数据科学介绍1.1： 课程简介

Introduction to Data Science

Part1.1: Intro

Goals & Limitations

☐Goals:

- Basic data science methods
- Using Matlab to do data science
- Be a data scientist apprentice

☐Limitations:

- Only 2 lectures

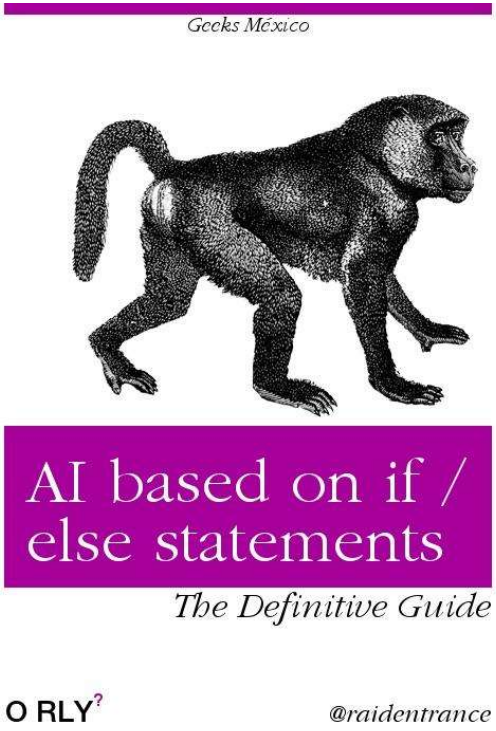
☐Then what? :

- You have to spend more time if you are interested.

What is data science, AI

☐What is data science ?

- Science that studies
- Data collection, stor tistics, and of course machine learning

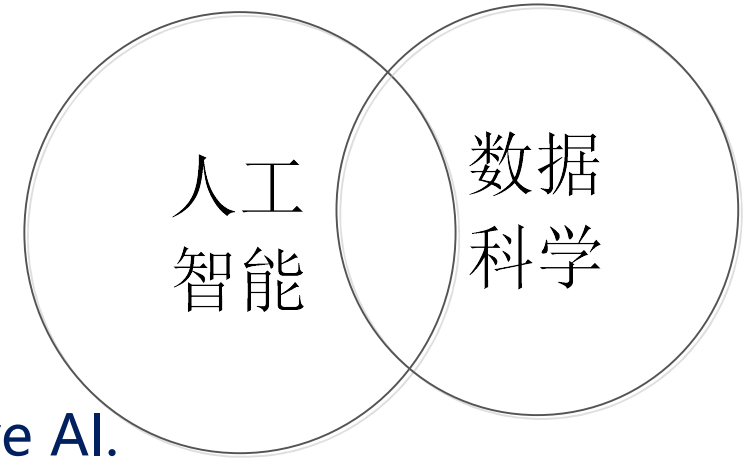


☐What is AI?

- Make machine do ir

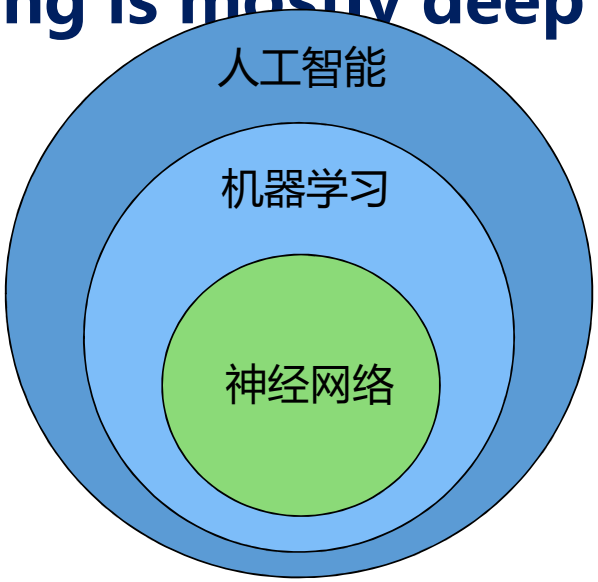
☐Relations:

- Some data science r achieve AI.



Machine Learning

- ML is a way of achieve AI
- artificial neural network is a way of achieve ML
- Deep Learning is mostly deep artificial neural network



How do we learn?

- Problem based
- Using Matlab, hands on

You will learn

- ☐ Simple linear regression model
- ☐ Simple linear time series model
- ☐ Basic classification
- ☐ Basic clustering
- ☐ Basic neural network

