What Now? Some Brief Reflections on Model-Free Data Analysis[®]

Richard Berk

Department of Statistics
Department of Criminology
University of Pennsylvania

ABSTRACT

David Freedman's critique of causal modeling in the social and biomedical sciences was fundamental. In his view, the enterprise was misguided, and there was no technical fix. Far too often, there was a disconnect between what the statistical methods required and the substantive information that could be brought to bear. In this paper, I briefly consider some alternatives to causal modeling assuming that David Freedman's perspective on modeling is correct. In addition to randomized experiments and strong quasi-experiments, I discuss multivariate statistical analysis, exploratory data analysis, dynamic graphics, machine learning and knowledge discovery.

Key words: *Causal Modeling, Regression Analysis, Exploratory Data Analysis, Data Science*
JEL Classifications: C81, C50

1. INTRODUCTION

David Freedman began his professional career as a probabilist and mathematical statistician. He spent most of his career in the Department of Statistics at the University of California, Berkeley. His theoretical work was extremely broad, including important contributions to such topics as Martingale inequalities, Markov processes, de Finettis theorem, consistency of Bayes estimates, sampling, and the bootstrap. For these contributions, he was justly recognized. He was an elected fellow of Institute of Mathematical Statistics and of the American Statistical Association and was an elected member of the American Academy of Arts and Sciences. In 2003, he received John J. Carty Award for the Advancement of Science from the National Academy of Sciences "for his profound contributions to the theory and practice of statistics, including rigorous foundations for Bayesian inference and trenchant analysis of census adjustment."

The reference to census adjustment is telling. It highlights the kind of work for which David Freedman is probably most widely know, at least outside of discipline of statistics. Towards the middle of his professional career, teaching of statistics at various levels led him to introduce a variety of real-world applications into his courses. The more he examined such applications, the more concerned he became. In his view, there was commonly a demonstrable disconnect between what the statistical tools required and what applied researchers provided. The result was assume-and-proceed statistics presented as science.

But the problems were much deeper. The statisticians and econometricians responsible for the tools that were so often misapplied were frequently guilty of selling defective goods.

[®] Thanks go to Jasjeet Sekhon for helpful comments on an earlier draft of this paper.

Performance claims were unproved, or proved with convenient but unreasonable assumptions, or proved incorrectly. Moreover, such tools were often introduced into the research community with a hype usually associated with Hollywood productions or with an academic puffery that was at least unseemly. These overlays were as offensive to Professor Freedman as the shoddy work. The day before he died, when he knew he had not long to live, he urged his friends to "keep after the rascals."

I began working with David Freedman over a decade ago. We collaborated on a number of projects through which I received a welcomed and ongoing tutorial in statistics. I think I understand his perspective and much of his applied work. I do not think I can do it justice. In broad, summary form, his critiques of statistical practice can seem like old news. Stated that way, many of his concerns *are* old news. Much of David Freedman's impact is in the fine details of how, in his words, "the train jumped the tracks" (e.g., Freedman, 2008ab; Freedman and Berk, 2008) and in his back and forth with individuals who tried to deflect or dilute his critiques (e.g., Freedman and Humphries, 1999; Freedman, 2004). Equally important were his efforts to show the ways in which applied statistical research had been severely compromised in a wide variety of concrete settings. His concerns about statistical adjustments to the census were perhaps the best example (Brown et al., 2000; Freedman et al., 2001; Freedman and Wachter, 2001; 2003; Wachter and Freedman, 1996; 2000a; 2000b).

I think it is most appropriate to let David Freedman speak for himself in his past writings. Any summary that I could write risks trivializing his thinking. Rather, I will offer some ideas about where one might go from here if the core of his writings is accepted.

2. WHERE DO WE GO FROM HERE?

A model can be seen as a quantitative theory of how the data on hand were generated. A causal model is special case that conveys how the conditional distribution of a response variable will change when one or more causal variables are manipulated. "Casual parameters" contain this information (Heckman, 2000: 52-53).

