Lecture 1 and 2 : Introduction to Statistical Learning

(Compulsory reading after lectures: Chap.1+ Chap.2)

Certain content of the slides refers to
listed on: http://www-bcf.usc.edu/~gareth/ISL/

Outlines

- Machine learning and Statistic Learning
- Supervised and Unsupervised Learning
- Introduction to Supervised Learning
- Prediction and Inference
- Approaches for Model Estimation
- Assessing Model Accuracy

Machine Learning VS Statistical Learning

Machine Learning (I)

- Arthur Samuel (1959): “The field of study that gives computers the ability to learn without being explicitly programmed”.

- Tom Mitchell (1998): “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E”

Machine Learning (II)

- It applies algorithms that iteratively learn from data, and allows computers to find hidden insights without being explicitly programmed where to look.

- A subfield of **computer science** and **artificial intelligence**. It emphasizes computational algorithms, letting the data to speak out self without much initial hypothesis/assumptions.

Machine Learning (III)

- Emphasizes **accuracy** and **prediction**. It treats an algorithm like a black box: given the input, the black box give out the output (prediction) based on certain optimization (accuracy) rules. The generalization, interpretation, uncertainty as well as further inference is not very important.

- To many statisticians, the Black Box mechanism is less than satisfactory, because of the difficulties in interpreting such models and the lack of consideration of **uncertainty**.

Statistical Learning (I)

- Statistical Learning is a branch of **applied statistics** that emerged in response to machine learning.

- Statistical Learning is the formalization of relationships between variables in the form of **mathematical model**. It will consider the **assumption** of model, the **generalization** of the model to larger dataset. It emphasizes the model **interpretability**, **precision** and **uncertainty (randomness)**.

Statistical Learning (II)

- Instead of a Black-box in Machine learning, Statistical learning will try to understand the “cogs” inside the box and their interactions.
- Statistical learning will also understand the **underlying distribution** of data, **statistical properties** of the estimator and parameters (**probability and inference**).

Histories

- Machine Learning is the science of getting computers to act without being explicitly programmed thus no human interaction. It was defined more recently by computer scientists like Arthur Samuel and Tom Mitchell in the mid-to-late 1950's and flourishing in the 1990's.

- Statistical models are more about mathematics, model building and require the modeller to understand the relation between variables. The theory on statistical modelling has been around for centuries (Least squares fitting were independently derived by Gauss and Legendre at beginning of 19th century).

Both are "Learning From the Data"

- There is no clear-cut boundary between the two. Nowadays, machine learning and statistics techniques are converging more and more and have much overlap.
- Both are part of data science: extraction of knowledge from data, using ideas from mathematics, statistics, machine learning, computer science, engineering, ...
- Discussion:

  https://www.r-bloggers.com/whats-the-difference-between-machine-learning-statistics-and-data-mining/

Supervised and Unsupervised learning

Notations

- Output variable ***Y***, input variables **X**
- *Y*: **outcome** measurement, correspond to a *output* variable. *Y* is also called dependent variable, response, target.

- **X**: $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ which contains p **predictor** measurements, each measurement $X_i, i = 1,2, \ldots p$ corresponds to a *input* variable. **X** is also called regressors, covariates, features, independent variables.

Supervised Learning (I)

- We can observe both the predictors **X**, and the response *Y*. The observation are **training set.**
- Given a training set, we will estimate a model that relates **X** and *Y*: **learn** a mapping function *f* from **X** to *Y*.
- The model is then used for further **prediction** and **inference.**

Supervised Learning (II)

- Learning from the training set is similar to a teacher supervising the learning process: The teacher provides solution/feedback (**Output**) to students for each questions (**Input**).

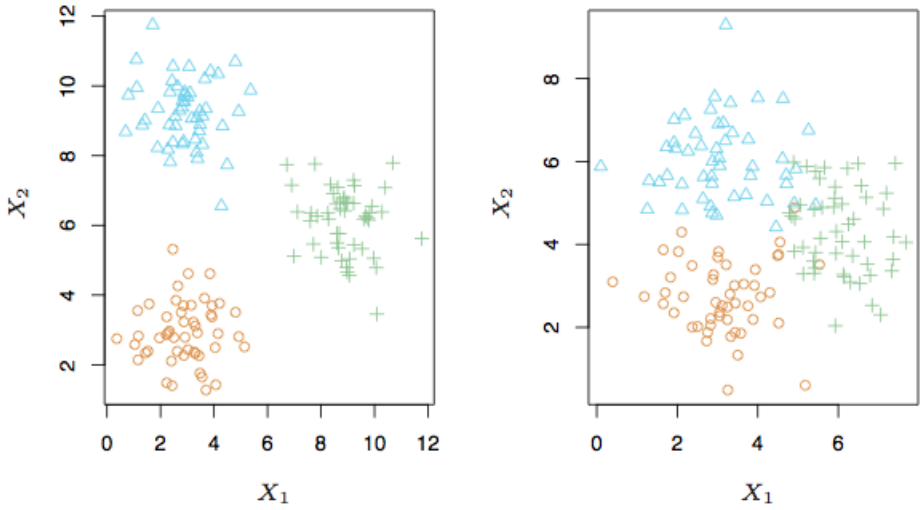
- The students can give solutions/feedback for any new question after sufficient training (**prediction**), and also consider the relationship as well as uncertainty in the solution (**inference**).

Unsupervised Learning (I)

- Only some feature variables **X** are observed, no solution/feedback (output) is given.
- One aim is to divide the observations into relatively distinct groups, such as **clustering**:
  - Divide 1,000,000 different genes into groups that are related by the similarity of proteins.
  - Market segmentation where we try to divide potential customers into groups based on their characteristics.

