MATH 478/678: Introduction to Statistical Methods in Data Science

Week 2: Chapter 2 Statistical Learning

By Yixin Fang

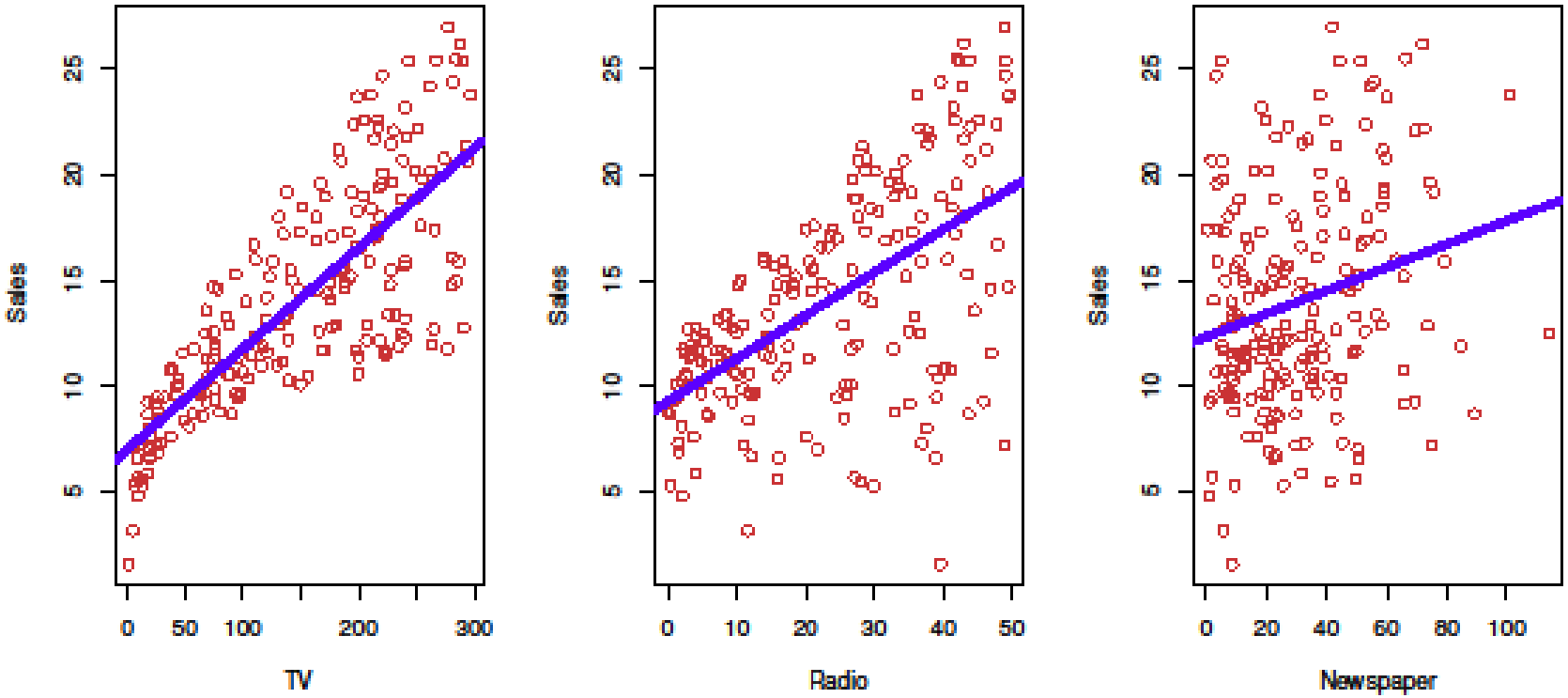
Outline of Week-2 Lectures

- Chapter 2: Statistical learning
  - 2.1 What is statistical learning?
  - 2.2 Assessing model accuracy

Example 1: Advertising data

- Suppose that we are statistical consultants hired by client to provide advice on how to improve sales of a particular product
- The Advertising dataset consists of the sales of that product in 200 different markets, along with advertising budgets in each those media: TV, radio, and newspaper
- If we determine that there is an association between advertising budgets and sales, then we can instruct our client to adjust advertising budgets and then increase sales
- In other words, our goal is to develop an accurate model that can be used to predict sales on the basis of the three media budgets

