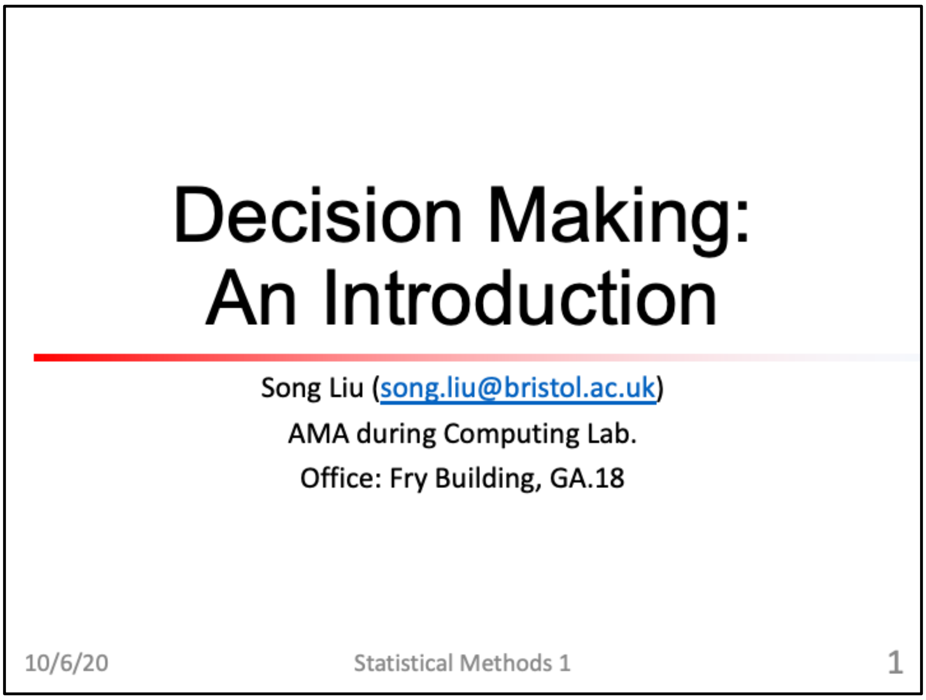
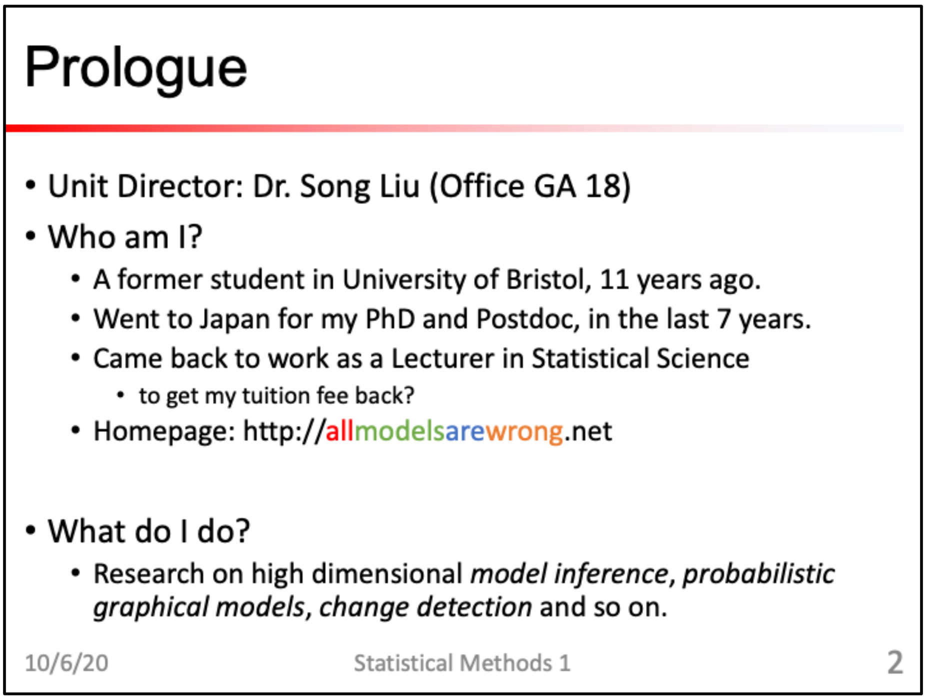


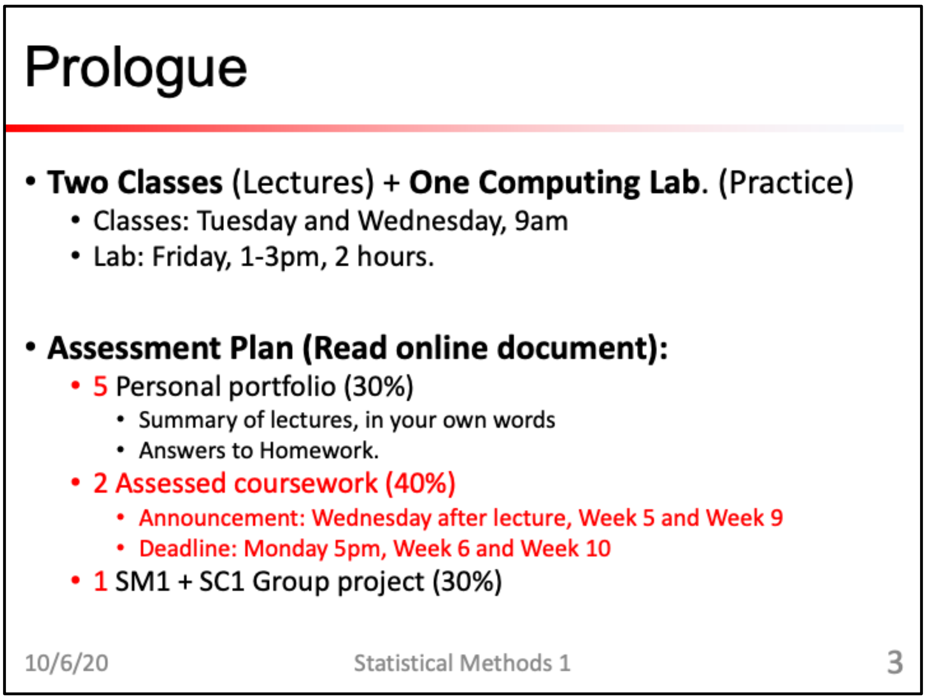
Decision Making: An Introduction

Song Liu (song.liu@bristol.ac.uk)
AMA during Computing Lab.
Office: Fry Building, GA.18



Prologue

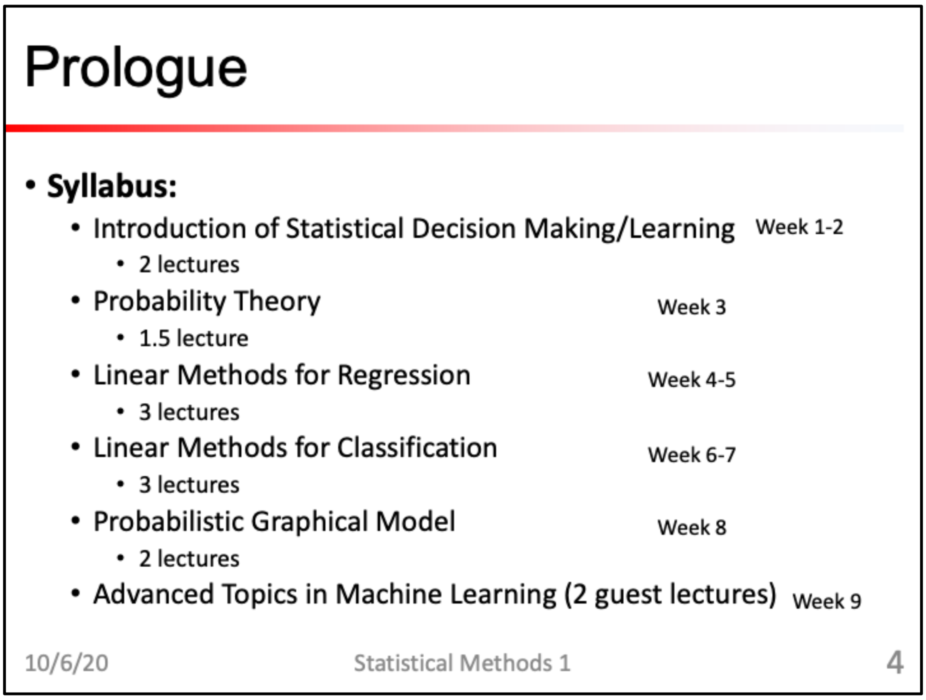
- Unit Director: Dr. Song Liu (Office GA 18)
- Who am I?
  - A former student in University of Bristol, 11 years ago.
  - Went to Japan for my PhD and Postdoc, in the last 7 years.
  - Came back to work as a Lecturer in Statistical Science
    - to get my tuition fee back?
  - Homepage: http://allmodelsarewrong.net
- What do I do?
  - Research on high dimensional *model inference*, *probabilistic graphical models*, *change detection* and so on.



