DATA 572: Supervised Learning

2023W2

Shan Du

Dr. Shan Du

- Ph.D. (The University of British Columbia Vancouver, 2009)
- Spin-off company in Computer Vision
- Associate Editor of IEEE Trans. on Circuits and Systems for Video Technology
- Area Chair and Session Chair of IEEE International Conference on Image Processing (ICIP)

Course Information

- **Instructor**: Dr. Shan Du FIP 324
  shan.du@ubc.ca
- **Department**: Computer Science
- **Class Time**: Monday and Wednesday 9:30AM – 11:00AM
- **Location**: EME 1153
- **Office Hour**: Friday 11:00AM – 12:00PM on Canvas
- **Course Website:**

  https://canvas.ubc.ca/courses/133403

Course Information

- **Teaching Assistant**:

  Yining Zhou, zhou258@student.ubc.ca

Course Information

- **Course Description**:
  - Analysis of data with categorical responses. Logistic regression, k-nearest-neighbours classification, discriminant analysis, decision trees and random forests. Restricted to students in the MDS program.
  - Prerequisite: DATA 571.

Course Information

- **Learning Materials:**
  - Lecture Notes and Tutorials (available electronically)
  - Recommended Textbooks :
    - An Introduction to Statistical Learning with Applications in Python (Chapter 1-9), Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Jonathan Taylor, 2023: https://www.statlearning.com/
    - Deep Learning (Chapter 5), Ian Goodfellow, Yoshua Bengio and Aaron Courville, MIT Press, 2016: https://www.deeplearningbook.org/

- **Programming Language:** Python

Course Information

Learning Outcomes:

Upon successful completion of this course, students will be able to:

- Demonstrate a theoretical and practical understanding of basic supervised machine learning models
- Develop models to solve classification problems
- Apply existing machine learning packages to develop classification models
- Understand how to quantify performance of machine learning models

Course Information

Grading:

Programming Assignments: 40%

Project: 60%

Regulations

- Students MUST achieve a passing grade in overall in order to pass the course.
- The course website contains the most up-to-date information and important dates for main events such as assignments and project due dates. Students need to check regularly.
- Attendance is mandatory in lectures and labs.
- **The use of artificial intelligence (AI) assistance for any assessed portions of this course is not permitted.**
- If you feel any mark was unfair or incorrectly recorded, ensure that I am aware of the problem before the last week of classes.

Regulations

- **Late Penalty:** Late assignments and report will be deducted 10% per day up to 3 days (after which they will receive 0 marks).

- **Plagiarism:** is forbidden (0 mark).

Course Contents and Schedule

Week	Contents
1	Introduction to Supervised Learning, Recap of Basic Concepts, Logistic Regression
2	Discriminant Analysis, Naive Bayes, K-Nearest Neighbors Classifier
3	Moving Beyond Linearity, Classification Trees and Random Forests, Boosting
4	Support Vector Machines
5	Projects and Presentations

Acknowledgment

- The course materials are based on:
  - The recommended textbooks/references
  - Some online resources and similar courses offered in other top-ranked universities
  - Some published papers and datasets

An Overview of Statistical Learning

- Statistical learning refers to a vast set of tools for *understanding data*.
- These tools can be classified as *supervised* or *unsupervised*.
- Broadly speaking, supervised statistical learning involves building a statistical model to associate some *input* with some *output*, given a training set of examples of inputs x and outputs y.
- With unsupervised statistical learning, there are inputs but *no supervising output*; nevertheless we can learn relationships and structure from such data.

