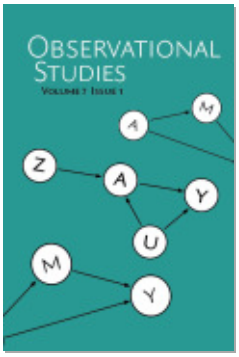
Statistical Modelling in the Age of Data Science

Stijn Vansteelandt



Observational Studies, Volume 7, Issue 1, 2021, pp. 217-228 (Article)

Published by University of Pennsylvania Press
DOI: *https://doi.org/10.1353/obs.2021.0013*

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

Statistical modelling in the age of data science

Stijn Vansteelandt **stijn.vansteelandt@ugent.be**
Department of Applied Mathematics, Computer Science and Statistics
Ghent University, 9000 Ghent, Belgium
and Department of Medical Statistics
London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK

Abstract

Twenty years after Leo Breiman's wake-up call on the use of data models, I reconsider his concerns, which were heavily influenced by problems in prediction and classification, in light of the much vaster class of problems of estimating effects and (conditional) associations. Viewed from this perspective, one realises that the statistical community's commitment to the use of data models continues to be dominant and problematic, but that algorithmic modelling (machine learning) does not readily provide a satisfactory alternative, by virtue of being almost exclusively focused on prediction and classification. The only successful way forward is to bridge the two cultures. It requires machine learning skills from the algorithmic modelling culture in order to reduce model misspecification bias and to enable pre-specification of the statistical analysis. It moreover requires data modelling skills in order to choose and construct interpretable effect and association measures that target the scientific question; in order to identify those measures from observed data under the considered sampling design by relating to minimal and well-understood assumptions; and finally, in order to reduce regularisation bias and quantify uncertainty in the obtained estimates by relating to asymptotic theory.

Keywords: Causal machine learning, Data science, Double machine learning, Targeted learning

1. Introduction

It has been a joy to reread the late Prof. Breiman's provocative paper (Breiman, 2001) and to see how relevant, timely and important his wake-up call continues to be twenty years later. Surely much has changed in our profession over those twenty years. But the fact that new and important developments on 'big data' and 'data science' are taking place largely outside statistics, should remind us of Breiman's alert that *'[the growth in algorithmic modelling applications and methodology] has occurred largely outside statistics in a new community - often called machine learning - that is mostly computer scientists'*. It signals once more an (at least partial) failure of our statistics profession to embrace new developments and engage (young) scientists with diverse backgrounds who are interested in data analysis.

In my opinion, a core reason for this lies in the predominance of mathematics in our discipline, which is both a blessing and a curse. It can be a curse in basic statistics education which, despite having become less mathematics-focused over the past twenty years, still has some way to go in stimulating and preparing students to 'think with data' (Hardin et al., 2015). It can be a curse in statistics research, which can be conservative and resistant

towards algorithmic methods while they are lacking a strong mathematical foundation, and can moreover be lured into elegant mathematical curiosities that have little or no relevance for data analysis. It can be a curse in decision processes for awarding funding, which can be biased towards the more mathematically-oriented project proposals that may seem to promise greater quality.

The mathematical foundation of our discipline is indisputably also a blessing, however, as also various discussants of Breiman (2001) express. Let me note in particular the increasingly common use of machine learning for estimating association and effect measures (as opposed to making predictions) (van der Laan and Rubin, 2006; van der Laan and Rose, 2011; Chernozhukov et al., 2018), a profound development which I will discuss in this article and which is precisely dependent upon asymptotic theories from mathematical statistics (Wager and Athey, 2018). It is eminently here where Breiman's two cultures meet, and where the skills of both 'data modellers' (statisticians) and 'algorithmic modellers' (computer scientists) are indispensable. It is precisely at this vast intersection of machine learning and estimating association and effect measures that statisticians can be at the forefront of data science.