Prologue

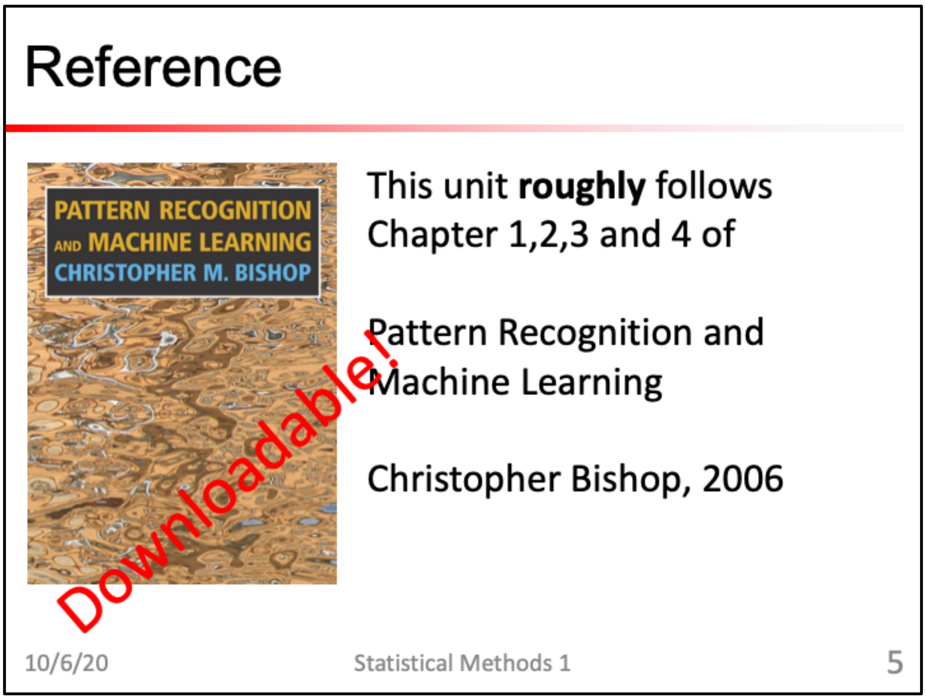
- **Two Classes** (Lectures) + **One Computing Lab**. (Practice)
  - Classes: Tuesday and Wednesday, 9am
  - Lab: Friday, 1-3pm, 2 hours.

- **Assessment Plan (Read online document):**
  - 5 Personal portfolio (30%)
    - Summary of lectures, in your own words
    - Answers to Homework.
  - 2 Assessed coursework (40%)
    - Announcement: Wednesday after lecture, Week 5 and Week 9
    - Deadline: Monday 5pm, Week 6 and Week 10
  - 1 SM1 + SC1 Group project (30%)



Prologue

- **Syllabus:**
  - Introduction of Statistical Decision Making/Learning Week 1-2
    - 2 lectures
  - Probability Theory Week 3
    - 1.5 lecture
  - Linear Methods for Regression Week 4-5
    - 3 lectures
  - Linear Methods for Classification Week 6-7
    - 3 lectures
  - Probabilistic Graphical Model Week 8
    - 2 lectures
  - Advanced Topics in Machine Learning (2 guest lectures) Week 9



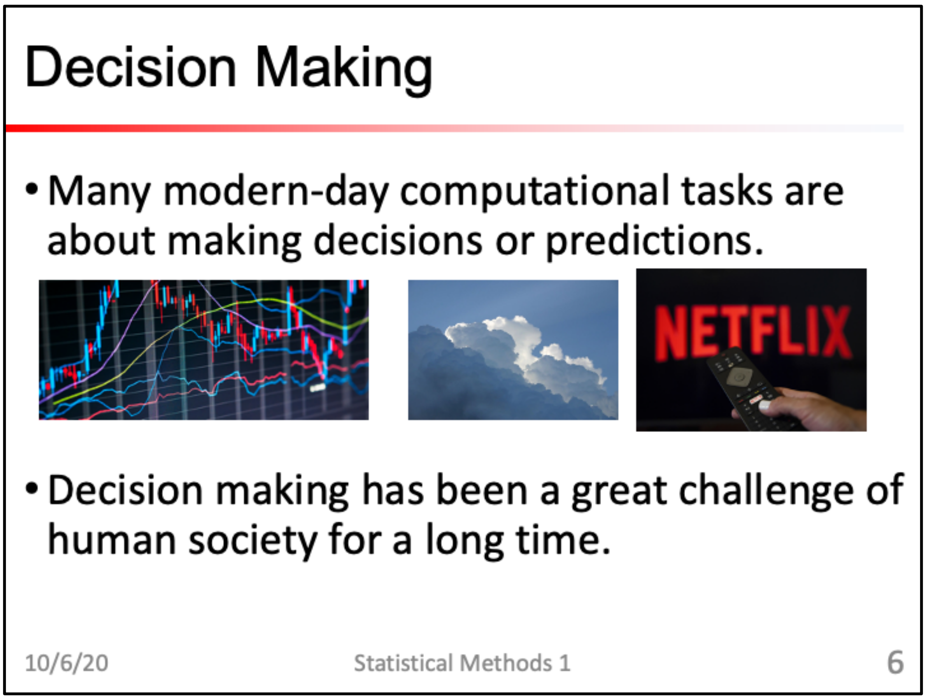
Reference

Downloadable!

This unit **roughly** follows Chapter 1,2,3 and 4 of

Pattern Recognition and Machine Learning

Christopher Bishop, 2006



Decision Making

- Many modern-day computational tasks are about making decisions or predictions.

- Decision making has been a great challenge of human society for a long time.

◀◀ A Look back … in China

"Oracle Bones"

- Emperor has a **question**.
- **Write it down** on the bones of large animals and **toss** it to **flame**.
- **Cracks on bones** reveal "Gods' will".
- **Priest deciphers** the patterns of cracks and provides an answer.

◀◀ A Look back … in Greece

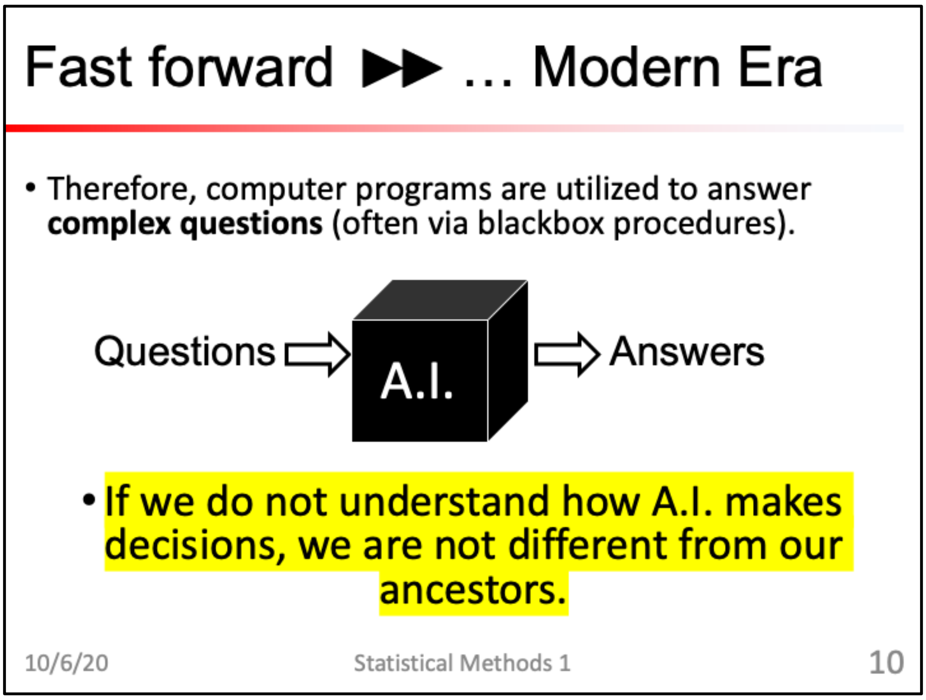
Pythia

- Supplicant has a **question**.
- He/she travels to Delphi **asks Pythia**.
- **Pythia** inhales vapors at Temples of Apollo, speaks gibberish.
- Priest deciphers her gibberish and provides supplicant an answer.

Fast forward ▶▶ … Modern Era

- No one believes in Pythia or Oracle bones anymore.
- However, the modern-day society faces another great challenge on decision making.

- Human cannot digest information fast enough and make rationalized decisions.



Fast forward ▶▶ … Modern Era

- Therefore, computer programs are utilized to answer **complex questions** (often via blackbox procedures).

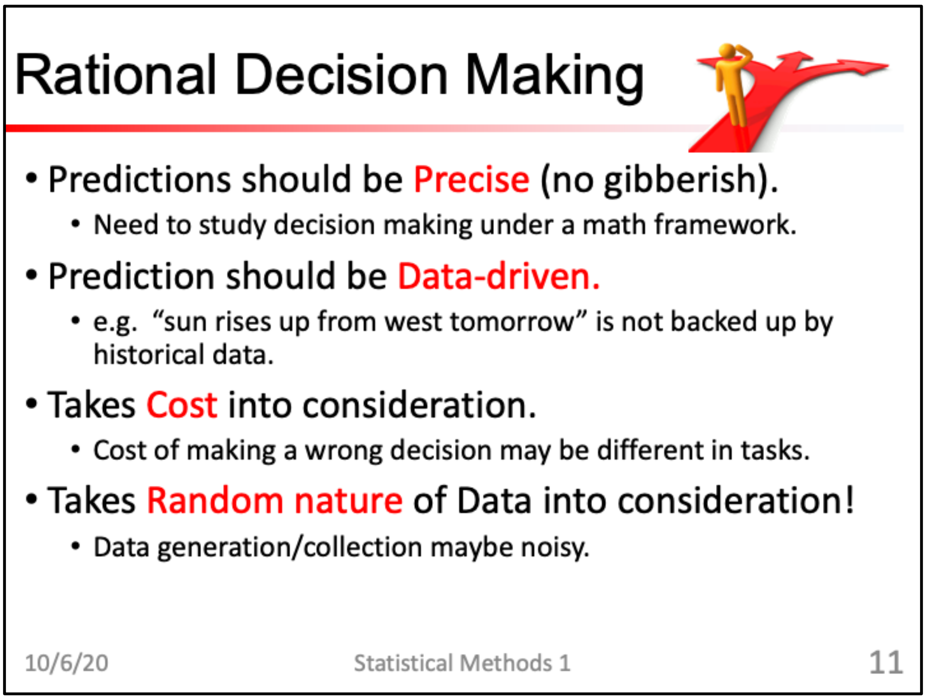
Questions ⇨ A.I. ⇨ Answers

- **If we do not understand how A.I. makes decisions, we are not different from our ancestors.**

10/6/20 Statistical Methods 1 10

Can we say we are doing better than our ancestors who are looking for answers from cracks of bones?

