Supervised learning and Unsupervised Learning

Hao Luo

Department of Social Work and Social Administration
The University of Hong Kong

Oct 4, 2021

Learning from data

- Vast amounts of data are being generated in many fields.
- Our job: make sense of it...
  - to extract important patterns and trends
  - identify relationships
  - understand what the data says
- The challenges in learning from data have led to a revolution in the statistical sciences.

What is statistical learning?

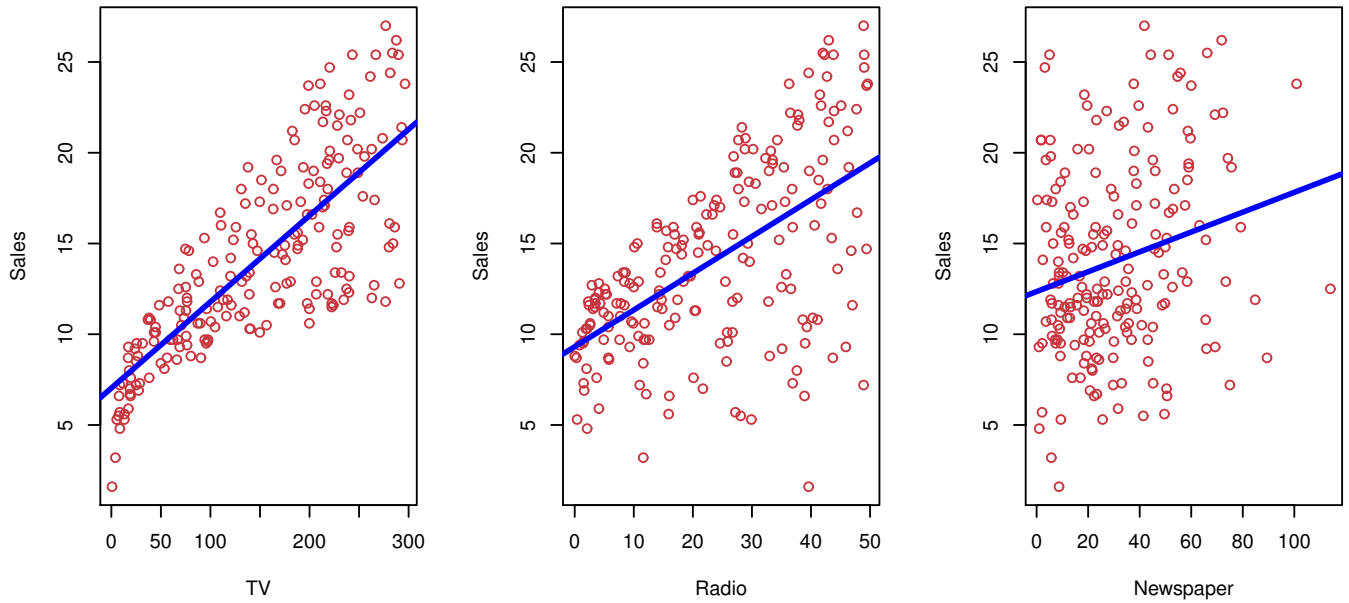
A simple statistical learning problem

Suppose that you are a consultant

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.

The goal: to develop an accurate model that can be used to predict sales on the basis of the three media budgets

- *input variables (predictors, independent variables, features)*: TV (X_1), radio (X_2), and newspaper (X_3)
- *output variable(response, dependent variable)*: sales (Y)



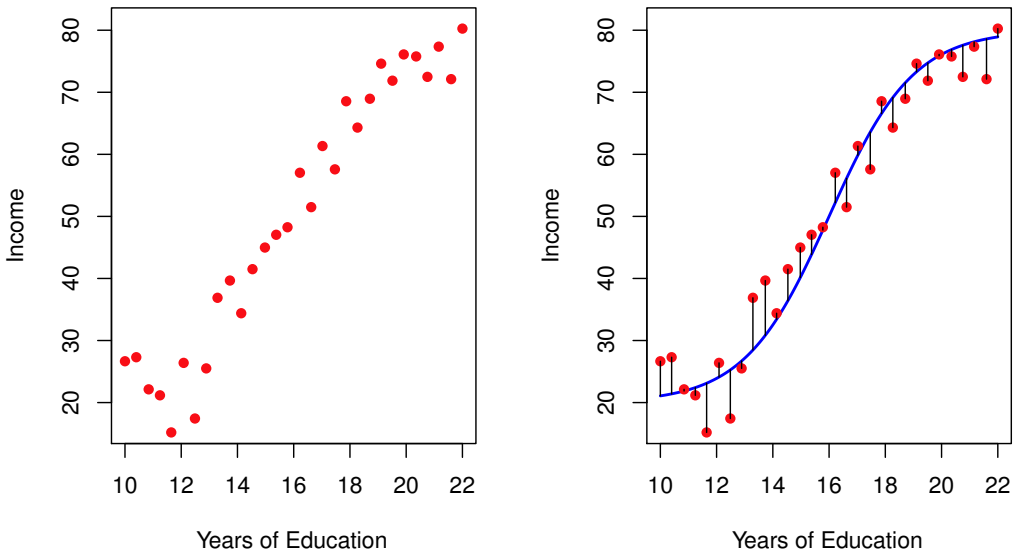
Suppose that we observe a quantitative response Y and p different predictors, X_1, X_2, ..., X_p. We assume that there is some relationship between Y and $X = (X_1, X_2, \ldots, X_p)$, which can be written in the very general form

