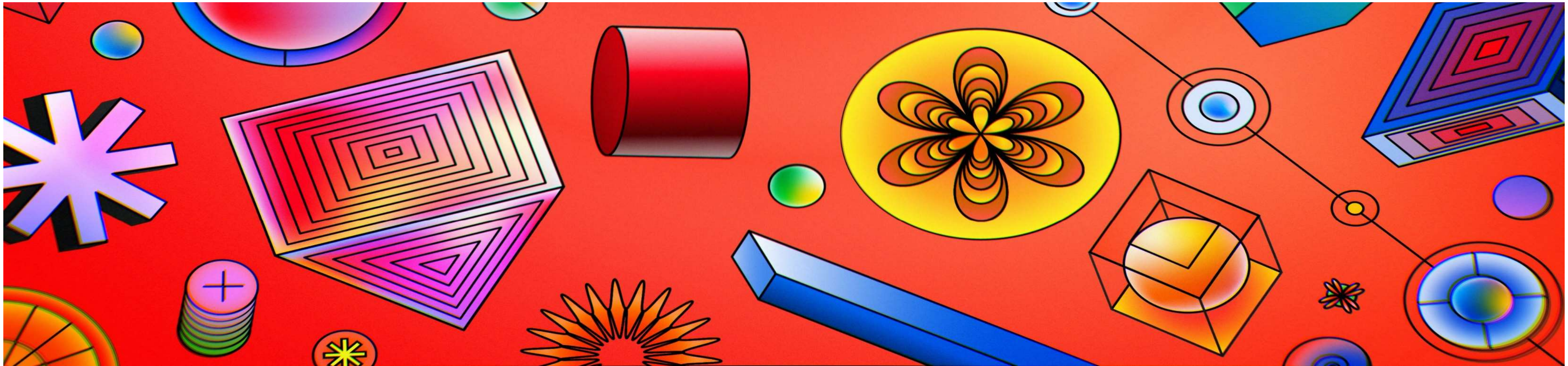
Fall 2023

BIF524/CSC463 Data Mining

Statistical Learning

Eileen Marie Hanna, *PhD*

05/09/2023



Attribute

- A data field representing a **characteristic** of a data object.
- Values of attributes are also called **observations**.
- A set of attribute describing an object is called **attribute vector or feature vector**.
- The type of an attribute is determined by its possible values.

patient ID	gender	. . .	occupation	Insurance coverage	phone number	temperature	pain level	weight	smoker	heart disease history
357	F	. . .	dentist	3	123456	38.2	moderate	62	0	1
358	M	. . .	architect	2	234567	40.3	severe	78	0	0
359	F	. . .	designer	1	345678	37.5	minimal	56	1	1
340	M	. . .	manager	3	456789	41	unbearable	88	1	0
341	F	. . .	nurse	1	567890	38.8	moderate	64	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Qualitative attributes

- **Nominal** attributes – also referred to as categorical
  - can have **symbols** or **name** of things as values.
  - can also be represented as **numbers coding** for possible names/categories.
  - It makes no sense to compute the mean or median for such attributes – the mode can however be calculated.

patient ID	gender	. . .	occupation	Insurance coverage	phone number	temperature	pain level	weight	smoker	heart disease history
357	F	. . .	dentist	3	123456	38.2	moderate	62	0	1
358	M	. . .	architect	2	234567	40.3	severe	78	0	0
359	F	. . .	designer	1	345678	37.5	minimal	56	1	1
340	M	. . .	manager	3	456789	41	unbearable	88	1	0
341	F	. . .	nurse	1	567890	38.8	moderate	64	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Qualitative attributes

- **Binary** attributes can only **two possible values**.

  - Also called **Boolean** attributes when states are true (1) and false (0) which typically mean that the attribute is present or absent for a certain object, respectively.

patient ID	gender	. . .	occupation	Insurance coverage	phone number	temperature	pain level	weight	smoker	heart disease history
357	F	. . .	dentist	3	123456	38.2	moderate	62	0	1
358	M	. . .	architect	2	234567	40.3	severe	78	0	0
359	F	. . .	designer	1	345678	37.5	minimal	56	1	1
340	M	. . .	manager	3	456789	41	unbearable	88	1	0
341	F	. . .	nurse	1	567890	38.8	moderate	64	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Qualitative attributes

- **Ordinal** attributes
  - have possible values of **meaningful order** or ranking among them.
  - The magnitude between successive values is not specified.
  - The median and mode values make sense here, unlike the mean.

patient ID	gender	. . .	occupation	Insurance coverage	phone number	temperature	pain level	weight	smoker	heart disease history
357	F	. . .	dentist	3	123456	38.2	moderate	62	0	1
358	M	. . .	architect	2	234567	40.3	severe	78	0	0
359	F	. . .	designer	1	345678	37.5	minimal	56	1	1
340	M	. . .	manager	3	456789	41	unbearable	88	1	0
341	F	. . .	nurse	1	567890	38.8	moderate	64	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Quantitative attributes

- Numeric, i.e., **measurable quantity** that can be represented by integer or real values.
- **Interval-scaled** attributes are measured on an equal-size units.
  - Values of interval-scaled attributed **do not have a zero-point** (e.g., 0℃ does not mean that there is no temperature).

patient ID	gender	. . .	occupation	Insurance coverage	phone number	temperature	pain level	weight	smoker	heart disease history
357	F	. . .	dentist	3	123456	38.2	moderate	62	0	1
358	M	. . .	architect	2	234567	40.3	severe	78	0	0
359	F	. . .	designer	1	345678	37.5	minimal	56	1	1
340	M	. . .	manager	3	456789	41	unbearable	88	1	0
341	F	. . .	nurse	1	567890	38.8	moderate	64	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Quantitative attributes

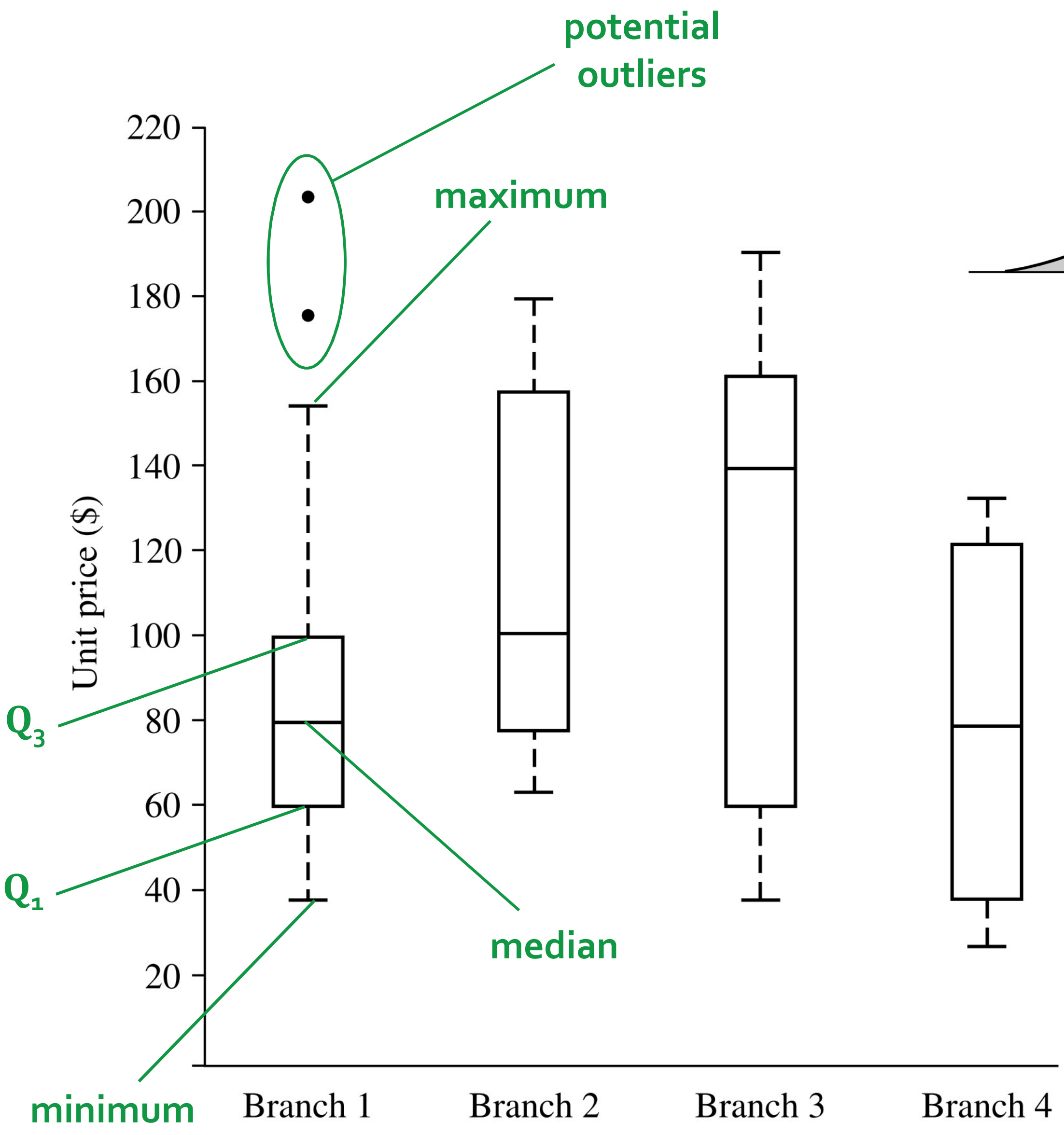
- **Ratio-scaled** attributes have ordered integer values with **inherent zero-point**.

