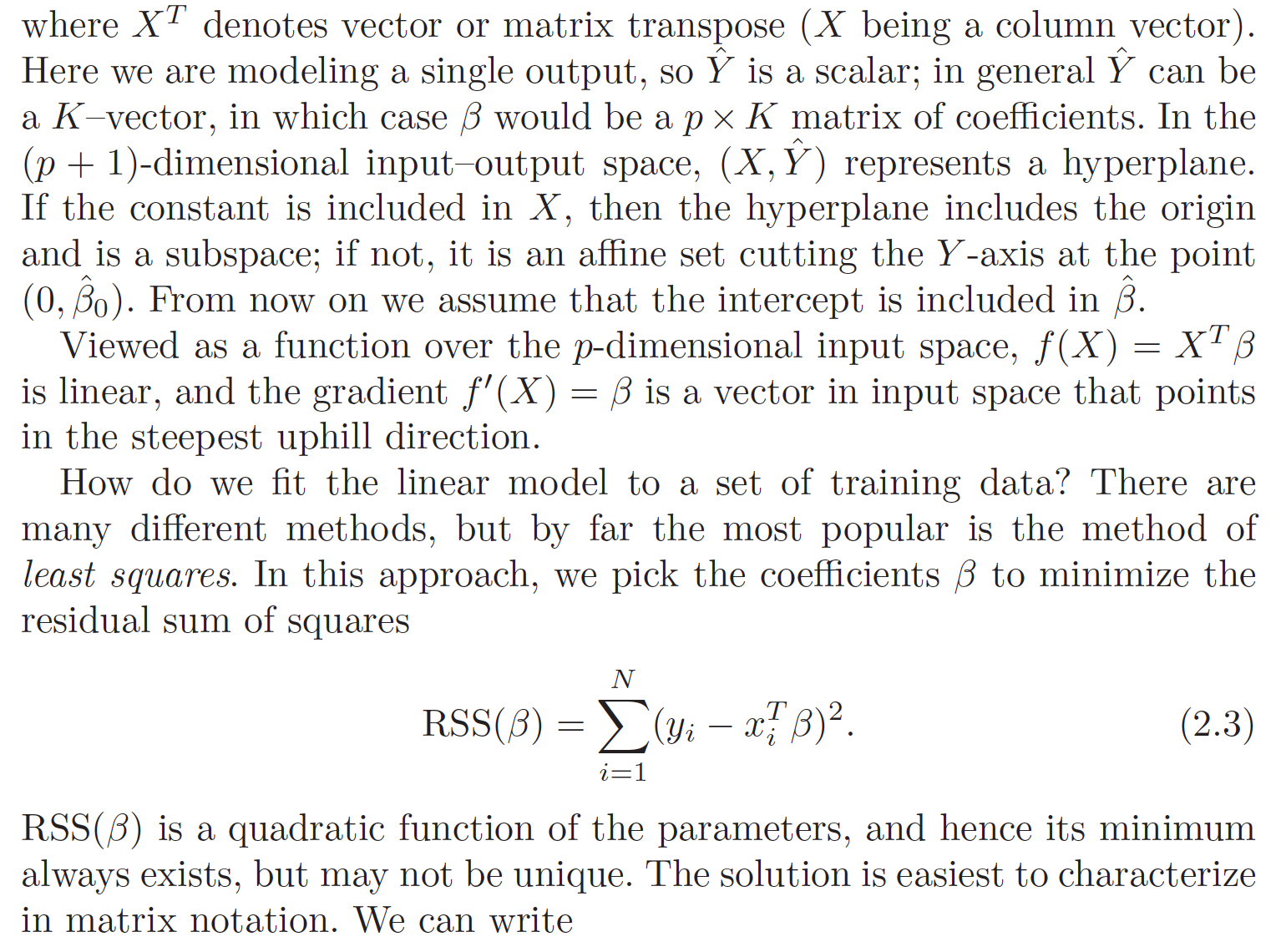
**Classification to Regression :**

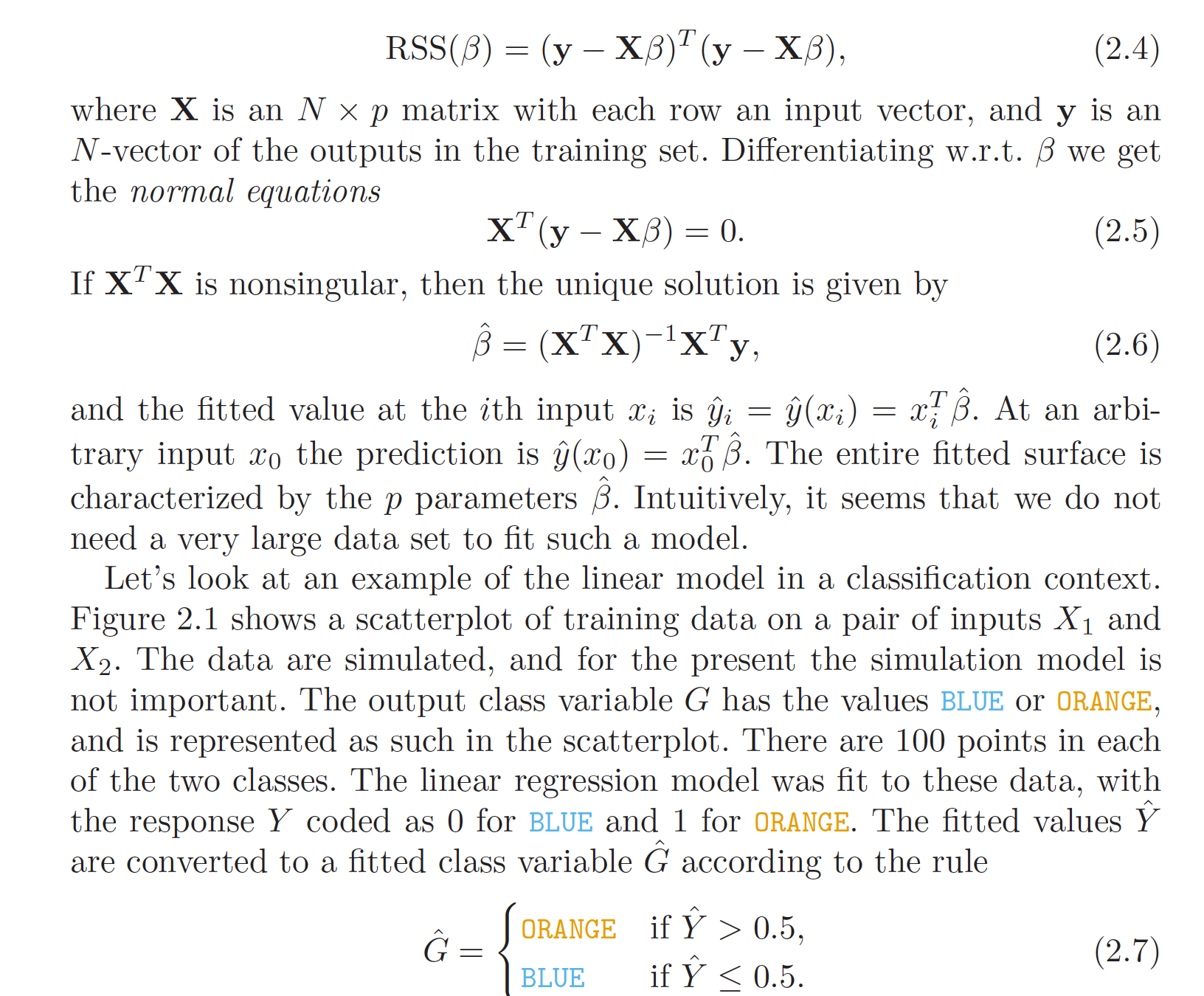
For a two-class G , one approach is to denote the binary coded target as Y , and then treat it as a quantitative output. The predictions ˆ Y will typically lie in [0, 1], and we can assign to ˆG the class label according to whether ˆy > 0. 5. This approach generalizes to K -level qualitative outputs as well.

**Two Simple Approaches to Prediction:**

* **Least Squares And Linear Models**









**Scenario 1:** The training data in each class were generated from bivariate Gaussian distributions with uncorrelated components and different means.

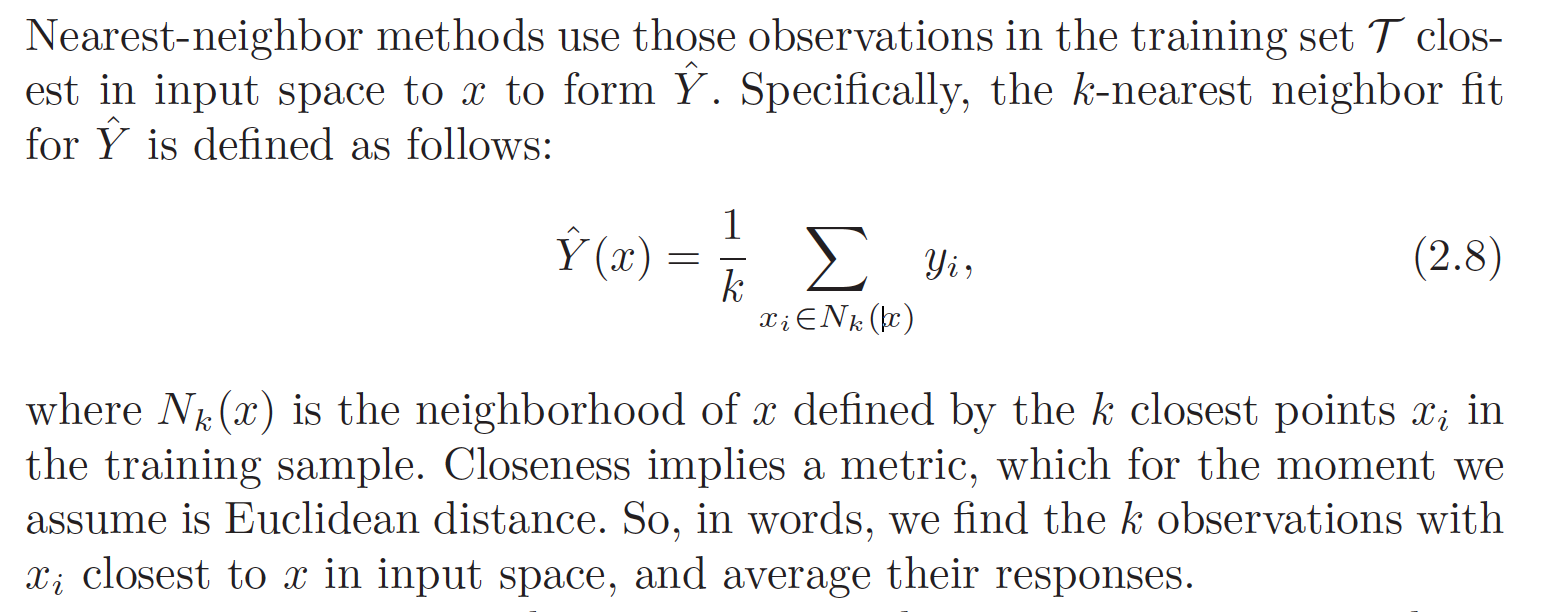
-- a linear decision boundary is the best one can do, and that our estimate is almost optimal. The region of overlap is inevitable, and future data to be predicted will be plagued by this overlap as well.