Imagine a world of applied statistics and applied econometrics in which causal modeling was only initiated when one could write down a nearly right model of how the data were generated. That is, imagine one where we are able to do what the standard statistics textbooks require. Some very light tuning of the model might be allowed, but the main tasks would be estimating the values of some interesting parameters and characterizing the impact of uncertainty on those estimates. I suspect that for the social and biomedical sciences, such a world is many years in the future, but for nearly-right models, many of Freedman's applied concerns evaporate[1]. How do we get to such a world? Or, if we don't want to get there, or do not believe it is possible to get there, what are the alternatives?

2.1. Experiments and Quasi-Experiments

One alternative to or precursor for conventional causal modeling has been around at least since the 1960s and needs little discussion: randomized field experiments. Although not the "gold standard" some claim−there is no gold standard−randomized field experiments can be very powerful if causal inference is the primary concern (Berk, 2005). And the analysis of such experiments can be and should be largely model-free (Freedman, 2006).

[1] Some will no doubt claim that we are already there. They need to read and digest Freedman's work (e.g., Freedman, 2005) and those of a large number of skeptics Freedman cites.

Strong quasi-experiments, such as those based on a regression-discontinuity design, have also been used since the 1960s. They fall between randomized experiments and observational studies and can sometimes produce credible casual inferences (e.g., Imbens and Lemieux, 2008). They also do not need to be discussed here. They too can be and should be analyzed in a largely model-free fashion (Imbens, 2004; Rubin, 2008).

Some might argue that randomized experiments and strong quasi-experiments are ends in themselves. They are not a step toward nearly-right models. This is not the venue in which examine the issues. Suffice it say, there are good arguments on both sides.

What does one do if there are no nearly-right models for the problem at hand, and randomized experiments or strong quasi-experiments are not feasible or are premature? Perhaps then, the pursuit of cause-and-effect needs to postponed. Perhaps "mere" description, which currently can be undertaken in many different ways, is then an appropriate alternative.

2.2. Multivariate Statistical Analysis

There was a time when multivariate statistical analysis was central in many of the social and biomedical sciences (Anderson, 1958). Put somewhat too simply, the goal was to find systematic patterns in the data without much concern about cause and effect. Statistical inference was common, often capitalizing on the multivariate normal distribution. Thus, models played a central role, but the models were usually not causal. Classification tools such as discriminant function analysis were popular. Data reduction tools such as factor analysis, clustering, canonical correlation and multidimensional scaling were also common.

Over the past two decades the underlying machinery of such procedures has become increasingly sophisticated so that the field is now populated with such tools as multiple correspondence analysis, categorical and nonlinear principal components analysis, nonlinear generalized canonical correlation analysis, and projection pursuit (Gifi, 1990). Description remains primary. But for some flavors of multivariate statistical analysis, statistical inference and the models it requires are given far less prominence.

> A technique is defined, more or less, as a tool into which you feed data of a particular type and format, and which then reproduces output of a particular type and format. This very instrumental interpretation of data analysis is contrasted, in many places, with the model-oriented approach of classical statistics. Models do have a place in Gifi's philosophy: not as tentative approximations to the truth, but as devices to sharpen and standardize data analytic techniques… (Gifi, 1990: v)[2].

Loosely translated, Gifi's position is that multivariate statistical analysis should be about organizing, summarizing, and displaying data to aide in the search for structure. The enterprise is inductive, significantly because deductive approaches depend far too heavily on assumed truths that are not demonstrably true. As such, multivariate statistical analysis has much in common with exploratory data analysis to which we turn next.

[2] The author A. Gifi is actually Jan de Leeuw, Distinguished Professor of Statistics at UCLA. The statement just quoted is rather gentle compared to his current views. Of late, he sometimes speaks of models as "codified prejudice." (personal communication).

