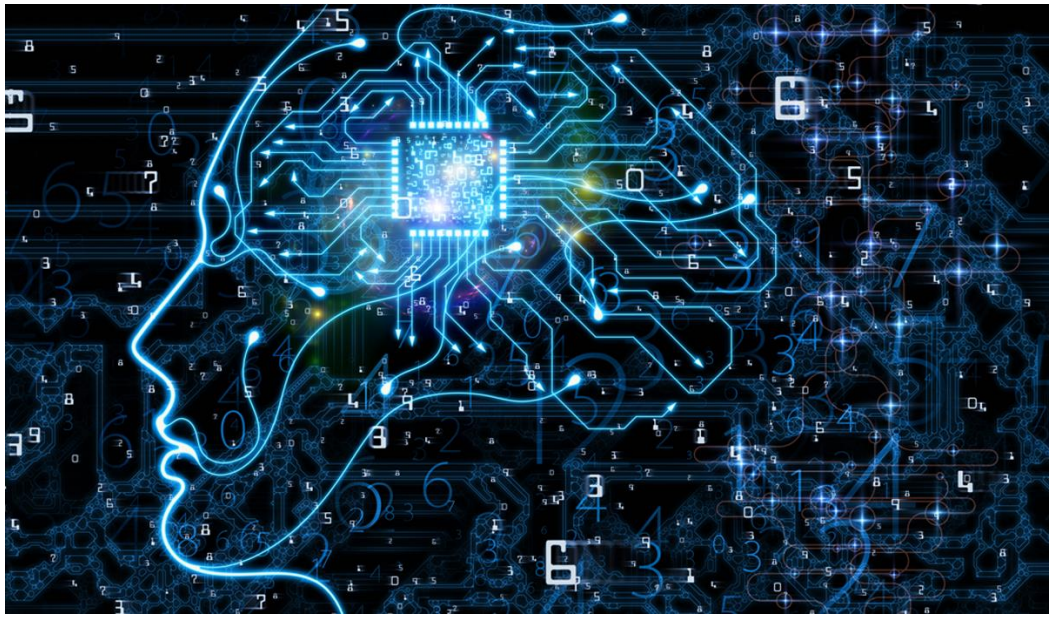
University of Ljubljana, Faculty of Computer and Information Science

Statistical Predictive Modeling



Prof Dr Marko Robnik-Šikonja

Intelligent Systems, October 2021

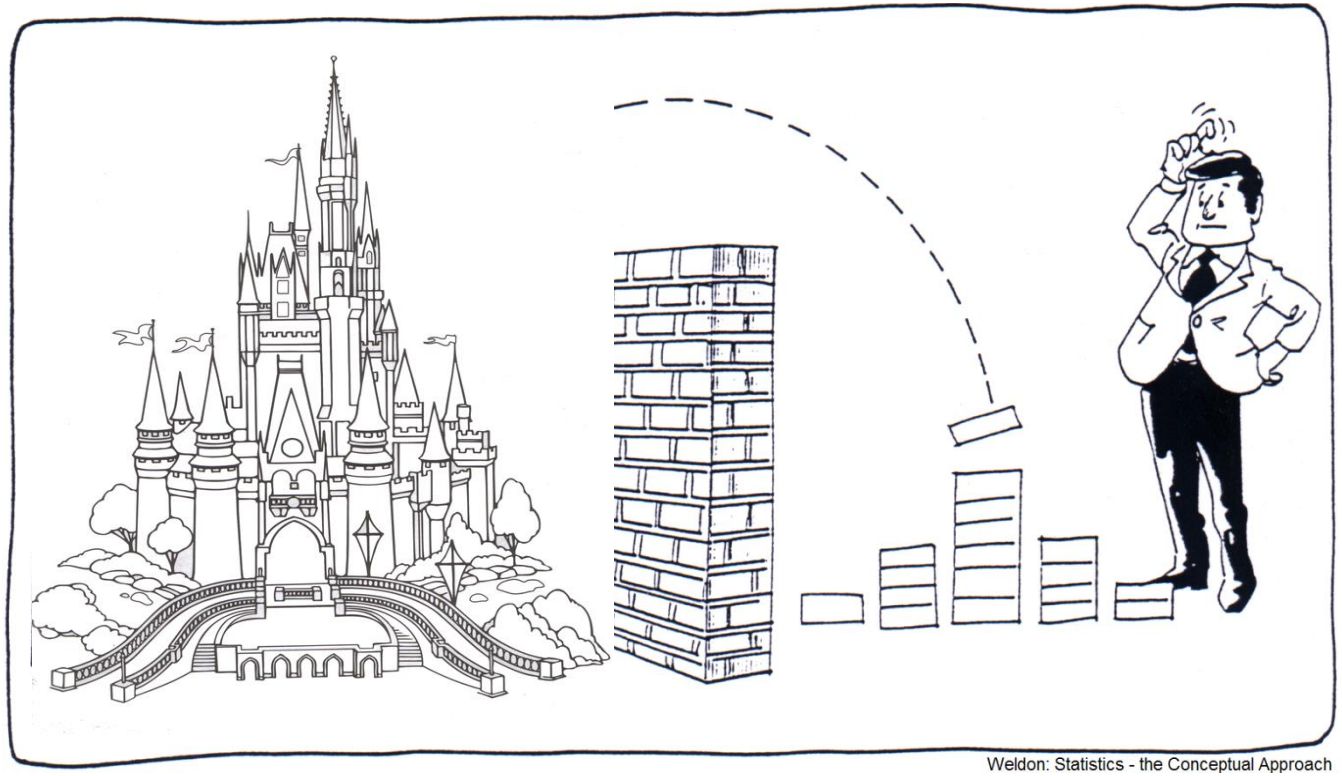
Learning

- **Learning** is the act of acquiring new, or modifying and reinforcing existing, knowledge, behaviors, skills, values, or preferences and may involve synthesizing different types of information.

- **Statistical learning** deals with the problem of finding a predictive function based on data.
- The goals of statistical learning: prediction and understanding.

Statistics and machine learning

- Definition from Wikipedia:
  ML algorithms operate by building a model from example inputs *i.e., samples.*



- ML can also be viewed as compression

The Data

Provided by the Institute of Oncology, Ljubljana

Post-surgery data for about 1000 breast cancer patients.

+

Recurrence and time of recurrence.

The Data

	class1	class2	menop	stage	grade	hType	PgR	inv	nLymph	cTh	hTh	famHist	LVI	ER	maxNode	posRatio	age
300	11.82	0	1	2	2	1	0	0	1	1	0	3	0	1	2	3	2
301	4.89	1	0	1	2	1	0	0	2	1	0	0	0	2	1	4	3
302	14.63	0	1	1	4	2	0	0	0	0	0	1	0	1	1	1	3
303	21.83	0	0	1	4	2	1	0	1	0	0	9	0	4	1	2	2
304	19.87	0	0	1	2	1	0	0	0	0	0	0	0	1	2	1	2
305	7.54	0	1	2	3	1	9	2	1	0	1	1	0	3	3	3	4
306	15.15	0	0	1	4	2	1	0	0	0	0	2	0	4	1	1	2
307	0.30	1	0	2	2	1	0	0	3	0	0	9	0	1	1	4	2
308	12.49	0	1	2	2	3	1	0	0	0	0	0	0	4	1	1	5
309	1.77	1	0	2	3	1	1	2	2	1	0	9	1	3	3	3	2

Each patient is described with 17 values:

- 15 patient's features
- 2 values, which describe the outcome

1 instance = 1 patient

	class1	class2	menop	stage	grade	hType	PgR	inv	nLymph	cTh	hTh	famHist	LVI	ER	maxNode	posRatio	age
300	11.82	0	1	2	2	1	0	0	1	1	0	3	0	1	2	3	2
301	4.89	1	0	1	2	1	0	0	2	1	0	0	0	2	1	4	3
302	14.63	0	1	1	4	2	0	0	0	0	0	1	0	1	1	1	3
303	21.83	0	0	1	4	2	1	0	1	0	0	9	0	4	1	2	2
304	19.87	0	0	1	2	1	0	0	0	0	0	0	0	1	2	1	2
305	7.54	0	1	2	3	1	9	2	1	0	1	1	0	3	3	3	4
306	15.15	0	0	1	4	2	1	0	0	0	0	2	0	4	1	1	2
307	0.30	1	0	2	2	1	0	0	3	0	0	9	0	1	1	4	2
308	12.49	0	1	2	2	3	1	0	0	0	0	0	0	4	1	1	5
309	1.77	1	0	2	3	1	1	2	2	1	0	9	1	3	3	3	2

- Menopause?
- Tumor stage
- Tumor grade
- Histological type
- Progesterone receptor lvl.
- Invasive tumor type
- Number of positive lymph nodes

- Hormonal therapy?
- Chemotherapy?
- Family medical history
- Lymphovascular invasion?
- Estrogen receptor lvl.
- Size of max. removed node
- Ratio of positive lymph nodes
- Age group

Prognostic Features

	class1	class2	menop	stage	grade	hType	PgR	inv	nLymph	cTh	hTh	famHist	LVI	ER	maxNode	posRatio	age
300	11.82	0	1	2	2	1	0	0	1	1	0	3	0	1	2	3	2
301	4.89	1	0	1	2	1	0	0	2	1	0	0	0	2	1	4	3
302	14.63	0	1	1	4	2	0	0	0	0	0	1	0	1	1	1	3
303	21.83	0	0	1	4	2	1	0	1	0	0	9	0	4	1	2	2
304	19.87	0	0	1	2	1	0	0	0	0	0	0	0	1	2	1	2
305	7.54	0	1	2	3	1	9	2	1	0	1	1	0	3	3	3	4
306	15.15	0	0	1	4	2	1	0	0	0	0	2	0	4	1	1	2
307	0.30	1	0	2	2	1	0	0	3	0	0	9	0	1	1	4	2
308	12.49	0	1	2	2	3	1	0	0	0	0	0	0	4	1	1	5
309	1.77	1	0	2	3	1	1	2	2	1	0	9	1	3	3	3	2

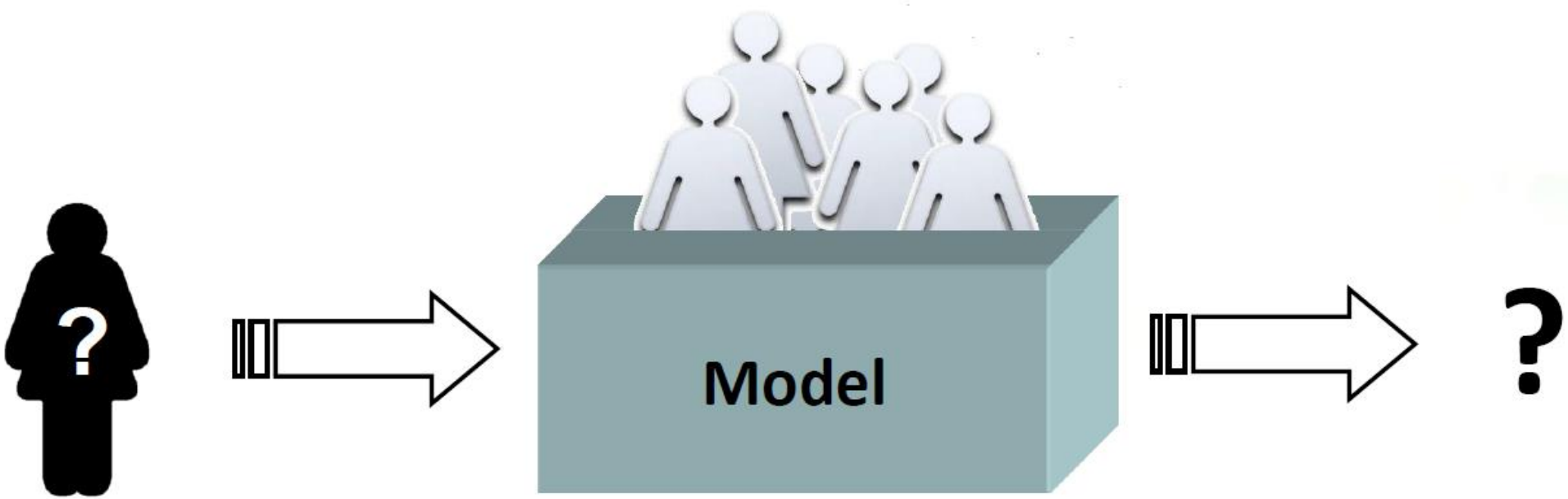
- Menopause?
- Tumor stage
- Tumor grade
- Histological type
- Progesterone receptor lvl.
- Invasive tumor type
- Number of positive lymph nodes

- Hormonal therapy?
- Chemotherapy?
- Family medical history
- Lymphovascular invasion?
- Estrogen receptor lvl.
- Size of max. removed node
- Ratio of positive lymph nodes
- Age group

Oncologists use these attributes for prognosis in every-day medical practice.

Basic Task in ML

We want to learn from past examples,with known outcomes.



To predict the outcome for a new patient.

Basic notation of predictive modelling

- Recurrence is a statistical variable named response or target or prediction variable that we wish to predict. We usually refer to the response as Y .
- Other variables are called attributes, features, inputs, or predictors; we name them X_i.
- The input vectors forms a matrix **X**

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$$

- The model we write as

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

where ∈ is independent from X, has zero mean and represents measurement errors and other discrepancies.

