

STATISTICAL LINEAR REGRESSION - PART I:

WHAT IS STATISTICAL LEARNING?

(Rafael Alcalá)

Bibliography:

Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani
An Introduction to Statistical Learning with Applications in R
Springer, 2013

Chapter 02

(Some of the figures in this presentation are taken from this book and some slides are based on Abbass Al Sharif's slides for his course DSO 530)

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

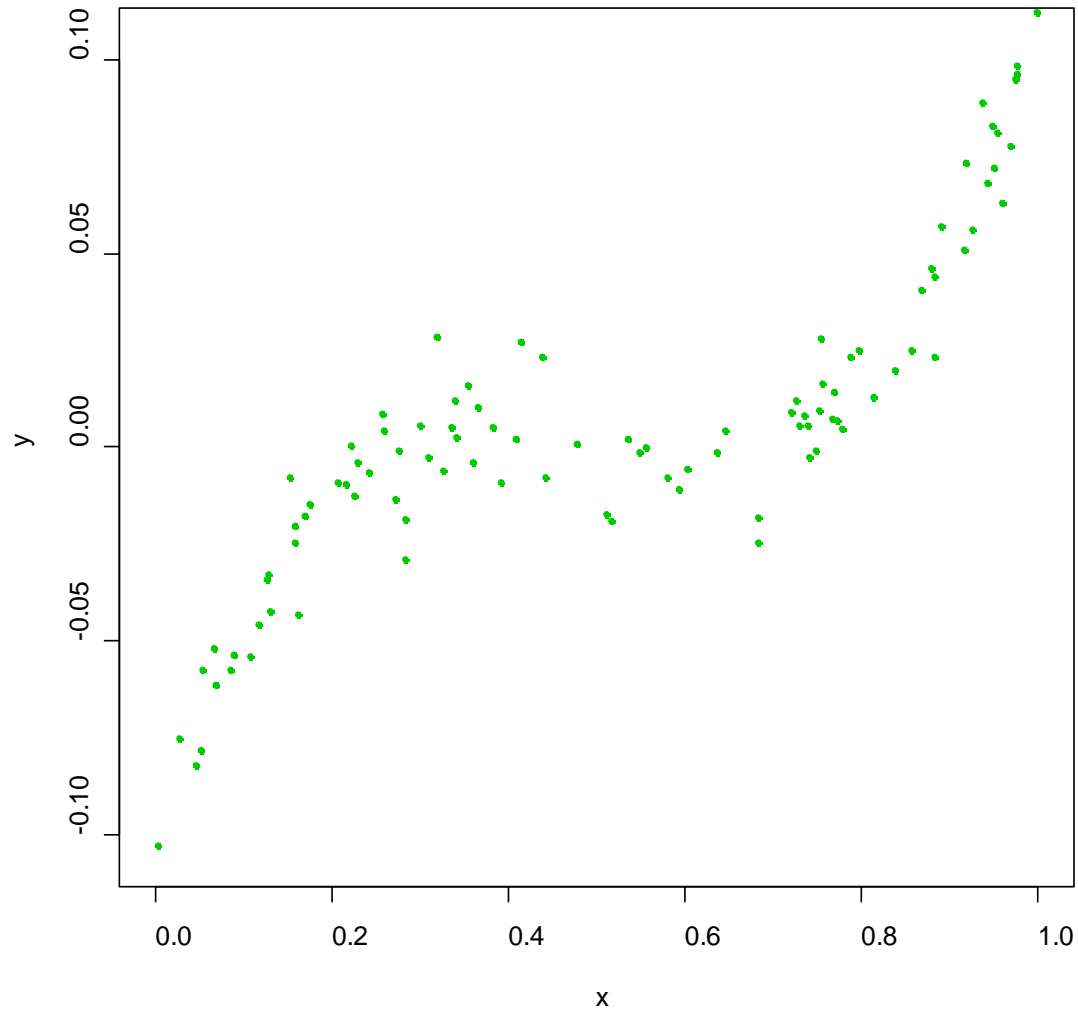
What is Statistical Learning?

