Supervised and Unsupervised Learning

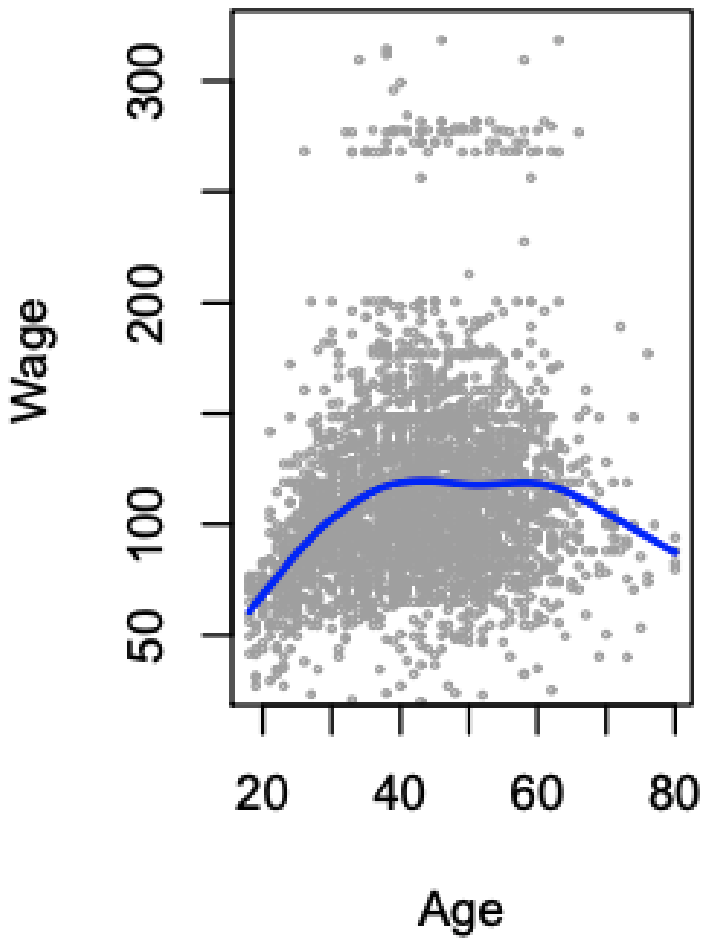
Labor economics

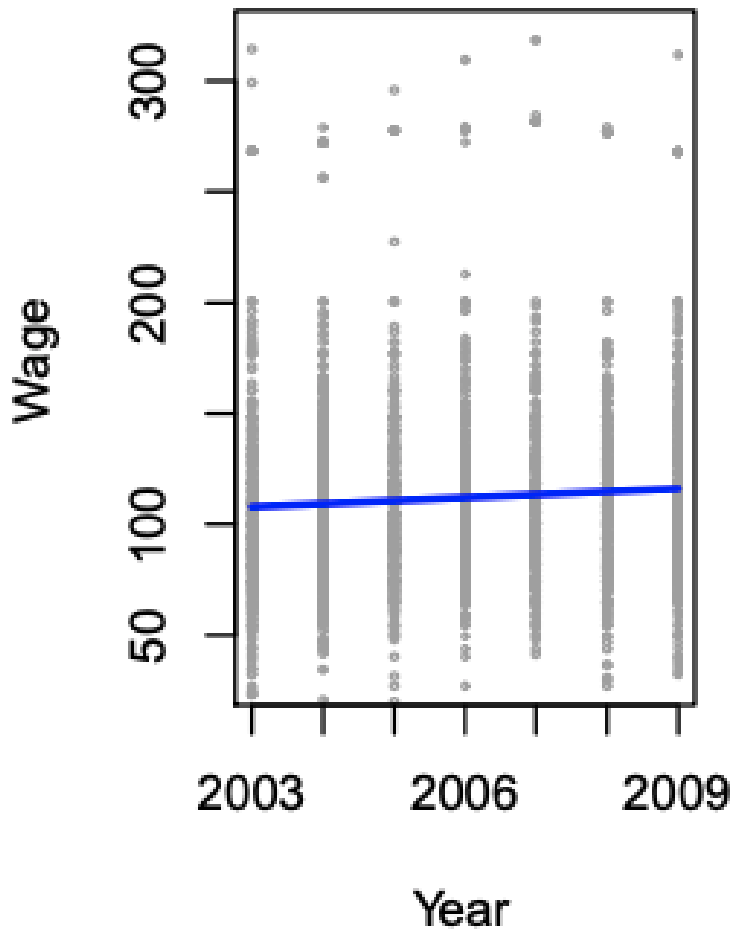
Instructor: Haoran LEI

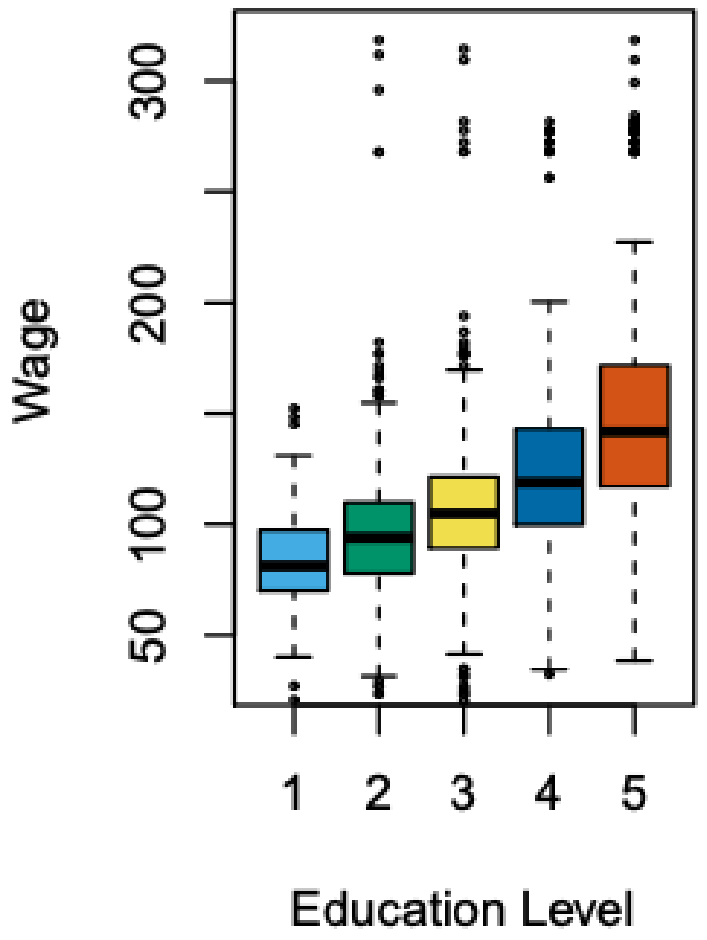
Hunan University

A statistical learning problem:

- income survey data for males from the central Atlantic region of the USA in 2009.
- Goal: *Establish the relationship between* ***salary*** *and* ***demographic variables*** *in population survey data*
- wage = f(Age, Year, Education Level)


Wage
300
200
100
50
20
40
60
80
Age


Wage
300
200
100
50
2003
2006
2009
Year


Wage
300
200
100
50
1
2
3
4
5
Education Level

The Supervised Learning Problem

1. **Outcome** measurement Y

- also called *dependent variable, response, target, ...*
- In the wage example, Y = wage.

2. Vector of p **predictor** measurements X

- also alled *inputs, regressors, covariates, features, independent variables, ...*
- In the wage example, p = 3 and $X =$ (Age, Year, Education).

The Supervised Learning Problem

3. We have N **observations**

- also called *examples*, *instances* of these measurements

Our data set in the wage example:

- $(X_1, Y_1), (X_2, Y_2), ...(X_N, Y_N)$
- X_i is the i-th observation.
- Each X_i is a (Age, Year, Education) tuple, usually written as (X_i^1, X_i^2, X_i^3).

Regreassion vs classification

In the **regreassion problem**, Y is quantitative.

- In the wage example, Y = wage and is quantitative.

In the **classification problem**, Y takes values in a finite, unordered set (eg, died/survived, **employed/unemployed**, digit, ...)

Objectives

On the basis of the **training data** we would like to:

- Accurately predict unseen test cases.
- Understand which inputs affect the outcome, and how.
- Assess the quality of our predictions and inferences

General rules

- It is important to understand the ideas behind the various techniques, in order to know how and when to use them.
- One has to understand the simpler methods first, in order to grasp the more sophisticated ones.
- It is important to accurately assess the performance of a method, to know how well or how badly it is working

Statistical methods are widely used beyond economics: science, industry and finance.

Unsupervised learning

- Both regressions and classications are **supervised learning**
- All datas are "labeled":
  - you know the wage of each individual in the sample; or know whether she is employed or not.
