Affaires mondiales Canada | Global Affairs Canada

Canada

CANADIAN FOREIGN SERVICE INSTITUTE | L'INSTITUT CANADIEN DU SERVICE EXTÉRIEUR

Introduction to Data Analysis

BASIC DATA ANALYSIS

Patrick Boily
Data Action Lab | uOttawa | Idlewyld Analytics
pboily@uottawa.ca

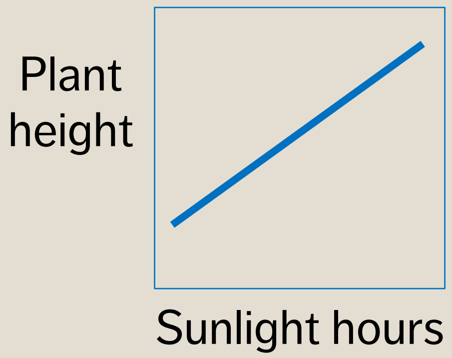
[with files from Jen Schellinck | Sysabee]

DEPENDENT VS. INDEPENDENT VARIABLES

In an *experimental setting*:

- **control/extraneous variables:** we do our best to keep these controlled and unchanging while other variables are changed
- **independent variables:** we control their values as we suspect they influence the dependent variables
- **dependent variables:** we do not control their values; they are generated in some way during the experiment, and presumably are dependent on everything

How do these translate over to other datasets?



DATA TYPES

Numerical data: integers or continuous numbers

- 1, 7, 34.654, 0.000004

Text data: strings of text – may be restricted to a certain number of characters

- “Welcome to the park”, “AAAAA”, “345”, “45.678”

Categorical data: a fixed number of values, may be numeric or represented by strings. **There is no specific or inherent ordering**

- (‘red’,’blue’,’green’), (‘1’,’2’,’3’)

Ordinal data: categorical data with an inherent ordering. Unlike integer data, the spacing between values is **not** defined

- (very cold, cold, tepid, warm, super hot)

DATA SUMMARIZING

Min: smallest value

Max: largest value

Median: “middle” value

Mode: most frequent value

Unique Values: list of unique values

etc.

Signal	Type
4.31	Blue
5.34	Orange
3.79	Blue
5.19	Blue
4.93	Green
5.76	Orange
3.25	Orange
7.12	Orange
2.85	Blue

ROLLING-UP DATA

We can perform operations over a set (or subset) of the data, typically over its **columns**.

Such an operation is akin to **compressing** or '**rolling-up**' the many data values into a single representative value.

Examples: 'mean', 'sum', 'count', 'variance', etc.

We can apply the same roll-up function to many different columns, providing a **mapping** (list) of columns to values.

Signal	Type
4.31	Blue
5.34	Orange
3.79	Blue
5.19	Blue
4.93	Green
5.76	Orange
3.25	Orange
7.12	Orange
2.85	Blue

Count	Signal avg	Signal stdev	Type mode
9	4.73	1.33	Blue/ Orange

CONTINGENCY/PIVOT TABLES

Contingency table: a table which examines the relationship between two categorical variables *via* their relative (**cross-tabulation**).

Pivot table: a table generated by applying operations (sum, count, mean, etc.) to variables, possibly based on another (categorical) variable. Contingency tables as special cases of pivot tables.