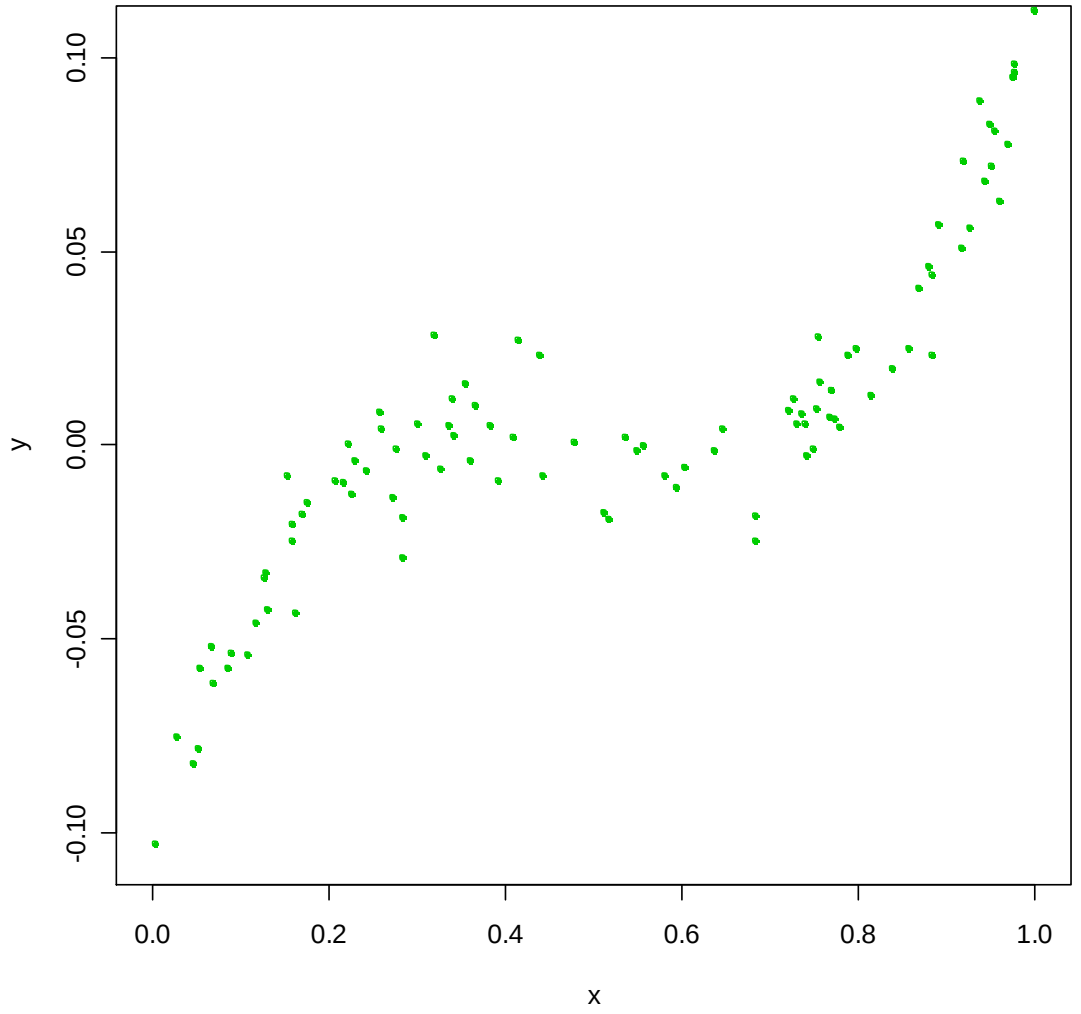
Further notation for instances

- 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

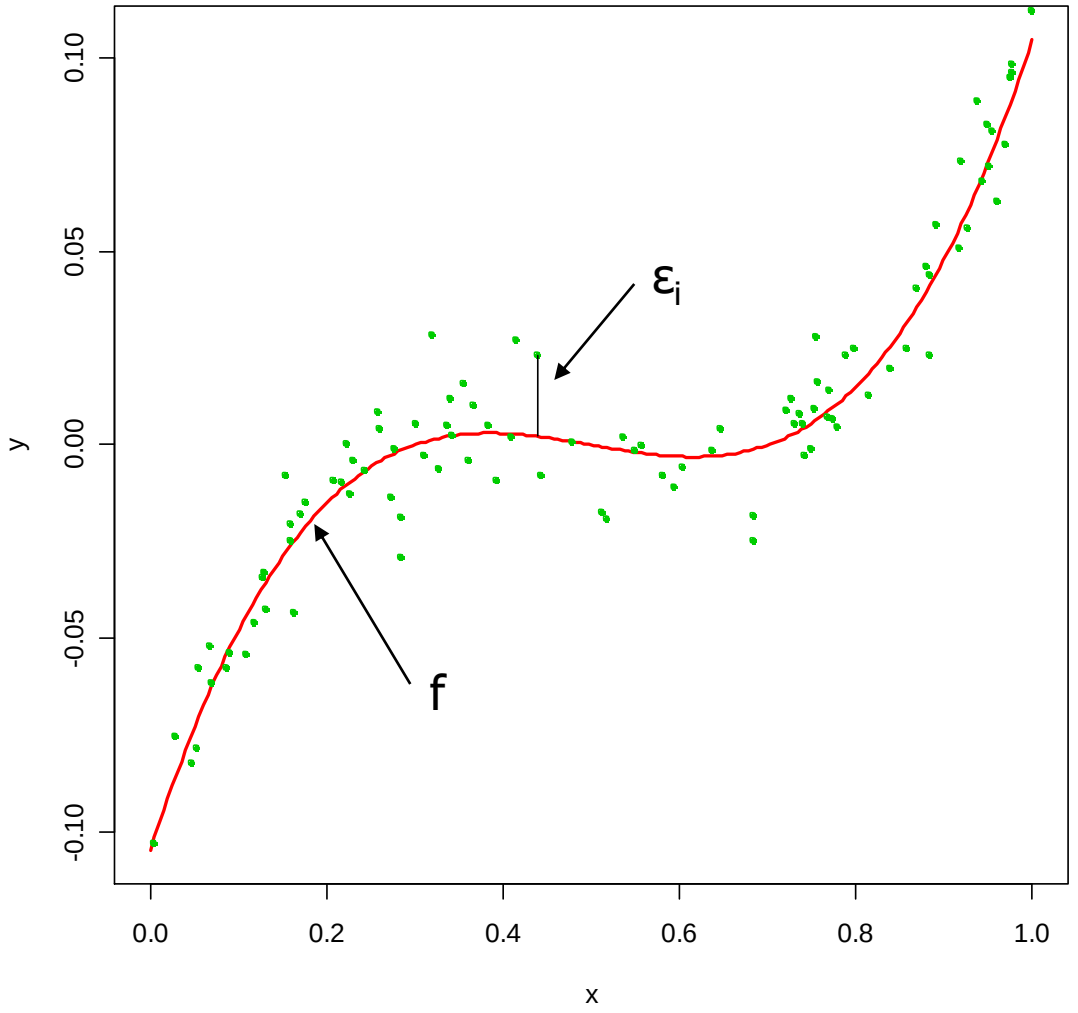
$$Y_i = f(\mathbf{X}_{i.}) + \varepsilon_i$$