Sales vs. advertising budgets

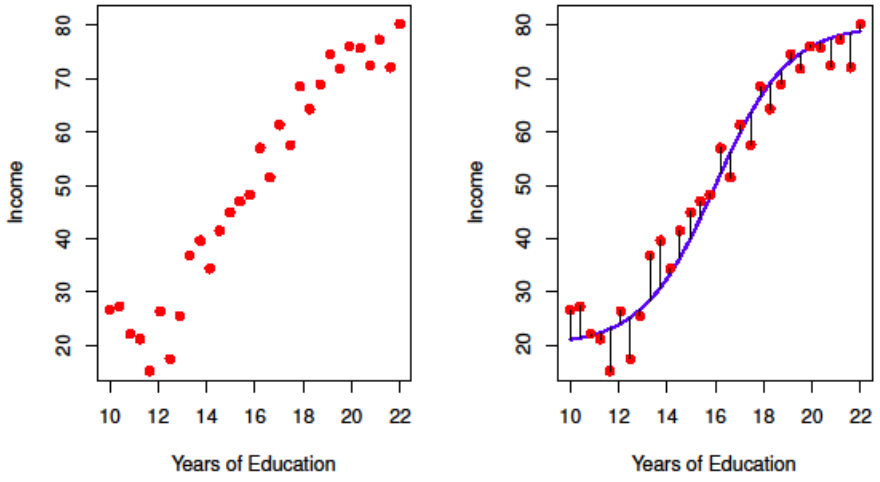


Variables Y and X

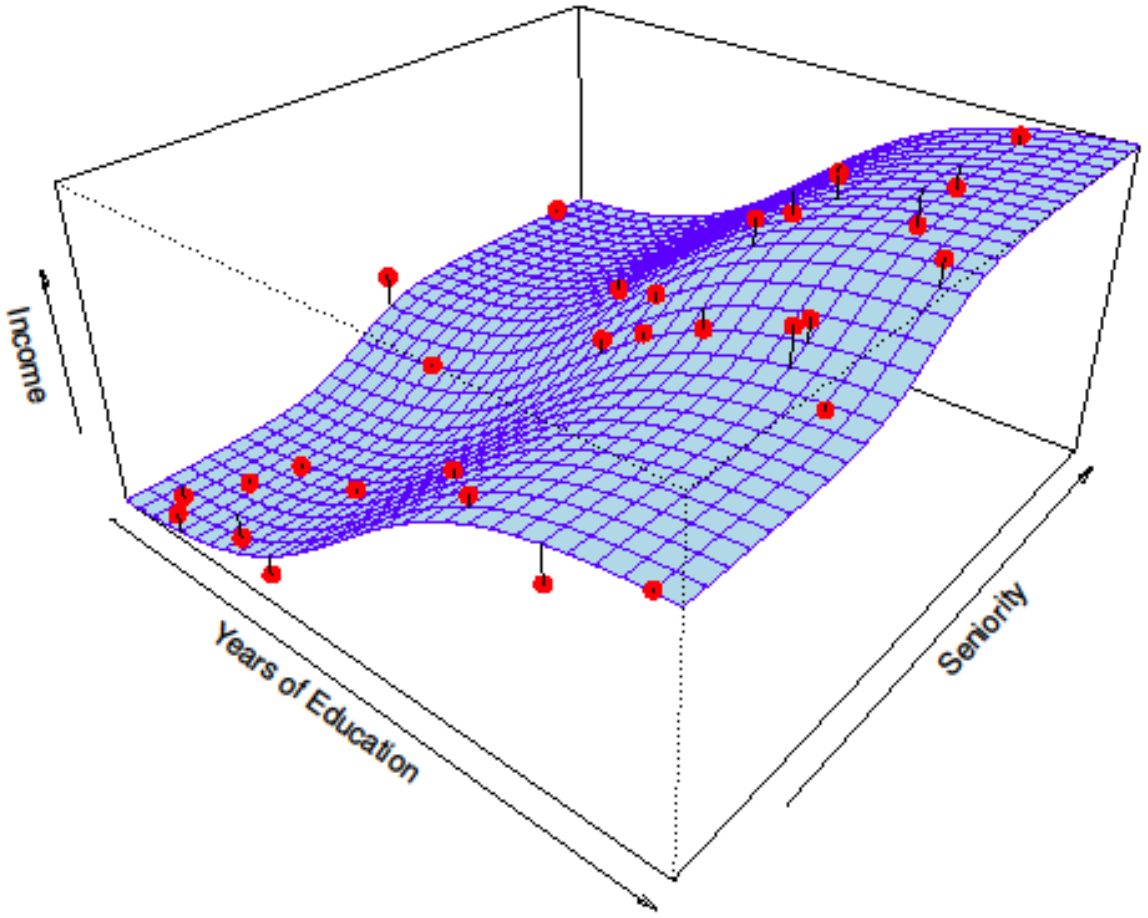
- The *Y* variable: sales
  - Output variable
  - Response variable
  - Outcome variable
  - Dependent variable
  - Target variable
- The *X* variables: X_1 for TV budget, X_2 for radio budget, and X_3 for newspaper variable
  - Input variables
  - Explanatory variables
  - Predictors
  - Independent variables
  - Features

Example 2: Income data

- Consider the income dataset consists of 30 individuals. It is a simulated dataset
- Output variable is income, and input variables are years of education and seniority



Two-dimensional surface of income vs. years of education and seniority



A general statistical model

- Suppose that we observe a quantitative outcome *Y* and *p* different predictors, $X_1, \cdots, X_p$. We assume that there is some relationship between *Y* and $X = (X_1, \cdots, X_p)$:

$$Y = f(X) + \epsilon$$

- Here *f* is some fixed but unknown function of the predictors and ϵ is a random error term, which is assumed to be independent of predictors and has mean zero. In this formula, *f* represents the systematic information that *X* provides about *Y*

Estimation of f

- There are two main reasons that we may wish to estimate f
  - Prediction
  - Inference (explanation)

Goal 1: Prediction

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

$$\hat{Y} = \hat{f}(X)$$

- Here $\hat{f}$ represents our estimate for *f*, and $\hat{Y}$ represents the resulting predictor for *Y*. In this setting, $\hat{f}$ is often treated as a ***black box***, in the sense that one is not typically concerned with its exact form, provided that it yields accurate predictors for *Y*

The prediction accuracy

- The accuracy of $\hat{Y}$ as prediction for *Y* depends on two quantities, which we will call the reducible error and the irreducible error

$$E\left(Y-\hat{Y}\right)^2 = E\left[f(X)+\epsilon-\hat{f}(X)\right]^2$$
$$= E\left[f(X)-\hat{f}(X)\right]^2 + Var(\epsilon)$$

- The left-hand-side represents the average, or expected value, of the squared difference between the predicted and actual value of *Y*
- The right-hand-side consists two terms: the ***reducible error*** and ***irreducible error***. The focus of this book is on techniques for estimating *f* with the aim of minimizing the reducible error

Goal 2: Inference

- We are often interested in understanding the way that *Y* is affected as $X_1, \cdots, X_p$ change. In this situation, we wish to estimate *f* as well, but $\hat{f}$ 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 explanatory variables are associated with the response?
  - What is the relationship between the response and each explanatory variable?
  - Can the relationship between *Y* and each explanatory variable be adequately summarized using a linear equation, or is the relationship more complicated?

Selecting statistical methods

- There are two common types of statistical inference methods
  - Confidence interval estimate
  - Hypothesis testing
- 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* allows for relatively simple and interpretable inference, but may not yield as accurate predictions as some other approaches; some of the highly *non-linear approaches* have high prediction accuracy, but at the expense of being less interpretable

How do we estimate *f* ?