An Example of Wage Data

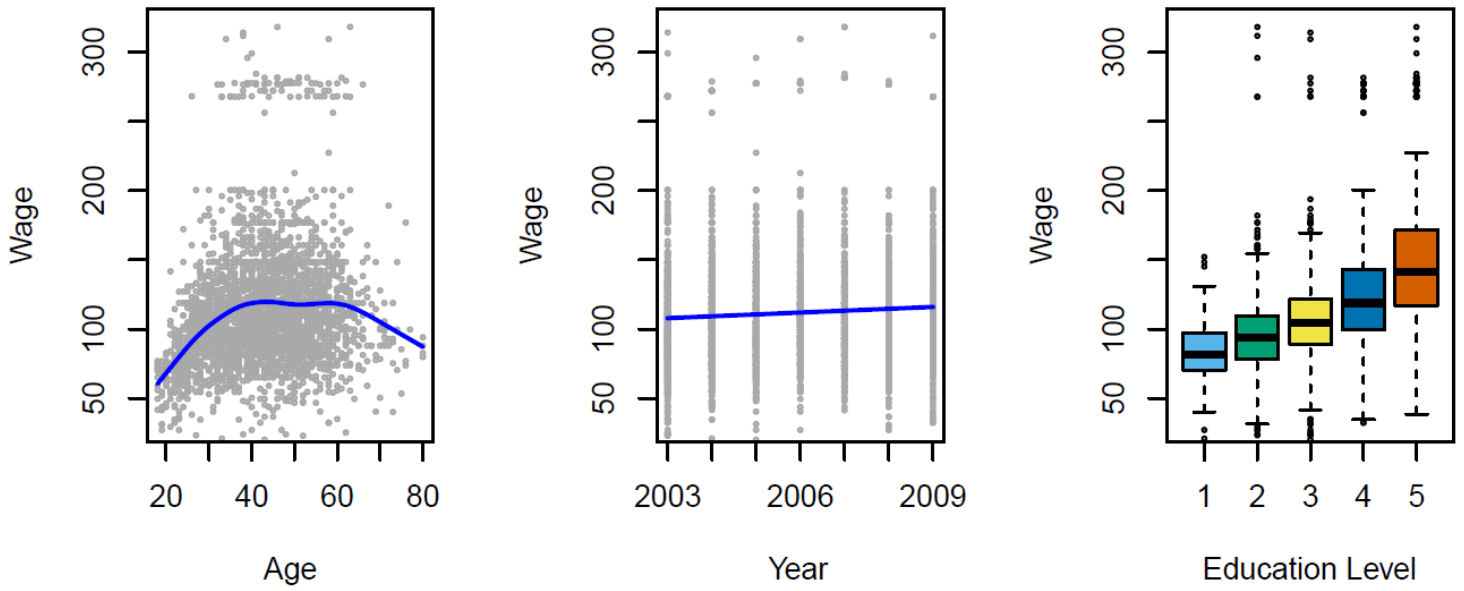


FIGURE 1.1. `Wage` *data, which contains income survey information for men from the central Atlantic region of the United States.* Left: `wage` *as a function of* `age`. *On average,* `wage` *increases with* `age` *until about* 60 *years of age, at which point it begins to decline.* Center: `wage` *as a function of* `year`. *There is a slow but steady increase of approximately* $10,000 *in the average* `wage` *between* 2003 *and* 2009. Right: *Boxplots displaying* `wage` *as a function of* `education`, *with* 1 *indicating the lowest level (no high school diploma) and* 5 *the highest level (an advanced graduate degree). On average,* `wage` *increases with the level of education.*

An Example of Wage Data

- Clearly, the most accurate prediction of a given man's wage will be obtained by combining his *age*, his *education*, and the *year*.
- The *Wage* data involves predicting a continuous or quantitative output value. This is often referred to as a *regression* problem.

An Example of Stock Market Data

- In certain cases, we may instead wish to predict a non-numerical value—that is, a *categorical* or *qualitative* output.
- For example, we examine a stock market data set that contains the daily movements in the Standard & Poor's 500 (S&P) stock index over a 5-year period between 2001 and 2005, the *Smarket* data, to predict whether the index will *increase* or *decrease* on a given day, using the past 5 days' percentage changes in the index.
- This is known as a *classification* problem.