Target of this class

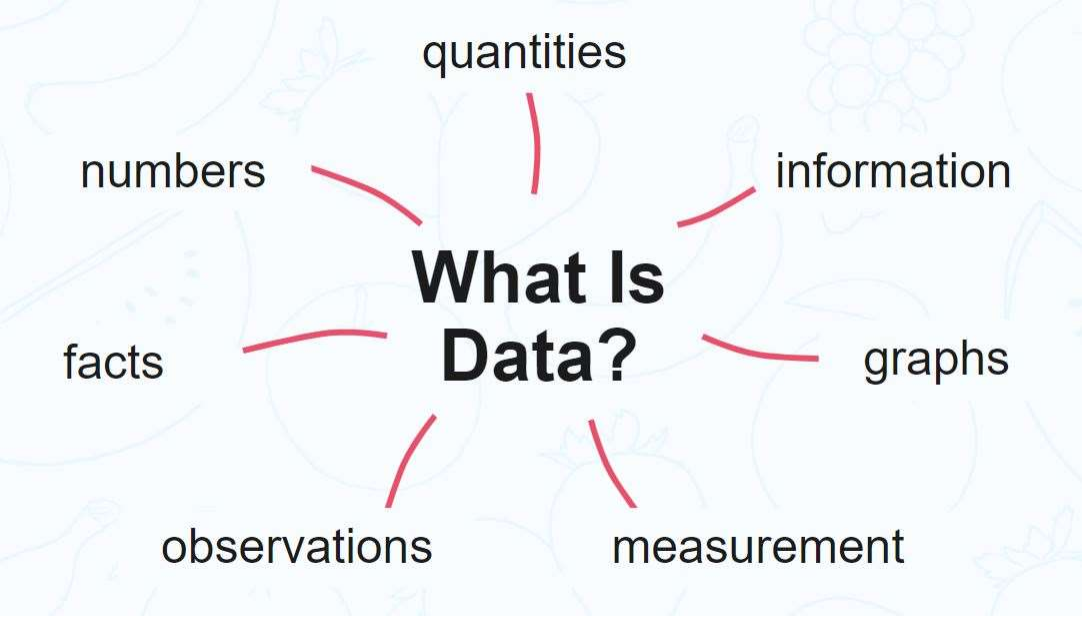
- Visualize you data and get a basic idea of it
- Build a model to predict something or into the future
- All in Matlab

数据科学介绍，探索性数据分析

Introduction to Data Science Part
1.1.1: Exploratory Data Analysis

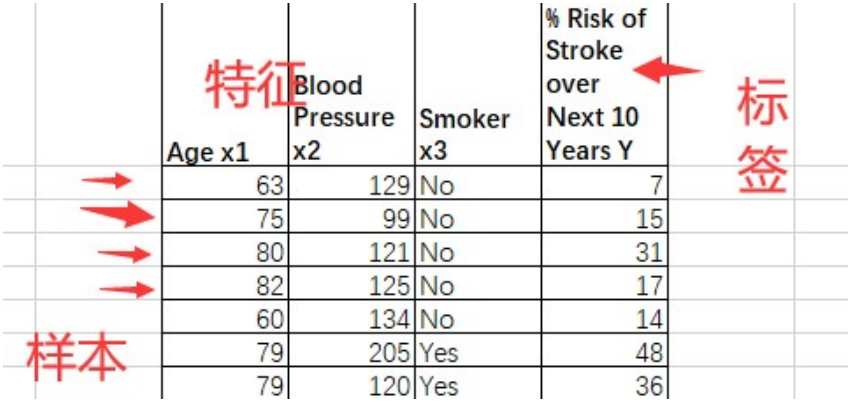
What is data

- Data is a formal representation of information
- Data is obtained by observation, and often numeric, qualitative or quantitative.