2. The two cultures

Breiman's description of two model cultures appears strongly influenced by his experience as a consultant working on problems of prediction, and by the era, twenty years ago, when the use of machine learning was entirely focussed on prediction and classification. I agree with other discussants of Breiman (2001) that prediction problems constitute only a minority of the problems that scientists face. A majority of empirical studies rather aims to develop insight into the effects of exposures on certain outcomes, into causal mechanism, or at a less ambitious level, into associations (e.g., detecting subgroups of the population that are more vulnerable to certain diseases). Predictions are not directly useful for this purpose (although they will turn out to be indirectly useful - see Section 3). I therefore see a slightly different divide into two model cultures than Breiman (2001) within this broader domain of problems. I will discuss these two cultures, each with their strengths and weaknesses, so as to find a compromise in Section 3.

2.1 The tradition within computer science

The tradition within computer science is to let the data speak, by making use of flexible, black-box models. Prioritising correctness is obviously a good thing, but comes with the risk of failing to summarise the data in an interpretable way. With the exception of pure prediction and classification problems, providing insight is nonetheless the main purpose of a data analysis. This partly motivates the increasing calls for machine learning algorithms to be 'interpretable' or 'explainable' (Molnar, 2020) - see Rudin (2019) for a relevant discussion - and the reluctance of statisticians to adopt these techniques outside of the context of prediction. This reluctance is also motivated by a number of more subtle concerns on which I will expand in Section 3.2.

2.2 The tradition within statistics

The tradition within statistics is to use (semi)parametric models with the primary aim to provide insight into data. It is a successful tradition in that sense. Note, for instance, the success of regression models for describing conditional associations between possibly continuous exposures and outcomes, and of random effect models for delivering insight into complex data structures. Such modelling leads to a simplification of reality, which is unavoidable and even desirable when the primary aim is to 'summarise' and provide insight. Indeed, when studying the (conditional) association between a continuous exposure and a continuous outcome, we would often not be interested to know exactly what the expected outcome is at each exposure level; we would usually content ourselves with - and even prefer - a linear or other approximation that is easy to interpret and communicate. The problem with the modelling tradition lies not in its use of models with the aim to approximate the patterns in the data, but rather in its positioning of the model above everything else. This is damaging in at least the following two broad ways.

First, many traditional statistical analyses focus more on the model than on the problem one is trying to solve (Breiman, 2001). This should be no surprise if one considers how textbooks and classes on regression are typically organised. We are trained to use logistic regression for dichotomous outcomes and Cox regression for survival endpoints. We are not trained to first think what estimand (i.e., association or effect measure) is relevant for answering the scientific question at stake; note, for instance, the generally poor understanding that the meaning of many regression coefficients changes due to non-collapsibility as one adjusts for more and more covariates (even in the absence of confounding, selection or mediation effects) (Greenland et al., 1999). We are moreover trained to view a fitted regression model as a representation of the data-generating mechanism, and to report all of its regression coefficients in a table. We are not trained that inferring the data-generating mechanism is an overly ambitious undertaking because different regression models will often fit the data nearly equally well; the latter is commonly mentioned in textbooks, but then subsequently 'forgotten' in how we report the model and draw inference. Ideally, the statistical analysis (and model) should be less ambitious, being centred around the one (or few) exposure variable(s) one wishes to relate to outcome, with careful consideration of which additional variables should versus should not be adjusted for (which may be different depending on the considered exposure) (Westreich and Greenland, 2013). This practice has become mainstream within causal inference (Hernan and Robins, 2020), but would often be regarded with suspicion outside of it.

Undue reliance on models is also evident from the abundant focus on estimands that cannot be defined in a model-free way. One example is the popular product-of-coefficient method in mediation analysis, where the indirect effect of an exposure on an outcome via a given mediator is defined as a product of the exposure coefficient in a model for the mediator, times the mediator coefficient in a model for the outcome. Such model-based definition is problematic because no theory justifies that a product of coefficients can generally be viewed as an indirect effect, rendering it essentially meaningless in many cases (Robins and Greenland, 1992). Similar problems are seen in many other contexts (e.g., instrumental variables estimation). The resulting lack of interpretability - consider for instance how difficult it would be to interpret the product of two log odds ratios when

mediator and outcome are dichotomous - can only be resolved by starting the analysis with the choice of an interpretable estimand (see Section 3 how this can be done), rather than the choice of a model.