An Example of Stock Market Data

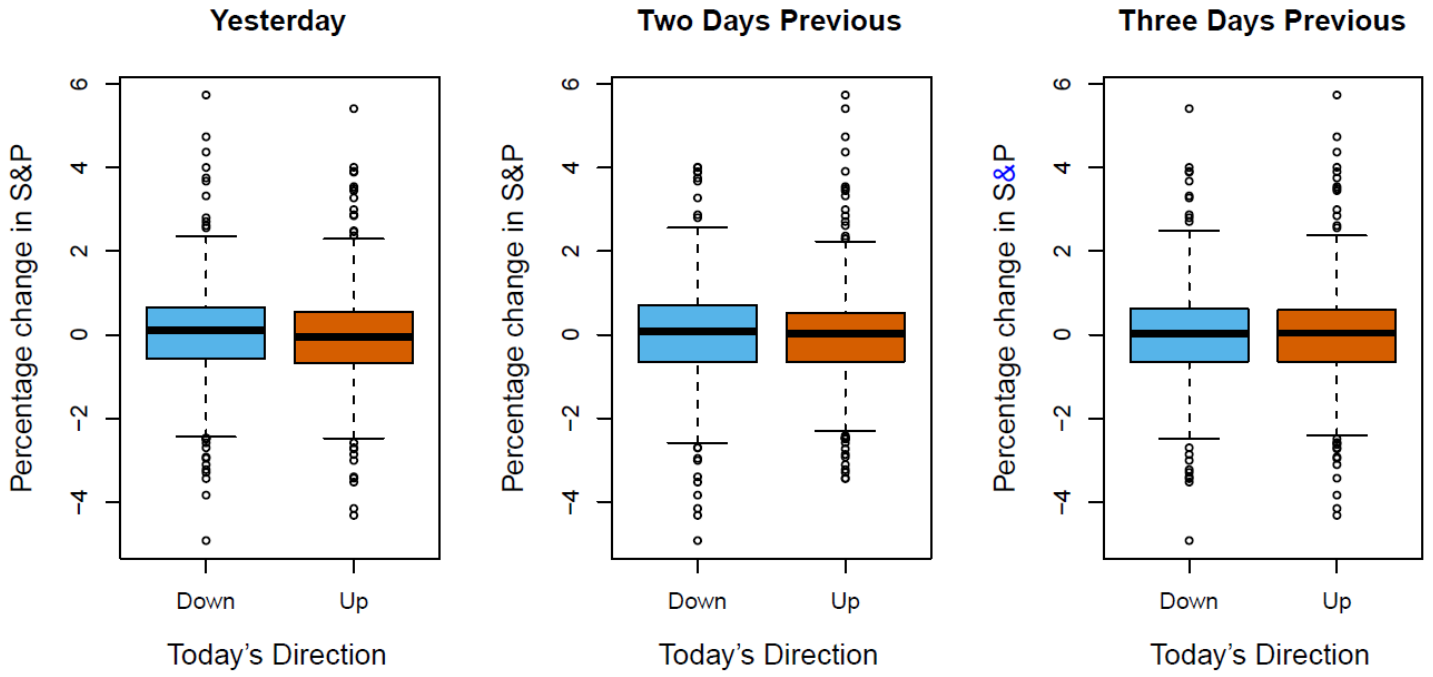


FIGURE 1.2. Left: *Boxplots of the previous day's percentage change in the S&P index for the days for which the market increased or decreased, obtained from the* `Smarket` *data.* Center and Right: *Same as left panel, but the percentage changes for 2 and 3 days previous are shown.*

A Brief History of Statistical Learning

- At the beginning of the nineteenth century, the method of *least squares* was developed, implementing the earliest form of what is now known as *linear regression*. Linear regression is used for predicting quantitative values.
- In order to predict qualitative values, such as whether a patient survives or dies, or whether the stock market increases or decreases, *linear discriminant analysis* was proposed in 1936.

A Brief History of Statistical Learning

- In the 1940s, various authors put forth an alternative approach, *logistic regression*.
- In the early 1970s, the term *generalized linear model* was developed to describe an entire class of statistical learning methods that include both linear and logistic regression as special cases.

A Brief History of Statistical Learning

- By the end of the 1970s, many more techniques for learning from data were available. However, they were almost exclusively linear methods because fitting non-linear relationships was computationally difficult at the time.
- By the 1980s, computing technology had finally improved sufficiently that non-linear methods were no longer computationally prohibitive.
- In the mid 1980s, classification and regression trees were developed, followed shortly by generalized additive models.
- Neural networks gained popularity in the 1980s, and support vector machines arose in the 1990s.

What is Statistical Learning?

- A simple example:
  - Suppose we are statistical consultants hired by a client to investigate the association between advertising and sales of a particular product.
  - The *Advertising* data set consists of the *sales* of that product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: *TV*, *radio*, and *newspaper*.