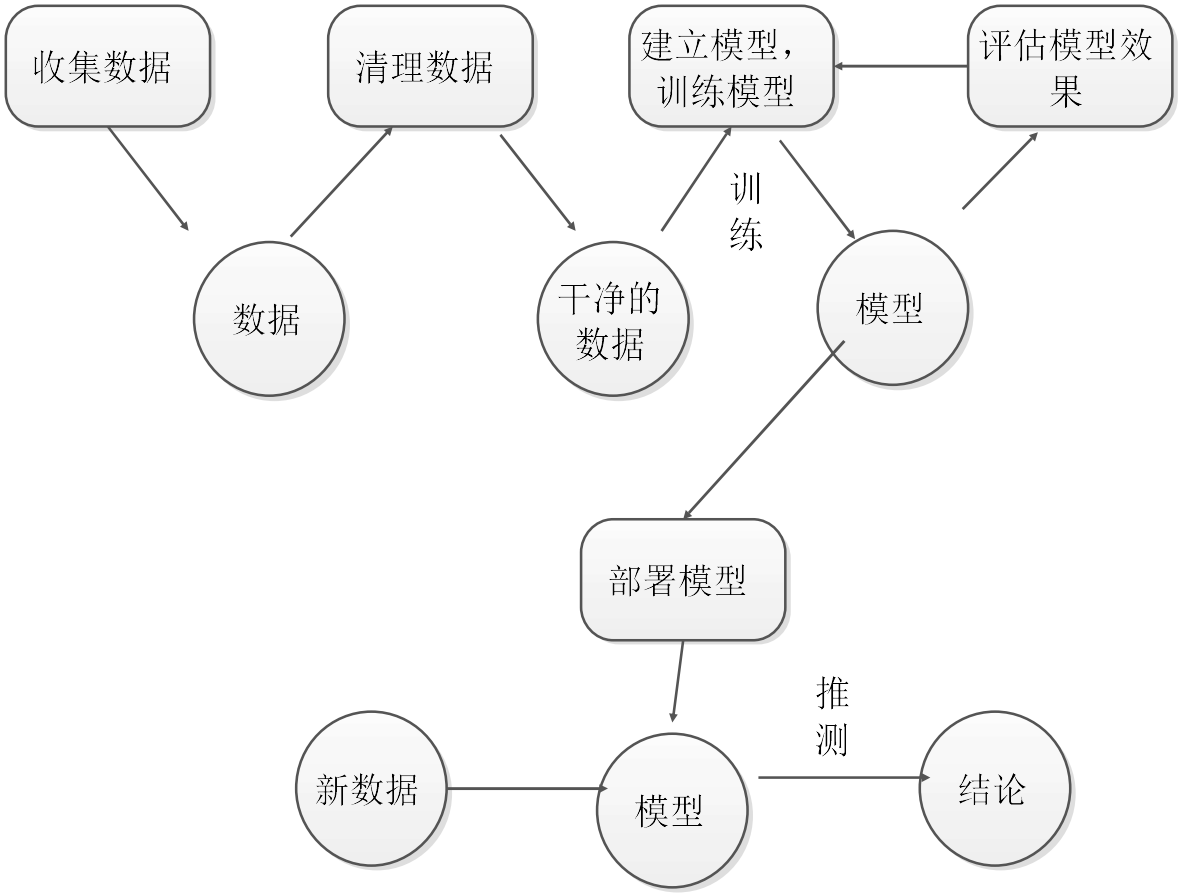
What is sample

- Data is collected via observation
- Observation is also knowns as sampling
- The result of each sampling process is a sample



Age x1	Blood Pressure x2	Smoker x3	% Risk of Stroke over Next 10 Years Y
63	129	No	7
75	99	No	15
80	121	No	31
82	125	No	17
60	134	No	14
79	205	Yes	48
79	120	Yes	36

The routine of data science

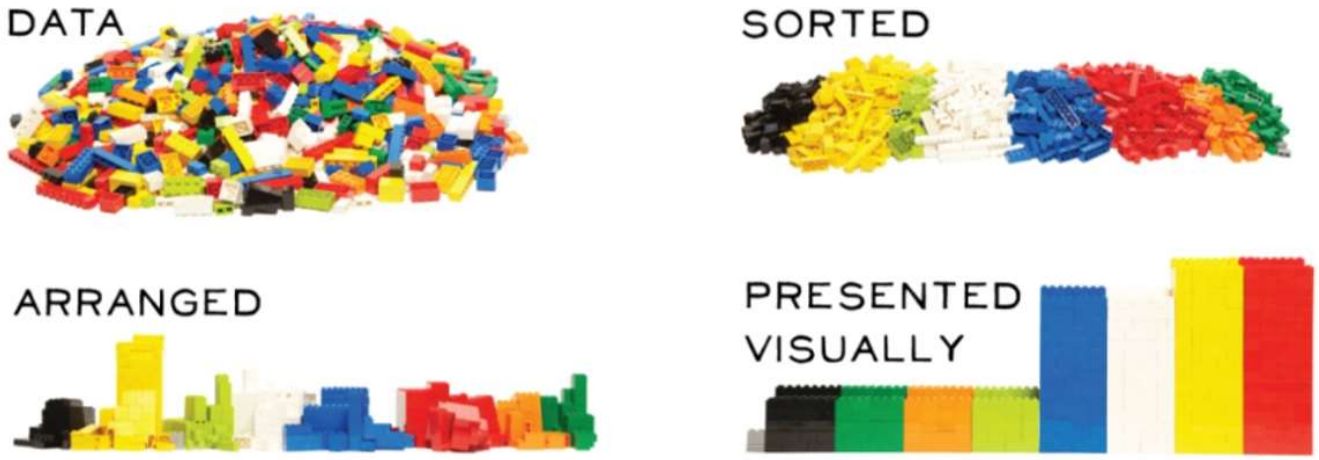


Types of data

- **Categorical**
  - Nominal
  - Ordinal
- **Continuous**
  - Interval
  - Ration

