Models and misunderstanding:

statistics, data science, and big data in the modern world

David J. Hand

Imperial College, London

and

Winton Capital Management

Underlying motivation for this talk

Proximate motivation for this talk

On the nature of statistical models

Two types of model - and why it matters

Underlying motivation for this talk

The world – the media, politicians – catching up on the importance of data and objective modelling

Many recent high profile success stories

- customer understanding
- managing passenger flow through London
- optimise call centre processes
- reduce impact of adverse weather events
- genomics
- Higgs boson
- .....

And some failure stories

- Google flu trends
- Company X – no clear objective
- Gaming and feedback, Crimemaps?
- ....

I predict that half of all big data projects will fail to deliver against their expectations.

Bernard Marr

http://www.forbes.com/sites/bernardmarr/2015/03/17/where-big-data-projects-fail/

The important thing is to learn from the mistakes

Lack of appreciation of the central importance of statistics in data science?

Tension between statisticians and ‘data scientists’ without a statistical background ?

All analysis of data carries assumptions

If your assumptions are wrong, your conclusions may be faulty

‘Assumptions’ = ‘model’

The classical perspective on statistical models

(S,P)
S – the sample space
P – a set of probability distributions on S, which contains some distribution which closely approximates the “true” distribution

Models are often parameterised by indexing P by a parameter set Θ
(A Bayesian model also requires a prior distribution on Θ)

Elaborate mathematical extensions of what a model is

e.g. McCullagh, 2002

Extensions to nonparametric models

the number and nature of parameters not fixed in advance, but determined by the data

e.g. kernel regression

e.g. ensemble models

The important question is

what is the model for?

Main uses:

- to make inferences: about unobserved cases
  - to the future
  - to other cases drawn from the same population
- to make inferences: about underlying mechanisms
  - about the overall characteristics of the underlying population
  - causal inferences
- summaries, for ease of comprehension
  - about the overall characteristics of a collection of data

Clearly the mathematical definition above is ***not adequate for all purposes***

For example, sometimes, instead of ***a sample*** from a population, we have ***the entire population***

e.g. data on all the countries in the world

Does this mean we cannot build statistical models?

That statistics, machine learning, data mining, etc, are irrelevant?

Proximate motivation for this talk

All models are wrong, but some are useful

GeorgeBox

Today companies like Google, which have grown up in an era of massively abundant data, don't have to settle for wrong models. ***Indeed, they don't have to settle for models at all***

Chris Anderson

Need to distinguish between

- substantive models

- empirical models

I claim that the distinction matters,
and should be taught to students

Substantive: theory-driven, mechanistic, ...

Based on some (simplified representation of a) theory about the underlying mechanism.

Use statistical methods to determine precise form

- estimate parameters
- fit model to data
- ....

Models as *approximations to something underlying*
- that is, to a "truth"

Models as *approximations to something underlying - that is, to a "truth"*

All models are wrong, but some are useful. GeorgeBox

Models as *approximations to something underlying - that is, to a "truth"*

All models are wrong, but some are useful. GeorgeBox

The normal model:

The unicorn, the normal curve, and other improbable creatures
Theodore Micceri, 1989

Normality is a myth; there never was, and never will be, a normal distribution Geary, 1947

Goes on to say "*This is an over-statement from the practical point of view*"

Empirical: data driven, descriptive, ...

Aim to summarise, identify, extract the relationships in data

No ***substantive*** theoretical base
But will have a ***statistical, mathematical*** theoretical base

Find a mathematical function/structure which fits the data

Search through variables, arbitrary transformations, optimisation, ...

Overfitting issues; regularisation; ...

Commonly used for *prediction*

Parametric and nonparametric methods

- kernel regression
- nearest neighbour
- ANNs
- classification trees
- random forests
- ensemble methods
- deep learning
- ....