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

A simple example

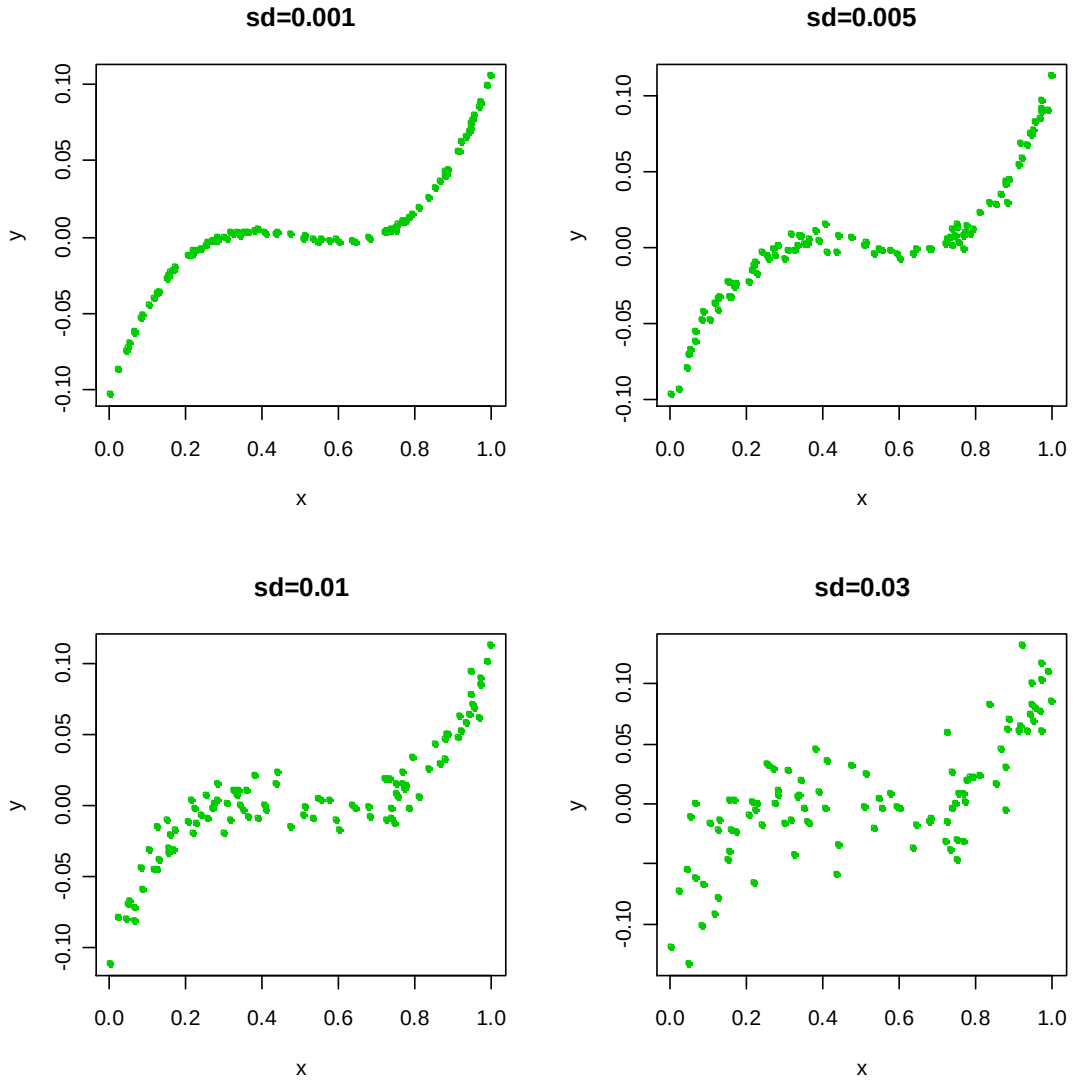


A simple example

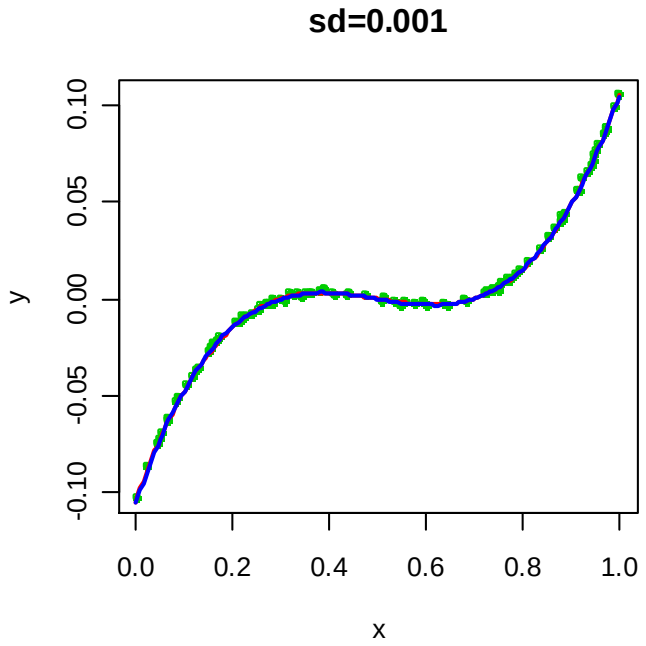


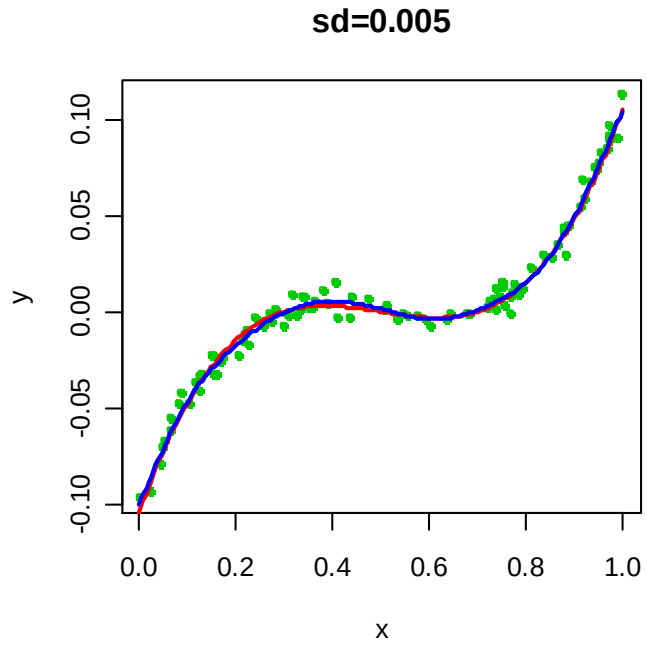
Different standard deviations

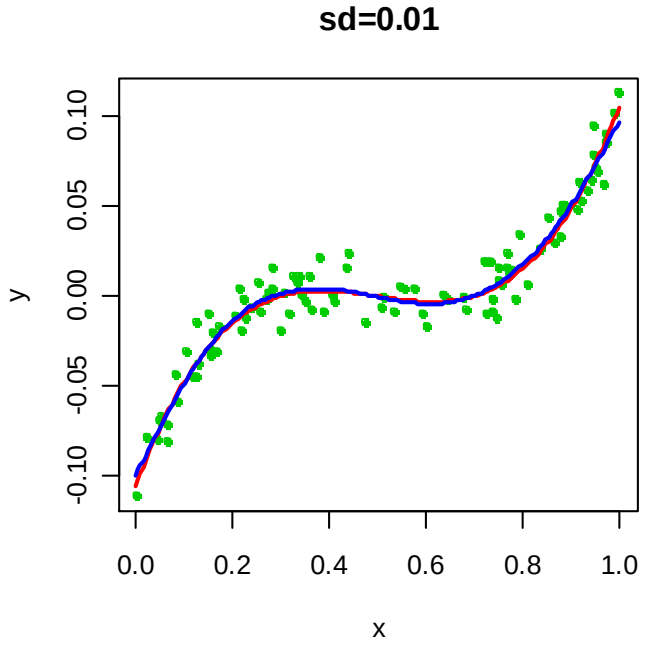
- The difficulty of estimating f will depend on the standard deviation of the ε's.