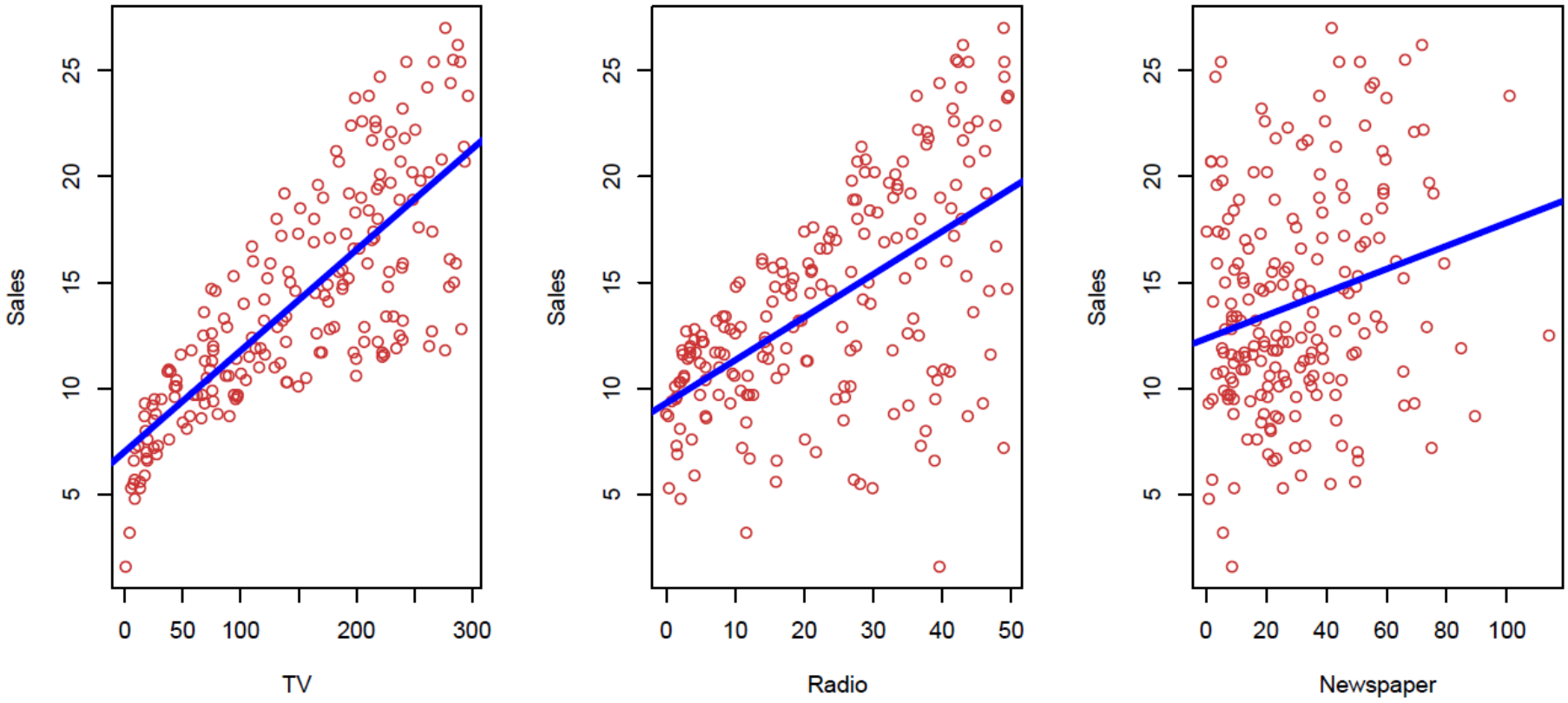


FIGURE 2.1. *The* `Advertising` *data set. The plot displays* `sales`*, in thousands of units, as a function of* `TV`*,* `radio`*, and* `newspaper` *budgets, in thousands of dollars, for* 200 *different markets. In each plot we show the simple least squares fit of* `sales` *to that variable, as described in Chapter* 3*. In other words, each blue line represents a simple model that can be used to predict* `sales` *using* `TV`*,* `radio`*, and* `newspaper`*, respectively.*

What is Statistical Learning?

- A simple example:
  - If we determine that there is an association between advertising and sales, then we can instruct our client to adjust advertising budgets, thereby indirectly increasing 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.
  - In this setting, the advertising budgets are *input variables* while sales is an *output variable*.

What is Statistical Learning?

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

What is Statistical Learning?

- A simple example:
  - Suppose we observe a quantitative response Y and p different predictors, $X_1, X_2, \ldots, X_p$. We assume that there is some relationship between Y and $X_1, X_2, \ldots, X_p$, which can be written in the very general form

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

  - Here f is some fixed but unknown function of $X_1, X_2, \ldots, X_p$, and ϵ is a random *error term*, which is independent of X and has mean zero. In this, f represents the *systematic* information that X provides about Y.

What is Statistical Learning?

