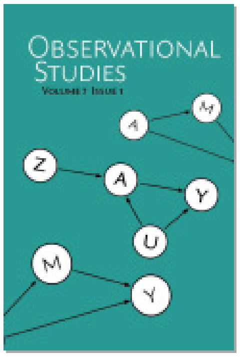
Statistical Modeling: Returning to its Roots

Qingyuan Zhao



Observational Studies, Volume 7, Issue 1, 2021, pp. 229-234 (Article)

Published by University of Pennsylvania Press
DOI: https://doi.org/10.1353/obs.2021.0014

Statistical Modeling: Returning to its roots

Qingyuan Zhao **qyzhao@statslab.cam.ac.uk**
Statistical Laboratory
University of Cambridge
Centre for Mathematical Sciences, Wilberforce Road
Cambridge, CB3 0WB, United Kingdom

Abstract

Leo Breiman's "Statistical Modeling: The Two Cultures" is a treasure for any statistician who engages with real world problem. I argue that there is a more fundamental dichotomy between the principles of statistical modeling and the techniques for statistical modeling. Focusing entirely on the techniques in statistical education and research can be dangerous. I join Breiman's call for statistics to return to its roots.

Keywords: Leo Breiman; Causal modeling; Philosophy of statistics.

How my views have evolved

Leo Breiman's "Statistical Modeling: The Two Cultures" is still an incredibly inspiring article 20 years after it was first published. What I most appreciate is Breiman's detailed and frank recount of how his opinions on statistics changed over his career. It is a treasure for young statisticians like me who started their careers not too long ago.

I have read Breiman's article three times. The first time was during the early years of my Ph.D., after seeing an emotional tribute to Breiman by Jerome Friedman in a workshop in 2013. He recounted the difficulties that he and Breiman had during the 1980s to promote machine learning (or the "algorithm modeling" culture, following the terms in Breiman's article) in statistics and mentioned how delighted Breiman would be to see the explosion of machine learning, had he not passed away in 2005. However, I didn't gain much from my first reading of Breiman's article. Most of the ideas in Breiman's article seemed very natural to me, because of the courses I had just taken and seminar talks I had heard in the Statistics Department at Stanford, a department heavily influenced by Breiman's ideas.

I read Breiman's article again when I finished my Ph.D. in 2016. By then I had learned more about causal inference and was intrigued by how the ideas there are so different. Causal inference was a niche research area (if not still a niche area now); at least very few people around me knew much about it. I was desperately looking for connections between causal inference and the "mainstream" statistics (whatever I thought it was), but I couldn't find many beyond a coincidence that Friedman's partial dependence plot for interpreting black-box algorithms is exactly the same as the confounder adjustment (or backdoor adjustment) formula for causal identification (Zhao and Hastie, 2021). If I may follow Breiman's terminology and call this much smaller area the "causal modeling" culture, then my estimated population of "causal modeling" in 2016 would be the same as Breiman's estimate for "algorithmic modeling" in 2001: "2% of statisticians, many in other fields".

Because I had seen many more bad examples of "data modeling", this time I could appreciate Breiman's points much better.

The last time I read Breiman's article was, of course, while I was preparing to write this contribution to *Observational Studies.* I have learned much more about causal inference, and I just finished teaching statistical modeling to the undergraduate students and causal inference to the postgraduate students at Cambridge. These experiences helped me to look at Breiman's argument more critically. Although I still agree with many points in the article (in particular, Breiman's call for statistics to "return to its roots"), I also start to feel unsure about some other points. Below I will write about the things I agree or disagree with Breiman.

Techniques versus Principles

Breiman's dichotomy of "data modeling" versus "algorithmic modeling" is a powerful one and is still widely present in today's textbooks and literature. Although people's views are almost always continuously distributed, I can understand and sympathize with why Breiman decided to create a sharp contrast between two extremes. However, as the importance of machine learning becomes more and more well recognized in statistics, it is no longer helpful to think about statistical modeling as "data models" versus "algorithmic models". I think the boundary between these two cultures is in fact very blurry now, if not nonexistent.