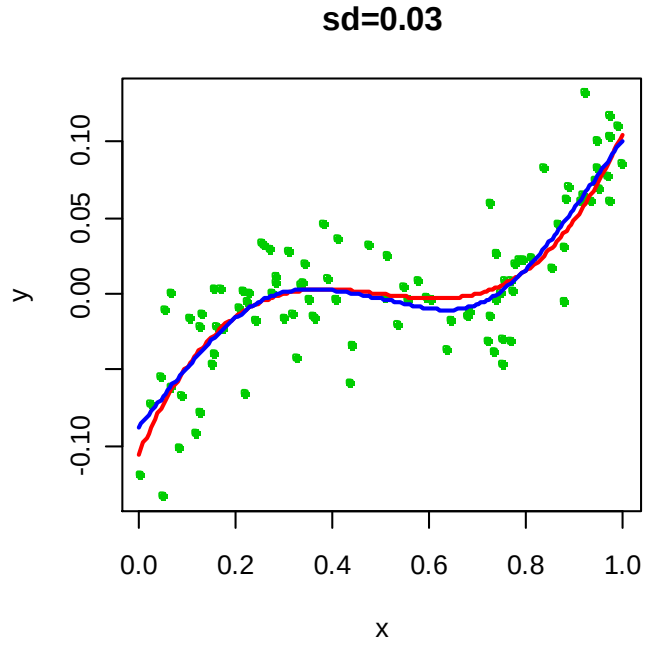


Different estimates for f



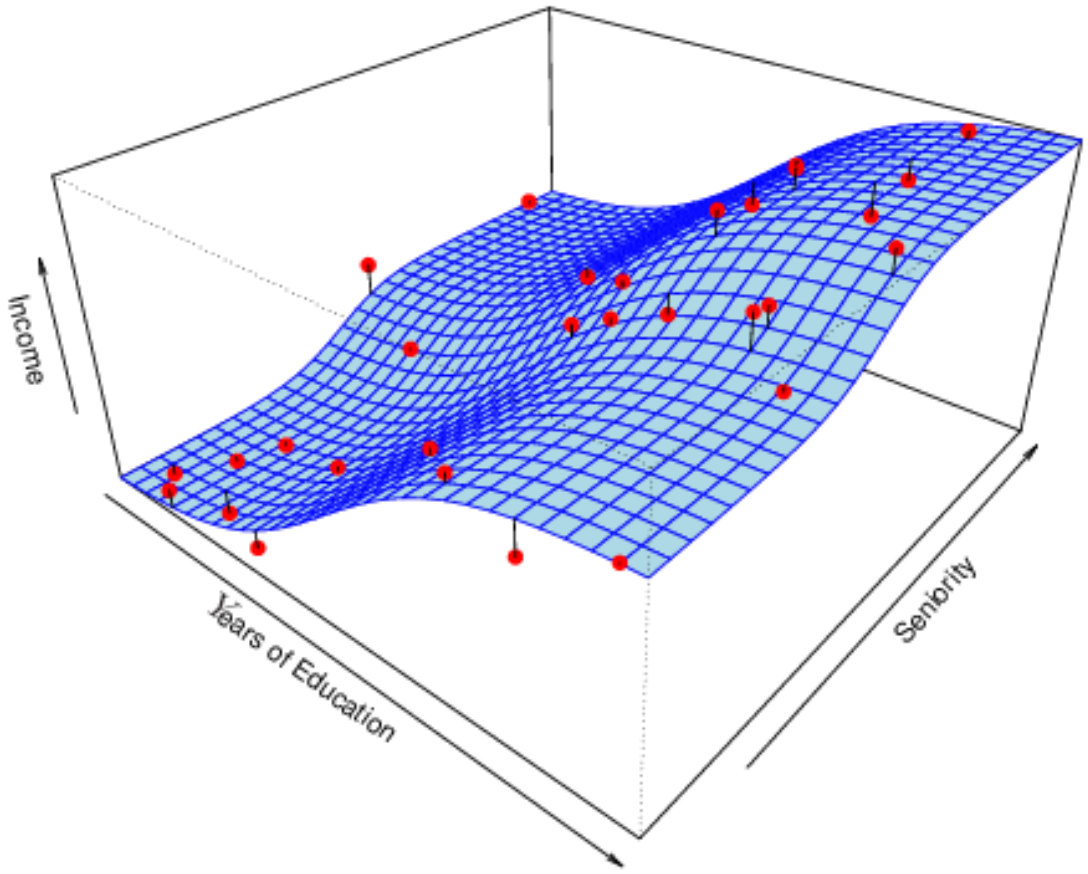






Income vs. Education and Seniority

Multidimensional **X**



1^{st} goal of learning: 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?

2nd goal of learning: inference

- often we are 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.?
- Sometimes more important than prediction, e.g., in medicine.
- 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**

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

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

Parametric methods

- They reduce 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.

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

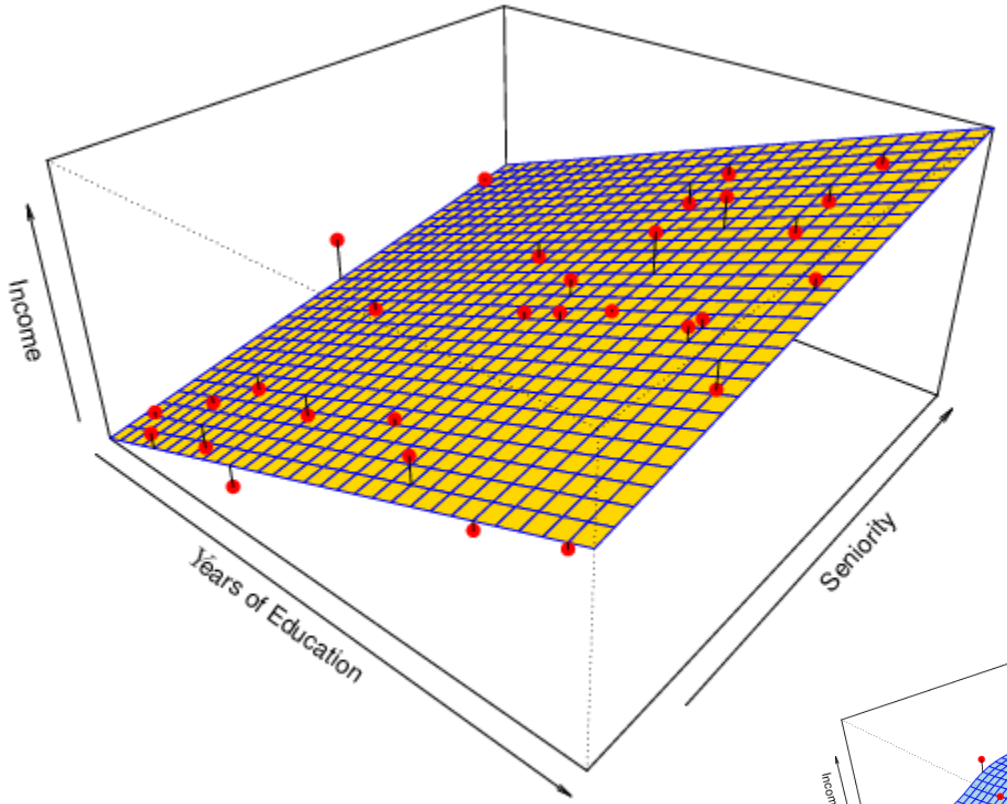
More complicated and flexible models for *f are often* more realistic.

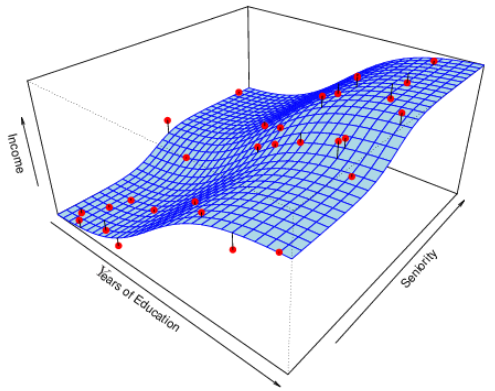
STEP 2:

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
For linear model the most common method uses ordinary least squares (OLS).

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.





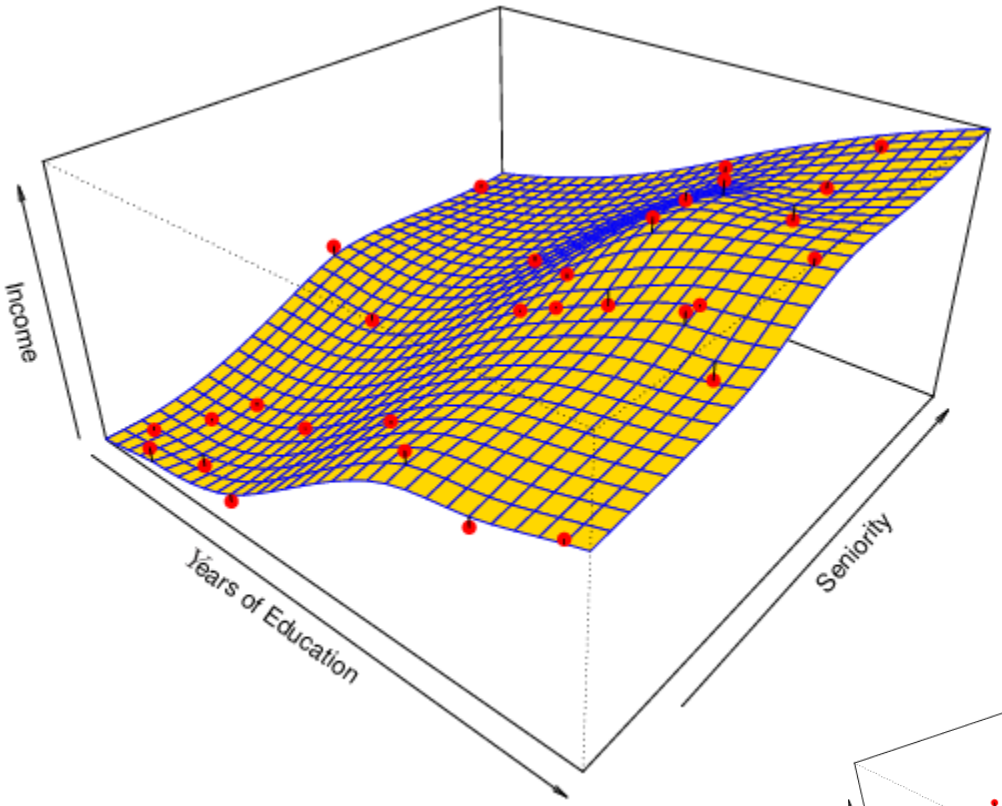
$$f = b_0 + b_1 \times Education + b_2 \times Seniority$$

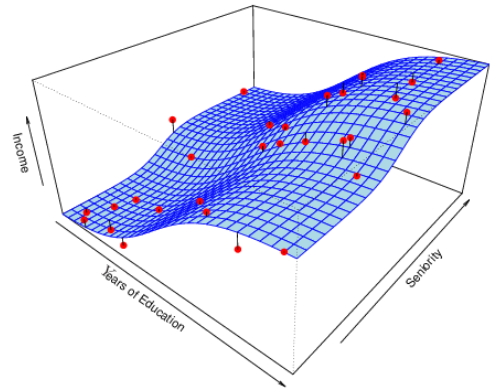
Non-parametric methods

- 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 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.