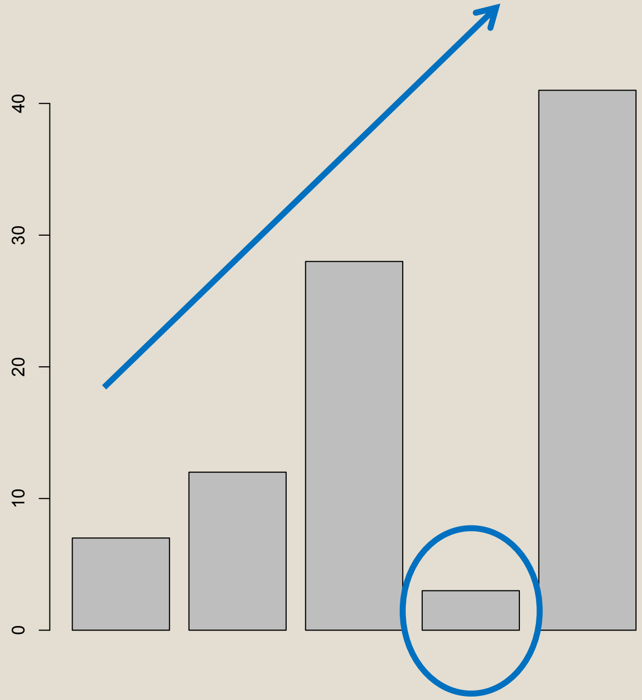
	Large	Medium	Small
Window	1	32	31
Door	14	11	0

Type	Count	Signal avg	Signal stdev
Blue	4	4.04	0.98
Green	1	4.93	N.A.
Orange	4	5.37	1.60

ANALYSIS THROUGH VISUALIZATION

Analysis (broad definition):

- identifying patterns or structure
- adding meaning to these patterns or structure by **interpreting** them in the context of the system.



Option 1: use analytical methods to achieve this.

Option 2: visualize the data and use the brain's analytic power (perceptual) to reach meaningful conclusions about these patterns.

We will discuss further.

DATA DESCRIPTIONS (REPRISE)

In a sense, the underlying reason for any analysis is to reach **data understanding**.

Studies and experiments give rise to **units**, which are typically described with **variables** (and measurements).

Variables are either **qualitative** (categorical) or **quantitative** (numerical):

- categorical variables take values (levels) from a finite set of **classes**
- numerical variables take values from a (potentially infinite) set of **quantities**

NUMERICAL SUMMARIES

In a first pass, a variable can be described along 2 dimensions: **centrality** & **spread** (**skew** and **kurtosis** are also used sometimes).

Centrality measures include:

- **median**, **mean**, **mode** (less frequently)

Spread (or **dispersion**) measures include:

- **standard deviation** (sd), **variance**, **quartiles**, **inter-quartile range** (IQR), **range** (less frequently).

The median, range and the quartiles are easily calculated from an **ordered list** of data.

VISUAL SUMMARIES – BOXPLOT

The **boxplot** is a quick way to present a graphical summary of a univariate distribution.

Draw a box along the observation axis, with endpoints at Q_1 and Q_3, and with a “belt” at the median.

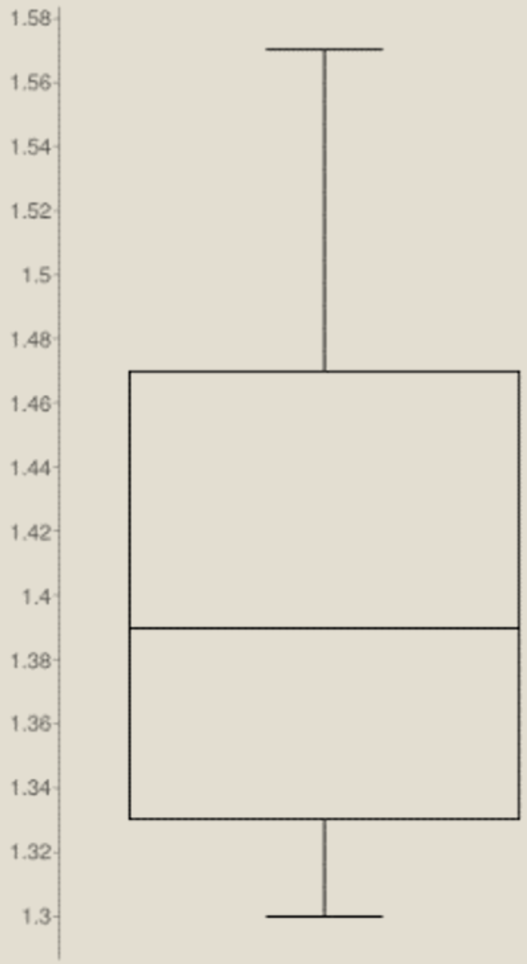
Plot a line extending from Q_1 to the smallest obs. less than $1.5 \times \mathrm{IQR}$ to the left of Q_1.

