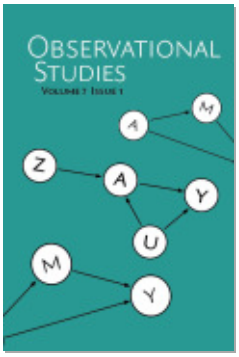
# Breiman's Two Cultures: A Perspective from Econometrics

Guido Imbens, Susan Athey




Observational Studies, Volume 7, Issue 1, 2021, pp. 127-133 (Article)

Published by University of Pennsylvania Press
*DOI:* *https://doi.org/10.1353/obs.2021.0028*

Observational Studies Vol. 7.1 (2021): Published July 2021

# Breiman's Two Cultures: A Perspective from Econometrics

**Guido Imbens** **imbens@stanford.edu**
*Graduate School of Business and Department of Economics*
*Stanford University*
*Stanford, CA 94305*

**Susan Athey** **athey@stanford.edu**
*Graduate School of Business*
*Stanford University*
*Stanford, CA 94305*

## Abstract

Breiman's "Two Cultures" paper painted a picture of two disciplines, data modeling, and algorithmic machine learning, both engaged in the analyses of data but talking past each other. Although that may have been true at the time, there is now much interaction between the two. For example, in economics, machine learning algorithms have become valuable and widely appreciated tools for aiding in the analyses of economic data, informed by causal/structural economic models.

**Keywords:** Econometrics, Causal Inference, Instrumental Variables

## 1. Introduction

When we first read Breiman's "Two Cultures" paper (Breiman, 2001) in the early 2000's it baffled us. In econometrics, as in much of statistics (98% according to Breiman), the modeling culture was dominant, and a purely prediction-based focus seemed very alien to what econometricians were doing. You can see this clearly in the history of econometrics (see for a description, *e.g,* Hendry and Morgan (1997)). From the founding days of the field many researchers were more focused on identification and estimation of causal effects (*e.g.,* the parameters of structural models) than on prediction. This causality-based focus led econometricians to de-emphasize $R^2$ values as measures of success, and instead aim to build a credible case for estimation of causal effects. To illustrate the difference between the predictive and causal approaches, we discuss two examples. Both examples are part of what Josh Angrist and Steve Pischke later called the credibility revolution (Angrist and Pischke (2010)) that since the late 1980s has been a major influence in empirical work in economics. Then we discuss how more recently researchers in econometrics have started appreciating the benefits of the algorithmic machine learning approaches. A rapdily growing literature attempts to combine the benefits of the prediction-focused machine learning algorithm approaches with the traditional focus on causal model-based approaches.

The first illustration focuses on the problem of estimating supply and demand functions, an important example of a simultaneous equations problem. The study of such problems goes back to the founding of econometrics as a separate discipline (see the example of the demand for potato flour in Tinbergen (1930), with a translation in Hendry and Morgan

(1997) and further discussion in Imbens). We illustrate this with an application taken from Angrist et al. (2000). They study the demand for fish as a function of price, using daily data from the Fulton fish market in New York. The demand function is a causal object, defined as the quantity as a function of the price that buyers would be willing to buy if the price was set exogenously. We denote the average demand function, averaged over buyers, by the potential outcome function $Q_t^d(p)$, where $t$ indexes the markets, days in this illustration. It is of fundamental interest in economics. For example, it is of interest to policy makers who may be interested in the effect of different market structures (*e.g.,* imposing a tax), or to sellers, who may be interested in the effect of charging higher prices. It is quite different from the predictive relation between quantities and prices. The latter may be of interest for different purposes. In the Angrist et al study the researchers have observations for a number of days at the Fulton fish market. The two main variables observed by the researcher are the quantity of fish traded, $Q_t$, and the average price at which it was traded on that particular day, $P_t$. Figure 1 shows the data, with each dot denoting the log quantity and log price combination for a particular day at the Fulton fish market. What should we do with such data? To estimate the demand function, one might be tempted to try tofit a *predictive* model, predicting the quantity sold as a function of the price. The best predictor would be the conditional expectation $E[Q_t|P_t = p]$. But for an economist such an exercise would make little sense as an attempt to estimate the demand function. There is little reason to believe that the conditional expectation of the quantity as a function of price, no matter how cleverly estimated, and no matter how well it predicts (how small the residual sum of squares), would correspond to the demand function. Price are not set randomly, not in this particular market, and not in most markets. To make sense of these data on quantities and prices and what they reveal about the demand function, one needs an economic model that explains how the demand function relates to the data we see, and in particular why prices took on the values that were observed. In this case a standard, perhaps too simple, economic model is that in addition to the demand function $Q_t^d(p)$, presumably decreasing in price, there is a supply function that relates price to the quantity supplied:

$$Q_t^s(p),$$

presumably increasing in price. The final piece of the economic model is the market clearing assumption that the price and quantity we actually see on day $t$ is the equilbrium/market-clearing price $P_t$ that equates supply and demand