Trade-off between prediction accuracy and model interpretability

- Why not just use a more flexible method if it is more realistic?

Reason 1:

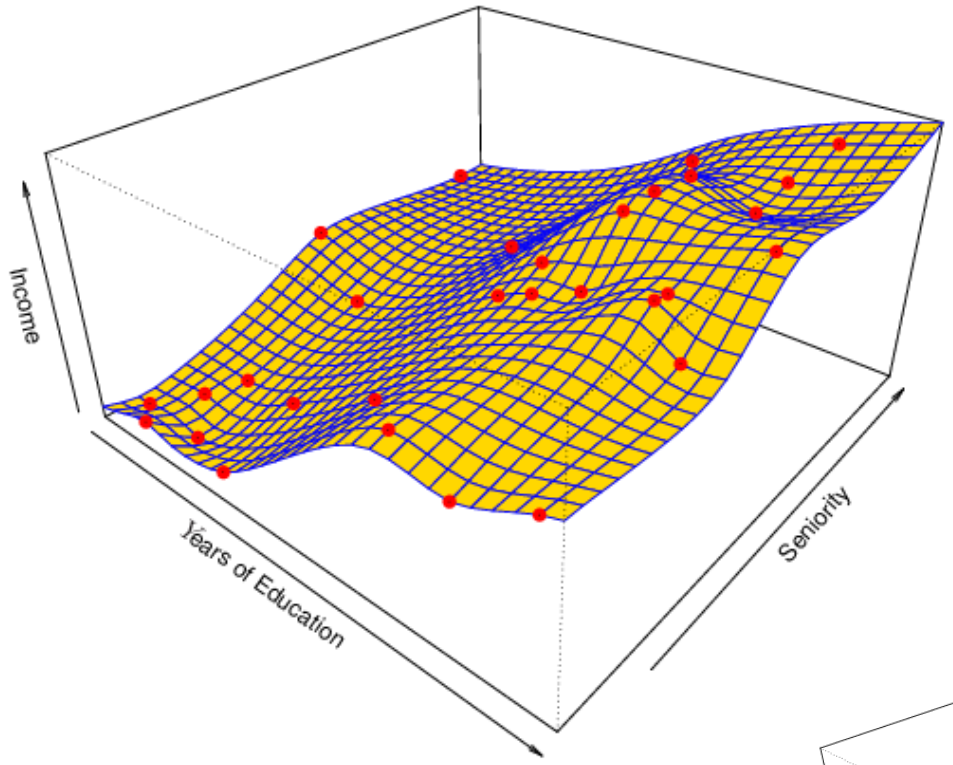
A simple method such as linear regression produces a model which is much easier to interpret (the inference part is better). For example, in a linear model, β_j is the average increase in Y for a one unit increase in X_j holding all other variables constant.

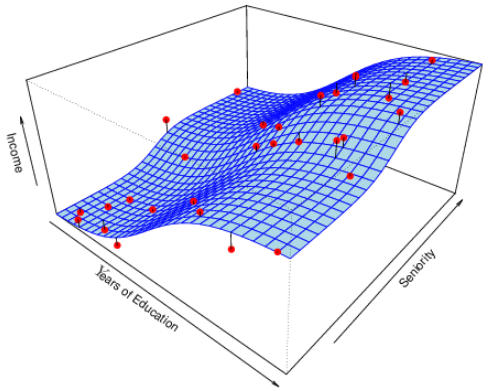
Reason 2:

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.