Second, there is a great danger in drawing false conclusions when viewing the fitted model as a representation of the ground truth (i.e., as a data-generating model). Misspecification of (semi)parametric models is indeed likely. For instance, when estimating the effect of an exposure on an outcome, the high dimensionality of confounders makes it difficult to postulate parametric models for exposure or outcome. Model misspecification is a serious threat to causal inference in particular, as it necessitates counterfactual predictions (Hernán et al., 2019): predictions of what the outcome would have been for an unexposed individual, had s/he been exposed. As exposed and unexposed individuals may have very different characteristics, these counterfactual predictions can be prone to extrapolation. The problem of model misspecification is especially severe when there is little overlap between exposure groups. Even mild model misspecifications may then induce severe bias whilst being difficult to diagnose (Vansteelandt and Daniel, 2014). In particular, they may not be signaled by standard machine learning metrics such as mean squared (prediction) error, which evaluate prediction error over random samples where only minor extrapolations are made. The resulting model misspecification bias or extrapolation bias is nearly always ill understood. In particular, standard estimators of the (conditional) association (or effect) between an exposure on an outcome will then typically converge to a limit that is no longer even capturing the intended association, e.g., it may suggest a conditional association when in truth there is none, or vice versa. Even if it does, then overly optimistic inferences are typically obtained (notably even when robust standard errors are used). This is the result of excess variability that most estimators exhibit when models are misspecified (Buja et al., 2019; Vansteelandt and Dukes, 2020) or when variable selection procedures are employed to construct a well-fitting model (Leeb and Pötscher, 2006; Dukes and Vansteelandt, 2020).

3. Bridging the two cultures

In view of the concerns about model misspecification raised in the previous section, there has been a rapidly growing interest over the past two decades within primarily the causal inference literature in harnessing the power of data-adaptive methods for inferring the effects of exposures on outcomes (e.g., van der Laan (2015); Mooney and Pejaver (2018); Hernán et al. (2019)). Here, I will focus on a 'data-adaptive method' as being any method that uses the data to learn structure (as opposed to fitting a pre-specified parametric model without any element of variable selection, model building, ...) and can be automated, so that it can be pre-specified without ambiguity. This includes random forest regression, support vector machines, gradient boosting, ... but also flexible parametric models with variable selection (e.g., stepwise variable selection, lasso, ...), with or without the inclusion of splines, which can often give competitive prediction performance (Rudin, 2019). Below, I will review these developments (van der Laan and Rose, 2011; Chernozhukov et al., 2018), which I believe are transforming the way how we will estimate associations and effects in the future. In the next section, I will argue that these developments accommodate Breiman's concerns.

3.1 Step 1. Choosing an estimand

The starting point of these developments is not the choice of a model, but the choice of an estimand. Such estimand represents an interpretable summary that we wish to target in the analysis (van der Laan and Rose, 2011; Hernan and Robins, 2020). It is connected to the scientific question in the sense that knowing the value of the estimand helps answer that question. It is defined in a model-free way so that it can be unambiguously specified well before seeing the data. In that sense, typical regression parameters cannot be viewed as estimands, as their meaning is dependent on the choice of model and variables that we include in it, which often is not pre-specified.

