Further Thoughts on The Two Cultures

David Zynda

December 14, 2017

Beauty characterized in mathematical expressions should render nothing short of sheer wonder given the nature of the discipline. Largely constructed in the mind, or even sometimes postulated from the Platonic heavens above, the correspondence of theoretical math to the natural world evokes more than a fortuitous and incidental instantiation, but rather seems to approach near truth itself. However, the deductive and formal evaluation of one formula via proof to another, no matter the elegance, still falls short of a perfect material match to the outside world. Long dictated but yet seemingly poorly understood is the fact that, "all models are wrong, but some are useful"[1]. No matter the beauty and creativity expressed step by step through a proof, the marriage of the theoretical domain to that of the physical world remains a daunting task to fully unite. There is no doubt, however, that the enormous success characterized by the mathematization of the world over the last few hundred years has fulfilled far more promises than not.

Isolated, controlled experiments, such as those found in agriculture and biology, no doubt prove useful to the traditional paradigm of experimental design, model validation, and goodness of fit tests, among the whole host of other assumptions. Moreover, it may even be the case that useful is too weak of a word to use. Indeed, it seems as the perfect match: the mathematical partitioning through designs of experiments harbor great insight and power to understand an underlying phenomenon of interest. Yet, the situations where such controls and measures can be executed seem rather limited, all else considered. Concerning economics, James Buchanan writes, "A market is not competitive by assumption or construction. A market becomes competitive, and competitive rules come to be established as institutions emerge to place limits on individual behavior patterns"[2]. In such case where it may not be always prudent to assume a certain set of assumptions, the question arises as to how one can best witness this "becoming" that Buchanan speaks of, be it in economics, atmospheric science, and many others.

Could the time be approaching in recognizing the limits of purely theoretical methods and models subject to rigid proofs and constraints? Given the rise of advanced computational power, should holding oneself to a set of prescribed assumptions be the norm instead of the exception? Certainly, there is a place and time for all methods. Yet, the rise of simulation in place of mathematically elegant optimization should be a force to be reckoned with, especially in the domain of modern statistics and economics.

In 1984, the beginning of a statistical revolution began its ascent. However, as Paul Feyerabend duly noted, no theory is born into a world that is wholly ready for it. *Classification and Regression Trees* by Breiman, Friedman, Stone and Olshen made a startling appearance on the traditional statistics discipline formerly characterized by model assumptions and mathematical restrictions. Telling, indeed, was that this publication would not be taken to a journal, but rather a book. No journal, at the time would take it[3]. Contrary to the work that was being done at the time, it mandated that the data should speak for itself compared to older assumptions that a model should be fitted upon the data and forced into a set of assumptions.

Dominant thought before and even today assumes that a stochastic data model exists within a black box representing the particular phenomenon of interest[4]. The black box then became a regression, or similar process which can be subject to model validation and goodness of fit and residual tests. The new way, fiercely advanced by Breiman, assumes no data generating process and leaves the black box to the unknown. Instead of representing the data through model imposition, algorithms such as Classification and Regression Trees (CART), Support Vector Machines (SVM), or Random Forests are advanced to solely predict the outputs of the black box given the inputs. In place of goodness of fit, predictive accuracy becomes the bastion and the proper test of evaluation.

[1] Box, G.E.P. (1976), "Science and Statistics", *Journal of the American Statistical Association*, 71: 791-799.

[2] Buchanan, James M., "What Should Economists Do" *Southern Economic Journal*, Vol. 30, No. 3 (Jan., 1964), pp 218.

[3] From a lecture in February 2017 delivered by Breiman's colleague Dr. Adele Cutler at Utah State University.

[4] Breiman, Leo, "Statistcal Modeling: The Two Cultures" *Statistical Science* Vol. 16, No. 3, 2001, pp. 199.

New methods of algorithmic modeling based on classification and regression trees allowed for the possibility of increasing predictive accuracy unlike linear and logistic regressions which came before it. Moreover, the promise to sift through multidimensional data and overcome the curse of dimensionality[5] would lead to unprecedented capabilities to do that which was never done before. Yet, it could never be accepted in a peer reviewed journal in statistical science. No matter the effort put forth, the pursuit to establish prescribed mathematical proofs and definitive assumptions and constraints could not be done as with ordinary least square and asymptotic methods of an older age. In abandoning the tradition preceding it, CART gave rise to so-called black box algorithms which have advanced technology and computer science in context of newer and more interesting problems statisticians could have formerly never touched. .

As late as 2001, Breiman estimated that still only 2% of academic statisticians embraced the methods and their subsequent successors he created beginning with his 1984 book[6]. He sees an undue fascination with fitting data models rather than understanding the underlying natural mechanism the model seeks to represent. To "invent a reasonably good parametric class of models for a complex mechanism devised by nature" lends itself to an undeniable conclusion: any results from such a model do not represent that nature's mechanism but rather the model's mechanism[7]. Consequently, if the model is not a good emulation of nature, then the conclusions may be wrong.[8]. No matter the beauty of the parametric model underlying the phenomena of interest, there may not in fact exist a one to one correspondence of the model's structure to that of a more complex phenomenon seeking to be understood. More importantly, if the traditional data models used for the last century do not adequately approximate more complex states of affairs, attention should be drawn to the concern of using such models generated through traditional methods to the evaluation of policy and scientific conclusions alike.

