WHAT IS STATISTICAL LEARNING?

Chapter 02 – Part I

Slides Inspired by content from IOM 530 “Applied Modern Statistical Learning Methods” – Gareth James (one of the authors of our book)

Outline

- What Is Statistical Learning?
  - Why estimate f?
  - How do we estimate f?
  - The trade-off between prediction accuracy and model interpretability
  - Supervised vs. unsupervised learning
  - Regression vs. classification problems

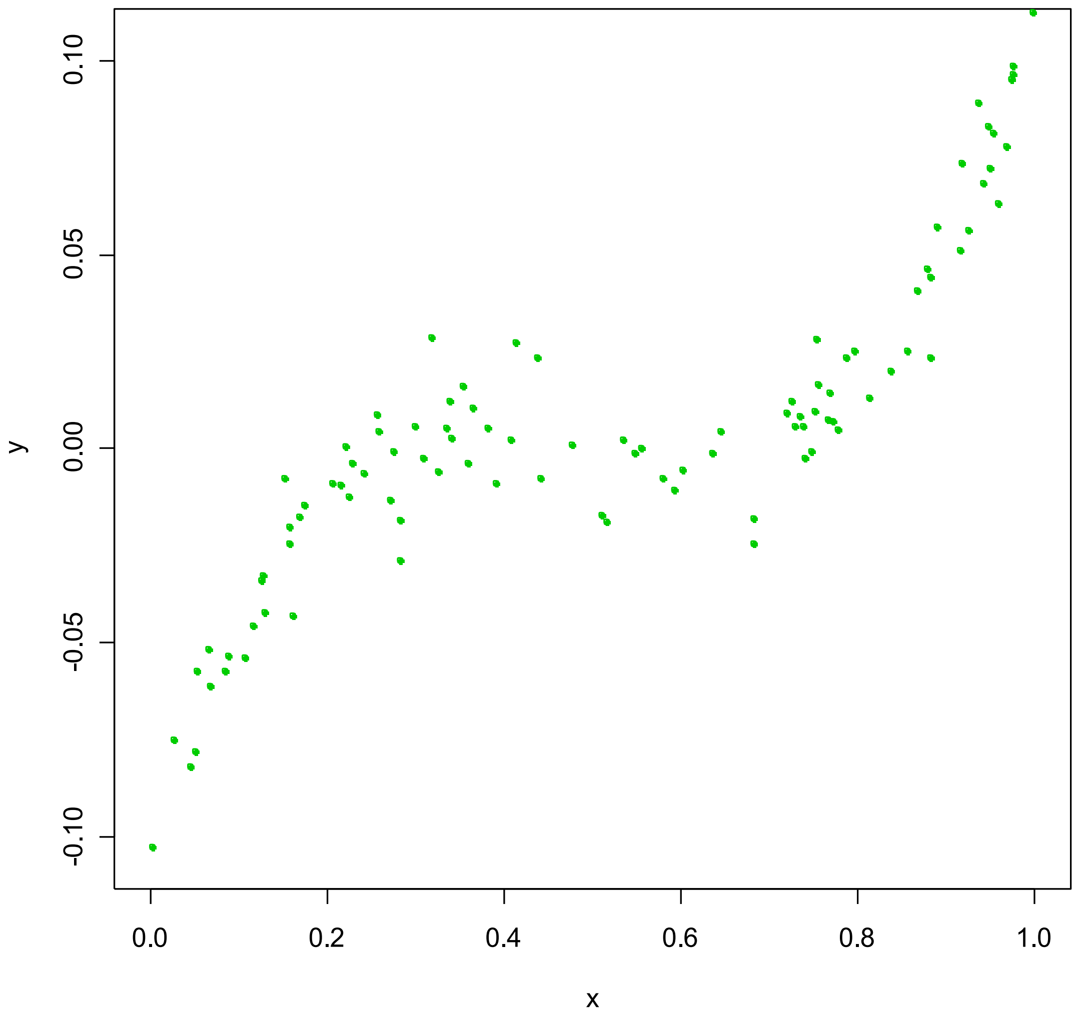
What is Statistical Learning?