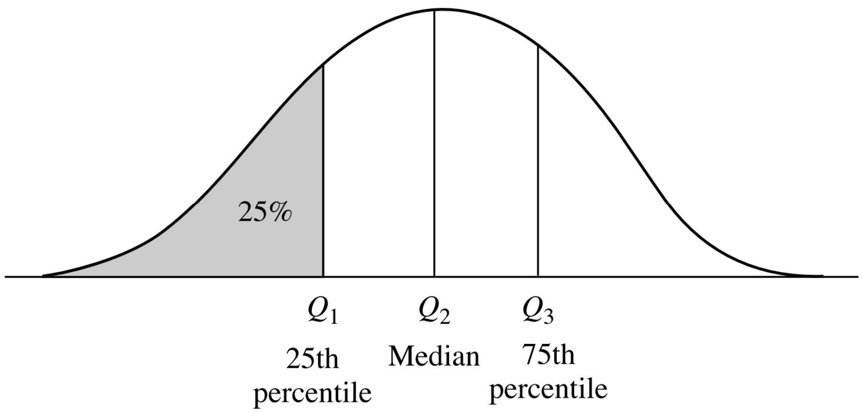
patient ID	gender	. . .	occupation	Insurance coverage	phone number	temperature	pain level	weight	smoker	heart disease history
357	F	. . .	dentist	3	123456	38.2	moderate	62	0	1
358	M	. . .	architect	2	234567	40.3	severe	78	0	0
359	F	. . .	designer	1	345678	37.5	minimal	56	1	1
340	M	. . .	manager	3	456789	41	unbearable	88	1	0
341	F	. . .	nurse	1	567890	38.8	moderate	64	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Discrete vs continuous attributes

- Another classification of attributes could be:
  - **discrete:** has a finite (e.g., hair_color) or countably infinite set of values, that may or may not be represented as integers (e.g., ZIP_code, customerID).
  - **continuous:** i.e., numeric values represented by integers or real numbers.

Boxplots – five-number summary of a distribution



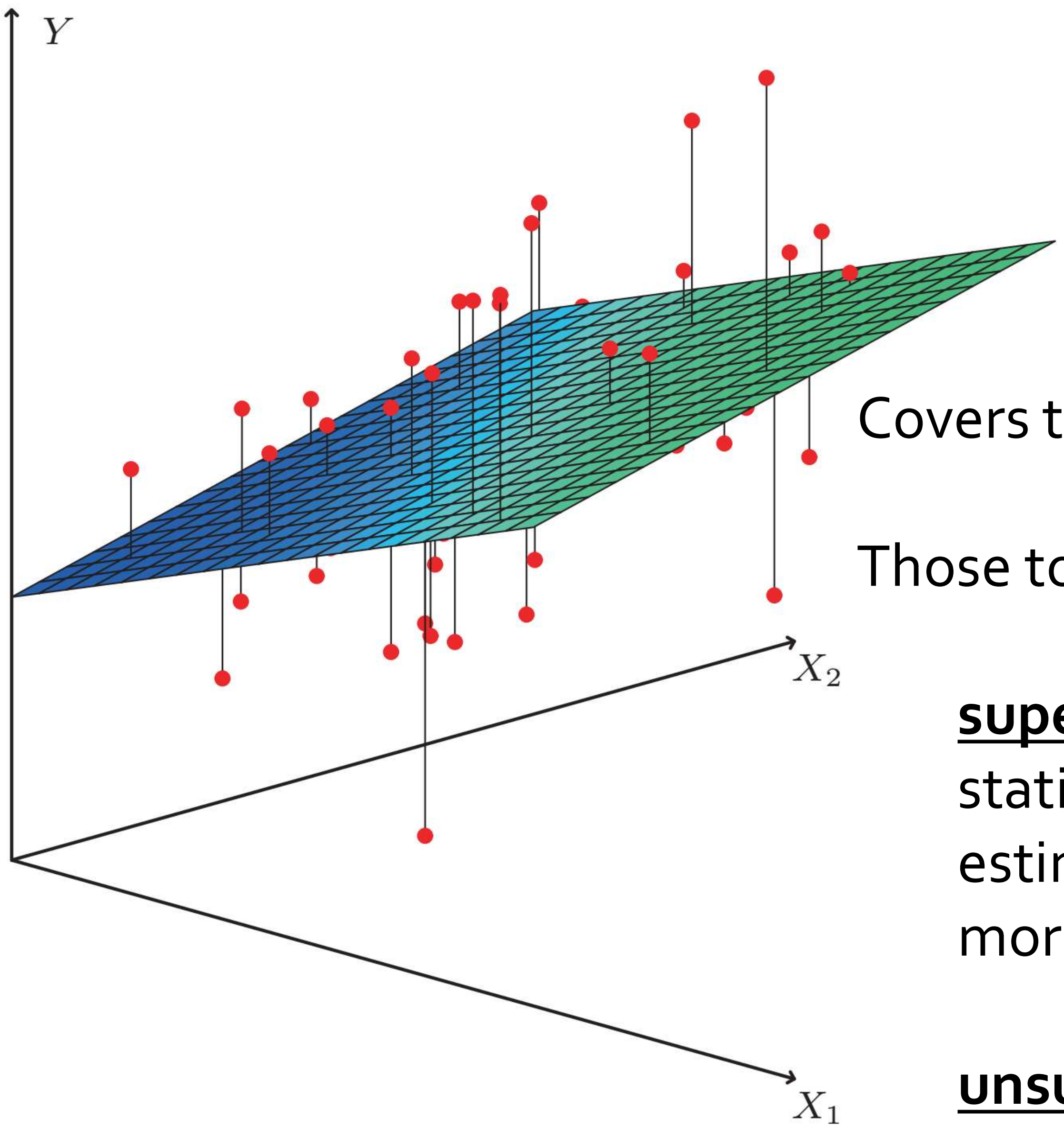


1 2 2 3 3 4 4 5 6 8 9

$$\frac{\Sigma(\,)}{9}$$

$$\frac{3+4}{2}$$

Statistical Learning



Covers tools for understanding data.

Those tools can be categorized as:

supervised: involves building a statistical model to predict or estimate an output, given one or more inputs.

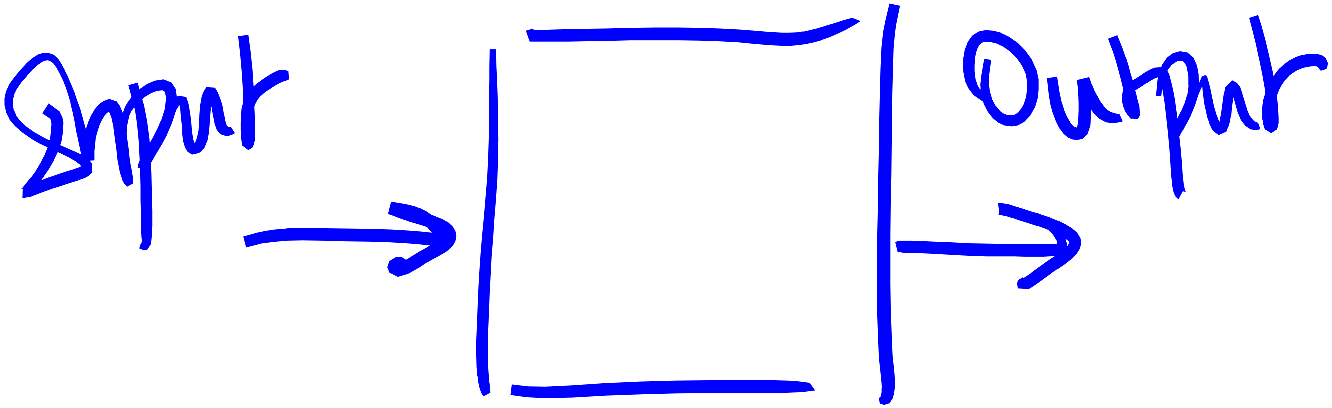
unsupervised: involves learning the structure and relationships in given inputs, with no supervised output.