The choice of an estimand forms a natural starting point within the causal inference literature (Daniel et al., 2016). For instance, the effect of a dichotomous exposure A (coded 0 or 1) on a dichotomous outcome Y is often quantified as $E(Y^1 - Y^0)$, where $Y^a, a = 0, 1$ denotes the potential outcome that would have been seen for a randomly chosen individual if the exposure of that individual had been set to a (Hernan and Robins, 2020). Given a pre-exposure covariate vector L that is sufficient to adjust for confounding (and the consistency assumption that $Y = Y^a$ if $A = a$), this estimand can be identified (or linked to the observed data distribution) as (Hernan and Robins, 2020)

$$E(Y^1 - Y^0) = E\left\{E\left(Y|A = 1, L\right) - E\left(Y|A = 0, L\right)\right\}. \tag{1}$$

Choosing an estimand is less common outside of the causal inference literature, but much more broadly conceivable. For instance, the conditional association between a possibly continuous exposure A and an outcome Y, given a specific vector of covariates L, can be quantified as

$$\frac{E\left[\mathrm{Cov}\left\{A, E\left(Y|A, L\right)|L\right\}\right]}{E\left[\mathrm{Var}\left(A|L\right)\right]} \tag{2}$$

or, when Y is dichotomous, as

$$\frac{E\left[\mathrm{Cov}\left\{A, \mathrm{logit}E\left(Y|A, L\right)|L\right\}\right]}{E\left[\mathrm{Var}\left(A|L\right)\right]} \tag{3}$$

(Vansteelandt and Dukes, 2020). These estimands appears less insightful, but (2) reduces to the standard mean difference β when

$$E\left(Y|A, L\right) = \beta A + \omega(L),$$

for some unknown function $\omega(.)$, and likewise (3) reduces to the standard log odds ratio β when

$$\mathrm{logit}E\left(Y|A, L\right) = \beta A + \omega(L),$$

for some unknown function $\omega(.)$. Both estimands thus have a clear connection to standard regression models and parameters with familiar interpretation. However, they continue to be relevant when the above models do not hold. In that case, (2) and (3) are obtained by summarising, for each level l of L, the dependence of $E\left(Y|A, L = l\right)$ or $\mathrm{logit}E\left(Y|A, L = l\right)$ onto A by means of a population-least-squares projection (i.e., a linear approximation)

among individuals with $L = l$, and next taking a weighted average of the resulting l-specific association measures using the weights

$$\frac{\text{Var}\,(A|L=l)}{E\left[\text{Var}\,(A|L)\right]}. \tag{4}$$

In the special case where A is dichotomous, (2) can therefore be written as a weighted average of mean differences:

$$E\left[W(L)\left\{E(Y|A=1,L)-E(Y|A=0,L)\right\}\right]$$

for weights

$$W(L)=\frac{P(A=1|L)P(A=0|L)}{E\left\{P(A=1|L)P(A=0|L)\right\}};$$

likewise, (3) can be written as a weighted average of conditional log odds ratios:

$$E\left[W(L)\left\{\text{logit}E(Y|A=1,L)-\text{logit}E(Y|A=0,L)\right\}\right].$$

In this formalism, statistical models remain useful to assign meaning to a complex estimand, but the estimand continues to capture the intended conditional association when the model is misspecified. Furthermore, by having thus preset the choice of association measure that will be obtained from the analysis, we have control over the interpretability of the analysis result, regardless of the complexity that we wish to invoke in the next section for estimating $E(Y|A=a,L=l)$ or $E(A|L=l)$ for given a and l (via machine learning or flexible parametric models).

3.2 Step 2. Estimating the estimand

It is tempting to employ flexible parametric models with variable selection, or even existing machine learning algorithms in the estimation of the unknown nuisance parameters (e.g., $E(Y|A=a,L=l)$ or $E(A|L=l)$ for given a,l) characterising a given estimand. For instance, to infer (1), one may train machine learning algorithms for the outcome in exposed as well as unexposed individuals, and then average the difference in predictions obtained for all individuals:

$$\frac{1}{n}\sum_{i=1}^{n}\hat{E}\left(Y_i|A_i=1,L_i\right)-\hat{E}\left(Y_i|A_i=0,L_i\right).$$