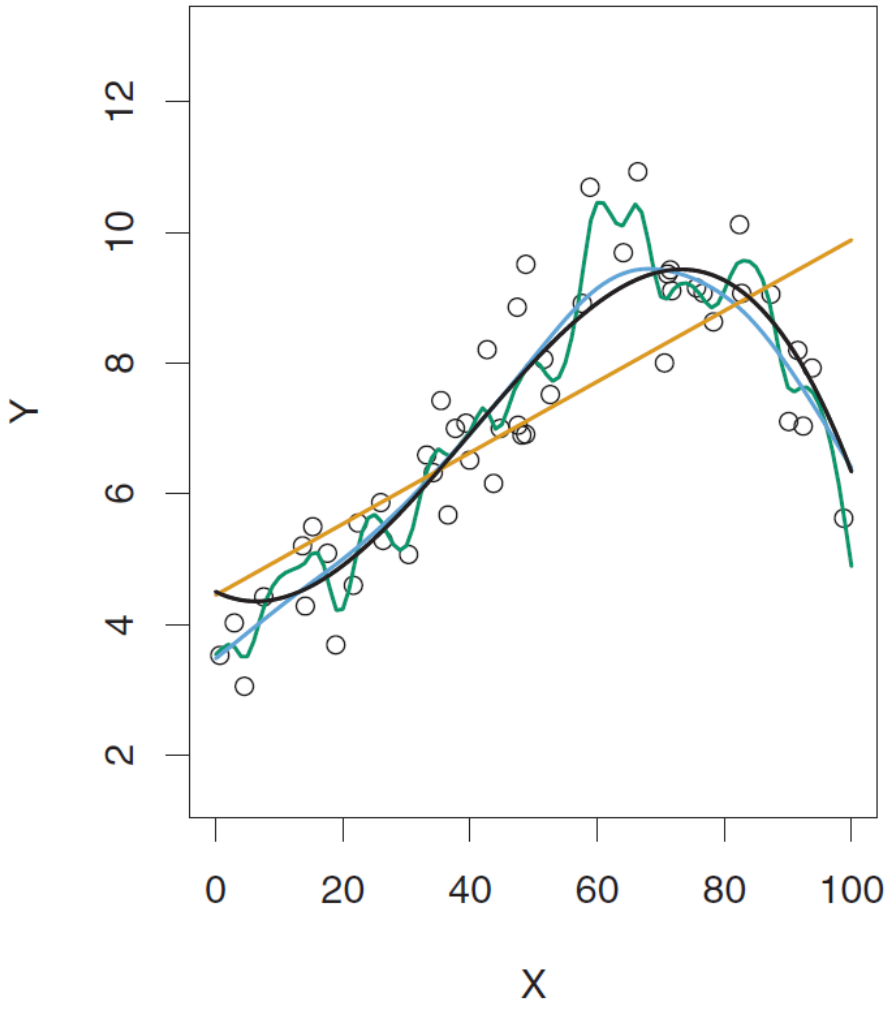
A poor estimate: overfitting

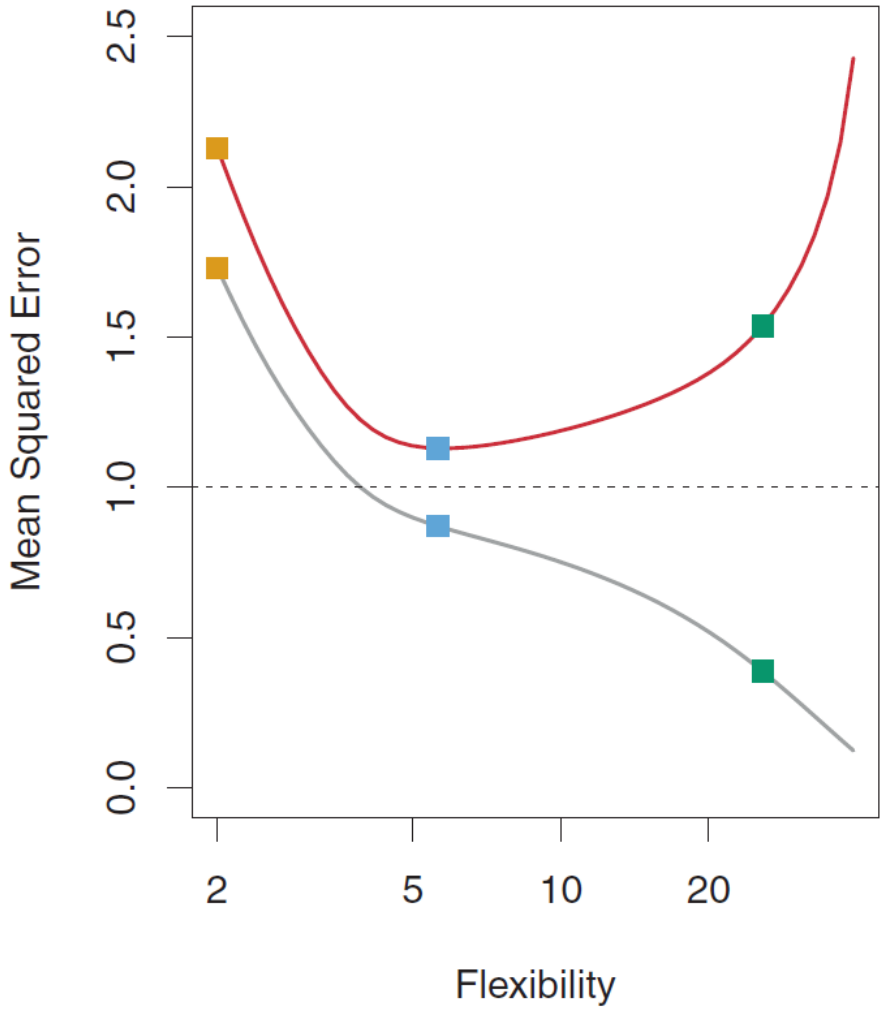
- Non-linear regression methods can also be too flexible and produce poor estimates for *f*.





Goodness of fit for three models





LEFT

Black: Truth

Orange: Linear Estimate

Blue: smoothing spline

Green: smoothing spline (more flexible)

RIGHT

RED: Test MSE

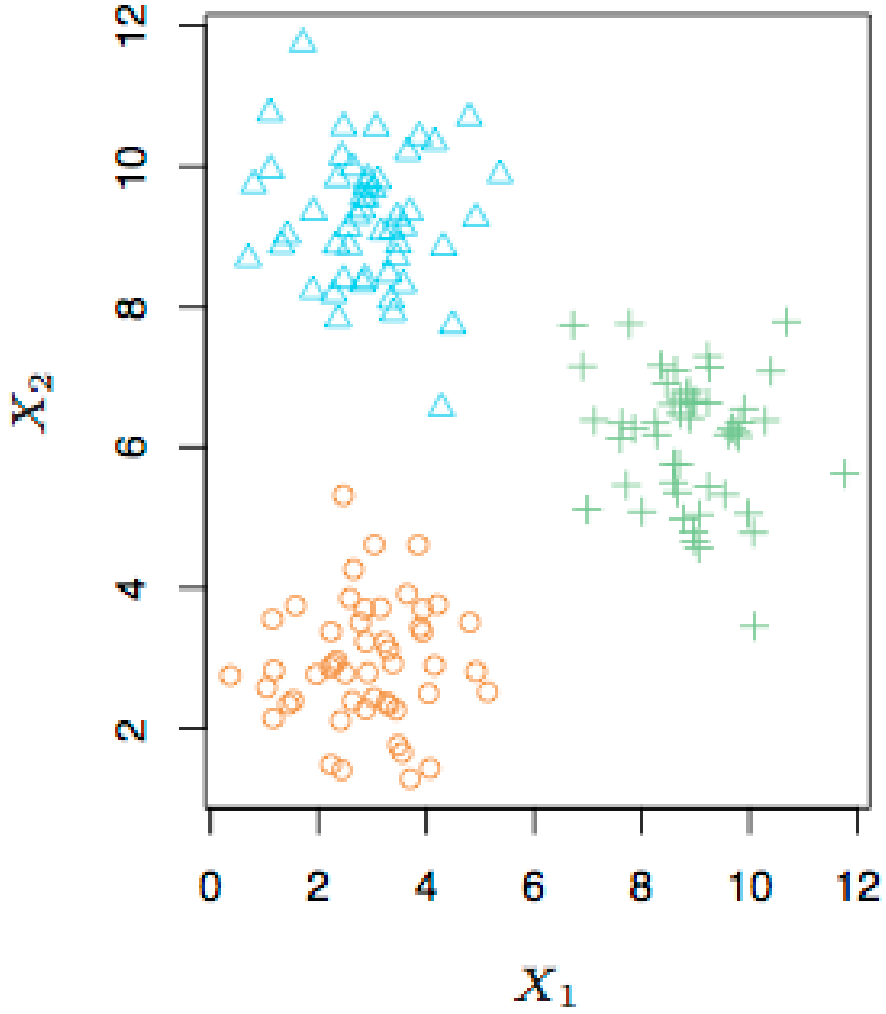
Grey: Training MSE

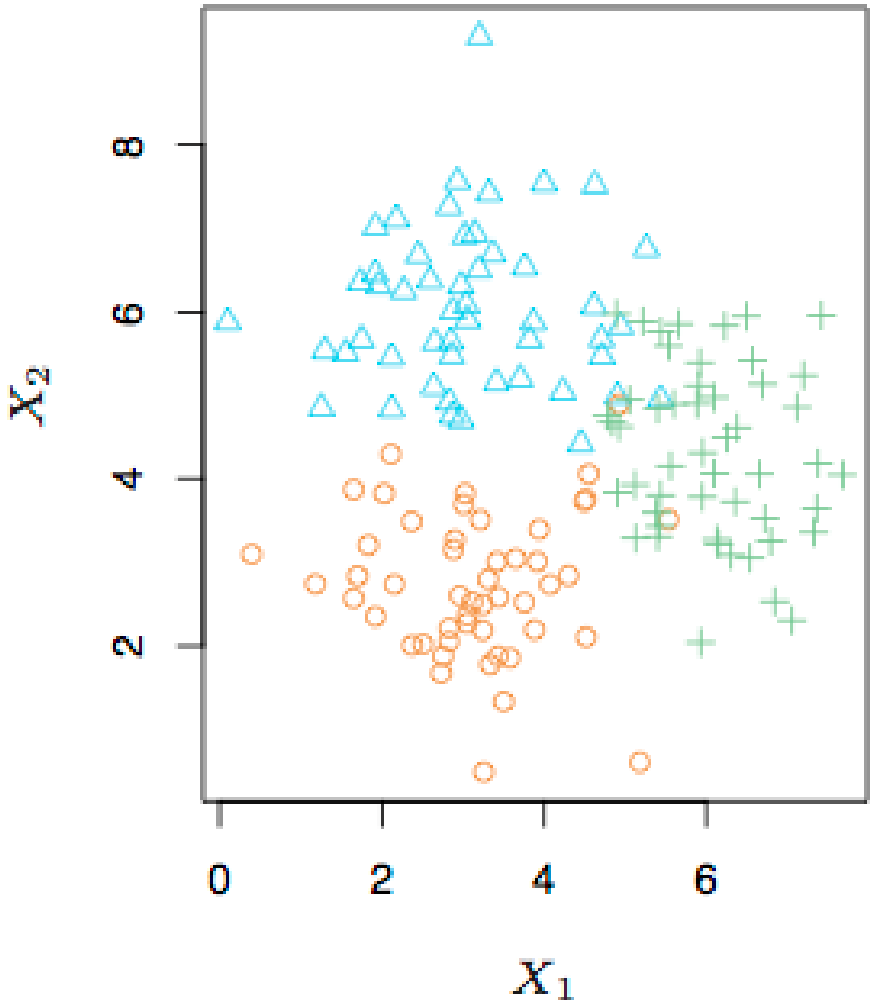
Dashed: Minimum possible test MSE (irreducible error)