Current work in economics continues to maintain a simple data model approach described by Breiman. Standard econometrics courses, based on such books as those by Wooldridge[9], yet maintain a rudimentary process of model fit and p-value check guided decision making to decide what variables get placed on the chopping block and which get to inform policy. Imposing models on economic data with hope to get below the prized 5% p-value results in potentially misguided results and limits the discipline to dated methods. No doubt a student needs to begin at square one, but refering to t-tables in an appendix to see if a variable is significant seems 50 years out of fashion.

If not quite evidenced enough, the author writes of his own experience using data models at the insistence of those partial to them for original research. With the work in development by Hilton[10], it was desired to be known what characteristics of a city contributed to the implementation of a minimum wage ordinance. Framing the model in terms of classification, a city was given the value of 1 if it had implemented a minimum wage and 0 otherwise. Twenty-some covariates were explored for potential significance using a logistic regression. However, with only 100 cities, the degrees of freedom ran out quickly. Furthermore, of the 100 cities, only 15 had implemented a minimum wage. In implementing the logistic regression which was mandated by those overseeing the research, stepwise selection, forward selection, and backward selection were used and performed in both R and SAS. Of the six different models, not one resembled the other in terms of variables retained. Although they might share one or two variables, the picture was never clear.

Fitting a traditional logistic regression model to such data like this minimum wage could be useful. Assuming that the usual specifications are met, an economist or political scientist could say, "holding all else constant, raising w by v results in an $x\%$ increase in y". However, fitting a classification tree to this minimum wage data allowed for something more powerful. With messy data that skewed the logistic regression every which way, the classification tree allowed for 85% predictive accuracy and ordinal rankings of important factors. The algorithm and crossvalidation utilized simulation and advanced methods unlike the logistic regression.

[5] *Ibid.*, 208.

[6] *Ibid.*, 199.

[7] *Ibid.*, 202.

[8] *Ibid.*

[9] Wooldridge, Jeffrey M., 1960-. Introductory Econometrics : a Modern Approach. Mason, Ohio: South-Western Cengage Learning, 2012.

[10] Hilton, Nicholas S., "The Determinants of Minimum Wage Ordinances: An Analysis of the 100 Largest Cities from 2012-2017". This is my partner's Plan A Master's Thesis which has been successfully defended and to be published in USU Digital Commons in the comming months.

Although there are no concrete parameter values and p-values, the model much better represented the underlying phenomenon of interest.

Perhaps the most formidable claim is then that advanced algorthmic methods and other simulated experiments do not allow for certainty and understanding. Breiman certainly recognized this as he wrote, “Accuracy generally requires more complex prediction methods. Simple and interpretable functions do make the most accurate predictors”[11]. In wake of this reality, an economist may immentaly closed to such kind of modeling. However, there is another side to this coin equally important to emphasize: “Increased certainty is always bought at the expense of reduced scope”[12]. As Reiss elegant puts it, “we never know where the results hold once we leaved the model world”[13].

The question then becomes should one choose an interpretable model poorly fit to messy data or a more elusive model with better predictive accuracy. Given that traditional models require “idealizations, simplifications, approximations”, they may be a good start to begin some underlying theory as a *prior*[14]. Yet, when push comes to shove, it seems rather precarious to base inherently important policy decisions on models which are not entirely representative of a phenomena of interest and entail poor predictive accuracy when put to the test. Such models can, as Breiman puts it, “lead to irrelevant theory and questionable scientific conclusions”[15].

Furthermore, Buchanan’s problem with the optimization of economics may be better mended with such newer algorithmic methods. Although Hayek’s harkening is duly noted (that market complexity may never quite be harnessed by any statistical model) [16], simulations and advanced black box algorithms may better represent the “becoming” process Buchanan speaks of. Leaving behind the linearity and normality allows for formerly unimaginable model fitting capabilities that could better reflect finer intricacies of the market. All in all, Breiman’s perception of statistical analysis[17] appears to be a sound insight to both statistics and economics duly considered:

1) Focus on finding a good solution - that’s what consultants get paid for
2) Live with the data before you plunge into modeling
3) Search for a model that gives a good solution either algorithmic or data
4) Predictive accuracy on test sets is the criterion for how good the model is
5) Computers are an indispensable partner

If the discipline of not just statistics, but more so economics is willing to abide by these guidelines, the field can only be enriched much more. This does not mean to abandon the traditional paradigm, but to be open to that which in fact gives the best solution. A 1 out of 20 shot for identifying an insignificant result as significant falls short compared with saying “given new data, this model predicts with 95% accuracy”. If the goal is to understand more honestly a phenomena of interest all the while tackling new and interesting problems, methods nearly a century old may not be always appropriate.

[11] Breiman, Leo, “Statistcal Modeling: The Two Cultures” *Statistical Science* Vol. 16, No. 3, 2001, pp. 208.

[12] Reiss, Julian, “A Plea for (Good) Simulations: Nudging Economics Toward an Experimental Science”, *Simulation & Gaming* Vol 42, No. 2, pp. 249.

[13] *Ibid.*

[14] *Ibid.*, 254

[15] Breiman, Leo, “Statistcal Modeling: The Two Cultures” *Statistical Science* Vol. 16, No. 3, 2001, pp. 199.

[16] F. A. Hayek. The American Economic Review, Volume 35, Issue 4 (Sep., 1945), 519-530.

[17] Breiman, Leo, “Statistcal Modeling: The Two Cultures” *Statistical Science* Vol. 16, No. 3, 2001, pp. 201.