Two Simple Clustering Examples, $p = 2$

- Both have 150 observations, on two feature variable. Those observations need to be divided into 3 groups
- We do not know which group each observation should belong to
- After unsupervised learning, get following clustering result



Left: *The three groups are well-separated.*
Right: *There is some overlap among the groups.*

Unsupervised Learning II

- We are **not** interested in prediction: no *Y* here.
- Another aim is to model the underlying structure or distribution in the data in order to learn more about the data, eg. principle components analysis (PCA).
- Certain unsupervised learning can be viewed as a "pre-processing" before a supervised learning. Especially for big data set: PCA is a dimension reduction method.

Regression vs. Classification

- Supervised learning problems can be further divided into regression and classification problems.
- Regression: Y is continuous/numerical, eg.:
  - The value (Y) of a given house based on the size (X_1) and building years (X_2) .
  - The heights (Y) of 18 year girl given her mother's height (X_1) and father's height (X_2).
- Classification: Y is discrete/categorical, eg.:
  - Given a tumor size (X_1) and patient's age (X_2) as predictor variables $\mathbf{X} = (X_1, X_2)$, whether the tumor is malignant (Y=1) or benign (Y=0).

Introduction to Supervised Learning

Training dataset (I)

- Let y_i denote the observed value for output variable of the ith individual, where $i = 1,2,\ldots,n$.
- Let $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ denote the observed values for the p input variables of the ith individual, where $i = 1,2,\ldots,n$.
- We have observed data for n individuals, then $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$ construct a training data set.

Training dataset (II)

- We believe that there is a relationship between Y and **X,** where:

$$Y = (y_1, \ldots, y_n)^T$$

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}.$$

- We use $X_j = (x_{1j}, \ldots, x_{nj})^T$ to denote the observed individual data for the jth input variable, where $j = 1,2,\ldots p$.

(H.L. Handwriting on the Lecture)

Model

- The **relationship** can be mathematically described as $Y = f(\mathbf{X}) + \varepsilon$: f is an unknown function and ε is a random error with mean 0:

(1) f represents the *systematic* information that **X** provides about *Y, it is fixed but unknown.*

(2) The random error ε is independent with **X** and is **irreducible** error. This randomness is where the statistical probability theory comes to our aid.

(The source of random error ε H.L.)

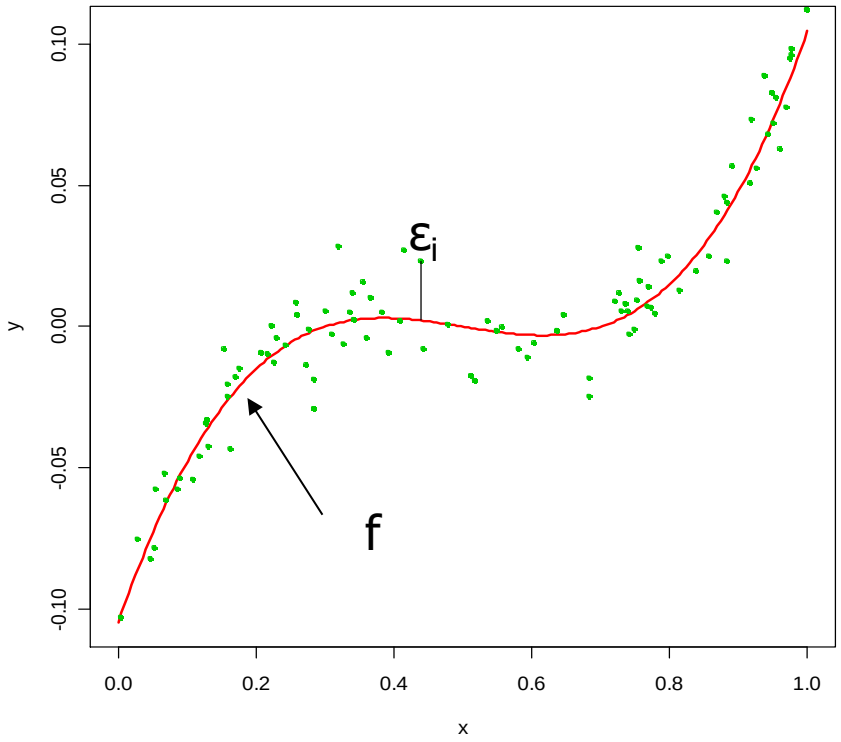
(3) If the mean of ε is not 0, we can always move the constant mean into f and get a 0 mean.

Model Estimation

- Estimate f !!
  - Statistic learning is to use training data and statistical method to estimate f for further prediction and inference.
  - Assumptions of data distribution and model structure can be important.
  - The estimated function is represented as $\hat{f}$.
  - The whole course is about how we can get a "good" estimation $\hat{f}$ by proper and rigorous statistical methodology.