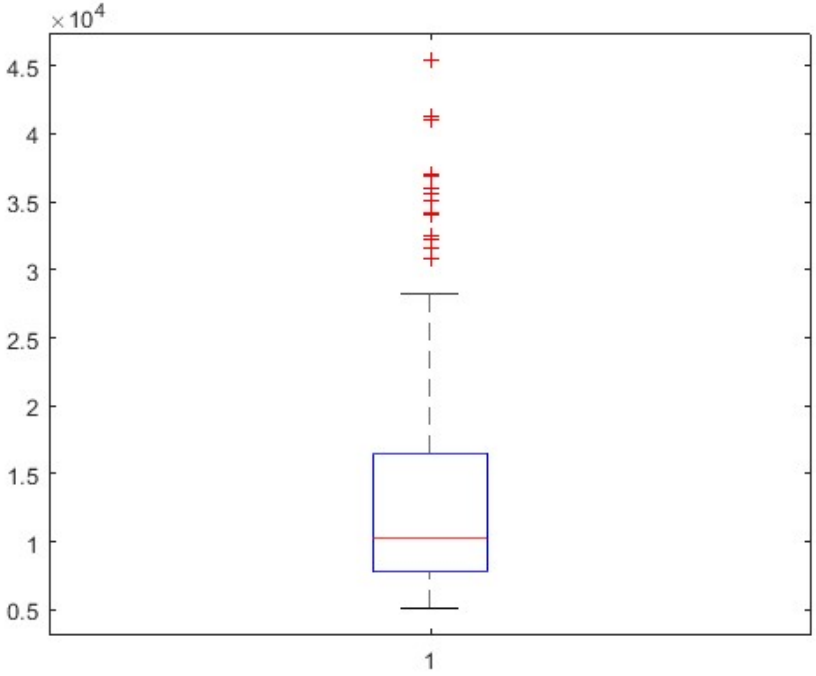
Exploratory Data Analysis

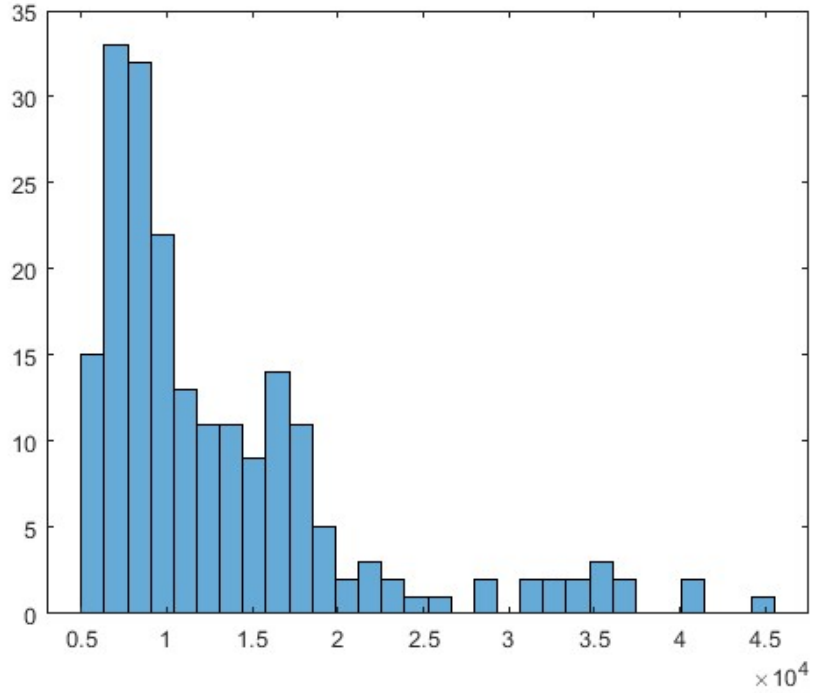
- Know you data without building any model
- Extremely helpful to model building
- Mostly done via data visualization



Descriptive statistics

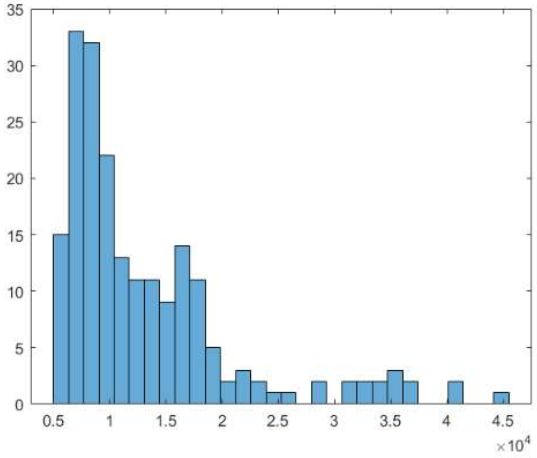
- Mean, median, range, variability

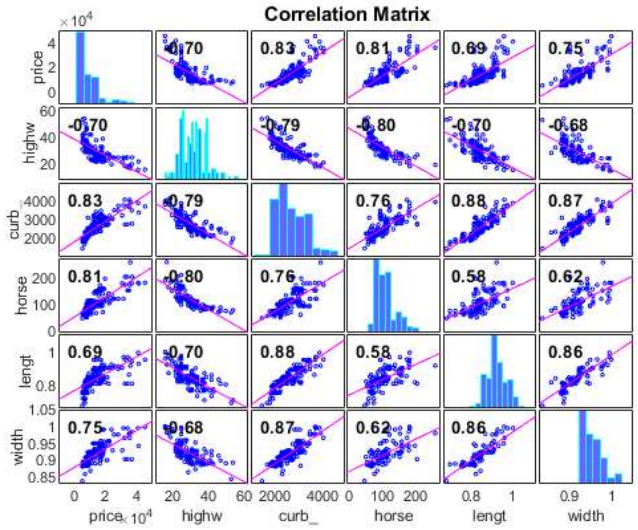


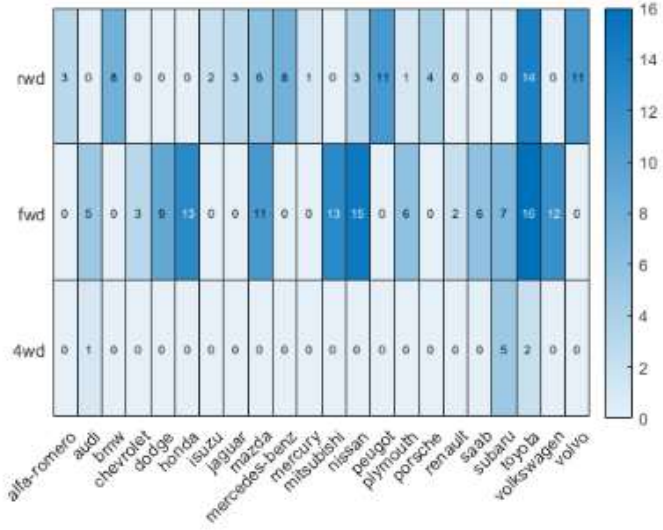


Relations between multiple variables

☐Case 1 – Exploratory Data Analysis







Relations between multiple variables

☐Continuous vs. Continuous

- Correlation (corrplot，scatter)

☐Categorical vs. Categorical

- Contingency table (crosstab)

☐Continuous vs. Categorical

- Boxplot，histogram，pie

数据科学入门1.2：回归的基本概念

Introduction to Data Science
Part 1.2: Basic definitions

What is Regression models

- A statistical procedure u
  showing how two or more
- Regression Analysis does
  relationship, but rather it
  (equation) that can help
  predictions.