- Suppose we observe Y_i and $X_i = (X_{i1},\ldots,X_{ip})$ for $i = 1,\ldots,n$
- We believe that there is a relationship between Y and at least one of the X's.
- We can model the relationship as

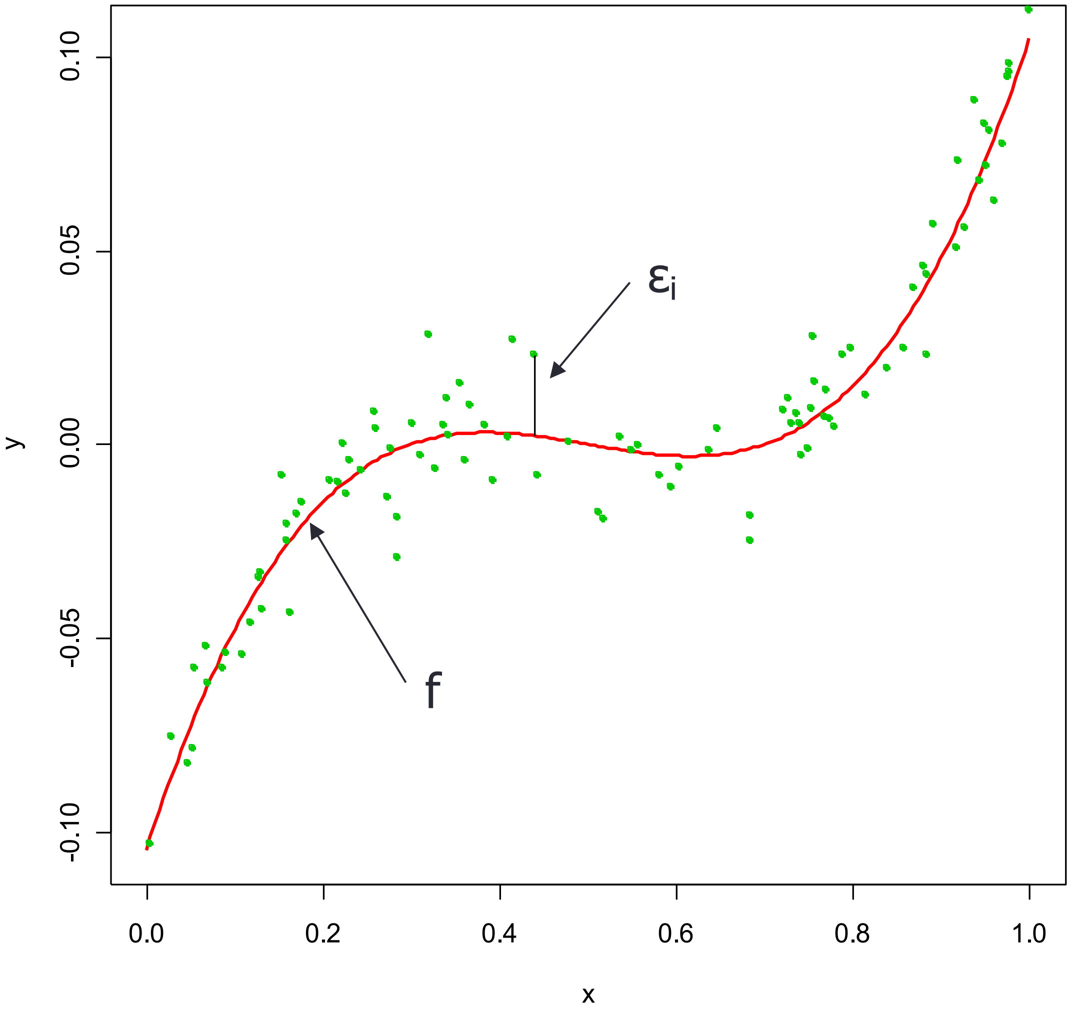
$$Y_i = f(\mathbf{X}_i) + \varepsilon_i$$

- Where f is an unknown function and ε is a random error with mean zero.