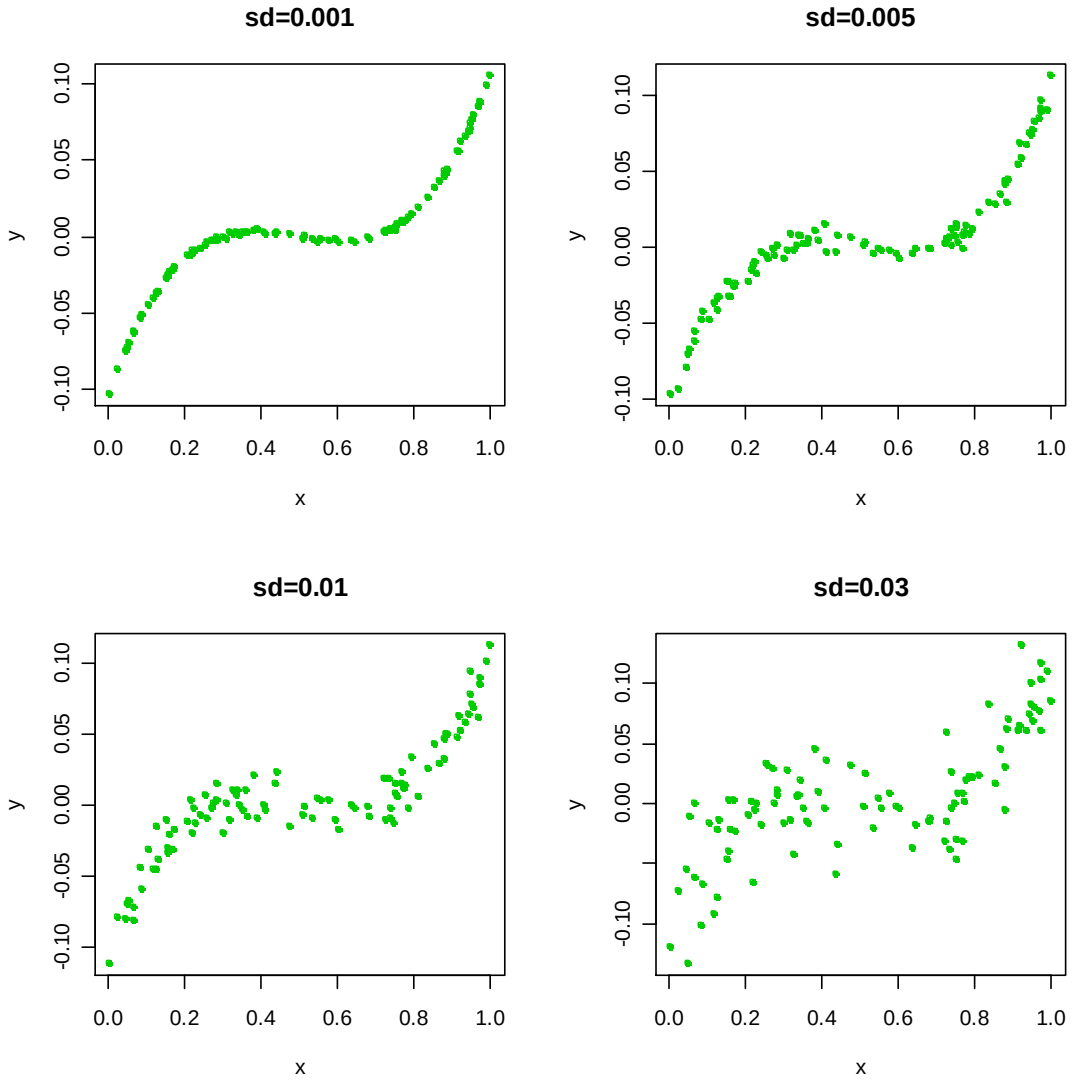
A Simple Example p=1, simulated data

$$Y = f(\mathbf{X}) + \varepsilon, \ \varepsilon \sim \mathrm{N}(0, \sigma^2)$$

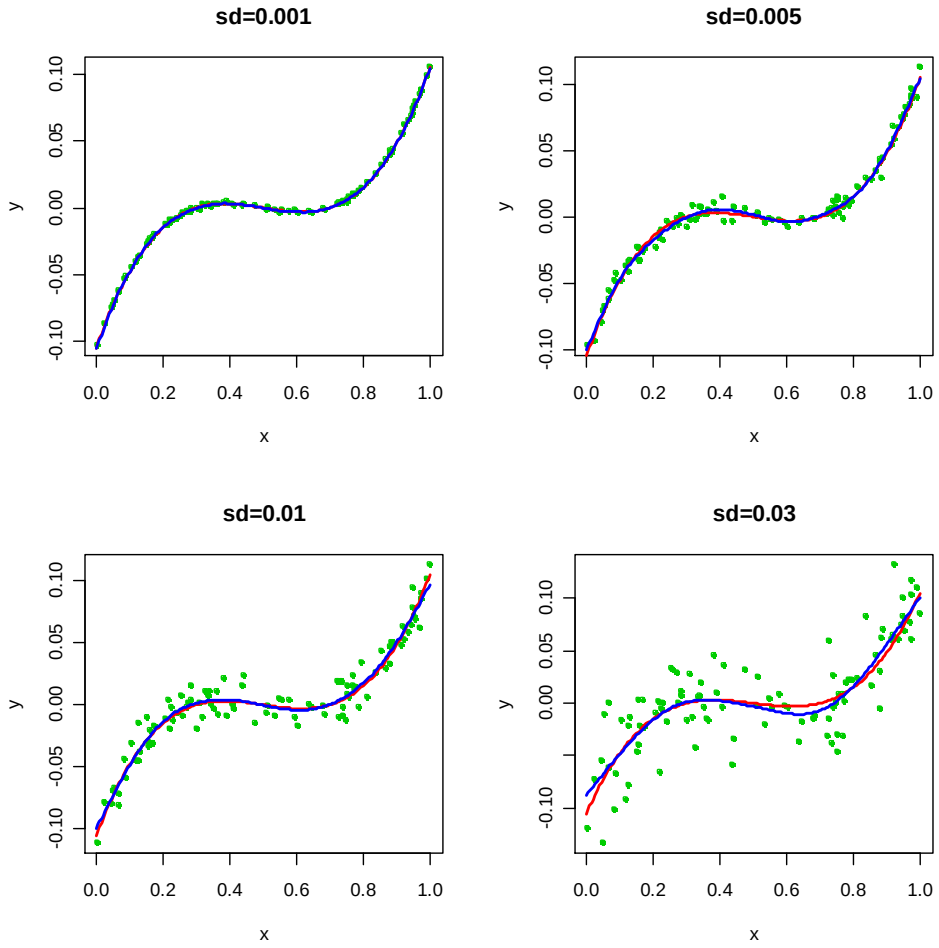


Different standard deviation of ε

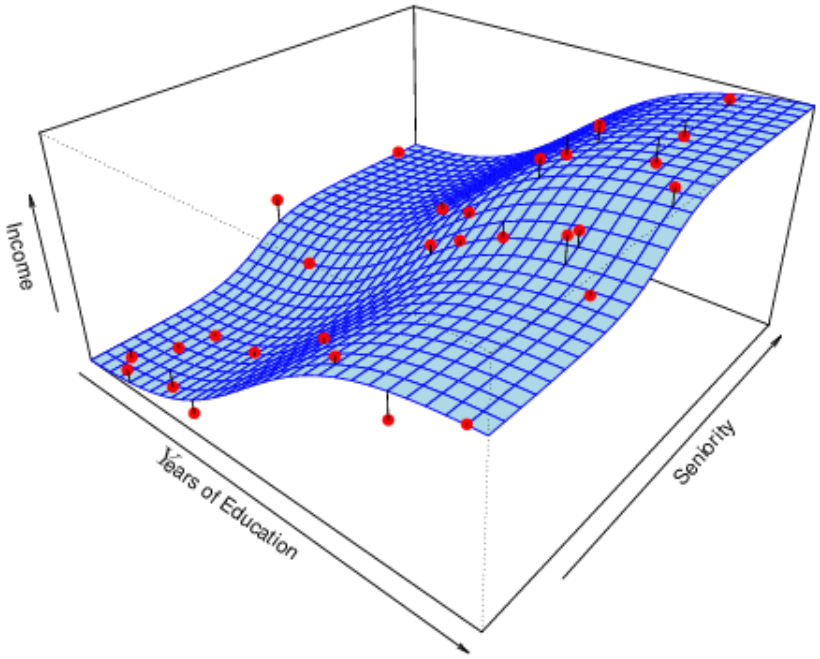
- The difficulty of estimating f will depend on the standard deviation of the ε's:



Different Estimation $\hat{f}$



Example with $p = 2$, n=30, simulated data. Income (Y) vs. Education (X_1) and Seniority (X_2)



f is the blue surface. Red points are the 30 observations, vertical lines are the error term. As data here is simulated, we know the shape of f.

Aim of Model Estimation---Prediction and Inference

Prediction

- A "good" estimation $\hat{f}$ can make "accurate" predictions for the response based on new input **X:** $\hat{Y} = \hat{f}(\mathbf{X})$.
  - (1) In many situations, inputs **X** can be observed, but the output Y cannot be easily obtained (Income example).
  - (2) The accuracy of prediction $\hat{Y}$ can be measured by the expected squared difference (error) between $\hat{Y}$ and Y: $E\left(Y - \hat{Y}\right)^2$, which can be decomposed into "***reducible error***" and "***irreducible.***

(H.L.)

- If the model is only for prediction, $\hat{f}$ can be a "black box" and we need not to know the detail form: just input **X**, and output prediction $\hat{Y}$. This is a common situation in Machine Learning.

Inference

- We may also be interested in the type of relationship between Y and $\mathbf{X}$:
  - How do we consider the uncertainty of the response?
  - 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.?