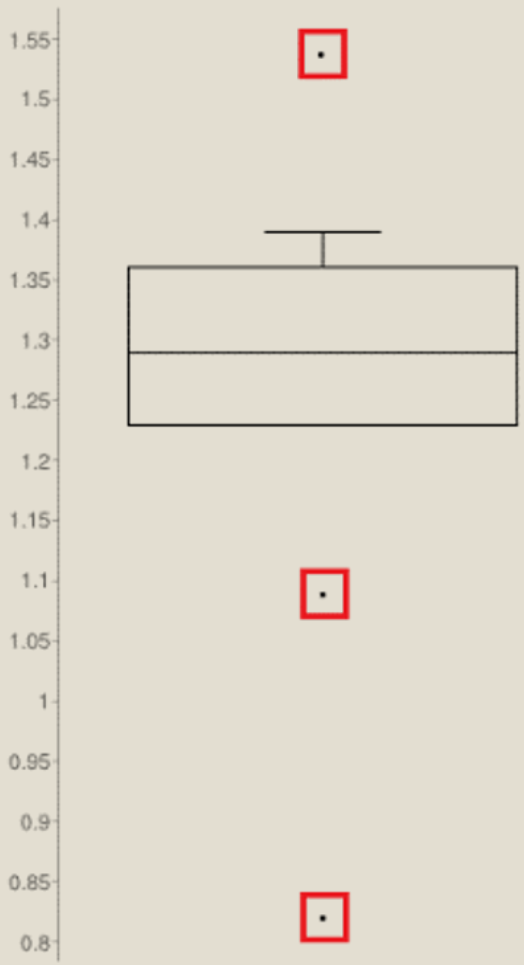
Plot a line extending from Q_3 to the smallest obs. more than $1.5 \times \mathrm{IQR}$ to the right of Q_3.

Any suspected outlier is plotted separately.

Queuing dataset: arrival rates (left), processing rates (right)

EXAMPLES





VISUAL SUMMARIES – HISTOGRAM

Histograms can also provide an indication of the distribution of a variable.

They should include/contain the following information:

- the range of the histogram is $r = Q_4 - Q_0$;
- the number of bins should approach $k = \sqrt{n}$, where n is the number of obs.;
- the bin width should approach r/k, and
- the frequency of observations in each bin should be added to the chart.

EXAMPLE

Consider the daily number of car accidents in Sydney over a 40-day period:

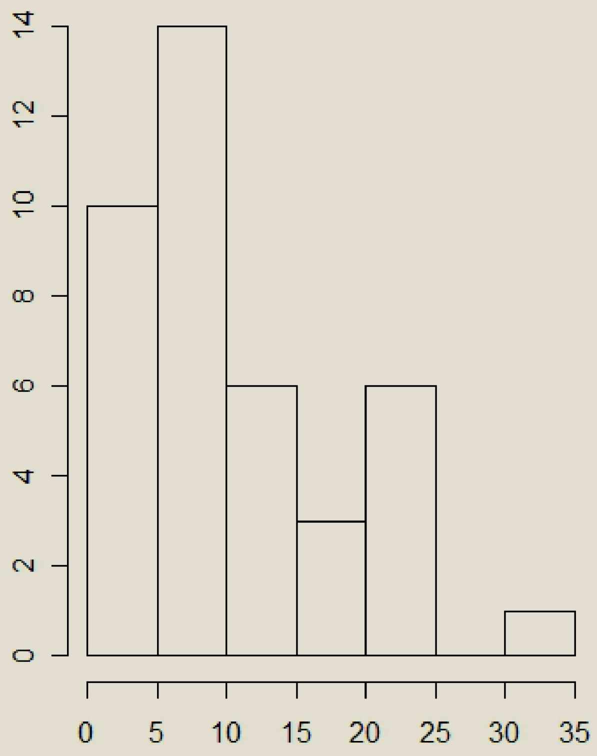
6, 3, 2, 24, 12, 3, 7, 14, 21, 9, 14, 22, 15, 2, 17, 10, 7, 7, 31, 7, 18, 6, 8, 2, 3, 2, 17, 7, 7, 21, 13, 23, 1, 11, 3, 9, 4, 9, 9, 25