- Throughout this course, we explore many linear and non-linear approaches for estimating *f*
- We will always assume that we have observed a set of *n* different data points. These observations are called the *training data*, because we will use these observations to train, or teach, our method how to estimate *f*
- Let x_{ij} represent the value of the *j*th predictor for observation *i*, where $i = 1, \cdots, n$ and $j = 1, \cdots, p$. Correspondingly, let y_i represent the response variable for the observation *i*. Then our training data consist of $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$, where $x_i = (x_{i1}, \cdots, x_{ip})$
- Our goal is to apply a statistical learning method to the training data in order to estimate the unknown function *f*

Parametric methods

- Parametric methods involve a two-step model-based approach
  - First we make an assumption about the function form, or shape, of *f.* For example,

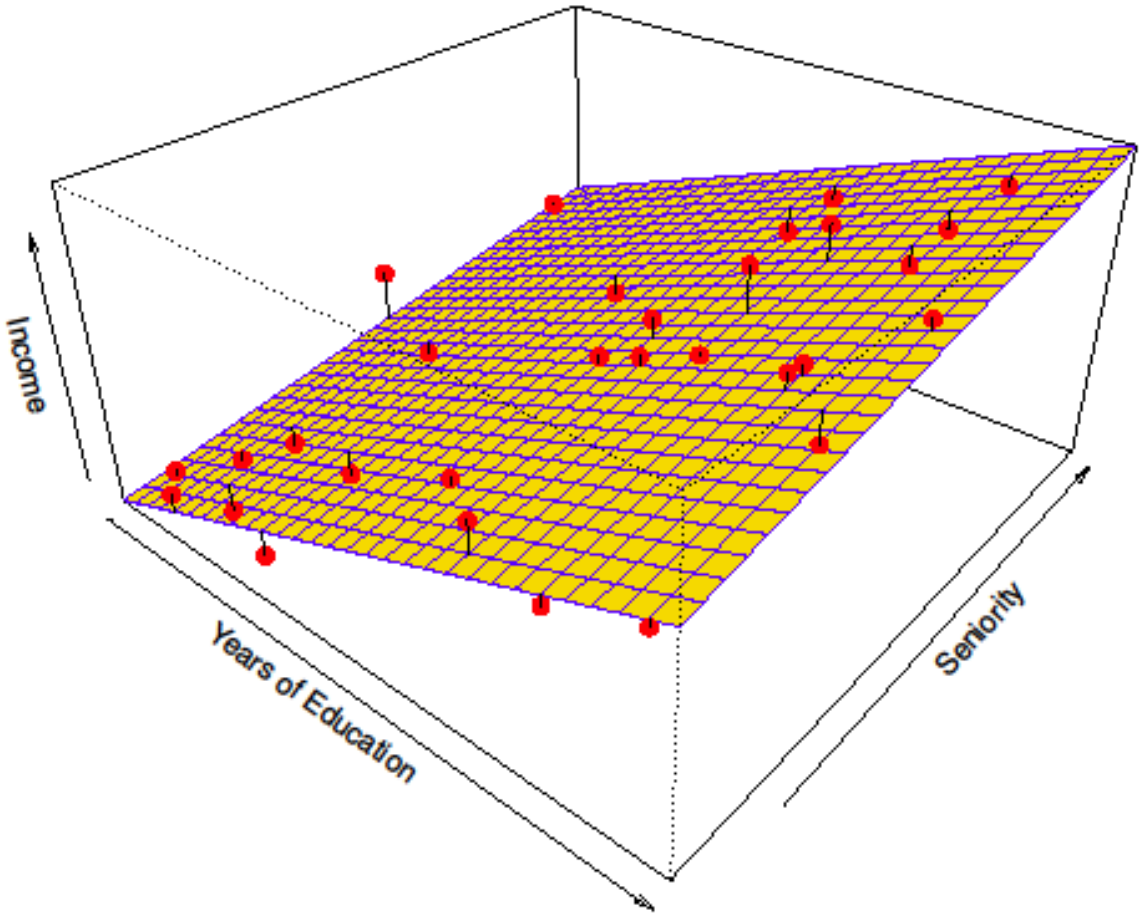
$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

  - After a model has been selected, we need a procedure that uses the training data to fit or train the model. For example, we use the *least squares* algorithm to estimate the coefficients

$$\hat{f}(X) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$$

- The model-based approach is referred as parametric. It reduces the problem of estimating *f* down to one of estimating a set of unknown parameters. The potential disadvantage of a parametric approach is that the model we choose will usually not match the true unknown form *f*