A more fundamental dichotomy, I think, is between the *principles* of statistical modeling and the *techniques* for statistical modeling. Chinese call the former *Dao* (the "way", "doctrine", or "holistic beliefs") and the latter *Shu* (the "technique", "skill", or "method"). It is difficult to describe the philosophical differences precisely, but here is roughly what I have in mind:

1. Techniques or *Shu*: The analysis in this mindset starts with some datasets curated by others. The goal is to analyze the dataset as best as one can, but the specific task usually depends on how sophisticated the analyst is. The model can be as simple as a simple linear regression and as complicated as a neural network with a billion parameters.
2. Principles or *Dao*: The analysis in this mindset starts with a scientific, engineering, or business problem. The goal is to understand the mechanisms involved in the problem and use that insight to make better decisions. This could be about estimating the causal effect of an intervention or understanding the limitations of the dataset at hand.

Let me try to give a more concrete example. I serve regularly in the Statistics Clinic organized by our Lab, which offers free consulting to members of the University of Cambridge. In the clinic, I have met many people who need help with statistical methodology. Our conversation typically starts with the client asking me: "*How* do I fit this model to my dataset?" And my immediate response is: "*Why* do you want to fit that model?" I usually then ask the client to describe their scientific problem. Surprisingly, they often tell me that their collaborator or supervisor just gave them the dataset and wanted them to use that model. Occasionally, I get a better understanding of their problem and end up suggesting a different model to them. Unfortunately, my client is rarely interested in the better model, usually because there is no software that can be immediately used by them, as the model I suggested is tailored to their problem. How ironic is that!

I suspect that most statisticians who engage with real world problems share similar experiences. It seems that the majority of data analysts put far more emphasis on techniques rather than principles. But techniques are only good if they are guided by good principles and applied to the right problems. Fundamentally, the difference between technique-driven modeling and principle-driven modeling is that the former culture views statistical modeling as choosing the best model for the data, while the latter culture views statistical modeling as a means to make better decisions. These goals are not always conflicting (for example, if a better decision means a better prediction for the next observation). However, most of the time these two mindsets lead to very different models and analyses.

A cyclic view of statistical modeling

The technique-driven culture can be quite dangerous, especially when stakes are high. A good example is the early analyses of the COVID-19 pandemic in the last year, on which I have written several articles and commentaries (Zhao et al., 2021; Bacallado et al., 2020; Zhao, 2020). A particularly striking example is that some extremely influential studies did not even estimate the rate of an exponential growth curve correctly (Zhao, 2020). A simple Poisson log-linear regression would serve this purpose well, but infectious disease modeling is dominated by compartmental dynamic models. Such models are useful for making forecasts, but are usually over-identified when the analyst only have early outbreak data. To solve this issue, most analysts just plug in very rough estimates of some of the model parameters and ignore the uncertainty in these estimates. This often leads to poor fits, deceptively small confidence intervals, and ridiculous results. An infamous example is a study by Oxford's Evolutionary Ecology of Infectious Disease group (Lourenço et al., 2020). By setting the proportion of population at risk of severe disease at an impossibly small value—0.1% (data from China and Italy by then already suggest the infection fatality ratio is much higher than 0.1%), the paper suggested that it is possible that half of the UK population had already been infected by March 19, 2020. This conclusion is obviously ridiculous in hind sight, but the study was instantly picked up by the media in headlines. It created unnecessary confusion at a critical time and was used by supporters of the passive "herd immunity" approach (immunity acquired by viral infection instead of vaccine) to criticize government's public health interventions.

Incidentally, this last study was perhaps also partly driven by preconceptions and politics, judging by the corresponding author's choice to publicize their extremely preliminary results and later advocacy of the notorious Great Barrington Declaration (Mandavilli and Stolberg, 2020). Every statistician knows how easy and dangerous it can be to selectively report and analyze data, especially if one already has a desirable conclusion in mind. I hope this will never become a "culture" in statistical modeling.

When trying to understand the techniques versus principles dichotomy of statistical modeling, I find the following quote from Box (1957) particularly incisive:

> Scientific research is usually an iterative process. The cycle: conjecture-design-experiment-analysis leads to a new cycle of conjecture-design-experiment-analysis and so on. It is helpful to keep this picture of the experimental method in mind

> when considering statistical problems. Although this cycle is repeated many times during an investigation, the experimental environment in which it is employed and the techniques appropriate for design and analysis tend to change as the investigation proceeds.

So the principles or the *Dao* are important, because they provide us a holistic view of statistics and guide the usage of techniques. On the other hand, the techniques or the *Shu* are equally important, because they allow us practice and improve the principles or *Dao* of statistical modeling.