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

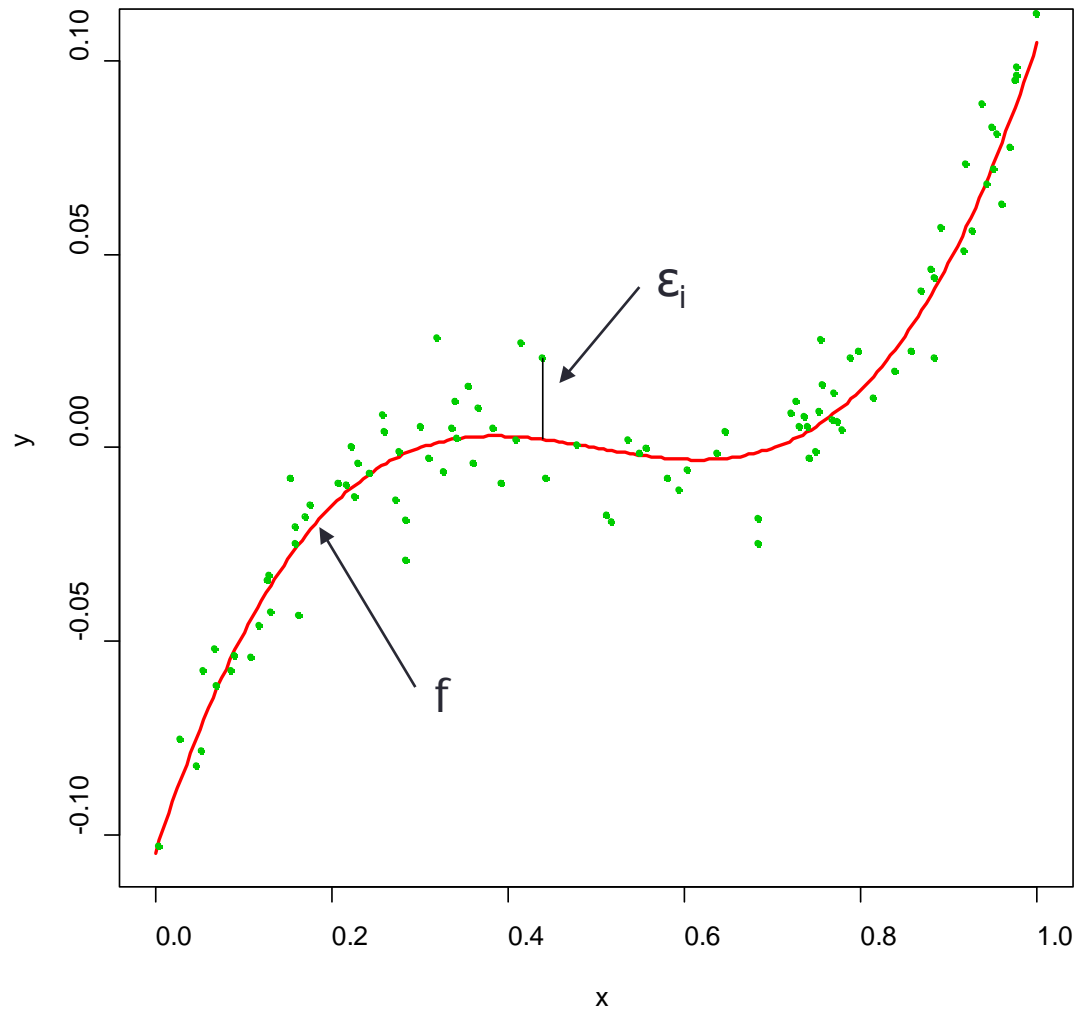
$$Y_i = f(\mathbf{X}_i) + \varepsilon_i$$