Likewise, to infer (2) or (3), one may fit flexible parametric models (possibly involving interactions, higher-order terms, splines, ...) or train machine learning algorithms for exposure and outcome, and then calculate

$$\frac{\sum_{i=1}^{n}\left\{A_i-\hat{E}(A_i|L_i)\right\}\left\{\hat{E}(Y_i|A_i,L_i)-\hat{E}(Y_i|L_i)\right\}}{\sum_{i=1}^{n}\left\{A_i-\hat{E}(A_i|L_i)\right\}^2}$$

or

$$\frac{\sum_{i=1}^{n}\left\{A_i-\hat{E}(A_i|L_i)\right\}\left(\text{logit}\left\{\hat{E}(Y_i|A_i,L_i)\right\}-\hat{E}\left[\text{logit}\left\{\hat{E}(Y_i|A_i,L_i)\right\}|L_i\right]\right)}{\sum_{i=1}^{n}\left\{A_i-\hat{E}(A_i|L_i)\right\}^2}, \tag{5}$$

respectively, where $\hat{E}\left[\text{logit}\left\{\hat{E}(Y_i|A_i, L_i)\right\}|L_i\right]$ denotes a prediction of logit $\left\{\hat{E}(Y_i|A_i, L_i)\right\}$ based on L_i, where $\hat{E}(Y_i|A_i, L_i)$ represents an outcome prediction. This is indeed attractive because it enables arbitrarily complex models or algorithms to be used without distorting the interpretation of the end result. However, these estimators will typically do a poor job because the variable selection or machine learning algorithms on which they rely have been optimally tuned for prediction, but not for estimating association measures (van der Laan and Rose, 2011). For instance, the above estimator of (1) may not work well when untreated individuals are quite different from treated individuals in their measured characteristics, as the used algorithms may not be ideally tuned towards the need for extrapolation. Standard assessments of prediction error give no insight into this. In particular, the accuracy of the (counterfactual) predictions $\hat{E}(Y|A = a, L)$ for $a = 0, 1$ cannot be readily assessed on validation samples, since one can only observe how untreated individuals would fare without treatment, but not with treatment. This is even more problematic when considering the above estimators of (2) or (3), where it becomes difficult or even impossible to assess whether the obtained predictions are sufficiently accurate in those parts of the data space that matter for the accuracy of this estimator. A further problem is that the set of features that is optimal for inclusion when the aim is prediction, may not be optimal when the aim is to estimate effects or associations. For instance, the omission of variables that are strongly predictive of exposure but moderately predictive of outcome may result in estimators of their association that have substantial bias. When the aim is prediction instead, then their omission may reduce prediction error by mitigating multicollinearity.

Use of the above estimators is further hindered by the uncertainty of predictions obtained using variable selection or machine learning algorithms. This uncertainty is difficult to quantify (except at a population level in terms of average prediction error), and so is the extent to which this uncertainty propagates into the estimator of the considered estimand. While the bootstrap may appear as an attractive way out, it has no validity in high-dimensional settings (Leeb and Pötscher, 2006; Dukes and Vansteelandt, 2020). In such settings, variable selection and machine learning procedures tend to deliver non-regular estimators, whose behaviour is sensitive to even minor changes in the data-generating mechanism. This makes them in particular sensitive to the distinction between the empirical versus population distribution of the data, a distinction which the bootstrap 'ignores'. This is troublesome as it is impossible to come to well-informed decisions without reliable measures of uncertainty.

The above problems partly motivate statisticians' reluctance to adopt machine learning procedures. They make the development of variable selection and machine learning methods for the estimation of association or effect measures dependent upon results from asymptotic statistics. Such results have been developed in the literature on nonparametric statistics (Pfanzagl, 1990; Bickel et al., 1993) and have paved the way for revolutionary developments on targeted learning (van der Laan and Rubin, 2006; van der Laan and Rose, 2011) and double machine learning (Chernozhukov et al., 2018). In particular, nonparametric inference for the considered estimand can proceed based on its so-called canonical gradient (Pfanzagl, 1990; Bickel et al., 1993) under the nonparametric model. For the estimands (2) and (3), this is given by (Vansteelandt and Dukes, 2020)

$$\frac{\{A - E(A|L)\}\left[\mu(Y, A, L) - \beta\{A - E(A|L)\}\right]}{E\left[\{A - E(A|L)\}^2\right]}$$

