Topic 1. Statistical Learning

Statistical Learning vs. Machine Learning

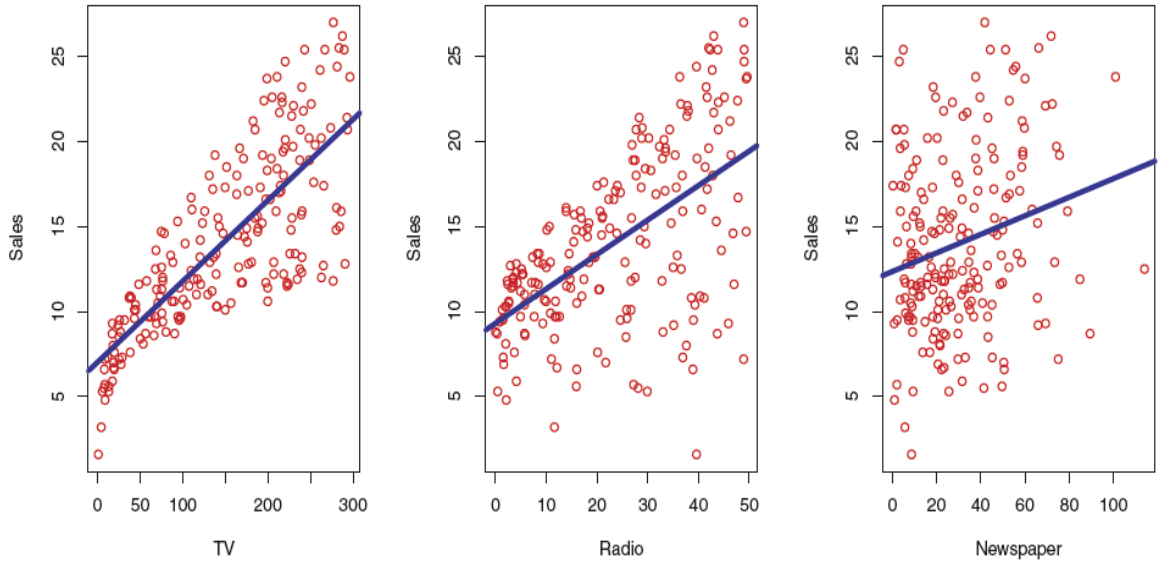
- Statistics vs. Machine learning
  - ---- Modeling/assumptions vs. learning from data
  - ---- Inference vs. prediction
  - ---- Smaller dataset vs. big data
- Statistical learning and machine learning are broadly the same
- Statistical learning: prediction in statistical views

Outline

- Motivating example
- Statistical learning
- Supervised learning
- Reasons for learning
- Restrictive vs. flexible methods
- Comparing different methods
- Fundamental understanding

Motivating Example

- Suppose we are statistical consultants hired by a client to provide advice on how to improve sales of a product
- Advertising dataset: sales as a function of TV, radio, and newspaper budges for 200 markets



- If we determine that there is an association between advertising and sales, then we can instruct our client to adjust advertising budgets, thereby increasing sales.
- Our goal is to **develop an accurate model that can be used to predict sales on the basis of the three media budgets**.

Variables

Statistics	Dependent variable Response	Independent variables predictors
Machine learning	Output (variable)	Inputs
Pattern recognition	Output	Features

$$Y \longleftarrow X$$

$Y: sales$

$X_1: TV\ budget$
$X_2: radio\ budget$
$X_3: newspaper\ budget$

Statistical Learning

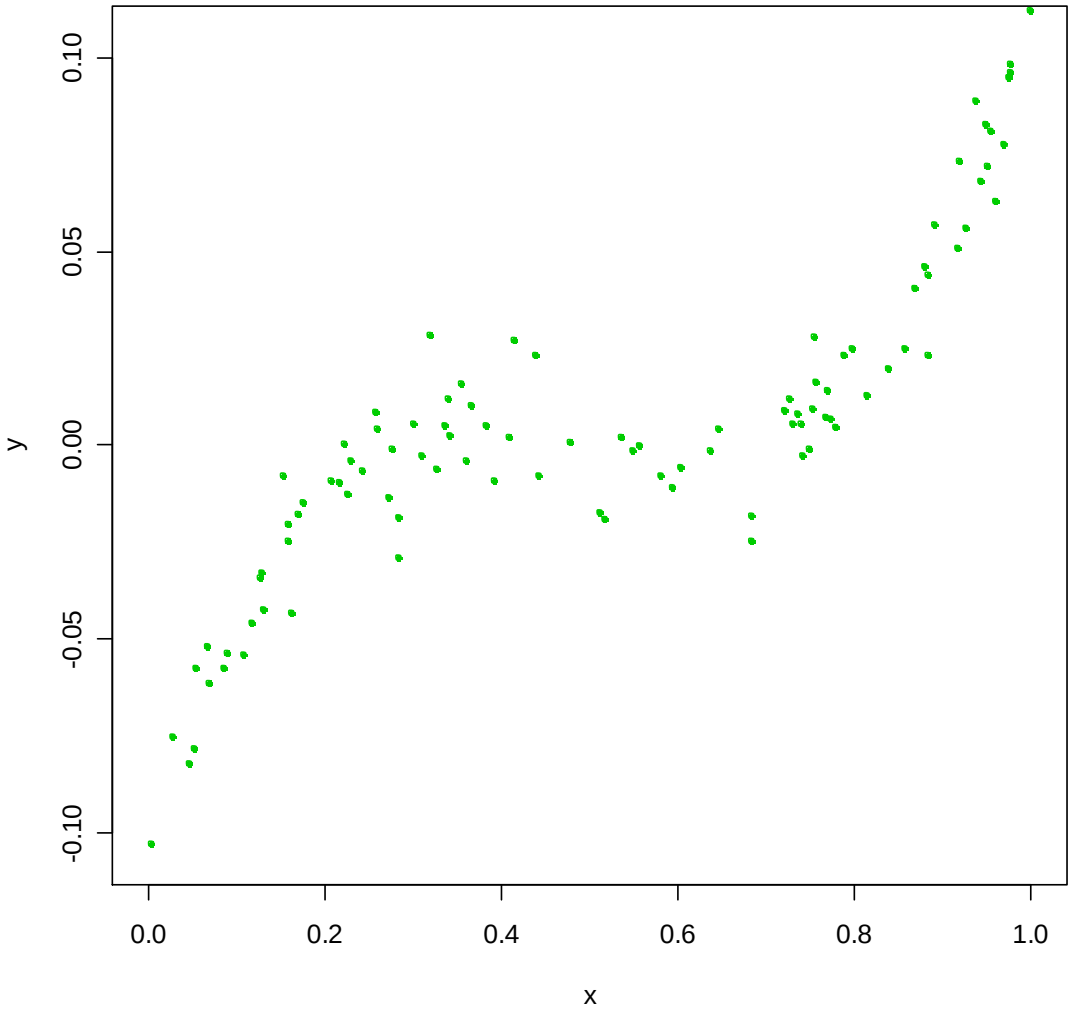
- Suppose we observe a quantitative response/output Y and p different predictors/inputs $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$
- We believe that there is a relationship between Y and at least one of the Xs.
- We can model the relationship as

$$Y = f(\boldsymbol{X}) + \varepsilon$$

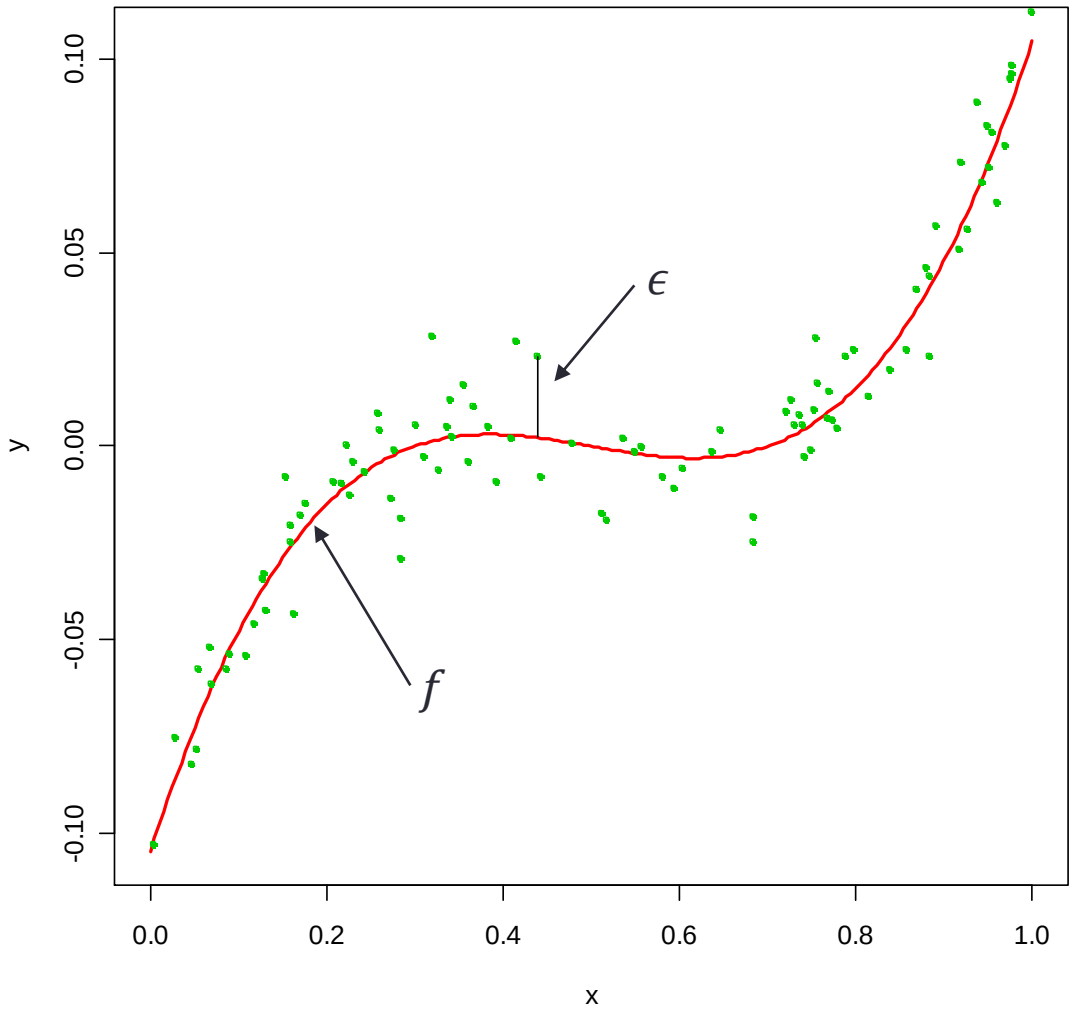
where f is some fixed but unknown function and ε is a random error (noise) with mean zero.