2.3. Exploratory Data Analysis

Exploratory data analysis (EDA) has evolved from rather different traditions than multivariate statistical analysis, but the overlap can be striking. Persi Diaconis observes that

> Exploratory data analysis (EDA) seeks to reveal structure, or simple description. We look at numbers or graphs and try to find patterns. We pursue leads suggested by background information, imagination, patterns perceived, and experience with other data analyses (Diaconis, 1985: 1)

Unlike multivariate statistics, EDA has never relied heavily on models of any sort, and there has been greater emphasis on clever, low tech approaches and visualization methods. John Tukey, the statistician most associated with EDA, says it this way.

> This book is about exploratory data analysis, about looking at data to see what it seems to say. It concentrates on simple arithmetic and easy-to-draw pictures. It regards whatever appearances we have recognized as partial descriptions, and tries to look beneath them for new insights. Its concern is with appearance, not confirmation (Tukey, 1977: v)

The problem with using models is not that they are too often fatuous. The problem is that they are too often premature. The descriptive goals of EDA are seen as necessary prerequisites to the development and use of models, often causal models. EDA is in part offered as a way to help identify problems in the data that could undermine model-based approaches and formal statistical tests; diagnostics are an important component of the EDA tool kit (Dasu and Johnson, 2003). Finally, the walls around EDA are very permeable so that some kinds of EDA are actually efforts to make traditional model-based procedure more robust (Hoaglin et al., 2000; 2006).

2.4. Dynamic Graphics

A third source of descriptive tools and procedures is found in the recent developments around computer driven visualization and dynamic graphics. This work builds substantially on earlier graphical approaches to data (Cleveland, 1993), EDA and multivariate statistics so that all three traditions are sometimes combined in the same procedure. For example, drawing on and dramatically extending some of the ideas from principal component analysis, there are a variety of projection methods that can reduce the dimensionality of a given data set. Linear combinations of variables are constructed in response to a variety of constraints and loss functions. Orthogonality, for instance, may or may not be imposed on the linear combinations, and there are loss functions that can combine variables so that clusters of observations in the projection space are revealed (Cook et al, 2007: 26-34). One is not limited to combining variables that are highly correlated.

But there is more. There is virtually a limitless number of projections one can construct from a given data set, and one is not formally constrained to projections in three dimensions or less. A means must be found to translate the information contained in such projections into a digestible form. Recent developments in computer technology can help.

> Unlike the passive paper medium, computers allow us to manipulate pictures, to pull and push their content in continuous motion like a moving video camera, or

> to poke at objects in one picture and see them light up in other pictures. This book features many illustrations of the power of these linking technologies. The diligent reader may come away ``seeing" high-dimensional data spaces! (Cook et al, 2007:3)

In short, increasing computing power and new algorithms can be harnessed to provide instructive visualizations of data analysis procedure combining many of the strengths of multivariate statistical analysis and exploratory data analysis. A key is being able to interact in real time with visual summaries of the data. An interesting byproduct is an evolving skill in how one looks at and thinks about high dimensional data.

2.5. Machine Learning

A fourth source of descriptive approaches is very recent developments in machine learning, also called statistical learning (Bishop, 2006; Berk, 2008; Hastie et al., 2009). It has its roots in computer science as well as statistics. Superficially, one can view much of this work as function estimation. One assumes something like $Y = f(X)+\varepsilon$, and the goal is to determine the $f(X)$. As such, it has the same look and feel as nonparametric or semiparametric regression that have been popular in some circles for well over a decade. But that neglects several key points.

1. To estimate the $f(X)$, one needs X. All of the usual omitted variables problems reappear. One also needs to make the usual sorts of regression assumptions about ε. And there are rarely any claims that the $f(X)$ has a causal interpretation to begin with. In practice, therefore, these procedures are less about function estimation and more about description.

2. The loss functions can be highly unconventional. Under support vector machines, for example, the loss function for binary outcomes is shaped much like a hockey stick (Berk, 2008: 314).