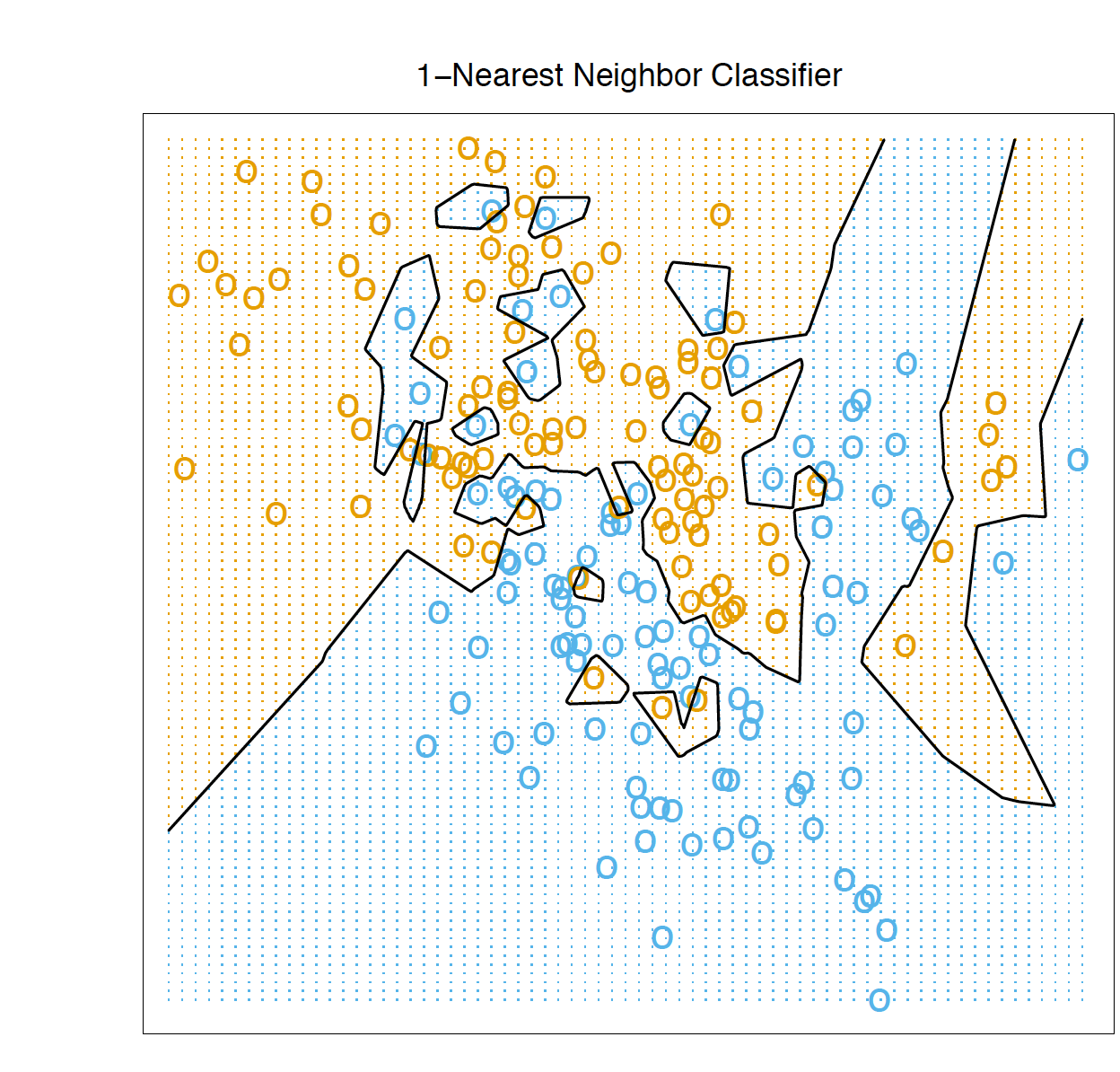
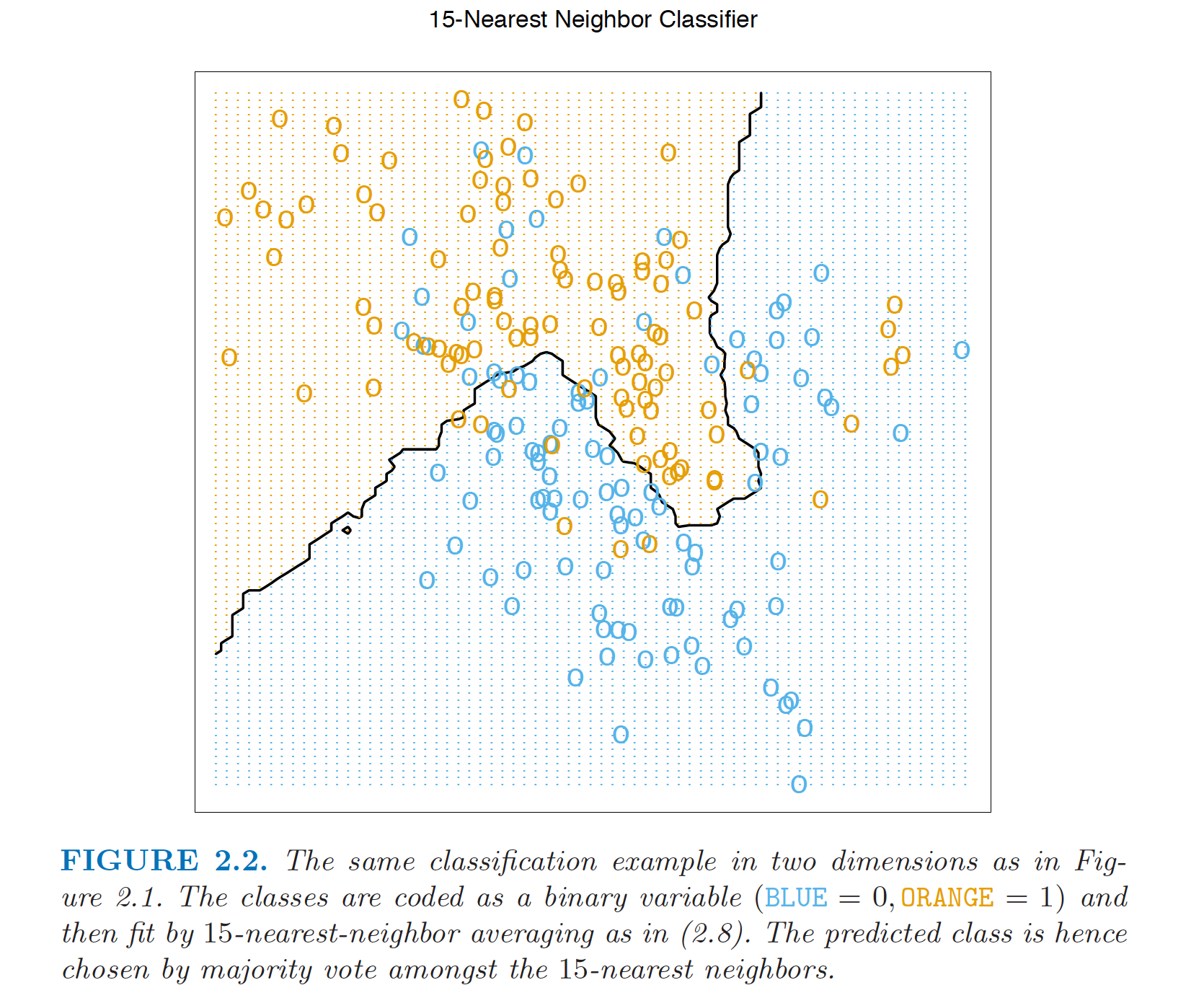
**Scenario 2:** The training data in each class came from a mixture of 10 low variance Gaussian distributions, with individual means themselves distributed as Gaussian.

-- In the case of mixtures of tightly clustered Gaussians the story is different. A linear decision boundary is unlikely to be optimal, and in fact is not. The optimal decision boundary is nonlinear and disjoint, and as such will be much more difficult to obtain.

**We now look at another classification and regression procedure that is in some sense at the opposite end of the spectrum to the linear model, and far better suited to the second scenario.**

* **Nearest Neighbors:**





Least Squares v/s Nearest Neighbors

The linear decision boundary from least squares is very smooth, and apparently

stable to fit. It does appear to rely heavily on the assumption that a linear decision boundary is appropriate. In language we will develop shortly, it has **low variance and potentially high bias.**

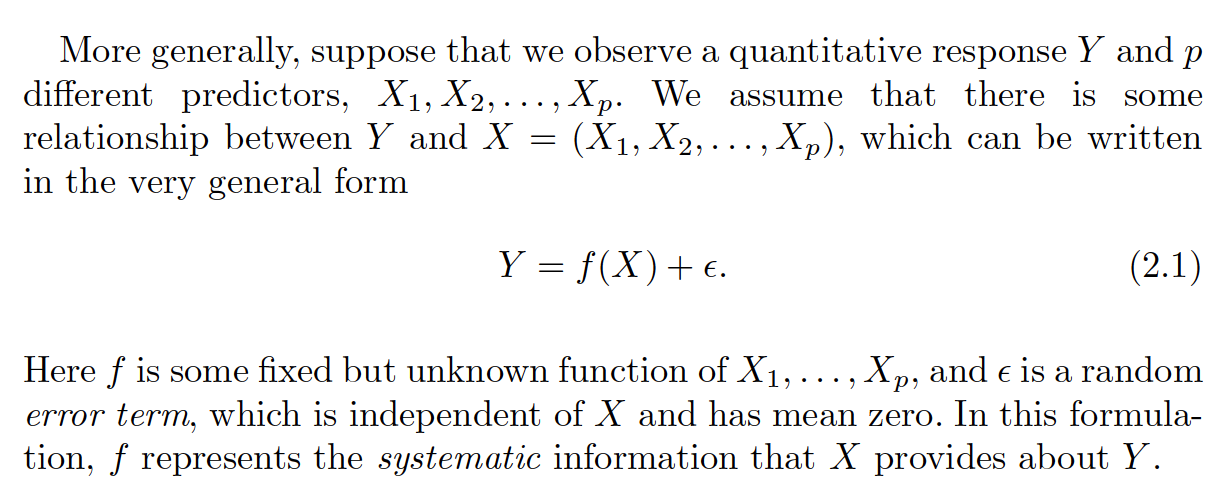
On the other hand, the k -nearest-neighbor procedures do not appear to rely on any stringent assumptions about the underlying data, and can adapt to any situation. However, any particular subregion of the decision boundary depends on a handful of input points and their particular positions, and is thus wiggly and unstable, **high variance and low bias.**

\*\****Each method has its own situations for which it works best.***

**Introduction to statistical Learning:**

**Statistical learning refers to a vast set of tools for understanding data.**

The inputs go by different names, such as *predictors , independent variables , features or sometimes just variables* . The output variable—in this case, sales —is variable often called the response or dependent variable , and is typically denoted using the symbol Y.

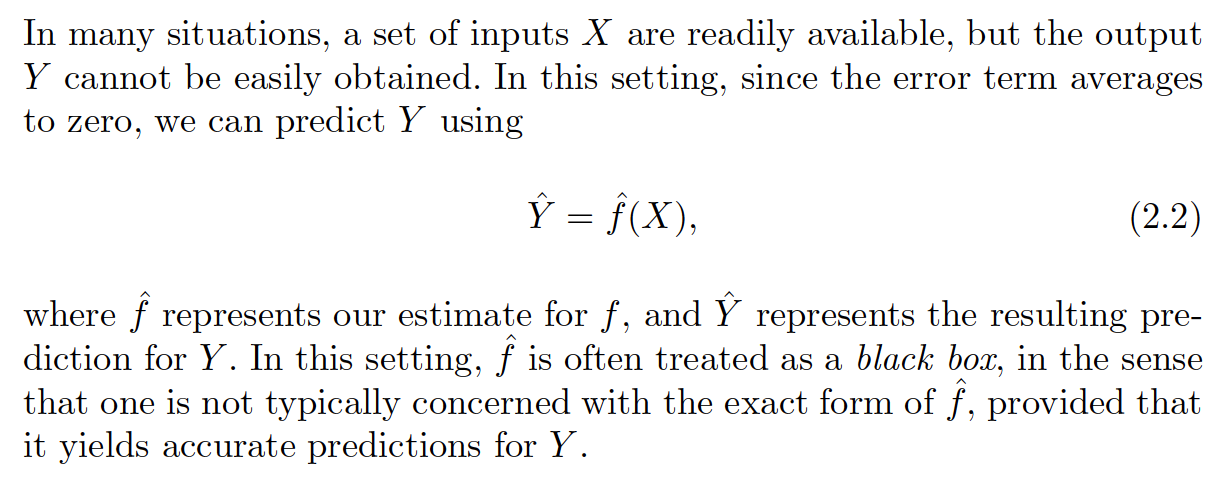


**In essence, statistical learning refers to a set of approaches for estimating .**

**Why Estimate ?**

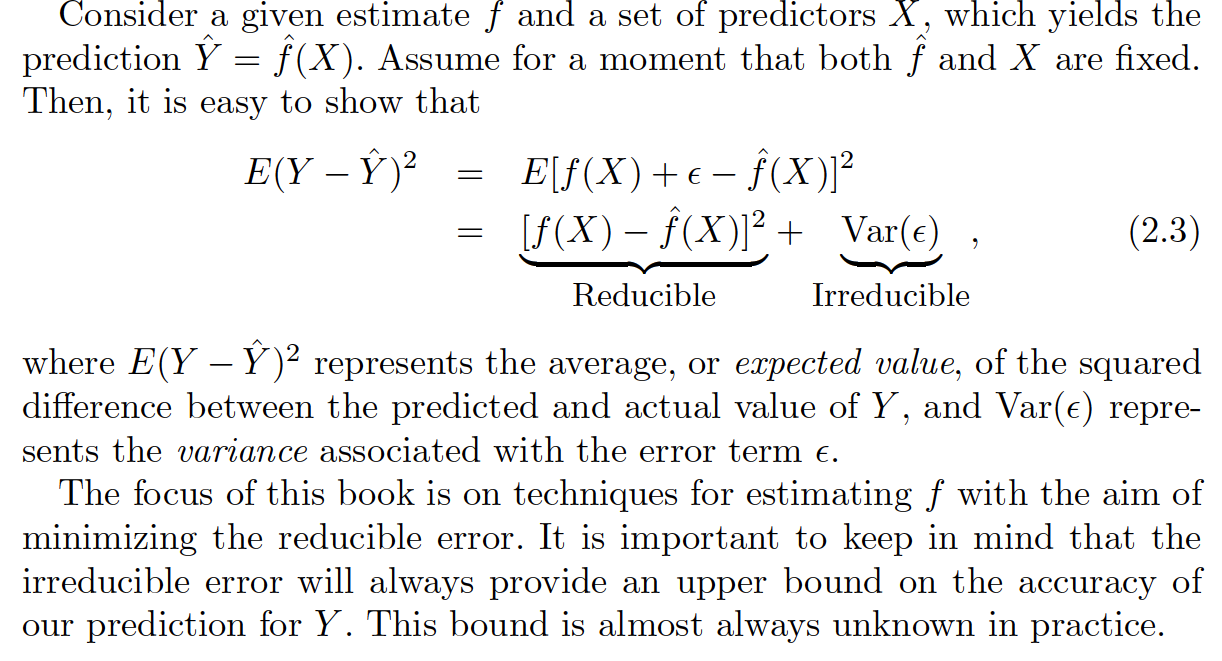
There are two main reasons that we may wish to estimate f : ***prediction*** and ***inference.***

* **Prediction**



**Page 18:**

The accuracy of ˆ Y as a prediction for Y depends on two quantities, which we will call the ***reducible error and the irreducible error*** . In general,ˆ f will not be a perfect estimate for f , and this inaccuracy will introduce some error. This error is reducible because we can potentially improve the accuracy of ˆ f by using the most appropriate statistical learning technique to estimate f . However, even if it were possible to form a perfect estimate for f , so that our estimated response took the form ˆ Y = f (X ), our prediction would still have some error in it! This is because Y is also a function of , which, by definition, cannot be predicted using X . Therefore, variability associated with also affects the accuracy of our predictions. This is known as the irreducible error, because no matter how well we estimate f , we cannot reduce the error introduced by .



* **Inference**

We are often interested in understanding the way that Y is affected as X1, . . . , Xp 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 X1, . . .,Xp . Now (ˆ 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:

1. Which predictors are associated with the response?
2. What is the relationship between the response and each predictor?
3. Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?