- In this case, the $\hat{f}$ should not be "Black box", and we often need to give out a mathematic formulation of $\hat{f}$.

Example: Housing Inference

- Wish to predict house price based on $p = 14$ variables.
- 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.

Approaches for Model Estimation

How Do We Estimate f ?

- We have observed a set of **training data:**

$$\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$$

where $\mathrm{x}_i = (x_{i1}, \ldots, x_{ip})$ is the observed input values for the ith individual, $i = 1, \ldots, n$.

- The model construction and estimation process is the Statistical Learning process:
  - Parametric Methods
  - Non-parametric Methods

Parametric Methods

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

STEP 1:

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

$$f(\mathbf{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p,$$

$$Y = f(\mathbf{X}) + \varepsilon$$

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.

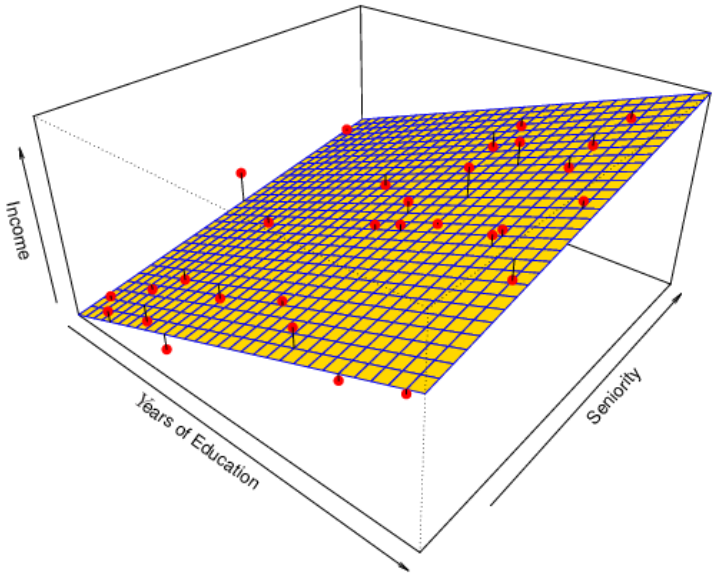
- As long as β_0, β_1, β_2,…, β_p are estimated, the model is specified.

- Common approaches for estimating the parameters in a linear model: ordinary least squares (OLS) and maximum likelihood method (MLE))

Example: A Linear Regression Estimate

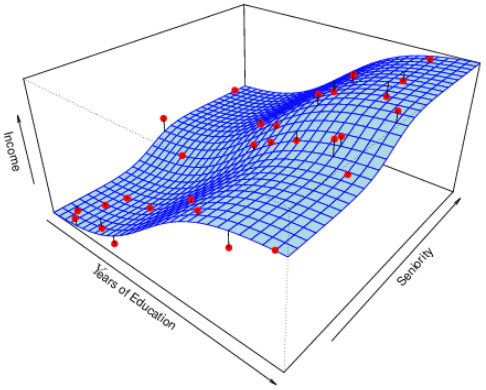
$$f = \beta_0 + \beta_1 \times \text{Education} + \beta_2 \times \text{Seniority}$$

- In real life, there is almost no cases where the relationship is pure linear: linear model is **underfitting** the data
- But the linear model is very easy to estimate and interpret.