Example: ***Substantive***

Model the relationship between the height from which a stone is dropped and the time it takes to hit the ground

e.g. drop stones from various heights

→ Data $\left(H_i, t_i\right)\ i = 1, \ldots, n$

→ Model From *Newton's Laws*

$$t = \sqrt{2H/a} + \varepsilon \qquad H = a\left(t + \varepsilon\right)^2 / 2$$

Use statistical methods to estimate *a*

Example: ***Empirical***

Logistic regression model for the probability that someone will purchase a product

based on the values of their characteristics $x_1, x_2, \ldots, x_d$

$$\log\left(\frac{p}{1-p}\right) = \sum_{j=1}^{d} \beta_j x_j$$

Models may start as empirical and then become substantive

This follows the classic picture of science in which

- we begin by collecting data and noticing regularities
- and then formulate theories which encapsulate those regularities

Some history of predictive empirical models

Judgemental approaches:

- in medical diagnosis
- in credit granting decisions
- in personnel selection
- in customer relationships
- ....

The example of retail credit decisions

Judgemental:

- personal knowledge
- standing in the community
- subjective characteristics

Empirical statistical models (scorecards)

- objective
- speed – number of cases
- replicability
- tirelessness
- incremental improvement
- not affected by prior probability failures; overconfidence

or by the host of other behavioural characteristics described by Daniel Kahneman, Gerd Gigerenzer, Dan Ariely, etc

But doubt in the early days that statistical methods could perform as well as human judgement

Parallelling the sudden ‘discovery’ of the power of data by the mass media – even though you have long known

→ experimental comparisons of judgemental vs statistical in several domains

Credit scoring

Myers and Forgy, 1963, reviewing comparisons made between 1941 and 1960

> *While results from these studies have differed, all have shown that a properly constructed numerical rating system can offer at least some degree of improvement over the purely subjective or judgmental approach to evaluating credit.*

Clinical psychology

Meehl, 1954, reviewed 20 comparisons

> and concluded that 19 of the 20 demonstrated superior or equal performance by the statistical methods

Psychology and medicine

Grove et al, 2000, meta-analysis of 136 studies

> *on average, mechanical prediction techniques were about 10% more accurate than clinical predictions*

Paul Meehl's simile:

when you check out at a supermarket, you don't eyeball the heap of purchases and say to the clerk, "well, it looks to me as if it's about $17 worth; what do you think?" The clerk adds it up

Paul Meehl's simile:

when you check out at a supermarket, you don't eyeball the heap of purchases and say to the clerk, "well, it looks to me as if it's about $17 worth; what do you think?" The clerk adds it up

A cautionary comment: I suspect that the individual studies are subject to considerable publication bias

A comment on interpretability

Interpretability is reassuring:

the model "makes sense", "is reasonable", etc

Good for convincing people

- customers
- managers
- people who have to use the model

Substantive models, by definition, have an interpretation

But: post hoc narratives are easy to construct

Simple empirical models are also easy to report

In many situations interpretability is a ***driver*** behind such models

e.g. weighted sums of predictors

But what do you mean by 'interpretability'?

In many contexts: *how important is variable v_i to the model?*

For elaborate models (e.g. ANNs, RFs, etc) compare the predictive power of the complete model with the model without v_i

Horses for courses?

Management: mainly empirical models
(incl. government)

Natural sciences: mainly substantive models

Medicine: both

Engineering: both, often substantive simplified to empirical heuristic

Not a sharp distinction

e.g. Regression

Empirical: aim for a simply interpretable predictive model

??

Not a sharp distinction

e.g. Regression

Empirical: aim for a simply interpretable predictive model

??

"the increase in *y* for unit increase in *x*, holding *z* constant"

Not a sharp distinction

e.g. Regression

Empirical: aim for a simply interpretable predictive model

??

"the increase in *y* for unit increase in *x*, holding *z* constant"

but this is not always meaningful: $y = \alpha + \beta x + \gamma x^2 + e$

Not a sharp distinction

e.g. Regression

Empirical: aim for a simply interpretable predictive model

??