3. There is nothing in the usual output requiring a causal story. For example, the “importance” of predictors is often represented as contributions to forecasting skill. A variable that forecasts well may or may not have any causal connection to the response.

4. Statistical inference is rarely employed and is likely to be fundamentally misguided in any case (Leeb and Pötscher, 2005; 2006; 2008).

5. The algorithms are often novel. In boosting, for instance, there can be thousands of passes through the data. After each pass, the data are reweighted so that observations that are fitted less well are given greater weight before fitting procedure is reapplied. The final collection of fitted values is a weighted average of each set of earlier fitted values, with these weights a function of the quality of the fit for that set. Also, there is no search for convergence in the usual sense (Berk, 2008: 258-259).

6. In addition to “supervised learning” for which there is a response variable, there is “unsupervised” learning in which there is no response variable. Thus, machine/statistical learning applies much the same approach whether in a setting that looks like regression or in a setting that looks like clustering.

To summarize, machine/statistical learning in practice can be seen as a form of data exploration in which description is the primary goal. Many of the procedures work well on very high dimensional data, and some work well even with there are more predictors than observations (e.g., random forests). They are also often well suited for the analysis of enormous data sets with hundreds of thousands of observations. For example, parallelization is often straightforward.

2.6. Knowledge Discovery

"Knowledge discovery" is really just a way some computer scientists and applied mathematicians summarize the series of research processes that are at least implicit in much empirical science. The enterprise begins with developing substantive expertise in the research area, proceeds through data collection and cleaning, then to data analysis, and finally to interpretation of the results. For purposes of this discussion, the key feature of knowledge discovery is the data analysis, sometimes broadly characterized as "data mining", which is the search for patterns in the data. Thus, data mining includes virtually any quantitative data analysis having a significant inductive component (Maimon and Rokach, 2005). Each of the descriptive approaches just described qualify.

There are many additional inductive approaches from computer science having a distinctive style. Perhaps the most well known are various query methods used with large, complex data sets. For example, how might one discover and then characterize the collect of consumer goods that certain kinds of individuals purchase (Höppner, 2005; Boulicaut and Masson, 2005)? Should one define the collection of goods through the estimated marginal probability of each item being purchased. And if so, what might be a loss function to determine membership? Or should the estimated joint probability of all possible bundles of purchased items use used? And if so, how best does one efficiently search through enormous data bases to consider all possible bundles? Query methods are essentially about how to pose clear, well-bounded questions to a data set and then about algorithms for finding the answers.

More provocative is very recent work in which the application of computer algorithms to data on physical system apparently finds not just regularities, but regularities that are essentially rediscoveries of known physical laws. In a recent issue of *Science*, the authors explain:

> Our goal is to find natural relations where they exist, with minimal restrictions on their analytical form (i.e., free-form). Many methods exist for modeling scientific data: Some use fixed-form parametric models derived from expert knowledge, and others use numerical models (such as neural networks) aimed at prediction. Still others have explored restricted model spaces using greedy monomial search. Alternatively, we seek the principal unconstrained analytical expression that explains symbolically precise conserved relations, thus helping distill data into scientific knowledge (Schmidt and Lipdson, 2009:81).

The algorithm proceeds in six steps (Schmidt and Lipdson, 2009: 82).

1. Collect experimental data.

2. Numerically calculate partial derivatives for every pair of variables.

3. Generate candidate symbolic functions. They are random at first. Subsequently, small variations on the best equations are selected (in step 5).

4. Derive the symbolic partial derivatives for each pair of variables implied by each candidate equation.

5. Compare the derived partial derivatives to the corresponding numerical partial derivatives and select the best equations.

6. When the fit between the two sets of partial derivatives is sufficiently good, return the most parsimonious equations.

In applications to date, this algorithm correctly rediscovers several non-trival physical laws. Successes with somewhat similar implications are reported in the same issue of *Science* by other authors working on biological systems. The intent is to automate routine science so that researchers can be released to undertake more creative and challenging tasks.