where

$$\mu(Y,A,L) \equiv Y - E(Y|L)$$

for estimand (2), and

$$\mu(Y,A,L) \equiv \frac{\{Y-E(Y|A,L)\}}{E(Y|A,L)\{1-E(Y|A,L)\}} + \text{logit}\{E(Y|A,L)\} - E[\text{logit}\{E(Y|A,L)\}|L]$$

for estimand (3). An estimator can now be obtained as the value of β that makes the sample average of these canonical gradients equal to zero, in which the 'unknownsÕ can be estimated using data-adaptive procedures (e.g., variable selection, machine learning). For instance, an estimator of the 'linear regression' estimand (2) is given by

$$\frac{\sum_{i=1}^n \left\{A_i - \hat{E}(A_i|L_i)\right\}\left\{Y_i - \hat{E}(Y_i|L_i)\right\}}{\sum_{i=1}^n \left\{A_i - \hat{E}(A_i|L_i)\right\}^2},$$

for given exposure and outcome predictions, $\hat{E}(A_i|L_i)$ and $\hat{E}(Y_i|L_i)$, respectively. An estimator of the 'logistic regression' estimand (3) is obtained by adding to the 'plug-in' estimator (5) the following bias correction term

$$\frac{\sum_{i=1}^n \left\{A_i - \hat{E}(A_i|L_i)\right\}\left\{Y_i - \hat{E}(Y_i|A_i,L_i)\right\} / \left[\hat{E}(Y_i|A_i,L_i)\left\{1-\hat{E}(Y_i|A_i,L_i)\right\}\right]}{\sum_{i=1}^n \left\{A_i - \hat{E}(A_i|L_i)\right\}^2}$$

for given exposure and outcome predictions, $\hat{E}(A_i|L_i)$ and $\hat{E}(Y_i|A_i,L_i)$, respectively.

The resulting estimators possess a small-bias property (Newey et al., 2004), which means that their bias converges to zero faster (as the sample size grows larger) than the (regularization) bias that affects the data-adaptive predictions. Under relatively weak conditions on the rate of convergence of these predictions towards the truth, and provided that sample splitting (Zheng and van der Laan, 2011; Chernozhukov et al., 2018) is used (whereby the estimator is calculated on a different part of the data than that on which the data-adaptive predictions are trained), standard root-n convergence to a normal limit is then obtained, with the estimation errors in the predictions not affecting the asymptotic behaviour of the resulting estimator. Interestingly, this implies that inference can proceed as if those predictions were a priori known, rather than data-driven. The obtained estimators are therefore asymptotically unbiased and come with (uniformly) valid confidence intervals that are obtainable through relatively simple analytical calculations. In particular, standard errors can be consistently estimated as one over root-n times the sample standard deviation of the canonical gradients, evaluated at the obtained estimates and predictions. This is extremely advantageous, given that the uncertainty in the machine learning predictions is so difficult to assess.

4. Discussion

We have gone a long way since Breiman's seminal paper, and I am positive about what is yet to come. I remember the many frustrations as a young student that several strategies

of building a regression model would lead to quite different results, leaving me paralysed what to report. I remember discussions whether or not to use Cox regression models with or without covariate adjustment in randomised experiments, not realising that the comparison was muddled by the difference in estimands targeted by Cox regression models with versus without covariates. Many of those concerns and confusions fade when starting data analyses with the choice of an estimand. This helps ensure that the data analysis targets the scientific question. It moreover prevents data analysts from making models overly simplistic on purpose just to secure a simple and easy-to-communicate result. Indeed, the analysis strategy of Section 3 targets the same estimand, no matter the complexity of the underlying data-generating mechanism. It thereby enables the use of flexible parametric models with variable selection, splines, ... or even machine learning. This flexibility attenuates model misspecification bias and ensures that the analysis is purely evidence-based, can be pre-specified and, unlike standard model-based analyses, acknowledges all uncertainties related to model building, variable selection, ... Arguably, it is harder to incorporate biological knowledge into machine learning procedures, and I believe that hand-made statistical models therefore remain valuable. But to minimise the concerns about models raised in Section 2.2, it seems wise to combine the resulting hand-made predictions with objective, algorithmic procedures via the use of ensemble learners (Van der Laan et al., 2007).