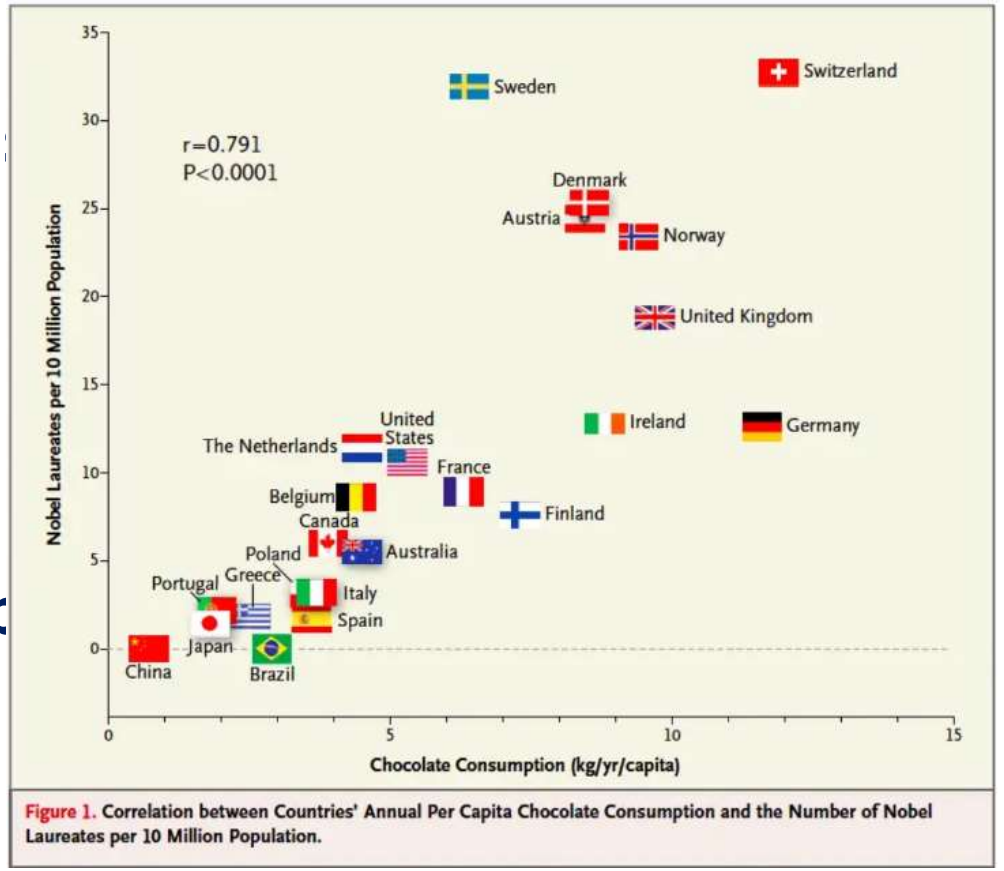


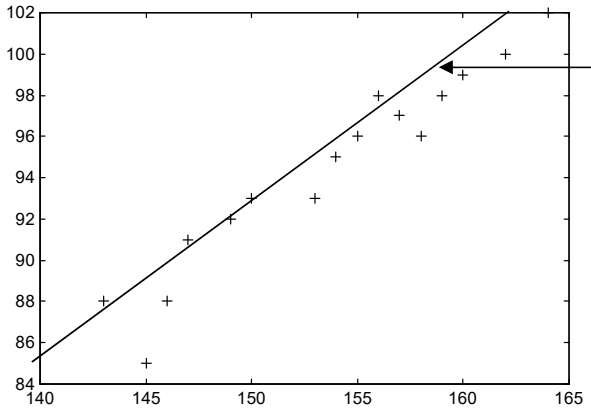
Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

An Example

例1 测16名成年女子的身高与腿长所得数据如下：

身高	143	145	146	147	149	150	153	154	155	156	157	158	159	160	162	164
腿长	88	85	88	91	92	93	93	95	96	98	97	96	98	99	100	102

以身高x为横坐标，以腿长y为纵坐标将这些数据点（xi, yi）在平面直角坐标系上标出.



$$y = \beta_0 + \beta_1 x + \varepsilon$$

散点图 Scatter Plot

Regression models

- **Regression models involve the following parameters and variables:**
  - The unknown parameters, denoted as β
  - The independent variables, X
  - The dependent variable, Y.

$$\boldsymbol{E(Y|X) = f(X, \beta)} \quad \boldsymbol{\widehat{Y} = f(X, \beta)}$$

  - If $\boldsymbol{f}$ is a linear function of β then it's linear regression (not X)

$$\boldsymbol{\widehat{Y} = \beta_0 + \beta_1 x} \qquad \boldsymbol{\widehat{Y} = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3}$$

$$\boldsymbol{\widehat{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}$$

Simple Linear Regression

- $\widehat{\boldsymbol{Y}} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x}$

$$\min\left[\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2\right]$$

$$\beta_1 = \frac{\sum_{i=1}^{n}\left(X_i - X_{bar}\right)\left(Y_i - Y_{bar}\right)}{\sum_{i=1}^{n}\left(X_i - X_{bar}\right)^2}$$

$$\beta_0 = Y_{bar} - \beta_1 X_{bar}$$

It is a function of β, not X.