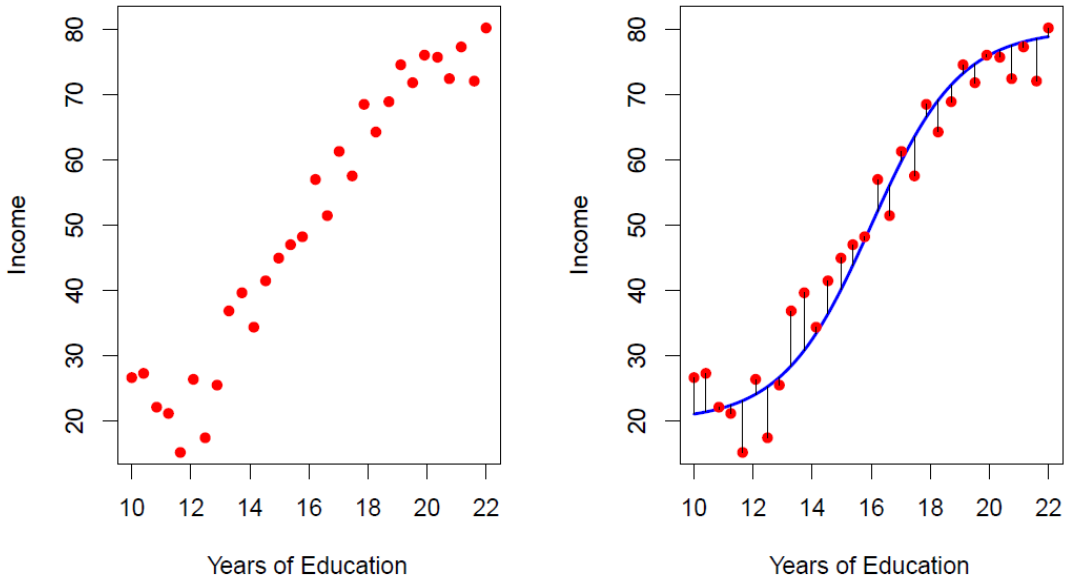


FIGURE 2.2. *The* `Income` *data set.* Left: *The red dots are the observed values of* `income` *(in thousands of dollars) and* `years of education` *for* 30 *individuals.* Right: *The blue curve represents the true underlying relationship between* `income` *and* `years of education`*, which is generally unknown (but is known in this case because the data were simulated). The black lines represent the error associated with each observation. Note that some errors are positive (if an observation lies above the blue curve) and some are negative (if an observation lies below the curve). Overall, these errors have approximately mean zero.*

What is Statistical Learning?