“Wage” dataset

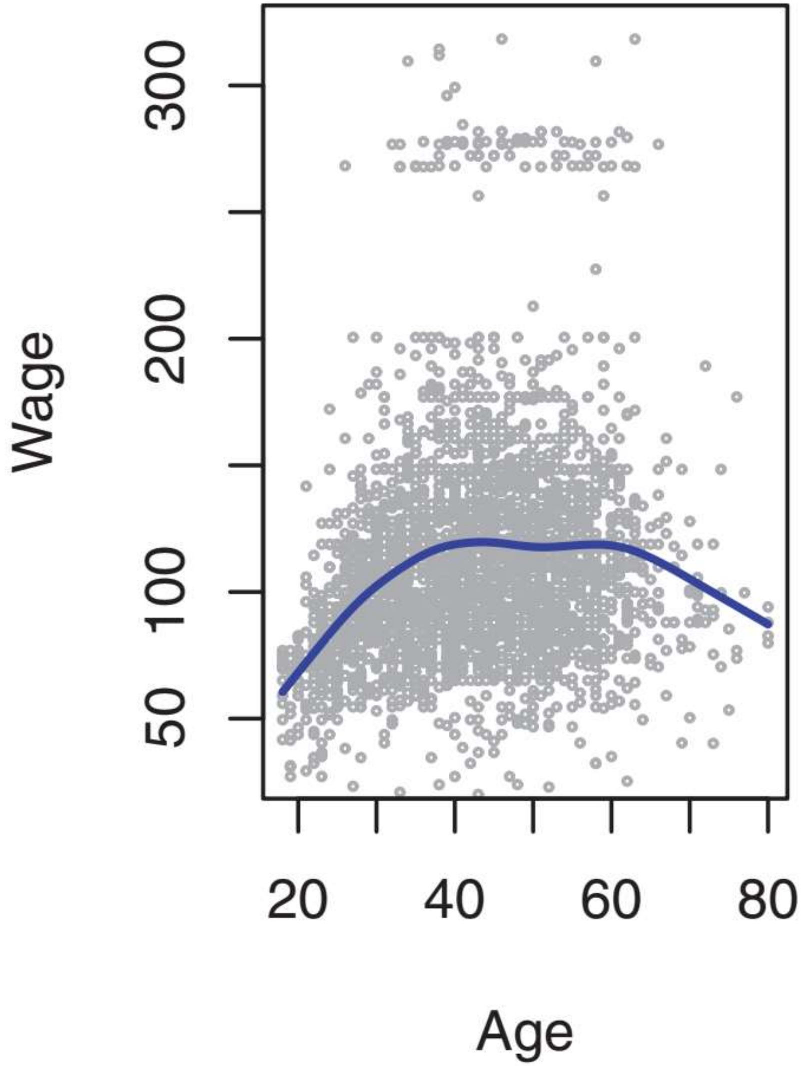
- Includes factors believed to be related to wages of a group of males from the Atlantic region in the US, e.g., age, education level, ..etc.

	year	age	maritl	race	education	region	jobclass	health	ealth_ins	logwage	wage
231655	2006	18	1. Never Married	1. White	1. < HS Grad	2. Middle Atlantic	1. Industrial	1. <=Good	2. No	4.318063335	75.04315402
86582	2004	24	1. Never Married	1. White	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Good	2. No	4.255272505	70.47601965
161300	2003	45	2. Married	1. White	3. Some College	2. Middle Atlantic	1. Industrial	1. <=Good	1. Yes	4.875061263	130.9821774
155159	2003	43	2. Married	3. Asian	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Good	1. Yes	5.041392685	154.685293
11443	2005	50	4. Divorced	1. White	2. HS Grad	2. Middle Atlantic	2. Information	1. <=Good	1. Yes	4.318063335	75.04315402
376662	2008	54	2. Married	1. White	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Good	1. Yes	4.84509804	127.1157438
450601	2009	44	2. Married	4. Other	3. Some College	2. Middle Atlantic	1. Industrial	2. >=Very Good	1. Yes	5.133021279	169.528538
377954	2008	30	1. Never Married	3. Asian	3. Some College	2. Middle Atlantic	2. Information	1. <=Good	1. Yes	4.716003344	111.7208494
228963	2006	41	1. Never Married	2. Black	3. Some College	2. Middle Atlantic	2. Information	2. >=Very Good	1. Yes	4.77815125	118.8843593
81404	2004	52	2. Married	1. White	2. HS Grad	2. Middle Atlantic	2. Information	2. >=Very Good	1. Yes	4.857332496	128.6804882
302778	2007	45	4. Divorced	1. White	3. Some College	2. Middle Atlantic	2. Information	1. <=Good	1. Yes	4.763427994	117.1468169
305706	2007	34	2. Married	1. White	2. HS Grad	2. Middle Atlantic	1. Industrial	2. >=Very Good	2. No	4.397940009	81.28325328
8690	2005	35	1. Never Married	1. White	2. HS Grad	2. Middle Atlantic	2. Information	2. >=Very Good	1. Yes	4.494154594	89.49247952
153561	2003	39	2. Married	1. White	4. College Grad	2. Middle Atlantic	1. Industrial	2. >=Very Good	1. Yes	4.903089987	134.7053751
449654	2009	54	2. Married	1. White	2. HS Grad	2. Middle Atlantic	2. Information	2. >=Very Good	1. Yes	4.903089987	134.7053751
447660	2009	51	2. Married	1. White	3. Some College	2. Middle Atlantic	1. Industrial	2. >=Very Good	1. Yes	4.505149978	90.48191336
160191	2003	37	1. Never Married	3. Asian	4. College Grad	2. Middle Atlantic	1. Industrial	2. >=Very Good	2. No	4.414973348	82.6796373
230312	2006	50	2. Married	1. White	5. Advanced Degree	2. Middle Atlantic	2. Information	2. >=Very Good	2. No	5.360551762	212.8423523
301585	2007	56	2. Married	1. White	4. College Grad	2. Middle Atlantic	1. Industrial	1. <=Good	1. Yes	4.861026342	129.156693
153682	2003	37	1. Never Married	1. White	3. Some College	2. Middle Atlantic	1. Industrial	2. >=Very Good	1. Yes	4.591064607	98.59934386