- Disadvantage of Parametric model:

  If the chosen model form is far away from the real form *f*, then the estimation will be poor .



Non-parametric Methods (I)

- Do not make explicit assumptions about the functional form of *f*.
- Seek an estimate of *f* as close to the training data set as possible, under certain smooth restrictions.
- Advantages:
  - Fit a wider range of possible shapes of *f*, less restriction, more **flexible**
  - Can *potentially* provide more accurate estimates

Non-parametric Methods (II)

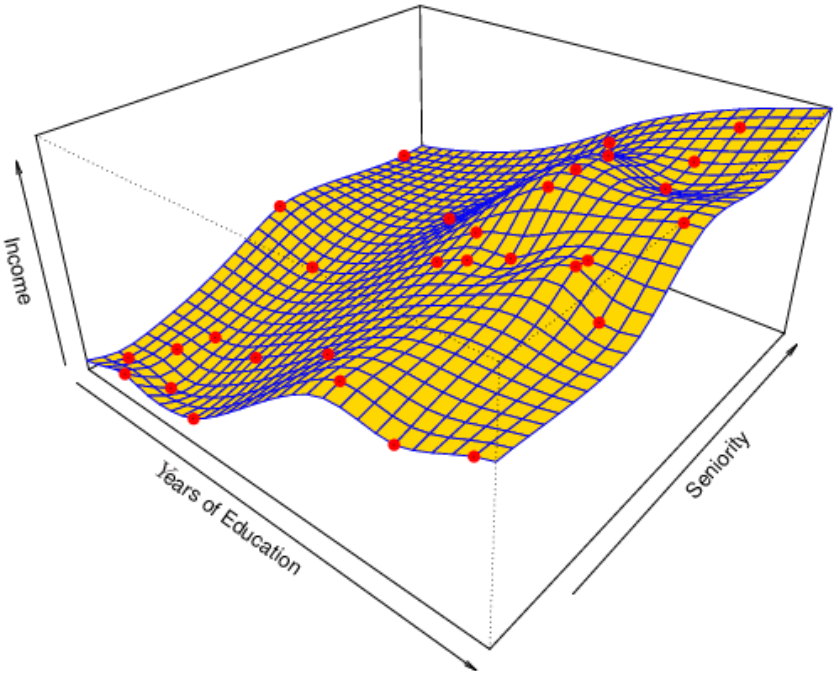
- Disadvantages:
  - Large number of observations is required to obtain an accurate estimate of *f*.
  - Suffered from “curse of dimensionality”.
  - Hard to interpret the result and give out further inference: black box.
  - Can result in overfitting, and can not be used in prediction.

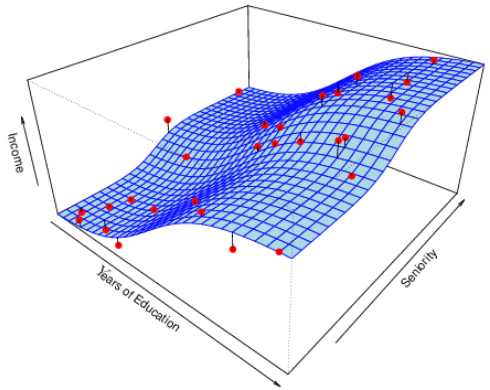
A Poor Estimate: rough thin-place spline fit

- Thin-plate spline fit **without** smooth restriction.

(H.L.)

- It makes no errors on the training data: it even fit the random error as the systematical information *f*.

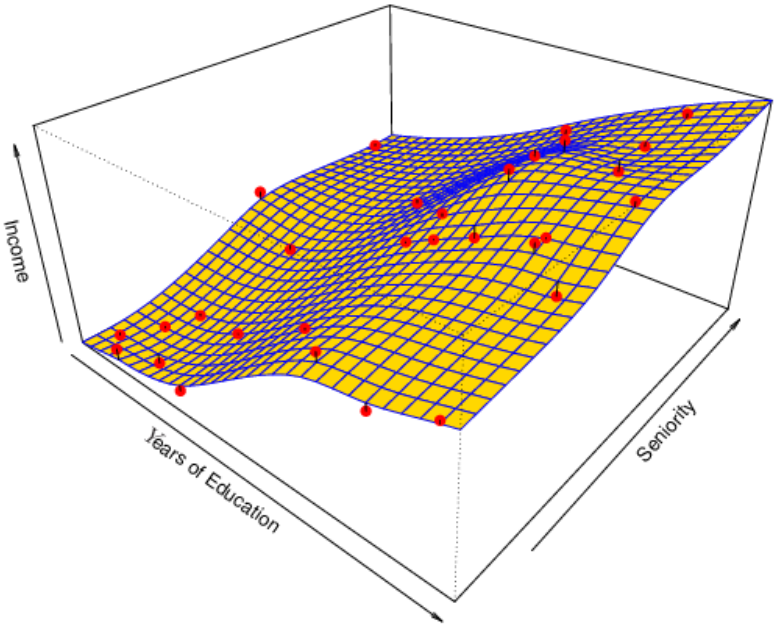


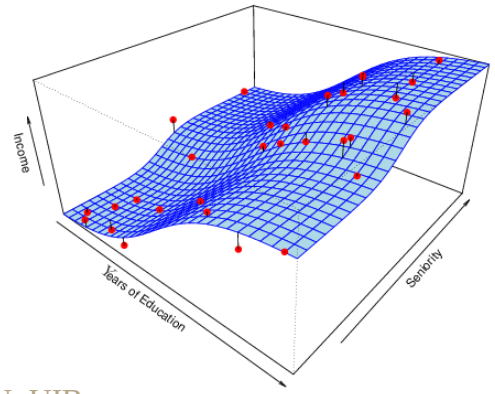


A smooth thin-plate spline fit

- It does not assume any pre-specified functional form of model.
- It just give a estimate for *f* which is close as possible to training data—subject to a **smooth plate restriction**.