The real issue of ancient methods is that it lacks a rational process of decision making, so what are the rationale behind decision making?



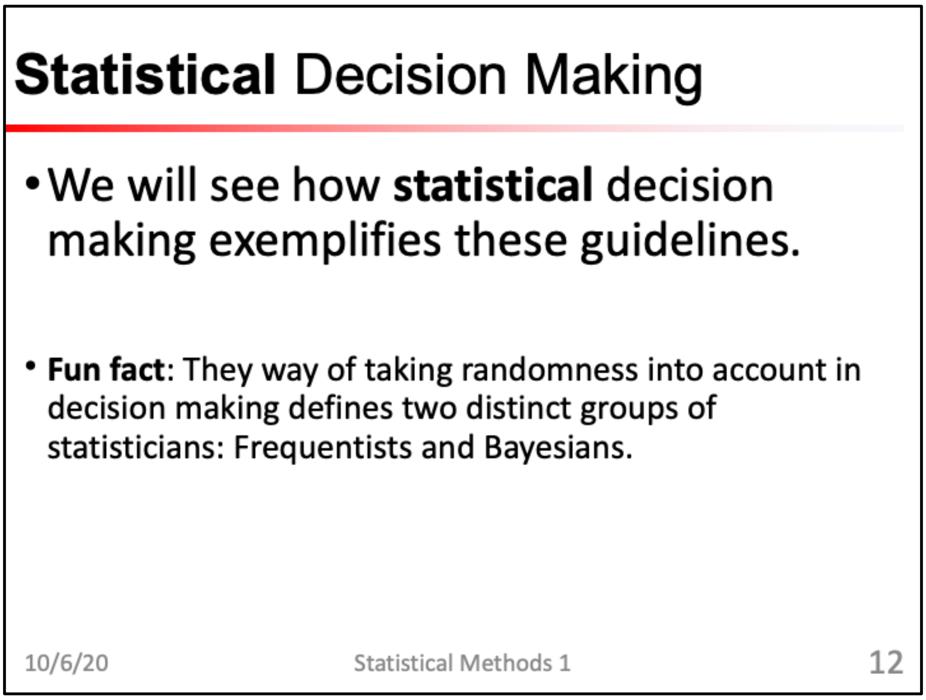
Rational Decision Making

- Predictions should be Precise (no gibberish).
  - Need to study decision making under a math framework.
- Prediction should be Data-driven.
  - e.g. “sun rises up from west tomorrow” is not backed up by historical data.
- Takes Cost into consideration.
  - Cost of making a wrong decision may be different in tasks.
- Takes Random nature of Data into consideration!
  - Data generation/collection maybe noisy.

10/6/20 Statistical Methods 1 11

Preciseness demands mathematics, Data driven requires big data, Cost-sensitive requires us to introduce proper measure of risks and randomness is where statistics comes in.

This is why, modern day data science is interesting: it is at the intersection of mathematics, statistics, big data and risk analysis.

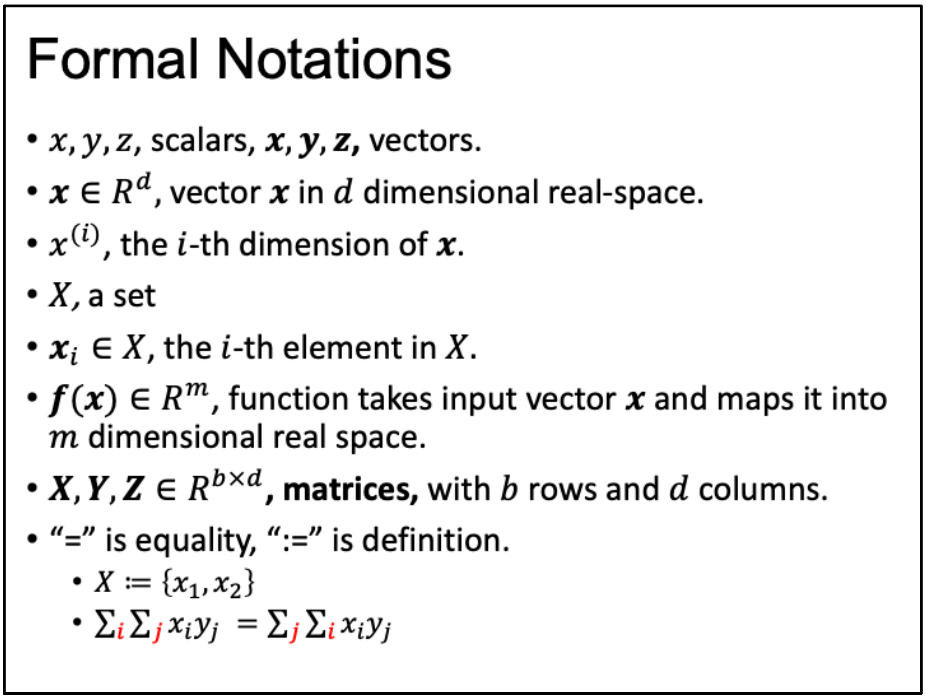


Statistical Decision Making

- We will see how **statistical** decision making exemplifies these guidelines.
- **Fun fact**: They way of taking randomness into account in decision making defines two distinct groups of statisticians: Frequentists and Bayesians.

10/6/20 Statistical Methods 1 12

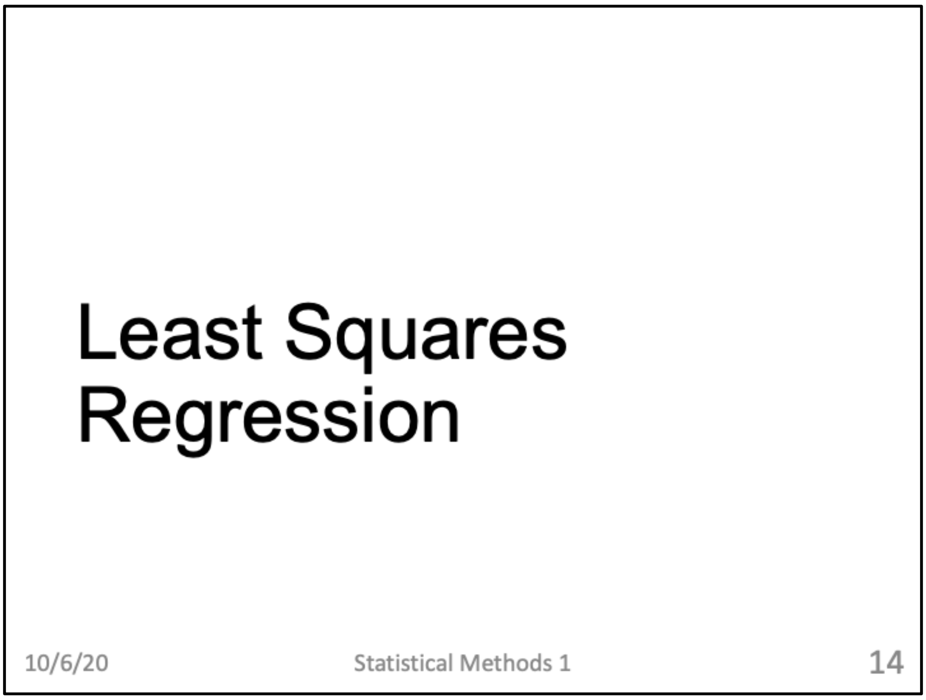
Do not take the division of frequentists and Bayesians too serious now. It does not matter for this lecture.



Formal Notations

- x, y, z, scalars, $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$, vectors.
- $\boldsymbol{x} \in R^d$, vector $\boldsymbol{x}$ in d dimensional real-space.
- $x^{(i)}$, the i-th dimension of $\boldsymbol{x}$.
- X, a set
- $\boldsymbol{x}_i \in X$, the i-th element in X.
- $\boldsymbol{f}(\boldsymbol{x}) \in R^m$, function takes input vector $\boldsymbol{x}$ and maps it into m dimensional real space.
- $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z} \in R^{b\times d}$, **matrices,** with b rows and d columns.
- “=” is equality, “:=” is definition.
  - $X \coloneqq \{x_1, x_2\}$
  - $\sum_i \sum_j x_i y_j \ = \sum_j \sum_i x_i y_j$

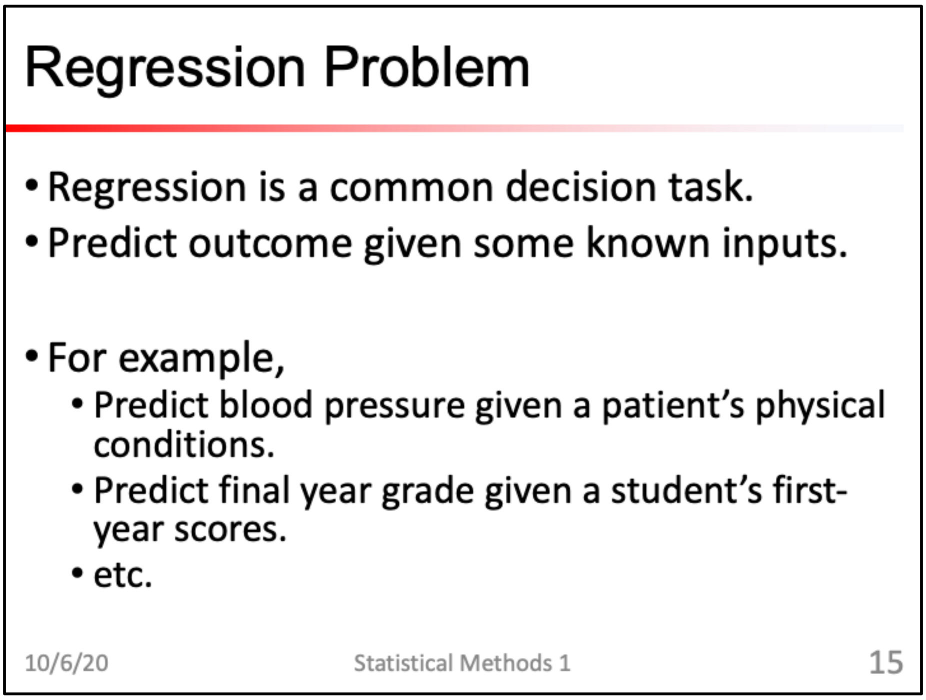
Please use follow these notations in your portfolio, homework and group project reports. Failure to do so might result reduction of scores.