Model



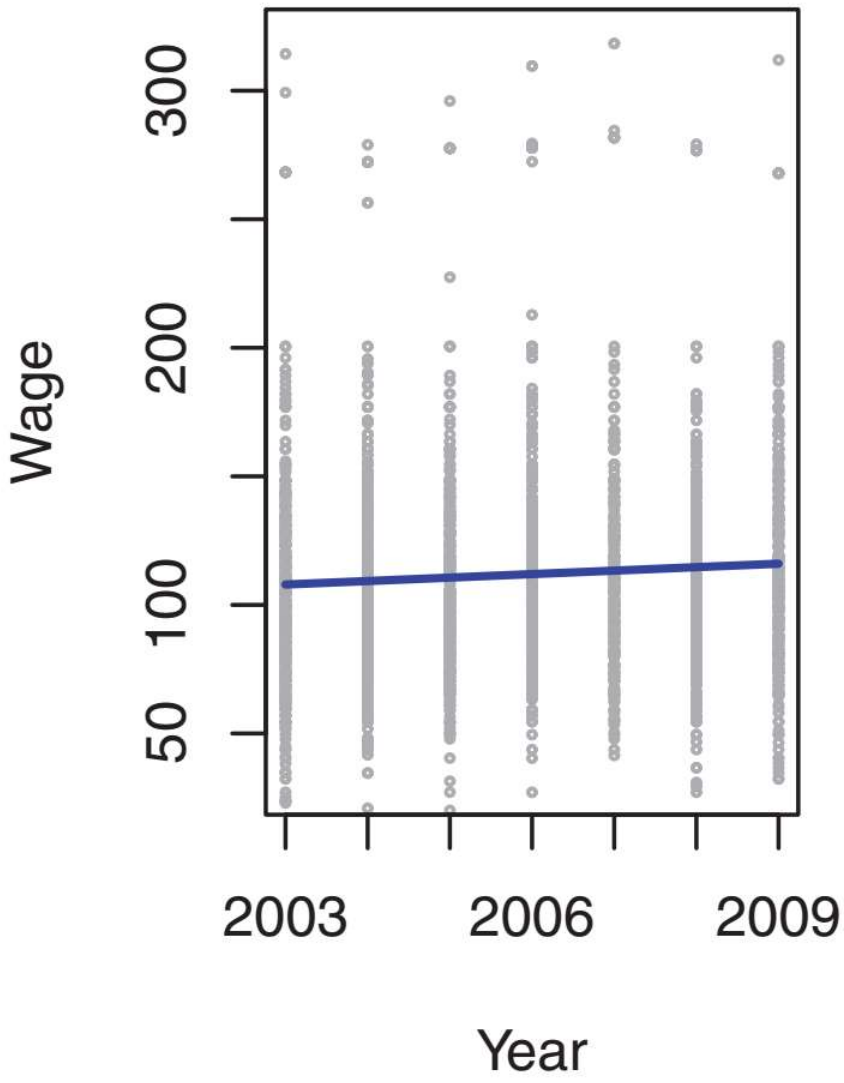
“Wage” dataset



estimate of the average wage at each age

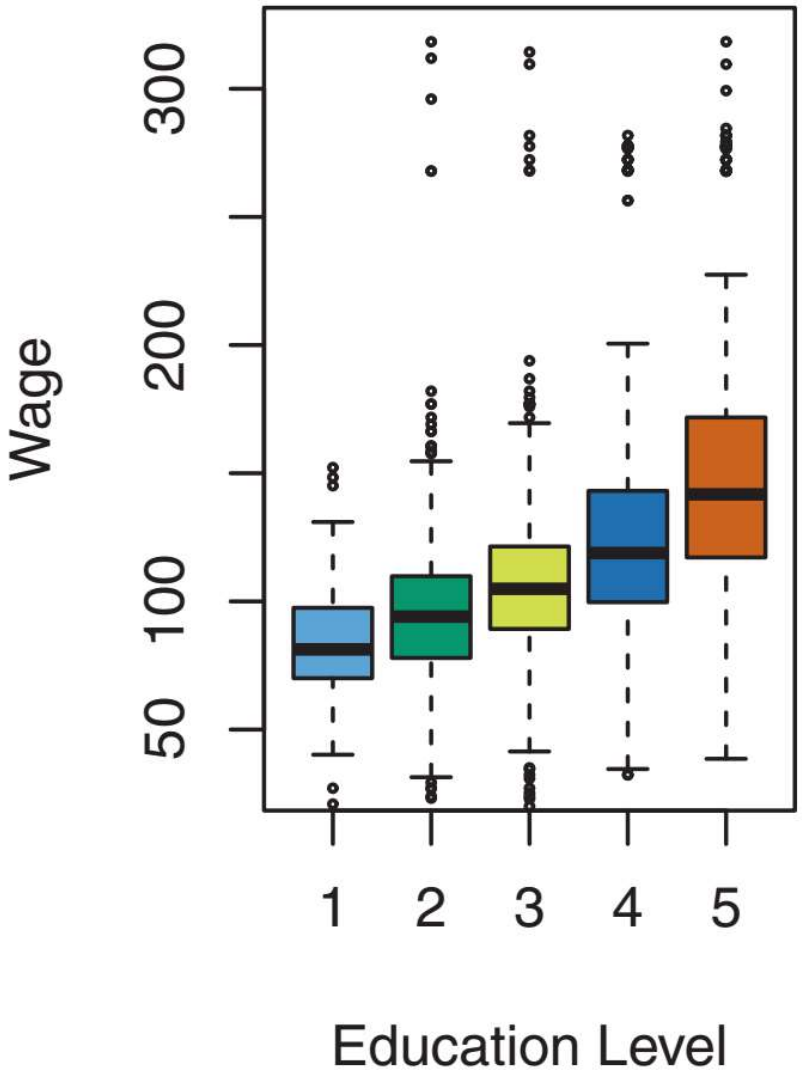
- Wage as a function of age.
- On average, wage increases with age until around 60 years and starts to declines afterwards.

"Wage" dataset



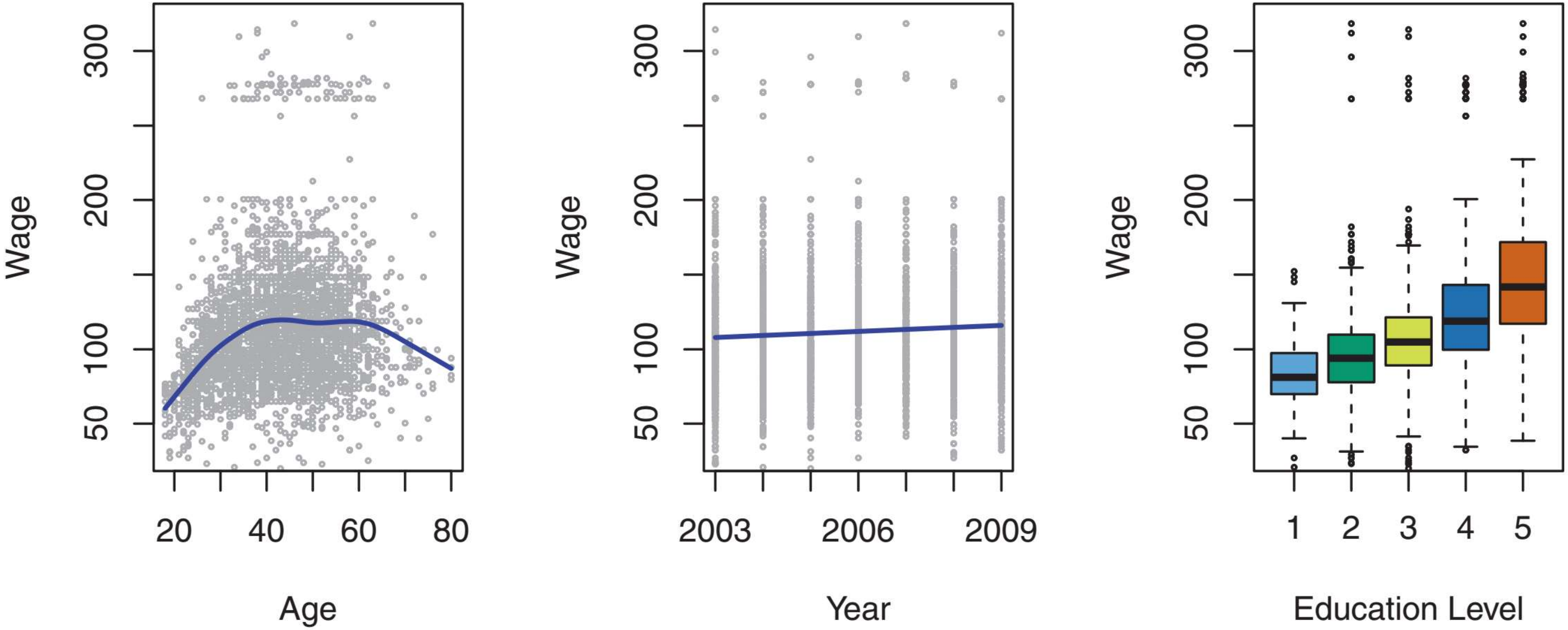
- Wage as a function of year.
- A slow and steady (roughly linear) increase in wages.
- approx. $10,000 between 2003 and 2009.

“Wage” dataset

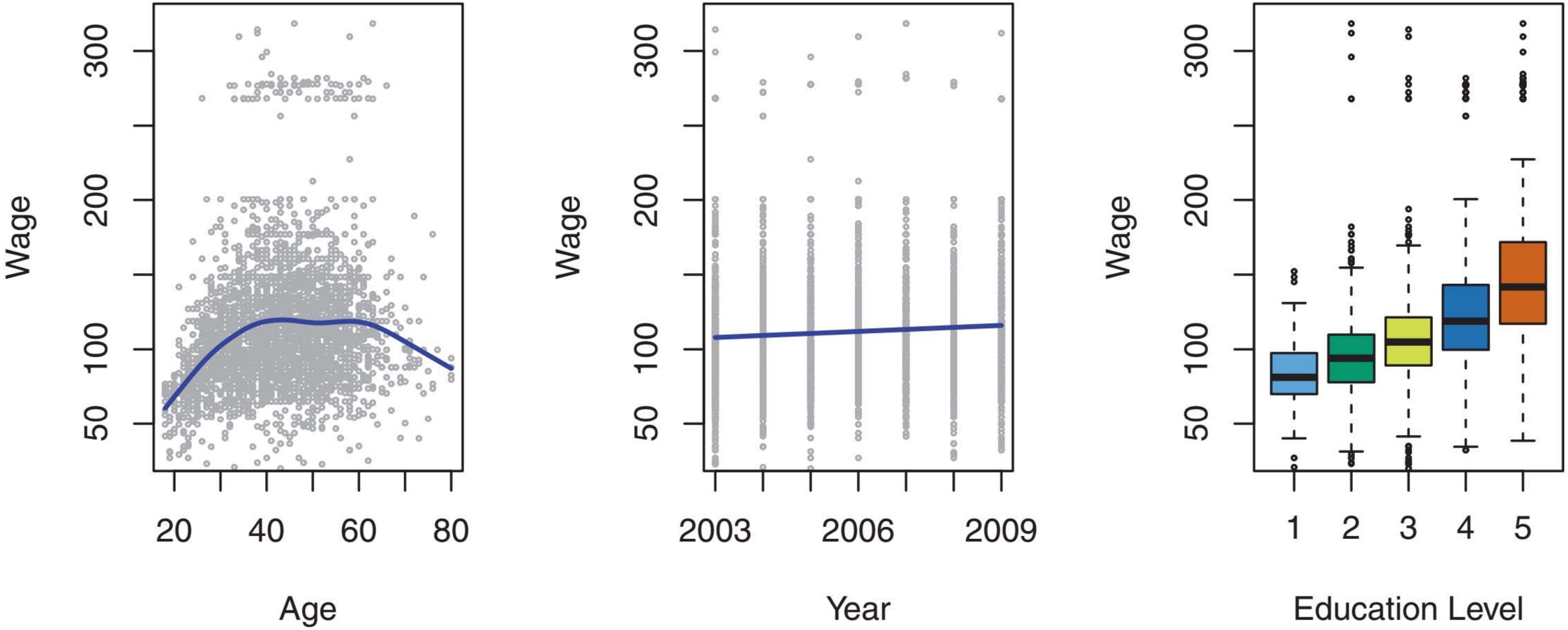


- Wage as a function of education level:
  - 1 being the lowest (no high school diploma)
  - 5 being the highest (advanced graduate degree).
- On average, wage increases with education level.

Which of those factors can be used to predict the wage of an employee?



Which of those factors can be used to predict the wage of an employee?



