Statistical Modeling: The Two Cultures

Leo Breiman

Leda Liang, Yash Nair, Zitong Yang

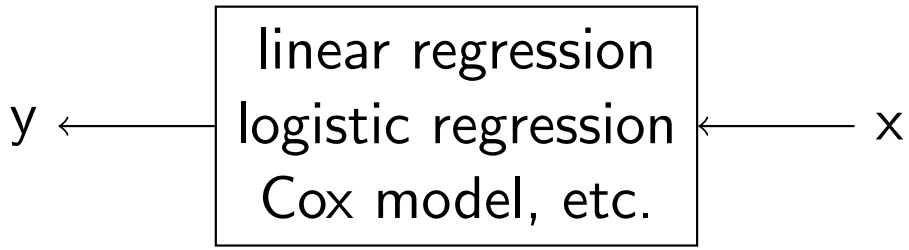
Two Goals in Statistics

- Prediction – Predict for future inputs
- Information – Learn about the underlying nature of the process

Two cultures:

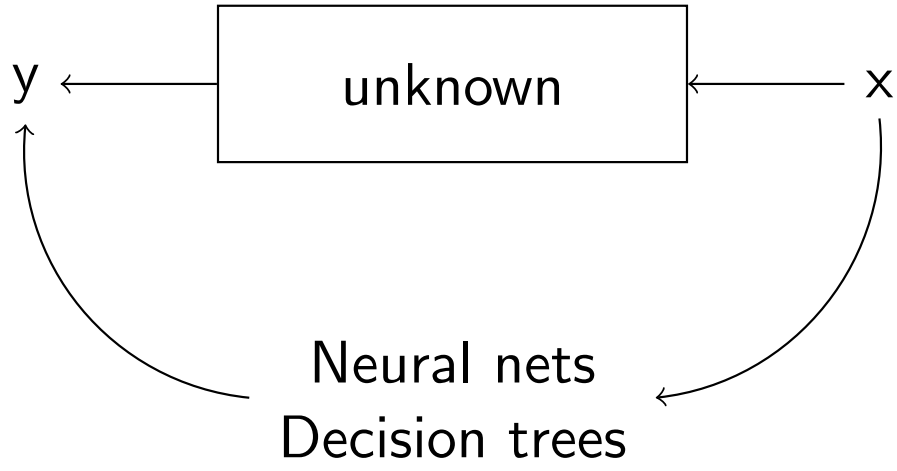
- Data modeling
- Algorithmic modeling

Data Modeling



- 98% of statisticians
- Data is generated by some model + some noise
- Estimate parameters of the model
- Validation is yes/no decision based on goodness of fit and residual examination

Algorithmic Modeling



- 2% of statisticians
- Nature is complex and data does not come from a model that can be simply described.
- Use any algorithm to predict y
- Evaluate by measuring predictive accuracy

"Modern" Problems[1]

- The size of data we collect keeps growing.
- The data modeling assumptions are more restrictive than the algorithmic modeling approach.

Examples

- Predicting ozone levels
- Determining chemical toxicity

[1] This paper was published in 2001

The Ozone Project

The goal is to predict ozone levels to warn the public of days with hazardous air quality conditions in Los Angeles. Ideally would like to reduce driving and time spent outside on such days.

Data consists of:

- Measurements from dozens of weather stations and different layers of the atmosphere
- Hourly readings of 450 variables including temperature, humidity, and wind speed
- 7 years of historical data

Large linear regression models were used to make predictions but gave too many false alarms

The Chlorine Project

The chemical structure of chemical compounds can tell us about its toxicity. Mass spectra can be cheaply obtained, but analyzing mass spectra manually is expensive. The goal is to predict the toxicity from mass spectra.

Data:

- 30,000 compounds with known chemical structure and mass spectra
- Mass spectra consists of the frequencies at each molecular weight

Dimension was too large for linear regression, but decision trees had 95% accuracy