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

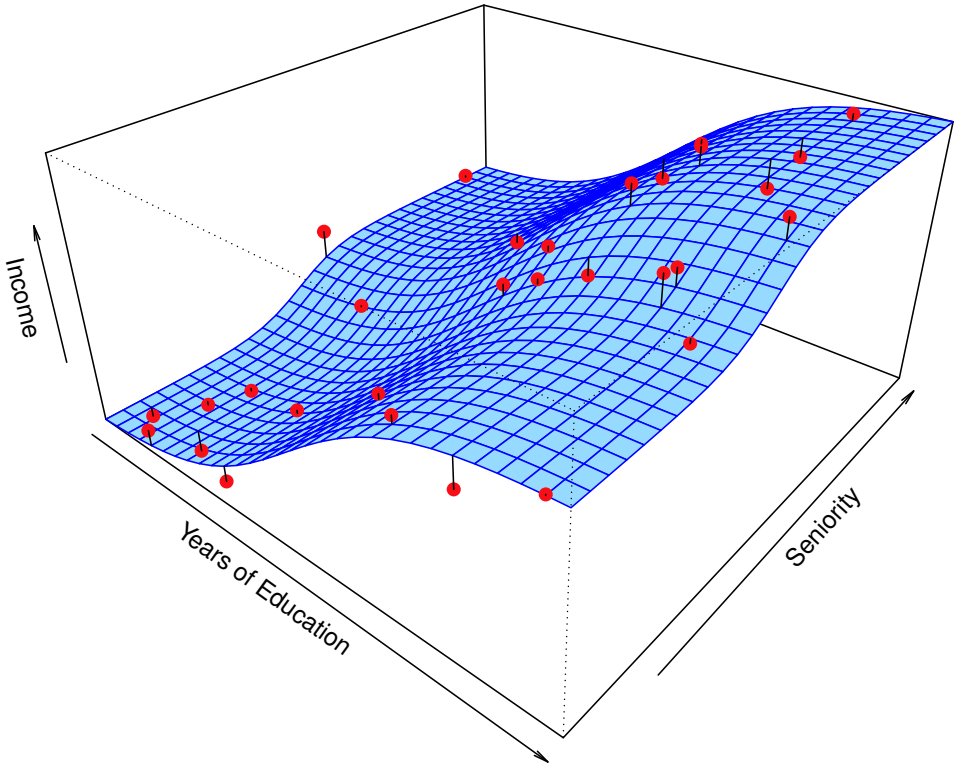
Another example

Educational capital

Income versus years of education for 30 individuals in the Income data set. The function f that connects the input variable to the output variable is in general unknown.



f may also be a two-dimensional surface that must be estimated based on the observed data.



The ultimate goal

estimating f!

Two main reasons that we may wish to estimate f

Prediction

- Inputs X are readily available, but the output Y cannot be easily obtained.
- 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.
- $\hat{f} \Rightarrow$ black box.

Black boxes at large

- Patients blood sample $\Rightarrow$ the patient's risk for a severe adverse reaction to a particular drug.
- Correctional Offender Management Profiling for Alternative Sanctions (COMPAS):
  137 feature $\Rightarrow$ a criminal defendants likelihood of committing a crime
- Netflix Prize
- .....

The accuracy of $\hat{Y}$ is determined by...

$$\begin{aligned} E(Y-\hat{Y})^2 &= E[f(X)+\epsilon-\hat{f}(X)]^2 \\ &= [f(X)-\hat{f}(X)]^2+\text{Var}(\epsilon) \end{aligned}$$

- *reducible error* + *irreducible error*

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

- reducible error: in general $\hat{f}$ will not be a perfect estimate for f;
- irreducible error: ϵ cannot be predicted using X.
- ϵ may contain:
  1. Unmeasured variables that are useful in predicting Y;
  2. Unmeasurable variation/measurement error.

Inference

How Y changes as a function of $X_1, \ldots, X_p$.

- Which predictors are associated with the response? Identifying the few important predictors.
- What is the relationship between the response and each predictor?
  - Positive?
  - Negative?
  - Depending on values of other predictors?
- Linear or more complicated?

How Do We Estimate f?

- Throughout this course, we explore many linear and non-linear approaches for estimating f.
- They generally share certain characteristics.
- We will always assume that we have observed a set of n different data points. $\Rightarrow$ training data.
- Our goal is to apply a statistical learning method to the training data in order to estimate the unknown function f.
- Most statistical learning methods for this task can be characterized as either *parametric* or *non-parametric*.

Parametric Methods

A two-step model-based approach:

1. We make an assumption about the functional form, or shape, of f.
$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$
2. A procedure that uses the training data to fit or train the model (estimation method)
   - E.g., (ordinary) least squares, etc.

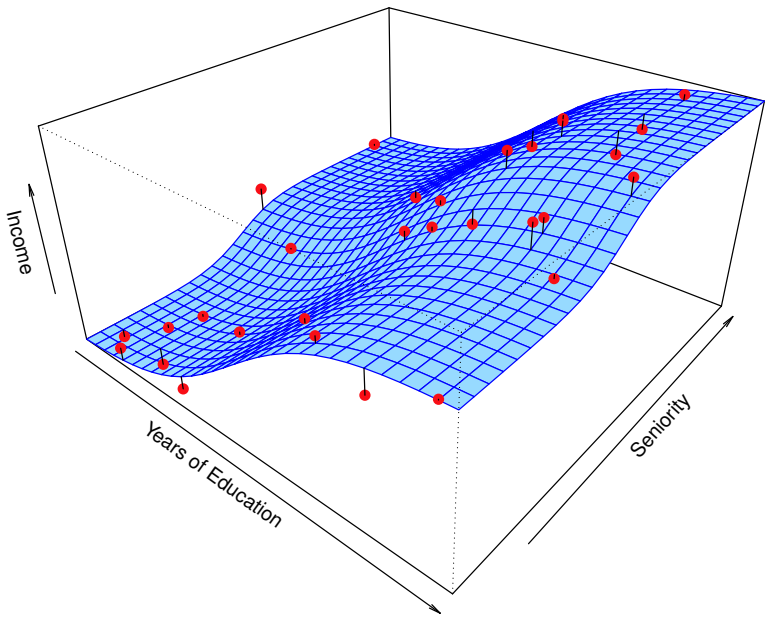
The model-based approach just described is referred to as *parametric*; it reduces the problem of estimating f down to one of estimating a set of parameters.