- Statistical learning refers to a set of approaches for estimating f.

A Simple Example

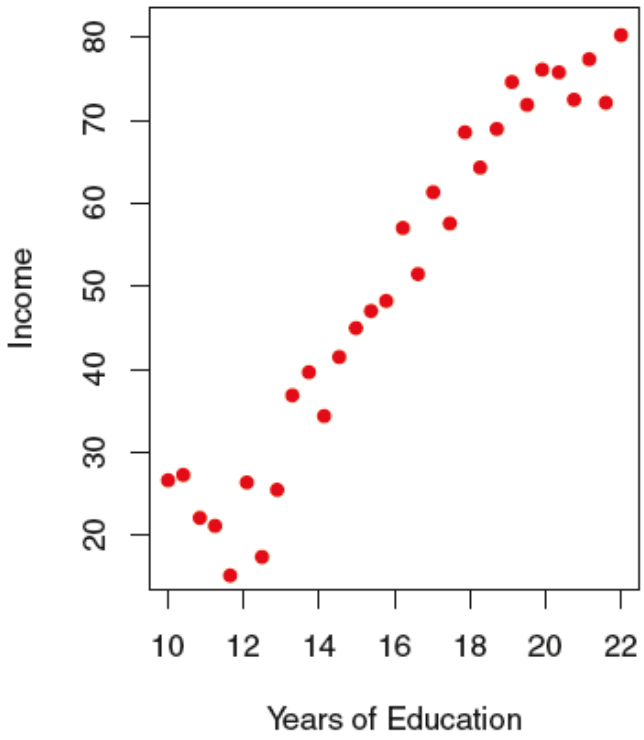


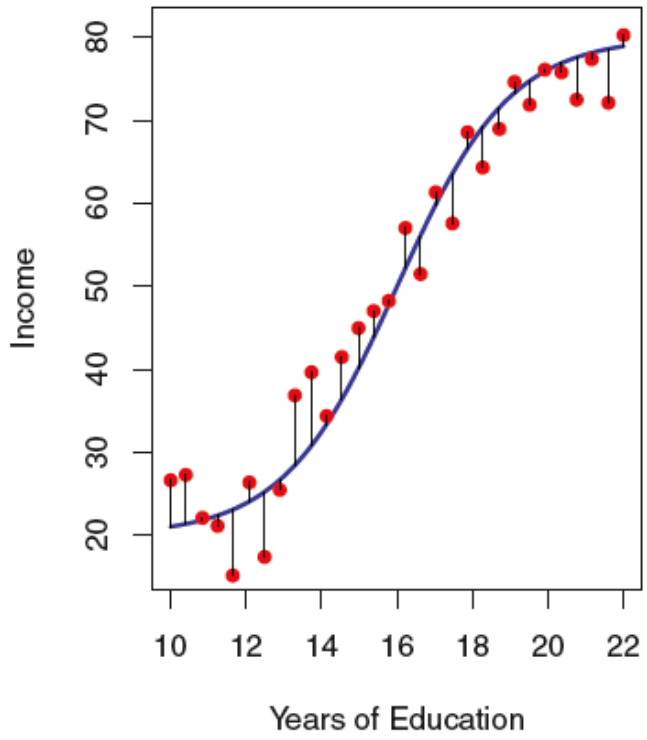
Estimate of f



Income Dataset

A simulated dataset with income (Y) vs. years of education (X) for 30 individuals