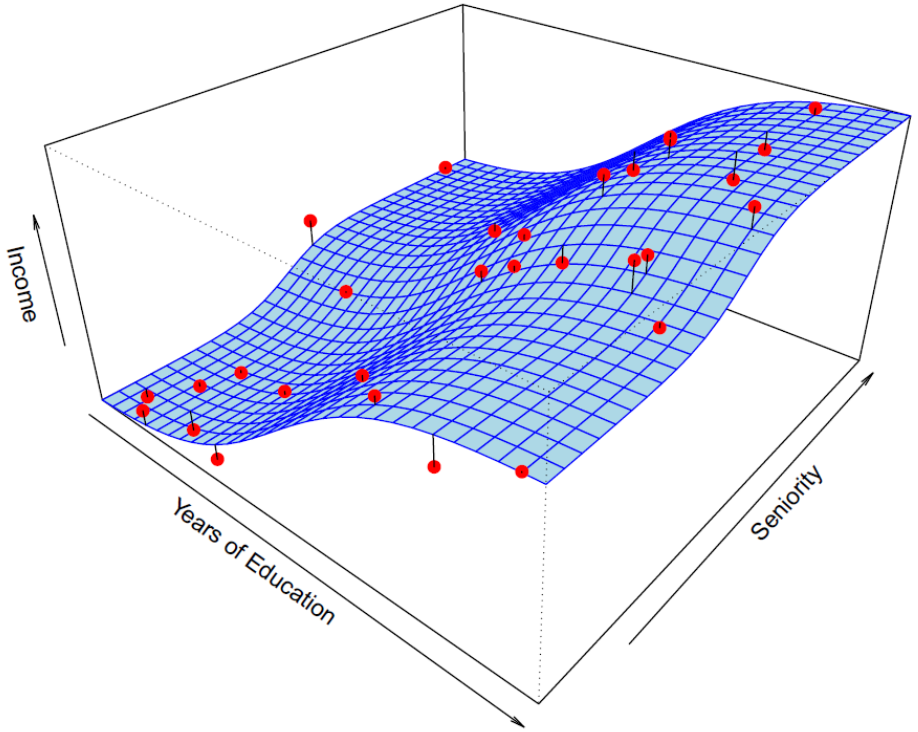


FIGURE 2.3. *The plot displays* `income` *as a function of* `years of education` *and* `seniority` *in the* `Income` *data set. The blue surface represents the true underlying relationship between* `income` *and* `years of education` *and* `seniority`*, which is known since the data are simulated. The red dots indicate the observed values of these quantities for* 30 *individuals.*

What is Statistical Learning?

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

Why Estimate f?

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

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

    where $\hat{f}$ represents our estimate for f, and $\hat{Y}$ represents the resulting prediction for Y.

Why Estimate f?

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

- In general, $\hat{f}$ 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 $\hat{f}$ by using the most appropriate statistical learning technique to estimate f.

Why Estimate f?

- However, even if it were possible to form a perfect estimate for f, so that our estimated response took the form $\hat{Y} = f(X)$, our prediction would still have some error in it!
- This is because Y is also a function of ϵ, which, by definition, cannot be predicted using X. Therefore, variability associated with ϵ 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 ϵ.

Why Estimate f?

– Inference

We are often interested in understanding the association between Y and $X_1, X_2, \ldots, X_p$. In this situation we wish to estimate f, but our goal is not necessarily to make predictions for Y. 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?*

  It is often the case that only a small fraction of the available predictors are substantially associated with Y. Identifying the few important predictors among a large set of possible variables can be extremely useful, depending on the application.

Why Estimate f?

- *What is the relationship between the response and each predictor?*

Some predictors may have a positive relationship with Y, in the sense that larger values of the predictor are associated with larger values of Y. Other predictors may have the opposite relationship. Depending on the complexity of f, the relationship between the response and a given predictor may also depend on the values of the other predictors.

- *Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?*

How Do We Estimate 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 training observations to train, or teach, our method how to estimate f.
- Our goal is to apply a statistical learning method to the training data in order to estimate the unknown function f.
- Broadly speaking, most statistical learning methods for this task can be characterized as either *parametric* or *non-parametric*.

How Do We Estimate f?

- Parametric Methods

  Parametric methods involve a two-step model-based approach.

  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:

  $$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

  2. After a model has been selected, we need a procedure that uses the training data to fit or train the model. We need to estimate the parameters $\beta_0, \beta_1, \ldots, \beta_p$. That is,

How Do We Estimate f?

we find values of these parameters such that

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

The potential disadvantage of a parametric approach is that 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 flexible 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 follow the errors, or noise, too closely.

How Do We Estimate f?

- 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.

How Do We Estimate f?

- Any parametric approach brings with it the possibility that the functional form used to estimate f is very different from the true f, in which case the resulting model will not fit the data well. In contrast, non-parametric approaches completely avoid this danger, since essentially no assumption about the 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.

How Do We Estimate f?

- The Trade-Off Between Prediction Accuracy and Model Interpretability
  - In general, as the flexibility of a method increases, its interpretability decreases.
  - When inference is the goal, there are clear advantages to using simple and relatively inflexible statistical learning methods. 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.
  - However, we have to deal with the potential for overfitting in highly flexible methods.

How Do We Estimate f?

- Supervised Versus Unsupervised Learning
  - Most statistical learning problems fall into one of two categories: supervised or unsupervised.
  - For each observation of the predictor measurement(s) $x_i, i = 1, \dots, n$ there is an associated response measurement y_i. 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).

How Do We Estimate f?

- By contrast, unsupervised learning describes the somewhat more challenging situation in which for every observation $i = 1, \ldots, n$, we observe a vector of measurements x_i but no associated response y_i.
- 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.

How Do We Estimate f?

– What sort of statistical analysis is possible? We can seek to understand the relationships between the variables or between the observations. One statistical learning tool that we may use in this setting is cluster analysis, or clustering. The goal of cluster analysis cluster is to ascertain, on the basis of $x_1, \ldots, x_n$ analysis , whether the observations fall into relatively distinct groups.

How Do We Estimate f?