Least Squares Regression

10/6/20 Statistical Methods 1 14

German mathematician Gauss used least squares to discover a dwarf planet Ceres using Kepler’s observational data.



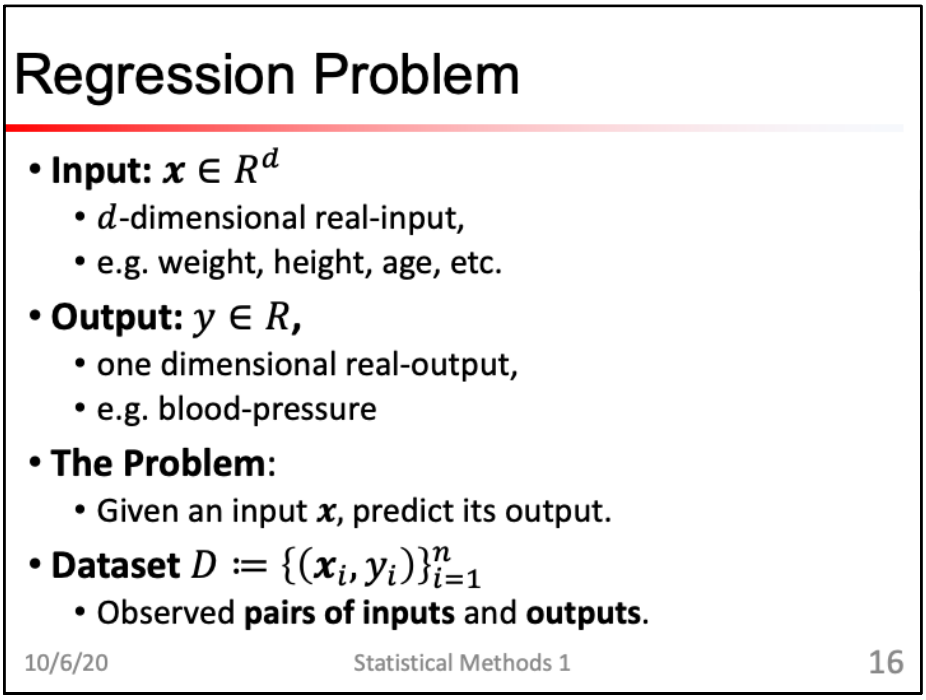
Regression Problem

- Regression is a common decision task.
- Predict outcome given some known inputs.

- For example,
  - Predict blood pressure given a patient’s physical conditions.
  - Predict final year grade given a student’s first-year scores.
  - etc.

10/6/20 Statistical Methods 1 15

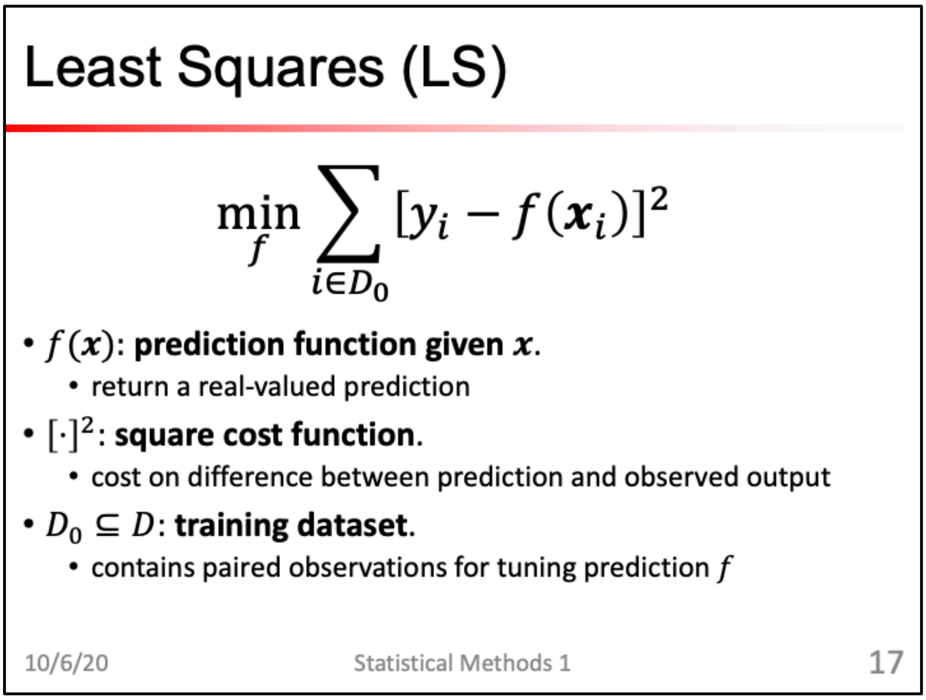
The regression is a very well studied problem.



Regression Problem

- **Input:** $\boldsymbol{x} \in R^d$
  - d-dimensional real-input,
  - e.g. weight, height, age, etc.
- **Output:** $y \in R$**,**
  - one dimensional real-output,
  - e.g. blood-pressure
- **The Problem**:
  - Given an input $\boldsymbol{x}$, predict its output.
- **Dataset** $D \coloneqq \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$
  - Observed **pairs of inputs** and **outputs**.

Let us look at a standard definition machine learning problem



Least Squares (LS)

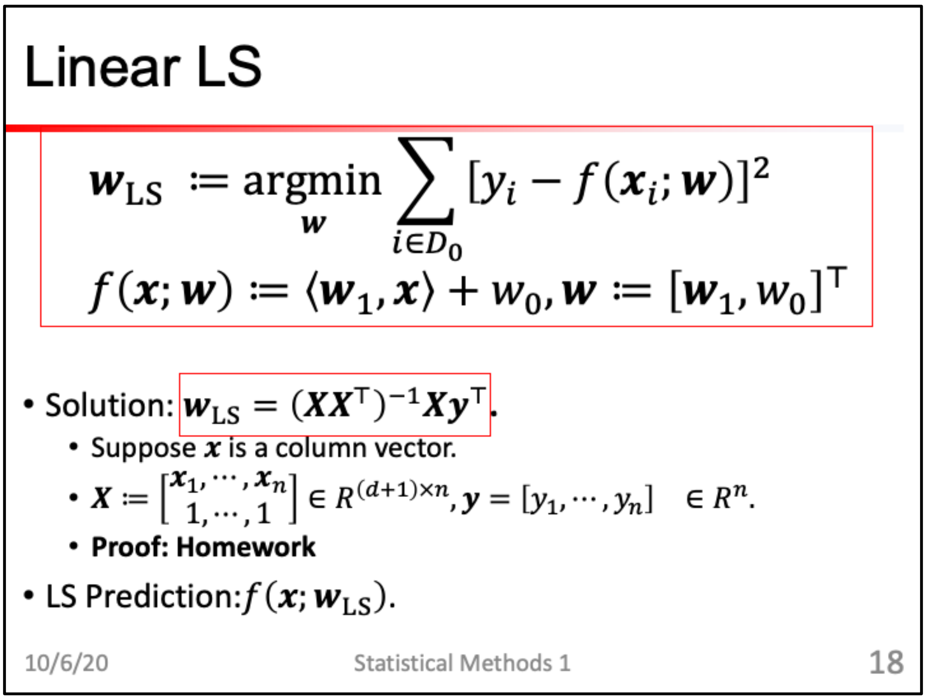
$$\min_{f} \sum_{i \in D_0} [y_i - f(\boldsymbol{x}_i)]^2$$

- $f(\boldsymbol{x})$: **prediction function given $\boldsymbol{x}$.**
  - return a real-valued prediction
- $[\cdot]^2$: **square cost function.**
  - cost on difference between prediction and observed output
- $D_0 \subseteq D$: **training dataset.**
  - contains paired observations for tuning prediction f

10/6/20 Statistical Methods 1 17

This formulation of LS gives us a mathematical procedure of regression problem.

It is clearly cost sensitive, data driven, how does least squares take randomness of data into account?



Linear LS

$$\boxed{\begin{aligned} \boldsymbol{w}_{\mathrm{LS}} &:= \operatorname*{argmin}_{\boldsymbol{w}} \sum_{i \in D_0} [y_i - f(\boldsymbol{x}_i; \boldsymbol{w})]^2 \\ f(\boldsymbol{x}; \boldsymbol{w}) &:= \langle \boldsymbol{w}_1, \boldsymbol{x} \rangle + w_0, \boldsymbol{w} := [\boldsymbol{w}_1, w_0]^\top \end{aligned}}$$

- Solution: $\boxed{\boldsymbol{w}_{\mathrm{LS}} = (\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}\boldsymbol{y}^\top}$.
  - Suppose $\boldsymbol{x}$ is a column vector.
  - $\boldsymbol{X} := \begin{bmatrix} \boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \\ 1, \cdots, 1 \end{bmatrix} \in R^{(d+1) \times n}, \boldsymbol{y} = [y_1, \cdots, y_n] \in R^n$.
  - **Proof: Homework**
- LS Prediction: $f(\boldsymbol{x}; \boldsymbol{w}_{\mathrm{LS}})$.