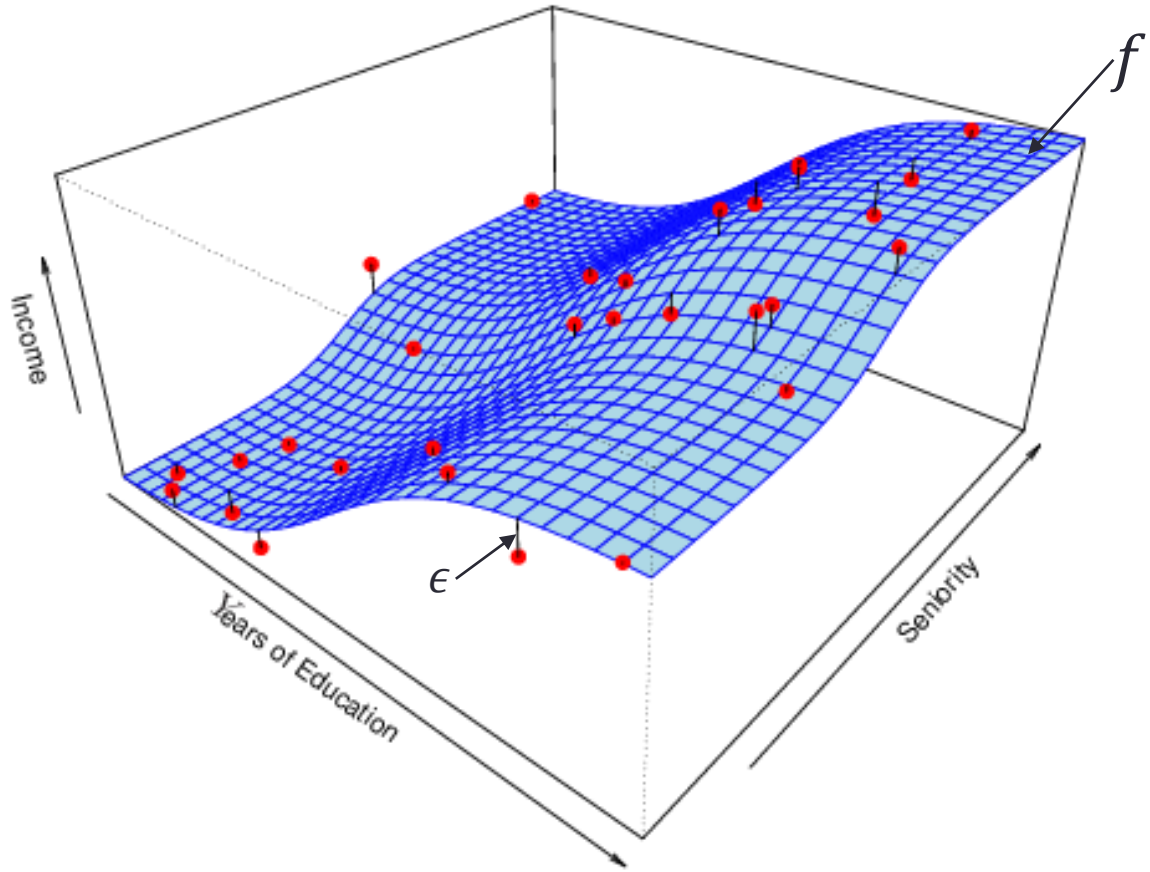




Two-dimensional Case

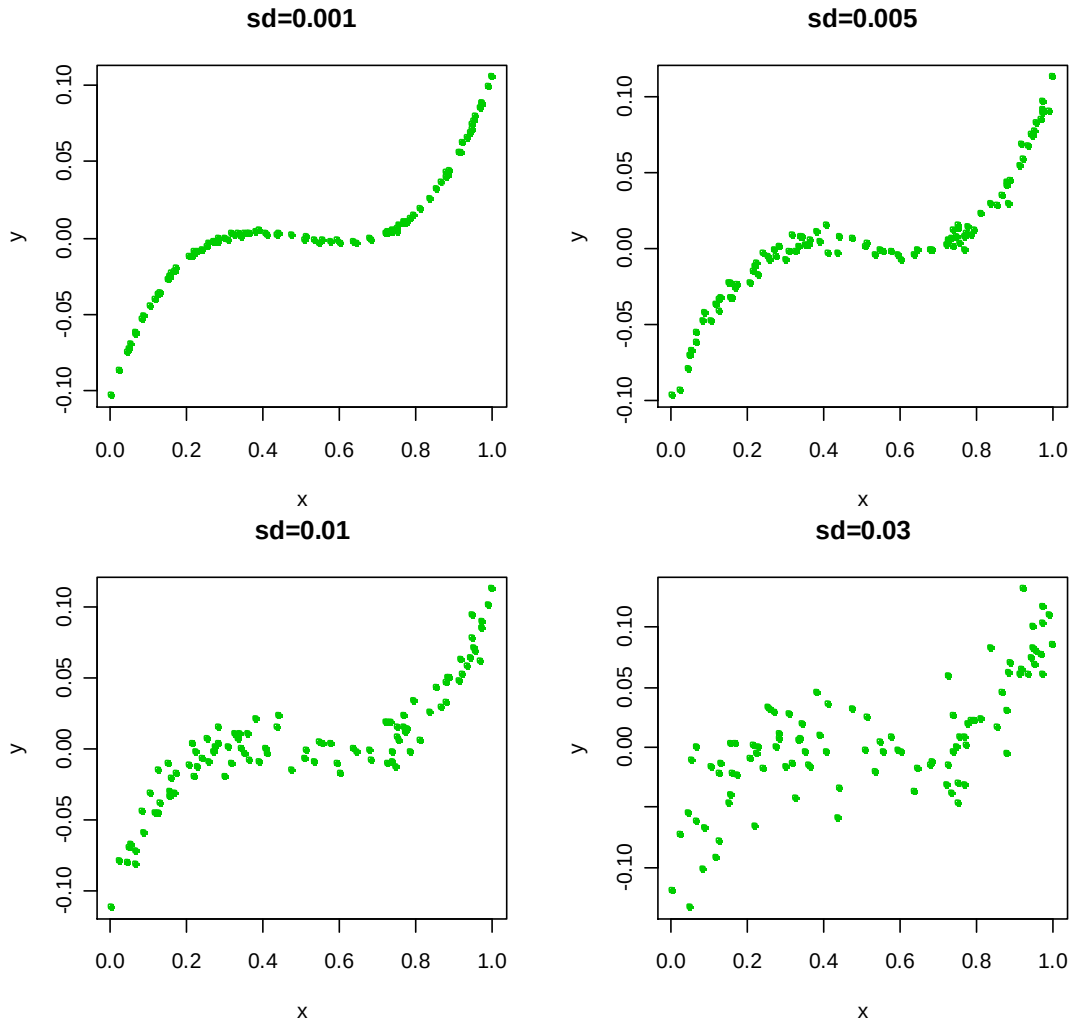
$Y = Income, X_1 = Years\ of\ education, X_2 = Seniority$

$f: two - dimensional\ surface$

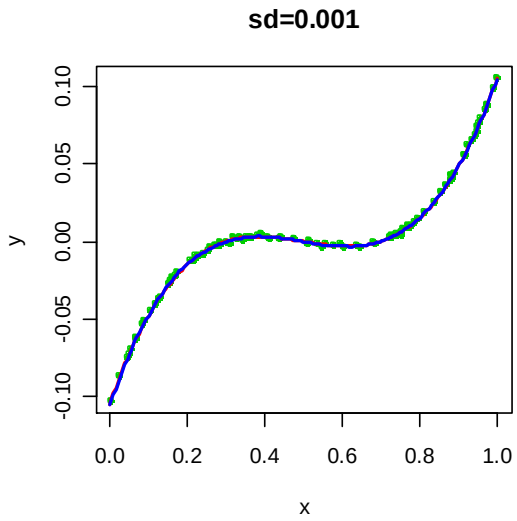


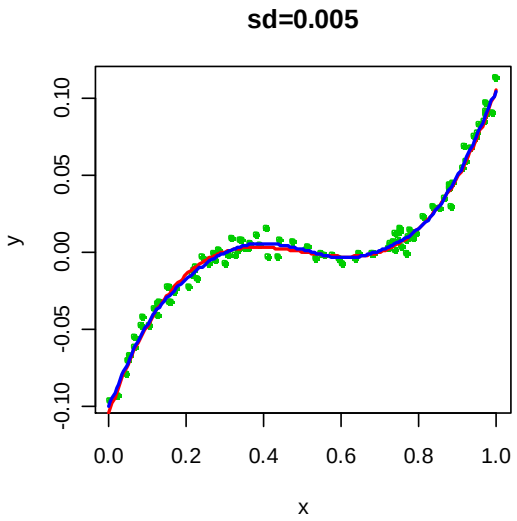
Different Noise Levels

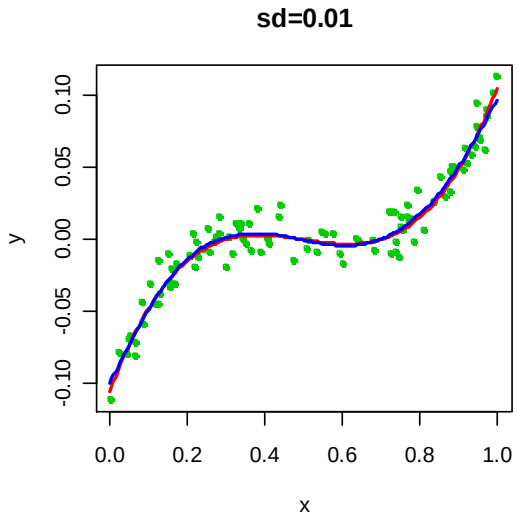
Difficulty of estimating f depends on the noise level, i.e., standard deviation of ε.

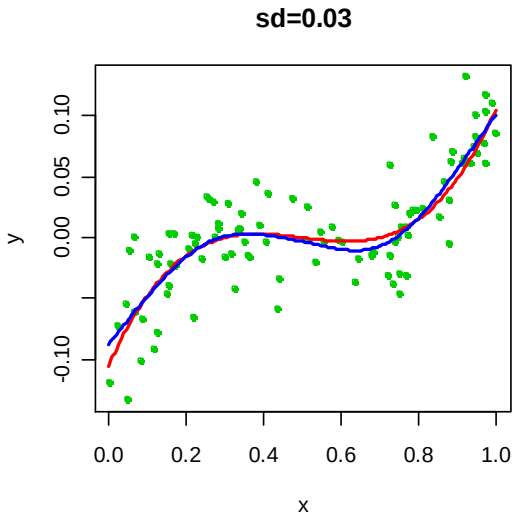


Estimates For f









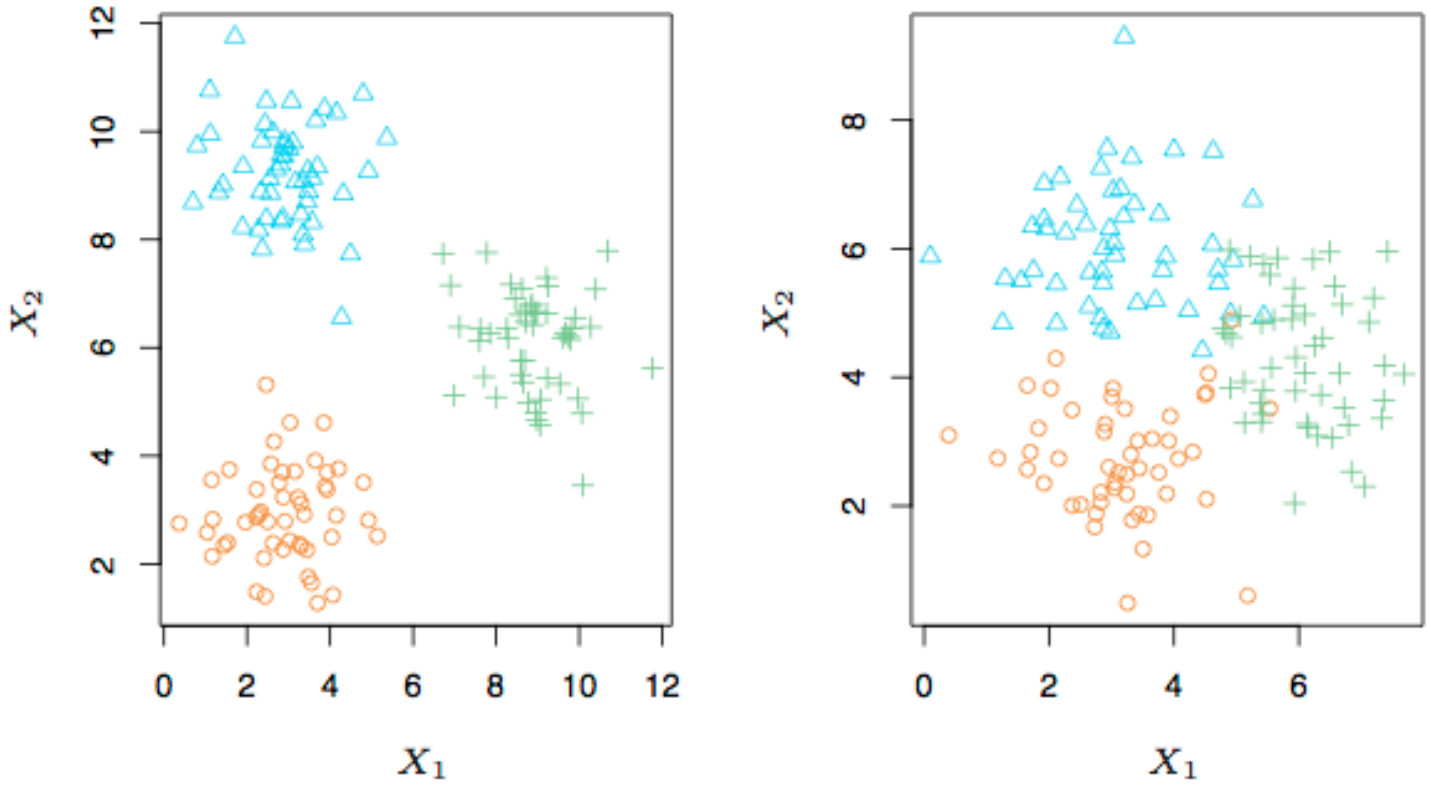
Supervised vs. Unsupervised Learning

- We can divide statistical learning problems into two categories: Supervised and Unsupervised

- Supervised Learning
  - Supervised Learning is where both the predictors, $\boldsymbol{X}$, and the response, Y, are observed, the situation we deal with in linear regression.
  - We wish to fit a model that relates the response to the predictors.
  - This course will focus on supervised learning.

