Can algorithms themselves be biased?

Momin M. Malik
Apr 24 · 22 min read

Problems viewing on mobile? (Does "β" show up as a box?) Check out a mobile-friendly pdf at https://www.mominmalik.com/algo_bias.pdf.

Edited May 2, 2019, adding attribution for the quote with permission, and making minor corrections.

This is a comprehensive response to a Berkman Klein Center colleague, Nagla

Rizk, who asked the following on our mailing list:

> *I am aware of data biases, but am trying to single out what can be biases in algorithms over and above what is/are already in the data, and beyond the point that algorithms are designed by experts without incorporating the input of communities affected. Can algorithms themselves be marginalizing, or does it all depend on the data? And if they can be, how so?*

First, while much of the conversation about AI and machine learning is around “algorithms,” and this is the term that people in machine learning use, this is misleading (I am working on an academic article about this). Machine learning “algorithms” are actually statistical *models*, and are better understood as such.

So, the better question is can *statistical models* be marginalizing, apart from data? The short answer is, “yes, but it doesn’t really matter when compared to the choice to use machine learning.”

The long answer is that there are three layers to consider: (1) whether to use quantitative modeling at all, (2) whether to do “explanatory” modeling or “predictive” modeling (machine learning falls into the latter) and, lastly, (3) aspects of the model. The direct answer to the question that motivates this post falls into (3), but I think (1) and (2) give crucial context.

Caution: I will be throwing out technical terms when I give examples, and do not explain most of these, as doing so would require making this piece into a primer about modeling. But hopefully, an interested reader can look up specific terms to learn more.

(1) Choosing to use quantitative (and, specifically, statistical) modeling

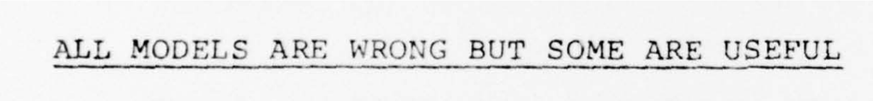
The choice to use quantitative modeling *at all* has consequences. For a discussion of the kind of things that quantitative models can never capture, I enjoy this quote from Michael Quinn Patton (2014) [bold emphasis added]:

> *“During the writing of this book, my first grandchild was born, and this book is dedicated to her. The hospital records document her weight, height, health, and* Apgar *score—activity (muscle tone), pulse, grimace (reflex response), appearance, and respiration. The mother’s condition, length of labor, time of birth, and hospital stay are all documented. These are physiological and institutional metrics. When aggregated across many babies and mothers, they provide trend data about the beginning of life—birthing.* ***But nowhere in the hospital records will***

> ***you find anything about what the birth of Calla Quinn*** **means**. *Her name is recorded but not why it was chosen by her parents and what it means to them. Her existence is documented but not what she means to our family, what decision-making process led up to her birth, the experience and meaning of the pregnancy, the family experience of the birth process, and the familial, social, cultural, political, and economic context that is essential to understanding what her birth means to family and friends in this time and place. A qualitative case study of Calla's birth would capture and interpret the story and meaning of her entry into the world from the*

> *perspectives of those involved in and touched by her coming into our lives.”*

In the immortal words of George Box, “all models are wrong, but some are useful.”



ALL MODELS ARE WRONG BUT SOME ARE USEFUL

"All models are wrong, but some are useful." From George E. P. Box's 1979 technical report, "Robustness in the Strategy of Scientific Model Building," Technical Summary Report #1954, University of Madison-Wisconsin Mathematics Research Center.

All quantitative models simplify the world. If they didn’t, they wouldn’t be models! (as in the famous dictum, “the map is not the territory”—but, to Box’s point, that doesn’t mean maps are useless.) Simplification means choices about what is and is not important, and these choices always have consequences.

We could phrase this simplification, and the impossibility of capturing everything in the world, as a trivial way in which all are models are "biased";[1] but below, especially in section (3), I consider if there can be properties of a given statistical model that introduce a *specific* bias.

There are a few types of quantitative modeling, but for modeling social systems, statistical modeling is the dominant type.[2]

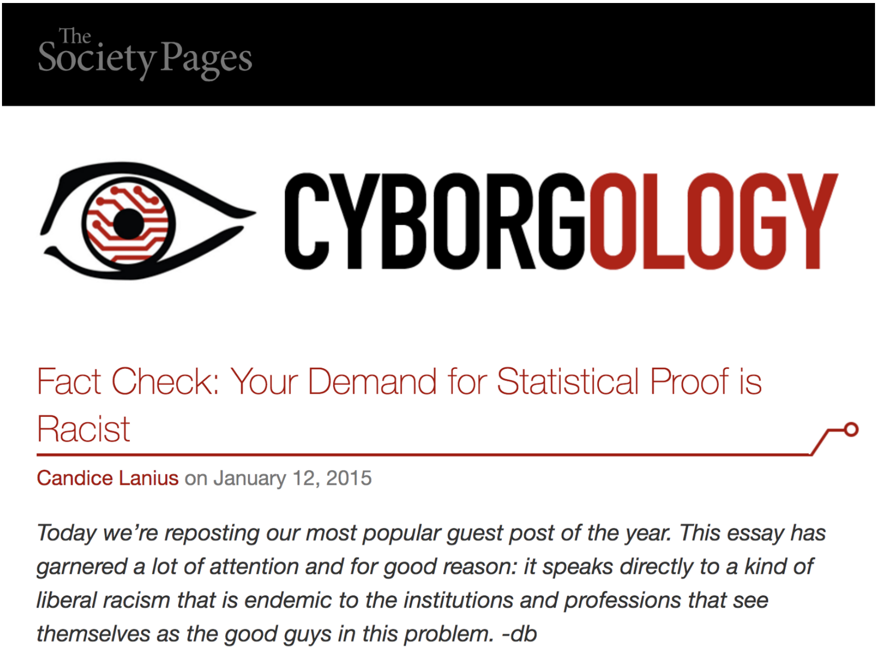
Something that is mostly inescapable once the choice has been made to specifically use *statistical* models (which includes machine learning) is we assume the world can be divvied up into entities (observations), and properties of those entities (i.e., variables). This is not inevitable. In a classic article, sociologist Andrew Abbott (1988) remarked [emphasis added],