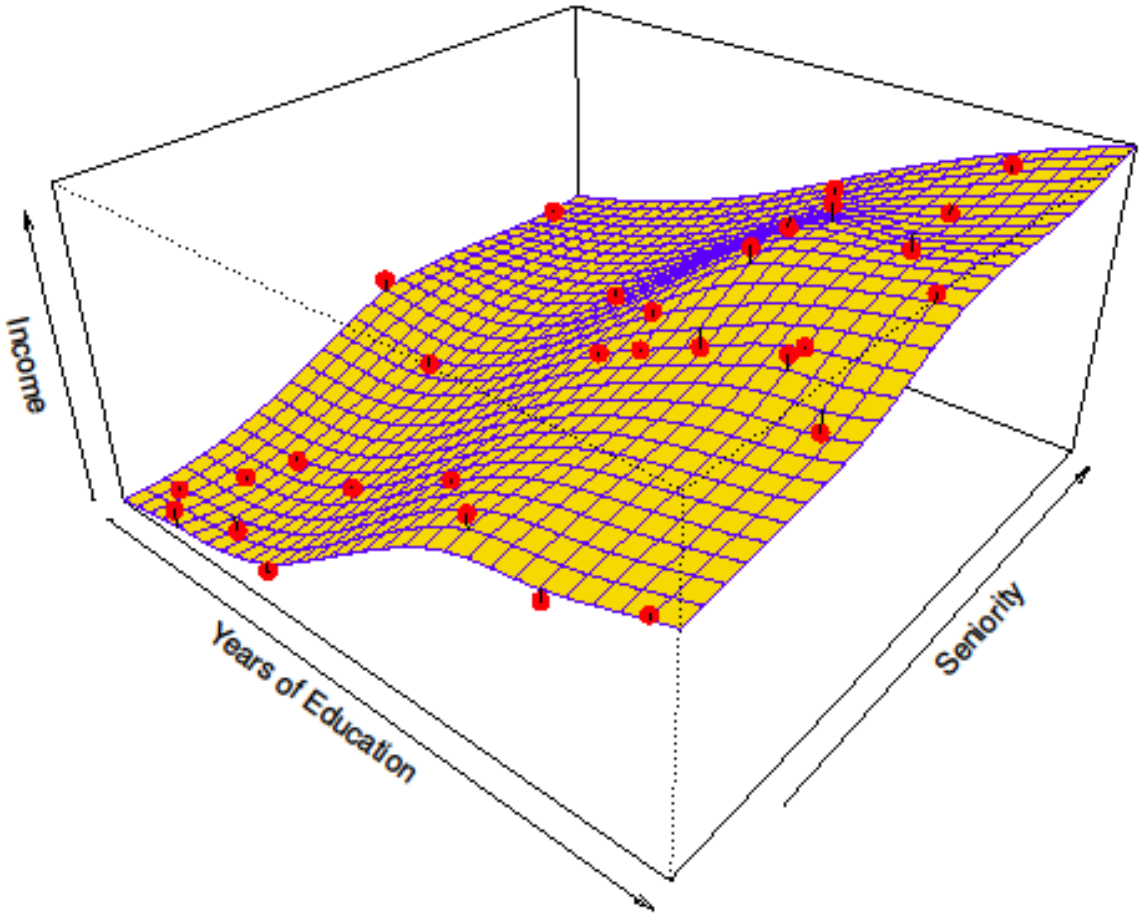
A linear model fit to the income data



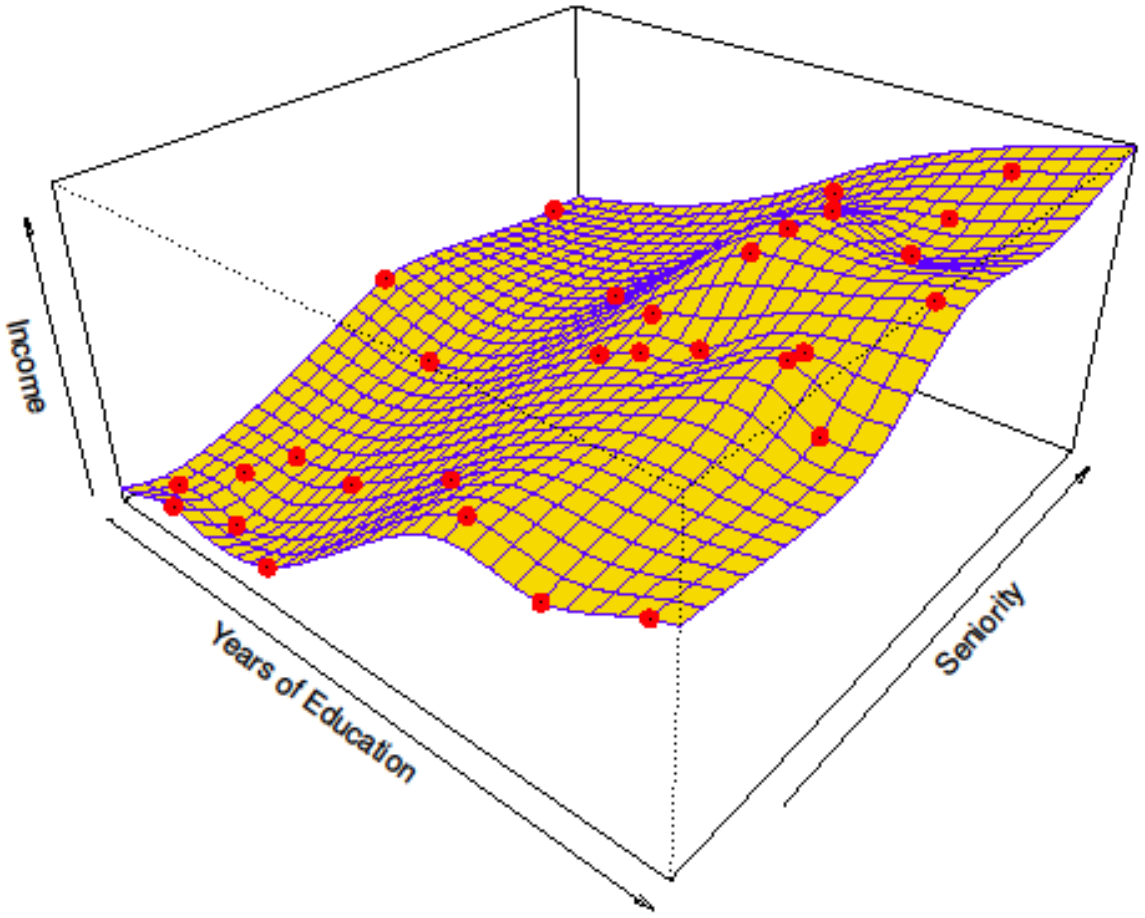
Non-parametric methods

- Non-parametric methods do not make explicit assumptions about the functional form of *f*. Instead they seek an estimate of *f* that gets as close to the data points as possible without being too rough or wiggly
- Such approaches can have a major advantage over parametric approaches: by avoiding the assumption of a particular functional form for *f*, they have the potential to accurately fit a wider range of possible shapes for *f*
- 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 is required in order to obtain an accurate estimate for *f*