A Simple Example



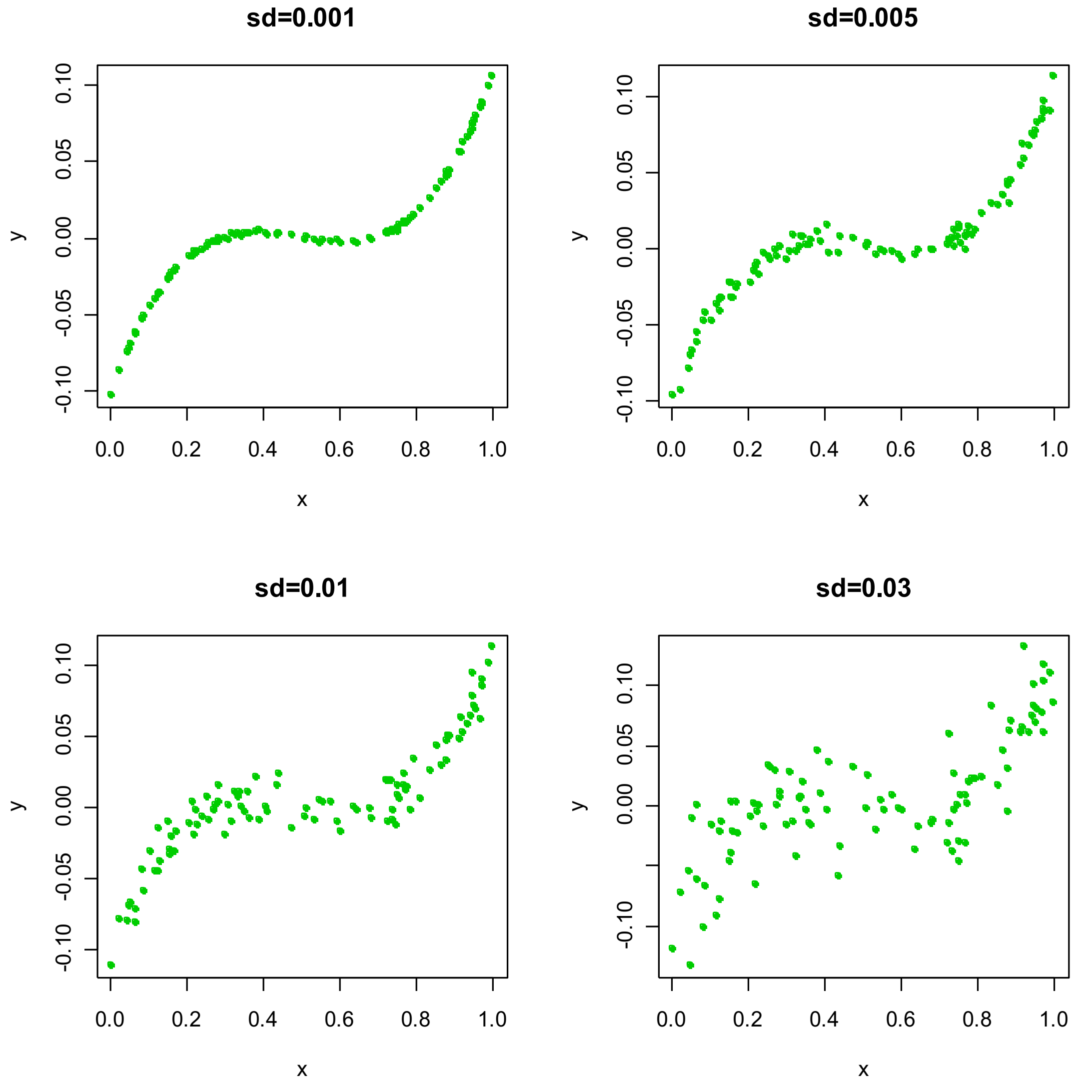
A Simple Example $Y_i = f(\mathbf{X}_i) + \varepsilon_i$