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

A Simple Example

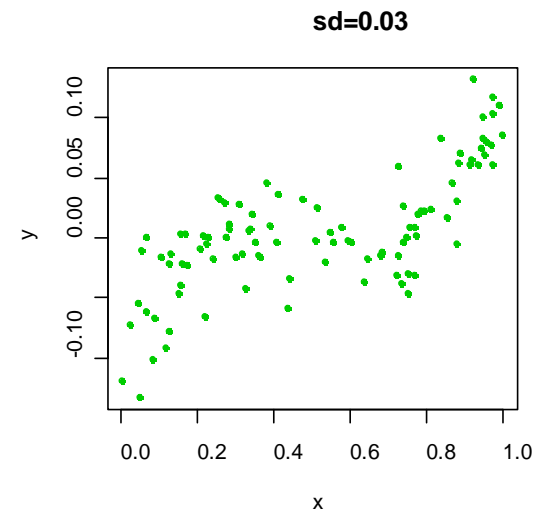
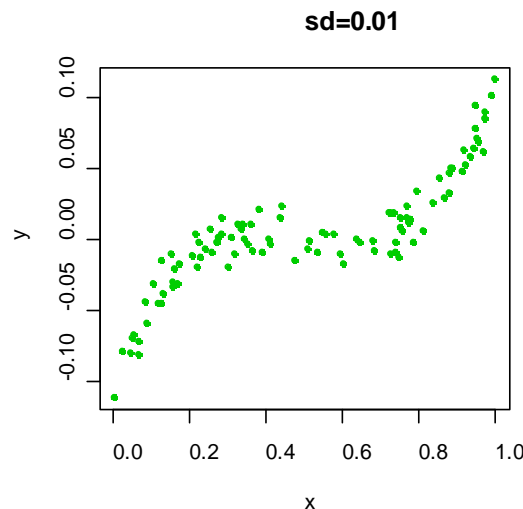
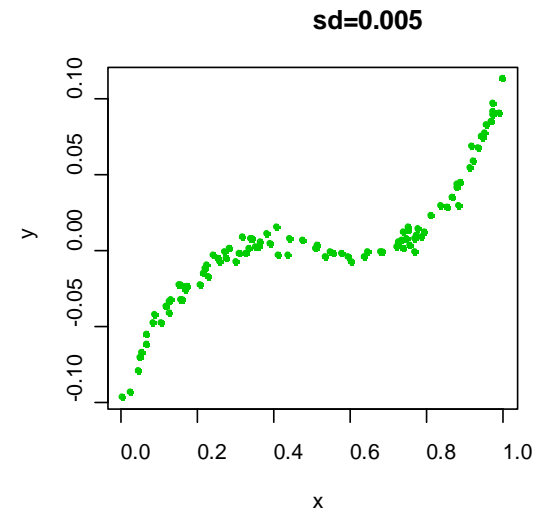
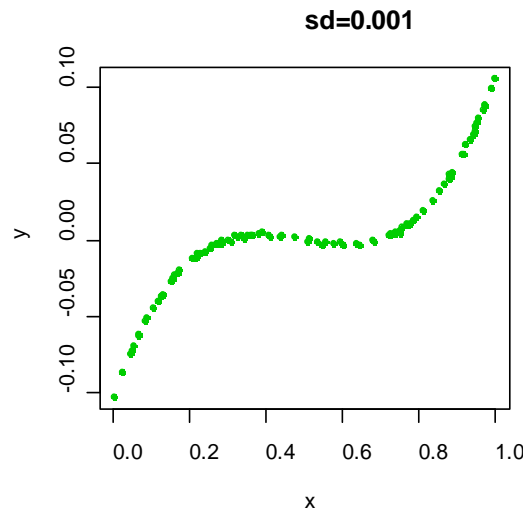


A Simple Example

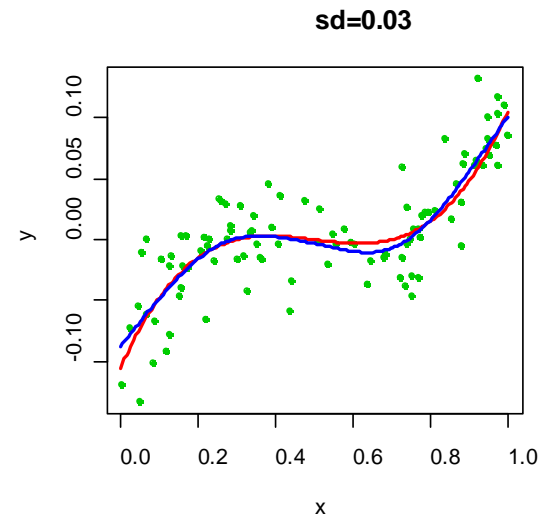
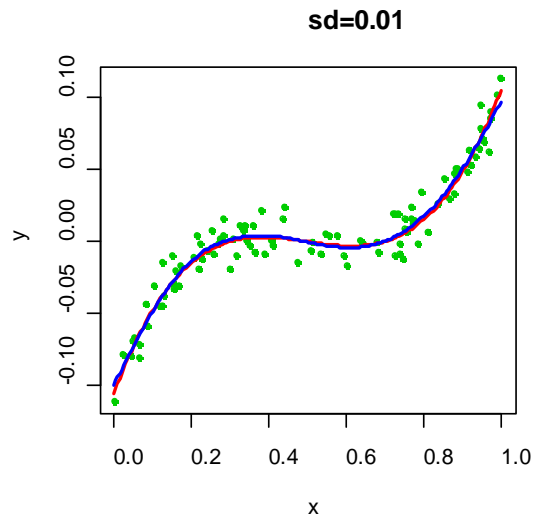
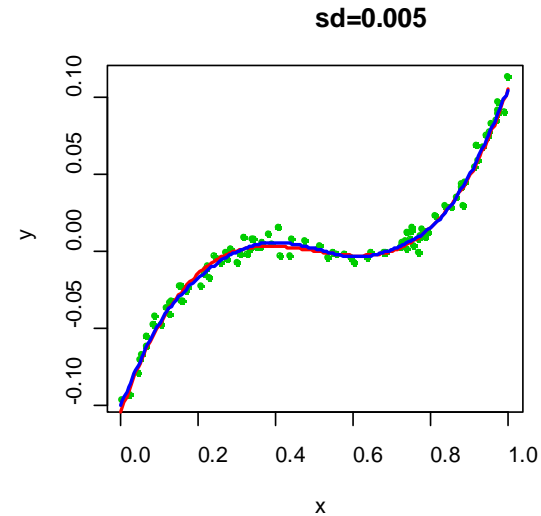
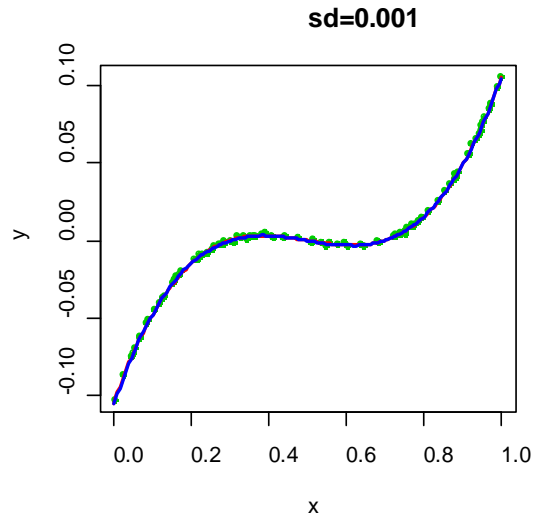