- Regression Versus Classification Problems
  - Variables can be characterized as either quantitative or qualitative (also known as categorical).
  - Quantitative variables take on numerical values.
  - In contrast, qualitative variables take on values in one of K different classes, or categories.
  - 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.

Assessing Model Accuracy

- 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

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

where $\hat{f}(x_i)$ is the prediction that $\hat{f}$ gives for the ith observation.

Assessing Model Accuracy

- Measuring the Quality of Fit
  - Training MSE: computed using the training data that was used to fit the model
  - Test MSE: the accuracy of the predictions that we obtain when we apply our method to previously unseen test data
  - We want to choose the method that gives the *lowest test MSE*.

Assessing Model Accuracy

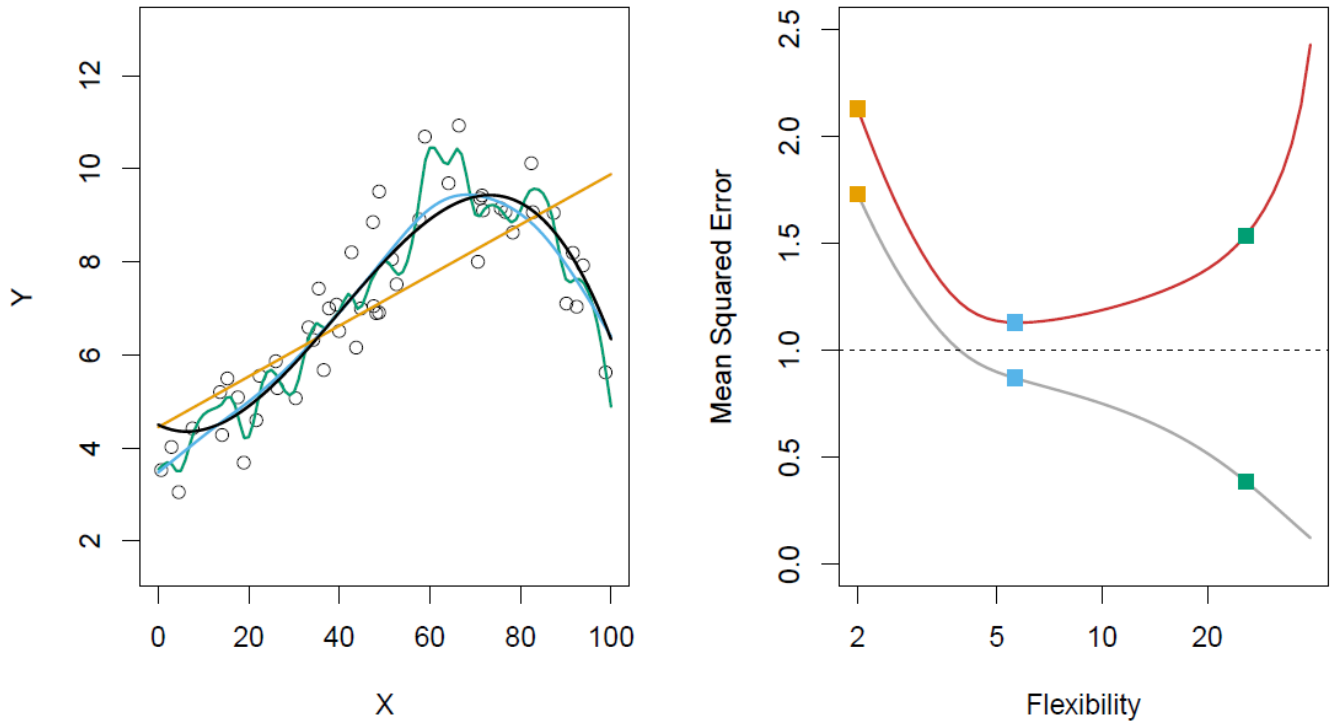
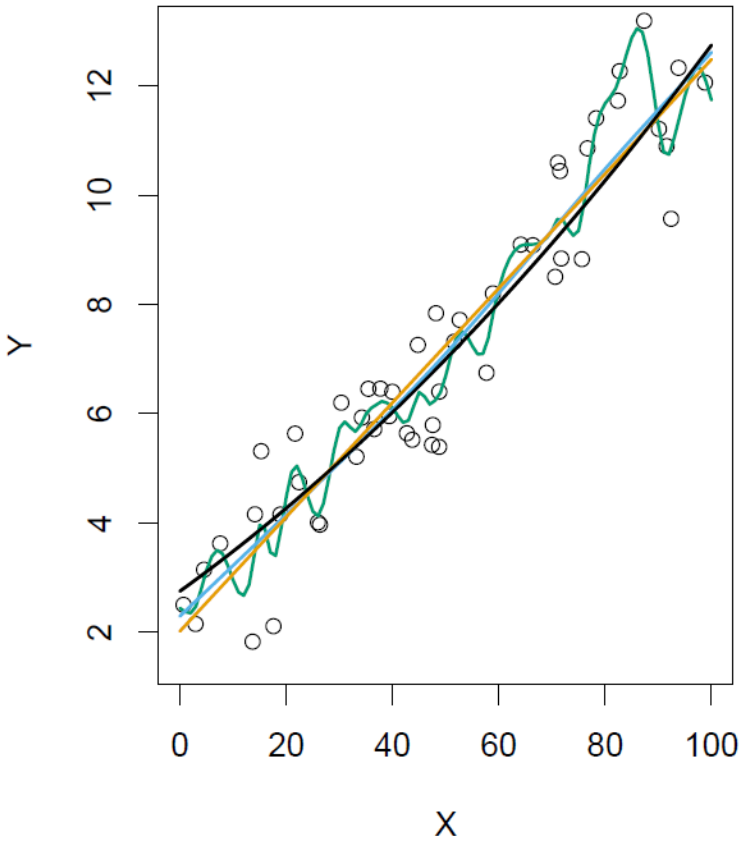
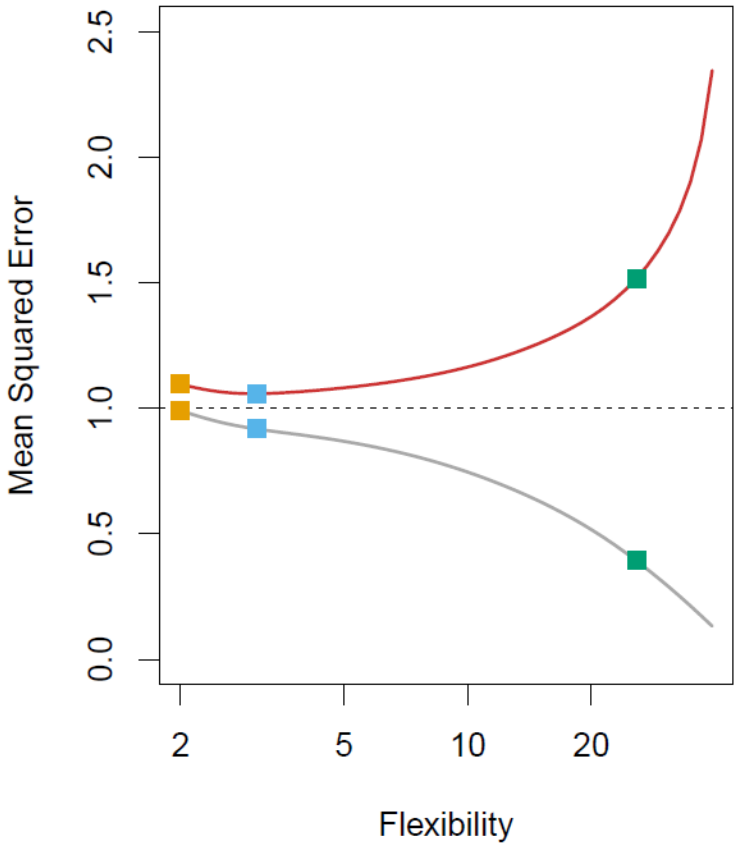


FIGURE 2.9. Left: *Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves).* Right: *Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.*

Assessing Model Accuracy





Assessing Model Accuracy

- The Bias-Variance Trade-Off

The expected test MSE, for a given value x_0, can always 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 terms ϵ.

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

Assessing Model Accuracy

- The Bias-Variance Trade-Off

We need to select a statistical learning method that simultaneously achieves low variance and low bias.

Assessing Model Accuracy

- The Classification Setting

Suppose that we seek to estimate f on the basis of training observations $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, where now $y_1, \ldots, y_n$ are qualitative. The most common approach for quantifying the accuracy of our estimate $\hat{f}$ is the training *error rate*

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

Assessing Model Accuracy

- The Classification Setting

The *test error* is associated with a set of test observations in a similar way.

A good classifier is one for which the test error is smallest.