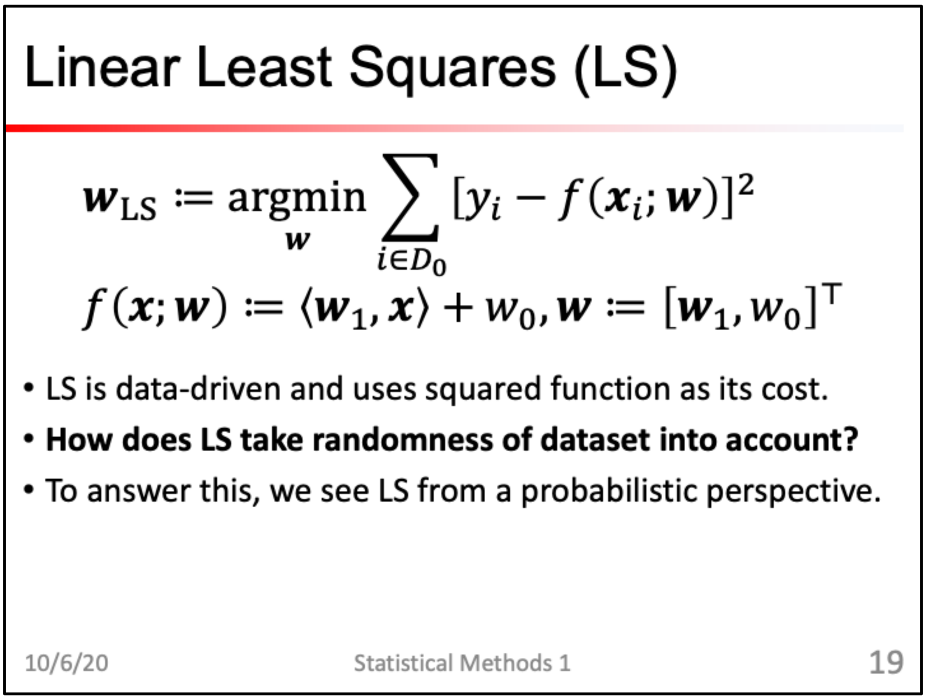
10/6/20 Statistical Methods 1 18

W1 is called **coefficients**, and W0 is called **bias**.

$f(\boldsymbol{x}; \boldsymbol{w})$ means an f function parametrized by parameter $\boldsymbol{w}$.

You can roughly think f assigns different weights along each dimension of $\boldsymbol{x}$ and sum them up.

Hint of the proof: construct $\tilde{\boldsymbol{x}} := [\boldsymbol{x}, 1]^\top$.

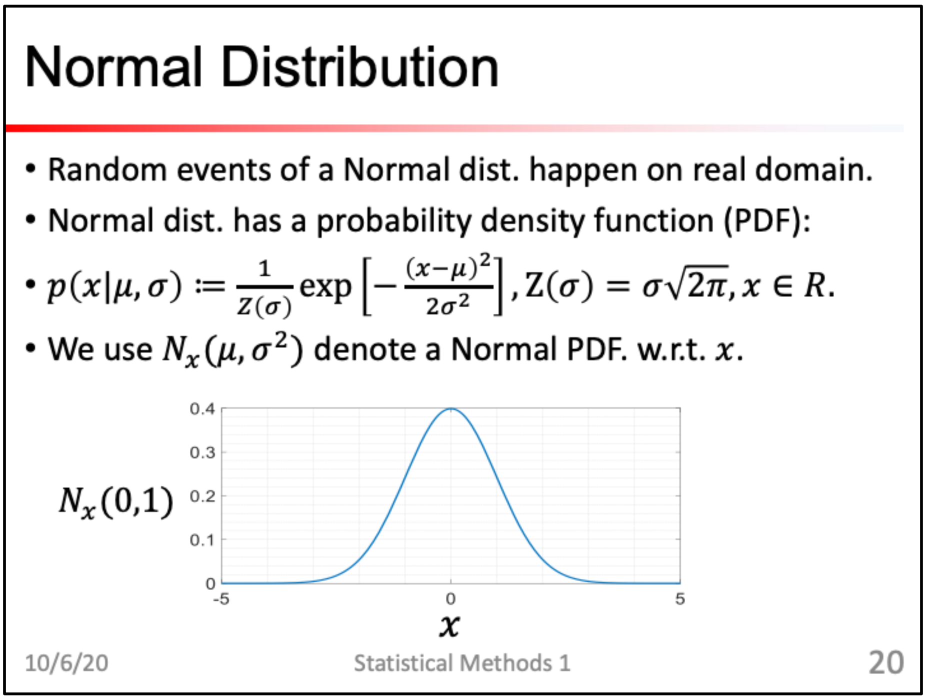


Linear Least Squares (LS)

$$\boldsymbol{w}_{\mathrm{LS}} \coloneqq \underset{\boldsymbol{w}}{\operatorname{argmin}} \sum_{i \in D_0} [y_i - f(\boldsymbol{x}_i; \boldsymbol{w})]^2$$

$$f(\boldsymbol{x}; \boldsymbol{w}) \coloneqq \langle \boldsymbol{w}_1, \boldsymbol{x} \rangle + w_0, \boldsymbol{w} \coloneqq [\boldsymbol{w}_1, w_0]^\top$$

- LS is data-driven and uses squared function as its cost.
- **How does LS take randomness of dataset into account?**
- To answer this, we see LS from a probabilistic perspective.



Normal Distribution

- Random events of a Normal dist. happen on real domain.
- Normal dist. has a probability density function (PDF):
- $p(x|\mu,\sigma) := \frac{1}{Z(\sigma)}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], Z(\sigma) = \sigma\sqrt{2\pi}, x \in R.$
- We use $N_x(\mu,\sigma^2)$ denote a Normal PDF. w.r.t. x.

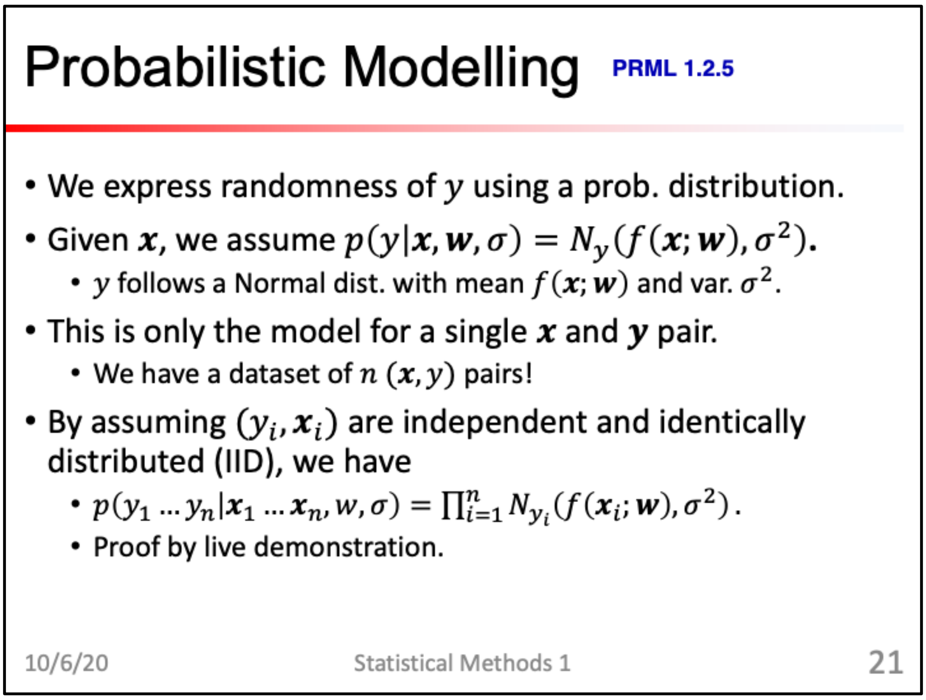
10/6/20 Statistical Methods 1 20

Here μ is mean and σ is the standard deviation of the normal distribution.

You can check the definition of 1-dimension Normal distribution here.
https://en.wikipedia.org/wiki/Normal_distribution

If you are not very familiar with the definition, do not worry. We will review some basics of probability theory in week 3. For now, just think the normal probability density function as a bell-shaped function.

The function value evaluated at a point $\boldsymbol{x}$ tells the relative likelihood of the occurrence of $\boldsymbol{x}$.



Probabilistic Modelling PRML 1.2.5

- We express randomness of y using a prob. distribution.
- Given $\boldsymbol{x}$, we assume $p(y|\boldsymbol{x}, \boldsymbol{w}, \sigma) = N_y(f(\boldsymbol{x}; \boldsymbol{w}), \sigma^2)$.
  - y follows a Normal dist. with mean $f(\boldsymbol{x}; \boldsymbol{w})$ and var. σ^2.
- This is only the model for a single $\boldsymbol{x}$ and $\boldsymbol{y}$ pair.
  - We have a dataset of n $(\boldsymbol{x}, y)$ pairs!
- By assuming $(y_i, \boldsymbol{x}_i)$ are independent and identically distributed (IID), we have
  - $p(y_1 \dots y_n | \boldsymbol{x}_1 \dots \boldsymbol{x}_n, w, \sigma) = \prod_{i=1}^{n} N_{y_i}(f(\boldsymbol{x}_i; \boldsymbol{w}), \sigma^2)$.
  - Proof by live demonstration.

10/6/20 Statistical Methods 1 21

Here, we make a distributional assumption on our data generating process.

We assume that given a specific $\boldsymbol{x}$**,** the distribution of $\boldsymbol{y}$ is a normal distribution with mean $f(\boldsymbol{x}; \boldsymbol{w})$ and variance σ^2.

The IID-ness is a crucial **assumption** in many statistical analysis. However, it may fail in many cases. For example when your dataset comes from a time-series, all your datapoints, y_i and y_{i+1} are always correlated.