Different Standard Deviations

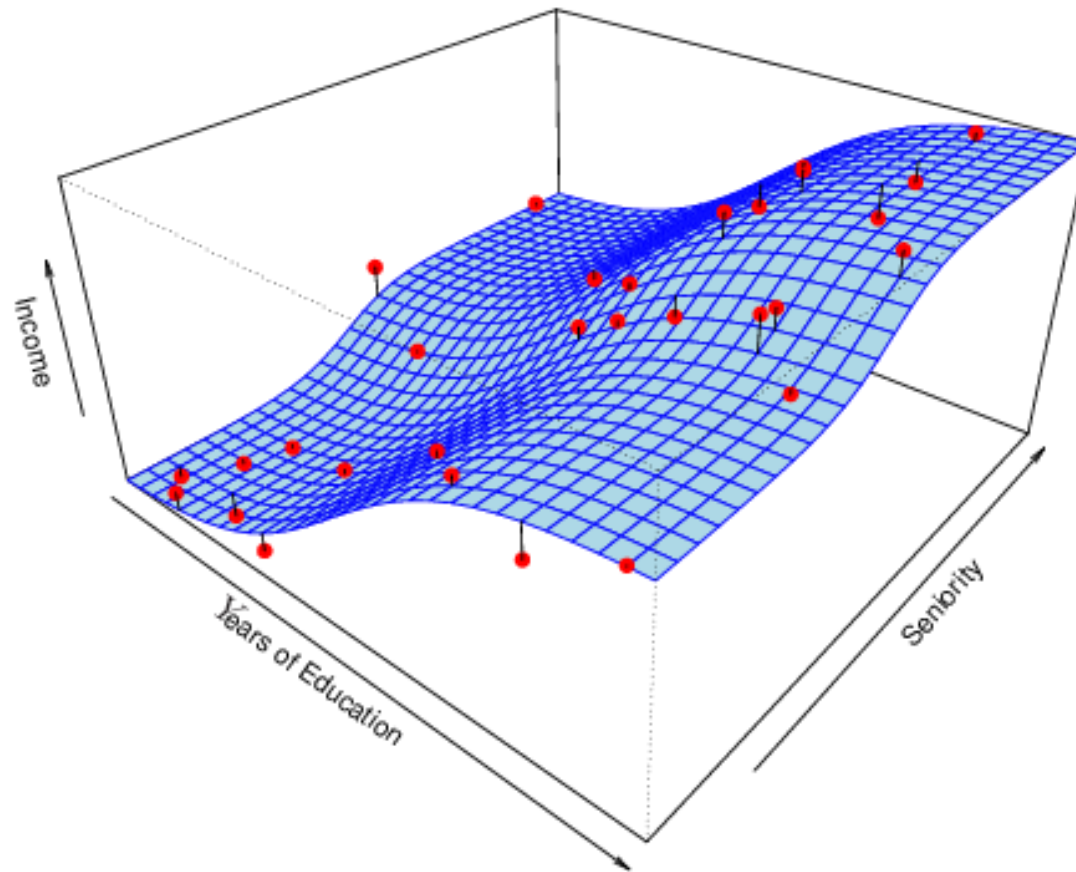
- The difficulty of estimating f will depend on the standard deviation of the ε 's.