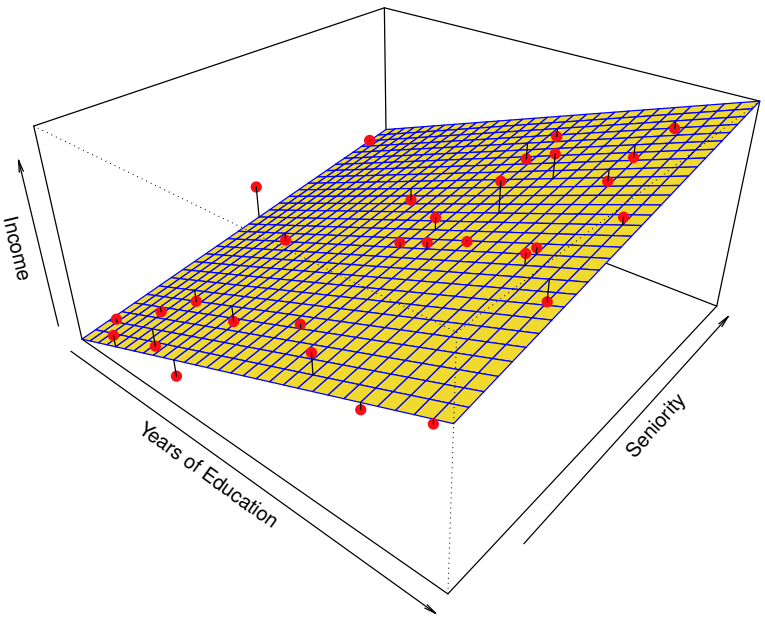
Pros and cons?

- Simplifies the problem of estimating f.
- The model we choose will usually not match the true unknown form of f.
- Solution: choosing *flexible* models
- New problem: *overfitting* $\rightarrow$ follow the noise too closely

An example of the parametric approach - Income data

$$\text{income} \approx \beta_0 + \beta_1 \times \text{education} + \beta_2 \times \text{seniority}$$



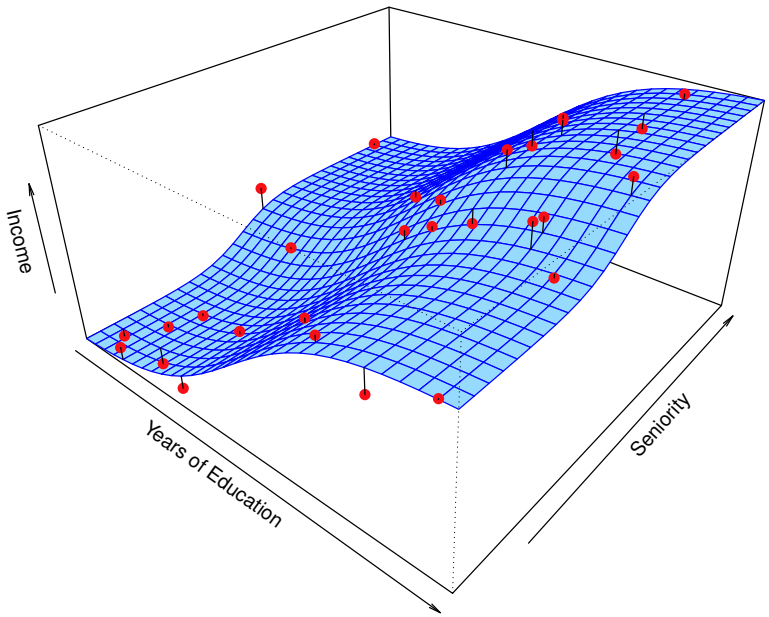


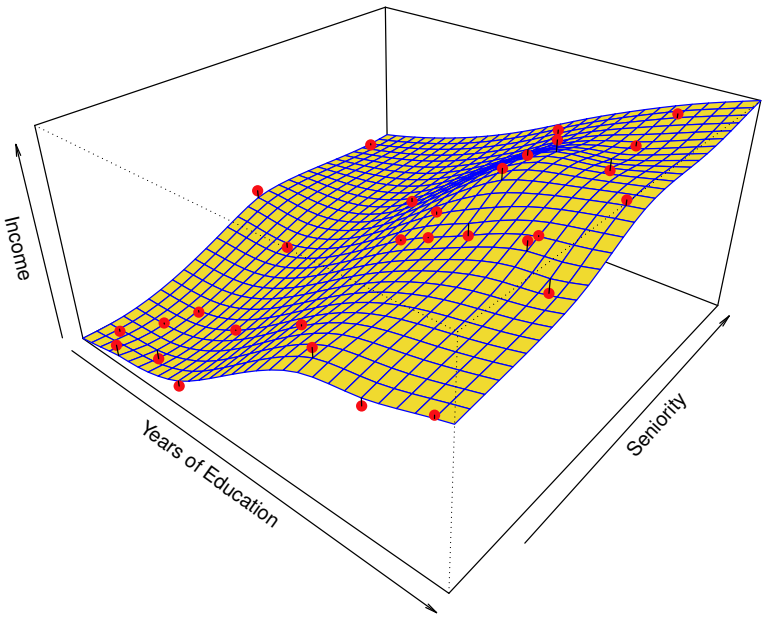
Non-parametric Methods

- Do not make explicit assumptions about the functional form of f.
- They seek an estimate of f that gets as close to the data points as possible without being too rough or wiggly.
- No need to worry about the assumption.
- Accurately fit a wider range of possible shapes for f.
- Need a very large number of observations!

An example of the non-parametric approach - Income data

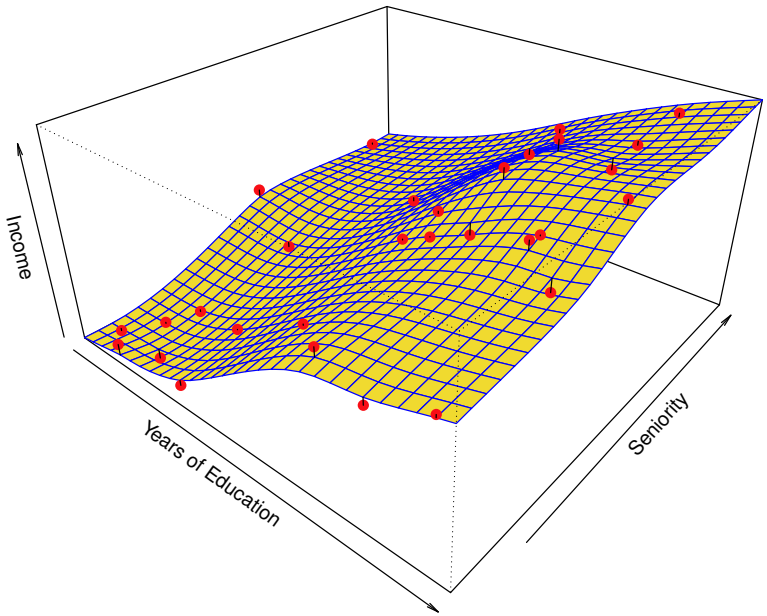
A *thin-plate spline* is used to estimate f. It attempts to produce an estimate of f that is as close as possible to the observed data, subject to the fit being *smooth*.

