Data Analytics with Python

Lecture 2

Ilias Suvanov

ilias.suvanov@gmail.com

2020 年 10 月 20 日

Causality and Prediction

Content

Statistics

History of Statistics

- For at least two millennia, these data were mainly tabulations of human and material resources that might be taxed or put to military use.

History of Statistics

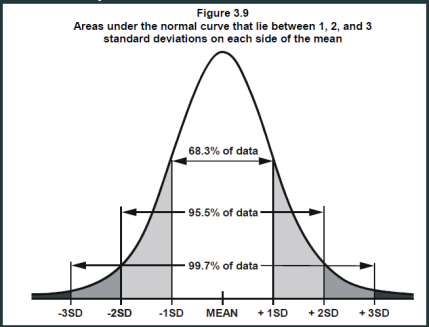
- For at least two millennia, these data were mainly tabulations of human and material resources that might be taxed or put to military use.
- AD 2 Chinese census under the Han dynasty finds 57.67million people in 12.36million households –the first census from which data survives, and still considered by scholars to have been accurate.

History of Statistics

- For at least two millennia, these data were mainly tabulations of human and material resources that might be taxed or put to military use.
- AD 2 Chinese census under the Han dynasty finds 57.67million people in 12.36million households –the first census from which data survives, and still considered by scholars to have been accurate.
- AD 7 Census by Quirinus, governor of the Roman province of Judea, is mentioned in Luke’s Gospel as causing Joseph and Mary to travel to Bethlehem to be taxed.

History of Statistics

- For at least two millennia, these data were mainly tabulations of human and material resources that might be taxed or put to military use.
- AD 2 Chinese census under the Han dynasty finds 57.67million people in 12.36million households –the first census from which data survives, and still considered by scholars to have been accurate.
- AD 7 Census by Quirinus, governor of the Roman province of Judea, is mentioned in Luke’s Gospel as causing Joseph and Mary to travel to Bethlehem to be taxed.
- 1808 Gauss, with contributions from Laplace, derives the normal distribution –the bell-shaped curve fundamental to the study of variation and error.



- 1908 William Sealy Gosset, chief brewer for Guinness in Dublin, describes the t-test. It uses a small number of samples to ensure that every brew tastes equally good.

- 1908 William Sealy Gosset, chief brewer for Guinness in Dublin, describes the t-test. It uses a small number of samples to ensure that every brew tastes equally good.
- 1950 Richard Doll and Bradford Hill establish the link between cigarette smoking and lung cancer. Despite fierce opposition the result is conclusively proved, to huge public health benefit.

- 1908 William Sealy Gosset, chief brewer for Guinness in Dublin, describes the t-test. It uses a small number of samples to ensure that every brew tastes equally good.
- 1950 Richard Doll and Bradford Hill establish the link between cigarette smoking and lung cancer. Despite fierce opposition the result is conclusively proved, to huge public health benefit.
- 1958 The Kaplan–Meier estimator gives doctors a simple statistical way of judging which treatments work best. It has saved millions of lives.