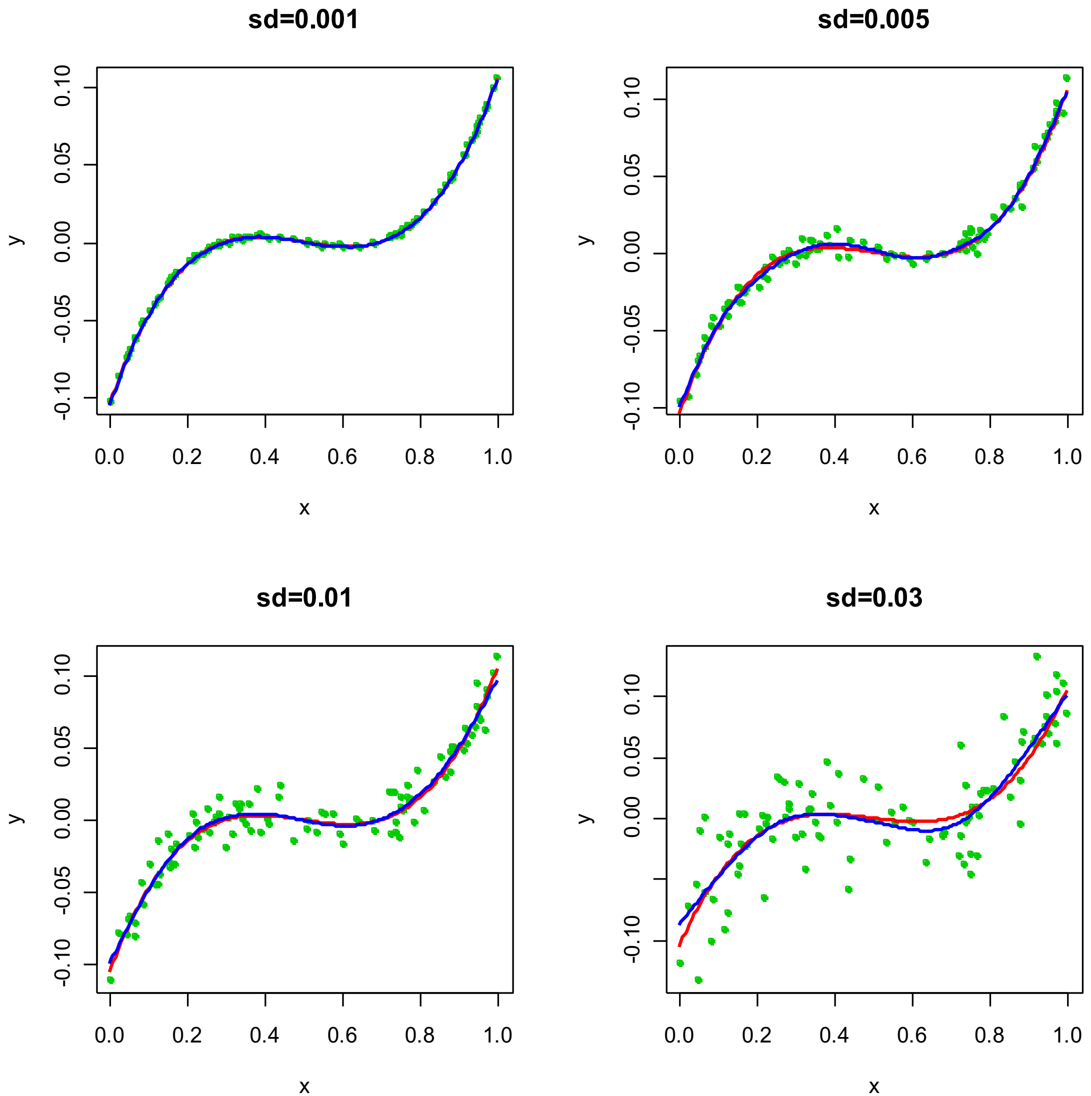
Different Standard Deviations

- The difficulty of estimating f will depend on the standard deviation of the ε's.

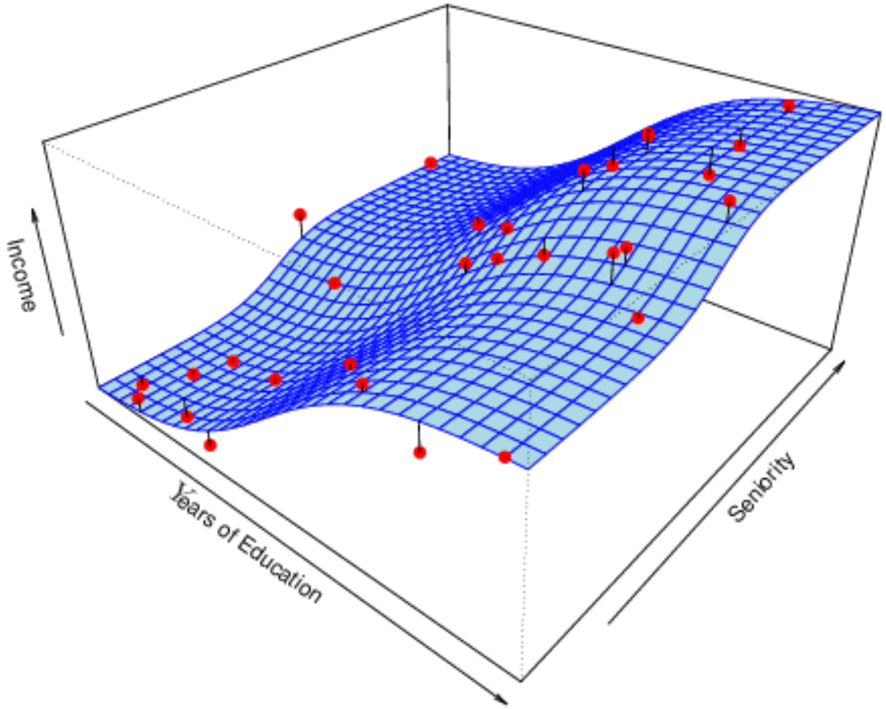
$$Y_i = f(\mathbf{X}_i) + \varepsilon_i$$



Different Estimates For *f*



Income vs. Education & Seniority



- Shown above is the “true” relationship between the variables Years of Education, Seniority, and Income.
- **CONCEPT CHECK: Describe the relationship between income, years of education and seniority that you see here**

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 (Estimation)**
  - **Inference (Explanation)**

Prediction (Estimation)

- 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 $\boldsymbol{X}$.

Prediction / Estimation Example: Direct Mailing Decision

- How much money an individual will donate to a charity?
- Data:
  - ***X***: 400 characteristics about each person
  - *Y*: How much they donated.
- Business Question: For a given individual should I send out a mailing?
  - Is the expected value of taking the action greater than the cost of the action?
- Assume that there is no desire to know what features are associated with people who contribute.