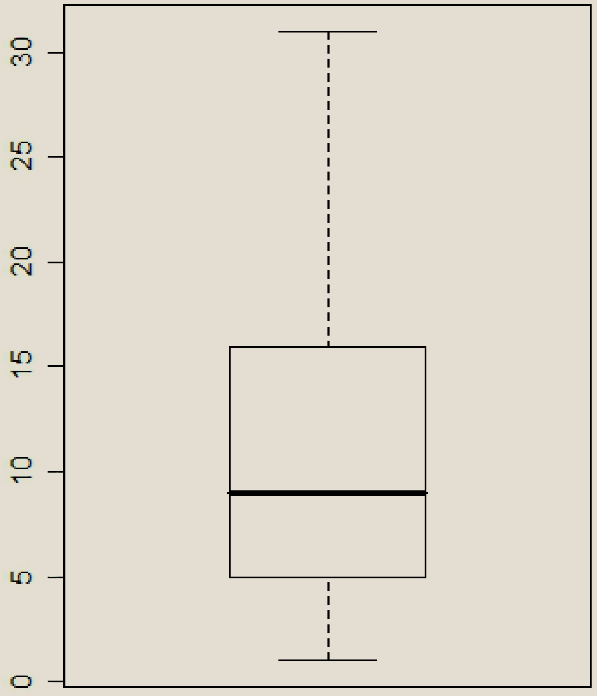
The sorted values are:

1 2 2 2 2 3 3 3 3 4 6 6 7 7 7 7 7 7 8 9 9 9 9 10 11 12 13 14 14 15 17 17 18 21 21 22 23 24 25 31

min	Q_1	med	Q_3	max
1	5.5	9	15.5	31

Is it more likely that one would see between 5-15 accidents on a given day, or between 25-35?





MOTIVATING EXAMPLE

Consider the following data, consisting of $n = 20$ paired measurements (x_i, y_i) of hydrocarbon levels (x) and pure oxygen levels (y) in fuels:

```
x:  0.99  1.02  1.15  1.29  1.46  1.36  0.87  1.23  1.55  1.40
y: 90.01 89.05 91.43 93.74 96.73 94.45 87.59 91.77 99.42 93.65

x:  1.19  1.15  0.98  1.01  1.11  1.20  1.26  1.32  1.43  0.95
y: 93.54 92.52 90.56 89.54 89.85 90.39 93.25 93.41 94.98 87.33
```