Poverty

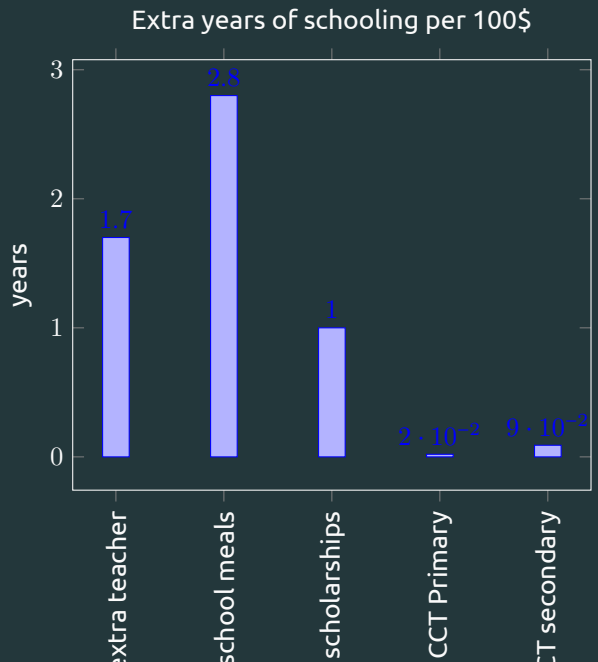
Fighting Poverty

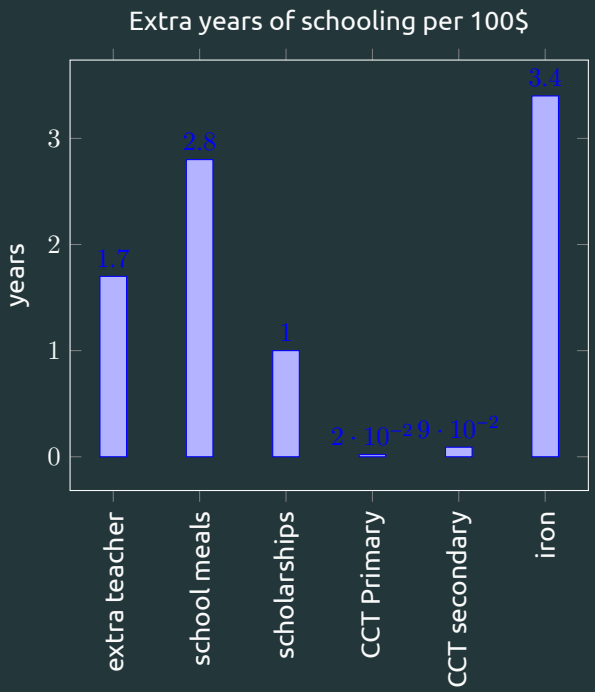
Immunization

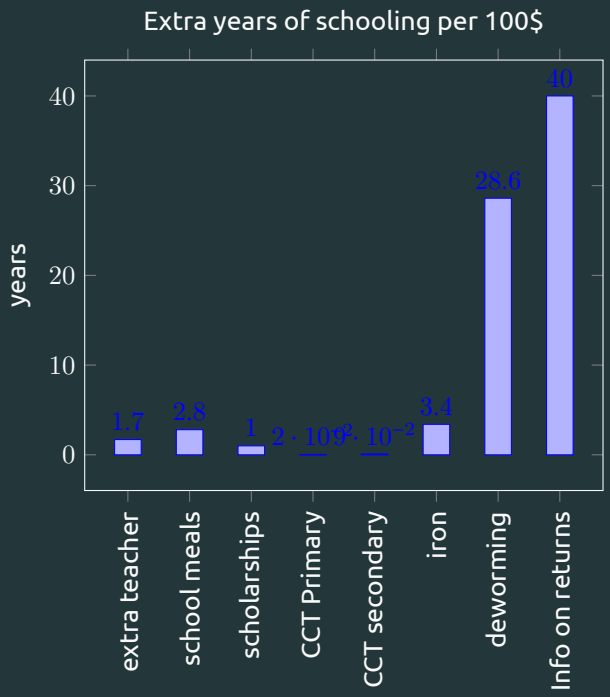
- Corruption
- In 2018, about 86% of the world' s children received vaccines that would protect them against polio, diphtheria, tetanus, pertussis, and measles. Immunizations currently prevent 2 million to 3 million deaths every year. Despite this success, more than 1.5million people worldwide die from vaccine-preventable diseases each year.
- Education

Graph are taken from lecture of Esther Duflo [1]

Education


Extra years of schooling per 100$
years
3
2
1
0
1.7
2.8
1
$2 \cdot 10^{-2}$
$9 \cdot 10^{-2}$
xtra teacher
school meals
scholarships
CCT Primary
CT secondary