Let’s say a company’s goal is to identify individuals who will respond positively to a mailing, based on observations of demographic variables measured on each individual. In this case, the demographic variables serve as predictors, and response to the marketing campaign (either positive or negative) serves as the outcome. The company is not interested in obtaining a deep understanding of the relationships between each individual predictor and the response; instead, the company simply wants an accurate model to predict the response using the predictors. This is an example of ***modeling for prediction.***

Another example involves modeling the brand of a product that a customer might purchase based on variables such as price, store location, discount levels, competition price,and so forth. In this situation one might really be most interested in how each of the individual variables affects the probability of purchase. For instance, what effect will changing the price of a product have on sales? This is an example of ***modeling for inference.***

Finally, some modeling could be conducted both for prediction and inference. For example, in a real estate setting, one may seek to relate values of homes to inputs such as crime rate, zoning, distance from a river, air quality, schools, income level of community, size of houses, and so forth. In this case one might be interested in how the individual input variables affect the prices—that is, *how much extra will a house be worth if it has a view of the river?*  This is an **inference problem.** Alternatively, one may simply be interested in predicting the value of a home given its characteristics: *is this house under- or overvalued?* This is a **prediction problem.**

**What kind of model, should we choose for prediction or inference?**

Depending on whether our ultimate goal is prediction, inference, or a

combination of the two, different methods for estimating f may be appropriate.

For example, linear models allow for relatively simple and interpretable inference, but may not yield as accurate predictions as some other approaches. In contrast, some of the highly non-linear approaches that we discuss in the later chapters of this book can potentially provide quite accurate predictions for Y , but this comes at the expense of a less interpretable model for which inference is more challenging

**How do we estimate ?**

Most statistical learning methods for this task can be characterized as either parametric or non-parametric.

* **Parametric Methods:** Parametric methods involve a two-step model-based approach.

First, we make an assumption about the functional form, or shape, of ‘**’.** Whether it needs linear or more complex model

Second, After the model has been selected, we need a procedure that uses the training data to fit or train model.like, in the case of linear model we need to estimate the parameters for input variables in linear equation (i.e least square is one of many possible to fit the linear model.

Example : Smooth thin plate spline. Linear or lower order regressions

***Potential Disadvantage of parametric methods***: the model we choose will usually not match the true unknown form of f . If the chosen model is too far from the true f , then our estimate will be poor. We can try to address this problem by choosing

flexible models that can fit many different possible functional forms for f . But in general, fitting a more flexible model requires estimating a greater number of parameters. These more complex models can lead to a phenomenon known as ***overfitting*** the data, which essentially means they overfitting follow the errors, or noise , too closely.

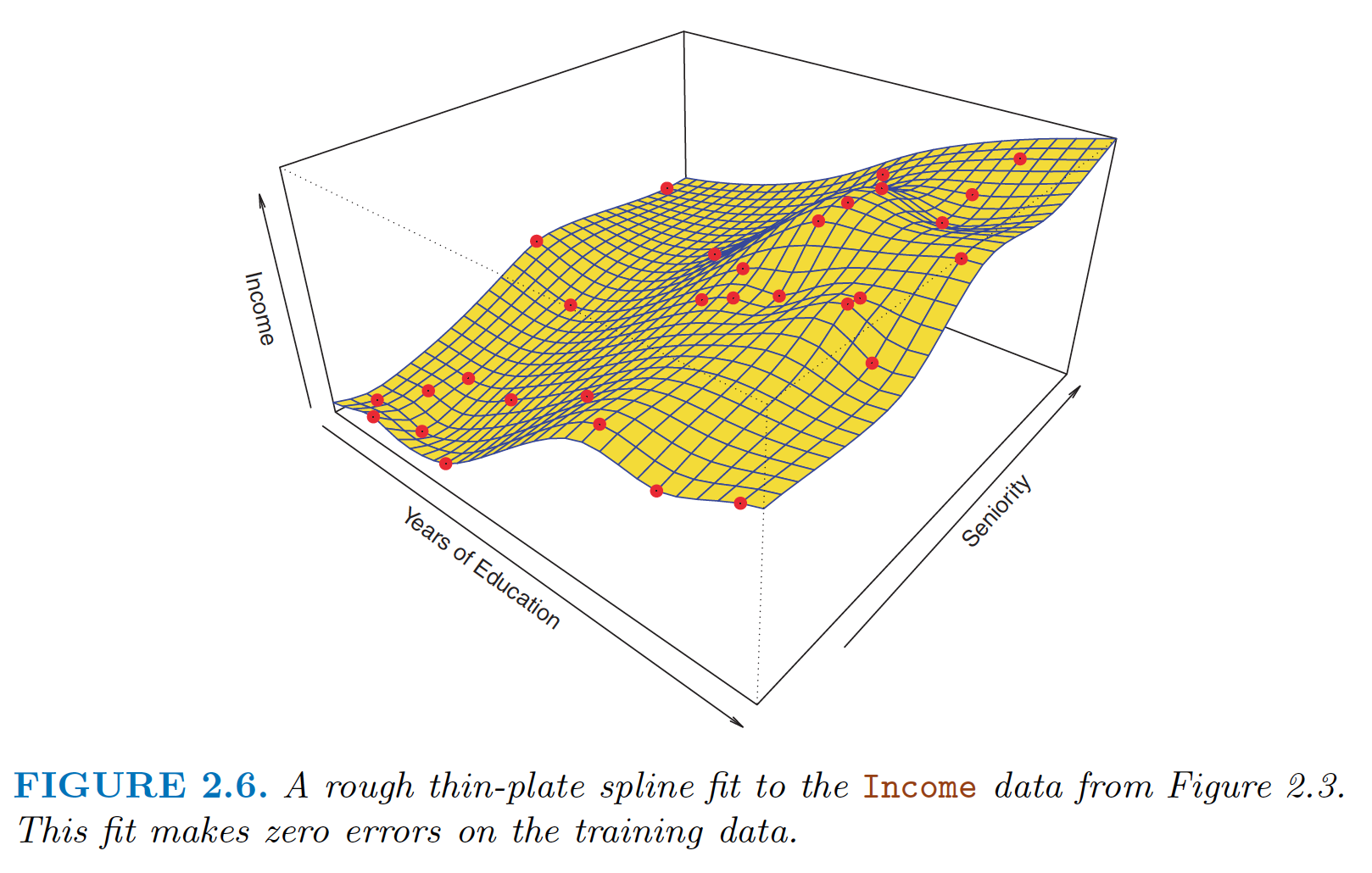
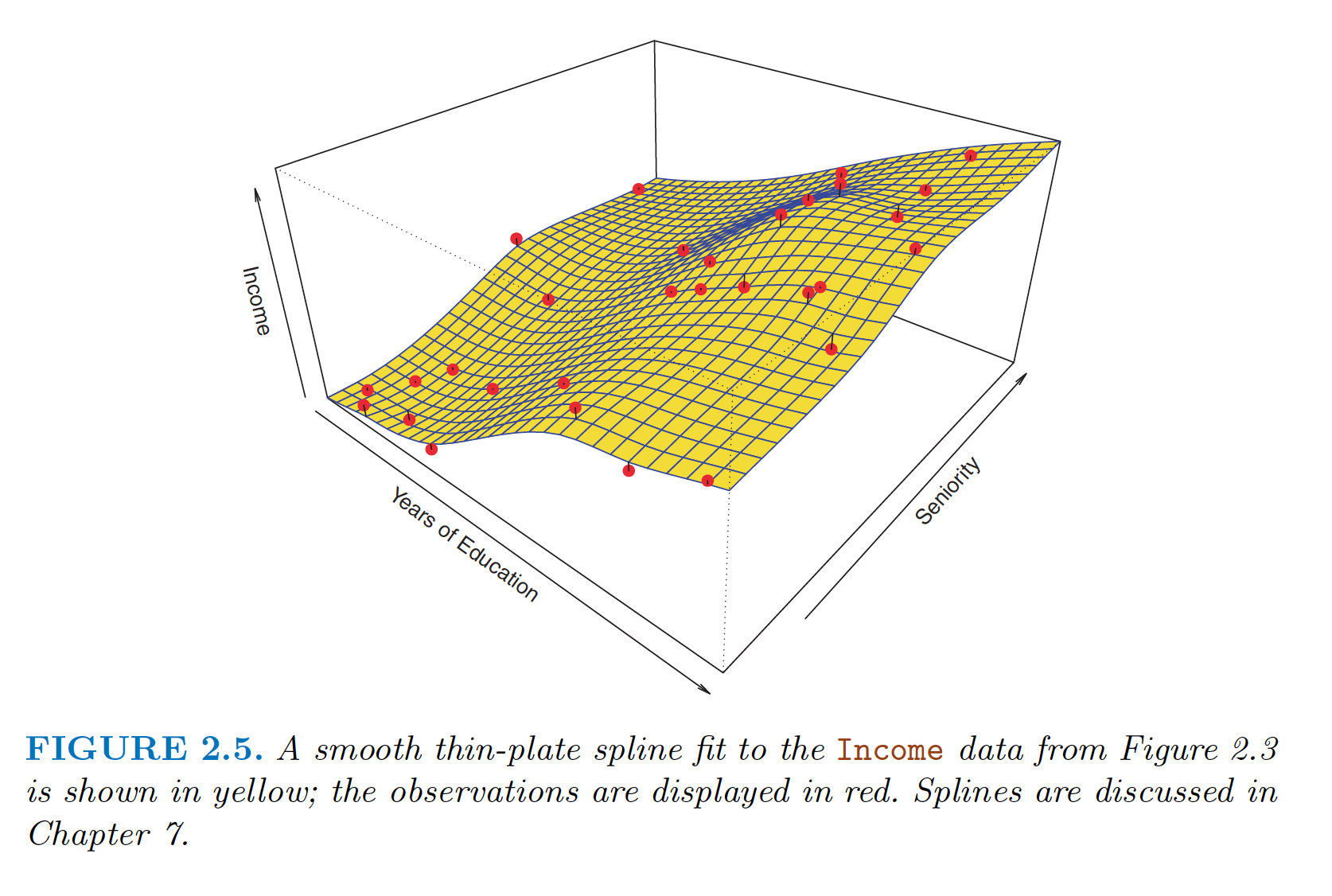
* **Non Parametric methods:** Non-parametric methods do not make explicit assumptions about the functional form of . Instead they seek an estimate of f that gets as close to the data points as possible without being too rough or wiggly .

*Advantage* : By avoiding the assumption of a particular functional form for , they have the potential to accurately fit a wider range of possible shapes for

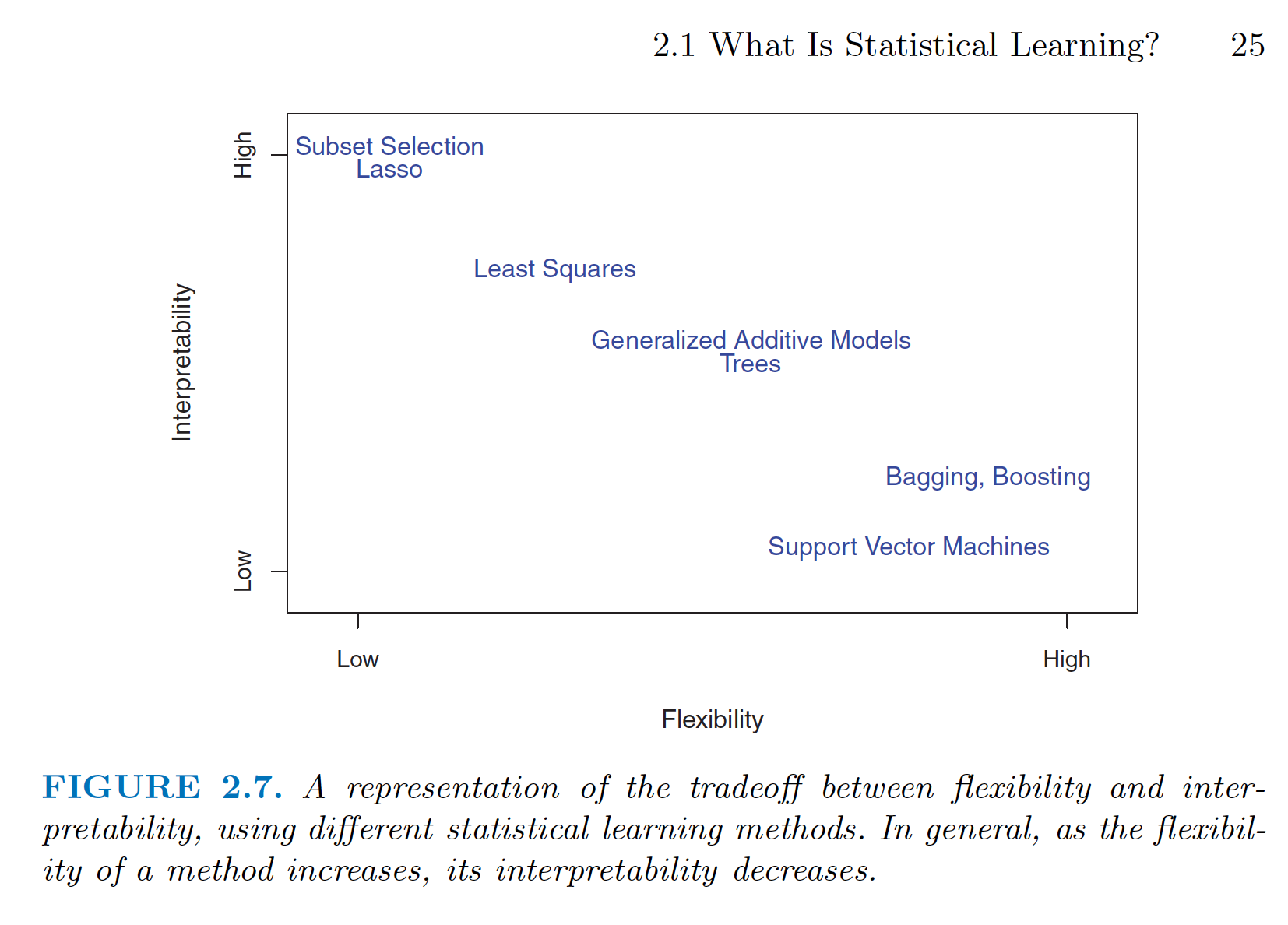
*Disadvantage:*  Since they do not reduce the problem of estimating f to a

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 . Also, it suffers from **overfitting**.

Example : Rough thin plate spline. KNN



**The Trade-Off Between Prediction Accuracy and Model Interpretability :**



One might reasonably ask the following question: *why would we ever choose to use a more restrictive method instead of a very flexible approach?*

There are several reasons that we might prefer a more restrictive model.

If we are mainly interested in inference, then restrictive models are much

more interpretable.

In some settings, however, we are only interested in prediction, and the interpretability of the predictive model is simply not of interest. 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 at first glance, has to do with the potential for overfitting in highly flexible methods.