Returning to its roots

On the surface, it may seem that I am criticizing Breiman's promotion of the "algorithmic modeling" culture. After all, decision trees and neural nets are just better *techniques* for predictive analyses than linear regression, logistic regression, and Cox's model. However, I believe that the dichotomy of principles versus techniques usually take different forms at different times. In 2001, the most notable division in statistics was between the "data modeling" culture and the "algorithmic modeling" culture, but the situation has drastically changed over the past two decades. For my generation of statisticians, ideas from machine learning are perhaps more influential than ideas from classical statistics. Some of my peers already see "data models" like linear and logistic regressions as obsolete techniques and make neural networks their default choice for real world problems.

But Breiman's article was more than just "data models" versus "algorithmic models" or classical statistics versus machine learning. In his final remarks, he makes it clear that he is "not against data models per se ... but the emphasis needs to be on the problem and on the data." In the very last paragraph, he concludes that: "The roots of statistics, as in science, lie in working with data and checking theory against data. I hope in this century our field will return to its roots." I agree with these points wholeheartedly.

At the moment, there is a contentious debate in the machine learning community about where it is heading to, with a minority of researchers calling for a "causal revolution" (Pearl, 2018). This is perhaps less controversial in statistics. Most statisticians I met recognize the importance of causal inference and are genuinely interested in learning more about it. Unfortunately, this is not always easy as the concept of causality is largely missing in statistics education.

A even bigger and more fundamental problem is that most Ph.D. statisticians (including me) are trained to develop new techniques and most data analysts are trained to apply the techniques. This is not just a contemporary problem; Tukey (1961) has already noticed it 60 years ago:

> The research problem involving statistical and quantitative methodology ... is a problem in higher education and in the cultural anthropology of scientists: Why do so few learn to analyze data well?

Tukey suggests that every Ph.D. student in social and behavioral science should go through Box's cycle of statistical research, most wisely in the order analysis-conjecture-design-experiment-analysis. I concur with this, but also would like to note that this is not how most papers in flagship statistics journals are being written or evaluated.

I would like to conclude with the opening sentence of the Chinese classic *Daodejing* (more widely known as the *Tao-te Ching* in the western world): "The *Dao* that can be said is not the eternal *Dao*". I think the same applies to statistic modeling, and there is never a right way of practicing or teaching statistics. But this is exactly why statistical modeling is so enchanting.

References

Sergio Bacallado, Qingyuan Zhao, and Nianqiao Ju. Letter to the editor: Generation interval for COVID-19 based on symptom onset data. *Eurosurveillance*, 25(29), 2020. doi: 10.2807/1560-7917.es.2020.25.29.2001381. URL https://doi.org/10.2807/1560-7917.es.2020.25.29.2001381.

George E P Box. Iterative experimentation. *Biometrics*, 13(2):240–241, 1957. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2527806.

José Lourenço, Robert Paton, Craig Thompson, Paul Klenerman, and Sunetra Gupta. Fundamental principles of epidemic spread highlight the immediate need for large-scale serological surveys to assess the stage of the sars-cov-2 epidemic. *medRxiv*, 2020. doi: 10.1101/2020.03.24.20042291. URL https://www.medrxiv.org/content/early/2020/12/22/2020.03.24.20042291.

Apoorva Mandavilli and Sheryl Gay Stolberg. A viral theory cited by health officials draws fire from scientists. *The New York Times*, 2020. URL https://www.nytimes.com/2020/10/19/health/coronavirus-great-barrington.html.

Judea Pearl. Theoretical impediments to machine learning with seven sparks from the causal revolution. 2018. URL http://arxiv.org/abs/1801.04016v1.

John W Tukey. Statistical and quantitative methodology. In D. P. Ray, editor, *Trends in Social Sciences*, pages 84–136. Philosophical Library, New York, 1961.

Qingyuan Zhao. Small data, big time-a retrospect of the first weeks of covid-19, 2020. Unpublished manuscript.

Qingyuan Zhao and Trevor Hastie. Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, 39 (272-281):1–10, 2021. doi: 10.1080/07350015.2019.1624293. URL https://doi.org/10.1080/07350015.2019.1624293.

Qingyuan Zhao, Nianqiao Ju, Sergio Bacallado, and Rajen D. Shah. BETS: The dangers of selection bias in early analyses of the coronavirus disease (COVID-19) pandemic. *The Annals of Applied Statistics*, 15(1):363–390, 2021. doi: 10.1214/20-aoas1401. URL https://doi.org/10.1214/20-aoas1401.