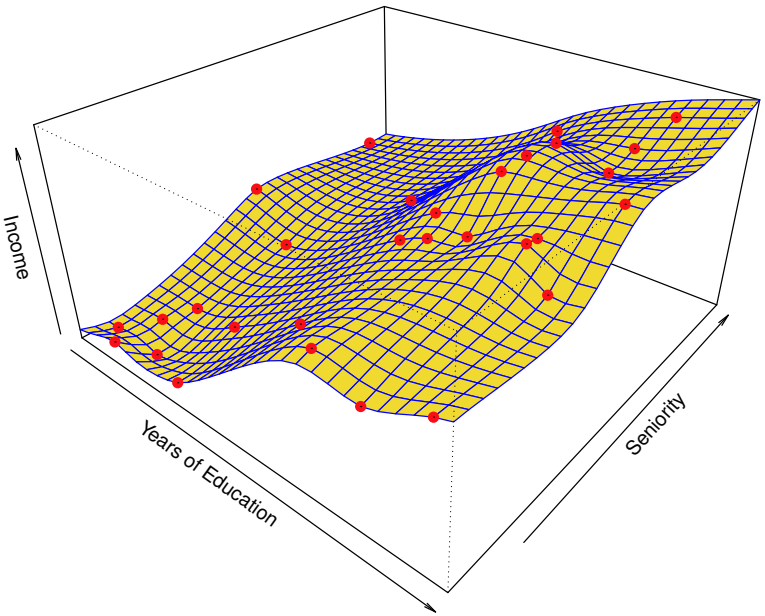




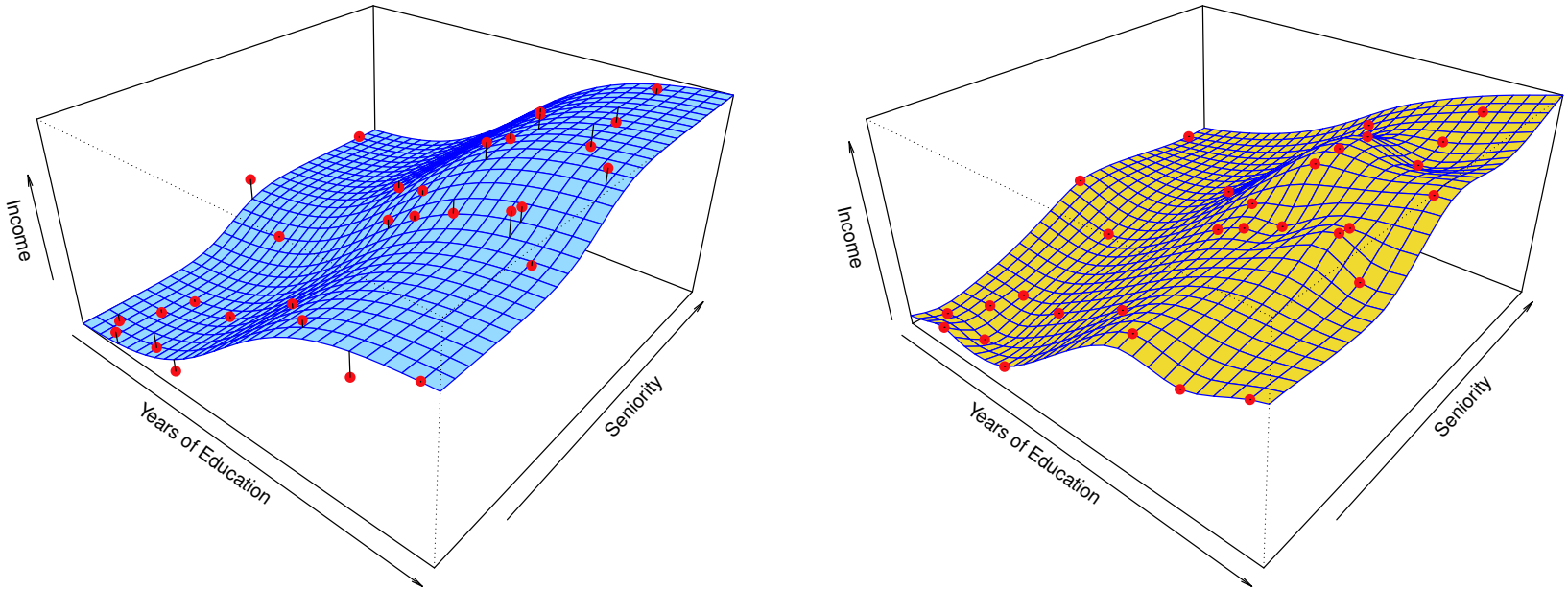
How smooth?

Select a level of smoothness.