**Supervised Versus Unsupervised Learning:**

*Supervised Learning* : For each observation of the predictor measurement(s) xi , i = 1, . . . , n there is an associated response measurement yi . We wish to fit a model that relates the response to the predictors, with the aim of accurately predicting the response for future observations (prediction) or better understanding the relationship between the response and the predictors (inference).

*Unsupervised Learning:*  For every observation i = 1, . . . , n , we observe a vector of measurements xi but no associated response yi . It is not possible to fit a linear regression model, since there is no response variable to predict. In this setting, we are in some sense working blind; the situation is referred to as unsupervised because we lack a response variable that can supervise our analysis.

*Semi-Supervised Learning:*  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 for which they are not.

**Regression v/s Classification:**

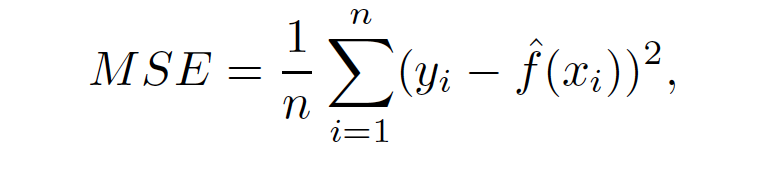
We tend to refer to problems with a quantitative response as regression problems, while those involving a qualitative response are often referred to as classification problems.However, the distinction is not always that crisp. Least squares linear regression is used with a quantitative response, whereas logistic regression (Chapter 4) is typically used with a qualitative (two-class, or binary ) response. As such it is often used as a classification method. But binary since it estimates class probabilities, it can be thought of as a regression method as well. Some statistical methods, such as K -nearest neighbors and boosting can be used in the case of either quantitative or qualitative responses.

We tend to select statistical learning methods on the basis of whether the response is quantitative or qualitative. **“However, whether the predictors are qualitative or quantitative is generally considered less important.”**

**Assessing Model Accuracy (For regression setting):**

It is an important task to decide for any given set of data which method produces the best results. Selecting the best approach can be one of the most challenging parts of performing statistical learning in practice.

**Measuring the Quality of Fit:** We need to quantify the extent to which the predicted response value for a given observation is close to the true response value for that observation. In the regression setting, the most commonly-used measure is the mean squared error (MSE), given by

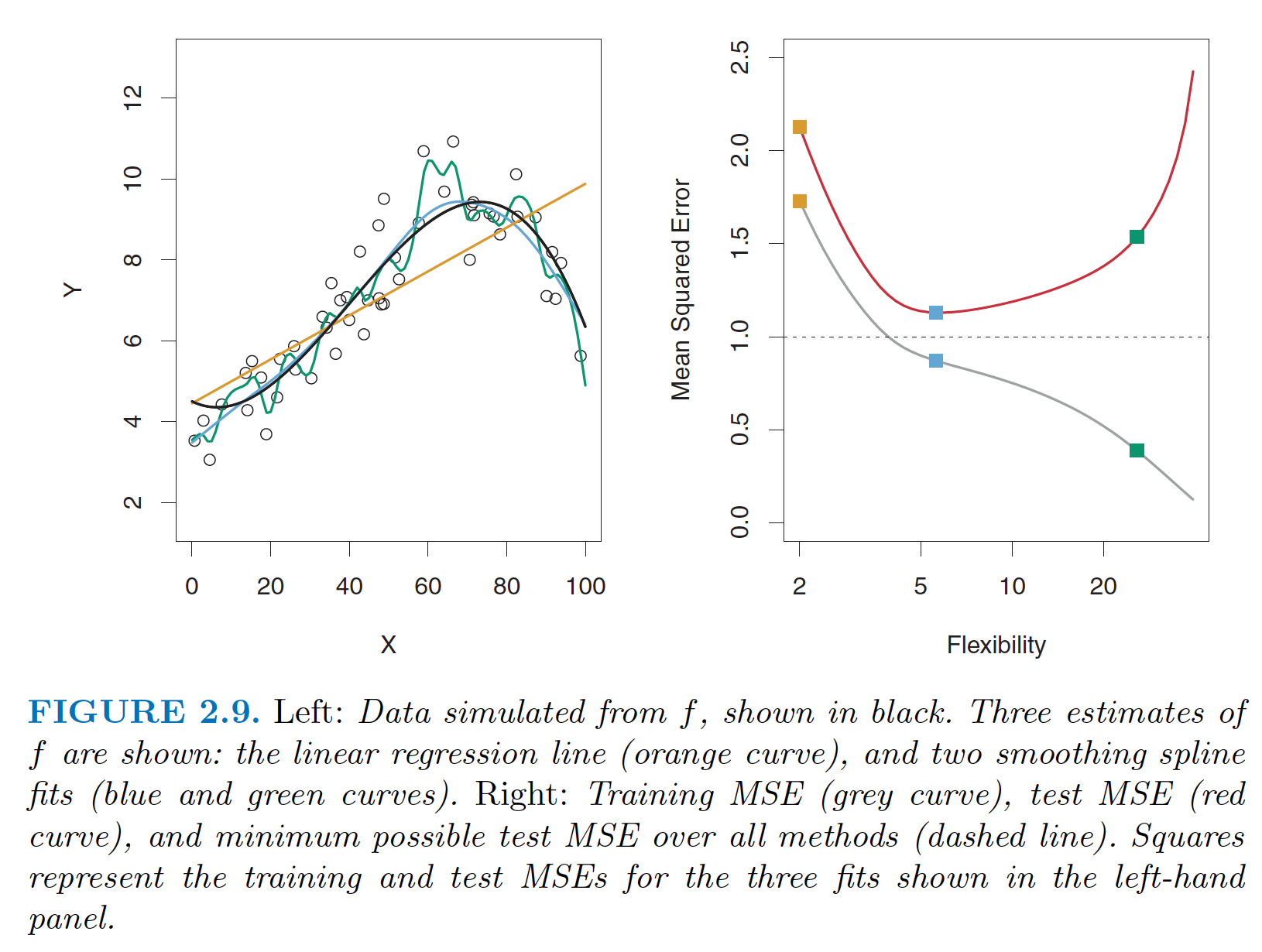
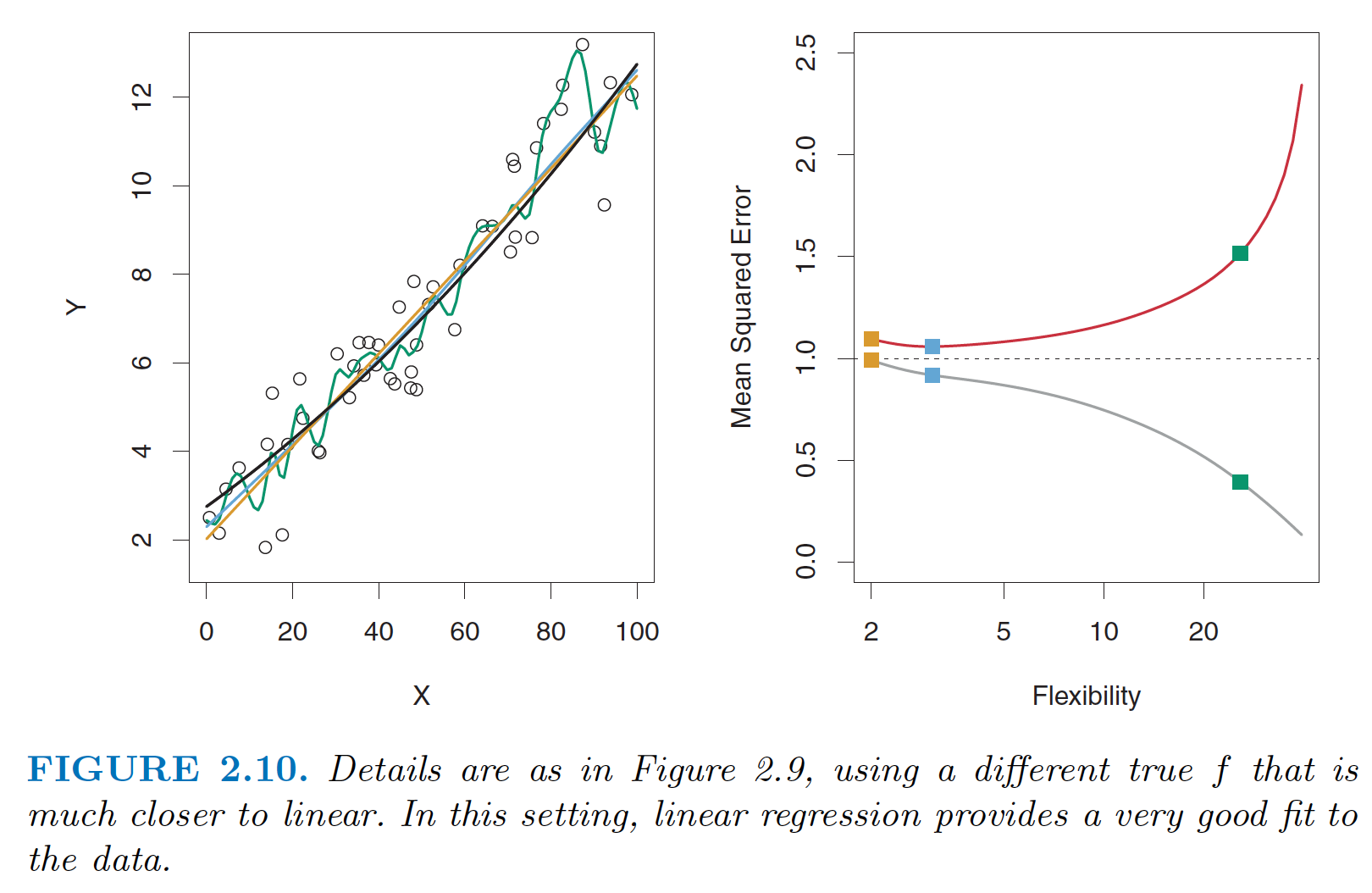


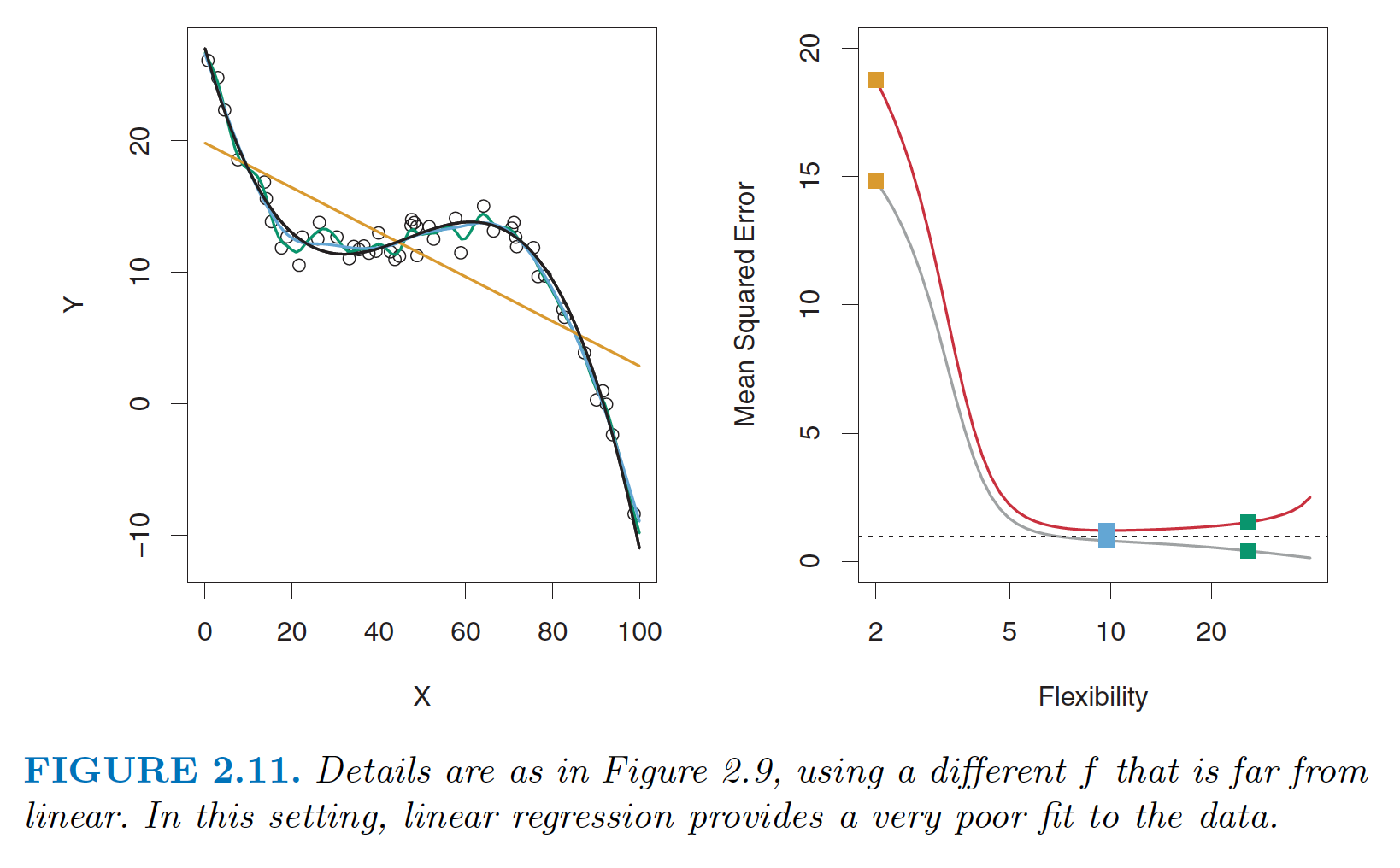
**The *degrees of freedom* is a quantity that summarizes the *flexibility* of a curve.**

A more restricted and hence smoother curve has fewer degrees of freedom than a wiggly curve. linear regression is at the most restrictive end, with two degrees of freedom. The training MSE declines monotonically as flexibility increases. In this example the true f is non-linear, and so the orange linear fit is not flexible enough to estimate f well. The green curve has the lowest training MSE of all three methods, since it corresponds to the most flexible of the three curves fit in the left-hand panel.