The use of machine learning (or more general data-adaptive procedures) for estimating associations and effects as described in Section 3, accommodates the three key concerns raised by Breiman (2001). It addresses his concerns about the 'multiplicity of good models' (i.e., the fact that multiple models may be fitting the data nearly equally well) because different well-fitting models will often generate rather similar predictions; these predictions are all that is being employed by the considered estimators in Section 3. Moreover, the possibility to aggregate over a large set of competing models or algorithms when obtaining machine learning predictions (Van der Laan et al., 2007), enables one to make well-informed compromises. Further, the strategy of Section 3 overcomes the tension between using complex, correct models versus interpretable, misspecified models. The reason is that the same 'simple' estimand is targeted no matter the complexity of the predictions consumed by the estimation procedure. Breiman's final concern about the need for model-based strategies to drastically reduce the number of covariates in high-dimensional settings is readily accommodated via the use of machine learning.

Despite these important developments, pioneered by van der Laan and Rubin (2006) (see van der Laan (2015) for a gentle overview), we still have a long way to go in accommodating Breiman's concerns. Much work remains to be done to extend this methodology, which is almost exclusively developing within the domain of causal inference, to more general settings where regression models are commonly used and less ambitious association measures are of interest. To change applied practice, the development of user-friendly software that can handle a wide array of estimands is key (Gruber and Van der Laan, 2011). Critical is also to rethink our statistics education, which is heavily centred around the use of models. In particular, I believe we must train our students to translate scientific questions into interpretable estimands, that may or may not be linked to a statistical model. This is more relevant and careful than instead training them (willingly or unwillingly) to view the fitted model as the primary result of the statistical analysis, especially considering that different models may fit the data nearly equally well. I believe we must restore the statistical model

in its original purpose to provide insight, rather than to reflect some ground truth in terms of how data have been generated by nature. Rethinking our statistics education is a challenging endeavour that need not make it more mathematical, quite on the contrary. Indeed, many introductory statistics courses now continue to have a strong focus on collections of test statistics and their distributions, standard error and confidence interval derivations and formulae, ... which is useful for the minority who will engage in methods development, but not for the majority who will end up analysing data. In times where software is abundant and accessible, the focus of introductory statistics courses should be primarily on statistical reasoning, as made most clear by computational, simulation-based methods such as bootstrap, permutation tests, simulation-based sample size calculation, ..., on concepts of bias (e.g., selection bias, confounding bias) and imprecision, on the translation of scientific questions into statistical estimands, on key assumptions linked to study design (e.g., independence assumptions), on flexible (statistical or machine) learning methods for prediction, which form a cornerstone of the methods of Section 3, ... For the smaller minority of students who are mathematically-minded and/or wish to engage in methods development, a core training in asymptotic statistics obviously remains indispensable. I believe that a central focus of such training should lie on nonparametric estimation and efficiency theory (Bickel et al., 1993; Kennedy, 2016; Fisher and Kennedy, 2020).

Acknowledgments

I wish to thank the Editor for the opportunity to contribute, and hope this contribution will help stimulate the reader to join others and myself in this exciting endeavour. I am also grateful to Josephina Argyrou, Oliver Dukes, Pawel Morzywolek and Jan De Neve for helpful discussions.

References

Peter J Bickel, Chris AJ Klaassen, Peter J Bickel, Ya'acov Ritov, J Klaassen, Jon A Wellner, and YA'Acov Ritov. *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore, 1993.

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

Andreas Buja, Lawrence Brown, Richard Berk, Edward George, Emil Pitkin, Mikhail Traskin, Kai Zhang, and Linda Zhao. Models as Approximations I: Consequences Illustrated with Linear Regression. *Statistical Science*, 34(4):523–544, November 2019. ISSN 0883-4237. doi: 10.1214/18-STS693.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, February 2018. ISSN 1368-423X.