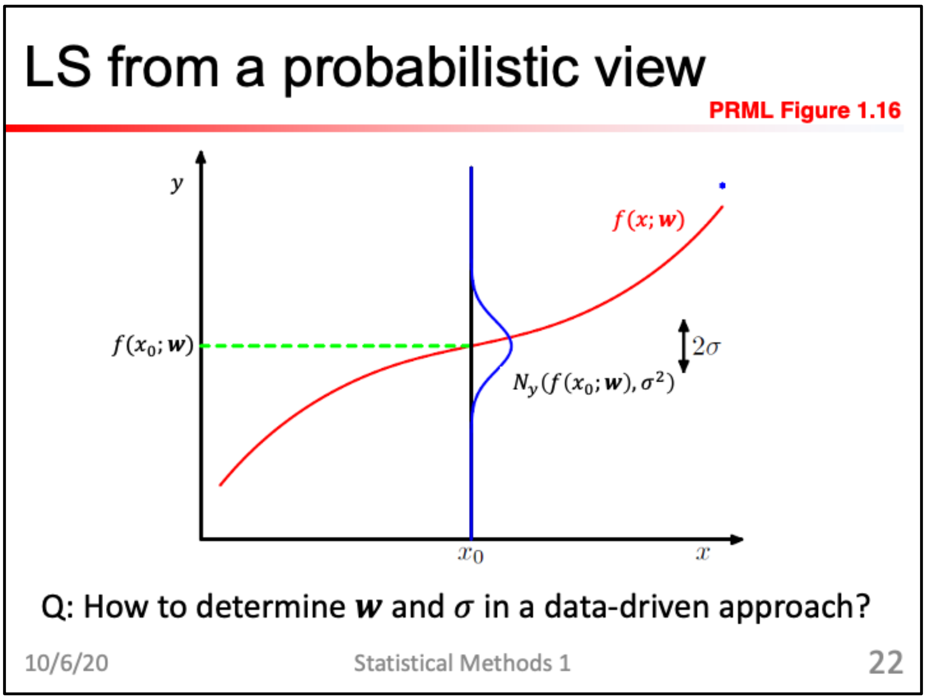
Hint, to prove above equality under IID assumption, we need the following formulas:

A ind B implies
P(A,B) = P(A)P(B)

What's more,

A,B ind C,D implies
A ind D, B ind D, A ind C, B ind C.

In prediction problems, it is common that we put model assumptions only on $p(y|\boldsymbol{x})$ as we only care about the distribution on the output given a specific $\boldsymbol{x}$ and modelling the distribution of $\boldsymbol{x}$ is also too hard ($\boldsymbol{x}$ is in d dimensional space, while y is in one dimensional space).



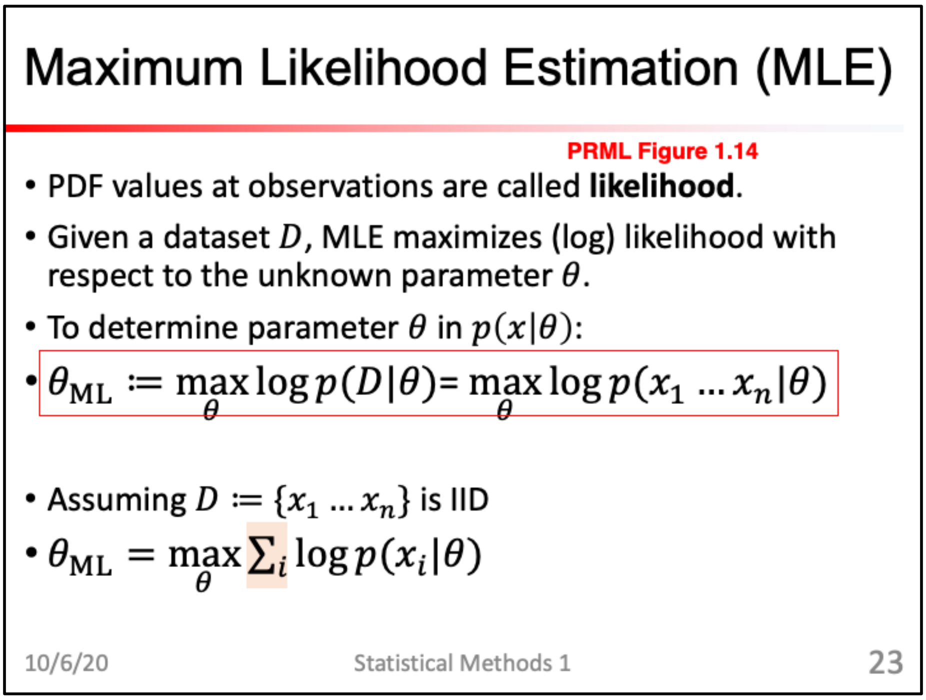
LS from a probabilistic view

PRML Figure 1.16

Q: How to determine $\boldsymbol{w}$ and σ in a data-driven approach?

This probabilistic model does not allow us to predict anything by itself.

How do you tune the parameter of your model in a data driven approach?



Maximum Likelihood Estimation (MLE)

PRML Figure 1.14

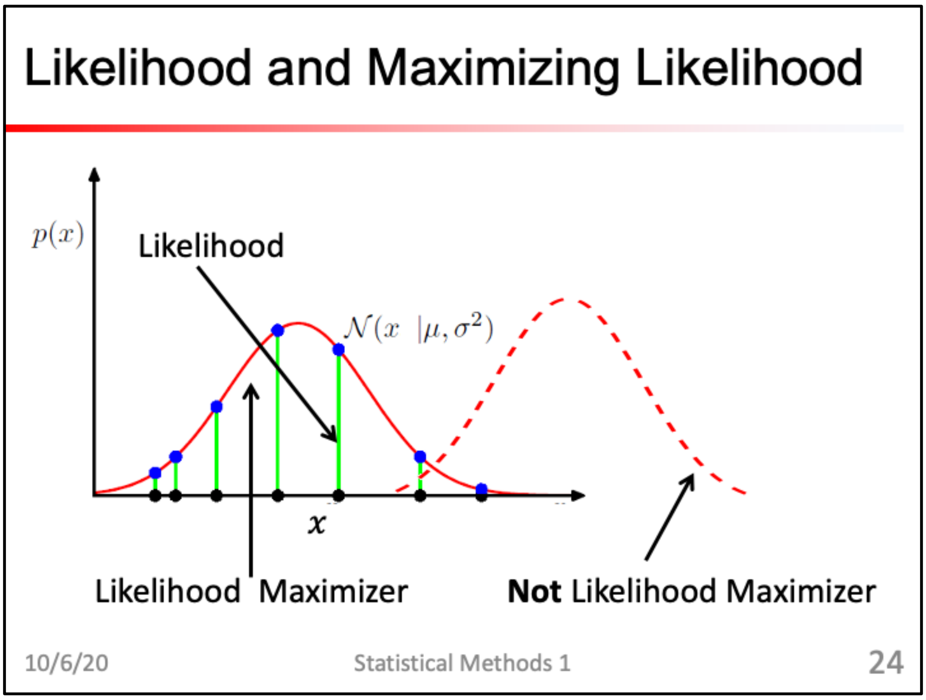
- PDF values at observations are called **likelihood**.
- Given a dataset D, MLE maximizes (log) likelihood with respect to the unknown parameter θ.
- To determine parameter θ in $p(x|\theta)$:
- $\boxed{\theta_{\mathrm{ML}} \coloneqq \max_{\theta} \log p(D|\theta) = \max_{\theta} \log p(x_1 \ldots x_n|\theta)}$
- Assuming $D \coloneqq \{x_1 \ldots x_n\}$ is IID
- $\theta_{\mathrm{ML}} = \max_{\theta} \sum_i \log p(x_i|\theta)$

10/6/20 Statistical Methods 1 23

Maximum Likelihood Estimation is one of classic choices for tuning probabilistic model parameters using a dataset.

There are also other ways of doing parameter tuning using a dataset, as we will see later.

In statistics, the process of tuning your probabilistic model parameter is usually referred as **model inference**.



Likelihood and Maximizing Likelihood

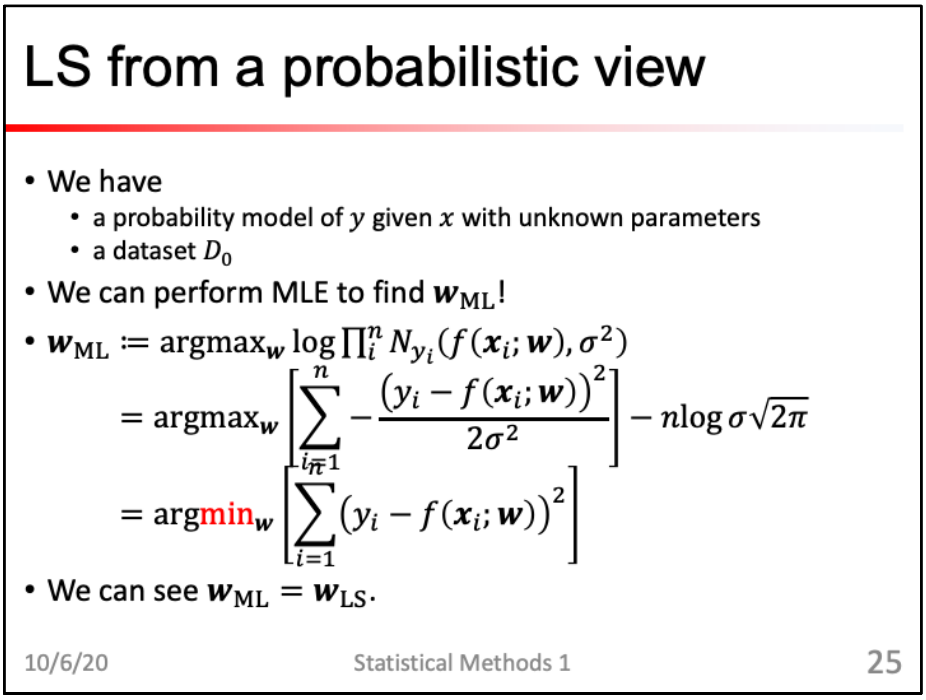
Likelihood

Likelihood Maximizer

Not Likelihood Maximizer

10/6/20 Statistical Methods 1 24

Maximum Likelihood, intuitively, is looking for a **probability density function**, so that its function value at your data-points, are **maximumly elevated**.