As model flexibility increases, training MSE will decrease, but the test MSE may not. When a given method yields a small training MSE but a large test MSE, we are said to be overfitting the data. This happens because our statistical learning procedure is working too hard to find patterns in the training data, and may be picking up some patterns that are just caused by random chance rather than by true properties of the unknown function f .

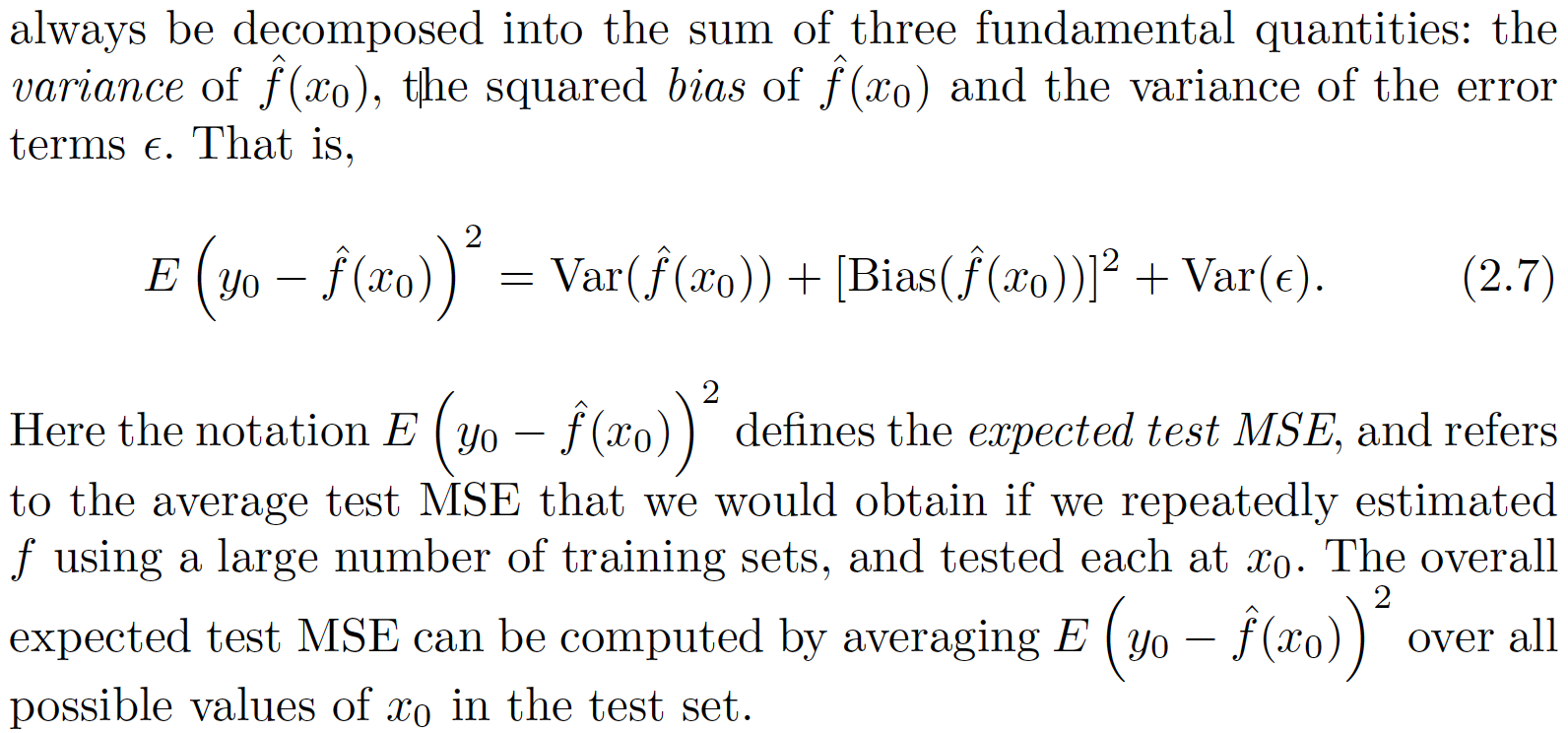
***Overfitting :*** Overfitting refers specifically to the case in which a less flexible model would have yielded a smaller test MSE than more flexible model.





**The Bias-Variance Trade-Off:**

The expected test MSE, for a given value x0 , can

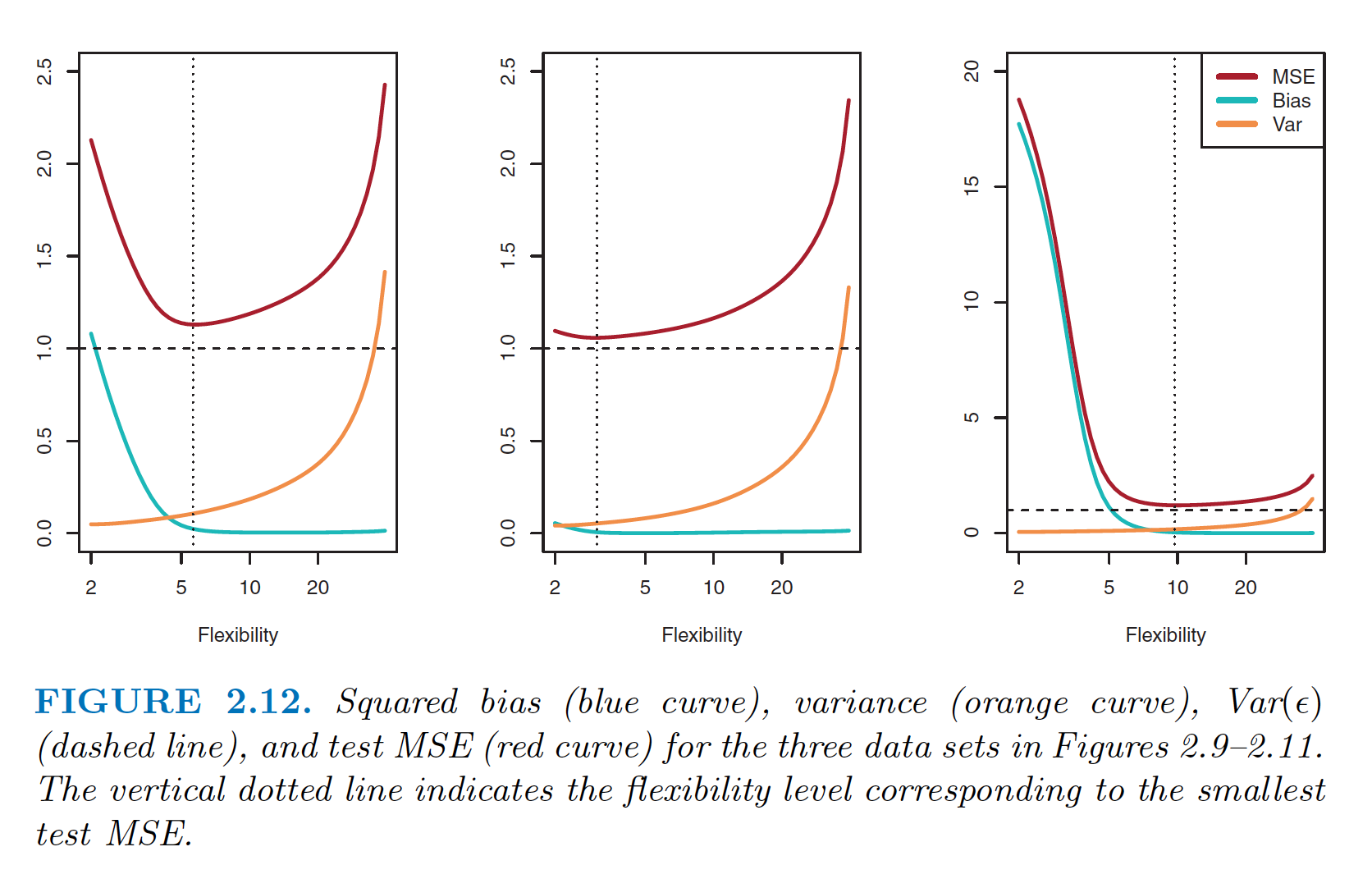


***Variance:*** refers to the amount by which ˆ would change if we estimated it using a different training data set. Since the training data are used to fit the statistical learning method, different training data sets will result in a different ˆ. But ideally the estimate for should not vary too much between training sets. However, if a method has high variance then small changes in the training data can result in large changes in ˆ. ***In general, more flexible statistical methods have higher variance*.** Consider the green and orange curves in Figure 2.9. The flexible green curve is following the observations very closely. It has high variance because changing any one of these data points may cause the estimate ˆ to change considerably. In contrast, the orange least squares line is relatively inflexible and has low variance, because moving any single observation will likely cause only a small shift in the position of the line.

***Bias:***  The error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model. For example, linear regression assumes that there is a linear relationship between Y and X1,X2, . . . , Xp . It is unlikely that any real-life problem truly has such a simple linear relationship, and so performing linear regression will undoubtedly result in some bias in the estimate of f . In Figure 2.11, the true f is substantially non-linear, so no matter how many training observations we are given, it will not be possible to produce an accurate estimate using linear regression. In other words, linear regression results in high bias in this example. However, in Figure 2.10 the true f is very close to linear, and so given enough data, it should be possible for linear regression to produce an accurate estimate.

***Generally, more flexible methods result in less bias.***

\*\*\*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 we increase the flexibility of a class of methods, the bias tends to initially decrease faster than the variance increases. Consequently, the expected test MSE declines. However, at some point increasing flexibility has little impact on the bias but starts to significantly increase the variance. When this happens the test MSE increases. Note that we observed this pattern of decreasing test MSE followed by increasing test MSE in the right-hand panels of Figures 2.9–2.11.



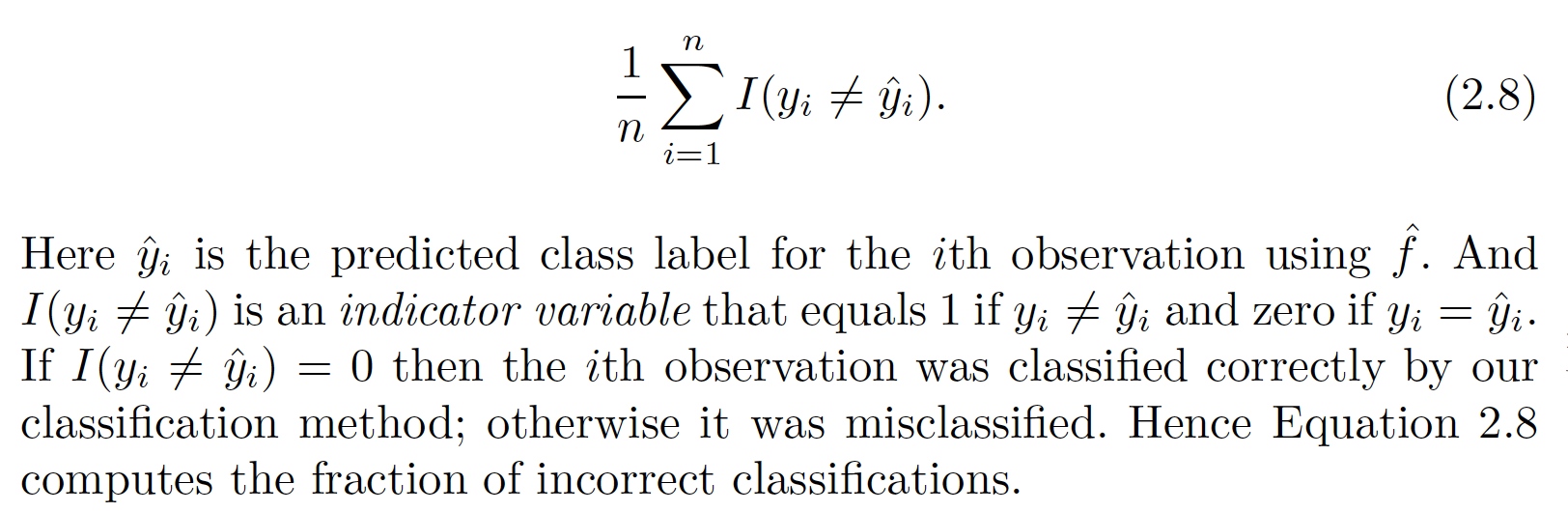
The horizontal dashed line represents Var( ), the irreducible error.

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.

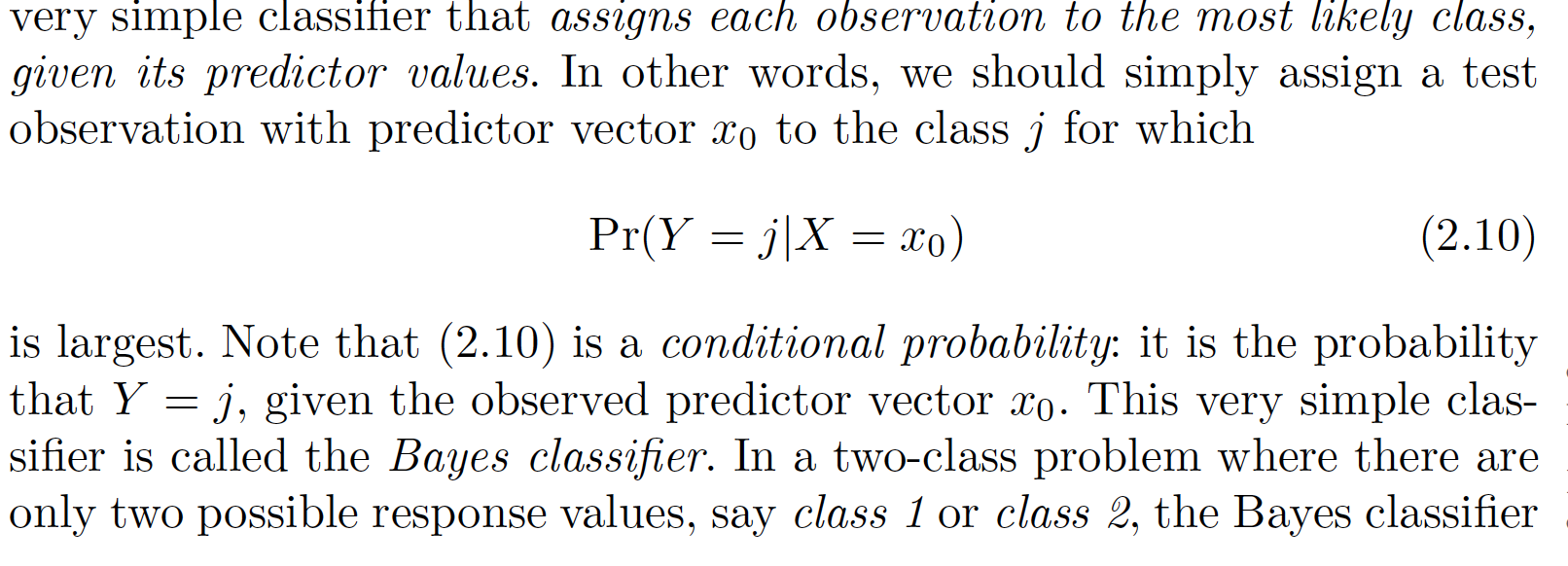
In a real-life situation in which f is unobserved, it is generally not possible to explicitly compute the test MSE, bias, or variance for a statistical learning method. Nevertheless, one should always keep the bias-variance trade-off in mind. Read ***cross-validation*** in future slides, which is a way to estimate the test MSE using the training data.