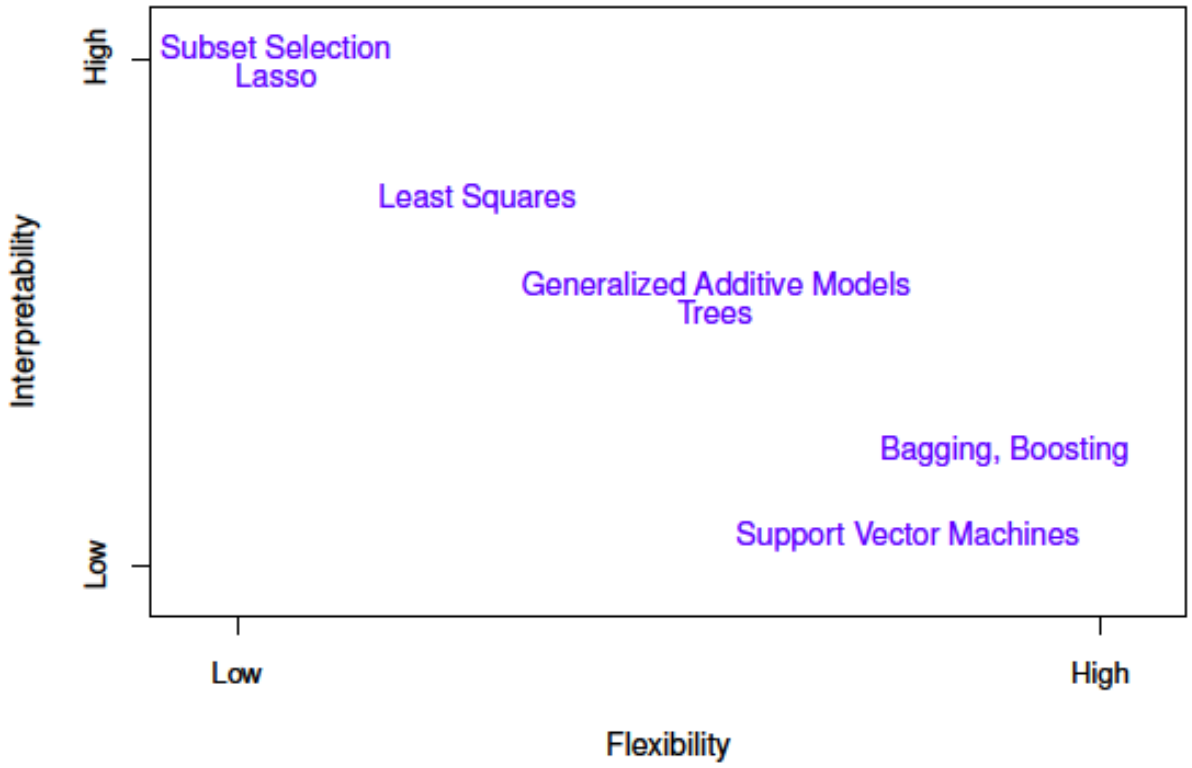
A smooth thin-plate spline fit to the income data



A rough thin-plane spline fit to the income data



The trade-off between prediction accuracy and model interpretability



Statistical learning methods

- Most statistical learning problems fall into one of two categories: *supervised learning* or *unsupervised learning*
  - Supervised learning: for each observation of the predictor measurements x_i, there is an associated response measurement $y_i, i = 1, \cdots, n$
  - Unsupervised learning: for every observation $i = 1, \cdots, n$, we observe a vector of measurements x_i, but no associated response y_i
- Variables can be characterized as either quantitative or categorical. We tend to refer to problems with a quantitative response as *regression problem*, while those involving a categorical response as *classification problem.* But sometimes there is overlap. For example, logistic regression is for binary response, and it is actually classification; some statistical methods, such as k-nearest neighbors and boosting, can be used in the case of either quantitative responses or categorical responses

Part (b): Assessing model accuracy

- This course will introduce many statistical learning methods that extend far beyond the standard linear regression approach
- No one method dominates all others over all possible datasets. On a particular dataset, one specific method work best, but some other method may work better on a similar but different dataset
- Hence it is an important task to decide for any given dataset 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

- In order to evaluate the performance of a statistical learning method on a given dataset, we need some way to measure how well its predictions actually match the observed data
- In the regression setting, the most commonly-used measured is the **mean squared error** (MSE) – actually it will be called **training MSE** in the next slide:

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

- Here $\hat{f}(x_i)$ is the prediction that $\hat{f}$ gives for the *i*th observation, and $\hat{f}$ is trained based on training data $\{(x_i, y_i), i = 1, \cdots, n\}$

Training MSE

- The MSE in the previously slide will be small if the predicted response are very close to the true responses, for training data
- This MSE is computed using the training data that was use to fit the model, and so should be referred to as ***the training MSE***
- But in general, we do not really care how well the method works on the training data. Rather, *we are interested in the accuracy of the predictions that we obtain when we apply our method to previously unseen test data*

Two examples

- Suppose we are interested in developing an algorithm to predict a stock's price based on previous stock returns. We train the method using stock returns from the past 6 months. But we don't really care how well our method predict last week's stock price. We instead care about how well it will predict tomorrow's price or next month's price
- Suppose we are interested in developing an algorithm to predict a patient's risk of diabetes based on clinical measurements (e.g. weight, blood pressure, family history). We are not very interested in whether or not the method accurately predicts diabetes risk for patients used to train the model, since we already how which of those patients have diabetes. We instead care about how likely some new patients will develop diabetes

Test MSE

- Suppose we fit our statistical learning method on our training data $\{(x_i, y_i), i = 1, \cdots, n\}$ and we obtain the estimate $\hat{f}$
- We want to know how close the prediction $\hat{f}(x_0)$ is to the true response 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***,

$$Ave\left(y_0 - \hat{f}(x_0)\right)^2$$

- Here the average is based on a large number of test observations

How to estimate the test MSE

- In some settings, we may have a test data available; that is, we may have access to a set of observations $\{(x_i^0, y_i^0), i = 1, \dots N\}$ that were not used to train the statistical learning method. Then we can simply estimate the test MSE using the test observations

$$Test\ MSE = \frac{1}{N}\sum_{i=1}^{N}\left(y_i^0 - \hat{f}(x_i^0)\right)^2$$

- But what if no test observations are available? We will discuss some methods in Chapter 5
- The question is: if no test observations are available, can we simply selecting a statistical learning method that minimizes the training MSE? The answer is NO