Different Estimates For f



Income vs. Education Seniority



Why Do We Estimate f ?

- Statistical Learning is 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 \mathbf{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.

How Do We Estimate f ?

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

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

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

Parametric Methods

- It reduces the problem of estimating f down to one of estimating a set of parameters.
- They involve a two-step model based approach

STEP 1:

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

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

However, there are far more complicated, and flexible, models for f . In a sense the more flexible the model the more realistic it is.

Parametric Methods (cont.)

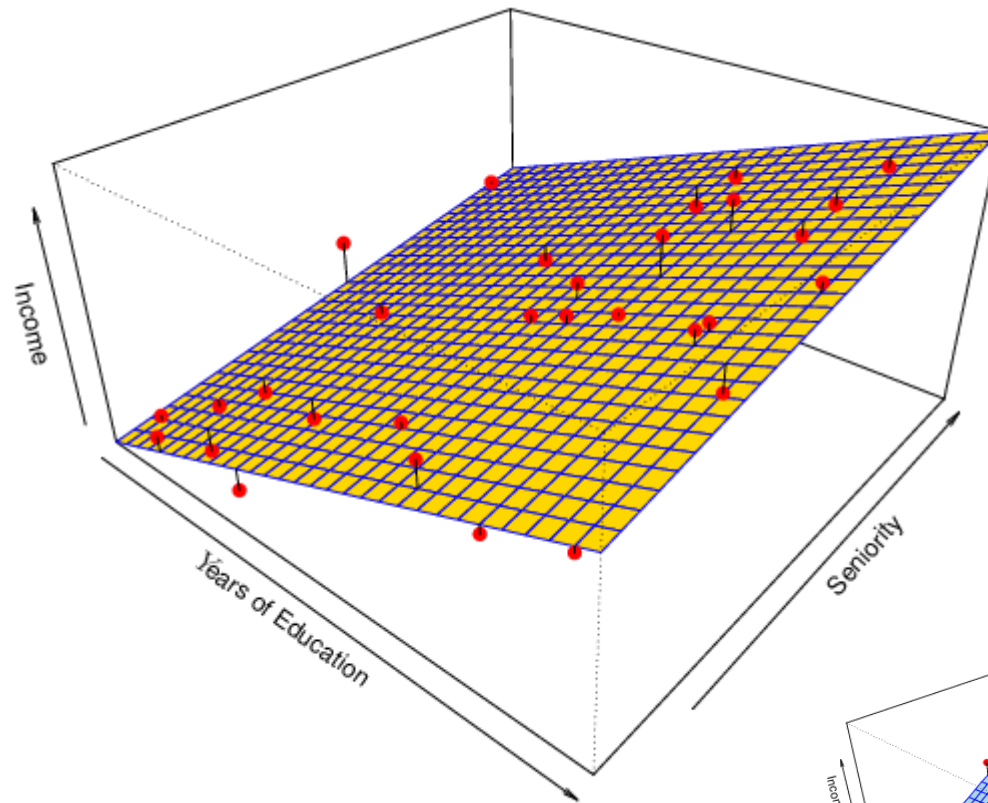
STEP 2:

Use the training data to fit the model i.e. estimate f or equivalently the unknown parameters such as $\beta_0, \beta_1, \beta_2, \dots, \beta_p$.

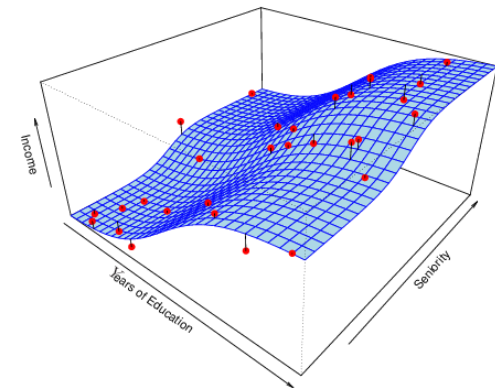
- The most common approach for estimating the parameters in a linear model is ordinary least squares (OLS).
- However, this is only one way.
- There are often superior approaches, even based on linear estimations.