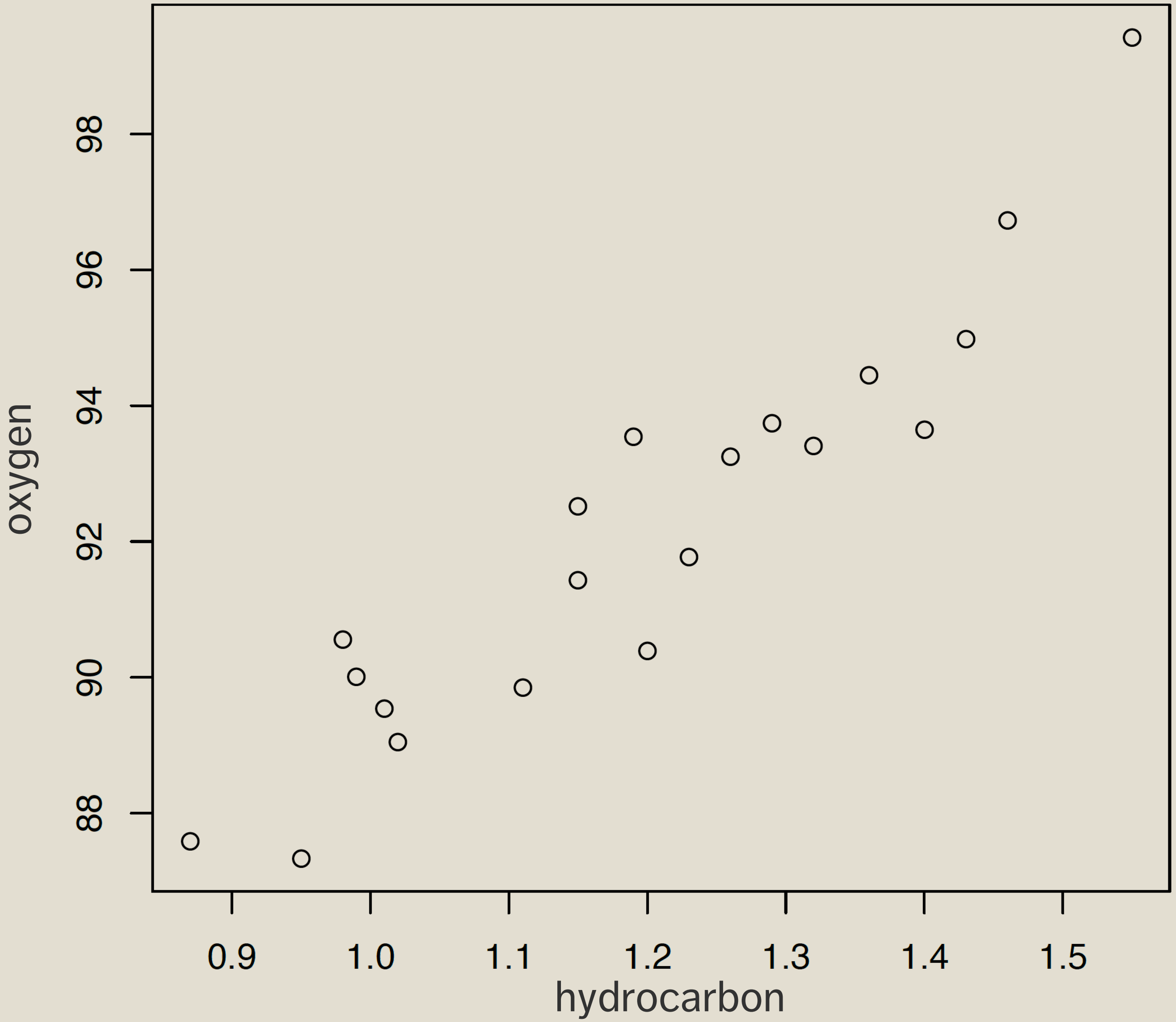
Goals:

- measure the **strength of association** between x and y
- **describe** the relationship between x and y

MOTIVATING EXAMPLE

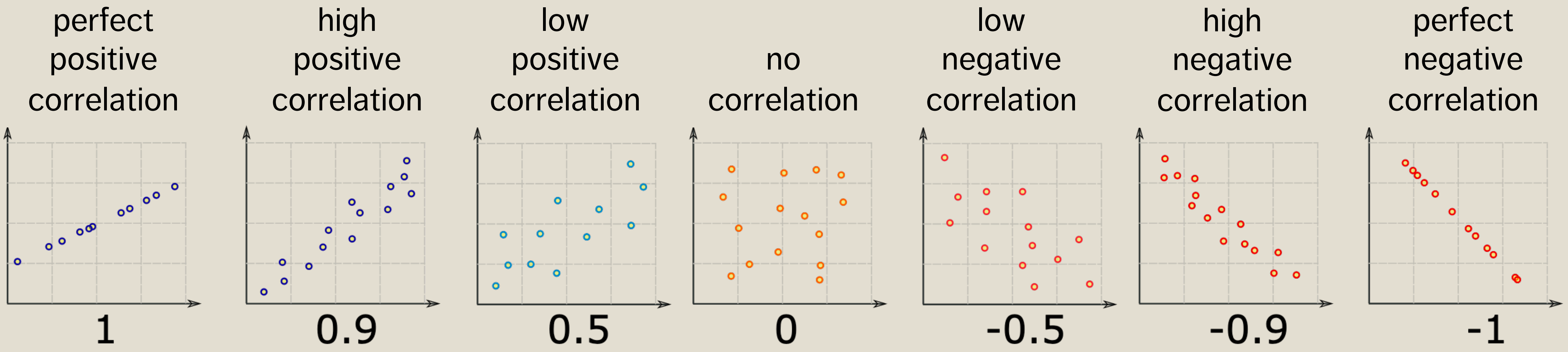
A graphical display provides an initial description of the relationship.