数据科学入门1.3：简单线性回归

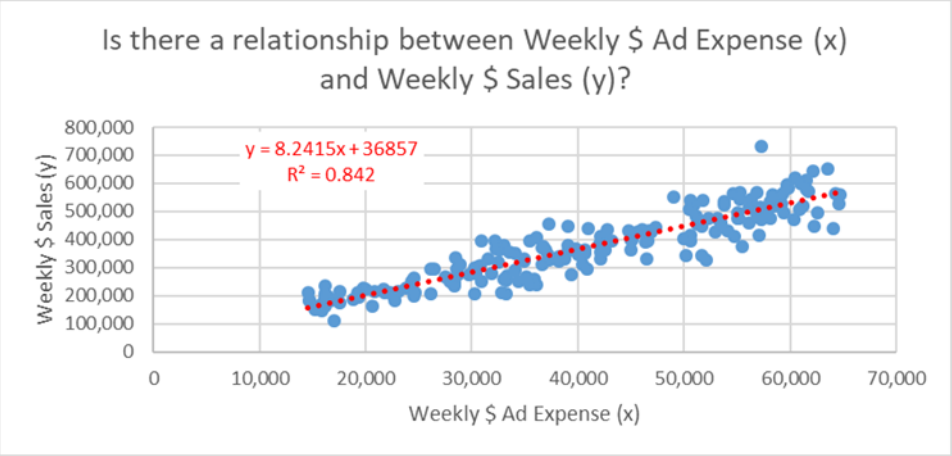
Introduction to Data Science

Part 1.3: Basic Regression

Simple Linear Regression

☐Case 2 – Expense vs. Sales

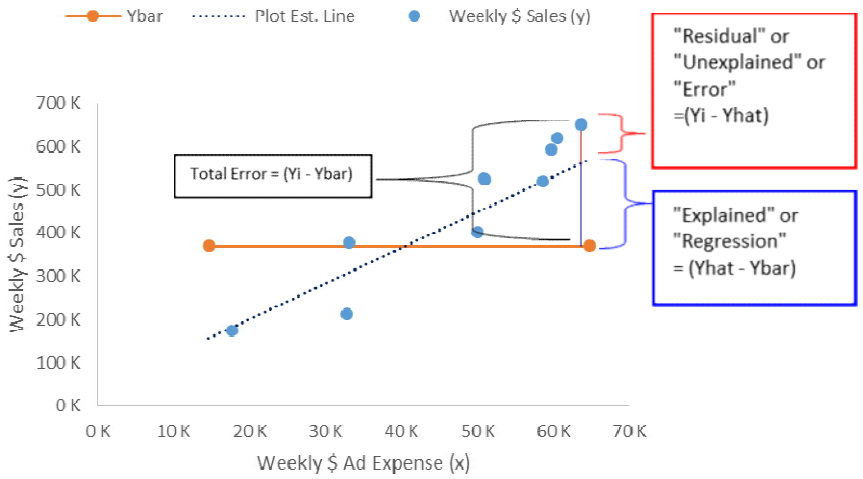
Weekly $ Ad Expense (x)	Weekly $ Sales (y)
63,566	651,334
50,762	527,670
50,941	523,751
17,597	175,467
33,029	377,978
58,543	520,100
60,492	620,856
59,686	593,739
16,432	181,949
17,262	184,644
39,118	379,374
36,078	238,688
42,113	410,066
50,562	413,541
38,240	340,242
59,870	582,843



Evaluation of the regression result

☐Coefficient of determination

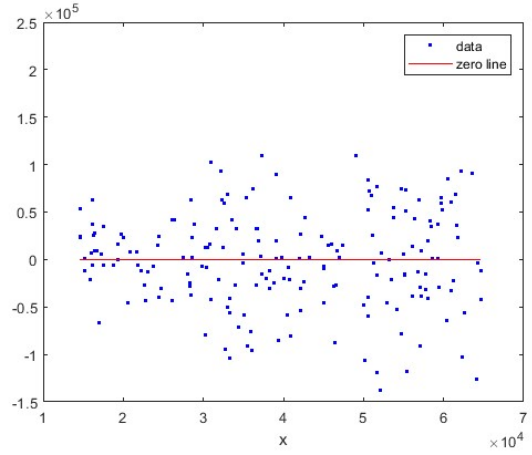
- $R^2 = \frac{SSR}{SST}$, $SSR = \sum(\hat{y}_i - \bar{y})^2$ $SST = \sum(y_i - \bar{y})^2$



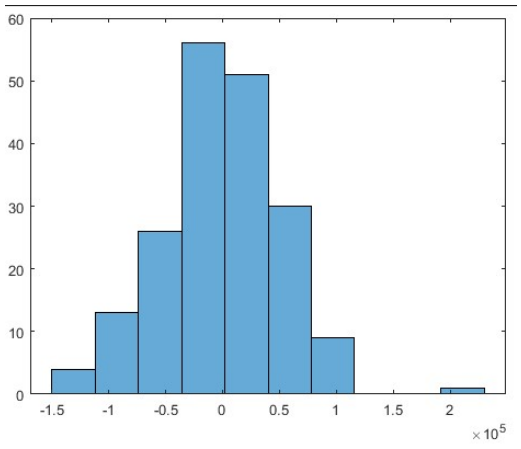
The closer is Adjusted-R-Squared to 1, the better

Evaluation of the regression result

- A good regression should give normally distributed residual with mean of 0 across all independent variable values.