Example: A Linear Regression Estimate

- Even if the standard deviation is low we will still get a bad answer if we use the wrong model.



$$\hat{f} = \beta_0 + \beta_1 \times \text{Education} + \beta_2 \times \text{Seniority}$$

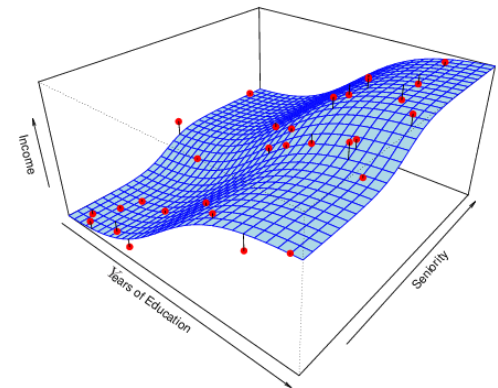
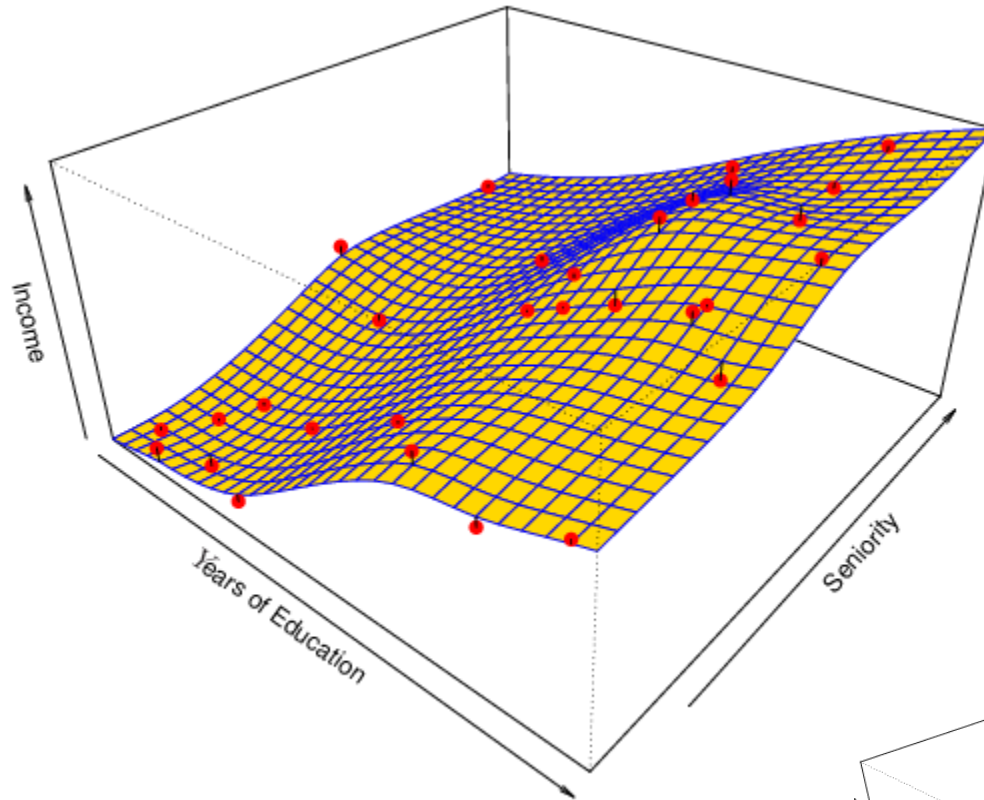


Non-parametric Methods

- They do not make explicit assumptions about the functional form of f .
- Advantages: They accurately fit a wider range of possible shapes of f .
- Disadvantages: A very large number of observations is required to obtain an accurate estimate of f

Example: A Thin-Plate Spline Estimate

- Non-linear regression methods are more flexible and can potentially provide more accurate estimates.



Tradeoff Between Prediction Accuracy and Model Interpretability

- Why not just use a more flexible method if it is more realistic?
- There are three main 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.