> *“…**it is striking how absolutely these assumptions contradict those of the major theoretical traditions of sociology.** Symbolic interactionism rejects the assumption of fixed entities and makes the meaning of a given occurrence depend on its location—within an interaction, within an actor’s biography, within a sequence of events.*

> *Both the Marxian and Weberian traditions deny explicitly that a given property of a social actor has one and only one set of causal implications. Marx's dialectical causality makes events produce an opposite as well as a direct outcome, while Weber and the various hermeneutic schools treat attributes as infinitely nuanced and ambiguous. Marx, Weber, and work deriving from them in historical sociology all approach social causality in terms of stories, rather than in terms of variable attributes."*

Abbott notes that some of these assumptions can be relaxed (e.g., a time series lets entities change), but those relaxations don't fundamentally change this schema for the world (e.g., in a time series, at a point in time an entity is still fixed; there is, for example, no notion of it changing depending on who is perceiving it).

Neither is this view of the world natural, or obvious. The application of statistical modeling to the social world was one that developed over time, and was not necessarily appropriate (Freedman, 2005). The idea that we can use an *average* of a population to characterize it was critiqued when first introduced (Donnelly, 2016). And statistical modeling has normative consequences (Rose, 2016) and political uses, like delegitimizing lived experiences: as Candice Lanius (2015) put it bluntly, “Your Demand for Statistical Proof is Racist.”



Fact Check: Your Demand for Statistical Proof is Racist

Candice Lanius on January 12, 2015

Today we’re reposting our most popular guest post of the year. This essay has garnered a lot of attention and for good reason: it speaks directly to a kind of liberal racism that is endemic to the institutions and professions that see themselves as the good guys in this problem. -db

Header of Candice Lanius, "Fact Check: Your Demand for Statistial Proof is Racist," Cyborgology blog, 2015, https://thesocietypages.org/cyborgology/2015/01/12/fact-check-your-demand-for-statistical-proof-is-racist/.

She writes:

> *A white woman can say that a neighborhood is "sketchy" and most people will smile and nod. She felt unsafe, and we automatically trust her opinion. A black man can tell the world that every day he lives in fear of the police, and suddenly everyone demands statistical evidence to prove that his life experience is real. Anything approaching a "post-racial society" would not require different types of evidence to tell our life stories: anecdotal evidence for white people, statistics for black people.*

That is, it is not inevitable that we use quantitative data and modeling for, say, demonstrating racism: we could just believe people when they describe their lived experiences.

(2) The distinction between prediction and explanation

Machine learning is done almost entirely with statistical machinery, although used in ways that have been anathema to classical statisticians (see Breiman, 2001, and Jones, 2018). Despite the same underlying math, one important distinction between machine learning and statistics is the distinction between models that are are "explanatory" and models that are "predictive", something that is enormously consequential and I am always surprised to see not discussed constantly.

“Explanatory” models try to tell us something about the underlying process. “Predictive” models try only to produce a reliable output (one that matches what happens in the world) given an input. The “learning” in machine learning refers to a program combining data with some assumptions to “learn” to produce a reliable output (which has almost nothing to do with the learning studied with, say, sociocultural learning theory).

While we might assume that prediction and explanation are the same task, it turns out to not be the case. To explain why, I will first explain the definitions of “prediction” and “bias” as *technical* terms (as the technical terms do not map onto colloquial understandings or dictionary definitions).

First, in statistics and machine learning, “prediction” is a technical term, defined as minimizing the error between a combination of input variables (also called covariates, independent variables, explanatory variables, or predictors), and a target output variable, *y* (also called the dependent variable or the response). *This can be achieved through correlations alone* (Lipton, 2015). Minimizing error is called “prediction” because we assume that the way in which correlations minimized errors in the past is a reliable guide to how they might do so in the future. Of course, if something in the world changes, then this assumption will fail; previously observed correlations will not be a reliable guide, and “predictions” will fail to *actually* predict.

Second, in statistics and machine learning, there is a formal definition of *bias* in models. This maps somewhat onto what we might mean by a model being biased, but relates to a metaphysical idea of what modeling does rather than to lived experience.