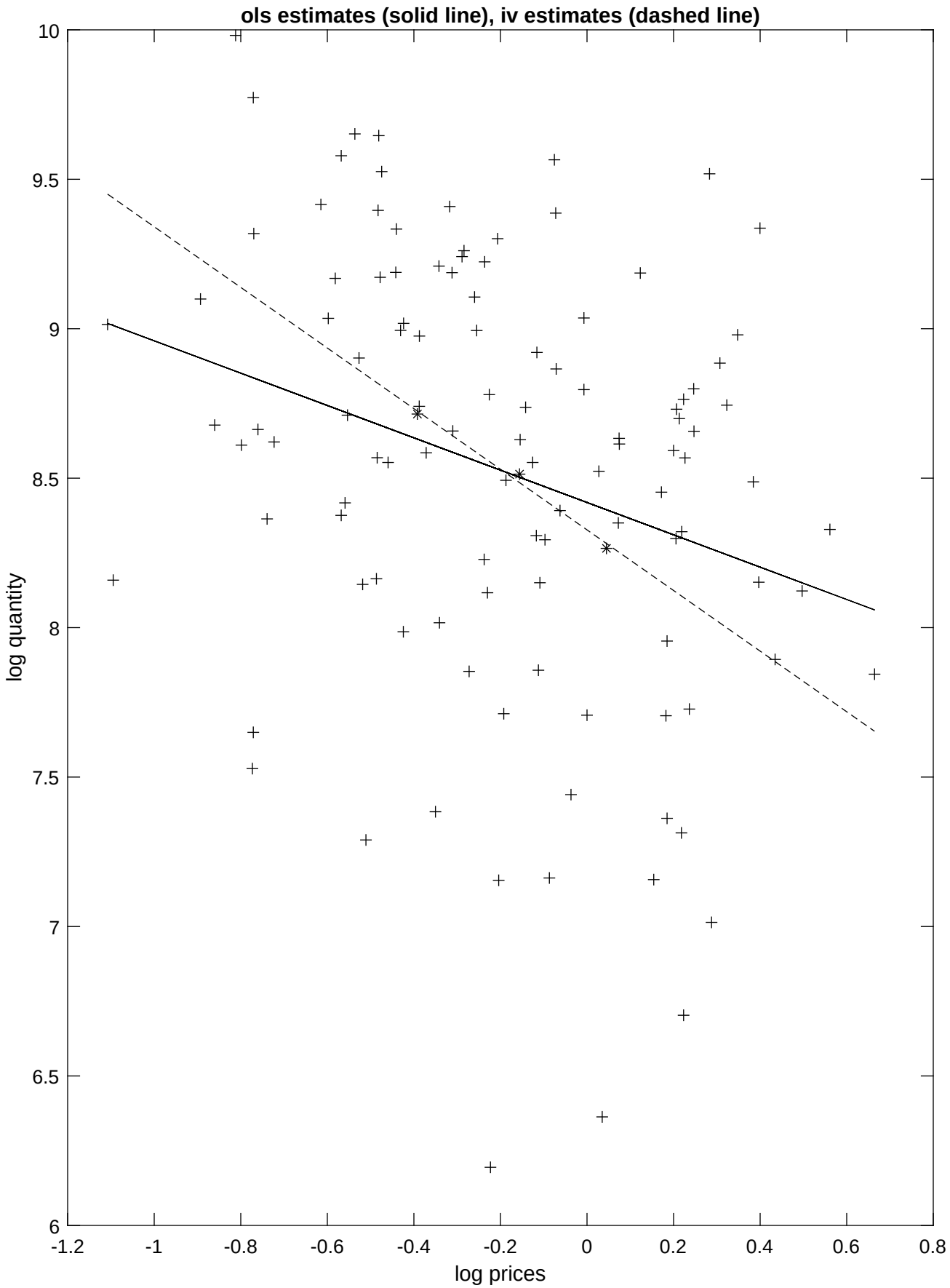
$$P_t \text{ solves } Q_t^d(p) = Q_t^s(p),$$

and that the observed quantity $Q_t$ is equal to the supply and demand at the equilibirum price:

$$Q_t = Q^d(P_t) = Q^s(P_t).$$

Under fairly mild conditions these equilibrium prices and quanties will be unique. The relation between the observed quantities and prices combines the supply and demand function. To separate the supply and demand functions econometricians often rely on instrumental variables methods. In this particular case (Angrist et al., 2000) use weather conditions (wave height and wind speed) at sea as instruments that directly affects supply, but does