- Unsupervised Learning
  - In this situation only $\boldsymbol{X}$ are observed. It is called "unsupervised" because we lack a response variable that can supervise our analysis.
  - Example: in market segmentation, we try to divide potential customers into groups based on their characteristics. Statistical learning tools such as clustering algorithms can be used to find whether the observations fall into distinct groups.

A Simple Clustering Example



Two Types of Supervised Learning

- Supervised learning problems can be further divided into regression and classification problems.
- Regression covers situations where Y is continuous/numerical. e.g.
  - Predicting the value of the Dow in 6 months
  - Predicting the value of a given house based on various inputs
  - Predicting the throughput of an automobile manufacturer
- Classification covers situations where Y is categorical. e.g.
  - Will the Dow be up (U) or down (D) in 6 months?
  - Is this email a SPAM or not?
  - What will be the status of the patient 1-month after the surgery (normal, weak, or dead)?

Select Statistical Learning Approaches

- Some methods work well on both types of problem, e.g., Neural Networks.
- Other methods work best on one of them, e.g., linear regression for regression, and logistic regression for classification.
- Select statistical learning methods on the basis of whether the response is *quantitative* or *qualitative*.
- Whether the predictors are qualitative or quantitative is generally considered less important. Most of the methods discussed in this course can be applied regardless of the predictor variable type, provided that qualitative predictors are properly coded before the analysis is performed.

Reasons for Estimating f

- There are two reasons for estimating f
  - **Prediction**
  - **Inference**

Prediction

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

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

- $\hat{f}$ represents estimate for f, $\hat{Y}$ represents resulting prediction.
- $\hat{f}$ is often treated as a black box, in the sense that one is not typically concerned with the exact form of $\hat{f}$, provided that it yields accurate predictions for Y.

"Essentially all models are wrong, but some are useful."

"…..there is no need to ask the question "Is the model true?". If "truth" is to be the "whole truth" the answer must be "No". The only question of interest is "Is the model illuminating and useful?"

---- George Box

Examples of Prediction

- **Example 1: direct mailing**

  --- Interested in predicting how much money an individual will donate (Y) based on over 400 different characteristics ($\boldsymbol{X}$) of people.

  --- Don't care too much about each individual characteristic.

  --- Just want to know: for a given individual should I send out a mailing?

- **Example 2: adverse reaction to particular drug**

  --- Interested in predicting the patient's risk for adverse reaction to a particular drug (Y) based on the characteristics ($\boldsymbol{X}$) of the patient's blood sample

  --- Don't care the blood characteristics of patients.

  --- Just want to know: for a given patient, can this drug be used?

Inference

- Alternatively, we may want to understand the relationship between $\boldsymbol{X}$ and Y.
- How Y changes when $\boldsymbol{X}$ change?
- For example,
  - **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.
  - **What is the relationship between the response and each predictor?**
    Some predictors may have a positive relationship with Y (increasing the predictor is associated with increasing Y), while some may have a negative relationship.
  - **Is the relationship a simple linear one or more complicated?**
- In this case $\hat{f}$ cannot be treated as a black box because we need to know its exact form.

Examples of Inference

- **Example 1: house price**

  --- Wish to predict median house price based on 14 variables.

  --- Probably want to understand which factors have the biggest effect on house price and how big the effect is. For example, how much impact does a river view have on the house price?

- **Example 2: advertising**

  --- Wish to predict sales of a product based on advertising budgets for three different media: TV, radio, and newspaper.

  --- One may be interested in answering questions such as: which media contribute to sales? Which media generate the biggest boost in sales? How much increase in sales is associated with a given increase in TV advertising?

Predictive Modeling vs. Inferential Modeling

- Predictive modeling: modeling for prediction
  Inferential modeling: modeling for inference
- Depending on whether our ultimate goal is prediction, inference, or a combination of the two, **different methods for estimating f may be appropriate**.
- Linear models allow for relatively simple and interpretable inference, but may not yield accurate predictions.
- Highly nonlinear approaches 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.

Estimating f

- Assume we have observed a set of **training data**

$$\{(\boldsymbol{X}_1, Y_1), (\boldsymbol{X}_2, Y_2), \dots, (\boldsymbol{X}_n, Y_n)\}$$
$$\boldsymbol{X} = (X_1, X_2, \dots, X_p)$$

- p = #predictors, n = #observations.
- We will use the training data and a statistical method to estimate f.
- Statistical learning methods:
  - Parametric Methods
  - Non-parametric Methods

Parametric Methods

- Reduce the problem of estimating f down to one of estimating a set of parameters.

- Involve a two-step model-based approach:

STEP 1:

Make some assumption about the *functional form* or *shape* of f, i.e., come up with a model structure. The most common example is a linear model

$$f(\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

STEP 2:

Using the training data to *fit* or *train* the model, i.e., estimate the parameters of the model $\beta_0, \beta_1, \ldots, \beta_p$. That is, we want to find values of these parameters such that

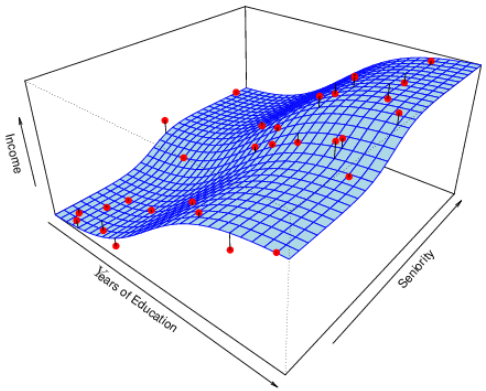
$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

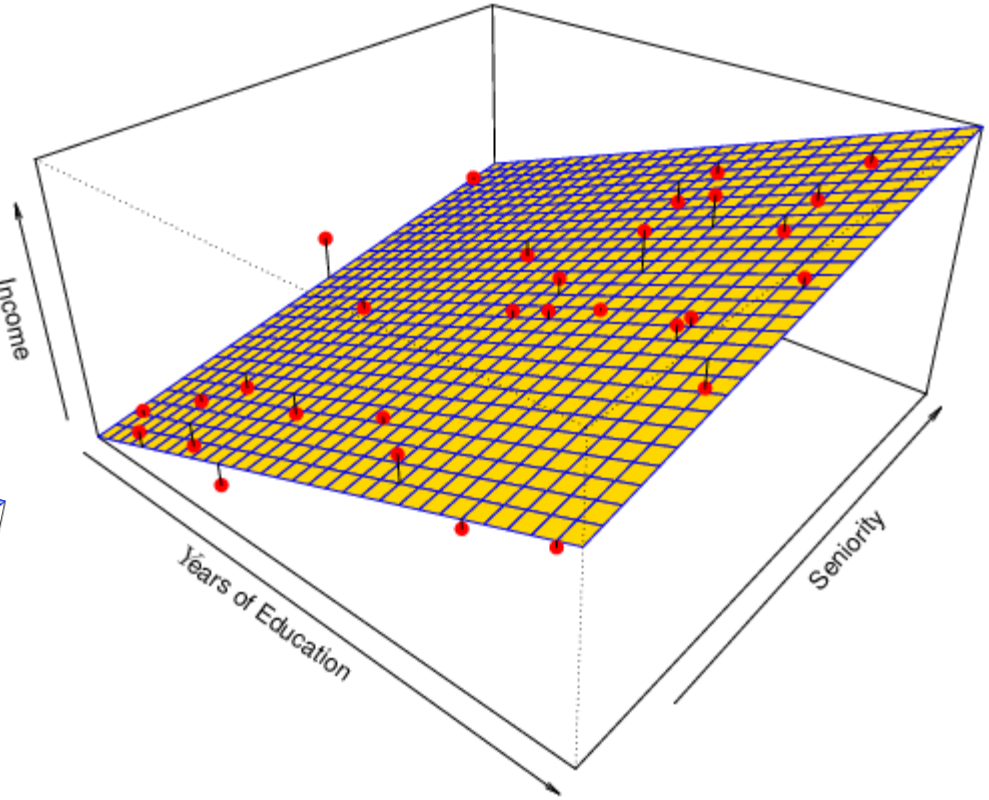
Pros and Cons of Parametric Methods

- **Pros:** simplify the problem of estimating f: it is generally much easier to estimate a set of parameters, such as the βs in the linear model, than to fit an entirely arbitrary function f.
- **Cons:** the model we choose may 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.
- This problem can be solved 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.