OK



OK

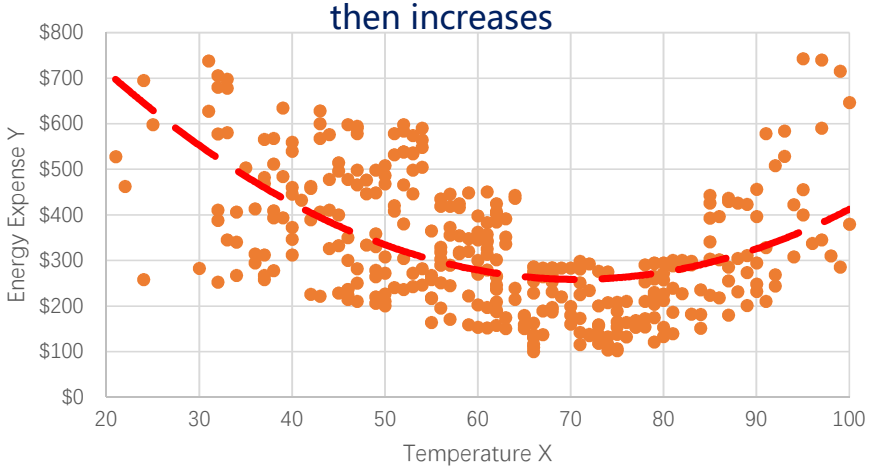
NOT OK

Polynomial Linear Regression

☐Case 3 – Temperature vs. Expense

Date	Temperature X	Energy Expense Y
2015/1/12	46	$236
2015/1/13	52	$304
2015/1/14	55	$164
2015/1/15	46	$214
2015/1/16	47	$210
2015/1/17	50	$508
2015/1/18	36	$295
2015/1/19	47	$250
2015/1/20	40	$372
2015/1/21	46	$478
2015/1/22	55	$258
2015/1/23	40	$559
2015/1/24	53	$536
2015/1/25	44	$576

Relationship Between Temperature and Energy Expense, relationship looks nonlinear: as x increases, y decreases for a while and then increases



$y = 0.1782x^2 - 25.164x + 1147.5$

$R^2 = 0.3069$

数据科学入门1.4：多变量回归

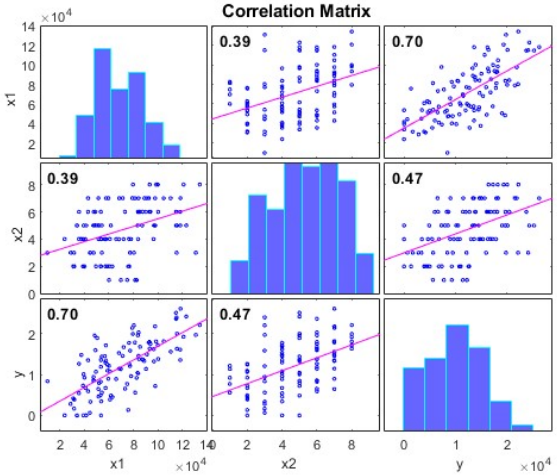
Introduction to Data Science

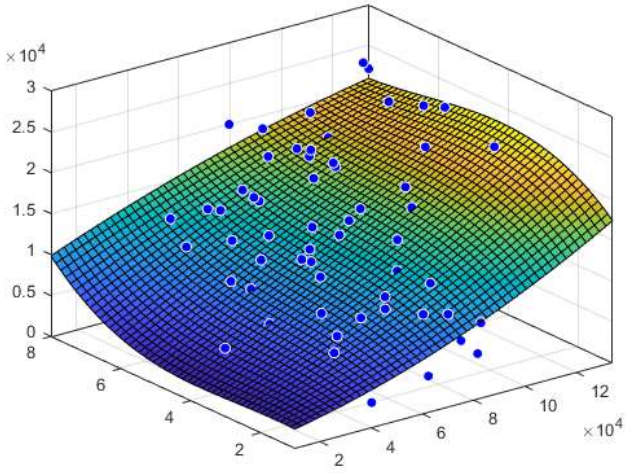
Part 1.4: Multi variable Regression

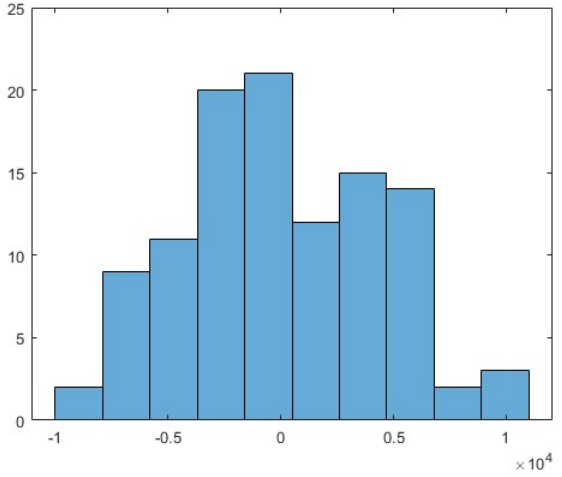
- Continuous variables
- Categorical variables