  (H.L.)
- It is more accuracy than the linear model: follow the original shape of *f* better.





Key points in chosen of $\hat{f}$ (I)

- For the same dataset, we may have different approaches to estimate f based on different assumption of f, which lead to different structure of estimation $\hat{f}$.
- Choosing among different $\hat{f}$ is Trade-Off between:
  - Prediction accuracy and Model interpretability.
  - Parsimony and Black box.
  - Training MSE and test MSE.
  - Bias and Variance.

Key points in chosen of $\hat{f}$ (II)

- Try to avoid over-fit or under-fit: how can we choose a $\hat{f}$ which can achieve a «good» fit?
  - No one approach/model dominates others over all possible data.
  - Select the best approach/model is one of the most challenging parts of performing Statistical Learning.

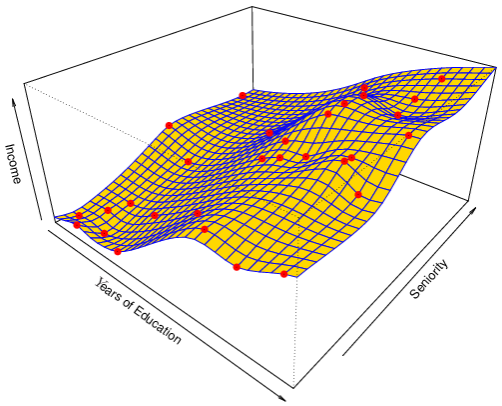
Assessing Model Accuracy

Outline

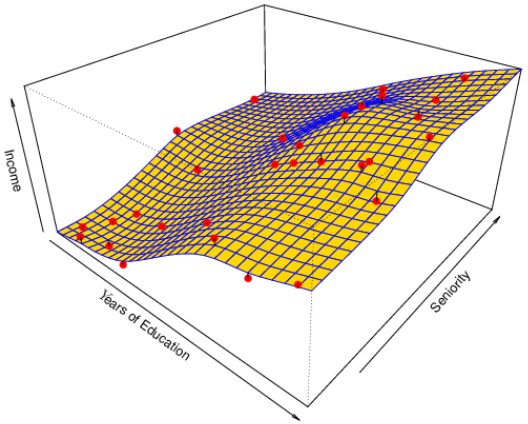
- Assessing Model Accuracy
  - Measuring the Quality of Fit
  - The Bias-Variance Trade-off
  - The Classification Setting

Model structure selection (I)

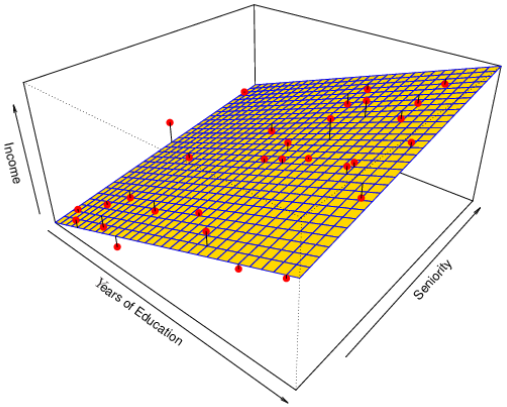
Rough Spline model



Smooth Spline model



Linear model



Model structure selection (II)

- If we have several suggestions of the model structure: several candidate estimations $\hat{f}$:
  - More flexible model has less restrictions and is often more complex and "wiggly": eg. Smooth thin – plate spline.
  - Less flexible model has more restrictions and is often more parsimony and "smooth": eg. Linear regression.
- Trade-Off between flexible (complex) and smooth (simple) models is trade- off between:
  - Fitting accuracy and model interpretability.
  - Fitting accuracy and prediction accuracy.

Measuring Quality of Fit for $\hat{f}$ (Regression)

- One common measure of accuracy of $\hat{f}$ is the mean squared error (MSE) i.e.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2;\ i = 1,...,n.$$

Where $\hat{y}_i = \hat{f}(\mathbf{x}_i)$ is the fitted value/prediction based on the estimated model $\hat{f}$ for the ith individual.

?Why we use $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2;\ i = 1,\ldots,n$ instead of $\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|;\ i = 1,\ldots,n$ to measure the error?

Training MSE and test MSE

- Given **training data** set and model structure, we aim at estimate a model which give out lowest *training MSE*: eg. with linear regression we choose the line such that MSE is minimized.

(H.L.)

- What we **really** care about is how well the method can be **generalized** to new data (data which is not used to fit the model). We call this new data "**Test Data**" which produce *test MSE*.

(H.L.)