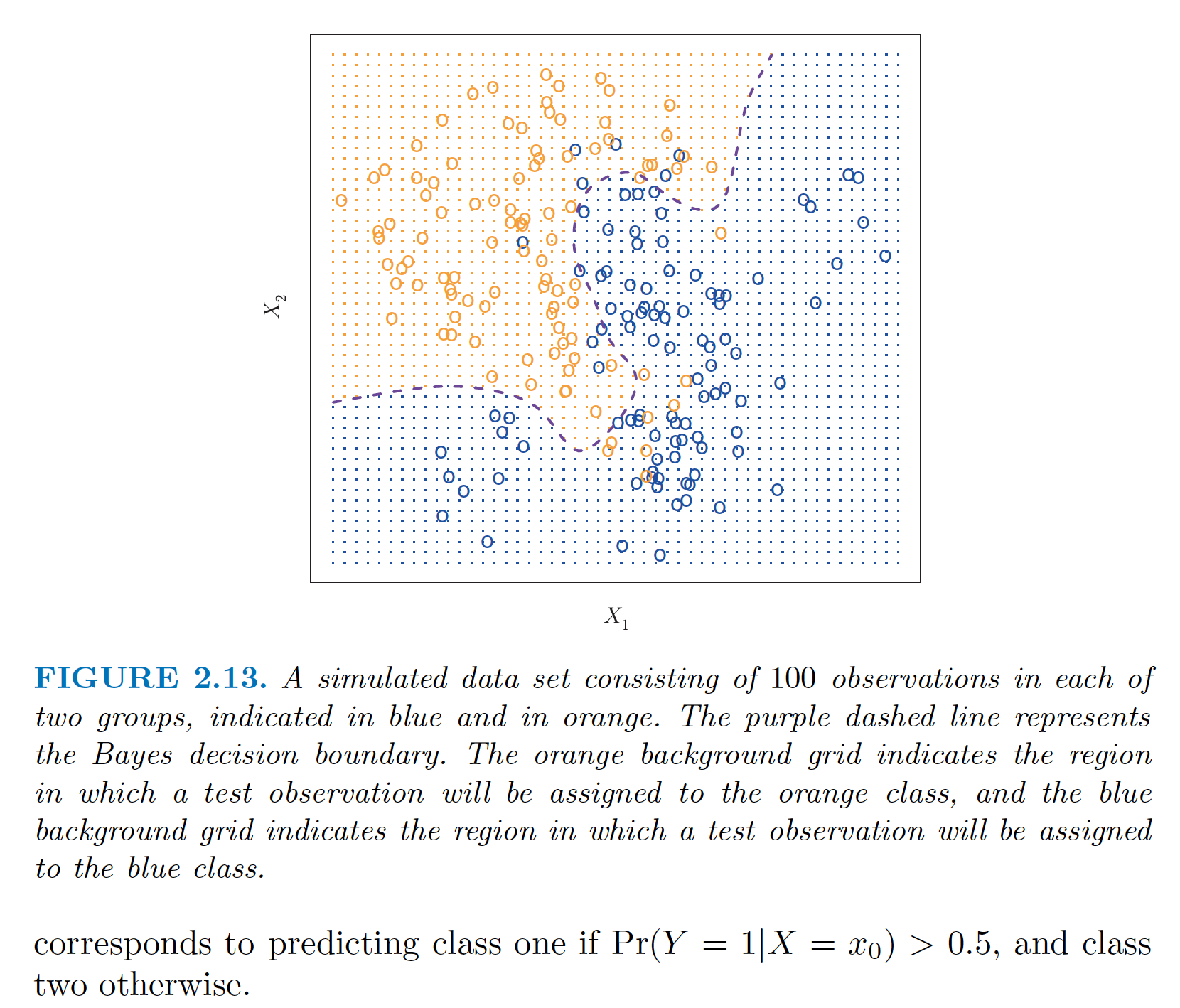
**Assessing Model Accuracy (For classification setting):**