Quantitative (or continuous) output
-> regression problem

“Smarket” – stock market dataset

	Year	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today	Direction
1	2001	0.381	-0.192	-2.624	-1.055	5.01	1.1913	0.959	Up
2	2001	0.959	0.381	-0.192	-2.624	-1.055	1.2965	1.032	Up
3	2001	1.032	0.959	0.381	-0.192	-2.624	1.4112	-0.623	Down
4	2001	-0.623	1.032	0.959	0.381	-0.192	1.276	0.614	Up
5	2001	0.614	-0.623	1.032	0.959	0.381	1.2057	0.213	Up
6	2001	0.213	0.614	-0.623	1.032	0.959	1.3491	1.392	Up
7	2001	1.392	0.213	0.614	-0.623	1.032	1.445	-0.403	Down
8	2001	-0.403	1.392	0.213	0.614	-0.623	1.4078	0.027	Up
9	2001	0.027	-0.403	1.392	0.213	0.614	1.164	1.303	Up
10	2001	1.303	0.027	-0.403	1.392	0.213	1.2326	0.287	Up
11	2001	0.287	1.303	0.027	-0.403	1.392	1.309	-0.498	Down
12	2001	-0.498	0.287	1.303	0.027	-0.403	1.258	-0.189	Down
13	2001	-0.189	-0.498	0.287	1.303	0.027	1.098	0.68	Up
14	2001	0.68	-0.189	-0.498	0.287	1.303	1.0531	0.701	Up
15	2001	0.701	0.68	-0.189	-0.498	0.287	1.1498	-0.562	Down
16	2001	-0.562	0.701	0.68	-0.189	-0.498	1.2953	0.546	Up
17	2001	0.546	-0.562	0.701	0.68	-0.189	1.1188	-1.747	Down
18	2001	-1.747	0.546	-0.562	0.701	0.68	1.0484	0.359	Up
19	2001	0.359	-1.747	0.546	-0.562	0.701	1.013	-0.151	Down
20	2001	-0.151	0.359	-1.747	0.546	-0.562	1.0596	-0.841	Down
21	2001	-0.841	-0.151	0.359	-1.747	0.546	1.1583	-0.623	Down

"Smarket" – stock market dataset

- Daily movements in S&P stock index over a 5-year period, between 2001 and 2005.

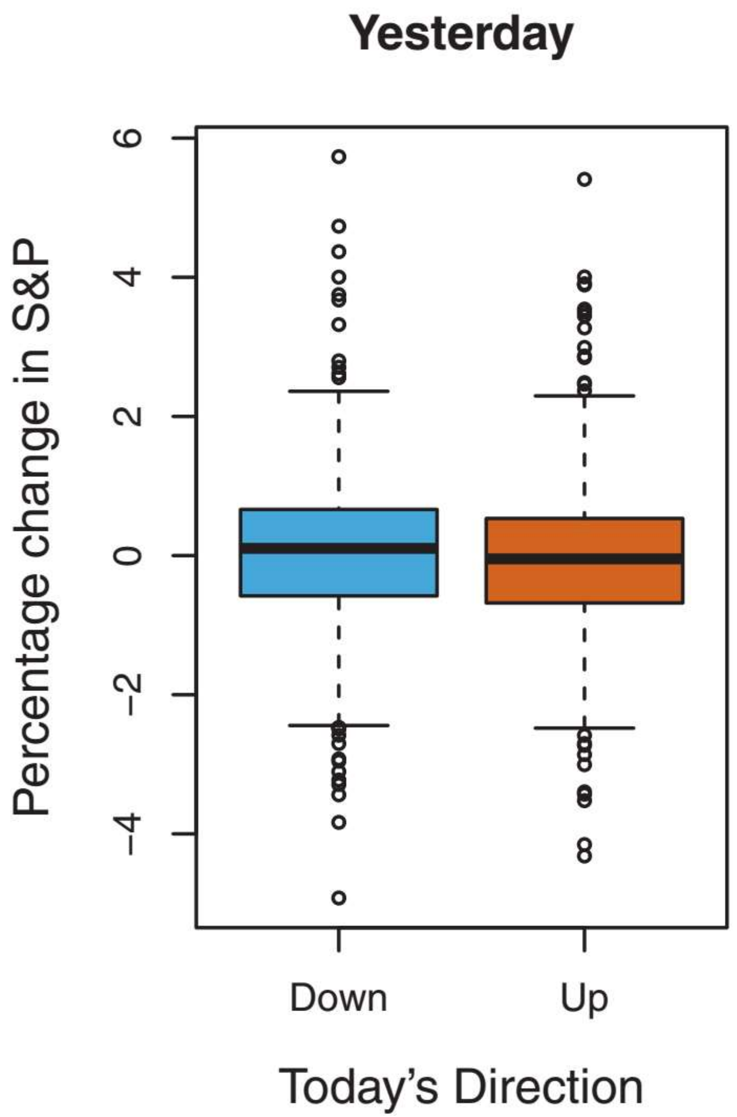
- **Predict whether the index will increase or decrease based on the percentage of change in the past 5 days**.

  - in this case, we are **not predicting a numerical value**.

- We are predicting whether a certain day's stock performance falls into the **Up bucket or the Down bucket**

    - **-> classification problem**.

“Smarket” – stock market dataset



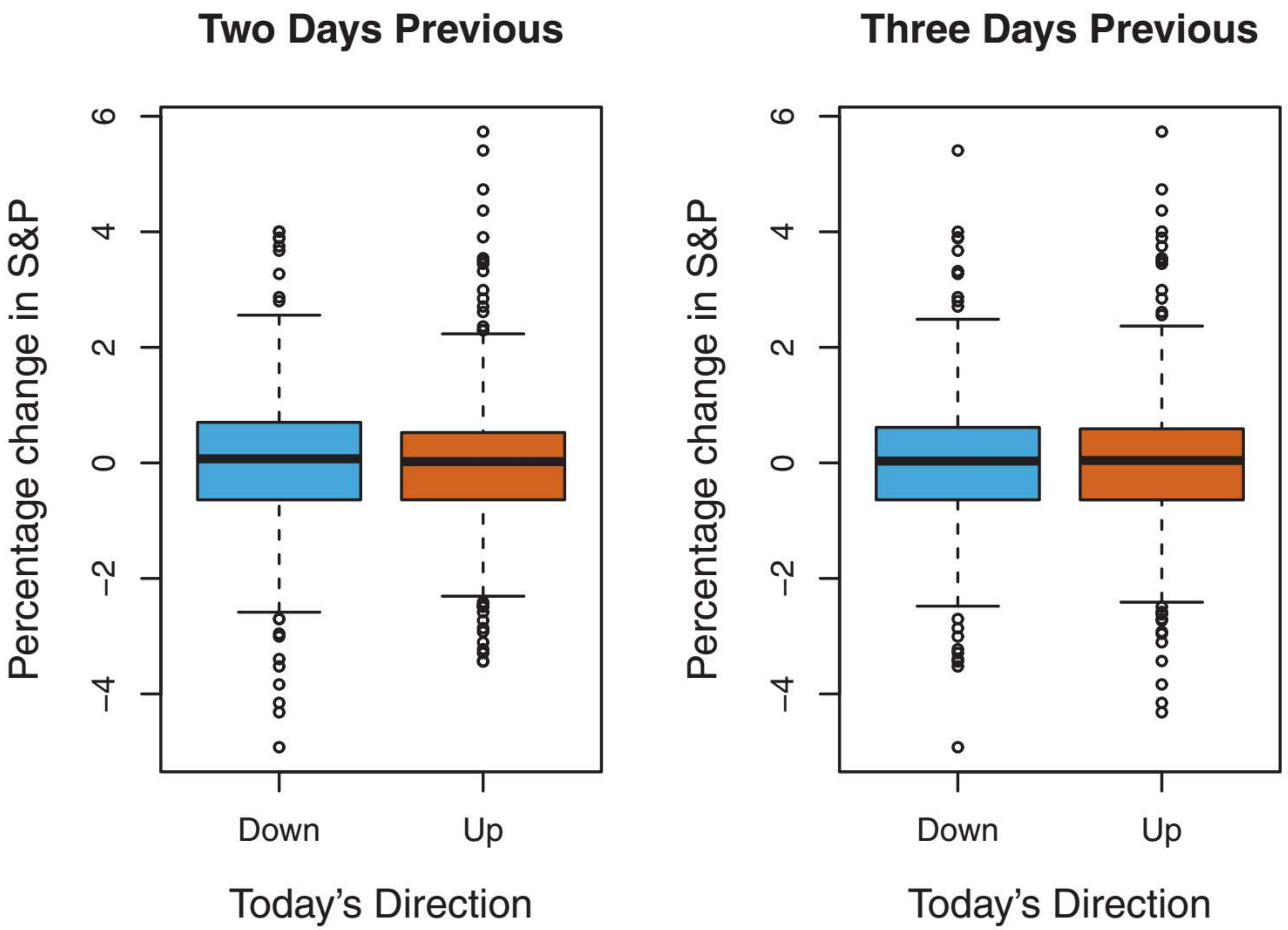
The percentage change in the stock index on the pervious day.