It seems that points lie around a hidden line!

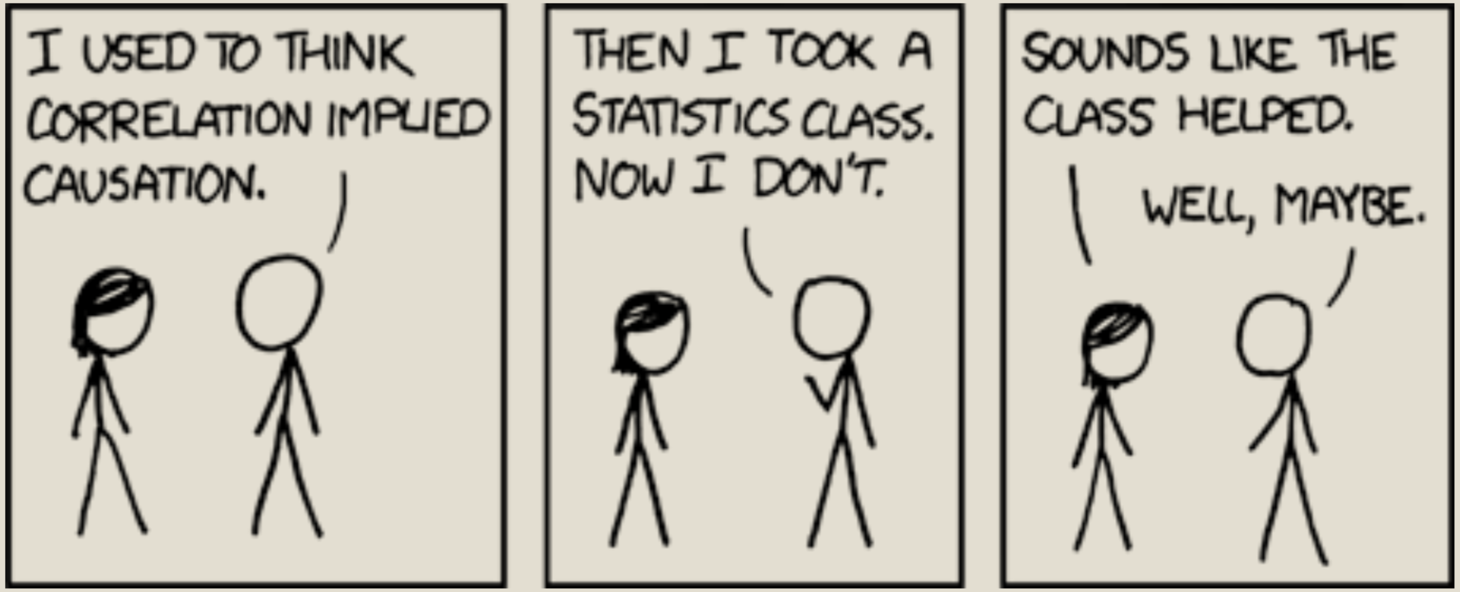


[Unknown author]

PROPERTIES AND INTERPRETATION



[https://xkcd.com/552/]



Correlation doesn't imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing 'look over there'.

LINEAR REGRESSION

If $\hat{\beta}_i$ is the estimate of the true coefficient β_i, the **linear regression** model associated with the data is

$$\widehat{Y}(\boldsymbol{x}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p = \boldsymbol{\beta x}$$