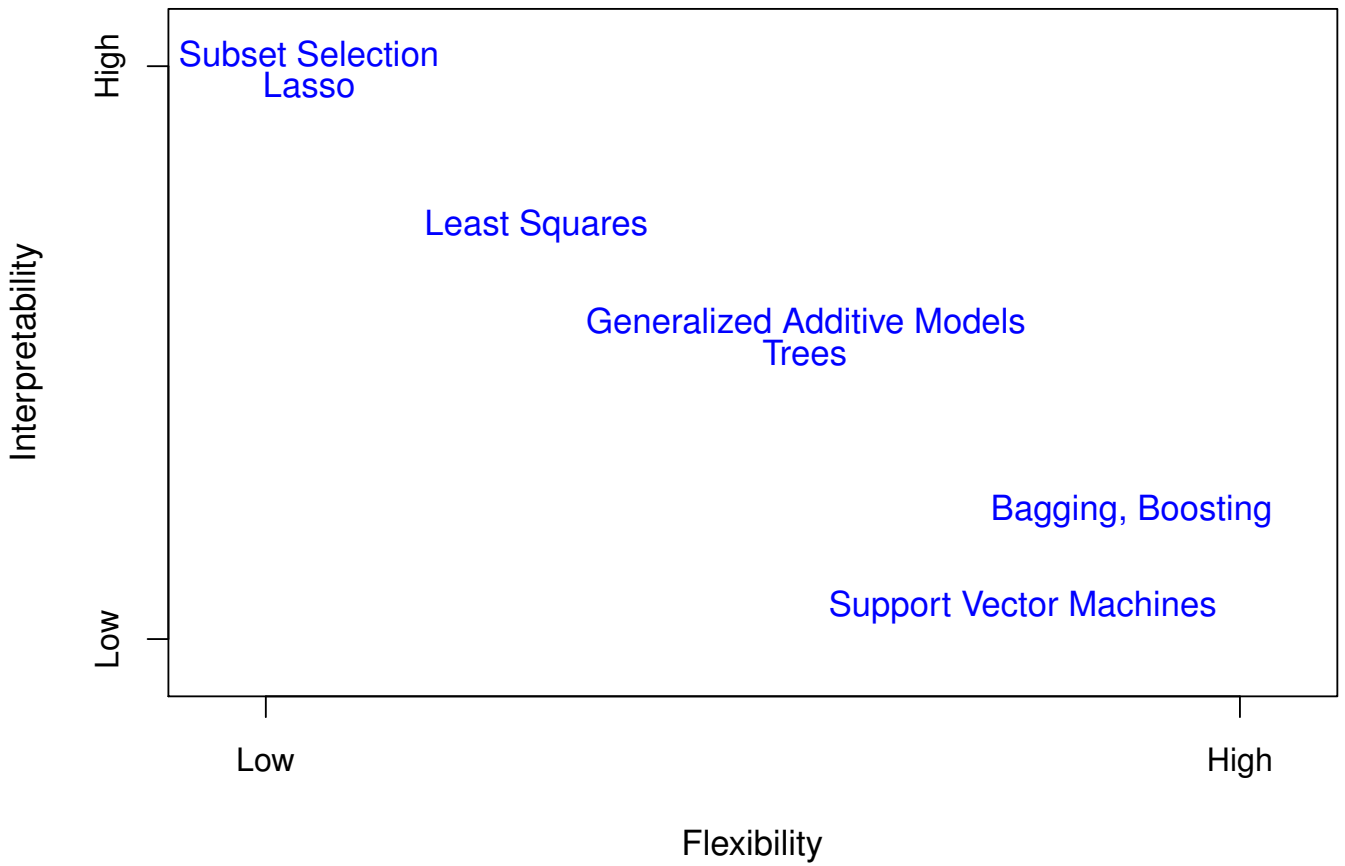
However...



Far more variable than the true function $f \rightarrow$ overfitting

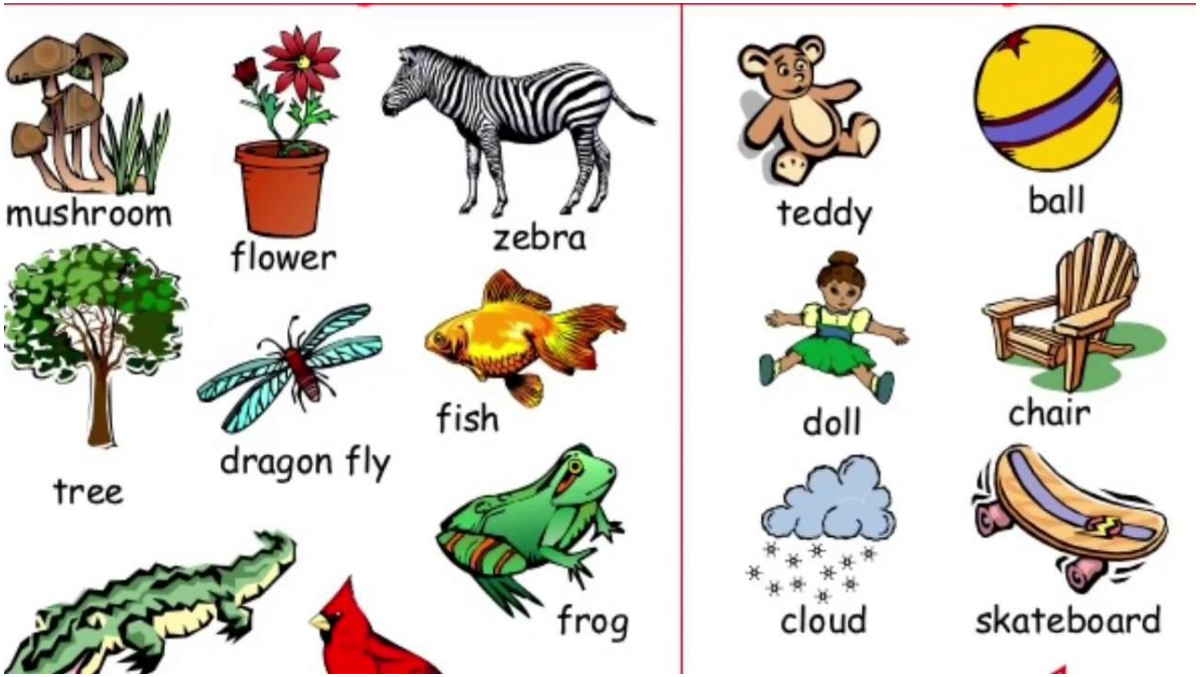
How to choose the *correct* amount of smoothness? Splines?