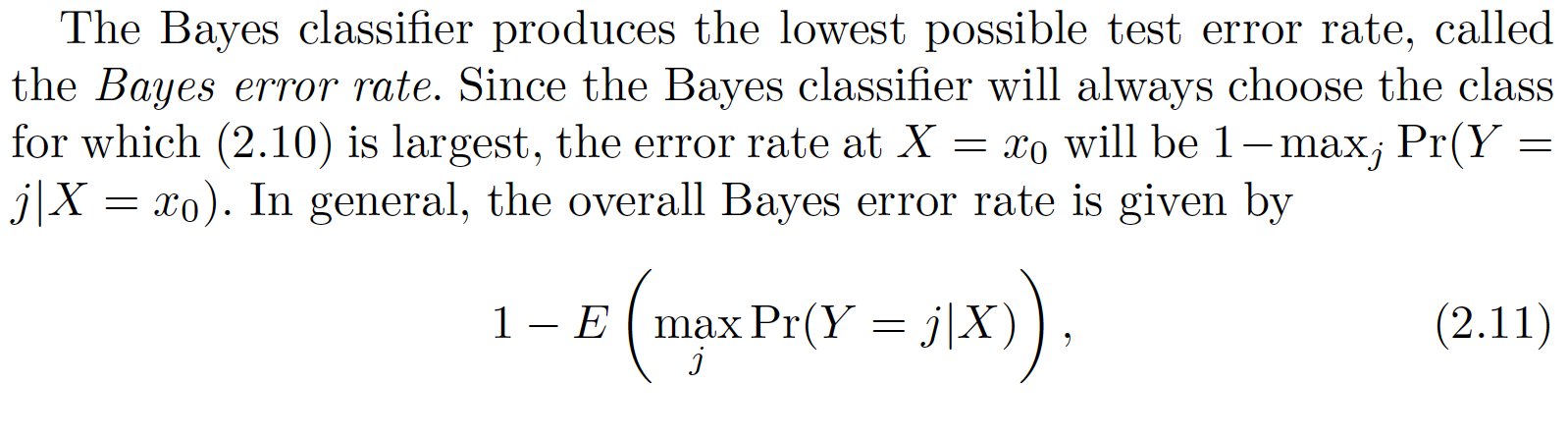
**Error rate:**

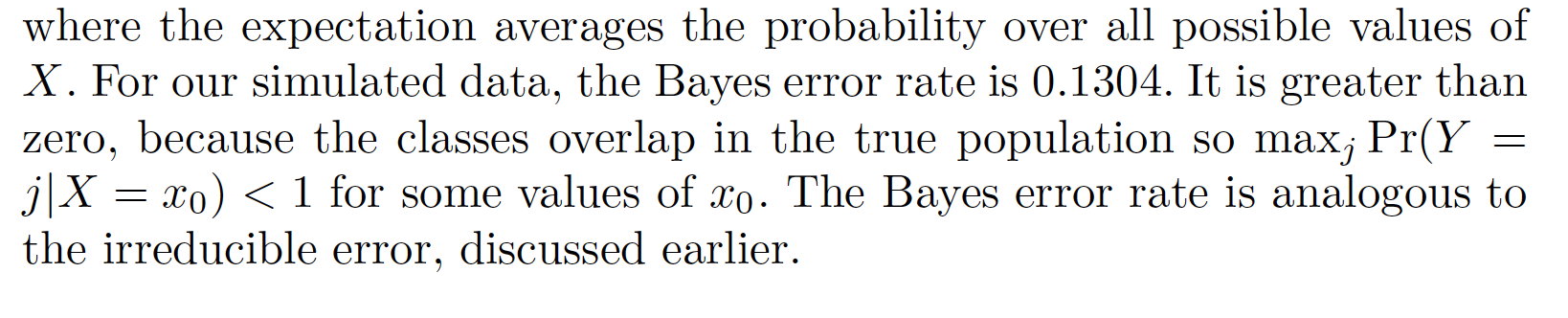


**The Bayes Classifier:**



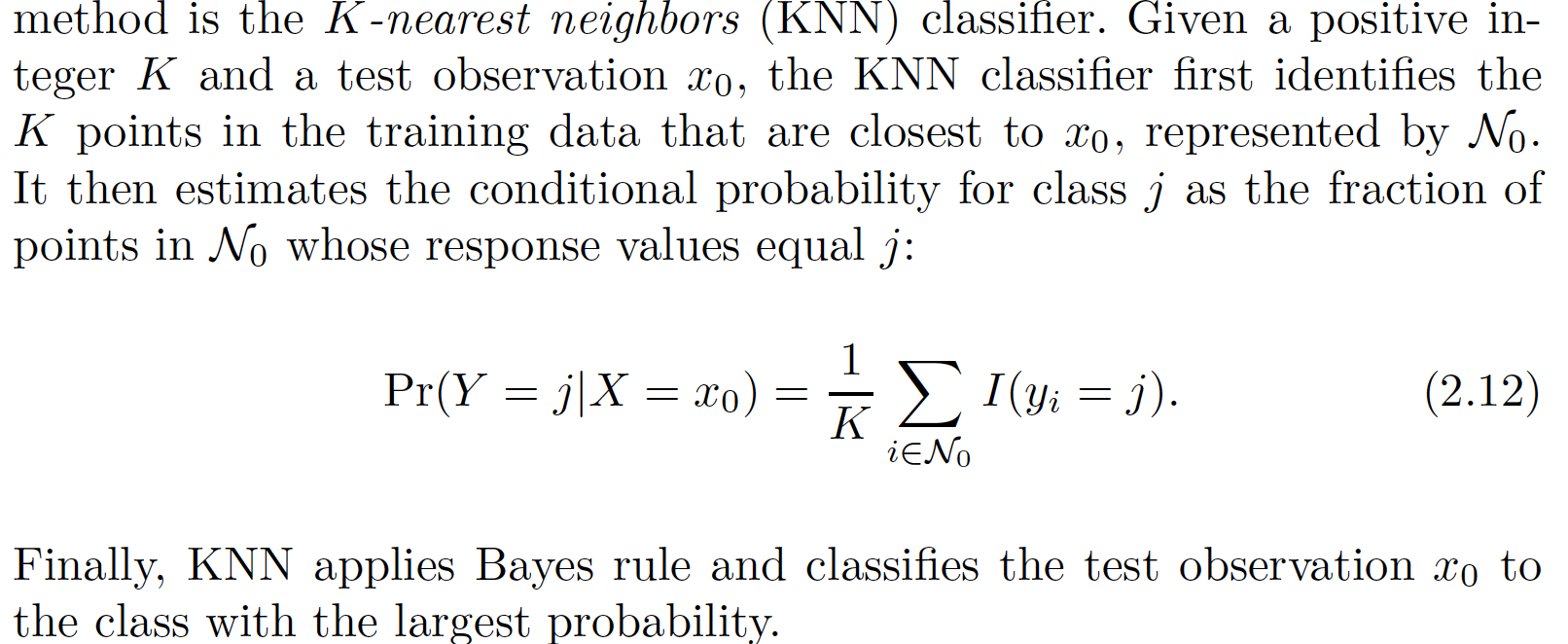


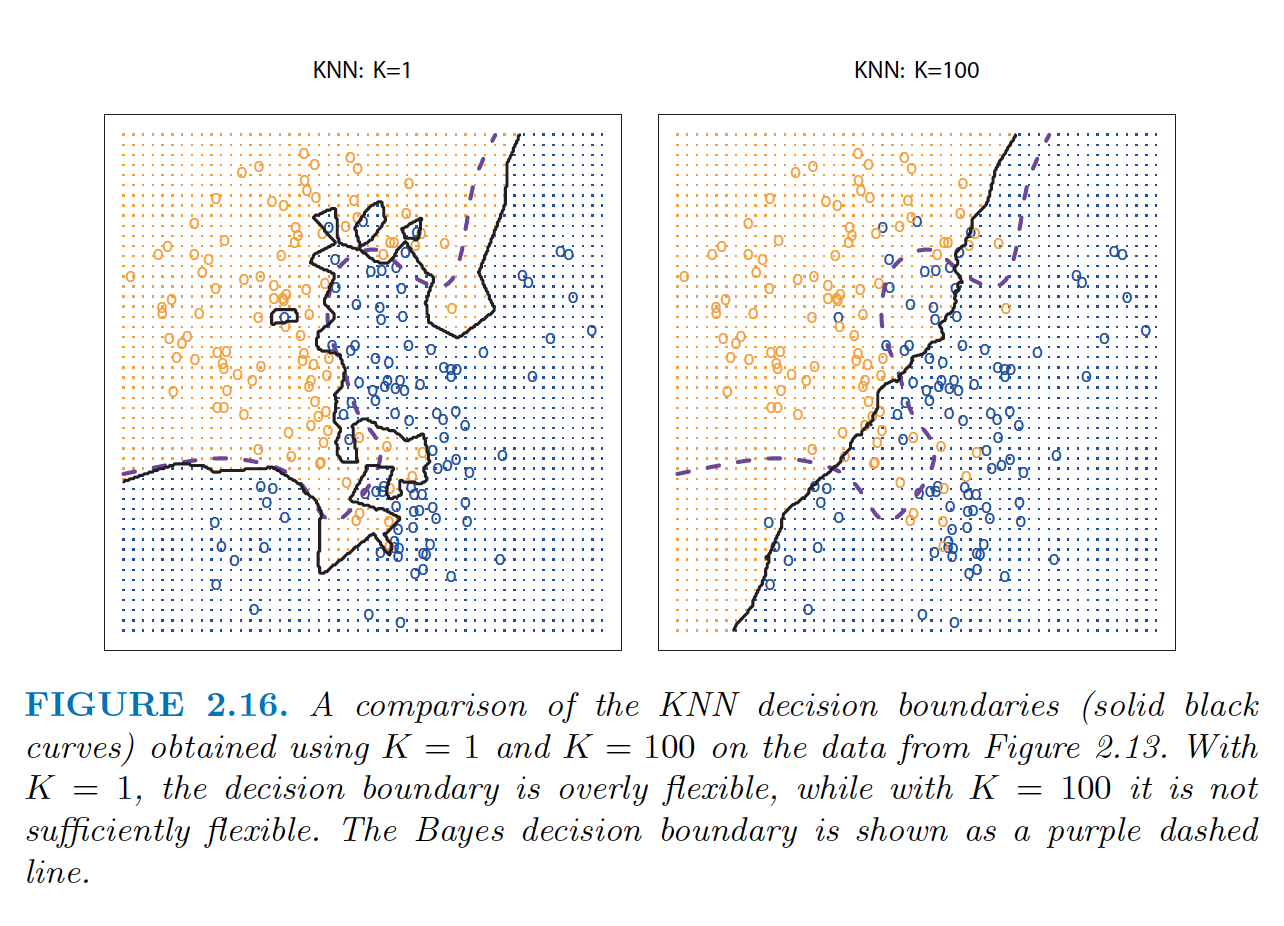


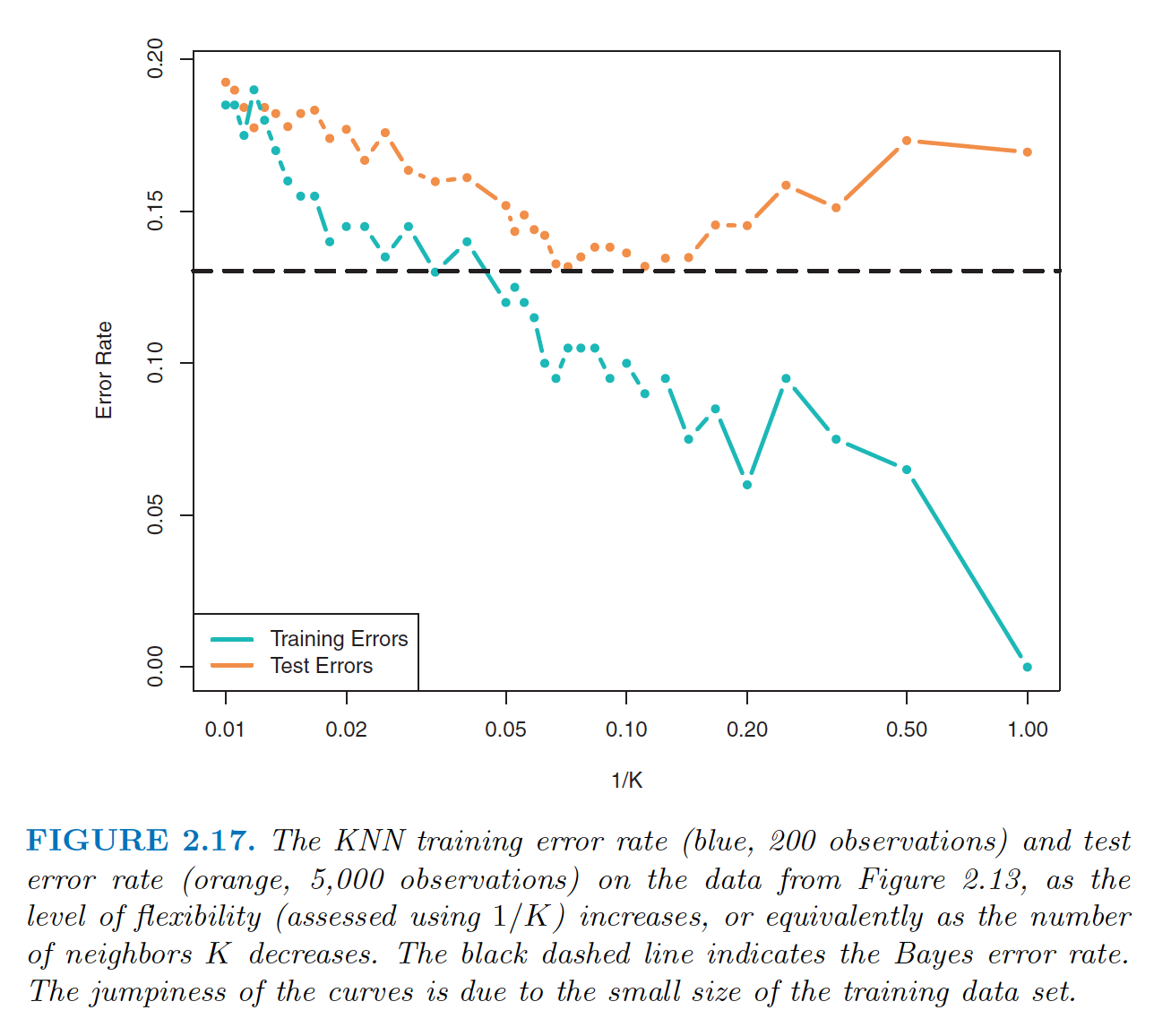
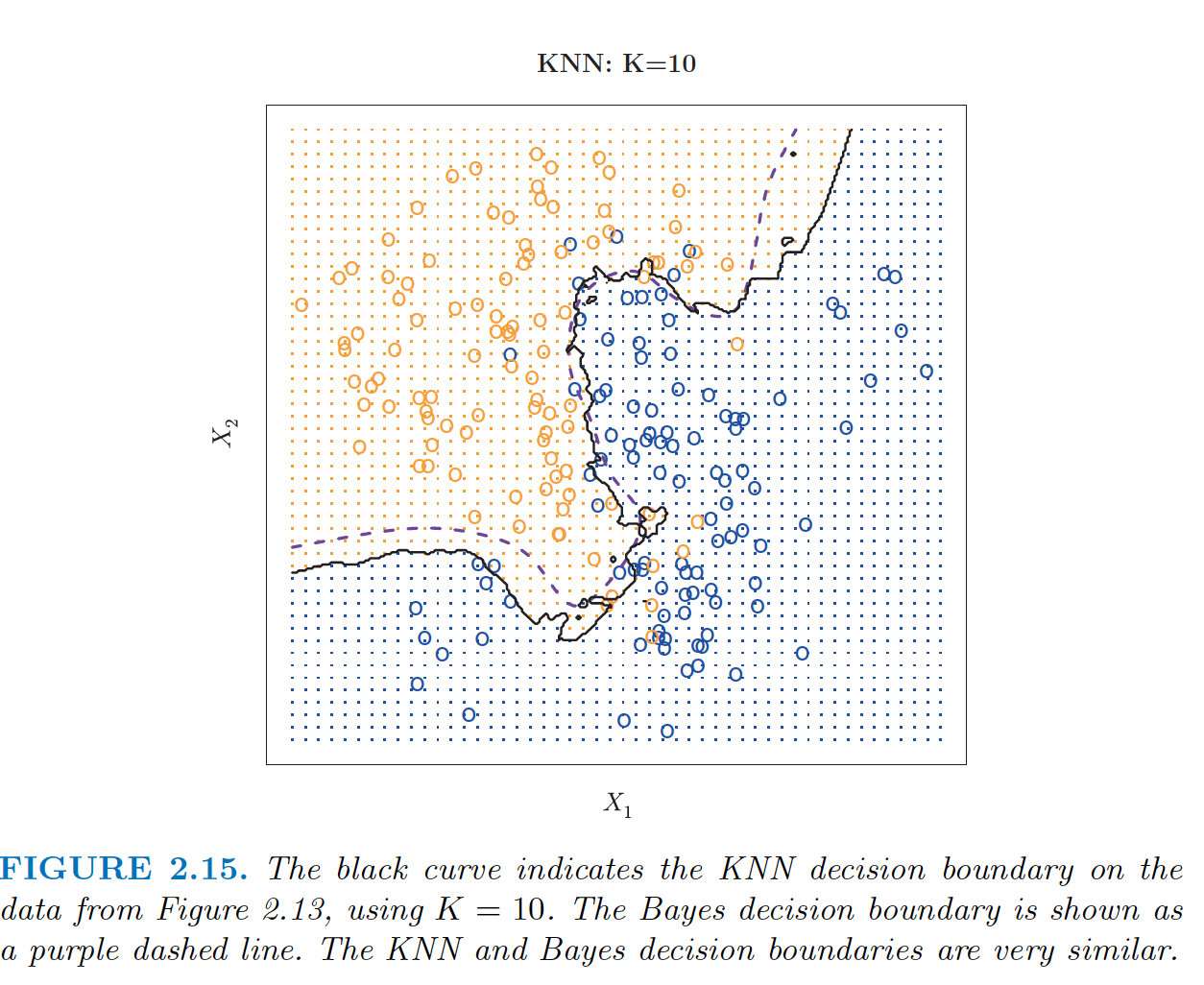


*In theory we would always like to predict qualitative responses using the Bayes classifier.* ***But for real data, we do not know the conditional distribution of Y given X , and so computing the Bayes classifier is impossible.*** *Therefore, the Bayes classifier serves as an unattainable gold standard against which to compare other methods. Many approaches attempt to estimate the conditional distribution of Y given X , and then classify a given observation to the class with highest estimated probability. One such method is the K-nearest neighbors (KNN) classifier.*

**K-Nearest Neighbors:**



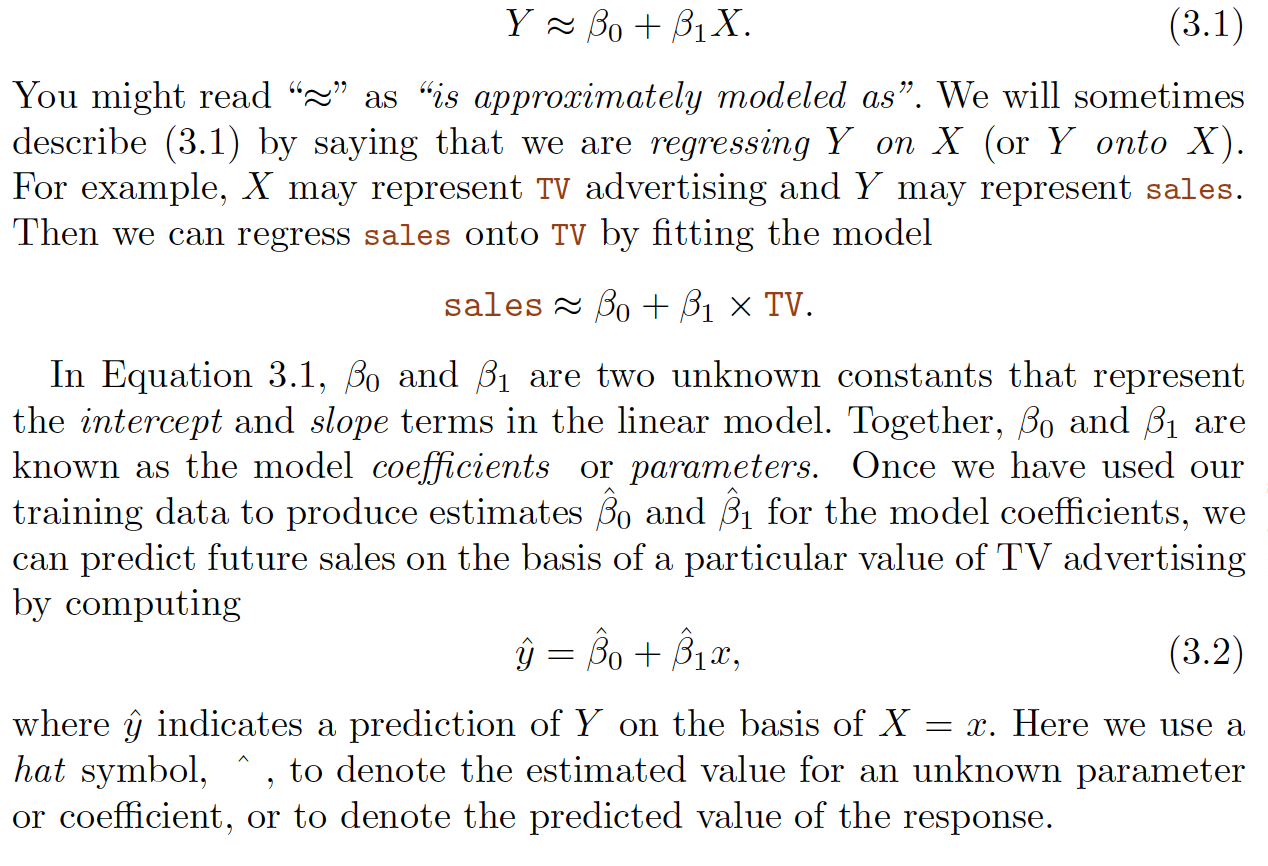




In both the regression and classification settings, choosing the correct level of flexibility is critical to the success of any statistical learning method. The bias-variance tradeoff, and the resulting U-shape in the test error, can make this a difficult task. In Chapter 5, we return to this topic and discuss various methods for estimating test error rates and thereby choosing the optimal level of flexibility for a given statistical learning method.