Example: Linear Model Fit

$$Income \approx \beta_0 + \beta_1 \times Education + \beta_2 \times Seniority$$



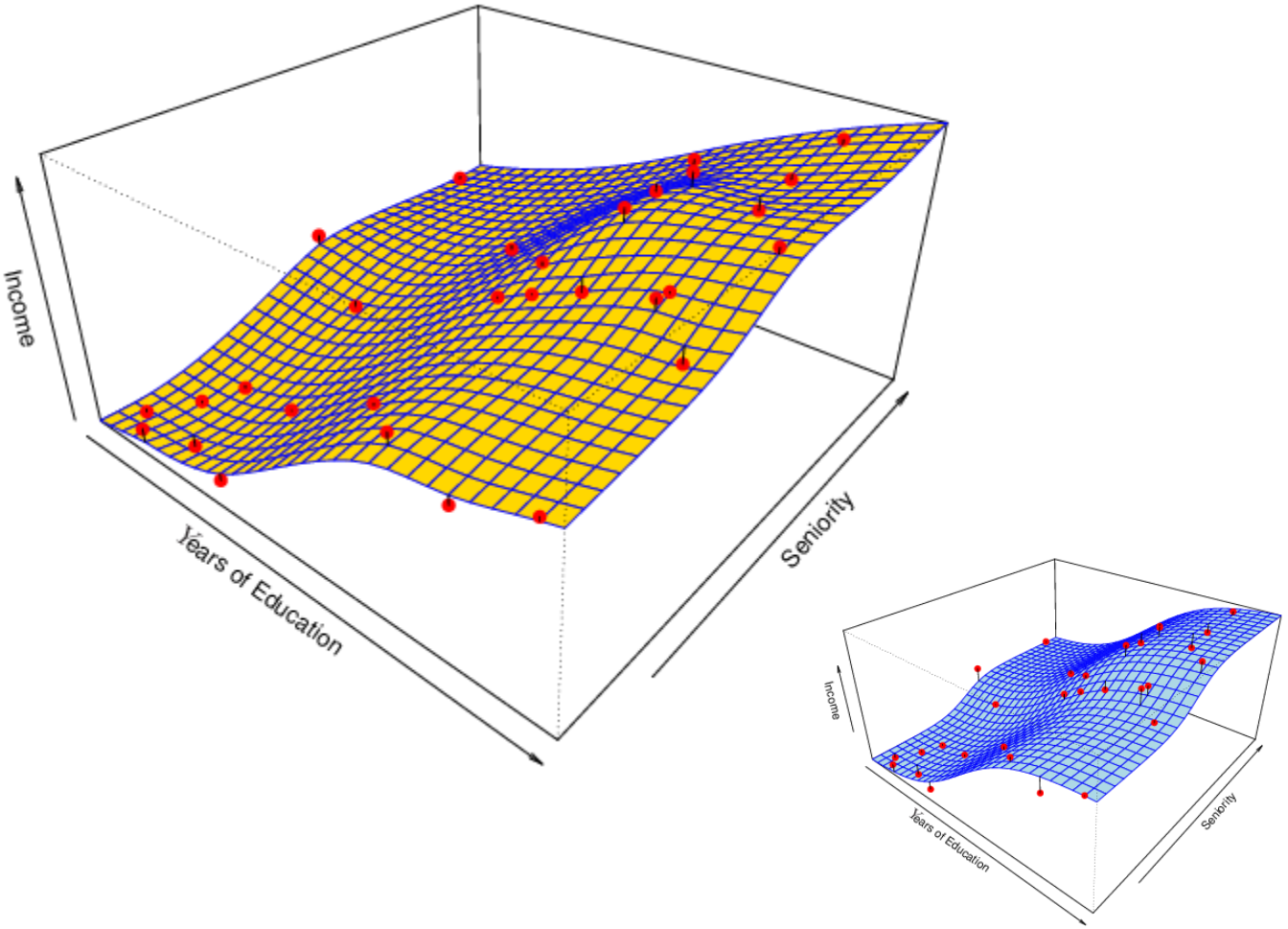


- The linear fit is not quite right; curvature in the true function is not captured in the linear fit.
- However, the linear fit still appears reasonable in capturing the positive relationship between education and income, and the slightly less positive relationship between seniority and income.

Non-parametric Methods

- They 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.
- Pros:
  - By avoiding the assumption of a particular functional form, they have the potential to accurately fit a wider range of possible shapes for f.
  - No bias induced due to forcing the model structure
- Cons
  - Since they don't reduce the problem of estimating f to a small number of parameters, a very large number of observations is required to obtain an accurate estimate for f.

Example: Thin-Plate Spline Fit



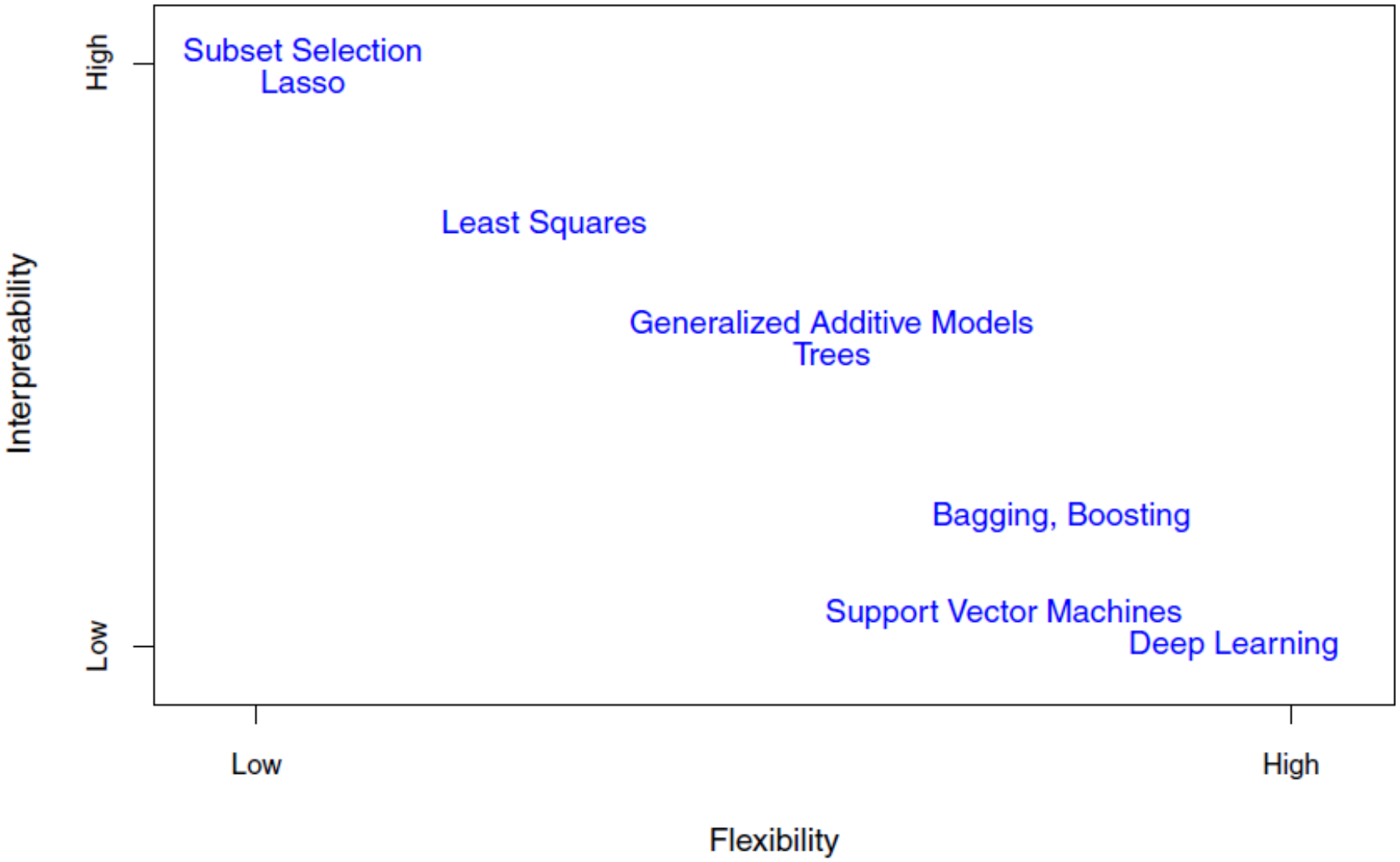
- The non-parametric fit produces a remarkably accurate estimate of the true f

Why Use a Restrictive Method?

- Restrictive methods: produce a relatively small range of shapes to estimate f, e.g., linear regression
- Flexible methods: generate a much wider range of possible shapes to estimate f, e.g., thin-plate splines
- *Why would we ever choose a restrictive method instead of a very flexible one?*
- There are two reasons

  Reason 1: If we are mainly interested in inference, then a simple model may be good choice since it is easy to understand the relationship between the response and predictors (better interpretability).