Specifically, we hypothesize that there is a “true”, underlying relationship between x and y, $y = f(x)$. For example, in a linear model, f is multiplying x by β_1 and adding β_0. We take some assumptions about what forms f can take, and combine this with data to make an *estimate* of f (or things within a specific form of f, like the β’s), which we notate as $\hat{f}$ and read as “eff-hat” (or, for the specific form f in terms of β’s, estimate as $\hat{\beta}$’s, and read as the “beta-hats”). Much of statistical theory, and machine learning theory, is devoted to studying the relationship between $\hat{f}$ and f: do our assumptions about f, and the way in which we use data, give us an $\hat{f}$ that is close to f?

The *bias of an estimator* is defined as the amount by which $\hat{f}$ departs from f (in expected value).

The way in which this somewhat maps onto what we colloquially mean by bias is, if our $\hat{f}$ is *biased* in comparison to f, it can lead to marginalization by not treating people “correctly”. I give an example of this below in section (3), where I set up a toy example where I set $f(x) = 2 + \sin(x)$, and consider $\hat{f}$'s that assume some relationship between x and y other than the true, sinusoidal one.

But when actually doing modeling in the real world, there is no such thing as the "true" f, let alone one that can be known! Consequently, we can never know if a model is biased under this theoretical notion of bias (of not matching a "true", "underlying" process). Whereas we can know, and should strive to know, if a particular $\hat{f}$ leads to social bias and marginalization. So the theoretical notion is only so useful.

With these definitions of "prediction" and "bias", I can get to the central weirdness about how explanatory modeling is different from predictive modeling: quite surprisingly, it turns out that biased, "wrong" models, that are *actually worse at reflecting causal processes than other models*, can do better at "prediction"! (Again, so long as the world doesn't change, and under this metaphysical notion of a "true" model such that we can say whether a given model is right or wrong.) Explanatory modeling seeks out unbiased models, but predictive modeling is happy to sacrifice being unbiased in favor of better predictions.

The reasons why this divide exists are complicated and deeply counter-intuitive, and have to do with the "bias-variance tradeoff" (see Shmueli, 2010)[4] as well as the difficulty of making models that reflect causal processes in the first place (see again

Breiman, 2001; but also, perhaps no modeling can ever truly get at causality). Less philosophically, it is sometimes the case that isolating the precise relationship between one specific input variable and the target output variable sometimes requires sacrificing how well we can model the output variable as a whole (as is frequent in the use of "instrumental variables" in econometrics).

This means that machine learning models that predict better than theory-driven statistical models do not do so because they are secretly "more right"; they often do so *despite being less right*. And, their predictive success is always fragile.

On the one hand, this leads to the arguments from statistics, econometrics, and “causal learning” that, even if our only goal is prediction, if we know about a causal relationship it will lead to us making predictions that are *robust* to changes in context (even if the short-term predictions aren’t as good).

But on the other hand, and more profoundly, this means that predictive modeling may be useless for finding out how to *intervene* in a system. As a concrete illustration, very different models, suggesting very different causal mechanisms, can make equally good predictions (for an example of this, see Sendhil Mullainathan's 2017 article with Jann Spiess).

Machine learning is entirely "predictive" (with the possible exception of some things in the area of probabilistic graphical models). In contrast, statistics is predominantly "explanatory", although (for better or worse) modern statistics has followed machine learning and has been taking up predictive modeling more and more.

(3) Aspects of the model

For *explanatory* modeling, there are some specific choices to be made about which assumptions to make (see again Abbott), including but not limited to:

- Is the relationship between the [mean of the] response and the covariates linear, or are there nonlinear relationships? Which covariates have nonlinear relationships, and what are the forms of the nonlinearity?

- Does the model have a parametric form, or is it “nonparametric”?

- Is the model additive, or are there interaction effects? Between which covariates?

- Is imputation done, and how? Otherwise, how are missing values handled?

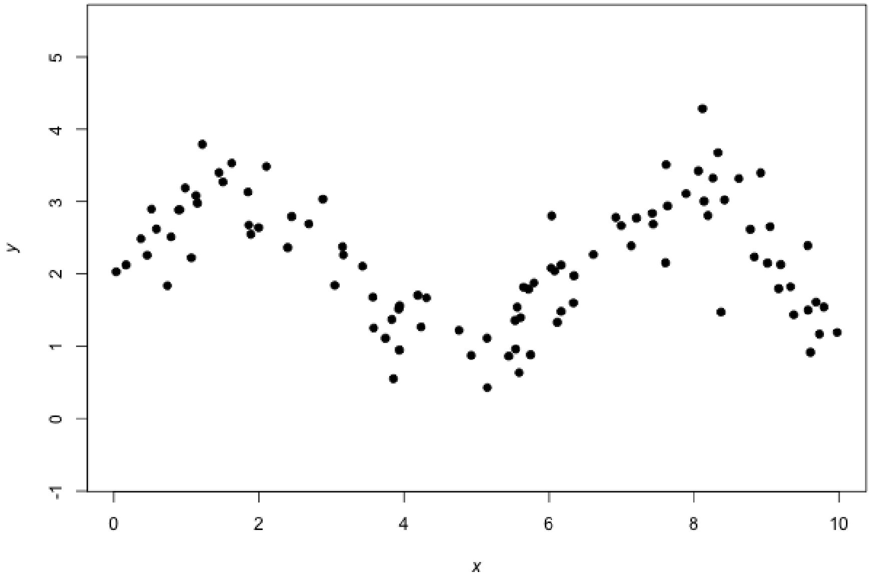
- What is the assumed distribution of the response variable? (E.g., normal, Binomial, Poisson, etc.)