□Case 4 - Continuous variables

Income & Education vs. Annual Credit Card Charges







☐Case 5 - Categorical variables

Age & Blood Pressure & Smoker vs. Risk of Stroke

Age x1	Blood Pressure x2	Smoker x3	% Risk of Stroke over Next 10 Years Y
63	129	No	7
75	99	No	15
80	121	No	31
82	125	No	17
60	134	No	14
79	205	Yes	48
79	120	Yes	36
82	138	Yes	37
64	192	No	28
53	159	No	13
59	151	Yes	18
88	177	Yes	56
80	130	Yes	34
64	209	Yes	37
69	131	Yes	15
68	172	Yes	36

数据科学介绍1.5：非线性回归和logistic 回归

Introduction to Data Science

Part 1.5: Non-linear regression and logistic regression

Non-linear regression

□Case 6 -Chemical kinetic reaction

在化学动力学反应过程中，建立了一个反应速度和反应物含量的数学模型，形式为：

$$y = \frac{\beta_4 x_2 - \frac{x_3}{\beta_5}}{1 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}$$

- 今测得一组参考数据，求$\beta_1 \sim \beta_5$

表 7.16 反应数据

序号	反应速度 y	氢 x_1	n 戊烷 x_2	异构戊烷 x_3	序号	反应速度 y	氢 x_1	n 戊烷 x_2	异构戊烷 x_3
1	8.55	470	300	10	8	4.35	470	190	65
2	3.79	285	80	10	9	13.00	100	300	54
3	4.82	470	300	120	10	8.50	100	300	120
4	0.02	470	80	120	11	0.05	100	80	120
5	2.75	470	80	10	12	11.32	285	300	10
6	14.39	100	190	10	13	3.13	285	190	120
7	2.54	100	80	65					

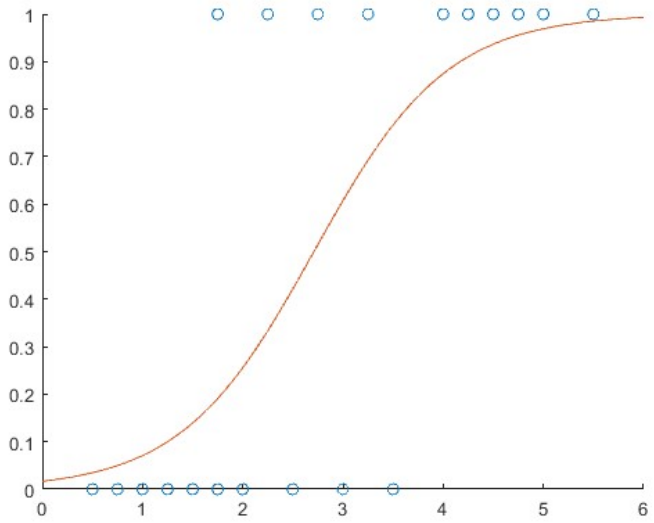
Logistics Regression

☐Case 7 –Learning Time vs. Passing Rate

Hours	0.5	0.75	1	1.25	1.5	1.75	1.75	2	2.25	2.5	2.75	3	3.25	3.5	4	4.25	4.5	4.75	5	5.5
Pass	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1

$$\hat{\theta} = argmax L_n(\theta; y)$$

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$



□Exercise-1

- 训练数据：Exercise_1_Training_Data.xlsx
- 测试数据：Exercise_1_Test_Data.xlsx

利用训练数据建立预测泰坦尼克号上乘客生存概率的模型，并用测试数据进行检验。

□Exercise-2

建立包含不超过4个特征的汽车价格预测模型，并通过已有数据检验你的模型（利用Case-1的数据）。

数据科学入门1.6：简单时间序列预测与总结

Introduction to Data Science
Part 1.6: Time Series and Summary

Time series and ARIMA model

What is a “time series”

首先我们定义一下这个模型。对于一个随机事件，我们每个一段时间观测一次，或者每隔一段时间按顺序发生的一个随机事件，我们叫他随机序列。对于第t次观测值，我叫做 y_t ，前一次就叫做 y_{t-1} ，前一次的钱一次叫做 y_{t-2} ，那么 y_{t-k} 能够理解了吧。

Auto-regressive model

$$y_t = c + \sum_{i=1}^{k} \beta_i * y_{t-i} + \epsilon$$

ARIMA（auto-regeressive integrated moving average）

$$y_t^d = c + \sum_{i=1}^{p} \beta_i L^i y_t^d + \sum_{i=1}^{q} \theta_i L^i \epsilon_t$$

1. lag算子：超简单就是 $L^i y_t = y_{t-i}$

2. d差分算子：y_t^d y的d阶差分，d=1是 就是 $y_t - y_{t-1}$，d是2的时候就是 $(y_t - y_{t-1}) - (y_{t-1} - y_{t-1})$，简单吧。

Vector Autoregression Models

- Just like AR model
- But with a Vector for Y

$$\hat{Y} = C + \sum_{i=1}^{p} AR_i * L^i Y + \sum_{i=1}^{p} \mathrm{B}_i * L^i X$$

Summary

How to do regression analysis?

- Do EDA and get a basic idea of your data
- Decide what is the features (independent variables) and what is the target (dependent variables)
- Decide which type of model you are going to make (use linear model if you are not sure)
- Fit (train, estimate) the model
- Check adjusted-r-square and plot the residual
- Use the model to do prediction or focasting.

Assignments

- **Try the examples in class**
- **Choose from the 3 problems below**
  - Predict car price from no more then 4 features can compare you model performance (auto_clean.csv)
  - Using logistic regression to predict the chance of survival for the Titanic passengers (exp-6-1.train/test.csv)
  - Predict the wind speed using the historical data (exp8.csv)