Inference (Explanation)

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

Inference Example: Understanding Home prices

- How do characteristics affect home prices
- Housing data:
  - **X**: 14 characteristics (e.g. number of beds; baths; square feet)
  - Y: cost of the home
- Business Question:
  - How would altering the variables affect my home's value?
- For example
  - Would installing an in-ground pool increase my home's value?
  - What is the financial impact of turning my 1-car garage into a woodworking shop?
  - What is the most cost-effective thing I could do to improve my home's value before I sell it?

How Do We Estimate f?

- We will assume we have observed a set of **training data**

$$\{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \ldots, (\mathbf{X}_n, Y_n)\}$$

- We must then use the training data and a statistical method to estimate f.
- Statistical Learning Methods:
  - Parametric Methods
  - Non-parametric Methods

$|\text{Features}| = m$ (or p)

$|\text{Observations}| = n$

	$\mathbf{X}_{:,1}$	$\mathbf{X}_{:,2}$	…	$\mathbf{X}_{:,m}$	Target Label
$\mathbf{X}_1$					Y_1
$\mathbf{X}_2$					Y_2
…					…
$\mathbf{X}_n$					Y_n

Parametric Methods

- Reduces the problem of estimating f to estimating a set of parameters.

- Two-step model based approach

<u>STEP 1:</u>
Make some assumption about the functional form of f, i.e. come up with a model. For example, a linear model:

$$f(\mathbf{X}_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 \cdot X_{i2} + \ldots + \beta_p X_{ip}$$

Parametric Methods (cont.)

STEP 2:

Use the training data to fit the model i.e. estimate f or equivalently the unknown parameters such as β_0, β_1, β_2,…, β_p.

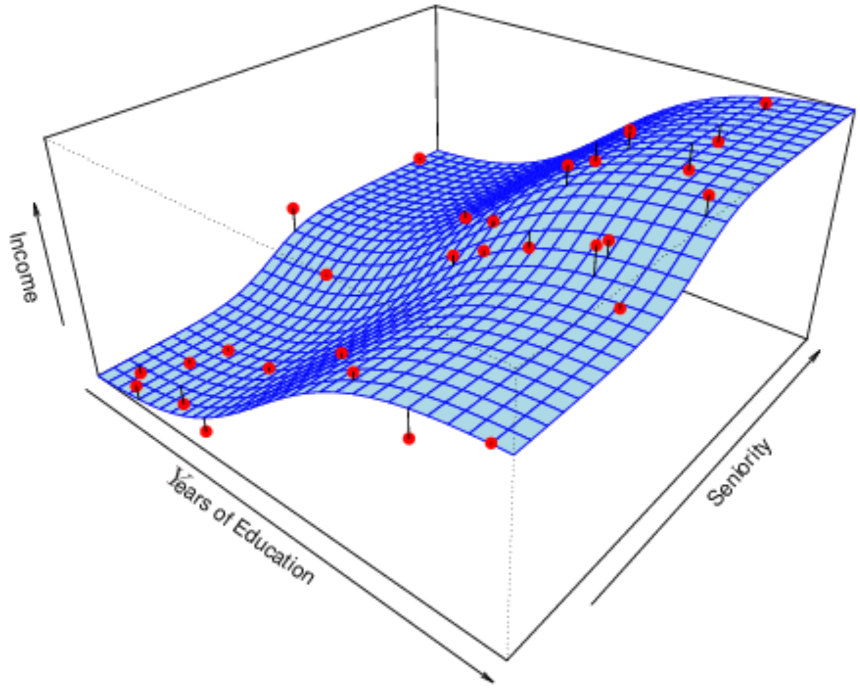
Individual prediction error terms can be computed:

$$\varepsilon_i = Y_i - f(\mathbf{X}_i)$$

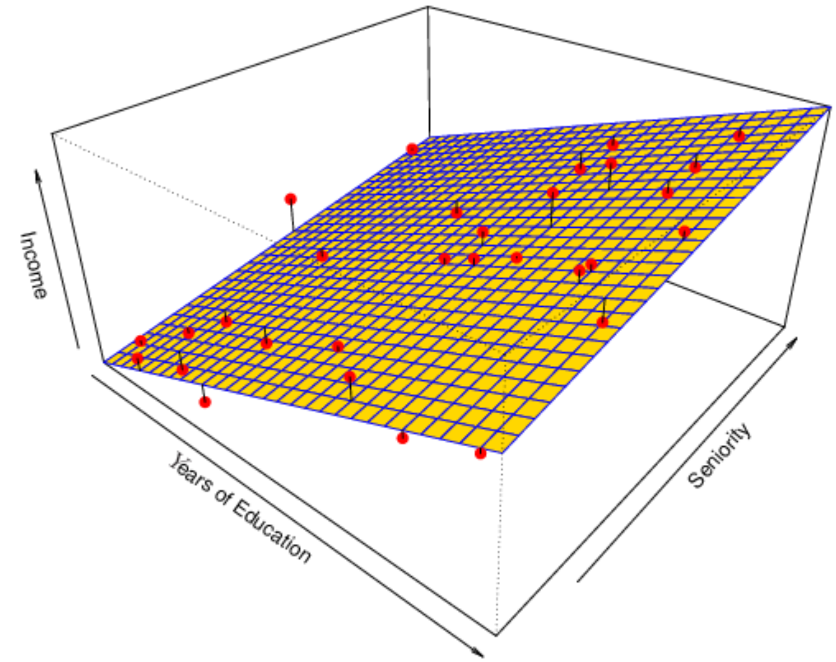
- One common approach for estimating the parameters in a linear model is ordinary least squares (OLS) – which minimizes the square of the sum of the error terms.
  - Has limitations due to computational tractability of inverting a matrix
- That there are other approaches