The Trade-Off Between Accuracy and Interpretability



Supervised and unsupervised learning

Supervised and unsupervised learning

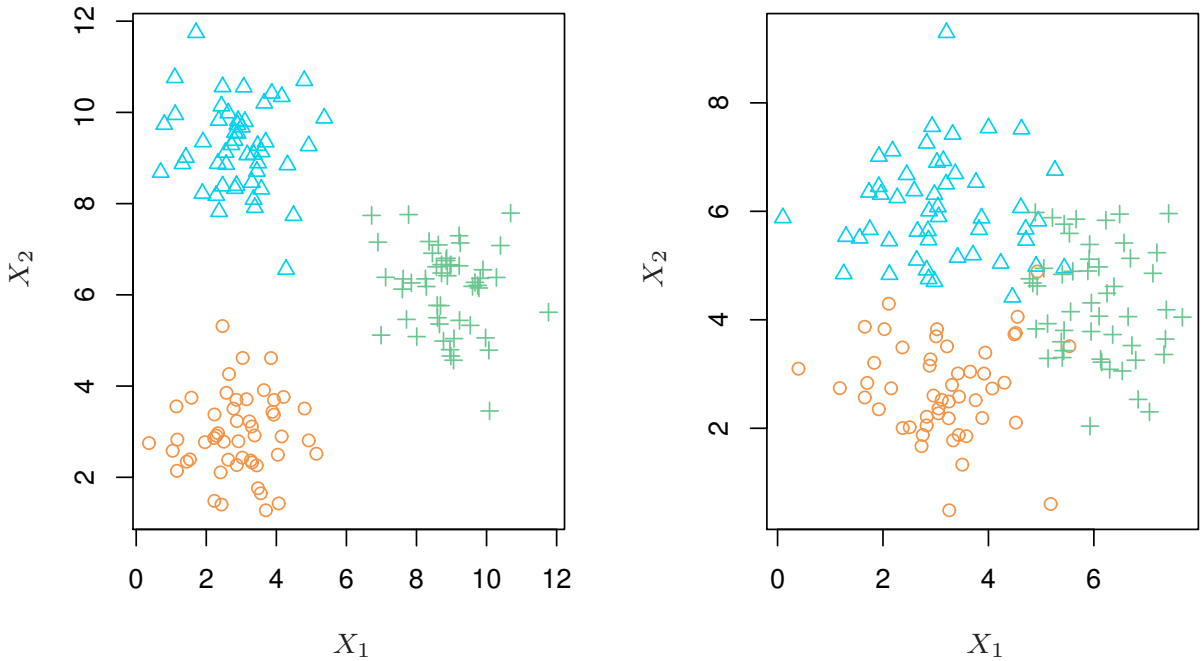


Supervised and unsupervised learning

- Supervised learning
  - outcome (quantitative/categorical) $\Leftarrow$ a set of features
  - a training set of data $\rightarrow$ observe outcome and features $\rightarrow$ build a prediction model
  - use this model to predict the outcome for new unseen objects
- Unsupervised learning
  - no outcome; observe only the features
  - describe how the data are organized or clustered

Supervised Versus Unsupervised Learning

- Supervised: linear regression, logistic regression, GAM, boosting, and support vector machine.
- Unsupervised: cluster analysis, principal component analysis.



Statistical Learning v.s. Machine Learning?

What are the differences???

Assessing Model Accuracy

Assessing Model Accuracy

- Why so many?
- No free lunch in statistics...
- No one method dominates all others over all possible data sets.
- Select the best approach for a given data set.

Measuring the Quality of Fit

- Quantify the extent to which the predicted response value for a given observation is close to the true response value for that observation.
- Mean squared error (MSE)

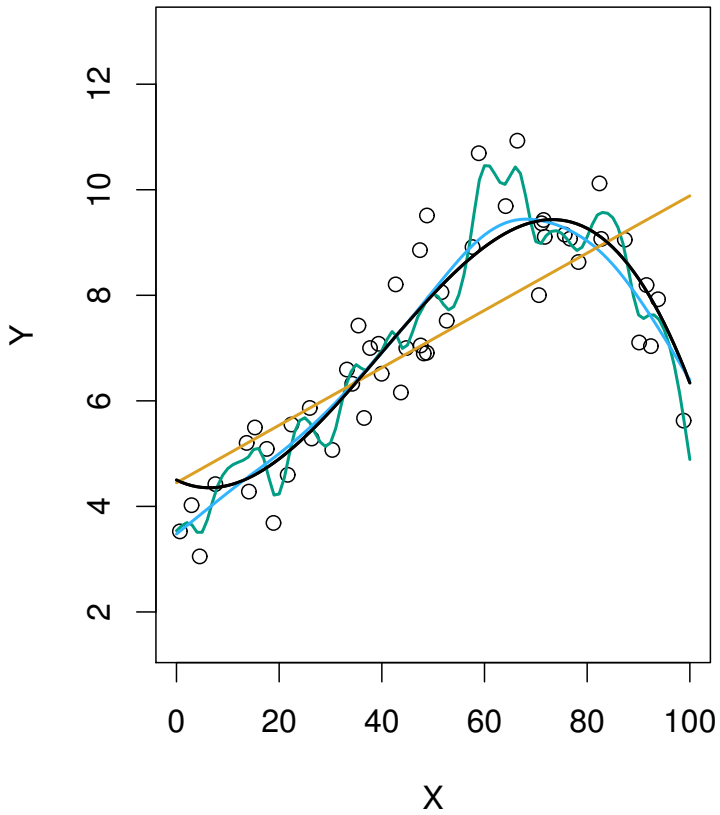
$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

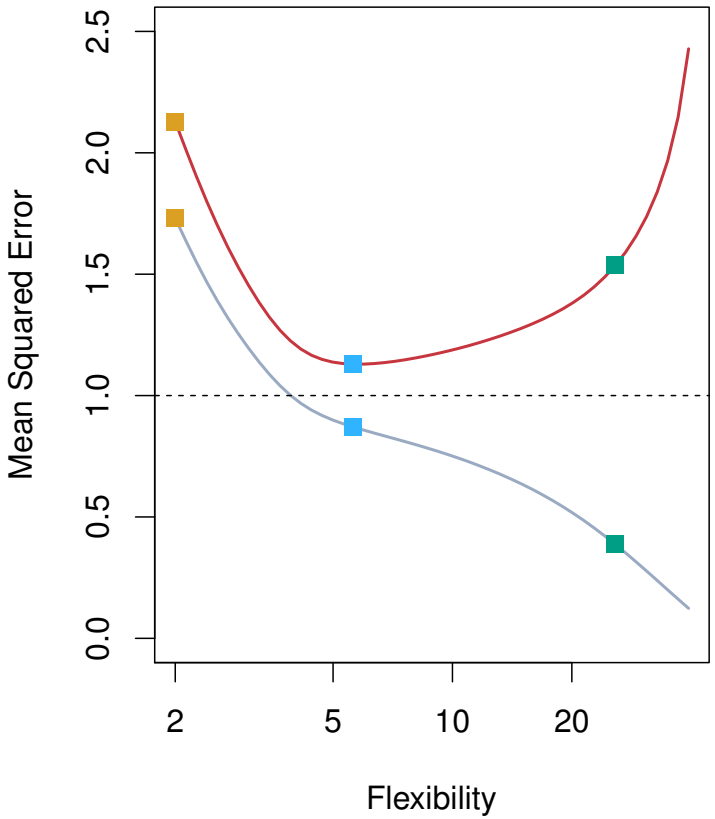
- Training MSE versus test MSE
  - Risk of diabetes; Stock's price
- Interested:

$$min\ \text{Ave}(y_0 - \hat{f}(x_0))^2$$

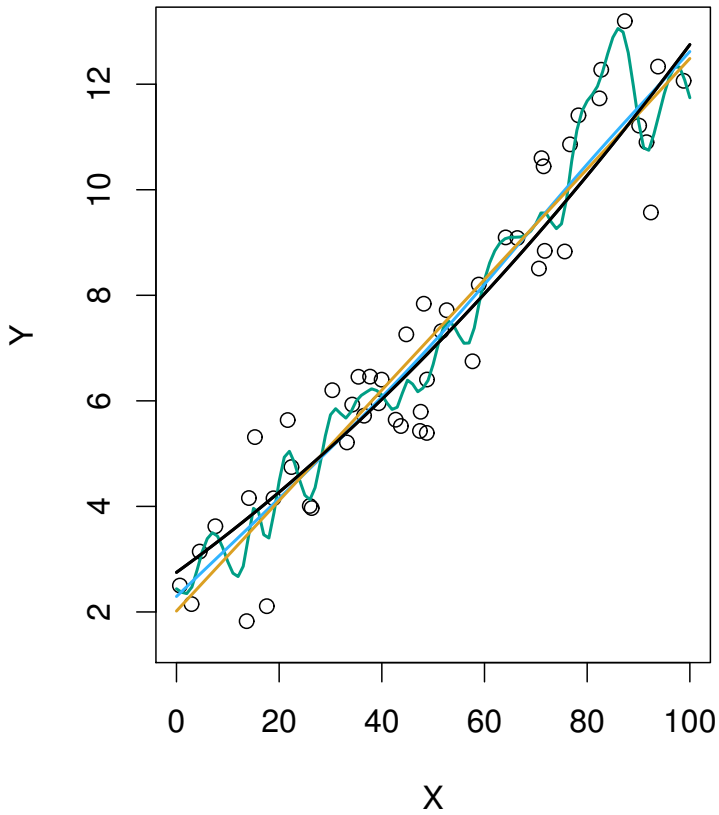
where x_0 and y_0 are previously unseen test observations not used to train the statistical learning method.

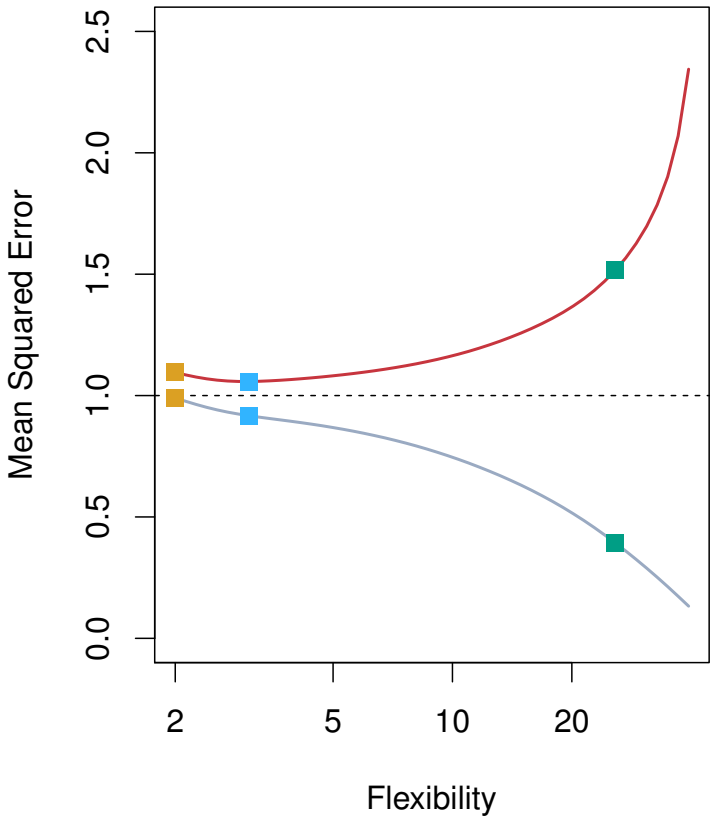
Relationship between training MSE and test MSE