data from 602 days for which the market decreased on the following day

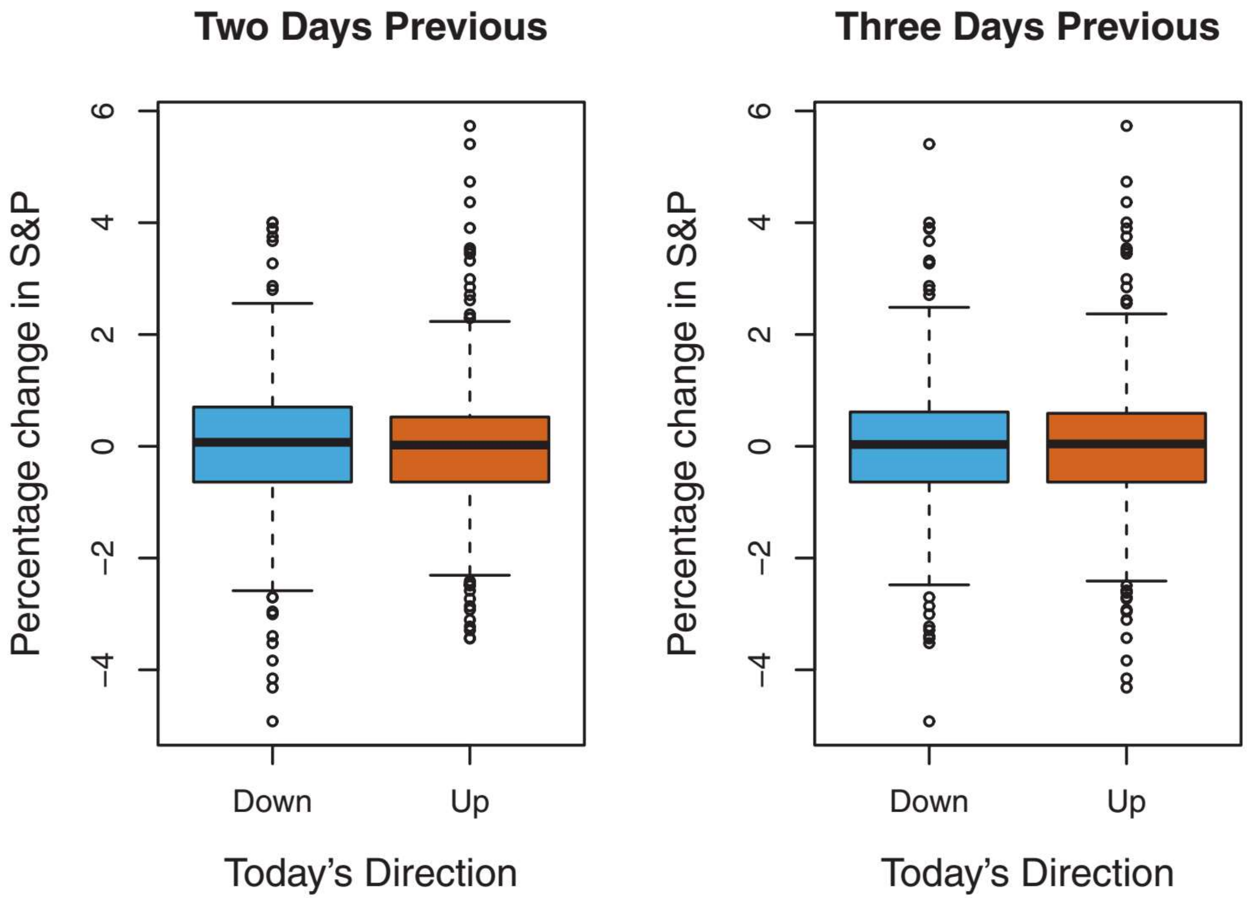
data from 648 days for which the market increased on the following day

Is it enough to make our predictions based on previous day changes only?

"Smarket" – stock market dataset



"Smarket" – stock market dataset



- **Little association** between pervious days and present returns.
- That is somehow **expected** due to **strong correlations between returns on successive days**.

- What more can we say through mining techniques? – later

Gene expression dataset

- A case where we **only have inputs variables with no output** -> **clustering problem**.

- The $NCI60$ dataset consists of the expression values of 6830 genes for each of 60 cell lines.

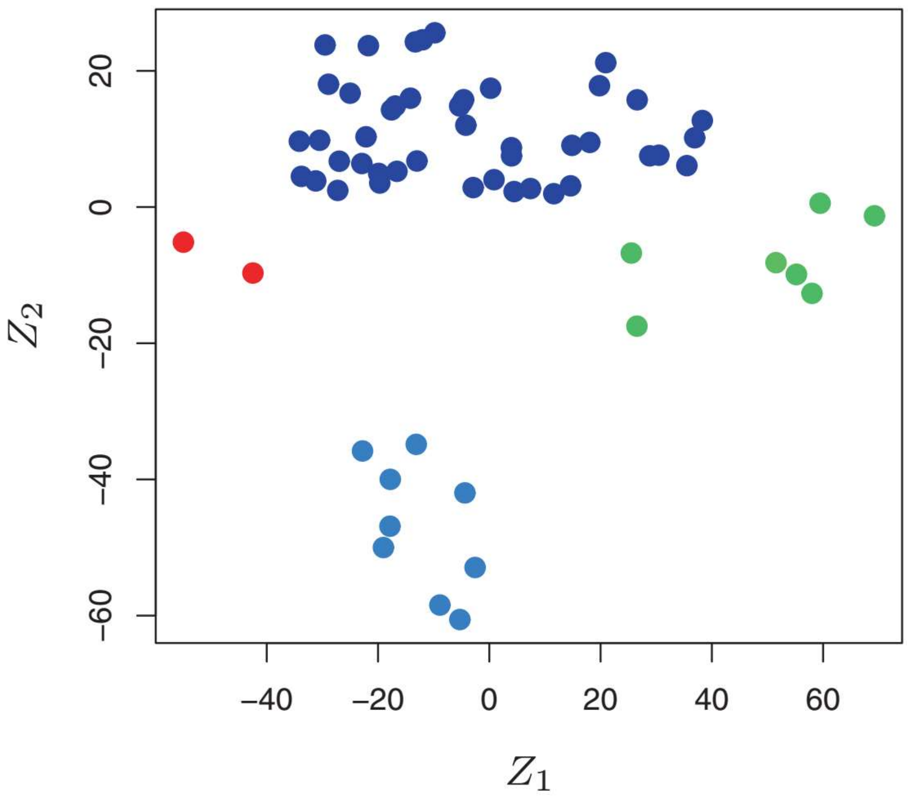
    - Can we group cell lines based on their gene expression measurements?

    - Knowing that we have thousands of values per cell line, **how can we visualize such data**?

    - **Principal components** summarize data in smaller dimensions.

	Sample 1	2	3	...	60
g1	*				
g2	*				
g3	*				
⋮					
g~20k					

Gene expression dataset

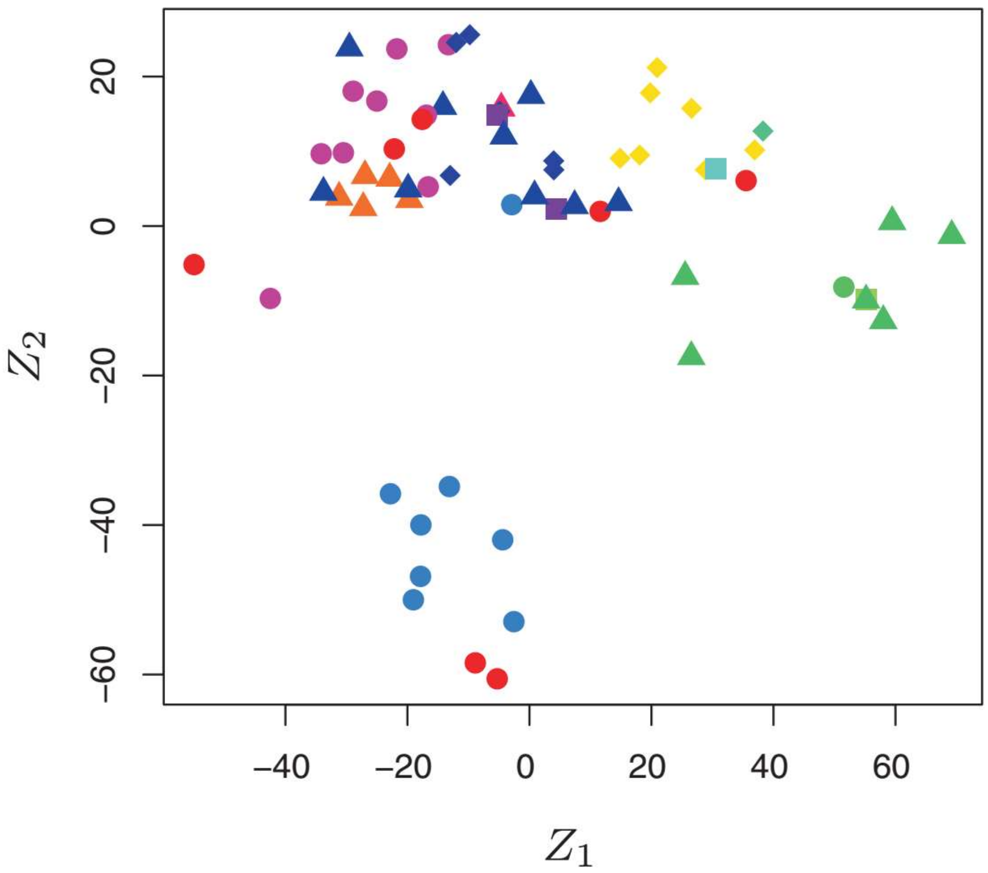


- Here, the first two components Z_1 and Z_2 summarize the expression of 6380 measurements for each cell line in just two numbers (or dimensions)
- Tradeoff as some information will be lost, but efficient visualization is acquired.
- **4 groups (clusters) of cell lines identified** and can then be further examined for **similarities in their cancer types, …, relationship between gene expression and cancer, …**

Gene expression dataset

- We also know that the cell lines come from **14 different cancer types**, but this information was **not used in the previous graph**.