META: Why do you think we would want to use OLS in a linear model? (hint – why is *squaring* the error terms mathematically important?)

Parametric Example: Linear Regression



True Phenomenon



Linear Model Approximation

$$f = \beta_0 + \beta_1 \times Education + \beta_2 \times Seniority$$

In-Class coding exercise

- This exercise will explore 1-dimensional linear regression (one feature is used to predict the target variable)
- You will manually fit the 2 parameter model by guessing and testing different betas until you get a low error

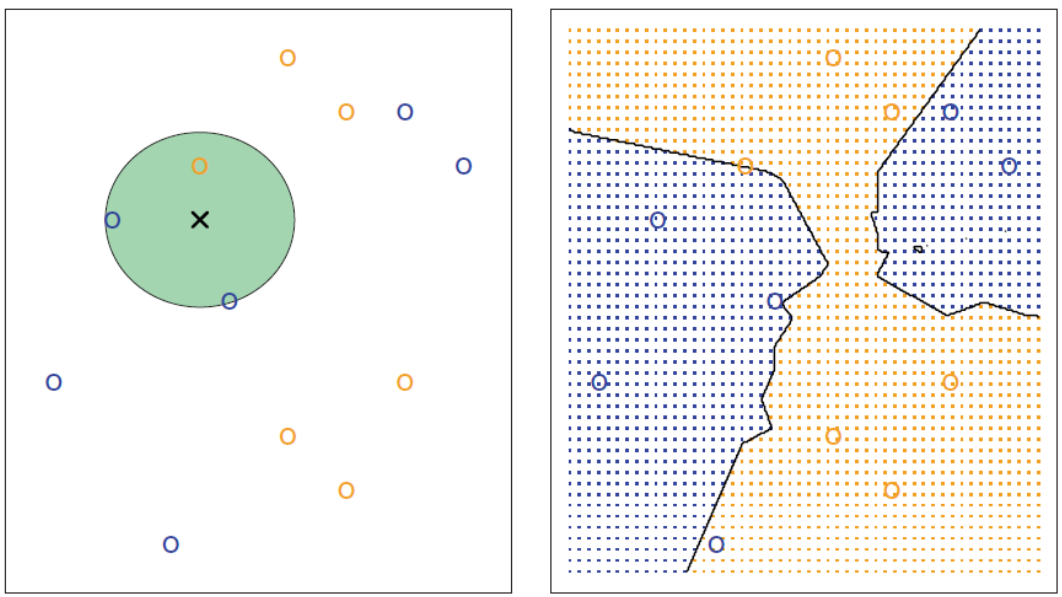
- Directions:
  - Obtain the python starter code from canvas (“in class” tab)
  - Read the directions for
    “Simple Linear Regression as Matrix Algebra, part 1”
  - Complete the steps 1 – 6 individually – ask your neighbor or the instructor for help if needed

Non-parametric Methods

- Non-parametric methods do not make explicit assumptions about the functional form of f
  - There is no parametric model, and no model parameters are fit from the data during the training process
  - Instead, (some) data observations from the training set are stored and used (directly) during prediction
- Advantages: accurately fit a wider range of possible f
- Disadvantages: Slower to train; Risk of overfitting; A very large number of observations are required to obtain an accurate estimate of f

Non-parametric Example: K-Nearest Neighbors

- Datapoints from the training set are stored
- A new datapoint's membership depends on what training set members it is close to (k members are considered)



ISLR page 40 / Figure 2.14

Model Flexibility

- Flexibility refers to a model's capacity to represent a complex mapping between the underlying data and the target variable
- Low flexibility models make the assumption that the relationship between the data and the target variable is simpler (e.g. linear)
- Higher-flexibility models allow for more elaborate relationships (e.g. polynomial)

Model Flexibility Tradeoff (1/2)

- Why not just use a more flexible method if it has a higher capacity?