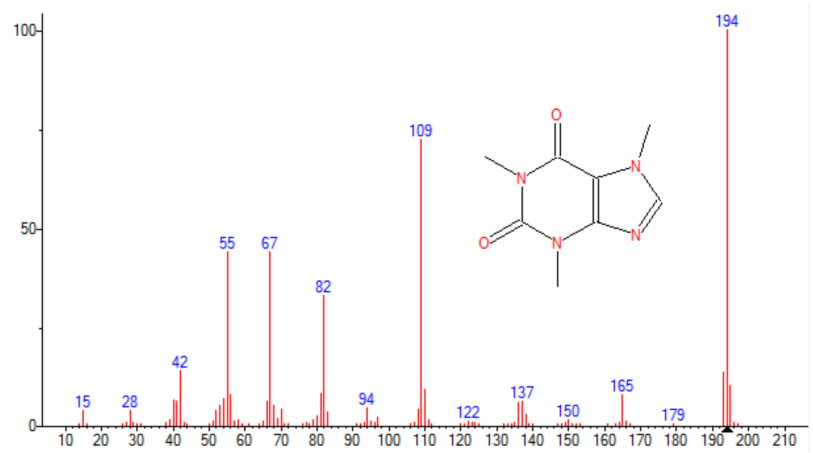


Figure: Example of Mass spectra[a]

[a] https://www.nist.gov/image/mass-spectra-caffeine

Brieman's Perceptions

Some realizations after working on consulting projects:

- Data modeling is motivated by an academic setting
- Algorithmic modeling is more useful for consulting and practical settings
- Search for the model that gives the best solution – either algorithmic or data
- Using restrictive models can prevent statisticians from working on exciting new problems
- Incorrect modeling assumptions can lead to questionable scientific discoveries

Data modeling: pro and cons

Example: Linear regression model

$$y = b_0 + \sum_{m=1}^{M} b_m x_m + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

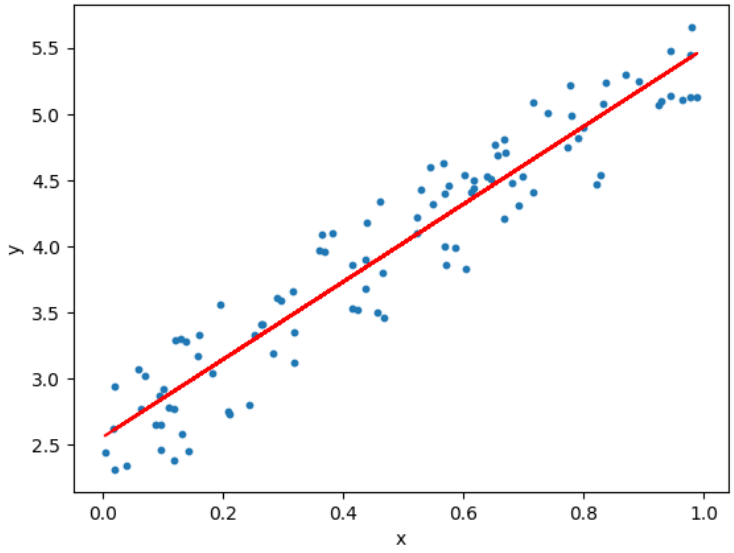


Figure: Example of linear regression

Pros:

- Nice theory
- Simple/elegant hypothesis tests/CIs

Cons:

- Goodness-of-fit evaluated by R^2

Overall con: "The conclusions are about the model's mechanism, not about nature's mechanism"

Gender discrimination case study

A study on the statistics department of a university was done to assess whether there was gender discrimination in the salaries of the faculty.

Assume LM as data model:

- $y =$ salary
- $x_1, \ldots, x_{24}$: measures of academic performance
- $x_{25} =$ gender

Questions:

- Is data adequate for answering the question?
- Does model accurately describe the data?

Detecting if data model is applicable: Problems

Goodness-of-fit test:

- Test if data follow data model (null hypothesis) or not (alternative)
- Problem: lack of power unless direction of alternative prespecified

Residual Analysis:

- Cannot detect lack of fit if #covariates > 5 — William Cleveland

Few application papers in JASA even bother to discuss/analyze model fit

Multiplicity of data models

1. Suppose two different statisticians use different models for the same problem:
   - Both run goodness of fit tests that fail to reject
   - Each concludes his/her model fits data...draws different conclusions
   - Who is right? Who is wrong?
2. Example:
   - Cox model in medical journals
   - Different (well-fitting) models could yield different answers

Predictive accuracy as a measure of model fit

- Use a "black-box" model *only* to make predictions
- In particular: *do not assume that data actually follows the "black-box" model*
- Compare closeness of model's output to nature's output:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- Eliminating bias:
  - Held-out test set
  - Cross-validation

Limitations of data models

- "If all a man has is a hammer, then every problem looks like a nail."
- More complex data $\implies$ More complex data models
- Insistence on using data models limits using other tools (e.g., algorithmic models)

Algorithmic modeling

- Development in mid-80s, characterized by powerful algorithms such as random forest and neural networks.
- Exciting new research community: young computer scientist, physicists and engineers, a few aging statisticians...
- Different venue of publication: *Neural Information Processing* and *Journal of Machine Learning Research*.