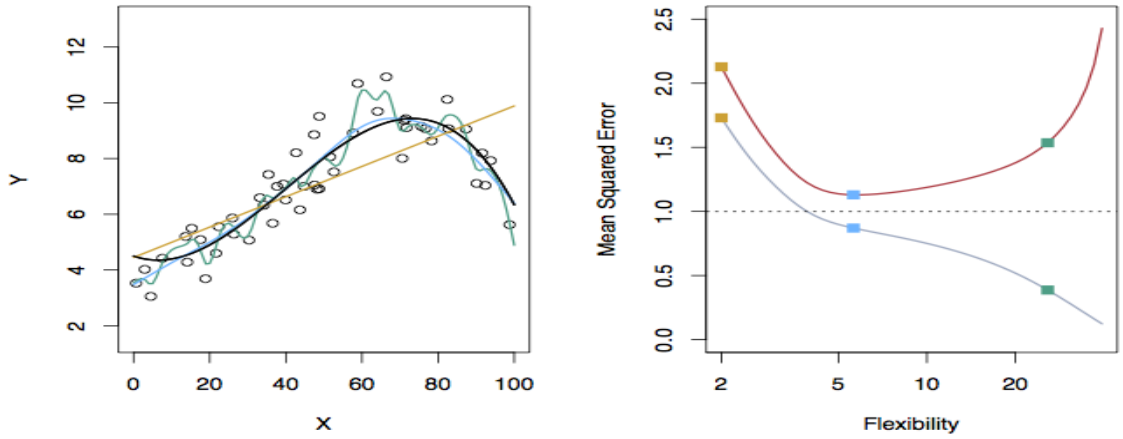
Training MSE

- More flexible/complex model often gives lower training MSE as it can generate a wider range of possible shapes to fit the training dataset.
- More flexible model has higher “fitting accuracy” of the training dataset.
- More flexible model can be harder to interpret than smoother model (eg. Income example).

Testing MSE

- More flexible model can give higher **test MSE** (bad!).
- We often want a model with **lowest test MSE** which can give out best prediction.
- Test MSE against degree of flexibility: U shape.

Example with Different Levels of Flexibility:
Simulated both training data and test data from $Y=f(X)+\varepsilon$



LEFT

Black line: Truth f

Based on training dataset, we try three different estimates $\hat{f}$:

Orange line: Linear Estimate, flexibility is 2.

Blue line: smoothing spline, flexibility is 5

Green line: less smoothing spline, flexibility is 25

RIGHT

Grey: Training MSE calculated by training data set (data set used to estimate $\hat{f}$) .

RED: Test MSE calculated by test data set (data set not used to estimate $\hat{f}$).

Dashed: Minimum possible test MSE (irreducible error).

Bias/ Variance Trade-Off

- Trade- Off between smooth and flexible model corresponds to the trade - off between **bias** and **variance**.
- If our model is too “simple/smooth” and has very few parameters, then it may have large bias (but small variance); if it is too “complex/flexible” and has very many parameters, then it may suffer from large variance (but have smaller bias).
- High bias correspond to underfitting, high variance correspond to overfitting.

Bias of Model (I)

- Bias refers to the error that is introduced by modelling a data/real life problem by a much simpler model.
- A estimated model with large bias underfit the true model and fail to capture structure exhibited by the data.

Bias of Model (II)

- For example, linear regression assumes that there is a linear relationship between Y and $\mathbf{X}$. 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 (but not all the cases).

Variance of Model (I)

- Variance refers to how much $\hat{f}$ would **change (vary)** if we had a different training data set, larger change means higher variance.
- A model with high variance will mistakenly fit the random patterns of the training set as systematic information.

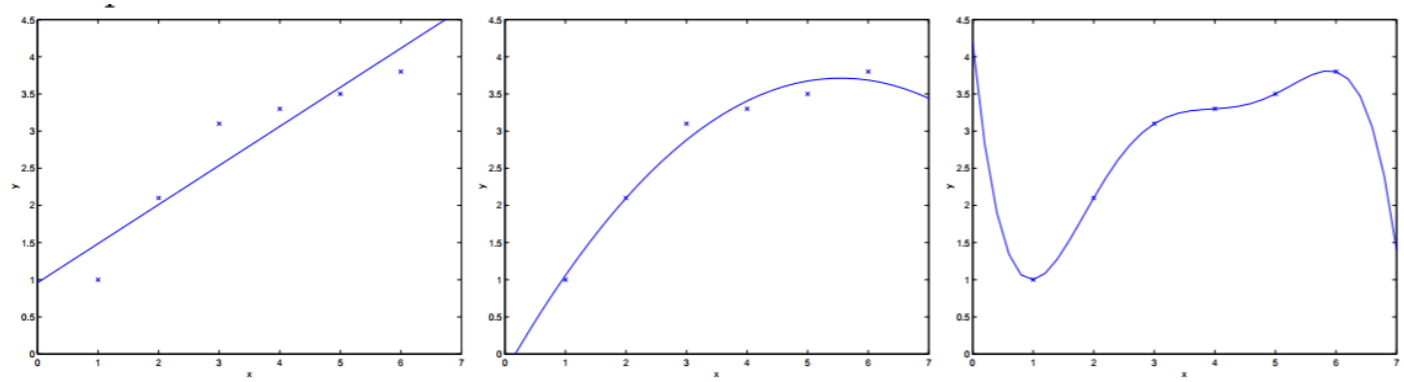
Variance of Model (II)

- Generally, the more flexible a method is the more variance it has, as it is not robust to the random error ε.
- As it will fit patterns in the data that happened to be present in our small, finite training set, but that do not reflect the wider pattern of the relationship between Y and $\mathbf{X}$.

A simple Example of bias and variance

- X: living area of a house, Y: price of a house.
  We have 6 data as training set, and 3 different $\hat{f}$

Linear model | 2nd order polynomial | 5th order polynomial



(H.L.)

- One tips to check if the estimated model $\hat{f}$ is suffered from high bias.

(H.L.)

Test MSE, bias and variance (I)

- For any given new input x_0, the expected test MSE is defined as:

$$\text{Expected Test MSE} = E\left(y_0 - \hat{f}(x_0)\right)^2$$

$$= [\text{Bias}(\hat{f}(x_0))]^2 + \text{Var}(\hat{f}(x_0)) + \underbrace{\text{Var}(\varepsilon)}_{\text{Irreducible Error}}$$

(Proof please refer to material under «Lecture note»)

- Expected test MSE refers to the average test MSE that we would obtain if we repeatedly estimated f using a large number of training sets, and tested each at x_0.

Test MSE, bias and variance (II)

- For model $\hat{f}$ which is more complex, the bias of $\hat{f}$ will decrease and the variance of $\hat{f}$ will increase but expected test MSE may go up or down, based on the rate of changes for bias and variance.