# Figure 1



not directly affect demand. The traditional approach would then assume that the demand function was linear in logarithms,

$$\ln(Q_t^d(p)) = \alpha + \beta \ln(p) + \varepsilon_t,$$

lead to an instrumental variables estimator for the price elasticity, $\beta$, equal to

$$\hat{\beta}_{\text{iv}} = \frac{\text{Cov}(Q_t, z_t)}{\text{Cov}(P_t, z_t)}.$$

The solid line in Figure 1 presents the regression line from a least squares regression of $Q_t$ on $P_t$, with a slope of -0.54. The dashed line presents the instrumental variables estimate of the demand function, with a slope of approximately -1.01. Whether the price explains much of the quantity traded here is not viewed as being of central importance. Certainly one can obtain a better fit, as measured by the residual sum of squares, by simple linear regression of the quantity on the price. However, that would not be viewed as meaningful here by most economists for the purpose of estimating the demand function. The stars in this figure represent the average log quantities and log prices on days where the weather was fair, mixed, or stormy. The instrumental variables estimates are essentially trying to fit a straight line through these points.

A second example is the returns to education. There is a large literature in economics devoted to estimating the causal effect of formal education (as measured by years of education) on earnings (see Card (2001) for a general discussion). For individual $i$ let $Y_i$ be the logarithm of earnings, and $X_i$ be years of education, and let $Y_i(x)$ be the potential earnings function for this individual, measuring log earnings for this individual if this individual were to receive level of education equal to $x$. Initially researchers would estimate the returns to education by estimating a linear regression of log earnings on years of education. Much of the literature has been concerned with the fact that educational choices are partly driven by unobserved individual characteristics (say, unobserved skills) that may be related to the potential earnings outcomes. As a result the linear regression of log earnings on years of education may be biased for the causal effect of education on log earnings, even after conditioning on observed individual characteristics. Building a better predictive model does not directly deal with this concern because it cannot adjust for unobserved covariates. Angrist and Krueger (1991) propose a clever research strategy to estimate the causal effect of education without this omitted variable bias. They suggest using compulsory schooling laws as an instrument. The idea is that compulsory schooling laws exogenously shift education levels without directly affecting earnings. In practice, of course compulsory schooling laws explain very little of the variation in education levels, so little that Angrist and Krueger needed to use Census data in order to get precise estimates. So, from a predictive perspective compulsory schooling laws appear to be largely irrelevant for modeling earnings. But, there is a reasonable argument that the compulsory schooling laws generate variation in education levels that is not associated with the unobserved skills that create the biases in least squares regressions of log earnings on years of education. In other words, there is a reasonable argument that it is a valid instrument in the sense of satisfying the exclusion restrictions (Imbens and Angrist (1994); Angrist et al. (1996)).

These two examples are essentially an elaboration of the point that David Cox (Cox, 2001) and Brad Efron (Efron, 2001) make in their comments on (Breiman, 2001) that much

of statistics is about causal effects of interventions, rather than predictions. This distinction may often be implicit, but that does not take away from the fact that causal effects are the ultimate goal. This view may have been part of the reason the econometrics community was initially slow in adopting the algorithmic methods that Breiman was advocating. However, although perhaps slower than one might have hoped, many of these methods are now enthusiastically been adopted in econometrics, from deep learning methods to generative adversarial networks (Athey et al. (2019a); Kaji et al. (2019)). See Athey and Imbens (2019) for a general discussion of the use of these methods in economics. A key insight is that although economic theory may be helpful, and in fact essential, for part of the model (*e.g.,* the exclusion restrictions that are the core of the instrumental variables methods, or the equilibrium assumptions that underly supply and demand models), there are parts of the model where economic theory is silent, and where the modern machine learning methods can be extremely effective in assisting in model specification, substantially more so than traditional econometric methods. The challenge is to incorporate the economic causal restrictions and non-prediction objectives into the algorithms.