- Are priors and/or regularization and/or variable selection used? What are the respective distributions and/or amounts and/or methods?
- Do we assume there is no dependency structure between errors? If not, how are we accounting for dependencies?

There are consequences for all of the above choices. We can get a very different picture of the world from different choices, and in many cases (unlike in the example below) there isn't necessarily a "right" choice. (Ideally, we get results across different modeling decisions that broadly agree; see Silberzahn et al., 2018.) Introducing an interaction effect or a nonlinear term, or using a different functional form, can give very different results.

Here is an example of some "synthetic", or "toy" data (in statistics, often we use simulated data to ask, "if the world worked the way we assume it works, do our techniques do what we want them to do?" This is a low bar to clear, but it can be a helpful exercise).

I take 200 values between 0 an 10, and choose a deterministic relationship, $y = 2 + \sin(x)$. I run these 200 values through the x in the equation, and each time add a “noise” term ε (to partially disguise the relationship) to get a y corresponding to each x, and then plot these pairs.

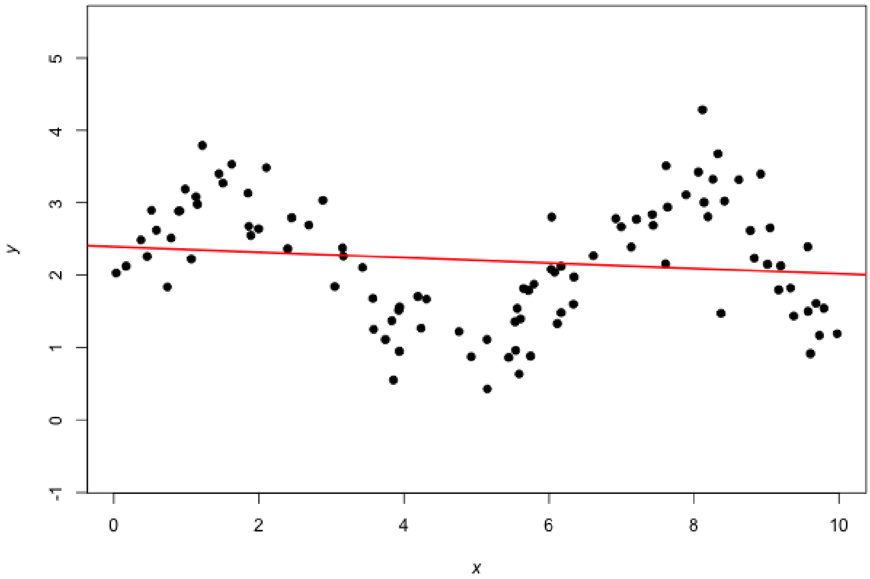


A scatterplot of synthetic data, drawn from y = 2 + sin(x) + ε, where $\varepsilon \sim \mathcal{N}(0, 2^{-2})$.

If we didn’t look at a scatterplot, and just fit the model $y = \hat{\beta}_0 + \hat{\beta}_1 x$ and looked at the results, we would see that the slope is not significantly different from zero; the estimated $\hat{\beta}_1$ is -0.03755, $p = 0.185$ (i.e., we cannot reject the null hypothesis that $\beta_1 \neq 0$). We might erroneously conclude that that there is no relationship between x and y.

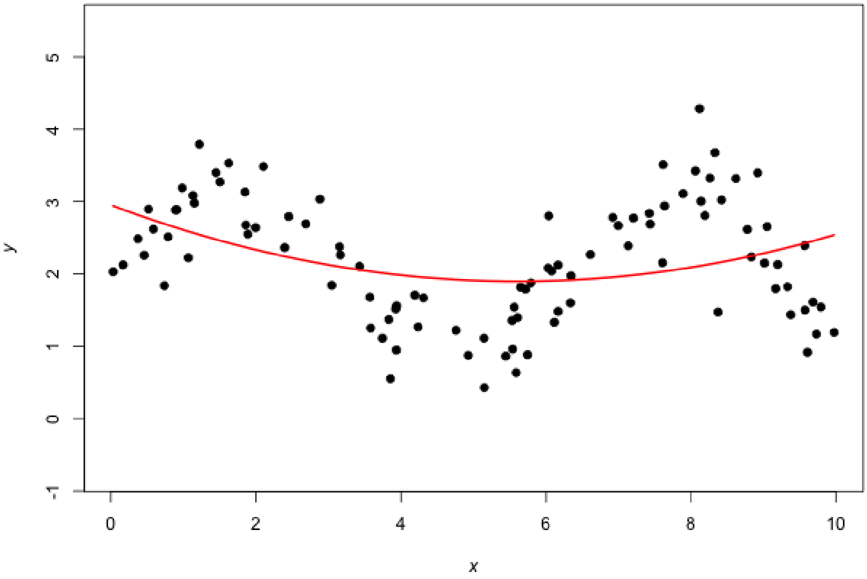
(Note that if we were doing predictive modeling/machine learning, we would be looking at the mean squared error on held-out data, and not at statistical significance; and we would likely try models other than ones only considering linear relationships. But if, for some reason, out of the models we considered this was the best-performing one and we selected it, then using this in the world would potentially marginalize people in its inaccuracy of failing to recognize the relationship between *x* and *y*.)

But the assumption of a linear relationship is one we are able to test by looking at the plot. We can see the fit is quite poor.