- When added, we get a similar graph, but this time it **shows that cell lines from the same cancer type tend to be grouped together -> independent verification of the analysis**.



Notations

- $\boldsymbol{n}$: number of distinct data points (observations) in a sample
- $\boldsymbol{p}$: number of available variables (attributes)

- $\boldsymbol{x_{ij}}$: j^{th} variable for the i^{th} observation, $i = 1,2,\dots,n$ and $j = 1,2,\dots,p$

$\boldsymbol{n} \times \boldsymbol{p}$ matrix

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

$\boldsymbol{x_i}$: vector containing p variables of the i^{th} observation, represented as column by default

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$$

$\boldsymbol{x_j}$: vector of length n containing observations of values of variable j for n observations

$$\mathbf{x}_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}$$

Notations

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

$$\mathbf{X}^T = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{pmatrix}$$

$$x_i^T = \begin{pmatrix} x_{i1} & x_{i2} & \cdots & x_{ip} \end{pmatrix}$$

Notations

- $\boldsymbol{y_i}$: the $\boldsymbol{i^{th}}$ **observation of the variable on which we want to make predictions** (e.g., wage) -> the set of all observations in vector form

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

- The observed data can be represented as:

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

where x_i is a vector of length p

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$$

Advertising dataset

- The goal is to **increase the sales** of a certain product.
- The dataset consists of:
  - **The sales of that product in 200 markets**
  - **The advertising budgets** for the product **in those markets** (**TV, radio, and newspaper**).

Advertising dataset

- **What is a suitable expenditure strategy on advertising?**

	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9
6	8.7	48.9	75	7.2
7	57.5	32.8	23.5	11.8
8	120.2	19.6	11.6	13.2
9	8.6	2.1	1	4.8
10	199.8	2.6	21.2	10.6
11	66.1	5.8	24.2	8.6
12	214.7	24	4	17.4
13	23.8	35.1	65.9	9.2
14	97.5	7.6	7.2	9.7
15	204.1	32.9	46	19
16	195.4	47.7	52.9	22.4
17	67.8	36.6	114	12.5
18	281.4	39.6	55.8	24.4
19	69.2	20.5	18.3	11.3
20	147.3	23.9	19.1	14.6
21	218.4	27.7	53.4	18
22	237.4	5.1	23.5	12.5
23	13.2	15.9	49.6	5.6
24	228.3	16.9	26.2	15.5
25	62.3	12.6	18.3	9.7
26	262.9	3.5	19.5	12
27	142.9	29.3	12.6	15
28	240.1	16.7	22.9	15.9

Advertising dataset

- We need to develop **an accurate model that can be used to predict sales based on budgets for the three media**.

- What are the input and output variables?

    - **Input variables:** advertising budgets

        - let $\boldsymbol{X_1}, \boldsymbol{X_2}$**, and** $\boldsymbol{X_3}$ be the TV, radio, and newspaper budgets, respectively.

    - **Output variable:** sales, denoted by $\boldsymbol{Y}$.

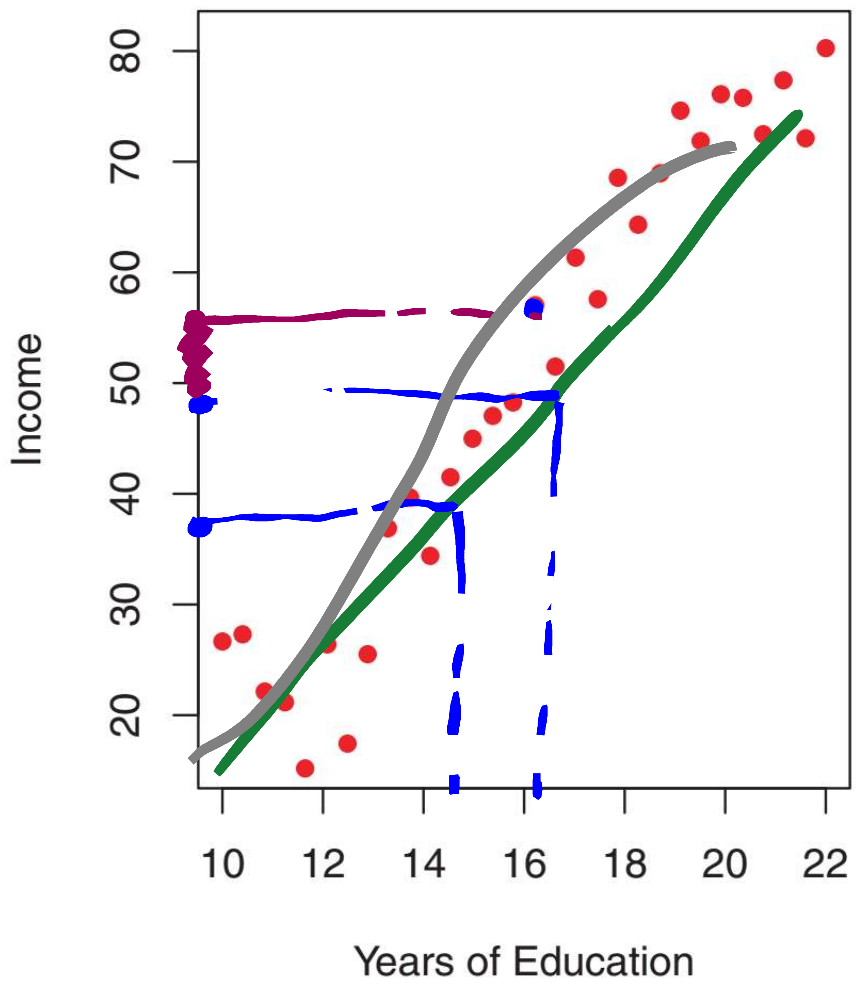
Advertising dataset

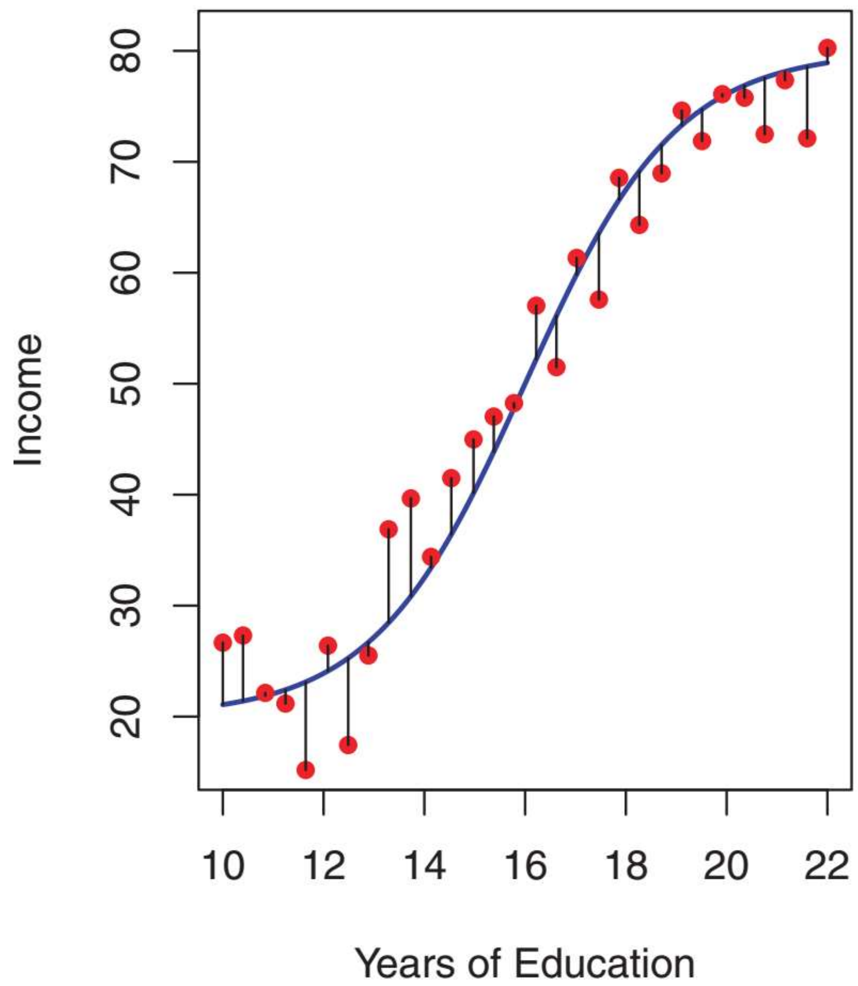
- The **relationship between a quantitative response** $\boldsymbol{Y}$ (here sales) **and** $\boldsymbol{p}$ **predictors** $X = (X_1, X_2, \ldots, X_p)$ (here $X = (X_1, X_2, X_3)$) can be written as:

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

$\in$ is a random error term that is independent of X and has mean zero.