(H. L.)

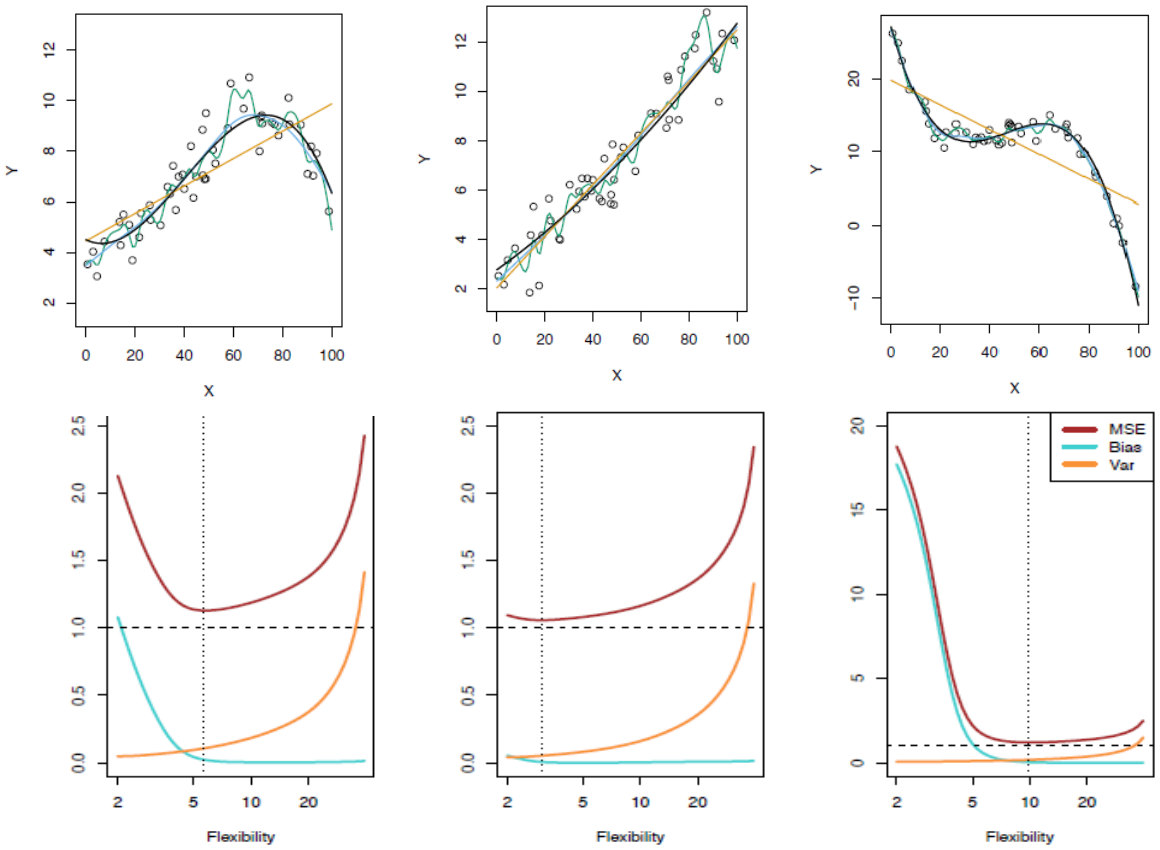
- We want a model with lowest test MSE for prediction, thus we need to distinguish whether **bias** or **variance** is the problem contributing to bad predictions.

- How the test MSE change with the model complexity/flexibility will depend on the data structure.

Test MSE, Bias and Variance (II)

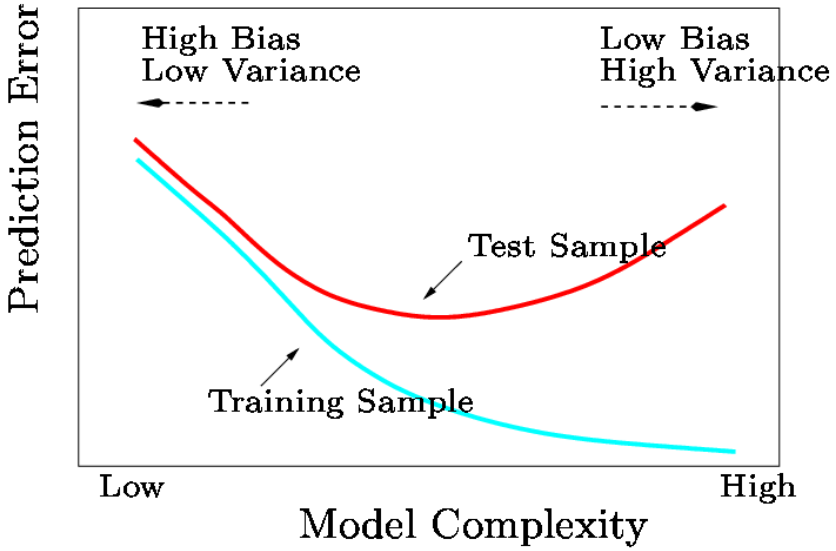
Horizontal dashed line: Var(ε)

Blue curve: Bias ($\hat{f}(\mathbf{x}_0)$), Orange curve: Var ($\hat{f}(\mathbf{x}_0)$), red curve: test MSE



A Fundamental Picture

- In general training errors will always decline when model is more complex.
- 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!

The Classification Setting

- For a regression problem, we used the MSE to assess the accuracy of the statistical learning method.
- For a classification problem we can use the error rate i.e.

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

- $I(y_i \neq \hat{y}_i)$ is indicator function, which will give 1 if the condition $y_i \neq \hat{y}_i$ is correct, otherwise it gives a 0.
- Thus the error rate represents the fraction of incorrect classifications, or misclassifications.