"the increase in *y* for unit increase in *x*, holding *z* constant"

but this is not always meaningful: $y = \alpha + \beta x + \gamma x^2 + e$

How y responds, on average, to change in x1, after allowing for simultaneous linear change in x2, for the data at hand

Terence Speed

Substantive:

- because you think it is linear
- or because you regard it as the first term in a Taylor series

"Treatment A is better than treatment B"

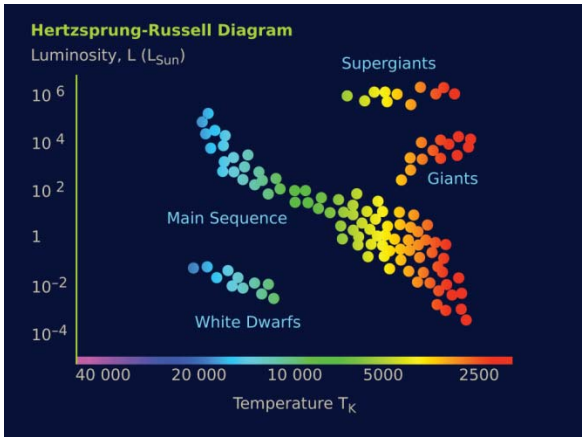
Could be a simple model, or a mere empirical description

e.g. Cluster analysis

Dissection vs partitioning

Dissection = substantive

- carving nature at the joints
- finding natural clusters
    - e.g. bipolar vs unipolar depression
    - e.g. Herzsprung-Russell diagram for stars



Partitioning = empirical

- finding a convenient way to split the data

e.g. 1: in an advertisement in the *Sunday Times* of 18th April 1999, James Meade Limited, a shirt manufacturer, gave a choice of sizes, as follows, where a ✓ denotes sizes available, and ✓ denotes standard size.

	Sleeve length (inches)						
	31	32	33	34	35	36	37
Collar size							
14½	✓	✓	✓	✓	✓	✓	✓
15	✓	✓	✓	✓	✓	✓	✓
15½	✓	✓	✓	✓	✓	✓	✓
16	✓	✓	✓	✓	✓	✓	✓
16½	✓	✓	✓	✓	✓	✓	✓
17	✓	✓	✓	✓	✓	✓	✓
17½	✓	✓	✓	✓	✓	✓	✓
18	✓	✓	✓	✓	✓	✓	✓

Weaknesses of substantive models

- *need a theory*
  - *OK in theory-rich domains*
  - *less so in others*
- *model (might be) OK if the theory is right*
- *but can be misleading if theory is wrong*
  - *and it's systematic bias, not random variation*

Weaknesses of empirical models

Assume:

- *the future is like the past*
- *choose criterion to fit model to data*
- *good quality data*
- *no selection bias*
- *no gaming, feedback, etc*
- *...*

The future is like the past

'In times of change, learners inherit the Earth, while the learned find themselves beautifully equipped to deal with a world that no longer exists'