The fit of $y = \hat{\beta}_0 + \hat{\beta}_1 x$. Clearly it does not fit well!

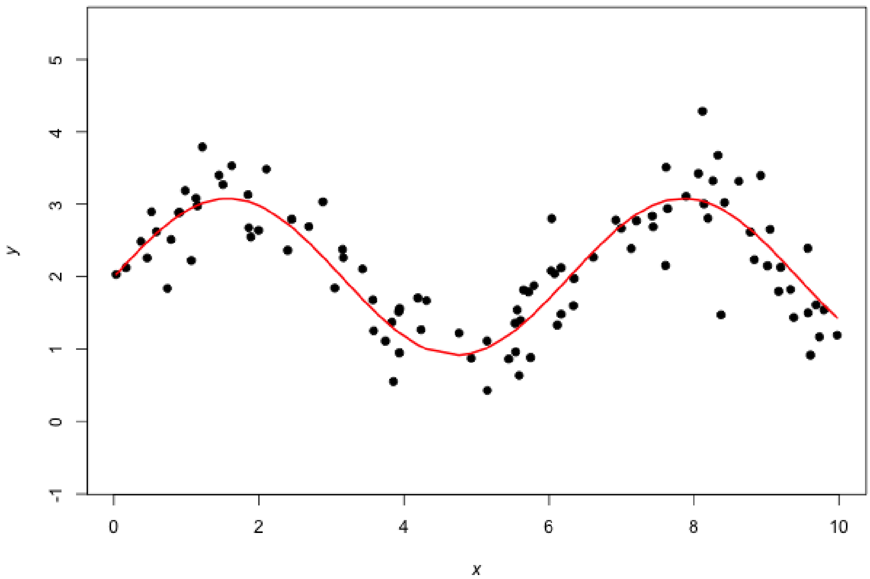
It seems that the relationship between *x* and *y* is not a linear one. So, we can try adding a quadratic term, x^2, to the regression.



The fit of $y = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$, a model with a quadratic term.

This is not much of an improvement.

But, driven by theory or by looking at the scatterplot, we might try to fit $y = \hat{\beta}_0 + \hat{\beta}_1 \sin(x)$, a model with a trigonometric term. This is fitting to the “correct” model, as it is the same as the formula that originally generated the data with $\beta_0 = 2$ and $\beta_1 = 1$.[3] This works perfectly:

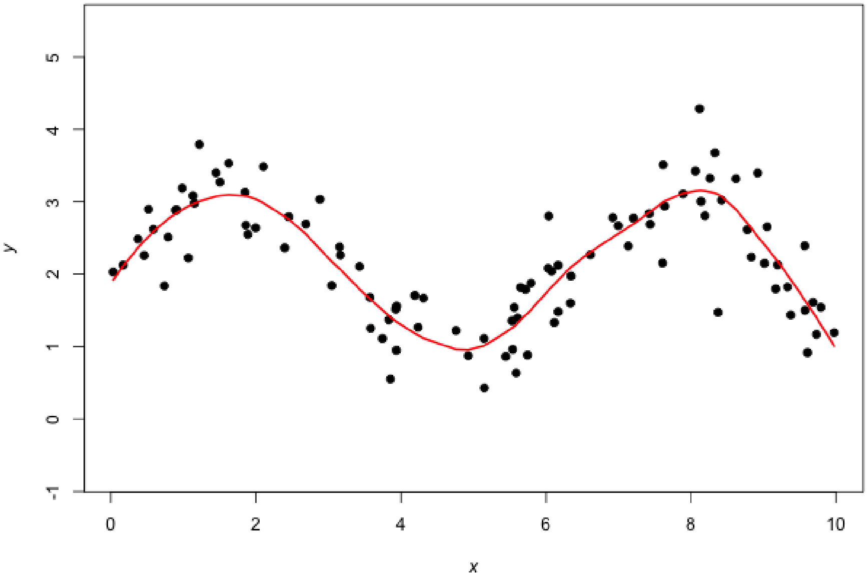


The fit of $y = \hat{\beta}_0 + \hat{\beta}_1 \sin(x)$, a model with a trigonometric term.

In this fit, the estimated intercept $\hat{\beta}_0$ is 1.996, with 95% confidence interval [1.904, 2.088] and the estimated $\hat{\beta}_1$ for $\sin(x)$ is 1.083, with 95%

confidence interval [0.947, 1.219]. Both intervals contain the true value, and the estimates are significant at $p < 0.001$.

Another option is a *nonparametric* fit. Nonparametric techniques make fewer assumptions about functional form of the relationship of x to y, and instead fit curves to the “shape” of the data.



A ″nonparametric″ fit (span chosen from minimum MSE on a held-out 10%).

This is much better than the linear and quadratic term fits, but not as good as the fit of the “correct” model. And, if our goal was to *understand* the relationship between *x* and *y*, this may not be a good approach as we don’t get any information other than what is in the graphical representation above (e.g., no estimated $\hat{\beta}$'s, not even an intercept term).[5]

In explanatory modeling, we try to use a mixture of theory, reasoning, intuition, testing (including looking at scatterplots, like above!), and experience to make these choices.[6]

For *predictive* modeling, the answers to the above questions will be, "whichever gives the best predictive performance."

Choices that have to be made for both explanatory and predictive modeling:

- The choice of optimization objective or "loss function" (this determines how a model is fit to data).