Flexibility vs. Interpretability

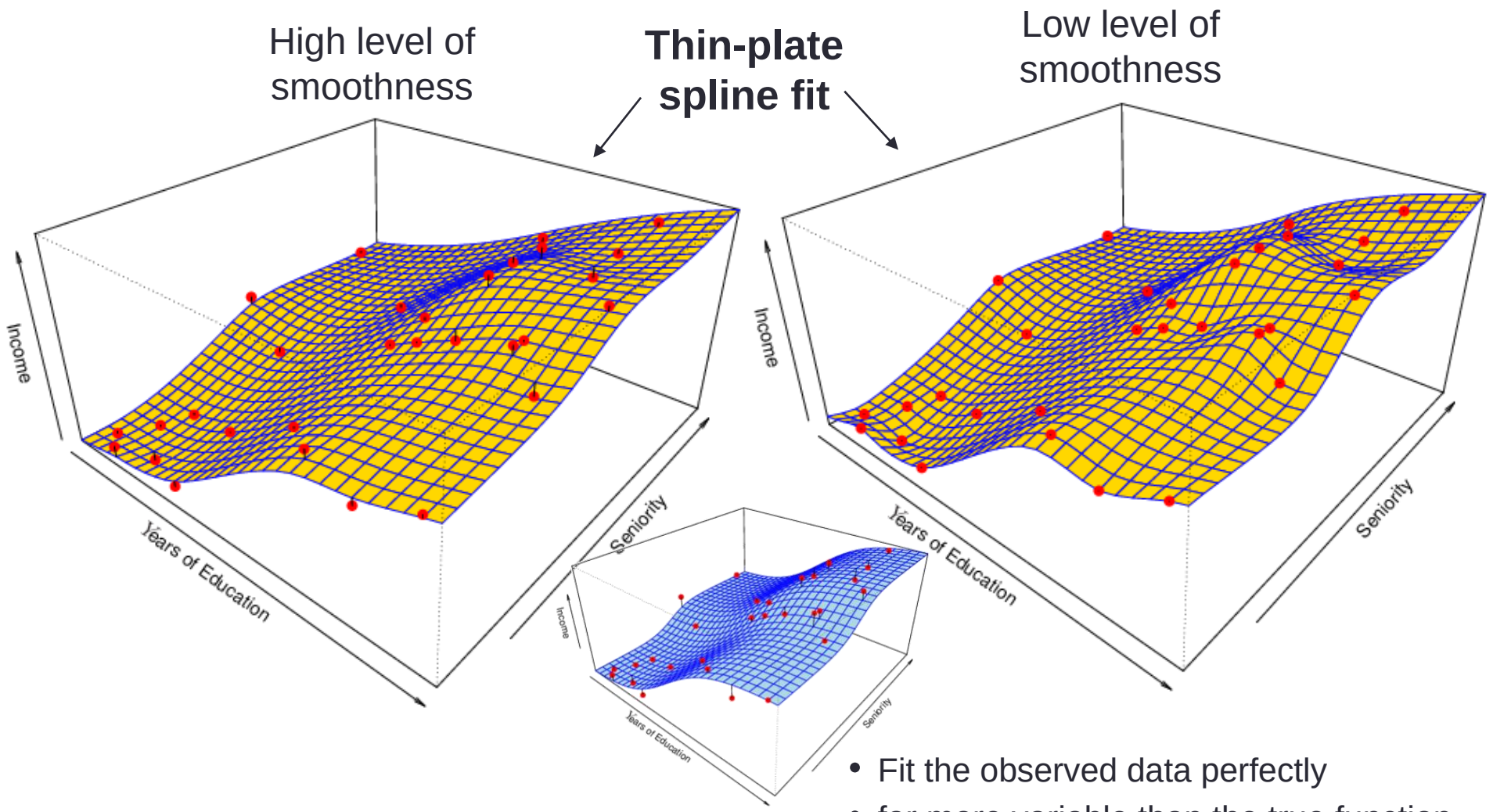


Why Use a Restrictive Method?

Reason 2: Even if we are only interested in prediction, it is often to obtain more accurate predictions using a less flexible method.

- This seems counterintuitive but has to do with the potential for *overfitting* in highly flexible methods.
- Overfitting: the model follows errors, or noise, too closely.

Example of Overfitting



- Fit the observed data perfectly
- far more variable than the true function
- will not yield accurate estimates of the response on new observations

Predictive Analytics

- This course will focus on predictive analytics, i.e., data analysis for prediction.

Selecting A Learning Method

- This course will cover a wide range of statistical learning methods, including the standard linear regression and many others that are more complex.
- *Why is it necessary to learn so many different approaches, rather than just a single best method?*
- No free lunch in statistics: no one method dominates all others over all data sets.
- It is important to decide for any given data set which method produces the best predictions.
- However, selecting the best approach can be one of the most challenging parts of performing statistical learning in practice!

Comparing Different Approaches

- To select the best approach for a given application, we need to evaluate the performance of each learning method.
- That is, to ***measure how well its predictions actually match the observed data***, or quality of fit.

Learning Process

- **1: Training:** Fit the learning method on the ***training data set***

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

to obtain the estimate $\hat{f}$.

Example: in linear regression, we find the estimates of the parameters $\hat{\beta}_0, \hat{\beta}_1$, and then the estimate of f: $\hat{f}(x) \approx \hat{\beta}_0 + \hat{\beta}_1 x$.

Predictions on the training data: $\hat{f}(x_1), \hat{f}(x_2), \ldots, \hat{f}(x_n)$

Performance measure: ***mean squared error*** (MSE)

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

The ***training MSE*** will be small if the predicted responses are very close to the true responses.

Learning Process

- **2. Prediction:** Use the estimate $\hat{f}$ to predict on the ***test data set***

$$\{(x_{01}, y_{01}), (x_{02}, y_{02}), \ldots, (x_{0m}, y_{0m})\}$$

Note: the test data are previously unseen observations not used to train the statistical learning method.

Example: in linear regression, the prediction is: $\hat{f}(x_0) \approx \hat{\beta}_0 + \hat{\beta}_1 x_0$.

Predictions on the test data: $\hat{f}(x_{01}), \hat{f}(x_{02}), \ldots, \hat{f}(x_{0m})$

Performance measure: mean squared error (MSE)

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_{0i} - \hat{f}(x_{0i}))^2$$

The **test MSE** will be small if the predicted responses are very close to the true responses.

Question: Training MSE vs. Test MSE

- Training MSE measures accuracy of the method in predicting the training data.
- Test MSE measures accuracy of the method in predicting the test data that are not used in training.

Which one should be taken as the performance measure of the method in prediction?

In other words, should we choose the method that minimizes the training MSE or the one that minimizes the test MSE?

Training MSE vs. Test MSE

- We do not care about how well the method works on the training data. Rather, we are interested in the accuracy of the predictions that we obtain when apply our method to previously unseen test data.
- **Example 1:** suppose we want to develop an algorithm to predict a stock's price based on previous stock returns. We can train the method using stock returns from the past 6 months. But we really don't care how well our method predicts last week's stock price; we instead care about how well it will *predict tomorrow's price or next month's price*.
- **Example 2:** suppose we have clinical measurements (e.g., weight, blood pressure, height, age) for a number of patients, as well as information about whether each patient has diabetes. We can use these data to train a statistical learning method to predict risk of diabetes based on clinical measurements. In practice, we want this method to accurately predict diabetes risk for *future patients*; we are not interested in whether or not the method accurately predicts diabetes risk for patients used to train the method, since we already know which of those patients have diabetes.

Guidelines

- **For the best prediction performance, select the statistical learning method that leads to the lowest test MSE.**
- **There is no guarantee that the method with the lowest training MSE will also have the lowest test MSE.**

Question: MSE vs. Flexibility

Statistical learning methods

Restrictive methods (e.g., linear regression)

Flexible methods (e.g., deep learning)

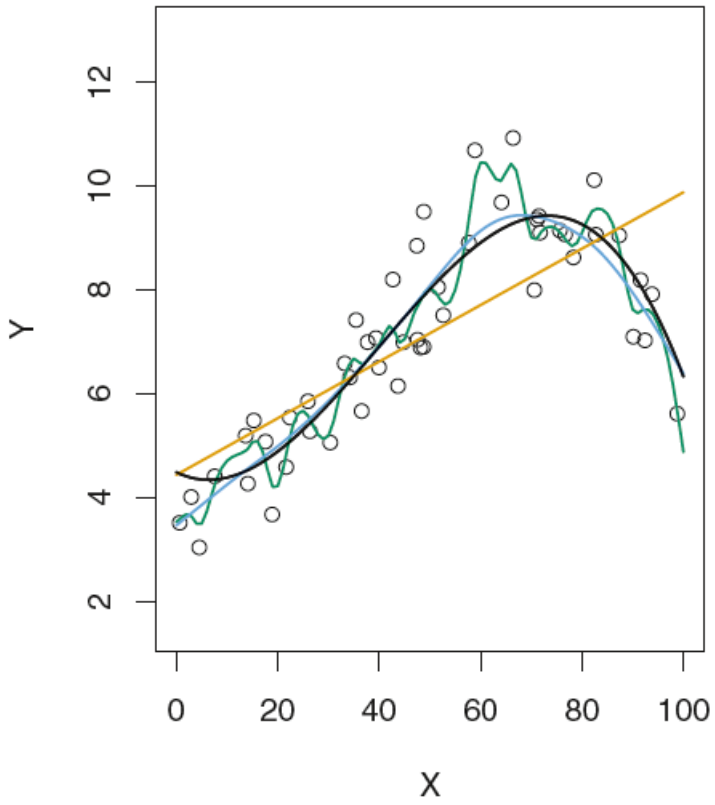
?