LS from a probabilistic view

- We have
  - a probability model of y given x with unknown parameters
  - a dataset D_0
- We can perform MLE to find $\boldsymbol{w}_{\mathrm{ML}}$!
- $\boldsymbol{w}_{\mathrm{ML}} \coloneqq \mathrm{argmax}_{\boldsymbol{w}} \log \prod_i^n N_{y_i}(f(\boldsymbol{x}_i; \boldsymbol{w}), \sigma^2)$

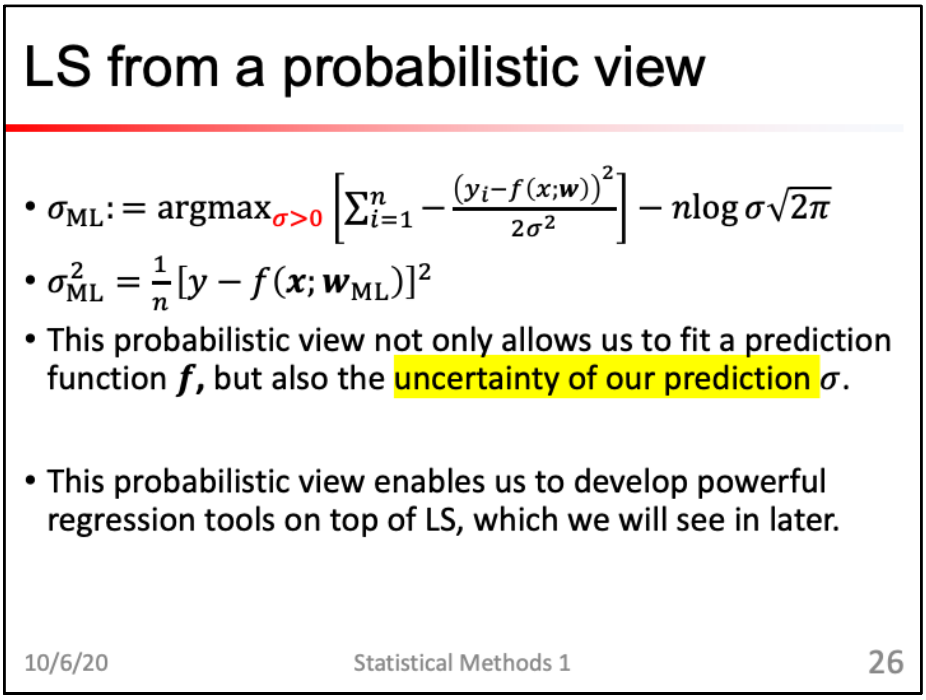
$$= \mathrm{argmax}_{\boldsymbol{w}} \left[\sum_{i=1}^{n} - \frac{\left(y_i - f(\boldsymbol{x}_i; \boldsymbol{w})\right)^2}{2\sigma^2} \right] - n\log \sigma\sqrt{2\pi}$$

$$= \mathrm{arg}\mathrm{min}_{\boldsymbol{w}} \left[\sum_{i=1}^{n} \left(y_i - f(\boldsymbol{x}_i; \boldsymbol{w})\right)^2 \right]$$

- We can see $\boldsymbol{w}_{\mathrm{ML}} = \boldsymbol{w}_{\mathrm{LS}}$.

10/6/20 Statistical Methods 1 25

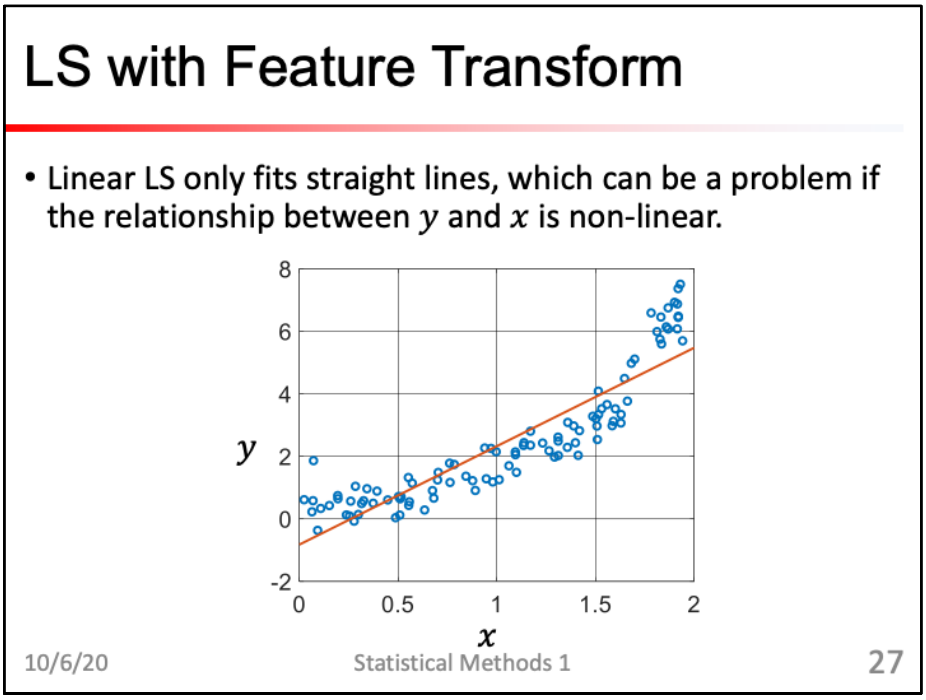
This is the first grand conclusion on this lecture, we introduced an **iid + normal distribution assumption** on our dataset, and we can derive LS using MLE on our dataset.



LS from a probabilistic view

- $\sigma_{\mathrm{ML}} := \mathrm{argmax}_{\sigma>0} \left[\sum_{i=1}^{n} -\frac{\left(y_i - f(x;\boldsymbol{w})\right)^2}{2\sigma^2} \right] - n\log \sigma\sqrt{2\pi}$
- $\sigma_{\mathrm{ML}}^2 = \frac{1}{n}[y - f(\boldsymbol{x};\boldsymbol{w}_{\mathrm{ML}})]^2$
- This probabilistic view not only allows us to fit a prediction function $\boldsymbol{f}$, but also the uncertainty of our prediction σ.
- This probabilistic view enables us to develop powerful regression tools on top of LS, which we will see in later.

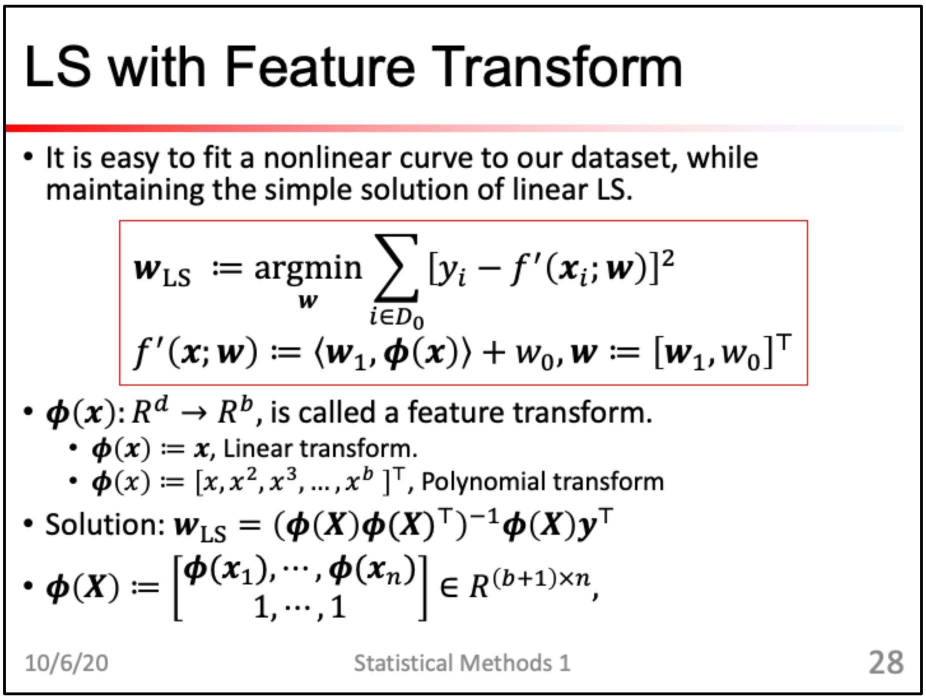
This is the second grand conclusion on this lecture. By using a distributional assumption on our dataset and probabilistic modelling, not only we can derive a prediction function, but also the uncertainty of our prediction, as quantified by σ_{ML} here.



LS with Feature Transform

- Linear LS only fits straight lines, which can be a problem if the relationship between y and x is non-linear.

Linear model does not work well on nonlinear dataset.



LS with Feature Transform

- It is easy to fit a nonlinear curve to our dataset, while maintaining the simple solution of linear LS.