Eric Hoffer

Some situations can reasonably assume stationarity

- the laws of physics do not change over time

But others are intrinsically non-stationary

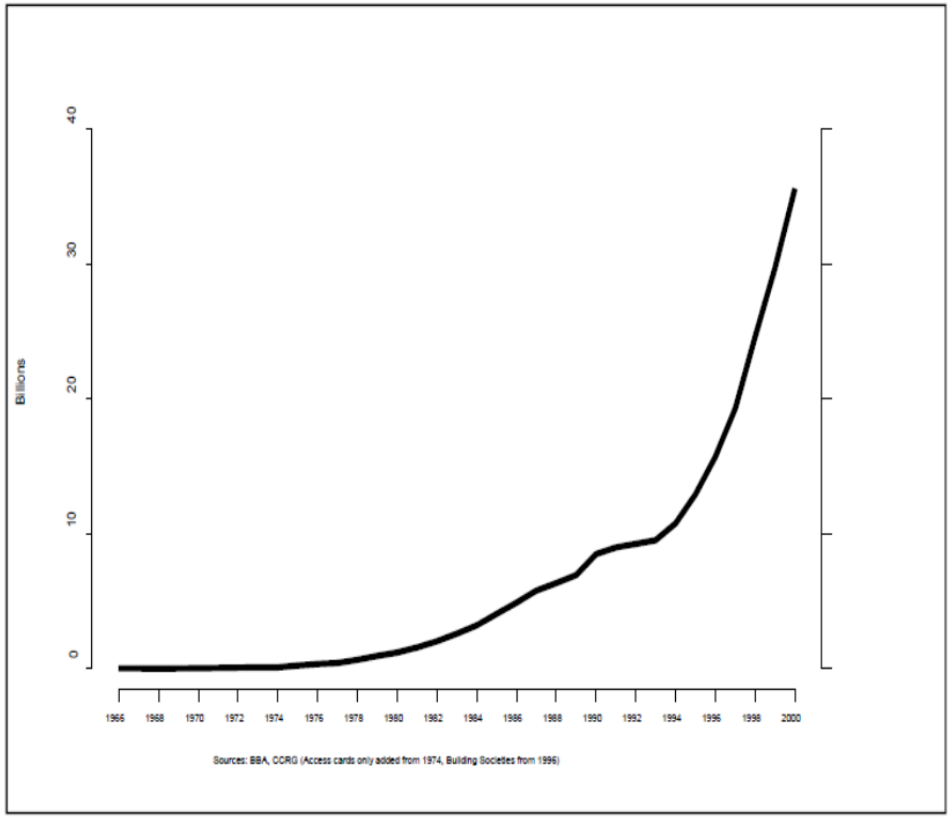
- human behaviour

And some involve feedback and gaming

- human behaviour (e.g. modelling financial markets)

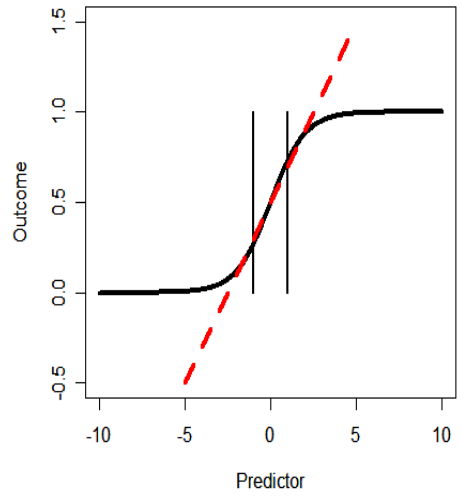
Failure to model nonstationarity

e.g. empirical economic models built in early 2000s



Failure to model range of variation:

e.g. *cosmic variance*

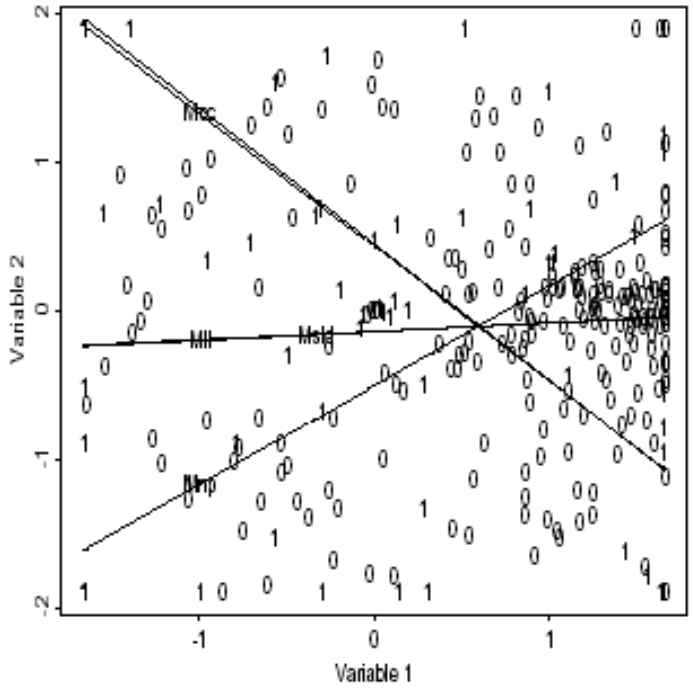


→ Predictive models may break down

→ Regression to the mean because underestimating variation

Substantive models are less vulnerable to this

Choose criterion to fit model to data



Ionosphere data, Benton (2001)