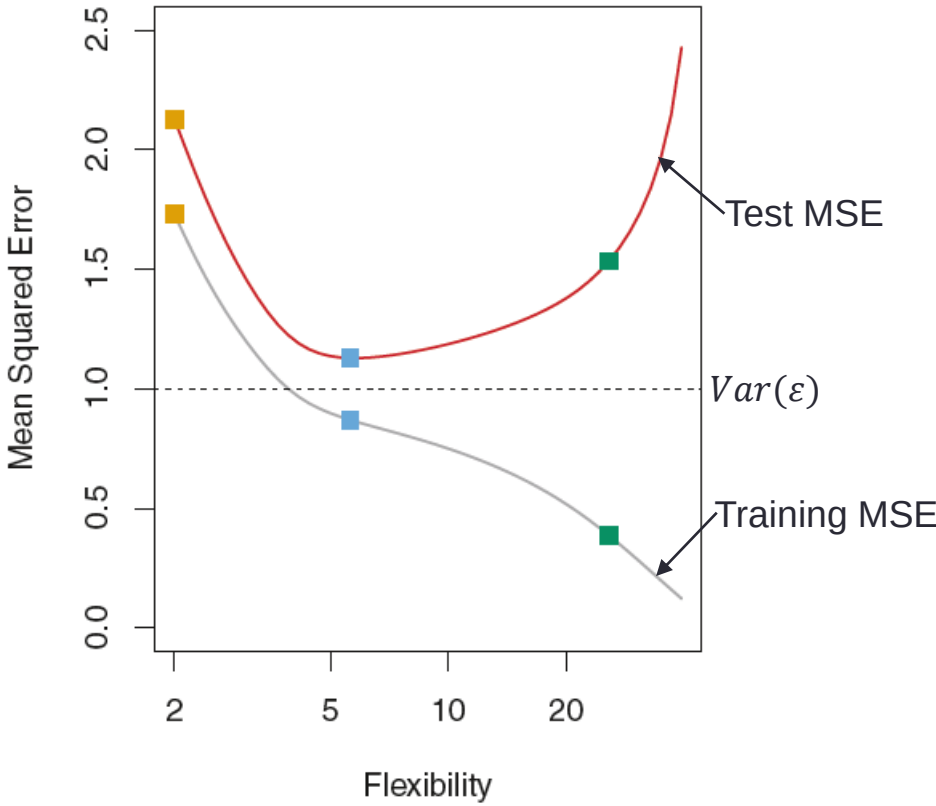
lowest training MSE

lowest test MSE

Simulation Example 1

True function: nonlinear
Noise level: high





LEFT

Black: True function

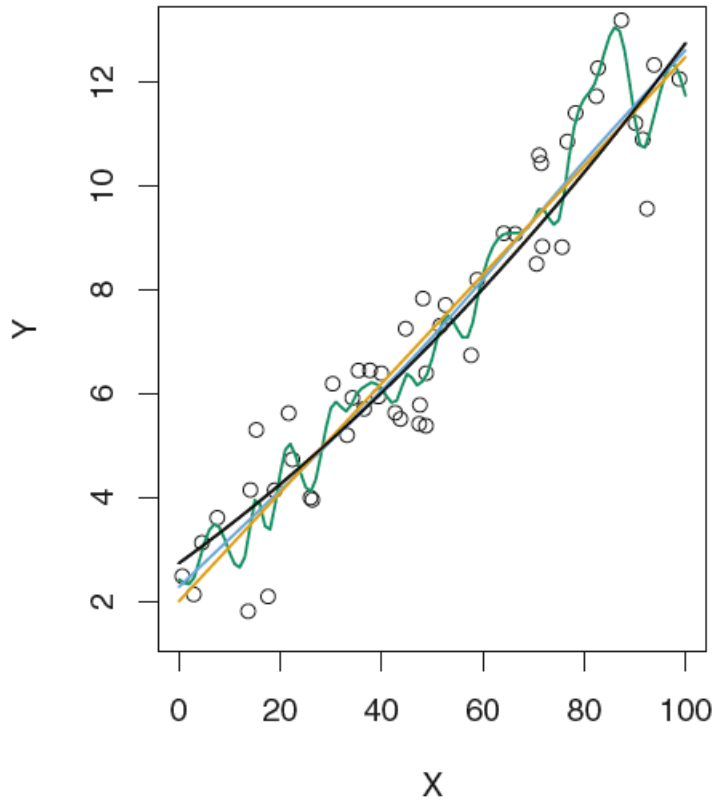
Orange: Linear regression (restrictive)

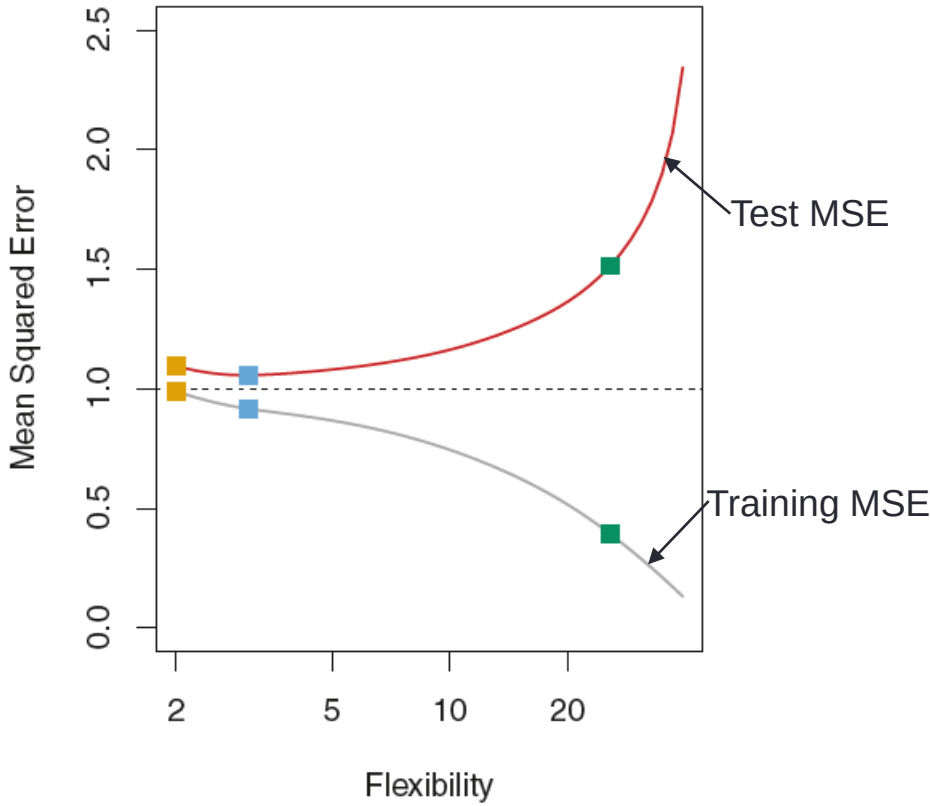
Blue: Smoothing spline (flexible)

Green: Smoothing spline (more flexible)

Simulation Example 2

True function: approximately linear
Noise level: high





LEFT

Black: True function

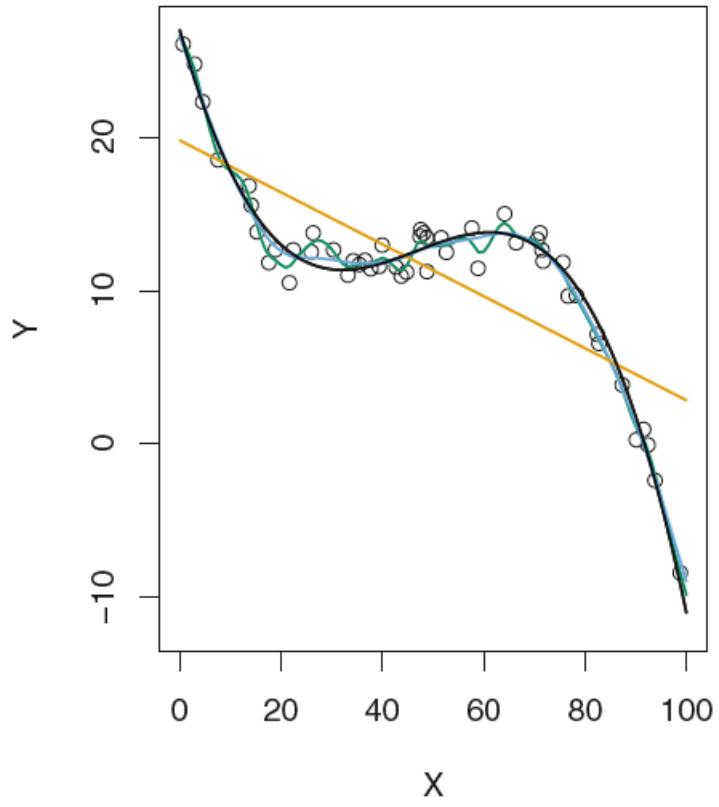
Orange: Linear regression (restrictive)

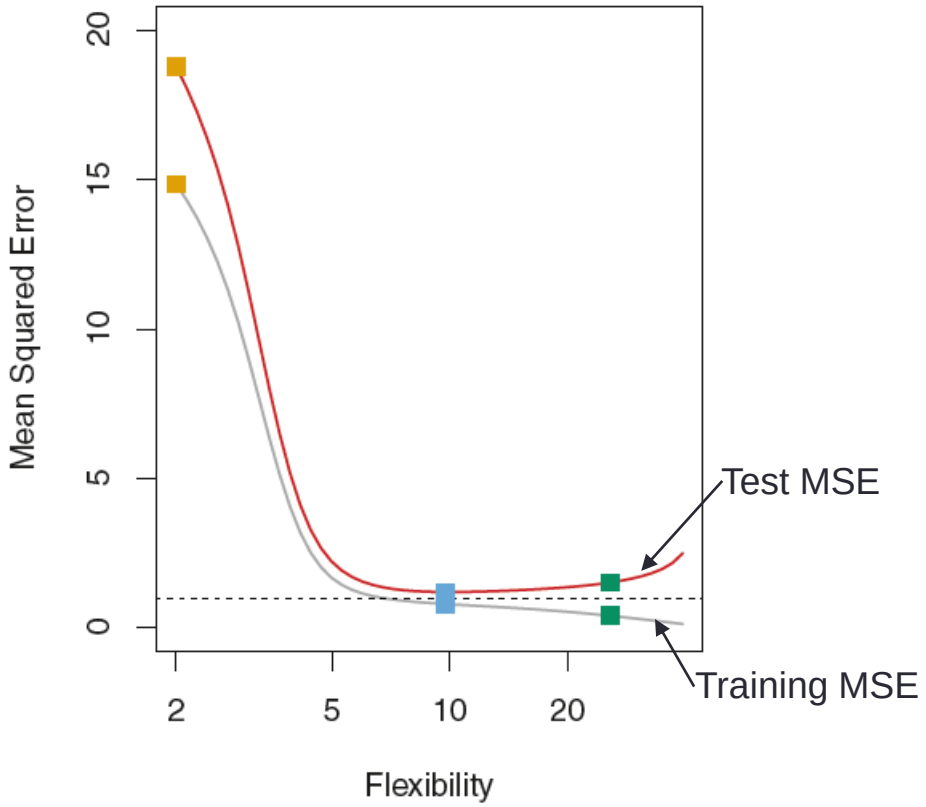
Blue: Smoothing spline (flexible)