Reason 1: Interpreting is easier with less flexible model

A simple method such as linear regression produces a model which is much easier to interpret (Inference is easier). 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.

Model Flexibility Tradeoff (2/2)

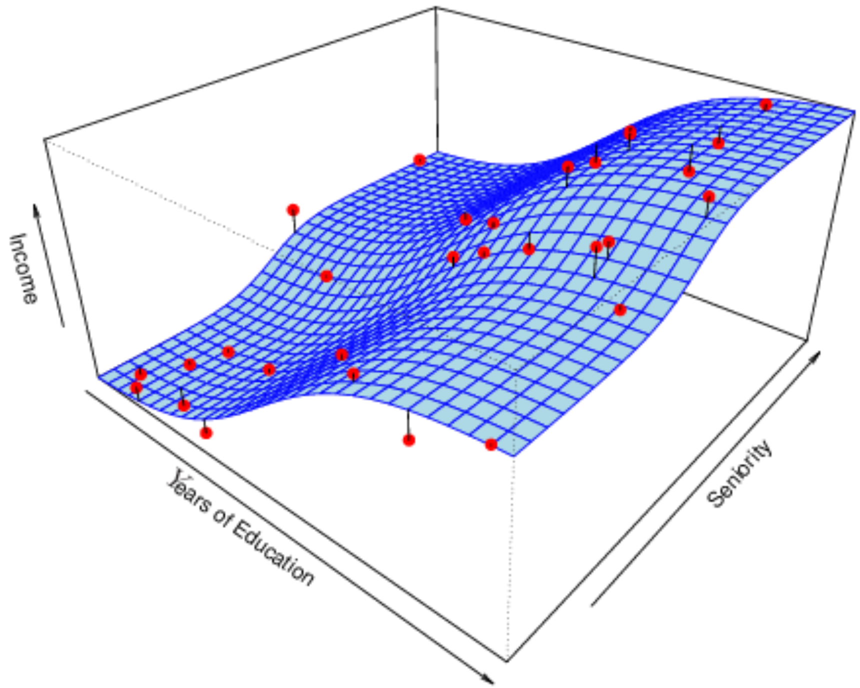
- Why not just use a more flexible method if it has a higher capacity?

Reason 2: Risk of overfitting during training

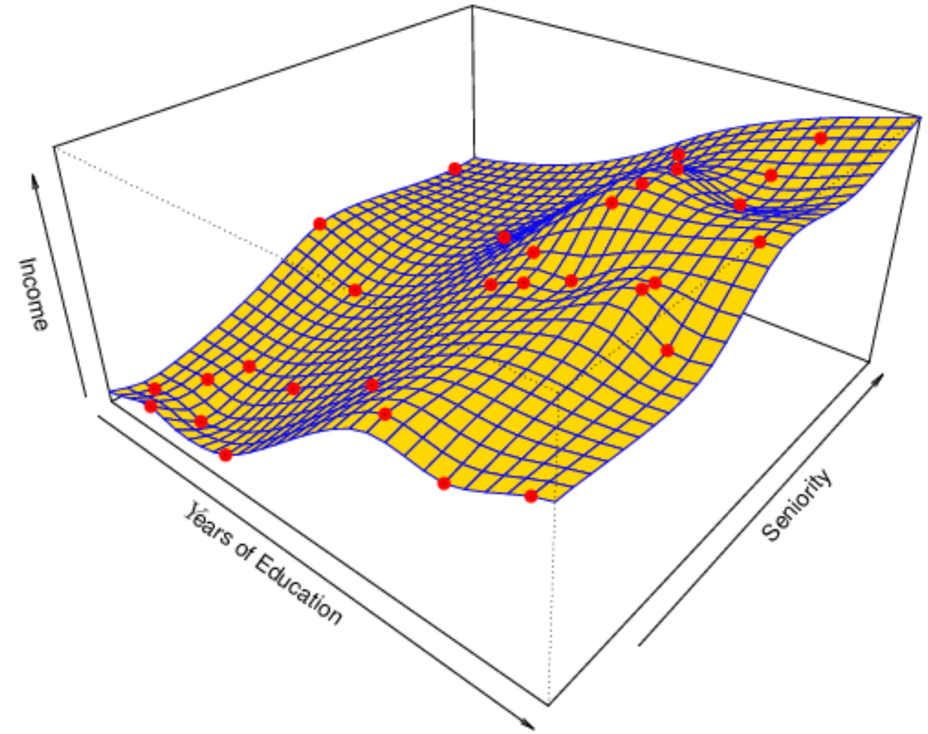
When data availability is limited, it is often possible to get more generalizable predictions with a simple model… a complicated model requires more data to properly train. With less data the higher capacity model essentially replicates a lookup function.