- The criteria we use for choosing one out of multiple models (perhaps the criteria is accuracy, but maybe it's based on precision and recall; in explanatory modeling, maybe it's based on *F*-tests for nested models). This can sometimes be folded into the loss function (e.g., we can have a loss function that penalizes false negatives more heavily than false positives to get a weighted accuracy we can use), but not always.
- How to do model goodness-of-fit-checking (or, in machine learning, more narrowly, where goodness-of-fit is always "predictive performance" and always checked

by “cross-validation”, how that cross-validation is carried out).

These decisions have consequences, especially if the loss function can’t or doesn’t map perfectly onto what we want in terms of social outcomes (O’Neil, 2014).

Still, even for predictive modeling, whether the best-performing model is linear or not, additive or not, parametric or nonparametric, uses regularization or not, etc., can have consequences for what kinds of errors are made, and what kinds of processes are picked up or missed.

I don’t have real-world examples (and I suspect, beyond the explanatory/predictive distinction, it would be hard to identify a choice of model as uniquely responsible for some unjust outcomes), but I can give a hypothetical example of logistic regression versus a decision tree.

Logistic regression fits a *continuous* relationship between some covariate and the probability of having a specific label. So, a logistic regression would report something like, for each year older somebody is, their "odds" of having a label increases by 1.1 times.

In contrast, a decision tree needs to discretize variables. So, a fitted decision tree would have to have some division, like age < 16.5, along which to split data in the process of arriving at an outcome. (Multiple discrete splits can approximate a continuous relationship, but this is not something decision trees are good at doing).

Decision trees also *automatically* pick up on interaction effects, for example splitting along age at different points for different sexes, and nonlinear (or more accurately, nonmonotonic) effects, for example splitting along age < 16.5 at one point and then further down splitting at age > 29.5.[7]

(For more on decision trees, see a draft paper I have: Malik, 2019).

While interaction and nonlinear terms can be used in a logistic regression, respectively here by including an “age×sex” term, and an “age^2” term, such terms need to be specifically (and manually) put into the model to be considered, and the number of possible such terms grows exponentially in the number of variables which makes it hard to do automatically.[8]

Perhaps a logistic regression that did not consider interaction effects between race and gender would be marginalizing.[9] Perhaps a decision tree has one boundary where being slightly above or below a cutoff leads to vastly differential treatment, which could also be marginalizing.

But these consequences are hard to anticipate before fitting the model to data. A better way is often to look at the kinds of errors a model makes after it is already fitted (and comparing it to the errors of other candidate models). What might this look like? One example is http://aequitas.dssg.io/example (disclosure: I was a 2017 Data Science for Social Good fellow, but am not involved in Aequitas).

Conclusion

I return to another part of the motivating question: can there be problems when models “are designed by experts without incorporating the input of communities affected”?

What would it mean to incorporate the input of affected communities?

I think the most important thing when working with a community is first deciding to use quantitative modeling in the first place. If a community doesn’t agree, then *the ethical thing to do is probably to not use modeling*. This is a perfectly legitimate stance to take, as quantitative modeling is not always applicable or even helpful, and can have negative repercussions (again, see Rose, 2016, and Lanius, 2015).

Next, the community should have input as to whether the model should try and capture or reflect causal processes and intuitions, or if predictive performance (based, for example, on non-causal correlations) is the only goal. If the former, explanatory statistical modeling should be used, and only in the latter case should machine learning “algorithms” be used. Making a decision about this may require educating people about the counter-intuitive peculiarities of modeling, the difficulty of which might end up being another reason to not use modeling.[10]

So, should data scientists seek community input when deciding whether to use, say, a decision tree or a logistic regression? I’m not sure how much this might matter. In predictive modeling/machine learning, the domain knowledge of "subject matter experts" is valued for proposing covariates and interactions/transformations to include in a model. Perhaps input can help decide on variables to *exclude* (note, however, that sensitive/protected attributes are often correlated with every other input variable, such that exclusion is not sufficient for protection), but for inclusion, if a given suggestion doesn't help prediction, it shouldn't be included (unlike in explanatory modeling). And in many cases, domain expertise turns out to not help the task of prediction, and purely data-driven approaches produce better-performing models (although they will likely be less robust to changes over time and in context). *So long as the decision has been made to use predictive modeling*, this is the place where I would say the modeler's expertise should be given priority. Plus, as I said above, it's hard to anticipate how specific aspects of a given model might lead to marginalization when used for a specific set of data.

But a community should absolutely be involved in looking at the results of a predictive model—and perhaps, on that basis, deciding not to use predictive modeling (or any modeling) after all.

Endnotes

[1] While phrased as specific to abstractions in machine learning, a 2019 article by Andrew Selbst, danah boyd, Sorelle Friedler, Suresh Venkatasubramanian, and Janet Vertesi applies quite generally to “traps” of modeling. Mainly: just because abstractions seem powerful does not mean that abstraction can do everything, which can be hard to accept for those trained in modeling.

[2] One alternative is simulation modeling, although I think this ultimately has more drawbacks than

benefits; see my discussion in Pfeffer & Malik (2017). Mathematical sociology, microeconomics, and physicists' approaches to social systems is frequently another type of quantitative modeling that doesn't have a specific label: this modeling frequently does not make use of data except for coarse confirmatory comparisons, instead relying on arguments made with equations and derivations, and perhaps from a specific underlying framework such as game theory.

