Why Do We Estimate f?

- Statistical Learning, and this course, are all about how to estimate f.
- The term statistical learning refers to using the data to "learn" f.

- Why do we care about estimating f?
- There are 2 reasons for estimating f,
  - **Prediction** and
  - **Inference.**

1. Prediction

- If we can produce a good estimate for f (and the variance of ε is not too large) we can make accurate predictions for the response, Y, based on a new value of **X**.

Example: Direct Mailing Prediction

- Interested in predicting how much money an individual will donate based on observations from 90,000 people on which we have recorded over 400 different characteristics.
- Don't care too much about each individual characteristic.
- Just want to know: For a given individual should I send out a mailing?

2. Inference

- Alternatively, we may also be interested in the type of relationship between Y and the X's.
- For example,
  - Which particular predictors actually affect the response?
  - Is the relationship positive or negative?
  - Is the relationship a simple linear one or is it more complicated etc.?

Example: Housing Inference

- Wish to predict median house price based on 14 variables.
- Probably want to understand which factors have the biggest effect on the response and how big the effect is.
- For example how much impact does a river view have on the house value etc.

Tradeoff Between Prediction Accuracy and Model Interpretability

- Why not just use a more flexible method if it is more realistic?
- There are two reasons

Reason 1:

A simple method such as linear regression produces a model which is much easier to interpret (the Inference part is better). For example, in a linear model, β_j is the average increase in Y for a one unit increase in X_j holding all other variables constant.

Reason 2:

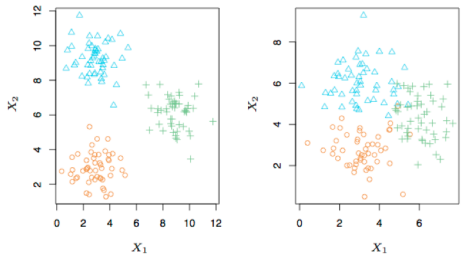
Even if you are only interested in prediction, so the first reason is not relevant, it is often possible to get more accurate predictions with a simple, instead of a complicated, model. This seems counter intuitive but has to do with the fact that it is harder to fit a more flexible model.

Supervised vs. Unsupervised Learning

- We can divide all learning problems into Supervised and Unsupervised situations
- Supervised Learning:
  - Supervised Learning is where both the predictors, $\mathbf{X}_i$, and the response, Y_i, are observed.
  - This is the situation you deal with in Linear Regression
  - Most of this course will also deal with supervised learning.

- Unsupervised Learning:
  - In this situation only the $\mathbf{X}_i$'s are observed.
  - We need to use the $\mathbf{X}_i$'s to guess what Y would have been and build a model from there.
  - A common example is market segmentation where we try to divide potential customers into groups based on their characteristics.
  - A common approach is clustering.

A Simple Clustering Example



Regression vs. Classification

- 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.
- 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?

Bias/ Variance Tradeoff

- The previous graphs of test versus training MSE's illustrates a very important tradeoff that governs the choice of statistical learning methods.
- There are always two competing forces that govern the choice of learning method i.e. bias and variance.

Bias of Learning Methods

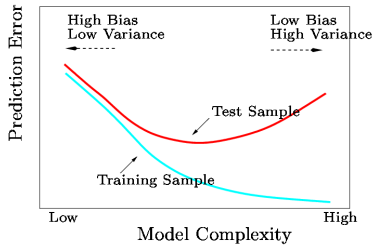
- Bias refers to the error that is introduced by modeling a real life problem (that is usually extremely complicated) by a much simpler problem.
- For example, linear regression assumes that there is a linear relationship between Y and X. It is unlikely that, in real life, the relationship is exactly linear so some bias will be present.
- The more flexible/complex a method is the less bias it will generally have.

Variance of Learning Methods

- Variance refers to how much your estimate for f would change by if you had a different training data set.
- Generally, the more flexible a method is the more variance it has.

A Fundamental Picture

- In general training errors will always decline.
- However, test errors will decline at first (as reductions in bias dominate) but will then start to increase again (as increases in variance dominate).



We must always keep this picture in mind when choosing a learning method. More flexible/complicated is not always better!