Overfitting

- A model can be too flexible, and make poor estimates of f on unseen data. This is also known as failure to *generalize*



True Phenomenon in Blue



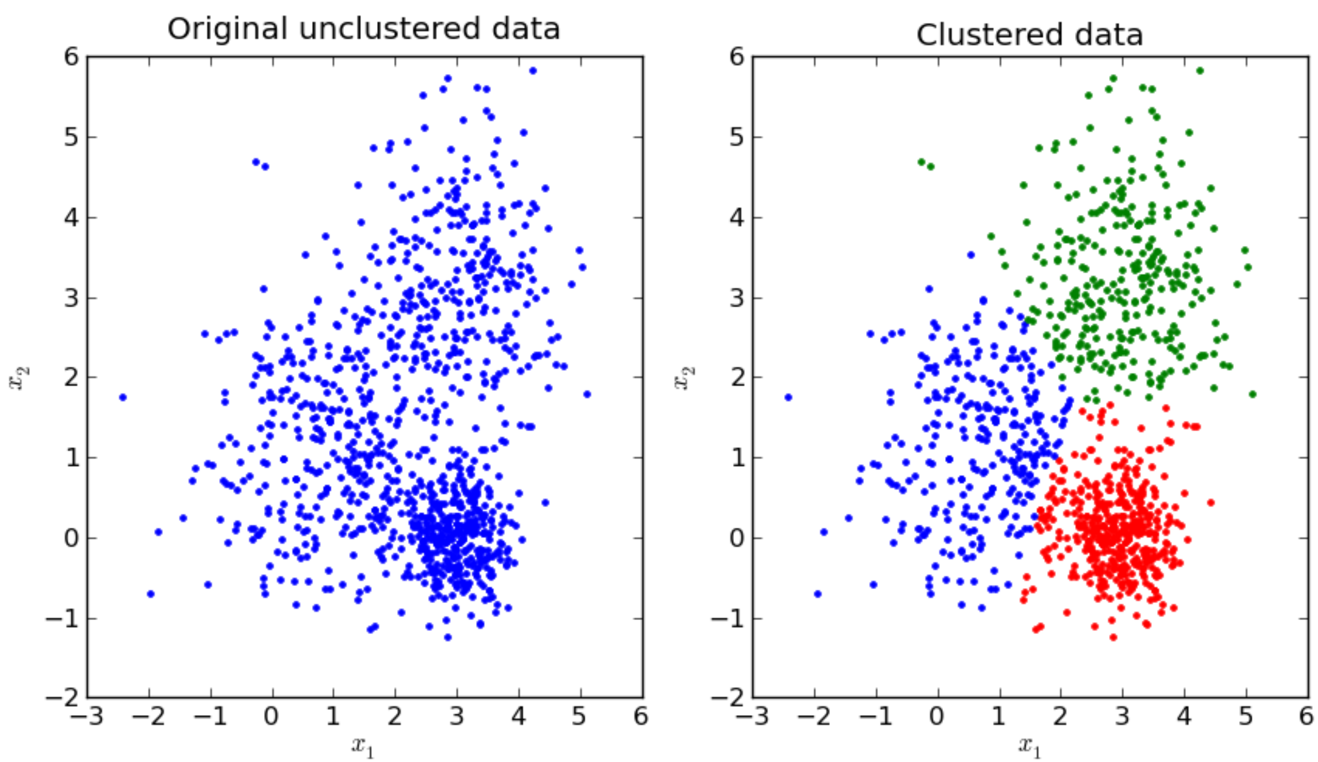
Overfit model
(Fitted to the noise in the data)

Supervised vs. Unsupervised Learning

- Supervised Learning:
  - Supervised Learning is where the predictors, $\mathbf{X}_i$, and the response, Y_i, are observed
  - Example task: Income prediction
  - Example technique: linear regression
- Unsupervised Learning:
  - Only the X_i's are observed.
  - We need to use the relationships among the X_i's to draw conclusions about the data
  - Example task: market segmentation - divide potential customers into groups based on their characteristics
  - Example technique: clustering

Clustering Example

- Clustering requires a distance measure to be defined for the data elements so that closeness can be determined in the original data feature space
- Clustering algorithms are sometimes evaluated using intra-member cohesion and inter-member separation



http://stackoverflow.com/questions/24645068/k-means-clustering-major-understanding-issue

Regression vs. Classification

- Supervised learning problems can be further divided into regression and classification problems.
- Regression: Y is continuous/numerical:
  - Predicting the value of a stock 6 months from today.
  - Predicting the price of a given house based on characteristics.
- Classification: Y is categorical:
  - Is this email SPAM or not?
  - Is this a picture of a cat, a dog, or a mouse?