Rhian M Daniel, Bianca L De Stavola, and Stijn Vansteelandt. Commentary: The formal approach to quantitative causal inference in epidemiology: misguided or misrepresented? *International journal of epidemiology*, 45(6):1817–1829, 2016.

Oliver Dukes and Stijn Vansteelandt. How to obtain valid tests and confidence intervals after propensity score variable selection? *Statistical methods in medical research*, 29(3): 677–694, 2020.

Aaron Fisher and Edward H Kennedy. Visually communicating and teaching intuition for influence functions. *The American Statistician*, pages 1–11, 2020.

Sander Greenland, James M Robins, and Judea Pearl. Confounding and collapsibility in causal inference. *Statistical science*, pages 29–46, 1999.

Susan Gruber and Mark J Van der Laan. tmle: An r package for targeted maximum likelihood estimation. 2011.

Johanna Hardin, Roger Hoerl, Nicholas J Horton, Deborah Nolan, Ben Baumer, Olaf Hall-Holt, Paul Murrell, Roger Peng, Paul Roback, D Temple Lang, et al. Data science in statistics curricula: Preparing students to Òthink with dataÓ. *The American Statistician*, 69(4):343–353, 2015.

Miguel A Hernan and James M Robins. *Causal inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020.

Miguel A Hernán, John Hsu, and Brian Healy. A second chance to get causal inference right: a classification of data science tasks. *Chance*, 32(1):42–49, 2019.

Edward H Kennedy. Semiparametric theory and empirical processes in causal inference. In *Statistical causal inferences and their applications in public health research*, pages 141–167. Springer, 2016.

Hannes Leeb and Benedikt M Pötscher. Performance limits for estimators of the risk or distribution of shrinkage-type estimators, and some general lower risk-bound results. *Econometric Theory*, pages 69–97, 2006.

Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

Stephen J Mooney and Vikas Pejaver. Big data in public health: terminology, machine learning, and privacy. *Annual review of public health*, 39:95–112, 2018.

Whitney K. Newey, Fushing Hsieh, and James M. Robins. Twicing Kernels and a Small Bias Property of Semiparametric Estimators. *Econometrica*, 72(3):947–962, 2004. ISSN 0012-9682. URL `https://www.jstor.org/stable/3598841`.

Johann Pfanzagl. Estimation in semiparametric models. In *Estimation in Semiparametric Models*, pages 17–22. Springer, 1990.

James M Robins and Sander Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, pages 143–155, 1992.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Mark van der Laan. Statistics as a science, not an art: the way to survive in data science. *Amstat News*, 1, 2015.

Mark J. van der Laan and Sherri Rose. *Targeted Learning*. Springer Series in Statistics. Springer New York, New York, NY, 2011. ISBN 978-1-4419-9781-4 978-1-4419-9782-1. doi: 10.1007/978-1-4419-9782-1.

Mark J. van der Laan and Daniel Rubin. Targeted Maximum Likelihood Learning. *The International Journal of Biostatistics*, 2(1), 2006. ISSN 1557-4679. doi: 10.2202/1557-4679.1043.

Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

Stijn Vansteelandt and Rhian M Daniel. On regression adjustment for the propensity score. *Statistics in medicine*, 33(23):4053–4072, 2014.

Stijn Vansteelandt and Oliver Dukes. Assumption-lean inference for generalised linear model parameters. *arXiv preprint arXiv:2006.08402*, 2020.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Daniel Westreich and Sander Greenland. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *American journal of epidemiology*, 177(4):292–298, 2013.

Wenjing Zheng and Mark J. van der Laan. Cross-Validated Targeted Minimum-Loss-Based Estimation. In *Targeted Learning*, pages 459–474. Springer New York, New York, NY, 2011. ISBN 978-1-4419-9781-4 978-1-4419-9782-1.