Optimum error rate: top-left to bottom-right

Optimum Gini: bottom-left to top right

Substantive models have *meaning*

Example: Lindsey model for mean number of reported AIDS cases

$$\log\left(\mu_{ij}\right) = \phi + \xi_1 \log\left(t\right) + \xi_2 / t + \theta_1 \log\left(u\right) + \theta_2 / u + v_1 t/u + v_2 t \log\left(u\right)$$
$$+ v_3 \log\left(t\right)/u + v_4 \log\left(t\right)\log\left(u\right)$$

u is reporting delay time $\Rightarrow$ ratio scale

$\Rightarrow$ can express in different units by rescaling

$$v_2 t \log\left(u\right) \rightarrow v_2 t \log\left(ku\right)$$
$$= v_2 t \log\left(u\right) + v_2 t \log\left(k\right)$$
$$= v_2 t \log\left(u\right) + v_5 t$$

So an extra term, $v_5 t$, is introduced into the model

Purely on measurement grounds, Lindsey's model *must be wrong*

Lindsey's time units (quarter years):

- Incorrect model: penalised deviance = 800.8
- Corrected model: penalised deviance = 803.8

Change time units to days:

- Incorrect model: penalised deviance = 814.1
- Corrected model: penalised deviance = 803.8

Changing the units in which time is measured changes the goodness of fit of Lindsey's model!

All models may be wrong, but they must be wrong in the right way to be useful

Example 1: Wrong in the right way

A multiple regression model is probably wrong (a linear relationship? Really?) but may be very effective

Example 2: Wrong in the wrong way

Preis T, Moat H.S., and Stanley H.E. (2013) Quantifying trading behavior in financial markets using Google trends. *Scientific Reports*, 3, 1684.

‘*By analyzing changes in Google query volumes for search terms related to finance, we find patterns that may be interpreted as ‘‘early warning signs’’ of stock market moves.*’

Chose 98 search terms related to stock markets

Implemented a strategy which

- sold DJIA following an increase in searches
- and bought following a decrease in searches

Concluded:

"*We find that returns from the Google Trends strategies are significantly higher overall than returns from the random strategies (<R>=0.60; t=8.65, df=97, p<0.001, one sample t-test).*"

But this is an inappropriate test

Telling us nothing about the effectiveness of the strategy

1) the 98 terms were *purposively* sampled, not a random sample from a population of possible terms: the authors confused *fixed* and *random* effects;

2) the standard deviation used in the denominator of their *t*-test is irrelevant to the variability of the mean return: it's the wrong statistic;

3) they ignored the correlation between the random aspect of the different words' returns: anyone working in the financial sector will be aware of the critical importance of correlation!

In summary

Broadly speaking

Substantive models if aim is to understand

Empirical models if aim is prediction

But there may be other constraints:

- legal constraints on choice of predictors [e.g. anti-discrimination]
- legal constraints on complexity (interpretability)
- accuracy vs transparency
- need for your boss to understand
- good enough vs best

Other taxonomies of model types

Cox 1990:

- substantive models
- empirical models
- indirect models

Dempster 1998

- empirical models
- stochastic models
- predictive models

Shmueli 2010

- predictive vs explanatory models

Gifi 1981/1990

there is, according to many statisticians, one appropriate way to analyse data: first formulate a model on the basis of prior knowledge, then compute the likelihood of the data given the model by maximizing the likelihood function.... [In contrast] the prescription seemingly advocated by Gifi. First choose a technique... on the basis of the format of your data, then apply this technique and study the output

Breiman 2001

- data models

  Assume (***tentatively entertain***) a particular functional form relating input data to response variables, via parameters

- algorithmic models

  Find some function which predicts response from input.

Conclusion

Chris Anderson's comments

Today companies like Google, which have grown up in an era of massively abundant data, don't have to settle for wrong models. ***Indeed, they don't have to settle for models at all.***

apply only for empirical models

Conclusion

Chris Anderson's comments

Today companies like Google, which have grown up in an era of massively abundant data, don't have to settle for wrong models. ***Indeed, they don't have to settle for models at all.***

apply only for empirical models

- but empirical models are only half the story

Conclusion

Chris Anderson's comments

Today companies like Google, which have grown up in an era of massively abundant data, don't have to settle for wrong models. ***Indeed, they don't have to settle for models at all.***

apply only for empirical models

- but empirical models are only half the story

- a story entirely ignoring substantive models

thank you