> A natural extension of the trend to ever-greater computer involvement in science is the concept of a robot scientist. This is a physically implemented laboratory automation system that exploits techniques from the field of artificial intelligence to execute cycles of scientific experimentation. A robot scientist automatically originates hypotheses to explain observations, devises experiments to test these hypotheses, physically runs the experiments by using laboratory robotics, interprets the results, and then repeats the cycle (King et al, 2009: 85)

With these developments and more to come, one has to wonder if we are falling into a trap somewhat like the one that David Freedman identified so well. Are we being seduced by technological advances that can undermine the very activities that good science requires?

3. CONCLUSIONS

David Freedman believed that significant portions of the social and biomedical sciences had been hijacked by causal modeling tools that too many researchers did not understand and that were in many cases misrepresented to begin with. The solution was not better modeling tools−that response had failed too many times already−but a fundamental reconsideration of the enterprise. That reconsideration would be based heavily on looking back to successful social and biomedical scientific practice in the past. That past depended on many different kinds of data, both qualitative and quantitative, a gradual accumulation rich descriptive material, opportunistic exploitation of natural variation and natural experiments, true experiments and strong quasi-experiments where possible, and a thoughtful integration of the information obtained in a context of real substantive expertise. At the same time, he was not hostile to technological advances that could further the enterprise. Over the years, he integrated many into his research. In that spirit, a case can be made for continuing the support of randomized experiments and strong quasi-experiments, and for resurrecting scientific description as one essential element in the social and biomedical sciences, aided where possible by recent technical developments.

REFERENCES

Anderson, T.W. (1958). *An Introduction to Multivariate Statistical Analysis*, First Edition. New York: John Wiley and Sons.

Berk, R.A. (2005). Randomized Experiments as the Bronze Standard. *Journal of Experimental Criminology*, 1(4): 417-433.

Berk, R.A. (2008). *Statistical Learning from a Regression Perspective*, New York: Springer.

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.

Brown, L.D., M.L. Eaton, D.A. Freedman, S.P. Klein, R.A. Olshen, K.W. Wachter, M.T. Wells and D. Ylvisaker (1999). Statistical Controversies in Census 2000. *Jurimetrics*, 39: 347-375.

Boullicaut, J-F., and C. Masson (2005). “Data Mining Query Languages”, (in: O. Maimon and L. Rokach -Ed. *The Data Mining and Knowledge Discovery Handbook*), New York: Springer.

Cleveland, W.E. (1993). *Visualizing Data*. Summit, Jew Sersey: Hobart Press.

Cook, D., Swayne, D.F., Buja, A., and T. Lang (2007). *Interactive Dynamc Graphics and Data Analysis*. New York: Springer.

Dasu, T., and T, Johnson (2003). *Exploratory Data Mining and Data Cleaning*. New York: John Wiley and Sons.

Diaconis, P. (1985). “Theories of Data Analysis: From Magical Thinking through Classical Statistics”. (in: D.C. Hoaglin, F. Mosteller, and J. Tukey -Ed., *Exploring Data Tables, Trends, and Shapes*), New York: John Wiley and Sons.

Freedman, D.A. (2004). On Specifying Graphical Models for Causation, and the Identification Problem. *Evaluation Review*, 26: 267-293.

Freedman, D.A. (2005). *Statistical Models Theory and Practice*. Cambrige University Press.

Freedman, D.A. (2006). Statistical Models for Causation: What Inferential Leverage Do They Provide? *Evaluation Review*, 30: 691-713.

Freedman, D.A. (2008a). Randomization Does not Justify Logistic Regression. *Statistical Science*, 23: 237-249.

Freedman, D.A. (2008b). On Regression Adjustments to Experimental Data. *Advances in Applied Mathematics*, 40: 180-193.

Freedman, D.A. and R.A. Berk (2008). On Weighting Regressions by Propensity Scores. *Evaluation Review*, 32: 392-409.