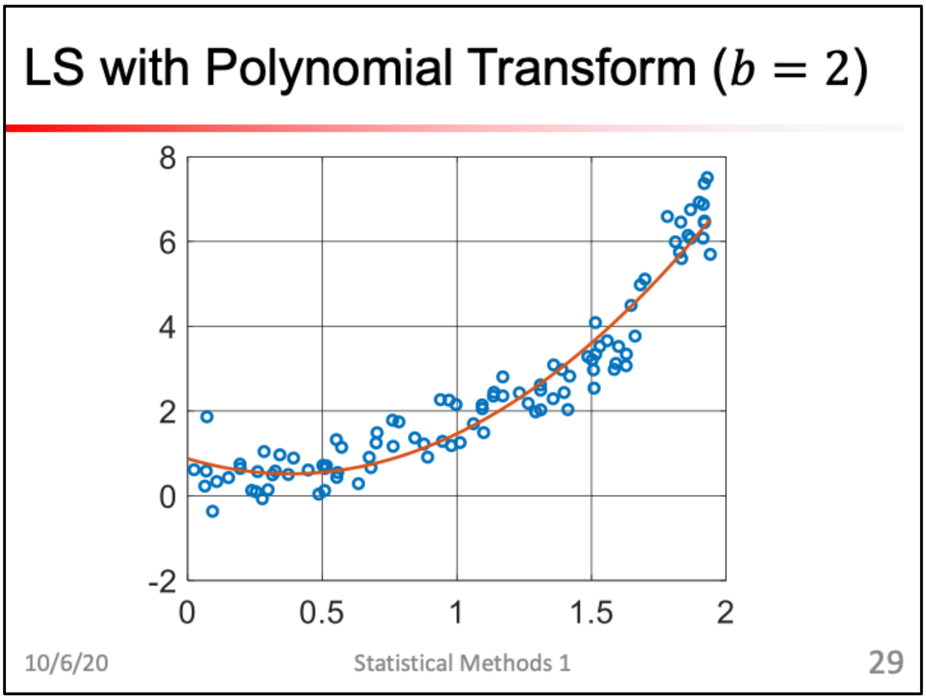
$$\boxed{\begin{aligned} &\boldsymbol{w}_{\mathrm{LS}} \coloneqq \operatorname*{argmin}_{\boldsymbol{w}} \sum_{i \in D_0} [y_i - f'(\boldsymbol{x}_i; \boldsymbol{w})]^2 \\ &f'(\boldsymbol{x}; \boldsymbol{w}) \coloneqq \langle \boldsymbol{w}_1, \boldsymbol{\phi}(\boldsymbol{x}) \rangle + w_0, \boldsymbol{w} \coloneqq [\boldsymbol{w}_1, w_0]^\top \end{aligned}}$$

- $\boldsymbol{\phi}(\boldsymbol{x}): R^d \to R^b$, is called a feature transform.
  - $\boldsymbol{\phi}(\boldsymbol{x}) \coloneqq \boldsymbol{x}$, Linear transform.
  - $\boldsymbol{\phi}(x) \coloneqq [x, x^2, x^3, \dots, x^b\ ]^\top$, Polynomial transform
- Solution: $\boldsymbol{w}_{\mathrm{LS}} = (\boldsymbol{\phi}(\boldsymbol{X})\boldsymbol{\phi}(\boldsymbol{X})^\top)^{-1}\boldsymbol{\phi}(\boldsymbol{X})\boldsymbol{y}^\top$
- $\boldsymbol{\phi}(\boldsymbol{X}) \coloneqq \begin{bmatrix} \boldsymbol{\phi}(\boldsymbol{x}_1), \cdots, \boldsymbol{\phi}(\boldsymbol{x}_n) \\ 1, \cdots, 1 \end{bmatrix} \in R^{(b+1)\times n}$,

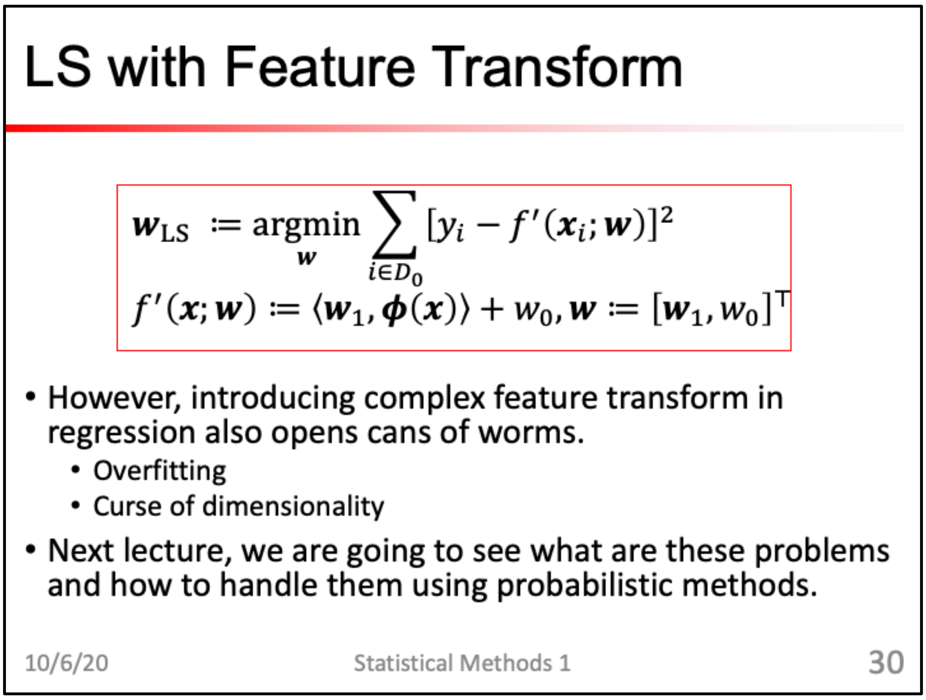
10/6/20 Statistical Methods 1 28

It is easy to adapt a linear model using feature transform functions.

Now $f(\boldsymbol{x}; w)$ is non-longer linear with respect to $\boldsymbol{x}$, but it is still linear with respect to $\boldsymbol{w}$! It allows us to reuse the (almost) same solution used in linear least square for $\boldsymbol{w}_{\mathrm{LS}}$.



LS with Polynomial Transform ($b = 2$)

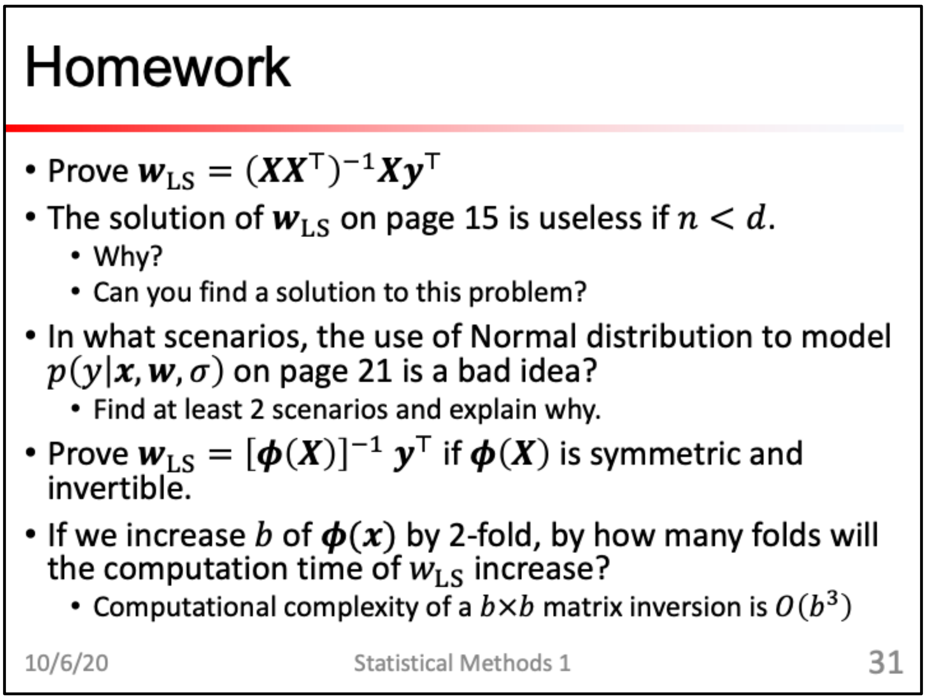


LS with Feature Transform

$$\boldsymbol{w}_{\mathrm{LS}} := \operatorname*{argmin}_{\boldsymbol{w}} \sum_{i\in D_0} [y_i - f'(\boldsymbol{x}_i;\boldsymbol{w})]^2$$

$$f'(\boldsymbol{x};\boldsymbol{w}) := \langle \boldsymbol{w}_1, \boldsymbol{\phi}(\boldsymbol{x})\rangle + w_0, \boldsymbol{w} := [\boldsymbol{w}_1, w_0]^\top$$

- However, introducing complex feature transform in regression also opens cans of worms.
  - Overfitting
  - Curse of dimensionality
- Next lecture, we are going to see what are these problems and how to handle them using probabilistic methods.



Homework

- Prove $\boldsymbol{w}_{\mathrm{LS}} = (\boldsymbol{X}\boldsymbol{X}^{\top})^{-1}\boldsymbol{X}\boldsymbol{y}^{\top}$
- The solution of $\boldsymbol{w}_{\mathrm{LS}}$ on page 15 is useless if $n < d$.
  - Why?
  - Can you find a solution to this problem?
- In what scenarios, the use of Normal distribution to model $p(y|\boldsymbol{x}, \boldsymbol{w}, \sigma)$ on page 21 is a bad idea?
  - Find at least 2 scenarios and explain why.
- Prove $\boldsymbol{w}_{\mathrm{LS}} = [\boldsymbol{\phi}(\boldsymbol{X})]^{-1}\ \boldsymbol{y}^{\top}$ if $\boldsymbol{\phi}(\boldsymbol{X})$ is symmetric and invertible.
- If we increase b of $\boldsymbol{\phi}(\boldsymbol{x})$ by 2-fold, by how many folds will the computation time of w_{LS} increase?
  - Computational complexity of a $b \times b$ matrix inversion is $O(b^3)$

10/6/20 Statistical Methods 1 31

Please include answers to homework in your portfolio with brief explanation.