Data Science for Smart Cities

CE88

Prof: Alexei Pozdnukhov
GSI: Madeleine Sheehan

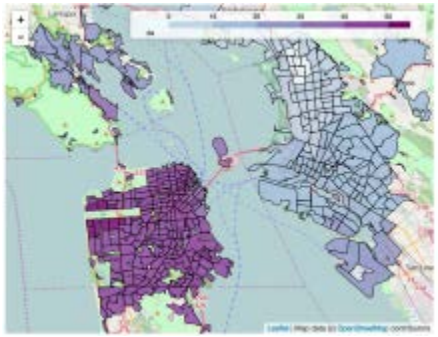
115 McLaughlin Hall

alexeip@berkeley.edu
m.sheehan@berkeley.edu
CE88 in title

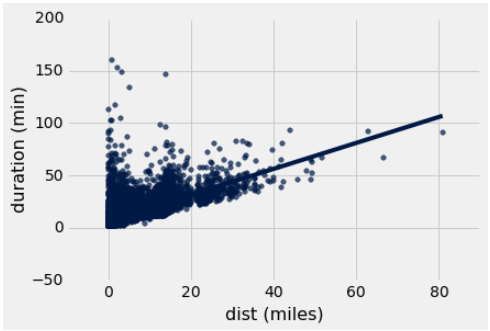
Today

9:10-9:30 **Model specification for regression analysis**

9:30-10:00 Mini Lab 6: Discuss the goals, explore the data



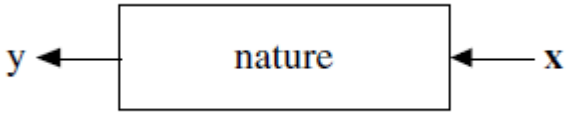
+



10:10-11:00 Mini Lab 6: exploratory data analysis that illustrates model design choices, and facilitates model specification

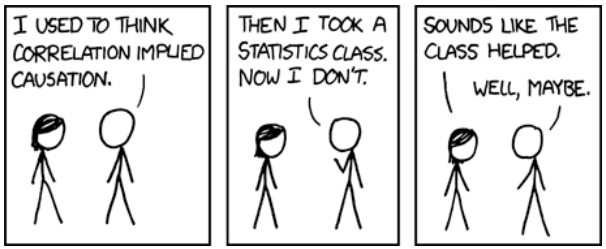
“Statistics starts with data”

Think of the data as being generated by a black box in which input variables *x* (predictor, or independent variables) go in one side, and on the other side the response variables *y* come out. Inside the black box, nature functions to associate the predictor variables with the response variables, so the picture is like this:



There are at least **two** goals in analyzing the data:

- To be able to predict what the responses are going to be to future input variables;
- To extract some information about *how* nature is associating the response variables to the input variables.



Statistical modelling

Nowhere it is set in stone how nature black box should be modelled when the only thing you got is empirical data. It is **our** choice to specify an appropriate model. How the translation from subject-matter problem to statistical model is done is often the most critical part of an analysis.

Best thing one can do is to specify a model in a way that achieves both goals we just stated.

One can assume a stochastic data model for the inside of the black box, i.e. assume that data are generated by independent draws from:

response variables = *f* (predictor variables, random noise, parameters).

Then the values of the parameters can be estimated from the data and the model then used for information (inference) and/or prediction.

Statistical modelling

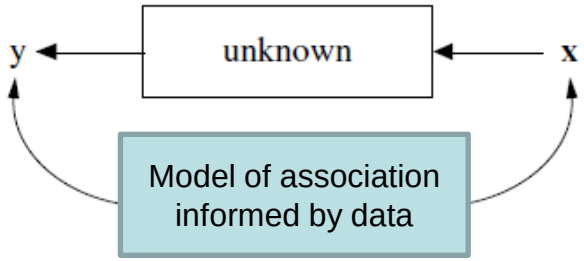
If one starts with “assume that the data are generated by the following model: … “, this implies that by imagination and by looking at the data, one can invent a reasonably good parametric class of models for a complex mechanism devised by nature.