Theoretical aspect of algorithmic modeling

- Supervised algorithmic modeling is a statistical phenomena!
- Low training error + more data than "degrees of freedom" = low test error
- Why?

$$\Pr_{S\sim D^{|S|}}\left[|\text{Test}_D(f) - \text{Train}_S(f)| \leq \sqrt{\frac{\log|F| + \log\frac{1}{\delta}}{|S|}} \text{ for all } f \in F\right] > 1-\delta$$

Theoretical aspect of algorithmic modeling

Low training error + more data than "degrees of freedom" = low test error

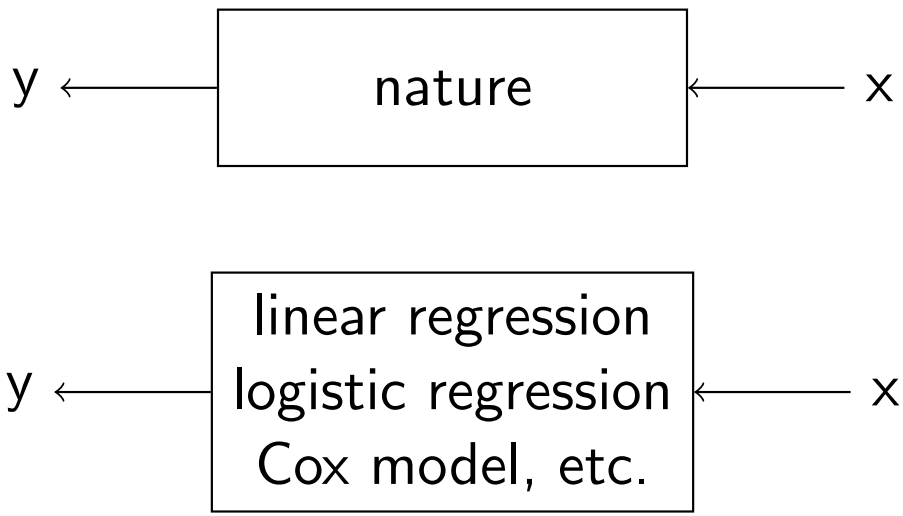
$$
\begin{aligned}
&\Pr[\text{Test}_D(f) - \text{Train}_S(f) > t \text{ for some } f \in F] \\
&= \Pr\left[\bigcup_{f \in F} \{\text{Test}_D(f) - \text{Train}_S(f) > t\}\right] \\
&\leq \sum_{f \in F} \Pr[\text{Test}_D(f) - \text{Train}_S(f) > t] \\
&\leq |F| \exp(-|S| t^2)
\end{aligned}
$$

There was some discussion around the failure of this for $p > n$ scenario in deep learing, but can be fixed.

Three lessons from algorithmic modeling

- Rashomon: The multiplicity of good models;
- Occam: the conflict between simplicity and accuracy
- Bellman: dimensionality – curse or blessing?