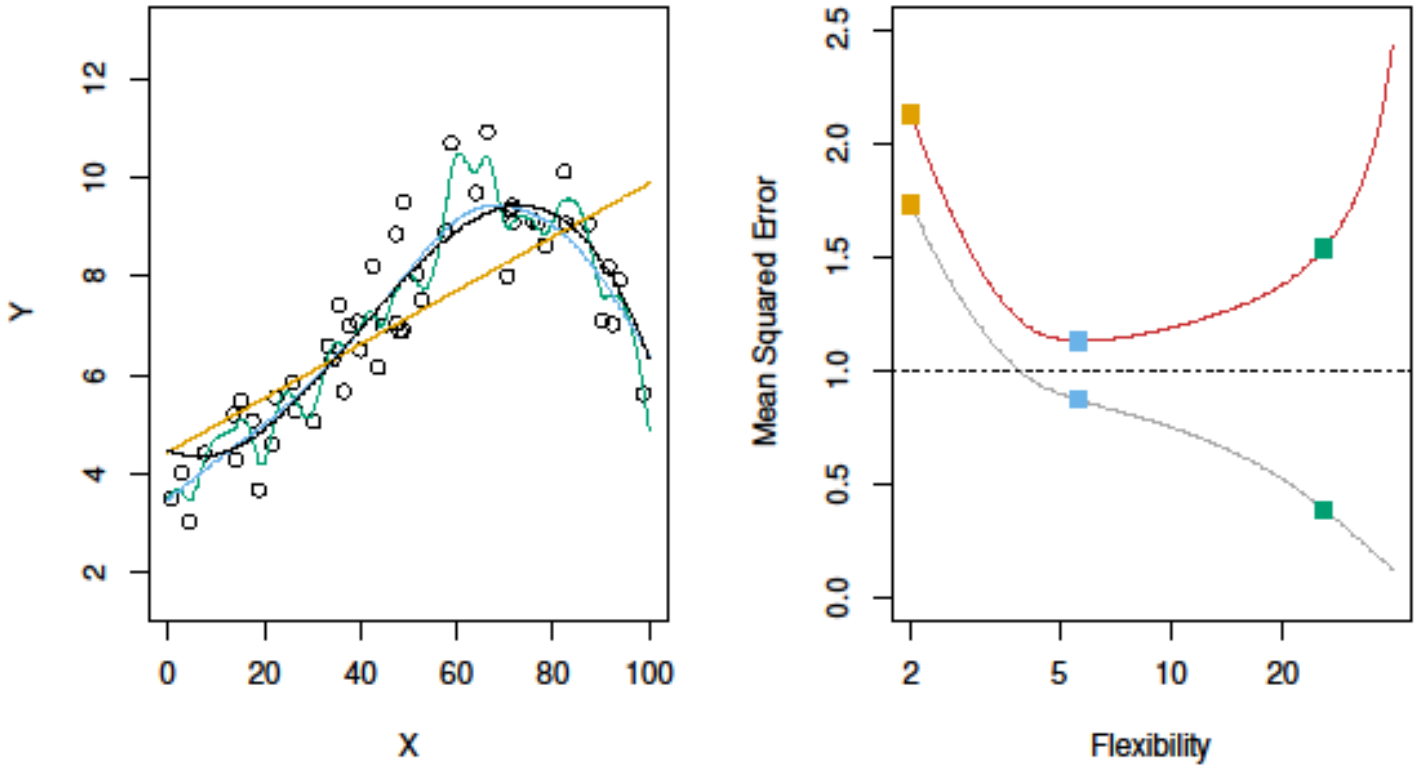
Training MSE vs. Test MSE

- Figure 2.9 illustrates an important phenomenon on a simple example
- In the left-hand panel, we have generated observations from the model, with the true f given by the black curve

$$Y = f(X) + \epsilon$$

- The orange, blue and green curves illustrate three possible estimates for f obtained using methods with increasing levels of flexibility
  - The orange line is the linear regression fit, which is relatively inflexible
  - The blue (appropriately flexible) and green (the most flexible) curves were produced using smoothing splines, discussed in Chapter 7, with different levels of smoothness

Figure 2.9


Y
X
0
20
40
60
80
100
2
4
6
8
10
12
Mean Squared Error
Flexibility
0.0
0.5
1.0
1.5
2.0
2.5
2
5
10
20

Training-MSE curve

- The grey curve on the right-hand panel displays the average training MSE as a function of flexibility, or more formally, the degrees of freedom, for a number of smoothing splines
- 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 it 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
- Therefore, if we use the training MSE to select a “best” method, we will always select the most flexible one