- In **unsupervised learning**, there is no outcome:
  - just a set of inputs/predictors on a set of samples

Unsupervised learning: clustering

Examples of unsupervised learning:

- Amazon classifies customers into different groups based on there purchasing history
- Banks grade their customers based on their reputation history (eg, whether the customer defaults on the loan or not, how much is the loan, ...)

The examples above are called **clustering**.

Objective of unsupervised learning

- The objective is to find groups of samples that behave similarly, find features that behave similarly, find linear combinations of features with the most variation.
- Unsupervised learning is different from supervised learning, but can be useful as a pre-processing step for supervised learning.

Case study: the Netflix prize

- competition started in October 2006. Training data is ratings for 18, 000 movies by 400, 000 Netflix customers, each rating between 1 and 5.
- training data is very sparse---about 98% missing.
- objective is to predict the rating for a set of 1 million customer-movie pairs that are missing in the training data.
- Netflix's original algorithm achieved a root MSE of 0.953. The first team to achieve a 10% improvement wins one million dollars.

Question: is this Supervised or Unsupervised learning?

Question: is this Supervised or Unsupervised learning?

- This is an **unsupervised learning** problem.
- Netflix has the whole data, participants are expected to recover those erased data using statistical learning methods.
- Data shape is as follows (`x` means missing data):

```
        Customer 1, Customer 2, Customer 3, Customer 4,...
Film 1      4           3           x           4       ...
Film 2      3           x           3           2       ...
Film 3      2           2           x           3       ...
...         ...         ...
```

Leaderboard

Showing Test Score. Click here to show quiz score

Display top 20 leaders.

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries !	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09
8	Dace	0.8612	9.59	2009-07-24 17:18:43
9	Feeds2	0.8622	9.48	2009-07-12 13:11:51
10	BigChaos	0.8623	9.47	2009-04-07 12:33:59
11	Opera Solutions	0.8623	9.47	2009-07-24 00:34:07
12	BellKor	0.8624	9.46	2009-07-26 17:19:11

BellKor's Pragmatic Chaos wins 1 million (Wiki).