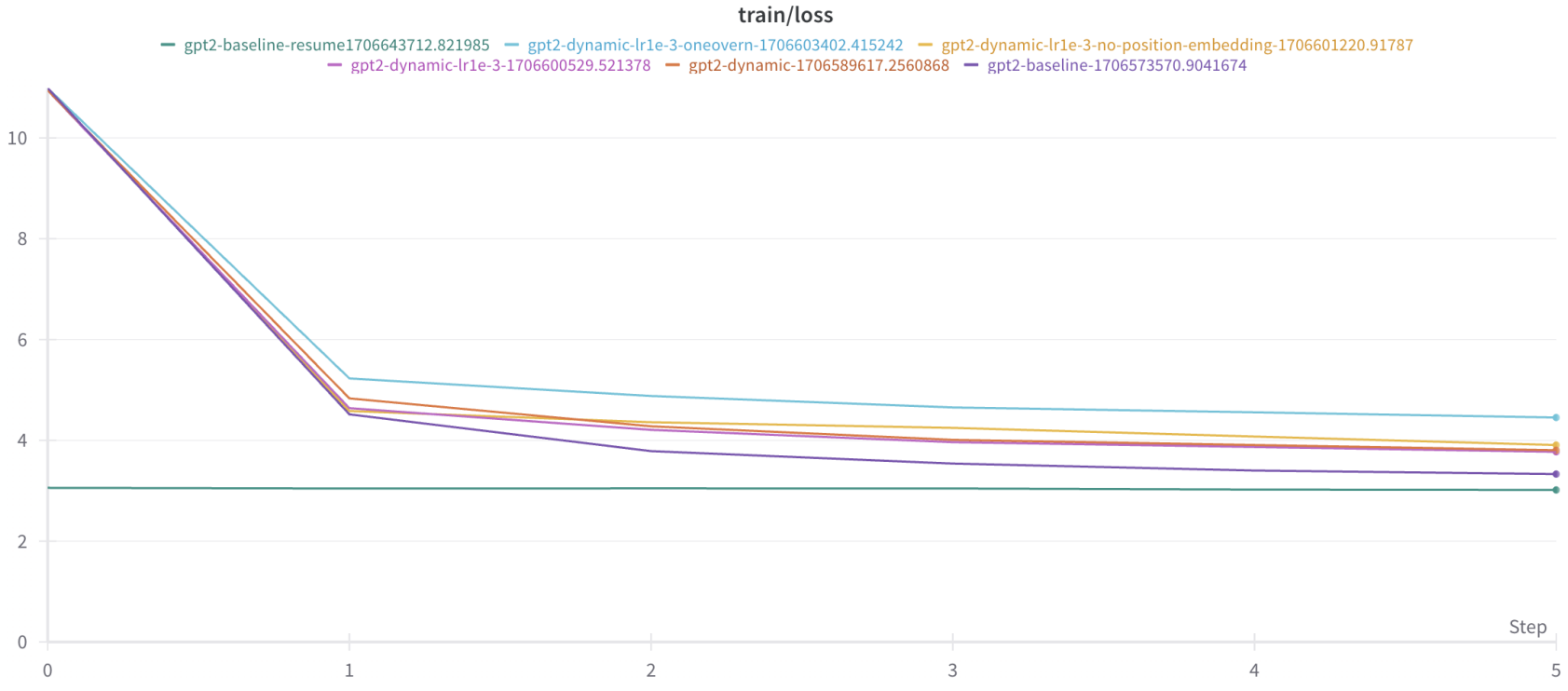
Is simplicity the right question to ask?



Statistical modeling: a third culture

- Algorithmic model: only care about predictive accuracy $\text{Test}_D(f)$ given that we trained $\text{Train}_S(f)$.
- Even this perspective is not enough in the modern domain

For example, in language modeling, we have virtually infinite data and can generate synthetic data. So $|S|$ is very large, and the gap between $\text{Test}_D(f)$ and $\text{Train}_S(f)$ is small.



Statistical modeling: a third culture

However, we don't care about $\text{Test}_D(f)$, which merely depicts how closely f mimics the distribution of language. Instead we care about

- The capability of our f – reasoning, planning, problem-solving.
- For example, the probability that the model f can solve Riemann hypothesis.

Average ⬆️	ARC	HellaSwag	MMLU	TruthfulQA	Winogrande	GSM8K	#Params (B)
80.48	76.02	89.27	77.15	76.67	85.08	78.7	72.29
80.48	76.02	89.27	77.15	76.67	85.08	78.7	72.29
78.55	70.82	85.96	77.13	74.71	84.06	78.62	72.29
77.91	74.06	86.74	76.65	72.24	83.35	74.45	60.81
77.44	74.91	89.3	64.67	78.02	88.24	69.52	12.88
77.38	73.72	86.46	76.72	71.01	83.35	73.01	60.81
77.29	74.23	86.76	76.66	70.22	83.66	72.18	34.39
77.29	70.14	86.03	77.4	69	84.37	76.8	72.29
77.28	72.87	86.52	76.96	73.28	83.19	70.89	60.81
77.1	74.32	89.5	64.47	78.66	88.08	67.55	12.88

Statistical modeling: a third culture

- The third culture deviates from the second culture in that we eventually care about the capability of our f.
- However predictive accuracy is the only thing we can optimize for, which is different from the capability of f.

Consider a simple thought experiment:

- The latest GPT4 was trained on Nov. 16th, 2023.
- If we feed the some world event that happens post that date to the model, the model will learn that knowledge and be able to reason about it.
- But the distribution of natural langugage before and after Nov. 16th are not different.
- The capability of f is no longer a pure statistical phenomena, like supervised learning.
- Some new ideas are here needed!