[3] In general, when we interpret fitted β parameters, we interpret them under the assumption that the model is "correct", or approximately so: if but if the model is completely wrong, the

fitted parameters are worthless for interpretation.

[4] The bias-variance tradeoff is a derivation that shows that the expected value of the overall quantity $(y - \hat{f}(x))^2$, which we call the *loss*, is sometimes minimized by an $\hat{f}$ that is biased, but has lower variance than an unbiased $\hat{f}$. While the derivation is clear enough, the implication is very strange and counter-intuitive indeed, which is explored quite well in Shmueli (2010).

[5] Nonparametrics are powerful when we don't have any insight into the functional form of the data, but the "correct" model, insofar as there can be

such a thing, is better. Another drawback is that nonparametrics can be harder to use for explanation, e.g., we don't get fits of the β parameters that we can interpret substantively to learn about the underlying process. Nonparametrics are popular in modern statistics if mostly unknown in social science, and much of machine learning can be understood as a "rebranding of nonparametric statistics" (Shalizi, 2018).

[6] More complex considerations are in more specific functional form of the relationship between the response, the covariates, and the errors: choices to be made here includes using binning or otherwise transforming the coding of

the response or the covariates, taking logarithmic or other “variance-stabilizing” transformations, and using survival models, time series models, spatial models, network models, item-response models, hierarchical/multilevel models, and many more.

[7] Decision trees are also highly unstable to changes in data: slightly different data can produce vastly different trees, although often with similar predictive performance in the end, but I’m not sure I can think of a way in which this could be marginalizing.

[8] Possible approaches include polynomial features with the lasso, and kernel methods—although with enough variables even these would be infeasible.

[9] There is a connection to intersectionality here, but note that true intersectional feminist theory would not accept the discretization/quantification of race and sex/gender that would be required for use in a statistical model.

[10] I explicitly avoided talking about problems with data, as there is lots of critical material about this already; but I emphasize that having community input or even determination about

which variables and measures to use, and incorporating community knowledge about what measures fail to capture and sources of bias in data, is critical. Qualitative research, and specifically Participatory Action Research, can be a systematic way of going about this.

References

Abbott, Andrew. “Transcending General Linear Reality.” *Sociological Theory* 6, no. 2 (1988): 169–186. https://dx.doi.org/10.2307/202114.

Breiman, Leo. “Statistical Modeling: The Two Cultures.” *Statistical Science* 16, no. 3 (2001): 199–231.

https://dx.doi.org/10.1214/ss/1009213726.

Donnelly, Kevin. *Adolph Quetelet, Social Physics, & the Average Men of Science, 1796–1874*. Pittsburgh, PA: University of Pittsburgh Press, 2016.

Freedman, David A. *Statistical Models and Causal Inference: A Dialogue with the Social Sciences*. Cambridge, UK: Cambridge University Press, 2009.

Freedman, David A. “Linear Statistical Models for Causation: A Critical Review.” In *Encyclopedia of Statistics in Behavioral Science*, edited by Brian S. Everitt and David C. Howell, 1061–1073. John Wiley & Sons, Ltd., 2005.

https://dx.doi.org/10.1002/0470013192.bsa598.

Jones, Matthew L. “How We Became Instrumentalists (Again): Data Positivism since World War II.” *Historical Studies in the Natural Sciences* 48, no. 5 (2018): 673–684. http://www.columbia.edu/~mj340/HSNS4805_12_Jones.pdf.

Lanius, Candice. “Fact Check: Your Demand for Statistical Proof is Racist.” *Cyborgology* blog, January 15, 2015. https://thesocietypages.org/cyborgology/2015/01/12/fact-check-your-demand-for-statistical-proof-is-racist/.

Lipton, Zachary C. “The Myth of Model Interpretability.” *KDnuggets* 15, no. 13

(April 2015). https://www.kdnuggets.com/2015/04/model-interpretability-neural-networks-deep-learning.html. See also the updated version: Zachary C. Lipton, “The Mythos of Model Interpretability,” *ACM Queue* 16, no. 3 (June 2018): 31–57, https://dx.doi.org/10.1145/3236386.3241340.

Macy, Michael A., and Robert Willer. “From Factors to Actors: Computational Sociology and Agent-Based Modeling.” *Annual Review of Sociology* 28 (2002): 143–166. https://dx.doi.org/10.1146/annurev.soc.28.110601.141117.

Pfeffer, Jürgen, and Momin M. Malik. "Simulating the Dynamics of Socio-Economic Systems." In *Networked Governance: New Research Perspectives*, edited by Betina Hollstein, Wenzel Matiaske, and Kai-Uwe Schnapp, 143–161. Cham, Switzerland: Springer, 2017. https://dx.doi.org/10.1007/978-3-319-50386-8_9.

Malik, Momin M. "Interpretability is a Red Herring: Grappling with 'Prediction Policy Problems'." Paper presented at the 17th Annual Information Ethics Roundtable: Justice and Fairness in Data Use and Machine Learning, April 5–7, 2019, Northeastern University, Boston, MA.

Draft at https://www.mominmalik.com/ier2019draft.pdf, slides + draft at https://www.mominmalik.com/ier2019.pdf.