Green: Smoothing spline (more flexible)

Simulation Example 3

True function: nonlinear
Noise level: low





LEFT

Black: True function

Orange: Linear regression (restrictive)

Blue: Smoothing spline (flexible)

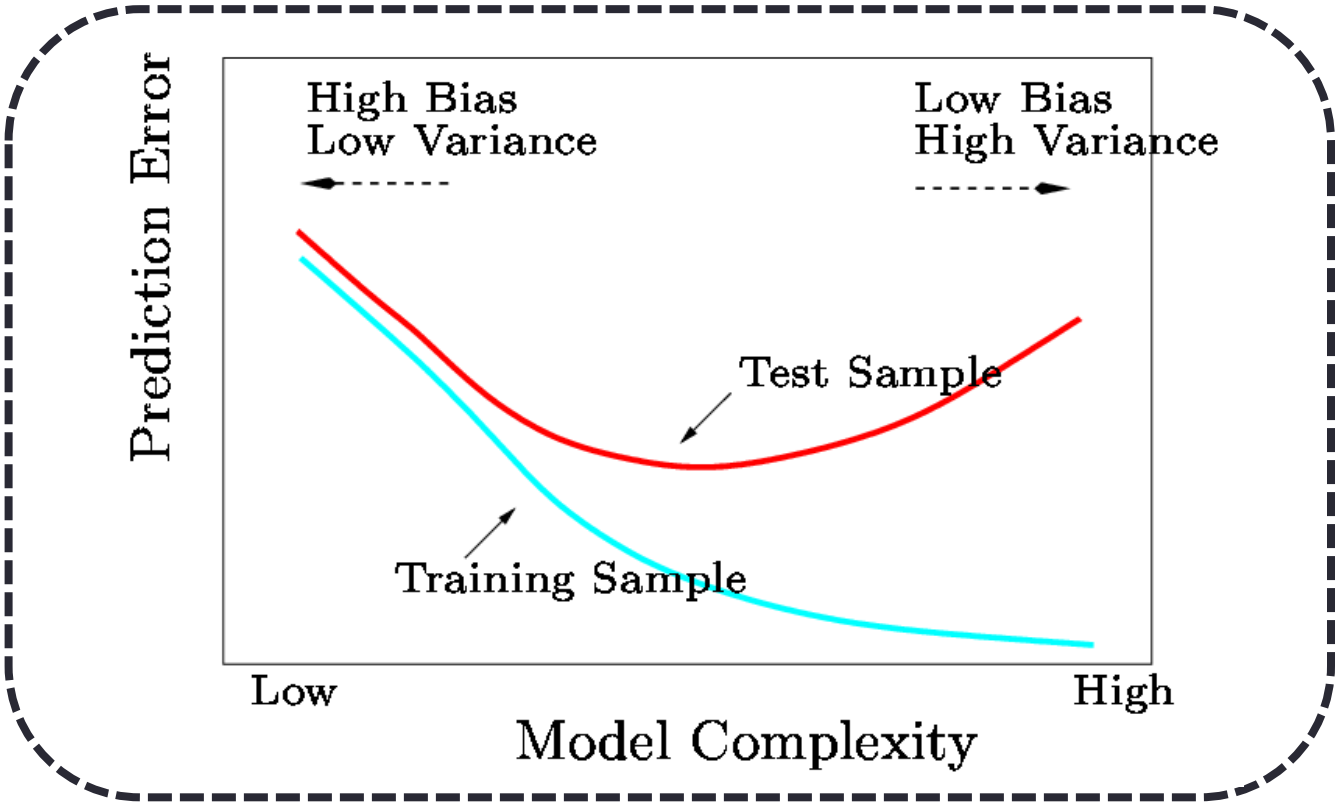
Green: Smoothing spline (more flexible)

Question: General Patterns of MSE

- What is the shape of the training MSE based on the three examples? In other words, in general, how does the training MSE change when a more flexible method is used?
- What is the shape of the test MSE based on the three examples? In general, how does the test MSE change when a more flexible method is used?

A Fundamental Picture

- In general, training errors always decline as flexibility (complexity) increases.
- However, test errors show a **U-shape**:
  - decline at first (as reductions in bias dominate)
  - then start to increase again (as increases in variance dominate).



Always keep this picture in mind when choosing a learning method.

More flexible/complicated is not always better!

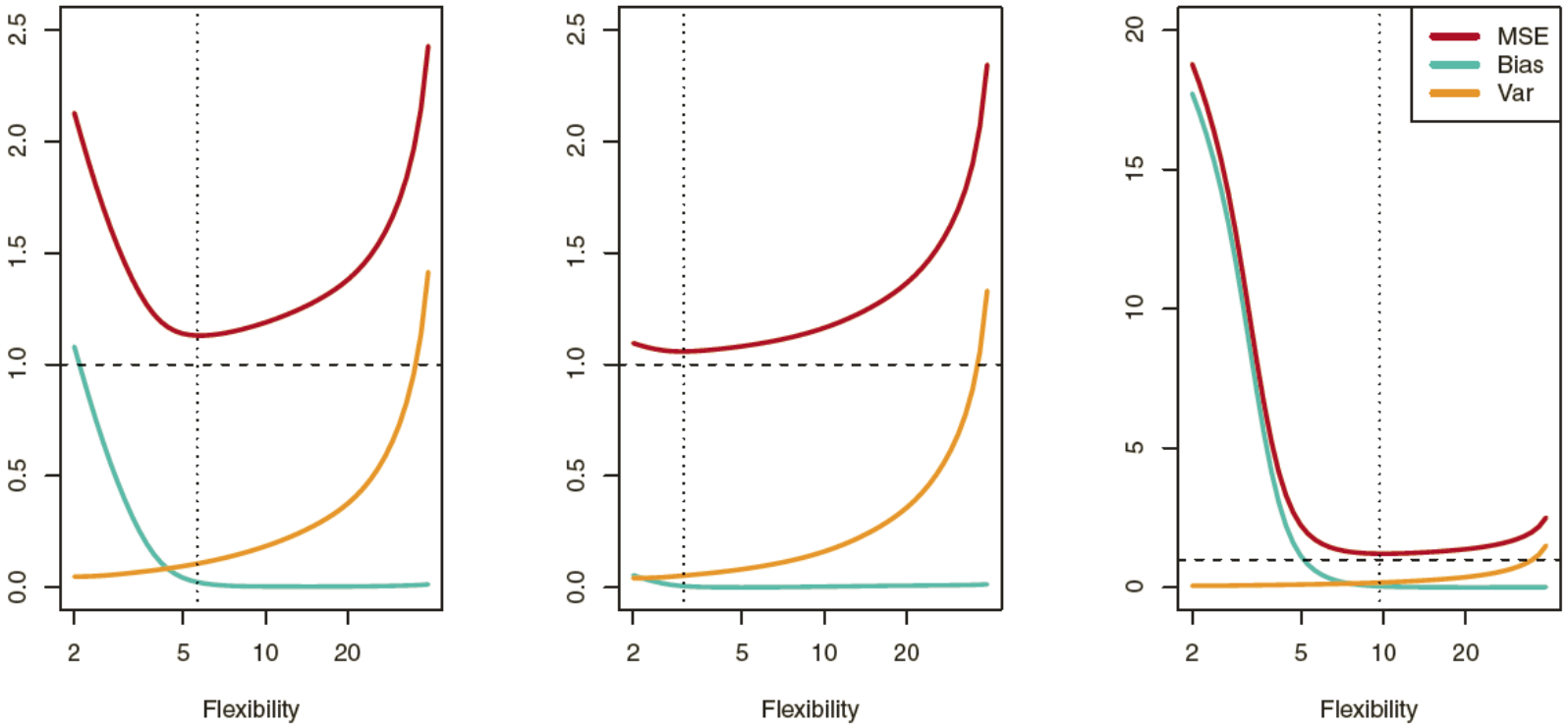
Bias vs. Variance

- The U-shape of the test MSE curve results from two competing properties of statistical learning methods: bias and variance.

$$Test\ MSE = Bias^2 + Variance + Var(\varepsilon)$$

- **Bias:** the error caused by approximating a real-life problem.
  Example: when the true function is substantially nonlinear, linear regression results in high bias.
- **Variance:** uncertainty due to the randomness of the training data ($\hat{f}$ would change if we estimated it using a different training data set).

Bias — Variance Trade-off



- In general, more flexible methods have lower bias and higher variance.
- One should always keep the bias-variance trade-off in mind in choosing statistical learning methods in practice.