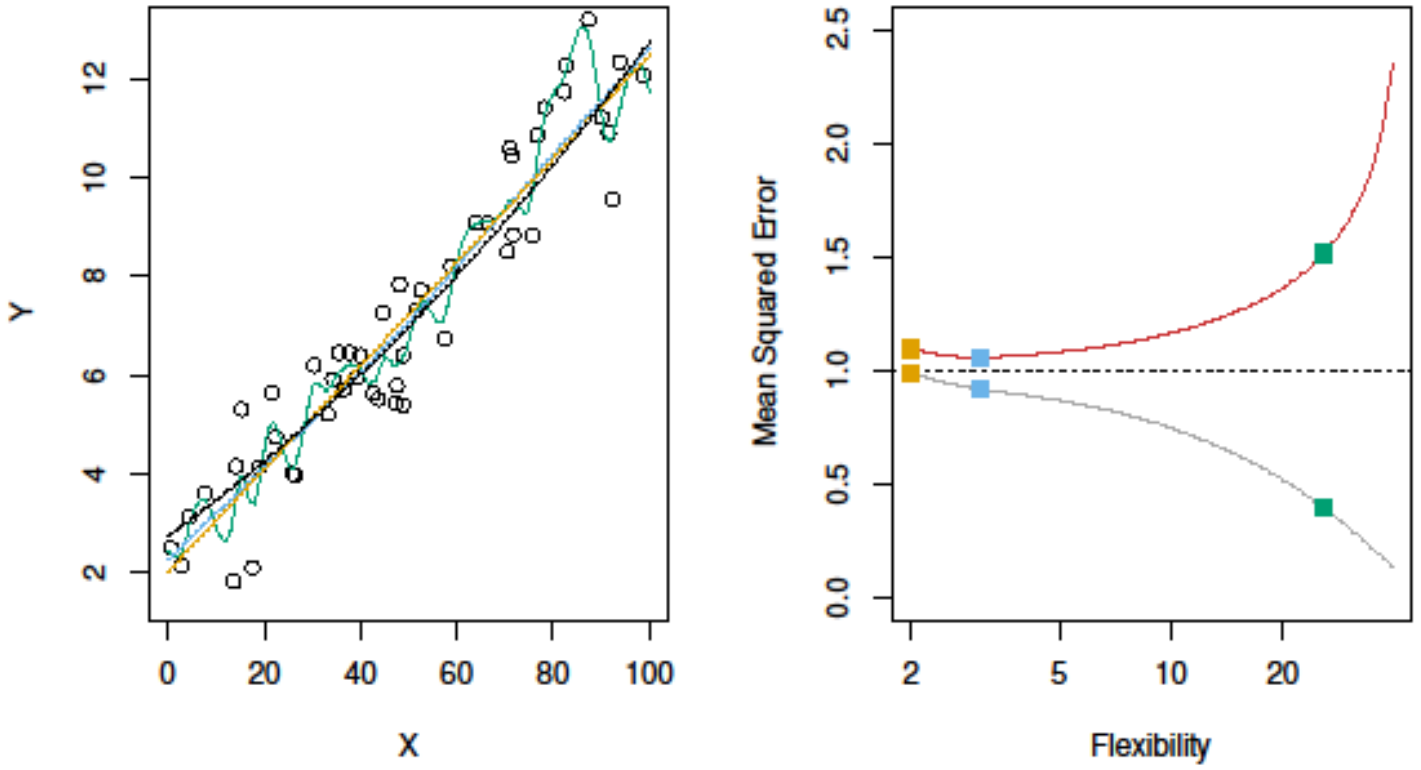
Test-MSE curve

- In this simulation example, we know the true f and therefore we can compute the test MSE on a very large test dataset, as a function of flexibility
- The red curve on the right-hand panel of Figure 2.9 displays the test-MSE curve
- The test MSE initially declines as the level of flexibility increases. However, at some point the test MSE levels off and then starts to increases again
- The blue curve minimizes the test MSE, which should not be surprising given that visually it appears to estimate f the best in the left-hand panel
- The horizontal dashed line indicates $Var(\epsilon)$, the irreducible error, the lowest achievable MSE among all methods

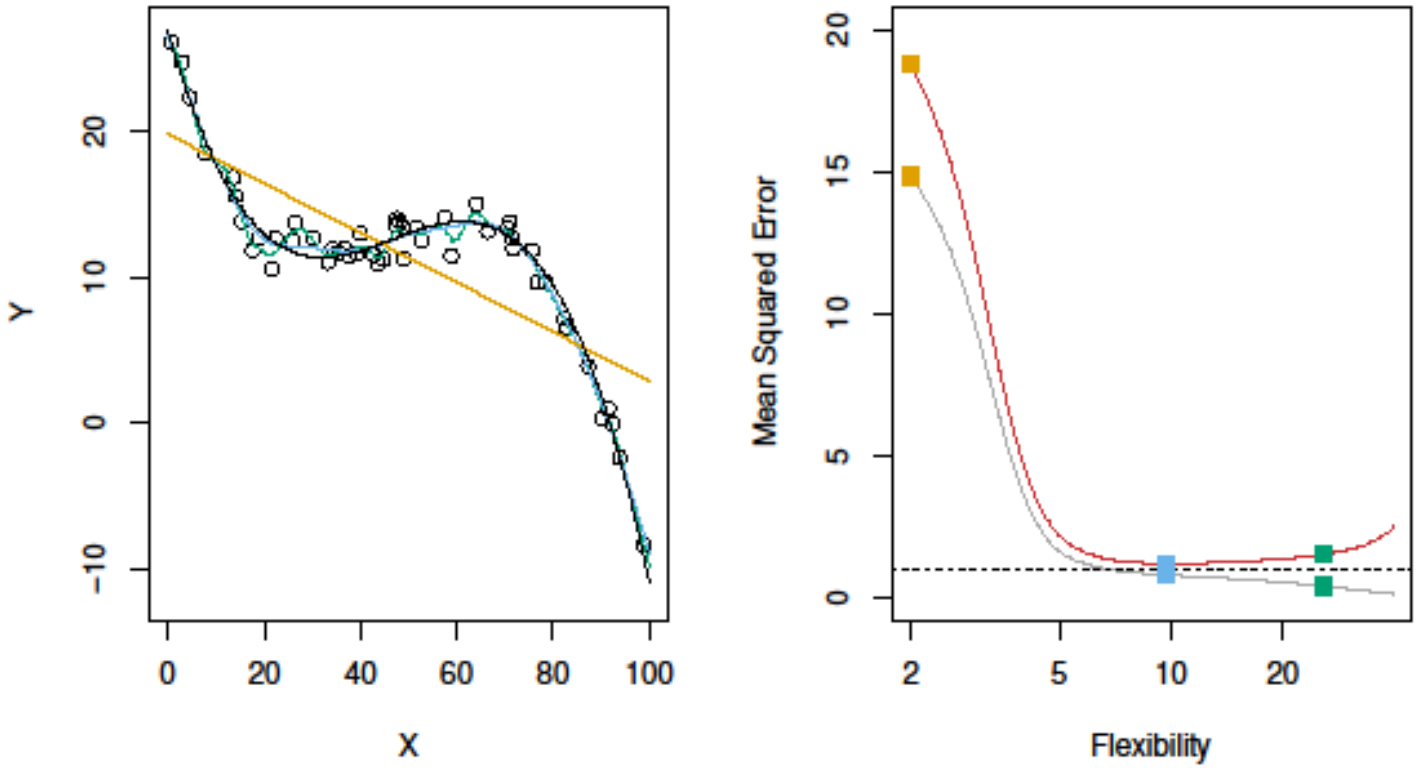
U-Shape

- As the flexibility of the statistical learning method increases, we observe a monotone decrease in the training MSE, while we observed an U-shape in the test MSE
- This is a fundamental property of statistical learning that holds regardless of the particular dataset at hand and regardless of the statistical method being used
- When a given method yields a small training MSE but a large test MSE, we are said to be ***overfitting*** the data. Overfitting refers specially to the case in which a less flexible model would have yielded a smaller test MSE