Let’s go back to the “Wages” dataset

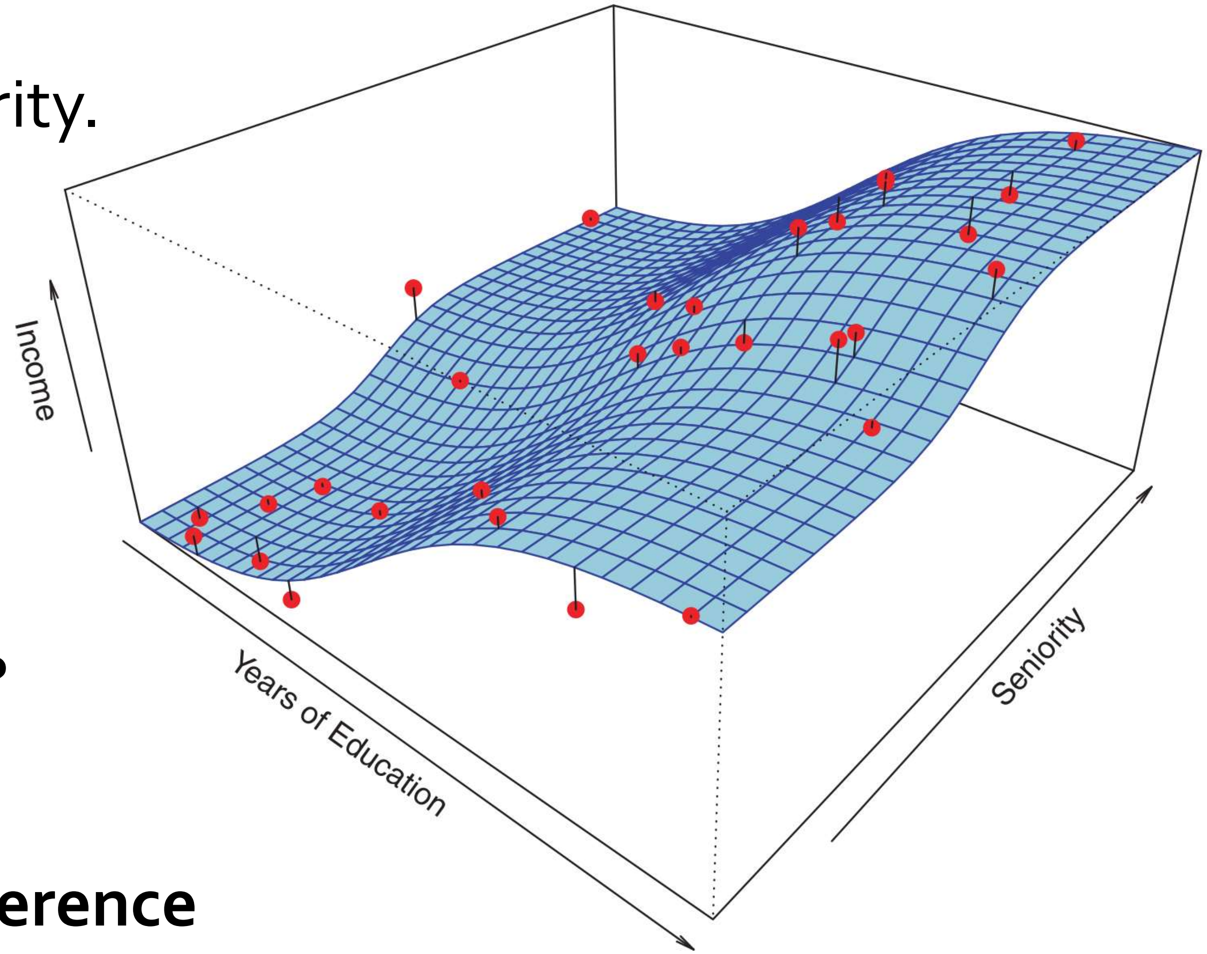




- The blue curve represents the true relationship (which is usually unknown) based on the observed points.
- Vertical lines correspond to ∈ (positive if above the curve)
  - overall, the error mean is approx. zero.

For the "Wages" dataset

- More/Which input variables?
- Here, income as a (true) function of years of education and seniority.



Why estimate *f*,
generally speaking**?**

prediction **inference**

Prediction

- Predict Y given a set of inputs X.

- Y can be predicted using:

$$\hat{Y} = \hat{f}(X)$$

resulting prediction for Y

estimate for f

Prediction – example

$$\hat{Y} = \hat{f}(X)$$

- Let $X_1, X_2, \ldots, X_p$ be the measured **characteristics of a blood sample**
    - Let Y be a variable corresponding to the **patient's risk of a severe adverse reaction to a drug**.
- In such settings, **two factors determine the accuracy of** $\hat{Y}$:

Prediction – example

$$\hat{Y} = \hat{f}(X)$$

- Let $X_1, X_2, \ldots, X_p$ be the measured **characteristics of a blood sample** and let Y be a variable corresponding to the **patient's risk of a severe adverse reaction to a drug**.

- In such settings, **two factors determine the accuracy of $\hat{Y}$:**

  - **<u>reducible error:</u>** usually $\hat{f}$ is not expected to be a perfect estimate of f -> error.

    - **Reducible because we can improve the accuracy** (i.e., reduce the error) by using more appropriate learning techniques.

Prediction – example

$$\hat{Y} = \hat{f}(X)$$

- Let $\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_p}$ be the measured **characteristics of a blood sample** and let $\boldsymbol{Y}$ be a variable corresponding to the **patient's risk of a severe adverse reaction to a drug**.

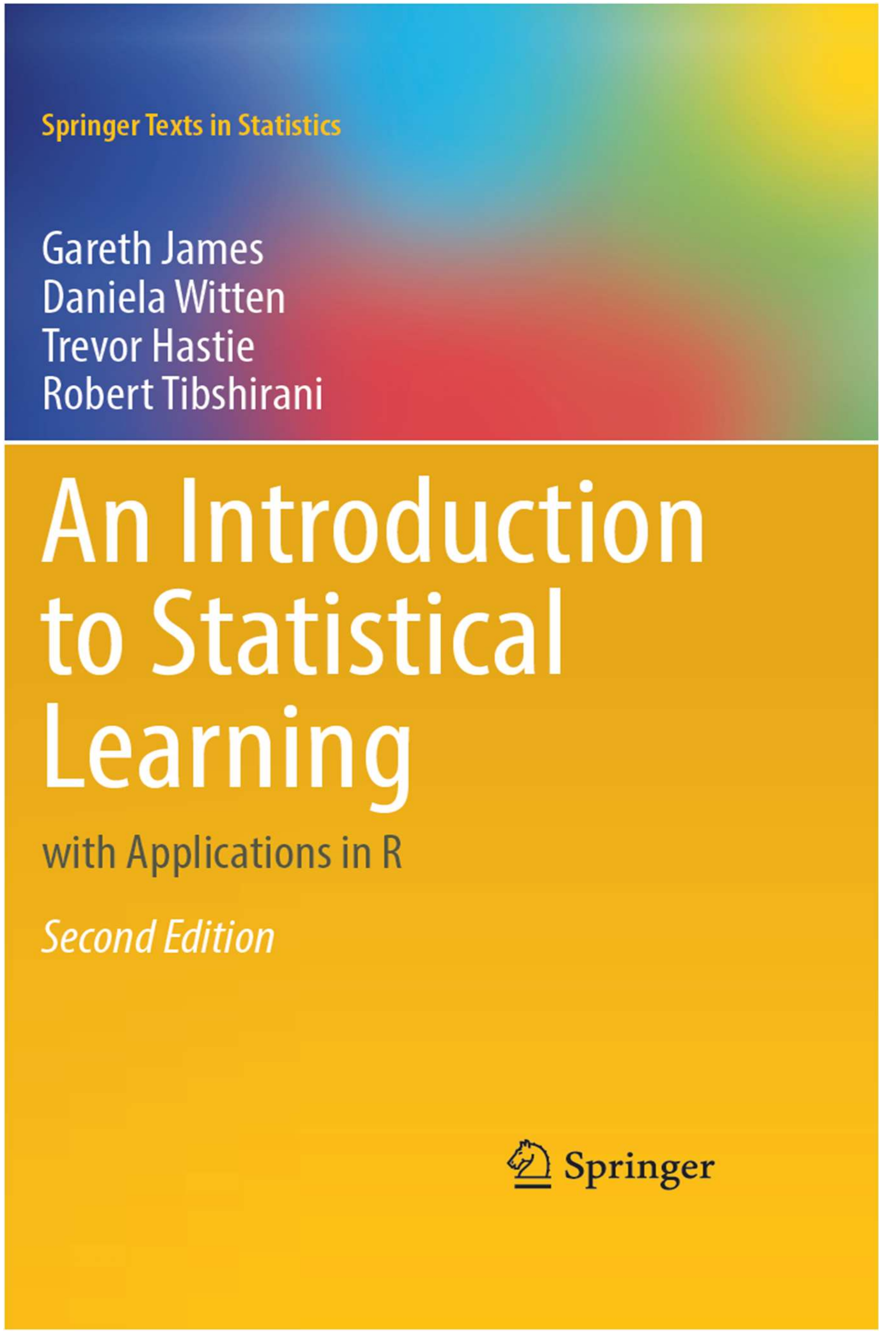
- In such settings, **two factors determine the accuracy of $\widehat{\boldsymbol{Y}}$:**

  - **reducible error**

  - <u>**irreducible error:**</u> there **will always be** an irreducible error introduced by $\in$ because Y is also a function of $\in$, which cannot be predicted by X.

    - Due to some **unmeasured or unmeasurable factors**
      - e.g., the manufacturing variation of the drug itself or the patient's wellbeing

Reference


Springer Texts in Statistics
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R
Second Edition
Springer