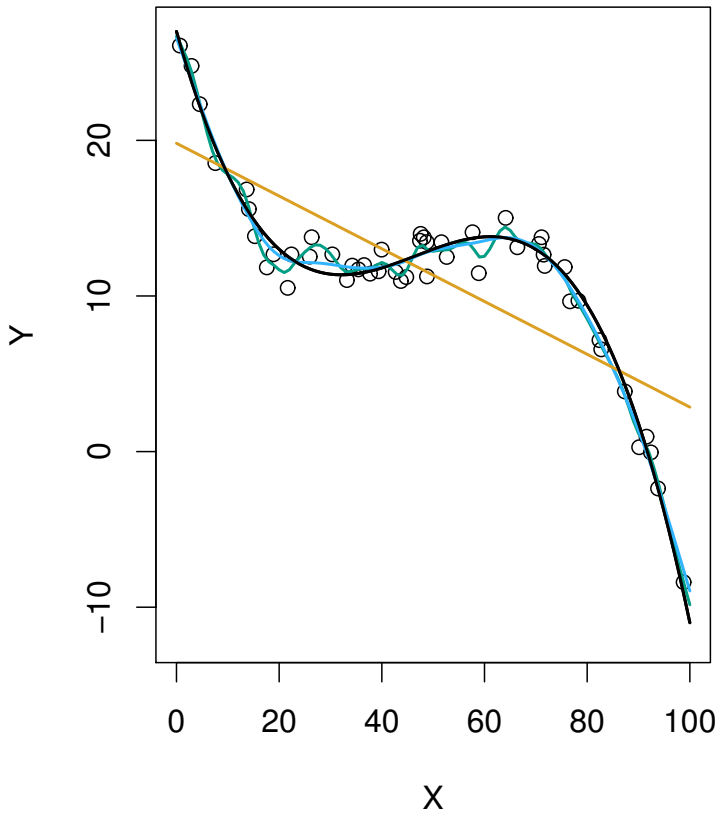


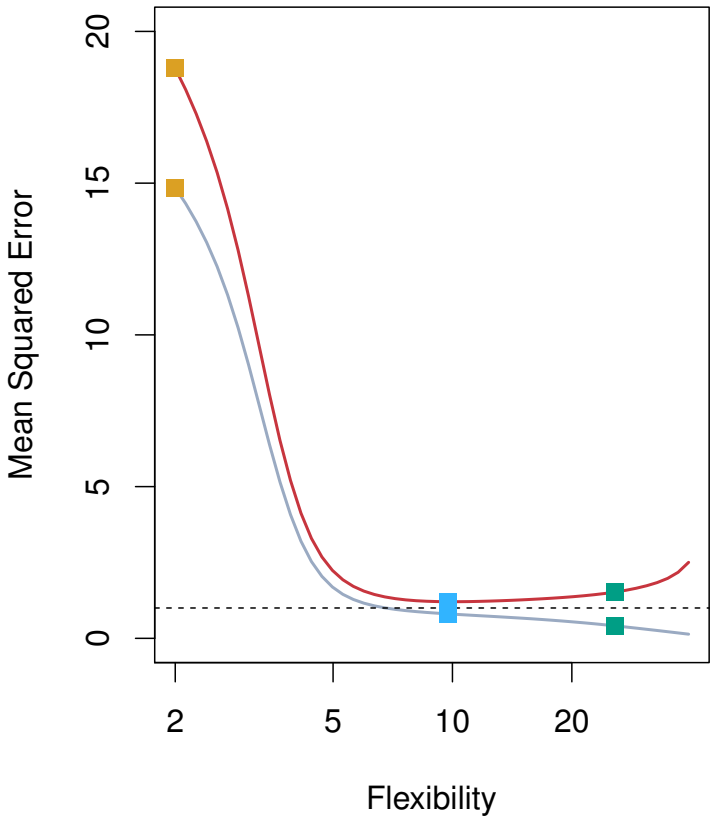
Not necessarily U-shape





Not necessarily U-shape





Two competing properties: bias versus variance

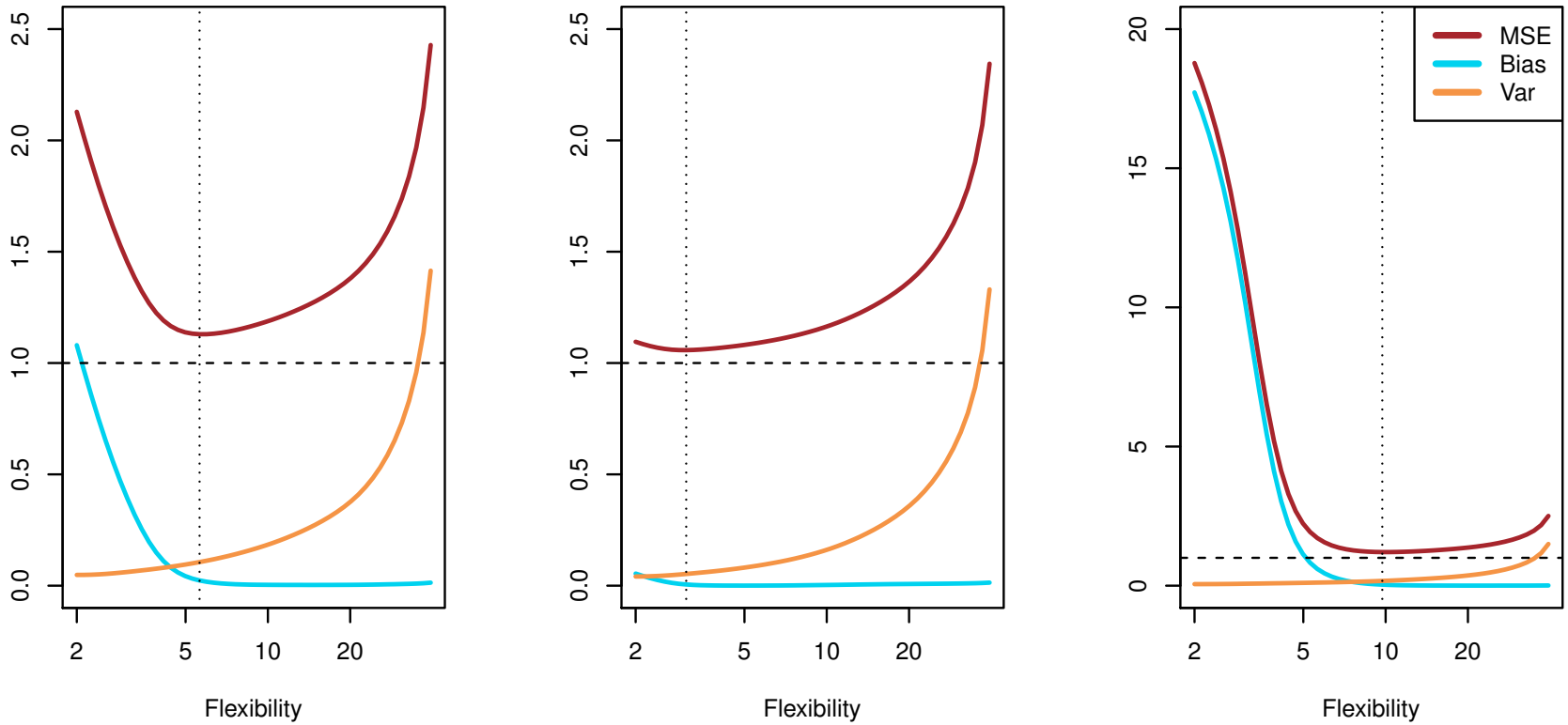
- The expected test MSE:

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

- *Variance* refers to the amount by which $\hat{f}$ would change if we estimated it using a different training data set.
- *Bias* refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model.

The relationship between bias, variance, and test set MSE

The bias-variance trade-off



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.
- How to quantifying the accuracy of our estimate $\hat{f}$ now?
- The training *error rate*

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

  $\hat{y}_i$ is the predicted class label and $I(y_i \neq \hat{y}_i)$ is an indicator variable.
- The test error

$$\text{Ave}(I(y_0 \neq \hat{y}_0))$$