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 due to the following two main reasons:

Reason 2

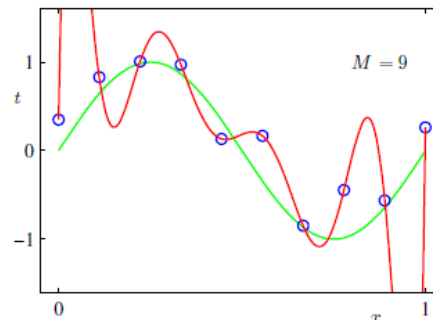
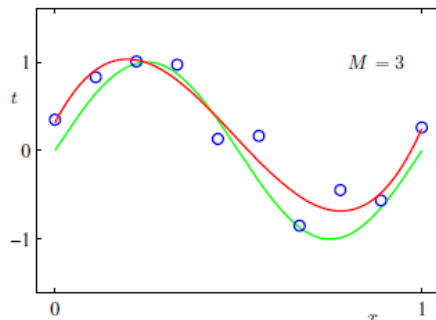
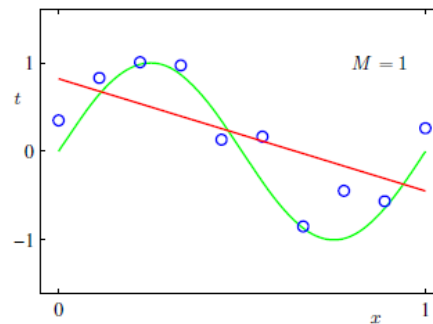
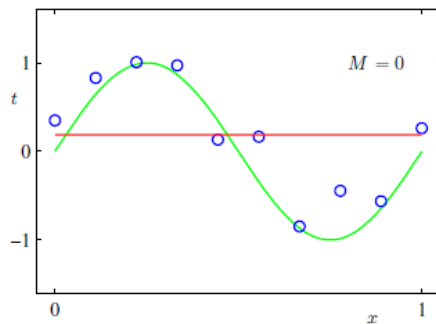
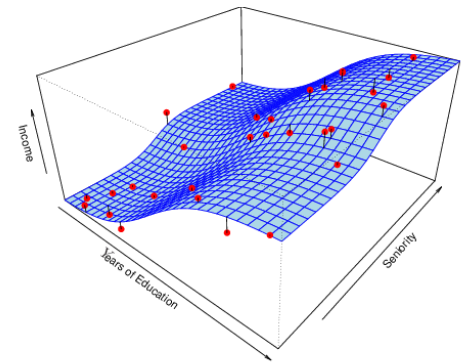
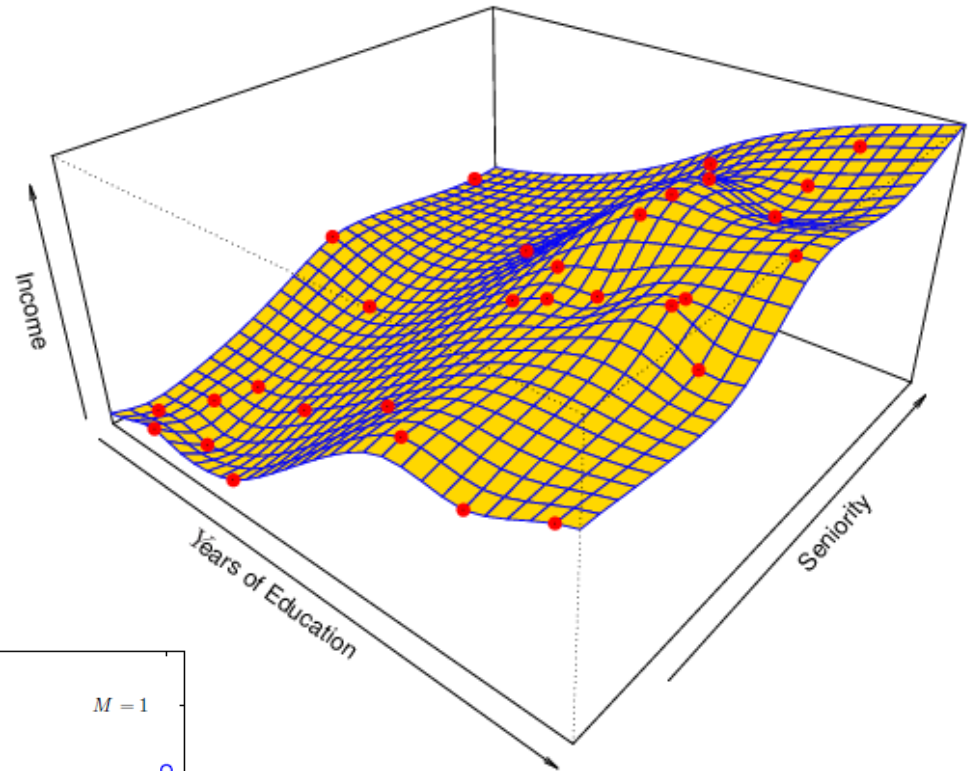
This seems counter intuitive but has to do with the fact that it is harder to fit a more flexible model.