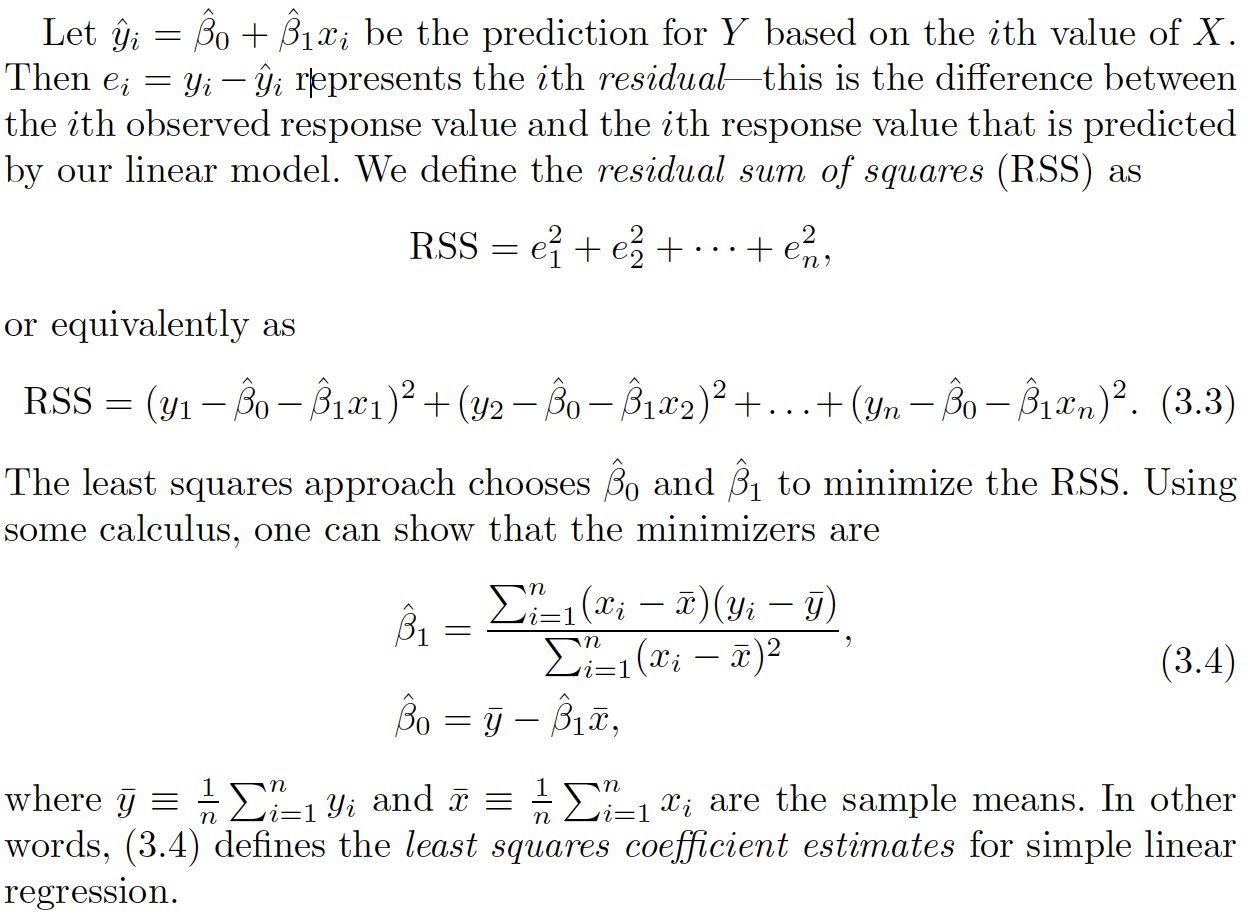
**Chapter 3-Linear Regression:**

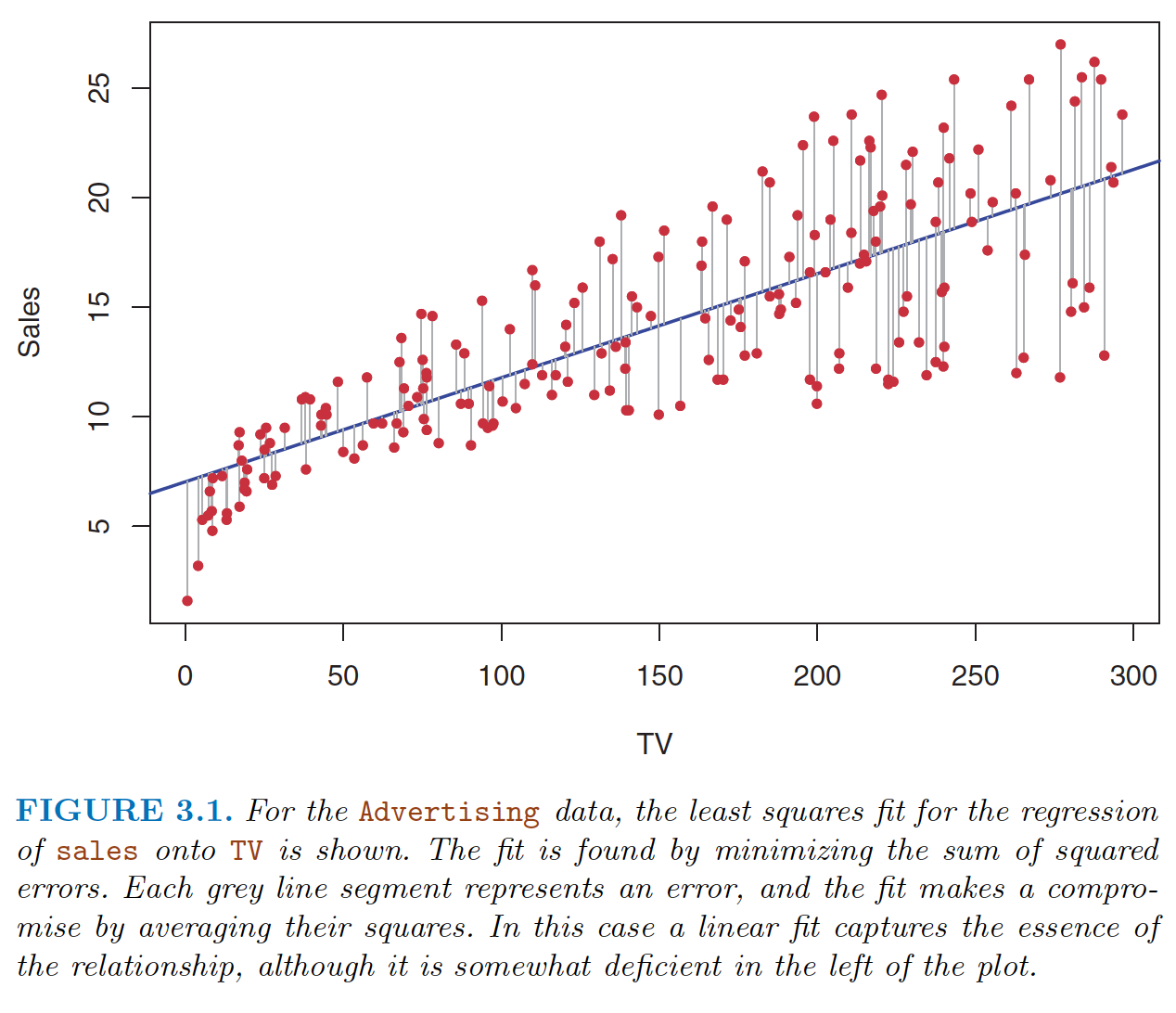
**Simple Linear Regression:** It is a very straightforward simple linear approach for predicting a quantitative response Y on the basis of a single predictor variable X . It assumes that there is approximately a linear relationship between X and Y . Mathematically, we can write this linear relationship as

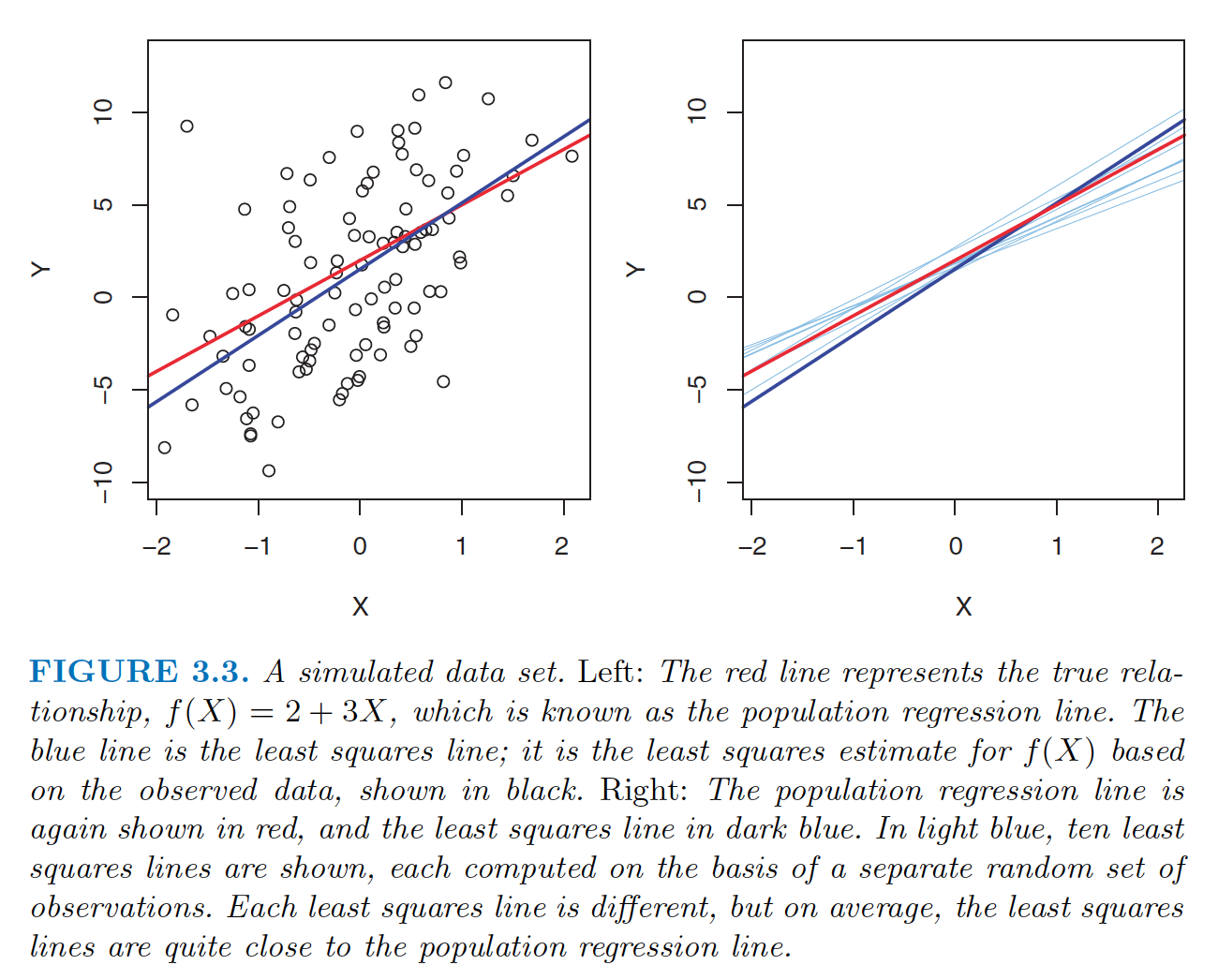
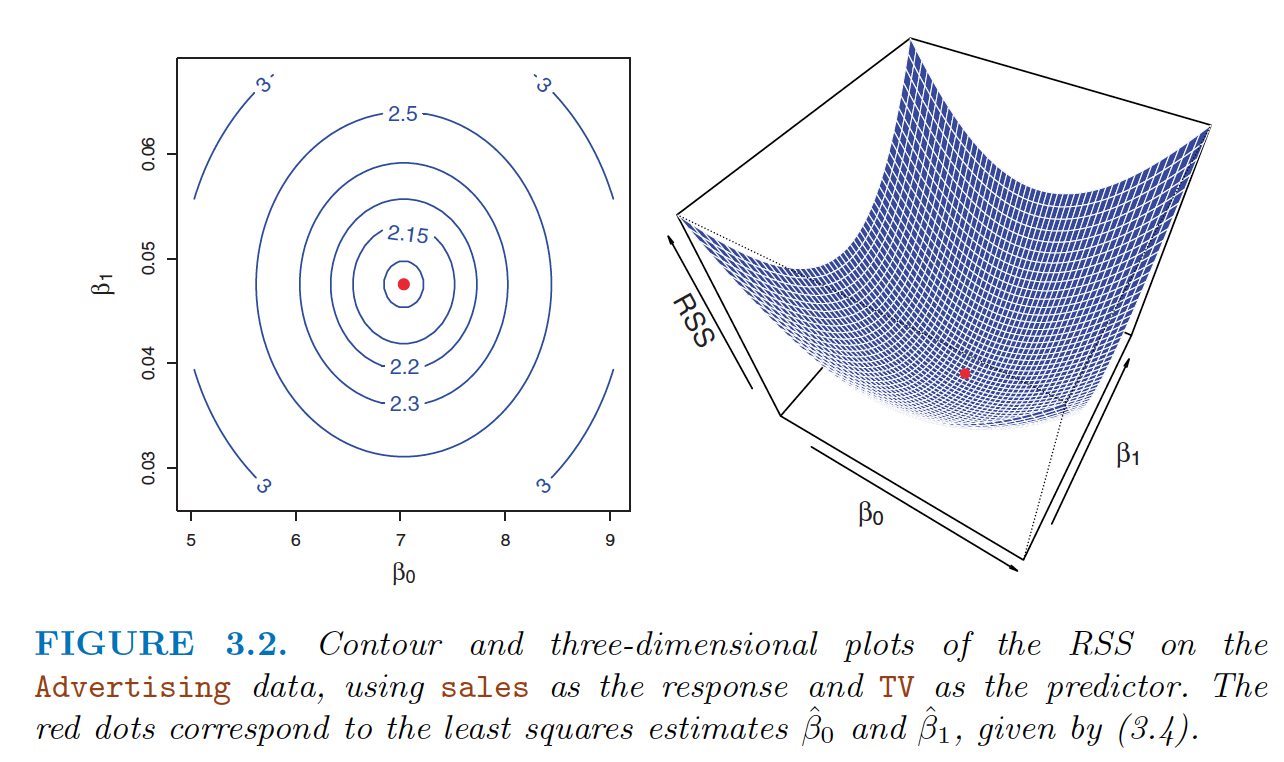


*\*\*****Advertising data*** *from Chapter 2. Figure 2.1 displays sales (in thousands of units) for a particular product as a function of advertising budgets (in thousands of dollars) for TV, radio, and newspaper media.*

We want to find an intercept and a slope such that resulting line is as close as possible to training points. By far the common approach is ‘least squares’.







The true relationship is generally not known for real data, but the least squares line can always be computed using the coefficient estimates.