The conclusions made with such approach could be true about the model’s mechanism, but not necessarily about the nature’s mechanism. If the model is a poor emulation of nature, the conclusions maybe wrong.

This approach is most solid when indeed the nature of the observed association between x and y can be identified, for example, when a known physical process is observed under noise (noise in the process parameters or as additive random factors to the observable).

Statistical modelling

Most often, the inside of the nature box is complex and unknown.

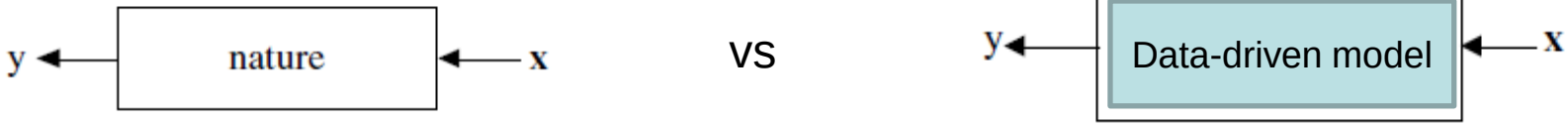


The approach to find an algorithm that operates on **x** to predict the responses y must be well thought through. Particularly, to make statistical inference about the nature of the x to y association, one has to:

- control for confounding factors
- use randomized controlled trials
- validate and test models on previously unobserved (out-of-sample) data

Analysis of models

When the model is built, we are comparing:



The great advantage of a model is if it produces a simple and understandable picture of the relationship between the input variables and responses. Examples are:

a decision rule based on thresholds,

```
if  (travel distance X > 10 miles):
            person Y travels by car
else:
            person Y travels by bike
```

or a model linear in parameters,

$$y = b_0 + \sum_{1}^{M} b_m x_m + \varepsilon$$

Such models may seem oversimplified, and not fitting the data perfectly. The choice between accuracy and interpretability is a difficult one. In a choice between accuracy and interpretability, practitioners often go for interpretability. However, a model does not have to be simple to provide reliable information about the relation between predictor and response variables.

EDA for model specification

Exploratory data analysis help to specify a model, aiming to:

- Suggest hypotheses about the causes of observed phenomena
- Support the selection of appropriate statistical tools and techniques
- Assess assumptions on which statistical inference will be based
- Provide a basis for further data collection through surveys or experiments

Minilab 6: explore a taxi fare dataset, think how to specify a model

id	departure time	arrival time	fare ($)	num	dep lon	dep lat	arr lon	arr lat	deptaz	arrtaz	dist (miles)
0	9/1/12 0:11	9/1/12 0:20	13.2	1	-122.414	37.8027	-122.421	37.7854	38	30	1.98084
1	9/1/12 0:23	9/1/12 0:31	10.65	1	-122.42	37.7861	-122.435	37.7622	30	94	2.40224
2	9/1/12 0:45	9/1/12 0:49	9	1	-122.415	37.7747	-122.408	37.7826	10	11	0.479348
3	9/1/12 0:41	9/1/12 0:54	13.95	2	-122.419	37.8066	-122.415	37.7781	40	10	2.12241
4	9/1/12 1:09	9/1/12 1:13	7.35	1	-122.43	37.7978	-122.418	37.789	45	32	1.03807
5	9/1/12 1:40	9/1/12 1:52	11.75	1	-122.433	37.7841	-122.411	37.787	77	7	0.960851
6	9/1/12 2:49	9/1/12 2:51	5.15	1	-122.409	37.7856	-122.412	37.791	7	29	0.414315
7	9/1/12 3:29	9/1/12 3:47	43.65	1	-122.403	37.7927	-122.386	37.6181	2	239	14.5105
8	9/1/12 0:33	9/1/12 0:57	46.75	1	-122.387	37.6174	-122.407	37.7889	239	5	14.2747
9	9/1/12 4:39	9/1/12 4:43	6.25	1	-122.422	37.7977	-122.418	37.789	34	31	0.685312