Reason 3

Allowing too many freedom degrees can lead to overfitted models, which do not represent the real nature of the data.

Poor Estimates

- Non-linear regression methods can also be too flexible and produce poor estimates for f .



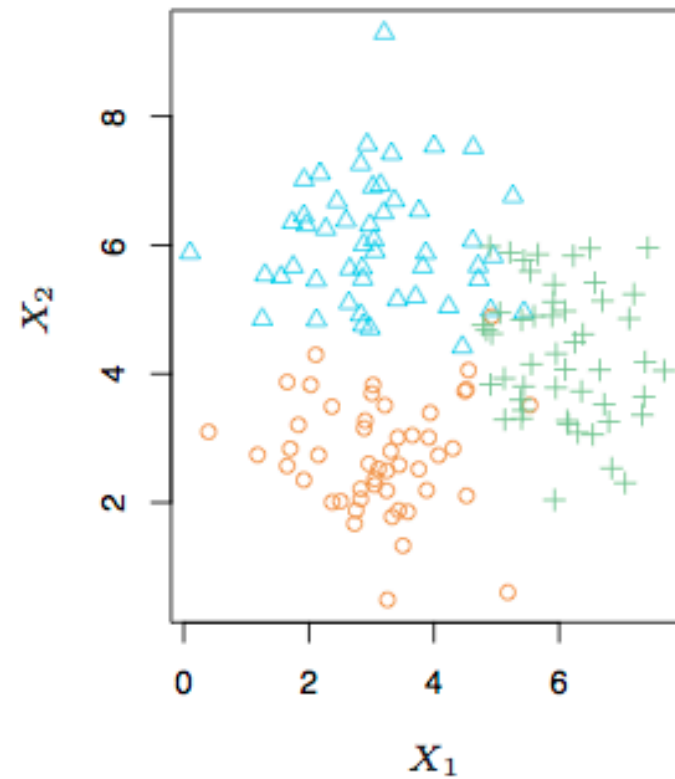
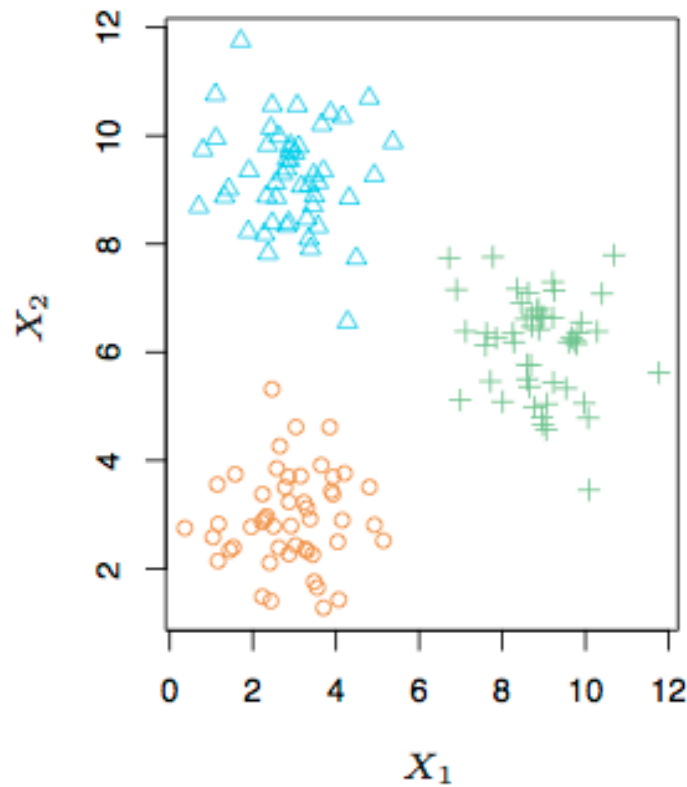
Supervised vs. Unsupervised Learning

- 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 classes.
 - **We will 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 (unsupervised classification).
- Another common example is *a priori* input space segmentation in regression problems where Y_i 's are initially excluded or considered as additional to the \mathbf{X}_i 's.
- 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, it is, the values in Y are categorized into different classes. e.g.
 - Will the Dow be up (U) or down (D) in 6 months?
 - Is this email a SPAM or not?

Different Approaches

- We will deal with both types of problems in this course.
- Some methods work well on both types of problem e.g. Neural Networks
- Other methods work best on Regression, e.g. Linear Regression, or on Classification, e.g. logistic regression, k-Nearest Neighbors.
- **We will deal with Regression in this part of the course**