Let me discuss two examples of this integration of machine learning methods into causal modeling. First, there is a large literature focusing on estimating average treatment effects under ignorable treatment assignment (Rosenbaum and Rubin (1983)). Under the ignorability/unconfoundedness assumption the target (the average treatment effect) can be written as a functional of number of conditional expectations, that of the outcome given the treatment and covariates, and that of the treatment given the covariates (the propensity score). Traditionally these conditional expectations were estimated using nonparametric regression methods. Building on Robins et al. (1994) that introduced double robust estimation, Van der Laan and Rose (2011); Chernozhukov et al. (2017); Athey et al. (2018) and others use algorithmic machine learning methods for recovering these conditional expectations. These methods are more effective at doing so than the traditional econometric methods, leading to more accurate estimate of the average treatment effect.

Second, a literature has developed using machine learning techniques to estimate average treatment effects conditional on observable characteristics in a variety of settings, including those where instrumental variables can be used to estimate treatment effects. For example, Athey et al. (2019b) develop a generalized random forest method that targets treatment effect heterogeneity. Building on an application by Angrist and Evans (1998), Athey et al. (2019b) analyze the question of how having additional children affects the labor supply of women. The instrumental variable is an indicator for whether a woman's first two children were of the same gender. Similarly, Hartford et al. (2017) make use of neural nets to analyze conditional average treatment effects in instrumental variables settings.

Since the publication of the Breiman paper much progress has been made. Modellers have embraced many of the algorithms developed in the machine learning literature. The algorithm builders have expanded beyond the original prediction problems and are now actively exploring methods for including causal objectives and restrictions into their algorithms using both graphical (Pearl (2000)) and potential outcome perspectives (Imbens and Rubin (2015)), and going into new directions such as causal discovery (Peters et al. (2017)). The two cultures have found they have much in common and much to learn from each other.

## Acknowledgments

Generous support from the Office of Naval Research through ONR grant N00014-17-1-2131 is gratefully acknowledged.

## References

Joshua D Angrist and William N Evans. Children and their parents' labor supply: Evidence from exogenous variation in family size. *American Economic Review*, pages 450–477, 1998.

Joshua D Angrist and Alan Krueger. Does compulsory schooling affect schooling and earnings. *Quarterly Journal of Economics*, CVI(4):979–1014, 1991.

Joshua D Angrist and Jörn-Steffen Pischke. The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of economic perspectives*, 24(2):3–30, 2010.

Joshua D Angrist, Guido W Imbens, and Donald B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91:444–472, 1996.

Joshua D Angrist, Kathryn Graddy, and Guido W Imbens. The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. *The Review of Economic Studies*, 67(3):499–527, 2000.

Susan Athey and Guido W Imbens. Machine learning methods that economists should know about. *Annual Review of Economics*, 11:685–725, 2019.

Susan Athey, Guido W Imbens, and Stefan Wager. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):597–623, 2018.

Susan Athey, Guido W Imbens, Jonas Metzger, and Evan M Munro. Using wasserstein generative adversarial networks for the design of monte carlo simulations. Technical report, National Bureau of Economic Research, 2019a.

Susan Athey, Julie Tibshirani, Stefan Wager, et al. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019b.

Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

David Card. Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica*, 69(5):1127–1160, 2001.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65, 2017.

David R Cox. [statistical modeling: The two cultures]: Comment. *Statistical Science*, 16 (3):216–218, 2001.

Brad Efron. [statistical modeling: The two cultures]: Comment. *Statistical Science*, 16(3): 218–219, 2001.

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep iv: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*, pages 1414–1423. PMLR, 2017.

David F Hendry and Mary S Morgan. *The foundations of econometric analysis*. Cambridge University Press, 1997.

Guido W Imbens. Book review: The foundations of econometric analysis by hendry and morgan.

Guido W Imbens and Joshua D Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 61:467–476, 1994.

Guido W Imbens and Donald B Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

T Kaji, Elena Manresa, and Guillaume Poulio. Artificial intelligence for structural estimation. Technical report, New York University, 2019.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, USA, 2000. ISBN 0-521-77362-8.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Jan Tinbergen. Determination and interpretation of supply curves: an example. *Zeitschrift fur Nationalokonomie*, 1(5):669–679, 1930.

Mark J Van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.