Mullainathan, Sendhil, and Jann Spiess. “Machine Learning: An Applied Econometric Approach.” *Journal of Economic Perspectives* 31, no. 2 (2017): 87–106. https://dx.doi.org/10.1257/jep.31.2.87.

O’Neil, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York, NY: Crown, 2016.

Patton, Michael Quinn. “The Nature, Niche, Value, and Fruit of Qualitative Inquiry.” In *Qualitative Research & Evaluation Methods: Integrating Theory and Practice,* 4th edition, 2–44. SAGE Publications, Inc., 2014. https://uk.sagepub.com/sites/default/files/upm-binaries/64990_Patton_Ch_01.pdf.

Rose, Todd. *The End of Average: How We Succeed in a World That Values Sameness.* New York, NY: HarperOne, 2016. See excerpt at https://www.thestar.com/news/insight/2016/01/16/when-us-air-force-discovered-the-flaw-of-averages.html, and animated video at https://vimeo.com/237632676.

Selbst, Andrew D., danah boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. “Fairness and Abstraction in Sociotechnical Systems.” In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (FAT* '19), 59–68. New York, NY: ACM. https://dx.doi.org/10.1145/3287560.3287598.

Shalizi, Cosma R. “Revised and Extended Remarks at ‘The Rise of Intelligent Economies and the Work of the IMF’.” *Three-Toed Sloth* blog, October 18, 2018. http://bactra.org/weblog/imf-2017-talk.html.

Shmueli, Galit. “To Explain or To Predict?” *Statistical Science* 25, no. 3 (2010): 289–310. https://dx.doi.org/10.1214/10-STS330.

Silberzahn, R., E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. Awtrey, Š. Bahník, et al. “Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results.” *Advances in Methods and Practices in Psychological Science* 1, no. 3 (September 2018): 337–56. https://dx.doi.org/10.1177/2515245917747646.

Thanks to the Berkman Klein Center community for this question and for

giving me positive feedback on my response, and to B Cavello for encouraging me to turn this into a blog post and helping me with some feedback!

Appendix

R code for generating the images, and doing associated analysis.

```
install.packages("magrittr")
library(magrittr) # Enables piping, %>%, for more readable code

set.seed(201904) # For reproducibility
set.seed(runif(1)*10000000) # But mixing it up a bit
n <- 100 # Number of observations
```

```
x <- sort(runif(n = n, min = 0,
max = 10))
y <- 2 + rnorm(n = n, mean =
sin(x), sd = .5)
summary(lm(y~x)) # For
reporting slope and
significance
(fit <- summary(lm(y ~
sin(x)))) # Same
coef(fit) %>% round(3) # Get
estimates and confidence
intervals
c(coef(fit)[1,1] -
1.96*coef(fit)[1,2],
  coef(fit)[1,1] +
1.96*coef(fit)[1,2]) %>%
round(3)
c(coef(fit)[2,1] -
1.96*coef(fit)[2,2],
  coef(fit)[2,1] +
1.96*coef(fit)[2,2]) %>%
round(3)

w <- 600 # image width
h <- 400 # image height
```

```
par(mar = c(3.5,3.5,0,0)+.5) #
decrease borders

png("scatterplot.png", width =
w, height = h)
plot(x, y, pch = 19, asp = 1,
     xlab =
expression(italic(x)),
     ylab =
expression(italic(y)))
dev.off()

png("linear_fit.png", width =
w, height = h)
plot(x, y, pch = 19, asp = 1,
     xlab =
expression(italic(x)),
     ylab =
expression(italic(y)))
lm(y~x) %>%
  abline(col = 2, lwd = 2)
dev.off()

png("quadratic_fit.png", width
= w, height = h)
plot(x, y, pch = 19, asp = 1,
     xlab =
expression(italic(x)),
     ylab =
expression(italic(y)))
lm(y ~ x + I(x^2)) %>%
  predict %>%
  lines(x, ., col = 2, lwd = 2)
dev.off()

png("trigonometric_fit.png",
width = w, height = h)
plot(x, y, pch = 19, asp = 1,
     xlab =
expression(italic(x)),
     ylab =
expression(italic(y)))
lm(y ~ sin(x)) %>%
  predict %>%
  lines(x, ., col = 2, lwd = 2)
dev.off()

# Tuning parameter selection
from cross-validation.
# Randomly partition data to
fit/train, and tune/validate.
train <-
```

```
sample(c(rep(T,floor(.9*n)),

rep(F,ceiling(.1*n))))
mse <- function(x1,x2)
mean((x1-x2)^2)
spans <- (10:90)/100 # Tuning
parameter range
loss <- lapply(spans, #
Calculate loss for all spans
         function(i)
           loess(y[train] ~
x[train], span = i) %>%
           predict(newdata =
x[!train]) %>%
           mse(y[!train])) %>%
unlist
plot(spans, loss, type = "l", #
Plot validation loss by span
     xlab = "span", ylab =
"MSE")
abline(v=spans[which.min(loss)]
, col = 2, lty = 2)
abline(h=min(loss), col = 2,
lty = 2)
opt <- spans[which.min(loss)] #
Select optimum span
```

```
png("nonparametric_fit.png",
width = w, height = h)
plot(x, y, pch = 19, asp = 1,
     xlab =
expression(italic(x)),
     ylab =
expression(italic(y)))
loess(y ~ x, span = opt) %>%
  predict %>%
  lines(x, ., col = 2, lwd = 2)
dev.off()
```