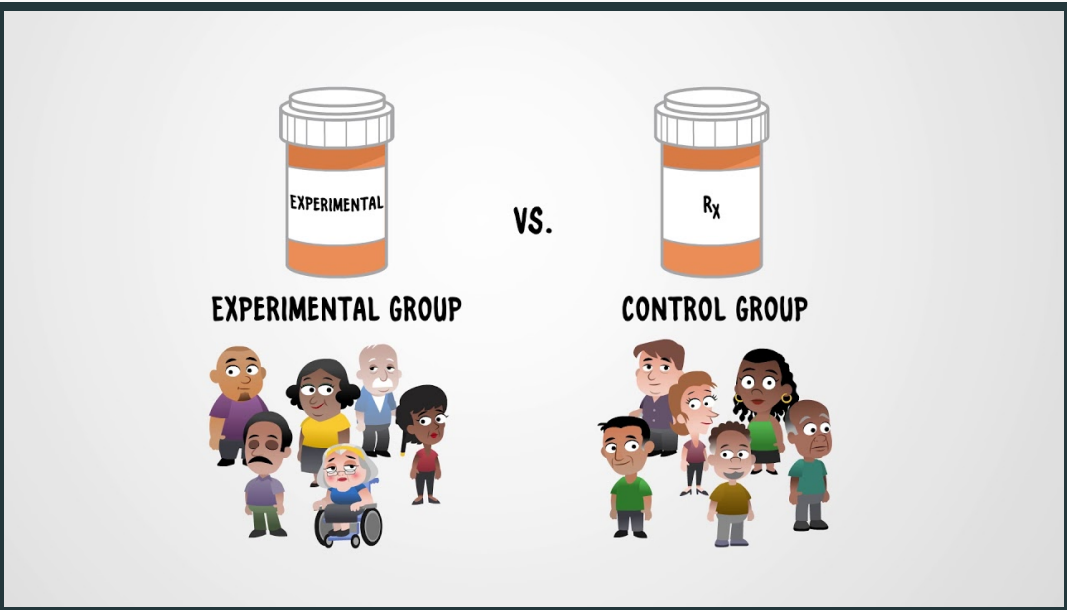

Extra years of schooling per 100$
years
0
1
2
3
1.7
2.8
1
$2 \cdot 10^{-2}$
$9 \cdot 10^{-2}$
3.4
extra teacher
school meals
scholarships
CCT Primary
CCT secondary
iron


Extra years of schooling per 100$
years
0
10
20
30
40
1.7
2.8
1
$2 \cdot 10^{-9}$
$9 \cdot 10^{-2}$
3.4
28.6
40
extra teacher
school meals
scholarships
CCT Primary
CCT secondary
iron
deworming
Info on returns

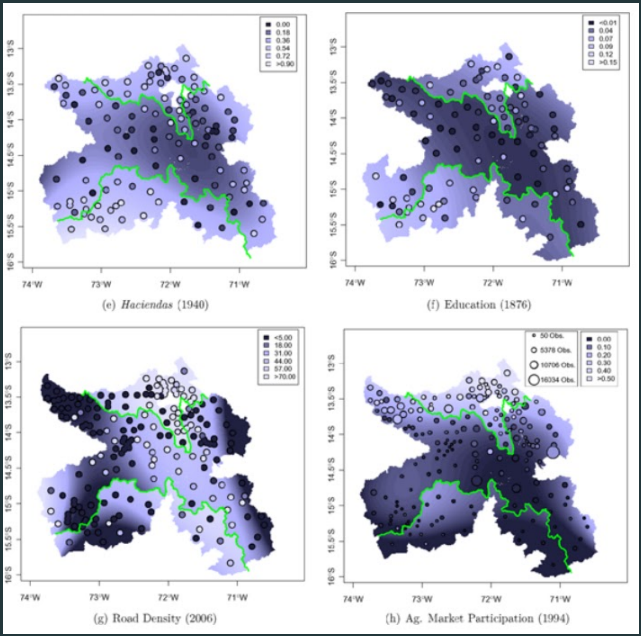
Methods to find causality

- A randomized controlled trial (or RCT)
- Statistical Inference

RCT



Statistical inference



(e) *Haciendas* (1940)

(f) Education (1876)

(g) Road Density (2006)

(h) Ag. Market Participation (1994)

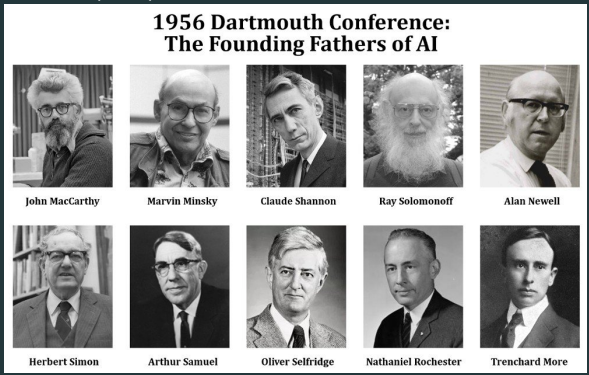
Machine Learning

History of Machine Learning(AI)

- The field of AI research was born at a workshop at Dartmouth College in 1956, where the term ”Artificial Intelligence” was coined by John McCarthy