Supervised, unsupervised, semi-supervised, self-supervised, weakly-supervised learning 1/2

- We can divide learning problems into Supervised and Unsupervised situations
- <u>Supervised learning:</u>
  - Supervised Learning is where both the predictors, $\mathbf{X}_i$, and the response, Y_i, are observed.
  - e.g., linear regression
- <u>Unsupervised learning:</u>
  - 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.
    - Idea: Maximizing initra-cluster similarity & minimizing inter-cluster similarity
- <u>Semi-supervised learning</u>
  - only a small sample of labelled instances are observed but a large set of unlabeled instances
  - an initial supervised model is used to label unlabeled instances
  - the most reliable predictions are added to the training set for the next iteration of supervised learning

A simple clustering example





Supervised, unsupervised, semi-supervised, self-supervised, weakly-supervised learning 2/2

- Self-supervised learning
  - a mixture of supervised and unsupervised learning
  - learns from unlabelled data
  - the labels are obtained from related properties of the data itself, often leveraging the underlying structure in the data
  - usually predicts any unobserved or hidden part (or property) of the input from any observed or unhidden part of the input.
  - e.g., in NLP, we can hide part of a sentence and predict the hidden words from the remaining words
  - e.g., in video processing, we can predict past or future frames in a video (hidden data) from current ones (observed data)
- Weakly-supervised data
  - noisy, limited, or imprecise sources are used to provide supervision signal for labeling large amounts of training data to do supervised learning
  - reduces the burden of obtaining hand-labeled data sets, which can be costly or impractical
  - e.g., using smart electricity meter to estimate household occupancy

Regression vs. classification

- Supervised learning problems can be further divided into
- Regression problems: Y is continuous/numerical. e.g.
    - Predicting the value of certain share on stock market
    - Predicting the value of a given house based on various inputs
    - The duration in years till cancer recurrence
- Classification problems: Y is categorical e.g.,
    - Will the price of a share go up (U) or down (D)?
    - Is this email a SPAM or not?
    - Will the cancer recur?
    - What will be an outcome of a football match (Home, Away, or Draw)?
    - Credit card fraud detection, direct marketing, classifying stars, diseases, web-pages, etc.
- Some methods work well on both types of problem, e.g., neural networks or kNN

Data mining: on what kinds of data?

- Database-oriented data sets and applications
  - Relational database, data warehouse, transactional database
- Advanced data sets and advanced applications
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data (incl. bio-sequences)
  - Structure data, graphs, social networks and multi-linked data
  - Object-relational databases
  - Heterogeneous databases and legacy databases
  - Spatial data and spatiotemporal data
  - Multimedia database
  - Text databases
  - The World-Wide Web

Association and correlation analysis

- Frequent patterns (or frequent itemsets)
  - What items are frequently purchased together in the supermarket?
- Association, correlation vs. causality
  - A typical association rule
    - Diaper → Beer [0.5%, 75%] (support, confidence)
  - Are strongly associated items also strongly correlated?
- How to mine such patterns and rules efficiently in large datasets?
- How to use such patterns for classification, clustering, and other applications?

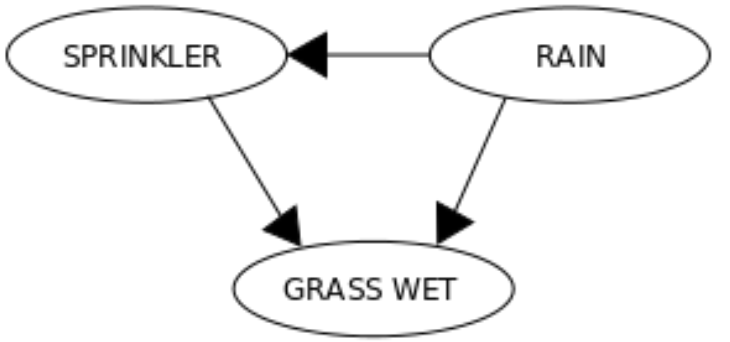
Outlier analysis

- Outlier: A data object that does not comply with the general behavior of the data
- Noise or exception? — One person's garbage could be another person's treasure
- Methods: byproduct of clustering or regression analysis, …
- Useful in fraud detection, rare events analysis

Relational learning

- Several variants:
  - Bayesian networks,
  - inductive logic programming
  - graph learning, e.g., link prediction

	SPRINKLER	
RAIN	T	F
F	0.4	0.6
T	0.01	0.99



	RAIN	
T	F	
0.2	0.8	

		GRASS WET	
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

Another view on learning: generalization as search

- Inductive learning: find a concept description that fits the data
- Example: rule sets as description language
  - Enormous but finite search space
- Simple solution:
  - enumerate the concept space
  - eliminate descriptions that do not fit examples
  - surviving descriptions contain target concept

Learning as optimization

- Usually the goal of classification is to minimize the test error
- Therefore, many learning algorithms solve optimization problems, e.g.,
  - linear regression minimizes squared error on the training set
  - AntMiner algorithms minimize the classification accuracy of decision rules on the training set using ACO
  - to find a good architecture of neural networks, GAs can be applied and minimize the prediction error on the validation set

Criteria of success for ML

- No single best method (no free lunch theorem)
- How to select the best model?
  - measure the quality of fit, i.e. how well the predictions match the observed data
  - measure on previously unseen data (called test set). Why?
- In regression the most popular measure is mean squared error

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

- in classification the *classification accuracy = 1 – error rate* is the most popular criterion

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

- We will say more about the topic later

No-Free-Lunch theorem

- In the "no free lunch" metaphor,
  each "restaurant" (problem-solving procedure)
  has a "menu" associating
  each "lunch plate" (problem)
  with a "price" (the performance of the procedure in solving the problem).
- The menus of restaurants are identical except in one regard – the prices are shuffled from one restaurant to the next.
- For an omnivore who is as likely to order each plate as any other, the average cost of lunch does not depend on the choice of restaurant.
- But a vegan who goes to lunch regularly with a carnivore who seeks economy might pay a high average cost for lunch.
- To methodically reduce the average cost, one must use advance knowledge of
  - a) what one will order and
  - b) what the order will cost at various restaurants.
- That is, improvement of performance in problem-solving hinges on using prior information to match procedures to problems.

No-free-lunch theorem

For any two learning algorithms $P_1(h|D)$ and $P_2(h|D)$, the following are true, independent of the sampling distribution P(x) and the number n of training points:

1. Uniformly averaged over all target functions F,

   $\mathcal{E}_1(E|F,n) - \mathcal{E}_2(E|F,n) = 0.$

2. For any fixed training set D, uniformly averaged over F,

   $\mathcal{E}_1(E|F,D) - \mathcal{E}_2(E|F,D) = 0$

3. Uniformly averaged over all priors P(F),

   $\mathcal{E}_1(E|n) - \mathcal{E}_2(E|n) = 0$

4. For any fixed training set D, uniformly averaged over P(F),

   $\mathcal{E}_1(E|D) - \mathcal{E}_2(E|D) = 0$

Consequences of the NFL theorem

If no information about the target function F(x) is provided:

- No classifier is better than some other in the general case.
- No classifier is better than random in the general case.

- ML practicioners poses implicit or explicit knowledge about the prices in different restaurants
- Meta-learning
- Automatic ML