Freedman, D.A. and P. Humphreys (1999). Are There Algorithms That Discover Causal Structure? *Synthese*, 121: 29-54.

Freedman, D.A., P.B. Stark and K.W. Wachter (2001). A Probability Model for Census Adjustment. *Mathematical Population Studies*, 9: 165-180.

Freedman, D.A. and K.W. Wachter (2001). Census Adjustment: Statistical Promise or Statistical Illusion? *Society*, 39: 26-33.

Freedman, D.A. and K.W. Wachter (2003). "On The Likelihood of Improving the Accuracy of The Census Through Statistical Adjustment". (in D. R. Goldstein -Ed., *Science and Statistics: A Festschrift for Terry Speed.* Institute of Mathematical Statistics Monograph 40: 197-230.)

Gifi, A. (1990). *Nonlinear Multivariate Analysis*, New York: John Wiley and Sons.

Hastie, T., R. Tibshirani and J. Friedman (2009). *The Elements of Statistical Learning*, Second Edition. New York: Springer.

Heckman, J. (2000). Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective. *Quarterly Journal of Economics*, 88 (2): 47-97.

Hoaglin, D.C., F. Mostellor and J. Tukey (2006), *Understanding Robust and Exploratory Data Analysis*, Wiley Classics Library. New York: John Wiley and Sons.

Hoaglin, D.C., F. Mostellor and J. Tukey (2000). *Exploring Data Tables, Trends, and Shapes*, Wiley Classics Library. New York: John Wiley and Sons.

Höppner, F. (2005). "Association Rules", (in O. Maimon and L:. Rokach -Ed., *The Data Mining and Knowledge Discovery Handbook*), New York: Springer.

Imbens, G. (2004). Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. *The Review of Economics and Statistics*, 86 (1): 4-29.

Imbens, G. and T. Lemieux (2008). Regression Discontinuity Designs: A Guide to Practice. *Journal of Econometrics*, 142: 611-614.

King, R.D., J. Rowland, S.G. Oliver, M. Young, W. Aubrey, E. Byrne, M. Liakata, M. Markham, P. Pir, L.N. Soldatova, A. Sparkes, K.K. Whelan and A. Clare (2009). The Automation of Science. *Science*, 324 (3): 85-89.

Leeb, H. and B.M. Pötscher (2005). Model Selection and Inference: Facts and Fiction, *Econometric Theory*, 21: 21–59.

Leeb, H. and B.M. Pötscher (2006). Can one Estimate the Conditional Distribution of Post-Model-Selection Estimators? *The Annals of Statistics*, 34 (5): 2554–2591.

Leeb, H. and B.M. Pötscher (2008). "Model Selection", (in T.G. Anderson, R.A. Davis, J.P. Kreib, and T. Mikosch -Ed., *The Handbook of Financial Time Series*), New York, Springer: 785–821.

Maimon, O., and L. Rokach (2005). *The Data Mining and Knowledge Discovery Handbook*. New York: Springer.

Rubin, D.B. (2008). For Objective Causal Inference, Design Trumps Analysis. *Annals of Applied Statistics*, 2 (1): 808-840.

Schmidt, M., and H. Lipson (2009). Distilling Free-Form Natural Laws from Exeprimental Data. *Science*, 324 (8): 81-85.

Tukey, J. (1977). *Exploratory Data Analysis*. Reading, MA: Addison -Wesley.

Wachter, K.W. and D.A. Freedman (2000). The Fifth Cell: Correlation Bias in U.S. Census Adjustment. *Evaluation Review*, 24: 191-211.

Wachter, K.W. and D.A. Freedman (2000). Measuring Local Heterogeneity with 1990 U.S. Census Data. *Demographic Research*, 3 (10): 1-22.

Wachter, K.W. and D.A. Freedman (1996). Planning for the Census in the Year 2000. *Evaluation Review*, 20: 355-377.