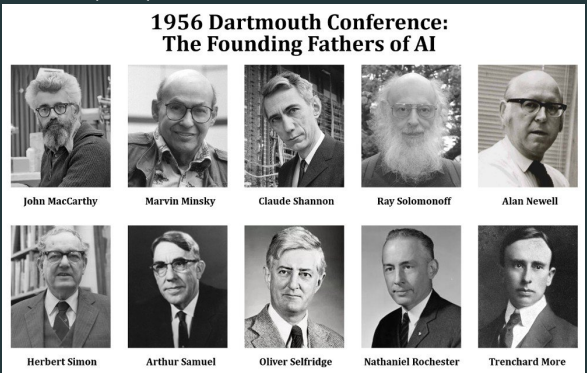
History of Machine Learning(AI)

- The field of AI research was born at a workshop at Dartmouth College in 1956, where the term ”Artificial Intelligence” was coined by John McCarthy
- Allen Newell (CMU), Herbert Simon (CMU), John McCarthy (MIT), Marvin Minsky (MIT) and Arthur Samuel (IBM)



History of Machine Learning(AI)

- The field of AI research was born at a workshop at Dartmouth College in 1956, where the term ”Artificial Intelligence” was coined by John McCarthy
- Allen Newell (CMU), Herbert Simon (CMU), John McCarthy (MIT), Marvin Minsky (MIT) and Arthur Samuel (IBM)



- By 1954 computers were learning checkers strategies and by 1959 were reportedly playing better than the average human, solving word problems in algebra, proving logical theorems.

Initial success

> ”Machines will be capable, within twenty years, of doing any work a man can do”.
>
> —-Herbert Simon(1965)

> ”within a generation ... the problem of creating ’artificial intelligence’ will substantially be solved”
>
> ——Marvin Minsky(1967)

Computer vision

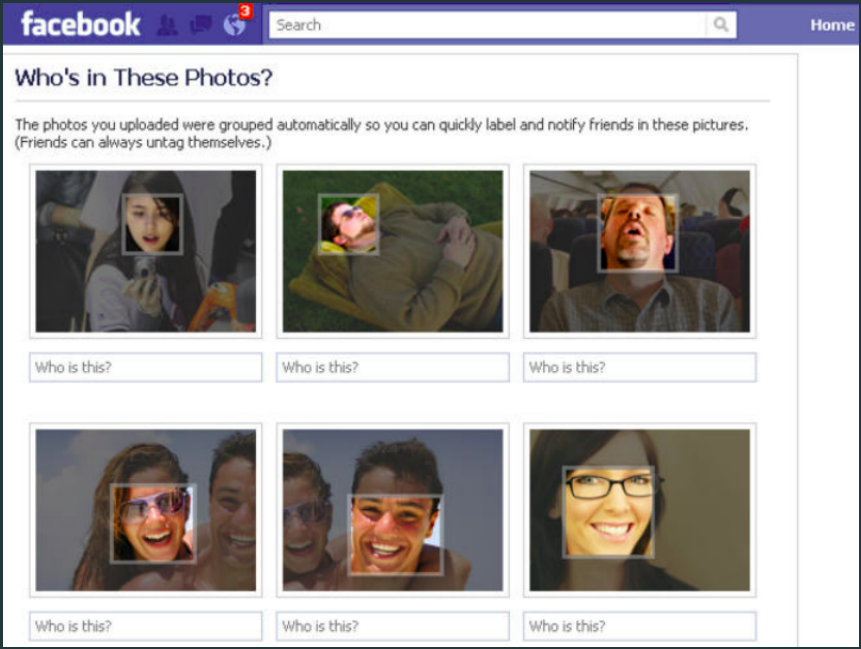


Image Recognition

Sentiment Analysis

Public Policy Applications of Machine Learning

Social Science Applications

- Medicine
- Extracting data(Sattelites, cell Phones)
- Predicting Recidivism
- Isolating violent police officers
- Credit Score
- Building an instrument

Connection between Statistics and Machine Learning

Casuality

1. You want to know structural relationship between x and y, or in other words, you want to find coefficients of function f.

$$y \xleftarrow{f} x$$

$$y = f(x)$$

Prediciton

1. You want to predict y variable, given x variable.

$$y \xleftarrow{f} x$$

$$y = f(x)$$

Statistics(Econometrics)

- Linear Regression
- Logistics Regression
- SVM

Causalty

- Hypothesis testing
- Significant coefficient
- t-statistics

Machine Learning

1. Linear Regression
2. Logistics Regression
3. SVM

Prediction

1. Accuracy
2. R^2
3. Loss function

Thank you for your attention!

Appendix

Materials i

- Esther Duflo.
  Social experiments to fight poverty.
- Sendhil Mullainathan.
  Smarter Algorithms, Better Policy.
  https://www.youtube.com/watch?v=cuGWl3t$_1MI$.