Another example: Figure 2.10 where the true *f* is much closer to linear



One More example: Figure 2.11 where the true *f* is far from linear



The bias-variance trade-off

- The U-shape observed in the test MSE curves turns out to be the results of two competing properties of statistical learning methods: bias and variance
- The expected test MSE at a given value x_0 can be decomposed into the sum of three fundamental quantities: the variance of $\hat{f}(x_0)$, the squared bias of $\hat{f}(x_0)$, and the variance of the error ϵ,

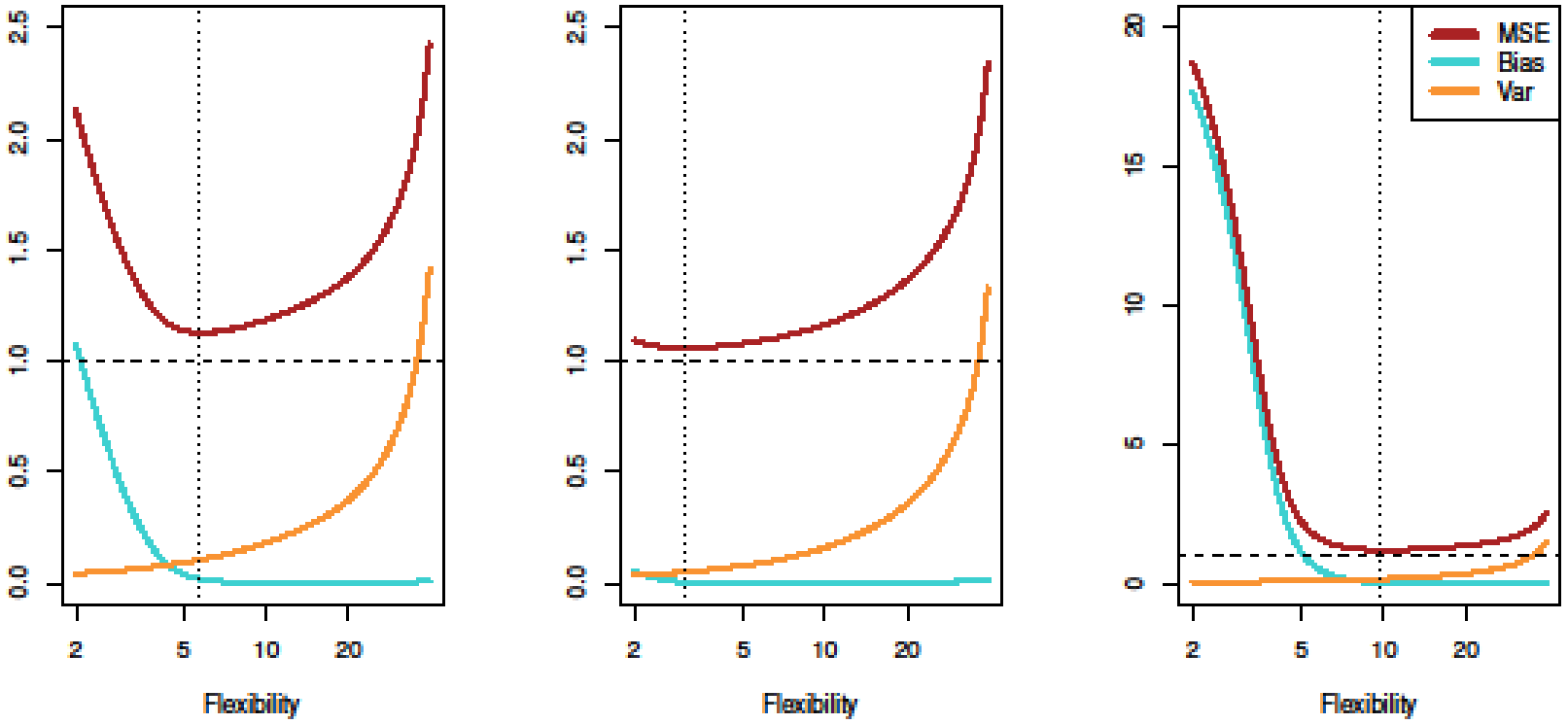
$$E\left(y_0 - \hat{f}(x_0)\right)^2 = E\left(f(x_0) - \hat{f}(x_0)\right)^2 + Var(\epsilon)$$
$$= Var\left(\hat{f}(x_0)\right) + \left[f(x_0) - E(\hat{f}(x_0))\right]^2 + Var(\epsilon)$$

- Therefore, in order to minimize the expected test error, we need to select a statistical learning method that simultaneously achieves low variance and low bias

Variance and Bias

- Variance refers to the amount by which $\hat{f}$ would change if we estimated it using a different training data set, noting that different training datasets will result in different $\hat{f}$. In general, more flexible statistical methods have higher variance
- Bias refers to the error that is introduced by approximating a real life problem, which may be extremely complicated, by a much simpler model. In general, more flexible statistical methods result in less bias
- As a general rule, as we use more flexible methods, the variance will increase and the bias will decrease. As we will increase the flexibility of a class of methods, the bias tends to initially decrease faster than the variance increases; therefore, the expected MSE declines initially. But at some point, increasing flexibility has litter impact on the bias but starts to significantly increase the variance; therefore, after some point, the expected MSE starts to increase

The bias-variance trade-off



Cross-validation

- 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 method
- In Chapter 5, we will discuss cross-validation, which is a way to estimate the test MSE using the test MSE